1. Describe your problem and solution
    a. The problem we wanted to address was that of seeing the data of the ridesharing apps and visualizing so we could gather some insights about the pricing, and the busiest times of the day and the most expensive times of the day, and which weekdays there is more ridership. All this along with a regression model to predict the price of the ride.
    b. To answer these questions we decided regression would be the best option to predict the price using ‘Trip Seconds’, ‘Trip Miles’ and ‘Additional Charges’ as the higher these values are, the total price of the ride would be higher as well. And we played around with a combination of all these to build models to obtain the optimal score.
    c. Along with this, we wanted to visualize the data so everyone could be able to see the trends and we decided on the time series as suggested by the professor during our consultation for the project. You can see the visualization for the daily, weekly, and monthly trends.
2. Describe your datasets

    The dataset we are using is the Transportation Network Providers - Trips which includes all the trips and their relevant data starting November 2018 reported by the rideshare companies to the City of Chicago as part of routine reporting required by ordinance. Below is a snapshot of the dataset columns and values.

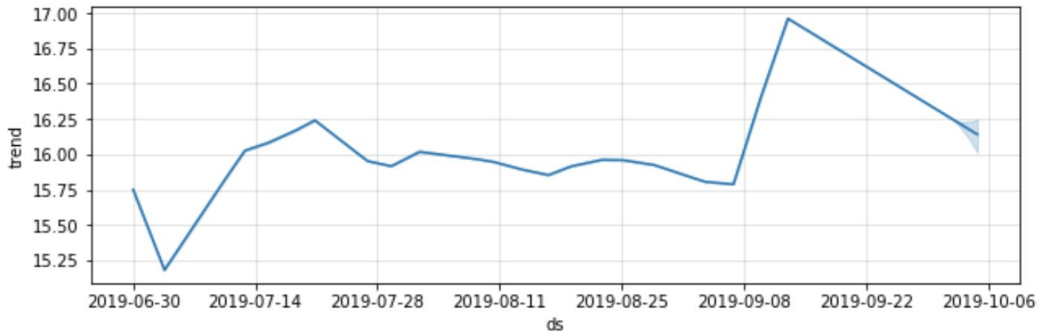
	Trip ID	Trip Start Timestamp	Trip End Timestamp	Trip Seconds	Trip Miles	Pickup Census Tract	Dropoff Census Tract	Pickup Community Area	Dropoff Community Area	Fare	Additional Charges	Trip Total	Shared Trip Authorized	Trips Pooled	Pickup Centroid Latitude	Pickup Centroid Longitude	Pickup Centroid Location	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	9c7214acfd1aded46abeebe48493923 3d618c02a	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)

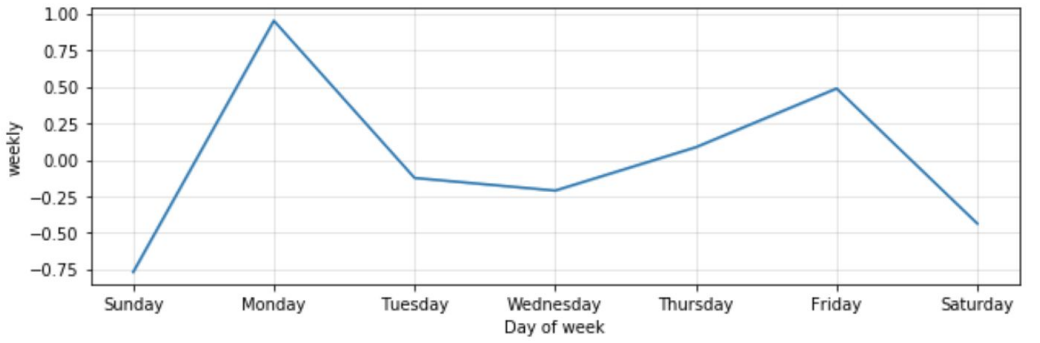
3. Describe your data preparation process and report the results obtained
    a. Gather Data: This was the trickiest part because we had to decide on a dataset to build the project around and we settled on the ridership data provided by the city of Chicago for ride-sharing apps like Uber and Lyft. We felt like this was a good fit because of how many people use these services and whatever insights we could gather from the data would be actually useful.
    b. Cleanse and Validate Data: Removing some data was challenging as our dataset was very large so we decided to only use 50,000 of the values instead of the millions available which would have taken hours and a supercomputer to visualize. Filling in missing values was another challenge as the things we wanted to use did not accept NaN values so then we just filled those values with the average thinking we could change it if we had to later and make a better model. Masking private or sensitive data entries was not of a problem as there were no names or anything just a user id which was made up of numbers and letters.
4. Describe your data exploration process and report the results obtained
    a. Data Exploration starts with discovering and assessing the data. This was interesting because we had to pick certain variables from the dataset that we could

