Visualizing NJ Transit Rail Commutes to NYC

Lisa Hu
Information Visualization 17:610:554
Spring 2021

Intro

- ➔ For five long months in 2018, I commuted to New York City almost every weekday using NJ Transit trains.
- ➔ NJ Transit is rather notorious for their delays, and I experienced a few of them myself, which made me curious to find out:
  - **How likely is it that the train arrives on time at New York Penn Station?**
  - **Is there a difference in lateness between trains scheduled to arrive on weekends versus during peak hours on weekdays?**

Data Source

- ➔ https://www.kaggle.com/pranavbadami/nj-transit-amtrak-nec-performance
  - ◆ Scraped from NJ Transit DepartureVision and compiled & saved in monthly CSV files.
- ➔ I decided to use a full 2-year range, from March 2018 to February 2020.
- ➔ Each row in the file was an individual stop. The default order of rows kept stops of a single train together.

	date	train_id	stop_sequence	from	from_id	to	to_id	scheduled_time	actual_time	delay_minutes	status	line	type
0	2018-03-01	3805	1.0	New York Penn Station	105	New York Penn Station	105	2018-03-02 01:22:00	2018-03-02 01:21:05	0.000000	departed	Northeast Corrdr	NJ Transit
1	2018-03-01	3805	2.0	New York Penn Station	105	Secaucus Upper Lvl	38187	2018-03-02 01:31:00	2018-03-02 01:31:08	0.133333	departed	Northeast Corrdr	NJ Transit
2	2018-03-01	3805	3.0	Secaucus Upper Lvl	38187	Newark Penn Station	107	2018-03-02 01:40:00	2018-03-02 01:40:07	0.116667	departed	Northeast Corrdr	NJ Transit

Data Processing

- ➔ All data processing was done in Python.
- ➔ First, joined all of the monthly CSV files together.
- ➔ Filtered data for only NJ Transit trains, valid statuses, and valid values in all columns.
- ➔ Original data only had dates, but I needed to know day of week to answer second question so added a new column with that information.
- ➔ Filtered for inbound trains stopping at New York Penn Station.
- ➔ Separate CSV files saved for different train lines.

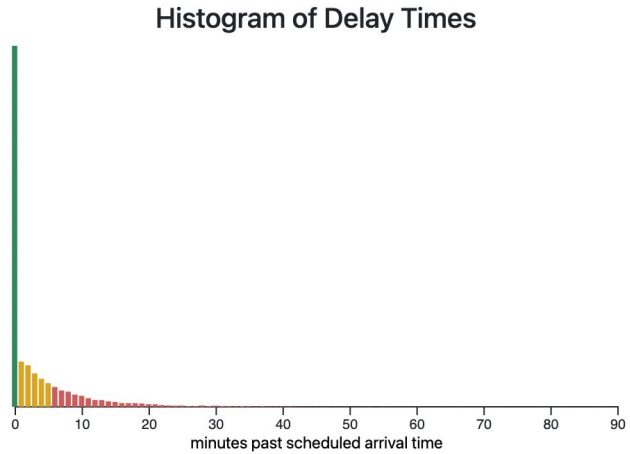
	week_day	hour	delay
0	Wed	06	4
1	Wed	13	6
2	Wed	19	0
3	Wed	15	0
4	Wed	22	2

excerpt of processed CSV file

DataStory Link

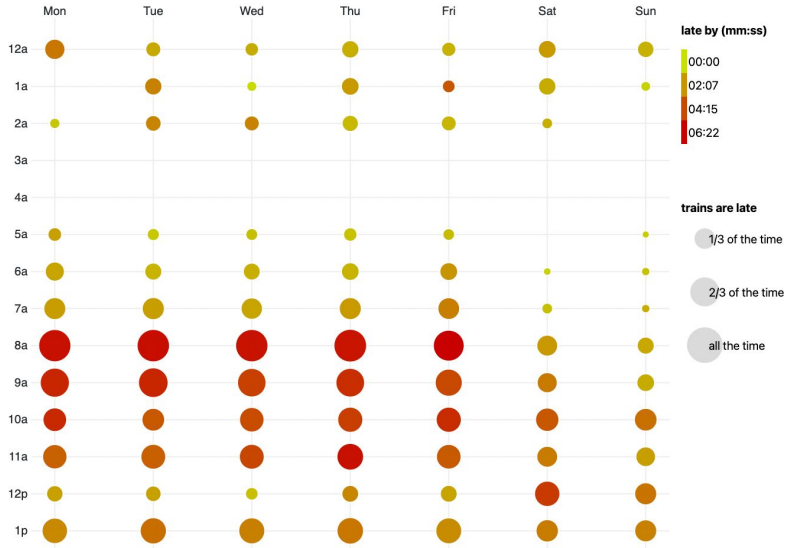
http://lh656.rutgers-sci.domains/infovis/finalproj/results.html

Histogram Vis: How likely is it that the train will arrive on time?



- ➔ Readers choose which line to look at.
- ➔ Color-coded
  - Green - on time
  - Yellow - 1 to 5 minutes late
  - Red - more than 5 minutes late
- ➔ For most lines, only half of the trains arrive on time.

Scatterplot Heatmap Vis: Do time and day of week matter?



- ➔ Readers choose which line to look at.
- ➔ Size of circles indicates percentage of late trains, color indicates amount of time late by.
- ➔ Interactive: Hover over circles to get numerical information.
- ➔ Most lines experience increase in frequency of delays around 8am on weekdays (rush hour).

Conclusion

- ➔ For most NJ Transit rail lines, trains arrive at New York Penn Station at the scheduled time only around half of the time.
  - ◆ Of the late trains, roughly half arrive within 5 minutes of the scheduled time.
- ➔ Commuters looking to arrive in New York City between 8am and 9am should expect trains scheduled at those times to arrive later with greater frequency and sometimes by greater amounts of time.
- ➔ More spikes in frequency of late arrivals occurs around 6pm on weekdays.

Looking Further

- From looking at the scatterplot heatmap, although most lines seem to perform slightly better on weekends, trains running on the North Jersey Coast line actually perform worse.
  - Could this be the result of New Yorkers going down to the beach for a day and then causing lines to be more busy in their return to the city?
  - A hint to the answer can be found by looking at lateness for separate seasons, dividing using the dates in the dataset.
- Much more can still be explored in terms of analysis and visualization of this data.