Fare Forecaster

Save Time, Make Money

CONTENTS

INTRO

Dillon Diatlo
Data Scientist
dillondiatlo@gmail.com

INTRO

Dillon Diatlo

Data Scientist

dillondiatlo@gmail.com

CREDENTIALS

Lyft Rating

4.90

Uber Rating

4.78

Taxi Rating

N/A

OBJECTIVE

OBJECTIVE

A

B

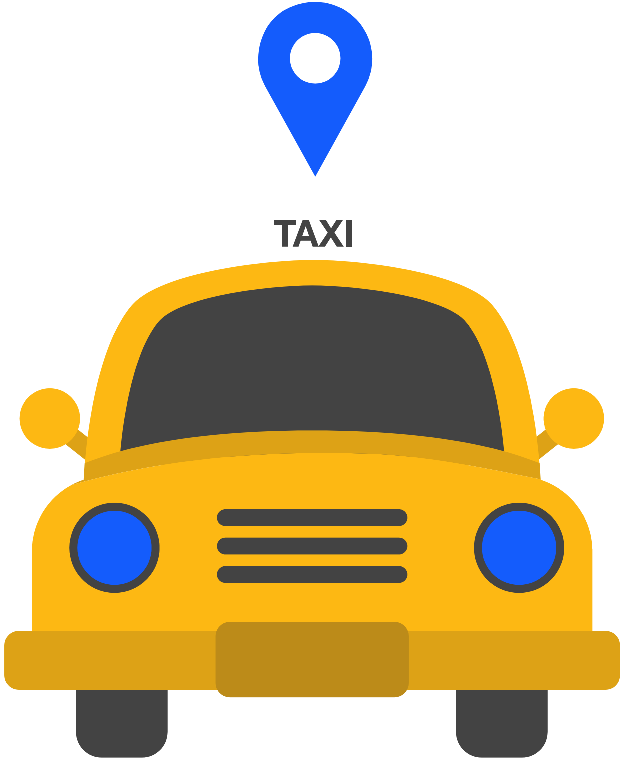
CHALLENGE

Determine where in NYC for-hire vehicle drivers can make the most money per trip, at that moment.

OBJECTIVE

Sample 212M rows of 2022 NYC for-hire-vehicle trip data to predict total driver revenue by trip.

Deploy an app to help drivers determine which zone to go to for the highest average fare.



