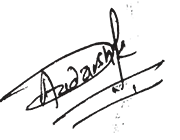
**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | *Programming for DA*  *Statistics for Data Analytics*  *Machine Learning for Data Analysis*  *Data Preparation & Visualisation* |
| **Assessment Title:** | MSC\_DA\_CA2 |
| **Lecturer Name:** | Sam Weiss/ David Gonzalez  Taufique Ahmed  Muhammad Iqbal  David McQuaid |
| **Student Full Name:** | Adarsh Kushwaha  **https://github.com/AadarshKushwaha/CA2-Integrated-Assessment** |
| **Student Number:** | 2023398 |
| **Assessment Due Date:** | 07/01/2024 |
| **Date of Submission:** | 05/01/2024 |

****

**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. |

Table of Contents

[Public Transport Trends in Ireland 4](#_Toc155535571)

[Licensing or permissions associated 4](#_Toc155535572)

[Introduction 4](#_Toc155535573)

[ Library Used 4](#_Toc155535574)

[ Ignore the warning message. 5](#_Toc155535575)

[Project Management Framework 5](#_Toc155535576)

[Data Preparation and Visualization 5](#_Toc155535577)

[ Acquisition of Dataset 5](#_Toc155535578)

[o Traffic Volume October 2023 (Council, 2023-07-01 - 2023-12-31) 5](#_Toc155535579)

[o Passenger Journey by Luas 2018-2022 (Ireland, 2023) 5](#_Toc155535580)

[o Passenger Journey by Public Transport (Bus, Dublin Metro, Rail) 2019-2023 (Dorgan, 2023-10-18) 5](#_Toc155535581)

[o Passenger handled by Main Airports 2019-2023 (Dorgan, 2023-09-19) 5](#_Toc155535582)

[o National Route Length 2015-2022 (Dorgan, 2022-12-02) 6](#_Toc155535583)

[o Passenger travel by Rails in Paris, France 2013-2021 (CML/DRC/ATTR, 17 may 2021) 6](#_Toc155535584)

[o Passenger journey by Rail in thousands 1982-2022 (Dorgan, 2022-11-30) 6](#_Toc155535585)

[o Public review on Public Transport of Ireland (u/Longjumping-Stretch5, 2023) 6](#_Toc155535586)

[o Public review on Public Transport of Netherlands (u/R\_Blitzie, 2023) 6](#_Toc155535587)

[ Positive and/or negative aspects of your research and Acquisition 6](#_Toc155535588)

[ EDA (Exploratory Data Analysis) 6](#_Toc155535589)

[o EDA for traffic\_volumes\_oct\_2023 6](#_Toc155535590)

[o EDA And Insight gain from passenger travel by Luas 7](#_Toc155535591)

[o EDA And Insight gain from passenger journey by Public Transport 8](#_Toc155535592)

[o EDA And Insight gain from passenger handled by main airports of Ireland. 8](#_Toc155535593)

[o EDA And Insight gain from National Route Length of Ireland 9](#_Toc155535594)

[o EDA of Paris Rail dataset 10](#_Toc155535595)

[o EDA of Irish Rail dataset 11](#_Toc155535596)

[Statistical Analysis 11](#_Toc155535597)

[ Paris Rail Dataset and Irish Rail Dataset 11](#_Toc155535598)

[o Descriptive statistics 11](#_Toc155535599)

[o Inferential statistics 13](#_Toc155535600)

[ Hypothesis Testing 14](#_Toc155535601)

[o One Way ANOVA 14](#_Toc155535602)

[o Two Way ANOVA 15](#_Toc155535603)

[o Significance level: 15](#_Toc155535604)

[o p-Value 15](#_Toc155535605)

[o t-Test 15](#_Toc155535606)

[o Wilcoxon signed rank test 15](#_Toc155535607)

[ Insights from Paris Rail Dataset 15](#_Toc155535608)

[ Insights from Irish Rail Dataset 16](#_Toc155535609)

[Comparison Between Passenger Journey in Ireland and Paris 17](#_Toc155535610)

[Machine Learning 18](#_Toc155535611)

[ Data Pre processing 18](#_Toc155535612)

[o Encoding Data for Machine Learning Models 18](#_Toc155535613)

[o One Hot Encoder: 18](#_Toc155535614)

[o Label Encoder: 18](#_Toc155535615)

[ Selection of Machine Learning 19](#_Toc155535616)

[o Supervised Learning 19](#_Toc155535617)

[o Feature Scaling: 19](#_Toc155535618)

[ Selection of Random state, and data for Training and Testing 19](#_Toc155535619)

[ Simple Linear Regression 19](#_Toc155535620)

[ Multiple Linear Regression 20](#_Toc155535621)

[ Performance Evaluation of Simple and Multiple Linear Regression 20](#_Toc155535622)

[ Ridge Regression 20](#_Toc155535623)

[ Lasso Regression 20](#_Toc155535624)

[ Performance Evaluation of Ridge Regression and Lasso Regression 21](#_Toc155535625)

[o R^2 score 21](#_Toc155535626)

[o Root Mean Square Error (RMSE): 21](#_Toc155535627)

