

# Passenger utilization on railway network using timeseries forecasting

BSc (Honours) in Computing  
Data Analytics  
Year 2020/2021

Alejandro Diaz  
X17104050

## Contenido

Passenger utilization on railway network using timeseries forecasting .....	3
Introduction.....	3
Objectives.....	4
Aims .....	4
Technologies.....	5
Background and data details .....	5
Data Analysis .....	7
IDE preparation .....	7
Loading the dataset .....	7
Filtering the data per station.....	8
Declaring data as timeseries.....	8
Plotting the current data .....	8
Analysing the current data .....	10
Normality tests .....	10
Seasonality.....	15
Forecasting passenger utilization on the selected Stations .....	17
ARIMA.....	17
AUTOTS.....	23
Forecasting results compared and conclusions .....	27
Appendix.....	29
Project Proposal .....	29
Background.....	29
Objectives .....	29
Technical Approach .....	30
Technical Details.....	30
Project Plan.....	31
Technical details .....	31
Evaluation.....	31
Invention disclosure form.....	31
Monthly Journals .....	32
October 2020.....	32
November 2020.....	33
December 2020 .....	34
January 2021.....	35
February 2021 .....	35

April 2021 ..... 36  
May 2021 ..... 37  
June 2021 ..... 38  
July 2021 ..... 39

# Passenger utilization on railway network using timeseries forecasting

## Introduction

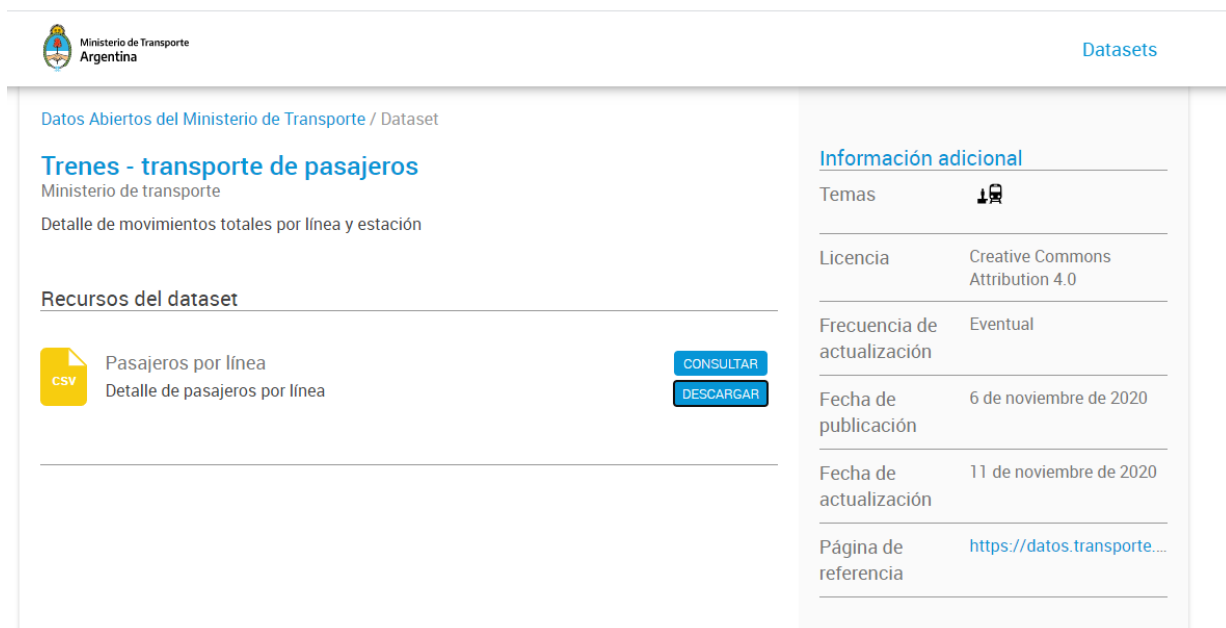
The project is going to be oriented about one of my passions, which is trains. I have been fascinated with trains since I was a kid. I could not understand how the human was capable to move and stop a 157 Ton huge metallic snake so easily.

Considering the tough times that we live I wanted to take a detour if the most common topic, which the is COVID-19. I consider that the virus not only has infected a good part of the human beings but all news, conversation and mostly our lives. That is my main reason to talk about trains, in this moment.

The purpose of my research was going to be oriented to providing forecasted data to the Spanish Railway Network Association (RENFE) to react to possible events on different dates and various train routes.

Once I conducted an initial research on the RENFE webpage I notice the existence of an API endpoints to retrieve data. Only after a several hours investigating that API I was able to find the endpoint to retrieve a list of contents under that database. Once I returned the list, I realized that only 1 day per route was offered as dataset, making the data too small to be able to forecast an actual day. To predict railway usage on a certain day we need a minimum of a month, and since we only had 1 day, I had to look for other options on the internet.

I was able to find a decent option thanks to the Argentinian government, that offers an up-to-date dataset of the railway usage on their network.



The screenshot shows the 'Datos Abiertos del Ministerio de Transporte / Dataset' page for 'Trenes - transporte de pasajeros'. The page includes the logo of the Ministerio de Transporte Argentina, the title 'Trenes - transporte de pasajeros', and a description 'Detalle de movimientos totales por línea y estación'. Under 'Recursos del dataset', there is a CSV file named 'Pasajeros por línea' with a 'DESCARGAR' button. On the right, the 'Información adicional' section lists: Temas (Trains icon), Licencia (Creative Commons Attribution 4.0), Frecuencia de actualización (Eventual), Fecha de publicación (6 de noviembre de 2020), Fecha de actualización (11 de noviembre de 2020), and Página de referencia (https://datos.transporte...).

Fig.1 – Ferrocarriles Argentinos Web site.

The Argentinian Ministry of Transport offers multiple datasets about railway usage, incidents, revenue data, and many other interesting datasets for other ways of public transport, like buses, undergrounds, planes and others.

The dataset that I found most relevant and most usable to my study was the one called “Trenes – transporte de pasajeros” which stands for “Trains – passenger usage”.

Despite that it looks close to the datasets offered by the Spanish counterpart, there were differences on how the data was presented. Whilst in the Spanish one we were given the data split in hours by station, on the Argentinian one the data is presented by station and showing the number of passengers per station over a 30-day period. Despite that at a first glance the data looks to be less usable, as per the much higher range of time (24h multiplied by 30 days, makes 720 hours per value), the number of stations listed, and the number of months flipped the coin against this data set. The service offers data since for the past 28 months, which in terms of a statistical point of view, makes it much more interesting than just a 24h timeframe.

## Objectives

The purpose of this research is to provide a series of Forecasts to the Argentinian Railway Network to mitigate any possible events around passenger utilization spikes over multiple routes across Argentina. The study will be focused on the stations having most present data. From all the relevant data found, only three stations met the minimum volumes to provide significant forecasts. On this study we will compare three different forecasting models to see which one provides the most accurate and relevant results.

The problem that the project is going to address is to allow the Argentinian Railway Network to tackle any impact on their services during peak times by forecasting usage on these dates and increase capacity on their trains effectively. That will avoid the company to lose ticket sales as for a possible out of seats scenario, and to choose the best way to increase the number of seats depending on the results of the forecast; would adding wagons to the train will be more effective than sending another train?

## Aims

Below there is a list of aims for this project:

Aim 1: to obtain datasets that are suitable for the research.

Aim 2: to be able to obtain datasets in a programmatical way.

Aim 3: to use the R programming language and Microsoft Excel to model, obtain and present the results in the clearest way.

Aim 4: using libraries and programming scripts, to be able to clean the datasets and make them usable to the research.

Aim 5: to pre-process the data eliminating, non-useful data and to clear redundancies or empty data points.

Aim 6: to use known and proven methodologies for Data analytics in order to obtain the desired results and to be able to provide an output easy to interpret to multiple audiences.

Aim 7: to forecast time slots to predict passenger usability, using different techniques.

Aim 9: to use different forecasting techniques and to compare them.

Aim 10: once having the results obtained from comparing the forecasting techniques, to choose the best model.

Aim 11: to make visualisations of the forecasted data to reach multiple audiences.

Aim 12: to provide results, choosing the best model and technique used and presenting the future datapoints calculated.

## Technologies

R – R is a programming language used mainly in data analytics, statistics, finance and Maths. There are multiple IDE's in the market to be able to work and use R, but the most known and used is R Studio. I get familiar with RStudio thanks to our lecturers on Data Application Development lecturer and Advanced Business Data Analysis, which gave us a lot of resources and techniques to get as much as possible of R and Studio. Within R I will going to clean and transform the data to make it readable and usable, then I will generate future data using Arima, Prophet and Holt Winters models with them corresponding software binaries.

Python – Python is considered one of the most flexible programming languages in the current scene. My intention was to use it on this research but found R & R studio a more adequate package to work on a data analysis project. The most relevant factor that made me change my mind about Python was the fact that Prophet can be used in R seamlessly, and that data analysis package was the most attractive factor of Python to me; since I can use it in R, I decided to not use Python this time.

Excel – Who does not know Excel these days? I am going to make use of it to store the datasets I have found on the internet as Excel supports xlsx and comma separated values.

## Background and data details

I considered a background section needed on this project despite of being unorthodox, since I swapped completely the country, and end customer that this document was directed to.

I was born in Spain and only moved to Ireland 5 years ago, so at the time I was required to work on a project directed o any kind of customer or any country, my fist desire was to build something directed to improve or analyse something from my motherland, and something according to my interests; Spain and Trains were the two factors selected and they can only mean RENFE (Red Nacional de Ferrocarriles Españoles).

What I never thought of back in 2020 was the huge constraint I was going to overcome when conducting an analysis; I had a terrible and irreversible lack of data to work with.

I try to reach that Spanish organisation from any possible angle; I called them, I emailed them, I even visited a regional office and try to speak with the management, but they would not attend individuals. All I needed was for them to update their database or give me access to more chunks of data for academic purposes, but nothing ever worked, and there was me, hanging there with ONE day of data in the past.

Since I could not obtain anything from them after several months, I had to pivot to another Country, and thankfully the Republic of Argentina was there for the rescue (Dataset below in Excel format).

	A	B	C	D
1	mes	linea	estacion	cantidad
2	01/12/2019	belgranonorte	Florida	37226
3	01/12/2019	belgranosur	KM. 12	2763
4	01/12/2018	belgranosur	KM. 12	1355
5	01/12/2018	mitre	El Talar	70
6	01/12/2019	mitre	El Talar	15
7	01/12/2018	mitre	Florida	19344
8	01/12/2019	mitre	Florida	17548
9	01/12/2018	mitre	Maq. Savio	2008
10	01/12/2019	mitre	Maq. Savio	954
11	01/12/2018	mitre	Matheu	907
12	01/12/2019	mitre	Matheu	922
13	01/12/2019	mitre	Zelaya	286
14	01/12/2018	mitre	Zelaya	308
15	01/12/2018	roca	Haedo	2214
16	01/12/2019	roca	Haedo	31
17	01/12/2018	roca	Tablada	277
18	01/12/2019	sarmiento	Haedo	148186
19	01/12/2018	sarmiento	Haedo	146098
20	01/12/2019	sarmiento	Miserere	307
21	01/12/2018	sarmiento	Miserere	336
22	01/11/2019	belgranonorte	Florida	38316

Fig.2 – Passenger utilization Dataset Ferrocarriles Argentinos.

The data set above consists of the following:

Observations: 5304

Attributes: 4

“cantidad” (Quantity) – Integer, shows the number of passengers using the service on a given month.

“mes” (Month) – Date in YYYY/MM/DD format, month corresponding to the datapoint.

“linea” (Line) – Character, categorical, corresponds to the name of the line passing that station.

“estación”(Station) – Character, categorical, refers to the station on the datapoint.

