PPPA 6007: Microeconomics for Public Policy I
Fall 2020

Use Numbers: Assignment 2 of 3
Determinants of Demand

Due October 27, 2020

For this assignment, we are going to look at data that helps us focus on demand. Generally, it is difficult to empirically separate the influence of demand from supply because what we observe in the world is the equilibrium. Equilibrium is a function of both demand and supply. Due to this, we are going to focus on a market where (over a short period) supply is relatively fixed: the New York City subway.[1]

The idea here is that when we compare two of the same days (e.g., two Wednesdays) at the same time, differences in transit ridership should be due to demand, rather than supply factors.

The data I have prepared allow you to make the following comparisons

1. ridership for the entire system by day
    - csv format: ridership_by_day_20200923.csv
    - rds format: ridership_by_day_20200923.rds
2. ridership by station and day
    - csv format: ridership_by_station_20200924.csv
    - rds format: ridership_by_station_20200924.rds

These data come from NY's Metropolitan Transit Authority (MTA) and I have pre-processed them for your ease of use.

You can find a map of the NYC subway system here. This map shows subway stations with dots. Below the station name, the map notes to which line (e.g., 1/2/3) the station belongs. For example, the 79th Street stop in Manhattan's Upper West Side is on the 1 line.

Importantly, there are different points on the map that have the same station name. For example, there are multiple 96th Street stations: one on the 1, 2 or 3; one on the B and C lines; one on the 6 train, and one on the Q train. To differentiate between these 96th Street stations, you need to know which line the station is on. Our dataset has one variable for the station and another variable for the line to help you with this issue.

As with the previous assignment, Google and news archives should be sufficient to answer these questions. This is not a major research paper, so please scale your effort accordingly.

[1] I wanted to do this for the DC Metro, but similar data are not available, to the best of my knowledge.

Also as in the previous assignment, you are welcome to discuss parts of this assignment with other students. However, any work you turn in must be your own and written in your own words.

To make graphs, you can use Excel, R, or the software of your choice. We can support technical questions in Excel or R (Brooks and Bayar only).

1 Questions

1. Assume (and this is likely true) that the amount of transit supplied was about the same in both February and March 2020. Graph subways boardings for the days included in the dataset for these two months. What demand feature explains the difference? How do you see this?

2. Even absent large-scale public health disasters, there is also variation in subway usage across station. Choose three stations and make a graph of pre-pandemic daily ridership data that illustrates these differences (pre-pandemic daily ridership has two separate weeks in September 2019 and February 2020; use whichever week you prefer). Explain whether differences in ridership across stations are due to transit demand, transit supply, or both.

3. Compare ridership at two stations on the 1 train – 238th St. and 86th Street – for the two weeks in September in the dataset. (Note: to choose these stations, you need to select both by station name and by `LINENAME`.)

   (a) Which station has a higher level of ridership? Give two hypotheses for this difference. (Remember that they are on the same line, so supply should not be an explanation.)

   (b) Which station has a larger change in ridership after the pandemic? Which features of demand explain this relative change (that is, bigger at one station than the other)?

4. Use the data provided to make a final graph that shows one more dimension of demand using these data. Write a brief paragraph that discusses how your graph shows this additional element of demand.

2 How to turn it in

Turn this assignment in to the google folder: for_students → use_numbers → assignment_2
Name the assignment "lastname_use_numbers_2". So mine would be "brooks_use_numbers_2."

3 Data

Daily Ridership

The daily ridership dataset for the assignment has the following variables (columns in excel-speak).

Variable	Definition and Source
`DATE`	Date as text in format MM/DD/YYYY
`date_format`	Date in date format
`total_boardings`	Total number of entries into subway
`year`	Total number of entries into subway
`month`	Total number of entries into subway
`day`	Total number of entries into subway

Daily Ridership by Station

The daily ridership dataset for the assignment has the following variables (columns in excel-speak).

Variable	Definition and Source
`DATE`	Date as text in format MM/DD/YYYY
`date_format`	Date in date format
`total_boardings`	Total number of entries into subway
`year`	Year
`month`	Month
`day`	Day of month (number)
`day_of_week`	Day of the week (Monday, Tuesday, . . .)
`STATION`	Name of station (note that some names repeat across lines; you may also need `LINENAME`)
`LINENAME`	Name of subway line (note that some `STATION` names repeat across lines)

Please be sure to carefully read the information about stations and lines in the introduction to the assignment.

4 R Commands

If you're interested, the R file that creates the data for this assignment is available here.

For this assignment, here are some R commands that may be helpful.

