A logo with blue text

Description automatically generated

**“CA1 Graph and AI Report”**

**B9AI101 Graph and AI**

**Student:                           Shubham A Solse**

**Student ID:                      20042764**

**Module Lecturer:          Terri Hoare**

**Date:                               14-Dec-2024**

**Git-Hub link:** [**https://github.com/ShubhamSolse/CA01\_Graph\_and\_AI/tree/master**](https://github.com/ShubhamSolse/CA01_Graph_and_AI/tree/master)

**Note: Code cannot be submitted through zip file as it has size of 700+ MB please git clone the git link**

Table of Contents

[Exercise 1 (Individual) 3](#_Toc185198222)

[Exercise 2 (Group)**Irish Transport Dataset** 11](#_Toc185198223)

[Business Understanding 11](#_Toc185198224)

[Data understanding 11](#_Toc185198225)

[Data Preparation 17](#_Toc185198226)

[Modelling 17](#_Toc185198227)

[Evaluation 19](#_Toc185198228)

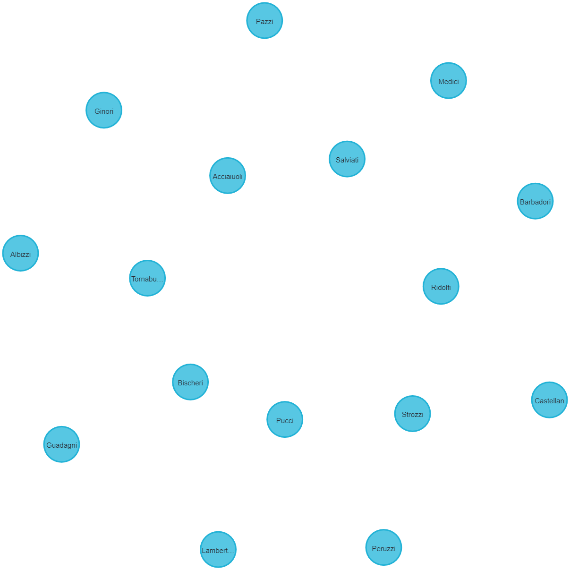
[Deployment 19](#_Toc185198229)

# Exercise 1 (Individual)

Padgett’s data on marriage alliances among leading Florentine families in the latter Renaissance is given in the figure below. Each node is a family, and each edge denotes a connection by marriage. Load the data as a Neo4j Knowledge Graph.

Answer:

* Creating nodes:



* Creating relationships

A blue circles with white text

Description automatically generated

* Compute Centrality Measures in Neo4j
  1. Degree Centrality

|  |  |
| --- | --- |
| Family | Degree Centrality |
| Medici | 10 |
| Tornabuoni | 6 |
| Guadagni | 6 |
| Peruzzi | 4 |
| Strozzi | 4 |
| Castellan | 4 |
| Ridolfi | 4 |
| Salviati | 4 |
| Albizzi | 4 |
| Bischeri | 4 |
| Barbadori | 2 |
| Pazzi | 2 |
| Acciaiuoli | 2 |
| Ginori | 2 |
| Lambertes | 2 |
| Pucci | 0 |

