Smoke and Die

Leanne (Yuen Wan) Lee, Yeonghwan Son (Tony)

December 7, 2015

The causal relation between cigarette smoking and lung cancer have been a field of wide interest and extensive research for many existing organizations and individuals. In this particular project, we plan to establish the relationship between cigarette smoking and lung cancer. Furthermore, we look into other factors that could affect one's chance of being diagnozed with lung cancer such as one's age, race and gender.

To begin with, we compare the population of cigarette smokers and the number of lung cancer patients in each of the 51 states in the United States of America.

Downloading Raw Data

```
download.file(url = "https://raw.githubusercontent.com/leanne8/smoke_and_die/master/rawdata/smoke_df.csv
              destfile = "../rawdata/smoke_df.csv")

download.file(url = "https://raw.githubusercontent.com/leanne8/smoke_and_die/master/rawdata/lung_cancer_
              destfile = "../rawdata/lung_cancer_df.csv")
```

Data Cleaning and Preparation

```
smoke_df <- read.csv(file = "../rawdata/smoke_df.csv",
                     header = TRUE, stringsAsFactors = FALSE)
lung_cancer_df <- read.csv(file = "../rawdata/lung_cancer_df.csv",
                           header = TRUE, stringsAsFactors = FALSE)

smoke_df <- smoke_df[ , c(1, 2)]
colnames(smoke_df) <- c("state", "smokers(%)")

lung_cancer_df <- lung_cancer_df[ , c(1, 3)]
colnames(lung_cancer_df) <- c("state", "cancer(%)")
lung_cancer_df[ , 2] <- as.numeric(lung_cancer_df[ , 2])
```

```
## Warning: NAs introduced by coercion
```

```
lung_cancer_df[lung_cancer_df$state == "NV", 2] <-
  round((1683/2685000) * 100000, digits = 1)
lung_cancer_df[ , 2] <- lung_cancer_df[ , 2] / 1000

smoke_cancer_df <- cbind(smoke_df, lung_cancer_df)
smoke_cancer_df[ , 3] <- NULL

write.table(smoke_df, file = "../data/smoke_cdf.csv", sep = ",",
            row.names = FALSE, col.names = TRUE)
write.table(lung_cancer_df, file = "../data/lung_cancer_cdf.csv", sep = ",",
```

```
            row.names = FALSE, col.names = TRUE)
write.table(smoke_cancer_df, file = "../data/smoke_cancer_cdf.csv",
            row.names = FALSE, col.names = TRUE, sep = ",")
```

Basic Clean Data Inspection

Looking into the data frame 'smoke_cdf'

```
smoke_cdf <- read.csv(file = "../data/smoke_cdf.csv",
                      header = TRUE, stringsAsFactors = FALSE)
str(smoke_cdf)
```

```
## 'data.frame':    51 obs. of  2 variables:
##  $ state     : chr  "Alabama" "Alaska" "Arizona" "Arkansas" ...
##  $ smokers...: num  21.5 22.6 16.3 25.9 12.5 17.7 15.5 19.6 18.8 16.8 ...
```

```
summary(smoke_cdf)
```

```
##     state             smokers...
##  Length:51          Min.   :10.30
##  Class :character   1st Qu.:16.70
##  Mode  :character   Median :19.00
##                     Mean   :19.31
##                     3rd Qu.:21.45
##                     Max.   :27.30
```

Inspecting the data frame 'lung_cancer_cdf'

```
lung_cancer_cdf <- read.csv(file = "../data/lung_cancer_cdf.csv",
                            header = TRUE, stringsAsFactors = FALSE)
str(lung_cancer_cdf)
```

```
## 'data.frame':    51 obs. of  2 variables:
##  $ state    : chr  "AL" "AK" "AZ" "AR" ...
##  $ cancer...: num  0.0696 0.0572 0.0494 0.0751 0.0442 0.0449 0.061 0.0679 0.0612 0.0612 ...
```

```
summary(lung_cancer_cdf)
```

```
##     state              cancer...
##  Length:51          Min.   :0.02990
##  Class :character   1st Qu.:0.05565
##  Mode  :character   Median :0.06120
##                     Mean   :0.06138
##                     3rd Qu.:0.06780
##                     Max.   :0.09240
```

