COMP2200 Data Science

Project Proposal

Sydney Public Transport Analysis

Project Goal:

One key objective we would like to explore with our chosen dataframe is to take advantage of the various sources to reach an optimal conclusion that will help with either decision making/support.

We are eager to provide an analysis that will enable us to correlate trends/patterns with real world activities. From this, we aim to provide a sustainable solution that in hope answers the distinction between an array of transportation rush hours, off-peak hours and outliers. We will also aim to challenge ourselves

Data Source and Background

The dataset is from the Bureau of Transport Statistics. This dataset contains NSW trains official train utilisation figures for Intercity train lines only. These figures include the data of passengers who on/tap off when entering and exiting the transportation service stations. It is set out as a line and aggregated into a monthly figure for a passenger for their estimated times of travel during that month.

The data was collected from opal train trips by month,line and card type, from July 2016 to August 2021. The format of the data set is in CSV. Here is a link to the relevant document page: https://opendata.transport.nsw.gov.au/dataset/opal-trips-train.

The Techniques expect to use in this project

Techniques that are expected to be used though the project are logistic regression and linear regression. These techniques can help apply a better understanding of the relationships between trainline, train card type and amount of uses per month.

Logistic regression will be used in understanding and predicting the probability in volume of people a trainline is expected to have. The technique of linear regression will help display the connection between the train card type and the frequency of its use throughout a set time period.

The techniques mentioned may change though the course of the project however for now are expected to be in use.

Project Plan

As we are analysing data of different transportation in Sydney our first milestone will be to observe the data on several basis. We will look into heat maps during peak times and non peak times, more or less used stations etc.

Our second analysis will be on analysing the data based on logistic regression and linear regression of the different observations we have done. Finally we can make plotting to support our analysis.

Week 9 complete data analysis/data cleaning
Week 10 Complete required tasks
Week 11 Run test and Complete the project