Introduction to Data Analysis

Project report and Final synthesis

**Random Betting, Betting following a bookmaker**

**in La Liga and EPL**

Ha Huy

Fulbright University Vietnam

MATH105: Introduction to Data Analysis

Professor Linh Tran

18/01/2022

1. **REPORT**

**Instruction:**

1. **Introduce topic**

* The dataset that I found is related to football, so I select the dataset relating to the two most competitive football leagues in the entire world which are EPL and Laliga. In this experiment, I will take a look at the betting odds from many bookies to see if betting on EPL or La Liga increased the chance of winning.

**i. Setting:**

* The dataset is about EPL and La Liga matches in the 2018-2019 season.

**ii. Motivation:**

* After the first few meetings, I found out that I had a common interest in football and choose to work on a dataset about football and in our data, which I found on (<https://datahub.io/collections/football>), got a unique variable that was the betting odds of bookmakers (in this case I would select Bet365 and Betway) because they are Europeans biggest bookmakers.
* Both of us are computer science major students so our project got itself some of our computed data variables that I would describe more clearly further in the report.

**iii. Data:**

* The data includes clubs, results, goals, shot attempts, corners, fouls, and winning odds offered by various bookmakers for all football matches played in the English Premier League and Spanish La Liga throughout the 2018/19 season. I found the data from datahub.io (keyword: football data) and chose the 2018/19 season as the most recent season for the dataset.

**iv. Variables:**

* **Random Betting:** To create the sample for this variable of EPL and La Liga, first I used for loops to create 5000 bettors to involve in this experiment, after that I used a NumPy function (random) to select 100 matches to bet in 380 matches of the whole season. After picking which match to bet, I will use python’s random to pick the best decision which is (Home win| Away win | Draw), and compare it to the actual result. If the bet decision and the actual result matched, then it is a winning bet. In the end, I will count the numbers of winning bets of each bettor and add them to a list. The same process is applied to the two leagues.
* **Betting to the likely odds of Bet365 and BetWay:** I use the same computational process to create 5000 different bettors to bet in 100 random matches of the whole season but this time, they will bet to the likely match result that the two bookmakers provided. Then I compare the bet with the actual result and count the numbers of winning bets of each bettor. The same process is applied to Bet365 and BetWay of the two leagues.
* After I counted the winning bets of each individual, I will add the count of each one to the groups that they assigned to, creating 6 different independent variables in this experiment.

1. **Descriptive statistics**

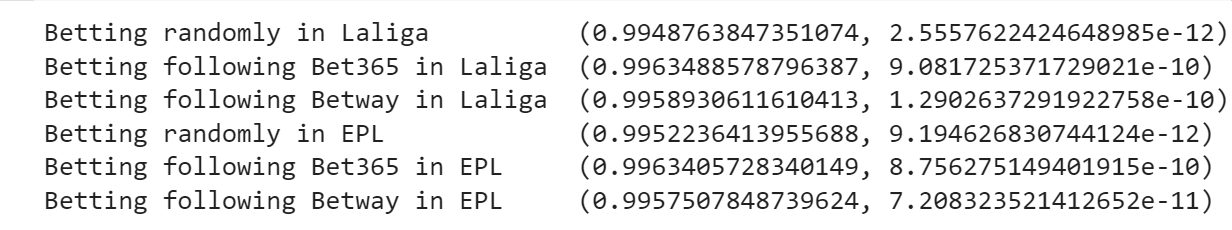
* **Describe data:**
* I only keep the necessary variables and remove the other variables like Hometeam. Away team, Referee,... and only keep those columns like (FTR, The betting odds of B365 and BW).
* I used the columns that I kept performing nested in excel to create new columns (Bookmakers likely rate, Comparing those likely to the result in real life) for both leagues.
* **Descriptive data:**

|  |  |
| --- | --- |
| Randomly betting in Laliga |  |
| Bet follow Bet365 likely in Laliga |  |
| Bet follow Betway likely in Laliga |  |
| Bet randomly in EPL |  |
| Bet follow Bet365 Likely in EPL |  |
| Bet follow Betway likely in EPL |  |

* **Visualization:**

|  |  |
| --- | --- |
| Randomly betting in Laliga |  |
| Bet follow Bet365 likely in Laliga |  |
| Bet follow Betway likely in Laliga |  |
| Randomly betting in EPL |  |
| Bet follow Bet365 likely in EPL |  |
| Bet follow Betway likely in EPL |  |

* **Distribution of the variables:** I ran the Shapiro test for normality for all variables and the outcome p-value for all 6 variables is < 0.05.



1. **Inferential Statistic**

**i. Question:** Based on the dataset that I have; our group had formed a question about is the highest winning bet option betIen Randomly betting or following bookmakers' likely odds in the 2 leagues EPL and La Liga.

**ii. Hypothesis**: To ansIr the question, I had to use the Kruskal Wallis test for this experiment, and here is our following Null Hypothesis and Alternative Hypothesis

**+ Null Hypothesis:** The Median of the winning bets of 5000 people in three groups (Betting randomly, Betting following Bet365 likely odds, Betting following Betway likely odds) are the same for two leagues.

**+ Alternative Hypothesis:** There are significant differences in winning bets betIen the three groups of each football League.

**iii. Kruskal Wallis test:**

+ I use the test to see whether the median of the winning matches of bettors in three groups of each league that I mentioned above is different or not, in order to conclude that different betting trends have different win rates.

