CZ4124 Assignment #1 - Technical Report

Name: Andrew Wiraatmaja **Title**: Most Entertaining F1 Circuit

1. Objectives

Formula 1 is the pinnacle of motorsports and has been one of the oldest world championships for motorsport. It already began in 1950 and still exists until now. Currently, it keeps growing and attract a lot of new fans every year. In fact, during the weekend of United States Grand Prix, more than 400,000 people come to circuit to attend the race weekends.

One of the reasons of why people love Formula 1 is because they love to see entertaining races. However, it may not seems be the case everytime there is a race. There are sometimes races that people called boring, which usually come to nature because nature of the circuit.

We know that during those 72 years, Formula 1 already race in a lot of circuits and not all of them are entertaining as stated above. Every circuit is different in terms of layout, length, type of circuit, and on a lot more aspects. Hence, there might be some type of which circuit can provide the most entertaining races, which motivates us to look at the previous race held to determine which circuit is the most entertaining.

This report aims to find out which circuit can provide the most entertaining race based on previous races held on the circuit. Aspects that we are looking is about actions in the race that could make the race interesting, such as Overtakes, Interruption (Safety Car/Virtual Safety Car/Red Flag) and also Retirement.

2. Design Rationale

For the visualization, the most common visualization that is used is bar plot as shown in Figure 1 and Figure 2. This is because we want to know that which circuit have the most numbers of some aspects. Since we are focused on the rank of the circuit, we use horizontal bar plot for most of them as it shows us easily about which circuit comes on top and which circuit are in the bottom. However, for data that do not have much circuit involved such as Interruption, we use vertical bar plot because it's quite easy to differentiate when not much data involved.

Some different colors are used for different circuit because it can help to differentiate and makes the chart more beautiful. Moreover, most color used for the bar plot is based on diverging palette because it can help us to know the rank from top to bottom.

Other visualization that is used is Geo Plot, which is plotting in the exact world map, pie chart, line chart and stacked bar chart.

Figure 3 is used to see how the distribution of the locations of the circuit around the world. Geo Plot is used because the easiest way people see where the circuit are is using the world map. However, the map is quite small, and people might not know how many of them are in certain continent. That's when pie chart, shown in Figure 4, come to play, as we want to know proportion of number of circuits in each continent. Pie chart will help to determine that.

Other visualization technique is stacked bar chart, shown in Figure 5, which shows number of overtakes shown in TV and all overtakes. Stacked bar chart is used because of number of overtakes shown in TV are subset of number of all overtakes. To show how they are related to each other, stacked bar chart would help and show how many of them are shown in TV.

Lastly is the line plot, shown in Figure 5, that is used to track the number of overtakes form the past years. Line plot is useful because you can see it clearly the progress of number of overtakes start from 2004 until now in just 1 graph.

The data contributed much to the progress because there are a lot of circuit that have been in Formula 1, and so it would be hard putting it on 1 graph. Usually, bar chart used in vertical way to show how many of the aspects for one category. However, this doesn't work for data that have a lot of categories. So, most of them bar chart used must be in horizontal way. Also, since there are a lot of csv file used for the raw data, some manipulation must be used in order to provide one graph. However, the data already structured nicely so joining two dataframe is not hard.

Overall strategy used is simplicity, because we do not want to confuse the viewer with complicated visualization. Basic visualization such as bar chart, pie chart and line plot are used. For each graph also there are not unnecessary information. Only information that helps is shown and this is to make viewer more focus on what to get from the visualization.

3. Knowledge Application

Some principles used is mostly Gestalt Principles, which are used in:

- Proximity → some labels are put near to their bar plot to show the value of each bar, and also because we sorted them in descending order, the one at the top are closer because they are "top in something" and the other in middle have same aspect, while the bottom is "bottom of something". Also, this principle used in Geo Plot, because we want to know distribution and plotting there in map will easily determine because if they are in same continent they will be close to each other
- Simplicity → data are ordered so that interpretation is easier for most of the bar plot and the data-to-ink ratio is maximize, and for interruption the color is the same because they belong to same category
- Connectedness → It's on the line plot with all of data that belong to same circuit are connected to each other
- Enclosure → some top 3 are highlighted together and means they are in same group, also in line plot earlier year and recent year are group together so viewer can know easily
- Figure-Ground → some of them are made so they are outstanding from the other

Other principles or law that is used:

- Bertin's Level of Organization are used so we know how to emphasize for certain category, which is like use position for location in map, colour to differentiate
- Weber's Law also play a part by creating a contrast in visualization
- Colour models that used are diverging palette
- Andrew Abela Chart Chooser used to make the bar plot horizontal for multiple data categories
- Eight-Fold Way to connect more with the audience

4. Novelty

Some original contributions are the data collected is not only from one source, but multiple sources. Kaggle data is used for two databases, which is the full Formula 1 World Championship Data and Formula 1 Race Events data. There also external sources, which there are some fans who collected number of overtakes during a Grand Prix and save it to Google Sheets. Also, circuit information is collected from their Wikipedia page, which require scraping. There are several data, and some cleaning must be used so that they can be joined and become one visualizations.

