



Changes in NYC Cab Activity - *Stat 405*

Jonathan, Youngwoo, Aleeya, Kalynn, Nur

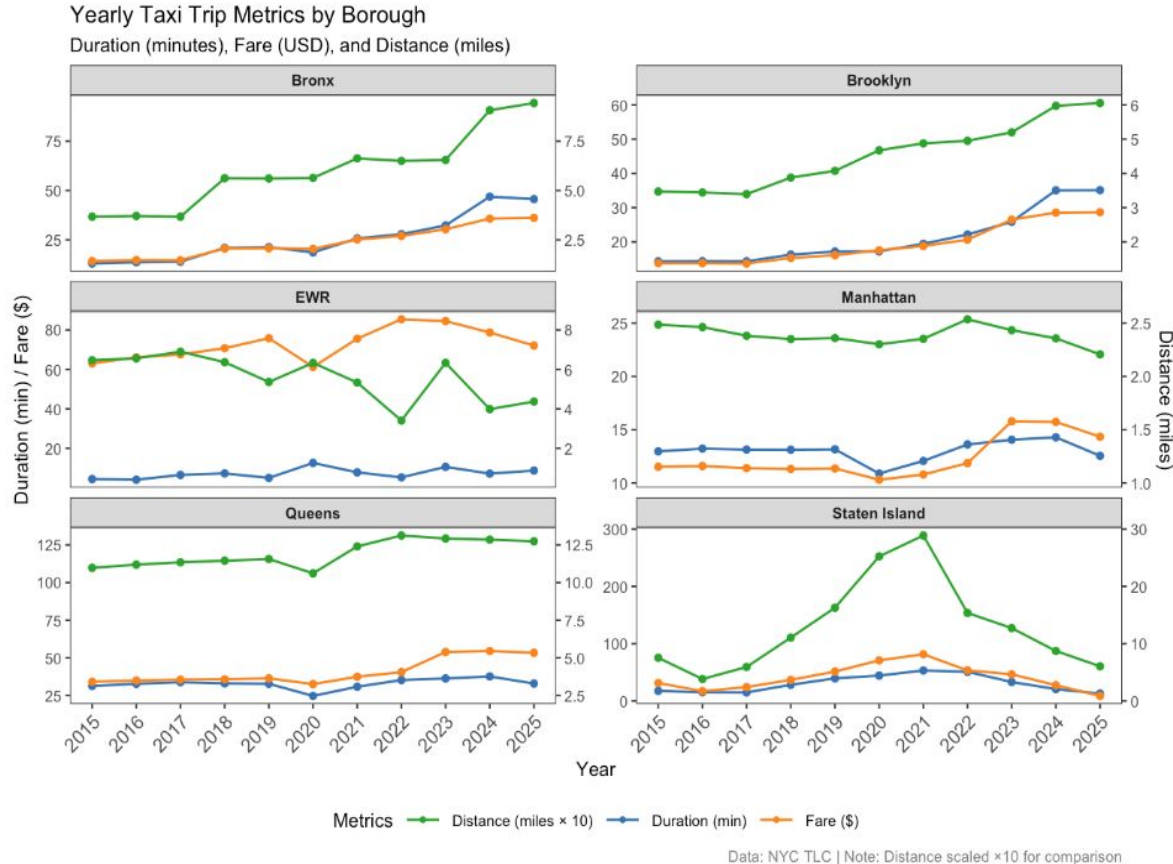
Introduction

- Data from NYC Taxi and Limousine Commission (NYC).
 - 30GB
 - Ranging from January 2015 - January 2025.
 - Relevant Variables:
 - Trip duration.
 - Pickup / dropoff location.
 - Fare amount.
 - Date and time of ride.
- Critical Questions:
 - 1. How do trends and pricing of taxi cab rides in NYC fluctuate based on daily, monthly and yearly trends?**
 - 2. What impact did the pandemic have on cab rates, fares, and usage?**

