



**Using NHL Shot Data to Predict Shot  
Success and Measure Player Ability**

## Table of Contents

Abstract/Executive Summary	3
Project Plan	4
Exploratory Data Analysis	12
Methodology	21
Data Visualization and Analytics	26
Ethical Recommendations	37
Challenges	40
Recommendations	42
References	44
Appendix	47
Code	75

### Abstract/Executive Summary:

The National Hockey League (NHL) is the premier professional hockey league in North America. Thirty-two teams compete for the Stanley Cup, with victory dependent on scoring more goals than the other team. Scoring more goals requires a combination of strong offenses and strong goaltending. Our analysis aims mainly to help NHL teams and their front offices make better informed roster decisions by effectively measuring players' shooting data, balancing shot quality and quantity, and identifying consistency and sustainability within goaltenders, a position with notoriously inconsistent performance.

Our models have other potential applications too. They can be used by underrated players' agents to negotiate higher paying contracts for their clients and employed by players to identify weaknesses to improve within their own play, as well as opponents' strengths to avoid and weaknesses to exploit.. Finally, our models can be used by fantasy league players and bettors to make better informed decisions about their own roster moves and bets.

We sourced our data from MoneyPuck.com. We utilized their shot data to answer our first research question. Starting with the 2007-2008 season, MoneyPuck has data for every shot attempt in every season. We used data from the 2021-2022 season through February 3rd of the 2024-2025 season, totaling 437,826 shot attempts with 137 columns per observation. These shot attempts were tracked over 4,780 regular season games and 3 playoff runs (2022, 2023, 2024). We utilized their goaltender data to answer our second research question. We only used data from the 2023-2024 season, totaling 98 goalies. The dataset provides 49 columns per observation.

Using methods such as logistic regression, k-Nearest Neighbors, random forest, k-means clustering, DBSCAN and hierarchical clustering, we wanted to find which variables were most predictive of a shot being a goal during the regular season, and which goaltending performance

based clusters could be identified based on effectiveness against different shot variables. We found that the random forest model was best suited for predicting the accuracy of a shot, with a tuned accuracy score of 0.93, a tuned precision score of 0.63, a tuned recall score of 0.05 and a tuned f1 score of 0.09.

We also clustered NHL goalies according to their statistics, using the same source for goalie data as we did for shot data. We compared three clustering analysis methods - k-means clustering, DBSCAN, and hierarchical clustering - and found that k-means clustering was the best method, with a silhouette score of .645. Due to the features we included in creating our cluster analysis, the outcome was a clustering of goalies mainly based on their icetime and number of shot attempts faced, separating starters, backups, and goalies with few appearances.

Project Plan:

Team Lead - Tyler Briles

Member - Kel Lau

Member - JP Stabner

Member - Wendy Price

Company Details:

Founded– November 26, 1917

Founders– Montreal Canadiens, Montreal Wanderers, Ottawa Senators, Quebec Bulldogs

Headquarters– New York City, New York

Category– Sports

Address:

New York

1 Manhattan West

395 Ninth Ave.

New York, NY 10001

Company Communication:

Phone: 212-789-2000

Fax: 212-789-2020

### Business Description/Profile

The National Hockey League (NHL) is a professional ice hockey league based in North America. It comprises 32 teams: 25 located in the United States and 7 located in Canada. Each team plays an 82-game regular season, split evenly between 41 home and 41 away games. At the end of the season, the top 3 teams from each division along with the next two highest-ranked teams in each conference (wild cards), 16 teams total, qualify for the Stanley Cup Playoffs. The playoffs follow a best-of-seven, single-elimination format, culminating in one team being crowned the Stanley Cup champion.

### Key Executives:

Commissioner– Gary Bettman

Deputy Commissioner– William Daly

Senior Executive Vice President of Hockey Operations– Colin Campbell

### Analysis Opportunity:

Leveraging MoneyPuck's datasets on shots from 2021-2025, our data team will analyze key factors influencing shot success and goaltending performance. This dataset includes roughly 435,000 shots, annotated with variables such as shot distance, shot angle and game state. Using these insights, our team will develop segmentation and prediction models to identify critical performance-driven indicators, enabling more accurate predictions of shot success and goaltending effectiveness.

### Research Questions:

Prior to the 2004-2005 lockout, NHL teams were able to spend as much money as they wanted, and the wealthier teams had a competitive advantage. However, starting in 2005-2006, NHL teams have faced a salary cap when building their Stanley Cup contenders. The salary cap is currently \$88 million, and over 50% of teams are within \$1 million of the salary cap (Puckpedia). Now that spending more money than other teams is no longer an option, and with so many teams spending to the cap, finding more value per dollar has never been more important to teams looking to win Lord Stanley. Not only does icing a competitive team keep stadiums full and generate more revenue during the regular season, playoff ticket sales generate additional stadium revenue, and attention from winning brings higher ad revenue for a team. The further teams can stretch money for players, the more money the organization generates off the ice. We will use the following questions to research how teams can earn more value for each of their salary cap dollars.

#### **RQ1: What variables are most predictive for a shot being a goal during the regular season?**

Hockey games are won by scoring more goals than the other team. Goals can only be scored by shooting the puck on net. However, not every shot is created equal. Low quality shots can waste possession, but teams' defensive structures are designed to prevent high quality opportunities, and high quality shots are not always available. Finding the balance between not wasting possession and taking available shots is important to scoring more goals than the other team. This research question attempts to identify key predictive factors - such as shot distance, angle, game situation, and defensive pressure - that influence the likelihood of a shot resulting in a goal. By uncovering patterns in shot data, our analysis aims to provide insights into optimal shot selection and how teams can generate more efficient scoring chances.

**RQ2: What performance-based clusters can be identified among NHL goaltenders based on their effectiveness against different shot variables?**

The goaltender position is one of the hardest to quantify in sports. Driven more by mental strength than physical ability, goaltenders' performances are notoriously fickle. Teams tend to be cautious when signing goalies to contracts for fear of these inconsistencies, yet many a desperate team still signs a goalie to a contract they quickly regret. We want to help teams identify consistency and sustainability in goaltenders' play, give goalies actionable areas for improvement, and provide offensive players with goalies' weaknesses to exploit.

Hypotheses:

H1:

Certain shot qualities will have a greater correlation to the shots' results and can be used to create a predictive model. We will use these key predictor variables, such as shot location, type, speed, and angle, to create shot prediction models and test our hypothesis.

H2:

Goalies with similar performances will be grouped together. These clusters will provide meaningful comparisons of goalies to their peers. Outliers may suggest future regression to the mean. We will use the predictive model variables created for RQ1 and other key shot variables to group goalies and test our hypothesis.

Data:

Our data is sourced from MoneyPuck, a hockey advanced analytics website. The dataset contains structured data measuring NHL shots from all games starting in 2021-2022 continuing through the present, totaling in roughly 435,000 instances. Variables include data about the shot,



including location, speed, type of shot, shooter, goalie, and many more. While the dataset also includes precomputed variables derived from Money puck's existing models, we will exclude these from our analysis. Instead, we will generate our own derived variables, such as a shot's expected result, goal probability, and rebound probability, a shot pressure score measuring defensive pressure to the shot, and a shooter fatigue score derived more a shooter's shift time at the the time of the shot, using the machine learning models developed for Research Question 1. These derived metrics will allow us to assess shot quality and aim to optimize predictive accuracy based on our own modeling techniques.

#### Measurements:

Our in-game measurements include the shot's result, goal or no goal, and the shot's context, like location, type, speed, and situation. We will analyze the relationship between these contexts and the shots' results to generate a predicted result based on a shot's context.

Money puck's dataset provides all in-game shot measurements, including results, and their definitions. Every shot's raw qualities are captured and provided by the NHL.

#### Methodology:

RQ1:

We want to measure shot qualities to create predictive shot result models. We will use exploratory data analysis to measure correlations between predictor variables and a shot's outcome, use principal component analysis to identify variance within the significantly correlated predictor variables, then use classification models on the principal components to quantify shots' qualities and make shot result predictions.

RQ2:

We want to analyze goaltenders' performance to other goalies' performances within and outside of their groups. We also want to analyze movement within these groups over our datasets' years and investigate whether group placement can be used to predict future results. We will measure goalie performance by comparing their shot results to the principal components identified in research question 1 and predictive models used to answer research question 1. We will cluster results against each variable and use visualization techniques to analyze the results and draw conclusions.

#### Computational Methods and Outputs:

Our RQ1 outputs will be the predictions made by our shot quality models, denoting a shot's likelihood for success. Our RQ2 outputs will be clusters of goalies grouped by performance and our visualizations of these groups.

We feel AUC/ROC will be the best metric to evaluate our classification models. We will use a validation set to prevent overfitting and measure the results on a test dataset.

#### Output Summaries:

RQ1: What variables are most predictive for a shot being a goal during the regular season?

This analysis will be able to identify which factors are the most important in a shot's success. We will use a feature importance bar chart to show how important a variable is (x-axis) compared to the variable itself (y-axis) (see appendix A for an example). We will also feature a correlation heatmap to show the strength of correlation of variables compared to goals (see appendix B for an example).

RQ2: How can we group NHL goalies based on their performance against different shot variables to identify patterns or elite performance clusters?

This analysis will be able to identify which goaltenders are performing better compared to different shot variables. We will use a clustered scatter plot to display shot variables and a goalies performance against them (see appendix C for an example).

### Implementation:

Hockey teams and analysts are increasingly turning to data-driven decision-making to gain a competitive edge. While expected goals models provide valuable insight, machine learning offers a more advanced approach to assess shot success and goaltender performance.

Each of our research questions has the ability to benefit both the NHL and its teams. First, developing a prediction model for research question 1: “What variables are most predictive for a shot being a goal during the regular season?” allows a team to gain deeper insight into shot success. Understanding these factors can help teams optimize their offensive strategies, improve shot selections and increase scoring efficiency. Additionally, coaches and analysts can use this model to evaluate player tendencies and develop game plans that maximize scoring chances.

Second, developing a cluster model for research question 2: “What performance-based clusters can be identified among NHL goaltenders based on their effectiveness against different shot variables?” enables teams to identify which shot types are most successful against specific goaltenders. This can help scouting departments evaluate potential acquisitions and provide valuable insights for contract negotiations, assist goaltenders in refining their training based on weaknesses, and help offensive systems create efficient attack strategies.

These research questions not only help teams and coaches optimize offensive strategies and goaltender preparation but also aid scouts and analysts in identifying underrated talent.

Beyond team strategy, they offer fans and media a deeper understanding of why certain shots succeed and provide a more comprehensive evaluation of goalie performance beyond traditional metrics like save percentage.

### Exploratory Data Analysis:

For this project, we sourced our data from MoneyPuck.com, downloading data from all shot attempts per season. MoneyPuck has this data starting with the 2007-2008 season through the present, but we chose to only use data from the 2021-2022 season through the present. This includes data from the current season, updated every night after all the day's games are finished. We pulled our data on Monday, February 3rd, and our dataset will not include any observations from the current year after this date.

We chose to exclude data prior to the 2021-2022 season to limit historical bias within the model. Because of the COVID-19 pandemic, the 2019-2020 and 2020-2021 seasons were both limited, and the 2020-2021 season included atypical division alignment to accommodate Canada's border restrictions. Additionally, offense has trended upward since 2007-2008, including a large jump between the 2020-2021 and 2021-2022 seasons (Hockey Reference). Furthermore, with 437,826 observations within our 3.5 years of data, we have more than enough observations to analyze for a project of this scope.

The dataset is all structured data and includes 437,826 observations with 137 columns per observation. We identified 13 columns that are only present in the 2024-2025 dataset and removed these to keep our data consistent. We then reviewed the remaining 124 columns and decided whether to include each column. We removed another 90 columns during this analysis, including 7 calculated columns based on MoneyPuck's predictive models. While many of these columns contained valuable information, analyzing them required greater resources than the scope of this project allows. For example, the dataset has a column denoting whether a shot was taken from a rebound and 4 more columns containing data about the type of rebound a shot was taken from. Using these additional 4 columns would have required running classification

algorithms on the rebound subset: more effort and time than we have for this project. As a result, we removed the 4 rebound specific columns and only kept the column denoting whether a shot was taken from a rebound or not.

Our 37 Columns are below:

**shotID:** Unique identification number for each shot.

**home/awayTeamCode:** A three letter string with each team's abbreviation.

**season:** Season the shot took place in. The year the season began is used.

**isPlayoffGame:** Binary classifier. 0 if regular season game; 1 if playoff game.

**Game\_id:** Unique identification number for the game the shot occurred in.

**id:** The event number of the shot in the game

**time:** How many seconds into the game the shot took place. Ranges from 1 to 8387 (4OT playoff game).

**timeUntilNextEvent:** Time between shot and the next game event.

**timeSinceLastEvent:** Time between shot and the previous game event.

**event:** Classifies whether the shot was on goal (SHOT), a goal (GOAL), or missed the net (MISS).

**goal:** Binary classifier. 0 if the shot isn't a goal. 1 if the shot is a goal.

**shotAngleAdjusted:** The absolute value of the shot's angle to the net. Ranges from 0 to 88.5 degrees.

- Not adjusted shot angle is positive from the left side of the ice and negative from the right side of the ice

**shotDistance:** The distance between the shot's location and the net. Ranges from 1 to 98.4 feet.

**shotType:** Identifies the shot type. Options are wrist shot, slapshot, snapshot, backhand shot, deflection, tip shot, and wraparound.

**shotOnEmptyNet:** Binary classifier. 0 if the shot was taken with a goalie in net. 1 if the shot was taken with the goalie pulled for an extra attacker.

- We used this column to identify and remove shots taken on an empty net from the dataset used to train and test our predictive model.

**shotRebound:** Binary classifier. 0 if the shot taken is not a rebound. 1 if the shot taken is a rebound.

- Rebound is defined as the last event being a shot and occurring within 3 seconds of this shot.

**shotRush:** Binary classifier. 0 if the shot taken is not a rush opportunity. 1 if the shot taken is a rush opportunity.

- A rush is defined as the last event being in another zone and occurring within 4 seconds of this shot.

**lastEventCategory:** The type of event before the shot. Options are Shot Block, Delayed Penalty, Coach's Challenge, Faceoff, Giveaway, Takeaway, Hit, Shot, Miss. 8 observations are classified as STOP and 1 observation is classified as EGT, but Moneypuck does not define these terms.

**playerPositionThatDidEvent:** The position of the player taking the shot. Options are L for Left Wing, R for Right Wing, C for Center, and D for Defenceperson.

**timeSinceFaceoff:** Number of seconds since the last faceoff before the shot.

**goalieIdForShot:** The unique identifier for the goalie the shot is taken on.

**goalieNameForShot:** The first and last name of the goalie the shot is taken on.

**shooterPlayerID:** The unique identifier for the shot's shooter.

**shooterName:** The first and last name of the shot's shooter.

**shooterLeftRight:** Identifies the shooter's handedness. Options are L for Left Shot and R for Right Shot.

**shooterTimeOnIce:** The number of game seconds that have passed since the shooter started their shift.

**shooterTimeOnIceSinceFaceoff:** The minimum number of game seconds that have passed since the shooter started their shift or since the previous faceoff.

**shootingTeamAverageTimeOnIce:** The average playing time in seconds the shooting team's players have been on the ice.

**shootingTeamAverageTimeOnIceOfForwards:** The average playing time in seconds the shooting team's forwards have been on the ice.

**shootingTeamAverageTimeOnIceOfDefencemen:** The average playing time in seconds the shooting team's defense players have been on the ice.

**defendingTeamAverageTimeOnIce:** The average playing time in seconds the defending team's players have been on the ice.

**defendingTeamAverageTimeOnIceOfForwards:** The average playing time in seconds the defending team's forwards have been on the ice.

**defendingTeamAverageTimeOnIceOfDefencemen:** The average playing time in seconds the defending team's defense players have been on the ice.

**offWing:** Binary classifier. 0 if the shooter is a left shot shooting from the left side of the ice or right shot from the right side of the ice. 1 if the shooter is a left shot shooting from the right side of the ice or a right shot shooting from the left side of the ice.

**shotWasOnGoal:** Binary classifier. 0 if the shot missed the net. 1 if the shot was on net.



**teamCode:** String identifier. The 3 letter team abbreviation.

We created a few new variables to compare divisions and conferences. Using the Team Codes, we separated each team into their respective conference and division.

**Conference:** Each team in the NHL belongs to either the Eastern or Western Conference. The teams in the Western conference include: Ana, Ari [now Uta - both are included], Cgy, Chi, Col, Dal, Edm, Lak, Min, Nsh, Sea, Sjs, Stl, Van, Vgk, and Wpg. The Eastern conference teams include: Bos, Buf, Car, Cbj, Det, Fla, Mtl, Njd, Nyi, Nyr, Ott, Phi, Pit, Tbl, Tor and Wsh.

**Division:** Each conference in the NHL is also split up into 2 divisions. Central and Pacific for the West, Atlantic and Metropolitan for the East. The central teams are: Ari & Uta, Chi, Col, Dal, Min, Nsh, Stl and Wpg. The Pacific teams are: Edm, Lak, Sea, Sjs, Van and Vgk. The Atlantic division consists of: Bos, Buf, Det, Fla, Mtl, Ott, Tbl and Tor. The Metropolitan division consists of: Car, Cbj, Njd, Nyi, Nyr, Phi, Pit and Wsh.

### Variable Correlations:

When considering which variables play the biggest role in a shot resulting in a goal, and therefore, our model, running a correlation matrix was a natural first step. The correlation matrix allows us to see which variables most strongly and weakly align with the result of the shot, as well as whether they were positively or negatively correlated. We may also notice variables with similar correlations and consider these variables' relationships.

Using Python, we loaded and analyzed our dataset and accounted for null values in a few categorical columns such as `goalieNameForShot`, `playerPositionThatDidEvent`, and `shooterLeftRight`. Additionally, we excluded strictly ID type columns as they provide no analytical value. We analyzed only our numeric and binary variables (see appendix D), then

analyzed all of our variables together which required one hot encoding our categorical variables (see appendix E), and finally analyzed only our one hot encoded variables against goals (see appendix F). Our heatmaps showed that no single variable has a terribly strong positive or negative correlation to our dependent variable goal.

No single variable having a strong correlation with  $\text{goal} = 1$  suggests several possibilities to explore. There is a strong likelihood that a shot resulting in a goal is due to a combination of several variables (shot angle + shot distance + goalie positioning + event immediately preceding the shot). It could also suggest non-linear relationships. If so, decision tree and random forest models will be appropriate. Additionally, there is a strong indication that feature engineering will be necessary to handle our variables. For example, we might need to create an “expected threat score” which might combine individual variables such as shot distance, shot angle, and time since the previous event.

### **Shot Type**

Since our heatmaps did not show obvious correlations for variables that lead to goals, we wanted to consider variables we thought would be likely to affect the shot’s outcome. The first variable we explored was the type of shot that was taken.

First, we looked at the number of shots taken compared to the result of the shot: a shot on goal, a shot that missed the net, or a goal. We noticed the high amount of wrist shots and the low amount of deflected and wrap shots. The low amount of deflected shots make sense as you would need a shot to first be deflected, and then to reshoot the puck. The wrap shots are harder to attempt as you have more time to be intercepted by either the enemy team or the goalie before you can even attempt a shot (see appendix G). Based on the first chart, we converted each shot type’s result counts to percentages of the total shots for that shot type (see appendix G).

