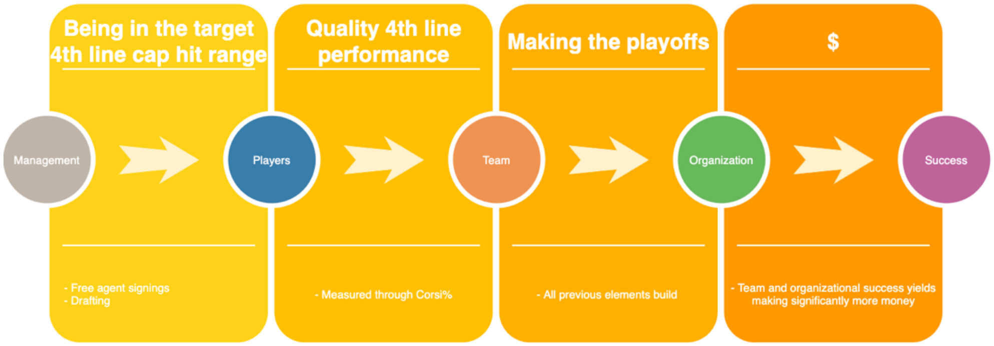
Navigating the NHL Salary Cap Crisis Through the Fourth Line

Intro:

The NHL is facing an unprecedented issue it has never faced before due to unpredictable circumstances that has left many of the teams in the league financially handcuffed. The issue is that in the NHL every team has a salary cap ($82.5 million dollars as of now) that they are allowed to spend on the players of their team annually. Typically, because of league revenue the salary cap rises from year to year or every few years by a few million. Because of this NHL front offices plan accordingly on how they will navigate staying under the salary cap threshold well into the future. However, when the league and every other business in the world is faced with an unpredicted global pandemic all of your plans go out the window. For the near future that salary cap will rise extremely minimally offering teams minimal flexibility. This is not what the league and teams' front offices had planned for and is now a huge issue for many teams as they have pressed themselves up to that cap ceiling with barely any room to budge as they had previously thought it would rise. This begs the question of how can we stay under that cap and field the best possible team on the ice to make the playoffs.

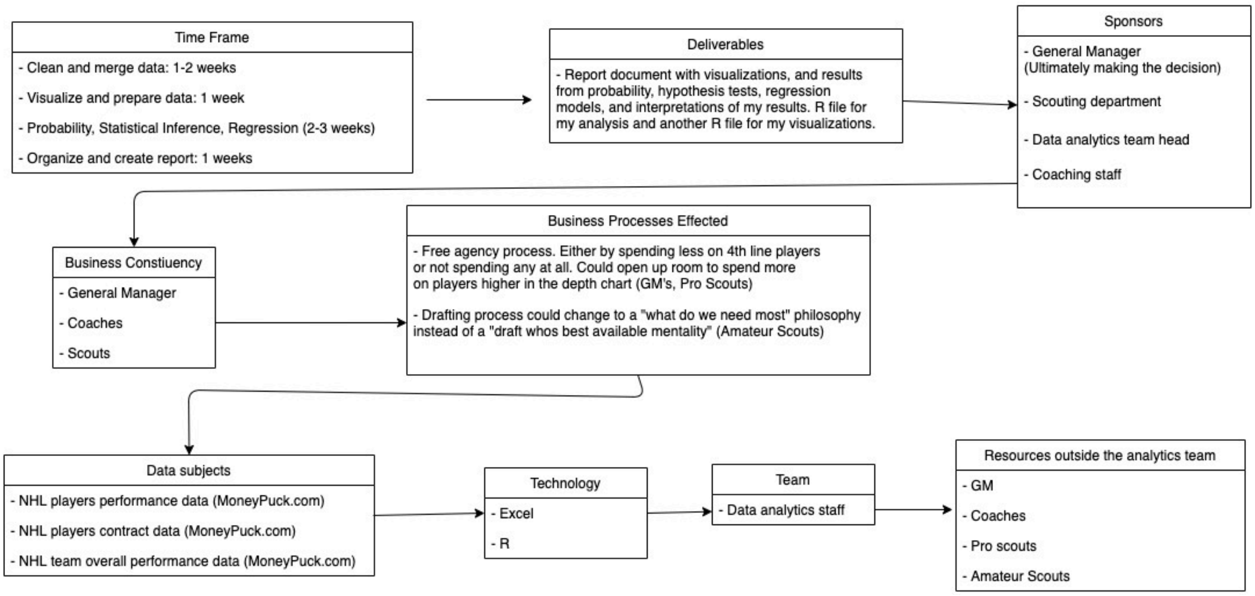
Starting by navigating this on a large scale on the piece by piece aspect on your team is almost nearly impossible when you consider all the moving parts, but there has always been a massive phenomenon in the NHL amongst coaches, analysts, media members, and general managers when it comes to fourth lines. A teams fourth line is typically composed of their "worst" forwards or in better terms the line that will play the least amount every night in games. While that may be true, these people loath great fourth line play and talk about how much of an impact they can have on games and on a team's success throughout the season. Longtime NHL media member Jeff Langridge who now writes for the *TheHockeyWriters* states "At any moment, the fourth line can change the momentum of a game at any moment. It could be a big hit, a fight or even, on the odd occasion, a goal. They may not be the most glorified members of the team, but they are important nonetheless". An even more interesting to component to these 4th lines is how they are composed which is not talked about nearly as often. The truth is these organizations need to field a quality fourth line, but at what cost? Especially with the modern salary cap realities. Some teams will go out and spend millions of dollars in the off season to hopefully improve their fourth line, but should it not improve then you are stuck with a massive 4th line cap hit (all three players on the 4th line yearly salaries added together) and a bad line. In addition to that you now have less money to allocate to much more vital spots on your team. This analysis will utilize NHL 4th line cap hits, whether NHL teams made the playoffs or not, and 4th line performance data from the most recent flat cap season to determine what is the most desirable amount NHL teams should be spending on their fourth line to yield them the best chances to make the playoffs and yield your organization much higher earnings. This formula is depicted to the right:


Being in the target 4th line cap hit range
Quality 4th line performance
Making the playoffs
$
Management
Players
Team
Organization
Success
- Free agent signings
- Drafting
- Measured through Corsi%
- All previous elements build
- Team and organizational success yields making significantly more money

Methods:

When looking at the business process of this analysis it is necessry to understand that I am doing this in the context that I work for one specific NHL teams analytics team. Our team will take roughly a seven week course from start to finish of gathering our data to articulating our results. We will deliver a report that includes all our visualizations and statistical test results along with two R files with one being for our analysis and the other being for our visualizations. For this project to be put in motion we will need approval from the key heads of our organzation and that being the general manager, scouting department, data analytics team head, and the coaching staff. They will have to believe we can improve through the 4th line. Our general manager will get the most use from this information as he is the one directly adding and removing players from the team, but also coaches and scouts who give input on this to the GM. All of the data provided in our research is publicly available on MoneyPuck.com and the technologies we will be using on it is R to do analysis and visualizations and Excel to just read our data into RStudio. The team will consist of just data analytics members of the organzization with additional help from the GM, coaches, and scouts. This project will come at no extra cost.

NHL Salary Cap Navigation Road Map



Time Frame
- Clean and merge data: 1-2 weeks
- Visualize and prepare data: 1 week
- Probability, Statistical Inference, Regression (2-3 weeks)
- Organize and create report: 1 weeks

Deliverables
- Report document with visualizations, and results from probability, hypothesis tests, regression models, and interpretations of my results. R file for my analysis and another R file for my visualizations.

Sponsors
- General Manager (Ultimately making the decision)
- Scouting department
- Data analytics team head
- Coaching staff

Business Constiuency
- General Manager
- Coaches
- Scouts

Business Processes Effected
- Free agency process. Either by spending less on 4th line players or not spending any at all. Could open up room to spend more on players higher in the depth chart (GM's, Pro Scouts)
- Drafting process could change to a "what do we need most" philosophy instead of a "draft whos best available mentality" (Amateur Scouts)

Data subjects
- NHL players performance data (MoneyPuck.com)
- NHL players contract data (MoneyPuck.com)
- NHL team overall performance data (MoneyPuck.com)

Technology
- Excel
- R

Team
- Data analytics staff

Resources outside the analytics team
- GM
- Coaches
- Pro scouts
- Amateur Scouts

