

**AY2023/24 SEMESTER 2**

**MH3511 - DATA ANALYSIS WITH COMPUTER**

**Title: Exploratory Data Analysis of Drivers’ Performance, Standings and Earnings in Formula 1 Dataset**

| **Group Members** | **Matriculation Number** |
| --- | --- |
| Garren Wee Qiming | U2221122F |
| Tan Jie Ning, Jolynn | U2220577B |
| Wong Liek Zhe | U2220416D |
| Pearlina Tan Qinlin | U2221690F |
| Greg Oh Wenbin | U2240893L |

# 

**Table of contents**

[**1. Introduction 3**](#_xpefbg616lac)

[**2. Data Preparation and Analysis 4**](#_wz2ppf6szgs)

[2.1 Summary statistics of driver’s experience level and average points earned per race for each driver in the most recent 10 years 4](#_nme3m0veewo5)

[2.2 Summary Statistics for Effects of starting from Pole Position on Final Position 5](#_xf10fktpq6m7)

[2.3 Summary statistics for drivers’ experience level and drivers’ salaries in 2023 6](#_hwqtx4uure3s)

[2.4 Summary statistics for drivers’ total points and drivers’ salaries in 2023 7](#_qaer7lvu43ig)

[**3. Statistical Analysis 10**](#_nvpkw4jg91t6)

[3.1 Analysis of the relationship between drivers' experience level and average points per race in most recent 10 years 10](#_dbgf5naxpa0g)

[3.2 Analysis of the relationship between pole position and final position 11](#_q82wogayegy8)

[3.3 Analysis of the relationship between drivers’ salaries and drivers’ experience levels 11](#_ib3qhw2g2ytq)

[3.4 Analysis of correlation between drivers’ salaries and drivers’ points 12](#_nvg00kmbu6tf)

[**4. Conclusion 13**](#_rmuqh6kzjlw5)

[**5. References 15**](#_dz9bqespivbo)

[**6. Appendix 16**](#_o0qd0wv8gkyd)

# 1. Introduction

Formula 1 (F1) is the most popular motorsport in the world today, demanding peak performances from the drivers, constructors and machines. The cars are capable of reaching 230 mph on the most difficult circuits around the world. Cars moving at such high speeds require years of experience and training from the driver to manoeuvre around the circuits. The most decorated F1 driver today is Max Verstappen who has a reported annual salary of $55 million in 2024. He has won 3 straight championships which requires him to have the most number of points at the end of the year (Fakas-Drosos, 2023).

Our report aims to use various statistical analysis including ANOVA test, fisher test, correlation coefficient, linear regression and Kruskal-Wallis test to study the relations between various factors that we determined to influence the success of a driver. Other than the performance of a driver, we will also delve into the financial aspect of the 20 drivers that make up the current F1 grid and see how their performance affects their salary. By analysing these statistical connections, we can gain valuable insights and make informed conclusions about the world of F1.

The dataset we mainly used is “Formula 1 World Championship (1950 - 2023)” from Kaggle.com. The dataset contains 14 data frames and the ones we employed are “drivers.csv”, “results.csv”, “driver\_standing.csv”. We also added additional sources that we scrapped from two other webpages for more reliable insights regarding drivers’ points and salaries specific to the year of 2023, taking salary figures from RacingNews365.com and the official Formula 1 website for drivers’ standings data (refer to References for links of the 3 sources).

Based on this data set, we seek to answer the following questions:

1. Does the driver’s experience affect the average points earned per race the driver participated in in the most recent 10 years?
2. Does the pole position affect the final position of the driver?
3. Does the driver’s experience affect the driver’s salary?
4. Does the driver’s average points affect the driver’s salary?

# 2. Data Preparation and Analysis

We will now look into the variables used in our statistical analysis.

## 2.1 Summary statistics of driver’s experience level and average points earned per race for each driver in the most recent 10 years

First, for F1 races that occurred in the most recent 10 years, we find the average points earned per race for each specific driver by with:

for the most recent 10 years in the dataset.

We define the experience level as the total number of races drivers participated in in the course of their career, which would be recorded in the vast dataset. The drivers are split into different experience\_level (Low, Med or High) according to how many races they have participated in their entire career. Low: 0-99 races; Medium: 100-199 races; High: more than 200 races.

Boxplot below shows the effect of the Experience Level on Average points earned per race in the last 10 years. It only contains data of drivers who participated in at least one race in the past 10 years.

| Figure 1: Boxplot of the effects of Experience Level on Average points earned per race in last 10 years before removing outliers | Figure 2: Boxplot of the effects of Experience Level on Average points earned per race in last 10 years after removing outliers |
| --- | --- |

We observe that boxplots, after removing outliers, the interquartile range of each boxplot is similar.

