**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | *Programming for DA*  *Statistics for Data Analytics*  *Machine Learning for Data Analysis*  *Data Preparation & Visualisation* |
| **Assessment Title:** | MSC\_DA\_CA1 |
| **Lecturer Name:** | Sam Weiss/ David Gonzalez  Taufique Ahmed  Muhammad Iqbal  David McQuaid |
| **Student Full Name:** | Ashish Royan Sanchís  <https://github.com/AshishSanctis/CA1_IMMIGRATION-ANALYSIS> |
| **Student Number:** | 2023444 |
| **Assessment Due Date:** | 12/11/2023 |
| **Date of Submission:** | 14/11/2023 |

****

**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. |

# Introduction

#### The Immigration of people from diverse nations significantly contributes to the economy and its expansion. As more people migrate towards a country, there is a corresponding increase in the country’s population. However, this poses a challenge with limited infrastructure capacity to accommodate the newcomers, This also impacts the public transport systems. Ireland, in particular has experienced a notable increase in the migrants recent times, Consequently altering the overall landscapes to address housing demand, strain on infrastructure, and provision of social service.

#### My objective is to determine the number of Immigrant's that have relocated to Ireland from neighbouring European countries and global nations. This involves analysing immigration datasets from 1996-2022. I have chosen to utilise Machine Learning models to predict upcoming trends in this migration pattern.

I’ve abstracted data from “Estimated Immigration(Persons in April).csv” file to analyse the immigration trends in Ireland. This Dataset serves as the basis for training and testing machine learning models, specifically for predicting migrant patterns. The data was sourced from the central statistics office, Ireland website(Central Statistics office, n.d.). for the development I made use Python programming language, Jupyter Notebook, and the Anaconda environment. The analysis and visualisation of insights from the data were facilitated by well-known Python libraries such as Pandas, NumPy, Matplotlib, Seaborn,Scikit-learn,Scipy among others.

#### The Immigration analysis and prediction, following tasks will be cin

## Importing Library

#### Here I imported some useful Python libraries for data analytics. After importing I named it as a short name like pandas as pd, NumPy as np, and Matplotlib as plt because that will be short and more convenient to use throughout the program and pd, np, and plt are recommended naming conventions.

##### pandas (Pandas, 2023).

##### Numpy (NumPy, 2023).

##### Matplotlib (Matplotlib, 2023).

##### seaborn (Waskom., 2023).

##### Sklearn (developers, 2023).

##### SciPy (community, 2023).

##### sklearn.linear\_model (developers, 2023).

##### sklearn.tree (developers, 2023).

##### sklearn.metrics (developers, 2023).

##### sklearn.ensemble (developers, 2023).

##### sklearn.model\_selection (developers, 2023).

##### statsmodels.tools.eval\_measures (developers, 2023).

##### sklearnks (developers, 2023).

##### warnings (Foundation, 2001-2023).

##### scikit-learn (developers, 2023)

#### All these libraries and modules are commonly used in data analytics and machine learning workflows to load, process, and visualize data, as well as to build and evaluate ML models.

## Pandas’ library to modify display options.

#### Here, I have used the *set\_option()* function from the panda's library to modify display options for rows and columns when viewing DataFrames. It sets the option to None which display all column, and all rows present in the data frame. By default, it shows a limited number of rows and columns which makes it difficult to view all observations and features at once.

## Ignore the warning message.

#### Here in this section, I have defined a function called *ignore warning ()* using Python that filters out (ignores) warning messages using *warnings.filterwarnings()* function. Also, *action=’ignore’* is a function call that modifies the behaviours of the Python warning module and sets the action to ignore meaning that warning messages will be suppressed and not displayed when encountered.pasted-image.png

Figure 1: loading the dataset into Jupyter Notebook

#### The dataset shows the following columns Statistic label, Year, Sex, Nationality, unit and Value. The data frame has been renamed to Emig.

## Data Visualisation and Preparation.(Exploratory Data Analysis)

A crucial step in the data analysis process is data visualisation, which involves displaying data graphically to identify trends, relationships, and other patterns. Choosing the appropriate visualisation techniques to match the goals and nature of the data is just as important to effective data visualisation as designing visually appealing charts and graphs. There are many different types of visualisations that can be used, such as histograms for data distribution, scatter plots for relationships between variables, bar charts for comparisons, line charts for trends over time, and heatmaps for pattern recognition. Data visualisation is an integral part of the data analysis process because its ultimate goal is to support data-driven decision-making and efficient dissemination of insights to stakeholders.

#### The data exploration is crucial phase in the data analysis journey for both visual and statistical examination. This helps us to understand and identify data, extract valuable insights and obtain fundamental statistic information about the data I have made use of some basic functions like shape(), head(), tail(), describe(), and info() to gain and understand the dataset.

### Shape

#### Here, ‘Emig*.shape’* is used to retrieve the dimensions (number of rows and columns). The result is a tuple where the first element is the number of rows i.e., 486 and the second element represents the number of columns i.e. 6.