Merging the two data frames 'smoke_cdf' and 'lung_cancer_cdf' to obtain 'smoke_cancer_cdf' for future use. Then, using 'smoke_cancer_cdf', we look for the state with the highest percentage of cigarette smoking population.

```
smoke_cancer_cdf <- read.csv(file = "../data/smoke_cancer_cdf.csv",
                             header = TRUE, stringsAsFactors = FALSE)
smoke_cancer_cdf$state[which.max(smoke_cancer_cdf$smokers...)]
```

```
## [1] "West Virginia"
```

Looking for the state with the lowest percentage of cigarette smoking population.

```
smoke_cancer_cdf$state[which.min(smoke_cancer_cdf$smokers...)]
```

```
## [1] "Utah"
```

Looking for the state with the highest percentage of lung cancer patients.

```
smoke_cancer_cdf$state[which.max(smoke_cancer_cdf$cancer...)]
```

```
## [1] "Kentucky"
```

Looking for the state with the lowest percentage of lung cancer patients.

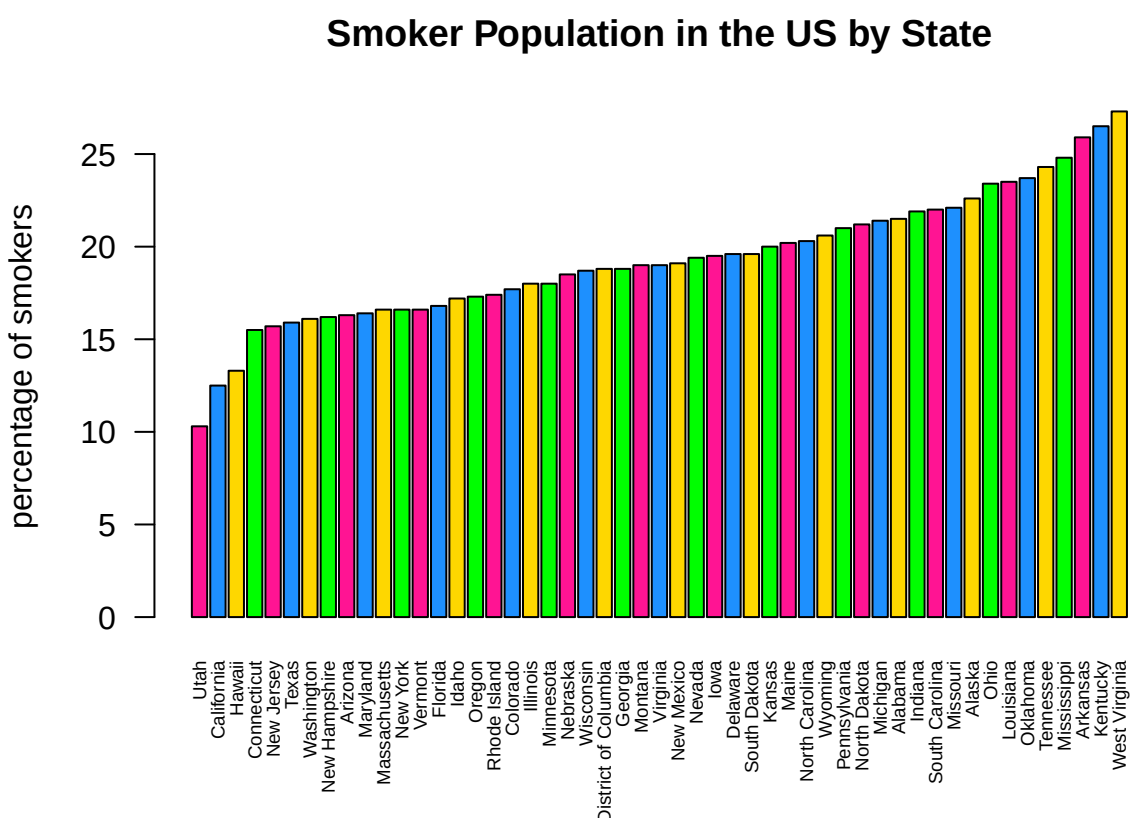
```
smoke_cancer_cdf$state[which.min(smoke_cancer_cdf$cancer...)]
```

```
## [1] "Utah"
```

