Death Rates of Cancer, Diabetes Mellitus, and Heart Disease in Eurozone

Clementine Surya

Load necessary package

First of all, I load in necessary package.

```
library("tidyverse")
library("dplyr")
library("tidyr")
library("ggplot2")
library("readxl") #To read xlxs file
library("gridExtra") #To show ggplot side by side
```

Part 1: Analysis

The aim for part 1 is to find dataset that contains of mix of categorical and numerical variables. I choose three dataset from Eurostat website which contains the death rates that were caused by Cancer, Diabetes Mellitus, and Heart Disease in Eurozone by Gender for Year 2017, 2018, and 2019.

The death rate describes mortality in relation to the total population. Expressed in deaths per 100, 000 inhabitants, it is calculated as the number of deaths recorded in the population for a given period divided by population in the same period and then multiplied by 100,000.

Below are the links to the datasets:

- Death due to Cancer: https://ec.europa.eu/eurostat/databrowser/view/TPS00116/default/table?lang=en&category=hlth.hlth_cdeath.hlth_cd_gmor (https://ec.europa.eu/eurostat/databrowser/view/TPS00116/default/table?lang=en&category=hlth.hlth_cdeath.hlth_cd_gmor)

- Death due to Diabetes Mellitus: https://ec.europa.eu/eurostat/databrowser/view/TPS00137/default/table?lang=en&category=hlth.hlth_cdeath.hlth_cd_gmor (https://ec.europa.eu/eurostat/databrowser/view/TPS00137/default/table?lang=en&category=hlth.hlth_cdeath.hlth_cd_gmor)

- Death due to heart disease: https://ec.europa.eu/eurostat/databrowser/view/TPS00119/default/table?lang=en&category=hlth.hlth_cdeath.hlth_cd_gmor (https://ec.europa.eu/eurostat/databrowser/view/TPS00119/default/table?lang=en&category=hlth.hlth_cdeath.hlth_cd_gmor)

In these dataset, the categorical columns are: Country and Year while the numerical columns are: Death Rates of Cancer, Diabetes Mellitus, and Heart Disease (by genders).

To sum up for the analysis part:
1. Death Rates that were caused by Cancer, Diabetes Mellitus, and Heart Disease were all higher among male than female.
2. By comparing death of rate caused by Cancer, Diabetes Mellitus, and Heart Disease, the cause of death by Cancer is the highest, followed by Heart Disease, and the last is Diabetes Mellitus in every year from 2017 to 2019.
3. The highest death rate caused by Cancer in 2019 for male was in Latvia, while for female was in Hungary.
4. The highest death rate caused by Diabetes Mellitus in 2019 for both male and female were in Croatia.
5. The highest death rate caused by Heart Disease in 2019 for both male and female were in Lithuania.
6. The top four most strongly correlated pairs for the numerical variables are shown below:

- Heart males death rate and heart females death rate have a strong positive correlation with correlation of 0.98.
- Diabetes males death rate and diabetes females death rate have a strong positive correlation with correlation of 0.97.
- Cancer males death rate and heart males death rate have a moderate positive correlation with correlation of 0.66.
- Cancer males death rate and cancer females death rate have a moderate positive correlation with correlation of 0.64.

7. When predicting the cancer males death rates, the predictor: cancer females death rate, diabetes males death rate, diabetes females death rate have p-value < 0.05 (using 5% significance level). Thus, these variables are considered as significant predictor. Furthermore, the adjusted R^2 in this fitted model which is 0.6963. Thus, 69.63% of the variation in cancer males death ratecan be explained by the fitted model.

Load the dataset

In this part, I load all the three dataset and combine them into one.

Cancer Death Rate Dataset

Load Cancer death rate dataset and tidy this dataset by removing rows that are not being used and removing the missing values.

```
#Load Cancer Death Rate dataset
df_cancer = read_excel("cancer.xlsx", skip = 9)
#Remove 7 rows in the bottom of this dataset
df_cancer = slice(df_cancer, 1:(n() - 7))
```

Here, I count the “:” symbol that needs to be removed

```
#Count ":" symbol in the data that we want to remove
sum(df_cancer == ":")
```

```
## [1] 24
```

Replace “:” value with NA and removing the the 24 missing values from this dataset.

```
#Replace ":" symbol with NA value
df_cancer <- replace(df_cancer, df_cancer==":", NA)
#Remove NA value from this dataset
df_cancer <- df_cancer %>% drop_na()
```

```
#Change the death rate from character to numeric value
i <- c(3, 4)
df_cancer[ , i] <- apply(df_cancer[ , i], 2,
                         function(x) as.numeric(as.character(x)))
#Rename the column names of this dataset
names(df_cancer) <- c('country','year','cancer_males','cancer_females')
#Then, I print the head, structure, and dimension of the cancer death rate dataset
head(df_cancer)
```

```
## # A tibble: 6 × 4
##   country  year  cancer_males cancer_females
##   <chr>    <chr>        <dbl>          <dbl>
## 1 Belgium  2017          309.           189.
## 2 Belgium  2018          300.           182.
## 3 Belgium  2019          295            183.
## 4 Bulgaria 2017          319            172.
## 5 Bulgaria 2018          320.           172.
## 6 Bulgaria 2019          335.           178.
```

```
str(df_cancer)
```

```
## tibble [99 × 4] (S3: tbl_df/tbl/data.frame)
##  $ country       : chr [1:99] "Belgium" "Belgium" "Belgium" "Bulgaria" ...
##  $ year          : chr [1:99] "2017" "2018" "2019" "2017" ...
##  $ cancer_males  : num [1:99] 309 300 295 319 320 ...
##  $ cancer_females: num [1:99] 189 182 183 172 172 ...
```

```
dim(df_cancer)
```

```
## [1] 99  4
```

Diabetes Death Rate Dataset

Load Diabetes death rate dataset and tidy the data using the same method as above.

```
#Load Diabetes Death Rate dataset
df_diabetes = read_excel("diabetes.xlsx", skip = 9)
#Remove 7 rows in the bottom of this dataset
df_diabetes = slice(df_diabetes, 1:(n() - 7))
```

Here, I count the “:” symbol that needs to be removed

```
#Count ":" symbol in the data that we want to remove
sum(df_diabetes == ":")
```

```
## [1] 24
```

Replace “:” value with NA and removing the the 24 missing values from this dataset.

```
#Replace ":" symbol with NA value
df_diabetes <- replace(df_diabetes, df_diabetes==":", NA)
#Remove NA value from this dataset
df_diabetes <- df_diabetes %>% drop_na()
```

```
#Change the death rate from character to numeric value
i <- c(3, 4)
df_diabetes[ , i] <- apply(df_diabetes[ , i], 2,
                          function(x) as.numeric(as.character(x)))
#Remove NA value from this dataset
df_diabetes <- df_diabetes %>% drop_na()
#Rename the column names of this dataset
names(df_diabetes) <- c('country','year','diabetes_males','diabetes_females')
#Then, I print the head, structure, and dimension of the diabetes death rate dataset
head(df_diabetes)
```

```
## # A tibble: 6 × 4
##   country  year  diabetes_males diabetes_females
##   <chr>    <chr>          <dbl>            <dbl>
## 1 Belgium  2017            16.4             11.0
## 2 Belgium  2018            15.8             10.7
## 3 Belgium  2019            14.2             10.6
## 4 Bulgaria 2017            23.1             19.4
## 5 Bulgaria 2018            25.5             20.8
## 6 Bulgaria 2019            23.9             18.5
```

```
str(df_diabetes)
```

```
## tibble [99 × 4] (S3: tbl_df/tbl/data.frame)
##  $ country         : chr [1:99] "Belgium" "Belgium" "Belgium" "Bulgaria" ...
##  $ year            : chr [1:99] "2017" "2018" "2019" "2017" ...
##  $ diabetes_males  : num [1:99] 16.4 15.8 14.2 23.1 25.5 ...
##  $ diabetes_females: num [1:99] 11.1 10.7 10.6 19.4 20.9 ...
```

```
dim(df_diabetes)
```

```
## [1] 99  4
```

Heart Disease Death Rate Dataset

The last dataset is Heart Disease dataset and here I load the data and tidy the data.

```
#Load Heart Disease Death Rate dataset
df_heart = read_excel("heart_disease.xlsx", skip = 9)
#Remove 7 rows in the bottom of this dataset
df_heart = slice(df_heart, 1:(n() - 7))
```

Here, I count the “:” symbol that needs to be removed

```
#Count ":" symbol in the data that we want to remove
sum(df_heart == ":")
```

```
## [1] 24
```

Replace “:” value with NA and removing the the 24 missing values from this dataset.

```
#Replace ":" symbol with NA value
df_heart <- replace(df_heart, df_heart==":", NA)
#Remove NA value from this dataset
df_heart <- df_heart %>% drop_na()
```

```
#Change the death rate from character to numeric value
i <- c(3, 4)
df_heart[ , i] <- apply(df_heart[ , i], 2,
                         function(x) as.numeric(as.character(x)))
#Remove NA value from this dataset
df_heart <- df_heart %>% drop_na()
#Rename the column names of this dataset
names(df_heart) <- c('country','year','heart_males','heart_females')
#Then, I print the head, structure, and dimension of the heart disease death rate dataset
head(df_heart)
```

```
## # A tibble: 6 × 4
##   country  year  heart_males heart_females
##   <chr>    <chr>       <dbl>         <dbl>
## 1 Belgium  2017         44.9          18.2
## 2 Belgium  2018         42.0          16.4
## 3 Belgium  2019         39.0          16.1
## 4 Bulgaria 2017        159.          104. 
## 5 Bulgaria 2018        141.           96.9
## 6 Bulgaria 2019        157.          107. 
```

```
str(df_heart)
```

```
## tibble [99 × 4] (S3: tbl_df/tbl/data.frame)
##  $ country      : chr [1:99] "Belgium" "Belgium" "Belgium" "Bulgaria" ...
##  $ year         : chr [1:99] "2017" "2018" "2019" "2017" ...
##  $ heart_males  : num [1:99] 44.9 42 39 159.3 140.8 ...
##  $ heart_females: num [1:99] 18.2 16.4 16.1 104.4 96.9 ...
```

```
dim(df_heart)
```

```
## [1] 99  4
```

