INT 353 FINAL REPORT

EDA

BY: VAMSHI ERUGADINDLA

FROM: K20CH

REG ID: 12017419

ROLL NO: RK20CHB39

WHAT IS THIS PROJECT?

Exploratory Data Analysis is ana approach to analyse the data using various visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representation. The dataset I have chosen has the data of F1 racing championships from the year 1983.

This dataset was taken from Kaggle and has information about all the tournaments from 1983 to 2021.

INTRODUCTION:

Formula one also popularly known as F1 racing is an early European grand prix championship of the 1920s and the 1930s.A point system is used at Grands Prix to determine two annual world championships: one for ***drivers*** and the other for ***constructors.*** Each driver may use no more than thirteen sets of ***dry-weather*** tyres, four sets of intermediate tyres, and three sets of ***wet-weather*** tyres during a race weekend.

QUALIFYING: Drivers would have one or more sessions in which to set their fastest time (qualifying time), with the grid order determined by each driver's best single lap.

DATA CLEANING:

Data cleaning is the process of fixing or removing incorrect, corrupted, duplicated, or incomplete data within a dataset. When combining multiple data sources, there are many ways that data can be duplicated.

Luckily the dataset I have chosen was already clean and it didn’t require any data cleaning.

Data cleaning

Finding null values in data

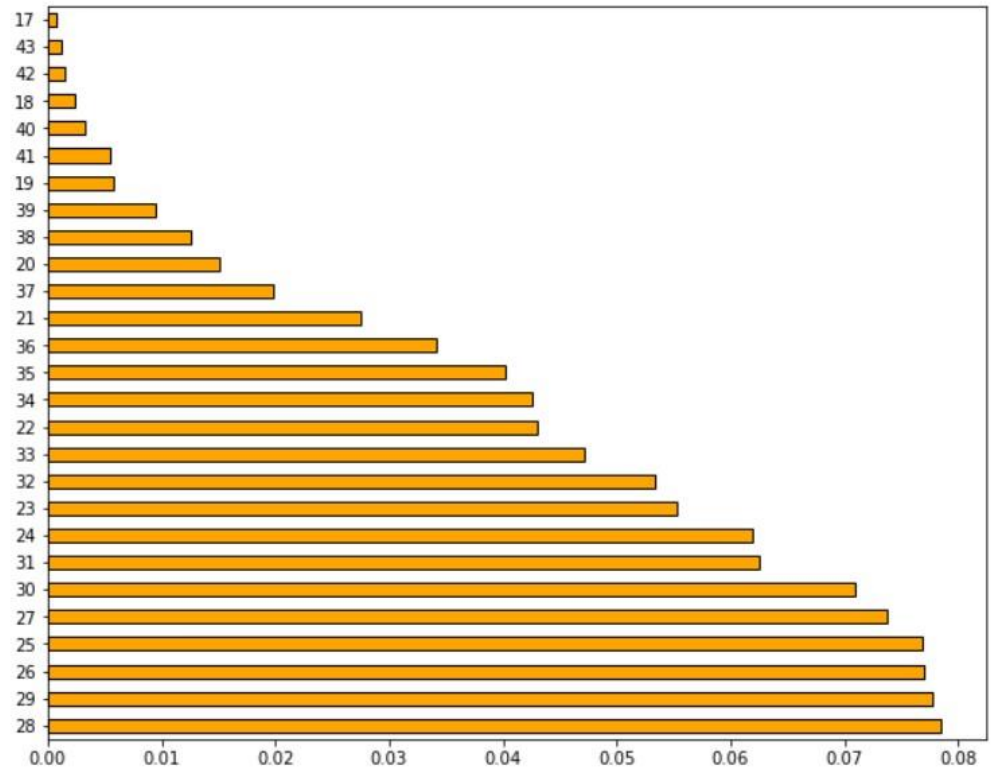
```
In [8]: df.isna().sum()

Out[8]: Unnamed: 0                0
        season                    0
        round                     0
        circuit_id                0
        weather_warm              0
        weather_cold              0
        weather_dry               0
        weather_wet               0
        weather_cloudy            0
        driver                    0
        nationality               0
        constructor               0
        grid                      0
        podium                    0
        driver_points             0
        driver_wins               0
        driver_standings_pos      0
        constructor_points        0
        constructor_wins          0
        constructor_standings_pos 0
        qualifying_time           0
        driver_age                0
        dtype: int64
```

UNIVARIATE ANALYSIS:

Univariate analysis is the simplest form to analyse single variable which does not have any relation with other variable and the purpose is to describe the data and the patterns that exist within a variable.

```
plt.figure(figsize=(10,8))
df.driver_age.value_counts(normalize=True).plot.barh(color="orange",edgecolor="black")
plt.show()
```