EDA

PIT STOP: DOWNSIZING

212,000,000

- Slow to download
- Forever to analyze

MONTH-A | MONTH-B

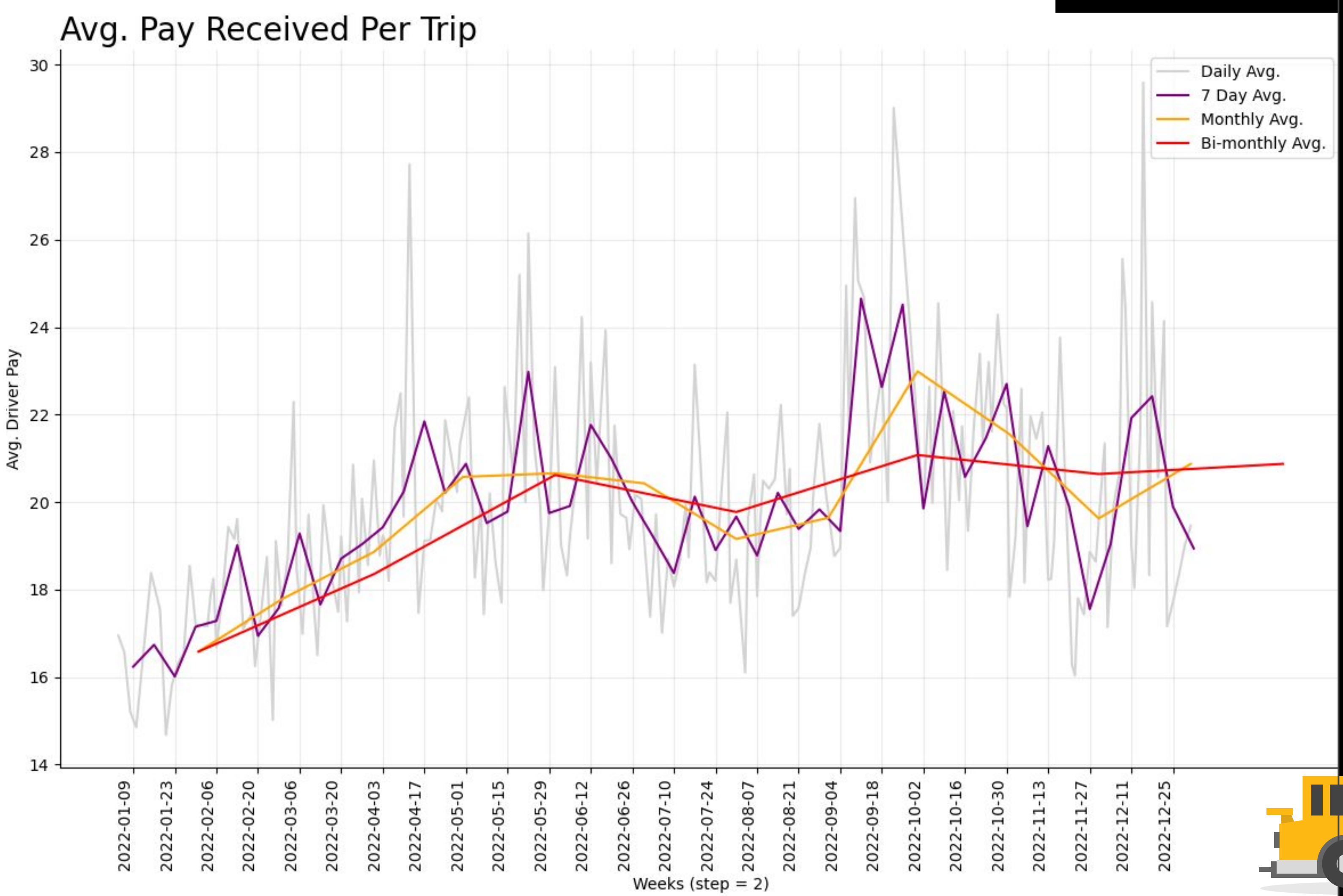
- Split ‘em up
- Chunksize param

CONCATENATE

- Put ‘em back together by month

READ | CREATE

- One BIG beautiful df
- ~4M rows



Outliers

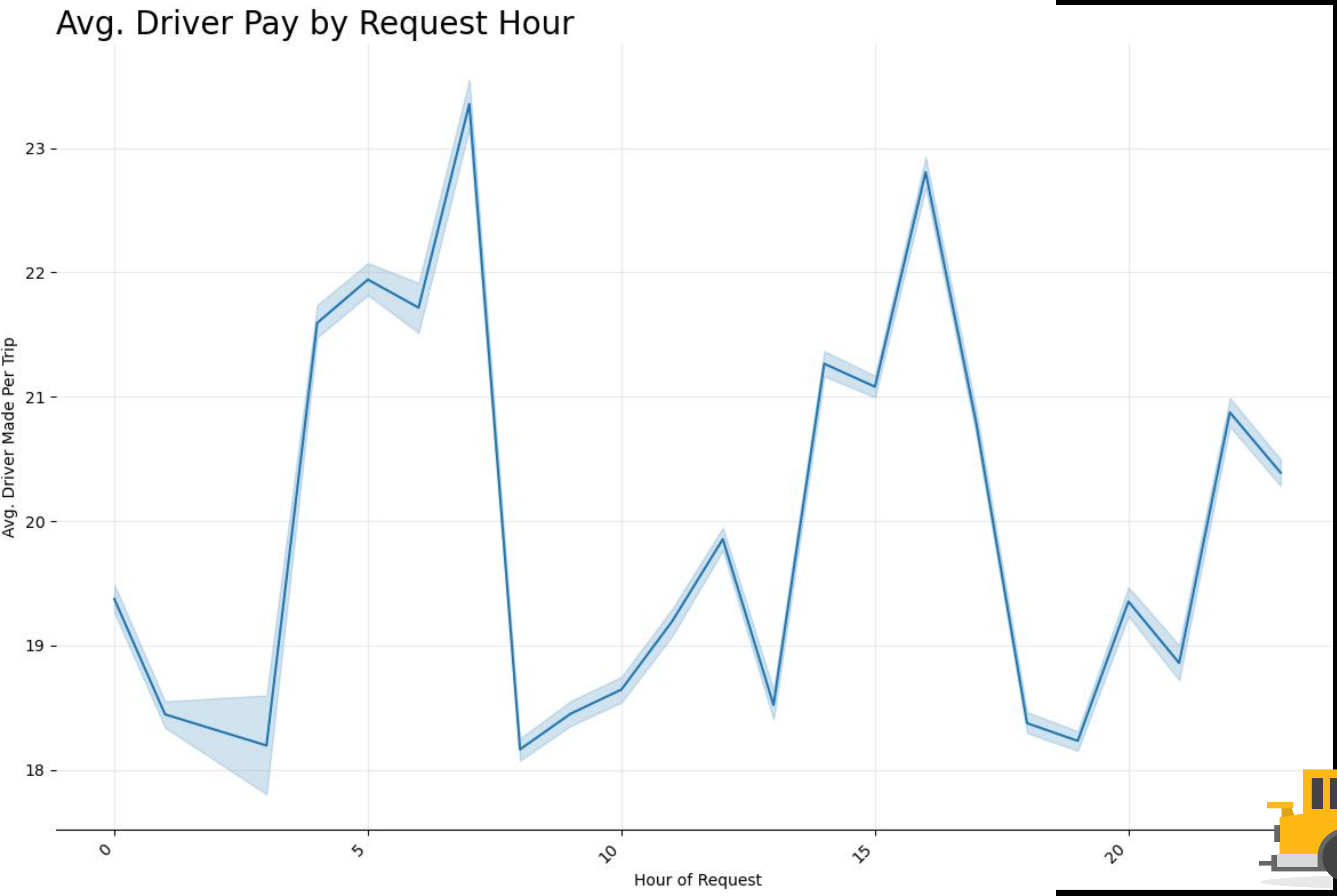
