Richa Rambhia
November 27, 2021

Final Project – Milestone 2
Hypothesis Testing

Introduction

Dataset:

The dataset chosen for the Final Project – Milestone 2 was the same dataset used in the Final Project – Milestone 1 which was picked from Kaggle. The dataset is related to the number of deaths by risk factor, and it 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.

In this assignment, the task is to answer some of the questions related to the data which will be done through inferential statistics and hypothesis testing. From the dataset the important parameters to work with and which will be helpful in answering the questions are the entity parameter which is nothing but the countries having various risk factors affecting the death rate, the year in which the risk factor affected the death rate, and all the various risk factors like the alcohol use, drug use, smoking, iron deficiency, diet low in fruits, diet low in vegetables, and many other factors.

Questions about the Data:

The dataset of number of deaths by risk factors gives us a lot of information related to the death rate taking place due to the risk factors in a particular country in the year. The questions that can be answered with respect to the dataset could be as follows.

a. Whether the overall mean of a particular risk factor is equal to assumed value or not which can be answered and performed through one-sample t-test.
b. Whether the mean of a particular risk factor in the entity i.e., the specific country is equal to assumed value or not which again can be solved through the one-sample t-test.
c. Which risk factor is higher in all the entities affecting the death rate more and that can be answered through the two-sample t-test?
d. Lastly, which country is affected more due to a particular risk factor as compared to another country? The two-sample t-test will again be helpful here in answering this question.

The data analysis and hypothesis testing of this dataset will give us more insights and help us in determining the various questions.

Data Analysis

After installing and importing all the libraries needed, the number of deaths by risk factor dataset was loaded into a variable and read. In order to know more about the dataset for further analysis, descriptive analysis was done which would give us information about the entire dataset, its attributes, and the various other parameters of the dataset. Descriptive analysis is one of the processes in the exploratory data analysis which helps in describing the data and computing the statistical values of the numeric variables of the dataset. Here, after reading the csv file, the data was described which includes the column names of the dataset that would help in knowing about the variables which would help us in analysis and testing. The starting and end records were displayed along with the summary of the entire dataset. The summary of the dataset helps in determining the statistical values of the variables which are numeric values.

Output:

1. Reading the csv file

```
Console  Terminal ×  Jobs ×
R 3.6.3 · ~/
> #reading csv file
> deathrate_dataset <- read.csv("number-of-deaths-by-risk-factor.csv")
> deathrate_dataset
        Entity Code Year Unsafe.water.source Unsafe.sanitation No.access.to.handwashing.facility
1  Afghanistan  AFG 1990          7554.04954        5887.74763                        5412.31451
2  Afghanistan  AFG 1991          7359.67675        5732.77016                        5287.89110
3  Afghanistan  AFG 1992          7650.43782        5954.80499                        5506.65736
4  Afghanistan  AFG 1993         10270.73138        7986.73661                        7104.62035
5  Afghanistan  AFG 1994         11409.17711        8863.01006                        8051.51595
6  Afghanistan  AFG 1995         12676.64742        9840.84868                        8770.68629
7  Afghanistan  AFG 1996         12154.94232        9426.89645                        8610.68686
8  Afghanistan  AFG 1997         12329.13171        9553.55600                        8722.94308
9  Afghanistan  AFG 1998         12133.60954        9390.04203                        8621.88492
10 Afghanistan  AFG 1999         11990.39633        9268.48961                        8502.72999
11 Afghanistan  AFG 2000         11121.65328        8588.51714                        7910.36627
12 Afghanistan  AFG 2001         10431.36444        8046.88412                        7501.54813
13 Afghanistan  AFG 2002          9764.76876        7522.61784                        7055.32209
14 Afghanistan  AFG 2003         11510.49271        8853.83887                        8214.45999
15 Afghanistan  AFG 2004         11656.61466        8950.64281                        8331.27082
16 Afghanistan  AFG 2005         10973.73633        8410.94881                        7928.46372
17 Afghanistan  AFG 2006         10040.77296        7673.53095                        7393.87332
18 Afghanistan  AFG 2007          8884.66772        6776.14823                        6631.32202
19 Afghanistan  AFG 2008          7703.60948        5857.35361                        5854.38195
20 Afghanistan  AFG 2009          6842.30998        5177.56491                        5279.97858
21 Afghanistan  AFG 2010          6436.21122        4843.77111                        4992.48476
22 Afghanistan  AFG 2011          6023.62912        4507.04169                        4725.82154
23 Afghanistan  AFG 2012          5613.05315        4170.82802                        4449.08079
24 Afghanistan  AFG 2013          5649.75574        4166.68608                        4474.55745
25 Afghanistan  AFG 2014          5533.13574        4048.98644                        4391.70827
26 Afghanistan  AFG 2015          5300.42444        3843.62273                        4231.83572
27 Afghanistan  AFG 2016          5158.40086        3723.50672                        4126.33647
28 Afghanistan  AFG 2017          5256.64928        3783.11112                        4156.20901
29     Albania  ALB 1990            50.12577          30.95006                          42.27688
30     Albania  ALB 1991            51.04275          31.82805                          44.27507
31     Albania  ALB 1992            48.82450          30.76490                          43.19028
   Household.air.pollution.from.solid.fuels Non.exclusive.breastfeeding Discontinued.breastfeeding Child.wasting
1                                 22388.497                   3221.1388                 156.097553    22778.849
2                                 22128.758                   3150.5596                 151.539851    22292.691
3                                 22873.769                   3331.3490                 156.609194    23102.198
4                                 25599.756                   4477.0061                 206.834451    27902.670
5                                 28013.167                   5102.6221                 233.930571    32929.006
6                                 29062.619                   5402.6604                 262.793340    35631.997
7                                 29407.322                   5263.6445                 253.668181    36114.593
8                                 29674.398                   5271.7718                 258.134146    36749.117
9                                 29807.453                   5165.9238                 254.708078    36569.466
10                                29484.612                   5044.3076                 251.898824    36124.047
11                                28329.380                   4566.9186                 235.101472    34320.365
12                                27721.641                   4219.4200                 220.529327    32430.507
13                                26679.508                   3852.4173                 207.900393    30516.319
14                                28325.887                   4184.9816                 254.024088    30053.871
15                                28511.343                   4067.0871                 262.000727    29568.254
16                                27621.503                   3761.3221                 245.140693    27799.719
17                                26765.727                   3549.8691                 216.945204    26233.636
```