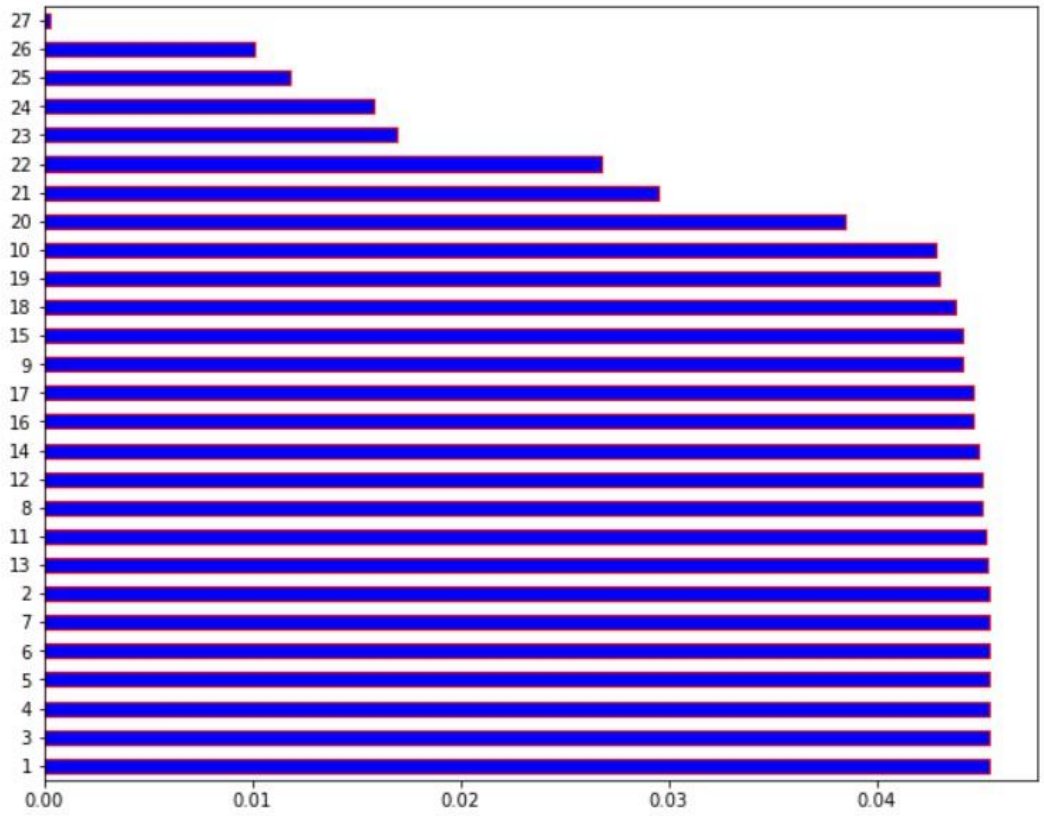


The above bar graph describes the frequency of each category in particular columns in separate bars.

It is to show the average number of drivers of a particular age.

The next graph depicts the same for grid number.

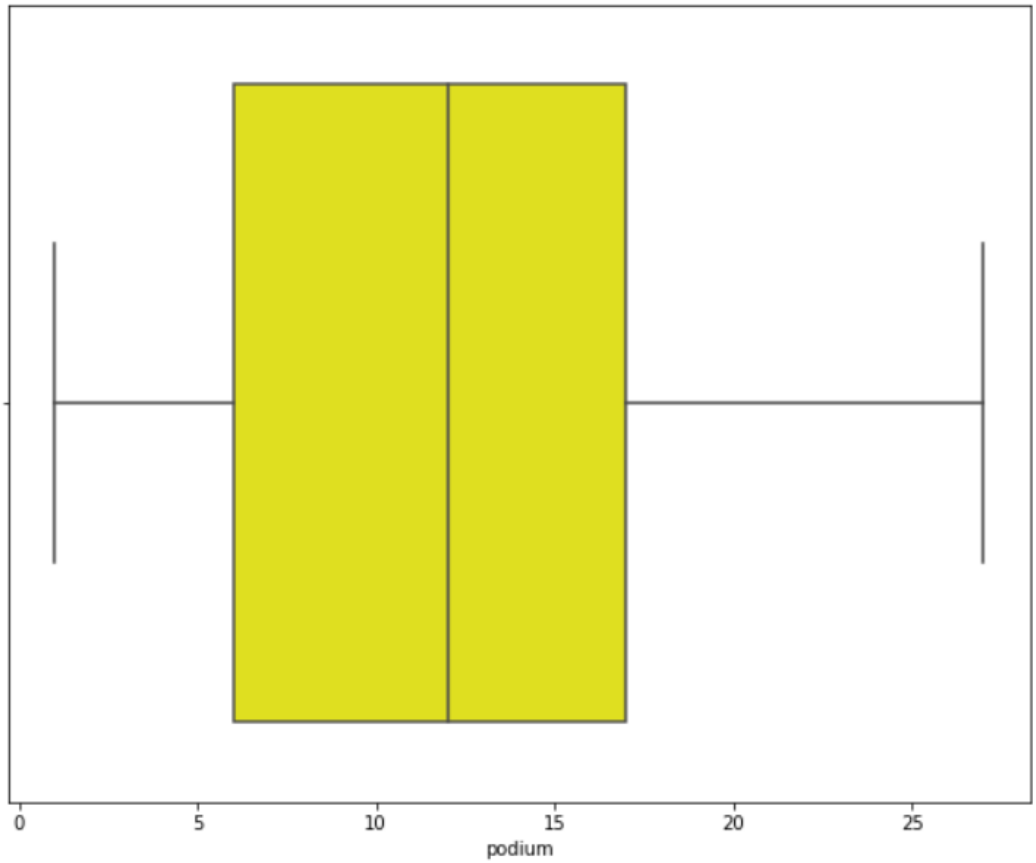
```
plt.figure(figsize=(10,8))
df.grid.value_counts(normalize=True).plot.barh(color="blue",edgecolor="red")
plt.show()
```



BOX PLOT:

The boxplot describes the lower and upper boundary of standard deviation which is called interquartile range.

