DAR F23 Hockey Analytics

Hockey Analytics

Jeff Jung

2023-10-16

Contents

Weekly Work Summary

NOTE: Follow an outline format; use bullets to express individual points.

- RCS ID: jungj6
- Project Name: Hockey Analytics

Summary of work since last week

- Last week, I have successfully categorized continuous variables into categories, and this week, I did further analysis with categorical data
- Created bar plots to show the modes categorized variable with saves data
- Created heat map, which required converting all categorical data into binary variables
- Heat map effectively showed the correlations between different categories of each feature

NEW: Summary of github issues added and worked

N/A

Summary of github commits

- One commit of HW4 to dar-jungj6

List of presentations, papers, or other outputs

N/A

- List of references (if necessary)
- Indicate any use of group shared code base
- Indicate which parts of your described work were done by you or as part of joint efforts
- **Required:** Provide illustrating figures and/or tables

Personal Contribution

- Analysis on the modes of failure with categorization
- Application of bar plots in illustrating different modes of failure with categorization
- Application of heat map in illustration different modes of failure with categorization

Analysis: Modes of Failure

Question being asked

Since we do not have much data for goals, I have decided to do further analysis on saves. Finding the modes of failure would allow me to identify what features of shots led to saves.

Data Preparation

Data preparation for hockeyTrain is the same as I did for other assignments. Since I am specifically examining the modes of failure, I have separated 'saves' from hockeyTrain. Also, I have categorized continuous variables into 3 groups (high, med, low), so that it is easier to visualize.

```
# Load necessary library
library(dplyr)

# Calculate quantiles to divide the data into thirds
puckDist_q <- quantile(hockeyTrain$puckDist, probs = c(1/3, 2/3))
puckAngle_q <- quantile(hockeyTrain$puckAngle, probs = c(1/3, 2/3))
puckSpeed_q <- quantile(hockeyTrain$puckSpeed, probs = c(1/3, 2/3))
shooterSpeed_q <- quantile(hockeyTrain$shooterSpeed, probs = c(1/3, 2/3))
goalieDist_q <- quantile(hockeyTrain$goalieDist, probs = c(1/3, 2/3))
goalieAngle_q <- quantile(hockeyTrain$goalieAngle, probs = c(1/3, 2/3))
posTime_q <- quantile(hockeyTrain$posTime, probs = c(1/3, 2/3))

# Create three categories based on quantiles
hockeyTrain <- hockeyTrain %>%
  mutate(puckSpeedCategory = case_when(
    puckSpeed <= puckSpeed_q[1] ~ "Low",
    puckSpeed <= puckSpeed_q[2] ~ "Medium",
    TRUE ~ "High"
  ))

hockeyTrain <- hockeyTrain %>%
  mutate(puckDistCategory = case_when(
    puckDist <= puckDist_q[1] ~ "Low",
    puckDist <= puckDist_q[2] ~ "Medium",
    TRUE ~ "High"
  ))

hockeyTrain <- hockeyTrain %>%
  mutate(puckAngleCategory = case_when(
    puckAngle <= puckAngle_q[1] ~ "Low",
    puckAngle <= puckAngle_q[2] ~ "Medium",
    TRUE ~ "High"
  ))

hockeyTrain <- hockeyTrain %>%
  mutate(shooterSpeedCategory = case_when(
    shooterSpeed <= shooterSpeed_q[1] ~ "Low",
```