Visualization Part 1

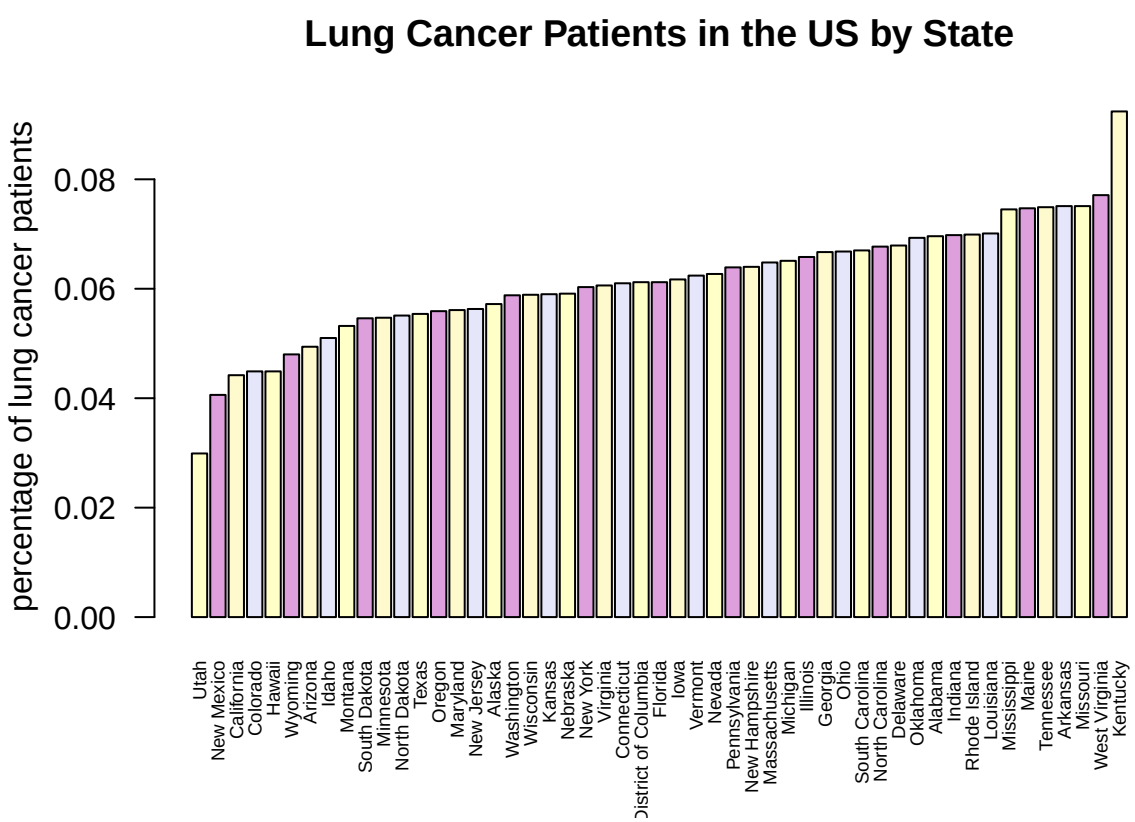
The first graph we introduce is a barplot of the percentage of cigarette smoking population in each state. The bars of the plot are arranged in an ascending order of the percentage. We observe from the graph, as we have inspected above, Utah has the lowest and West Virginia has the highest percentage of cigrette smoking population.

```
smoker_perc <- smoke_cancer_cdf$smokers...
names(smoker_perc) <- smoke_cdf[ , 1]
barplot(sort(smoker_perc), main = "Smoker Population in the US by State",
        cex.names = 0.6, las = 2, ylab = "percentage of smokers",
        col=c(rgb(255,20,147, maxColorValue = 255),
              rgb(30,144,254, maxColorValue = 255),
              rgb(254,215,0, maxColorValue = 255),
              rgb(0,254,0, maxColorValue = 255)))
```