Methods

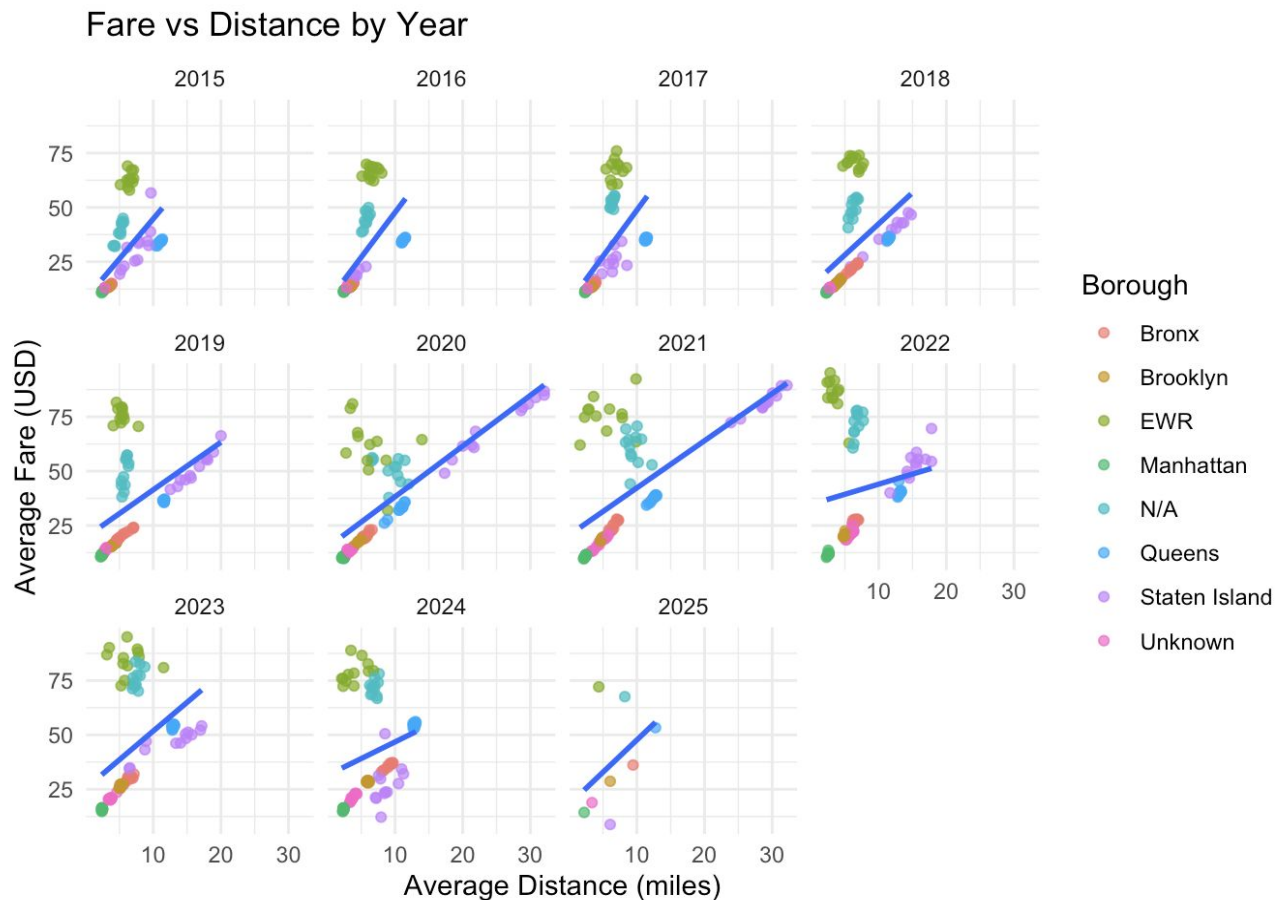
- Dataset Processing: Analyzed NYC taxi trip data stored in Parquet format, measuring trip duration, fare amounts, and passenger patterns across different boroughs
- Parallelization Strategy: 1,388 parallel jobs using HTCondor at CHTC
- Resource Allocation: Each job: 1 CPU core, 4GB RAM, and ~5GB disk space
- Statistical Analysis: Generated regression models predicting trip outcomes, t-tests comparing conditions (rush hour vs. non-rush hour, Manhattan vs. outer boroughs), and time-based pattern analysis
- Output Files: Produced 9 CSV reports including location summaries, hourly/daily patterns, and borough comparisons, all consolidated in a comprehensive RDS file
- Job Performance: Average execution time of ~11 seconds per job with minimal memory footprint
- Data Transfer: Each job processed ~33MB of outbound data and ~6MB of inbound data

