MS1: Project Proposal

1. Serena Jiao, Aydan Gooneratne, Constance Wang, Henry Chen
    a. Emails: sjiao@seas.upenn.edu, aydang@seas.upenn.edu, conswang@seas.upenn.edu, henchen@seas.upenn.edu
    b. GitHub usernames: serenajiao, aydangoon, conswang, hen-chen
2. Description
    a. We plan to make a website application that helps people plan their trips to NYC, specifically finding restaurants to go to based on their living location and historical traffic data. For example, after I enter where I live, the application will help me find restaurants near me to eat at given the expected traffic at that specific weekday and time. This application will help tourists save time.
3. Dataset Info
    a. AirBnb dataset:
        i. This dataset holds information about airbnb listings such as location, host data, . Ideally, we would combine this with our restaurant dataset to see the closet restaurants near a particular Airbnb location.
        ii. http://insideairbnb.com/get-the-data.html
        iii. 75472 rows for December 2021 data, NYC specific
        If this is not enough, we can augment with different cities' data OR the calendars/reviews/neighbourhood data sets
        Fields include:
        id,listing_url,scrape_id,last_scraped,name,description,neighborhood_overview,picture_url,host_id,host_url,host_name,host_since,host_location,host_about,host_response_time,host_response_rate,host_acceptance_rate,host_is_superhost,host_thumbnail_url,host_picture_url,host_neighbourhood,host_listings_count,host_total_listings_count,host_verifications,host_has_profile_pic,host_identity_verified,neighbourhood,neighbourhood_cleansed… many more fields
        iv. Summary statistics
            1. Average price of listing = $170.86 USD
            2. Average reviews per month = 1.721
    b. Restaurant dataset:
        i. https://data.cityofnewyork.us/Health/Restaurants-rolled-up-/59dk-tdhz
        ii. This dataset contains information about restaurants in New York. Location and name of restaurant are the more important attributes. We will combine this with the Airbnb dataset
        iii. Size statistics:
            1. Number of rows = 28,738
            2. Number of attributes = 6
            3. 1.7 MB
        iv. Summary statistics:
            1. Can find the most popular restaurants by sorting by the number of restaurants given a particular zipcode

2. Eg. 28 restaurants for zipcode 10462
3. Median Neighborhood = Bronx

c. Average traffic in a neighborhood on any given day of the year
   i. https://data.ny.gov/Transportation/Annual-Average-Daily-Traffic-AADT-Beginning-1977/6amx-2pbvhttps://data.cityofnewyork.us/widgets/59dk-tdhz?mobile_redirec
   ii. This dataset contains information about average daily traffic in New York based on roads and districts.
   iii. There are 2.59 M rows and 17 columns, with attributes including ones like Year, State Route, and Road Name.
   iv. Summary Statistics:
      1. Average Traffic Value = 40,153
      2. Median Neighborhood = Bronx

4. Queries

```
// See how long an airbnb locations are in a neighborhood
SELECT neighborhood, COUNT (*) AS num_locations
FROM Airbnb_Locations l
GROUP BY l.neighborhood

// Select all restaurants in a certain neighborhood around an airbnb location
SELECT *
FROM NATURAL JOIN Restaraunts r ON Airbnb_Locations l
WHERE r.neighborhood = l.neighborhood

// See how much average traffic there is around an airbnb location
SELECT AVG(avg_traffic_value) AS avg_traffic, day
FROM Airbnb_Locations l, Avg_Traffic t
WHERE l.address LIKE '%t.road%'
GROUP BY l.id

// Select all restaurants given a particular zipcode
SELECT *
FROM Restaurants r
WHERE r.zipcode = (Given zipcode)

// get all restaurants along an airbnb's road that have traffic below a certain amount
SELECT *
FROM Restaurants r, Airbnb_Locations l, Avg_Traffic t
WHERE l.address LIKE '%t.road%' AND r.address LIKE '%t.road'% AND t.avg_traffic_value < 30,000
```