Math-331 Project 1

2022-10-18

Predicting and Analysing Medical Insurance Costs

First importing the ggplot library and the tidyverse library to be able to plot data to analyze data and draw conclusions from the data set.

```
library(ggplot2)
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✔ tibble  3.1.8      ✔ dplyr   1.0.10
## ✔ tidyr   1.2.1      ✔ stringr 1.4.1
## ✔ readr   2.1.2      ✔ forcats 0.5.2
## ✔ purrr   0.3.4
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

Analyzing the different variables in the dataset and they affect each other

Reading the csv file insurance and storing it into a data set called medicalInsurance:

```
medicalInsurance <- read.csv("/Users/malaika/Desktop/MATH-331/insurance.csv")
#medicalInsurance
```

Summarizing the data set

```
summary(medicalInsurance)
```

```
##       age            sex                 bmi           children
##  Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000
##  1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000
##  Median :39.00   Mode  :character   Median :30.40   Median :1.000
##  Mean   :39.21                      Mean   :30.66   Mean   :1.095
##  3rd Qu.:51.00                      3rd Qu.:34.69   3rd Qu.:2.000
##  Max.   :64.00                      Max.   :53.13   Max.   :5.000
##     smoker             region             charges
##  Length:1338        Length:1338        Min.   : 1122
##  Class :character   Class :character   1st Qu.: 4740
##  Mode  :character   Mode  :character   Median : 9382
##                                        Mean   :13270
##                                        3rd Qu.:16640
##                                        Max.   :63770
```

There are 7 variables in this data set: * age = age of subject as an integer type variable.

- sex = sex of subject as a character type variable.
- bmi = Body Mass Index of subject as a float type variable.
- children = number of children of subject as an integer type variable.
- smoker = smoker status of subject as a character type variable.
- region = region of subject as a character type variable.
- charges = cost medical insurance for subject as a float type variable.

Changing the sex, children & smoker and region variables to factor as they are categorical variables having a limited number of different values

```
medicalInsurance$sex <- as.factor(medicalInsurance$sex)
medicalInsurance$children <- as.factor(medicalInsurance$children)
medicalInsurance$smoker <- as.factor(medicalInsurance$smoker)
medicalInsurance$region <- as.factor(medicalInsurance$region)
```

Summary to view the new variable types:

```
summary(medicalInsurance)
```

```
##       age            sex           bmi        children smoker
##  Min.   :18.00   female:662   Min.   :15.96   0:574    no :1064
##  1st Qu.:27.00   male  :676   1st Qu.:26.30   1:324    yes: 274
##  Median :39.00                Median :30.40   2:240
##  Mean   :39.21                Mean   :30.66   3:157
##  3rd Qu.:51.00                3rd Qu.:34.69   4: 25
##  Max.   :64.00                Max.   :53.13   5: 18
##        region       charges
##  northeast:324   Min.   : 1122
##  northwest:325   1st Qu.: 4740
##  southeast:364   Median : 9382
##  southwest:325   Mean   :13270
##                  3rd Qu.:16640
##                  Max.   :63770
```

Analyzing how age and sex affects the insurance charges

Calculating the average (mean) age of subjects in the data set, storing it in the variable age_mean and rounding the final answer:

```
age_mean=round(mean(medicalInsurance$age))
age_mean
```

```
## [1] 39
```

The average age of the subjects in the data set is 39 years.

Plotting the age through a stacked histogram to see which is the most common age group and the sex associated with it.

```
ggplot(data=medicalInsurance, aes(x=age,fill = sex)) +
  ylab("frequency")+
  geom_histogram() +
  ggtitle("Frequency of age compared to sex")+
  labs(caption = "Figure 1")
```

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

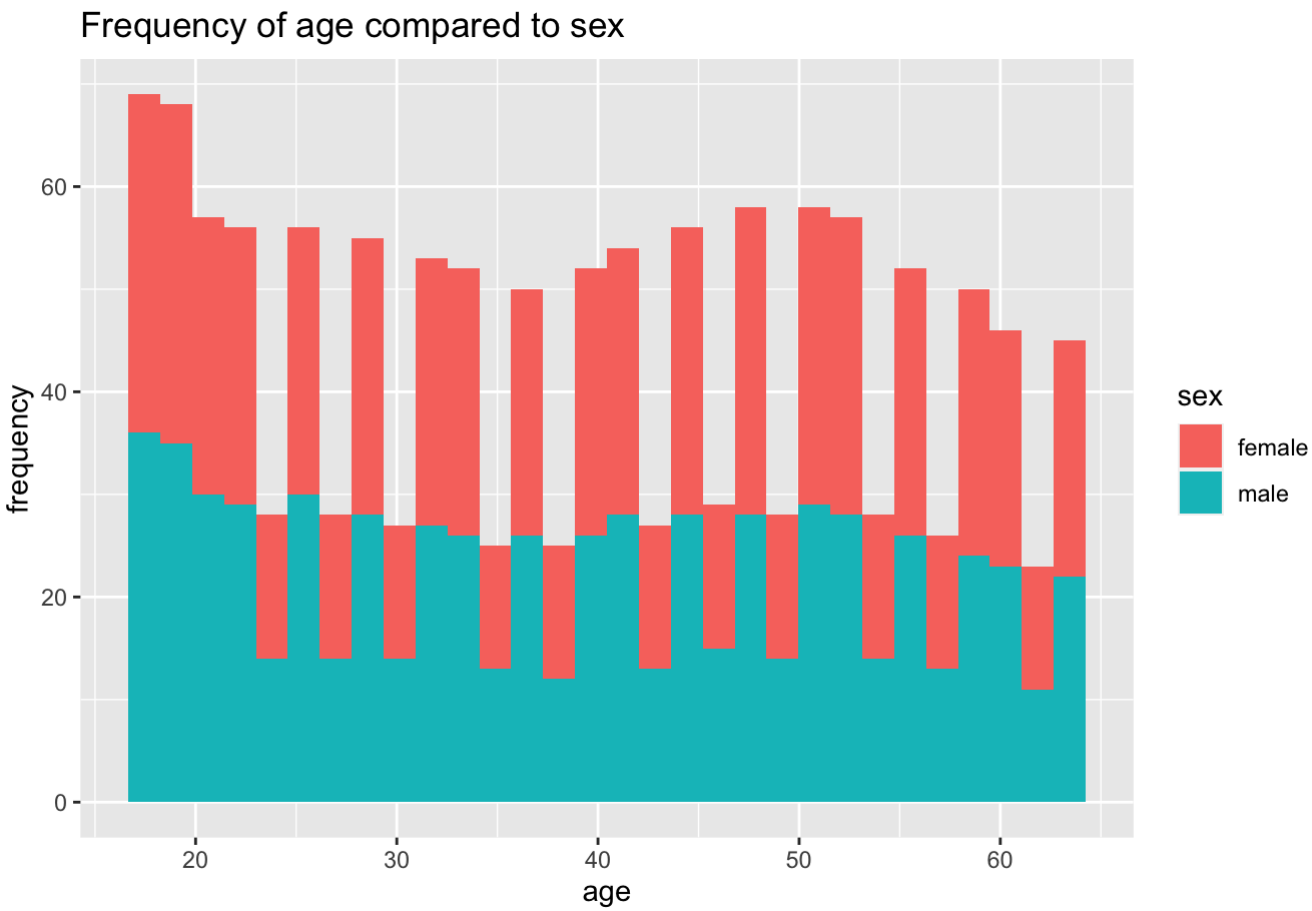

Figure 1

From the graph the most frequent age group in the data set is below 20. It's difficult to analyze how each sex compares to each other using this graph hence I used a bar graph to find the distinction between the number of male and female subjects. I also used facet_wrap to plot two individual graphs for each sex:

```
ggplot(data=medicalInsurance, aes(x=age,fill = sex)) +
  ylab("frequency")+
  geom_bar() +
  facet_wrap(~sex)+
  ggtitle("Frequency of age compared to sex")+
  labs(caption = "Figure 2")
```

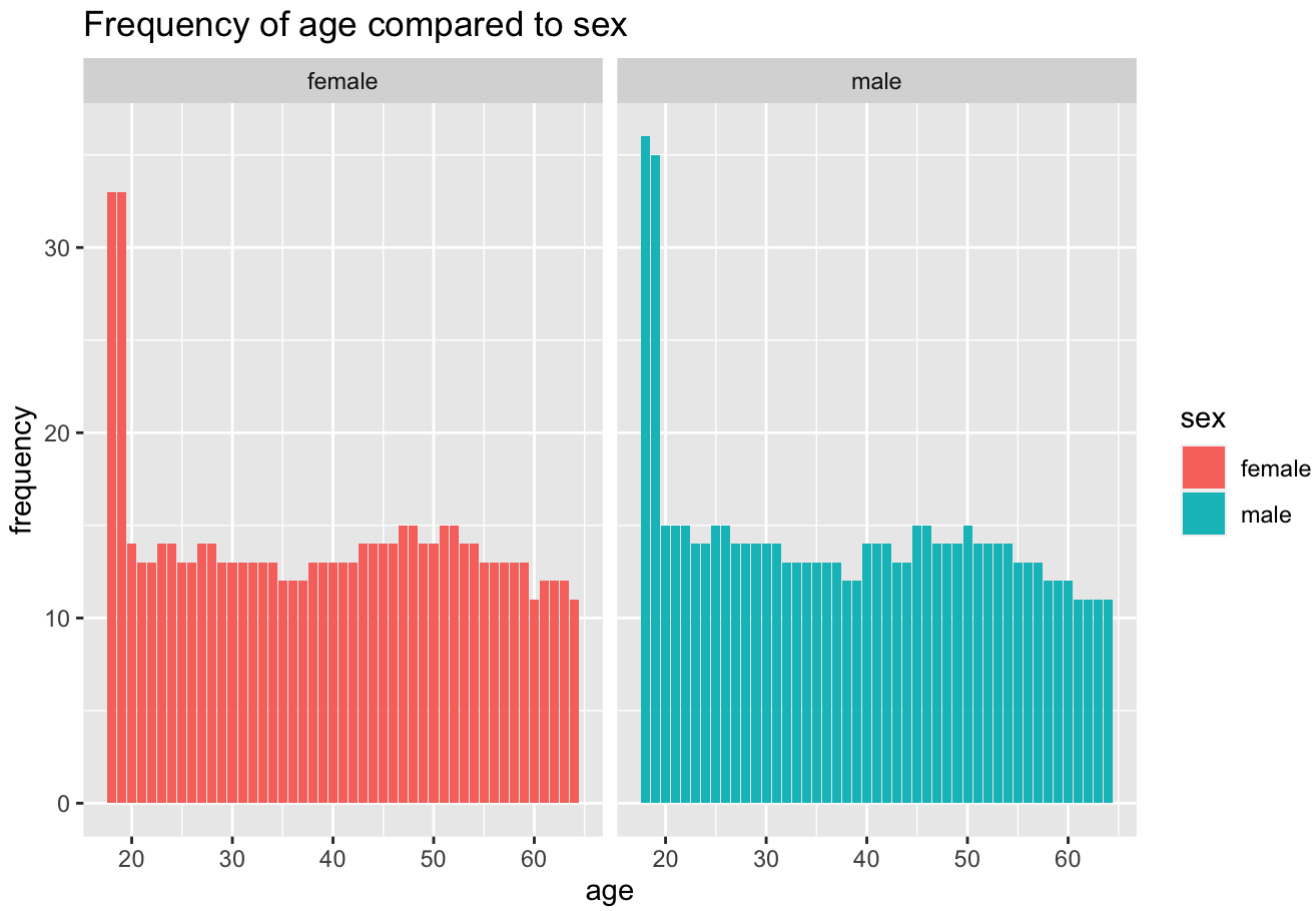

Figure 2

From the graph we can see that the number of male subjects is slight more than the females.

To find the exact number of male and female subjects:

```
length(which(medicalInsurance$sex=="male"))
```

```
## [1] 676
```

```
length(which(medicalInsurance$sex=="female"))
```

```
## [1] 662
```

Now that we found the distribution of age and sex in the dataset we can see how it affects the medical insurance charges.

