

**Data Driven Decision Making**

***Topic: Analysis of Formula 1***

*Adithya Vinod*

*Sneha Kumble Manohara*

**School of Graduate Professional Studies**

MPS/MS in Data Analytics

DAAN 881 – Data Driven Decision Making

(Spring, 2023)

# Document Control

## Work carried out by:

|  |  |  |
| --- | --- | --- |
| **Name** | **Email Address** | **Task description** |
| Adithya Vinod | avv5502@psu.edu |  |
| Sneha Kumble Manohara | skm6622@psu.edu |  |
|  |  |  |

## Revision Sheet

|  |  |  |
| --- | --- | --- |
| **Release No.** | **Date** | **Revision Description** |
| 1 | 3/19/2023 |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**TABLE OF CONTENTS**

**General Guidelines**

1. To complete all the homework assignments for this course please use this template document.
2. Each assignment has to be submitting by the following Sunday 11:59 PM EST.
3. Each figure should be followed by a brief description about the figure.
4. The figures can be hand drawn and scanned in some circumstances, but the hand drawn figure should be clear and legible to obtain full credits. Unclear hand drawn figures will receive partial credits. For constructing figures and diagrams it is advised to use tools.
5. Figures and tables should have appropriate captions. For documenting and referencing styles please follow the APA or MLA writing style.
6. Please make sure that you provide a reference section.
7. Any material text or figure taken from books, journals or Internet should be referenced. If you have a sentence or a figure that does not belong (authorship) to you, they need to be clearly referenced. If you fail to do so your report will be considered as a case for plagiarism. It is your duty to make sure that your report is free from any activity related to plagiarism. In case you are suspected of attempting plagiarism then you will be responsible for the cause. The penalty for plagiarism will be a “0” awarded to your report. So, it is good to keep simple, always have the principle to acknowledge people for their contributions.

Please go through the following instructions before submitting the report

#### **Academic Integrity**

Academic integrity — scholarship free of fraud and deception — is an important educational objective of Penn State. Academic dishonesty can lead to a failing grade or referral to the [Office of Student Conduct](http://www.sa.psu.edu/ja/).

Academic dishonesty includes, but is not limited to:

* cheating
* plagiarism
* fabrication of information or citations
* facilitating acts of academic dishonesty by others
* unauthorized prior possession of examinations
* submitting the work of another person or work previously used without informing the instructor and securing written approval
* tampering with the academic work of other students

#### How Academic Integrity Violations Are Handled

In cases where academic integrity is questioned, [procedure requires an instructor to notify a student](http://www.psu.edu/oue/aappm/G-9-academic-integrity.html) of suspected dishonesty before filing a charge and recommended sanction with the college. Procedures allow a student to accept or contest a charge. If a student chooses to contest a charge, the case will then be managed by the respective college or campus Academic Integrity Committee. If a disciplinary sanction also is recommended, the case will be referred to the [Office of Student Conduct](http://www.sa.psu.edu/ja/title=).

All Penn State colleges abide by this Penn State policy, but review procedures may vary by college when academic dishonesty is suspected. Information about Penn State's academic integrity policy and college review procedures is included in the information that students receive upon enrolling in a course.

Additionally, Penn State students are expected to act with civility and personal integrity; respect other students' dignity, rights, and property; and help create and maintain an environment in which all can succeed through the fruits of their own efforts. An environment of academic integrity is requisite to respect for oneself and others, and a civil community.

#### For More Information on Academic Integrity at Penn State

Please see the [Academic Integrity Chart](http://www.campuses.psu.edu/CAO.pdf)  for specific college contact information or visit one of the following URLs:

* Penn State Senate [Policy on Academic Integrity](http://www.psu.edu/dept/oue/aappm/G-9.html)
* [iStudy for Success!](http://istudy.psu.edu/tutorials/) — learn about plagiarism, copyright, and academic integrity through an educational module
* [Turnitin](http://tlt.its.psu.edu/turnitin) a web-based plagiarism detection and prevention system

**Week 1 assignment**

**Research Goal:**

Formula 1 (F1) is a popular motorsport that involves high-speed racing of specially designed cars on tracks. The F1 season consists of a series of races called Grands Prix, held in different countries around the world. F1 is one of the most popular motorsports in the world, with a large following of fans and enthusiasts. As such, there is a great deal of interest in the sport, and analyzing it can help to provide insights and information that fans and stakeholders may find useful.

Analyzing the sport of F1 can provide valuable insights into the strategies and tactics used by teams and drivers during races. This can help fans to better understand the sport, and it can also provide valuable information for researchers and analysts who are interested in studying the performance of drivers and teams.

**Business Questions:**

1. Do geographical factors like altitude and temperature have an effect on F1 engine?
2. Does a racer having Pole position for Qualifying ensure the fact that he would win the race?
3. As a F1 scout, which country should the team invest to find the next F1 Champion?

