WEB SCRAPPING AND DATA ANALYSIS

Name: Manish Mohapatara (045029)

Link: https://github.com/Manish-045029/045029_python-project.git

OBJECTIVE:

This managerial report provides an overview of a Python script designed to scrape data from a website and analyse the data. The objective of this project is to gather, analyse, and present historical data on Formula One World Drivers' Champions from the Simple English Wikipedia page.

The project aims to provide insights into the drivers who have won Formula One championships, including their names, nationalities, and the years in which they achieved their championships.

Additionally, it utilizes data analysis and statistical tools to explore patterns, trends, and statistics related to the champions' ages and the average ages of champions over different decades.

By compiling and presenting this data, the project seeks to offer a comprehensive overview of Formula One World Drivers' Champions throughout the years, facilitating a deeper understanding of the sport's history and evolution.

Methodology:

Data Collection:

- Web Scraping: Gathered data from the Simple English Wikipedia page containing a list of Formula One World Drivers' Champions.
- The Beautiful Soup library was used to parse the HTML content of the webpage, and the requests library was utilized to send HTTP GET request.

Data Extraction:

- The script identifies the table containing information on Formula 1 championships in each page.
- Headers are extracted from the table and cleaned, removing unwanted columns.
- Data from each row of the table is extracted, cleaned, and added to a Data Frame.

Data Cleaning and Preprocessing:

- Data Cleaning: Removed any unnecessary characters, whitespace, or special characters from the extracted data.
- Handling Missing Data: Checked for missing or incomplete data, and excluded any rows with missing information.

Result Presentation:

The final dataset is presented as a formatted table using the tabulate library.

General Description of Data:

The data extracted from the Simple English Wikipedia page provides information about the Formula One World Drivers' Champions. Formula One, often referred to as F1, is a renowned international motorsport that features a series of annual racing championships. The dataset specifically focuses on the list of drivers who have achieved the prestigious title of "Formula One World Drivers' Champion" throughout the history of the sport.

Columns:

The dataset contains several columns that provide specific details about each Formula One World Drivers' Champion. These columns include:

- Driver: The name of the Formula One driver who won the championship.
- Nationality: The nationality or country of origin of the driver.
- Championships: The number of Formula One World Drivers' Championships won by the driver.
- Seasons: The years or seasons in which the driver secured the championship.
- Constructor: The name of the Constructor who won the championship.
- Engine: Engines who won the most number of seasons
- Poles Podiums: Number of poles and podiums won by a specific driver
- Age: Age of the driver who won championship

Entries:

- Each row in the dataset represents a different Formula One driver who has won the championship at least once. The data spans multiple decades and includes champions from various countries.

Season	Driver	Age	Constructor	Engine	Tyres	Poles	Wins	Podiums	F. Laps	Points	% Points	Clinched
1950	Giuseppe Farina [1]	44	Alfa Romeo	Alfa Romeo	P	2	3	3	3	30	83.333 (47.619)	Race 7 of 7
1951	Juan Manuel Fangio [2]	40	Alfa Romeo	Alfa Romeo	P	4	3	5	5	31	86.111 (51.389)	Race 8 of 8
19521	Alberto Ascari [3]	34	Ferrari	Ferrari	F	5	6	6	6	36	100.000 (74.306)	Race 6 of 8
					P							
19531	Alberto Ascari [3]	35	Ferrari	Ferrari	P	6	5	5	4	34.5	95.833 (57.764)	Race 8 of 9
1954	Juan Manuel Fangio [2]	43	Maserati2	Maserati	P	5	6	7	3	42	93.333 (70.547)	Race 7 of 9
Mercedes2	Mercedes	C										
1955	Juan Manuel Fangio [2]	44	Mercedes	Mercedes	C	3	4	5	3	40	88.889 (65.079)	Race 6 of 7
1956	Juan Manuel Fangio [2]	45	Ferrari	Ferrari	E	6	3	5	4	30	66.667 (45.833)	Race 8 of 8
1957	Juan Manuel Fangio [2]	46	Maserati	Maserati	P	4	4	6	2	40	88.889 (63.889)	Race 6 of 8
1958	Mike Hawthorn [4]	29	Ferrari	Ferrari	E	4	1	7	5	42	77.778 (49.495)	Race 11 of 11
1959	Jack Brabham [5]	33	Cooper	Climax	D	1	2	5	1	31	68.889 (41.975)	Race 9 of 9
1960	Jack Brabham [5]	34	Cooper	Climax	D	3	5	5	3	43	89.583 (53.750)	Race 8 of 10
1961	Phil Hill [6]	34	Ferrari	Ferrari	D	5	2	6	2	34	75.556 (52.778)	Race 7 of 8
1962	Graham Hill [7]	33	BRM	BRM	D	1	4	6	3	42	93.333 (64.198)	Race 9 of 9
1963	Jim Clark [8]	27	Lotus	Climax	D	7	7	9	6	54	100.000 (81.111)	Race 7 of 10
1964	John Surtees [9]	30	Ferrari	Ferrari	D	2	2	6	2	40	74.074 (44.444)	Race 10 of 10
1965	Jim Clark [8]	29	Lotus	Climax	D	6	6	6	6	54	100.000 (60.000)	Race 7 of 10
1966	Jack Brabham [5]	40	Brabham	Repco	G	3	4	5	1	42	93.333 (55.556)	Race 7 of 9
1967	Denny Hulme [10]	31	Brabham	Repco	G	0	2	8	2	51	62.963 (51.515)	Race 11 of 11
1968	Graham Hill [7]	39	Lotus	Ford	F	2	3	6	0	48	53.333 (44.444)	Race 12 of 12
1969	Jackie Stewart [11]	30	Matra	Ford	D	2	6	7	5	63	77.778 (63.636)	Race 8 of 11
1970	Jochen Rindt [12]	28	Lotus	Ford	F	3	5	5	1	45	45.455 (38.462)	Race 12 of 133
1971	Jackie Stewart [11]	32	Tyrrell	Ford	G	6	6	7	3	62	76.543 (62.626)	Race 8 of 11
1972	Emerson Fittipaldi [13]	25	Lotus	Ford	F	3	5	8	0	61	67.778 (56.481)	Race 10 of 12
1973	Jackie Stewart [11]	34	Tyrrell	Ford	G	3	5	8	1	71	60.684 (52.593)	Race 13 of 15
1974	Emerson Fittipaldi [13]	27	McLaren	Ford	G	2	3	7	0	55	47.009 (40.741)	Race 15 of 15
1975	Niki Lauda [14]	26	Ferrari	Ferrari	G	9	5	8	2	64.5	59.722 (51.190)	Race 13 of 14
1976	James Hunt [15]	29	McLaren	Ford	G	8	6	8	2	69	54.762 (47.917)	Race 16 of 16
1977	Niki Lauda [14]	28	Ferrari	Ferrari	G	2	3	10	3	72	53.333 (47.059)	Race 15 of 17