Deflected shots have the highest chance of becoming a goal, which could be in part due to their lower sample size. However, when a player deflects a puck, it becomes much harder for goalies to predict where the puck will be next.

We next analyzed how shot types and their results varied between conferences and divisions. Each shot type had similar success rates between the two conferences. Additionally, each conference has higher success rates on half of the shot types (see appendix I). The differences between divisions are more uniform. The Atlantic and Central divisions have higher shot success rate than the Metropolitan and Pacific divisions, respectively, for the majority of shot types. Wrap shots and deflected shots have the biggest differences between divisions (see appendix J). These variables are also the shot types with the least amount of data.

### **Shot Distance**

In addition to analyzing the different shot types, we expected distance from the goal was another variable that will greatly impact a shot's success rate. One would generally assume that a shot from a shorter distance has the potential to be more successful for various reasons, such as the goalie having less time to react and adjust their positioning, or that the shooter is in the process of skating closer to the net, such as on a breakaway. A trend of the number of shots taken as distance increases could go several ways:

- The number of shots might decrease because the chances of the shot scoring decrease.
- The number of shots might increase because a shot from a further distance has a higher potential to be tipped or redirected on the way to the goal, making it more difficult to react to the path of the puck to the goal. There is also more potential for other players to block the goalies line of vision. Additionally, these shots are more readily available to offenses.

- A mix of these factors could cause the number of shots to be similar across varied distances.

In order to make our data easier to analyze, we created a categorical variable for shot distance ranges. These ranges were based on a diagram from Sandro Azerrad on Gaimday.com based on data from the 2018-2019 NHL season (figure 1). This allows for better comparison of shots across multiple distances in the offensive zone and gives perspective for where shots from these distances could be from. The visualizations were created using Power BI.

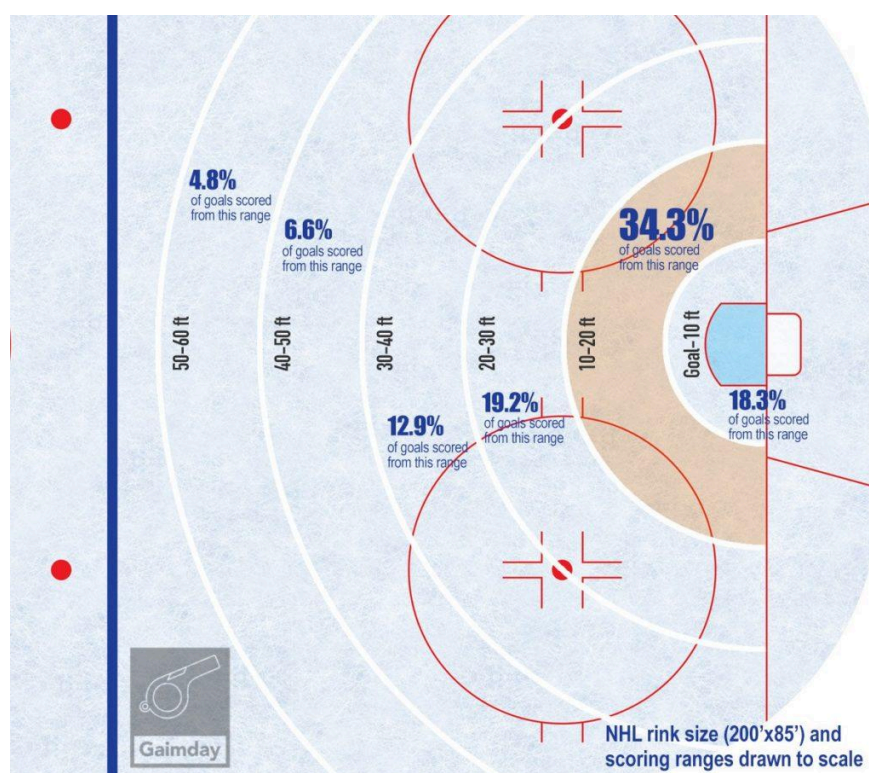


Figure 1

The chart in appendix K shows the distribution of total shots in the dataset by distance. Shots from 60 ft or more from the goal (which is generally at or beyond the blueline) or less than 10 ft from the goal are the least frequent, but not by a wide margin. There is only a difference of

approximately 9% between the least and most frequent shot distance. Additionally, shot frequency generally decreases as you increase the distance from the goal beyond 10 ft.

We also looked at the ways that shot distance may vary by division and conference, based on roster make-up, play-style, and frequently-faced opponents (see appendix L). The data did not support any differences in the distances of shots taken between conferences or divisions across the league.

We also wanted to consider some summary statistics on the distances of the shots in our dataset, such as the spread of the shot distances. These also showed little difference in things like the spread of shot distances, average, maximum and minimum distances across divisions or conferences (see appendix M and N).

The main result from our exploratory analysis of the shot distance data was intuitive - as the distance from goal increased, the number of shots that resulted in goals decreased. Shots closest to the goal also resulted in the fewest that missed the goal entirely.

### Methodology:

RQ1 - What variables are most predictive for a shot being a goal during the regular season?

Winning hockey games comes down to scoring more goals than the opponent. However, not every shot is created equal. A low-quality shot wastes possession, and the structure of the defense is designed to block or mitigate high quality opportunities. Being able to identify what is most predictive of a shot resulting in a goal is critical. To answer this question, we will develop predictive models that analyze the characteristics of a shot to determine its likelihood of resulting in a goal.

### **Data Preparation**

Our dataset was sourced from Money puck and it contained roughly 435,000 shot instances from the 2021 through part of the 2025 seasons. Each shot is recorded with multiple variables, including shot distance, shot angle, shooter identity, goaltender identity, and game situation such as powerplay, even strength, or penalty kill. Additionally, stats also included how long each player had been on the ice as well as what position the player who shot the puck played.

Before building our predictive models, we decided which columns would be kept as “in scope” for our project. Many of the columns added additional complexities to the data, but the scope of this current project would not allow for us to adequately analyze those variables. After careful consideration, 37 columns remained for inclusion into our analysis. Many of the missing values were in “id” type columns like goalie or player names. Since this information was not necessary for shot analysis, we changed those missing values to “unknown” so the shot data was kept in our data. We then converted categorical variables such as last event category, player position that did event, and shot type into numerical representations using one-hot encoding.

Using this information, we built separate correlation heatmaps to see which variables had the most impact on if a shot was a goal. The heatmaps helped us determine important variables that we'll focus on to build our predictive models such as shot distance, shot angle, time since last event, and player position.

## **Modeling Techniques**

### **Logistic Regression**

This is a model that is often used to help with binary classification problems such as ours - is a shot a goal (1) or not (0). It finds linear relationships between variables with the value of the coefficient indicating the strength and direction of these relationships. A positive coefficient would indicate an increased chance of scoring a goal while a negative coefficient would indicate a decreased chance.

### **Random Forest**

Random forest is a decision tree technique that works to identify non-linear relationships which is why it's often used along with logistic regression. A single decision tree can be prone to overfitting, so we will use random forest which averages multiple decision trees to improve accuracy while also reducing variance. Random forest will allow us a deeper understanding of the variables involved in a shot's success beyond linear relationships.

### **k-Nearest Neighbors (k-NN)**

k-NN is a distance-based classification method that predicts outcomes based on the similarity of new observation points to past observations. That similarity is measured in distance. K-NN is useful in our analysis because it does not assume linear relationships.

## **Model Evaluation and Selection**

After trying different modeling techniques, we will then evaluate the performance of those models based on accuracy, precision and recall, F1 score, and AUC/ROC (area under the ROC curve). To prevent overfitting, we will use cross-validation which is when a dataset is split into multiple training and testing sets. This helps prevent models from overfitting to a specific training set. Based on these evaluation metrics, we will determine the best modeling approach for our data. Additionally, we may explore ensemble methods to improve model performance.

### **RQ2 - What performance-based clusters can be identified among NHL goaltenders based on their effectiveness against shot variables?**

Much like a pitcher in baseball, the goaltender position in hockey is one of the most difficult to evaluate. A skater's performance can easily be measured through goals, assists, time of possession, number of hits, etc. However, a goaltender's skill is inclusive of metrics which are harder to measure such as mental resilience, reaction time and positioning in the net. This makes it harder for teams to evaluate goalies for long-term contracts, especially when metrics might only be available for regular season play rather than the playoffs. To help address this, we will develop unsupervised learning models to group goalies based on their performance against different shot variables. These clusters will help identify the truly elite goalies based on their performance trends.

## **Data Preparation**

We initially planned to use the same data from RQ1 for RQ2, but we realized this plan was out of scope for this project. The data we used for RQ1 is derived from individual shots' measurements; therefore, clustering goalie performance requires us to analyze each goalie's performance individually and create our own goaltender database to cluster.



We pivoted to Money puck's goaltender specific dataset (<https://moneypuck.com/data.htm> 2023-2024 season). Though their statistics are based on Money puck's own models, this dataset gives us the opportunity to demonstrate a future application of our own models. To prepare our data for clustering, we created 14 calculated columns using rates to normalize cumulative metrics (metric divided by timeOnIce). We also used feature scaling to normalize numeric variables.

## **Modeling Techniques**

### **K-Means Clustering**

K-Means is a centroid based method that will group goaltenders based on their similarity in key metrics. A K-Means model partitions goalies into k clusters, where each goalie is assigned to the nearest cluster center.

### **DBSCAN**

DBSCAN, which stands for density-based spatial clustering of applications with noise, is a clustering method that uses the density of points within a dataset to create clusters. Rather than specifying a number of clusters to create, DBSCAN creates clusters according MinPts - the minimum number of points required to form a region that is considered dense - and epsilon - the maximum distance that can be between two points to be considered neighbors. DBSCAN also reduces the impact of outliers by considering them noise.

### **Hierarchical Clustering**

Hierarchical clustering builds a dendrogram without requiring a certain number of clusters. This model is useful because it analyzes how goalies naturally group together based on shot outcomes.

## **Model Evaluation and Selection**

Since clustering is an unsupervised learning method, traditional evaluation metrics cannot be applied to evaluate the model's success. Evaluation techniques that will be considered are the silhouette score which is used to evaluate K-means clustering by measuring how similar a goaltender is to their assigned cluster compared to other clusters. A higher silhouette score (closer to 1) indicates that goalies are matched well to their cluster and also well-separated from the others. A dendrogram analysis is used in hierarchical clustering. Dendrograms are visually analyzed to determine the optimal number of clusters and their relationships.

### Data Visualizations:

#### Logistic Regression Analysis:

Confusion matrix for our first regression model: Figure 2

```
Confusion Matrix:
[[99636  137]
 [ 7236  423]]
Accuracy:0.931
```

Figure 2

We first ran logistic regression using all the predictor variables in our dataset. Initially, our regression had a 0.93 accuracy rate with 0.755 precision, 0.055 recall. And a 0.93 AUC-ROC score. However, we noticed a confounding variable while analyzing our features. The highest coefficient was `timeUntilNextEvent`. This variable measures the time between the event and the next event, and if a goal is scored, the next event should always be a faceoff, and `timeUntilNextEvent` should always be 0. 30,249 of the 31,078 goals in our dataset (97.3%) have a `timeUntilNextEvent` of 0. This column is essentially another target column.

After this discovery, we retrained the regression without `timeUntilNextEvent` (see appendix O). Without `timeUntilNextEvent`, our logistic regression model predicted every shot would not be a goal. Because most shots do not result in goals, our accuracy is high, but this provides an incomplete evaluation of our model. Additionally, recall, precision and F1 score are irrelevant evaluators because there are no true or false positives, and recall and precision are both 0 due to no predicted samples. The AUC-ROC curve score was 0.716, which is not great performance.

After that, we analyzed feature importance by looking at the coefficients associated with each feature (see appendix P). Shot distance has the highest negative coefficient. This can be explained by the puck being easier to shoot and be accurate from a shorter distance. This reflects our exploratory data analysis.

#### k-Nearest Neighbor Analysis:

We ran a kNN algorithm using the same variables as our logistic regression analysis. We used 1-10 neighbors in our model, limited by time and hardware constraints. Our best model used 10 neighbors and had an accuracy of 0.929 (see appendix Q). While its accuracy is identical to our logistic regression model, kNN was able to predict goals. However, like our logistic regression model, it is incredibly good at predicting no goal but still struggles to find goals. This is reflected in the precision, recall and F1 score, where the weighted accuracy is akin to a coin toss (see appendix R).

#### Random Forest Analysis

Our first few attempts to run the random forest model were hindered by categorical variables. After researching random forests, the literature we read suggested encoding these variables makes the model perform worse, not better. One-hot encoding would create too many binary variables for the model to effectively handle, and label and ordinal encoders would create an order in that data that does not exist.

After removing these categorical variables, we trained our first random forest model (see appendix S). Like our logistic regression and kNN models, our random forest model predicted the shot would not be a goal almost every time. As a result, our accuracy was high at .931, and our recall was abysmal at 0.049. This recall led to our poor F1 score as well (see appendix T). However, although the sample size was small, our random forest model did predict 399 more

goals than our kNN model and was 64.9% of the time when predicting goals. While this is lower than we like, we recognize its performance is better than our previously trained models.

Supporting our basic descriptive statistics, an AUC-ROC score of 0.724 suggests marginally better performance by our random forest model (see appendix U). Additionally, the histogram in appendix V demonstrates how skewed our model is to no goals. This suggests our model could be better served to produce goal probabilities rather than goal or no goal results.

After running our first random forest model, we performed hyperparameter tuning on it and repeated the process. The tuned model produced 38 more goal predictions but only 13 were correct. Our accuracy and precision scores ticked down, and our recall and F1 scores ticked up, but the performance was essentially the same, reflected in our equal AUC/ROC scores (see appendix W).

### Clustering Analysis

Initially we performed a correlation matrix of the variables in the goalie dataset to see which variables might be closely correlated, which may also indicate the features that will similarly affect the clusters that the goalies in our dataset are placed into. There are notable high positive correlations between goals saved above expected per 60 minutes and both save percentage and both low- and medium-danger goals saved above expected per 60 minutes. There are notably negative correlations between number of pucks frozen per save and GSAX/60, save percentage, and medium-danger GSAX/60 (see appendix X).

The first cluster analysis we performed on our dataset was k-means clustering. We first created a scree plot (appendix Y) to identify the number of principal components to use in our principal component analysis (PCA), which will help us identify the ideal number of clusters for

the k-means clustering. This is shown in appendix Z, where the ideal number is identified by the elbow method. Appendix AA is the resulting k-means cluster analysis.

The second type of cluster analysis we considered was DBSCAN. We initially used a k-distance graph (see appendix BB) to determine the ideal epsilon value for DBSCAN clustering. This was found to be around 375.5. We used this epsilon value in our DBSCAN clustering of NHL goalie (see appendix CC).

The third type of cluster analysis we performed to consider using was hierarchical clustering. We began this process by using a dendrogram (see appendix DD) to decide how many clusters we would create and how those clusters would be determined. Since the dendrogram showed 3 distinct clusters, this is the number we used for our agglomerative hierarchical clustering analysis (see appendix EE). We compared the silhouette scores of each type of clustering analysis to determine which type we will use to cluster our data.

### Data Analysis

#### Model Description

Each CSV file was input separately and combined to create one CSV file that included every shot and its qualities from the 2021-2022 season through February 3rd of the 2024-2025 season. We then investigated the dataset's variables and removed all variables based on Moneypuck's predictive models, as well as variables outside of our project's score and redundant variables. We wanted to limit our scope to the factors that we thought would have the most impact due to our time constraints.

We trained three classification models on this dataset. To ensure the accuracy of our classification models, we split the data into training and test datasets and used cross-validation to prevent overfitting. Our logistic regression and kNN models trained on 75% and tested on 25%

of our dataset, while our random forest and its tuned model trained on 80% and tested on 20% of our dataset. The training and testing datasets were randomly assigned from the original dataset for all three models using the independent variables described in appendix FF.

Using these independent variables, we ran our logistic regression and kNN models and started to run our random forest model before removing the categorical variables. The literature we read suggested one-hot encoding these categorical variables makes the random forest perform worse. It creates too many binary variables for the model to effectively handle, particularly lastEventCategory and shotType. Label and ordinal encoding these variables creates an order that does not exist.

We initially planned to cluster goaltenders using the same dataset as goal prediction models, but we realized this was out of the scope of our project. We were only able to cluster individual shots because the dataset focused on shots taken and their qualities. To cluster goalies using this dataset, we would have had to run our goal prediction models on each goalie's shots individually, then cluster the results. Instead of using the shot dataset, we used goalie specific performance data from another of Money puck's datasets. Though this data is based on Money puck's own models, clustering their models' performance demonstrates a future application of our own models.

Like the shot data, we first investigated this dataset, determined which variables were unnecessary, then added our own calculated columns to normalize the cumulative stats by rate. We removed columns that would not be relevant to our analysis or were redundant, such as the player ID number (since we have their names), team, position, and penalties taken. We also decided to include all gameplay situations, rather than delineating powerplay and short-handed situations, so these rows were removed from the data. We expanded on the variables provided to

create rates for each statistic in order to attempt to normalize the comparisons of the goalies. We began with converting icetime, which was in seconds, to icetime in minutes, and then expanding on the variables provided to make them rates per 60 minutes of gameplay. We also utilized the goals and xGoals statistics provided to expand the features to include goals saved above expected (GSAx) and GSAx/60 (per 60 minutes of icetime). We did the same with the low, medium, and high danger chance statistics. All of this initial analysis was performed on the CSV file obtained from MoneyPuck, and was performed in Microsoft Excel for convenience.

After loading and previewing the data, we began comparing our three cluster analysis methods. We began considering a k-means cluster analysis by doing a principal component analysis (PCA). Because many of the variables in our data can be derived by the same data or show similar events, a PCA would help us to narrow the features we would include in our clustering to reduce dimensionality and hopefully create clusters that create clearer patterns. We graphed the explained variance ratio of the PCA to look for the “elbow” in the graph, which would indicate to us how many components would be best to use in the data. The graph showed that two components is best. Then using a PCA with two components, we similarly graphed the “inertia” of the PCA, which is the sum of the squared distances data points to their respective clusters’ centers. Again using the elbow method, we determined from the graph that 3 clusters would be ideal for the k-means clustering analysis.

The second type of clustering analysis we looked at was a DBSCAN. In order to optimize our DBSCAN clustering, we first used a k-distance graph, which will show the sorted distances to the kth nearest neighbor for each data point in a dataset. Similarly with previous graphs used to optimize our models, the elbow method was used to determine the optimal epsilon value for our DBSCAN model. The resulting epsilon value is 375.5. This seems to be a good value due to



coinciding with the elbow of the k-distance graph. It was also verified with a loop searching for the max silhouette score at the range of values around the elbow in the graph, as well as a resulting silhouette score that is within a suitable range as our k-means clustering analysis silhouette score. We used this epsilon value in our DBSCAN clustering of NHL goalies.

