Data Analysis Project

Formula 1 Race Data

Ting-Hua, Yeh
2022-05-14

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies
  - Download data on Kaggle website
  - Data wrangling
  - EDA with Visualizing data

- Summary of all results
  - Visualizing analysis result

Introduction

- Project background
  - I started to watch Formula 1 in 2018, the fast pace and the teamwork really fascinated me. The race result usually includes many factors, not only the cars and the drivers but also the pit crew performance and the characteristics of the circuits. As a new fan of racing, I want to explore the historical data to see if something is interesting to help me to know the sport better.

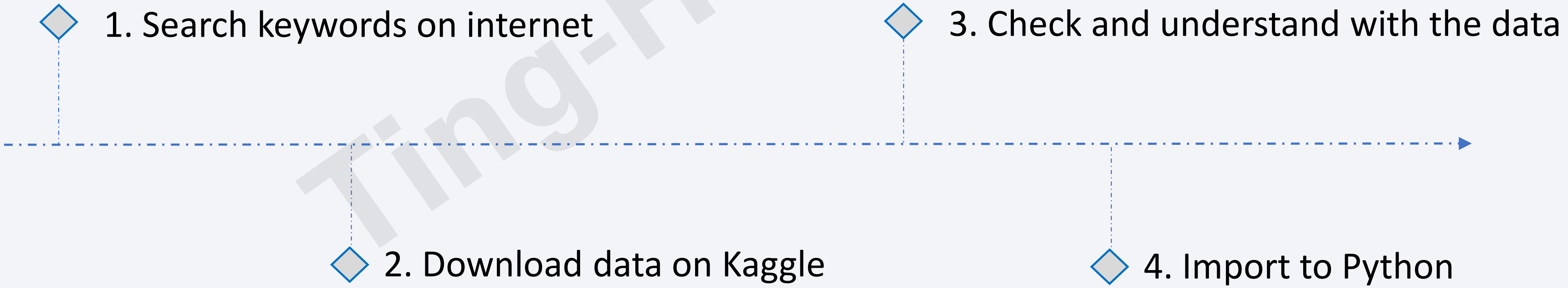
- Problems I want to find answers

  1. Which country has the most constructors of all the time?
  2. Which country has the most drivers of all the time?
  3. Which constructor has a better performance at pit lane in the 2018 to 2021 seasons?
  4. Some people said Red Bull’s car performed more successfully than other tier 1 constructors (Ferrari and Mercedes) at high altitude circuits, is that true in the 2018 to 2021 seasons?
  5. Does the better start position bring better racing results?

Methodology

Data Collection

- After searching relevant keywords on the internet, I found out there is a Formula 1 dataset on Kaggle website, and checked whether the information in the dataset are sufficient enough to answer my questions.



1. Search keywords on internet

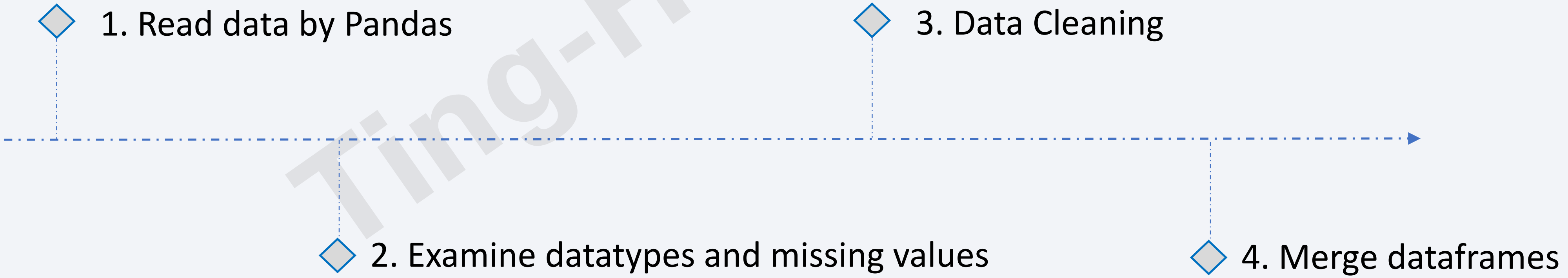
2. Download data on Kaggle

3. Check and understand with the data

4. Import to Python

Data Wrangling

- Used Pandas package to read the file, and examined the data type on each dataframe. Checked the missing value, determined to drop or to transform it. Finally, merge multiple dataframes for different questions.