## 2.2 Summary Statistics for Effects of starting from Pole Position on Final Position

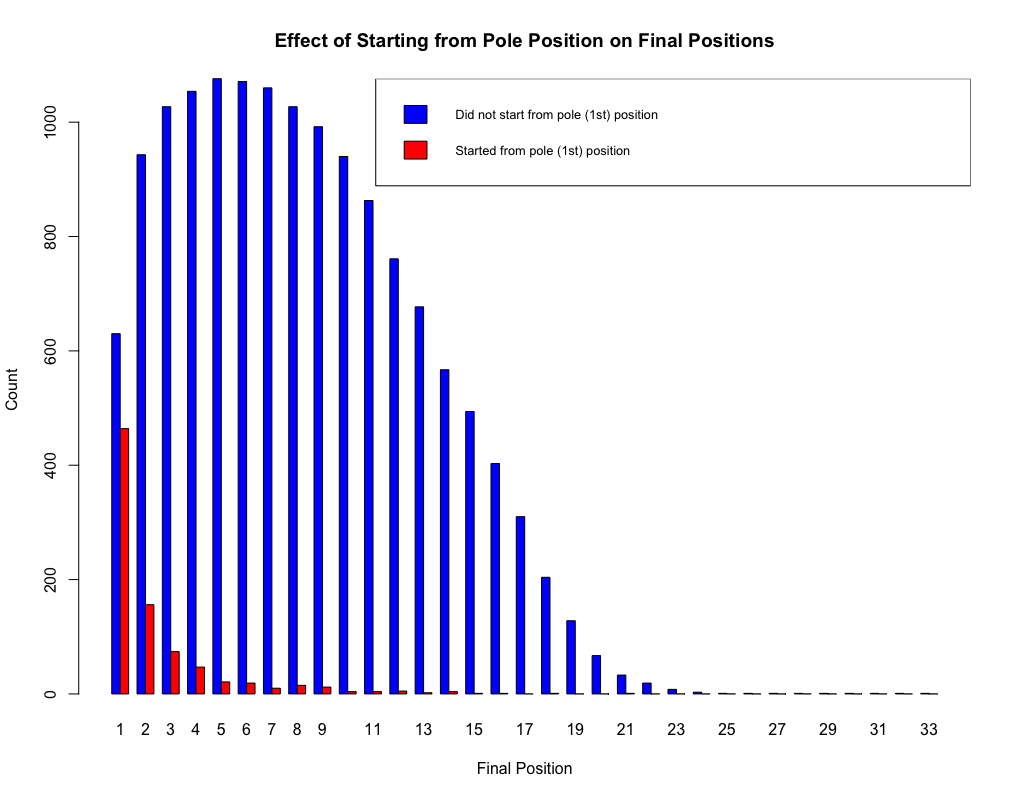


Figure 3: Effects of starting from pole position on final positions

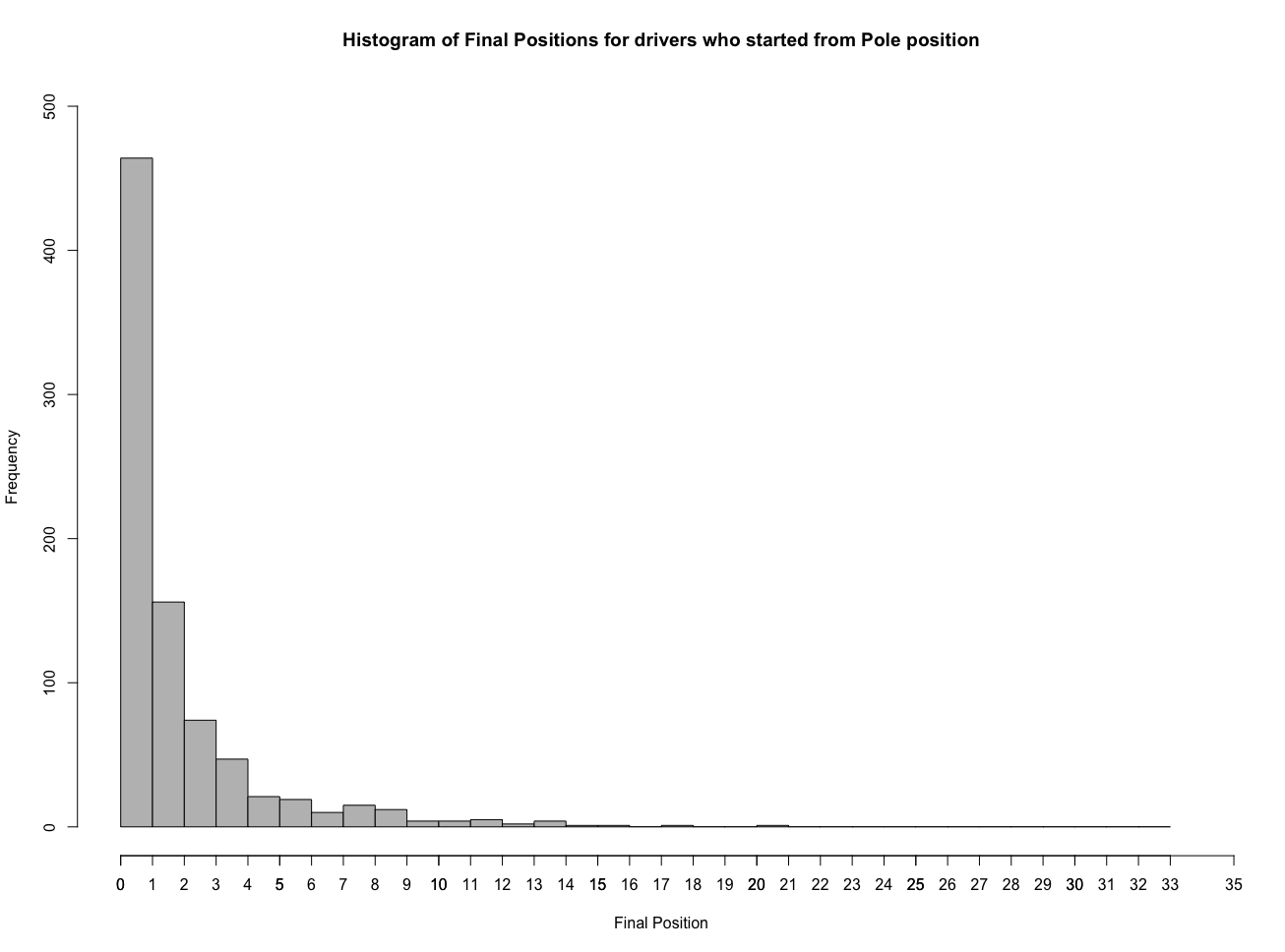
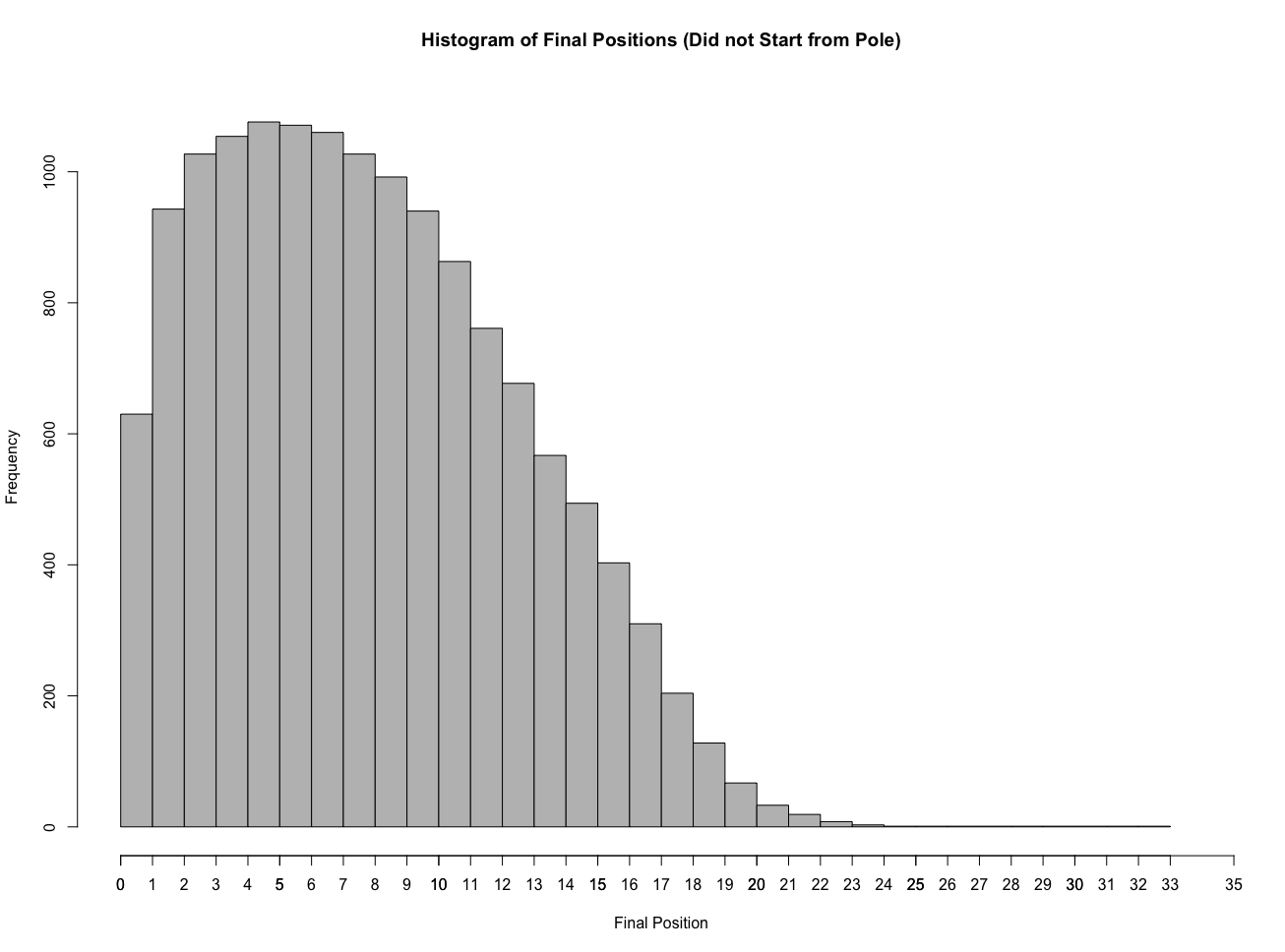
 

Figure 4: Histogram for final positions of Figure 5: Histogram for final positions of

drivers starting from pole position drivers who did not start from pole position

This shows that the most drivers who started from pole position will end up getting the first position and those who did not start from pole position mostly obtained 5th to 6th place.

