DAR F23 Project Status Notebook

Hockey Analytics

Amy Enyenihi

2023-10-30

Contents

Weekly Work Summary

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

- RCS ID: enyena
- Project Name: Hockey Analytics
- Summary of work since last week
  - I continued to work on visualization of rink and the players.
- Summary of github issues added and worked
  - N/A
- Summary of github commits
  - branch name: dar-enyena
  - Updated the shots_stats_goal file to remove repeat feature: https://github.rpi.edu/DataINCITE/Hockey_Fall_2023/blob/main/StudentData/shots_stats_goal.df.Rds
  - Added Caleb's Rds file, as per his request, with his clusters as a new feature: https://github.rpi.edu/DataINCITE/Hockey_Fall_2023/blob/main/StudentData/shot_stats_goal_clusters.df.Rds
- List of presentations, papers, or other outputs
  - https://docs.google.com/presentation/d/1EVZPEgD-1kW1jJnWyxOlaZB-rY22s-CnOWjcbfydQpw/edit#slide=id.gb75c90f927_0_22
  - https://docs.google.com/presentation/d/1fvrcLAgWpKUDxGLxeLrjXp56otHXgIafFzB0Xyr7bRc/edit#slide=id.gb75c90f927_0_22
  - https://docs.google.com/presentation/d/1w2C0_E5i9exIyJkIGHFStFG4Pb1rYR_b1zraEa_njVQ/edit#slide=id.gb75c90f927_0_22
  - https://docs.google.com/presentation/d/1YM2R4mTCLE08qst519VYvdoJfaP7ys00qXlIret2w3M/edit#slide=id.gb75c90f927_0_22
- 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
    - I started with code that was originally written by Mohamed (rink plot). This week, I continued to work with clusters that were developed by Caleb and began to work with the data frame created by Jeff.
- **Required:** Provide illustrating figures and/or tables

Personal Contribution

- Clearly defined, unique contribution(s) done by you: code, ideas, writing…
- Include github issues you've addressed
    - I have been the main contributor to the development of the rink plot displays. I have used features added by Caleb and Jeff, but I created the images.

Analysis: Question 1 - Improving the Rink Display of Outcomes

Question being asked

Provide in natural language a statement of what question you're trying to answer

How can we improve the previously made rink plot colors to be more distinct and ideally color blind friendly?

Data Preparation

Provide in natural language a description of the data you are using for this analysis

Include a step-by-step description of how you prepare your data for analysis

If you're re-using dataframes prepared in another section, simply re-state what data you're using

I am using the data frame that was created by Mohamed and Dr. Morgan. Because I am displaying the rink, I included the appropriate numbers for the rink size and coordinates and loaded in the jpegs.

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

#read in the data frame
shots_stats.df <- readRDS("../../StudentData/shots_stats_goal.df.Rds")

