**CCT College Dublin**

**Assessment Cover Page**

*To be provided separately as a word doc for students to include with every submission*

|  |  |
| --- | --- |
| **Module Title:** | Programming, Stadistics, Machine learning for Data Analytics and Data preparation and Visualization |
| **Assessment Title:** | MSC\_DA\_CA2 |
| **Lecturer Name:** | David McQuaid, Muhammad Iqba, Taufique Ahmed and Sam Weiss |
| **Student Full Name:** | Jesus Rodrigo Colina Nunez |
| **Student Number:** | 2017156 |
| **Assessment Due Date:** | 07/11/2024 |
| **Date of Submission:** | 07/11/2024 |

**Jesus Rodrigo Colina Nunez**

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

**MSc in Data Analytics**

 (SimpliLearn, 2023)

MSC\_DA\_CA1

Jesus Rodrigo Colina Nunez

Student Number: 2017156

**Table of Contents**

[Abstract 3](#_Toc155946490)

[Data preparation and Visualization 4](#_Toc155946491)

[Raw data 4](#_Toc155946492)

[Licensing and permissions 4](#_Toc155946493)

[Data Preparation and Programming 5](#_Toc155946494)

[Interactive dashboard 10](#_Toc155946495)

[Time Series Analysis for Individual Lines: 10](#_Toc155946496)

[Statistics for Data Analytics 12](#_Toc155946497)

[Descriptive statistics 12](#_Toc155946498)

[Inferential statistics 13](#_Toc155946499)

[Parametric and non-parametric inferential statistical techniques 13](#_Toc155946500)

[Tests conclusion 15](#_Toc155946501)

[Machine Learning 16](#_Toc155946502)

[Linear Regression 16](#_Toc155946503)

[K-Nearest Neighbors (KNN) 16](#_Toc155946504)

[Cross-Validation for Supervised Learning Models: 17](#_Toc155946505)

[Sentiment Analysis 18](#_Toc155946506)

[GitHub Link 20](#_Toc155946507)

[Bibliography 21](#_Toc155946508)

# Abstract

*This assignment and research aim on public transportation, I have chosen the datasets; Luas public transport in Dublin and the Subway called “Subte” in Buenos Aires Argentina. Both data sets focused on monthly ridership over the years.*

*Jupyter notebook will be used in this assignment to perform the analysis for the datasets selected. The proses model used is The Cross Industry Standard Process for Data Mining (CRISP-DM) and Exploratory Data Analysis (EDA) are performed to initiate investigations on the data.*

*After an extensive review on the data sets, It was decided that the area of fucus will be the forecasting the yearly ridership of the next 5 years. Linear Regression and KNN will be used to forecast the next 5.*

*A sentiment analysis is done by using Reddit API, the sentiment analysis is focused on Ireland’s Public transportation (for e.g., Dublin Bus, Luas, Tram, Irish Rail etc.) In the sentiment analysis 1000 post were analysed using TexBlob.*

# Data preparation and Visualization

## Raw data

Regarding the process of acquiring this data, it is important to note that it involved thorough research across different websites, which was time-consuming and not straightforward. This highlights the challenges often encountered in data acquisition, including the need to navigate through various sources and assess the reliability and relevance of the data found. The licensing and permissions associated with the data are typically governed by the terms set by the data provider, in this case, the Dirección General de Estadística y Censos. It's crucial to adhere to these terms, especially when using the data for research or publication, to ensure compliance with legal and ethical standards (Dirección General de Estadística y Censos, 2020).

Various datasets were downloaded and analysed, but it did not meet the criteria for comparison or were too complex to use. I think in my opinion the search for the new data set was one of the most challenging parts of this assignment, because not manly data sets were suitable to be compared but, it also had to come from a reliable source like form the government.

The Dublin dataset was easy to find as my research led me to the link that was provided by the lectures.