```
ggplot(data=medicalInsurance, aes(x=age,y=charges,fill=sex))+
  facet_wrap(~sex) +
  xlab("age")+
  ylab("charges")+
  geom_violin() +
  ggtitle("Charges vary depending on age and sex") +
  labs(caption = "Figure 3")
```

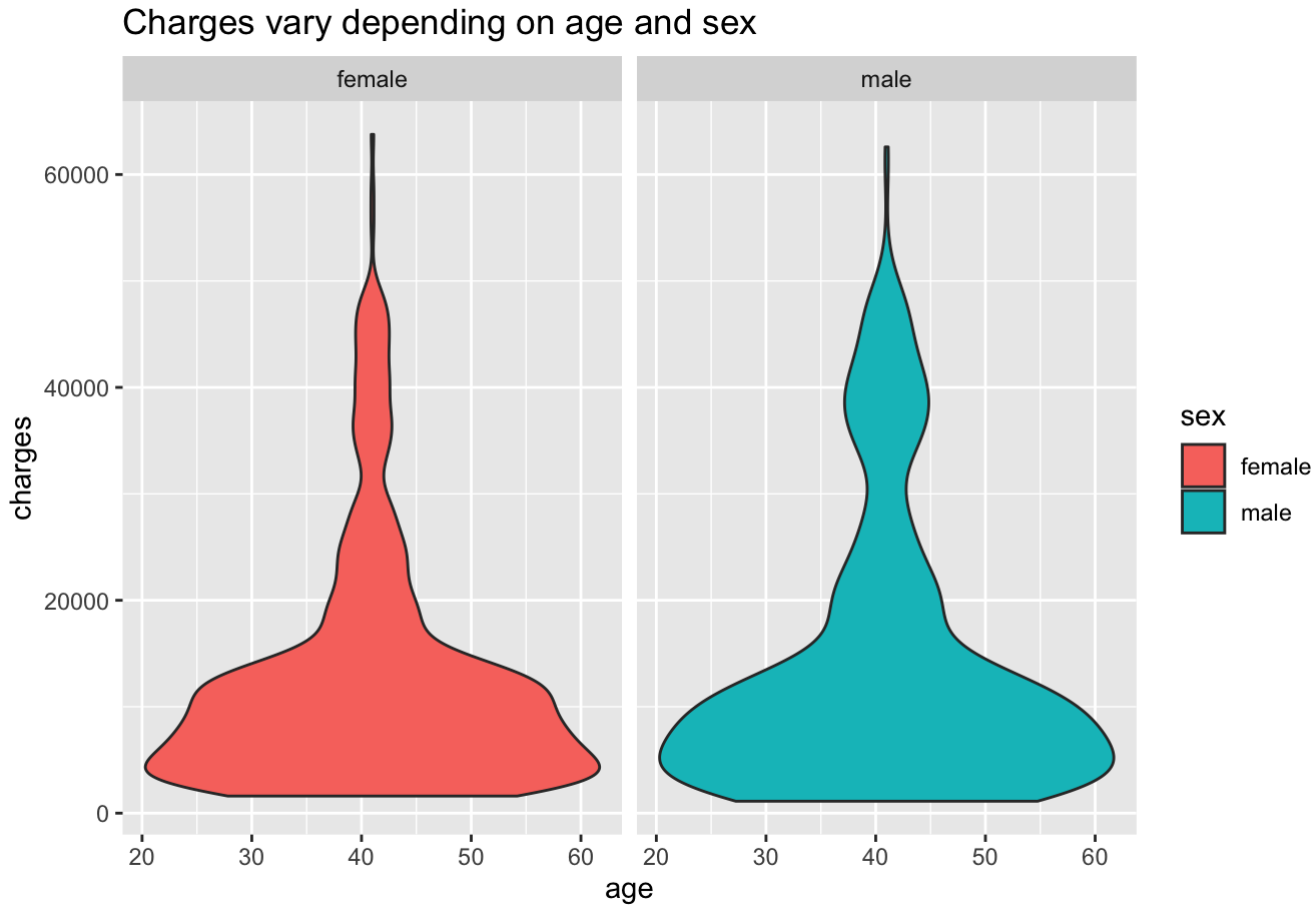

Figure 3

- Although it is difficult to analyze the plots by just looking at them I was able to make these conclusions:

1. Men and women between the ages 35 and 55 ages are charged the most
2. Women between 40 to 45 years are charged up to or more than $60000 whereas men between 35 to 45 are charged up to $60000.

To accurately find how the charges differ I'm using a t-test that will determine if there is a statistical difference between the means of two groups:

```
t.test(medicalInsurance$age,medicalInsurance$charges)
```

```
##
##  Welch Two Sample t-test
##
## data:  medicalInsurance$age and medicalInsurance$charges
## t = -39.965, df = 1337, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -13880.68 -12581.75
## sample estimates:
##   mean of x   mean of y
##    39.20703 13270.42227
```

The 95% confidence interval tells us the estimated difference between the means of the variables age and charges is between -13880.68 and -12581.75 and this compares to the plot in figure 3.

Grouping the data by sex to see how the charges vary depending on the sex of the subject

```
group_by(medicalInsurance, sex) %>%
  summarize(average_Charges=(round(mean(charges),digits=2)))
```

```
## # A tibble: 2 × 2
##   sex    average_Charges
##   <fct>            <dbl>
## 1 female          12570.
## 2 male            13957.
```

On average men are not charged significantly more than women. Men are charged $13956.75 whereas women are charged $12569.58.

Conclusions drawn by analyzing the Body Mass Index

Calculating the average (mean) body mass index of subjects in the data set and storing it in the variable bmi_mean:

```
bmi_mean=round(mean(medicalInsurance$bmi),digits = 2)
bmi_mean
```

```
## [1] 30.66
```

The average BMI (rounded up to 2 decimal points) is 30.66.

I researched the Adult Body Mass Index and the CDC categorizes ranges of BMI by the following:

- If your BMI is less than 18.5, it falls within the underweight range.
- If your BMI is 18.5 to <25, it falls within the healthy weight range.
- If your BMI is 25.0 to <30, it falls within the overweight range.
- If your BMI is 30.0 or higher, it falls within the obesity range.

Using the above information the average body mass index of the subjects in the data set it can be concluded that they cross the obesity range.

(source: https://www.cdc.gov/obesity/basics/adult-defining.html (https://www.cdc.gov/obesity/basics/adult-defining.html))

Analyzing the BMI to determine how it varies with the sex of the subjects

Using a bar plot as the variables used here (age and bmi) are both float data types.