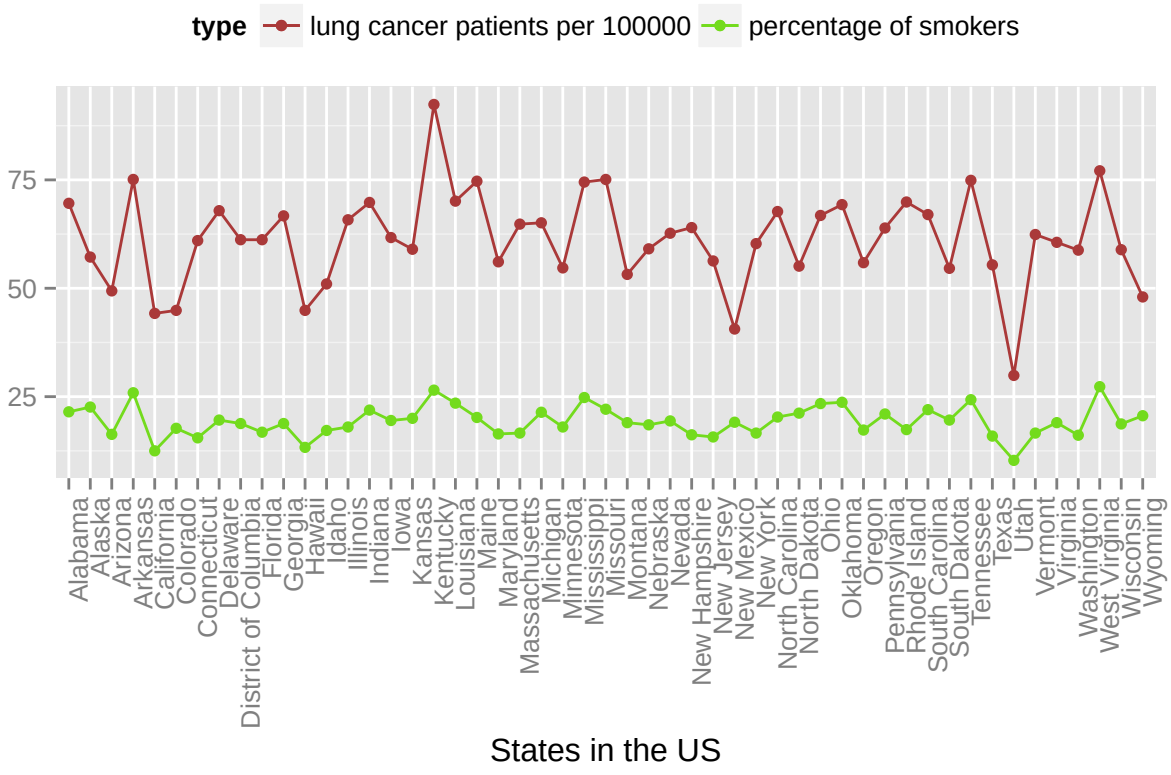
Then next graph is a similar barplot as the previous one, but it shows the percentage of lung cancer patients in each state. We see that Utah, which had the lowest smoking population percentage, also has the lowest percentage of lung cancer patients among all the states. Kentucky has has the highest rate of lung cancer patients followed by West Virginia which had the highest cigarette smoking population percentage.

```
lung_cancer_perc <- smoke_cancer_cdf$cancer...
names(lung_cancer_perc) <- smoke_cdf[ , 1]
barplot(sort(lung_cancer_perc),
        main = "Lung Cancer Patients in the US by State",
        cex.names = 0.6, las = 2, ylab = "percentage of lung cancer patients",
        col=c(rgb(255,255,200, maxColorValue = 255),
              rgb(221,160, 221, maxColorValue = 255),
              rgb(255,250,205, maxColorValue = 255),
              rgb(230,230,250, maxColorValue = 255)))
```



In this graph, we compared the rate of smoking population of each state with the rate of lung cancer patients. It shows that the higher smoking population in the state lead to a greater number of lung cancer patients.

```
ggplot(smoke_cancer_cdf) +
  geom_point(aes(x = names(smoker_perc),
                 y = smoke_cancer_cdf$smokers..., col = "red")) +
  geom_line(aes(x = names(smoker_perc),
                y = smoke_cancer_cdf$smokers..., col = "red", group = 1)) +
  geom_point(aes(x = names(smoker_perc),
                 y = smoke_cancer_cdf$cancer... * 1000, col = "green")) +
  geom_line(aes(x = names(smoker_perc),
                y = smoke_cancer_cdf$cancer... * 1000, col = "green", group = 2)) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  scale_color_manual(values = c("#AA3939", "#73DB1D"), name = "type",
                     labels = c("lung cancer patients per 100000", "percentage of smokers")) +
  theme(legend.position = "top") +
  xlab("States in the US") +
  ylab("")
```