The following box plot describes the median value of the podium number.



It is clearly visible that the interquartile range of the podium number is between 10 and 15, and approximately 12.

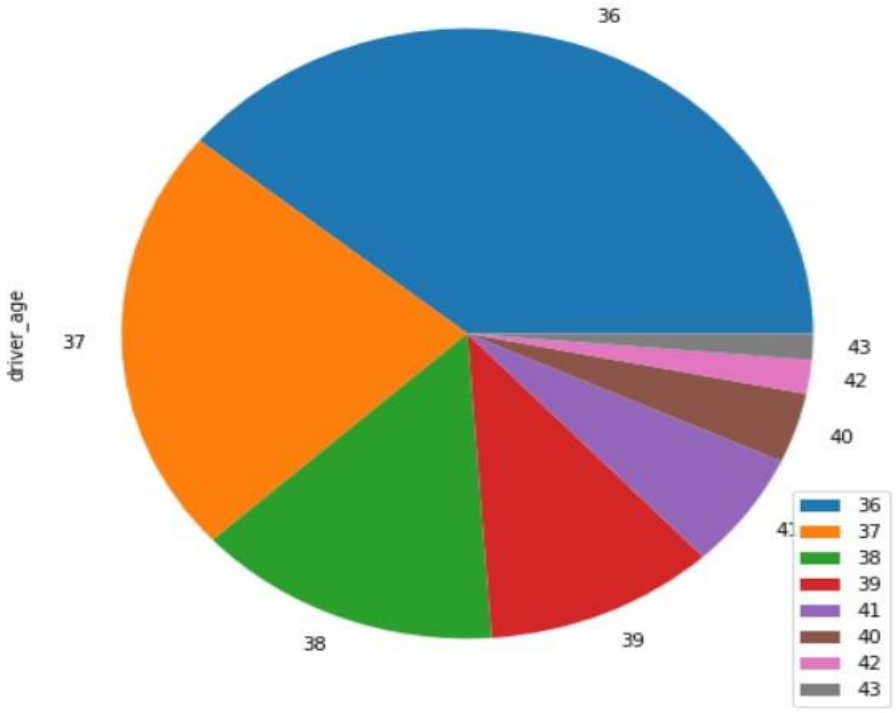
PIE CHART:

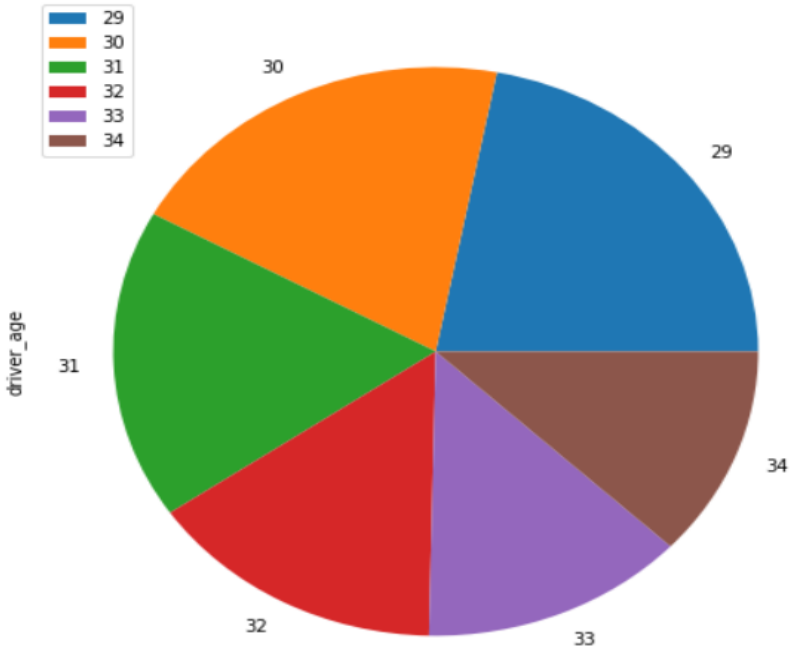
A Pie chart is a statistical which is divided into slices to illustrate numerical proportions.

In a pie chart the arc length of each slice is proportional to the quantity it represents.

Pie charts can be easily understood.

The following pie chart represents univariate ordered analysis of driver age greater than or equal to 36.

