**CCT College Du CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Data Preparation & Visualisation, Machine Learning for Data Analytics, Programming for Data Analytics, Statistics for Data Analytics |
| **Assessment Title:** | CA 1 |
| **Lecturer Name:** | David McQuaid, Muhammad Iqbal, Saw Weiss, Taufique Ahmed |
| **Student Full Name:** | Lylkend Silva Albuquerque |
| **Student Number:** | 2023290 |
| **Assessment Due Date:** | 12 /11/2023 |
| **Date of Submission:** | 12/11/2023 |

**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.** |

# MSC\_DA\_CA1

Author: L. Albuquerque

e-mail: [2023290@student.cct.ie](mailto:2023290@student.cct.ie)

Student ID: 2023290

Ireland population changes over the past 10 years  
*With a close look at immigration and Migration*

## Acknowledgements

I would like to thank the following people who have helped me undertake this research:

My Professor David McQuaid, for his enthusiasm in all of our classes, for his support, encouragement and patience and of course to all the others professors who strive to put us on the path of knowledge and development such as:

Sam Weiss

Dr. Muhammad Iqbal

Dr. Bharathi Chakravarthi

Taufique Ahmed

The CCT college, for input throughout this MSc programme.

All my colleagues which in their way had tried to give some contributions to data collection and some expertise was given in our class group.

To the Central Statistics Office (CSO) who concisely made available to us all the databases in which it was possible to develop this research.

My best friend Naisa de Almeida Cunha I simply couldn’t have done this without your words of encouragement, availability and care, special thanks.

Dear housemates which always respected my space while I was researching.

And to my parents, who set me off on the road to this MSc a long time ago.

## Abstract

*Ireland has experienced significant population changes over the past few decades. One of the most prominent trends has been a steady increase in population, largely attributed to factors such as immigration, a growing economy, and improved living conditions.*

*Ireland has attracted a substantial number of immigrants, both from within the European Union and further afield. This influx of diverse cultures and backgrounds has contributed to the country's changing demographics. Ireland's strong economic growth, particularly in the technology and pharmaceutical sectors, has driven employment opportunities. This has not only retained Irish citizens but also attracted skilled workers from abroad. The majority of population growth has been concentrated in urban areas, especially Dublin. This urbanization trend is indicative of economic and employment opportunities, drawing people from rural regions to cities. Alongside population growth, Ireland is also experiencing an aging population. The proportion of elderly citizens is increasing, leading to healthcare and pension challenges in the years to come. Fertility rates in Ireland have fluctuated, with periods of decline followed by a slight increase. Government policies and social support measures have influenced family planning decisions. Managing the evolving population presents challenges in infrastructure, healthcare, education, and housing. However, it also offers opportunities for cultural enrichment, a skilled workforce, and economic development. While this abstract is based on data up to 2023 collected from Central Statistics Office (CSO), it is essential to recognize that population trends are subject to change due to unforeseen events, such as economic shifts or global health crises. The COVID-19 pandemic, for instance, has demonstrated how unexpected events can significantly impact population dynamics, with effects on migration patterns, healthcare infrastructure, and social behaviour. As we move forward, staying vigilant and adaptable in response to such unforeseen challenges will continue to be critical in understanding and shaping Ireland's demographic landscape.*

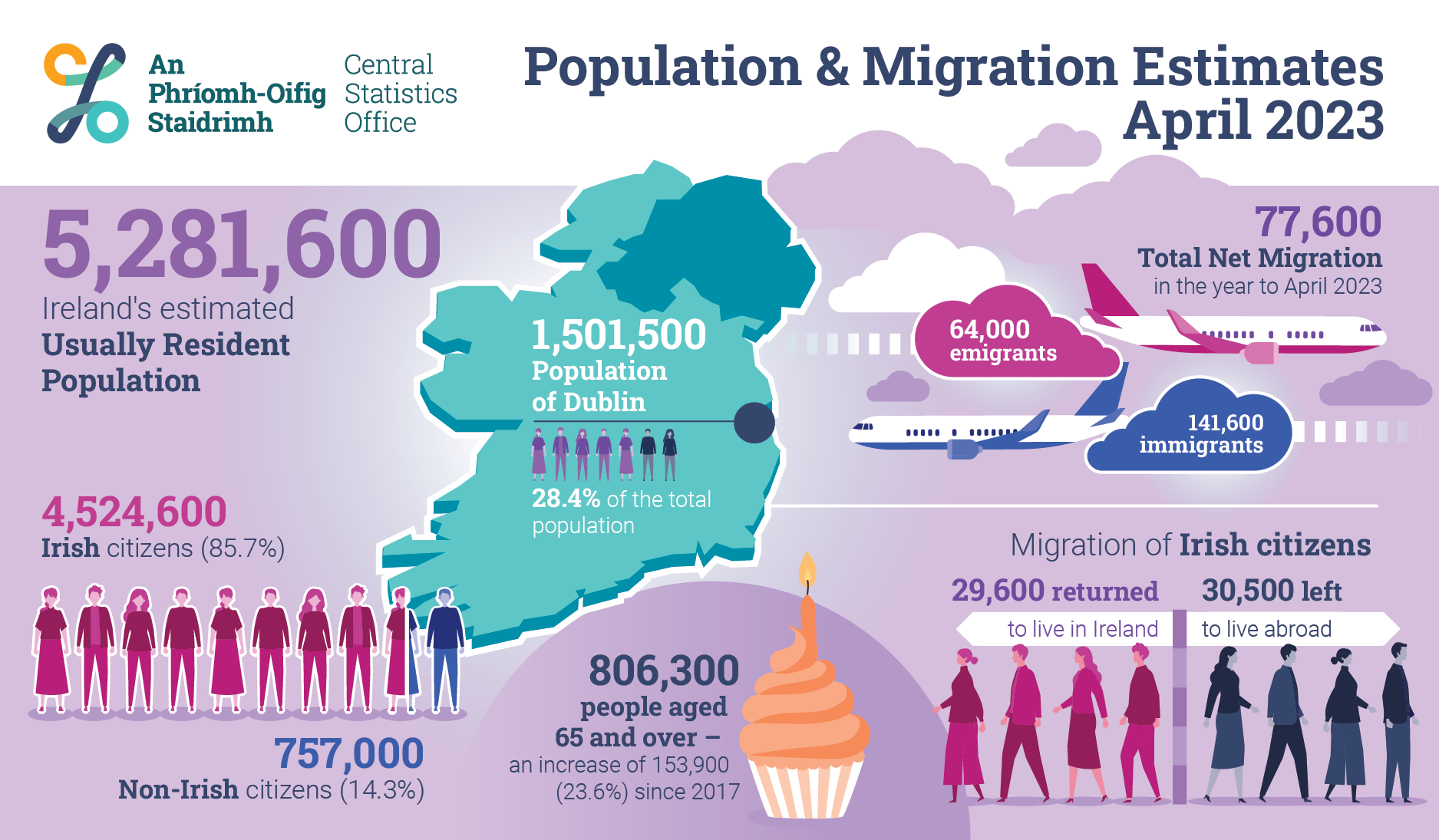
****

Figure 1 - cso.ie\_populationandmigrationestimates\_2023

**Introduction**

We are going to dive into a detailed examination of migration trends using the years 2014 to 2023 as a basis for analysis, unravelling the complex dynamics that shape the ebb and flow of people within the studied region. With a particular emphasis on key migration-related variables such as "Immigrants," "Emigrants," "Net migration," and their cumulative impact on "Population change," this analysis aims to uncover patterns, variations, and implications embedded in the migration landscape.