2. Descriptive Analysis

```
Console   Terminal ×   Jobs ×
R 3.6.3 · ~/
> 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"
>
> start_records <- head(deathrate_dataset,10)
> start_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 Child.wasting
1                                  22388.50                    3221.139                   156.0976      22778.85
2                                  22128.76                    3150.560                   151.5399      22292.69
3                                  22873.77                    3331.349                   156.6092      23102.20
4                                  25599.76                    4477.006                   206.8345      27902.67
5                                  28013.17                    5102.622                   233.9306      32929.01
6                                  29062.62                    5402.660                   262.7933      35632.00
7                                  29407.32                    5263.644                   253.6682      36114.59
8                                  29674.40                    5271.772                   258.1341      36749.12
9                                  29807.45                    5165.924                   254.7081      36569.47
10                                 29484.61                    5044.308                   251.8988      36124.05
   Child.stunting Low.birth.weight.for.gestation Secondhand.smoke Alcohol.use Drug.use Diet.low.in.fruits
1        10408.44                       12168.56         4234.808    356.5293 208.3254           8538.964
2        10271.98                       12360.64         4219.597    320.5985 217.7697           8642.847
3        10618.88                       13459.59         4371.908    293.2570 247.8333           8961.526
4        12260.09                       18458.43         4863.559    278.1298 285.0362           9377.118
5        14197.95                       19958.39         5292.380    250.6916 306.6468           9688.449
6        15243.02                       20444.71         5491.018    220.1991 324.6103           9876.717
7        16009.92                       21072.04         5595.951    197.9284 342.8512          10102.415
8        16473.90                       21262.69         5701.812    181.6741 361.9349          10330.181
9        16665.71                       21214.16         5762.015    203.5203 379.1053          10522.664
10       16729.54                       20972.04         5774.820    170.0059 388.7336          10639.183
   Diet.low.in.vegetables Unsafe.sex Low.physical.activity High.fasting.plasma.glucose High.total.cholesterol
1                7678.718   387.1676              4221.303                    21610.07               9505.532
2                7789.773   394.4483              4252.630                    21824.94                     NA
3                8083.235   422.4533              4347.331                    22418.70                     NA
```

```
Console   Terminal ×   Jobs ×
R 3.6.3 · ~/
> summary(deathrate_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 Secondhand.smoke
 Min.   :    0.00           Min.   :      0   Min.   :      0.0   Min.   :      0.3              Min.   :      2.9
 1st Qu.:    0.26           1st Qu.:     41   1st Qu.:      1.9   1st Qu.:    144.6              1st Qu.:    278.1
 Median :    6.62           Median :    730   Median :     77.9   Median :   1220.7              Median :   1196.2
 Mean   :  409.11           Mean   :  43446   Mean   :  11767.7   Mean   :  30948.0              Mean   :  24282.3
 3rd Qu.:   78.28           3rd Qu.:  10235   3rd Qu.:   1971.6   3rd Qu.:   8708.1              3rd Qu.:   5963.7
 Max.   :34850.40           Max.   :3365309   Max.   :1001277.4   Max.   :1976612.5              Max.   :1260994.2

  Alcohol.use        Drug.use          Diet.low.in.fruits Diet.low.in.vegetables   Unsafe.sex
 Min.   :  -2315   Min.   :     1.2   Min.   :      1.6   Min.   :      0.8      Min.   :      1.0
 1st Qu.:    364   1st Qu.:    92.9   1st Qu.:    536.0   1st Qu.:    413.0      1st Qu.:    136.1
 Median :   2803   Median :   408.6   Median :   2452.9   Median :   1837.8      Median :    831.8
 Mean   :  50203   Mean   :  8890.2   Mean   :  45452.6   Mean   :  28742.0      Mean   :  26764.5
 3rd Qu.:  12891   3rd Qu.:  2170.8   3rd Qu.:  10521.8   3rd Qu.:   7612.3      3rd Qu.:   5949.0
 Max.   :2842854   Max.   :585348.2   Max.   :2423447.4   Max.   :1462367.4      Max.   :1771140.7

 Low.physical.activity High.fasting.plasma.glucose High.total.cholesterol High.body.mass.index
 Min.   :      2.4     Min.   :     21             Min.   :     10        Min.   :     20
 1st Qu.:    261.6     1st Qu.:   2035             1st Qu.:    839        1st Qu.:   1141
 Median :   1189.4     Median :   7820             Median :   4005        Median :   4740
 Mean   :  21141.5     Mean   :  99556             Mean   :  51628        Mean   :  68685
 3rd Qu.:   5694.7     3rd Qu.:  34705             3rd Qu.:  17423        3rd Qu.:  21601
 Max.   :1263051.3     Max.   :6526028             Max.   :4392505        Max.   :4724346
                                                   NA's   :4907
 High.systolic.blood.pressure    Smoking           Iron.deficiency     Vitamin.A.deficiency Low.bone.mineral.density
 Min.   :      21              Min.   :     12   Min.   :     0.01   Min.   :     0.0     Min.   :     0.4
 1st Qu.:    2665              1st Qu.:   1293   1st Qu.:     2.26   1st Qu.:     1.9     1st Qu.:    40.6
 Median :   10993              Median :   5936   Median :    31.99   Median :    70.5     Median :   246.8
 Mean   :  174383              Mean   : 133548   Mean   :  1878.75   Mean   : 11908.6     Mean   :  4579.1
 3rd Qu.:   47323              3rd Qu.:  31638   3rd Qu.:   421.38   3rd Qu.:  2081.9     3rd Qu.:  1096.1
 Max.   :10440818              Max.   :7099111   Max.   :125242.95   Max.   :986995.0     Max.   :327314.3

 Air.pollution     Outdoor.air.pollution Diet.high.in.sodium Diet.low.in.whole.grains Diet.low.in.nuts.and.seeds
 Min.   :      9   Min.   :      5       Min.   :      3     Min.   :      9.3        Min.   :      5.2
 1st Qu.:   1077   1st Qu.:    554       1st Qu.:    356     1st Qu.:    798.7        1st Qu.:    553.3
 Median :   6125   Median :   2242       Median :   1946     Median :   3504.3        Median :   2279.2
 Mean   :  95736   Mean   :  55573       Mean   :  54241     Mean   :  53348.8        Mean   :  34967.0
```

Now, after the describing about the entire dataset, a subset of the dataset was created which had a randomly selected sample of dataset. This is done because the dataset has 6000+ entries and so rather than testing on the entire dataset, a sample of the dataset was created on which the testing can be performed. The sample_n() function was used which randomly selects some amount of data from the original dataset. Here, from the 6000+ entries, 1500 sample of data was extracted from the dataset. Along with that, the summary of this randomly selected dataset was also displayed as this was the dataset which would be used further.

