Big Data in Formula 1 and Statistical Findings from Past Records

By Paul Sweeney

Student Number: x12343601

Declaration Cover Sheet for Project Submission

SECTION 1 *Student to complete*

Name:
Student ID:
Supervisor:

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:_____ Date:_____

NB. If it is suspected that your assignment contains the work of others falsely represented as your own, it will be referred to the College's Disciplinary Committee. Should the Committee be satisfied that plagiarism has occurred this is likely to lead to your failing the module and possibly to your being suspended or expelled from college.

Complete the sections above and attach it to the front of one of the copies of your assignment,

What constitutes plagiarism or cheating?

The following is extracted from the college's formal statement on plagiarism as quoted in the Student Handbooks. References to "assignments" should be taken to include any piece of work submitted for assessment.

Paraphrasing refers to taking the ideas, words or work of another, putting it into your own words and crediting the source. This is acceptable academic practice provided you ensure that credit is given to the author. Plagiarism refers to copying the ideas and work of another and misrepresenting it as your own. This is completely unacceptable and is prohibited in all academic institutions. It is a serious offence and may result in a fail grade and/or disciplinary action. All sources that you use in your writing must be acknowledged and included in the reference or bibliography section. If a particular piece of writing proves difficult to paraphrase, or you want to include it in its original form, it must be enclosed in quotation marks
and credit given to the author.

When referring to the work of another author within the text of your project you must give the author's surname and the date the work was published. Full details for each source must then be given in the bibliography at the end of the project

Penalties for Plagiarism
If it is suspected that your assignment contains the work of others falsely represented as your own, it will be referred to the college's Disciplinary Committee. Where the Disciplinary Committee makes a finding that there has been plagiarism, the Disciplinary Committee may recommend

- that a student's marks shall be reduced
- that the student be deemed not to have passed the assignment
- that other forms of assessment undertaken in that academic year by the same student be declared void
- that other examinations sat by the same student at the same sitting be declared void

Further penalties are also possible including

- suspending a student college for a specified time,
- expelling a student from college,
- prohibiting a student from sitting any examination or assessment.,
- the imposition of a fine and
- the requirement that a student to attend additional or other lectures or courses or undertake additional academic work.

Final Report Contents:

Qualifying Lap Times and Race Lap Times in Monaco ('86-'15)

For this section I have decided to look at the fastest qualifying lap times and the fastest race times between 1986 and 2015 to see if there is a statistical difference between them. The reason why I have chosen these years is because, prior to 1986, the circuit and a different layout and would not be relevant for this test. The null hypothesis for this test is that there is no difference in lap times between qualifying and the race and the alternative hypothesis is that there is a statistical difference in lap times. In order to find these results, we first have to conduct a test for normality to see if the data is normally distributed or not. There is also another null and alternate hypothesis in this section. The null hypothesis states that the data for both sets are normally distributed and the alternate hypothesis states that the data for both sets are not normally distributed. The alpha value that I will be using for both of these tests is .05. This means that there is only a 5% chance of a Type 1 error being committed. A type 1 error occurs when the null hypothesis is being reject when it should actually be accepted.

Here is the result of the test for normality conducted in SPSS:

Tests of Normality

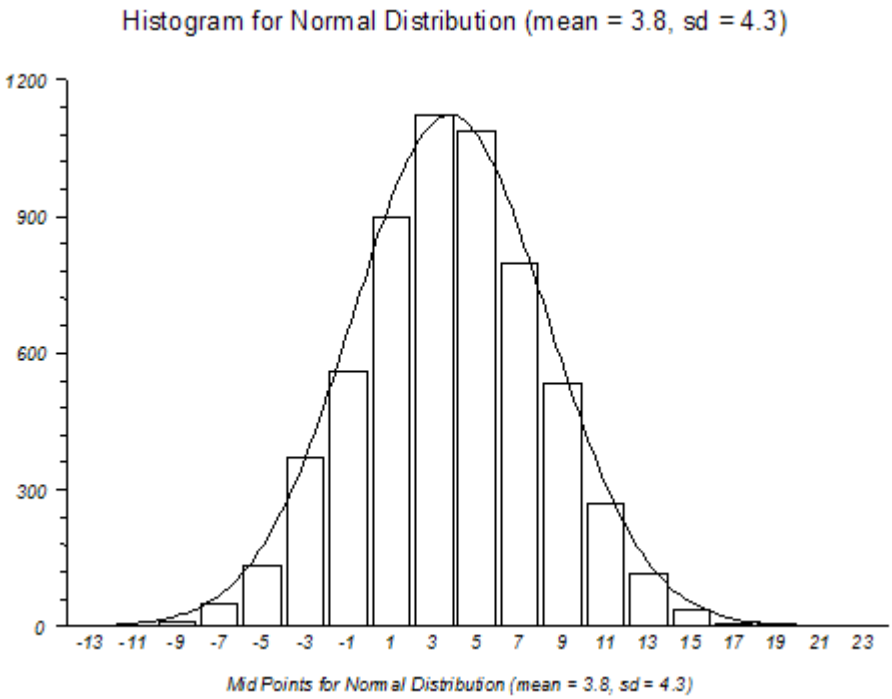
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Fastest_Qualifying_Lap_Times	.147	29	.108	.926	29	.043
Fastest_Race_Lap_Times	.123	29	.200*	.922	29	.035

*. This is a lower bound of the true significance.

