Analysis of NYC Taxi & Limousine Commission Dataset (2012 - 2021)

Alec Peterson

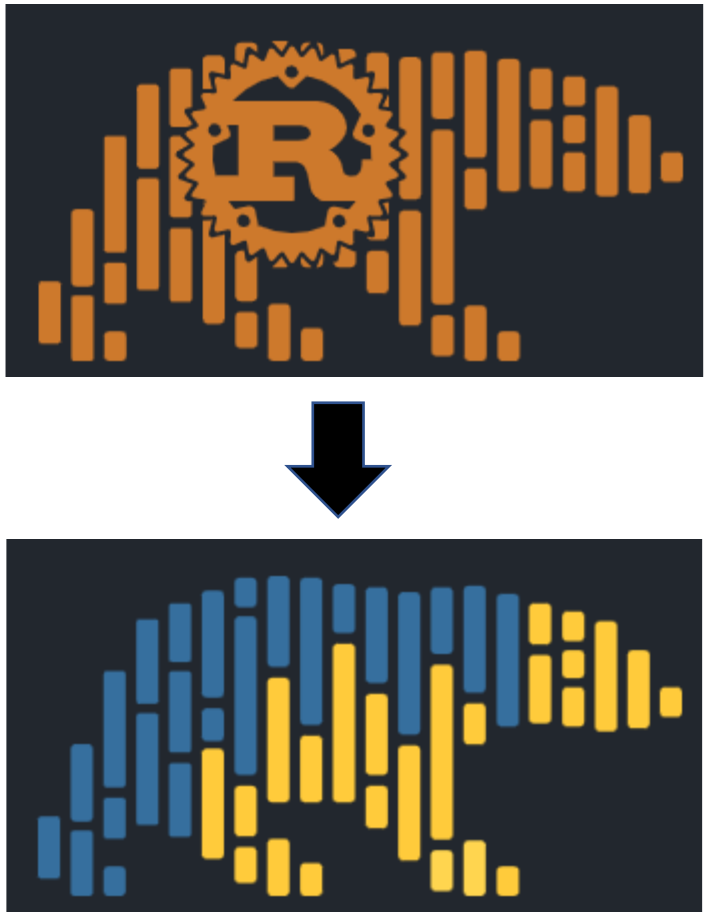
DSCI 521-900

Summer 2022

Project Overview

`polars` for “big” data Processing

- To experiment and process even a year’s worth of data (~9 GB), used the `polars` module
- `polars` is a columnar query engine written in the low-level and increasingly popular **Rust** programming language, and is also made available as a Python package
- Has a similar syntax to `pandas`, but is significantly more performant for large datasets due to:
  - Low-level optimizations
  - Parallelism (using more than one CPU core)
  - Optimized in-memory columnar representation via Apache Arrow
- `polars` worked better out-of-the-box than other packages like `dask` (which parallelizes python code) or `pyarrow` (which deals more directly with Arrow structures)



Memory limits still a challenge

- Combined Taxi and For-Hire Vehicle data from 2012 through 2021 would have exceeded computer's RAM limits (32 GB)

➔ Had to experiment on smaller datasets (1 mo – 1 yr)

- **Process:**
  1. Experiment with `polars` methods on small dataset
  2. Convert groupby-aggregated dataset (in MB) to `pandas` or write to more manageable .parquet file for subsequent `pandas` analysis
  3. Graph with `pandas`-compatible libraries (especially `plotly`)

plotly.express

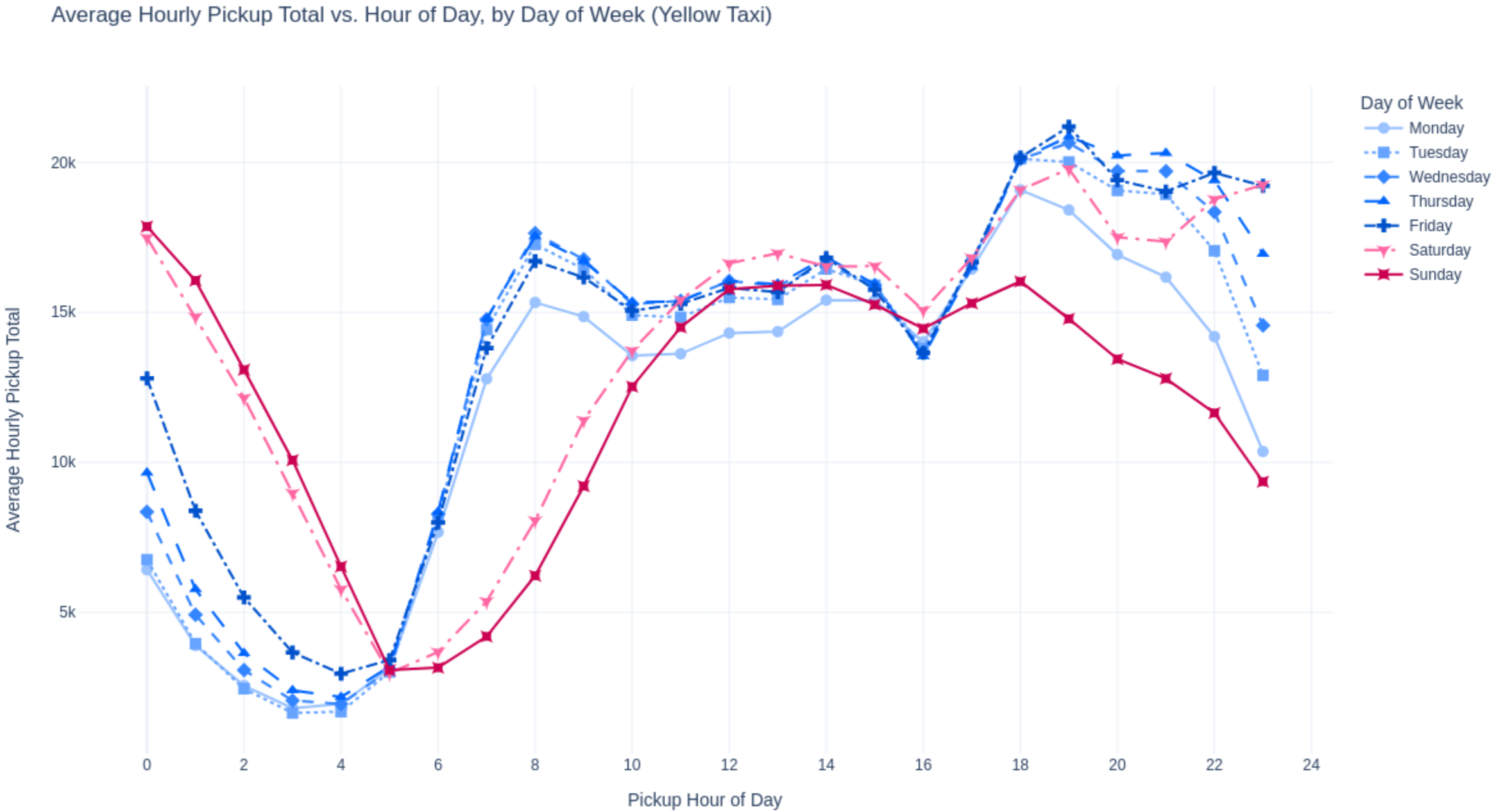
- plotly is an open-source JavaScript library with many interactive chart types, with many made easily available in Python with plotly.express
- Python API straightforward
  - Less clunky and more visually appealing than matplotlib
  - Like seaborn but with more customization and interactivity
- Professional-looking plots, and interactivity valuable for dashboarding applications or other reporting
- Includes charts for Maps, like choropleths which were best for visualizing NYC Taxi Zones (simpler than using matplotlib with geopandas)

