# STAT 301-3 Final Project EDA

*Albert Li*

*May 6, 2018*

## Load Packages

```
# Loading package(s)
library(tidyverse)
library(stringr)
```

## Read In and Modify Dataset

```
#remove unnecessary columns, factorize certain columns
shot_dat <- read_csv("data/unprocessed/shot_logs.csv") %>%
  select(-CLOSEST_DEFENDER_PLAYER_ID, -player_id, -GAME_ID) %>%
  mutate(LOCATION = factor(LOCATION, levels = c("A", "H")),
         W = factor(W, levels = c("W", "L")),
         SHOT_RESULT = factor(SHOT_RESULT, levels = c("made", "missed")),
         PTS_TYPE = factor(PTS_TYPE, levels = c(2, 3)),
         PERIOD = factor(PERIOD, levels = c(1, 2, 3, 4, 5, 6, 7)))
```

```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   MATCHUP = col_character(),
##   LOCATION = col_character(),
##   W = col_character(),
##   GAME_CLOCK = col_time(format = ""),
##   SHOT_CLOCK = col_double(),
##   TOUCH_TIME = col_double(),
##   SHOT_DIST = col_double(),
##   SHOT_RESULT = col_character(),
##   CLOSEST_DEFENDER = col_character(),
##   CLOSE_DEF_DIST = col_double(),
##   player_name = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
#create opposing team variable
shot_dat <- shot_dat %>%
  mutate(opposing_team = factor(str_sub(MATCHUP, start = -3)))

#replace na's with 0
shot_dat <- shot_dat %>% replace_na(replace = list(SHOT_CLOCK = 0))

#change game clock to just seconds
shot_dat <- shot_dat %>%
  mutate(min = as.numeric(str_sub(GAME_CLOCK, start = 1, end = 2)),
         sec = as.numeric(str_sub(GAME_CLOCK, start = 4, end = 5)),
         GAME_CLOCK_sec = 60*min + sec)
```

The data I am using consists of all shot attempts taken during the 2014-2015 NBA regular season. Some changes I made to the dataset include removing unnecessary fields, factorizing certain predictors, creating the opposing team variable, replacing NA's under SHOT_CLOCK to 0 (this occurs when the shot clock is turned off in game), and converting the time left in the quarter to seconds.

## Create Train and Test Datasets

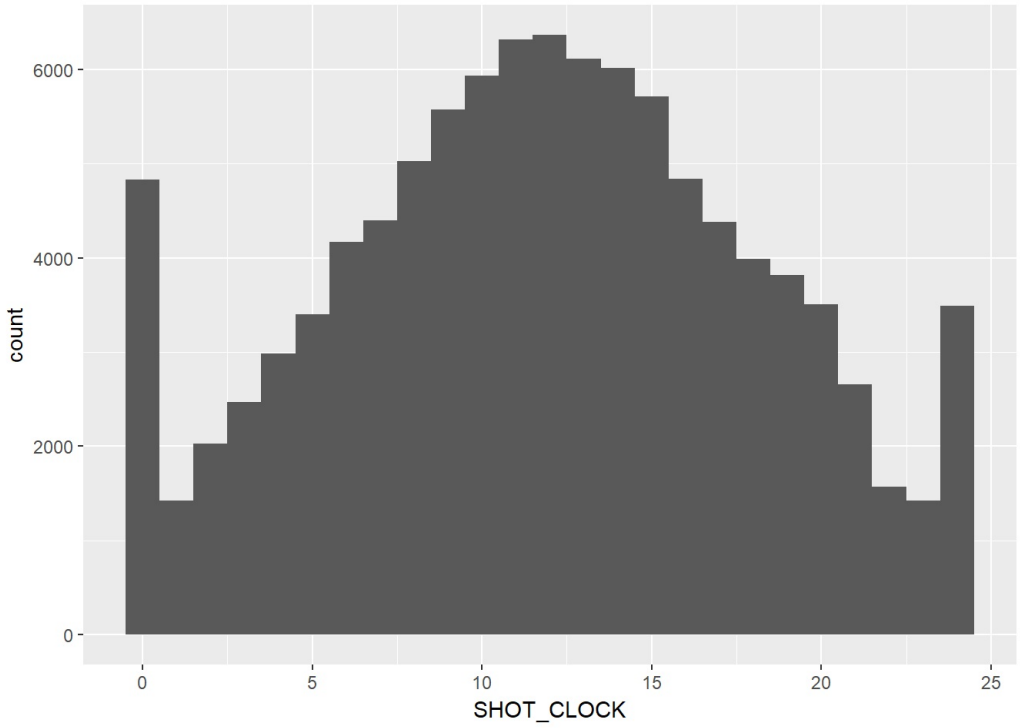
```
#set seed
set.seed(1)

#leave out 20% of data
shot_dat_train <- shot_dat %>% sample_frac(.8)
shot_dat_test <- shot_dat %>% setdiff(shot_dat_train)
```

I randomly sampled 80% of the dataset to be the training data while leaving out the other 20% to be the test data to be used during the modelling process. I will use shot_dat_train for the remainder of the EDA.