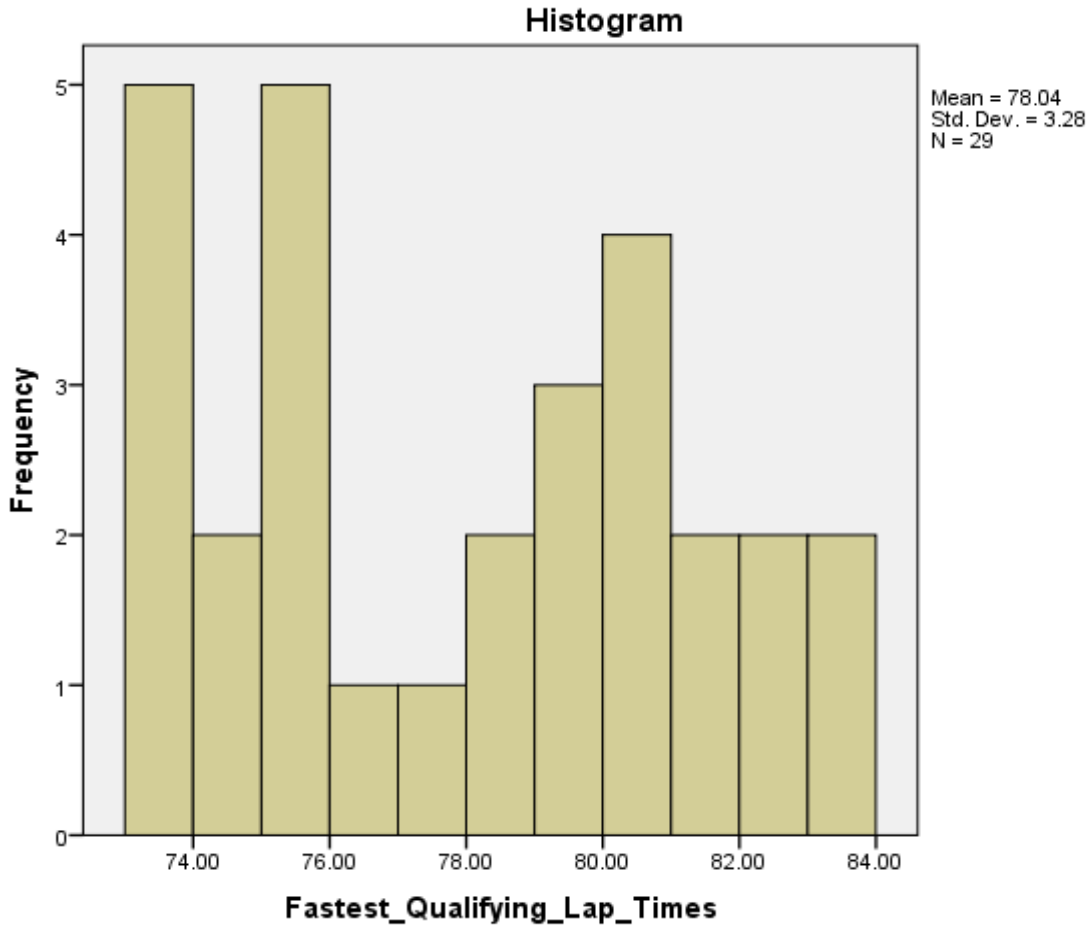
a. Lilliefors Significance Correction

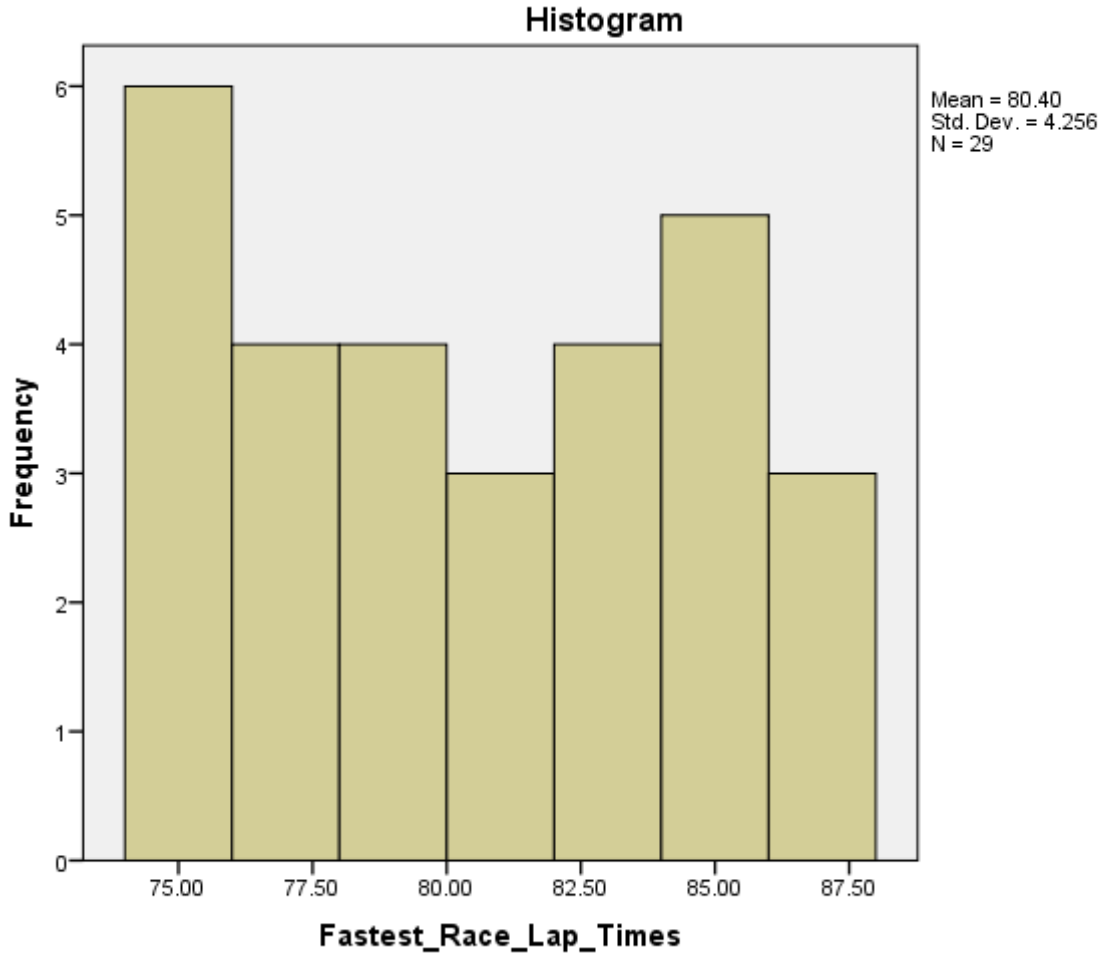
As you can see, the significance value for the Kolomgorov-Smirnov test is above .05 which tells us that we should accept the null hypothesis. But for the Shapiro-Wilk test, it is says that the null hypothesis should be rejected. Which one should be chosen?

The Shapiro-Wilk test is generally considered as the better test but it might be wrong to just assume that it is correct instead of the Kolomgorov-Smirnov test. There is a visual comparison that can be used: Histogram.



This is a picture of what a normally distributed histogram looks like. As you can see it has a bell-curve to it which shows normal distribution.





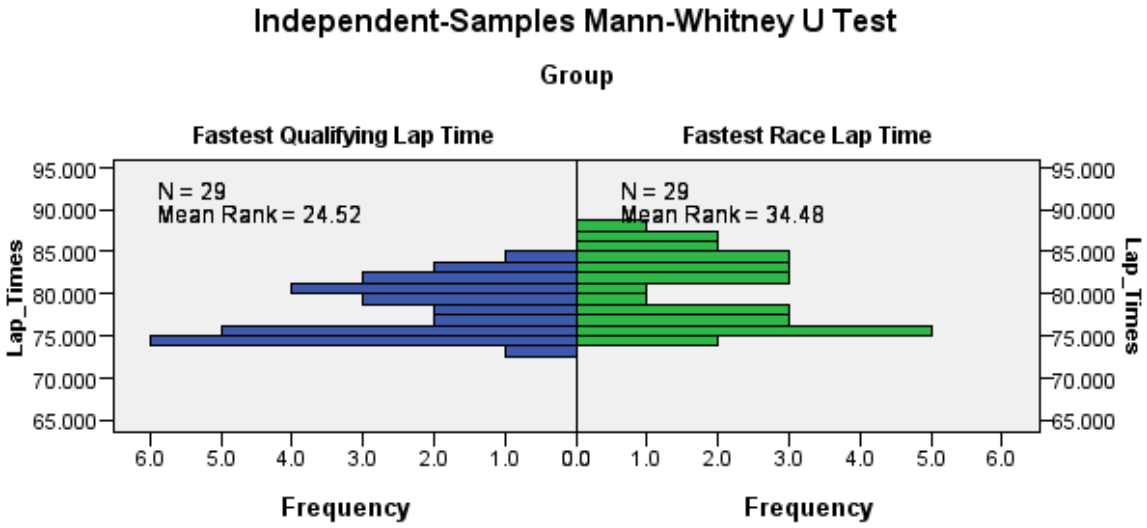
These two histograms for the Fastest Qualifying Times and Fastest Race Times show that they do not fit the bell-curve, which shows that the data in both cases are not normally distributed. Therefore, the Shapiro-Wilk test is correct and now we can reject the null hypothesis in favour of the alternate hypothesis as evidence has been found that the data is not normally distributed.

Now it is possible to check to see if both groups are different. Since the data is not normally distributed, a non-parametric test will be conducted. In this case, I will be doing a Mann-Whitney U test on these datasets. The results from the Mann-Whitney U test are as follows.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Lap_Times is the same across categories of Group.	Independent-Samples Mann-Whitney U Test	.025	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.



Total N	58
Mann-Whitney U	565.000
Wilcoxon W	1,000.000
Test Statistic	565.000
Standard Error	64.303
Standardized Test Statistic	2.247
Asymptotic Sig. (2-sided test)	.025

The results of this test show that the null hypothesis is to be reject in favour of the alternate hypothesis on the basis that the significance value is below our alpha value of .05. The graph also shows us that the lap times in qualifying at the Monaco GP between 1986 and 2015 were faster than the lap times during the race.

Record of Laps Led In a Season between Ferrari, McLaren and Williams

I decided to gather data based around the record numbers of laps led by the three of the oldest teams in the sport Ferrari, McLaren, Williams. I have taken 8 of each of the teams record setting years and decided to create some tests based around this data. The null hypothesis for this test is that there is no difference between each of the teams whereas the alternate hypothesis states that there is a difference between at least two of the groups.

Before that can be calculated, a test of normality will have to be done to see if each of the groups are distributed normally. The null hypothesis for this test states that the groups are normally distributed and the alternate hypothesis states that the groups are not normally distributed. The alpha value for both of these tests is set at .05.

