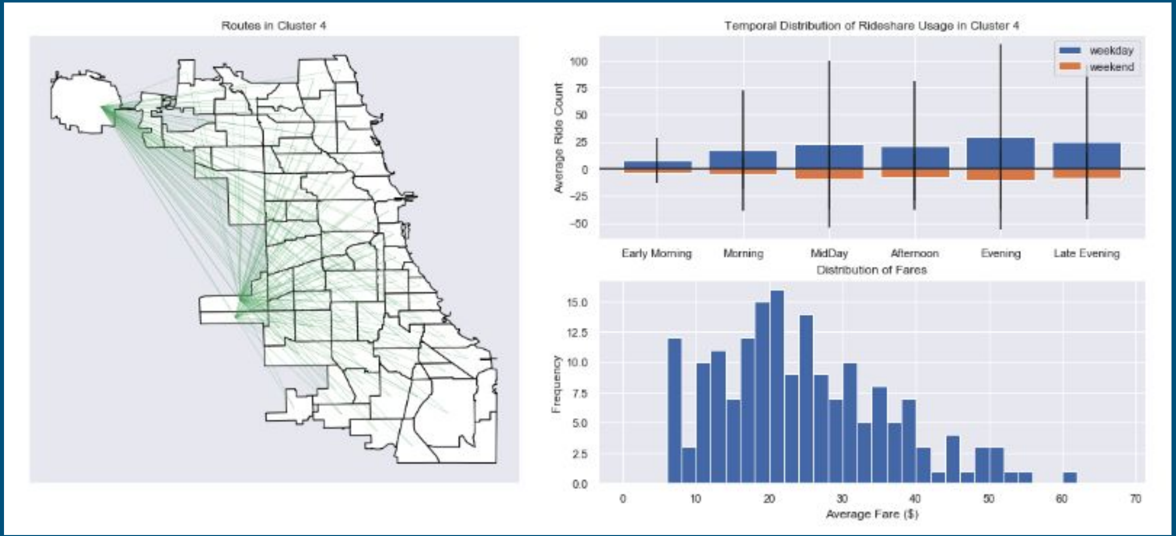
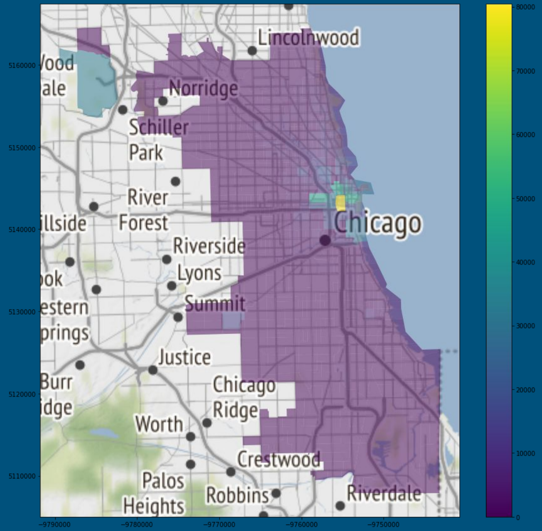




Machine Learning with the Chicago Ridesharing Trips Data Set

Lisa Taylor
Completed as Capstone Assignment 1
Springboard Data Science Bootcamp

Introduction

- Motivation
- Relevance
- Project Overview
- Data Acquisition and Cleaning
- Supplemental Data
- EDA
- ML applications
    - Clustering
    - Regression
- Final Thoughts

Trips Data Acquisition and Cleaning

- Trips
  - 12 GB csv file, 11/2018 - 3/2019, 40 M trips
  - Fields: pickup time/loc, dropoff time/loc, fare, tip, duration, distance, pooling
  - Anonymized and rounded
- Loading
  - “Chunksize”
  - 5% sample
- Cleaning and Enhancement
  - Remove unusable records (zero fare, no census tract)
  - Round datetimes to the hour
  - Derived datetime integer fields: DOW, Year, Month, Hour

Weather

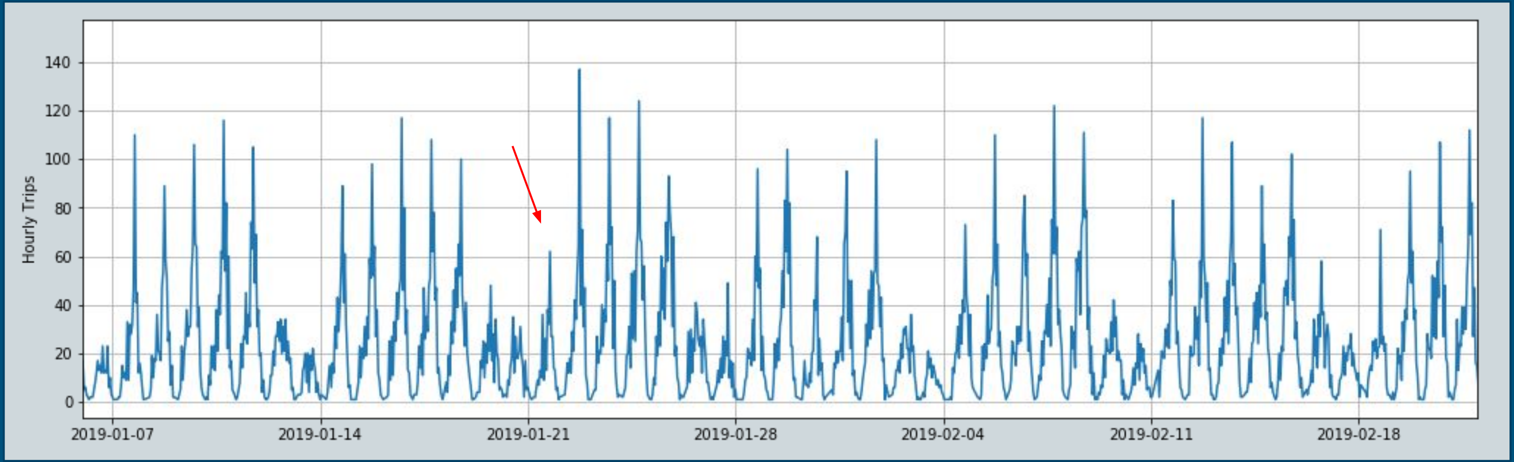
- Hypothesis: Weather influences rideshare demand
- NOAA data for Midway Airport (csv)
  - Fields: Precipitation, temperature, and wind speed
  - Divide into hourly and daily
  - Remove duplicates
  - Merge with rides

