**MSC\_DA\_CA2**

**Ireland Milk Production Comparison with the EU Countries**

**Mahender Kumar**

**SBA22212**

Table of Contents

[**1. Report Description** 3](#_Toc123228159)

[**2. Abstract** 4](#_Toc123228160)

[**3. Introduction** 4](#_Toc123228161)

[**3. Exploratory data analysis (EDA)** 4](#_Toc123228162)

[3.1 Dealing with missing data and data cleaning 6](#_Toc123228163)

[3.2 Extract countries data where human population is between +/- 1 million 7](#_Toc123228164)

[**4. Handling outliers** 7](#_Toc123228165)

[**5. Descriptive statistics** 9](#_Toc123228166)

[5.1 Measures of Central Tendency 10](#_Toc123228167)

[5.2 Variation Measures 11](#_Toc123228168)

[**6. Poisson distribution** 14](#_Toc123228169)

[**7. Normal Distribution** 16](#_Toc123228170)

[**60. Conclusion** 19](#_Toc123228171)

[**70. References** 19](#_Toc123228172)

[**80. Appendix** 20](#_Toc123228173)

[5.1 Gantt Chart 20](#_Toc123228174)

Table of Figures

[Figure 1 Display first five rows of the livestock products data frame 4](#_Toc123228188)

[Figure 2 Display first five rows of the world population data frame 5](#_Toc123228189)

[Figure 3 Common methods to get the insight of data 5](#_Toc123228190)

[Figure 4 Merging two datasets 5](#_Toc123228191)

[Figure 5 null, nan, zero value in data frame 6](#_Toc123228192)

[Figure 6 Histogram plot to find outliers in Milk production value 7](#_Toc123228193)

[Figure 7 Boxplot to find out outliers in milk production value 9](#_Toc123228194)

[Figure 8 Mean Value 10](#_Toc123228195)

[Figure 9 Mode value 11](#_Toc123228196)

[Figure 10 Median Value 11](#_Toc123228197)

[Figure 11 Variation and Standard Deviation 12](#_Toc123228198)

[Figure 12 Box Plot configuration 13](#_Toc123228199)

[Figure 13.2 Histogram plot 14](#_Toc123228200)

# **1. Report Description**

Git hub repository address  
<https://github.com/Mahender1007/MSC_DA_ContAssignment2>

other stuff

why use colour on the graph

**Project outcome**Create a datasheet of dairy product over time (1960 to 2020) for Ireland and other EU nations. Data frame will be created from different dairy products, types of animals where dairy products are collected and total population of the Countries.

**Data Description**This data, which is in a csv file format, consists of Year, Ireland and other European countries. These columns contain information about the columns mentioned from 1960 to 2020.

**How Was This Data Collected?**Crop and Food data were obtained from the Food and Agricultural Organisation of the United Nations. FAO collects the data from all the countries in the world. FAO is an international organisation and it is a part of the United Nations. I am using raw data to form the base of my data frames. FAO is a website that provides reliable data in various fields.For more detailed information, please click the link below. Link:<https://www.fao.org/faostat/en/#data/QCL>

Total yearly population data are collected from EUROSTAT which is freely available and there is no licensing requirement to access these data. This population is then compared to the other 27 countries in the EU. Four countries are identified as having a similar (+/- 1m people) population in 2020. The population data for these countries is downloaded. They can only be downloaded separately.  
For more detailed information, please click the link below. Link:  
<https://data.worldbank.org/indicator/SP.POP.TOTL?end=2020&start=1960>

**Is There a License for the Data Used?**The FAO and EUROSTAT websites where the data were obtained, has a Creative Commons Attribution 4.0 International license. This license type makes shared data publicly available according to open data standards and license datasets. Data are freely available and no licence is required to access or download the data.

# **2. Abstract**

# **3. Introduction**

Add introduction here

Why choose this data

Data explanation

# **3. Exploratory data analysis (EDA)**

EDA is first step in the data analysis process, it is an approach to analysing data sets to summarise their main characteristics. It allows to check and analyse the data before we make any conclusion or assumption. EDA also ensures that the results produced are valid and applicable to the business needs (Wes McKinney 2017).