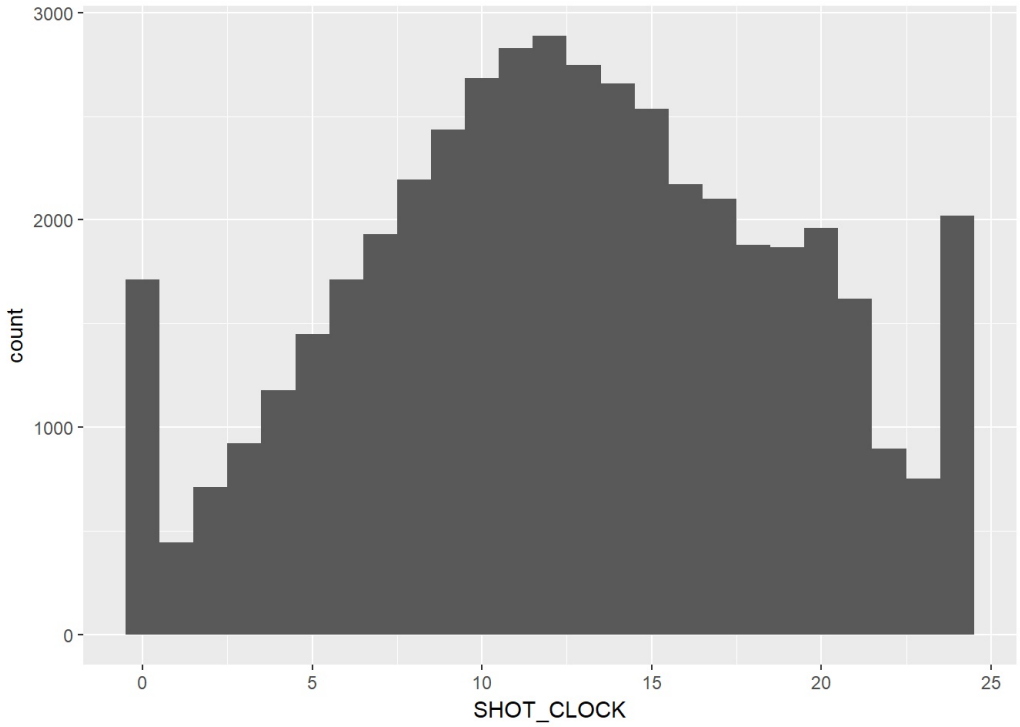
## Shot Clock Distribution

During an NBA game, when a team has possession of the basketball, they have 24 seconds to shoot the basketball. This 24 seconds is called the shot clock. The dataset includes the time left on the shot clock when every shot attempt is taken. Let's take a look at the distribution of time left on the shot clock.

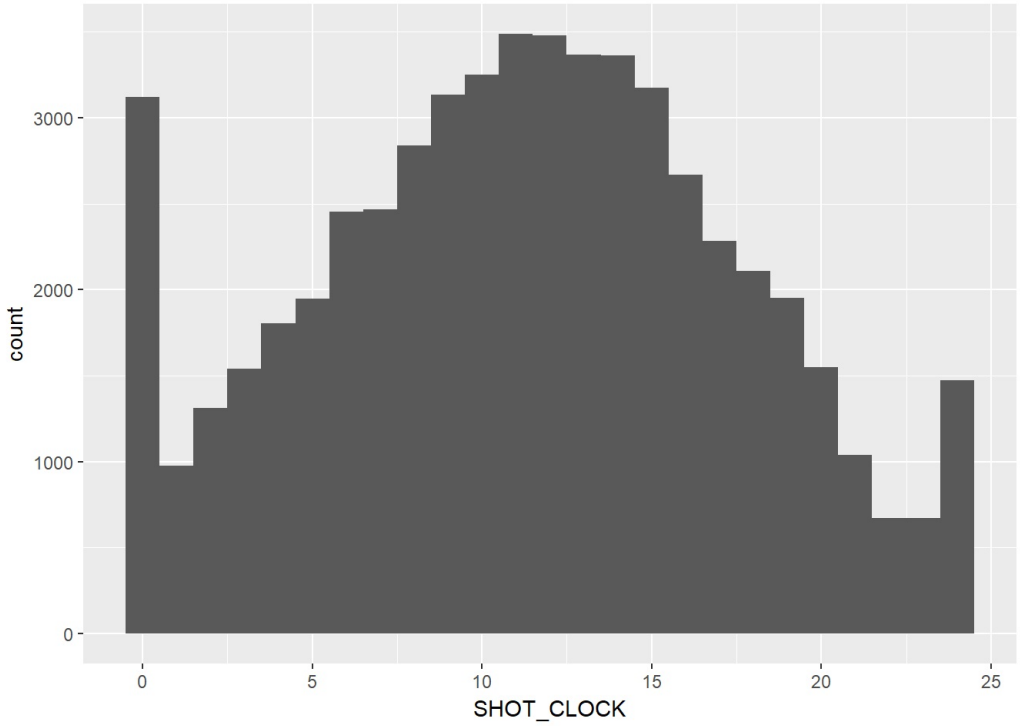
```
#shot clock distribution
shot_dat_train %>%
  ggplot(aes(x = SHOT_CLOCK)) +
  geom_histogram(binwidth = 1)
```




```
#makes
shot_dat_train %>%
  filter(FGM == 1) %>%
  ggplot(aes(x = SHOT_CLOCK)) +
  geom_histogram(binwidth = 1)
```




```
#misses
shot_dat_train %>%
  filter(FGM == 0) %>%
  ggplot(aes(x = SHOT_CLOCK)) +
  geom_histogram(binwidth = 1)
```



