Project Title: NBA Project

Group Members: Graham, Shakeel

**Final Report**

Extraction:

Our original data sources were a csv file called ‘draft78.csv’ that contained all NBA players’ draft rank and draft year dating back to 1978, an xlsx file called ‘Player - Salaries per Year (1990 - 2017).xlsx’ containing all NBA players’ salaries per year dating back to 1990 and up to 2017, and our final data set was a csv file called ‘player\_data.csv’ which contained basic player information such as height, position, weight etc.  We loaded all three datasets into pandas dataframes.

Transform:

Once loaded into dataframes, we first took took the ‘player\_data’ dataframe and dropped the fields ‘year start’ and ‘year end’ as those fields were already represented in the ‘Salaries per Year’ data frame.  Next we dropped the ‘Yrs’ field from our draft dataframe as this information was also included in the ‘Salaries per Year’ dataframe. Next we merged the ‘Salaries per Year’ dataframe with the ‘draft’ dataframe, then took that new merged dataframe and merged it with the ‘player\_data’ dataframe.  We chose to keep the ‘Salaries per Year’ dataframe as the main one because it had the most unique rows since it had each player’s salary for a given year which we would need for our later analysis. Once we assembled this final dataframe, we dropped the fields ‘Register Value’ as it was not pertinent to any of our analyses and the two fields ‘Team’ and ‘name’ as they were both redundant.

Load:

Once we had our final dataframe created, we renamed all of the fields so that they were lowercase and separated by an underscore rather than a space to prepare to load it into MySQL.  Next we created a database in MySQL called nba\_db and created a table with the corresponding field names. After this, we connected our jupyter notebook file to our MySQL relational database using the pymysql method. We were successfully able to load our data in.

Analysis:

Once we had our core data stored safely in a database, we were able to start manipulating our final dataframe for analysis. We were curious to see what we could learn from the data we gathered.

Our first analysis was to compare a player’s max salary gained to their height.  To do this, we aggregated our data frame so that it was grouped by player name with their max salary for their career.  Next, since their height was stored as a string, we had to split the string into feet and inches, then combine the two values to get total inches in height which we stored in a new dataframe.  Once we had the name of every player and their height in inches, we ran a linear regression with player height along the x-axis and max salary received along the y. The slope of the line of best fit is 105591.20009693105 and the correlation coefficient is 0.06395703097085799

Next we aggregated players’ max salaries by height and took the average of their max salaries for each height group.  We plotted the height values along the x-axis and the average max salaries along the y and ran a linear regression. For this regression the slope of the line of best fit is 87847.06378979368 and the correlation coefficient is 0.29499621506247503 which suggests that there is some positive relationship between height and average salary.

Our next analysis consisted of comparing a player’s draft rank (what place they were drafted) to their average max salary.  For this we aggregated players’ max salaries by draft rank and took the average of their max salaries for each pick. We plotted the draft pick along the x-axis and average max salary along the y and ran a linear regression.  For this regression, we found The slope of the line of best fit is -36946.44053206166 the correlation coefficient is 0.5768357125308243, which suggests draft rank has a stronger relationship to a player salary than height which is expected, however our model suggests that on average a player were increase their average max salary by a greater amount if they gain an inch in height vs. moving up a spot in the draft order.

Lastly, we wanted to compare a player’s position against their respective salary. We approached this two fold. First, we looked the average salary compared to each position. In order to do this, we grouped our data frame by position and took the average of salary. We generated a bar graph to visualize each position and their respective salary. From this bar graph, we observed that the Forward-Guard position had the highest average salary and that the Guard position had the lowest average salary.

While this was informative, we believed we could gather more information by looking at how the average salary for each position changed over time. To examine this trend, we needed to plot a line graph broken by each position. First, we created a series of years from 1991 to 2018. This represented all the “Season End” years from our final dataframe. Next, we needed to gather all the salaries for each respective position. We created another dataframe that grouped each position as well as the season end. Next, we created a series for each position by using the loc function. This returned all the data we needed for each position across all the years. Finally, we did plotted each series as a line graph. From our visualization, we can gather that as of 2018, the Forward-Guard position is in fact the highest average salary; however, our graph shows that this position has increased over the last few years. Overall, all the positions have generally trended up. We can infer that NBA salaries generally increase over time due to a growing market and increase in salary capacity.

The entirety of the ETL project as a whole really taught us the core procedure a Data Scientist follows in their profession. Through this project, we were able to dive into each step that’s part of ETL.