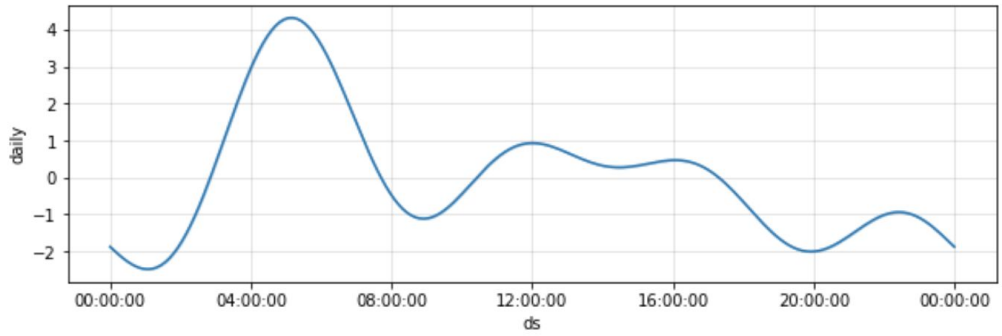
use to find solutions to the problems we mentioned in Q1, to build the regression model we settled on using ‘Trip Seconds’, ‘Trip Miles’ and ‘Additional Charges’ as the variables as those were the ones that made the most sense. To visualize some of the trends mentioned above we decided on using the timestamps as the variables as that was the variable that offered us the most flexibility and options to visualize the trends which date we could group in different months, and then group them according to the weekdays and then group them one more time depending on the time of the day.

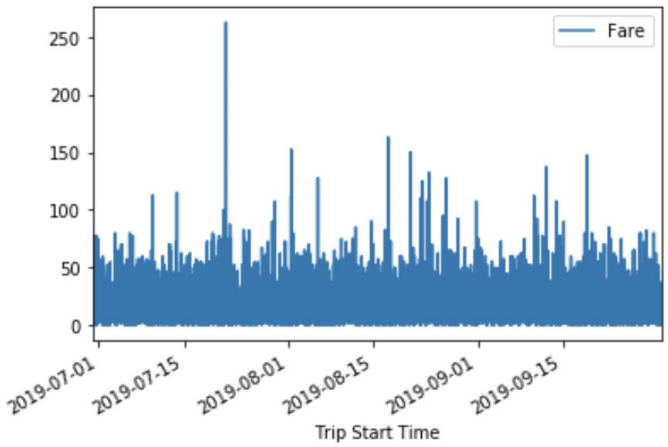
5. Describe your data modeling process and report the results obtained
    a. The data modeling process basically transforms and enriches data. To transform the data we started by building a regression model like we had in the previous project and this includes partitioning the data into a training and test set which one can see in the project3.ipynb file. After this, another part of transforming and enriching was the visualization that needed to be done to see the answers to the problems mentioned in Q1, and this was done using time series. Both the regression model and the visualizations came out how expected and the regression model got better as we included more variables that affected the fare price. The time-series suggested by the professor helped us visualize all of the problems we wanted to be visualized in Q1. During this data modeling process and especially the visualization we see a lot of trends such as ridership is high on Mondays and then again on Fridays, Its lowest on Wednesdays. The peak price occurs around 6:00 am. There is also a price increase in fare prices starting September which

