DATA606 DATA PROJECT PROPOSAL

Title: CUNY SPS MDS DATA606_Project Proposal"

Author: Charles Ugiagbe

Date: 11/6/2021

Data Preparation

Load required libraries:

```
library(tidyverse)
library(patchwork)
library(ggforce)
library(statsr)
```

load data from Github

```
url <- "https://raw.githubusercontent.com/omocharly/DATA606_PROJECT/main/insurance.csv"
insurance <- read.csv(url)
```

Take a look at the head of the data

```
head(insurance)
```

```
##   age    sex    bmi children smoker    region   charges
## 1  19 female 27.900        0    yes southwest 16884.924
## 2  18   male 33.770        1     no southeast  1725.552
## 3  28   male 33.000        3     no southeast  4449.462
## 4  33   male 22.705        0     no northwest 21984.471
## 5  32   male 28.880        0     no northwest  3866.855
## 6  31 female 25.740        0     no southeast  3756.622
```

Take a glimpse look at the dataset

Dataset has 7 variable and 1338 Observation

```
glimpse(insurance)
```

```
## Rows: 1,338
## Columns: 7
## $ age      <int> 19, 18, 28, 33, 32, 31, 46, 37, 37, 60, 25, 62, 23, 56, 27, 1~
## $ sex      <chr> "female", "male", "male", "male", "male", "female", "female",~
## $ bmi      <dbl> 27.900, 33.770, 33.000, 22.705, 28.880, 25.740, 33.440, 27.74~
## $ children <int> 0, 1, 3, 0, 0, 0, 1, 3, 2, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0~
## $ smoker   <chr> "yes", "no", "no", "no", "no", "no", "no", "no", "no", "no", ~
## $ region   <chr> "southwest", "southeast", "southeast", "northwest", "northwes~
## $ charges  <dbl> 16884.924, 1725.552, 4449.462, 21984.471, 3866.855, 3756.622,~
```

Research question

This project aims to:

1. *Determine if the mean insurance charges of male individuals in the dataset is significantly different from the mean charges of female*
2. *Determine if the mean insurance charges of Smokers in the dataset is different from the mean charges of Non - smokers*
3. *Formulate a multiple Regression model or predicting the insurance charges of individuals*

Cases

There are 7 variables and 1338 observations in the dataset. six(6) of the Variable in the dataset are potential predictor of the of the 7th variables (Insurance charges). There are no missing value in any of the observation. Each observation represents the likely variable that play vital roles in determining the insurance charge

Data collection

This dataset was downloaded from kaggle and then uploaded to my github repository. The data can be accessed directly from the repository at Github

Type of study

This is an observational study as there is no control group.

Data Source

Data is from kaggle public datasets and can be found online here: https://www.kaggle.com/mirichoi0218/insurance

Response Variable (Dependent Variable)

The Dependent variable is the Insurance Charges and its numerical

Predictor Variables (Independent Variables)

There are six(6) independent used. They independent variables are: Age(numeric), sex(numeric), BMI(numeric), Children(numeric), Smoker(categorical), Region(categorical)

