DAB 501

EDA of data using RStudio

SUBMITTED TO: -

Adegoke Ojeniyi

SUBMITTED BY: -

Manish Kataria (W0865937)

Introduction

In this project we discuss about various function used in R studio to calculate different values and relation in columns. First, we explain our data the find minimum, maximum, mean, mode, median, percentiles, variance and standard deviation for each column and then find correlation and covariance between all columns.

Clean, Format and prepare the dataset

First, we have to clean the data by removing unwanted values. Then format the data by choosing necessary and that columns where we can perform various R functions at last we have to save as “csv” file then import into in R studios by following command.

```r
```{r}
data <- read.csv("C:/Users/manis/Desktop/DAB_501_tabacoo.csv", header = TRUE)
```
```

This command is use to call or import the data from their location.

List and explain the variables in the dataset

Now, we have a data that is based on tobacco usage in India. In this data we have data for each state and union territory, and for county. In this data we have 10 variables or columns. We can get the name of that column by using function str(). It tells us about structure of data.

```r
str(data)
```

```
'data.frame':   38 obs. of  10 variables:
 $ State.UT                                                 : chr  "India" "India" "India" "Andaman and Nicobar Islands" ...
 $ Area                                                     : chr  "Total" "Urban" "Rural" "Total" ...
 $ Ever.tobacco.users....                                   : num  18.1 13.5 19.5 21.5 7.3 63.1 16.2 21.5 10.3 18.8 ...
 $ Current.tobacco.users....                                : num  8.5 5.5 9.4 4.4 2.6 57.9 11.9 7.3 3 8 ...
 $ Ever.cigarette.users....                                 : num  4.6 4.1 4.7 4.4 2.6 47.9 5.3 4.8 1.1 6.3 ...
 $ Current.cigarette.users....                              : num  2.6 2.3 2.6 1.3 1.1 45.9 4.3 2.3 0.6 4.5 ...
 $ Median.age.of.initiation.of.Cigarette..in.years.         : num  11.5 11.2 11.6 12 12.6 8.7 10.9 7.9 9.5 8.2 ...
 $ Ever.tobacco.smokers.who.quit.in.last.12.months....      : num  10.6 14.6 9.8 19.8 12.8 22.6 11.6 5.1 10.4 18.5 ...
 $ Current.tobacco.smokers.who.tried.to.quit.smoking.in.the.past.12.months....: num  20 29.3 18.6 32.9 36.6 35.1 14.9 23.6 22.6 23.1 ...
 $ Current.tobacco.smokers.who.wanted.to.quit.smoking.now.....              : num  20.6 25.7 19.8 30.9 30.4 23.8 22.9 28.9 13.4 29 ...
```

We have columns: -

- State/UT
- Area
- Ever tobacco users (%)
- Current tobacco users (%)
- Ever cigarette users (%)
- Current cigarette users (%)
- Median age of initiation of Cigarette (in years)
- Ever tobacco smokers who quit in last 12 months (%)
- Current tobacco smokers who tried to quit smoking in the past 12 months (%)
- Current tobacco smokers who wanted to quit smoking now (%)

Minimum value for each variable in the dataset

Here we find minimum value for each variable by using function min ().

```r
min(data$Ever.tobacco.users....)
min(data$Current.tobacco.users....)
min(data$Ever.cigarette.users....)
min(data$Current.cigarette.users....)
min(data$Median.age.of.initiation.of.Cigarette..in.years.)
min(data$Ever.tobacco.smokers.who.quit.in.last.12.months....)
min(data$Current.tobacco.smokers.who.tried.to.quit.smoking.in.the.past.12.months....)
min(data$Current.tobacco.smokers.who.wanted.to.quit.smoking.now......)
```

```
[1] 3.8
[1] 1
[1] 1.1
[1] 0
[1] 7
[1] 0
[1] 0.1
[1] 0.1
```

