Final Project

Nabhit Arora, Aahan Anand

Date: 19th April, 2024

```
library(tidyverse)
library(readr)
library(dplyr)
library(ggplot2)
library(broom)
library(knitr)
```

Introduction

This project conducts a thorough investigation into the correlation between the qualifying positions of Formula 1 drivers and their final race outcomes across multiple races. The analysis seeks to ascertain the impact of qualifying positions on race performance and evaluate other contributing factors.

Guiding Question:

Does the qualifying position significantly impact the final race results in Formula 1, and what other factors might influence race outcomes?

Data Access and Preparation

```
# Load the datasets
qualifying_data <- read_csv("/Users/nabhitarora/Desktop/STAT184/FinalProject/qualifying.csv")
race_results <- read_csv("/Users/nabhitarora/Desktop/STAT184/FinalProject/results.csv", na = "\\N") %>%
  mutate(
    raceId = as.numeric(raceId),
    driverId = as.numeric(driverId),
    positionOrder = as.numeric(positionOrder),
    statusId = as.numeric(statusId)  # Include status to account for DNFs etc.
  )

# Initial data inspection
glimpse(qualifying_data)
```

```
## Rows: 9,815
## Columns: 9
## $ qualifyId     <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1~
## $ raceId        <dbl> 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, ~
## $ driverId      <dbl> 1, 9, 5, 13, 2, 15, 3, 14, 10, 20, 22, 4, 18, 6, 17, 8, ~
## $ constructorId <dbl> 1, 2, 1, 6, 2, 7, 3, 9, 7, 5, 11, 4, 11, 3, 9, 6, 10, 5,~
## $ number        <dbl> 22, 4, 23, 2, 3, 11, 7, 9, 12, 15, 17, 5, 16, 8, 10, 1, ~
## $ position      <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1~
## $ q1            <chr> "1:26.572", "1:26.103", "1:25.664", "1:25.994", "1:25.96~
## $ q2            <chr> "1:25.187", "1:25.315", "1:25.452", "1:25.691", "1:25.51~
## $ q3            <chr> "1:26.714", "1:26.869", "1:27.079", "1:27.178", "1:27.23~
```

```
glimpse(race_results)
```

```
## Rows: 26,080
## Columns: 18
## $ resultId        <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,~
## $ raceId          <dbl> 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18~
## $ driverId        <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,~
## $ constructorId   <dbl> 1, 2, 3, 4, 1, 3, 5, 6, 2, 7, 8, 4, 6, 9, 7, 10, 9, 11~
## $ number          <dbl> 22, 3, 7, 5, 23, 8, 14, 1, 4, 12, 18, 6, 2, 9, 11, 20,~
## $ grid            <dbl> 1, 5, 7, 11, 3, 13, 17, 15, 2, 18, 19, 20, 4, 8, 6, 22~
## $ position        <dbl> 1, 2, 3, 4, 5, 6, 7, 8, NA, NA, NA, NA, NA, NA, NA, NA~
## $ positionText    <chr> "1", "2", "3", "4", "5", "6", "7", "8", "R", "R", "R",~
## $ positionOrder   <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,~
## $ points          <dbl> 10, 8, 6, 5, 4, 3, 2, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ laps            <dbl> 58, 58, 58, 58, 58, 57, 55, 53, 47, 43, 32, 30, 29, 25~
## $ time            <chr> "1:34:50.616", "+5.478", "+8.163", "+17.181", "+18.014~
## $ milliseconds    <dbl> 5690616, 5696094, 5698779, 5707797, 5708630, NA, NA, N~
## $ fastestLap      <dbl> 39, 41, 41, 58, 43, 50, 22, 20, 15, 23, 24, 20, 23, 21~
## $ rank            <dbl> 2, 3, 5, 7, 1, 14, 12, 4, 9, 13, 15, 16, 6, 11, 10, 17~
## $ fastestLapTime  <time> 01:27:00, 01:27:00, 01:28:00, 01:28:00, 01:27:00, 01:~
## $ fastestLapSpeed <dbl> 218.300, 217.586, 216.719, 215.464, 218.385, 212.974, ~
## $ statusId        <dbl> 1, 1, 1, 1, 1, 11, 5, 5, 4, 3, 7, 8, 5, 4, 10, 9, 4, 4~
```

