**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Data Preparation and Visualization  Statistics for Data Analysis  Programming for Data Analysis  Machine Learning for Data Analysis |
| **Assessment Title:** | Prediction of Mortality Rates for Cardiovascular Disease, Cancer, Diabetes, and Chronic Respiratory Diseases in the Population of Ireland |
| **Lecturer Name:** | 1. David 2. Sam 3. Mohammed |
| **Student Full Name:** | DIANA FLORA NAMAEMBA |
| **Student Number:** | 2023385 |
| **Assessment Due Date:** | 12th November 2023 |
| **Date of Submission:** | 11th November 2023 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

**Contents**

[**1.0 INTRODUCTION** 2](#_Toc150417633)

[1.1 Background 2](#_Toc150417634)

[1.2 Problem Statement 3](#_Toc150417635)

[**2.0 METHOD** 3](#_Toc150417636)

[2.1 Data Selection 3](#_Toc150417637)

[2.2 Level of Measurements of the Data features 4](#_Toc150417638)

[2.3 Data Preparation and Visualization 4](#_Toc150417639)

[**2.3.1 Early Data Analysis (EDA)** 4](#_Toc150417640)

[**2.3.2 Data Visualization** 4](#_Toc150417641)

[**2.3.3 Data Preparation** 5](#_Toc150417642)

[**2.3.4 Data Visualization after Data Cleaning** 6](#_Toc150417643)

[**2.3.5 Programming in Data preparation and Visualization** 7](#_Toc150417644)

[2.4 Statistics 8](#_Toc150417645)

[**2.4.1 Descriptive Statistics** 8](#_Toc150417646)

[**2.4.2 Coefficient of variation** 8](#_Toc150417647)

[**2.4.3 Density Distribution** 9](#_Toc150417648)

[**2.4.4 Histogram** 9](#_Toc150417649)

[**2.4.5 Skewness and Kurtosis** 9](#_Toc150417650)

[**2.4.6 Cumulative Density Function** 10](#_Toc150417651)

[2.4.**3 Normal Distribution** 11](#_Toc150417652)

# **1.0 INTRODUCTION**

## 1.1 Background

Mortality rate is the number of deaths in a given population at a particular period with a particular disease as the underlying cause. The study of mortality rates is vital in public health research. Neoplasms, Diseases of the circulatory system, diseases of the respiratory system and external causes of injury and poisoning contribute to high mortality rates in Ireland. According to (*New research reveals where and how people die in Ireland - News & Events | Trinity College Dublin*, 2021) cancers which form part of neoplasms account for 31% of deaths, Diseases of the circulatory system account for 29% of deaths, Diseases of the respiratory system account for 13% of deaths and 4% of the deaths are as a result of external injuries and poisoning. These diseases are ranked as 2,9,10 and 19 respectively in the International Statistical Classification of Diseases and Related Health Problems (ICD)-10 disease classification.(*ICD-10 Version:2019*, no date) .

A neoplasm is an abnormal growth of tissues. Neoplasms can either be cancerous or benign(*ICD-10 - Wikipedia*, no date). Diseases of the circulatory system are diseases that affect the heart and blood vessels and make it hard for blood to flow throughout the body.(*Circulatory System Diseases: Risk Factors & Symptoms*, no date). Diseases of the respiratory system are diseases or disorders that affect the lungs and airways in turn affecting the human respiration.(*Respiratory disease | Definition, Causes, & Major Types | Britannica*, 2023). An injury is either a physical or physiological harm of the body that is caused by the body interacting with energy either in thermal, mechanical, electrical, chemical or radiant form or from extreme pressure in an amount that the body in either its physiological or physical form cannot tolerate. Lack of elements such as oxygen can also cause Injury. Poisoning on the other hand is damage to the body by toxins.(*22 - Injury, poisoning or certain other consequences of external causes - ICD-11 MMS*, 2023).

Modelling mortality is very important for the economy, life, demography and social insurance because mortality rates help determine various things such as insurance products prices, insurance liabilities’ etc.(Deprez, Shevchenko and Wüthrich, 2017) The use of machine learning models in modelling mortality has recently emerged. There has been application of various machine learning models, such as stochastic mortality models for estimation and forecasting the mortality rates.(Deprez, Shevchenko and Wüthrich, 2017). Machine learning models have also been used for mortality trend prediction. (‘11-23-22\_Mortality-trend-prediction-using-ML’, 2022). Machine learning techniques have also allowed for the study of the adequacy of the mortality rates that have been estimated.(Deprez, Shevchenko and Wüthrich, 2017) .Mortality rate modelling also helps assess the quality fit of the estimated mortality rates estimated using stochastic methods.

## 1.2 Problem Statement

The CSO OF Ireland collected 1892 mortality rate values for cardiovascular disease, cancer, diabetes or chronic respiratory diseases for the population of Ireland for 42 areas of residence. Certain Features of the data have been defined. The aim is to build a predictive model and find the mortality rate for each disease at a particular area of residence at a particular year using various machine learning approached and using past data.

