# Report on Taxi Fare Visualization and Prediction


## Christopher Hendra


## Data Description

We are given the dataset Taxi Trips - 2021 describing reported taxi trips from the year 2021 by taxi companies in Chicago. Each row of the dataset represents the start time and end time for the journey rounded to the nearest 15 minutes, masked taxi id, ride id, and total payment made for the ride, which includes the fare, tip, toll, and other extra charges. The city of Chicago can be divided into 77 districts and in most cases, we have access to pickup and dropoff locations based on these 77 districts from the columns pickup community area and dropoff community area. Due to privacy concerns, the pickup and dropoff area can often be missing, along with the census tract or the latitude and longitude of the centroid of the census tract from which the trip originates.

In this report, we will summarize some key statistics from the dataset above, along with several other datasets that we have downloaded from the same website for the purpose of comparison, with the goal of evaluating the decision to launch a ride-hailing service in the city of Chicago. Eventually, we will also present our modeling approach toward predicting taxi fares in Chicago and evaluate the usefulness of our predicted fares on a held-out test dataset as well as several other datasets from the years prior to 2021.

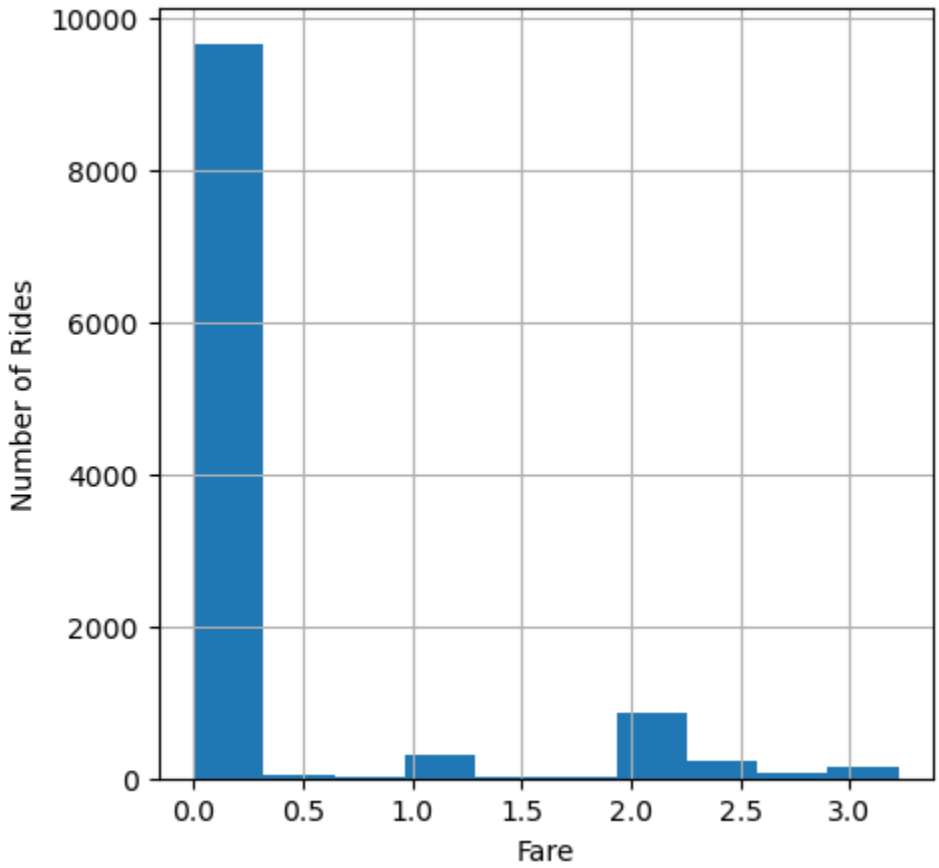
## Data Cleaning

### Fare Distribution

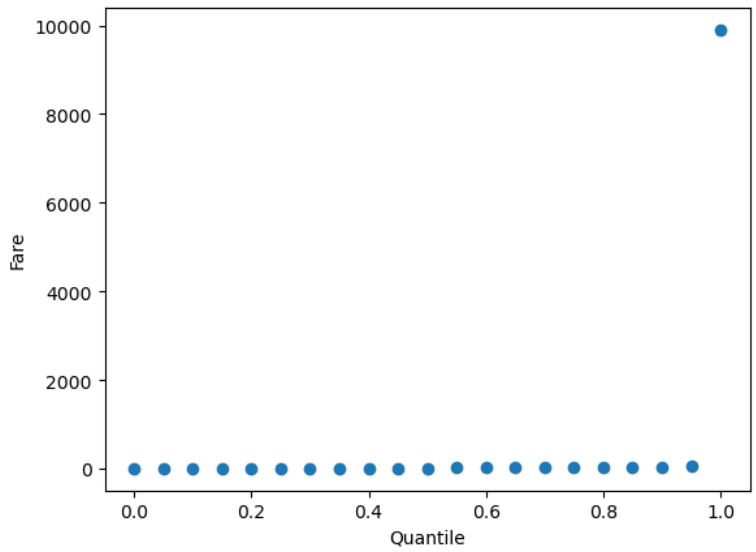
We begin by identifying a set of anomalous rides from the dataset so as to ensure the robustness of our analysis. The city of Chicago describes the following guideline for taxi fare (Chicago Taxi Fare):

- Base Fare of $3.25
- Each additional mile is $2.25
- Every 36 seconds of elapsed time $0.20
- First additional passenger (aged 13 through 64) $1.00
- Each additional passenger is $0.50
- Convenience Fee for electronic payment $0.50
- Vomit Clean-up Fee $50.00
- Illinois Airport Departure Tax $4.00 (for taxis leaving the airports)

Firstly, we note that any ride should have a minimum fare of $3.25. We identify roughly 0.29% of the rides in the 2021 taxi dataset to have fares below $3.25. Furthermore, most of these 0.29% trips have zero fares, which is clearly anomalous and so we decide to remove these rides from our analysis

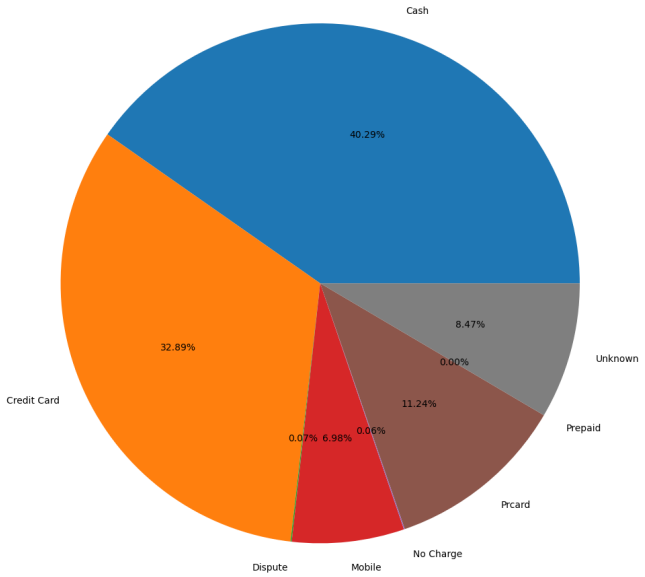



Next, we visualize the remaining rides with fares above $3.25 and plot the fare corresponding to the higher quantiles on the dataset, and notice that some of the fares can be as high as $1000. According to a report by Schaller consulting, an average long trip in Chicago will cost $25.30 anything way beyond this number seems suspicious. The 95th quantile fare on our dataset has a value of around $47 and so we decided to be conservative with our fares and set an upper limit of $100 fare for the rides in our dataset.




## Payment Type

After removing rides with anomalous fares, we visualize the distribution of payment type in our dataset




Some of the transactions are disputed or not being charged at all. We decided to remove these transactions as they are often indicative of mischarges or errors in the reporting. We decided to keep unknown transactions since they constitute roughly 8.5% of the remaining data and only exclude them when visualizing statistics that involve payment methods.

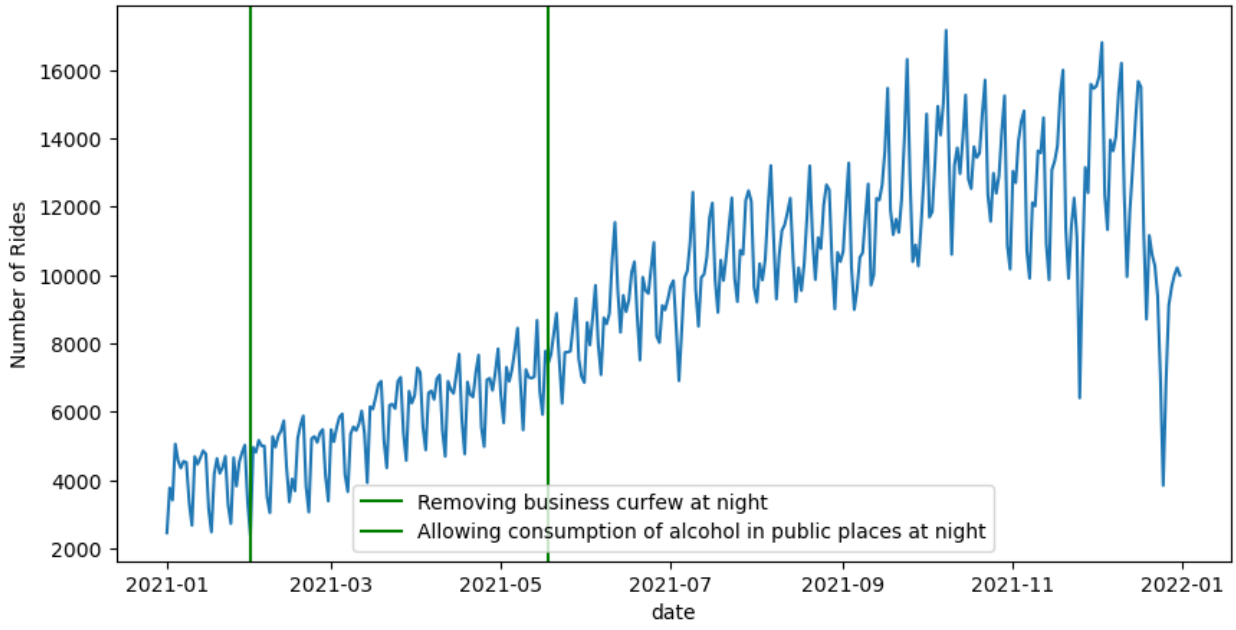
### Trip Distance, Duration, and Speed

Next, we remove rides with abnormal duration. A quick look at the dataset reveals a number of rides with a duration of more than 2 hours (up to 23 hours) or rides with a duration of less than 1 minute. We decided to remove anything less than 1 minute or more than 2 hours. Afterward, we also remove rides without any trip distance (zero or null). We allow for short trips (0.1 miles and above) but remove rides with speeds exceeding a certain limit. The speed limit in Chicago varies from 70 mph on interstate highways outside urban areas, 65 mph on rural interstates, 55 mph on interstate highways near or in major cities and on other highways, and 30 mph in the urban area unless some other speed restriction is established. As such, we remove any rides with speeds greater than 70 mph, which make up only 0.12% of the remaining rides.