### Head

#### The head() method in Pandas is commonly used to quickly inspect the initial rows of the dataframe and get a sense of its structure and content.

### Tail

#### *Tail* is used to display the last 5 rows of the data frame. If we don’t pass any value inside the tail or head, it will display 5 observations by default.

### Rename column names and dropping the columns

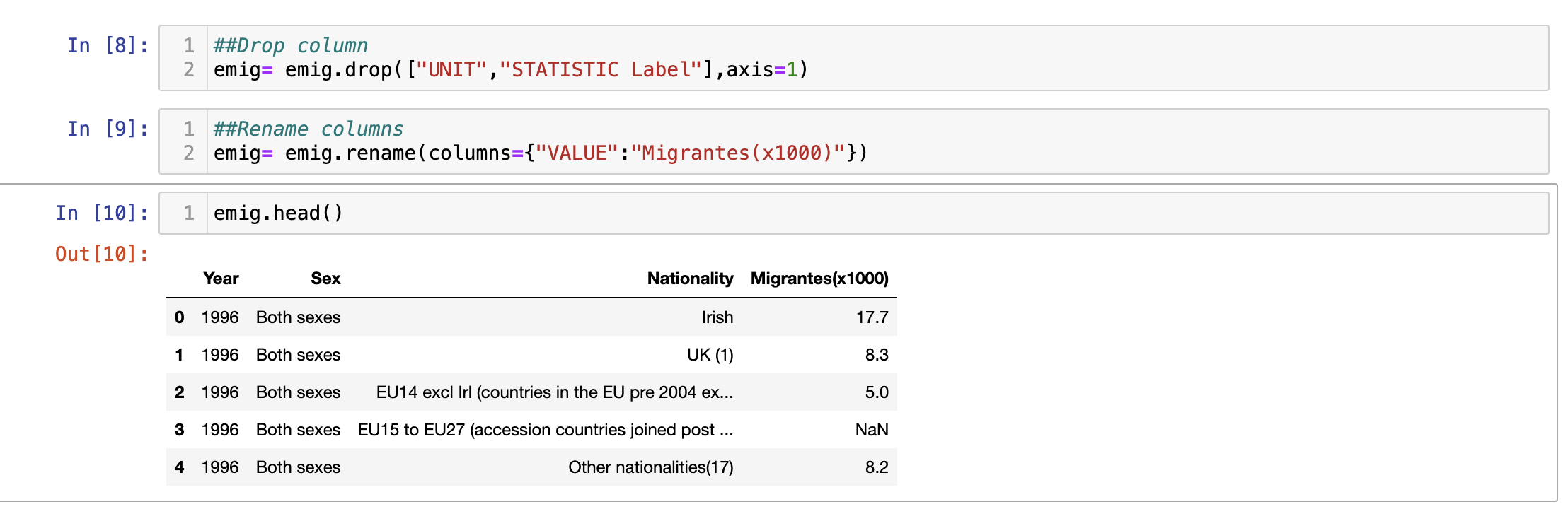


Figure 2: DataFrame after renaming and droping columns

#### The above command .Drop() is used to delete the columns of dataframe as the data will not be required for our analysis. Asia will help us to see the data only which is required.

The command .Rename() is used to rewrite the column headings for easier understanding and user convenience.

Figure 3: Drooping the unnecessary columns

### Columns Data typespasted-image.png

Figure 4: Types of data present in the data frame

#### The command is used to check the data type of each column in the data-frame. This help us to understand the their types and changes required during mathematical calculation to avoid errors.

### Check the Null value.

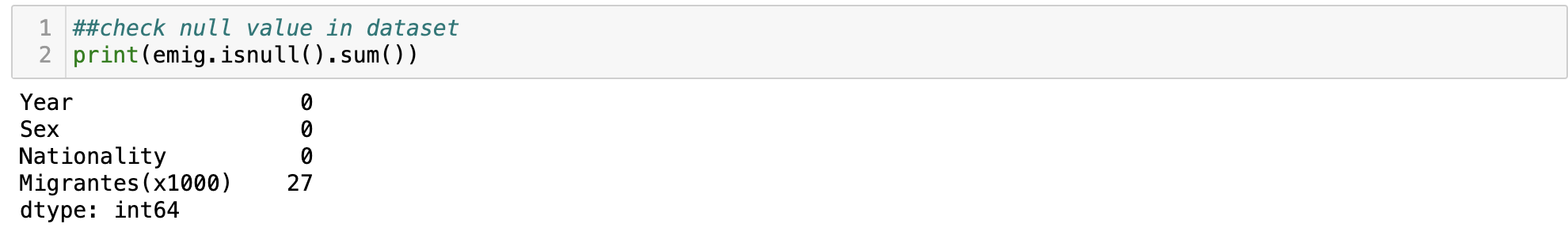


Figure 5: Checking null values in the data frame

#### The command is used to find the number of missing (null) values in each column. This is a common step in data analysis to identify and handle missing data. As shown in the above figure the all the columns have filled data expect for Migrantes(x1000) which has around 27 numbers of missing values.