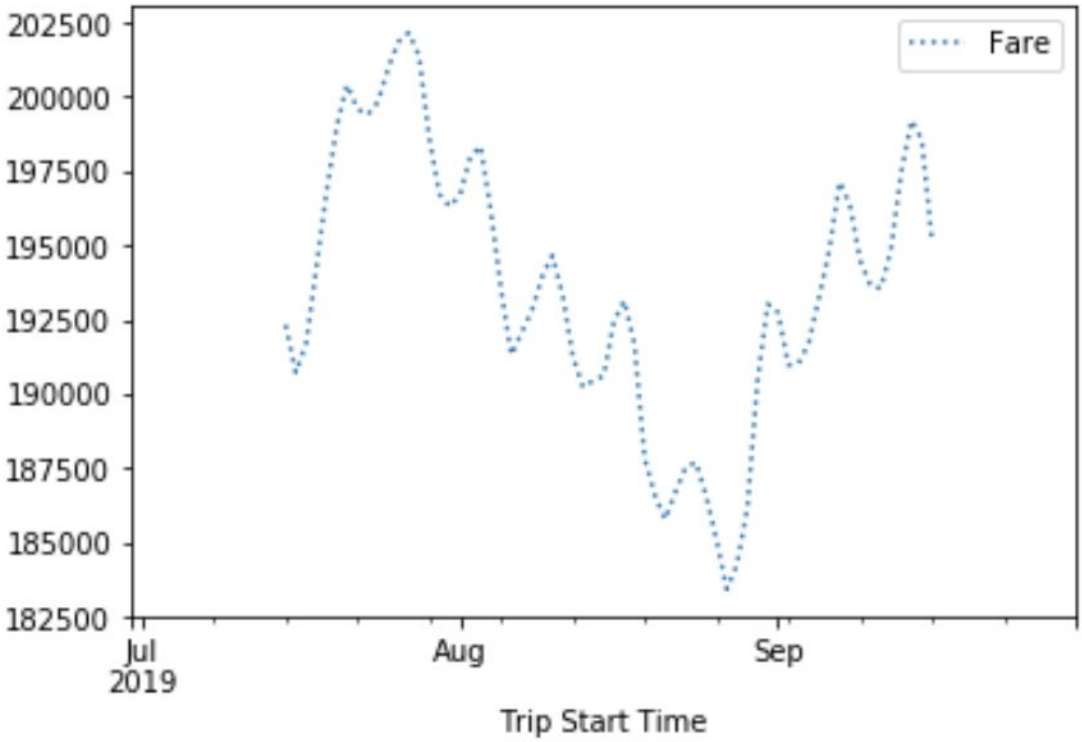
would lead us to believe that these are because of schools starting up again.