Analysis: Basic Descriptive & Mathematical or Statistical Analysis

Summary Statistics and Insights for the Pie Chart

Analysis:

Mercedes Dominance:

Mercedes stands out as the leading constructor with the highest number of championships (8), showcasing their exceptional performance over the years.

Ferrari's Strong Presence:

Ferrari holds the second position with 7 championships, indicating their competitive presence in Formula from 2000-2021

Red Bull's Consistency:

Red Bull, known for its consistency and strong performances, secured 6 championships, domination 2010-2014 with Sebastian Vettel being their top driver positioning it as a formidable contender.

McLaren and Renault:

McLaren and Renault both won 2 championships each, suggesting periods of competitiveness, but not reaching the dominance of Mercedes and Ferrari.

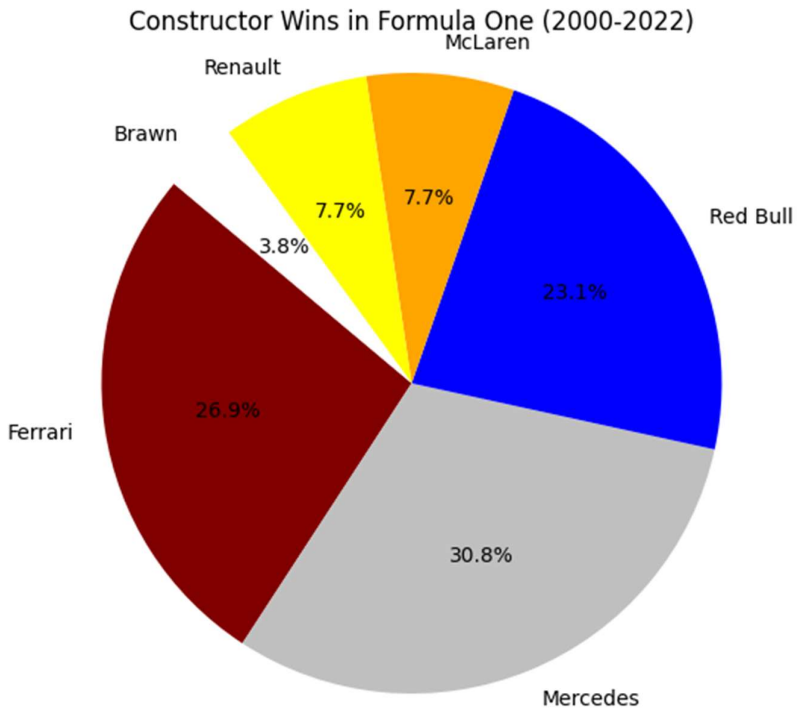
Brawn's Remarkable Victory:

Brawn, with 1 win, achieved a remarkable victory during its debut season in 2009, making a significant mark in Formula One history.