Output:

```
Console  Terminal  Jobs
R 3.6.3 · ~/
> #creating a randomly selected sample data set
> randomly_selected_dataset <- sample_n(deathrate_dataset,1500)
> randomly_selected_dataset
                                       Entity Code Year Unsafe.water.source Unsafe.sanitation
1                                       Ghana  GHA 2004        6.774547e+03      5.087796e+03
2                                     Belarus  BLR 2011        4.932151e+00      4.076170e+00
3                                      Norway  NOR 1996        2.840490e+00      2.540556e+00
4                                  Kazakhstan  KAZ 2007        8.436492e+01      8.425726e+01
5                                 El Salvador  SLV 2011        2.184164e+02      1.604209e+02
6  Central Europe, Eastern Europe, and Central Asia 2011        2.048725e+03      1.750887e+03
7                               Low-middle SDI      2000        7.175085e+05      5.452516e+05
8                       Southern Latin America      1990        1.082295e+03      7.199234e+02
9                       Southern Latin America      2005        4.160199e+02      1.850361e+02
10                            Solomon Islands  SLB 1999        5.139498e+01      4.151498e+01
11                                     Malawi  MWI 1991        1.752858e+04      1.361341e+04
12                                Saint Lucia  LCA 1990        6.927555e+00      4.993922e+00
13                                   Dominica  DMA 1992        2.760149e+00      1.887322e+00
14                                       Iran  IRN 2002        7.085992e+02      3.743948e+02
15                                   Slovenia  SVN 2000        1.111840e+00      4.126796e-01
16                                 South Asia      1992        1.052140e+06      8.164646e+05
17                                     Canada  CAN 1990        6.102564e+00      5.800429e+00
18                                     Panama  PAN 2009        1.370335e+02      9.281407e+01
19                                  Caribbean      1998        7.578784e+03      5.512117e+03
20                                   Ethiopia  ETH 2014        4.732326e+04      3.822108e+04
21                         Sub-Saharan Africa      2004        5.939449e+05      4.512267e+05
22                                    Jamaica  JAM 1993        1.435510e+02      9.928261e+01
23                                 Uzbekistan  UZB 1999        5.777967e+02      6.362908e+02
24                   High-income Asia Pacific      2008        4.097661e+02      1.076408e+02
25                                    Croatia  HRV 2011        6.513234e+00      1.277415e+00
26                                       Togo  TGO 1990        4.626621e+03      3.557044e+03
27                                    Ecuador  ECU 2015        1.899780e+02      7.352081e+01
28                                     Norway  NOR 2012        8.013725e+00      5.687740e+00
29                                 Kazakhstan  KAZ 2001        2.153017e+02      1.866798e+02
30                                   Djibouti  DJI 2007        2.891552e+02      2.185937e+02
31                                    Belgium  BEL 2012        2.046808e+01      1.430574e+01
   No.access.to.handwashing.facility Household.air.pollution.from.solid.fuels Non.exclusive.breastfeeding
1                       5.767985e+03                              10979.21135                8.733465e+02
2                       1.785100e+01                                195.64280                6.421849e+00
3                       1.482855e+01                                 81.26595                5.303032e-01
4                       1.388347e+02                               3873.01086                1.732305e+02
5                       1.662218e+02                                991.59127                3.748717e+01
6                       2.411897e+03                              41754.36523                2.853640e+03
7                       4.131042e+05                             750161.63293                1.391549e+05
8                       8.363328e+02                               6478.10531                4.759511e+02
9                       6.302267e+02                               3753.41566                1.446759e+02
10                      3.049067e+01                                394.77380                7.012577e+00
11                      1.112276e+04                              10141.26477                3.941146e+03
12                      3.560395e+00                                 40.97462                1.533856e+00
13                      2.691443e+00                                 25.11541                5.113791e-01
14                      4.115367e+02                                755.72755                4.872197e+02
15                      4.198093e+00                                172.17948                4.998021e-01
```

```
Console  Terminal  Jobs
R 3.6.3 · ~/
> summary(randomly_selected_dataset)
                           Entity           Code           Year       Unsafe.water.source Unsafe.sanitation
 Congo                        :  14           : 225   Min.   :1990   Min.   :      0.0   Min.   :      0.0
 Kazakhstan                   :  13    COG    :  14   1st Qu.:1996   1st Qu.:     10.5   1st Qu.:      4.6
 Nigeria                      :  13    KAZ    :  13   Median :2004   Median :    311.1   Median :    188.8
 Cape Verde                   :  12    NGA    :  13   Mean   :2004   Mean   :  36558.3   Mean   :  27081.1
 Somalia                      :  12    CPV    :  12   3rd Qu.:2011   3rd Qu.:   5171.2   3rd Qu.:   3739.1
 Democratic Republic of Congo:  11    SOM    :  12   Max.   :2017   Max.   :2095066.5   Max.   :1622958.9
 (Other)                      :1425   (Other):1211
 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.:     19.2                 1st Qu.:    101.1                        1st Qu.:     4.8
 Median :    262.4                 Median :   1251.6                        Median :   127.7
 Mean   :  22237.0                 Mean   :  47294.1                        Mean   :  7148.4
 3rd Qu.:   3809.5                 3rd Qu.:   9576.9                        3rd Qu.:  1292.5
 Max.   :1230318.6                 Max.   :2677805.7                        Max.   :502180.0

 Discontinued.breastfeeding Child.wasting     Child.stunting     Low.birth.weight.for.gestation Secondhand.smoke
 Min.   :    0.00          Min.   :      0   Min.   :     0.0   Min.   :      0.4               Min.   :      3.0
 1st Qu.:    0.29          1st Qu.:     46   1st Qu.:     2.0   1st Qu.:    159.9               1st Qu.:    327.6
 Median :    8.43          Median :    956   Median :    97.4   Median :   1351.3               Median :   1289.6
 Mean   :  480.65          Mean   :  50016   Mean   : 13689.5   Mean   :  34616.6               Mean   :  27209.7
 3rd Qu.:   80.74          3rd Qu.:  10262   3rd Qu.:  2047.8   3rd Qu.:   8793.8               3rd Qu.:   6255.4
 Max.   :33853.62          Max.   :3296648   Max.   :975802.7   Max.   :1960013.1               Max.   :1253577.0

  Alcohol.use          Drug.use         Diet.low.in.fruits  Diet.low.in.vegetables   Unsafe.sex
 Min.   :   -789.5   Min.   :     1.4   Min.   :      1.9   Min.   :      1.3      Min.   :      1.4
 1st Qu.:    472.8   1st Qu.:   110.1   1st Qu.:    621.1   1st Qu.:    473.8      1st Qu.:    152.3
 Median :   3165.2   Median :   435.8   Median :   2587.9   Median :   1932.6      Median :    889.2
 Mean   :  58064.6   Mean   : 10118.6   Mean   :  51964.4   Mean   :  32135.3      Mean   :  29658.8
 3rd Qu.:  13745.3   3rd Qu.:  2324.9   3rd Qu.:  11560.1   3rd Qu.:   8172.8      3rd Qu.:   6344.5
 Max.   :2817609.6   Max.   :572923.0   Max.   :2420710.2   Max.   :1459655.9      Max.   :1771140.7

 Low.physical.activity High.fasting.plasma.glucose High.total.cholesterol High.body.mass.index
 Min.   :      2.6    Min.   :     23              Min.   :     19        Min.   :     21
 1st Qu.:    300.5    1st Qu.:   2344              1st Qu.:    834        1st Qu.:   1346
 Median :   1282.8    Median :   8052              Median :   3587        Median :   4791
 Mean   :  24482.4    Mean   : 113066              Mean   :  61757        Mean   :  79754
 3rd Qu.:   5992.5    3rd Qu.:  35358              3rd Qu.:  15192        3rd Qu.:  24128
 Max.   :1242710.4    Max.   :6399257              Max.   :4392505        Max.   :4614666
                                                   NA's   :1129
 High.systolic.blood.pressure    Smoking          Iron.deficiency      Vitamin.A.deficiency Low.bone.mineral.density
 Min.   :      23              Min.   :     12   Min.   :     0.01   Min.   :     0.0   Min.   :     0.4
 1st Qu.:    3157              1st Qu.:   1448   1st Qu.:     2.39   1st Qu.:     2.0   1st Qu.:    48.1
 Median :   11466              Median :   6285   Median :    39.75   Median :    72.9   Median :   253.8
 Mean   :  200920              Mean   : 151515   Mean   :  2117.62   Mean   : 13788.5   Mean   :  5220.2
 3rd Qu.:   51580              3rd Qu.:  30897   3rd Qu.:   406.50   3rd Qu.:  2143.6   3rd Qu.:  1135.0
 Max.   :10293778              Max.   :7046703   Max.   :123752.23   Max.   :971721.1   Max.   :320874.9

 Air.pollution     Outdoor.air.pollution Diet.high.in.sodium Diet.low.in.whole.grains Diet.low.in.nuts.and.seeds
 Min.   :      9   Min.   :      5      Min.   :      6     Min.   :      9.4      Min.   :      5.3
 1st Qu.:   1205   1st Qu.:    661      1st Qu.:    415     1st Qu.:    913.0      1st Qu.:    622.4
 Median :   6400   Median :   2472      Median :   2030     Median :   3625.7      Median :   2425.6
 Mean   : 108381   Mean   :  64454      Mean   :  63374     Mean   :  61604.8      Mean   :  40172.2
 3rd Qu.:  23919   3rd Qu.:  14060      3rd Qu.:  10377     3rd Qu.:  16032.8      3rd Qu.:  10955.4
 Max.   :4875231   Max.   :3318503      Max.   :3163601     Max.   :3039445.4      Max.   :2042618.7
```