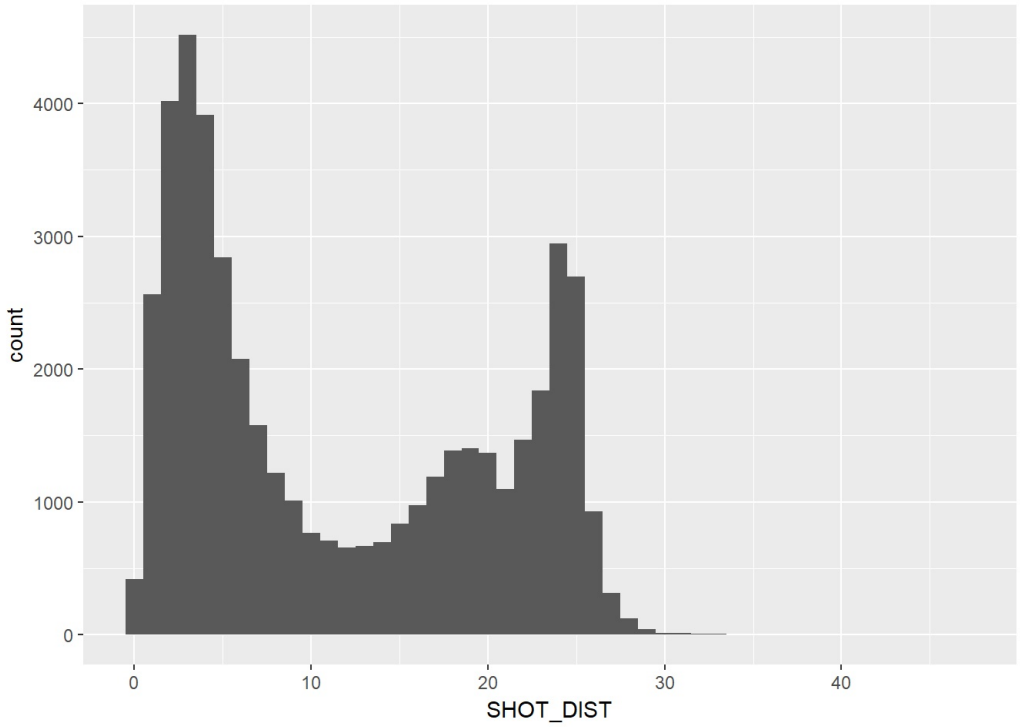

we can see there is a fairly normal distribution with the exception being a large number of shots taken near the expiration of the shot clock and a large number of shots taken at the start of the shot clock. This makes sense since I categorized as shot attempts when the shot clock is turned off as 0 seconds on the shot clock. The large number of shots taken at the start of the shot clock (24 seconds) is most likely due to a player getting a rebound off a missed shot and shooting the ball instantly. Most of these shots typically occur near the basket. We also see more misses when the shot clock is at 0 and more makes when the shot clock is at 24.

## Shot Distance Distribution

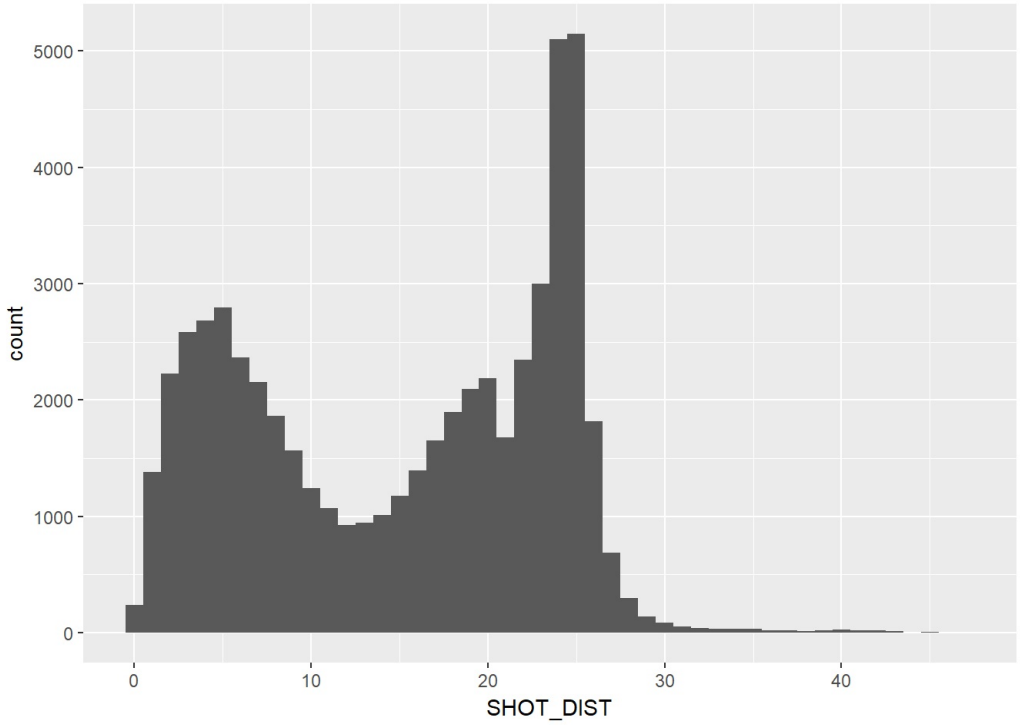
Shot attempts are taken at a variety of ranges. Let's take a look at the distributions. Shot distance is measured in feet.

```
#shot distance distribution
shot_dat_train %>%
  ggplot(aes(x = SHOT_DIST)) +
  geom_histogram(binwidth = 1)
```


```
#makes
shot_dat_train %>%
  filter(FGM == 1) %>%
  ggplot(aes(x = SHOT_DIST)) +
  geom_histogram(binwidth = 1)
```




```
#misses
shot_dat_train %>%
  filter(FGM == 0) %>%
  ggplot(aes(x = SHOT_DIST)) +
  geom_histogram(binwidth = 1)
```