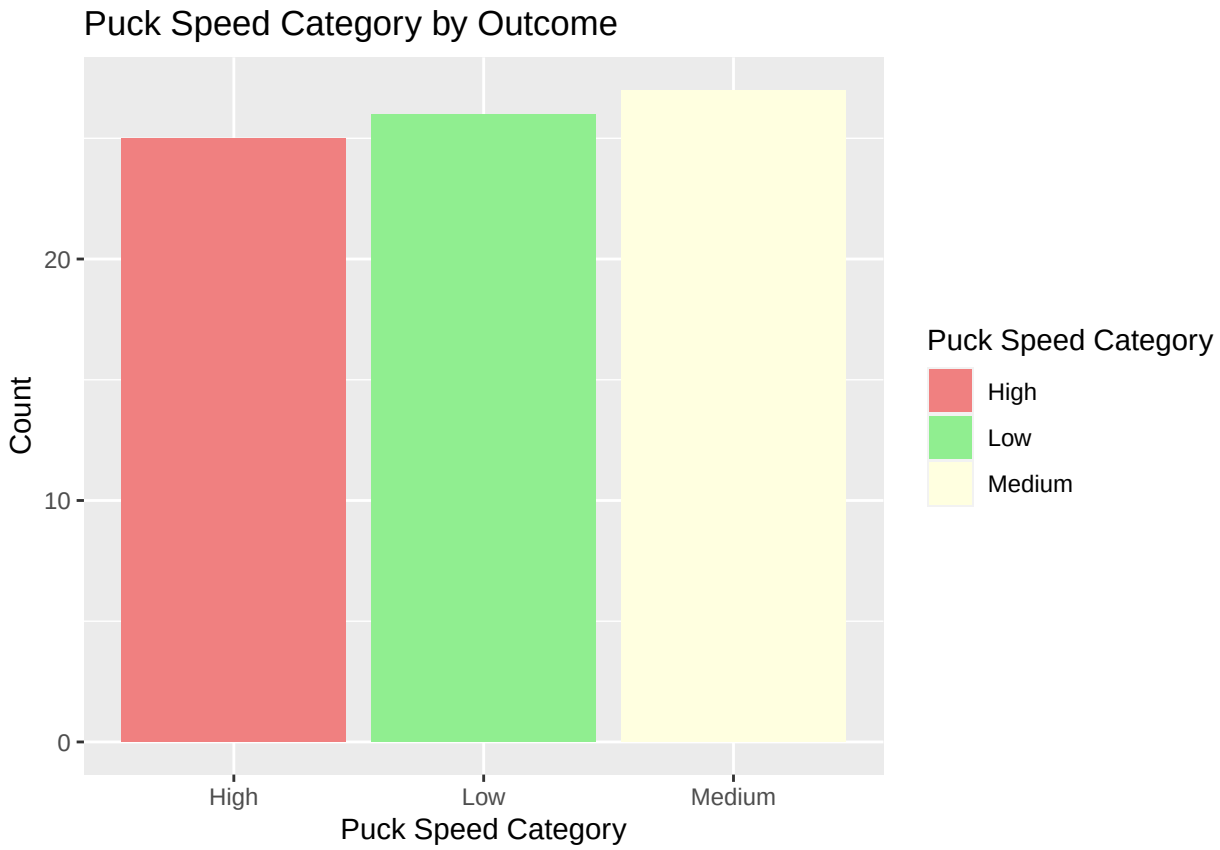
```
    shooterSpeed <= shooterSpeed_q[2] ~ "Medium",
    TRUE ~ "High"
  ))

hockeyTrain <- hockeyTrain %>%
  mutate(goalieDistCategory = case_when(
    goalieDist <= goalieDist_q[1] ~ "Low",
    goalieDist <= goalieDist_q[2] ~ "Medium",
    TRUE ~ "High"
  ))

hockeyTrain <- hockeyTrain %>%
  mutate(goalieAngleCategory = case_when(
    goalieAngle <= goalieAngle_q[1] ~ "Low",
    goalieAngle <= goalieAngle_q[2] ~ "Medium",
    TRUE ~ "High"
  ))

hockeyTrain <- hockeyTrain %>%
  mutate(posTimeCategory = case_when(
    posTime <= posTime_q[1] ~ "Low",
    posTime <= posTime_q[2] ~ "Medium",
    TRUE ~ "High"
  ))

# Subset the data for not goals (e.g., "Save")
saves_data <- hockeyTrain[hockeyTrain$outcomes.goal != "1", ]
```

Analysis: Methods and results

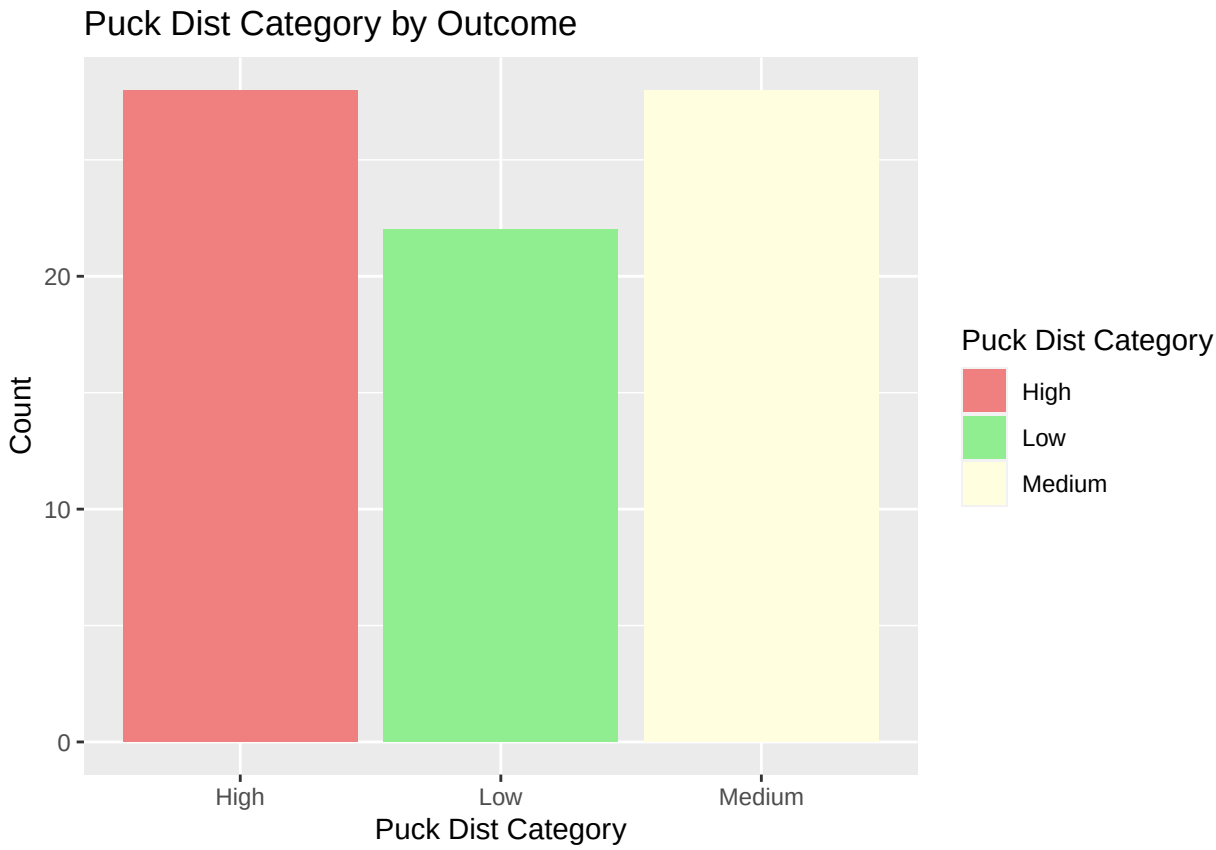
First, I have created a bar chart to show what the mode groups are for each variable. And then, to show how each feature is correlated with other features, I have created a heatmap.

```
# Include all analysis code, clearly commented
# If not possible, screen shots are acceptable.
# If your contributions included things that are not done in an R-notebook,
#   (e.g. researching, writing, and coding in Python), you still need to do
#   this status notebook in R.  Describe what you did here and put any products
#   that you created in github. If you are writing online documents (e.g. overleaf
#   or google docs), you can include links to the documents in this notebook
#   instead of actual text.

# Find modes for each variable
mode_puckSpeed <- names(sort(table(saves_data$puckSpeedCategory), decreasing = TRUE)[1])
mode_puckDist <- names(sort(table(saves_data$puckDistCategory), decreasing = TRUE)[1])
mode_puckAngle <- names(sort(table(saves_data$puckAngleCategory), decreasing = TRUE)[1])
mode_shooterSpeed <- names(sort(table(saves_data$shooterSpeedCategory), decreasing = TRUE)[1])
mode_goalieDist <- names(sort(table(saves_data$goalieDistCategory), decreasing = TRUE)[1])
mode_goalieAngle <- names(sort(table(saves_data$goalieAngleCategory), decreasing = TRUE)[1])
mode_posTime <- names(sort(table(saves_data$posTimeCategory), decreasing = TRUE)[1])

# Print the modes for each variable
cat("Mode for puckSpeedCategory:", mode_puckSpeed, "\n")
```