# Size of rink image and of all plots
xsize <- 2000
ysize <- 850

# Coordinates to the goal pipes
pipes_x <- 1890
lpipe_y <- 395
rpipe_y <- 455

# This file path should contain the hockey rink images and all the sequences
filepath <- '../../FinalGoalShots/'

# Read the rink images and format them to a raster used for graphing
rink_raster <- makeRaster(filepath, 'Rink_Template.jpeg')
half_rink_raster <- makeRaster(filepath, 'Half_Rink_Template.jpeg')
```

Analysis: Methods and results

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!

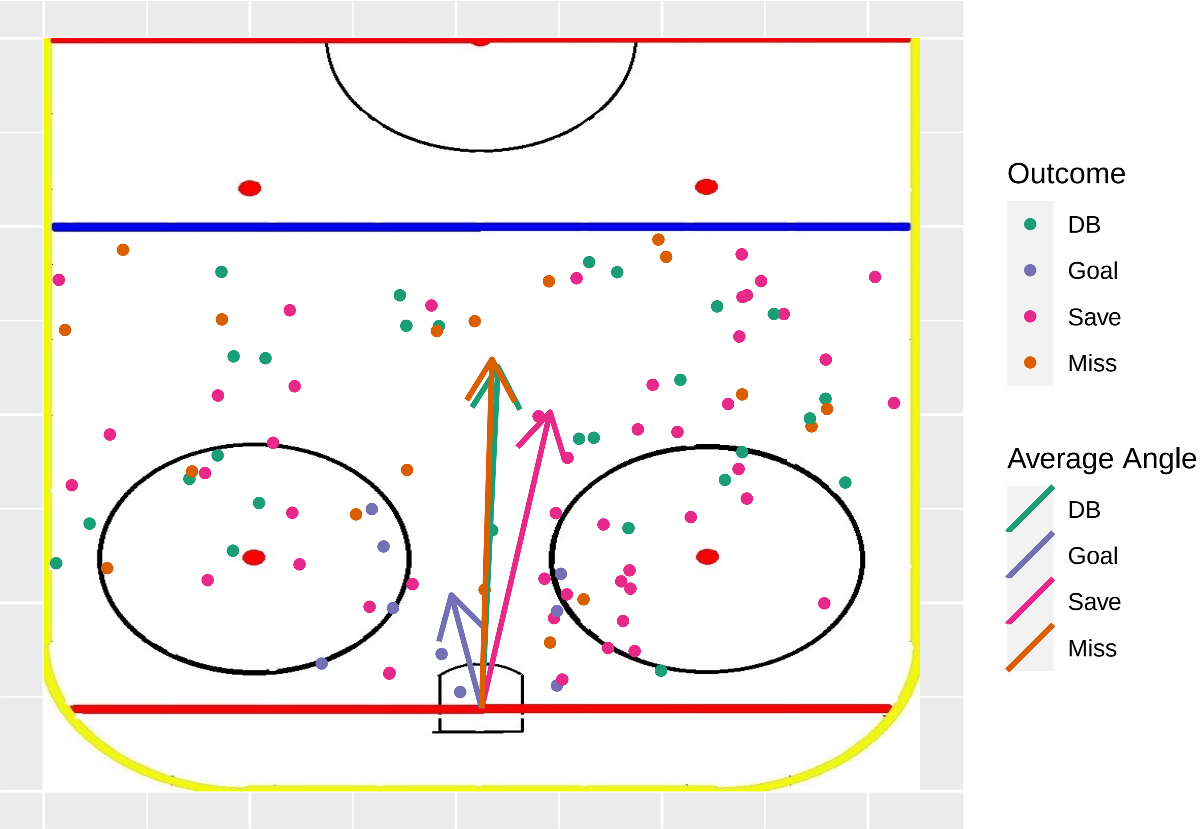
I want to use ColorBrewer to develop the rink image with new colors.

```
# 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.

shot_outcomes <- shots_stats.df %>%
  group_by(shotOutcome) %>%
  summarise(mean(puckAngle), mean(puckDist)) %>%
  set_names(c('outcome', 'meanAngle', 'meanDist')) %>%
  # Data for graphing
  mutate(xstart = ysize / 2) %>%
  mutate(ystart = pipes_x) %>%
  mutate(radius = meanDist)

halfRinkGraph(shots_stats.df) +
  # Graph players colored by shot type
  geom_point(aes(color = shotOutcome, x = shotStatX(shots_stats.df), y =
    shotStatY(shots_stats.df))) + scale_color_discrete('Outcome', type =
    c('#1b9e77', '#7570b3', '#e7298a', '#d95f02'), labels = c('DB', 'Goal',
    'Save','Miss')) + new_scale_color() +
  # Arrow pointing to average direction
  geom_spoke(data = shot_outcomes, aes(x = xstart, y = ystart, angle =
    torad(meanAngle), radius = radius, color = outcome), key_glyph = 'abline',
    linetype = 'solid', arrow = arrow(), linewidth = 1) +
  scale_color_discrete('Average Angle', type = c('#1b9e77', '#7570b3', '#e7298a',
    '#d95f02'), labels = c('DB', 'Goal', 'Save','Miss')) +
  labs(x = NULL, y = NULL)
```

```
## Warning: Removed 6 rows containing missing values (`geom_point()`).
```



```
# to show the raw numbers used in the creation of the image
print(shot_outcomes)
```