[o Mean Absolute Percentage Error (MAPE): 21](#_Toc155535628)

[Sentiment Analysis 21](#_Toc155535629)

[ Sentiment Analysis on Public Transport 21](#_Toc155535630)

[o Sentiment Question for ‘Dublin’ is. 22](#_Toc155535631)

[o Sentiment Question for ‘Netherlands’ is. 22](#_Toc155535632)

[ Comparison between Sentiment Analysis on Public Transport of Ireland Vs Netherlands 22](#_Toc155535633)

[Python 23](#_Toc155535634)

[Dashboard 23](#_Toc155535635)

[Reference 24](#_Toc155535636)

# Public Transport Trends in Ireland

#### Trends of Public Transport in Ireland is produced as a part of Integrated assignment for MSc in Data Analytics to gain some insights on use of public vehicles like Bus, Luas, Irish rail, Dublin Metro etc. Also, it has comparison between use of Rails by public in Ireland with foreign country like France, Paris city is one of the crowed cities of France. Moreover, in this report a sentiment analysis is covered with Neutral, Positive, Negative sentiment of public regarding public transport provided in Ireland and Netherlands.

# Licensing or permissions associated

#### In academic or research contexts, ethical approval might be necessary, especially for studies involving human subjects. And it is free to use for study, experiment, share and available for republish the content in any format or medium for any purpose, even for profit under certain terms and conditions like *appropriate credit,* as per mentioned on the website from where data is collected. Adhere to the terms specified in the license (International, 2023). The data is produced annually by the Ireland’s Open Data Portal (Ireland, 2023); a constituent of the Irish Government. The Data utilized in this report has been primarily collected from the external published source listed on Ireland’s Open Data Portal (Ireland, 2023); As such the department doesn’t hold any responsibility for the accuracy or robustness of the underlying data. Numerical Values may be rounded. References to data are highlighted in this document and cited the original source from where data are collected.

# Introduction

#### In Ireland, the transportation sector is experiencing continuous growth, as evidenced by key indicators such as increased public transportation usage, rising airport and city arrivals, and a surge in commuter numbers. After a period of decline and stagnation in previous years, the industry has robustly rebounded, underscoring its vital role in a nation's development. Efficient transportation not only fosters economic prosperity but also facilitates the movement of goods, tourists, and commuters within and beyond city and country borders. My research focuses on assessing the everyday utilization of public transport in Ireland and the diverse modes of transportation chosen by individuals for their travel needs. Additionally, a comparative analysis between Ireland's rail transport system and that of Paris, a bustling city in France, was conducted. Sentiment analysis was employed to gauge public opinions regarding the use of public transportation in Ireland and the Netherlands. To predict future travel trends and identify the busiest railway stations, I employed machine learning models. Utilizing Python programming language, Jupiter Notebook, and the Anaconda environment, along with popular libraries such as Pandas, NumPy, Matplotlib, Seaborn, Sklearn, and SciPy, I conducted in-depth analyses and visualized insights from various datasets.

## Library Used

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

## Ignore the warning message.

#### In this section, a Python function named ignore\_warning() is defined. It utilizes the warnings.filterwarnings() function, with action='ignore', to filter out warning messages, suppressing their display during execution.

# Project Management Framework

#### The adoption of the KDD (Knowledge Discovery in Databases) framework for project management is apt for this endeavor. The rationale behind selecting KDD aligns with its comprehensive approach, encompassing data preparation, feature engineering, and diverse data analyses, including machine learning and data mining. This framework proves suitable for projects aiming to extract valuable insights from extensive datasets. For instance, in the healthcare sector, utilizing KDD, a hospital can analyze electronic health records to discern patterns in patient data, thereby enhancing care delivery. KDD's structured stages, such as data cleaning, integration, selection, transformation, mining, pattern evaluation, and knowledge representation, make it well-suited for systematic and effective management of the project. KDD is an *iterative process* where evaluation measures can be enhanced (Accredian, 2022)

# Data Preparation and Visualization

## Acquisition of Dataset

### Traffic Volume October 2023 (Council, 2023-07-01 - 2023-12-31)

#### This dataset has been used to analysis number of vehicle running on road in month of October 2023. This dataset has size of 381.1 MB.

### Passenger Journey by Luas 2018-2022 (Ireland, 2023)

#### The 14KB dataset analyzes Luas passenger travel from 2018 to 2022, detailing Green and Red Line ridership monthly, providing insights into yearly travel patterns.

### Passenger Journey by Public Transport (Bus, Dublin Metro, Rail) 2019-2023 (Dorgan, 2023-10-18)

#### The 107KB dataset examines passenger travel by Bus, Dublin metro, and Rail, excluding Luas, from 2019 to 2023.

### Passenger handled by Main Airports 2019-2023 (Dorgan, 2023-09-19)

#### This dataset has been used to analysis number of passenger handle by main airports of Ireland like Dublin, Cork, Shannon, Knock, Kerry in given year 2019-2023. This dataset has size of 28KB.