Migration, a critical component of demographic dynamics, reflects not only the movement of individuals across borders but also plays a pivotal role in influencing population growth or decline. Understanding the nuances of migration trends is essential for policymakers, demographers, and researchers as they navigate the complexities associated with population shifts.

In the upcoming sections, we will employ statistical measures and visualizations to delve into the intricate details of migration patterns. By scrutinizing variables like "Net migration," which encapsulates the net gain or loss of population due to migration, we aim to discern the underlying factors that contribute to demographic changes over time.

Through a critical lens, we will explore questions such as the impact of immigration and emigration on population change, the trends in net migration over the years, and any notable deviations from expected patterns. This focused analysis on migration trends is poised to unearth valuable insights, enabling a comprehensive understanding of how migration shapes the demographic landscape of the region under study.

As we navigate through the subsequent sections of this analysis, the aim is not only to describe observed trends but also to interpret their implications, contributing to a broader understanding of the societal, economic, and policy dimensions influenced by migration dynamics.

**Exploratory Data Analysis -** **EDA**

“The main purpose of EDA is to help look at data before making any assumptions”(What is Exploratory Data Analysis?, no date) Exploratory Data Analysis (EDA) also called *Early Data analysis* is a crucial and fundamental step at the beginning of any data analysis, including data analysis for machine learning since it is possible to better understand the data in a broad view, which will certainly help the analyst to choose better ways to explore the database in a more cohesive way with the reality of that data set.

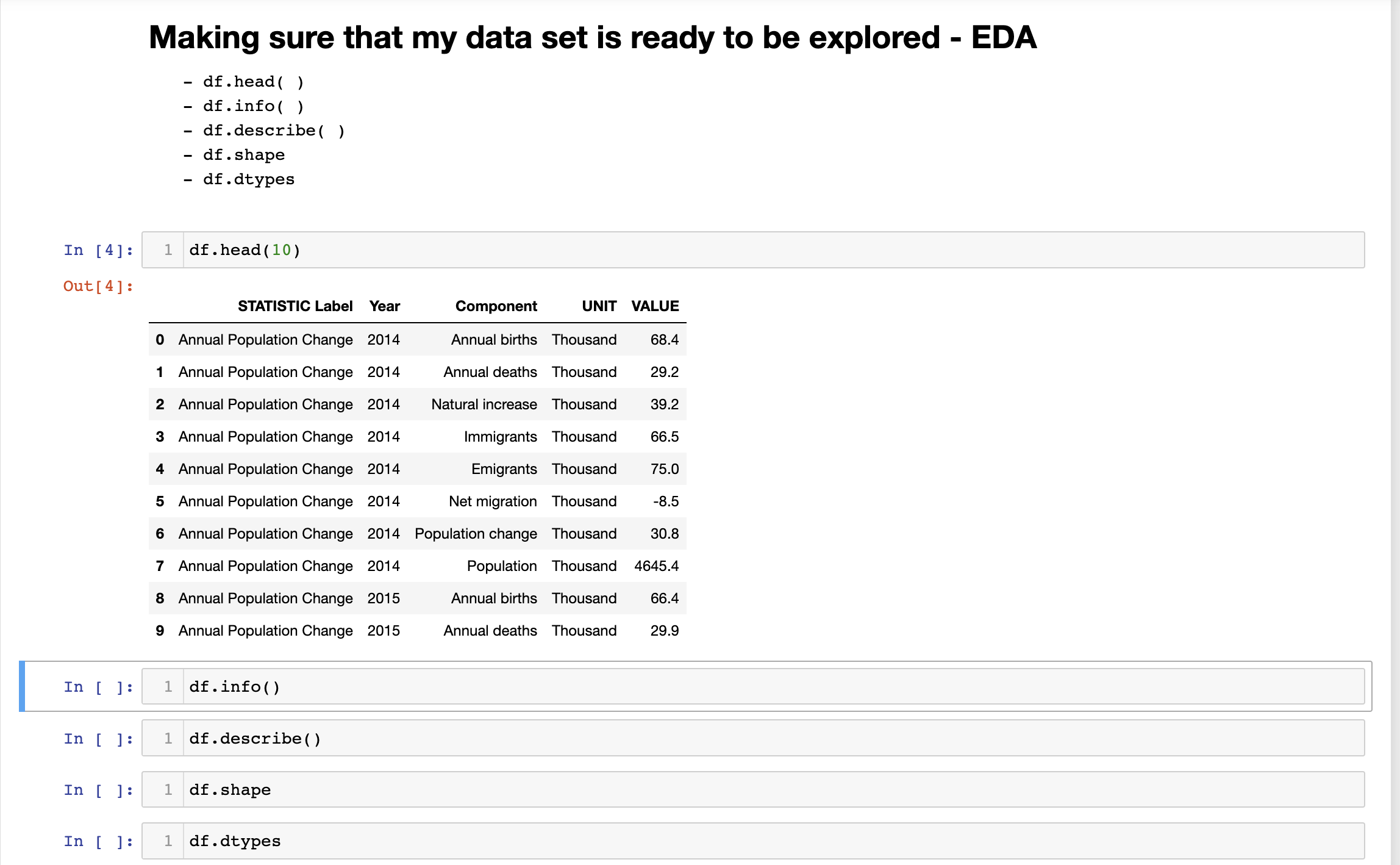


Figure 2 – EDA

* *Data quality check*: EDA allows us to identify and resolve issues related to data quality, such as missing values, outliers, and inconsistencies. Detecting and addressing these issues early is critical to ensuring the integrity of our analysis. As we can see in the image below, we are checking whether we have duplicity, then deleting (*which is part of Data Cleaning*) columns that were initially identified as not being interesting for our analysis and finally doing a simple conversion of units to have more realistic numbers.

