**DAB 501**

**EDA of data using RStudio**

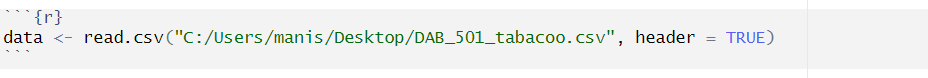
**SUBMITTED TO: -**

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.  
  
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. A screenshot of a computer

Description automatically generated  
  
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 ().   
A screenshot of a computer

Description automatically generated

**Maximum value for each variable in the dataset**Same as minimum here we find maximum value for each variable by using function max().  
A screenshot of a computer

Description automatically generated

|  |  |  |
| --- | --- | --- |
| 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. A screenshot of a computer

Description automatically generated  
**Median value for each variable in the dataset.**  
Median value is the middle value of column. In R studios we use mean() function.  
A screenshot of a computer

Description automatically generated  
  
  
**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.  
A screenshot of a computer program

Description automatically generated  
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 smoking in the past 12 months (%) | 34.23 | 26.5 | 0.1 Count 1 |
| Current tobacco smokers who wanted to quit smoking now (%) | 34.22 | 26..35 | 0.1 Count 1 |

A screenshot of a computer program

Description automatically generated  
  
**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,  
A screenshot of a computer

Description automatically generated

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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.  
A screenshot of a computer

Description automatically generated  
  
**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.  
A screenshot of a computer

Description automatically generated  
**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.  
A computer screen shot of a computer code

Description automatically generated  
Here we give a name to each column.

A computer screen shot of a computer code

Description automatically generated

Now we create a function to find covariance and correlation in one time for a pair.  
A close up of text

Description automatically generated

We apply that function on possible pair of variables.

A close up of text

Description automatically generated

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.