Hockey Analytics Final Project Report

Jeff Jung

Fall 2023 Nov 26th

Contents

DAR Project and Group Members

Jeff Jung Caleb Smith Liebin Zhang Ashley Woodson Amy Enyenihi

Project

Hockey Analytics

Abstract

This research project explores the application of data analytics to gain insights into the performance of the RPI women's hockey team based on shot data. The project involves a multi-step analysis, beginning with

thorough data cleaning to eliminate outliers that may skew the results. Subsequently, continuous variables such as puck distance, puck angle, and player speeds are discretized through careful categorization determined by the examination of density plots. The discretized variables are then used in the creation of cluster models using the k-means algorithm. This cluster analysis aims to unveil patterns and relationships within the data, shedding light on factors influencing successful goals or saves. By employing a systematic approach that combines data cleaning, categorization, and clustering, this project contributes to the understanding of key variables impacting the outcomes of hockey shots.

Introduction and Background

The purpose of hockey analytics is to utilize statistical analysis and advanced metrics to gain deeper insights into player and team performance, enabling data-driven decision-making in various aspects of the sport. By examining detailed statistics and metrics beyond traditional measures, hockey analytics helps teams evaluate player contributions, optimize strategic approaches, and identify areas for improvement. Ultimately, the goal is to enhance overall team performance, inform coaching decisions, and provide a more comprehensive understanding of the game.

The dataset utilized for this project was sourced from the RPI women's hockey team and comprises a total of 105 shots. It encompasses nine continuous and five categorical variables. A more detailed explanation of the dataset will be provided in subsequent sections.

In light of limited data on goals, my research primarily concentrated on exploring datasets associated with saves. Initially, I sought to understand the central tendency and variability of saves by calculating the median and standard deviation. However, these summary statistics lacked granularity. Consequently, I turned to density plots for visualization, revealing interesting patterns in the data distribution. Notably, each variable exhibited distinct concentration ranges, prompting me to categorize continuous variables based on the observed density plot shapes.

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## -- Attaching packages -------------------------------------- tidymodels 1.1.1 --

## v broom        1.0.5     v rsample      1.2.0
## v dials        1.2.0     v tibble       3.2.1
## v ggplot2      3.4.3     v tidyr        1.3.0
## v infer        1.0.5     v tune         1.1.2
## v modeldata    1.2.0     v workflows    1.1.3
## v parsnip      1.1.1     v workflowsets 1.0.1
## v purrr        1.0.2     v yardstick    1.2.0
## v recipes      1.0.8

## -- Conflicts ----------------------------------------- tidymodels_conflicts() --
## x purrr::discard() masks scales::discard()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## x recipes::step()  masks stats::step()
## * Use tidymodels_prefer() to resolve common conflicts.

