Data Set Title
Exploratory Analysis

Nick Harshaw, nharshaw@bellarmine.edu
Ellie Landoch, elandoch@bellarmine.edu

I. INTRODUCTION

Short description of the data set including a reference to where it can be found and why you chose it.

This data set focuses on career stats for all drivers who have ever raced in Formula 1 from 1950 to 2023. For this project, we filtered the data to focus more on just the race winners. This is because their stats will be more notable to get an understanding of how different drivers perform in the sport as opposed to some of the driver that may have only driven in one or two races.
Link to data set on Kaggle: https://www.kaggle.com/datasets/petalme/f1-drivers-dataset

II. DATA SET DESCRIPTION

Narrative summary of the data set: e.g. this data set contains 113 samples with 22 columns with various data types. A complete listing is shown in **Table 1**. For data types you want to indicate two things (nominal, ordinal, interval, or ratio) and the Pandas data type. For example, age might be ratio/int32. For missing data, indicate what percentage of data from that column are missing. Ensure you check to for NaN, NA, or any other indicators that actually mean missing data.

Data set contains 113 samples with 22 columns with various data types.

Table 1: Data Types and Missing Data

Variable Name	*Data Type*	*Missing Data (%)*
Driver	Nominal/object	0%
Nationality	Nominal/object	0%
Seasons	Ordinal/object	0%
Championships	Ratio/float64	0%
Race_Entries	Ratio/float64	0%
Race_Starts	Ratio/float64	0%
Pole_Positions	Ratio/float64	0%
Race_Wins	Ratio/float64	0%
Podiums	Ratio/float64	0%
Fastest_Laps	Ratio/float64	0%
Points	Ratio/float64	0%
Active	Nominal/bool	0%
Championship Years	Ordinal/object	59.4%
Decade	Ordinal/int64	0%
Pole_Rate	Ratio/float64	0%
Start_Rate	Ratio/float64	0%
Win_Rate	Ratio/float64	0%
Podium_Rate	Ratio/float64	0%
FastLap_Rate	Ratio/float64	0%
Points_Per_Entry	Ratio/float64	0%
Years_Active	Ordinal/int64	0%
Champion	Nominal/bool	0%

III. Data Set Summary Statistics

Narrative introduction to the section.

This data set has statistics about every F1 race winner such as their name, nationality, number and rate of starts, wins, podiums, fastest laps, and points, their active status, and whether or not they've won any championships.

Table 2: Summary Statistics for XXX (name of dataset)

Variable Name	*Count*	*Mean*	*Standard Deviation*	*Min*	*25th*	*50th*	*75th*	*Max*
Championships	113	0.65	1.33	0	0	0	1	7
Race_Entries	113	120.75	86.23	3	52	109	165	359
Race_Starts	113	118.38	85.61	2	49	108	163	356
Pole_Positions	113	9.27	15.38	0	0	3	13	103
Race_Wins	113	9.58	15.67	1	1	4	11	103
Podiums	113	26.81	31.41	1	7	18	33	191
Fastest_Laps	113	9.20	12.36	0	1	5	13	77
Points	113	381.15	645.45	8	71	180	329	4415.5
Pole_Rate	113	.07	.10	0	0	.03	.11	.56
Start_Rate	113	.97	.06	.67	.97	.99	.99	1
Win_Rate	113	.08	.09	.004	.02	.05	.10	.46
Podium_Rate	113	.23	.16	.01	.11	.20	.29	.86
FastLap_Rate	113	.08	.09	0	.02	.05	.11	.50
Points_Per_Entry	113	2.49	2.30	.20	1.17	1.81	3	14.2
Years_Active	113	9.84	3.97	2	7	9	12	19

There should be a table for **EACH** categorical variable.