## Licensing and permissions

The data on Buenos Aires subway usage, from the period January 2010 to October 2023, can be found on the provided webpage. This website includes statistical data on the number of passengers using the subway and premetro services in Buenos Aires. The data is sourced from the Dirección General de Estadística y Censos, based on information from CNRT (until August 2020) and SBASE (from September 2020 onwards) (Dirección General de Estadística y Censos, 2020).

The website data.gov.ie provides data on passenger journeys on the Luas. The dataset is published by Transport Infrastructure Ireland and is licensed under the Creative Commons Attribution 4.0 license. The data is available in various formats, including CSV, JSON-STAT, PX, and XLSX, making it accessible for different uses. The open data license is a positive aspect, as it allows for easier use and distribution of the data. (Open Data Unit, 2023)

# Data Preparation and Programming

Data preparation is the process of taking raw data and getting it ready for ingestion in an analytics platform. To achieve the final stage of preparation, the data must be cleansed, formatted, and transformed into something digestible by analytics tools (Chen, 2023).

3 Data set will be used for this process:

* LuasPassengerNumbers.csv
* PassengerJourneysbyLuas.csv
* SubteBuenosAires.xlsx

For these analysis snipped screen shots were taken from my Jupyter Notebook, these are only examples of what I did, full code, images, plots, charts and results can be found on my Jupyter Notebook File or Github.

**1.- Overview the data**

Firstly, we'll look at the basic structure of the dataset, including the number of rows and columns, types of data in each column, and a glimpse of the first few rows.

A screenshot of a computer code

Description automatically generated

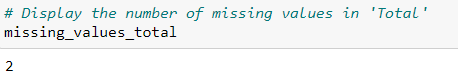
This will give us the size of the data set SubteBuenosAires.xlsx.

**2.- Summary Statistics**

Summary statistics will be generated for each column. This includes count, mean, standard deviation, min, max, and percentiles for numeric columns. It helps to understand the distribution and variability of the data.

**3.- Data Cleaning**

Here, missing values are checked, and then decide how to handle them (e.g., imputation or removal). Anomalies are also looked or outliers that may need attention.



2 values are missing in the column "Total", I will replace those missing values with the median, it would be better option rather than removing both rows, because we would be missing two months.

**4.- Analysis of Individual Variables**

Each variable will be explored separately, for numerical variables distribution histograms are used. For categorical variables like months, we might use count plots to see the frequency of each category.

A graph of a number of data

Description automatically generated with medium confidence

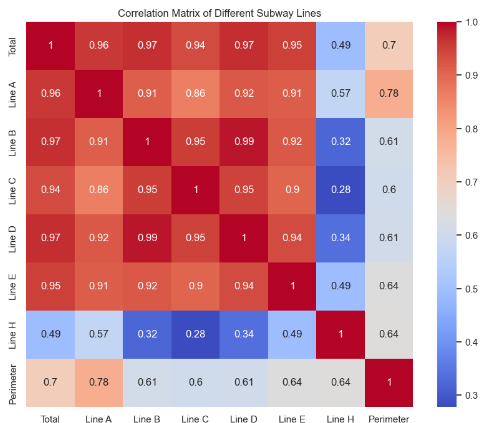
Histograms for Each Line:

These histograms show the distribution of ridership for each subway line.

We observe different patterns in ridership across lines, indicating varying usage intensities and possibly different user demographics or purposes.

**5.- Correlation Analysis**

Relationships between different numerical variables will be examinated. This is typically done using a correlation matrix and visualized through a heatmap.



The heatmap presents the correlation matrix of different subway lines and total ridership:

There is a strong positive correlation between the ridership of different lines, indicating that when one-line experiences high ridership, others tend to as well.

The 'Total' ridership is highly correlated with each individual line, showing that the total ridership is a good representation of the overall system usage.

Some lines have stronger correlations with each other, suggesting they might serve interconnected or similarly trafficked areas.

**6. Time Series Analysis**

Since the data has a time component (year and month), we can explore trends, seasonality, and patterns over time.