The third type of cluster analysis we considered was hierarchical clustering. We began this process by using a dendrogram to decide how many clusters we would create and how those clusters would be determined. We created dendrograms using both the Ward and Average methods. Since both dendrograms showed 3 distinct clusters, this is the number we used for our agglomerative hierarchical clustering analysis. Because the Ward method showed the clusters being more compact and closely related, this was the method we moved forward with for the hierarchical clustering.

#### Model Results:

RQ1: What variables are most predictive for a shot being a goal during the regular season?

The following table represents the results from each model.

<b><u>Algorithm Name</u></b>	<b><u>ROC</u></b>	<b><u>Precision</u></b>	<b><u>Recall</u></b>	<b><u>F1 Score</u></b>	<b><u>Accuracy</u></b>
Logistic Regression	0.716	0	0	0	0.929
kNN	0.609	0.365	0.003	0.006	0.929
Random Forest	0.72	0.626	0.051	0.094	0.931

Our random forest model was far and away the highest performing model. However, this is mainly because our logistic regression and kNN models produced little to no predicted goals

and is not because our Random Forest model is a much better predictor of goals. Our models successfully find true negatives but struggle greatly to identify true positives.

Our random forest model may have selected predicted more goals because its classification method was better suited toward our analysis, but it also may have selected more goals because we removed the categorical variables from its predictor variables. This inconsistency makes true comparison between our models more challenging.

When looking at our random forest model, the area under the curve of the Receiving Operating Characteristic was 0.72. This means that our model has a 72% higher chance of determining whether or not a goal will be scored correctly. This result is disappointing. In a sport where the difference between a goaltender saving 90% or 91% of their shots against is massive, only correctly predicting a shot's result 72% of the time is disappointing. Although our accuracy is at 93%, we recognize this accuracy is over inflated due to lack of goal predictions.

RQ2: What performance-based clusters can be identified among NHL goaltenders based on their effectiveness against different shot variables?

We used the silhouette scores of each model to compare which was best in our cluster analysis of NHL goalies' performances. The comparison of the three methods is shown below:

<b><u>Method</u></b>	<b><u>Silhouette Score</u></b>
K-means clustering	0.645
DBSCAN	0.554
Hierarchical clustering	0.644

Though two of our models had very similar silhouette scores, k-means clustering was highest.

## Results

RQ1: What variables are most predictive for a shot being a goal during the regular season?

When looking at which variables are most predictive for a goal being scored, three variables stand out. Shot distance, shot angle and shooting team average time on ice. With shot distance, the closer a shot is to the net, the more likely it is that it will go in. This is also reflected in our EDA, where we found the same thing. With shot angle, shots taken from a sharper angle toward the net are likely more difficult to score because the goalie can more effectively block the puck's line of sight. As the shooting angle becomes more extreme, the available net space decreases, making it harder to find an opening.

The shooting team's average time on ice is less clear. Our logistic regression model suggests a positive correlation between the shooting team's average time on ice and a goal being scored, but there are certainly limits to this. A player that is hemmed in the defensive zone for minutes at a time will be exhausted and unlikely to score, but they will also be unlikely to shoot the puck in the first place. Furthermore, a team's best players typically stay out for the last few minutes of the game when their team is trailing. These players pull their goalie and have an advantage of 6 offensive players against 5 defensive players.

Furthermore, while player position and shot type may not have a direct correlation with goals being scored, a goal's qualities are likely different depending on player position and shot type. We identified shot distance as the most important variable for a goal's success when comparing all shots against each other, but this likely caused slapshots taken by defensepersons to be underrepresented in the goal predictions. Slapshots require more time and space to take and are ineffective once a shooter gets too close to the net. Wrist shots account for 53.7% of this

dataset, and the models' predictions are likely skewed by wrist shots' ideal qualities. Segmenting further based on shot type and player position provides a great opportunity to improve our models.

## Results

RQ2: What performance-based clusters can be identified among NHL goaltenders based on their effectiveness against different shot variables?

After applying this method to our data, we found that the model tended to cluster goalies according to the amount of icetime they had, despite attempting to normalize the data for rates. (The final k-means clustering with cluster centroid is shown in appendix GG.) This was due to still including the icetime feature, as well as static features such as number of shots faced or number of pucks frozen in our model. The goalies, along with their associated clusters and features that played the biggest role in the analysis, is shown in the goalie\_clusters.csv file. The feature importances (see appendix HH) show that the icetime was the biggest contributor to how goalies were clustered, as well as the static variables for total numbers of shot attempts faced, blocked and unblocked. Cluster 2 (see appendix KK) showed goalies who were clear starters for their teams, having the bulk of the icetime and shot attempts against them. Cluster 0 (see appendix II) showed goalies who were clear backups for their team and played consistently throughout the season, but in a much lesser capacity than the starters. Cluster 1 (see appendix JJ) showed goalies who played the fewest minutes of icetime during the season. These goalies likely only played in cases of injury or fatigue of both of the starting and backup goalies and did not see consistent playing time during the season. Future analysis done in this way, or looking to cluster goalies by performance, would likely show more by removing all static variables (like

icetime and shot attempts faced) and only include features that are normalized for the rate of the statistic per unit of icetime (commonly 60 minutes).

### Ethical Recommendations:

The use of and reliance on predictive modeling in NHL shot data creates ethical implications for owners, coaches, players, staff, and even fans. Predictive models can be a game-changer for a team's front office, helping them make smarter decisions—from determining which players deserve long-term contracts to deciding who should skate on the top line each night. However, using predictive models also raises concerns about fairness, transparency, and unintended consequences for players and their families. Beyond the rink, predictive models also affect fan engagement and activities such as sports betting and fantasy leagues. At its core, the NHL is both a sport and a business, and while team owners may be passionate about hockey, their ultimate goal is profitability and to “bring home the cup”. Advanced analytics has become an invaluable tool to help them achieve those goals, but it's imperative the use of analytics remains ethical.

In order to provide ethical recommendations around the use of advanced analytics in the NHL, utilitarianism provides a useful framework. This ethical theory, proposed by English philosophers Jeremy Bentham and John Stuart Mill in the early 1800s, is based on the idea that an action is morally right if it produces the greatest benefit and wrong if it causes the most harm. Known as the Greatest Happiness Principle, utilitarianism prioritizes outcomes that benefit the largest number of people (Quinn, 2020). When considering the use of predictive models in the NHL, this means taking into account the impact on not just teams and owners, but also the players, staff, and fans.

From a utilitarian perspective, predictive analytics in hockey should be used in a way that benefits the most people while minimizing harm. When used correctly, these models can create positive effects for teams, players, and fans. However, if misused, predictive models could lead

to unfair treatment of players when it comes to their job security and wages. The use of fair and transparent predictive models during contract negotiations can help ensure a player is accurately compensated based on objective data rather than bias or favoritism. Utilitarianism also allows for consideration of the impact of cutting a veteran player from the team based on a model's predictions, weighing the effect on both the player's livelihood as well as team culture.

However, players may suffer unintended consequences if models are used without context. For example, a player with declining statistics could be cut from a team if the team solely focused on data rather than any other temporary factors such as an injury or other external situations. To align with utilitarian ethics, advanced analytics should be supplemented by human-centered decision making to ensure fair treatment of players and maximize the benefit to the most people.

In addition to teams and players, advanced analytics also impacts the business side of hockey, such as with sports betting and fantasy leagues. Predictive models can provide fans with an increased understanding of the game, making fantasy hockey and betting more inviting and fun. However, these same models can create ethical problems, especially if they lead to an unfair advantage for people who have access to exclusive information. From a utilitarian perspective, this scenario would not be creating the greatest benefit for the most people, instead only benefiting a few while causing harm to the most. Also, if predictive models heavily influence betting, it could potentially lead to game fixing concerns. To help mitigate these worries, the NHL, sportsbooks, and teams should be transparent about how their models were developed and used.

To ensure that predictive analytics are used in a way that maximizes benefit and minimizes harm, several recommendations should be considered. Teams should be transparent and openly communicate how models are being used to inform roster moves so that trust is

preserved with players, staff, and fans. During player evaluations, analytics should be supplemented by qualitative considerations so players are not unfairly impacted solely by a model's predictions. The NHL should regulate how predictive models influence sports betting in order to ensure fairness. Lastly, teams should use advanced analytics in support of enhanced strategy without making decisions that might impact the integrity of the sport so many love. While predictive models provide valuable insights, they should be a complement to human judgment rather than a replacement of them.



### Challenges:

The first challenge we faced was with combining the datasets together. When doing so, there was a corrupted row that didn't show up visually, but blocked us from running any exploratory analysis. When merging csv files, we used command prompt and we believe that the corrupted row came when combining the files. The second challenge we faced was that in our analysis, all of our regression models would mainly predict no goal, with the logistic regression never predicting a goal at all. This is due to the fact that most of our data are shots and misses. This challenge may be mitigated by applying class weight to a future model.

Communication and organization was the biggest challenge for this project. We had team members spread across 3 different time zones, working vastly different professional and academic schedules. Some of us had more classes than others. Some of us completed work for this project early in the week, and others completed work later in the week. We even had two team members with surgeries the week of the data visualization submission. This caused more segmentation of the work than we would have preferred. We used different sets of variables and training/testing proportions in our classification models, and all of these communication obstacles are part of why.

We feel pleased with how we navigated the challenges of this project. We planned well and had our work comfortably completed without any fires to put out along the way. We successfully navigated three check-ins, eight submissions, and one final project through innumerable discord chats, group calls, and one-on-one working sessions when we realized the existing dataset did not answer our clustering research question. We pushed each other to be better and helped each other grow.

We were disappointed in our models' results, but we recognize these are the first steps to building a fully fleshed out model like MoneyPuck's. We started the project with a massive scope and with each passing week realized we did not have the time to achieve our grand plans. However, our final result leaves us with a great first step to creating original, meaningful models that innovate how the game we love is played.

We faced additional challenges in our goalie clustering analysis. We initially planned to utilize the same dataset for both our shot prediction model, as well as our clustering analysis, but realized that using the shot data for a clustering analysis would cluster the shots themselves, whereas we were interested in clustering the goalies facing shots with differing attributes. In order to continue our clustering analysis, we were forced to find a new dataset based on the goalies themselves, which we quickly found through the same source with MoneyPuck. While this second dataset was able to be used for our analysis, it also came with its own challenge. Much of the statistics in the goalie dataset were raw data, which needed to be normalized for the rate of icetime each goalie played to account for the differing roles each goalie plays on their team. It also included breakdowns of game situations and other features we were not interested in, such as penalties taken by goalies or their statistics while their team was on the powerplay. Much of the challenges faced with the second dataset required for this project were handled through data cleaning and preprocessing of the CSV file obtained from MoneyPuck in Microsoft Excel for convenience.

### Recommendations and Next Steps:

Our project identified a few key variables critical to predicting a shot's success such as shot distance, shot angle, and the scoring team's time on the ice. However, there are many additional areas for analysis. Since the dataset was heavily skewed towards no goal, future models should explore weighted variables, oversampling goals, or undersampling missed shots. Another viable option is to engineer new composite features such as a score that relates to a high danger shot or a threat score. This composite variable could be a combination of multiple shot variables and could provide model enhancements. Other options would be to use different modeling techniques such as neural networks to gain deeper understanding of complex relationships among the variables. The data is also present to try segmentation - segmentation could be attempted with either game situations (powerplay or playoff game) or even player position of the shooter.

Our goaltender clustering analysis mainly categorized goalies based on their ice time and workload. While this is valuable information, it did not provide the deeper insights we were hoping for. Future research would include refining the features. Instead of using raw data such as saves and goals against, incorporating derived metrics such as expected goals, rebound control, or expected save percentage, could provide more meaningful information. Also, using a data set from multiple years would provide the ability to assess a goaltender's performance over multiple seasons. For the purposes of this project, we were only able to utilize a couple clustering methods, but other research could use additional methods such as Spectral Clustering (unsupervised learning using graph theory) or Gaussian Mixture Modeling (soft clustering by assigning probabilities to different data points in clusters).

Whatever direction future research takes, it's imperative that any models are used with ethical consideration and transparency. Advanced analytics are a part of NHL hockey today and are involved in team roster decisions as well as fan engagement with sports betting. Predictive models should always complement and never replace human expertise and decision making.

## References

Azerrad, S. (2023, November 21). How Most NHL Goals are Scored (Shoot More Like This) -

Gaimday. Gaimday. <https://www.gaimday.com/blog/how-are-most-nhl-goals-scored/>

Clustered Scatter Plot. (n.d.). Retrieved from

[https://blogger.googleusercontent.com/img/b/R29vZ2xl/AVvXsEgKvvYqTMfovfBGRtHJZvtX7iJdGM26MI88gnlIHq2wQ1kuDyMmRlZ\\_Y\\_IMSRlI8QWN5wQBeNUVbSXO7KXkY-W9p1k9yTQmuFVftz9iPOE2p1GNcfCiYJK6uvxyposKow66\\_TmK3K2yeeyk/s1600/agglomerative\\_clustering.png](https://blogger.googleusercontent.com/img/b/R29vZ2xl/AVvXsEgKvvYqTMfovfBGRtHJZvtX7iJdGM26MI88gnlIHq2wQ1kuDyMmRlZ_Y_IMSRlI8QWN5wQBeNUVbSXO7KXkY-W9p1k9yTQmuFVftz9iPOE2p1GNcfCiYJK6uvxyposKow66_TmK3K2yeeyk/s1600/agglomerative_clustering.png)

Correlation Heatmap. (n.d.). Retrieved from

[https://miro.medium.com/v2/resize:fit:1400/1\\*8ca2B3abftiPSv9MZE\\_9Cw.png](https://miro.medium.com/v2/resize:fit:1400/1*8ca2B3abftiPSv9MZE_9Cw.png)

Feature Importance Bar Chart. (n.d.). Retrieved from

[https://www.scikit-yb.org/en/latest/\\_images/importances-2.png](https://www.scikit-yb.org/en/latest/_images/importances-2.png)

Garmat, D. (2018, June 12). Clustering NHL goalies. Clustering NHL Goalies – Dan Garmat’s

Analytics Blog – Statistician - Data Scientist from Oregon.

<https://dgarmat.github.io/NHL-Goalies/>

Johansson, Ulf & Wilderoth, Erik & Sattari, Arsalan. (2022). How Analytics is Changing Ice

Hockey. Linköping Hockey Analytics Conference. 49-59. 10.3384/ecp191006.

Johnson, D. (2024a, April 5). Down the xG rabbit hole. Hockey Analysis.

<https://hockeyanalysis.com/2024/04/05/down-the-xg-rabbit-hole/>

Johnson, D. (2024b, April 8). Quick comparison of four public expected goal models. Hockey Analysis.

<https://hockeyanalysis.com/2024/04/08/quick-comparison-of-four-public-expected-goal-models/>

Khan, R., Schena, P., Park, K., & Pinsky, E. (2022). A clustering approach to analyzing NHL goaltenders' performance. In Proceedings of the International Conference on Data Science and Advanced Analytics (pp. 1-10).

[https://doi.org/10.1007/978-3-031-17292-2\\_1](https://doi.org/10.1007/978-3-031-17292-2_1)

Lukan, A. (2022, September 6). Beyond the box score - An intro to hockey analytics. NHL.com.

<https://www.nhl.com/kraken/news/beyond-box-score-intro-to-hockey-analytics-335471754>

Nandakumar, N., & Jensen, S. T. (2019). Historical perspectives and current directions in hockey analytics. *Annual Review of Statistics and Its Application*, 6(1), 19–36.

<https://doi.org/10.1146/annurev-statistics-030718-105202>

NHL Corporate Data. Retrieved from: <https://www.nhl.com/>

NHL League Averages | Hockey-Reference.com. (n.d.). Hockey-Reference.com.

<https://www.hockey-reference.com/leagues/stats.html>

NHL Salary Cap Data. Retrieved from: <https://puckpedia.com/teams>

NHL Shot and Goalie Data. Retrieved from: <https://moneypuck.com/data.htm>

Piccolo, N. (2022, July 3). Inside the stat: Expected goals. Inside The Rink.

<https://insidetherink.com/inside-the-stats-expected-goals/>

Quinn, M. J. (2020). *Ethics for the information age* (8th ed.). Pearson.

Whitmore, J. (n.d.). Introducing expected goals on target (xGOT). Stats Perform. Retrieved February 1, 2025, from

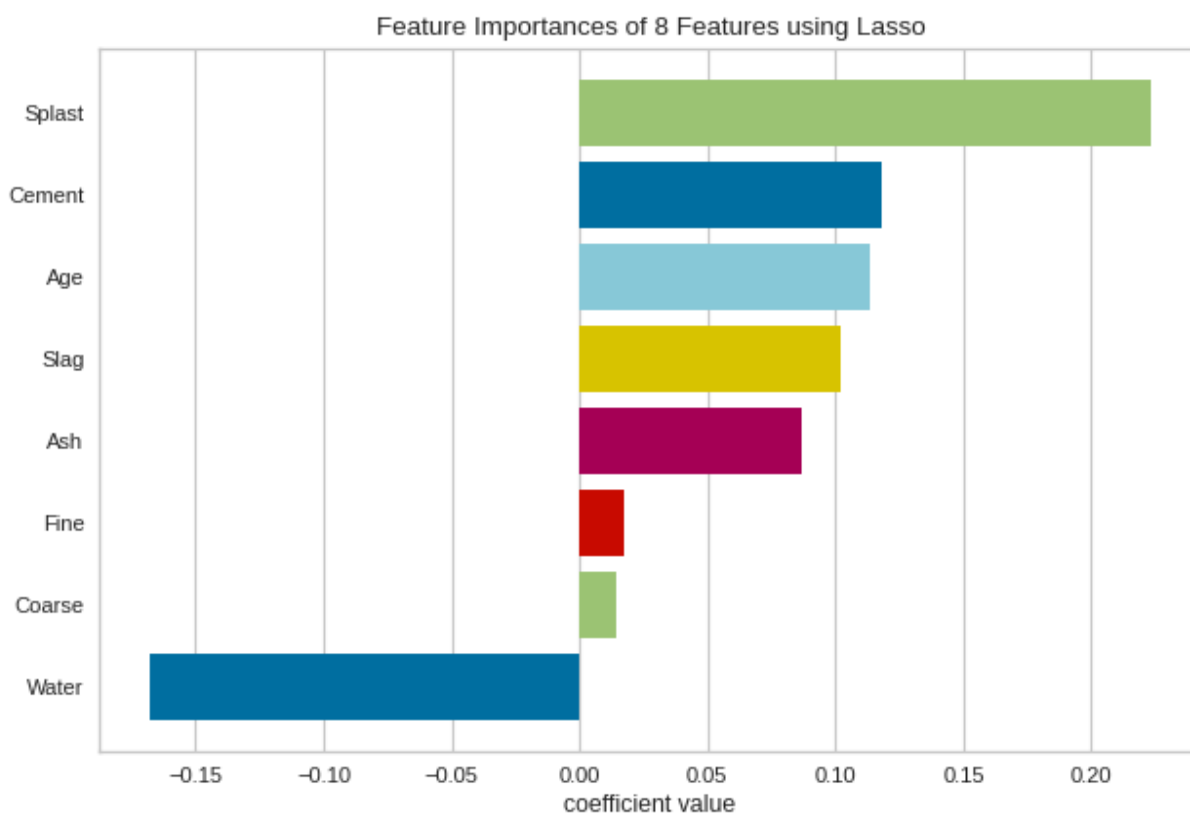
<https://www.statsperform.com/resource/introducing-expected-goals-on-target-xgot/>

Younggren, J., & Younggren, L. (2021, April 25). A new expected goals model for predicting goals in the NHL. Evolving Hockey.