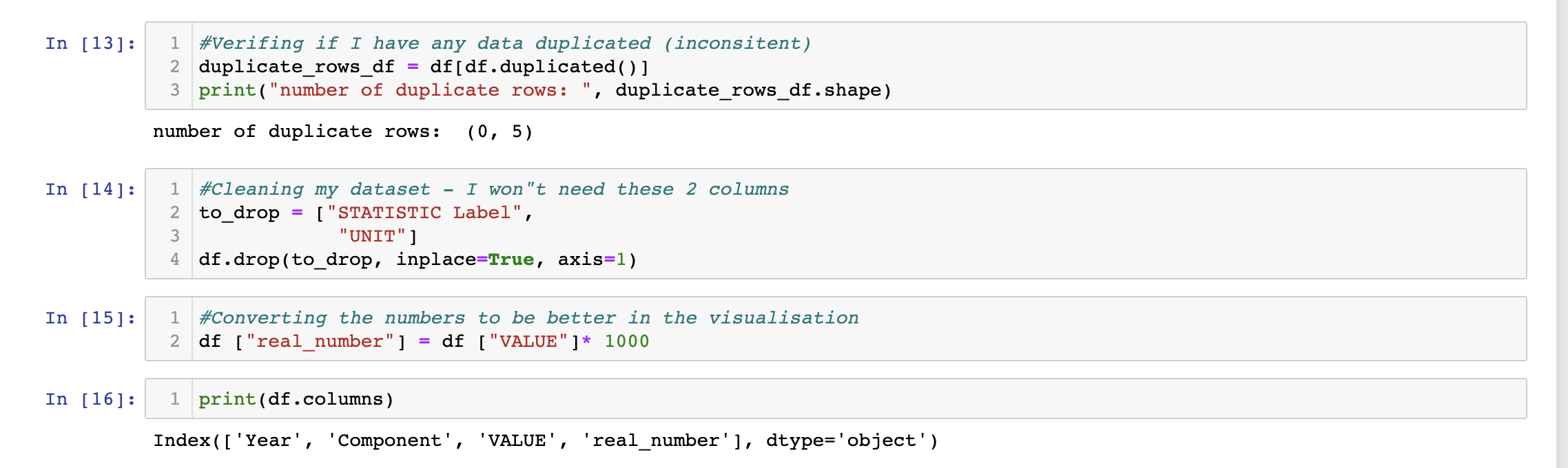


Figure 3 – Data Cleaning

* *Feature Selection:* EDA can help you identify which characteristics (variables) are relevant and important for your analysis. By exploring the relationships between features and the target variable, you can make informed decisions about feature selection and engineering. So we just realized that within the feature called "Component" we have eight different variables which later on they will be explored to greater potential.

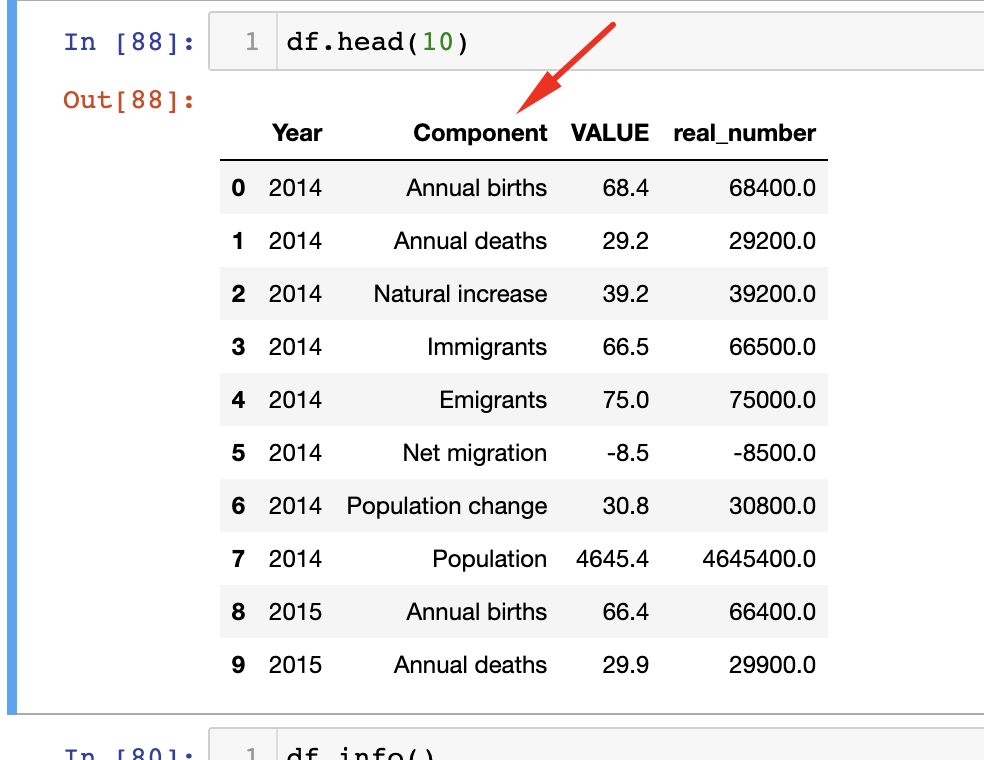


Figure 4 – First 10 lines of DF data set

* *Pattern Discovery:* EDA helps you discover patterns, trends, and relationships within the data as already said in the previous paragraph. Understanding these patterns can lead to insights that are valuable for the analysis and for making data-driven decisions in the analyses process and so on.

In summary, EDA is fundamental because it provides the necessary context and insights to make informed decisions throughout the data analysis process. It sets the stage for data pre-processing, model selection, and hypothesis testing. Skipping EDA can lead to uninformed choices and suboptimal results in subsequent analysis steps. Therefore, it is highly advisable to perform EDA at the beginning of any data analysis project.

### **Programming Paradigms**

was used a combination of various programming paradigms, including imperative, procedural, data analysis, functional programming and elements of object-oriented. It's common to use multiple paradigms when working on data analysis and visualization tasks as we can see in the peace of code below:



Figure 5 – Programming paradigms

* *Imperative Programming*: using imperative programming in this code to instruct the computer on how to perform a series of steps to create bar plots. You specify the sequence of operations explicitly in a step-by-step manner.
* *Procedural Programming*: This code also follows a procedural structure where it’s defined separate procedures or steps to perform specific tasks, such as creating bar plots for each unique year. That piece is structured the code into several procedures or blocks to manage different parts of the process.
* *Data Analysis Paradigm*: This code mainly follows the data analysis paradigm, where it is working with data in a structured way. Where we can notice the filtering and grouping of data, performing calculations and visualizing the results. This is in line with the principles of data analysis and manipulation.
* *Functional Programming*: There are elements of functional programming in this code, specifically in how you use functions from libraries like pandas and matplotlib. For example, I am using methods like “**groupby(“Component”)**” and “**.sum()”,** which are functional in nature.
* *Object-Oriented Programming (OOP):* Although not explicitly evident in this code snippet, libraries such as pandas and matplotlib, which were declared at the beginning of the project (as we can see in the figure below) are built on object oriented principles. However, the use in this code is more focused on procedural aspects.

The choice of these programming paradigms and techniques influenced the structure, organization and problem-solving capabilities, one example that it is given in the picture below is the definition of the "Label graph" function that reduces five lines of code into just one. So that when called it reduces time and keeps the code organized and clean. However, I believe that in the future, more and more codes with structures that are easy to use, read and maintain can be implemented. Each paradigm serves a specific purpose and combining them can lead to efficient and sustainable data analysis and visualization solutions.

