**Project#3 Final Report**

**Executive Summary**

The Minnesota Wild have experienced multiple years of poor performance in the regular and postseason. One thing that has stayed true is multiple game losses landing directly in the middle of the season. Sports analysis and laymen alike have described this occurrence as a mid season slump. While there is a plethora of advanced statistics and reports to dig through, the organization and casual analytical gurus are left clueless as to why this happens. As we tried to create a model that would give us some insights as to how to specifically stop a midseason slump, we ended up stumped.

Deciding to take a different approach to stop the midseason slump looked a lot like just getting a better team. Getting a large dataset allows us to test multiple metrics against against player performance. Trying to get the best valued players was the key to improving our team given the salary cap set upon NHL organizations. Creating the metrics to measure player performance compared to salary created a value that we could set to individual players and teams as a whole. This metric is the cornerstone to our analysis but could be modified to fit other needs or leverage other performance metrics other than just goals.

Going forward, we can see the use of this model to help individual teams decide what players are worth their current salary. It may also provide insights to recruiters or the front office as to which players would provide the organization with the best value, ultimately eliminating the midseason slump altogether and create a more value packed roster. Although our current analysis only measures offensive players such as forwards, it can be modified to fit each organization's needs. All in all the insights provided by our analytical model will prove to pave the way for hockey analytics and bring to the sport the data-science it currently lacks.

**Table of contents**

**Executive Summary……………………………………………………………….p.1-2**

**Data Understanding……………………………………………………………….p. 3**

**Data Preparation…………………………………………………………………...p.5**

**Descriptive Analytics……………………………………………………………...p.6-9**

**Modeling and Predictive Analytics……………………………………………...p. 9-11**

**Comparison with Documented Results………………………………………..p. 11-13**

**Data Understanding**

Our dataset consisted of 120 columns, each representing a different variable or metric. Some of the metrics were simple descriptors such as name, number, team but many of the metrics were advanced in nature such as Goals vs Salary(GVS) which was an advanced calculation of a player's goals to their current salary. This was all laid out in one spreadsheet which congregated different metrics from different sources. This proved helpful to have every player in the league in one dataset with almost every metrics currently available.

Breaking our dataset down and understanding it proved to be more challenging than just reading the legend that was provided. The data was vast and the metrics were advanced. Most of the time we would have to research a statistic to figure out how exactly it was being calculated and in which ways we would be able to apply it to our end goal. The metrics usually included one or more other advanced metrics to calculate the final statistic. This required us to fully grasp what each variable was and how it was calculated and applied in the field.

The statistics provided were almost always in the form of numeric data. Aside from the team name, player name and number information, the statistics were calculated by either direct statistics taken during the game and sometime included normalizing the metrics based on average ability or pay. This allows us to harness not only the simple data that is collected such as goals and time on ice but also the more advanced statistics that take into account the player's replacement level and how they should be performing compared to the rest of the league.

**Data Preparation**

The quality of the data provided was surprisingly good. Taking into account the breadth of different statisticians and websites that this data was collected from, we were expecting to have to do a lot more cleaning than we did. There were some minor issues that needed to be fixed but overall the quality was above average. The real challenge came when we were attempting to figure out which of the 120 metrics were going to be the most useful and informative.

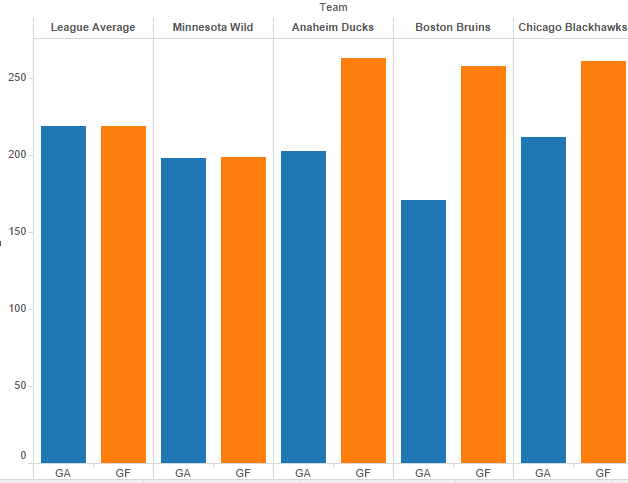
Our data did require us to remove characters that were causing errors in our programs such as Weka but that did not require much work. The dataset did contain some missing values and such but was not problem as either a missing data point represented a zero value or it was insignificant altogether. In cases where we needed to fill in the values, we pulled them from the NHL.com stats page or provided the average number depending on what was approriate. Narrowing down our metrics was the only task that was significantly difficult but did prove to give us a better insight as to how these metrics will be used to predict and prescribe different actions and detect player performance.