Following the benefits, we begin using EDA on the data sets.

* It helps to understand the dataset variable and the relationship among the data
* It provides better knowledge of the data set
* Helps to identify if there are any errors in our dataset. I.e., Duplicates, missing data
* Helps detect outliers or anomalies

In my project, data are collected from 3 datasheets. These datasheets consist milk production data from all around the world. Data are gathered over a period of 61 years (1961-2021).

The following are 3 types of datasets.

1. Livestock products
2. World population
3. World population in farming

After loading the data from csv files into the data frame using Pandas library, simple tasks were performed to get an insight of data.

Initially, I worked on following the two data sets  
a. lp\_df (Livestock products)

b. wpt\_df (World population)

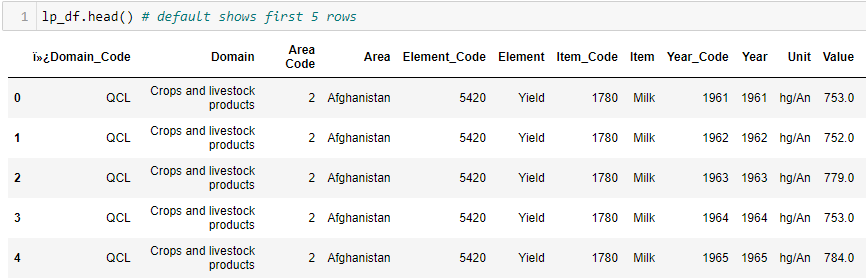


Figure 1 Display first five rows of the livestock products data frame

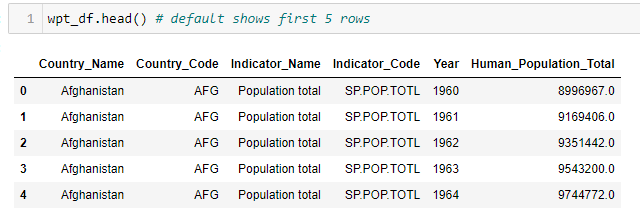


Figure 2 Display first five rows of the world population data frame

I also performed the following methods on the DataFrame to get more information

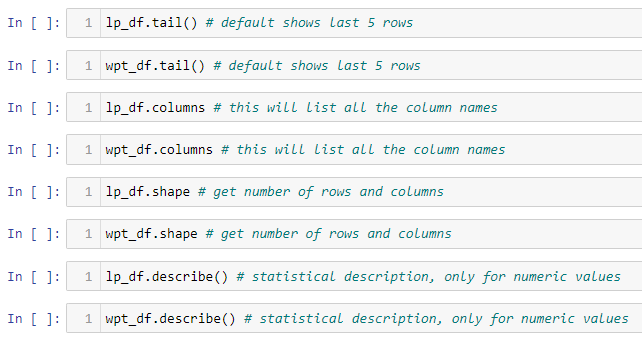


Figure 3 Common methods to get the insight of data

I also did some cleaning of data fin data frame lp\_df (Livestock products). I merged all unique “Element” row into column. [Please see section 2.0.3 in jupyter notebook]

*Merging the data frames*

Data frame lp\_df (Livestock products) and wpt\_df (World population) were merged into one dataframe (Please see section 2.0.4 in jupyter notebook)

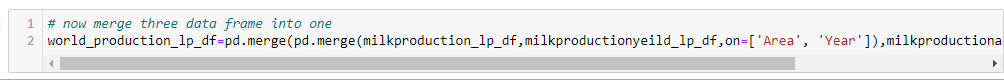


Figure 4 Merging two datasets

*Extraction of European country data from the main data frame*

This project evaluates milk production only in European countries so I need to extract European countries data from the main data frame world\_production\_inc\_lp\_df. Library countrygroups was used to get the list of countries in EU union. (Please see section 2.0.5 in jupyter notebook)

## 3.1 Dealing with missing data and data cleaning

One of the most common issues in dataset are missing values. Data can either be missing during data extraction or collection due to corruption or data not being recorded. I checked both datasheets for missing and empty(zero) data.