**9. Comparative Analysis**

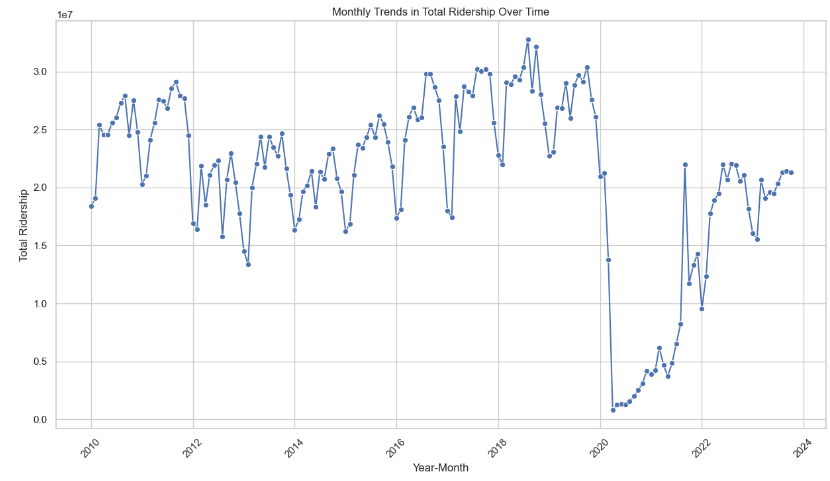
Here comparation different subway lines is done in terms of ridership and look for any interesting patterns or differences.

A chart of different lines

Description automatically generated with medium confidence

**8. Visualization**

Throughout the EDA, various plots and charts are used to visualize our findings, making them easier to understand and interpret.



The line plot shows the monthly trends in total over time:

There are noticeable fluctuations within each year, suggesting a potential seasonal pattern in ridership.

Some years show significant peaks or troughs, which could be due to various external factors for example (events, weather conditions, operational changes and the most important in 2020 we see it goes down because of COVID – 19 and now it is slowing increasing).

The trend over the entire period can be better understood by observing changes in peak and trough levels across different years.

## Interactive dashboard

The dashboard is designed to provide interactive visualizations of the Buenos Aires subway system ridership data. It allows to explore ridership trends both yearly and over time for individual subway lines, including a comprehensive view of the total ridership.

**Yearly Ridership Visualization:**

It is possible to select multiple subway lines from a checklist. The dashboard displays a bar chart showing the aggregate yearly ridership for the selected lines.

This feature helps in comparing the performance of different subway lines year-over-year and understanding overall trends.

## Time Series Analysis for Individual Lines:

A dropdown menu enables users to select a specific subway line or the total ridership.The dashboard then presents a line chart illustrating the ridership trends of the selected line over time.

This is useful for detailed analysis of ridership patterns, identifying peak and off-peak periods, and assessing the impact of specific events.

**A graph of a subway rider dashboard

Description automatically generated with medium confidence**

**A graph showing a line

Description automatically generated with medium confidence**

This dashboard could serve as a powerful tool, offering insights into the usage patterns of the Buenos Aires subway system. Its interactive nature allows for customized analysis, aiding in decision-making and public awareness.

# Statistics for Data Analytics

## Descriptive statistics

The descriptive statistics of the Buenos Aires Subway (Subte) dataset provide the following insights:

A screenshot of a graph

Description automatically generated

The 'Total' column, representing total ridership, has a mean of approximately 21 million, with a standard deviation of 7.4 million. This indicates significant variability in total ridership over time.

A graph of blue bars

Description automatically generated with medium confidence

Here is the bar chart displaying the average monthly ridership for each subway line (A, B, C, D, E, H) of the Buenos Aires Subway (Subte) from 2010 to 2023.

From these statistics, it is evident that there is significant variability in the monthly ridership across different lines and over the years. Lines B and D consistently have higher ridership, while Line E and Line H have comparatively lower ridership. The large standard deviations in the total and line-specific riderships suggest fluctuations possibly due to seasonal trends, operational changes, or external factors affecting public transportation usage.

## Inferential statistics

The confidence interval is the range of values that you expect your estimate to fall between a certain percentage of the time if you run your experiment again or re-sample the population in the same way (Bevans, 2024).