### Info()pasted-image.png

Figure 6: Information about dataset

#### The info() method provides information about the data frame including datatypes, the number of not-null values, and memory usage. Here we got information about the column name, the datatype of the column, the number of not-null (not-missing values), total number of values in the column. Overall, it provides a high-level overview of the dataset's characteristics.

### Describe()pasted-image.png

Figure 7: Statistical description of the dataset

#### The describe method is used to generate descriptive statistics of the Pandas data frame. It provides a summary of central tendency, dispersion, and shape of the distribution, mean, standard deviation, min and max value, and quartile value (i.e. Q1, Q2, Q3) of each numerical column, which clarifies the distribution and range of values in each column.

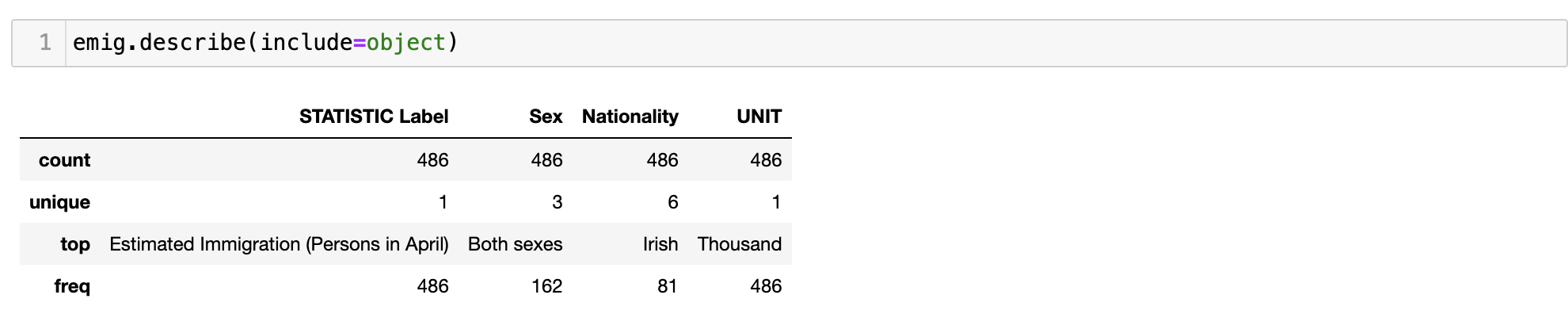
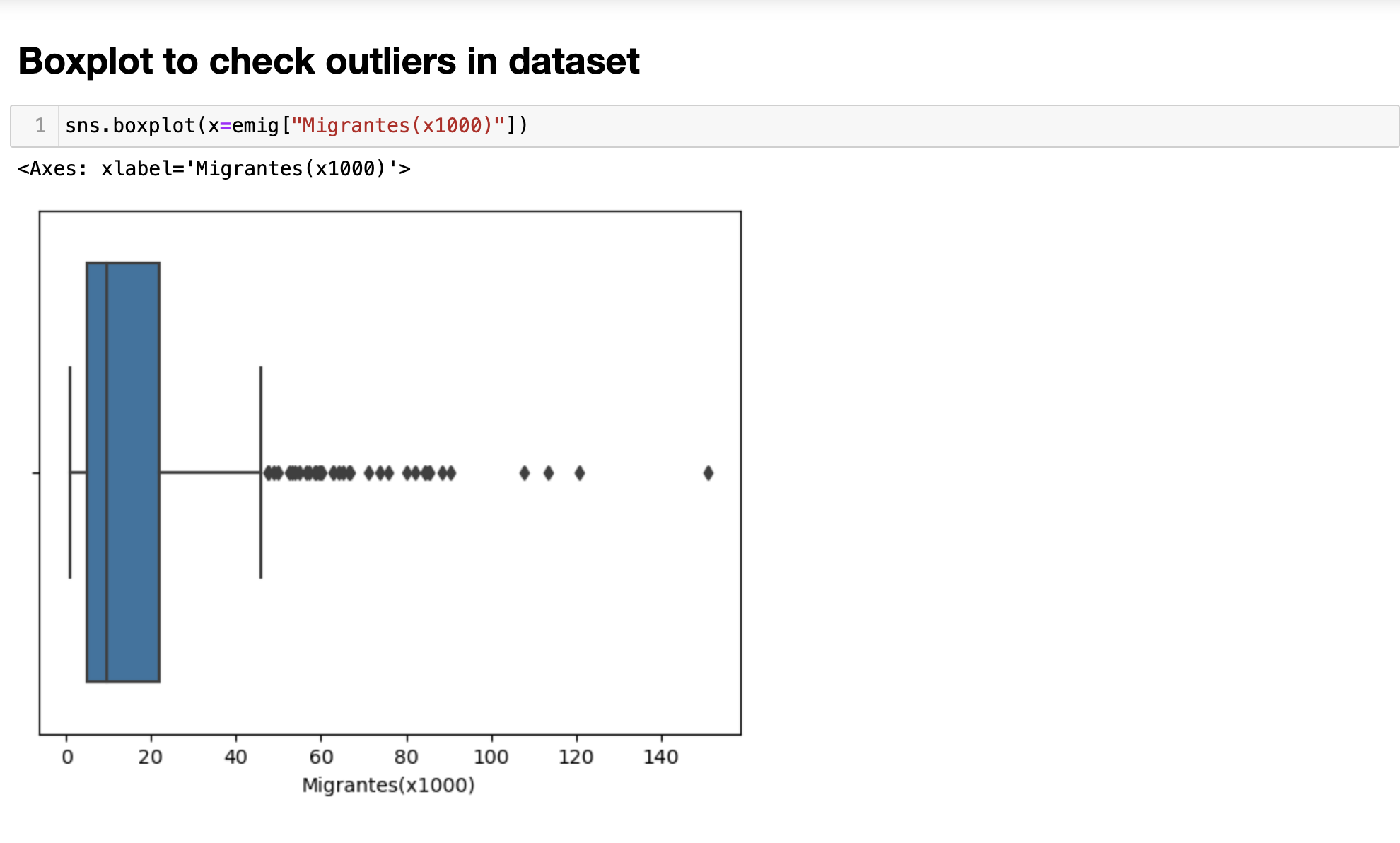
* Describe(Include=object)