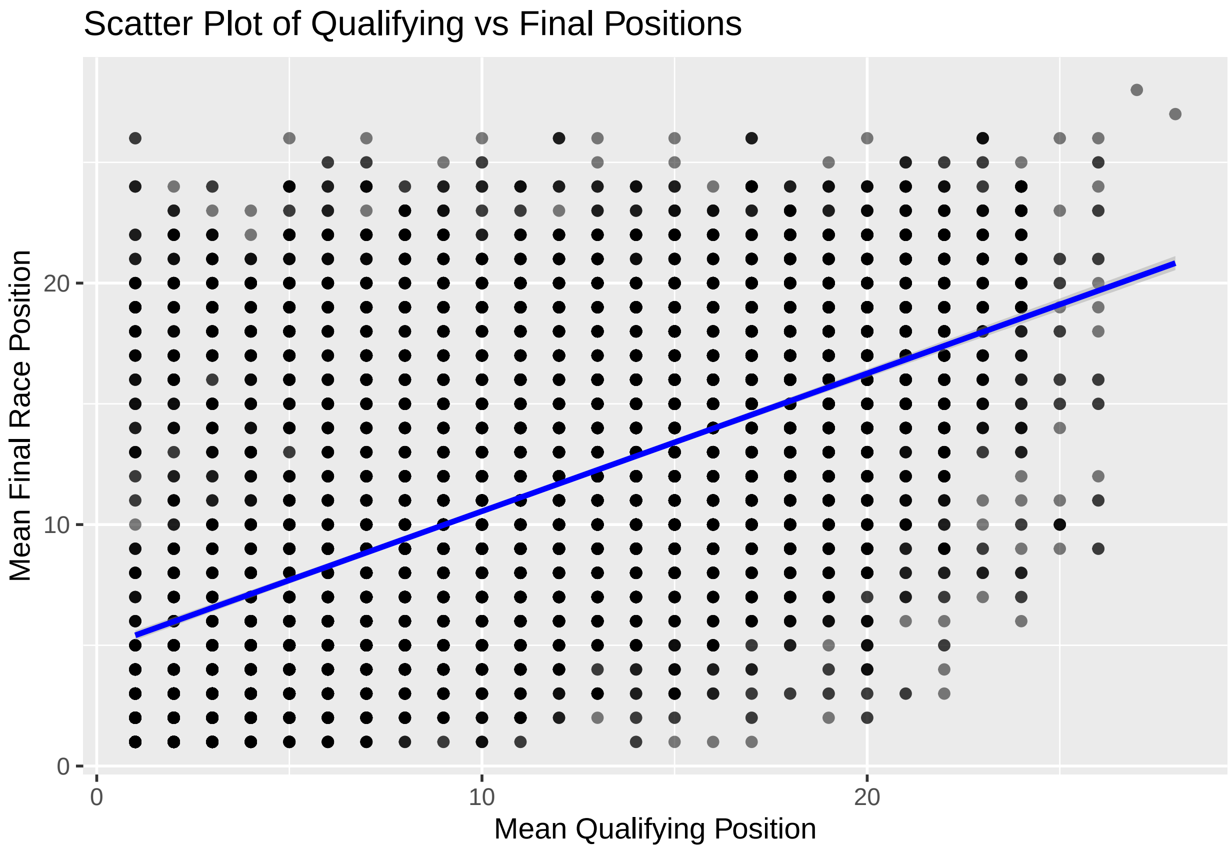
The first figure shows the output of the glimpse() function from the dplyr package for two datasets: qualifying_data and race_results.