```
## # A tibble: 4 x 6
##   outcome        meanAngle meanDist xstart ystart radius
##   <fct>              <dbl>    <dbl>  <dbl>  <dbl>  <dbl>
## 1 Defender Block      88.0     454.    425   1890   454.
## 2 Goal               101.      153.    425   1890   153.
## 3 Save                80.4     399.    425   1890   399.
## 4 Miss                88.8     464.    425   1890   464.
```

Discussion of results

Provide in natural language a clear discussion of your observations.

As nothing but the colors have changed for the display, there are no new observations. I shifted away from using red, black, and blue as they are colors used in the image that match a real hockey rink.

Analysis: Question 2 - Visualizing Clusters Created by Caleb

Question being asked

Provide in natural language a statement of what question you're trying to answer

How can I visualize the clusters, previously developed by Caleb, on the rink?

Data Preparation

Provide in natural language a description of the data you are using for this analysis

Include a step-by-step description of how you prepare your data for analysis

If you're re-using dataframes prepared in another section, simply re-state what data you're using

All items necessary to visualize the rink are already loaded in, so no preparation necessary for that portion. As for the data itself, I will be using the data frame Caleb uploaded to Github. It is the same as the original data frame but it contains an additional column for Caleb's clusters (numbered 1 through 5). All cluster labels were provided by Caleb and a detailed explaination of each can be found in his assignment 4 notebook.

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

#read in the data frame
shots_stats_clusters <- readRDS("../../StudentData/shot_stats_goal_clusters.df.Rds")
```

Analysis: Methods and Results

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!

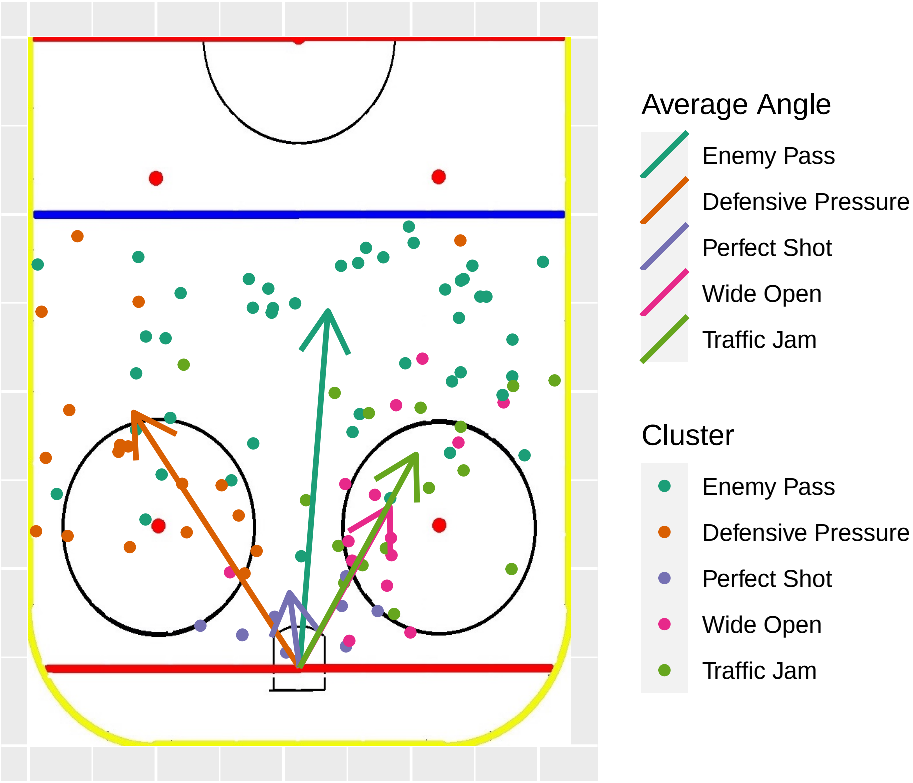
Caleb has analyzed and provided his own explanation as to how each cluster was established. I intend to simply provide a visual to the analysis he started. The numbers used to generate the arrows are the average of each cluster so while the information will not add statistical evidence, it will tell us about how the clusters differ from one another, on average.

```
# 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.

c_clusters <- shots_stats_clusters %>%
  group_by(Cluster) %>%
  summarise(mean(puckAngle), mean(puckDist)) %>%
  set_names(c('cluster', 'meanAngle', 'meanDist')) %>%
  # Data for graphing
  mutate(xstart = ysize / 2) %>%
  mutate(ystart = pipes_x) %>%
  mutate(radius = meanDist)

halfRinkGraph(shots_stats_clusters) +
  # Graph players colored by shot type
  geom_point(aes(color = Cluster, x = shotStatX(shots_stats_clusters), y =
    shotStatY(shots_stats_clusters))) + scale_color_discrete('Cluster', type =
    c('#1b9e77', '#d95f02', '#7570b3', '#e7298a', '#66a61e'), labels =
    c('Enemy Pass', 'Defensive Pressure', 'Perfect Shot','Wide Open',
    'Traffic Jam')) + new_scale_color() +
  # Arrow pointing to average direction
  geom_spoke(data = c_clusters, aes(x = xstart, y = ystart, angle = torad(meanAngle),
    radius = radius, color = cluster), key_glyph = 'abline', linetype = 'solid',
    arrow = arrow(), linewidth = 1) +
  scale_color_discrete('Average Angle', type = c('#1b9e77', '#d95f02', '#7570b3',
    '#e7298a', '#66a61e'), labels = c('Enemy Pass', 'Defensive Pressure',
    'Perfect Shot','Wide Open', 'Traffic Jam')) +
  labs(x = NULL, y = NULL)
```

```
## Warning: Removed 3 rows containing missing values (`geom_point()`).
```



```
# to show the raw numbers used in the creation of the image
print(c_clusters)
```

```
## # A tibble: 5 x 6
##   cluster meanAngle meanDist xstart ystart radius
##   <fct>       <dbl>    <dbl>  <dbl>  <dbl>  <dbl>
## 1 1            84.9     505.    425   1890   505.
## 2 2           126.      444.    425   1890   444.
## 3 3            98.4     107.    425   1890   107.
## 4 4            58.1     268.    425   1890   268.
## 5 5            58.8     352.    425   1890   352.
```