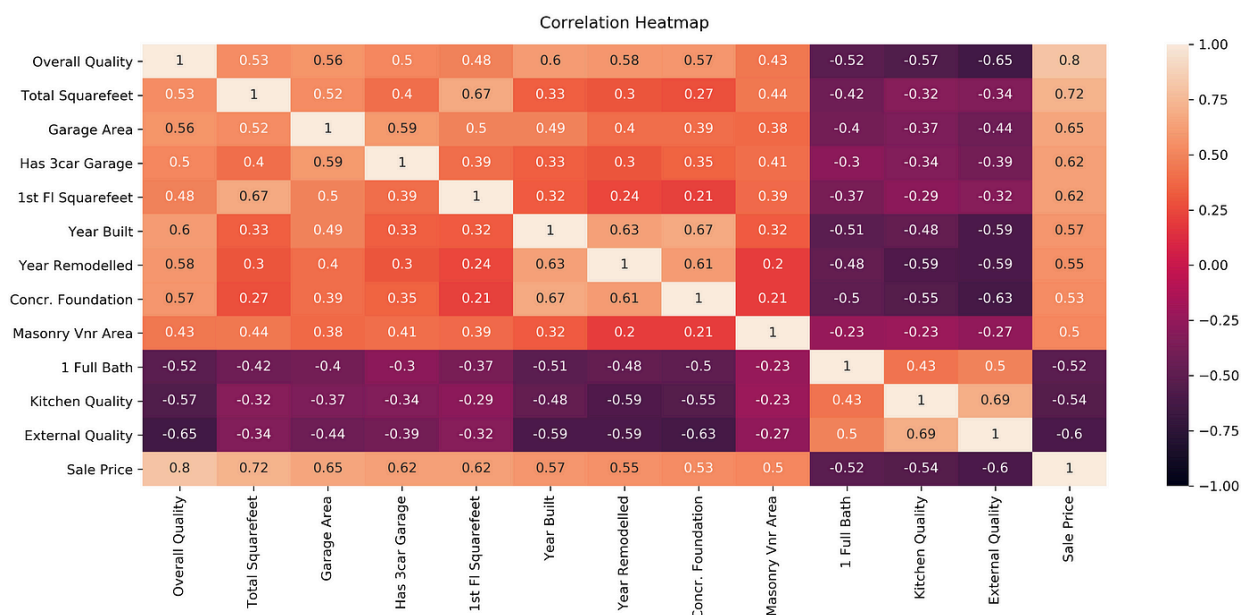
<https://evolving-hockey.com/blog/a-new-expected-goals-model-for-predicting-goals-in-the-nhl/>

## Appendix

### A: Feature Importance Bar Chart Example

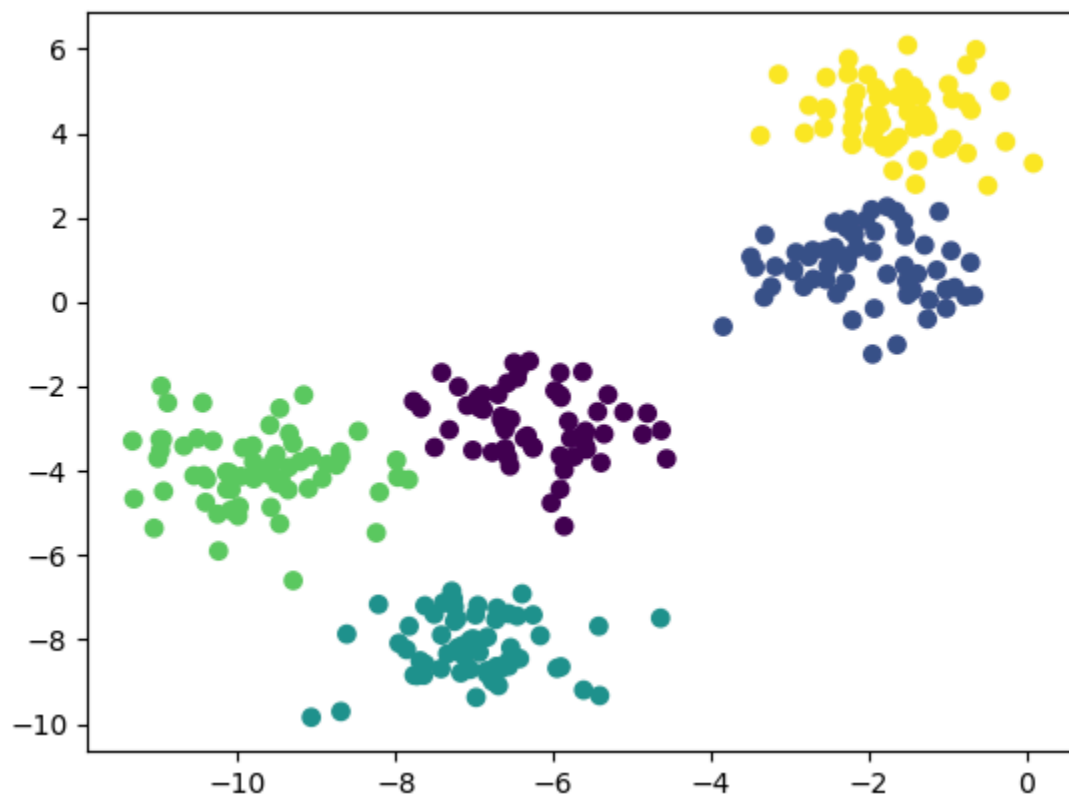


### B: Correlation Heatmap Example

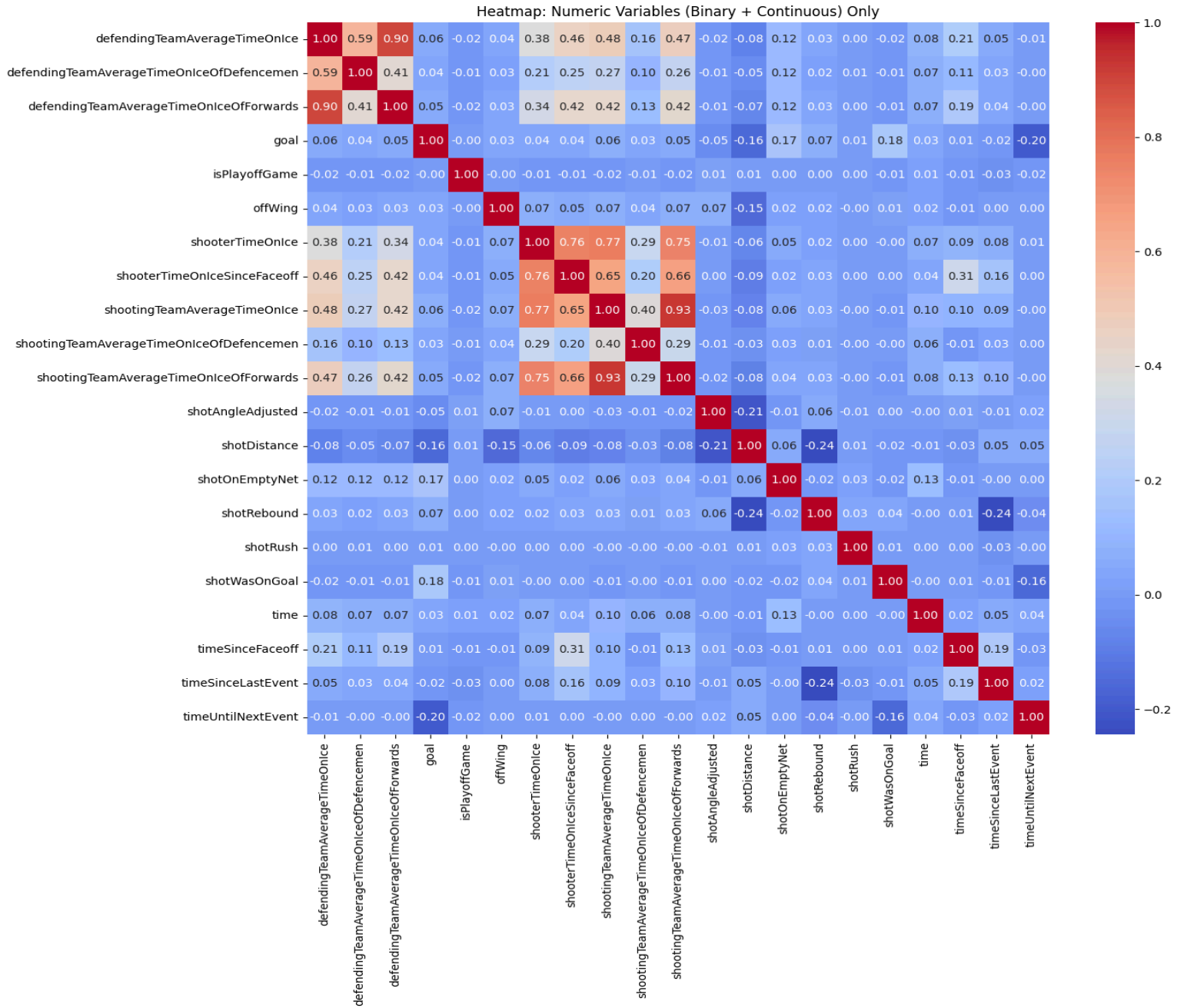




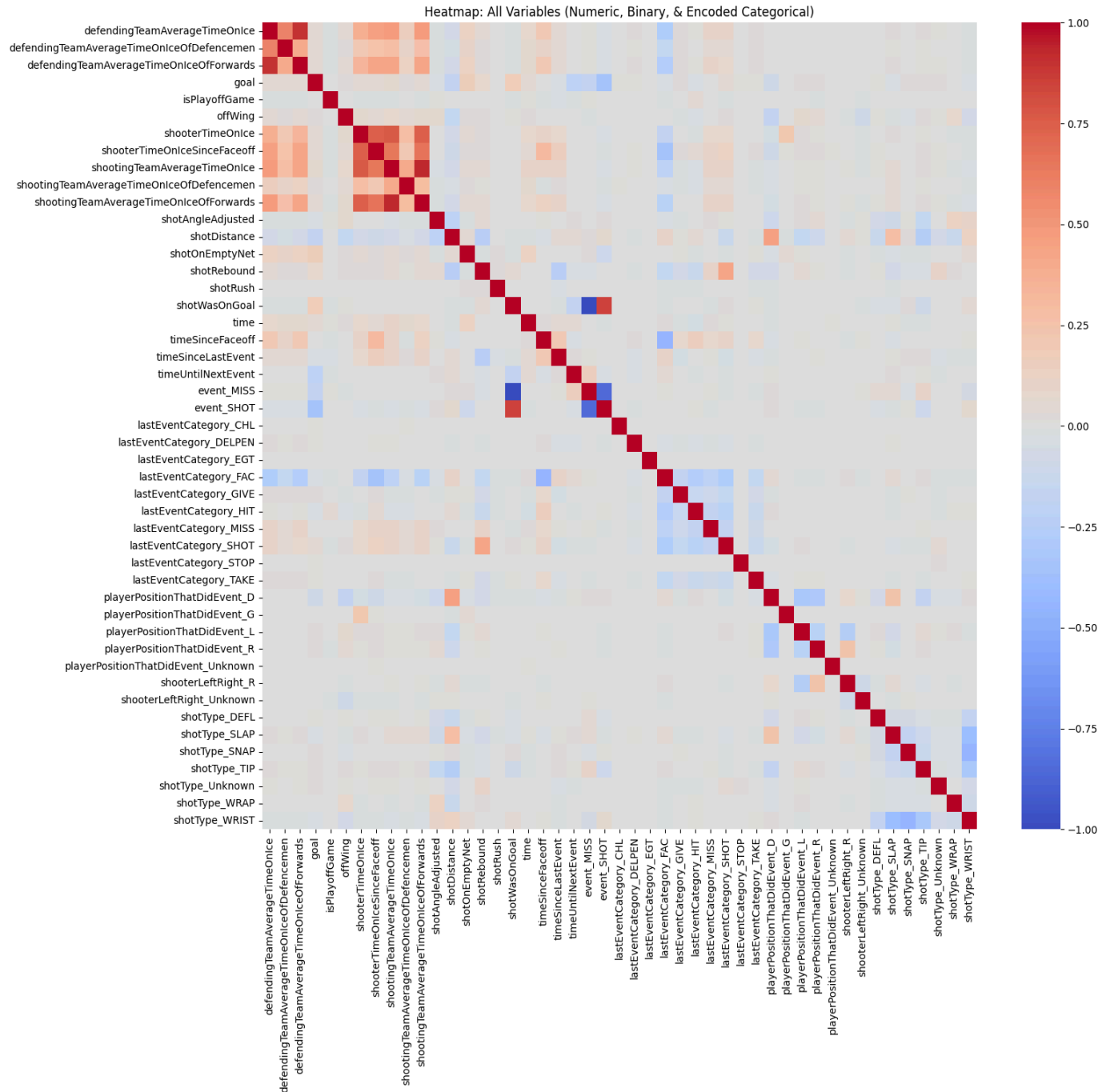
C: Scatterplot example



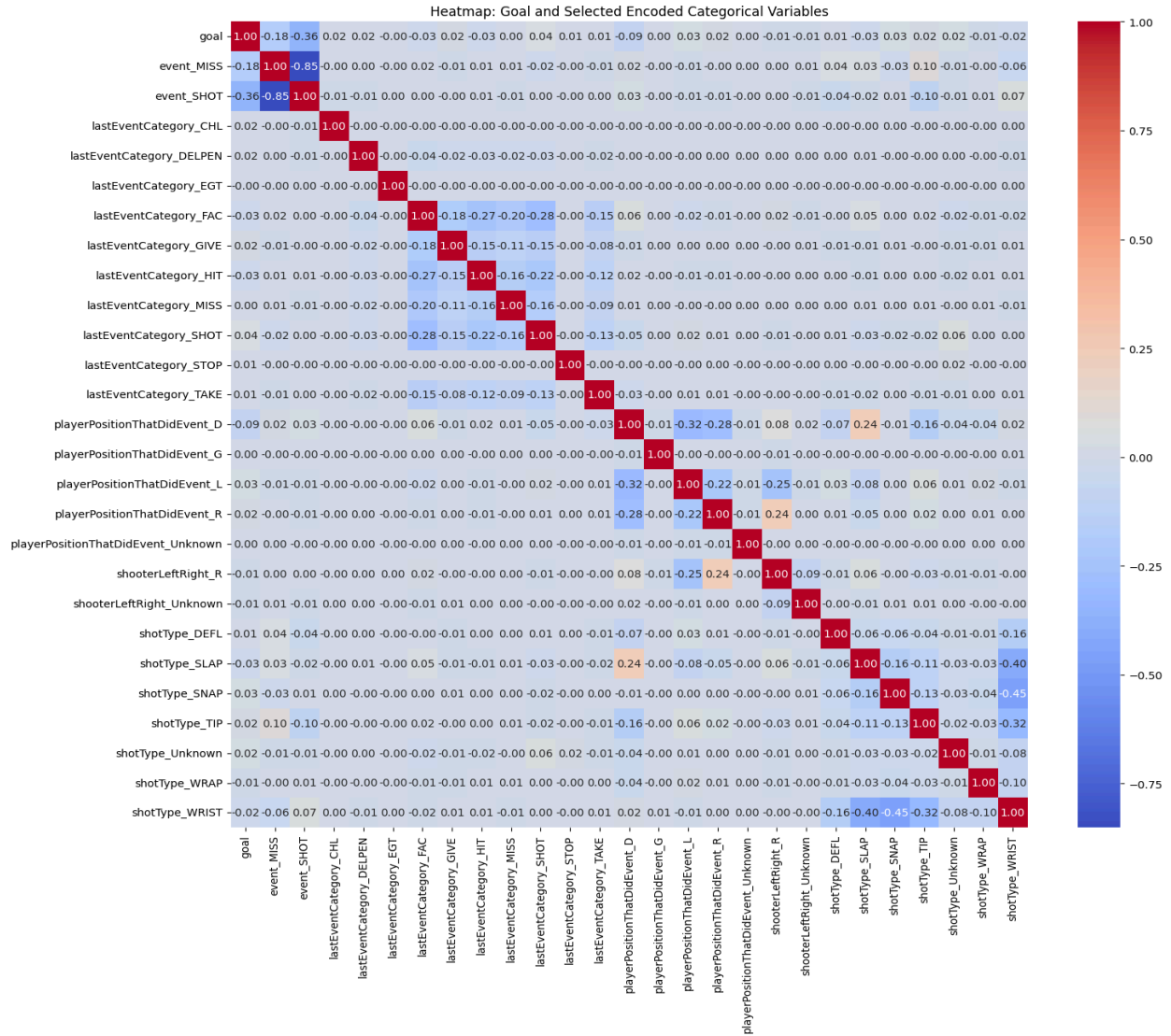
## D: Numeric and Binary Variables Heatmap