**Dataset:**

The dataset for the above problem statement can be found on Kaggle. The link to the dataset is:

[Formula 1 World Championship (1950 - 2023) | Kaggle](https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020/code?datasetId=468218&sortBy=voteCount)

**Week 2 assignment**

The dataset consists of all information on the Formula 1 races, drivers, constructors, qualifying, circuits, lap times, pit stops, championships from 1950 till the latest 2023 season.

We are considering 6 tables(drivers, driver\_standings, circuits, results, status, races) for our analysis. A total of 40 columns are taken into consideration.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable Name** | **Data Type** | **Data Classification** | **Description** | **Missisng Values** |
| driverId | Numeric | Interval | Driver ID | 0% |
| driverRef | Numeric | Nominal | Reference name of driver eg. 44 | 0% |
| code | Numeric | Nominal | 3-letter alphabet code of driver eg. HAM | 0% |
| forename | Categorical | Nominal | First name of the driver eg. hamilton | 0% |
| surname | Categorical | Nominal | Last name of the driver eg. Lewis | 0% |
| nationality | Categorical | Nominal | Country eg. British | 0% |
| driverStandingsId | Numeric | Interval | Id | 0% |
| raceId | Numeric | Nominal | Id of race eg. 18 | 0% |
| driverId | Numeric | Nominal | Id of driver eg. 2 | 0% |
| points | Numeric | Nominal | Points | 0% |
| position | Numeric | Nominal | Final position | 0% |
| wins | Numeric | Nominal | Number of wins | 0% |
| circuitId | Numeric | Interval | Id | 0% |
| circuitRef | Categorical | Nominal | Reference name of circuit eg. albert\_park | 0% |
| name | Categorical | Nominal | Actual name of circuit. eg. Albert Park Grand Prix Circuit | 0% |
| location | Categorical | Nominal | City of the circuit eg. Melbourne | 0% |
| country | Categorical | Nominal | Country of the circuit eg. Australia | 0% |
| lat | Categorical | Interval | Latitude eg. -37.8497 | 0% |
| lng | Categorical | Interval | Longitute eg. 144.968 | 0% |
| alt | Categorical | Interval | Altitude eg. 10 | 0% |
| resultId | Numeric | Interval | Id | 0% |
| grid | Numeric | Nominal | Position in starting grid eg. 1 | 0% |
| laps | Numeric | Interval | Lap number eg. 23 | 0% |
| time | DateTime | Time series | Total time | 0% |
| rank | Numeric | Nominal | Final Rank | 0% |
| statusId | Numeric | Interval | Status Id | 0% |
| raceId | Numeric | Interval | Id of race | 0% |
| status | Categorical | Nominal | Status of the race. Eg. Finished, Finished | 0% |
| year | Numeric | Nominal | Year of the race. Eg. 2009 | 0% |
| round | Numeric | Interval | Race number. Eg. 2 | 0% |
| date | DateTime | Time series | Date. Eg. 2009-03-29 | 0% |
| name | Categorical | Nominal | Name of circuit. Eg. Australian Grand Prix | 0% |

**Week 3 assignment**

We conducted a preliminary analysis to check for any error and inconsistencies in the dataset. First, we removed all the unnecessary columns that were not required for our analysis.

As a part of data cleaning, we looked for missing values, duplicate values, outliers and skewness in the desired tables. There seems to be no null values or duplicate values in any of the tables.

**Feature Removal:**

Variables not required for the analysis were dropped from every table.

|  |  |
| --- | --- |
| Table | Variables dropped |
| circuit | url |
| drivers | 'number','url' |
| driver\_standings | 'positionText' |
| results | 'constructorId', 'number', 'positionText', 'position', 'time', 'milliseconds', 'fastestLap', 'rank', 'fastestLapTime','fastestLapSpeed' |
| races | 'url','fp1\_date', 'fp1\_time', 'fp2\_date', 'fp2\_time', 'fp3\_date', 'fp3\_time','quali\_date', 'quali\_time', 'sprint\_date', 'sprint\_time' |

The final tables included the following varialbles.

Text

Description automatically generated

**Outliers in table driver\_standings:**

Points, Position and wins columns had outliers which were detected using the box and whisker plot. Points had 5, position had 39 and wins had 61 outliers. Instead of removing outliers, we will be using Quantile based flooring and capping method, where the data points that are lesser than the 10th percentile are replaced with the 10th percentile value and the data points that are greater than the 90th percentile are replaced with 90th percentile value.**Chart, box and whisker chart

Description automatically generated** **Chart, box and whisker chart

Description automatically generated**

Chart, box and whisker chart

Description automatically generated

**Outliers in table races:**

Table races had no outliers.

Chart

Description automatically generatedChart, histogram

Description automatically generated

**Outliers in table results:**

Grid had 28, points had 2 and laps had 20 outliers.

Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

**Skewness:**

In table driver\_standings we did notice some skewness in the data. To reduce the skewness, we will be implementing box-cox transformation.

