NYC Taxi data project

Presentation

Ruize Xuan

Outline

- Introduction
- Data Visualization
- Data Cleaning
- Idea / Data validation
- Method and model
- Result
- Conclusion

Introduction

The goal for this analysis to use the New York City Taxi dataset to evaluate the potential information leakage from the Federal Reserve around FOMC meetings along unofficial channel.

The filtered sample is from more than ten years taxi records in New York. Each row contains one taxi trip.

Data Visualization

I find:

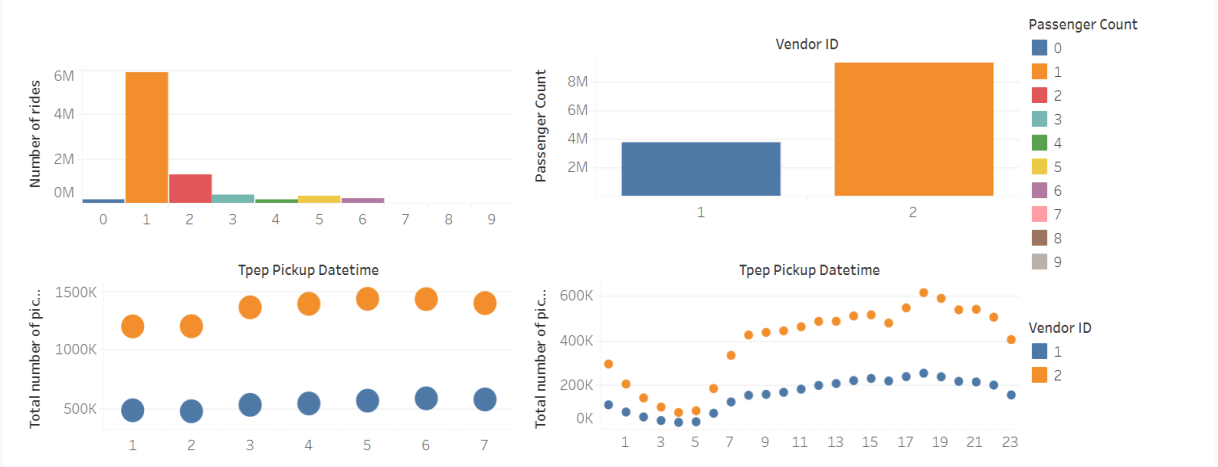
There are a few trips with zero, or seven to nine passengers.

The vast majority of rides had only a single passenger, with two passengers being the second most popular option.

Towards larger passenger numbers we are seeing a smooth decline through 3 to 4.

Vendor 2 has significantly more trips in this data than vendor 1.