Using Excel function UNIQUE, we can observe 237 different stations, however only a very few stations have been updated on the site's database. It's a fact that the Argentinian railway network is a heavily outdated country resource, and the country is only digitalizing that service data since a few years back, therefore, I will only focus on the three stations having most years of usage Data, all of them having more than 43 data points. Each observation contains passenger usage data from the first day of each month to the first day of the next month.

- Haedo – Station located in Haedo city, on the district of Buenos Aires, we can observe 56 datapoints of monthly usage data.
- Retiro – Station located in the city of Buenos Aires, we can observe 45 datapoints of monthly usage data.
- Florida – Station located on the town of Florida, district of Buenos Aires, we can observe 43 datapoints of monthly usage data.

## Data Analysis

On this section I am going to cover all the steps involved in obtaining the data, cleaning the data and making it usable to run statistical tests on it.

### IDE preparation

Most people think that IDE's are full ready to compute the data and present it in the most convenient way, but in reality that is not the case. Like most of the IDE's in the market there is some "tweaking" required to obtain the desired results.

In my case, I have used R-Studio mainly and I noticed that using plotting methods such as Autoplot or anyone from the GGLOT2 package, once presenting the ranges on the X and Y axis it was shows in Scientific notation, something highly unreadable. To prevent that I used the function and arguments below:

```
69
70 ##the code below is to avoid scientific notation appearing for big values
71 options(scipen=999)
72
```

Fig.3 – Overriding scientific notation

### Loading the dataset

R is a very popular language and therefore there are infinite different packages to give functionality to the IDE. Since I was able to choose the dataset format once I downloaded it from the MTA I preferred to have it in Excel. After doing some research I found the most solid and stable package, called Tidyverse; within Tidyverse there is readxl library, which contains read\_excel function.

```
78 #reading the file using read_excel function from the readXL Library
79 install.packages("Tidyverse")
80 library(readxl)
81
82 trainArg <- read_excel("C:\\Users\\alex\\OneDrive\\Escritorio\\SUMMER\\Datasets finales\\arg1.xlsx")
```

Fig.4 – Loading the dataset in R



## Filtering the data per station

The examples below show how to pipe data from the dataset and filter it by a selected criteria using functionality from the dplyr package in R. “%>%” gets the data from the left side and pipes it on the right side where I applied filter() using “estacion” (Station) argument to select the observations from all the three selected stations in the dataset

```
90
91 RetTrain <- trainArg %>% filter(estacion == "Retiro" )
92 HaedoTrain <- trainArg %>% filter(estacion == "Haedo" | estacion == "HAEDO")
93 FloridaTrain <- trainArg %>% filter(estacion == "Florida" | estacion == "FLORIDA")
94
```

Fig.5 – Piping and filtering desired stations using dplyr functions and operators.

## Declaring data as timeseries

Having the data declared as timeseries enables us to use a series of analytical packages and functions that only work for this type of data. In the snippet below I am using the function “ts” to declare a timeseries object, selecting the filtered data in Fig.6 located in the fourth column of the Data set. We want to capture the first date present on the dataset and leaving the last datapoint (March 2020) to make a prediction over that month in a further stage of the analysis. Since the data points enclose months, the frequency has to be 12. Timeseries is a list of indexed datapoints in a time order.

```
95 #creating timeseries objects. We leave the last datapoint 2020-03 out of the analysis
96
97 RetiroTS<- ts(RetTrain[,4], start=c(2017,5), end=c(2020,2), frequency = 12)
98 HaedoTS<- ts(HaedoTrain[,4], start=c(2017,5), end=c(2020,2), frequency = 12)
99 FloridaTS<- ts(FloridaTrain[,4], start=c(2017,5), end=c(2020,2), frequency = 12)
```

Fig. 6 – Declaring relevant data as timeseries.

## Plotting the current data

In this section I am going to plot the data extracted directly from the dataset. The range of the data will cover from March 2017 to February 2020, leaving the last datapoint for the Forecasting area. The data plotted comes in form of Timeseries of indexed datapoints ordered in time. To beautify the plots, I have used the R package “ggthemes” and from there I have found particularly adequate the package “theme\_solarized” which adds a look of old statistical papers of financial journals by having a nice sepia background and a beautiful font.

```
111
112 autoplot(RetiroTS) +
113   theme_solarized() +
114   labs(title = "Passenger utilization in Retiro station", y="Number of passengers", x="Date")
115
116 autoplot(HaedoTS) +
117   theme_solarized() +
118   labs(title = "Passenger utilization in Haedo station", y="Number of passengers", x="Date")
119
120 autoplot(FloridaTS) +
121   theme_solarized() +
122   labs(title = "Passenger utilization in Florida station", y="Number of passengers", x="Date")
123
124
```

Fig. 7 – Coding snippet plotting values for the three stations.

### Retiro plot (current)

We can see a clearly observe strong variations of passenger usage between months, dropping almost to zero passengers in some datapoints. We can also see a slight increment of passenger usage over the years and just from looking at the plot is difficult to see if we have any type of seasonality on this data.

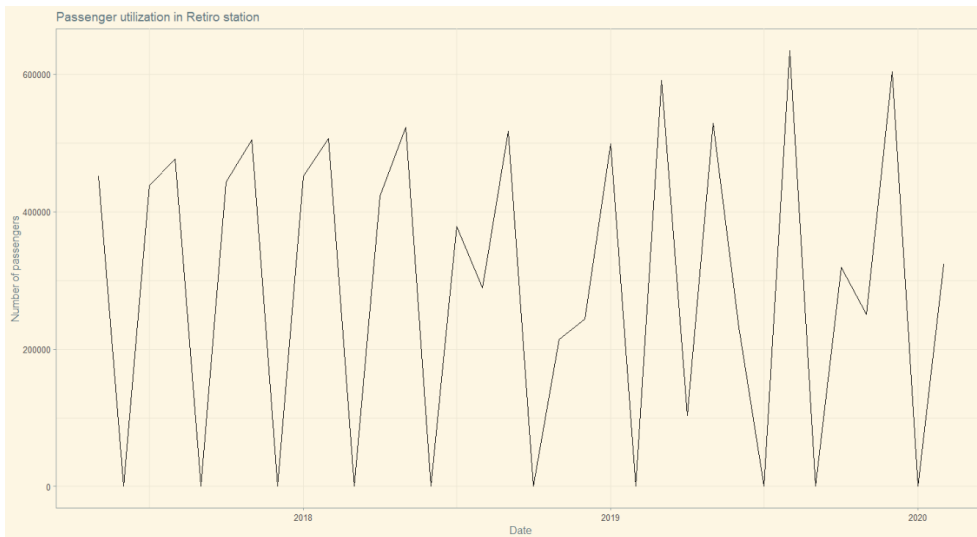


Fig. 8 – Plotting values of Retiro

### Haedo plot (current)

We can see some tendency in Haedo for passenger usage to drop next to zero in some months in the year, right now we can't clearly state any seasonality but later we will have a deeper look using other techniques. We can see a significant increase of usage in the last months of 2019.

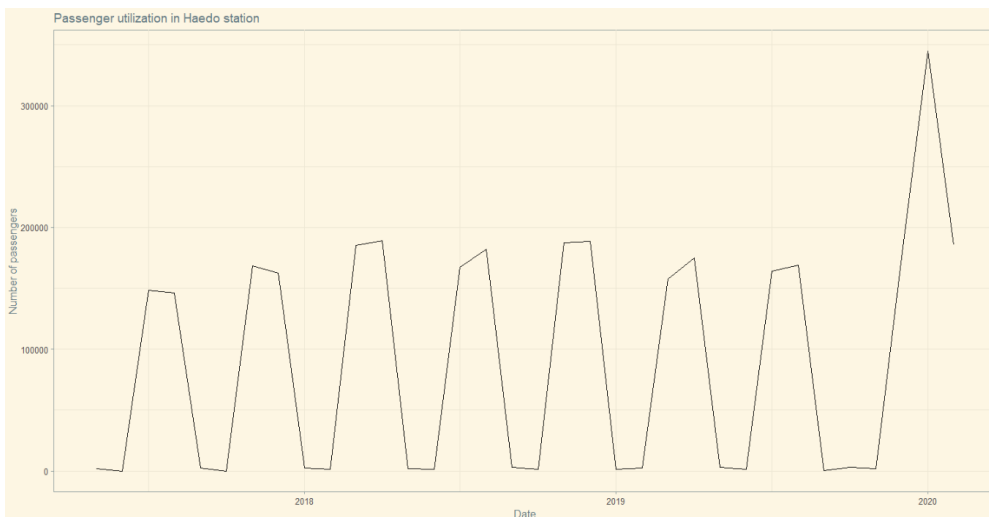


Fig. 9 – Plotting values of Haedo

### Florida plot (current)

At a Glance we can see the Florida has much more consistency in terms of passenger usage, since the number of passengers remain always over 15.000 per month with the exception of the last months of 2019.

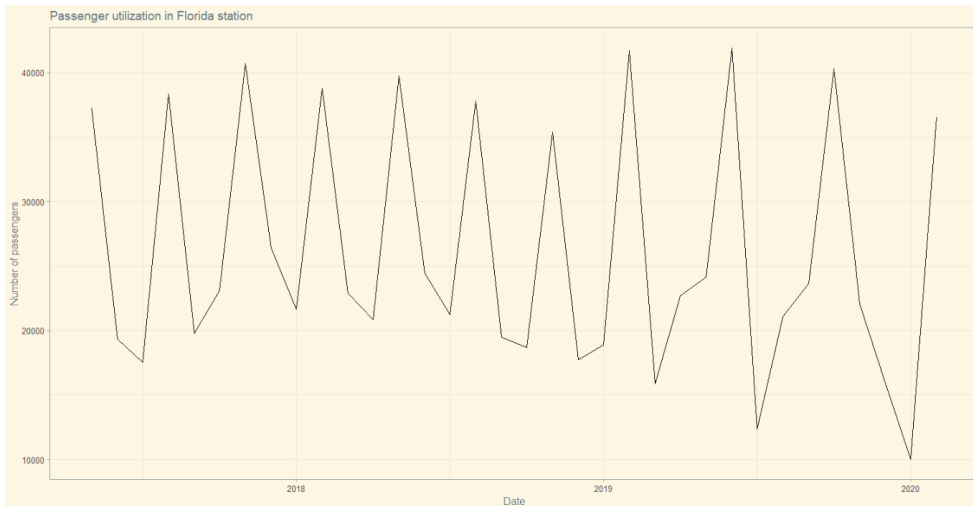


Fig. 10 – Plotting values of Florida

### Analysing the current data

On this section I am going to dive deep into the data and analyse normality and seasonal trends it might have. To do that I am using methods contained within specific data analysis packages.

#### Normality tests

To check for normality I will use Quantile-Quantile plots, histogram and Shapiro test. The null hypothesis formulated on this analysis is the assumption of normal distributed data. Confidence is 95%.

