­CCT College Dublin

|  |  |
| --- | --- |
| Module Titles: | Data Preparation & Visualisation, Machine Learning for Data Analytics, Programming for Data Analytics, Statistics for Data Analytics |
| Assessment Title: | Integrated CA, Ireland Tourism |
| Lecturer Names: | David McQuaid, Dr. Muhammad Iqbal, Sam Weiss, Taufique Ahmed |
| Student Full Name: | Ricardo Alves de Souza |
| Student Number: | 2018499 |
| Assessment Due Date: | 07/04/2024 |
| Date of Submission: | 12/04/2024 |

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. |

Title: Integrated CA, Ireland Tourism

Author: Ricardo Alves de Souza

E-mail: [2018499@student.cct.ie](mailto:2018499@student.cct.ie?subject=Integrated%20CA1)

GitHub Address: <https://github.com/ricardoasouz/2024---MSc-in-Data-Analytics---Feb---FT/tree/main/MSC_DA_CA1>

Table of Contents

[Abstract 5](#_Toc163562969)

[Introduction 5](#_Toc163562970)

[EDA Exploratory Data Analysis 6](#_Toc163562971)

[Obtaining The Data 6](#_Toc163562972)

[Observing the “.CSV” file 7](#_Toc163562973)

[Preparing the Dataset for Machine Learning and Statistics 9](#_Toc163562974)

[Plotting the Dataset 14](#_Toc163562975)

[TIME SERIES ANALYSIS 17](#_Toc163562976)

[Descriptive Statistics 21](#_Toc163562977)

[Describing the data 21](#_Toc163562978)

[Common descriptive statistics in general include: 22](#_Toc163562979)

[Difference between Discrete Data and Continuous Data. 27](#_Toc163562980)

[Definition of discrete data. 27](#_Toc163562981)

[Definition of continuous data. 27](#_Toc163562982)

[Definition of Descriptive Statistics 28](#_Toc163562983)

[Some common formulas and calculations used in Descriptive Statistics 28](#_Toc163562984)

[Minimum 28](#_Toc163562985)

[Maximum 28](#_Toc163562986)

[Range 28](#_Toc163562987)

[Sum 29](#_Toc163562988)

[Size / Count 29](#_Toc163562989)

[Mean 29](#_Toc163562990)

[Median 30](#_Toc163562991)

[Mode 31](#_Toc163562992)

[Standard Deviation 31](#_Toc163562993)

[Variance 31](#_Toc163562994)

[Midrange 32](#_Toc163562995)

[Quartiles 32](#_Toc163562996)

[Interquartile Range 33](#_Toc163562997)

[Outliers 33](#_Toc163562998)

[Sum of Squares 33](#_Toc163562999)

[Mean Absolute Deviation 34](#_Toc163563000)

[Root Mean Square 34](#_Toc163563001)

[Standard Error of the Mean 35](#_Toc163563002)

[Skewness 35](#_Toc163563003)

[Kurtosis 37](#_Toc163563004)

[Coefficient of Variation 38](#_Toc163563005)

[Relative Standard Deviation 38](#_Toc163563006)

[Frequency 39](#_Toc163563007)

[Reference list 41](#_Toc163563008)

­

[Figure 1 - Table of Some Descriptive Statistics 7](#_Toc162356614)

­­

# Abstract

*An initial examination of the dataset in data science is crucial for gaining insights into its features, uncovering interconnections, detecting emergent trends, and conditioning the dataset for more advanced analytical processes, which might include predictive modelling or time series forecasting.*

# Introduction

An initial examination of the dataset in data science is crucial for gaining insights into its features, uncovering interconnections, detecting emergent trends, and conditioning the dataset for more advanced analytical processes, which might include predictive modelling or time series forecasting

# EDA Exploratory Data Analysis

Exploratory Data Analysis is an important process performed in data science to understand the data and its characteristics, to identify correlations, to extract patterns and to prepare the data for further stages of analysis such as forecasting or prediction. (Aloorravi, 2024)

## Obtaining The Data

This study has been utilised, to acquire data to analyse from the website of the government of Ireland, “CSO Central Statistics Office” in March of 2024.

https://www.cso.ie/en/statistics/tourismandtravel/

The segment chosen for analysis was:

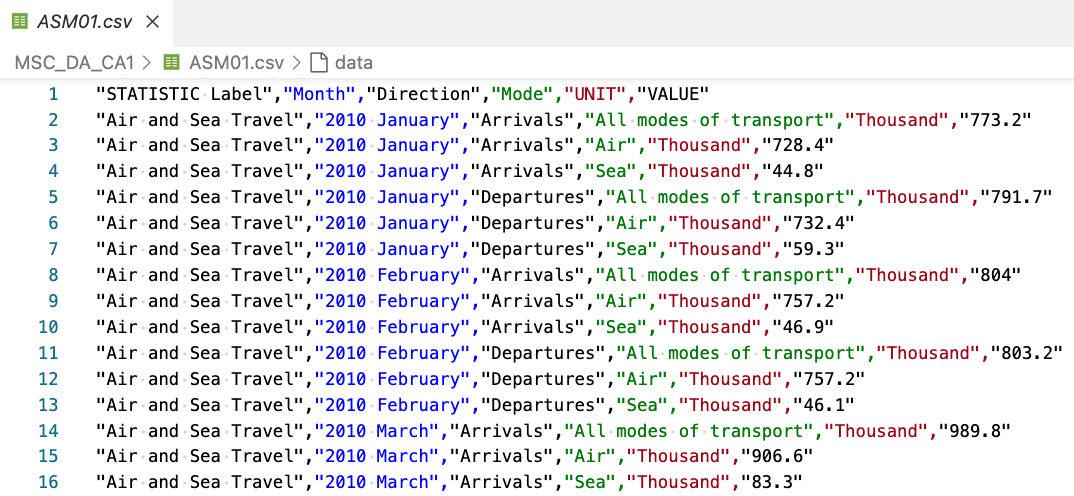
“Air and Sea Travel Statistics”

The file can be obtained from this internet address: “https://data.cso.ie/table/ASM01”, in multiple formats like, (“CSV 1.0”, “JSON-stat 2.0”, “JSON-stat 1.0”, “PX 2013”, “XLSX 2007”).

The format chosen in this case is “CSV 1.0”, an acronym for Comma Separated Values, and the file are “ASM01.csv”, Air and Sea Travel.

This dataset brings to us quantities of Arrives and Departures by Air and Sea in Ireland in the period about January 2010 to January 2024, separated by month with your values in thousands.