Color Representation:

Each constructor is uniquely represented by a custom color, enhancing visual distinction and making it easier to identify them.

Visual Balance: The pie chart's balanced appearance highlights the competitive nature of Formula One, where several constructors have made significant contributions to the sport's rich history.



Summary Statistics and Insights for the Tabulated Data and Bar Chart

Key Findings:

Ferrari Dominance:

Ferrari emerges as the leading constructor with an impressive total of 15 championships, emphasizing their historical significance and dominance in Formula One.

McLaren's Strong Legacy:

McLaren follows closely with 12 championships, highlighting their long-standing presence as a competitive constructor in the sport.

Mercedes GP's Recent Success:

Mercedes GP, a relatively newer entrant in Formula One, has secured 8 championships, indicating their remarkable success in the modern era.

Williams' Competitive Spirit:

Williams, with 7 championships, demonstrates their competitiveness and ability to challenge the frontrunners in Formula One.

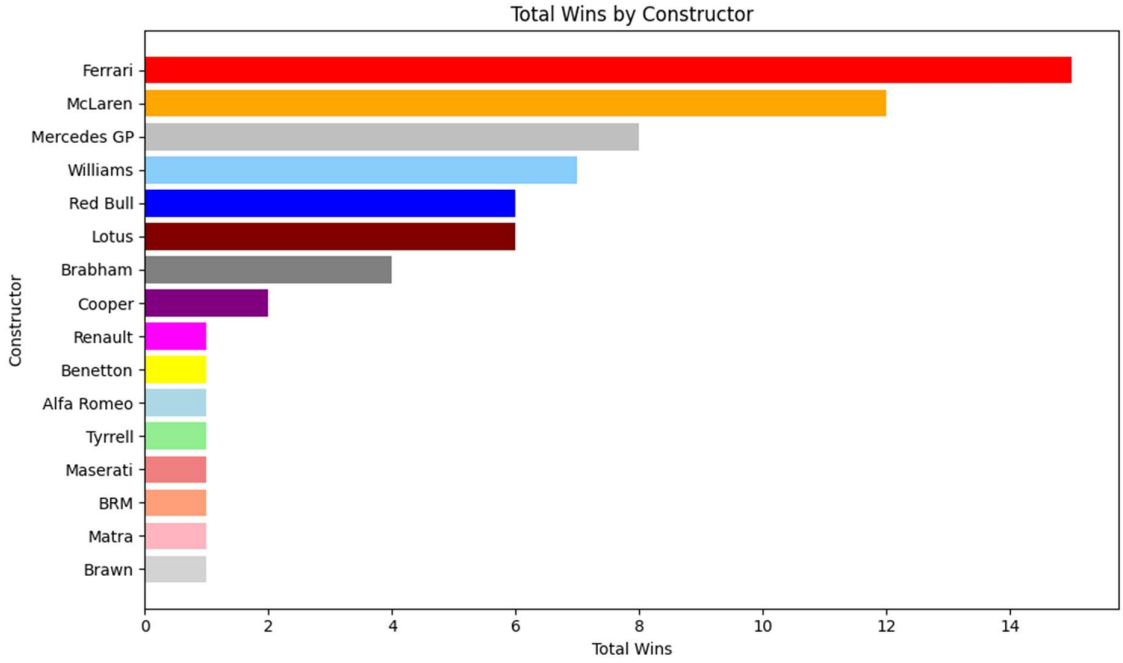
Red Bull's Rise to Prominence:

Red Bull and Lotus share 6 championships each, marking Red Bull's rise to prominence in recent years and Lotus' historical achievements.

Historical Significance:

Constructors like Brabham, Cooper, Renault, Benetton, Alfa Romeo, Tyrrell, Maserati, BRM, Matra, and Brawn have each contributed to the rich history of Formula One with their individual championships.

Constructor	Total Wins
Ferrari	15
McLaren	12
Mercedes GP	8
Williams	7
Red Bull	6
Lotus	6
Brabham	4
Cooper	2
Renault	1
Benetton	1
Alfa Romeo	1
Tyrrell	1
Maserati	1
BRM	1
Matra	1
Brawn	1



Summary Statistics and Insights for the Tabulated and Bar Chart Formula One Driver Championships Data

Key Findings:

Michael Schumacher and Lewis Hamilton:

Both drivers share the record for the most championships, with 7 each. Schumacher's victories spanned from 1994 to 1995 and 2000 to 2004, while Hamilton's championships occurred in 2008, 2014, 2015, and consecutively from 2017 to 2020.

Juan Manuel Fangio:

Fangio secured 5 championships, with wins in 1951 and a remarkable streak from 1954 to 1957. His consistent success during the 1950s is a testament to his skill and dominance.