1. Read data by Pandas

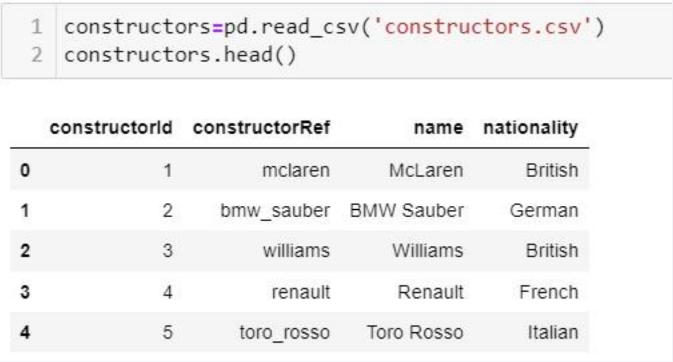
2. Examine datatypes and missing values

3. Data Cleaning

4. Merge dataframes

Data Wrangling

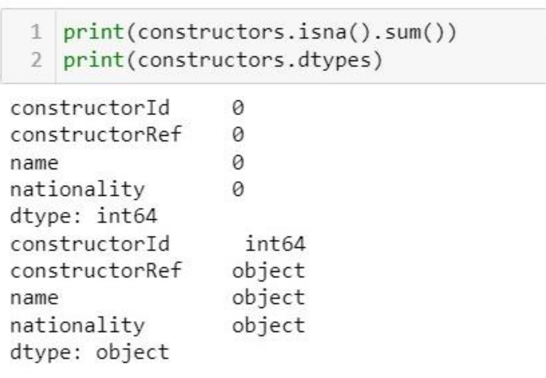
1. Read data by Pandas



```
constructors=pd.read_csv('constructors.csv')
constructors.head()
```

	constructorId	constructorRef	name	nationality
0	1	mclaren	McLaren	British
1	2	bmw_sauber	BMW Sauber	German
2	3	williams	Williams	British
3	4	renault	Renault	French
4	5	toro_rosso	Toro Rosso	Italian

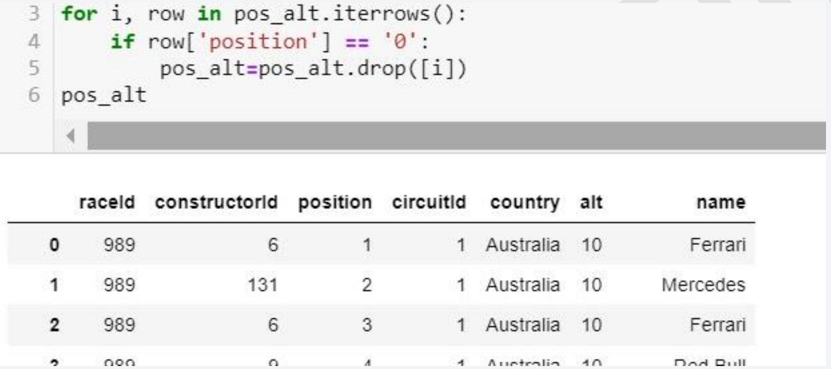
2. Examine datatypes and missing values



```
print(constructors.isna().sum())
print(constructors.dtypes)
```

```
constructorId     0
constructorRef    0
name              0
nationality       0
dtype: int64
constructorId      int64
constructorRef    object
name              object
nationality       object
dtype: object
```

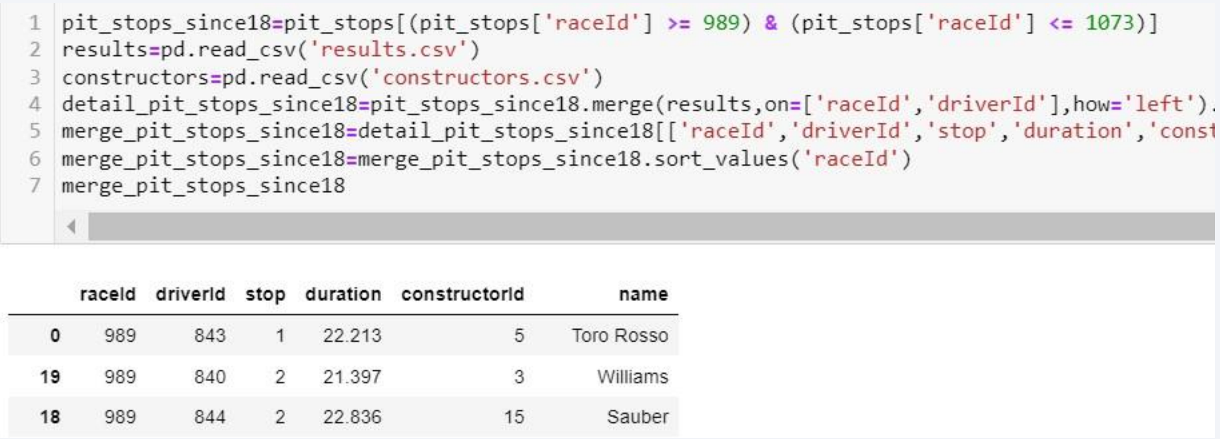
3. Data Cleaning



```
for i, row in pos_alt.iterrows():
    if row['position'] == '0':
        pos_alt=pos_alt.drop([i])
pos_alt
```

	raceId	constructorId	position	circuitId	country	alt	name
0	989	6	1	1	Australia	10	Ferrari
1	989	131	2	1	Australia	10	Mercedes
2	989	6	3	1	Australia	10	Ferrari
3	989	9	4	1	Australia	10	Red Bull

