Chicago Taxi Analysis

Christopher Hendra

Chicago City Taxi Data

- Each row describing the information of a single taxi ride made in the city of Chicago
- Details such as company name, trip miles, trip duration, payment type, fare, tips, tolls, are present
- Taxi ID information is present but masked
- Census tract information is often masked but pick up “community area” and dropoff “community area” is mostly present in the data

Chicago City Taxi Data: Tasks

1. Assuming that you are working for a mobility company that is evaluating launching ride-hailing services in Chicago, analyse, visualize and summarize the data for presentation to a city launch team.

2. Using the dataset, build a model to predict the fare for that trip.

Chicago City Taxi Data: Data Analysis and House-Keeping

- Identify anomalous ride events
- Describe ride and fare trend throughout the year, month, and time of the day as well as area within the city of Chicago
- Describe market leaders along with their shares in the taxi industry in Chicago

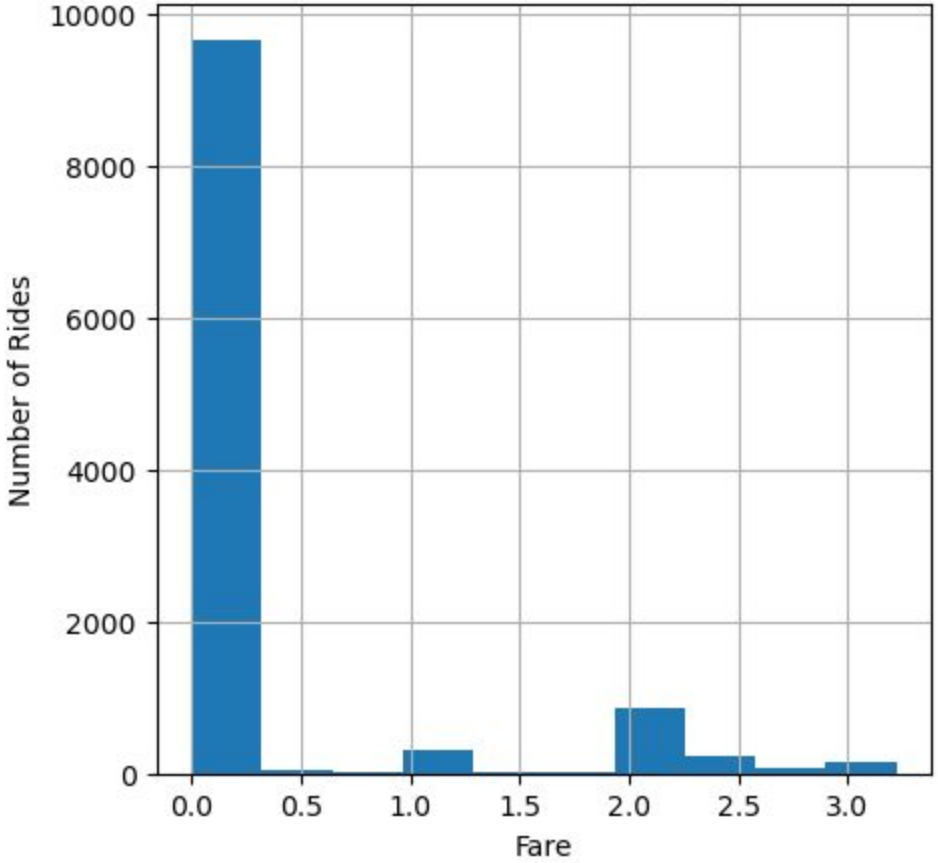
Chicago City Taxi Data: Fare Guidelines

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)

