project

Cool people

November 23, 2015

```r
library(readr)
library(ggplot2)
download.file(url = "https://raw.githubusercontent.com/leanne8/smoke_and_die/master/smoke_df.csv",
              destfile = "smoke_df.csv")
smoke_df <- read.csv(file = "smoke_df.csv", header = TRUE, stringsAsFactors = FALSE)

# Data on lung cancer patients
download.file(url = "https://raw.githubusercontent.com/leanne8/smoke_and_die/master/lung_cancer_df.csv"
              destfile = "lung_cancer_df.csv")
lung_cancer_df <- read.csv(file = "lung_cancer_df.csv", header = TRUE, stringsAsFactors = FALSE)


# Extracting only the necessary data from 'smoke_df' data frame
smoke_df <- smoke_df[ , c(1, 2)]
colnames(smoke_df) <- c("state", "cigarette smokers(%)")

# Checking smoke_df data frame
str(smoke_df)
```

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

```r
summary(smoke_df)
```

```
##     state            cigarette 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
```

```r
head(smoke_df)
```

```
##        state cigarette smokers(%)
## 1    Alabama                 21.5
## 2     Alaska                 22.6
## 3    Arizona                 16.3
## 4   Arkansas                 25.9
## 5 California                 12.5
## 6   Colorado                 17.7
```

```r
tail(smoke_df)
```

```
##             state cigarette smokers(%)
## 46        Vermont                 16.6
## 47       Virginia                 19.0
## 48     Washington                 16.1
## 49 West Virginia                  27.3
## 50      Wisconsin                 18.7
## 51        Wyoming                 20.6
```

```r
lung_cancer_df <- lung_cancer_df[ , c(1, 3)]
colnames(lung_cancer_df) <- c("state", "lung cancer patients(per 100,000)")
lung_cancer_df[ , 2] <- as.numeric(lung_cancer_df[ , 2])
```

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

```r
# Extracting only the necessary data from 'lung_cancer_df' data frame
# since the data doesn't have Neveda data, we have to manually look it up in the Neveda cancer webpage
# source from : Lung cancer in Neveda
# In 2009, there were 1,683 people diagnosed with lung cancer in Neveda
lung_cancer_df[lung_cancer_df$state == "NV", 2] <- round((1683/2685000) * 100000, digits = 1)

# Changing the number of patients with lung cancer(per 100,000) into the percentage of lung cancer pati
lung_cancer_df[ , 2] <- lung_cancer_df[ , 2] / 1000
colnames(lung_cancer_df) <- c("state", "lung cancer patients(%)")

# Checking lung_cancer_df
str(lung_cancer_df)
```

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

```r
summary(lung_cancer_df)
```

```
##     state            lung cancer patients(%)
##  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
```

```r
head(lung_cancer_df)
```

```
##   state lung cancer patients(%)
## 1    AL                  0.0696
## 2    AK                  0.0572
## 3    AZ                  0.0494
## 4    AR                  0.0751
## 5    CA                  0.0442
## 6    CO                  0.0449
```

```r
tail(lung_cancer_df)
```

```
##    state lung cancer patients(%)
## 46    VT                  0.0624
## 47    VA                  0.0606
## 48    WA                  0.0588
## 49    WV                  0.0771
## 50    WI                  0.0589
## 51    WY                  0.0480
```

```r
# Merging the two data frames
smoke_cancer_df <- cbind(smoke_df, lung_cancer_df)
smoke_cancer_df[ , 3] <- NULL

# state with highest percentage of smokers
smoke_cancer_df$state[which.max(smoke_cancer_df$`cigarette smokers(%)`)]
```

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

```r
# state with lowest percentage of smokers
smoke_cancer_df$state[which.min(smoke_cancer_df$`cigarette smokers(%)`)]
```