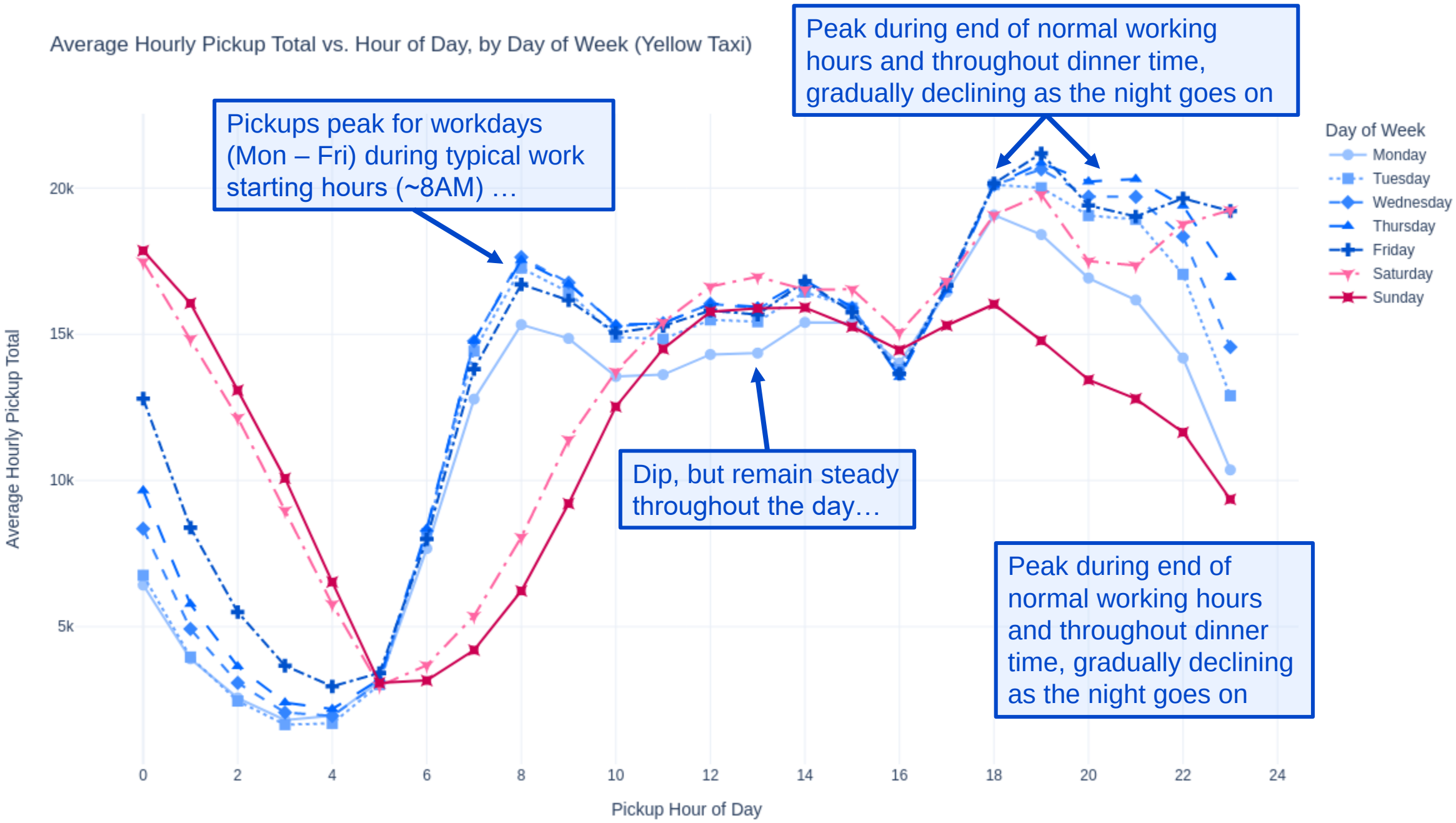
Data Analysis

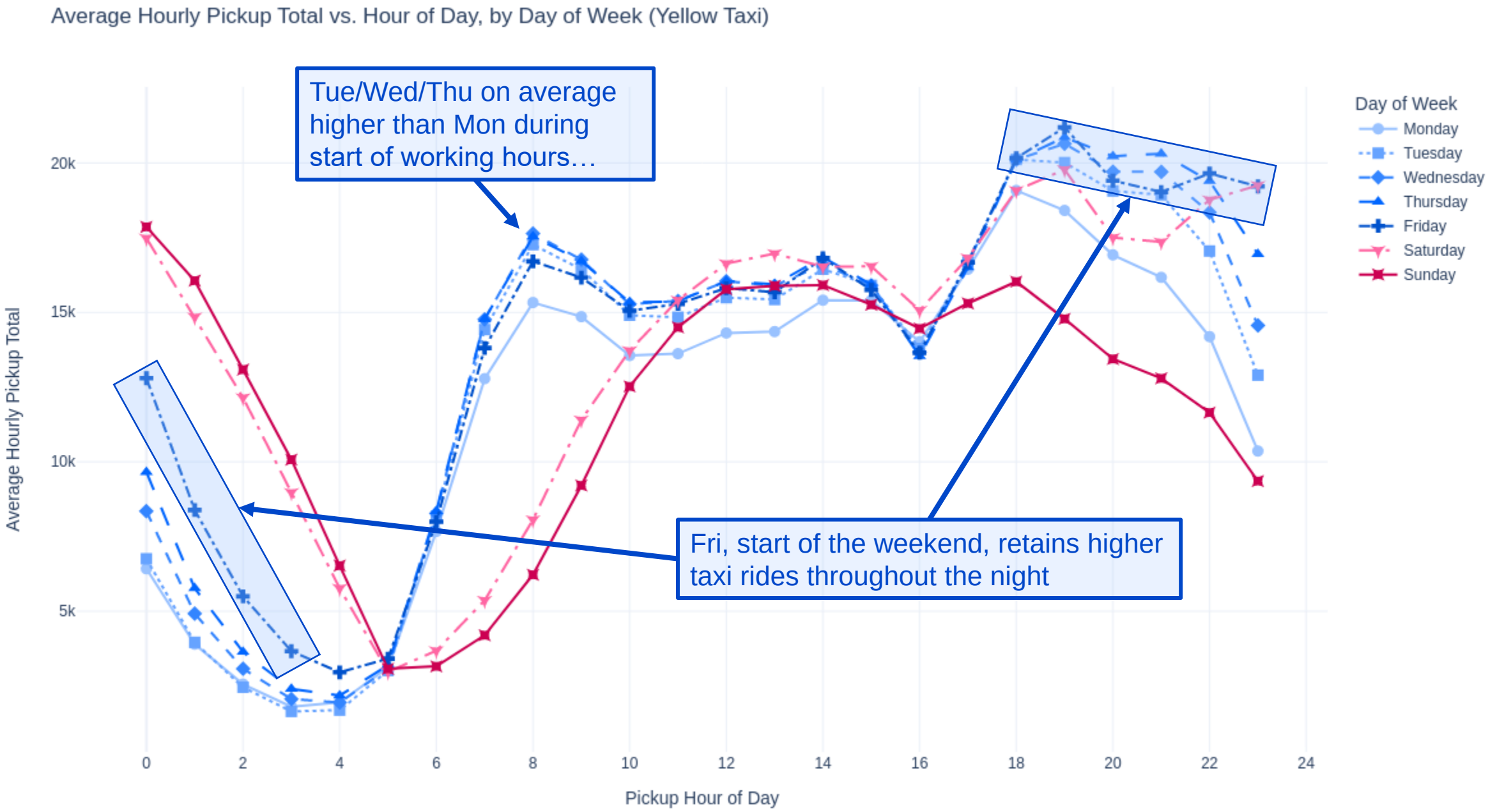
Analysis Goals

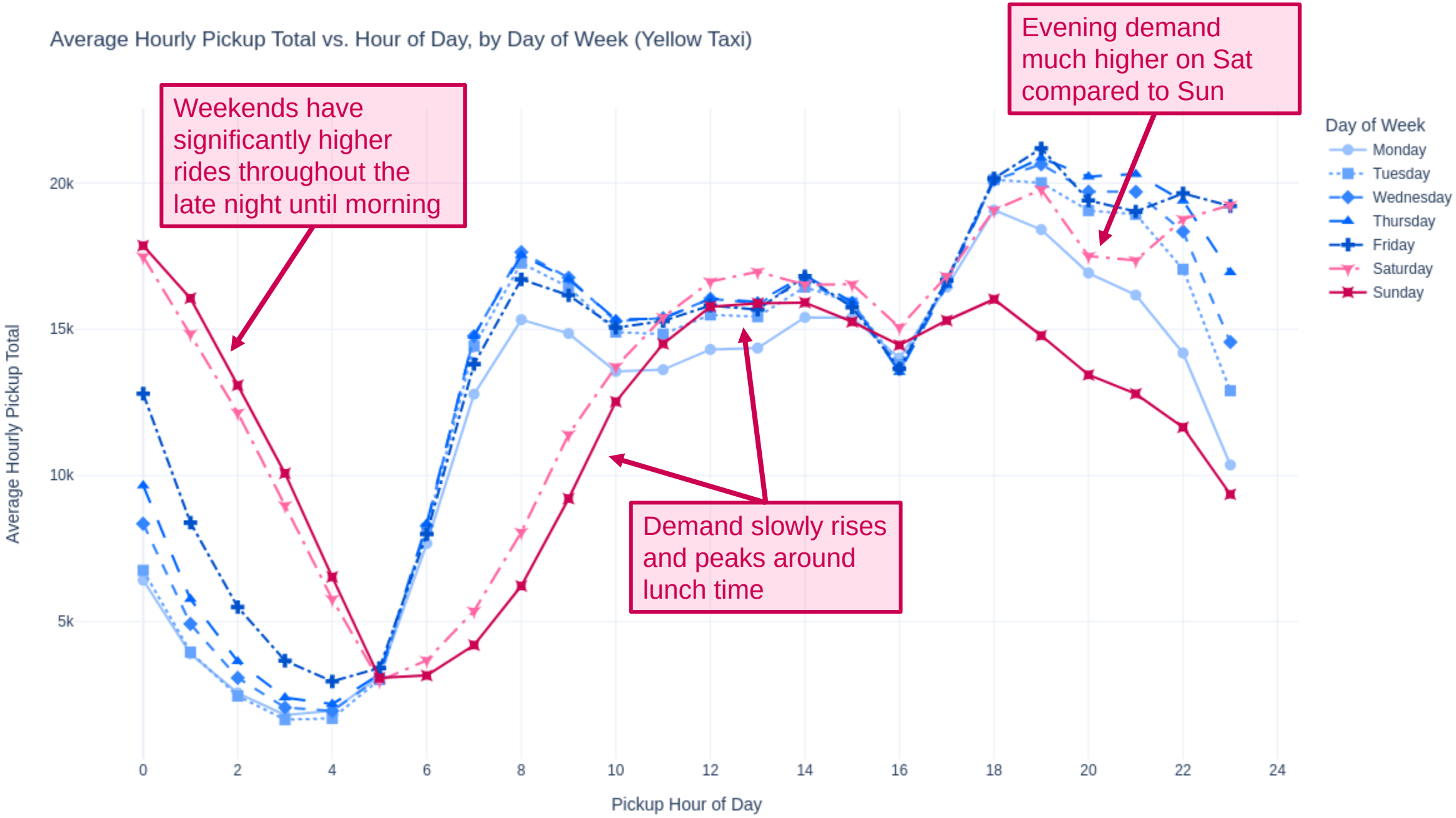
1. **Temporal Analyses**
   A. “Busy” periods over multiple time scales (week, year, multiple years)