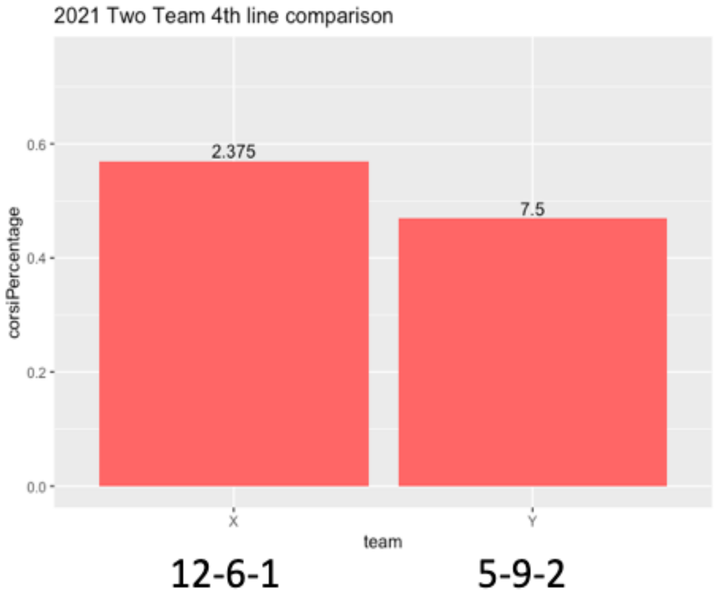
When looking at the readiness of our organization to take out this project it is clear we are capable and ready. We have a great dataset from every game and team of last season along with ample people on the analytics staff to carry out the analysis. This organzation believes analytics has a spot in hockey and wants to use it to our advantage. The financial risk of this project is embedded in the amount you spend on your players on your fourth line which is required regardless. We are currently pressed against the cap like every other team and in order to navigate and be successful we will bridge that gap by doing a thorough investigation on last seasons data.

Should we be successful our organzation can anticipate two business processes changing. The first being on how we plan to build our fourth line. We may now resort to filling that spot with cheaper players from our minor league team or on the contrast going and spending more in free agency. Second, we may also draft on more of a need basis compared to what the best available player is. Finally we can assume more organizational and team success coupled with earing much more money in the future seasons as we will now be a playoff team.

Our analytics team feels extremely confident in the results we will be able to produce from this project because almost everything from the NHL today is measured and quantified. From MoneyPuck.com we were able to gather data on every NHL teams most frequently deployed 4th line from the 2020-2021 season. This data origionally came from two data sets, one being the 4th line with their teams results from the season (wins, losses playoff outcome, etc). The other data set being that teams 4th line with the lines performance stats (cap hit, corsi %, hits, etc). While its amazing for us to have this much data it also comes with two inherint problems. The data sets are not merged and they are filled with loads of variables and we need to determine which variables are most important to us. By merging them and determining which are most important to us we had a clear scope on how to carry on with the visualization and analysis of our datasets. For non hockey fans a data dictionary on what we selected is provided below:

- Cap: The amount every team is allowed to spend on their players (79.5 million USD).
- 4th Line Cap Hit: How much money all three players on the fourth line costs against the overall on cap.
- Playoff indicator: Shows a 1 if the team made the playoffs and a 0 if they did not.
- Corsi %: measured by taking the number of shot attempts at even strength a line has and then dividing it by the shot attempts of the opponents line at even strength. In easier to understand terms this is measuring the amount of shots your line is getting compared to the amount they are letting up while they are often on the ice. This can indicate how much they are controlling play and chances they are getting. Above 55% is considered elite.

Corsi percentage is something that is very vital to us predicitng overall team success through our fourth line because it can tell the story of if your line is controlling the play when they're on the ice or not. Relating back to the quote from Jeff Langridge he denotes that 4th lines are not always scoring, but if they are controlling play and most importantly not getting scored on then you are in a great space for the rest of your team to succeed.



Through this data you can see the type of in depth visualizations we are allowed to make. This visualization is from the current season, but it adds to the emphasis on how teams are struggling to be successful and navigate the flat cap. Team X has an elite record (denoted below X axis) and their fourth line has an elite corsi perentage. We are also able to lay each teams 4th line cap hit on top of these bars and you can see team X has a much lower 4th cap hit meaning they also have a ton of money to spend on the rest of their team. Team Y could not be a more opposite story and they were expected to be amongst the NHL elite this season.

