EDAV Fall 2019 PSet 5, part A

Jake Stamell (jfs2167), Hariz Johnson (hsj2114), Samrat Halder (sh3970)

This assignment is designed to help you get started on the final project. Be sure to review the final project instructions (https://edav.info/project.html), in particular the new section on reproducible workflow (https://edav.info/project.html#reproducible-workflow), which summarizes principles that we've discussed in class.

1. The Team

[2 points]

a) Who's on the team? (Include names and UNIs)

Jake Stamell (jfs2167), Hariz Johnson (hsj2114), Samrat Halder (sh3970)

b) How do you plan to divide up the work? (Grading is on a group basis. The point of asking is to encourage you to think about this.)

To start off, Jake has worked on manually pulling and cleaning data from the NHL API. Given Jake's familiarity with the data, he will focus on exploratory data anlaysis to identify views of the data and filters to include in the final output.

Samrat and Hariz are setting up the shiny app and creating the initial UI and hockey rink visualization. They will focus on creating the visualizations for shiny and building in the interactive features.

Our remaining tasks include: - Plot data onto the rink - Add interactive filters to the UI - Structure the data and filter it based on the user's inputs - Make the app visually appealing - Once we have achieved minimum viable product we can continue adding additional filters, plots, etc

2. The Questions

[6 points]

Recently, play-level data in hockey has become more detailed and accessible. If we filter the data available in different ways and plot it we hope to better understand the game and find answers for the following questions:

a) What are the indicators of success in hockey?

b) Can we distinguish different play styles from our visualization and if so, which play styles are most effective?

c) Can we identify the characteristics of shots that are more likely to result in a goal (location, type of shot, etc)?

3. Which output format do you plan to use to submit the project?

[2 points]

We plan to submit a shiny app.

4. The Data

What is your data source? What is your method for importing data? Please be specific. Provide relevant information such as any obstacles you're encountering and what you plan to do to overcome them.

[5 points]

We are working directly with the Nhl.com API, using the httr package to make data calls and the jsonlite package to parse the raw data. One obstacle has been the lack of documentation for the API provided by the NHL. Luckily, there is an online community that is working on documenting it, which has helped to identify some endpoints. Using these endpoints has also yieled other endpoints for retrieving data. Beyond this, it has primarily been manual work to go through the data provided by the various calls to identify what is useful.

One other challenge has been missing data values. For example, we are focusing on play-level data; however, some games are missing all records of these. Unfortunately, we cannot resolve this particular issue so will have to focus our analysis on the existing records. This will require some additional checks to make sure the missing data is not skewing our analysis. Another area of missing data arises naturally from the data we are using. Some information is specific to the type of play. For example, the information recorded for a shot is different from the info recorded for a hit. To overcome this, we will have to be careful about how we dynamically filter our data for the interactive display when dealing with different play types.

5. Provide a short summary, including 2-3 graphs, of your initial investigations.

[10 points]

So far, we have had a parallel process of 1) collecting/cleaning data and 2) setting up the shiny UI. Below, we have included a few exploratory visualizations of the data as well as a screenshot of our current UI.