Table 3: Proportions for XXX (n=yyy)

Nationality	*Frequency*	*Proportion (%)*
United Kingdom	20	17.7%
Italy	15	13.3%
United States	15	13.3%
France	14	12.4%
Brazil	6	5.3%
Finland	5	4.4%
Germany	5	4.4%
Australia	4	3.5%
Sweden	3	2.7%
Austria	3	2.7%
Argentina	3	2.7%
Countries w/ 2 (count: 7)	2	1.8%
Countries w/ 1 (count: 6)	1	0.9%

Active	*Frequency*	*Proportion (%)*
False	103	91.2%
True	10	8.8%

Decade	*Frequency*	*Proportion (%)*
1950	13	11.5%
1960	20	17.7%
1970	15	13.3%
1980	23	20.4%
1990	8	7.1%

2000	14	12.4%
2010	11	9.7%
2020	9	8.0%

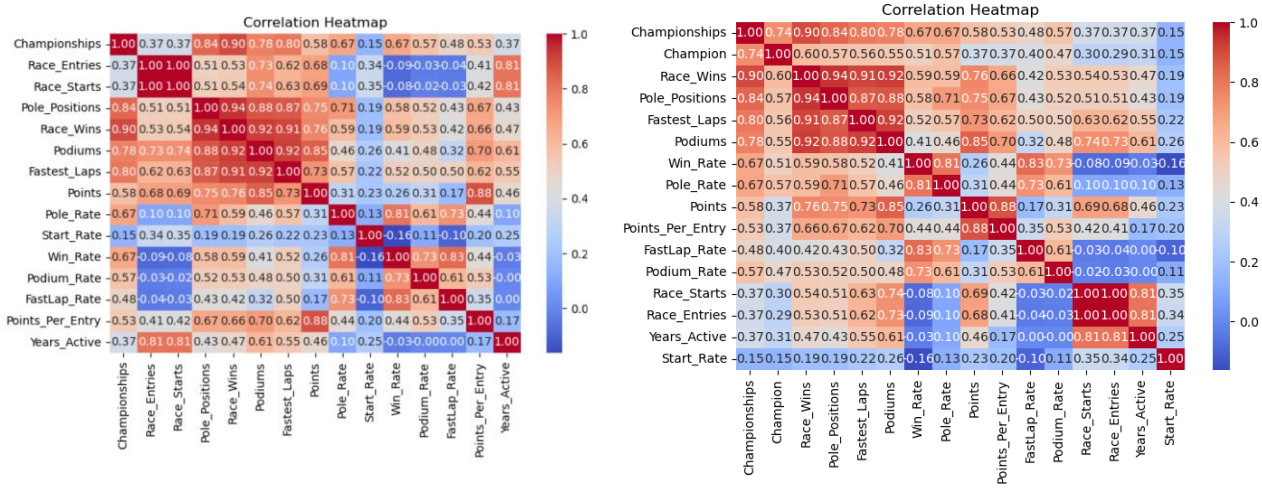
Champion	*Frequency*	*Proportion (%)*
False	79	69.9%
True	34	30.1%

After you summarize the categorical variables, generate a correlation matrix for all continuous variables (not categorical – this doesn't make sense)

Table 4: Correlation Table/Tables

	Championships	**Champion**
Championships	1	0.744218
Champion	0.744218	1
Race_Wins	0.898141	0.597453
Pole_Positions	0.840856	0.57172
Fastest_Laps	0.795553	0.556616
Podiums	0.775573	0.547697
Win_Rate	0.667412	0.508742
Pole_Rate	0.671982	0.574039
Points	0.578689	0.369397
Points_Per_Entry	0.53238	0.368874
FastLap_Rate	0.484186	0.398755
Podium_Rate	0.568407	0.474437
Race_Starts	0.370001	0.295445
Race_Entries	0.366162	0.294751
Years_Active	0.368481	0.314657
Start_Rate	0.149499	0.145798

After the table with the raw data, include a heatmap of the correlation matrix as a figure.