Below are some formats that could be in the missing data:

* n/a
* NA
* —
* na
* NaN

Why is it important to deal with missing data?

Missing data are important because, depending on the type, they can sometimes bias your results. This means results may not be generalisable outside of study because data come from an unrepresentative sample.

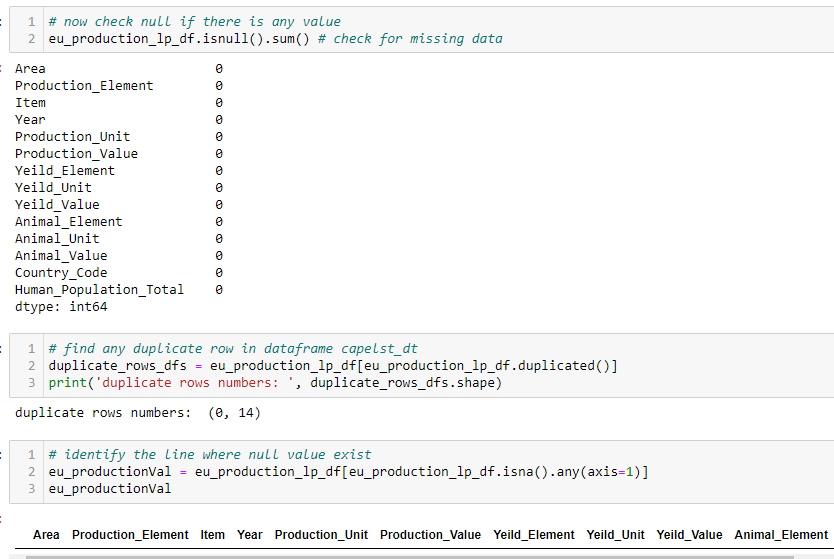


Figure 5 null, nan, zero value in data frame

From the above (see figure 5) observation I found that are no missing or null value in the data frame. However, if there are missing or null value then they can be fixed by deleting the row or replacing it by mean, mode or with the other values. If data base is small then it best to replace the null value instead deleting it. Mean value can be used to replace the missing value columns if data distribution is symmetric.

## 3.2 Extract countries data where human population is between +/- 1 million

I also overserved the dataset and extracted the data where countries population are +/- 1 million in year 2020.

Upon checking I identified the following 4 countries where human population are close to Ireland human population.

1. Slovakia
2. Denmark
3. Croatia
4. Finland

# **4. Handling outliers**

This is one of the most important parts of the data pre-processing and treating the outliers, as they can negatively affect the statistical analysis resulting in less accuracy. The main reason they occur is due to variability in the data, or human error. It is good practice to detect and remove these outliers from the data as it could also give biased results. The simplest way to detect an outlier is by graphing the features or the data points. Visualisation is one of the best and easiest ways to have an inference about the overall data and the outliers. Scatter plots and box plots are the most preferred visualisation tools to detect outliers. Other method to detect outliers is using IQR (Interquartile Range) techniques (Chun-houh, Wolfgang, and Antony 2008).