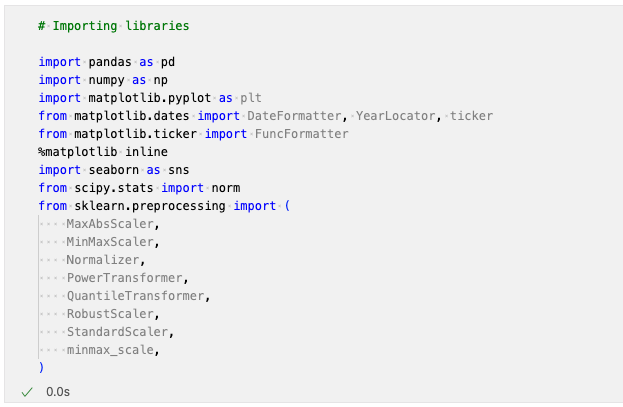
## Observing the “.CSV” file



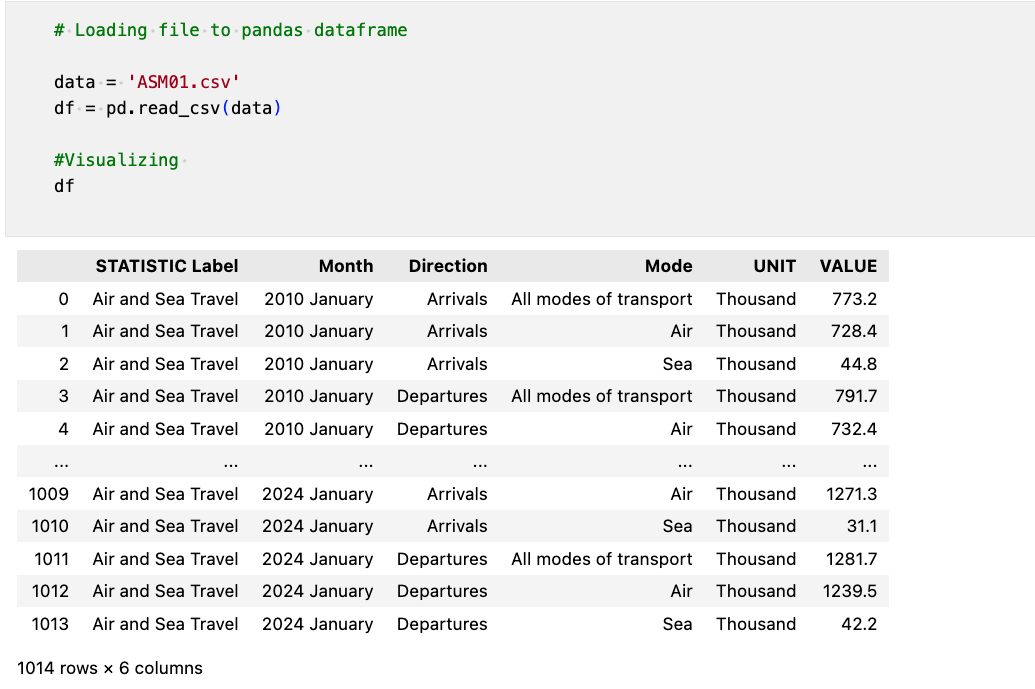
The ASM01.csv came with 1015 lines, with this structure can be seen in the previous picture. The first line with labels of columns: “STATISTIC Label”, “Month”, “Direction”, “Mode”, “UNIT”, “VALUE”, and on sequence your respective values.

Python is the computer program language chosen for this study of the case, with code in the format of Jupiter Notebook and with a complement of some libraries like Pandas, NumPy, MatPlotLib, SciPy, SeaBorn…

The “as” method it’s to create an alias name for librarie.

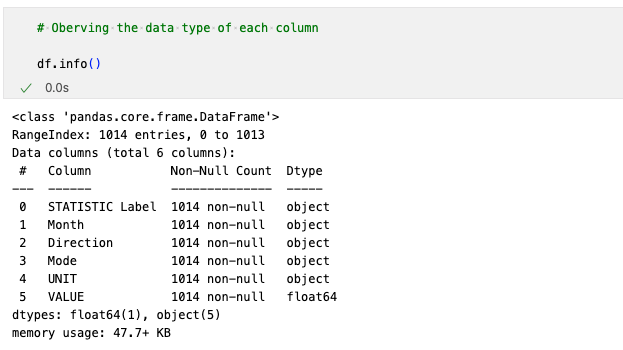
­­

Let’s create a “Pandas” data frame and visualize, for it only needs to use your alias “pd”.



­ After importing some libraries, and loading the file ‘ASM01.csv’ to a “Pandas” data frame, “Jupiter Notebook” can show us the five first rows and the last five rows of the data frame and say he has 1014 rows with 6 columns.

Let’s observe the data type of each column utilizing the “info()” method from “Pandas”



It is possible to observe that almost all columns are in categorical format “object”, just one column “VALUE”, is in numerical format “float64”. And all columns don’t have any null value, “non-null”.

To perform statistical calculations, summarization and or start creating comprehensive graphics of this dataset it needs to convert these values to numerical format and organize them in a way to makes more easier to do these tasks.

## Preparing the Dataset for Machine Learning and Statistics

In his book The Visual Display of Quantitative Information, Edward Tufte discusses his famous six principles of graphical integrity. The six principles of graphical integrity are (comparisons, causality, multivariate, integration, documentation, and context).

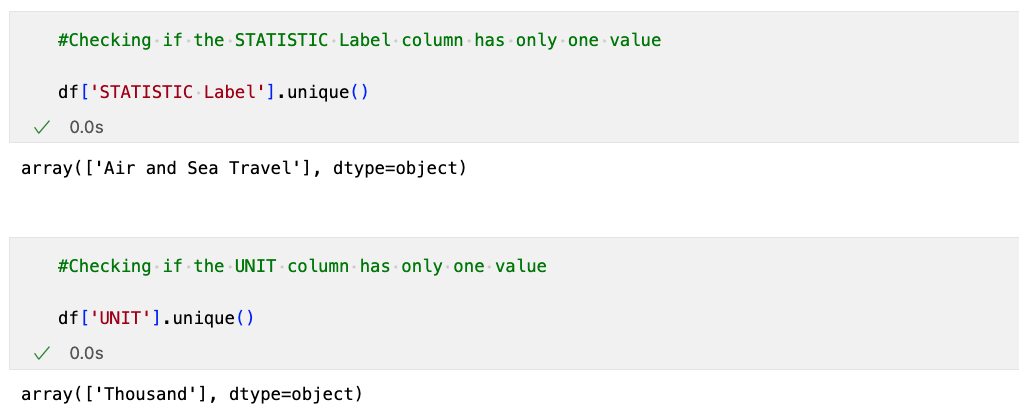
Also, he describes the best practices for data visualization:

* Above all else, show the data. Create the simplest graph that conveys the information you want to present.
* Maximize the data-ink ratio. Every bit of ink requires a reason. ...
* Erase non-data-ink.
* Erase redundant data-ink.
* Revise and edit. (Tufte, 1997)

Let’s try to prepare the data set keeping these principles in mind.

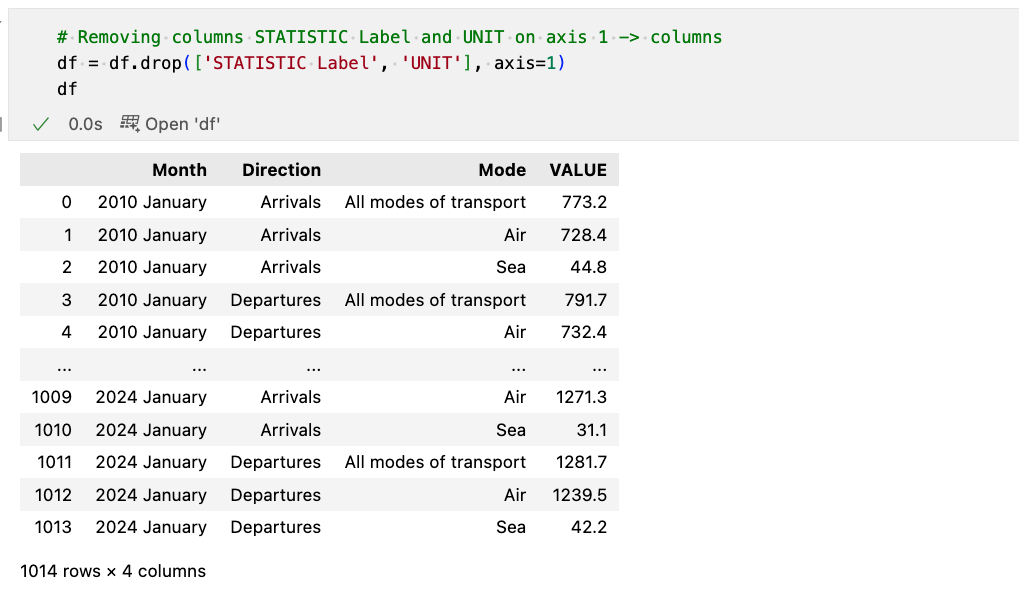
First of all, let’s confirm if the columns, “STATISTIC Label”, and “UNIT”, have just the same value for all rows, making them “Air and Sea Travel”, and “Thousand” redundant values.

For this, it’s possible to utilize the method “Unique()”, he answers us with a list of unique values found in the column.