   B. Impact of for-hire vehicles (e.g. Uber and Lyft)
   C. Impact of the COVID-19 pandemic

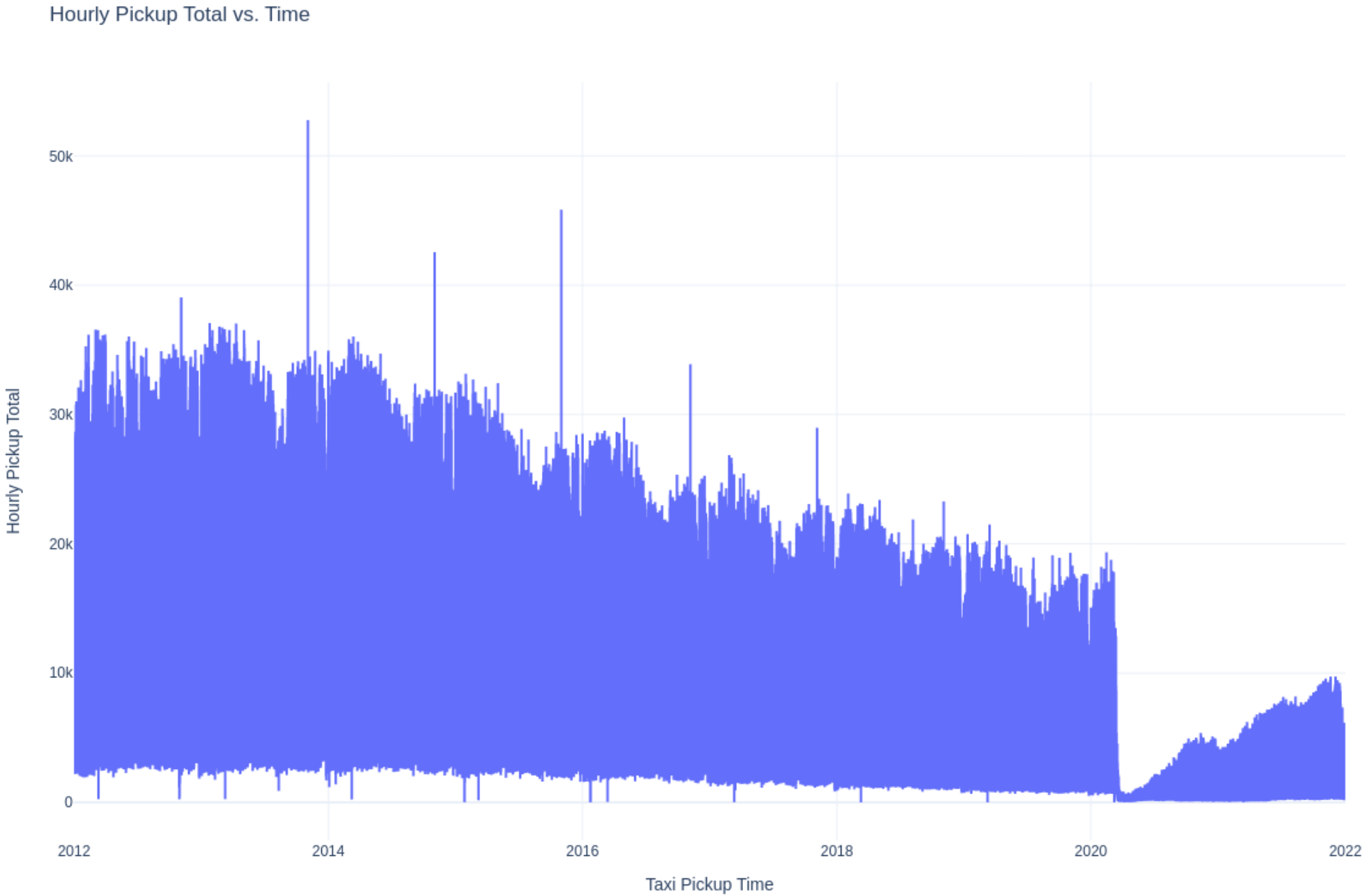
2. **Geospatial Analysis** - busiest areas/boroughs (also over time)

