Joe Gallagher
GES 486
Final Project Report

Introduction

This project builds on my thesis research about the adoption of new fare collection technology by transit agencies in the U.S. The goal was to create a cartographic product that could illustrate some of the study background – that is, the variegated landscape of transit fare collection in the U.S. But in a larger sense, this project is about getting intimate with data: inspecting, cleaning, joining and representing it.

I started with a simple goal: to map the fare collection modes in use at transit agencies in the U.S. Where are tokens still used? How widespread are smartcards? Is mobile ticketing taking over? The map could illustrate trends I saw in the course of my thesis research. Smartcards, while expensive, helped facilitate interoperability of transit agencies operating in the same metropolitan region. Mobile ticketing applications were cheap and fast to implement by agencies of all sizes. Non-digitized fare payment media were on the way out. These could be mapped alongside key performance measures like annual ridership, annual revenue, and annual operating costs.

The task turned out to be more complex than initially imagined. The final map represents just the tip of the iceberg, below which lie a number of data cleaning and manipulation tasks.

Overall the project can be broken down into two separate problems/solutions, resulting in the final dataset and maps.

> *Problem*: NTD does not collect data on fare collection modes
> *Solution*: Join NTD to APTA - robust info on the FTA on annual revenue and ridership, fine-grained detail on each fare collection mode in place at APTA's member agencies.
> *Problem*: APTA data and NTD data do not share a primary key
> *Solution*: Iterative data cleaning + creation of a recoding table to match data on agency name.

Data and methods

This project used two datasets. The first is the 2022 APTA Public Transportation Fare Database, published by American Public Transportation Association. The APTA Fare Database has information about where fares are sold, what forms of currency are accepted, the equipment used to collect fares, costs of base fares and transfers, inter-agency reciprocity agreements etc.

The second is The National Transit Database and is administered by the Federal Transit Administration. Federally funded transit agencies report monthly figures including ridership (in unlinked passenger trips AKA a measure of people boarding buses and passing through turnstiles), fare revenue, vehicles in service, and operating costs.

Adjacent to the NTD is the National Transit Map: Agencies shapefile. This file features locations headquarters of all participating transit agencies. This is the spatial data I eventually linked the fares data to.

The NTD collects data on revenue from fares collected on each mode within an agency, however it does not require respondents to indicate *how* revenue is collected. It might be interesting to know, and map, fare payment media – how people pay to ride transit – at the busiest transit agencies in the U.S. by unlinked passenger journeys.

All of my methods are a prelude to a very tricky merge. The processing and cleaning begin easily enough: reading in data.

Because each dataset compiles information about all transit modes – including on demand service and vanpool, which handle fares very differently – I filtered out those modes, leaving in the "typical" mix of heavy rail, commuter buses and rail, buses, light rail and so-on. Both datasets use the same mode codes - APTA has adopted the FTA's codes.