# **2.0 METHOD**

## 2.1 Data Selection

The data used for this study was obtained from the Central Statistics Office (CSO) - Ireland. The Central Statistics Office (CSO) is Ireland's national statistical office and it impartially collects, analyzes and makes available statistics about Ireland’s people, society and economy. (*Population Changes - CSO - Central Statistics Office*, no date). The office houses a database with open access datasets for use. Data used was the Mortality rate attributed to cardiovascular disease, cancer, diabetes or chronic respiratory disease Dataset- (G0315). The data had 6 features and 1892 observations. The features included one continuous variable and 5 categorical variables. All the categories were nominal features.

## 2.2 Level of Measurements of the Data features

There was a total of 6 features but only 4 were used in this study

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Variable Type** | **Level of Measurement** | **Study Use (Y/N)** |
| Statistic Label | Independent variable | Nominal | N |
| Area of Residence | Independent variable | Nominal | Y |
| ICD 10 Diagnostic Group | Independent variable | Nominal | Y |
| Year | Independent variable | Nominal-Integer | Y |
| Unit | Independent variable | Nominal | N |
| Mortality Rate (per 1000) | dependent variable | Numerical | Y |

## 2.3 Data Preparation and Visualization

### **2.3.1 Early Data Analysis (EDA)**

Early data analysis was performed to better explore the data. The primary aim of the EDA was to examine the data’s distribution, outliers and any anomalies that would be used to generate specific hypothesis testing. (*Secondary Analysis of Electronic Health Records*, 2016). EDA was important because it aimed to assist in pattern recognitions. The steps involved in EDA included: -

1. Viewed the datasets head and tail, this displayed the top and bottom rows of the dataset giving an overview of the data contents and structure.(*head () and tail () Functions Explained with Examples and Codes*, 2023)
2. Checking the shape of the dataset, how many rows and columns the dataset had. The dataset had 1892 observations and 6 observations
3. Checking the dataset’s data types (continuous or discrete). The dataset had 5 categorical variables and only one continuous variable.
4. Calculate the summary statistics of both numeric and categorical variables.
5. Checking for any missing data and duplicates.
6. Checking for outliers

Ref: JupyterNotebook Line 1-12

### **2.3.2 Data Visualization**

Data visualization was then performed. This is because it’s the fastest way to learn about the data. (*OReilly.Media.Machine.Learning.and.Data.Science.Blueprints.for.Finance.1492073059*, 2020). Various data visualisation techniques were used. A pair plot was plotted because it helps understand the relationship between variables. There was no relationship between the continuous variables year and Mortality Rate. A heatmap was also preferred because it visualises the relationship between an independent variable and dependent variable.(*Secondary Analysis of Electronic Health Records*, 2016) . A heatmap for the correlation of the variables in the dataset was plotted. The results showed there was a weak negative correlation between year and mortality rate.

Ref. Python JupyterNotebook Line 13-14.

### **2.3.3 Data Preparation**

Data preparation involved data preprocessing. Data preprocessing steps included: Data cleaning, feature selection and data transformation.

(*OReilly.Media.Machine.Learning.and.Data.Science.Blueprints.for.Finance.1492073059*, 2020)

#### **2.3.3.1 Data Cleaning**

Data Cleaning involved checking the following: -

1. ***Validity***- This involved checking the data types and the data ranges. This involved using the python libraries pandas and NumPy. The data types for the four variables were integers and objects.
2. ***Completeness***- This involved checking the degree to which all the required data were known or available.
3. ***Uniformity***- This involved checking the degree to which the data was specified using the same units of measure.

(*OReilly.Media.Machine.Learning.and.Data.Science.Blueprints.for.Finance.1492073059*, 2020)

The data cleaning steps included: -

**Step 1: Handling missing data.**

Initially the data consisted of 1892 observations and 6 features. There was one feature mortality rate that had missing data for the whole year of 2017. The data was missing completely at random (MCAR) (Jakobsen *et al.*, 2017) therefore the missing data points were dropped because any other missing data handling technique would make the data incorrect because the mortality rate values were obtained for each year, each area of residence and for each disease. These process of handling missing data ensured the data was reliable, meaningful in analysis and most importantly unbiased(Kang, 2013). As a result, the data was reduced to 1720 observations and 6 features.

**Step 2: Removing features that were not used**

From the 6 features a total of 2 featured “statistic label” and “unit” were dropped because they were labels for 2 features in the dataset. The “area of residence” had a category called “state” that was an average of the mortality rate for each year for all area of residence for each disease. The ‘area of residence’ also had 8 provinces in the data, i.e., border, midland, west, Dublin, Mid-east, Mid-West, South-East, and South-West.(*cso ireland regions - Google Search*, no date) The provinces’ mortality rate values equalled the average of all the areas in that particular province.