Coming up with a threshold to measure players was imperative and needed to be very accurate as it is the basis for making our decisions on players. The first step was to break the data down into different binary bins and runs analysis on which metrics were associated with positive outcomes such as goals, assists or shots. This proved to be difficult as most of the metrics would only be associated with negative outcomes such as low shots for is associated with low goals for a player. Trying to find positive associations was something that did not work well for us but we were able to find out which metrics would be least likely to lead to low numbers of what would be positive metrics.

**Descriptive Analytics**

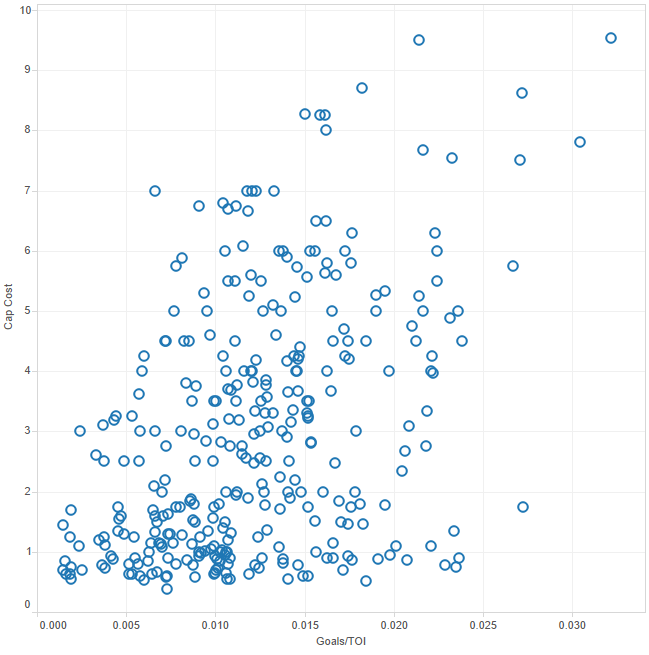
Our use of descriptive analytics included presenting charts showing the Goals Against(GA) and Goal For(GF) for 3 different teams including the Wild. We also included the league average of both of the statistics. GA for the wild is consistent or better than the league average and other teams signifying that our defending players are up to par and don’t require the analysis we are providing. On top of this performing analysis on defensive players is considerably more difficult and would require us to completely change our model.This showed that the wild significantly lacked the goals for compared to some of our closest rivals and the league average. This presented us with the ability to focus on strengthening the value of the Wilds offensive players.

**Bar graph showing Goals For and Goals Against for the entire NHL, MN Wild, and other teams:**



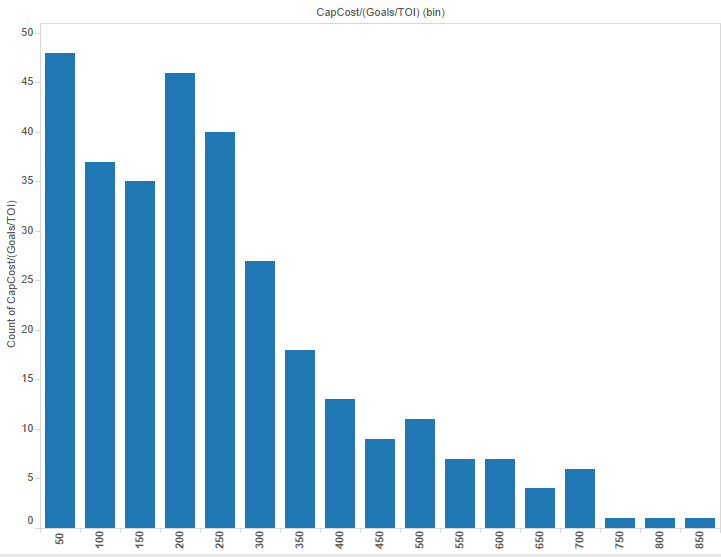
Another model that we created plotted all of the players in the NHL illustrating their goals over time on ice(G/TOI) against their cap cost otherwise known as salary. This laid out the groundwork for our model and gave us an understanding of players value as we were comparing each player's performance to their time on ice. Because this statistic took into account a player's goal divided by their time on ice, the scatter plot shows how effective a player is compared to how much they are played. We were able to locate the spot where we would look at players and consider trading for them.