4. Merge dataframes



```
pit_stops_since18=pit_stops[(pit_stops['raceId'] >= 989) & (pit_stops['raceId'] <= 1073)]
results=pd.read_csv('results.csv')
constructors=pd.read_csv('constructors.csv')
detail_pit_stops_since18=pit_stops_since18.merge(results,on=['raceId','driverId'],how='left').
merge_pit_stops_since18=detail_pit_stops_since18[['raceId','driverId','stop','duration','const
merge_pit_stops_since18=merge_pit_stops_since18.sort_values('raceId')
merge_pit_stops_since18
```

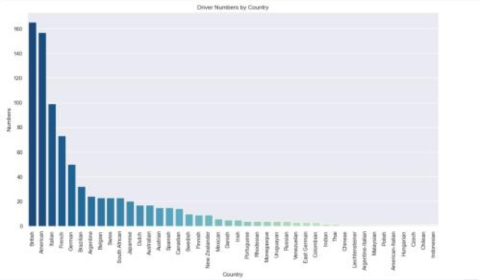
	raceId	driverId	stop	duration	constructorId	name
0	989	843	1	22.213	5	Toro Rosso
19	989	840	2	21.397	3	Williams
18	989	844	2	22.836	15	Sauber

EDA with Data Visualization

- For different questions, I chose a different plot to visualize it. Here are plots I used:

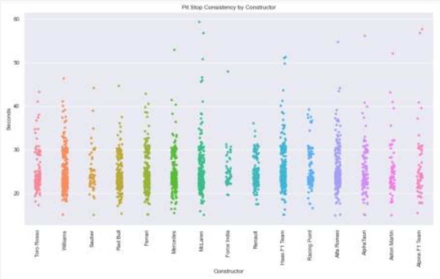
1. Barplot

For question 1 & 2



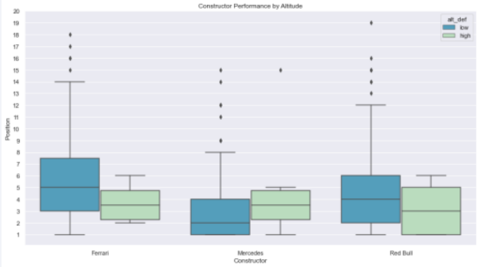
2. Stripplot

For question 3



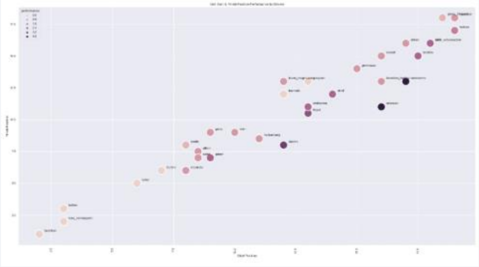
3. Boxplot

For question 4



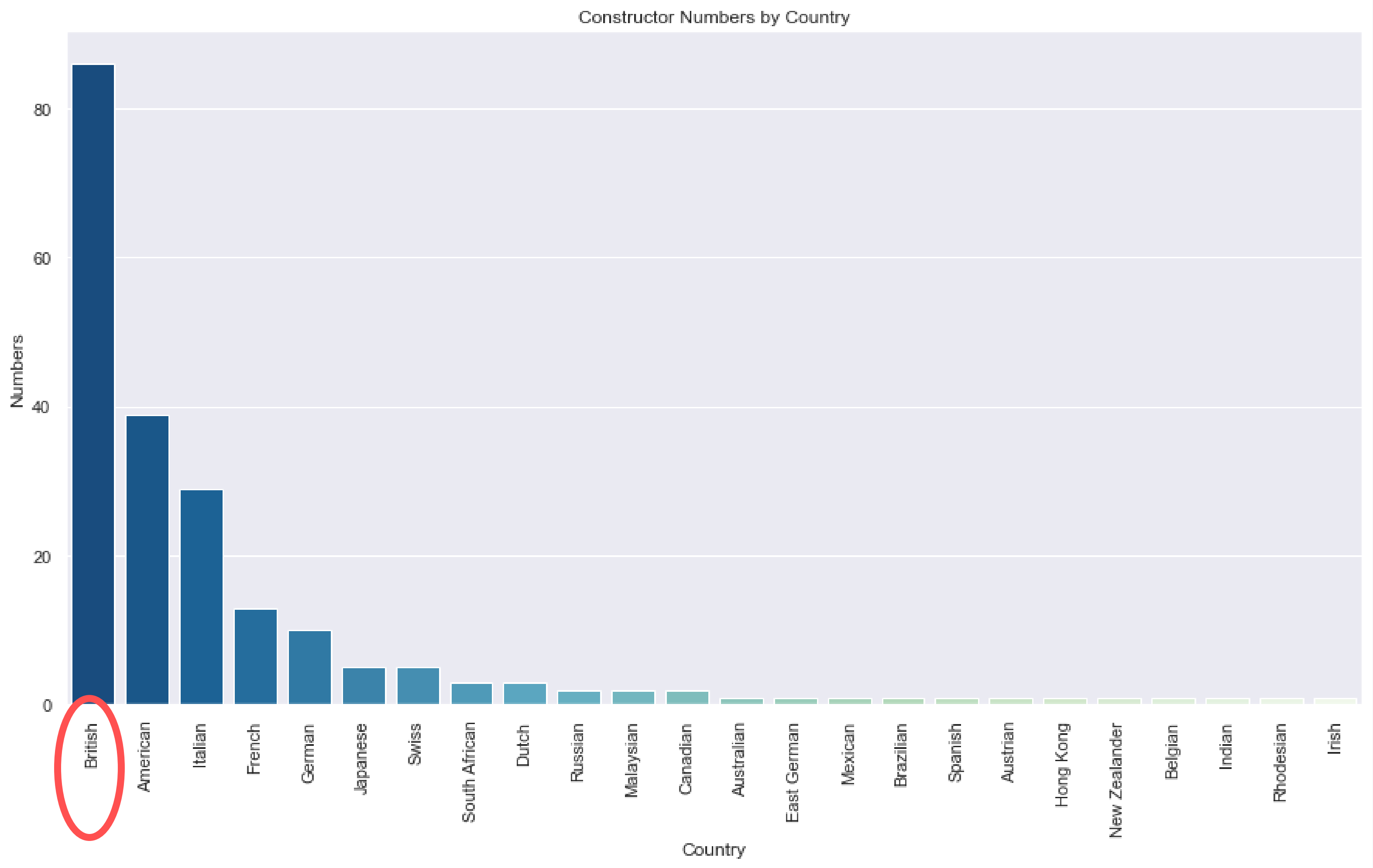
4. Scatterplot

For question 5



