Data Exploration Analysis

Amanda Gomez
04/01/2023

WHAT DOES THE DATA REPRESENT?

F1 DATA SET

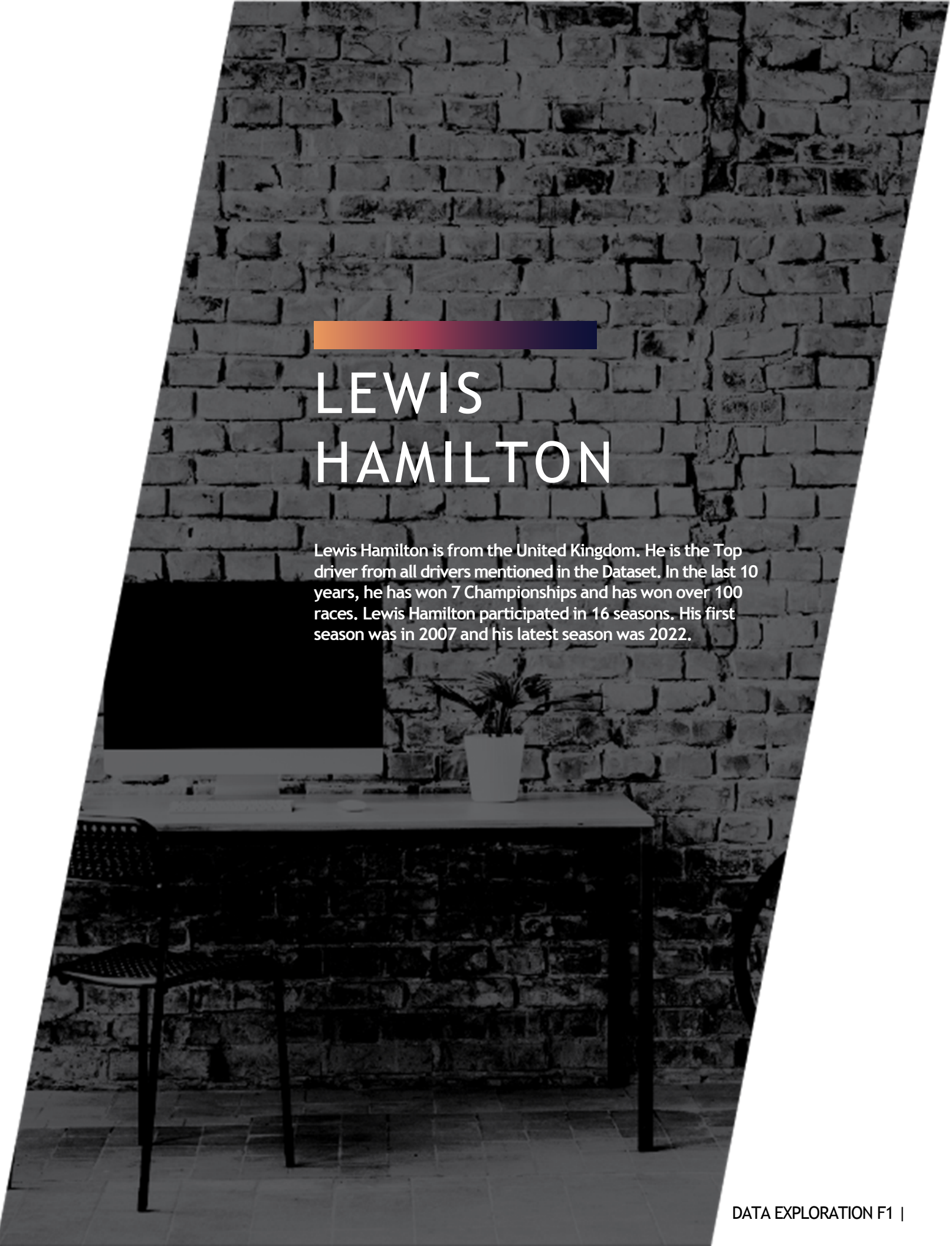
The F1 Dataset represents information about the Formula 1 Race Car contests. The F1 is an international auto racing sport. In the file, the columns list numerical data about the driver, nationality, season, championships, race entries, race starts, pole positions, race wins, podiums, fastest laps, points, active, championship years, decade, pole rate, start rate, win rate, podium rate, fast lap rate, points per entry, years active, and champion.

WHAT ARE IMPORTANT STATISTICS YOU FOUND?

I did most of my analysis around the nationality and the driver data. I found that based on the top 10 nationalities with the most championship wins, the United Kingdom has the highest with 30.3%. Second place belongs to the United States with 18.2%. Third place is Brazil with 12.1%. In relation to this I found that the United Kingdom also has the highest amount of race wins among the top 10 drivers. The last statistic I want to add is that 18.6% of all drivers are from the United Kingdom and 18.4% are from the United States. The highest number of nationalities are from United Kingdom.

#1 DRIVER

LEWIS HAMILTON



LEWIS HAMILTON

Lewis Hamilton is from the United Kingdom. He is the Top driver from all drivers mentioned in the Dataset. In the last 10 years, he has won 7 Championships and has won over 100 races. Lewis Hamilton participated in 16 seasons. His first season was in 2007 and his latest season was 2022.