### National Route Length 2015-2022 (Dorgan, 2022-12-02)

#### The 103KB dataset assesses the growth of national route length in kilometers from 2019 to 2023, including local authority involvement.

### Passenger travel by Rails in Paris, France 2013-2021 (CML/DRC/ATTR, 17 may 2021)

#### The 172KB dataset analyzes passenger travel in Paris stations from 2013 to 2021, detailing rail types (Metro, RER) in 9 files per year.

### Passenger journey by Rail in thousands 1982-2022 (Dorgan, 2022-11-30)

#### The 22KB dataset analyzes Irish Rail passenger travel from 1982 to 2022, covering Mainline, Dublin suburban, DART, and international journeys.

### Public review on Public Transport of Ireland (u/Longjumping-Stretch5, 2023)

#### The 73KB dataset analyzes public sentiment on 'Dublin is worst capital in Europe for public transport ticket pricing' from Reddit's r/Ireland subreddit.

### Public review on Public Transport of Netherlands (u/R\_Blitzie, 2023)

#### The 110KB dataset evaluates public sentiment on free public transport in the Netherlands, sourced from Reddit comments within the r/Netherlands subreddit.

## Positive and/or negative aspects of your research and Acquisition

#### Research heavily relies on data quality, and publicly accessible datasets often undergo rigorous maintenance, validation, and curation. Academic databases, subject to peer review, maintain higher standards. Ethical considerations in data collection underscore the importance of open sources with explicit usage guidelines, ensuring cleanliness and ethical approval for research, experiments, analysis, and study.

#### However, drawbacks exist. The certainty of data robustness and correctness is uncertain. Publicly available data may carry biases, and crucial information might be missing. For instance, in a Traffic volume dataset with millions of observations, only three regions (North city, South city, and Center city of Dublin) provide clear analysis, while others like IRL, IRL3 lack precise information. These challenges highlight negative aspects of research and data acquisition from open-source platforms, urging caution and thorough scrutiny in their application.

## EDA (Exploratory Data Analysis)

#### Exploring a dataset is a vital data analysis step, encompassing visual and statistical examination to comprehend characteristics, detect anomalies, and glean insights. Employing functions like shape, head(), tail(), describe(), and info() provides a comprehensive dataset description.

### EDA for traffic\_volumes\_oct\_2023

#### The dataset comprises 1,06,20,233 observations with 9 features, including End\_Time, Region, Site, Detector, Sum\_Volume, and Avg\_Volume. Notably, Weighted\_Avg, Weighted\_Var, and Weighted\_Std\_Dev are null and subsequently removed. The dataset is devoid of any null values except for these three features. Column names are refined for clarity. **6415051** duplicated rows are excluded to prevent potential skewing in the analysis. Mean and Median analysis of sum\_of\_traffic\_volume reveals average traffic distribution ranging between **10,00,000 to 1,50,00,00,** sometimes peaking at **20-25 Hundred Thousand**. Scatterplot analysis confirms a strong correlation between sum of traffic volume and sum of traffic volume per 5 minutes in Dublin's North, South, and Center City regions.

##### Heatmap

#### A heatmap is utilized to visualize traffic volume distribution by weekdays and hours, aiding in identifying peak times for potential vehicle increase. The Red Zone indicates maximum road activity, while purple signifies minimal traffic. Busiest hours occur from **17-19 in the evening** and **9-10 in the morning, Monday to Saturday**. Further analysis reveals North City experiences the highest traffic, reaching **242.036K** on **October 5 at 18:00**, while South City and Center City peak at **150K and 131K**, respectively, during the month.

A screen shot of a heatmap

Description automatically generated

Figure 1: Traffic volume based on Days and Hours in Day.

### EDA And Insight gain from passenger travel by Luas

#### The dataset, with 195 observations and 8 features, focuses on relevant aspects: line\_type (Red, Green, All Luas line), years, months, and passengers. It's clean, lacking missing or duplicated values. Despite outliers reaching **4 million**, the average passenger distribution ranges from **100,000 to 200,000**, with significant passenger travel by red and green lines. COVID-19's impact is evident, with the lowest Red line passengers (**139.27K**) and Green line passengers (**89.35K**) in April 2020 during the peak of the pandemic (Ireland, 2021). (Ireland., 2021).

A graph of different colored lines

Description automatically generated

Figure 2: Passenger Travel by Luas

### EDA And Insight gain from passenger journey by Public Transport

#### The dataset comprises 1024 observations and 8 features, with only 4 relevant for analysis: mode\_of\_transport (Dublin Metro Bus, Bus excluding Dublin Metro, Rail, and All transport excluding Luas), passenger count, year, and week number. No missing or duplicate values exist, and the data distribution is right-skewed, ranging from **100K to 400K**. The bar graph illustrates that Dublin Metro Bus had an average of **3.12M** passengers in 2023, with Buses Excluding Dublin Metro and Rail showing comparable numbers, except in 2019. Overall, excluding Luas, **4.81M** people used public transport in 2023.