Combine the dataset

After loading and cleaning all the three datasets, I combine them into one dataset which is called “df_all”. This new dataset has 99 rows and 8 columns.

```
#First, I combine Cancer and Diabetes Death Rate datasets together
df_cancer_diabetes =  df_cancer %>% right_join(df_diabetes, by=c("country","year"))
#Then, I combine all the data together into one dataset. This dataset will be used throughout the analysis p
art below
df_all =  df_cancer_diabetes %>% right_join(df_heart, by=c("country","year"))
#Print the head, structure, and dimension of the combined dataset
head(df_all)
```

```
## # A tibble: 6 × 8
##   country  year  cancer_males cancer_females diabetes_…¹ diabe…² heart…³ heart…⁴
##   <chr>    <chr>        <dbl>          <dbl>       <dbl>   <dbl>   <dbl>   <dbl>
## 1 Belgium  2017          309.           189.        16.4    11.0    44.9    18.2
## 2 Belgium  2018          300.           182.        15.8    10.7    42.0    16.4
## 3 Belgium  2019          295            183.        14.2    10.6    39.0    16.1
## 4 Bulgaria 2017          319            172.        23.1    19.4   159.    104. 
## 5 Bulgaria 2018          320.           172.        25.5    20.8   141.     96.9
## 6 Bulgaria 2019          335.           178.        23.9    18.5   157.    107. 
## # … with abbreviated variable names ¹diabetes_males, ²diabetes_females,
## #   ³heart_males, ⁴heart_females
```

```
str(df_all)
```

```
## tibble [99 × 8] (S3: tbl_df/tbl/data.frame)
##  $ country         : chr [1:99] "Belgium" "Belgium" "Belgium" "Bulgaria" ...
##  $ year            : chr [1:99] "2017" "2018" "2019" "2017" ...
##  $ cancer_males    : num [1:99] 309 300 295 319 320 ...
##  $ cancer_females  : num [1:99] 189 182 183 172 172 ...
##  $ diabetes_males  : num [1:99] 16.4 15.8 14.2 23.1 25.5 ...
##  $ diabetes_females: num [1:99] 11.1 10.7 10.6 19.4 20.9 ...
##  $ heart_males     : num [1:99] 44.9 42 39 159.3 140.8 ...
##  $ heart_females   : num [1:99] 18.2 16.4 16.1 104.4 96.9 ...
```

```
dim(df_all)
```

```
## [1] 99  8
```

Numerical Summary

I begin the analysis part by showing the numerical summary for each year:

```
by(df_all, df_all$year, summary)
```

```
## df_all$year: 2017
##    country              year            cancer_males   cancer_females
##  Length:34          Length:34          Min.   :214.7   Min.   :121.7
##  Class :character   Class :character   1st Qu.:291.9   1st Qu.:176.9
##  Mode  :character   Mode  :character   Median :323.3   Median :201.0
##                                        Mean   :342.0   Mean   :198.6
##                                        3rd Qu.:381.0   3rd Qu.:226.4
##                                        Max.   :473.1   Max.   :263.0
##  diabetes_males  diabetes_females  heart_males     heart_females
##  Min.   : 0.00   Min.   : 0.00    Min.   : 35.13   Min.   : 13.42
##  1st Qu.:18.65   1st Qu.:12.42    1st Qu.: 61.73   1st Qu.: 30.68
##  Median :23.34   Median :16.29    Median : 95.03   Median : 46.79
##  Mean   :28.54   Mean   :22.11    Mean   :149.83   Mean   : 91.17
##  3rd Qu.:38.48   3rd Qu.:29.32    3rd Qu.:176.68   3rd Qu.:100.00
##  Max.   :61.59   Max.   :54.39    Max.   :636.88   Max.   :408.00
## ------------------------------------------------------------
## df_all$year: 2018
##    country              year            cancer_males   cancer_females
##  Length:33          Length:33          Min.   :226.1   Min.   :117.6
##  Class :character   Class :character   1st Qu.:284.1   1st Qu.:174.0
##  Mode  :character   Mode  :character   Median :320.0   Median :198.2
##                                        Mean   :338.2   Mean   :196.8
##                                        3rd Qu.:380.1   3rd Qu.:224.0
##                                        Max.   :464.8   Max.   :258.2
##  diabetes_males  diabetes_females  heart_males     heart_females
##  Min.   : 7.63   Min.   : 0.00    Min.   : 33.12   Min.   : 12.31
##  1st Qu.:17.67   1st Qu.:12.08    1st Qu.: 64.35   1st Qu.: 29.62
##  Median :22.76   Median :16.68    Median : 92.71   Median : 49.71
##  Mean   :29.58   Mean   :22.32    Mean   :145.74   Mean   : 88.34
##  3rd Qu.:40.71   3rd Qu.:28.82    3rd Qu.:178.88   3rd Qu.: 96.86
##  Max.   :77.27   Max.   :63.92    Max.   :600.43   Max.   :381.51
## ------------------------------------------------------------
## df_all$year: 2019
##    country              year            cancer_males   cancer_females
##  Length:32          Length:32          Min.   :244.2   Min.   :106.2
##  Class :character   Class :character   1st Qu.:276.7   1st Qu.:174.5
##  Mode  :character   Mode  :character   Median :321.9   Median :191.6
##                                        Mean   :331.5   Mean   :192.8
##                                        3rd Qu.:377.6   3rd Qu.:218.3
##                                        Max.   :454.6   Max.   :253.0
##  diabetes_males   diabetes_females  heart_males     heart_females
##  Min.   :  7.30   Min.   : 5.88    Min.   : 32.57   Min.   : 12.91
##  1st Qu.: 19.13   1st Qu.:13.96    1st Qu.: 56.89   1st Qu.: 26.69
##  Median : 24.04   Median :16.64    Median : 88.59   Median : 46.19
##  Mean   : 30.31   Mean   :22.81    Mean   :137.90   Mean   : 82.20
##  3rd Qu.: 36.52   3rd Qu.:28.55    3rd Qu.:159.27   3rd Qu.:105.29
##  Max.   :106.97   Max.   :86.56    Max.   :558.28   Max.   :349.88
```

First of all, I analyse the death rate that was caused by Cancer per year for both male and female:
- In 2017, male had death rate caused by cancer with mean 342.0 and median 323.3 (indicating a right skewed distribution) while female had death rate caused by cancer with mean 198.6 and median 201 (indicating a left skewed distribution).
- In 2018, male had death rate caused by cancer with mean 338.2 and median 320.0 (indicating a right skewed distribution) while female had death rate caused by cancer with mean 196.8 and median 198.2 (indicating a left skewed distribution).
- In 2019, male had death rate caused by cancer with mean 331.5 and median 321.9 (indicating a right skewed distribution) while female had death rate caused by cancer with mean 192.8 and median 191.6 (indicating a right skewed distribution).

From above analysis of cancer death rate, we can see in all the years, the cancer date rate was higher among men than women.

Second of all, I analyse the death rate that was caused by Diabetes per year for both male and female:
- In 2017, male had death rate caused by diabetes with mean 28.54 and median 23.34 (indicating a right skewed distribution) while female had death rate caused by diabetes with mean 22.11 and median 16.29 (indicating a right skewed distribution).
- In 2018, male had death rate caused by diabetes with mean 29.58 and median 22.76 (indicating a right skewed distribution) while female had death rate caused by diabetes with mean 22.32 and median 16.68 (indicating a right skewed distribution).
- In 2019, male had death rate caused by diabetes with mean 30.31 and median 24.04 (indicating a right skewed distribution) while female had death rate caused by diabetes with mean 22.81 and median 16.64 (indicating a right skewed distribution).

From above analysis of diabetes death rate, we can see in all the years, the diabetes death rate was higher among men than women.

Lastly, I analyse the death rate that was caused by Heart Disease per year for both male and female:

- In 2017, male had death rate caused by heart disease with mean 149.83 and median 95.03 (indicating a right skewed distribution) while female had death rate caused by heart disease with mean 91.17 and median 46.79 (indicating a right skewed distribution).
- In 2018, male had death rate caused by heart disease with mean 145.74 and median 92.71 (indicating a right skewed distribution) while female had death rate caused by heart disease with mean 88.34 and median 49.71 (indicating a right skewed distribution).
- In 2019, male had death rate caused by heart disease with mean 137.90 and median 88.59 (indicating a right skewed distribution) while female had death rate caused by heart disease with mean 82.20 and median 46.19 (indicating a right skewed distribution).

From above analysis of heart disease death rate, we can see in all the years, the heart disease death rate was higher among men than women.

In conclusion, the death rate caused by cancer, diabetes, and heart disease were all higher among men than women.