## 2.3 Summary statistics for drivers’ experience level and drivers’ salaries in 2023

We defined the experience of a driver as the number of races they have participated in. We took the 20 drivers that make up the current F1 grid from RacingNews365.com. The mean number of races the 20 drivers participated in is 147.65, maximum is 370 and minimum is 12.

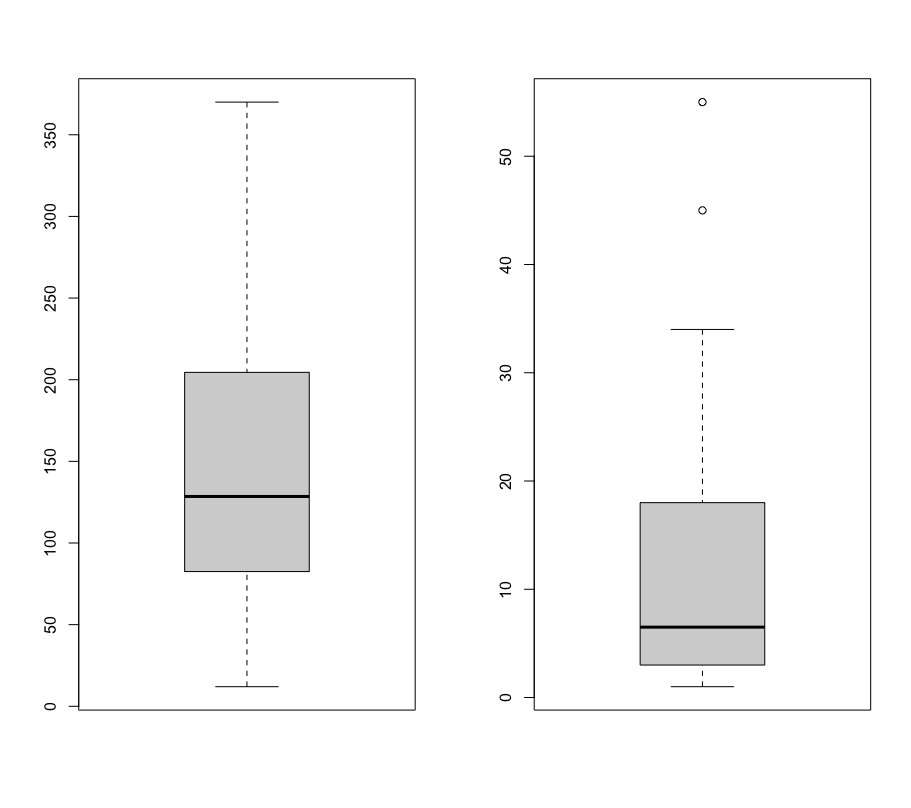


Figure 6: Boxplot of number of races participated in Figure 7: Boxplot of drivers’ salaries

The drivers are split into different experience levels (Low, Medium or High) according to how many races they have participated in. Low: 0-99 races; Medium: 100-199 races; High: more than 200 races.

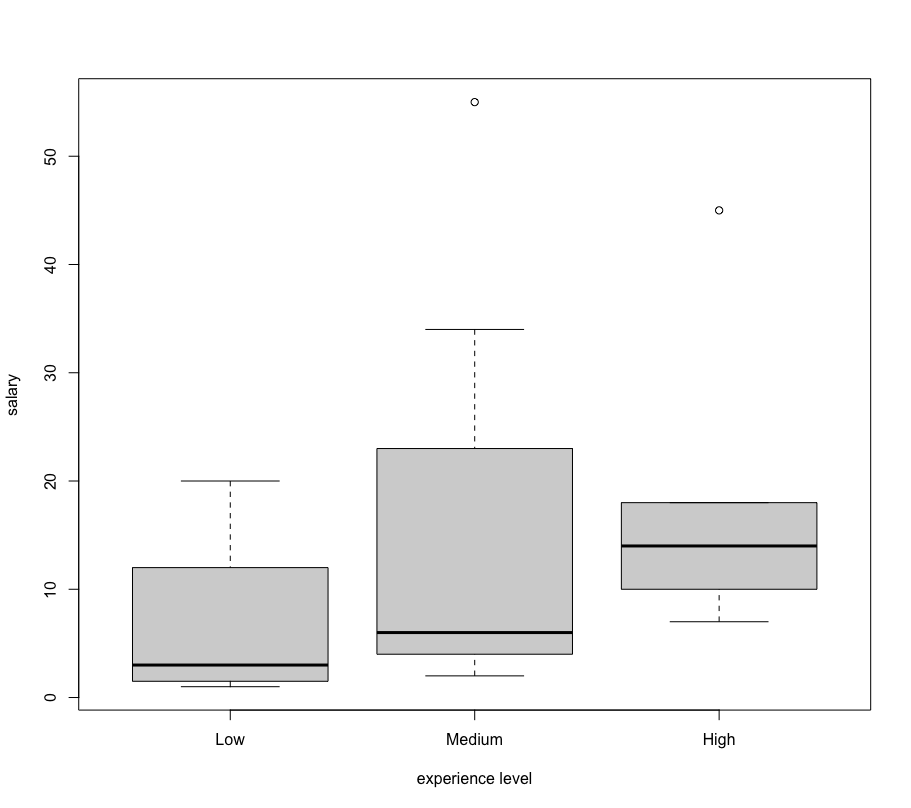


Figure 8: Boxplot of drivers’ salaries by experience level

## 2.4 Summary statistics for drivers’ total points and drivers’ salaries in 2023

After scraping, drivers’ points data from formula1.com had initially one outlier and was relatively skewed from a normal distribution, as observed in the qqplot and boxplot performed on the dataset.

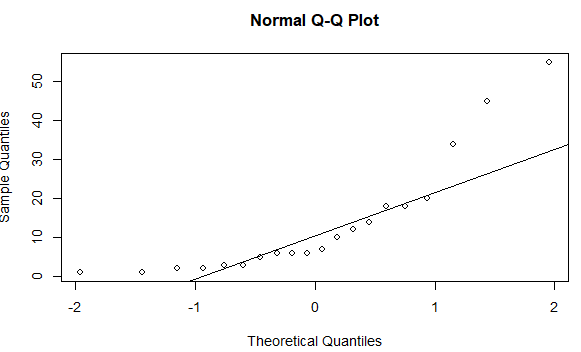
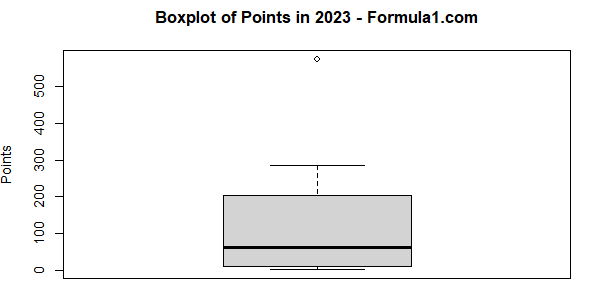


Figure 9: box plot and qq plot of drivers’ points

To clean the data, the outlier was removed and logarithmic transformations were applied to the data. The following qqplot shows improvements in normality and reduced skew of the dataset.