Here is the result from the test for normality:

Tests of Normality

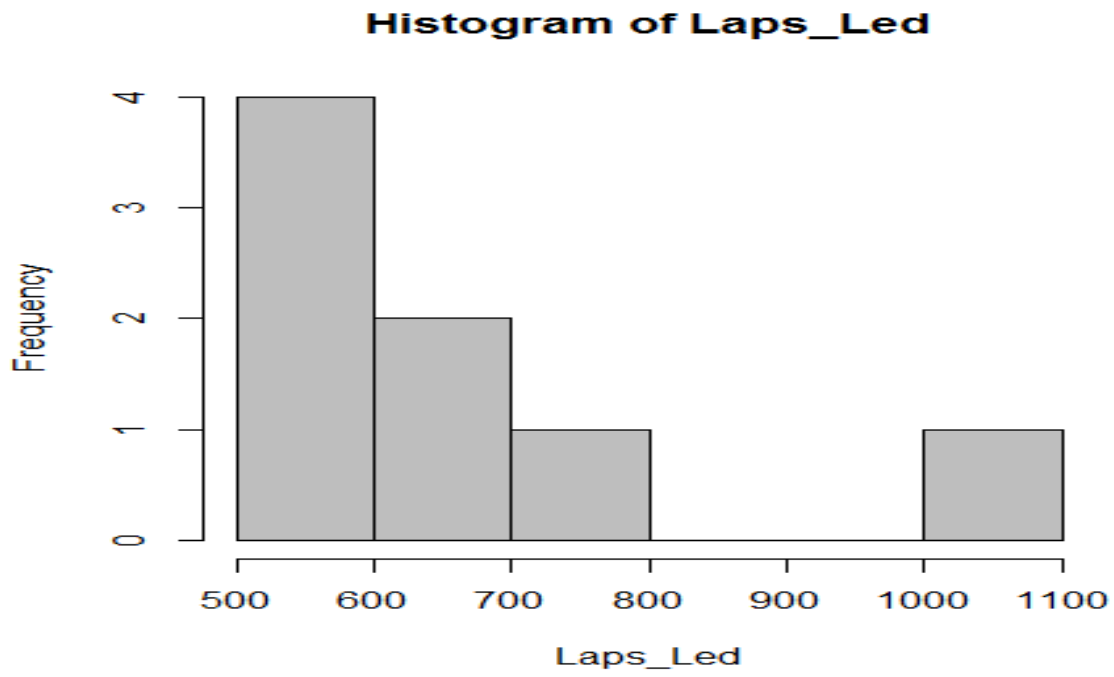
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
McLaren	.196	8	.200*	.837	8	.071
Williams	.199	8	.200*	.937	8	.579
Ferrari	.253	8	.140	.857	8	.113

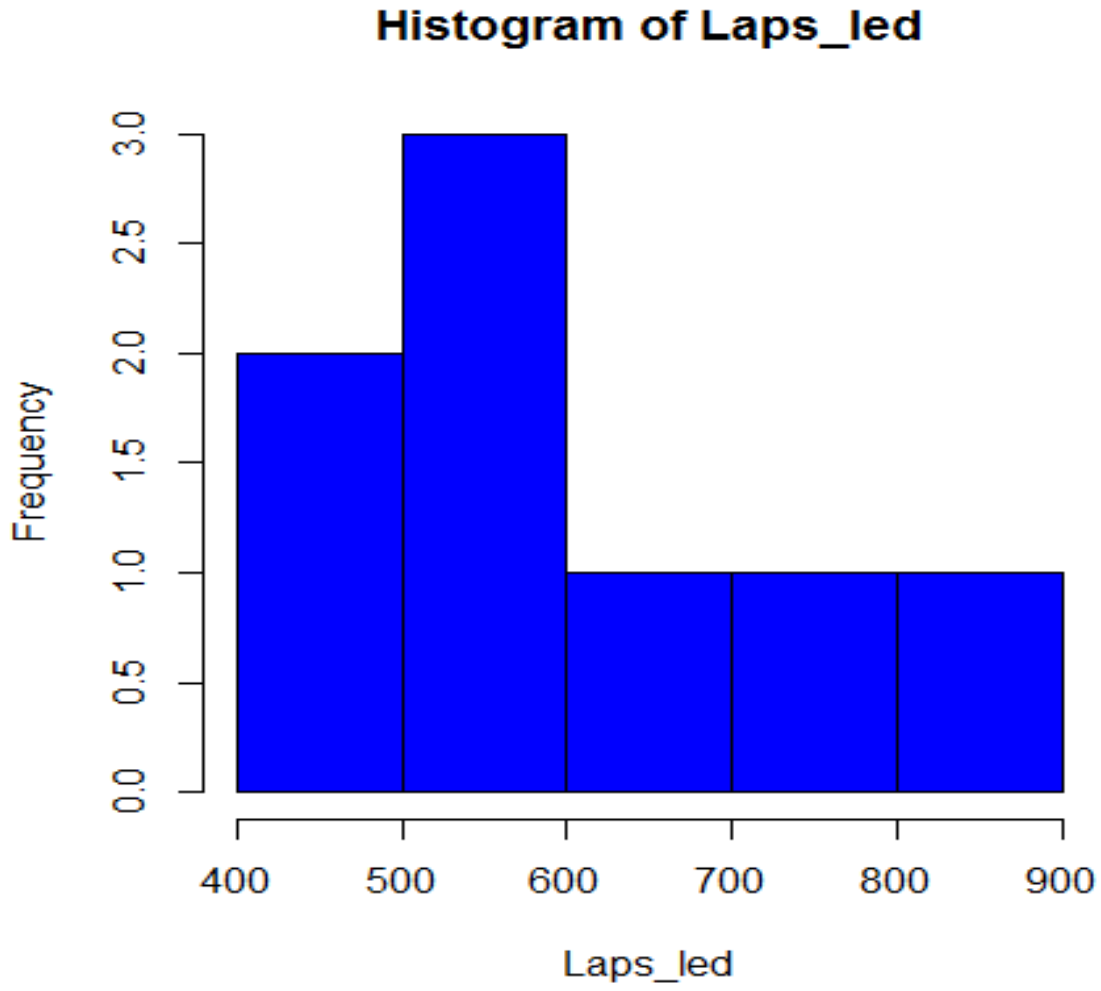
*. This is a lower bound of the true significance.

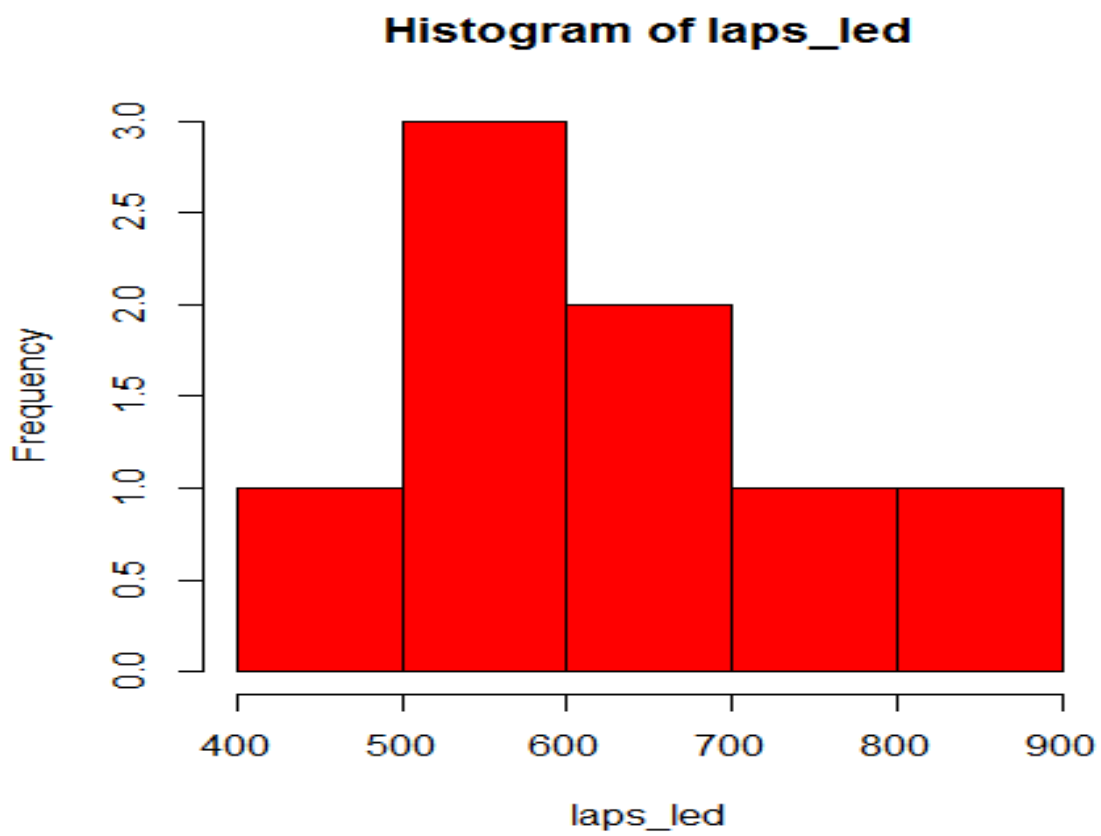
a. Lilliefors Significance Correction

In this test I am going to use the statistics from the Shapiro-Wilk test. The Shapiro-Wilk test shows that the data for McLaren, Ferrari and Williams is normally distributed as they all have a significance value above .05. This means that we have failed to reject the null hypothesis.