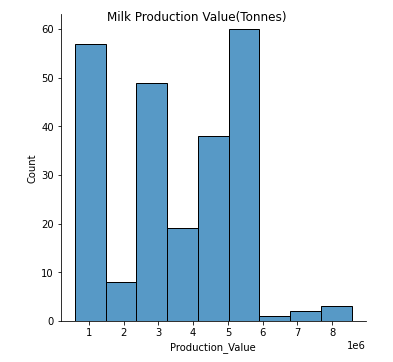
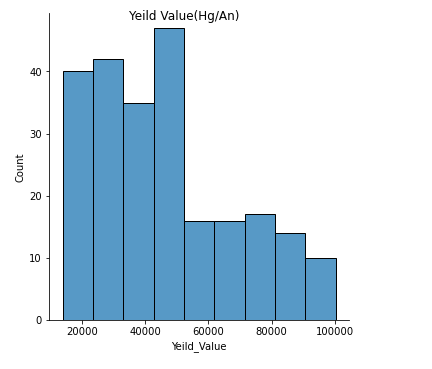
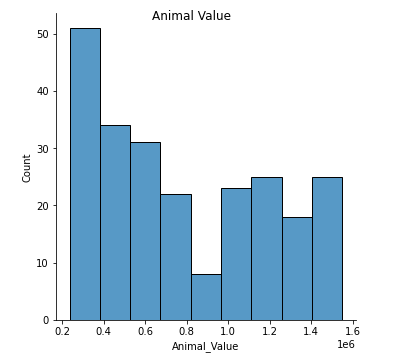
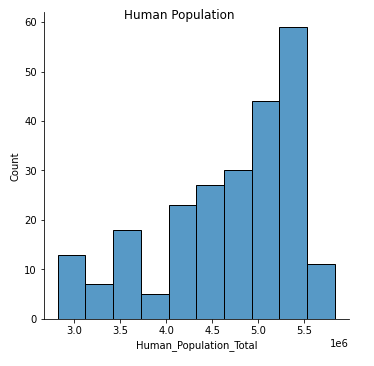


Figure 6 Histogram plot to find outliers in Milk production value







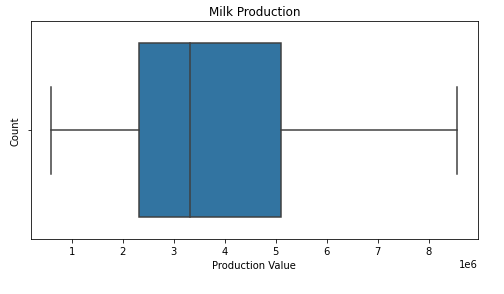


Figure 7 Boxplot to find out outliers in milk production value

After visualising and analysing the data frames, (See figure 6 and 7) I noticed there was no outliers in the data frame. Luckily data sets I am working have no anomalies and I did not have to take any further steps. If there are outliers in the datasets then following steps can be taken to handle them.

1. Remove or change outliers during post-test analysis.
2. Change the value of outliers.

# **5. Descriptive statistics**

Descriptive statistics were used to summarise and describe the characteristics of the data set. It was important to use these descriptive statistics as they can be useful for two purposes:   
a. to provide basic information about variables in a dataset and   
b. to highlight potential relationships between variables

## 5.1 Measures of Central Tendency

*Mean*This is a typical average in dataset column. The following formula is used to calculate the mean value;

\bar{x} = \dfrac{\sum x}{n}

Mean value is calculated in Ireland data frame Ireland\_production\_inc\_lp\_df on column Production\_Value

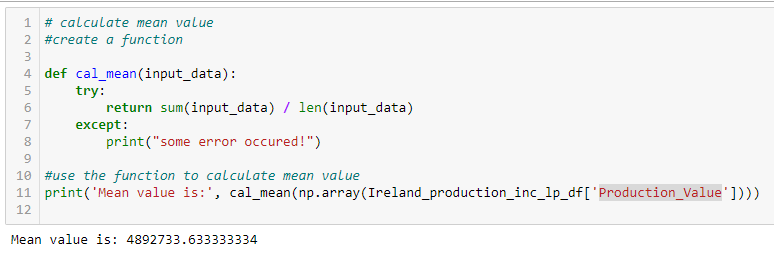


Figure 8 Mean Value

**Mean value can be used to replace null or empty columns in dataset**

*Mode*The mode value of a data set is the most frequently occurring value. It tells that the most popular choice or the most common characteristic in the datasets.

Mode value is calculated in data frame Ireland\_production\_inc\_lp\_df on column Production\_Value

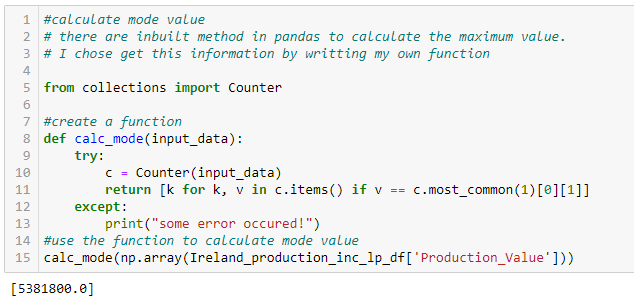


Figure 9 Mode value

**Mode value can be used to replace null or empty columns in dataset**

*Median*Median represents the middle value for the dataset. It is the point at which half of the data is more and half the data is less. Median helps to represent a large number of data points with a single data point.

