Richa Rambhia
December 12, 2021

ALY 6010 – FINAL PROJECT
EDA, Hypothesis Testing, Regression Model

Introduction

Dataset:

The dataset chosen for the Final Project was the dataset which was picked from Kaggle. The dataset is related to the number of deaths by risk factor and has various risk factors mentioned which would help in determining the cause of death rate in a particular entity which is nothing but in a particular country. The dataset contains 6468 rows of data and 32 data fields, and the data fields are entity, code, year, alcohol use, drug use, air pollution, no access to handwashing facility, smoking, outdoor air pollution, iron deficiency, etc. This dataset contains numeric as well as categorical data and is helpful in determining which risk factor in a particular country would affect the death rate. The reason for choosing this dataset is because while analyzing the dataset it can be found that which risk factor would highly affect the death rate and what is the death rate due to these factors in each country.

Dataset Source:

https://www.kaggle.com/pavan9065/air-pollution?select=number-of-deaths-by-risk-factor.csv

Exploratory Data Analysis:

In this section, exploratory data analysis would be performed where the dataset is analyzed statistically and visually to have a better understanding of the data for further analysis. Descriptive analysis and data exploration are the phases that are performed in the exploratory data analysis where we try to understand the dataset and find out the outliers which would be done by data cleaning.

Hypothesis Testing:

After the exploratory data analysis is performed, hypothesis testing comes to picture which helps us in answering various questions about the dataset and this would be done through inferential statistics. This would give us more insights about the dataset and help in determining the answers to the questions of the data.

Regression Model:

Once the hypothesis testing is completed, we built a regression model to identify the relationship between the variables of the dataset. Regression model is a statistical model which helps in determining the relationship between the dependent and independent variables and the fitted line to plot the model is a regression plot line. Thus, we will build a regression model and understand the coefficients to examine the relationship between the variables of our dataset.

Exploratory Data Analysis

Exploratory Data Analysis consists of various phases which include data collection, data cleaning, data exploration or data visualization and data analysis. Once the dataset was read into a variable named deathrate_dataset, describing the dataset was done to understand what the dataset looks like after which there were new variables and subset of dataset created which would be used in the further phases of the data analysis part.

Descriptive analysis is the statistical analysis of the dataset done to understand the statistical values of the attributes of the dataset on which the analysis needs to be performed. It basically means to compute the statistical values of the numeric variables of the dataset. The data description part included functions like displaying the column names, the starting and ending records of the dataset, the dimensions of the dataset, summary and structure of the dataset, and the type of class of each attribute in the dataset whereas the descriptive analysis included computing the statistical values like minimum, maximum, mean, median, mode, range, standard deviation, and summary of each of the risk factor attribute. The summary function would return all the statistical values of a particular attribute of the dataset.

Data cleaning is another important phase of the data analysis process which helps in dealing with the incorrect data and the missing data values. Here, data cleaning step was performed to check for the null values and remove them. After data cleaning, data visualization was performed which help in the data analysis of the dataset visually. Subsets of dataset was created in order to perform the visualizations on the subset as the actual dataset contained huge volume of data where it would become difficult in understanding the data values.

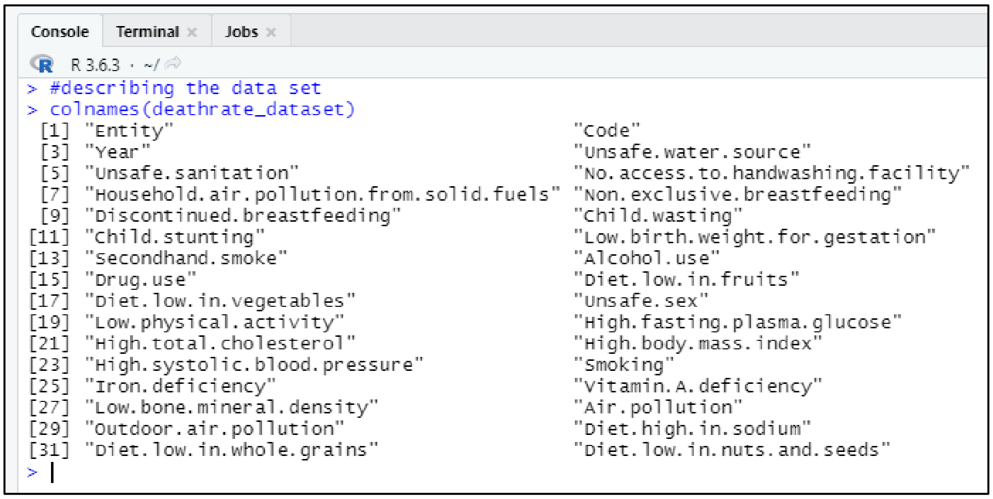
Data visualization plays an important role in the exploratory data analysis as majorly we understand the data based on the visual analysis and make our conclusions on the same. In the EDA, first a bar plot was created for the death rate due to smoking in the country. Similarly, a scatterplot was created for the death rate due to alcohol use in the country and coordinate polar for the death rate due to no access to hand wash facility in the country. Next visualization created was a boxplot for the risk factor due to air pollution and then using a line graph, visualization comparing the two risk factors in a particular country was plotted. These visualizations thus helped in answering some of the basic questions and understanding which risk factor would lead to an increase in death rate in the country or which country is the most affected by the risk factor.

The analysis of the visualizations in the entity Afghanistan gave us insights that the risk factor of drug use increased during the years whereas the alcohol use had a varying graph which indicated a decrease in the year from 2000 to 2005 but had an increase in the year 2015. Similar analysis was done for the other countries as well.

Results

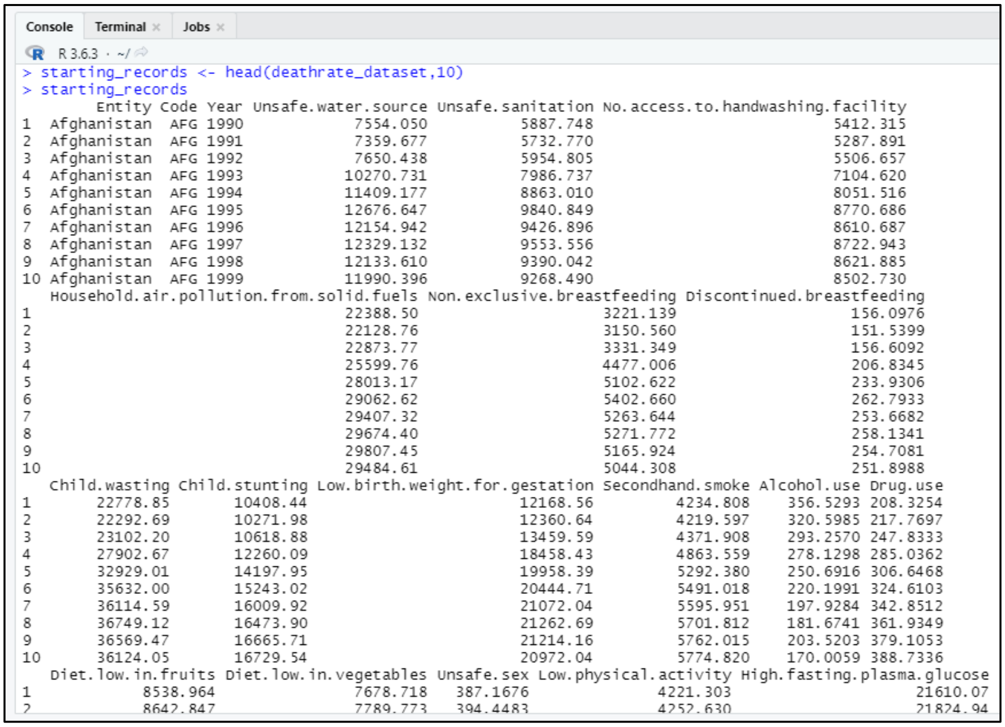
1. Describing the dataset:

a. Column names



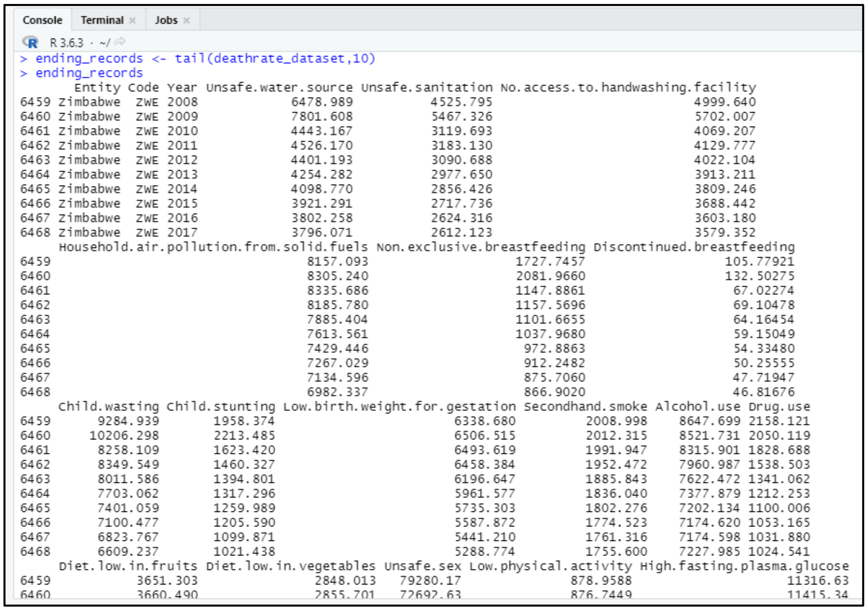
```
> #describing the data set
> colnames(deathrate_dataset)
 [1] "Entity"                                   "Code"
 [3] "Year"                                     "Unsafe.water.source"
 [5] "Unsafe.sanitation"                        "No.access.to.handwashing.facility"
 [7] "Household.air.pollution.from.solid.fuels" "Non.exclusive.breastfeeding"
 [9] "Discontinued.breastfeeding"               "Child.wasting"
[11] "Child.stunting"                           "Low.birth.weight.for.gestation"
[13] "Secondhand.smoke"                         "Alcohol.use"
[15] "Drug.use"                                 "Diet.low.in.fruits"
[17] "Diet.low.in.vegetables"                   "Unsafe.sex"
[19] "Low.physical.activity"                    "High.fasting.plasma.glucose"
[21] "High.total.cholesterol"                   "High.body.mass.index"
[23] "High.systolic.blood.pressure"             "Smoking"
[25] "Iron.deficiency"                          "Vitamin.A.deficiency"
[27] "Low.bone.mineral.density"                 "Air.pollution"
[29] "Outdoor.air.pollution"                    "Diet.high.in.sodium"
[31] "Diet.low.in.whole.grains"                 "Diet.low.in.nuts.and.seeds"
> |
```

b. Start records



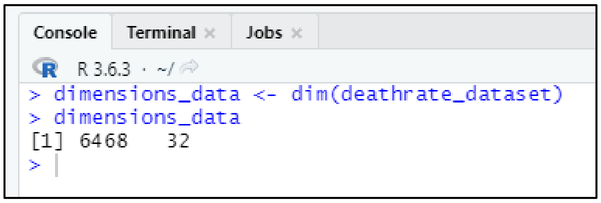
```
> starting_records <- head(deathrate_dataset,10)
> starting_records
        Entity Code Year Unsafe.water.source Unsafe.sanitation No.access.to.handwashing.facility
1  Afghanistan  AFG 1990            7554.050          5887.748                         5412.315
2  Afghanistan  AFG 1991            7359.677          5732.770                         5287.891
3  Afghanistan  AFG 1992            7650.438          5954.805                         5506.657
4  Afghanistan  AFG 1993           10270.731          7986.737                         7104.620
5  Afghanistan  AFG 1994           11409.177          8863.010                         8051.516
6  Afghanistan  AFG 1995           12676.647          9840.849                         8770.686
7  Afghanistan  AFG 1996           12154.942          9426.896                         8610.687
8  Afghanistan  AFG 1997           12329.132          9553.556                         8722.943
9  Afghanistan  AFG 1998           12133.610          9390.042                         8621.885
10 Afghanistan  AFG 1999           11990.396          9268.490                         8502.730
   Household.air.pollution.from.solid.fuels Non.exclusive.breastfeeding Discontinued.breastfeeding
1                                  22388.50                    3221.139                   156.0976
2                                  22128.76                    3150.560                   151.5399
3                                  22873.77                    3331.349                   156.6092
4                                  25599.76                    4477.006                   206.8345
5                                  28013.17                    5102.622                   233.9306
6                                  29062.62                    5402.660                   262.7933
7                                  29407.32                    5263.644                   253.6682
8                                  29674.40                    5271.772                   258.1341
9                                  29807.45                    5165.924                   254.7081
10                                 29484.61                    5044.308                   251.8988
   Child.wasting Child.stunting Low.birth.weight.for.gestation Secondhand.smoke Alcohol.use Drug.use
1       22778.85       10408.44                       12168.56         4234.808    356.5293 208.3254
2       22292.69       10271.98                       12360.64         4219.597    320.5985 217.7697
3       23102.20       10618.88                       13459.59         4371.908    293.2570 247.8333
4       27902.67       12260.09                       18458.43         4863.559    278.1298 285.0362
5       32929.01       14197.95                       19958.39         5292.380    250.6916 306.6468
6       35632.00       15243.02                       20444.71         5491.018    220.1991 324.6103
7       36114.59       16009.92                       21072.04         5595.951    197.9284 342.8512
8       36749.12       16473.90                       21262.69         5701.812    181.6741 361.9349
9       36569.47       16665.71                       21214.16         5762.015    203.5203 379.1053
10      36124.05       16729.54                       20972.04         5774.820    170.0059 388.7336
   Diet.low.in.fruits Diet.low.in.vegetables Unsafe.sex Low.physical.activity High.fasting.plasma.glucose
1            8538.964               7678.718   387.1676              4221.303                    21610.07
2            8642.847               7789.773   394.4483              4252.630                    21824.94
```

c. End records



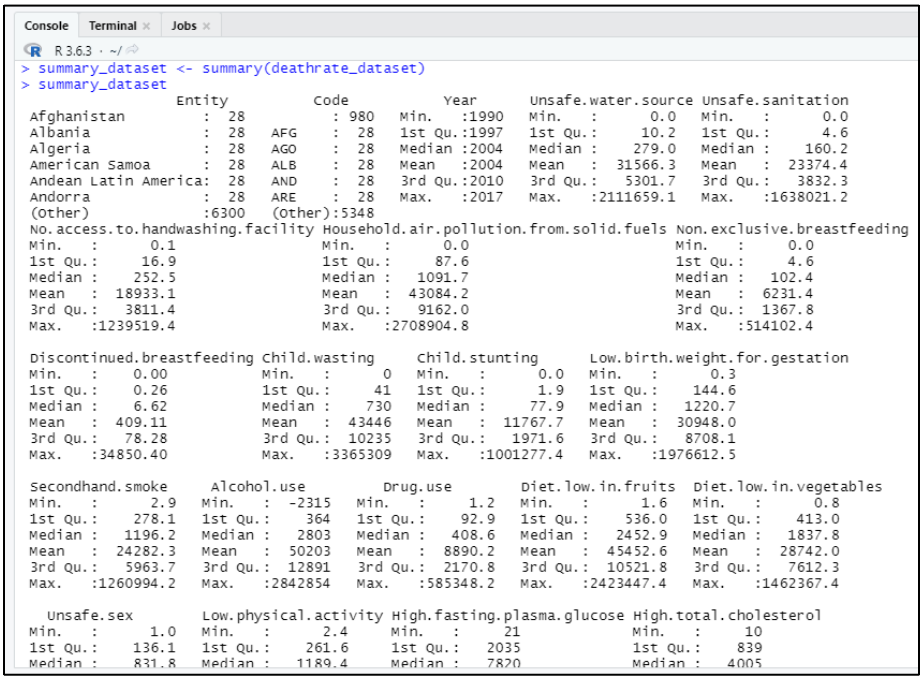
```
> ending_records <- tail(deathrate_dataset,10)
> ending_records
       Entity Code Year Unsafe.water.source Unsafe.sanitation No.access.to.handwashing.facility
6459 Zimbabwe  ZWE 2008            6478.989          4525.795                          4999.640
6460 Zimbabwe  ZWE 2009            7801.608          5467.326                          5702.007
6461 Zimbabwe  ZWE 2010            4443.167          3119.693                          4069.207
6462 Zimbabwe  ZWE 2011            4526.170          3183.130                          4129.777
6463 Zimbabwe  ZWE 2012            4401.193          3090.688                          4022.104
6464 Zimbabwe  ZWE 2013            4254.282          2977.650                          3913.211
6465 Zimbabwe  ZWE 2014            4098.770          2856.426                          3809.246
6466 Zimbabwe  ZWE 2015            3921.291          2717.736                          3688.442
6467 Zimbabwe  ZWE 2016            3802.258          2624.316                          3603.180
6468 Zimbabwe  ZWE 2017            3796.071          2612.123                          3579.352
     Household.air.pollution.from.solid.fuels Non.exclusive.breastfeeding Discontinued.breastfeeding
6459                                 8157.093                   1727.7457                  105.77921
6460                                 8305.240                   2081.9660                  132.50275
6461                                 8335.686                   1147.8861                   67.02274
6462                                 8185.780                   1157.5696                   69.10478
6463                                 7885.404                   1101.6655                   64.16454
6464                                 7613.561                   1037.9680                   59.15049
6465                                 7429.446                    972.8863                   54.33480
6466                                 7267.029                    912.2482                   50.25555
6467                                 7134.596                    875.7060                   47.71947
6468                                 6982.337                    866.9020                   46.81676
     Child.wasting Child.stunting Low.birth.weight.for.gestation Secondhand.smoke Alcohol.use Drug.use
6459      9284.939       1958.374                       6338.680         2008.998    8647.699 2158.121
6460     10206.298       2213.485                       6506.515         2012.315    8521.731 2050.119
6461      8258.109       1623.420                       6493.619         1991.947    8315.901 1828.688
6462      8349.549       1460.327                       6458.384         1952.472    7960.987 1538.503
6463      8011.586       1394.801                       6196.647         1885.843    7622.472 1341.062
6464      7703.062       1317.296                       5961.577         1836.040    7377.879 1212.253
6465      7401.059       1259.989                       5735.303         1802.276    7202.134 1100.006
6466      7100.477       1205.590                       5587.872         1774.523    7174.620 1053.165
6467      6823.767       1099.871                       5441.210         1761.316    7174.598 1031.880
6468      6609.237       1021.438                       5288.774         1755.600    7227.985 1024.541
     Diet.low.in.fruits Diet.low.in.vegetables Unsafe.sex Low.physical.activity High.fasting.plasma.glucose
6459           3651.303               2848.013   79280.17              878.9588                    11316.63
6460           3660.490               2855.701   72692.63              876.7449                    11415.34
```

d. Dimensions of the dataset



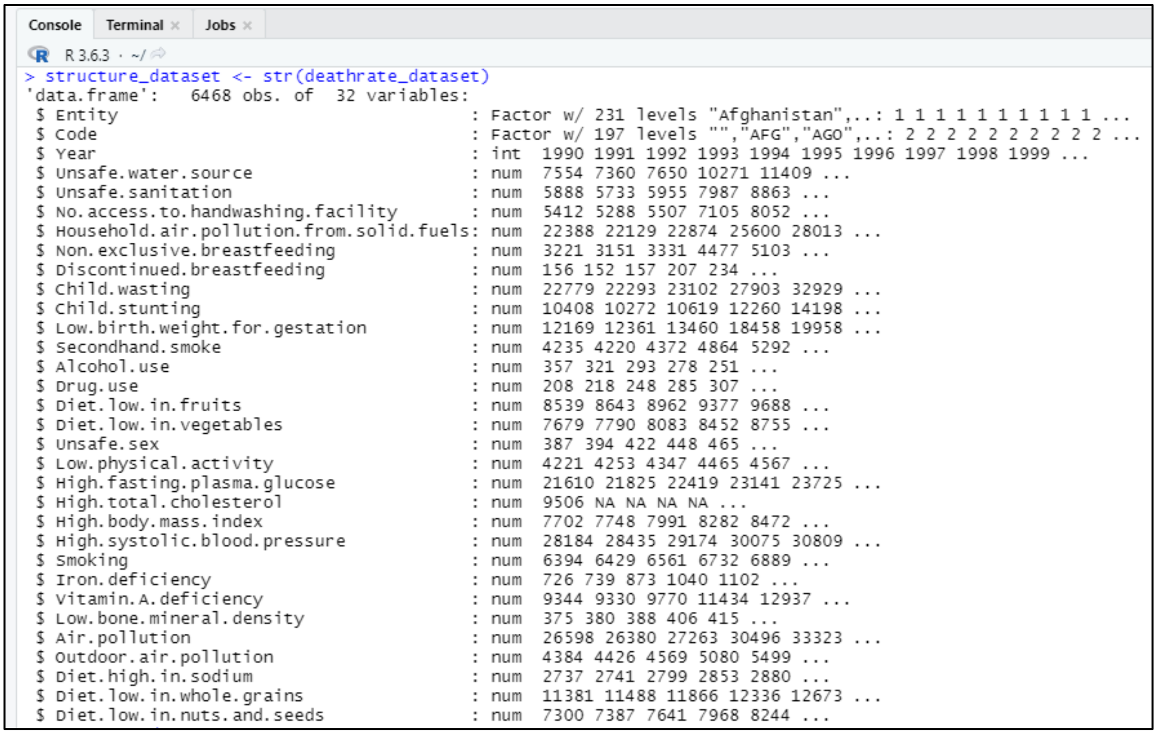
```
> dimensions_data <- dim(deathrate_dataset)
> dimensions_data
[1] 6468   32
> 
```