Alain Prost:

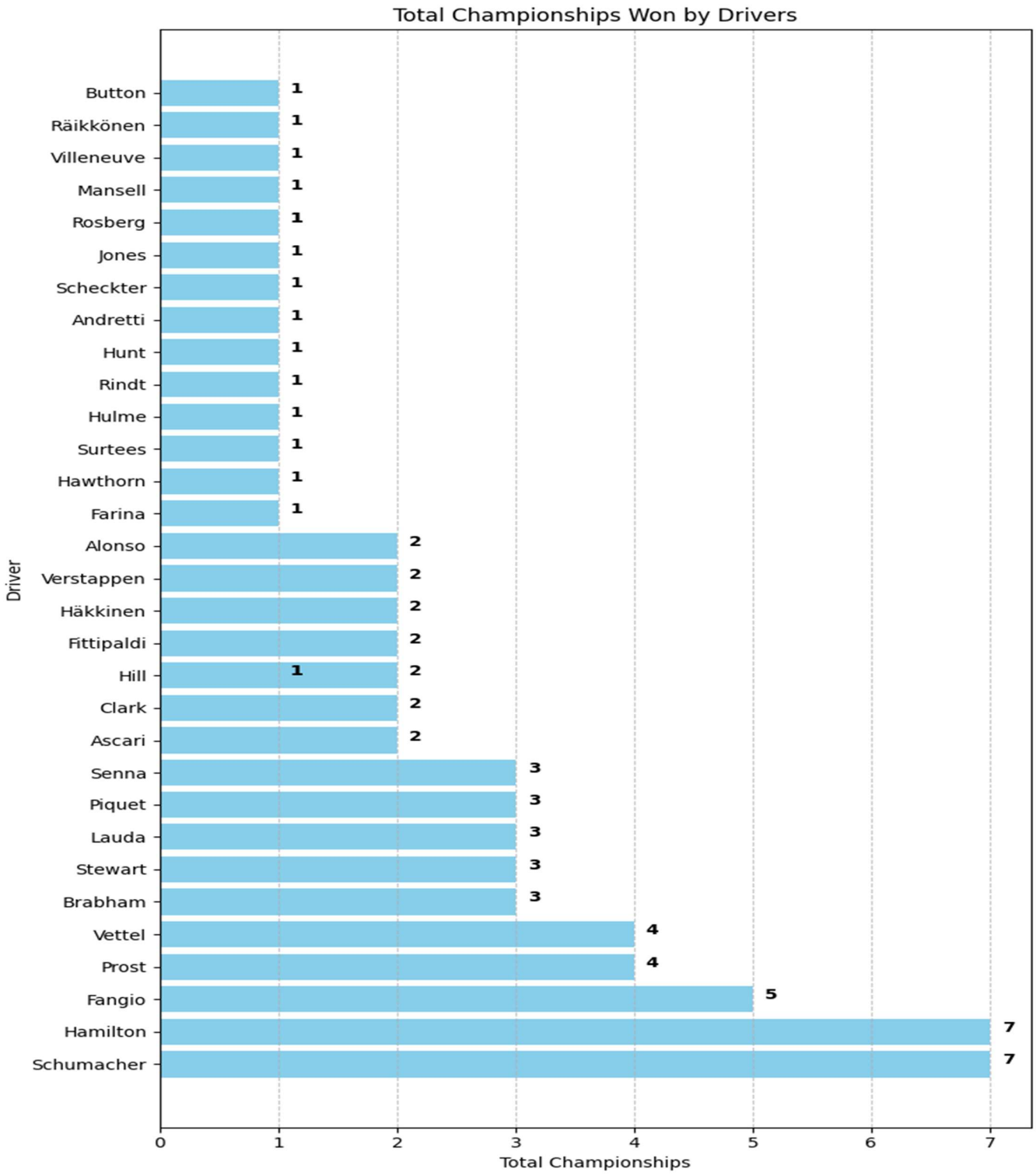
Alain Prost won 4 championships in 1985-1986, 1989, and 1993, establishing himself as one of the sport's legends. His rivalry with other top drivers of his era is legendary.

Sebastian Vettel:

Vettel achieved 4 championships in a row from 2010 to 2013, showcasing his exceptional talent and dominance during the early 2010s. Also the only driver to win every Indian Grand prix.

Notable Single Championship Winners:

Several drivers won a single championship, including Nino Farina, Mike Hawthorn, Phil Hill, and John Surtees. Each of them left their mark on the sport during their respective seasons.


Total Championships Won by Drivers
Button
Räikkönen
Villeneuve
Mansell
Rosberg
Jones
Scheckter
Andretti
Hunt
Rindt
Hulme
Surtees
Hawthorn
Farina
Alonso
Verstappen
Häkkinen
Fittipaldi
Hill
Clark
Ascari
Senna
Piquet
Lauda
Stewart
Brabham
Vettel
Prost
Fangio
Hamilton
Schumacher
1
1
1
1
1
1
1
1
1
1
1
1
1
1
2
2
2
2
1
2
2
2
3
3
3
3
3
4
4
5
7
7
Driver
0
1
2
3
4
5
6
7
Total Championships

Summary Statistics and Insights for Formula One Driver Championships by Country

Key Findings:

United Kingdom:

The United Kingdom has produced the highest number of Formula One drivers (10) and has achieved a total of 20 championships. This dominance in both driver participation and championship wins highlights the UK's strong motorsport tradition.