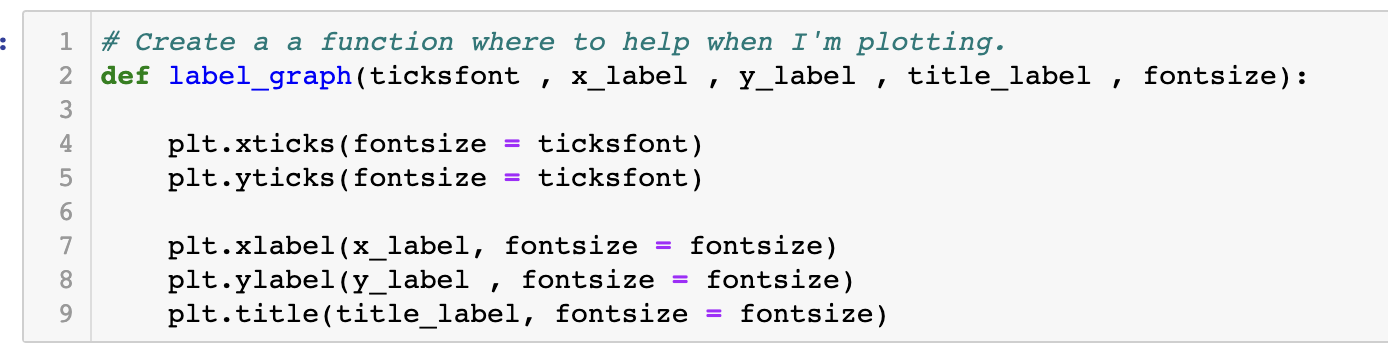


Figure 6 – Function to create graphs

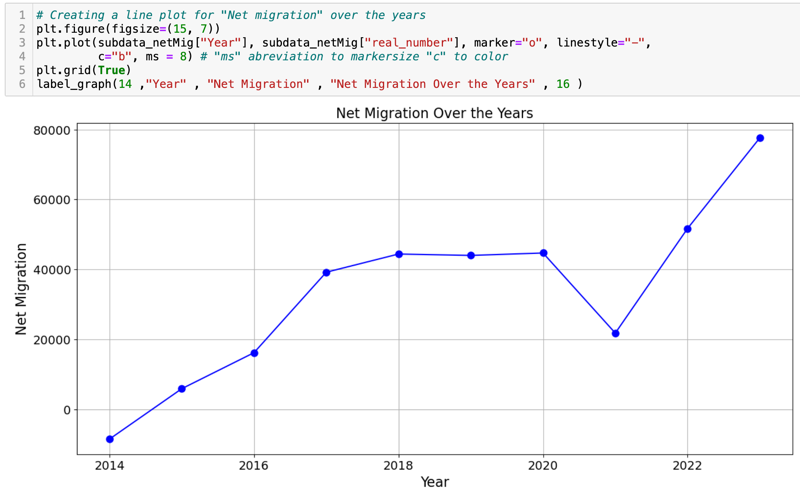
****The line graph on the side was given as a good example of the use of this type of graph where the emphasis is to point out the data in a clear and direct way in terms of 2 simple constants where none of the variables are compared with each other.

Figure 7 - Line chart

**Distributions**

Before delving into the intricacies of the distributions, a pivotal step was taken: the dataset underwent a thorough transposition. The primary objective of this transposition was to meticulously scrutinize each individual feature in isolation. This strategic manoeuvre was employed to enrich our comprehension of the unique indicators within every cell. The overarching intention is to revisit and delve into these indicators in the future, thereby cultivating a more nuanced and comprehensive perspective.