Census and Geospatial Data

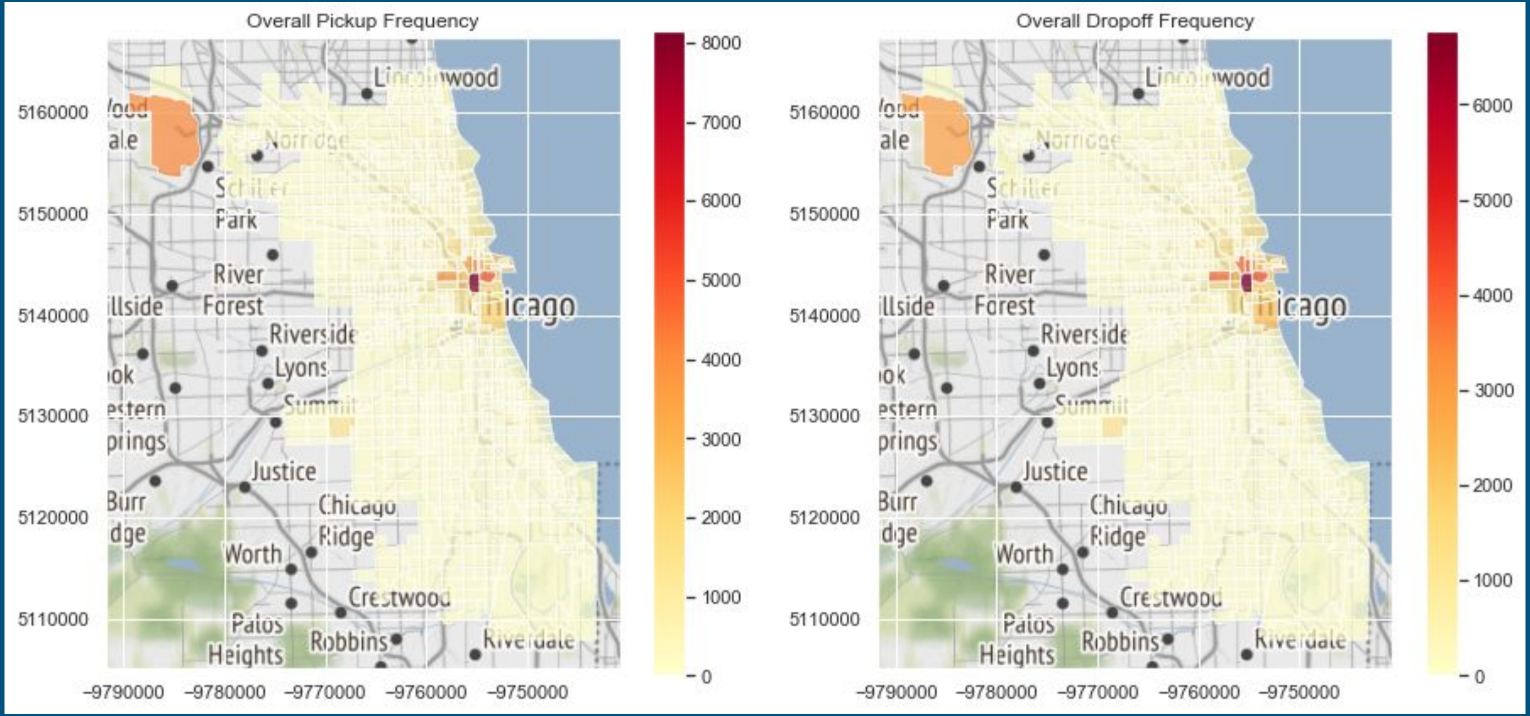
- Proxy for driver information
- Data Sources:
  - Census tract characteristics for 2017 (census API, json→ df)
    - Median income, Population
  - Census tract and community area polygons (geojson → geopandas gdf)
- Derived fields:
  - Distance from downtown, direction relative to downtown (bearing), population density
- Cleaning
  - Missing income -> 0, check outliers
- Merge to rides data on ride pickup tract

EDA Findings: Temporal Pattern

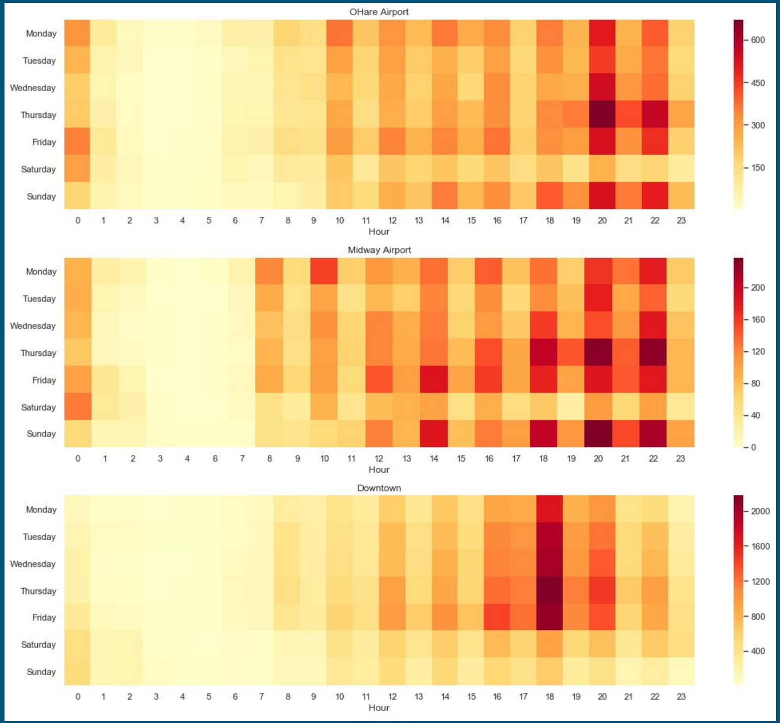
- Usage is cyclical
- Daily pattern superimposed on weekly pattern
- Holiday effects