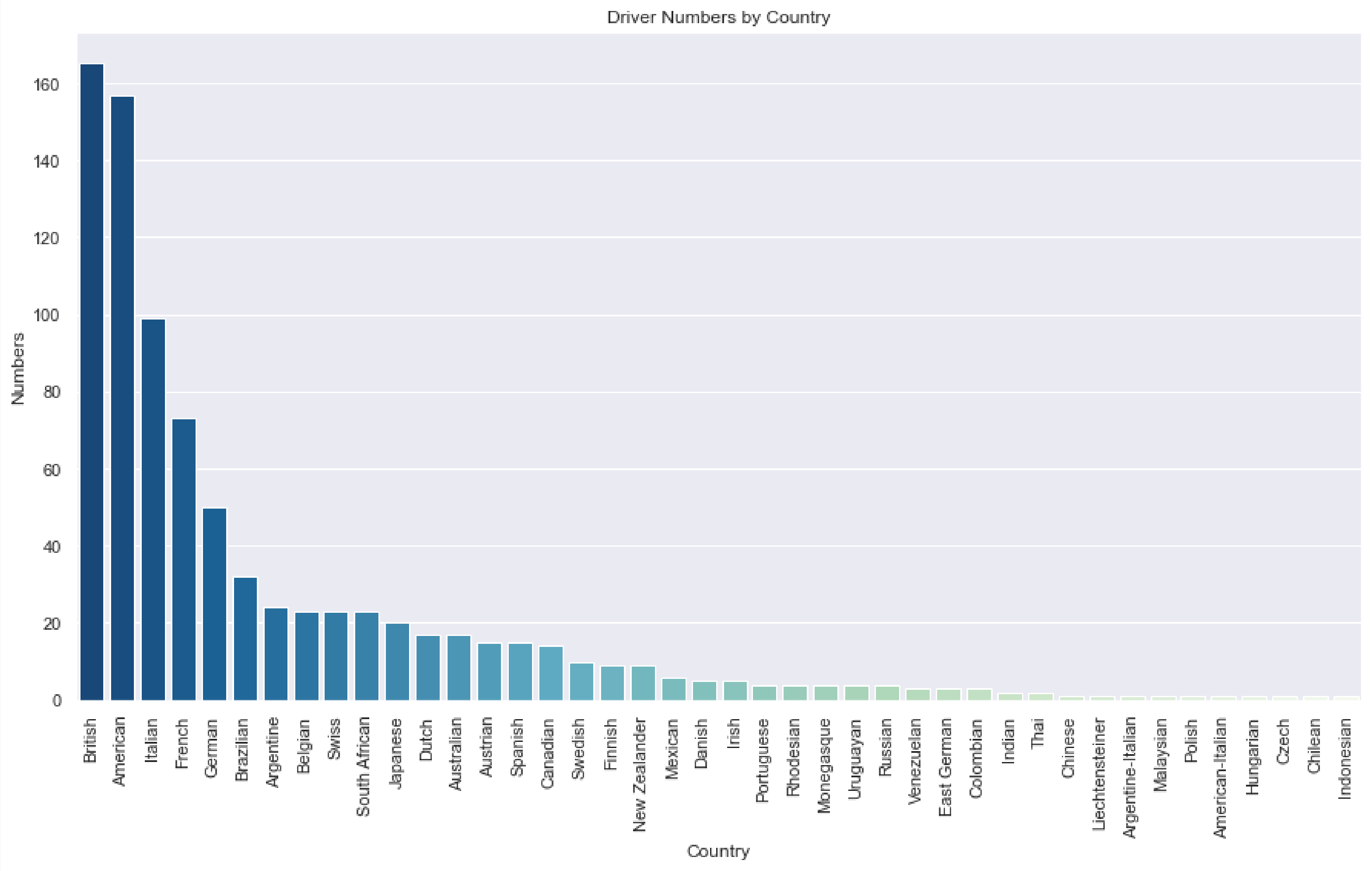
Results

Constructors Numbers by Country



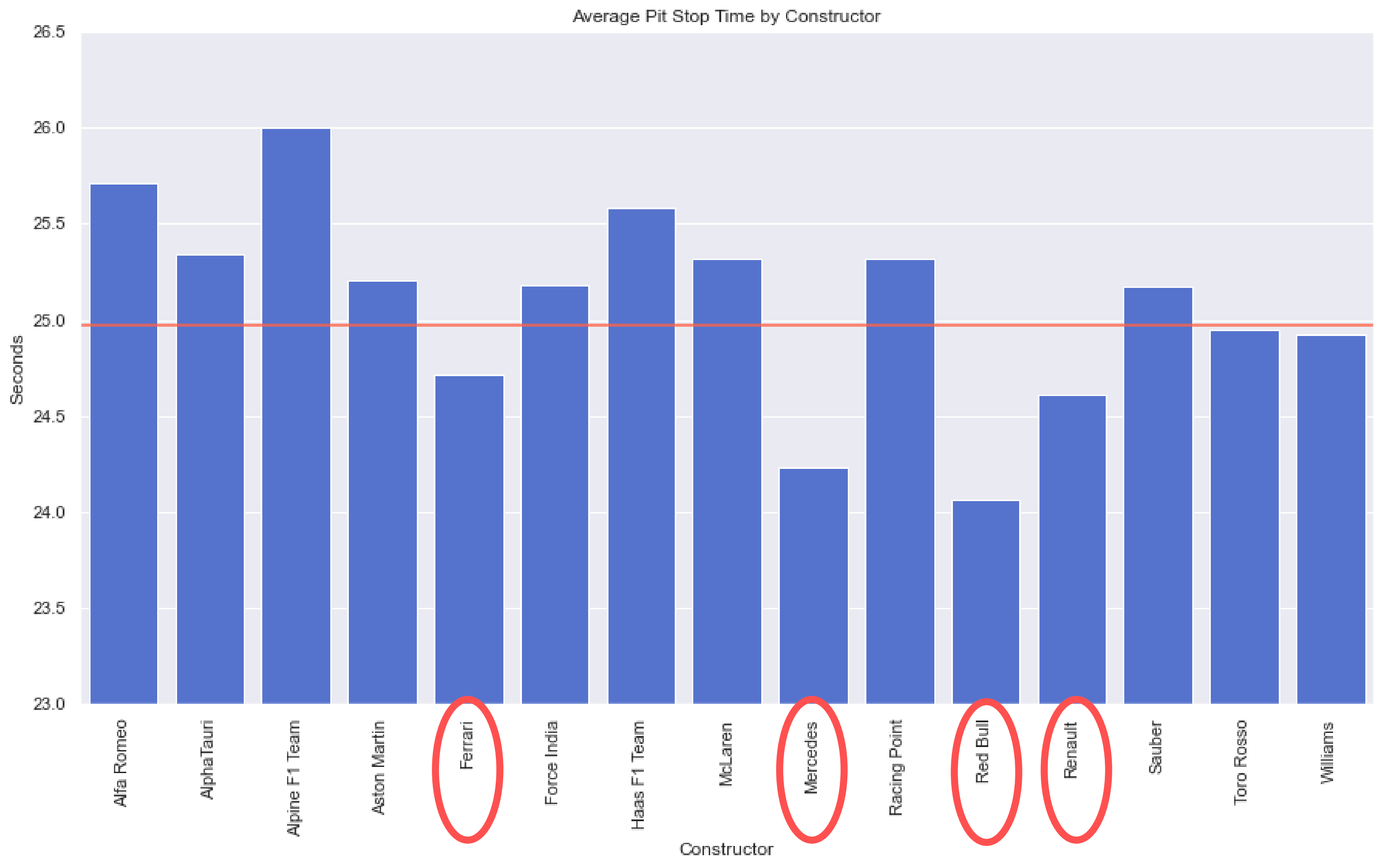
- After analyzing the constructors dataset, it shows that there were 86 British constructors up to now, more than 2 times compared to the second one – American, which has 39 constructors.