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

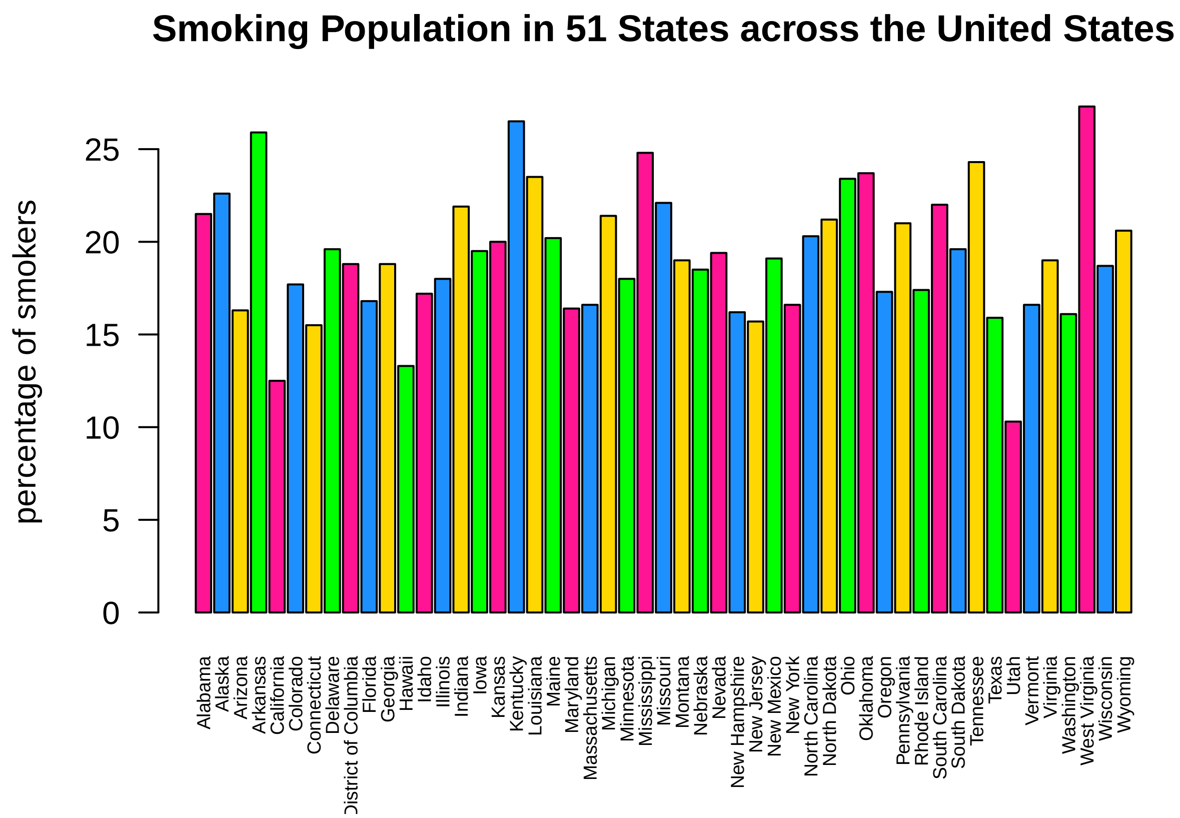
```r
# state with highest percentage of lung cancer patients
smoke_cancer_df$state[which.max(smoke_cancer_df$`lung cancer patients(%)`)]
```

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

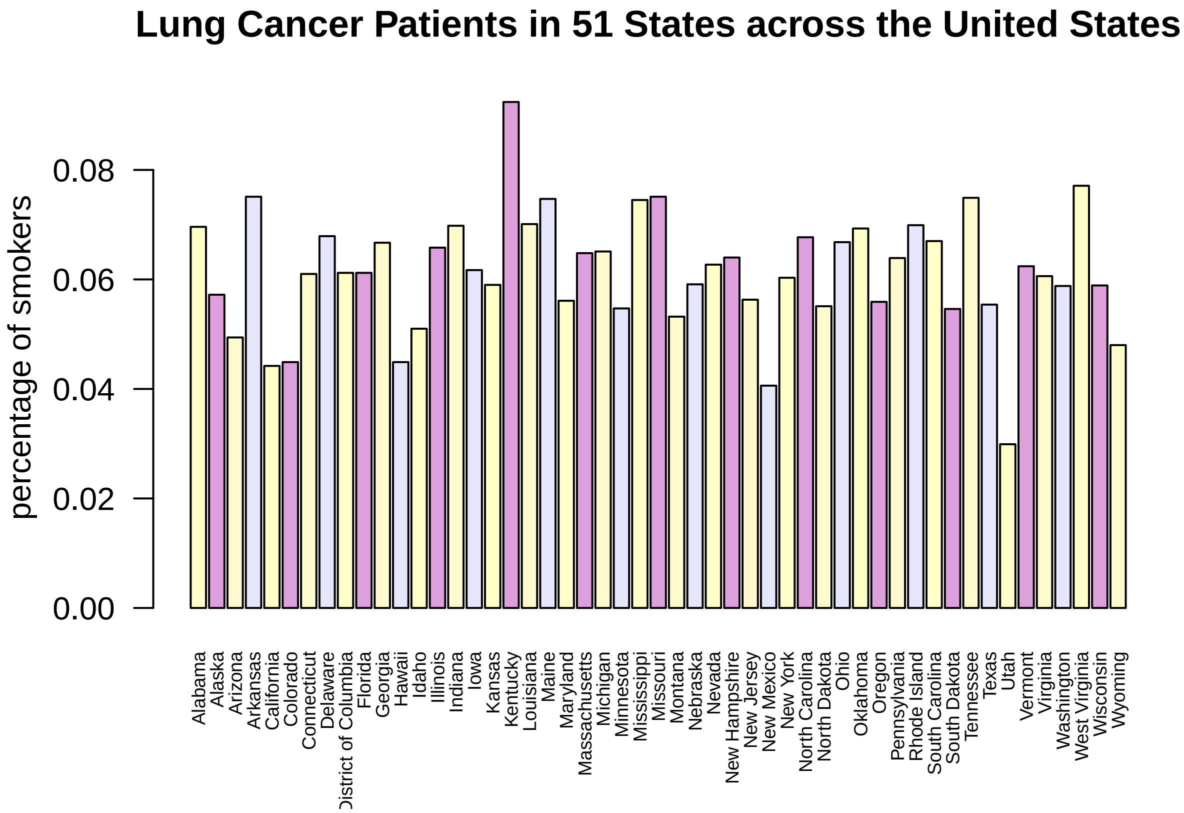
```r
# state with lowest percentage of lung cancer patients
smoke_cancer_df$state[which.min(smoke_cancer_df$`lung cancer patients(%)`)]
```

