# Hockey Analytics Research

Heatmap Analysis and App Developement

Ashley Woodson

2023-12-04

## Contents



## DAR Project and Group Members

- Project name: Hockey Analytics
- Project team members: Amy Enyenihi, Jeff Jung, Caleb Smith, Lieben Zhang

Required packages will be installed and loaded as necessary (code suppressed) if the notebook is run.

## Abstract

This project's purpose is to analyze data gleaned from motion capture images of RPI Women's Ice Hockey games. The use of heatmap visualization results in clusters which are supported through the other analysis methods performed by other group members. This analysis has found that puck speed, number of players involved in the play, and goalie positioning have the highest relevance to defensive success or failure.

## Introduction and Background

Our team's investigation into the motion capture data serves as a continuation of the work of the previous research team, who developed and applied the model which analyzed the video images to create our initial data set. Our aim is to evaluate these provided features: Number of offensive and defensive players, the puck speed, distance and angle, the goalie distance and angle from the goal, and the handedness, speed and possesion time of the shooter. We want to be able to provide this analysis in a format hockey players and coaches can use to both gain insight through provided conclusions, as well as explore the data to draw their own conclusions. This is achieved through the app linked in the appendix.

## Problems Tackled

The problem focused on in this research is to identify what feature or features has the greatest impact on goal outcome. In essence, what factors prevent goals?

The organization of analysis below aims to answer this question by introducing the data and processing the dataset. Then, the heatmap is created and analyzed. Throughout the semester many different graphing techniques were attempted, finding this to be the most successful. Conclusions and next steps follow.

## Data Description/Preparation

- The source of the dataset is Jeff's categorized data, linked in the Files section.
- The dataset was then additionally processed by me, to create a dataframe with 105 observations (shots) and 11 features.
- All of these variables are categorical (though some were originally continuous, like speed)
- I have additionally applied min/max scaling to the dataset by column.

The data file is to be read in, have the excess features removed, have the columns be min/max scaled, and converted to a matrix. When Jeff categorized the data, he left the continuous variables in the dataframe, so since they did not need to be included in the heatmap, I removed them from my dataframe. The categorical variables were chosen over the combination of continuous and categorical variables for consistency of analysis. the heatmaps consistently took the data type (categorical or continuous) as the primary level of feature importance, which is not useful for feature analysis. Since categorical variables like handedness cannot be made continuous, the use of all categorical variables was deemed optimal for equal weighting the variables. The columns are min/max scaled, something which was decided upon through multiple analyses. It was clear that the min/max scaling resulted in the most readable and informative visualization, compared to other normalization methods. Lastly, conversion to a matrix is necessary for the use of the pheatmap function later on.

```
#read in the data file. THis line assumes it has been placed in the same directory you are currently in
shots_cat.df <- read_rds("categorized_shots_stats_goal.df.Rds")

#remove the continuous and outcome features
features <- subset(shots_cat.df, select = -puckDist)
features <- subset(features, select = -puckAngle)
features <- subset(features, select = -puckSpeed)
features <- subset(features, select = -shooterSpeed)
features <- subset(features, select = -goalieDist)
features <- subset(features, select = -posTime)
features <- subset(features, select = -goalieAngle)
features <- subset(features, select = -closestDef)
features <- subset(features, select = -defDist)
features <- subset(features, select = -defAngle)
features <- subset(features, select = -shotOutcome)
```

```
features <- subset(features, select = -outcomes.goal)

#make the goal line
goal <- subset(shots_cat.df , select = outcomes.goal)
goal$outcomes.goal <- as.numeric(goal$outcomes.goal)

#normalize data using custom function
pmmScale <- as.data.frame(lapply(features, minMax))

#convert to a matrix in order to input to heatmap function
pmmScale <- as.matrix(pmmScale, rownames.force = TRUE)
```

Infographic: Viewing dataframe statistics. (the heatmap is also an infographic but belongs in analysis)