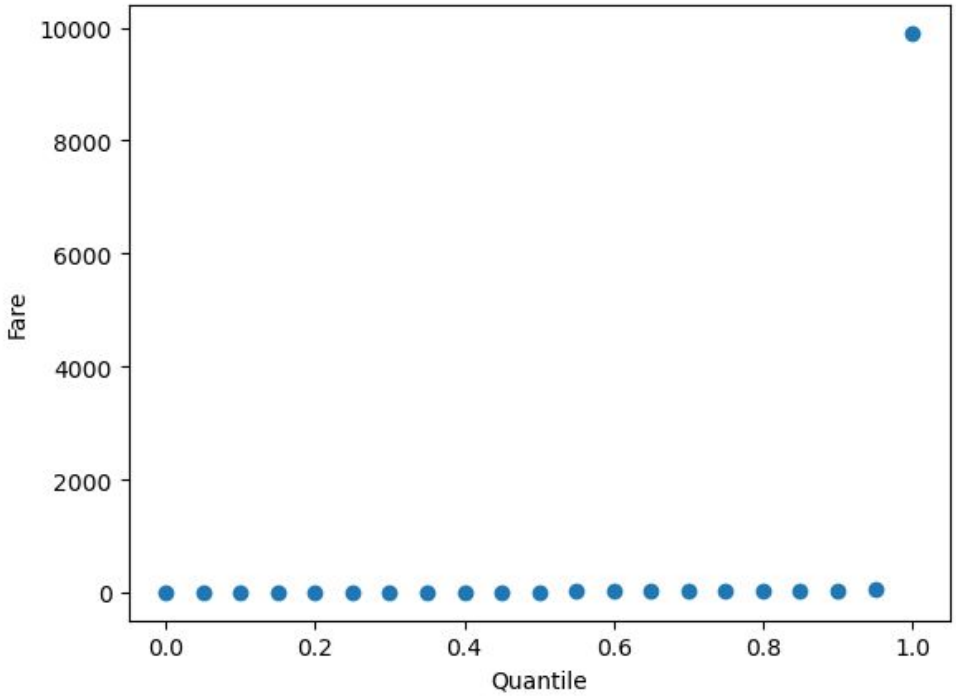
Chicago City Taxi Data: Ride Fare

- Each ride should have a minimum fare of $3.25
- Anything below this might not be representative of the general trend in the city
- Only 0.29% of rides in the 2021 taxi dataset is below this $3.25 baseline