## Median of Puck Distance for Saves: 454.3136
## Standard Deviation of Puck Distance for Saves: 174.835
```

Problems Tackled

The main problem that was considered was how we can improve the analysis. One way to do that is by discretizing continuous variables, specifically puckDist, puckAngle, puckSpeed, shooterSpeed, goalieDist, goalieAngle, posTime, defDist, and defAngle. The initial step involves cleaning the data to mitigate potential skewness. Subsequently, we examine density plots using the cleaned data to identify suitable ranges for discretization. Employing the k-means method, we create two distinct sets of clusters: one for continuous variables and another for categorical variables. This process aims to enhance the granularity of our analysis and facilitate a more nuanced exploration of the dataset.

Data Description

The main dataset utilized, named "shots," is derived from games played by the RPI women's hockey team, comprising a total of 105 observations or shots. It encompasses nine continuous variables: puckDist, puckAngle, puckSpeed, shooterSpeed, goalieDist, goalieAngle, posTime, defDist, and defAngle. Additionally, there are five categorical variables: NumOffense, NumDefense, rightHanded, closestDef, and shotOutcome.

In the initial stage of data processing, a two-step approach to data cleaning is implemented. The process involves two functions: detect_outlier and remove_outlier. The detect_outlier function identifies all outliers in the dataset, while the remove_outlier function eliminates the data points classified as outliers by the detect_outlier function. This systematic approach ensures the removal of any anomalous or extreme values, contributing to a cleaner and more reliable dataset for subsequent analyses.

In the process of analyzing continuous variables, density plots were utilized to identify suitable cutoffs for categorization. This involved assessing the concentration areas within each density plot to determine distinct value ranges. For instance, in the case of puckDist, the cutoff value of 310 was chosen because an approximately equal number of data points lie both below and above this value in the density plot. The number of categories established was contingent on the density plot's shape, allowing for a more nuanced categorization approach. As an illustration, defenderAngle was categorized using cutoffs at 63 and 112, aligning with the observation of three concentrated data ranges discerned from its density plot.