- A few daily outliers can skew

Plateau

- Increase until May
- Plateau could be holidays

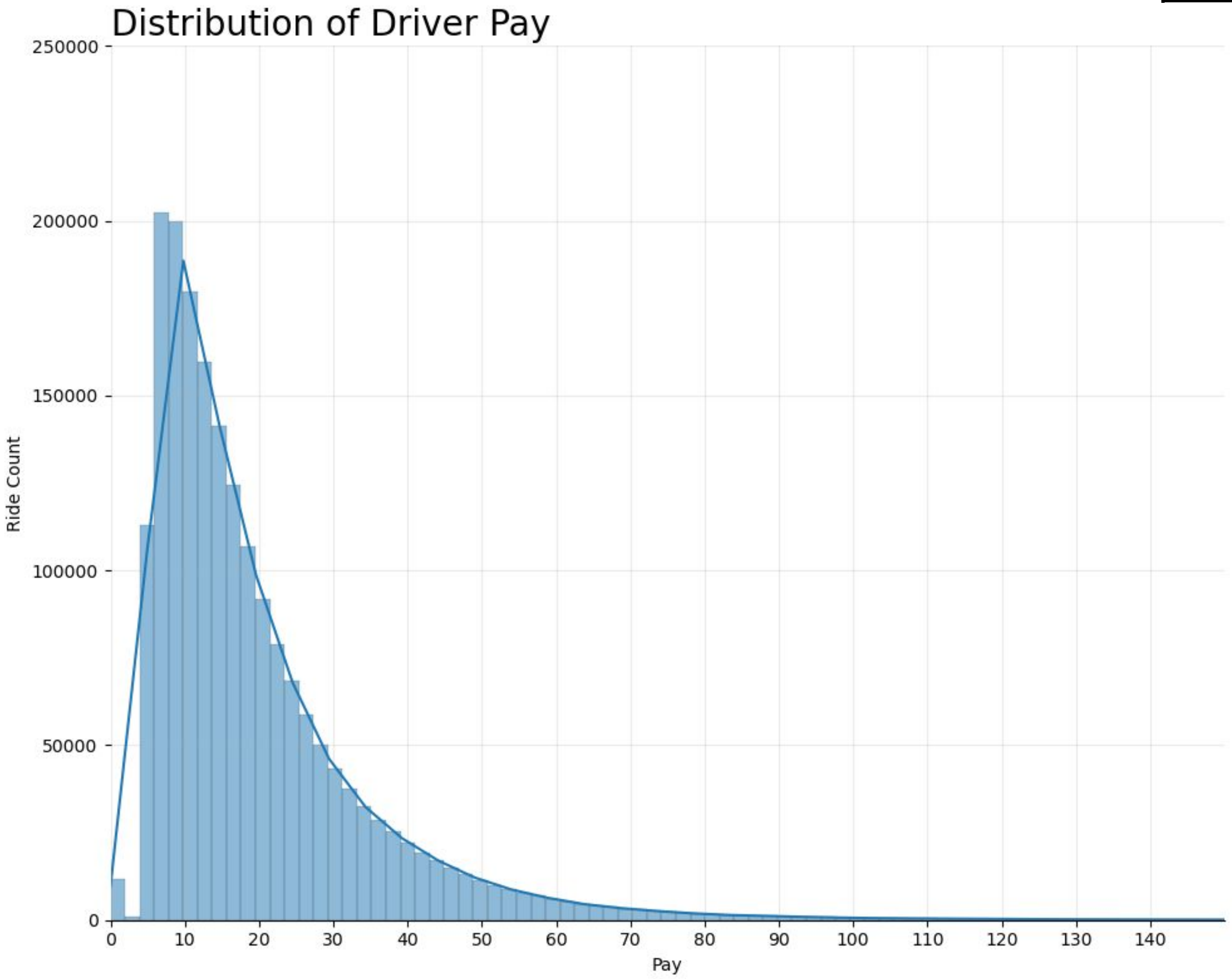
Seasonality

- Difficult to track w/ 1 year of data



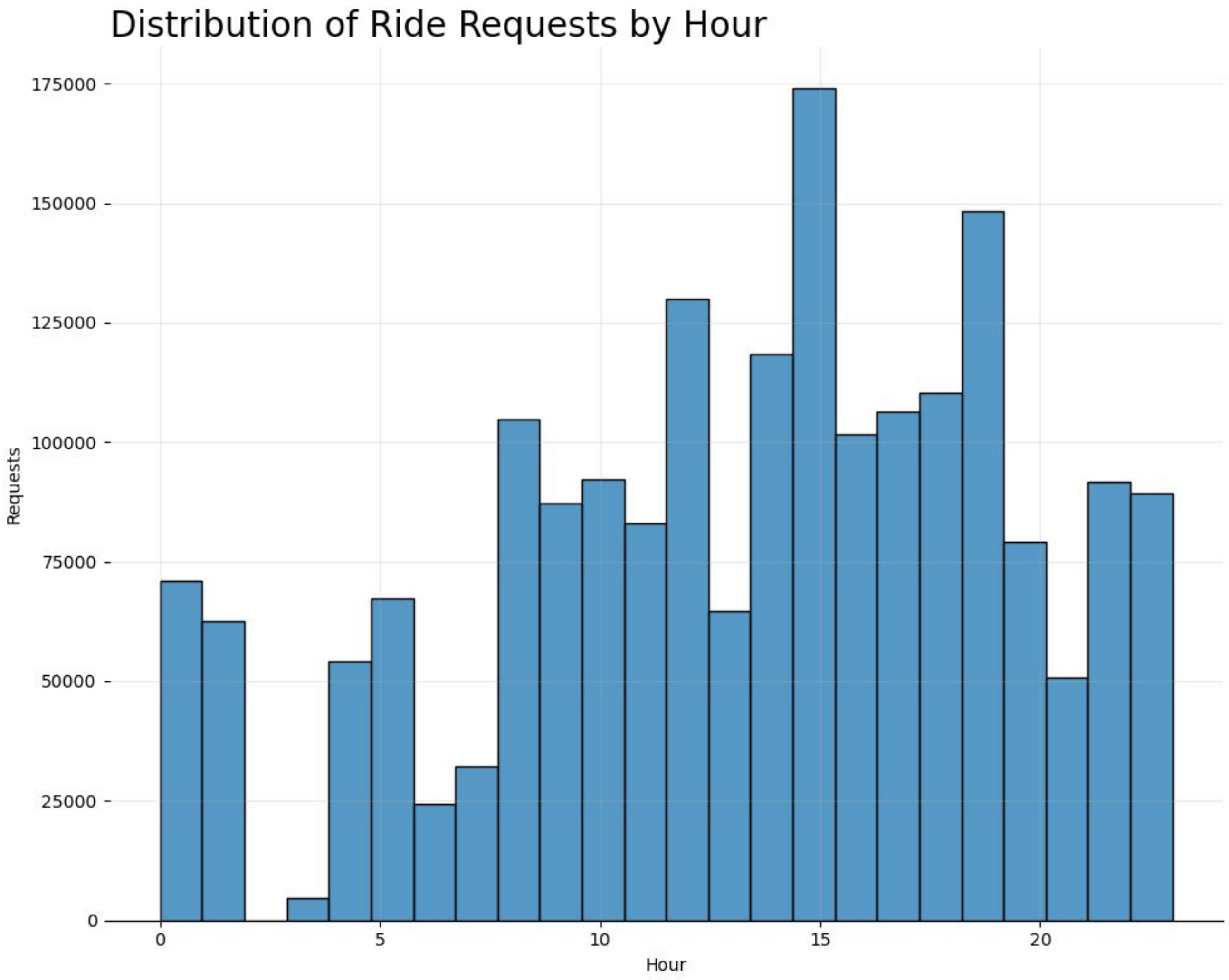
Bi-modal

- 7am & 5pm



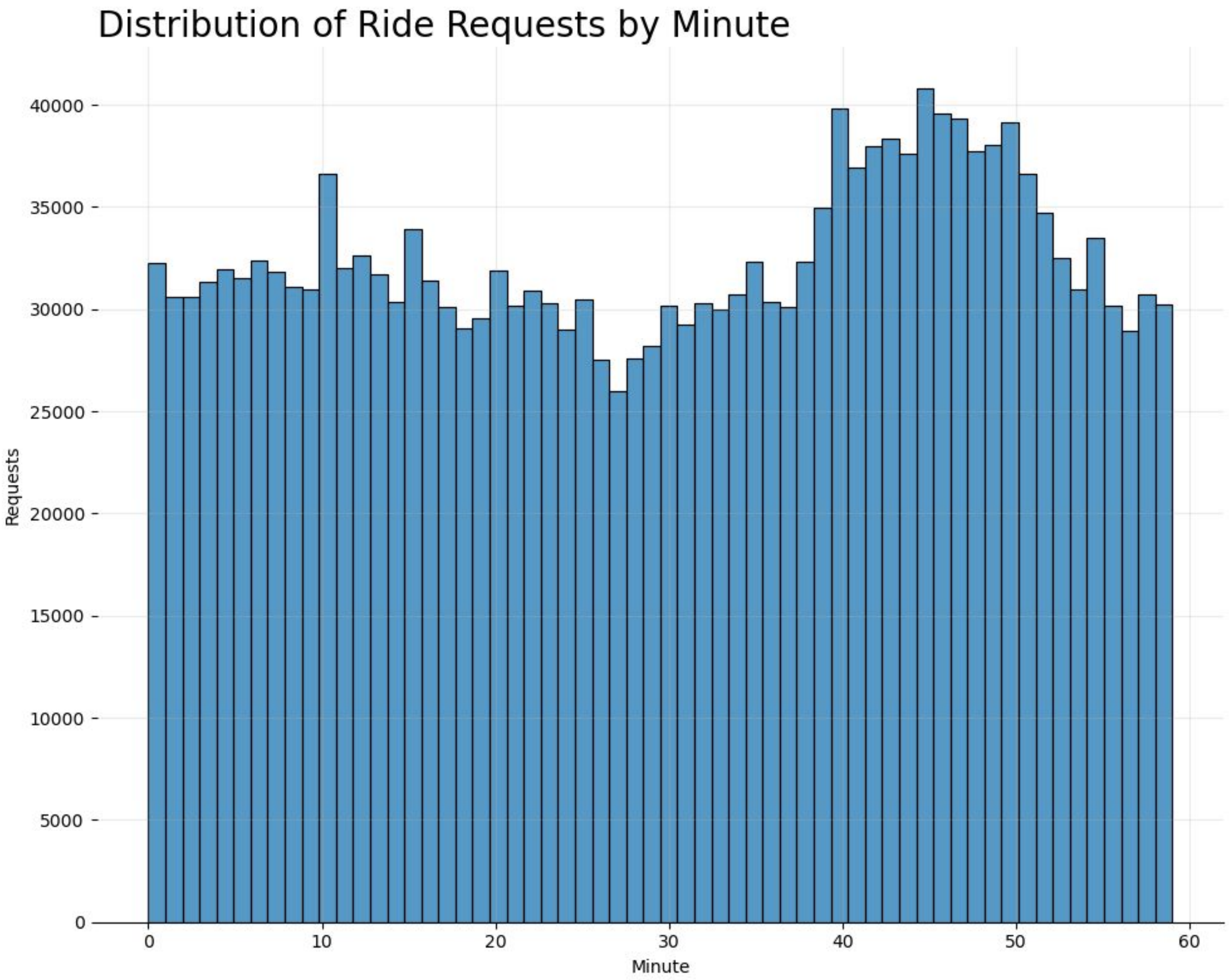
Distribution

- Right skewed