## E: All Variables with One Hot Encoded Categorical Variables Heatmap

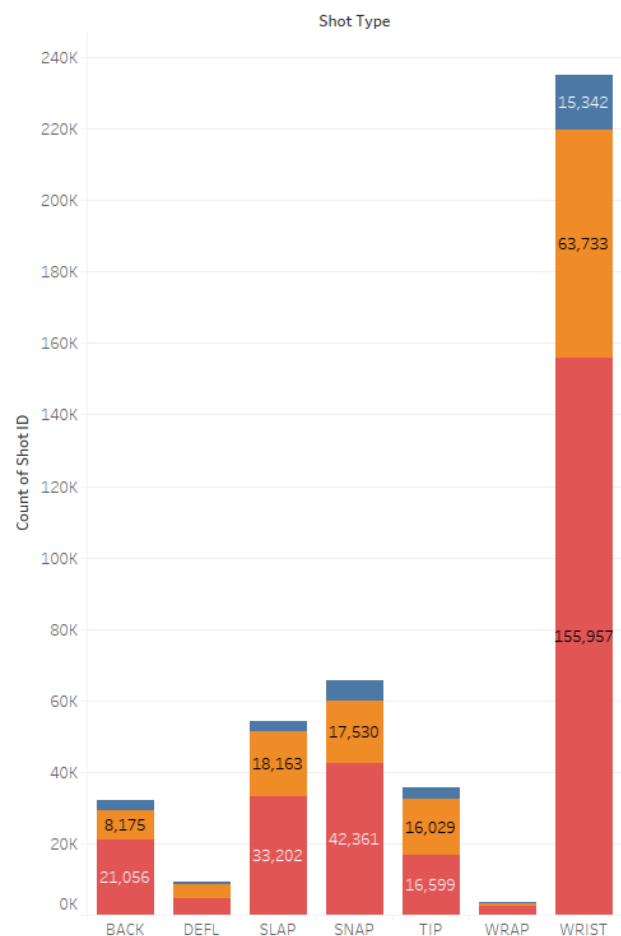


## F: Goals and One Hot Encoded Variables Heatmap

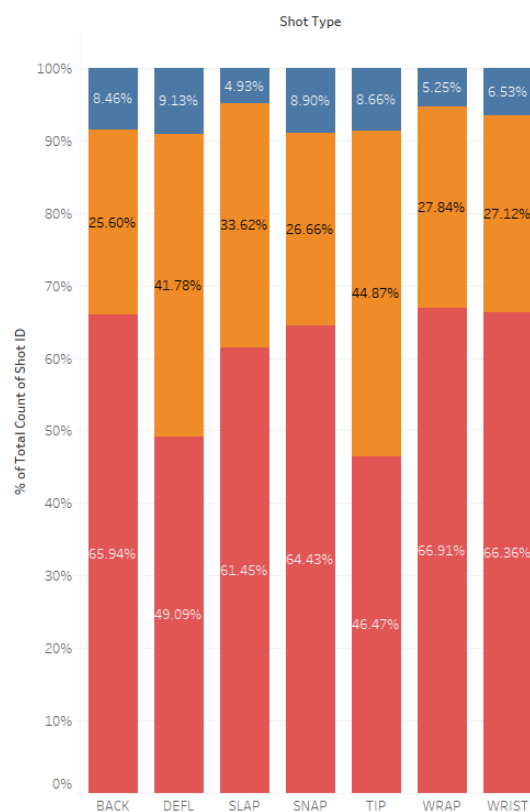


## G and H: Shots vs results

Count of Each Shot Compared to Event



Percentage of Each Shot Compared to Event

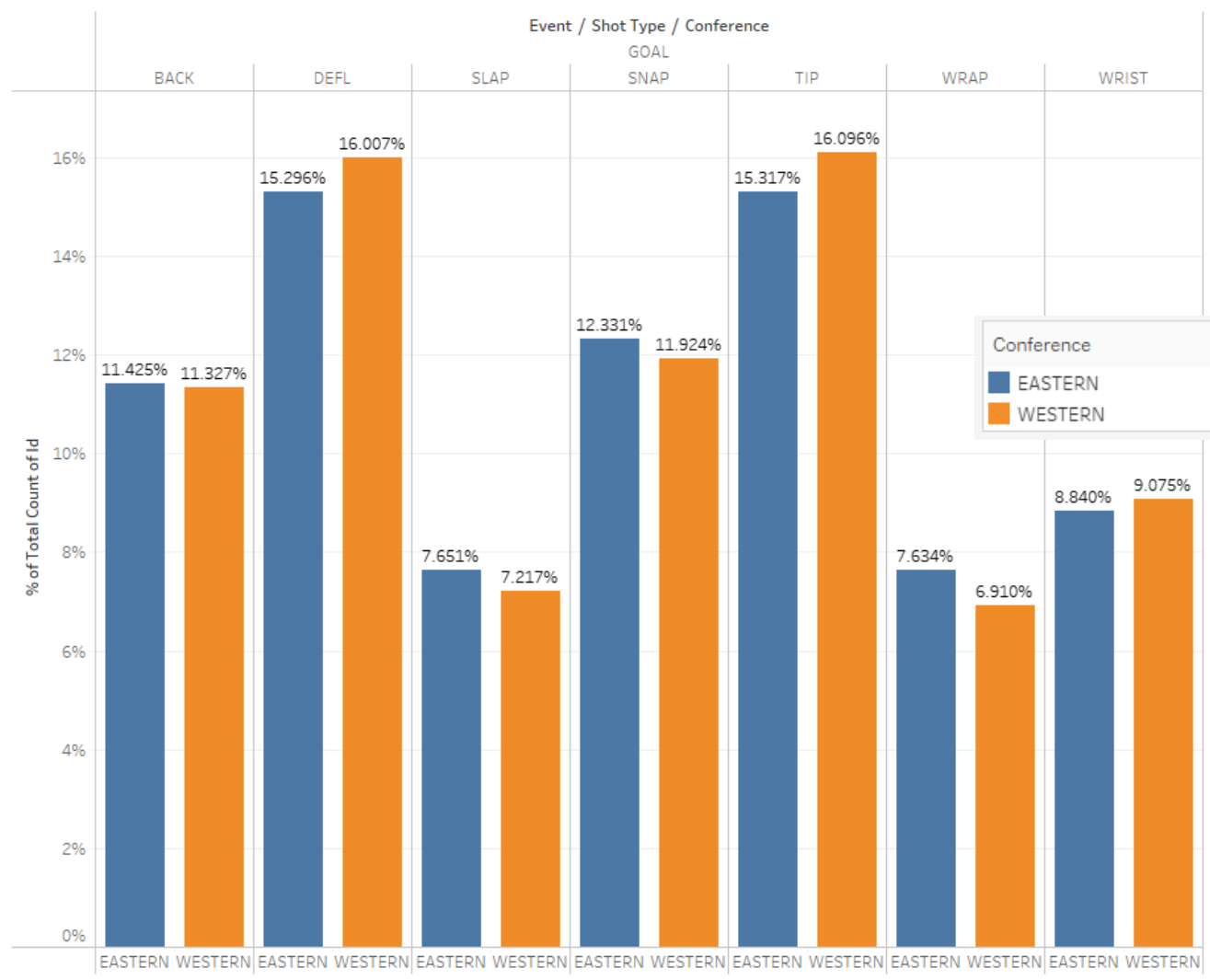


Event

- GOAL
- MISS
- SHOT

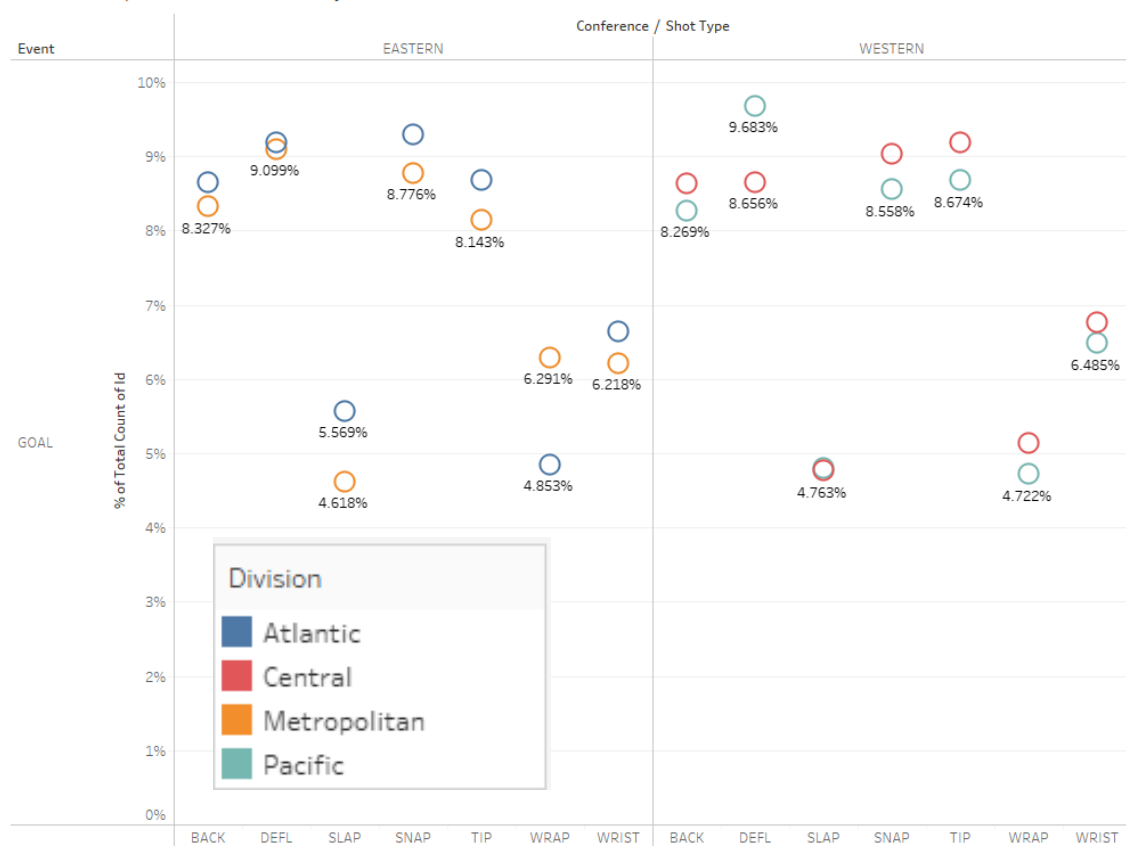
## I: Percentage of Goals by Conference

Percentage of Goal Compared to Type of Shot by Region

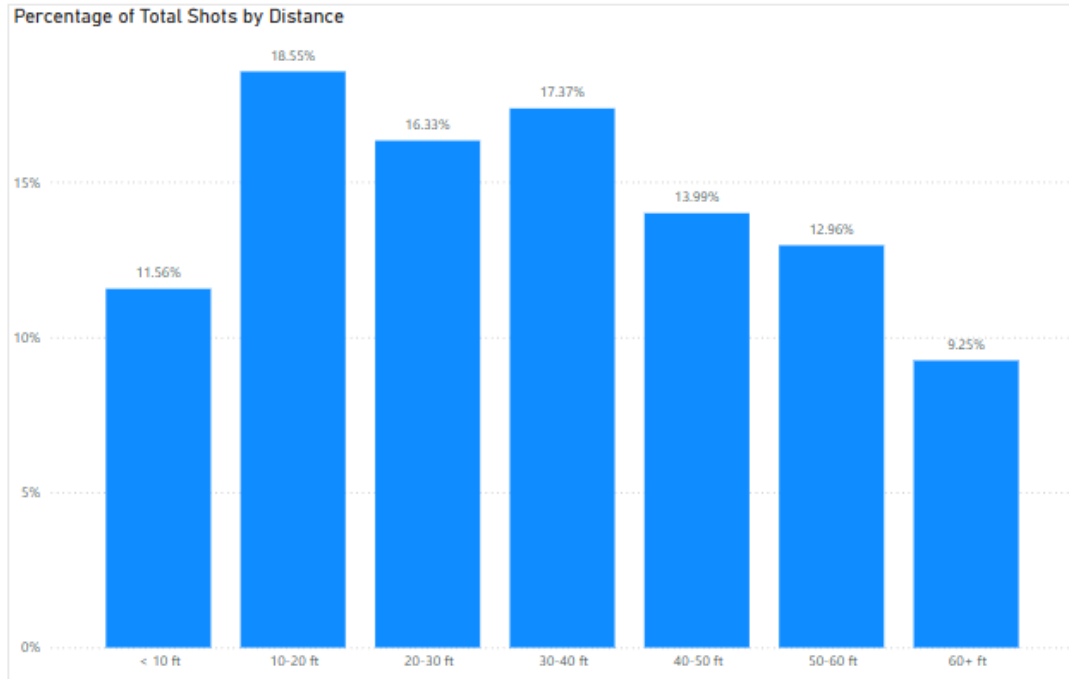


## J: Shots by Division

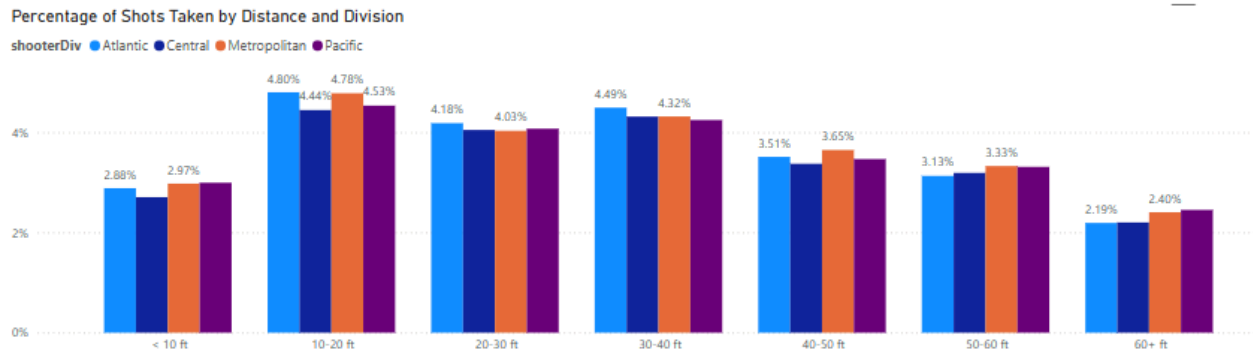
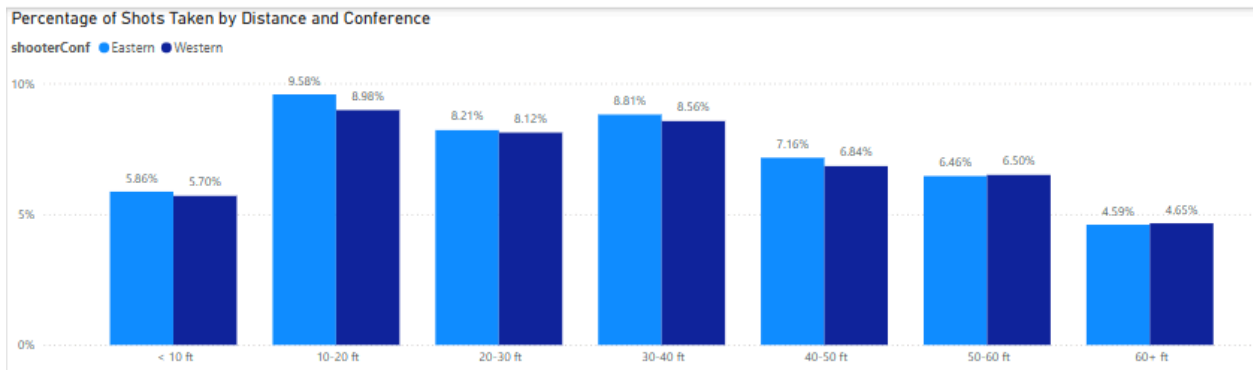
Goals Compared to All Shots by Conference



