Homework 5

Kevin Akers

September 25, 2018

Problem 3

A good figure should have clearly labeled axes. Using different colors or shading can also make it easier to illustrate what you're trying to convey with the figure.

Problem 4

```
success <- function(x){
  sum(x)/length(x)
}

set.seed(12345)
P4b_data <- matrix(rbinom(10, 1, prob = (30:40)/100), nrow = 10, ncol = 10)

#Proportion of successes in each column
apply(P4b_data, 2, success)
```

```
##  [1] 0.6 0.6 0.6 0.6 0.6 0.6 0.6 0.6 0.6 0.6
```

```
#Proportion of successes in each row
apply(P4b_data, 1, success)
```

```
##  [1] 1 1 1 1 0 0 0 0 1 1
```

All the columns are the same, while the rows are either a 0 or a 1.

```
prob <- function(x){
  rbinom(10, 1, prob = x)
}

prob_vector <- (30:40)/100

prob_matrix <- sapply(prob_vector, prob)

apply(prob_matrix, 2, success)
```

```
##  [1] 0.2 0.3 0.4 0.3 0.4 0.6 0.3 0.3 0.5 0.6 0.5
```

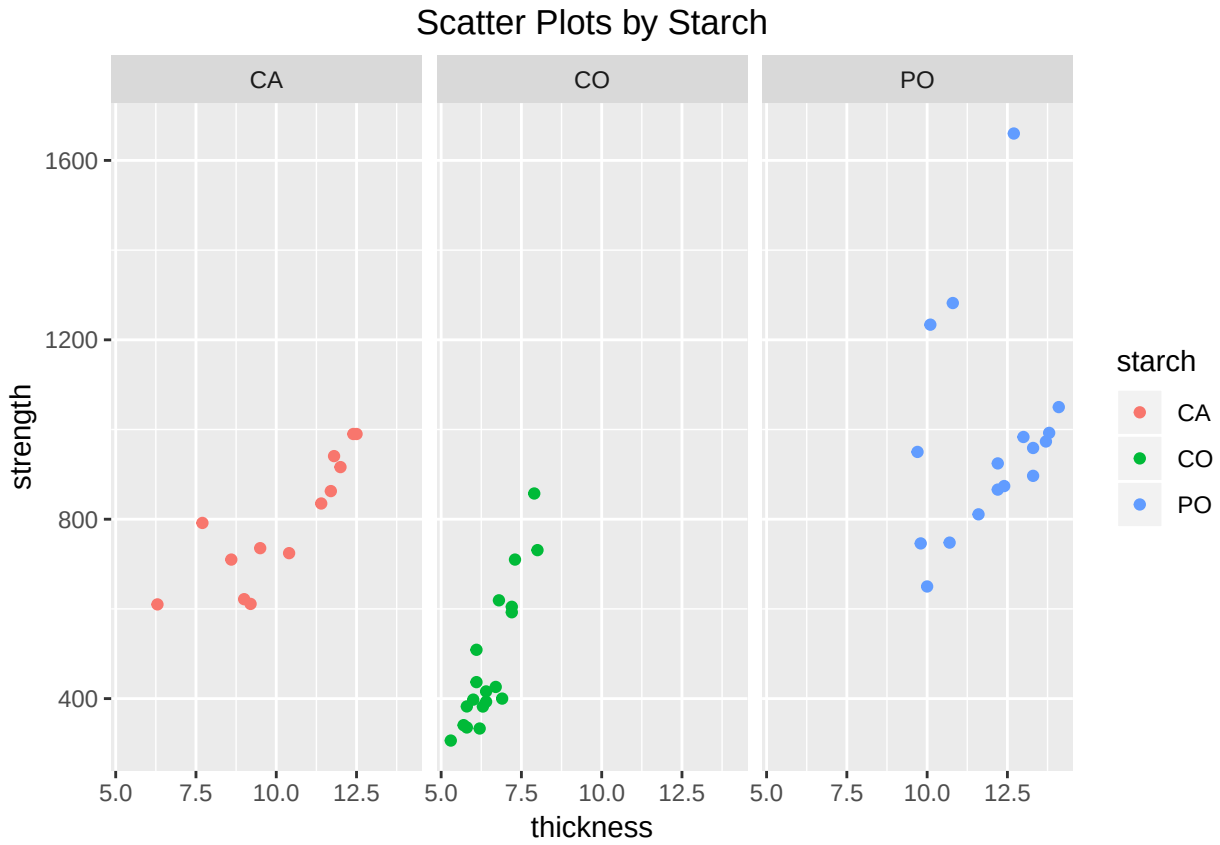
Problem 5

```
url <- "http://www2.isye.gatech.edu/~jeffwu/book/data/starch.dat"
starch <- read.table(url, header = T)
knitr::kable(summary(starch), caption = "Data Summary")
```

Table 1: Data Summary

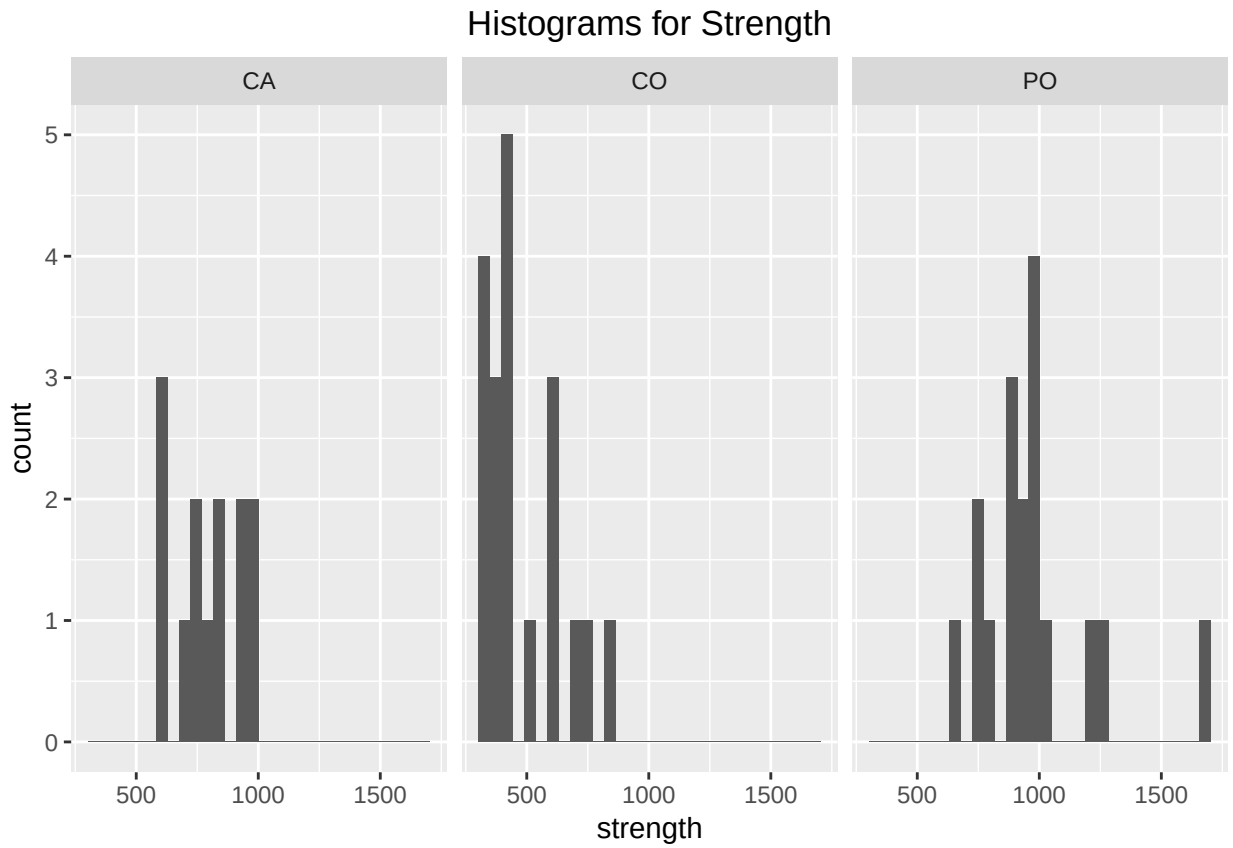
starch	strength	thickness
CA:13	Min. : 306.4	Min. : 5.300
CO:19	1st Qu.: 508.8	1st Qu.: 6.700
PO:17	Median : 735.4	Median : 9.500
NA	Mean : 737.0	Mean : 9.388
NA	3rd Qu.: 924.4	3rd Qu.:12.000
NA	Max. :1660.0	Max. :14.100

```
#First we graph strength vs thickness divided up by starch.
ggplot(data = starch, aes(x = thickness, y = strength, color = starch)) +
  geom_point() +
  facet_wrap(~starch) +
  ggtitle("Scatter Plots by Starch") +
  theme(plot.title = element_text(hjust = 0.5))
```



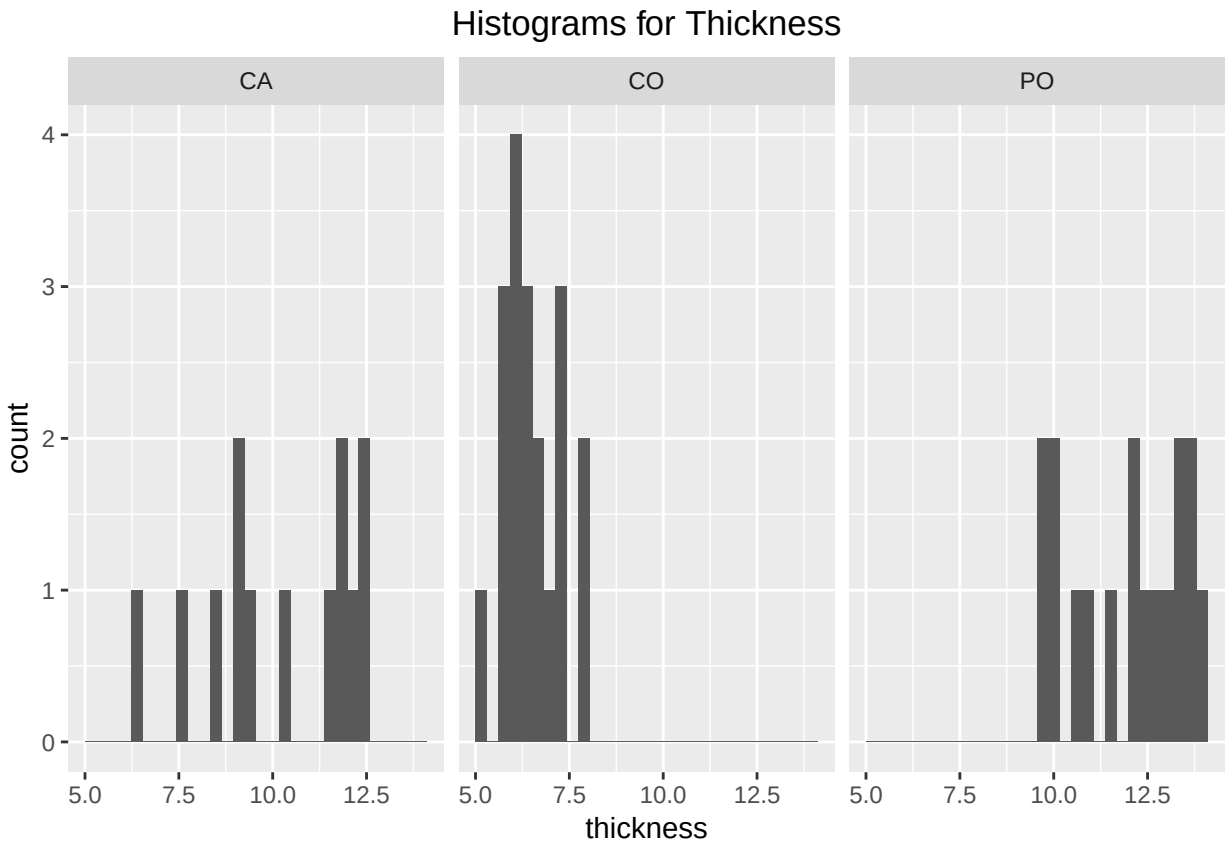

```
#Then we produce histograms for strength and thickness, respectively.
ggplot(data = starch, aes(x = strength)) +
  geom_histogram() +
  facet_wrap(~starch) +
  ggtitle("Histograms for Strength") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(data = starch, aes(x = thickness)) +
  geom_histogram() +
  facet_wrap(~starch) +
  ggtitle("Histograms for Thickness") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Problem 6

```
download("http://www.farinspace.com/wp-content/uploads/us_cities_and_states.zip",
         dest = "us_cities_states.zip")
unzip("us_cities_states.zip", exdir = "./")
states <- fread(input = "./us_cities_and_states/states.sql",
                sep = "'", sep2 = ",", header = F, select = c(2, 4))
colnames(states) <- c("state_name","state")
cities <- fread(input = "./us_cities_and_states/cities_extended.sql", header = F,
                sep = "'", sep2 = ",", select = c(2, 4, 6, 8, 10, 12))
colnames(cities) <- c("city","state","zip","lat","long","county")
unique_cities <- fread(input = "./us_cities_and_states/cities.sql", header = F,
                sep = "'", sep2 = ",", select = c(2, 4))
colnames(unique_cities) <- c("city","state")

#Part b.
count_cities <- unique_cities %>% group_by(state) %>%
                summarise(count = n()) %>%
                arrange(state)
count_cities1 <- cbind(count_cities[1:10,], count_cities[11:20,],
                       count_cities[21:30,], count_cities[31:40,], count_cities[41:50,])
colnames(count_cities1) <- paste(rep(c("State", "Cities"), 5))
kable(count_cities1, caption = "Number of Cities in Each State")
```

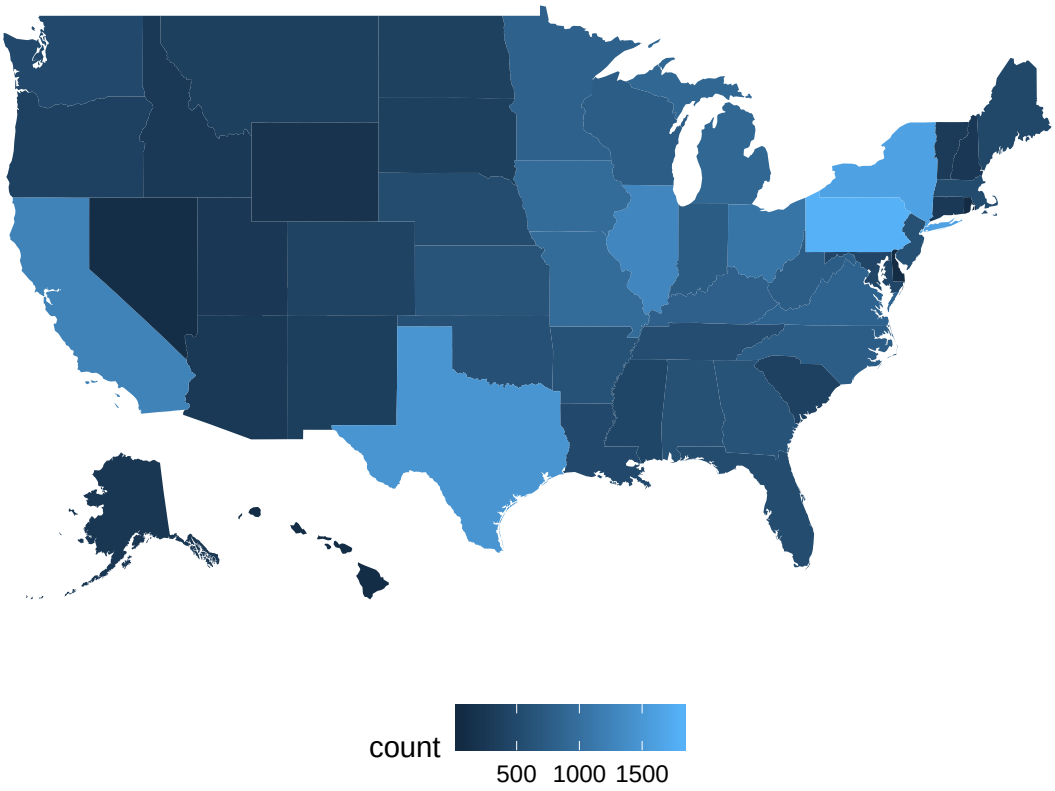
Table 2: Number of Cities in Each State

State	Cities	State	Cities	State	Cities	State	Cities	State	Cities
AK	229	GA	629	MD	430	NH	255	RI	70
AL	579	HI	92	ME	461	NJ	579	SC	377
AR	605	IA	937	MI	885	NM	346	SD	364
AZ	264	ID	266	MN	810	NV	99	TN	548
CA	1239	IL	1287	MO	942	NY	1612	TX	1466
CO	400	IN	738	MS	440	OH	1069	UT	250
CT	269	KS	634	MT	360	OK	585	VA	839
DC	3	KY	803	NC	762	OR	379	VT	288
DE	57	LA	479	ND	373	PA	1802	WA	493
FL	524	MA	511	NE	528	PR	99	WI	753

```
#Part c.
letter_occ <- function(letter, state_name){
  letter_vec <- unlist(strsplit(state_name,""))
  sum(letter_vec == letter)
}

letter_count <- data.frame(matrix(NA, nrow = nrow(states), ncol = 26))

for(i in 1:50){
  letter_count[i,] <- sapply(letters, FUN = letter_occ, state = tolower(states[i,1]))
}

data("fifty_states")
crimes <- data.frame(state = tolower(rownames(USArrests)), USArrests)
# map_id creates the aesthetic mapping to the state name
# column in your data
# create the datasets needed for the city counts map
map_data1 <- count_cities %>% left_join(states) %>%
  na.omit() %>% filter(state != "DC") %>%
  mutate(state_name = tolower(state_name)) %>%
  select(state_name, count)
```

```
## Joining, by = "state"
```

```
p1 <- ggplot(map_data1, aes(map_id = state_name)) +
  geom_map(aes(fill = count), map = fifty_states) +
  expand_limits(x = fifty_states$long,
                y = fifty_states$lat) + coord_map() +
  scale_x_continuous(breaks = NULL) + scale_y_continuous(breaks = NULL) +
  labs(x = "", y = "") +
  theme(legend.position = "bottom", panel.background = element_blank(),
        plot.title = element_text(hjust = 0.5))
p1
```


count
500
1000
1500