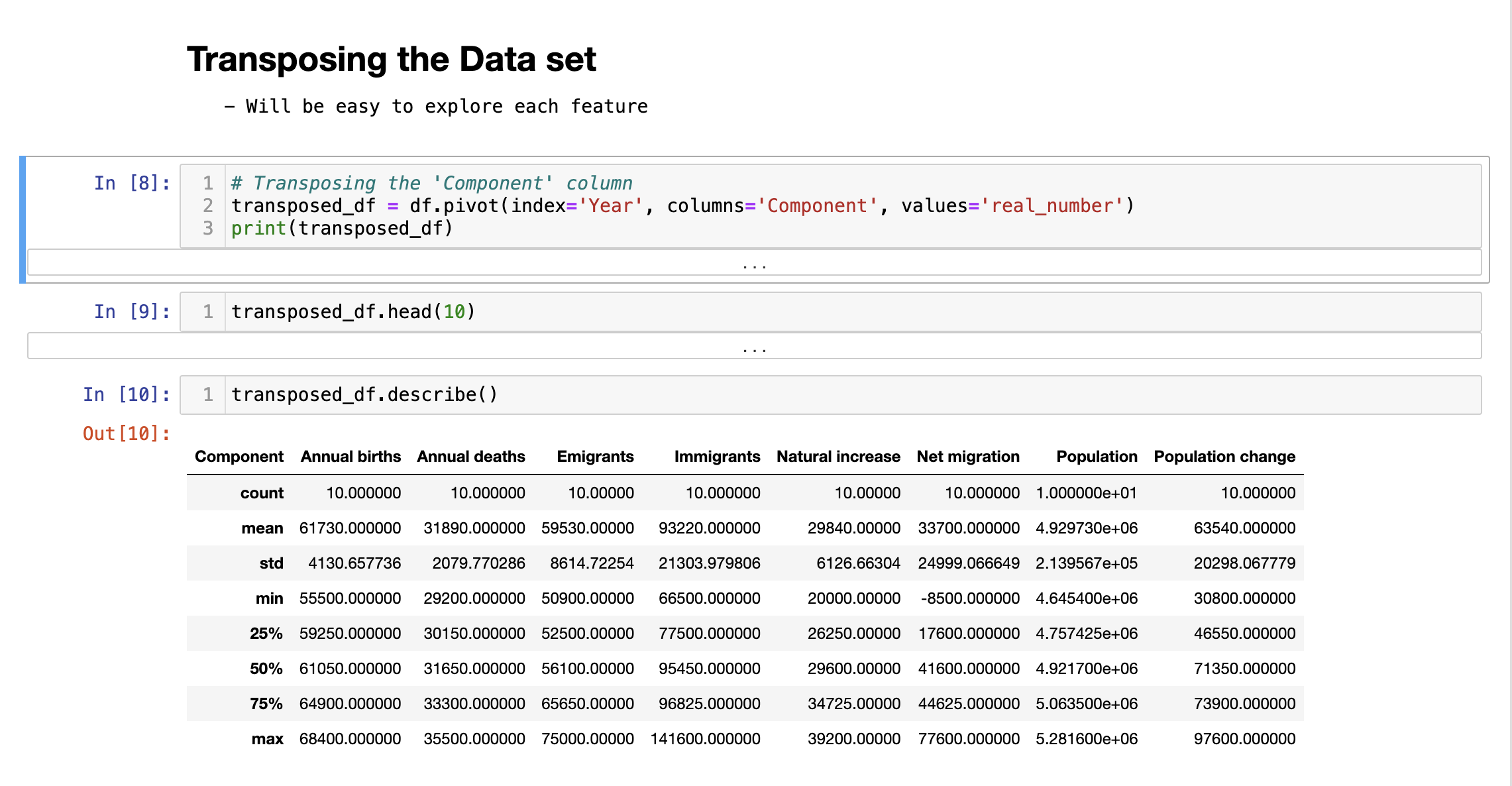
****

Figure 8 – Transposition of Data set

* Normal distribution

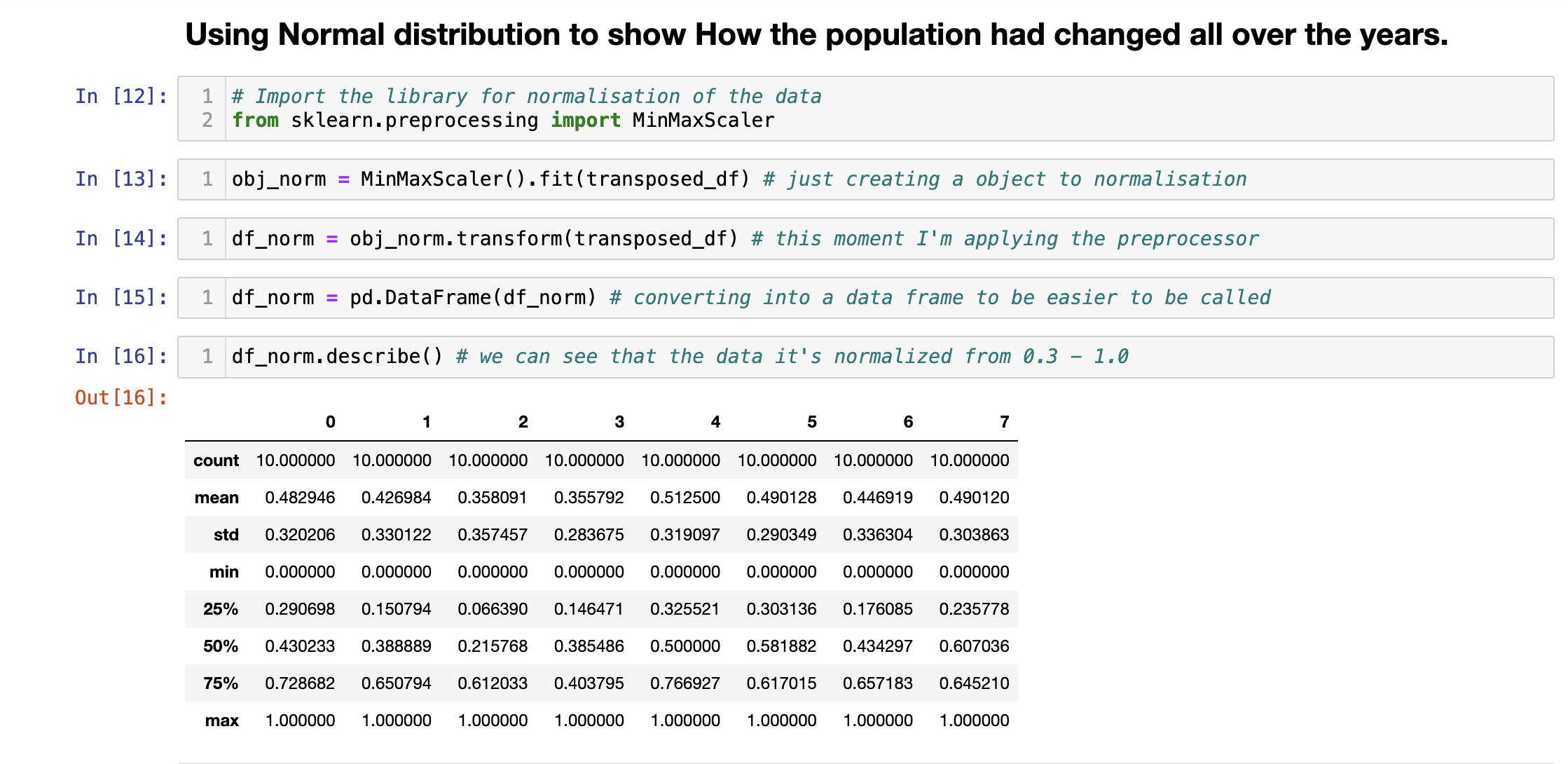
****

Figure 9 - Normalization steps

So we going to use the “*MinMaxScaler*” is a data pre-processing technique commonly used in machine learning to scale and normalize the features of a dataset within a specified range, typically [0, 1]. This scaler linearly transforms the original data, ensuring that the minimum value of the transformed data corresponds to the specified minimum (usually 0), and the maximum value corresponds to the specified maximum (usually 1). Upon completion of the normalization procedure, a discernible transformation is observed in the dataset. Specifically, the data undergoes a scaling operation, bringing it into conformity with the predefined range of 0 to 1, in accordance with the standard norms for normalization. It is imperative to underscore that normalization, in this context, does not entail any modification of the intrinsic data values. Rather, its principal function lies in the systematic scaling of the dataset, allowing for more effective and insightful analysis while maintaining the authenticity of the original data. This process ensures that the numerical representation of the data aligns with the desired scale without distorting the inherent information encoded within the dataset.

In the following image, a box plot is employed to illustrate that while all the data has undergone normalization, there exists a notable disparity in the proximity of the means. The selection of boxplots is motivated by their ability to provide a swift and efficient means of comprehending the fundamental characteristics of a dataset. Boxplots excel in facilitating the identification of outliers, enabling the comparison of distributions, and supporting informed decision-making in the realm of data analysis and interpretation. Renowned for their versatility, boxplots find application across diverse fields, including statistics, data science, and exploratory data analysis.