Median value is calculated in the data frame Ireland\_production\_inc\_lp\_df on column Production\_Value



Figure 10 Median Value

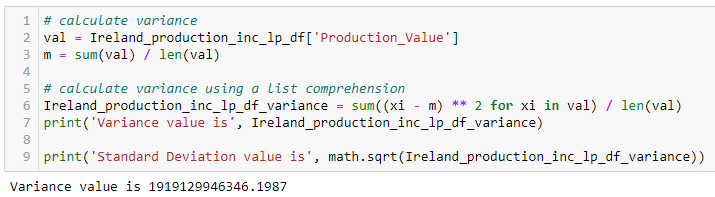
**Median value can be used to replace null or empty columns in dataset**

## 5.2 Variation Measures

*Variance*The variance is a measure of variability. It is calculated by taking the average of squared deviations from the mean. Variance tells the degree of spread in the data set. The more spread the data is, the larger the variance is in relation to the mean.

s^2= \dfrac{\sum (X - \bar{x})^2}{n - 1}

Variance value is calculated in data frame Ireland\_production\_inc\_lp\_df on column Production\_Value



*Standard Deviation*Standard deviation is the square root of the variance. The Standard Deviation is a measure of how spread-out the numbers are.  
Standard Deviation value is calculated in the data frame Ireland\_production\_inc\_lp\_df on column Production\_Value

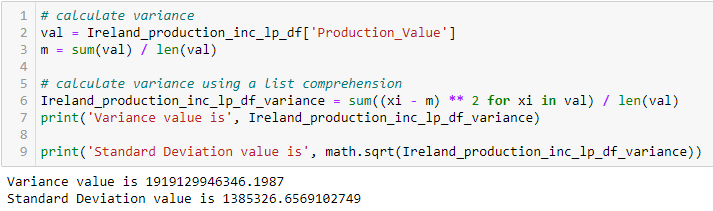


Figure 11 Variation and Standard Deviation

*Box Plot*I used Box and Whisker or Box Plot to statistically represent a column in the data set. A Box Plot is the visual representation of the statistical five number summary of a given data set.

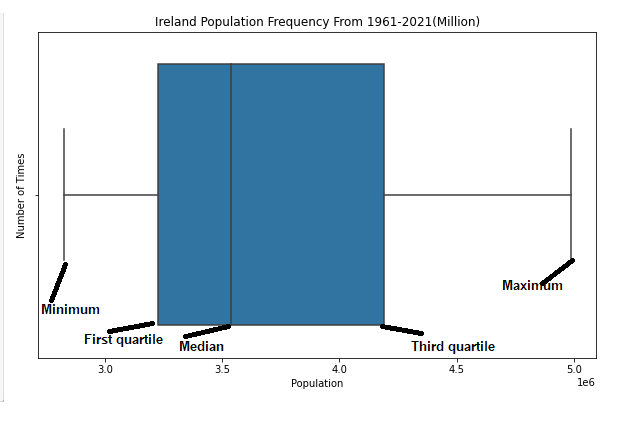


Figure 12 Box Plot configuration

The following are the dimensions:

1. Minimum
2. First Quartile
3. Median
4. Third Quartile
5. Maximum

Box plots portray the distribution of data, outliers, and the median. The box within the chart displays where approximately 50 percent of the data points fall. As shown in fig 12 it summarises a data set in five marks. The mark with the greatest value is called the maximum, and it will likely fall far outside the box. The mark with the lowest value is called the minimum, and it will likely fall outside the box on the opposite side of the maximum.