**A network of green circles and white text

Description automatically generated**

**Insights**

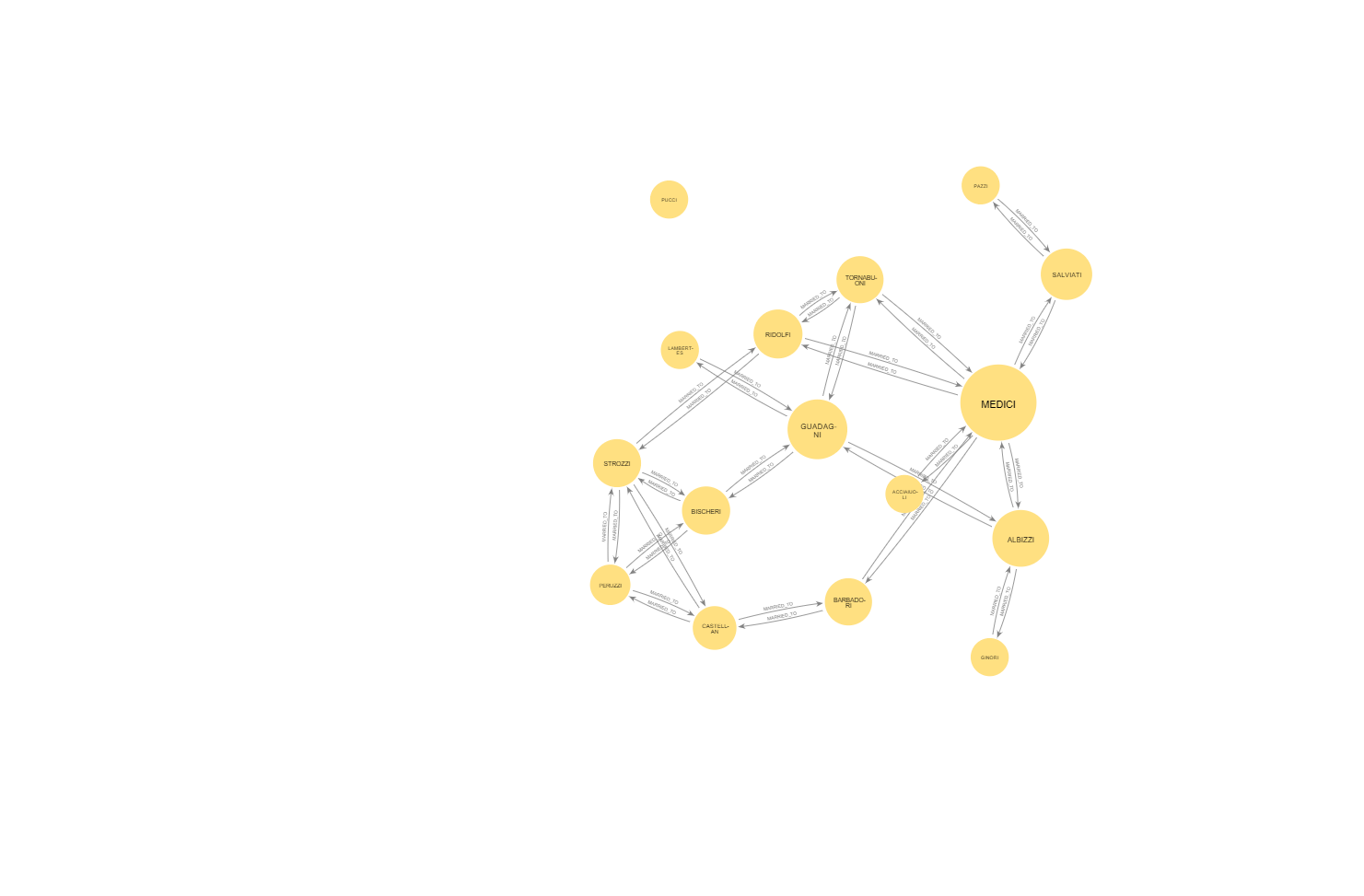
* **Medici's Centrality**: Medici is the central node in the network, very well connected and influential.
* **Middlemen**: The secondary hubs, Tornabuoni and Guadagni, assist in maintaining the network cohesive and impact mid-level families.
* **Mid-Level Families**: Peruzzi, Strozzi, and others moderately connected, suggesting regional influence but dependence on stronger hubs for broader access.
* **Peripheral Vulnerability**: The families such as Barbadori, Pazzi, and similar ones are weakly connected, indicating limited influence and higher vulnerability in the network.
* **Pucci’s Isolation**: Pucci is cut off from the network, which shows a chance to improve inclusivity and lessen the dependence on central families.
  1. Closeness Centrality

|  |  |
| --- | --- |
| Family | ClosenessCentrality |
| Tornabuoni | 0.482758621 |
| Medici | 0.466666667 |
| Guadagni | 0.4 |
| Ridolfi | 0.378378378 |
| Salviati | 0.341463415 |
| Albizzi | 0.341463415 |
| Barbadori | 0.325581395 |
| Acciaiuoli | 0.325581395 |
| Bischeri | 0.325581395 |
| Castellan | 0.311111111 |
| Lambertes | 0.291666667 |
| Peruzzi | 0.274509804 |
| Strozzi | 0.269230769 |
| Pazzi | 0.259259259 |
| Ginori | 0.259259259 |
| Pucci | 0 |