View the head of the dataframe.

```
##   NumOffense NumDefense rightHanded puckSpeedCategory puckAngleCategory
## 1        0.5  0.3333333           0               1.0               0.0
## 2        0.5  0.3333333           0               0.5               0.5
## 3        0.0  0.3333333           0               0.5               0.0
## 4        0.0  0.3333333           0               0.5               1.0
## 5        0.5  0.3333333           1               1.0               0.0
## 6        1.0  0.6666667           0               1.0               0.5
##   puckDistCategory posTimeCategory goalieDist_qCategory shooterSpeed_qCategory
## 1              1.0             0.0                  1.0                    0.5
## 2              0.5             0.0                  0.0                    0.0
## 3              0.5             0.0                  0.5                    0.5
## 4              0.5             1.0                  0.0                    0.5
## 5              0.0             1.0                  1.0                    1.0
## 6              1.0             0.5                  0.5                    0.0
##   goalieAngleCategory defDistCategory
## 1                 0.0             1.0
## 2                 0.5             0.0
## 3                 0.0             1.0
## 4                 1.0             0.0
## 5                 0.5             0.5
## 6                 0.5             0.5
```

View the summary of the data.

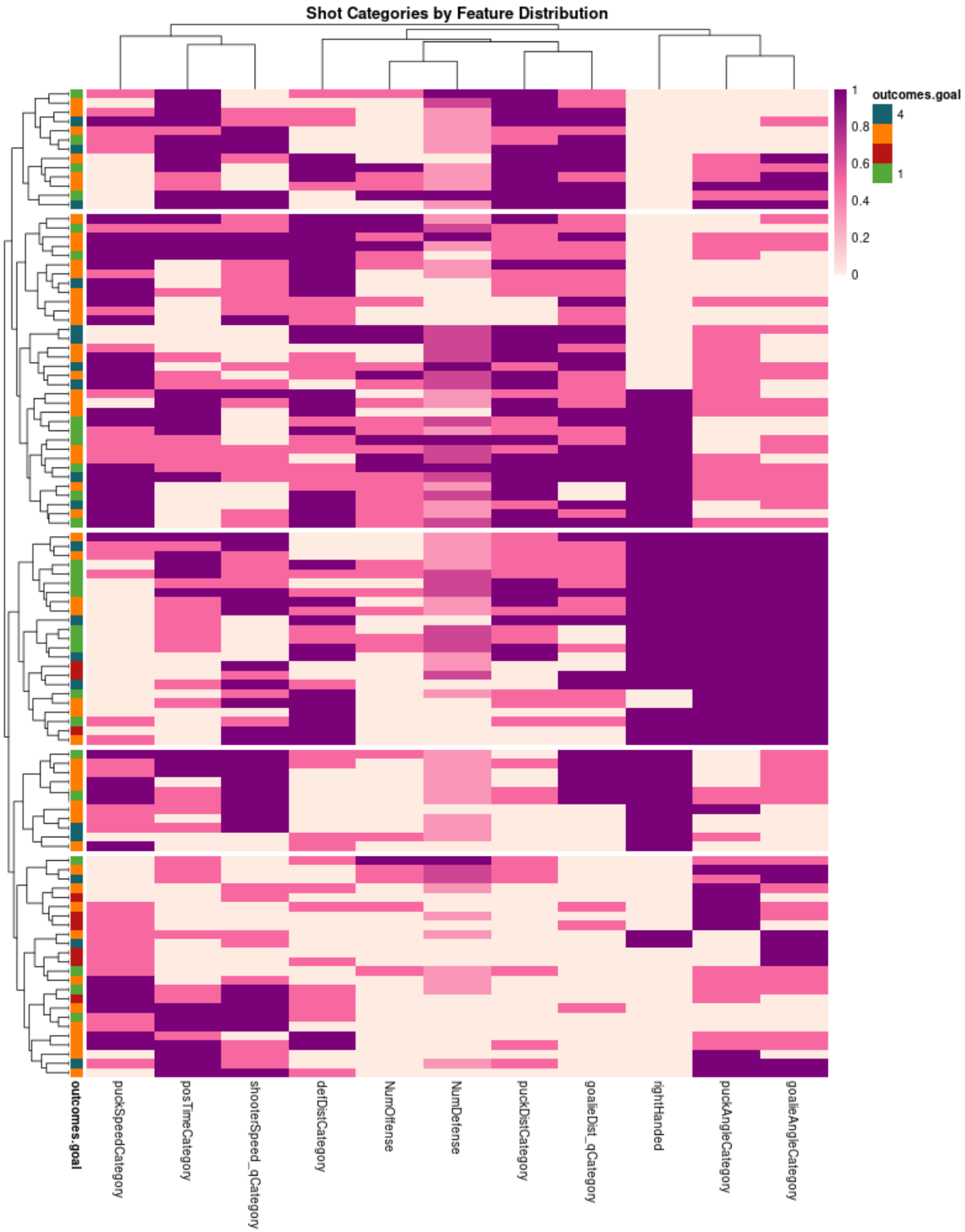
```
##    NumOffense       NumDefense      rightHanded     puckSpeedCategory
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0
##  Median :0.0000   Median :0.3333   Median :0.0000   Median :0.5
##  Mean   :0.2619   Mean   :0.3524   Mean   :0.4667   Mean   :0.5
##  3rd Qu.:0.5000   3rd Qu.:0.6667   3rd Qu.:1.0000   3rd Qu.:1.0
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0
##  puckAngleCategory puckDistCategory posTimeCategory  goalieDist_qCategory
##  Min.   :0.0       Min.   :0.0      Min.   :0.0000   Min.   :0.0
##  1st Qu.:0.0       1st Qu.:0.0      1st Qu.:0.0000   1st Qu.:0.0
##  Median :0.5       Median :0.5      Median :0.5000   Median :0.5
##  Mean   :0.5       Mean   :0.5      Mean   :0.4905   Mean   :0.5
##  3rd Qu.:1.0       3rd Qu.:1.0      3rd Qu.:1.0000   3rd Qu.:1.0
##  Max.   :1.0       Max.   :1.0      Max.   :1.0000   Max.   :1.0
##  shooterSpeed_qCategory goalieAngleCategory defDistCategory
##  Min.   :0.0            Min.   :0.0         Min.   :0.0
```

```
##  1st Qu.:0.0           1st Qu.:0.0         1st Qu.:0.0
##  Median :0.5           Median :0.5         Median :0.5
##  Mean   :0.5           Mean   :0.5         Mean   :0.5
##  3rd Qu.:1.0           3rd Qu.:1.0         3rd Qu.:1.0
##  Max.   :1.0           Max.   :1.0         Max.   :1.0
```