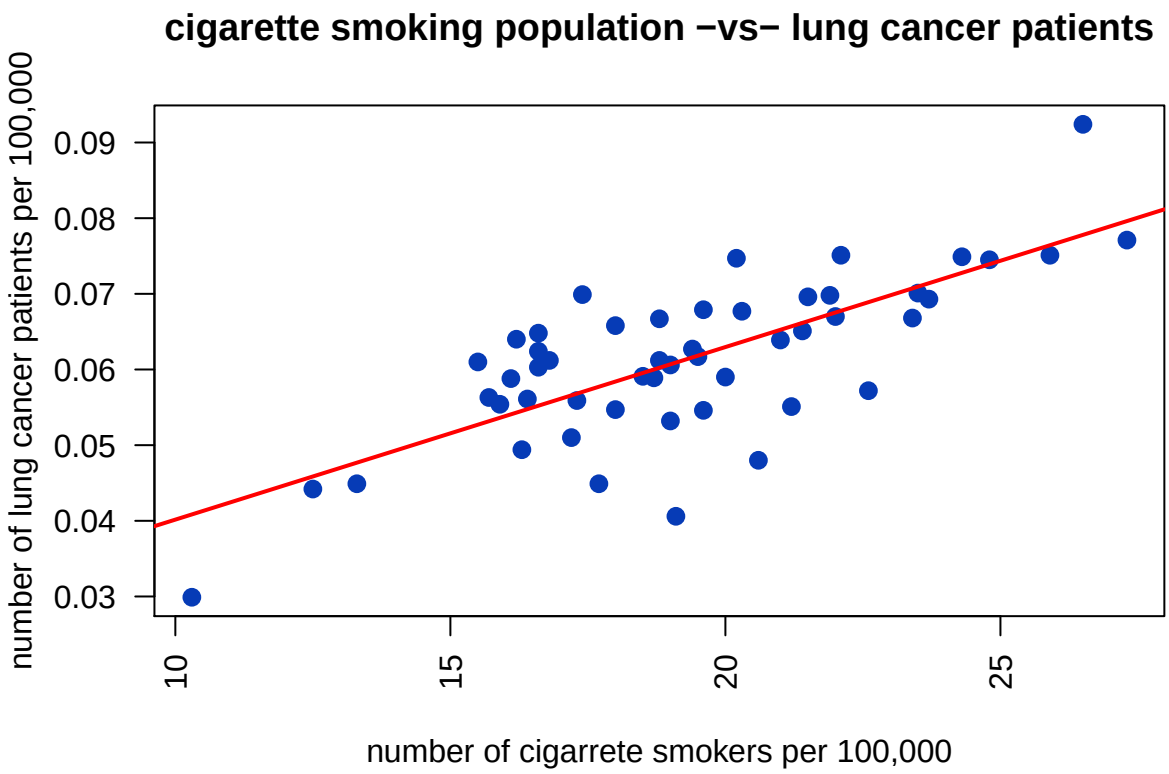


In this linear regression model, we analyzed the correlation between the rate of smoking population with the number of lung cancer patients. It clearly show that there is a positive correction between the two factors.

```
smoke_cancer_fit <- lm(smoke_cancer_cdf$cancer... ~ smoke_cancer_cdf$smokers...)
smoke_cancer_fit
```

```
##
## Call:
## lm(formula = smoke_cancer_cdf$cancer... ~ smoke_cancer_cdf$smokers...)
##
## Coefficients:
##                   (Intercept)  smoke_cancer_cdf$smokers...
##                      0.017352                     0.002281
```

```
plot(smoke_cancer_cdf$smokers..., smoke_cancer_cdf$cancer...,
     pch = 16, cex = 1.3, col = "#063BB6",
     main = "cigarette smoking population -vs- lung cancer patients", las = 2,
     xlab = "number of cigarrete smokers per 100,000",
     ylab = "number of lung cancer patients per 100,000")
abline(smoke_cancer_fit, col = "#FF0000", lwd = 2)
```



Downloading male and female lung cancer raw data

```
download.file(url = "https://raw.githubusercontent.com/leanne8/smoke_and_die/master/rawdata/lung_cancer
              destfile = "../rawdata/lung_cancer_male.csv")

download.file(url = "https://raw.githubusercontent.com/leanne8/smoke_and_die/master/rawdata/lung_cacner
              destfile = "../rawdata/lung_cancer_female.csv")
```

