CS 418

Introduction to Data Science
Final Project

Problem Statement

Gather insights on ridership in Chicago

- for Uber
- for Lyft

Ridership data available from the Chicago data portal

Use regression and clustering to find trends between the time of day and ride duration and price to see what conclusions we can make.

Data Source

	Trip ID	Trip Start Timestamp	Trip End Timestamp	Trip Seconds	Trip Miles	Pickup Census Tract	Dropoff Census Tract	Pickup Community Area	Dropoff Community Area	Fare	Additional Charges	Trip Total	Shared Trip Authorized	Trips Pooled	Pickup Centroid Latitude	Pickup Centroid Longitude	Pickup Centroid Location	Dropoff Centroid Latitude	Dropoff Centroid Longitude	Dropoff Centroid Location
0	9c620428487cada88130ab08e2ed063c4824852d	9/12/19 12:45	9/12/19 12:45	24	0.0	1.703184e+10	1.703184e+10	28.0	28.0	10.0	0.00	10.00	True	1	41.870415	-87.675086	POINT (-87.6750856208 41.8704150003)	41.870415	-87.675086	POINT (-87.6750856208 41.8704150003)
1	9c62bd8cad97bd27430e2e3ffc07628a7fa00d5d	8/18/19 19:00	8/18/19 19:00	242	0.0	1.703108e+10	1.703108e+10	8.0	8.0	2.5	2.55	5.05	False	1	41.898332	-87.620763	POINT (-87.6207628651 41.8983317935)	41.898332	-87.620763	POINT (-87.6207628651 41.8983317935)
2	9c62fa73e82e32a54c9f2aca47e11f370b9bb80c	7/8/19 17:15	7/8/19 17:15	6	0.0	NaN	NaN	16.0	16.0	15.0	0.00	17.00	True	1	41.953582	-87.723452	POINT (-87.7234523905 41.9535821253)	41.953582	-87.723452	POINT (-87.7234523905 41.9535821253)
3	9c69c7bba2eb3f2846988fed0916b1dc245b0b14	9/27/19 6:30	9/27/19 6:30	10	0.0	NaN	NaN	2.0	2.0	2.5	2.55	5.05	False	1	42.001571	-87.695013	POINT (-87.6950125892 42.001571027)	42.001571	-87.695013	POINT (-87.6950125892 42.001571027)
4	9c7214acfd1aded46abeebe484939233d618c02a	9/28/19 22:45	9/28/19 22:45	21	0.0	1.703107e+10	1.703107e+10	7.0	7.0	2.5	2.55	5.05	False	1	41.929078	-87.646293	POINT (-87.6462934762 41.9290776551)	41.929078	-87.646293	POINT (-87.6462934762 41.9290776551)

Link: https://data.cityofchicago.org/Transportation/Transportation-Network-Providers-Trips/m6dm-c72p/data

Description of Data Science Solution

Dataset was of 25 GB.
Millions of rows, hence we partitioned it to just 50,000 rows.
Filled missing values (NaN) with average values or 0.
No need to mask data as there was no sensitive or private data.

Description of Data Science Solution

To calculate the Trip Total Charges, we built regression models using the variables

- Trip Seconds
- Trip Miles
- Additional Charges

Score obtained when:

- 1 predictor variable used: 79.16%
- 2 predictor variables used: 83.88%
- All predictor variables used: 85.68%

Description of Data Science Solution

Understanding time-based patterns is critical for any business.
Time-series forecasting important for a data scientist.

