National College of Ireland

BSc in Computing

Data Analytics

2017/2018

Robbie Jenkinson

X14379546

X14379546@student.ncirl.ie

Irish Tourism Trends

Technical Report

Declaration Cover Sheet for Project Submission

SECTION 1 *Student to complete*

Name: Robbie Jenkinson
Student ID: X14379546
Supervisor: Simon Caton

SECTION 2 Confirmation of Authorship

The acceptance of your work is subject to your signature on the following declaration:

I confirm that I have read the College statement on plagiarism (summarised overleaf and printed in full in the Student Handbook) and that the work I have submitted for assessment is entirely my own work.

Signature: Robbie Jenkinson Date: 09/05/2018

Robbie Jenkinson 09/05/2018

Table of Contents

Table of Figures

Executive Summary

How did the recession impact airports and the Irish tourism industry? What are the most popular airports that fly to Ireland? And, how much do tourists spend when they come to Ireland? These are among the many questions that are answered in the analysation of the Irish tourist industry with the developed analytics platform Irish Tourism Trends. Being able to analyse and see why people come to Ireland and how much they're spending gives key information to governments and councils to target and use to their advantage. The Irish tourism industry is key to the Irish economy thriving at this moment in time. However, with data giving priceless insights to companies and sports teams for a competitive advantage, the same must be done for the Irish tourism industry and act on specific trends that are occurring. This report and project addresses this issue with the use of data from the CSO as the data source.

Keywords: Tourism, Ireland, Data Analytics, Future, Information, The CSO.

1 Introduction

The Irish Tourism Trends platform is created to give key insights and information to the user in the form of interactive and informative visualisations on two key areas, the passenger intake into Irish airports and where the most passengers are coming from, and the visitors area which uncovers why exactly people are coming to Ireland, where they're coming from and how much money they're spending when they come to Ireland. The overall project uncovers some striking stats which have not been previously known, such as correlations, the busiest month in Ireland and the impact of the recession on our tourism industry.

The project is extremely important in the context that the tourism industry is vital to Irelands booming economy. It also uncovers trends in tourism and passenger intake that many wouldn't have known before. It can also provide vital information as to what is important for the Irish economy to keep going the way it is. This information can provide key knowledge to the tourist industry in Ireland as well as governments and local councils.

1.1 Background & Motivation

After seeing my town of Skerries win the tidiest town in Ireland in 2016, the tourism erupted in the town. Housing prices in Skerries were reported to have gone up purely because of the award of tidiest town in Ireland. The Infrastructure on the harbour (Most popular area of Skerries) boomed, where cafes, restaurants and shops popped up. It's intriguing as to why some places have droves of tourists per year and others don't. With the analysation of the data that the CSO has released since 2007, many questions about Irish tourism can be answered. Not only can this be helpful for a user to learn about the industry, but it can be even more helpful for local governments to see the trends for themselves and make educated decisions based of the visualised information.

1.2 Aims

The aims of the Irish Tourism Trends platform are to answer any question that a user may have about the Irish tourism industry whether that’s about how the recession impacted the intake of passengers into airports or how Brexit could potentially impact our economy, the visualisations allow for more than just one question at a time to be answered as each individual user will have their own curiosities about the industry. The platform will also be deployed to a cloud-based server where users can access this information anytime they wish which is an extremely important function of the platform.

1.3 Technologies

R – The language R is used for the scraping, processing, and transformation of data. It can only be used in a software platform called RStudio which has four windows to use while analysing data. The data window where the datasets, subsets, and any value is held, the console where anything that gets run pops up, the utility window which can show packages, plots and much more and the fourth window which is where the coding takes place (R-project.org, n.d.).

Shiny – A shiny app allows multiple data visualizations to be put together and run on a platform/webpage where the user can go and interact with the data, whereas a normal webpage wouldn’t have interactive data. A shiny app is very dynamic and can be deployed to the cloud on the Shiny apps platform which hosts the app and allows users to access the platform and view and interact with the data visualisations (RStudio, n.d.).

2 System

2.1 *Requirements*

2.1.1 Functional requirements

Functional requirements have changed quite a lot since the beginning of development of the application. Originally this application was to be a platform developed in PHP, it is now a Shiny application and so this has changed each use case diagram and requirement but also the process of data mining has changed from KDD to CRISP-DM as the function of this application has evolved.

Online platform to display data and findings by the using visualisations. The user interface needs to be fully functioning if the project is to be successful and the user needs to be able to navigate through the site properly to view the different data visualizations. It is essential that each visualization is appropriately visualised to allow full interaction of the user with the app.

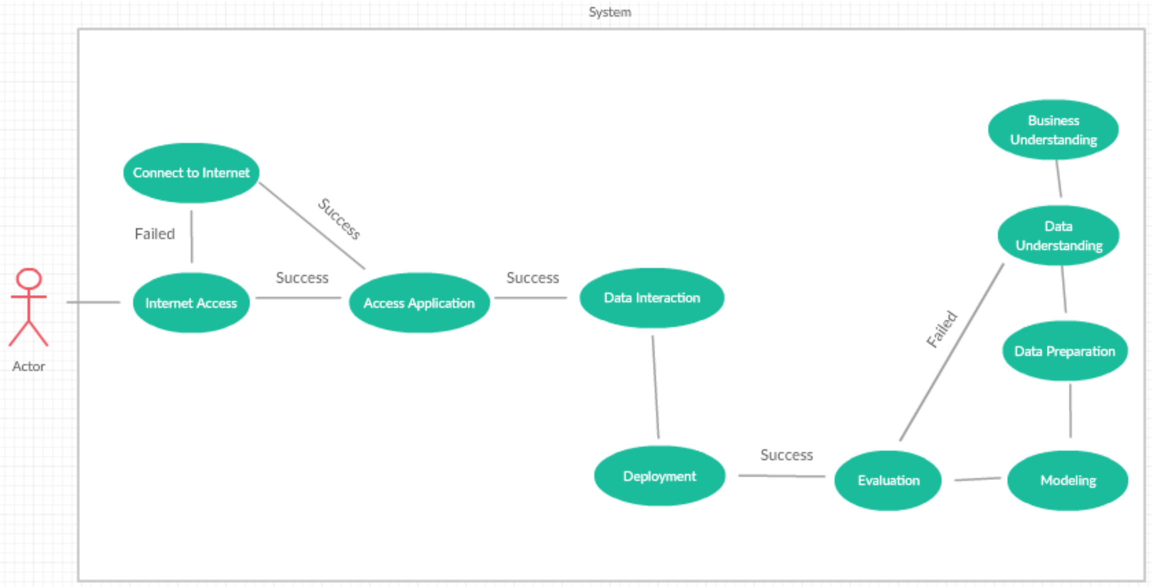


Figure 1: Functional Req. Use Case

2.1.1.1 Requirement 1: User Access

2.1.1.1.1 Description & Priority

The user must have access to the internet to access the online application platform. They must enter the URL of the application correctly into the search bar in the browser or the browser will return a blank page if the URL should be wrong. Once the user has correctly entered the URL into the browser they will be brought to the homepage of ‘Irish tourism trends’ where they can then interact with the platform.

2.1.1.1.2 Use Case

Scope

The scope of this use case is to clearly define how important it is for the user to have access to the internet and have a device that can access the internet for them to access the online application.

Description

This use case describes why it is necessary for the user to have access to the internet to access the online application.

Use Case Diagram

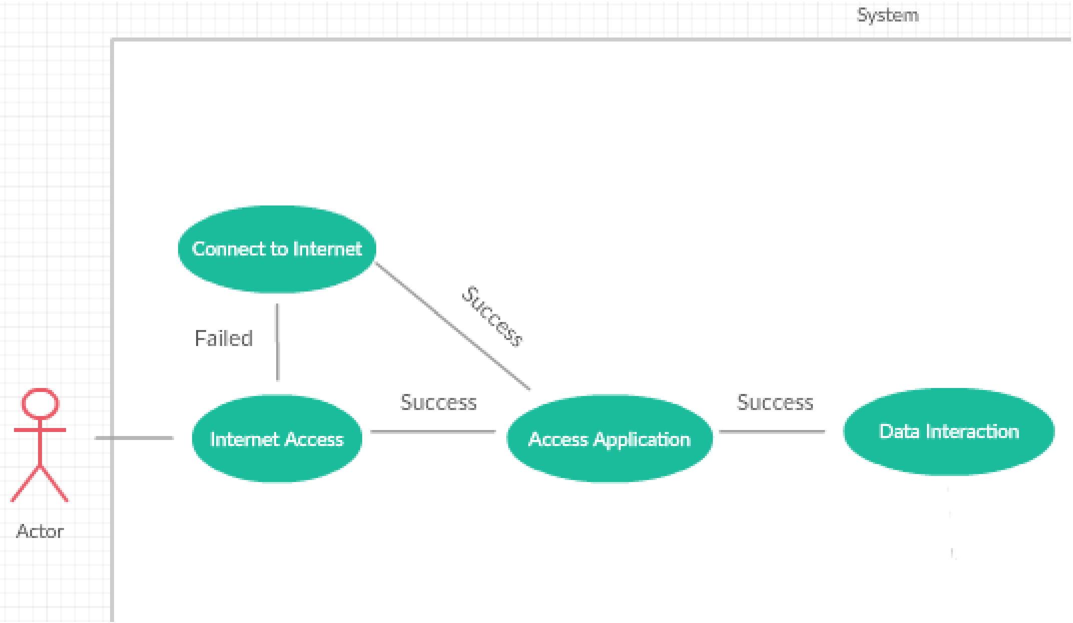


Figure 2: User Access Use Case

Flow Description

Precondition

The user is in the company of a device which can connect to the internet.

Activation

The use case starts when the user powers on the device capable of connecting to the internet.

Main Flow

1. The user navigates to a browser of their choice.
2. The user identifies the URL needed to access the online application and inserts it into the URL bar.
3. The user hits the return button and is brought to the online application landing page.
4. The user navigates through the system, interacting with the visualized data.

Alternate Flow

A1: The user fails to reach the online application landing page

1. The user has incorrectly entered the URL into the task bar and is denied access to any webpage.
2. The user must re-enter the URL into the task bar.
3. The user correctly enters the URL into the task bar and is granted access to the online application.

Exceptional Flow

E1: The user cannot connect to any webpage

1. The user has not been granted internet access and cannot access the internet and any online application.
2. The user powers off the machine.

Termination

The use case is terminated when the user leaves the platform.

Post condition

The system goes into a wait state.

2.1.2 Data requirements

To fulfil the objectives of the project, it is important to comply with the CRISP-DM process. After the business objectives have been identified to examine and analyse the Irish tourism industry, the next and arguably most important part of the process is the data understanding stage. It's important that the potential data that will be used be verified and explored to the extent that the data will sufficiently enable the achievement of the business objectives. It was established that data from The CSO (central statistics office) would allow for the execution of analysing the Irish tourist industry. This data source is also extremely trustworthy and accurate in their data presentation.

Quite a lot of time needed to be allocated to the cleaning of the scraped data from The CSO as the data tables that presented the data were untidy. Creating data frames that could be easily queried was more difficult than first thought, but in the end was achieved.

The use of an SQL database enabled the insertion of such data frames to save as a finite state. This meant it was possible to pull on the data and create subsets and merge data without disrupting the main datasets in the SQL database. This was also a key functionality as any new data that is scraped can easily be added to the existing datasets in the database, while the data can be pulled from the SQL database into RStudio to be manipulated into sub sets for the different visualisations of charts, maps and graphs on the Irish Tourism Trends Platform.

2.1.3 User requirements

For users to be able to access and interact with the visualised data they must have a device that can connect to the internet and accessing webpages on the internet. They must know the URL of the platform to access the app. Once those requirements are met, the user will be granted access to the online platform where they can interact with the visualised charts, figures and maps. It is somewhat important that the user have prior knowledge of some things regarding the Irish tourism industry, for example what area that would be thought to be the busiest or what airport would be thought to be the busiest overall.

2.1.4 Performance/Response time requirement

This online platform is hosted on the shiny cloud server which allows for the deployment of data analytics projects from RStudio to the cloud. This enables the interaction of the user with the data, to change the dynamic of a visualisation, for example of a chart. From the time that the user enters the site, there will be a slight lag of loading every chart, which will be no longer than 10 seconds. The performance and response time of the platform will be very quick from then on as the data is hosted in the cloud also which makes it quicker for interaction.

2.1.5 Availability requirement

As mentioned in the performance/response time requirement section, the app is hosted on shiny apps server which allows the availability of the app around the clock. However, for a student it is tough to pay hosting once the usage time of the app goes past 24 hours, as it’s free for usage time under 24 hours however it is a priced plan from then on. It must be decided after that whether to pay that or not.

2.1.6 Recover requirement

The project will be stored and backed up in 3 places. GitHub allows files and project to be stored in the cloud. It is very efficient as it can be on your local machine where you can push your code to when you’re finished with it for the time being. Shiny apps stores and hosts the project, it is not only where the project can be accessed online but the project can also be downloaded if something were to happen to your local version. And finally, google drive which is a drag and drop file holder. This is very safe and easily accessible if something were to go wrong.

2.1.7 Maintainability requirement

The project can be easily updated and upgraded within RStudio which can then be re-deployed to the cloud within minutes. Shiny apps is integrated into RStudio so the re-deployment of an app after the fixing of a bug or just maintenance is incredibly easy. The code must be organized efficiently however, as shiny apps won't run the code the same way the local machine will run the code which works perfectly. Working directory must be identified and the correct path to the data which will be in the shiny folder along with the server.r and ui.r file.

2.1.8 Extendibility requirement

The maintenance of the Shiny app online is incredibly easy when done and laid out correctly, this also makes the extendibility of the app easy too. The ui.r file handles the lay out of the app so in order to add another tab for example, restaurants, it would only take a few lines of code and a push of the re-publish button to re-deploy it to the cloud.

2.1.9 Reusability requirement

Any code used to visualise the data can be easily re-used and transformed to visualise a completely different aspect of data. The package 'plotly' for example is an exciting tool that can be integrated into a shiny app and allows for the reusability of any code previously written.

2.1.10 Resource utilization requirement

The data source that will be scraped is the CSO, a well trusted databank and well known to have accurate data. The CRISP-DM stage of data understanding points out to scrape trustworthy data and data that will enable the fulfilment of the business objectives. The CSO is a very reliable data source which enables the use of their data under the Creative Commons act.

2.2 Data Process Methodology

The data process as previously mentioned was changed from KDD to CRISP-DM which is more suited to the requirements of the project. It is more concerned with the findings and informativeness of the system and allows more of a business approach to the data process, the key stages include:

1. Business understanding & determining business objectives is a key aspect to the overall function of the application especially when dealing with important issues such as tourism where questions must be asked that can enable potential faults or problems to be addressed when working with the data & it is where the understanding of the project is critical. It is essential to understand what the overall project should be addressing (Smart Vision - Europe, n.d.).

2. Data understanding, a vital stage after determining business objectives is the stage which involves the identification and obtaining of the data which will allow the business objectives to be fulfilled efficiently. In the case of Irish Tourism Trends, it was identified that The CSO would give the best possible and most accurate data to enable the answering of questions regarding the Irish tourist industry (Smart Vision - Europe, n.d.).

3. Stage three involves the preparation of data. This includes the cleaning and combining of data for different types of data exploration, be it statistical analysis or predictive analysis. The data from The CSO was extremely difficult and time consuming when trying to break the dataset down into a manageable load that could be easily queried (Smart Vision - Europe, n.d.).

4. Modelling is the fourth stage of the CRISP-DM process and involves the deciding of what is the best way to analyse the given data. Data mining techniques such as time series and other such techniques are involved in predictive analysis, and the use of other descriptive statistics such as

visualizations can be vital to explaining the data. In terms of the Irish Tourism Trends platform, the use of time series prediction analysis was used a couple of times where as in reporting stats, random forest, conditional inference tree, and statistical tests were used to analyse the data (Smart Vision - Europe, n.d.).