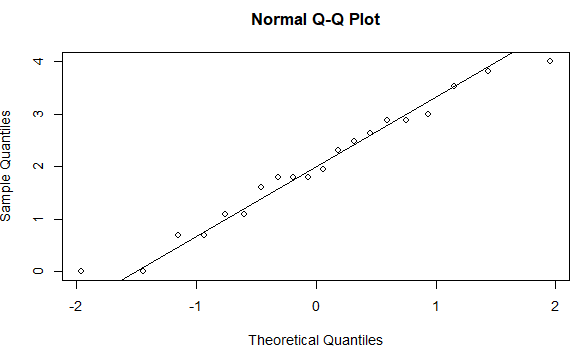


Figure 10: qq plot of log drivers’ points

Upon web scraping, drivers’ salaries data was observed to be skewed towards lower values and did not follow a normal distribution. 2 outliers were observed, before logarithmic transformation.

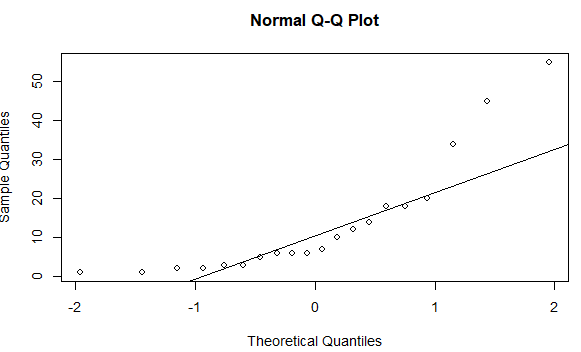
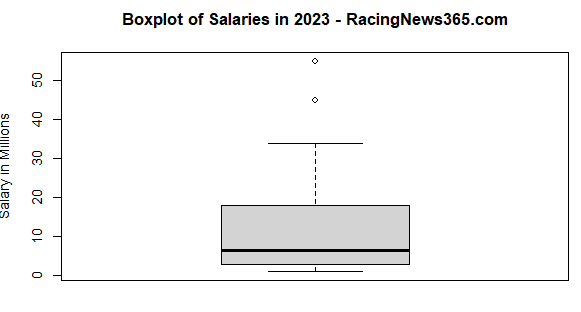


Figure 11: box plot and qq plot of drivers’ salaries

After logarithmic transformation, no outliers were observed subsequently in the data. qqplot shows that the data has reduced skew and fit closer to a normal distribution than before.

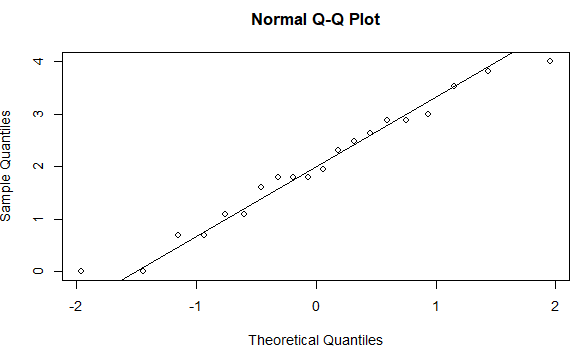


Figure 12: qq plot of log drivers’ salaries

Histogram to show the distribution of drivers’ points upon logarithmic transformation. Total sample size is 20 drivers with a mean of 3.531, maximum of 5.652 and minimum of 0.

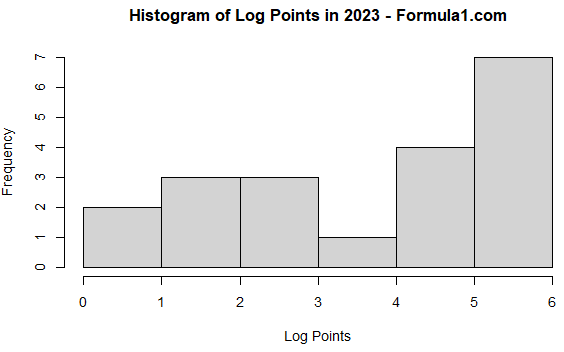


Figure 13: Histogram of log points

Histogram to show the distribution of drivers’ salaries upon logarithmic transformation. Total sample size is 20 drivers with a mean of 2.003, maximum of 4.007 and minimum of 0.

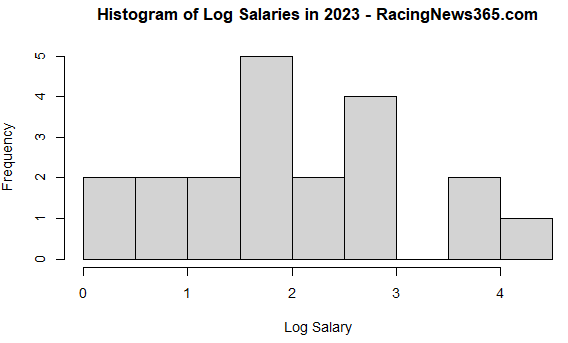


Figure 14: Histogram of log salaries

Data for this section pertains exclusively to the year 2023, which stands as the most recently completed year of races and events, offering a more up-to-date and accurate perspective on the analysis.

