## línea horizontal

Icon

Description automatically generated

Individual Delivery

22/12/2020 to 10/08/2021

Javier Gil Antuñano Foncillas

Data Science

The Bridge

# General Vision

Formula 1 has been around for 70 years, in those 70 years technology has evolved, safety has been made a priority, science has become an art, and drivers have evolved. In these 70 years formula one has also affected the daily life of all humanity.

These radical changes make this topic exciting to study, if all the evolutions are considered and there is still a recognizable pattern this will mean two main things:

1st That formula one still conserves the essential value that created it.

2nd That changing ever environments have some constant with in them.

# Goals

The goal to look for and identify the constant patterns that make a world champion. If such patterns exist, the secondary goal would be to be able to predict the who the following champion would be.

The goals would be to achieve an A+

The goal of the project is to determine in this ever-changing sport there is a pattern that defines a world champion, there are certain constant characteristic that world champions have, and if there is one can it predicts future world champions.

# Specifications

1. Document all steps. Structure your code to keep it cleaned using good practices.

2. Use Trello

3. Collect the data. Try to do each call, it collects the last updated data.

4. Determine and explain if the data is cleaned. If not, then clean it.

5. Show different tendencies for each column in your dataset.

6. Represent, in a pie chart, the time you needed for each point in the project.

steps section.

7. Answer the questions:

a. Was it possible to demonstrate the hypothesis? Why?

b. What can you conclude about your data study?

c. What would you change if you need to do another EDA project?

8. Show the histogram of each column of your dataset with bins =5. How are the ranges painted.

9. Which are the columns with the highest correlation Draw the correlation matrix.

10. Use Matplotlib functions to show all graphs. No pandas directly.

11. Research to save each plot in local files.

12. Use distribute modules for each functionality. The jupyter notebooks must not have any loop or functions. It only must have the initials imports and the call to necessary functions.

13. Apart from matplotlib, use seaborn to show the graphs.

14. Answer the questions:

a. Are their outliers or some rare data

b. What are the columns that have more repeated values.

d. What do you learn doing this project?

## Software

As it is a software program, we must specify the minimal software that we need to have to execute the software program. If it is multiplatform, we must specify the software for each OS.

For this project, the software programs that have been used are the following:

For the coding:

* The computer runs in Windows 10.
* Python 3.0
* Jupiter 2.3.3
* Visual Studio Code
* And the following python libraries:
  + Numpy
  + Pandas
  + Seaborn
  + Matplotlib
  + Geopy
  + Plotly

For the presentation side of the project:

* PowerPoint

## Hardware

As it is a software program, we must specify the minimal hardware that we need to have to execute the software program.

The Hardware used for this project was a:

* Windows Surface Pro 6:
  + with 8 gigs of RAM.

## Requirements

1. The project must give an answer to a hypothesis (explained below).

2. The student will do a presentation and must document all steps he/she does.

3. The student must divide the tasks that he/she has to do.

4. It is mandatory for the student to use trello (or other related) to manage the tasks in

different status: TODO, DOING and DONE. BACKLOG and REVIEW are optional.

5. The delivery must be sent before 08/01/2021 at 23:59.

6. The delivery must be sent in a .zip file by email/classroom with this structure:

a. A folder src/ that contains all the source code.

b. A folder documentation/ that contains all the documents related to

documentation.

c. A folder resources/ that contains other useful content.

d. A folder reports/ that contains all related to created reports such as figures,

html, pdf, etc

e. A folder notebooks/ that contains notebooks for your tests.

f. A folder src/utils/ that contains all the modules used by the main file.

g. A file src/main.ipynb that contains all the functionality. This file must only

contain imports, pandas, matplotlib, requests, and calls to your src/utils/\*

modules.

h. There are, at least, these modules inside src/utils/ :

i. “folders\_tb.py” that contains the generic functionality related to open,

create, read, and write files.

ii. “visualization\_tb.py” that contains the generic functionality related to

pandas, matplotlib, seaborn and other libraries focus on

visualizations.

iii. “mining\_data\_tb.py” that contains the generic functionality related to

collect data, clean data, and others (wrangling methods such us

working with multiples jsons)

# Steps

## Research the context

I research the knowledge of my subject and found that Formula 1 (F1) is a sport highly dependent on Data, as stated previously it is also a sport that has evolved in the last 70 years.