Figure 8: Statistical description of the dataset

The command helps us to understand the number of different types of variables that are included in the columns and the number of times they have been used.

## Box plots

#### Box plots offer a swift visual representation of the dispersion of values within a dataset. They display key statistics such as the median, upper and lower quartile, minimum and maximum value, and any value way from maximum and minimum are considered as outliers which are not good for further analysis because they provide incorrect analysis (Insights, 2023). Here in the figure below, we can see that some points are far away from the maximum point, and I removed them by using the Interquartile Range (IQR) technique which is nothing but a formula to remove outliers from data sets.

Q1= first quartile,

Q3 = third quartile,

IQR = Interquartile Range (Anon., 2023)

#### IQR = Q3-Q1

#### lower limit = Q1 - 1.5\*IQR

#### upper limit = Q3 + 1.5\*IQR

Remove all the value that is less than the lower limit and more than that upper limit.

Were,

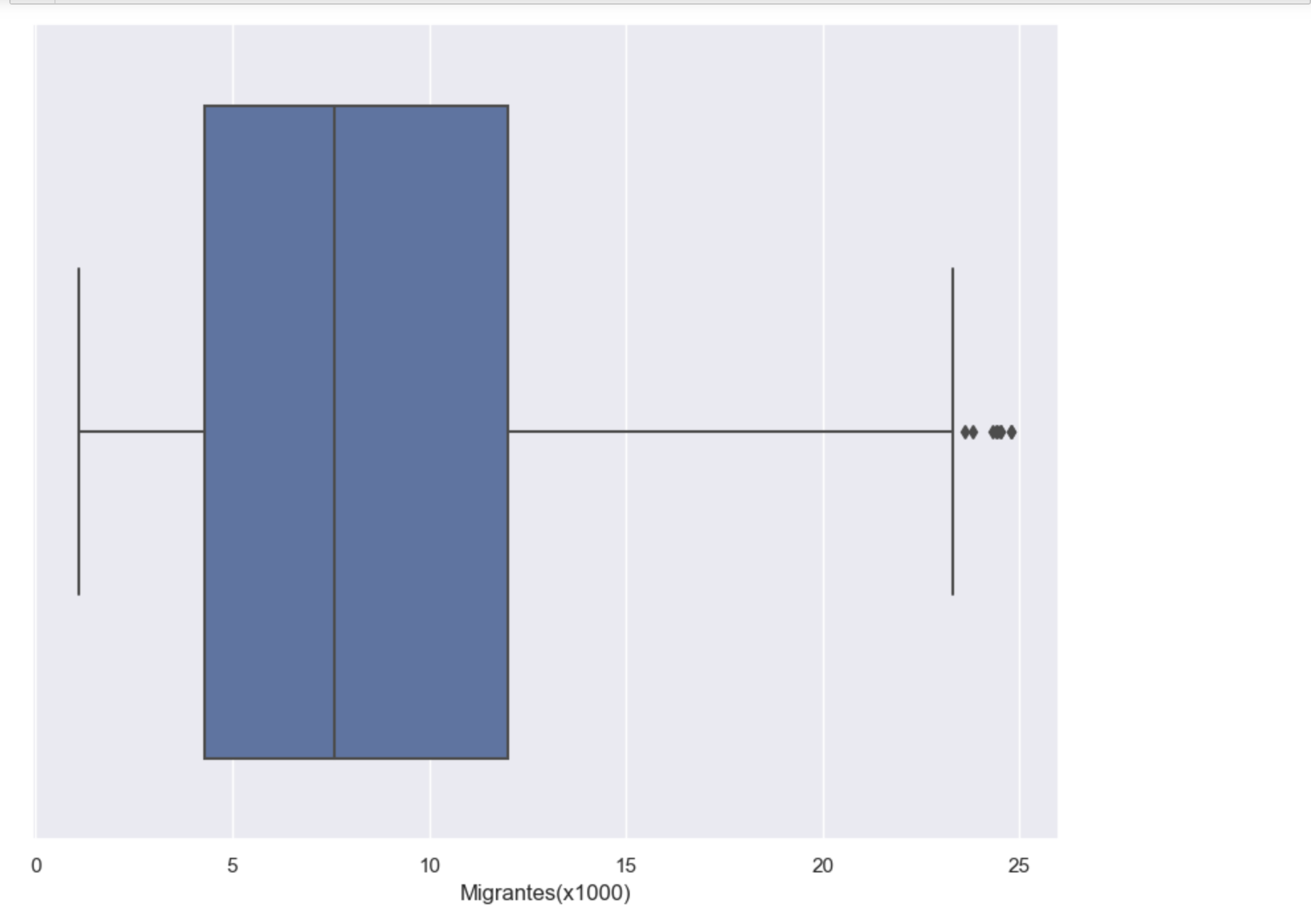


Figure 8: Removing outliers present in the dataset

#### After removing outliers, In my case,

#### IQR =16.799999999999997

#### lower limit = -20.199999999999996

#### upper limit = 46.99999999999999 After removing the outlier the very least value near the maximum point of the whisker box plot.

#### pasted-image.png

Figure 9: Data frame after removing outlier

## Heatmap

#### I explore the correlation between the year and Migrantes feature of the dataset by visualising through Heat-map which is commonly used to display correlation matrices. And I found that year and migrant”s have -0.06 of correlation.

## Statistical Analysis

Because it's widely used and can fit many situations, the normal distribution, also known as the bell curve, is popular in statistics and science. It's like a basic model that helps describe how data spread out in nature and in man-made situations. The central limit theorem is a big deal in statistics. It says that even if the original data doesn't follow a normal pattern, the average of lots of random samples from any group will usually follow a normal pattern. This is important because it lets researchers and statisticians test ideas, figure out numbers accurately, and make strong conclusions about populations.