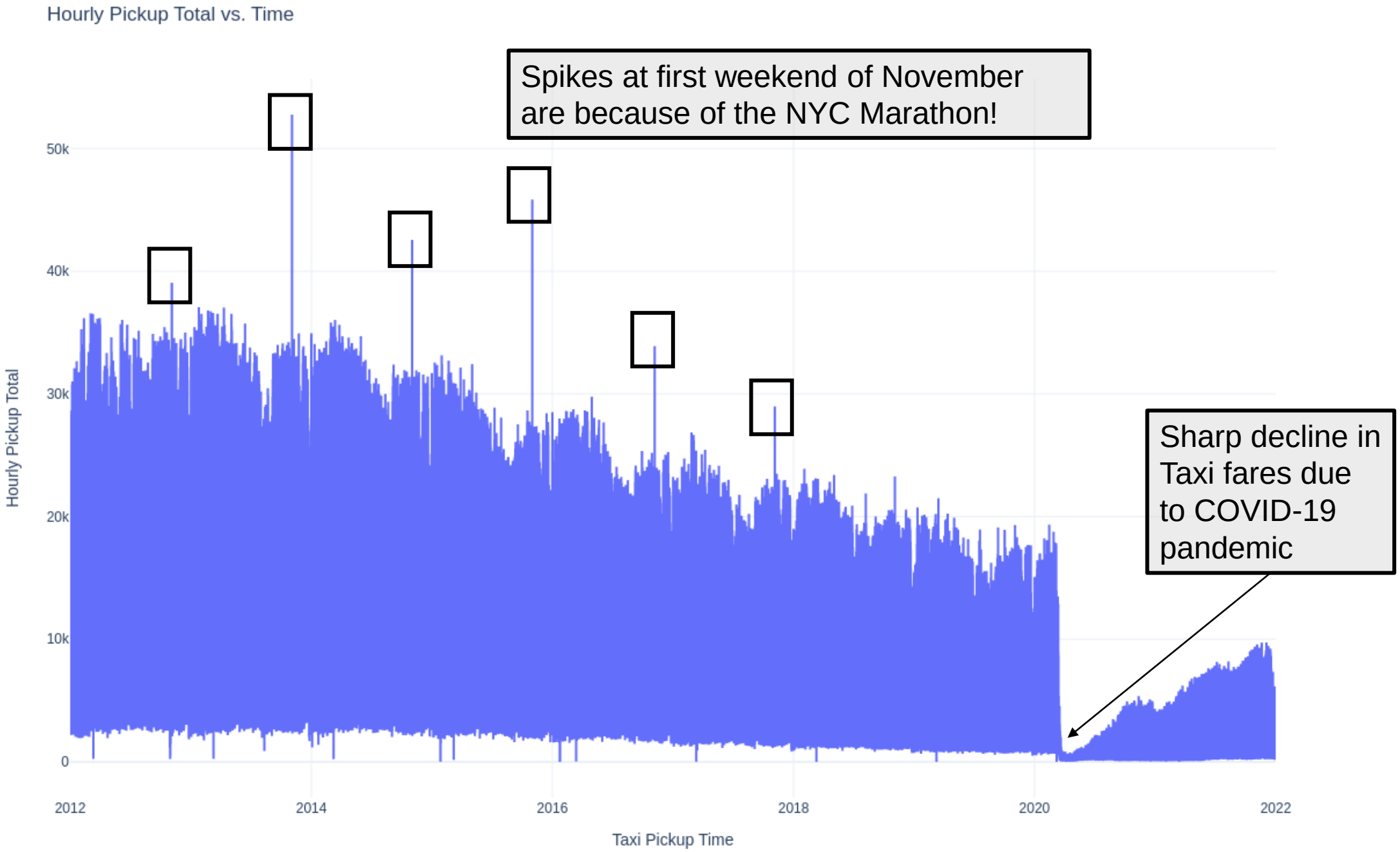

Average Hourly Pickup Total vs. Hour of Day, by Day of Week (Yellow Taxi)
Average Hourly Pickup Total
20k
15k
10k
5k
0
2
4
6
8
10
12
14
16
18
20
22
24
Pickup Hour of Day
Day of Week
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday

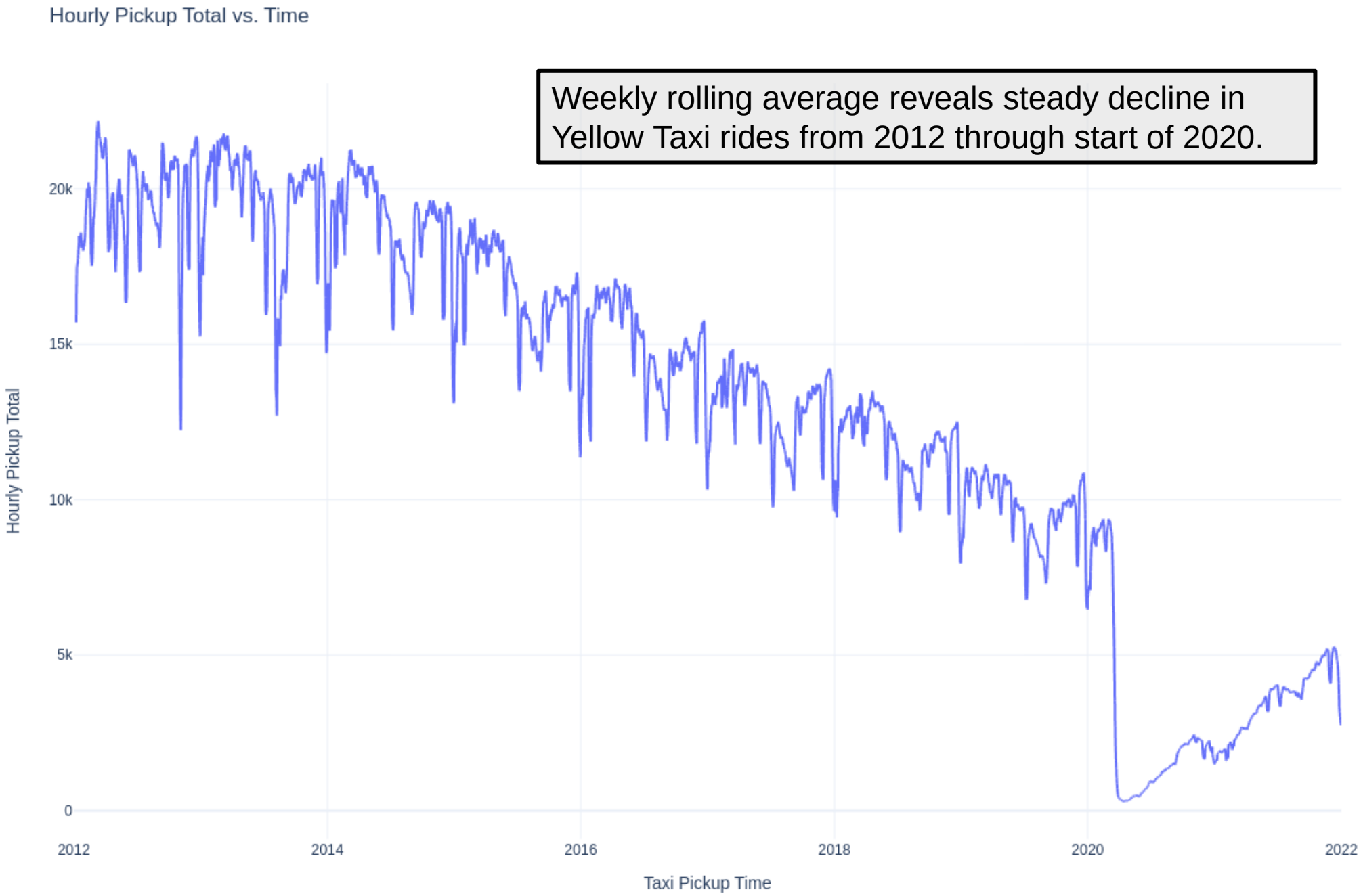

Average Hourly Pickup Total vs. Hour of Day, by Day of Week (Yellow Taxi)
Peak during end of normal working hours and throughout dinner time, gradually declining as the night goes on
Pickups peak for workdays (Mon – Fri) during typical work starting hours (~8AM) …
Dip, but remain steady throughout the day…
Peak during end of normal working hours and throughout dinner time, gradually declining as the night goes on
Day of Week
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Average Hourly Pickup Total
20k
15k
10k
5k
0
2
4
6
8
10
12
14
16
18
20
22
24
Pickup Hour of Day


Average Hourly Pickup Total vs. Hour of Day, by Day of Week (Yellow Taxi)
Tue/Wed/Thu on average higher than Mon during start of working hours…
Fri, start of the weekend, retains higher taxi rides throughout the night
Day of Week
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Average Hourly Pickup Total
20k
15k
10k
5k
0
2
4
6
8
10
12
14
16
18
20
22
24
Pickup Hour of Day


Average Hourly Pickup Total vs. Hour of Day, by Day of Week (Yellow Taxi)
Weekends have significantly higher rides throughout the late night until morning
Evening demand much higher on Sat compared to Sun
Demand slowly rises and peaks around lunch time
Day of Week
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Average Hourly Pickup Total
20k
15k
10k
5k
0
2
4
6
8
10
12
14
16
18
20
22
24
Pickup Hour of Day


Hourly Pickup Total vs. Time
Hourly Pickup Total
50k
40k
30k
20k
10k
0
2012
2014
2016
2018
2020
2022
Taxi Pickup Time


Hourly Pickup Total vs. Time
Spikes at first weekend of November are because of the NYC Marathon!
Sharp decline in Taxi fares due to COVID-19 pandemic
Hourly Pickup Total
50k
40k
30k
20k
10k
0
2012
2014
2016
2018
2020
2022
Taxi Pickup Time