5. Evaluation is the stage where models are assessed to see if they have in fact met the goals of the analysis. It involves the evaluation of the models related the aims that were specified in the business understanding phase. As with predictive techniques such as a random forest or conditional inference tree, if the models are extremely inaccurate then the results must not be published (Smart Vision - Europe, n.d.).

6. The deploying stage is where the findings of an analysis are released for interpretation of others. In terms of Irish Tourism Trends, the results were published on a shiny application via data visualisations where users can go and interact with the data (Smart Vision - Europe, n.d.).

2.3 Design and Architecture

The system architecture consists of 4 key components which enable the app to be run successfully. The ui.R file is extremely important and can be thought as the front end of the platform. It is used to create the user interface and enhance the user experience when using the platform. The overall structure and visualizations are put in place in this file also. The server.R can be thought of as the engine room of the platform where all the queries and merging of data for visualisations takes place. It uses functions and packages such as ‘plotly’ and ‘leaflet’ to create the visualisations. The overall app combines these two files and creates the Shiny app which shows a user interface which was created in the ui.R file and allows interaction which was created in the server.R file. The final stage is the cloud deployed platform itself. This is a combination of the server.R, ui.R files and all the data used. These files and data are deployed to the cloud which allows users to access the platform online.

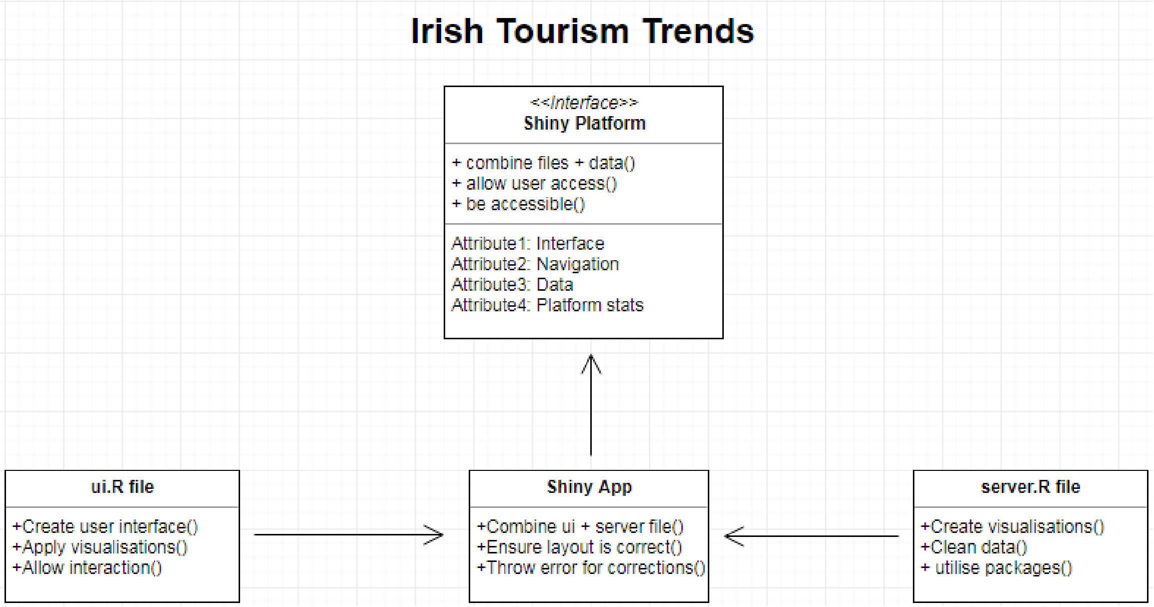


Figure 3: System Architecture

2.4 *Implementation*

The implementation of data analysis in the project followed the stages of the CRISP-DM data process and started with determining the business objectives and what exactly was to be learned from the project and report. Once it was decided what was to be learned from the project and analysing the Irish tourism, it was important to identify a credible data source to scrape and use to fulfil these business objectives.

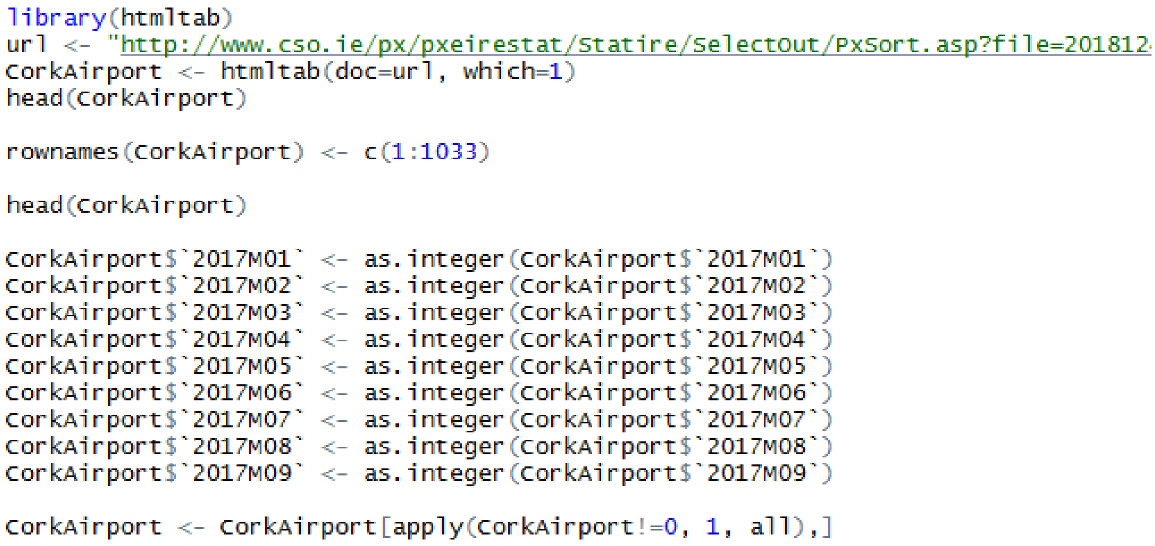
2.4.1 Software tool

RStudio was the chosen software tool to scrape, clean, and analyse the data concerning Irish tourism. RStudio is a data analytics tool which can only be used with the language R. This language is extremely powerful and easy to use when learned correctly. It enables the answering of data related questions and provides tools where data can be visualised for greater understanding. (R-project.org, n.d.).

2.4.2 Scraping & Cleaning data

As the chosen data source, to fulfil the specified business objectives such as The CSO does not provide APIs, it was essential to go in and scrape the data manually using Rstudio libraries rvest & htmltab into RStudio to be cleaned and merged for the analysis that was to follow.

Scraping data from The CSO involved selecting what parameters that were needed. So, it was selected that the last 10 years of data should be returned and only incoming flights from every possible airport to the selected airport in Ireland. For example, the code in Fig.4 shows the scraping of data for Cork airport. The table of data is scraped and cleaned when it enters RStudio. The clean includes changing the months of the year to integer and removing 0's from the data frame. The cleaning process was very labouring as the scraped data came with attribute headers of each month of the year. These months had to be combined into the year they were accompanied by.



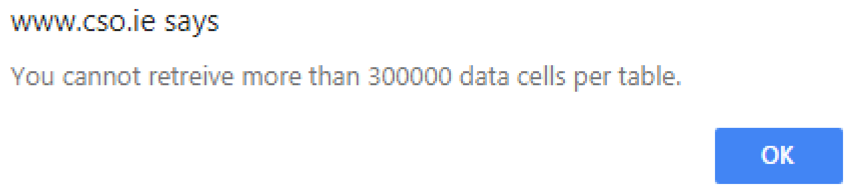
```
library(htmltab)
url <- "http://www.cso.ie/px/pxeirestat/Statire/SelectOut/PxSort.asp?file=201812
CorkAirport <- htmltab(doc=url, which=1)
head(CorkAirport)

rownames(CorkAirport) <- c(1:1033)

head(CorkAirport)

CorkAirport$`2017M01` <- as.integer(CorkAirport$`2017M01`)
CorkAirport$`2017M02` <- as.integer(CorkAirport$`2017M02`)
CorkAirport$`2017M03` <- as.integer(CorkAirport$`2017M03`)
CorkAirport$`2017M04` <- as.integer(CorkAirport$`2017M04`)
CorkAirport$`2017M05` <- as.integer(CorkAirport$`2017M05`)
CorkAirport$`2017M06` <- as.integer(CorkAirport$`2017M06`)
CorkAirport$`2017M07` <- as.integer(CorkAirport$`2017M07`)
CorkAirport$`2017M08` <- as.integer(CorkAirport$`2017M08`)
CorkAirport$`2017M09` <- as.integer(CorkAirport$`2017M09`)

CorkAirport <- CorkAirport[apply(CorkAirport!=0, 1, all),]
```

Figure 4: Scraping Code

This process is then carried out for the remaining airports Ireland such as Shannon, Dublin, Knock etc. All airport data was then merge into one data frame called inwardFlights. This would be the basis for the analysis of the Irish tourist industry regarding inward flights and the amount of people coming to the Island.

2.4.3 Creating Data Frames

From the cleaned data of the CSO, what was then necessary was to create appropriate data frames that would enable the visualisation and analysation of the data. Data scraped from the CSO was very large in size and query parameters to select data on the CSO were sometimes denied as the data being requested was too big.



www.cso.ie says

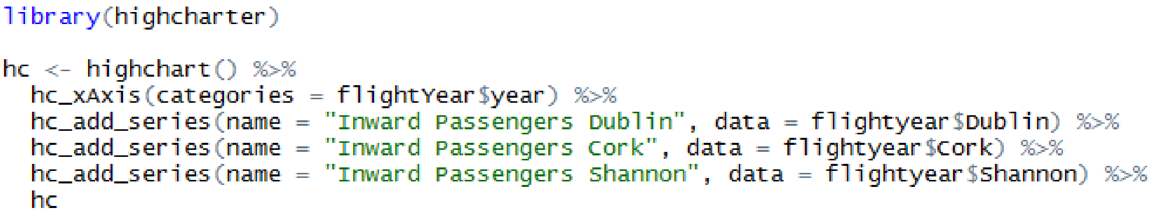
You cannot retreive more than 300000 data cells per table.

Figure 5: Maximum Data Selection

So once each airport was scraped individually and cleaned appropriately, the data could then be combined for the next stages of development to take place. To combine each separate data frame together the rbind() function was used. This function is essentially binding data frames based on their columns, however each column must be identical for the use of this function or else it won't work. Having every airports data in one data frame would make the development and querying of data much easier as it saves the hassle of querying 8 or 9 separate data frames for the same thing where as one query will do when they are combined.

2.4.4 Testing data

To ensure that the data that was to be used for visualisation and other analysis in the project was well structured, it was important to test out the data using simple visualisation techniques that put out graphs and charts to allow the data to be visualised. For testing visualisations, the package 'highcharter' was used. Highcharts is a package in R and other software tools that enable the visualisation of data in different formats.



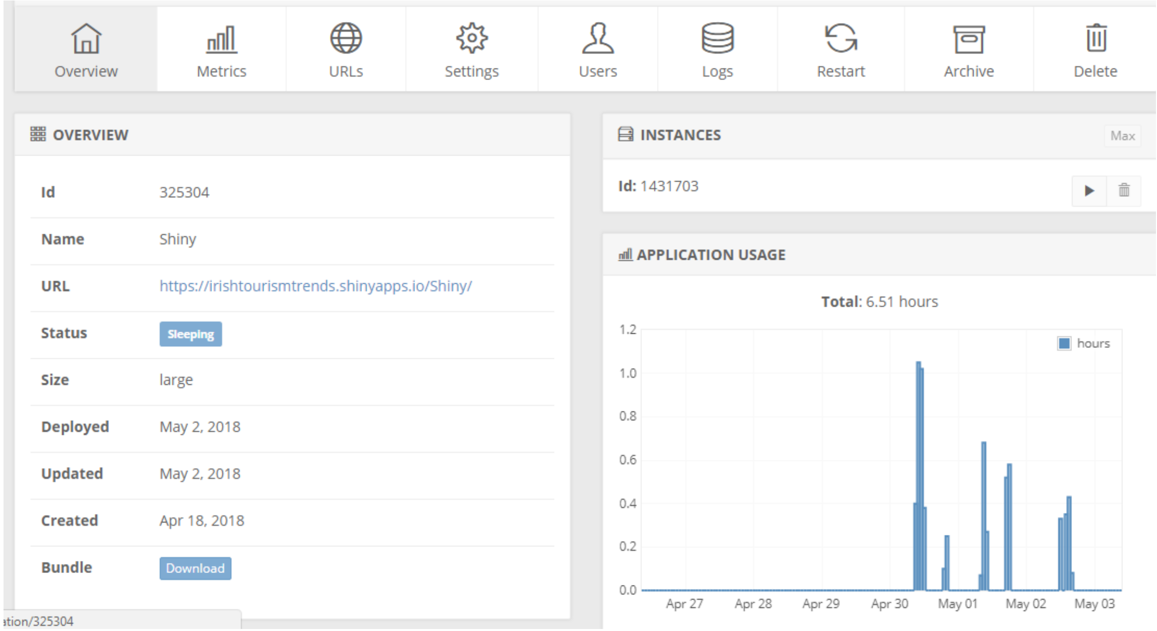
```
library(highcharter)

hc <- highchart() %>%
  hc_xAxis(categories = flightYear$year) %>%
  hc_add_series(name = "Inward Passengers Dublin", data = flightyear$Dublin) %>%
  hc_add_series(name = "Inward Passengers Cork", data = flightyear$Cork) %>%
  hc_add_series(name = "Inward Passengers Shannon", data = flightyear$Shannon) %>%
  hc
```

Figure 6: Test Visualisation

The sample of code in Figure 6 shows that the library highcharter has been called on to visualise the data for Dublin, Cork and Shannon. The year at which these passengers came to these airports would be the x-axis, while the quantity of passengers each year would be calculated along the y-axis. There would be 3 lines which we can differentiate between Dublin, Cork, and Shannon.

2.4.5 Shiny Application

Shiny apps allow for data and visualisations to converted into an interactive platform of charts and graphs which users can themselves go and play with the data. Shiny apps will be used to develop the platform for the data to be displayed. Furthermore, the application will be deployed to the shiny apps server which enables shiny apps to be stored on the cloud and accessed online (RStudio, n.d.).



OVERVIEW

Id	325304
Name	Shiny
URL	https://irishtourismtrends.shinyapps.io/Shiny/
Status	Sleeping
Size	large
Deployed	May 2, 2018
Updated	May 2, 2018
Created	Apr 18, 2018
Bundle	Download

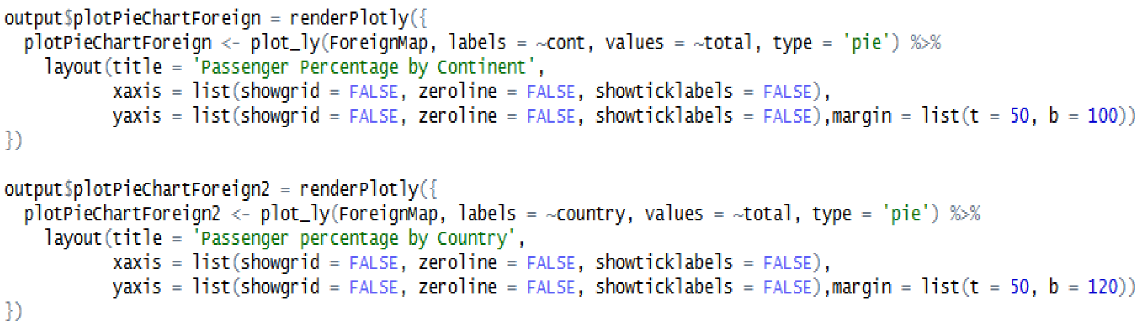
INSTANCES

Id: 1431703

Figure 7: Shiny Apps Admin Panel

It provides a very helpful tool such as running time and metrics. As you can see from the status in the overview section of Figure 7, the application is sleeping which means it's not in use by anyone for the moment. If somebody was to log on, it would change to running.