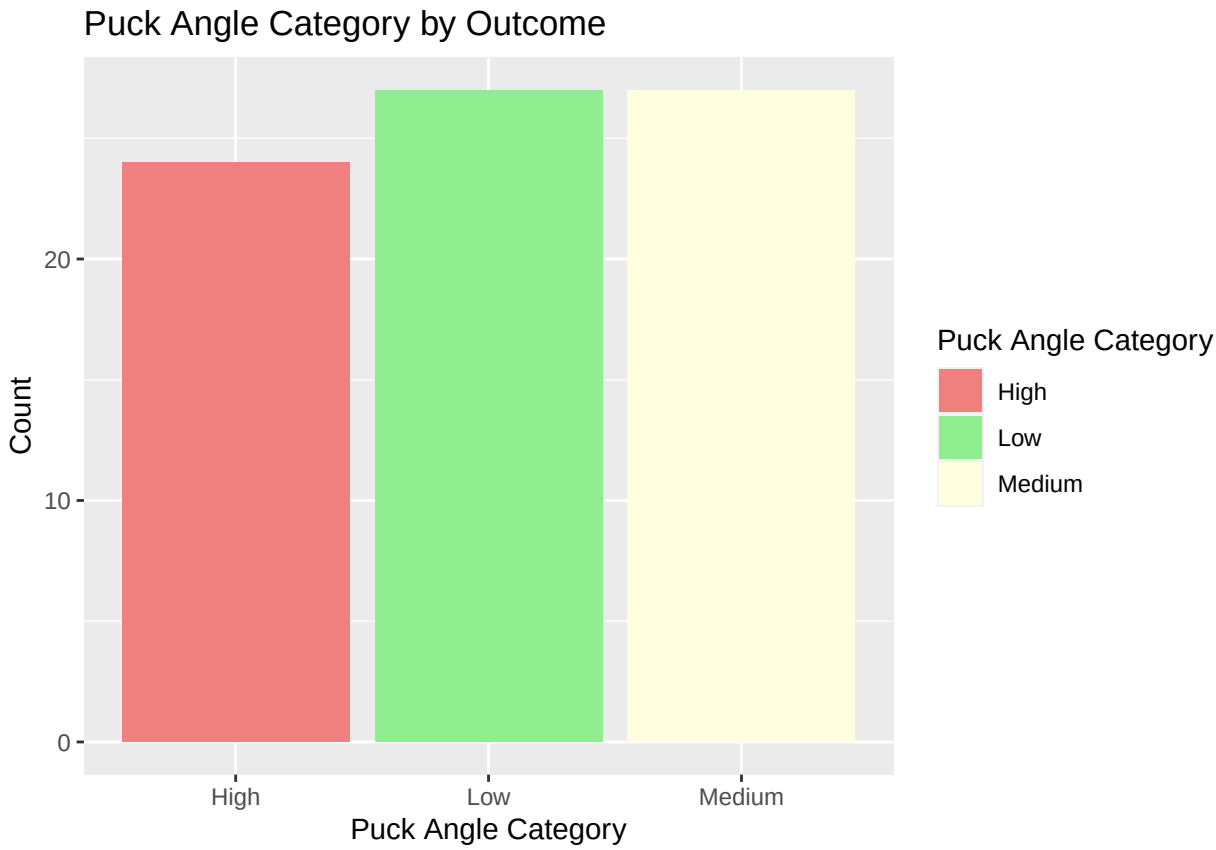
```
## Mode for puckSpeedCategory: Medium
cat("Mode for puckDistCategory:", mode_puckDist, "\n")

## Mode for puckDistCategory: High
cat("Mode for puckAngleCategory:", mode_puckAngle, "\n")

## Mode for puckAngleCategory: Low
cat("Mode for shooterSpeedCategory:", mode_shooterSpeed, "\n")

## Mode for shooterSpeedCategory: Medium
cat("Mode for goalieDistCategory:", mode_goalieDist, "\n")

## Mode for goalieDistCategory: High
cat("Mode for goalieAngleCategory:", mode_goalieAngle, "\n")

## Mode for goalieAngleCategory: Medium
cat("Mode for posTimeCategory:", mode_posTime, "\n")

## Mode for posTimeCategory: High
# Define custom colors for each category
category_colors <- c("Low" = "lightgreen", "Medium" = "lightyellow", "High" = "lightcoral")

# Create a bar plot
ggplot(saves_data, aes(x = puckSpeedCategory, fill = puckSpeedCategory)) +
  geom_bar(position = "dodge") +
  scale_fill_manual(values = category_colors) +  # Set custom colors
  labs(
    title = "Puck Speed Category by Outcome",
    x = "Puck Speed Category",
    y = "Count",
    fill = "Puck Speed Category"  # Legend title
  )
```