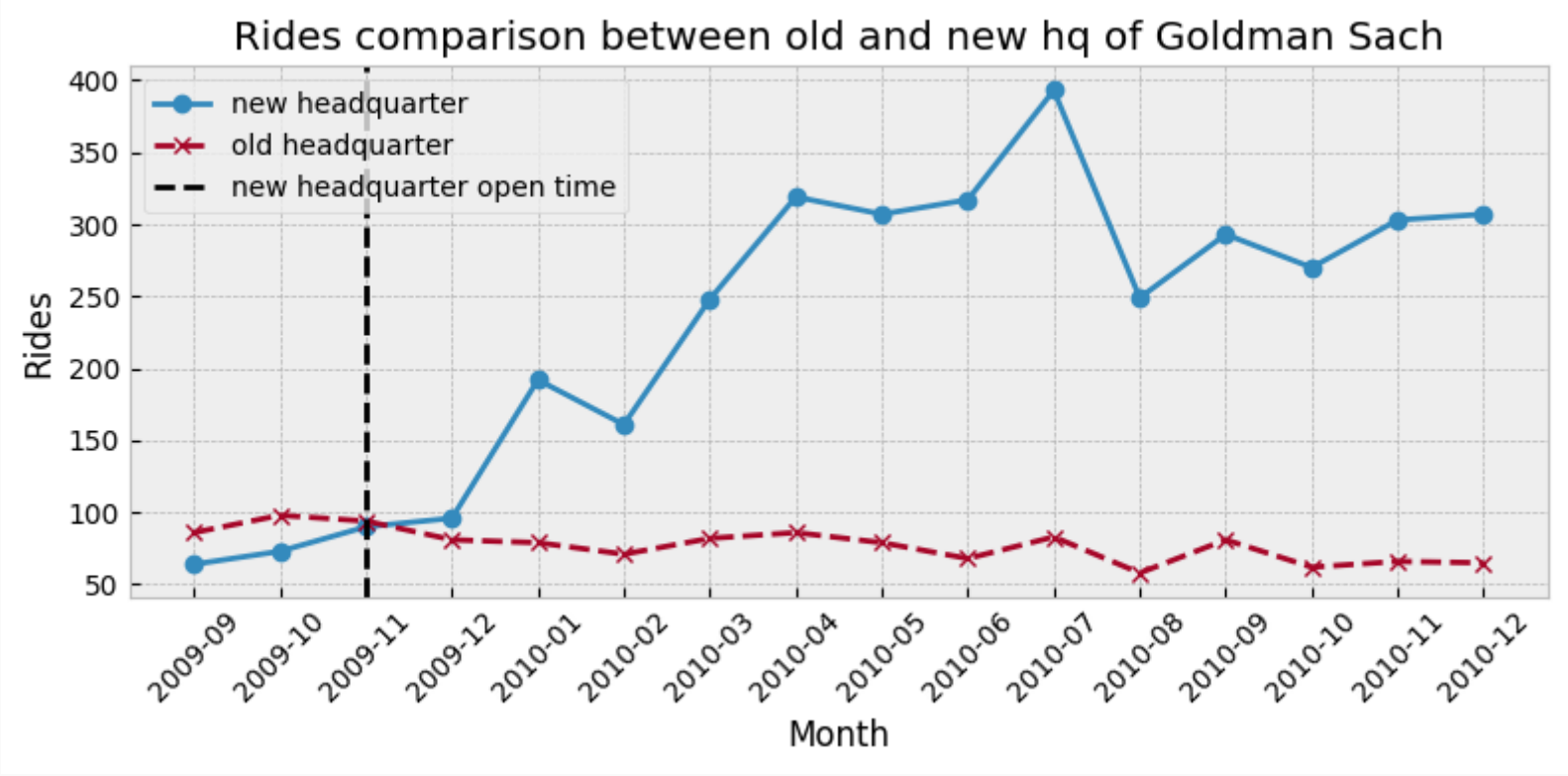
We find an interesting pattern with Monday being the quietest day and Friday very busy.



Data Cleaning

- Extreme trip duration
  - Longer than one day
    - These value should be removed
  - Close to 24 hours
    - We will remove trip_durations longer than 22 hours
  - Shorter than a few minutes
    - Zero distance
    - Short trips (limits)

Data Validation



Method and model

$$\lambda = \exp(\delta\beta + \alpha_{ym(t)} + \theta_{wd(t)})$$

It's a Poisson regression.

λ is the number of rides.

δ is the indicator if the date t is within FOMC window.

α_{ym} is the effect of year of month.

θ_{wd} is the effect of weekday.

Result

Generalized Linear Model Regression Results

Dep. Variable:	Passenger_Count	No. Observations:	16910
Model:	GLM	Df Residuals:	16907
Model Family:	Poisson	Df Model:	2
Link Function:	Log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-25481.
Date:	Thu, 23 May 2024	Deviance:	12393.
Time:	02:18:17	Pearson chi2:	1.72e+04
No. Iterations:	5	Pseudo R-squ. (CS):	-0.02141
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Month	0.0351	0.001	30.378	0.000	0.033	0.037
Weekday	0.0632	0.003	21.543	0.000	0.057	0.069
FOMC	0.1116	0.028	3.943	0.000	0.056	0.167

Conclusion

- Employed a rich dataset constructed from anonymous trip-level taxi data to examine interactions between insiders of the Federal Reserve Bank of New York and major commercial banks around FOMC meetings.
- Found evidence suggestive of an increase in rides between them both at the New York Fed's offices and in areas.
- More improvement can be made by further segmenting time of day, especially during lunch and dinner time.