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

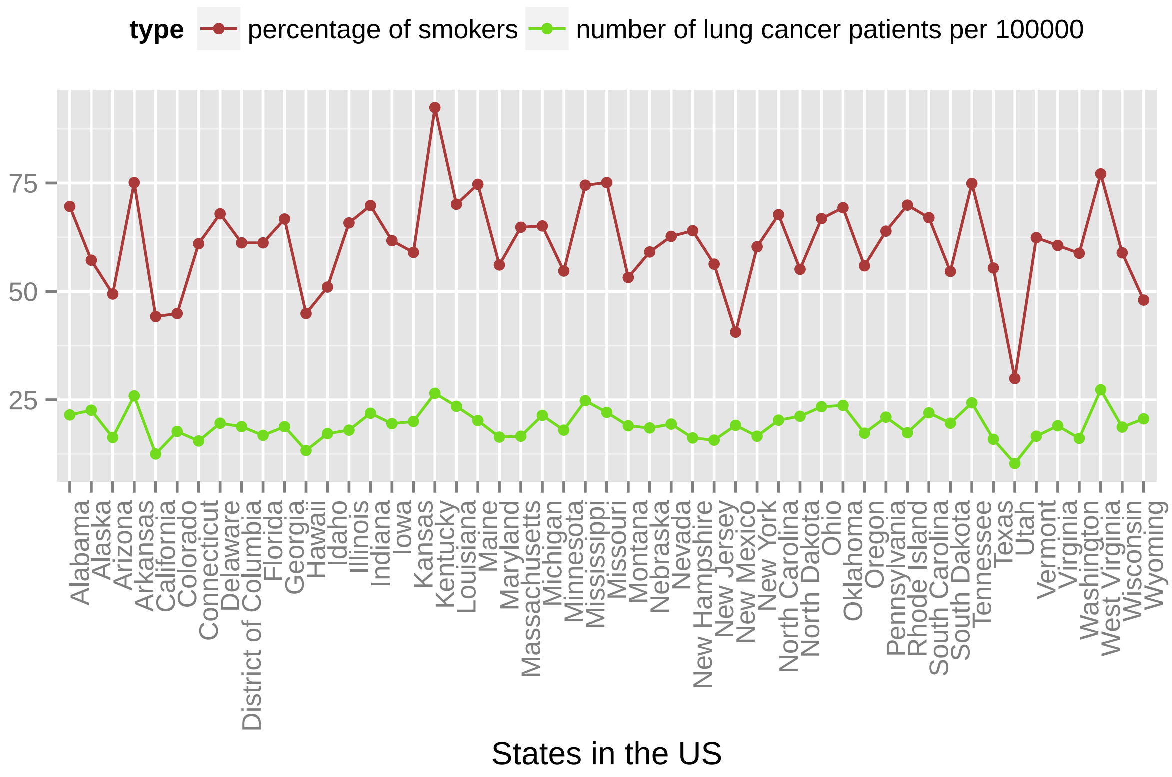
```r
# visual representation of the percentage of smokers in each state
smoker_perc <- smoke_cancer_df$`cigarette smokers(%)`
names(smoker_perc) <- smoke_df[ , 1]
barplot(smoker_perc, main = "Smoking Population in 51 States across the United States",
        cex.names = 0.6, las = 2, ylab = "percentage 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)))
```



```
# visual representation of the percentage of people with lung cancer in each state
lung_cancer_perc <- smoke_cancer_df$`lung cancer patients(%)`
names(lung_cancer_perc) <- smoke_df[ , 1]
barplot(lung_cancer_perc,
        main = "Lung Cancer Patients in 51 States across the United States",
        cex.names = 0.6, las = 2, ylab = "percentage of smokers",
        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)))
```