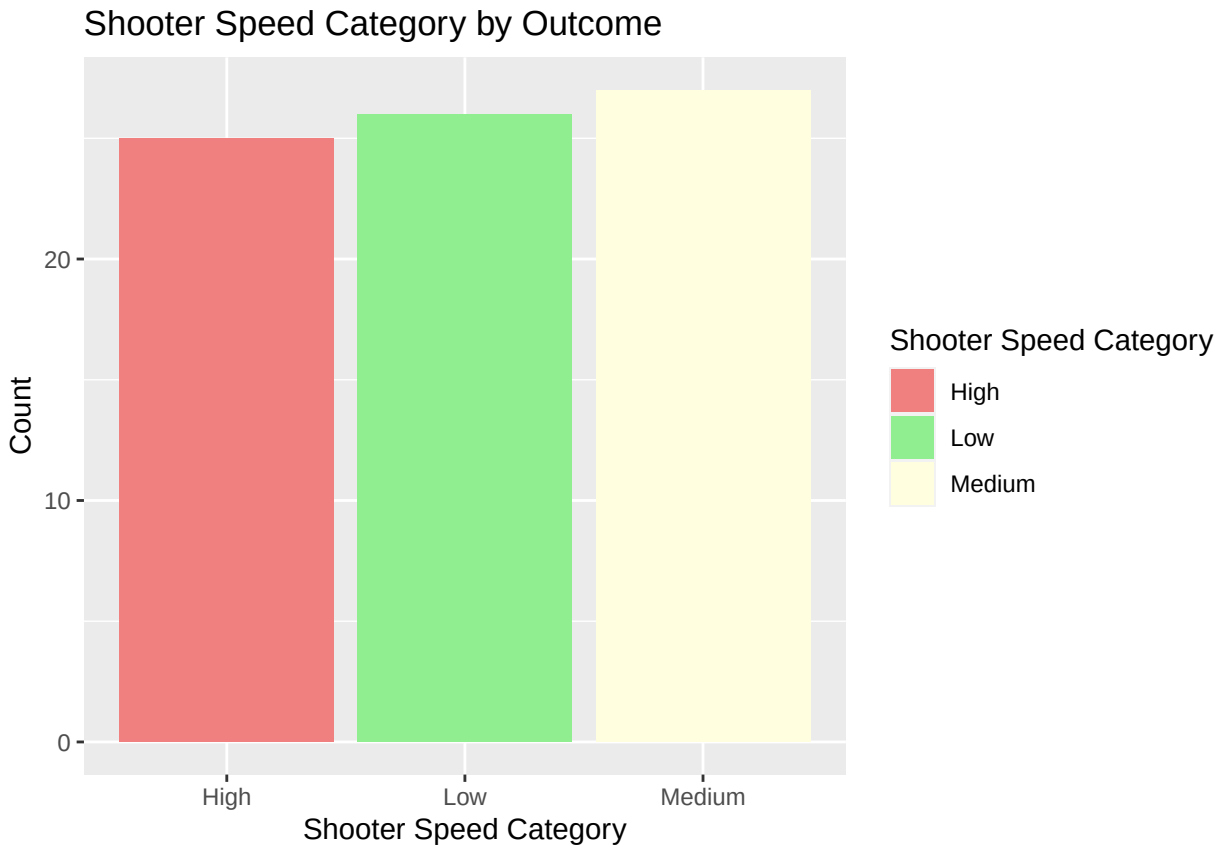

```
ggplot(saves_data, aes(x = puckDistCategory, fill = puckDistCategory)) +
  geom_bar(position = "dodge") +
  scale_fill_manual(values = category_colors) +  # Set custom colors
  labs(
    title = "Puck Dist Category by Outcome",
    x = "Puck Dist Category",
    y = "Count",
    fill = "Puck Dist Category"  # Legend title
  )
```