R commands	Description
Make total ridership graph	
`# load packages` `library(tidverse)` `# read data` `d3 <- readRDS(ridership_by_day_20200923.rds)` `# make dataset just for feb and march` `d4 <- d3[which(year(d3$date_format) == 2020 &` `month(d3$date_format) %in% c(2,3)),]` `# make graph` `e1 <- ggplot() +` `geom_point(data = d4[which(month(d4$date_format) == 2),],` `mapping = aes(x = date_format, y = total_boardings), color =` `‘‘blue’’) +` `geom_line(data = d4[which(month(d4$date_format) == 2),],` `mapping = aes(x = date_format, y = total_boardings), color =` `‘‘blue’’) +` `geom_point(data = d4[which(month(d4$date_format) == 3),],` `mapping = aes(x = date_format, y = total_boardings), color =` `‘‘red’’) +` `geom_line(data = d4[which(month(d4$date_format) == 3),],` `mapping = aes(x = date_format, y = total_boardings), color =` `‘‘red’’) +` `scale_y_continuous(limits = c(0,5000000), labels =` `scales::comma) +` `labs(x = ‘‘’’, y = ‘‘total daily boardings’’)+` `theme_minimal()`	The first command loads the “tidyverse” packages (if you did this in R last time, you already installed these; if not, see the instructions from the previous assignment). The final command plots the weeks in red or blue. We use a point and a line for each week.

R commands	Description
Plot three stations	
`# read station data` `d2 <- readRDS(‘‘ridership_by_station_20200924.rds’’)` `# just grab three stations` `table(d2$STATION)` `table(d2[which(d2$STATION == ‘‘125 ST’’),]$LINENAME)` `e3 <- d2[which(d2$STATION %in% c(‘‘GRD CNTRL-42 ST’’,‘‘VERNON-JACKSON’’) | d2$STATION == ‘‘125 ST’’ & d2$LINENAME == ‘‘23’’),]` `# just use 2019` `e3 <- e3[which(substr(e3$DATE,start=7,stop=10) == ‘‘2019’’),]` `# make the graph` `e3p <- ggplot(data = e3) +` `geom_point(mapping = aes(x = date_format, y = total_boardings, color = STATION)) +` `geom_line(mapping = aes(x = date_format, y = total_boardings, color = STATION)) +` `scale_y_continuous(labels = scales::comma) +` `labs(x = ‘‘’’, y = ‘‘total daily boardings’’)+` `theme_minimal()`	The first command reads the data. I then create dataframe $e3$ that has only the relevant stops. The next lines of code limit to 2019. Having prepared the dataset, I then make a graph, with one color for each station.

R commands	Description
Plot two stations in two Septembers	
`# read station data` `d2 <- readRDS(‘‘ridership_by_station_20200924.rds’’)` `# limit to two stations` `e4 <- d2[which(d2$STATION %in% c(‘‘238 ST’’,‘‘86 ST’’) &` `d2$LINENAME == ‘‘1’’ & month(d2$date_format) == 9),]` `# you may or may not need this command` `# re-order days of the week` `e4$day.of.week <- factor(e4$day.of.week, levels =` `c(‘‘Monday’’,‘‘Tuesday’’,‘‘Wednesday’’,‘‘Thursday’’,` `‘‘Friday’’,‘‘Saturday’’,‘‘Sunday’’))` `# just use 2019` `e4 <- e4[which(substr(e4$DATE,start=7,stop=10) == ‘‘2019’’),]` `# make color values` `cvals <- c(‘‘#74c476’’,‘‘#006d2c’’,‘‘#9ecae1’’,‘‘#3182bd’’)` `# make a station-year variable for grouping` `e4$grouper <- paste0(e4$STATION,‘‘ in ’’,year(e4$date_format))` `# make the graph` `e3q <- ggplot(data = e4,` `mapping = aes(x = day.of.week, y = total_boardings, color =` `grouper)) +` `geom_point() +` `geom_line(mapping = aes(group = grouper)) +` `scale_y_continuous(labels = scales::comma) +` `scale_color_manual(values = cvals) +` `labs(x = ‘‘day of the week’’, y = ‘‘total daily boardings’’,` `color = ‘‘station-year’’)+` `theme_minimal()`	The first command reads the data. The next commands limit to the two stations of interest, re-order the factor variable day of the week so that the days are ordered appropriately in the graph, limits data to 2019, and makes a vector of colors that I’ll use in the graph. The last command before the graph creates a station-year variable so that we can differentiate the stations’ ridership in the two years. Having prepared the dataset, I then make a graph, with one color for each station-year combo.