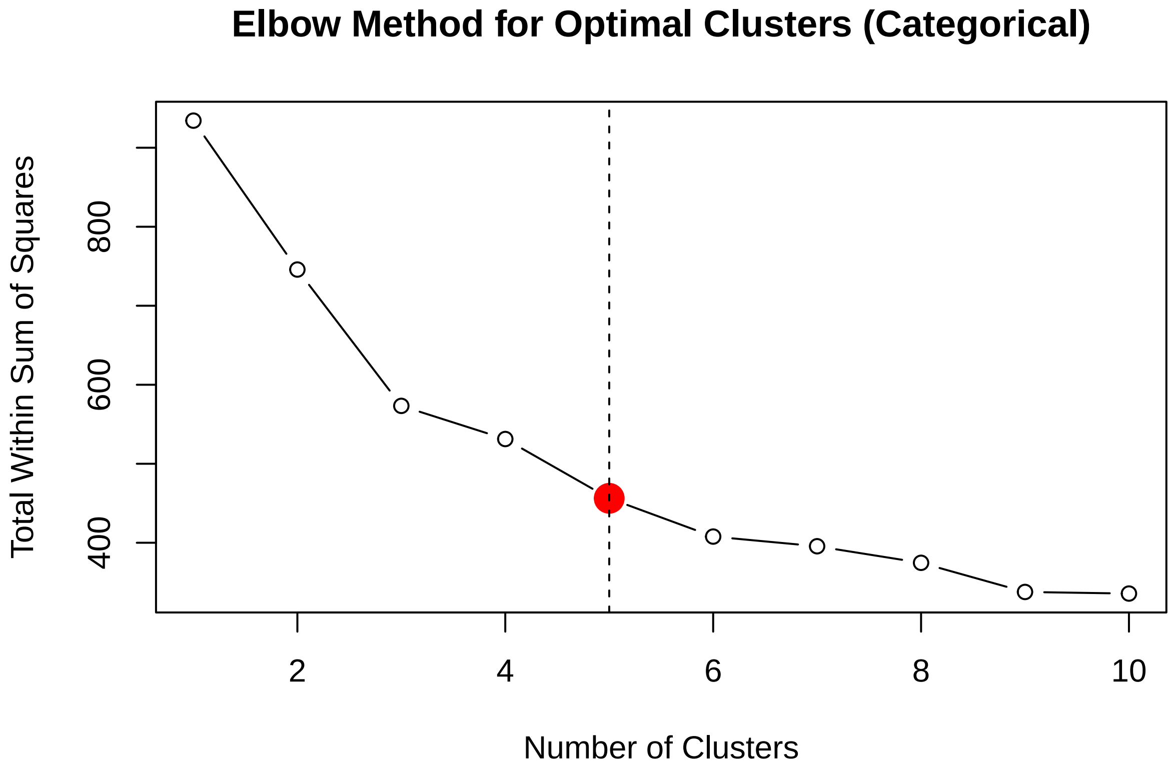
The assignment of discretized values follows a straightforward process. For instance, all puckDist values below 310 are assigned 0, and those above 310 are assigned 1. In cases where there are three categories, an additional value is assigned. For example, the categorized defenderAngle variable assigns values below 63 to 0, values between 63 and 112 to 1, and values above 112 to 2. The specific cutoffs for each feature are detailed in the "Variable Categorization Based on Quantiles" table. To ensure these variables function as discretized categories, the assigned values are converted into factors. This systematic approach enhances the interpretability and utility of the discretized variables in subsequent analyses.

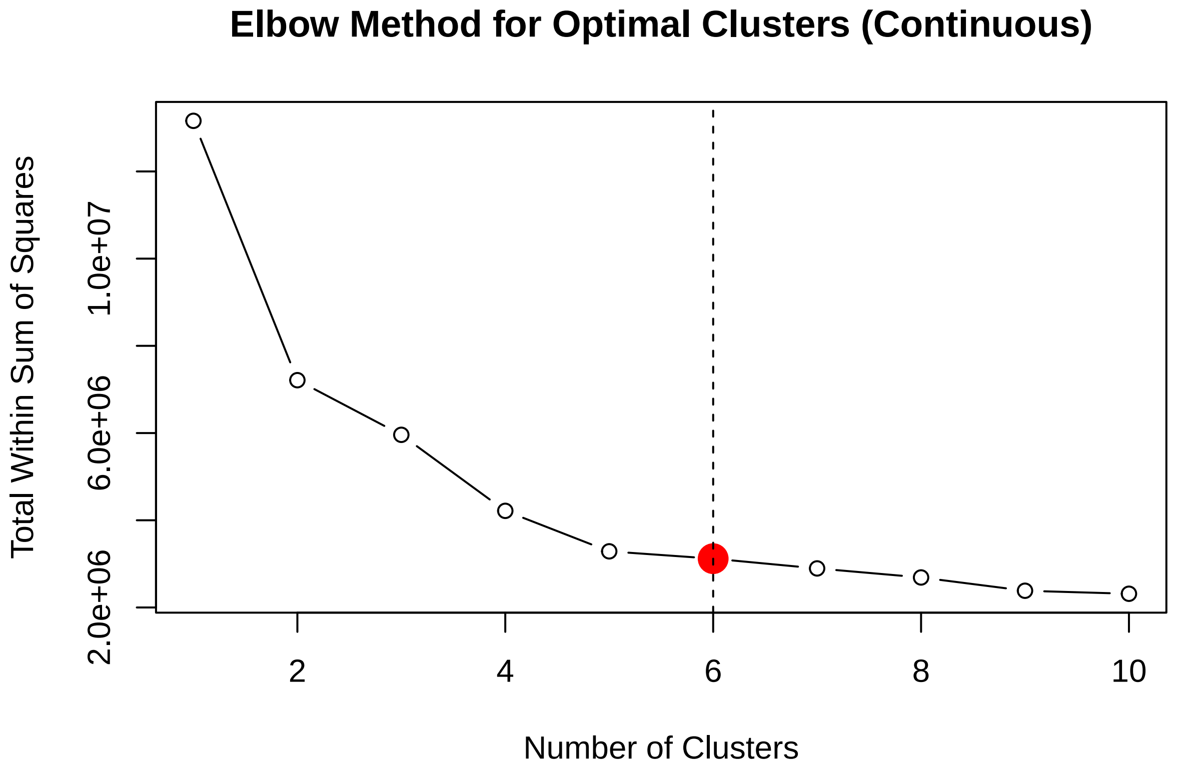
Table 1: Variable Categorization

Variable	Cutoffs
puckDist	310
puckAngle	90
puckSpeed	15, 50
shooterSpeed	17, 25
goalieDist	58, 80
goalieAngle	80
posTime	45
defDist	250
defAngle	63, 112

Data Analytics Methods

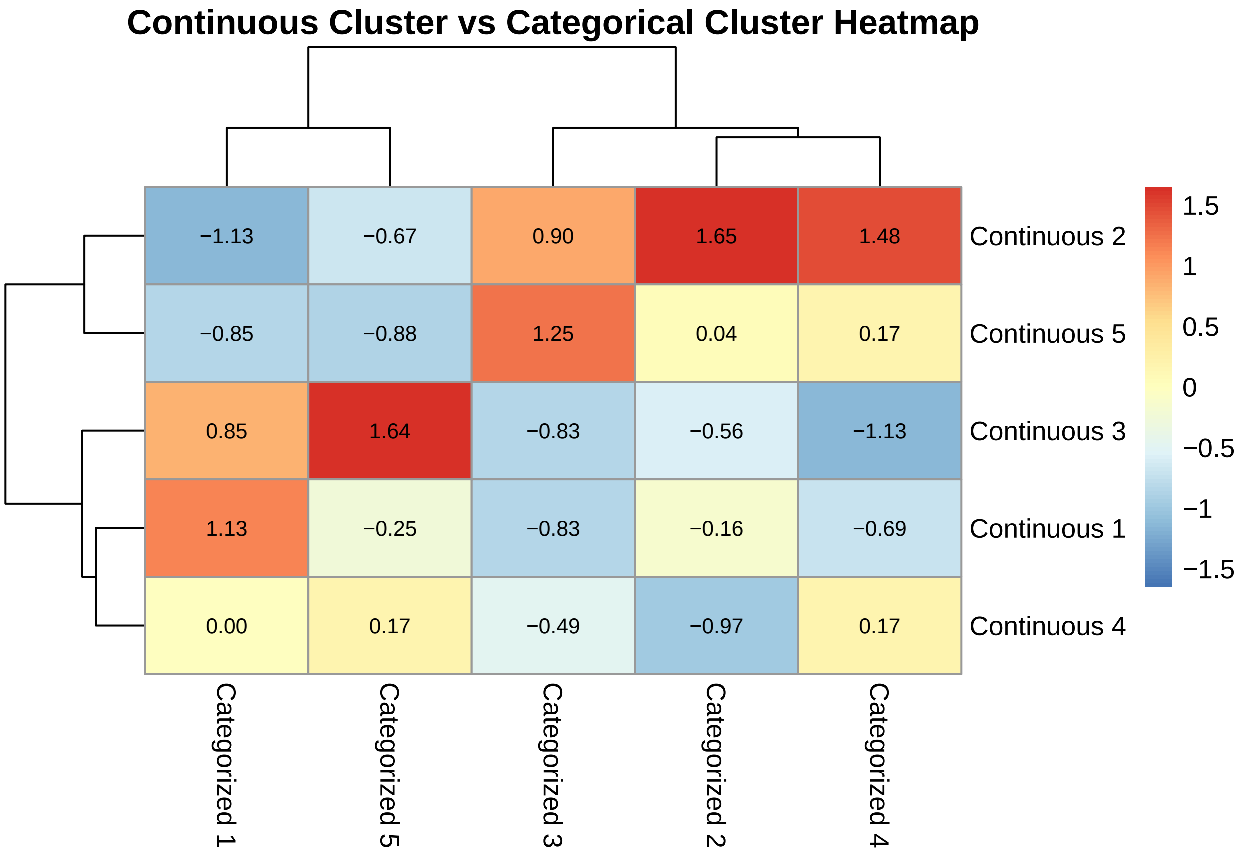
After obtaining a cleaned dataset devoid of outliers, I proceeded to exclude the original categorical variables from the shot__stats__goal data. This facilitated a focused comparison between the newly created categorical variables and their corresponding continuous counterparts. Employing the k-means machine learning algorithm, I developed two models—one for the original data and another for the categorical data. To ensure optimal cluster identification, an elbow test was conducted. Despite this, I chose to utilize 5 clusters to align with Caleb's cluster data, given the inherent variability in cluster assignments by the k-means algorithm. The resulting models were saved as files for future reference.





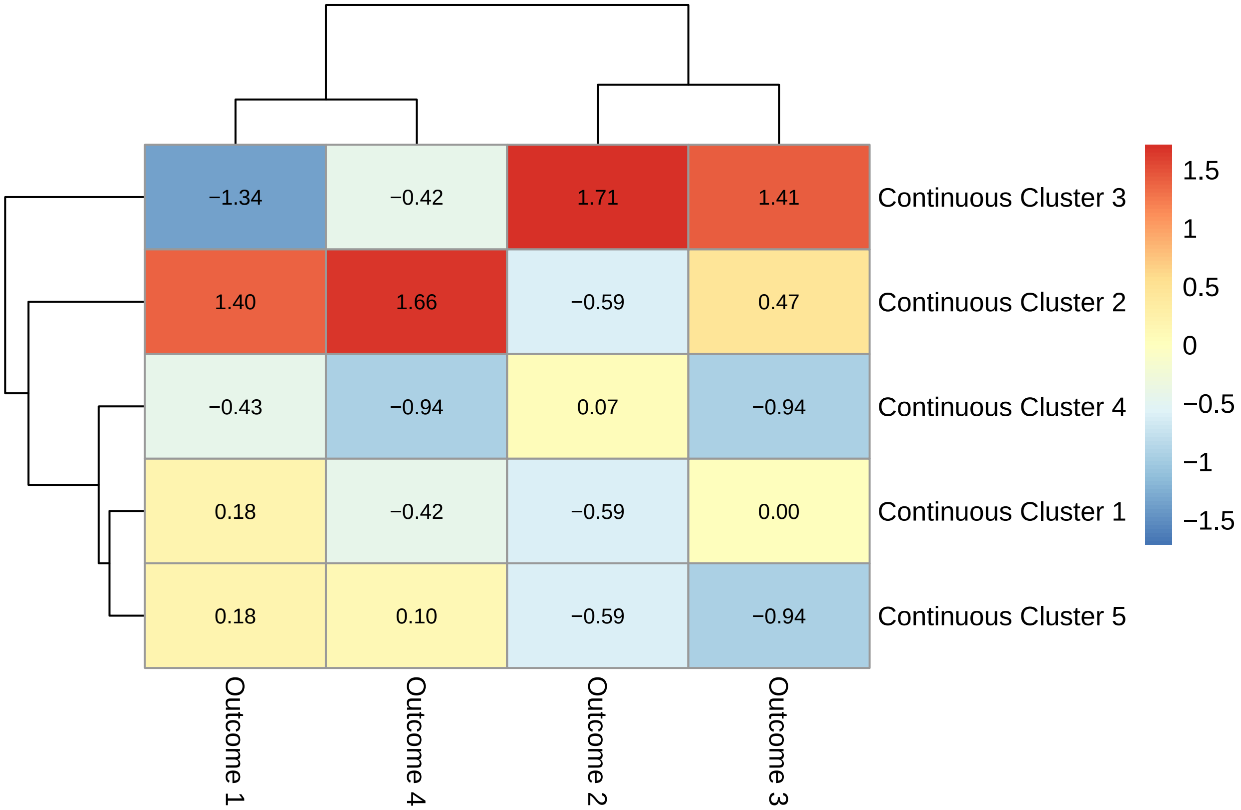
Experimental Results

