MATH411 | Fall 2018 | Homework 2 (Due: Friday in class, 9/28/2018)

Your Name Here

Date Submitted Here

Problem 1 (Writing Function and Plotting Curve)

Imagine a monopolist selling a specific product with demand curve $Q(p)$, where $Q(p)$ is the quantity sold given a specific price p. To simplify things, let's suppose that $Q(p)$ is a linear function:

$$Q(p) = \alpha p + \beta$$

The total revenue will be given by:

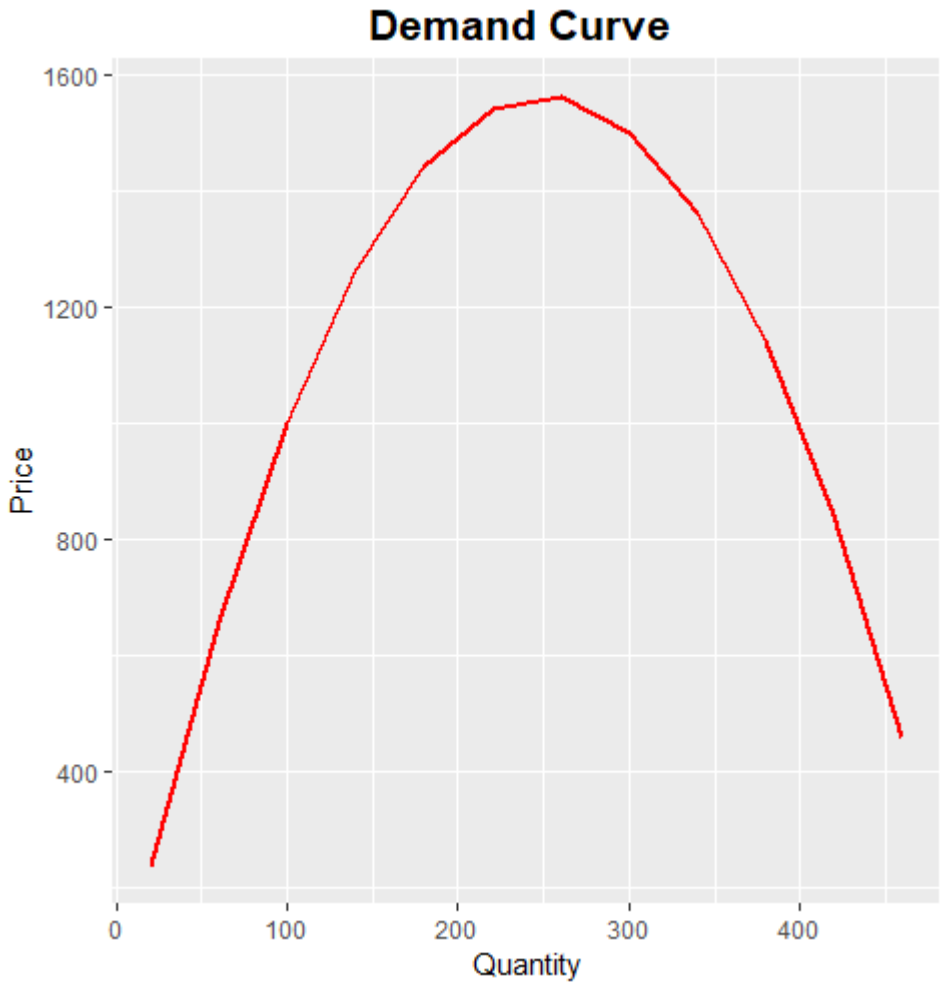
$$R(p) = pQ(p) = \alpha p^2 + p\beta$$

(a) Code $R(p)$ in R by using $\alpha = -40$ and $\beta = 500$.

(b) Plot $R(p)$ VS p for p between 1 and 12. (Make your graph as nice as possible)

```
library(tidyverse)
price = 1:12
alpha = -40
beta  = 500
Demand = function(p,a,b){
  demand = a*p + b
  return(demand)
}
TotalRev = function(Q,p){
  Revenue = p*Q
  return(Revenue)
}
tibble(x = Demand(price, alpha, beta),
       y = TotalRev(quantity, price)) %>%
  ggplot(aes(x = x, y = y)) +
  geom_line(size = 1, color = 'red')+
  theme(legend.position = "top",
        plot.caption = element_text(hjust = 0.5),
        plot.subtitle = element_text(face = "italic"),
        plot.title = element_text(size = 16, face = "bold", hjust = 0.5))+
  labs(x = "Quantity", y = "Price",
       title = "Demand Curve ")
```

problem 1 output:



Problem 2 (Categorical Variable vs. Numerical Variable)

Amazon's new headquarters Scrape the table (i.e., the `Twenty-six cities` data) from the `cbsnews` website at https://www.cbsnews.com/news/amazon-hq2-cities-location-choices-new-second-headquarters/. Tidy the data, then

(a) Print the First 5 and bottom 5 rows of your data.