```
ggplot(data=medicalInsurance,aes(x=age, y=bmi, fill=sex)) +
   ggtitle("BMI depending on age and sex")+
  facet_wrap(~sex)+
  geom_boxplot()+
  labs(caption = "Figure 4")
```

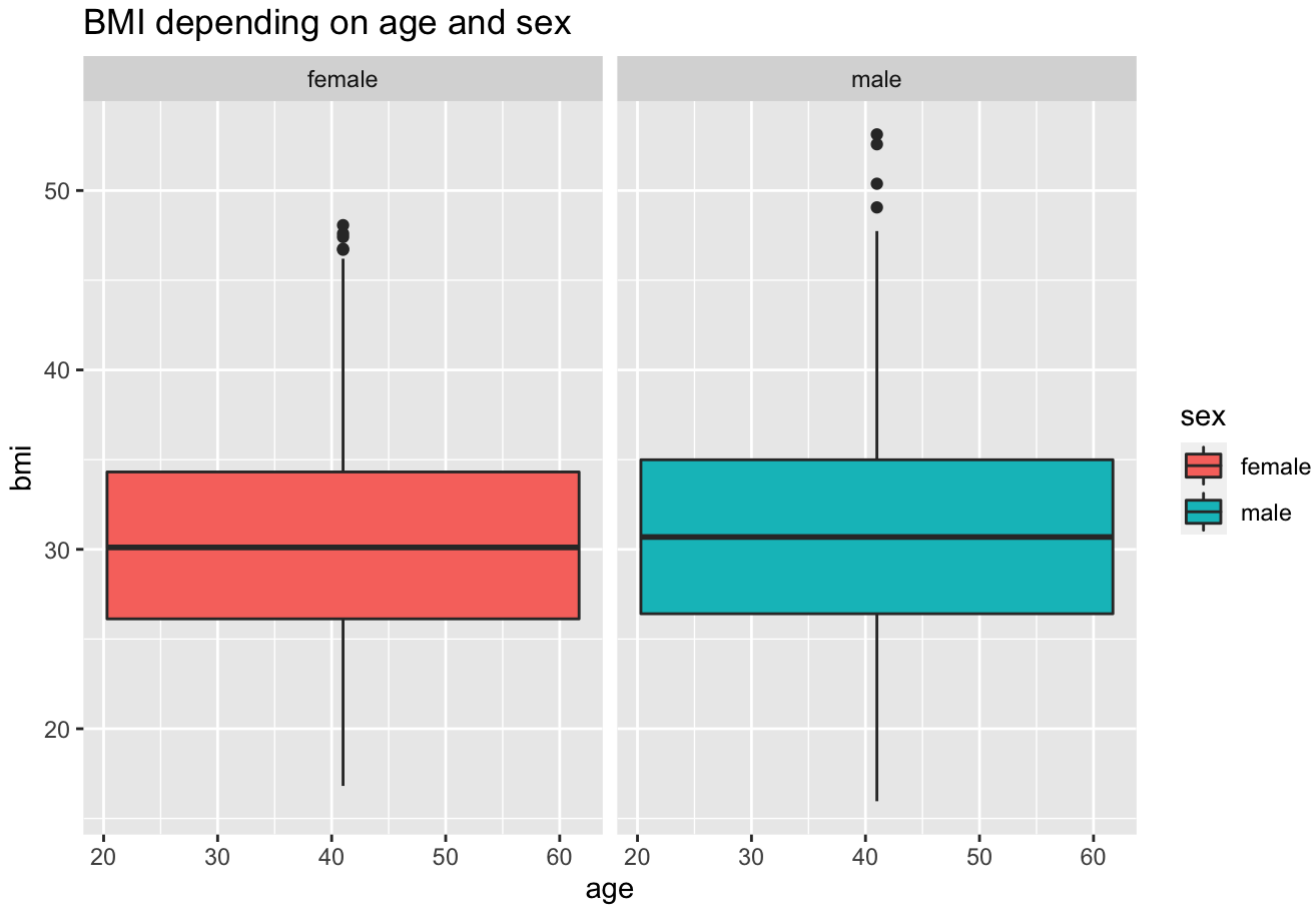

Figure 4

The median line for the male BMI is slightly higher than that for th female sex but other than that the graph doesn't tell us much regarding how age affects bmi.

Grouping the data by sex to further determine how the BMI varies:

```
group_by(medicalInsurance, sex) %>%
  summarize(average_BMI=(round(mean(bmi),digits=2)))
```

```
## # A tibble: 2 × 2
##   sex    average_BMI
##   <fct>        <dbl>
## 1 female        30.4
## 2 male          30.9
```

The average BMI for male and female subjects in the data barely differ as the average BMI for female subjects= 30.38 and male subjects= 30.94.

Determining how the charges would vary depending on the average BMI of the male and female subjects

```
ggplot(data=medicalInsurance,aes(x=bmi, y=charges, fill=sex,alpha=0.5)) +
   ggtitle("Charges depending on BMI")+
  geom_violin() + labs(caption = "Figure 5")
```

```
## Warning: position_dodge requires non-overlapping x intervals
```

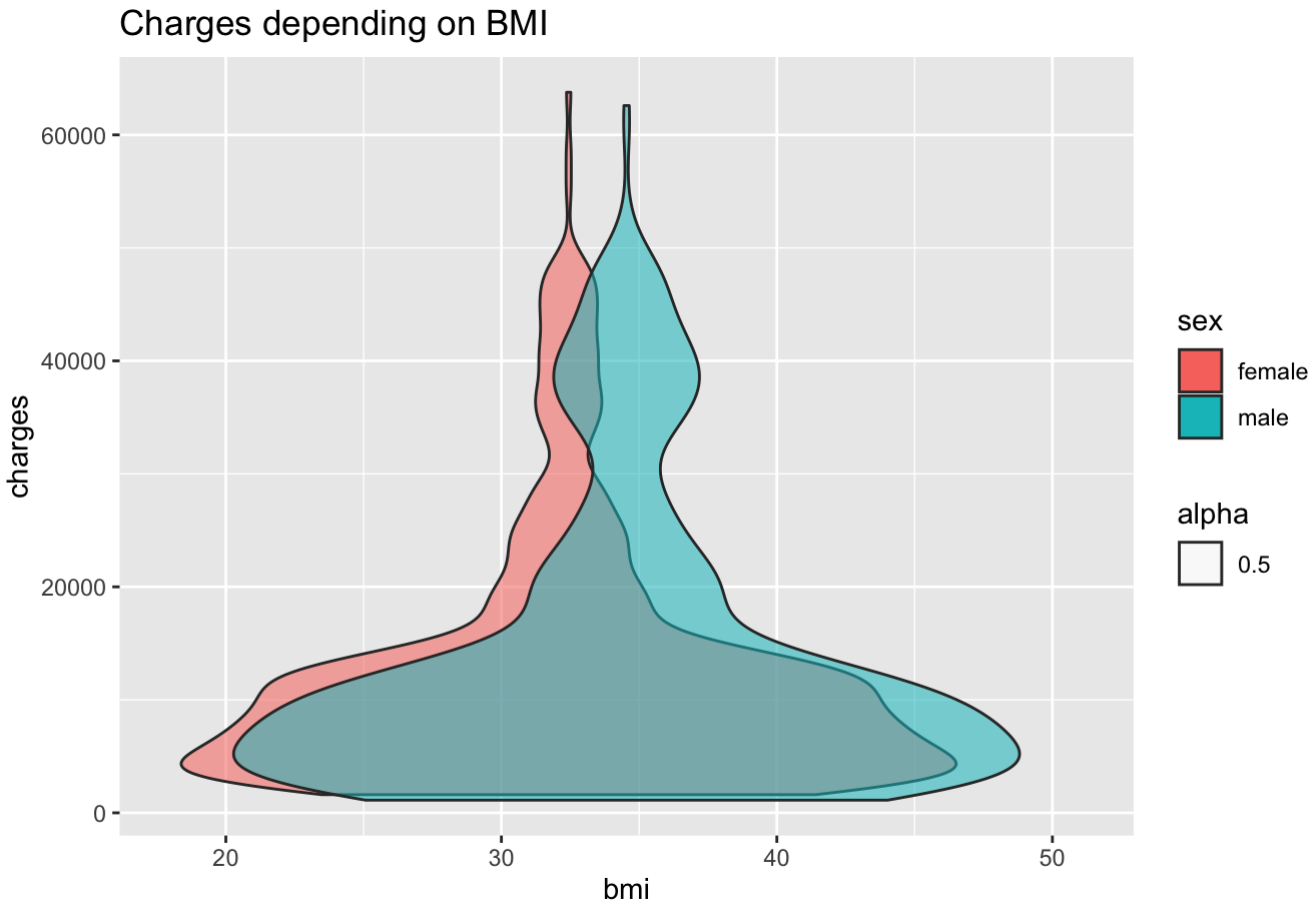

Figure 5

Although it is difficult to analyze the plots by just looking at them I was able to make these conclusions:

1. Males with a BMI between 30-40(obese), especially those with a BMI around 35 are charged significantly more.
2. Females with a BMI between 30-35(obese) are charged significantly more.
3. Females have a higher maximum charge amount compared to males.

To accurately find how the charges differ I'm using a t-test that will determine if there is a statistical difference between the means of the two groups:

```
t.test(medicalInsurance$bmi,medicalInsurance$charges)
```

```
##
##  Welch Two Sample t-test
##
## data:  medicalInsurance$bmi and medicalInsurance$charges
## t = -39.991, df = 1337, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -13889.23 -12590.29
## sample estimates:
##  mean of x  mean of y
##    30.6634 13270.4223
```

The 95% confidence interval tells us the estimated difference between the means of the two variables bmi and charges is between -13889.23 -12590.29 and this compares to the plot in figure 3.

Since the BMI for the subjects in the data is categorized as obese how would this affect the charges?

Grouping the data set into two groups: one where the BMI is less than 25 and another where it is greater than or equal to 25 to determine how the charges vary

```
group_by(medicalInsurance, bmi>30) %>%
  summarize(average_BMI_Charges=(round(mean(charges),digits=2)))
```

```
## # A tibble: 2 × 2
##   `bmi > 30` average_BMI_Charges
##   <lgl>                    <dbl>
## 1 FALSE                   10719.
## 2 TRUE                    15561.
```

The average annual medical charge for those with a BMI below 30 (not considered obese)= $10719.39

The average annual medical charge for those with a BMI above 30 (considered obese)= $15560.93

The difference is significantly more for those who are considered obese than those who are not.

Conclusions drawn by analyzing the number of children

Determining if having children and the number of children affects how much medical insurance costs.

Using a bar chart to find the distribution of children:

```
ggplot(data=medicalInsurance, aes(x=children,fill="red"))+
  xlab("number of children")+
  ylab("frequency")+
  geom_bar() +
  ggtitle("Frequency of children") + labs(caption = "Figure 6")
```

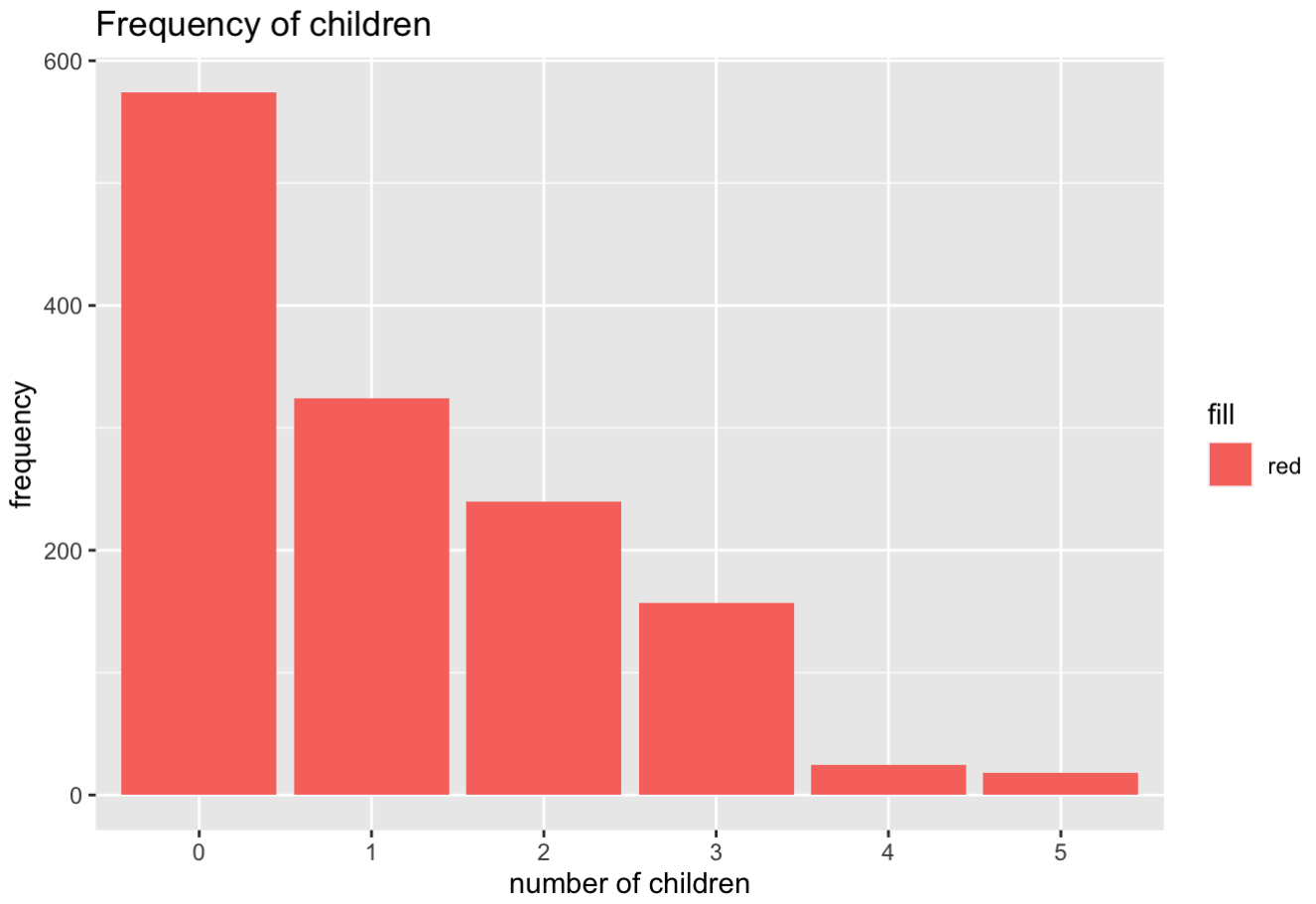

Figure 6

From the graph it can be concluded that the minimum number of children is 0 and the maximum is 5 and 0 kids is the most frequently occurring in the data set.

Using the data found above to graph a violin plot to determine how children affect medical insurance cost:

```
ggplot(data=medicalInsurance,aes(x=children, y=charges,fill="red")) +
   ggtitle("Charges depending on number of children")+
  geom_violin()+labs(caption = "Figure 7")
```

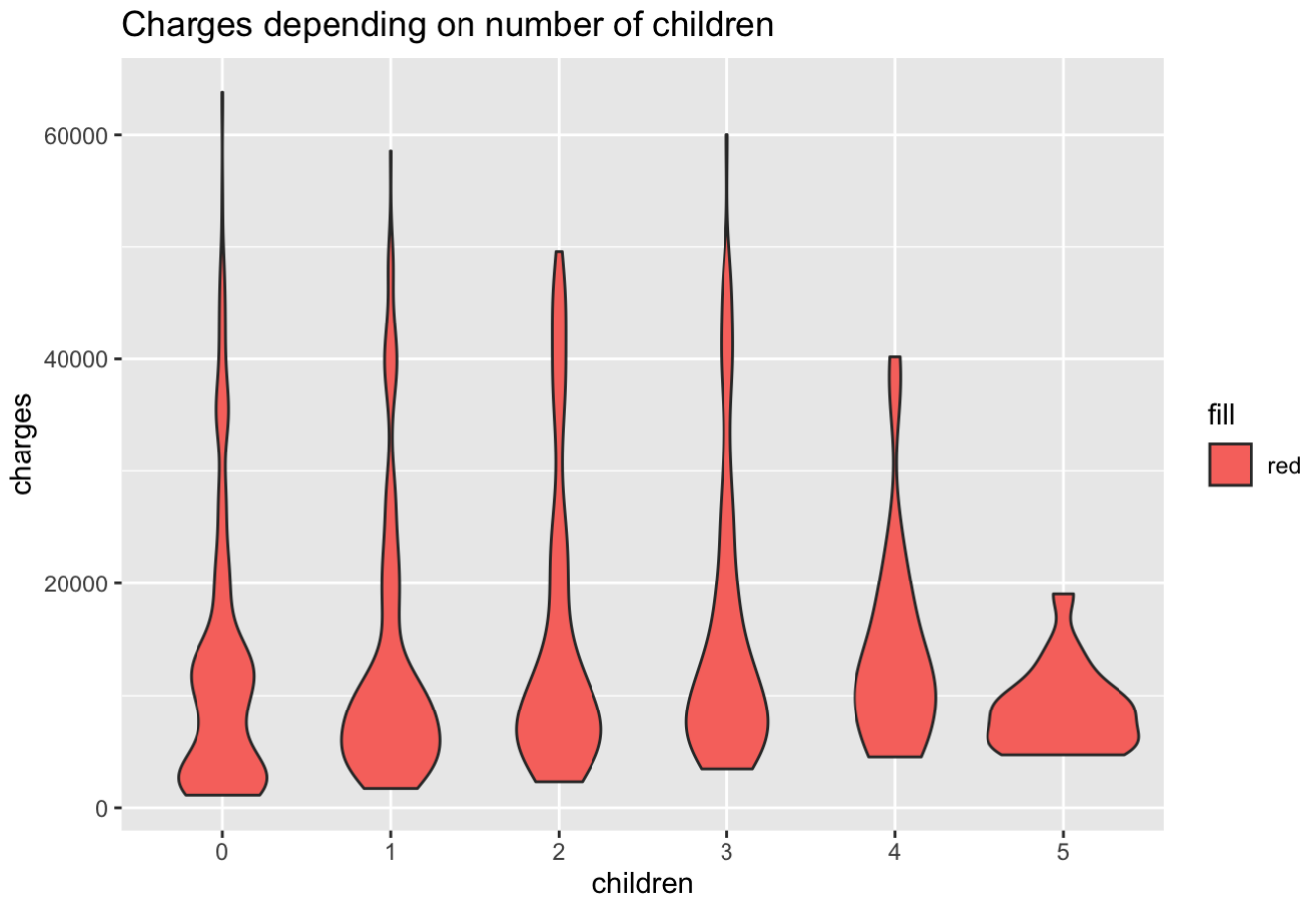

Figure 7

From the violin plot it can be concluded that having the most amount of children (in this case 5 from the data set) has the lowest charge.

```
group_by(medicalInsurance, children) %>%
  summarize(average_children_charges=(round(mean(charges),digits=2)))
```

```
## # A tibble: 6 × 2
##   children average_children_charges
##   <fct>                       <dbl>
## 1 0                          12366.
## 2 1                          12731.
## 3 2                          15074.
## 4 3                          15355.
## 5 4                          13851.
## 6 5                           8786.
```

The subjects that have 3 or 4 children are charged the most whereas those who have 5 are charged the least which I found really odd since having more children should increase the medical insurance charges.

A question that can be raised through this data set is analyzing why having more children decreases the total medical insurance charges instead of increasing it.

Conclusions drawn by analyzing the number of subjects who smoke

Questions to answer:

1. What age smokes the most?

2. Which gender smokes the most?

I looked at the age and sex of the subjects in the dataset and there was not a significant difference hence I chose to remove it from my analysis.

Using which and length to find the percentage of subjects that smoke compared to the total subjects in the data set.