The normal distribution gives a common way to understand and model how things vary, making it easier to handle complex real-world problems and helping with predictions and decisions. It's a universal and important idea in statistics and research, used in physics, and predicting data. It's the foundation of probability and statistics because of its even, engineering, economics, biology, and more. This pattern is a powerful tool for studying well-defined features and wide usefulness. This helps us understand the world better and is used in countless everyday situations.

### Mean

#### Mean is the average value of a given dataset. And the average migrant’s to reland from 1996 to 2022 are **8.944875 (in thousands)**

### Standard deviation

#### It is a measure of how dispersed the data is concerning the mean. A low standard deviation signifies that the data is closely packed around the mean, while a high standard deviation suggests that the data is more widely dispersed. Here in my case, the Standard Deviation value is **5.684793**

#### Overall, it suggests that the data values are clustered closely around the man. If it is relatively large, it indicates that the data values are more spared out (Medicine, n.d.).

#### pasted-image.pngpasted-image.png

Figure 10: shows the symmetric distribution of the dataset

#### 

## The bar graph to visualise the total migrant’s to Ireland from 1996-2022

#### This bar graph depicts the migration pattern to Ireland from 1996 to 2022. As shown in the graph the migrant’s to Ireland have been reduced significantly after 2009 which is from around 25**K** to 12**k in 2022.** It means that the migrant’s to Ireland has been decreased to approximately **0.55% annually** and will predict the migrantion in our further analysis by our machine learning models.

## Statistic Graph

### Normal Distribution

#### This Normal distribution graph depicts a bell-shaped curve which means our data distribution follows normal distribution. Here approximately 68% of the data falls within one standard deviation of the mean, 95% falls within two standard deviations and 99.7% of data falls within three standard deviations. This highlights the predictability of values within different ranges (CHEN, 2023).