The box itself contains the first or lower quartile, the third or upper quartile, and the median in the centre. The median is the value separating the higher half from the lower half from the data frame where a total number of pedestrians crossed in a day throughout the year. The median value shown can also be deemed as "the middle" value in a set of numbers based on a count of column values rather than the middle based on a numeric value. The sections in a box plot help the viewer see where the median falls within the distribution. The lower quartile is the 25th percentile, while the upper quartile is the 75th percentile. The median is the middle, but it helps give a better sense of what to expect from these measurements. The whiskers (the lines extending from the box on both sides) typically extend to 1.5\* the Interquartile Range (the box) to set a boundary beyond which would be considered outliers. The outliers are individual dots that occur outside the upper and lower extremes (Chun-houh, Wolfgang, and Antony 2008).

*Histogram Plot*A histogram is a graph that shows the frequency of numerical data using rectangles. The height of a rectangle (the vertical axis) represents the distribution frequency of a variable (how often that variable appears). The width of the rectangle (horizontal axis) represents the value of the variable (Total number of Pedestrians).

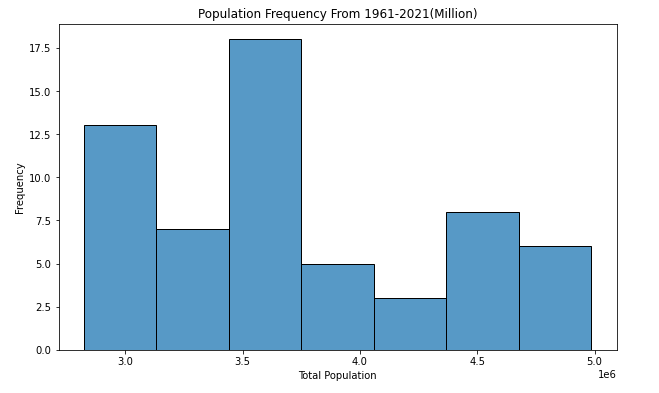


Figure 13 Histogram plot

Histograms were useful to statistically represent the data as there was a need to display a comparison of the highest number of pedestrians that crossed throughout a year and the number of times it occurred on Capel Street. In fig 5.2, it can be observed that there was a skewed distribution of data (Google 2022)

The anatomy of a histogram fig 5.2  
It was a visual representation of a data set which show how often each value in the data set pdnfootfall\_dt on column capelst\_pdntotal occurs. The values are grouped into bins along the x-axis. The height of the bar indicates how many values of the data set fall into that bin. In above fig 5.2, a histogram shows all of the Capel Street total pedestrian number values from a data set of pedestrian footfalls. These 365 row values are sorted into 17 bins; the first bin includes the total pedestrian count that occurred are 1000 to 2000; the second and third bin includes the total pedestrian count that occurred are 3000 to 4000 and so on and so forth. The height of the blue bar indicates the monotonal pedestrians crossed in a day, along in each bin. There were more than 60 times when 6500 or more pedestrians crossed the street in a day.

# **6. Poisson distribution**

A Poisson distribution is a discrete probability distribution, it gives the probability of a discrete (i.e., countable) outcome.

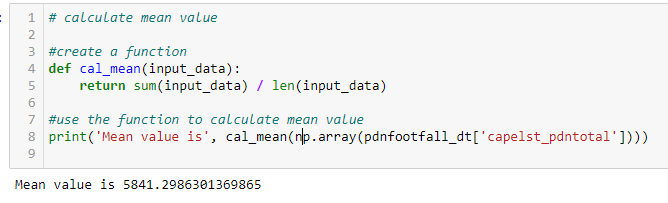
I used Poisson distribution to predict the number of pedestrians crossing within a given interval of time. In my dataset, individual counts occurred at random and independently. That is, the probability of one event was not affecting the probability of another event.

Furthermore, I was aware of the mean number of events occurring within one year and this number is called λ (lambda), and it was a constant number.

The following is the formula to calculate the Poisson distribution.