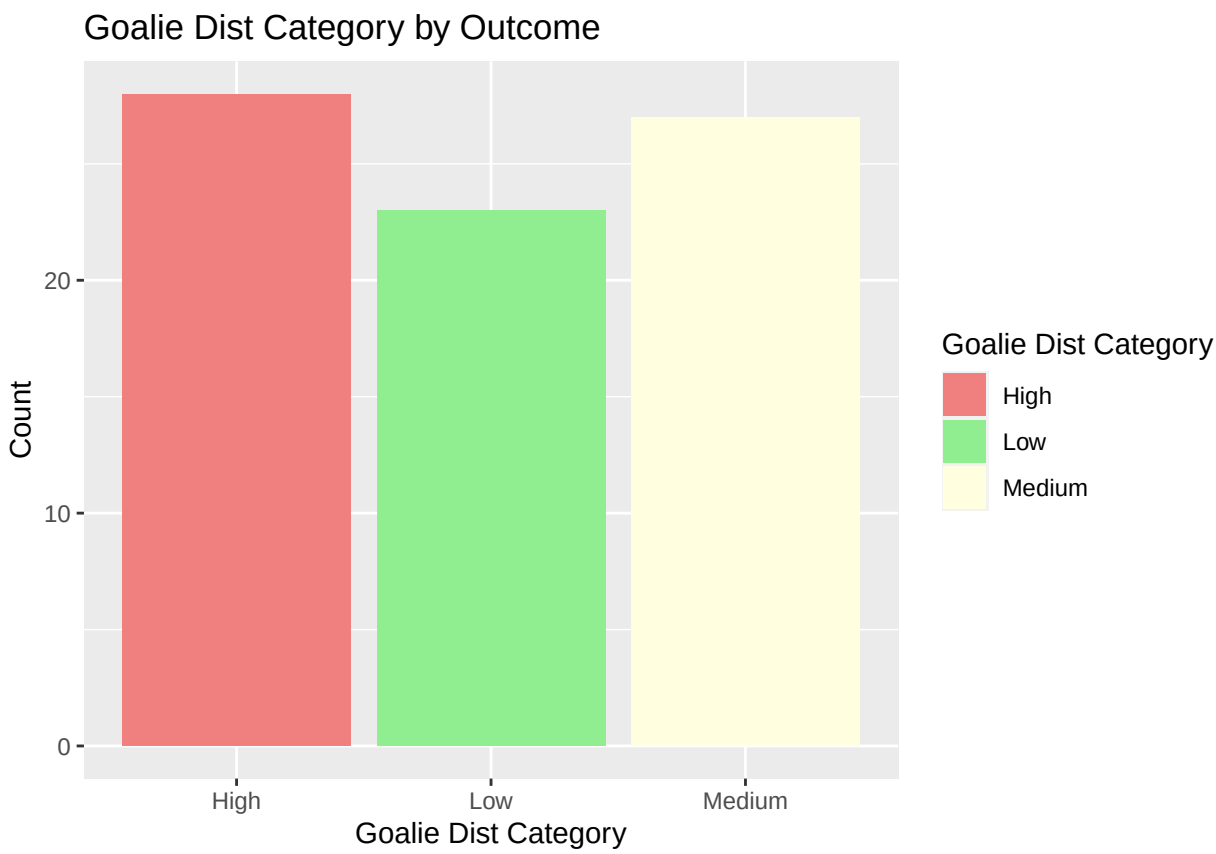

```
ggplot(saves_data, aes(x = puckAngleCategory, fill = puckAngleCategory)) +
  geom_bar(position = "dodge") +
  scale_fill_manual(values = category_colors) +  # Set custom colors
  labs(
    title = "Puck Angle Category by Outcome",
    x = "Puck Angle Category",
    y = "Count",
    fill = "Puck Angle Category"  # Legend title
  )
```



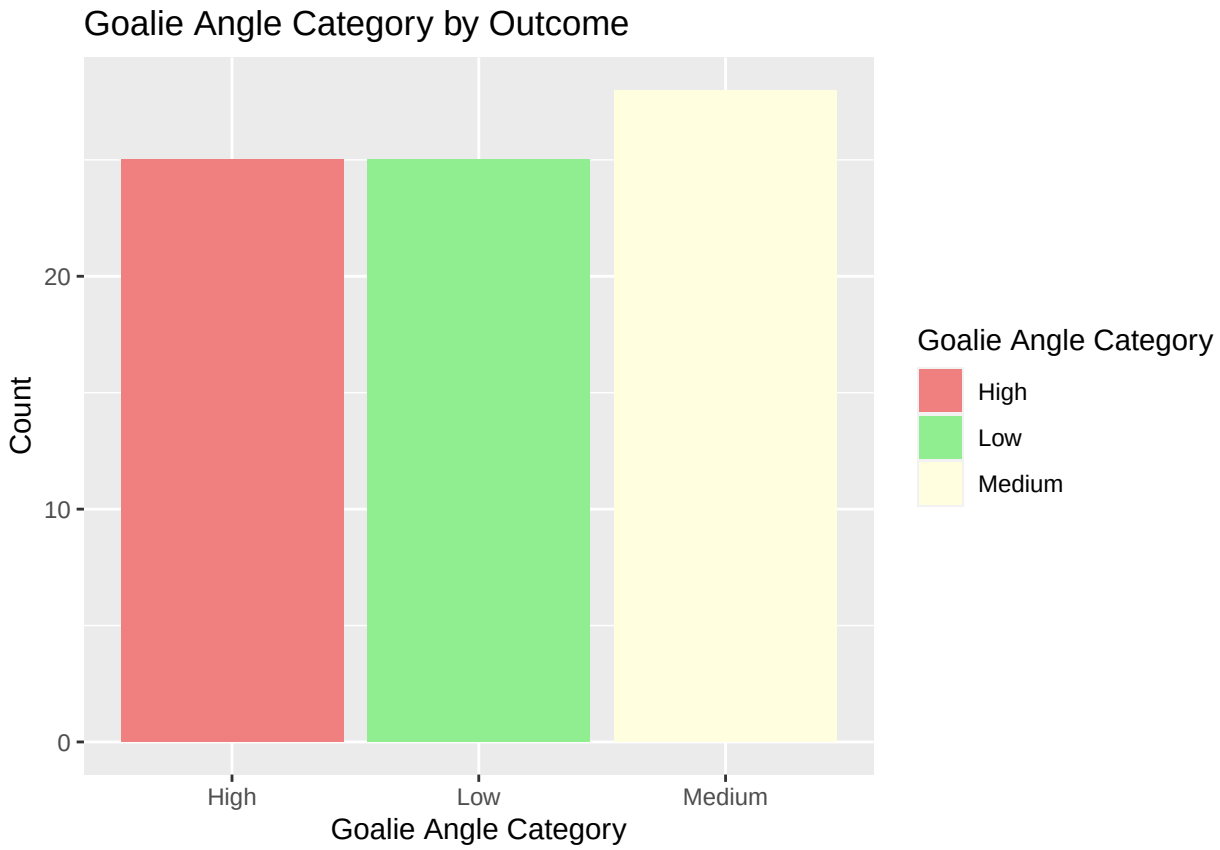

```
ggplot(saves_data, aes(x = shooterSpeedCategory, fill = shooterSpeedCategory)) +
  geom_bar(position = "dodge") +
  scale_fill_manual(values = category_colors) +  # Set custom colors
  labs(
    title = "Shooter Speed Category by Outcome",
    x = "Shooter Speed Category",
    y = "Count",
    fill = "Shooter Speed Category"  # Legend title
  )
```