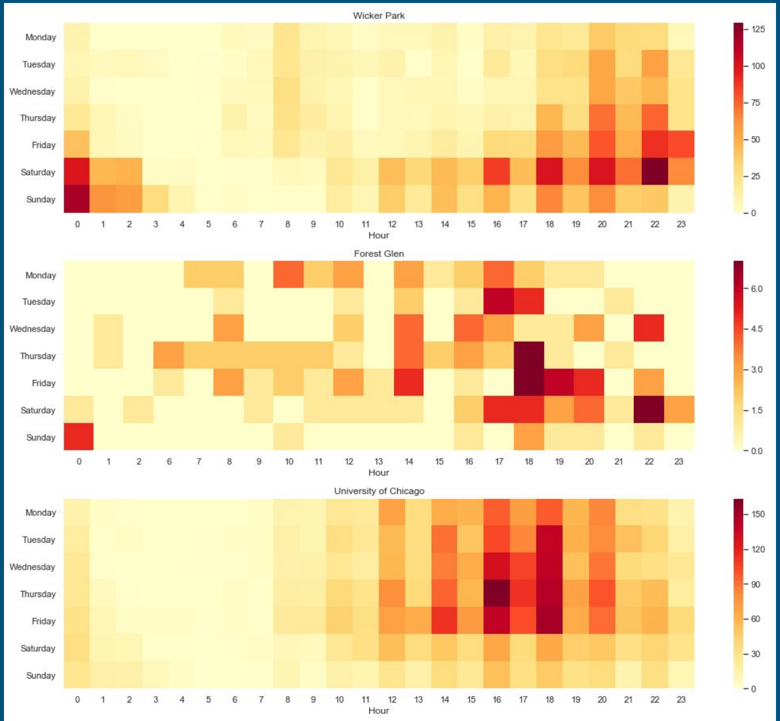


EDA Findings: Spatial Pattern



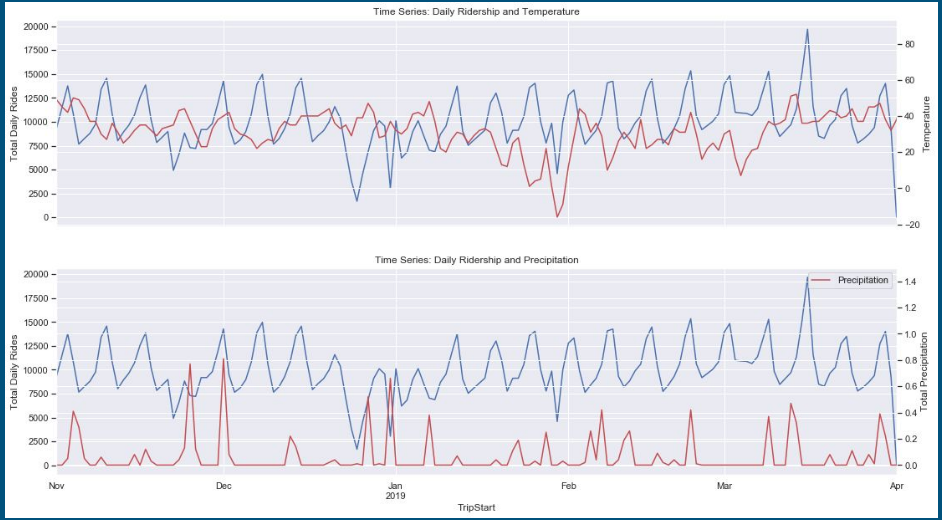
EDA Findings: Spatiotemporal Patterns

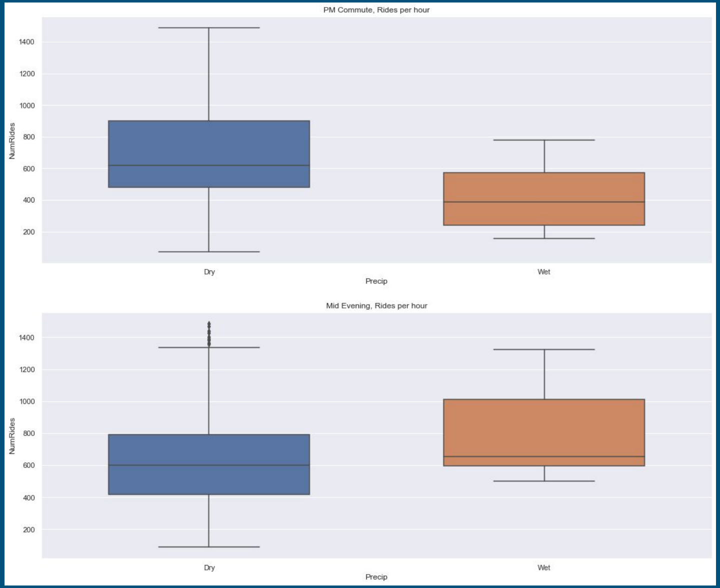




EDA Findings: Weather Effects?

- Not apparent at daily level
- Focused effect within specific time windows





Machine Learning Implementation

Business Goal	ML Application
Target improvements in public transit offerings	Clustering analysis to uncover common utilization patterns
Improve infrastructure planning	Regression to predict rideshare utilization by location and time

Clustering - Data Preparation

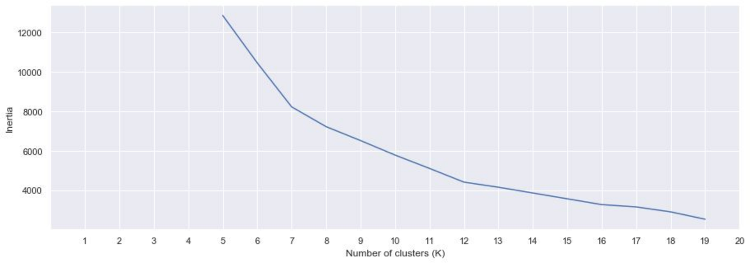
- Customer Segmentation: cluster routes by behavior
- Data structure:
    - Rows indexed by Route (MultiIndex, unique pairing of Pickup and Dropoff Community Area)
    - Aggregated ride counts by time of day and week period (10 categories)
        - Weekday vs Weekend
        - Early morning, morning, mid-day,afternoon, evening, late evening
    - Average Fare
    - Is Airport (pickup or dropoff)
- All fields scaled