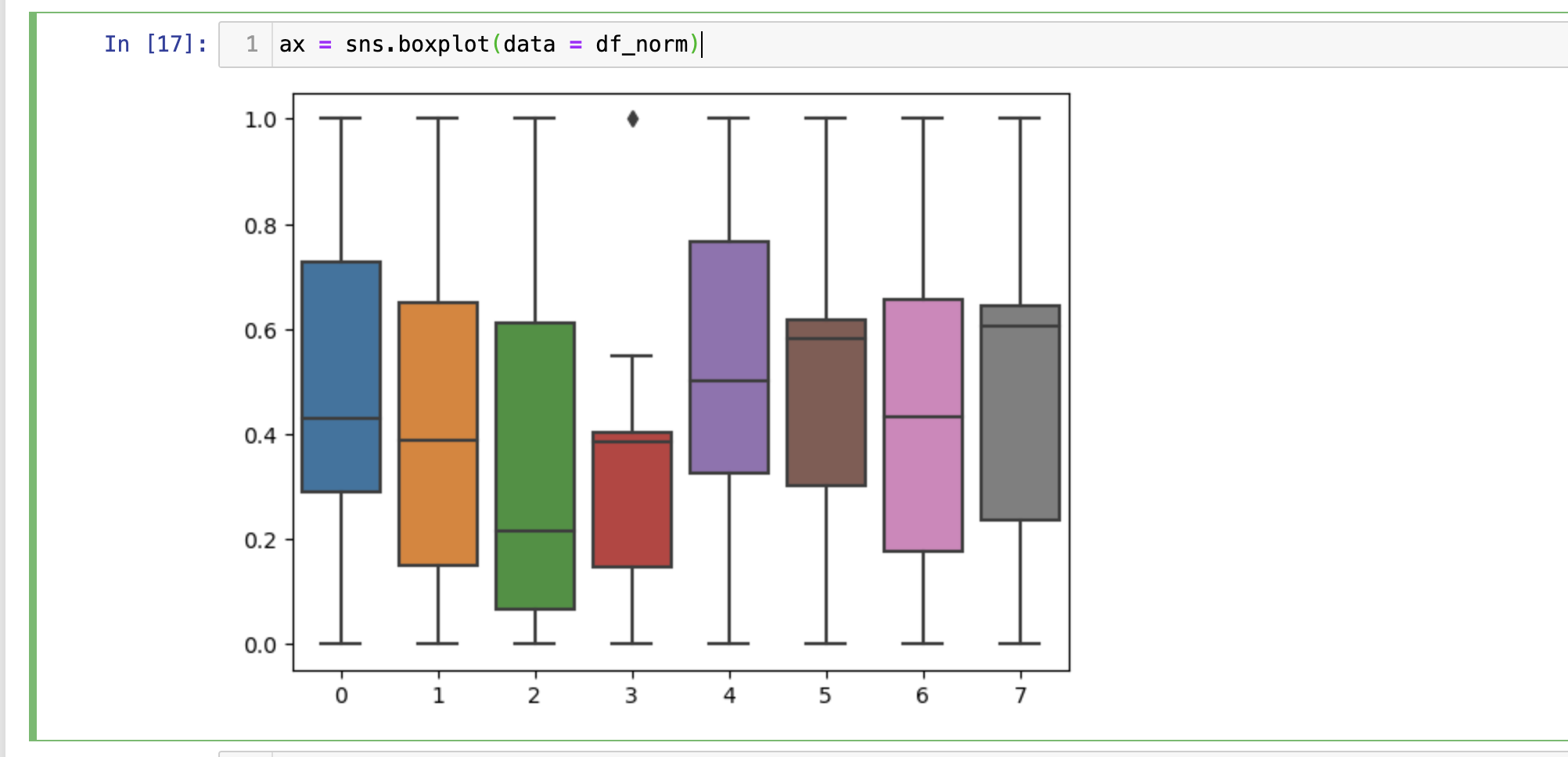


Figure 10 - Boxplot of normalization

The normal distribution is characterized by two parameters: mean (loc) and standard deviation (scale). In the code snippet below, the mean and standard deviation for variable 7 are explicitly provided by the *MinMaxScaler* technique in the in the previous step. These parameters play a crucial role in defining the shape and characteristics of the normal distribution. The code employs the “*norm.pdf”* function from the *scipy.stats* module to calculate the probability density function of the normal distribution for a range of values on the x-axis. This PDF represents the likelihood of observing different values of variable 7 in the dataset.

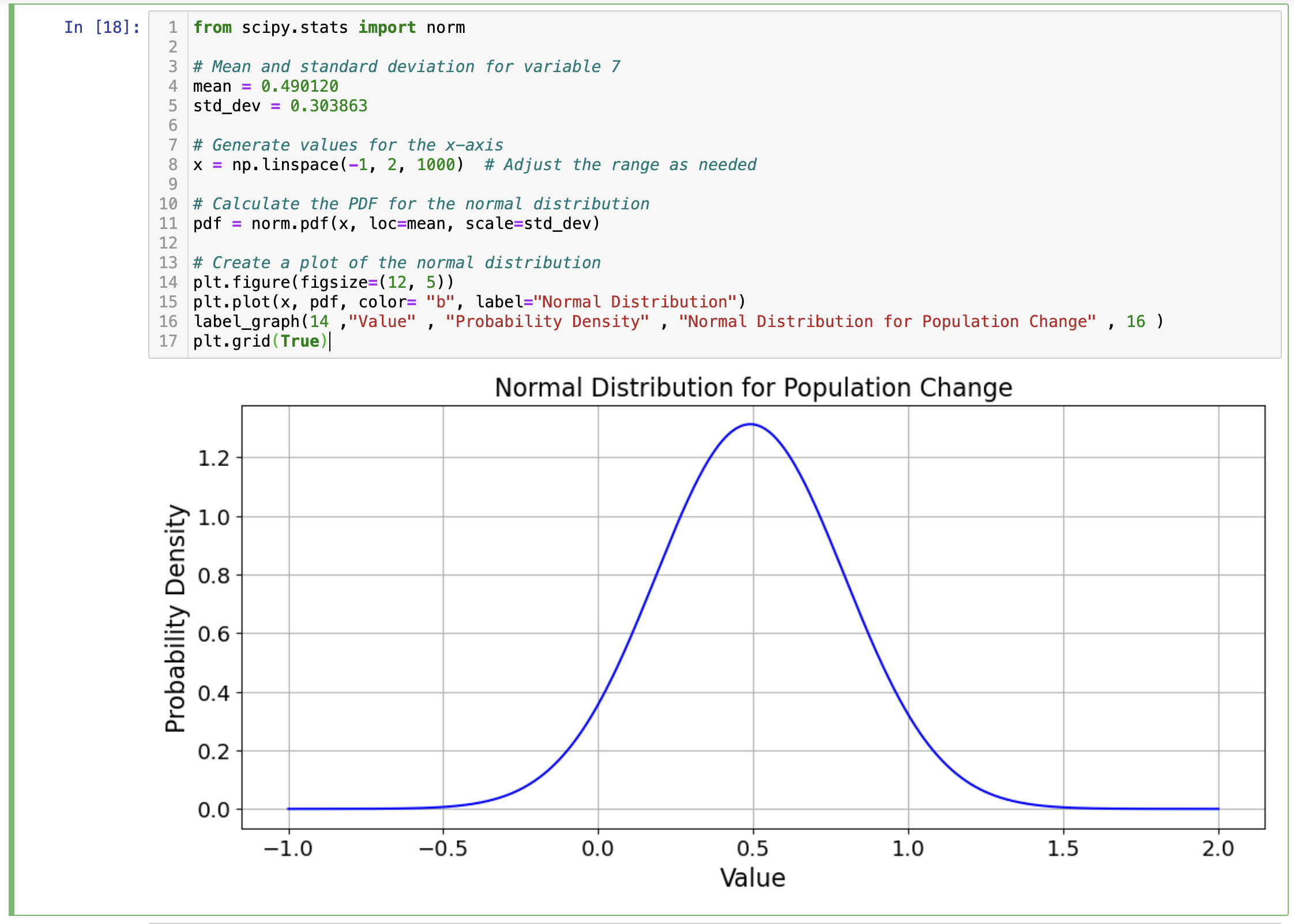
****

Figure 11 - PDF calculator

* Binomial distribution

Binomial distribution is a common discrete distribution used in statistics, going in the opposite direction of the normal distribution is computed by taking the probability of success to the power of the number of successes, multiplying it by the probability of failure to the power of the difference between the number of successes and the total number of trials. The result is then multiplied by the combination of the number of trials and successes. “*Binomial distribution is often used in social science statistics as a building block for models for dichotomous outcome variables. It also has applications in finance, banking, and insurance, among other industries*.” Said (Barone 2023) in his article.



Figure 12 - Piece of project code

In the code above performs a simulation of a binomial distribution to model the occurrence of non-citizens in a population.

The total population is 5,281,600 according our data set , and the number of non-citizens is 757,000. The probability of a person being a non-citizen is calculated by dividing the number of non-citizens by the total population. The probability of a person being a non-citizen is printed with a precision of four decimal places along with information about the total population and the number of non-citizens.