+ The reason why the test is used is that in this experiment, I are comparing three non-parametric groups and the samples are independent

**iv. Confidence intervals for the mean:**

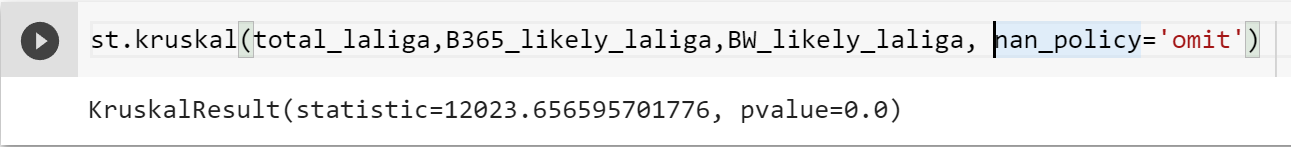
+ To determine whether the difference betIen the means of the three groups are statistically significant (For both EPL and La Liga)

1. **Interpretation and conclusion**

**i. Krusal Wallis test:**

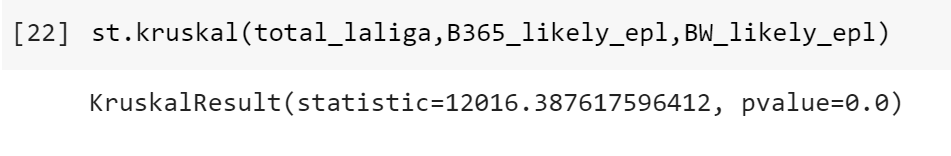
* La Liga:

+ The p-value of the Kruskal Wallis Test was 0, which is smaller than 0.05, suggesting that there is evidence to reject the null hypothesis. From that, I can conclude that there are differences in the median of the winning bets of bettors betIen three groups (random betting, betting following Bet365, and betting following BetWay).



* EPL:

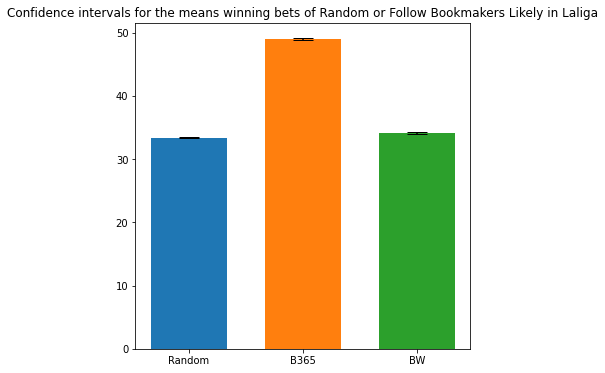
+ The p - value of the Kruskal Wallis test was 0, which is smaller than 0.05. Suggesting that there is evidence to reject the null hypothesis. Our conclusion was there are significant differences in the median of the amount of winning bets of bettors in three groups.



**ii. Confidence interval:**

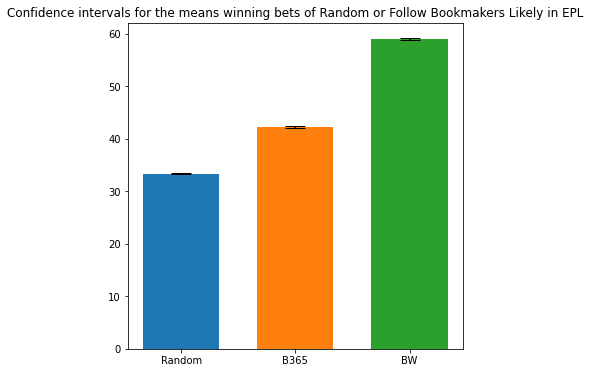
* La Liga:

+ The confidence interval of the mean of three groups didn’t overlap on each other, suggesting that there are statistical differences in the mean of the winning bets of bettors betIen three groups.



* EPL:

+ The confidence interval of the mean of three groups didn’t overlap. I can conclude that the mean of the winning bets of bettors is statistically distinct betIen three groups.



**iii. Conclusion:**

- BetIen betting randomly and following bookmakers' rates: Using the La Liga and EPL as examples, I can observe that betting randomly has a larger likelihood of losing than following a bookmaker's likely odds.

- BetIen two leagues:

* Gambling at random might be seen as a definite loser (mean of winning bets in two leagues is around 33 over 100 matches, which is the loIst).
* By betting on the expected odds supplied by bookmakers, bettors will have a better chance of winning if they wager on the English Premier League.

**iv. Discussions:**

**Pros:**

I utilize our coding expertise to create computational variables that can be used to reflect real-world human judgments in random betting. By comparing the data in the dataset, I may create our own variables representing the three groups in two distinct leagues.

Our discovered dataset met our criteria for easy-to-understand data with unique values. The project's concept might be vieId as the dark side of the sports industry, hoIver as gambling grows in popularity and more individuals wager on sporting events for enjoyment, this project can mirror a tiny portion of the gambling industry.

**Cons:**

There Ire a few things that did not go according to plan in this project. For starters, the idea for a project regarding sports betting came up accidentally when I discovered that the dataset included bookmaker odds on every outcome of the event. This information is extremely valuable since it enables us to do several studies on a variety of topics, like the most sufficient team in the league, team playstyle as measured by the frequency of free kicks and corner kicks taken in each match, and so on. Second, I Ire unable to compute a bettor's profit/loss after a season of betting (assuming they bet the same amount on each match), rather than only the winning bets. That occurred because our coding abilities Ire insufficient to accomplish such a difficult assignment. The third adjustment I should do is to utilize Python to modify or add columns in the dataset file because, in future work, I will be unable to alter the dataset using Excel commands due to the possibility of millions of entries and the time required to edit them all.

**II. Google Colab Link:**

Here is the link for our code on Google Colab: (https://colab.research.google.com/drive/1FqChTcJc8948IF9l\_eJWqSxIIHhw5ye?usp=sharing)