The areas in each province area as shown below: -

* **Border** – Cavan, Donegal, Leitrim, Louth, Monaghan, Sligo.
* **Midland** – Laois, Longford, Offaly, Westmeath.
* **West** – Galway, Mayo, Roscommon.
* **Dublin** – Dublin City, Dún Laoghaire-Rathdown, Fingal, South Dublin.
* **Mid-East** – Kildare, Meath, Wicklow.
* **Mid-West** – Clare, Limerick, North Tipperary.
* **South-East** – Carlow, Kilkenny, South Tipperary, Waterford, Wexford.
* **South-West** – Cork, Kerry.

The provinces were therefore dropped because they skewed the data and were not meaningful for the analysis since the areas were already present in the data.

**Step 3: Encoding categorical data**

The dataset had two variables (“Area of Residence” and “The IDC-10 Diagnostic group”) that were categorical variables. The variables were encoded using the one-hot encoding method. The one-hot encoding was preferred because the categorical variables were nominal i.e., they did not have any order. Only 2 variables were encoded that is “area of residence” and “The ICD-10 Diagnostic group”.(Dahouda and Joe, 2021). A total of 38 new columns were created, changing the shape of the data to 1360 observations and 40 features.

#### **2.3.3.2 Feature selection**

Feature selection is the process of selecting features that are useful to the model. Multicollinearity was used to determine which features to keep. Features with a correlation coefficient of 0.95 and above would be removed.

(*OReilly.Media.Machine.Learning.and.Data.Science.Blueprints.for.Finance.1492073059*, 2020)

A heatmap for the correlation of the features was plotted. The variable Mortality rate and year were checked for collinearity. There was no relationship between the two variables hence both of them were included in the model.

#### **2.3.3.3 Data Transformation**

Data transformation is a data preparation technique that makes sure the data is in the best possible manner in the machine learning algorithms.

(*OReilly.Media.Machine.Learning.and.Data.Science.Blueprints.for.Finance.1492073059*, 2020).

Data transformation can be achieved through the following steps: -

1. ***Rescaling***- is rescaling the scale of all attributes if they are not of the same scale to the same scale.
2. ***Standardization***- is a technique that transforms attributes to standard normal distribution.
3. ***Normalization***- is rescaling the observations to have a length of one.

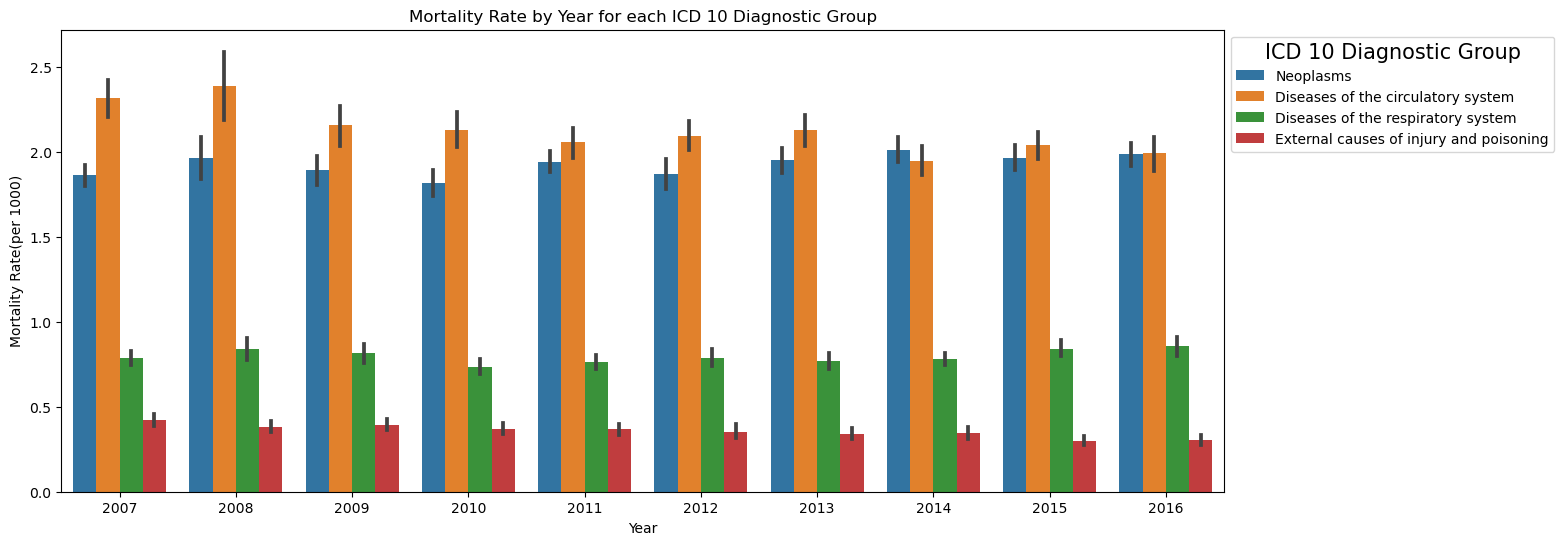
No data transformation technique was applied because of the following reasons: -

1. The Year variable was a nominal integer, scaling it would make the feature lose meaning.
2. The area of residence and ICD-10 diagnosis group were encoded using one hot-encoding making them either 0 or 1. Thus no need to scale them.