Here are the histograms for each team that have been produced in R:


Histogram of Laps_Led
Frequency
0
1
2
3
4
500
600
700
800
900
1000
1100
Laps_Led


Histogram of Laps_led
Frequency
0.0
0.5
1.0
1.5
2.0
2.5
3.0
400
500
600
700
800
900
Laps_led



And the R code to produce these histograms.

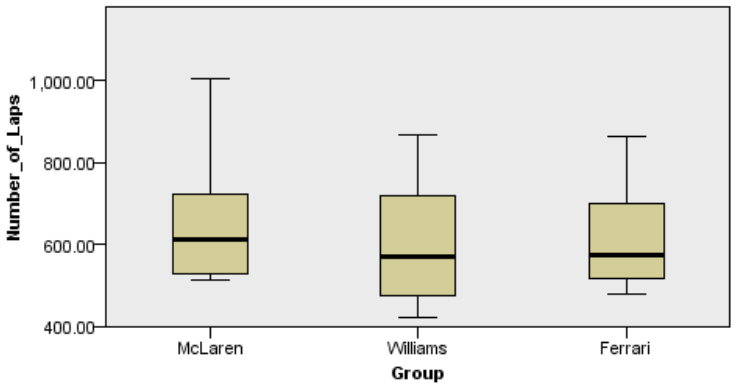
```
LapsInaYear <- read.csv(file= "Laps-Led-In-a-Year-
Constructors.csv", head = TRUE, sep = ',')

LapsInaYear


# Display Groups Individually

LapsInaYear[1:8, 1]

LapsInaYear[1:8, 2]

LapsInaYear[1:8, 3]


# Create Histograms

Laps_Led=LapsInaYear[1:8, 1]

hist(Laps_Led)

hist(Laps_Led, col = 'grey')


Laps_led=LapsInaYear[1:8, 2]
```

```
hist(Laps_led)
hist(Laps_led, col = 'blue')


laps_led=LapsInaYear[1:8, 3]
hist(laps_led)
hist(laps_led, col = 'red')
```

The graphs that are displayed seem to show signs of non-normal distribution but the statistics say that the data is normally distributed. This is one of the problems that can occur with having a small sample file size.

I have decided to conduct a Kruskal-Wallis test for this data due to its small sample size. The results for the Kruskal-Wallis test are as follows.

Independent-Samples Kruskal-Wallis Test



Total N	24
Test Statistic	.455
Degrees of Freedom	2
Asymptotic Sig. (2-sided test)	.797

1. The test statistic is adjusted for ties.
2. Multiple comparisons are not performed because the overall test does not show significant differences across samples.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Number_of_Laps is the same across categories of Group.	Independent-Samples Kruskal-Wallis Test	.797	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

As the test shows, we have failed to reject the null hypothesis as the significance value is above the alpha value of .05. I have also included a box-plot diagram as part of this test. The box-plot diagram displays the distribution of the data. The lowest line represents the minimum value, the bottom edge represents the first quartile, the line just above represents the median number, the top of the box represents the third quartile and finally the line at the top represents the maximum number.

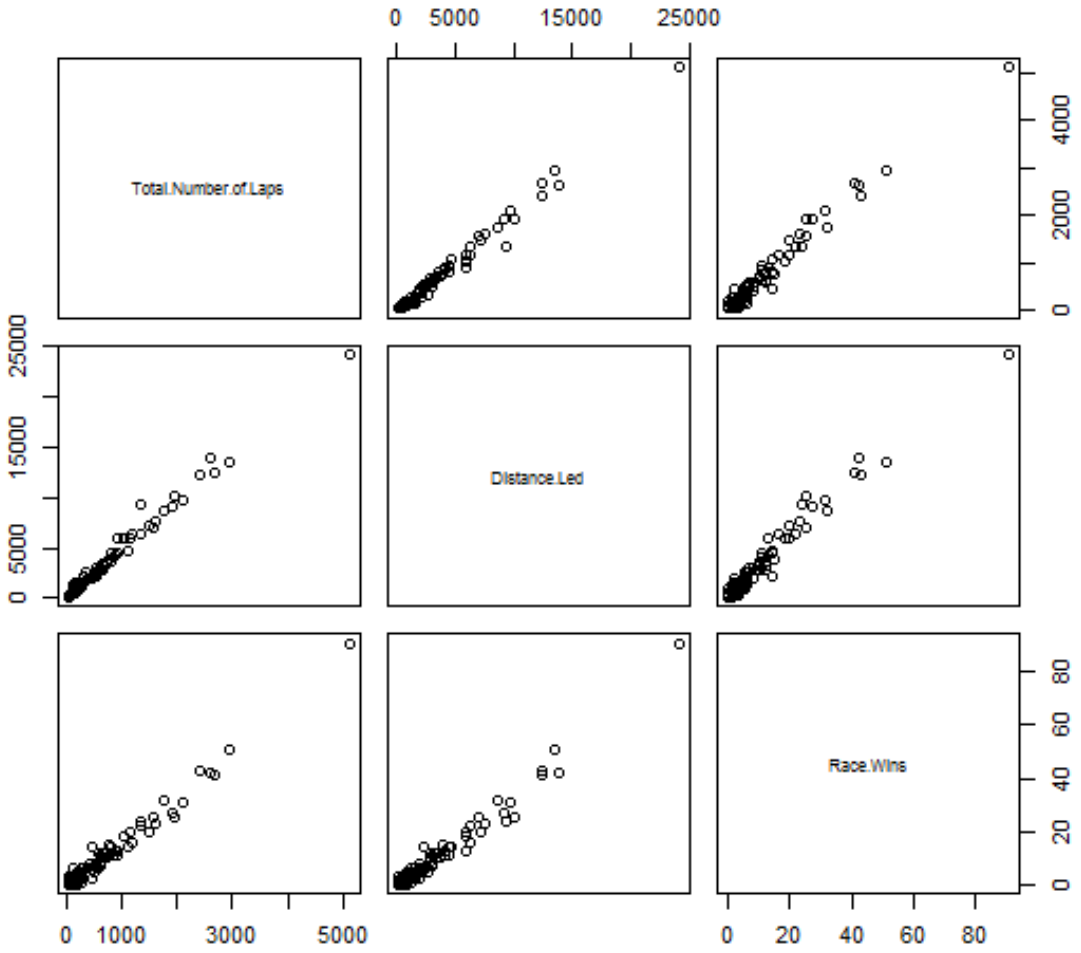
Correlation between Total Number of Laps, Distance Led and Race Wins for Drivers (MLR Test in R)

For this section, I have decided to see if there is any correlation between the total number of laps some drivers have led, the distance that they led for and the race wins that have come out of it. I have taken a sample of 100 drivers with the most laps led and filled out the rest of the data accordingly. Here is the code that I used to read the data file into R and conduct this experiment.

```
setwd(file.path ( "H:", "Project"))
Driversinfo <- read.csv(file= "Laps-led-Distance-and-Race-
Wins-(drivers).csv", head = TRUE, sep=',')

# Display info
Driversinfo

# Display names
names(Driversinfo)
# Run Correlation
cor(Driversinfo)
# Show Graphically
pairs(Driversinfo)

# Run Model
Model <- lm( Total.Number.of.Laps ~ Distance.Led + Race.Wins,
data=Driversinfo)
Model
```

Also, here are the results that have been generated and the graph to give a visualisation.

```
                     Total.Number.of.Laps Distance.Led Race.Wins
Total.Number.of.Laps            1.0000000    0.9930557 0.9867319
Distance.Led                    0.9930557    1.0000000 0.9825063
Race.Wins                       0.9867319    0.9825063 1.0000000
```



As the graph and the statistics show, there is a very strong correlation between the Total Number of Laps led, Distance Led and Race Wins. This test shows that the more laps you lead, the more distance you will led and increase the chances of winning.

Constructors Scores 2007

Since McLaren were disqualified from the Constructors Championship in 2007, I decided to gather the data on what their score would have been and compare it to the likes of Ferrari to see if there would have been a difference between the two scores. The null hypothesis states that there would have been no difference in the score and the alternative hypothesis states that there would have been a difference between the scores. The alpha value is set at .05 for this test and I will be using RStudio to conduct my findings.