For the testing and analysis, the variables to be tested needed to be extracted for which new variables were created. Also, since the analysis needed to be done for a particular country as well, a subset of dataset was created using the subset function which extracted all the values for the particular country selected. The frequency table for the entity parameter was also displayed which helped in knowing about the count of the particular entity.

Output:

1. Creating new variables

```
Console  Terminal ×  Jobs ×
R 3.6.3 · ~/
> air_pollution <- randomly_selected_dataset$Air.pollution
> air_pollution
  [1] 1.366547e+04 1.048875e+04 1.782017e+03 1.076745e+04 2.646657e+03 3.351480e+05 1.143126e+06 2.106755e+04
  [9] 2.310631e+04 4.309076e+02 1.101027e+04 6.968843e+01 4.216256e+01 2.022309e+04 1.125406e+03 1.305939e+06
 [17] 7.914285e+03 7.835422e+02 2.091592e+04 4.257607e+04 5.980784e+05 9.347711e+02 1.762088e+04 6.488292e+04
 [25] 3.433451e+03 3.269600e+03 3.678523e+03 1.340432e+03 1.141125e+04 3.642237e+02 5.392358e+03 1.362682e+04
 [33] 8.433248e+03 7.144399e+04 9.235907e+05 4.011379e+05 4.372973e+04 1.474935e+04 2.365816e+03 4.426324e+03
 [41] 1.749689e+04 8.855626e+02 1.858695e+03 7.220803e+03 1.618441e+06 2.337884e+01 3.025891e+01 1.150166e+04
 [49] 2.800639e+03 2.231142e+02 1.093685e+02 3.855375e+02 8.261433e+02 7.197227e+01 1.246410e+03 1.275060e+05
 [57] 1.076550e+03 6.563414e+02 5.932213e+03 2.560146e+03 3.037211e+03 1.553417e+03 6.957051e+03 5.456347e+03
 [65] 1.205726e+03 9.877219e+03 4.625740e+05 1.201931e+03 7.376189e+04 8.191378e+02 1.364878e+03 6.217892e+03
 [73] 9.593034e+03 5.169398e+01 2.204646e+04 4.931891e+03 1.144669e+05 6.751020e+02 1.272710e+03 5.088409e+04
 [81] 8.212551e+02 1.724931e+04 1.286482e+06 2.282124e+02 9.212487e+05 5.807697e+01 3.507699e+03 2.555999e+04
 [89] 9.988788e+05 8.365891e+03 6.844667e+01 4.834532e+04 3.974817e+05 1.686135e+03 1.943930e+04 5.487277e+03
 [97] 1.141044e+03 4.957125e+02 2.711287e+04 2.493496e+03 2.020285e+02 3.289130e+03 1.503562e+04 9.603585e+04
[105] 8.589861e+00 5.580201e+04 1.647555e+05 8.121153e+02 1.999879e+01 1.115049e+05 6.792177e+03 1.649786e+03
[113] 1.449654e+03 6.739228e+03 4.651112e+03 1.697557e+04 1.058561e+05 6.898560e+03 1.319406e+04 2.166007e+02
[121] 1.432113e+05 3.489499e+02 2.939112e+04 1.152902e+05 1.930503e+01 3.497105e+01 4.141861e+04 1.338478e+03
[129] 7.821007e+03 2.184877e+04 4.994613e+03 1.687499e+04 2.530742e+03 5.054380e+02 2.396171e+04 3.668709e+03
[137] 1.332597e+05 9.109526e+03 2.200935e+02 1.963455e+04 2.208886e+02 8.742609e+02 4.648689e+05 1.876227e+05
[145] 2.669864e+03 2.165944e+04 1.440512e+04 7.157915e+04 2.019588e+04 5.938379e+03 4.806998e+04 7.091779e+01
[153] 2.654274e+04 2.531568e+05 4.738358e+04 4.226085e+03 4.559174e+05 8.452644e+03 4.077715e+03 1.075783e+04
[161] 1.546512e+04 2.024796e+03 6.685441e+03 1.103567e+04 1.300072e+04 1.147425e+02 3.044792e+04 9.426956e+02
[169] 1.015834e+04 8.339517e+02 8.768228e+03 4.699713e+02 2.734347e+04 5.558394e+01 7.336519e+01 1.221800e+02
[177] 2.068176e+04 4.637194e+02 3.177152e+03 3.139844e+03 6.748911e+01 2.533940e+04 2.366541e+03 1.230315e+02
[185] 5.860547e+02 3.337813e+02 4.232574e+02 8.987410e+03 1.574336e+03 3.469333e+03 1.979679e+04 3.666725e+04
[193] 1.183411e+05 1.169231e+05 5.722712e+04 6.805425e+02 5.449992e+01 1.174923e+03 3.386621e+03 2.323015e+03
[201] 1.103738e+05 1.528929e+03 1.886454e+03 5.798551e+04 6.842591e+03 5.789552e+03 1.128969e+04 1.535207e+04
[209] 1.001666e+02 1.350287e+05 6.871577e+04 4.058903e+02 6.345811e+04 4.811159e+03 5.687049e+04 7.302806e+01
[217] 4.134022e+01 1.330298e+04 1.746314e+03 2.380711e+03 5.957300e+02 1.930560e+02 1.949086e+03 5.868851e+03
[225] 1.300262e+05 3.762561e+03 2.590856e+03 8.612219e+02 1.592960e+03 6.903372e+03 2.352131e+01 3.914039e+03
[233] 4.443285e+03 2.290130e+04 6.127350e+03 3.948759e+03 8.356363e+03 6.598215e+04 3.706861e+03 1.807298e+03
[241] 9.028577e+03 5.862240e+03 1.810417e+04 3.600081e+03 1.234818e+04 4.203168e+04 2.023356e+01 3.797880e+03
[249] 2.650569e+03 4.720094e+06 1.943202e+04 1.664792e+03 8.825819e+03 5.784835e+04 5.144867e+03 1.964639e+03
[257] 1.315585e+06 7.142549e+03 6.636194e+04 3.226585e+04 2.268992e+02 4.764171e+05 3.073412e+04 3.048229e+04
[265] 5.504382e+01 2.755287e+03 5.093347e+03 8.072676e+02 1.394775e+04 3.702878e+02 2.150954e+05 1.518705e+03
[273] 6.828805e+04 1.126930e+03 6.444189e+02 1.128751e+05 1.521137e+04 1.012920e+04 5.403441e+01 3.655090e+02
[281] 2.336058e+04 1.630576e+04 2.751620e+03 3.145077e+03 5.387096e+03 8.024157e+03 5.601803e+04 4.775123e+06
[289] 6.864566e+02 4.615828e+06 8.610612e+01 4.863822e+06 2.401362e+03 4.095088e+01 5.807428e+02 3.775672e+01
[297] 3.925437e+02 3.405028e+03 1.642428e+06 1.784084e+04 6.642704e+04 6.510996e+02 1.255258e+05 1.679222e+06
[305] 3.634272e+04 1.315671e+06 1.269473e+05 7.875022e+03 1.523353e+04 2.423547e+03 8.315757e+03 1.317682e+06
[313] 2.519583e+05 4.956929e+03 1.297203e+04 8.786103e+03 1.504450e+04 3.109673e+01 9.178810e+02 1.153636e+05
[321] 1.284042e+06 6.406535e+03 4.945033e+03 1.985623e+04 8.296752e+03 5.865867e+05 4.967454e+03 4.875231e+06
```