Load the Data

- Pandas
- Print Info

F1 Data

- Latest season 2023

Participation

- Driver Name
- Season active

CLEANING AND PREPARATION

LOAD THE DATA

In order to prepare for the analysis, I first loaded the data and printed some of the information. I did this to understand the information in the columns. I also checked for any missing values and the shape of the data frame.

IS THE DATA RECENT?

I wrote a code that would tell me about the latest year in the dataset to understand how recent the data is. Most of my analysis is based on the top 10 aspect of the data, so my code reflects that information.

PARTICIPATION

I was also curious to how many seasons each driver participated in, so I wrote a code to filter the data based on driver name entered and season. This made it cleaner and easier to see that data in comparison to scrolling the actual dataset.

WHAT CAN YOU TELL ABOUT THE DATASET SO FAR?

```
Driver
Nationality
Seasons
Championships
Race_Entries
Race_Starts
Pole_Positions
Race_Wins
Podiums
Fastest_Laps
Points
Active
Championship Years
Decade
Pole_Rate
Start_Rate
Win_Rate
Podium_Rate
FastLap_Rate
Points_Per_Entry
Years_Active
Champion
```

By running some analysis of the data, I was able to see that the dataset is recent. The latest season in the data is 2023. This suggest that the results obtained in this analysis our up to date and accurate. Aside from this, the dataset list 22 different columns and 868 entries of information which is helpful to run a create analysis. I wish the creator of the dataset would have added age as a column. I think that could have been an interesting aspect to analysis as well.

What might your next steps to model the data be?

WHAT'S NEXT

I wanted to try a different type of chart to test out the data. I was trying to figure out how I could incorporate a Heatmap into this data. I was planning on using the library seaborn to create it. I wanted to show the distribution of driver nationalities across different seasons, with the number of drivers for each nationality in each season represented by a red color.

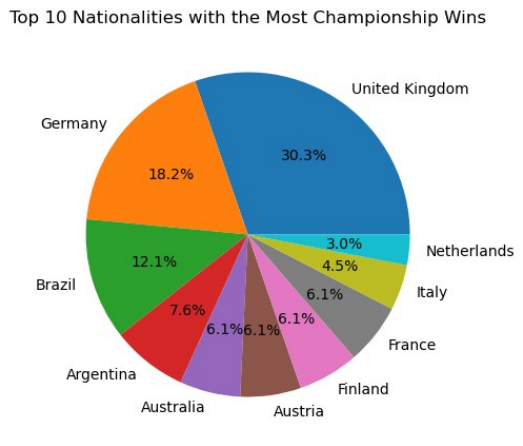
PATTERNS, TRENDS, CORRELATIONS?

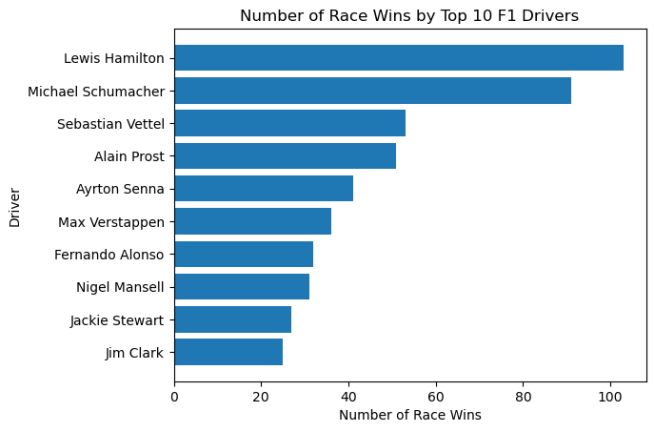
I created a scatterplot to see if there is a relationship between the top 10 driver race wins and their fastest laps. Based on the chart, there seems to be a weak positive correlation between race wins and fastest lap times for the top 10 drivers. This could be because there are some drivers with relatively low fastest lap times despite having a high number of race wins, and vice versa. There are also some outliers in the data, for example the driver with the highest number of race wins and the fastest lap time. Overall, more data and analysis would be needed to create a more accurate conclusion about the relationship between race wins and fastest lap times.

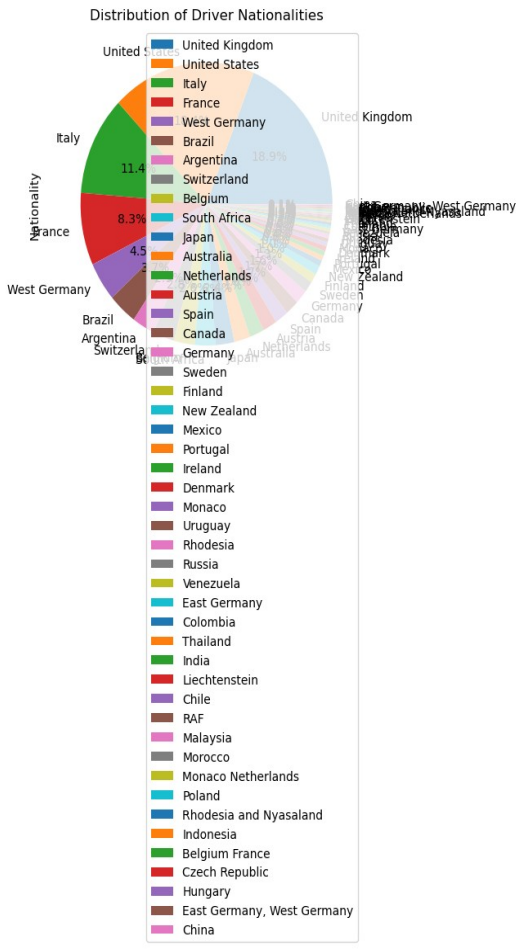
I also found it interesting that the United States were not one of the top nationalities on the championship pie chart. Even though 18.4% of all the drivers are from the United States, which is the second highest percentage on the chart.