```
numOfSmokers<-length(which(medicalInsurance$smoker=="yes"))
numOfSmokers
```

```
## [1] 274
```

```
round(numOfSmokers/(length(medicalInsurance$smoker))*100,digits=2)
```

```
## [1] 20.48
```

From the 1338 subjects in the data set the 274 subjects are smokers, which is approximately only 20.48% of the dataset

Comparing the charges of the 274 smokers (20.48% of the subjects) to the rest of the non smokers in the dataset

Using a bar graph to determine the charges for those who smoke compared to those who do not. The charges for those who do not smoke should be considerbly less than those who do.

```
ggplot(data=medicalInsurance,aes(x= smoker,y=charges, fill=sex)) +
  geom_boxplot() + #facet_wrap(~Species)+
  xlab("Do the subjects smoke?")+
  ggtitle("Charges for smokers compared to non smokers") + labs(caption = "Figure 8")
```

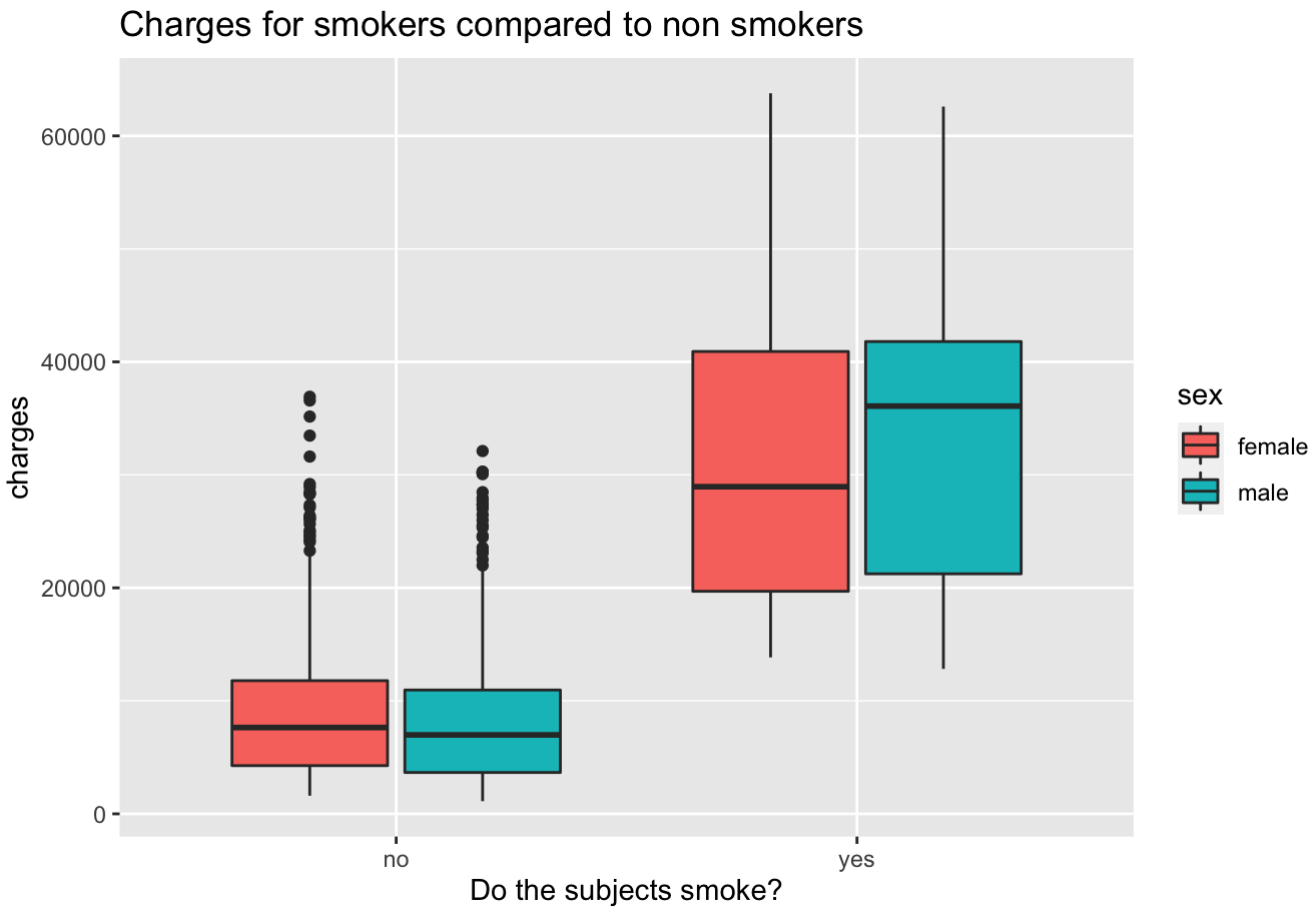

Figure 8

- Those who smoke are charged drastically more for their insurance compared to those who do not.
- The median line for male smokers is considerably higher than the one for females who smoke.
- There are some outliers for those who do not smoke but are charged higher. A future analysis could be made on why these outliers exist and what are their causes.
- Finding the average charge for smokers and how it varies for non-smokers and smokers

```
group_by(medicalInsurance, smoker=="yes") %>%
  summarize(average_smoker_charges=(round(mean(charges),digits=2)))
```

```
## # A tibble: 2 × 2
##   `smoker == "yes"` average_smoker_charges
##   <lgl>                             <dbl>
## 1 FALSE                              8434.
## 2 TRUE                              32050.
```

- The average annual medical charge for smokers is= $32050.23
- The average annual medical charge for nonsmokers is= $8434.27
- Those who smoke are have a significantly higher medical insurance charge than those who do not which agrees with the hypothesis that I made.

How do the charges for medical insurance vary depending on the region

Determine what are the different regions given in the data set by using unique():

```
unique(medicalInsurance$region)
```

```
## [1] southwest southeast northwest northeast
## Levels: northeast northwest southeast southwest
```

Determining the charges distributed over the region using a violin plot:

```
ggplot(data=medicalInsurance, aes(x=region,y=charges,fill=sex))+
  xlab("region")+
  ylab("charges")+
  geom_violin() +
  ggtitle("Charges depending on region") + labs(caption = "Figure 9")
```

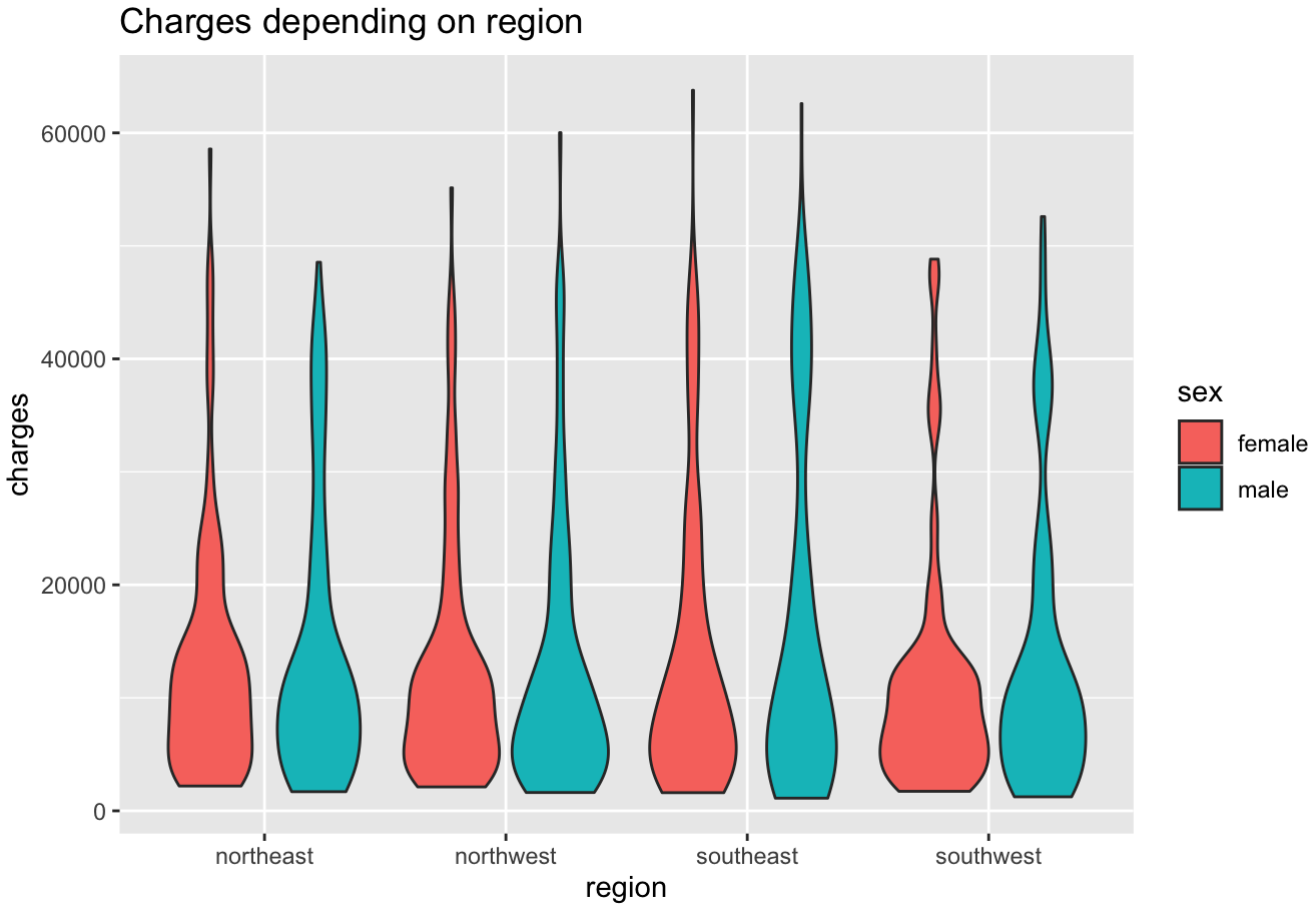

Figure 9

- The plot doesn't tell us much about the distribution of charges but some of the conclusions I derived were:

1. Subjects in the southeast pays the most for their medical insurance with women being charged more than men but not by a significant amount.
2. Female subjects in the northeast pay significantly more than the male subjects.
3. In the northwest and southwest male subjects pay significantly more than the female subjects.

To be able to determine more from this plot I will be using an anova test as it will tell me if there are any statistical differences between the means of the different regions (as they are more than 2 groups).

```
aov.mod<-aov(charges~region,data=medicalInsurance)
summary(aov.mod)
```

```
##               Df    Sum Sq   Mean Sq F value Pr(>F)
## region         3 1.301e+09 433586560    2.97 0.0309 *
## Residuals   1334 1.948e+11 146007093
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA tells us there is a significant difference between the means.

Using TukeyHSD to further summaries the individual relationship between the grouped variables and the charges.

```
TukeyHSD(aov.mod)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = charges ~ region, data = medicalInsurance)
##
## $region
##                            diff         lwr        upr     p adj
## northwest-northeast  -988.8091 -3428.93434 1451.31605 0.7245243
## southeast-northeast  1329.0269 -1044.94167 3702.99551 0.4745046
## southwest-northeast -1059.4471 -3499.57234 1380.67806 0.6792086
## southeast-northwest  2317.8361   -54.19944 4689.87157 0.0582938
## southwest-northwest   -70.6380 -2508.88256 2367.60656 0.9998516
## southwest-southeast -2388.4741 -4760.50957  -16.43855 0.0476896
```

The p-value here tells us there is a high significant difference between the mean charges of southwest and southeast.

To find the breakdown of the charges based on region, I grouped the the data by region to see the charges associated with each region.

```
group_by(medicalInsurance, region) %>%
  summarize(average_region_charges=(round(mean(charges),digits=2)))
```

```
## # A tibble: 4 × 2
##   region    average_region_charges
##   <fct>                      <dbl>
## 1 northeast                 13406.
## 2 northwest                 12418.
## 3 southeast                 14735.
## 4 southwest                 12347.
```

The medical insurance charges do not vary significantly over the region distribution. My hypothesis from the graph above on southeast being charged the highest is correct as it has the highest increased average region charges= $14735.41.

Conclusions from the calculations done on the data set:

- Sex of the subjects by itself doesn't play an important role in the determination of charges however it is important when analyzing other factors such as age, bmi and smoking.

- The average BMI of the subjects in the database being categorized as obese influenced the total medical insurance charged to them as it was higher especially when age was also considered as a factor
- The variable children doesn't affect this data set however other calculations can be done with the variable children if I had more data as the number of children and the amount charged doesn't co-relate the way I hypothesized and hence this could be a future analysis that can be conducted.
- Using plots was a great way to visualize the data set however some of them were too similar to read/compare hence using hypothesis testing such as t-tests and anova gave me a better statistical reading.
- Breaking down region determined which region is charged the most and how sex effects those charges however over all the average charges were not significantly different but could be explored further with more data findings.