e. Summary of the dataset



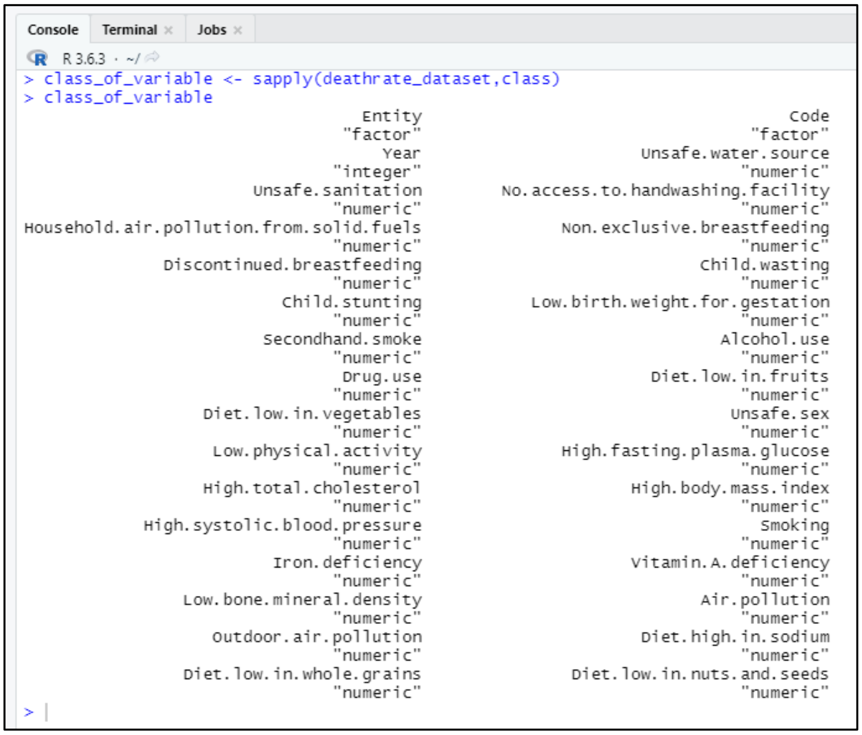
```
> summary_dataset <- summary(deathrate_dataset)
> summary_dataset
                  Entity           Code           Year       Unsafe.water.source Unsafe.sanitation
 Afghanistan          :  28            : 980   Min.   :1990   Min.   :      0.0   Min.   :      0.0  
 Albania              :  28   AFG      :  28   1st Qu.:1997   1st Qu.:     10.2   1st Qu.:      4.6  
 Algeria              :  28   AGO      :  28   Median :2004   Median :    279.0   Median :    160.2  
 American Samoa       :  28   ALB      :  28   Mean   :2004   Mean   :  31566.3   Mean   :  23374.4  
 Andean Latin America:  28   AND      :  28   3rd Qu.:2010   3rd Qu.:   5301.7   3rd Qu.:   3832.3  
 Andorra              :  28   ARE      :  28   Max.   :2017   Max.   :2111659.1   Max.   :1638021.2  
 (Other)              :6300   (Other):5348                                                             
 No.access.to.handwashing.facility Household.air.pollution.from.solid.fuels Non.exclusive.breastfeeding
 Min.   :      0.1                 Min.   :      0.0                        Min.   :     0.0           
 1st Qu.:     16.9                 1st Qu.:     87.6                        1st Qu.:     4.6           
 Median :    252.5                 Median :   1091.7                        Median :   102.4           
 Mean   :  18933.1                 Mean   :  43084.2                        Mean   :  6231.4           
 3rd Qu.:   3811.4                 3rd Qu.:   9162.0                        3rd Qu.:  1367.8           
 Max.   :1239519.4                 Max.   :2708904.8                        Max.   :514102.4           
                                                                                                        
 Discontinued.breastfeeding Child.wasting     Child.stunting      Low.birth.weight.for.gestation
 Min.   :    0.00           Min.   :      0   Min.   :      0.0   Min.   :      0.3              
 1st Qu.:    0.26           1st Qu.:     41   1st Qu.:      1.9   1st Qu.:    144.6              
 Median :    6.62           Median :    730   Median :     77.9   Median :   1220.7              
 Mean   :  409.11           Mean   :  43446   Mean   :  11767.7   Mean   :  30948.0              
 3rd Qu.:   78.28           3rd Qu.:  10235   3rd Qu.:   1971.6   3rd Qu.:   8708.1              
 Max.   :34850.40           Max.   :3365309   Max.   :1001277.4   Max.   :1976612.5              
                                                                                                  
 Secondhand.smoke    Alcohol.use        Drug.use         Diet.low.in.fruits Diet.low.in.vegetables
 Min.   :      2.9   Min.   :  -2315   Min.   :     1.2   Min.   :      1.6   Min.   :      0.8     
 1st Qu.:    278.1   1st Qu.:    364   1st Qu.:    92.9   1st Qu.:    536.0   1st Qu.:    413.0     
 Median :   1196.2   Median :   2803   Median :   408.6   Median :   2452.9   Median :   1837.8     
 Mean   :  24282.3   Mean   :  50203   Mean   :  8890.2   Mean   :  45452.6   Mean   :  28742.0     
 3rd Qu.:   5963.7   3rd Qu.:  12891   3rd Qu.:  2170.8   3rd Qu.:  10521.8   3rd Qu.:   7612.3     
 Max.   :1260994.2   Max.   :2842854   Max.   :585348.2   Max.   :2423447.4   Max.   :1462367.4     
                                                                                                     
   Unsafe.sex         Low.physical.activity High.fasting.plasma.glucose High.total.cholesterol
 Min.   :      1.0   Min.   :      2.4     Min.   :     21               Min.   :     10      
 1st Qu.:    136.1   1st Qu.:    261.6     1st Qu.:   2035               1st Qu.:    839      
 Median :    831.8   Median :   1189.4     Median :   7820               Median :   4005      
```

f. Structure of the dataset



```
> structure_dataset <- str(deathrate_dataset)
'data.frame':	6468 obs. of  32 variables:
 $ Entity                                  : Factor w/ 231 levels "Afghanistan",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Code                                    : Factor w/ 197 levels "","AFG","AGO",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ Year                                    : int  1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 ...
 $ Unsafe.water.source                     : num  7554 7360 7650 10271 11409 ...
 $ Unsafe.sanitation                       : num  5888 5733 5955 7987 8863 ...
 $ No.access.to.handwashing.facility       : num  5412 5288 5507 7105 8052 ...
 $ Household.air.pollution.from.solid.fuels: num  22388 22129 22874 25600 28013 ...
 $ Non.exclusive.breastfeeding             : num  3221 3151 3331 4477 5103 ...
 $ Discontinued.breastfeeding              : num  156 152 157 207 234 ...
 $ Child.wasting                           : num  22779 22293 23102 27903 32929 ...
 $ Child.stunting                          : num  10408 10272 10619 12260 14198 ...
 $ Low.birth.weight.for.gestation          : num  12169 12361 13460 18458 19958 ...
 $ Secondhand.smoke                        : num  4235 4220 4372 4864 5292 ...
 $ Alcohol.use                             : num  357 321 293 278 251 ...
 $ Drug.use                                : num  208 218 248 285 307 ...
 $ Diet.low.in.fruits                      : num  8539 8643 8962 9377 9688 ...
 $ Diet.low.in.vegetables                  : num  7679 7790 8083 8452 8755 ...
 $ Unsafe.sex                              : num  387 394 422 448 465 ...
 $ Low.physical.activity                   : num  4221 4253 4347 4465 4567 ...
 $ High.fasting.plasma.glucose             : num  21610 21825 22419 23141 23725 ...
 $ High.total.cholesterol                  : num  9506 NA NA NA NA ...
 $ High.body.mass.index                    : num  7702 7748 7991 8282 8472 ...
 $ High.systolic.blood.pressure            : num  28184 28435 29174 30075 30809 ...
 $ Smoking                                 : num  6394 6429 6561 6732 6889 ...
 $ Iron.deficiency                         : num  726 739 873 1040 1102 ...
 $ Vitamin.A.deficiency                    : num  9344 9330 9770 11434 12937 ...
 $ Low.bone.mineral.density                : num  375 380 388 406 415 ...
 $ Air.pollution                           : num  26598 26380 27263 30496 33323 ...
 $ Outdoor.air.pollution                   : num  4384 4426 4569 5080 5499 ...
 $ Diet.high.in.sodium                     : num  2737 2741 2799 2853 2880 ...
 $ Diet.low.in.whole.grains                : num  11381 11488 11866 12336 12673 ...
 $ Diet.low.in.nuts.and.seeds              : num  7300 7387 7641 7968 8244 ...
```

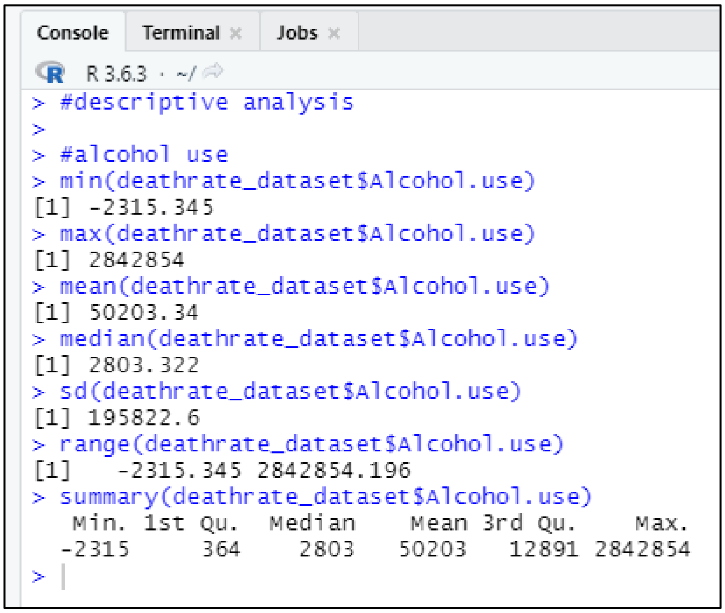
g. Class type of the attributes



```
> class_of_variable <- sapply(deathrate_dataset,class)
> class_of_variable
                                  Entity                                      Code 
                                "factor"                                  "factor" 
                                    Year                       Unsafe.water.source 
                               "integer"                                 "numeric" 
                       Unsafe.sanitation         No.access.to.handwashing.facility 
                               "numeric"                                 "numeric" 
Household.air.pollution.from.solid.fuels               Non.exclusive.breastfeeding 
                               "numeric"                                 "numeric" 
              Discontinued.breastfeeding                             Child.wasting 
                               "numeric"                                 "numeric" 
                          Child.stunting            Low.birth.weight.for.gestation 
                               "numeric"                                 "numeric" 
                        Secondhand.smoke                               Alcohol.use 
                               "numeric"                                 "numeric" 
                                Drug.use                        Diet.low.in.fruits 
                               "numeric"                                 "numeric" 
                  Diet.low.in.vegetables                                Unsafe.sex 
                               "numeric"                                 "numeric" 
                   Low.physical.activity               High.fasting.plasma.glucose 
                               "numeric"                                 "numeric" 
                  High.total.cholesterol                      High.body.mass.index 
                               "numeric"                                 "numeric" 
            High.systolic.blood.pressure                                   Smoking 
                               "numeric"                                 "numeric" 
                         Iron.deficiency                      Vitamin.A.deficiency 
                               "numeric"                                 "numeric" 
                Low.bone.mineral.density                             Air.pollution 
                               "numeric"                                 "numeric" 
                   Outdoor.air.pollution                       Diet.high.in.sodium 
                               "numeric"                                 "numeric" 
                Diet.low.in.whole.grains                Diet.low.in.nuts.and.seeds 
                               "numeric"                                 "numeric" 
> |
```

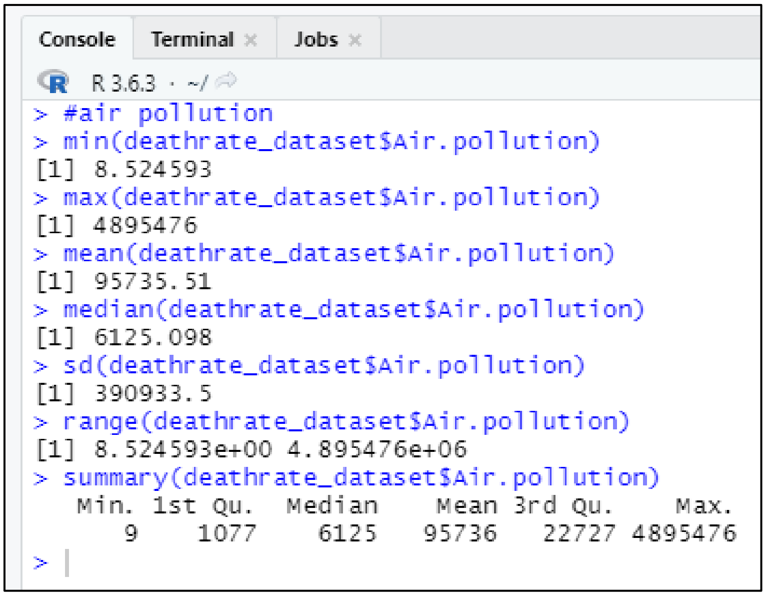
2. Descriptive Analysis:

a. Risk factor – Alcohol Use



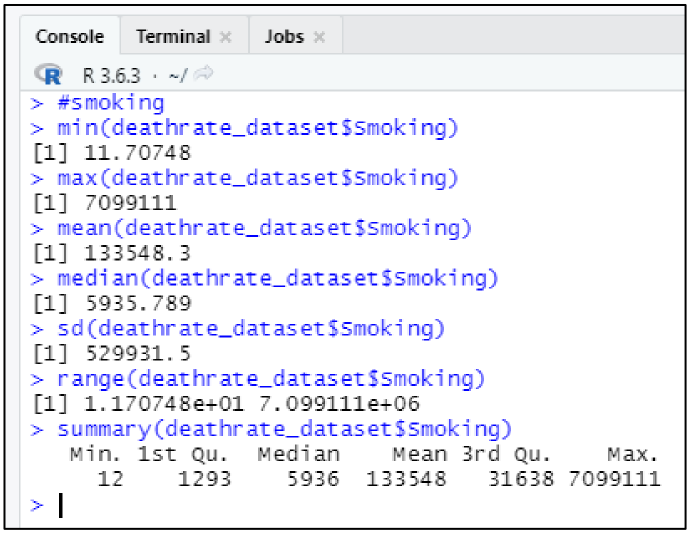
```
> #descriptive analysis
> 
> #alcohol use
> min(deathrate_dataset$Alcohol.use)
[1] -2315.345
> max(deathrate_dataset$Alcohol.use)
[1] 2842854
> mean(deathrate_dataset$Alcohol.use)
[1] 50203.34
> median(deathrate_dataset$Alcohol.use)
[1] 2803.322
> sd(deathrate_dataset$Alcohol.use)
[1] 195822.6
> range(deathrate_dataset$Alcohol.use)
[1]   -2315.345 2842854.196
> summary(deathrate_dataset$Alcohol.use)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  -2315     364    2803   50203   12891 2842854 
> 
```

b. Risk factor – Air Pollution



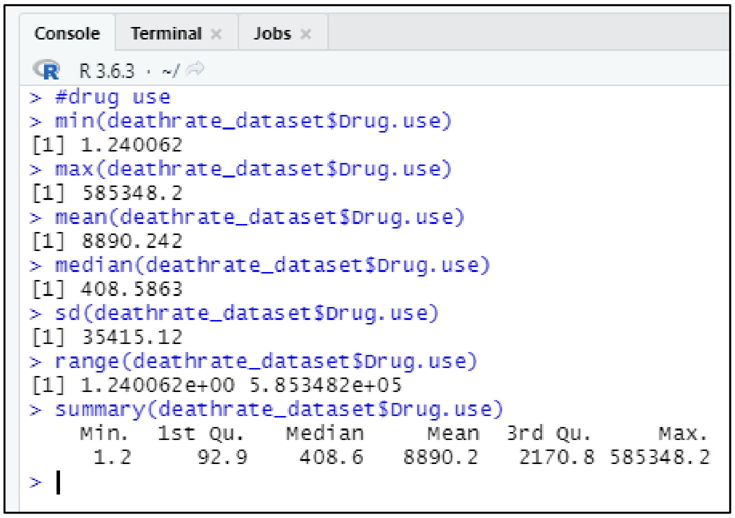
```
> #air pollution
> min(deathrate_dataset$Air.pollution)
[1] 8.524593
> max(deathrate_dataset$Air.pollution)
[1] 4895476
> mean(deathrate_dataset$Air.pollution)
[1] 95735.51
> median(deathrate_dataset$Air.pollution)
[1] 6125.098
> sd(deathrate_dataset$Air.pollution)
[1] 390933.5
> range(deathrate_dataset$Air.pollution)
[1] 8.524593e+00 4.895476e+06
> summary(deathrate_dataset$Air.pollution)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
      9    1077    6125   95736   22727 4895476 
> 
```

c. Risk factor – Smoking



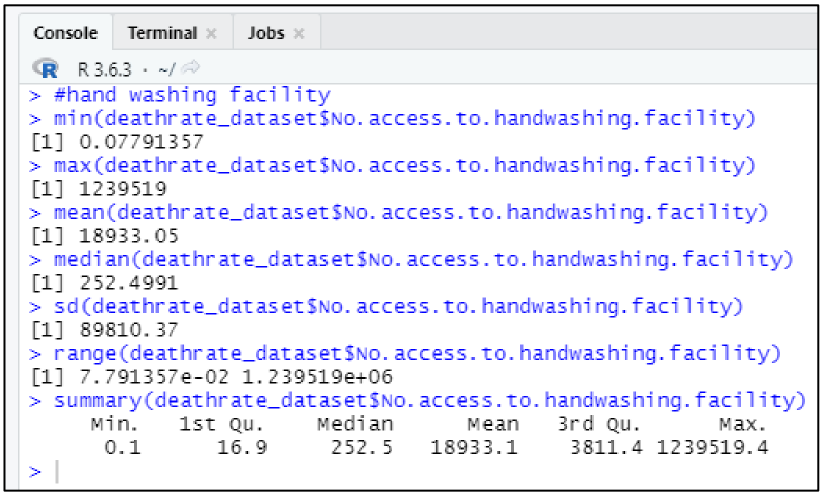
```
> #smoking
> min(deathrate_dataset$Smoking)
[1] 11.70748
> max(deathrate_dataset$Smoking)
[1] 7099111
> mean(deathrate_dataset$Smoking)
[1] 133548.3
> median(deathrate_dataset$Smoking)
[1] 5935.789
> sd(deathrate_dataset$Smoking)
[1] 529931.5
> range(deathrate_dataset$Smoking)
[1] 1.170748e+01 7.099111e+06
> summary(deathrate_dataset$Smoking)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
     12    1293    5936  133548   31638 7099111 
> 
```

d. Risk factor – Drug Use



```
> #drug use
> min(deathrate_dataset$Drug.use)
[1] 1.240062
> max(deathrate_dataset$Drug.use)
[1] 585348.2
> mean(deathrate_dataset$Drug.use)
[1] 8890.242
> median(deathrate_dataset$Drug.use)
[1] 408.5863
> sd(deathrate_dataset$Drug.use)
[1] 35415.12
> range(deathrate_dataset$Drug.use)
[1] 1.240062e+00 5.853482e+05
> summary(deathrate_dataset$Drug.use)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
     1.2     92.9    408.6   8890.2   2170.8 585348.2 
> 
```

e. Risk factor – No access to handwash facility



```
> #hand washing facility
> min(deathrate_dataset$No.access.to.handwashing.facility)
[1] 0.07791357
> max(deathrate_dataset$No.access.to.handwashing.facility)
[1] 1239519
> mean(deathrate_dataset$No.access.to.handwashing.facility)
[1] 18933.05
> median(deathrate_dataset$No.access.to.handwashing.facility)
[1] 252.4991
> sd(deathrate_dataset$No.access.to.handwashing.facility)
[1] 89810.37
> range(deathrate_dataset$No.access.to.handwashing.facility)
[1] 7.791357e-02 1.239519e+06
> summary(deathrate_dataset$No.access.to.handwashing.facility)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
      0.1      16.9     252.5   18933.1    3811.4 1239519.4 
> 
```

Hypothesis Testing

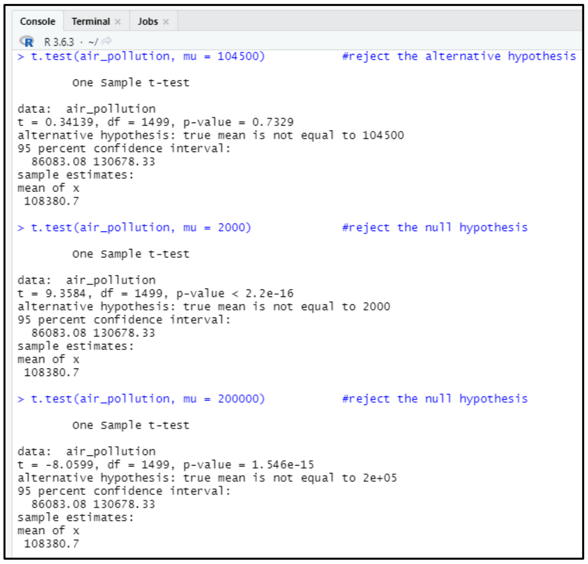
After the exploratory data analysis was performed, it was now time to proceed with the hypothesis testing to answer the questions related to the dataset based on inferential statistics. Hypothesis testing is used to test an assumption regarding a population mean which helps to determine whether a specific treatment has an effect on the individuals in a population. T-test is a statistical test which is used to compare the means of two groups used in hypothesis testing to determine whether a process actually has an effect on the population of interest or whether two groups are different from one another. Therefore, by using the t-test we try to find the answers to our questions for the number of deaths by risk factors dataset.

One sample and two sample tests were performed based on the requirements of the dataset and these concepts helped in understanding the analysis better. For the one sample test, the mean values were kept varying to check and understand the hypothesis. It was observed that for each of the attribute tested, the mu value was assumed and depending on the p-value we rejected the null hypothesis or not. For two-sample t-test, we would check whether the means of two populations is equal or not. Here, the two attributes were compared to check which risk factor is higher as compared to the other for the overall dataset. One another analysis that can be done with respect to the two-sample t-test is which country is most affected by a particular risk factor as compared to the other.