A graph of different colored bars

Description automatically generated

### EDA And Insight gain from passenger handled by main airports of Ireland.

#### The dataset, containing 336 observations and featuring year, month, passengers, and airport, provides insights into airport traffic. With a right-skewed distribution, it ranges **from 0 to 4 million** passengers. Dublin Airport emerges as the busiest, peaking at **3.4 million** passengers in 2019 and 2023, and experiencing lows of **42K** in April 2020 and **95K** in March 2021. July and August consistently stand out as the busiest months. Following Dublin, Cork ranks as the second busiest, while Kerry Airport maintains the lowest passenger numbers, hitting a maximum of **45K** in 2023.

#### A graph of a number of people Description automatically generated with medium confidence

Figure 3: Dublin and Cork Airport of Ireland

#### A graph of different colored squares Description automatically generated

Figure 4: Passenger handled by Main Airports of Ireland

### EDA And Insight gain from National Route Length of Ireland

#### The dataset contains 1080 rows and 4 columns, featuring data from 2015 to 2022, 26 local authorities, and 4 carriage types. The dataset had two empty rows, removed using the isna() method. It exhibits a right-skewed distribution. Clare and Roscommon have the highest National route lengths at **3.73KM and 3.95KM**. Mayo records 0KM, while Donegal and Kerry **have 48KM and 88KM**. Ireland predominantly features Single Carriageways, with Roscommon having 1**800KM**. Kildare boasts the longest Motorway (**864KM**), Dublin (**656KM**), and Cork (**504KM**) for Dual Carriageways.

A graph of different colored lines

Description automatically generated with medium confidence

Figure 5: Type of Carriageway in Local Authority

From year 2015 to 2022 the length of single carriageway has been seen slightly decreasing from **2206 KM to 2139KM,** Similarly, **Motorway** has been increased from **918KM to 996KM.** Alsoslightly increments is seen in **Dual carriageway** from **305KM in 2020 to 332KM in 2022.** Increase in traffic volume will directly affect the length of National Route as well and Increments is seen in both Motorway and Dual Carriageway.

A graph with different colored lines

Description automatically generated

Figure 6: National Route Length over Years

### EDA of Paris Rail dataset

The Paris Rail Dataset comprises nine datasets spanning 2013 to 2021. By merging them with pd.concat(), a new dataframe is formed with 3325 observations and 5 features. These features include type\_of\_journey (Metro or RER), station count, city count, year, and passenger records. The dataset is devoid of missing values or duplicates but contains outliers, with Metro and RER journeys ranging from 0 to 5 billion passengers. The data exhibits a right-skewed distribution and is tightly clustered around the first 1 million passengers, as depicted in the distribution graph.

### EDA of Irish Rail dataset

#### The Irish Rail Dataset comprises 120 observations and 3 features. The 'Year' ranges from 1981 to 2022, and the 'Passenger' column denotes the number of travelers. 'Type\_of\_journey' classifies journeys into five categories. Null values were eliminated using the isna() method, and duplicates were removed with isduplicated(). Outliers in the 'DART' category were addressed, resulting in a normally distributed dataset after their removal.

# Statistical Analysis

#### Statistical Analysis is the process of gathering and examining data to identify patterns and trends. It serves to eliminate bias in data evaluation through different methods like extracting insights from research findings, constructing statistical models etc. (Simplilearn, n.d.).

## Paris Rail Dataset and Irish Rail Dataset

### Descriptive statistics

Descriptive statistics entails the use of numerical and graphical methods to portray and illustrate data effectively. Its application involves describing provided sample sets or populations, contributing significantly to our comprehension of various data aspects.

1. **Central tendency of data**: - This represents the central point within the data distribution, elucidating the data's location and focusing on its concentration. The three most utilized metrics to determine the "center" of the data are:

* **Mean of Passenger**

#### The mean, or average, is derived by summing all data values and dividing by the number of observations. In Paris (2013-2021), average Metro and RER passenger travel is **44,17,983.34**, while Irish Rail averages **16,015.28** (1981-2022).

* **Median of Passenger**

#### The median, representing the 50th percentile, resides precisely at the midpoint of ordered data. Unaffected by outliers, it effectively pinpoints the center. In Paris (2013-2021), the median passenger travel for Metro and RER is **3043606**. For Irish rail, **it's 13027.0**.

* **Mode of Passenger**

#### The mode represents the most frequently occurring data point in a dataset. If a specific element has the highest frequency, it becomes the mode. In cases without repeated numbers, the dataset has no mode. The mode for passenger travel with both Metro and RER in Paris from **2013 to 2021 is 394,292**, while the highest occurrence in Irish Rail is **850 passengers**.

**Dispersion of Data**: - Dispersion, often referred to as the "spread of the data," quantifies the extent to which data values are distributed. In many datasets, values tend to be closely clustered around the mean. Conversely, in some datasets, values exhibit a broader distribution away from the mean. The measurement of these variations in data spread can be accomplished by

* **Inter Quartile Range (IQR)**

Quartiles are distinctive percentiles. The 1st Quartile (Q1) aligns with the 25th percentile, the 2nd Quartile (Q2) corresponds to the 50th percentile, and the 3rd Quartile (Q3) mirrors the 75th percentile.