- Quantile Quantile or QQ Plots is a plot where each datapoint is plotted as a single point. A line is drawn to indicate what an ideal normal distribution would be against the datapoints. Having the datapoints next close to the line or on the line is a good sign of normality, as opposed to having the dots far from the line, which indicates non-normality.
- Histograms shows the data represented in bars; on this project the passenger usage will be represented on the Y axis and the number of repetitions will be represented on the X axis. Having a small difference between bars will represent a sign of normality against having big differences between bars will represent non-normality. Having a shape of a bell indicates normality.
- Shapiro Wilk test is used for datasets having not many observations against Kolmogorov S. that is used for large number of observations. On this project the I will choose a standard of 95% of confidence, therefore our P-Value for this test will be 0.05

## Normality tests for Retiro station

### Shapiro-Wilk

Shapiro-Wilk normality test

```
data: RetTrain$cantidad  
W = 0.87171, p-value = 0.0001394
```

Fig. 11 – Coding snippet Shapiro-Wilk Retiro

### GGQQPLOT

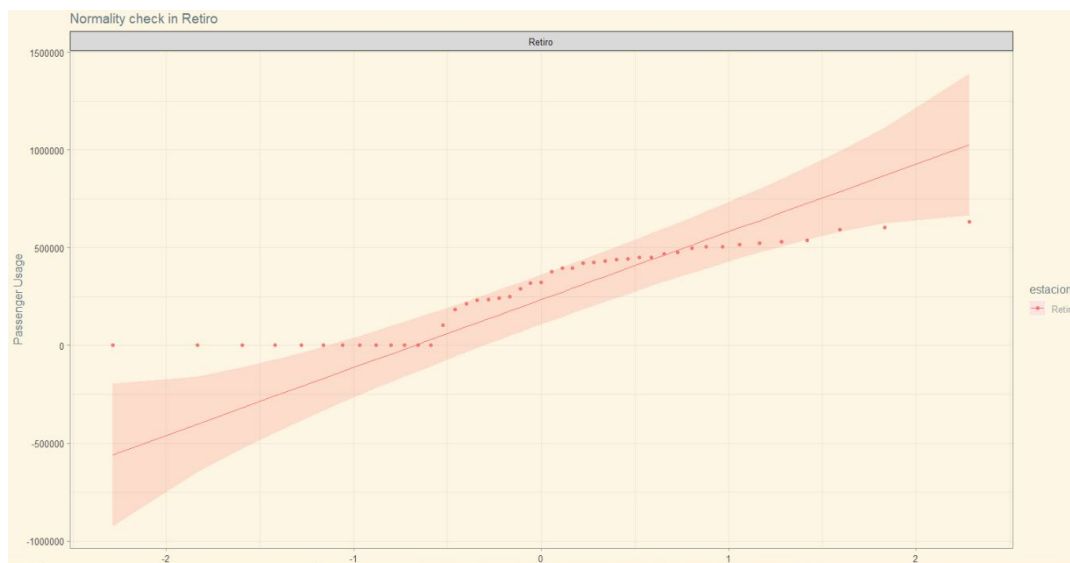


Fig. 12 – Quantile-Quantile for Retiro

### Histogram

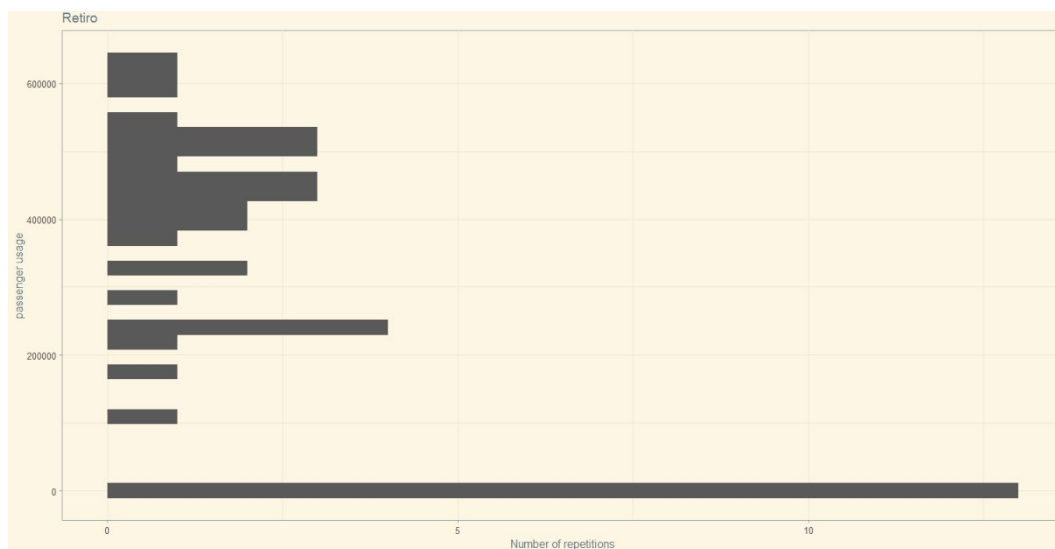


Fig. 13 – Histogram for Retiro

### Interpretation of the normality tests for Retiro Station

Having a look at the Shapiro-Wilk test we can reject the Null Hypothesis since P-Value is less than 0.05 which is the first sign of non-distributed data; the Quantile Quantile plot shows most of the observations far from the 45-degree line that indicated normality, which gives us another sign of non-normally distributed data. Lastly the histogram does not show anything like a bell or a sine wave shape, as repetitions of the observations have an erratic pattern, also we can see a lot of datapoints having zero, which indicates that on that month in the year there was no passenger usage on Retiro station. We can conclude saying that data is non-normally distributed.

### Normality tests for Haedo station

#### Shapiro-Wilk

```
Shapiro-Wilk normality test
data: HaedoTrain$cantidad
W = 0.78045, p-value = 0.00000009717
```

Fig. 14 – Coding snippet Shapiro-Wilk for Haedo

#### GGQQPLOT

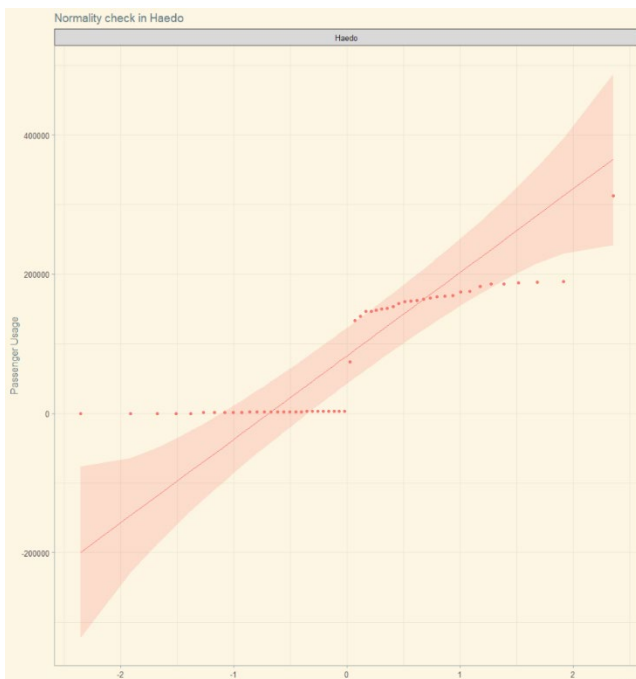


Fig. 15 – Quantile-Quantile for Haedo

## Histogram

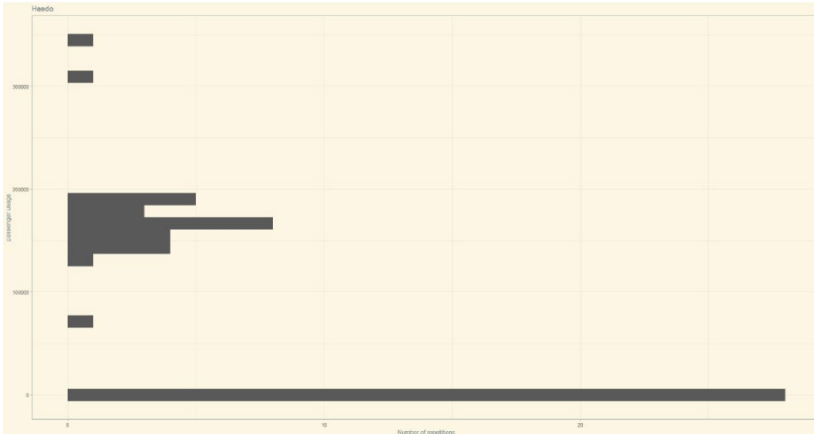


Fig. 16 – Histogram for Haedo

### Interpretation of the normality tests for Haedo Station

Having a look at the Shapiro-Wilk test we can reject the Null Hypothesis since P-Value is less than 0.05 which is the first sign of non-distributed data; the Quantile Quantile plot shows most of the observations far from the 45-degree line, this plot is very characteristic as we can see and horizontal line on the zero value intersecting the line which indicated absence of passengers on these datapoints; these events gives us another sign of non-normally distributed data. Lastly the histogram does not show anything similar to a bell or a sine wave shape, as repetitions of the observations have an erratic pattern, again can see a lot of datapoints having zero, which indicates that on that month in the year there was no passenger usage on Haedo Station. We can conclude that data is non-normally distributed.

### *Normality tests for Florida station*

#### Shapiro-Wilk

```
Shapiro-Wilk normality test
data: FloridaTrain$cantidad
W = 0.88255, p-value = 0.0003854
```

Fig. 17 – Coding snippet Shapiro-Wilk for Florida

# GGQQPLOT

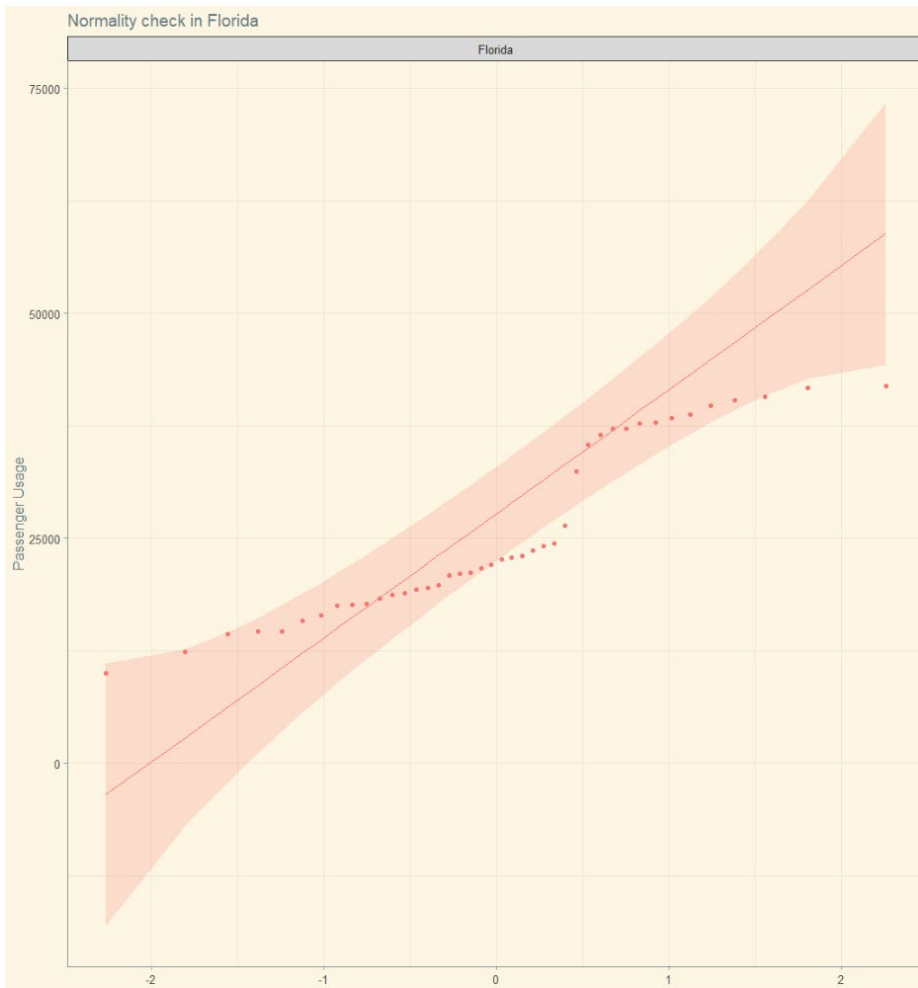


Fig. 18 – Quantile-Quantile for Florida

# Histogram



Fig. 19 – Histogram for Florida

## Interpretation of the normality tests for Florida Station

The Shapiro-Wilk on this test concluded as Florida being the station having the data most close to normal, however we can still reject the null hypothesis since P-Value continues to be way beyond 0.05. Checking the Quantile-Quantile plot shows us datapoints very spaced across the plane but and far from the line that indicates ideal normality; additionally, we can see no zero values plotted, which is a sign of high utilization of the station. Finally, the histogram does not show any hints of bell or sine wave lines across the columns representing the data and we can see big jumps from a utilization perspective. We can conclude this analysis stating that Florida's data is not normally distributed.