### **2.3.4 Data Visualization after Data Cleaning**

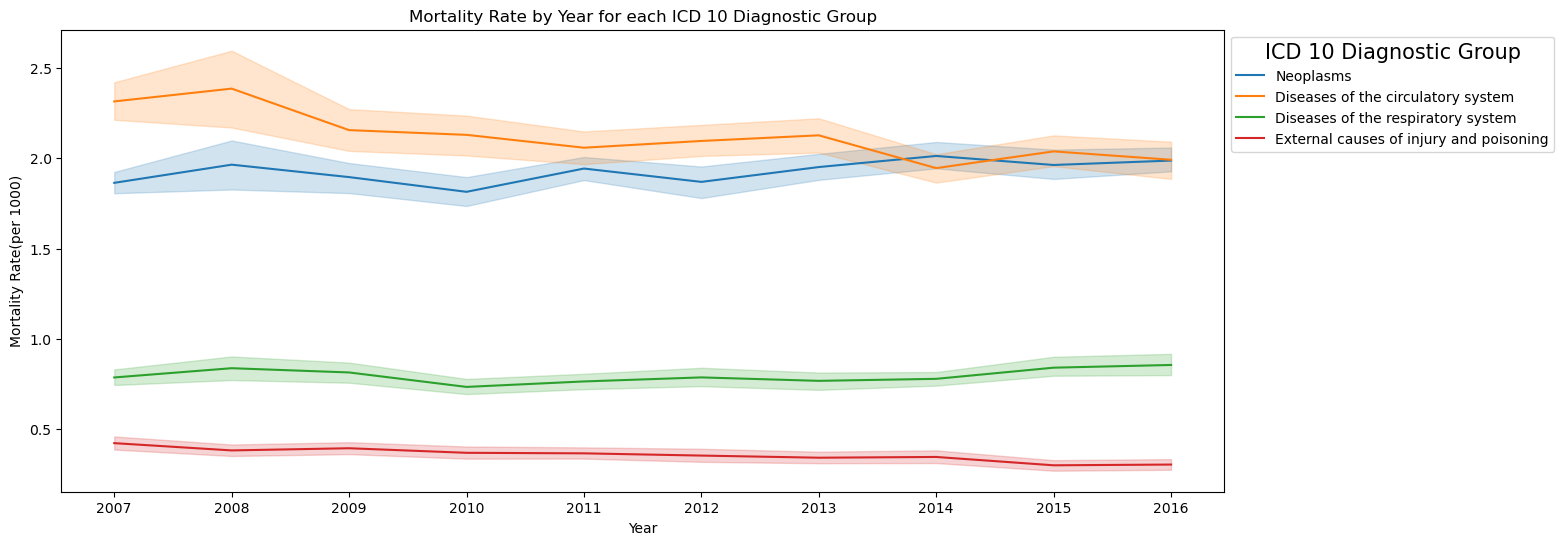
After cleaning the data several plots were made to answer the following questions: -

1. ***Which year had the highest mortality rate for each disease?***



The diseases of the circulatory system had the highest mortality rate over the 10 years while the External causes of injury and poisoning had the lowest mortality rate over the 10 years.

1. ***What is the mortality trend of each disease over the years?***



The mortality rate of the diseases of circulatory system and neoplasms were very high compared to diseases of the respiratory system and external causes of injury and poisoning. The mortality rate of the diseases of the circulatory system is decreasing over the years. The mortality rate of neoplasms increases over the years. The mortality rate of diseases of the respiratory system decreases then increases over the years. The mortality rate of external causes of injury and poisoning decreases over the years

### **2.3.5 Programming in Data preparation and Visualization**

**Programming: :** (Gra ded out of 100)

1. The project must be explored programmatically, this means that you must implement suitable Python tools (code and/or libraries) to complete the analysis required. All of this is to be implemented in a Jupyter Notebook. Your codebook should be properly annotated. The project documentation must include sound justifications and explanation of your code choices (code quality standards should also be applied). **[0-50]**

**Please recall that simply performing the analyses is a requirement to achieve a grade of PASS. Critical analysis and independent research are required for higher marks.**

1. Briefly discuss your use of aspects of various programming paradigms in the development of your project. For example, this may include (but is not limited to) how they influenced your design decisions or how they helped you solve problems. Note that marks may not be awarded if the discussion does not involve your specific project. **[0-50]**

## Statistics

Descriptive and predictive statistics were used in this study. Descriptive was used to understand the past data while predictive analysis was used to make predictions of the mortality rate for the four diseases.