**Scatter plot showing Cap Cost versus Goals/TOI:**



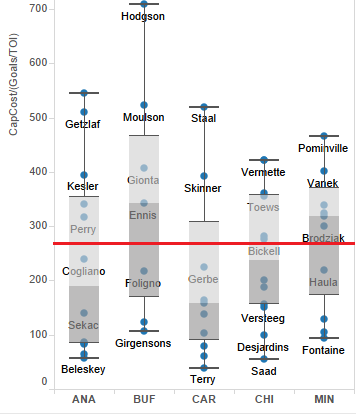
One more illustration we created shows a complete statistic we made, cap cost over goals over time on ice(cap cost/(G/TOI). This histogram shows the distribution of players ratings of our statistic. The horizontal axis represent the value of our statistic associated with each player and is distributed into bins of 50. The vertical axis represents the number of player in each respective bin. Counterintuitively, the players with a higher statistical value are actually being paid more than they are producing. As you may notice there are a lot more players with a lower value, this is where a scatterplot and regression line will give a better indication of who exactly to pick but nonetheless there are considerably more players to potentially acquire and little players to potentially let go or trade.

**Histogram Showing Frequencies of Players with Different Cap Cost/(Goals/TOI):**



Finally, we have created a team based box-and-whisker plot to visualize which teams are getting the most value out of their offensive players. The red line represents the league average for our statistic. Our Minnesota Wild fall in line with most of the NHL but the upper whisker signifies that we may be spending too much money on certain players, namely Pominville and Vanek. This illustration does a great job of aligning our team with others. If you look at the Chicago Blackhawks, they have a much more efficient team that gets the most value out of their players. The Blackhawks have recently started using analytics to base their decisions and it seems to be working very well for them.

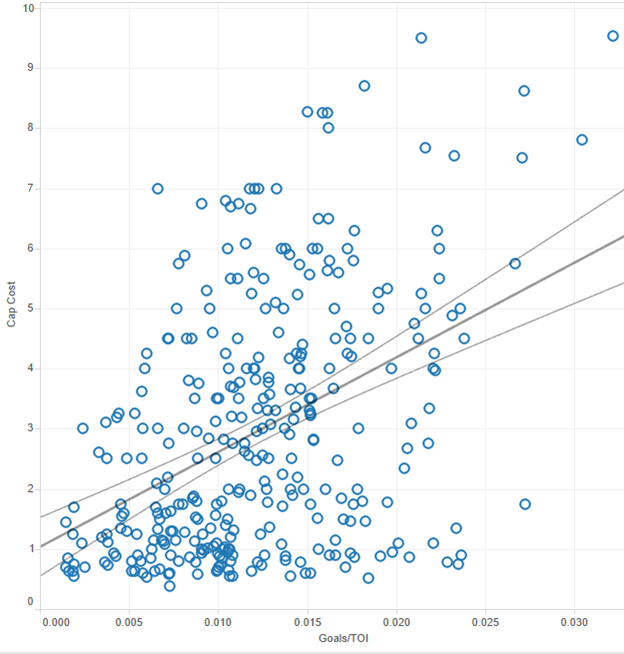
**Box and Whisker Plot comparing MN Wild to Other Teams’ Relative Effectiveness of Dollars Spent on Offense (Red Line Denotes League Average):**



**Modeling and Predictive Analytics**

We have used a combination of models to achieve different aspects of our overall goal during the predictive analytics portion of this project. First off, we used a simple linear regression of Cap Cost versus Goals/Time on Ice to look at every offensive player in the NHL. This model gives us an idea of a player’s offensive performance relative to their salary. This model is very useful because it allows us to find the best “value players”, or players with a relatively good performance per dollar spent on their salary.

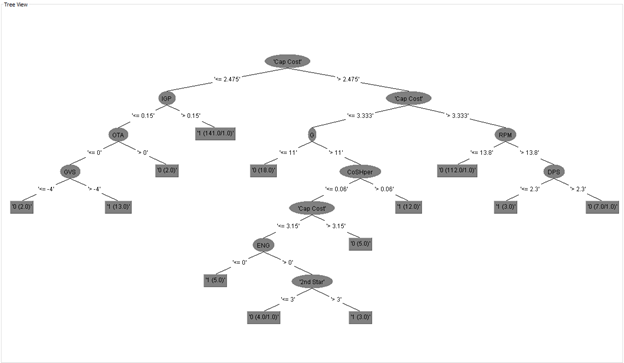
**Linear regression showing Cap Cost vs. Goals/TOI:**



Next, we used the regression line from our simple linear regression to create a binary class variable of whether we should be interested or not be interested in any given player. If the player is on the right side of the regression line we assign them a “1” and any player above the regression line gets assigned a “0”. We then used a J48 decision tree to find variables that most strongly predict a player's that we want to look at or want to avoid.

The first split on our tree is Cap Cost. There are different characteristics that are more or less predictive when looking at different ranges of Cap Cost. For instance, When looking at players that makes less than 2.465 million dollars per year, the vast majority of players are going to have an IGP of above .15, which means they will be scoring at least 15% of the total goals scored on their respective teams. On the split above 2.45 million dollars Cap Cost, there are a couple other variables that are deemed more important. Having over 11 goals is the biggest split for the players making between 2.45 and 3.3 million.