Data Cleaning and Preparation for lung cancer data

```
male_df <- read.csv("../rawdata/lung_cancer_male.csv", header = TRUE,
                    sep = "\t", col.names = c("X", "male_age", "male_all",
                                              "white", "black", "asian",
                                              "native_american", "hispanic"),
                    stringsAsFactors = FALSE)
male_df[ , 1] <- NULL

female_df <- read.csv("../rawdata/lung_cancer_female.csv", header = TRUE,
                      sep = "\t", col.names = c("X", "female_age",
                                                "female_all", "white", "black",
                                                "asian", "native_american",
                                                "hispanic"),
```

```
                      stringsAsFactors = FALSE)
female_df[ , 1] <- NULL

male_df[male_df == "~"] <- NA
female_df[female_df == "~"] <- NA

for (i in 2:length(colnames(male_df))) {
  male_df[ , i] <- as.numeric(male_df[ , i])
  female_df[ , i] <- as.numeric(female_df[ , i])
}

write.table(male_df, file = "../data/male_cdf.csv",
            row.names = FALSE, col.names = TRUE, sep = ",")
write.table(female_df, file = "../data/female_cdf.csv",
            row.names = FALSE, col.names = TRUE, sep = ",")
```

Data preparation for plotting lung cancer rate vs race

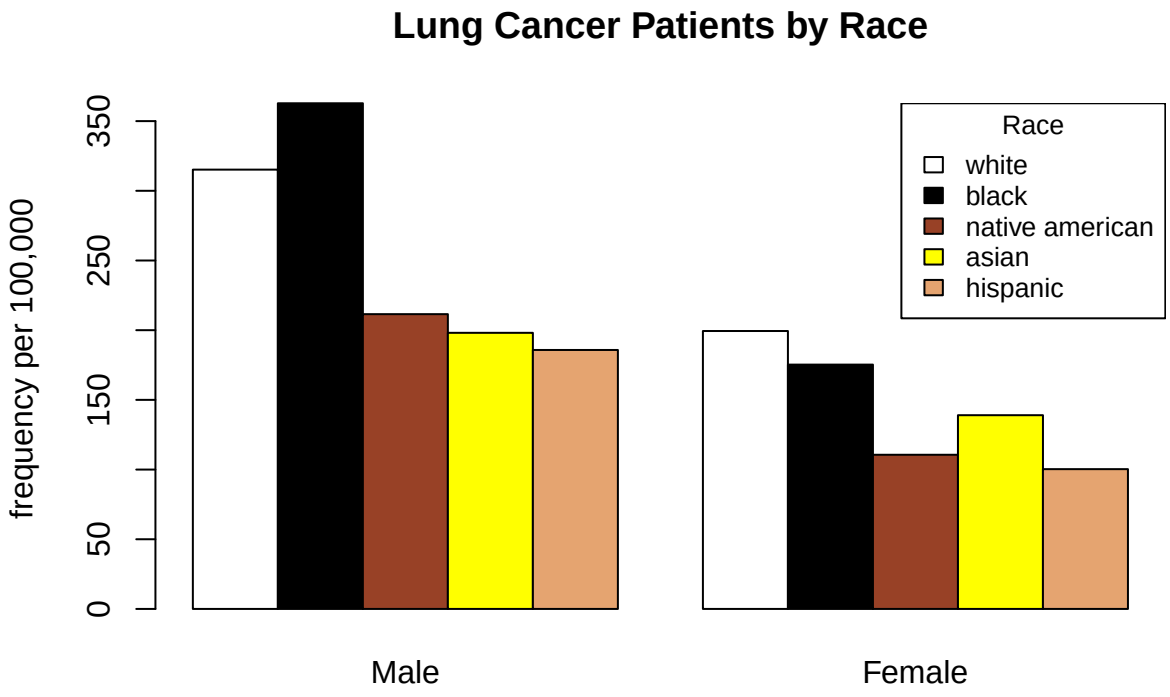
```
male_cdf <- read.csv(file = "../data/male_cdf.csv",
                     header = TRUE, stringsAsFactors = FALSE)
female_cdf <- read.csv(file = "../data/female_cdf.csv",
                       header = TRUE, stringsAsFactors = FALSE)

male_fifty_df <- male_cdf[12:19, ]

total_rate_male <- c()
for (i in 1:5){
  total_rate_male[i] <- round(sum(male_fifty_df[ , (i+2)]) / 8, digit = 1)
}
names(total_rate_male) <- colnames(male_fifty_df)[3:7]

female_fortyfive_df <- female_cdf[11:19, ]

total_rate_female <- c()
for (i in 1:5) {
  total_rate_female[i] <- round(sum(female_fortyfive_df[ , (i+2)]) / 9, digit = 1)
}
names(total_rate_female) <- colnames(female_fortyfive_df)[3:7]
```