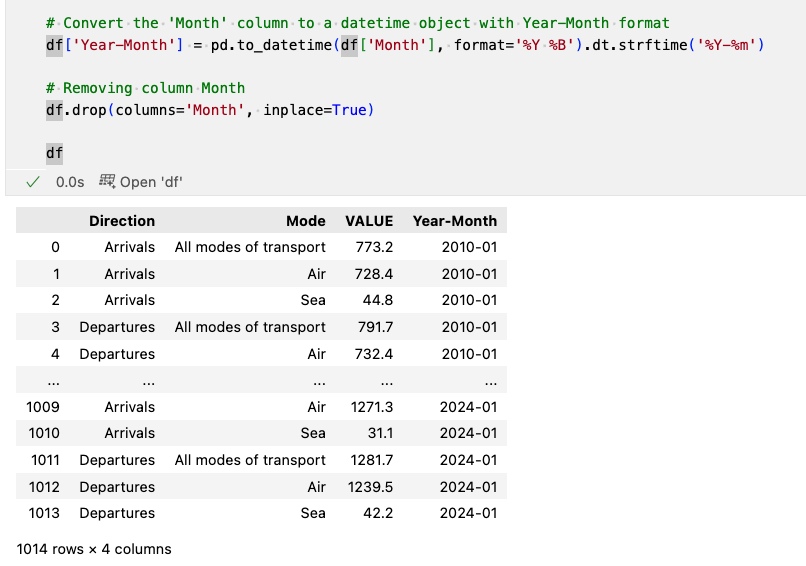
The answer is yes, having just one value in the entire table makes these values redundant. But do not lose this important information “Thousand”, means the scale of values, let’s move this to the title label of columns later.

Let’s perform the remotion of these columns utilizing the method “drop()”, cleaning the dataset and visualizing the result.

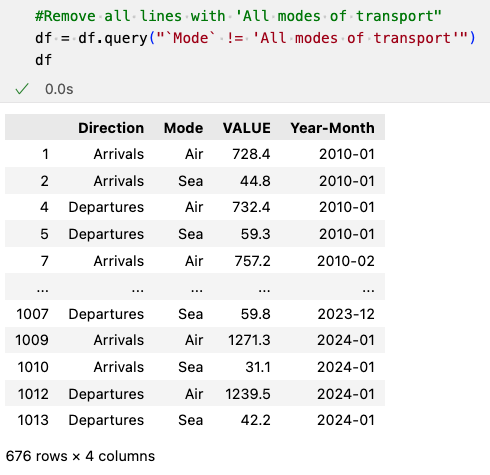


The result is now the data frame has only four columns, but still with too many categorical values.

The column “Month” can be transformed into, the “DateTime” type, for this let’s utilize the method “to\_datetime()” of Pandas, converting the values to the string format “Year-Month”, perfect to visualize without the “day”, redundant information on this case, and removing the original column “Month” after this.

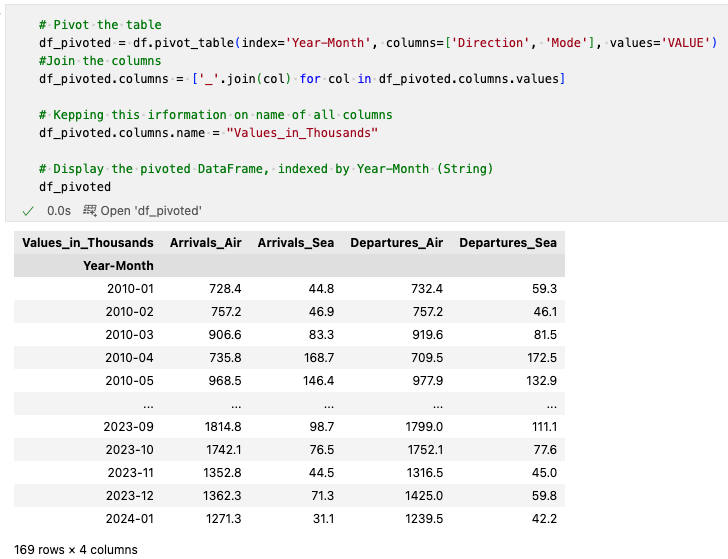


The column “Mode” has rows with the value “All modes of transport” and the “VALUE” column has an apparent sum of the two rows below but with some round differences. This sum is redundant and unnecessary at the moment and easy to have later by calc. Let’s remove all rows with the value, “All modes of transport” from the data frame. For this let’s use the method “query()” from “Pandas”, it works similarly to an “SQL Language”, in this case filtering and returning only rows when column “mode” is different “!=” to value “All modes of transport”.

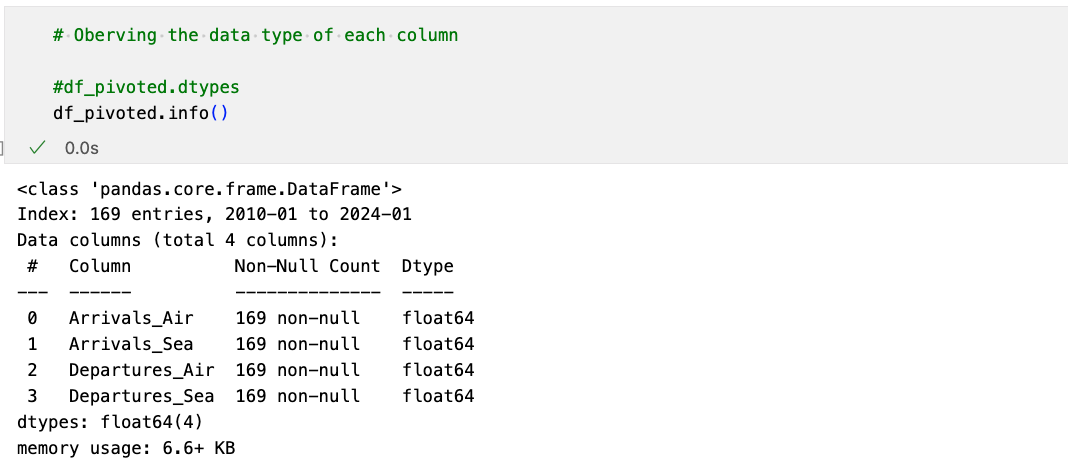


The column “Year-Month” can start to be the index of the data set, making every month one row and all these values on the “Direction” and “Mode” columns can be joined e.g., “Arrivals” + “Air”, creating news columns, and your respective values on the column “VALUE” below of the column created.

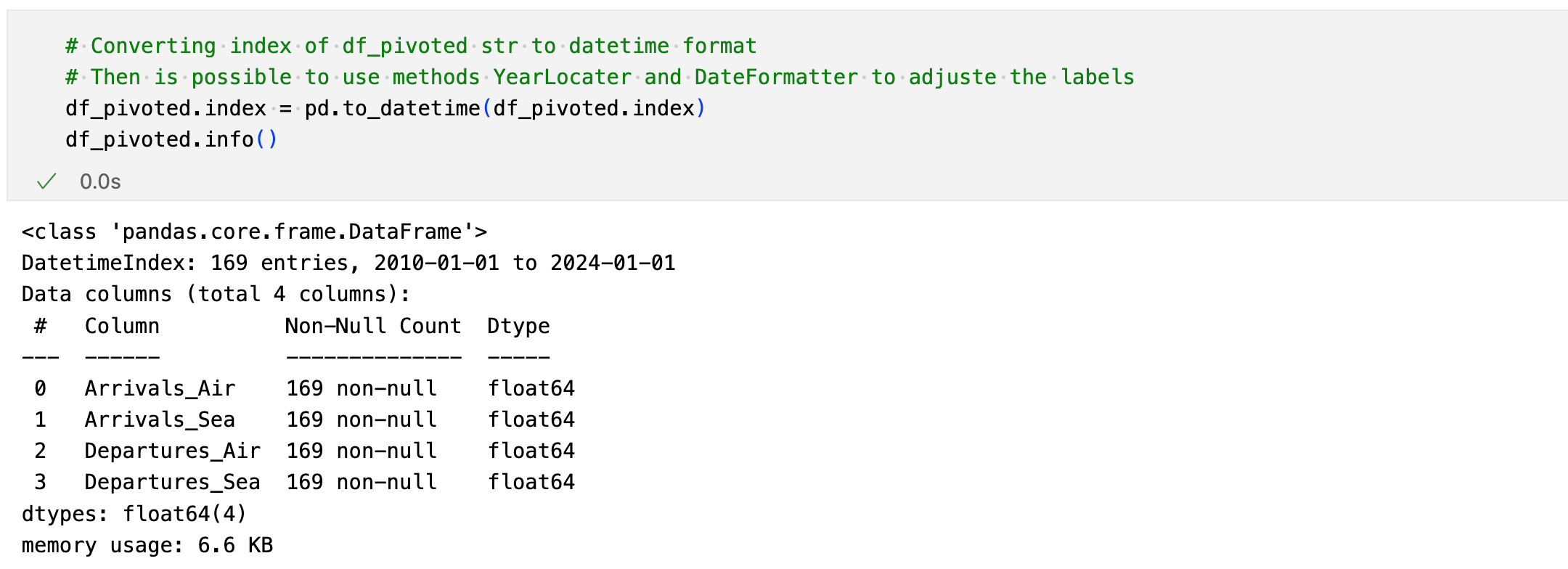
For this let’s utilize the method “pivot\_table()”, passing these specific parameters, and creating then a new data frame called “df\_pivoted”. They are labelling the columns with the “UNIT” value removed before “Thousands” so, as not to forget or lose this parameter.