Other original contribution is this is the original idea of how to determine whether a circuit are entertaining or not using data analysis. Mostly people used their judgement to know which circuit are the best, but now in this report, we used facts from previous races to judge whether a circuit is entertains enough or not. We also used some facts that can help entertain the race.

Major reference source is Kaggle and it helps to bring a new idea, because some people already put on their notebook there. It shows some visualizations used and how to process and join that. From several notebook, it has been noticed that not many people are into the circuits analysis and it motivates us to do outside the box which analyzing the circuit.

5. Technical Challenges and Innovation

Visualization tools that mainly used is Matplotlib and Seaborn package in Python. For data wrangling, Pandas are used to manipulate the data. For further improvement, as it is presented in Power Point, some tools are used such as shapes and text box to emphasize more. Also, slides transition and animation were use so the stories can be told in step-by-step matter.

Some technical contributions used are how to modify the data so that it can be visualize in proper manner. Technique used is group by, joining, and sorting. After performing some manipulations, the data can be passed into Matplotlib/Seaborn and use some simple technique there to visualize the data according of what we want.

Technically challenging aspects is how to modify the data so it can be passed to visualization tools. Also, because there are a lot of raw data involved, we must know where to find the information required and join those dataframe. Also, as I have limited knowledge for data visualization, some researching must be done in how to visualize the data. Sometimes you also cannot find what you want from the data and so improvisation must be done so it can still deliver the meaning.

References

[1] https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020References and weblinks (if available)

[2] https://www.kaggle.com/datasets/jtrotman/formula-1-race-events

[3] https://docs.google.com/spreadsheets/d/1XueNI7ZawEX0RLDq5dAGVqsEb1-DBOK2kUWGwM1OMKs/edit#gid=0