Utilizing the original and categorical clusters derived from the preceding analyses, a heatmap was generated by constructing a correlation matrix between these two groups of clusters. The heatmap visually represents the degree of correlation between pairs of variables, employing a numeric scale reflected in the color intensity within each cell. Larger numerical values indicate a higher correlation, while smaller values signify a lower correlation. This visual representation facilitates the interpretation of relationships and associations between variables within the context of both original and categorical cluster groupings.

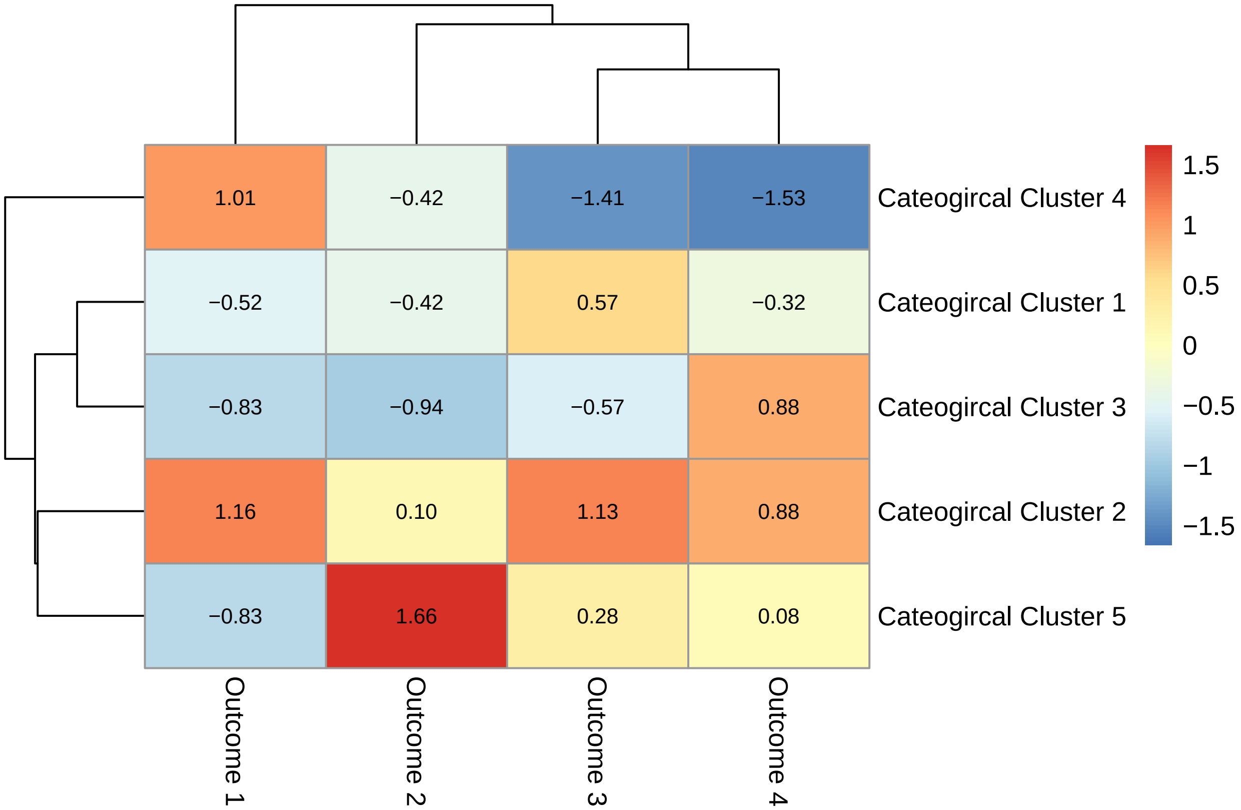


Furthermore, I generated heatmaps by constructing correlation matrices between each outcome and the two groups of clusters. These visualizations aim to illuminate the likelihood of specific outcomes associated with each cluster. This classification approach provides insights into the characteristics and patterns associated with different clusters, offering a comprehensive understanding of the potential outcomes and their correlations with specific shot attributes.

Original Clusters vs Outcomes Heatmap



Categorical Clusters vs Outcomes Heatmap



Upon examining the correlation matrix between the original and categorical clusters, it is observed that they exhibit robust correlations. Ideally, a high correlation is desired between each original cluster and precisely one categorical cluster. Original Cluster 1 demonstrates a notable correlation with Categorical Cluster 1, while all other correlations are slightly negative. This pattern suggests a strong similarity between shots in these two clusters. Original Cluster 2 exhibits high correlations with Categorical Clusters 2 and 4, and a moderate correlation with Categorical Cluster 3. For Original Cluster 3, a high correlation is observed with Categorical Cluster 5. However, Original Clusters 4 and 5 do not exhibit strong correlations with categorical clusters, indicating dissimilarity. In summary, with the exception of Original Clusters 2 and 4, the remaining original clusters display a one-to-one correlation with their corresponding categorical clusters.

For each cluster for discretized data, the mode discrete value for each feature is shown in the "Discrete Mode Values" table.

Table 2: Discrete Model Values

cat_cluster	outcomes.goal	puckSpeedCategory	puckAngleCategory	puckDistCategory	posTimeCategory	goalieDistCategory	shooterSpeedCategory	goalieAngleCategory	defDistCategory	defAngleCategory