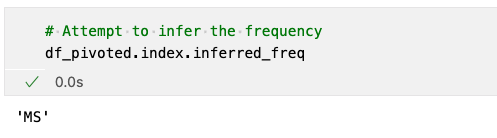
Let’s use the method “info()” on the new data frame “df\_pivoted” to check some of your parameters.



Index of “df\_pivoted”, is still in string format. Let’s change to the “DateTime” format then it’s possible to utilize methods like, “YearLocator” and “DateFormatter” to simplify after procedures like adjustments on labels.



With the index in “DateTime” format now is possible to check and certify your frequency. For this let’s utilize the method “inferred\_freq”, from “Pandas”.

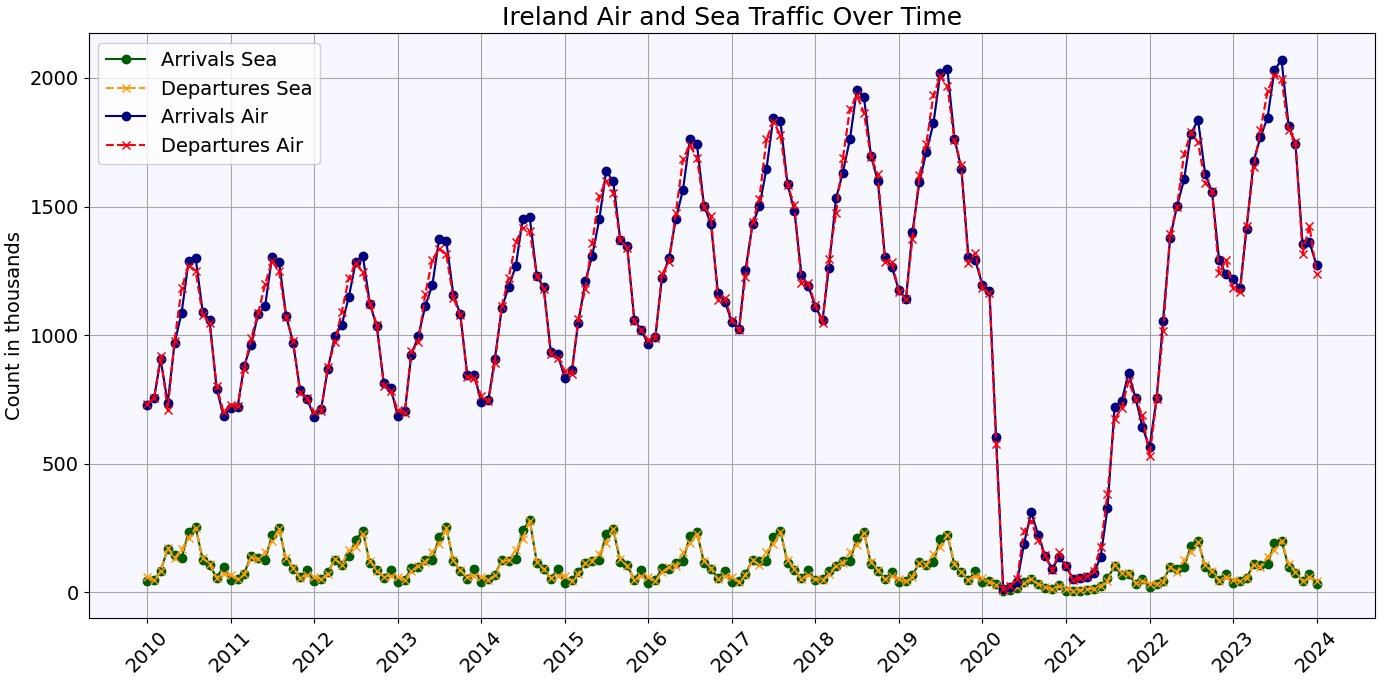
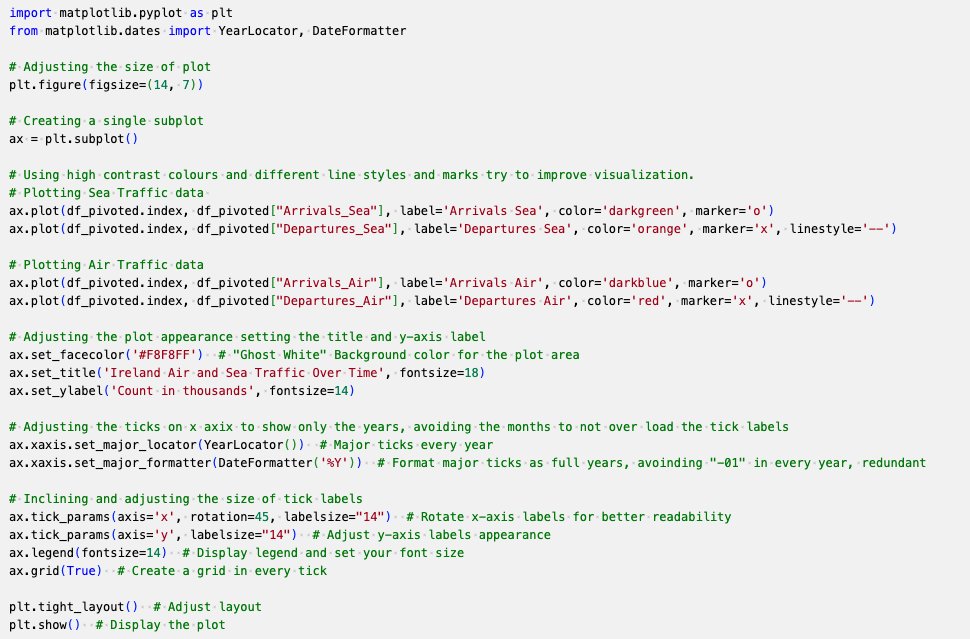


The return “MS” from “month sequence”, confirms to us that all rows have a perfect chronological sequence of months.

Now we have a new data frame called “df\_pivoted” with the same data, but indexed in date and time format and organized in numeric columns with your referent titles, which makes it easier to plot graphs, carry out subsequent statistics calculations analyses and or apply Machine Learning models.

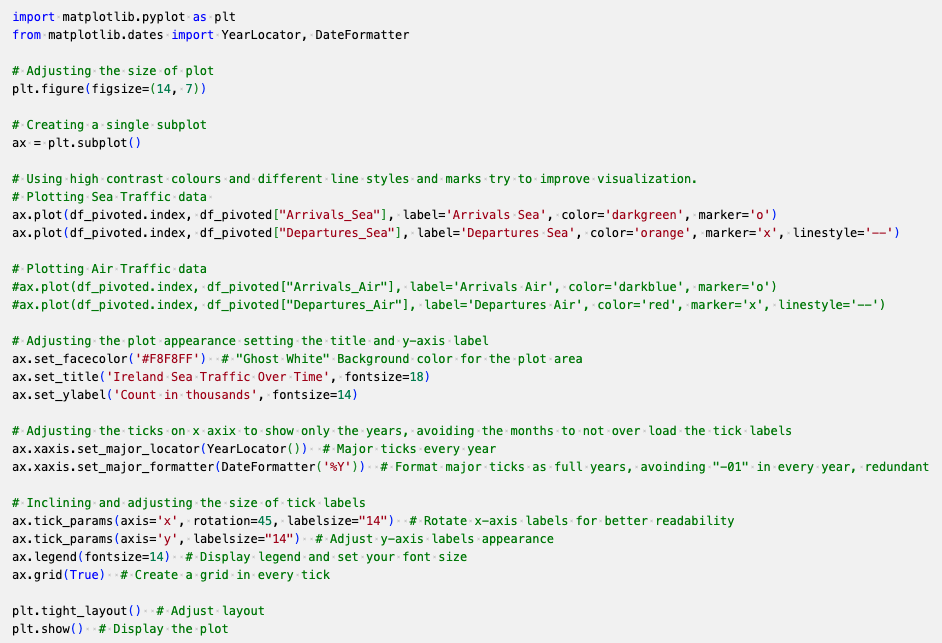
Let’s plot this data set and see. For this let’s import the library “matplotlib.pyplot” and set your alias name as “plt”, and from “matplotlib.dates” let’s import “YearLocator” and “DateFormatter”, these methods are going to help us adjust the labels on graph plotted. The comments on the code have more details and explanations about the “pyplot” parameters.