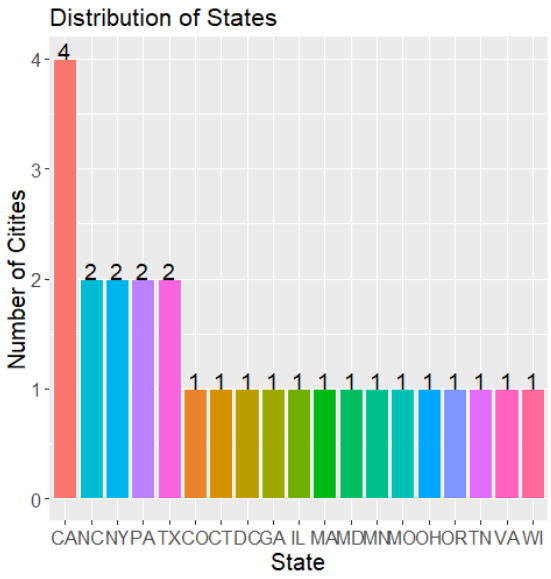
(b) Make a bar plot to show the distribution of states in the data, rank the states by the number of cities in it from highest to lowest.

(c) Make a horizontal bar plot of `Percent with bachelor's degree` VS `Metro area`. Rank the `Metro area` by their `Percent with bachelor's degree` and label the percentage, i.e., %, on top of each `Metro area`.

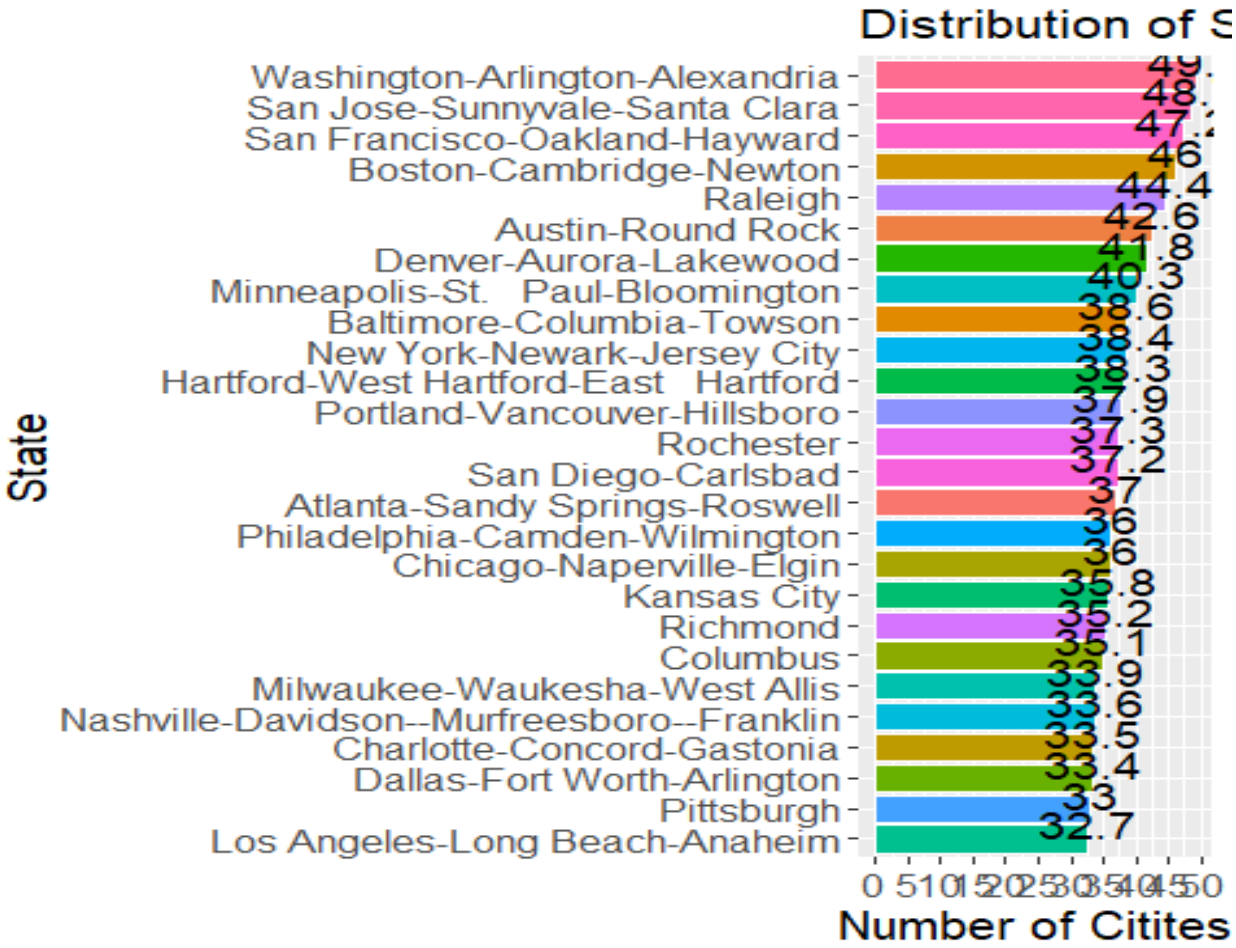
```
library(rvest)
library(stringr)
URL = "https://www.cbsnews.com/news/amazon-hq2-cities-location-choices-new-second-headquarters/"


cities = read_html(URL) %>%
  html_table() %>%
  .[[1]] %>%
  as_tibble()
colnames(cities) = c("Metro_Area", "State", "Pop_Total", "Bach_Deg_Percent")
cities = cities %>%
  slice(-1)
cities = cities %>%
  mutate(Pop_Total = parse_number(Pop_Total),
         Bach_Deg_Percent = as.numeric(Bach_Deg_Percent))
  # Reading the data:
  cities %>% colnames()
  cities
  cities %>% head (5)
  cities %>% tail(5)
cities %>%
  count(State) %>%
  rename(freq = n) %>%
  ggplot(aes(x = reorder(State, -freq), y = freq, fill = State)) +
  geom_bar(stat = "identity", color = "white") +
  scale_y_continuous(breaks = seq(0, 5, 1)) +
  guides(fill = FALSE) +
```

```
 geom_text(aes(label = freq),
           vjust = 0,
           color = "black",
           size = 5) +
 labs(title = "Distribution of States",
      x    = "State",
      y    = "Number of Cities") +
 theme(axis.text.x = element_text(size = 12),
       axis.text.y = element_text(size = 12),
       axis.title.x = element_text(size = 15),
       axis.title.y = element_text(size = 15),
       plot.title = element_text(hjust = 0, size = 16),
       plot.subtitle = element_text(hjust = 0, size = 12))
```