Results

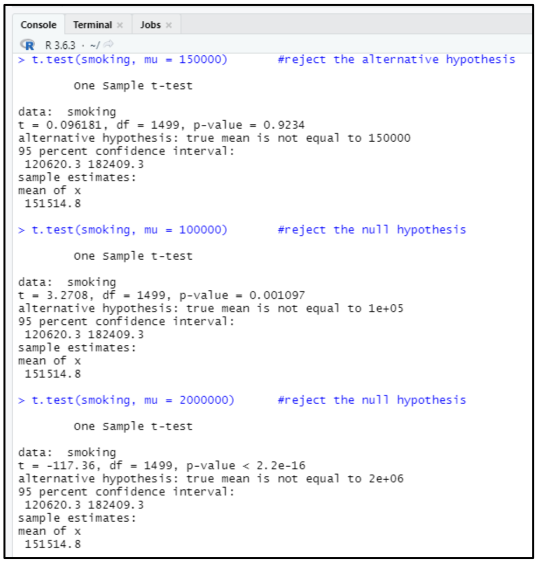
1. <u>One sample t-test:</u>

   a. Risk Factor - Air Pollution



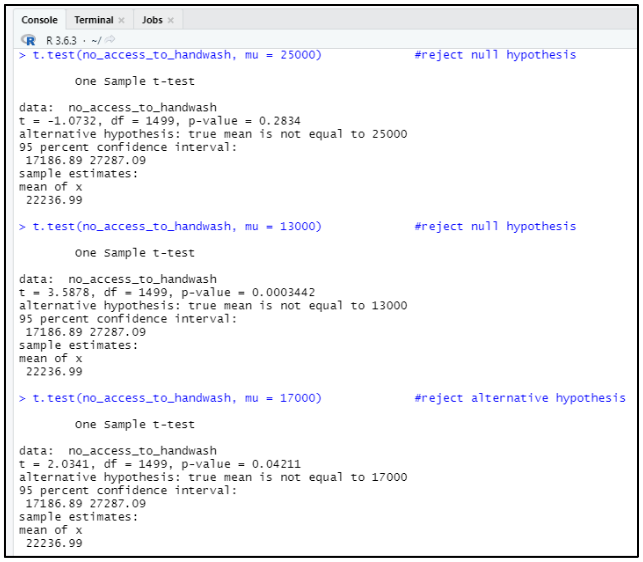
```
> t.test(air_pollution, mu = 104500)          #reject the alternative hypothesis

        One Sample t-test

data:  air_pollution
t = 0.34139, df = 1499, p-value = 0.7329
alternative hypothesis: true mean is not equal to 104500
95 percent confidence interval:
  86083.08 130678.33
sample estimates:
mean of x
 108380.7

> t.test(air_pollution, mu = 2000)            #reject the null hypothesis

        One Sample t-test

data:  air_pollution
t = 9.3584, df = 1499, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 2000
95 percent confidence interval:
  86083.08 130678.33
sample estimates:
mean of x
 108380.7

> t.test(air_pollution, mu = 200000)          #reject the null hypothesis

        One Sample t-test

data:  air_pollution
t = -8.0599, df = 1499, p-value = 1.546e-15
alternative hypothesis: true mean is not equal to 2e+05
95 percent confidence interval:
  86083.08 130678.33
sample estimates:
mean of x
 108380.7
```

   b. Risk Factor - Smoking



```
> t.test(smoking, mu = 150000)       #reject the alternative hypothesis

        One Sample t-test

data:  smoking
t = 0.096181, df = 1499, p-value = 0.9234
alternative hypothesis: true mean is not equal to 150000
95 percent confidence interval:
 120620.3 182409.3
sample estimates:
mean of x
 151514.8

> t.test(smoking, mu = 100000)       #reject the null hypothesis

        One Sample t-test

data:  smoking
t = 3.2708, df = 1499, p-value = 0.001097
alternative hypothesis: true mean is not equal to 1e+05
95 percent confidence interval:
 120620.3 182409.3
sample estimates:
mean of x
 151514.8

> t.test(smoking, mu = 2000000)      #reject the null hypothesis

        One Sample t-test

data:  smoking
t = -117.36, df = 1499, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 2e+06
95 percent confidence interval:
 120620.3 182409.3
sample estimates:
mean of x
 151514.8
```

c. Risk Factor - No access to handwash facility



```
> t.test(no_access_to_handwash, mu = 25000)            #reject null hypothesis

        One Sample t-test

data:  no_access_to_handwash
t = -1.0732, df = 1499, p-value = 0.2834
alternative hypothesis: true mean is not equal to 25000
95 percent confidence interval:
 17186.89 27287.09
sample estimates:
mean of x
 22236.99

> t.test(no_access_to_handwash, mu = 13000)            #reject null hypothesis

        One Sample t-test

data:  no_access_to_handwash
t = 3.5878, df = 1499, p-value = 0.0003442
alternative hypothesis: true mean is not equal to 13000
95 percent confidence interval:
 17186.89 27287.09
sample estimates:
mean of x
 22236.99

> t.test(no_access_to_handwash, mu = 17000)            #reject alternative hypothesis

        One Sample t-test

data:  no_access_to_handwash
t = 2.0341, df = 1499, p-value = 0.04211
alternative hypothesis: true mean is not equal to 17000
95 percent confidence interval:
 17186.89 27287.09
sample estimates:
mean of x
 22236.99
```

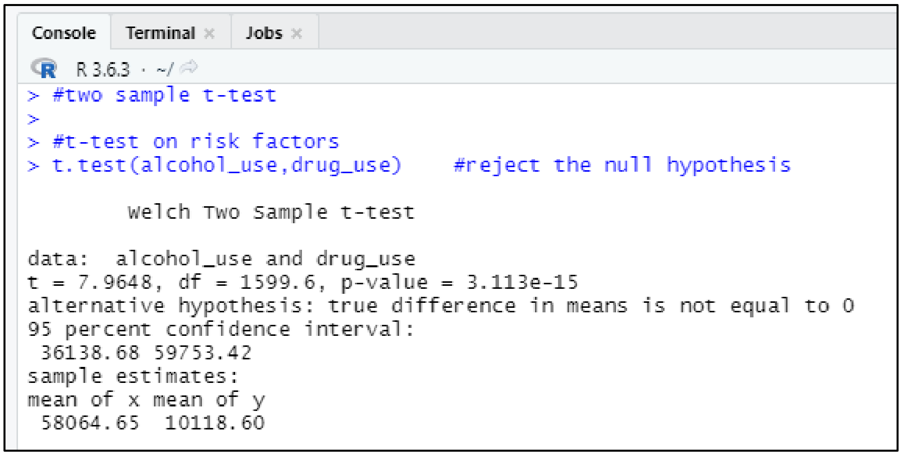
From the outputs of the one-sample t-test, it is observed that for each of the risk factor, depending on the p-value we either reject the null hypothesis or fail to reject the null hypothesis which means that if we fail to reject the null hypothesis we conclude that the data is supporting the assumption whereas if we reject the null hypothesis we conclude that the two populations do not have the same value of parameter.

With respect to all the risk factors, depending on our mu value we determine the p-value which in turn gives us the hypothesis of rejecting the null hypothesis or not. If the p-value is greater than 0.05 which is the level of significance, we reject the alternative hypothesis and if the p-value is less than 0.05, we reject the null hypothesis. In this case, when the mu value was approx. closer to the mean value, the p-value was greater, and we rejected the alternative hypothesis and so the assumed value was same as the population mean. Also, we can conclude that changing the level of significance value from 0.05 to 0.1 does not affect the hypothesis and the conclusion remains the same. The values and analysis of each population tested can be clearly seen in the outputs displayed above.

Therefore, our first question is answered here which is to test whether the mean of population is equal to the assumed value or not.

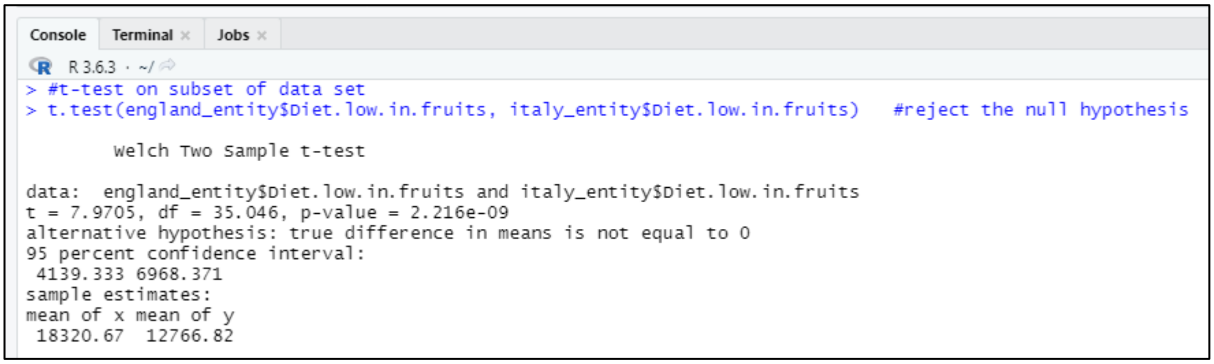
2. Two sample t-test:

   a. t-test on Risk Factors



```
> #two sample t-test
>
> #t-test on risk factors
> t.test(alcohol_use,drug_use)    #reject the null hypothesis

        Welch Two Sample t-test

data:  alcohol_use and drug_use
t = 7.9648, df = 1599.6, p-value = 3.113e-15
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 36138.68 59753.42
sample estimates:
mean of x mean of y 
 58064.65  10118.60 
```

   b. t-test on subset of data



```
> #t-test on subset of data set
> t.test(england_entity$Diet.low.in.fruits, italy_entity$Diet.low.in.fruits)   #reject the null hypothesis

        Welch Two Sample t-test

data:  england_entity$Diet.low.in.fruits and italy_entity$Diet.low.in.fruits
t = 7.9705, df = 35.046, p-value = 2.216e-09
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 4139.333 6968.371
sample estimates:
mean of x mean of y 
 18320.67  12766.82 
```

From the first output, which was based on comparing the two risk factors for the overall entities, it was observed that since the p-value is less than 0.05, we reject the null hypothesis which states that the alcohol use in different entities is high than that of the drug use and that the risk factor affecting the death rate is more due to the alcohol use as compared to the drug use in different countries. The level of significance considered here was 0.05 but changing the value to 0.1 also does not change our conclusion.

The second output based on comparing a risk factor for the two entities tells us that since the p-value is less than 0.05, we reject the null hypothesis which implies that diet low in fruits, a risk factor leading to the death rate is more in the entity England as compared to the entity Italy. And so, England is affected more due to this risk factor as compared to Italy leading to the death rate factor.

Also, it can be noted that a paired t-test cannot be performed here in this case since we reject the null hypothesis in the two-sample t-test.

Regression Model

Regression model is a statistical model which estimates the relationship between one dependent variable and one or more independent variables which uses a line to examine the relationship called as the regression line. The regression model is therefore utilized to assess the strength of the relationship between variables and for modeling the future relationship between them. In the final project, a simple linear regression model is built to examine the relationship between a dependent variable and an independent variable and the equation for the same is, y = c + mx, where y is the dependent variable, x is the independent variable, c is the intercept and m is the slope.

Correlation coefficient is a statistical relationship between the two variables which is a numerical measure of some type of correlation that is a number between -1 and 1 telling you the strength and direction of a relationship between the variables. Depending on the value of the correlation coefficient we have the following results:

- Correlation coefficient = 1, type of a perfect positive correlation
- Correlation coefficient = 0, zero correlation
- Correlation coefficient = -1, type of a perfect negative correlation

Now, based on the dataset of number of deaths by risk factors we examine the relationship between various attributes using the regression model and understand the correlation coefficient. The risk factors like air pollution, alcohol use, drug use, smoking, etc. would be used to examine the relationship with respect to the year in a particular entity. Here, since the dataset contains huge volume of data having all the entities, we create a subset of the data and create a sample from the original dataset which has all the values extracted of a particular entity in general. This would help us analyze the relationship between the year and risk factors of that particular entity and hence this extraction from the dataset was done which would give us a clear idea about the risk factors in that entity. We can also examine the relationship between the year and different risk factors for all the countries in all to check and analyze the overall risk factors in the world.

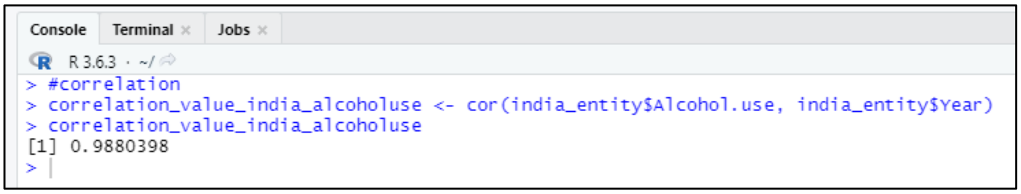
Data visualization was performed where plots like boxplot and density plot for the risk factors were created to get an understanding over the data before performing the regression model. These visualizations help in knowing about the risk factors and their statistical values along with the bandwidth. Once the visualizations were created, regression model was built for the attributes of the risk factors against the year. The first relationship examined was between the year and the risk factor of alcohol use in the entity India. The correlation coefficient value was computed and based on the value of the coefficient we decide the result. Regression table for each of the relationship was also created to understand the different values like the p-value confidence interval, beta value, etc. Similarly, the relationship was examined between the other risk factors like drug use, smoking, etc. with respect to the year for entities of North America and Maldives.

MV Regression helps to measure the angle of more than one independent variable and more than one dependent variable finding the relationship between the variables. It is used to predict the behavior of the outcome variable and the relationship of the predictor variable and how the variables are changing. Here, in the case of applying a MV regression, we test the relationship between two risk factors against a year for each entity and print the summary of the model which would give us the basic analysis of the model generated. Other statistical values of the regression analysis like the coef, sigma, vcov and anova were also displayed.

Results

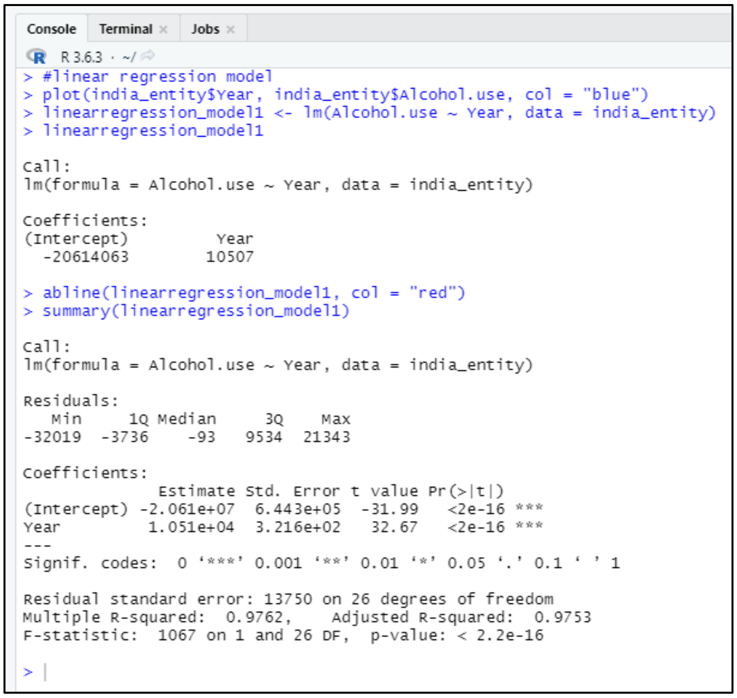
1. Regression model 1: Entity – India

a. Correlation Coefficient:



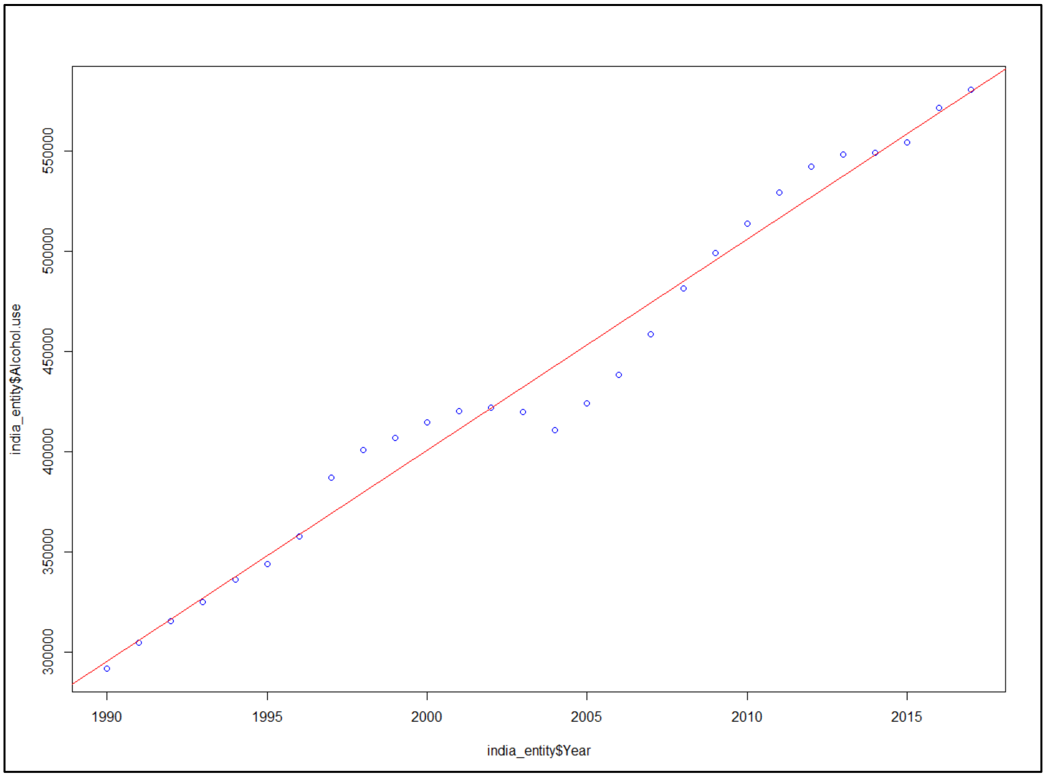
```
> #correlation
> correlation_value_india_alcoholuse <- cor(india_entity$Alcohol.use, india_entity$Year)
> correlation_value_india_alcoholuse
[1] 0.9880398
> 
```

b. Regression Model:

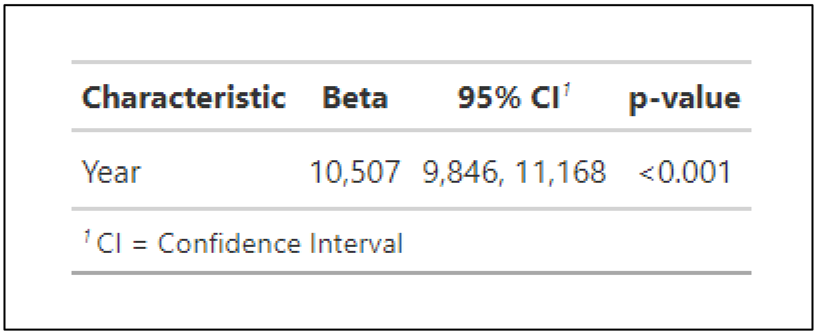


```
> #linear regression model
> plot(india_entity$Year, india_entity$Alcohol.use, col = "blue")
> linearregression_model1 <- lm(Alcohol.use ~ Year, data = india_entity)
> linearregression_model1

Call:
lm(formula = Alcohol.use ~ Year, data = india_entity)

Coefficients:
(Intercept)         Year  
  -20614063        10507  

> abline(linearregression_model1, col = "red")
> summary(linearregression_model1)

Call:
lm(formula = Alcohol.use ~ Year, data = india_entity)

Residuals:
   Min     1Q Median     3Q    Max 
-32019  -3736    -93   9534  21343 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.061e+07  6.443e+05  -31.99   <2e-16 ***
Year         1.051e+04  3.216e+02   32.67   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13750 on 26 degrees of freedom
Multiple R-squared:  0.9762,	Adjusted R-squared:  0.9753 
F-statistic:  1067 on 1 and 26 DF,  p-value: < 2.2e-16

> 
```

c. Regression Plot:



d. Regression Table:

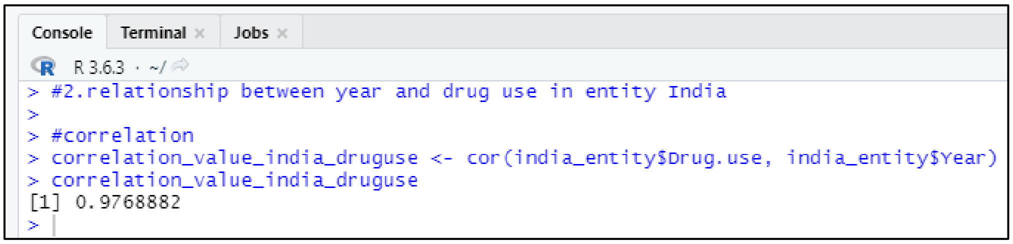


Characteristic	Beta	95% CI[1]	p-value
Year	10,507	9,846, 11,168	<0.001

[1] CI = Confidence Interval

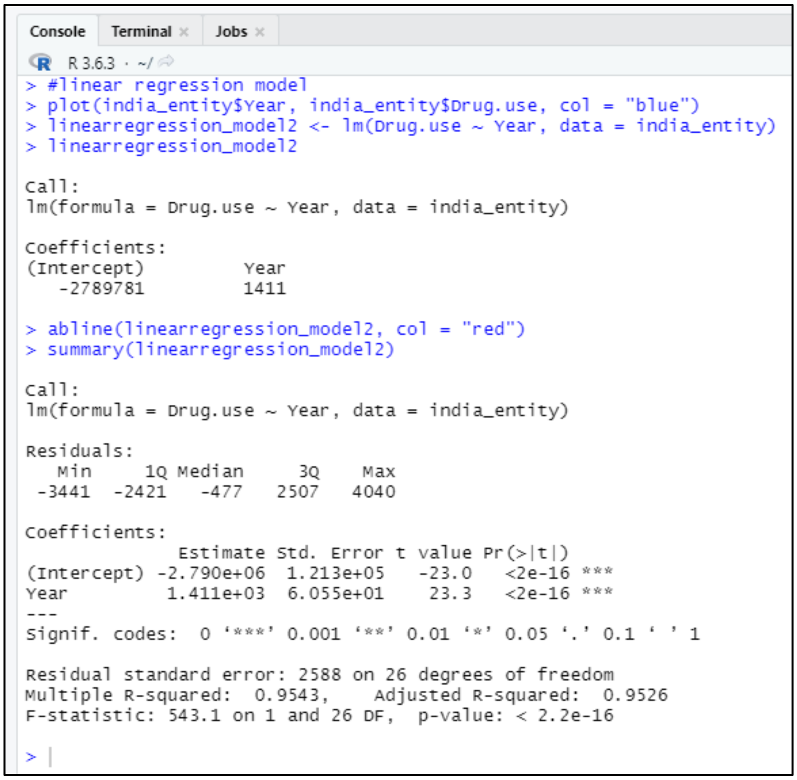
2. <u>Regression model 2: Entity – India</u>

a. Correlation Coefficient:



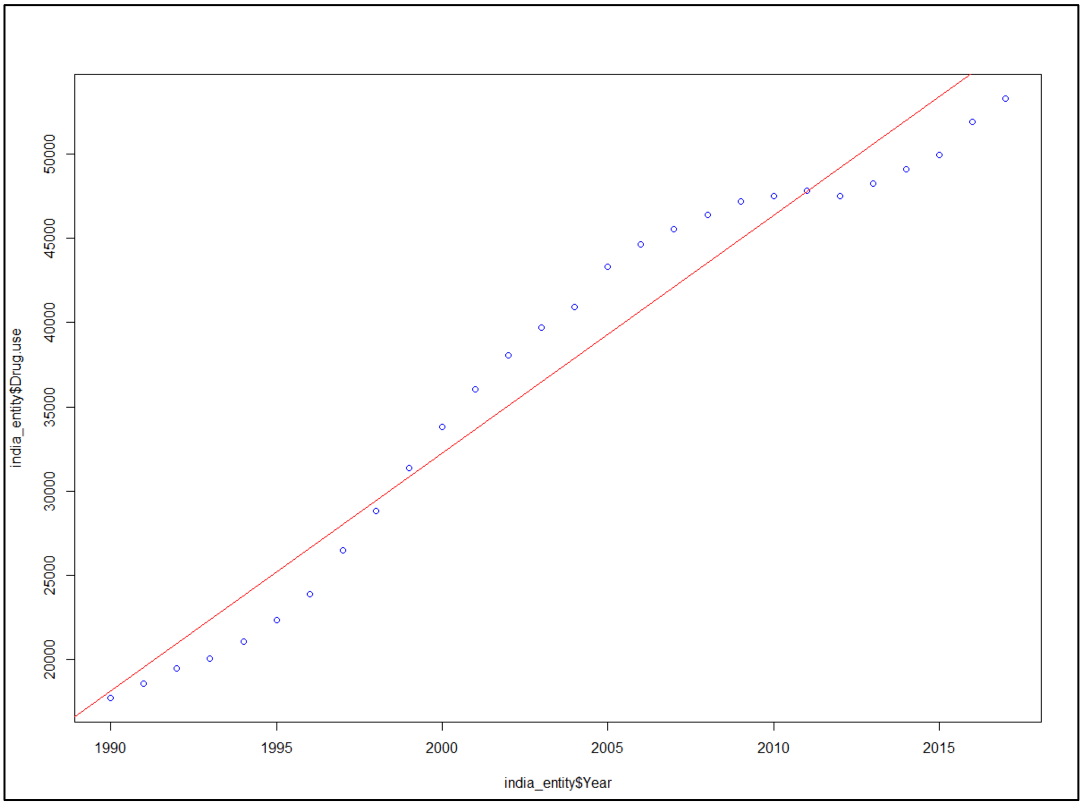
```
> #2.relationship between year and drug use in entity India
>
> #correlation
> correlation_value_india_druguse <- cor(india_entity$Drug.use, india_entity$Year)
> correlation_value_india_druguse
[1] 0.9768882
>
```

b. Regression Model:

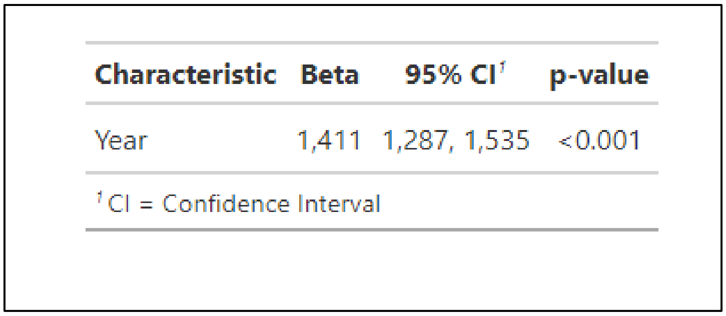


```
> #linear regression model
> plot(india_entity$Year, india_entity$Drug.use, col = "blue")
> linearregression_model2 <- lm(Drug.use ~ Year, data = india_entity)
> linearregression_model2

Call:
lm(formula = Drug.use ~ Year, data = india_entity)

Coefficients:
(Intercept)         Year
   -2789781         1411

> abline(linearregression_model2, col = "red")
> summary(linearregression_model2)

Call:
lm(formula = Drug.use ~ Year, data = india_entity)

Residuals:
   Min     1Q Median     3Q    Max
 -3441  -2421   -477   2507   4040

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.790e+06  1.213e+05   -23.0   <2e-16 ***
Year         1.411e+03  6.055e+01    23.3   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2588 on 26 degrees of freedom
Multiple R-squared:  0.9543,    Adjusted R-squared:  0.9526
F-statistic: 543.1 on 1 and 26 DF,  p-value: < 2.2e-16

> |
```

c. Regression Plot:



d. Regression Table:

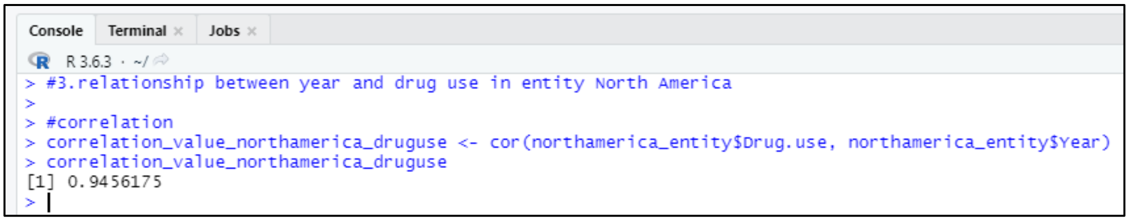


Characteristic	Beta	95% CI[1]	p-value
Year	1,411	1,287, 1,535	<0.001

[1] CI = Confidence Interval

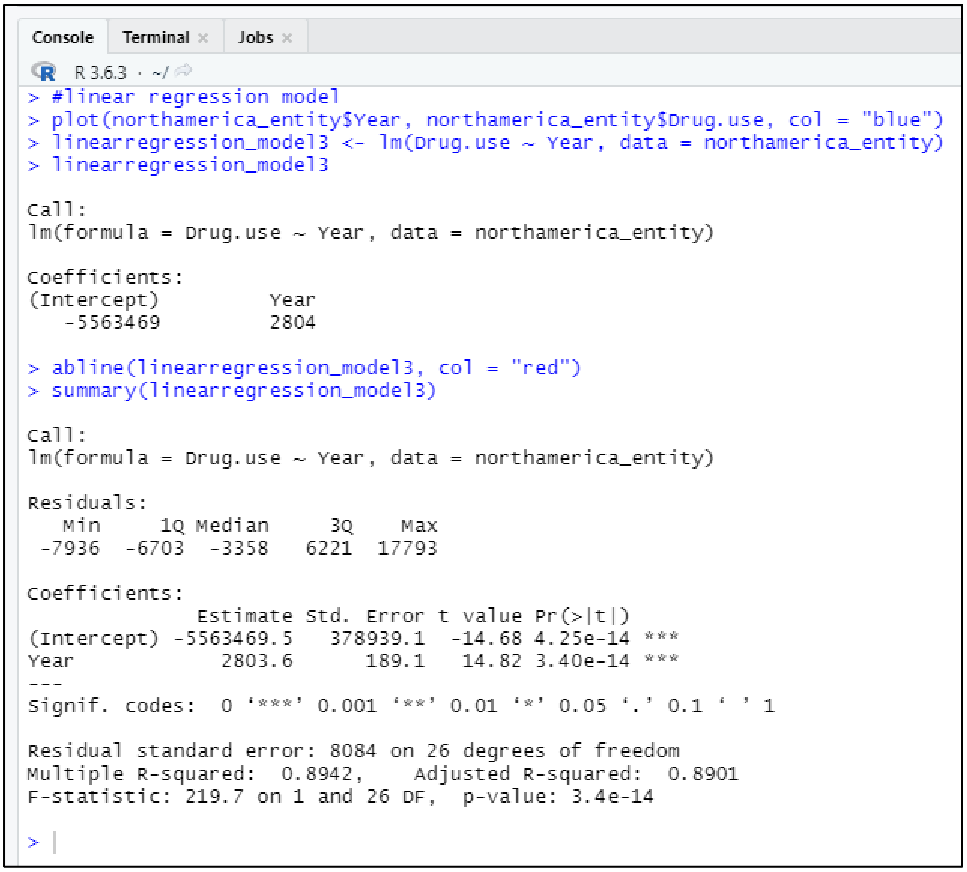
3. Regression model 3: Entity – North America

a. Correlation Coefficient:



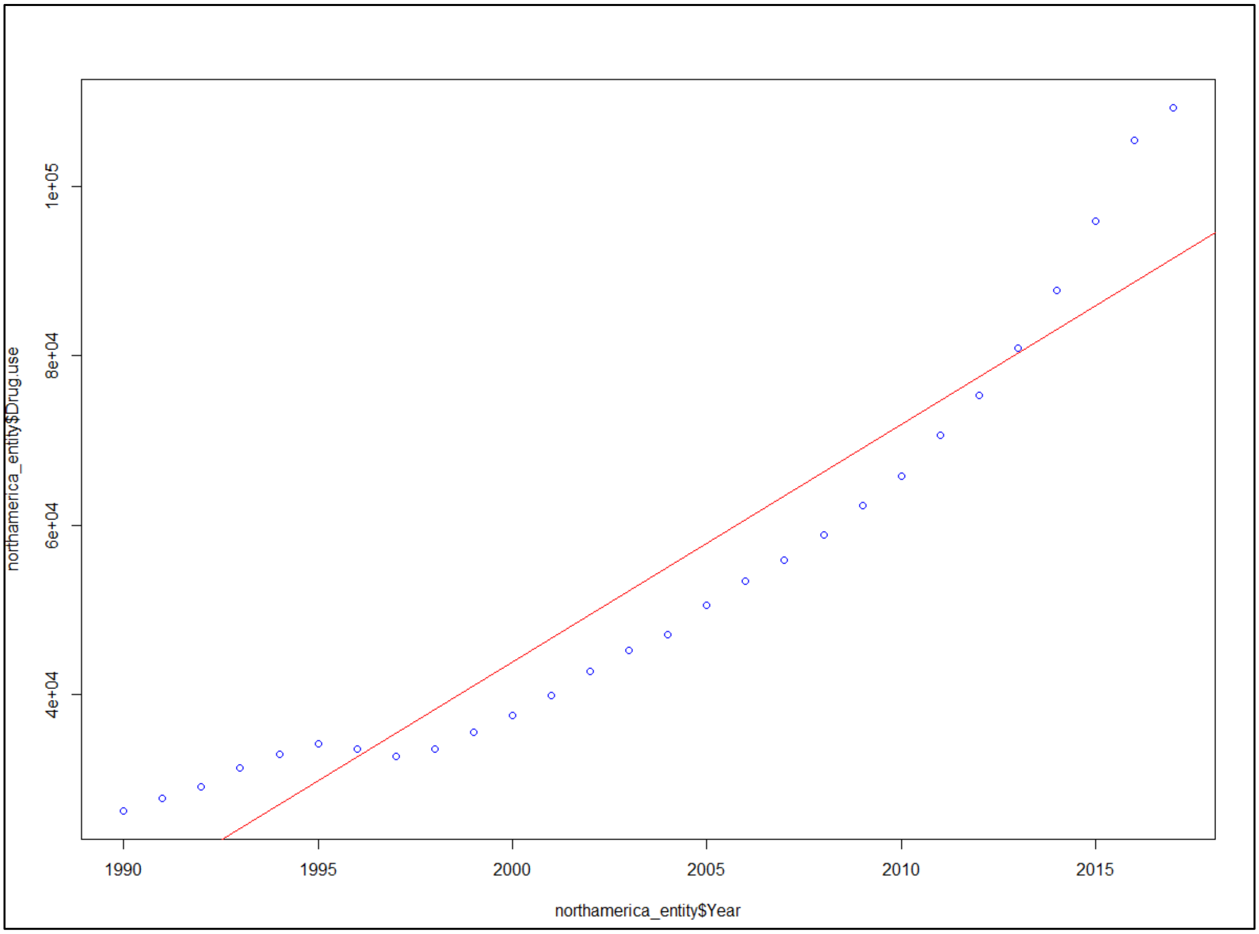
```
> #3.relationship between year and drug use in entity North America
>
> #correlation
> correlation_value_northamerica_druguse <- cor(northamerica_entity$Drug.use, northamerica_entity$Year)
> correlation_value_northamerica_druguse
[1] 0.9456175
> 
```

b. Regression Model:

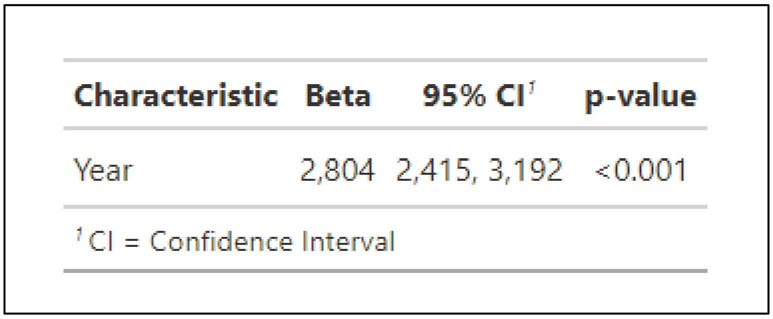


```
> #linear regression model
> plot(northamerica_entity$Year, northamerica_entity$Drug.use, col = "blue")
> linearregression_model3 <- lm(Drug.use ~ Year, data = northamerica_entity)
> linearregression_model3

Call:
lm(formula = Drug.use ~ Year, data = northamerica_entity)

Coefficients:
(Intercept)         Year  
   -5563469         2804  

> abline(linearregression_model3, col = "red")
> summary(linearregression_model3)

Call:
lm(formula = Drug.use ~ Year, data = northamerica_entity)

Residuals:
   Min     1Q Median     3Q    Max 
 -7936  -6703  -3358   6221  17793 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -5563469.5   378939.1  -14.68 4.25e-14 ***
Year            2803.6      189.1   14.82 3.40e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8084 on 26 degrees of freedom
Multiple R-squared:  0.8942,	Adjusted R-squared:  0.8901 
F-statistic: 219.7 on 1 and 26 DF,  p-value: 3.4e-14

> 
```

c. Regression Plot:



d. Regression Table:

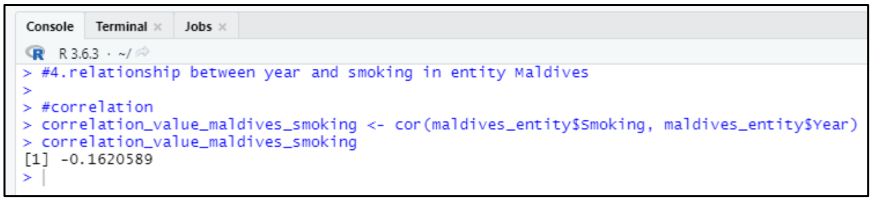


Characteristic	Beta	95% CI[1]	p-value
Year	2,804	2,415, 3,192	<0.001

[1] CI = Confidence Interval

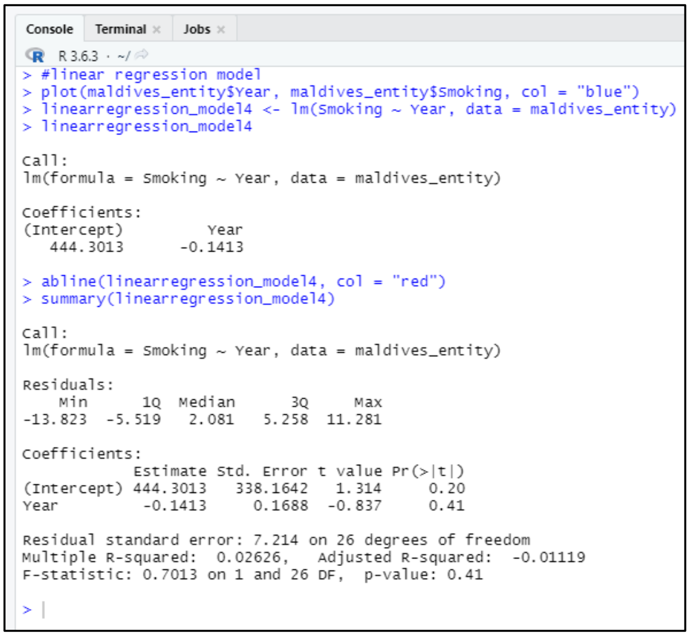
4. <u>Regression model 4: Entity – Maldives</u>

a. Correlation Coefficient:



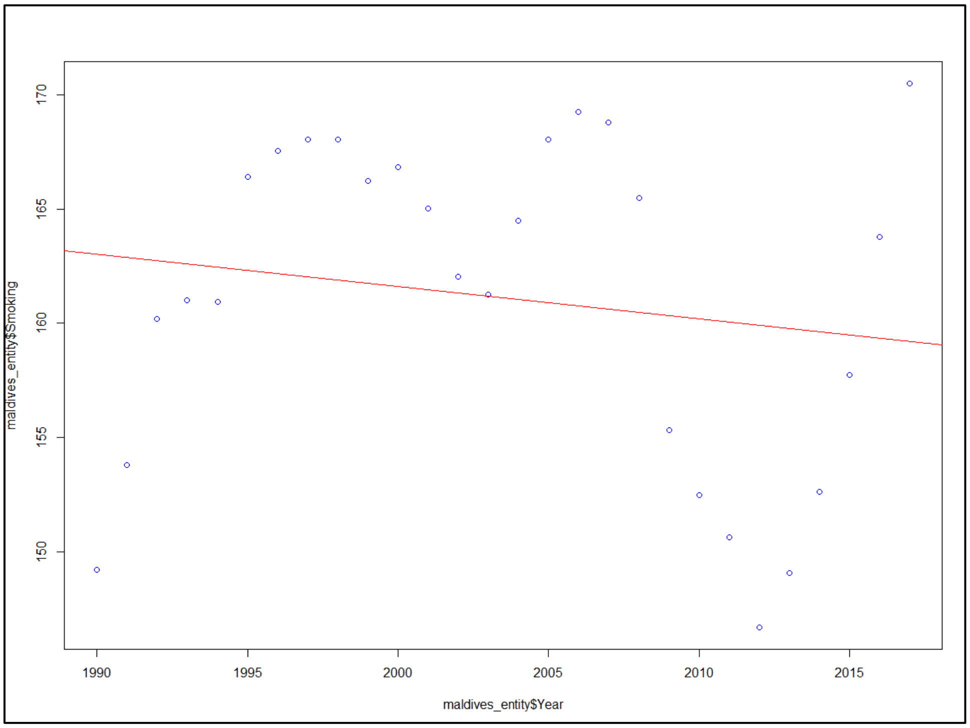
```
> #4.relationship between year and smoking in entity Maldives
>
> #correlation
> correlation_value_maldives_smoking <- cor(maldives_entity$Smoking, maldives_entity$Year)
> correlation_value_maldives_smoking
[1] -0.1620589
>
```

b. Regression Model:

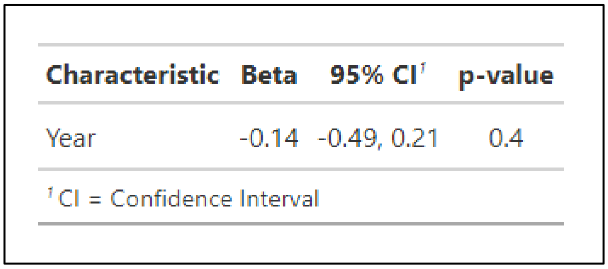


```
> #linear regression model
> plot(maldives_entity$Year, maldives_entity$Smoking, col = "blue")
> linearregression_model4 <- lm(Smoking ~ Year, data = maldives_entity)
> linearregression_model4

Call:
lm(formula = Smoking ~ Year, data = maldives_entity)

Coefficients:
(Intercept)         Year
   444.3013      -0.1413

> abline(linearregression_model4, col = "red")
> summary(linearregression_model4)

Call:
lm(formula = Smoking ~ Year, data = maldives_entity)

Residuals:
    Min      1Q  Median      3Q     Max
-13.823  -5.519   2.081   5.258  11.281

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 444.3013   338.1642   1.314     0.20
Year         -0.1413     0.1688  -0.837     0.41

Residual standard error: 7.214 on 26 degrees of freedom
Multiple R-squared:  0.02626,	Adjusted R-squared:  -0.01119
F-statistic: 0.7013 on 1 and 26 DF,  p-value: 0.41

>
```

c. Regression Plot:



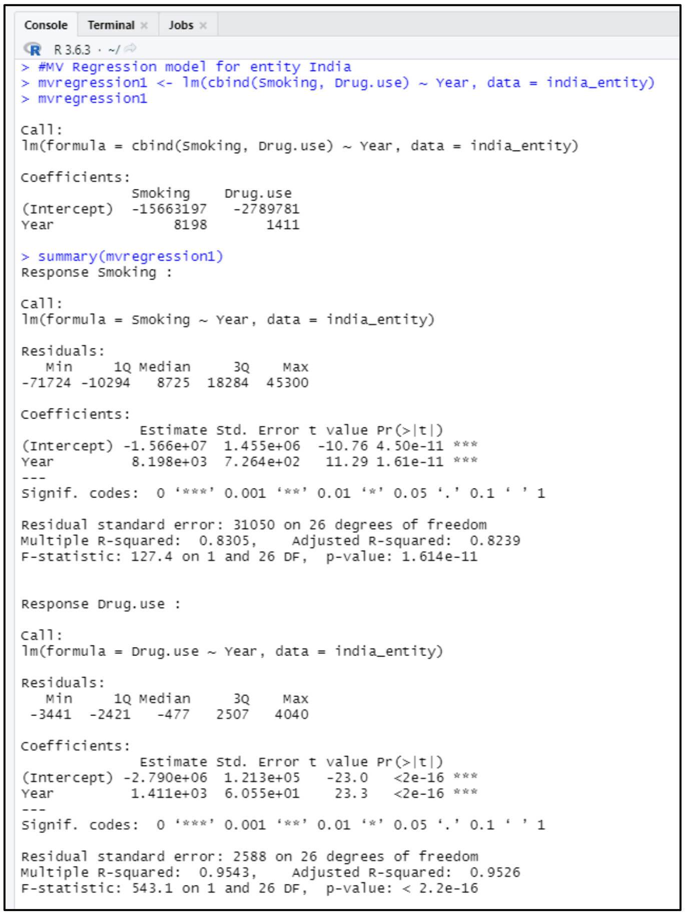
d. Regression Table:



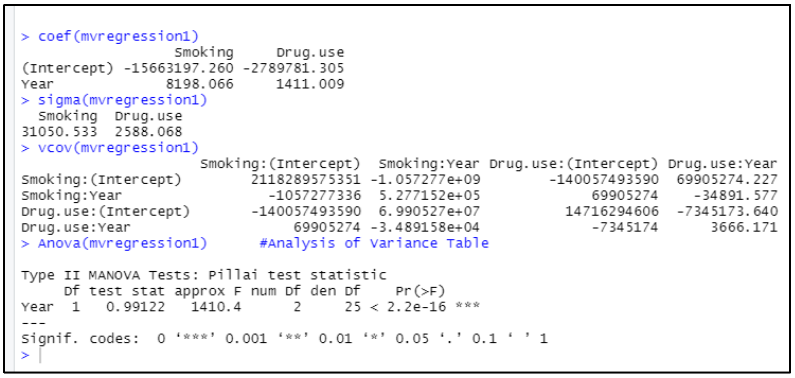
Characteristic	Beta	95% CI[1]	p-value
Year	-0.14	-0.49, 0.21	0.4

[1] CI = Confidence Interval

5. MV Regression Model 1: Entity – India

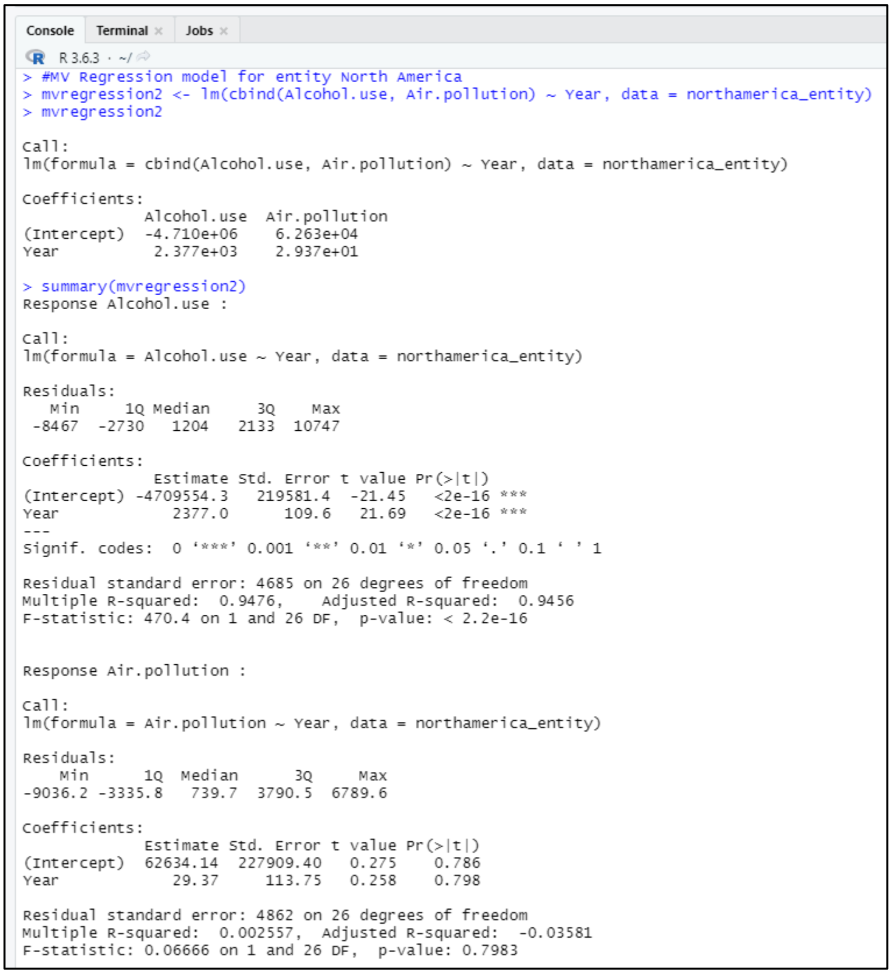


```
> #MV Regression model for entity India
> mvregression1 <- lm(cbind(Smoking, Drug.use) ~ Year, data = india_entity)
> mvregression1

Call:
lm(formula = cbind(Smoking, Drug.use) ~ Year, data = india_entity)

Coefficients:
             Smoking    Drug.use
(Intercept)  -15663197   -2789781
Year              8198       1411

> summary(mvregression1)
Response Smoking :

Call:
lm(formula = Smoking ~ Year, data = india_entity)

Residuals:
   Min     1Q Median     3Q    Max 
-71724 -10294   8725  18284  45300 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.566e+07  1.455e+06  -10.76 4.50e-11 ***
Year         8.198e+03  7.264e+02   11.29 1.61e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31050 on 26 degrees of freedom
Multiple R-squared:  0.8305,	Adjusted R-squared:  0.8239 
F-statistic: 127.4 on 1 and 26 DF,  p-value: 1.614e-11


Response Drug.use :

Call:
lm(formula = Drug.use ~ Year, data = india_entity)

Residuals:
   Min     1Q Median     3Q    Max 
 -3441  -2421   -477   2507   4040 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.790e+06  1.213e+05   -23.0   <2e-16 ***
Year         1.411e+03  6.055e+01    23.3   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2588 on 26 degrees of freedom
Multiple R-squared:  0.9543,	Adjusted R-squared:  0.9526 
F-statistic: 543.1 on 1 and 26 DF,  p-value: < 2.2e-16
```

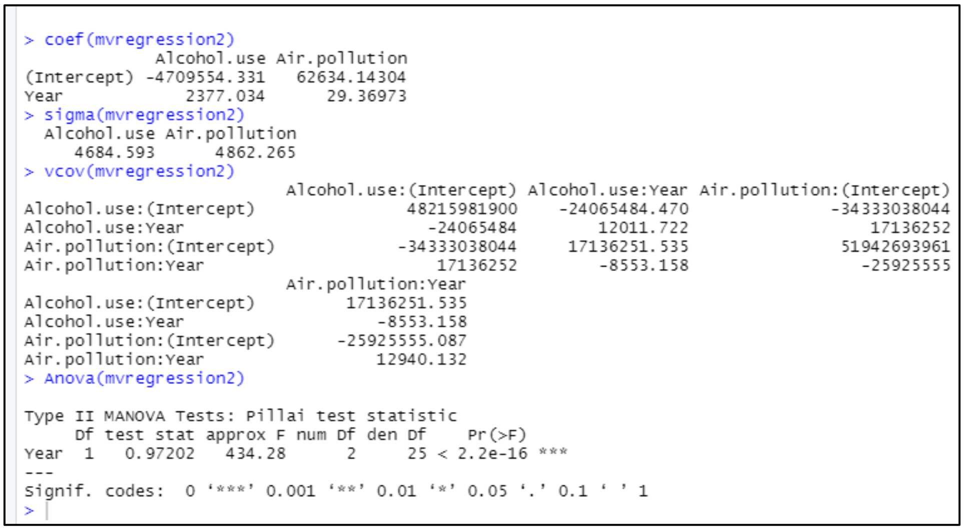


```
> coef(mvregression1)
                 Smoking     Drug.use
(Intercept) -15663197.260 -2789781.305
Year             8198.066     1411.009
> sigma(mvregression1)
  Smoking  Drug.use 
31050.533  2588.068 
> vcov(mvregression1)
                    Smoking:(Intercept)  Smoking:Year Drug.use:(Intercept) Drug.use:Year
Smoking:(Intercept)       2118289575351 -1.057277e+09         -140057493590  69905274.227
Smoking:Year                -1057277336  5.277152e+05             69905274    -34891.577
Drug.use:(Intercept)      -140057493590  6.990527e+07          14716294606  -7345173.640
Drug.use:Year                  69905274 -3.489158e+04             -7345174      3666.171
> Anova(mvregression1)      #Analysis of Variance Table

Type II MANOVA Tests: Pillai test statistic
     Df test stat approx F num Df den Df    Pr(>F)    
Year  1   0.99122   1410.4      2     25 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

6. MV Regression Model 2: Entity – North America

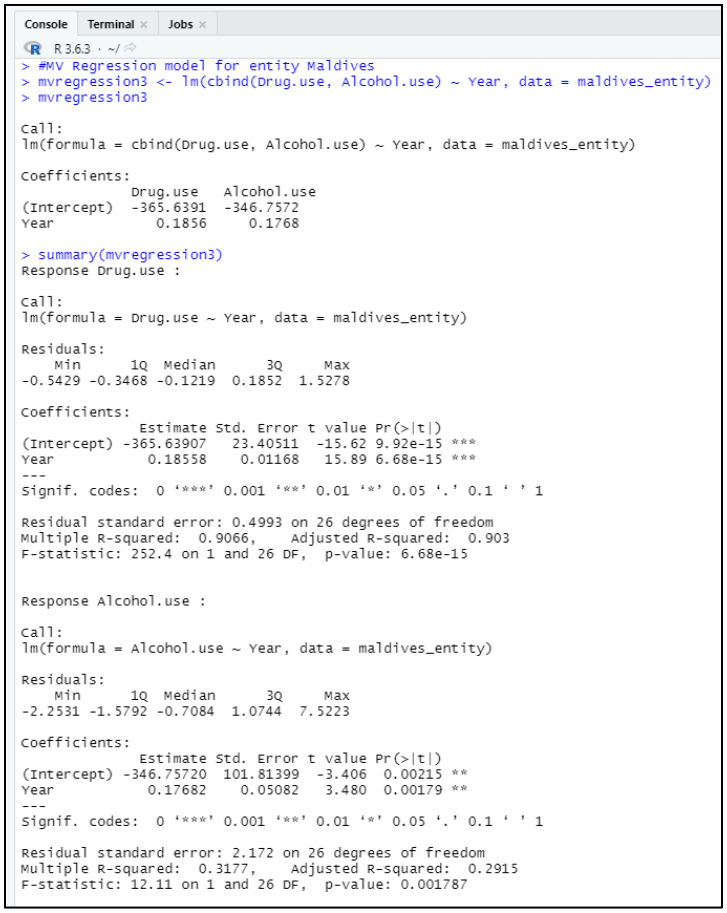


```
> #MV Regression model for entity North America
> mvregression2 <- lm(cbind(Alcohol.use, Air.pollution) ~ Year, data = northamerica_entity)
> mvregression2

Call:
lm(formula = cbind(Alcohol.use, Air.pollution) ~ Year, data = northamerica_entity)

Coefficients:
             Alcohol.use  Air.pollution
(Intercept)  -4.710e+06    6.263e+04   
Year          2.377e+03    2.937e+01   

> summary(mvregression2)
Response Alcohol.use :

Call:
lm(formula = Alcohol.use ~ Year, data = northamerica_entity)

Residuals:
   Min     1Q Median     3Q    Max 
 -8467  -2730   1204   2133  10747 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -4709554.3   219581.4  -21.45   <2e-16 ***
Year            2377.0      109.6   21.69   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4685 on 26 degrees of freedom
Multiple R-squared:  0.9476,	Adjusted R-squared:  0.9456 
F-statistic: 470.4 on 1 and 26 DF,  p-value: < 2.2e-16


Response Air.pollution :

Call:
lm(formula = Air.pollution ~ Year, data = northamerica_entity)

Residuals:
    Min      1Q  Median      3Q     Max 
-9036.2 -3335.8   739.7  3790.5  6789.6 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  62634.14  227909.40   0.275    0.786
Year            29.37     113.75   0.258    0.798

Residual standard error: 4862 on 26 degrees of freedom
Multiple R-squared:  0.002557,	Adjusted R-squared:  -0.03581 
F-statistic: 0.06666 on 1 and 26 DF,  p-value: 0.7983
```

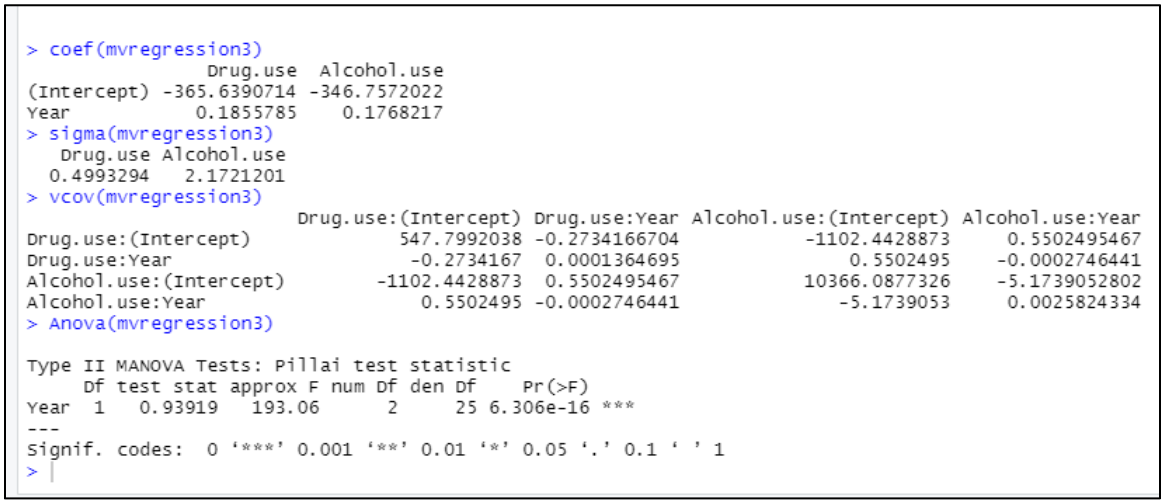


```
> coef(mvregression2)
             Alcohol.use Air.pollution
(Intercept) -4709554.331   62634.14304
Year            2377.034      29.36973
> sigma(mvregression2)
  Alcohol.use Air.pollution 
     4684.593      4862.265 
> vcov(mvregression2)
                           Alcohol.use:(Intercept) Alcohol.use:Year Air.pollution:(Intercept)
Alcohol.use:(Intercept)                48215981900    -24065484.470              -34333038044
Alcohol.use:Year                         -24065484        12011.722                  17136252
Air.pollution:(Intercept)             -34333038044     17136251.535               51942693961
Air.pollution:Year                        17136252        -8553.158                 -25925555
                           Air.pollution:Year
Alcohol.use:(Intercept)          17136251.535
Alcohol.use:Year                    -8553.158
Air.pollution:(Intercept)       -25925555.087
Air.pollution:Year                  12940.132
> Anova(mvregression2)

Type II MANOVA Tests: Pillai test statistic
     Df test stat approx F num Df den Df    Pr(>F)    
Year  1   0.97202   434.28      2     25 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