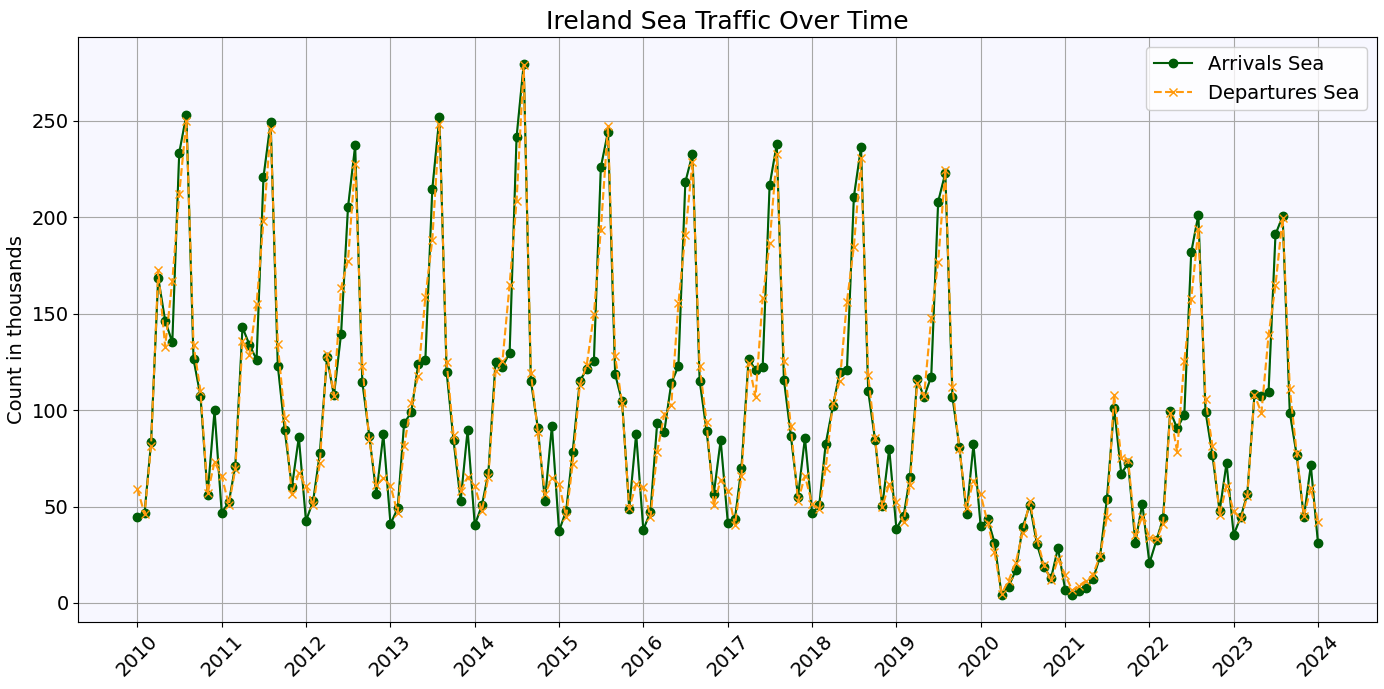
## Plotting the Dataset

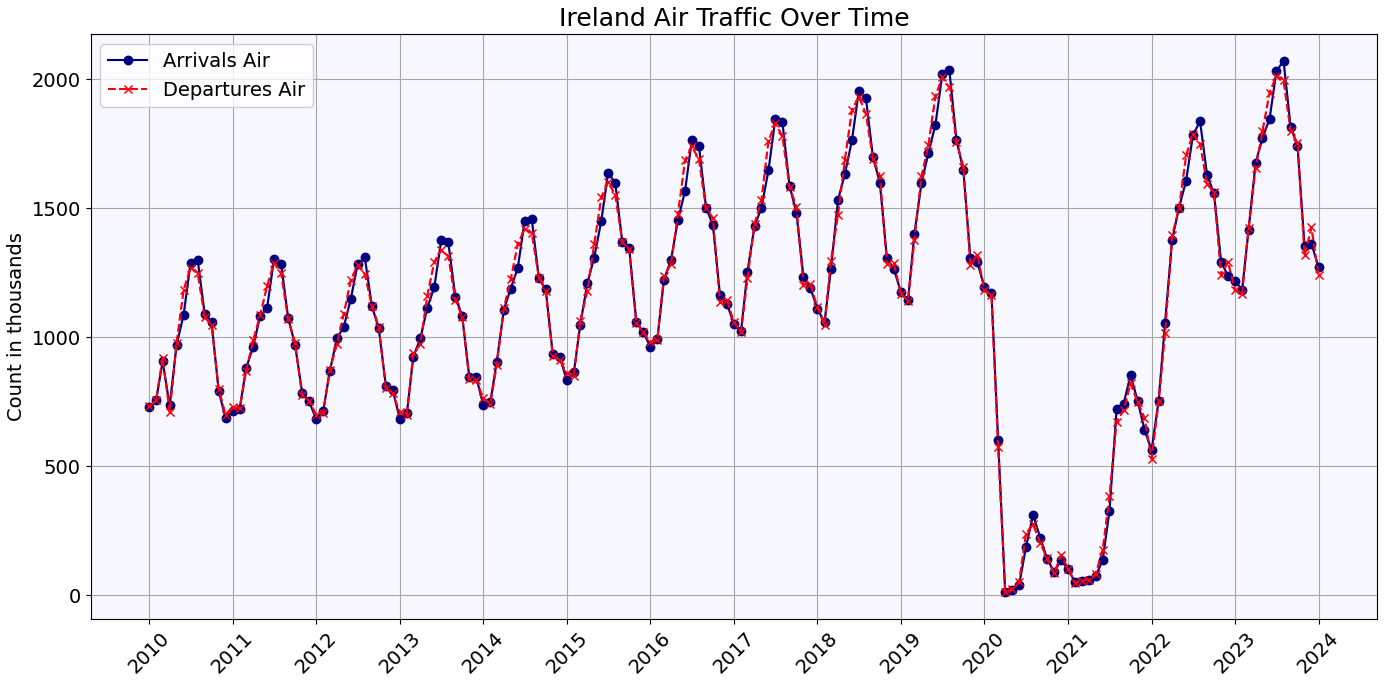


The full data is plotted on the graph above and the first thing possible to view is the big trend caused by COVID-19 in the air sector after the year 2020 but keeping the plots on the same scale makes it harder to view what happened in the sea sector.

Let’s plot “arrivals\_sea” and “departures\_sea” together but in on separate graphic fitting your respective scales to check what can see. For this let’s just put a comment marker “#” on the lines about “Plotting Air Traffic Data”, and adjust the title. And do the same to “Arrivals\_air” and “Departures\_air” to compare.







Now adjusting the graphs to your respective scale, and putting both together on the same page, it is clear that maritime traffic has the same trend after the year 2020.

It is also possible to see that maritime traffic is decreasing in this sample of the year, variant reaching a peak of 250 thousand in the middle of 2010 to 200 thousand in mid-2023. Otherwise, is possible to view the Air traffic as increasing from the approximated 1250 thousand to over 2000 thousand.

Also, it is possible to see dual modes of transport Sea and Air have the same seasonality with a peak in the middle of the year and arrivals and departures have very small differences.