The long duration and the dependence on Data are a very interesting combination that provides the Data scientist with the tools that it needs to study the hypothesis.

## Get Data

The Data can be found inn Kaggle. In the site are a variety of Datasets you could opt for, but for this project and to simplify this step, the optimal Dataset can be found in the following link: <https://www.kaggle.com/aadiltajani/fia-f1-19502019-data> . This decision was made because this Dataset has data for the past 70 years, that allow the Data scientist to detect the patterns that are specified above. And if such patterns are present for the Data scientist to predict whom the next world champion will be with the 2020 data.

## Data Wrangling

The Data Wrangling aspect of this project was relatively simple. This was possible because of formula 1 being a very data and statistic driven sport. This results in all the races being documented and accounted for. There is more data that we could possibly use.

The Data Wrangling aspect resulted to be choosing the most appropriate datasets that could result in useful Dataframes, that could be used for the goals and purposes of this project.

To create the graphs shown in the main file, there was a need to drop some of the columns and generate new datasets by merging two or more Dataframes. This was done to stablish relationships and to be able to present the graphs that are shown in the presentations.

## Data Mining / Clean Data

Even before we start with the process of Data Minig and cleaning the Data we need to create seven files within the folder of our project.

These files will hold the Datasets that we will use to study the patterns that all the world champions have in common. Five of the seven files will be “.ipynb” we will use this files to clean the data and make our graphs. The remaining files will be a “.py” files these files will be used to import dataset from one file to another to study the correlation between the information.

The “.ipynb” receive the name of the DataFrame they are holding. They also have a tag ranging form Part 0 – Part 3, this is just for organizational purposes and it may not be needed.

Part 0: Race winners.

In this Dataframe the data scientist will only want to keep the Date, the Venue, the Name of the driver and the Team. This Dataframe will be used for graphs and correlations.

There will be another dataset with only the Date and the Name of the driver, this second Dataframe will be used to study correlation.

Part 1: Race results.

In this Dataframe, as in Part 0, the data scientist will only want to keep the Date, the Venue, the Name of the driver and the Team. This Dataframe will be used for graphs and correlations.

Part 2: Fastest Laps

In this Dataframe the relevant information is in the columns named, Year, Venue, and name. This will be then used for correlation and for graphs.

Part 3: Qualy times

For this case we have two Dataframes. The first one ranging from 1950- 1999, and the second ranging from 2000-2020. The need of having two Dataframes is due to changes in regulations in the sport.

The first Dataframe is left untouched. To the second Dataframe we need to drop the columns named: “Q1”, “Q2” and “Laps”, and rename “Lap Time” : “Q3”. After doing such we can merge both Dataframes.

Ones the Dataframes have been merged, we stablish a mask to see who the first qualifiers were. Then we keep the columns of 'Year', "Venue", 'Name', ‘Team’.

This Dataframe will be used for both correlation and plotting purposes.

Part 4: World Champions.

This Dataframe will be divided in several Dataframes.

The first Dataframe will be world\_champions, it would be the first position of every year, and drooping everything except the Name and the number of titles. This will be used for plotting and correlation.

The second Dataframe will be equal to the world\_champions Dataframe and adding the nationality of the world champion. This Dataframe will be used for plotting purposes.

The remaining Dataframes are simply the result of dividing the original Dataframe into the different years from 1950-2020. These DataFrames will be used for correlation.

The “.py” files have both the world champions and the Race winners; these have been created to be able to import these two Dataframes to other files to work with it to stablish correlations and many other functions.

This was done in the notebooks once this process was concluded the relevant code was exported to the main source code and functions were created to alleviate the code in the notebook.

For this project, the Data Scientist must clean the data, for this he or she must get rid of columns that are not as relevant as others.

## Others...

In part 0 and part 4 perform a pivot table to be able to visualize de data, this way you open the door to visualize how many champions there have been in the past, how many races they won, what is the min.

It was not possible to demonstrate the hypothesis, did not see any clear patterns.

I can conclude that with more time and expertise could have gotten some more interesting patterns.

I would study how the champions changed from beginning to end of their career.

I learned a lot about how to get the data to do what I wanted, new ways to present the data, and geoplotting.

The outliers can be seen in the Race wins, Fast laps, and Pole positions. Both in the graphs and in the presentation.

There are repeated elements in all the Dataframes and they are delt with in the scr/main.