These confidence intervals provide an estimated range for the true average ridership of the TOTAL for each year.

A graph with a line going up

Description automatically generated

There is a significant drop in ridership in 2020, likely due to the global COVID 19 pandemic, and a partial recovery in the subsequent years. Similar calculations can be made for other subway lines to compare trends.

Further analysis could explore the reasons behind the fluctuations in ridership, such as external events or changes in the subway system.

This analysis helps in understanding the variability and trends in ridership over different years, providing a more detailed perspective on how subway usage has evolved

## Parametric and non-parametric inferential statistical techniques

For these tests I had to make some hypotheses and make some scenarios in order to perform them.  
All these tests can be found on my Jupyter Notebook file with their code.

1. **One-sample t-test**

Hypothesis: The mean ridership in 2010 for Line A is different from 4 million.  
A close up of a number

Description automatically generated

Conclusion: With a p-value > 0.05, we fail to reject the null hypothesis. This suggests that the mean ridership in 2010 for Line A is not significantly different from 4 million.

**2. Paired t-test**

Hypothesis: There is a significant difference in ridership between Line A and Line B in 2010.

A close up of numbers

Description automatically generated

Conclusion: With a p-value < 0.05, we reject the null hypothesis. There is a significant difference in ridership between Line A and Line B in 2010.

**3. Analysis of Variance (ANOVA)**

Hypothesis: The mean ridership among Line A, Line B, and Line C are equal.



Conclusion: With a p-value < 0.05, we reject the null hypothesis. This suggests that there are significant differences in the mean ridership among these lines.

**4. Wilcoxon Test**

Hypothesis: There is a significant difference in ridership between Line A and Line B in 2010.

A black and white text

Description automatically generated

Conclusion: With a p-value < 0.05, we reject the null hypothesis. This indicates a significant difference in ridership, supporting the paired t-test result but under the assumption of non-normality.

**5. Chi-squared Test**

Hypothesis: There is an association between 'Year' and 'Total' ridership being above the median.

A screenshot of a computer

Description automatically generated

Conclusion: With a p-value < 0.05, we reject the null hypothesis. There is a significant association between the year and whether the total ridership is above or below the median.

**6. The Mann-Whitney U test**

This test can be used to compare the distributions of the two datasets. This test does not assume a normal distribution and is suitable for comparing two independent samples.

Null Hypothesis: The distribution of monthly passenger counts is the same for Buenos Aires Subte and Dublin Luas.

Alternative Hypothesis: The distribution of monthly passenger counts is different between Buenos Aires Subte and Dublin Luas.

A black text on a white background

Description automatically generated

Conclusion: The null hypothesis is rejected. There is a statistically significant difference in the distribution of monthly passenger counts between the Buenos Aires Subte and the Dublin Luas.

## Tests conclusion

Deciding which statistical test to apply for each hypothesis required a thorough understanding of the tests and their applicability to different types of data also ensuring data was correctly formatted and translated was crucial, particularly for accurate interpretation and analysis.

This analysis not only provided statistical insights into the subway system's ridership but also highlighted the importance of contextual understanding in data analysis. It underscores the potential of using statistical methods to inform urban planning and public transport policies. The challenges faced during this process emphasize the need for careful data preparation and thoughtful selection of statistical tools.

# Machine Learning

The choice of K-Nearest Neighbors (KNN) and Linear Regression as machine learning models for forecasting passenger numbers in the Dublin Luas and Buenos Aires Subte datasets is grounded in their suitability for regression tasks and their distinct modeling characteristics.

Rationale and justification for selecting these models:

## Linear Regression

Linear Regression is one of the most fundamental and widely used machine learning models for regression tasks. It assumes a linear relationship between the input variables and the target variable.

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable (IBM, 2024).

It is computationally efficient and can be quickly implemented, making it suitable for datasets where a linear approximation of relationships is reasonable.

Linear Regression can effectively forecast values when the relationship between features and target variable is linear or close to linear.