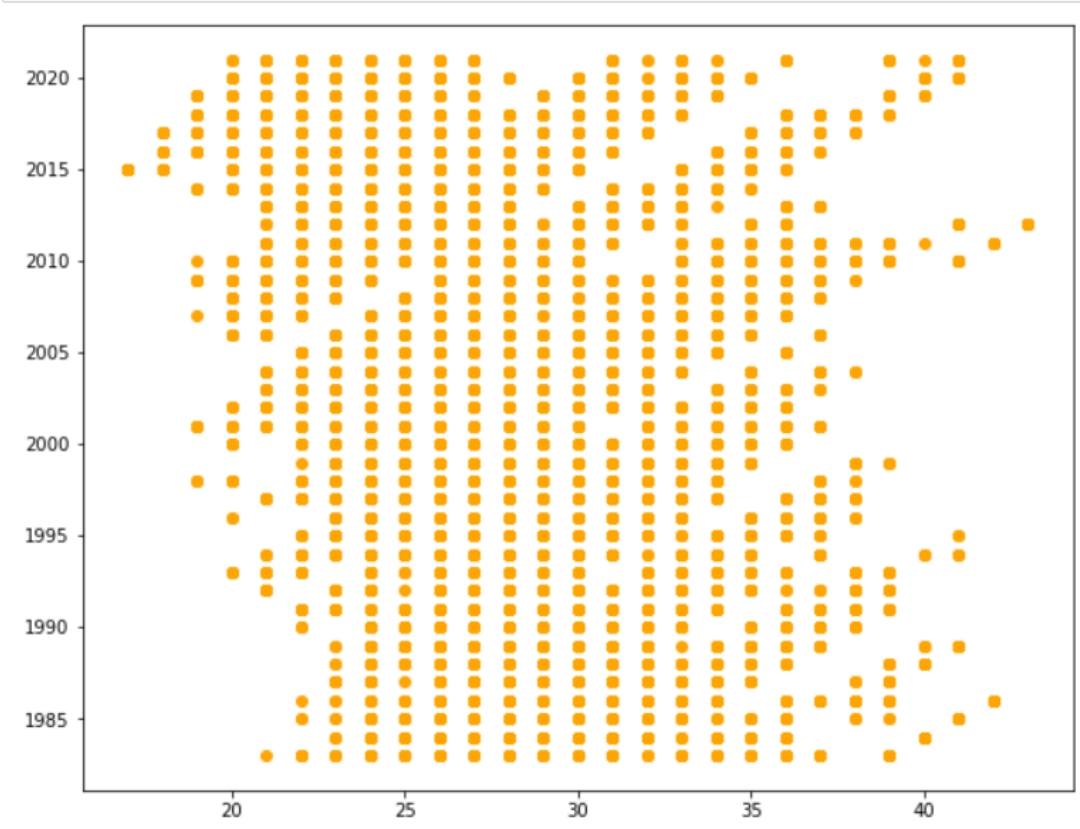




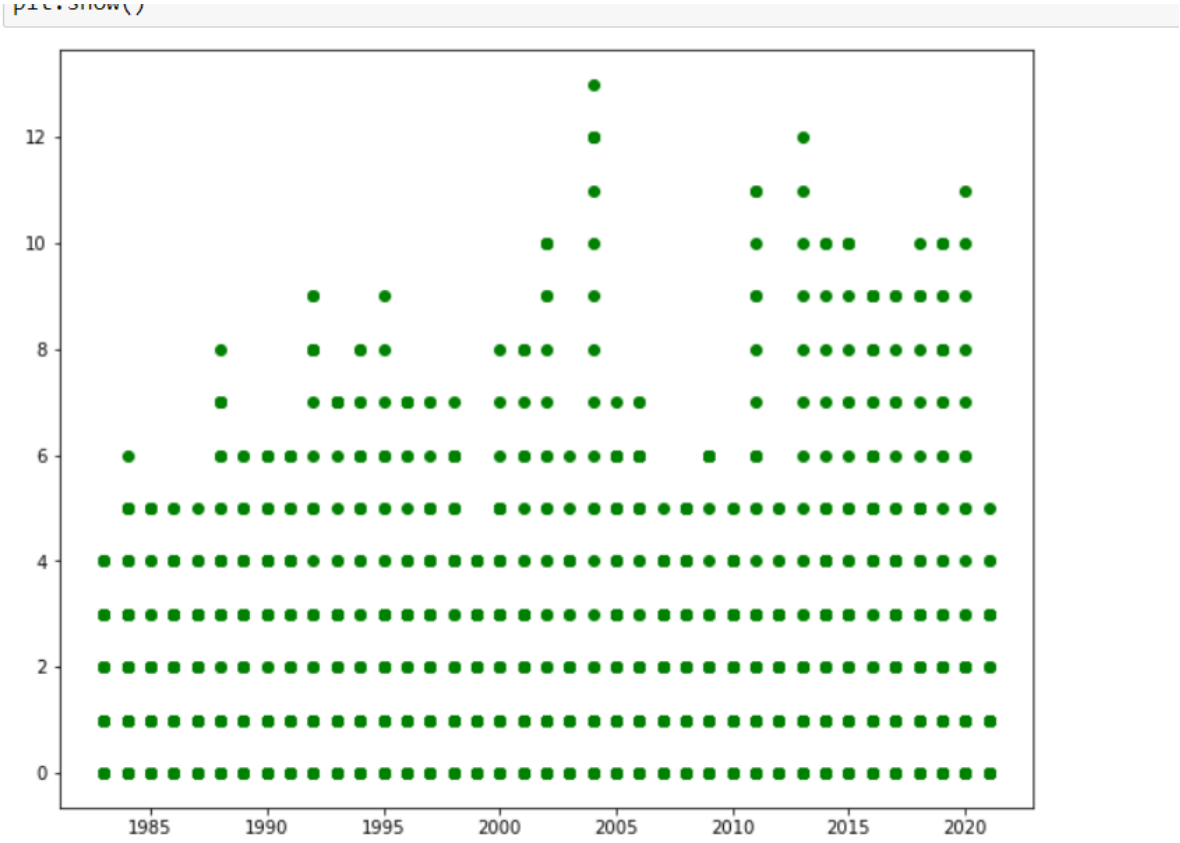
BI VARIATE ANALYSIS:

Bivariate analysis is one of the simplest forms of quantitative analysis. It involves the analysis of two variables. The purpose is determining empirical relationship between them.