# STATISTICS

# 

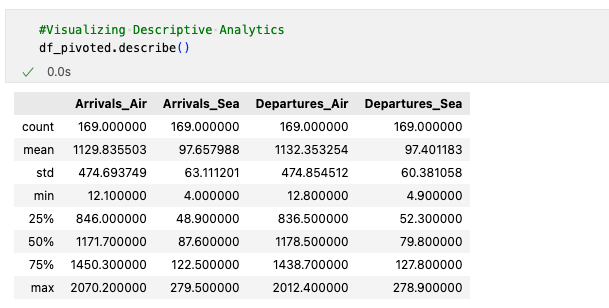
## Descriptive Statistics

Descriptive statistics provide a numerical summary of the characteristics of a set of collected data, in a population or in a sample “subset thereof”. These calculations are designed to convey the central tendency, dispersion and shape of the distribution of the data set among other characteristics.

## Common descriptive statistics in general include:

* Count / Size
* Minimum
* Maximum
* Sum
* Mean
* Median
* Average
* Midrange
* Standard deviation
* Quartiles

With the data frame now prepared and organized in numerical columns “df\_pivoted”, is easy to get the most common statistics descriptions from the data frame. For this is just needed to execute some methods already implemented on the libraries, like “describe()” from “Pandas”



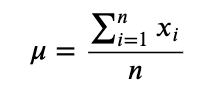
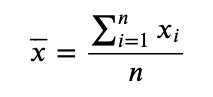
How is possible to see after preparing the dataset is easy to obtain some statistics numbers, but let’s perform some of these calculations, one by one to try to understand the behaviours of this data, like “central tendency”, “skewness” and “kurtosis”.

Let’s start obtaining the mean, median and mode of each column.

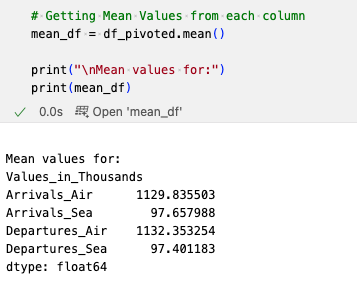
The mean, also commonly referred to as the average, is a measure of the central tendency of a dataset. It is calculated by adding all the data values together to find the sum and then dividing this total by the number of data points in the set, which is the size or count.

The mean formula:

For a Population: For a Sample:

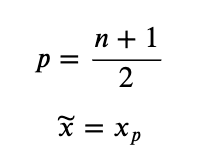
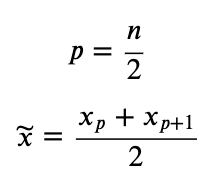
To collect this let’s use the method “mean()” from “Pandas”, creating a new data frame “mean\_df”, only with the mean values of each column.



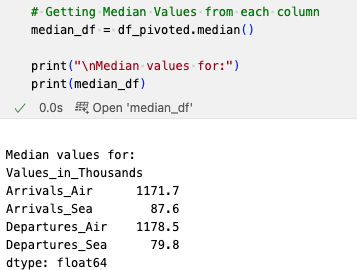
Median, when arranging a dataset from smallest to largest, the median is the middle value that divides the dataset into two halves. For datasets with an odd number of entries, the median is the central value. However, if the dataset has an even number of entries, the median is found by calculating the average of the two middle values.

The median formula:

If n is an odd number: If n is even the median is the average of the values at positions p and p + 1 where

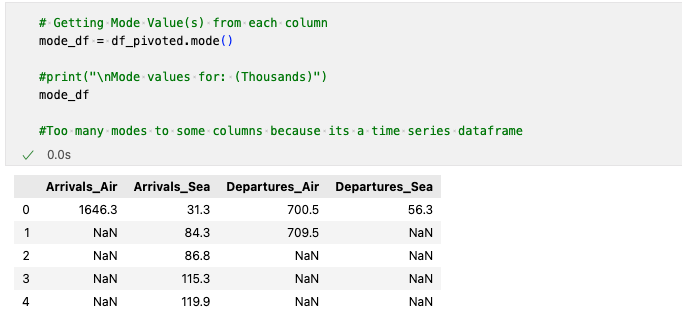
 

Obtaining these values for each column utilizing method “median()” from “Pandas”, creating a new data frame “median\_df”, only with the median values.



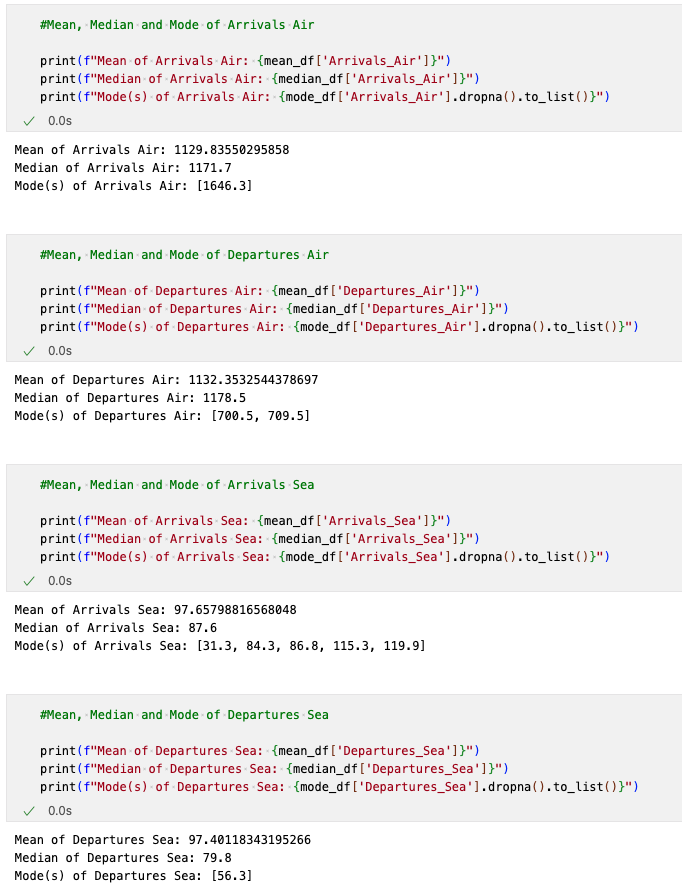
The mode in a dataset represents the value(s) that occur most frequently. A dataset can have one mode and or more than one mode calling it multi-modal. If multiple values occur with the same highest frequency, or no mode if all values are unique. Also can be called bimodal if have two modes.

Verifying these values from every column on the data set, utilizing the method “mode()” from “Pandas”.



Observe the column “Arrivals\_Sea” has a list of mode(s) and the column “Departures\_Air” is possible to see two values.

Let’s print these three measurements together and to better visualize the mode(s) values, let’s print him in a list format for the four columns on the data set.



Analysing these three attributes it is possible to analyse the dispersion of the data, “skewness”.

Skewness measures the asymmetry of a data distribution compared to the normal distribution. If the distribution has a longer tail on the left side, it is considered left-skewed or negatively skewed.

If it has a longer tail on the right side, it's right-skewed or positively skewed. This characteristic helps in understanding the direction and extent of distribution deviation from the symmetrical bell curve.

Also, there exists a common table to try to interpret these values based on the order of who is bigger than who.

| **Relationship** | **Distribution Skewness** | **Interpretation** |
| --- | --- | --- |
| **Mean > Median > Mode** | Positively Skewed (Right-Skewed) | The distribution's tail is longer towards the higher values, suggesting a significant number of observations are larger than the central values. A few high-value outliers are pulling the mean to the right of the median and mode. |
| **Mean < Median < Mode** | Negatively Skewed (Left-Skewed) | The distribution's tail is longer towards the lower values, indicating a significant number of observations are smaller than the central values. Some low-value outliers are dragging the mean to the left of the median and mode. |
| **Mean = Median = Mode** | Symmetrical (No Skew) | The distribution is balanced on both sides, often indicative of a normal distribution. The values are evenly distributed around the center, with no skewness. |
| **Mean = Median > Mode** | Moderately Positively Skewed | Similar to positively skewed, but the closeness of the mean and median suggests that while the distribution has a longer tail on the right, the skewness is not as pronounced. |
| **Mean = Median < Mode** | Moderately Negatively Skewed | Similar to negatively skewed, but the closeness of the mean and median indicates that while the distribution has a longer tail on the left, the skewness is mild. |

There are three basic approaches to generating forecasts: regression-based methods,

And heuristic smoothing methods. and general time series models.