Now that our data is clearly defined and merged we can begin to look at the statistical methodology we will use to yield the best possible results for our research and organization. We have defined that we want to use this data to give ourselves the best chance of making the playoffs because making the playoffs denotes franchise success and with that comes with a lot of additonal money. Getting additional money after losing so much from COVID-19 is paramount for franchines right now. To do this first we will utilize logistic regression models to predict an indicator variable which will yield a 1 if the team is predicted to make the playoffs or a 0 if the team is not predicted to make the playoffs. Our models will build in the sense that our first model will just utilize 4th line cap hit to predict our playoff indicator variable. Our second model will utilize the 4th line cap hit and the 4th line corsi percentage to predict the playoff indicator variable and finally our third model will utlize both of those again, but it will also include an interaction term between 4th line corsi percentage and 4th line cap hit. Once these models have been ran we will use our predicted playoff outcomes and compare them to the actual playoff indicator variables from last season to acquire an accuracy measure of our models. We will also use k fold cross validation as another means of getting an accuracy score.

Once we have selected the best model and we have the estimate variables for each predictor we can input variables into the model equation and see not only the effect they have have on the playoff prediciton, but also start to learn desirable values for corsi % and cap hit you will want to try to be at to yield the best chance of your team making the playoffs. Using that new found information we will run probability tests on them to see our liklihood of being inside a certain target range on cap hit to see if it is feesable for out organization to aspire to get there or not.

Moving on from probability testing we will move to affirming our beliefs in these stats and seeing their significance. We will run a two sample hypthesis test to see if playoff and non playoff teams are spending the same or different amounts on their fourth lines. Should this test fail to reject our null hypothesis we will not be able to completely affirm their is statistical proof that playoff and non playoff teams are spending different amounts on their fourth line, but it will still strengthen our belief that there may be some truth to it.

Our team had much debate on whether to utilze linear regression models to analyze things like 4th line corsi% or 4th line cap hit, but it was ulitmately determined that doing so does not set out to achieve the goal of this project. If we set out to use a linear regression model for those means then that would mean the goal of our project is to build a fourth line with a high corsi percentage and proper cap hit, but we do not want to have just a good fourth line, we want to have a good team that makes the playoffs. A fourth line with a high corsi% does not bring your organization more money. Therefor this idea was scrapped, but it also worth noting based off our beliefs that a playoff team will inherintly have a 4th line with a high corsi percentage and ideal 4th line cap hit. When doing probability testing we also debated doing probabilities for certain corsi percentages, but it does not seem fair because there are so many factors that go into the making of a corsi percentage.

Results:

Figure: 1.1

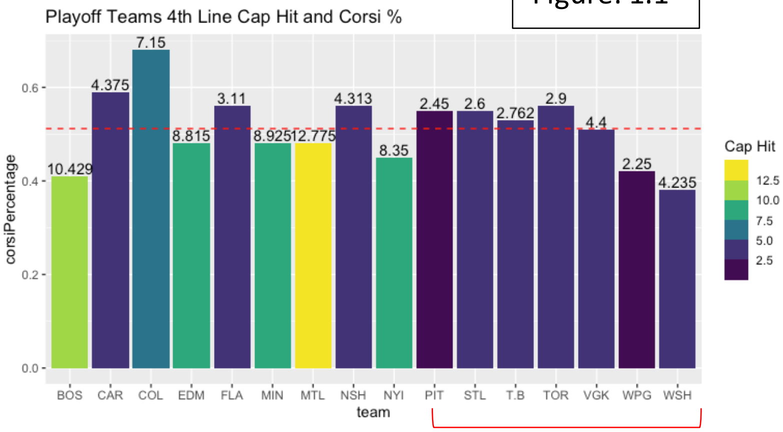
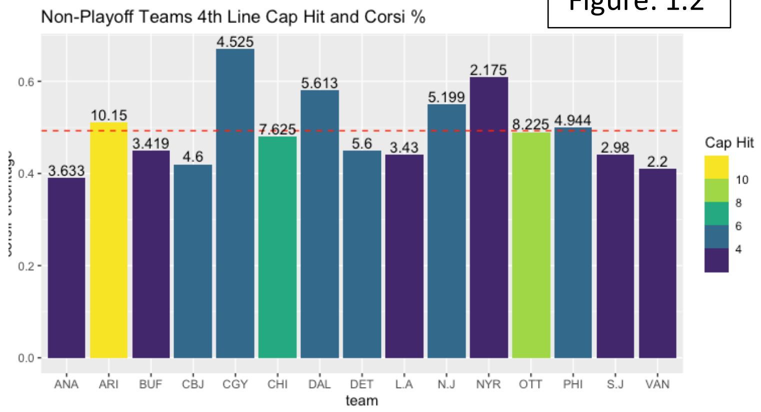