Analysis and Interpretation



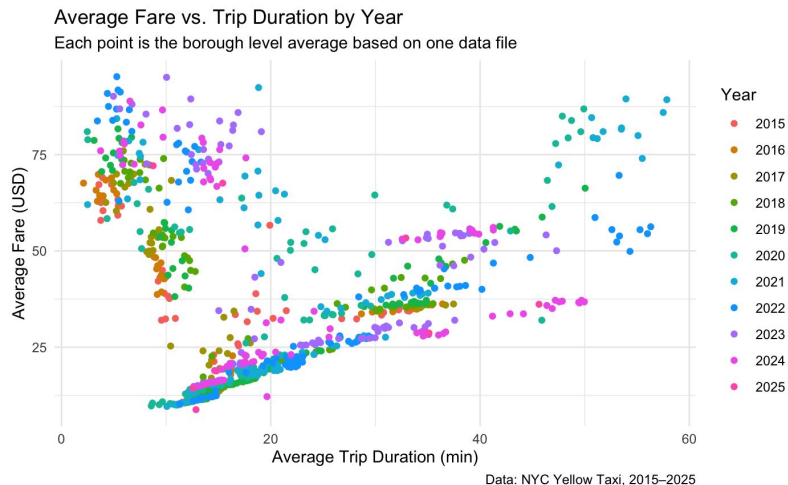
1. Queens has consistently high values across all metrics
2. Staten Island shows the longest distances but surprisingly moderate fares. There is a notable peak around **2020–2021**, indicating a sudden rise in trip distances, likely linked to **pandemic-related behavior** (e.g., fewer, longer trips).
3. Manhattan has the shortest trips (low duration/distance) but relatively high fares per mile
4. EWR has highest average fare possibly due to trips to and from Newark Airport are longer in distance compared to typical city trips, and also include extra fees such as airport surcharges and bridge or tunnel tolls.

Fare vs Distance by Year



Fare increases by distance as well as year increases.

Average Fare vs. Trip Duration



Strongest correlation between trip duration and fare during covid years.



Call:
lm(formula = avg_fare ~ avg_duration * year, data = borough_s)

Residuals:

Min	1Q	Median	3Q	Max
-35.00	-16.68	-9.25	17.27	64.73

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	34.82888	4.88533	7.129	2.00e-12	***
avg_duration	-0.49937	0.29838	-1.674	0.094536	.
year2016	1.96117	6.75080	0.291	0.771491	
year2017	2.51530	6.81679	0.369	0.712221	
year2018	1.61308	6.93468	0.233	0.816115	
year2019	0.48577	6.55890	0.074	0.940976	
year2020	-21.16724	6.57406	-3.220	0.001326	**
year2021	-12.62860	6.53013	-1.934	0.053424	.
year2022	14.12253	6.50880	2.170	0.030273	*
year2023	18.56271	7.19607	2.580	0.010043	*
year2024	16.62242	6.69198	2.484	0.013166	*
year2025	2.92185	16.03326	0.182	0.855436	
avg_duration:year2016	-0.17064	0.41075	-0.415	0.677915	
avg_duration:year2017	-0.05423	0.41078	-0.132	0.894995	
avg_duration:year2018	0.23189	0.38597	0.601	0.548119	
avg_duration:year2019	0.44765	0.35770	1.251	0.211078	
avg_duration:year2020	1.53768	0.35362	4.348	1.52e-05	***
avg_duration:year2021	1.30591	0.33935	3.848	0.000127	***
avg_duration:year2022	0.18228	0.33766	0.540	0.589435	
avg_duration:year2023	0.10738	0.36444	0.295	0.768325	
avg_duration:year2024	0.09774	0.34144	0.286	0.774742	
avg_duration:year2025	0.48716	0.66638	0.731	0.464926	