First, I created some basic statistics for each of the groups. The results for each group are as follows:

```
> mean(ConstructorScores$Ferrari)
[1] 12
> median(ConstructorScores$Ferrari)
[1] 11
> max(ConstructorScores$Ferrari)
[1] 18
> min(ConstructorScores$Ferrari)
[1] 4
> range(ConstructorScores$Ferrari)
[1]  4 18
> diff(ConstructorScores$Ferrari)
 [1]  -3   6  -6  -3  -3   7   7  -4
 [9]  -6   0  10 -12  12  -9   7   2
> var(ConstructorScores$Ferrari)
[1] 22.25
> sd(ConstructorScores$Ferrari)
[1] 4.716991
> kurtosis(ConstructorScores$Ferrari)
[1] 1.657398
> skewness(ConstructorScores$Ferrari)
[1] -0.01473197
> summary(ConstructorScores$Ferrari)
   Min. 1st Qu.  Median    Mean
      4       8      11      12
3rd Qu.    Max.
     16      18


> mean(ConstructorScores$McLaren)
[1] 12.82353
> median(ConstructorScores$McLaren)
[1] 12
> max(ConstructorScores$McLaren)
[1] 18
> min(ConstructorScores$McLaren)
[1] 8
> range(ConstructorScores$McLaren)
[1]  8 18
> diff(ConstructorScores$McLaren)
 [1]   4  -6   2   4  -6   6 -10   6  -4   5  -5   8  -7  -1  -2   0
> var(ConstructorScores$McLaren)
[1] 13.40441
> sd(ConstructorScores$McLaren)
[1] 3.661204
> kurtosis(ConstructorScores$McLaren)
[1] 1.746983
> skewness(ConstructorScores$McLaren)
[1] 0.2082459
> summary(ConstructorScores$McLaren)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   8.00   10.00   12.00   12.82   15.00   18.00
```

These basic descripitves show some very basic calculations being done for example the mean, median, mode etc. However some basic descriptives such as Kurtosis are not as easy to calculate and might be slightly hard to understand at first. Kurtosis is a statistical measure that can be used to describe how the observed data has been distributed. Var (or Variance) can be described as a

measurement of how the numbers are spread out in a dataset. SD (or Standard Deviation) shows the dispersion of a set of data from the mean. Skewness is a way of describing the asymmetry from the normal distribution in a dataset. The data between McLaren and Ferrari look very similar but a further test still needs to be conducted. I will conduct a t-test (Unpaired) to see if there is a difference between the two scores.

Here are the results:

```
data:  ConstructorScores$Ferrari and ConstructorScores$McLaren
t = -0.5687, df = 30.145,
p-value = 0.5738
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.780578  2.133519
sample estimates:
mean of x mean of y
 12.00000  12.82353
```

Here is the line of code that helped produced the result:

```
ConstructorScores <- read.csv(file = "2007 Constructor
Scores.csv", head = TRUE, sep = ',')

ConstructorScores

t.test(ConstructorScores$Ferrari, ConstructorScores$McLaren)
```

The program is telling me that a difference has been found in the score so in this case I will be rejecting the null hypothesis in favour of the alternative as there is enough evidence to prove that a difference has been found.

ANOVA Test on 2007 Constructor Scores

In this section, I have decided to do another test based on the Constructor Scores in 2007 but this time adding another team, BMW Sauber, to the dataset. I will be conducting an ANOVA test to see if there is a difference between these three teams and the groups that differ. The null hypothesis for this test is that there is no difference between any of the groups and the alternate hypothesis is that there is a difference between two or more groups. The alpha value for this test is .05.

Here are the results of the ANOVA test:

```
             Df Sum Sq Mean Sq
ind           2  488.9  244.47
Residuals    48  694.2   14.46
            F value   Pr(>F)
```

```
ind            16.9 2.77e-06 ***
Residuals
---
Signif. codes:
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’
  0.05 ‘.’ 0.1 ‘ ’ 1
```

The code that produced the result:

```
s_anova_test <- stack(anova_test)

model <- aov(values ~ ind, data = s_anova_test)

summary(model)
```

So a difference has been found in the data which means we have enough evidence to reject the null hypothesis in favour of the alternative hypothesis.

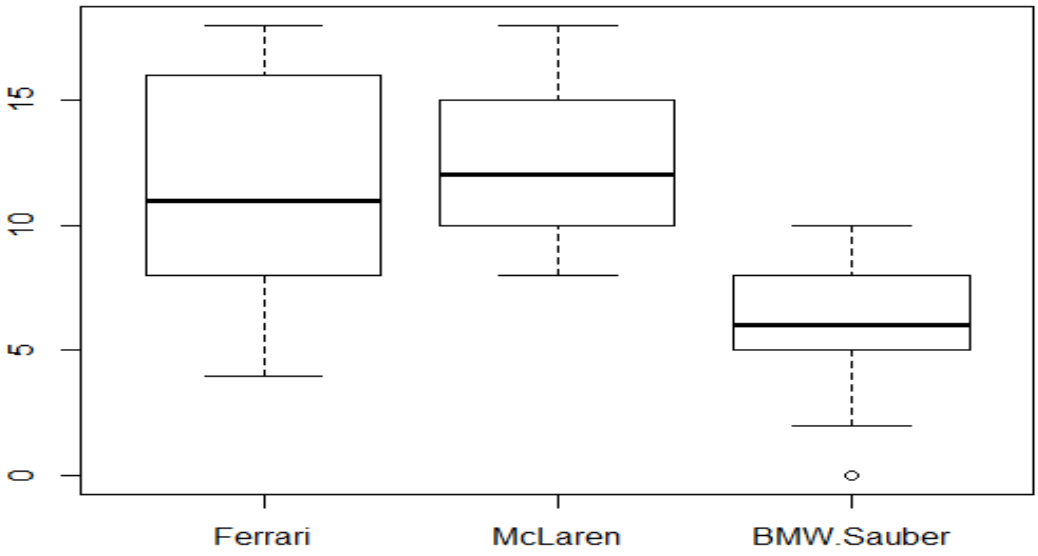
Since that a difference was found, I will now create a Tukey HSD model which will show me what groups differ and provide me with more information.

Result:

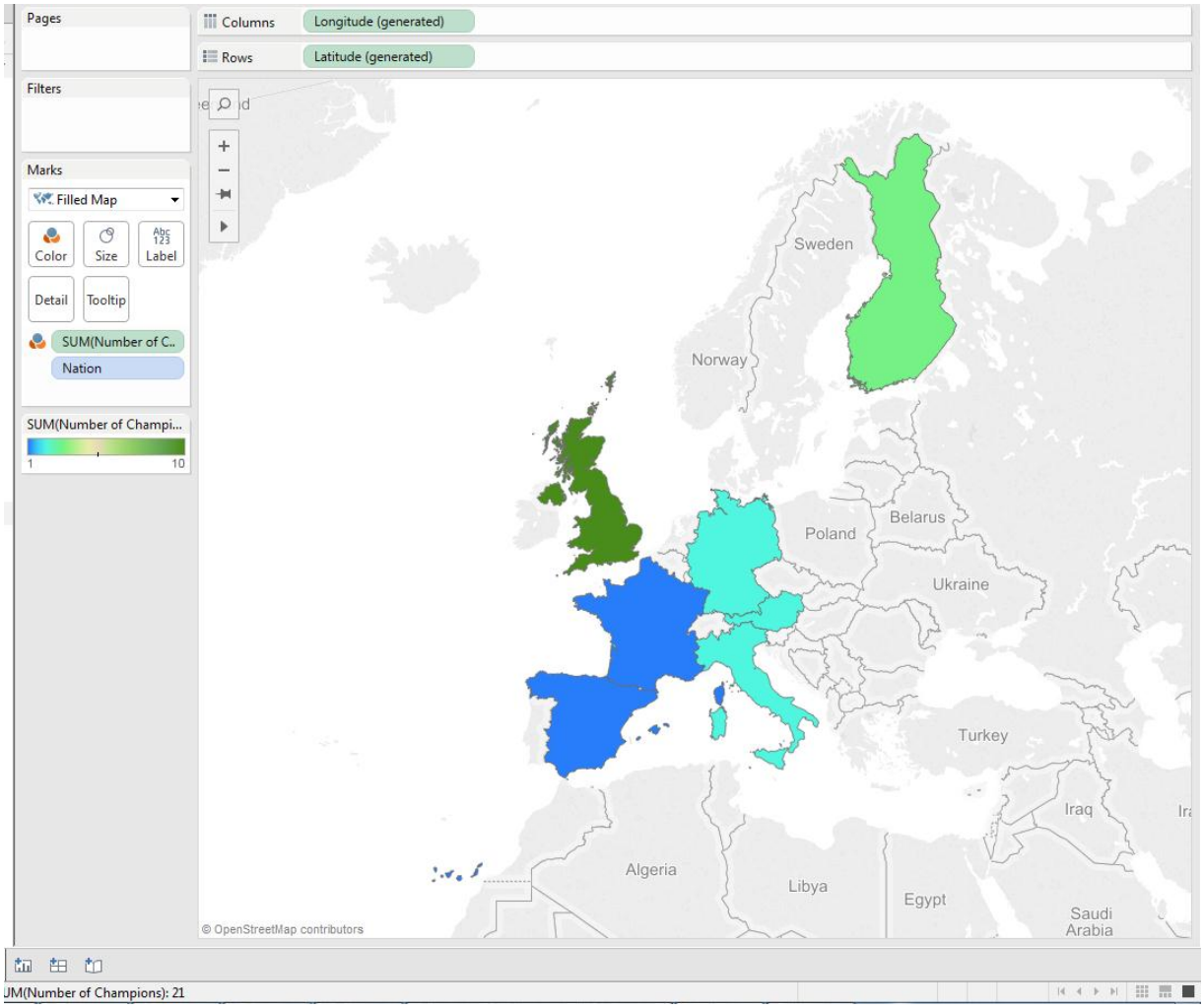
```
Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = values ~ ind, data = s_anova_test)

$ind
                         diff
Ferrari-BMW.Sauber 6.1176471
McLaren-BMW.Sauber 6.9411765
McLaren-Ferrari    0.8235294
                          lwr
Ferrari-BMW.Sauber  2.962881
McLaren-BMW.Sauber  3.786411
McLaren-Ferrari    -2.331236
                          upr
Ferrari-BMW.Sauber  9.272413
McLaren-BMW.Sauber 10.095942
McLaren-Ferrari     3.978295
                        p adj
Ferrari-BMW.Sauber 0.0000674
McLaren-BMW.Sauber 0.0000079
McLaren-Ferrari    0.8036765
```