**Insights**

* **Tornabuoni and Medici's Influence**: Tornabuoni (0.48) and Medici (0.47) are crucial places to connect quickly across the network.
* **Role of Guadagni**: Guadagni (0.40) acts as a very good broker, which connects mid-level families to the central hubs for connectivity.
* **Mid-Level Connectors**: Ridolfi (0.37), Salviati, and Albizzi (0.34) are mid-level connectors that connect important and less important families in the network.
* **Peripheral Weakness**: Families like Pucci (0) and Ginori (0.26) are cut off, showing they have few roles and possible weaknesses in the network.
* **Opportunities for Growth**: Connecting Pucci, Castellan (0.31), and Lambertes (0.29) to core families would increase the strength and reach of the network.
  1. Betweenness Centrality

|  |  |
| --- | --- |
| Family | Betweenness Centrality |
| Medici | 61 |
| Tornabuoni | 54 |
| Guadagni | 30 |
| Ridolfi | 18 |
| Salviati | 13 |
| Albizzi | 13 |
| Bischeri | 12 |
| Castellan | 10 |
| Peruzzi | 4 |
| Strozzi | 3 |
| Pucci | 0 |
| Barbadori | 0 |
| Pazzi | 0 |
| Acciaiuoli | 0 |
| Ginori | 0 |
| Lambertes | 0 |



Insights

* **Medici's Dominance**: Medici (61) is first in betweenness centrality, reflecting the fact that it is the most important intermediary/influencer in the network.
* **Tornabuoni's Strategic Position**: Tornabuoni (54) is a second important connector to Medici that ties together several families in the network.
* **Bridging Role of Guadagni**: There are also a number of bridge nodes, one bridging between mid-level families (30-Guadagni).
* **Moderate Influencers**: Ridolfi (18), Salviati (13), and Albizzi (13) serve as the secondary connectors and link small clusters in the network.
* **Peripheral Families' Isolation**: the Pucci, Barbadori, Pazzi, and others have a betweenness centrality of 0, so they hold no key mediation positions.
  1. Eigenvector Centrality

|  |  |
| --- | --- |
| Family | Eigenvector Centrality |
| Medici | 0.585052928 |
| Tornabuoni | 0.427817702 |
| Guadagni | 0.280522315 |
| Salviati | 0.272698747 |
| Albizzi | 0.272698747 |
| Barbadori | 0.229087686 |
| Acciaiuoli | 0.229087686 |
| Ridolfi | 0.225403074 |
| Bischeri | 0.148422585 |
| Castellan | 0.124010166 |
| Pazzi | 0.111452107 |
| Ginori | 0.111452107 |
| Lambertes | 0.110190801 |
| Peruzzi | 0.09604474 |
| Strozzi | 0.088750699 |
| Pucci | 6.41E-127 |

**A network of green circles and dots

Description automatically generated**

**Insights**

* **Medici's Prominence**: Medici has the highest eigenvector centrality among all families, with 0.59, showing its strong influence and links with other powerful families.
* **Tornabuoni's Key Role**: Tornabuoni (0.43) follows Medici, highlighting its importance in the network through connections with highly connected families.
* **Guadagni's Significant Influence**: Guadagni (0.28) ranks third with a moderate level of influence, showing meaningful ties to the central families in the network.
* **Mid-Tier Influencers**: Salviati (0.27), Albizzi (0.27), Barbadori (0.23), and Acciaiuoli (0.23) are more in supportive roles with somewhat decent network influence.
* **Peripheral Families**: Pucci (6.41E-127) and others with low scores have minimal influence, meaning weak ties to central or influential families.
  1. PageRank

|  |  |
| --- | --- |
| Family | PageRank |
| Medici | 2.294305949 |
| Guadagni | 1.354791189 |
| Tornabuoni | 1.299931459 |
| Salviati | 1.031239473 |
| Albizzi | 1.031239473 |
| Strozzi | 0.936036072 |
| Peruzzi | 0.935590748 |
| Castellan | 0.927413664 |
| Bischeri | 0.925422261 |
| Ridolfi | 0.907033674 |
| Pazzi | 0.58568466 |
| Ginori | 0.58568466 |
| Barbadori | 0.536741655 |
| Acciaiuoli | 0.536741655 |
| Lambertes | 0.530750444 |
| Pucci | 0.15 |