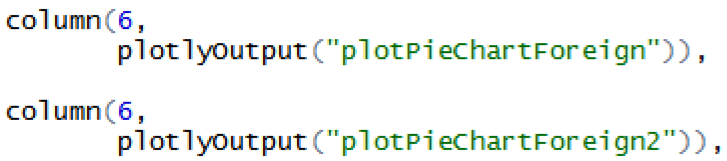
To develop a shiny application, there's a few things that need to be met. It is not the same as running a normal file, in fact, shiny is made up by 2 files, the ui.R file and the server.r file. As mentioned, the ui.r file provides the layout of the user interface for the platform, this is what the user will see. The file itself is an extension of the server.r which provides the back-end of the application. In the server file, the charts, graphs and maps are created and then pulled on by the ui file. The server file is far more complicated looking and hectic than the ui file.



```
output$plotPieChartForeign = renderPlotly({
  plotPieChartForeign <- plot_ly(ForeignMap, labels = ~cont, values = ~total, type = 'pie') %>%
    layout(title = 'Passenger Percentage by Continent',
           xaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE),
           yaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE),margin = list(t = 50, b = 100))
})

output$plotPieChartForeign2 = renderPlotly({
  plotPieChartForeign2 <- plot_ly(ForeignMap, labels = ~country, values = ~total, type = 'pie') %>%
    layout(title = 'Passenger percentage by Country',
           xaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE),
           yaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE),margin = list(t = 50, b = 120))
})
```

Figure 8: Shiny Server Code

From the code example of the server.r file in Figure 8, it can be seen that it is quite technical the way each visualisation must be laid out to allow the data to be called on by the ui.r file and visualised in the Shiny app.



```
column(6,
       plotlyOutput("plotPieChartForeign")),

column(6,
       plotlyOutput("plotPieChartForeign2")),
```

Figure 9: Shiny UI Code

And as from the code example of the ui.r file in Figure 9, it can be seen how easy it is to call on the server code. These two pieces of code actually work together. These two files working together enable the shiny app to be created and allow users to interact with the visualised data on the platform.

2.5 Graphical User Interface (GUI) Layout

The GUI of the shiny platform was developed with the intention of providing the user with the best experience possible and allowing them to absorb the knowledge in the best way possible way. The GUI of the platform follows the standard guidelines of HCI (Human-Computer Interaction) which state how a design can have the best possible impact on a user.

2.5.1 Human-Computer Interaction

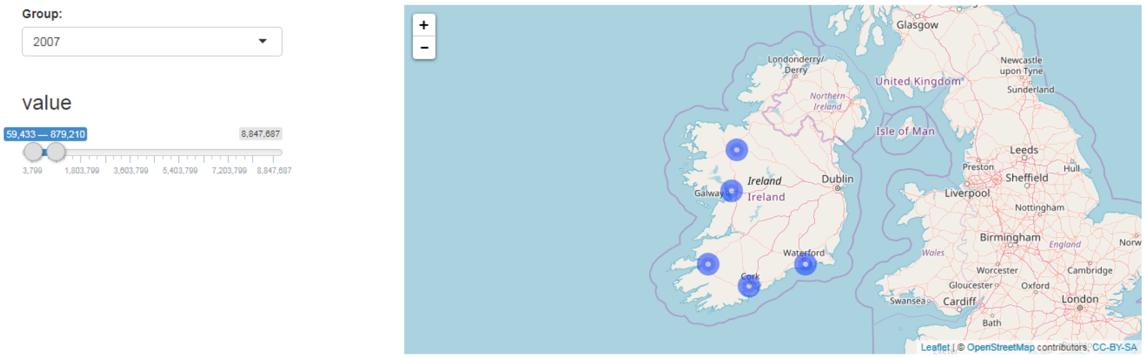
HCI (Human-Computer Interaction) relates to how a human will view and use the user interface of the Irish Tourism Trends platform. Laying out the foundations for Human-Computer Interaction can make an easy job of creating a better UX of an application. HCI guidelines were followed in the development of the Irish Tourism Trends Platform (The Interaction Design Foundation, n.d.)

2.5.2 Simplistic

When working with data and visualising data, especially big data, it's important to keep things simple and not overly complicated as it might be easy to understand what’s going on where for a developer, but not to a user that has no previous knowledge of the platforms layout or function. It was decided for the Irish Tourism Trends platform to keep things simple with just 3 pages. The home page described to the user what they would be seeing as they worked through the platform, the flights page contained data that was concerned with incoming flights to Ireland and where they’re going. The visitors page described everything there was to know about visitors coming to Ireland. The colour of the platform was kept simple with colours of white and grey (apart from charts colours) (Devaney, 2016).

2.5.3 Hierarchy of Visualisations

This is very similar to the logic of keeping the layout simple when bearing in mind that a user with little experience of the platform won’t know what to be looking for and where. With that said, the visual hierarchy ensures that the most important features are the most important things to be seen on a page. In terms of the platform for this project the visual hierarchy is applied by ensuring that the most important, impressive and interesting features are at the top of the page whereas the not so important features are after.



Group: 2007

value 59,433 — 879,210

Figure 10: Leaflet Map

On the flights page of the platform, it’s ensured that the interactive map which involves the slider and drop-down menu is at the top of the flights page. As this is the most interactive and visually impressive feature, this has been placed at the top above others.

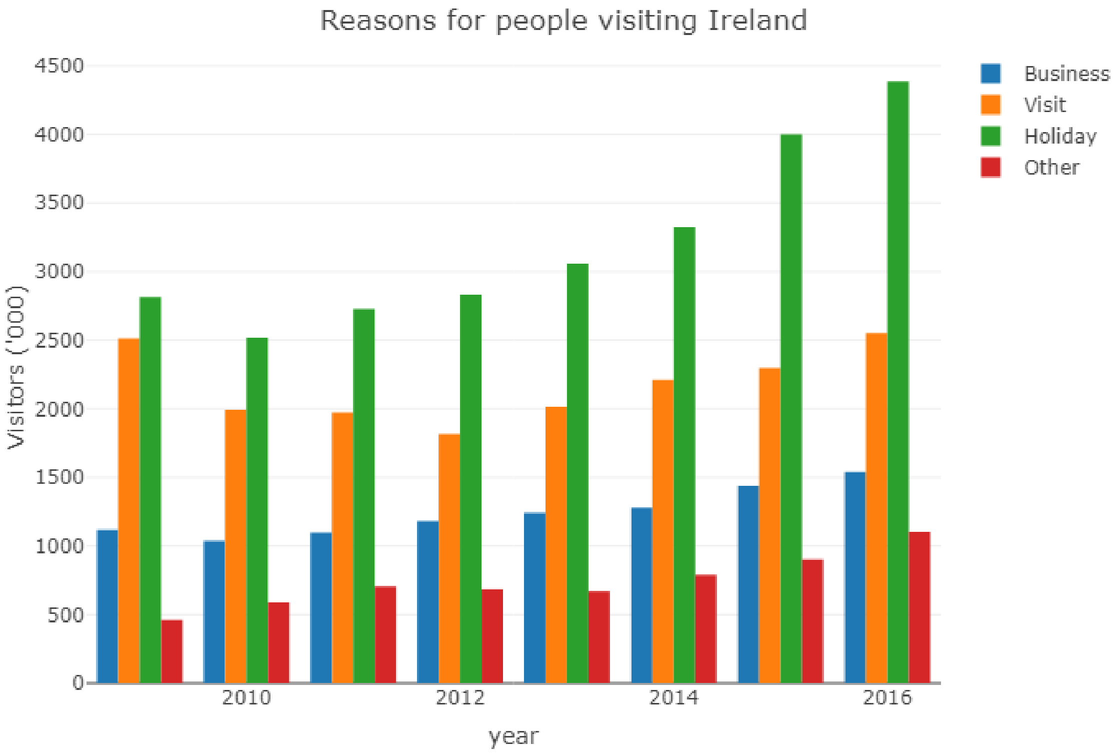


Figure 11: Tourists Travel Reasons

On the other hand, on the visitor's page, this graph has been placed at the top. It may not be as interactive as others however, it is the most informative, and so it was decided that this is essential to get the information across amongst other features (Devaney, 2016).

2.5.4 Navigable

The guidelines for the navigability of a webpage state that it's essential for the user to be able to navigate a platform with ease, and so the use of a navigation bar is vital. Preferably the navigation of a platform would be at the top left of the page. This is exactly where the navigation for the Irish Tourism Trends platform has been placed.

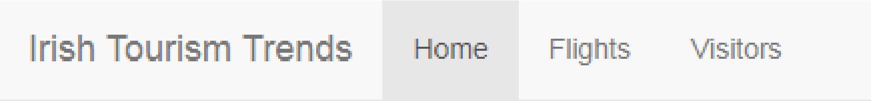


Figure 12: Shiny Navigation Bar

The navigation for the platform is simple. 'Home' is the landing page of the platform where users are able to view the exact aim of the platform, 'Flights' takes the user to the flights page and 'Visitors' takes users to the visitor's page. The navigation is kept very simple for ensuring a pleasant UX (Devaney, 2016).

2.5.5 Consistent

This guideline also relates to the simplicity of the platform stating that there shouldn't be a whole lot of difference in design between each page but rather a consistency where they are similar enough in design, so the user experience is also kept consistent. It also makes for a cleaner look in an application. As stated before, the consistency of this platform was simply kept to grey and white except for the coloured charts and maps (Devaney, 2016).

2.5.6 Accessible

With user testing still being conducted, it is difficult to decipher if the application is being used more on phone or desktop however, it is vital that the platform be responsive and readily available should a user want to access it on any device at any time. The platform was deployed using shiny apps server enabling the application to be available around the clock via browser.

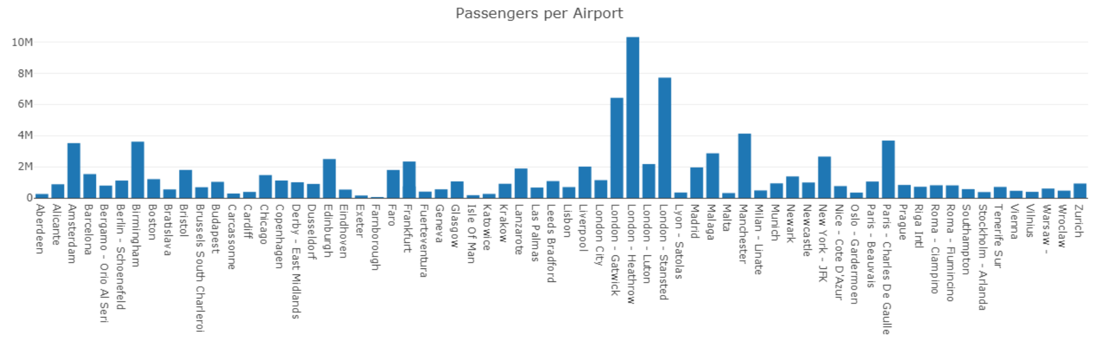


Figure 13: Passengers Per Airport Bar chart

From the chart about passengers per airport in Figure 13, it's seen that this chart is being viewed in a browser on desktop mode. This is the original state of the application where charts maps and graphs fir perfectly. To ensure accessibility, the platform must be made responsive.

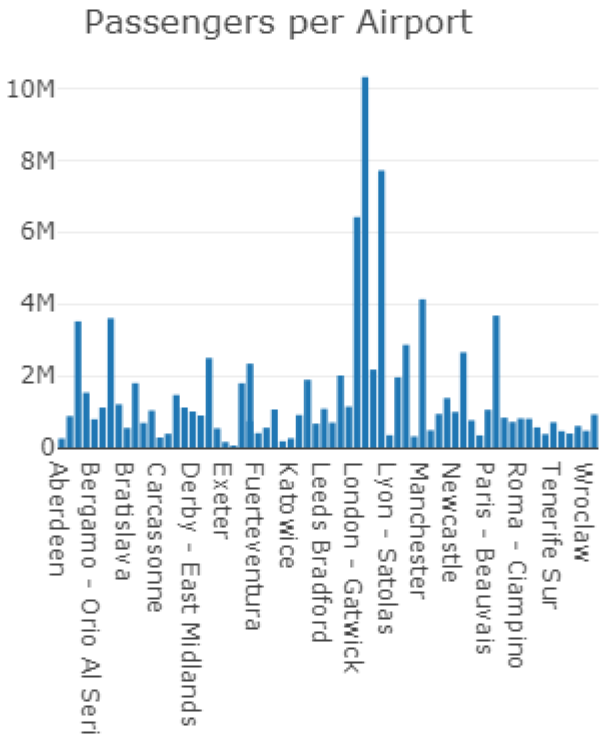


Figure 14: Passengers Per Airport Bar chart Mobile

The chart above in Figure 14, is the exact same as the previous chart about passengers per airport, however it is in mobile mode where the chart has become responsive and shrunk to fit the frame of the phone (Devaney, 2016).

2.6 Testing

Testing of the data and code is essential for having accurate charts, maps and graphs. Testing such as white box testing and black box testing enable the testing of the system from a technical perspective and a visual perspective.

White Box Testing: In the case of the Irish Tourism Trends platform, this type of testing is carried out by someone who has knowledge of the inner functions of the R code (Software Testing Fundamentals, n.d.).

Black Box Testing: In the case of the Irish Tourism Trends platform, this type of testing is carried out by someone who has no knowledge of the inner functions of the R code (Software Testing Fundamentals, n.d.).

The industry standard for documenting test cases has been followed. It consists of Test ID, Test Priority, Test Date, Test Title, Test Summary, Pre-Condition, Test steps, Test Data, Expected Result, Post-Condition, Actual result.

2.6.1 White Box Testing

Test ID: 1

Priority: High

Test Date: 05/05/2018

Test Title: passenger charts and map

Test Summary: Testing if data injected incorrectly into the data set will be output on the platform.

Pre-Condition: The data is output ranging from years 2007-2016, and longitude and latitude coordinates to use for mapping.

Test Steps: Inject data into the data frame that contains the year 2019, as it is in the future. Read the data into RStudio where the data will be run to create the shiny app. Run every line of code to ensure the inclusion of libraries and cleaning techniques.

Test Data: The data frame contains the attributes Airport, Year, Visitors, long, lat. The data injected will be Airport = Dublin, Year = 2019, Visitors = 1000000, long = 0.0000, lat = 0.0000.

Expected Result: The injected code will be removed from the data frame once the file is run. The file does not allow years from 2019 on or zero values for longitude or latitude.

Post-Condition: The data is output ranging from years 2007-2016, and longitude and latitude coordinates to use for mapping.

Actual Result: The platform is not updated by the injected data.

Test ID: 2

Priority: High

Test Date: 05/05/2018

Test Title: passenger count by month

Test Summary: Testing if data injected incorrectly into the data set will be output on the platform.

Pre-Condition: The data is ranging from months January to December outputting the data of passenger intake during those months

Test Steps: Inject data into the data frame of a value of 0 into for visitors count.

Test Data: The data frame contains the attributes Month & Visitors. The data to be injected should be Month = July, Visitors = 0.

Expected Result: Once the entire file is run, the month that contains zero visitors will be removed from the data frame as there is no value for passengers to Ireland.

Post-Condition: The data is output ranging from Jan-Dec and displaying the number of passengers coming to Ireland.

Actual Result: The platform is not updated by the injected data.

Test ID: 3

Priority: High

Test Date: 05/05/2018

Test Title: Foreign airport count

Test Summary: Testing if data injected incorrectly into the data set will be output on the platform.

Pre-Condition: The data frame contains various airports from around the globe that fly to Ireland. These airports have had at least 1 passengers fly to an Irish airport in the last 10 years. This data is visualised in the shiny platform

Test Steps: Inject some data into the foreign airports data frame. Read the foreign airports data into RStudio and run every line of code in the file.