We took a sample size of 100 to simulate a binomial distribution. This represents the number of trials in the binomial experiment. Using NumPy's random module“*np.random”*, a binomial distribution is simulated with 100 trials (we had tested with a thousand trials which was taking too long), where each trial represents the probability of a person being a non-citizen. The results are stored in the “*binomial\_distribution”* array. The code then uses Matplotlib to create a histogram of the simulated binomial distribution. Bar charts were chosen for their ability to visually communicate categorical data, making them a widely used tool in visualizing data in various fields.

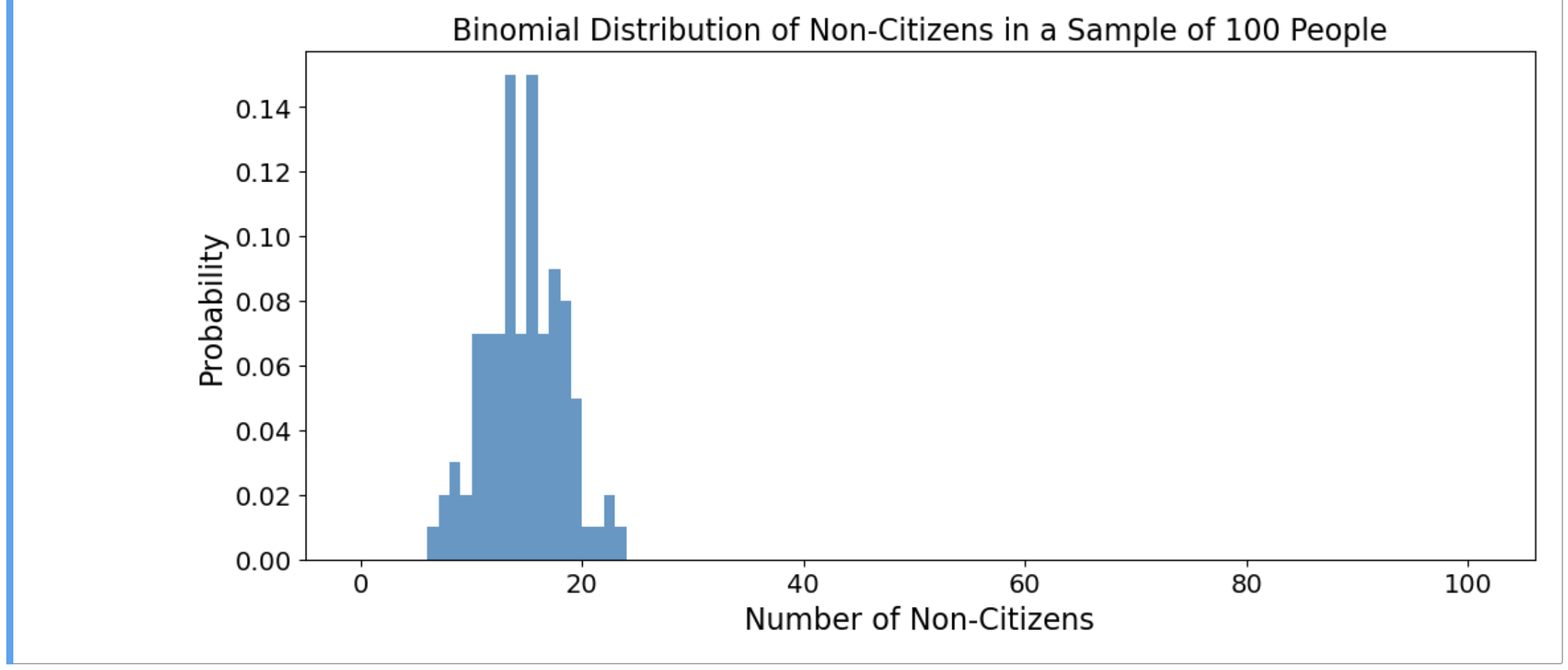


Figure 13 - Plot of Binomial Distribution

In summary, this code provides a practical demonstration of a binomial distribution by simulating the presence of non-citizens in a sample of 100 people from a larger population. The histogram visually represents the distribution of the number of non-citizens in the sample, showcasing the variability that can occur in such scenarios.

* Bernoulli Distribuition

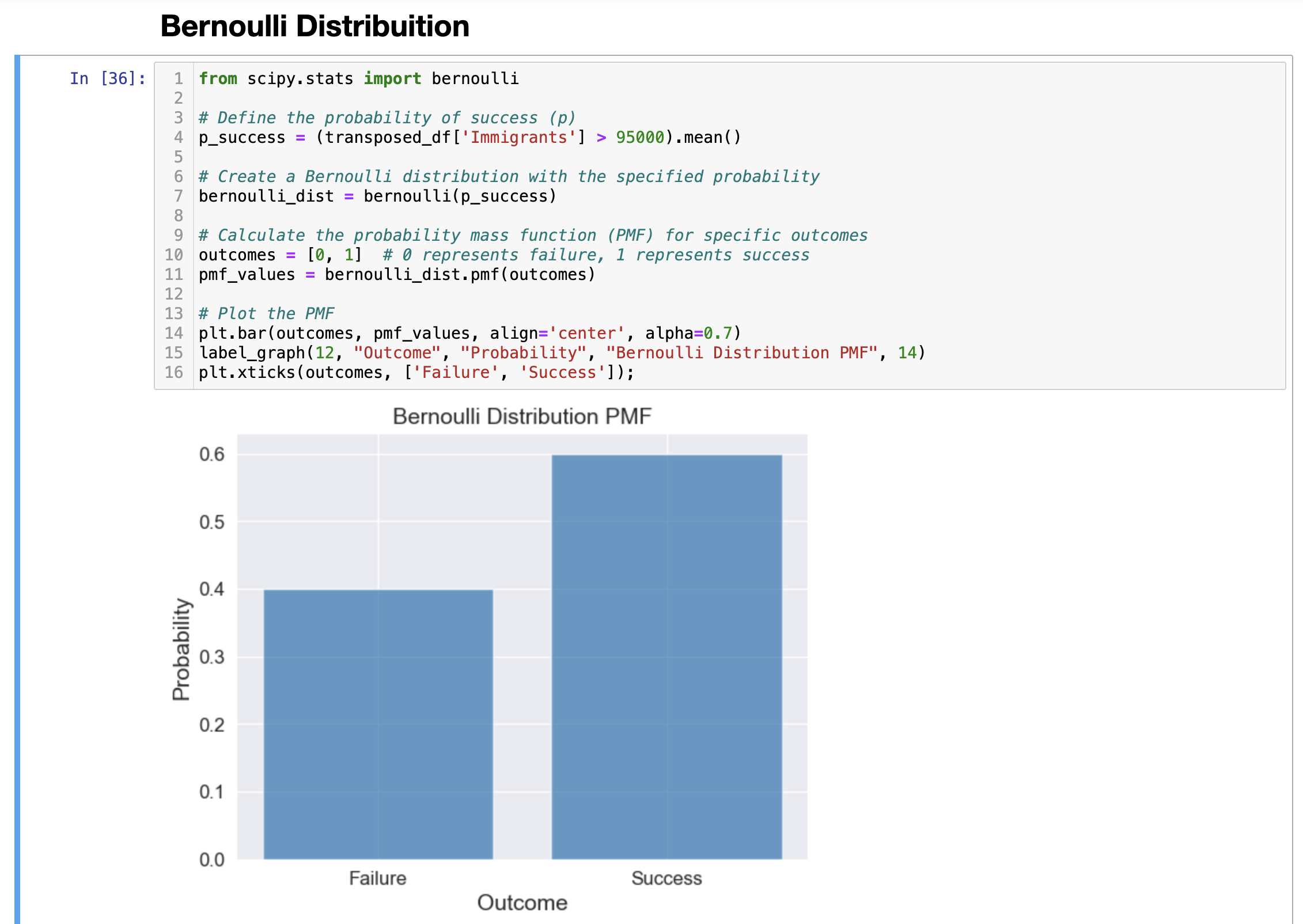
The Bernoulli distribution holds a pivotal role in probability theory, particularly as a discrete distribution, sharing notable similarities with the binomial distribution. Both distributions revolve around scenarios characterized by two possible outcomes in statistics: success (1) or failure (0).

