Mike Haldeman

# Multiple Regression with dummy variable and Cluster Analysis

This project will be analyzing National Hockey League statistics. In doing so, it is hoped to help predict the Salary for a player based on the various variables that are attributed to their play. In order to help predict the salary of a player, the model will be conducted using Regression. This data analysis method will optimize the variables within the data set to allow for an accurate prediction model for salary. Using the prediction model, the analyses of the five underpaid and overpaid players will be shown. Also within this project, a cluster analysis will be performed to cluster players together with players that are similar to each other. This will be performed using the data set and with the combination of regression it will help accurately cluster players into categories.

After analyzing the NHL data set in JMP, it was determined that the following variables were statistically significant to our regression model to help predict NHL salaries. These variables are: the position of the player, the points of the player, the penalty minutes of the player, the total on ice per game played, and the total on ice percentage. We can see this supported in the JMP output below(Figure 1).

An alternative for this model would be to include the games played variable. But for a more accurate model, the inclusion of this variable would allow for the prediction of salary to be weighted wrong. As if some players were injured or were not able to play within that season and had that be comparable within the season. This would cause the prediction to be off. For example, if one player plays 6 games while another player plays the full 82. If the player who plays 6 games is being paid more but has a season ending injury and can't play the rest of the games will not allow for comparables.

Moving on from the predictions, the next step is to take the predictions into ranking the five most underpaid and overpaid players. This can be found by taking the actual salary data for the player and the predicted value given from the model of statistically significant variables. (Figure 2) These variables are compared 0.05 > ”the variable”, with each variable reaching statistical significance. The only variable of question could be penalty minutes (PIM) as it is greater than 0.05. I felt the need to use this variable as it is very close to the comparable number of 0.05 along with it being a very important distinguisher between some players with the National Hockey League. From this data the model can be created. The variances can be shown for the five overpaid and underpaid players below and within (Figure 2).

Top 5 Underpaid: These are the players that have outperformed the estimated salary for that player.

1. Leon Draisaitl has outperformed his expectations by $4,308,087.27
2. Jack Eichel has outperformed his expectations by $3,905,912.68
3. Viktor Arvidsson has outperformed his expectations by $3,792,112.38
4. Alex Wennberg has outperformed his expectations by $3,668,965.04
5. Bo Horvat has outperformed his expectations by $3,622,951.94

Top 5 Overpaid: These are the players that have underperformed the estimated salary for that player.

1. Jonathan Toews has underperformed his expectations by $8,667,506.46
2. Patrick Kane has underperformed his expectations by $7,585,411.94
3. Shea Weber has underperformed his expectations by $6,795,094.09
4. P.K Subban has underperformed his expectations by $6,070,824.71
5. Alex Ovechkin has underperformed his expectations by $5,165,516.92

Analyzing the shown data above (Figure 4), it is evident that there are players on the underpaid list that have a very low actual salary considering all the salaries with the data set. This leads to the assumption that these players are rookies within the league. As these players have a rookie level contract salary. As shown in (Figure 4) these players have overperformed their salaries immensely. While on the other hand it can be seen that the players listed for being overpaid are those with large salaries. Given the large actual salary of the players it can be assumed that these players are veteran players in the NHL.

With the details known from the regression model, the data collected can be used to analyze a cluster analysis from this dataset. From the cluster analysis, there are 6 groups that can be clustered together for a formation on the dataset. (Figure 3)

1. “The Rookies
2. “The Enforcers”
3. “The Average Offensemen”
4. “The Average Defensemen”
5. “The All-Star Defensemen”
6. “The All-Star Offensemen”

With each number cluster name and number coinciding with the clusters shown in(Figure 3). Each cluster has their own distinguishing variables. For example, “The Rookies” differ from “The All-Star Forwards” in all categories, as “The All-Star Forwards” had much higher averages than all of “The Rookies” variables. Looking at “The Enforcers” variables compared to “The Rookies”, we can see that most variables are very similar in all variables except for penalty minutes (PIM). As “The Enforcers” penalty minutes per game were very high when compared to “The Rookies”, or really any other cluster. Cluster 5 and 6 differ only at position and the amount of points averaged. This is understandable as defensive players are less likely to gain points when compared to offensive players. Along with this clusters 3 and 4 were similar in this sense as all variables were the middle ground compared to the other clusters. The distinguishing factor between 3 and 4 were position, as 3 was the offensive players and 4 was the defensive players.