Data Wrangling

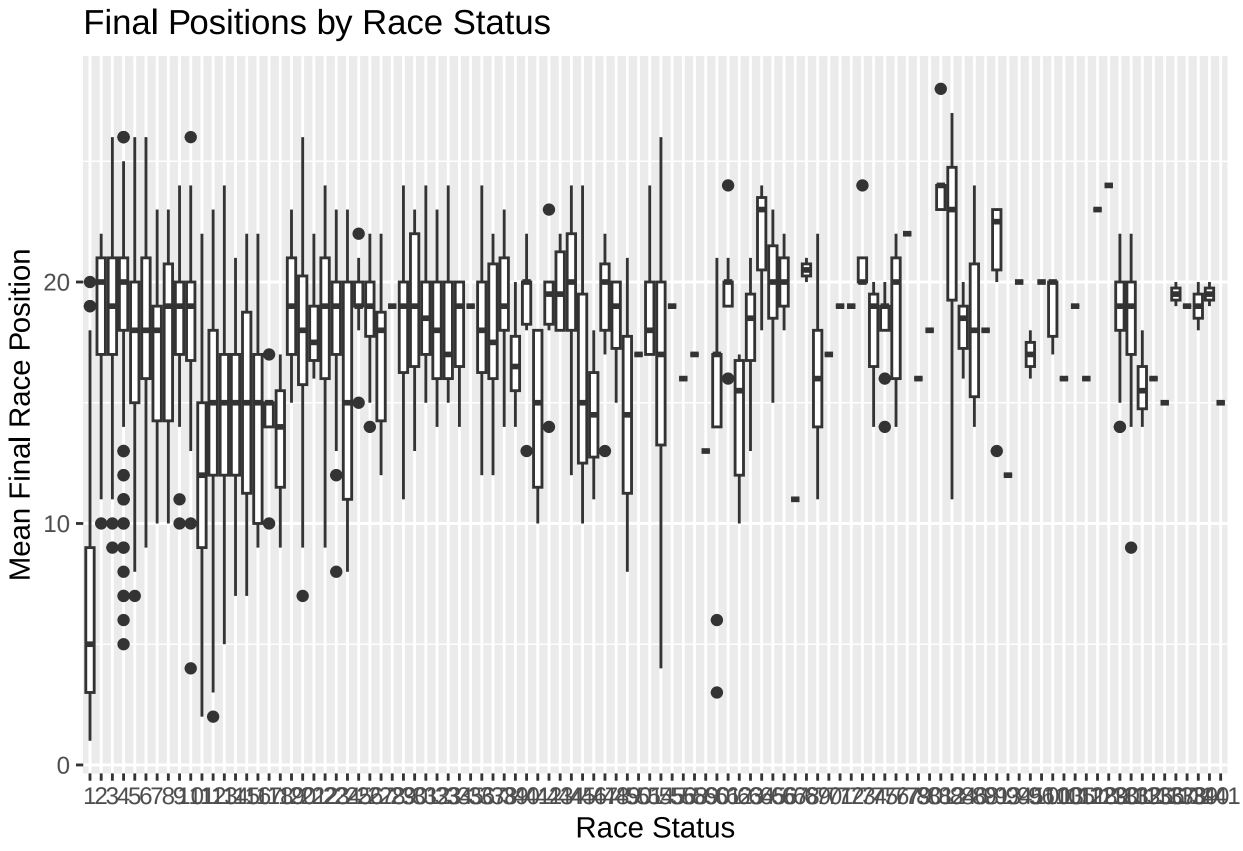
```
# Clean and prepare qualifying data
qualifying_data <- qualifying_data %>%
  filter(!is.na(position) & position > 0) %>% # Ensure the 'position' is valid
  mutate(quali_position = as.numeric(position)) # Rename and ensure 'position' from qualifying is numeric

# Clean and prepare race results data, making sure to keep the 'grid' column
race_results <- race_results %>%
  mutate(grid = as.numeric(grid), # Ensure 'grid' is numeric
         positionOrder = as.numeric(positionOrder)) # Ensure 'positionOrder' is numeric

# Join datasets based on raceId and driverId
full_data <- left_join(qualifying_data, race_results, by = c("raceId", "driverId"))

# Prepare data for analysis
analysis_data <- full_data %>%
  group_by(raceId, driverId) %>%
  summarise(
    mean_quali_position = mean(quali_position, na.rm = TRUE), # Use 'quali_position' from qualifying data
    mean_final_position = mean(positionOrder, na.rm = TRUE), # Use 'positionOrder' from race results data
    status = first(statusId)
  ) %>%
  ungroup()
```

Exploratory Data Analysis

```
# Explore data through various visualizations

# Qualifying position vs Final race position scatter plot
ggplot(analysis_data, aes(x = mean_quali_position, y = mean_final_position)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", color = "blue") +
  labs(title = "Scatter Plot of Qualifying vs Final Positions",
       x = "Mean Qualifying Position", y = "Mean Final Race Position")
```



```
# Boxplot to assess spread and outliers
ggplot(analysis_data, aes(x = as.factor(status), y = mean_final_position)) +
  geom_boxplot() +
  labs(title = "Final Positions by Race Status",
       x = "Race Status", y = "Mean Final Race Position")
```



Scatter Plot of Qualifying vs Final Positions

This scatter plot displays the relationship between mean qualifying positions and mean final race positions. Each dot represents an aggregation of these metrics, likely by driver and race.