Bivariate analysis can be helpful in testing simple hypothesis of association.



The above graph depicts the bivariate relation between season and the age of the age of the driver participated in that season.



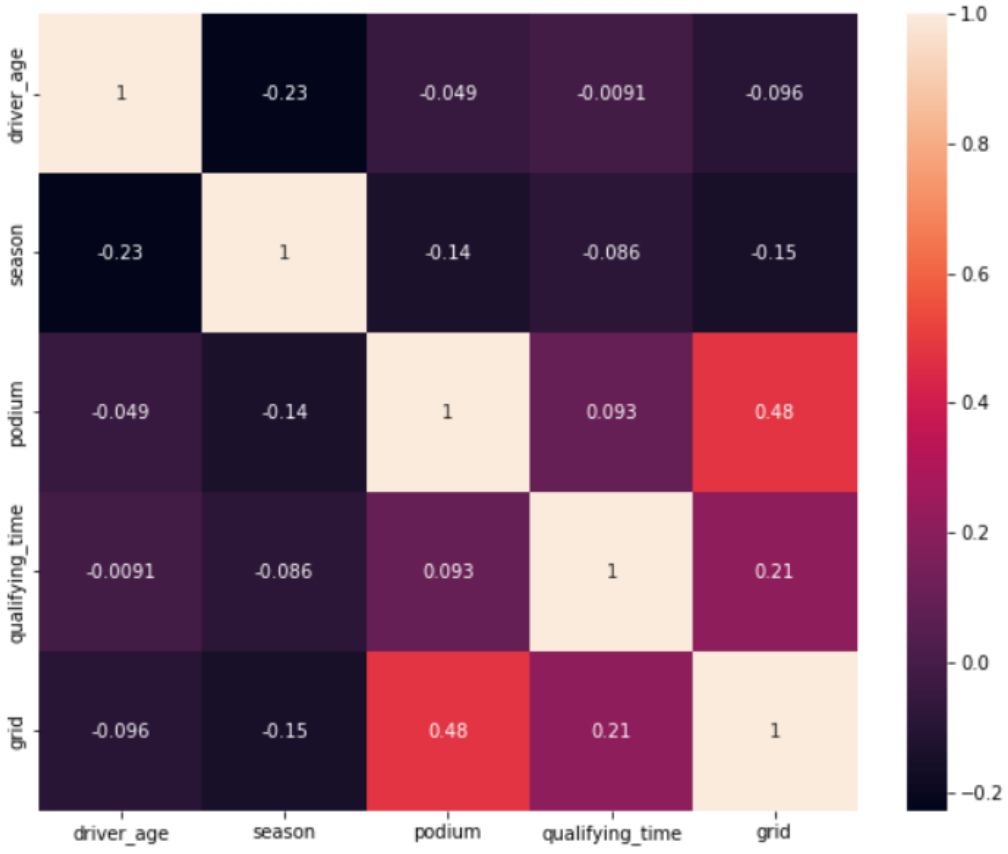
The above graph shows the bivariate relation between season and number of driver wins in that particular season.

CORRELATION HEAT MAP:

Correlation heat maps are a type of plot that visualises the strength of relationship between numerical variables.

They are used to understand which variables are related to each other.

Correlation heatmaps can be used to find both linear and nonlinear relationships between variables.



The above correlation heatmap shows the correlation between the age of the driver that won the season and his qualifying time.

STATISTICAL ANALYSIS:

The following images are the statistical analysis of my data set.

Stastical Analysis

```
df.describe()
```

	Unnamed: 0	season	round	grid	podium	driver_points	driver_wins	driver_standings_pos	constructor_points	constructc
count	14794.000000	14794.000000	14794.000000	14794.000000	14794.000000	14794.000000	14794.000000	14794.000000	14794.000000	14794.
mean	7465.306476	2001.589766	9.185278	11.761052	11.896782	19.940922	0.363864	10.664188	40.059010	0.
std	4350.153881	11.242589	5.116603	6.701385	6.766711	42.078820	1.175424	7.671286	81.624051	1.
min	0.000000	1983.000000	1.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.
25%	3698.250000	1992.000000	5.000000	6.000000	6.000000	0.000000	0.000000	4.000000	0.000000	0.
50%	7403.500000	2001.000000	9.000000	12.000000	12.000000	3.000000	0.000000	10.000000	8.000000	0.
75%	11232.750000	2012.000000	13.000000	17.000000	17.000000	19.000000	0.000000	17.000000	41.000000	0.
max	15085.000000	2021.000000	21.000000	27.000000	27.000000	387.000000	13.000000	30.000000	722.000000	18.

```
df.driver_age.describe()
```

```
count    14794.000000
mean        28.586184
std          4.730817
min         17.000000
25%         25.000000
50%         28.000000
75%         32.000000
max         43.000000
Name: driver_age, dtype: float64
```

```
df.podium.describe()
```

```
count    14794.000000
mean        11.896782
std          6.766711
min          1.000000
25%          6.000000
50%         12.000000
75%         17.000000
max         27.000000
Name: podium, dtype: float64
```

```
df.grid.describe()
```

```
count    14794.000000
mean        11.761052
std          6.701385
min          1.000000
25%          6.000000
50%         12.000000
75%         17.000000
max         27.000000
Name: grid, dtype: float64
```

```
df.qualifying_time.describe()
```

