Speeding to Victory: A Predictive Modeling Approach in Formula 1

Meida Rahma Al Kariim

12/2023

OUTLINE

01 **BACKGROUND**
AND GOALS

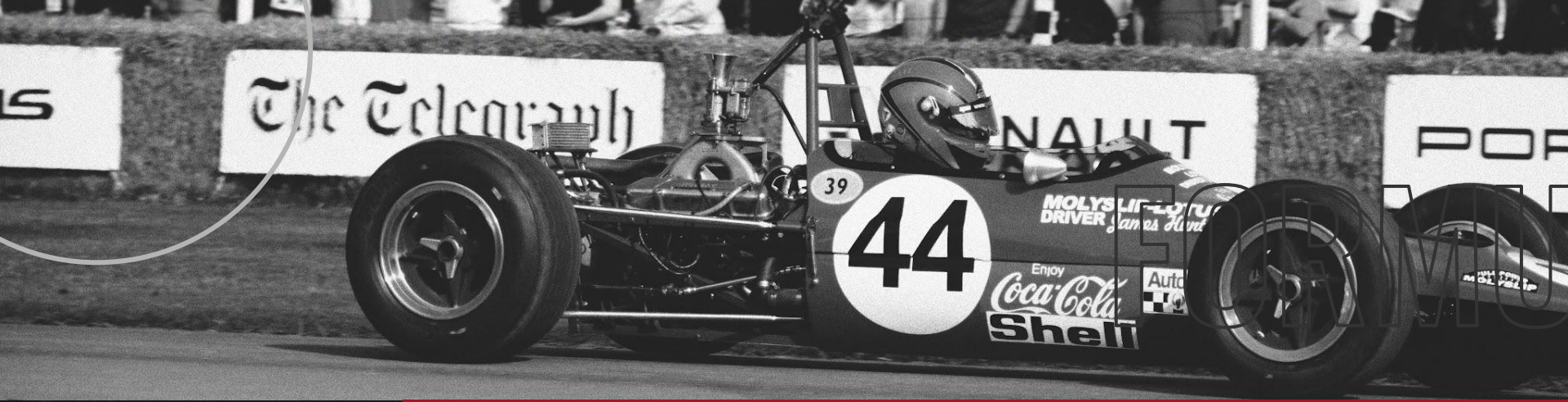
02 **DATASET**
AND DATA PREPARATION

03 **EDA**
EXPLORATORY DATA ANALYSIS

04 **MODELING**
AND EVALUATION

05 **CROSS VALIDATION**
AND EVALUATION

06 **CONCLUSION**
AND RECOMENDATION



12/2023

01 BACKGROUND

Formula 1 began in 1950 as a recognized world championship series by the International Automobile Federation (FIA).

BACKGROUND

CARS

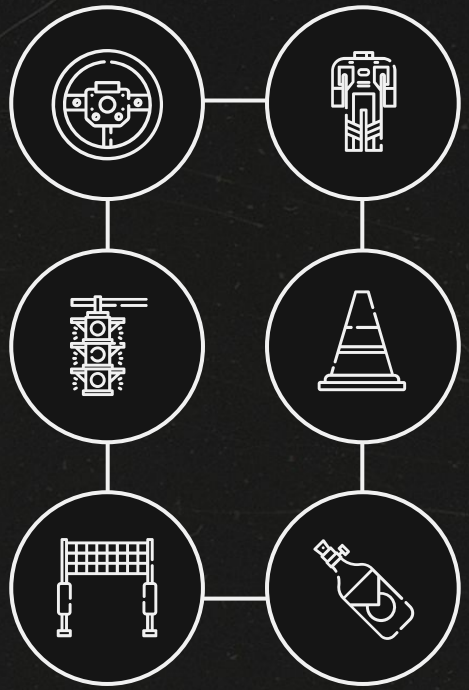
F1 cars are highly sophisticated and designed to achieve **high speeds.**

RACES

The F1 season consists of a series of Grand Prix (GP) held at **various circuits** worldwide.

RULE CHANGES

F1 rules and regulations may **change each season.**



TEAMS

Such as Mercedes, Ferrari, Red Bull Racing, and others, **compete against each other** in the championship.

POINTS

Drivers and teams earn **points based on race** and qualifying results

WORLD CHAMPIONS

After all the **points are collected,** Formula 1 World Championship is obtained

FORMULA 1

OBJECTIVES TO WIN

PROVIDE

Provide business insight related to the data.