## K-Nearest Neighbors (KNN)

KNN is a non-parametric, instance-based learning algorithm primarily used for classification but also suitable for regression. It predicts the output based on the 'k' closest training examples in the feature space.

The K-Nearest Neighbors (KNN) algorithm is a popular machine learning technique used for classification and regression tasks. It relies on the idea that similar data points tend to have similar labels or values (Srivastava, 2024).

Unlike Linear Regression, KNN does not assume a linear relationship between features and the target. It can capture more complex patterns.

In conclusion, the selection of Linear Regression and KNN, accompanied by systematic hyperparameter tuning and appropriate feature selection, provides a robust approach to forecast future passenger numbers. This combination offers both a simple linear approximation and a more flexible.

I tried to use the models to forecast the next 5 years and I noticed that the KNN did not work.

A white background with black text

Description automatically generated

A graph with lines and numbers

Description automatically generated

A graph with lines and numbers

Description automatically generated

**Conclusion of my 5-year forecast**

The Linear Regression model's prediction of a declining trend might be influenced by recent years' data, where a significant drop (possibly due to external factors like the COVID-19 pandemic) is evident.

The KNN model's constant prediction might be due to its reliance on the most recent and similar data points, which do not vary significantly in the last few years.

## Cross-Validation for Supervised Learning Models:

It is commonly used in applied machine learning to compare and select a model for a given predictive modeling problem because it is easy to understand, easy to implement, and results in skill estimates that generally have a lower bias than other methods (Brownlee, 2023).

I used Cross-validation to assess the generalizability and performance of the Linear Regression and K-Nearest Neighbors (KNN) models to get a more reliable assessment of the models' performance.

This process involves splitting the dataset into 'k' subsets and using each subset as a test set while training on the remaining 'k-1' subsets. This is repeated 'k' times with each subset serving as the test set once. The average performance across all 'k' trials is used as the overall performance metric.

Mean Squared Error (MSE) was calculated for each fold as a performance metric.

A graph with lines and numbers

Description automatically generated

Both models show high MSE values, indicating a significant variance between the predicted and actual values. This might be due to the simplicity of the models and their inability to capture more complex patterns in the data.

The average MSE values are quite similar for both models, suggesting that neither has a distinct advantage in terms of prediction accuracy for this dataset.

The wide range in MSE values across different folds indicates variability in the data, which could be due to year-to-year fluctuations in passenger numbers.

In summary, the cross-validation results provide a more comprehensive view of the models' performance, confirming that both models have limitations in predicting this particular dataset accurately. This suggests the need for more complex modelling approaches or additional features that could better capture the factors influencing passenger numbers.

## Sentiment Analysis

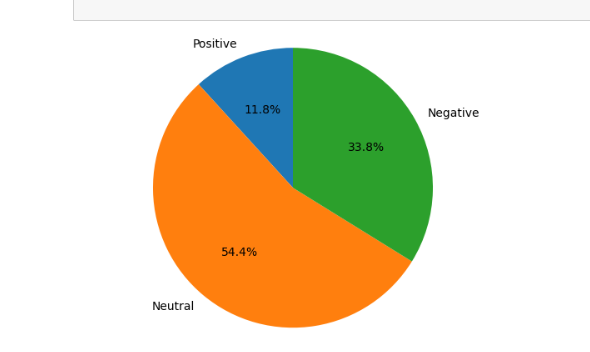
Sentiment analysis is extremely useful in social media monitoring as it allows us to gain an overview of the wider public opinion behind certain topics. (Bannister, 2023).

Public transport sentiment analysis is a method used to understand public opinions and feelings about public transportation systems, such as buses, Luas, trains, through the analysis of textual data. This data typically comes from social media platforms, online forums, reviews, or customer feedback. The analysis can provide insights into various aspects like customer satisfaction, common complaints, praises, or general public mood regarding the services.