Graphical Summary

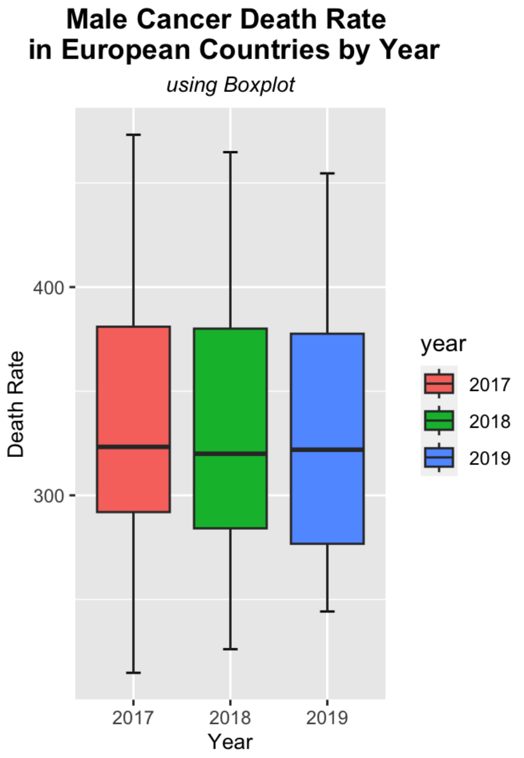
After showing the numerical summary, I create graphical summary by using boxplot for cancer, diabetes, and heart disease death rate distributions.

Cancer Death Rate Bloxpot:

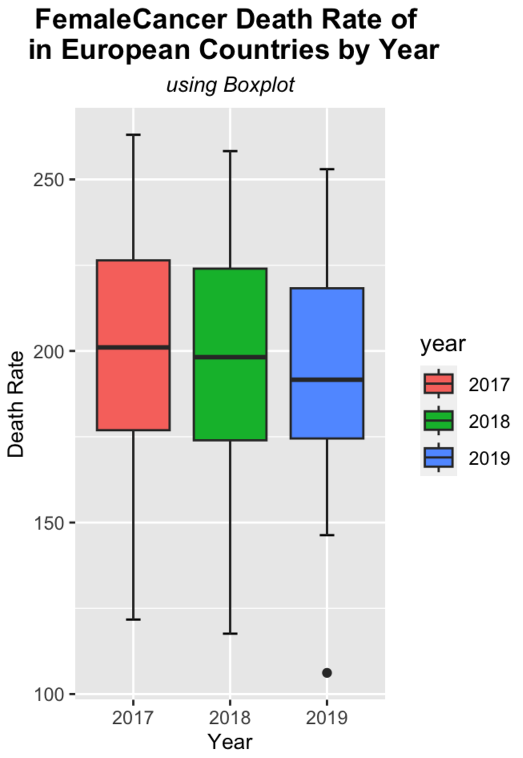
```
#Create boxplot for male cancer death rate by year
cancer_male_plot = ggplot(df_all, aes(x= as.character(year), y = cancer_males, fill = year))+
  stat_boxplot(geom = "errorbar",
               width = 0.15)+
  geom_boxplot()+
  labs(title = "Male Cancer Death Rate \n in European Countries by Year",
       subtitle = "using Boxplot",
       x = "Year",
       y = "Death Rate")+
  theme(plot.title = element_text(face="bold",hjust = 0.5),
        plot.subtitle = element_text(face="italic",hjust = 0.5, size = 10),
        axis.title.x = element_text(size = 10),
        axis.title.y = element_text(size= 10))

#Create boxplot for female cancer death rate by year
cancer_female_plot = ggplot(df_all, aes(x= as.character(year), y = cancer_females, fill = year))+
  stat_boxplot(geom = "errorbar",
               width = 0.15)+
  geom_boxplot()+
  labs(title = "FemaleCancer Death Rate of \n in European Countries by Year",
       subtitle = "using Boxplot",
       x = "Year",
       y = "Death Rate")+
  theme(plot.title = element_text(face="bold",hjust = 0.5),
        plot.subtitle = element_text(face="italic",hjust = 0.5, size = 10),
        axis.title.x = element_text(size = 10),
        axis.title.y = element_text(size= 10))

#Show the male and female cancer death rate plot side by side
grid.arrange(cancer_male_plot, cancer_female_plot, ncol=2)
```


Male Cancer Death Rate in European Countries by Year

using Boxplot


FemaleCancer Death Rate of in European Countries by Year

using Boxplot

In 2017, male death rate by cancer has right skewed distribution, which is also shown in the numerical summary part, where the mean is 342.0 and median is 323.3. The spread of the middle distribution is between 291.9 and 381.0 (IQR range of 89.1) and it is ranging from 214.7 to 473.1 (range value of 258.4). Where in 2018, male death rate by cancer has right skewed distribution, which is also shown in the numerical summary part, where the mean is 338.2 and median is 320.0. The spread of the middle distribution is between 284.1 and 380.1 (IQR range of 96) and it is ranging from 226.1 to 464.8 (range value of 238.7). And in 2019, male death rate by cancer has right skewed distribution, which is also shown in the numerical summary part, where the mean is 331.5 and median is 321.9. The spread of the middle distribution is between 276.7 and 377.6 (IQR range of 100.9) and it is ranging from 244.2 to 454.6 (range value of 210.4).
There are no outlier in these distributions.

In 2017, female death rate by cancer has left skewed distribution, which is also shown in the numerical summary part, where the mean is 198.6 and median is 201. The spread of the middle distribution is between 176.9 and 226.4 (IQR range of 49.5) and it is ranging from 121.7 to 263 (range value of 141.3). Where in 2018, female death rate by cancer has left skewed distribution, which is also shown in the numerical summary part, where the mean is 196.8 and median is 198.2. The spread of the middle distribution is between 174.0 and 224.0 (IQR range of 50.0) and it is ranging from 117.6 to 258.2 (range value of 140.6). And in 2019, female death rate by cancer has right skewed distribution, which is also shown in the numerical summary part, where the mean is 192.8 and median is 191.6. The spread of the middle distribution is between 174.5 and 218.3 (IQR range of 43.8) and it is ranging from 106.2 to 253.0 (range value of 146.8).
There is one outlier located below the minimum value of year 2019 distribution.

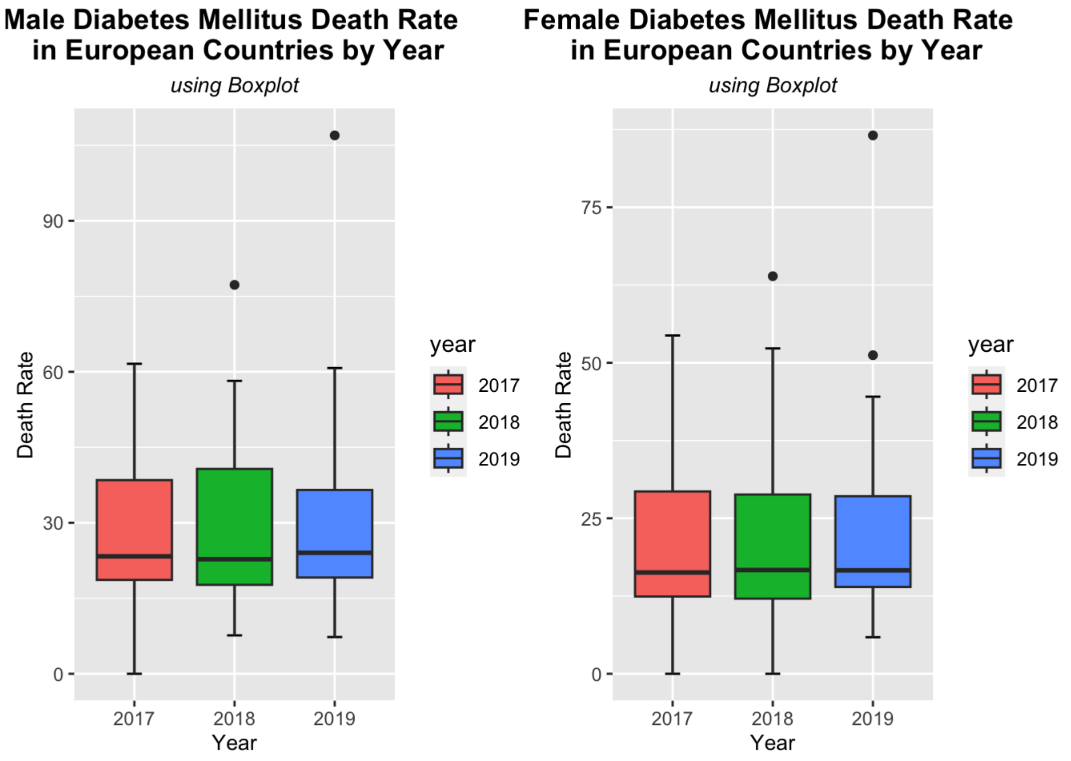
From the boxplots of cancer death rate above, we can see in all the years, the cancer death rate was higher among men than women, which is also shown in numerical analysis part.