## K: Percentage of Shots by Distance

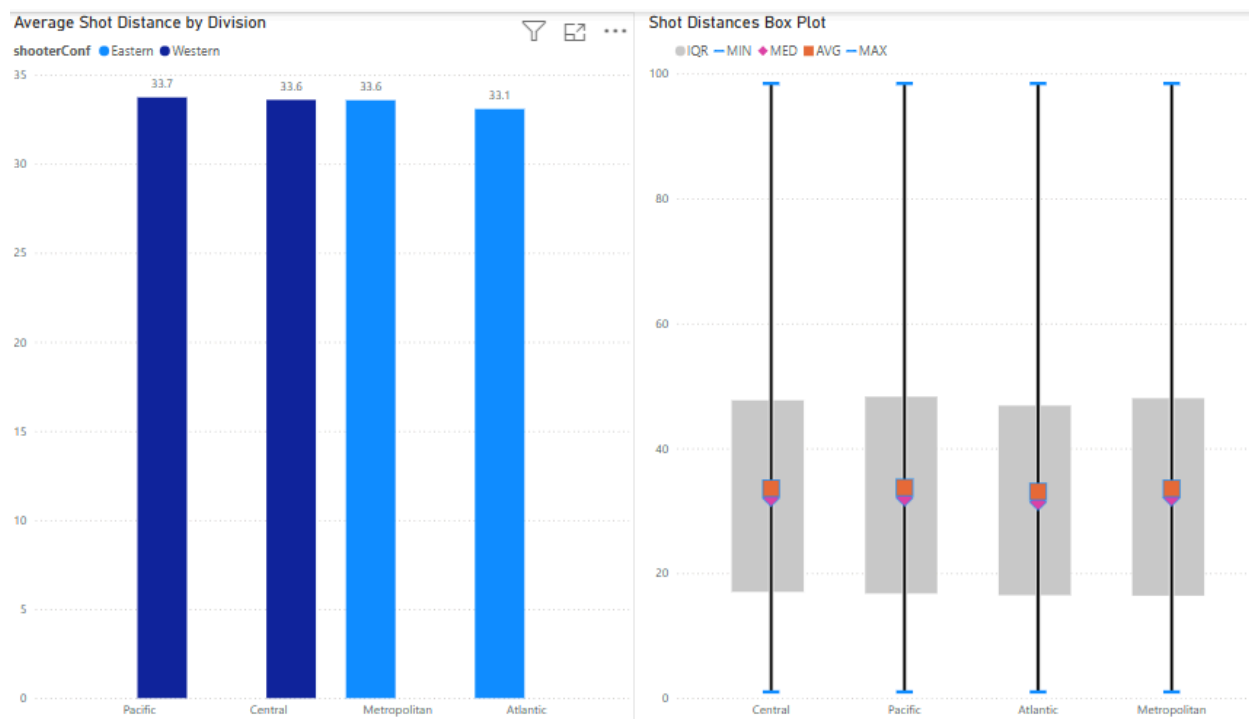


## L: Percentage of Shots by Distance and Conference and Division



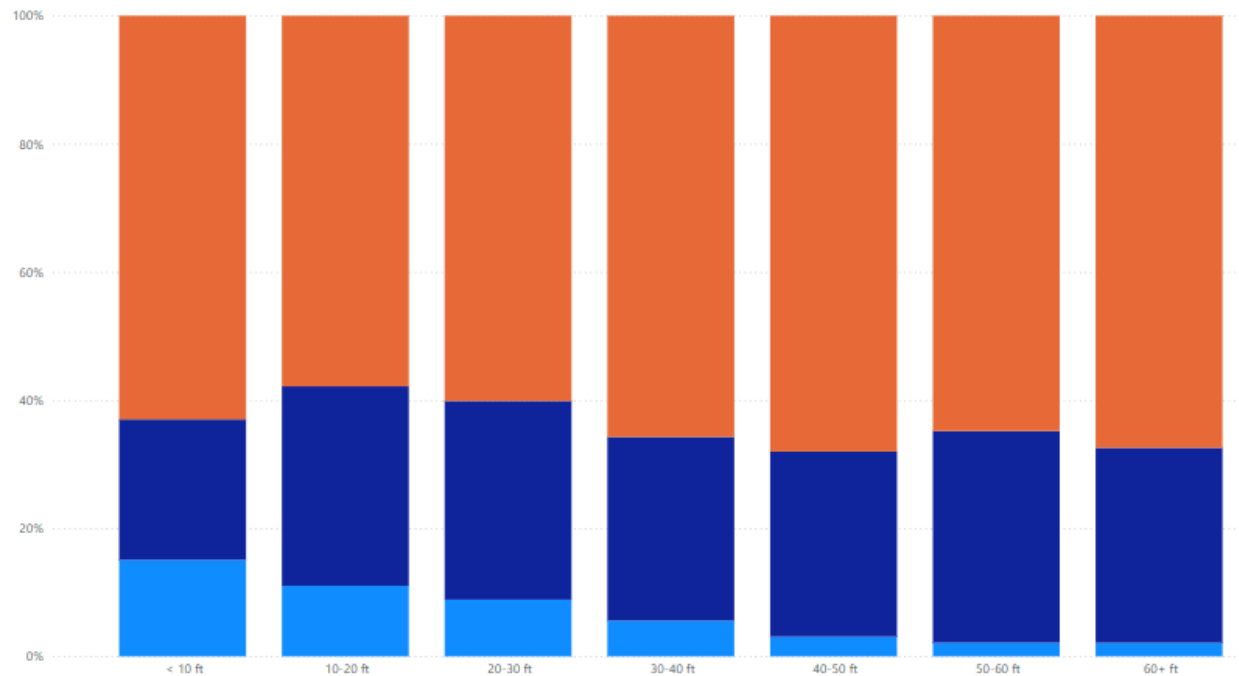


## M and N: Summary Statistics



Count of distanceRange by distanceRange and event

event ● GOAL ● MISS ● SHOT

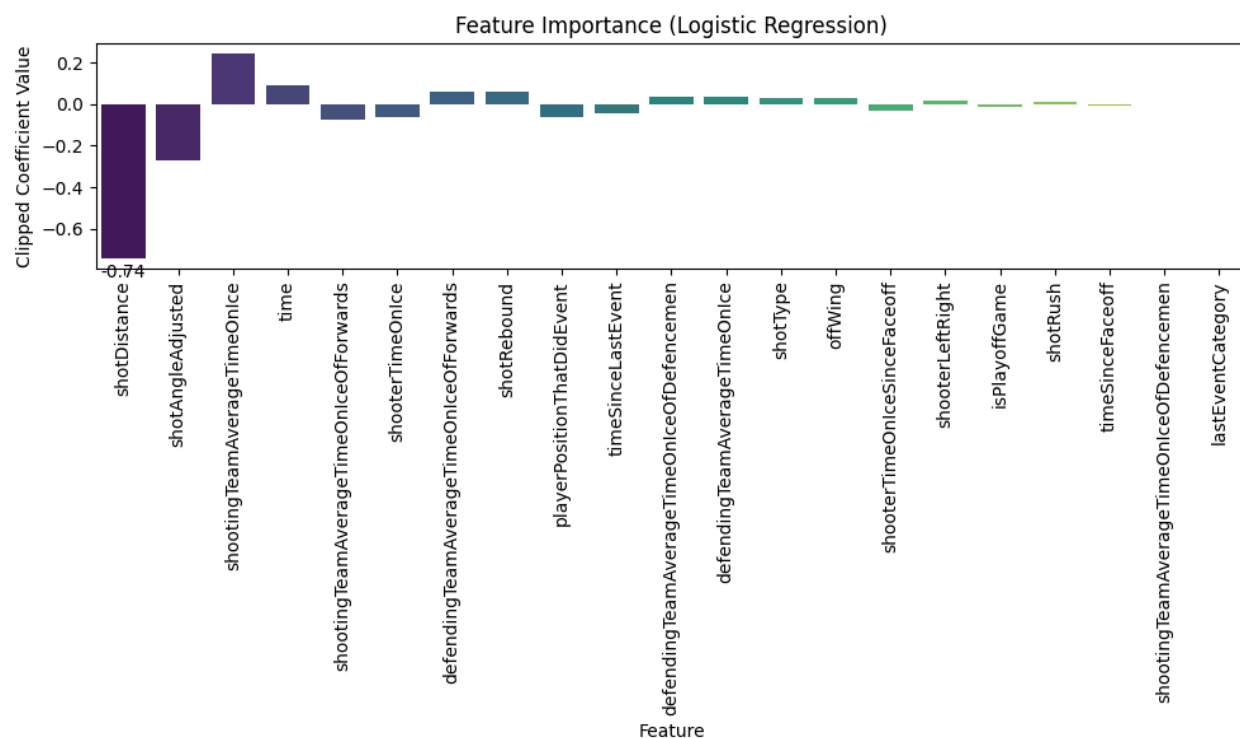


O: Logistic Regression without timeUntilNextEvent

Accuracy: 0.929

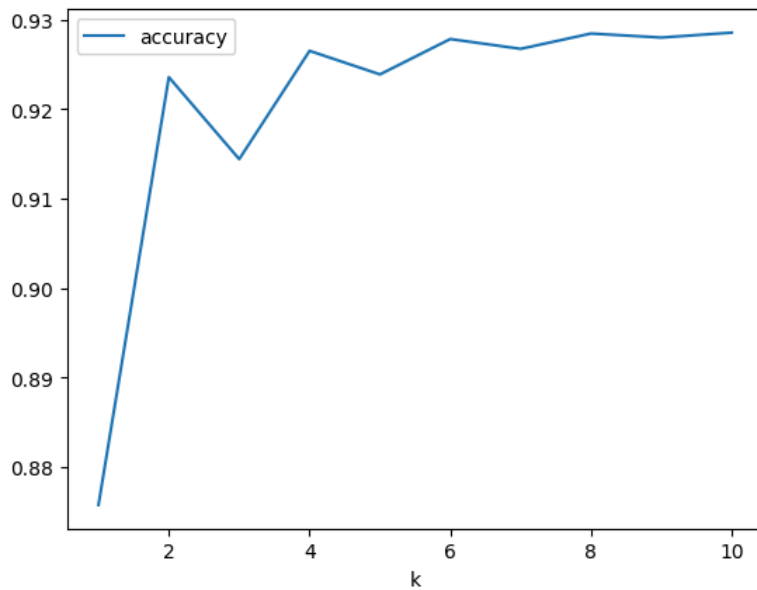


P: Feature Importance and Coefficients

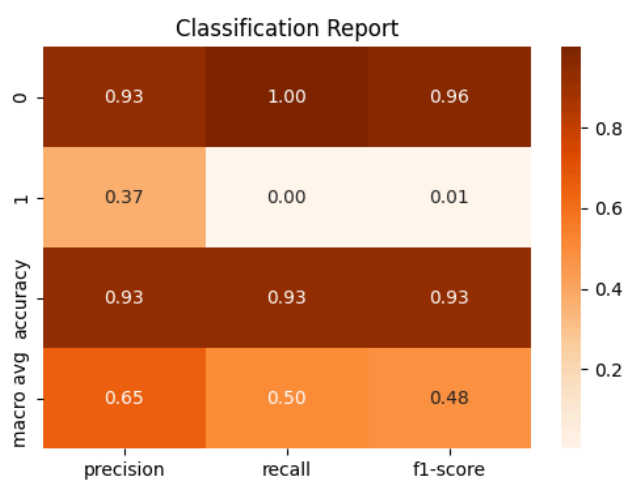


	Coefficients
shotDistance	-0.741491
shotAngleAdjusted	-0.273587
shootingTeamAverageTimeOnIce	0.242884
time	0.093669
shootingTeamAverageTimeOnIceOfForwards	-0.075452
shooterTimeOnIce	-0.063572
defendingTeamAverageTimeOnIceOfForwards	0.063045
shotRebound	0.060657
playerPositionThatDidEvent	-0.060481
timeSinceLastEvent	-0.045419
defendingTeamAverageTimeOnIceOfDefencemen	0.034431
defendingTeamAverageTimeOnIce	0.033777
shotType	0.029982
offWing	0.029611
shooterTimeOnIceSinceFaceoff	-0.029319
shooterLeftRight	0.015711
isPlayoffGame	-0.013919
shotRush	0.013230
timeSinceFaceoff	-0.006519
shootingTeamAverageTimeOnIceOfDefencemen	0.002350
lastEventCategory	-0.000622

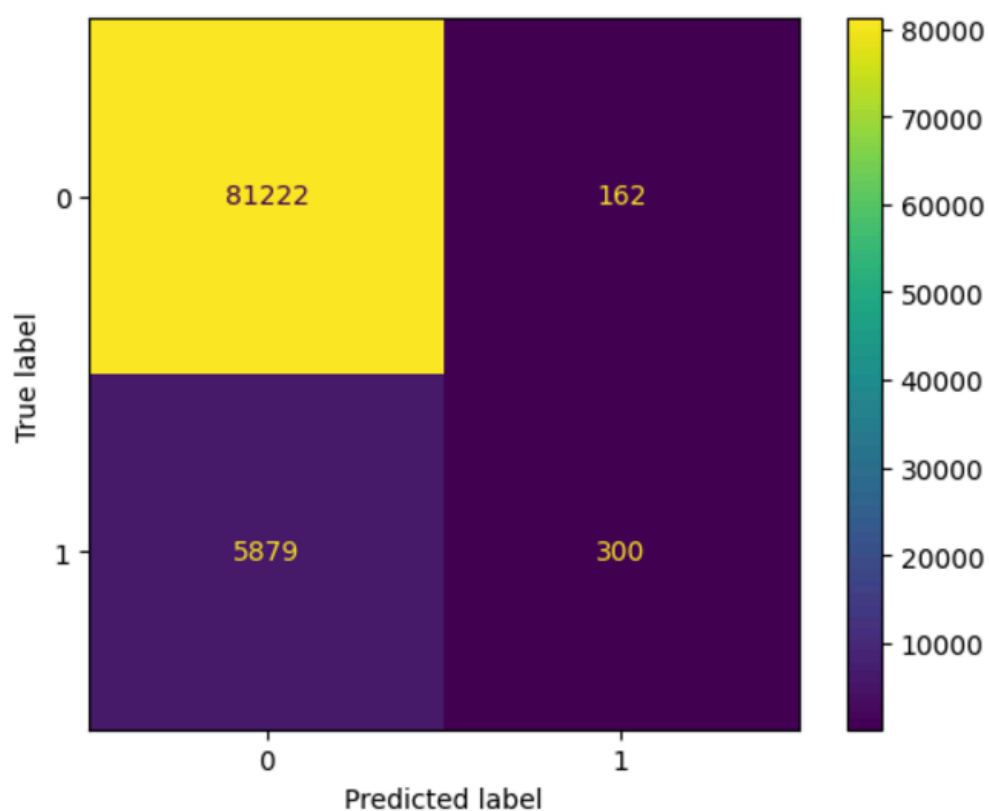
Q: k-NN hyperparameter tuning plot



## R: k-NN Confusion Matrix and Classification Report



S: Random Forest Confusion matrix, basic descriptive statistics, and feature importances

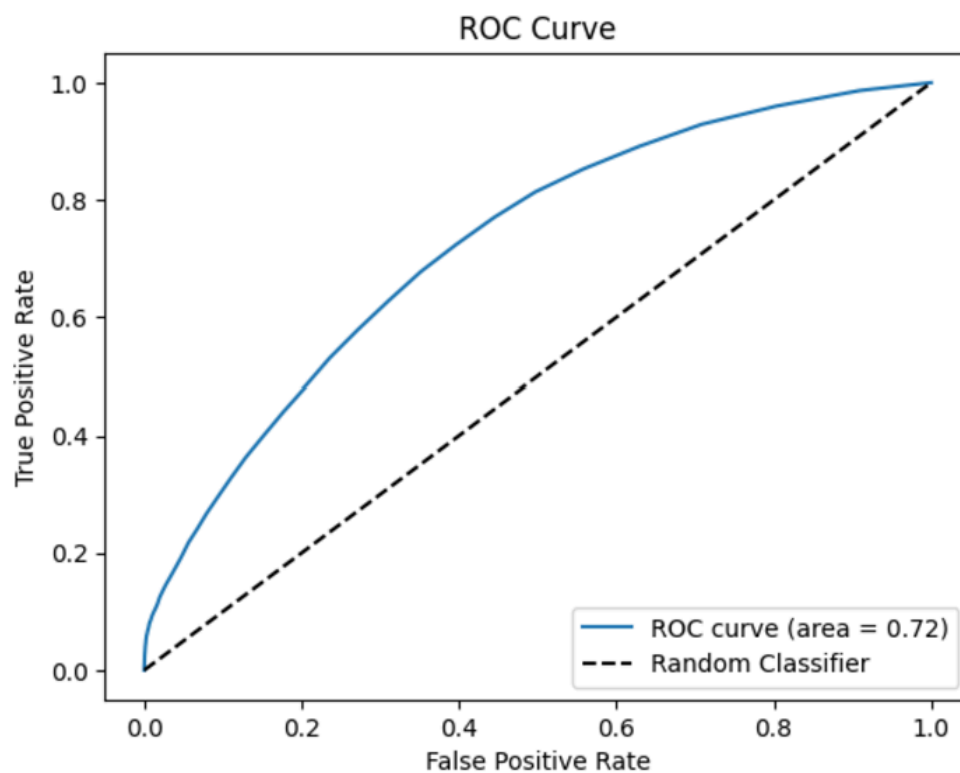


	Feature	Gini	Importance
13	shotDistance	0.098418	
3	goalieIdForShot	0.079921	
17	time	0.079860	
6	shooterPlayerId	0.075408	
12	shotAngleAdjusted	0.068284	
0	defendingTeamAverageTimeOnIce	0.062231	
9	shootingTeamAverageTimeOnIce	0.061468	
2	defendingTeamAverageTimeOnIceOfForwards	0.060219	
1	defendingTeamAverageTimeOnIceOfDefencemen	0.059928	
10	shootingTeamAverageTimeOnIceOfDefencemen	0.058623	
18	timeSinceFaceoff	0.058029	
11	shootingTeamAverageTimeOnIceOfForwards	0.057828	
19	timeSinceLastEvent	0.050476	
7	shooterTimeOnIce	0.049472	
8	shooterTimeOnIceSinceFaceoff	0.046418	
14	shotOnEmptyNet	0.014920	
5	offWing	0.008983	
4	isPlayoffGame	0.004877	
15	shotRebound	0.004148	
16	shotRush	0.000491	

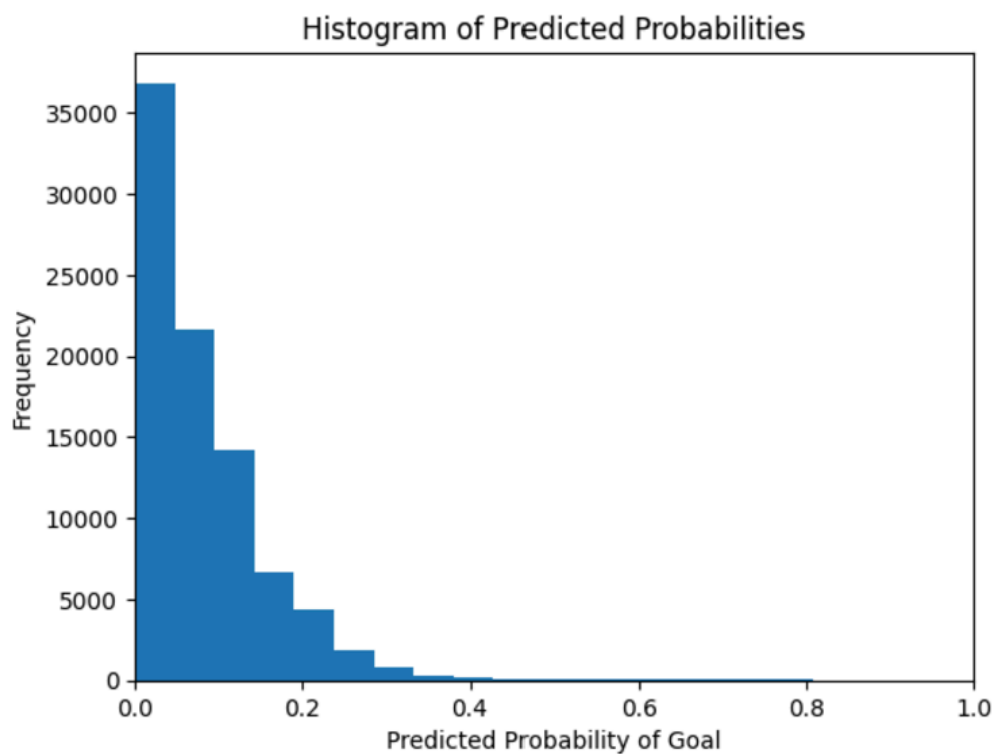
T: Random Forest Accuracy, Precision, Recall, F1 Scores

```
Accuracy: 0.9310096730354145  
Precision: 0.6493506493506493  
Recall: 0.04855154555753358  
F1 Score: 0.09034783918084625
```