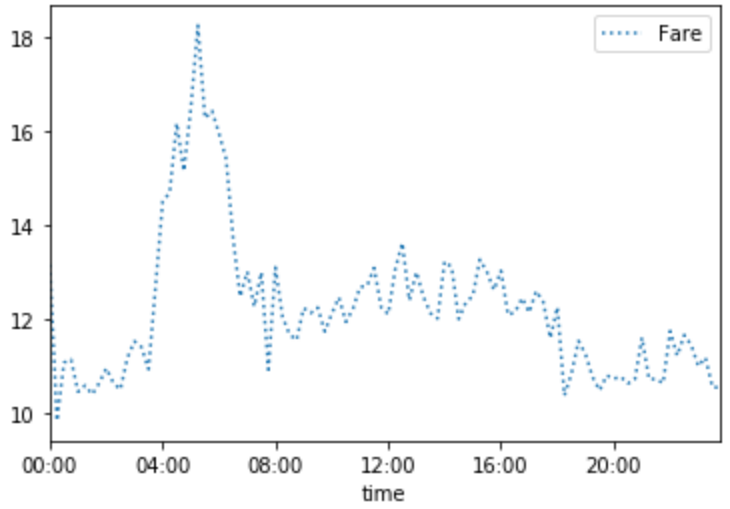
Facebook Prophet

- Open source library published by Facebook
- Based on decomposable (trend + seasonality + holidays) models.

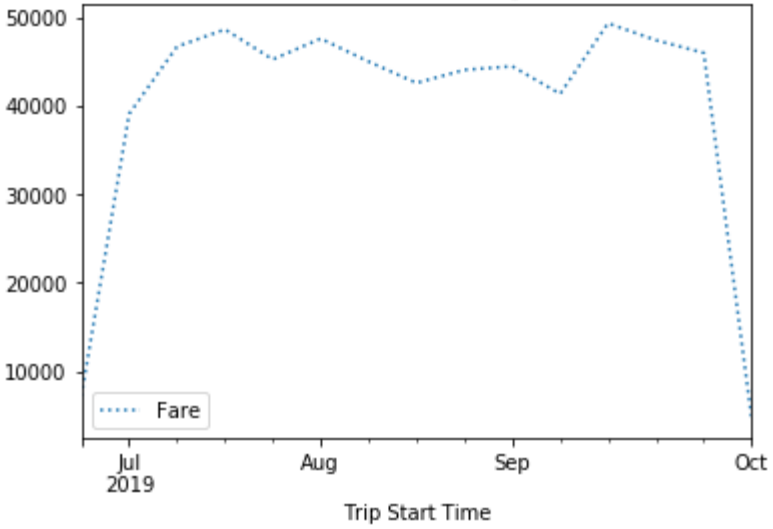
Results

From the data that we processed, it can be concluded that 04:00 to 08:00 is usually the highest fare period.