Driver Numbers by Country



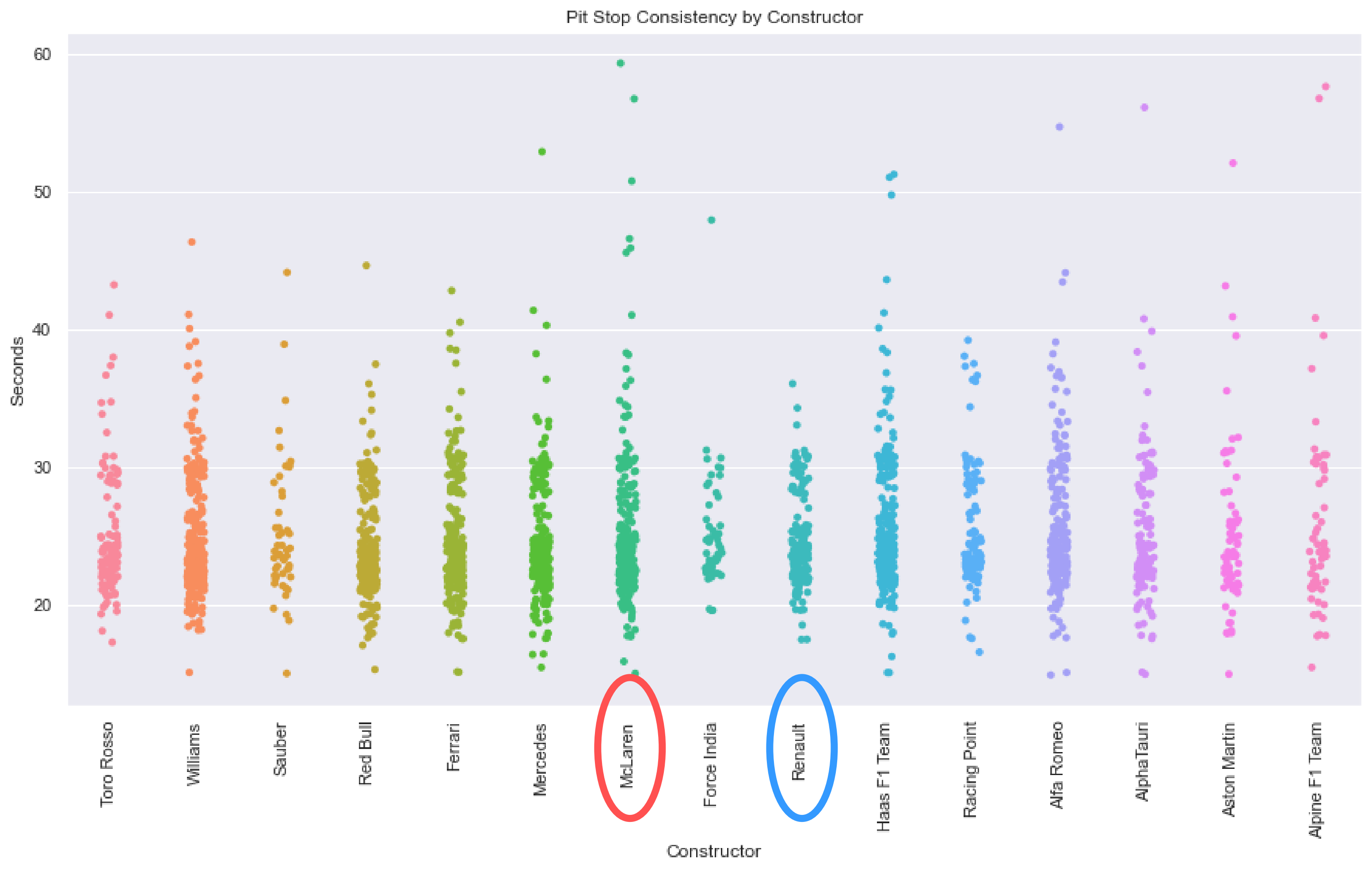
- Count the numbers of drivers by their nationality, it shows that the top 5 countries are the same as which has more constructors. Surprisingly, despite the huge difference on the constructors numbers, the numbers of American drivers are very close to the British drivers.