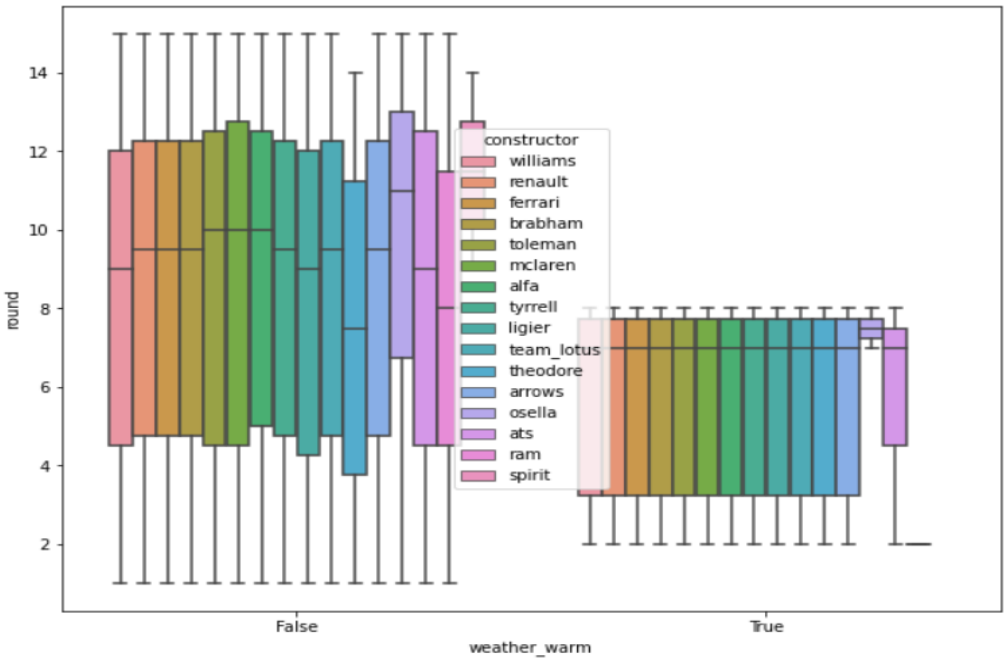
```
count    14794.000000
mean         2.553860
std          8.000276
min        -77.000000
25%          1.000000
50%          2.100000
75%          3.500000
max        904.600000
Name: qualifying_time, dtype: float64
```

MULTIVARIATE ANALYSIS:

Multivariate analysis is defined as: The statistical study of data where multiple measurements is made on each experimental unit and where the relationships among multivariate measurements and their structure are important. Multivariate analysis is part of Exploratory data analysis.