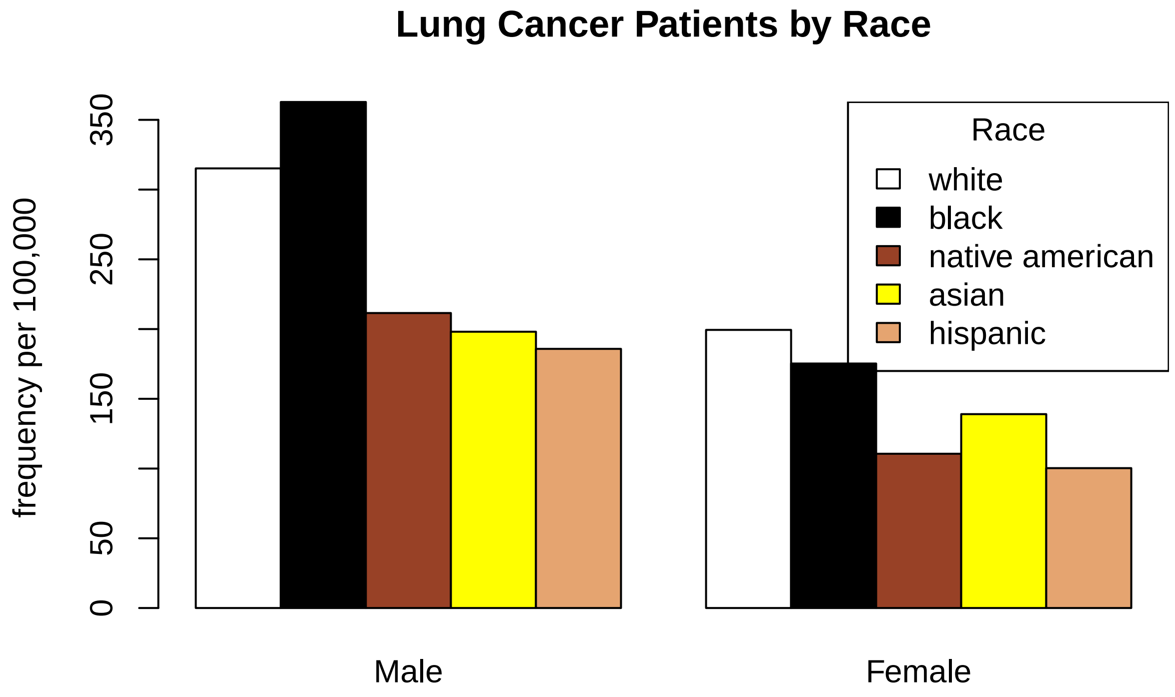


```r
# comparing the percentage of smokers and number of patients with lung cancer(per 100,000) in each stat
ggplot(smoke_cancer_df) +
  geom_point(aes(x = names(smoker_perc), y = smoke_cancer_df$`cigarette smokers(%)`, col = "red")) +
  geom_line(aes(x = names(smoker_perc), y = smoke_cancer_df$`cigarette smokers(%)`, col = "red", group =
  geom_point(aes(x = names(smoker_perc), y = smoke_cancer_df$`lung cancer patients(%)` * 1000, col = "g
  geom_line(aes(x = names(smoker_perc), y = smoke_cancer_df$`lung cancer patients(%)` * 1000, col = "gr
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  scale_color_manual(values = c("#AA3939", "#73DB1D"), name = "type", labels = c("percentage of smokers
  theme(legend.position = "top") +
  xlab("States in the US") +
  ylab("")
```