## Seasonality

On this section I am going to analyse the existing seasonality across the Argentinian train stations of Florida, Haedo and Retiro.

The term of seasonality refers to patterns found in a dataset overtime. These patterns can darken the result of the forecasts. Having a highly seasonal dataset needs of techniques to remove the patterns found before proceeding on forecasting data.

On this analysis I am going to use partial autocorrelation function to have a view of seasonality. PACF or Partial Autocorrelation Function is the correlation between a timeseries and the lagged version of itself after we subtract the correlation at smaller lags, giving us a better view of the seasonality in the data against the ACF or Autocorrelation Function.

In our data, the X axis will be a representation of the lag over time, time being 12 months in a year. The Y axis will give us the PACF for a specific time over the 12-month period.

To calculate the PACF on R I decided to use `ggpacf()` function from `ggplot2` package, as the regular `pacf()` function included in R did not allow themes.

```
ggPacf(RetiroTS) + scale_x_continuous(name="Lag", limits=c(1, 12)) + theme_solarized()
ggPacf(HaedoTS) + scale_x_continuous(name="Lag", limits=c(1, 12)) + theme_solarized()
ggPacf(FloridaTS) + scale_x_continuous(name="Lag", limits=c(1, 12)) + theme_solarized()
```

Fig. 20 – Coding snippet for Partial Autocorrelation function



## PACF RETIRO



Fig. 21 – PACF Retiro

## PACF HAEDO

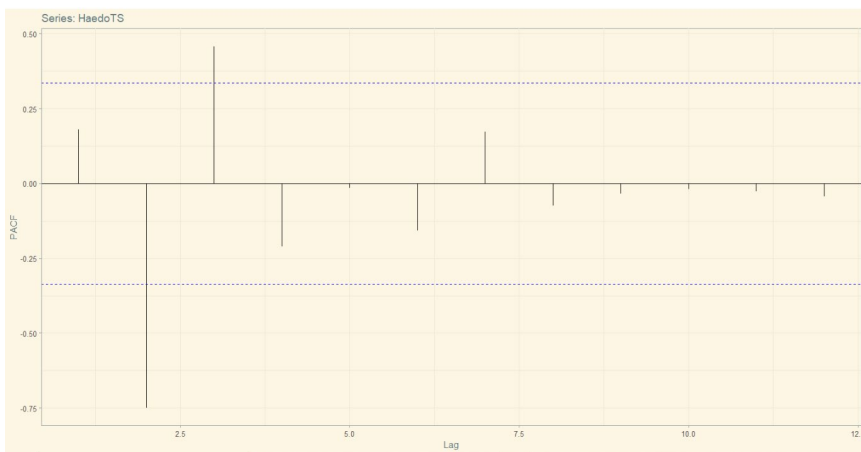


Fig. 22 – PACF Haedo

## PACF FLORIDA

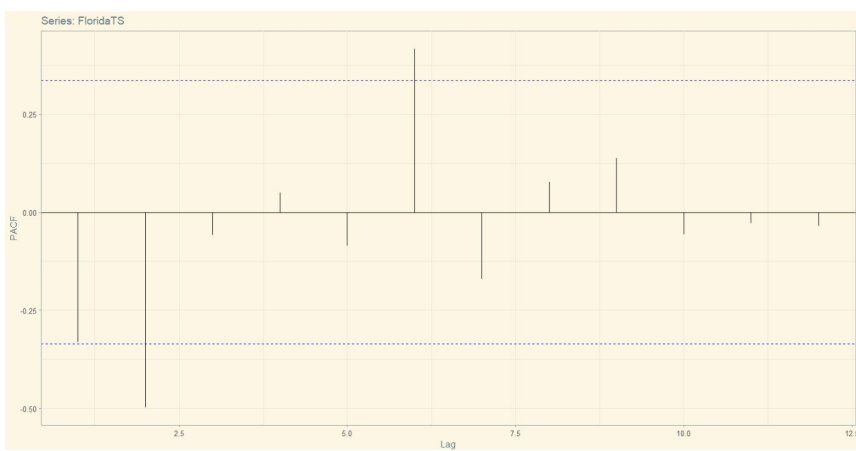


Fig. 23 – PACF Florida

## Interpretation of PACF on all three stations

Looking at the three PACF graphs we can observe only a few lag measures going over the positive and negative correlation control lines. Retiro has some correlation associated with the first two months of the year having two bars surpassing control lines. Haedo has two bars as well on the first and third month of the year, and lastly, Florida has a bar in February and another one mid-year passing the control line.

Based on observation of the PACF plots we can conclude that seasonality on the data selected is not statistically relevant to apply any correlation removal techniques, as it will not interfere in the forecasts.

## Forecasting passenger utilization on the selected Stations

In our last section of the Analysis, we are going to use two different Forecasting methods, and finally compare the output of them when making predictions of the data.

- Autoregressive Integrated Moving Average or ARIMA is a forecasting method used in time series data. The model supports autoregressive and moving average data points, however the downside of the model is how sensitive it is to the seasonality component over the data; for that reason I am going to build two different analysis and compare them. The first analysis will be based on using the `auto.arima()` function with no modificatory arguments, that will include some seasonality and trend found previously on the analysis of current data, secondly I will run another `auto.arima()` session using built in arguments of the `auto.arima` tool to make another model removing seasonality and trend (`d` and `D` arguments) that will created a differenced model with no seasonality and trend patterns on it.
- AutoTS is a collection of functions from Prophet and Forecast libraries. `getBestModel()` is the main and most complete function that recalls up to seven algorithms having a “bagged” argument that takes the data from all the algorithms computed and funnels into a single mean value. From all seven algorithms that `getBestModel` can recall I will only use two; For our study I will recall Prophet forecasting model and also SARIMA. Prophet is a model developed by Facebook that works particularly well with data having trend and seasonality. SARIMA is a model based on Arima that has been optimized to automatically remove seasonality of a dataset to compute the model

## ARIMA

We are going to commence our Forecasting session with Arima. First it is required to build a model using `auto.arima()` function, and then we can forecast straight away using `forecast()` function; of the first series of examples I will not remove trend and seasonality, I will compute arima “as is”. On the second series of results I will use the “`d`” and “`D`” arguments that will create a difference of the model and therefore get rid of any seasonal patterns found in the data.

```
185  
186 #Applying ARIMA having some trend & seasonality  
187  
188 Retiro_arima <- auto.arima(RetiroTTS)  
189 print(Retiro_arima)  
190
```

Fig. 24 – Coding snippet for Arima model with seasonality and trend

```

200
201 #forecasting using ARIMA with trend&seasonality
202
203
204
205 fcst_ar_Retiro <- forecast(Retiro_arima,h=1)
206
207 fcst_ar_Retiro
208
209 autoplot(fcst_ar_Retiro) +
210   labs(subtitle="Retiro station", y="Number of passengers", x="Date") +
211   theme_solarized()
212
213
214

```

Fig. 25 – Coding snippet for a forecast based on Arima model with seasonality and trend

*Arima – no trend and seasonality removed*

## RETIRO

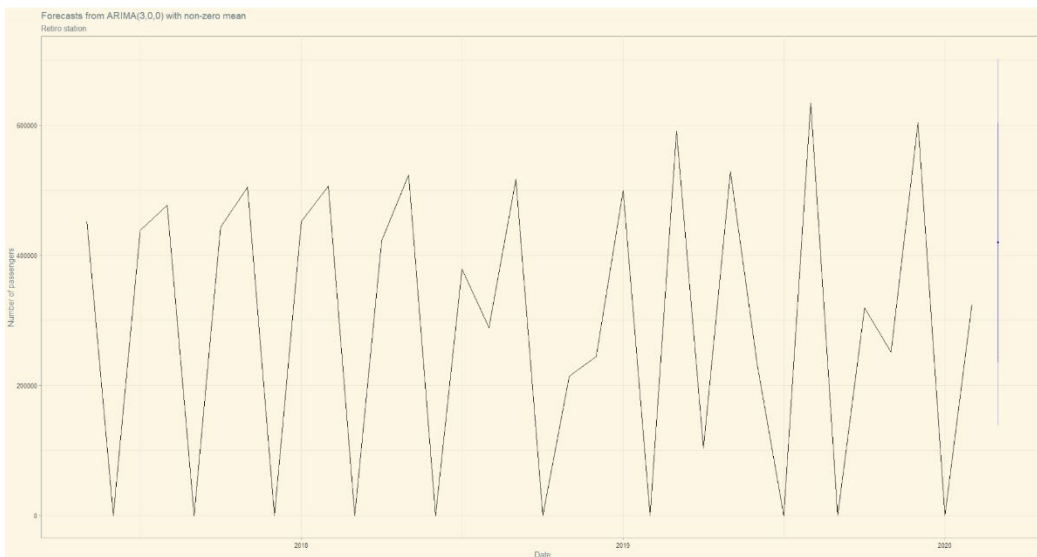


Fig. 26 – Plot of Forecasting the month of March 2020 in Retirousing ARIMA Model. No Trend and Seasonality has been removed.

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Mar 2020	419627	235604.6	603649.3	138189	701064.9

Fig. 27 – Results of forecasting March 2020 using ARIMA Model in Retiro. No Trend and Seasonality has been removed.

## Interpretation

Fig. 27 shows us a line plot showing a graphical representation of the passenger usage on the Retiro station and on the right side a Point forecast on 419627 passengers; the blue shaded line corresponds to the error threshold associated to the forecasting model. We can observe a minimum usability of zero passengers on

some periods and a maximum peak of more than 640.000 passengers in some periods over the years of 2019 and 2020. The results of the forecast seem to be ponder looking at the variability of the graph.

### HAEDO

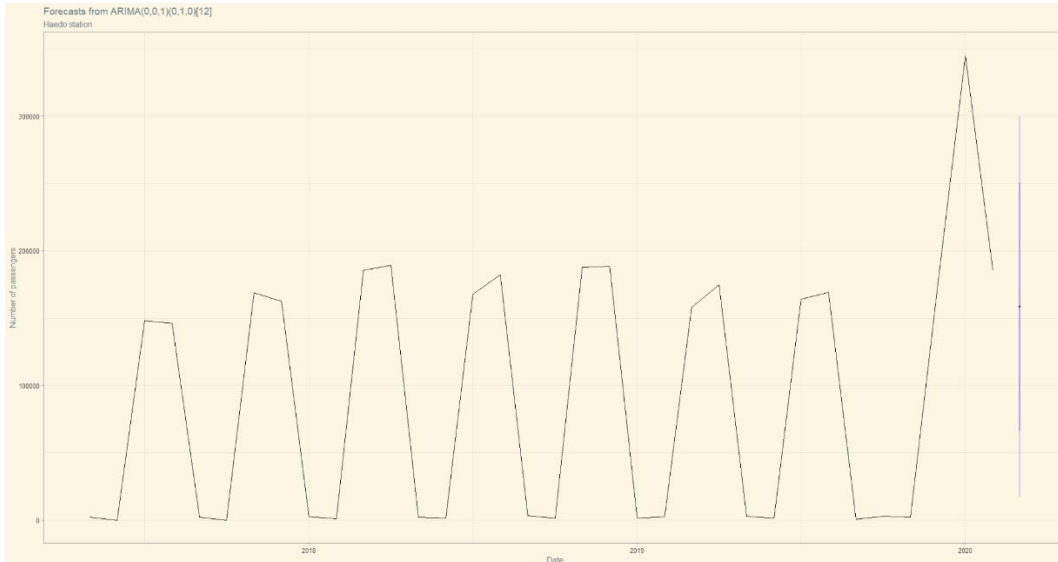


Fig. 28 – Plot of Forecasting the month of March 2020 in Haedo using ARIMA Model. No Trend and Seasonality has been removed.