1	3	3	1	2	2	2	2	1	1	3
2	3	2	1	2	1	2	1	1	1	3
3	3	2	2	2	1	3	1	2	2	1
4	1	2	2	2	1	2	1	2	1	1
5	3	2	2	1	1	1	1	2	1	1

Describing the characteristics of each cluster for continuous variables provides insights into distinct shot patterns. Shots in Cluster 1 are taken from a considerable distance from the goal, exhibiting high puck speed and possession time, indicating long and calculated shots. Cluster 2 comprises shots taken from an exceptionally far distance with a slow pace. Shots in Cluster 3 are characterized by intense pressure from the goalie and defenders. Cluster 4 represents mid-range shots with a moderate puck speed. Finally, shots in Cluster 5 are exceptionally far and fast, suggesting shots with both considerable distance and speed.

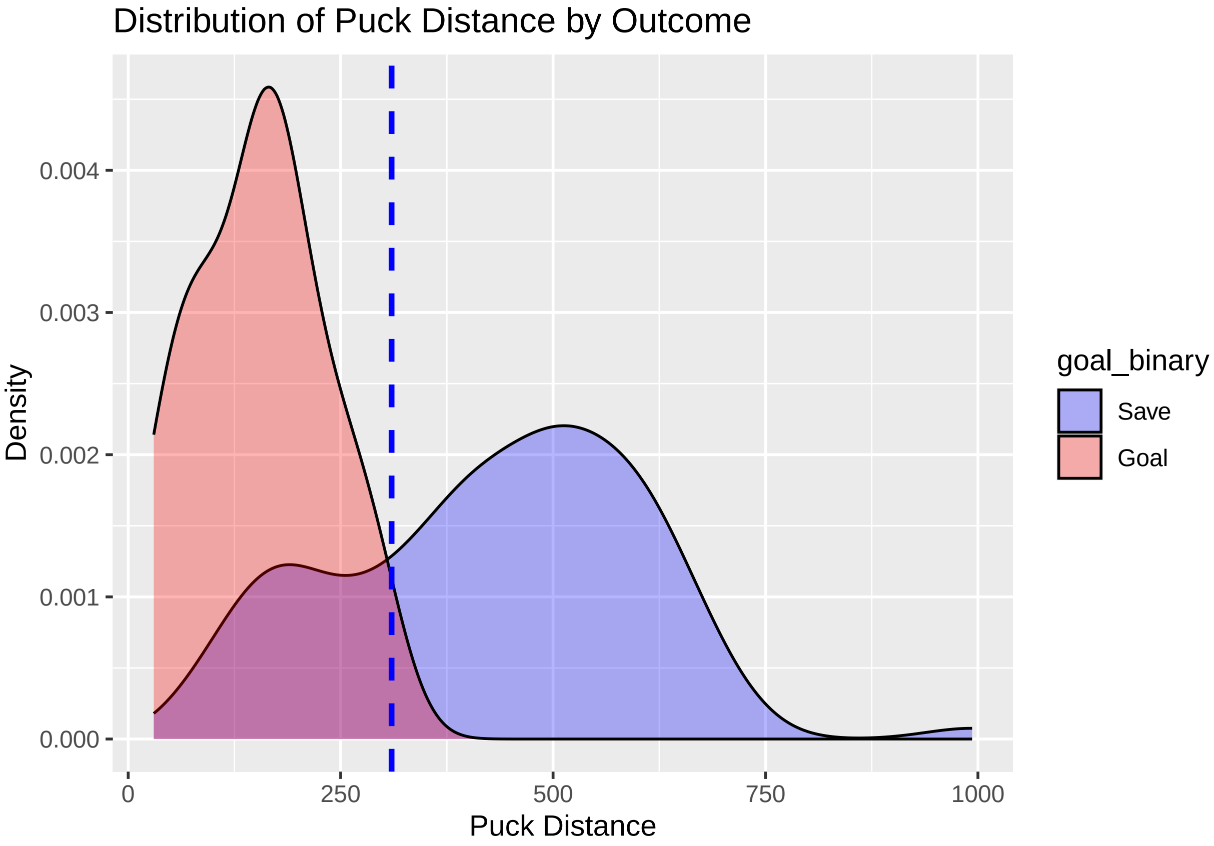
For each cluster for continuous data, the mean value for each feature is shown in "Continuous Mean Values" table.

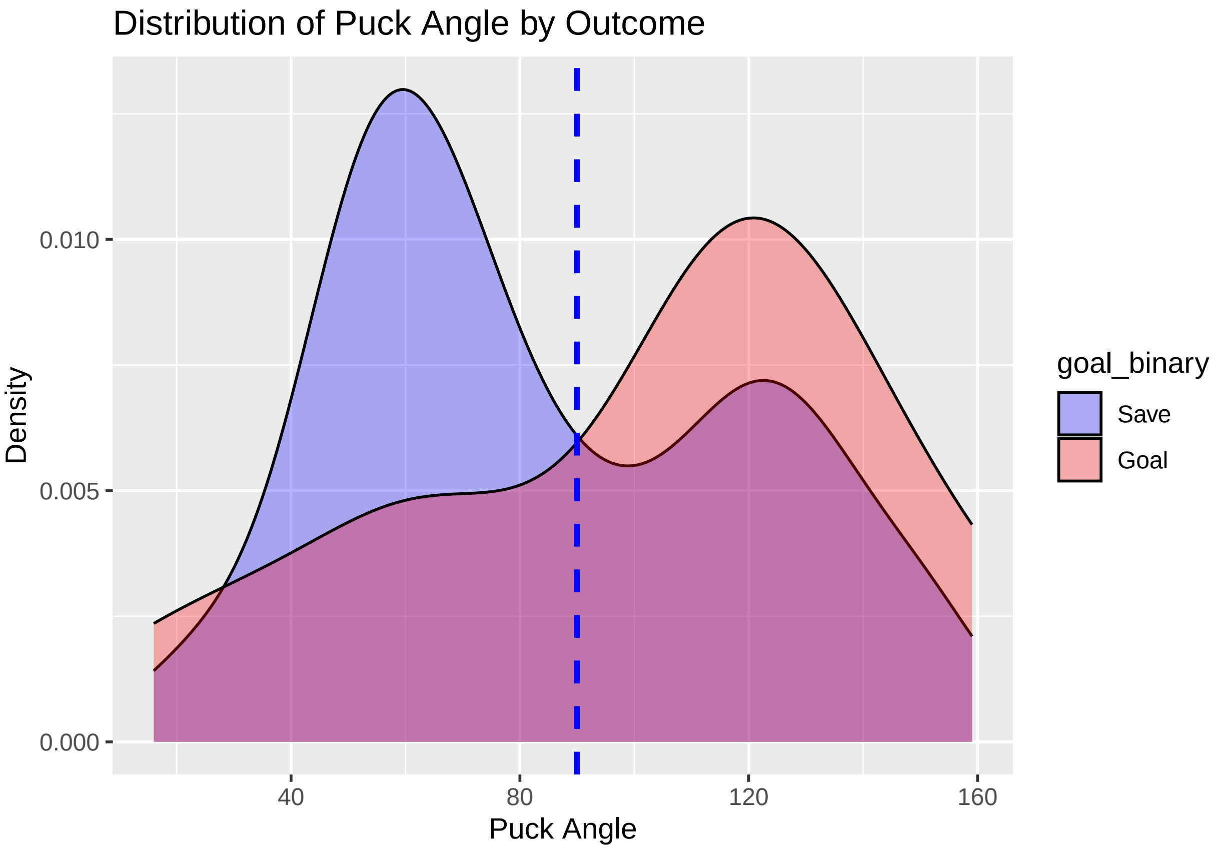
Table 3: Continuous Mean Values

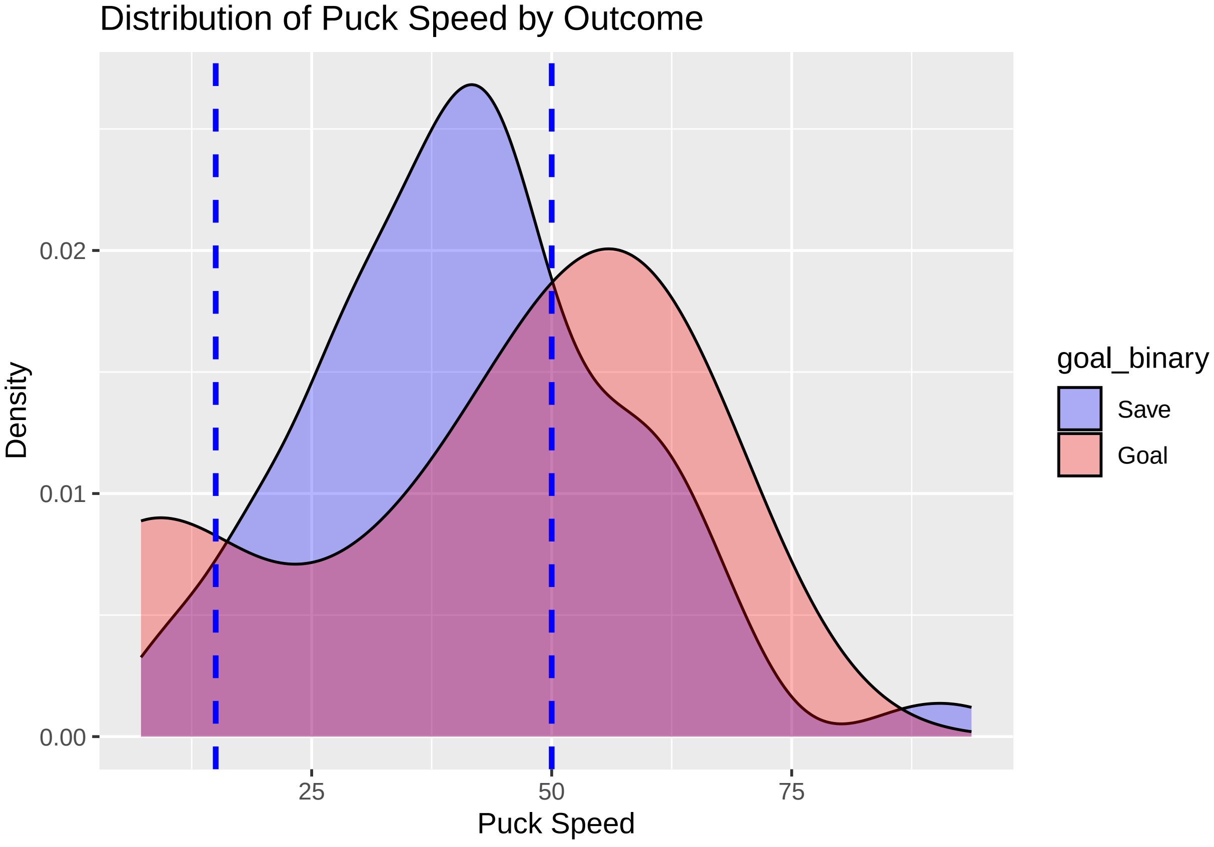
Group.1	puckDist	puckAngle	puckSpeed	shooterSpeed	goalieDist	goalieAngle	posTime	defDist	defAngle
1	429.0734	85.08309	43.24438	18.63413	74.60591	78.96551	42.78947	120.64397	116.34889
2	577.3543	82.26851	38.55734	12.54376	83.53557	75.38863	30.60714	161.56247	94.38426
3	147.5481	77.51730	38.48058	17.18447	45.33600	63.08071	20.66667	94.58971	102.01265
4	290.6015	98.85316	42.68814	18.39713	65.12851	96.68207	27.18750	159.86220	87.33023
5	549.3437	90.63352	47.81298	14.04467	75.97900	80.39762	25.44444	446.61064	92.61817

Describing the characteristics of each cluster for discretized variables reveals distinct patterns. Shots in Cluster 1 are characterized by a fast puck speed, a long possession time, and a rapid shooter speed. Cluster 2 encompasses mid-range shots with moderate speed, resembling Cluster 4, but predominantly taken from the right side. Shots in Cluster 3 are notable for being taken from a considerable distance from the goal and defender. Cluster 4 shares similarities with shots in Cluster 2 but is predominantly taken from the left side. Finally, shots in Cluster 5 represent high-pressure situations, taken in close proximity to the goalie and defender.

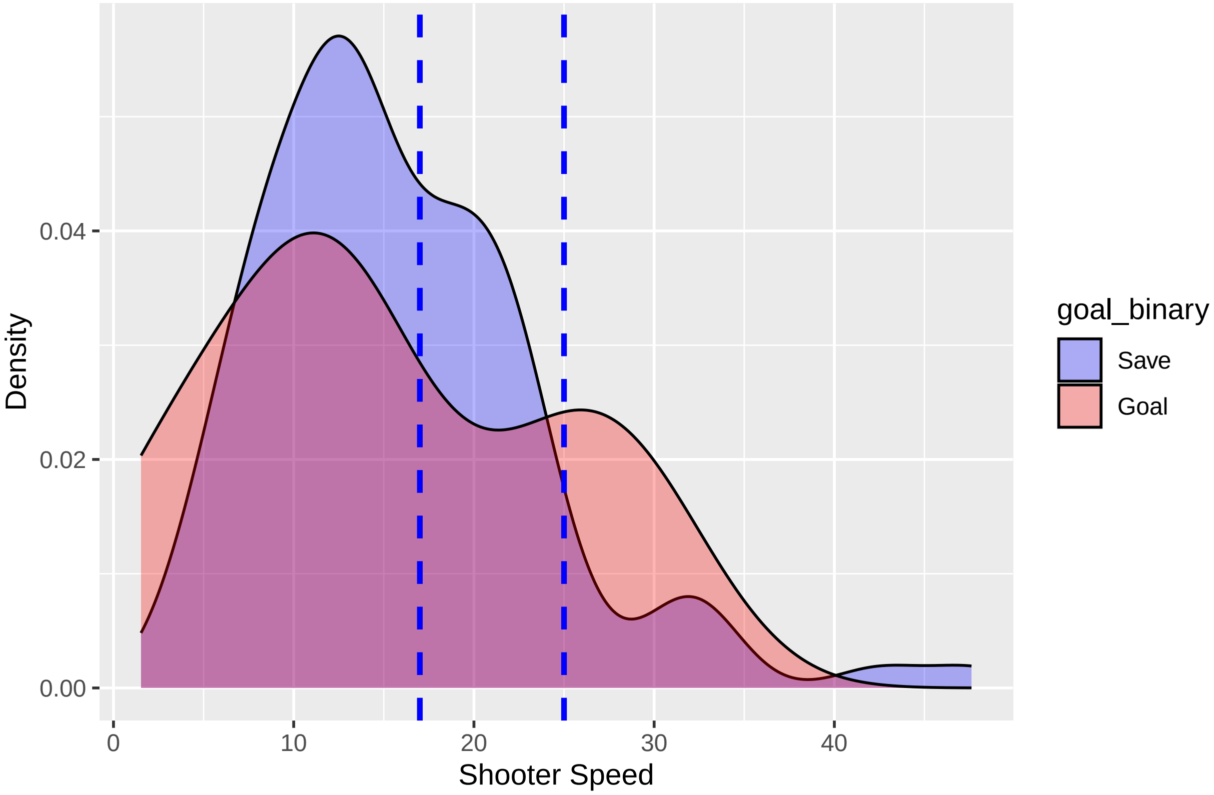
```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```