**J48 Decision Tree:**



There were a lot of alternatives for us to choose from and we tried many different models while working on this project. The main models we attempted to use and later figured out they were not the right models for the job. We tried to use the Naive Bayes model to use the Minnesota Wild test set against the NHL dataset. We ran into a few complications and figured that this model is mostly used for text mining and our data is almost entirely numeric. One model that we thought could potentially improve analysis would be multiple linear regression. Although we didn’t use this model with any amount of success, we think it could be applied to our dataset for a more in-depth analysis.

The model we chose is very well suited to solve the MN Wild’s problem of having a relatively weak offense largely due to a misallocation of funds on offensive players. Our main use for this model is to find players from the NHL to replace players on our team who are do not fit the criteria set by our model. This model will be expanded upon in our prescriptive section where we will actually make suggestions for players to drop and acquire based on this model.

**Evaluation**

Our original model that relied on GVS and GVT encountered overfitting quite often. This led the model to predict obvious results such as a player with higher GVS would have a higher goal record. In order to counteract overfitting we needed to create our own variable that would seek to avoid these problems. Instead of measuring a player solely by his goals vs salary, we were able to incorporate their time on ice to give the model a more realistic perception of what players do only while they are on the ice.

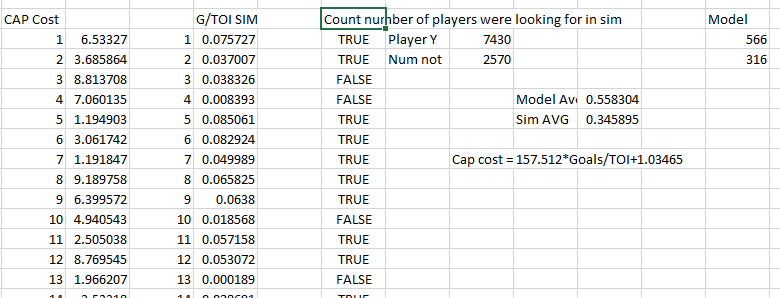
With the resources of a NHL hockey organization, we will have access to many more proprietary statistical databases. The increased access and financial backing would allow us to be able to run better models and identify players more accurately. Additionally, there is an increasing number of visual analysis that is being applied to hockey. This is where access to stadium camera systems give analysts the ability to track exactly where statistical event occur on the ice.

Models that can accurately value players by their salary is greatly needed in the world of hockey. All to often players are overpaid and management signs players on for contracts that will end up overpaying for them near the end of their contract. In order to create a more efficient team that is able to deliver results, analytics must be adopted at a faster rate in the NHL. This would provide the fans with a more satisfying game to watch and benefit everyone from fans to management and even the players as they are able to reach greater heights and possibly even win the coveted Stanley Cup.

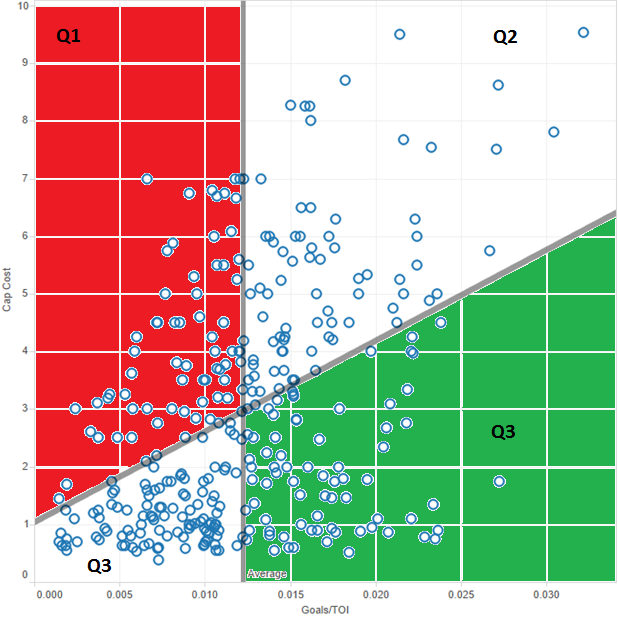
The ROI is not easily calculated when looking at letting players go, buying out players, and signing new players. There are just too many unpredictable factors in a real world scenario. In our predictive analytics section, the increased performance at roughly the same market rate for players is examined, and this is a more accurate representation of the ROI. In other words, for the same price of a few alternative players, we will show the increased performance, showing options to achieve a greater ROI in the same salary range for our underperforming players.