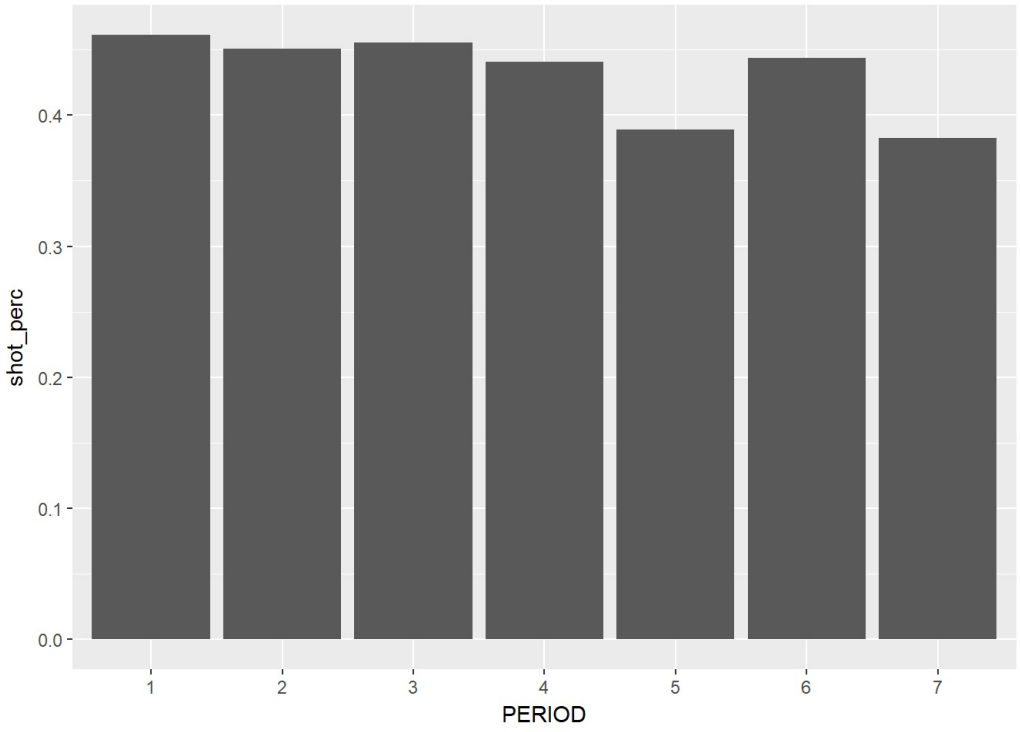



Overall, we see a clear bimodal distribution where a lot of shot attempts are either very close to the basket or very far, beyond the three point line. This makes a lot of sense, as the modern NBA strategy focuses on shooting threes and driving to the basket for a lay-up. From the shot distance distributions on made and missed shots, we see that most misses are from long range while most makes are from short range. Again, this makes sense as it is much easier to shoot a basketball as one moves closer to the basket.