Maximum value for each variable in the dataset

Same as minimum here we find maximum value for each variable by using function max().

```r
max(data$Ever.tobacco.users....)
max(data$Current.tobacco.users....)
max(data$Ever.cigarette.users....)
max(data$Current.cigarette.users....)
max(data$Median.age.of.initiation.of.Cigarette..in.years.)
max(data$Ever.tobacco.smokers.who.quit.in.last.12.months....)
max(data$Current.tobacco.smokers.who.tried.to.quit.smoking.in.the.past.12.months....)
max(data$Current.tobacco.smokers.who.wanted.to.quit.smoking.now......)
```

```
[1] 89.7
[1] 57.9
[1] 76.9
[1] 45.9
[1] 14.5
[1] 37.2
[1] 99.9
[1] 99.9
```

Column Name	Minimum value	Maximum value
Ever tobacco users (%)	3.8	89.7

Current tobacco users (%)	1	57.9
Ever cigarette users (%)	1.1	76.9
Current cigarette users (%)	0	45.9
Median age of initiation of Cigarette (in years)	7	14.5
Ever tobacco smokers who quit in last 12 months (%)	0	37.2
Current tobacco smokers who tried to quit smoking in the past 12 months (%)	0.1	99.9
Current tobacco smokers who wanted to quit smoking now (%)	0.1	99.9

Mean value for each variable in the dataset

Now we have to find the mean value for each variable. Mean value is the average value of a column. In R studios we use mean() function to find the mean value. Here we have mean value for each column.

```r
mean(data$Ever.tobacco.users....)
mean(data$Current.tobacco.users....)
mean(data$Ever.cigarette.users....)
mean(data$Current.cigarette.users....)
mean(data$Median.age.of.initiation.of.Cigarette..in.years.)
mean(data$Ever.tobacco.smokers.who.quit.in.last.12.months....)
mean(data$Current.tobacco.smokers.who.tried.to.quit.smoking.in.the.past.12.months....)
mean(data$Current.tobacco.smokers.who.wanted.to.quit.smoking.now......)
```

```
[1] 23.15
[1] 11.49211
[1] 9.681579
[1] 5.810526
[1] 10.91842
[1] 15.35263
[1] 34.23421
[1] 34.22632
```

Median value for each variable in the dataset.

Median value is the middle value of column. In R studios we use mean() function.

```r
median(data$Ever.tobacco.users....)
median(data$Current.tobacco.users....)
median(data$Ever.cigarette.users....)
median(data$Current.cigarette.users....)
median(data$Median.age.of.initiation.of.Cigarette..in.years.)
median(data$Ever.tobacco.smokers.who.quit.in.last.12.months....)
median(data$Current.tobacco.smokers.who.tried.to.quit.smoking.in.the.past.12.months....)
median(data$Current.tobacco.smokers.who.wanted.to.quit.smoking.now......)
```

```
[1] 16.9
[1] 5.6
[1] 4.65
[1] 2.4
[1] 11.15
[1] 13.15
[1] 26.5
[1] 29.35
```

Mode value for each variable in the dataset.

In general mode is the most frequent value in column. In R studios we do not have any pre define function to find the mode value. We

have create the mode function by self as below.

```{r}
mymode <- function(m){
sort(table(m), decreasing = TRUE)[1]
}
```

Here we create a function i.e. mymode to calculate the mode. It tells us about mode value and how many times that value occurs. Mode for each column is below.

Column Name	Mean	Median	Mode
Ever tobacco users (%)	23.15	16.9	10.3 Count 2
Current tobacco users (%)	11.49	5.6	4.3 Count 2
Ever cigarette users (%)	9.68	4.65	4.8 Count 3
Current cigarette users (%)	5.81	2.4	2.3 Count 4
Median age of initiation of Cigarette (in years)	10.92	11.15	11.1 Count 6
Ever tobacco smokers who quit in last 12 months (%)	15.35	13.15	11.6 Count 2
Current tobacco smokers who tried to quit	34.23	26.5	0.1 Count 1

smoking in the past 12 months (%)			
Current tobacco smokers who wanted to quit smoking now (%)	34.22	26..35	0.1 Count 1

```{r}
mymode(data$Ever.tobacco.users....)
mymode(data$Current.tobacco.users....)
mymode(data$Ever.cigarette.users....)
mymode(data$Current.cigarette.users....)
mymode(data$Median.age.of.initiation.of.Cigarette..in.years.)
mymode(data$Ever.tobacco.smokers.who.quit.in.last.12.months....)
mymode(data$Current.tobacco.smokers.who.tried.to.quit.smoking.in.the.past.12.months....)
mymode(data$Current.tobacco.smokers.who.wanted.to.quit.smoking.now......)
```

```
10.3
   2
4.3
  2
4.8
  3
2.3
  4
11.1
   3
11.6
   2
0.1
  1
0.1
  1
```

Percentiles value for each variable in the dataset.