- Most trips between $5-20



Distribution

- Left skew
- Less trips in the AM

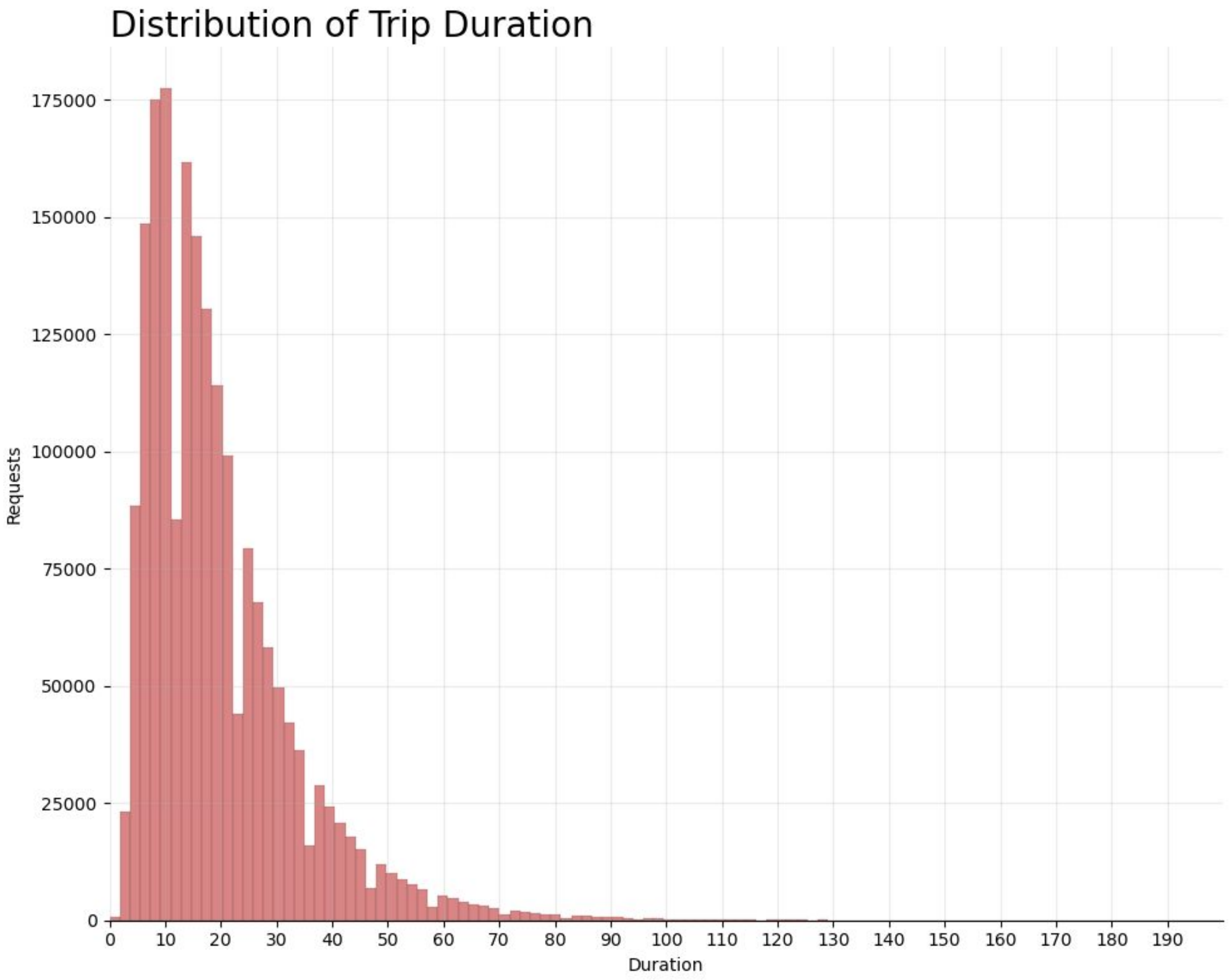


High Passenger Time

- 4:40pm - 4:55pm

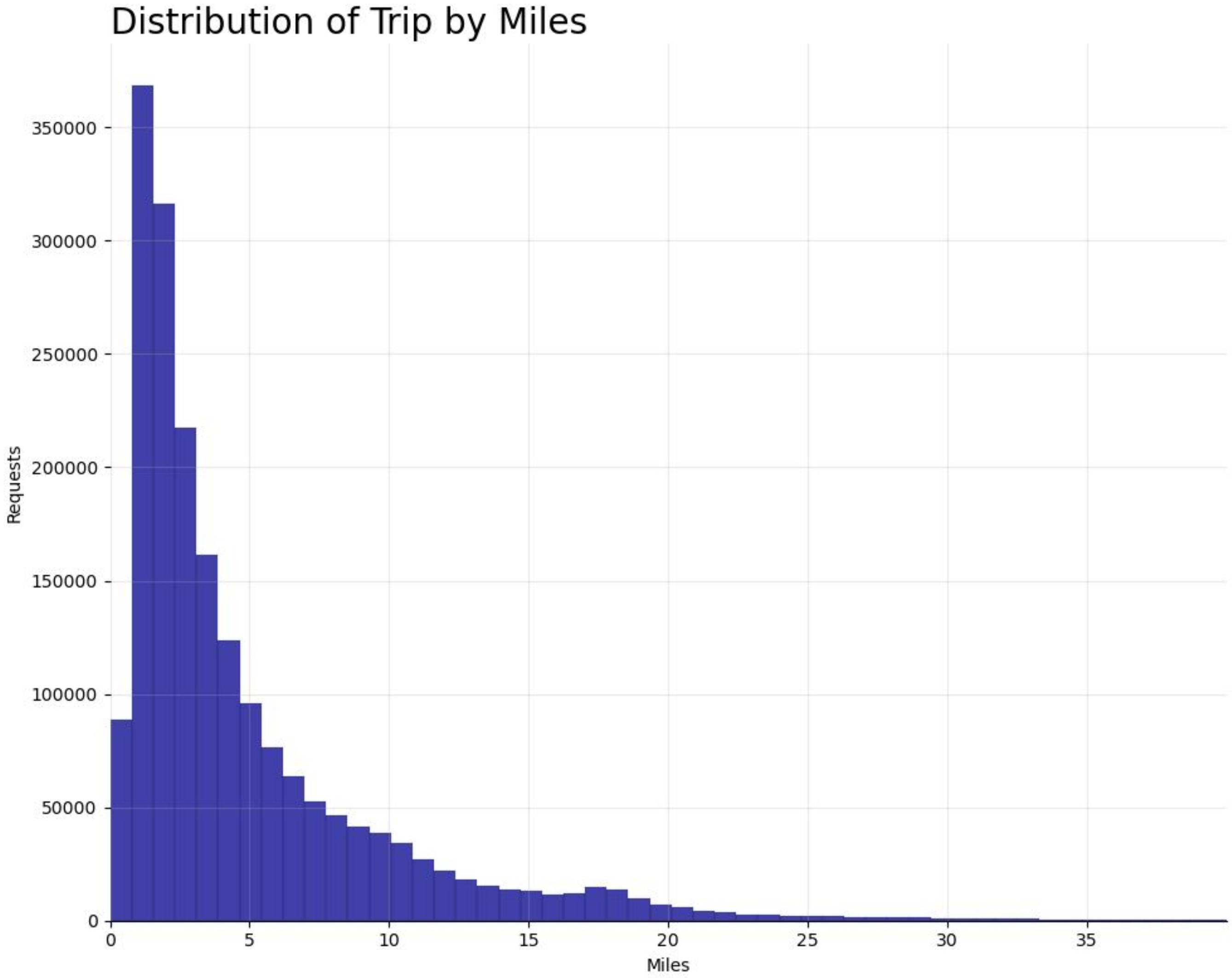
Distribution

- Almost bi-modal
- Mid-hour dip
- Most between 40min - 55min



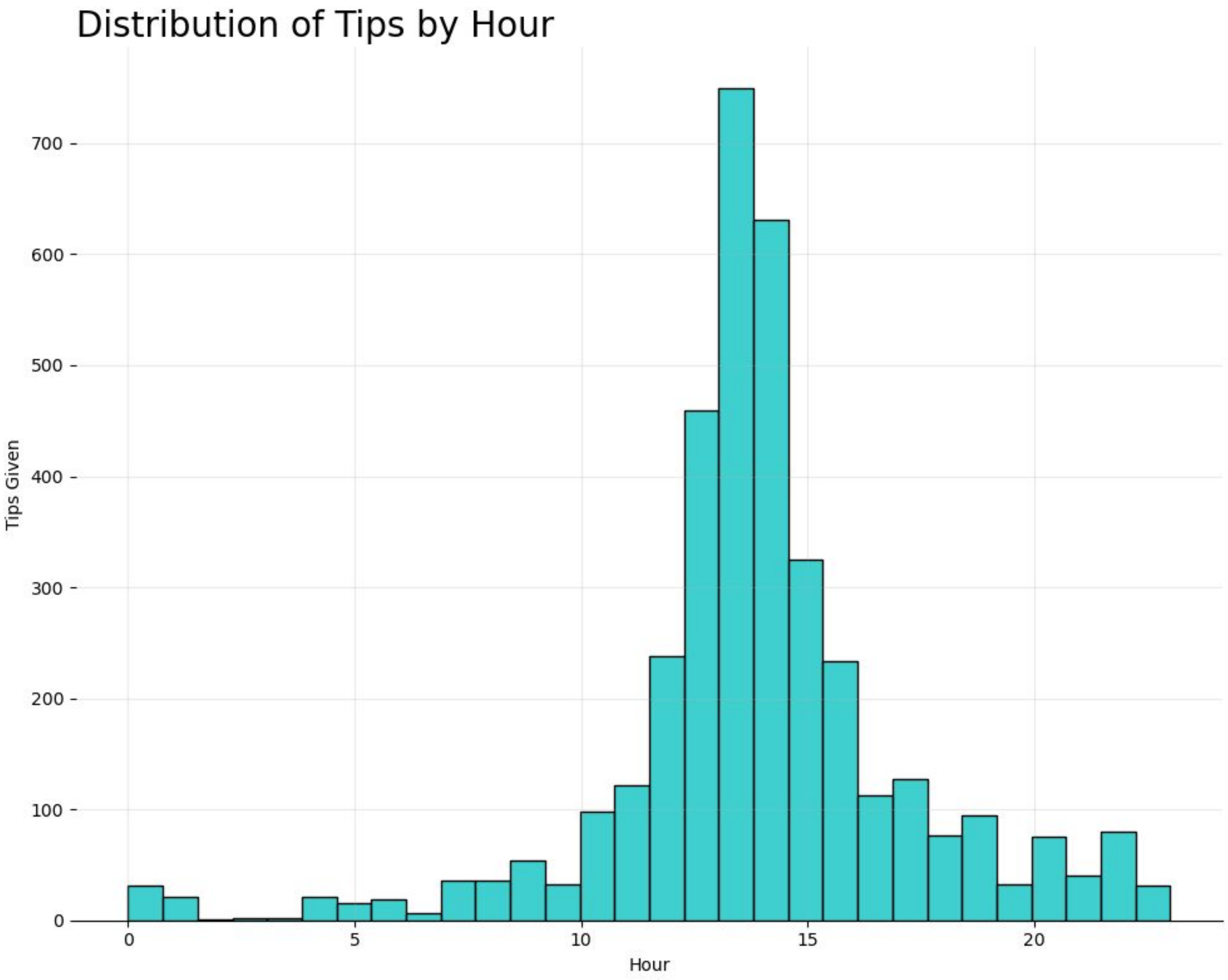