**Prescriptive analytics**

Simulating our model proved to be difficult. We used 10,000 trials to simulate our model. We were comparing whether the number of players that would fall under our regression curve would work in 10,000 random simulation. We chose 10,000 simulations because our whole dataset consisted of 880 instances. This gave us 10 times the amount of our dataset and should provide a meaningful test. Although the simulation predicted a smaller number of instances under the curve, we believe our model still hold up because of the complex nature of such a large dataset and the game in general. The salary and goal/TOI was not correlated to each other so the simulation was not 100% the same. Below you will find the results with model and sim averages present.

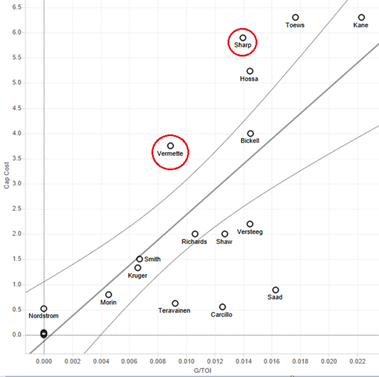


**Linear Regression Quadrant Analysis:**



The above visual shows our regression line broken down into quadrants, Q1, Q2, Q3, and Q4. As you can see, the Quadrant Q1 is marked as red, for high cost-low performance players. These are the players that should be avoided at all costs. Quadrant Q3 is marked with green to distinguish low cost, high performance players. These are typically the players that should be most sought after. Quadrants Q2 and Q4 and not color coded denote low cost low performance players and high cost high performance players. These quadrants are a bit trickier to distinguish but we generally want to avoid these players when aiming to pick up good value players. There are obviously exceptions to this rule, as someone like Alex Ovechkin (The best scorer in the NHL) would be an obvious addition to any team even for a high cost, but typically elite players are compensated highly because they are predictable and can make a huge difference for a team.

**Linear Regression, Chicago Blackhawks:**



The two players circled (Vernette and Sharp) were let go by the Chicago Blackhawks at the end of the 2014-2015 season. This that the Chicago Blackhawks, the most dominant team in the NHL for the last decade, is implementing a model very similar to the one described in this report.

**Linear Regression, Minnesota Wild:**



The 4 players circled (Stewart, Brodziak, Bergenheim, and Cooke) were released at the end of the 2014-2015 season. The single player circled to the left of our regression line is Vanek, who is in a contract that ends in one year. The reason he is circled but the other players are not is that they have 3+ year contracts and it is extremely expensive to buy out a player who has multiple years left on their contract. It costs 2/3 of their guaranteed total salary over 2x of the next years on their contract. For example, if We wanted to get rid of a player who makes 5 million per year and has 3 years left on their contract, we would have to pay them 1.66 million dollars per year over the next 6 years. While this would free up a lot of cap space, it would be an extremely bad financial decision. It makes more sense to buy out a player when they have 2 or less years left on their contract, which Vanek does.

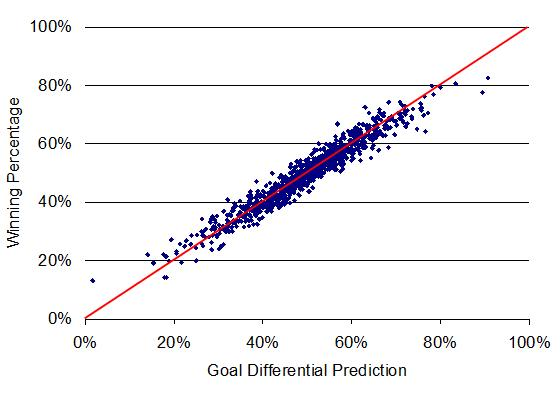
The below spreadsheet has the 4 players that our model has determined that the MN Wild should replace. There are four potential replacements for each player to meet our cap space of 9.2 million dollars.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Thomas Vanek Replacement Options (Average 39.2% increase in performance)** | | | | |
| **First Name** | **Last Name** | **Cap Cost** | **Goals/TOI** | **CapCost/(Goals/TOI)** |
| **Thomas** | **Vanek** | **6.5** | **0.016194956** | **401.3595344** |
| **Max** | **Pacioretty** | **4.5** | **0.023847889** | **188.6959471** |
| **Patric** | **Hornqvist** | **4.25** | **0.022125852** | **192.0829987** |
| **Jiri** | **Hudler** | **4** | **0.022068769** | **181.251614** |
| **Wayne** | **Simmonds** | **3.975** | **0.022211645** | **178.9601806** |
| **Kyle Brodziak Replacement Options (Average 83.6% increase in performance)** | | | | |
| **First Name** | **Last Name** | **Cap Cost** | **Goals/TOI** | **CapCost/(Goals/TOI)** |
| **Kyle** | **Brodziak** | **2.833** | **0.009451796** | **299.731395** |
| **Nick** | **Foligno** | **3.083** | **0.020838935** | **147.9442208** |
| **Mathieu** | **Perreault** | **3** | **0.017858915** | **167.9833293** |
| **Brock** | **Nelson** | **2.825** | **0.015356265** | **183.9640043** |
| **Kyle** | **Okposo** | **2.8** | **0.015341345** | **182.5133324** |
| **Chris Stewart Replacement Options (Average 62.6% increase in performance)** | | | | |
| **First Name** | **Last Name** | **Cap Cost** | **Goals/TOI** | **CapCost/(Goals/TOI)** |
| **Chris** | **Stewart** | **0.926** | **0.010883085** | **339.242044** |
| **Mark** | **Stone** | **0.873** | **0.01909378** | **45.72169576** |
| **Brendan** | **Gallagher** | **0.87** | **0.017644464** | **49.30725014** |
| **Jimmy** | **Hayes** | **0.925** | **0.017426396** | **53.08039597** |
| **Lee** | **Stempniak** | **0.9** | **0.016607617** | **54.19200118** |
| **Sean Bergenheim Replacement (Average 49.8% increase in performance)** | | | | |
| **First Name** | **Last Name** | **Cap Cost** | **Goals/TOI** | **CapCost/(Goals/TOI)** |
| **Sean** | **Bergenheim** | **2.65** | **0.012251565** | **224.4611199** |
| **Tomas** | **Tatar** | **2.75** | **0.021799594** | **126.1491384** |
| **Jaden** | **Schwartz** | **2.35** | **0.020460358** | **114.8562503** |
| **Chris** | **Kreider** | **2.475** | **0.016707773** | **148.1346437** |
| **Jannik** | **Hansen** | **2.5** | **0.014136773** | **176.8437535** |