Discussion of results

Provide in natural language a clear discussion of your observations.

This image supports the idea that cluster three is the "Perfect Shot". On average, shots from this cluster were, at minimum, two times closer than the average shot from other clusters. This would make sense as, the closer the shooter is to the goal, the "more perfect" the shot is. The average distances appear to make sense as all the shots in the Perfect Shot cluster are found close to the goal while others are spread across the half rink. Clusters 2, 4, and 5, (Defensive Pressure, Wide Open, and Traffic Jam) all have an average angle significantly far from 90 degrees. The average angle for clusters 1 and 3 (Enemy Pass and Perfect Shot) are much more centralized. While they are more similar, but as mentioned, the Perfect Shot cluster has an average distance that is about 5 times closer to the goal than the average shot of Enemy Pass. I would presume this is a distinctive feature that separates these two clusters.

While we can gain the information discussed above, I think there is some information lost with the way I chose to show the selected features. The current version of this plot does not have any indication of the goals vs non goals. So while it can be nice to see the clusters visualized this way, more meaningful information can be reported with representation of the goals and non goals.

Analysis: Question 3 - Visualizing Categorized Puck Speed

Question being asked

Provide in natural language a statement of what question you're trying to answer

What visuals can I create using the categorized variables generated by Jeff?

Data Preparation

Provide in natural language a description of the data you are using for this analysis

Include a step-by-step description of how you prepare your data for analysis

If you're re-using dataframes prepared in another section, simply re-state what data you're using

For this portion, I will be using the data frame that Jeff created. This data frame contains all the same features, but the continuous variables have been categorized into three levels.