Test Data: The data frame contains the attributes Airport, Numbers, long, lat and country. The injected data should contain Airport = Dublin, Numbers = 0, long = 2.5555, lat = 51.4700, country = Ireland

Expected Result: Once the entire file is run, this data injection will be removed as it contains the country Ireland and zero visitors. The data is only concerned with airports from foreign countries that carry at least one passenger to Ireland.

Post-Condition: The data is output to the shiny platform for visualisation concerning foreign airports.

Actual Result: The platform is not updated by the injected data.

Test ID: 4

Priority: High

Test Date: 05/05/2018

Test Title: Tourists years

Test Summary: Testing if data injected incorrectly into the data set will be output on the platform.

Pre-Condition: The reason and length data frames both visualise data from 2009-2016. This data is visualised in the Shiny platform.

Test Steps: Inject some data into the reason and length data frames. Read the data frames into RStudio and run every line of code in the file.

Test Data: The data frames both contain the attributes year, business, visit, holiday and other. The injected data should contain year= 2020, business = 34, visit = 454, holiday = 545, other = 345.

Expected Result: Once the entire file is run, this data injection will be removed as it contains the year 2020 which hasn't happened yet. The data is only concerned with data form the year 2009-2016.

Post-Condition: The data is output to the shiny platform for visualisation concerning the years that tourists visited Ireland.

Actual Result: The platform is not updated by the injected data.

Test ID: 5

Priority: High

Test Date: 05/05/2018

Test Title: Tourists reason for visiting

Test Summary: Testing if data injected incorrectly into the data set will be output on the platform.

Pre-Condition: The Exp and Cor data frames both visualise data that is associated with the reason why a tourist will come to Ireland. This data is visualised in the Shiny platform.

Test Steps: Inject some data into the Exp and Cor data frames. Read the data frames into RStudio and run every line of code in the file.

Test Data: The data frames both contain the attribute 'Type'. Data should be injected with type = Fun.

Expected Result: Once the entire file is run, this data injection will be removed as it contains a reason for travel that isn't 'business', 'visit, 'holiday' or 'other'.

Post-Condition: The data is output to the shiny platform for visualisation concerning travel reasons.

Actual Result: The platform is not updated by the injected data.

2.6.2 Black Box Testing

Test ID: 1

Priority: High

Test Date: 06/05/2018

Test Title: Visualisation of non-normal data

Test Summary: Testing to see if injection of non-normal distributed data would impact the dynamic of the visualisation of data.

Pre-Condition: The flightYear data frame visualises data of the number of passengers that comes to Ireland each year and into each airport. This data is visualised in the Shiny platform.

Test Steps: Inject some data into the flightYear data frame. Read the data frames into RStudio and run every line of code in the file.

Test Data: The data frame contains an attribute for each airport in Ireland and the year. Injected data, Dublin = 15000000, Year 2016.

Expected Result: Once the entire file is run, this data will be visualised in the shiny application, However the chart should be a log scale enabling the lower values and higher values to be perfectly visible on the chart.

Post-Condition: The data is output to the shiny platform for visualisation on a log scale dot plot.

Actual Result: The injected data is visualised; however, it does not change the user experience or the dynamic of the chart as it is a log scale.

Test ID: 2

Priority: High

Test Date: 06/05/2018

Test Title: User interaction with Leaflet

Test Summary: Testing to see if injection of data values will update the leaflet map and enable the user to change the visualisation based on year and passengers.

Pre-Condition: The flightsMap data frame allows the visualisation of all of Irelands airports on a map of Ireland and the number of passengers entering Ireland on a given year.

Test Steps: Inject some data into the flightsMap data frame. Read the data frames into RStudio and run every line of code in the file.

Test Data: The data frame contains an attribute for each airport in Ireland and the year. Injected data, Dublin = 15000000, Year = 2016.

Expected Result: Once the entire file is run, this data will be visualised in the shiny application on a map of Ireland. The injected data should update the slider bar which enables users to select the number of passengers

Post-Condition: The data is output to the shiny platform for visualisation on a map of Ireland with interaction from a dropdown menu and slider bar.

Actual Result: The injected data is visualised on the shiny platform and the slider bar has been updated by the injected data for the year 2016.

2.7 Customer testing

Customer testing was carried out with the use of surveys. The website, Survey Monkey was used to prepare questions and develop a survey so that users could evaluate the system of Irish Tourism Trends. Customers were sent a link to the platform and asked to go in and interact with the system before attempting the survey. The questions consisted of a signature of consent, the overall platform experience and the Irish tourism industry itself. Every question in the survey has been placed in the appendix of this document. There were 20 overall responses to survey and the results were analysed and reported below.

Signature of consent:

At the very beginning the users are asked to read a document of consent for their involvement in the testing. The document states that users can decide not to participate at any stage, that the questions and answers ensure the identity of a user is kept 100% confidential as it does not collect names, email addresses, or IP addresses, it explains what the applications aims are, and how the data is stored in a password protected electronic format. Users are asked to give electronic signature of consent and click 'Agree' if they have read the above information, they voluntarily agree to participate and are at least 18 years of age or decline participation by clicking 'Disagree'.

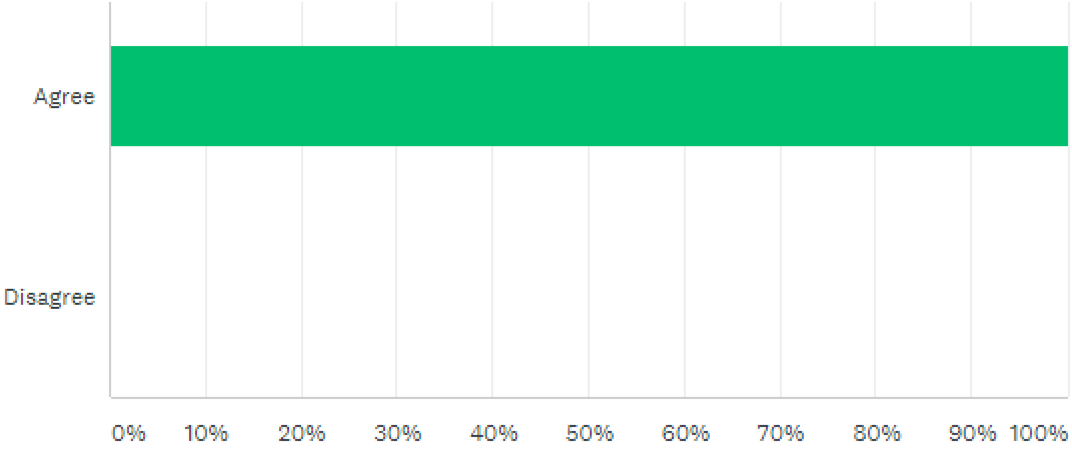


Figure 15: Participants Consent Form Agreement

All 20 of the participants clicked agree and gave their electronic signature of consent to participate in the survey study. This was a 100% answer rate for 'agree'. This meant that every participant could then carry on and answer the rest of the survey and allow the study of their response.

UI Rating:

After consenting to partaking in the survey study, users where asked if the user interface was pleasing to look at and navigate and to rate the user interface on a scale of 1-10, 1 being not very pleasing and 10 being very pleasing.

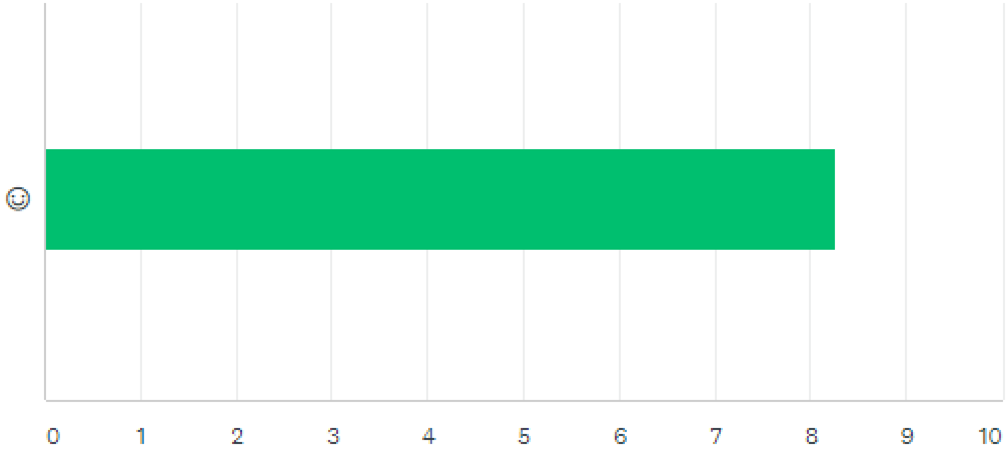


Figure 16: UI Rating

Overall, 19 users answered this question with 1 person skipping it. The ratings of the UI were, 2 people selected 6 for UI rating, which accounted for 10.53% of the overall percentage. 5 people selected 7 for UI rating which accounts for 26.32% of the overall percentage. 2 people selected 8 as the UI rating at 10.53% of the overall percentage. 6 people selected 9 as the UI rating at 31.58% for the overall percentage and 4 people selected a perfect 10 for 21.05% of the overall percentage. The average UI rating from a total of 19 submissions was 8.26 which is pretty good. The UI was deployed and can be accessed by any form of device.

Device used to access application:

Participants were then asked what device they were accessing the application on. The application is deployed to the web so it’s available on both mobile device and desktop.

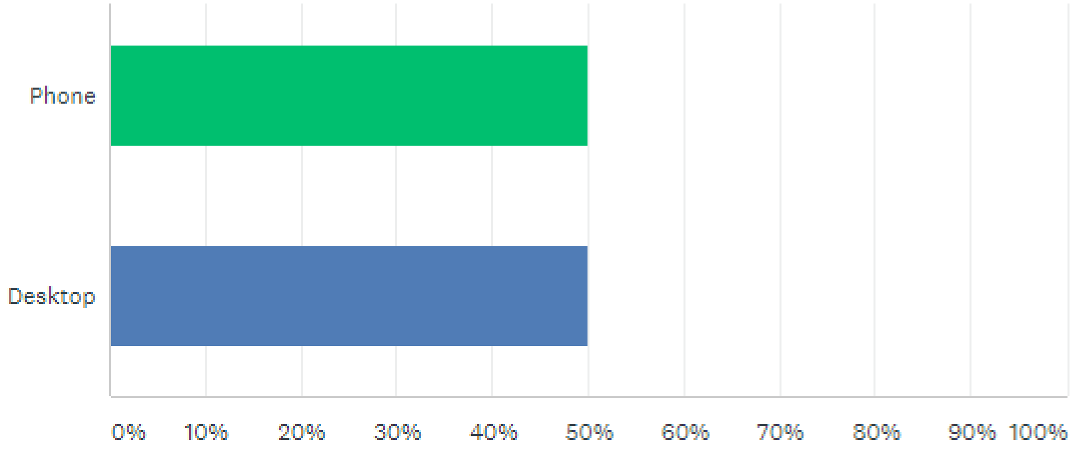


Figure 17: Device Used for Application

All 20 participants completed this question with 50% (10 people) accessing the application from desktop and 50% (10 people) accessing the application from their phone or mobile device

Most interesting feature:

Participants were then asked what part of the application was the most interesting. This question also had a comment field to allow for further specification of their answer. This was a multichoice question which consisted of the two tabs of the application which were flights and visitors.

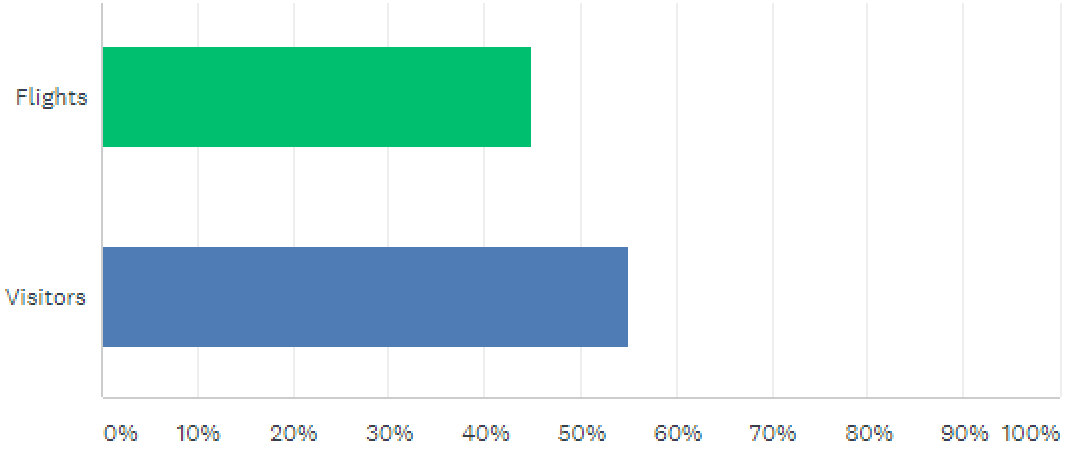


Figure 18: Most Interesting Feature

9 (45%) of the participants selected flights as the most interesting feature, whereas 11 (55%) of the participants selected visitors as the most interesting part of the

application. Comments were also enabled to allow further specification of the participants choice. There were 6 comments overall with all being anonymous:

- *“numbers coming from London/UK and possible effects that Brexit will have on those numbers.”*
- *“Passengers per Country”*
- *“none”*
- *“the increase in tourist umbers”*
- *“Regional airports feature”*

It was interesting to see how participants viewed the application and the possible knock on impact of certain political problems such as Brexit.

Recommendation:

Participants were then asked if they would recommend this application to their colleagues or friends. Although it isn’t a social platform or an entertainment platform, it does open eyes to what kind of nationalities are coming to Ireland, why they’re coming to Ireland and how much they’re spending. It also shows interesting facts and figures as to the number of passengers coming to Ireland.

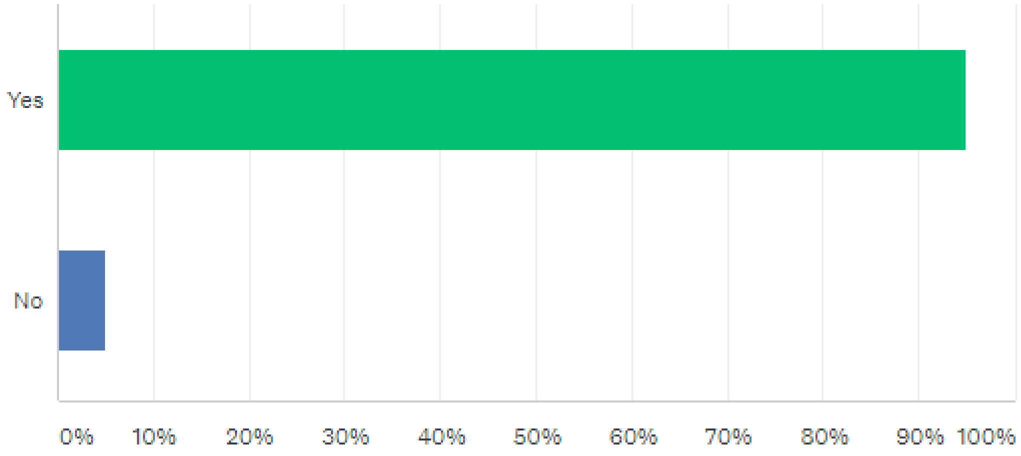


Figure 19: Recommendation

95% of the participants said they would recommend the application to their friends and colleagues and 5%, 1 person said they wouldn’t recommend the application to their friends and colleagues.

Impact of recession:

Participants were then asked if they were aware of the impact of the recession on passenger intake to Ireland. The impact of the number of tourists coming to Ireland is also displayed so participants are aware of both.