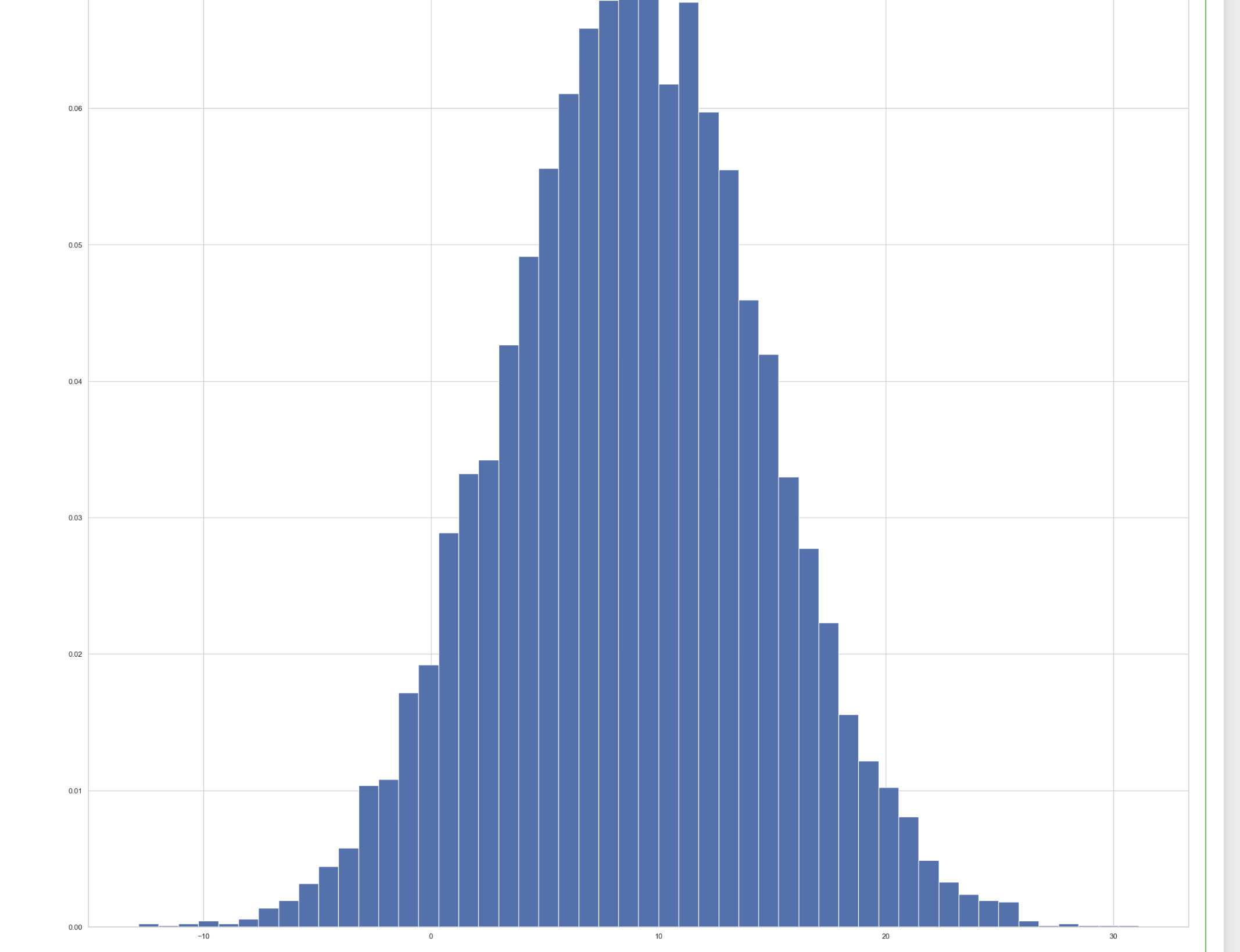


Figure 16: Population distribution of the irland

### Poisson Distribution

### The Poisson distribution is a probability distribution that describes the number of events that occur within a fixed interval of time or space. It is often used to model rare events that happen independently and at a constant average rate. It also shows the bell-shaped distribution of population data which means prediction of the population has more room to be accurate.pasted-image.png

Figure 17: Position distribution with lambda 5

# Data Preprocessing

## Encoding Data for Machine Learning Models

#### In the data frame columns like age group, and sex, are object datatypes. If we give that data to our machine learning model, there will high risk of inaccurate prediction because the machine learning model deals with integers rather than categorial or label data. So, to encode our data I will use several encoding techniques to encode our dataframe column.

### Ordinal Encoder:

#### OrdinalEncoder is used when the categorical features have an inherent order or ranking. It transforms categorical or ordinal data into a unique category assigning a unique integer value based on its order. Our age group column is an ordinal data category, so I used this encoder to encode that age group column.

### Label Encoder:

#### LabelEncoder is used when the categorical features do not have a meaningful order or ranking. It transfers categorical features into integers, assigning a unique integer to each unique category. The order is arbitrary and doesn’t have any raking. Our columns like sex, and region are nominal data, so we apply this encoding technique to that column.



Figure 18: Data set after label encoding

#### Here in the above image, after applying the encoding technique we can see sex and Nationality values have been changed from objects to float.

# Machine Learning

The choice of a project management framework in Data Analytics or Data Science depends on the project's nature and what the organization prefers. Let me explain one of the commonly used frameworks.

CRISP-DM (Cross-Industry Standard Process for Data Mining) is a comprehensive framework widely used in the data mining and data science community. It's great for projects with clear and well-defined processes, especially when the goal is things like predictive modeling, machine learning, or data mining.