problem 2 output 1



Problem 2 output 2



Problem 3 (Categorical Variables and Numerical Variables)

Pittsburgh Penguins Scrape the Pittsburgh Penguins' `Team Record By Season` data from https://www.nhl.com/penguins/team/season-by-season-record.

(a) Seperate the last column `Finish` into two columns `rank` and `region`. (Hint: YOu can use the `separate` function from `tidyverse`). Delete the `NA` values in your data (Hint: you need to delete the 2004-05 season). Then print out the first 5 and bottom 5 rows.

(b) Make a bar plot to show the distribution of `rank` over all the seasons. Rank the `rank` by its frequency from highest to lowest.

(c) Plot the distribution of `GF` (Goals for, ie.e, goals scored by the Penguins).

(d) Make a new variable called `win.probability`, which can be calculated by $\frac{W}{GP}$ (i.e., numberf of game wins divided by number of game played). Plot the density distribution of `win.probability` and hightlight the mean of `win.probability` on the density plot as a big point.

(e) Make another variable called `GFpergame` (GF/GP). Then make a scatter plot between `win.probability` and `GFpergame`. Comment on the pattern you can tell from the scatter plot.

(f) Make another variable called `GApergame` (GA/GP). Then make a scatter plot between `win.probability` and `GApergame`. Comment on the pattern you can tell from the scatter plot.

(g) Make a scatter plot between `GFpergame` and `GApergame`. Comment on the pattern you can tell from the scatter plot.

```
cities %>%
 ggplot(aes(x = reorder(Metro_Area, Bach_Deg_Percent),
        y = Bach_Deg_Percent, fill = Metro_Area)) +
 geom_bar(stat = "identity", color = "white") +
 scale_y_continuous(breaks = seq(0, 50, 5)) +
 guides(fill = FALSE) +
 geom_text(aes(label = paste0(round(Bach_Deg_Percent, 2))),
        vjust = 0,
```

```
        color = "black",
        size = 5) +
 labs(title = "Distribution of States",
     x    = "State",
     y    = "Number of Citites") +
 theme(axis.text.x = element_text(size = 12),
      axis.text.y = element_text(size = 12),
      axis.title.x = element_text(size = 15),
      axis.title.y = element_text(size = 15),
      plot.title = element_text(hjust = 0, size = 16),
      plot.subtitle = element_text(hjust = 0, size = 12))+
 coord_flip()
```

```
### Problem 3: Categorical Variable vs. Numerical Variable
################################################
```

```
#     Pittsburgh Penguins Scrape the Pittsburgh Penguins' Team Record By Season data
from
#     https://www.nhl.com/penguins/team/season-by-season-record.
```

```
penguins = read_html("https://www.nhl.com/penguins/team/season-by-season-record")
```

```
tbl = penguins %>%
 html_table(fill = TRUE) %>%
 .[[1]] %>%
```

```
 .[-1,]

tbl = tbl %>%
  separate("Finish", into = c("Rank", "Region"))



# 1.  Seperate the last column Finish into two columns rank and region. (Hint: YOu can
#     use the separate function from tidyverse). Delete the NA values in your data (Hint:
#     you need to delete the 2004-05 season). Then print out the first 5 and bottom 5 rows.

penguins %>% head (5)
penguins %>% tail(5)

# 2.  Make a bar plot to show the distribution of rank over all the seasons. Rank the rank
#     by its frequency from highest to lowest.

tbl %>%
  count(Rank) %>%
  rename(freq = n) %>%
  ggplot(aes(x = Rank, y = freq, fill = Rank)) +
  geom_bar(stat = "identity", color = "white") +
  scale_y_continuous(breaks = seq(0, 15, 1)) +
  guides(fill = FALSE) +
  geom_text(aes(label = freq),
            vjust = 1.5,
            color = "black",
            size = 5) +
```