Visualization Part 2

In this bar chart, it illustrates that the rate of lung cancer patients by race. It shows that black male and white female have higher chances of getting lung cancer.

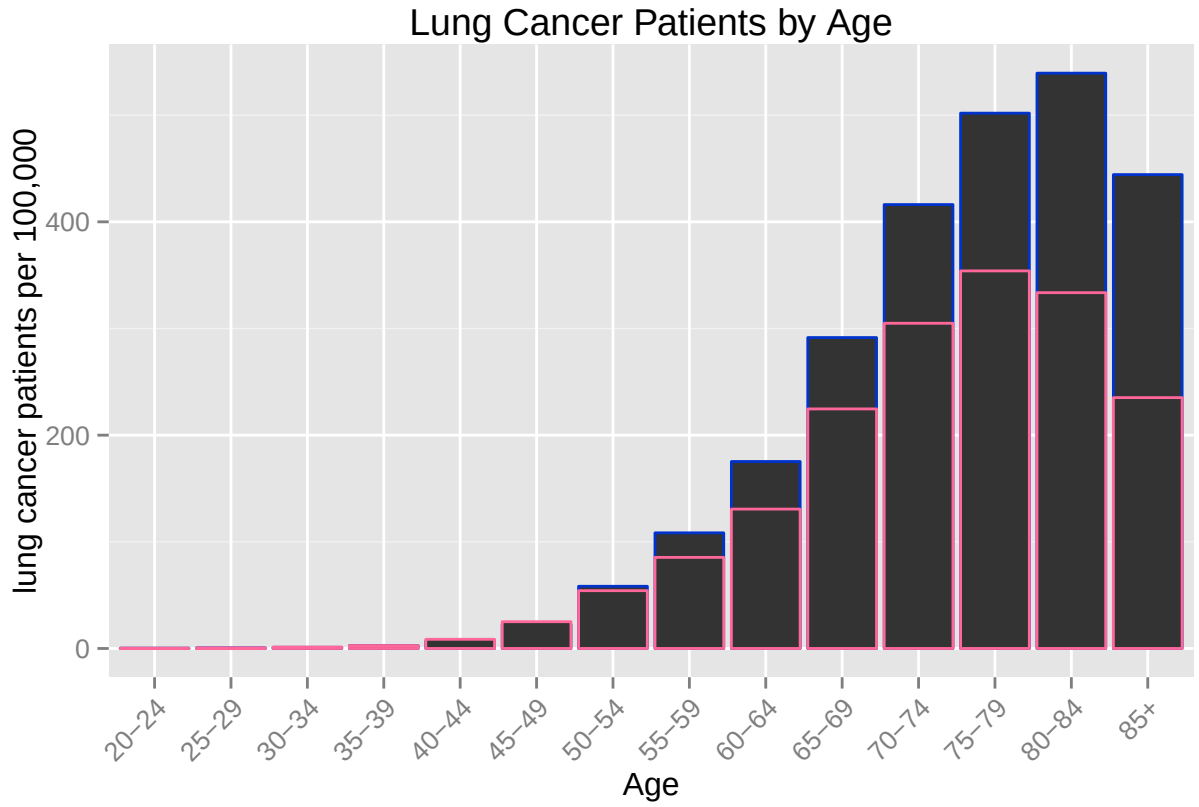
```
total_rate_combined <- cbind("Male" = total_rate_male,
                             "Female" = total_rate_female)
barplot(total_rate_combined,
        col = c("#FFFFFF", "#000000", "#984126", "#FFFF00", "#E5A470"),
        main = "Lung Cancer Patients by Race",
        ylab = "frequency per 100,000", beside = TRUE)
legend("topright",
       title = "Race",
```

```
       legend = c("white", "black", "native american", "asian", "hispanic"),
       fill = c("#FFFFFF", "#000000", "#984126", "#FFFF00", "#E5A470"),
       cex = 0.8)
```