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Mar 2020	158231.1	65920.95	250541.2	17054.92	299407.2

Fig. 29 – Results of forecasting March 2020 using ARIMA Model in Haedo. No Trend and Seasonality has been removed.

### Interpretation

Fig. 29 shows us a line plot showing a graphical representation of the passenger usage on the Haedo station and on the right side a Point forecast on 158231 passengers; the blue shaded line corresponds to the error threshold associated to the forecasting model. On this case we can observe a higher error threshold compared with Fig.27 . The usage is clearly lesser as well. The peak period corresponds to the last months of 2019. As opposed to Fig27 the forecasted result seem to be overly optimistic looking at the graphical representation.

## FLORIDA

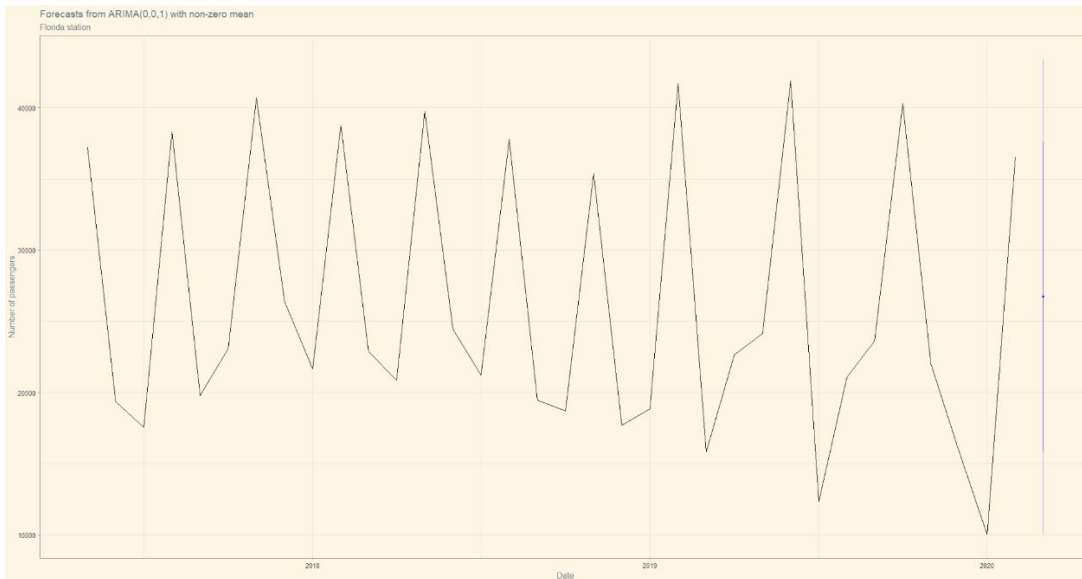


Fig. 30 – Plot of Forecasting the month of March 2020 in Florida using ARIMA Model. No Trend and Seasonality has been removed.

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Mar 2020	26735.06	15838.85	37631.26	10070.75	43399.37

Fig. 31 – Results of forecasting March 2020 using ARIMA Model in Florida. No Trend and Seasonality has been removed.

### Interpretation

Fig. 31 shows us a line plot showing a graphical representation of the passenger usage on the Florida station and on the right side a Point forecast on 26735 passengers; the blue shaded line corresponds to the error threshold associated to the forecasting model. By far we can notice Florida being the station with lowest passenger usage in terms of volume of passengers as opposed to the highest usage in terms of steadiness, since there is usage every month in the year but in the period very next to 2020. Looking at the point forecast we can observe a pessimistic forecast compared with Fig. 27 and Fig.29. Maximum and minimum errors forecasting are the highest observed compared to the previous graphs.

Arima – trend and seasonality removed from the model

## RETIRO

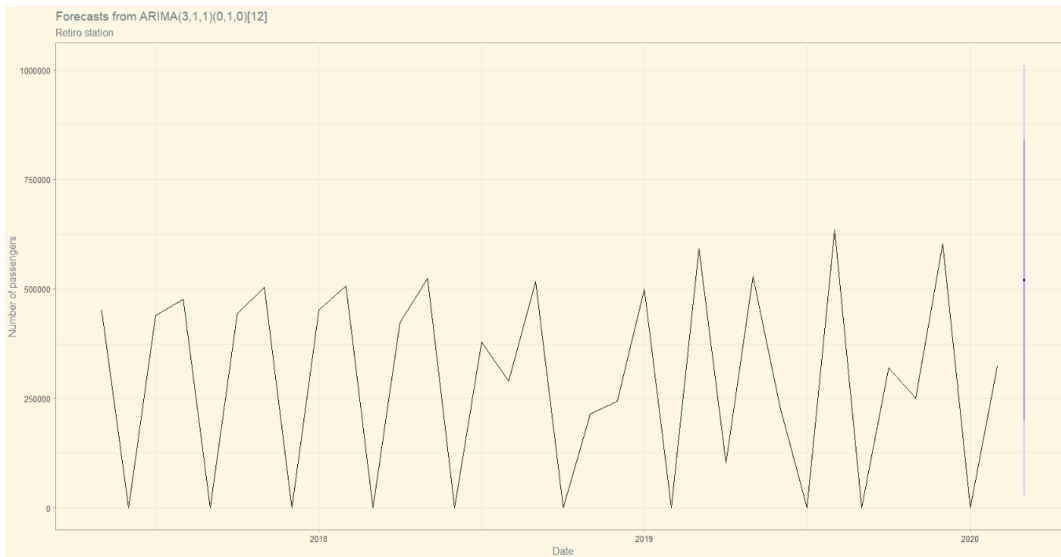


Fig. 32 – Plot of Forecasting the month of March 2020 in Retiro using ARIMA Model. T & S Removed from the model

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Mar 2020	519578.6	198295.7	840861.6	28218.73	1010939

Fig. 33 – Results of forecasting March 2020 using ARIMA Model in Retiro. T&S Removed from the model

## Interpretation

Fig. 33 shows us a line plot showing a graphical representation of the passenger usage on the Retiro station and on the right side a Point forecast on 519578 passengers. We can clearly see that removing trend and seasonality varied greatly the results of the forecast since Fig.27 shows 419627 passengers. The graph required a different scale to allocate a higher error as an impact of the additional arguments used in auto.arima function

## HAEDO

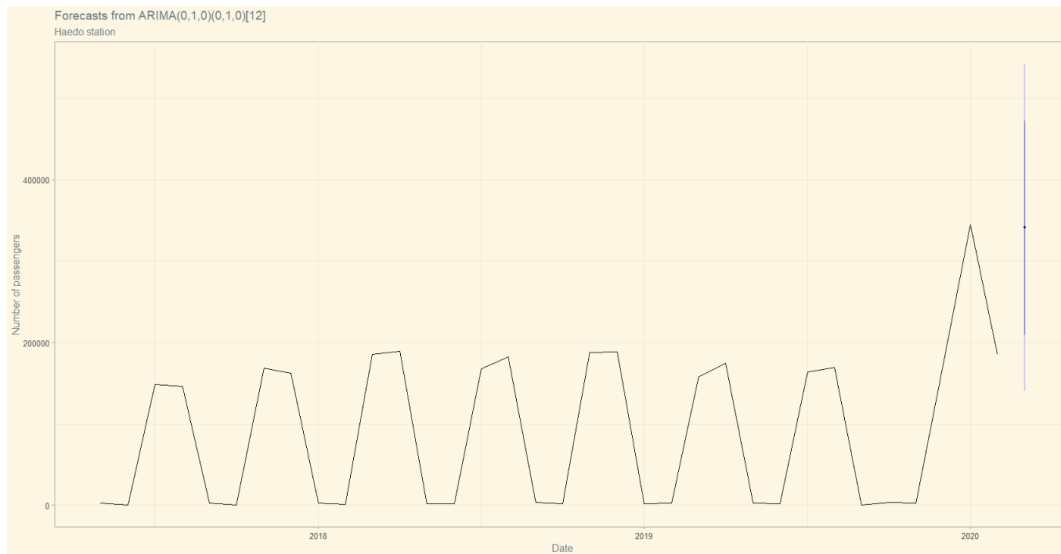


Fig. 34 – Plot of Forecasting the month of March 2020 in Haedo using ARIMA Model. T & S Removed from the model

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Mar 2020	340822	209190.1	472453.9	139508.4	542135.6

Fig. 35 – Results of forecasting March 2020 using ARIMA Model in Haedo. T&S Removed from the model

### Interpretation

Fig. 35 shows us a line plot showing a graphical representation of the passenger usage on the Haedo station and on the right side a Point forecast on 340822 passengers. We can clearly see that removing trend and seasonality varied greatly the results of the forecast since Fig.29 shows 158231 passengers, having a forecast with more than the double than observed with no seasonality and trend removed. The graph required a different scale to allocate a higher error as an impact of the additional arguments used in auto.arima function.

## FLORIDA

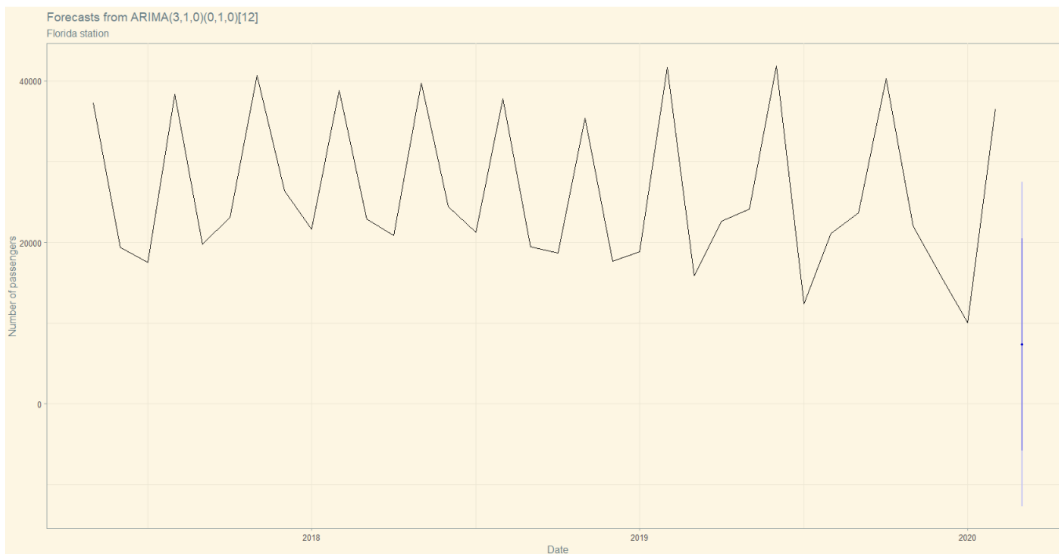


Fig. 36 – Plot of Forecasting the month of March 2020 in Florida using ARIMA Model. T & S Removed from the model

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Mar 2020	7357.427	-5788.416	20503.27	-12747.41	27462.26

Fig. 37 – Results of forecasting March 2020 using ARIMA Model in Florida. T&S Removed from the model

### Interpretation

Fig. 35 shows us a line plot showing a graphical representation of the passenger usage on the Florida station and on the right side a Point forecast on 7357 passengers. We can clearly see that removing trend and seasonality varied greatly the results of the forecast since Fig.31 shows 26735 passengers, having forecasted almost 4 times less than the previous model having Trend and Seasonality. The graph required a different scale to allocate a higher error as an impact of the additional arguments used in auto.arima function.