## Data Analytics Methods

The analysis method I chose was to utilize heatmaps. Throughout the semester I applied many different heatmap functions to both the continuous and discrete data, including heatmap2, heatmaply, and pheatmap. these use slightly different dendrogram algorithms, and thus produced slightly different graphs, but ultimately the same results once analyzed. As a result, the algorithm chosen was the one which created the most readable visualization. This was pheatmap.

This was then employed with the min/max scaling to create the heatmap below:


Shot Categories by Feature Distribution
outcomes.goal
4
1
1
0.8
0.6
0.4
0.2
0
outcomes.goal
puckSpeedCategory
posTimeCategory
shooterSpeed_qCategory
defDistCategory
NumOffense
NumDefense
puckDistCategory
goalieDist_qCategory
rightHanded
puckAngleCategory
goalieAngleCategory

## Discussion of Results and Key Findings

Analysis of the clusters above has resulted in the following names: 1. Too far, waited too long This cluster contains shots that were too far from the goal, and were also characterized by a high possetion time. This may be the result of defenders successfully denying the shooter clear access to the goal, forcing bad shots. 2. Far shots, crowded rink Too many players, both offensive and defensive, increase the odds of an interception before the puck reaches the goal. 3. Edge shots (some goals) With this category, all the shots were taken from the highest angle category for both the puck and goalie. However, all the successful goals were from close in, as opposed to many of the non-goals being from a further distance. Within this cluster we see how, although the angle is what clusters them, distance is still the defining factor of success. 4. Goalie face-off A well-aligned goalie is difficult to shoot around, resulting in a high proportion of saves. 5. Goalie out of place (most goals) When the goalie is not in the goal, or off to the opposite side of the shooter, then there is rarely someone to block the shot and most goals occur under this condition.

