# THE FORMULA FOR VICTORY:

exploring F1 statistics


# University Of Mumbai

## Department Of Statistics

**Head Of Department: Dr. Santosh P. Gite**

**Mentor: Dr. Prof. Alok D. Dabade**

Group Members:

Jyoti Jagdale (13)

Yisha Joshi (14)

Swarangi Jagdale (32)

Aishwarya Parab (37)

# INTRODUCTION:

## WHAT IS FORMULA 1?

Key aspects of Formula 1:

- Grand Prix
- Circuit
- Constructors
- Drivers
- Qualifying
- Racing
- Podium
- Points system
- Pit Stops
- Safety car

# INTRODUCTION:

**WHY did we choose formula 1 racing data for analysis?**

- EXTENSIVE AVALAIBILITY OF DATA gives EXCITING AND USEFUL INSIGHTS CATERING TO A LARGE INTERNATIONAL FANBASE

- ANALYTICAL TECHNIQUES POSSESSING WIDE RANGE OF APPLICATIONS BEYOND THE SPORT

# OBJECTIVES:

Our project aims to look at the insights of different formula 1 aspects using statistical methods:

- To build a winner prediction model for the current 2024 season of formula 1
- To evaluate driver’s performance index and study various drivers metrics.
- Identifying trends and patterns in the data of drivers, races and constructors with respect to various social factors
- To study casualty data and observe various trends and patterns
- Poisson is distribution for testing a hypothesis for casualty analysis by ump test
- To study the correlation between driver age v/s driver points and driver points v/s driver position.

# Data Pre-Processing

**Data Collection:** Secondary data from Formula1 official website, ergast.com, Kaggle.

- Datasets:
  - Races & results
  - Qualifying
  - Weather information
  - Driver & constructor standings

**Data Consolidation/merging:**

We merged all these dataset into a single dataset which will be used for predictive models.

The merged dataset has Variables:

- Year, race, Circuit, Weather: Conditions (warm, cloudy, dry, wet, cold), Driver’s Age, wins, position qualifying time, nationality and the constructor
- Champion: Binary variable (1 for max points driver, 0 for others)

**Data Cleaning:**

- Time Span: 1994 to 2023 (excluding 1950-1993 due to missing qualifying info)
- Data Preparation: Filtered out null values from the 'podium' column

# Data Pre-Processing

## Data

| | raceId | year | round | weather | weather_warm | weather_cold | weather_dry | weather_wet | weather_cloudy | resultId | ... | constructor_renault | constructor_sauber | constructor_toro_rosso | constructor_toyota | constructor_williams |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 257 | 1994 | 1 | Sunny | True | False | False | False | False | 5139 | ... | False | False | False | False | False |
| **124** | 268 | 1994 | 12 | Sunny | True | False | False | False | False | 5444 | ... | False | False | False | False | False |
| **125** | 268 | 1994 | 12 | Sunny | True | False | False | False | False | 5443 | ... | False | False | False | False | True |
| **126** | 268 | 1994 | 12 | Sunny | True | False | False | False | False | 5448 | ... | False | False | False | False | True |
| **127** | 268 | 1994 | 12 | Sunny | True | False | False | False | False | 5452 | ... | False | False | False | False | False |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **5063** | 1102 | 2023 | 5 | Partly cloudy | False | False | False | False | True | 25935 | ... | False | False | False | False | False |
| **5064** | 1102 | 2023 | 5 | Partly cloudy | False | False | False | False | True | 25933 | ... | False | False | False | False | False |
| **5065** | 1102 | 2023 | 5 | Partly cloudy | False | False | False | False | True | 25929 | ... | False | False | False | False | False |
| **5058** | 1101 | 2023 | 4 | Partly cloudy | False | False | False | False | True | 25916 | ... | False | False | False | False | False |
| **5165** | 1110 | 2023 | 12 | Partly cloudy, with a rain interval | False | False | False | True | False | 26081 | ... | False | False | False | False | False |

5166 rows × 78 columns

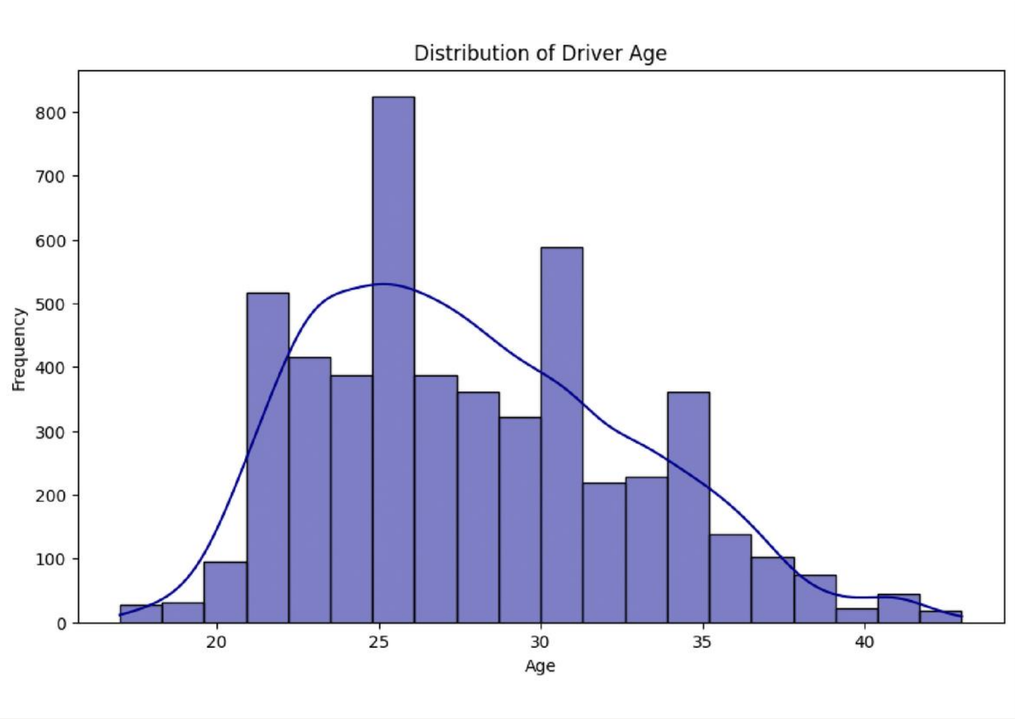
## Additional Data:

- Overtakes and Overtaken data of each driver  (1994-2020): Used to calculate performance index
- Casualty(accidents) data with virtual safety cars

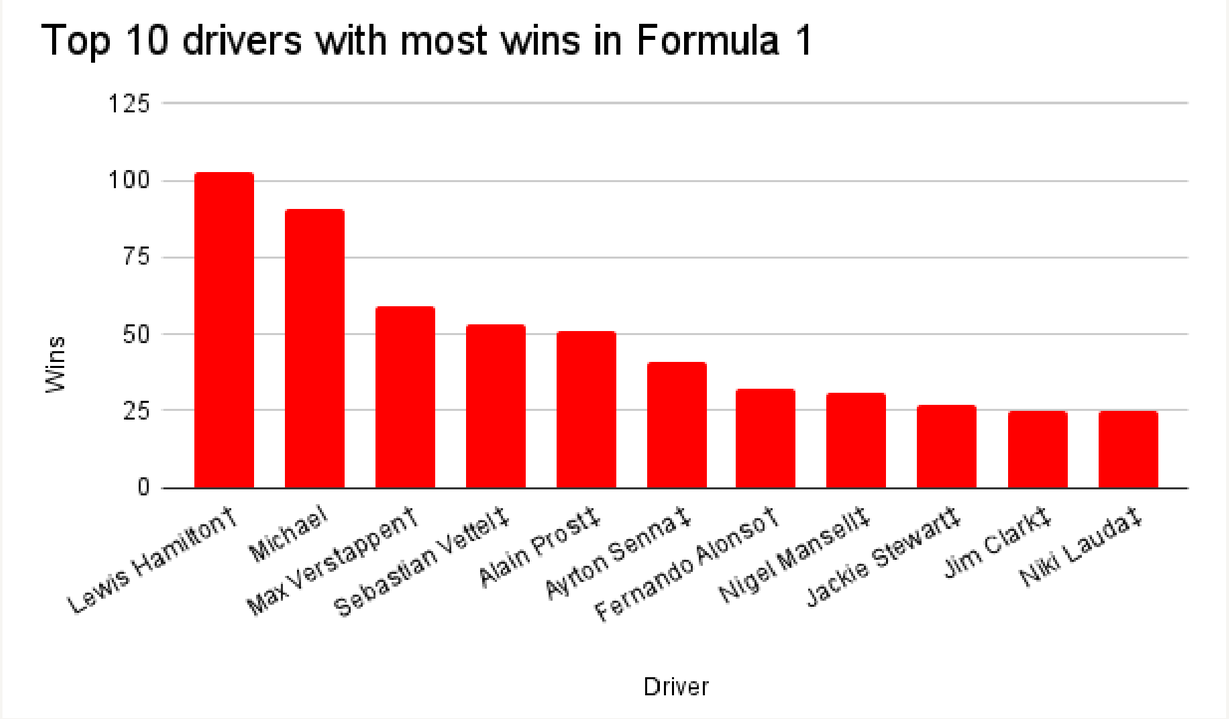
# DATA VISUALIZATION:

A. DRIVER METRICS

- age distribution of drivers



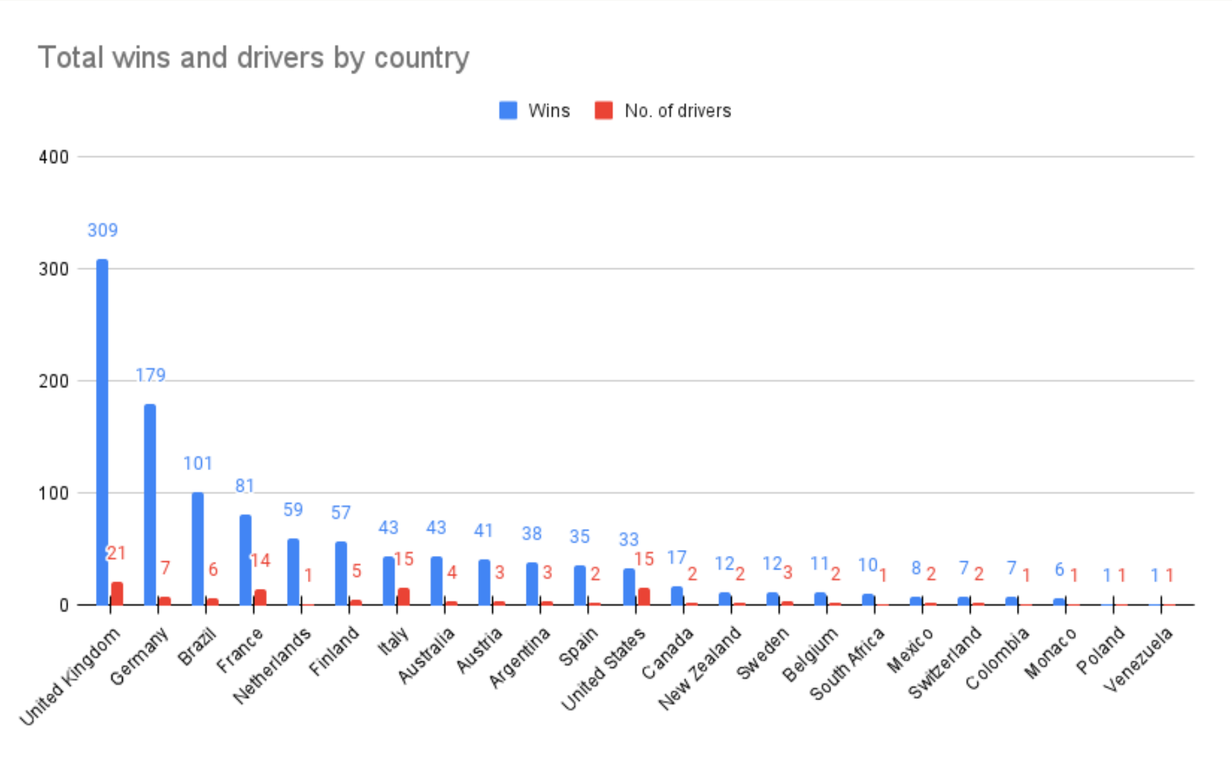

- Top 10 Formula 1 drivers with most wins in races



# DATA VISUALIZATION:

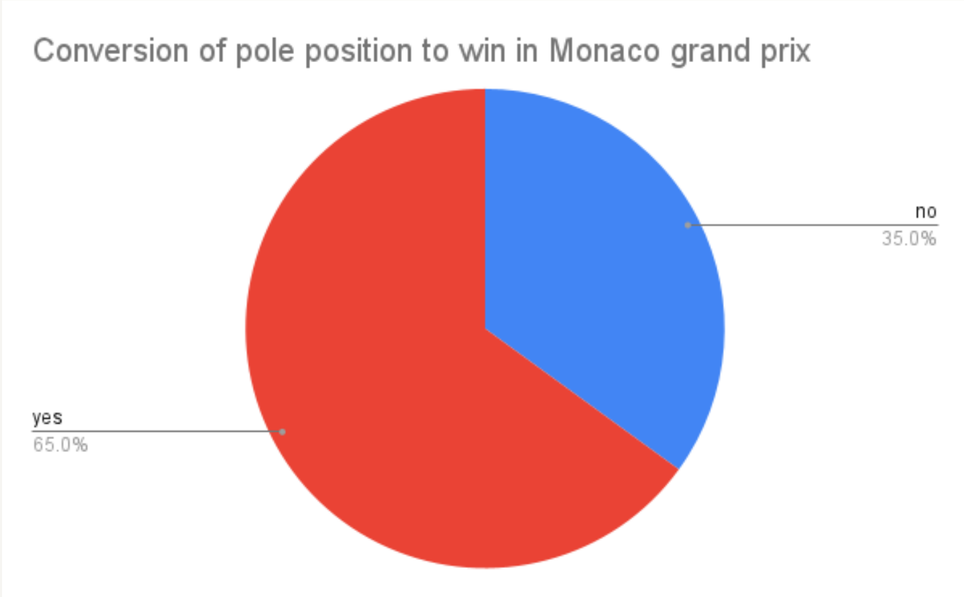
A. DRIVER METRICS

- Driver nationality and wins per country




- Drivers from UK have the highest winning rate among all.
- The Netherlands has the highest driver-to-win ratio, with one driver securing 59 wins.

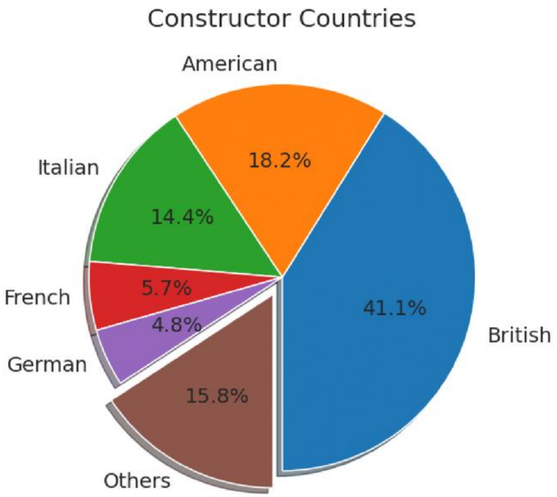
- Monaco pole conversion rate: the percentage of securing a win from pole position in Monaco grand prix



# DATA VISUALIZATION:

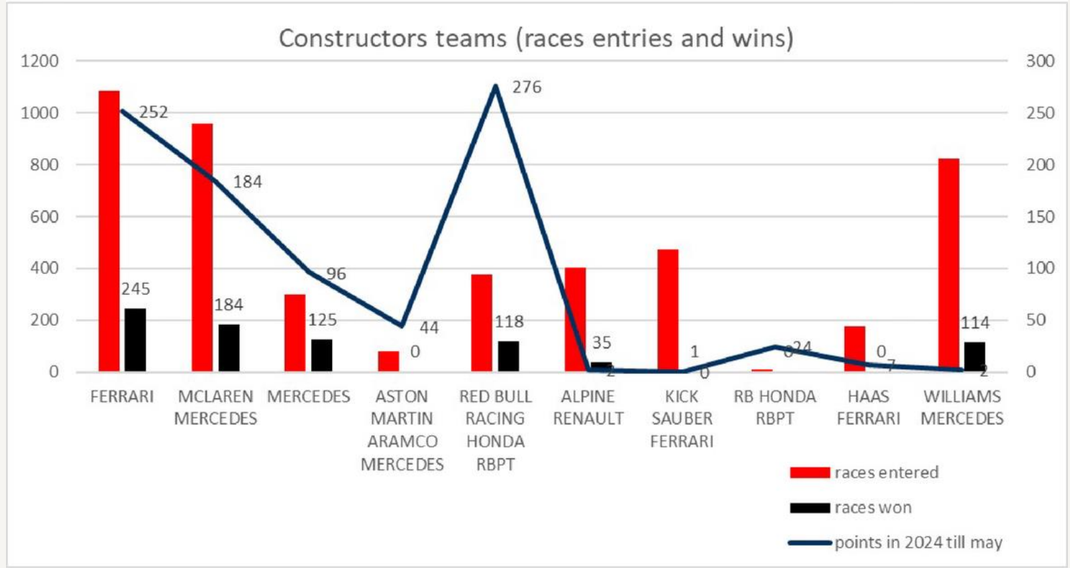
B. CONSTRUCTORS' METRICS:

Nationality of constructors in formula 1:




- Dominance of British constructors followed by American and European constructors

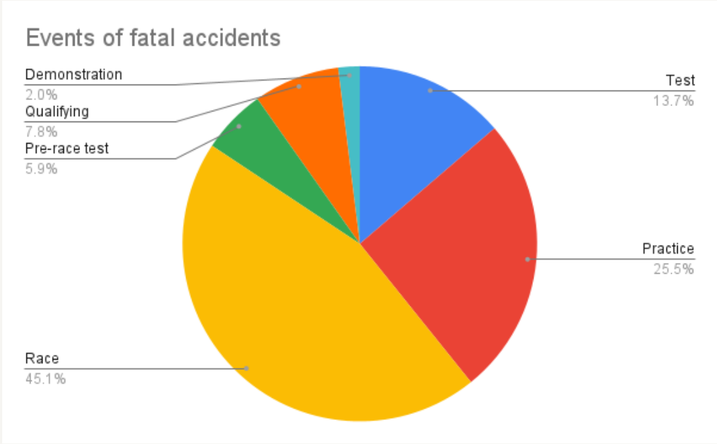
Constructors having highest wins and points till date.

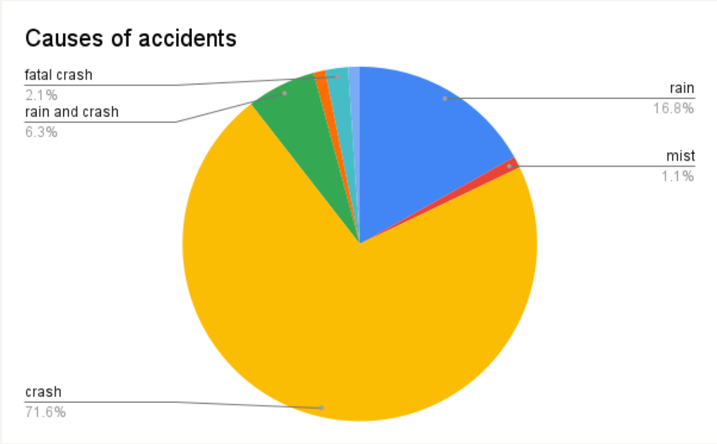



- Dominance of Ferrari and Redbull Racing in race wins and points scored followed by Mclaren and Mercedes
- Ferrai, Mclaren and Williams are some of the oldest constructors with maximum race entries

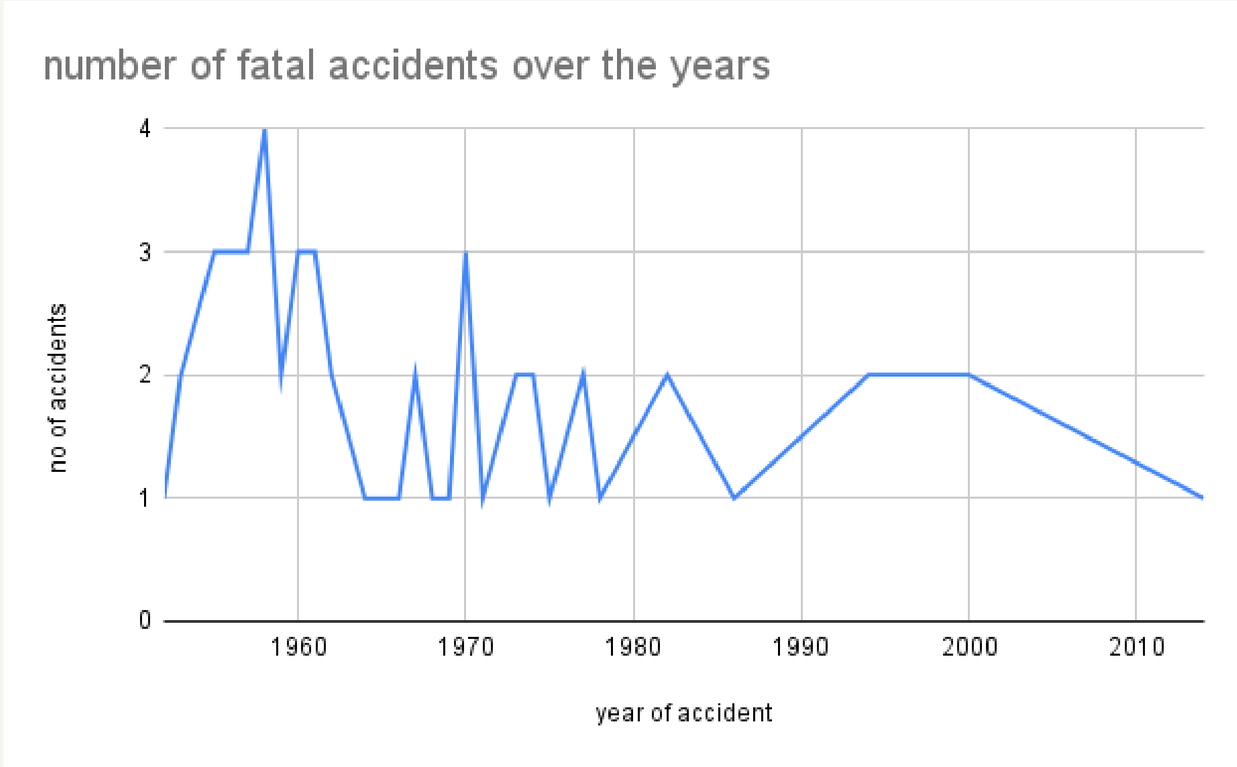
# DATA VISUALIZATION:

C. CASUALTY ANALYSIS:




- most accidents take place during races with car crashes being the main reason of these accidents




- Decline in number of fatal accidents happening on the racetrack due to implementation of numerous safety measures by the FIA

# Poisson Distribution for testing a hypothesis for Casualty analysis

## Objective:

Using chi square goodness of test will find out is this actually following poisson distribution.