Figure: 1.2



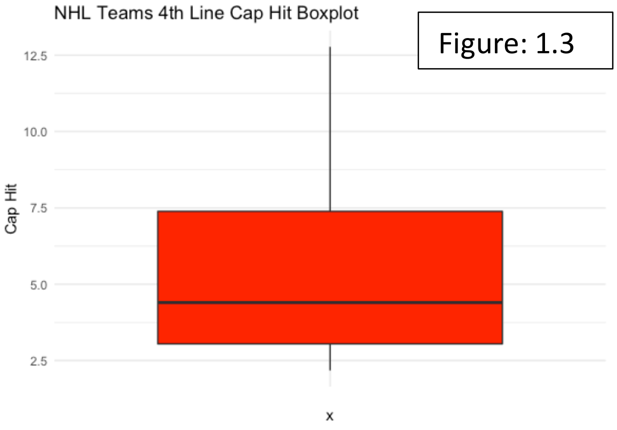
The first results we will discuss is those of our visualizations as they allowed us to see some potential trends in our data and helped clear the picture for what we wanted to dive deeper into for the project. Many of our visualizations had a focus on plotting 4th line cap hit, 4th line corsi percentage, and the playoff indicator from the 2020 – 2021 season together as they are what we desired to work with.

Figures 1.1 and 1.2 differ in the sense that they are broken up by teams that made the playoffs (1.1) and teams that did not (1.2). Along with that there is a dashed average value line running through each of them. This line denotes the average 4th line corsi percentage for the playoff teams and the average for the non playoff teams. While we do see the average 4th line corsi percentage just over .5 for playoff teams and just under .5 for non playoff teams one of the most intruiging things from these is the grouping of teams from PIT to WSH in figure 1.1. The league average for 4th line cap hit is 5.29 million and they are all below that along with 5 of those 7 teams having a corsi percentage at or above the mean for playoff teams 4th line corsi percentage. While it may be easier to start to draw the conclusion playoff teams have a smaller 4th line cap hit that is not necessarily the case. The playoff teams actually yield a much higher variabilty and mean than the non playoff teams as denoted in table 1.1.

Table 1.1	Mean 4th line Cap Hit	Cap Hit Standard Deviation	Mean 4th line Corsi %	Corsi % Standard Deviation
Playoff Teams	$5.61 Million	3.3	.51	.076
Non Playoff Teams	$4.95 Million	2.62	.49	.08

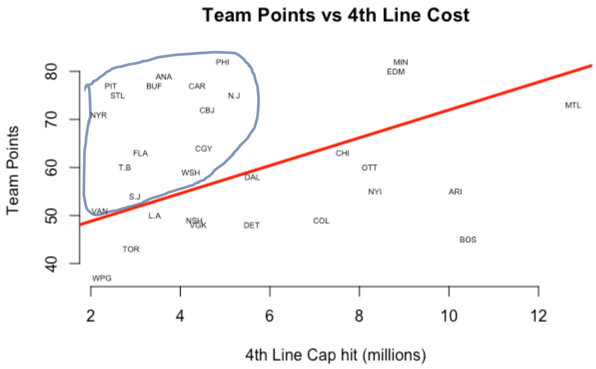
This led to ask the question whether teams like MTL could be potential outliers in the playoff data set that were causing the mean to be skewed and thus also have a much higher variability but in the histogram in figure 1.3 we see that they are indeed not an outlier. Should we have a larger sample size than just the 31 teams in the league that could potentially change.