Percentiles tells the range of the values lies in particular percentage range. Like 0%-25%,25%-50%,50%-75%,75%-100%. In R we don't have any pre-define function to find the percentiles. We find it by

using quantile() function like,

```r
mypercentile <- quantile(data$Ever.tobacco.users....)
print(mypercentile)
mypercentile <- quantile(data$Current.tobacco.users....)
print(mypercentile)
mypercentile <- quantile(data$Ever.cigarette.users....)
print(mypercentile)
mypercentile <- quantile(data$Current.cigarette.users....)
print(mypercentile)
mypercentile <- quantile(data$Median.age.of.initiation.of.Cigarette..in.years.)
print(mypercentile)
mypercentile <- quantile(data$Ever.tobacco.smokers.who.quit.in.last.12.months....)
print(mypercentile)
mypercentile <- quantile(data$Current.tobacco.smokers.who.tried.to.quit.smoking.in.the.past.12.months....)
print(mypercentile)
mypercentile <- quantile(data$Current.tobacco.smokers.who.wanted.to.quit.smoking.now.....)
print(mypercentile)
```

```
    0%    25%    50%    75%   100%
 3.800 11.925 16.900 24.275 89.700
    0%    25%    50%    75%   100%
 1.000  4.150  5.600  9.175 57.900
    0%    25%    50%    75%   100%
 1.100  3.175  4.650  6.825 76.900
    0%    25%    50%    75%   100%
 0.000  1.325  2.400  3.925 45.900
    0%    25%    50%    75%   100%
 7.000 10.025 11.150 11.900 14.500
    0%    25%    50%    75%   100%
 0.000  9.725 13.150 22.175 37.200
    0%    25%    50%    75%   100%
 0.100 20.175 26.500 46.675 99.900
    0%    25%    50%    75%   100%
 0.100 20.450 29.350 40.175 99.900
```

Column Name	0%-25%	25%-50%	50%-75%	75%-100%
Ever tobacco users (%)	3.800 - 11.925	11.925 - 16.900	16.900 - 24.275	24.275 - 89.700
Current tobacco users (%)	1.000 - 4.150	4.150 - 5.600	5.600 - 9.175	9.175 - 57.900
Ever cigarette users (%)	1.100 - 3.175	3.175 - 4.650	4.650 - 6.825	6.825 - 76.900
Current cigarette users (%)	0.000 - 1.325	1.325 - 2.400	2.400 - 3.925	3.925 - 45.900
Median age of initiation of Cigarette (in years)	7.000 - 10.025	10.025 - 11.150	11.150 – 11.900	11.900- 14.500
Ever tobacco smokers who quit in last 12 months (%)	0.000 - 9.725	9.725 – 13.15	13.15 – 22.175	22.175 – 37.200

Current tobacco smokers who tried to quit smoking in the past 12 months (%)	0.100 – 20.175	20.175 – 26.500	26.500 – 46.675	46.675 – 99.900
Current tobacco smokers who wanted to quit smoking now (%)	0.100 – 20.450	20.450 – 29.250	29.350 – 40.175	40.175 – 99.900

Variance for each variable in the dataset.