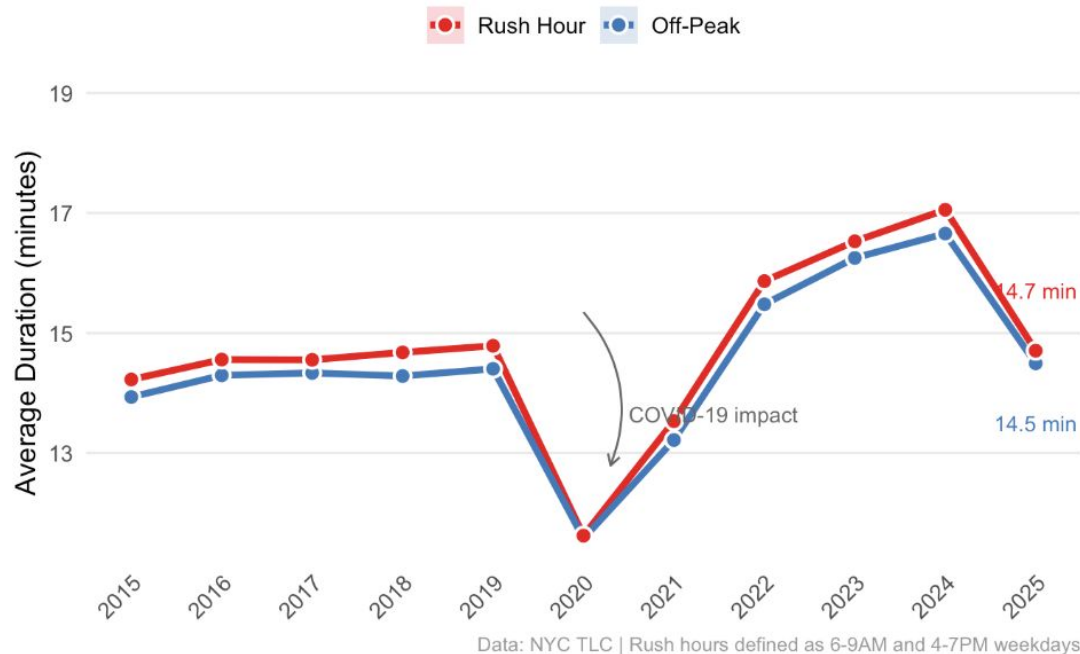
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' ' 1

Rush Hour vs. Non-Rush Hour Trip Duration by Year

2015-2025

Rush Hour vs Off-Peak Trip Durations

Yearly trends in New York City taxi trip durations with 95% confidence intervals



Analysis and Interpretation

Consistent Gap: Rush hour trips are consistently longer (red above purple), validating traffic congestion effects.

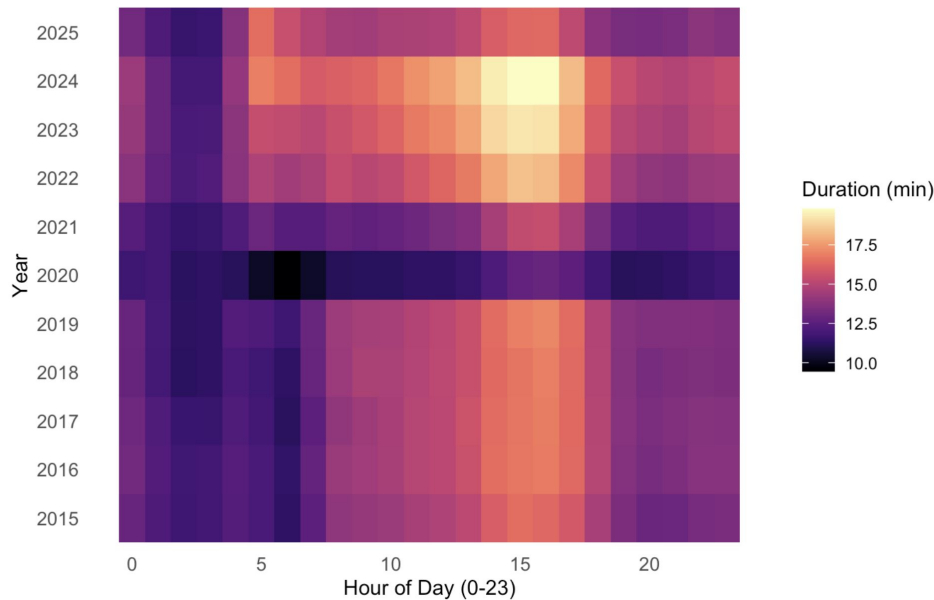
2015–2019: Gradual increase suggests worsening congestion.

2020: Likely a sharp drop (COVID-19 lockdowns reduced traffic).