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

# read in the data frame
shots_stats_cat <- readRDS("../../StudentData/categorized_shots_stats_goal.df.Rds")

# everything needs to be factored rather than numeric for imaging
# the cluster values need to be a factor, not numeric
shots_stats_cat$puckSpeedCategory <- shots_stats_cat$puckSpeedCategory %>% as.factor()
```

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!

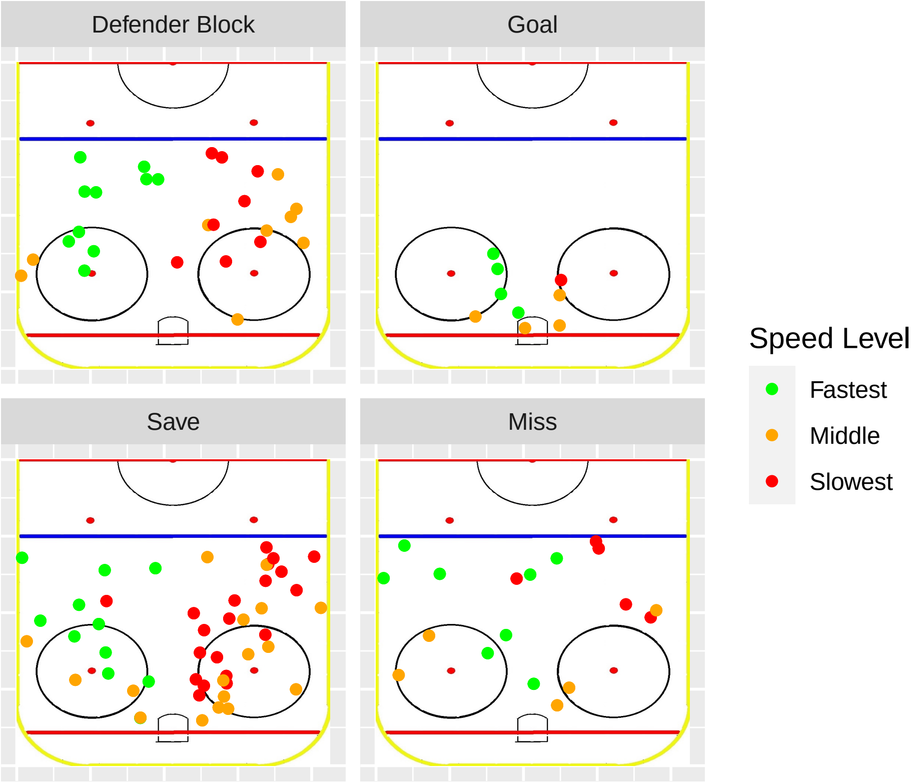
After hearing Jeff had categorized the continuous variables, I wanted to see how we could generate images that help tell a story. Even if they did not, I wanted to experiment and see where it could lead me. My hopes were to create something that could be reflected in the app, but first I started with a simple puck speed image sorted into the four outcomes.

```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.

# looking at puck speed categorized for each outcome
shot_outcomes_speed <- shots_stats_cat %>%
  group_by(shotOutcome) %>%
  summarise(mean(puckAngle), mean(puckDist)) %>%
  set_names(c('outcome', 'meanAngle', 'meanDist')) %>%
  # Data for graphing
  mutate(xstart = ysize / 2) %>%
  mutate(ystart = pipes_x) %>%
  mutate(radius = meanDist)

halfRinkGraph(shots_stats_cat) +
  # Graph players colored by shot type
  geom_point(aes(color = puckSpeedCategory, x = shotStatX(shots_stats_cat), y =
    shotStatY(shots_stats_cat))) + scale_color_discrete('Speed Level',type =
    c('green','orange','red'), labels = c('Fastest', 'Middle', 'Slowest')) +
  new_scale_color() + facet_wrap(~shotOutcome)
```

```
## Warning: Removed 6 rows containing missing values (`geom_point()`).
```



Discussion of results

Provide in natural language a clear discussion of your observations.

Much like other visuals, there is not a lot of new information to gain however some of our puck speed expectations are confirmed. As expected, the slowest shooting speeds (red) were observed frequently in the non goal outcomes and much less frequently among the goals. When presenting in class, Dr. Morgan pointed out the bias to the left and right sides relative to the shooting speed. Because of this, I intend to look further into potential sources or similar trends.

Summary and next steps

Provide in natural language a clear summary and your proposed next steps.

As for I as I can tell, the first image is okay as is. I will likely not be revisiting it unless there is a direct ask for changes relating to it.

As for the clustering images, I will be changing the image to provide a more clear story. The current image does not indicate any of the goals vs non goals and that information is key. Additionally, for some of the clusters, the placement of the other players is important. I believe for clusters like “Traffic Jam” and “Wide Open”, visualizing the other players relative to the shooter will provide a better idea of what is going on for the play. This inspired me to begin to work on creating visuals for individual plays so that I can create images that include all the players. Once I have that figured out, it could be a feature added to the app.

Finally, I plan on adding more features to the last image. Because of the feedback Dr. Morgan provided, I will be adding handedness as a feature to see how that reflects to the left/right side shooting bias. I want to use Jeff’s data frame more to visualize other categorized variables and look for trends between goals and non goals or even other features.