Diabetes Mellitus Death Rate Bloxpot:

```
#Create boxplot for male diabetes death rate by year
diabetes_male_plot = ggplot(df_diabetes, aes(x= as.character(year), y = diabetes_males, fill = year))+
  stat_boxplot(geom = "errorbar",
               width = 0.15)+
  geom_boxplot()+
  labs(title = "Male Diabetes Mellitus Death Rate \n in European Countries by Year",
       subtitle = "using Boxplot",
       x = "Year",
       y = "Death Rate")+
  theme(plot.title = element_text(face="bold",hjust = 0.5),
        plot.subtitle = element_text(face="italic",hjust = 0.5, size = 10),
        axis.title.x = element_text(size = 10),
        axis.title.y = element_text(size= 10))

#Create boxplot for female diabetes death rate by year
diabetes_female_plot = ggplot(df_diabetes, aes(x= as.character(year), y = diabetes_females, fill = year))+
  stat_boxplot(geom = "errorbar",
               width = 0.15)+
  geom_boxplot()+
  labs(title = "Female Diabetes Mellitus Death Rate \n in European Countries by Year",
       subtitle = "using Boxplot",
       x = "Year",
       y = "Death Rate")+
  theme(plot.title = element_text(face="bold",hjust = 0.5),
        plot.subtitle = element_text(face="italic",hjust = 0.5, size = 10),
        axis.title.x = element_text(size = 10),
        axis.title.y = element_text(size= 10))

#Show the male and female diabetes death rate plot side by side
grid.arrange(diabetes_male_plot, diabetes_female_plot, ncol=2)
```



In 2017, male death rate by diabetes has right skewed distribution, which is also shown in the numerical summary part, where the mean is 28.54 and median is 23.34. The spread of the middle distribution is between 18.65 and 38.48 (IQR range of 19.83) and it is ranging from 0.00 to 61.59 (range value of 61.59). Where in 2018, male death rate by diabetes has right skewed distribution, which is also shown in the numerical summary part, where the mean is 29.58 and median is 22.76. The spread of the middle distribution is between 17.67 and 40.71 (IQR range of 23.04) and it is ranging from 7.63 to 77.27 (range value of 69.64). And in 2019, male death rate by diabetes has right skewed distribution, which is also shown in the numerical summary part, where the mean is 30.31 and median is 24.04. The spread of the middle distribution is between 19.13 and 36.52 (IQR range of 17.39) and it is ranging from 7.3 to 106.97 (range value of 99.67).
There are one outlier located above the maximum value in 2018 distribution and one outlier located above the maximum value in 2019 distribution.

In 2017, female death rate by diabetes has right skewed distribution, which is also shown in the numerical summary part, where the mean is 22.11 and median is 16.29. The spread of the middle distribution is between 12.42 and 29.32 (IQR range of 16.90) and it is ranging from 0.00 to 54.39 (range value of 54.39). Where in 2018, female death rate by diabetes has right skewed distribution, which is also shown in the numerical summary part, where the mean is 22.32 and median is 16.68. The spread of the middle distribution is between 12.08 and 28.82 (IQR range of 16.74) and it is ranging from 0.00 to 63.92 (range value of 63.92). And in 2019, female death rate by diabetes has right skewed distribution, which is also shown in the numerical summary part, where the mean is 22.81 and median is 16.64. The spread of the middle distribution is between 13.96 and 28.55 (IQR range of 14.59) and it is ranging from 5.88 to 86.56 (range value of 80.68).
There are one outlier located above the maximum value in 2018 distribution and two outlier located above the maximum value in 2019 distribution.

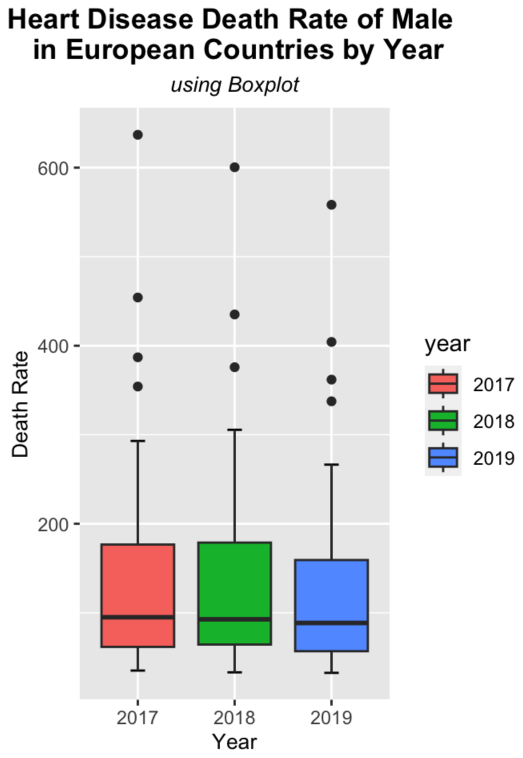
From the boxplots of diabetes mellitus death rate above, we can see in all the years, the diabetes mellitus death rate was higher among men than women, which is also shown in numerical analysis part.

Heart Disease Death Rate Bloxpot:

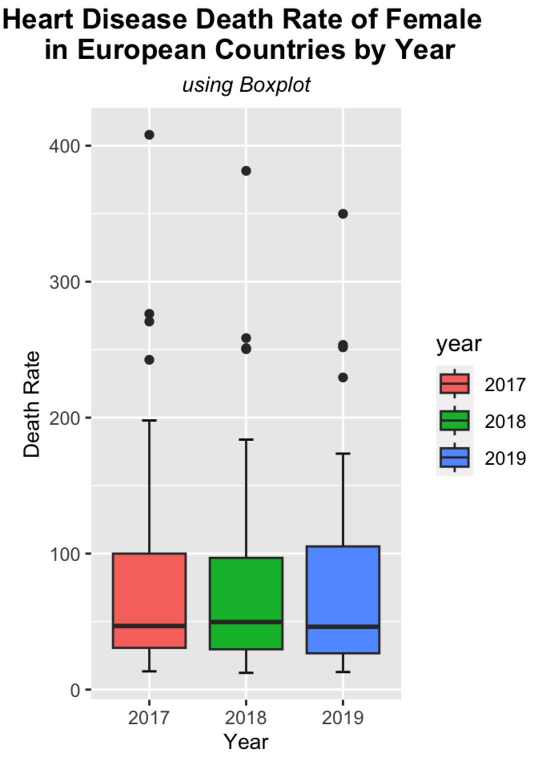
```
#Create boxplot for male heart disease death rate by year
heart_male_plot = ggplot(df_all, aes(x= as.character(year), y = heart_males, fill = year))+
  stat_boxplot(geom = "errorbar",
               width = 0.15)+
  geom_boxplot()+
  labs(title = "Heart Disease Death Rate of Male \n in European Countries by Year",
       subtitle = "using Boxplot",
       x = "Year",
       y = "Death Rate")+
  theme(plot.title = element_text(face="bold",hjust = 0.5),
        plot.subtitle = element_text(face="italic",hjust = 0.5, size = 10),
        axis.title.x = element_text(size = 10),
        axis.title.y = element_text(size= 10))

#Create boxplot for female heart disease death rate by year
heart_female_plot = ggplot(df_all, aes(x= as.character(year), y = heart_females, fill = year))+
  stat_boxplot(geom = "errorbar",
               width = 0.15)+
  geom_boxplot()+
  labs(title = "Heart Disease Death Rate of Female \n in European Countries by Year",
       subtitle = "using Boxplot",
       x = "Year",
       y = "Death Rate")+
  theme(plot.title = element_text(face="bold",hjust = 0.5),
        plot.subtitle = element_text(face="italic",hjust = 0.5, size = 10),
        axis.title.x = element_text(size = 10),
        axis.title.y = element_text(size= 10))

#Show the male and female heart disease death rate plot side by side
grid.arrange(heart_male_plot, heart_female_plot, ncol=2)
```


Heart Disease Death Rate of Male in European Countries by Year

using Boxplot


Heart Disease Death Rate of Female in European Countries by Year

using Boxplot

In 2017, male death rate by heart disease has right skewed distribution, which is also shown in the numerical summary part, where the mean is 149.83 and median is 95.03. The spread of the middle distribution is between 61.73 and 176.68 (IQR range of 114.95) and it is ranging from 35.13 to 636.88 (range value of 601.75). Where in 2018, male death rate by heart disease has right skewed distribution, which is also shown in the numerical summary part, where the mean is 145.74 and median is 92.71. The spread of the middle distribution is between 64.35 and 178.88 (IQR range of 114.53) and it is ranging from 33.12 to 600.43 (range value of 567.31). And in 2019, male death rate by heart disease has right skewed distribution, which is also shown in the numerical summary part, where the mean is 137.90 and median is 88.59. The spread of the middle distribution is between 56.89 and 159.27 (IQR range of 102.38) and it is ranging from 32.57 to 558.28 (range value of 525.71).
There are four outliers located above the maximum value in 2017 distribution, three outliers located above the maximum value in 2018 distribution, and four outliers located above the maximum value in 2019 distribution.