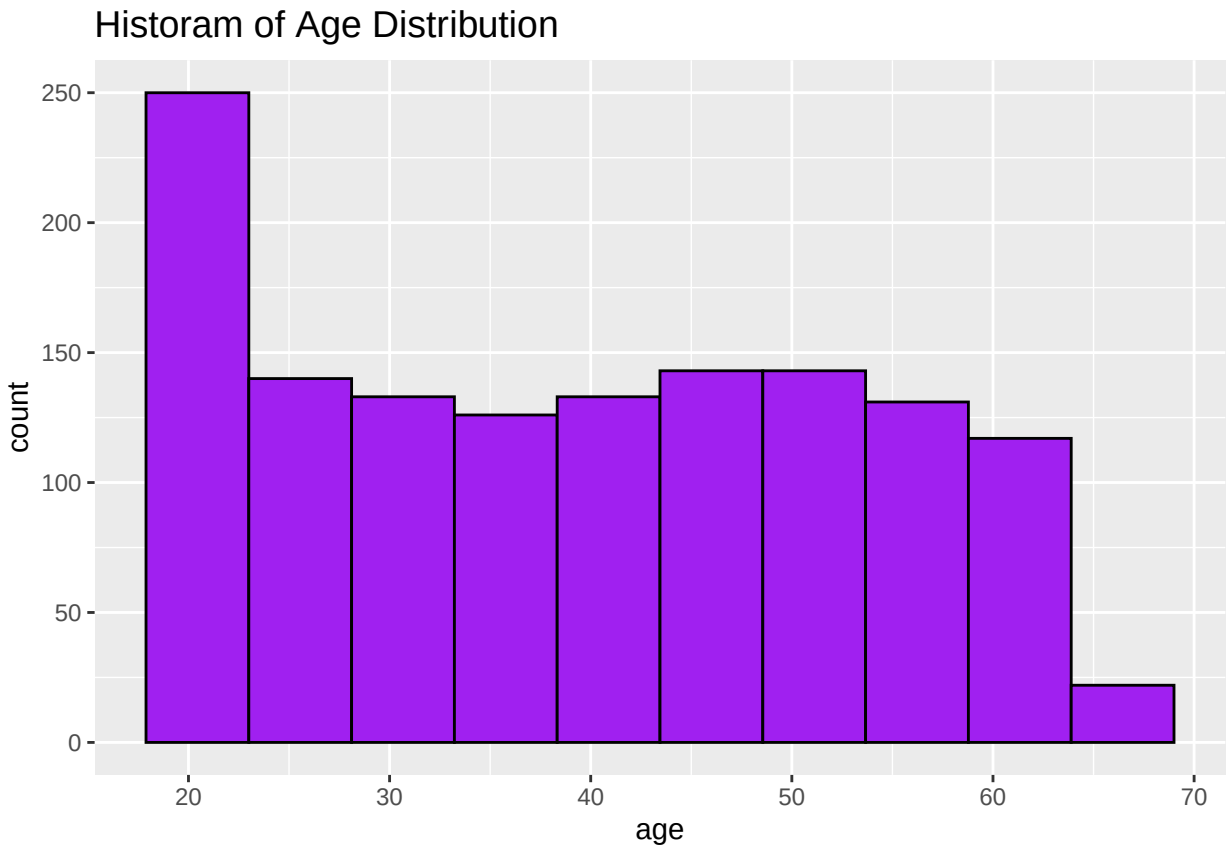
Relevant Summary Statistics

```
summary(insurance)
```

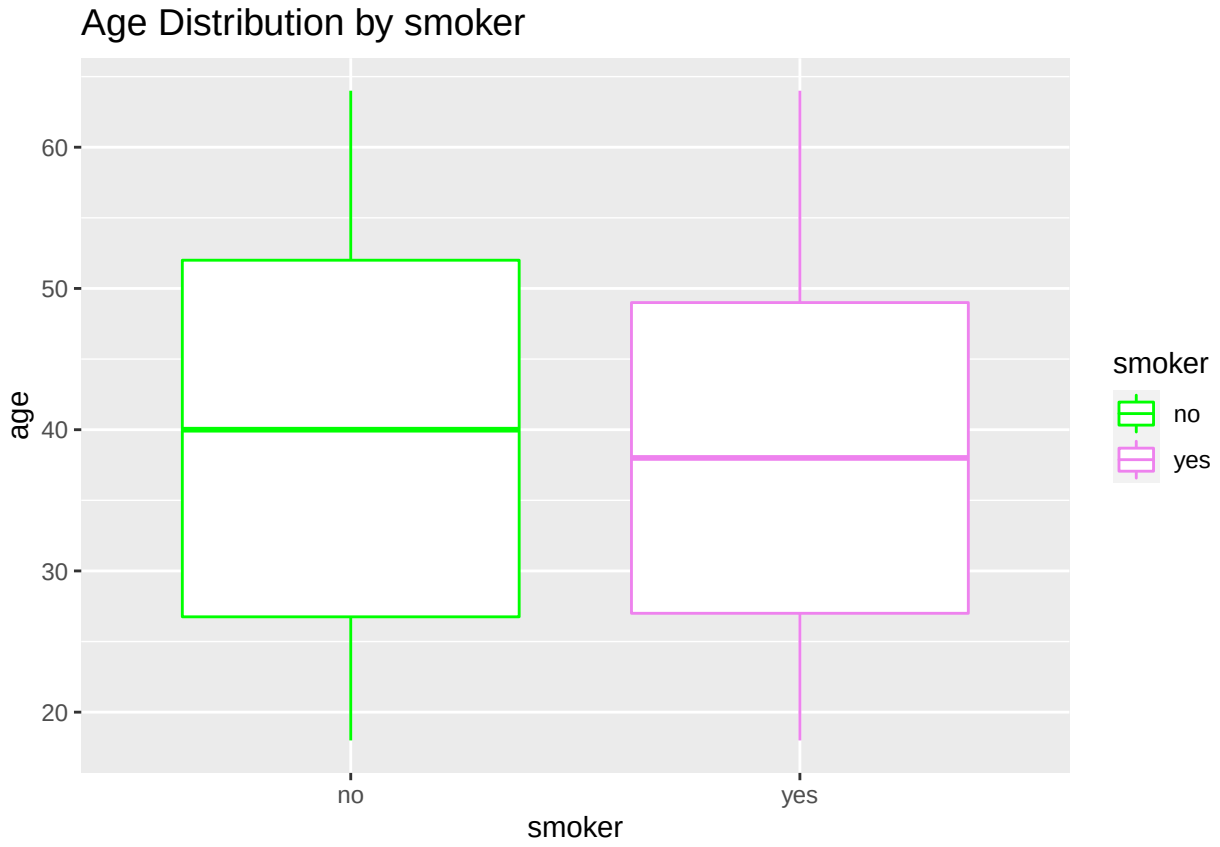
```
##       age            sex                 bmi           children
##  Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000
##  1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000
##  Median :39.00   Mode  :character   Median :30.40   Median :1.000
##  Mean   :39.21                      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
```