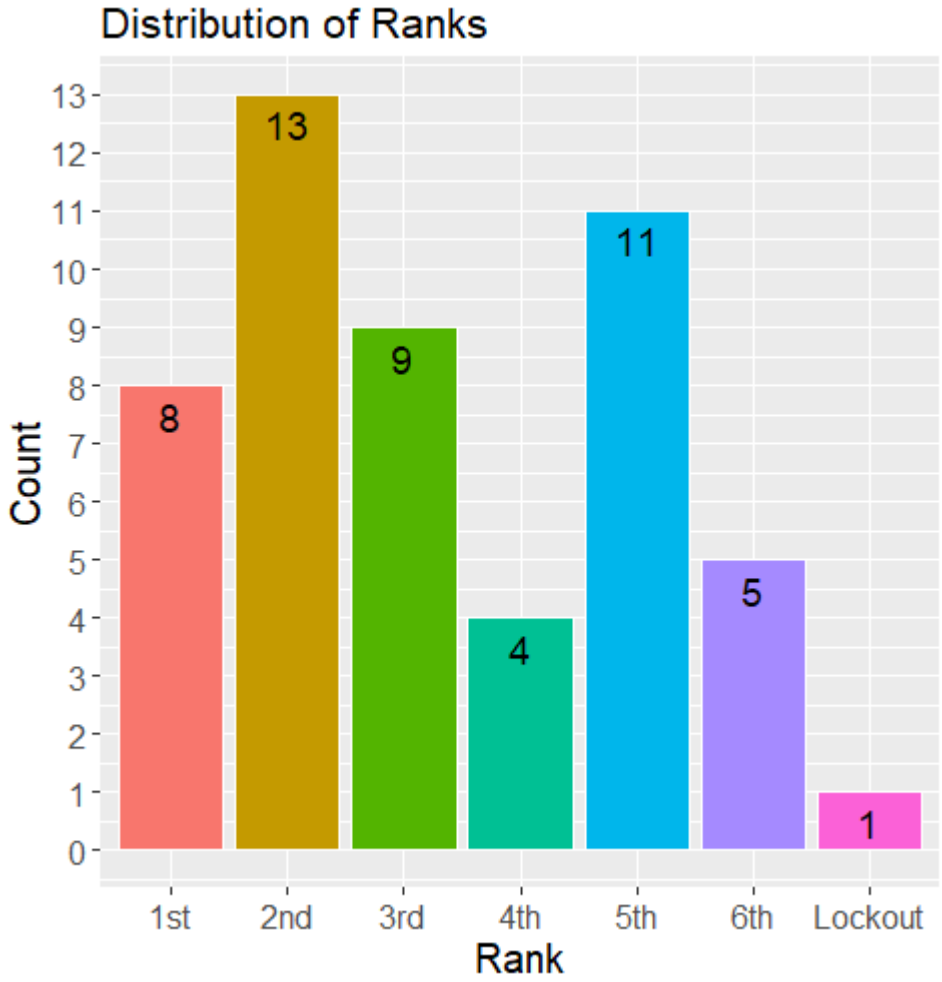
```
  labs(title = "Distribution of Ranks",
       x    = "Rank",
       y    = "Count") +
  theme(axis.text.x = element_text(size = 12),
        axis.text.y = element_text(size = 12),
        axis.title.x = element_text(size = 15),
        axis.title.y = element_text(size = 15),
        plot.title = element_text(hjust = 0, size = 16),
        plot.subtitle = element_text(hjust = 0, size = 12))

# 3.  Plot the distribution of GF (Goals for, ie.e, goals scored by the Penguins).

tbl %>%
  ggplot() +
  geom_histogram(aes(x = GF, y = ..density..), fill = "blue") +
  geom_density(aes(GF), color = "orange", size = 1) +
  scale_x_continuous(breaks = seq(160, 370, 25)) +
  labs(title = "Distribution of Goals by Penguins",
       x    = "GF",
       y    = "Density") +
  theme(axis.text.x = element_text(size = 12),
        axis.text.y = element_text(size = 12),
        axis.title.x = element_text(size = 15),
        axis.title.y = element_text(size = 15),
        plot.title = element_text(hjust = 0, size = 16),
        plot.subtitle = element_text(hjust = 0, size = 12))
```