IV. DATA SET GRAPHICAL EXPLORATION

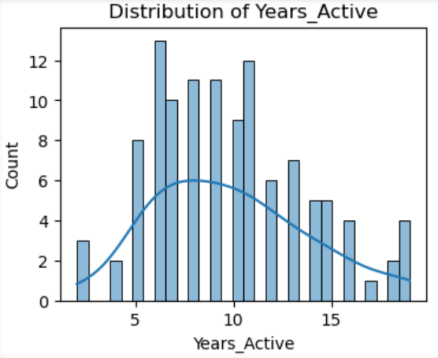
Narrative introduction to the section. In each section below, indicate any interesting distributions, anomalies, imbalance, etc. that you notice.

A. *Distributions*
B. *Scatterplots / Pairwise Plots (continuous variables)*
C. *BarCharts (categorical variables)*
D. *Other Plots - don't skimp – there are likely other plots that would be useful that I haven't already specified. Include those in this section.*

All figures should be cited formatted like this and mentioned in the text.

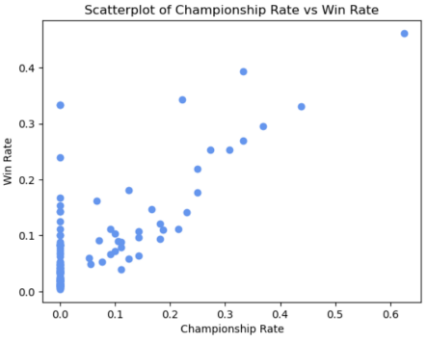
****See all graphs in the .ipynb file (as well as analysis of the data found)****

Histogram: Distribution of how many years drivers are active in the sport



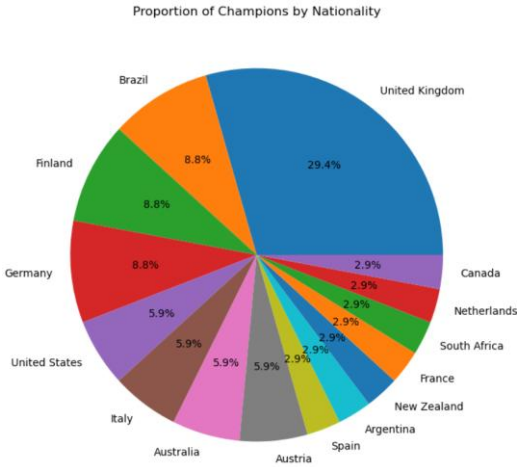
Summary – Most race-winning drivers have decently long careers in the sport due to their successes in their races. Based on the histogram, most drivers that win races spend 7-12 years driving in F1.

Scatter plot: Championship rate (# of championships/# of years active) vs. win rate



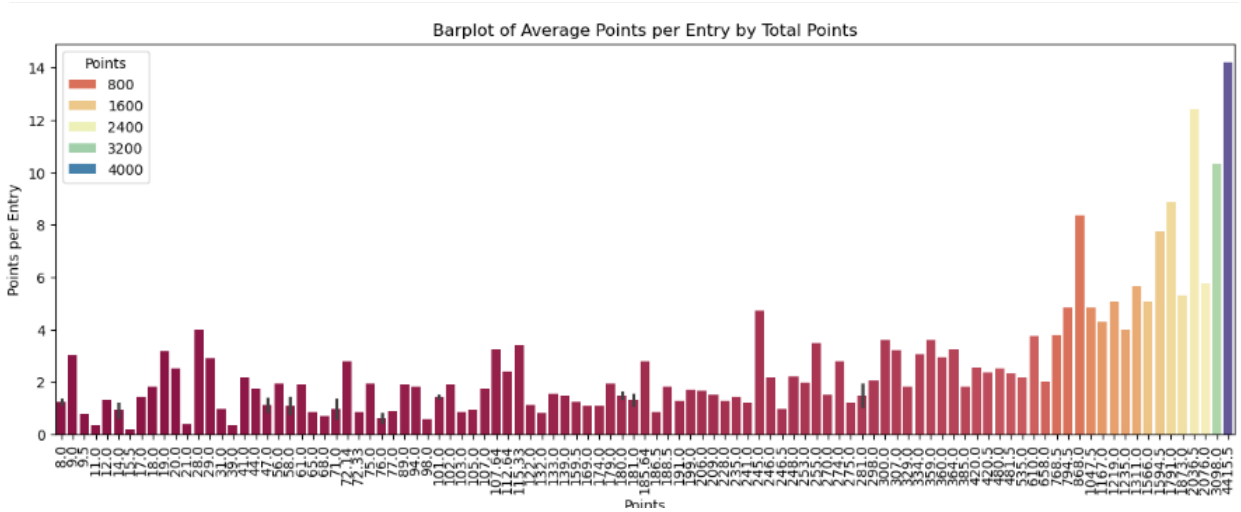
Summary – If we only focus on the drivers that have a championship rate > 0, we can see the trend that as a driver's championship rate increases, their probability of winning a each of their races also increases. The drivers with the highest win rates are much more likely to become world champions.

