**COMP2200 Data Science**

**Group Project Proposal**

**Data Analysis: The Usage of Various Transports within NSW, and The Relationship between Tap-on and Tap-off of Opal Cards**

*Group 68:* Hai Hoang Nguyen, Dong Ha Nguyen Luu, Xuyen Linh Huynh, Xuan Dung Vuong

**Project Goal**

­

Our goal for this project is to analyse different sections of the transport usage in NSW. We will build the model to see if there is a relationship between the tap-on and tap-off rates of Opal uses within four suburbs in NSW (Parramatta, Strathfield, Chatswood and Macquarie Park). In addition, we will compare the total uses of different transport modes in NSW, for example - the uses of trains and buses monthly, and analyse the relationship between transport uses and demography (such as age group or income).

Outline of the project:

1. Examine and explain the different patterns of transport usage including train, bus, light rail and ferry in specific parts and regions of NSW.
2. Identify if there is any relationship between the age and income of citizens and their preferred modes of public transport to travel to work. Additionally, predict the volume of use of public transport based on the age and income of citizens.
3. Perform an analysis of the number of tap-ons and tap-offs in three weeks in four different areas in NSW.

**Data Source and Background**

The datasets we will be using are data scraped from Transport for NSW and Australian Bureau of Statistics. The data from Transport for NSW focuses on analysing the amount of tap on and tap off each day during a week for each suburb, it is calculated hour by hour for each public transport. We can use this data to analyse the volume of use of public transport in each suburb.

The second data we will be using are income data from Census Table Builder in 2016. The data analyse the amount of income that each person earns per week. The data ranges from nil income (people who dont earn any income) to over $3,000 per week. The income has a slight impact on the usage of public transport. From this data, we can analyse and predict the relationship between public transport and income.

**What work to do**

* Data cleaning:

**Format:** We need to rearrange the data format including the referenced columns and rows to a better format to be read in the notebook environment.

**Missing values:** We clean the value of NaN in the column “Metro” and in the “Transport mode” dataset to 0 as there is no metro service until the middle of 2019.

* Data manipulation:

We need to calculate the average count of usages for every investigated transport type in a period of half a year. Areas to examine include:

* Bus, ferry, light rail and train as types of transport.
* Quantitative data of usage counts per transport type monthly.
* Grand total count of all types by month.

**Technique**

We use the linear regression model to predict the tap-on events at one specific suburb (for the stops of all types of transports) based on the tap-off events from one nearby suburb. We also expect to use the clustering technique in the project, and use logistic regression to classify binary objects (groups) then build the regression model.

**Project Plan**

Week 7: Decide project idea, collect necessary data sets and start with one analysis technique

Break period: Complete and submit the group project proposal. Complete one model building.

Week 8-10: Construct and complete model building.

Week 11: Evaluate all models.

Week 12: Prepare and record presentation.

Week 13: Review and submit the group project.

Datasets:

<https://opendata.transport.nsw.gov.au/dataset/opal-trips-all-modes>

<https://opendata.transport.nsw.gov.au/dataset/opal-tap-on-and-tap-off-release-3-2020>

<https://www.abs.gov.au/statistics/people/population/regional-population-age-and-sex/latest-release> (NSW age groups)

<https://www.transport.nsw.gov.au/data-and-research/travel-insights> (Travel Insights - Tap on stats)