## Data Visualization

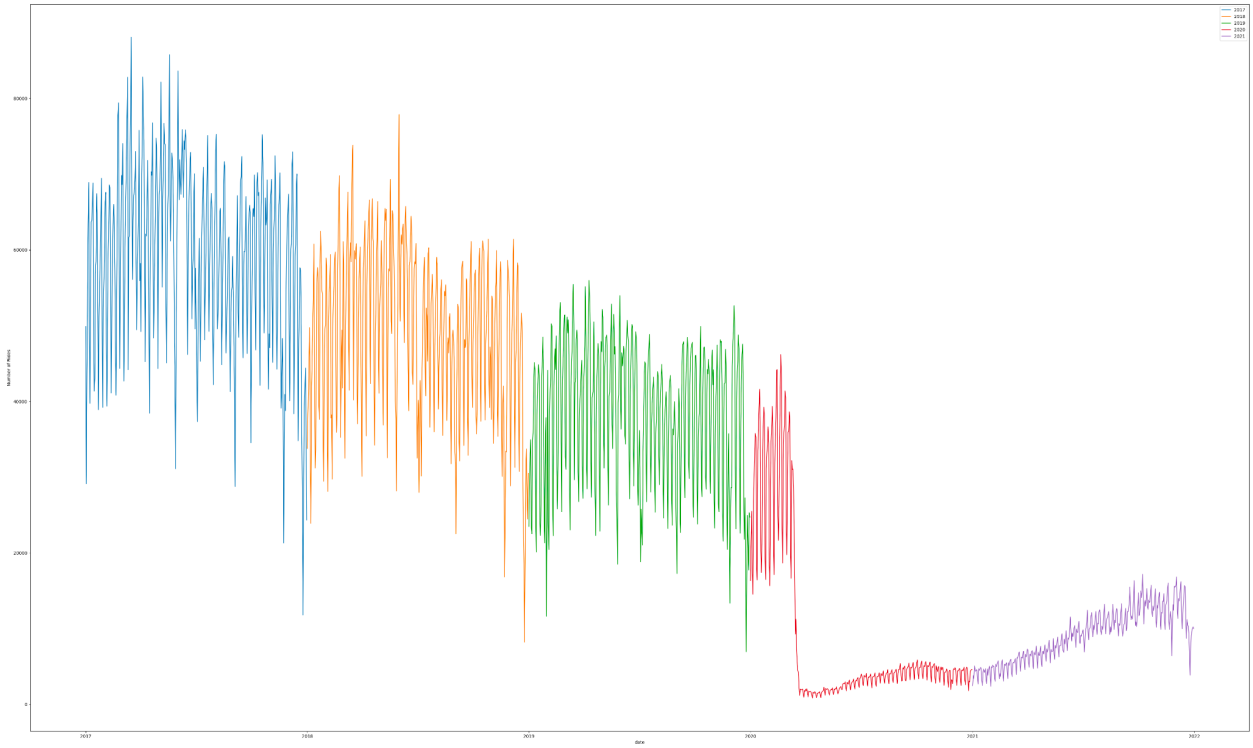
### Ride Patterns Throughout the Year

We first aim to understand the mobility patterns of riders and drivers in the city of Chicago. A naive look at the number of rides across the year will reveal the following pattern:



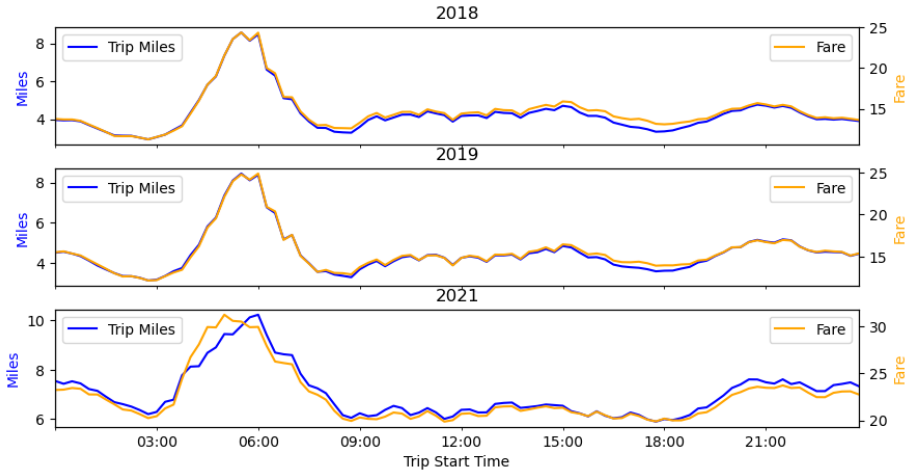

This plot seems to suggest that the number of rides increases throughout the year. However, we are aware of several covid restrictions being lifted by the city of Chicago such as the removal of the business curfew and the resumption of alcohol consumption in public places at night. In 2020, a couple of stay-at-home advisories and several movement restrictions were imposed by the city in order to curb the number of covid infections which theoretically should depress the number of rides throughout the year. In order to remove confounding effects on the observed

ride pattern, we decided to download the taxi dataset from **2017 to 2020** and visualize them alongside the 2021 dataset. When being plotted with the data from the pre-covid years, we observe the following trend:

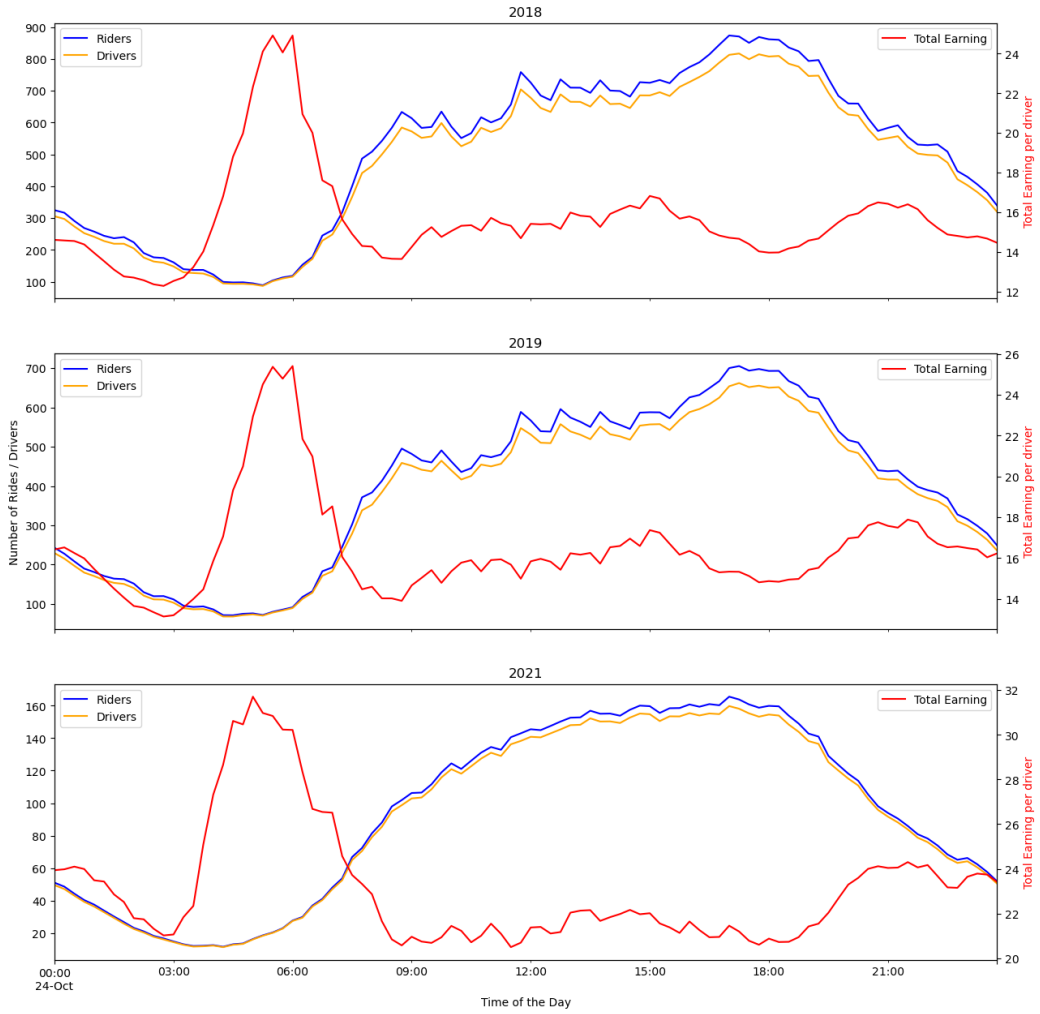


The blue-colored line plot represents rides from 2017, orange 2018, green 2019, red 2020, and purple 2021. We see that the number of rides while fluctuating, tends to increase slightly for the first half of the year. Furthermore, the number of rides during the pre-covid years is much higher than in our 2020 and 2021 datasets. The increase in the number of rides seems to be correlated with the resumption of activities following the lifting of some covid measures and the pent-up demands for travel that follows. While this dispels the idea that rides are more concentrated towards the end of the year, the plot suggests that the number of rides should continue to grow as more restrictions are being lifted and people are returning to the new normal, thus indicating a potential market share that can be captured