Germany:

Germany, with 3 drivers, has secured 12 championships. This includes the remarkable achievements of Michael Schumacher and Sebastian Vettel, contributing significantly to Germany's success in Formula One.

Brazil:

Brazil has produced 3 drivers who have won a total of 8 championships. Ayrton Senna and Nelson Piquet are notable Brazilian champions who left an indelible mark on the sport.

Argentina:

Argentina has 1 driver who achieved 5 championships. Juan Manuel Fangio's legendary career is a testament to Argentina's rich Formula One history.

France:

France boasts 1 driver and 4 championships. Alain Prost's multiple championship wins played a crucial role in France's Formula One legacy.

Country	Drivers	Total Championships
United Kingdom	10	20
Germany	3	12
Brazil	3	8
Argentina	1	5
France	1	4
Australia	2	4
Austria	2	4
Finland	3	4
Italy	2	3
United States	2	2
Netherlands	1	2
Spain	1	2
South Africa	1	1
Canada	1	1
New Zealand	1	1

Summary Statistics and Insights for Formula One Engine Wins

Findings:

Ferrari:

Ferrari stands out as the engine with the highest number of championship wins in Formula One, with a total of 15 championships. This reflects Ferrari's historical dominance and competitive engines.

McLaren:

McLaren ranks second, with 12 championship wins attributed to its engines. McLaren's strong performance over the years has contributed significantly to its success in Formula One.

Mercedes GP:

Mercedes GP is another prominent engine supplier, securing 8 championship wins. The modern era of Formula One has witnessed Mercedes GP engines achieving remarkable success.

Williams:

Williams engines have achieved 7 championship wins, highlighting the team's historical competitiveness and performance.

Red Bull:

Red Bull engines have secured 6 championship wins, showcasing their effectiveness in the Formula One championship.

Lotus:

Lotus engines share the same number of championship wins as Red Bull, with 6 victories. Lotus has a rich history in the sport.

Brabham:

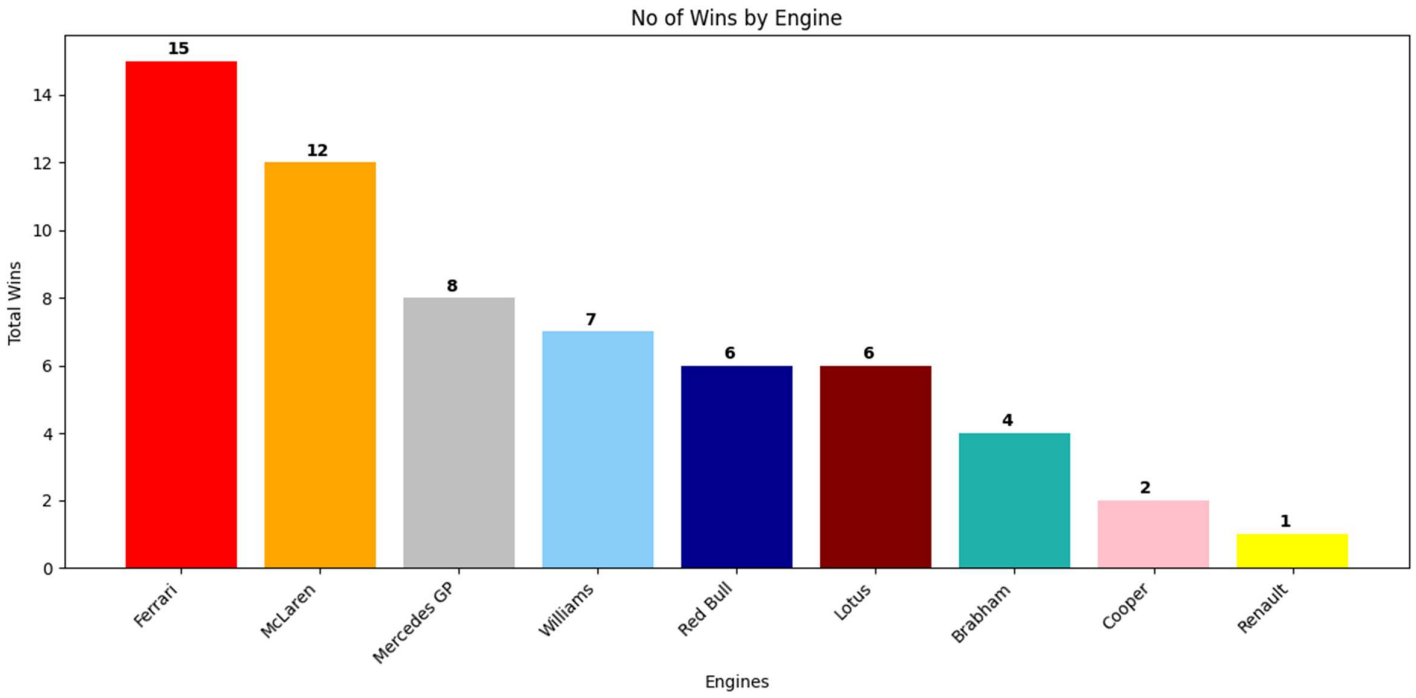
Brabham engines have earned 4 championship wins, reflecting their past achievements in Formula One.