### **2.4.1 Descriptive Statistics**

Descriptive statistics included calculating the measures of central tendencies: (mean, median and frequencies) and the measures of dispersion (Variance, standard deviation, kurtosis) for the mortality rate of each disease (Neoplasm, Diseases of the circulatory system, Diseases of the respiratory system and External causes of injury and poisoning).

**Mortality rate by ICD Diagnostic group**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ICD-10 Diagnostic group** |  | **Mortality rate (per 1000)** | | | |
| **n (Mean)** | **Median (SD)** | **Variance** | **IQR** | **Min-Max** |
| Diseases of the circulatory system | 420(2.12) | 2.11(0.35) | 0.13 | 1.93-2.30 | 1.00-3.92 |
| Diseases of the respiratory system | 420(0.80) | 0.79(0.14) | 0.02 | 0.71-0.87 | 0.35-1.48 |
| External causes of Injury and Poisoning | 420(0.36) | 0.35(0.10) | 0.01 | 0.29-0.41 | 0.10-0.75 |
| Neoplasms | 420(1.92) | 1.91(0.23) | 0.06 | 1.79-2.05 | 0.93-2.73 |

The diseases of the circulatory system had the highest mean mortality rate compared to the other diseases. The external causes of injury and poisoning had the lowest mean mortality rate compared to the other diseases. The measures of central tendency and dispersion were also calculated for the mortality rate of each disease by area of residence and mortality rate of each disease by year. Ref: python JupyterNotebook line 43-45.

### **2.4.2 Coefficient of variation**

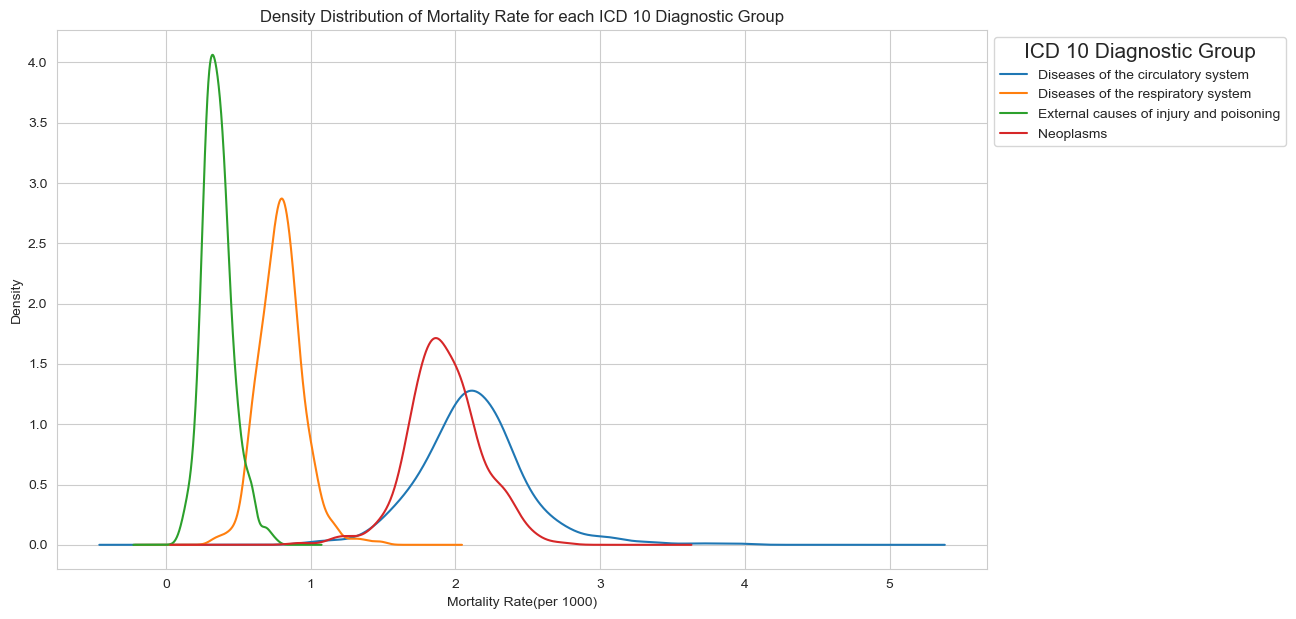
The coefficient of variation (CV) is the relative measure of variability for the size of the standard deviation in relation to its mean. (*Coefficient of Variation in Statistics - Statistics By Jim*, 2020). CV was calculated for the mortality rate of each disease.

|  |  |
| --- | --- |
| **Disease** | **CV (%)** |
| Disease of the circulatory system | 13.66 |
| Disease of the respiratory system | 6.71 |
| External cause of injury and poisoning | 25.60 |
| Neoplasm | 10.14 |