Distribution of Puck Distance by Outcome
Density
0.004
0.003
0.002
0.001
0.000
0
250
500
750
1000
Puck Distance
goal_binary
Save
Goal

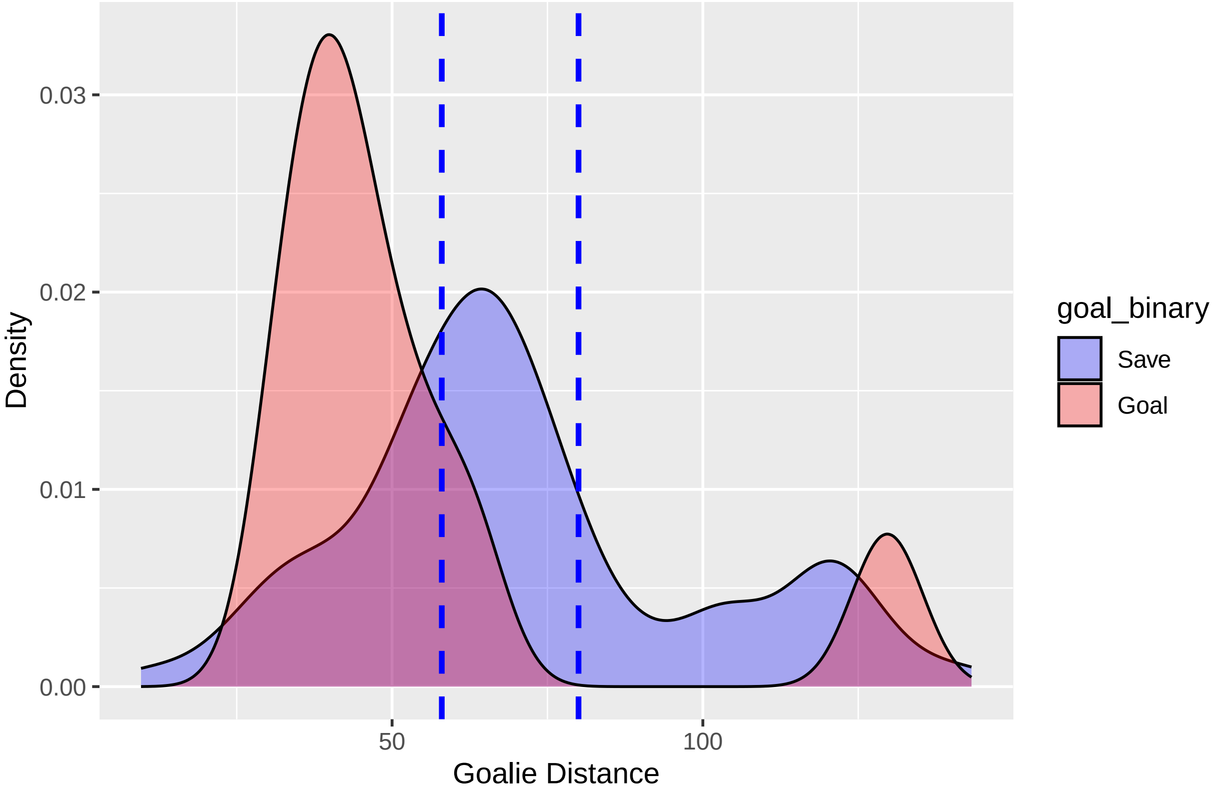

Distribution of Puck Angle by Outcome
Density
0.010
0.005
0.000
40
80
120
160
Puck Angle
goal_binary
Save
Goal