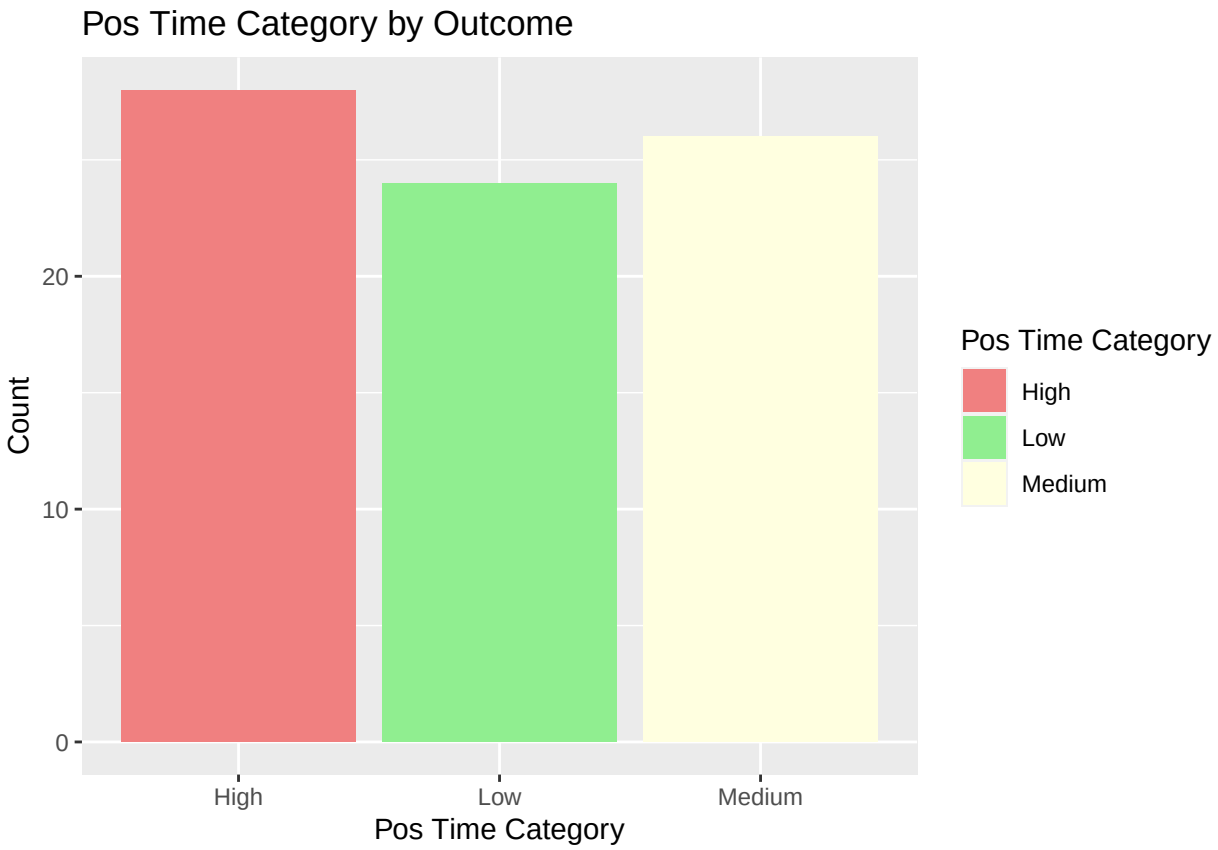

```
ggplot(saves_data, aes(x = goalieDistCategory, fill = goalieDistCategory)) +
  geom_bar(position = "dodge") +
  scale_fill_manual(values = category_colors) +  # Set custom colors
  labs(
    title = "Goalie Dist Category by Outcome",
    x = "Goalie Dist Category",
    y = "Count",
    fill = "Goalie Dist Category"  # Legend title
  )
```




```
ggplot(saves_data, aes(x = goalieAngleCategory, fill = goalieAngleCategory)) +
  geom_bar(position = "dodge") +
  scale_fill_manual(values = category_colors) +  # Set custom colors
  labs(
    title = "Goalie Angle Category by Outcome",
    x = "Goalie Angle Category",
    y = "Count",
    fill = "Goalie Angle Category"  # Legend title
  )
```




```
ggplot(saves_data, aes(x = posTimeCategory, fill = posTimeCategory)) +
  geom_bar(position = "dodge") +
  scale_fill_manual(values = category_colors) +  # Set custom colors
  labs(
    title = "Pos Time Category by Outcome",
    x = "Pos Time Category",
    y = "Count",
    fill = "Pos Time Category"  # Legend title
  )
```



Discussion of results

The results are pretty reasonable. I see that there is a tie between low and medium for puck angle, and some of other variables have a small difference between two categories. But this is due to lack of data, and it looks like this method can explain a lot about modes of failure.