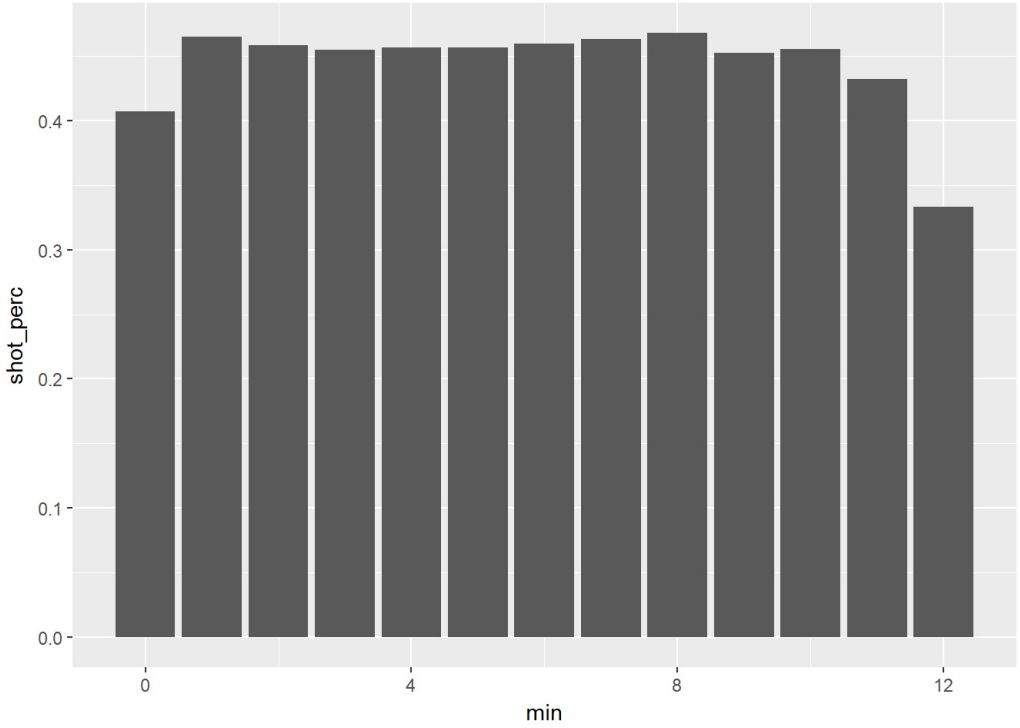
## NBA Clock and Period

Next, I wanted to see if the time has an effect on a player's shooting percentage. To observe this, I found shooting percentages by NBA period, and also by time remaining on the game clock of the period the shot was attempted.

```
#average makes by quarter
shot_dat_train %>%
  group_by(PERIOD) %>%
  summarize(shot_perc = mean(FGM)) %>%
  ggplot(aes(x = PERIOD, y = shot_perc)) +
  geom_bar(stat = "identity")
```



```
#shot percentage by minute of period
shot_dat_train %>%
  group_by(min) %>%
  summarize(shot_perc = mean(FGM)) %>%
  ggplot(aes(x = min, y = shot_perc)) +
  geom_bar(stat = "identity")
```




Looking at the NBA period, there isn't a whole lot of variation in shot percentages, though there does seem to be a slight dip starting in the fourth quarter. This may suggest that fatigue is a factor in a player's shot attempt. We also see that shot percentage dips even more in the first overtime period (period 5).

Looking at minutes remaining in the period, we also see shot percentage staying fairly consistent. The only inconsistencies occur towards the beginning of the period and towards the end of the period. I feel that the dip in shot percentage towards the last minute of a quarter makes sense as players may feel rushed or pressured to attempt a shot in order to beat the game clock from expiring. The dip in shot percentage at the start of the quarter could suggest players potentially starting to warm up before adjusting to the game.

## Opposing Team and Defender Analysis

Oftentimes when a player attempts a shot, they are being guarded by a defender. I want to see how the defender's distance and the actual defender affects the shot.

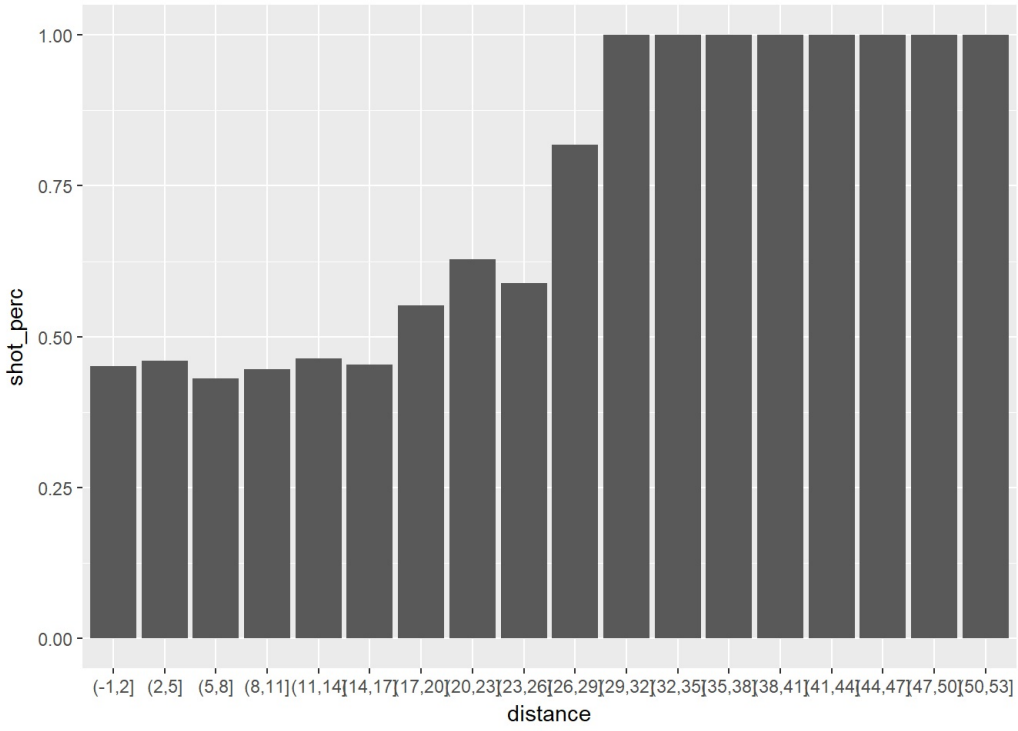
```
#opposing team
shot_dat_train %>%
  group_by(opposing_team) %>%
  summarize(shot_perc = mean(FGM)) %>%
  arrange(shot_perc) %>%
  head()
```

```
## # A tibble: 6 x 2
##   opposing_team shot_perc
##          <fctr>     <dbl>
## 1           GSW 0.4263974
## 2           OKC 0.4334079
## 3           IND 0.4349244
## 4           POR 0.4362287
## 5           MIL 0.4380336
## 6           ATL 0.4385343
```

```
#average closest defender distance on makes and misses
shot_dat_train %>%
  group_by(FGM) %>%
  summarize(avg_def_dist = mean(CLOSE_DEF_DIST))
```

```
## # A tibble: 2 x 2
##     FGM avg_def_dist
##   <int>        <dbl>
## 1     0     4.124268
## 2     1     4.118092
```

```
#grouping distances by increments of 3 feet
shot_dat_train %>%
  group_by(distance = cut(CLOSE_DEF_DIST, breaks = seq(-1, 57, by = 3))) %>%
  summarize(shot_perc = mean(FGM)) %>%
  ggplot(aes(x = distance, y = shot_perc)) +
  geom_bar(stat = "identity")
```




```
#average shot percentage: .452
mean(shot_dat_train$FGM)
```

```
## [1] 0.4519838
```

```
#top defenders
shot_dat_train %>%
  group_by(CLOSEST_DEFENDER) %>%
  summarize(count = n(),
            shot_perc = mean(FGM)) %>%
  filter(count >= 150) %>%
  arrange(shot_perc) %>% head()
```

```
## # A tibble: 6 x 3
##     CLOSEST_DEFENDER count shot_perc
##                <chr> <int>     <dbl>
## 1    Jones, Terrence   164 0.3536585
## 2 Galloway, Langston   152 0.3552632
## 3      Grant, Jerami   205 0.3560976
## 4       Cole, Norris   215 0.3581395
## 5  Livingston, Shaun   207 0.3671498
## 6        Allen, Tony   261 0.3678161
```