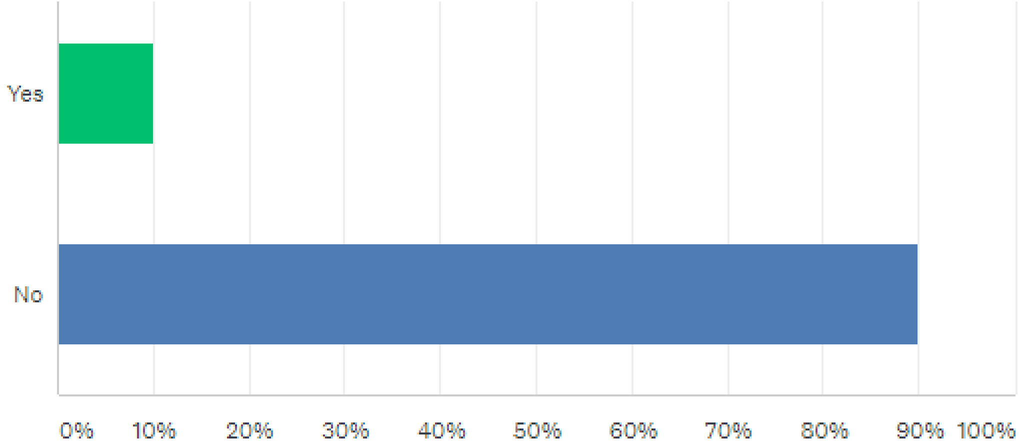


Figure 20: Impact of Recession Knowledge

90% of the participants were unaware of the impact of the recession on the intake of passengers into Irish airports where as 2 people, 10% of the participants were aware of the impact. This stat is one of the most interesting on the platform in my opinion.

Regional flights:

Participants were then asked if they have ever flown to anywhere in Ireland form an Irish airport. This is more of a tourism related question and to see if Irish people are flying to airports within Ireland.

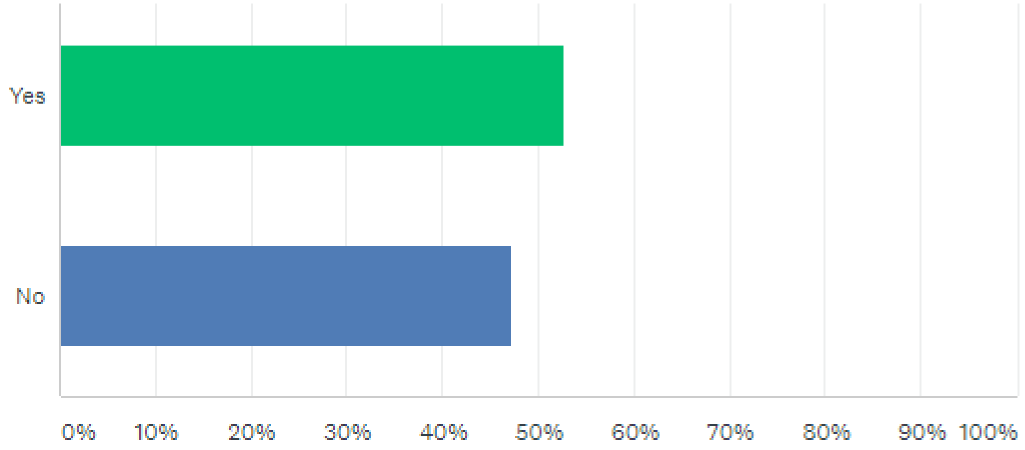


Figure 21: Regional Flights

Interestingly, 10 of the participants have flown to an Irish airport from within Ireland where as 9 people haven't. This number is well above what I thought would be the amount of people flying regional.

2.8 Evaluation

The system is evaluated by the accuracy of the data that the platform is displaying and how easy it is to interpret the data that is displayed. It was never intended to use a wide range of visualisations but visualisations that would be suit and enable the user to interpret the data shown effectively.

2.8.1 Insights

It is important to show users how many passengers that came into Ireland in the flights section for them to gauge the similarity of the number of tourists coming in with the amount of flights coming to Ireland. For example, the user might see that Great Britain brings the most tourists to Ireland but is that because British airports are the most popular when flying to Ireland or is there no difference between airports flying to Ireland it's just that British tourist are most frequent. It is important that the flights section can tie in and supplement the interpretation of the visitors tab.

Recession

As the historic data shows data from 2007 onward, it is vital that the impact of the recession be shown in one way or another. From flights data, the number of passengers coming to Ireland shows an interesting trend once the recession hit. This is correlated in the visitor's tab where it's seen why tourists are coming to Ireland, this also follows a similar trend once the recession hit.

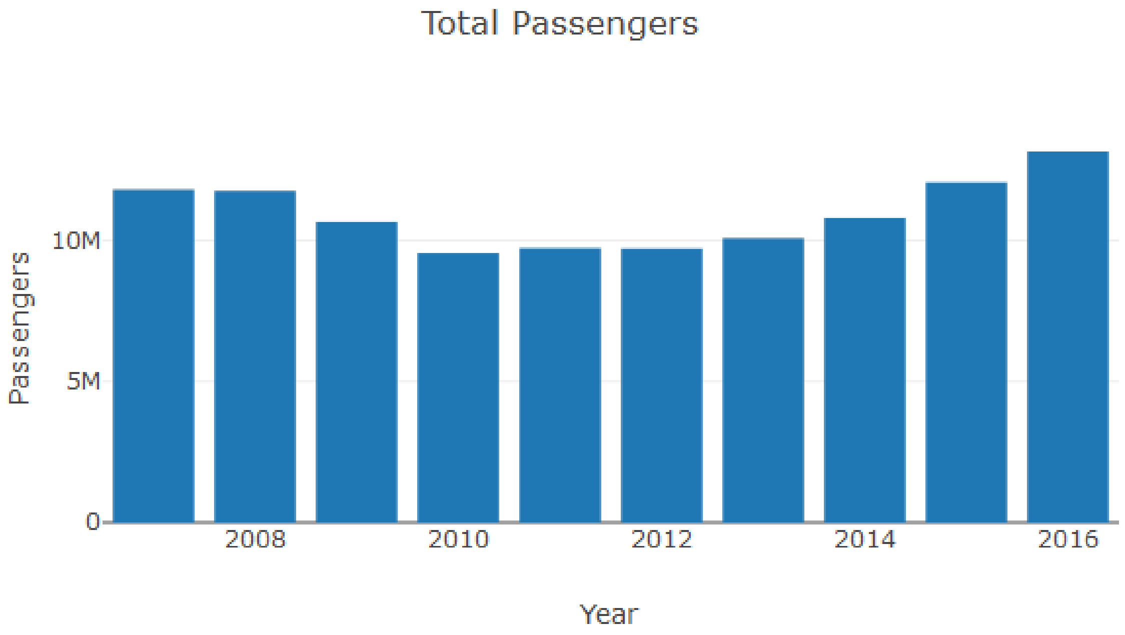


Figure 22: Total Passengers

From the bar chart above in Figure 22, the impact of the recession on the number of passengers coming to Ireland can be instantly seen. In one year, from 2008-2009, the passenger intake into Ireland drops by over 1 million passengers. This was the initial hit of the recession, and the impact can be seen clearly. The following year (2010) suffers another drop in passenger intake as from 2009 it drops by a staggering 1.1 million passengers. The total fall in passengers from the recession period came to 2.2 million passengers which incredibly only occurred over 2 years. Thankfully the passenger intake began to steadily rise to 2016 where it's the highest it's been in just under 10 years at 13.1 million.

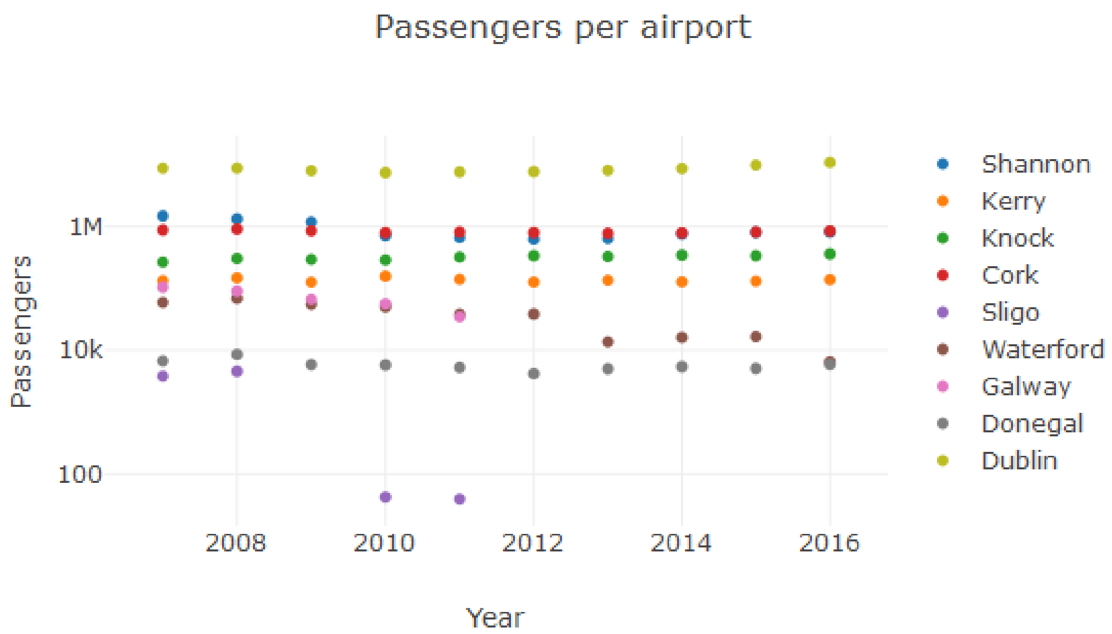


Figure 23: Total Passengers Per Airport

Where the previous bar chart shows the overall impact of the recession period on intake of passengers into Ireland. The analysation of the impact of the recession can be evidently seen on each individual Irish airport in Figure 23. Every Irish airport suffered from the year 2008-2010, however 2 Irish airports never recovered. Sligo passenger intake fell from 4570 passengers in 2008 to just 42 passengers in 2010. Service to this airport was then cancelled in 2011 when it took in just 39 passengers. Galway took a big fall from 90.359k passengers in 2008 to 56.52k passengers in 2010. Another fall to 34.547k in 2011 saw the cancellation of service to this airport in 2011. It's evident that the regional airports took the biggest hit over the recession period where Irish international airports quickly regained their status within a couple of years.

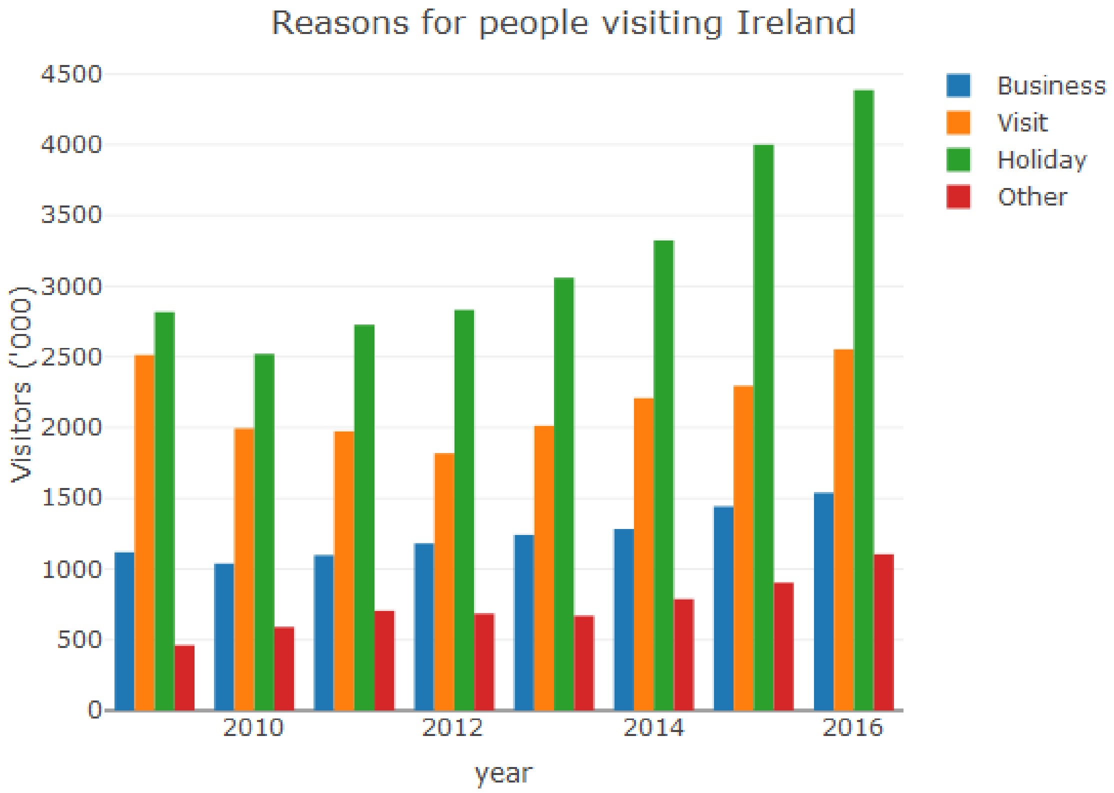


Figure 24: Visiting Reason

While it's important to see the impact of the recession on passengers coming to Ireland, it is equally important to see the impact on the reason why tourists are coming to Ireland. It's evident in Figure 24, that the main reason people are coming to Ireland is for a holiday, followed by visiting family and friends, then business and other. From 2009 to 2010, The number of tourists coming to Ireland took a steep drop. The number of non-residential tourists fell from 6.9 million in 2009 to 6.1 in 2010. People visiting family and friends took the biggest fall and struggled to begin steadily climbing until 2013, while business, holiday and other began to recover straight after 2010.

Busiest month for people to travel to Ireland

This can be an important stat for tourism in Ireland to target specific months of when people are most likely to be travelling to Ireland. For me, I was shocked at the busiest month for passenger intake in Ireland but when I realised a popular event is within that month, it would be thought as of as one of the busiest months.

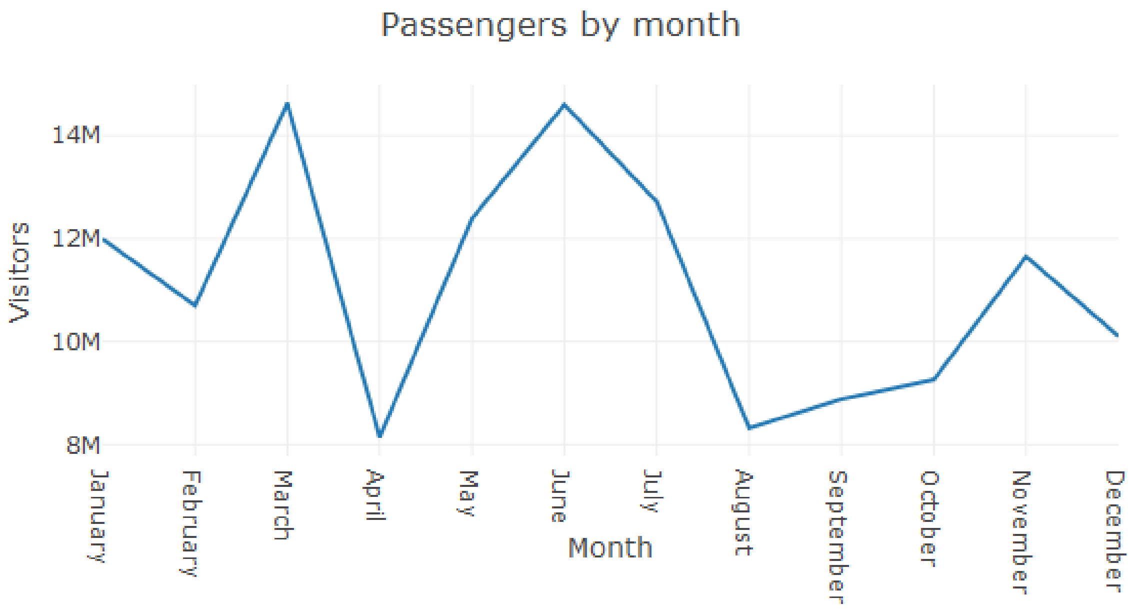


Figure 25: Busiest Month

The busiest month for passenger intake is March when St. Patrick's Day is held. What awe struck me was the sheer drop in passenger intake from March into April. From 2007-2016, 14.6 million people came to Ireland in March which then compared to April at 8.1 million, shows a difference of 6.5 million in just one month which is a staggering number. The summer months understandably are also just as busy which then falls quickly and builds back up in November.