### Ride Patterns Throughout the Day

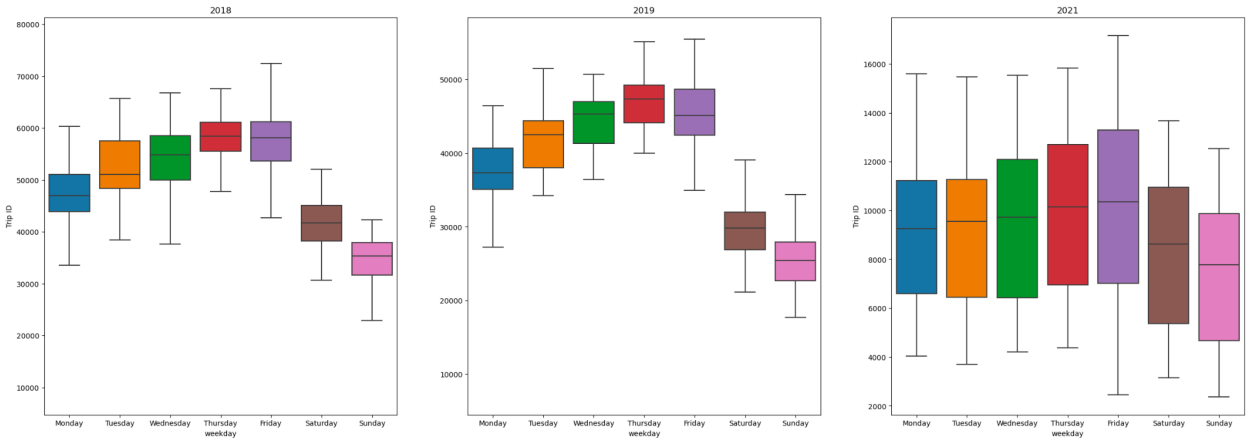



Next, we visualize the trend of trip distance and trip fare throughout the year. Here the blue lines represent the total miles traveled while the orange lines represent the fares paid during the ride. We visualize the 2018 and 2019 data as controls and note that the distribution of distance and fare seem to be equal across the years. We plot trip distance and fare together as a sanity check since the two are correlated and should display the same trend throughout the day. We observe a maximum average fare and trip distance in the morning which indicates that people tend to travel larger distances and pay higher fares in the morning

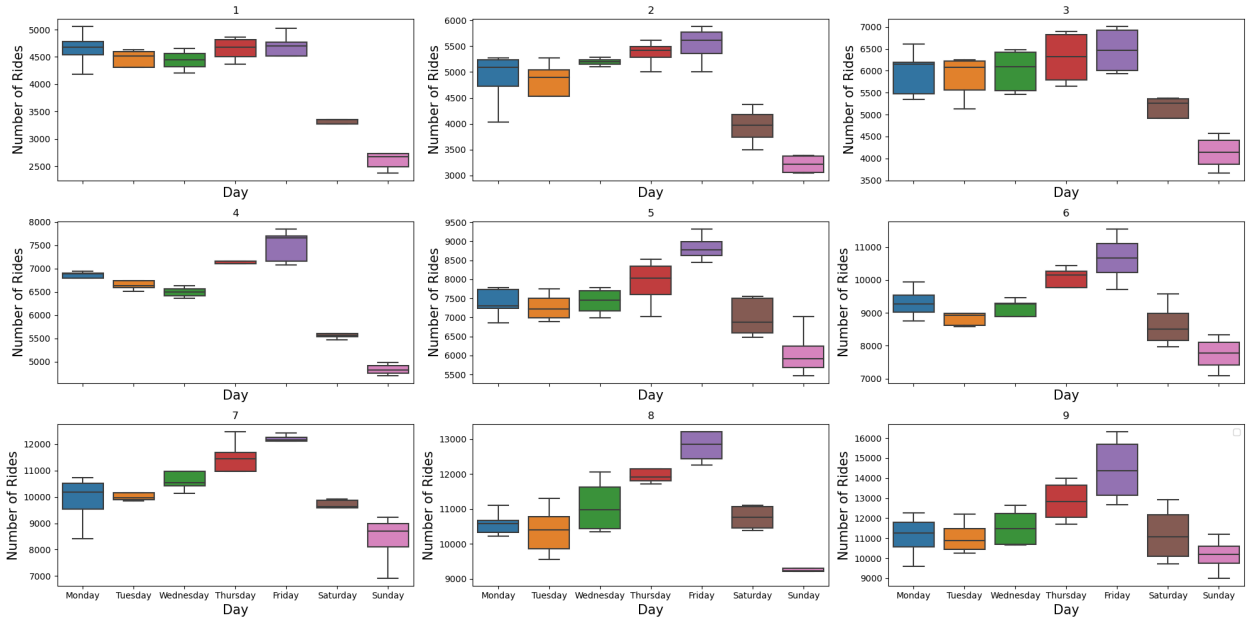



Afterward, we look at the trend for the number of rides, drivers, and total earnings throughout the day. The general trend seems to be the same between 2018, and 2019 when compared to 2021 but the number of rides and the number of drivers are definitely much lower during 2021. There are generally more rides during the later part of the day as compared to the earlier morning hours but the driver's earnings seem to be the highest during the morning hours. Consequently, in the morning, there is almost one taxi driver for every rider and the gap widens during the day when the driver’s earnings fall. This fall in earnings is most likely due to the trend of people taking shorter rides during the day as compared to the morning as shown previously

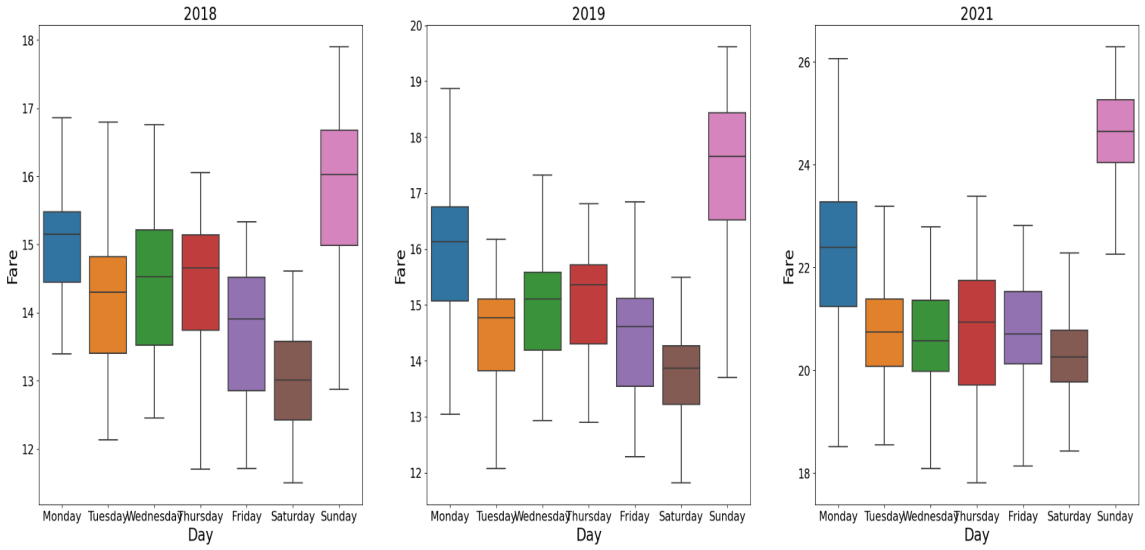
## Ride Patterns During the Week




We visualize the number of rides throughout the week for the years 2018, 2019, and 2021. Here we observe a general increase in the number of rides from Monday to Friday and a decrease during Saturday and Sunday for the years 2018 and 2019. The trend seems to be weaker in 2021 but this might also be confounded by the lifting of covid restrictions and so we visualize the number of rides during the week separately for the first 9 months of 2021