Variance is the average spread of values from the mean of that column. We use var() function to find the variance.

```
```{r}
var(data$Ever.tobacco.users....)
var(data$Current.tobacco.users....)
var(data$Ever.cigarette.users....)
var(data$Current.cigarette.users....)
var(data$Median.age.of.initiation.of.Cigarette..in.years.)
var(data$Ever.tobacco.smokers.who.quit.in.last.12.months....)
var(data$Current.tobacco.smokers.who.tried.to.quit.smoking.in.the.past.12.months....)
var(data$Current.tobacco.smokers.who.wanted.to.quit.smoking.now......)
```
```

```
[1] 331.4561
[1] 203.058
[1] 226.8264
[1] 95.23502
[1] 2.986949
[1] 90.63661
[1] 473.1953
[1] 466.5242
```

Standard deviation for each variable in the dataset.

It is the average difference between central point (we also called median) to other point. we use sd() function to find that.

```r
sd(data$Ever.tobacco.users....)
sd(data$Current.tobacco.users....)
sd(data$Ever.cigarette.users....)
sd(data$Current.cigarette.users....)
sd(data$Median.age.of.initiation.of.Cigarette..in.years.)
sd(data$Ever.tobacco.smokers.who.quit.in.last.12.months....)
sd(data$Current.tobacco.smokers.who.tried.to.quit.smoking.in.the.past.12.months....)
sd(data$Current.tobacco.smokers.who.wanted.to.quit.smoking.now.....)
```

```
[1] 18.20594
[1] 14.24984
[1] 15.06076
[1] 9.758843
[1] 1.728279
[1] 9.520326
[1] 21.75305
[1] 21.59917
```

Difference between Variance and Standard deviation is about that Variance is find with mean which means the value mean is not compulsory that is available in column but on other hand Standard deviation is calculate form median with is central point of the data and its is one of the observations in out data. And Standard deviation is also square root of Variance.

Column Name	Minimum value	Maximum value
Ever tobacco users (%)	331.4561	18.20594
Current tobacco users (%)	203.058	14.24984
Ever cigarette users (%)	226.8264	15.06076
Current cigarette users (%)	95.23502	9.758843

Median age of initiation of Cigarette (in years)	2.986949	1.728279
Ever tobacco smokers who quit in last 12 months (%)	90.63661	9.520326
Current tobacco smokers who tried to quit smoking in the past 12 months (%)	473.1953	21.75305
Current tobacco smokers who wanted to quit smoking now (%)	466.5242	21.59917

Covariance and correlation for pair of two variable in the dataset.

Correlation It illustrate is the how one column reacts when any value change in other column. This value always lies between -1 to 1 when we have value -1 to 0 it means each column is inversely proportional to each other and if value is between 0 to 1 it means columns are directly proportional to each other. Moreover, if value is negative or positive in between 0.0 to 0.3 it means it is week correlation, if value is between 0.3 to 0.5 it means Correlation is

moderate and if value lies between 0.5 to 1.0 it means it is strong Correlation. In R we use cor() function to find correlation.

Covariance Its just tells us direction of linear relation is it positive or negative. If value is above then 0 it means relation is positive and if value is less then 0 it means relation is negative. In R we use cov() function to find covariance.

Difference: - So it is clear from definition of both of them Correlation tells us about how strong the relation and Covariance tells us the direction or nature (positive or negative) of relation.

```
```{r}
# Define the columns
 col1 <- data$Ever.tobacco.users....
 col2 <- data$Ever.cigarette.users....
 col3 <- data$Current.cigarette.users....
 col4 <- data$Median.age.of.initiation.of.Cigarette..in.years.
 col5 <- data$Ever.tobacco.smokers.who.quit.in.last.12.months....
 col6 <- data$Current.tobacco.smokers.who.tried.to.quit.smoking.in.the.past.12.months.
 col7 <- data$Current.tobacco.smokers.who.wanted.to.quit.smoking.now......
```

Here we give a name to each column.

```
Cor_and_cov<- function(m, k, name_c1, name_c2) {
  # Compute covarince
  cov_value <- cov(m, k)
  print(paste("Covarience between", name_c1, "and", name_c2, ":", cov_value))

  # Compute correlation
  cor_value <- cor(m, k)
  print(paste("Correlation between", name_c1, "and", name_c2, ":", cor_value))
  print("  ")
}
```

Now we create a function to find covariance and correlation in one time for a pair.