Hourly

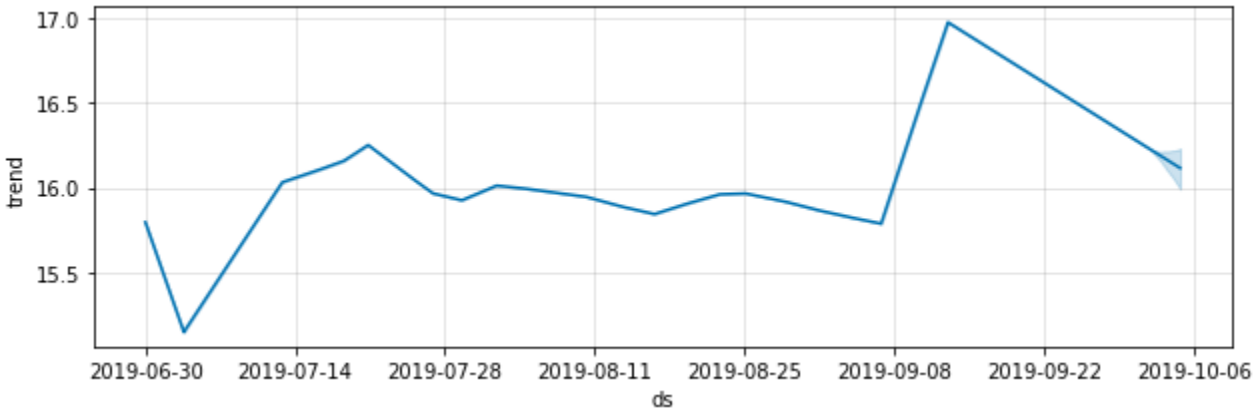


Monthly



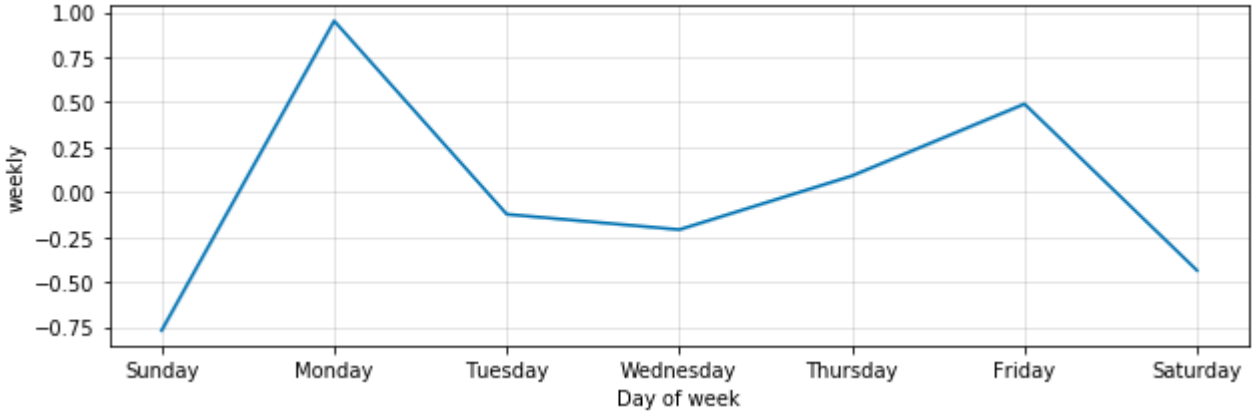
From the data that we processed, it can be concluded that July start to September end the total fare remained consistent, however, it gradually declined from October start.

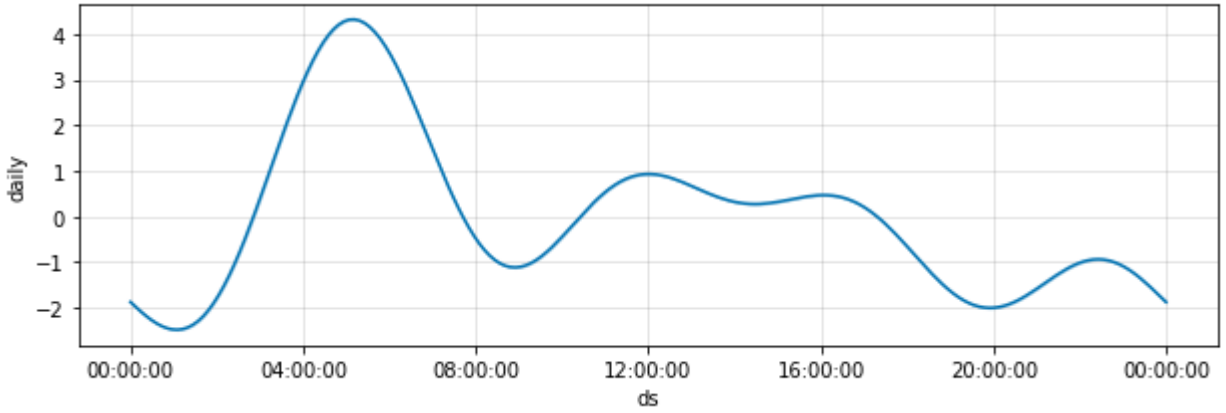
Results- Prophet Predictions



Monthly

Weekly





Hourly

Conclusion

Hence, we can conclude that:

1. All the predictor variables were required to increase the score of our regression model.
2. 04:00 to 08:00 hrs is the highest fare period.
3. Facebook Prophet helps us in prediction using timeseries where we can predict that September is the highest fare month, Sunday is the least busy and Monday is the busiest day of the week. 04:00 to 08:00 hrs is the highest fare period.