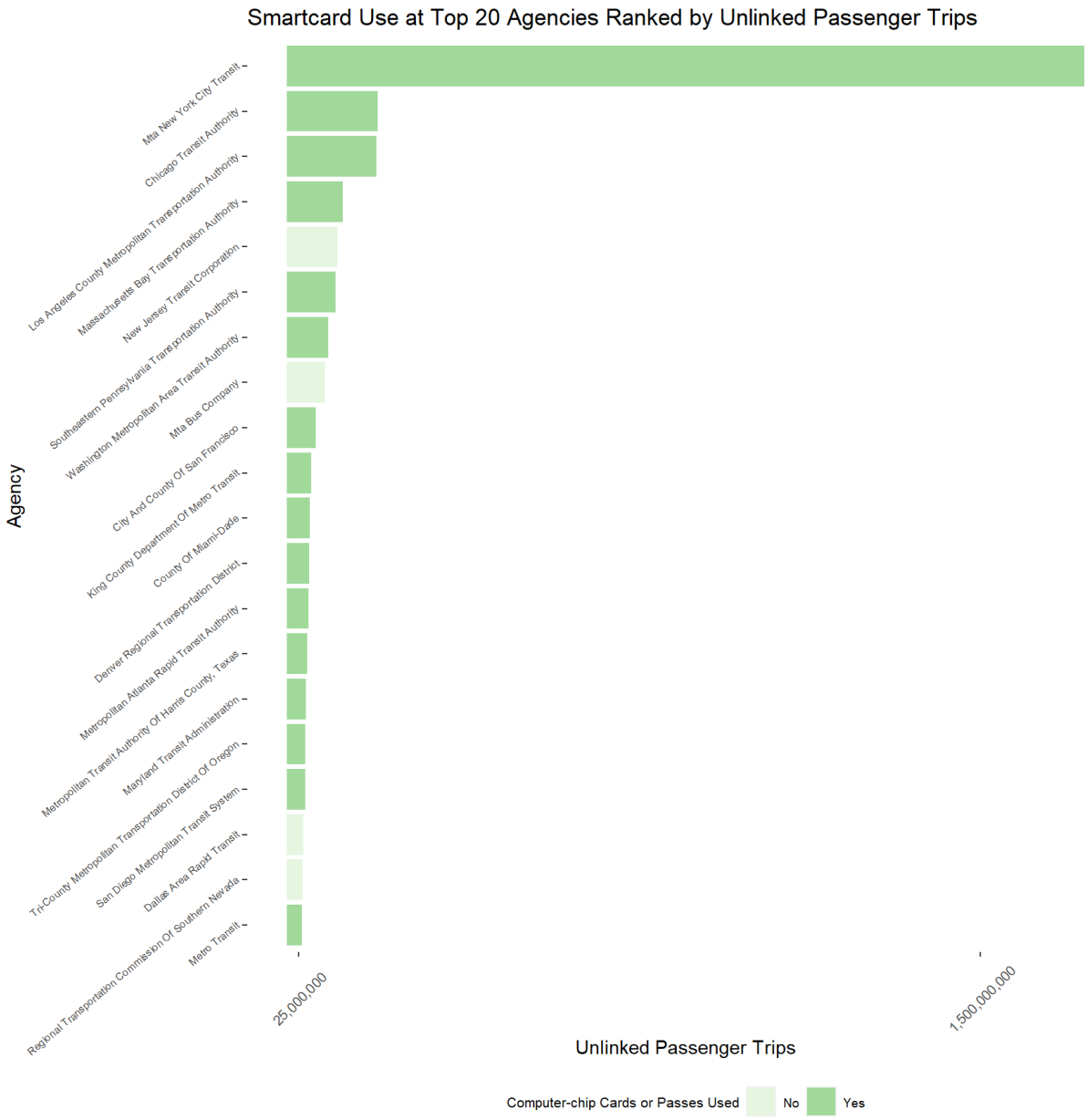
After filtering, I renamed the agency_name variable of each dataset to make them easier to handle. All the figures collected in these datasets are spread across modes; the APTA database has over 500 rows of data, and the NTD has over 2000. The next step was to summarize each table down to one row per agency. I grouped each table by their respective unique agency ID. Then I applied different functions across different variables within the summary function. For qualitative variables, I took the first non-NA value. For totals like revenue and ridership figures, I took the sum of those figures from the various modes. For averages like average costs and average fares, I took the mean of those figures from various modes. After removing inactive, rural, and small-systems reporters from the NTD, and their equivalents from the APTA database, I had simplified the APTA database down to 214 rows and the NTD down to 477 rows.

On my very first attempt at this task in March, I joined the data here using agency_name, and thought that would be that. However, the join only matched about 60 rows. Then, when I plotted the largest transit agencies by unlinked passenger trips, some major systems were missing while other strange systems (such as the Cape Cod Ferries) were present in the top 25. To make a useful map, I would have to figure out how to better-join these two databases that do not share a primary key.