Analysis: Heat Map of the Modes of Failures

Question being asked

We have analyzed the modes of failure. Now, I want to find out the relativity between one mode of failure with other modes of failures. More specifically, I want to find out what features are common and correlated when a shot has failed. Also, we will further examine by categorizing continous variables.

Data Preparation

First, I have created a subset of data which takes in continous variables. Then I found correlation between those varibles and converted it into a matrix so that it can be used for making the heatmap. To make an another heatmap for categorical data, since the heatmap can only take numerical values, I have first converted categorical variables that were continous into binary variables and made a subset of these variables. Then I found the correlation of these variables and graphed into a heatmap.

```
# Include all data processing code (if necessary), clearly commented

# Define the features that need to be included
keeps <- c("puckDist", "puckAngle", "puckSpeed", "shooterSpeed", "goalieDist", "goalieAngle")

# Create a data frame for the heatmap with
```

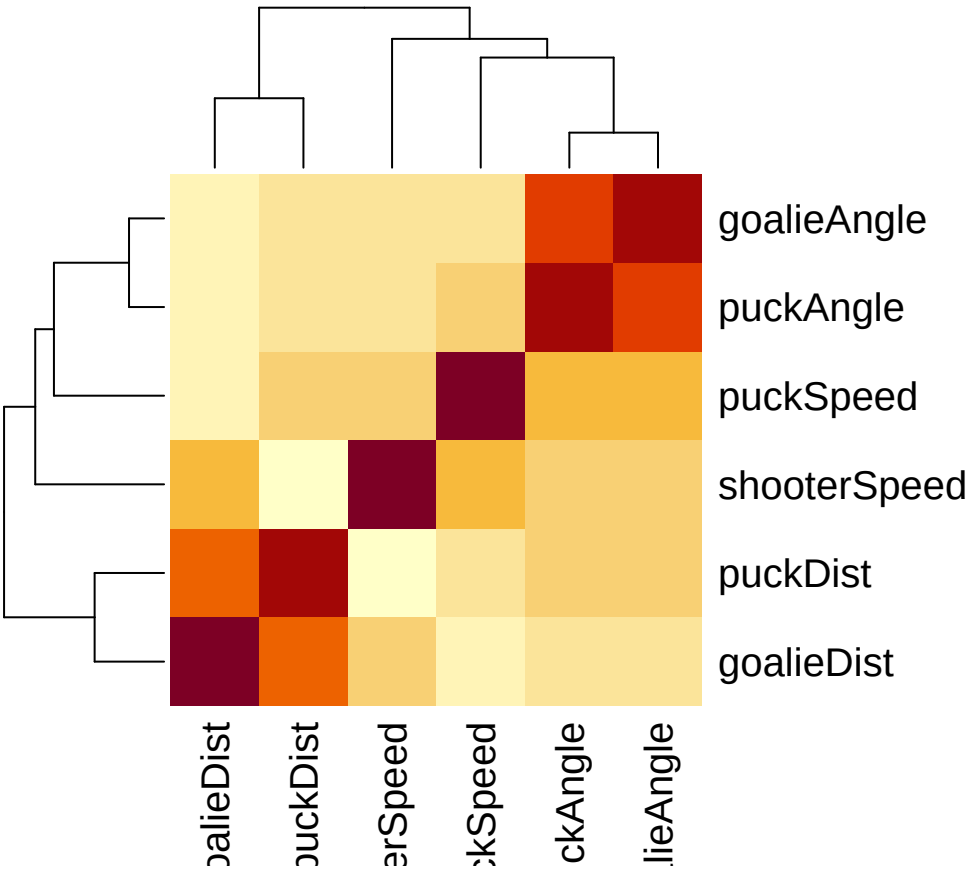
```r
heatmap_data <- saves_data[keeps]

# Find correlation
heatmap_matrix <- cor(heatmap_data[sapply(heatmap_data, is.numeric)])

# Converts categorical variables with 3 categories to binary binary variables
saves_data <- saves_data %>%
  mutate(
    puckSpeedHigh = ifelse(puckSpeedCategory == "High", 1, 0),
    puckSpeedMedium = ifelse(puckSpeedCategory == "Medium", 1, 0),
    puckSpeedLow = ifelse(puckSpeedCategory == "Low", 1, 0),

    puckDistHigh = ifelse(puckDistCategory == "High", 1, 0),
    puckDistMedium = ifelse(puckDistCategory == "Medium", 1, 0),
    puckDistLow = ifelse(puckDistCategory == "Low", 1, 0),

    puckAngleHigh = ifelse(puckAngleCategory == "High", 1, 0),
    puckAngleMedium = ifelse(puckAngleCategory == "Medium", 1, 0),
    puckAngleLow = ifelse(puckAngleCategory == "Low", 1, 0),

    shooterSpeedHigh = ifelse(shooterSpeedCategory == "High", 1, 0),
    shooterSpeedMedium = ifelse(shooterSpeedCategory == "Medium", 1, 0),
    shooterSpeedLow = ifelse(shooterSpeedCategory == "Low", 1, 0),

    goalieDistHigh = ifelse(goalieDistCategory == "High", 1, 0),
    goalieDistMedium = ifelse(goalieDistCategory == "Medium", 1, 0),
    goalieDistLow = ifelse(goalieDistCategory == "Low", 1, 0),

    goalieAngleHigh = ifelse(goalieAngleCategory == "High", 1, 0),
    goalieAngleMedium = ifelse(goalieAngleCategory == "Medium", 1, 0),
    goalieAngleLow = ifelse(goalieAngleCategory == "Low", 1, 0)
  )

# Subset of data to be represented in heatmap
catkeeps <- c("puckSpeedHigh", "puckSpeedMedium", "puckSpeedLow", "puckDistHigh", "puckDistMedium", "puc

# Now we have added binary variables to saves_data, we can take subset of saves_data
heatmap_catdata <- saves_data[catkeeps]

# Find correlation between binary categorical variables
heatmap_catmatrix <- cor(heatmap_catdata[sapply(heatmap_catdata, is.numeric)])
```