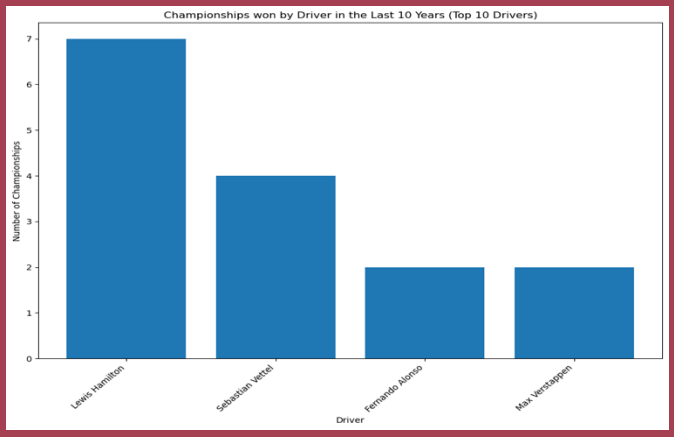
VISUALIZATION CHARTS

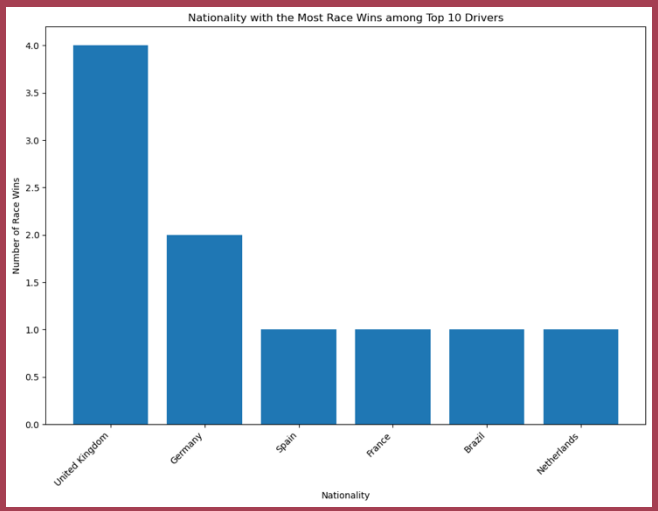
These are the charts that were used to get the information in this report. All Charts were coded using Python.

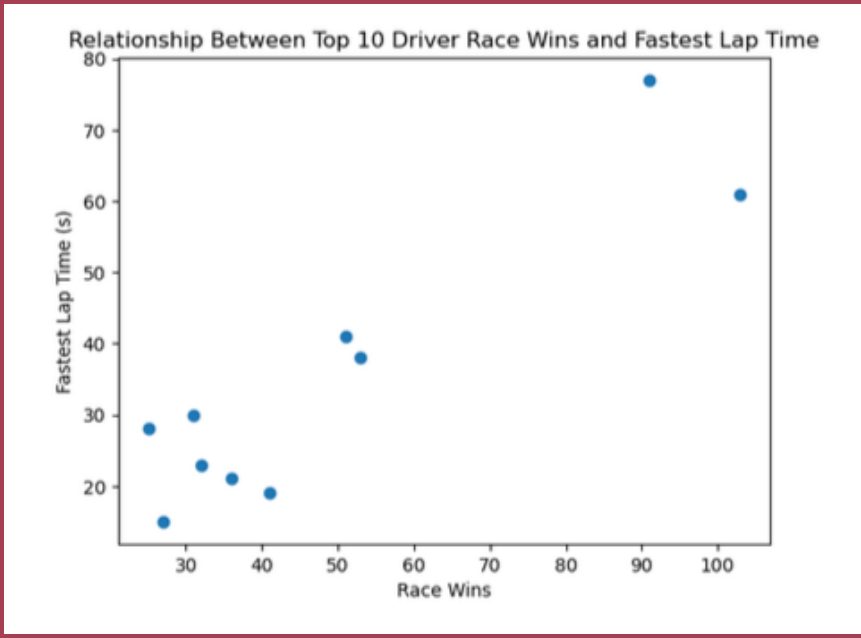











Relationship Between Top 10 Driver Race Wins and Fastest Lap Time
Fastest Lap Time (s)
Race Wins
80
70
60
50
40
30
20
30
40
50
60
70
80
90
100