PROVIDE

Provide business recommendations related to this matter.

PREDICT

Predict the best model to determine who is the next F1 World Championship winner.

02 DATASET

AND DATA PREPARATION

DATASET

ABOUT

Data about Formula 1 championships from 1950 -2023.

CLEANED

Cleaned dataset obtained 445.698 rows and 20 columns

HAVE

Have 16 numerical columns and 4 categorical columns

DATA DICTIONARY

For more info.

DATA PREPARATION

The dataset consists of all information on the Formula 1 races, drivers, constructors, result, stats, championships from 1950 till the latest 2023 season.

MISSING VALUE

NO	FEATURE	MISSING VALUE	PERCENTAGE
1	timetaken_in_millisec	2,172,879	61.42%
2	max_speed	2,140,657	60.51%
3	fastestLap	2,140,657	60.51%
4	rank	2,105,375	59.51%

DUPLICATED VALUE → 0

TARGET → DRIVER_NAME (CATEGORICAL COLUMN)

We have chosen our target variable, 'driver_name,' which falls under the classification category. Therefore, the machine learning approach we will employ is supervised learning with a focus on classification.

For more info:
KAGGLE

12/2023

12/2023

FORMULA 1

03

EXPLORATORY DATA ANALYSIS

Univariate Analysis, Bivariate Analysis, Deep Dive EDA

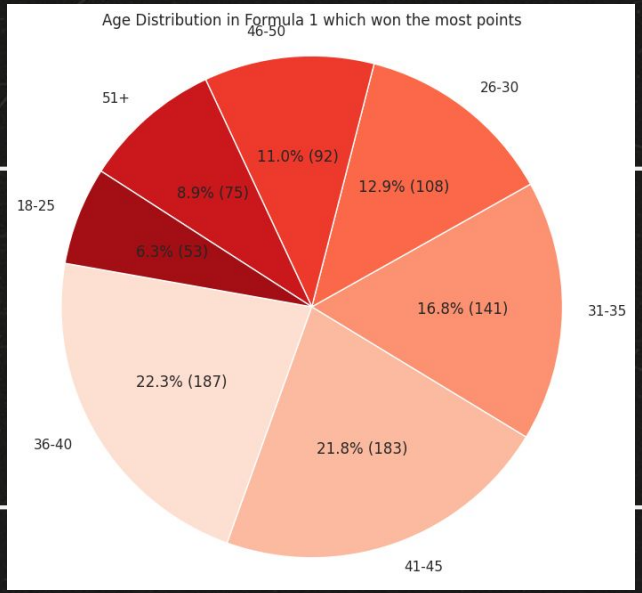
DISTRIBUTION OF DRIVER AGES IN FORMULA 1

22.3%

The age range 38-40 has 187 points.

11.0%

The age range 26-30 has 108 points.



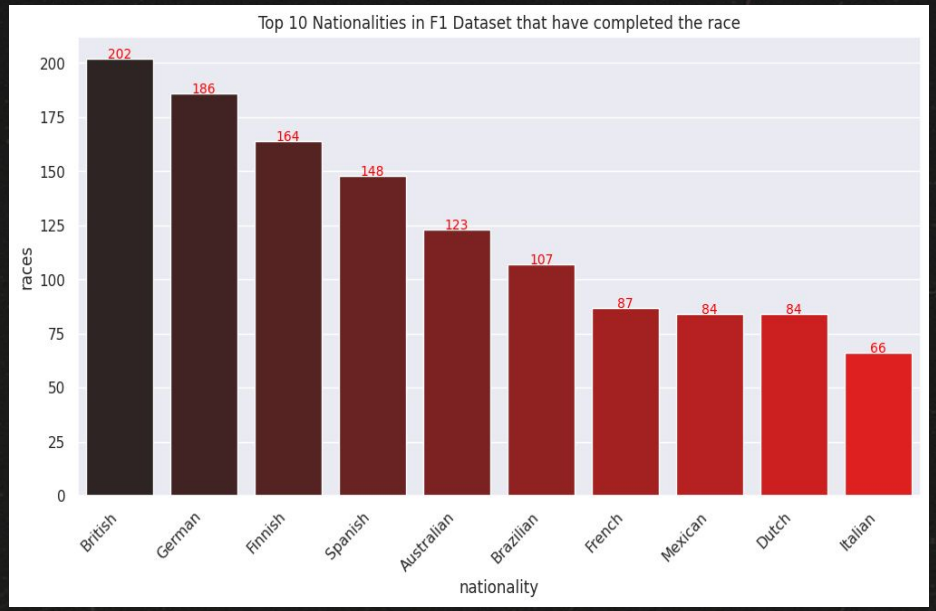
21.8%

The age range 41-45 has 183 points.