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

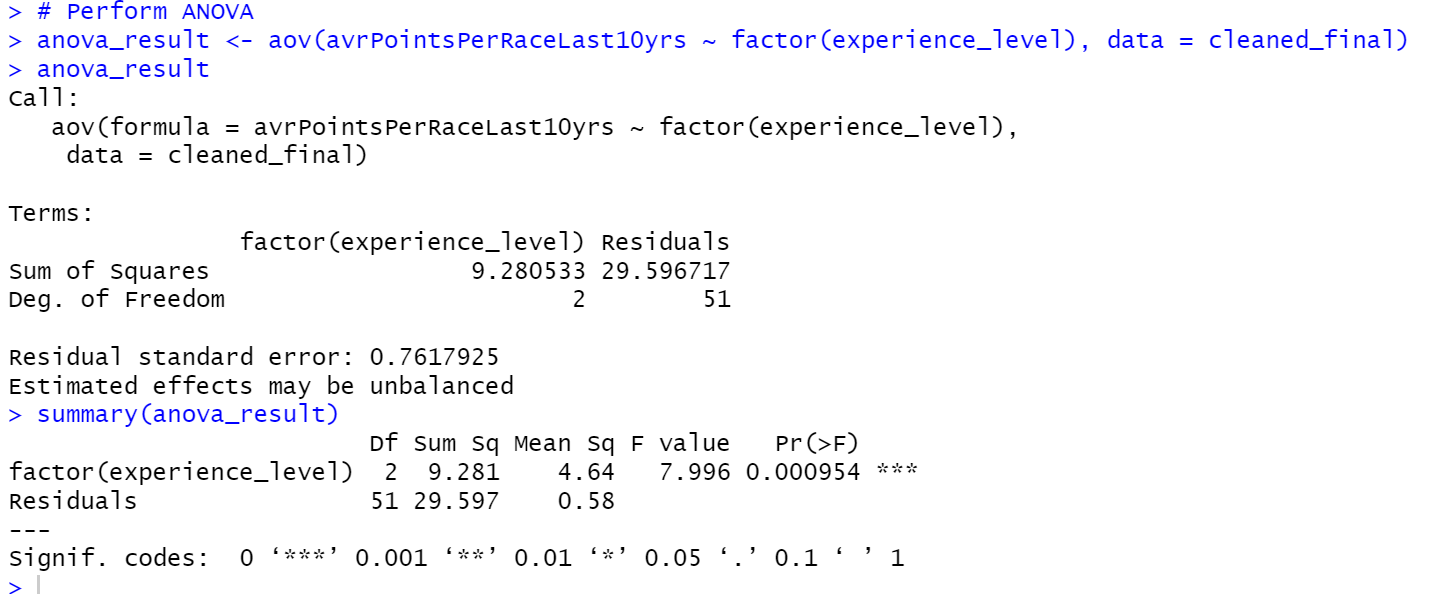
# 3. Statistical Analysis

## 3.1 Analysis of the relationship between drivers' experience level and average points per race in most recent 10 years

It is a common fact that “practice makes perfect”, thus we will assume that the drivers with more experience will tend to have higher average points (be placed in higher ranks) per race. Therefore, we will analyse the average points per race each driver participated in the most recent 10 years of the dataset with the number of F1 races they participated in over the course of their career.

Noting that the boxplots, after removing outliers (Fig. 1), the interquartile range of each boxplot is similar, we infer that the data for each experience level (rank at “Low”, “Med”, “High”) has similar variances.

Moreover, under the assumption that the data for each driver is independent and that it is normally distributed, we use ANOVA to find if the mean average points per race each driver participated in is affected by their different experience levels over the past 10 years. ANOVA is appropriate here when we are interested in knowing if there is an overall effect of the factor, experience level, on the outcome measure, the average points earned per race, where the data for the groups are independent, normally distributed and have similar variances.

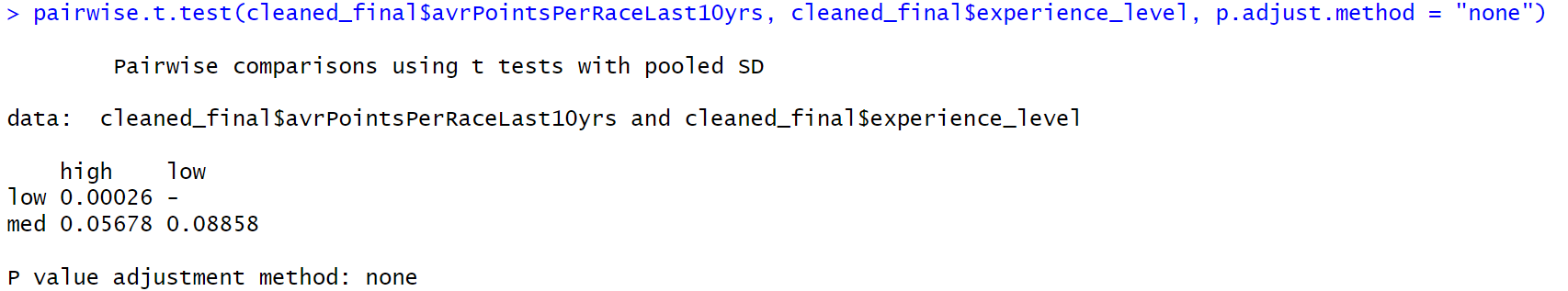


H0: mean average points per race for each race each driver of different experience levels over the past 10 years are equal.

H1: mean average points per race for each race each driver of different experience levels over the past 10 years are not equal.

The p-value (0.000954) from ANOVA test conducted on the average points per race for each driver is lower than 0.05, showing that there is sufficient evidence to reject H0, and thus the mean average points per race for each race each driver of different experience levels over the past 10 years are not equal.

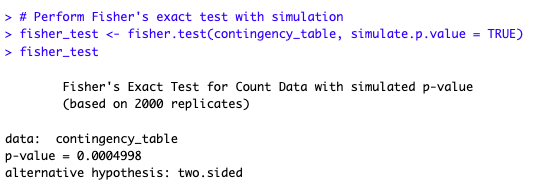
We proceed with a pairwise comparison by using a pairwise t-test to determine which group ( “Low”, “Med”, “High”) is different from the other groups.



From the pairwise t-test, we see that high and medium experience levels (with p-value = 0.05678 > 0.05), and low and medium experience levels (with p-value = 0.08858 > 0.05) are quite similar in average points earned per race. However, low and high experience levels are very different in average points earned, with p-value = 0.00026 < 0.05.

## 3.2 Analysis of the relationship between pole position and final position

For a F1 race, understanding the impact of pole position on race is essential as F1-teams and drivers rely on data and analysis to develop race strategies. If pole position has a significant impact on final positions, F1-teams can adapt their strategies by prioritising their qualifying performances.



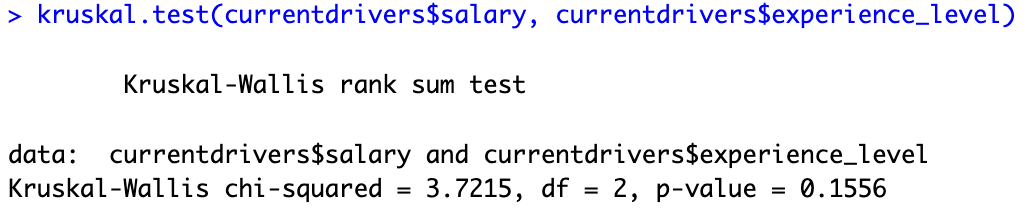
H0: Starting from pole position does not affect the final position.

H1: Starting from pole position does affect the final position

By performing Fisher's exact test with simulation, we are testing whether the observed association between starting from the pole position and final position is statistically significant, under the assumption that H0 is true. Since p-value is lesser than 0.05, we have sufficient evidence to reject H0 and conclude that pole position does affect the final position.

## 3.3 Analysis of the relationship between drivers’ salaries and drivers’ experience levels

When we look at the financial aspect of F1, we will assume that drivers with more experience tend to have higher salaries than drivers with lesser experience. This can also be observed from figure 8 which shows that drivers’ salaries are higher as they gain more experience. Therefore, we will analyse the salaries of drivers in 2023 with differing experience levels. We deduced that the variances in salaries of the drivers with differing experience levels are not equal as seen from the boxplot in section 2.3. We concluded that ANOVA test will not be suitable as it requires the variances of each group to be equal. Therefore, we decided to utilise the Kruskal-Wallis test which is a non-parametric test.



H0: Median salary of drivers of different experience levels are equal.

H1: Median salary of drivers of different experience levels are not equal.

The p-value (0.1556) from the Kruskal-Wallis test conducted on the drivers' salaries shows that H0 is not rejected and thus the salaries of drivers of different experience levels may be equal. We then take a look at the correlation between drivers’ salaries and their experience levels.



