NYC Taxi Activity

Introduction

This report examines how COVID-19 altered NYC's yellow taxi system by comparing monthly trip volumes, fares, and durations (2015-2025). Processing TLC trip records (≈100 billion rides) on CHTC (1,388 R jobs), we applied paired t-tests to detect shifts in ridership and fares, linear regressions to assess fare–duration relationships, and heatmaps to visualize spatio-temporal changes. We observed a significant drop in trips (mean – 15,112 trips/month per zone; p < 0.0001), a fare increase (+$5.24/trip; p < 0.0001), and stronger duration–fare correlations during the pandemic. These findings illuminate demand decline and pricing dynamics, motivating further study of seasonality, borough variation, and ride-hailing competition.

Data and Methods

We used monthly Parquet files from the TLC (via Kaggle), spanning January 2015–January 2025 (≈30 GB, 100 billion rides). Each record includes pickup/drop-off times, coordinates, fare, distance, and passenger count. In R we:

1. **Filtered** out rides with missing/implausible coordinates, zero fares, or durations < 30 s
2. **Aggregated** zone-month summaries: total trips; mean fare, duration, distance, and passengers; rush-hour and weekend shares
3. **Engineered** binary flags for rush hour/weekend, borough assignment (spatial join), and fare per mile

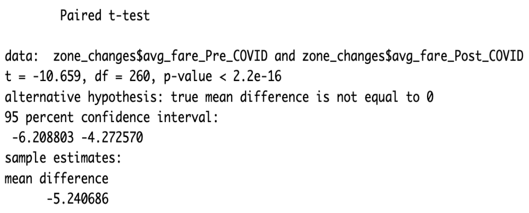
We partitioned the analysis into 1,388 independent jobs on CHTC. Each job executed an R script producing nine CSV outputs including summary statistics, t-test results, regression models, location analyses, and time-based patterns. Jobs used 4GB RAM, 5GB disk storage, and completed in approximately 11 seconds.

To clone our github repository use: git clone https://github.com/NurArsani/Project.git

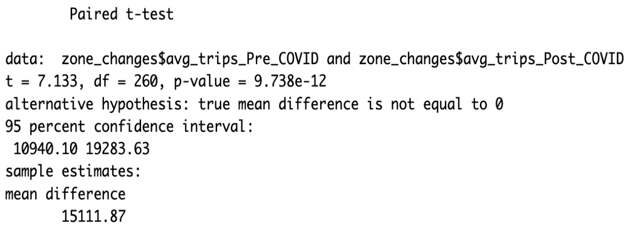
Results

1. **Paired t-tests**

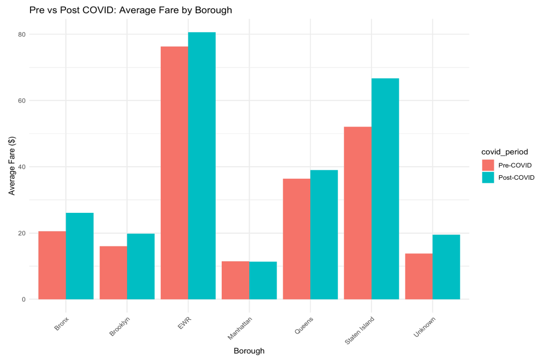
We compared pre-COVID (2019–2020) vs. post-COVID (2020–2024) monthly trip volumes and average fares across 261 zones.

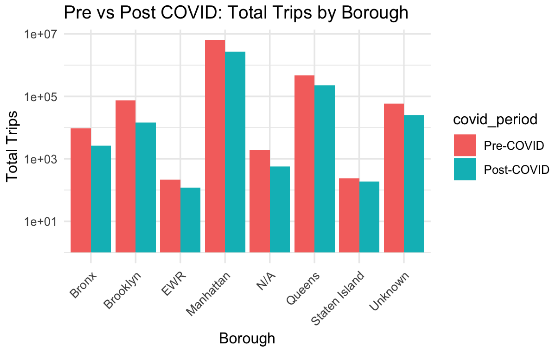


```
	Paired t-test

data:  zone_changes$avg_fare_Pre_COVID and zone_changes$avg_fare_Post_COVID
t = -10.659, df = 260, p-value < 2.2e-16
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -6.208803 -4.272570
sample estimates:
mean difference 
      -5.240686 
```



```
	Paired t-test

data:  zone_changes$avg_trips_Pre_COVID and zone_changes$avg_trips_Post_COVID
t = 7.133, df = 260, p-value = 9.738e-12
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 10940.10 19283.63
sample estimates:
mean difference 
       15111.87 
```