Foreign Countries

The data collected allowed some great insights into the number of passengers per airport, number of passengers per country, number of tourists per country and how much they're spending, and what type of reason for travel that people will spend the most money on.

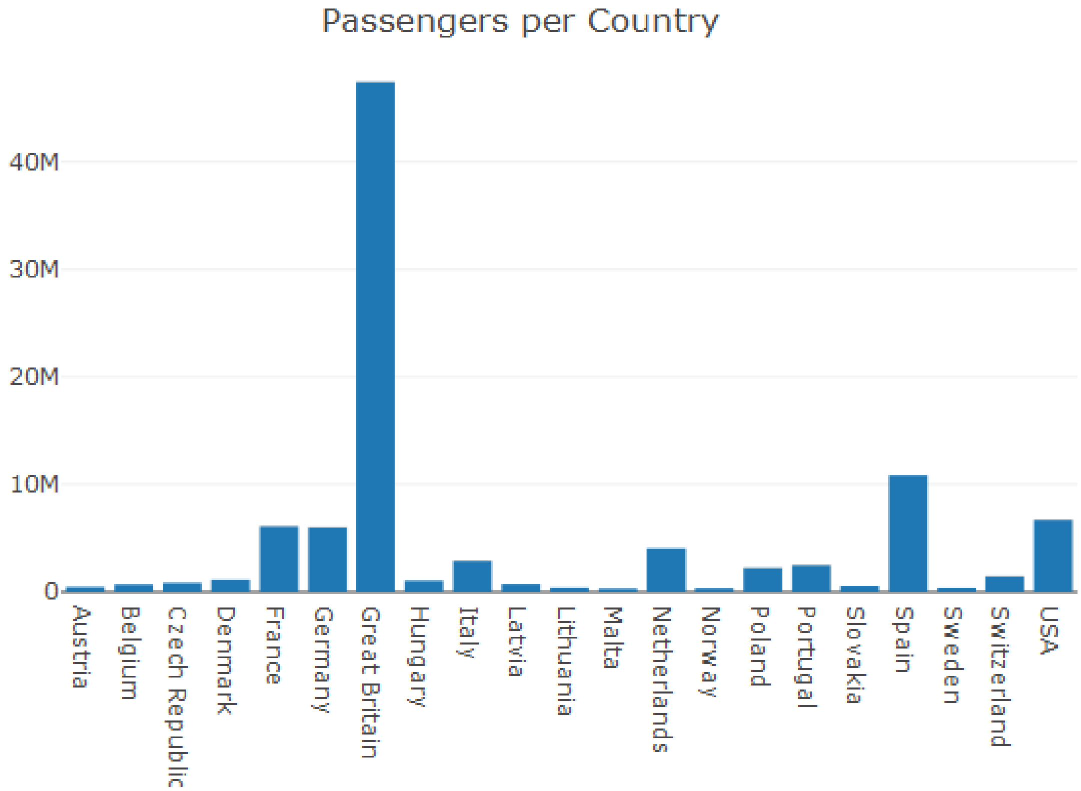


Figure 26: Passengers Per Country

From the bar chart on the number of passengers coming in from each country, it's easy to see how busy flights to Ireland from Great Britain are. Spain is Great Britain's closest competitor at 10.9 million and 36.6 million behind Great Britain, Great Britain's passenger intake is incredible and shows how important British airlines flying to Ireland is. However, it is worrying when thinking of the future of how Brexit might impact Ireland now that we see how important British flights to Ireland are. Since allowing comments in the user participants survey, one user identified how much of an impact Brexit could have on Ireland since seeing this chart.

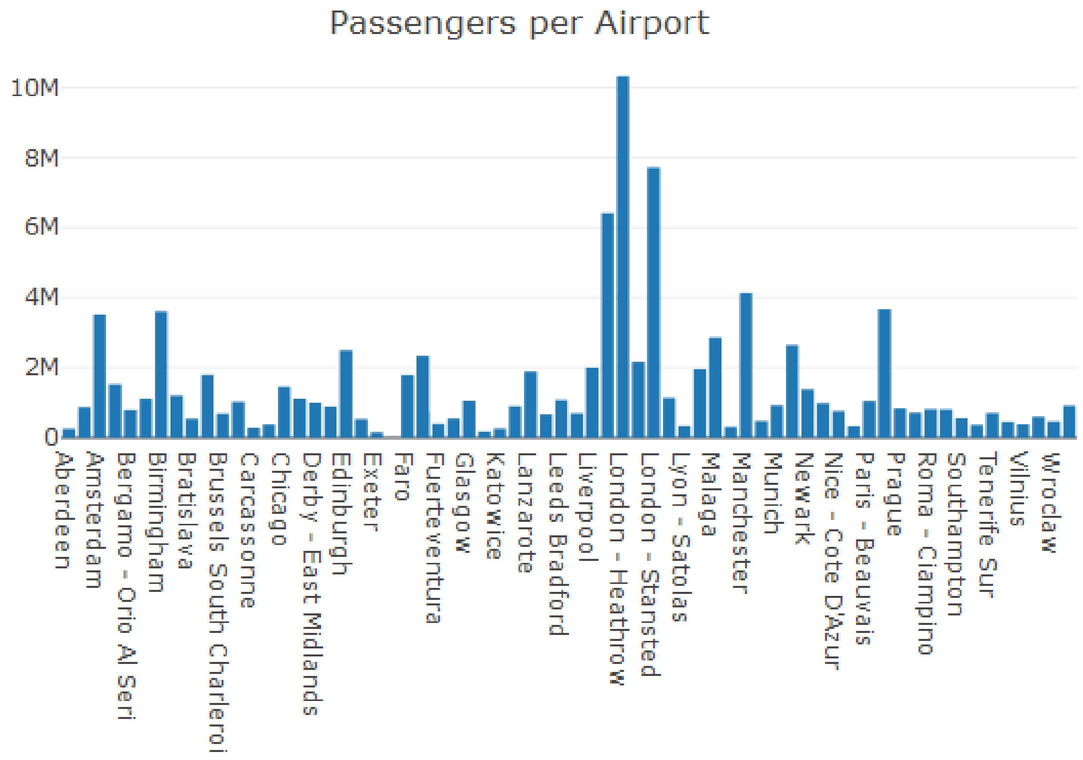


Figure 27: Passengers per Foreign Airport

Not surprisingly, British airports brought the most passengers to Ireland from 2007-2016. We can see in Figure 27, Manchester, Birmingham and Stansted rank among the highest in passenger numbers while London Heathrow brings in over 10.3 million over the last 10 years. It isn’t surprising that Heathrow airport carries the most passengers to Ireland as Heathrow to Dublin is one of the busiest routes.

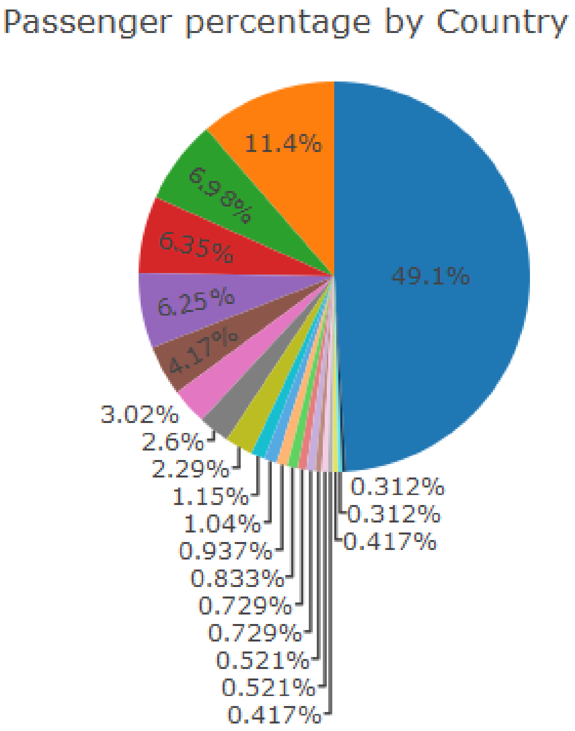


Figure 28: Passengers Country Pie Chart

It is astounding to see that Great Britain has accounted for almost 50% of the intake of passengers into Ireland which is scary and makes the thought of Ireland ending up on the wrong side of Brexit a catastrophe. Spain, France, Germany and other large EU countries also take a heavy percentage of the passenger intake to Ireland.

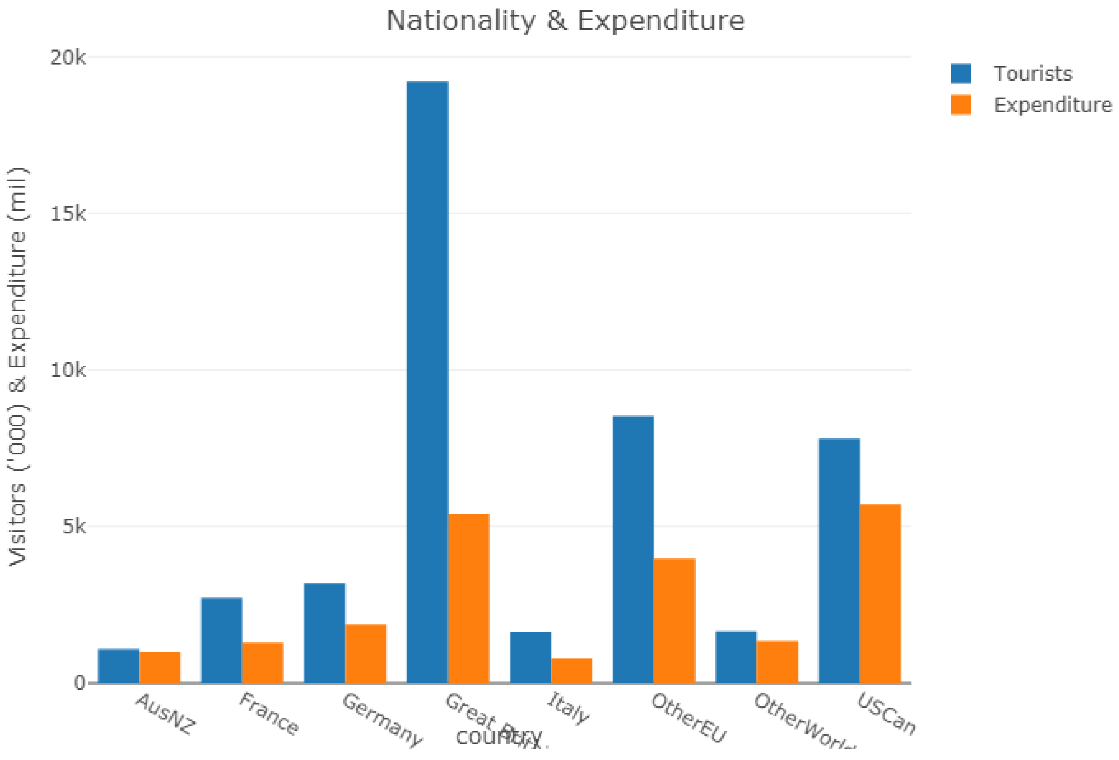


Figure 29: Nationality & Expenditure of Tourists

From the chart above in Figure 29, it's seen the nationality of tourists and how much they're spending when they come to Ireland. It's strange to see that there's many British tourists however they are second in terms of expenditure. It's the Canadians and Americans that spend the most money as British people have spent 5.4 billion euro and Americans and Canadians have spent 5.7 billion and have half the number of tourists compared to Great Britain when they come to Ireland. Unfortunately, the nationality of every tourist doesn't get released, rather it's classified into other EU, which is 14 countries combined and other World which is 14 countries combined.

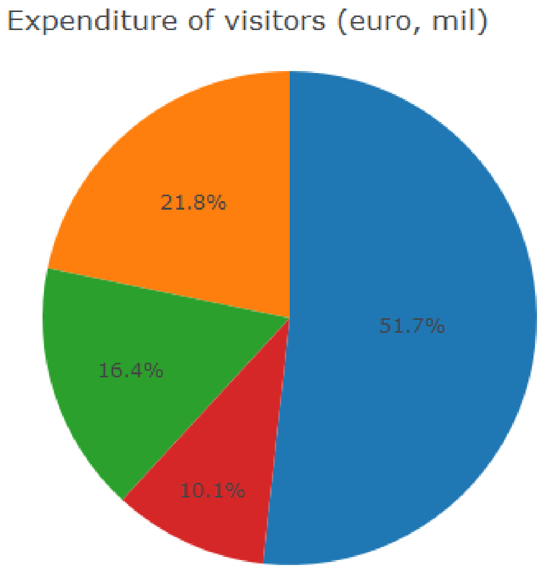


Figure 30: Expenditure Pie Chart

It’s also important to find out what reason people have for coming to Ireland and how much money they’re spending. 51.7% is the percentage of expenditure on a holiday from 2009-2016, this is not surprising as it would be worrying for people not to be spending the most amount of money on a holiday rather than anything else. Business and Other are at the other end of the scale at 16.4% for business and 10.1% for other, this was also expected as you would think that people visiting friends and family might spend more than somebody on a business trip.

2.8.2 Statistical Analysis

Irish Airports

Data was collected for the analysis of passenger intake into Irish airports from 2007-2016. From looking at the data initially it looks like there is a difference between the number of passengers each airport is taking in. However, this has not been statistically proven that there is a difference between Irish airports in terms of taking in passengers.

H0: Distribution of Irish airport data is normal

H1: Distribution of Irish airport data is NOT normal

Alpha = 0.05

A shapiro-wilk test was run on each attribute to see if the distribution of data for each Irish airport was normally distributed. At least one Irish airports data was not normally distributed. Shannon airport, **W** = 0.81827, **p-value** = 0.024, which means a non-parametric must be run to see if there is a difference between Irish airports.

H0: Cork = Kerry = Knock = Shannon = Sligo = Waterford = Galway = Dublin = Donegal.

H1: Cork ≠ Kerry ≠ Knock ≠ Shannon ≠ Sligo ≠ Waterford ≠ Galway ≠ Dublin ≠ Donegal.

A Kruskal-Wallis rank sum test was conducted to see if there was a difference between the airports in Ireland in terms of passenger intake from 2007-2016. **H** = 82.527, **p-value** = 0.3709. The null hypothesis is rejected that there is no difference between Irish airports. The alternate hypothesis that there is a significant statistical difference between Irish airports is accepted.

Tourist numbers and expenditure

Data was collected to enable the analysation of the reasons why people are coming to Ireland; how many are coming to Ireland and how much they're spending.

H0: Data for visitor numbers are normally distributed.

H1: Data for visitor numbers are NOT normally distributed.

A shapiro-wilk test was run to see if the data for the number of visitors coming to Ireland are normal. **W** = 0.94, **p-value** = 0.059. At an alpha value of 0.05, we can see the data are normally distributed for tourists coming to Ireland. We fail to reject the null hypothesis that data for number of visitors are normally distributed.

H0: Data for amount of money spent are normally distributed.

H1: Data for amount of money spent are NOT normally distributed.

A shapiro-wilk test was run to see if the data for the amount of money spent by tourists are normally distributed. **W** = 0.82, **p-value** = 9.715e-05. At an alpha value of 0.05, we reject the null hypothesis that data for the amount of money spent by visitors is distributed normally. We accept the distribution of data is not normal.

H1: Business = Holiday = Visit = Other

H0: Business ≠ Holiday ≠ Visit ≠ Other

As at least one of the attributes data is not normally distributed we must use a non-parametric test if there is a difference between the number of tourists and money spent by them. A Kruskal-Wallis test was conducted to test if there was a difference. **H** = 20.977, **p-value** = 0.4165. At an alpha value of 0.05, we fail to reject the null hypothesis that there is no difference between the number of visitors coming to Ireland and the amount of money they're spending.