2021–2025: Recovery trend

Average Trip Duration by Hour and Year

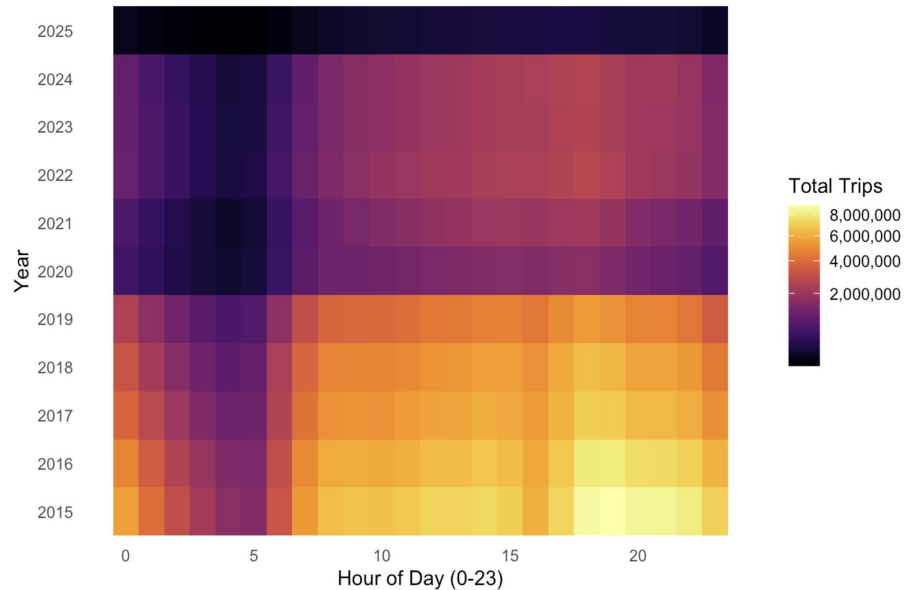
New York City Taxi Data



Average trip durations dropped sharply in 2020, likely due to COVID-19, and have since increased, especially during afternoon hours. Trips typically occur the longest between 3pm and 7pm.

Total Trips by Hour and Year

New York City Taxi Data (sqrt scale)



Total taxi trips collapsed in 2020 and have not fully recovered by 2025, indicating lasting changes in travel behavior. Most trips pre COVID-19 occur after 6am and before midnight, and peak between 4pm and 7pm.

