Proposal

- The problem I want to solve:

Predict the number of visitors in each CBG/neighborhood each day and each hour.

> Underlying questions: What are the most popular brands in a neighborhood? Are there regional preferences for some brands over others? What times do people visit certain census block groups (ex. Manhattan during the day vs. night)

- My potential client and why he cares about this problem:

Restaurants, shops/store, companies, mayor of a city could be interested to know which area attracts the most people to make decisions such as:

- where to implement a new restaurant, shop, … strategically (according to the brands also present in the area for example)?
- what type of business to implement in which area (depending on the amount of people during the day)?
- when to organize an event (depending on the amount of people per day): beginning, middle of week, or weekend?
- when and where do we need for public transportation (for a mayor, transportation company)?

- The data I will be using: How will you acquire the data?

A Census Block Group (CBG) is the most granular level the US Census Bureau reports data on, and covers ~1500 households.

SafeGraph derived the popularity of a CBG or distances traveled to a CBG by analyzing a large panel of GPS movement data.

The authors also combined this GPS data with our dataset of 5 million building footprints for Points-of-Interest in the U.S. (SafeGraph Places) to determine visits to places like stores or restaurants. They used these visit counts to derive consumer insights such as top brands (ex. McDonald's) visited in a CBG.

The data I will use is already put together in this way:

https://www.kaggle.com/safegraph/visit-patterns-by-census-block-group

- How I will solve this problem:

- My deliverables:

Typically, this includes code, a paper, or a slide deck.

Column description:

Column 1: census_block_group
The unique 12-digit FIPS code for the Census Block Group. Please note that some CBGs have leading zeroes.

Column 2: date_range_start (will be deleted)
Start time for measurement period as a timestamp in UTC seconds.

Column 3: date_range_end (will be deleted)
End time for measurement period as a timestamp in UTC seconds.

Column 4: raw_visit_count
Number of visits seen by our panel to this CBG during the date range.

Column 5: raw_visitor_count
Number of unique visitors seen by our panel to this POI during the date range.

Column 6: visitor_home_cbgs
This column lists all the origin home CBGs for devices that visited a destination in the CBG listed in the column census_block_group (the destination CBG). The number mapped to each home CBG indicates the number of visitors observed from this home CBG that visited census_block_group during this time period. Home CBGs with less than 50 visitors to census_block_group are not included.

Column 7: visitor_work_cbgs
This column lists all the work-location CBGs for devices that visited a destination in the CBG listed in the column census_block_group (the destination CBG). The number mapped to each work CBG indicates the number of visitors observed with this work CBG that visited census_block_group during this time period. Work CBGs with less than 50 visitors to census_block_group are not included.

Column 8: distance_from_home
Median distance from home traveled to CBG by visitors (of visitors whose home we have identified) in meters.

Column 9: related_same_day_brand
Brands that the visitors to this CBG visited on the same day as their visit to the CBG where customer overlap differs by at least 5% from the SafeGraph national average to these brands. Order by strength of difference and limited to top ten brands.

Column 10: related_same_month_brand
Brands that the visitors to this CBG visited on the same month as their visit to the CBG where customer overlap differs by at least 5% from the SafeGraph national average. Order by strength of difference and limited to top ten brands.

Column 11: top_brands
A list of the top brands visited in the CBG during the time period. Limited to top 10 brands.

Column 12: popularity_by_hour
A mapping of hour of the day to the number of visits in each hour over the course of the date range in local time.

Column 13: popularity_by_day
A mapping of day of week to the number of visits on each day (local time) in the course of the date range.