

## **CapProj1: Milestone Report**

**Objective:** As the new General Manager for the Seattle Supersonics, I want to be able to evaluate players appropriately, without bias. As the owner of an expansion team in a competitive league with less money to offer potential free agents, our approach to building a competitive team has to be to acquire undervalued players that we predict will have success.

### **Data Wrangling**

- Source of the Data and Method of Acquisition

The data was scraped and collected from Basketball-Reference and ESPN. The data includes statistics about all the NBA players from 2014-2018.

Links: <https://www.basketball-reference.com/>, <http://www.espn.com/nba/statistics/>

The goal of this project is to create an accurate predictor for the metric RPM. Real-Plus-Minus is an unbiased comparative metric to evaluate player value. Specifically, a player's estimated on-court impact on team performance, measured in net point differential per 100 offensive and defensive possessions. The data collected from these two sites, will help achieve the goal of creating this predictor.

- Cleaning Steps

Once the data was collected via python requests. The html format needed to be converted (BeautifulSoup), and then needed to be sorted through such that the contents of the tables were stored into some combination of lists and dictionaries. The data was then arranged into a DataFrame, where it would be easily analyzed. This process was done for both sites.

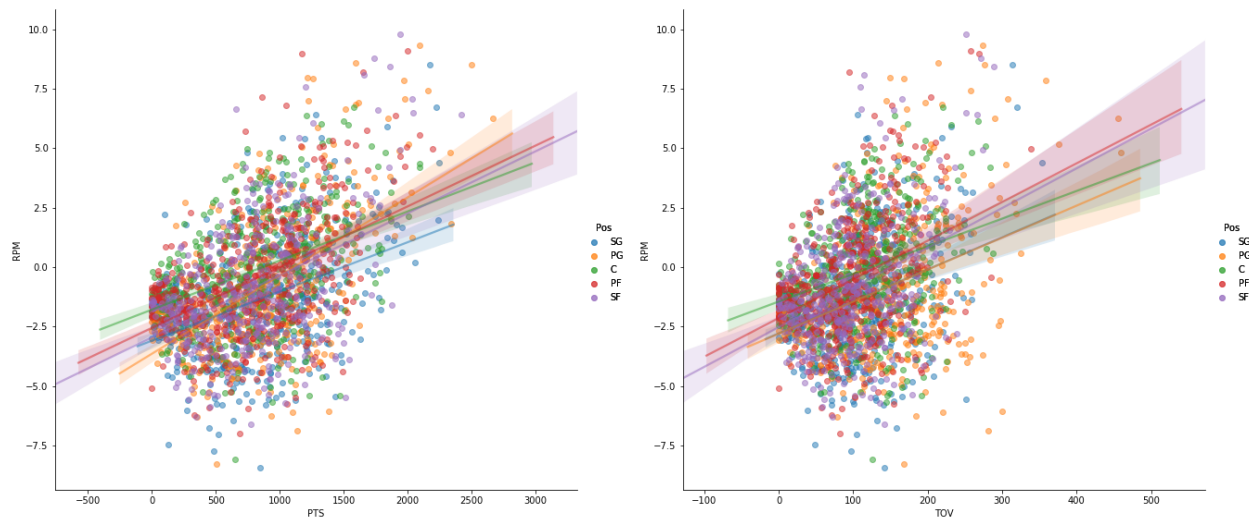
- Merging and Sorting

The two DataFrames would have to be merged together such that no data would be lost or changed. The data also had to be arranged such that all information for a player that year was on a single row, and then the two could be merged into one. Some players were missing from the data because the spelling of their names was different in each DataFrame. Those names had to be changed to match the others. Additionally, some positions were different, so the same had to be done for them. No outliers were removed, due to the high number of variables and proportionally lower number of datapoints. At this point, the data had been cleaned and sorted enough to be saved into a csv for EDA.

## Data Storytelling

From the data collected, containing information of the NBA players from 2014-2018, analysis on variables such as assists, points and rebounds are conducted to find their correlation to RPM (player rating). Here is some idea of what that data looks like now that it has been put together.