Figure 14 - Bernoulli Distribution

In the context of Ireland's immigration patterns over the past decade, the total number of immigrants has been substantial, amounting to 932,200 individuals. This data provides a basis for exploring the application of the Bernoulli distribution to predict the likelihood of a specific outcome for the next year.

The mean of the past immigration data is a crucial statistical metric, providing a central tendency around which the distribution of immigrant numbers tends to cluster. In this case, the mean is calculated as 93,220 immigrants per year. Now, the question arises: Is it likely for Ireland to receive more than 95,000 new immigrants in the next year? This question involves comparing a specific threshold (95,000) with the mean of the past data. The application of the Bernoulli distribution involves considering the probability of success, where success is defined as the event of receiving more than 95,000 immigrants in the next year. The mean becomes a reference point for assessing the likelihood of surpassing this threshold. By evaluating the probability of success using the Bernoulli distribution, one can quantitatively assess the likelihood of the next year's immigration exceeding 95,000. If the calculated probability is greater than 50%, it indicates a higher chance of success compared to failure, suggesting that surpassing the mean is a plausible outcome.

In summary, the Bernoulli distribution, when applied to Ireland's immigration data, offers a statistical lens through which one can assess the likelihood of surpassing a specific threshold in the next year. This analysis provides valuable insights for decision-making and future planning, grounded in the principles of probability and statistical inference.

General overview about the distribution used in this project is discrete distribution is a probability distribution illustrating the presence of distinct and individually countable outcomes, such as 1, 2, 3, yes, no, true, or false. Like the example seen previously about Bernoulli Distribution, which gave us a Boolean result that there is a “YES” 60% probability that next year the number of immigrants will be greater than 95 thousand people. In contrast to a discrete distribution, a continuous probability distribution encompasses outcomes with any conceivable value, including fractional and indeterminate values. An exemplar of this is the normal distribution, visualized as a bell-shaped curve characterized by a continuous line that spans the entire range of values within its probability function.

**Project Management Framework**

The choice of a project management framework for a data science project depends on various factors, including the nature of the project, the goals, and the specific requirements. We going to have a brief overview of three commonly used frameworks in the context of data science projects which are:

1. Cross-Industry Standard Process for Data Mining (CRISP-DM)
2. Knowledge Discovery in Databases (KDD)
3. Sample, Explore, Modify, Model, Assess (SEMMA).
4. Cross-Industry Standard Process for Data Mining(CRISP-DM)

CRISP-DM is one of the most widely used frameworks for data mining and data science projects. It provides a structured and iterative approach to guide the entire data mining process from understanding the business problem to deploying the model.

Stages:

* Business Understanding
* Data Understanding
* Data Preparation
* Modelling
* Evaluation
* Deployment

CRISP-DM is suitable for projects where the emphasis is on the entire data mining lifecycle, including understanding business objectives, exploring and preparing data, building and evaluating models, and deploying solutions(IBM Documentation 2021).

1. Knowledge Discovery in Databases (KDD)

KDD is a broader concept that encompasses the entire process of discovering knowledge from data. While not a specific methodology like CRISP-DM, KDD emphasizes the overall process of turning data into actionable knowledge.

Stages:

* Selection
* Pre-processing
* Transformation
* Data Mining
* Interpretation/Evaluation
* Knowledge Presentation

KDD is suitable for projects where the focus is on the overall knowledge discovery process. It's more of a conceptual framework that includes various techniques and methods, and it is often used in conjunction with specific methodologies like CRISP-DM(KDD Process in Data Mining 2018).

1. SEMMA (Sample, Explore, Modify, Model, Assess):

SEMMA is a framework developed by SAS for data mining and analytics. It provides a structured approach that is particularly well-suited for business analytics and predictive modelling.

Stages:

* Sample: Select a representative sample from the data.
* Explore: Explore and visualize the data to gain insights.
* Modify: Prepare and pre-process the data for modelling.
* Model: Build predictive models.
* Assess: Evaluate and assess the performance of the models.

SEMMA is suitable for projects where the primary focus is on predictive modelling and analytics. It's often used in situations where there is a need to build and deploy models for decision-making(Kumar 2022).

In the specific dataset which we are exploring in this project, as we would like to predict future immigration and migration numbers based on historical data. It is recommended to choose supervised learning because we have labelled data (past immigration and migration numbers) and it is desired make predictions for future years.

As for a real-life scenario, let's consider a government agency responsible for managing migration trends in different patterns. We can use a supervised machine learning model to predict future migration patterns. For example, if there is a consistent net outflow of migration in certain areas (such as highly qualified professionals), they can allocate resources to foster, support or review policies to resolve the problem. This predictive capability can help with proactive decision making.

In summary, we would apply the CRISP-DM framework in a real situation but, as we are exploring the dataset in academic field we would go for KDD which is focused in the deep intention of learning process, and for ML modelling, supervised learning is justifiable as we have label data and it is given us a goal of predicting future migration trends.

## References

Barone A. 2023. Binomial Distribution: Definition, Formula, Analysis, and Example. Investopedia [Internet]. [accessed 2023 Nov 12]. https://www.investopedia.com/terms/b/binomialdistribution.asp

IBM Documentation. 2021. [accessed 2023 Nov 10]. https://www.ibm.com/docs/sr/spss-modeler/saas?topic=dm-crisp-help-overview

KDD Process in Data Mining. 2018. GeeksforGeeks [Internet]. [accessed 2023 Nov 10]. https://www.geeksforgeeks.org/kdd-process-in-data-mining/

Kumar M. 2022. Project Management in Data Science using SEMMA. Accredian [Internet]. [accessed 2023 Nov 10]. https://medium.com/international-school-of-ai-data-science/project-management-in-data-science-using-semma-a51ddf99cbd9

What is Exploratory Data Analysis? What is exploratory data analysis? [Internet]. [accessed 2023 Nov 7]. https://www.ibm.com/topics/exploratory-data-analysis