16.8%

The age range 31-35 has 141 points.

Top 10 Nationalities in F1

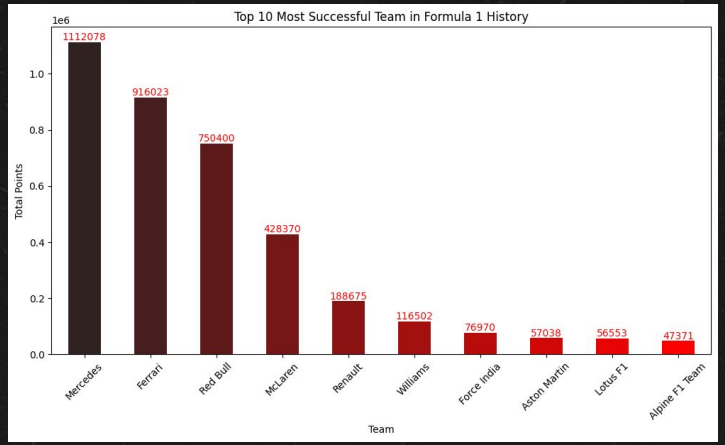


This Histogram based on the nationalities in F1 that have completed the race.

IMPACT FOR TEAM

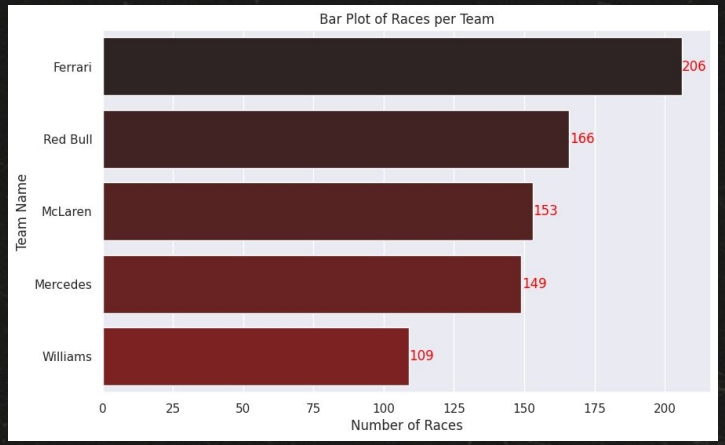
The team can choose the best drivers with the greatest points achieved based on their age and nationality. .

Top 10 Most Successful Team in Formula 1 History



The results of the stem diagram aside from the calculation of the number of points obtained from each team.

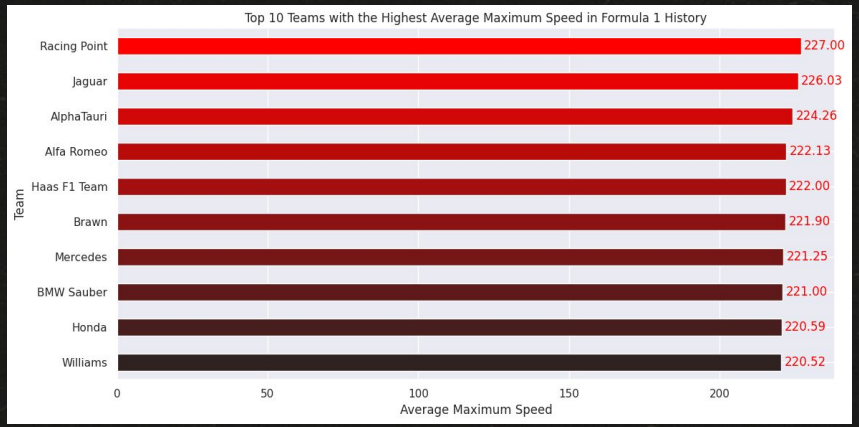
Bar Plot of Races per Team



The bar plot is based on the team and the number of races they followed.

12/2023

Top 10 Teams with the Highest Average Maximum Speed in Formula 1 History



The bar plot is based on the team and the highest average achieved by the car they made themselves.