Looking at opposing teams, we see that the Golden State Warriors held opponents to the worst shooting percentage. This is probably one of many reasons why the Warriors won the NBA Finals that season.

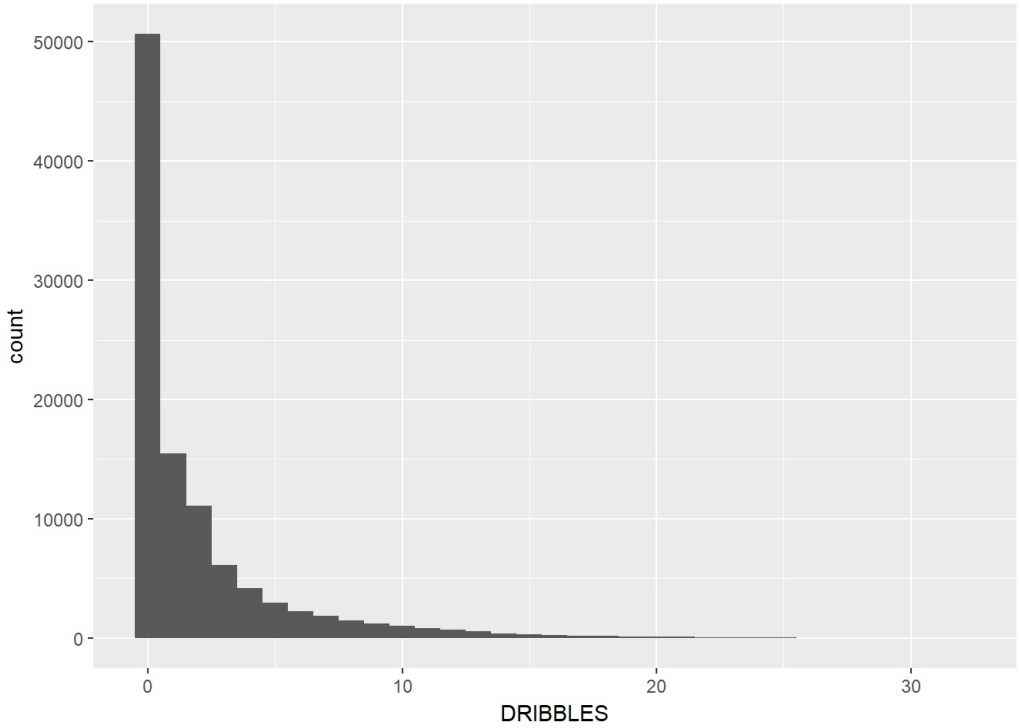
I found it interesting how the average defender distance between a made shot and missed shot was virtually the same. To investigate more, I decided to create the distribution of shot percentages based on defender distances in increments of 3 feet. Looking at this distribution, it looks like shot percentages do not seem to increase until defenders are at least 17 feet away.

Next I wanted to see the defenders who caused the worst shot percentages from the shooter. I restricted this criteria only to defenders who have defended at least 150 shot attempts. Judging from the top defenders in the resulting table, it suggests that a the defender himself does have an impact on a shooter. The top 6 defenders from my criteria allow a shot percentage much lower than the league average shot percentage (.452).