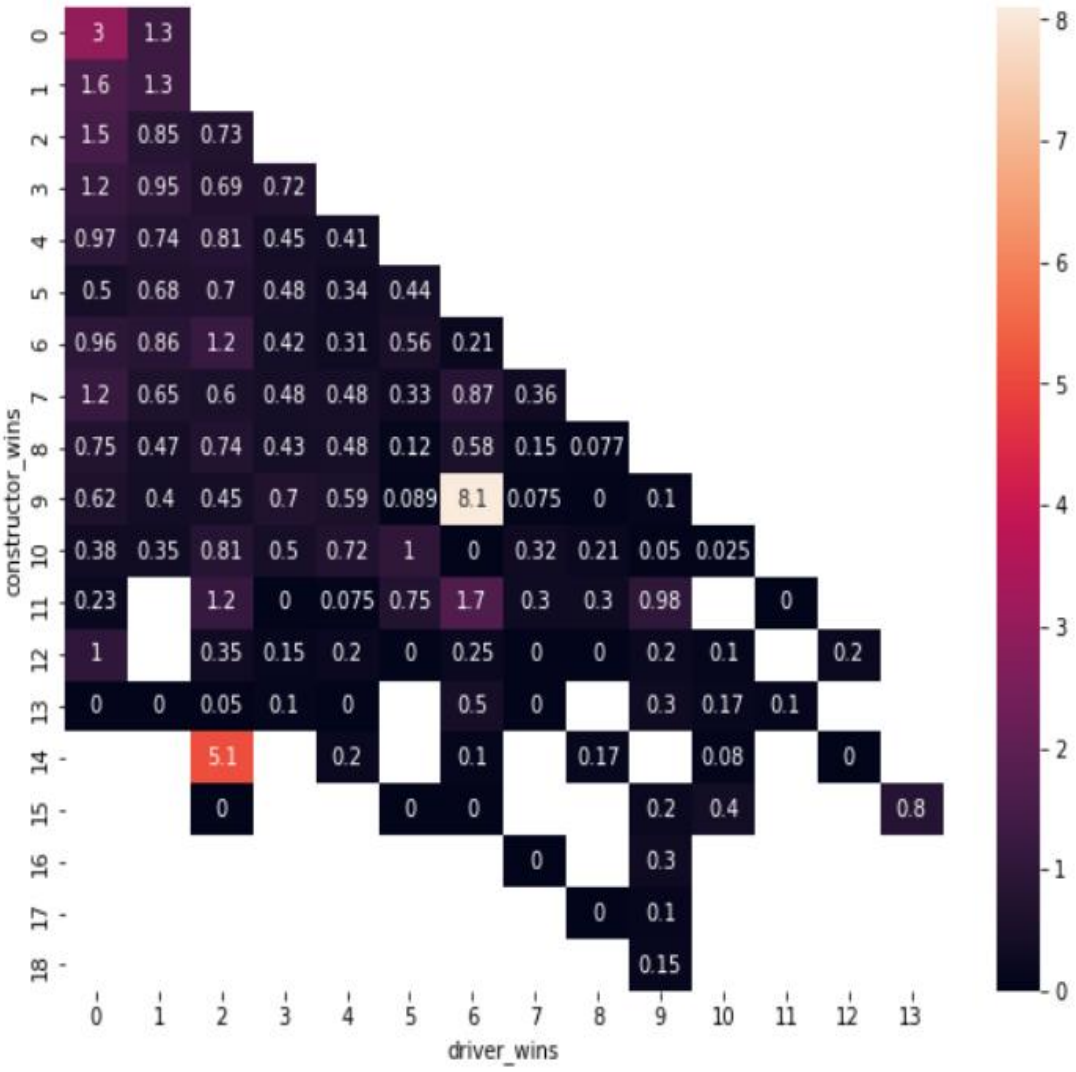
Based on MVA, we can visualize the deeper insight of multiple variables. There are more than 20 different methods to perform multivariate analysis and which method is best depends on the type of data and the problem you are trying to solve.

BOX PLOT: A box plot (or box-and-whisker plot) shows the distribution of quantitative data in a way that facilitates comparisons between variables or across levels of a categorical variable.

