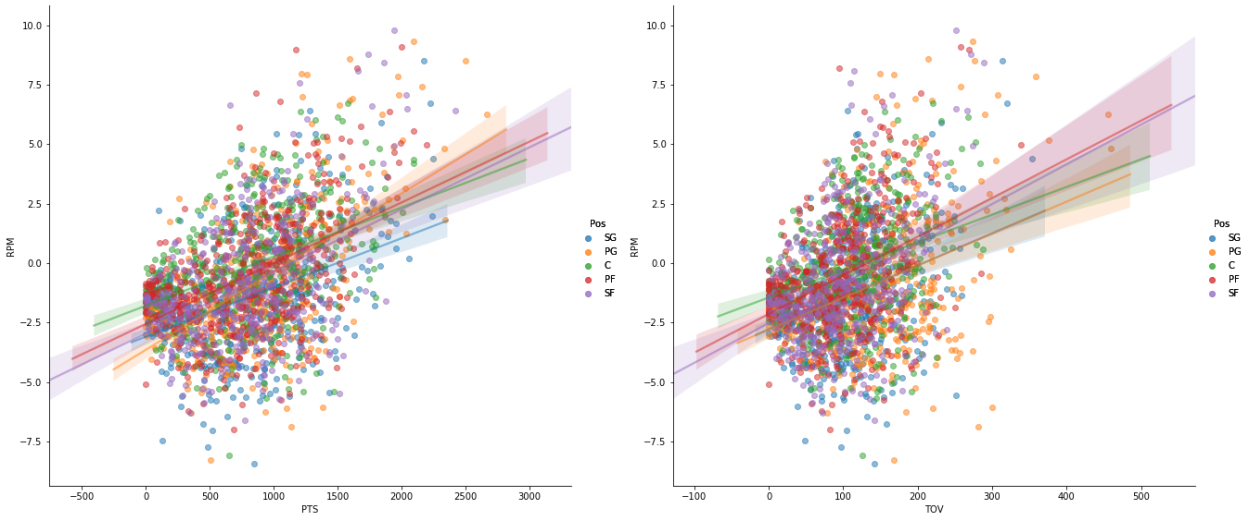
CapProj1 Mini-Proj: Data Storytelling

Revisiting where we left off…

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). By doing so, a best-fit model can be constructed to determine and therefore predict RPM.

- **Crossroads of our Data Journey**
    - Finding trends

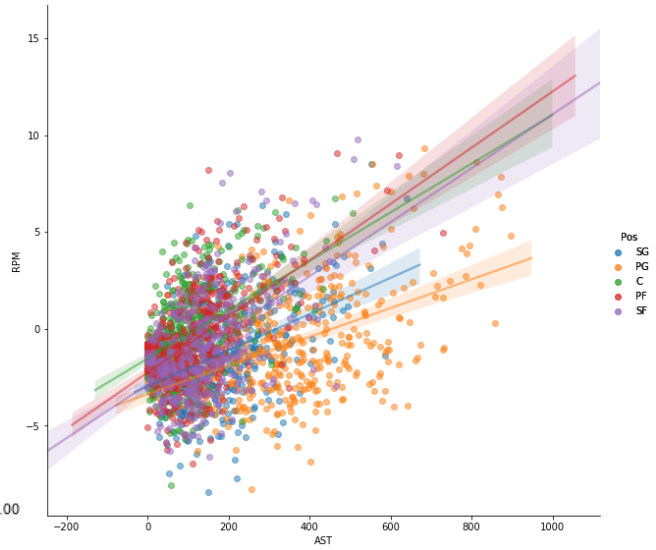


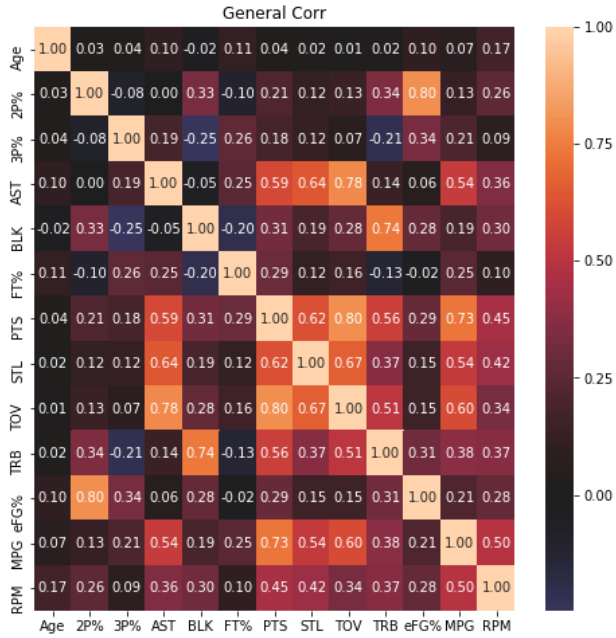
scatter plots comparing PTS (points) and TOV (turnovers) against RPM

- Looking at the plots, what are some insights you can make? Do you see any correlations? Is there a hypothesis you'd like to investigate further? What other questions do the insights lead you to ask?

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.





Our next obstacle is multicollinearity, which refers to the correlation that independent variables have with each other. We want strong correlations between these variables and the dependent variable RPM that we are trying to model. But when we see high correlation between the independent ones, this complicates the model and makes it less clear which variable is responsible for the changes in RPM. We can see examples of these strong relationships in the graph to the left. In order to create a model that encompasses this issue and cuts through to show the clear direct relationship between the independent variables and RPM, careful analysis for each individual variable must be done.

- Mission

It is important to take steps, such as the ones above, to cut and polish important information. By the time we come up with a finished model, we do not want a predictor that merely replicates the image we see with our own eyes. But instead we want a true predictor that accurately predicts the real life future outcomes of RPM that we have been analyzing. Continuing this process will lead to the best result.