From the above data it shows that both type of journey has outlier which leads to the skewness in data. Most of the data lies above the Upper fence in both cases.

* **Range**

The range is determined by subtracting the smallest value from the largest value in the dataset, expressed as Max – Min = Range. And range of this data is **51240385** For Paris Rail**. And 49612** For Irish Rails**.**

* **Standard Deviation**

The primary indicator of data spread is often the standard deviation, which measures the extent of data deviation from the mean. The formula for standard deviation differs for population and sample. While both formulas share similarities, they are not identical. The symbol for sample standard deviation is "s" (lowercase), while the symbol for population standard deviation is represented by the Greek letter "σ" (sigma, lowercase).

Paris Rail

**Mean: 4417983.349172932** and **Standard Deviation: 5600198.934991727** - It indicates that there is a wide range of passenger travel, and it can be varied significantly from average. Here it shows that people highly travel in Metro as compared to RER, contributing to the overall dataset.

Irish Rail

**Mean: 16015.28** and **Standard Deviation: 11747.91** - It indicates that there is a tight range of passenger travel, and it cannot be varied significantly from average. Here it shows that few people travel in different type of rail, contributing to the overall dataset.

1. **Shape of the Data**: - The configuration characterizes the nature of the graph. Understanding the shape of the data is crucial as decisions regarding the probability of the data are contingent on its form. There are two methods to quantify the shape of the data.

A graph of a passenger

Description automatically generated

Figure 7: Shape of Passenger of Paris Rail

It shows data is Paris Rails has tightly distributed to the right and loose distributed to the left. Hence, it is Right Skewed or Positively Skewed. **Mean:** **4417983.349172932** is greater than **Median: 3043606. And Skewed value is 4.89.**

A graph with blue lines and numbers

Description automatically generated

Figure 8: Distribution of Irish Rail

#### It show data Irish Rail data is partially normally distributed high variance to right. Hence, it is Symmetrical Distribution with **Mean: 16015.28, Median 13072 and Skewed value is 0.88.**

* **Kurtosis of data**

Kurtosis is the measure of describing the distribution of data.

**Here, kurtosis Value for Paris Rail Passenger is 29.48. Therefore, it has Leptokurtic i.e. Data is Heavy Tails**

**Here, kurtosis Value for Irish Rail Passenger is 0.177. Therefore, it has Leptokurtic i.e. Data is Heavy Tails**

After removal of outliers Data is Normally or Symmetrically distributed and its skewed value is tending to **0.66.** Also, it’s kurtosis value is **-0.209.** Hence, Data distribution is light tailed.

1. **Scatterplot Correlation**

We understand that correlation is a statistical metric indicating the connection between the relative movements of two variables. In cases of correlation, the data points align along a line or curve. The strength of the correlation is reflected in how closely the points adhere to the line.From above scatter plot. There is only positive correlation between city and station regarding passenger. Other than that, there is No Correlation and Negative Correlation.

### Inferential statistics

Inferential statistics is a branch of statistics employing analytical methods to draw conclusions about a population through the examination of random samples. The objective of inferential statistics is to extend findings to the broader population. This involves utilizing statistics derived from sample data, such as the sample mean, to infer characteristics of the population parameter, such as the population mean.

## Hypothesis Testing

Hypothesis testing serves as a statistical approach to assess whether there is sufficient evidence in sample data to make inferences about a population. The process entails creating two opposing hypotheses: the null hypothesis (H0) and the alternative hypothesis (Ha). Subsequently, data is gathered to evaluate the available evidence. More detail in jupyter notebook

1. **Hypothesis testing performed on dataset of Paris rail.**

After performing Hypothesis testing mention in Jupyter file. The results are given below:

*t-test/t-statistic: - 0.3263789137198776*

*Critical t-Value: - 2.009575234489209*

*p-Value: - 0.7455276532812358*

*Confidence of Interval: - (2645599.005, 3686307.234)*

**Fail to reject the null hypothesis: There is no significant impact of the new train on passenger means of traveling. Hence, Null Hypothesis is Accepted.**

1. **Hypothesis Test Parametric and Nonparametric on dataset of Paris rail**

After performing Hypothesis testing mention in Jupyter file. The results are given below:

***Parametric Test***

*(t-test) t-test/t-statistic: -1631411763492762.0*

*p-Value: - 0.0*

***Non-Parametric Test***

*(Wilcoxon Signed-Rank Test) t-test/t-statistic:- 0.0*

*p-Value:- 1.862645149230957e-09*

*Confidence of Interval:- (13959.983, 17928.61)*

***Parametric Test: - Reject the null hypothesis:***

*The new train has a significant impact on passenger means of traveling.*

***Non Parametric Test: Reject the null hypothesis:***

*The new train has a significant impact on passenger means of traveling.*

1. **ANOVA Test**

ANOVA, or Analysis of Variance, is applied to examine whether there are statistically significant differences in means among two or more groups.

### One Way ANOVA

It is performed when only 1 independent variable is considered. Here we perform One way ANOVA test of type\_of\_journey independent variable from Irish Rail data. ['Mainline and other services', 'Dublin suburban services', 'International journeys', 'DART'] are unique category or type of journey.