To conduct this analysis, is typically gather textual data related to Dublin public transport from platforms like Reddit, Twitter, or customer feedback forms. Then, using sentiment analysis tools (like TextBlob or VADER), can categorize each piece of text as positive, negative, or neutral based on the sentiment it expresses.

For this example Reddit´s API was used to perform the sentiment analysis about the Public transport in Dublin.

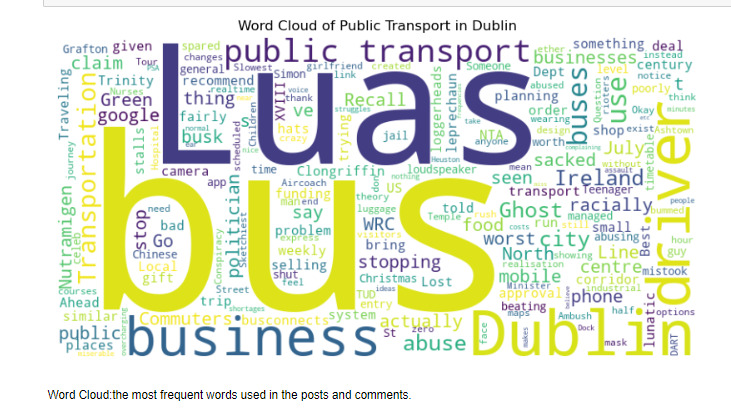
1000 Coments were analysed using TextBlob, and I plot its findings.

****

We can clearly see that there is a majority of the comments are neutral with the 54.4%, then proceeds the Negative with the 33. % and with a small part of 11.8% with positive comments. A predominance of negative sentiment could indicate widespread issues.

Recurring negative sentiments about delays, ticket prices, or overcrowding can pinpoint areas needing improvement.

For Dublin's public transport, such findings can guide improvements, policy-making, and public relations strategies.

****

# GitHub Link

**https://github.com/CCT2017156/MSC\_DA\_\_CA2**

# Bibliography

Anon., 2020. *estadisticaciudad.gob.ar.* [Online]   
Available at: https://www.estadisticaciudad.gob.ar/eyc/?p=46559  
[Accessed 08 December 2023].

Bannister, K., 2023. *brandwatch.com.* [Online]   
Available at: https://www.brandwatch.com/blog/understanding-sentiment-analysis/  
[Accessed 06 January 2024].

Bevans, R., 2024. *scribbr.com.* [Online]   
Available at: https://www.scribbr.com/statistics/confidence-interval/  
[Accessed 15 January 2024].

Brownlee, J., 2023. *machinelearningmastery.com.* [Online]   
Available at: https://machinelearningmastery.com/k-fold-cross-validation/  
[Accessed 09 January 2024].

Chen, M., 2023. *oracle.com.* [Online]   
Available at: https://blogs.oracle.com/analytics/post/what-is-data-preparation-and-why-is-it-important  
[Accessed 04 January 2024].

Dirección General de Estadística y Censos, 2020. *estadisticaciudad.gob.ar.* [Online]   
Available at: https://www.estadisticaciudad.gob.ar/eyc/?p=46559  
[Accessed 08 January 2023].

Dirección General de Estadística y Censos, 2020. *estadisticaciudad.gob.ar.* [Online]   
Available at: https://www.estadisticaciudad.gob.ar/eyc/  
[Accessed 06 December 2023].

IBM, 2024. *ibm.com.* [Online]   
Available at: https://www.ibm.com/topics/linear-regression#:~:text=Resources-,What%20is%20linear%20regression%3F,is%20called%20the%20independent%20variable.  
[Accessed 08 January 2024].

Open Data Unit, 2023. *data.gov.ie.* [Online]   
Available at: https://data.gov.ie/dataset/toa11-luas-passenger-numbers  
[Accessed 07 December 2023].

SimpliLearn, 2023. *SimpliLearn.* [Online]   
Available at: https://www.simplilearn.com/data-analytics-trends-article  
[Accessed 20 December 2023].

Srivastava, T., 2024. *analyticsvidhya.com.* [Online]   
Available at: https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/  
[Accessed 09 January 2024].