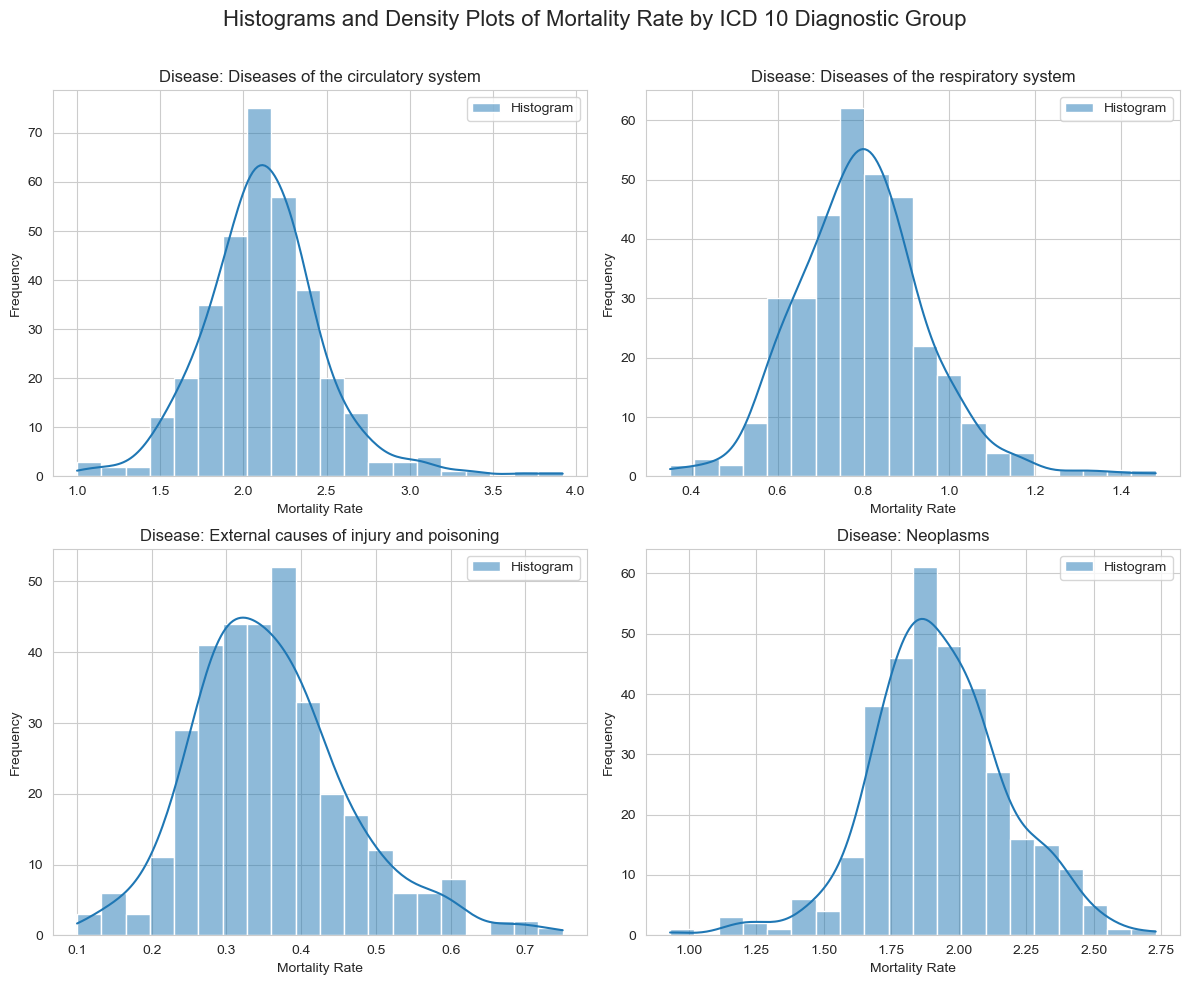
The CV of the four diseases is less than 100 this means that the SD is less than the mean, meaning there is low variability in the dataset and this is acceptable.

### **2.4.3 Density Distribution**

A density distribution and histogram were plotted to check the spread of the data. The density plots showed that the data had low variability and hence were not skewed.



