Revised Proposal

Group A - Felix Stetsenko, Enoch Shin

November 12, 2019

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:
    - 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)?
    - 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?
    - 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.

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:

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

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.

Sample of the Data

Below, we've just done a simple query of the TNP data with data from Valentine's Day 2019.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.

Ingest (Query)

City of Chicago data uses the Socrata API.

```
url <- "https://data.cityofchicago.org/resource/m6dm-c72p.json"
# names <- readRDS("ChicagoCommute/names.Rda")
mytoken <- "bJZzO3D3YpmgVEszO9oq241gH"
```

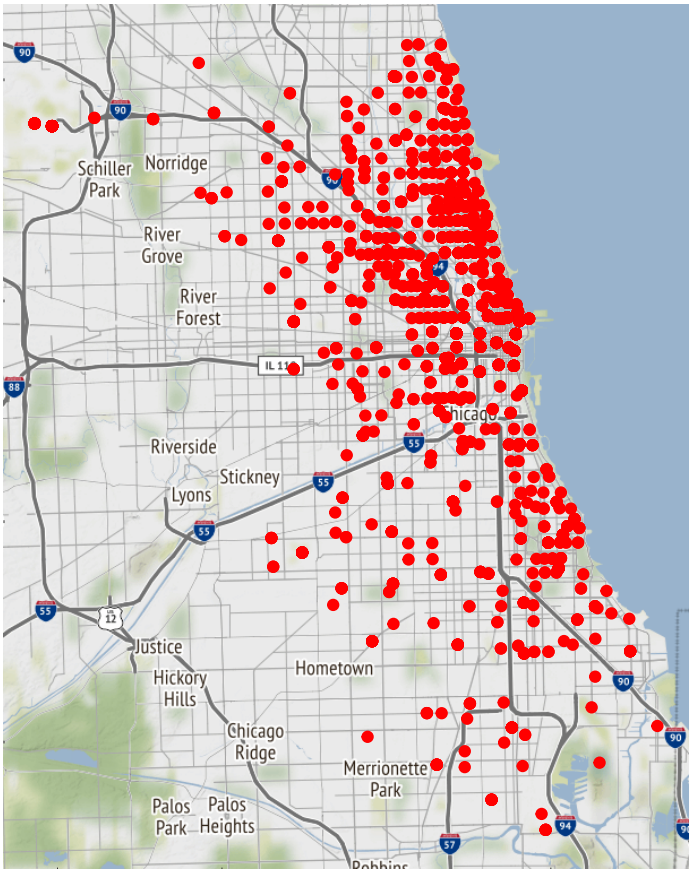
```
# valday <- read.socrata(paste0(url, "?","$where=trip_start_timestamp between '2019-02-14T00:00:00' and
# saveRDS(valday, "ChicagoCommute/RDA/valday.Rda")
valday <- readRDS("ChicagoCommute/RDA/valday.Rda")
```

Wrangle

```
# toCoord <- grep(".coordinates$", names(valday), value = TRUE)
# val_coord <- valday %>%
              # rowwise() %>%
              # mutate_at(.funs=function(x){paste(unlist(x), collapse=",")},
                          # toCoord)
toNum <- c("trip_seconds", "trip_miles", "fare", "tip",
          "additional_charges", "trip_total")

coordNum <- grep("_centroid_l[[:alpha:]]*e$", names(valday), value=TRUE)

valday <- valday %>%
            mutate_at(.funs=as.numeric, c(toNum, coordNum))
```

We need to do a sample of the Valentine's Day dataset since ~315k data points take prohibitively long to plot.

```
#quick conversion to numeric for some variables of interest
subval <- mosaic::sample(valday, size=2500)
toNum <- c("pickup_centroid_longitude", "pickup_centroid_latitude",
           "dropoff_centroid_longitude", "dropoff_centroid_latitude")
subval %<>% mutate_at(.funs=as.numeric, toNum)

qmplot(pickup_centroid_longitude, pickup_centroid_latitude,
       data = subval,
       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

Quitting from lines 266-267 (RevisedProposal.Rmd) Error in head(chicago_df) : object 'chicago_df' not found Calls: ... withCallingHandlers -> withVisible -> eval -> eval -> -> head In addition: Warning messages: 1: In knitr::knit(knit_input, knit_output, envir = envir, quiet = quiet, : The file "RevisedProposal.Rmd" must be encoded in UTF-8. Please see https://yihui.name/en/2018/11/biggest-regret-knitr/ for more info. 2: Removed 147 rows containing missing values (geom_point).