Hourly Pickup Total vs. Time
Weekly rolling average reveals steady decline in Yellow Taxi rides from 2012 through start of 2020.
20k
15k
10k
5k
0
Hourly Pickup Total
2012
2014
2016
2018
2020
2022
Taxi Pickup Time

Analysis Goals

1. **Temporal Analyses**
   A. "Busy" periods over multiple time scales (week, year, multiple years)
   B. Impact of for-hire vehicles (e.g. Uber and Lyft)
   C. Impact of the COVID-19 pandemic

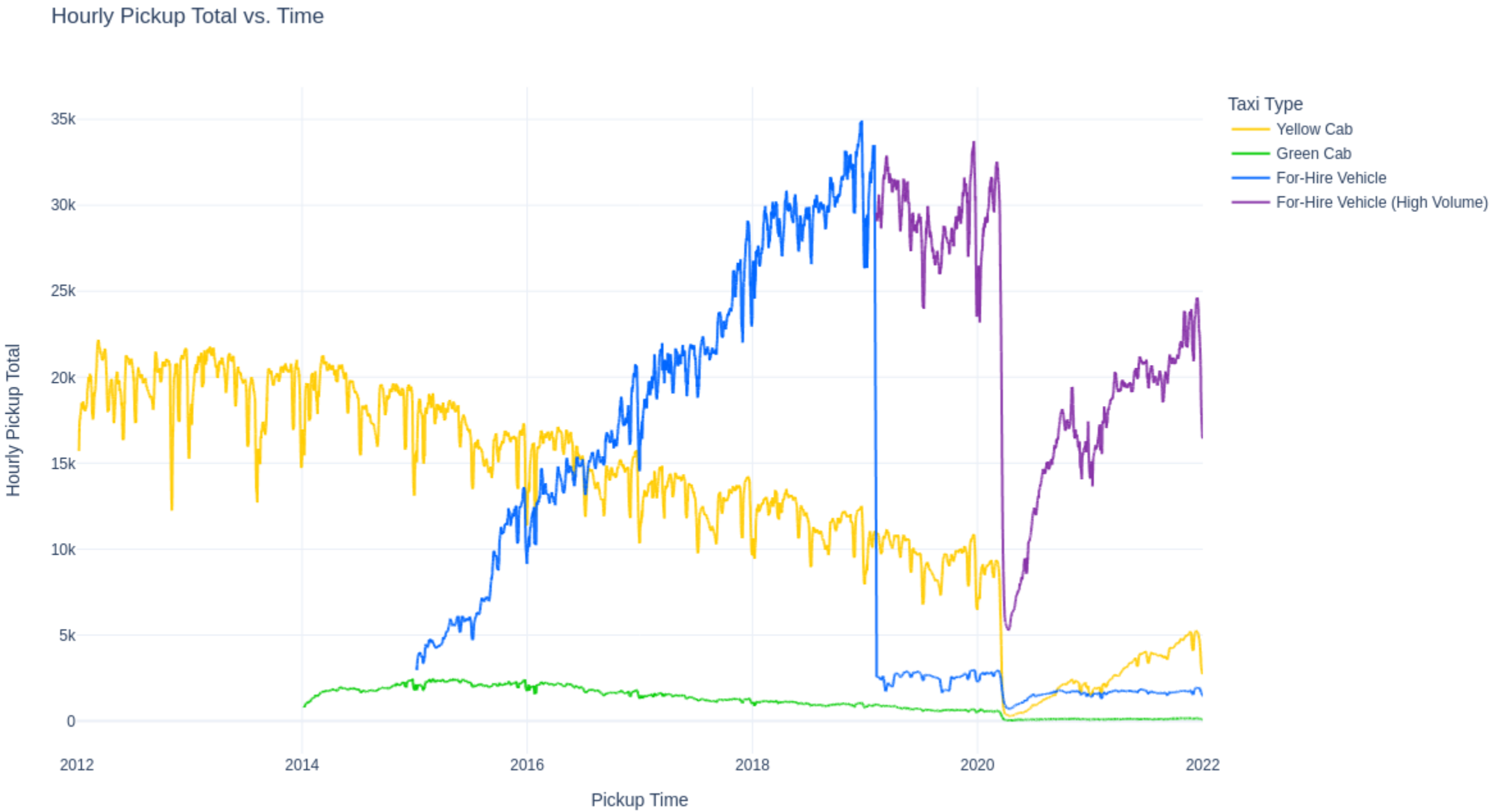
2. **Geospatial Analysis** - busiest areas/boroughs (also over time)

Green Cabs and For-Hire Vehicles

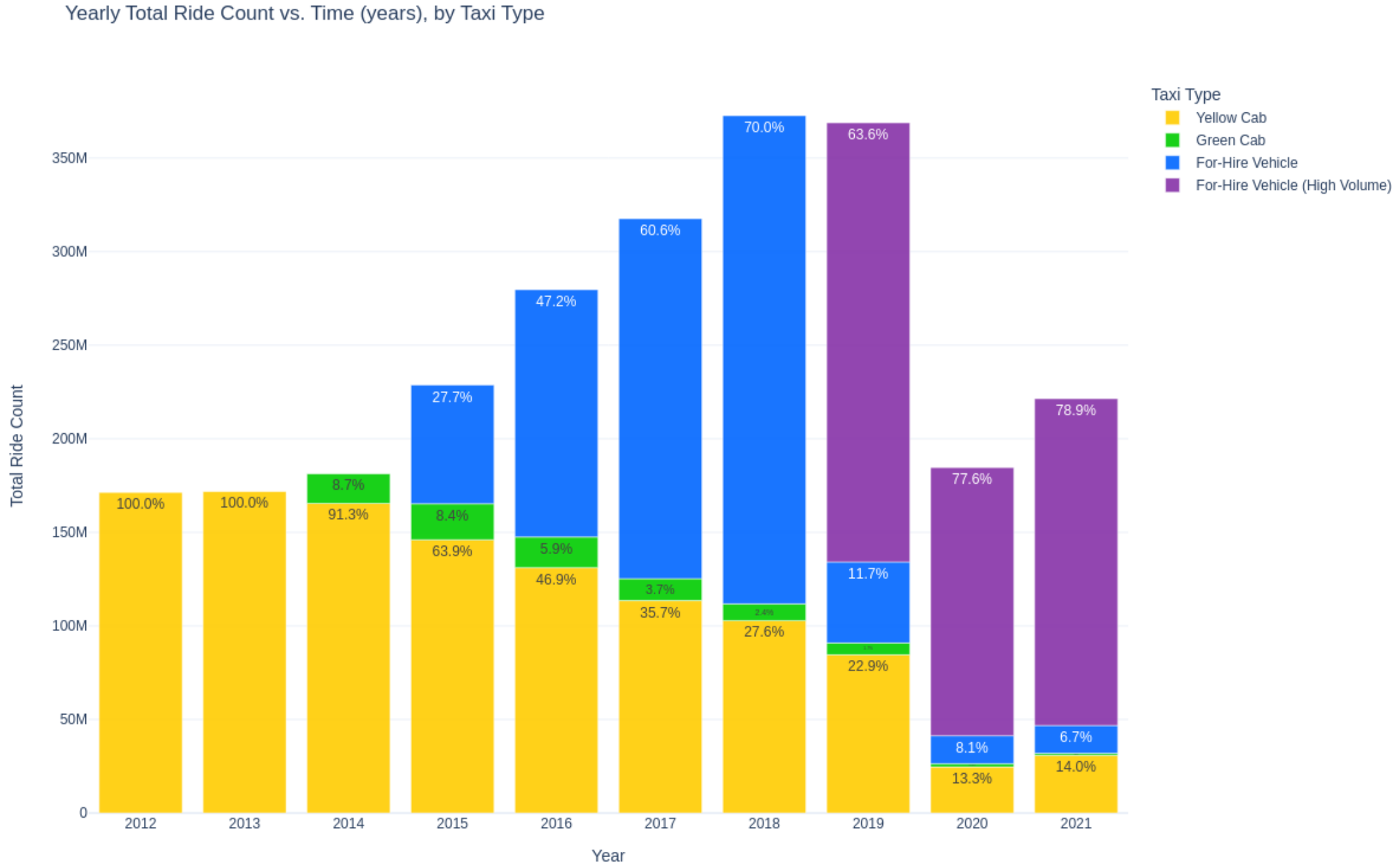
- In August 2013, apple-green painted taxis ("**Green Cabs**") were allowed to pick up passengers in Upper Manhattan, the Bronx, Brookyln, Queens (excluding LaGuardia and JFK Airports) and Staten Island.
- In ~2015, **For-Hire Vehicles (FHVs)** were allowed to provide pre-arranged rides (e.g. via ride-hailing apps).
- In 2019, a "**High Volume**" category of FHVs was created for dispatch bases with >10,000 trips per day. This largely represents cars from the most popular ride-hailing services, Uber and Lyft.

