**MLB Salaries in R-Studio**

**Introduction:**

In my project I looked at the salaries of all MLB players since 1985. I chose this data set because I enjoy baseball and I think going into a career of inspecting data from baseball players statistics and salaries would be very interesting. My data set had a lot of rows because of how many players have played in the MLB in this time span. The data set was made even bigger because each player’s row was repeated each time they signed with a new team or got a new contract giving them a new salary. When first inspecting my dataset, I was a bit nervous because of how few columns that had data that was useful. However, after data cleaning and manipulation I was able to get rid of unimportant columns and create a couple new and relevant statistics that could be used in my graphs.

**Data Collection:**

When choosing my data set, I knew that I wanted to do something with sports because that was something I knew would have big data sets, and it is interesting to me. It took me a while to find a data set that involved baseball I could work with. I first choose a data set that was far too complicated and had way too many columns of data. I then choose another set that didn’t seem to be compatible with R because of weird symbols in a few of the columns. I finally landed on the MLB Salaries data set, which I found on Kaggle.

**Data Cleaning:**

For my data cleaning step what I did was null the player ID column. This column didn’t do anything for me when looking at the data set since I already had the player’s names. This column just convoluted the data and had no importance in what I was meaning to do with the data.

**Data Manipulation:**

This step of the project was probably the hardest for me because of how few columns of data my data set had. I first made a new column with current data without hardcoding it so that the date would update according to my computer’s system. I then decided on my 2 new columns of data. My first column I created was called stature, which added each player’s height and weight together. This gave me an overall idea of the size of each player. This statistic was relevant to the data set because I wanted to see the salary differences between big and small players. The other category I created was called salary accounting for size differences. To create this category, I divided each player’s overall salary by their stature. This category was made thinking that players who were bigger would in general would have bigger salaries. Looking back this new category wasn’t as effective in giving me the information I wanted as I thought it would have been but later when making my graphs it still had some value. I thought that this category would be relevant because it would show smaller players having higher salaries and bigger players having smaller salaries hopefully leveling them out showing who would have a bigger salary if everyone was about the same size.

**Data Visualizations:**

For my fist visualization I created a dot plot which showed salary vs. season, (or year), which worked as my time component. To create this graph, I used ggplot with a geompoint and simply put in which category I wanted to be my x and y axis. I thought that a dot plot would be most optimal for this graph because it would show where the bulk of the salaries are for each year but also show outliers. As seasons got more and more recent salaries became more spread out and much higher. There was an obvious trend upwards in salary as seasons became more recent. This makes sense because of inflation and an increasing skill gap.

For my next graph I created a team vs. salary accounting for size difference bar graph. To create this graph, I used ggplot and geombar. I used a blue fill to add a color that would not affect people who are colorblind. This graph shows which teams pay out the most money to the smallest players. When looking at this graph it makes sense because the biggest market teams give out the largest salaries, like the Yankees, Dodgers and Red Sox.

I then created my count graph. To do this I made another bar graph using essentially the same code but without a category for the y axis so that it would just default to showing the count for my x-axis. I thought that the best category to put on the x axis would be height just to show about what the average size of an MLB player in this data set is. The graph came out in about a normal distribution with the data centered around 74 inches tall.

For my last graph I created a salary vs stature bar graph. This graph I think is the most interesting because it shows how your size effects how much money you make. When looking at the graph it is obvious that stature does influence your salary in the MLB, but it’s not necessarily just the bigger you are the more money you make. As players get larger, they begin to make more money until their stature gets to about 300 pounds plus inches. After this, players begin to make less and less money. I think this trend is due to people being around 300 pounds plus inches are the people who are the most in shape. Being more then 300 pounds plus inches most likely starts to reduce players speed making them less valuable because they can’t play important positions like shortstop or center field. Most of the time players who play these positions make the most money.

**Conclusion:**

Overall, I think that this project went well considering that my data set was lacking a bit in its categories. Looking back, I should have probably chosen a data set that had much less rows but more useful categories. However, this data set was able to do the job, after being a little creative in my data manipulation step, creating categories that had some use but could have had a little more relevance to them. I think that after cleaning up my data a little I was able to create some nice visuals that are relatively easy to interpret by anyone. My next steps could be to import more data to this set like players statistics. I could then compare different statistics to player’s salaries to see which stats are most important to be proficient in to make the most money.