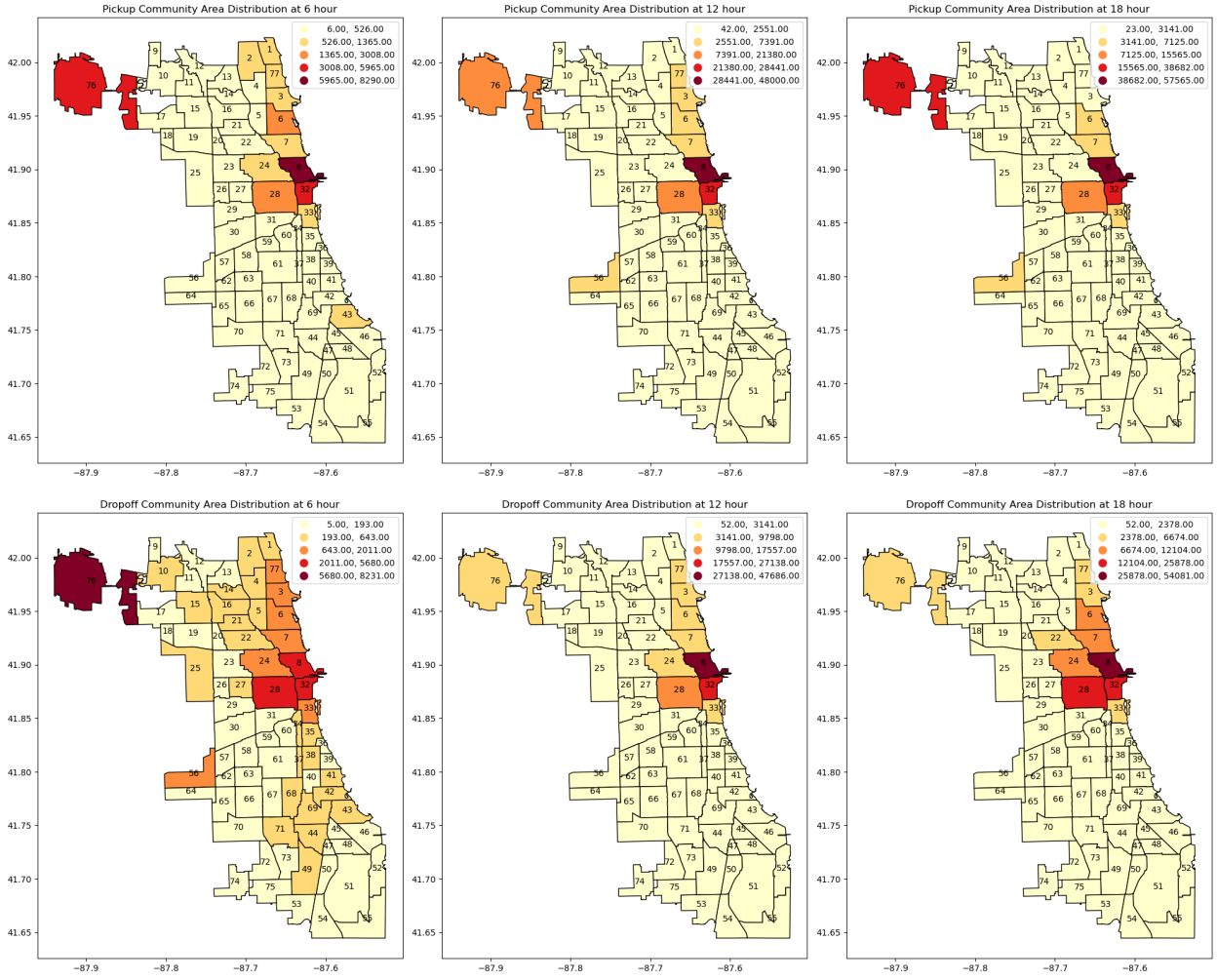



Here we see a clearer pattern throughout the week. This also suggests that the covid restrictions, while it affects the total number of rides during a period of time, it does not affect the short time periodicity of the number of rides.



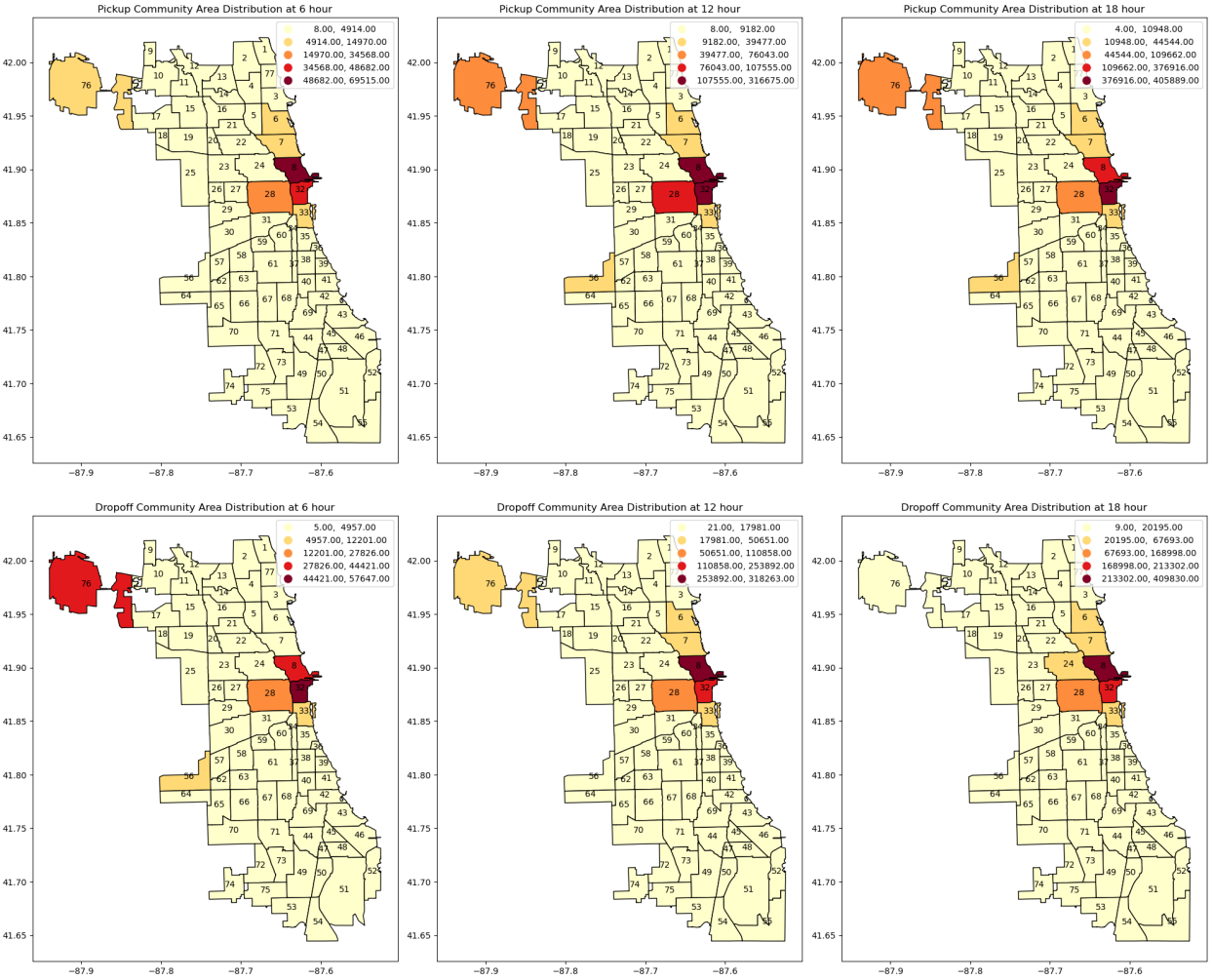

As opposed to the number of rides, the average fares throughout the day display the opposite pattern with people paying a higher fare on average on Sunday. This seems to suggest that people on average take longer trips on Sunday and less so during the other days of the week. We observe similar patterns in the years 2018 and 2019 as well.

## Ride Patterns by Community Area

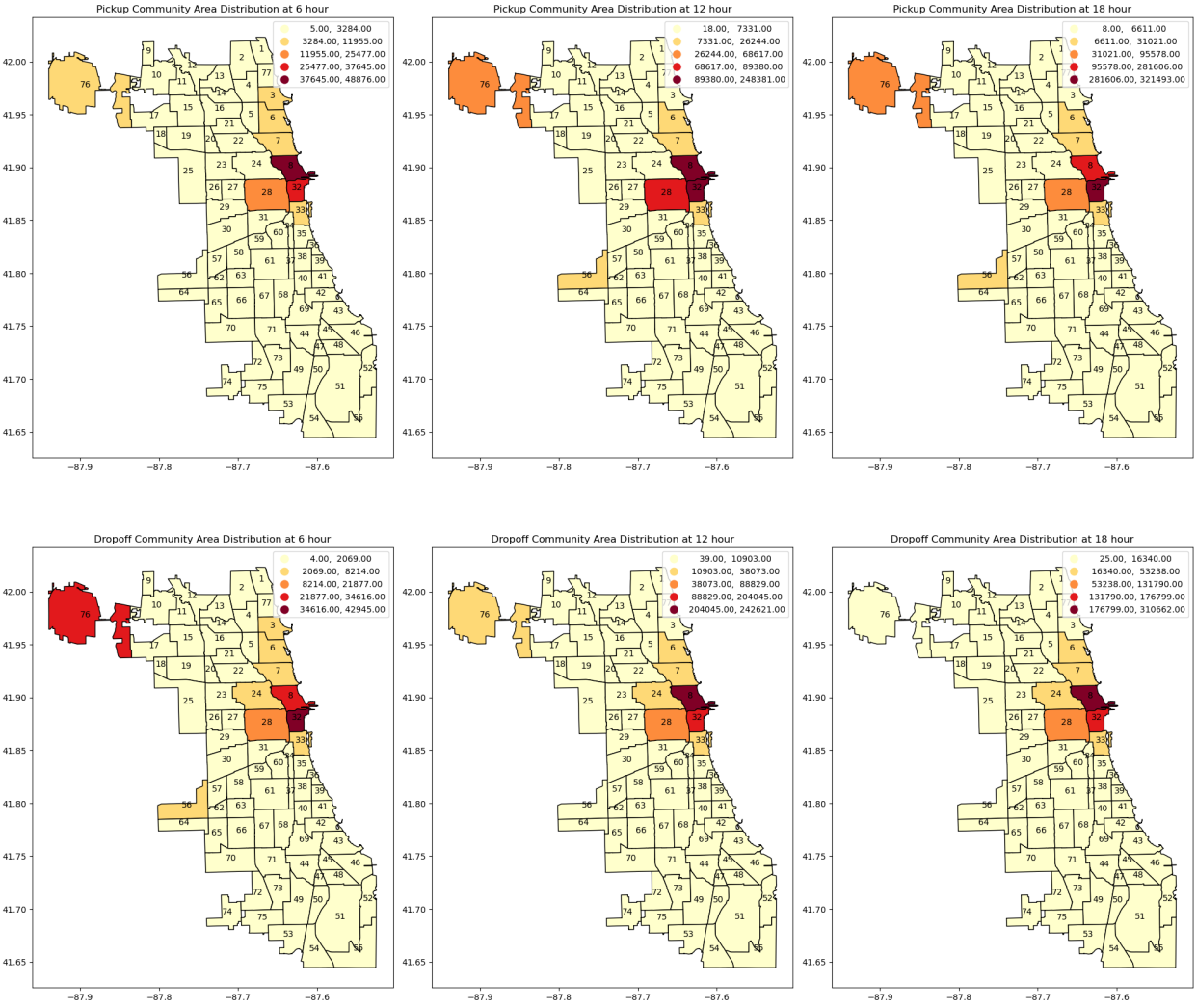


Here we visualize the ride patterns divided by the pickup and dropoff area during three key hours of the day (6 am, 12 pm, and 6 pm). Chicago city is divided into 77 districts and generally, we observe several key districts with the highest number of rides throughout the day. At these three hours of the day, the highest number of taxi pickups occur in districts 8, 28 and 32 which are the Loop, Near West Side, and Near North Side districts. These are the downtown area of Chicago which means that they are the most economically active districts. Another notable district will be district 76 which is the O'Hare airport. Generally, we observe similar pickup and dropoff patterns during these hours, even during the pre-covid years as shown below:

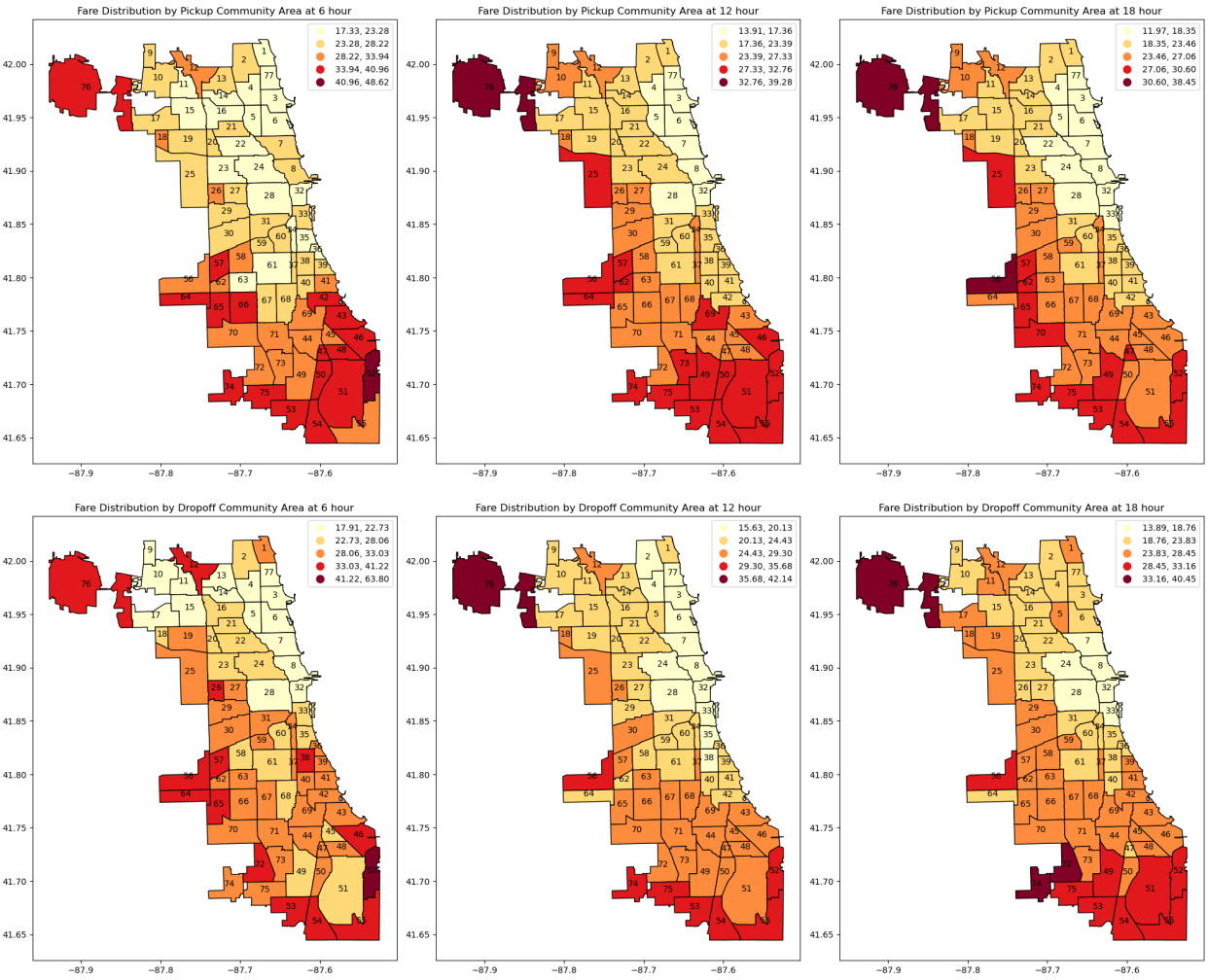
**Geographic Patterns of Ride 2018**



# Geographic Patterns of Ride 2019

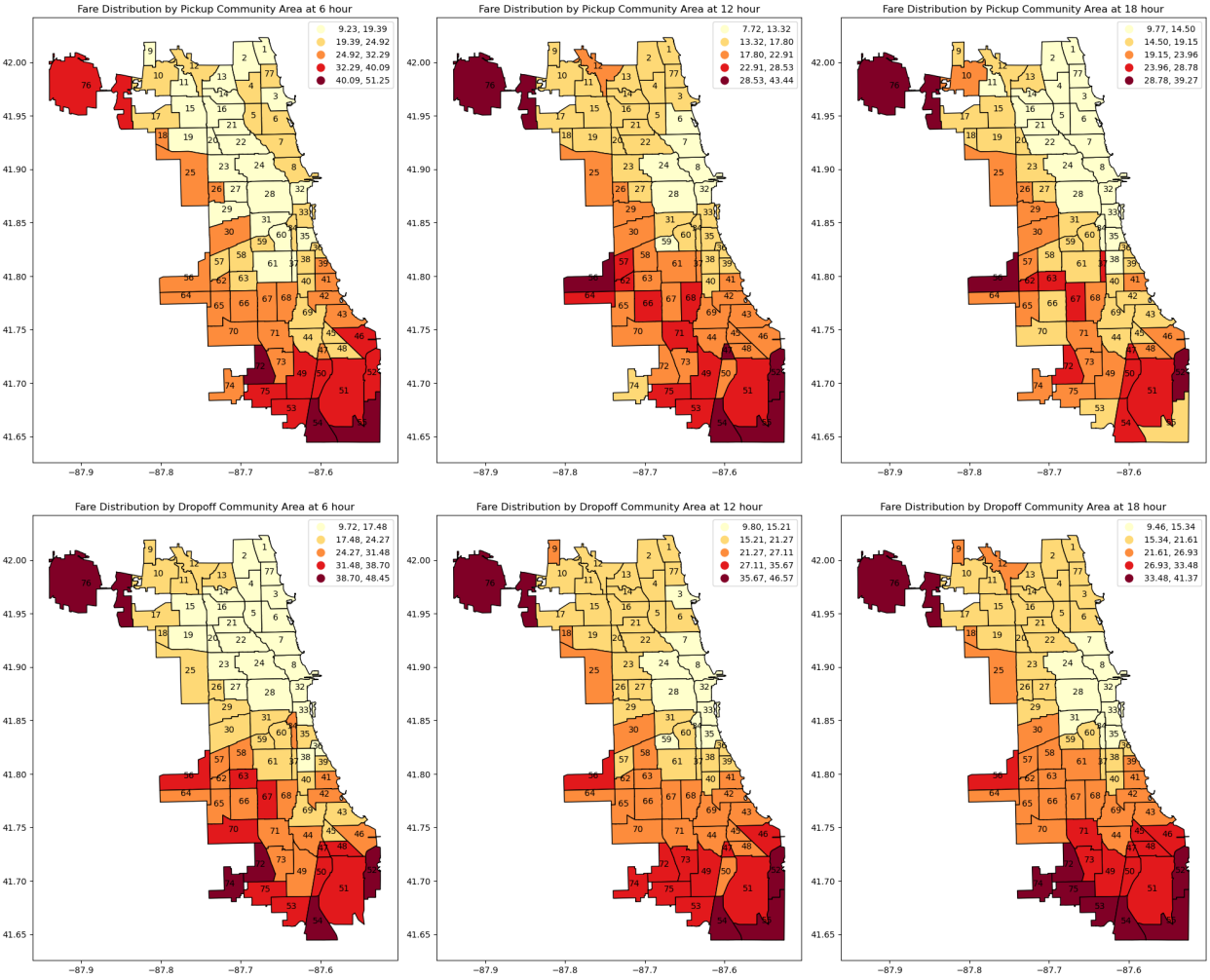


## Fare Patterns by Community Area

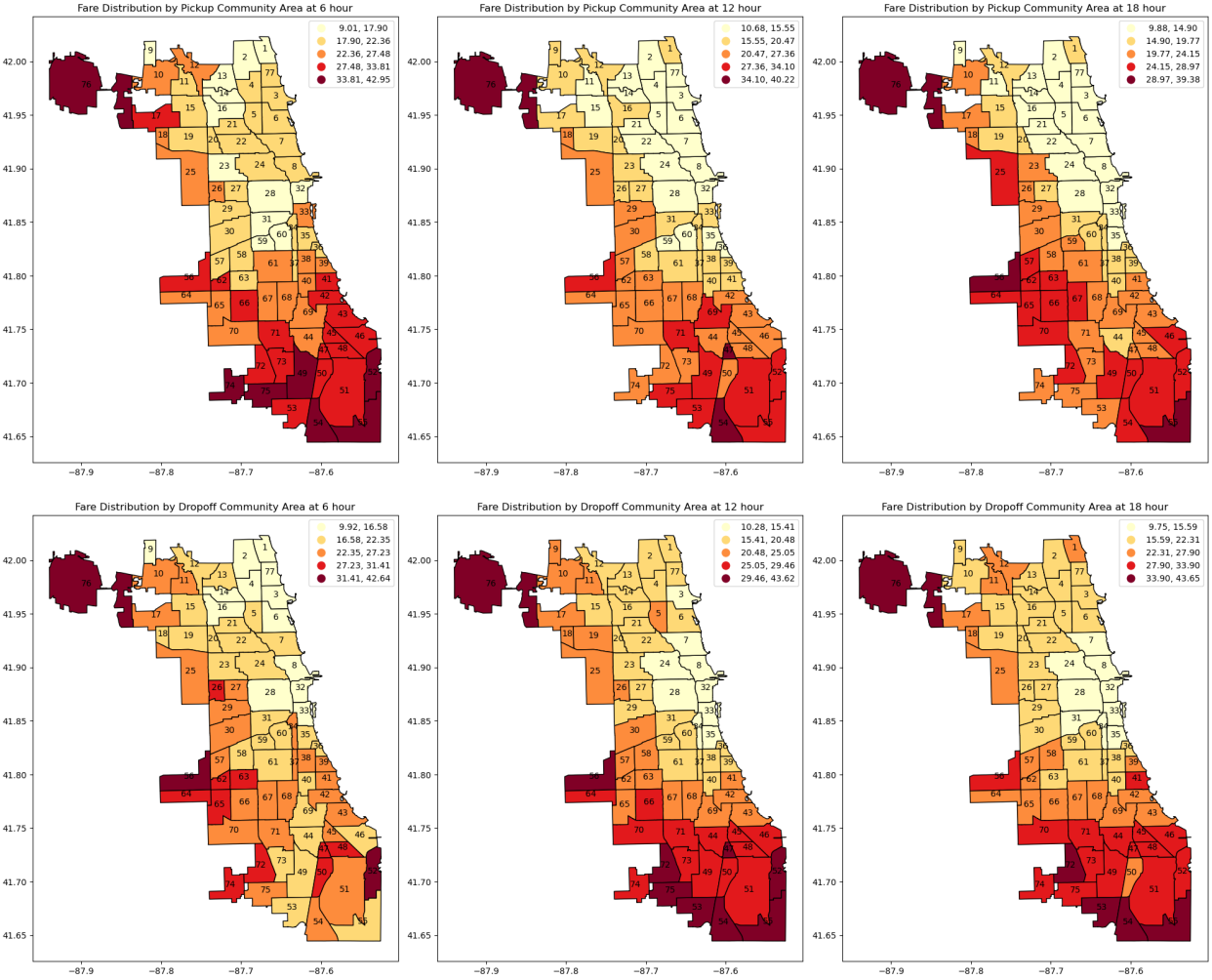



Next, we look at the fare distribution by geographical area during the same key hours as before. Here we observe higher-fare rides from the southern districts compared to the downtown areas. This suggests that people tend to travel longer distances from the southern part of the city, while those from the downtown areas tend to go for shorter rides. The average fare for trips starting and ending in the O'Hare airport (area 76) is also among the highest. As before, we visualize the fare distribution for the years 2018 and 2019 as controls and observe similar patterns.