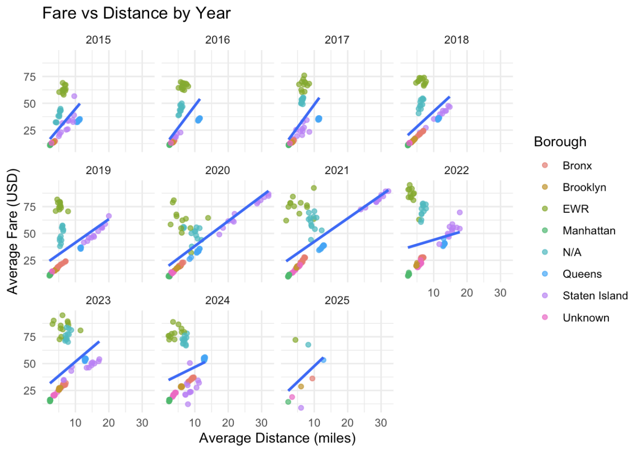




Paired t-tests confirm that 1) fares increased significantly post-COVID (p < 0.0001), with an average rise of $5.24 per trip per zone per month, and 2) a significant drop in trips post-COVID (p < 0.0001), with an average decrease of 15,112trips per zone per month.

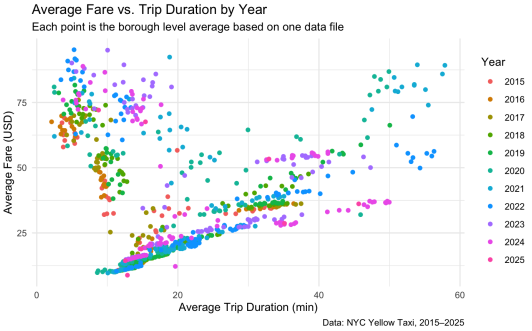
2. Linear Regression

We modeled average fare as a function of trip duration, year, and borough, incorporating interaction terms to detect temporal and spatial variations.



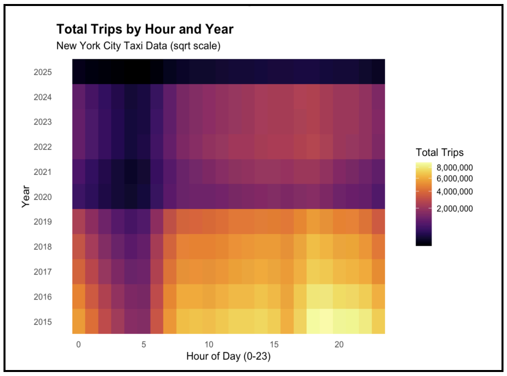
Regressions found a main effect of trip duration on average fare (β = 2.41, *p* < .001), with strongest effects specifically during years (e.g., 2019: β = 0.45, *p* = 0.21 vs. 2020: β = 1.54, *p* < .001).

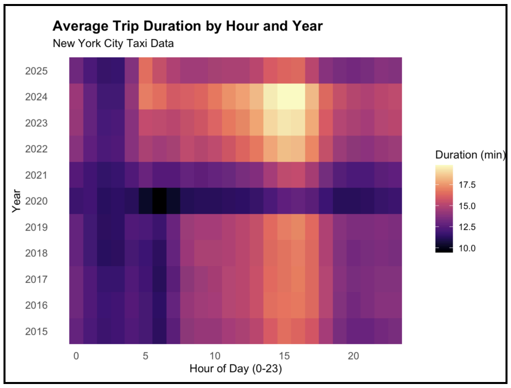
Duration effects varied by year and borough, with significant interaction terms (e.g., avg_distance × EWR: β = -0.98, *p* < .001; avg_distance × Queens: B = 0.59, *p* < .001), suggesting contextual variation in fare pricing.