For example, imagine a retail company like Amazon wants to create a recommendation system to make customers' experiences better and boost sales. In this case, they could use the CRISP-DM framework. This framework is made up of different stages like understanding the business, understanding the data, preparing the data, creating models, evaluating those models, and finally deploying them. Each stage has specific tasks and things to deliver, making sure the project is approached in a systematic way (Accredian, 2022).

KDD (Knowledge Discovery in Databases) is like a big framework that covers things like getting data ready, making new features, and doing different types of data analysis, including machine learning and data mining. It's good for projects where the main goal is to find important knowledge in really big sets of data. For example, in healthcare, a hospital might want to look at electronic health records to find patterns in patient data and make care better. KDD has different steps like cleaning the data, combining data, picking the right data, changing the data, doing data mining, checking the patterns, and showing the knowledge. It's a back-and-forth process where you keep making it better (GeeksforGeeks, n.d.).

SEMMA (Sample, Explore, Modify, Model, and Assess) is super useful for projects where the main goal is to make models that predict things and see how well they work. It's good when the data is clear, and you want to really focus on making models. For example, if a company that scores credit wants to make a model to see if people are risky to give loans to, SEMMA is a good way to do it. There are five steps in SEMMA: take a sample of the data, explore it, change things if needed, make the model, and then check how good the model is (Alliance, 2023).

## Selection of Machine Learning

### Unsupervised Learning

#### Since my dataset is unlabeled i.e., continuous data of total population based on years and region it doesn’t give any true or false that is labelled output. So, I’m using Unsupervised Learning Machine learning models which are used for prediction, clustering, dimensionality reduction and anomaly detection, where I’m using my continuous data set to predict population based on region and years.

### Feature Scaling:

#### Feature Scaling in machine learning is the process of standardizing or normalizing the numerical features of datasets. It involves transforming the values of features into a specific range or distribution to ensure that they contribute equally to the model’s performance. I’m using the Standardization (Z-score normalization) technique to scale my dependent and independent variables because my dataset has a lot of outliers, and it follows Gaussian Distribution. After scaling my dataset looks like the one given below (GeeksforGeeks, n.d.).pasted-image.png

## Linear Regression

#### Linear Regression analysis is used to predict the value of a variable based on the value of another variable. For example, the value that I want to predict is a dependent variable and the values that I’m using for prediction are called independent variable (Sklearn, n.d.).

#### X (Independent Variable) => Year, sex,Nationality

Y (Dependent Variable) => Migrantes(x1000)

#### After Training and Testing my model, I got the accuracy for Training **is 20.192%** and **Testing at 36.323%** which is very low. It shows that my model is underfitted due to less amount of data i.e., 422 observations.

#### These are my basic predictions from my machine learning model for upcoming years.

#### Further to enhance my machine learning model I have used Ridge Regression, Lasso Regression, and Elastic Net to improve the hyperparameter of my Model to get more accuracy their score is as given below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Approach | R^2 Training set | R^2 Test set | Root means square error for prediction | Mean absolute percentage error for prediction |
| Lasso Regression | 23.735% | 16.489% | 914.379% | 98.379% |
| Decision Tree Regressor | 1.0% | 0.885% | 3.3926% | 16.024% |
| Random Forest Regressor | 0.9868% | 0.90318% | 3.113% | 16.0436% |

#### These are the results I got from my different machine-learning models after applying several models to my dataset.

### R^2 score

#### A higher r^2 score indicates that the model is better at explaining the variance in the training and testing data. However, an excessively high r^2 score may suggest overfitting, where the model fits the training data too closely but may not generalized well to new unseen data. In my case, From the above r^2 for testing and training for both Ridge and Lasso regression, it is confirmed that my Machine learning model is underfitting (FERNANDO, 2023).

### Root Mean Square Error (RMSE):

#### These metrics are used to evaluate the performance of predictive models often in the context of regression problems. It represents the square root of the average squared difference between predicted and actual values. It provides a measure of how spared the errors are. If it has a lower value indicates better performance of the model. However, in my case, it has very high values that are 81.942% and 82.172% for Ridge and Lasso respectively (TO, 2023).

### Mean Absolute Percentage Error (MAPE):

#### These metrics are used to evaluate the performance of predictive models often in the context of regression problems. It represents the average percentage difference between predicted and actual values. It provides a measure of how spared the errors are. If it has a lower value indicates better performance of the model.

# conclusion

#### In this Report, I present the data for Ireland's population based on Nationality, sex, Migrantes(x1000), and years. Where I perform different Machine learning models like Linear Regression, lasso Regression, to predict the estimated migrants based on independent parameters like, sex, nationality, year against the Migrants(x1000), What similarity I found between my ML models are that they all tend to predict the population and accuracy for training and testing are not much different model that I mention in the table above in the score table for the model.The random forest and decision tree tend to show more accurate predictions compared to Lasso and Linear regression.