Pie graph: % of world champions per country



Summary – this graph gives us an idea of the distribution of world champion titles among different countries. This data can also be compared to the charts of the number of race-winning drivers from each country and how many wins have been scored by each of them. This data reveals which countries produce the most successful F1 drivers.

Bar Chart: Average # of points scored per race vs. driver's total points



Summary – This graph is skewed to the left. So, this shows how as a driver earns more total points throughout their career, they are also likely to be scoring more points in each individual race on average. This probably means that as a driver gains more experience, they will likely be scoring larger quantities of points more often than the other drivers.

****See all graphs in the .ipynb file (as well as analysis of the data found)****

V. SUMMARY OF FINDINGS

Finish up with a paragraph or two of summarizing your findings about this data set.

Our findings tell us a lot about Formula 1 and its drivers. One of the big things is that there isn't a whole lot of parity. If you look at all the people who have entered a race, the majority haven't won a single race. Out of the 868 total f1 racers only 113 racers have won a race. This leaves a staggering 87% of racers who had an F1 career never leaving a single racetrack as a winner. Going further down the line to podiums, or a top three spot in the race, still only 215 racers have a top three finish. Again, leaving a massive 75% of racers who have never even placed at a single racing event. This leaves a large portion of our data either incomplete or with many amounts of 0 for data making it much harder to draw real conclusions about what truly makes one racer better as most of the statistics are only about what separates the elite from the average. For this reason, we decided to limit our data frame to only racers that have won at least one race. While this doesn't affect our correlation coefficient graphs much as it only changes most of the coefficients but around .02 or .03 it does make a massive difference in our graphs and data truly allowing us to see the difference in data and graphs as to what separates the greats from the elite. (This also helps us get rid of most of the missing values in our dataset.)

After refining the dataset, we used a correlation coefficient table to see how all of data related to each other. First, it is essential to understand that F1 is about winning the world championship. It is the highest honor in the sport and winning or even simply challenging for one can define your career. Using this info, we tried to figure out what is most likely going to lead racers to win a championship (or even multiple). In F1 the person with the most points at the end of the season claims the championship title, and the better you do in races throughout the year the more points you get. Since points are what wins you the championship, one might assume it would have a very high correlation, this however isn't the truth because racers could build up high point totals if they had a long career without ever getting even close to being a champion. As far as our correlation data showed, the most effective way to tell which racer is better, other than the obvious championships, is race wins. This is because of its incredibly high correlation coefficient. The heatmap also shows positive correlations across all variables. Additionally, our univariate histograms reveal that most variables are right-skewed, indicating that higher values are less common among drivers. Bar plots depicting the count of wins, podiums, and points versus their respective rates are left-skewed, highlighting that a small group of drivers achieve higher success rates. Moreover, the data shows that most F1 successes originate from drivers hailing from a select few countries, highlighting the dominance of certain nations in the sport.