Distribution

- Right skewed
- Pattern of dips and peaks every 10-15 minutes



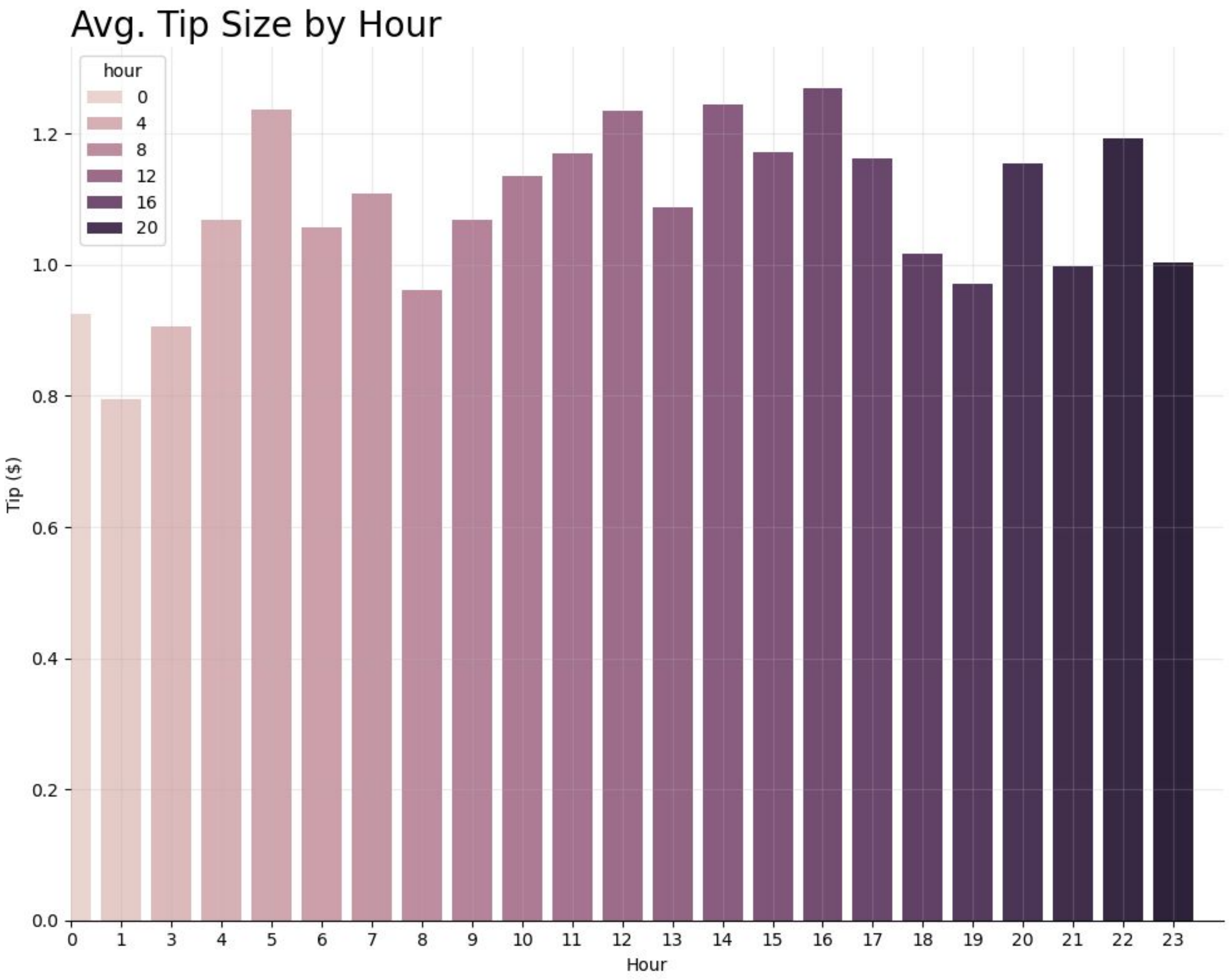
Distribution

- Right skewed
- 2-10 miles



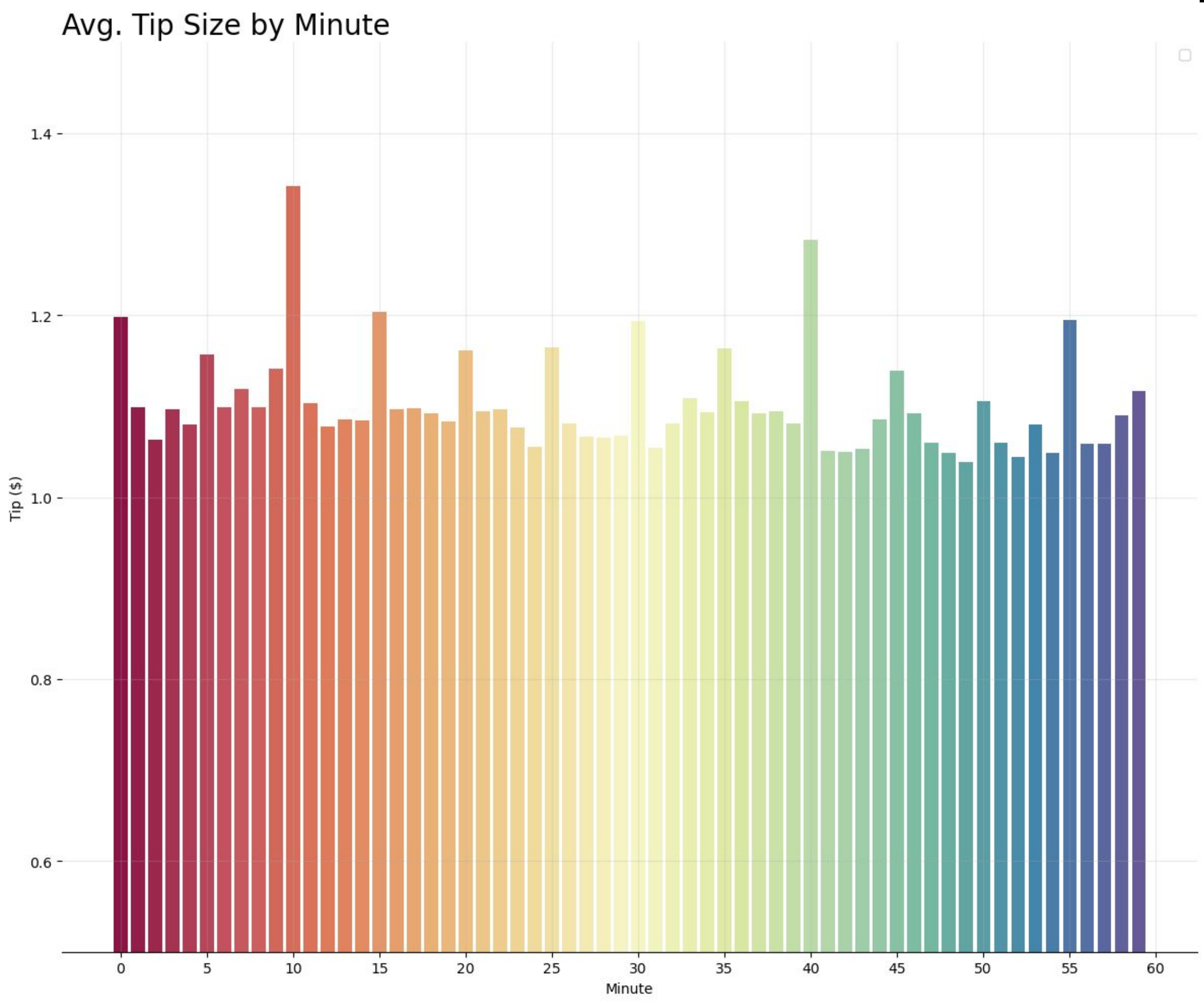
Distribution

- Closer to normal distribution, though a bit left skewed
- Follows similar distribution of trips per hour



