Capstone Two - Project Proposal

What is the problem you want to solve?

I wanted to investigate which team Valentino Rossi was most compatible with during 2000-2020. And which circuit is Valentino Rossi best at racing?

As we all know, in the motorcycle racing industry, the conditions for a rider to win do not depend solely on the rider's skills. Rider skills are the premise, but equally important is the degree of integration between the rider and the team and the degree of integration between the rider and his motorcycle.

I don't think it's a simple "problem". Because if you want to study which circuit Valentino Rossi is racing in is better, then you need to get the results to advance the team's decision-making on the race.

Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis?

The client is definitely Valentino Rossi's team. To be number one in the racing world is the pursuit of every team, and the team where Valentino Rossi belongs to will consider this issue to be very important. Because drivers also have their own weaknesses and advantages, only by making correct decisions through data can the driver's ability be brought out to the limit.

What data are you using? How will you acquire the data?

I'm going to use the data set called"grand-prix-race-winners" from Kaggle.com, and this dataset contains all information on the motogp containing races,drivers' finishing positions,constructors,championships,information of drivers from 1949 until 2022.

I'm going to use the pandas library, and functions like read_csv() to collect the data from the database.

Briefly outline how you'll solve this problem. Your approach may change later, but this is a good first step to get you thinking about a method and solution.

First, I will perform data collection and extract the csv file of the database into my notebook. Next I'll create a folder to store any data visualizations I'll get to later. Then I will go to understand the data characteristics. Use the formulas .columns, .dtypes, .info() to grasp the main characteristics of this data, and find out the two variables of "circuit" and "rider" to satisfy my analysis steps. Finally, data cleaning, check whether there are duplicate rows and make a conclusion.

What are your deliverables?

A code, and there are also data visualization charts in this notebook file.