2. Frequency table

```
Console  Terminal ×  Jobs ×
R 3.6.3 · ~/
> table_entity <- table(sample_entity)
> table_entity
sample_entity
                              Afghanistan                                   Albania
                                        2                                         8
                                  Algeria                            American Samoa
                                        6                                         2
                     Andean Latin America                                   Andorra
                                        4                                         8
                                   Angola                       Antigua and Barbuda
                                        7                                         8
                                Argentina                                   Armenia
                                        6                                         7
                              Australasia                                 Australia
                                        5                                        10
                                  Austria                                Azerbaijan
                                        9                                         8
                                  Bahamas                                   Bahrain
                                        6                                         8
                               Bangladesh                                  Barbados
                                        7                                         5
                                  Belarus                                   Belgium
                                        4                                         7
                                   Belize                                     Benin
                                        4                                         7
                                  Bermuda                                    Bhutan
                                        4                                         8
                                  Bolivia                    Bosnia and Herzegovina
                                        6                                         7
                                 Botswana                                    Brazil
                                        6                                         6
                                   Brunei                                  Bulgaria
                                        7                                         9
                             Burkina Faso                                   Burundi
                                        8                                         6
                                 Cambodia                                  Cameroon
                                        7                                         6
                                   Canada                                Cape Verde
                                        3                                        12
                                Caribbean                  Central African Republic
                                        5                                         5
                             Central Asia                            Central Europe
                                        8                                         6
Central Europe, Eastern Europe, and Central Asia                      Central Latin America
                                        9                                         6
                  Central Sub-Saharan Africa                                      Chad
                                        7                                         6
                                    Chile                                     China
                                        6                                         6
                                 Colombia                                   Comoros
                                        9                                         7
                                    Congo                                Costa Rica
                                       14                                         6
                            Cote d'Ivoire                                   Croatia
                                        6                                         9
                                     Cuba                                    Cyprus
                                        6                                         5
```

3. Subset of dataset

```
Console  Terminal  Jobs
R 3.6.3 · ~/
> #subset from original data set
> india_entity <- subset(deathrate_dataset, subset=(deathrate_dataset$Entity=="India"))
> india_entity
     Entity Code Year Unsafe.water.source Unsafe.sanitation No.access.to.handwashing.facility
2689  India  IND 1990            807723.2          636517.8                          430219.3
2690  India  IND 1991            811265.9          638456.6                          430207.6
2691  India  IND 1992            809832.1          636020.6                          428346.3
2692  India  IND 1993            798193.5          625142.5                          421085.4
2693  India  IND 1994            782064.1          611345.7                          411623.0
2694  India  IND 1995            770998.0          601637.2                          404314.4
2695  India  IND 1996            757174.2          590231.4                          395475.5
2696  India  IND 1997            756259.5          589697.0                          393456.1
2697  India  IND 1998            750757.1          585826.1                          388720.3
2698  India  IND 1999            741045.4          577967.5                          381265.2
2699  India  IND 2000            734455.9          572125.8                          376339.0
2700  India  IND 2001            718835.9          559478.2                          367703.6
2701  India  IND 2002            700750.5          544821.6                          357291.3
2702  India  IND 2003            683466.5          529869.5                          346840.4
2703  India  IND 2004            668881.3          515958.9                          338034.0
2704  India  IND 2005            660257.9          505825.7                          333250.2
2705  India  IND 2006            654674.3          497423.1                          329214.7
2706  India  IND 2007            650617.8          489607.0                          325868.8
2707  India  IND 2008            648163.4          481544.8                          322651.4
2708  India  IND 2009            651194.1          476657.0                          321282.4
2709  India  IND 2010            656580.0          473093.9                          321058.3
2710  India  IND 2011            645210.7          457656.6                          313755.2
2711  India  IND 2012            625665.6          434398.5                          302658.1
2712  India  IND 2013            604731.8          405370.5                          289350.1
2713  India  IND 2014            588807.0          377367.7                          276761.9
2714  India  IND 2015            573767.0          352445.6                          267021.1
2715  India  IND 2016            560745.7          332744.7                          258164.9
2716  India  IND 2017            569679.2          328720.0                          257783.9
     Household.air.pollution.from.solid.fuels Non.exclusive.breastfeeding Discontinued.breastfeeding Child.wasting
2689                                 691699.0                   122801.68                   6598.318      861136.0
2690                                 693852.4                   120834.70                   6663.947      847509.5
2691                                 693823.3                   118244.39                   6512.997      832721.1
2692                                 680356.2                   113499.59                   6086.902      810052.1
2693                                 670049.4                   108452.88                   5617.580      791943.4
2694                                 659136.6                   103304.98                   5169.259      774324.9
2695                                 653263.0                    97235.34                   4726.994      748657.1
2696                                 671861.4                    92509.36                   4401.079      721795.7
2697                                 668996.9                    88892.96                   4253.957      691683.7
2698                                 642748.2                    86196.16                   4293.700      661414.9
2699                                 629573.7                    84637.40                   4258.705      639991.3
2700                                 619161.0                    80677.92                   3782.869      606418.3
2701                                 604227.0                    76002.37                   3273.234      572178.6
2702                                 582220.5                    71788.55                   2924.533      536602.1
2703                                 552103.9                    68158.90                   2658.521      506246.0
2704                                 545624.6                    64578.38                   2344.105      479780.3
2705                                 546837.6                    61488.87                   1990.622      454782.8
2706                                 549219.8                    58591.65                   1731.430      433088.8
2707                                 548852.7                    56138.69                   1664.638      413314.0
2708                                 547817.3                    54242.16                   1707.459      393356.2
2709                                 550363.9                    53166.67                   1817.689      376323.4
2710                                 551432.3                    52162.28                   1916.172      356571.0
2711                                 563339.7                    51151.27                   2037.334      339512.3
2712                                 542067.4                    48605.79                   2097.361      313727.1
```