Visualizations and Exploratory data analysis

```
plot1<-ggplot(insurance, aes(x=age))+
geom_histogram(color="black", fill="purple", bins=10)+
labs(title="Historam of Age Distribution")+
theme(plot.title = element_text(size=14))
plot1
```




```
plot2<-ggplot(insurance, aes(x=smoker, y=age, color=smoker)) +
geom_boxplot()+
scale_color_manual(values=c('green', "violet"))+
labs(title="Age Distribution by smoker")+
theme(plot.title = element_text(size=14))
plot2
```



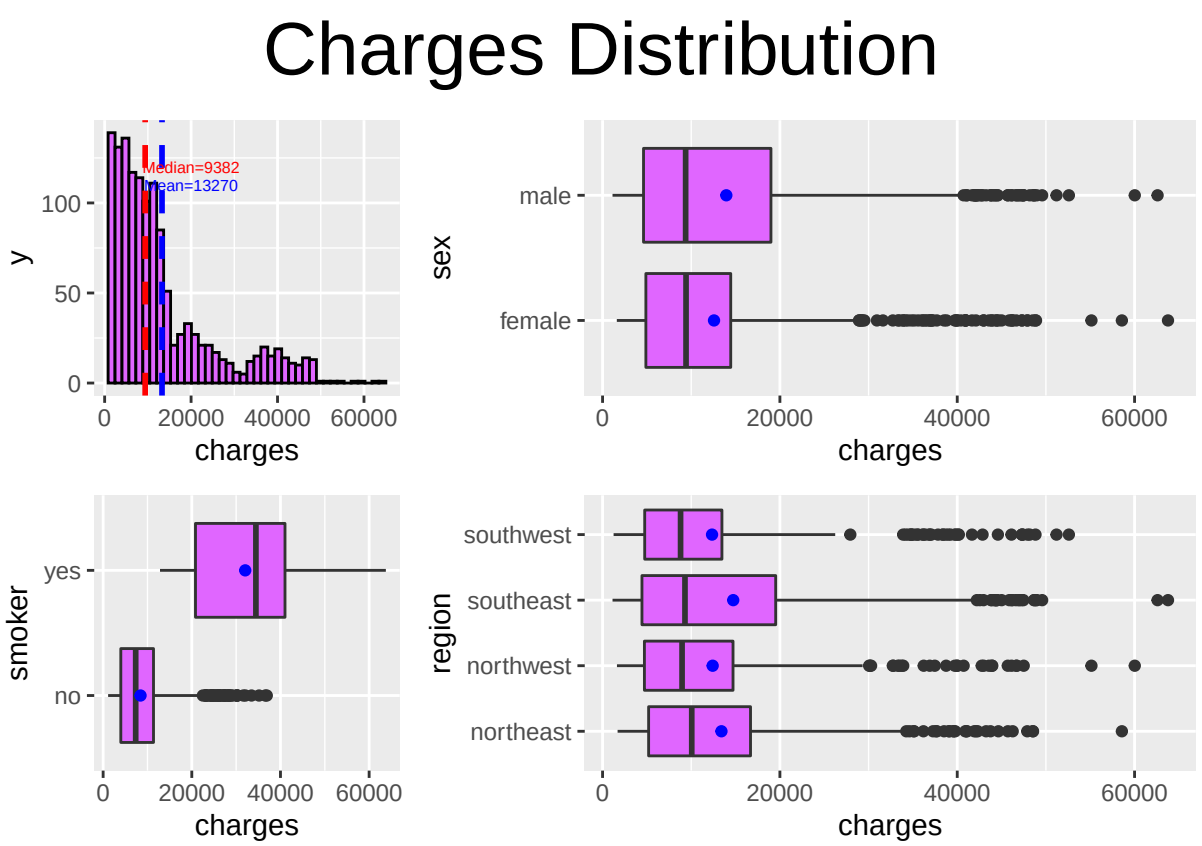

```
p1<-ggplot(insurance, aes(x=charges))+
geom_histogram(color="black", fill="mediumorchid1", bins=40)+
geom_vline(aes(xintercept= 13270), color="blue", linetype="dashed", size=1)+
geom_vline(aes(xintercept= 9382), color="red", linetype="dashed", size=1)+
annotate("text", x= 20000, y=110, size=2, label="Mean=13270", color="blue")+
annotate("text", x= 20000, y=120, size=2, label="Median=9382", color="red")


p2<-ggplot(insurance, aes(x=sex, y=charges)) +
geom_boxplot(fill="mediumorchid1")+
stat_summary(fun=mean, geom="point", color="blue")+
coord_flip()


p3<-ggplot(insurance, aes(x=smoker, y=charges)) +
geom_boxplot(fill="mediumorchid1")+
stat_summary(fun=mean, geom="point", color="blue")+
coord_flip()

p4<-ggplot(insurance, aes(x=region, y=charges)) +
geom_boxplot(fill="mediumorchid1")+
stat_summary(fun=mean, geom="point", color="blue")+
coord_flip()
```

```
options(repr.plot.width=9, repr.plot.height=20)
layout<-"
ABB
CDD"
p1 + p2 + p3 + p4 + plot_layout(design = layout)+
plot_annotation(title="Charges Distribution",
                theme = theme(plot.title = element_text(size = 28, hjust = 0.5)))
```




```
### age, charges, sex
plot3<-ggplot(insurance, aes(x=sex, y=age, color=sex)) +
geom_sina()+
scale_color_manual(values=c('hotpink', "royalblue"))+
labs(title="Age Distribution by sex")+
theme(plot.title = element_text(size=14))

plot4<-ggplot(insurance, aes(x=age, y=charges, color= sex))+
geom_jitter(alpha=0.3, size=2.5)+
scale_color_manual(values=c('hotpink', "royalblue"))+
geom_rug()+
geom_smooth(method=lm, formula=y~x)+
labs(title="Age x Charges by sex")+
theme(plot.title = element_text(size=14))

### age, charges, smoker
plot5<-ggplot(insurance, aes(x=smoker, y=age, color=smoker)) +
```

```
geom_sina()+
scale_color_manual(values=c('grey', "brown"))+
labs(title="Age Distribution by smoker")+
theme(plot.title = element_text(size=14))

plot6<-ggplot(insurance, aes(x=age, y=charges, color= smoker))+
geom_jitter(alpha=0.3, size=2.5)+
scale_color_manual(values=c('brown', "grey"))+
geom_rug()+
geom_smooth(method=lm, formula=y~x)+
labs(title="Age x Charges by smoker")+
theme(plot.title = element_text(size=14))


options(repr.plot.width=10, repr.plot.height=35)
layout<-"
ABB
CDD"
plot3 + plot4 + plot5 + plot6 + plot_layout(design = layout)
```