Figure: 1.3



In the NHL teams earn points to qualify for the playoffs, if your team wins you earn 2 points, if you lose in overtime you gain 1 point, and if you lose in regulation your team will earn 0 points. At the end of the season the 8 teams in each conference with the most points will make it to the playoffs. While our goal of this project is not to maximize points earned and it is to make the playoffs, the two do go hand in hand as making the playoffs requires earning points. In figure 1.4 you see team points plotted on the y-axis with 4th line cap hit plotted on the X-axis. What you will notice about this is you see a distinct group that is circled that above our regression line for cap hit and team points. This is the desirable group you want to fall in as your team point total is above what it is expected to be based off of your 4th line cap hit.

Figure 1.4



The results of our modeling improved as we continued to build our models off of eachother. Our logistic models were utilized to predict our playoff indicator and our first one started off simple and was: Playoff Indicator = B0 + B1Cap Hit + Error Value. What you will see from this model is that our coefficienct for Cap Hit is really close to zero and if it was zero we could drop it from the equation completely because it would not influence our playoff prediction. However, it barely has some influece. This regression model yielded an accuracy of 42% in predicting playoff the indicator variable when compared to the actual results from the 2020 season. Our kfold cross which ran 5 fold cross validation yielded an accuracy of 45% Once we have analyzed each individual model we will go into more depth on comparing their accuracies. Because are p-value for Cap Hit is above our signifiance level of .05 we fail to reject our null hypothesis for individual significance meaning that Cap hit is not indvidually significant in derterminig a teams playoff indicator. The results from our first model are summarized in table 2:

Table 2	Coefficient	Std. Error	Z-value	P-value
Intercept	-0.394	0.783	-0.504	.614
Cap Hit	0.087	0.133	0.657	.511

Our second model built off of using Cap Hit as a predictor variable and now also utilizes corsi percentage as a predictor too. The logistic regression equation is as follows: Playoff indicator = B0 + B1Cap Hit + B2Corsi% + Error Value. This model showed improvements over our first model as we have an improved accuracy to 58% now, but our kfold cross validation accuracy dropped to 42%. When testing for joint significance our null and alternative hypothesis were H0: Cap hit = corsi % = 0, and HA: at least one of Bj != 0. We still failed to reject our null but we did have p values that were starting to drop. Failing to reject our null means that together cap hit and corsi percentage are not significant in predicting our playoff indicator variable. Table 3 summarizes our data for this test below:

Table 3	Coefficient	Std. Error	Z-Value	P-Value
Intercept	-2.34	2.678	-8.75	.381
Cap Hit	0.09	0.134	.729	.466
Corsi %	3.76	4.934	.764	.445

Our third and final aims to build off model 2, but now to utilize an interaction term between Cap hit and Corsi%. It is now represented as follows: B0 + B1Cap Hit + B2Corsi% + (corsi% * Cap Hit) + Error Value. This model had our best accuracy yet of 68% and 58% for k fold cross validation.The results are noted in table 4 below:

Table 4	Coefficient	Std Error	Z-Value	P-Value
Intercept	-9.309	6.681	-1.393	.164
Cap Hit	1.551	1.276	1.215	.224
Corsi %	18.061	13.440	1.344	.179
Cap Hit * Corsi%	-2.996	2.603	-1.151	.250

When looking at table 5 below it is easy to see that model 3 is our best model to go with because of it achieving the highest accuracy in both standard accuracy and k-fold cross validation. It also was the closest to achieving joint significance by having the smallest p-values. Now that we know which is our best model we can start plugging in numbers to our equation and getting a desired Cap Hit range to be in.