Exploratory visualizations: Our first chart displays the frequency of play-types in our data. While goals are the most interesting events in hockey, they occur relatively infrequently. Shots on the other hand are more plentiful, and we will likely need to focus our future (offensive) analysis on these events. We can see this in charts 2-4. These all show the coordinates of where plays occur. Chart 2 shows all plays, 3 only face offs, and 4 only shots. When we look at all plays, the face offs dominate the visualization because those happen in fixed locations. (This can be seen in chart 3). The shots display shows clearly that shots are concentrated around the net.

```r
# loading packages
library(data.table)
library(ggplot2)
library(forcats)

# reading in data
load("2019-11-21_nhl-cleaned-data.RData")

# Plots
ggplot(vF_game_plays_2017[,.N,by=result.eventTypeId][order(N)]) +
  geom_col(aes(x=fct_inorder(result.eventTypeId),y=N)) +
  labs(x="Play type"
       ,y="Count"
       ,title="Overview of play types") +
  coord_flip()
```
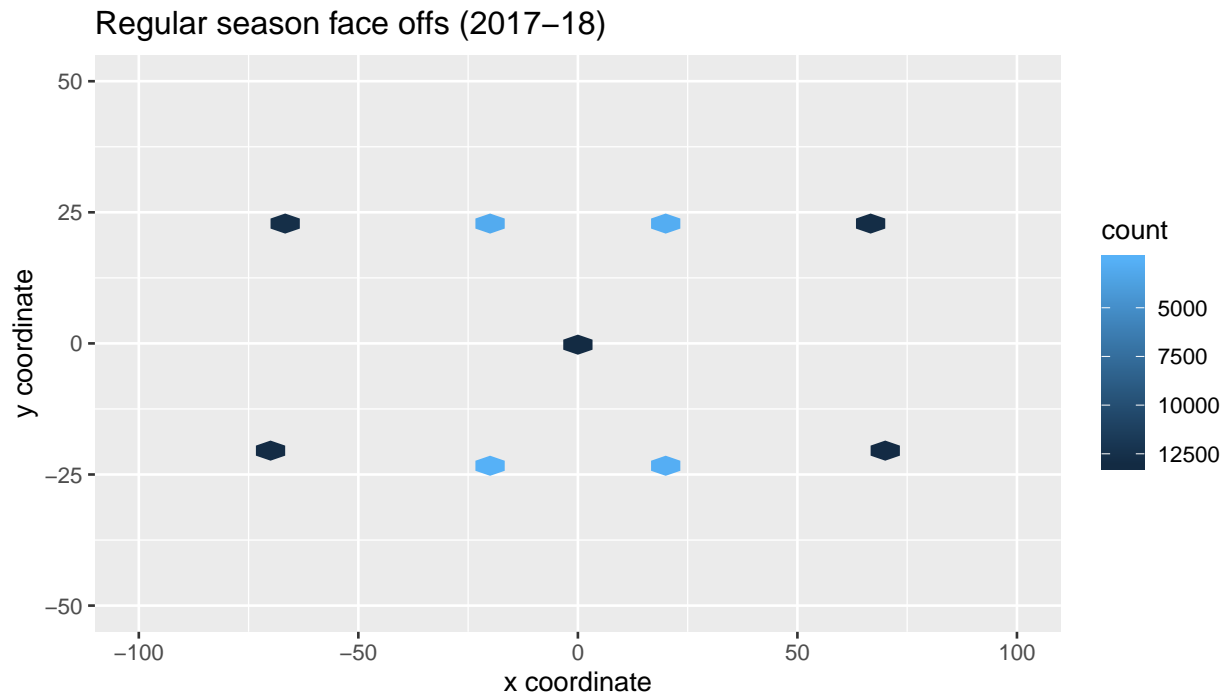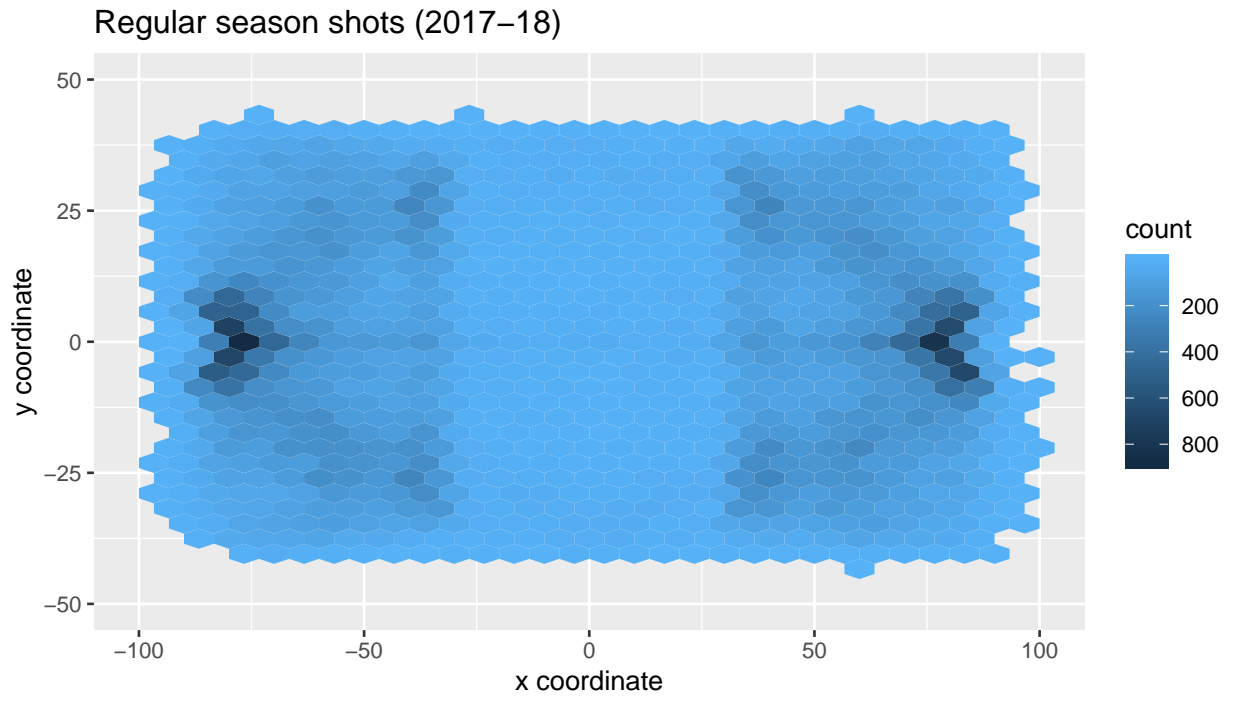
Overview of play types
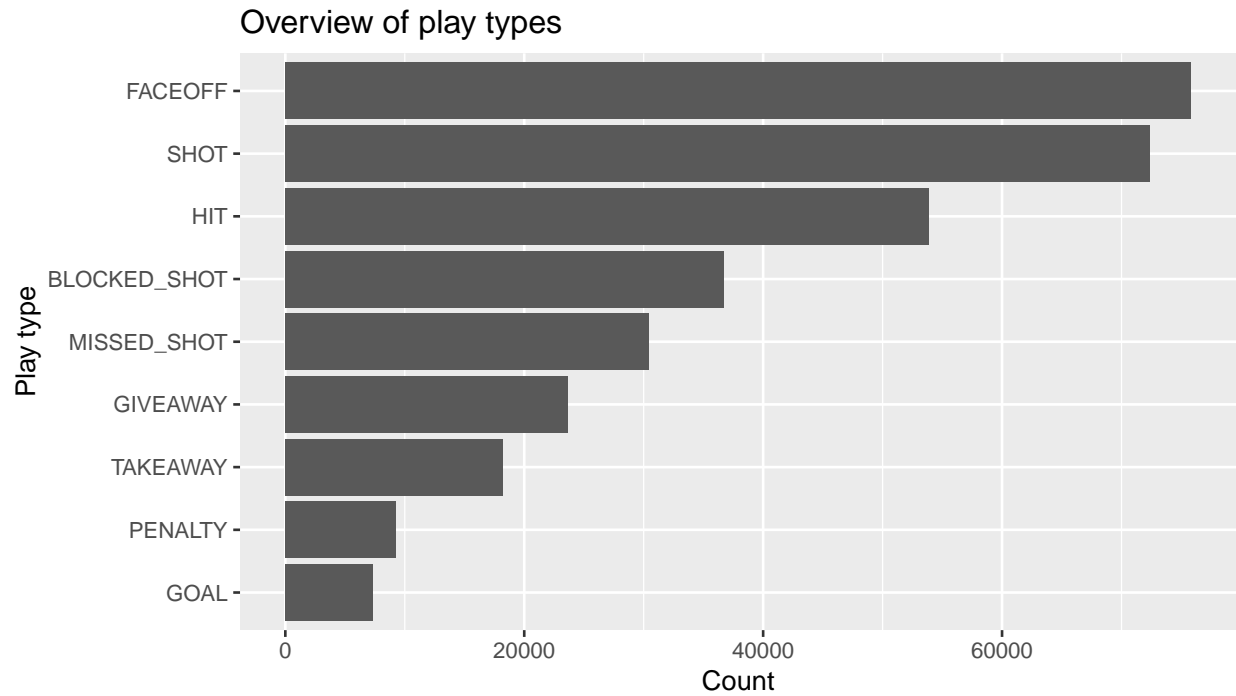

```
ggplot(vF_game_plays_2017) +
  geom_hex(aes(x=s.x,y=coordinates.y)) +
  scale_x_continuous(limits = c(-100,100)) +
  scale_y_continuous(limits = c(-50,50)) +
  scale_fill_continuous(trans="reverse") +
  labs(x="x coordinate"
      ,y="y coordinate"
      ,title="Regular season plays (2017-18)")
```