Model Selection

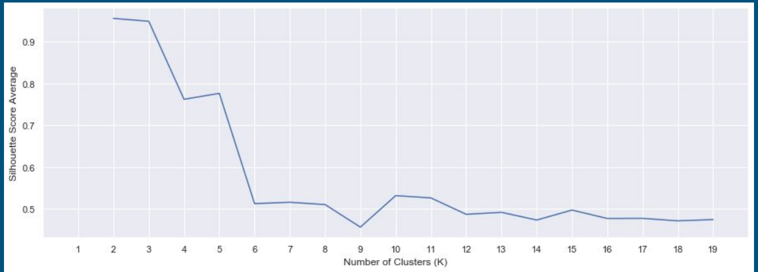
- KMeans ★
- DBSCAN
- Agglomerative Clustering
- Spectral Clustering

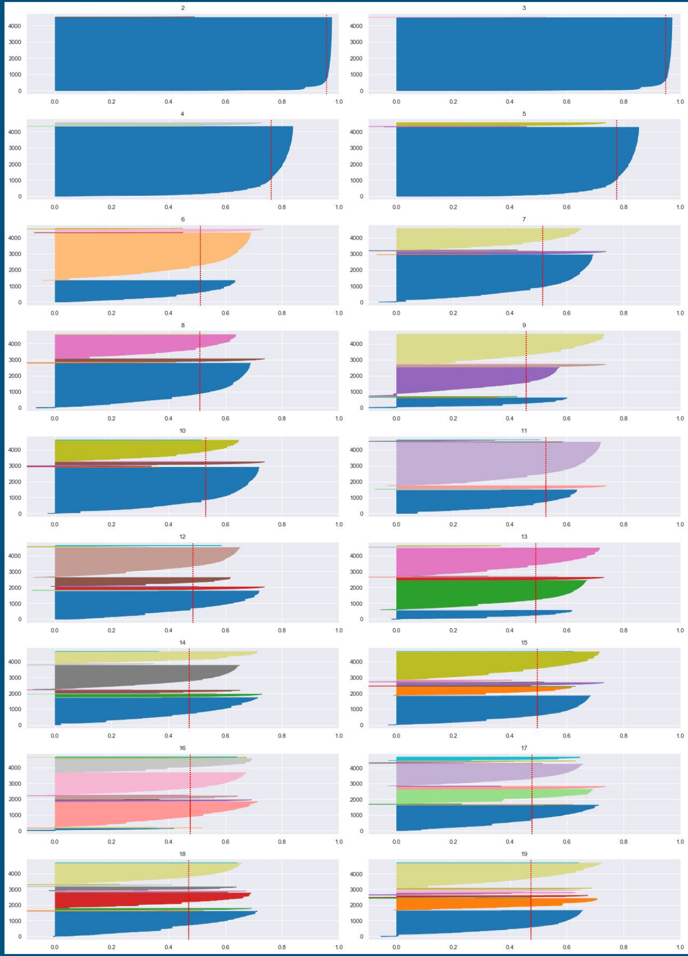
KMeans: Selection of K

- Elbow method

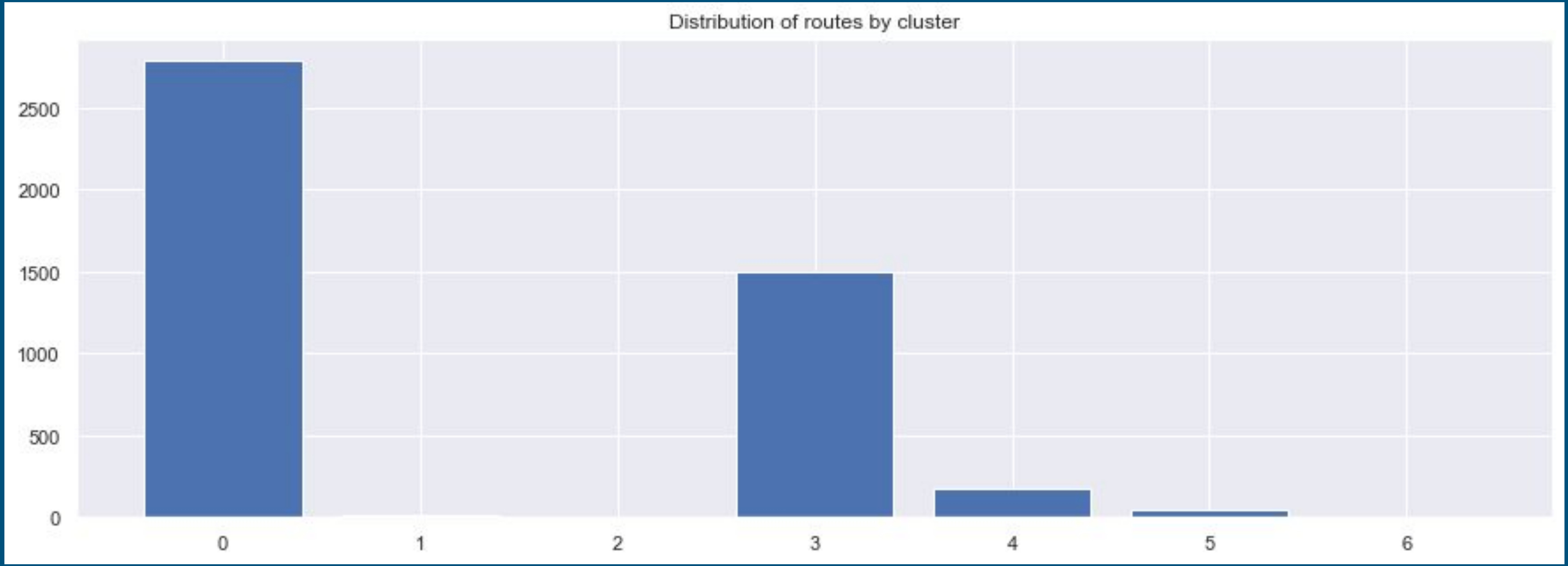


- Silhouette



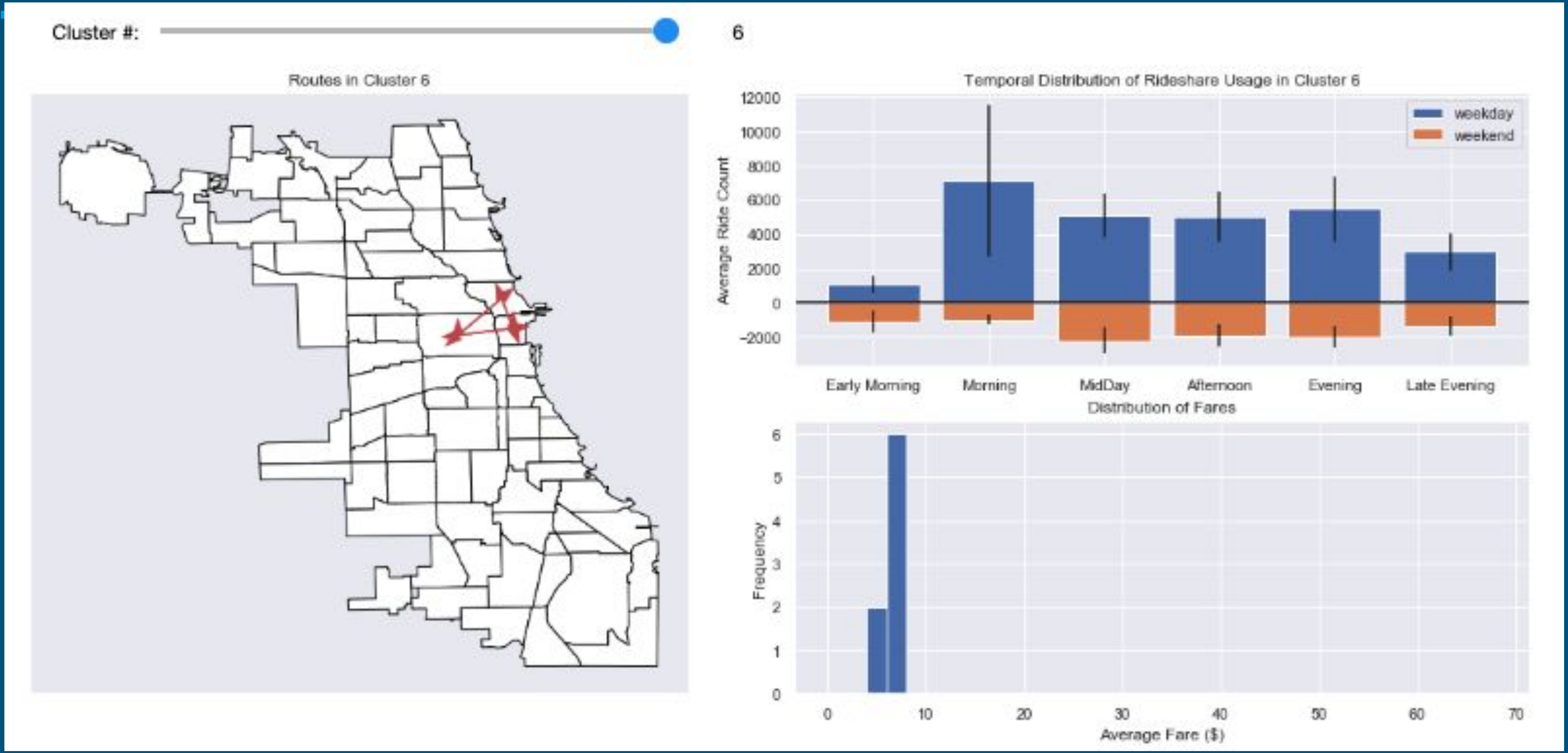


KMeans, K=7

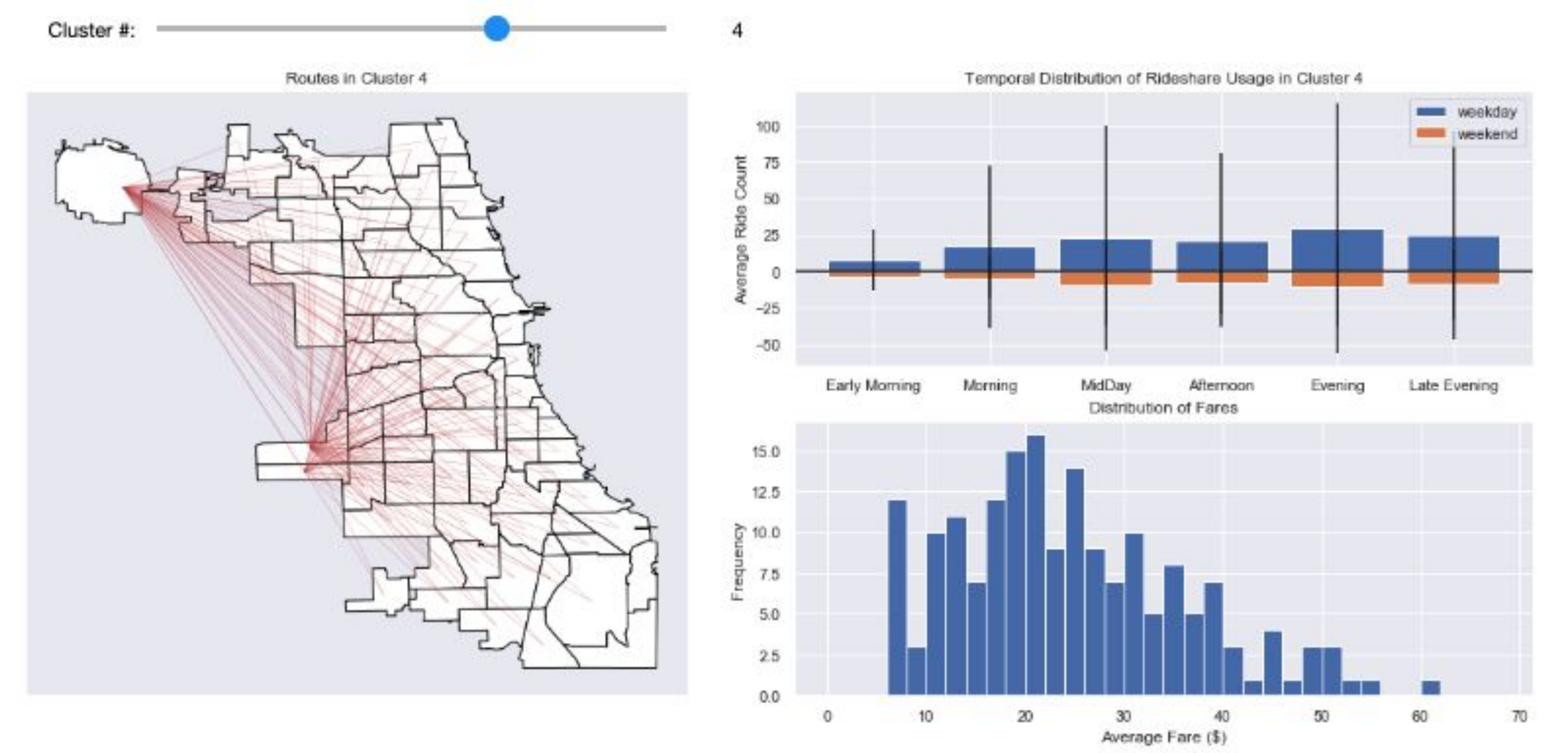


Distribution of routes by cluster
2500
2000
1500
1000
500
0
0
1
2
3
4
5
6

Visualizing Clusters - City Center



Visualizing Clusters - Airport



Clustering - Results

Found clusters representing routes with distinct rideshare patterns:

- Commuting near downtown: morning vs evening patterns
- Rides within the city center
- Airport rides
- Longer commute rides
- Low frequency routes with higher fares

Higher K $\longrightarrow$ more difficult to interpret

Clustering - Future Work

- Break out the airport rides
- Weighting ride count variables
- Finer spatial scale - focused regions
- Incorporate demographics

Usage Prediction by Regression

- Goal: predict ride pickup counts by location given DOW, hour
- EDA: cyclical, focus on downtown, airport
- Assumption: Future rideshare usage will be consistent with historical pattern.

Dataset Preparation

Group on pickup tract, ride start time (rounded)