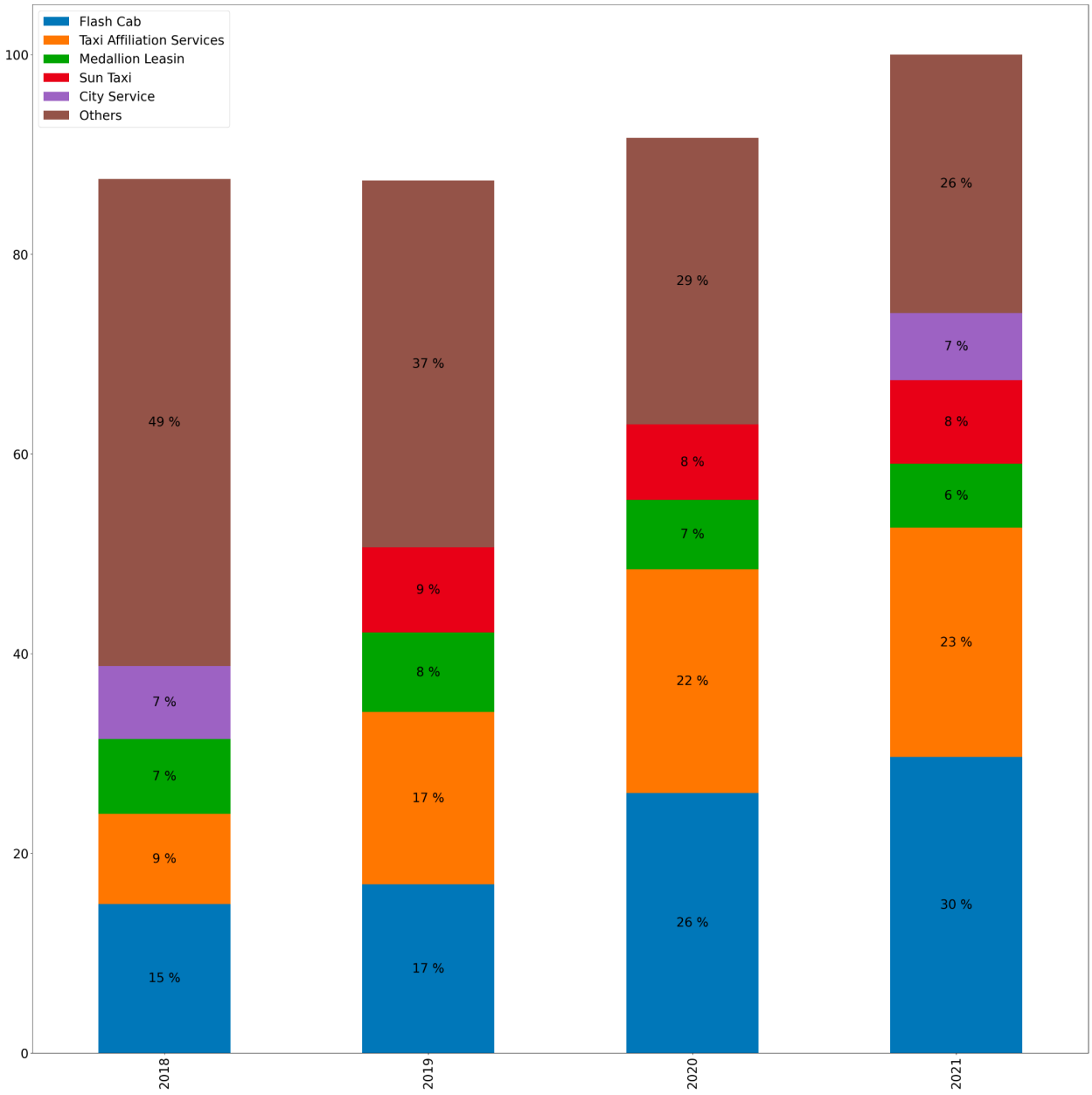
# Geographic Fare Patterns 2018



# Geographic Fare Patterns 2019



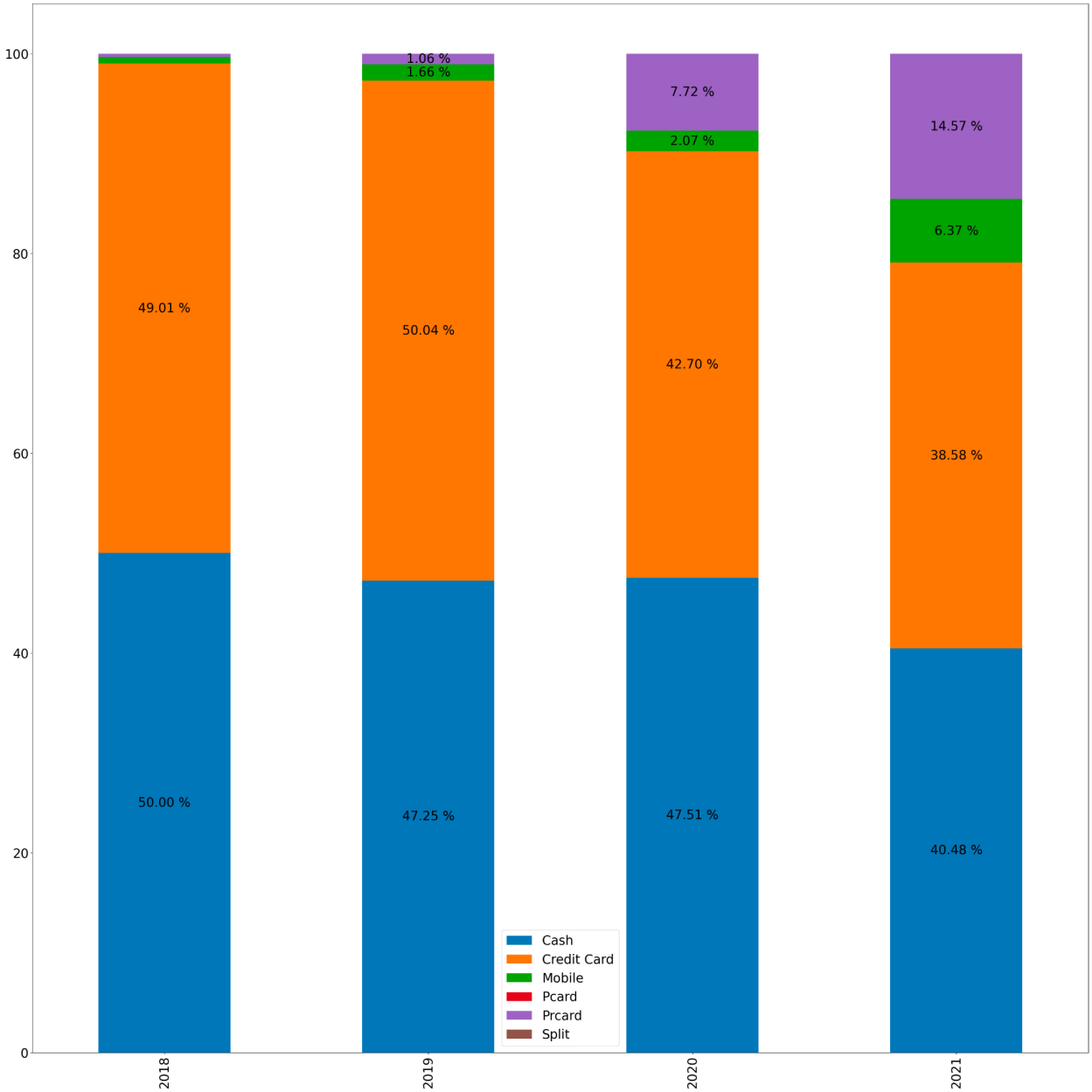
## Market Share



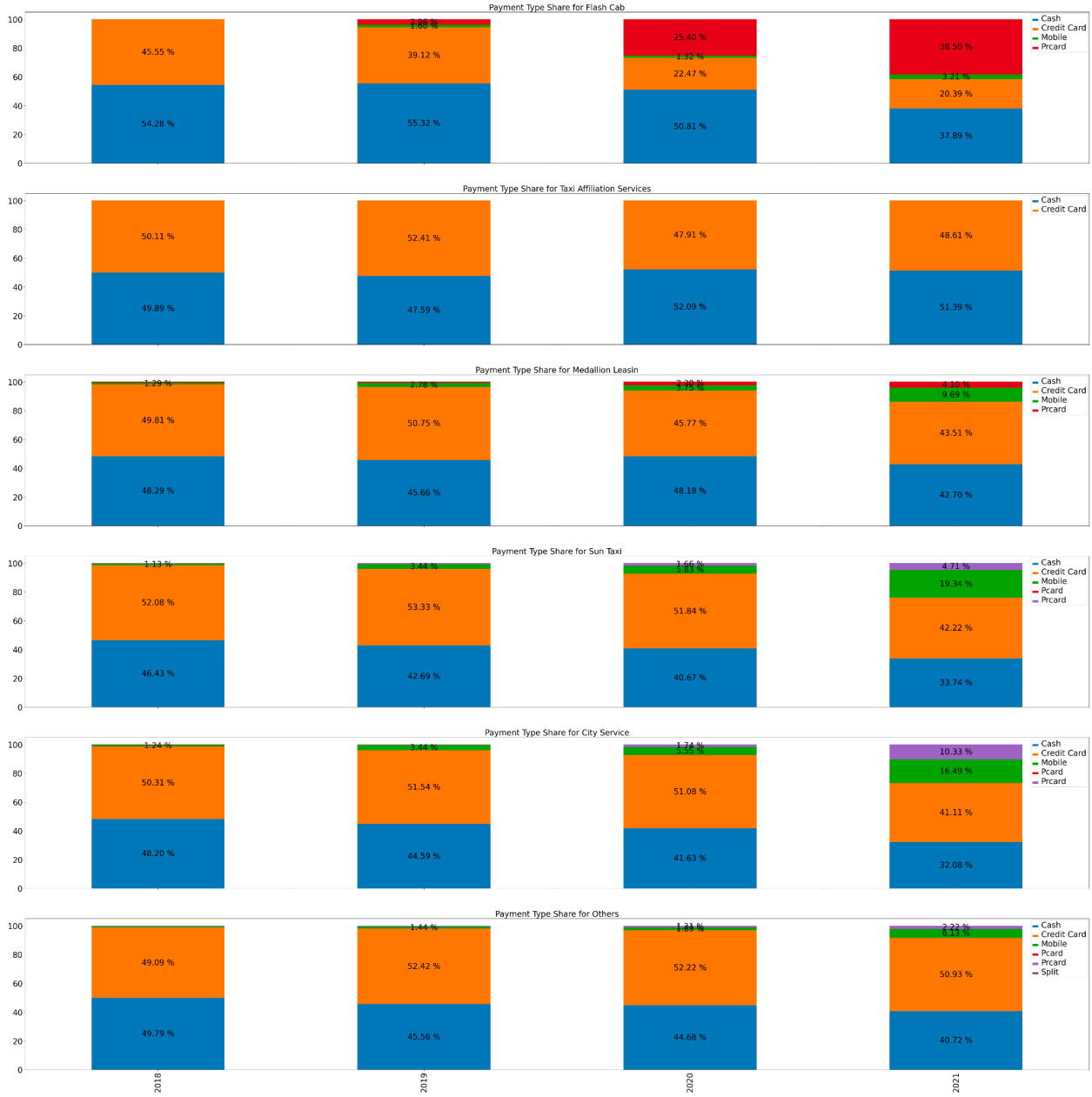

We visualize the market share of the taxi market for the years 2018, 2019, 2020, and 2021. Here we only display the top four market leaders for each year and group the rest as others. Generally, the pre-covid market shares in 2018 are distributed more evenly as evidenced by the higher collective market shares of smaller taxi companies. As the year progresses they consolidate into a couple of big companies, shown by the decreasing market share of ‘Others’. It is worth noting that the consolidation seems to speed up during the height of the pandemic and

lastly in 2021, most of the market shares of the smaller companies seem to go to Flash Cab and Taxi Affiliation Services.

**Payment Methods**




The payment type in general has a mixture of cash and non-cash payment with cash slowly declining from 2018 to 2021. Non-cash payment is split between credit card, mobile, and Prcard with credit card payment declining from 2018 to 2021 as riders prefer the use of mobile or Prcard payment.




In general most of the top-performing companies have a lower proportion of cash, followed by payment using Prcard or Pcard (except for taxi affiliation services). Most notably perhaps is Flash Cab which has the highest usage of Prcard proportion. This might suggest a strong incentive from the riders to utilize this method of payment which results in them grabbing a significant percentage of the market share

# Fare Prediction

After visualizing key statistics from the data, we move on to the problem of fare prediction. Firstly, in order to measure the performance of our model, we will define two metrics:

- Mean Absolute Error:
  - $MAE = \frac{\sum_{i=1}^{N} |\hat{y}_i - y_i|}{N}$
  - Here $y_i$ is the ground truth fare for the i-th trip, $\hat{y}_i$ is the i-th prediction, and $N$ is the total number of trips
  - Mean Absolute Error (MAE) measures the average difference between our prediction and the ground truth fare
- Root Mean Squared Error:
  - $RMSE = \sqrt{\frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}{N}}$
  - Same as before, $y_i$ is the ground truth fare for the i-th trip, $\hat{y}_i$ is the i-th prediction, and $N$ is the total number of trips
  - Root Mean Squared Error (RMSE) measures the standard deviation of the error in our prediction

High RMSE indicates high variability in the error of our model. As such, even with low MAE, high RMSE indicates a non-consistent prediction error. Therefore for a model to be reliable, it has to have low MAE coupled with low RMSE.

