Revised Proposal

Group A - Felix Stetsenko, Enoch Shin

October 31, 2019

(handwritten annotations: "① This is coming together nicely. ② I'm happy to approve the proposal ③ Can you pick one or two? I'm more interested in 1 and/or 2.")

1. **Group:** A

2. **Group Members:** Enoch Shin and Felix Stetsenko

3. **Title:** *Visualizing Transportation Network Provider and Taxi trips in Chicago*

4. **Purpose:** Our initial idea is to visualize travel patterns across the City of Chicago using Transportation Network Provider (TNPs referring to rideshare companies like Uber and Lyft) and Taxi trip data publicly available on Chicago's Data Portal website. We would visualize the origins and destinations (on the census tract level) of all taxi and TNP trips across the city from November 2018 to present. We could then use the data visualization for a number of possible analyses:

   (handwritten: #1)
   - Do TNP and taxi services serve the same market, or is one service preferred over the other for certain travel markets (e.g. trips originating from O'Hare Airport)?

   (handwritten: #2)
   - Can residential segregation be visualized through TNP and taxi service and trip patterns? For instance, is there a relationship between census tract demographics and TNP vs. taxi share of trips originating from that census tract?

   (handwritten: #3)
   - Using a clustering algorithm on the origin / destination data, how closely would the clusters reflect neighborhood boundaries and racial and socio-economic divides?

5. Data Overview

Data: The relevant data is all publicly available, either from the City of Chicago or the U.S. Census Bureau. We would create a visualization by using an overlay on top of the Google Maps API. We would consider adding the functionality to have the data continuously update as the City of Chicago posts updates on its Data Portal.

(handwritten: ④ did my example from yesterday help answer your questions regarding geolocation + merging?)

6. Variables

Transportation Network Provider Data

A TNP company is defined by the City of Chicago as a "ride share company, which provides prearranged transportation services for compensation through an Internet-enabled application or digital platform connecting passengers with drivers of vehicles for hire. TNP drivers and their vehicles join and become affiliated with TNP companies and are then available to be dispatched through the TNP's digital platform. Each TNP company must be licensed. The TNP license is an annual license which is not transferable."

Link: City of Chicago TNP data (available from November 2018 to present): https://data.cityofchicago.org/Transportation/Transportation-Network-Providers-Trips/m6dm-c72p

Codebook: Chicago TNP data (available from November 2018 to present)

Trip ID

A unique identifier for the trip

Trip Start Timestamp

When the trip started, rounded to the nearest 15 minutes (floating timestamp).

Trip End Timestamp

When the trip ended, rounded to the nearest 15 minutes (floating timestamp).

Trip Seconds

Time of the trip in seconds.

Trip Miles

Distance of the trip in miles.

Pickup Census Tract

The Census Tract where the trip began.

Dropoff Census Tract

The Census Tract where the trip ended.

Pickup Community Area

The Community Area where the trip began.

Dropoff Community Area

The Community Area where the trip ended.

Fare

The fare for the trip, rounded to the nearest $2.50.

Tip

The tip for the trip, rounded to the nearest $1.00.

Additional Charges

The taxes, fees, and any other charges for the trip (dollars).

Trip Total

Total cost of the trip (dollars).

Shared Trip Authorized

Whether the customer agreed to a shared trip with another passenger.

Trips Pooled

If customers were matched for a shared trip, how many trips, including this one, were pooled.

Pickup Centroid Latitude

The latitude of the center of the pickup census tract.

Pickup Centroid Longitude

The longitude of the center of the pickup census tract.

Pickup Centroid Location

The location of the center of the pickup census tract.

Dropoff Centroid Latitude

The latitude of the center of the dropoff census tract.

Dropoff Centroid Longitude

The longitude of the center of the dropoff census tract.

Dropoff Centroid Location

The location of the center of the dropoff census tract.

Taxi Data

Link: City of Chicago Taxi data (available from 2013 to present): https://data.cityofchicago.org/Transportation/Taxi-Trips/wrvz-psew

City of Chicago Taxi data (available from 2013 to present)

Trip ID

A unique identifier for the trip.

