Matt Stice

Project Plan Bikeshare

**Objective**:

Based on rider patterns, discover any stations that become excessively popular destination stations. Contrast those statistics with the most popular trip start locations. Recommend a program for redistributing the bike inventory to balance the supply with demand. Consider the factors surrounding the identified stations. Does that research provide insight on the variables involved? Might also consider distribution by time (more popular in morning/afternoon/evening?), seasonality (is college in session?) and proximity of stations to each other for redistribution.

**Data:**

Datasets used-

bluebikes\_2016, bluebikes\_2017, bluebikes\_2018, bluebikes\_2019: trip information, transactional

Information contained:

Specific aspects of each ride taken- which bike, trip start time, trip end time, trip starting station, trip ending station

-Could use total start station entries for a station id and subtract total end station entries for a station id in order to find which stations are more popular start vs end destinations

-Could further segment this by time of day (morning/afternoon/evening) and month of year (for seasonality) to help determine availability/redistribution

Basic customer info- not of real use for objective

Columns: bike #, start time, end time, start station, end station, user type (subscriber vs customer), user birth year (if given, “/N” if not), and user gender (as integer 1, 2 or 0- not known which is assigned to which)

Useful data for objective: start time and end time (check for daily and seasonal patterns of station use), start station, end station

Primary key: bike\_id

Dataset sizes: from approx. 1.25 million to 2.5 million trips/year timeframe

bluebikes\_stations: station information, referential

Information contained:

Information on stations- address, absolute location, neighborhood (easy way to know which stations are close to each other), # of docks, station ID

-Could use this to determine station proximity for possible redistribution

-Could also compare dock count to differences between start/end location to determine if redistribution was needed

Columns: [station] number (varchar of assigned station number), [station] name (address), latitude, longitude, district (neighborhood), public (yes/no if station is public- all are), total docks (number of bike docks at station), [station] id

Primary Keys(?): number AND id

Dataset size: 189 stations

Join key between tables: bluebikes\_[year]- start\_station\_id AND end\_station\_id, buebikes\_stations- id

-Would need to compare differences between start/end trip station to station location in order to efficiently redistribute.

Data dictionary:

-blue\_bikes\_2016, blue\_bikes\_2017, blue\_bikes\_2018, blue\_bikes\_2019: tables containing trip data

-bike\_id: the ID number of the bike used

-start\_time: the time the trip started

-end\_time: the time the trip ended

-start\_station\_id: the ID number of the station where the trip started

-end\_station\_id: the ID number of the station where the trip ended

-user\_type: whether the trip was made by a subscribed member or non-subscribed customer

-user\_birth\_year: the user’s birth year

-user\_gender: either a 0, 1, or 2 which means either male, female or unknown, but not stated which number corresponds to which.

-bluebikes\_stations: table with parking station information.

-number: unknown, but not the same as station\_id

-name: station name, which is also location

-latitude: station latitude

-longitude: station longitude

-district: specific city/town within Boston metropolitan area where station is

-public: is the station public (all listed are)

-total\_docks: how many docks a station has/how many bikes it can fit

-id: station\_id

**Analysis**:

KPIs-

Start/end balances: Differences between trip start and end counts for stations

-Further segmented by time of day and month of year for daily/yearly peaks and troughs

-Compare to dock counts to see if balances are accommodated by station volume

-Also compare to time of day to see if distribution is needed or if morning/afternoon/evening balances will be negated

Proximity of stations with large positive vs negative start/end balances to determine efficient redistribution

Questions-

Which stations have the largest imbalances between starts and ends from 6am to 11am?

Which stations have the largest imbalances between starts and ends from 11am to 3pm?

Which stations have the largest imbalances between starts and ends from 3pm to 8pm?

How do the imbalances compare to the # of docks for the station? (Is redistribution needed?)

Does the station’s morning/afternoon/evening imbalance get corrected for by the next period?

Which stations have the largest discrepancies from one month to another? Which months would need redistribution or not?

Which districts have stations with both large positive and large negative imbalances that could easily be redistributed to one another?

Secondary research: do seasonal/hourly imbalances make sense for location? (if a station has a large imbalance Sept-May is it near a college? Do stations near tourist attractions have seasonal imbalances? Do stations in business districts get more trips ending there in the morning when people come to work and less in the evening when people leave work?)

How do I ask these questions-

Which stations have the largest morning/afternoon/evening imbalance and largest seasonal imbalance? How does that relate to their dock count? Where are they?

Needed functions- where date\_part(month, start\_time), where date\_part(month, start\_time) , join, sum(start\_station\_id) - sum(end\_station\_id), group\_by(start\_station\_id)

Needed calculations w/o knowing how- subtract total end\_station entries for an individual station from its total start\_station entries during a certain time range (like 7am-11am) or date range (like June-August) in order to get the imbalance for the station.