Best Times for Bigger Tips

- 5am, 5pm

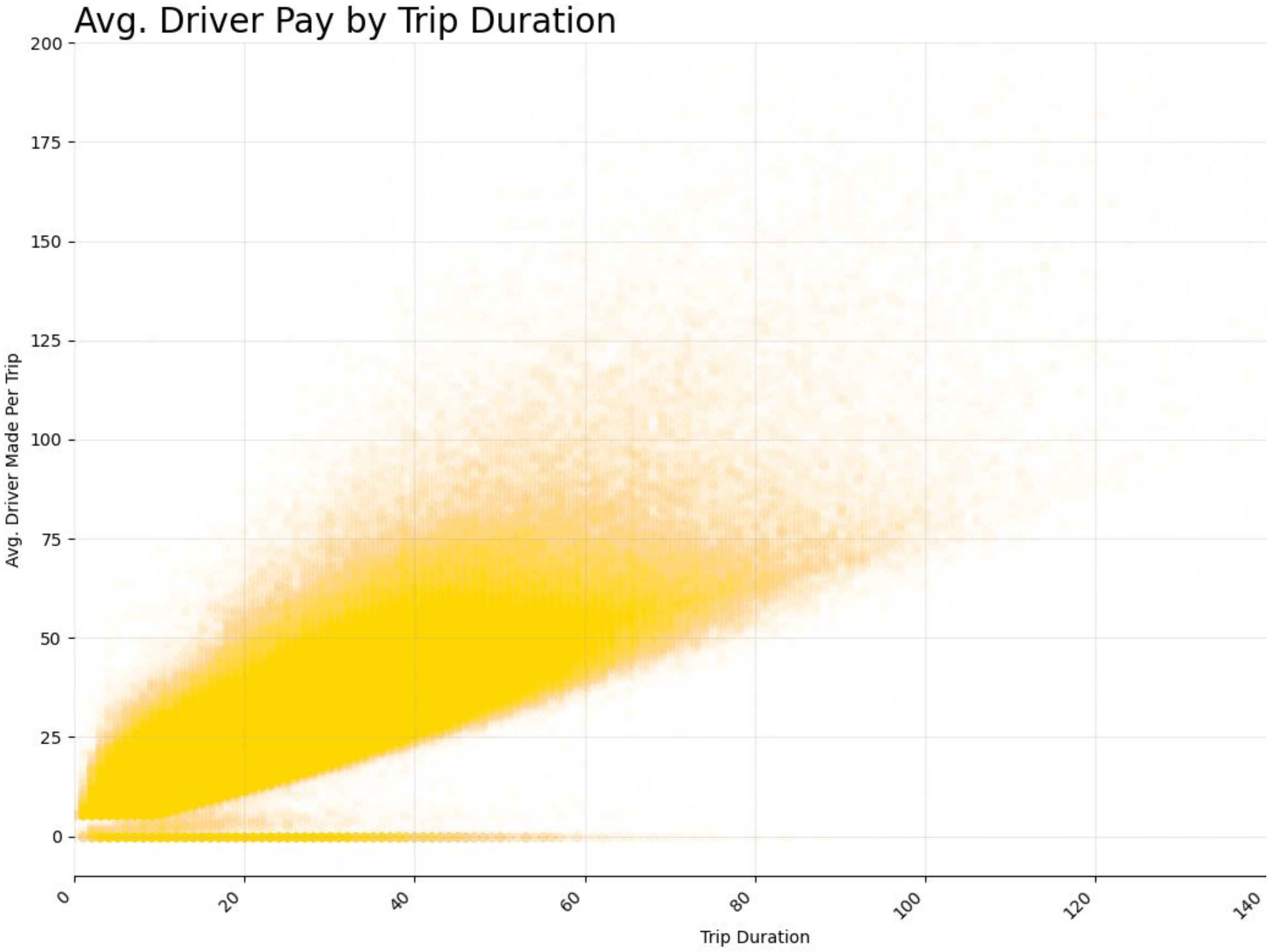


Biggest Tips Happen

- 10 minutes past the hour
40 minutes past the hour

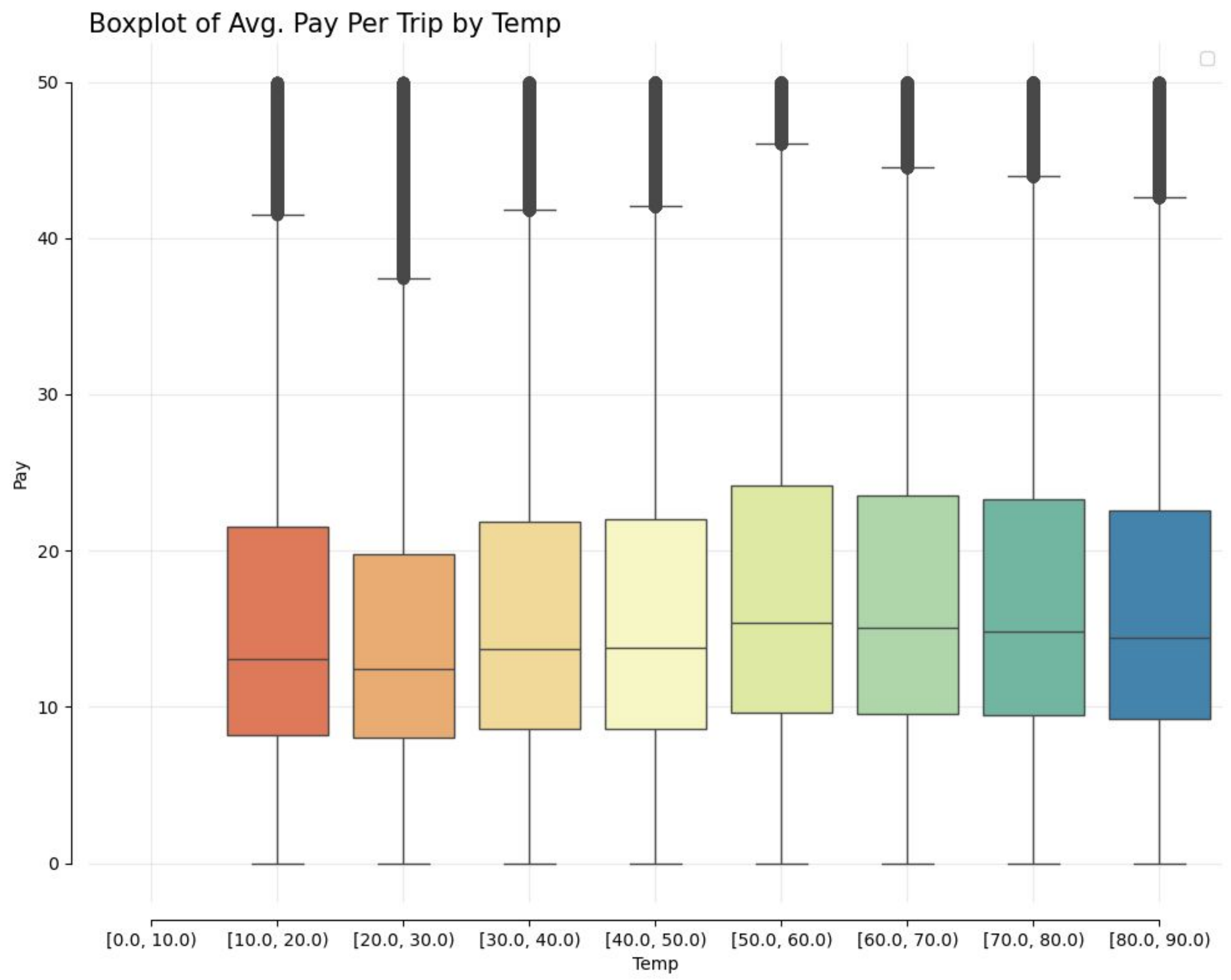
Patterns

- Avg size of tips seem to peak every 5 minutes



Patterns

- Affirms expectations of positive correlation
- Discover of negative pay, which then went back to fix



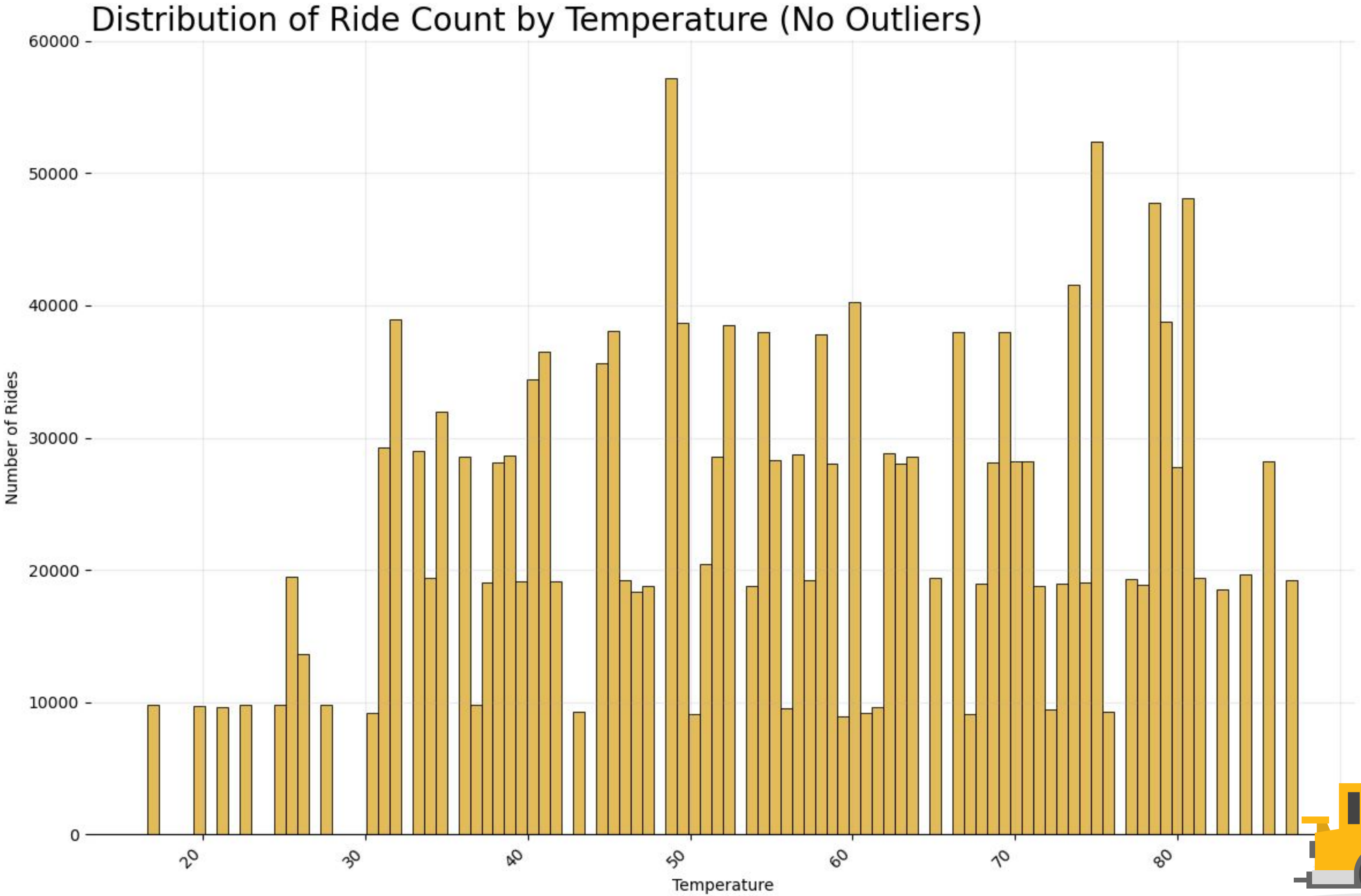
Biggest Tips Happen

- 50-60 degrees
- could be more trips at this time
- excited to get out after winter