**A network of green circles and dots

Description automatically generated**

**Insights**

* **Medici's Dominance**: The highest PageRank is Medici's, at 2.29, indicating this family was central and well connected within the whole network.
* **Guadagni and Tornabuoni's Importance**: Guadagni (1.35) and Tornabuoni (1.30) are also important hubs, though slightly below Medici in terms of network influence.
* **Contribution by Salviati and Albizzi**: Salviati (1.03) and Albizzi (1.03) both have high connectivity, hence are very important as mid-tier influencers in the network.
* **Peripheral Families**: Pucci (0.15) holds the lowest PageRank, showing a very minimal influence and also very weak connections with other families in the network.
* **Strong Secondary Nodes**: The families Strozzi (0.93) and Peruzzi (0.93) appear as strong connectors, linking critical clusters in the network.

Conclusion:

* **Medici's Dominance**: Medici is always at the top of the list for all centrality measures, which hints at a central position within the network. Its strong connections and positioning as a main intermediary make it the most influential family.
* **Importance of Secondary Hubs**: The families—Tornabuoni, Guadagni, Salviati, and Albizzi—act as significant secondary hubs. They connect important clusters and thus help to keep the network together; they facilitate connections among more influential families.
* **Peripheral Families' Weakness**: Indeed, the centrality scores of families like Pucci, Barbadori, Pazzi, and Ginori are low, which signals weak connections and less influence. Their isolation makes them vulnerable in the structure and decision-making processes of the network.
* **Opportunities for Growth**: Strengthening the connections between peripheral families (e.g., Pucci) and the central hubs could improve the overall cohesion of the network. This would decrease vulnerabilities and lead to a more connected and resilient network with better influence distribution.
* **Strategic Positioning for Influence**: Those families with high betweenness and eigenvector centrality, such as Tornabuoni and Guadagni, are in strategic positions linking key network clusters; strengthening their ties could enhance their influence and increase network stability and efficiency.

# Exercise 2 (Group)**Irish Transport Dataset**

## Business Understanding

1. Abstract

The following analysis and insights are the result of an in-depth study on the transport systems of Ireland, with an emphasis on LUAS, DART, and bus services. This project used actual data scraped from the official websites of these transport systems. Using data preparation, visualization in Neo4j, and exploratory data analysis, critical insights and patterns were identified. The study further investigated data integrity and then applied machine learning techniques for predictions on the transport metrics.

1. Introduction

Effective public transportation is a must if any city wants to move in sync and make the environment sustainable. In this project, the data for LUAS, DART, and bus services are used to infer a pattern in improving routes to elicit useful information. It will, therefore, put forth an integrated view of transport across Ireland using tools like Neo4j and Streamlit coupled with machine-learning techniques

## Data understanding

1. Data Collection

Data was scraped from the official websites of LUAS, DART, and bus services to ensure authenticity and reliability. The datasets included information such as:

* Timetables
* Routes
* Passenger counts
* Station connectivity

A map of a train and tram services

Description automatically generated

Figure Dart Service Route Map



Figure Dublin Bus Service Route Map

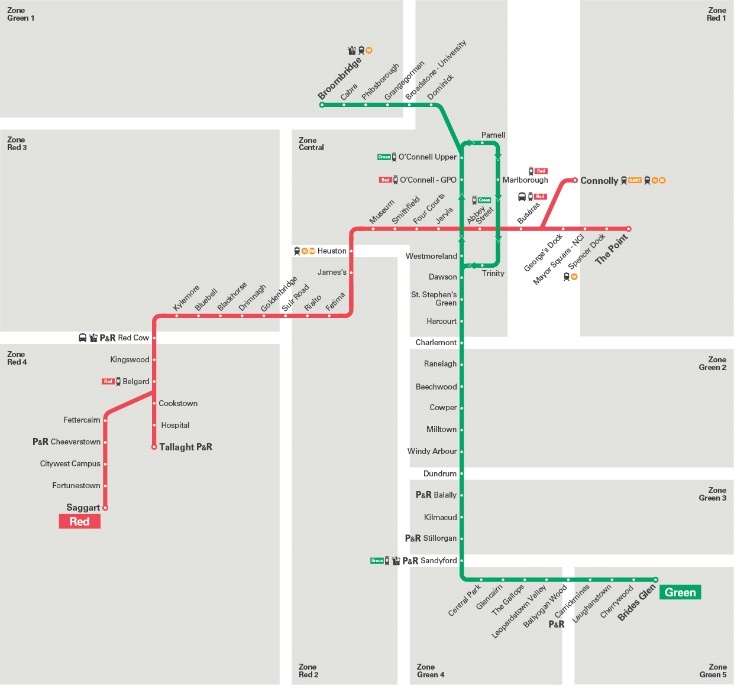


Figure Luas Service Route Map

1. Data Understanding

Understanding the dataset is one of the major steps for good analysis and creating models. Important steps in understanding data were done:

* **Exploratory Inspection**: A first look into the data sets to find out their structure, formats, and important variables. This involved learning how stations, routes, and schedules are related.
* **Attribute Analysis**: Considering how significant various attributes are, such as station names, routes, travel times, and passenger counts, for the objectives of the project.
* **Data Size and Variety**: Checking how big and varied the dataset is to ensure it covers LUAS, DART, and bus transport systems well. This step also involved looking at whether the dataset showed busy and quiet travel times.
* **Data Quality Checks**: Finding and fixing probable problems like double entries, alternative formats, or abnormal values within the data.