# Difference between Discrete Data and Continuous Data.

## Definition of discrete data.

Discrete data refers to a form of quantitative information characterized by countable figures and non-fractional values. Typically, discrete data is presented in the form of whole numbers that convey precise quantities. A common way to conceptualize discrete data is to preface it with "the number of," for instance, the number of patrons in a shop. This kind of data generally encapsulates distinct occurrences that are already in the past. In analysing discrete data, you can examine precise numbers, such as the quantity of products sold on a particular date or the duration of time an employee has worked in a given week.

## Definition of continuous data.

Continuous data is a quantitative data category that captures measurements that can be highly precise, extending to numerous decimal places as needed. It represents values that can be measured on a scale and can fall between any two amounts within a range. This data type is prevalent in sectors that demand exactness, such as healthcare, production, and research and development. Continuous data is dynamic, presenting the opportunity for organizations to scrutinize their processes and forecast upcoming patterns. An instance of its application could be a company monitoring the duration required by a team to fulfil assignments, providing insights into productivity and efficiency.

Top of Form

Bottom of Form

# Definition of Descriptive Statistics

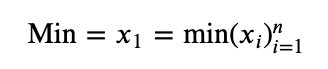
# Some common formulas and calculations used in Descriptive Statistics

## Minimum

Ordering a data set:

x1 ≤ x2 ≤ x3 ≤ ... ≤ xn

The minimum in a data set is the least value​, when the data is arranged in ascending order from the smallest to the largest value

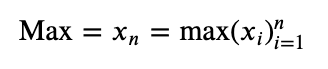


## Maximum

The maximum in a dataset represents the greatest value located at the far right when the data is ordered from the lowest to the highest value.

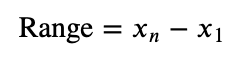
Ordering a data set:

x1 ≤ x2 ≤ x3 ≤ ... ≤ xn



Range

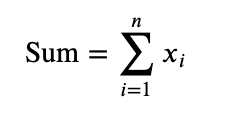
The difference between the minimum and maximum values in a dataset is known as the range. It is calculated by subtracting the minimum value, from the maximum value. The range provides a measure of the spread or dispersion of the data points within the set.



## Sum

The total of all data values in a dataset is known as the sum of the aggregate. It is calculated by adding together all the individual values in the dataset. This total is often symbolized by the Greek letter Sigma (Σ) followed by the expression for the data points, indicating the summation of the series of values.

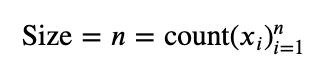
x1 + x2 + x3 + ... + xn



## Size / Count

The size or count of a dataset refers to the number of individual data points it contains. This is a measure of the dataset's magnitude in terms of its elements and is often denoted as *n* in statistical notation.

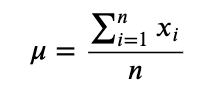
Top of Form

Bottom of Form

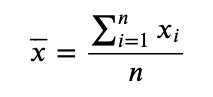
## Mean

The mean, also commonly referred to as the average, is a measure of the central tendency of a dataset. It is calculated by adding all the data values together to find the sum and then dividing this total by the number of data points in the set, which is the size or count.

For a Population:



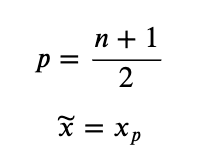
For a Sample:



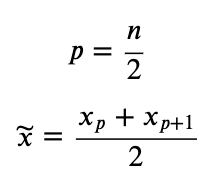
## Median

When arranging a dataset from smallest to largest, the median is the middle value that divides the dataset into two halves. For datasets with an odd number of entries, the median is the central value. However, if the dataset has an even number of entries, the median is found by calculating the average of the two middle values.

If n is an odd number:



If n is even the median is the average of the values at positions p and p + 1 where



## Mode

The mode in a dataset represents the value(s) that occur most frequently. A dataset can have one mode and more than one mode if multiple values occur with the same highest frequency, or no mode if all values are unique.

Top of Form

Bottom of Form

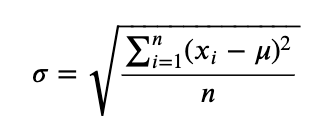
Top of Form

Bottom of Form

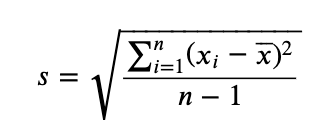
## Standard Deviation

Standard deviation quantifies how spread out the data points in a set are from their average value. This measure is calculated by taking the square root of the average of the squared deviations of each data point from the mean. A larger standard deviation indicates that the data points are more widely dispersed from the mean.Top of Form

For a Population



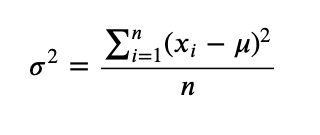
For a Sample



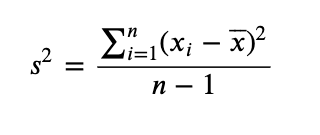
## Variance

Variance measures the spread of data points around the mean within a dataset. It's computed as the average of the squared deviations from the mean. A greater variance indicates that the data points are more dispersed from the mean, highlighting the variability within the dataset.

For a Population



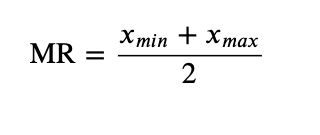
For a Sample



## Midrange

## 

The midrange in statistics is found by averaging the smallest and largest numbers in a dataset. It provides a quick sense of the centre or middle value of the data, especially useful for understanding the data's range.



## Quartiles

Quartiles divide a dataset into four equal parts.

The median, or second quartile (Q2), splits the data into upper and lower halves.

The first quartile (Q1) is the median of the data points below Q2

The third quartile (Q3) is the median of the data points above Q2.

These quartiles help in understanding the distribution of data by highlighting the *spread* and *central tendency*.

\**This is one of several methods for calculating quartiles.*

## Interquartile Range

The interquartile range (IQR) is defined as the distance between the first quartile (Q1) and the third quartile (Q3) in a dataset. It represents the range within which the middle 50% of the data points lie, effectively measuring the spread of the central portion of the dataset and minimizing the impact of *outliers*.

𝐼𝑄𝑅=𝑄3−𝑄1

## Outliers

Potential outliers are those values in a dataset that fall either below the Lower Fence or above the Upper Fence. These fences are determined by specific calculations that take into account the interquartile range (IQR), helping to identify data points that significantly differ from the rest of the dataset.

Upper Fence = 𝑄3+1.5×𝐼𝑄𝑅

Lower Fence = 𝑄1−1.5×𝐼𝑄𝑅

## Sum of Squares

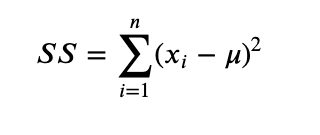
The sum of squares refers to the aggregated total of each data point's deviation from the mean, squared.

This calculation is a fundamental part of various statistical analyses, serving to quantify the variance within a dataset.

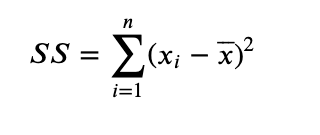
Top of Form

Bottom of Form

For a Population



For a Sample

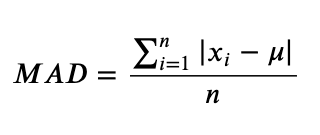


## 

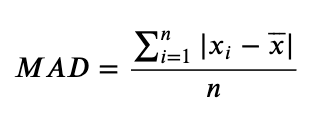
## Mean Absolute Deviation

Mean absolute deviation measures the average distance between each data point and the mean of the dataset. This is calculated by taking the absolute values of the differences between each data point and the dataset's mean, and then dividing by the number of data points. It provides insight into the variability of the dataset.

For a Population

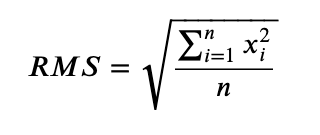


For a Sample



## Root Mean Square

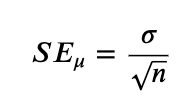
The root mean square (RMS) is a statistical measure that calculates the magnitude of a set of numbers. It is found by taking the square root of the average of the squares of the values in the set. This metric is especially useful in contexts where both positive and negative values in the dataset are treated equally, and it tends to give a higher value than the average due to the squaring of the values.