In this bar chart, it shows relationship between the rate of lung cancer patients by different age groups. The graph not only shows that older people have higher chance of being diagnosed with lung cancer, but also males have higher chance of getting lung cancer than females do.

```
both_gender_df <- cbind(male_cdf, female_cdf)
both_gender_df <- both_gender_df[-c(1:5), ]
ggplot (both_gender_df) +
  geom_bar(aes(x = both_gender_df$male_age, y = both_gender_df$male_all),
           stat = "identity", col= "#0033CC") +
  geom_bar(aes(x = both_gender_df$female_age, y = both_gender_df$female_all),
           stat = "identity", col= "#FF6699") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  xlab("Age") + ylab("lung cancer patients per 100,000") +
  ggtitle("Lung Cancer Patients by Age")
```



Exporting plots as png files

```
png(filename = "../images/smokepop.png")
barplot(sort(smoker_perc), main = "Smoker Population in the US by State",
        cex.names = 0.6, las = 2, ylab = "percentage of smokers",
        col=c(rgb(255,20,147, maxColorValue = 255),
              rgb(30,144,254, maxColorValue = 255),
              rgb(254,215,0, maxColorValue = 255),
              rgb(0,254,0, maxColorValue = 255)))
dev.off()

png(filename = "../images/lungcancer.png")
barplot(sort(lung_cancer_perc),
        main = "Lung Cancer Patients in the US by State",
        cex.names = 0.6, las = 2, ylab = "percentage of lung cancer patients",
        col=c(rgb(255,255,200, maxColorValue = 255),
              rgb(221,160, 221, maxColorValue = 255),
              rgb(255,250,205, maxColorValue = 255),
              rgb(230,230,250, maxColorValue = 255)))
dev.off()

png(filename = "../images/comp.png")
figure1 <- ggplot(smoke_cancer_cdf) +
```

```
  geom_point(aes(x = names(smoker_perc),
                 y = smoke_cancer_cdf$smokers..., col = "red")) +
  geom_line(aes(x = names(smoker_perc),
                y = smoke_cancer_cdf$smokers..., col = "red", group = 1)) +
  geom_point(aes(x = names(smoker_perc),
                 y = smoke_cancer_cdf$cancer... * 1000, col = "green")) +
  geom_line(aes(x = names(smoker_perc),
                y = smoke_cancer_cdf$cancer... * 1000, col = "green", group = 2)) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  scale_color_manual(values = c("#AA3939", "#73DB1D"), name = "type",
                     labels = c("lung cancer patients per 100000",
                                "percentage of smokers")) +
  theme(legend.position = "top") +
  xlab("States in the US") +
  ylab("")
plot(figure1)
dev.off()

png(filename = "../images/reg.png")
plot(smoke_cancer_cdf$smokers..., smoke_cancer_cdf$cancer...,
     pch = 16, cex = 1.3, col = "#063BB6",
     main = "cigartte smoking population -vs- lung cancer patients", las = 2,
     xlab = "number of cigarette smokers per 100,000",
     ylab = "number of lung cancer patients per 100,000") +
  abline(smoke_cancer_fit, col = "#FF0000", lwd = 2)
dev.off()

png(filename = "../images/race.png")
barplot(total_rate_combined, col = c("#FFFFFF", "#000000", "#984126", "#FFFF00", "#E5A470"),
        main = "Lung Cancer Patients by Race",
        ylab = "frequency per 100,000", beside = TRUE)
legend("topright",
       title = "Race",
       legend = c("white", "black", "native american", "asian", "hispanic"),
       fill = c("#FFFFFF", "#000000", "#984126", "#FFFF00", "#E5A470"),
       cex = 0.8)
dev.off()

png(filename = "../images/age.png")
figure2 <- ggplot (both_gender_df) +
  geom_bar(aes(x = both_gender_df$male_age, y = both_gender_df$male_all),
           stat = "identity", col= "#0033CC") +
  geom_bar(aes(x = both_gender_df$female_age, y = both_gender_df$female_all),
           stat = "identity", col= "#FF6699") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  xlab("Age") + ylab("lung cancer patients per 100,000") +
  ggtitle("Lung Cancer Patients by Age")
plot(figure2)
dev.off()
```