```
Cor_and_cov( col1,  col2, "Ever tobacco users", "Ever cigarette users")
Cor_and_cov( col1,  col8, "Ever tobacco users", "Current tobacco users ")
Cor_and_cov( col1,  col3, "Ever tobacco users", "Current cigarette users")
Cor_and_cov( col1,  col4, "Ever tobacco users", "Median age of initiation of Cigarette  in years")
Cor_and_cov( col1,  col5, "Ever tobacco users", "Ever tobacco smokers who quit in last 12 months")
Cor_and_cov( col1,  col6, "Ever tobacco users", "Current tobacco smokers who tried to quit smoking in the past 12 months")
Cor_and_cov( col1,  col7, "Ever tobacco users", "Current tobacco smokers who wanted to quit smoking now")
```

We apply that function on possible pair of variables.
```

```
[1] "Covarience between Ever tobacco users and Ever cigarette users : 253.928513513514"
[1] "Correlation between Ever tobacco users and Ever cigarette users : 0.926086741914152"
[1] "  "
[1] "Covarience between Ever tobacco users and Current tobacco users  : 242.915810810811"
[1] "Correlation between Ever tobacco users and Current tobacco users  : 0.936338084052909"
[1] "  "
[1] "Covarience between Ever tobacco users and Current cigarette users : 153.341081081081"
[1] "Correlation between Ever tobacco users and Current cigarette users : 0.863072310215107"
[1] "  "
[1] "Covarience between Ever tobacco users and Median age of initiation of Cigarette  in years : -3.52364864864865"
[1] "Correlation between Ever tobacco users and Median age of initiation of Cigarette  in years : -0.111986510760148"
[1] "  "
[1] "Covarience between Ever tobacco users and Ever tobacco smokers who quit in last 12 months : 51.9310810810811"
[1] "Correlation between Ever tobacco users and Ever tobacco smokers who quit in last 12 months : 0.29961429508991"
[1] "  "
[1] "Covarience between Ever tobacco users and Current tobacco smokers who tried to quit smoking in the past 12 months : 67.4617567567567"
[1] "Correlation between Ever tobacco users and Current tobacco smokers who tried to quit smoking in the past 12 months : 0.170343060842081"
[1] "  "
[1] "Covarience between Ever tobacco users and Current tobacco smokers who wanted to quit smoking now : 74.3864864864865"
[1] "Correlation between Ever tobacco users and Current tobacco smokers who wanted to quit smoking now : 0.189166393334916"
[1] "  "
```

Here is our results.

Column Name	Covariance	Correlation	
Current tobacco users (%)	242.9158	0.9260	**Ever tobacco users (%)**
Ever cigarette users (%)	253.9285	0.9260	**Ever tobacco users (%)**
Current cigarette users (%)	153.3410	0.8630	**Ever tobacco users (%)**
Median age of initiation of Cigarette (in years)	-3.5236	-0.111	**Ever tobacco users (%)**
Ever tobacco smokers who quit in last 12 months (%)	51.9310	0.2996	**Ever tobacco users (%)**

Current tobacco smokers who tried to quit smoking in the past 12 months (%)	67.4617	0.1703	**Ever tobacco users (%)**
Current tobacco smokers who wanted to quit smoking now (%)	74.38	0.1891	**Ever tobacco users (%)**

When we calculate, we can see that Correlation, Current tobacco users (%), Ever cigarette users (%), Current cigarette users (%) are directly proportional to Ever tobacco users (%) also they have strong correlation. These all are positive relation because we have covariance is in positive side on the numeric scale.

Moving forward, we have Correlation between Ever tobacco users (%) and Median age of initiation of Cigarette (in years) is inversely proportional to each other but they have week correlation. It is a positive relation because we have covariance is in negative side on the numeric scale.

At last, we can see that Correlation, Ever tobacco smokers who quit in last 12 months (%), Current tobacco smokers who tried to quit smoking in the past 12 months (%), Current tobacco smokers who wanted to quit smoking now (%) are directly proportional to Ever

tobacco users (%) but they have week correlation. These all are positive relation because we have covariance is in positive side on the numeric scale.