Taxi ID

A unique identifier for the taxi.

Trip Start Timestamp

When the trip started, rounded to the nearest 15 minutes.

Trip End Timestamp

When the trip ended, rounded to the nearest 15 minutes.

Trip Seconds

Time of the trip in seconds.

Trip Miles

Distance of the trip in miles.

Pickup Census Tract

The Census Tract where the trip began.

Dropoff Census Tract

The Census Tract where the trip ended.

Pickup Community Area

The Community Area where the trip began.

Dropoff Community Area

The Community Area where the trip ended.

Fare

The fare for the trip.

Demographic Data from Census Bureau and National Historical GIS (Work in Progress)

Link: 2017 American Community Survey 5-year Estimates Subject Tables (available at the census tract level): https://data.census.gov/cedsci/map?table=B00001&tid=ACSDT5Y2017.B00001& hidePreview=false&vintage=2017&layer=censustract&cid=B00001_001E&lastDisplayedRow= 34

Link 2: https://www.nhgis.org/ for shapefiles with race

We'll not include a detailed codebook since we're not sure if it'll be plausible to use this data, for the following reasons below:

[handwritten: 6]

[handwritten: Can you please confirm?]

The Census Bureau data appears to lack geographic information about race distribution (i.e. the data tells us summaries of counts of racial groups, but it's unclear if the Bureau releases the general locations of these demographics). The goal of the project is to see if there are any patterns in rideshare usage that are associated with racial distributions, so we're interested in demographic data.

The National Historical GIS data is more inline with what we're looking for, but we cannot query this data with an API: we have to go through the site and put in a request for the data through their portal, and then they send a notification via email that the data is ready to download. Turnaround only takes a couple minutes, though. However, the University of Virginia has already detailed their process for mapping distributions of demographic groups (link here), and it seems like an entire project of its own. I presume we can find a way of adapting the University's code to our project, but I also think we should put limits on the scope of this project since we're already working with large quantities of data with the TNP data set.

7. End Product

[handwritten: less critical in my opinion]

[handwritten: more relevant]

The final deliverable will be an interactive, animated visualization overlaid over the Google Maps API; it will show Taxi and TNP trips for a few selected days from the data set (selected from November 2018 to present).* The user will have the option of selecting specific time frames, origin / destination census tracts, and adding additional layers to the map (e.g. public transit lines, demographic and socio-economic variables).

*At the moment, we are focusing on a specific holiday (Valentine's Day: February 14, 2019), and two specific days (Saturday, February 2, 2019; Monday, February 4, 2018) for analysis. We may have to adjust the scope (i.e. make it on the order of several hours instead of a day). However, the actual querying with the API is mostly sorted out.

[handwritten: great!]

Sample of the Data

[handwritten: ✓]

Below, we've just done a simple query of the TNP data with 3,000 set as the limit. No other particular filters or constraints were done for the query. However, see our `chicagoIngest.Rmd` file to view our progress on navigating the API.

```
#quick conversion to numeric for some variables of interest
toNum <- c("pickup_centroid_longitude", "pickup_centroid_latitude")
chicago_df <- readRDS("ChicagoCommute/chicago_df.Rda")
chicago_df %<>% mutate_at(.funs=as.numeric, toNum)
#basic plot using open source maps -- will use Google API later
```

```
qmplot(pickup_centroid_longitude, pickup_centroid_latitude,
      data = chicago_df,
      maptype = "toner-lite",
      color = I("red"))
```
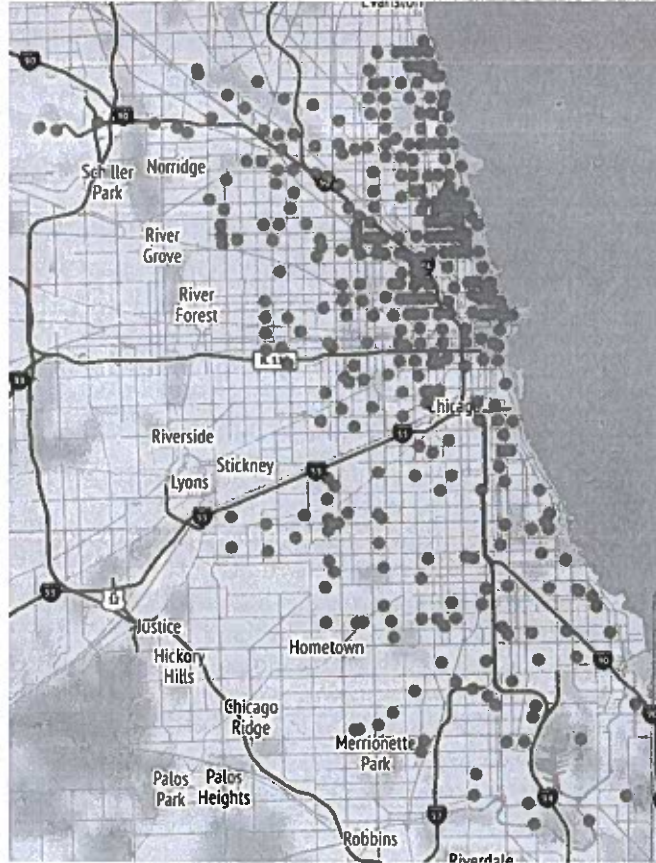