## AUTOTS

AutoTS is a series of methods that bring can bring up to seven different algorithms into play to produce a model and forecast. On this section of the project we are going to review how AutoTS behaves on the three selected stations and we are going to bring use the Prophet model and SARIMA. Prophet is a powerful and modern model developed by Facebook that allows forecasting on timeseries.



```

293
294 ## Using getBestModel() function from AutoTS package
295
296
297 AutoTSBMRet <- getBestModel(RetTrain3x$mes, RetTrain3x$cantidad, freq = "month", complete = 1,
298                             n_test = NA, graph = TRUE,
299                             algos = list("my.prophet", "my.sarima"),
300                             bagged = "custom",
301                             metric.error = my.rmse)
302 View(AutoTSBMRet)
303
304 AutoTSBMHae <- getBestModel(HaedoTrain3x$mes, HaedoTrain3x$cantidad, freq = "month", complete = 1,
305                             n_test = NA, graph = TRUE,
306                             algos = list("my.prophet", "my.sarima"),
307                             bagged = "custom",
308                             metric.error = my.rmse)
309
310 View(AutoTSBMHae)
311
312 AutoTSBMFlo <- getBestModel(FloridaTrain3x$mes, FloridaTrain3x$cantidad, freq = "month", complete = 1,
313                             n_test = NA, graph = TRUE,
314                             algos = list("my.prophet", "my.sarima"),
315                             bagged = "custom",
316                             metric.error = my.rmse)
317
318 View(AutoTSBMFlo)
319

```

Fig. 38 – Coding snippet of AutoTS method getBestModel for every station.

Fig.39 shows the use of getBestModel() function from AutoTS package and recalling both Prophet and SARIMA algorithmic models into play. This function creates a model based on a timeseries data and forecasts a full year in advance using all selected models; it does provide a “bagged” result which compares all the algorithms selected and gives a median value.

## RETIRO

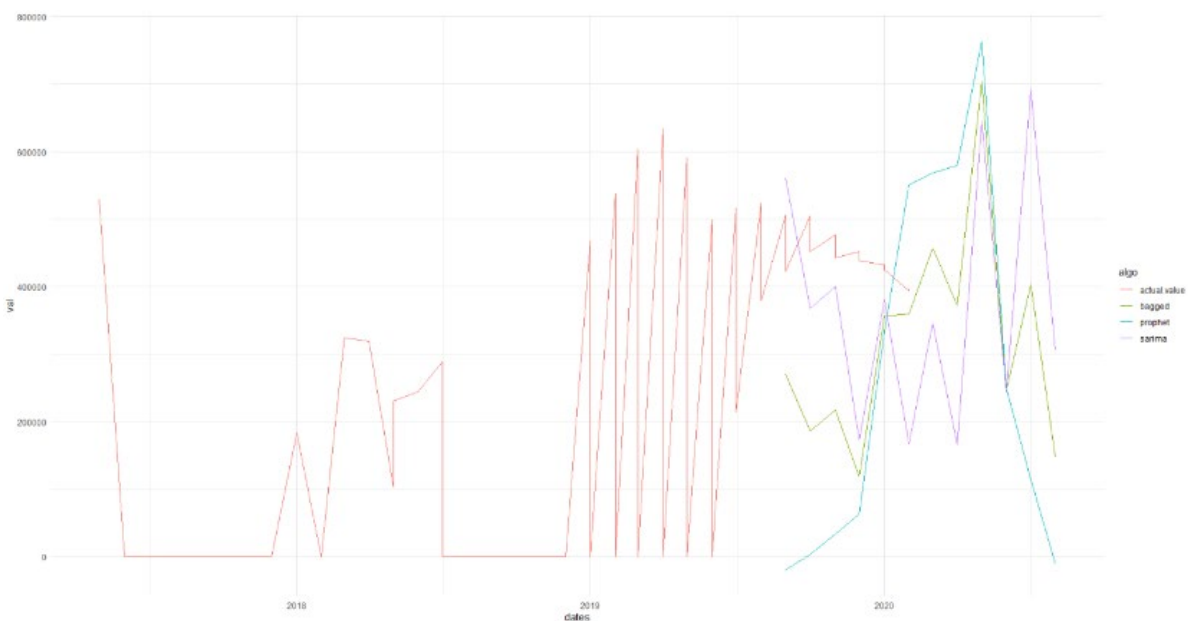


Fig. 39 – Graph generated by getBestModel showing results on Retiro for Prophet, SARIMA & Bagged of both.

	dates	type	prophet	sarima	bagged	actual.value
96	2020-03-01	mean	568500.073	345779.595	457139.83	NA

Fig. 40 – Table showing numeric results from getBestModel on Retiro for Prophet, SARIMA & Bagged of both

### Interpretation

We can observe in the graph on the Fig.39 the difference between the algorithms results. At a first glance, prophet seems to be very optimistic against SARIMA with turns to be the opposite. Prophet predicts 568500 passengers and SARIMA 345779 for the month of March 2020.

### HAEDO

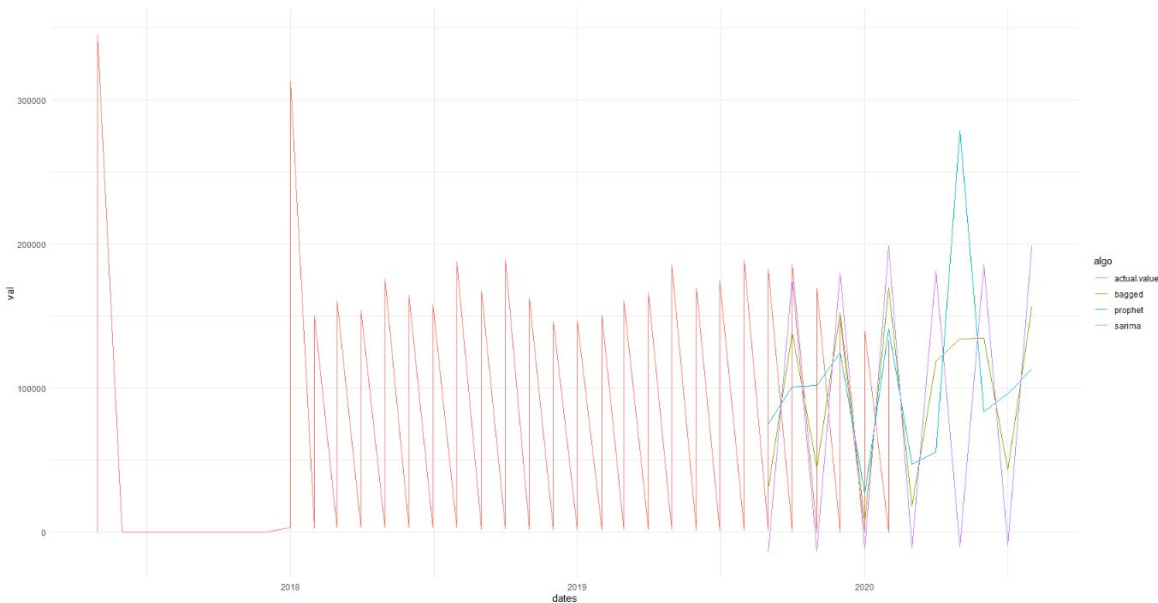


Fig. 41 – Graph generated by getBestModel showing results on Haedo for Prophet, SARIMA & Bagged of both.

	dates	type	prophet	sarima	bagged	actual.value
129	2020-03-01	mean	47343.754	-10779.789	18281.983	NA

Fig. 42 – Table showing numeric results from getBestModel on Haedo for Prophet, SARIMA & Bagged of both

### Interpretation

We can observe at a first glance, how prophet seems to be very optimistic against SARIMA with turns to provide a negative result, something that should be instantly discarded from our study, since passengers cannot be less than zero in any given period for obvious reasons Prophet predicts 47343 passengers and SARIMA -10779 for the month of March 2020.

## FLORIDA

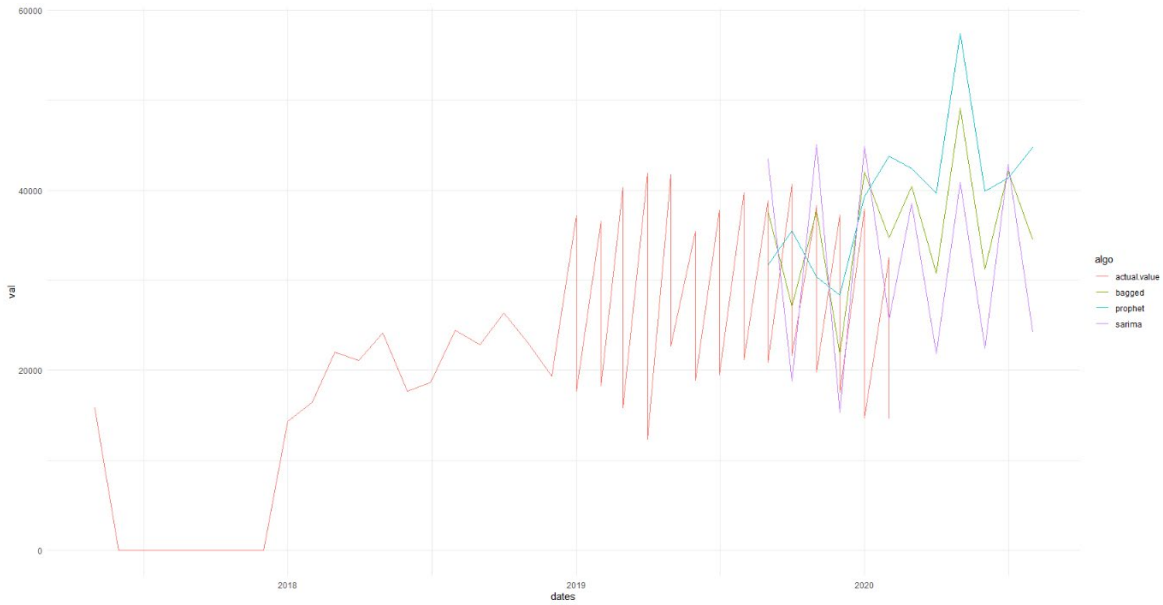


Fig. 43 – Graph generated by getBestModel showing results on Florida for Prophet, SARIMA & Bagged of both.

	dates	type	prophet	sarima	bagged	actual.value
90	2020-03-01	mean	42392.33	38507.963	40450.145	NA

Fig. 44 – Table showing numeric results from getBestModel on Florida for Prophet, SARIMA & Bagged of both

### Interpretation

We can observe in the graph on the Fig.43 the difference between the algorithms results. We can see again the trend on Prophet's optimism and SARIMA pessimism. Prophet predicts 42392 passengers and SARIMA 38507 over the month of March 2020.

## Forecasting results compared and conclusions

	Results			Percentual difference		
	RETIRO	HAEDO	FLORIDA	RETIRO	HAEDO	FLORIDA
ARIMA W/T&S	419627	158231	26735	72%	328%	59%
ARIMA WO/T&S	519578	340822	7357	113%	821%	-56%
PROPHET	568500	47343	42392	133%	28%	152%
SARIMA	345779	-10779	38507	42%	-129%	129%
BAGGED P&S	457139	18281	40450	88%	-51%	140%
<b>REAL VALUES 03/20</b>	<b>243621</b>	<b>37008</b>	<b>16831</b>			

Fig. 45 – Table showing numeric results and Delta percentages against real results of March 2020

After utilizing all the selected analysis techniques, I have gathered all the results in the tables over the Fig.45 in order to have another way of comparison, not only graphical.

The table in the right shows the results of ARIMA W/TS (Arima with Trend and Seasonality components in the data), ARIMA WO/TS (Arima with Trend and Seasonality removed using auto.arima arguments), Prophet and SARIMA, and the Bagged value of combining the results from Prophet and SARIMA on the Auto.TS Analysis.

On the bottom of the Fig.45 we can see the actual real results of the passenger utilization over the month of March 2020 and on the right we can see the difference between the Results on the left and the actual Real results on the bottom visualized in a percentual way having colour and measure bars for the ease of the interpretation.