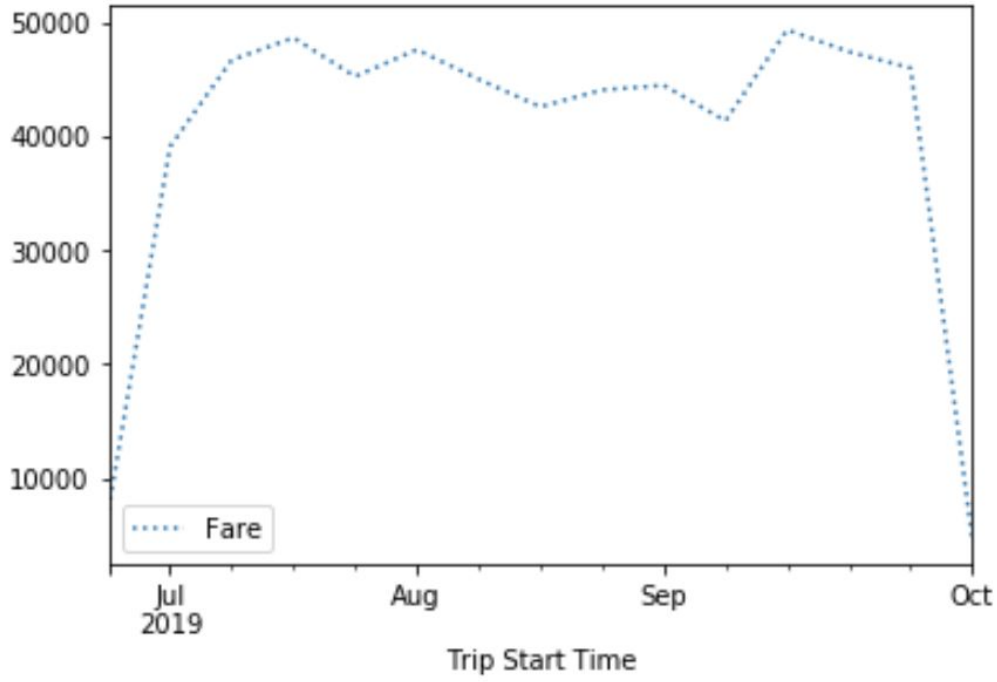


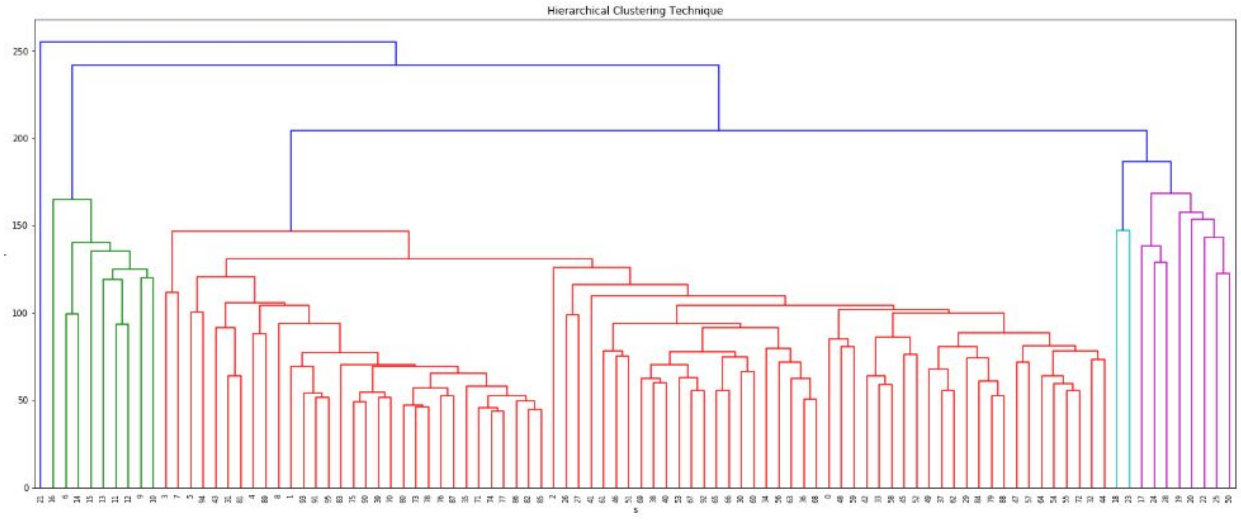







202500
200000
197500
195000
192500
190000
187500
185000
182500
Fare
Jul
2019
Aug
Sep
Trip Start Time


50000
40000
30000
20000
10000
Fare
Jul
2019
Aug
Sep
Oct
Trip Start Time


Hierarchical Clustering Technique
250
200
150
100
50
0