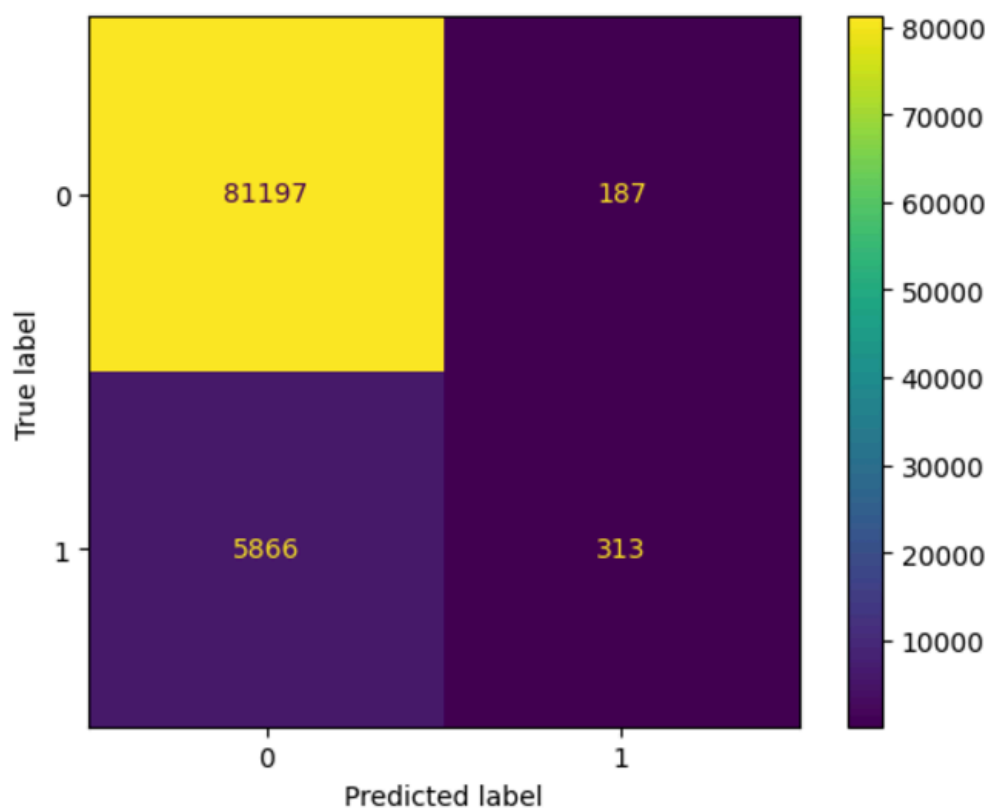
U: Random Forest AOC-ROC



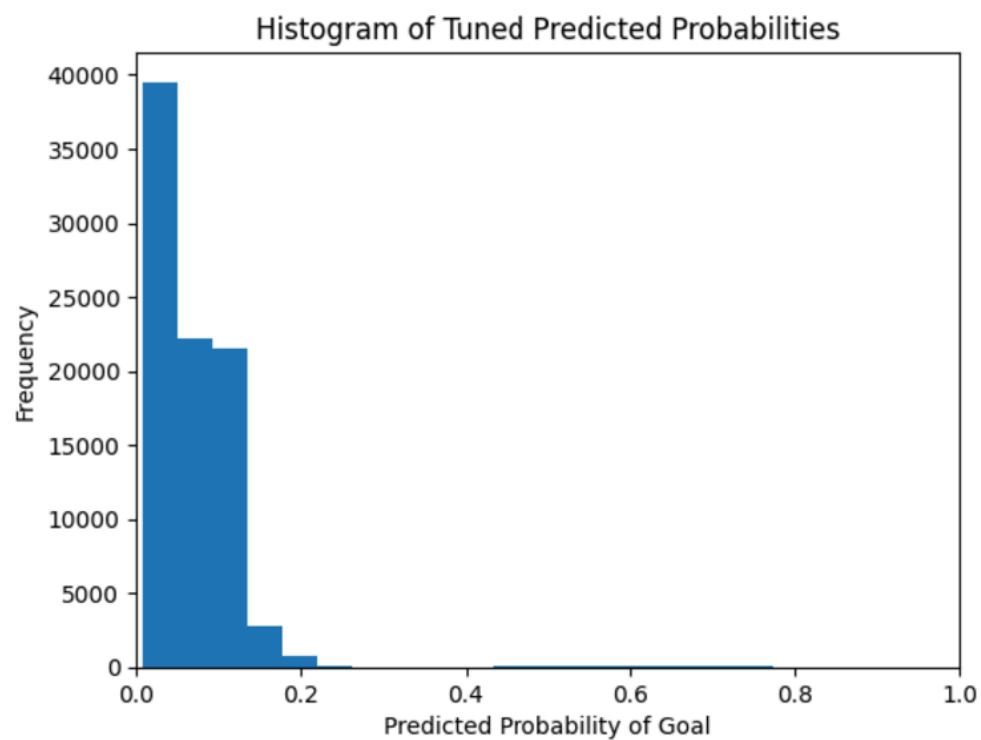
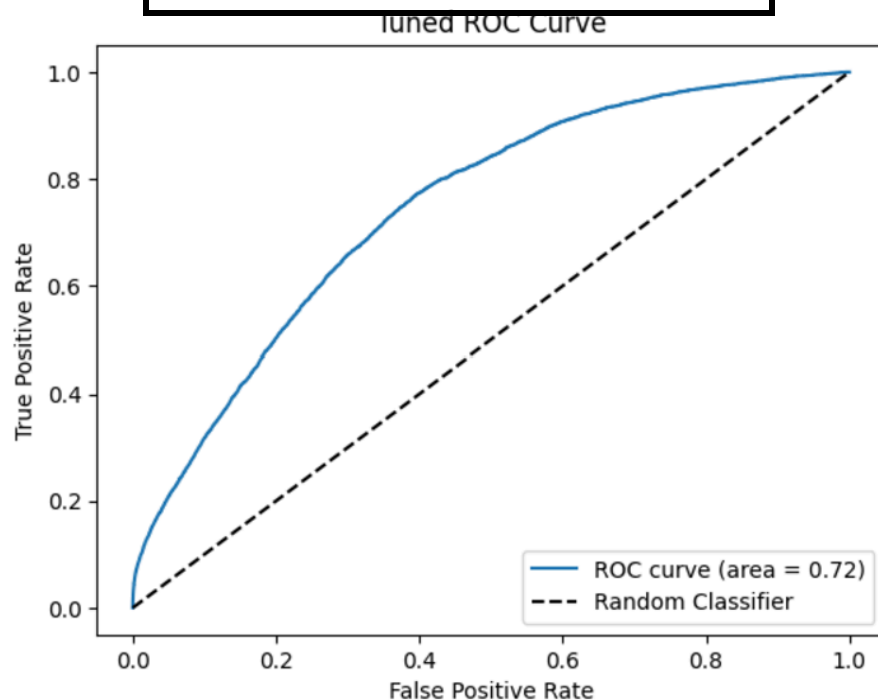
V: Histogram of Random Forest Predicted Probabilities



W: Random Forest Confusion Matrix, Accuracy Scores, AOC-ROC, and Histogram of Predicted Probabilities after Hyperparameter Tuning

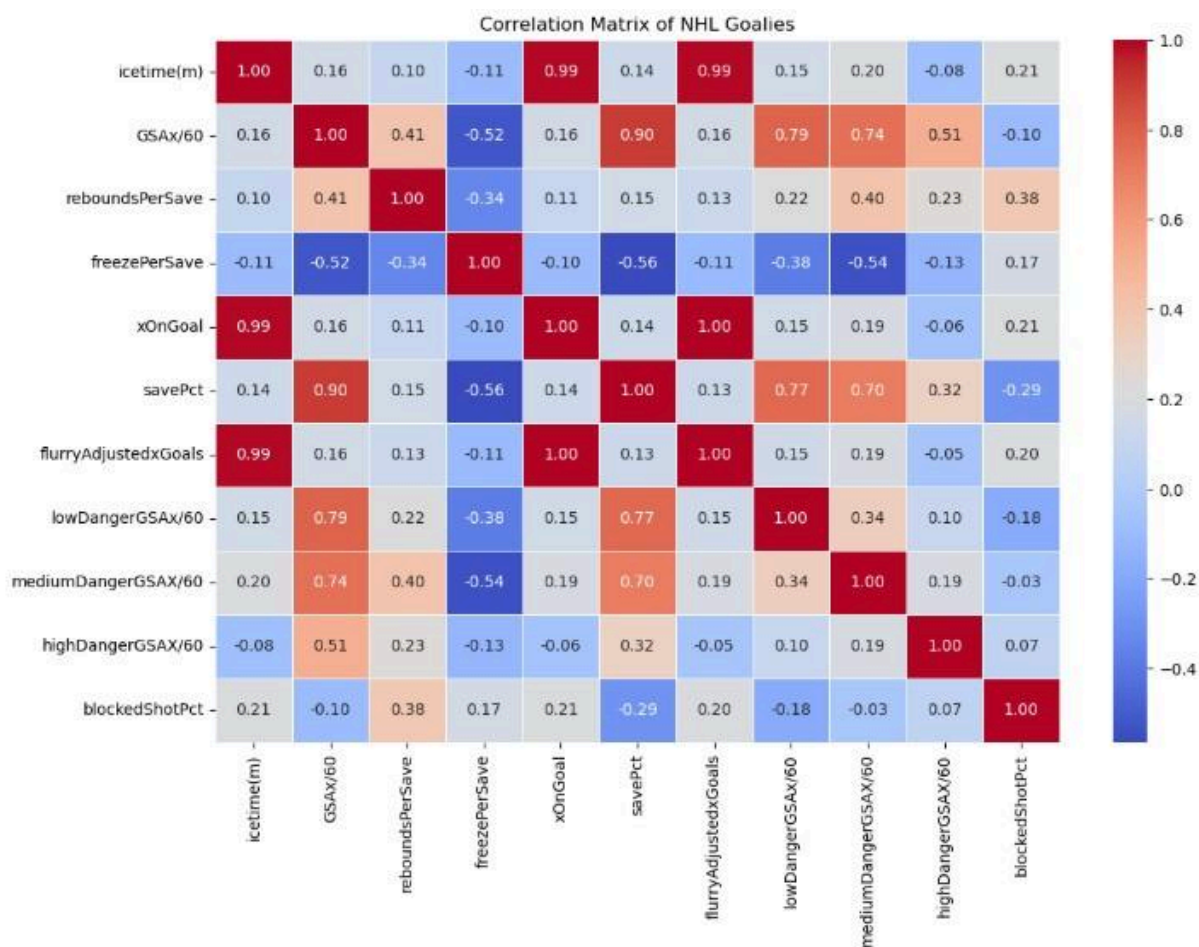


Tuned Accuracy: 0.930872628850085  
Tuned Precision: 0.626  
Tuned Recall: 0.0506554458650267  
Tuned F1 Score: 0.09372660577930828

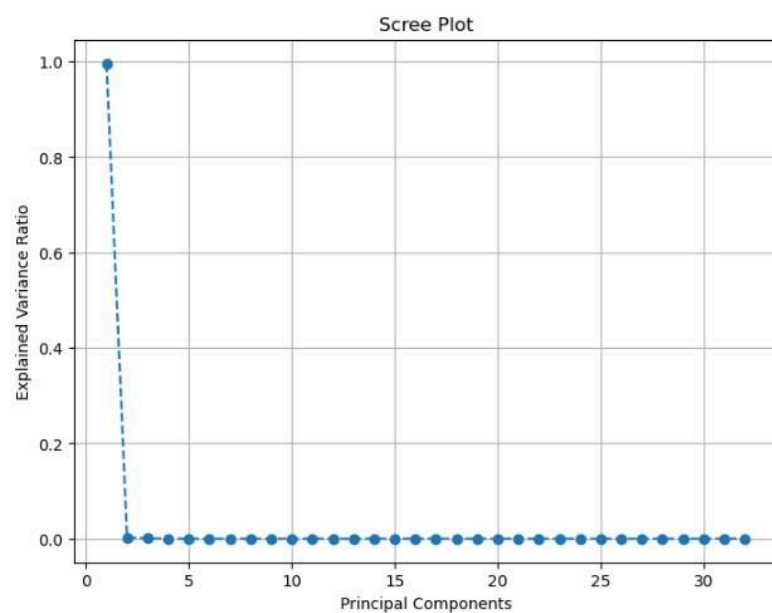




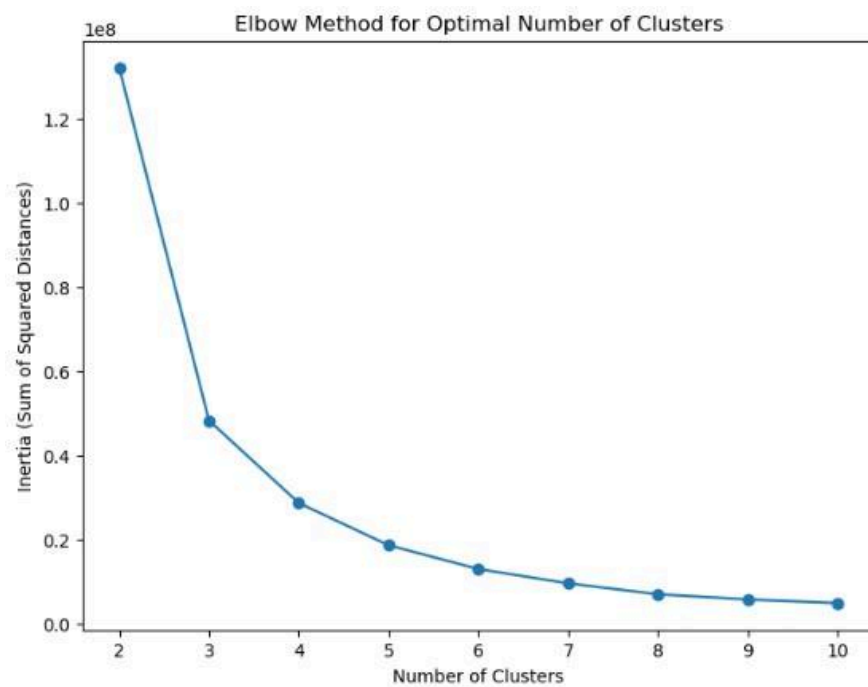
X: Correlation Matrix of Goalie Dataset



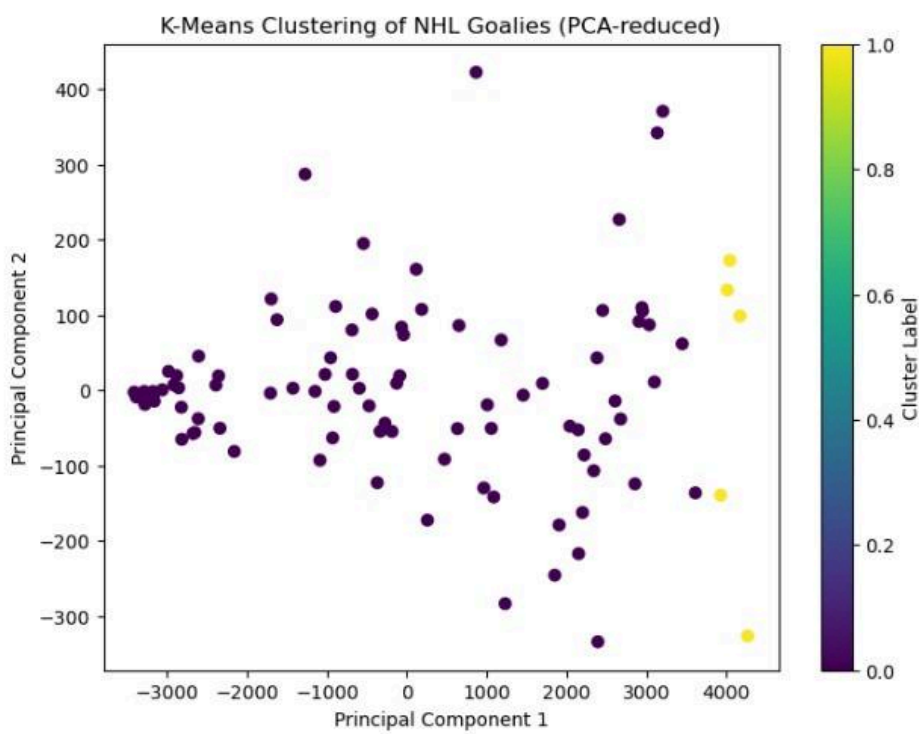
Y: K Means Clustering Scree Plot



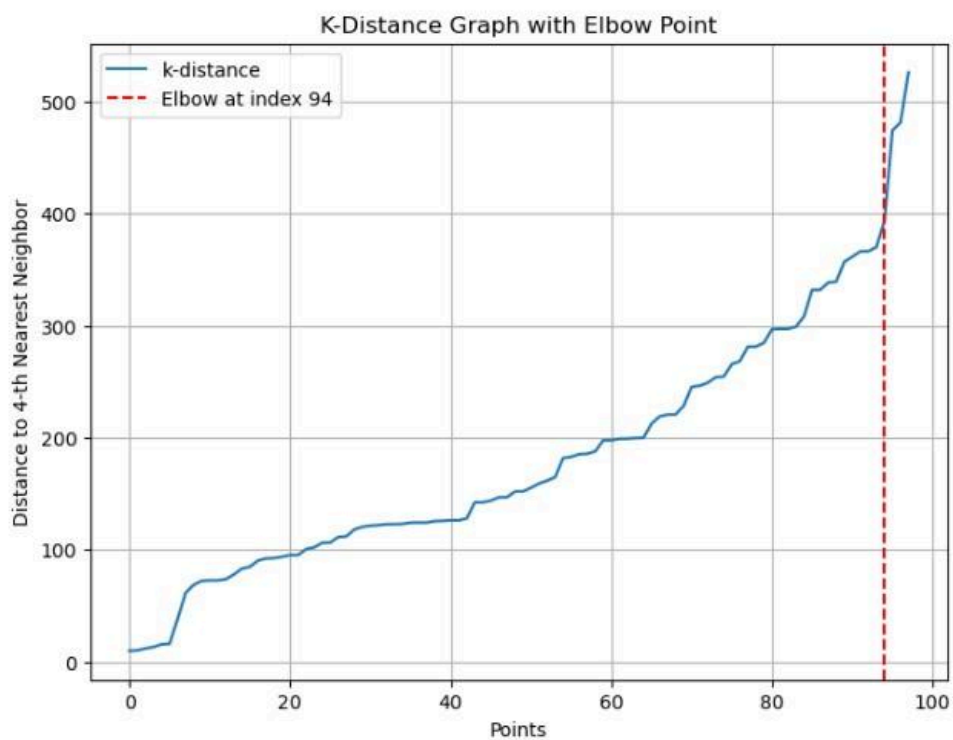
## Z: Elbow Method



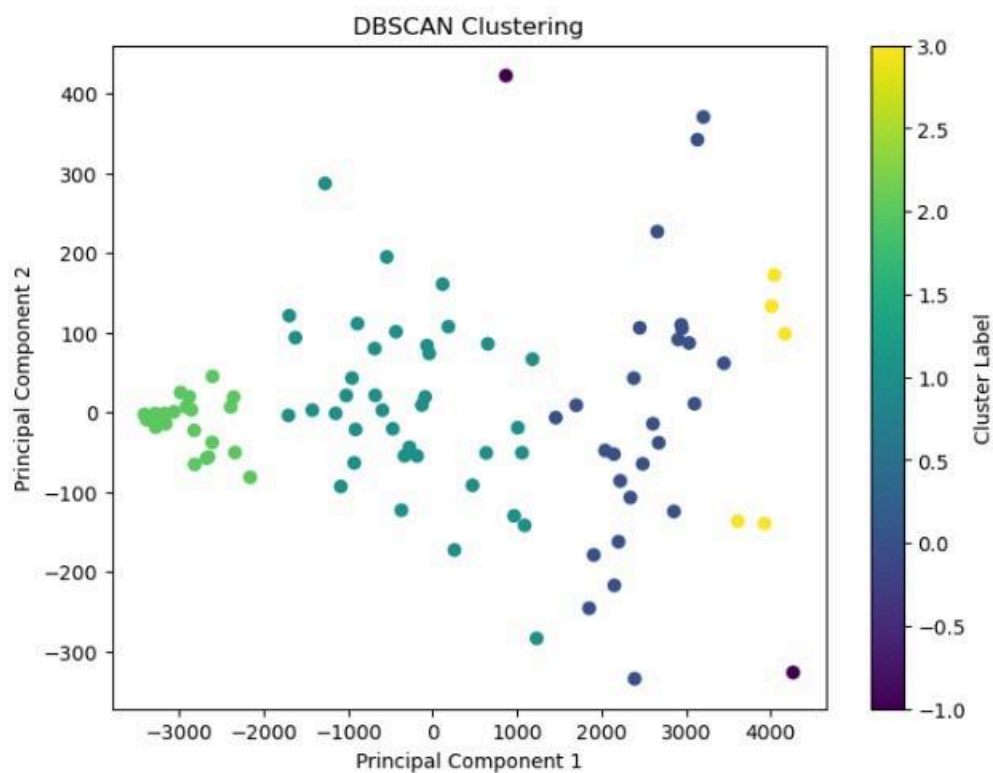
## AA: PCA-reduced K-Means Clustering



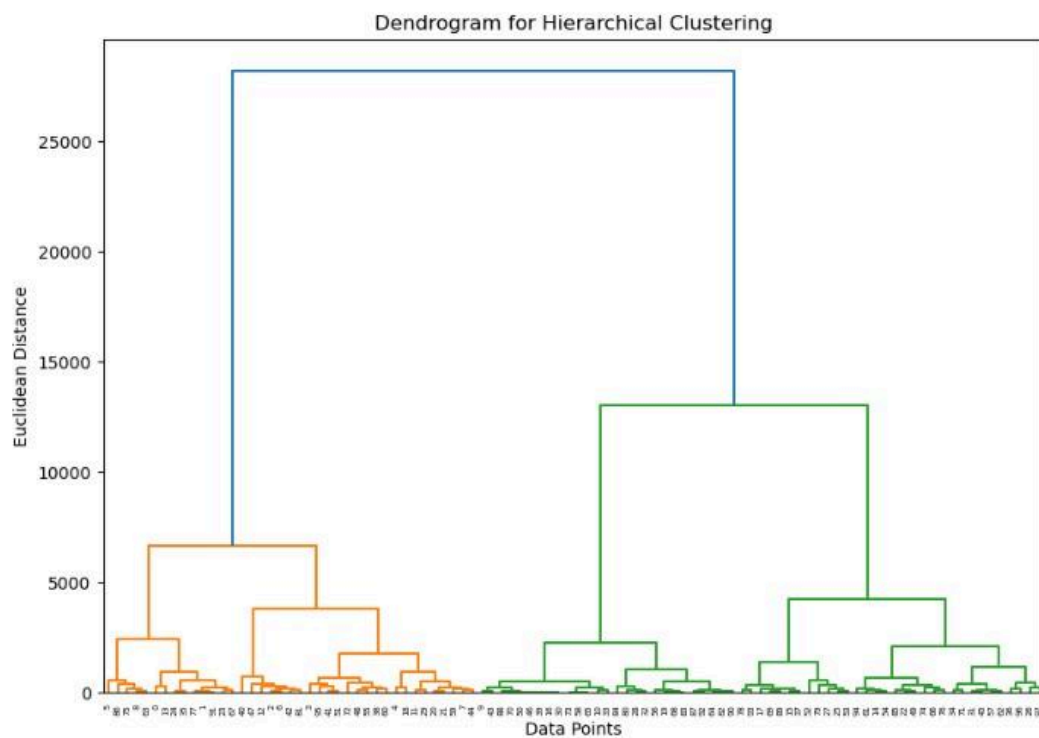
## BB: K-Distance Graph with Elbow Point