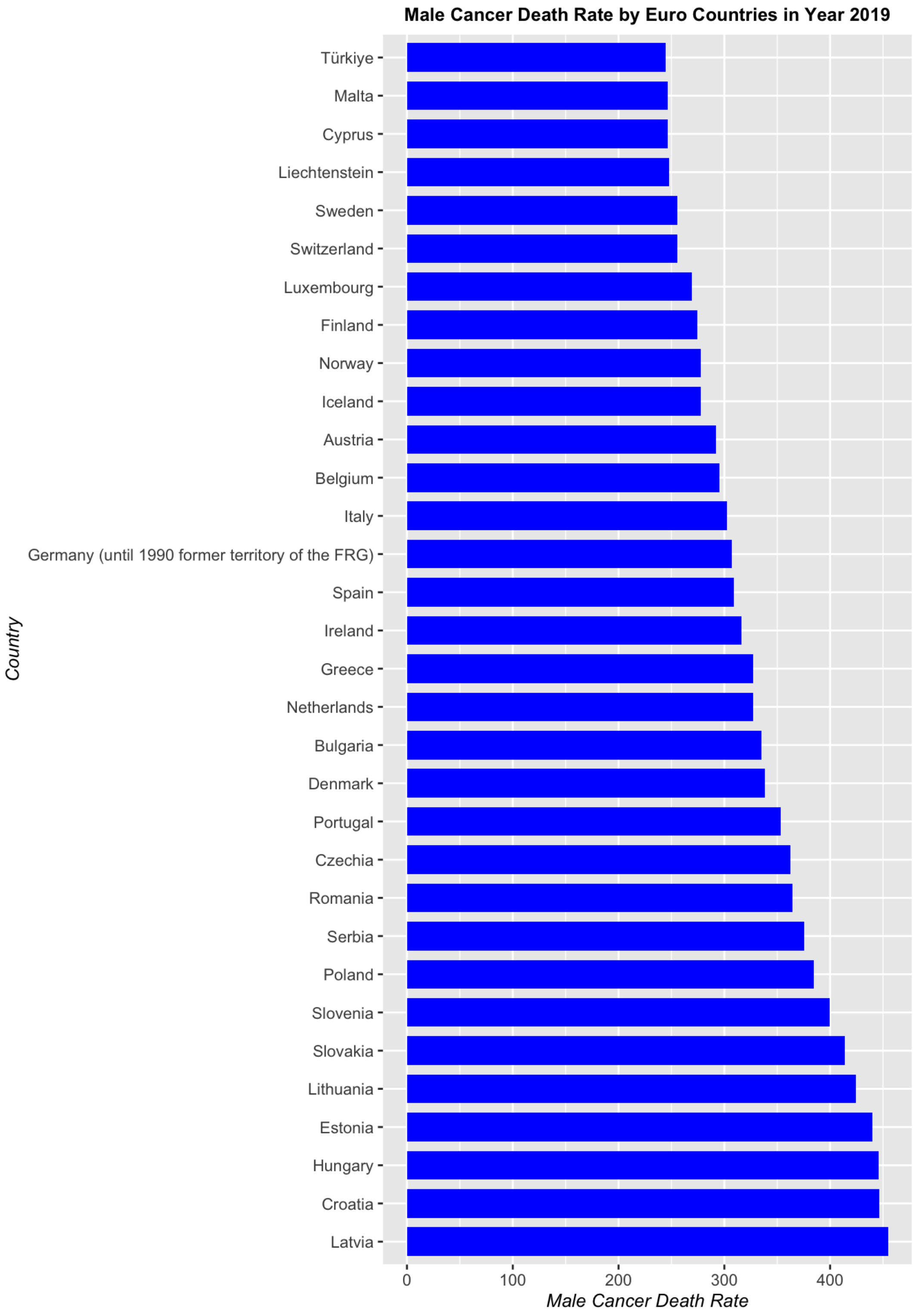
In 2017, female death rate by heart disease has right skewed distribution, which is also shown in the numerical summary part, where the mean is 91.17 and median is 46.79. The spread of the middle distribution is between 30.68 and 100.00 (IQR range of 69.32) and it is ranging from 13.42 to 408.00 (range value of 394.58). Where in 2018, female death rate by heart disease has right skewed distribution, which is also shown in the numerical summary part, where the mean is 88.34 and median is 49.71. The spread of the middle distribution is between 29.62 and 96.86 (IQR range of 67.24) and it is ranging from 12.31 to 381.51 (range value of 369.2). And in 2019, female death rate by heart disease has right skewed distribution, which is also shown in the numerical summary part, where the mean is 82.20 and median is 46.19. The spread of the middle distribution is between 26.69 and 105.29 (IQR range of 78.60) and it is ranging from 12.91 to 349.88 (range value of 336.97).
There are four outliers located above the maximum value in 2017 distribution, three outliers located above the maximum value in 2018 distribution, and four outliers located above the maximum value in 2019 distribution.

From the boxplots of Heart Disease death rate above, we can see in all the years, the Heart Disease Death rate was higher among men than women, which is also shown in numerical analysis part.

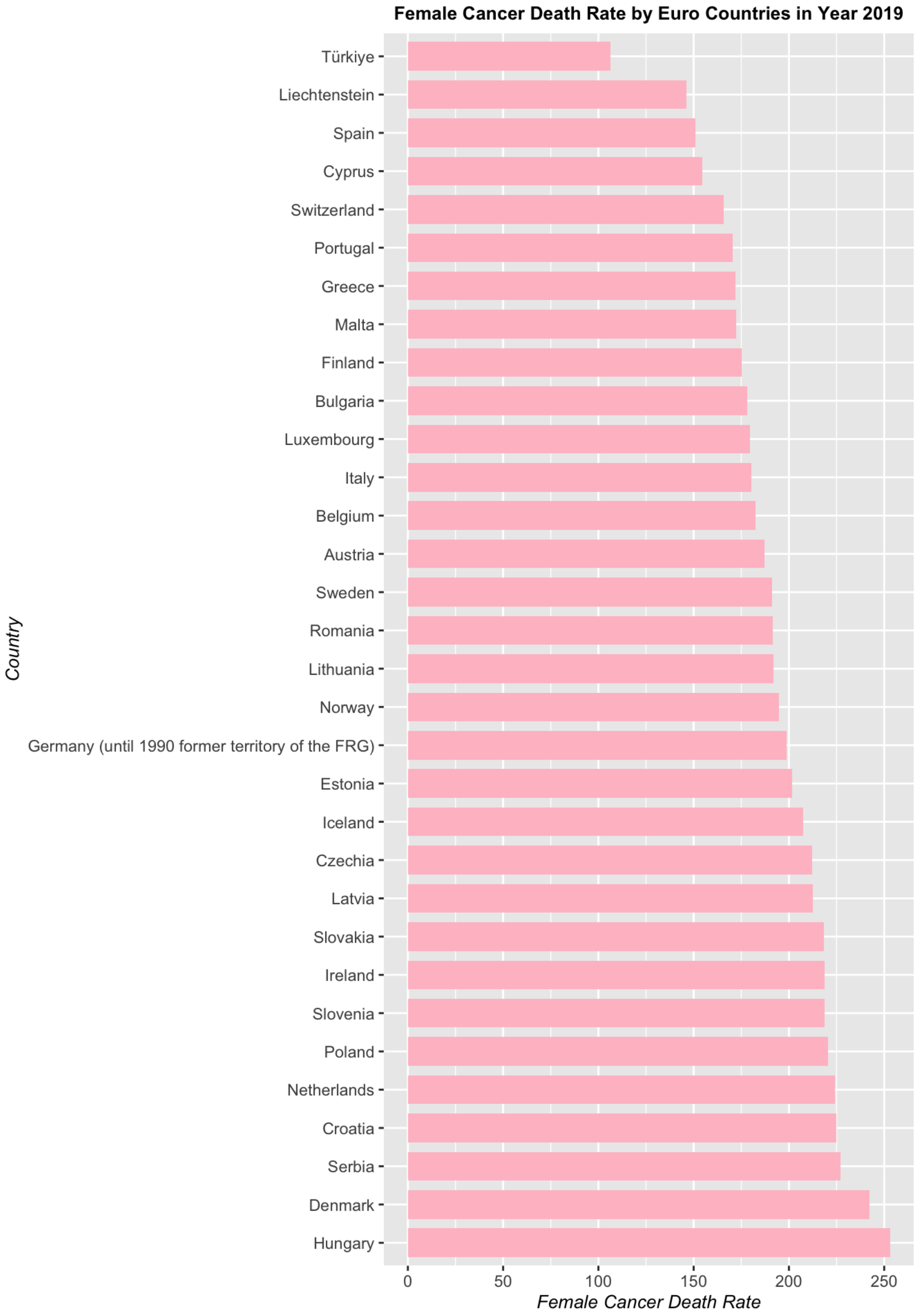
Barplot to show death rate in each Euro countries in Year 2019

In this part, we can see from the barplots below that:
- The highest death rate caused by cancer in 2019 for male was in Latvia, while for female was in Hungary.
- The highest death rate caused by diabetes mellitus in 2019 for both male and female were in Croatia.
- The highest death rate caused by heart disease in 2019 for both male and female were in Lithuania.