7. MV Regression Model 3: Entity – Maldives



```
> #MV Regression model for entity Maldives
> mvregression3 <- lm(cbind(Drug.use, Alcohol.use) ~ Year, data = maldives_entity)
> mvregression3

Call:
lm(formula = cbind(Drug.use, Alcohol.use) ~ Year, data = maldives_entity)

Coefficients:
             Drug.use   Alcohol.use
(Intercept)  -365.6391  -346.7572  
Year            0.1856     0.1768  

> summary(mvregression3)
Response Drug.use :

Call:
lm(formula = Drug.use ~ Year, data = maldives_entity)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.5429 -0.3468 -0.1219  0.1852  1.5278 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -365.63907   23.40511  -15.62 9.92e-15 ***
Year           0.18558    0.01168   15.89 6.68e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4993 on 26 degrees of freedom
Multiple R-squared:  0.9066,	Adjusted R-squared:  0.903 
F-statistic: 252.4 on 1 and 26 DF,  p-value: 6.68e-15


Response Alcohol.use :

Call:
lm(formula = Alcohol.use ~ Year, data = maldives_entity)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.2531 -1.5792 -0.7084  1.0744  7.5223 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)   
(Intercept) -346.75720  101.81399  -3.406  0.00215 **
Year           0.17682    0.05082   3.480  0.00179 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.172 on 26 degrees of freedom
Multiple R-squared:  0.3177,	Adjusted R-squared:  0.2915 
F-statistic: 12.11 on 1 and 26 DF,  p-value: 0.001787
```



```
> coef(mvregression3)
                Drug.use  Alcohol.use
(Intercept) -365.6390714 -346.7572022
Year           0.1855785    0.1768217
> sigma(mvregression3)
   Drug.use Alcohol.use 
  0.4993294   2.1721201 
> vcov(mvregression3)
                        Drug.use:(Intercept) Drug.use:Year Alcohol.use:(Intercept) Alcohol.use:Year
Drug.use:(Intercept)             547.7992038 -0.2734166704           -1102.4428873     0.5502495467
Drug.use:Year                     -0.2734167  0.0001364695               0.5502495    -0.0002746441
Alcohol.use:(Intercept)        -1102.4428873  0.5502495467           10366.0877326    -5.1739052802
Alcohol.use:Year                   0.5502495 -0.0002746441              -5.1739053     0.0025824334
> Anova(mvregression3)

Type II MANOVA Tests: Pillai test statistic
     Df test stat approx F num Df den Df    Pr(>F)    
Year  1   0.93919   193.06      2     25 6.306e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

Analysis

From the results it is observed that the independent variable is the year attribute, and the dependent variable is the risk factor of that particular entity. The correlation coefficient value for all the built models came out to be positive which implies that there is a similar relation between the two variables. The regression model built examines the relationship between the two variables of our dataset and the coefficient values helps to predict the dependent variable.

For example, in the first regression model built for the attribute year and alcohol use in the entity India, results show a positive correlation coefficient, and the values of the coefficients are -20614063 and 10507. Thus, the predicted equation for the alcohol use is, alcohol use = -20614063 + 10507 x year which helps in predicting the alcohol use in the entity India. We also observe that as the p-value is less than 0.05, we reject the null hypothesis which means that the year is not the only significant parameter to determine the risk factor in a particular entity because there could be other factors which would help in determining that risk factor. A regression plot and regression table for the same is also generated which helps in examining the relationship between the two variables and the table gives us information after the built of the model related to the p-value and confidence interval about whether the hypothesis is rejected or not.

Similarly, models were built for the other risk factors in various other entities to understand their relationship with the attributes of the dataset. The idea here was to extract all the information about a particular entity and have a model built on the same to understand the relationship between the variables among that particular entity rather than having an overall analysis for the same. This would thus help in having adequate information about an entity as a whole and to know which risk factor in particular affects the death rate in that entity.

Along with a simple regression model, multi variant regression model was also built where there were two dependent variables and an independent variable. This analysis here helps in examining the relationship between two dependent and independent variables. For example, here we built a MV regression model for the two risk factors drug use and alcohol use against the year attribute for the entity of Maldives. This analysis would be the same if done separately on individual parameters and thus MV regression helps in examining the relationship of the variables if they are to be examined against the same independent or dependent variable.

The reason for performing the regression analysis is because we want to examine the relationship between the year and the risk factor in a particular entity and find the analysis which year had the lowest or highest risk factor leading to the death rate and then predict the risk factor for the upcoming years. Therefore, the question of which year had the max or the min risk factor is answered by the regression plot and thus we can further predict the risk factors leading to the death rate for the coming years in a particular country which would help reducing that risk factor in that country minimizing the death rate.

Summary

Exploratory data analysis, hypothesis testing and building the regression model was performed here. EDA was the initial phase where we performed steps like data extraction, data cleaning, descriptive analysis, and data visualizations. These steps in the EDA helped in analyzing the data which gave us insights about the dataset and to better understand the data for further analysis. The next step after the EDA was hypothesis testing where one-sample, two-sample and paired t-test was performed which helped in answering the questions giving us information about the data related to the risk factor leading to the death rate in the country. This analysis was helpful in having an understanding about which risk factor was affected the most in a particular country and which risk factor needs to be reduced as it is affected the most in a country as compared to in another country.

Once the hypothesis testing was performed, we built a regression model to examine the relationship between the variables of the dataset. The regression model would thus help in predicting the risk factor in a particular entity in the year. Correlation coefficient chart and regression table gave us more insights about the relationship between each variable and the regression line plot distinguished the relationship between the attributes that were plotted.

In the EDA, descriptive analysis table and data visualizations were created where we got statistical information about the dataset with respect to the descriptive analysis and charts like boxplot, density plot and bar plot were created for the attributes of the dataset. In the hypothesis testing, we had one-sample and two-sample t-test being performed on the attributes and depending upon the p-value we either rejected the null hypothesis or failed to reject the null hypothesis. Also, depending on the hypothesis we decided whether a paired t-test could be performed or not and here in this case, it was observed that a paired t-test cannot be performed since we rejected the null hypothesis in the two-sample t-test.

Regression model was built after performing the hypothesis testing on the attributes. Here, we built a regression model on the risk factors with respect to the year which helped in analyzing and predicting the risk factor in the year in a particular country. Since the entire dataset had a huge volume of data and analyzing this would not be efficient, we extracted the data for each entity to analyze the relationship of the attributes within the entity. This would thus help in understanding and analyzing the risk factor in the year in each entity rather than an overall analysis for all entities together.

References

Prabhakaran, S. (n.d.). Linear Regression. R-Statistics.Co. Retrieved December 7, 2021, from http://r-statistics.co/Linear-Regression.html#:~:text=The%20most%20common%20metrics%20to%20look%20at%20while,zero%20the%20better%20%207%20more%20rows%20

Linear Regression in R. (n.d.). Datacamp. Retrieved December 7, 2021, from https://www.datacamp.com/community/tutorials/linear-regression-R

Zach, S. (2021, February 1). Correlation vs. Regression: What's the Difference? Statology.Org. Retrieved December 8, 2021, from https://www.statology.org/correlation-vs-regression/#:~:text=Differences%3A%20Regression%20is%20able%20to%20show%20a%20cause-and-effect,variable%2C%20based%20on%20the%20value%20of%20another%20variable.

Tutorial: tbl_regression. (2020, September 13). Danieldsjoberg.Com. Retrieved December 8, 2021, from https://www.danieldsjoberg.com/gtsummary/articles/tbl_regression.html#:~:text=The%20tbl_regression%20%28%29%20function%20takes%20a%20regression%20model,creates%20highly%20customizable%20analytic%20tables%20with%20sensible%20defaults.

Facer, C. (n.d.). How to Create a Correlation Matrix in R. DisplayR. Retrieved December 8, 2021, from https://www.displayr.com/how-to-create-a-correlation-matrix-in-r/

Getting started with Multivariate Multiple Regression. (n.d.). Data Library Virginia. Retrieved December 9, 2021, from https://data.library.virginia.edu/getting-started-with-multivariate-multiple-regression/#:~:text=Performing%20multivariate%20multiple%20regression%20in%20R%20requires%20wrapping,On%20the%20other%20side%20we%20add%20our%20predictors.

Zach, S. (2020, December 23). How to Plot Multiple Linear Regression Results in R. Statology.Org. Retrieved December 9, 2021, from https://www.statology.org/plot-multiple-linear-regression-in-r/

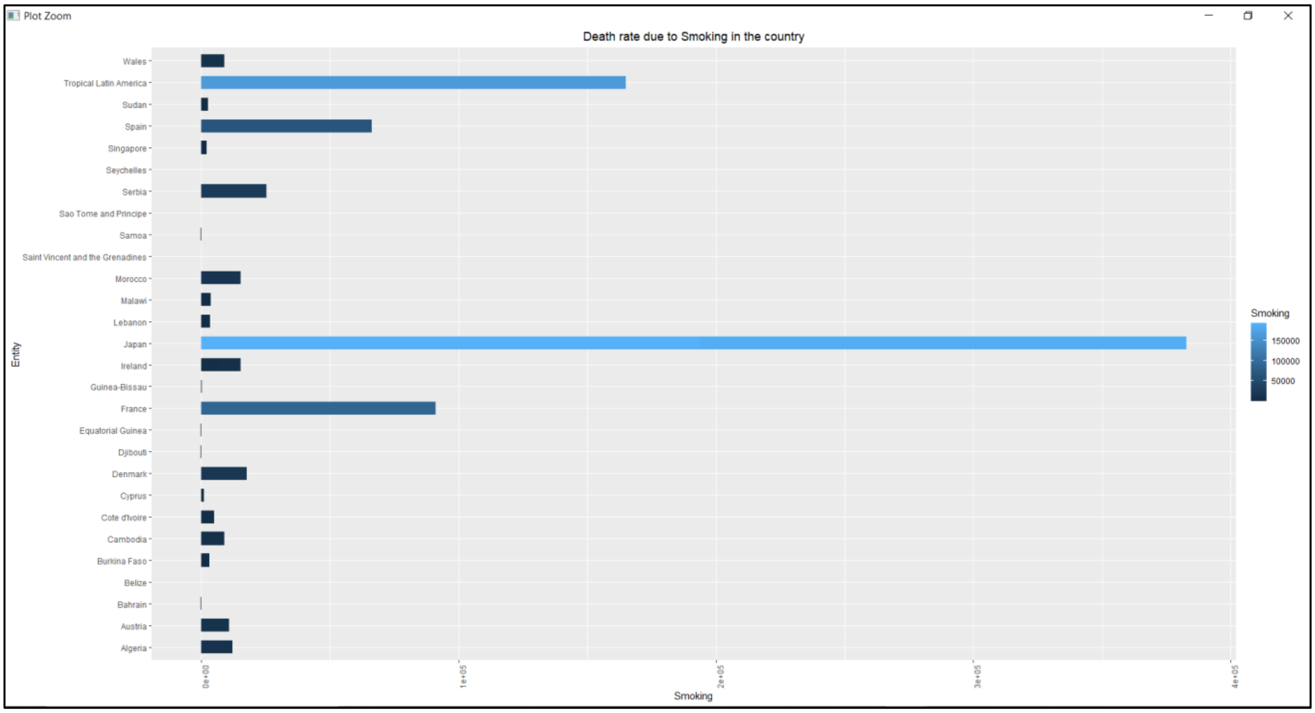
Appendix

Exploratory Data Analysis:

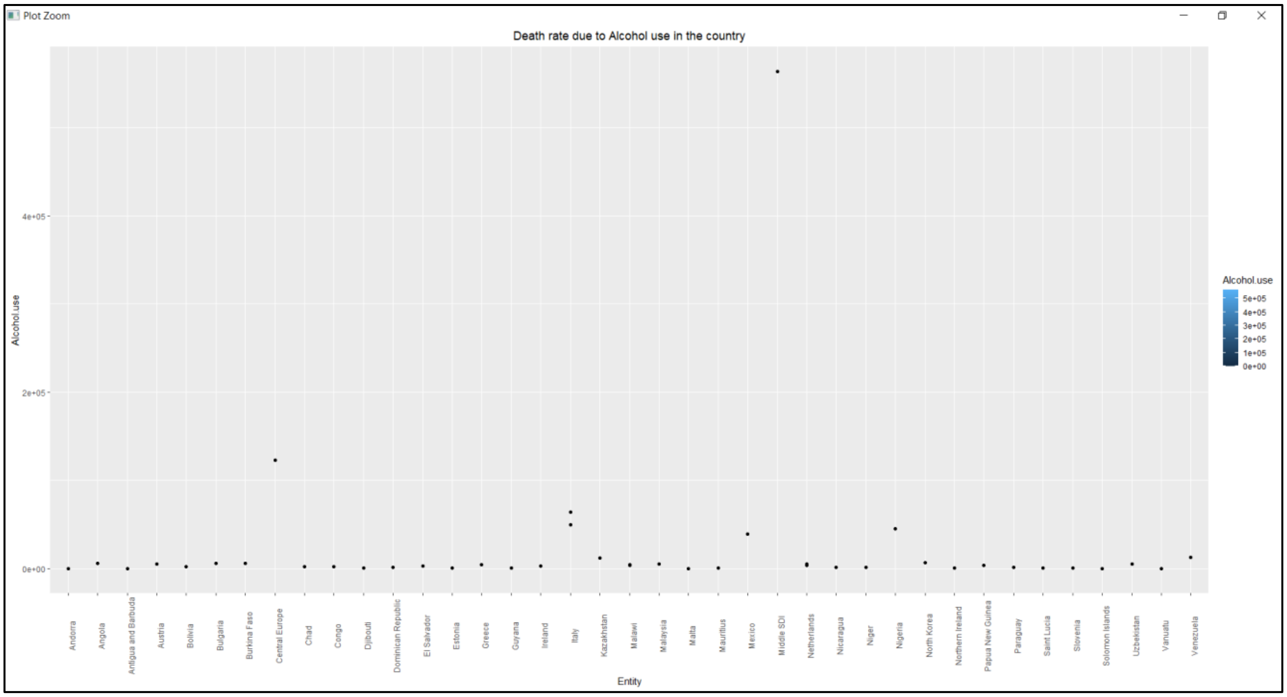
a. Descriptive Analysis Table

Statistical Value Risk Factor	*Min*	*Max*	*Mean*	*Median*	*Standard Deviation*	*Range*
Alcohol Use	-2315	2842854	50203	2803	195822	-2315 - 2842854
Air Pollution	8.52	4895476	95735.51	6125.09	390933.5	8.52 - 4895476
Smoking	11.707	7099111	133548.3	5935.789	529931.5	11.707 -7099111
Drug Use	1.240	585348.2	8890.242	408.5863	35415.12	1.240 - 585348.2
No access to handwash facility	0.0779	1239519	18933.05	252.4991	89810.37	0.0779 - 1239519

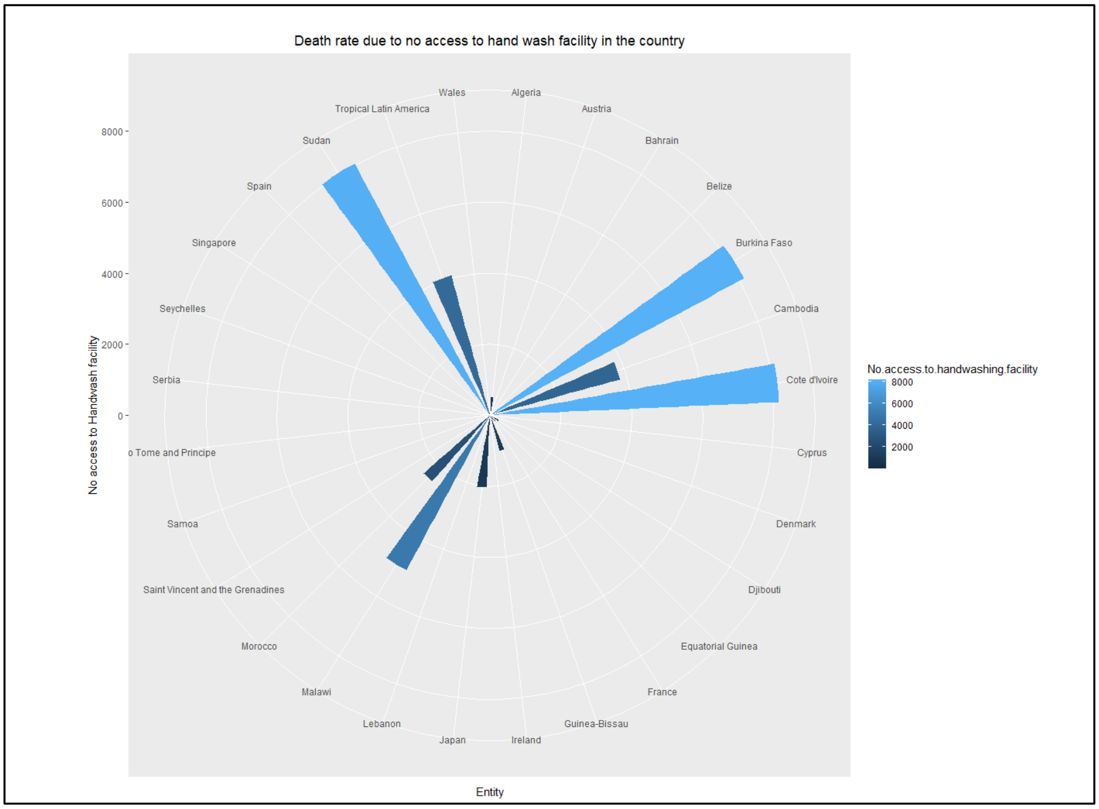
b. Graph 1 - Death rate due to Smoking in the country



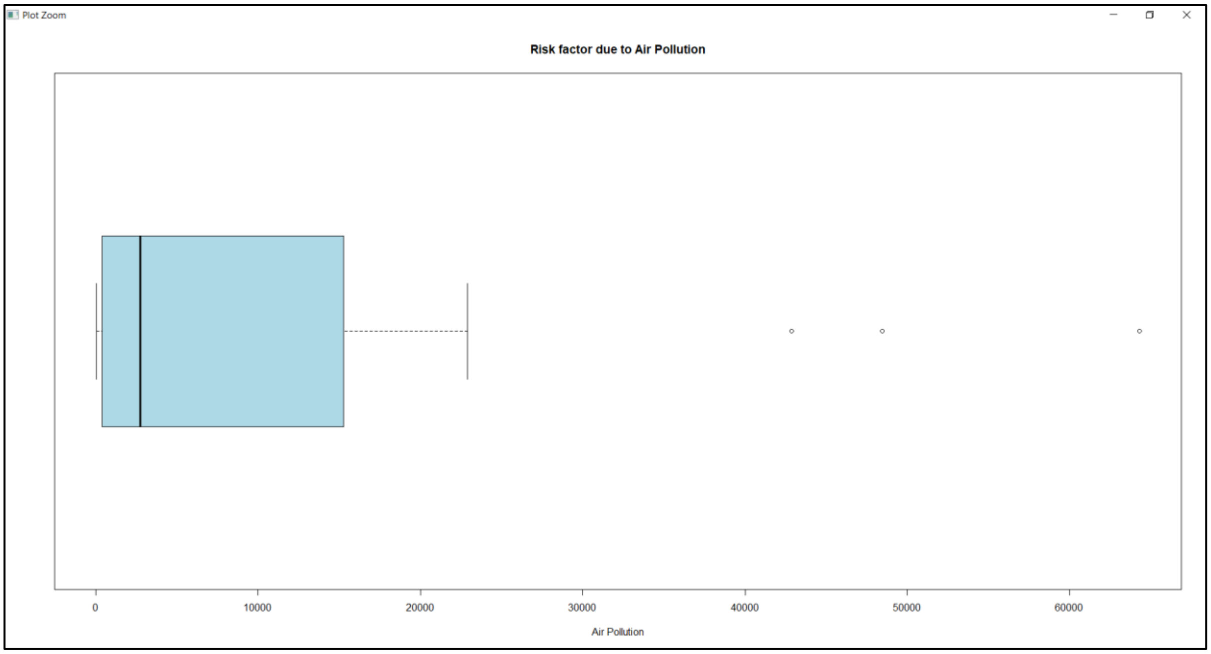
c. Graph 2 - Death rate due to Alcohol use in the country



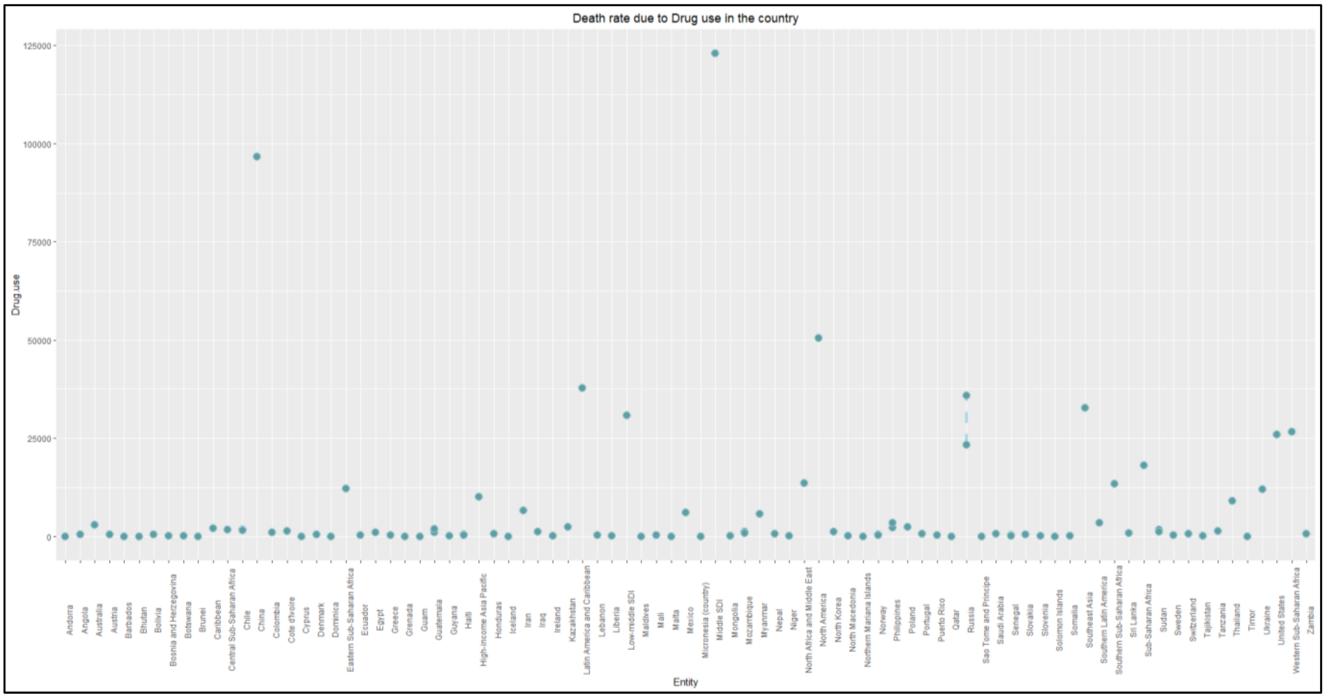
d. Graph 3 - Death rate due to no access to hand wash facility in the country



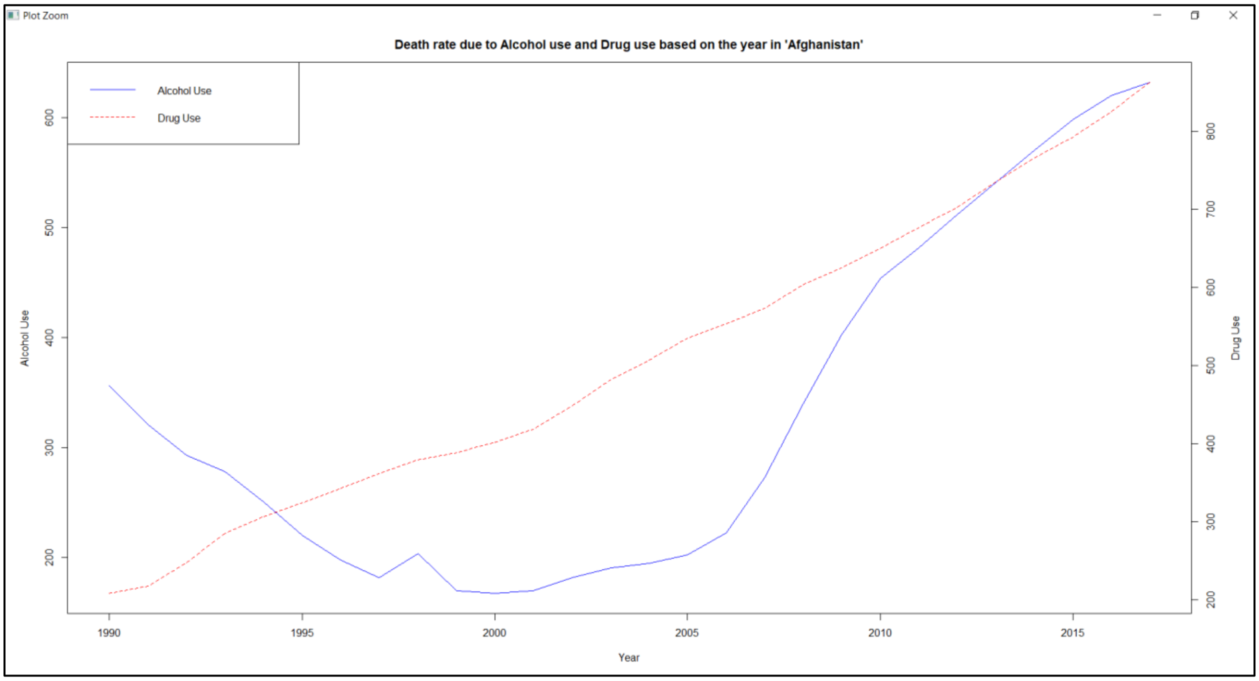
e. Graph 4 - Risk factor due to Air pollution



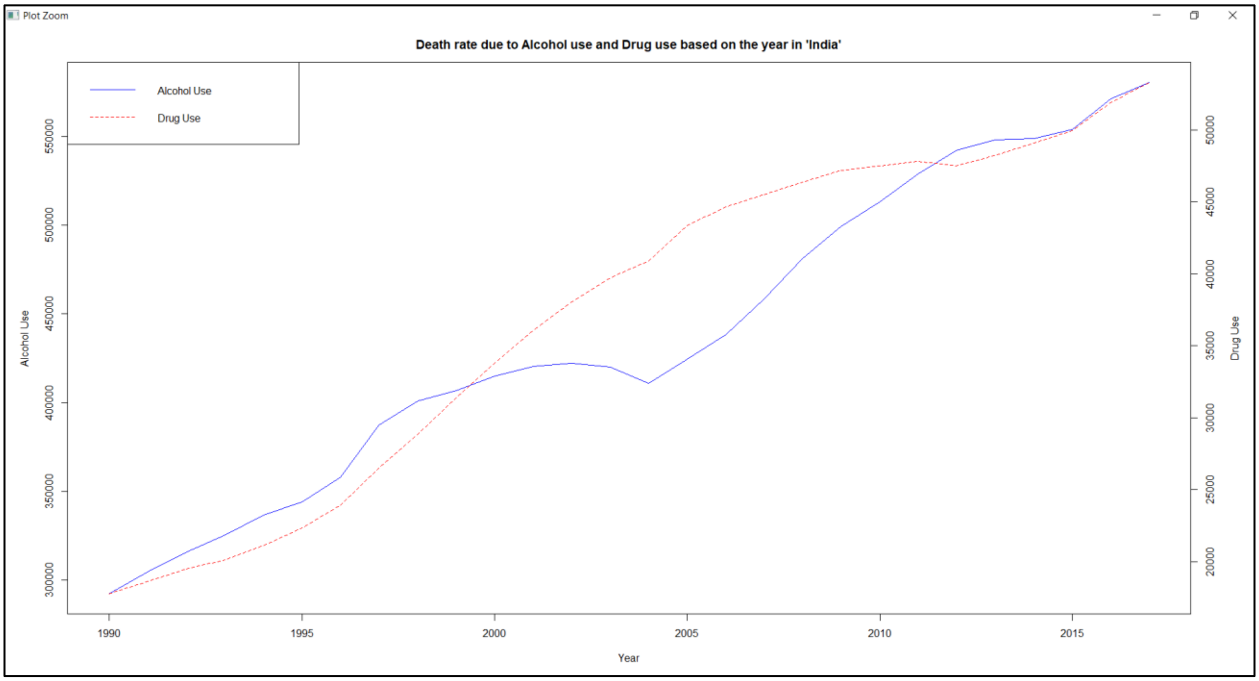
f. Graph 5 - Death rate due to Drug use in the country



g. Graph 6 - Death rate due to Alcohol use and Drug use based on the year in ‘Afghanistan’

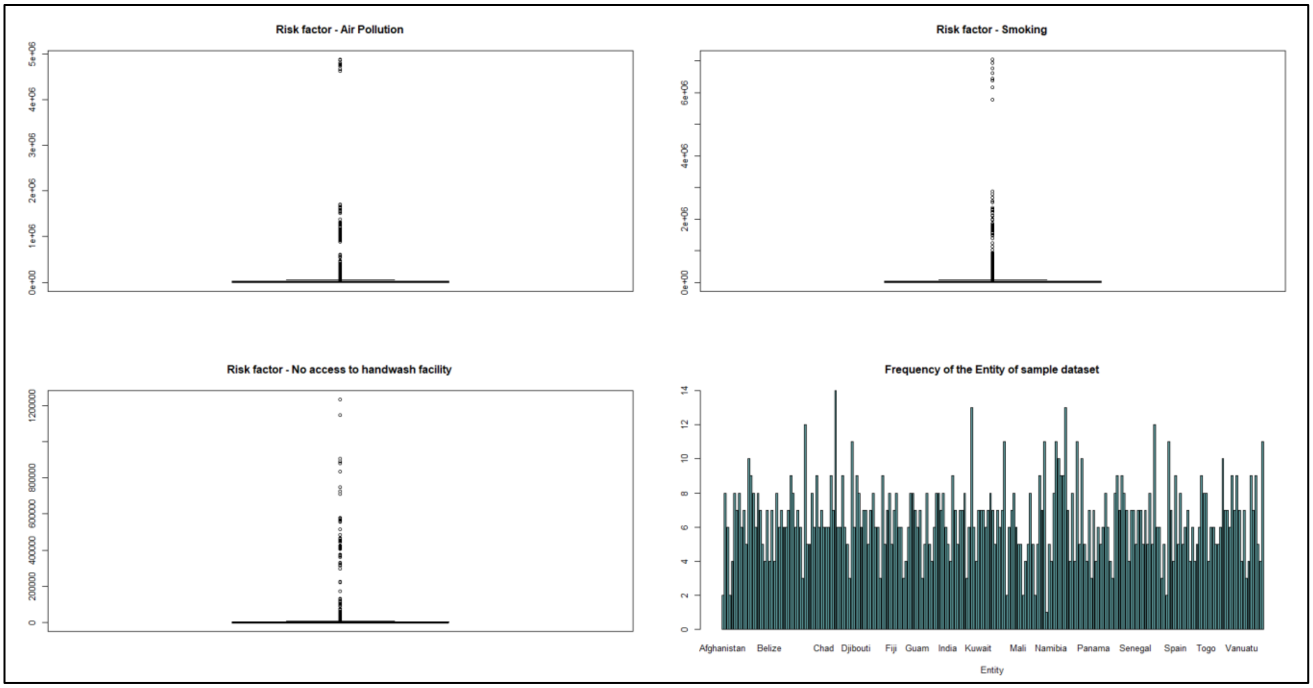


h. Graph 7 - Death rate due to Alcohol use and Drug use based on the year in ‘India’

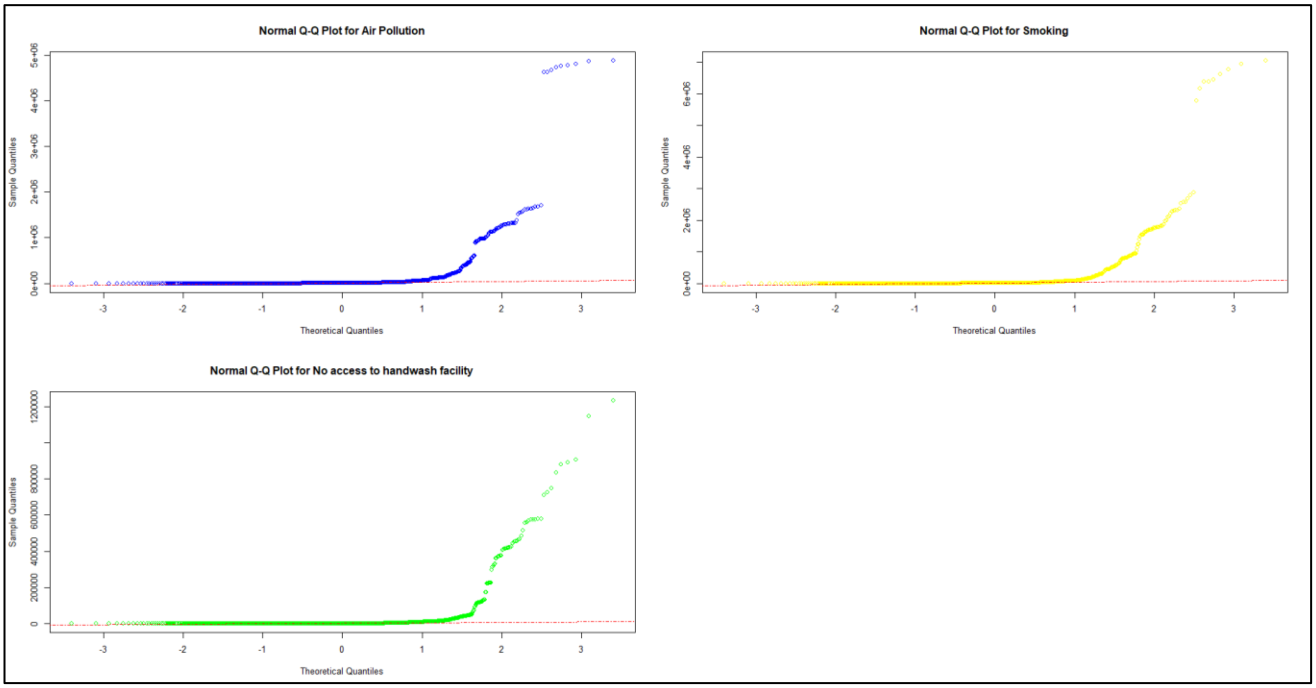


Hypothesis Testing:

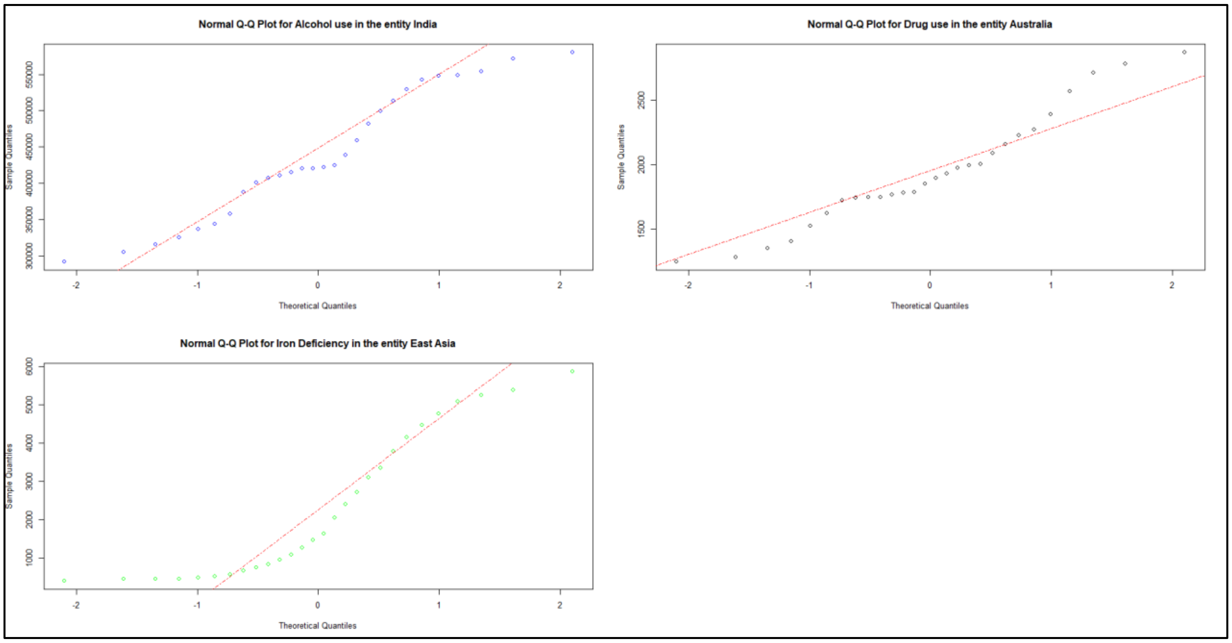
a. Data Visualization



b. One-sample t-test

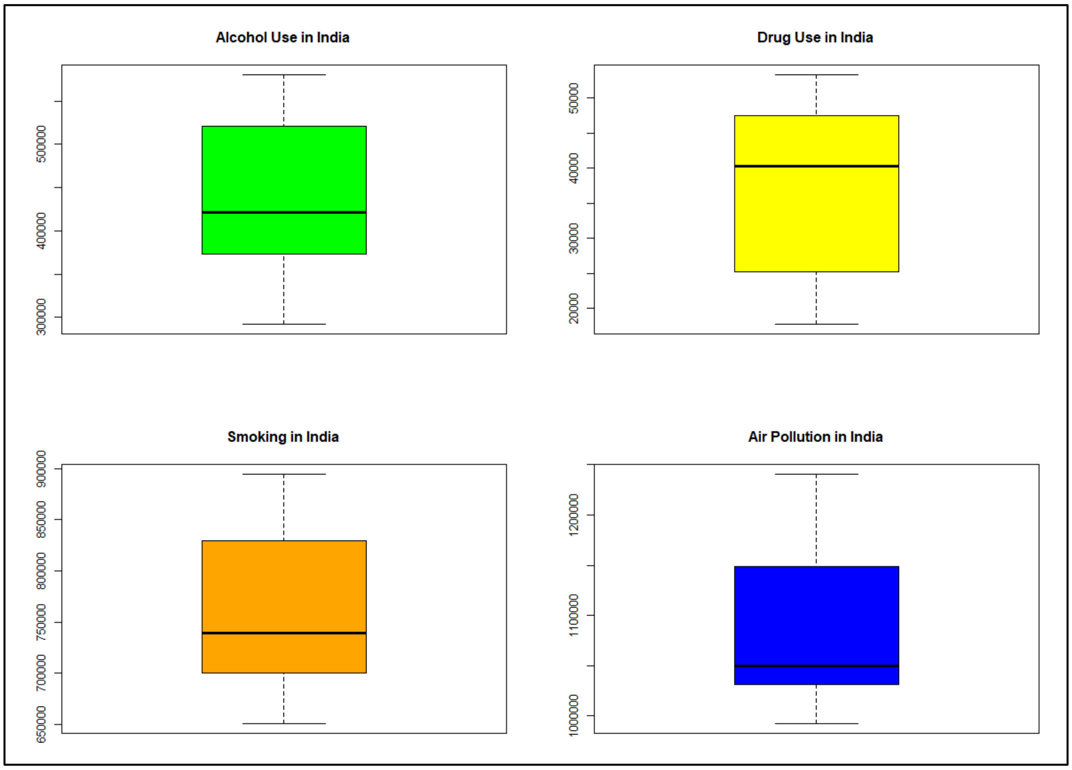


c. Two-sample t-test

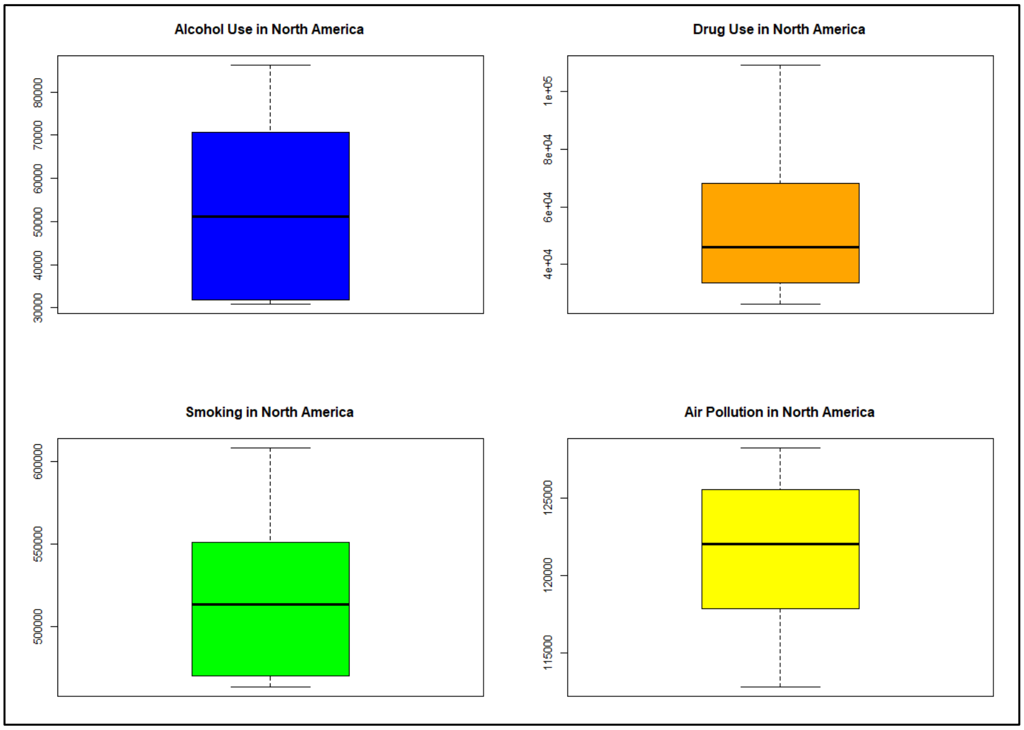


Regression Model:

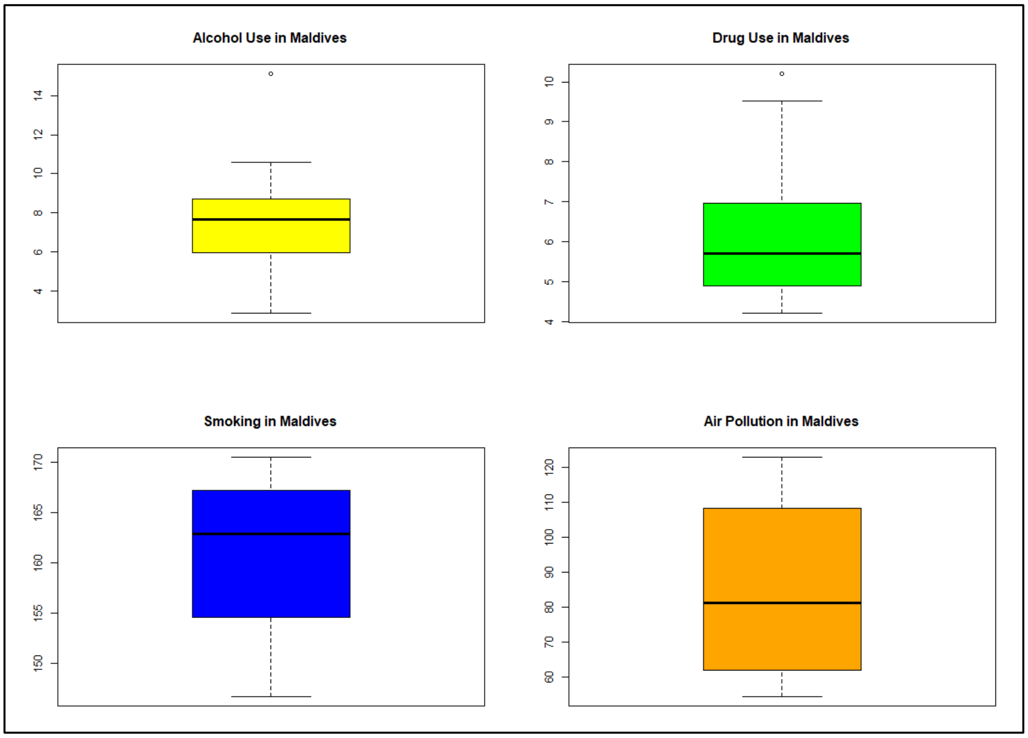
a. Graph 1: Boxplot of the attributes for Entity India



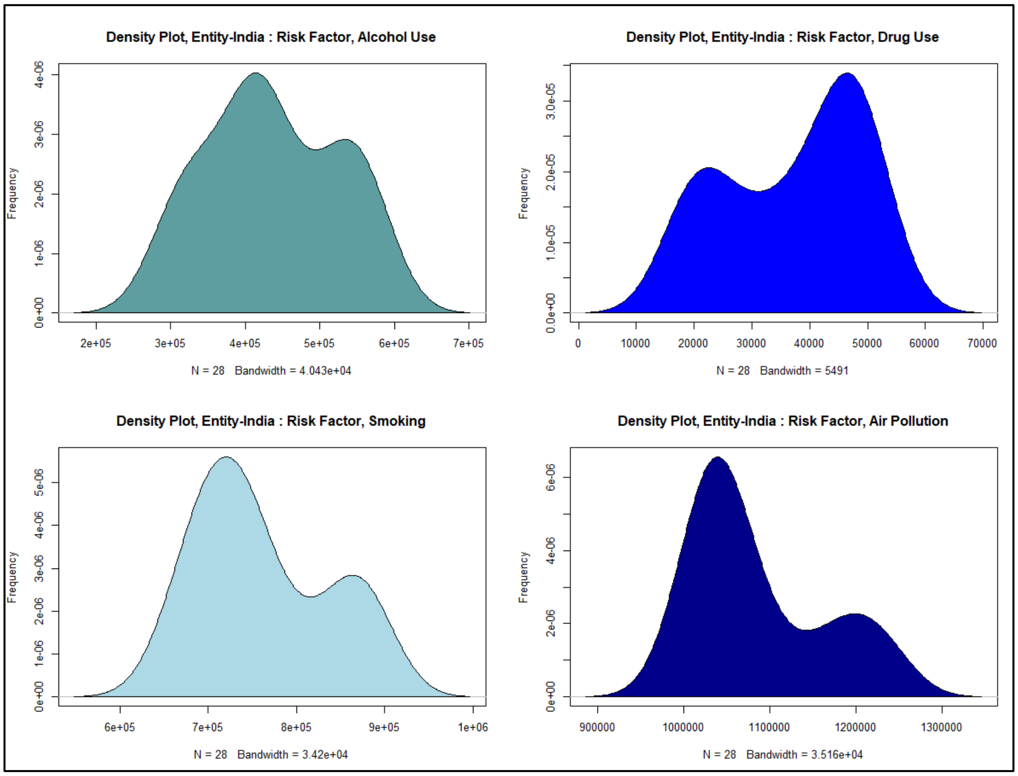
b. Graph 2: Boxplot of the attributes for Entity North America



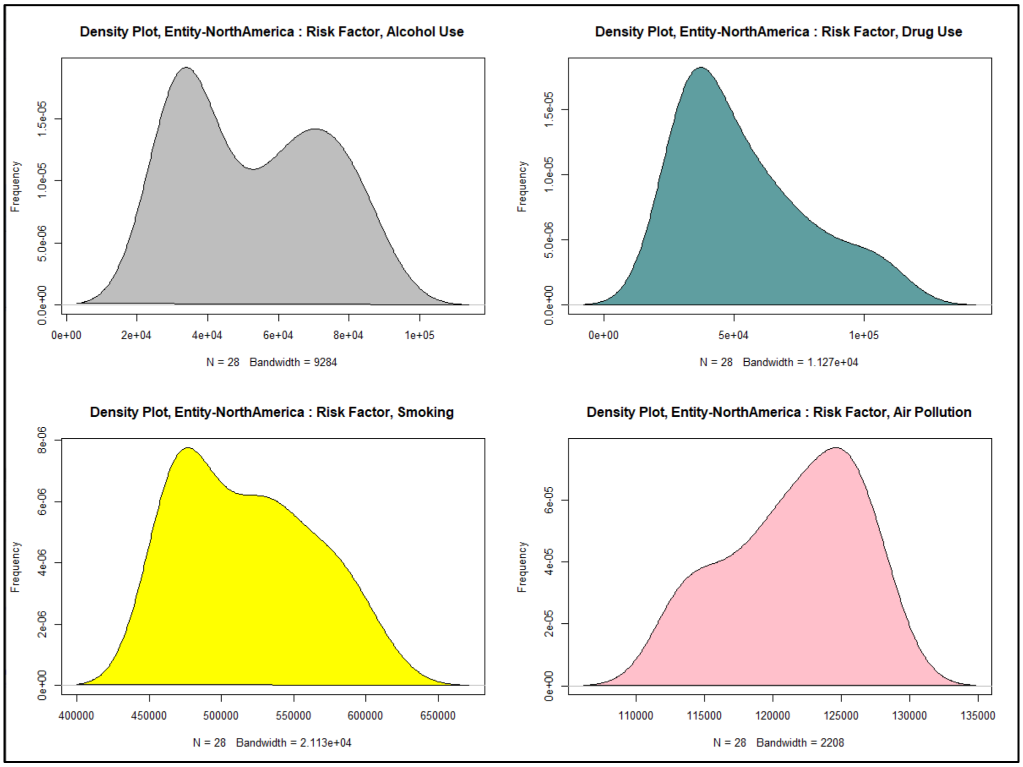
c. Graph 3: Boxplot of the attributes for Entity Maldives



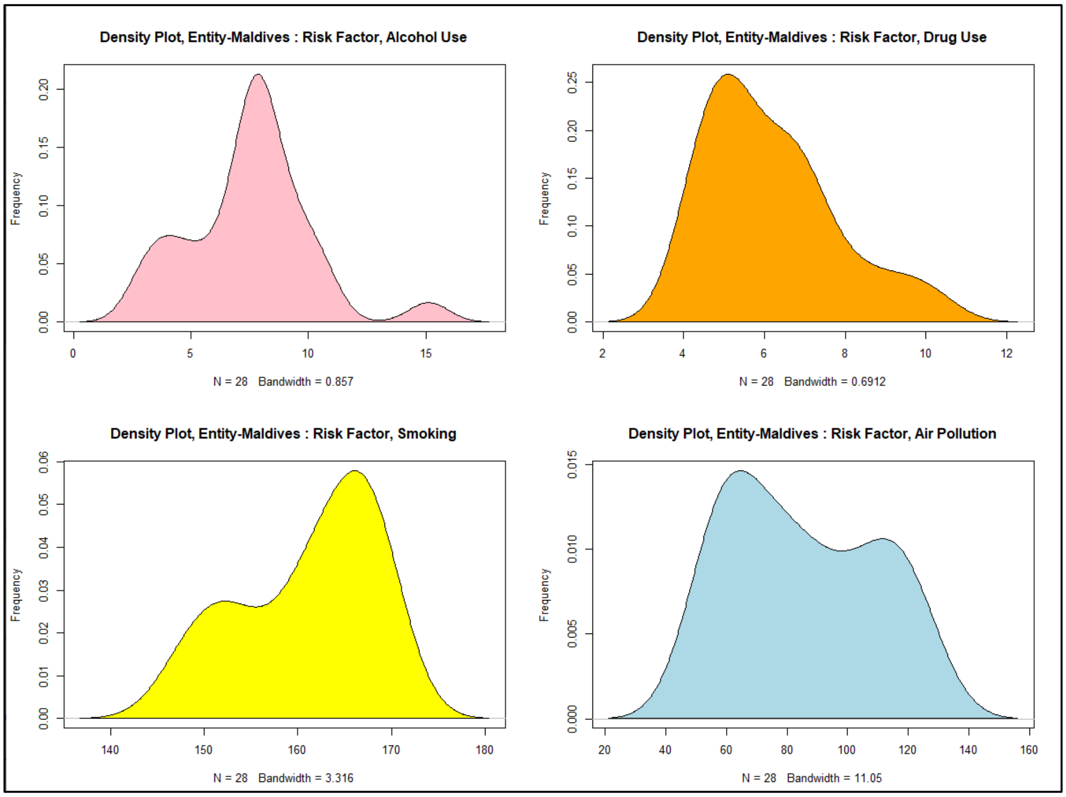
d. Graph 4: Density plot of the attributes for Entity India



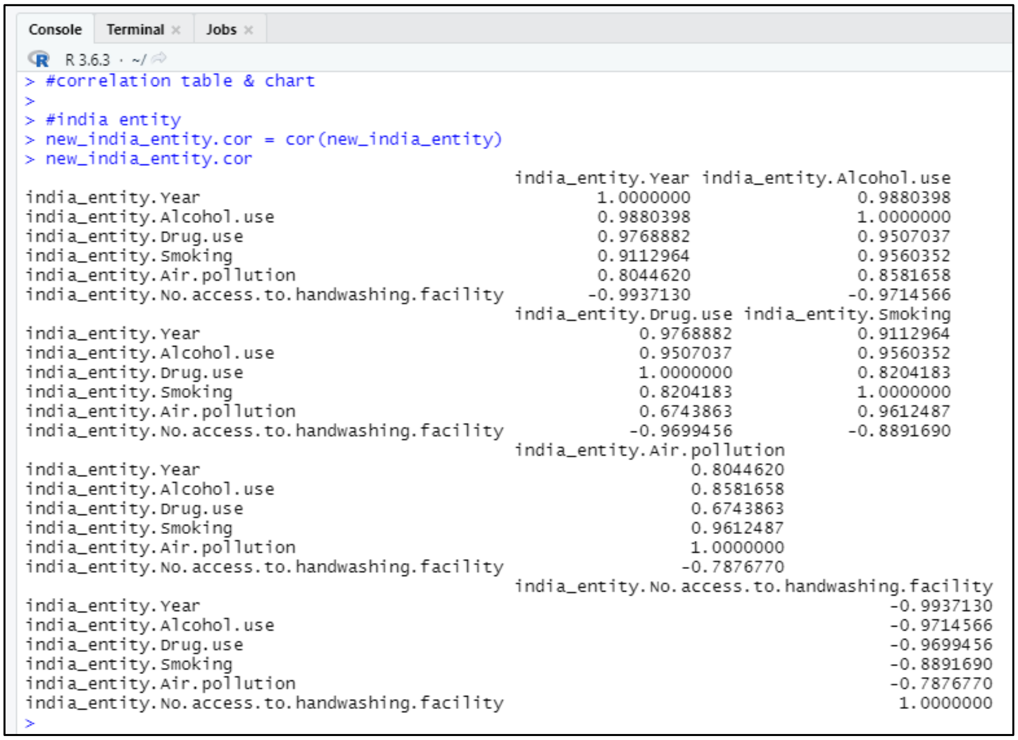
e. Graph 5: Density plot of the attributes for Entity North America



f. Graph 6: Density plot of the attributes for Entity Maldives

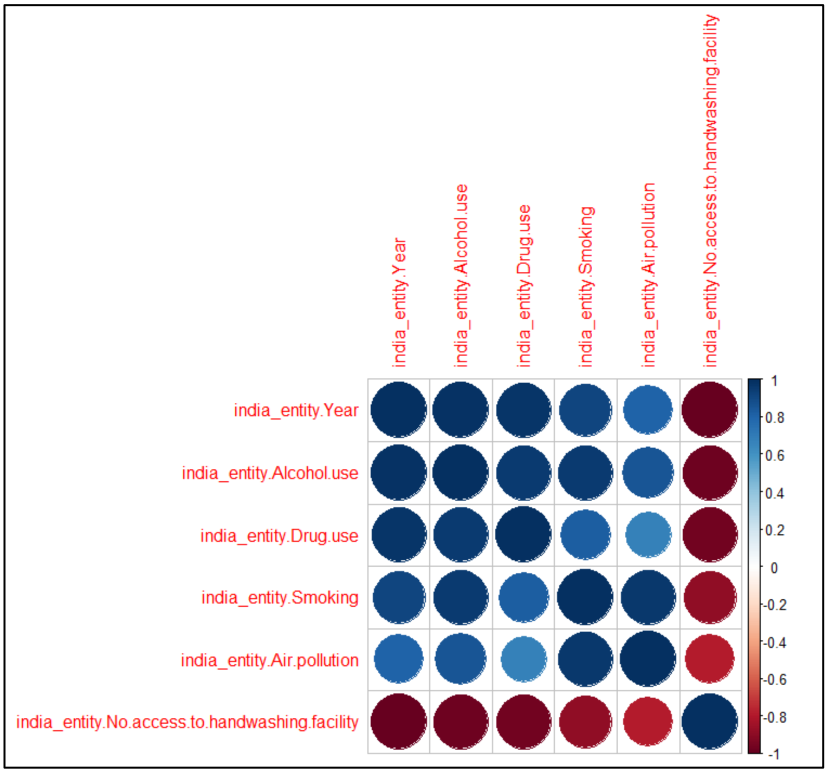


g. Correlation Table: Entity - India

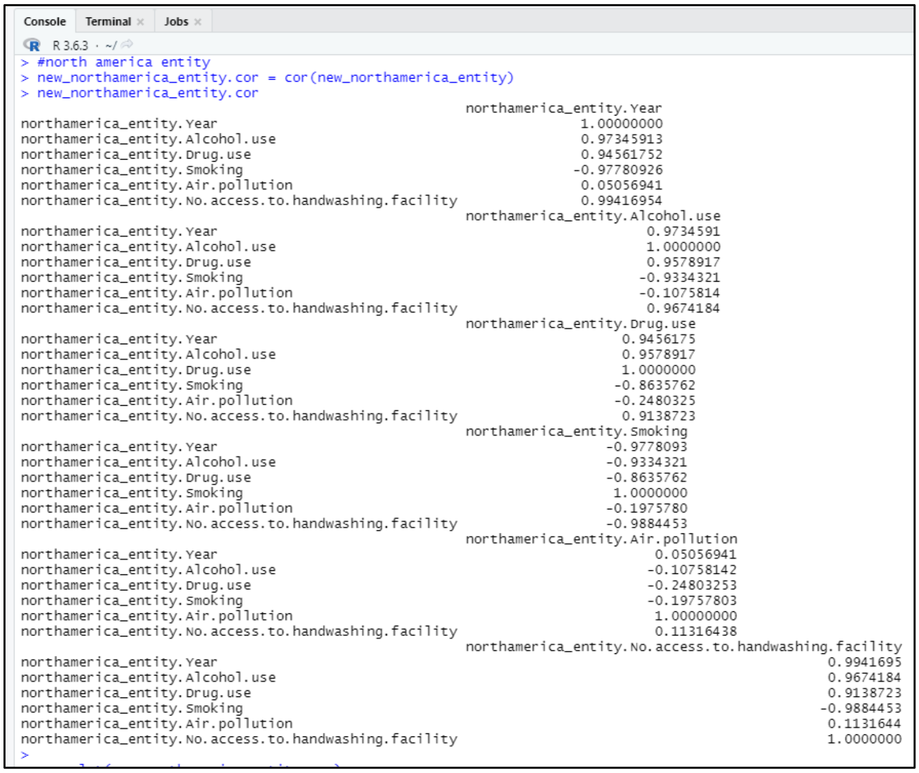