Pit Stop Performance by Constructor (2018 – 2021) -1



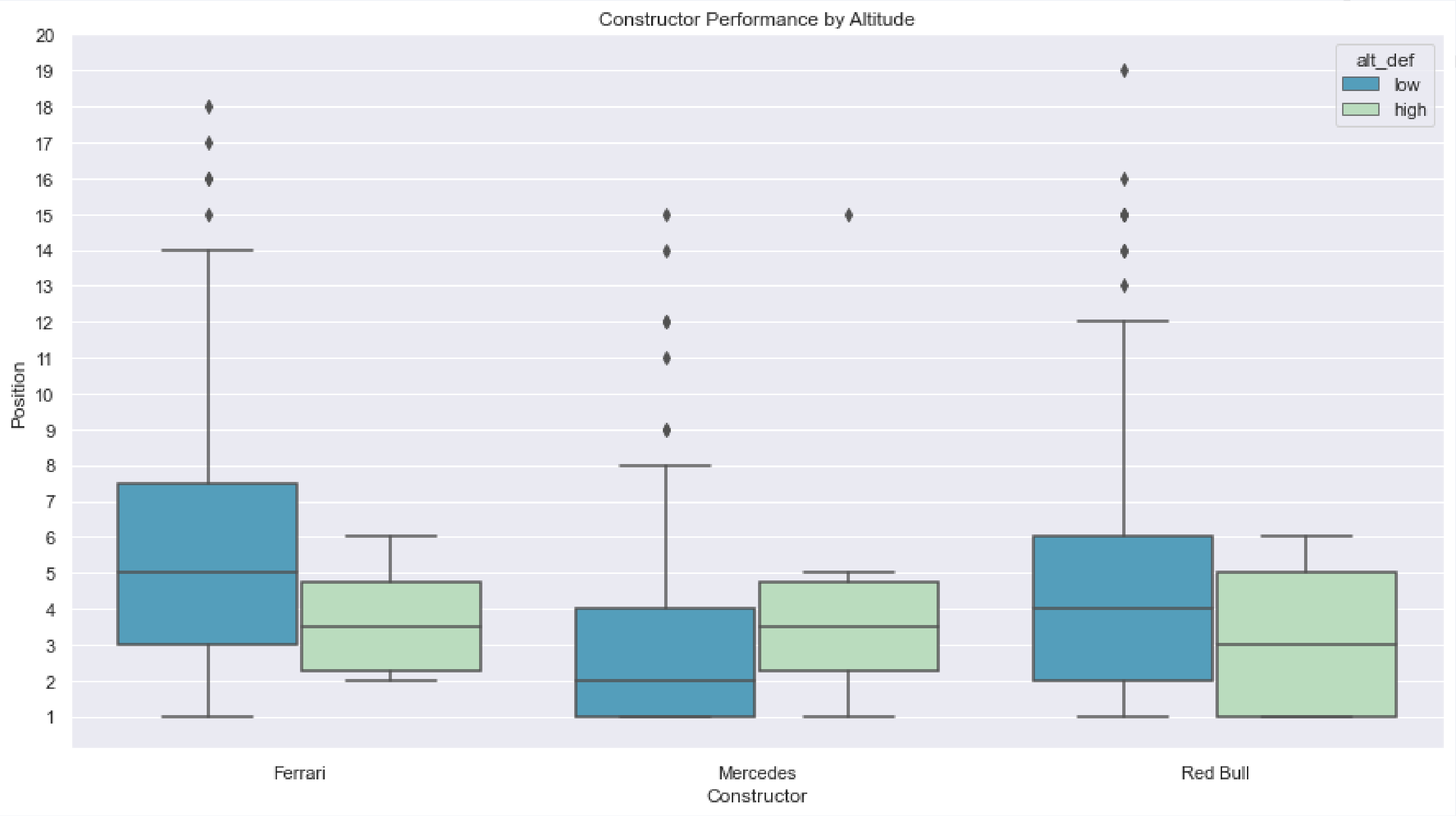
- Red Bull, Mercedes, Ferrari, and Renault have great performance at the pit stop, their average pit stop time were much better than the average performance (24.97 seconds) of all the constructors.