The descriptive analysis was again performed on these new variables and on the subset of the data to get a new idea about the statistical values further which would help us in hypothesis testing.

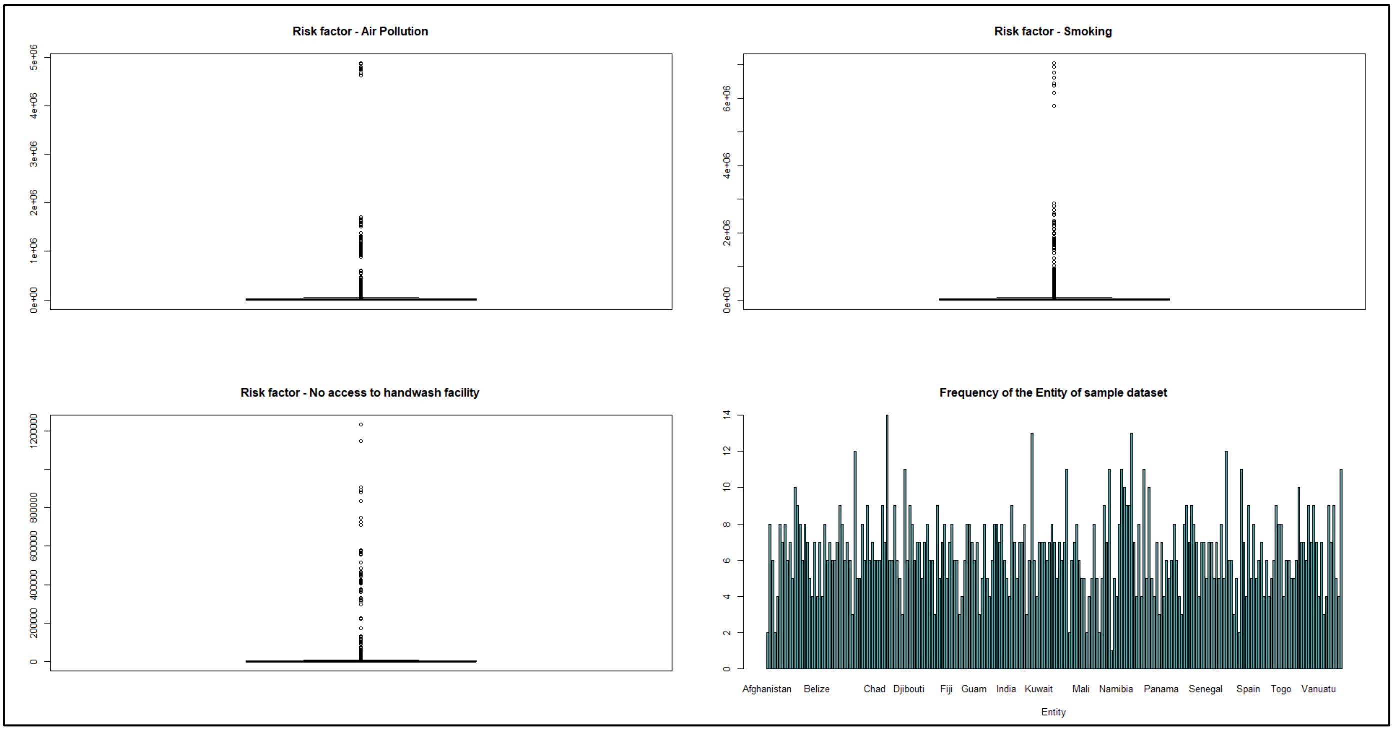
Output:

```
Console  Terminal  Jobs
R 3.6.3 · ~/
> #descriptive analysis of the variables created
> mean(air_pollution)
[1] 108380.7
> summary(air_pollution)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
      9    1205    6400  108381   23919 4875231 
> mean(smoking)
[1] 151514.8
> summary(smoking)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
     12    1448    6285  151515   30897 7046703 
> mean(no_access_to_handwash)
[1] 22236.99
> summary(no_access_to_handwash)
     Min.    1st Qu.     Median       Mean    3rd Qu.       Max. 
      0.1       19.2      262.4    22237.0     3809.5  1230318.6 
> mean(india_entity$Alcohol.use)
[1] 437495.6
> summary(india_entity$Alcohol.use)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 292070  379993  421276  437496  517488  580499 
> mean(australia_entity$Drug.use)
[1] 1945.317
> summary(australia_entity$Drug.use)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1251    1737    1874    1945    2175    2870 
> mean(eastasia_entity$Iron.deficiency)
[1] 2290.679
> summary(eastasia_entity$Iron.deficiency)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  389.9   632.3  1544.4  2290.7  3864.8  5856.6 
> 
```

Data Visualization

Before getting into further analysis and testing, data visualization was performed to get a better understanding of the data that was created earlier. The visualizations created were based on the variables of air pollution, smoking and no access to hand wash facility along with the frequency of the entity variable of the sample dataset. For the risk factors, boxplot was created to get a generalized idea about the data and bar plot was created for the frequency of the entity variable.

Output:



The above figure shows 4 graphs related to the dataset which explains the boxplot and bar plot for the variables of risk factors and the subset of entity data.

Hypothesis Testing

The data analysis and data visualizations gave us an idea about the dataset and now we further proceed to the hypothesis testing. Hypothesis testing is basically used to test an assumption regarding a population parameter by using a sample data. In the hypothesis testing there are two tests which are performed: one sample t-test and two sample t-test. One sample t-test is performed to test whether or not the mean of a population is equal to some value whereas two sample t-test is used to test whether or not the means of two populations are equal.

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

One sample t-test was performed on the new variables that were created earlier to test whether the mean of a population is equal to some assumed value or not. The risk factors that were considered for the one sample testing are, air_pollution, smoking, and no_access_to_handwash. The dataset as mentioned earlier was a sample dataset chosen at random.
The outputs of the one sample testing for the three risk factors are as below.