This shows that a difference has been found in each of the three groups.

I also produced a boxplot for each of these three teams to visually show what the difference looks like:

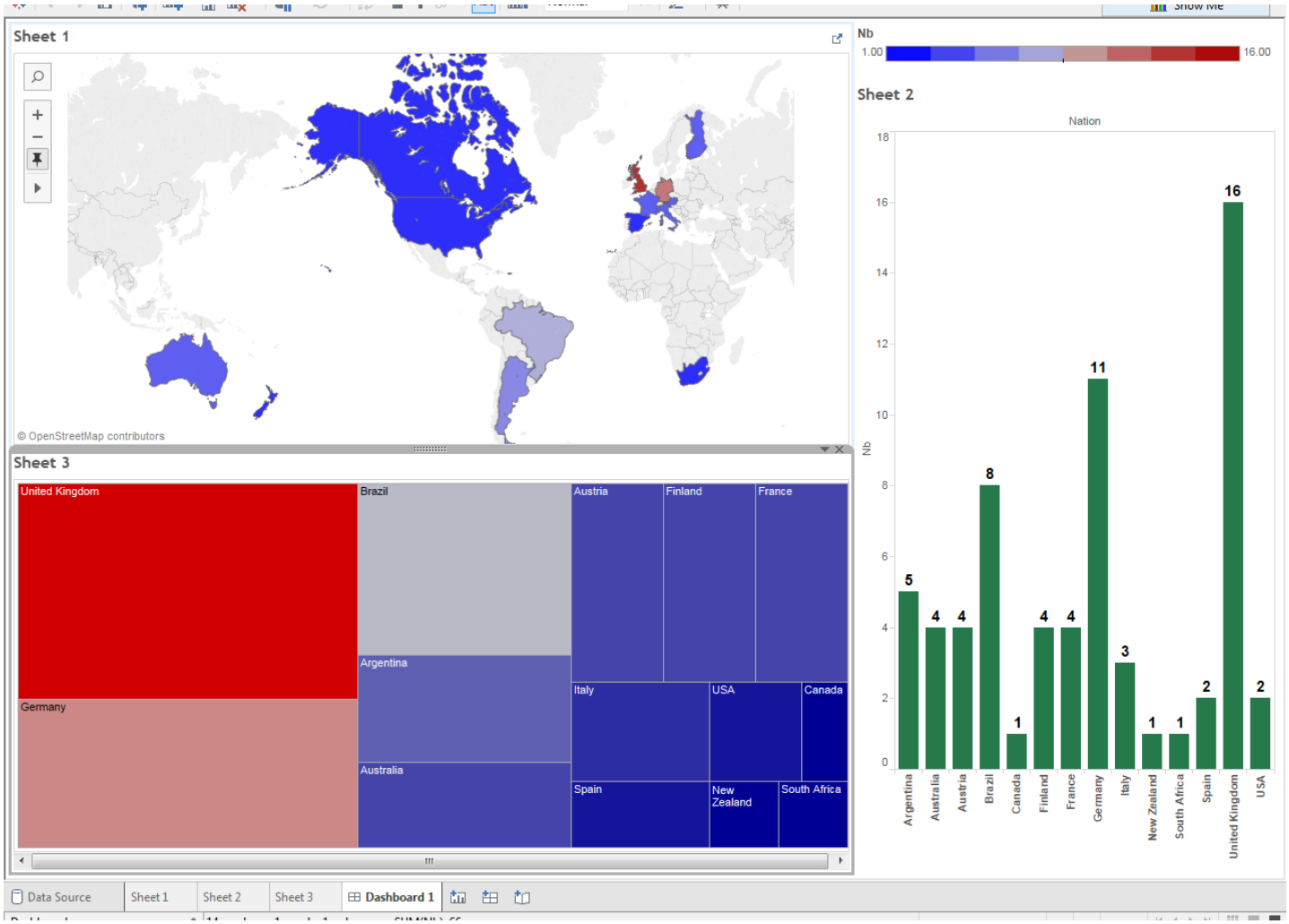


Champions per European Nation (Tableau)



This image from Tableau shows the number of European World Champion Drivers that there are currently. The blue colour indicates a low number of World Champions, dark green indicates a high number of World Champions and countries that have no colour have no World Champions currently.

Total Number of Drivers Championships per Nation



This image here shows a World Map, a bar chart and a tree map of the total number of Drivers Championships per Nation. As you can see there are a number of ways of visually showing the differences in numbers between each of the countries. Countries that are highlighted deep blue have a low number of Champions and countries that are red have a high number of champions.

BSHTM 4th Year Journal

Student Name: Paul Sweeney

Student Number: x12343601

2015/2016

Contents:

Introduction:

My name is Paul Sweeney. I am 22 years old and I am a student in the National College of Ireland studying Technology Management specializing in Data Analytics. For this project, we were asked to collect data on a certain area, create programmes that are able to turn that data into information and create a report on our findings. As part of this project, we were also asked to keep a monthly journal of our progress on our project throughout the year. This would help us keep track of the updates that we were making to our project and give some insight to the examiners as to how we worked on our project.

Month: September 2015

Unfortunately, due to circumstances beyond my control, I was unable to attend the first week of lectures. This meant that I already had less of an understanding about the project and what was required of me. I also questioned my own skills as a Data Analyst as, prior to this year, I had no understanding of what a Data Analyst does or what was required of me for my project. So I decided to ask my Project Manager (Eugene O'Loughlin) what the final year project was about. Eugene guided me very well by giving me suggestions on what I could base my project on, ideas on what I could do with my project and giving me information that I had missed out on in class. Eugene also informed me to base the project around something that I feel passionate about. So with that in mind I decided to base my project about searching for big data and information on Formula 1. With this project I intend to find trends and patterns with the data that I will have found and be able to make predictions about certain areas such as lap times around a certain track. After discussing my ideas for the project with Eugene, he thought that it would be a good idea to base my project around this area

Achievements for This Month:

Achievements for this month would be coming up with an idea for a project and submitting the Project Proposal. As it is just the beginning of the project, the Project Proposal will be changed and edited during the year but getting a start on it and submitting as much of it as possible for this month I would consider as a small achievement.

Reflections:

I have learned that time management is going to be vital for the duration of this project. There is a lot of work that is going to have to be put into this project and managing time is going to be key. I feel that it is going to be difficult since there are other subjects, projects and tests that I will have to focus on as well.