Target: Count of rides per grouping (~5% of total)

Features:

- Agg fields: Average fare, Average distance
- Derived fields: Hour, DOW, IsHoliday, IsAirportPU
- Enriched fields: DistToDowntown, Bearing, MedIncome, PopDensity, Temperature, Precipitation

Linear Regression - Statsmodels

Advantage: Explainable model, infer parameter effects

Potential Issues:

- Correlated features (distance, time, fare)
- Is underlying model linear?
- Target variable is count, non-negative

OLS Regression Results

```
results=smf.ols('NumRides ~ DistToDowntown + TripTotal + C(Precip)
        + C(IsAirportPU) + C(DayPeriod) + C(IsWeekday)'
        ,data=agg_hourly_all).fit()
```

Dep. Variable:	NumRides	**R-squared:**	0.149
Model:	OLS	**Adj. R-squared:**	0.149
Method:	Least Squares	**F-statistic:**	9429.
Date:	Fri, 25 Oct 2019	**Prob (F-statistic):**	0.00
Time:	14:30:18	**Log-Likelihood:**	-1.5281e+06
No. Observations:	537380	**AIC:**	3.056e+06
Df Residuals:	537369	**BIC:**	3.056e+06
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.5931	0.022	210.937	0.000	4.550	4.636
C(Precip)[T.Wet]	0.1839	0.031	6.003	0.000	0.124	0.244
C(IsAirportPU)[T.1]	9.2190	0.060	153.052	0.000	9.101	9.337
C(DayPeriod)[T.morning]	0.3546	0.020	17.369	0.000	0.315	0.395
C(DayPeriod)[T.midday]	0.4401	0.021	21.235	0.000	0.399	0.481