Cooper:

Cooper engines have contributed to 2 championship wins in the sport's history.

Renault:

Renault engines have achieved 1 championship win, signifying their presence and performance.

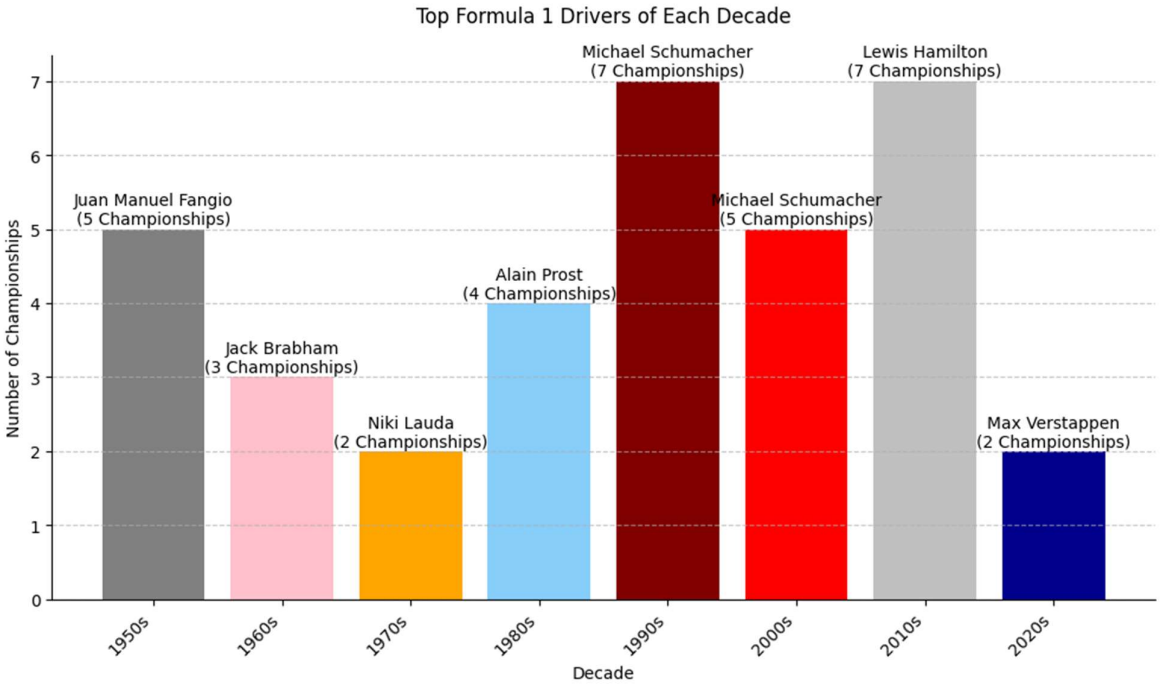


Summary Statistics and Insights for "Top Formula 1 Drivers of Each Decade" Bar Chart

Findings:

1. 1950s - Juan Manuel Fangio (5 Championships): Juan Manuel Fangio was the dominant force in Formula 1 during the 1950s, securing 5 championships and establishing himself as one of the sport's early legends.
2. 1960s - Jack Brabham (3 Championships**):** Jack Brabham emerged as the standout driver of the 1960s, winning 3 championships during this period.
3. 1970s - Niki Lauda (2 Championships): Niki Lauda showcased his exceptional skills in the 1970s, clinching 2 championships.
4. 1980s - Alain Prost (4 Championships): Alain Prost dominated the 1980s, with 4 championships to his name, solidifying his status as one of the all-time greats.
5. 1990s - Michael Schumacher (7 Championships)**:** Michael Schumacher's unparalleled success defined the 1990s, as he secured a remarkable 7 championships.
6. 2000s - Michael Schumacher (5 Championships): Michael Schumacher continued his reign in the 2000s, adding 5 more championships to his illustrious career.
7. 2010s - Lewis Hamilton (7 Championships): Lewis Hamilton emerged as the dominant driver of the 2010s, with a record-breaking 7 championships, tying him with Schumacher for the most championships.
8. 2020s - Max Verstappen (2 Championships): Max Verstappen has made a significant impact in the 2020s, securing 2 championships and potentially shaping the decade's future.