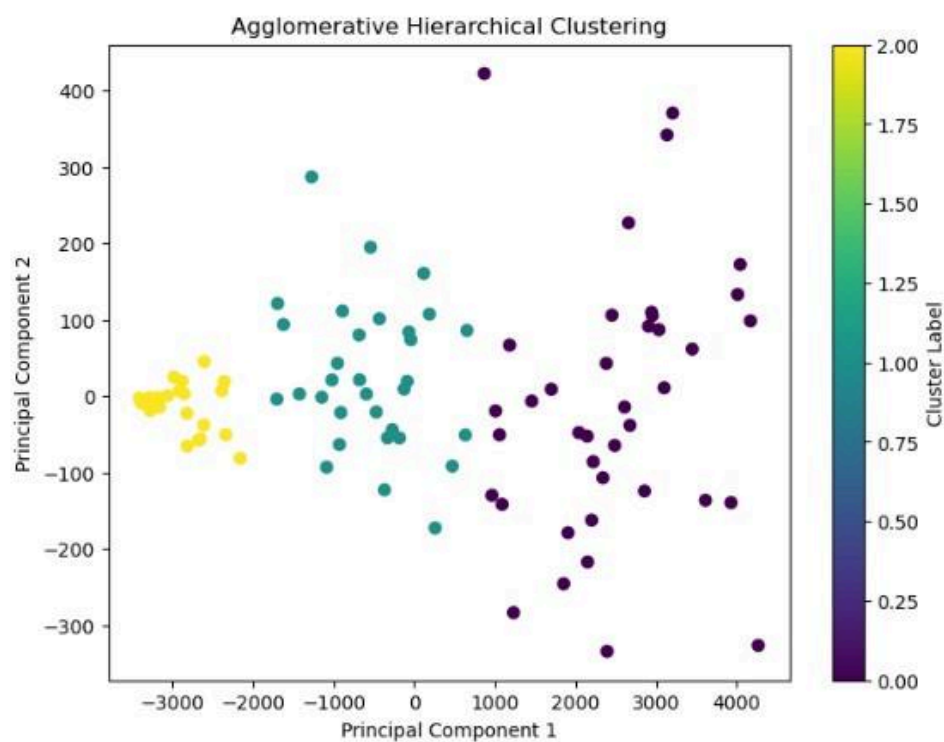
## CC: DBSCAN Clustering of NHL Goalies



### DD: Dendrogram for Hierarchical Clustering



### EE: Hierarchical Clustering of NHL Goalies

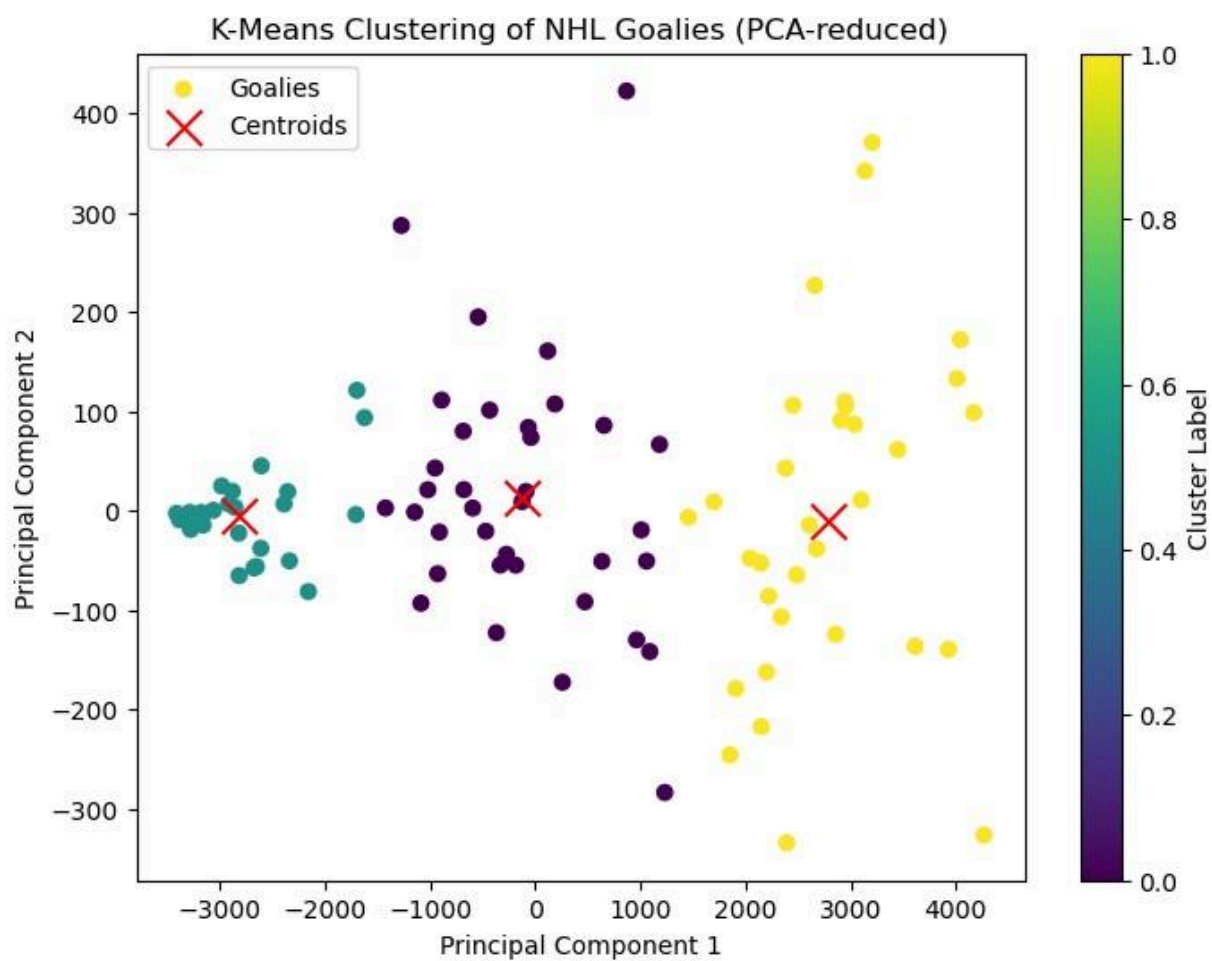


## FF: Independent Variables and Definitions used in Linear Regression analysis

<b><u>Variable:</u></b>	<b><u>Definition:</u></b>
defendingTeamAverageTimeOnIce	The average playing time in seconds the defending team's players have been on the ice
defendingTeamAverageTimeOnIceOfDefence men	The average playing time in seconds the defending team's defense players have been on the ice
defendingTeamAverageTimeOnIceOfForwards	The average playing time in seconds the defending team's forwards have been on the ice
isPlayoffGame	Binary classifier. 0 if regular season game; 1 if playoff game
lastEventCategory	The type of event before the shot. Options are Shot Block, Delayed Penalty, Coach's Challenge, Faceoff, Giveaway, Takeaway, Hit, Shot, Miss. 8 observations are classified as STOP and 1 observation is classified as EGT, but Money puck does not define these terms.
offWing	Whether a shot is being taken from a shooters' dominant side; 0 for dominant and 1 for nondominant
playerPositionThatDidEvent	The position of the player taking the shot. Options are L for Left Wing, R for Right Wing, C for Center, and D for Defenceperson.
shooterLeftRight	Identifies the shooter's handedness
shooterTimeOnIce	The number of game seconds that have passed since the shooter started their shift
shooterTimeOnIceSinceFaceoff	The minimum number of game seconds that have passed since the shooter started their shift or since the previous faceoff

shootingTeamAverageTimeOnIce	The average playing time in seconds the shooting team's players have been on the ice
shootingTeamAverageTimeOnIceOfDefence men	The average playing time in seconds the shooting team's defense players have been on the ice
shootingTeamAverageTimeOnIceOfForwards	The average playing time in seconds the shooting team's forwards have been on the ice
shotAngleAdjusted	The absolute value of the shot's angle to the net. Ranges from 0 to 88.5 degrees
shotDistance	The distance between the shot's location and the net. Ranges from 1 to 98.4 feet
shotRebound	Binary classifier. 0 if the shot taken is not a rebound. 1 if the shot taken is a rebound
shotRush	Binary classifier. 0 if the shot taken is not a rush opportunity. 1 if the shot taken is a rush opportunity.
shotType	Identifies the shot type. Options are wrist shot, slapshot, snapshot, backhand shot, deflection, tip shot, and wraparound.
time	How many seconds into the game the shot took place. Ranges from 1 to 8387 (4OT playoff game).
timeSinceFaceoff	Number of seconds since the last faceoff before the shot
timeSinceLastEvent	Time between shot and the previous game event

## GG K-Means Clustering Analysis with Centroids





## HH Feature Importances for Clustering Analysis

	PC1	PC2	importance
icetime(m)	4.668364e-01	8.486741e-01	1.315510e+00
totalShotAttemptsFaced	6.076013e-01	-4.299250e-01	1.037526e+00
unblockedShotAttempts	4.733441e-01	-2.339936e-01	7.073377e-01
blocked_shot_attempts	1.342572e-01	-1.959314e-01	3.301886e-01
xOnGoal	2.467045e-01	-1.345904e-02	2.601636e-01
ongoal	2.356902e-01	7.765605e-03	2.434558e-01
saves	2.139151e-01	2.138496e-02	2.353000e-01
xFreeze	5.999812e-02	2.815943e-03	6.281406e-02
freeze	4.813557e-02	-6.966410e-03	5.510198e-02
goals	2.177509e-02	-1.361936e-02	3.539444e-02
rebounds	2.502121e-02	-6.605992e-03	3.162721e-02
xGoals	2.281905e-02	-3.324124e-03	2.614318e-02
flurryAdjustedxGoals	2.169346e-02	-1.904225e-03	2.359768e-02
xRebounds	1.682621e-02	-4.718340e-03	2.154455e-02
lowDangerGoals	8.058842e-03	-1.174278e-02	1.980162e-02
mediumDangerGoals	7.580139e-03	-1.111064e-02	1.869077e-02
highDangerGoals	6.136105e-03	9.234058e-03	1.537016e-02
highDangerGSAX	2.083258e-03	-9.480646e-03	1.156390e-02
mediumDangerGSAX	-6.968283e-05	1.134356e-02	1.141324e-02
GSAX	1.043967e-03	1.029523e-02	1.133920e-02
lowDangerxGoals	7.089506e-03	-3.308414e-03	1.039792e-02
lowDangerGSAX	-9.693356e-04	8.434365e-03	9.403700e-03
highDangerxGoals	8.219363e-03	-2.465878e-04	8.465951e-03
mediumDangerxGoals	7.510456e-03	2.329206e-04	7.743377e-03
blockedShotPct	2.669550e-06	-3.689672e-05	3.956627e-05
freezePerSave	-3.920201e-06	-2.032802e-05	2.424822e-05
savePct	1.846528e-06	1.010351e-05	1.195003e-05
reboundsPerSave	1.402130e-06	-1.996854e-06	3.398984e-06
highDangerGSAX/60	-2.149958e-09	-1.063586e-07	1.085086e-07
mediumDangerGSAX/60	8.601447e-09	8.750247e-08	9.610391e-08
lowDangerGSAX/60	8.242002e-09	8.738524e-08	9.562724e-08
GSAX/60	1.470062e-08	6.844945e-08	8.315007e-08



## II Cluster 0

name	Cluster	icetime(m)	totalShotAttemptsFaced	unblockedShotAttempts	blocked_shot_attempts
Antti Raanta	0	1265.183333	1184	946	238
Dustin Wolf	0	950.5	1213	934	279
Spencer Martin	0	1072.216667	1377	1078	299
Nico Daws	0	1144.05	1435	1136	299
Ville Husso	0	1012.916667	1439	1136	303
Anthony Stolarz	0	1506.45	1640	1331	309
Kevin Lankinen	0	1191.416667	1466	1148	318
Calvin Pickard	0	1296.35	1497	1164	333
Devon Levi	0	1294.9	1631	1298	333
Martin Jones	0	1169.666667	1535	1196	339
Dan Vladar	0	1128.516667	1545	1191	354
David Rittich	0	1364.516667	1634	1273	361
Jonas Johansson	0	1477.483333	1747	1384	363
James Reimer	0	1352.683333	1771	1407	364
Daniil Tarasov	0	1376.466667	1855	1476	379
Joel Hofer	0	1628.65	1985	1594	391
Laurent Brossoit	0	1351.033333	1736	1329	407
Cayden Primeau	0	1324.833333	1893	1472	421
Vitek Vanecek	0	1791.816667	2065	1638	427
Jonathan Quick	0	1582.733333	2009	1579	430
Joseph Woll	0	1471.65	1980	1537	443
Casey DeSmith	0	1663.483333	2016	1564	452
Carter Hart	0	1455.183333	1943	1486	457
Scott Wedgewood	0	1789.133333	2143	1680	463
Anton Forsberg	0	1569.9	2010	1539	471
Pyotr Kochetkov	0	2371.483333	2419	1920	499
Arvid Soderblom	0	1744.333333	2389	1867	522
Philipp Grubauer	0	1996.6	2446	1907	539
Semyon Varlamov	0	1594.733333	2323	1773	550
Darcy Kuemper	0	1866.533333	2486	1936	550
Karel Vejmelka	0	2040.133333	2663	2110	553
Alex Nedeljkovic	0	2060.466667	2738	2138	600
Kaapo Kahkonen	0	1914.216667	2903	2289	614
Jake Allen	0	1986.883333	2783	2165	618
Marc-Andre Fleury	0	2232.666667	2783	2163	620
Adin Hill	0	1968.816667	2742	2102	640


## JJ Cluster 1

name	Cluster	icetime(m)	totalShotAttemptsFaced	unblockedShotAttempts	blocked_shot_attempts
Yaniv Perets	1	12.76666667	1	1	0
Kenneth Appleby	1	20	20	16	4
Michael Hutchinson	1	57.8	76	64	12
Matt Villalta	1	71.96666667	76	62	14
Georgi Romanov	1	59.25	80	66	14
Louis Domingue	1	60	78	63	15
Malcolm Subban	1	60	87	70	17
Matt Murray	1	60	73	54	19
Yaroslav Askarov	1	81.66666667	105	80	25
Magnus Hellberg	1	119.8833333	139	112	27
Chris Driedger	1	119.5833333	158	125	33
Ivan Fedotov	1	121.15	153	116	37
Jesper Wallstedt	1	179.2833333	213	166	47
Jack Campbell	1	266.7833333	297	250	47
Arturs Silovs	1	243.2166667	258	199	59
Felix Sandstrom	1	263.4833333	312	231	81
Mads Sogaard	1	281.5666667	342	258	84
Cal Petersen	1	276.9166667	375	284	91
Devin Cooley	1	301.4833333	460	368	92
Hunter Shepard	1	244.8333333	399	298	101
Pheonix Copley	1	436.2833333	478	372	106
Jiri Patera	1	316.7666667	479	373	106
Matt Tomkins	1	360.2166667	507	386	121
Eric Comrie	1	519.9	626	503	123
Ivan Prosvetov	1	494.1833333	612	486	126
Magnus Chrona	1	471.25	670	525	145
Jet Greaves	1	515.8666667	776	620	156
Frederik Andersen	1	912.8833333	982	788	194
Akira Schmid	1	914.15	1025	830	195
Justus Annunen	1	801.0833333	1029	810	219

## KK Cluster 2


name	Cluster	icetime(m)	totalShotAttemptsFaced	unblockedShotAttempts	blocked_shot_attempts
Ilya Samsonov	2	2300.966667	2981	2320	661
Linus Ullmark	2	2400.133333	3090	2428	662
Lukas Dostal	2	2322.816667	3277	2570	707
Tristan Jarry	2	2740.633333	3477	2765	712
Joey Daccord	2	2832.783333	3501	2781	720
Thatcher Demko	2	3015.883333	3558	2825	733
John Gibson	2	2561.483333	3442	2705	737
Elvis Merzlikins	2	2258.666667	3297	2546	751
Jeremy Swayman	2	2565.85	3400	2649	751
Filip Gustavsson	2	2526.55	3342	2589	753
Connor Ingram	2	2802.933333	3684	2917	767
Alex Lyon	2	2498.85	3482	2696	786
Sam Montembeault	2	2428.766667	3474	2685	789
Stuart Skinner	2	3361.35	3829	3021	808
Sergei Bobrovsky	2	3414.233333	3857	3046	811
Ukko-Pekka Luukkonen	2	3081.233333	3827	3001	826
Mackenzie Blackwood	2	2437.133333	3665	2827	838
Jacob Markstrom	2	2831.116667	3687	2848	839
Jake Oettinger	2	3084.583333	3825	2986	839
Logan Thompson	2	2645.466667	3591	2729	862
Andrei Vasilevskiy	2	3062.866667	3828	2954	874
Cam Talbot	2	3116.116667	3904	3021	883
Charlie Lindgren	2	2852.383333	3883	2996	887
Joonas Korpi	2	3080.466667	3971	3072	899
Igor Shesterkin	2	3277.1	4152	3231	921
Connor Hellebuyck	2	3567.333333	4420	3495	925
Petr Mrazek	2	3152.666667	4293	3366	927
Jordan Binnington	2	3291.483333	4477	3525	952
Alexandar Georgiev	2	3637.95	4449	3488	961
Samuel Ersson	2	2809.466667	3738	2772	966
Juuse Saros	2	3624.633333	4548	3554	994
Ilya Sorokin	2	3325.983333	4813	3697	1116


Source Code:

 Heatmap Code.pdf

 Logistic Regression and KNN.pdf

 KNN with ROC.pdf

 DAT490 Random Forest PDF.pdf

 GoalieClusterAnalysisCode.pdf