Month: October 2015

There has been a lot of confusion as to what has to be done for this project. Our class has had meetings with the people involved in this project and, in certain circumstances, left us slightly more confused as to what has to be done. This case was more severe for the students specialising in the Business Analysis rather than the Data Analytics, but by no means easier for the students doing Data Analytics. We had a class meeting with Eugene O'Loughin and Ron Elliot early in the month. We had a deep and in depth conversation about what has to be done for this project and by the end of the meeting, we all had more information about the project and the aims were made clearer.

Achievements for This Month:

Unfortunately, due to a lot of confusion about what has to be done for the project, I did not manage to get much work done on the project this month. Gathering more information about what is required for the project was really the only achievement I managed this month.

Reflections:

I felt that I had lost quite a bit of time in doing my project since there was some confusion about what has to be done. Not knowing fully about the project, I did not get as much work done on the project as I had hoped.

Changes for Next Month:

A lot more time is going to be spent putting work into this project. Now that I have more information about the project, I can begin working properly and putting my full potential towards this project.

Month: November 2015

At the start of the month, I had a meeting with Eugene. I was still confused as to what I was supposed to be focusing on at this stage of the year. We discussed some of the approaches that I could take towards this project. During the meeting I began to realise what I was supposed to be doing for my project. We discussed about how I should gather my data files, establish the source of where I got my data from and discussed the technologies that I could include such as SPSS, RStudio, Tableau etc. Eugene even suggested some of the comparisons that I could make and what I should use each of the technologies for. For example, Tableau should be used for displaying maps and SPSS can be used for certain statistical findings.

I found this meeting to be the most beneficial to me so far as it gave me a clearer sight as to what I had to do for this project. After that meeting I began to try and find appropriate data files that I could use. Unfortunately I found this very difficult. The reason being was because there was not many sites that had the kind of in-depth information that I was looking for and the sites that had some of the relative information I would have had to pay for, which I

was not willing to do. I thought I had found data files that I could download straight from one of the sites that I found but it was encrypted and I was not able to read them and searching for these files was taking up a lot of time.

Achievements for This Month:

Once again, I had not made as much progress as I had hoped. Even though a lot of the confusion had been cleared up, there were still certain areas that I was unsure about such as obtaining the correct data files. One thing that I can take from this month is that I now know the software that I can use for my project, once I have obtained the data files, and what the software can be used for.

Reflections:

I am still struggling to get a proper start on this project. Although the aims and objectives is becoming clearer, achieving them is a lot harder than I initially thought it would be.

Changes for Next Month:

The thoughts of changing my project have come across me a few times by this stage and might be a possibility for next month, but I feel that having more meetings with Eugene will help clarify some of the problems that I'm still having.

Month: December 2015

This month has proven to be the most successful month that I have had so far. After another meeting with Eugene this month I have realised that my approach to gathering data files was completely wrong. What I was doing initially was trying to download files directly from in the Internet. This turned out to be the wrong idea completely and what I should have done was extract some data from a site that I found (STATSF1), put them into Excel files and use those files for my project. Now that I knew the right approach to getting the data files that I needed, I then began searching for the relevant data that I required.

I found a site called STATSF1 which had a lot of data about Formula 1 that I felt I could put to good use. It had many areas in which I could obtain data from which I felt was relevant towards my project. Some of these areas include number of wins per driver, number of laps led by a Constructor etc. So what I decided to do was to gather the data that I was going to use and store it all in one Excel file. I did this because I felt that having all of my data stored in the one file was the right way to approach this project. December was also the first month that I had used Tableau. Tableau was shown to our class in one of our modules (Business Data Analysis) and we were shown some of the interesting features that Tableau has to offer. I will definitely learn how to use Tableau properly for my project as Eugene has suggested.

Achievements for This Month:

Fully understanding the project and having clear aims on what has to be done. Learning how to create a data file and finding a site that has the data that I will require.

Reflections:

I am disappointed that it has taken me this long to realise what is actually involved in the project and what is required. But now that I understand the project I can finally put some hard work and effort into it.

Changes for Next Month:

Beginning work on the project itself such as learning more information about RStudio and how to use it as part of our project. Begin on the preparation for the Mid-Point Presentations.

Month: January 2016

January proved to be less successful than December. I had to spend a lot of time concentrating on my modules as it was exam season. Therefore, I had less time to focus on my project which obviously did not help as I felt I was already behind at this stage. When the second semester began during the last week in January, we learned more about RStudio and how we could incorporate it into our project. We were taught how to read a file into RStudio and did some minor statistical findings using RStudio. I was thankful that we were shown this at an early stage as I could now use it as part of my Mid-Point Presentation. I also found out that storing all of my data in one Excel file was untidy. I cleansed the data as best as I could but I knew I was going to have to come up with a better solution quickly.

Achievements for This Month:

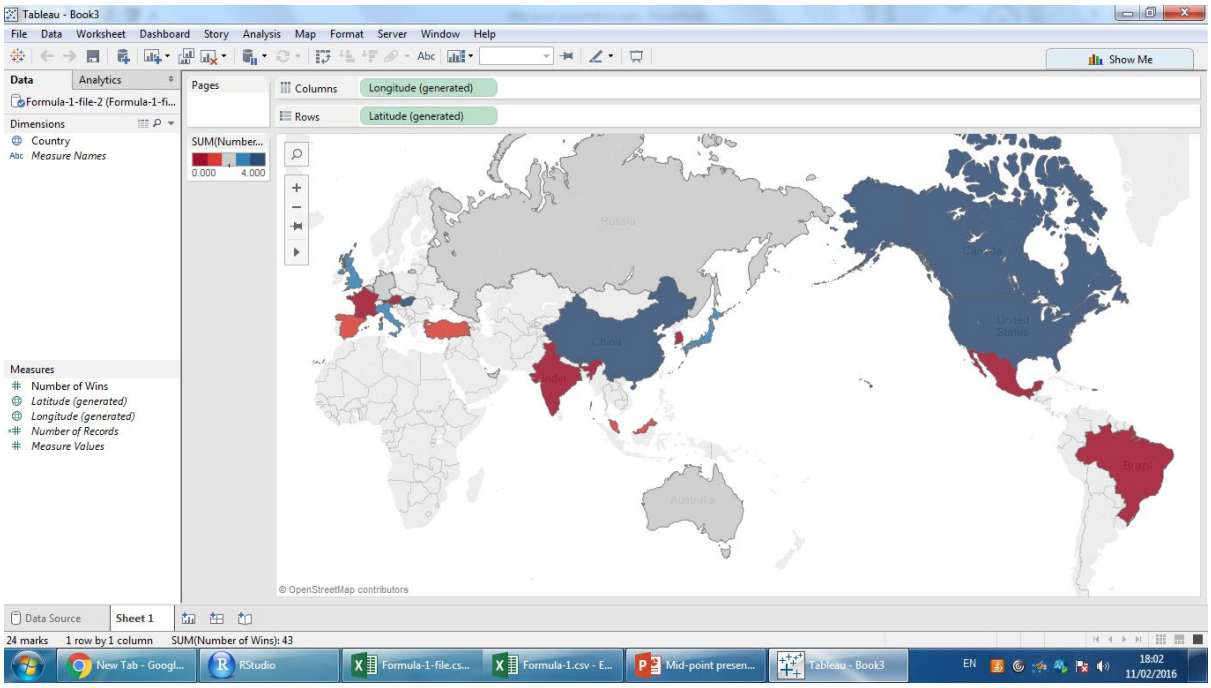
Finding out how to read files into RStudio and doing some statistical findings in R. Learning more about Tableau and how I could incorporate it into my project.

Reflections:

One positive that I can take from this month is that I began learning more about Tableau. I could see that Tableau would be a very good tool to use as part of my project and I began to do some work on it. I decided that, as a demonstration for my Mid-Point Presentation, I would show the countries that Lewis Hamilton has won in and the number of times he has won in those countries and have some basic descriptives ready in my R code.

Month: February 2016

At the beginning of this month, we had our Mid-Point Presentations. I managed to give some details about my project and show the examiners (Eugene and Ron) some of the progress I had made.



A world map showing the countries that Lewis Hamilton won races in. The number of races won is determined by the colour of the country itself. Tableau was used to display this world map.