STATISTICAL ANALYSIS

Trips per month per zone

NYC Yellow Cab Pickups by Zone — Jan 2019 vs Jan 2020 vs Jan 2021



2019-01

N/A

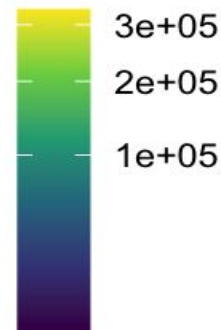


2020-04

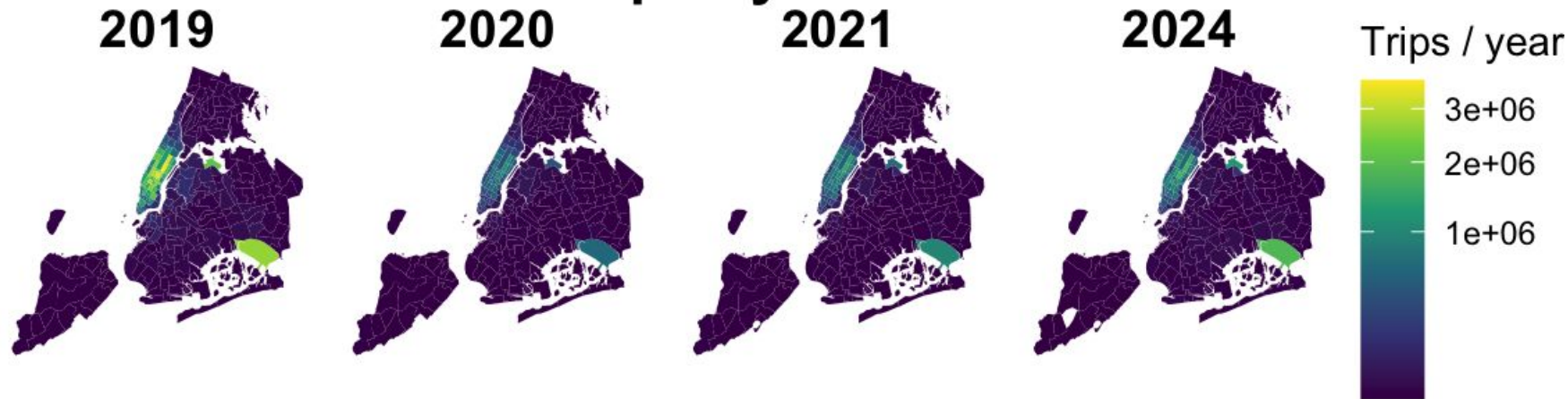


2021-01

Trips / month



NYC Yellow Cab Pickups by Zone — Annual Totals



Significant difference in the mean number of annual trips per zone in 2019 and in 2024

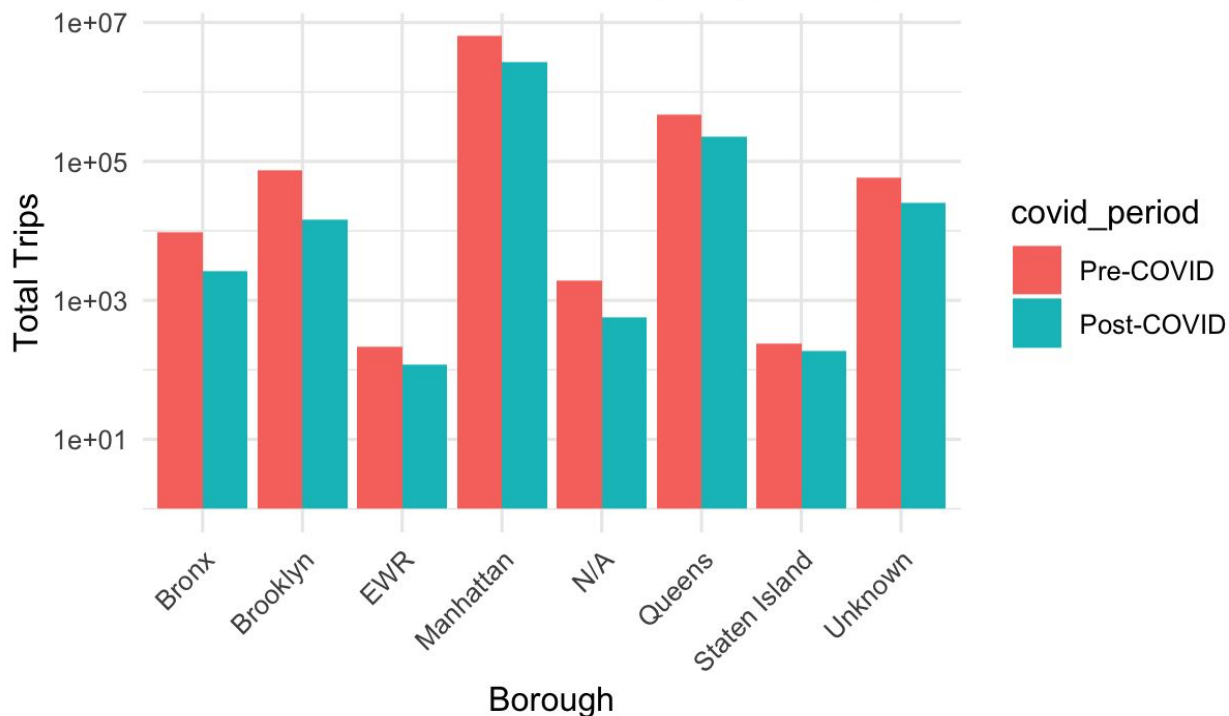
Paired t-test

```
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
```



Borough & Total Average Trips – pre and post COVID

Pre vs Post COVID: Total Trips by Borough



Pre-COVID: 2019


Post-COVID: 2021, 2022

Is this significant?