## Hypothesis:

- Null Hypothesis (H0): observed data follows a Poisson distribution with mean λ.
- Alternative Hypothesis (H1): The observed data does not follow a Poisson distribution with mean λ.

## Rejection criteria:

- If $\chi 2$> critical value, reject the null hypothesis (H0).

OR

- If p-value < α, reject the null hypothesis (H0).

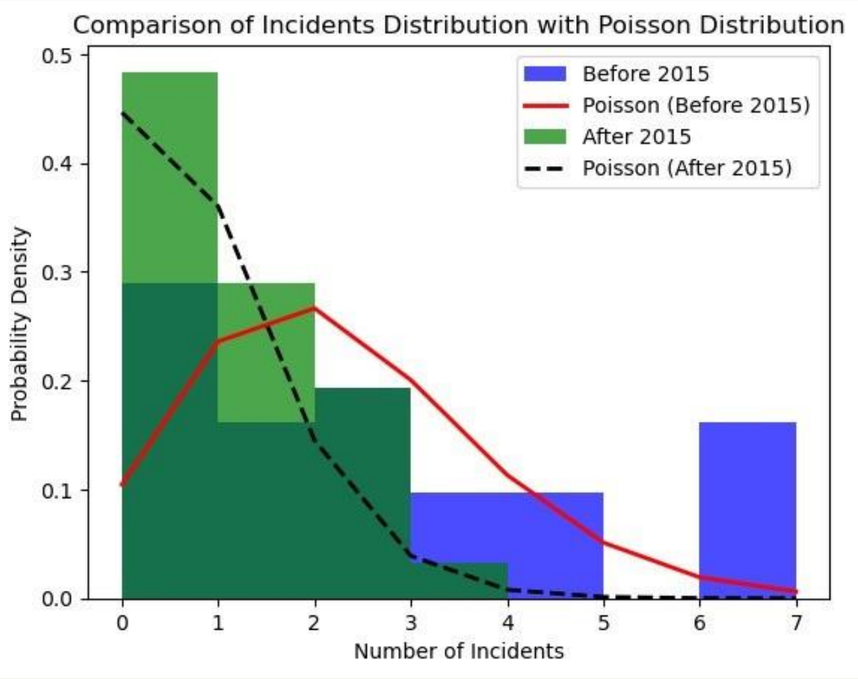
## Chi-square goodness-of-fit test results:

Before 2015 - Chi2 Statistic: 11.50, p-value: 0.12

After 2015 - Chi2 Statistic: 9.45, p-value: 0.22

Critical Values for Chi-square Test:

Critical Value: 14.07




Hence,

For Before and 2015 - Null hypothesis not rejected: Observed data follows Poisson distribution.

# Poisson Distribution for testing a hypothesis for Casualty analysis by UMP test.

**AIM: Compare casualty rates before and after 2015 (due to introduction of Virtual Safety Cars after 2015).**

The virtual safety cars are in the use since the year 2015.

**Hypothesis:**

- Null Hypothesis: H0: before = after =
  Significance of safety cars on casualty before 2015 and after 2015 are same
- Alternative Hypothesis: H1: before $\neq$ after
  the casualty after 2015 significantly decreased due to safety cars.

We use UMP test for our two sided hypothesis.

$$L(\lambda \mid \text{data}) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$$

$\Lambda = -2\ln(Lo/La)$

where L0 is calculated using under null hypothesis

and La is combination of before and after.

Compare $\Lambda$ with the Critical Value:

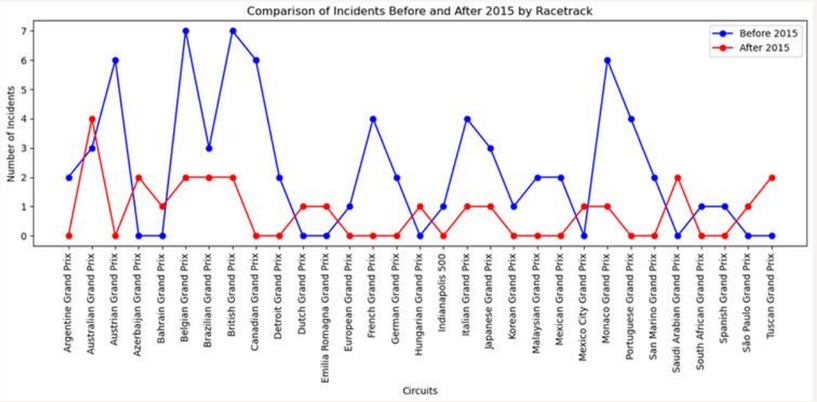
**Rejection Criteria:**

- If $\Lambda$ exceeds the critical value, reject the null hypothesis.

OR

- If p-value < α, reject the null hypothesis (H0).

On graph we can see the casualty has been decreased significantly due to safety cars from 2015.




But will see if it is really significant using ump test.

Using ump test we get the results as follows:

**Output:**

Sufficient Statistic (Sum of Incident Counts: ∑Xi): 95.0000

Likelihood Ratio Test Statistic (Lambda): 10.2625

P-value: 0.0014

Critical Value (alpha = 0.05): 3.8415

Decision: Reject the null hypothesis (there is a significant difference between the casualty before 2015 and incidents after 2015).