Output:

1. Air Pollution

```
Console  Terminal ×  Jobs ×
R 3.6.3 · ~/
> 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 
```

2. Smoking

```
Console  Terminal ×  Jobs ×
R 3.6.3 · ~/
> 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
```

3. No access to handwash facility

```
Console  Terminal ×  Jobs ×
R 3.6.3 · ~/
> 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.

Next analysis was tried with respect to one sample t-test with additional parameters. The additional parameter was the alternative which could be either less or greater. Basically, this parameter assumes that the actual mean value is either greater or lesser than the assumed mean value. The output obtained was as follows.

Output:

```
Console   Terminal ×   Jobs ×
R 3.6.3 · ~/
> #one sample testing for overall data set with additional parameters
> t.test(air_pollution, mu = 104520, alternative = "greater")          #reject alternative hypothesis

        One Sample t-test

data:  air_pollution
t = 0.33963, df = 1499, p-value = 0.3671
alternative hypothesis: true mean is greater than 104520
95 percent confidence interval:
 89671.5      Inf
sample estimates:
mean of x
 108380.7

> t.test(smoking, mu = 200000, alternative = "less")                   #reject null hypothesis

        One Sample t-test

data:  smoking
t = -3.0784, df = 1499, p-value = 0.001059
alternative hypothesis: true mean is less than 2e+05
95 percent confidence interval:
     -Inf 177437.4
sample estimates:
mean of x
 151514.8

> t.test(no_access_to_handwash, mu = 16000 , alternative = "less")     #reject alternative hypothesis

        One Sample t-test

data:  no_access_to_handwash
t = 2.4226, df = 1499, p-value = 0.9922
alternative hypothesis: true mean is less than 16000
95 percent confidence interval:
     -Inf 26474.36
sample estimates:
mean of x
 22236.99
```

Further, the answer to the second question is solved by testing for the entity variable. Similar approach is followed here with a difference that the test is conducted for a particular entity only and not the overall sample data. Here, again, depending on the p-value we either rejected the null hypothesis or not and the level of significance here too does not change our conclusion. The output obtained for the same is as follows.

Output:

```
Console   Terminal ×   Jobs ×
R 3.6.3 · ~/
> #one sample testing for subset of data set
> t.test(india_entity$Alcohol.use, mu = 430000)                         #reject the alternative hypothesis

        One Sample t-test

data:  india_entity$Alcohol.use
t = 0.4534, df = 27, p-value = 0.6539
alternative hypothesis: true mean is not equal to 430000
95 percent confidence interval:
 403574.5 471416.7
sample estimates:
mean of x
 437495.6

> t.test(india_entity$Alcohol.use, mu = 200000, alternative = "greater")           #reject the null hypothesis

        One Sample t-test

data:  india_entity$Alcohol.use
t = 14.366, df = 27, p-value = 1.823e-14
alternative hypothesis: true mean is greater than 2e+05
95 percent confidence interval:
 409336.6      Inf
sample estimates:
mean of x
 437495.6

>
> t.test(australia_entity$Drug.use, mu = 2000)                          #reject the alternative hypothesis

        One Sample t-test

data:  australia_entity$Drug.use
t = -0.66675, df = 27, p-value = 0.5106
alternative hypothesis: true mean is not equal to 2000
95 percent confidence interval:
 1777.039 2113.596
sample estimates:
mean of x
 1945.317

> t.test(australia_entity$Drug.use, mu = 3000, alternative = "less")              #reject the null hypothesis

        One Sample t-test

data:  australia_entity$Drug.use
t = -12.86, df = 27, p-value = 2.506e-13
alternative hypothesis: true mean is less than 3000
95 percent confidence interval:
    -Inf 2085.01
sample estimates:
mean of x
 1945.317
```

<u>Two sample t-test:</u>

Two sample t-test helped in answering the next two questions which was performed on the risk factors of the dataset and the subset of the entity parameter. Two sample t-test is performed when there are independent parameters, and we test whether the two population means are same or different. For the two-sample t-test, new variables and subset of the sample were created same as it was done for one sample testing for which the testing has to be done along with their descriptive analysis.

Output:

Descriptive analysis on the new variables.

```
Console   Terminal ×   Jobs ×
R 3.6.3 · ~/
> #descriptive analysis of the variables created
> mean(alcohol_use)
[1] 58064.65
> summary(alcohol_use)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
   -789.5     472.8    3165.2   58064.6   13745.3 2817609.6
> mean(drug_use)
[1] 10118.6
> summary(drug_use)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
     1.4    110.1    435.8  10118.6   2324.9 572923.0
> mean(england_entity$Diet.low.in.fruits)
[1] 18320.67
> summary(england_entity$Diet.low.in.fruits)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  14321   14861   18015   18321   21179   24282
> mean(italy_entity$Diet.low.in.fruits)
[1] 12766.82
> summary(italy_entity$Diet.low.in.fruits)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  11306   11616   12283   12767   13783   15572
> 
```

The two-sample t-test done for the new variables created and the subset of dataset for which we obtain the answers to our questions is as below.

Output:

1. t-test on risk factors

```
Console   Terminal ×   Jobs ×
R 3.6.3 · ~/
> #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
```

2. t-test on subset of dataset

```
Console   Terminal ×   Jobs ×
R 3.6.3 · ~/
> #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.

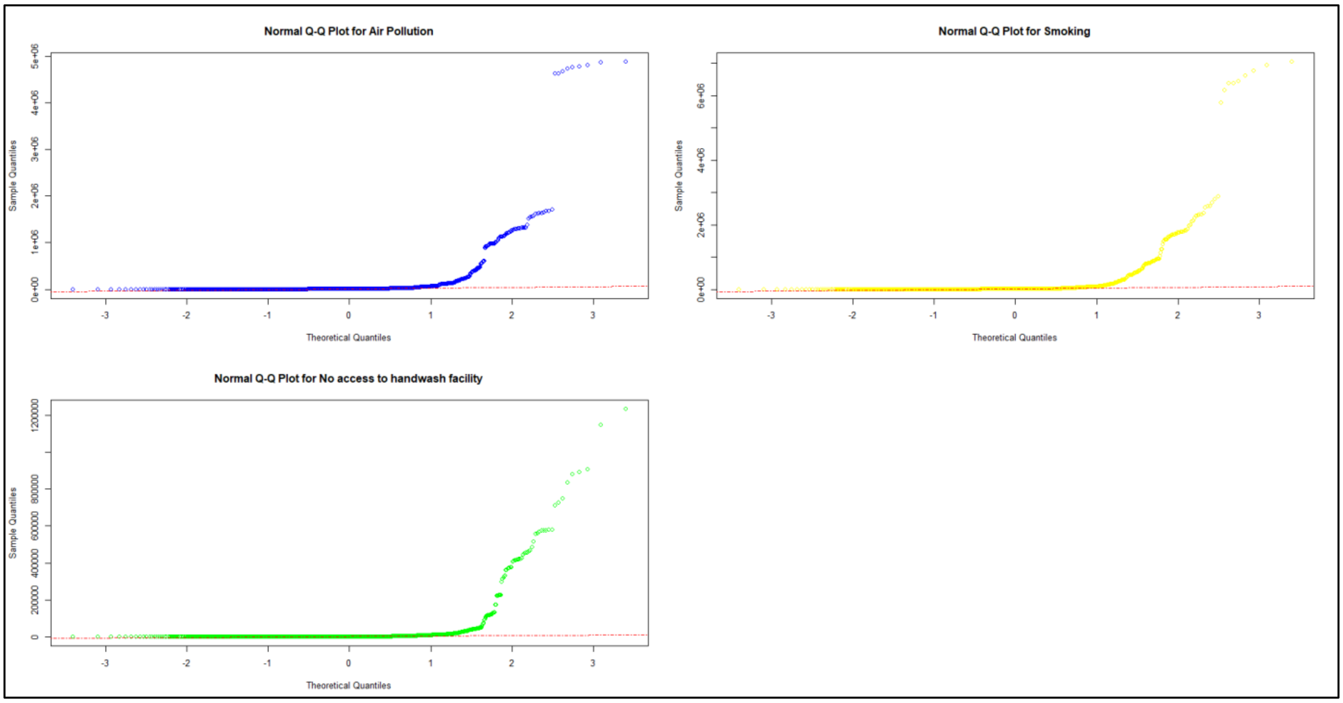
Also, it can be noted that a paired t-test cannot be performed here in this case since we reject the null hypothesis in the two-sample t-test.