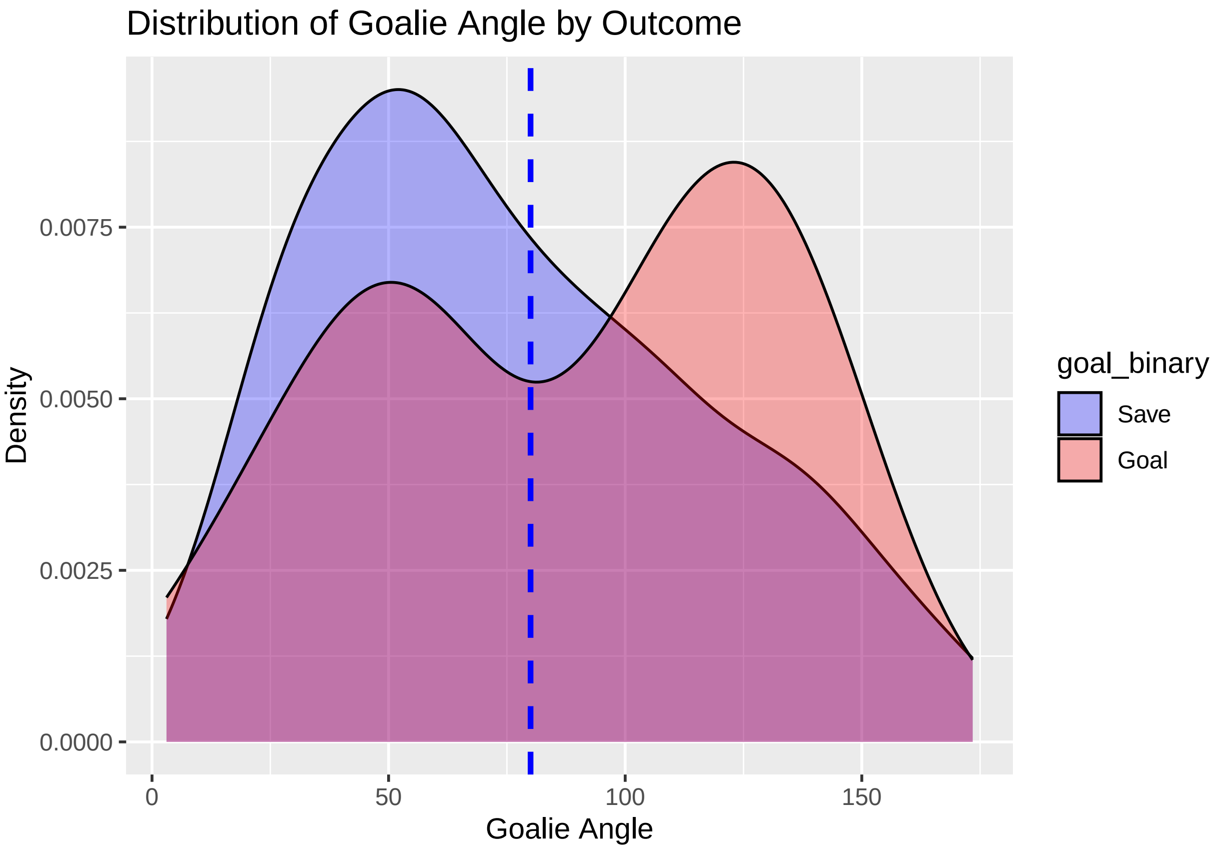

Distribution of Puck Speed by Outcome
Density
0.02
0.01
0.00
25
50
75
Puck Speed
goal_binary
Save
Goal

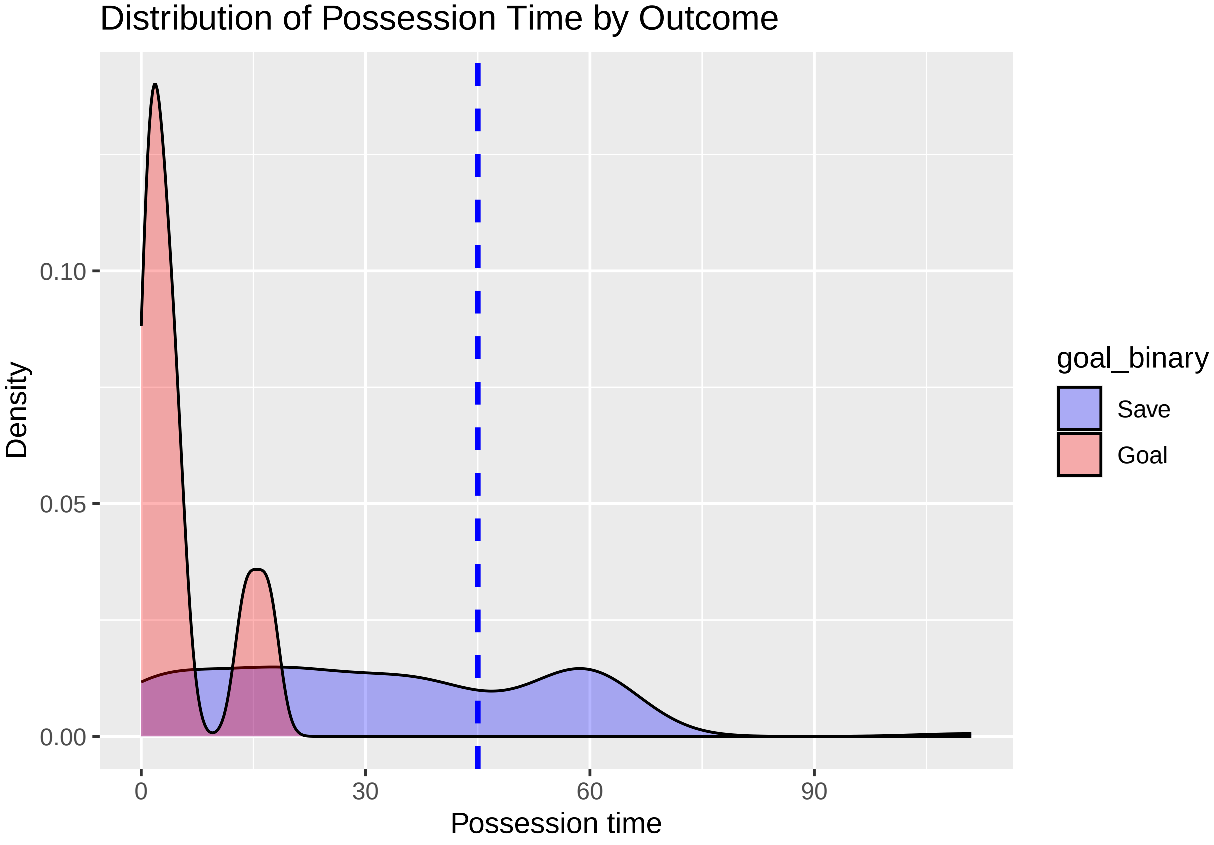
Distribution of Shooter Speed by Outcome