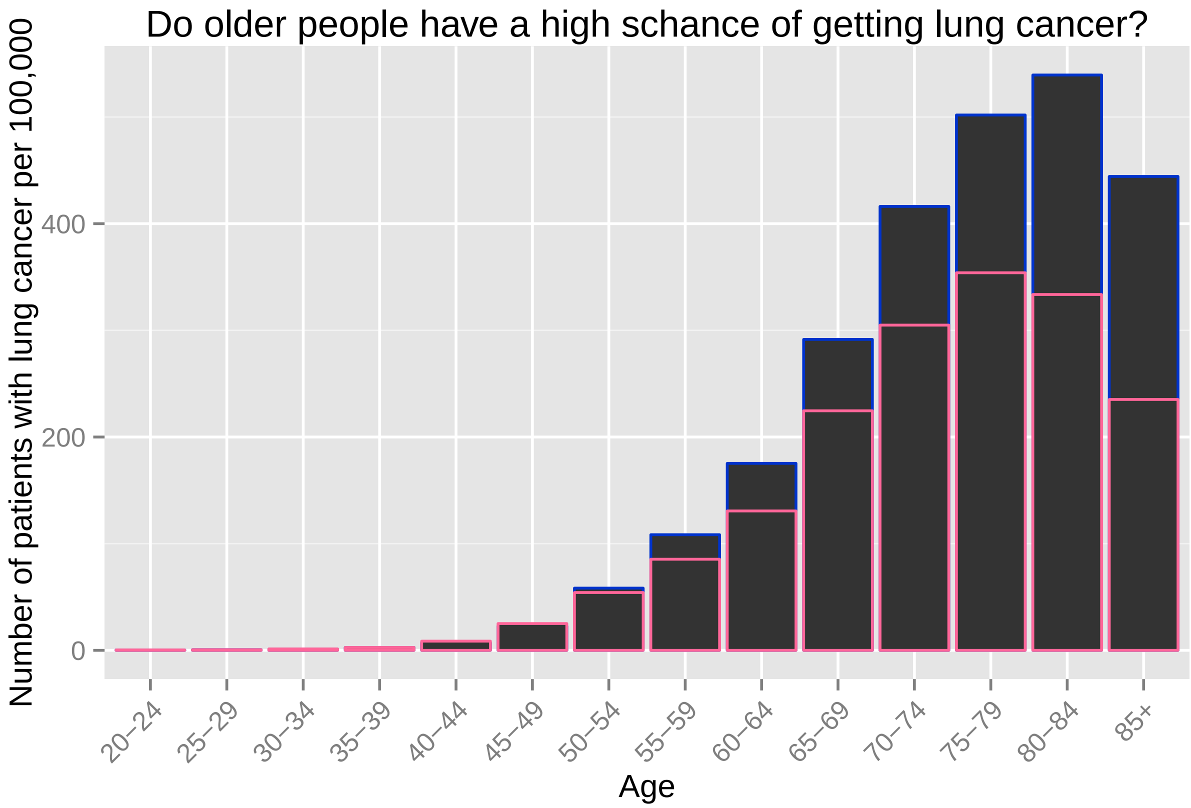


```
# We see that in general, states with higher percentage of smokers have greater number of lung cancer p

## Secondly, we look at the relationship between lung cancer and one's race as well as gender

# Downloading necessary files
download.file(url = "https://raw.githubusercontent.com/leanne8/smoke_and_die/master/lung_cancer_male.tx
              destfile = "lung_cancer_male.csv")
male_df <- read.csv("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

download.file(url = "https://raw.githubusercontent.com/leanne8/smoke_and_die/master/lung_cacner_%20femal
              destfile = "lung_cancer_female.csv")
female_df <- read.csv("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

# assigning NA to values that are not available in the data frames
male_df[male_df == "~"] <- NA
female_df[female_df == "~"] <- NA

# changing the values in the data frames into numbers
for (i in 2:length(colnames(male_df))) {
  male_df[ , i] <- as.numeric(male_df[ , i])
  female_df[ , i] <- as.numeric(female_df[ , i])
}
```

```
# comparing rate of lung cancer in patients over 50 years old by gender
male_fifty_df <- male_df[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_df[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]

# visual representation of the rate of lung cancer patients by race and gender
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"))
```



```
#visual representation of the rate of lung cancer patients by ages
both_gender_df <- cbind(male_df, female_df)
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("Number of patients with lung cancer per 100,000") +
  ggtitle("Do older people have a high schance of getting lung cancer?")
```


Do older people have a high schance of getting lung cancer?
Number of patients with lung cancer per 100,000
Age