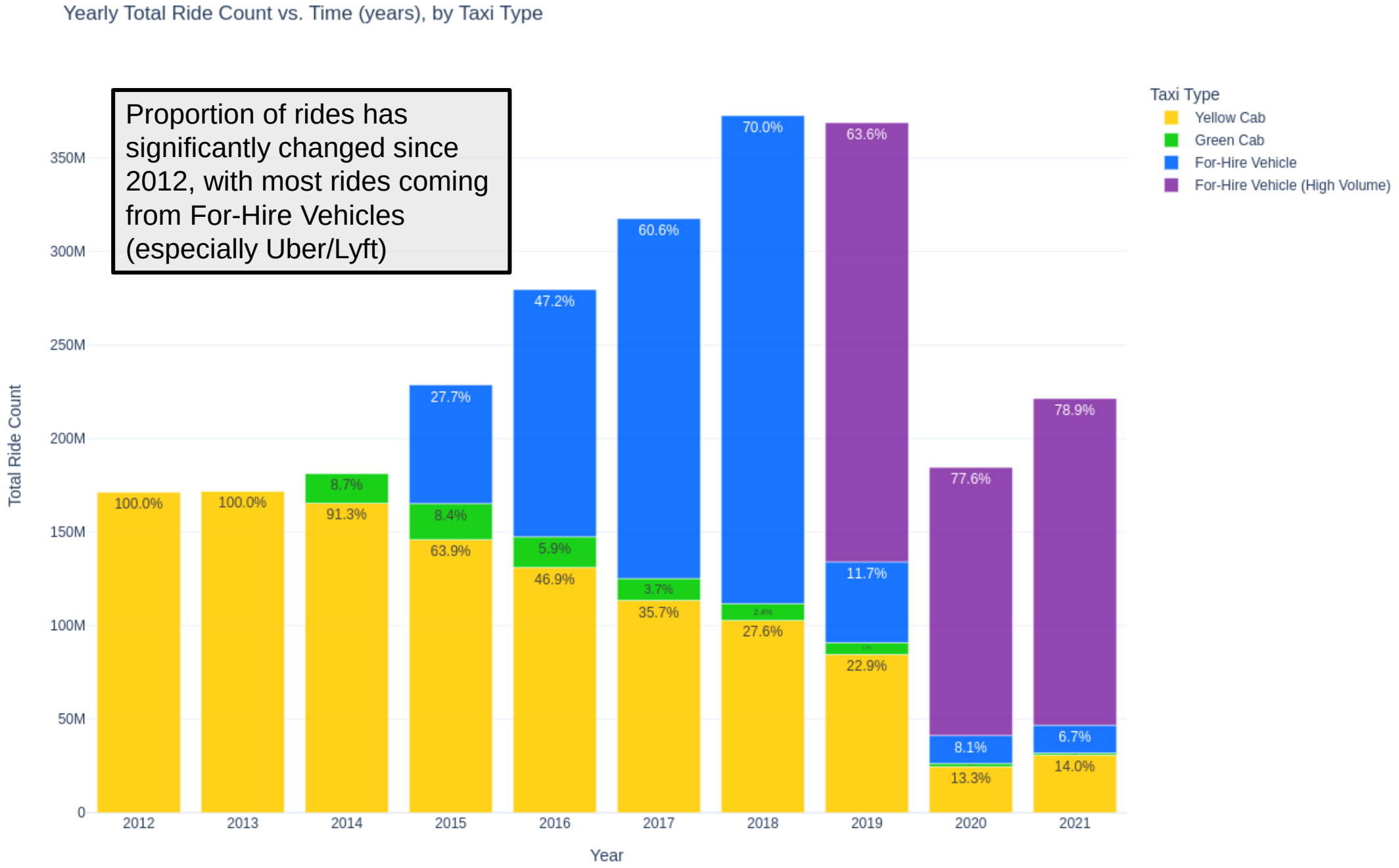
Uber

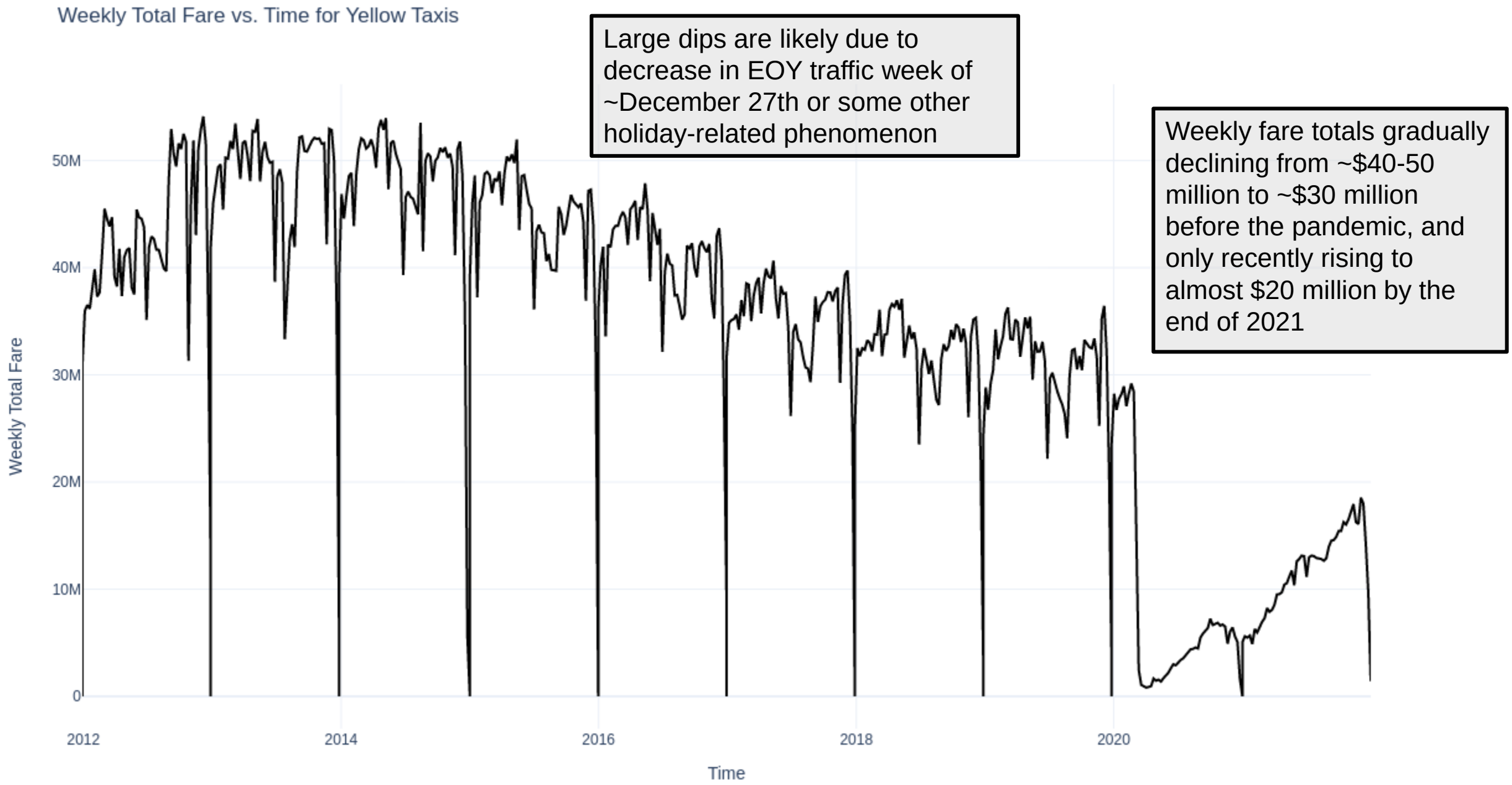
lyft

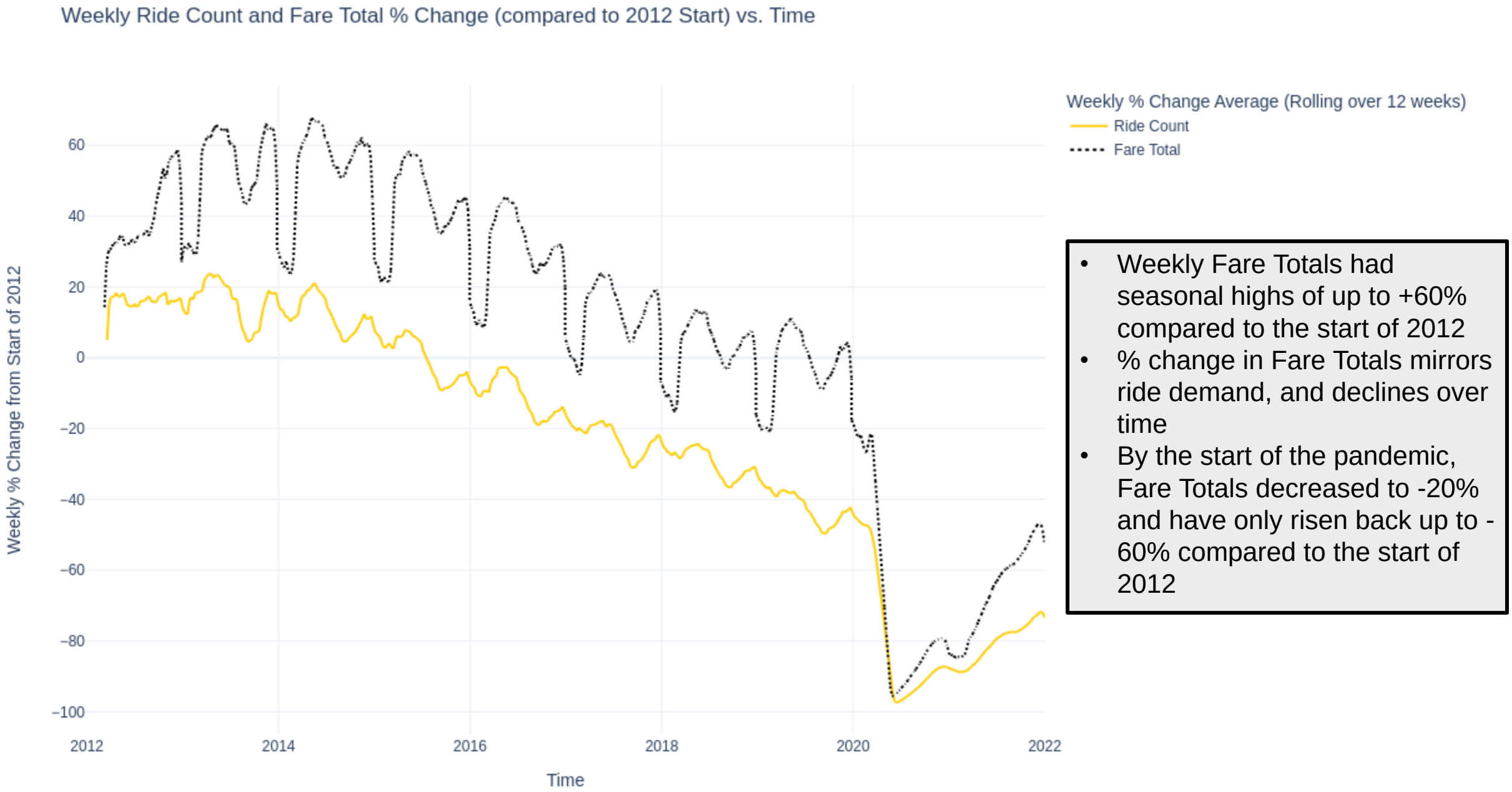

Hourly Pickup Total vs. Time
Taxi Type
Yellow Cab
Green Cab
For-Hire Vehicle
For-Hire Vehicle (High Volume)
35k
30k
25k
20k
15k
10k
5k
0
Hourly Pickup Total
2012
2014
2016
2018
2020
2022
Pickup Time


Hourly Pickup Total vs. Time
From 2015 onward, For-Hire Vehicle demand peaks and surpasses Yellow Taxi historical demand due to ride-hailing apps
High Volume FHVs retained the for-hire ride demand when the category was created in 2019
Introduction of Green Cabs not very significant…
Pandemic demand rose back much quicker for Uber/Lyft while ride demand for cabs and other For-Hire vehicles remain low
Taxi Type
Yellow Cab
Green Cab
For-Hire Vehicle
For-Hire Vehicle (High Volume)
Hourly Pickup Total
35k
30k
25k
20k
15k
10k
5k
0
2012
2014
2016
2018
2020
2022
Pickup Time


Yearly Total Ride Count vs. Time (years), by Taxi Type
Taxi Type
Yellow Cab
Green Cab
For-Hire Vehicle
For-Hire Vehicle (High Volume)
Total Ride Count
350M
300M
250M
200M
150M
100M
50M
0