Data Visualizations for Testing:

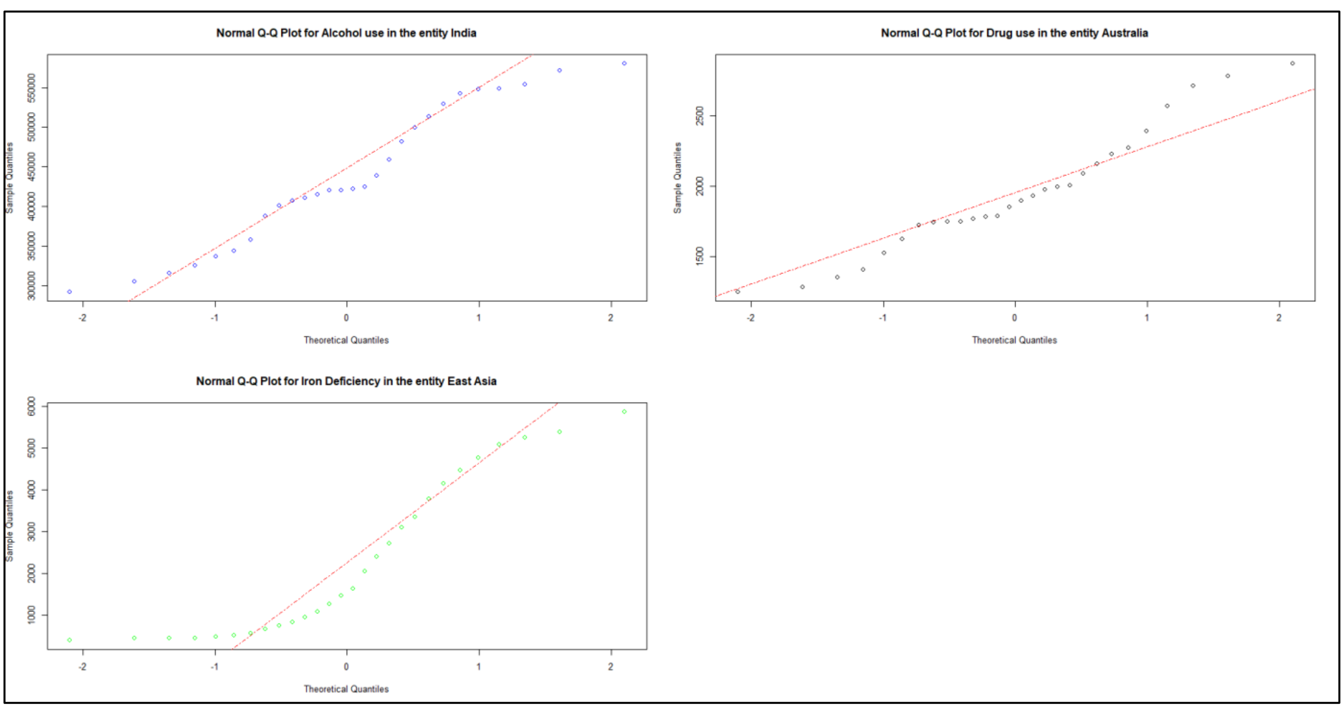
After the one-sample and two-sample t-test were conducted, visualizations were created in order to get a clear understanding about the parameters and the subset of the dataset. The visualization created was the normal Q-Q plot which is nothing but a probability plot for a graphical method to compare two probability distributions. The normal q-q plot plotted for the three variables and for the subset of the dataset is as shown below.

Output:

a. One sample testing



b. Two sample testing



Summary

Hypothesis testing is a technique to help determine whether a specific treatment has an effect on the individuals in a population. The task here was to answer some of the questions statistics and testing which would help in the analysis of the data. Descriptive analysis was performed first to get an overall idea about the dataset after which the hypothesis testing on particular attributes was performed. The summary function during the descriptive analysis helps in understanding the overall statistical values of the dataset and it returns the statistical values like the min, max, mean, median and the range. Data visualizations also performed gave insights about the dataset which further helped in analysis.

The 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. For this analysis, England was affected more due to the risk factor of diet low in fruits as compared to in Italy.

Thus, hypothesis testing using the t-test helped in answering the questions related to the dataset and for further analysis. It gave an overall idea from both the one sample t-test and two sample t-test. The testing and analysis on the number of death rate by risk factors helps in understanding the risk factors as an individual parameter and also with respect to a particular entity whether which entity is most affected by a risk factor as compared to the other.

References

Two Sample T-Test Unequal Variance. (n.d.). The Open Educator. Retrieved November 25, 2021, from https://www.theopeneducator.com/doe/hypothesis-Testing-Inferential-Statistics-Analysis-of-Variance-ANOVA/Two-Sample-T-Test-Unequal-Variance

How To Do Two-Sample T-test in R. (n.d.). Data Novia. Retrieved November 25, 2021, from https://www.datanovia.com/en/lessons/how-to-do-a-t-test-in-r-calculation-and-reporting/how-to-do-two-sample-t-test-in-r/

One Sample T Test – Clearly Explained with Examples | ML+. (n.d.). Machine Learning +. Retrieved November 25, 2021, from https://www.machinelearningplus.com/statistics/one-sample-t-test/#purpose-of-one-sample-t-test

Paired T-Test. (n.d.). Byjus. Retrieved November 25, 2021, from https://byjus.com/maths/paired-t-test/

Zach, S. (2019, May 15). How to Perform a Paired Samples t-test in R. Statology. Retrieved November 26, 2021, from https://www.statology.org/paired-samples-t-test-r/#:~:text=How%20to%20Conduct%20a%20Paired%20t-test%20in%20R.,that%20we%20want%20to%20compute%20a%20paired%20t-test