1. Exploratory Data Analysis
   1. DART

A graph of green rectangular bars

Description automatically generated

Figure Facilities availability in DART stations

A graph of stations by route names

Description automatically generated

Figure Top 10 Dart Stations by Routes serviced

A pie chart with numbers and a triangle

Description automatically generated

Figure Weekend Available Dart Stations

A screen shot of a computer screen

Description automatically generated

Figure Facility Distribution

* 1. LUAS

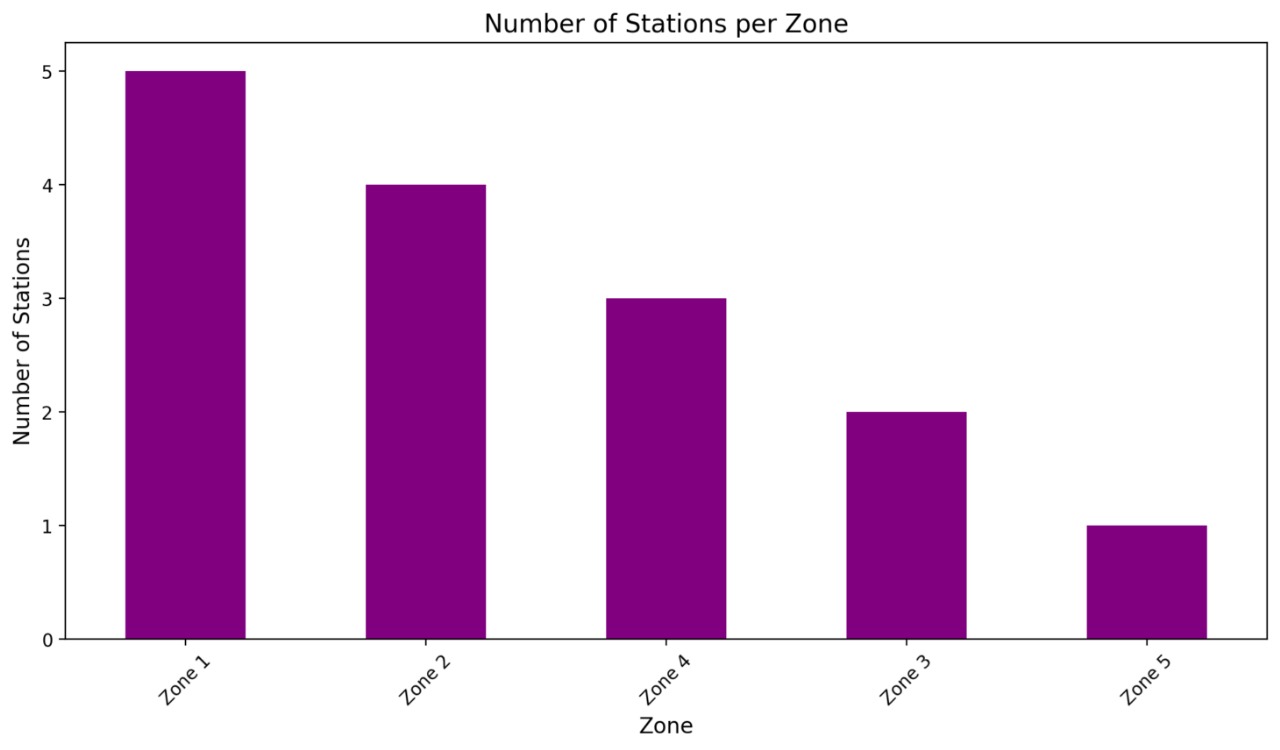


Figure Number of Stations per zone

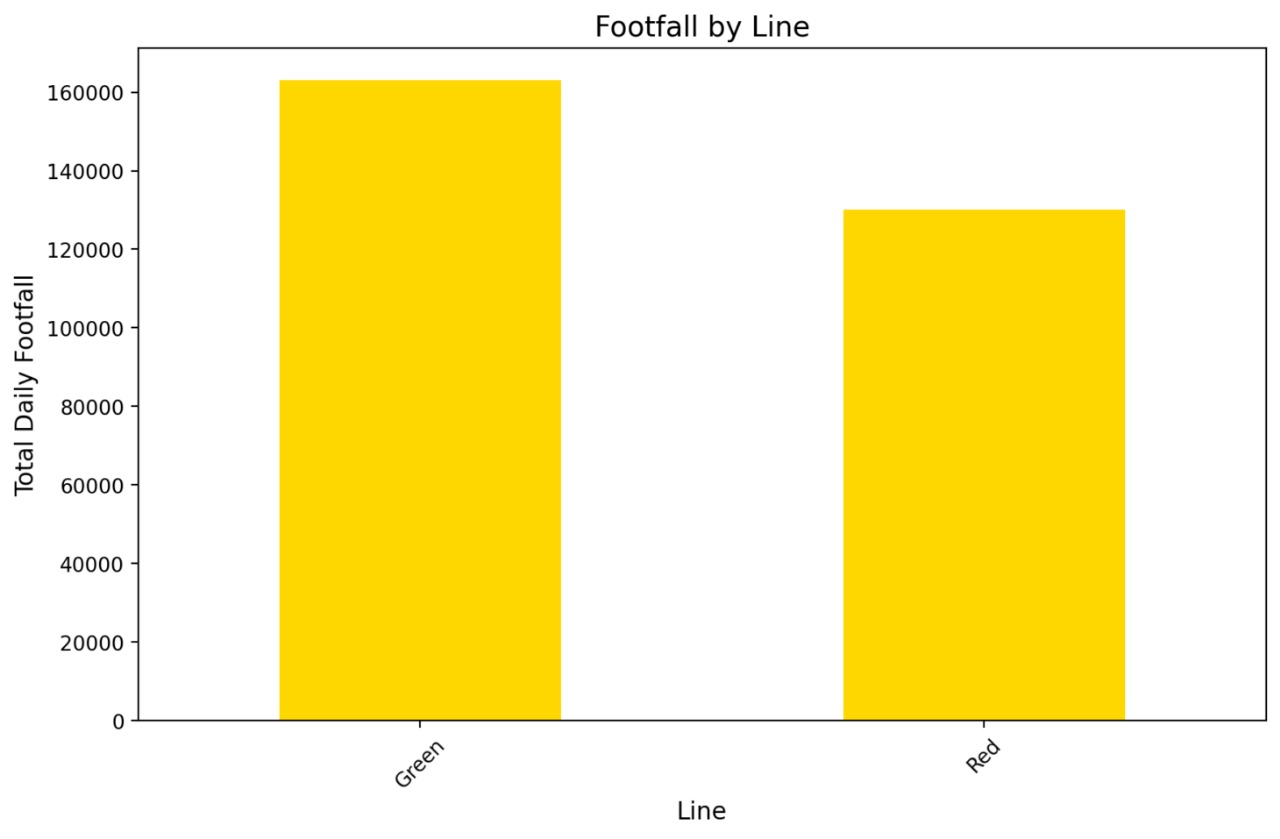


Figure Footfall by line

A screen shot of a number

Description automatically generated

Figure Stations Count per zone

* 1. BUS

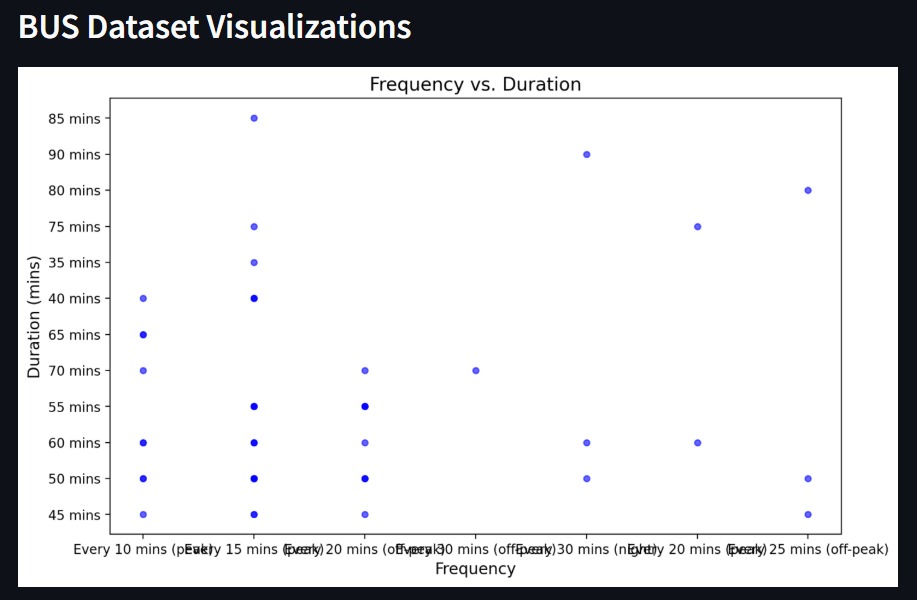


Figure Frequency vs Duration

A graph showing route numbers

Description automatically generated

Figure Top 10 routes by key Landmarks

## Data Preparation

The collected data was carefully cleaned and prepared to make sure it could be used for analysis. Important steps in preparing the data included:

* **Handling Missing Values**: Missing data points were analyzed to check if they have any significant effects. Imputation of the missing values was performed using techniques such as forward fill for time-series data and mean/mode imputation for numeric and categorical data, respectively.
* **Duplicate Removal**: Removed duplicate records to avoid duplication in the analysis.