Analysis:

There is a clear trend showing that a lower mean qualifying position (closer to pole position) tends to correlate with a lower mean final position (closer to 1st place in the race), implying that drivers who start at the front tend to finish at the front. The linear model (blue line) reinforces this trend, as it has a positive slope, indicating a direct correlation between the qualifying and final positions. The density of points is greater towards the lower end of the graph, which may suggest that it is more common for drivers to remain at the front if they start at the front, whereas variability increases with poorer starting positions.

Boxplot: Final Positions by Race Status

This boxplot categorizes the mean final race positions by the race status codes.

Analysis:

Race status seems to have a wide range of influences on the final race positions, as indicated by the variability in the boxplots. Some status codes correspond to more spread-out final positions, suggesting these statuses (possibly indicating retirements, disqualifications, or other incidents) are more disruptive to a driver's finishing position. There is a noticeable trend where certain statuses lead to consistently lower or higher positions, but due to the categorical nature of the status code, further investigation into what each code represents is needed to draw precise conclusions.

Statistical Analysis

```
# Multiple regression analysis
model <- lm(mean_final_position ~ mean_quali_position + status, data = analysis_data)
summary(model)
```

```
##
## Call:
## lm(formula = mean_final_position ~ mean_quali_position + status,
##     data = analysis_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.7201  -3.4270  -0.8924   2.6926  20.3601
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         4.493226   0.101219   44.39   <2e-16 ***
## mean_quali_position 0.514026   0.007979   64.42   <2e-16 ***
## status              0.105442   0.002943   35.83   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.873 on 9812 degrees of freedom
## Multiple R-squared:  0.4027, Adjusted R-squared:  0.4025
## F-statistic:  3307 on 2 and 9812 DF,  p-value: < 2.2e-16
```

```
# Output a detailed summary table of model coefficients
kable(tidy(model), caption = "Regression Analysis Summary")
```

Table 1: Regression Analysis Summary

term	estimate	std.error	statistic	p.value
(Intercept)	4.4932257	0.1012188	44.39120	0
mean_quali_position	0.5140258	0.0079787	64.42502	0
status	0.1054416	0.0029430	35.82790	0

Regression Analysis Summary

The regression output provides statistical details on the relationship between the average final position and both the average qualifying position and race status.

Analysis:

The intercept coefficient is significant and positive, suggesting that the expected final position without considering qualifying position and status is approximately 4.5. The coefficient for mean_quali_position is also significant and indicates that for each unit decrease in qualifying position (towards pole position), the final race position is expected to improve by about half a place, holding status constant. The status variable's positive coefficient implies that as the status code increases, which may be associated with less ideal conditions or occurrences, the final race position worsens slightly. The R-squared value of 0.4027 suggests that approximately 40% of the variability in the final race position can be explained by the model, which is a substantial but not complete explanation, indicating other factors also play significant roles.

Observations and Conclusion

The statistical analysis, including a multiple regression model, indicates that qualifying position is a significant predictor of final race outcomes, adjusted for factors like race status (e.g., Did Not Finish due to mechanical failure). Better qualifying positions are generally associated with better final race placements, emphasizing the strategic importance of qualifications in Formula 1 racing.

Challenges and Technical Issues

Data type inconsistencies and missing values were prevalent, requiring careful handling during the preprocessing stages. The mutate() and filter() functions from dplyr were instrumental in managing these issues efficiently.

Additional Insights

Further research could incorporate more external variables such as weather conditions, pit stop strategy, and team dynamics to broaden understanding of their impact on race results.

Conclusion

The analysis confirms the critical role of qualifying positions in Formula 1 races, with substantial effects on the race outcomes. Teams and drivers should focus on optimizing qualifying strategies to enhance their competitive advantage.