Next, we need to define a benchmark model. As mentioned before, Chicago has a fare guideline that depends on the trip miles with an additional charge per 36 seconds of travel and whether someone is using a non-cash payment or whether the trip involves going to or from the airport. Therefore we will define our benchmark model as

$$Fare = 3.25 + miles * 2.25 + \frac{trip\ seconds}{36} * 0.20 + 0.50 \times non\ cash\ payment + 4 \times airport$$

We randomly split 30% of our dataset as the held-out test set and further split the remaining 70% into 75% training and 25% validation set. The training set is used to train our model, the validation set is used for feature and model selection, and the test set is used to compare our final models against the benchmark model

We will test several variants of the benchmark models by either including or not including the additional 50 cents from cash payments as well as the extra $4 from trips to the airport. Below is the performance of our benchmark models on the validation set.

| Credit Card | Airport Trip | MAE | RMSE |
|---|---|---|---|
| Yes | Yes | 6.22 | 8.58 |
| Yes | No | 5.68 | 7.89 |
| No | Yes | 5.98 | 8.40 |
| No | No | 5.43 | 7.73 |

The result varies between RMSE of 7.73 to 8.57 and between MAE 5.43 to 6.22. This variation leads us to question the extent of variation in the guideline amounts and whether the different company is charging different rate for an additional mile or trip seconds. This also puts into question whether taxi providers are charging for non-cash payments and whether they absorb the additional $4 charge for leaving the airport.

We think that a good way to answer some of these questions is to start with a simple linear regression model so that we can see the importance of some of these variables before building more complicated models. We begin by comparing the coefficient of the models with trip miles and trip seconds vs models that incorporate non-cash payment and airport trips

## Linear Regression Model

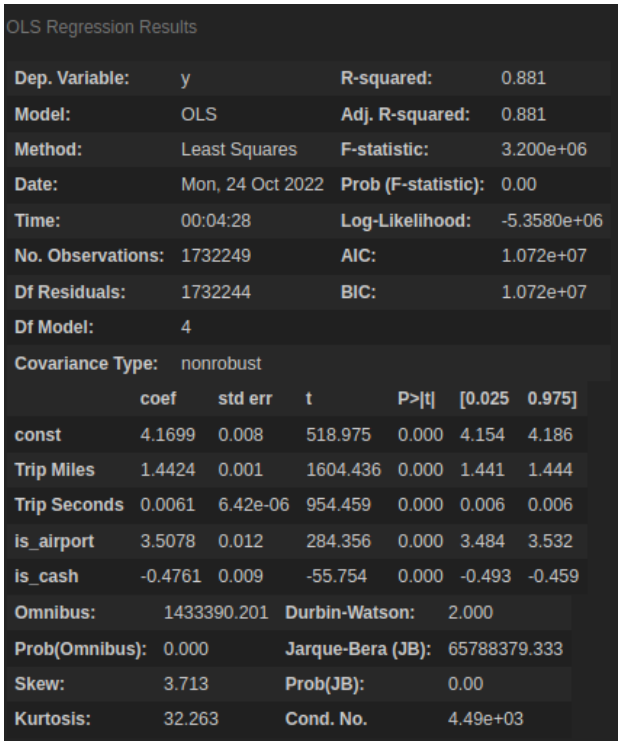
OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | y | **R-squared:** | 0.875 |
| **Model:** | OLS | **Adj. R-squared:** | 0.875 |
| **Method:** | Least Squares | **F-statistic:** | 6.066e+06 |
| **Date:** | Mon, 24 Oct 2022 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 00:04:27 | **Log-Likelihood:** | -5.3988e+06 |
| **No. Observations:** | 1732249 | **AIC:** | 1.080e+07 |
| **Df Residuals:** | 1732246 | **BIC:** | 1.080e+07 |
| **Df Model:** | 2 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **const** | 3.8246 | 0.007 | 544.269 | 0.000 | 3.811 | 3.838 |
| **Trip Miles** | 1.5357 | 0.001 | 1787.662 | 0.000 | 1.534 | 1.537 |
| **Trip Seconds** | 0.0063 | 6.54e-06 | 958.465 | 0.000 | 0.006 | 0.006 |

| | | | |
|---|---|---|---|
| **Omnibus:** | 1471455.104 | **Durbin-Watson:** | 1.999 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 66277527.113 |
| **Skew:** | 3.884 | **Prob(JB):** | 0.00 |
| **Kurtosis:** | 32.290 | **Cond. No.** | 2.49e+03 |