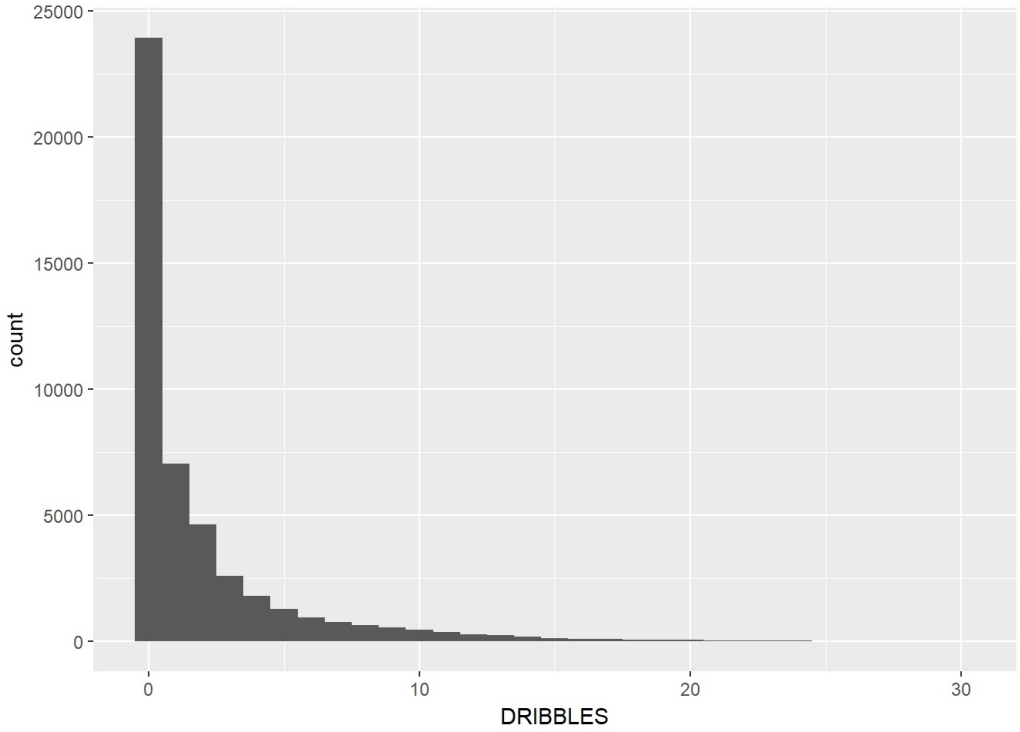
## Other Findings

Below are some visualizations of distributions of number of dribbles before a shot was taken as well as how the shot number of the attempt affects shot percentage.

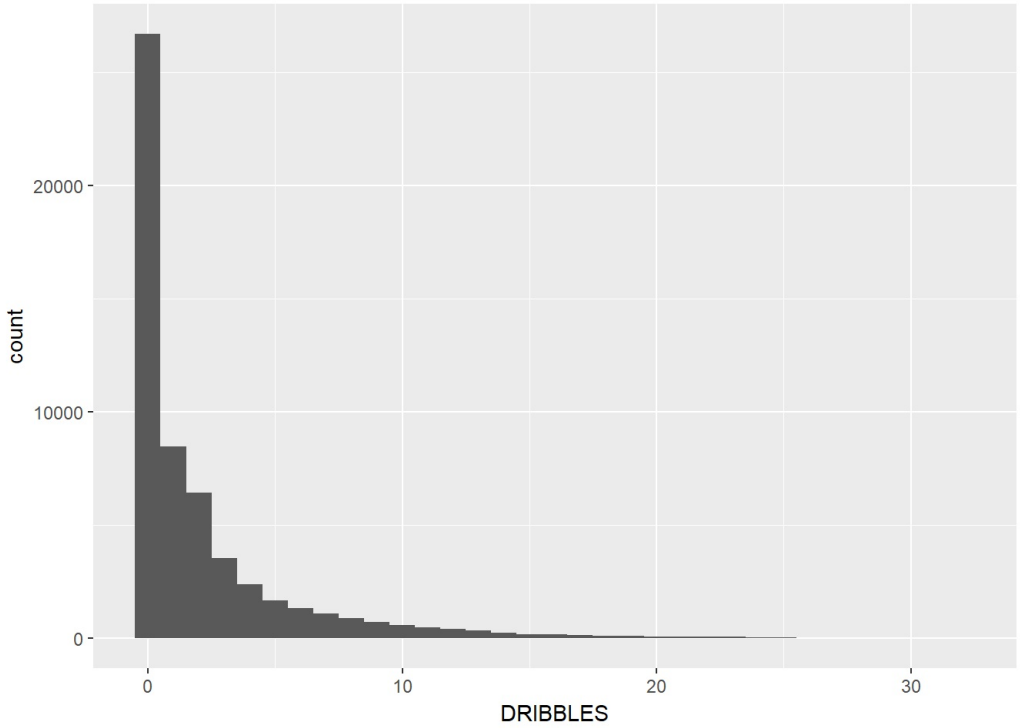
```
#dribbles distribution
shot_dat_train %>%
  ggplot(aes(x = DRIBBLES)) +
  geom_histogram(binwidth = 1)
```




```
#makes
shot_dat_train %>%
  filter(FGM == 1) %>%
  ggplot(aes(x = DRIBBLES)) +
  geom_histogram(binwidth = 1)
```