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
Year
100.0%
100.0%
8.7%
91.3%
27.7%
8.4%
63.9%
47.2%
5.9%
46.9%
60.6%
3.7%
35.7%
70.0%
2.4%
27.6%
63.6%
11.7%
22.9%
77.6%
8.1%
13.3%
78.9%
6.7%
14.0%


Yearly Total Ride Count vs. Time (years), by Taxi Type
Proportion of rides has significantly changed since 2012, with most rides coming from For-Hire Vehicles (especially Uber/Lyft)
Taxi Type
Yellow Cab
Green Cab
For-Hire Vehicle
For-Hire Vehicle (High Volume)
Total Ride Count
350M
300M
250M
200M
150M
100M
50M
0
100.0%
100.0%
8.7%
91.3%
27.7%
8.4%
63.9%
47.2%
5.9%
46.9%
60.6%
3.7%
35.7%
70.0%
2.4%
27.6%
63.6%
11.7%
22.9%
77.6%
8.1%
13.3%
78.9%
6.7%
14.0%
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
Year


Weekly Total Fare vs. Time for Yellow Taxis
Large dips are likely due to decrease in EOY traffic week of ~December 27th or some other holiday-related phenomenon
Weekly fare totals gradually declining from ~$40-50 million to ~$30 million before the pandemic, and only recently rising to almost $20 million by the end of 2021
Weekly Total Fare
50M
40M
30M
20M
10M
0
2012
2014
2016
2018
2020
Time