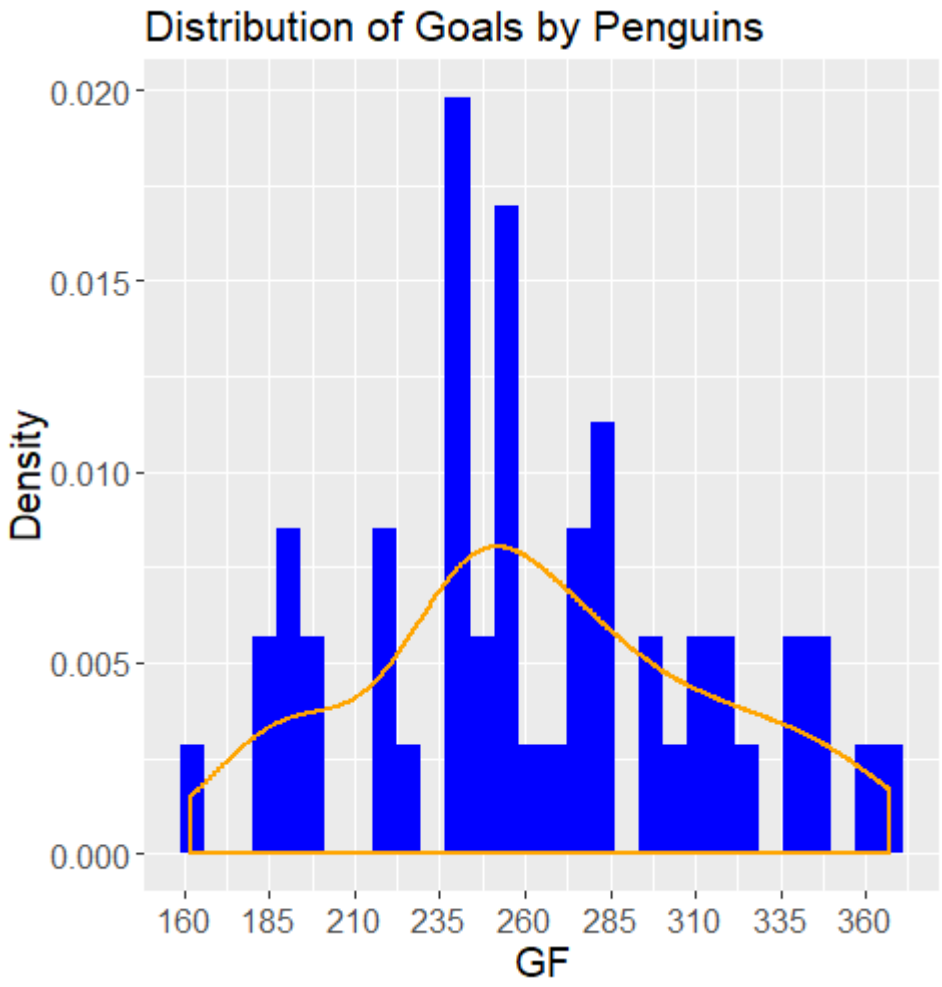
```
# 4.  Make a new variable called win.probability, which can be calculated by W
#     GP (i.e., numberf of game wins divided by number of game played).
#     Plot the density distribution of win.probability and hightlight the mean of
win.probability on the density plot
#     as a big point.

tbl = tbl %>% mutate(win.probability = W/GP) %>% drop_na()

tbl %>%
  ggplot() +
  geom_histogram(aes(x = win.probability, y = ..density..), fill = "blue") +
  geom_density(aes(win.probability), color = "orange", size = 1) +
  scale_x_continuous(breaks = seq(0.2, 0.8, 0.05)) +
  geom_point(aes(x = mean(tbl$win.probability), y=2.6), size = 5, color = "black")
  labs(title = "Distribution of Winning probability in a single game",
       x    = "Probability",
       y    = "Density") +
  theme(axis.text.x = element_text(size = 12),
        axis.text.y = element_text(size = 12),
        axis.title.x = element_text(size = 15),
        axis.title.y = element_text(size = 15),
        plot.title = element_text(hjust = 0, size = 16),
        plot.subtitle = element_text(hjust = 0, size = 12))

# 5.  Make another variable called GFpergame (GF/GP). Then make a scatter plot between
```