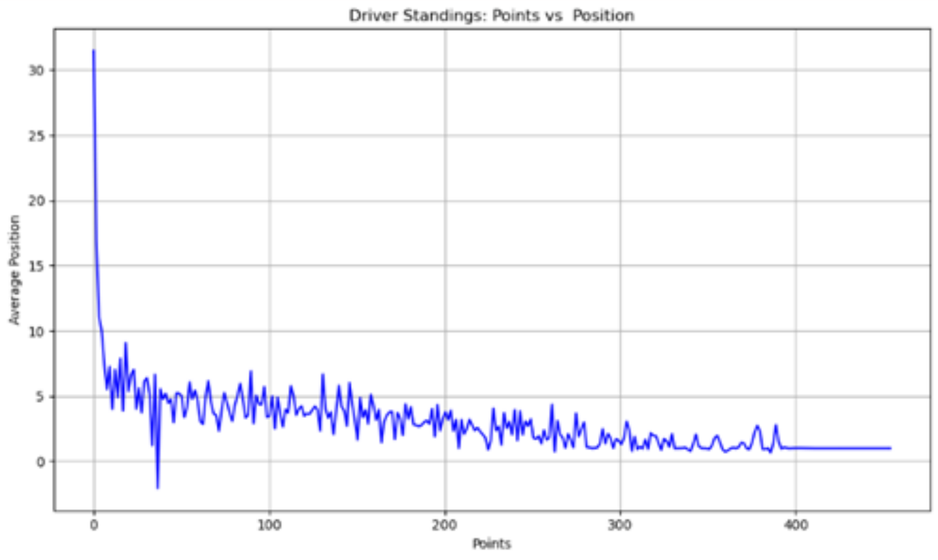
i.e. the casualty after 2015 significantly decreased due to virtual safety cars.

# Correlation between different variables of dataset:

## 1. Driver Points v/s Driver position:

### Output and Conclusion:

Spearman correlation coefficient: -0.862023048410559

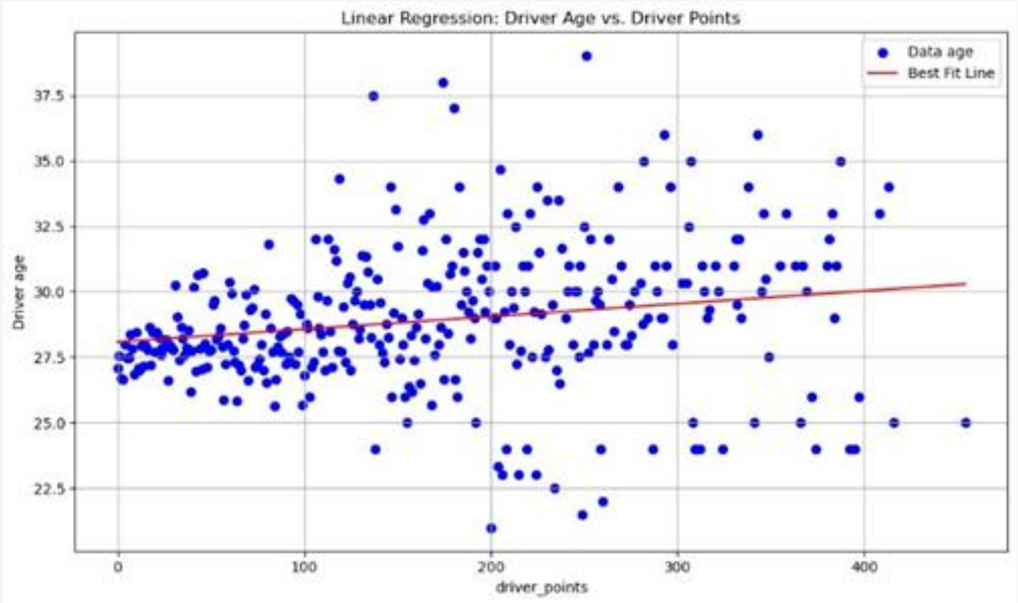



this is strongly negative correln graph it suggest that if a driver has a lower position(means a better position like a pole position) in the standings, they tend to have higher points.

## 2. Driver Points v/s Driver Age:

### Output and Conclusion:

Spearman correlation coefficient: 0.1926




this weakly positive graph roughly indicates that as the points scored by driver shows slight increase as the age ie. Expernice increases

# Objective: To build a winner prediction model for the current 2024 season of formula 1

## Why not Logistic Regression Model?

The dataset ranges from 1994 to 2023.

- **Variables :** Our response variable(y) is champion(driver that has maximum points at the end of the season) and independent variables are driver_wins, podium, grid, circuits, nationality, constructors, driver_age, qualifying_time, year, round.
- The data is split into training set(80%) and test set(20%).The training set consists of 4132 observations and test set consists of 1034 observations.

Conclusion:

The logistic model fails to validate the assumption of linearity of independence variable with the logit.