IMPACT TO DRIVER

The driver can choose the team based on the highest number of points obtained, the number of races followed, and the maximum average car that the team has made. Because their cars are made by the team themselves in accordance with the rules and regulations set by the FIA.

12/2023

04

MODELING

LogisticRegression, KNeighborsClassifier, GaussianNB, and SGDClassifier

MODELING On Base

MODEL	TRAINING DATA	TEST DATA
LogisticRegression	99.00577742876374	99.00044873233117
SGDClassifier	67.65481265425174	67.6553735696657
KNeighborsClassifier	99.97616109490689	99.98092887592551
GaussianNB	97.64976441552614	97.67332286291227

The results of data training and data test are not too different, it can be said to be a suitable model. Based on that the best model for base in this dataset is KNeighborsClassifier.

MODELING On IQR

MODEL	TRAINING DATA	TEST DATA
LogisticRegression	99.08671390782182	0.6708548350908683
SGDClassifier	63.77555946013786	0.03814224814897914
KNeighborsClassifier	99.96524492672472	0.869418891631142
GaussianNB	98.13481106756001	0.8896118465335427

The results of the data train become different when tried on the data test. All models can be said to be overfitting because the accuracy results given are lower compared to data train.

So we choose the base data for cross validation.

MOTORCYCLE

05 CROSS VALIDATION

On base data

12/2023

CROSS VALIDATION

FORMU

MODEL	SCORES
LogisticRegression	0.11342743828078232
SGDClassifier	0.22933253219871044
KNeighborsClassifier	0.9999935894353895

Cross-validation are used for check if the model are stable or not with the fold = 5. Overall, **KNeighborsClassifier** has the best performance based on the weighted average accuracy, with avalue of **0.999.**

06

CONCLUSION

And recommendation

CONCLUSION

- Dataset Formula 1 consists of 3,537,627 rows and 37 columns. The dataset has a missing value and outlier in several features. To change the non-numeric column to numeric, I labeled encoding for features and targets. After the Dataset Preprocessing data has 20 features. I divided the dataset into 80% of training data and 20% of test data, with training data divided into training and validation data.

- I chose Logisticsregression, Kneighborsclassifier, Gaussiannb, and SGDClassifier as a model to predict the Winner of the Race based on outlier data that is not handled and the outlier data is handled. Add a standard scaler to improve the performance of the model with the final result. In data that is not handled outliers, the results of data train and data tests are not too different, it can be said to be a suitable model.

- Whereas in the data handled by the outlier, the results of the data trains become different when tried on the data test. All models can be Said to be overfitting because the accuracy results given are lower comparated to data train. So we choose a data base for cross validation.

- From this study we can conclude that the performance of the KNN model plays a very good role for train and test data on the base and IQR for hypertuning. Based on this model with the fold = 5 KNeighborsClassifier has the best performance based on the weighted average accuracy, with avalue of 0.9999935

BUSINESS RECOMMENDATION

Based on the results of the analysis obtained:

1. Drivers can choose the team based on the number of points obtained and the maximum average of a car. Because their cars were made by the teams themselves in accordance with the rules and regulations set by the FIA.
2. As for the teams, if you want to choose the best drivers based on points can be seen based on their age and nationality. Although it does not rule out the possibility of an influence from other factors.
3. Sponsors can find the next sponsor target by using this prediction model, so that the sponsors given can be on target.
4. The audience can take part in gambling sites using this prediction model in order to maximize profits.
5. As a model performance test, it would be better if the model was applied in the next race match.