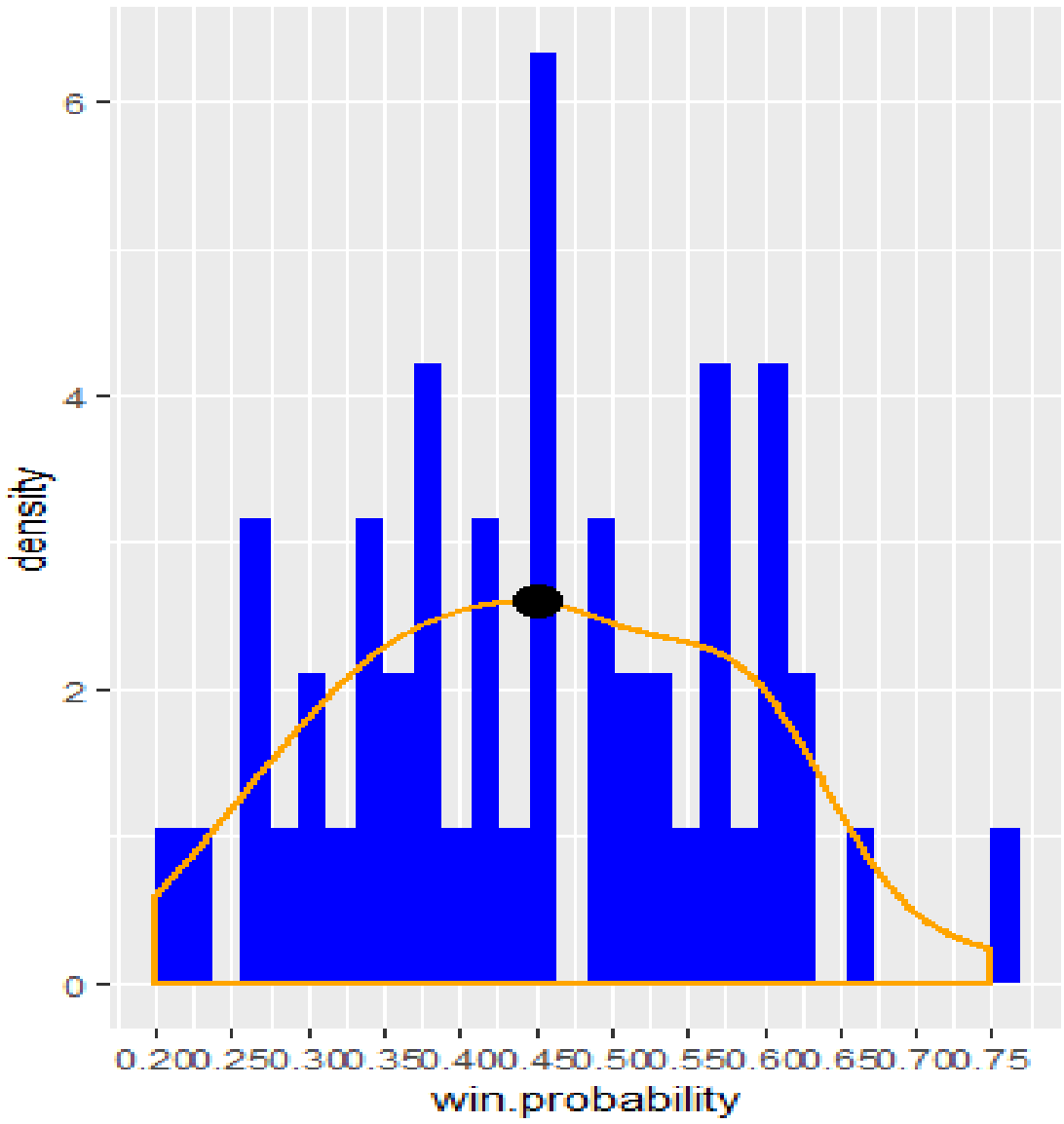
```
#     win.probability and GFpergame. Comment on the pattern you can tell from the
#     scatter plot.

tbl = tbl %>% mutate(GFpergame = GF/GP)

tbl %>%
 ggplot(aes(x = GFpergame, y = win.probability)) +
 geom_point(size = 5) +
 labs(title = "Relationship between GF and win Probability",
      x    = "GF per game",
      y    = "Win Probability") +
 theme(axis.text.x = element_text(size = 12),
       axis.text.y = element_text(size = 12),
       axis.title.x = element_text(size = 15),
       axis.title.y = element_text(size = 15),
       plot.title = element_text(hjust = 0, size = 16),
       plot.subtitle = element_text(hjust = 0, size = 12))



# 6.  Make another variable called GApergame (GA/GP). Then make a scatter plot between
#     win.probability and GApergame. Comment on the pattern you can tell from the
#     scatter plot.

tbl = tbl %>% mutate(GApergame = GA/GP)

tbl %>%
```

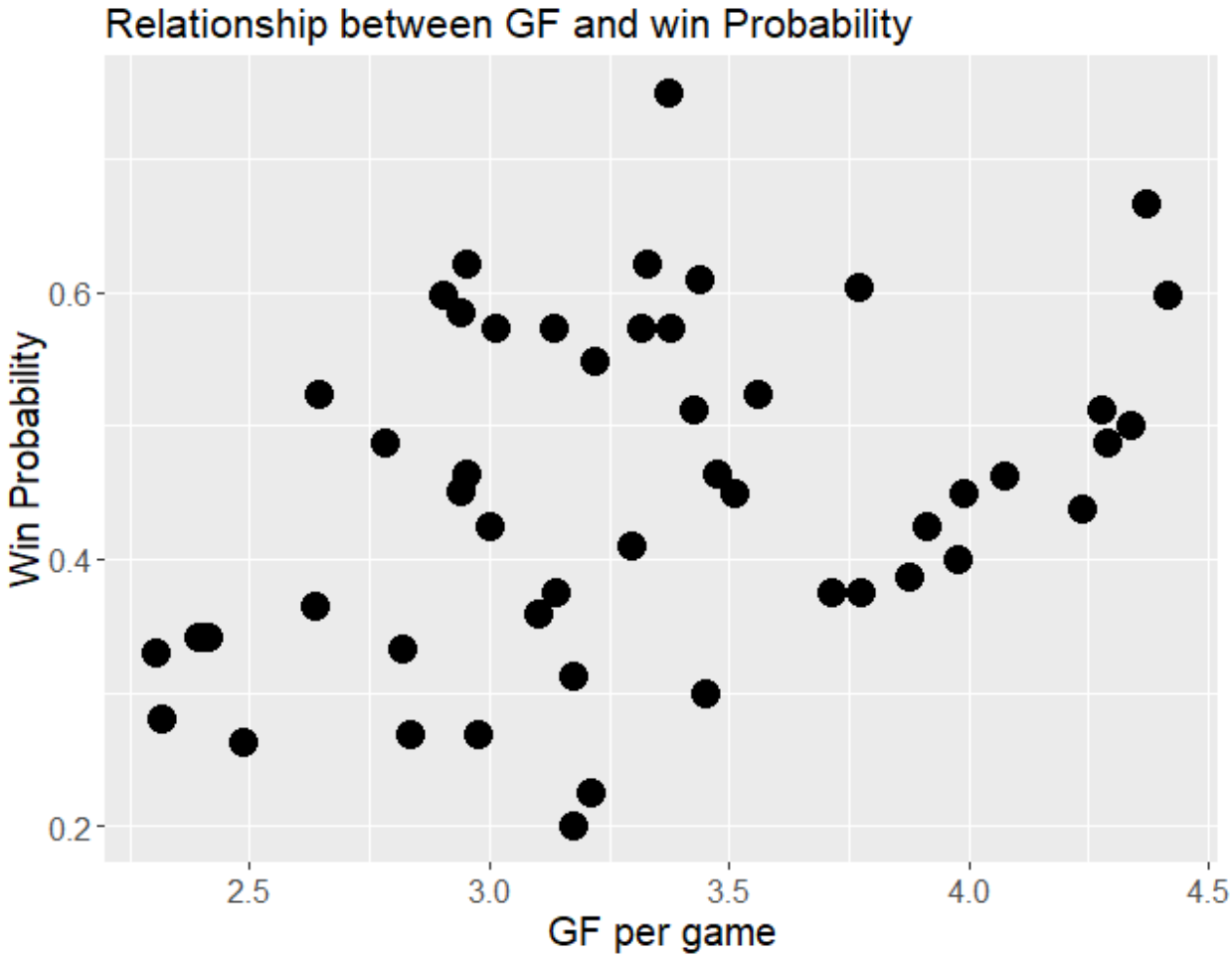
```
 ggplot(aes(x = GApergame, y = win.probability)) +
 geom_point(size = 5) +
 labs(title = "Relationship between GA and win Probability",
     x    = "GA per game",
     y    = "Win Probability") +
 theme(axis.text.x = element_text(size = 12),
       axis.text.y = element_text(size = 12),
       axis.title.x = element_text(size = 15),
       axis.title.y = element_text(size = 15),
       plot.title = element_text(hjust = 0, size = 16),
       plot.subtitle = element_text(hjust = 0, size = 12))

# 7.  Make a scatter plot between GFpergame and GApergame. Comment on the pattern you
#     can tell from the scatter plot.


tbl %>%
 ggplot(aes(x = GFpergame, y = GApergame)) +
 geom_point(size = 5) +
 labs(title = "Relationship between GF and GA",
     x    = "GF per game",
     y    = "GA per game") +
 theme(axis.text.x = element_text(size = 12),
       axis.text.y = element_text(size = 12),
       axis.title.x = element_text(size = 15),
       axis.title.y = element_text(size = 15),
       plot.title = element_text(hjust = 0, size = 16),
```

plot.subtitle = element_text(hjust = 0, size = 12))

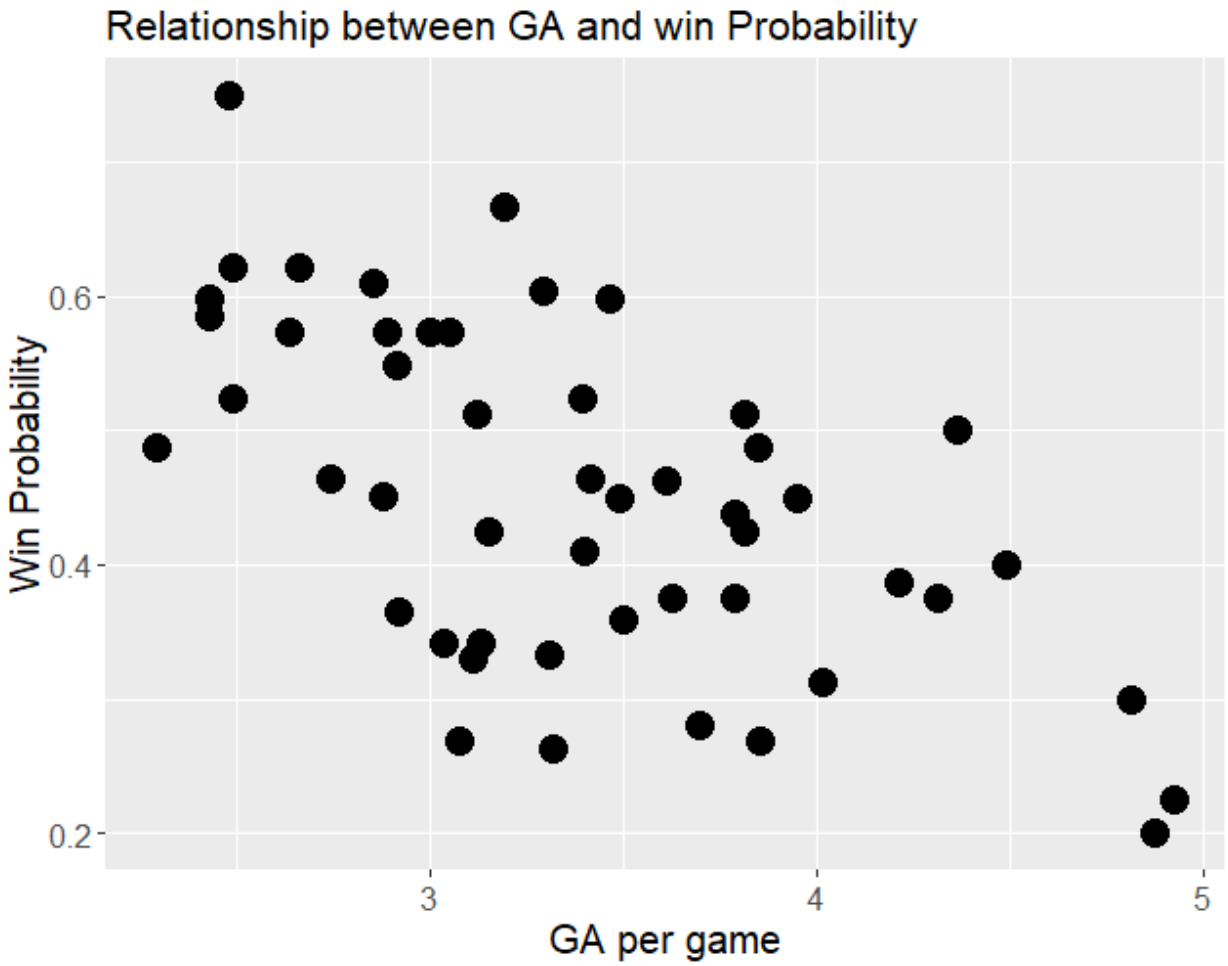



Distribution of Goals by Penguins
0.020
0.015
0.010
0.005
0.000
Density
160
185
210
235
260
285
310
335
360
GF


density
6
4
2
0
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
0.75
win.probability

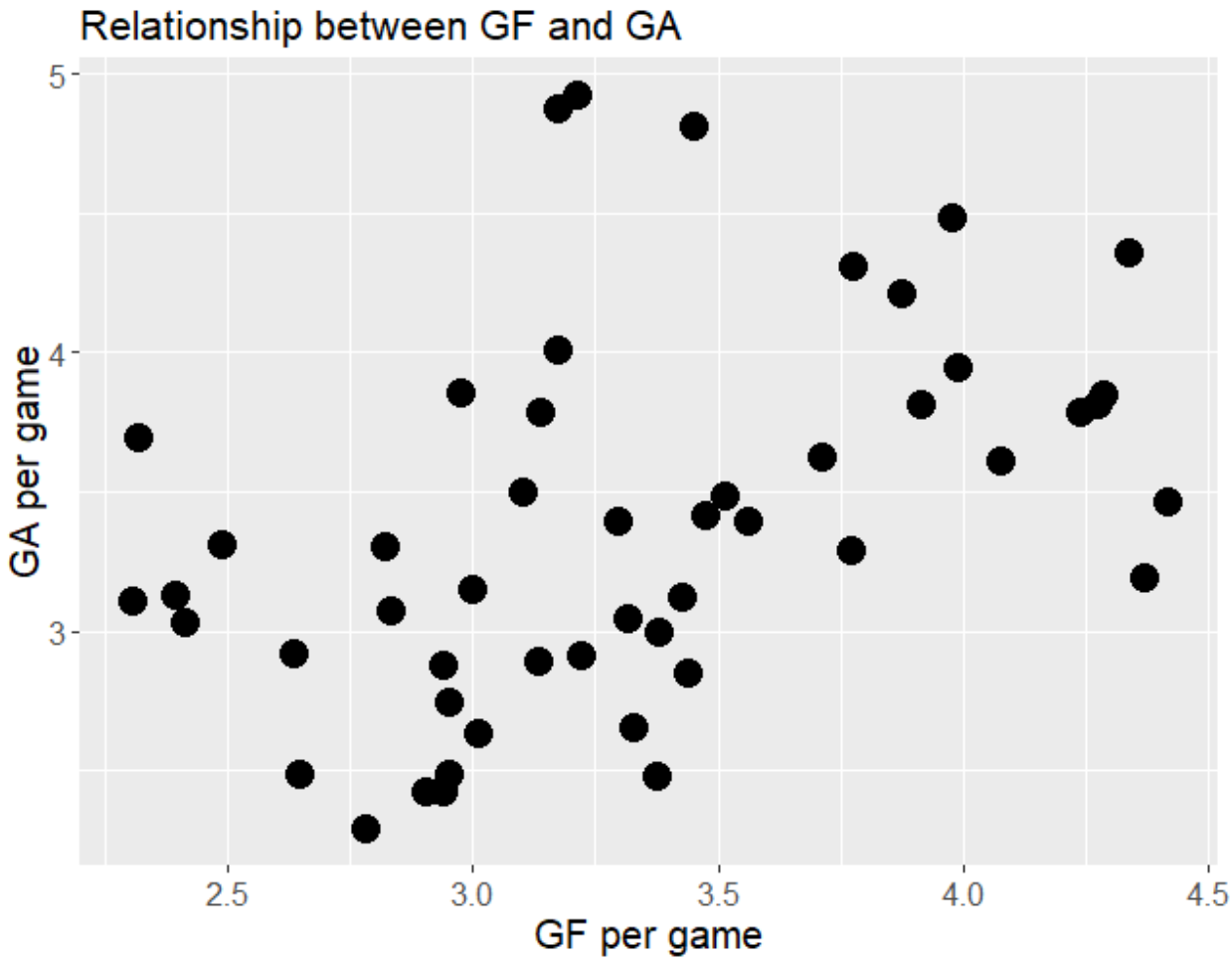


There seems to be a positive correlation between win probability and GF.



There seems to be a negative correlation between win probability and GA

There seems to be a positive correlation between win probability and GF.



There seems to be a positive correlation between GA and GF.