# Reference

Central Statistics Office, I., n.d. *Population Estimates / Annual Population Estimates.* [Online]   
Available at: https://data.cso.ie/product/pme   
[Accessed 16 10 2023].

NumPy, 2023. *About us.* [Online]   
Available at: https://numpy.org/about/  
[Accessed 1 11 2023].

Matplotlib, T., 2023. *Matplotlib: Visualization with Python.* [Online]   
Available at: https://matplotlib.org/  
[Accessed 1 11 2023].

Waskom., M., 2023. *seaborn.* [Online]   
Available at: https://seaborn.pydata.org/  
[Accessed 1 11 2023].

community, T. S., 2023. *Statistical functions.* [Online]   
Available at: https://docs.scipy.org/doc/scipy/reference/stats.html  
[Accessed 1 11 2023].

developers, s.-l., 2023. *1.11. Ensembles: Gradient boosting, random forests, bagging, voting, stacking.* [Online]   
Available at: https://scikit-learn.org/stable/modules/ensemble.html  
[Accessed 1 11 2023].

developers, s.-l., 2023. *index.* [Online]   
Available at: https://scikit-learn.org/stable/index.html  
[Accessed 1 11 2023].

Foundation, P. S., 2001-2023. *warning.* [Online]   
Available at: https://docs.python.org/3/library/warnings.html  
[Accessed 1 11 2023].

pydata, p., 2023. *pandas.* [Online]   
Available at: https://pandas.pydata.org/

Ptil, P., 2018. *Medium/ What is Exploratory Data Analysis?.* [Online]   
Available at: https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15  
[Accessed 2 11 2023].

Pandas, 2023. *pandas.DataFrame.* [Online]   
Available at: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.head.html#pandas.DataFrame  
[Accessed 5 11 2023].

Insights, A., 2023. *Visualize.* [Online]   
Available at: https://doc.arcgis.com/en/insights/latest/create/box-plot.htm#:~:text=Box%20plots%20provide%20a%20quick,or%20unusual%20occurrences%20in%20data.  
[Accessed 2 11 2023].

Anon., 2023. *Interquartile range.* [Online]   
Available at: https://en.wikipedia.org/wiki/Interquartile\_range#:~:text=Interquartile%20range%2C%20IQR%20%3D%20Q3,%3D%209%20%2B%203%20%3D%2012.  
[Accessed 1 11 2023].

Waskom, M., 2023. *seaborn.heatmap.* [Online]   
Available at: https://seaborn.pydata.org/generated/seaborn.heatmap.html  
[Accessed 1 11 2023].

Simplilearn, n.d. *The Complete Guide to Skewness and Kurtosis.* [Online]   
Available at: https://www.simplilearn.com/tutorials/statistics-tutorial/skewness-and-kurtosis  
[Accessed 3 11 2023].

Medicine, N. L. o., n.d. *Common Terms and Equations.* [Online]   
Available at: https://www.nlm.nih.gov/oet/ed/stats/02-900.html#:~:text=A%20standard%20deviation%20(or%20%CF%83,data%20are%20more%20spread%20out.  
[Accessed 3 11 2023].

RTE, 2021. *Why are birth rates falling in Ireland?.* [Online]   
Available at: https://www.rte.ie/brainstorm/2021/0519/1222517-ireland-birth-rates-falling/  
[Accessed 4 11 2023].

CHEN, J., 2023. *Normal Distribution: What It Is,.* [Online]   
Available at: https://www.investopedia.com/terms/n/normaldistribution.asp#:~:text=Error%20Code%3A%20100013)-,What%20Is%20a%20Normal%20Distribution%3F,as%20a%20%22bell%20curve%22.  
[Accessed 5 11 2023].

Accredian, M. K. p. a., 2022. *Project Management in Data Science using CRISP-DM.* [Online]   
Available at: https://medium.com/international-school-of-ai-data-science/project-management-in-data-science-using-crisp-dm-54ee35a5f4f3  
[Accessed 10 11 2023].

Alliance, N. H. b. D. S. P., 2023. *What is SEMMA?.* [Online]   
Available at: https://www.datascience-pm.com/semma/  
[Accessed 10 11 2023].

GeeksforGeeks, n.d. *Feature Engineering: Scaling, Normalization, and Standardization.* [Online]   
Available at: https://www.geeksforgeeks.org/ml-feature-scaling-part-2/  
[Accessed 10 11 2023].

Sklearn, n.d. *sklearn.linear\_model.LinearRegression¶.* [Online]   
Available at: https://scikit-learn.org/stable/modules/generated/sklearn.linear\_model.LinearRegression.html  
[Accessed 9 11 2023].

FERNANDO, J., 2023. *R-Squared: Definition, Calculation Formula, Uses, and Limitations.* [Online]   
Available at: https://www.investopedia.com/terms/r/r-squared.asp  
[Accessed 10 11 2023].

TO, S. H., 2023. *RMSE: Root Mean Square Error.* [Online]   
Available at: https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/  
[Accessed 10 11 2023].