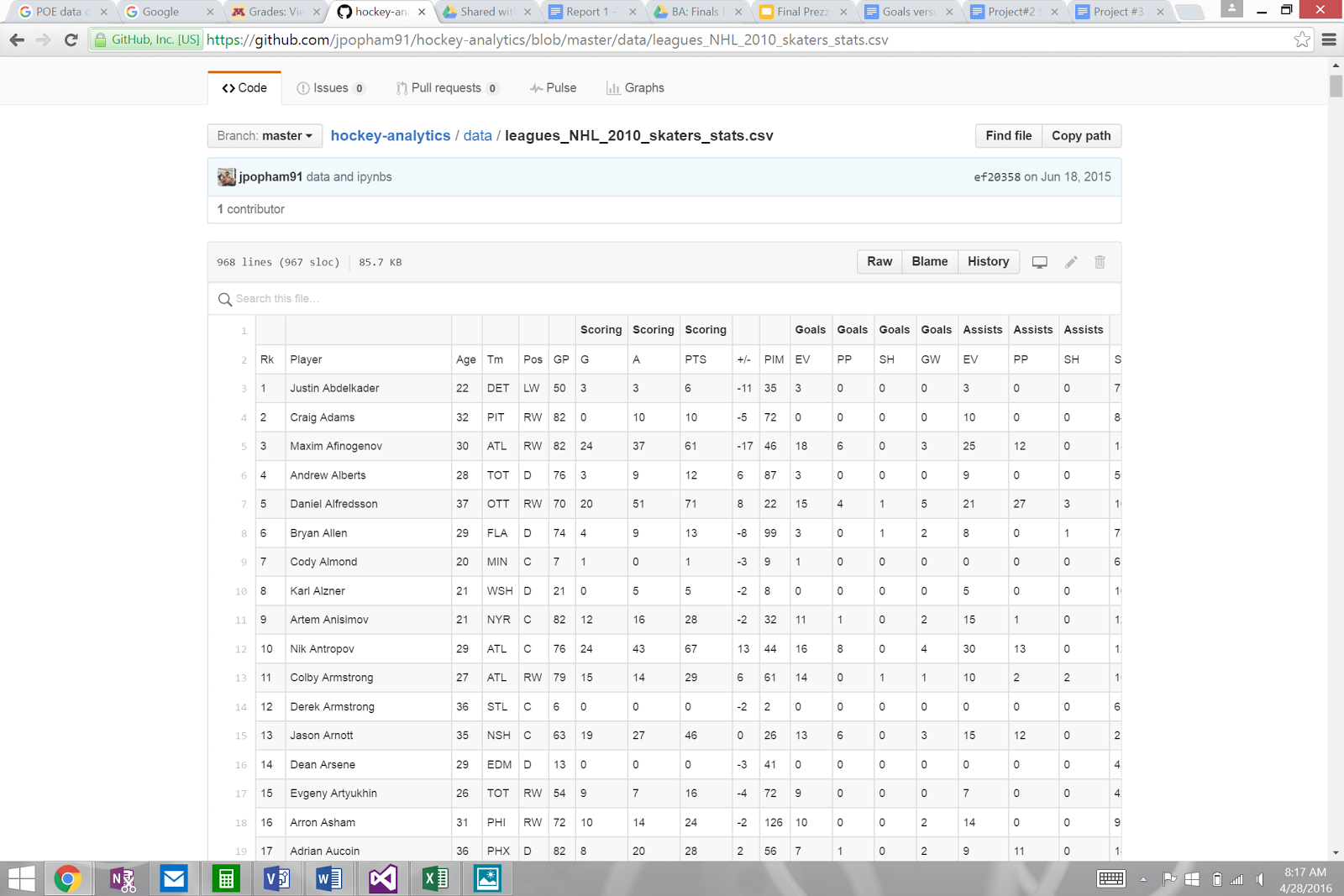
**Comparisons with documented results**

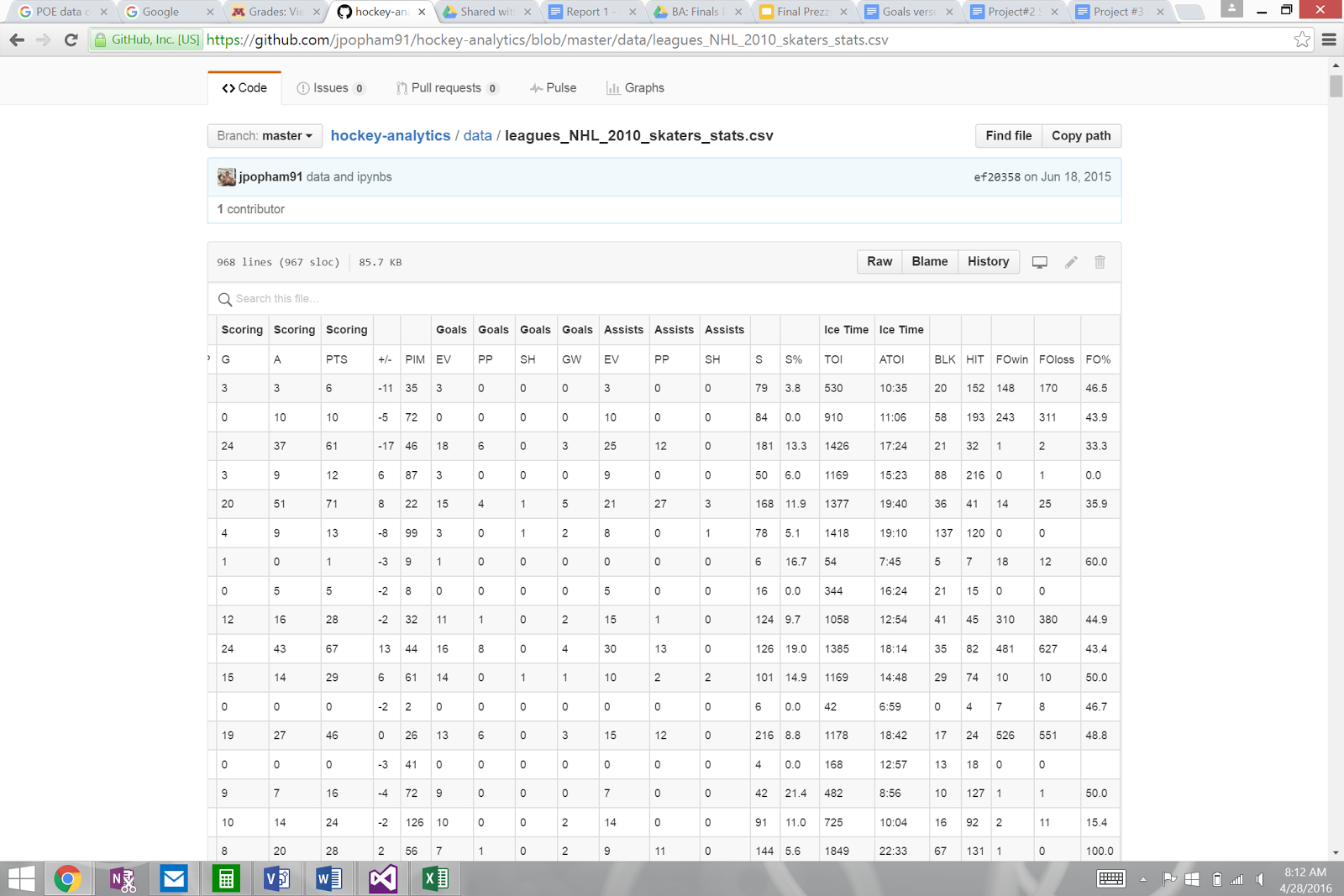
One thing that is very difficult for us to overcome is the fact that there is nothing like this published on the internet. Hockey analytics is a relatively new thing, and is one of the last sports to officially adopt the strategy. So we are going at this project blindly in a sense, equipped only with our limited data mining skills. Since there are no immediate published results on the internet to compare with, our next option was to look for something close to it in another sport. We did find examples on sites like kaggle but they were about baseball. We also found another promising dataset and model but it was also in baseball. Which is still in the sports realm, but hockey and baseball are very different, especially since baseball has a lot more statistics to track making it easier to create models that will lead to insights. So trying to compare the data has been more like trying to compare apples to oranges.