1. If p-Value < alpha (0.05): Reject the Null hypothesis **else** Fail to reject the Null hypothesis.

*One-way ANOVA Result*

*F- Statistics: - 32.50922377709635*

*p- Value: - 1.4020181420511712e-11*

***Reject the null hypothesis: The new train has a significant impact on passenger means of traveling.***

### Two Way ANOVA

It is performed when more than 1 independent variable is considered. Here we perform Two way ANOVA test of type\_of\_journey and city independent variable from Paris Rail data.

A screenshot of a computer

Description automatically generated

Figure 10: Result of Two Way ANOVA

***Reject the null hypothesis: The new train has a significant impact on passenger means of traveling.***

### Significance level:

#### The significance level is characterized as the likelihood of rejecting the null hypothesis when it is true. For instance, a significance level of 0.05 implies a 5% chance of erroneously concluding a difference exists when, in fact, there is none. This parameter is commonly denoted by the symbol alpha (α).

### p-Value

#### The p-value is the probability of observing a t-statistic as extreme as the calculated value under the assumption that the null hypothesis is true. If the p-value is sufficiently low, it provides a basis for rejecting the null hypothesis. Specifically, we reject the null hypothesis when the p-value is below the chosen significance level.

### t-Test

#### A t-test is employed when the data adheres to a Student's t-distribution, and the sample size is below 30. It serves the purpose of comparing the mean of a sample to the mean of a population when the population variance is unknown.

### Wilcoxon signed rank test

#### The Wilcoxon rank-sum test is utilized for comparing two independent samples, whereas the Wilcoxon signed-rank test is employed to compare two related samples, matched samples, or to conduct a paired difference test for repeated measurements within a single sample, assessing whether their population mean ranks exhibit differences.

## Insights from Paris Rail Dataset

In Paris, number of people preferred public transport as train is extremely high. In year 2013-2021 a total of **14,689,794,636** traveled by Metro and RER only in Paris. It is known that use of Metro is comparatively greater than RER as per data. Highest passenger travel by Metro is recorded as **1.03Billion in 2018** where **187.3921 Millions by RER** in 20219. And in 2020 downfall of passenger travel can be seen in both type of journey **580.9343M in Metro and 118.902M in RER.**  The effect of COVID-19 can be seen on the public transportation in Paris as well (Laroche, 07 April 2022). **Noisy-le-Grand Mont d'Est** train station is recorded highest number of passengers of **62.9M. Paris city** has recorded **6.51Billions** of passenger travel by both Metro and RER.

A graph of a travel by rail

Description automatically generated

Figure 11: People Travel By Rails in Paris

## Insights from Irish Rail Dataset

#### Rail passenger travel in Ireland from 1981-2022 reached **23,22,672**, displaying steady growth. In 1981, **15.374K** passengers used trains, facing a decline in 1982 to **12.81K**. From 1983 to 2007, there was consistent annual growth, peaking at **45.511K** in 2007. After a downturn in 2008, it took a decade to recover by 2017. In 2019, a peak of **50.06K** passengers was recorded, followed by a decline in 2020-21 during the COVID-19 period, reaching **17.41K**.

A graph showing the growth of travel

Description automatically generated

Figure 12: Passenger travel by Rail in Ireland

In Ireland as per data the average number of passengers took DART as per travel is **17K** which is highest number among other type of journey and Dublin Suburban services is second highest of **13.9K**.

A graph of different colored squares

Description automatically generated

Figure 13:Type of journey by Train in Ireland

# Comparison Between Passenger Journey in Ireland and Paris

The comparison of passenger’s uses trains as mean of transport in both Ireland and Paris from 2013-2021. From the below image the use of train in Paris is comparatively higher than in Ireland. Paris has highest number of passengers of **1.21Billions in 2018** and Ireland has **50.06K in 2019**. And downfall of volume of passenger can be seen in both place in 2020. As we know that was the time of COVID-19 where all the transportation are close except basic required one. In the given image it can be clear the rise and downfall in passenger over years.

A graph with a line going up

Description automatically generated

Figure 14.a : Comparison between Passenger Volume in Ireland Vs Paris

A graph with a line going up

Description automatically generated

Figure 14.b : Comparison between Passenger Volume in Ireland Vs Paris

# Machine Learning

Machine learning models play a crucial role in data analytics by leveraging algorithms and statistical models to analyze and interpret complex datasets. Here are several ways in which machine learning models contribute to data analytics like Predictive Analytics, Regression Analysis, Natural Language Processing (NLP) etc. Here in this project using Regression model for Prediction of passenger travel of Paris Rails.

## Data Pre processing

### Encoding Data for Machine Learning Models

#### In the data frame columns like type\_of\_journey, station, city are object datatypes. If we give that data to our machine learning model, there will be high risk of inaccurate prediction because the machine learning model deals with integers rather than categorial or label data. So, to encode data several encoding techniques are used to encode dataframe column.

### One Hot Encoder:

#### type\_of\_journey is column is best fit for One Hot Encoding and implemented on this column because it has two categories of data i.e. Metro an RER.