3. Heatmaps

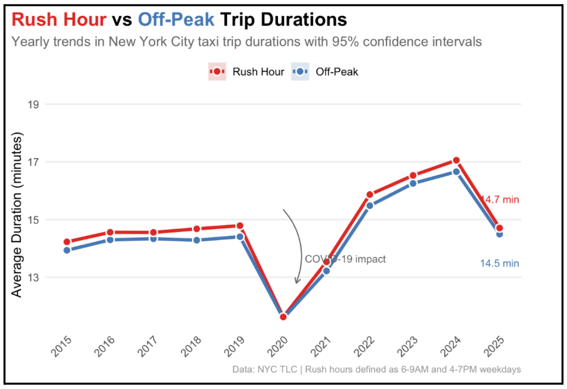
We mapped percent changes in monthly trip volumes by zone and hour to highlight spatial–temporal demand shifts.



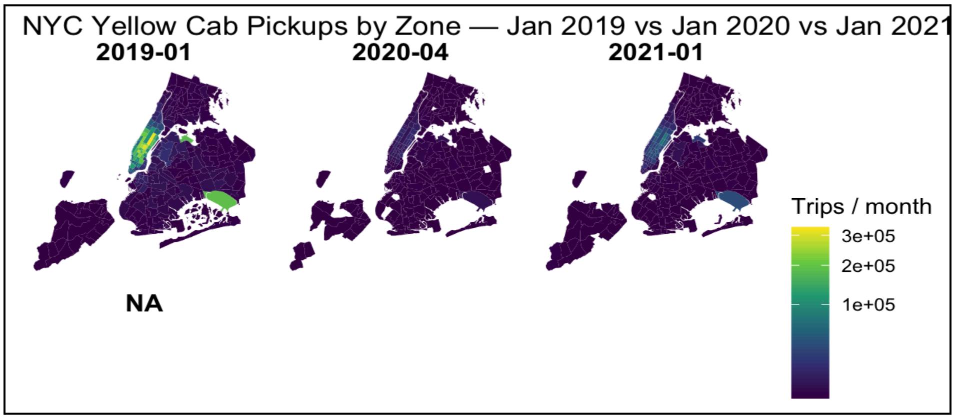


Heatmaps showed the most pronounced demand and trip duration drop in 2020–2021, especially during commuting hours.

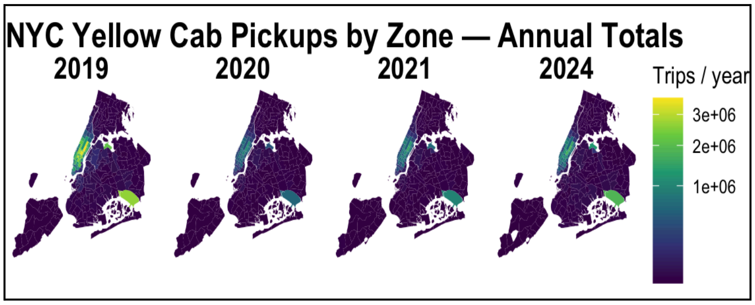
3. **Other**

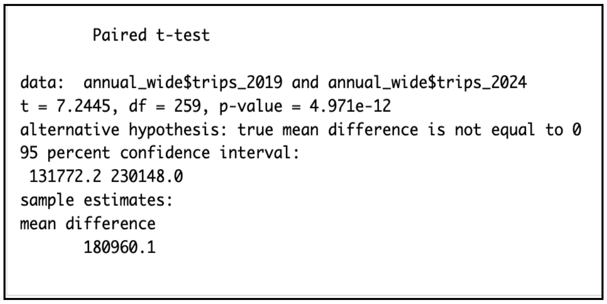


We see a consistent gap–rush hour trips are consistently longer.


NYC Yellow Cab Pickups by Zone — Jan 2019 vs Jan 2020 vs Jan 2021
2019-01
2020-04
2021-01
NA
Trips / month
3e+05
2e+05
1e+05

We see a slow rebound in January 2021–not back in full force.





```
	Paired t-test

data:  annual_wide$trips_2019 and annual_wide$trips_2024
t = 7.2445, df = 259, p-value = 4.971e-12
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 131772.2 230148.0
sample estimates:
mean difference 
       180960.1 
```

We see the significant statistical difference between trips in 2019 vs 2024.

Weaknesses:

- Our data included no data from Uber or Lyft,
- Seasonal confounding variables
- Model simplicity–only including duration, year, and borough.

Conclusion

Analyses demonstrated that NYC taxi usage decreased during the pandemic. Average fares rose substantially and relationships between fare and duration strengthened. These findings underscore how external shocks can simultaneously depress demand and shift pricing. Future work should incorporate ride-hailing data and apply advanced time-series models to disentangle seasonality.