One of our team members did find a site that had some interesting topics about hockey analytics, mostly ways on how to approach the science of analyzing hockey in a newer statistical manner. This website titled hockeyanalytics.com had a very interesting section on how the state called Goal Differential, which counts how many more or less goals you score compared to how many you give up, is related to winning percentage. Here is their representation:



We found an interesting example on the github.com website where someone had preprocessed hockey data for each player in the league. We didn’t follow the same analytical steps in analyzing the players but we did follow this example when preprocessing the data. In the data preparation, you can see that we originally had a lot of attributes for the data, but we took away some attributes just like this example has done to minimize the amounts of stuff we are looking at to find the KPIs(key performance indicators).





This was interesting and very encouraging news to discover, so we tried to possibly recreate this example and then use it create other linear regression between different variables to help us draw conclusions such as the one explained in the chart. But something that we did not realize initially was that the statistics they used were compiled over each season and was something that they had to do themselves because they are statistics that are not published on most hockey stat sites and books. So in order to either recreate this visualization with our variables and teams or create another similar one we would have to manually process all of the statistics into datasets that meet our needs for analysis.

This type of analysis takes time to complete, and it is time that we may not have. This is something that is of interest to us, we just feel that it may be more beneficial if we move in a different direction. So we decided to focus less on the team success as a whole and focus more on the individual player’s success compared to their cost to the team. We feel that we are moving in a much better direction if we focus on individual success rather than team success, because looking at each player’s performance will be easier to interpret rather than the team success because the team changes every year. We want to help the Wild accumulate the best players in a pre-emptive strike to stop the slumps that they are experiencing.

**Deployment**

Our model will be used in the meeting rooms of coaches and strategists for the MN Wild and hockey teams around the country. Our analytical work cannot give people the green light to trade player X for player Y but it can be a starting point for trying to figure out who to replace and who to pick up. This will allow NHL teams to realize which players are not contributing to the team as much as other compared to their respective salary cap distribution. When out analytical assessment detects that a player is make to much compared to their performance, we will be able to notify the management of such a team and suggest alternative players that are playing above their pay grade. With these analytical tools the teams will be able to offer players less than they are worth saving the team money and allowing it to flourish as the team has more money to spend on other players.

With the complex nature of the game of hockey there is potential for our model to be incorrect and a player could go into a slump as soon as we pick them up or go on a hot streak just after we let them go. This is inherent with the game of hockey as there are a lot more variable that factor into every game then a player's statistics and team and coach cohesion plays a large role in this sport as compared to baseball where there is just a pitcher vs. a batter to compete for points. The way hockey flows leaves it opened for a lot of luck to play a part and people have started to study the luck statistic and how to track it but as of now we believe that there is a strong argument to be made for our analytical model to be used in practice.

There could be some minor ethical problems in the use of our model but they can stem from some of the problems that are already inherent in deal making and negotiations. With information that players are being underpaid and signing them onto a deal where they will be paid at the same level or lower could be argued to be unethical but the fact is that this practice happens everyday and in realms other than sports. Our model will give organization the opportunity and an insight but we are taking these statistics from publicly available sources so there is no exclusivity to our data other than our model that we created and interpreted.

The risks associated with applying our model include signing on players who eventually lose their value and stop producing what they should compared to their salary. This risk can be mitigated by tracking the player's performance across multiple years giving us a more accurate depiction of the players value trajectory. We can also leverage the players age into our statistic to give younger or older players a respective increase or decrease in worth. This will counteract the effect of players losing their value at the end of their contracts.

Another risk is that we recommend an organization to pick up a certain player only to have them face a season ending injury. This risk is something that is inherent in the game and is somewhat unpredictable. What we can do is take into account the player's past injuries and age in order to increase or decrease their value according to how injury prone they are. This will give our model more accurate prediction of a player's worth.

Like the real world example the Moneyball exhibits, fans will become dissatisfied if our model reccomend we release or trade on of their favorite players. Fans could also just not trust computer systems making their teams hiring and firing decisions. We can mitigate this risk by being opened about the fact that we will be using analytics to provide managers with the most information possible to make the best decisions for their team. If you leverage it and provide information to people about how this will help the organization, fans will likely embrace the chance to have a better team.