The result above indicates the summary of a simple linear regression model with the variables trip miles and trip seconds. Firstly we see that the trip seconds coefficient at 0.0063 is similar to the recommended guideline by the city which charges 0.20 for every 36 seconds passed (0.0063 * 36 = 0.2268). The constant value of 3.82 is a bit off from the surcharge of 3.25 and the additional mile charge amounts to 1.54. The model also indicates a good fit with Adj. R-squared of 0.875. Next, we try to incorporate the airport trip information along with non-cash payment information



OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | y | **R-squared:** | 0.881 |
| **Model:** | OLS | **Adj. R-squared:** | 0.881 |
| **Method:** | Least Squares | **F-statistic:** | 3.200e+06 |
| **Date:** | Mon, 24 Oct 2022 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 00:04:28 | **Log-Likelihood:** | -5.3580e+06 |
| **No. Observations:** | 1732249 | **AIC:** | 1.072e+07 |
| **Df Residuals:** | 1732244 | **BIC:** | 1.072e+07 |
| **Df Model:** | 4 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **const** | 4.1699 | 0.008 | 518.975 | 0.000 | 4.154 | 4.186 |
| **Trip Miles** | 1.4424 | 0.001 | 1604.436 | 0.000 | 1.441 | 1.444 |
| **Trip Seconds** | 0.0061 | 6.42e-06 | 954.459 | 0.000 | 0.006 | 0.006 |
| **is_airport** | 3.5078 | 0.012 | 284.356 | 0.000 | 3.484 | 3.532 |
| **is_cash** | -0.4761 | 0.009 | -55.754 | 0.000 | -0.493 | -0.459 |

| | | | |
|---|---|---|---|
| **Omnibus:** | 1433390.201 | **Durbin-Watson:** | 2.000 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 65788379.333 |
| **Skew:** | 3.713 | **Prob(JB):** | 0.00 |
| **Kurtosis:** | 32.263 | **Cond. No.** | 4.49e+03 |

Despite the strong multicollinearity, the coefficient of the variables seems to be convincing enough. For example, the trip second is close to the 0.2 charges per 36 seconds (0.0061 * 36 = 0.22), while paying with cash saves us roughly 0.47 cents, which is similar to the electronic payment surcharge of 0.50 cents in the Chicago city guidelines. When the pickup point or dropoff is at the airport, there is an additional 3.5 to the fare as well, which is close to the guideline of a $4 fee for entering the airport.

The cost per mile, however, is a bit off from the baseline of 2.25 per mile and the minimum fare as indicated by the constant parameter is off from the 3.25 baseline as indicated inin the guidelines. Again, these discrepancies might suggest measurement errors in the trip miles or seconds, or perhaps different companies charging different base charges or cost per mile.

In addition to the above results, we have tried a linear model that considers different costs per mile for a different company and in general, we do not see an improvement to the model fit and the learned model coefficient seems to be relatively similar for each company. We present the

validation results of all three linear models, along with the best-performing benchmark model below:

| Model | MAE | RMSE |
|---|---|---|
| Benchmark | 5.43 | 7.73 |
| Trip per mile + Trip Duration | 2.88 | 5.47 |
| Trip per mile + Trip Duration + non-cash payment + airport | 2.88 | 5.35 |
| Trip per mile for different company + Trip Duration + non-cash payment + airport | 2.88 | 5.32 |

In general, a linear model can outperform our benchmark model quite significantly. This might suggest that the additional charge for the different components of the trip that we learned from the Chicago city guidelines do not match the actual charges on average. We do not see any performance gain from including airport trip information and non-cash payment or by incorporating the taxi company identity, but this might also be because of the limitation of a linear regression model. We decided to keep the non-cash payment and airport trip information since they are rather simple to model but choose to leave out the company information which is one-hot encoded in our model

### Tree-based Ensemble Regressor

After trying out the linear model, we decided to try RandomForest and GradientBoosting models on the four features (trip per mile + trip duration + non-cash payment + airport trip).

| Model | MAE | RMSE |
|---|---|---|
| Benchmark | 5.43 | 7.73 |
| Linear Regression | 2.88 | 5.35 |
| GradientBoosting Regressor | 1.38 | 3.64 |
| RandomForest Regressor | 1.13 | 3.43 |
| XGBoost Regressor | 1.07 | 3.20 |

We generally observe a much better validation set performance than the linear model with the ensemble classifier. The tree-based ensemble models might capture non-linear relationships in the data better than the linear regression model and the benchmark model. In order to improve

this model further, we decided to one-hot encode the pickup community area and the dropoff community area, since based on the visualization, they might influence the fare calculation as well. It is not clear if the visualized relationship was purely due to distance or if there is another factor at play, such as the likelihood of having more passengers for trips from a certain area, or other non-obvious relationships that might be present in the data. Regardless, we try to include this in the data and also consider missing districts as a feature on its own.

| Model | MAE | RMSE |
|---|---|---|
| GradientBoosting Regressor | 1.38 | 3.64 |
| RandomForest Regressor | 1.13 | 3.43 |
| XGBoost Regressor | 1.07 | 3.20 |
| GradientBoosting Regressor with Community Area | 1.45 | 3.60 |
| RandomForest Regressor with Community Area | 0.978 | 3.00 |
| XGBoost Regressor with Community Area | 1.07 | 3.03 |

While we do not observe a performance improvement on the GradientBoosting Regressor model, we observe a significant performance improvement on the RandomForest Regressor model with the pickup and dropoff community area. We decided to use this model as our final model and compare its performance on the test dataset and other datasets against the benchmark classifier

| Dataset | XGBoost Regressor with Community Area MAE | XGBoost Regressor with Community Area RMSE | RandomForest Regressor with Community Area MAE | RandomForest Regressor with Community Area RMSE | Benchmark MAE | Benchmark RMSE |
|---|---|---|---|---|---|---|
| 2017 | 0.73 | 1.76 | 0.63 | 2.02 | 3.47 | 5.24 |
| 2018 | 0.69 | 1.63 | 0.60 | 1.81 | 3.66 | 5.61 |
| 2019 | 0.75 | 1.76 | 0.67 | 2.06 | 4.00 | 6.16 |
| 2020 | 0.80 | 1.96 | 0.75 | 2.32 | 5.87 | 3.93 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 2021 | 1.07 | 3.01 | 0.98 | 3.00 | 5.43 | 7.71 |

We observe a much better MAE and RMSE from our RandomForest Regressor compared to the benchmark. This means that on average our model has less than $1 error when predicting fares with a standard deviation of $3 and below on most of these datasets. XGBoost Regressor in general is not as good as RandomForest Regressor in terms of these metrics but it has faster implementation and takes up less storage.

**Neural Networks Embedding**

One drawback of our approach above is the use of so many categorical variables with a one-hot encoding that might not be optimal with tree-based approaches. Here we try to learn the embeddings of these features instead with a neural network and use the embedding results as features for an XGBoost Regressor model.

We use the following categorical variables:

- 5 payment type variables (including Unknown payment type)
- 79 community area type variables (including missing data)
- Months of the trip
- Days of the week

Next we fit a neural network with 3 hidden layers and encode all the categorical variables above into 2-dimensional vectors. We compare the performance of this neural network model against our tree-based models and also use the embedding vectors of the neural network model as features for both RandomForest Regressor and XGBoost Regressor.

| Model | MAE | RMSE |
|---|---|---|
| Neural Network | 0.976 | 2.738 |
| RandomForest Regressor with Community Area | 0.978 | 3.00 |
| XGBoost Regressor with Community Area | 1.07 | 3.03 |
| RandomForest Regressor with NN Embeddings | 0.84 | 2.62 |
| XGBoost Regressor with NN Embeddings | 0.83 | 2.54 |

The neural network model has a more stable performance compared with the other models from its lower RMSE but has higher MAE. Nevertheless, the embedding vectors help to improve the performance of our RandomForest Regressor and XGBoost Regressor.
Next, we focus on performing hyper-parameter tuning for the XGBoost model only since this model takes up a much smaller storage space compared to RandomForest Regressor (3.9 MB vs 8.4GB). We ran a randomized grid search on 50 different parameter configurations and report the results on the test set from 2017 to 2021:

| Dataset | XGBoost Regressor with NN Embedding MAE | XGBoost Regressor with NN Embedding RMSE | RandomForest Regressor with Community Area MAE | RandomForest Regressor with Community Area RMSE | Benchmark MAE | Benchmark RMSE |
|---|---|---|---|---|---|---|
| 2017 | 0.52 | 1.69 | 0.63 | 2.02 | 3.47 | 5.24 |
| 2018 | 0.50 | 1.56 | 0.60 | 1.81 | 3.66 | 5.61 |
| 2019 | 0.56 | 1.69 | 0.67 | 2.06 | 4.00 | 6.16 |
| 2020 | 0.62 | 2.06 | 0.75 | 2.32 | 5.87 | 3.93 |
| 2021 | 0.81 | 2.52 | 0.98 | 3.00 | 5.43 | 7.71 |

## Further Improvements

We believe that fare prediction can be used for determining a competitive price to set given an estimated trip distance, duration, and destination. Some care needs to be taken when using this prediction as a price estimate by setting some buffer related to the error distribution of our model and more experiments should be done to determine the optimal fare

Beyond this, there are several other improvements that can be made to this project. It is evident that our XGBoost Regressor is able to model the relationship between the trip fare against the other variables here relatively well. There can be several other variables that can be included in this model such as the companies or weather data that might also improve the model since the presence of rainfall or snow might affect the trip fare as well. Other improvements include dockerizing this whole project so that it is easier to reproduce and deploy

## Code Availability

The notebooks to reproduce our results and scripts for downloading, cleaning, and performing inference on the datasets used in this assignment are available at https://github.com/chrishendra93/taxi_analysis