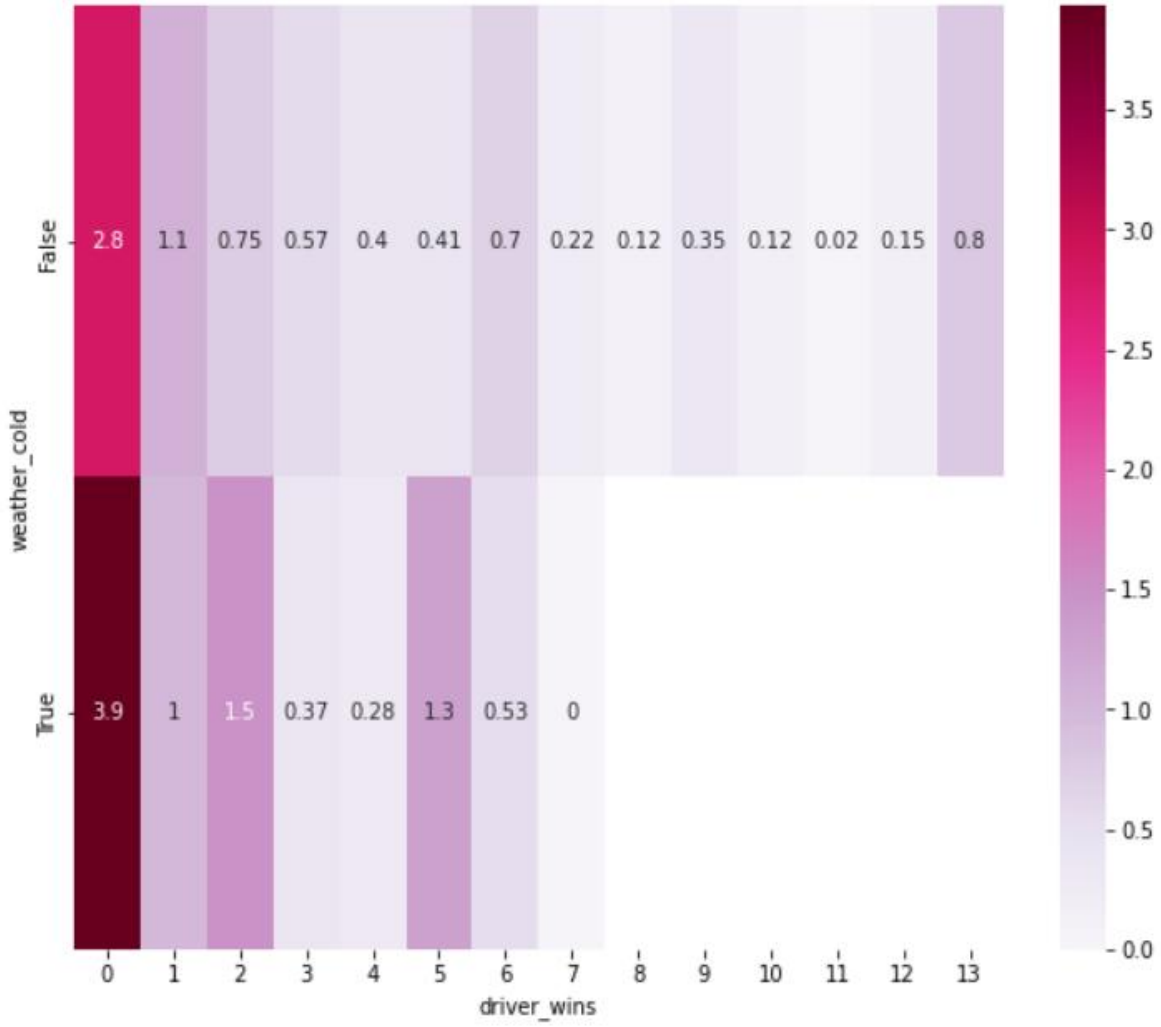


The preceding boxplot explains the multivariate analysis between the weather, number of rounds and the name of the constructor that won the number of rounds in that weather. It clearly shows that in warm weather the median wins of the teams are nearly equal.

HEAT MAP: A heatmap is a useful visualization method to illustrate multivariate data when there are many variables to compare, such as in a big data analysis. It is a plot that displays values in a colour scale in a grid.

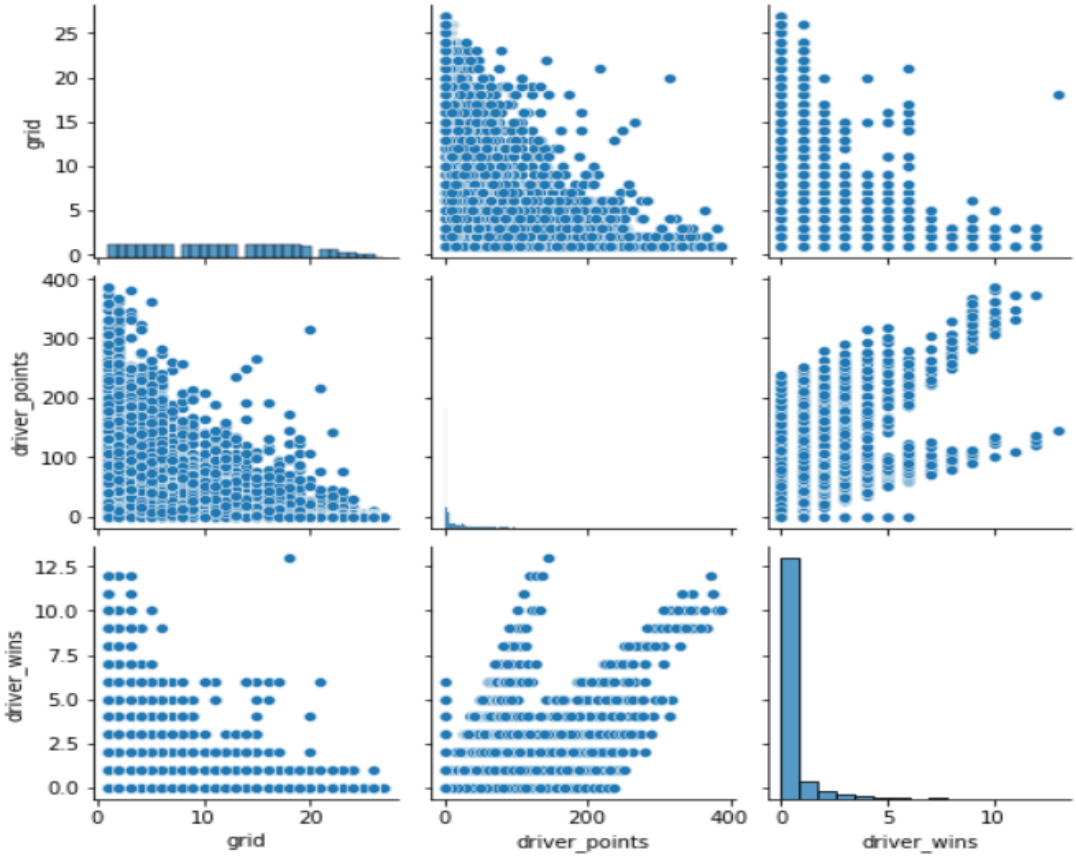


The above heatmap gives the number of drivers wins in the qualifying time of a particular constructor.



The above heatmap describes the average number of driver wins in a cold weather and their grid numbers. It shows in cold weather there are not more than 8 wins in a cold weather in the same season.

PAIR PLOT: A pair plot creates a grid of scatter plots to compare the distribution of pairs of numeric variables. It also features a histogram for each feature in the diagonal boxes.



The above graphs are called pair plots that forms grids of scatterplot distributions. The above graphs describe the pair plots between the grid number, driver points and the number of wins by a particular driver.

CONCLUTION

By the following project I was able to analyse any given data using univariate, bivariate, and multivariate data analysis, I was also able to visualise the given data using different techniques.

In my dataset I can conclude that as the time passed the competition for the f1 racing was increased, it seems that qualifying time for each season was gradually decreased by the increased competition.

It is also observed that weather conditions greatly effect the number of wins scored by the team.

I can finally conclude that analysing a huge amount of data was made easy by EDA.