Table 5	Accuracy	K Fold accuracy	Joint Signifcance	Individual Significance
Model 1	42%	45%	No	No
Model 2	58%	42%	No	No
Model 3	68%	58%	No	No

When beginnig to plug corsi percentages and cap hits into our equation it became extremely apparent that corsi% had a massive weight on the overall indication. Plugging in a near average corsi percentage of .52 and a near average cap hit of $4.5 million only yields a 5% chance of making the playoffs. We then looked to see what the probabilty a randomly selected team had a cap hit below $4.5 million and that showed that it is roughly a 39% chance of happening. It is apparent that we need to pump our corsi percentage up so we can get a better chance of making the playoffs so we bumped it up to an elite percentage of .58. By doing this we found that if your Cap hit was $2 million it indicated you have a 76% chance of making the playoffs and if it was $3.75 million you had a 46% of making the playoffs. Now that we have two more appropriate ranges of a cap hit to make the playoffs given you have an elite corsi percentage of 58% we can run probability on that to see what the chances are a randomly selected team will fall in that range of $2 to $3.75 million for their 4th line cap hit. We found that selecting a random team yields about a one in four chance of having a cap hit in that range. With all of our probability we can invoke the Central Limit Theorem because our number of observations is greater than 30. This states that the sum of observations from are distribution take on an approximately normal distribution.

Over the course of this project we as a team have been curious to see if playoff teams are spending a significantly different amount than non playoff teams. We will form a hypothesis test on this at the 95% signficance level where our null hypothesis is HO: Playoff teams are spending the same amount on their fourth line compared to non playoff teams and our alternative hypothesis is HA: Playoff teams are spending a different amount on their fourth line compared to non playoff teams. Mathematically written as H0: (μ playoff $) = ($\mu$ Non-Playoff $) and Ha: ($\mu$ playoff $) $\neq$ (μ Non-Playoff $). From running a two saple two tailed t-test on this we see get a p-vaule of .52 which is loads higher than our significance level of 0.05 meaning that we fail to reject our null hypothesis which states that playoff teams are spending the same amount on their fourth lines as non playoff teams. We cannot say whether playoff teams are spending a significantly different amount on their 4th line compared to non playoff teams.

Conclusion:

Through our analysis we can first take pride in our logistic regression model which hit nearly 70% accuracy for prediciting if a team will make the playoffs or not utilizing their 4th line cap hit, 4th line corsi percentage, and both of them together as an interaction term. However, we were not able to achieve joint significance meaning we cannot state that together cap hit, corsi%, and them as a interaction term are significant in predicting our playoff indicator. Despite this fact, our third model still was the closest to joint significance and had the best accuracy. When testing input variables on our model it become obvious the weight that corsi percentage had in the prediction for playoffs in our model. So if we find a desired cap hit value by plugging numbers into our equation it will be fairly easy to set out and get players who will sum up to that desired cap hit, but one weakness of this model is that when you go out and get players you truly do not know what their corsi percentage will be until they play together. So based off inputs in our model you could decide you want a 4th line with a $3 million dollar cap hit and a 56% corsi percentage. Well the only thing you can truly control as a front office is that cap hit, the corsi percentage is a fate you will find out once you put that product on the ice. So you could end up spending $3 million and getting a 48% for corsi. With that said if our team would like to target a desired corsi percentage when building a fourth line we would need to dive into other advanced stats for individual players and see if we can model out three players that yield a high corsi percentage for the potential fourth line. This would be similar to the linear regression process we decided not to do. Another shortcoming is that our model only utilizes data from the fourth line, a hockey team is composed of three other lines, three defensive pairs, and two goalies. A teams chances of making the playoffs does not soley ride or die on their fourth line factors. While there are shortcomings it is still extremely valuable information to take in as it has proven accuracy and should you be able to use this build a quality 4th line you will have one huge stone in place to building a succesful team.

Our probability analysis shows that you have a fairly low chance of falling in the desired cap hit range to be in assuming you have an elite corsi percentage, but should you be able to do that you will be setting yourself up for a very high chance of making the playoffs.

The hypothesis testing derived that we are not able to confirm that playoff teams are different amounts on their fourth lines compared to non playoff teams. This does not strenghten our beliefs as their difference is not significant, but it is still worth looking into. Based on our findings through the model, probability testing, and hypothesis testing we can state that in order for our franchise to maximize playoff chances we want to aim to have a high corsi percentage, but aire on having the lowest cap possible to achieve that.

Because of this flat cap world we now live in it has forced our analytics team and organization as whole to seek new ways to put a playoff team on the ice. Change brings innovation at all levels and in the this aspect the change of the cap not changing has forced our team to begin building itself through a completely new road that is the 4th line of our team. By doing this we are helping ensure our best possible chances of success in making the playoffs and increasing franchise revnue by doing so.