Data Analysis in R

Truong Thi An Hai

11/3/2020

Load data into R and rename column names

```
#Read the zipIncome into R
My_Data = read.delim("https://hyper.mephi.ru/assets/courseware/v1/94f633ca057a1aa84db0364cf4bfa81d/asset-v1:MEPhIx+CS712DS+2020Fall+type@asset+block/zipIncome.txt", sep='|')
#Remove last row (which contains the total number of rows)
My_Data = My_Data[-nrow(My_Data),]
#Display the column names of the data
colnames(My_Data)
```

```
## [1] "zip_prefixes"       "meanhouseholdincome"
```

```
#Change the column names
names(My_Data) <- c("zipCode","income")
#Results
colnames(My_Data)
```

```
## [1] "zipCode" "income"
```

Analyze the summary of the data

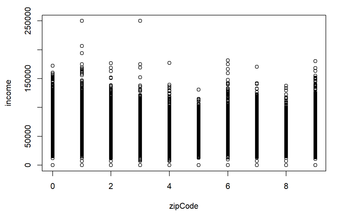
```
summary(My_Data)
```

```
##    zipCode              income
##  Length:32038       Min.   :     0
##  Class :character   1st Qu.: 37644
##  Mode  :character   Median : 44163
##                     Mean   : 48245
##                     3rd Qu.: 54373
##                     Max.   :250000
```

The numerical value of mean for mean household income is 48245 The numerical value of median for household income is 44163

Plot a scatter plot of the data

```
plot(My_Data)
```



There seem to be two outlier values are 0 and 250000

Create a subset of the data

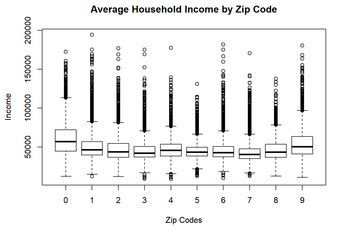
```
newData = subset(My_Data, income<200000 & income >7000)
#Analyze the summary of the new data
summary(newData)
```

```
##    zipCode              income
##  Length:31871       Min.   :  8465
##  Class :character   1st Qu.: 37755
##  Mode  :character   Median : 44234
##                     Mean   : 48465
##                     3rd Qu.: 54444
##                     Max.   :194135
```

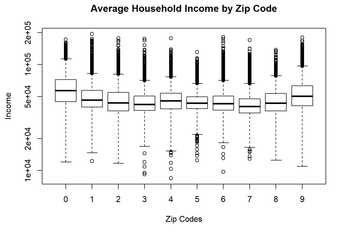
The numerical value of the mean after filtration is 48465

Create a simple box plot

```
boxplot(col="white", data = newData, income ~ zipCode, main = "Average Household Income by Zip Code", xlab = "Zip Codes", ylab = "Income")
```

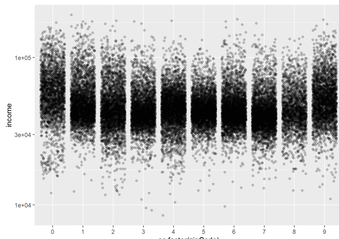



```
#Create a new box plot where the y-axis uses a log scale
boxplot(col="white", data = newData, income ~ zipCode, main = "Average Household Income by Zip Code", xlab = "Zip Codes", ylab = "Income", log='y')
```

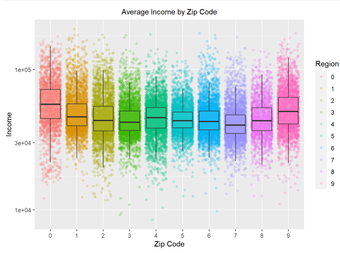


Use the ggplot library in R

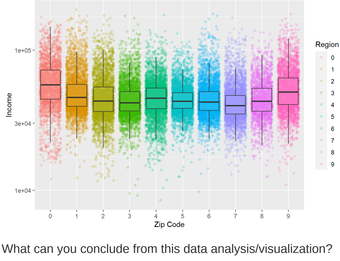
```
#Make a ggplot that consists of just a scatter plot using the function geom_point() with position = "jitter"
ggplot(newData, aes(x = as.factor(zipCode), y=income))+geom_point(position="jitter", alpha=0.2)+scale_y_log10()
```




```
#Create a new ggplot by adding a box plot layer to your previous graph.  To do this, add the ggplot function geom_boxplot(). Also, add color to the scatter plot so that data points between different zip codes are different colors
ggplot(newData,aes(x=as.factor(zipCode),y=income))+geom_point(aes(colour=factor(zipCode)),position = 'jitter',alpha=0.2)+ geom_boxplot(alpha=0.1,outlier.size =-Inf) + scale_y_log10()+labs(color="Region",x="Zip Code",y="Income",title="Average Income by Zip Code") + theme(plot.title = element_text(size =11, face="plain",hjust = 0.5))
```




```
#Another method
ggplot(newData,aes(x=as.factor(zipCode),y=income))+geom_point(aes(colour=factor(zipCode)),position = 'jitter',alpha=0.2)+ geom_boxplot(alpha=0.1,outlier.size =-Inf) + scale_y_log10()+ ylab("Income") + xlab("Zip Code") + ggtitle ("Average Income by Zip Code") + labs(color="Region") + theme(plot.title = element_text(size =11, face="plain", hjust = 0.5))
```



What can you conclude from this data analysis/visualization?

- It is important to visualize your data in different ways.
- Visualization enables you to better understand what your data is telling you.
- Visualization enables you to better communicate your results to stakeholders.
- Zip codes starting in 0 (New England) and 9 (West Coast) have higher average household incomes.

THE END