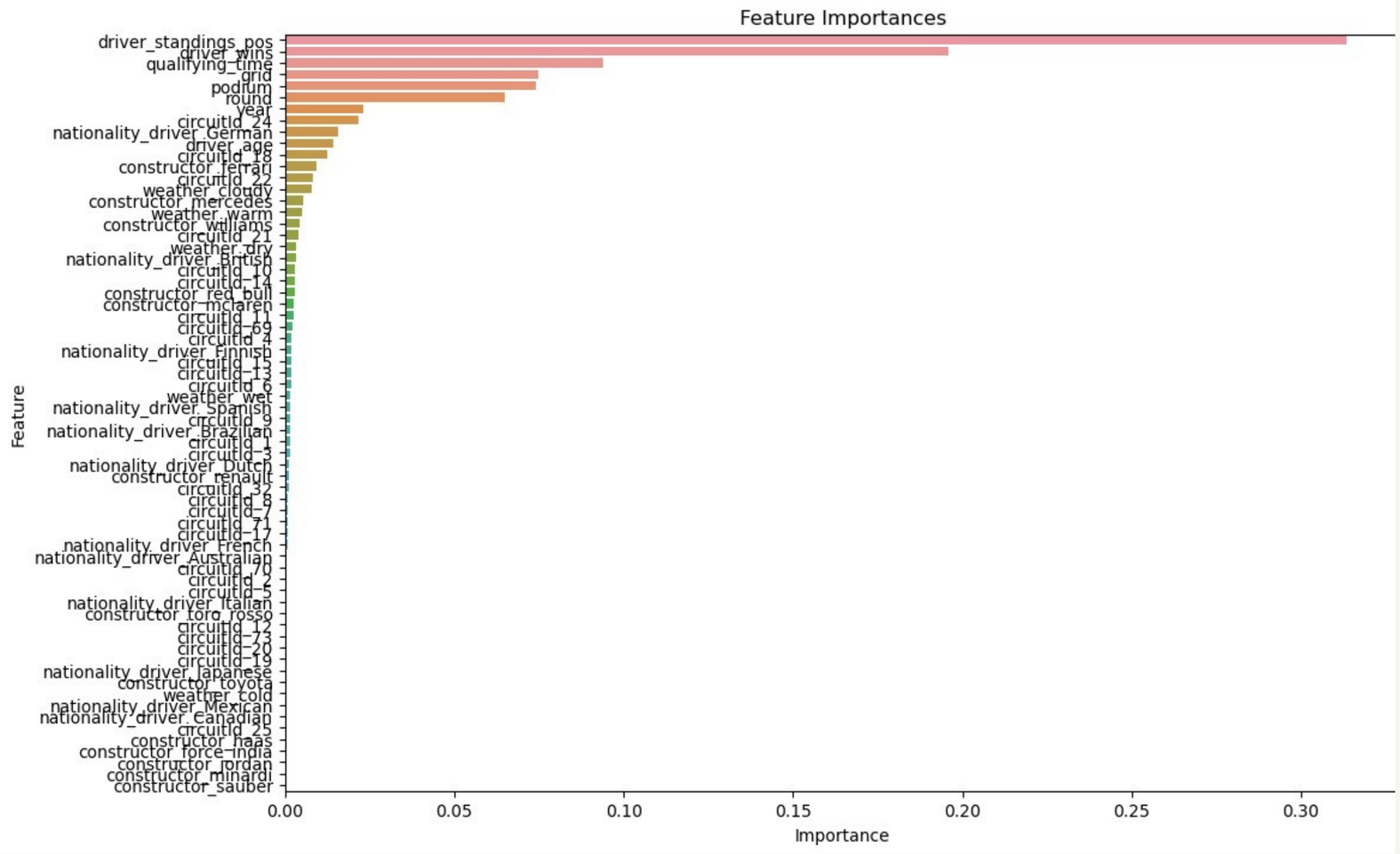
Objective: To build a winner prediction model for the current 2024 season of formula 1

## Random Forest model

The dataset ranges from 1994 to 2023.

- **Variables** : Our response variable(y) is champion(driver that has maximum points at the end of the season) and independent variables are driver_wins, podium, grid, circuits, nationality, constructors, driver_age, qualifying_time, year, round.
- The data is split into training set(80%) and test set(20%).The training set consists of 4132 observations and test set consists of 1034 observations.
- **Model Training**: A RandomForest classifier is trained on the resampled training data.
- Feature Importance
- **Prediction for 2024**: we used the 2023 data structure for prediction and made sure that the features in the 2024 data match the training data and used it to predict probabilities of each driver belonging to winning class .

Feature importance display the importance of each variable in predicting the target variable