T-Test Analysis – Covid and Total Average Trips

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
```

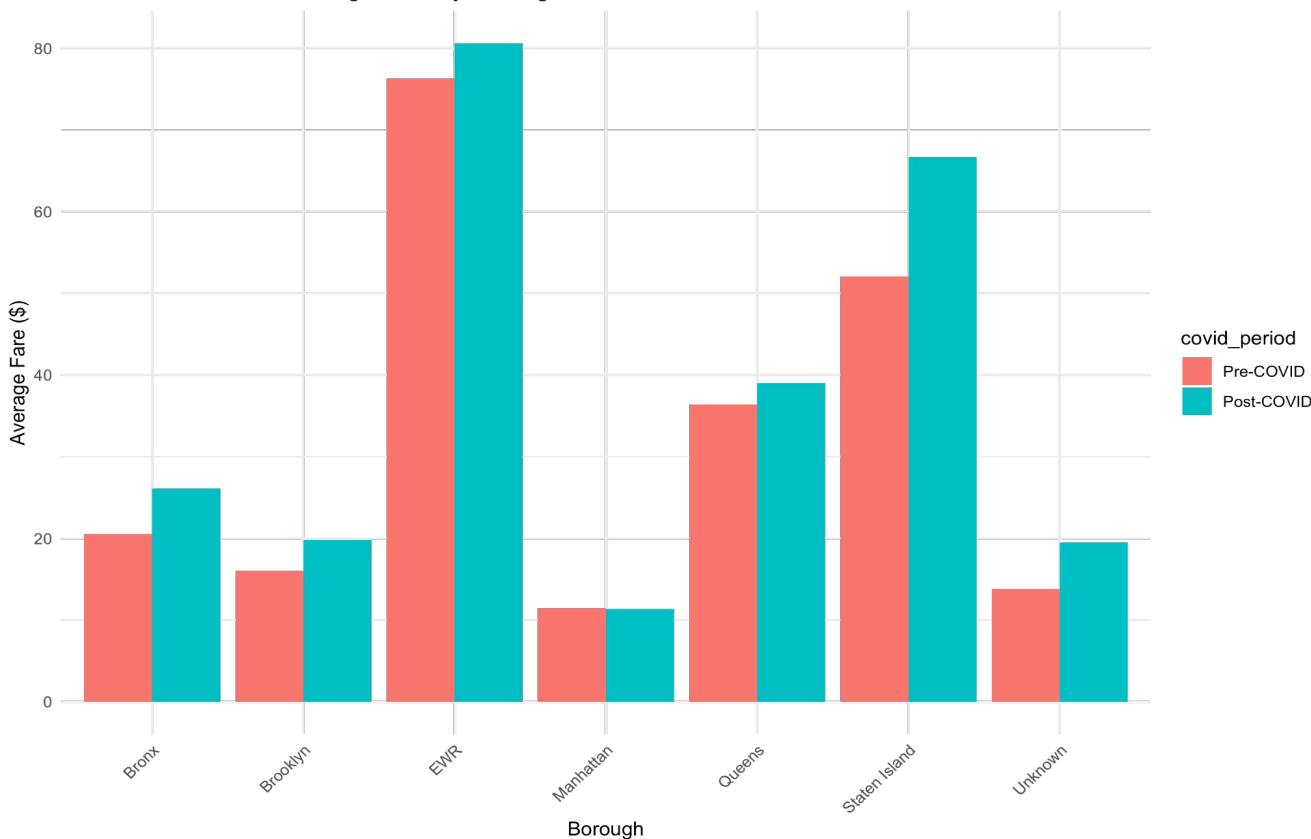


YES!

The paired t-test shows a **significant drop in trips post-COVID** ($p < 0.0001$), with an average decrease of **15,112 (pre_covid - post_covid) trips** (per zone per month)?

Borough & Average Fare/Trip - Pre and Post COVID

Pre vs Post COVID: Average Fare by Borough



Pre-COVID: 2019


Post-COVID: 2021, 2022

Is this significant?

T-Test Analysis-Covid & Average Fare/Trip

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
```



YES!

The paired t-test confirms **fares increased significantly post-COVID** ($p < 0.0001$), with an average rise of **\$5.24** per trip)per zone per month)?

Conclusion

- **Explicit decreases in trip duration during covid.**
 - **Borough-specific differences.**
 - **EWR has the highest average fare compared to all borough.**
- **Differences in rush-hour vs. non rush-hour cab demand.**
- **Explicit and significant increase in fare price after covid, with an average rise of \$5.24 per trip (per zone per month)?**
- **Explicit and significant decrease in trips after covid, with an average decrease of 15,112 trips per zone per month**
- **Unique borough and year differences in correlations between average fare and trip duration / distance.**