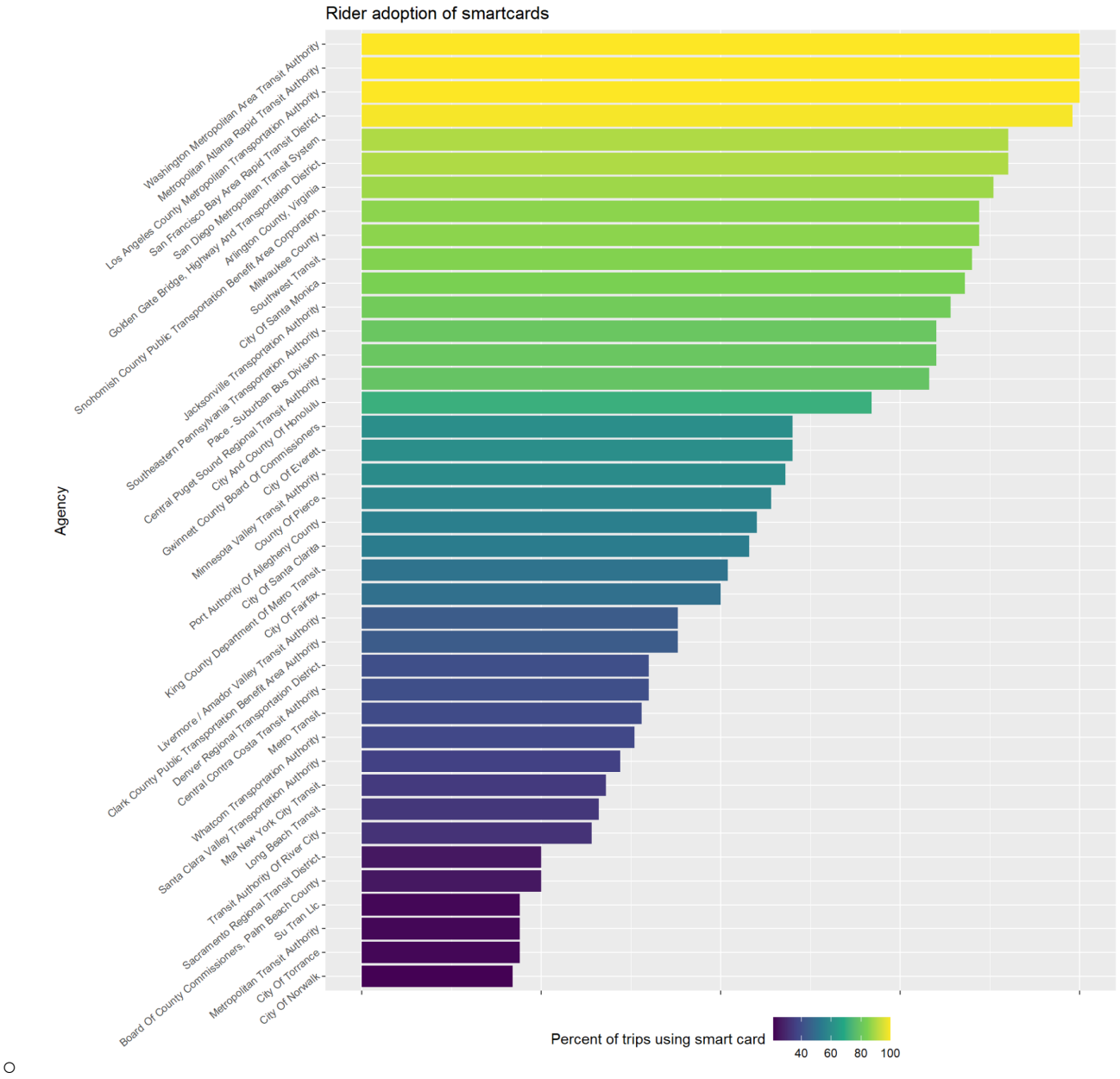
On the next join, I attempted a match on partial strings, thinking that hyphens and agency suffixes might account for the lack of matches. That "caught" a few dozen more rows, but a number of agency names were now duplicated (such as each of the agencies in Connecticut). My next attempt joined the data based on city name. This appeared pretty successful though I was left with a number of duplicate matches since city names are not unique, nor do large cities typically operate just one transit agency. At this point, I compiled, by hand, two character vectors: one was the APTA agency name; the other was its equivalent in the NTD. A total of 107 of the 214 agencies in the APTA table needed to be recoded. I merged the character vectors into a data frame, which I joined to the APTA table using the shared APTA agency columns. Naturally, it matched 109 rows, with the remainder of rows having a new column for equivalent NTD agency name populated with NAs. I used mutate with if_else to re-populate the APTA agency column with the value from the NTD agency name column for all rows that did not have NA in the NTD agency column. The recode was a success: when joining the recoded APTA table to the original NTD table, it matched 214 rows. 1:1.