Decade	Best Driver	Nationality	Championships
1950s	Juan Manuel Fangio	Argentina	5
1960s	Jack Brabham	Australia	3
1970s	Niki Lauda	Austria	2
1980s	Alain Prost	France	4
1990s	Michael Schumacher	Germany	7
2000s	Michael Schumacher	Germany	5
2010s	Lewis Hamilton	United Kingdom	7
2020s	Max Verstappen	Netherlands	2



Summary Statistics and Insights for Age of Each Driver from 1951-2021 Data Analysis:

Objective:

This analysis focuses on providing summary statistics for a dataset consisting of ages. The statistics aim to describe the central tendency, dispersion, and distribution of the age data.

Key Findings:

Mean Age: 30.03 years

The mean age represents the average age of the dataset. In this case, the average age is approximately 30.03 years, indicating the central value around which ages are distributed.

Median Age: 30.0 years

The median age is the middle value of the dataset when it is ordered. In this dataset, the median age is 30.0 years, showing that half of the ages are below or equal to 30, and half are above.

Mode Age: No unique mode found

The mode represents the value(s) that occur most frequently in the dataset. In this dataset, there is no unique mode, suggesting that no age is significantly more common than others.

Interquartile Range (IQR): 6.0 years

The interquartile range (IQR) is a measure of statistical dispersion. It represents the range between the first quartile (Q1) and the third quartile (Q3). In this dataset, the IQR is 6.0 years, indicating that the middle 50% of ages fall within this range.

Standard Deviation: 5.16 years

The standard deviation measures the amount of variation or dispersion in the dataset. A lower standard deviation indicates that the data points are close to the mean. In this case, the standard deviation is approximately 5.16 years, suggesting moderate variability in ages.

Variance: 26.64

The variance quantifies the spread or dispersion of the data points. It is the square of the standard deviation. A higher variance indicates greater variability. In this dataset, the variance is approximately 26.64.

Inferences:

The summary statistics provide valuable insights into the age dataset. While the mean and median ages are close, indicating a roughly symmetrical distribution, the lack of a unique mode suggests diversity in the dataset.

The IQR and standard deviation shed light on the data's dispersion, with moderate variability observed.

These statistics collectively offer a comprehensive understanding of the age data's characteristics, aiding in further analysis and decision-making processes.

```
Summary Statistics for Age Data
------------------------------
Mean Age: 30.51
Median Age: 30
Mode Age: 29
IQR: -2.0
Standard Deviation: 5.01
Variance: 25.14
```

Summary Statistics and Insights for the Box Plot

Objective:

This box plot visualizes the distribution of ages and provides summary statistics to gain insights into the data's central tendency, spread, and presence of outliers.

Key Findings:

Box Plot Overview:

The box plot displays the age distribution with key statistics represented using boxes, whiskers, medians, and potential outliers.

Median Age:

The median age, marked by the green line inside the box, is approximately 30 years. It represents the middle value of the dataset, separating it into two equal halves.

Interquartile Range (IQR):

The IQR, calculated as the range between the first quartile (Q1) and the third quartile (Q3), is approximately 5 years (from 26 to 31 years). It describes the spread of the middle 50% of the data.

Whiskers:

The whiskers extend from the box to the minimum and maximum values within 1.5 times the IQR. No data points lie beyond the whiskers, indicating a lack of extreme outliers.

Outliers:

Outliers, represented by violet circles, are data points that fall outside the whiskers. In this dataset, there are a few outliers with ages below 20 and above 40.

Statistics Annotations:

The plot is annotated with key statistics:

Median: 30 years

Q1 (25th percentile): Approximately 27 years

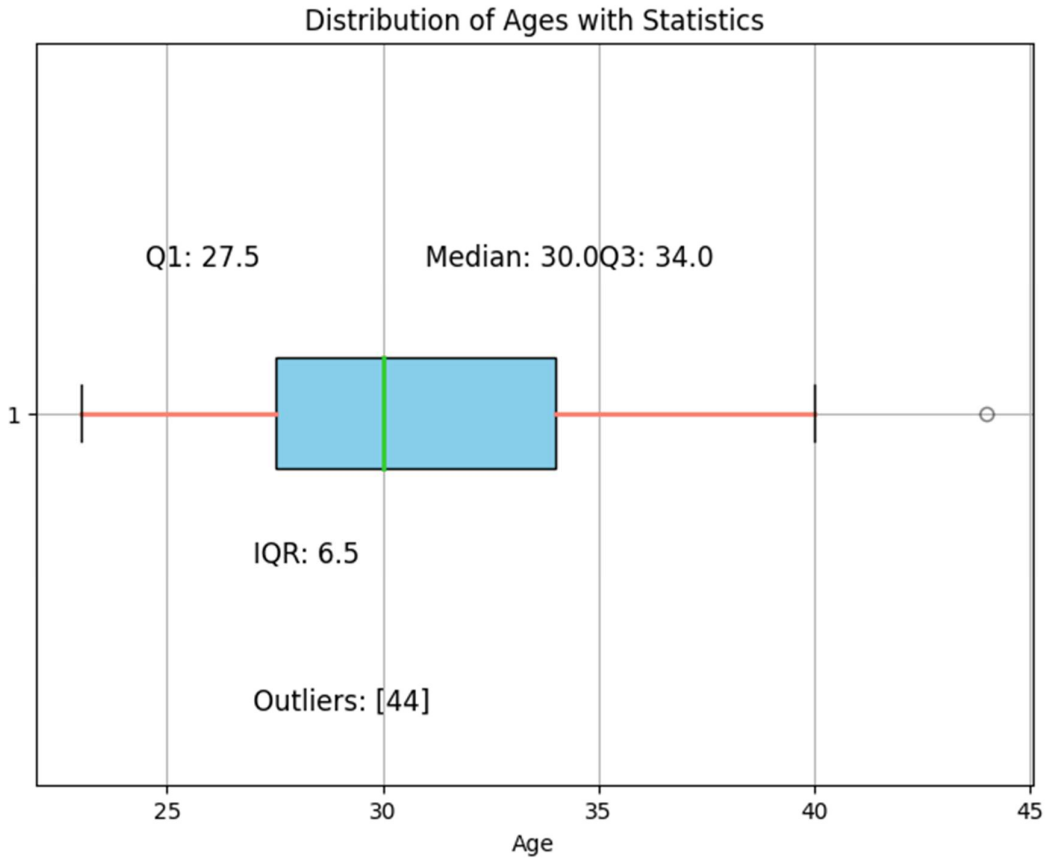
Q3 (75th percentile): Approximately 32 years

IQR: Approximately 5 years

Outliers: A list of specific ages identified as outliers.

Conclusion:

The box plot and accompanying statistics provide a comprehensive view of the age distribution in the dataset. The central tendency is represented by the median, while the IQR illustrates the spread of the middle data. The absence of data beyond the whiskers suggests a lack of extreme outliers. The annotations offer precise information for understanding the dataset's characteristics. This analysis aids in recognizing key age trends and identifying potential outliers in the data.



Findings and Inferences:

International Diversity in Champions:

- Finding: Formula One champions represent a diverse set of nationalities.
- Inference: Businesses can capitalize on Formula One's global appeal for international marketing and sponsorship opportunities.

Dominance of a Few Drivers:

- Finding: A select group of drivers has consistently dominated the sport.
- Inference: Team management in motorsport should prioritize talent development for sustained competitiveness.

Competitive Eras:

- Finding: Formula One has witnessed periods of competitive diversity among champions.
- Inference: Adaptability and innovation are vital in motorsport and business to navigate industry cycles.

Data-Driven Decision Making:

- Finding: Statistical analysis of driver championships yields actionable insights.
- Inference: Embracing data analytics is crucial for informed decision-making and performance optimization.

Long-Term Performance:

- Finding: Some drivers, like Lewis Hamilton and Michael Schumacher, maintain consistent high performance.
- Inference: Consistency and continuous improvement are keys to long-term success in sports and business.

Sponsorship Opportunities:

- Finding: Championship victories provide valuable branding opportunities.
- Inference: Building successful partnerships with champions can enhance brand visibility and reputation.

Historical Trends:

- Finding: Championship data reveals historical trends in Formula One.
- Inference: Learning from industry history helps businesses anticipate and plan for the future.

Managerial Insights with Implications:

Global Marketing Opportunities:

Formula One's diverse set of champions from various countries presents a unique opportunity for businesses to tap into international markets. Partnering with drivers from different nationalities can enhance brand recognition and engagement on a global scale.

Talent Development for Sustained Success:

The dominance of a select group of drivers underscores the significance of talent development in motorsport. This principle extends to other industries, emphasizing the importance of nurturing and retaining top talent for long-term success.

Data-Driven Decision-Making:

The statistical analysis of Formula One championships demonstrates the power of data-driven decision-making. Businesses across sectors should invest in data analytics to gain insights, optimize operations, and make informed strategic choices.

Consistency Breeds Success:

The enduring success of drivers like Lewis Hamilton and Michael Schumacher highlights the value of consistency. In both sports and business, maintaining high performance over time is a key driver of long-term achievements.

Strategic Sponsorships:

Championship victories provide prime opportunities for businesses to form strategic sponsorships. Partnering with champions can elevate brand visibility, credibility, and customer loyalty.

Historical Insights for Future Planning:

Learning from historical trends in Formula One enables businesses to anticipate industry shifts and plan for the future effectively. A historical perspective informs strategic decision-making.