[4] CZ4124 Lecture

Appendix – Figures and Tables

Figure 1 – Horizontal bar plot

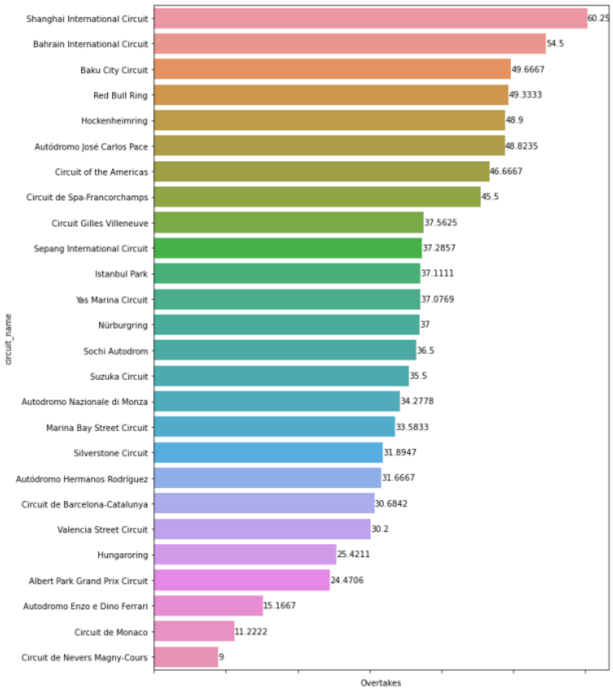


Figure 2 – Vertical bar plot

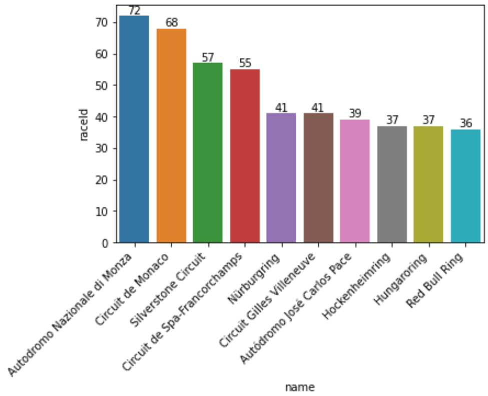


Figure 3 – Geo plot

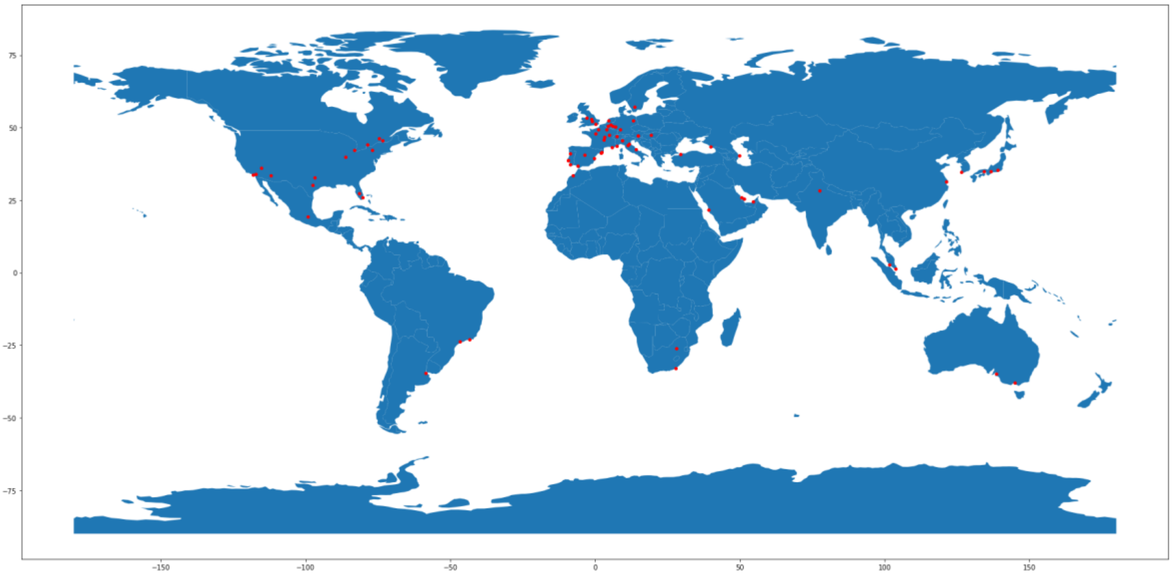


Figure 4 – Pie chart

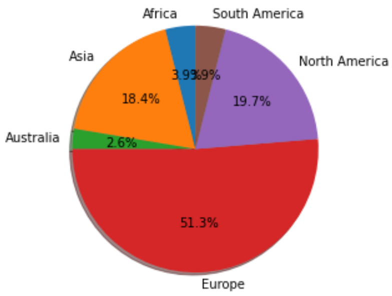


Figure 5 – Stacked bar chart

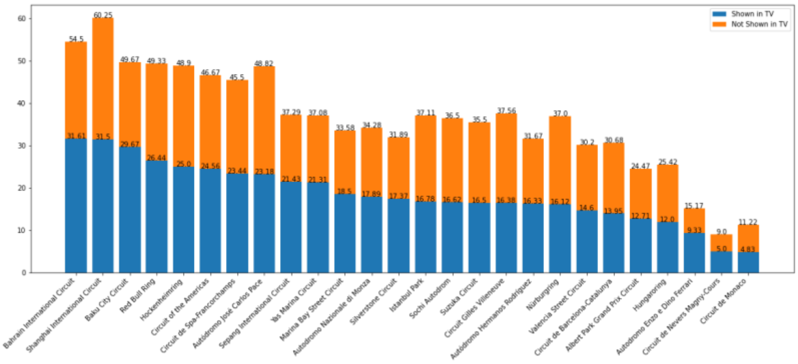
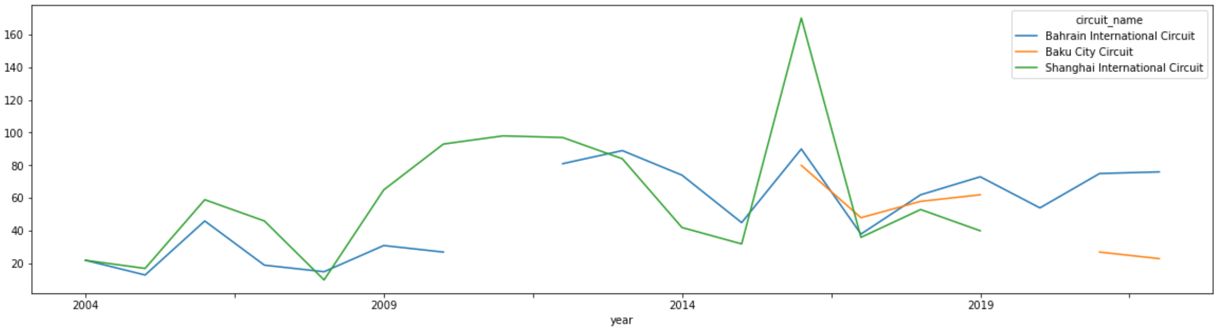


Figure 6 – Line chart


circuit_name
Bahrain International Circuit
Baku City Circuit
Shanghai International Circuit
year