### Label Encoder:

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

#### A white background with black and white clouds Description automatically generated

Figure15: Data set after label encoding

#### Here in the above image, after applying the encoding technique we can see type of journey column has changes categorical text to numerical 0.0, 1.0. station, and city column has been converted from labeled value to numeric value with different numbers for each observation and so on.

## Selection of Machine Learning

### Supervised Learning

#### The dataset's dependent variable is the unlabeled continuous data of total passengers traveling in Paris, categorized by years, stations, and cities. With no binary labels (true or false), Supervised Learning Machine learning models, including Simple Linear Regression, Multi-Linear Regression, Ridge Regression, and Lasso Regression, are employed for predicting rail passenger travel in Paris based on city, station, and years. These models are components of the Regression methodology.

### Feature Scaling:

#### In machine learning, Feature Scaling standardizes or normalizes numerical features to ensure uniform contribution to model performance. Employing Standardization (Z-score normalization) is chosen for scaling dependent and independent variables due to dataset outliers and adherence to Gaussian Distribution. The scaled dataset exhibits improved consistency and uniformity in feature values. (GeeksforGeeks, n.d.).

A white background with black and white clouds

Description automatically generated

Figure 15: Standardizing of Dataset

## Selection of Random state, and data for Training and Testing

#### I have used 20% (0.2) for testing the data set and the rest 80% of the data for training, Also I have selected the random state 42 known as best random value for all the models.

## Simple Linear Regression

Simple linear regression, a fundamental statistical model, examines the linear association between two quantitative variables. In this case, the independent variable (predictor) is the Year column, and the dependent variable (response) is the Passenger column. The best-fit line has an intercept of 0.0 and a slope of [1.]. both model training and testing achieved accuracy is 100% for both Training and Testing phases.

## Multiple Linear Regression

Multiple linear regression is an extension of simple regression, differing in the inclusion of multiple independent variables to predict a single dependent variable (e.g., year, city, station predicting passengers). **The best-fit line has an intercept of 0.0089 and slopes of [0.0932, -0.0273, -0.0225, -0.1977].** Post-training and testing, the model achieved 100% accuracy for both training and testing datasets.

## Performance Evaluation of Simple and Multiple Linear Regression

A table with numbers and symbols

Description automatically generatedA graph of multiple linear regression

Description automatically generated with medium confidence

Figure 16.a: Simple and Multiple Linear Regression, dataframe. Figure 17.b. Graphical representation of score

## Ridge Regression

#### To counter multicollinearity in linear regression, ridge regression serves as a regularization technique by reducing coefficients of highly correlated predictor variables. This prevents overfitting by introducing a penalty term to the least squares objective. The regularization strength is controlled by the hyperparameter alpha. GridSearchCV is employed to identify the optimal alpha from a range of values. In this case, the best alpha is 55, yielding a ridge score of -0.9860. The selection of dependent and independent variables remains consistent with multiple linear regression.

## Lasso Regression

It is like Ridge regression. It works by introducing a bias term but instead of squaring the stop, the absolute value of the slope is added as a penalty term. It helps reduce overfitting and it is particularly useful for feature selection. Ridge regression can reduce the slop close to zero (but not exactly zero). But, Lasso regression can reduce the slope to be exactly equal to zero.

## Performance Evaluation of Ridge Regression and Lasso Regression

A table with numbers and a number

Description automatically generatedA graph of a bar graph

Description automatically generated with medium confidence

RepresentationFigure 18.a: Score of Ridge and Lasso in Dataframe. Figure 18:b: Score of Ridge and Lasso Graphical

### R^2 score

#### The R^2 score gauges the model's capability to explain variance, with higher scores indicating better performance. However, excessive scores may hint at overfitting, compromising generalization to new data. (FERNANDO, 2023).

### Root Mean Square Error (RMSE):

#### Root Mean Square Error (RMSE) measures the average squared difference between predicted and actual values, reflecting error spread. A lower RMSE signifies superior model performance. (TO, 2023).

### Mean Absolute Percentage Error (MAPE):

#### Root Mean Square Error (RMSE) measures the average squared difference between predicted and actual values, reflecting error spread. A lower RMSE signifies superior model performance. (TO, 2023).

All Score are mentioned above in form of Dataframe and represented in form of Bar Graph for both Ridge and Lasso Regression.

# Sentiment Analysis

## Sentiment Analysis on Public Transport

#### Sentiment Analysis is performed on public transport of Dublin and Netherlands to get thought, view on public transport and data has been collected from reddit by using **PRAW** python library. A Python module called PRAW provides an easy-to-use interface for interacting with the Reddit API. It simplifies operations like posting on Reddit, getting information from the network, and accessing its data. PRAW makes it easier for developers to integrate Reddit features into Python applications by streamlining the complex processes of handling HTTP requests and processing JSON answers.

To get data of public transport people’s review. ‘**ireland**’ is used as subreddit name and **topic keyword is “Dublin is worst capital in Europe for public transport ticket pricing, finds Greenpeace”.** And ‘**Netherlands**’ is used as subreddit name and **topic keyword is “Thoughts on free public transport in Netherlands**?”. After using several Natural Language Processing technique like text processing, tokenization, bag of words, stemming, term frequency, inverse document frequency sentiment analysis is performed.