Now that there are names to each cluster with distinguishing factors, we can use this to analyze the Regression model with finding where the five underpaid and overpaid players fall within the clusters. Seeing if the regression model holds truth with the predictions.

Analyzing the cluster data with the regression model for the 5 underpaid players. It can be seen within (Figure 4)(Figure 3) that the colors next to each players name match up with the cluster that each player is belonging to. From this we can analyze further. It is present that each player within the underpaid category within the regression model falls in cluster 6 or “The All-Star Offensemen”. The cluster is made up of players that have the highest average in all categories, also aligning with being an offensive player rather than a defensive player. Showing that this further proves the truth from the regression model as each of these players are outperforming their contract as to be clustered with the league leaders in each variable. When it comes to the overpaid players (Figure 5)(Figure 3) we see that each of these players are clustered respectively with the All-Star categories aligning with their positions. As these five players make up parts of clusters 5 and 6. The clusters also show that each of these players are all very good players within the NHL and would be considered All-Stars within the league. As we see from (Figure 2) that each of these players are paid very high actual salaries emphasizing that these players must be somewhat good to be paid that high of a salary. What can be gathered from this is that the regression model still predicts for each player to have a high salary, but the actual is much higher than the predicted. This leads to the question why these players are being paid so much more than the underpaid players. It could be leadership, being a veteran in the league, or the influence the player has within the team and city they play for. All of these could be answers as to the much higher salary than all others within the dataset. But as shown in (Figure 5) (Figure 3) all five players fall within the All-Star clusters showing that they are of high skill level and earned the salaries being made. Along with the five underpaid players, have made names for themselves and given reasons as to be earning the salaries similar to highest within the National Hockey League. With each player belonging to cluster 6, being amongst the greatest players within the league.

Figure 1:

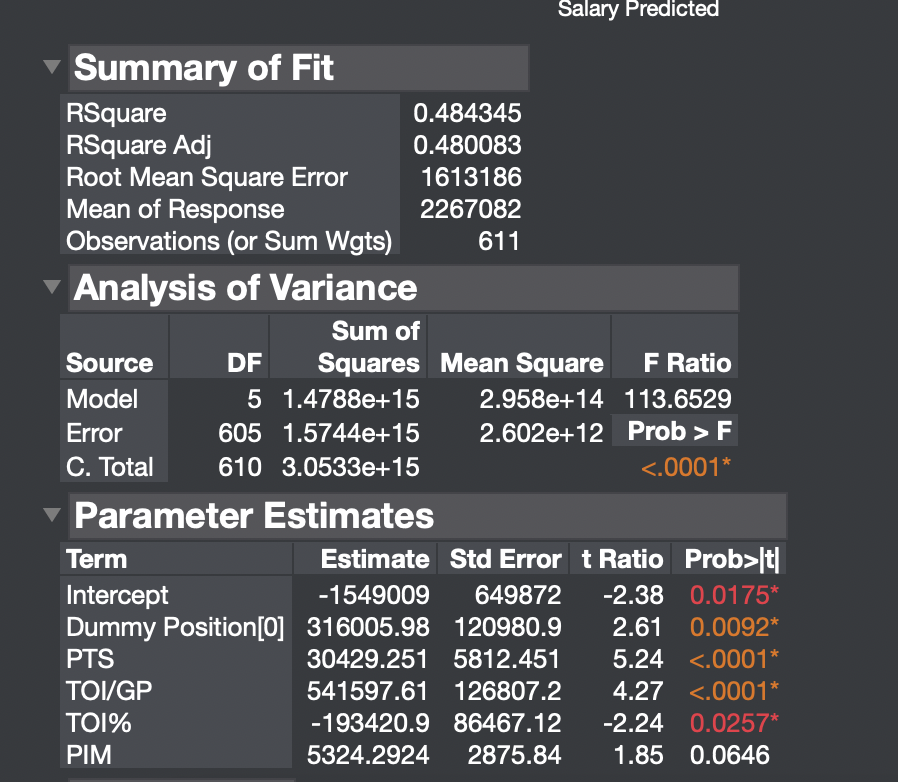


Figure 2:

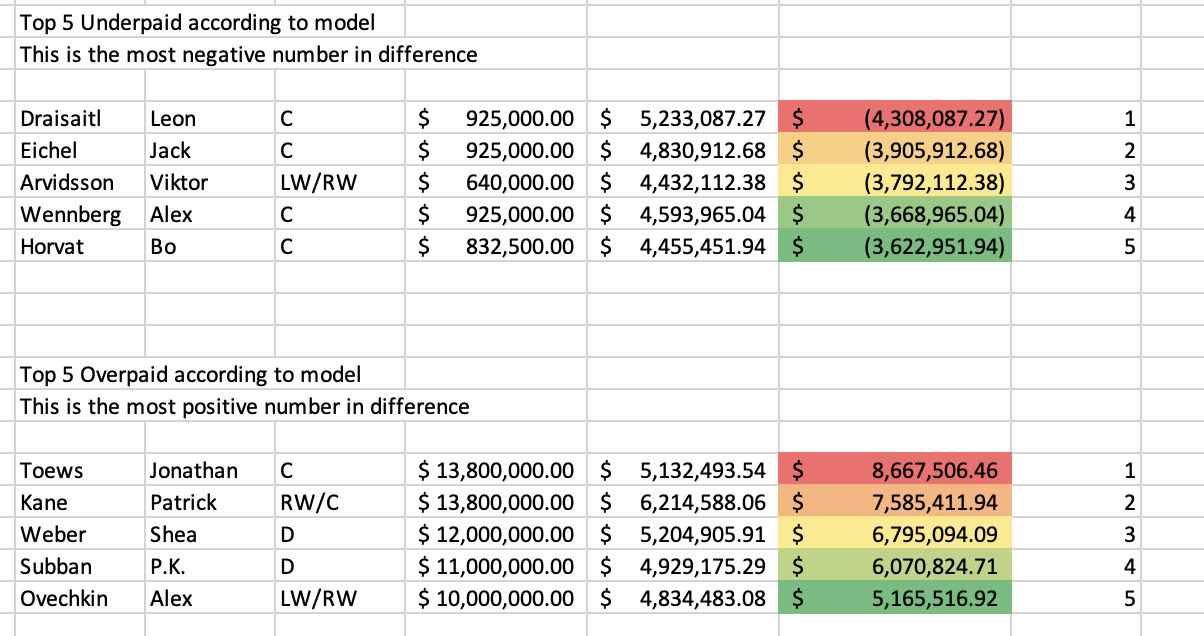


Figure 3:

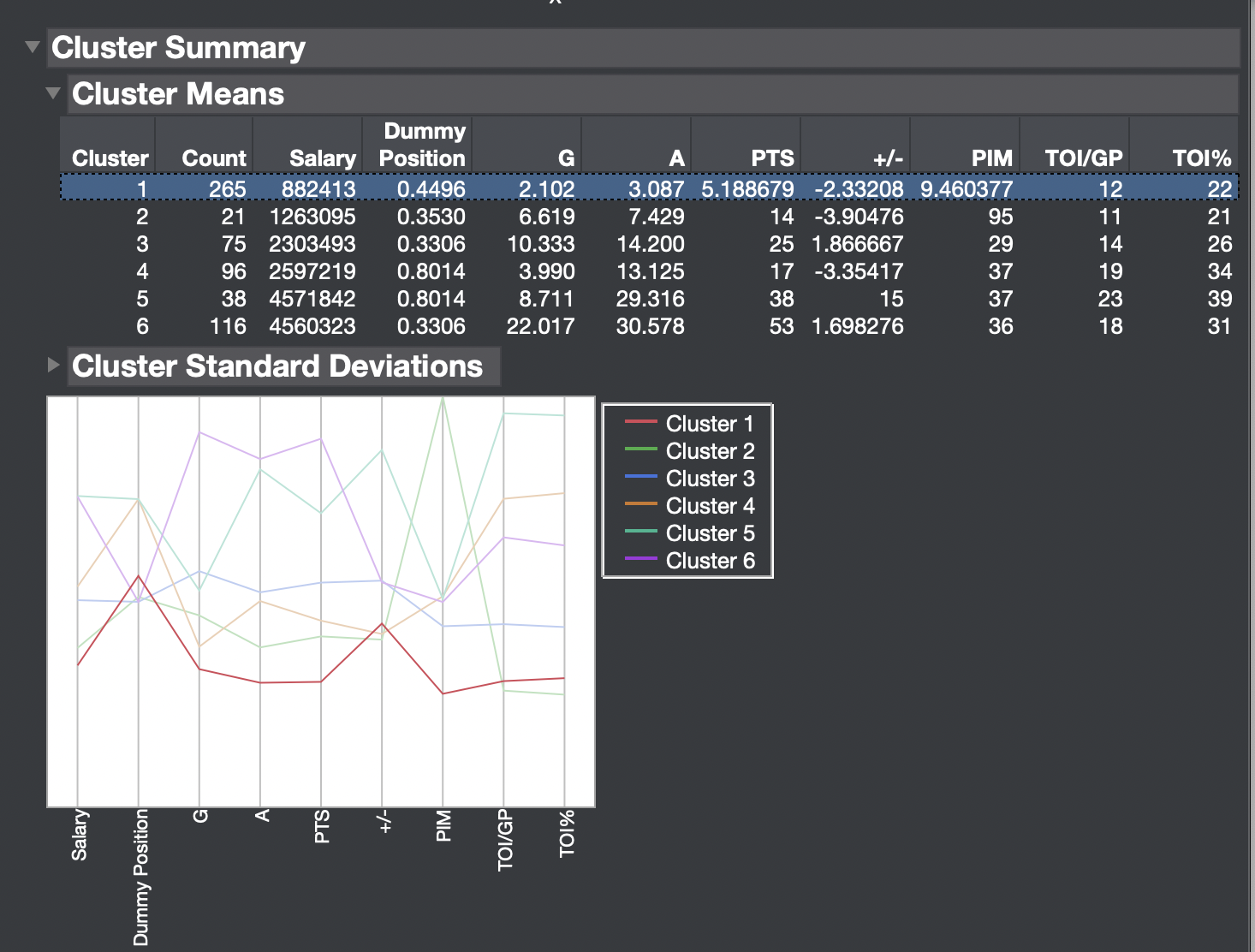


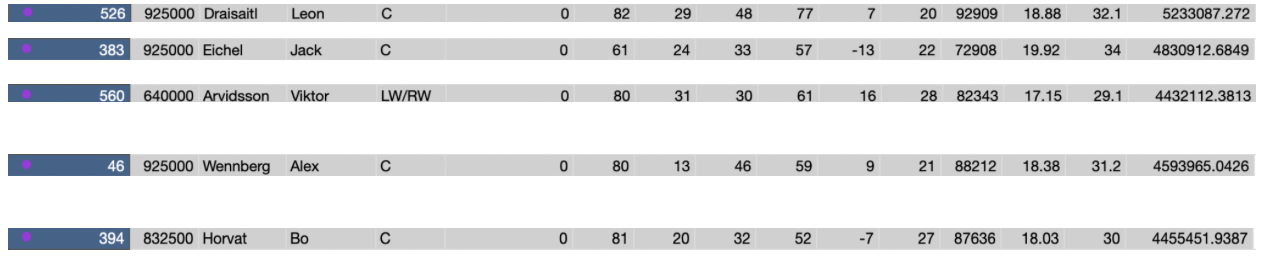
Figure 4: 

Figure 5:

