Homework5

Yueyao Wang

9/30/2017

Problem 3

I think a good figure should be:

- Focus on one or several related aspects of data. I think trying to put too much information into one plot doesn't make interpretation easy.
- Label the key elements in the plot clearly. Try to put a descriptive title and provide legend for your encoding of the data.
- Use appropriate scale to capture all information in the data.

Problem 4

Table 1: Apply by Row

p_0.31	p_0.32	p_0.33	p_0.34	p_0.35	p_0.36	p_0.37	p_0.38	p_0.39	p_0.4
0	1	0	0	0	0	1	0	0	0

Table 2: Apply by Col

p_0.31	p_0.32	p_0.33	p_0.34	p_0.35	p_0.36	p_0.37	p_0.38	p_0.39	p_0.4
0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2

Table 3: First 5 Columns of Matrix in Part b

	flip1	flip2	flip3	flip4	flip5
p_0.31	0	0	0	0	0
p_0.32	1	1	1	1	1
p_0.33	0	0	0	0	0
p_0.34	0	0	0	0	0
p_0.35	0	0	0	0	0
p_0.36	0	0	0	0	0
p_0.37	1	1	1	1	1
p_0.38	0	0	0	0	0
p_0.39	0	0	0	0	0
p_0.4	0	0	0	0	0

From above tables we can see that if apply by row the proportion we calculated using matrix in Part b are either 0 or 1 and if we apply by column, the proportion are all the same. If we look at the columns of the matrix created in Part b, we can find that all columns are the same. So the matrix in Part b didn't generate 10 flips of a coin with various p. It generate 1 flip of a coin with 10 probabilities and then repeated that column 10 times. The function in Part d fixed this error.

Table 4: Success Proportion in Part d

p_0.31	p_0.32	p_0.33	p_0.34	p_0.35	p_0.36	p_0.37	p_0.38	p_0.39	p_0.4
0.2	0.4	0.5	0.1	0.4	0.7	0.2	0.2	0.6	0.1

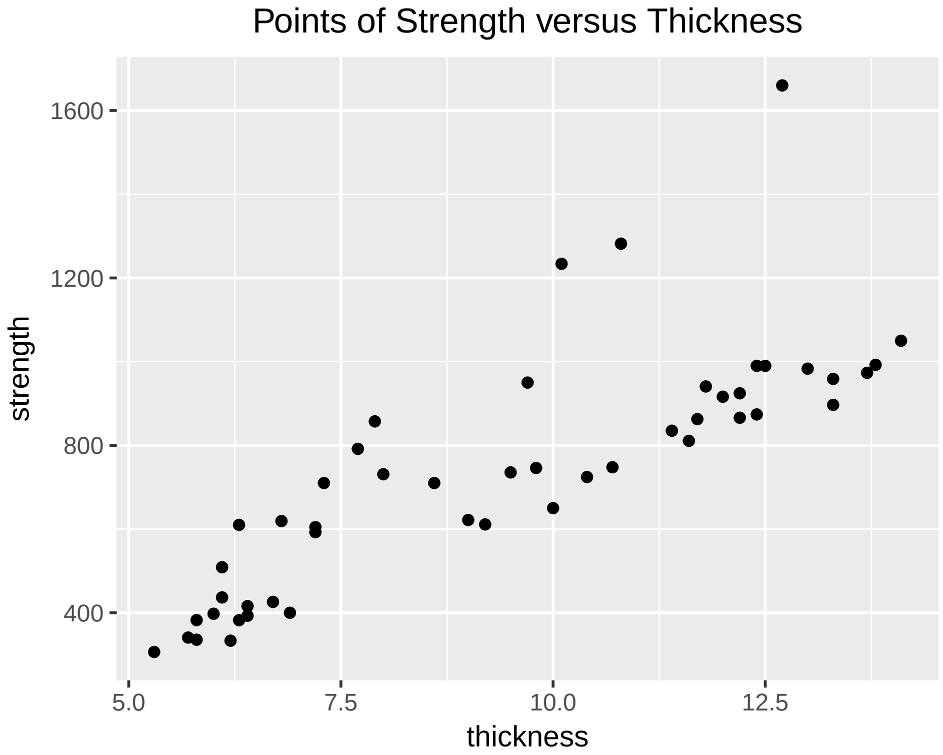
Problem 5

I first read the data into R and have a look at the head of the data. It's already tidy data and the starch is a factor.

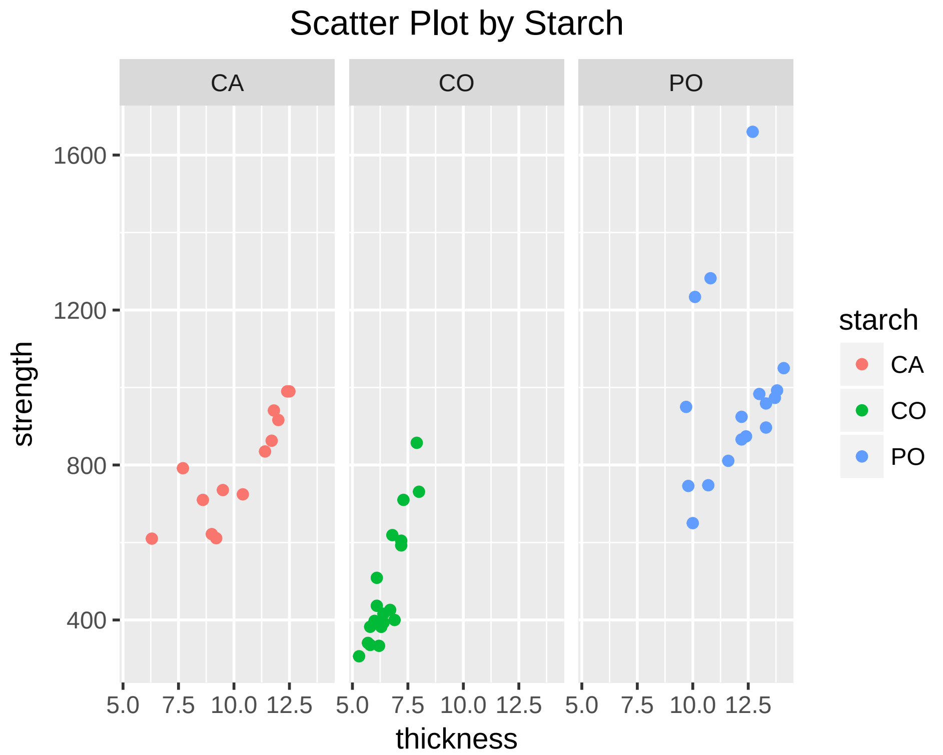
I first draw the scatter plot of thickness and strength across all starch to see the relationship roughly. And by the following figure we know that as the thickness increasing, the strength also increasing.

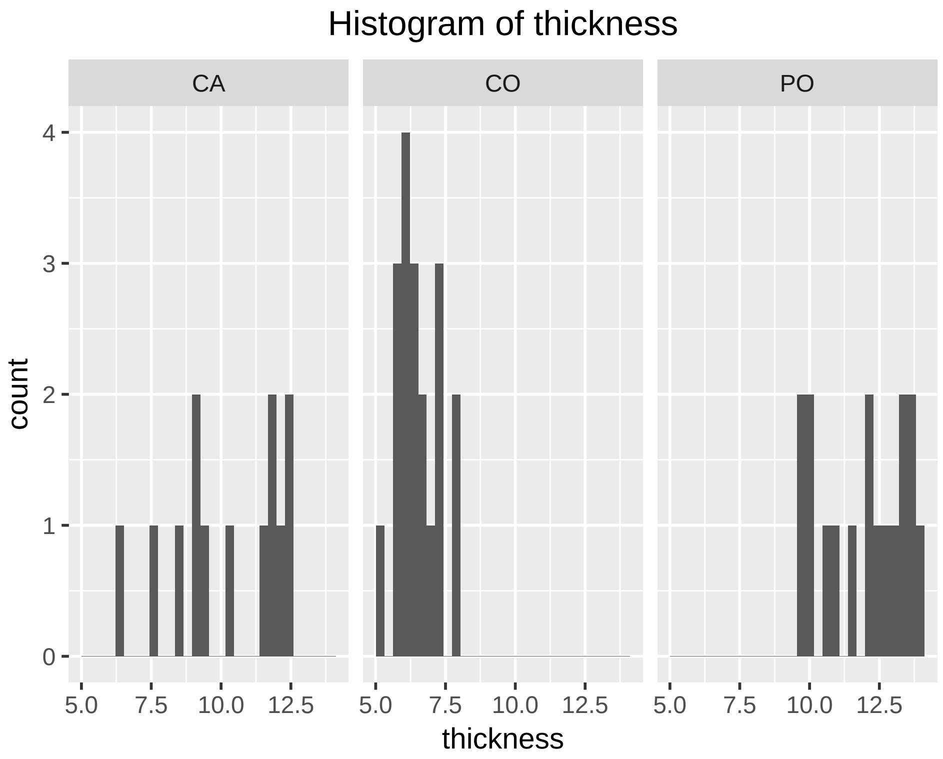
Table 5: Head of Starch Data

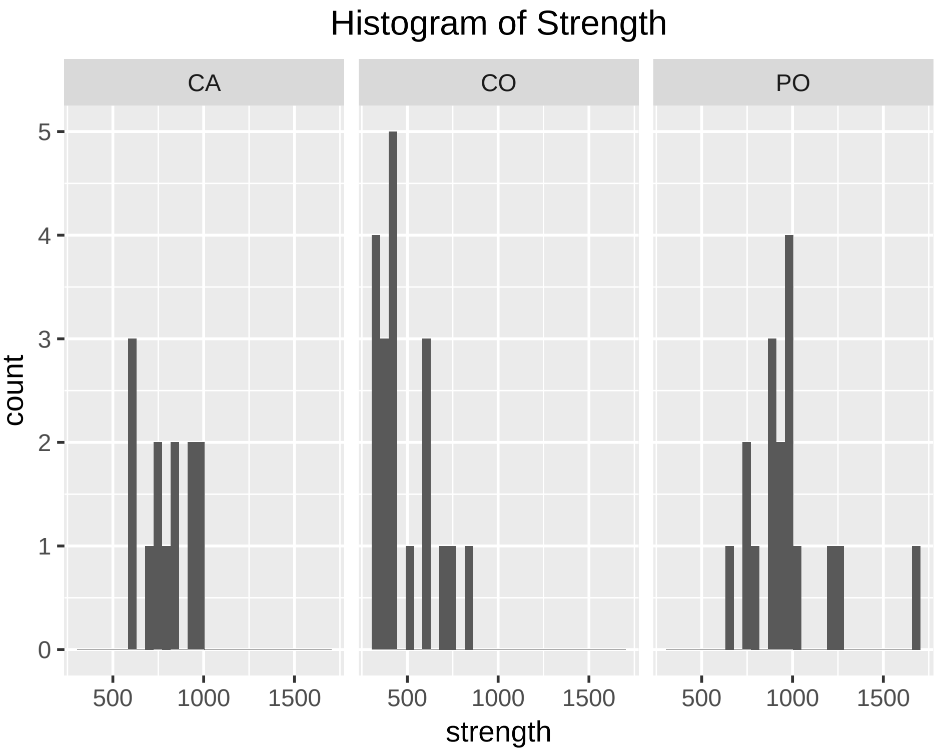
starch	strength	thickness
CA	791.7	7.7
CA	610.0	6.3
CA	710.0	8.6
CA	940.7	11.8
CA	990.0	12.4
CA	916.2	12.0



Then I redo the scatter plot by starch to see whether the kind of starch will impact the relationship between thickness and strength.


Scatter Plot by Starch
CA
CO
PO
strength
1600
1200
800
400
thickness
5.0
7.5
10.0
12.5
starch
● CA
● CO
● PO


Histogram of thickness
CA
CO
PO
count
4
3
2
1
0
thickness
5.0
7.5
10.0
12.5



We can see that the trend doesn't change for each kind of starch. From the histograms, we know that the range of thickness and strength for each starch is different. The strength and thickness of CO is relatively small while the thickness of PO are relatively large.

Problem 6

Part a: Get and import a database of US cities and states.

Part b: Create a summary table of the number of cities included by state

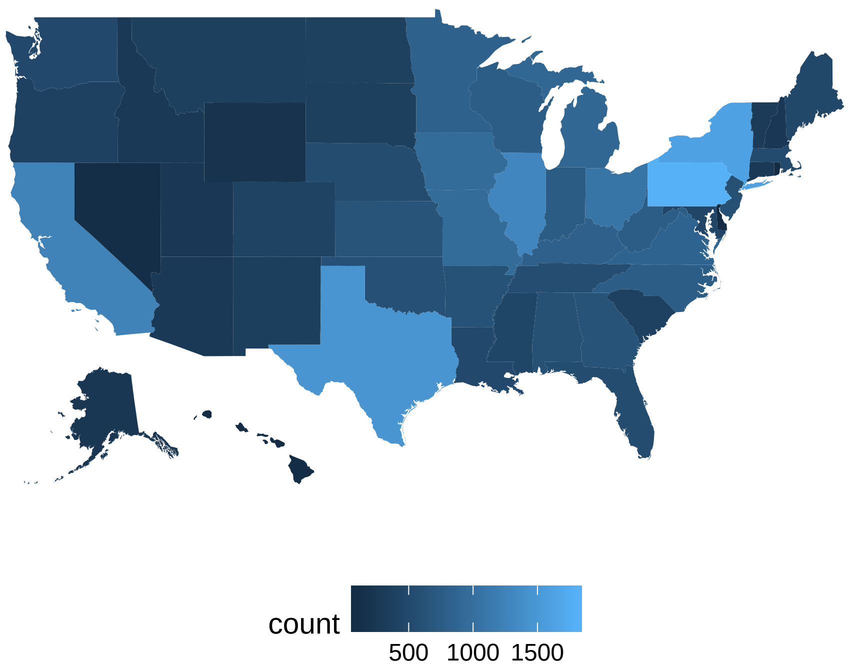
Table 6: Number of Cities in Each State

state 1	count 1	state 2	count 2	state 3	count 3	state 4	count 4	state 5	count 5
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

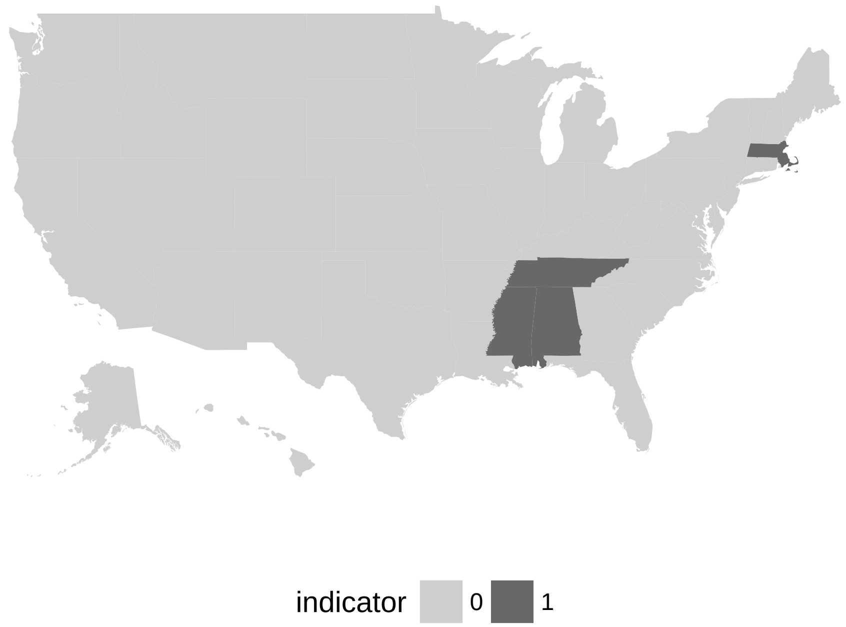
Part d: Create Maps

Count of Cities in Each State



U.S. states

Have More Than 3 Occurances of ANY Letter in Thier Name



Appendix: R code

```
####### Problem 4##########
set.seed(9302017)
#-----a compute success proportion-------
```

```r
succ_prop <- function(x) {
    # input: x is a vecotr of 0 and 1, suppose 1 stands for
    # success output: proportion of success in x vector
    prop <- sum(x == 1)/length(x)
    return(prop)
}
#------b flip coins matrix---------
P4b_data <- matrix(rbinom(10, 1, prob = (31:40)/100), nrow = 10,
    ncol = 10)

#------c compute success proportion by row and show in table------
P4b_succ_prop1 <- apply(P4b_data, MARGIN = 1, FUN = succ_prop)
P4b_succ_prop1 <- as.data.frame(t(P4b_succ_prop1))
colnames(P4b_succ_prop1) <- paste("p", (31:40)/100, sep = "_")
kable(P4b_succ_prop1, caption = "Apply by Row")

#------ compute success proportion by col and show in table ------
P4b_succ_prop2 <- apply(P4b_data, MARGIN = 2, FUN = succ_prop)
P4b_succ_prop2 <- as.data.frame(t(P4b_succ_prop2))
colnames(P4b_succ_prop2) <- paste("p", (31:40)/100, sep = "_")
kable(P4b_succ_prop2, caption = "Apply by Col")
#---show first 5 colnums in coin flips in Part b---
P4b <- P4b_data[, 1:5]
rownames(P4b) <- paste("p", (31:40)/100, sep = "_")
colnames(P4b) <- paste("flip", 1:5, sep = "")
kable(P4b, caption = "First 5 Columns of Matrix in Part b")
#----d coin flips--------
coinflip_vec <- function(p) {
    c <- rbinom(10, 1, p)
    return(c)
}
desired_prop <- (31:40)/100
P4d_data <- sapply(desired_prop, coinflip_vec)

#----compute the success proportion---
P4d_succ_prop <- apply(P4d_data, MARGIN = 2, FUN = succ_prop)
P4d_succ_prop <- as.data.frame(t(P4d_succ_prop))
colnames(P4d_succ_prop) <- paste("p", (31:40)/100, sep = "_")
kable(P4d_succ_prop, caption = "Success Proportion in Part d")

####### Problem 5##########
#---read in data----
url <- "http://www2.isye.gatech.edu/~jeffwu/book/data/starch.dat"
starch_data <- read.table(url, header = T)
kable(head(starch_data), caption = "Head of Starch Data")
ggplot(data = starch_data, aes(x = thickness, y = strength)) +
    geom_point() + ggtitle("Points of Strength versus Thickness") +
    theme(plot.title = element_text(hjust = 0.5))
### Problem 6 ### Part a we are grabbing a SQL set from
### here
### http://www.farinspace.com/wp-content/uploads/us_cities_and_states.zip
### download the files, looks like it is a .zip
### download('http://www.farinspace.com/wp-content/uploads/us_cities_and_states.zip',
### dest = 'us_cities_states.zip')
unzip("us_cities_states.zip", exdir = "./")
# read in the states data
states <- fread(input = "./us_cities_and_states/states.sql",
    sep = "'", sep2 = ",", header = F, select = c(2, 4))
colnames(states) <- c("state_name", "state")
# read in the cities data with zip
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")
# read in the cities without zip
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: this table includes Puerto Rico and
# Washington, D.C.
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", "count"),
    5), rep(1:5, each = 2))
kable(count_cities1, caption = "Number of Cities in Each State")
#---Part c-----
letter_freq <- function(letter, state) {
    letter_vec <- unlist(strsplit(state, ""))
    return(sum(letter_vec == letter))
}

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

for (i in 1:nrow(states)) {
    letter_count[i, ] <- sapply(letters, FUN = letter_freq,
        state = tolower(states[i, 1]))
}
# Part d: in this part I delete the 'state' Puerto Rico
# and Washington, D.C.
data("fifty_states")
# this line is optional due to lazy data loading
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)
# first map
```

```
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)) + ggtitle("Count of Cities in Each State")
p1
# create dataset for the second map
morethan3_indicator <- apply((letter_count > 3), MARGIN = 1,
    FUN = sum)
map_data2 <- states %>% mutate(indicator = as.factor(as.numeric(morethan3_indicator >=
    1))) %>% filter(state != "DC") %>% mutate(state_name = tolower(state_name))
# second map
p2 <- ggplot(map_data2, aes(map_id = state_name)) + geom_map(aes(fill = indicator),
    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 = "",
    title = "U.S. states", subtitle = "Have More Than 3 Occurances of ANY Letter in Thier Name") +
    theme(legend.position = "bottom", panel.background = element_blank(),
        plot.title = element_text(hjust = 0.5), plot.subtitle = element_text(hjust = 0.5)) +
    scale_fill_manual(values = c("#cecece", "#686868"))
p2
```