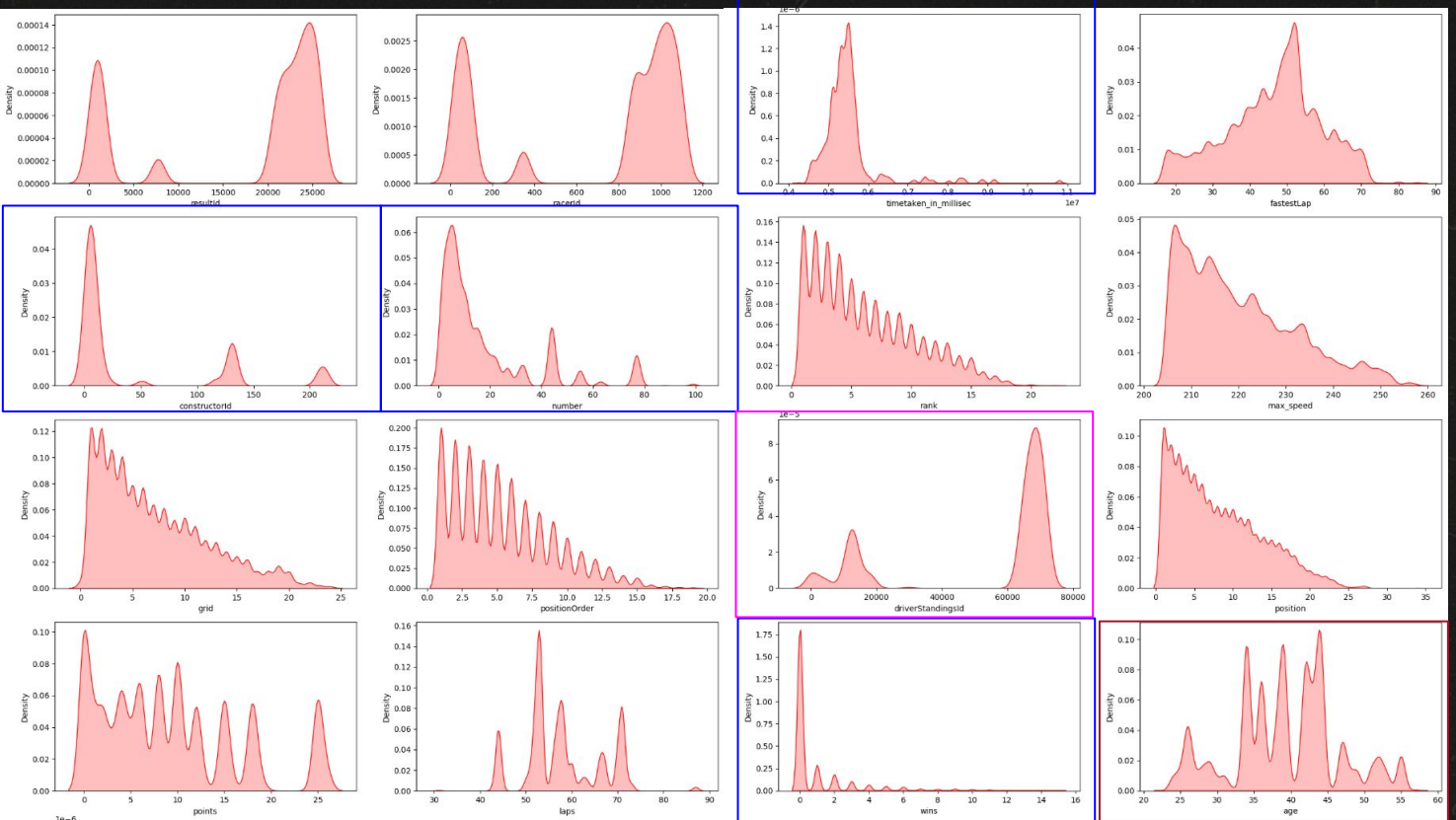
12/2023

THANKS!

Do you have any questions?

meidarahma1105@gmail.com

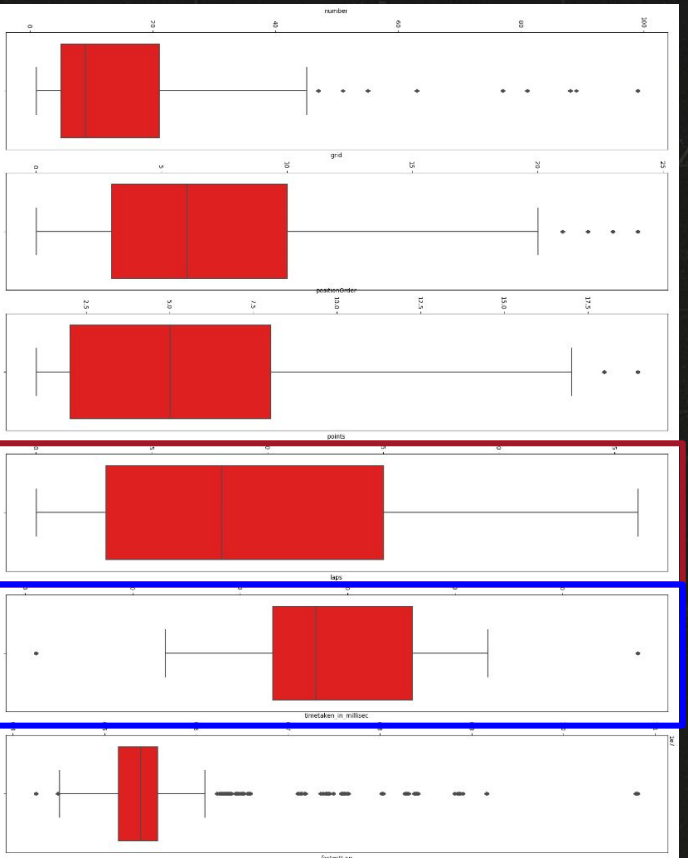
KDE PLOT



Observation:

1. 'Age' most at the age of 30 to 45 (which is seen in the red outline box.).
2. " Constructorid ", “Number ", Timetaken ", “Wins” indicated Right Skew. (which is seen in the blue outline box).
3. Whereas 'drivertanding' indicated Left Skew. (which is seen in the purple outline box.)

Boxplot To Detect Outliers



Almost all have an outlier except the 'Points' and 'Age' columns (which is seen in the red outline box). The outlier is very diverse, choosing to save it.

There's nine features have extreme outliers and feature 'laps' that has outlier in both side (which is seen in the blue outline box).

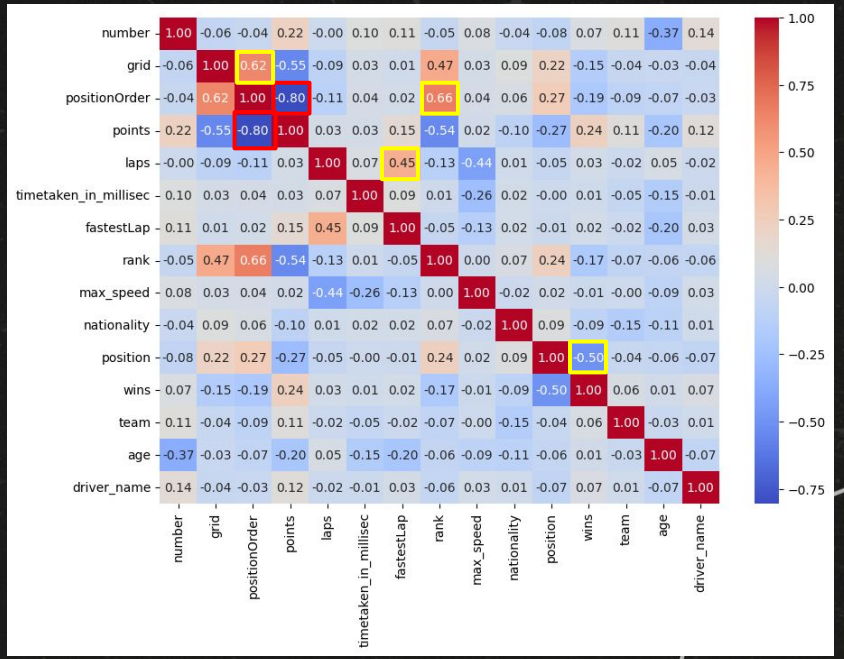
We will call **Baseline**, for dataframes that we do not do outlier and **IQR** handling for the dataframe that we do handling.

MULTICOLLINEARITY

BASELINE

Feature correlates each other, in line with the regulations in F1 (See shape with yellow outline.).

- The higher the grid, the higher the positionOrder
- The more laps, the faster the car droves.
- The higher the position means the higher the Wins.
- The higher the positionOrder, the higher the rank.
- etc.



HANDLING:

In features that are highly correlated, it can be deleted columns that do not really affect the target data.

On BASELINE:

- The ‘positionOrder’ column is very correlated with ‘position’ so we have to bring down one of them. Choose to maintain the ‘positionOrder’ . Therefore, drop ‘position’ .

MULTICOLLINEARITY

IQR

Feature correlates each other, in line with the regulations in F1 (See shape with yellow outline.).

- The higher the positonOrder, the higher the points
- The higher the rank, the higher the points.
- The faster the car, the more laps are completed.
- The higher the position, the higher the rank.
- etc.

HANDLING:

In features that are highly correlated, it can be deleted columns that do not really affect the target data.

On IQR:

- The 'Max_speed' column is very correlated with 'timetich_in_millisecond' so we have to bring down one of them. Choose to maintain the positionorder. Therefore, drop `Timetich_in_millisecond.