* **Data Integration**: Combined datasets from LUAS, DART, and bus services to create a unified transport network dataset. Merging the datasets together was done by common attributes like station names and route connections.
* **Validity Check**: It has been checked against the official transportation schedules and the known data to ensure cleaned data are correct and complete.\

## Modelling

**Graph Analysis Techniques**

* **Degree Centrality**: Identified the most connected nodes (key transport hubs).
* **Shortest Path Algorithm**: Determined optimal routes between stations.
* **PageRank Algorithm**: Ranked nodes based on their importance in the network.

To ensure data integrity, the graph was analyzed to identify isolated nodes or stations. These nodes were flagged for further investigation to determine their relevance and connectivity in the transport network.

A diagram of a network

Description automatically generated with medium confidence

Figure Graph of Irish Transport

A screenshot of a phone

Description automatically generated

Figure Degree Centrality for Dart

A screenshot of a computer

Description automatically generated

Figure Shortest path

A screenshot of a computer

Description automatically generated

Figure PageRank

## Evaluation

We evaluated and found the average value to understand the travel time based on our models and what the real travel time is.

## Deployment

We deployed the graph project using PyCharm and Neo4j on Streamlit. Streamlit is an open-source platform that helped us visualize the data and deploy it while running. It also allowed us to run the models while they were running.