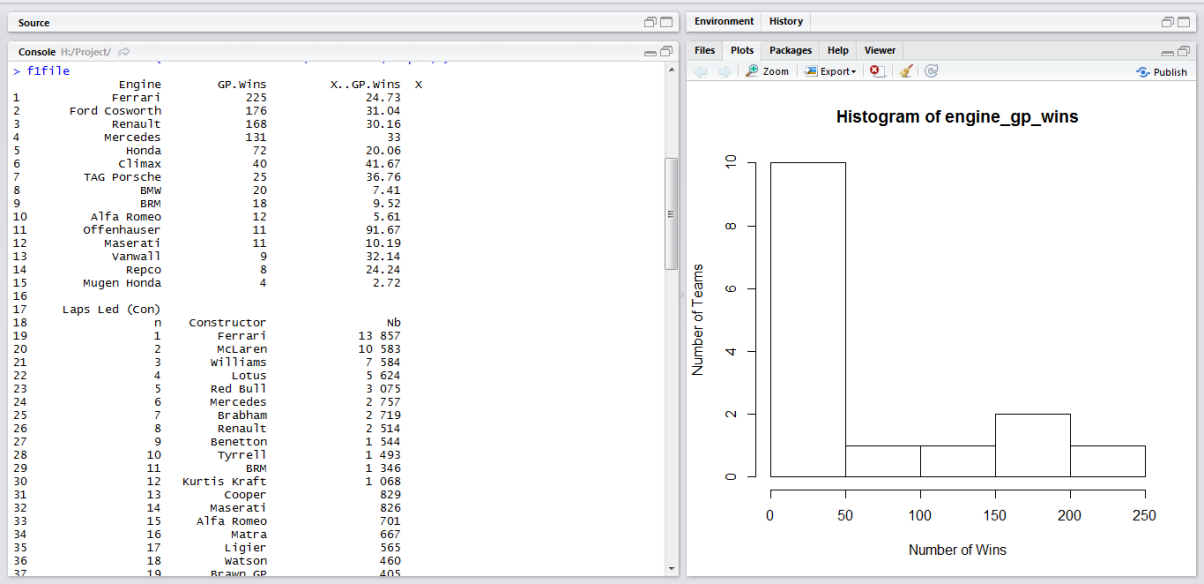
I felt that my learning of RStudio has progressed. Here is an example of the code that I had produced for this presentation:

```
#set working directory

setwd(file.path ("H:" , "Project"))

f1file <- read.csv(file="Formula-1.csv", head=TRUE, sep=",")

f1file


#read in data

engine_gp_wins <- c(225, 176, 168, 131, 72, 40, 25, 20, 18,
12, 11, 11, 9, 8, 4)
```

```
#Basic Descriptives

mean(engine_gp_wins)

max(engine_gp_wins)

min(engine_gp_wins)

range(engine_gp_wins)

var(engine_gp_wins)

sd(engine_gp_wins)

kurtosis(engine_gp_wins)

skewness(engine_gp_wins)

summary(engine_gp_wins)


# Histogram

hist(engine_gp_wins, xlab = "Number of Wins", ylab = "Number
of Teams")
```

This code here is able to read a data file into RStudio from its working directory and is able to display basic descriptives of the data and display a histogram.

Here is what some of the code displays when it is running:



This shows the data file that has been read into RStudio and a histogram of the number of teams that have won at least one grand prix and the number of teams that have achieved a certain number of wins. For example there are 10 teams in the history of Formula 1 who

have won more than one race but less than fifty and only one team who has won between 200 and 250 races.

I received some positive feedback but the feedback also acknowledged the fact that I was behind and had some catching up to do. I was happy that the feedback I had received was honest and took the advice that was given to me on board and will use it to help further improve my project.

With the feedback that I received in mind, I began to learn more about SPSS and how to do certain functions. The Advanced Business Data Analysis module was helping my understanding of SPSS more and more as we went along. But I did not make a start on the SPSS section of my project until the end of the month just so I had a clear view on what I wanted from my SPSS and how I could achieve it without making too many mistakes.

Achievements for This Month:

Having the presentation ready in time to a certain standard was a minor achievement. But where I felt I had progressed the most was in my learning of RStudio and SPSS. I felt that I had made big gains on my knowledge in both of these areas and could use what I have learned in my project. I was now able to create an R programme that could read in a data file and successfully create some basic statistics.

Reflections:

I am taking this month as a positive month towards my project. The positive but yet honest feedback that I have received from Eugene and Ron has encouraged me to focus more on my project. They said that this project has potential and I plan on completing this project to the fullest degree possible.

Changes for Next Month:

Begin working on the SPSS aspect of the project and begin working on WordPress for my web application. Also finding a different way to cleansing my data or storing my data in a different way.

Month: March 2016

The very first thing that I wanted to get done this month was to sort out my data properly. Initially, I decided to store all of the data in a single Excel file. This proved to be problematic however as the more data I added, the more unclean it became. When I would read the file into RStudio, the display I got from R was very messy and was unable to perform any of the calculations that I had wanted to do. So what I decided to do then was to store the data in multiple Excel files. This proved to be a much better idea than the one previous as, although I would have to read in more data files, the calculations became much easier to do. The program was now able to read the files in successfully and cleanly without having too much data cleansing to do.

After sorting that problem out, I began looking at WordPress. WordPress is an online tool that can help you create a web application or blog for free. Obviously, I intend to use this for creating my web application. I have never used WordPress before and my skills for creating a site would not be of the highest standard, so I was very nervous about this section of the project. It would take a lot of work and learning to do in order to create a good web application. As I began working on WordPress I quickly realised that I found it as a difficult tool to use since I had no previous experience of using it.

I quickly moved on to SPSS just so I would not get delayed by my difficulties of using WordPress. As the semester went on, I gradually learned about how to use SPSS and more of the functions and statistics that it was capable of providing me with. So I began to incorporate some of the things that I had learned in the Advanced Business Data Analysis class. I was able to display basic descriptives of some data and able to conduct relevant tests.

Achievements for This Month:

Organising relative data into separate, functioning data files that could be read in RStudio and be able to conduct tests on the data. Advance my learning on SPSS and create basic descriptives and perform tests.

Reflections:

Despite struggling with WordPress, I feel that a lot of work has been completed this month. A lot of knowledge has been gained and a good amount of progress has been made this month.

Changes for Next Month:

Learning more about WordPress and creating some statistical findings using Excel.

Month: April 2016

My struggles with WordPress continues. I am finding it very difficult to create my web application using this tool. I have come across tutorial videos on YouTube to try and guided me through this process, but to no avail. I just keep on making the same mistakes that I had been in the previous month and is getting slightly frustrating by this stage. It is also hindering my progress on other parts of my project which is the most frustrating part. However, I was recommended by a friend to use W3Schools which helps you learn how to code, so I will be using W3Schools to hopefully help me improve my skills in coding.

On the plus side of things, I have managed to create a world map and a complete dashboard using Tableau. I am very happy with this as it is working in the way that I had intended it to work. Progress has also been made in RStudio and SPSS where more tests have been developed and created.

Achievements for This Month:

Progress continued in RStudio and SPSS and the completion of the Tableau section.

Reflections:

I am disappointed with my lack of progress in WordPress. I seem to be having a great amount of difficulty trying to use it as my skills in coding need to be improved vastly. Other than that I would say the project is coming along nicely but I feel that I am still behind and have a lot of work ahead of me for next month.

Changes for Next Month:

Fully competing the project.

Month: May 2016

It is weird to think that this is my final diary entry. The college year has been full of ups and downs, inside and outside of the project, but alas I am here typing up what is to be my final diary entry.

Despite having to re-learn some coding through W3Schools, I have not been successful in trying to develop my web application through WordPress. I am finding it very difficult to use and time is running out. I am going to have to use my last resort as a way of completing my web application, Wix. It is an easier tool to use and I feel I will be able to create something good out of it but I cannot help but feel that it is going to cost me some marks. But I would rather have something to show for at the presentation and the showcase instead of having nothing at all. I will have to create my application to a high standard in order to obtain good marks. The final report has been created with all the statistical findings that I used for the duration of the project and with RStudio fully functioning.

Final Diary Entry

It is incredible what has happened over the past year of this project. From so much confusion at the start of the college to having the project fully completed and ready on-time, it has been an experience that I will not forget (for both good and bad reasons). But what is most important is the knowledge that I have gained from creating the project and the workload that will be expected of me outside of college life.

The End

References:

Michael Schumacher Picture:

Bibliography: Nipun (2014) *Role of Michael Schumacher's parents in his racing career*. Available at: http://inspireicons.com/his-father-did-two-jobs-and-mother-worked-in-food-canteens-to-ensure-he-race-michael-schumacher/ (Accessed: 11 May 2016).**In-line Citation:**(Nipun, 2014)

Formula 1 logo:

Bibliography:*File: F1 logo.svg* (2015) in *Wikipedia*. Available at: https://en.wikipedia.org/wiki/File:F1_logo.svg (Accessed: 11 May 2016).**In-line Citation:**(*File: F1 logo.svg*, 2015)