- **Embarking on our Data Journey**
  - Finding trends

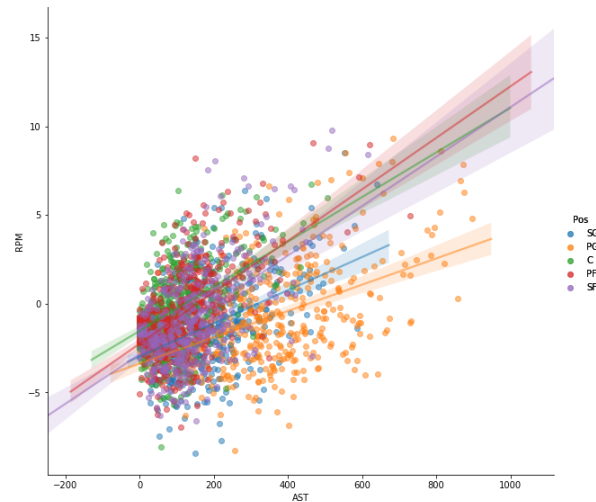


*scatter plots comparing PTS (points) and TOV (turnovers) against RPM*

- **Insights / Further Questions**

If we look at the graph on the left, we see that PTS and RPM have a positive relationship for all positions. This of course is a very intuitive relationship. If you score points, you directly impact your team in a positive way. However, we see in the right graph that the relationship between TOV and RPM is also positive. This means that players who lose the ball more often also have a higher positive impact on their team. This is initially counterintuitive, but once you consider that players who lose the ball are more often in possession of the ball ie. the “ball-carriers” are more often the talented players, it fits the data. This leads us to ask the question of which other variables have trends counter to belief.

Another interesting note, is that when comparing assists (AST) to RPM, we get very different slopes for each position. Above, we can see that the regression lines for each position generally are close to one another and have similar slope. We must consider which variables when evaluated produce better results in a general model with more data, and which ones ought to be compared separately.

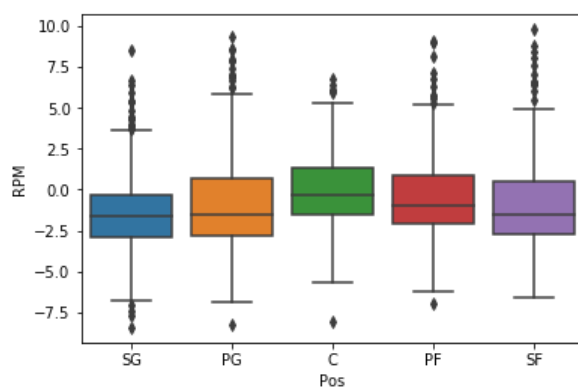


## Exploratory Data Analysis - Inferential Statistics

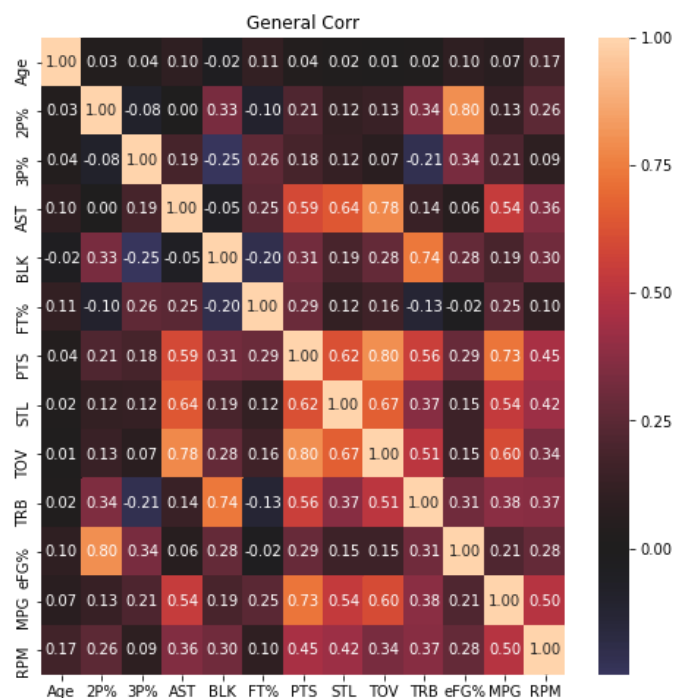
- Initial Analysis:

We see from our analysis of the data that indeed each variable has some value of correlation to the compared variable RPM. Most values that are expected to have a strong correlation to RPM such as points and assists do in fact have a high correlation coefficient. However, some values counterintuitively have high correlation coefficients such as turnovers (which decreases offensive impact and likely increases defensive impact negatively). These trends are most likely the cause of many variables acting simultaneously.

Plot 1 (below-left): Box-Pot of RPM by Position (all years)



Plot 2 (right): Correlation Matrix of all variables against RPM (all years)



- Primary Plot Analysis:

RPM is a good unbiased tool to evaluate players comparatively versus one another, but the process of evaluation occurs first within the team. A player's RPM should not be directly compared to a player from another team until the quality of the team as a whole is taken into consideration first. Additionally, in the first plot, we can see that RPM also differs between positions. To understand this skewness, we must analyze our data by separating players by position and to seeing which variables are more or less impactful than the general model with accounts for all the positions.