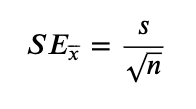
## Standard Error of the Mean

The standard error of the mean (SEM) is derived by dividing the standard deviation of the dataset by the square root of the number of observations (*n*). This metric indicates how much the sample mean is expected to vary from the true population mean.

For a Population



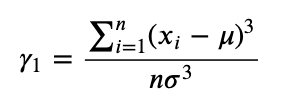
For a Sample



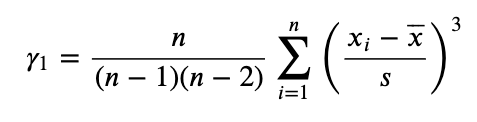
## Skewness

Skewness measures the asymmetry of a data distribution compared to the normal distribution. If the distribution has a longer tail on the left side, it is considered left-skewed or negatively skewed. Conversely, if it has a longer tail on the right side, it's right-skewed or positively skewed. This characteristic helps in understanding the direction and extent of distribution deviation from the symmetrical bell curve.

For a Population



For a Sample



Common nomenclatures about different kinds of Skewness coming from different Relationships between Mean, Median and Mode:

| **Relationship** | **Distribution Skewness** | **Interpretation** |
| --- | --- | --- |
| **Mean > Median > Mode** | Positively Skewed (Right-Skewed) | The distribution's tail is longer towards the higher values, suggesting a significant number of observations are larger than the central values. A few high-value outliers are pulling the mean to the right of the median and mode. |
| **Mean < Median < Mode** | Negatively Skewed (Left-Skewed) | The distribution's tail is longer towards the lower values, indicating a significant number of observations are smaller than the central values. Some low-value outliers are dragging the mean to the left of the median and mode. |
| **Mean = Median = Mode** | Symmetrical (No Skew) | The distribution is balanced on both sides, often indicative of a normal distribution. The values are evenly distributed around the center, with no skewness. |
| **Mean = Median > Mode** | Moderately Positively Skewed | Similar to positively skewed, but the closeness of the mean and median suggests that while the distribution has a longer tail on the right, the skewness is not as pronounced. |
| **Mean = Median < Mode** | Moderately Negatively Skewed | Similar to negatively skewed, but the closeness of the mean and median indicates that while the distribution has a longer tail on the left, the skewness is mild. |

Example of Skewness Interpretation:

| **Revenue Distribution** | **Skewness Interpretation** |
| --- | --- |
| **Mean > Median > Mode** | The revenue data is positively skewed. This suggests that while most companies generate revenue within a similar range (around the mode and median), there are a number of companies with revenue significantly higher than the average, which pulls the mean upwards. |

Top of Form

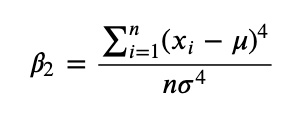
## Kurtosis

Kurtosis measures the "tailedness" of a distribution, indicating how outlier-prone a dataset is. High kurtosis suggests more extreme outliers than a normal distribution, while low kurtosis indicates fewer extreme outliers. This helps assess the extremity and concentration of tail data compared to a normal bell curve.

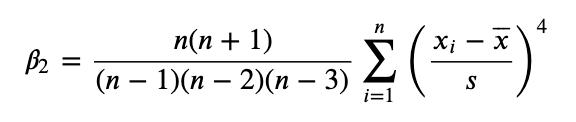
Top of Form

Bottom of Form

For a Population



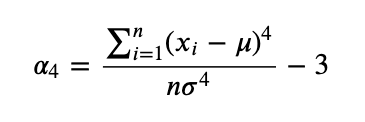
For a Sample



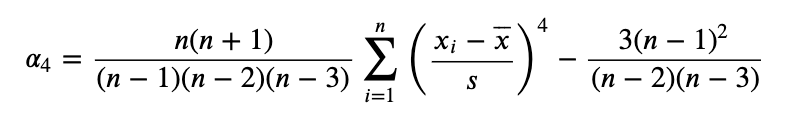
Kurtosis Excess

Excess kurtosis gauges the peak height of a distribution's tails, focusing on the concentration of outliers rather than their extremity. A distribution with high excess kurtosis indicates a significant presence of outlier data, pointing to more frequent extreme deviations from the mean compared to a normal distribution.

For a Population



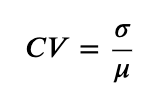
For a Sample #(This is just Kurtosis in MS Excel and Google Sheets)



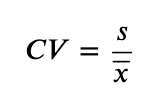
## Coefficient of Variation

The coefficient of variation (CV) measures the relative spread of data points around the mean, expressed as a ratio of the standard deviation to the mean. It's a useful statistic for comparing the degree of variability from one data series to another, even if the means are drastically different. The CV is calculated by dividing the standard deviation by the mean. This measure is particularly helpful in assessing the risk or variability in different contexts, such as finance and scientific research, where understanding relative dispersion is crucial.

For a Population



For a Sample



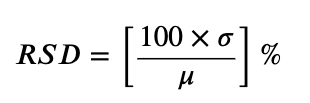
## Relative Standard Deviation

The relative standard deviation (RSD) quantifies the variation in a data set relative to its mean, presented as a percentage. It's computed by multiplying the standard deviation by 100 and then dividing by the mean. This statistic is valuable for comparing the variability of datasets with different units or means, providing a normalized measure of dispersion.

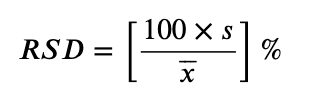
Top of Form

Bottom of Form

For a Population



For a Sample



## Frequency

Frequency measures how often each value appears in a dataset, essential for determining the mode, the value that occurs most frequently. This statistical concept helps in understanding the distribution and concentration of data points.

# Trend and Seasonality: In time series analysis, it's also important to consider trend and seasonality.

# These factors can significantly affect your central tendency measures.

# For example, a steadily increasing trend could make the mean over the entire series less representative of any specific point in time.

# Outliers: Especially with mean calculations, consider the impact of outliers.

# If your time series data includes extreme values, they could skew the mean.

# As can be seen in the described statistics, column Arrivals Air got close to 0

Kernel Density Estimate (KDE) Explained

#KDE is a non-parametric way to estimate the probability density function (PDF) of a random variable.

#It's smooth and not limited to a specific distribution shape (like normal or binomial distributions).

#KDE works by placing a kernel (a smooth, bell-shaped curve) on each data point and then summing all these kernels to produce a smooth estimate of the data's density function.

#The final curve is a weighted sum of those kernels and gives a smoothed estimate of the dataset's density.

#The bandwidth (bw) parameter controls the width of the kernels and thus the smoothness of the density estimate:

#a larger bandwidth leads to a smoother curve,

#while a smaller bandwidth leads to a curve that closely follows the data.

#Handling Multimodal\* Distributions

#For datasets with more than one mode,

#the KDE will show peaks at each mode, depending on the chosen bandwidth.

#A well-chosen bandwidth can reveal the multimodal nature of the data.

#The KDE does not calculate a normal distribution;

#rather, it estimates the data's density based on the existing data points.

#If the data is multimodal,

#the KDE will reflect those modes in its estimate.

Bottom of Form

# Reference list

Tufte, E.R. (1997) *The visual display of quantitative information / Edward R. Tufte*. Cheshire, CT: Graphics Press.

*Descriptive statistics* *from Wolfram MathWorld*. Available at: https://mathworld.wolfram.com/topics/DescriptiveStatistics.html (Accessed: 26 March 2024).

Aloorravi, S. (2024) ‘CHAPTER 4 Exploratory Analysis of Time Series Data’, in *Mastering Time Series Analysis and Forecasting with Python*. London, United Kingdom: Copyright © 2024 Orange Education Pvt Ltd, AVATM, pp. 142–142.

*Seaborn seaborn.heatmap - seaborn 0.13.2 documentation*. Available at: https://seaborn.pydata.org/generated/seaborn.heatmap.html (Accessed: 07 April 2024).

*Pandas* *pandas.DataFrame.corr - pandas 2.2.1 documentation*. Available at: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.corr.html (Accessed: 07 April 2024).

*Sklearn.preprocessing.MinMaxScaler*. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html (Accessed: 07 April 2024).

*Matplotlib.pyplot#* (no date) *matplotlib.pyplot - Matplotlib 3.5.3 documentation*. Available at: https://matplotlib.org/3.5.3/api/\_as\_gen/matplotlib.pyplot.html (Accessed: 07 April 2024).