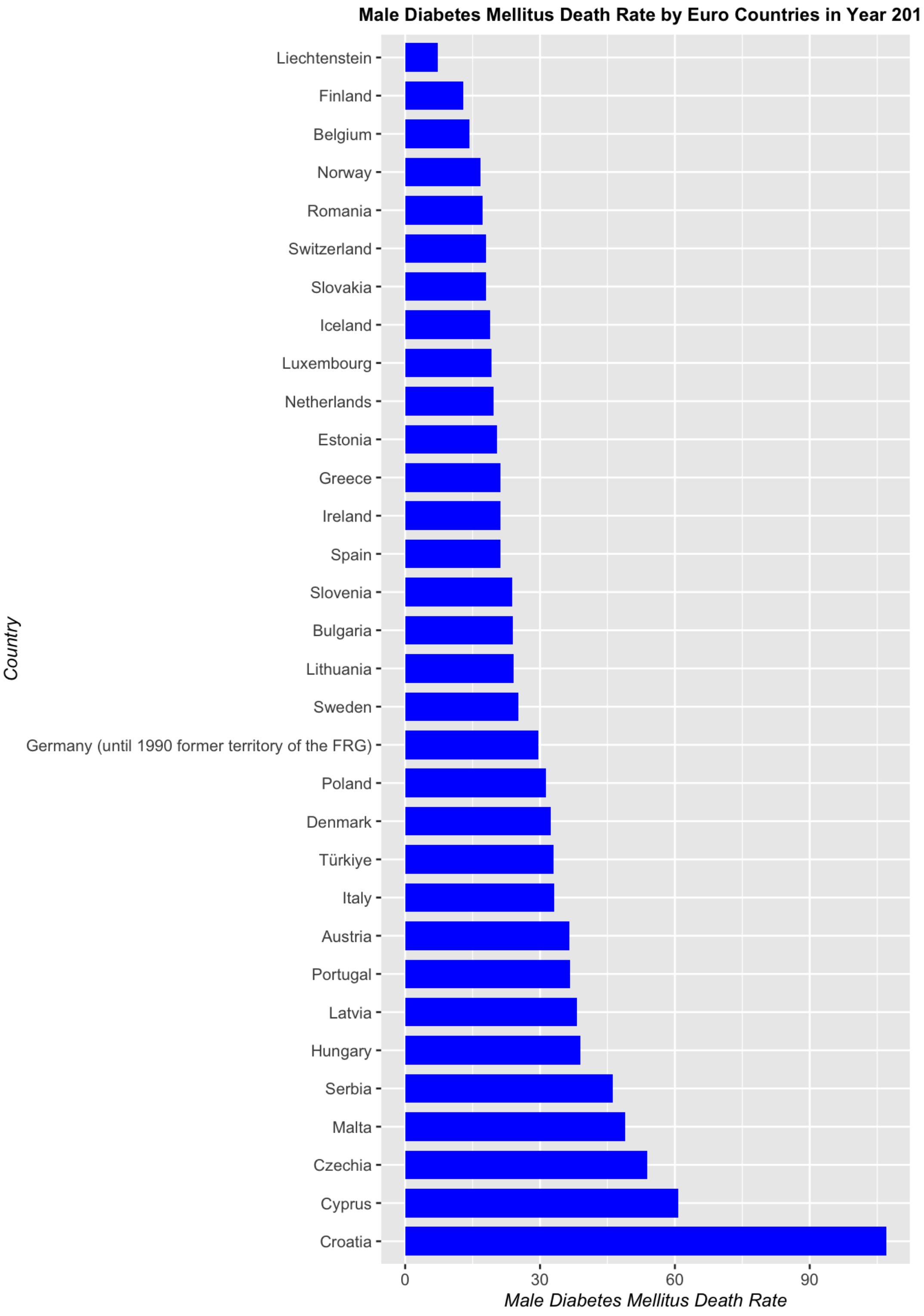
```
#First of all, I create dataset that contains only year 2019
df_2019 <- subset(df_all,year == "2019")
#I create a ascending barplot for cancer death rate male and female in 2019 with below code.
#Male cancer death rate barplot in 2019:
ggplot(data = df_2019, aes(x = reorder(country, -cancer_males),
                           y = cancer_males, fill = country)) +
  geom_bar(stat = "identity", width = 0.75,fill= "blue")+
  coord_flip() +
  labs(x = "Country",
       y = "Male Cancer Death Rate",
       title = "Male Cancer Death Rate by Euro Countries in Year 2019") +
  theme(plot.title = element_text(face="bold",hjust = 0.5, size= 10),
        axis.title.x = element_text(face="italic", size = 10),
        axis.title.y = element_text(face="italic", size = 10))
```



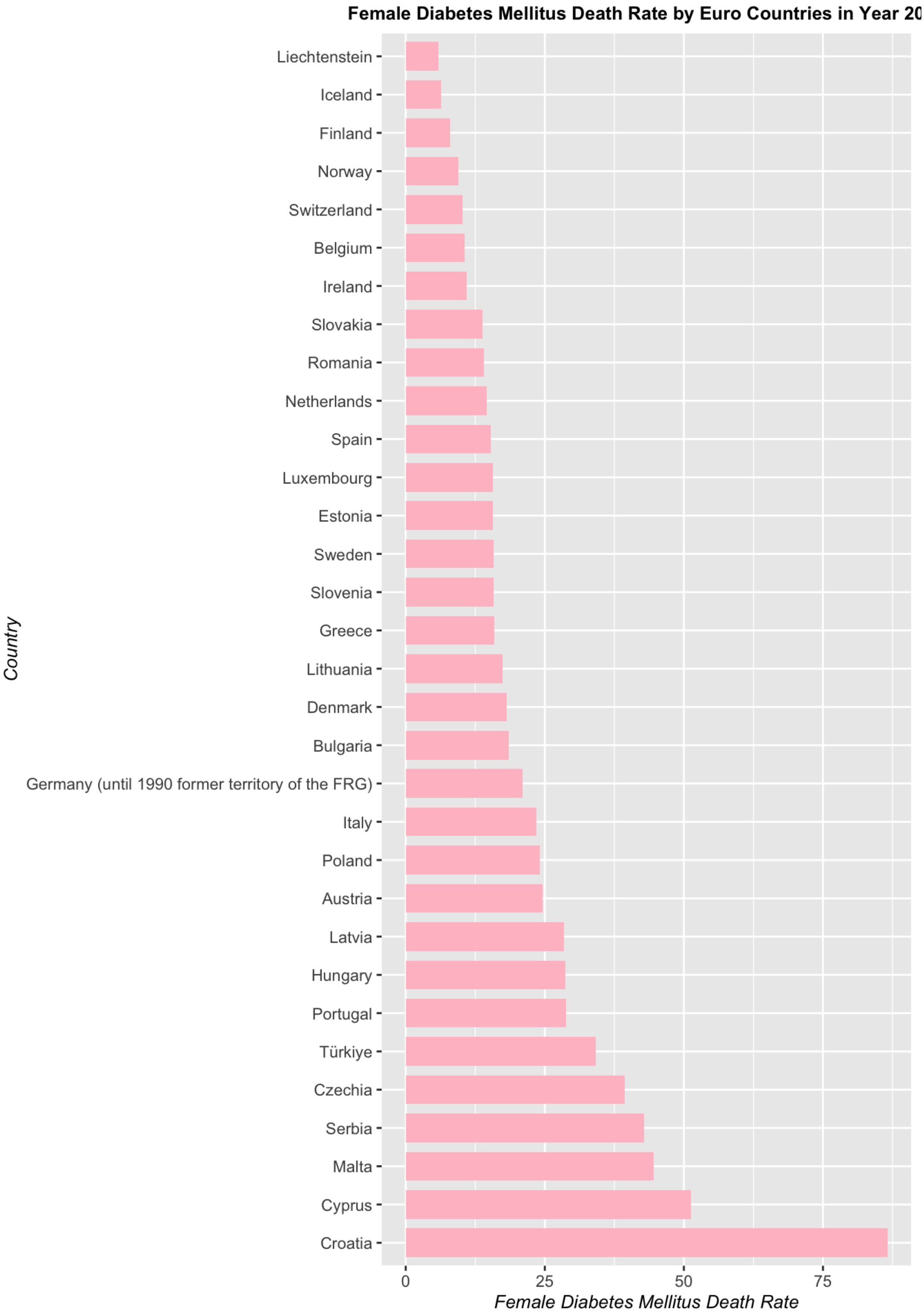
```
#Female cancer death rate barplot in 2019:
ggplot(data = df_2019, aes(x = reorder(country, -cancer_females),
                           y = cancer_females, fill = country)) +
  geom_bar(stat = "identity", width = 0.75,fill= "pink")+
  coord_flip() +
  labs(x = "Country",
       y = "Female Cancer Death Rate",
       title = "Female Cancer Death Rate by Euro Countries in Year 2019") +
  theme(plot.title = element_text(face="bold",hjust = 0.5, size= 10),
        axis.title.x = element_text(face="italic", size = 10),
        axis.title.y = element_text(face="italic", size = 10))
```