### Sentiment Question for ‘Dublin’ is.

#### Dublin is worst capital in Europe for public transport ticket pricing, finds Greenpeace and sentiment on this topic is given below:

**neutral 178**

**positive 58**

**negative 44**

### Sentiment Question for ‘Netherlands’ is.

#### Thoughts on free public transport in Netherlands? and sentiment on this topic is given below:

**neutral 279**

**positive 87**

**negative 35**

## Comparison between Sentiment Analysis on Public Transport of Ireland Vs Netherlands

**A graph with numbers and a bar

Description automatically generated with medium confidence**

Figure 19: Sentiment Analysis on public transport review of Ireland(Dublin) vs Netherlands

According to comment Data collected from reddit, Netherland has more positive review of the people on public transport than that of Dublin. Also, fewer negative sentiment over Dublin regarding public transport.

# Python

#### I have explained all the requirement in jupyter notebook file in Python section before Dashboard because report file is already more words than required.



Figure 20: Dashboard video

# Dashboard

It is also available in Jupyter Notebook File

**Double Click on Image to play** Its Video of dashboard from Jupyter Notebook.

# Reference

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

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

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

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

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

CML/DRC/ATTR, F., 17 may 2021. *Incoming annual traffic per station.* [Online]   
Available at: https://data.ratp.fr/explore/?sort=modified  
[Accessed 7 12 2013].

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

Council, D. C., 2023-07-01 - 2023-12-31. *Traffic Volumes from SCATS Traffic Management System Jul-Dec 2023 DCC.* [Online]   
Available at: https://data.gov.ie/dataset/dcc-scats-detector-volume-jul-dec-2023  
[Accessed 10 12 2023].

Council, D. C., 2023. *Creative Commons Attribution 4.0.* [Online]   
Available at: https://creativecommons.org/licenses/by/4.0/  
[Accessed 7 12 2023].

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

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

Dorgan, N., 2022-11-30. *TCA01 - Passenger Journeys by Rail (Thousand).* [Online]   
Available at: https://data.gov.ie/dataset/tca01-passenger-journeys-by-rail-thousand  
[Accessed 7 12 2023].

Dorgan, N., 2022-12-02. *NRA03 - National Route Length.* [Online]   
Available at: https://data.gov.ie/dataset/nra03-national-route-length  
[Accessed 8 12 2023].

Dorgan, N., 2023-09-19. *TAM06 - Passengers handled by main airports.* [Online]   
Available at: https://data.gov.ie/dataset/tam06-passengers-handled-by-main-airports  
[Accessed 7 12 2023].

Dorgan, N., 2023-10-18. *THA24 - Passenger Journeys by Public Transport.* [Online]   
Available at: https://data.gov.ie/dataset/tha24-passenger-journeys-by-public-transport  
[Accessed 7 12 2023].

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

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

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

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

International, A. 4., 2023. *CC BY 4.0 DEED.* [Online]   
Available at: https://creativecommons.org/licenses/by/4.0/  
[Accessed 8 12 2023].

Ireland., H. (. E. o. C.-1. r. i., 2021. *The Pandemic One Year on: Trends and Statistics Between Three Waves of the COVID-19 Pandemic in Ireland.* [Online]   
Available at: https://publicpolicy.ie/covid/the-pandemic-one-year-on-trends-and-statistics-between-three-waves-of-the-covid-19-pandemic-in-ireland/  
[Accessed 15 12 2023].

Ireland, G. o., 2023. *IRELAND'S OPEN DATA PORTAL.* [Online]   
Available at: https://data.gov.ie/dataset/?theme=Transport  
[Accessed 5 12 2023].

Ireland, G. o., 2023. *Accessibility.* [Online]   
Available at: https://data.gov.ie/pages/accessibility  
[Accessed 5 12 2023].

Ireland, T. I., 2023. *TOA11 - Luas Passenger Numbers.* [Online]   
Available at: https://data.gov.ie/dataset/toa11-luas-passenger-numbers/resource/3915960b-fe0d-41b8-9316-d9b9ce0ca816#&r=TLIST(A1)&c=STATISTIC  
[Accessed 7 12 2023].

Laroche, F., 07 April 2022. *The effect of COVID-19 on long-distance transport services in France.* [Online]   
Available at: https://rsaiconnect.onlinelibrary.wiley.com/doi/10.1111/rsp3.12534  
[Accessed 10 12 2023].

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

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

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

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

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

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

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

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

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

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

u/Longjumping-Stretch5, 2023. *Times, The Irish.* [Online]   
Available at: https://www.reddit.com/r/ireland/  
[Accessed 7 12 2023].

u/R\_Blitzie, 2023. *r/Netherlands/.* [Online]   
Available at: https://www.reddit.com/r/Netherlands/comments/15kelue/thoughts\_on\_free\_public\_transport\_in\_netherlands/  
[Accessed 7 12 2023].

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

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