Patterns

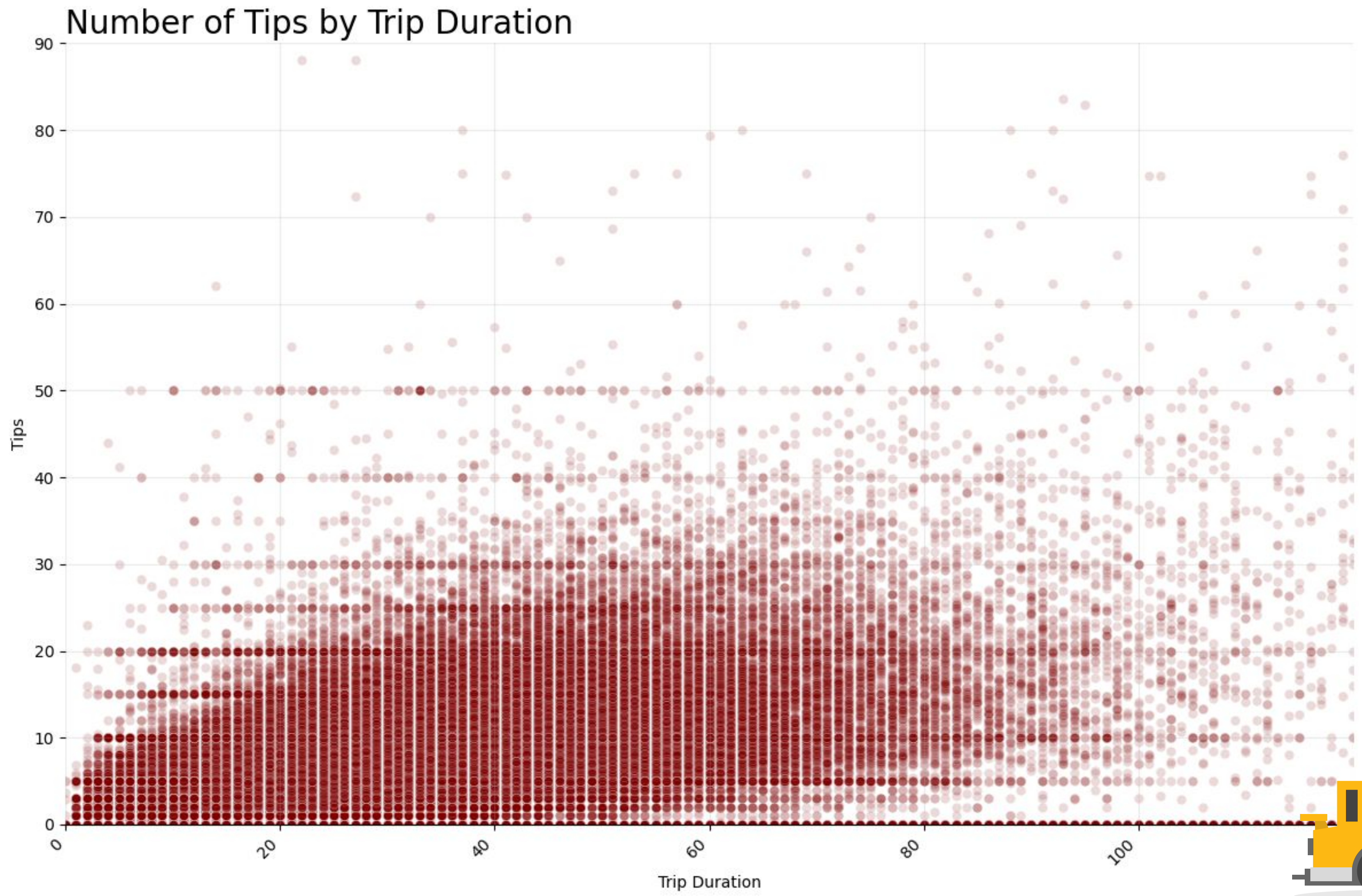
- Almost uniform distribution

Ever so slight positive correlation



Patterns

- Slight left skew
-bi-modal
-peak at 48/49 and 75

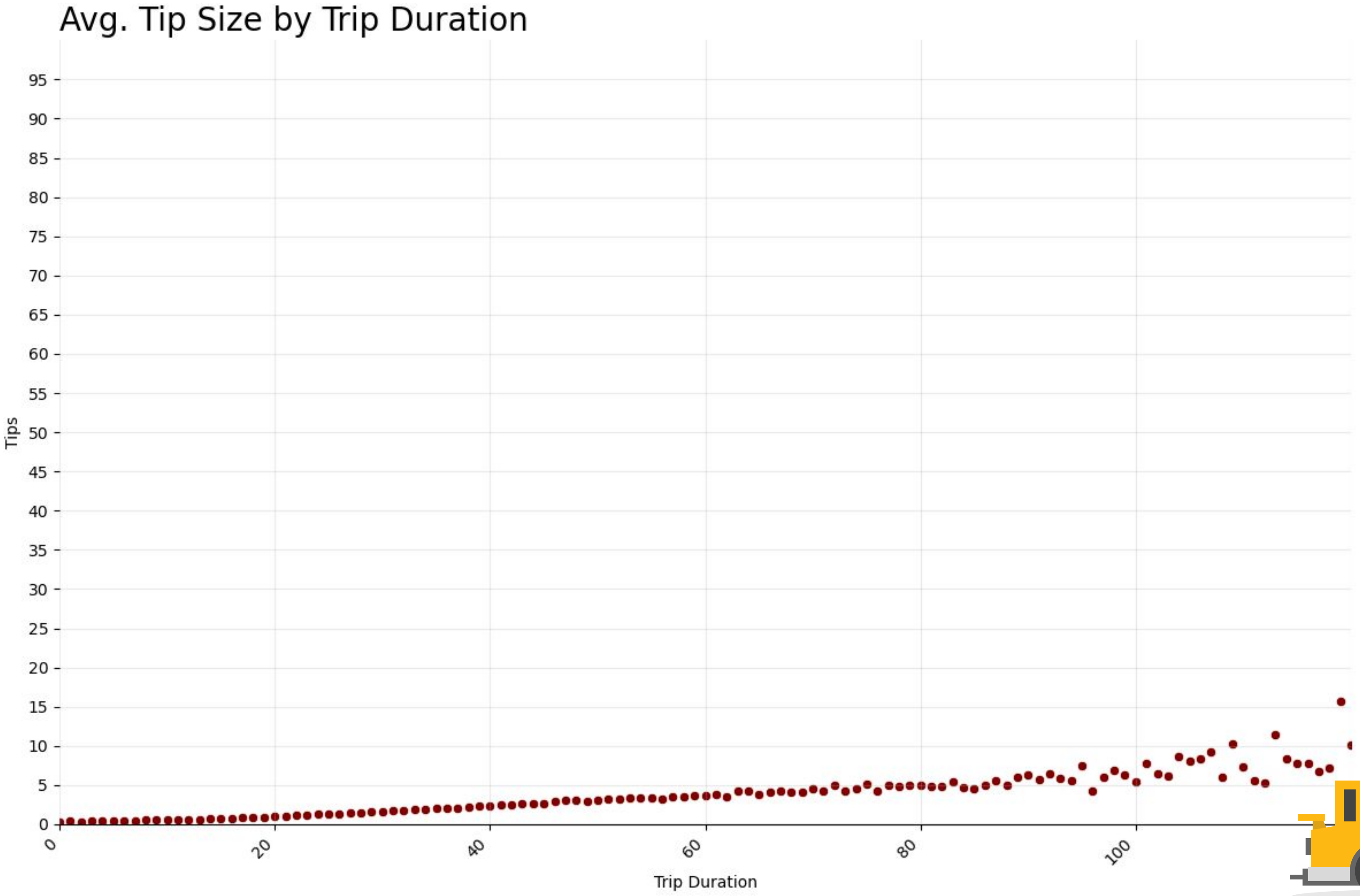


Biggest Tips Happen

- As duration goes up number of tips goes up
- Majority of tips shrink after about 60-70 minutes

Patterns

- Horizontal lines cutting across



Patterns

- Tip size increases by trip duration
- plateaus a bit at 60, similar to before
- Tip size begins to flatten out rarely above $15 avg

MODELS

	TYPE	Preprocessor	Params	Metric SCORE
MODEL 1	GradientBoosting-Regressor	ColumnTransformer OneHotEncoder	random_state=2024	**Train r^2: 86.8%** **Test r^2: 87.4%** **Train RMSE: $5.90** **Test RMSE: $6.00**
MODEL 2	RandomForest-Regressor	ColumnTransformer OneHotEncoder StandardScaler	n_estimators=250 max_depth=30 min_samples_split=300 max_features='sqrt' n_jobs=4)	**Train r^2: 81.2%** **Test r^2: 78.6%**
MODEL 3	LassoCV	StandardScaler	alphas= np.logspace(-3, 0, 100) cv=5 max_iter=10	**Train r^2: 86%** **Test r^2: 86%**
MODEL 4	XGBoostRegressor	ColumnTransformer OneHotEncoder	n_estimators=500 max_depth=10 min_samples_split=200 min_child_weight=1 max_features=TKTKTKTKT enable_categorical=True	**Test r^2: 12%** **Test RMSE: $15.95**