### **2.4.4 Histogram**

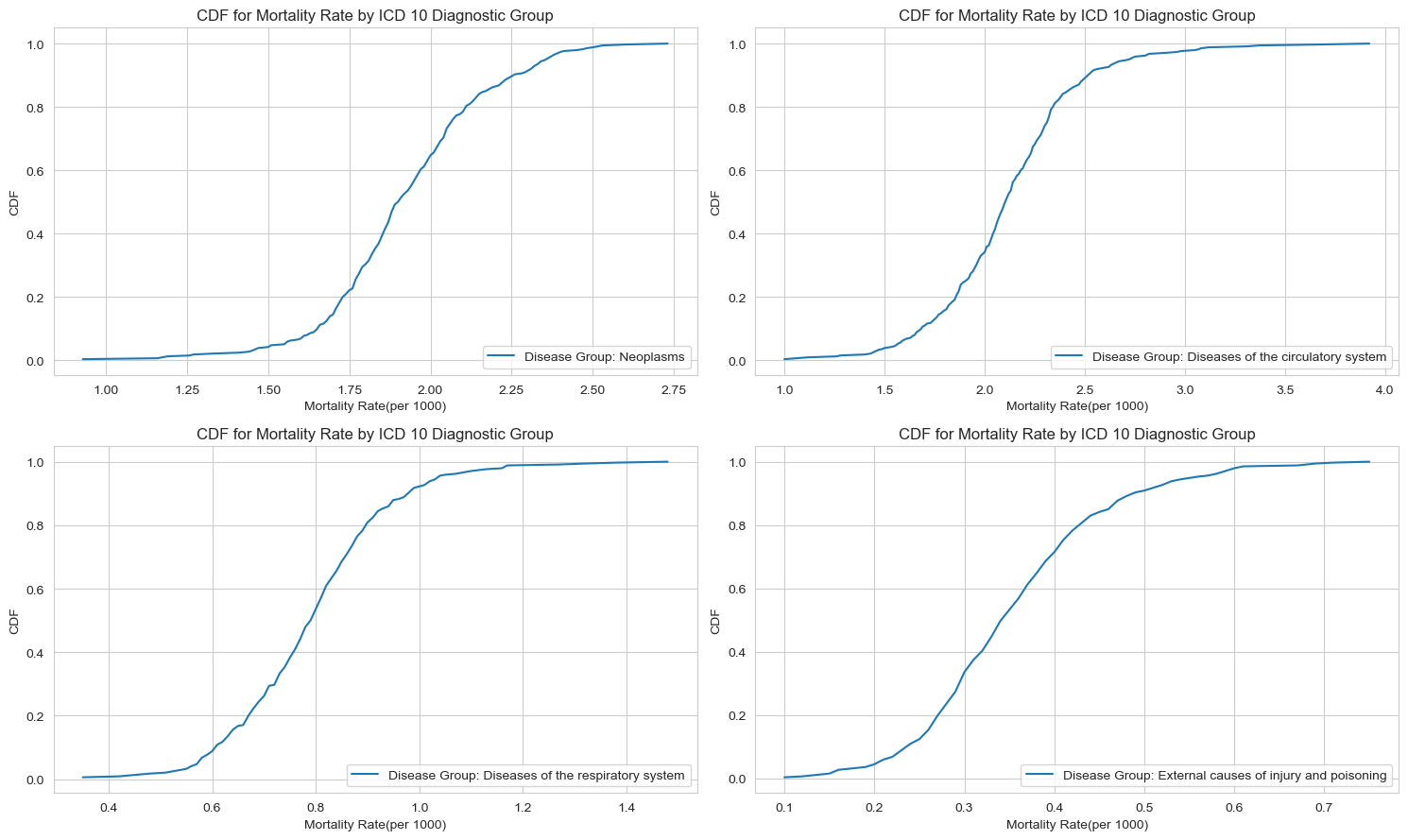
A histogram was also plotted to check the spread of the mortality rate for the four diseases. The plots showed the data was not skewed.

### 

### **2.4.5 Skewness and Kurtosis**

|  |  |  |
| --- | --- | --- |
| **Disease** | **Skewness** | **Kurtosis** |
| Disease of the circulatory system | 0.70 | 2.98 |
| Disease of the respiratory system | 0.55 | 1.88 |
| External cause of injury and poisoning | 0.67 | 0.95 |
| Neoplasm | -0.09 | 1.05 |

### **2.4.6 Cumulative Density Function**

****A CDF was plotted to visualize the shape of the distribution of the mortality rate for each disease

**2.4.7 Normality Tests (Normal Distribution)**

Normality test was done using Q-Q plots and the Kolmogorov-Smirnov test to test if the mortality rate variable for each disease was normally distributed.

**2.4.7.1 Kolmogorov-Smirnov Test**

Kolmogorov-Smirnov test was used to test for normality using p-values. It was used because it is better for sample sizes greater than or equals to 50 (Mishra *et al.*, 2019).

Set the hypothesis to test for normality:

H0: The Mortality rate variable for each disease does not follow a normal distribution

H1: The mortality rate variable for each disease follows a normal distribution

|  |  |  |  |
| --- | --- | --- | --- |
| Disease | **Test-Statistic** | **P-Value** | **Decision Rule** |
| Disease of the circulatory system | 0.91 | 0.0 | The Mortality rate variable follow a normal distribution (reject H0) |
| Disease of the respiratory system | 0.68 | 5.05e-159 | The Mortality rate variable follow a normal distribution (reject H0) |
| External cause of injury and poisoning | 0.55 | 1.13e-96 | The Mortality rate variable follow a normal distribution (reject H0) |
| Neoplasm | 0.90 | 0.0 | The Mortality rate variable follow a normal distribution (reject H0) |

**Programming: :** (Graded out of 100)

1. The project must be explored programmatically, this means that you must implement suitable Python tools (code and/or libraries) to complete the analysis required. All of this is to be implemented in a Jupyter Notebook. Your codebook should be properly annotated. The project documentation must include sound justifications and explanation of your code choices (code quality standards should also be applied). **[0-50]**

