**Project 2: Proposal Outline**

## Group Members

1. Catherine Tratnik
2. Manisha Lal
3. Myfanwy Brown-Robinson
4. Rubalpreet Bhullar
5. Sharanvika Jegatheeswaran

**Topic**: Bikeshare program

**Note**: We have decided to utilise our dataset from project 1 and elaborate on it

**Outline**: <https://docs.google.com/document/d/1-mMdbSi_jF9I5rVU-d-zTT7ojQ4HlXlhYWM44jIPrI0/edit?usp=sharing>

**Table breakdown:** <https://docs.google.com/spreadsheets/d/1SI5cIHukRNwWpbsVuDQRwRXds-chaqYyod6EBZw1QFM/edit?usp=sharing>

Extraction:

## Data sources we will be working with (identify the data source):

* Toronto, Canada 2019 - <https://ckan0.cf.opendata.inter.prod-toronto.ca/download_resource/e69c5761-e0ed-4999-9426-8174f848b11a>
* Vancouver, Canada 2019 - <https://www.mobibikes.ca/en/system-data> OR <https://bikeshare-research.org/#bssid:vancouver> (station info)
* Antwerp, Belgium - <https://bikeshare-research.org/#bssid:antwerp> OR <https://www.velo-antwerpen.be/en>
* Sao Paolo, Brazil 2018 - <https://www.kaggle.com/joseguilhermelopes/bike-sharing-system-in-brasilia-brazil?select=df_rides.csv>

**Note**: Amsterdam data is an API & we may require data to be scraped for data from EU

Transform:

## Types of Transformation

* Cleaning
  + Identify which variables are relevant?
  + Created wanted Dataframe
* Joining
  + Join files for each city as needed
* Filtering
  + Drop irrelevant/NaN columns
* Aggregating

Load:

## Required Fields

* Final database
  + Load data into relational or non-relational
* Tables/collections
* Reason – build on project 1 to encounter data from EU as comparison

## Types of tables required for DB

* A membership table that contains the columns: membership type and membership id
* A Bike Station table that contains the columns: bike id and location
* A Bike ID table that contains the columns: id, station name, city, longitude, & latitude
* A Ridership table that contains the columns: column trip id, start station id, end station id, start date, end date, and membership id
* A Location table that contains the columns: id, city, and country

**Due Date**: December 17, 2020