```
> #correlation table & chart
>
> #india entity
> new_india_entity.cor = cor(new_india_entity)
> new_india_entity.cor
                                         india_entity.Year india_entity.Alcohol.use
india_entity.Year                                1.0000000                0.9880398
india_entity.Alcohol.use                         0.9880398                1.0000000
india_entity.Drug.use                            0.9768882                0.9507037
india_entity.Smoking                             0.9112964                0.9560352
india_entity.Air.pollution                       0.8044620                0.8581658
india_entity.No.access.to.handwashing.facility  -0.9937130               -0.9714566
                                         india_entity.Drug.use india_entity.Smoking
india_entity.Year                                    0.9768882            0.9112964
india_entity.Alcohol.use                             0.9507037            0.9560352
india_entity.Drug.use                                1.0000000            0.8204183
india_entity.Smoking                                 0.8204183            1.0000000
india_entity.Air.pollution                           0.6743863            0.9612487
india_entity.No.access.to.handwashing.facility      -0.9699456           -0.8891690
                                         india_entity.Air.pollution
india_entity.Year                                         0.8044620
india_entity.Alcohol.use                                  0.8581658
india_entity.Drug.use                                     0.6743863
india_entity.Smoking                                      0.9612487
india_entity.Air.pollution                                1.0000000
india_entity.No.access.to.handwashing.facility           -0.7876770
                                         india_entity.No.access.to.handwashing.facility
india_entity.Year                                                            -0.9937130
india_entity.Alcohol.use                                                     -0.9714566
india_entity.Drug.use                                                        -0.9699456
india_entity.Smoking                                                         -0.8891690
india_entity.Air.pollution                                                   -0.7876770
india_entity.No.access.to.handwashing.facility                                1.0000000
>
```

h. Correlation Chart: Entity – India

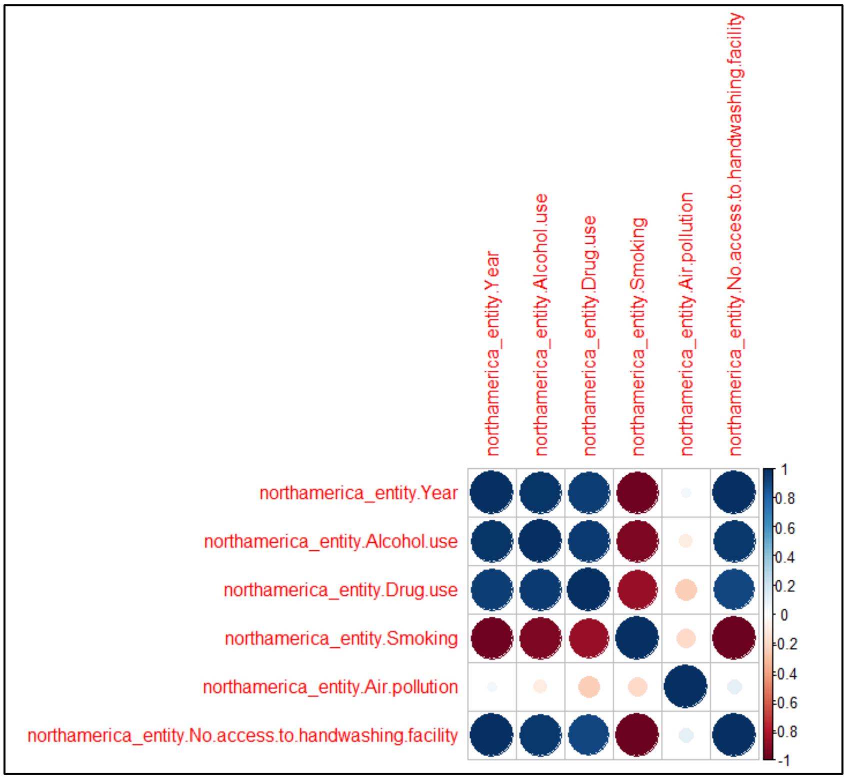


i. Correlation Table: Entity – North America

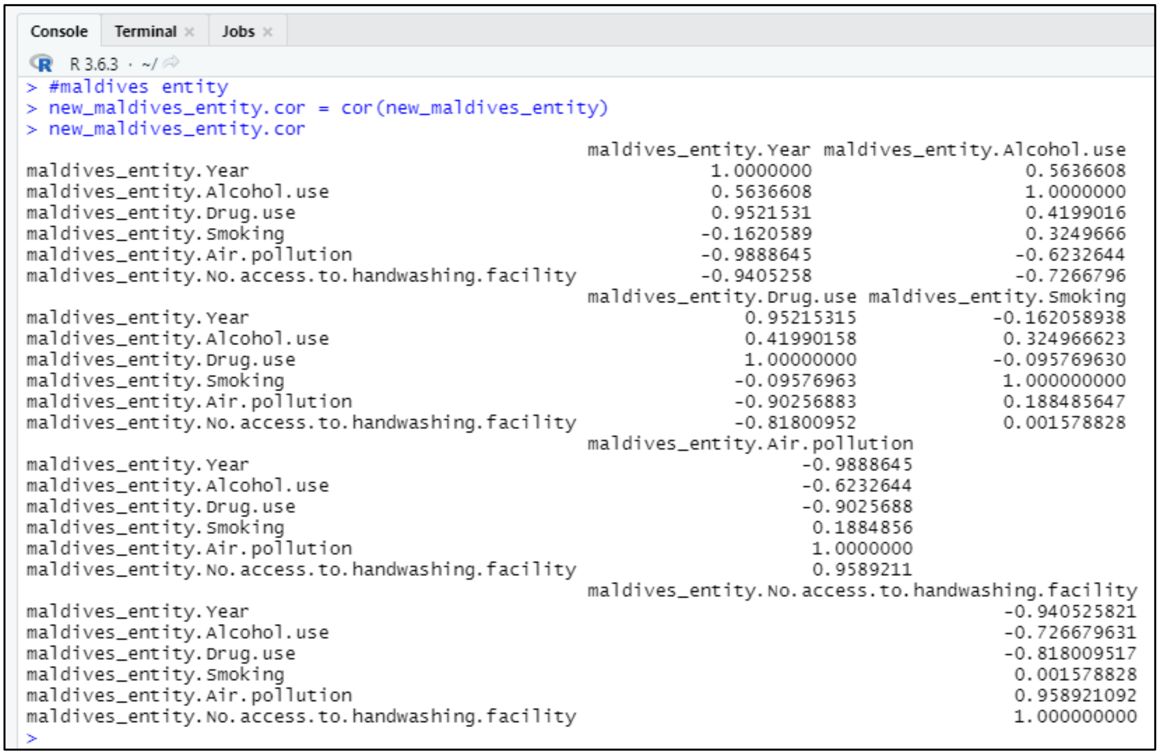