*Calculating mean value*

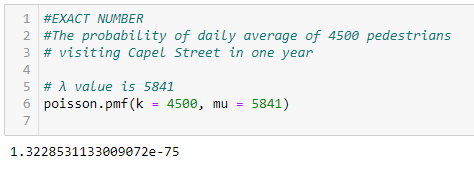


In my data sets column λ value is 5841

The below criteria were tested

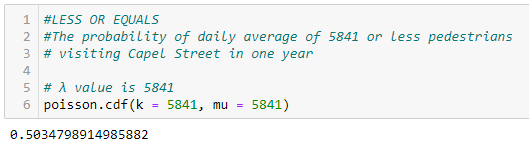
1. The probability of the daily average of 4500 pedestrians visiting Capel Street in one year

*Answer: 1.32*



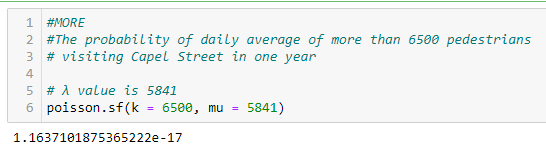
1. The probability of daily the average of 5841 or less pedestrians visiting Capel Street in one year

*Answer: 0.50*



1. The probability of the daily average of more than 6500 pedestrians visiting Capel Street in one year

*Answer: 1.16*



# **7. Normal Distribution**

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequently occur than data that is further from the mean. The normal distribution is easy to work with mathematically.

I implemented normal distribution on data frame pdnfootfall\_dt on column henryst\_pdntotal. These data refer to total pedestrian visited during the year 2015.

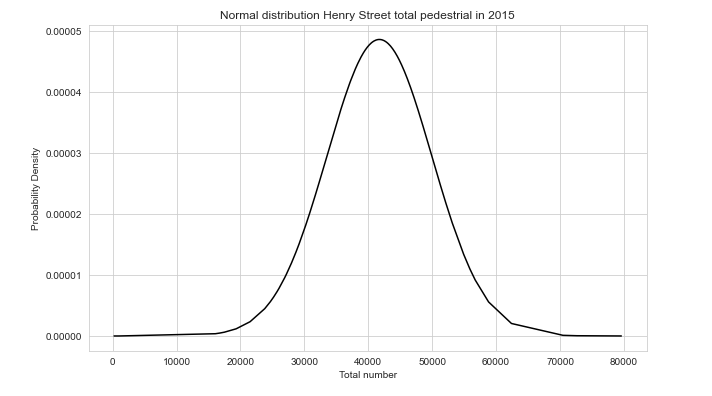
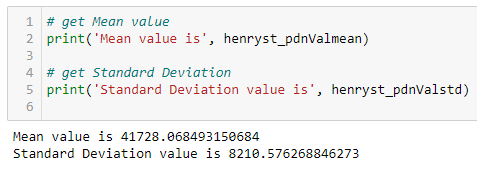


Figure 7.1 Normal distribution



Mean value is 41728 and the Standard deviation is 8210.

The normal distribution density function simply accepts a data point along with a mean value and a standard deviation and throws a value which we call probability density. The shape can be altered by changing the mean and the standard deviation. Changing the mean will shift the curve towards that mean value, this means we can change the position of the curve by altering the mean value while the shape of the curve remains intact. The shape of the curve can be controlled by the value of Standard deviation. A smaller standard deviation will result in a closely bounded curve while a high value will result in a more spread-out curve.

*Some properties of a normal distribution:*

1. The mean, mode, and median are all equal.
2. The curve is symmetric around the mean.

To find the probability of a value occurring within a range in a normal distribution, I needed to find the area under the curve in that range. i.e., I need to integrate the density function. Since the normal distribution is a continuous distribution, the area under the curve represents the probabilities. A standard normal distribution is just similar to a normal distribution with mean = 0 and standard deviation = 1.

Following is the formula to calculate the z value;



The z value above is also known as a z-score. A z-score gives you an idea of how far from the mean a data point is.

1. To find the probability of a total number of people that visited the streets that has a value of less than or equal to 30000. (Using Cumulative Density Function)

X = 30000  
μ = 41728  
σ = 8210

*Answer:*To calculate this probability I firstly needed to follow standardisation process. This process will calculate the requested data points.