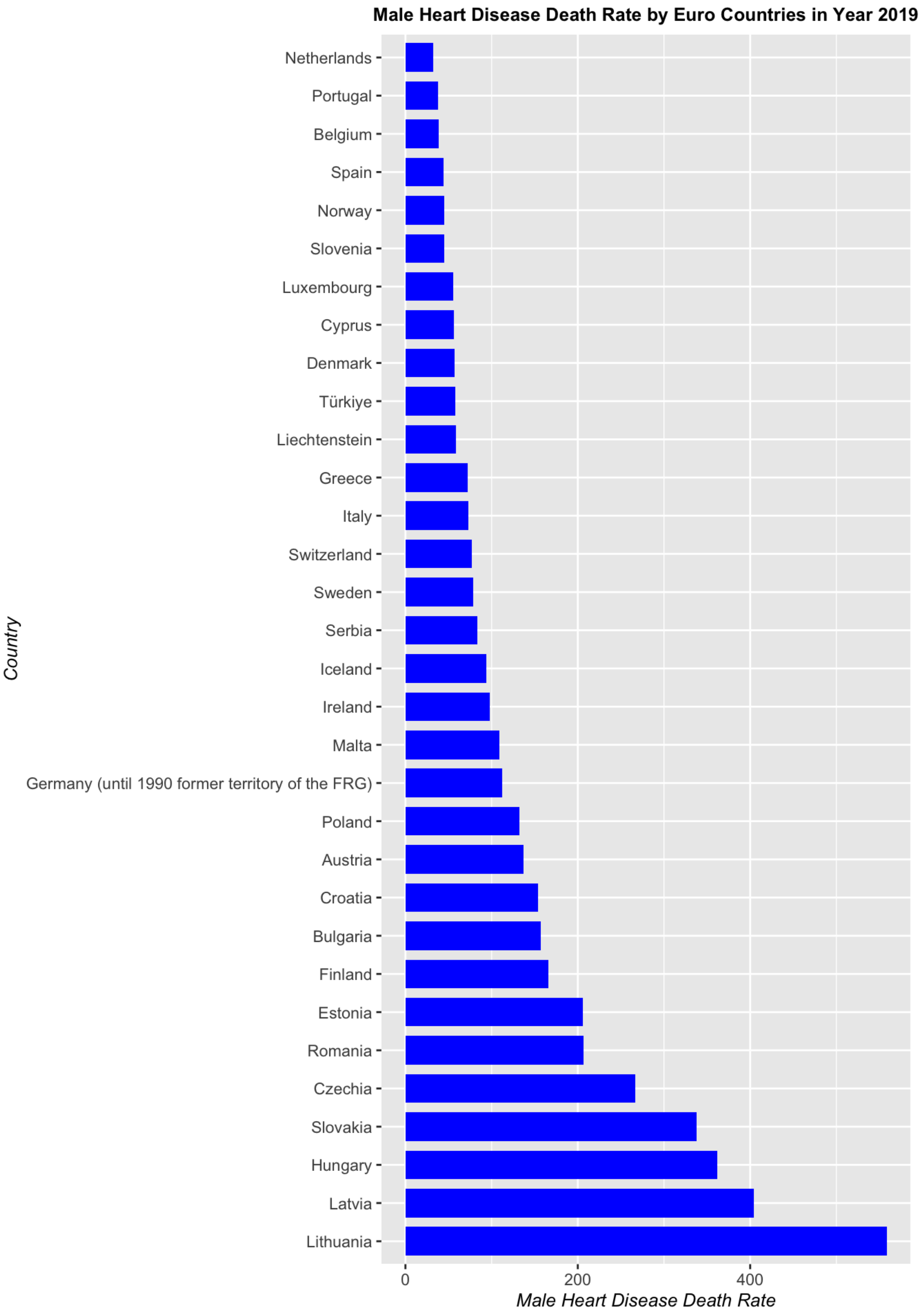
```
#I create a ascending barplot for diabetes death rate male and female in 2019 with below code.
#Male diabetes death rate barplot in 2019:
ggplot(data = df_2019, aes(x = reorder(country, -diabetes_males),
                           y = diabetes_males, fill = country)) +
  geom_bar(stat = "identity", width = 0.75,fill= "blue")+
  coord_flip() +
  labs(x = "Country",
       y = "Male Diabetes Mellitus Death Rate",
       title = "Male Diabetes Mellitus Death Rate by Euro Countries in Year 2019") +
  theme(plot.title = element_text(face="bold",hjust = 0.5, size= 10),
        axis.title.x = element_text(face="italic", size = 10),
        axis.title.y = element_text(face="italic", size = 10))
```



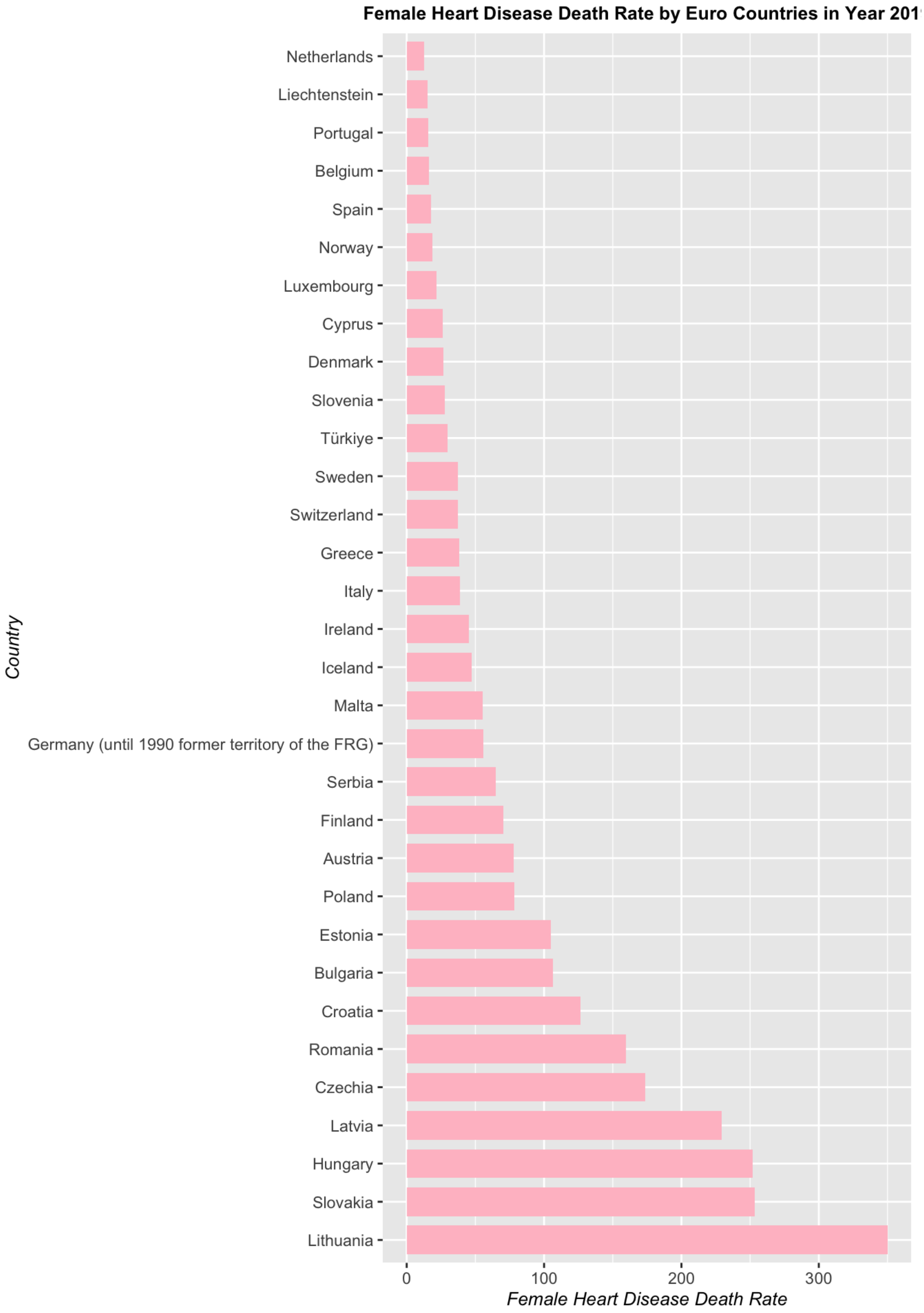
```
#Female diabetes death rate barplot in 2019:
ggplot(data = df_2019, aes(x = reorder(country, -diabetes_females),
                           y = diabetes_females, fill = country)) +
  geom_bar(stat = "identity", width = 0.75,fill= "pink")+
  coord_flip() +
  labs(x = "Country",
       y = "Female Diabetes Mellitus Death Rate",
       title = "Female Diabetes Mellitus Death Rate by Euro Countries in Year 2019") +
  theme(plot.title = element_text(face="bold",hjust = 0.5, size= 10),
        axis.title.x = element_text(face="italic", size = 10),
        axis.title.y = element_text(face="italic", size = 10))
```



```
#I create a ascending barplot for heart disease death rate male and female in 2019 with below code.
#Male heart disease death rate barplot in 2019:
ggplot(data = df_2019, aes(x = reorder(country, -heart_males),
                           y = heart_males, fill = country)) +
  geom_bar(stat = "identity", width = 0.75,fill= "blue")+
  coord_flip() +
  labs(x = "Country",
       y = "Male Heart Disease Death Rate",
       title = "Male Heart Disease Death Rate by Euro Countries in Year 2019") +
  theme(plot.title = element_text(face="bold",hjust = 0.5, size= 10),
        axis.title.x = element_text(face="italic", size = 10),
        axis.title.y = element_text(face="italic", size = 10))
```



```
#Female heart disease death rate barplot in 2019:
ggplot(data = df_2019, aes(x = reorder(country, -heart_females),
                           y = heart_females, fill = country)) +
  geom_bar(stat = "identity", width = 0.75,fill= "pink")+
  coord_flip() +
  labs(x = "Country",
       y = "Female Heart Disease Death Rate",
       title = "Female Heart Disease Death Rate by Euro Countries in Year 2019") +
  theme(plot.title = element_text(face="bold",hjust = 0.5, size= 10),
        axis.title.x = element_text(face="italic", size = 10),
        axis.title.y = element_text(face="italic", size = 10))
```



Correlation Test

In this part, I would like to create correlation test to see the relationship between two numerical variables in the dataset using Pearson method and identify which variables are strongly correlated.