```
> #north america entity
> new_northamerica_entity.cor = cor(new_northamerica_entity)
> new_northamerica_entity.cor
                                                      northamerica_entity.Year
northamerica_entity.Year                                            1.00000000
northamerica_entity.Alcohol.use                                     0.97345913
northamerica_entity.Drug.use                                        0.94561752
northamerica_entity.Smoking                                        -0.97780926
northamerica_entity.Air.pollution                                   0.05056941
northamerica_entity.No.access.to.handwashing.facility               0.99416954
                                                      northamerica_entity.Alcohol.use
northamerica_entity.Year                                                    0.9734591
northamerica_entity.Alcohol.use                                             1.0000000
northamerica_entity.Drug.use                                                0.9578917
northamerica_entity.Smoking                                                -0.9334321
northamerica_entity.Air.pollution                                          -0.1075814
northamerica_entity.No.access.to.handwashing.facility                       0.9674184
                                                      northamerica_entity.Drug.use
northamerica_entity.Year                                                 0.9456175
northamerica_entity.Alcohol.use                                          0.9578917
northamerica_entity.Drug.use                                             1.0000000
northamerica_entity.Smoking                                             -0.8635762
northamerica_entity.Air.pollution                                       -0.2480325
northamerica_entity.No.access.to.handwashing.facility                    0.9138723
                                                      northamerica_entity.Smoking
northamerica_entity.Year                                              -0.9778093
northamerica_entity.Alcohol.use                                       -0.9334321
northamerica_entity.Drug.use                                          -0.8635762
northamerica_entity.Smoking                                            1.0000000
northamerica_entity.Air.pollution                                     -0.1975780
northamerica_entity.No.access.to.handwashing.facility                 -0.9884453
                                                      northamerica_entity.Air.pollution
northamerica_entity.Year                                                     0.05056941
northamerica_entity.Alcohol.use                                             -0.10758142
northamerica_entity.Drug.use                                                -0.24803253
northamerica_entity.Smoking                                                 -0.19757803
northamerica_entity.Air.pollution                                            1.00000000
northamerica_entity.No.access.to.handwashing.facility                        0.11316438
                                                      northamerica_entity.No.access.to.handwashing.facility
northamerica_entity.Year                                                                              0.9941695
northamerica_entity.Alcohol.use                                                                       0.9674184
northamerica_entity.Drug.use                                                                          0.9138723
northamerica_entity.Smoking                                                                          -0.9884453
northamerica_entity.Air.pollution                                                                     0.1131644
northamerica_entity.No.access.to.handwashing.facility                                                 1.0000000
>
```

j. Correlation Chart: Entity – North America

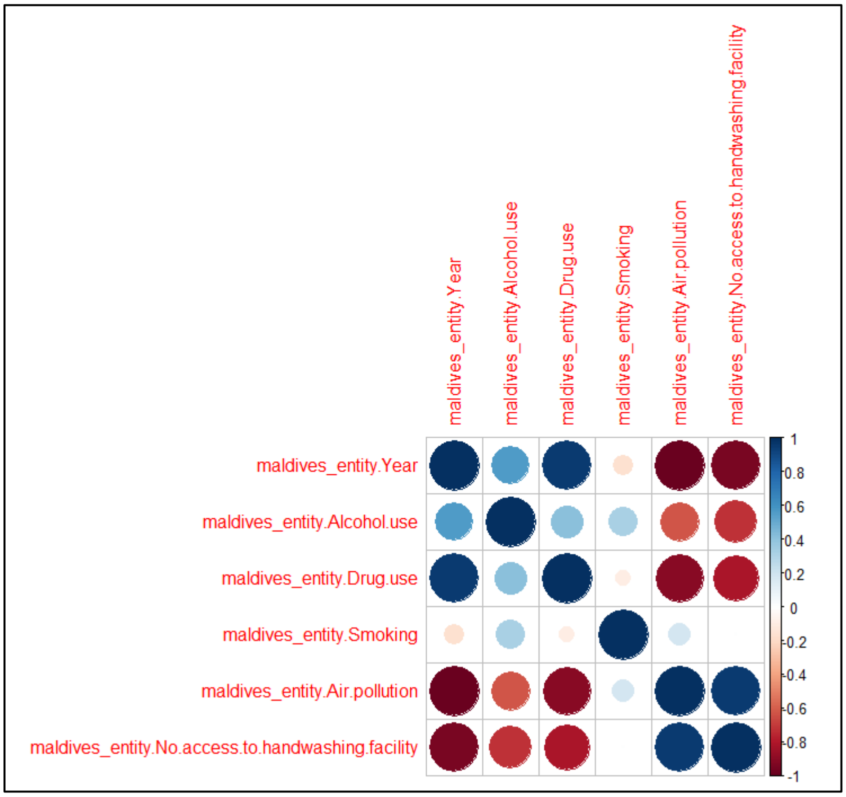


k. Correlation Table: Entity – Maldives



```
> #maldives entity
> new_maldives_entity.cor = cor(new_maldives_entity)
> new_maldives_entity.cor
                                                  maldives_entity.Year maldives_entity.Alcohol.use
maldives_entity.Year                                         1.0000000                   0.5636608
maldives_entity.Alcohol.use                                  0.5636608                   1.0000000
maldives_entity.Drug.use                                     0.9521531                   0.4199016
maldives_entity.Smoking                                     -0.1620589                   0.3249666
maldives_entity.Air.pollution                               -0.9888645                  -0.6232644
maldives_entity.No.access.to.handwashing.facility           -0.9405258                  -0.7266796
                                                  maldives_entity.Drug.use maldives_entity.Smoking
maldives_entity.Year                                            0.95215315            -0.162058938
maldives_entity.Alcohol.use                                     0.41990158             0.324966623
maldives_entity.Drug.use                                        1.00000000            -0.095769630
maldives_entity.Smoking                                        -0.09576963             1.000000000
maldives_entity.Air.pollution                                  -0.90256883             0.188485647
maldives_entity.No.access.to.handwashing.facility              -0.81800952             0.001578828
                                                  maldives_entity.Air.pollution
maldives_entity.Year                                                 -0.9888645
maldives_entity.Alcohol.use                                          -0.6232644
maldives_entity.Drug.use                                             -0.9025688
maldives_entity.Smoking                                               0.1884856
maldives_entity.Air.pollution                                         1.0000000
maldives_entity.No.access.to.handwashing.facility                     0.9589211
                                                  maldives_entity.No.access.to.handwashing.facility
maldives_entity.Year                                                                   -0.940525821
maldives_entity.Alcohol.use                                                            -0.726679631
maldives_entity.Drug.use                                                               -0.818009517
maldives_entity.Smoking                                                                 0.001578828
maldives_entity.Air.pollution                                                           0.958921092
maldives_entity.No.access.to.handwashing.facility                                       1.000000000
>
```

l. Correlation Chart: Entity – Maldives


maldives_entity.Year
maldives_entity.Alcohol.use
maldives_entity.Drug.use
maldives_entity.Smoking
maldives_entity.Air.pollution
maldives_entity.No.access.to.handwashing.facility
maldives_entity.Year
maldives_entity.Alcohol.use
maldives_entity.Drug.use
maldives_entity.Smoking
maldives_entity.Air.pollution
maldives_entity.No.access.to.handwashing.facility