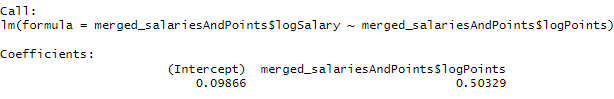
We observe that the correlation coefficient (0.4143743) indicates that there is a moderate level of correlation between experience levels and salaries. While the p-value shows that it is probable for the median salary to be equal, it does not mean that there is no relation between salaries and experience levels as we can see from the correlation coefficient that there is still some correlation between salaries and experience levels.

## 3.4 Analysis of correlation between drivers’ salaries and drivers’ points

Understanding the potential correlation between drivers’ salaries and the points they have achieved during a specified timeframe allows a deeper understanding of how effective compensation structures are put in place for drivers to achieve their respective performances. This can prove useful in projecting drivers’ salaries based on their achieved points, and understanding the general magnitude of the incentive strategies put in place to reward players

in the industry.

Linear regression is used because it is able to express the relationship between the two factors in concern as an equation, even allowing the projection of salaries with increased points. Logarithmic transformations are accounted for in the final regression equation.

****

Calculating the equation for the best-fit line between the transformed salary and points data results in the following:

log(y) = 0.50329 \* log(x) + 0.09866

Where y is the salary (in millions, USD) obtained by a driver

x is the number of points obtained by a driver

Our observations of a significant R2 value suggests the relationship between drivers' salaries and their performance points, and how these variables are closely linked to each other. Points achieved can therefore serve as a reliable predictor of salaries pertinent to the dataset, and possibly for subsequent years as well.

# 4. Conclusion

In conclusion, analysing the data we have acquired provides several interesting insights into the world of Formula 1, with regards to various factors that impact drivers’ performances, standings and their earnings.

Firstly, we utilised ANOVA to observe the possible correlation between drivers' experience levels and the average points they earned for each race over the past decade. Our findings found a significant relationship between the two factors, suggesting that more experienced drivers were consistent in achieving more points than other peers in the industry.

Fisher’s exact test was also useful in helping us to observe that drivers who begin races in the pole position will have a significantly higher advantage compared to their competitors, according to corresponding analysis with their final positions achieved, signifying the qualifying performance of each driver.

Regarding finances, even though the Kruskal-Wallis test showed an insignificant difference in median salaries across experience levels, a moderately positive correlation suggests a possible relationship between the factors. We inferred that experience hence played somewhat of a role in determining one’s earnings, even if not a major determinant.

We finally employed linear regression to examine the correlation between drivers' salaries and their points, using the year 2023 as a recent example of observation. Our analysis revealed a strong linear relationship, and that drivers' on-track performance directly impacts their financial compensation.

Thus, our study allowed us to explore the roles leading to various types of success in Formula 1, in terms of both performance and financial factors.

# 5. References

​‌‌‍‍​‍Fakas-Drosos, J. (2023, November 5). Why is F1 so popular? - John Fakas-Drosos - Medium. Medium. <https://medium.com/@annarchitect.ddr/why-is-f1-so-popular-2b0c3d6273a9#:~:text=The%20Thrill%20of%20Speed%20in%20F1%20Racing&text=It%20is%20this%20adrenaline%2Dpumping,the%20world's%20most%20challenging%20circuits>

Formula 1 World Championship (1950 - 2023). (2023, August 24). Kaggle. https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020

Results. (n.d.). Formula 1® - the Official F1® Website. <https://www.formula1.com/en/results.html/2023/drivers.html>

Rencken, D. (2024, February 6). F1 driver salaries 2023. RacingNews365. <https://racingnews365.com/f1-driver-salaries-2023>