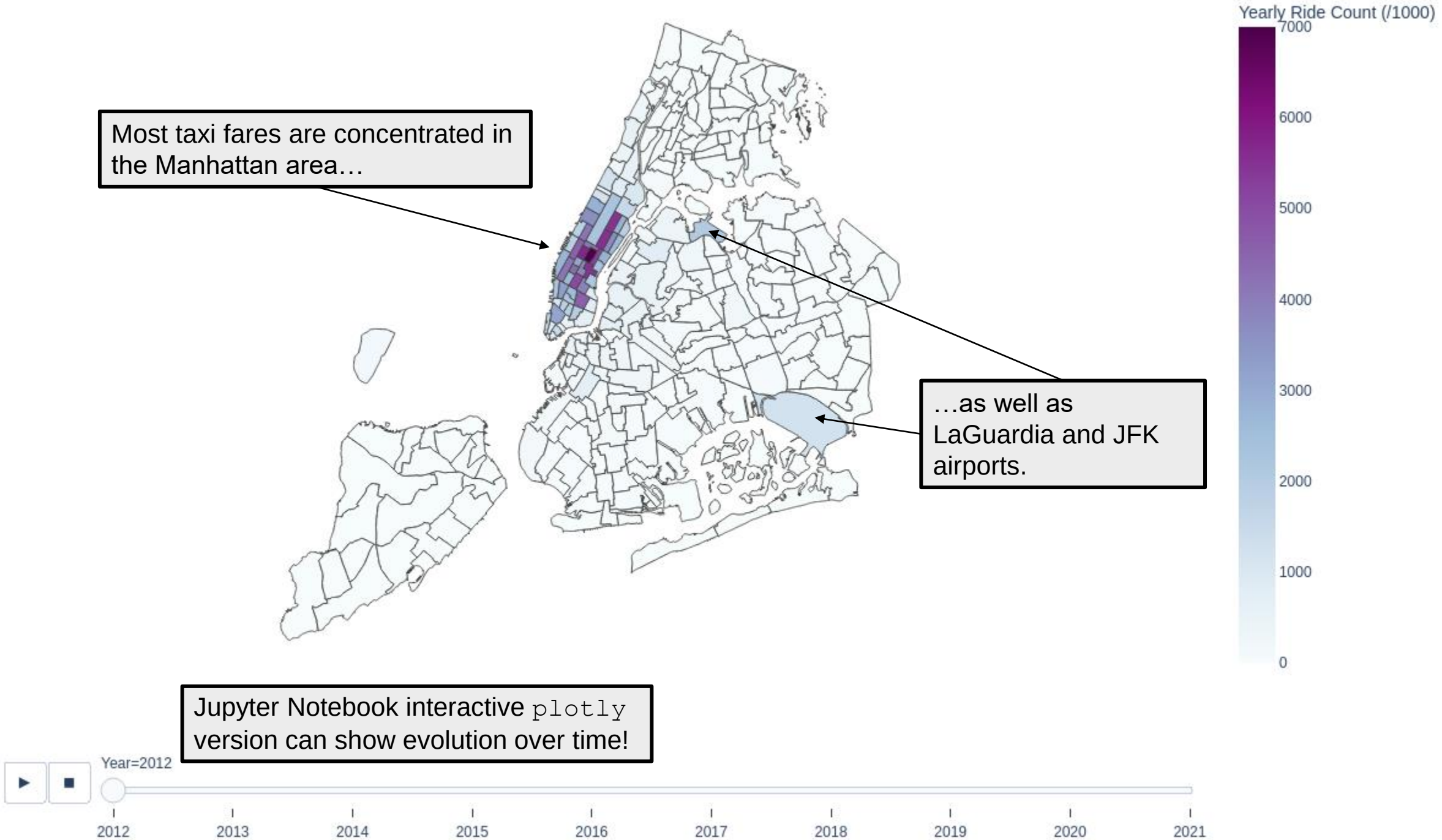


- Weekly Fare Totals had seasonal highs of up to +60% compared to the start of 2012
- % change in Fare Totals mirrors ride demand, and declines over time
- By the start of the pandemic, Fare Totals decreased to -20% and have only risen back up to -60% compared to the start of 2012

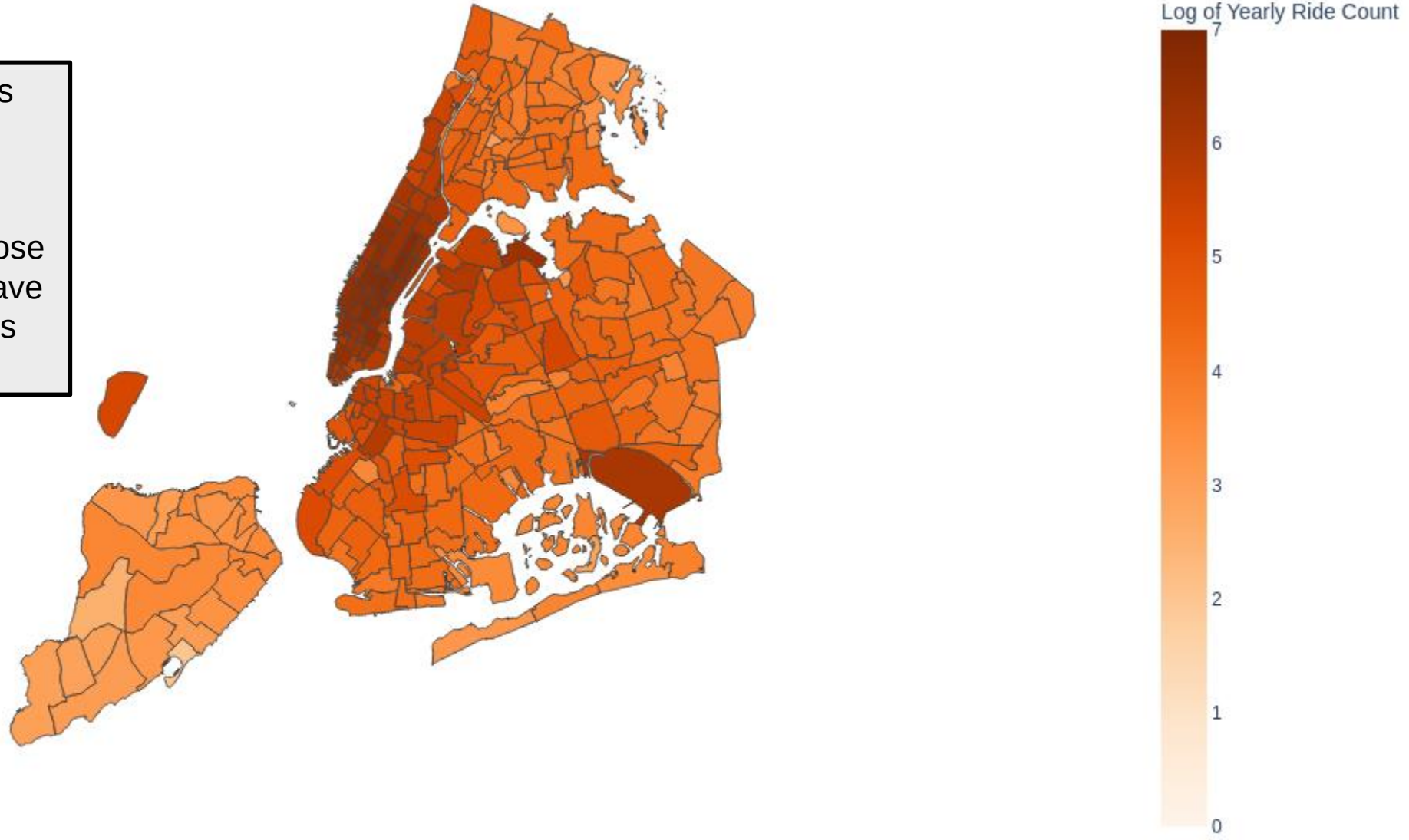
Analysis Goals

1. **Temporal Analyses**
   A. “Busy” periods over multiple time scales (week, year, multiple years)
   B. Impact of for-hire vehicles (e.g. Uber and Lyft)
   C. Impact of the COVID-19 pandemic

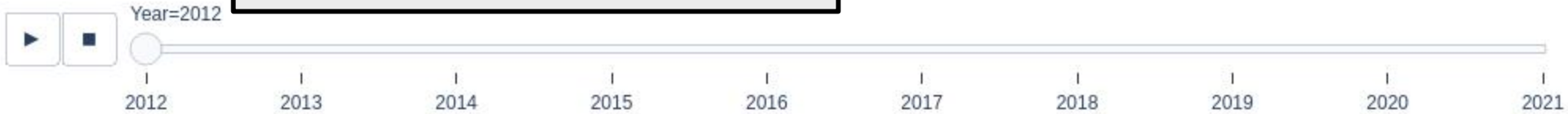
2. **Geospatial Analysis** - busiest areas/boroughs (also over time)


Most taxi fares are concentrated in the Manhattan area…
…as well as LaGuardia and JFK airports.
Yearly Ride Count (/1000)
7000
6000
5000
4000
3000
2000
1000
0
Jupyter Notebook interactive plotly version can show evolution over time!
Year=2012
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021

- A log analysis of taxi zones gives a better sense of magnitude of rides throughout NYC
- Other areas, especially those close to Manhattan, still have significant Yellow Cab rides per year



Jupyter Notebook interactive `plotly` version can show evolution over time!



Conclusions

1. **Takeaways**
   - Yellow taxi industry significantly impacted by ride-hailing apps, representing only 14% of total rides in 2021 compared to 100% in 2012
   - COVID-19 pandemic significantly decreased
   - Temporal and geospatial analysis gives insight to periods and areas of peak demand

2. **Future Analysis Directions**
   - Analysis of airport taxi pickups/drop-offs throughout the day and proportion of total rides in NYC
   - Analysis of busiest taxi zones within each borough
   - Parallel computing to work with dataset more directly (multiple computers), GPU acceleration
   - Slides with revealjs to embed visualizations within slides

Analysis of NYC Taxi & Limousine Commission Dataset (2012 - 2021)

Alec Peterson

DSCI 521-900

Summer 2022