This is very significant because they line up very closely with the clusters developed by Caleb through his U-Map/k-means analysis. The increases confidence that these results are accurate. They also follow closely with the generally understood conventions of ice hockey, where when the goalie is not able to block the shot, being out of position, this is the best way to score. Long distance shots, too many players close to the goal, and a well-placed goalie are all known to cause difficulty in scoring.

## Conclusions

Conclusions that can be drawn lie in the cluster groupings, where we can see that, as per the above analysis, puck distance is the primary factor in predicting goal success. For RPI's defense, keeping shooters at a high distance guarantees that no goal will be scored. THe next most significant factor is the number of players involved, meaning the more players, the more chaotic teh rink becomes and the more interceptions occur. Finally, goalie misplacement accounts for the bigest indication of defensive faliure: a goal. Recommendations to make use of these results would be to provide our analysis to the hockey players so they can draw their own conclusions about the technical hockey causes of each of these scenarios, and how they can best adapt thier game play to utilize this knowledge.

## Directions for Future Investigation

A good direction to move this project in would primarily include increasing data. We have only 105 shots-on-goal on RPI's goalie, so having more instances of this would help us to increase accuracy. With only one third per game on offense, we lack significant data there; the results of this analysis vaaried too much because of the different goalies. More instances of RPI on offense would allow us to answer quesetions like, "How can RPI increase scoring success?" As opposed to the defensive-only questions, like we have answered in our collective work of "How can RPI prevent opposing goals?"

Ongoing data capture will be necessary as well, since RPI is a college the roster changes every year, and completely every four years, such that new goalies, defenders, and other players with different strengths and weaknesses will be on the ice, and may garner different results and recommendations for their particular skill set.

In addition to this, the application itself (reviewed below in the Appendix) could use more fine-tuning before it is ready to publish to our hockey players. For specifically my heatmap page, interactive text boxes would increase readability, and having more feedback from hockey players might allow for more comprehensive cluster descriptions.

A dynamic heatmap, which can display various clusters, or update depending on different datasets as future work is completed would also help.

# Bibliography

R packages most utilized for analysis and developement: pheatmap: https://www.rdocumentation.org/packages/pheatmap/versions/1.0.12/topics/pheatmap shinydashboard: https://rstudio.github.io/shinydashboard/structure.html#column-based-layout

# Files

Files referenced for reproducability:

- Jeff's Categorized Data
  - https://github.rpi.edu/DataINCITE/Hockey_Fall_2023/blob/main/StudentData/categorized_shots_stats_goal.df.Rds
  - This serves as the base input data, which could change in future implementations with additional motion-capture data from the new cameras and future games.
- Hockey Dashboard App
  - https://github.rpi.edu/DataINCITE/Hockey_Fall_2023/tree/dar-woodsa4/ShinyApps/HockeyDashboard
  - Explained below in Appendix, this links to the Hockey Dashboard App where my heatmap is displayed.
- This notebook
  - https://github.rpi.edu/DataINCITE/Hockey_Fall_2023/tree/dar-woodsa4/StudentNotebooks/Assignment07/dar_final_draft_woodsa4_01dec2023.pdf
  - https://github.rpi.edu/DataINCITE/Hockey_Fall_2023/tree/dar-woodsa4/StudentNotebooks/Assignment07/dar_final_draft_woodsa4_01dec2023.Rmd
  - These are the pdf and R notebook respectively, where the pdf has been submitted to gradescope, and the both on GitHub for easy access and reproducability.

# Appendix

I have created a tab in the final organized application known as "Hockey Dashboard", so i have included a link here. There was no additional content presented there so it did not seem relevant to include in the data analysis sections, but was additional work I completed towards our group goal. The app can be found here: https://github.rpi.edu/DataINCITE/Hockey_Fall_2023/tree/dar-woodsa4/ShinyApps/HockeyDashboard