Results

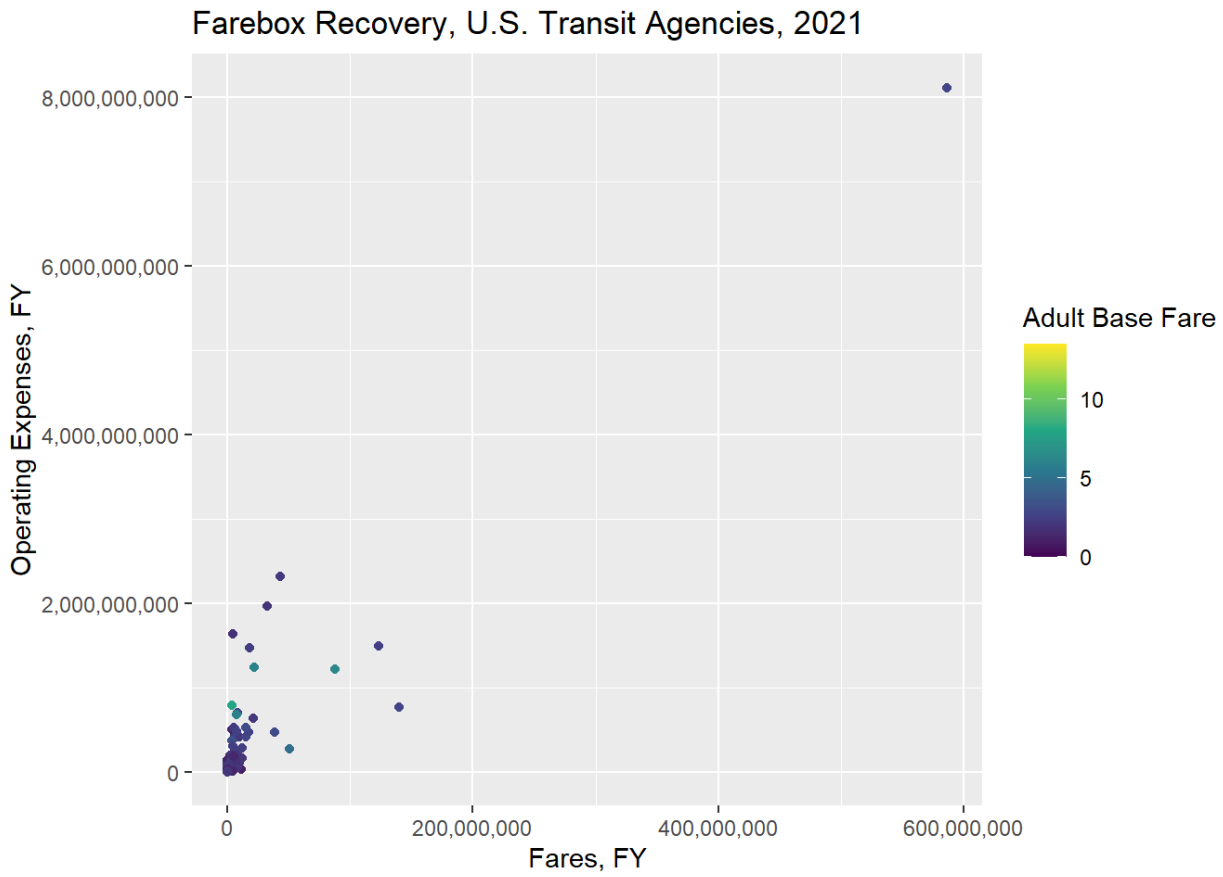
With the merge complete, I could represent variables from both the NTD and the APTA fares database. The first analysis I did was smartcard use at the top 20 transit agencies ranked by unlinked passenger trips. Since unlinked passenger trips are essentially a measure of how many bodies pass through a turnstile or otherwise board a transit vehicle, it could be interesting to know how common smartcards are at the busiest agencies.



According to the data, 16 of the 20 busiest agencies by unlinked passenger trips accept fares via smartcard. New York's MTA clearly smokes every other transit agency in terms of ridership. But how does it *fare* with smartcard adoption?