Regular season plays (2017−18)
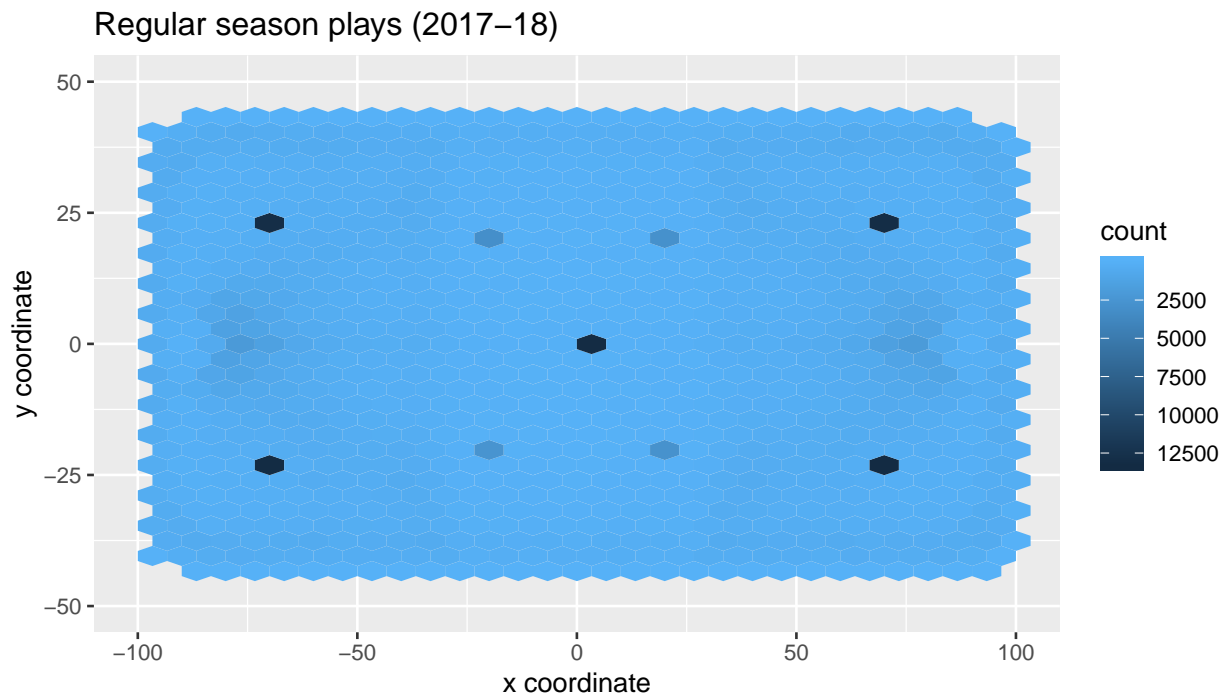
```
ggplot(vF_game_plays_2017[result.eventTypeId=="FACEOFF"]) +
  geom_hex(aes(x=s.x,y=coordinates.y)) +
  scale_x_continuous(limits = c(-100,100)) +
  scale_y_continuous(limits = c(-50,50)) +
  scale_fill_continuous(trans="reverse") +
  labs(x="x coordinate"
       ,y="y coordinate"
       ,title="Regular season face offs (2017-18)")
```


Regular season face offs (2017−18)

```
ggplot(vF_game_plays_2017[result.eventTypeId=="SHOT"]) +
  geom_hex(aes(x=s.x,y=coordinates.y)) +
  scale_x_continuous(limits = c(-100,100)) +
  scale_y_continuous(limits = c(-50,50)) +
  scale_fill_continuous(trans="reverse") +
  labs(x="x coordinate"
       ,y="y coordinate"
       ,title="Regular season shots (2017-18)")
```

Regular season shots (2017–18)

Figure 1: UI sample