The conclusions we can get from the Analysis is as follows.

- From the chosen models, we can clearly ponder the balance in favour or any for the three stations.
- Prophet worked well on Stations having more periods without passengers (Haedo) and did not perform that well with the other two stations having high utilization more often, as the model clearly performed as too optimistic.
- Arima worked well with the Stations capturing passenger utilization more often and performed poorly on stations with stations having more extended periods without utilization (Haedo). This statement can be used for any of the two Arima models used, nevertheless, standard Arima ran using auto.arima function with no arguments proved to perform much better than Arima using D and d arguments to eliminate trend and seasonality; as stated in the analysis on the Current Data, the Datasets did not contain excessive seasonal patterns, therefore applying differential arguments to remove the seasonality and trend did not help on getting better performance, and in consequence accuracy.
- Looking at the results, if I had to choose a model over another to forecast on the three stations I would move for Prophet as the model adapted well to every situation shown in the data, unlike Arima that shown poor accuracy when having patterns in the data next to real zero.

END



# Appendix

## Project Proposal

### Background

The project is going to be oriented about one of my passions, which is trains. I've been fascinated with trains since I was a kid. I could not understand how the human was capable to move and stop a 157 Ton huge metallic snake so easily.

Considering the tough times that we live I wanted to take a detour if the most common topic, which the is COVID-19. I consider that the virus not only has infected a good part of the human beings but all news, conversation and mostly our lives. That is my main reason to talk about trains, in this moment.

The purpose of my research is going to be oriented to providing forecasted data to the Spanish Railway Network Association (RENFE) in order to react to possible events on different dates and various train routes.

### Objectives

The purpose of this research is to provide a series of Forecasts to the Spanish Railway Network Association in order to mitigate any possible events around passenger utilization spikes over multiple routes across Spain. The study will be focused on the most frequented routes, which are Barcelona, Madrid, Bilbao and Valencia. The report will be divided into two sections. First section will cover a comparison between current passenger utilization and data extracted from 2018, presented in a graphical form and interpreted. The second section will go throughout a range of forecasts of passenger utilization spikes on weekends, bank holidays, events like concerts or football matches, and peak times like Christmas or July and August holidays.

The problem that the project is going to address is to allow RENFE to tackle any impact on their services during peak times by forecasting usage on these dates and increase capacity on their trains effectively. That will avoid the company to lose ticket sales as for a possible out of seats scenario, and also to choose the best way to increase the number of seats depending on the results of the forecast; would adding wagons to the train will be more effective than sending another train?

## Technical Approach

After an initial research on different Railway entities of several countries, I have found RENFE's one as the one with more usable resources, as various datasets presented in .CSV and XSLX formats, across with an API endpoint for each of the routes, which will be used for the forecasting part of the report.

The data sets used can be found on <https://data.renfe.com/> and also some examples to API consults.

## Technical Details

### Workstation:

- Lenovo Thinkpad T480:
  - RAM: 32 GB
  - Processor: Intel Core i7-8466U @ 1.90GHz

### Programming languages, IDE's and Libraries used.

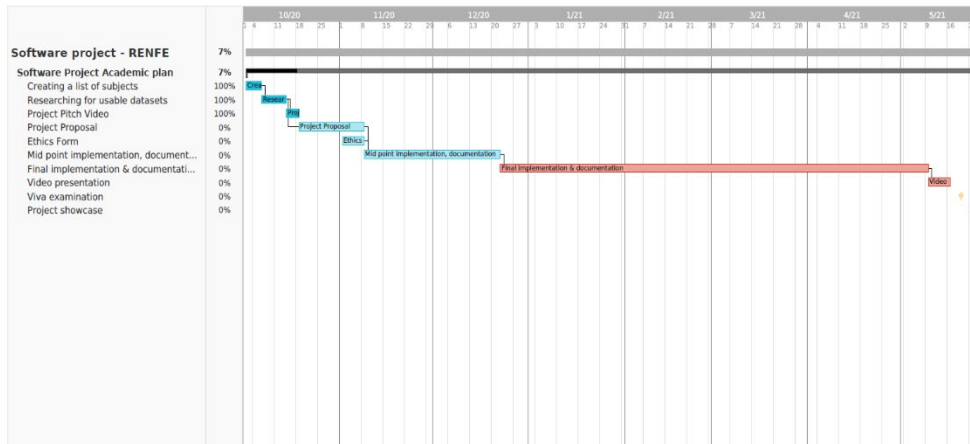
- R
  - GGLOT2 – Data processing and visualisation library
  - Forecast – library containing AUTOARIMA function, that enable us forecasting
  - readXL – Library used to read excel spreadsheets amongst other tools to manipulate excel sheets
  - ggthemes – library containing preloaded themes for better data visualisation and make the experience richer to the user.
  - RStudio – IDE for R
- Python
  - Prophet – Open source library developed by Facebook designed to work on timeseries data
  - IDLE / PyCharm – both most commonly used IDEs for Python, still to be defined which one is more suitable to work on this project.

### Software

- Windows 10
- Word
- Excel
- Lucidchart

## Project Plan

Below is the Gantt chart showing the dates to complete the academic Project.



### Technical details

To be filled up for the next submission of the document. I have not technical details yet as I have not started to work on the data sets just yet.

### Evaluation

To be filled up once having data to evaluate. At this stage of their project there is no data to process and no content can be added on this section

### Invention disclosure form

Will be removed for the mid-implementation submission.



## Monthly Journals

October 2020

Software Project Journal: October 2020

Student: Alejandro Diaz Salgado

Programme: BSHC in Computing, year 4, Data Analytics

Times are tough with the Covid-19 problem and I am pursuing a career in software engineering while working full time in a technical support job and trying to get technical certifications to get promoted in my job.

As we are unable to attend class, they have had to adapt the educational program to the current situation, and it is not possible to take written exams. The alternative to written exams is projects, which makes this year's workload completely unsustainable.

Currently I do not find much motivation to get to work in all the CAs for the subjects, but I think I have reached the point where I have to get to work, or I will not be able to cope with the workload that I have before Christmas.

Regarding the software project, we have been assigned to Enda Stafford, a teacher who gave us two subjects in the previous years. In general, I don't consider him a bad teacher, although from my point of view he only sticks to the program and the content they give him, which makes the class extremely boring and linear. They promised us that they would assign tutors in the second week of class, and they have assigned them in week 6, which does not give much time to change the direction of the project if it turns out to be unworkable.

My tutor is called Giovanni, she has given me feedback on my project, and I am not very happy with his comments. She says that the video is not clear as to what the objective of my project is going to be, although I do seem to have mentioned it. My goal is to improve train routes from the point of view of the number of passengers that get on at each train stop.

In the last week of the month, I have been working full time and attending a Kepner Tregoe course on troubleshooting, which has not left me much extra time to discuss the project with Giovanni. We have agreed to discuss the project on Monday, November 2, so I hope to make you change your mind and not have to go for other options.

November 2020

Software Project Journal: November 2020

Student: Alejandro Diaz Salgado

Programme: BSHC in Computing, year 4, Data Analytics

November has been a very tough month for me, as I had some personal issues related to my family and COVID19. I find 4<sup>th</sup> year having twice as the workload as we had from 3<sup>rd</sup> year, and the reason behind is the college not having the ability to perform closed book assessments, which impact us, the students in having one or two major project per subject, adding up a lot of extra hours of work outside classes; for us, part time students that work full time and study, it's a bit of an unsustainable situation.

The time passed and we are next to Christmas, and unfortunately, I could make barely any progress at all on my project, since I focused in delivering the assignments for subjects due this semester.

The only time I spent on the final project was to research how to implement the deliverables in the cloud, and I decided to use Hadoop & MapReduce for it. The idea is to spread all the computing load into several nodes deployed in the cloud using Hadoop's architecture. For now I just checked some simulations for pricing and determined that Google Cloud Dataproc will be the most suitable cloud platform to implement the structure, it has the lowest cost per hour regardless of the number of nodes we use and it provides the maximum amount of free-tier usability, with up to 300\$ with absolutely no commitments to purchase the paid version afterwards.

About how to do visualisations on the data, I mentioned in October's Journal the idea of using R and ggPlot2, but after working with that stack on the project for Data application development I realised how much I hate R, as it's a programming language I find completely different if I have to compare it with Python or Java, therefore it's more likely that I will run Python for visualisation and data pre-processing parts.

I have investigated also about the two main frameworks for Data Analysis, as I had to run that investigation also for the Data Application development subject, and I'm far more comfortable using KDD rather than CRISP-DM, as it makes much more sense to me, the idea of cleaning, pre-processing, model, mine and present the data.

## December 2020

Software Project Journal: December 2020

Student: Alejandro Diaz Salgado

Programme: BSHC in Computing, year 4, Data Analytics

It is third of January and today is my last day to submit this Journal to gather all the progress and any comments about how December was for me. I took notes over December in a separate document and now I will reflect them on this paper.

There you go 2020, I bet most people were waiting for to say these words anxious after all facts that happened on this historical year, not good facts unfortunately.

December was another hectic academic month. Since the college was not able to support closed book exams, December meant a month loaded of tremendous projects for us part time students. After all these years in college, December 2020 meant the most though month for me so far, but as I am writing these words right now, on a 3<sup>rd</sup> January, looks like we made it!

On terms of the Final year Data analytics project, we were forced to submit some part of the project to show the progress we have made so far on it. In my humble opinion, I don't think that it was the right way to ask for it. In any of the past years were students had closed book exams that did not require a great amount of time typing in a keyboard, that should have been a natural requirement, but on this academic year, that closed book exams are just not possible, and we are required to deliver great wordcount project with a lot of technical complexity added, the fact of having to present progress, and the fact that progress was graded, made a complete non-sense in my view. We are required to submit projects due on a December worth 2, 3 or 4 thousand words, in order to pass the module in January, but also to deliver a good chunk of progress in a project due in May? I think the right way to put this ask to part time students was to decrease the grading from 25% of the final grading to NULL, not just because of the workload, more over because of the fact that we have 2 more modules in semester 2, and one of it is KEY to deliver the final year project in Data analytics, which is Data Mining, just as a side note, these words above are part of my thinking and opinion and hope not to offend anyone in the way, it was an honesty exercise so far.

About the project, my project supervisor was a great help, as he pointed me in the right direction in terms of what and why questions to be addressed in the results. Once I started hands on the project, I had a big roadblock, that made me change the course of it. My first intention was to analyse the Spanish railway network and provide a time series forecasting over several stations in Spain, but the inability of accessing the data was a turning point. The Spanish service had an API endpoint to retrieve information, and my initial though was that API had stored historical information, and therefore accessible just changing parameters, but after tweaking it, I realised that data was only accessible from 1 day, yes 24 hours; because of this reason I had to research a new target, and fortunately I found a similar service on the Argentinian railway network.

As per the above, my studies are going to be swapped from the Spanish railway network to the Argentinian railway network that offers several years of data. The data is not as granular as Spanish website, but still big enough to use it for prediction analysis.

## January 2021

Software Project Journal: January 2021

Student: Alejandro Diaz Salgado

Programme: BSHC in Computing, year 4, Data Analytics

Merry xmas! This is a time of the year I personally enjoy as I get to see my distanced relatives and some friends that would only show up during these times. In terms of the project, I decided to take some break of college over January, as per the general and personal situation. Working and studying is never easy, so I think us “professional students” need to take some time off to avoid any brownouts and to freshen the mind. Aside from the sabbatical time I went ahead and made some research in useful technologies for the project. Found a nice forecasting pack called Prophet, which is developed and distributed by Facebook for free. We can find features like Saturating Forecasts, Seasonality, Trend Changepoints, Outliers and others.