Analysis: Methods and Results

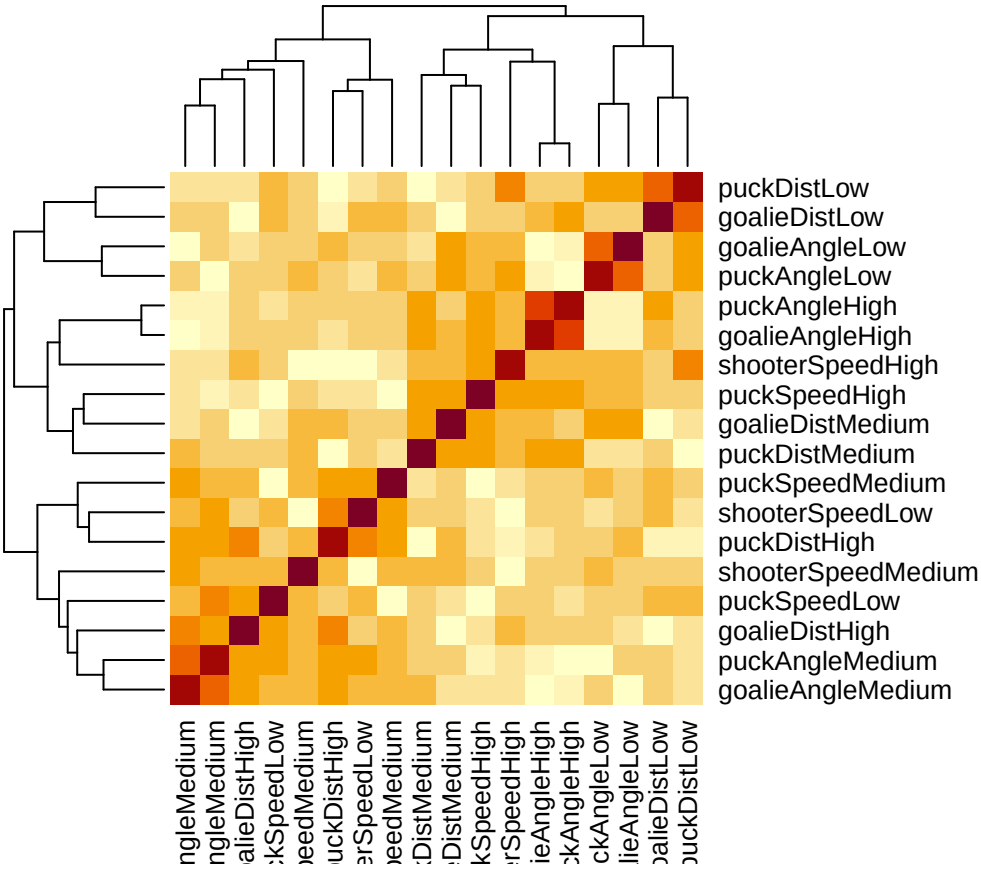
All data have been prepared above, and heatmap_matrix and heatmap_catmatrix are in appropriate matrix form only with numeric values. We can use the heatmap function to produce a heatmap.

```r
# Include all analysis code, clearly commented
# If not possible, screen shots are acceptable.
# If your contributions included things that are not done in an R-notebook,
#   (e.g. researching, writing, and coding in Python), you still need to do
#   this status notebook in R.  Describe what you did here and put any products
#   that you created in github. If you are writing online documents (e.g. overleaf
#   or google docs), you can include links to the documents in this notebook
```

```
#   instead of actual text.

# Heatmap without categorization
heatmap(heatmap_matrix)
```




```
# Heatmap with the categorization
heatmap(heatmap_catmatrix)
```



Discussion of results

We can clearly see that some variables are highly correlated, and there exist variables that are not so correlated to other variables. What is more interesting is that the correlation between the same variables in different categories can vary a lot. For example, we can see that the correlation between puckAngleHigh and goalieAngleHigh is very high, meaning that a shot within these two categories is likely to be a save. However, the correlation between puckAngleHigh and goalieAngleLow is very low, meaning a shot within these two categories is likely to not be a save.

Analysis: N/A

Question being asked

Data Preparation

```
# Include all data processing code (if necessary), clearly commented
```

Analysis methods used

Describe in natural language a statement of the analysis you're trying to do

Provide clearly commented analysis code; include code for tables and figures!

```
# Include all analysis code, clearly commented
# If not possible, screen shots are acceptable.
# If your contributions included things that are not done in an R-notebook,
#   (e.g. researching, writing, and coding in Python), you still need to do
#   this status notebook in R.  Describe what you did here and put any products
```

```
#   that you created in github. If you are writing online documents (e.g. overleaf
#   or google docs), you can include links to the documents in this notebook
#   instead of actual text.
```

Discussion of results

Provide in natural language a clear discussion of your observations.

Summary and next steps

Next week, I will find some way to incorporate categorical data into cluster analysis.