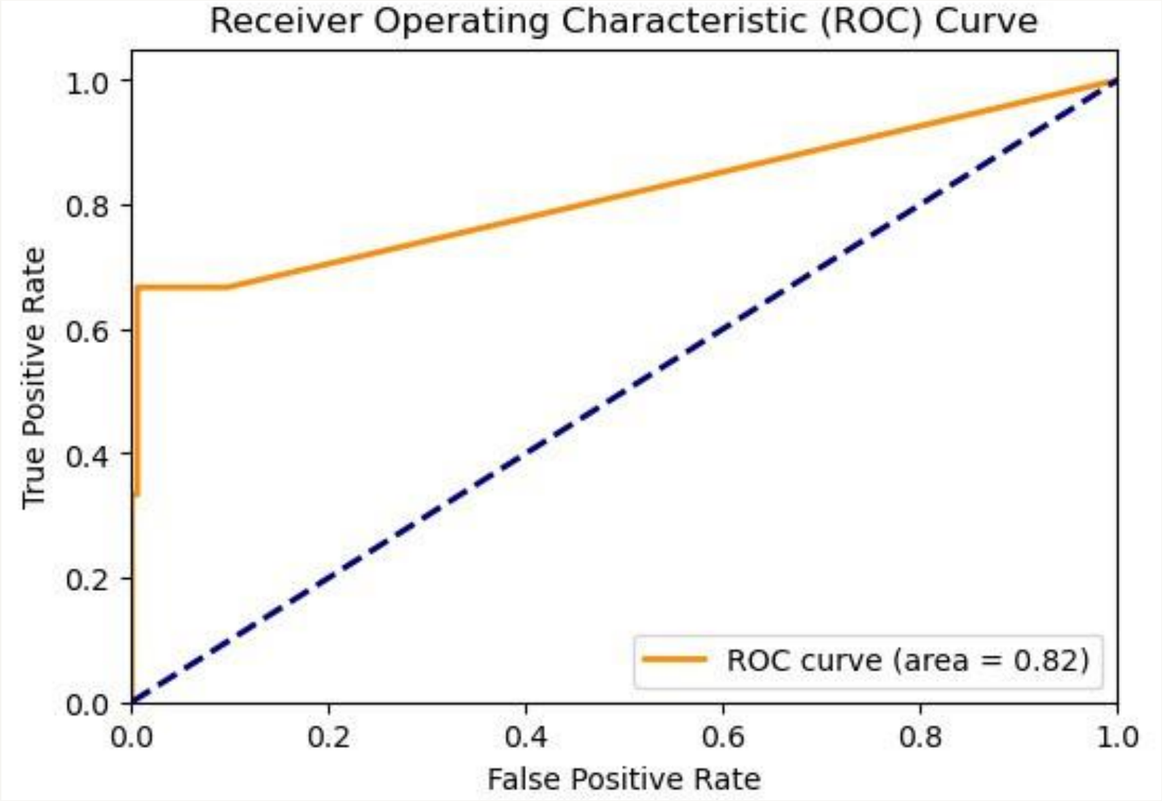


# Output and Interpretation:

**Accuracy:** 0.998, indicates high performance.
**ROC-AUC Score:** 0.815
**Confusion matrix for test data:**

| | | Predicted | |
|---|---|---|---|
| | | 0 | 1 |
| Actual | 0 | 1031 | 0 |
| | 1 | 2 | 1 |

True Positives (TP): 1, True Negatives (TN): 1031
False Positives (FP): 0,False Negatives (FN): 2




Sensitivity (True Positive Rate): 0.3333 or 33.33%, Specificity (True Negative Rate) : 1 or 100%, Precision: 1 or 100%

# Output and Interpretation:

**Confusion matrix for train data:**

**True Positives (TP):** 27
**True Negatives (TN):** 4105
**False Positives (FP):** 0
**False Negatives (FN):** 0

**Sensitivity** (True Positive Rate): 1 or 100%
**Specificity** (True Negative Rate) : 1 or 100%
**Precision:** 1 or 100%

| | | Predicted | |
|---|---|---|---|
| | | 0 | 1 |
| Actual | 0 | 4105 | 0 |
| | 1 | 0 | 27 |

# Output and Interpretation:

- Each tree indepaendently predicts a class for a given input. To obtain probabilities, each decision tree in the random forest can estimate the probability that a particular sample belongs to each class and then aggregates the probabilities from all the individual trees for that class.

| Driver | Probability of the driver belonging to winning class |
|---|---|
| max_verstappen | 0.109000 |
| norris | 0.001667 |
| hamilton | 0.001429 |
| perez | 0.001111 |

- Max Verstappen is predicted by the model to have the highest chance of winning the championship in 2024, with a win probability of 10.9%.
- Most drivers have negligible win probabilities, highlighting a competitive gap according to model's predictions
- Even the highest probability (10.9% for Verstappen) is relatively low, suggesting that the model might be quite uncertain about its predictions. This could be due to the inherent unpredictability of the sport or potential limitations in the dataset and features used.

# Conclusion:

We can incorporate more data or additional relevant features that could enhance the model's accuracy and reliability in predicting championship outcomes.

For eg. 1) Car specifications such as engine,tyres,top speed
2) Team strategy
3) pitstop strategy, Time taken for each pitstop
4) collisions occurred during the race, cars involved in the collision
5) tyre strategy in case of rain(intermediate and full wet tyre)

The performance metrics like accuracy, ROC-AUC score, and the classification report can give a sense of how well the model is performing overall, particularly in distinguishing between champions and non-champions.

These are the driver standings of 2024 which validate that max Verstappen could win the championship as he is currently leading the driver.

**Objective:To evaluate Driver’s performance Index and study various drivers metrics**

## Principal Component analysis

Why are we using Principal Component analysis?

↓

What metrics are we using for the Driver's performance index?
(Overtakes Count, Overtaken Count, points, driver standings pos, driver wins, driver age, and Races Participated)

↓

How PCA is Applied in This Study?

↓

Performance_score = normalized metrics * weights (1st PC).

# Results

Weights Derived from 1st PC (1994-2020):

- Overtakes: 0.240
- Overtaken: 0.240
- Points: 0.469
- Standings: 0.543
- Wins: 0.400
- Age: 0.171
- Races Participated: 0.426

1st principal component explains 73% of the total variation in the dataset

Higher weights indicate more significant metrics in explaining the variance in the data

Driver Performance Index (1994-2020)

Top Drivers:

- Hamilton: 3.16
- Vettel: 2.29
- Raikkonen: 1.74
- Alonso: 1.68
- Schumacher: 1.28

Hamilton has the highest performance index

Hamilton has most World Drivers' Championships

# DPI for year 2019-2020

**Weights Derived from PCA(2019-2020)**

- **Overtakes: 0.0961**
- **Overtaken: 0.3673**
- **Points: 0.5453**
- **Standings: 0.5401**
- **Wins: 0.4496**
- **Age: 0.0268**
- **Race Participated: -0.2528**

**Driver Performance Ratings(2019-2020)**

- **Top Drivers:**
  - **Hamilton: 4.82**
  - **Bottas: 2.56**
  - **Verstappen: 2.28**
  - **Leclerc: 1.20**
  - **Ocon: 0.67**
  - **Sainz: 0.55**
  - **Vettel: 0.51**
  - **Gasly: 0.14**
  - **Perez: 0.12**
  - **Hulkenberg: 0.05**



# Conclusion

- The machine learning model predicted that Max Verstappen has the highest chance of winning the championship in 2024, with a win probability of 10.9% with respect to other drivers.
- Hamilton has the highest Drivers performance index suggesting he is the best formula 1 driver.
- In Drivers performance index; the Points, driver standings position, and Races Participated are the most significant metrics in explaining the variation in the dataset.
- There is a significant difference in the incident rate after 2015 which suggests that introduction of Virtual Safety cars is effective.
- If a driver has a lower standing position, they tend to have higher points.
- The points scored by driver shows slight increase as the age increases.

# Applications /Future scope

- FIA foundations are working on research that focuses on safety improvements and sustainability. If the data is available we can work on finding the effectiveness of the measures taken.
- The data about Car engines and speeds because of confidentiality is not available but including that data we can increase the reliability of the prediction models.
- More metrics availablity about the driver can also give us better index.

# THANK YOU!