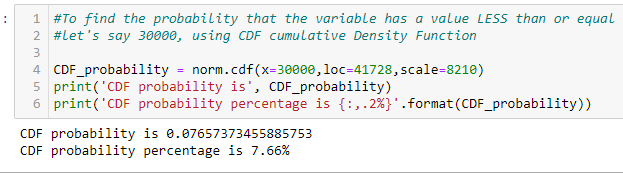


Figure 7.2 CDF calculation

As shown in fig 7.2 the probability of the total number of people that visited the streets that has a value of less than or equal to 30000 is 7.66%.

1. To find the probability of the total number of people visited the streets that has a value of greater than or equal to 50000. (Using Survival Function)

X = 50000  
μ = 41728  
σ = 8210

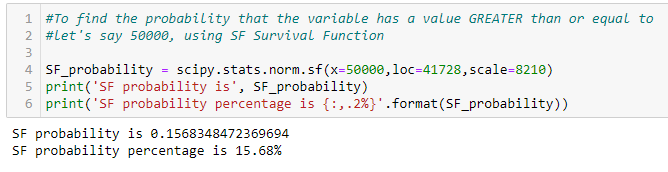
*Answer:*  


Figure 7.3 SF calculation

As shown is fig 7.3 the probability shows the total number of people that visited the streets that has a value of greater than or equal to 50000 is 15.68%.

# **60. Conclusion**

# **70. References**

***Books***Andreas, C. M. and Sarah, G. and O'Reilly (2016) Introduction to Machine Learning with Python, Media, Inc.

Hands on Machine Learning with Scikit Learn, Keras, and TensorFlow, 2nd Edition, Aurélien Géron, O'Reilly Media, September 2019, ISBN: 9781492032649.

Python Machine Learning Third Edition, Sebastian Raschka , Vahid Mirjalili , Copyright © 2017 Packt Publishing.

Discovering Knowledge In Data: An Introduction To Data Exploration, Second Edition, By Daniel Larose And Chantal Larose, John Wiley And Sons, Inc., 2014.

Chun-houh, C. and Wolfgang, H. and Antony, U. (2008) Handbook of Data Visualization, Springer publishing.

Wes, M. (2017). Python for Data Analysis, 2nd Edition.

Perez (2021). Data Mining. The CRISP-DM Methodology. The CLEM language and IBM SPSS MODELER.

Ojala and Garriga (2010) Permutation Tests for Studying Classifier Performance.

***Websites***

Google (2022) Machine Leaning. Available at: https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML#:~:text=Machine%20learning%20(ML)%20is%20a,to%20predict%20new%20output%20values. (Accessed: 17 Nov 2022).

Google (2022) Exploratory data analysis. Available at: https://businessanalyst.techcanvass.com/objective-of-exploratory-data-analysis/ (Accessed: 11 Nov 2022).

Google (2022) Handling outliers. Available at: https://towardsdatascience.com/exploratory-data-analysis-topic-that-is-neglected-in-data-science-projects-9962ae078a56. (Accessed: 12 Nov 2022).

Google (2022) Normal Distribution. Available at: https://www.askpython.com/python/normal-distribution (Accessed: 12 Nov 2022).

Google (2022) Pedestrian footfall. Available at: https://www.retailsensing.com/people-counting/count-pedestrian-footfall-town-centre/ (Accessed: 22 Nov 2022).

Google (2022) Linear regression. Available at: https://realpython.com/linear-regression-in-python/(Accessed: 20 Nov 2022).

Google (2022) Supervised Leaning. Available at: https://www.sciencedirect.com/topics/computer-science/supervised-learning (Accessed: 15 Nov 2022).

Google (2022) Data Science project management methodologies. Available at: https://medium.datadriveninvestor.com/data-science-project-management-methodologies-f6913c6b29eb (Accessed: 13 Nov 2022).

Google (2022) Descriptive analysis. Available at: https://www.analyticssteps.com/blogs/overview-descriptive-analysis (Accessed: 18 Nov 2022).

Google (2022) Descriptive analysis. Available at: https://www.labxchange.org/library/items/lb:LabXchange:10d3270e:html:1 (Accessed: 22 Nov 2022).

# **80. Appendix**

## 5.1 Gantt Chart