## February 2021

Software Project Journal: February 2021

Student: Alejandro Diaz Salgado

Programme: BSHC in Computing, year 4, Data Analytics

We all had high expectations over 2021 but by the looks of it is not going to be a normal year at all. I’m finding studying from home as something awkward and difficult; I find making questions to the lecturers not as easy as if we were in class in front of them, the relationship with my colleagues from the past years suffered a lot for the lack of seeing each other, and on top of that concentrating at your room surrounded by distractions is a lost cause (The NCI library is really missed at the moment I believe).

Since we have a deadline of May to hand off the project, I found needed to have a bit of slack in January and take it a bit of sabbatical, just in terms of the final project. I could not stop attending to classes otherwise I would lose the track of the events and end up in February with no clue of what they are talking about. I mentioned about attending to classes as I found them as a really good resource for my project, we are covering advanced analytics in both the modules. Advanced Business Data Analysis is covering more the interpretation of the analytics rather than include completely new concepts if compared to the last semester. This is really good since I found a bit of lack of interpretation on the analyses, we made in class last semester, and will help hugely in order to expose the deeps of my project. For the Data Mining module, I find that we are covering very advanced data analytics procedures and hoping to get exposure to an algorithm that allows me to incorporate some data mining into my project. For now, we covered mainly K-means algorithm, which I don’t find it a good fit to make predictions from time series. I hope to get more research done and start showing some actual work done over the month of March.

April 2021

Software Project Journal: April 2021

Student: Alejandro Diaz Salgado

Programme: BSHC in Computing, year 4, Data Analytics

March and April are usually tough months in terms of workload, and 2021 was no different. Was not able to progress much on my project through April because of having the TABA's for Data mining and Advanced Business Data Analysis.

As we are approaching the date to deliver the Final Project, I want to leave here some reflections that I gathered through all this time.

We started the year four with an expectation of producing an analysis on some dataset of our choice by using Data Analytics and Data Mining Techniques. During the past three years we, as Computing grad students, have learned concepts related to Networking, Software Engineering, Systems, Operative systems, and the industry of software, and then we were "Forced" to pick up a branch of specialisation. Speaking about myself, from all the options given, I chose Cybersecurity as my primary option, as I consider myself as someone fluent in terms of computer networking and network security, however the only options available we had (being Part Time students) were plain Software Engineering and Data Analytics. Since the past three years we have studied Software Engineering through various modules, I found a bit redundant to choose that option, so I head to Data Analytics instead.

I remember my thoughts back in September 2020 and my idea of a fourth year was to gain some knowledge in Data Analytics itself thought some module specific lectures to produce a Final Project over 2021, sadly we had to deliver our progress by December and that was inexplicably worth 25% off the marks off the whole project.

I can say with conviction that my knowledge about data analytics and forecasting was barely minimum back in December and it was reflected on a mediocre delivery of it. At this stage, after completing the TABA and both Advanced Business Data Analysis and Data Mining I can say that I am in a completely different position that back then. I have been exposed to Linear Regression, Multilinear Regression, Principal Component Analysis, ANOVA, Decision Trees, Random Forest amongst other Data Mining and Business Analysis techniques.

I am wondering myself if the order of events were wrongly placed on the programme, as having a better knowledge in Data Analytics and Data Mining would have helped me to understand what Data Analytics is, and not only chose a better subject and plan a better Final Project but to use richer techniques. I would ask for the sake of the ones that run this Course in the future to consider an approach were a Student it given the first semester to get the knowledge across the specialisation and not being partially marked on something that it is not ready to develop until these first DA modules have finished. Who was first the chicken or the Egg? Thanks for reading.

May 2021

Software Project Journal: May 2021

Student: Alejandro Diaz Salgado

Programme: BSHC in Computing, year 4, Data Analytics

After a hectic end of the semester, having to deliver two big projects as TABAS for both projects I had, I was forced to defer the submission of the Final Year Project due to personal circumstances. I took the month of May to think about the biggest mistakes I have made throughout the transition of this journey, and how to surpass them to deliver the best possible product in the next submission in August.

First and most prominent mistake was procrastination. I am possibly the best procrastinator I know of, and when it comes to undertake a 60% module that requires hours and hours of research and documentation, it was not on my favour. It is something I really cannot help, as I apply that perspective to most parts of my life, where I am normally punctual and efficient, I associate my obsession with efficiency with procrastination; why spend X number of hours on something if I can work on other things and I still have X+Y number of hours left to use. Well, that perspective worked great for me over the past four years, as I never had any noise or interference on the equation, sadly, over this course I had several sources of noise that proved that procrastinating things is stupid and not a valid argument.

My second huge mistake was related to assumptions. I assumed that the crew behind RENFE's data management was going to update their data sources publicly. Wrong.

Then I assumed that I was going to be able to create some program to access that data over RENFE's API. Wrong again.

Lastly, I assumed that even RENFE presented only one day of data, they would update it several times a month. Of course, I was leveraging everything on another assumption and when I wanted to manually pull data, they never updated it. We are in mid-2021 and the last update was in mid-2020. I overlook the fact that half of the world was affected by COVID and a Public data repository aimed for research would have zero priority over other million things to update.

For all these reasons above, I doomed myself and never aimed for success since the beginning of the project.

Luckily, thanks to the mentors we were assigned, I was able to get my stuff together again and steer towards other directions. I was given the option of applying for an extension for the project and luckily, I was granted for it, so I have some golden months to work again on this and try not to lower down my results of Y4.

After a chat with the programme coordinator, I proposed the option of dumping completely the idea of the project and work towards a different analysis, but wisely, the mentor suggested for me to continue with that concept, even if I presented a Failed project. Undertaking a completely new project was not going to be feasible and needed of many

hours, also all the work behind was going to be worthless; based on that, I decided to give another chance to the passenger utilization of trains and try to find data in other sources and other countries.

June 2021

Software Project Journal: June 2021

Student: Alejandro Diaz Salgado

Programme: BSHC in Computing, year 4, Data Analytics

Working on a project over the summer whilst working and living in Spain is hard. Fact number 1.

Since I am literally unable to procrastinate more on this project, I was hands on and started to figure the best way out of the labyrinth.

To start, I thought it was a brilliant idea to have a catch up with my mentor, Giovanni. We set up a team meeting and reviewed the status of my project. Inevitably he flagged the status of my project as being poor, I did not have a reliable data source, something critical to develop a data analysis project. To solve that problem, I proposed the idea of manipulating the current data I had adding several passengers and subtracting them from the figures in a random way. To do that, I researched for already made functions on R and Excel and found something on the spreadsheet's program. The function is called `RANDBETWEEN` and it allows to add or subtract randomly from a given column based on a given range of numbers and present the results in an absolute manner to avoid negative values. Giovanni, did not like the idea of adding randomness to the dataset, because that would completely defeat the purpose of the project, which is providing a forecast. Forecast randomness is something not valid for a project, so I decided to terminate that idea based on Giovanni's advice.

The next approach was to find for other data we could use and related to trains. Months ago, I was able to find some datasets over the Argentinian database of railway utilization, something that I did not expect to be useful, but it was exactly what I was looking for due to the structure it has. The dataset has a structure showing the station, the train line (which I learned that in Argentina each train has a specific name), the month, year and day, grouped in just one value, and the number of passengers using that station over a month's period. Based on that data we can generate a time series forecast, which is going to be the direction I am going to take from now until the end of the project.

July 2021

Software Project Journal: July 2021

Student: Alejandro Diaz Salgado

Programme: BSHC in Computing, year 4, Data Analytics

It is 30th of July and I cannot believe I am completing a paper ahead of time.

After a difficult journey completing the Final Project, I have reached an end, and I would never thought back in May about me saying these words, but I have to admit I enjoyed writing the thesis and I learnt a handful of techniques to apply on my recently promoted job as a Manager on an known Data storage company.

On this last month of college, I have learnt a lesson that I was not able to learn in my past 34 years of life: spending one hour a day in a huge project allows you to finish it on time; just one hour.

Aside of my two minutes of reflections, this on this past month I came across very interesting technologies super useful for my day to day and to make events more predictable. Forecasting timeseries is something powerful. It can be applied so something that meaningful as the stock markets. After completing this project, I can say with confidence that I can have a slight better chance of winning money in stock or blockchain than before, as now I can "at least" run a forecast of how the stock will behave (within significant margins of error based on market trends and a million different external events, of course). That in my view is kind of a superpower only available for people that studied it or got interested on the matter.

In terms of technical knowledge gathered over the past months, R is with any sight of doubt, the most interesting tool I have come across on the degree. I could describe it as the programming language for programmers that use programming to be more productive an efficient but are people not creative and not really interested in programming, but other topics that need of it. I ran my whole thesis using R and sometimes using excel, and I found it really useful on a daily basis and a great tool to basically any business that need to analyse, filter and visualize data in a very easy way and without complications and compatibilities like other tools have (Tableau or SPSS in example). On the analytics side applied on the past month it was funny to see how little progress has been made on the past years in terms of forecasting. Arima models have many years and compared to Prophet, which was developed by a super corporation not long ago, has performed better in some situations. The advantage of modern models against old models resides in the adaptability.

Wrapping up the Final Project, I hope you enjoyed reading it, and I hope you I was not too cryptic when talking about some topics.

Thanks for reading

Alex



## References

- Prieto, G., 2021. *El ascenso y ocaso de la red ferroviaria argentina - Geografía Infinita*. [online] Geografía Infinita. Available at: <<https://www.geografiainfinita.com/2020/03/el-ascenso-y-ocaso-de-la-red-ferroviaria-argentina/>> [Accessed 1 August 2021].
- Wild, C., 2021. *Differences between timeseries*. [online] Stat.auckland.ac.nz. Available at: <<https://www.stat.auckland.ac.nz/~wild/wildaboutstatistics/>> [Accessed 1 August 2021].
- Youtube.com. 2021. *Holt Winters Forecasting Model in R* [online] Available at: <<https://www.youtube.com/watch?v=CB5HPpUzB0o>> [Accessed 1 August 2021].
- Youtube.com. 2021. *AFC and PACF Explained*. [online] Available at: <[https://www.youtube.com/watch?v=IcI9\\_46\\_RZY](https://www.youtube.com/watch?v=IcI9_46_RZY)> [Accessed 1 August 2021].
- E. E. Holmes, a., 2021. *4.4 Correlation within and among time series / Applied Time Series Analysis for Fisheries and Environmental Sciences*. [online] Nwfsc-timeseries.github.io. Available at: <<https://nwfsc-timeseries.github.io/atsa-labs/sec-tslab-correlation-within-and-among-time-series.htm>> [Accessed 1 August 2021].
- Docs.oracle.com. 2021. *Identifying Seasonality with Autocorrelations*. [online] Available at: <[https://docs.oracle.com/cd/E57185\\_01/CBPUG/PRHistData\\_Autocorr.htm](https://docs.oracle.com/cd/E57185_01/CBPUG/PRHistData_Autocorr.htm)> [Accessed 1 August 2021].
- Brownlee, J., 2021. *How to Identify and Remove Seasonality from Time Series Data with Python*. [online] Machine Learning Mastery. Available at: <<https://machinelearningmastery.com/time-series-seasonality-with-python/>> [Accessed 1 August 2021].
- Brownlee, J., 2021. *A Gentle Introduction to SARIMA for Time Series Forecasting in Python*. [online] Machine Learning Mastery. Available at: <<https://machinelearningmastery.com/sarima-for-time-series-forecasting-in-python/>> [Accessed 1 August 2021].
- Prophet. 2021. *Trend Changepts*. [online] Available at: <[https://facebook.github.io/prophet/docs/trend\\_changepts.html#automatic-change-point-detection-in-prophet](https://facebook.github.io/prophet/docs/trend_changepts.html#automatic-change-point-detection-in-prophet)> [Accessed 1 August 2021].
- Cran.r-project.org. 2021. *AutoTS*. [online] Available at: <<https://cran.r-project.org/web/packages/autoTS/autoTS.pdf>> [Accessed 1 August 2021].