Pit Stop Performance by Constructor(2018 – 2021) -2



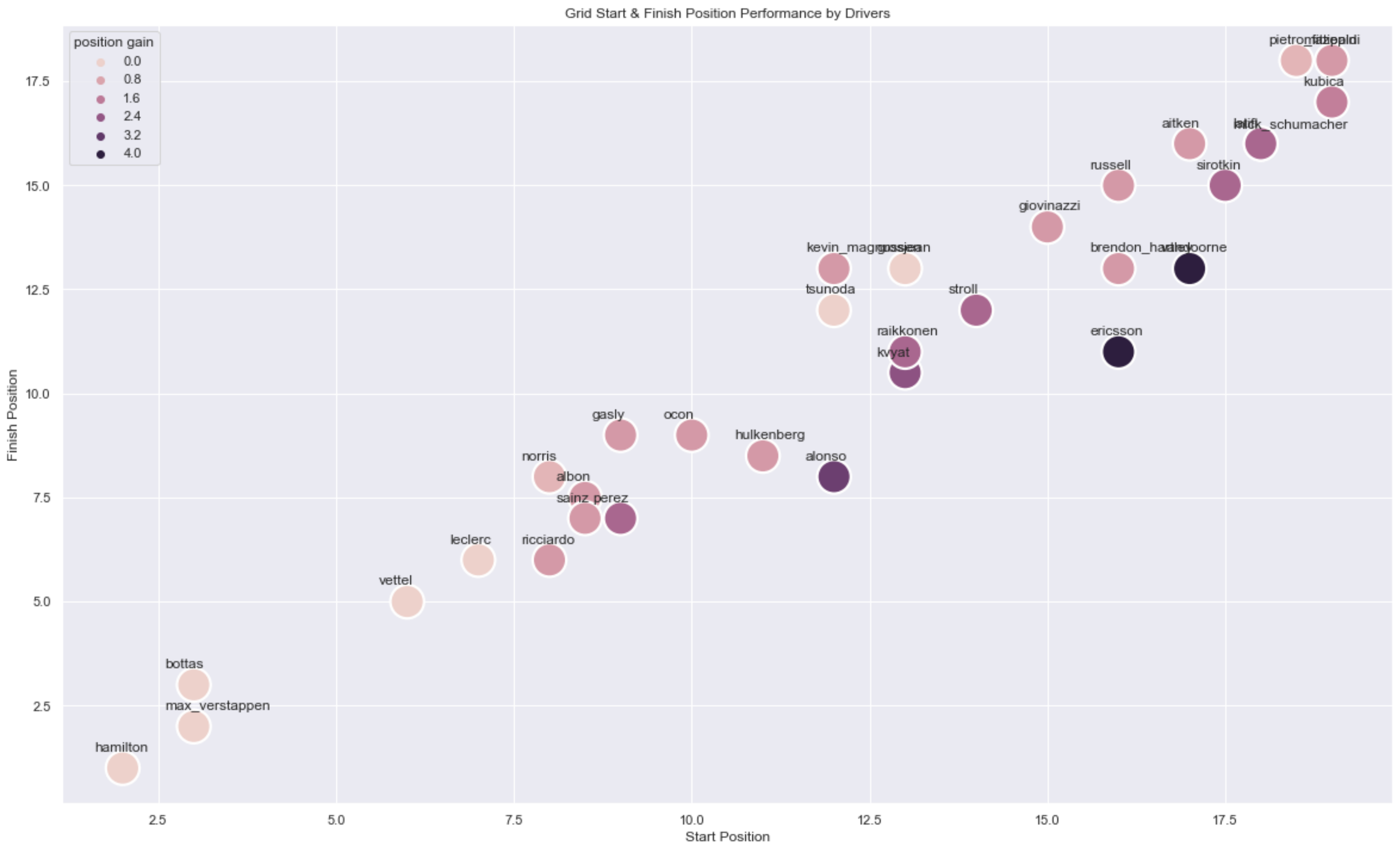
- After reviewing the distribution of every constructor’s pit performance, it shows that Renault is very consistent, they seldom had big mistakes at the pit stop. On the other hand, Mclaren seems unstable on the pit performance.

Constructor Performance by Altitude



- Comparing the performance between high altitude circuits and low altitude circuits. It shows Red Bull did perform better at high altitude circuits, but Ferrari also slightly progress at high altitude circuits. The interesting insight is that Mercedes really not performed well at high altitude circuits.

Start and Finish Performance by Drivers



- From the distribution, it shows that the start and the finish position did have a positive relationship. Vandoorne, Ericsson, and Alonso usually finished better than the start.

Conclusions

- England, USA, Italy, France, and Germany are the top 5 countries that have more constructors than others. The industry seems to prevail in England, there are 86 British constructors from the past to now, the number is 2.2 times than the 2nd country. Also, the top 5 nationality of the driver are identical to the constructor, but the gap between the 1st and the 2nd country is not huge.
- Red Bull, Mercedes, Ferrari, and Renault performance well at the pit stop, and Renault's pit crew did a very consistent job. Although Mclaren didn't stable at the pit stop, they only not be in the top 5 constructors in the 2018 season. So, maybe this is what I can explore more in the future, maybe pit stop performance doesn't have a big impact on the race result? For example, if you had a bad pit stop performance during the season, but your cars were not retired as much as others did.
- Red Bull did perform better at high altitude circuits than other tier 1 constructors, and also better than itself at low altitude circuits. But the interesting difference is on Mercedes, their performance really not well at the high altitude circuits.
- The start position did have a positive relationship with the finish position, and most of the time the position change is within 5 places, this made qualifying important. The position around 10th to 17th seems to have more opportunities to swap positions.

Thank You