**Please recall that simply performing the analyses is a requirement to achieve a grade of PASS. Critical analysis and independent research are required for higher marks.**

1. Briefly discuss your use of aspects of various programming paradigms in the development of your project. For example, this may include (but is not limited to) how they influenced your design decisions or how they helped you solve problems. Note that marks may not be awarded if the discussion does not involve your specific project. **[0-50]**

## 2.5 Machine learning for Data Analytics

Machine learning involved consideration of three project management frameworks: CRISP-DM, KDD and SEMMA. Knowledge discovery in databases (KDD) is the process of selecting the target data, pre-processing, transforming, data mining and interpreting it. The sample, explore, modify, model and assess (SEMMA) is the process of conducting a data mining project by sampling the data, exploring the data, modifying the data, modelling the data and assessing the data by evaluating the results. The Cross industry standard process for data mining (CRISP-DM) is a process that uses business understanding, data understanding, data preparation, data modelling, evaluation and deployment of the results. (Martins, Pesado and García-Martínez, 2016).

The project management framework used for data science projects is the CRISP-DM. The CRISP-DM framework was applied in a study on Data Quality Improvement to Support Machine Learning of Stunting Prediction in Infants and Toddlers.(Purbasari *et al.*, 2021)

Supervised machine learning technique was used because our data was labelled data. Secondly majority of work done or research done using mortality rate data has leaned towards supervised learning as the best approach for modelling mortality.(Saroj *et al.*, 2022), (Deprez, Shevchenko and Wüthrich, 2017; Krittanawong *et al.*, 2021; ‘11-23-22\_Mortality-trend-prediction-using-ML’, 2022)

**Machine learning process involved the following steps: -**

1. Splitting the dataset.
2. Supervised ML Model selection.
3. Making predictions

### **2.5.1 Data splitting**

The pop\_data was split into independent variables called X and the dependent variable called y. The X and y variables were then split into Training and test sets as shown below: -

Training set: X\_train and y\_train which includes 70% of the X data and y data respectively.

Test set: X\_test and y\_test which include 30% of the X data and y data respectively.

Data Splitting in training and test set was very important because it helped find the most efficient set of model parameters that had the correct balance between the model complexity and the model’s generalization capabilities.(Eliane Birba, 2020).

**2.5.2 Supervised Machine Learning Algorithms Selection**

Regression algorithms were selected because the target variable/ independent variable i.e., Mortality rate was continuous in nature.(Shetty *et al.*, 2022). The following Regression models were applied:-

* ***K-Nearest Neighbors regression-***
* ***Decision Tree Regression***
* ***Random Forest Regression***
* ***Linear Regression***
* ***Ridge Regression***
* ***Ridge Regression (gridSearchCV)***
* ***Lasso Regression***
* ***Lasso Regression (gridSearchCV)***
* ***Support Vector Machine***
* ***Support Vector Machine (gridsearchCV)***

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ML Algorithm** | R-squared | | **Data(n)** | | **Training Set (n)** | | Test Set (n) | |
|  | **Training set** | **Test set** | **X** | **Y** | **X\_train** | **Y\_train** | **X\_test** | Y\_test |
| K-Nearest Neighbors regression | 0.92 | 0.90 | 1360 observations  39 features | 1360 observations  1 feature | 952 observations  39 features | 952 observations  1 feature | 408 observations  39 features | 408 observations  1 feature |
| Decision Tree Regression | 0.89 | 0.91 |
| Random Forest Regression | 0.99 | 0.94 |
| Multiple Linear Regression | 0.93 | 0.94 |
| Ridge Regression | 0.93 | 0.94 |
| Ridge Regression(alpha=10) | 0.93 | 0.93 |
| Ridge Regression(alpha=0.1) | 0.93 | 0.93 |
| Ridge Regression (gridSearchCV) | 0.93 | 0.94 |
| Lasso Regression | 0.00 | -0.01 |
| Lasso Regression(alpha=0.01) | 0.90 | 0.91 |
| Lasso Regression  (alpha=0.0001) | 0.93 | 0.94 |
| Lasso Regression (gridSearchCV) | 0.90 | 0.91 |
| Support vector Machine | -0.02 | -0.05 |
| Support Vector Machine (gridsearchCV) |  | 0.95 |

**Programming: :** (Graded out of 100)

1. The project must be explored programmatically, this means that you must implement suitable Python tools (code and/or libraries) to complete the analysis required. All of this is to be implemented in a Jupyter Notebook. Your codebook should be properly annotated. The project documentation must include sound justifications and explanation of your code choices (code quality standards should also be applied). **[0-50]**

**Please recall that simply performing the analyses is a requirement to achieve a grade of PASS. Critical analysis and independent research are required for higher marks.**

1. Briefly discuss your use of aspects of various programming paradigms in the development of your project. For example, this may include (but is not limited to) how they influenced your design decisions or how they helped you solve problems. Note that marks may not be awarded if the discussion does not involve your specific project. **[0-50]**