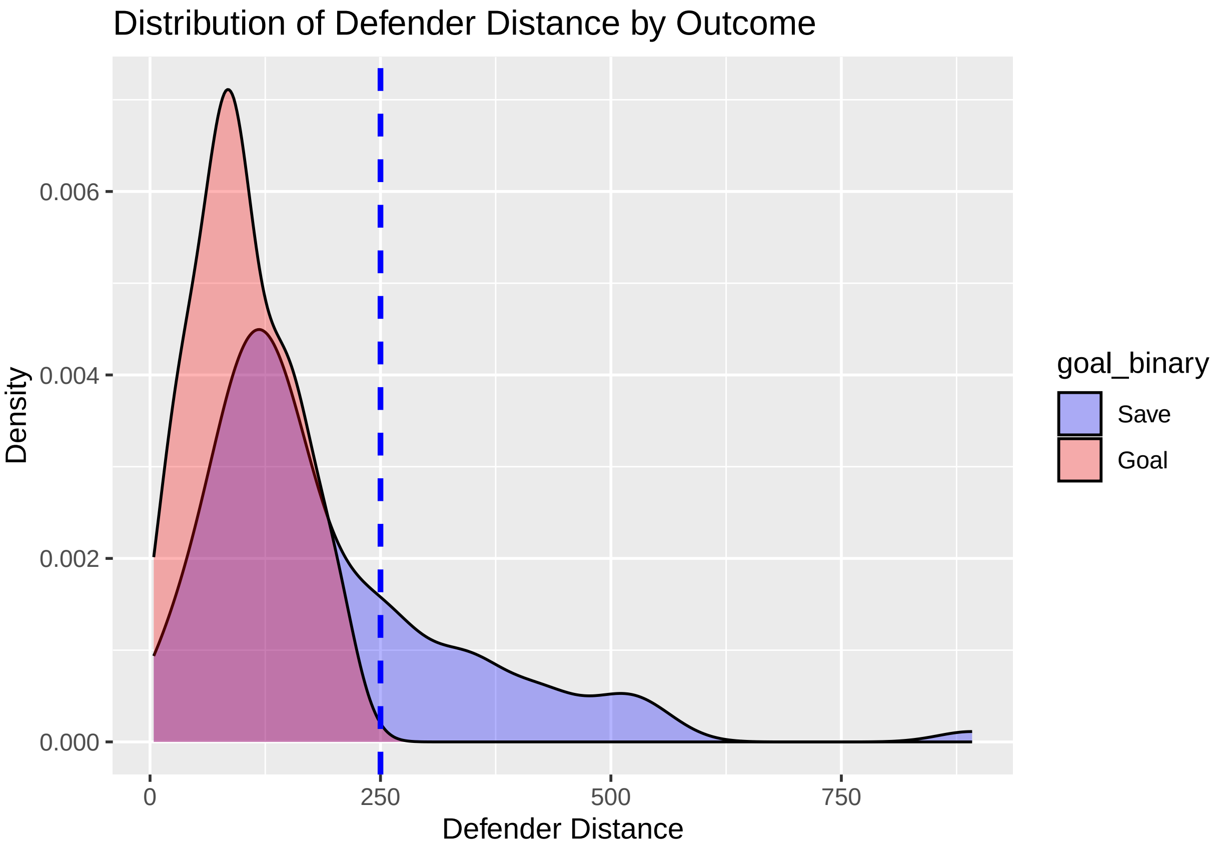


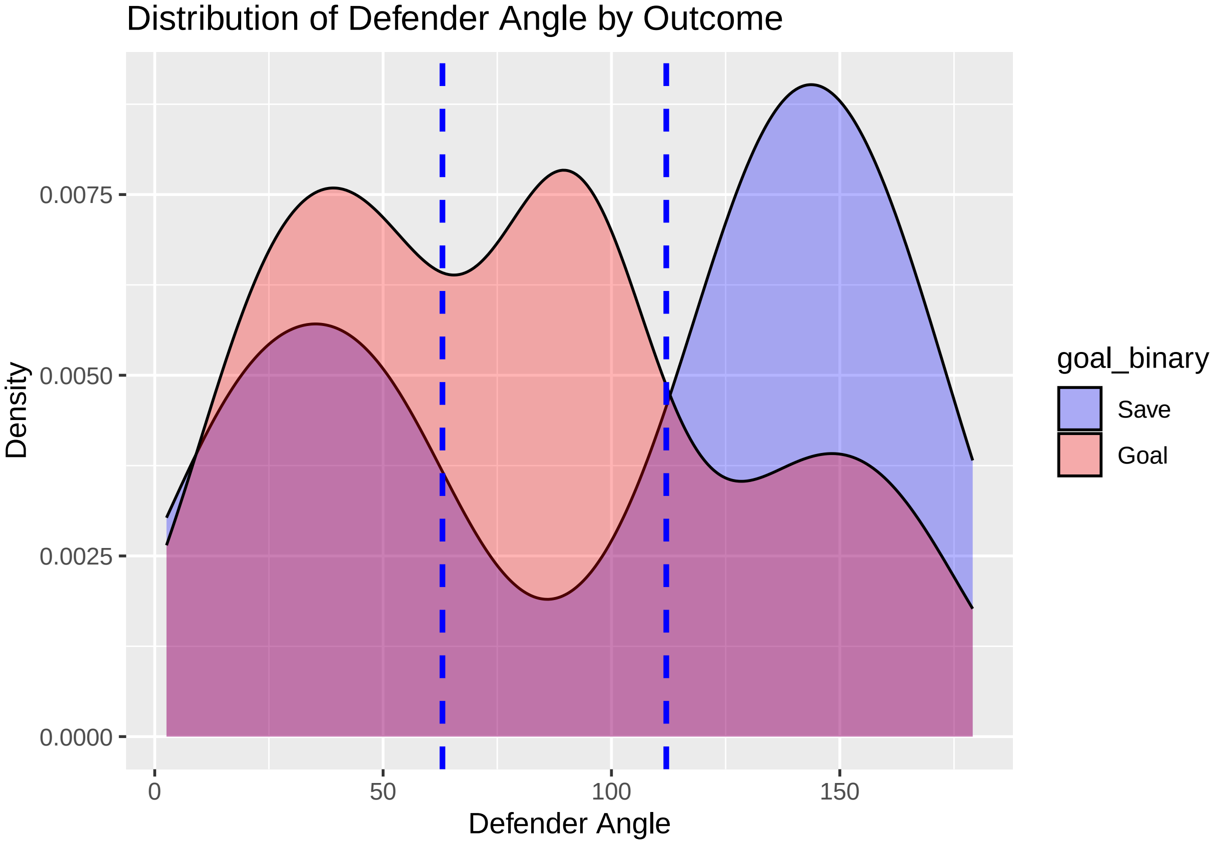
Distribution of Goalie Distance by Outcome




Distribution of Goalie Angle by Outcome
Density
0.0075
0.0050
0.0025
0.0000
0
50
100
150
Goalie Angle
goal_binary
Save
Goal


Distribution of Possession Time by Outcome
Density
0.10
0.05
0.00
0
30
60
90
Possession time
goal_binary
Save
Goal


Distribution of Defender Distance by Outcome


Distribution of Defender Angle by Outcome

Lastly, the density plots above were used to find the cutoffs for discretization. Marks were added where the cutoffs are.

Discussion of Results and Key Findings

In exploring the dataset, it became evident that transforming continuous variables into categorical ones adds a layer of granularity to the analysis. While continuous variables provide a single value for optimal shots, categorizing them allows for a more nuanced understanding of the combinations of features leading to specific outcomes. Employing the k-means clustering method on both categorical and continuous variables revealed discernible patterns in shots. The correlation matrix highlighted consistent one-to-one correlations between the two groups of clusters. This consistency suggests that the observed shot patterns are not arbitrary but rather a recurring phenomenon in hockey games. The comprehensive analysis of both categorical and continuous variables provides valuable insights into the dynamics of successful shots.

The insights gained from this analysis hold significant implications for hockey strategy. By delving into shot clusters using categorical variables, players can derive optimal shot combinations tailored to specific game scenarios. This analytical approach not only answers nuanced questions but also unveils patterns that may remain obscure without data-driven exploration. For instance, it allows us to determine the ideal goalie angle category when facing a shot with a considerable puckDist. This knowledge empowers players and teams to make informed decisions, enhancing their strategic prowess on the ice.

Conclusions

The distinctive patterns observed in these clusters underscore their meaningful nature, dismissing the notion of mere chance. This clustering aligns with the prevalent trends in shot-taking scenarios. The application of hockey data analytics promises to revolutionize our comprehension of the sport, unveiling insights that are uniquely attainable through analytical methodologies. The process of data discretization proves pivotal in this context, as categorical variables offer a nuanced understanding beyond singular numerical values. By adopting a more scientific approach to shot selection, informed by the concealed insights within the data, hockey players and teams can elevate their strategic decision-making on the ice.

Directions for Future Investigation

With an increased volume of data, we can enhance categorization methodologies, possibly identifying additional clusters that capture diverse shot scenarios. The primary objective remains the same – to discern the optimal shot selection in situations where making a shot is imperative. This iterative process, fueled by more extensive data, promises to unveil deeper insights into the nuanced dynamics of hockey shots, contributing to a more comprehensive understanding of strategic decision-making in the sport.

Bibliography

http://www.sthda.com/english/wiki/saving-data-into-r-data-format-rds-and-rdata#google_vignette
https://stackoverflow.com/questions/32684931/how-to-aggregate-data-in-r-with-mode-most-common-value-for-each-row

Files and Github Commits

Uploaded the final project notebook to FinalReport folder.

Contribution

Density plots other group mates to better understand the data set. The data set with categorized variables helped other group members to experiment with analytical methods.