C(DayPeriod)[T.afternoon]	0.7442	0.022	34.596	0.000	0.702	0.786
C(DayPeriod)[T.evening]	1.6513	0.021	79.566	0.000	1.611	1.692
C(DayPeriod)[T.lateevening]	0.9898	0.022	44.611	0.000	0.946	1.033
C(IsWeekday)[T.1]	-0.2841	0.012	-22.974	0.000	-0.308	-0.260
DistToDowntown	-0.3834	0.001	-280.178	0.000	-0.386	-0.381
TripTotal	0.0166	0.001	18.116	0.000	0.015	0.018

Non-Parametric Regression

Does not assume functional form for Y=f(X)

Candidates:

- KNN Regression
- RBF-kernel SVR
- Tree-Based Models

Baseline Models

NumRides = F(DistToDowntown,Hour,IsWeekday,TripMiles,Precip)

TrainTestSplit: (*Time series*) Before: 80%, After 20%

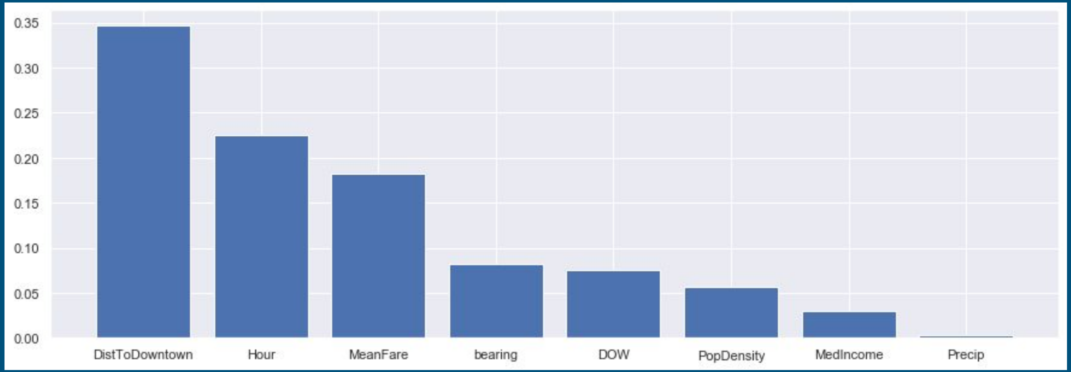
Remove airport pickups and holidays

Test Results:

Baseline Model	R-squared	RMSE
Random Forest	0.81	1.9
Gradient Boosting	0.68	2.5
Bagging	0.81	1.9

Feature Selection

- Modifications to Base Model Feature Set:
    - + Bearing
    - IsWeekday —> DOW
    - Trip Miles —> Mean Fare
    - + MedIncome, PopDensity



- Best combination of features for RandomForestRegressor:
    - DistToDowntown, Hour, MeanFare, Bearing, DOW, PopDensity
- Test R-Squared: 0.84 (was 0.81), RMSE: 1.8 (was 1.9)

Hyperparameter Tuning

Cross-Validation with Time Series Split applied to training data, 4 partitions:

- Number of Estimators

n_estimators	10	20	30	50
R-squared	0.825	0.831	0.832	**0.834**

- Grid Search: Parameter Selection Method, Tree Depth (5 to 25)

For max_depth=16:

max_features	Log2	Square Root	None (Bagging)
R-squared	0.830	0.829	**0.839**

Final Model

Hyperparameters:

N_estimators = 50

Max_depth = 16

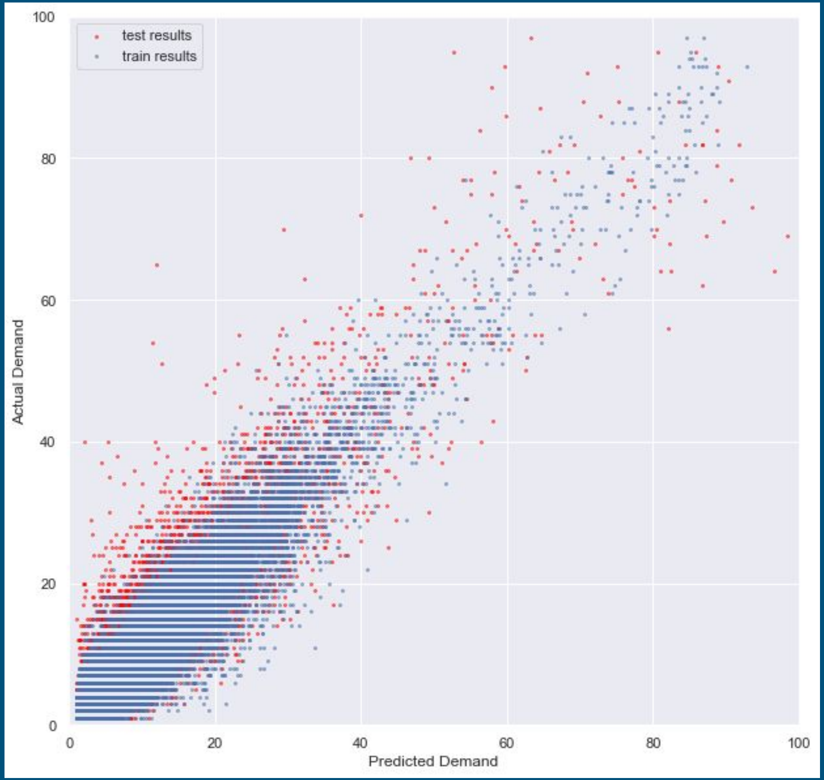
Feature_Selection = None

Fit to training set and score on test:

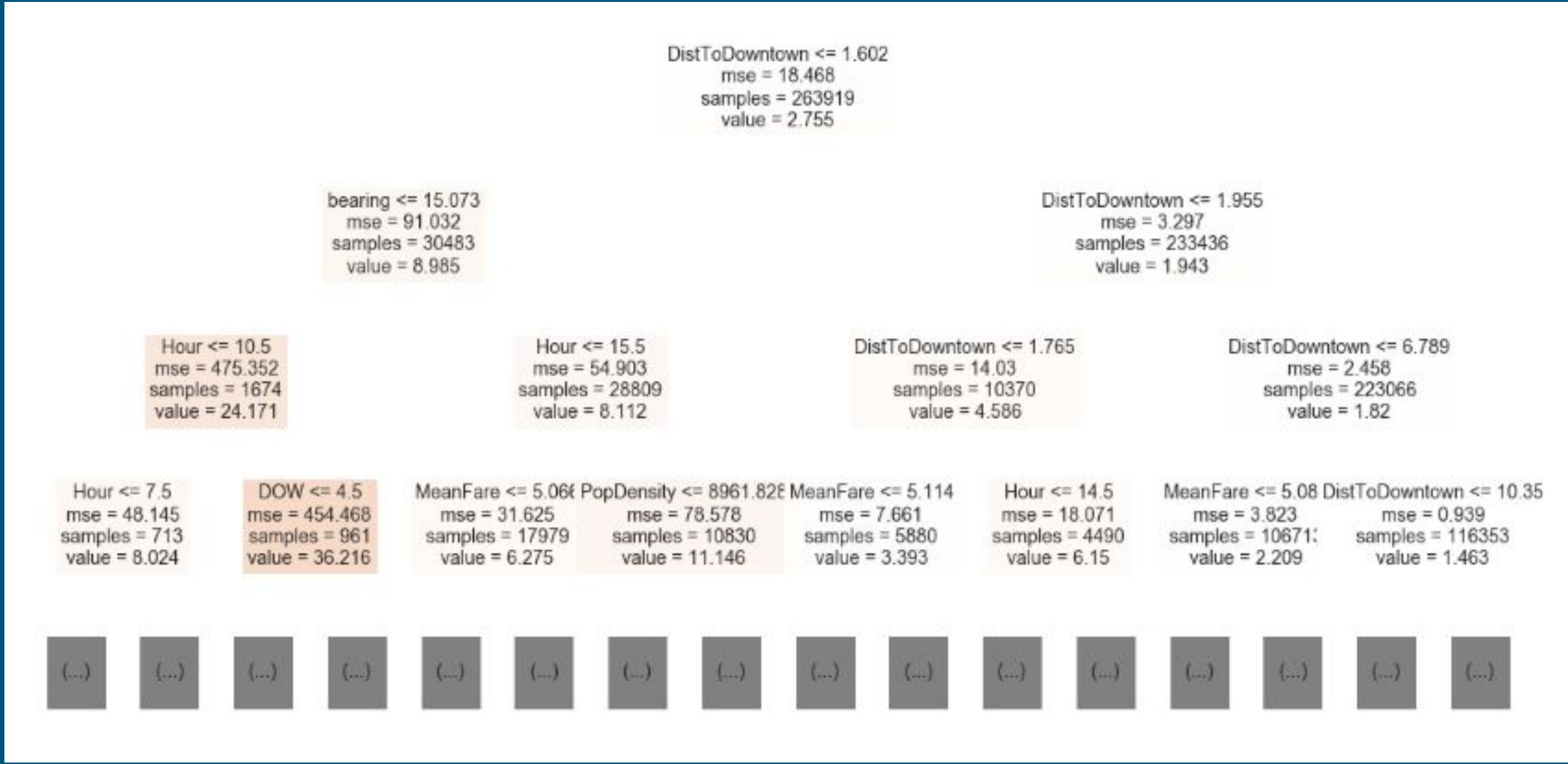
R-squared: 0.85

RMSE: 1.69

Evaluation: Predicted Vs Actual



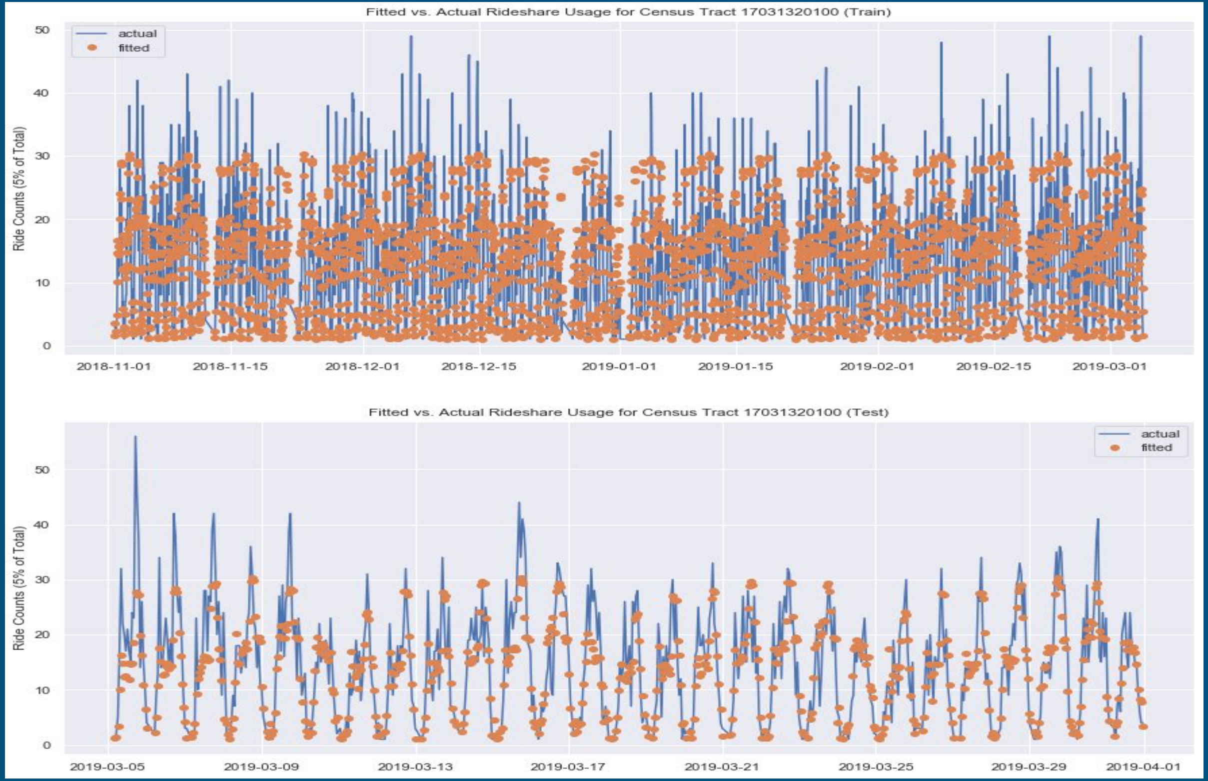
Evaluation: Top Feature Splits



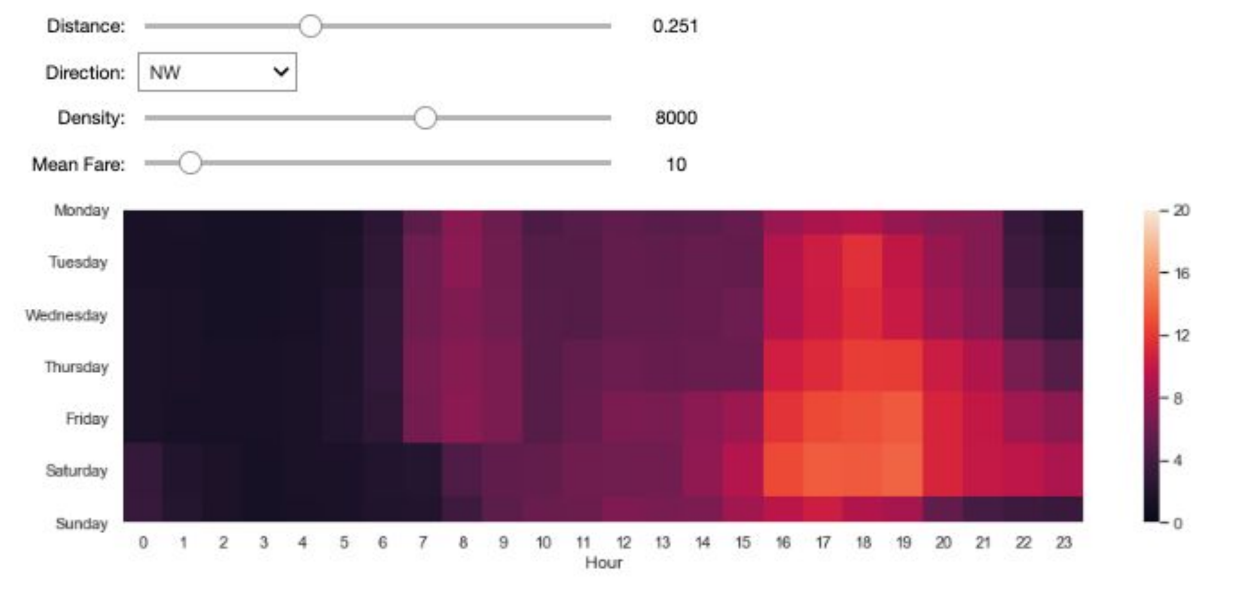
Evaluation - Single Census Tract

Test

Train



Communication of Findings - Dashboard



Future Work - Usage Prediction

- Use updated Trips dataset, larger sample
- Revisit boosting
- Spatiotemporal visualizations
- Time series forecasting models for each census tract, including airports
- Integrate important City features (train stations, major job centers, sports arenas)

Conclusion

- Developed rideshare customer segmentation and usage prediction tools
- Built dashboards to enable inspection of results and prediction
- Demonstrated methods for gaining insight from rideshare datasets