Sports Betting Final Thoughts

One thing that I learned in this project is there may be an issue with different API sources that do not allow you to use them for your statistics. I was able to make the transformations to the API data, but when it came to visualizations, I was not able to use the data. For some reason, the integers were saved as a “nonetype.” I would have needed to manually add the data back in to be able to use it, but I could not find a quick way to do it. Another thing I was able to learn to complete the project was how to use pandas to push an entire data set into SQL. I was initially thinking I would need to create each column name and type before I can loop the data in. However, I was able to find out this can be pushed in using one line of code.

Diverse types of changes were made to clean or manipulate the data. These would include adding time columns, correcting casing, removing columns/rows, removing outliers, and removing duplicates. No legal or regulatory guidelines were located for the data I collected. A risk associated with my transformations is that they could affect the combined 2023-2024 averages. This could lead to false readings for projects completed with the data. For this project, I tried to make conscious decisions to minimize the risks to maintain data integrity.

I did make assumptions when transforming the data. One instance would be in removing outliers in the CSV data. I ran a z-score and compared it to the raw data to see what 1 game played would be equal to. I added a little bit more to this value to grab up to only 2 games played for all of 2023 and removed them. Another assumption was when I was checking for duplicates in the HTML and API data. The data set was large, and I was unable to view each column value to verify it was false. Later I was able to find a better method to check for duplicates and was able to use this to confirm that there were no duplicates.

The CSV data was sourced through Kaggle. Kaggle seems to have a good reputation when it comes to data credibility. However, I have previously found data sets that lose credibility as it does not provide definitions to the shorthand used for the column values. I did compare different 2023 statistics to other sites, and they all seemed to match, which led me to believe that this information was credible. The HTML and API sources were sourced through google searches. These may not be as credible as the official sites, but I was unable to gain access to the official MLB sites. Since I was limited in what I could use, I needed to trust that they were credible.

The CSV Data seems to have been obtained ethically as well as the HTML data. I saw nothing about how the CSV data was obtained, but the HTML data was held on a website that created their own data tables based off MLB.com. The data I was not able to pull was less ethical as it was directly linked to the MLB website. Since they were taking data in this method it would not have been very ethical of me to use it. Considering the API sourced the data from the MLB API, I would not say it was obtained in an ethical way. The MLB site requires you to have an API key and I could not get a student access key for the site. They are sharing the MLB API access for profit. The best way for me to mitigate these two ethical issues would be to go directly to ESPN and MLB for access to their official sites.