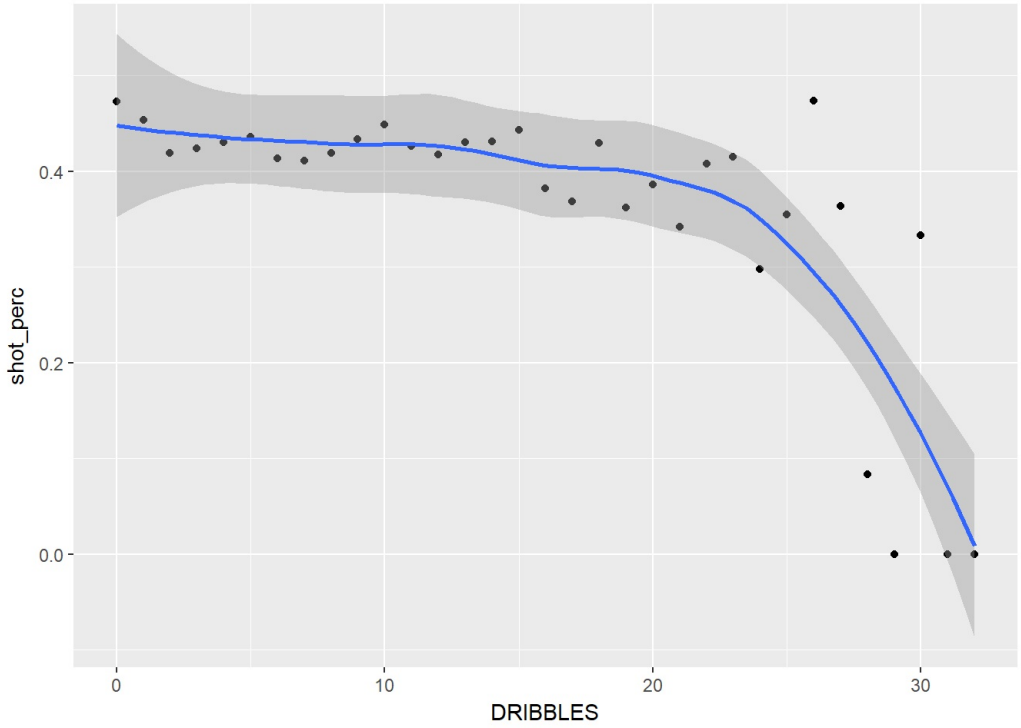



```
#misses
shot_dat_train %>%
  filter(FGM == 0) %>%
  ggplot(aes(x = DRIBBLES)) +
  geom_histogram(binwidth = 1)
```




```
#plotting shot percentage vs dribbles
shot_dat_train %>%
  group_by(DRIBBLES) %>%
  summarize(shot_perc = mean(FGM)) %>%
  ggplot(aes(x = DRIBBLES, y = shot_perc)) +
  geom_point() +
  geom_smooth()
```

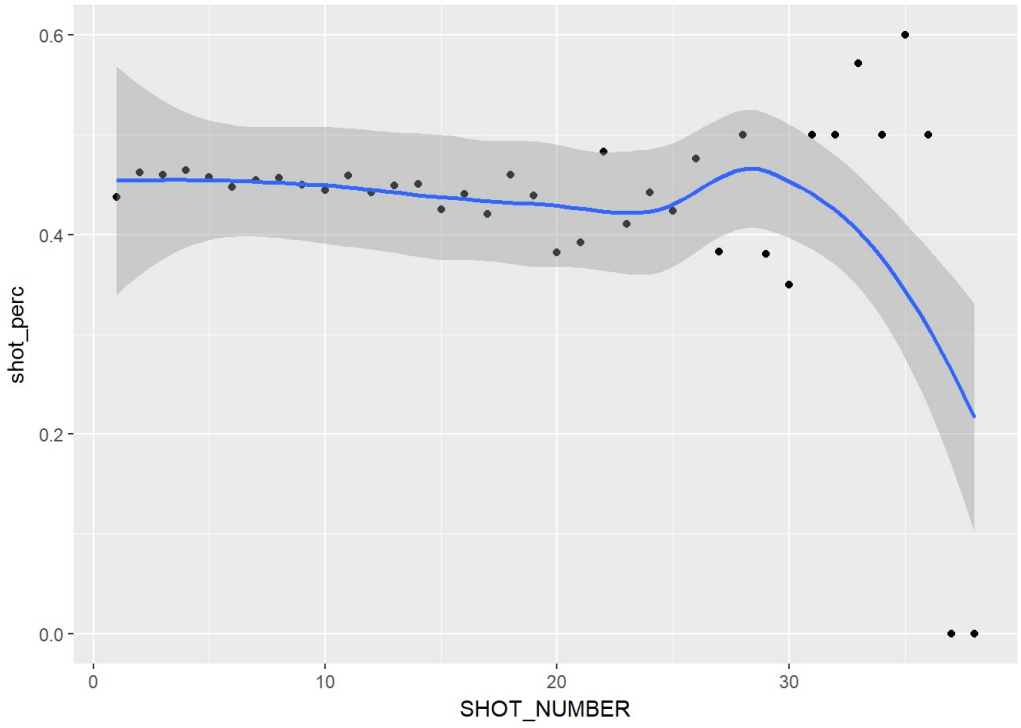
```
## `geom_smooth()` using method = 'loess'
```




Looking at the distributions above, there doesn't seem to be a noticeable difference. All I can gather is that the vast majority of shots come from very few dribbles. When looking at the shot percentage vs dribbles plot, it does look like shot percentage starts to noticeably dip at around 20 dribbles.

```
#shot number and average makes
shot_dat_train %>%
  group_by(SHOT_NUMBER) %>%
  summarize(shot_perc = mean(FGM)) %>%
  ggplot(aes(x = SHOT_NUMBER, y = shot_perc)) +
  geom_point() +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'loess'
```




The shot number of a player's shot attempt does not seem to affect shot percentage a whole lot. There seems to be a slight increase at around 25 attempts, but not many players will reach 25 shots in a game, and if they do, it's most likely because they are shooting really well that particular game.

# Summary and Next Steps

Overall, the most interesting findings to me was observing a fairly normal distribution in the shot clock time at the shot attempt, a very noticeable bimodal distribution in shot distances, the duration of a game affecting shot percentage, and the effect opposing teams and opposing defenders have on a shot attempt. When taking the next step in creating different models, I will keep the knowledge I have learned from this EDA in mind.