- Insights of Primary Plots / Further Analysis / Additional Visuals:

	C	PF	SF	PG	SG	General
Position						
GS	0.2166	0.269	0.285	0.3788	0.2622	0.2769
MP	0.2336	0.2981	0.2597	0.3334	0.2587	0.2571
MPG	0.2379	0.2904	0.2804	0.3379	0.2529	0.2478
FT	0.1868	0.2369	0.212	0.3231	0.1556	0.2227
FTA	0.182	0.1995	0.1973	0.2955	0.1523	0.2141
PTS	0.1592	0.2333	0.1987	0.3161	0.1723	0.2066
FG	0.1326	0.2044	0.1753	0.2638	0.1567	0.1851
STL	0.1804	0.1895	0.2233	0.2681	0.2284	0.1793
DRB	0.175	0.1524	0.1862	0.2259	0.1196	0.1673
2P	0.1128	0.1503	0.1414	0.1722	0.1268	0.1533
TRB	0.1405	0.1147	0.163	0.2212	0.1305	0.1388
FGA	0.1038	0.1717	0.1313	0.2132	0.1184	0.1321
AST	0.218	0.3101	0.2678	0.24	0.2006	0.1284
2PA	0.0846	0.1218	0.106	0.1305	0.1014	0.1191
TOV	0.0879	0.1436	0.1326	0.1298	0.1201	0.1125
BLK	0.0843	0.1203	0.108	0.0815	0.1259	0.093
eFG%	0.063	0.0515	0.0741	0.1149	0.0528	0.0759
FG%	0.0465	0.0319	0.0796	0.078	0.0695	0.0752
2P%	0.0493	0.0338	0.0532	0.0624	0.0385	0.0604
ORB	0.0536	0.0252	0.0437	0.1317	0.0914	0.0603
3P	0.0356	0.0738	0.0825	0.2502	0.0886	0.0432
PF	0	0.0216	0	0.0439	0.0286	0.03
3PA	0.0298	0.0626	0.0664	0.2138	0.063	0.0294
Age	0.0315	0.0638	0	0	0.0487	0.0288
FT%	0	0.0458	0.0203	0.0373	0	0.0143
3P%	0.0402	0.0425	0.0325	0.0622	0.0395	0.0082

The end goal is to create a best-fit model that accurately predicts RPM. To do that and be able to see clear connections, we must strip away unrelated information and adjust the related information in a way that can be understood. From our analysis above of the primary plots, we know that we must first strip away the unrelated variables. By sorting these variables by their r-squared values (the variable's correlation to RPM) and using their p-value (in this case whether the relationship is statistically significant), we can determine which variables we should use. In the chart to the left, the variables are sorted by their r-squared value for all positions and given value zero (0) if insignificant (p-value greater than 0.01).

From this chart we now know which variables have relevance. To create a good model, we must use the right combination of variables to produce the most accurate prediction of what RPM is based on those variables. For most of the variables, the general model has an r-squared value

averaging somewhere close to and between the values for each position. A general

model would in those cases be sufficient to describe them. But if we look at assists (AST), we see that all the r-squared values are higher for each position than the general model. If we look back at the scatter plot of 'AST' and 'RPM', we can see the data is much different for each position. Therefore, we need to keep in mind, during model testing, which variables must be evaluated separately and which ones can follow a general analysis. During the model testing process, we see that the more variables we have in our model, the higher the correlation is to the RPM data. But that is merely because the model is from one dataset, and having more variables just means the model can describe itself better. To be able to predict future RPM accurately, we must determine which variables have a direct impact.

To determine which variables to use, we initially include all the variables in our model. The model will initially contain variables that are redundant (Multicollinearity). To get the right combination of variables for our model, we have to determine which of these variables are redundant of one another. Then one group of redundant variables at a time, only use the variable with highest correlation. Using the Variance Inflation Factor (VIF), we can determine which variables are likely correlated to one another. Variables with higher VIFs are likely related in this way. When you systematically remove one variable, every other variable's VIF value decreases, significantly more so for the variable related to the removed variable. For example, the variables "PTS" (points), "FG" (field-goals), "2P" (two-pointers) and "3P" (three-pointers) all describe scoring two/three points for a made basket for that individual. These are essentially combinations of the same thing. We then leave only one variable in the model, that which has the highest correlation, in this case "PTS". After repeating this process, we get a model without insignificant or redundant independent variables (highest r-squared first).

$$\text{RPM} \sim \text{GS} + \text{MPG} + \text{FT} + \text{PTS} + \text{STL} + \text{DRB} + \text{AST} + \text{TOV} + \text{BLK} + \text{eFG\%}$$