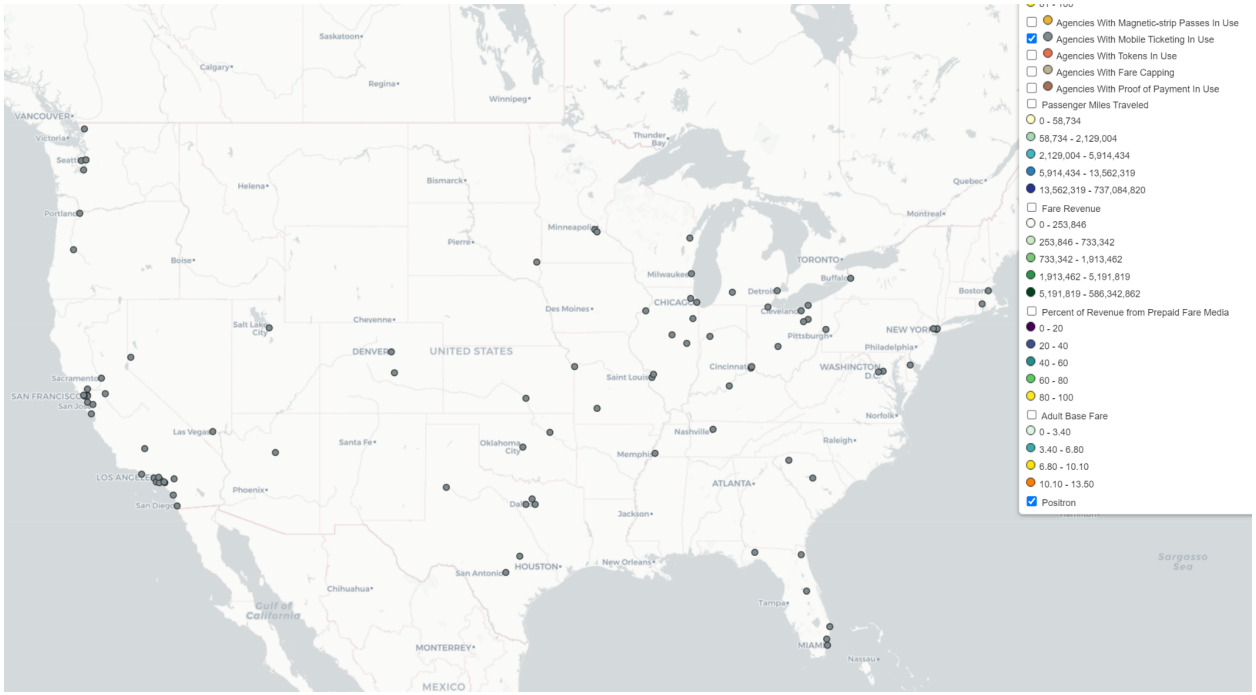


Here we have an interesting contrast. This plot presents the top 40 transit agencies by percent of trips using a smartcard. In systems like WMATA and MARTA, smartcard usage is near total – those technologies have been in place there for nearly 20 years and you essentially cannot board a vehicle without one. In NYC, only 34% of trips used a smartcard in 2022. However, Omny was only rolled out in 2019, and riders may also use their credit or debit cards, cell phones and the classic Metrocard to pay for transit there.



I also wanted to visualize farebox recovery. Farebox recovery is typically represented as a ratio of total revenue from fares to operating expenses. What does it look like on a scatter for agencies nationwide? Here we can see just how little fares contribute to operating expenses. Nationwide, the average is about 30 percent. The one major outlier here pulling out the graph is of course NYC. For its 8 billion dollars in operating expenses, NY MTA pulls in just under 600 million in fares. I classified the points based on adult base fares – as you can see most agencies keep fares below 5 dollars. The lighter dots are more likely commuter rail and bus services.

Finally, I joined the combined NTD and APTA datasets to the National Transit Map. I created an interactive thematic map that visualizes transit agencies by fare collection mode, along with revenue and other performance measures.



Conclusion

There are a few takeaways. The major one concerns data. Variables such as agency name and ID should be standardized across industry organizations and state reporting agencies such as the APTA and the FTA. While the whole spectrum of fare collection is not in the FTA's purview, and that's totally fine, it should be easier for scholars, transit industry professionals and others to look at these data side by side. Another data-related takeaway is that in my own cleaning operations, I took risks and made sacrifices that could compromise my analysis in some ways. For example, I recoded NAs for the smartcard and other variables to "No"; this is a big assumption. I assumed this was survey data with binary yes/no/no response values. No response does not necessarily mean "No". Summarizing the data for all modes comes with its own risks and sacrifices as well. Particularly with fare collection; many agencies require different fare payment media depending on mode, even within the same system. This analysis would not be able to capture that variety.

The second takeaway concerns fare collection. Smartcards are in use at more agencies than I even imagined. It is not a technology solely for large, well-funded systems like WMATA or MTA; they are even in use at systems as marginal as Norwalk, CA's bus system. Mobile ticketing, as illustrated by one of the layers on my interactive map is also being adopted quickly. Agency size does not necessarily seem to be a barrier to adopting a smartcard or similar "new" fare collection technology.

Reflection

While my data clean and recode was effective, it was also occasionally clunky and inefficient. I wonder if there is a more efficient way to capture unmatched rows and re-iterate the matching and recoding in a function. I am happy with my plots, but only as "drafts" and will definitely improve on them in the coming weeks, including faceting/small multiples of my scatters.

As for my final map product, I enjoyed working with QGIS desktop but there are many limitations to the QGIS2web plug-in like lack of support for grouped layers and removal of

various layer styles. The project needed to be completed quickly and so my web map is a good "first draft" but I will be exploring a more effective, clean and interactive webmap using shiny and leaflet or equivalents in the coming weeks as well.

Finally, for this project I used very few variables contained in the APTA fares database. I would like to spend time exploring other variables including different fare collection modes and inter-agency reciprocity agreements.