# 

# 

# 

# 

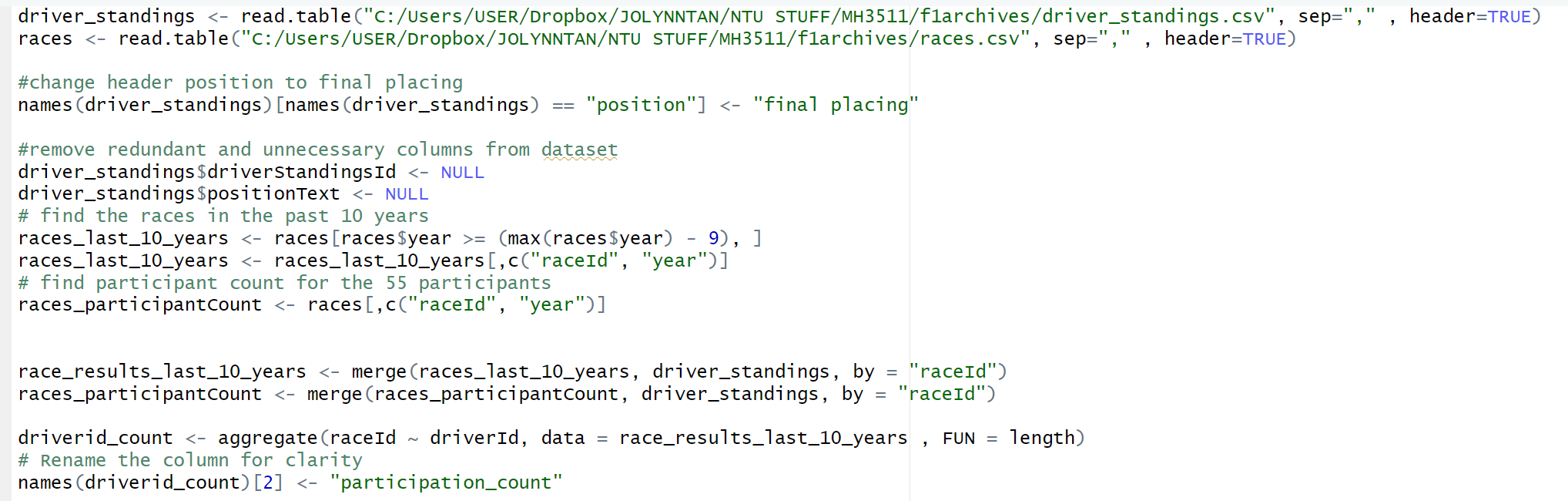
# 

# 

# 

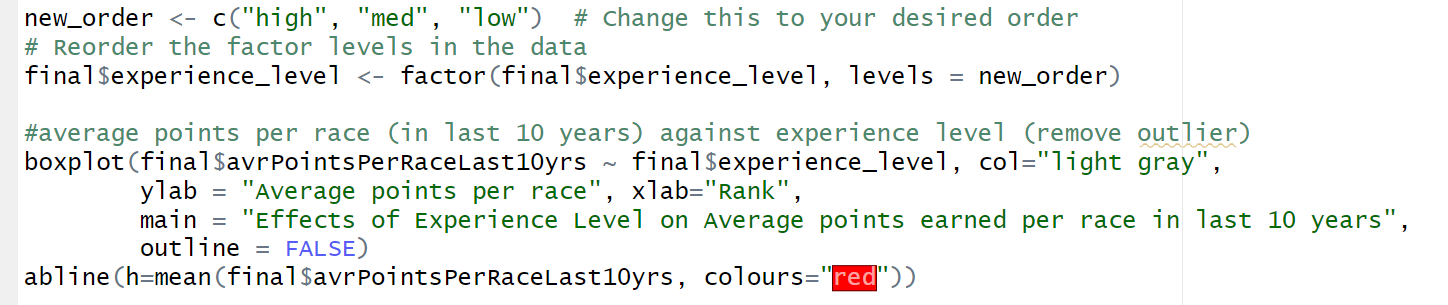
# 6. Appendix

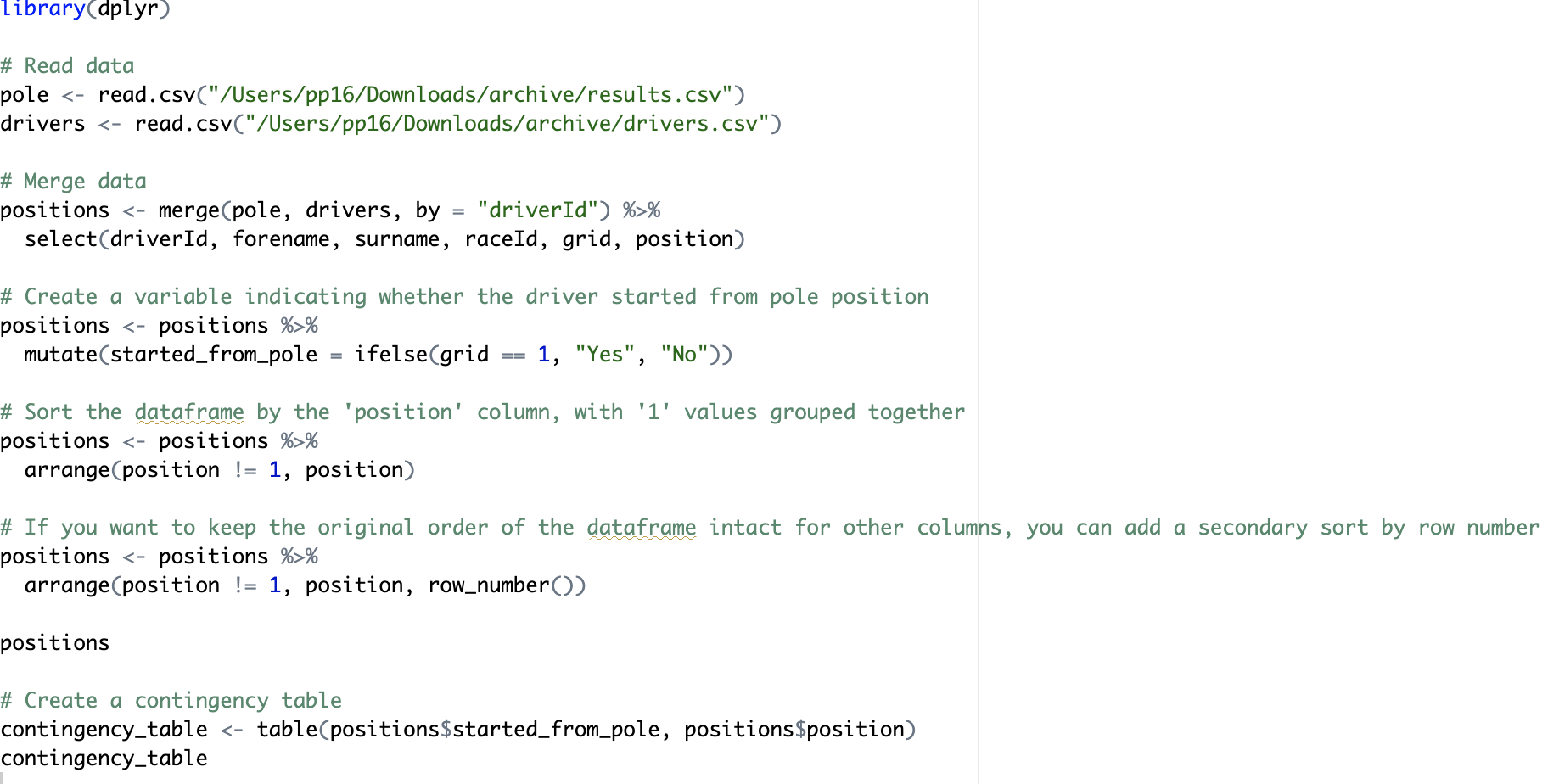
Section 2.1 preparation of data:



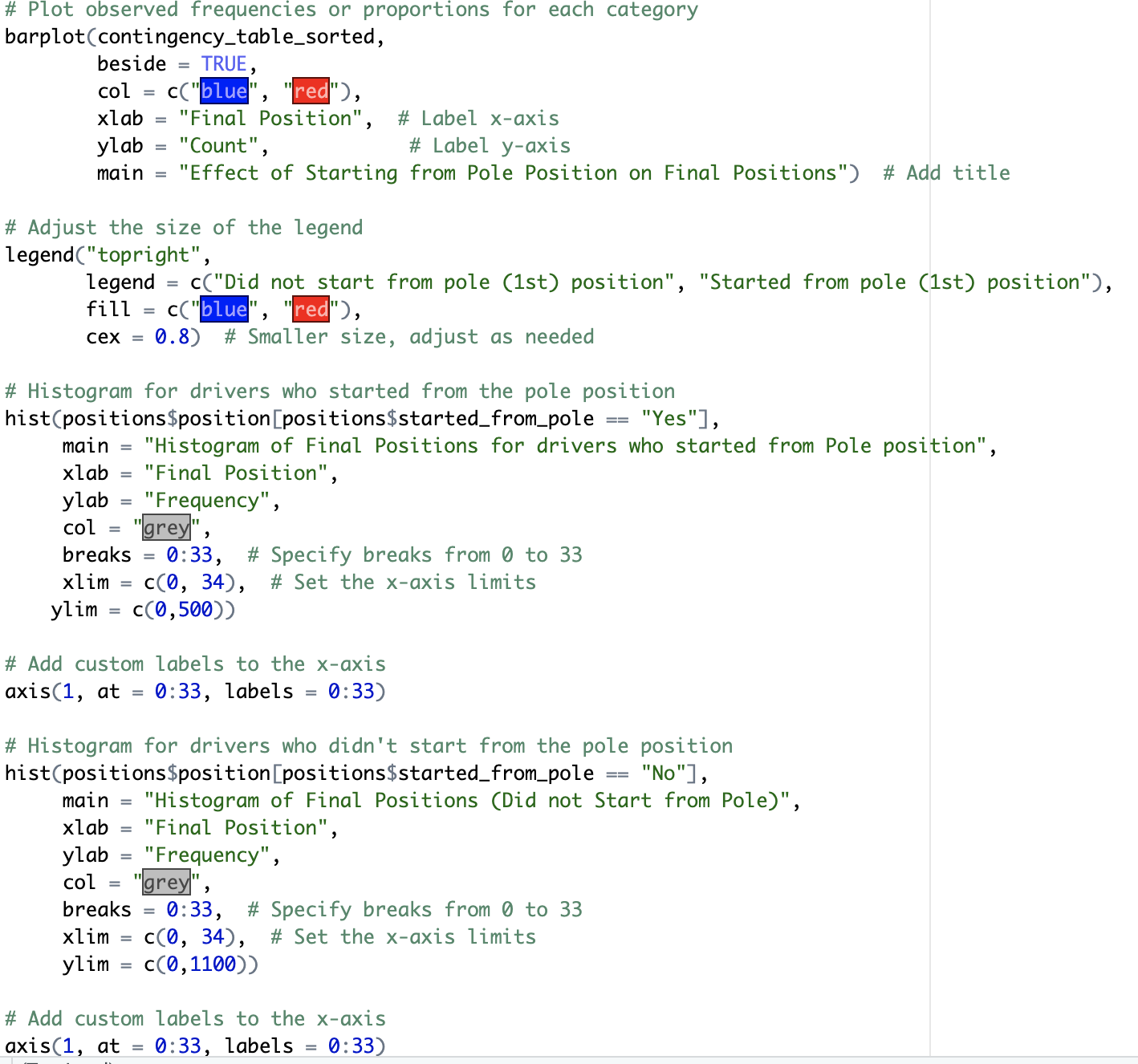


Section 2.1 plotting of boxplot:

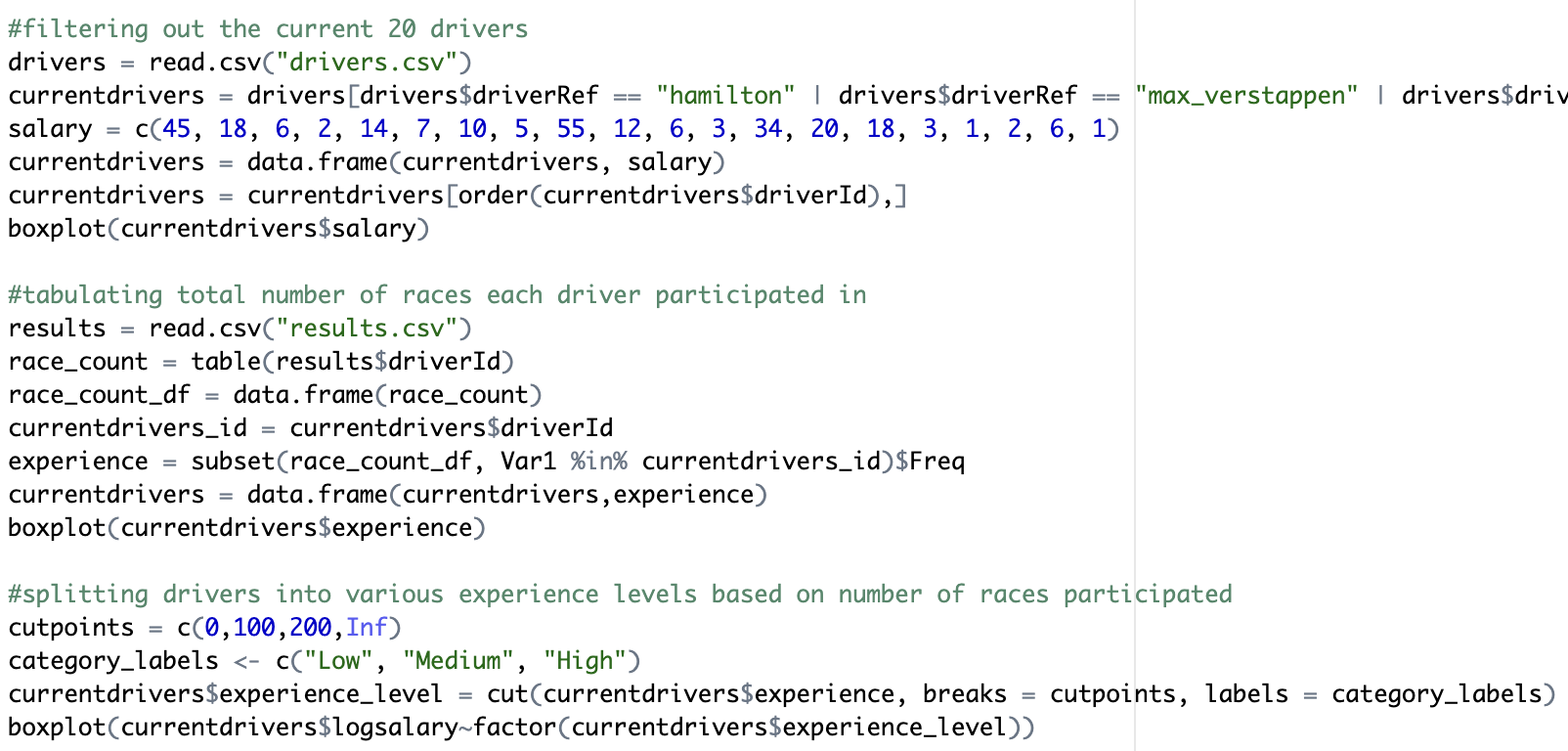


Section 2.2: Preparation of data

Section 2.2: Plotting of boxplot and histograms

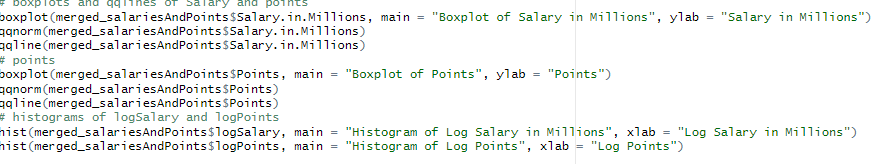


Section 2.3:

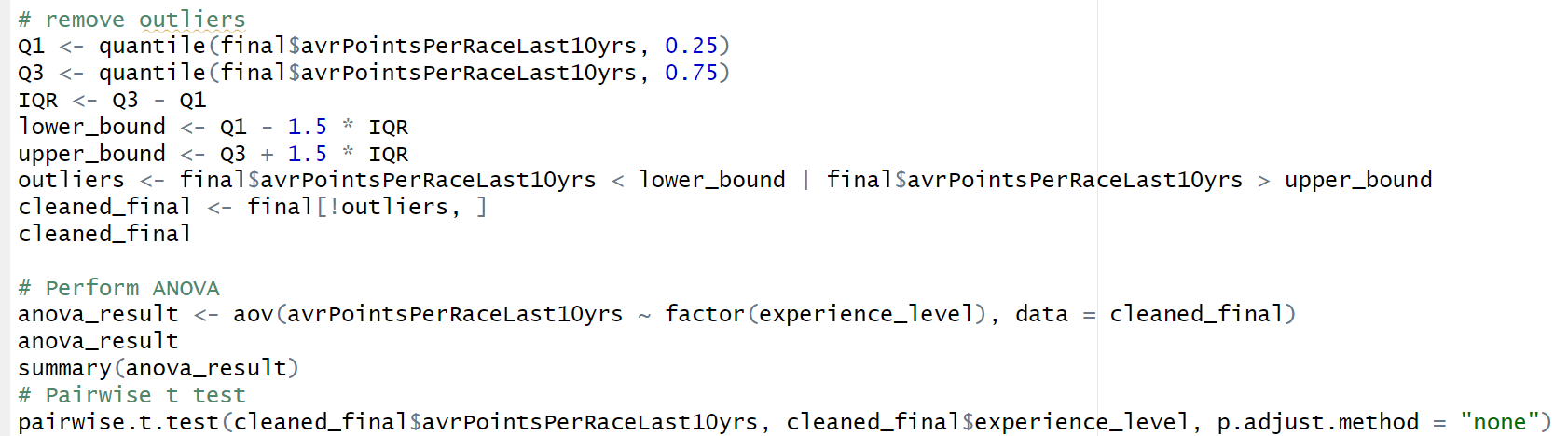


2.4 Preparation of Data and generation of figures





Section 3.1:



3.4 Calculation of linear regression between points and salaries