Here, I list the four most strongly correlated pairs:

- Heart males death rate and heart females death rate have a strong positive correlation with correlation of 0.98.
- Diabetes males death rate and diabetes females death rate have a strong positive correlation with correlation of 0.97.
- Cancer males death rate and heart males death rate have a moderate positive correlation with correlation of 0.66.
- Cancer males death rate and cancer females death rate have a moderate positive correlation with correlation of 0.64.

```
#Create dataset for numerical variables only
df_numerical <- df_all %>%
  select(-country, -year)
#Standardized the numerical variables
df_numerical<- as.data.frame(scale(df_numerical))
```

```
#Show the correlation of the numerical variables in the dataset using Pearson Method
round(cor(df_numerical),
  digits = 2 # rounded to 2 decimals
)
```

```
##                  cancer_males cancer_females diabetes_males diabetes_females
## cancer_males             1.00           0.64           0.23             0.23
## cancer_females           0.64           1.00           0.15             0.05
## diabetes_males           0.23           0.15           1.00             0.97
## diabetes_females         0.23           0.05           0.97             1.00
## heart_males              0.66           0.31           0.06             0.07
## heart_females            0.70           0.34           0.12             0.13
##                  heart_males heart_females
## cancer_males            0.66          0.70
## cancer_females          0.31          0.34
## diabetes_males          0.06          0.12
## diabetes_females        0.07          0.13
## heart_males             1.00          0.98
## heart_females           0.98          1.00
```

Linear Regression

I create a linear regression model to predict the cancer death rate of males with the predictor variables as follow: cancer females, diabetes males, diabetes females, heart males, and heart females.

```
mod <- lm(cancer_males ~ ., data = df_numerical)
summary(mod)
```

```
## 
## Call:
## lm(formula = cancer_males ~ ., data = df_numerical)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -1.1602 -0.3317 -0.1323  0.1830  1.3085 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      1.235e-16  5.539e-02   0.000 1.000000    
## cancer_females   5.415e-01  6.787e-02   7.978 3.77e-12 ***
## diabetes_males  -7.612e-01  2.547e-01  -2.988 0.003590 ** 
## diabetes_females  8.789e-01  2.558e-01   3.436 0.000884 ***
## heart_males      1.599e-02  2.987e-01   0.054 0.957430    
## heart_females    4.692e-01  3.061e-01   1.533 0.128685    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.5511 on 93 degrees of freedom
## Multiple R-squared:  0.7118, Adjusted R-squared:  0.6963 
## F-statistic: 45.93 on 5 and 93 DF,  p-value: < 2.2e-16
```

Here, I interpret the coefficients that are shown in the figure above:

- The cancer females death rate coefficient in the regression equation is 5.415×10^{-1}. This coefficient represents the mean increase of cancer males death rate for every additional unit in cancer females death rate. If the cancer females death rate increases by 1 unit, the average cancer males death rate increases by 5.415×10^{-1} (holding all the other independent variables constant).
- The diabetes males death rate in the regression equation is -7.612×10^{-1}. This coefficient represents the mean decrease of cancer males death rate for every additional unit in diabetes males death rate. If the diabetes males death rate increases by 1 unit, the cancer males death rate decreases by 7.612×10^{-1} (holding all the other independent variables constant).
- The diabetes females death rate in the regression equation is 8.789×10^{-1}. This coefficient represents the mean increase of cancer males death rate for every additional unit in diabetes females. If the diabetes females death rate increases by 1 unit, the average cancer males death rate increases by 8.789×10^{-1} (holding all the other independent variables constant).
- The heart males death rate in the regression equation is 1.599×10^{-2}. This coefficient represents the mean increase of cancer males death rate for every additional unit in heart males. If the heart males death rate increases by 1 unit, the average cancer males death rate increases by 1.599×10^{-2} (holding all the other independent variables constant).
- The heart females death rate in the regression equation is 4.692×10^{-1}. This coefficient represents the mean increase of cancer males death rate for every additional unit in heart males. If the heart males death rate increases by 1 unit, the average cancer males death rate increases by 4.692×10^{-1} (holding all the other independent variables constant).
- When all the predictor measurements are equal to zero, the cancer males death rate is 1.235×10^{-16}.
- Then, I observe the p-value and compare it with significance level of 5%. We can see that cancer females death rate, diabetes males death rate, diabetes females death rate have p-value < 0.05. Thus, these variables are considered as significant predictor. This also indicates that relations exist between them with the cancer males death rate.
- Lastly, I observe the adjusted R^2 in this fitted model which is 0.6963. Thus, 69.63% of the variation in cancer males death ratecan be explained by fitted model.

Part 2: R Package

In this part, I find an existing R package, that we didn't use extensively in the course, and write a report demonstrating its use using R Markdown. The package that I choose is tidyr package.

The sole purpose of the tidyr package is to simplify the process of creating tidy data. Tidy data describes a standard way of storing data that is used wherever possible throughout the tidyverse.

```
#Download and install the tidyverse package
#install.packages("tidyr")
#Load the package
#library(tidyr) -> This has been done in the load package part
#Running the following code will give the description of the tidyr package functions that I use below
?gather
?separate
?unite
?spread
```

1. Gather function from tidyr package

gather() function: It takes multiple columns and gathers them into key-value pairs. Basically it makes "wide" data longer. The gather() function will take multiple columns and collapse them into key-value pairs, duplicating all other columns as needed.

```
#Using gather() function on df_cancer dataset
df_cancer_gather <- df_cancer %>%
  gather(Gender, Rate, cancer_males:cancer_females)

#Print the head and tail of the data in a long format
head(df_cancer_gather)
```

```
## # A tibble: 6 × 4
##   country  year  Gender        Rate
##   <chr>    <chr> <chr>        <dbl>
## 1 Belgium  2017  cancer_males  309.
## 2 Belgium  2018  cancer_males  300.
## 3 Belgium  2019  cancer_males  295
## 4 Bulgaria 2017  cancer_males  319
## 5 Bulgaria 2018  cancer_males  320.
## 6 Bulgaria 2019  cancer_males  335.
```

```
tail(df_cancer_gather)
```

```
## # A tibble: 6 × 4
##   country year  Gender          Rate
##   <chr>   <chr> <chr>          <dbl>
## 1 Serbia  2017  cancer_females  233.
## 2 Serbia  2018  cancer_females  233.
## 3 Serbia  2019  cancer_females  227
## 4 Türkiye 2017  cancer_females  122.
## 5 Türkiye 2018  cancer_females  123.
## 6 Türkiye 2019  cancer_females  106.
```

2. Separate function from tidyr package

separate() function: It converts longer data to a wider format and turns a single character column into multiple columns.

```
#Use separate() function to make data wider
df_cancer_separate <- df_cancer_gather %>%
            separate(Gender, c("Disease",
                              "Gender"))

#Print the head and tail of the data
head(df_cancer_separate)
```

```
## # A tibble: 6 × 5
##   country  year  Disease Gender  Rate
##   <chr>    <chr> <chr>   <chr>  <dbl>
## 1 Belgium  2017  cancer  males   309.
## 2 Belgium  2018  cancer  males   300.
## 3 Belgium  2019  cancer  males   295
## 4 Bulgaria 2017  cancer  males   319
## 5 Bulgaria 2018  cancer  males   320.
## 6 Bulgaria 2019  cancer  males   335.
```

```
tail(df_cancer_separate)
```

```
## # A tibble: 6 × 5
##   country year  Disease Gender   Rate
##   <chr>   <chr> <chr>   <chr>   <dbl>
## 1 Serbia  2017  cancer  females  233.
## 2 Serbia  2018  cancer  females  233.
## 3 Serbia  2019  cancer  females  227
## 4 Türkiye 2017  cancer  females  122.
## 5 Türkiye 2018  cancer  females  123.
## 6 Türkiye 2019  cancer  females  106.
```

3. Unite function from tidyr package

unite() function: It is the compliment of separate. To undo separate(), we can use unite(), which merges two variables into one. Here, we will merge two columns Disease and Gender with a separator “_”

```
#Use unite() function to glue Disease and Gender columns
df_cancer_unite <- df_cancer_separate %>%
            unite(Gender, Disease,
                  Gender, sep = "_")

#Print the head of the data
head(df_cancer_unite)
```

```
## # A tibble: 6 × 4
##   country  year  Gender        Rate
##   <chr>    <chr> <chr>        <dbl>
## 1 Belgium  2017  cancer_males  309.
## 2 Belgium  2018  cancer_males  300.
## 3 Belgium  2019  cancer_males  295
## 4 Bulgaria 2017  cancer_males  319
## 5 Bulgaria 2018  cancer_males  320.
## 6 Bulgaria 2019  cancer_males  335.
```

4. Spread function from tidyr package

spread() function: It helps in reshaping a longer format to a wider format. The spread() function spreads a key-value pair across multiple columns.

```
# use spread() function to make data wider
df_cancer_spread <- df_cancer_unite %>%
            spread(Gender, Rate)

#Print the head of the data
head(df_cancer_spread)
```

```
## # A tibble: 6 × 4
##   country year  cancer_females cancer_males
##   <chr>   <chr>          <dbl>        <dbl>
## 1 Austria 2017            187.         303.
## 2 Austria 2018            188.         299.
## 3 Austria 2019            187.         292.
## 4 Belgium 2017            189.         309.
## 5 Belgium 2018            182.         300.
## 6 Belgium 2019            183.         295
```

References

- Tidyr package in R programming. GeeksforGeeks. (2020, August 6). Retrieved December 20, 2022, from https://www.geeksforgeeks.org/tidyr-package-in-r-programming/ (https://www.geeksforgeeks.org/tidyr-package-in-r-programming/)

- Death Caused. Database - Eurostat. (n.d.). Retrieved December 20, 2022, from https://ec.europa.eu/eurostat/web/main/data/database (https://ec.europa.eu/eurostat/web/main/data/database)