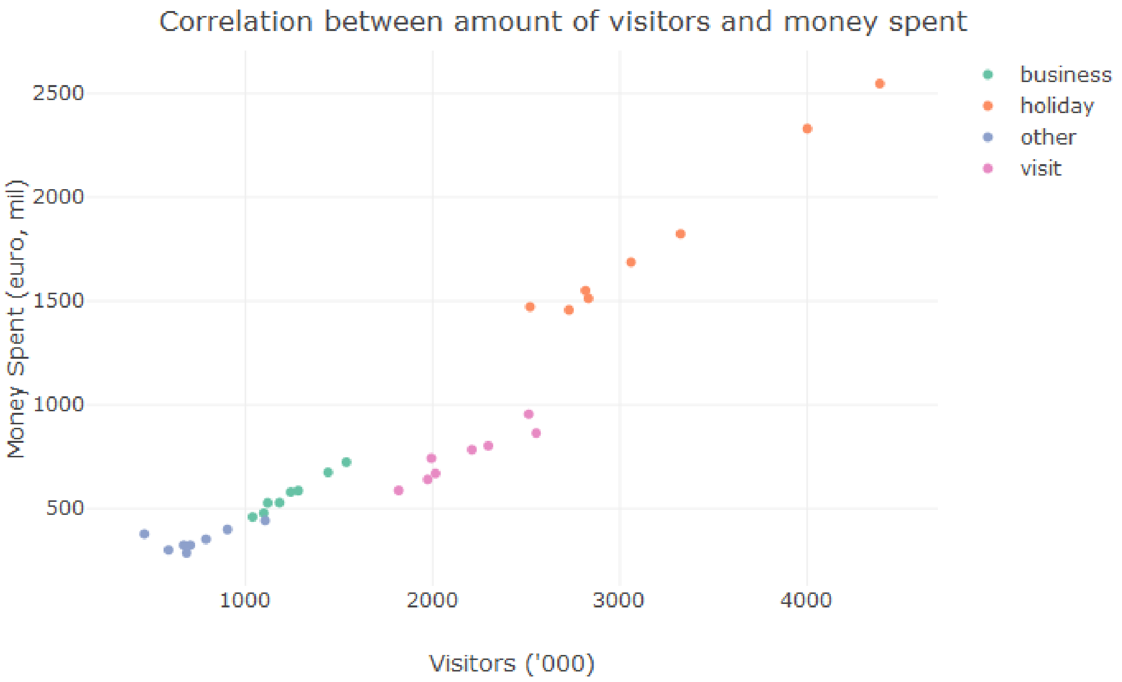


Figure 31: Correlation Plot

The correlation between the 3 different types of travel can be seen on the correlation graph in Figure 31. Although each different type of travel reason has different volumes of tourists, they still correlate to the amount of money those tourists are spending. There's a sharp rise in the people visiting Ireland on a holiday and the amount of money they're spending, whereas there is a steady rise in the other reasons for travelling and the amount of money being spent.

2.8.3 Predictive Analysis & Data Mining

When dealing with tourism, the numbers of passengers coming to Ireland, and the money being spent, it's important to identify trends and predict where the future lies for the industry. Especially with theories and predictions of a 10-year cycle for a recession it is particularly important to see where the expenditure and passenger intake will be in years to come. A time series prediction analysis was run in RStudio using Arima to predict the future trend of passenger intake into Ireland, however, the prediction was logically way off so Excel was used for a second opinion and gave a reasonable prediction compared to what was output by Arima

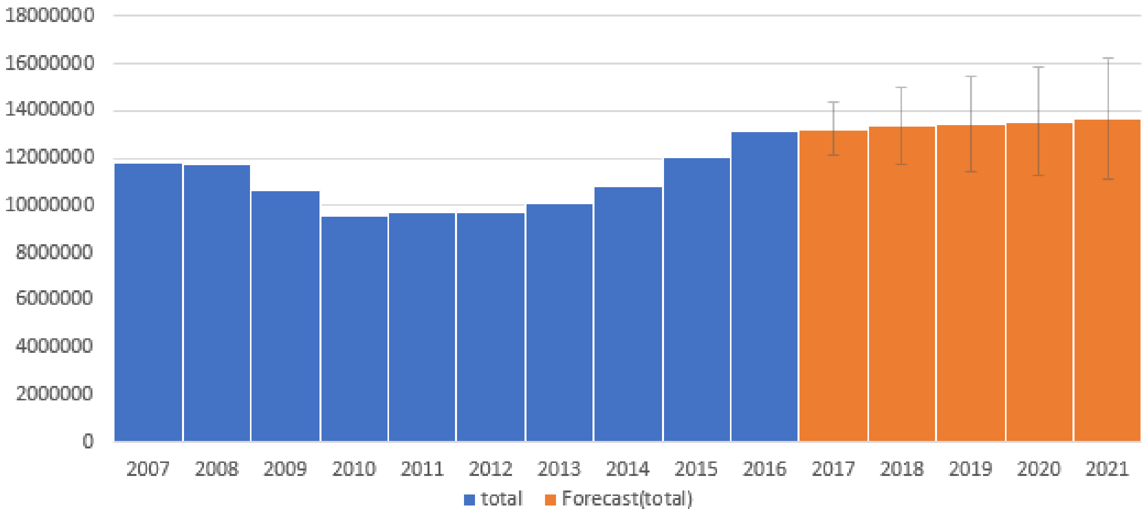


Figure 32: Passenger Time Series Prediction

A steady rise can be seen in Figure 32, for the passenger intake into Ireland with the highest predicted value being over 16 million by 2021, and the lowest being under 12 million by 2021, which could well happen if a 10-year recession cycle theory and prediction happens. The median prediction however is that the passenger intake will rise by 500k by the time 2021 comes around.

Data Mining methods were then run on to determine and classify Irish airports based on the number of passengers they have brought in from 2007-2016. A decision tree was created to classify the most popular airports in Ireland, these airports also fly internationally so they are of course the most popular when compared to regional airports. A decision tree visualises the classification decisions based on the passenger intake and divides each Irish Airport into its own leaf.

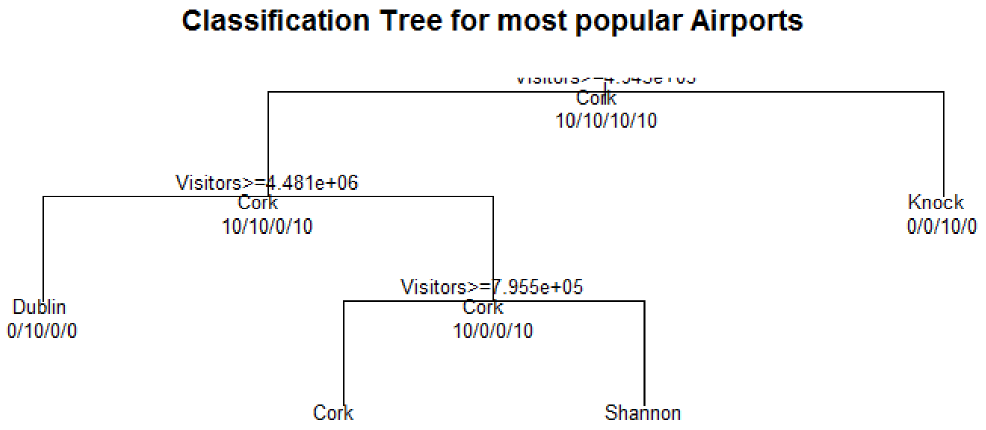


Figure 33: Decision Tree Airports

From the decision tree in Figure 33, we can see that the first decision is whether passenger intake is more than or less than Cork. Knock airport is the only airport that took in less passengers than Cork, whereas Dublin took in the most passengers out of all the Irish airports. The decision tree correctly classified each leaf of the tree showing that Dublin takes the most passengers in and Knock taking in the least number of passengers.

A conditional inference tree was then created to classify the popular airports in Ireland. From the decision tree, we could see that Dublin had the most passengers and Knock had the least passengers. The conditional inference tree, much like the decision tree, creates a visualisation of the decisions made to classify each airport with each leaf being a classification.

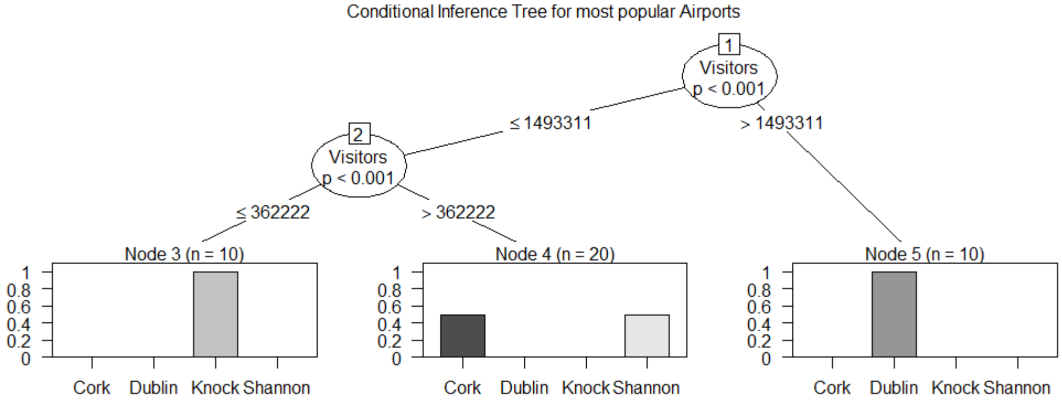
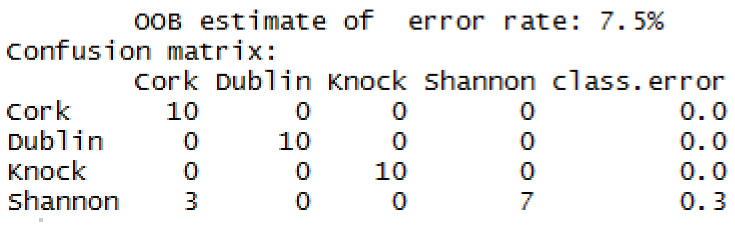


Figure 34: Conditional Inference Tree Airports

The conditional inference tree in my opinion has a much greater explanatory power than the decision tree as it's much easier to see the decisions being made. We can see in Figure 34, that Dublin is classified into node 5 as the number of passengers is greater than 1493311, whereas the rest of the popular Irish airports are less than 1493311. We can see that Knock is then classified by itself in node 3 as it is less than or equal to 362222, whereas both Cork and Shannon are classified together in node 4 which is greater than 362222 but less than or equal to 1493311. The conditional inference tree is much better at explaining its decisions when compared to a decision tree in my opinion.

A Random Forest model was then created to classify the most popular airports in Ireland. The Random Forest is effectively a combination of many decision trees in deciding which is used to conclude on the classification of attributes based on values.



```
        OOB estimate of  error rate: 7.5%
Confusion matrix:
        Cork Dublin Knock Shannon class.error
Cork      10      0     0       0         0.0
Dublin     0     10     0       0         0.0
Knock      0      0    10       0         0.0
Shannon    3      0     0       7         0.3
```

Figure 35: Random Forest Classification

From the Random Forest in Figure 35, we can see that it did not classify the Irish airports 100% correctly as 3 instances from Shannon airport were predicted for Cork airport. This is because Cork and Shannon airport values are very similar in terms of passenger intake. Dublin passenger intake is very high whereas Knock airport passenger intake is relatively low, so they were 100% correctly predicted.

Another important prediction that needs to be identified and acted on is the prediction of the expenditure of tourists when they're in Ireland. In the Irish Tourism Trends platform, it's shown that a different reason for travelling to Ireland will result in a different to type of expenditure for a tourist, however, there is a correlation between all reasons for travel and expenditure. It is however, very important to predict to the future of expenditure of tourists in Ireland.

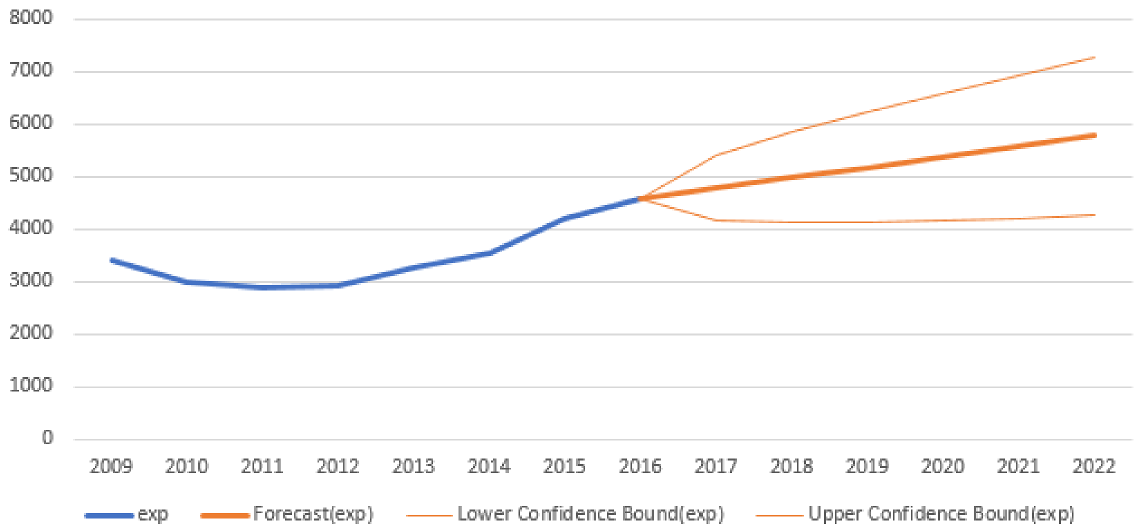


Figure 36: Time Series Tourist Expenditure Prediction

The tourist's expenditure was predicted with Arima in RStudio however the prediction did not seem like an accurate prediction, so a time series prediction analysis chart was created in excel to see the future of tourist's expenditure in Ireland. It's estimated that the median expenditure of tourist's in Ireland by 2022 will rise to just below 6 Billion euro which is astonishing when compared to the years 2010-2012 when the total expenditure was under 3 Billion. To double the expenditure of tourists in Ireland in 10 years would be an incredible stat.

It's incredibly important to use data mining techniques and predictive analysis when working with data that is so important to a country's economy. In my opinion the future of Irelands economy looks extremely strong with the passenger intake into Irish airports and the number of tourists coming to Ireland and how much they're spending in terms of money. The Irish economy recovered extremely well in terms of tourists and expenditure and looks to stay strong for the next few years. However, it's frightening to see the impact of the recession period on Irish airports, with Irish regional airports especially suffering from the that period as Sligo, and Galway have stopped taking passengers since 2011 as they took a serious drop form 2008-2011 in terms of passenger intake.

3 Conclusions

3.1 Challenges

In terms of the collection of data, the project was very challenging. The CSO was used for the data source which provided a table of all the historic data to do with the passenger's intake into Irish airports, and the number of tourists and their expenditure in Ireland. The data was laid out poorly especially in the passenger intake table where each month of the year was an attribute, for example '2007M01' was an attribute, so going up to 2016 accounted for an awful lot of attributes. The cleaning of each dataset was extremely time consuming also where attributes had to be combined into year, months and airports. Once cleaned however, sub setting and analysing the data was an easier job.

Another challenge came late in the project where it was identified that some of the projects data could not be used. Initially it was intended that TripAdvisor would provide data about accommodation and activities in Ireland, however, it was thankfully found that their terms of use document states that no data from TripAdvisor can be scraped and used without their permission. This made the final week of the project extremely stressful as TripAdvisor data accounted for half of the project. Every piece of data to do with TripAdvisor was removed from the project. This was a blessing in disguise however, as it forced the search for more data for the project. Data to do with why tourists come to Ireland, how much they spend and where they're coming from was found on the CSO. In my opinion, this data is more meaningful and more interesting towards other parts of the project and made the overall project better.