Text

Description automatically generated

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Table races did not have a lot of skewness.

A picture containing text

Description automatically generated

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Table results showed some skewness as well.

Text

Description automatically generated

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

**Week 4 assignment**

As a part of data cleaning, we looked for missing values, duplicate values, outliers and skewness in the desired tables. Variables not required for the analysis were dropped from every table. There were no missing values and duplicate values.

**Outliers:**

In order to remove the outliers, we used Quantile based flooring and capping, where the outlier is capped at a certain value above the 99th percentile value or floored at a factor below the 1st percentile value.

* **driver\_standings:**

Chart, histogram

Description automatically generated Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

* **races:** Table races had no outliers.
* **results:**

Chart, histogram

Description automatically generatedChart, box and whisker chart

Description automatically generatedChart, histogram

Description automatically generatedChart

Description automatically generated

**Skewness:**

We have considered a feature to be skewed if the skewness value lies outside the range: [-0.5, 0.5].

Boxcox transformation was implemented to reduce the skewness in the data.

* **driver\_standings:**

Skewness after transformation

Text

Description automatically generated

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

* **races:**

Skewness after transformation

A picture containing text

Description automatically generated

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

* **results:**

Skewness after transformation

Text

Description automatically generated

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

**Week 5 assignment**

We performed several steps like variable selection, transformation, data reduction etc. As a part of preprocessing, we created new features from the existing ones and removed some of the features that were not necessary for our analysis.

To reduce the outlier, we used Quantile based flooring and capping, where the outlier is capped at a certain value above the 99th percentile value or floored at a factor below the 1st percentile value. Box-cox transformation was implemented to reduce the skewness in the data.

Since we are having three different business queries to solve, we will be creating three different datasets by joining the six different tables in different combinations. We have performed PCA for dimensionality reduction. The goal of PCA is to identify the directions of maximum variance in the data and project the data onto a new set of axes, which are the principal components. But the results of the PCA weren’t of use as they create new dimensions as we lose the information related to the columns we initially started with. So we will be manually selecting the variables we will be needing for all the three datasets.

**Data Modelling**

We plan to use regression analysis technique and ANOVA model for our analysis.

The ANOVA model assumes that the independent is categorical and that the dependent variable is continuous and regularly distributed. The objective of an ANOVA is to determine if the means of the dependent variable is the same at all levels of the independent variable. The alternative hypothesis in an ANOVA is that at least one of the means is different, while the null hypothesis is that all the means are equal. This hypothesis is tested using the F-statistic, and the p-value is computed to assess the result’s statistical significance.

Regression analysis is a statistical technique used to model the relationship between a dependent variable and one or more independent variables.

There are several performance measures that are commonly used to evaluate the accuracy of regression models:

1. Mean Absolute Error (MAE): The average absolute difference between the predicted values and the true values.
2. Mean Squared Error (MSE): The average of the squared differences between the predicted values and the true values.
3. Root Mean Squared Error (RMSE): The square root of the mean squared error, which provides a more interpretable result in the same units as the response variable.
4. R-squared (R2): A measure of how well the model fits the data, with a value close to 1 indicating a good fit.
5. Mean Absolute Percentage Error (MAPE): The average absolute difference between the predicted values and the true values, expressed as a percentage of the true values.

The metric chosen for evaluation will be the R-square, a statistical measure that represents the proportion of the variance of a dependent variable that is explained by an independent variable or variables in regression problems. Thus, the larger the R², the more explanatory the linear model is, that is, the better it fits the sample.

**Week 6 assignment**

This week the focus will be on data modeling. You will have to update the template document with the following:

* Describe all the steps performed for data modelling.
* Discuss about the performance measure(s) used to determine the goodness of fit for the proposed model.
* Do you suspect your initial effort (model) suffers overfitting? If so, discuss what steps you took to overcome overfitting.
* Discuss about your findings or inferences for each of the research/ business queries you identified in week 1 assignment.

Submit your completed week 6 assignment to the dropbox in Canvas.

**Week 7 assignment**

**Purpose:**

To provide a demonstration of your team’s project in this course

**Tasks:**

1. Before making the final submission make sure that you have implemented all my feedbacks from week 1 to week 6.
2. Please ensure that the title page, document control section, table of contents and the reference section is updated and is up-to-date. Any appendix section should also be updated.
3. The team should prepare to demonstrate the designed data-driven Analytics System in the class (if determined by the instructor).
4. The demonstration should include a power point presentation and/or demonstration.
5. Every team will have an opportunity to go through the demonstration of the data-driven Analytics system designed by other teams.
6. The team should submit the template document, dataset(s), code/script(s) and the power point presentation deck in the box account set up by the instructor. Details about the box account will be provided in the class or by e-mail.
7. Team members should make all submissions latest by April 27, 2023, at 11:59 PM EST.
8. More instructions (if needed) will be provided by the instructor over e-mail or in class.

Submit your completed week 7 assignment to the dropbox in Canvas.