WINNER

GradientBoostingRegressor

LinearRegression Baseline

-6%

Train RMSE: $5.90
Test RMSE: $6.00

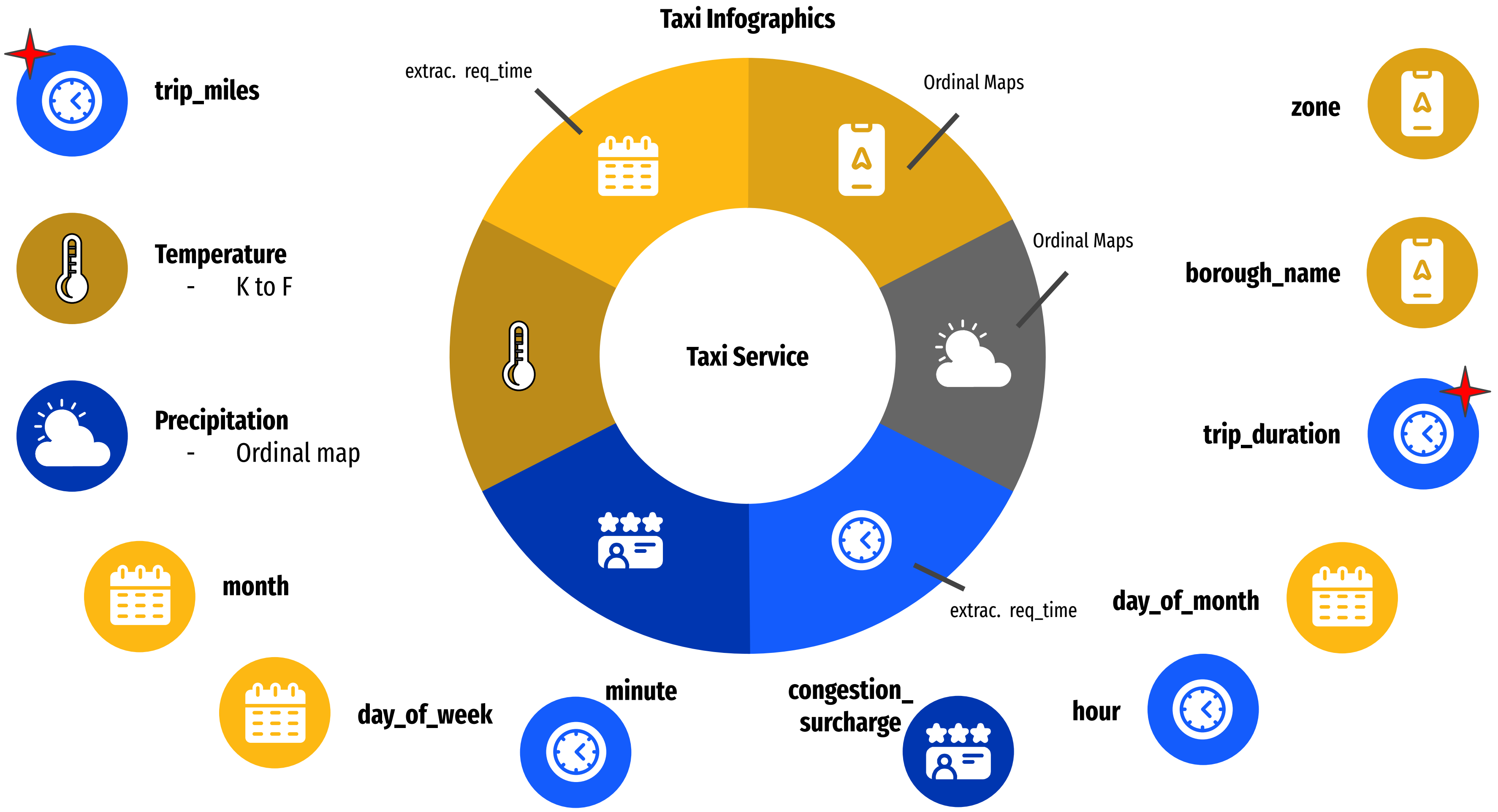
87.4%

Of the variability of average driver revenue per trip can be explained by the features in this model

Features

1. trip_miles
2. temp
3. preciptype
4. zone
5. borough_name
6. trip_duration
7. month
8. day_of_month
9. day_of_week
10. hour
11. minute
12. congestion_surcharge

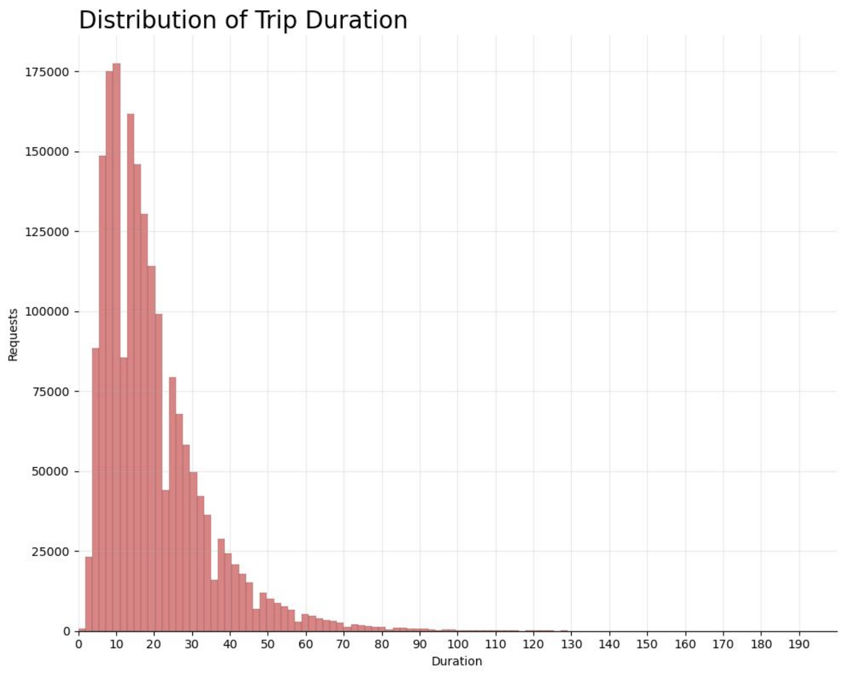
APP



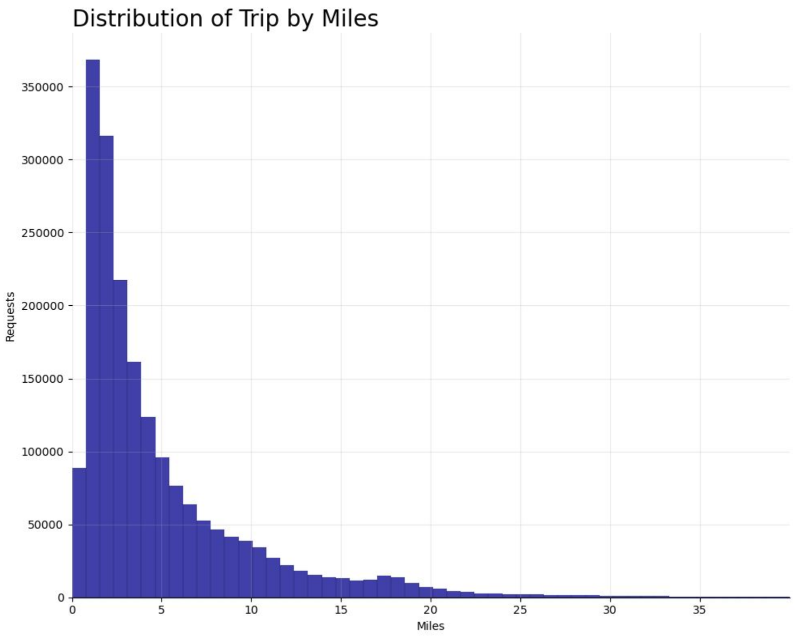
Taxi Infographics
Taxi Service
extrac. req_time
Ordinal Maps
Ordinal Maps
extrac. req_time
trip_miles
Temperature
- K to F
Precipitation
- Ordinal map
month
day_of_week
minute
congestion_
surcharge
hour
day_of_month
trip_duration
borough_name
zone

FEATURES

5-15 minutes

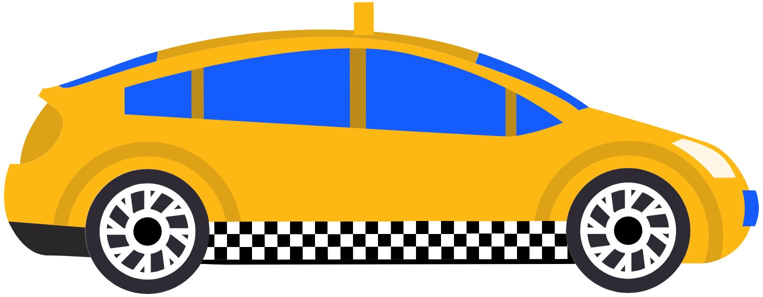


2-10 miles



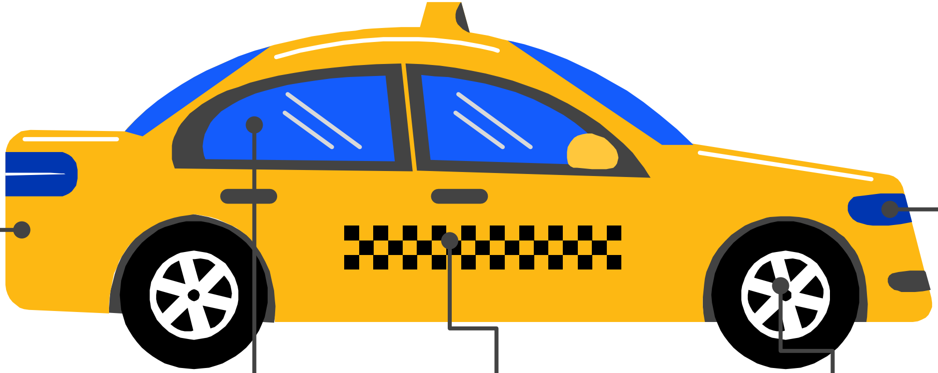
DEMONSTRATION

[Click Here to Try]



CONCLUSION

NEXT STEPS/RECS



1

TOOLS

Run experiment again w/ Spark, BigQuery, etc.

2

GEO-GRANULAR

Mapped by taxi zone, can probably do by block

3

TIME FRAME

Get data from multiple years

4

BIASIS

Initially not, but downsizing made it more bias

5

DEMAND

In future versions, include demand data

THANK YOU!

Questions?

Dillon Diatlo

Data Scientist

dillondiatlo@gmail.com
[Portfolio]()