This shows us the pickup locations for the rideshare records on a selection of 3,000 queried observations.

Table of Head of Data

Table 1: Table continues below

trip_id	trip_start_timestamp
022ad3b7f1320d4e52cce3d5931eb0a0cee16c48	2018-11-01 08:45:00
03a2ac30a46af881e6a2e6af06a3a779b67c0802	2018-11-01 20:15:00
040590c0bf5b22f8ccf7d8f19873c612bebfd480	2018-11-01 17:00:00
0483215f5d0fa7c8598ce17f34865e0e55075292	2018-11-01 21:30:00
054ad756d76b71bce5741b527d2474847e926365	2018-11-01 21:30:00
06507e974eb09e9aaf8a64ef4a45bae1627ec4c7	2018-11-01 16:45:00

Table 2: Table continues below

trip_end_timestamp	trip_seconds	trip_miles	dropoff_census_tract
2018-11-01 09:30:00	3140	38.5419281243957	17031980100
2018-11-01 21:15:00	4125	49.8706525619584	NA
2018-11-01 18:00:00	3383	12.2729895391704	17031839100
2018-11-01 22:15:00	2142	24.1096309527866	NA
2018-11-01 22:15:00	2052	21.6391838824985	17031281900
2018-11-01 18:00:00	4395	21.7475684185008	17031031700

Table 3: Table continues below

dropoff_community_area	fare	tip	additional_charges	trip_total
56	47.5	10	9.1	66.6
NA	62.5	0	3.8	66.3
32	27.5	0	2.5	30
NA	27.5	5	2.5	35
28	25	2	2.5	29.5
3	40	0	2.5	42.5

Table 4: Table continues below

shared_trip_authorized	trips_pooled	dropoff_centroid_latitude
FALSE	1	41.785998518
FALSE	1	NA
FALSE	1	41.8809944707
TRUE	1	NA
TRUE	1	41.8792550844
FALSE	1	41.9654457842

Table 5: Table continues below

dropoff_centroid_longitude	dropoff_centroid_location.type
-87.7509342894	Point
NA	NA
-87.6327464887	Point
NA	NA
-87.642648998	Point
-87.6631958495	Point

Table 6: Table continues below

dropoff_centroid_location.coordinates	pickup_census_tract
-87.75	NA
41.79	17031081700
-87.75	NA
41.79	17031081600
-87.75	NA

dropoff_centroid_location.coordinates	pickup_census_tract
41.79	NA

Table 7: Table continues below

pickup_community_area	pickup_centroid_latitude	pickup_centroid_longitude
NA	NA	NA
8	41.89	-87.63
NA	NA	NA
8	41.89	-87.63
NA	NA	NA
NA	NA	NA

pickup_centroid_location.type	pickup_centroid_location.coordinates
NA	NULL
Point	NULL
NA	NULL
Point	NULL
NA	NULL
NA	NULL