This Capstone project involves analyzing data from NBA players from 1991-2017 in order to determine patterns between various NBA stats and salaries.The original data consisted of three datasets - NBA player salaries, NBA player career spans, and NBA player stats.

The first major step was to combine the salaries and career datasets. However, both sets needed to be cleaned before that could be done. Since the salaries data begins in 1991, the career data would also have to be cropped so that the only players who would be included in the final data set would be those who began their careers in 1991 or after. Unnecessary columns were dropped from both columns and they were both reindexed to the 'Player' column. The career set required further cleaning, however, because there was duplicate player data, partly because of data entry errors and partly because several pairs of players shared the same name. All of those players were dropped because there were only a few instances and the cleanup required would have been time consuming. Their removal would influence the final results only minimally.

In the salary dataset, the 'Salary' column's numbers were string values that needed to be formatted properly to convert into float. There were also several 'Unknown" values in the 'Salary' column. It proved difficult to find all of the missing salary values, especially since they were from the 1998-1999 season, which was shortened due to a strike. Instead, all players that had unknown salary values were dropped from the dataset. As for the remaining numerical values, they had to be stripped of extra spaces on both the left and the right sides.

The salary and career sets were then inner joined by the 'Player' column. The new combined salary/career set was checked for null values and none were found. However, there were some other issues. For a few players, there were duplicate rows when taking into account player names, year, and salary. The second instance of these rows were dropped. Looking for duplicates just based on player name and year indicated that several players who played for multiple teams in a year were paid different salaries by each team they played for that specific year. Those salaries were summed using groupby on player and year columns. The index was then reset so that set was no longer multiindex.

The stats data was then cleaned, first by looking at the player/year duplicate rows. In the set, when players played for multiple teams in a given year, the totals for that year were listed in the first row, while subsequent rows for the same year divided the totals by team. Those subsequent rows were dropped while the yearly totals were kept.

The new salary/career data was then combined with the stats data through inner join on the ‘Player’ and ‘Year’, and columns. A search for null values revealed several in multiple columns. Three players had null values for stats that would not be calculable with the given data, so those players were dropped. A closer look at the remaining null values, and the formulas required to calculate each columns respective stats, indicated that they were due to division by zero. Those null values were replaced by 0.0 floats.An additional column for points per game was added to the set by dividing the number of points scored in a year divided by the number of games played in that same year. It was a surprising omission for the set given that it’s a widely used and cited stat when assessing player value. A final inspection for null values and duplicates found nothing. At that point the final dataset was considered fully cleaned.