Using Shiny apps was also a massive challenge. Learning to use it took a few weeks and identifying what I wanted to answer on the platform took time but once the general layout of Shiny was known, then it was much easier to begin to create the platform. Deploying the application was a different story, as Shiny apps cloud requires the data to be in the same folder of the Shiny app as well as having the layout perfect too. When the project was run locally, it would run perfectly, however, once it was tried to be deployed there would always be an error which made the process extremely stressful. The showlogs() function was very useful as it would show what is breaking and stopping the application from being successfully deployed.

3.2 Advantages

The insights gained in the development of the platform are incredibly interesting. Seeing the impact of the recession on both airports intake of passengers to Ireland and the impact of the recession on the tourists themselves was evidently huge. It was personally interesting to find out and visualise different charts, graphs and maps that would answer questions regarding tourism, but it was just as cool to see how people were interpreting the visualisations during user testing as they have their own questions about the Irish tourist industry and it seems that the visualisations aided in the answering of multiple questions.

The development of this the platform also improved my personal data mining skills, analysing skills, visualisations skills and overall development skills during the undertaking of the project.

4 Further development or research

Given more time, quite a lot of meaningful insights and stats could’ve been found for example, which foreign airports are flying to the different airports around Ireland. It would be extremely helpful for local governments to find these stats out and maybe help them aim for a specific target market. It’s also important to keep scraping the trends of the number of passengers coming to Ireland as a recession in the next few years could deteriorate Irish airports intake of passengers, however, seeing as how well the popular airports or Ireland recovered well from the recession period, there is no doubt that it would recover even faster this time around as passenger intake is at the highest it’s been since 2007.

The Central Statistics Office announced that is interested in gathering data about visitors to Ireland with the main focus being on mobile phone data to see exactly what they do in Ireland and allow further insights into how tourists operate when they come to Ireland. Although this would be very interesting, it is very close to the line in terms of privacy as tourists would be watched with every step they take in Ireland. If laws in Europe that concern data were to change, then this would be an incredible step in the research of Irish tourism trends (O'Reilly, n.d.).

Further development of the Shiny application would be that the data is live and updated regularly. This is a tough task now as it relies solely on The CSO releasing this data and it is only annual data for some parts, however it would be an interesting and cool feature which would really help local governments and councils in Ireland to improve by giving them an advantage with this data.

5 Bibliography

[1] Smart Vision - Europe. (n.d.). What is the CRISP-DM methodology? [online] Available at: https://www.sv-europe.com/crisp-dm-methodology/ [Accessed 10 May 2018].

[2] The Interaction Design Foundation. (n.d.). What is Human-Computer Interaction (HCI)?. [online] Available at: https://www.interaction-design.org/literature/topics/human-computer-interaction [Accessed 10 May 2018].

[3] Devaney, E. (2016). 8 Guidelines for Exceptional Web Design, Usability, and User Experience. [online] Blog.hubspot.com. Available at: https://blog.hubspot.com/blog/tabid/6307/bid/30557/6-guidelines-for-exceptional-website-design-and-usability.aspx [Accessed 10 May 2018].

[4] Software Testing Fundamentals. (n.d.). Differences Between Black Box Testing and White Box Testing - Software Testing Fundamentals. [online] Available at: http://softwaretestingfundamentals.com/differences-between-black-box-testing-and-white-box-testing/ [Accessed 10 May 2018].

[5] O'Reilly, Q. (n.d.). The CSO wants to use mobile phone data to figure out tourism patterns. [online] TheJournal.ie. Available at: http://www.thejournal.ie/cso-mobile-data-tourists-2940965-Aug2016/ [Accessed 10 May 2018].

[6] R-project.org. (n.d.). R: What is R?. [online] Available at: https://www.r-project.org/about.html [Accessed 13 May 2018].

[7] RStudio. (n.d.). Shiny. [online] Available at: https://www.rstudio.com/products/shiny-2/ [Accessed 13 May 2018].

6 Appendix

6.1 Project Proposal

6.1.1 Objectives

Original:

For my 4th year data analytics project I will be analyzing the Irish tourism industry trying to answer questions such as, why is there mass tourism hot spots such as Dublin compared to other cities, how did the recession effect the Irish tourism industry and subsequently how has the Irish tourism industry progressed since then. This project and research could potentially be invaluable to cities and regions that are struggling to attract tourist as it will allow them to visualize the data on what attracts tourists to other such regions which they can then act on.

I aim to scrape data that contains all the information i will need to complete and answer the questions I'm asking. I'll will develop a platform that I can embed my visualized data onto, which will also allow users that are interested, to go on a view the data that they are particularly interested in.

Apart from trying to answer the questions mentioned in the first paragraph, another objective I would like this project to achieve is being able to assist local governments, shops, hotels etc. in viewing and processing factual data that can help local governments and councils make educated decisions when developing their area.

Revised:

Apart from finding the hotspots of Irish tourism, the other questions plus many more and objectives have been answered in the platform. The objective to help governments, shops etc. identify why people are coming to Ireland, and when they're spending the most money still stands.

6.1.2 Technical Approach

I intend to research, and scrape provided data from site the CSO. I will be looking to implement my recent learnings of R into my research in the coming months. I will use Shiny apps, which enables data analytics applications to be created and allow for user interaction with the data visualisations.

The project plan clearly lays out what things are most necessary for my project to be successful. The most important sequences are the ones that concerns the data, as that is what the project revolves around. The scraping, cleaning, storing and visualization of the data are vital for the project to meet the standard of a Data Analytics project.

6.1.3 Technical Details

To gather data, R in RStudio will be used to scrape data to visualize and analyse the data that is scraped. Shiny apps will be used to deploy the platform to the web for users to go on and interact with the platform themselves. Github and google drive will be used to store the project should the system crash and everything is lost.

6.2 Project Plan

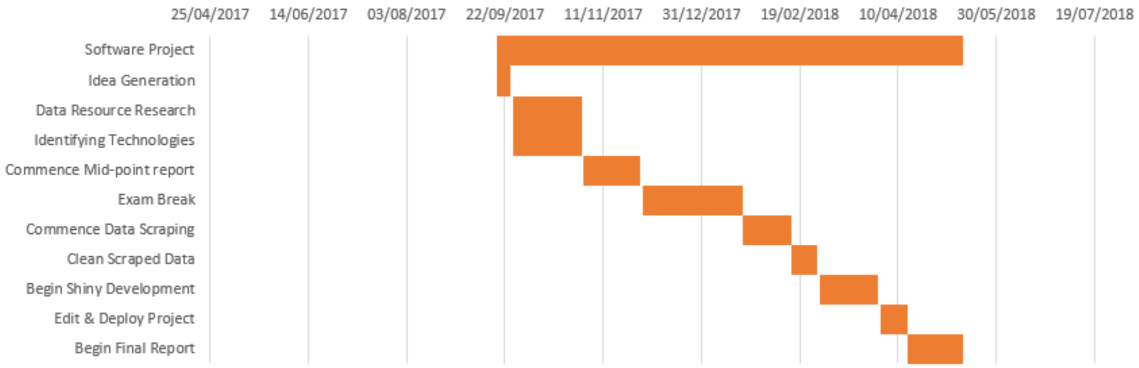


Figure 37: Project Plan

6.3 Monthly Journals

September

My Achievements:

This month was the start of the Software project campaign and was huge month for everybody as we had to identify, specify and pitch our project idea in a dragon's den type presentation.

In the data analytics stream there are two main things that are essential for a viable software project and these include, the ability to gather and interpret data, and the ability to answer a question with that data. For example, "why Ireland's weather is so bad?", here data would need to be gathered, manipulated and interpreted to answer this question.

My software project pitch was on Monday the 2nd of October and was thankfully accepted so I can now push forward and start identifying my requirements, and technologies that I'll need to use. My idea is to identify WHY is it that there are areas of mass tourism in Ireland (such as Cork, Dublin, Galway etc.). Thankfully there are plenty of datasets on recent Irish tourism that will be useful to me, in particular Failte Ireland have all sorts of Irish Tourism going back 10 years starting in 2006. Having such a huge amount will allow me to answer so interesting questions such as "How did the recession affect Irish Tourism?" and "How far has Irish tourism come in the last 10 years?". I will also hopefully be able to predict the future of Irish Tourism for years to come.

I'm delighted that my idea was accepted and that I could study the Irish tourism, to try and answer some important questions which I believe would provide some valuable information and data visualisation to local governments, to address some problems they may or may not have identified or to identify some trends.

October:

My Achievements:

This month I had to write up my project proposal which was very useful as we had to draw up a Gantt chart for the development of our projects. This gave me a good insight into actually how long I should be spending on one particular piece of development and where I should be development wise at a certain time.

This month I was also able to conduct further research on the where I'll be getting my data from and how I'll go about getting it. I have identified various sources such as failte Ireland, tourism Ireland and the CSO. Getting data from failte Ireland will be tricky as all of their data is stored in PDF's online. However, I will be extracting, cleaning and transforming the data I gather with the language R so that I can then visualize the data collected using charts made in Shiny which is a visualization tool that can be integrated with R, or a very good tool called 'Highcharts' which I found while researching the best ways to visualize data.

I am delighted that there is sufficient data available for me to work with, but now it's about how to go about efficiently scraping to be cleaned and transformed.

I also contacted my project supervisor this month, Simon Caton. We have arranged to meet after reading week. Hopefully he can point me in the right direction for what I should be doing with my project and hopefully I can get some tips on what and what not to do.

My Reflection:

I feel I am on track and everything is in place for me to really crack on with the most important part of the project which is gathering the data to be used. It will be tough to start that in the next few weeks as assignments and CA's will be piling up after reading week. Once I weather that storm, I will continue.

November:

My Achievements:

This month I carried out further research on possible data sources and possible questions for me to be able to answer with the data. It is important in my project that I show data that can be efficiently visualized and easily interpreted. This month

I also had to get going on my mid-point document and presentation. I had to explain in my document that I wanted the user to be able to sign into the platform and easily view the data and further technically describe my project and data sources.

I had to put together a small presentation and prototype to show at my mid-point presentation. I scraped data from failte Ireland and visualized it using the highcharts package in R. It showed the number of visitors coming to Ireland and where they are coming from. My presentation wasn't until the end of the semester in December, so I had everything ready.

My Reflection:

November was a good month in terms of getting my data to be visualized into a prototype for my midpoint presentation. I was able to get most of my document done for the mid-point upload to so that was good.

January:

My Achievements:

In December I had my mid-point presentation. I had compiled some data about the number of visitors into Ireland for my prototype presentation and visualized it using the highchart package in R. I pitched my presentation but after a Q&A with the examiners, realised that my data that I had chosen was not suitable as it was already aggregated into tables and already explored. I was advised to find other data sources, adjust my idea, or chose another idea.

I was also given some constructive feedback regarding my output of the data and how I intend to display my idea. My original idea was to visualize charts of different data and display them in a platform made from php where the user could sign in and view the data. Simon, my supervisor advised that I use a shiny app to display my data. This is an application where the data is fully interactive from the back-end. This was a much better idea than my original as my original charts would've been static images.

I was also advised the idea of displaying my data with the use of the map of Ireland which would enable me to show the different counties of Ireland. However, I was anxious I wouldn't find any other data source I could use that wasn't already aggregated and thus would end my project idea.

Over the Christmas break and after exams I had fully identified my data sources. I had TripAdvisor for my data on accommodation and activities in Ireland and The CSO for data on incoming flights to Ireland over the last 10 years. This data showed what month people were coming to Ireland and what airport they're going to.

Towards the end of January, I had begun to scrape all of the data I would need. It was very draining scraping data from TripAdvisor as no two URLs are the same and must be scrape individually using Rvest. Getting data from the CSO was much easier as all you have to do is pick the years and airports you want the numbers for. I would then scrape the table that was output into a data frame in Rstudio.

My Reflection:

December was a tough month, given the adjustments I was given to make to my project. Also, to have to go and find other data sources to fulfil my projects needs would be a tough task. I was delighted to be fully underway by the end of January and scraping the data that I needed.

February:

My Achievements:

By the end of February, I had scraped most of the data I would be using for my project and began to analyse parts and visualise parts to make sure the user would be able to view the data correctly and I was answer questions appropriately with the data.

I met with Simon to discuss the data that I had scraped, and he advised me to use a database in order to keep my finite original datasets in a database as the more I worked on the data, the more subsets etc. would be created so It would get confusing. I wrote a piece of code in RStudio to insert my data into a database to be held. An advantage of keeping my data in a database was that if I was to scrape more data, it would be much easier to insert that data into one of my data sets.

I met with Simon again to discuss the best possible way of outputting my data in my shiny app. I proposed two tabs which would hold the transport and accommodation/activities separately and we both agreed on the different questions I should be answering such as "is there a correlation between the number of activities and accommodation in the counties of Ireland?".

I concentrated my efforts for the rest of February on cleaning my data and preparing my data for the development of the shiny application.

My Reflection:

February was very good month in terms of kicking on with my project. I had scraped and cleaned all my data and I was ready to begin development of my application.

March:

My Achievements:

In March I began to create my Shiny application where I would output several interactive plots and charts. I began with the data of the accommodation and activities in Ireland. I used the plotly package to display most of my data as it has plenty of possibilities to output data.

I thought it would be a good idea to have two maps of Ireland side by side with one displaying a heatmap of the accommodation in the different counties in Ireland and the other displaying a heatmap a heatmap of the activities in the different counties in Ireland. It was interesting to see the differences between the counties.

I made an interactive correlation plot also, so users could test out the data and see what kind of activity was correlated with a type of accommodation in Ireland. I was shocked to see that there were relationships between some activities and accommodation.

My Reflection:

I was delighted to finally developing my application to display my data. This took a lot of pressure and stress off my back coming towards the end of my time in college and exam time.

April:

My Achievements:

April was an up and down month. I had finished my shiny application and had begun to try and deploy it and start my technical report, so everything was on track until I found out that some data that I would be using in my project wasn't allowed to be used. I had intended to use TripAdvisor's data on the number of accommodation and activities in Ireland however, their terms of use document states that data form their site cannot be used without their permission. This sent me into a panic as I had to remove this data which took quite a substantial part of my project up. After trying to get permission from their UK press office, I was denied, so I removed every bit of data I had. This, however, was actually a blessing in disguise as it forced me to find some new data to have on my platform. The CSO have a databank on the number of tourists coming to Ireland, why they're coming from and how much they're spending in Ireland. This, in my opinion was more interesting and allowed me to create a new tab on my platform for visitors.

My Reflection:

I have almost finished my project and I couldn't be prouder of myself as I've put in countless hours into this over the year. I look forward to the finished product and being able to showcase this to employers and my fellow students at the showcase at the end of May.

6.4 Survey Questions

Consent Form

ELECTRONIC CONSENT: Your participation in this research study is voluntary. You may choose not to participate. If you decide to participate in this research survey, you may withdraw at any time. The procedure involves filling an online survey that will take approximately 2 minutes. Your responses will be confidential, and we do not collect identifying information such as your name, email address or IP address. The survey questions will be about the Irish tourist industry and the application analysing Irish tourism. We will do our best to keep your information confidential. All data is stored in a password protected electronic format. To help protect your confidentiality, the surveys will not contain information that will personally identify you. The results of this study will be used for scholarly purposes only and may be shared with National College of Ireland representatives. Please select your choice below. Clicking on the "Agree" button below indicates that: • you have read the above information • you voluntarily agree to participate • you are at least 18 years of age. If you do not wish to participate in the research study, please decline participation by clicking on the "Disagree" button.

Question 1:

Is the user interface pleasing to look at & navigate? Rate the UI design on a scale of 1–10 based on design preference?

Question 2:

Has the application been accessed by Phone or Desktop?

Question 3:

What part of the application was most interesting?

Question 4:

Would you recommend this app to your colleagues and friends?

Question 5:

Did you previously know about impact on passenger intake from the recession?

Question 6:

Have you ever flown to anywhere in Ireland from an Irish airport?