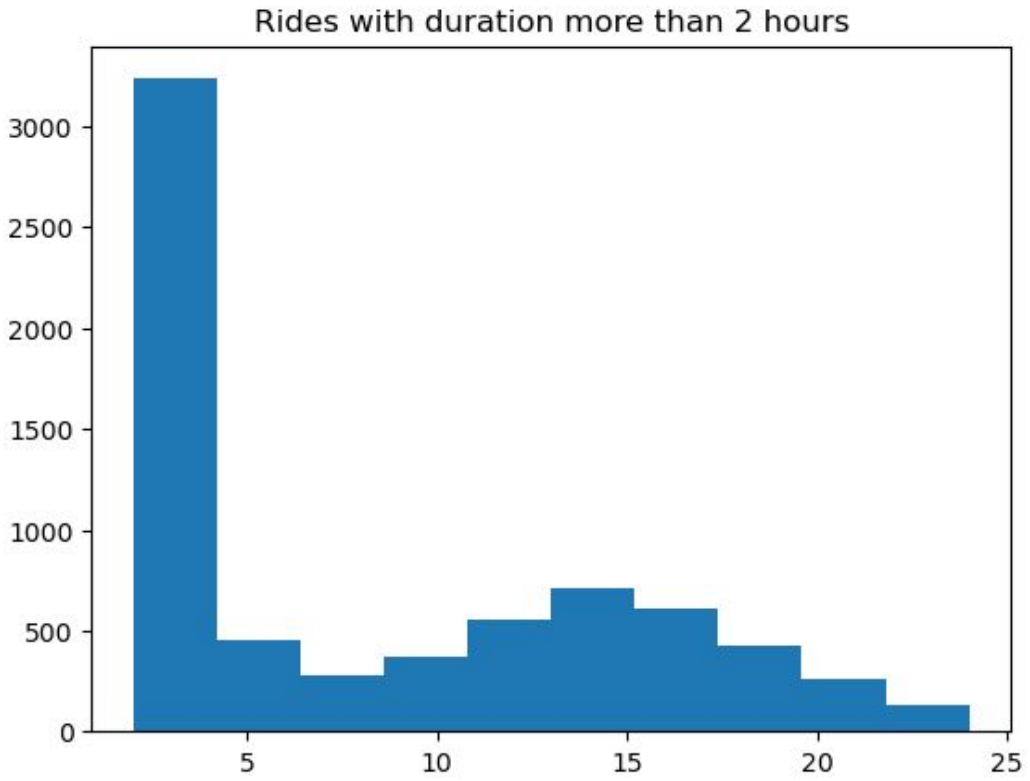
Chicago City Taxi Data: Ride Fare

- According to a report by Schaller consulting, an average long trip in Chicago will cost $25.30
- 95th percentile is about $47
- Set a threshold of $100 on our data



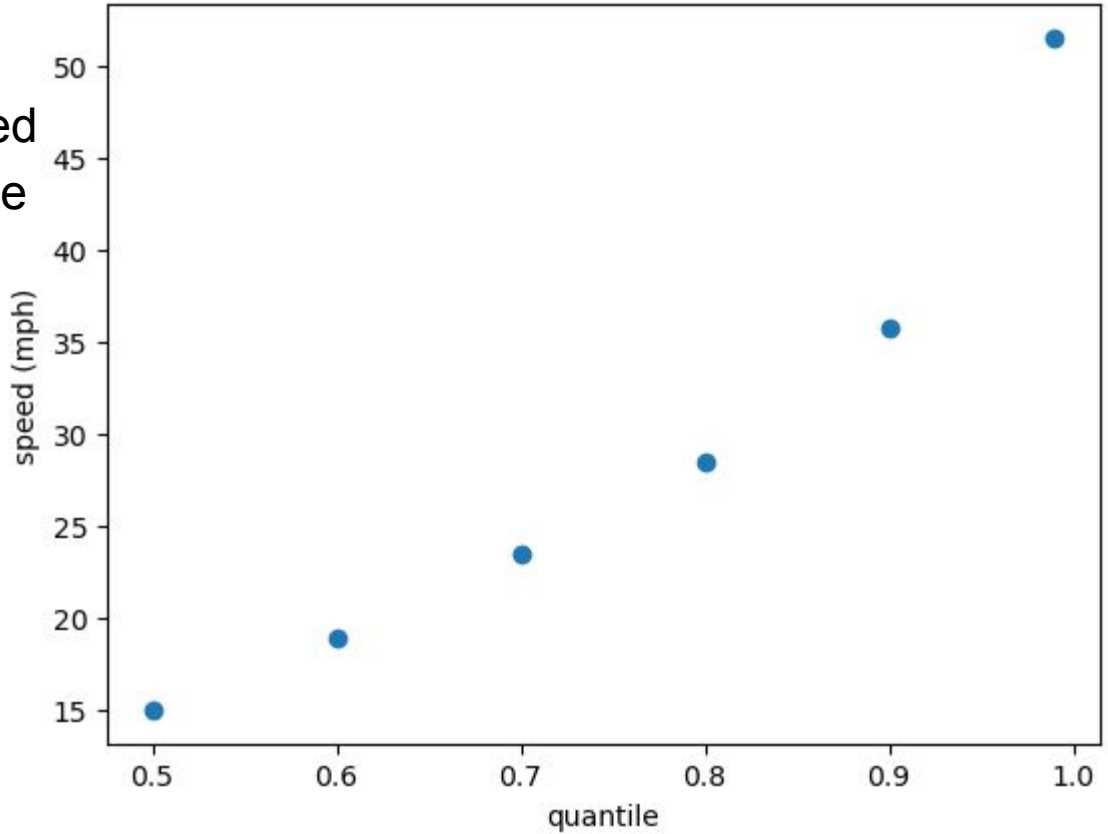
Chicago City Taxi Data: Ride Duration

- Chicago city is about 25 miles from north to south, 25 miles from east to west
- Trips in general should not take too long even when traffic is busy
- Exclude rides with duration more than 2 hours from our analysis (roughly 0.18% of the data)



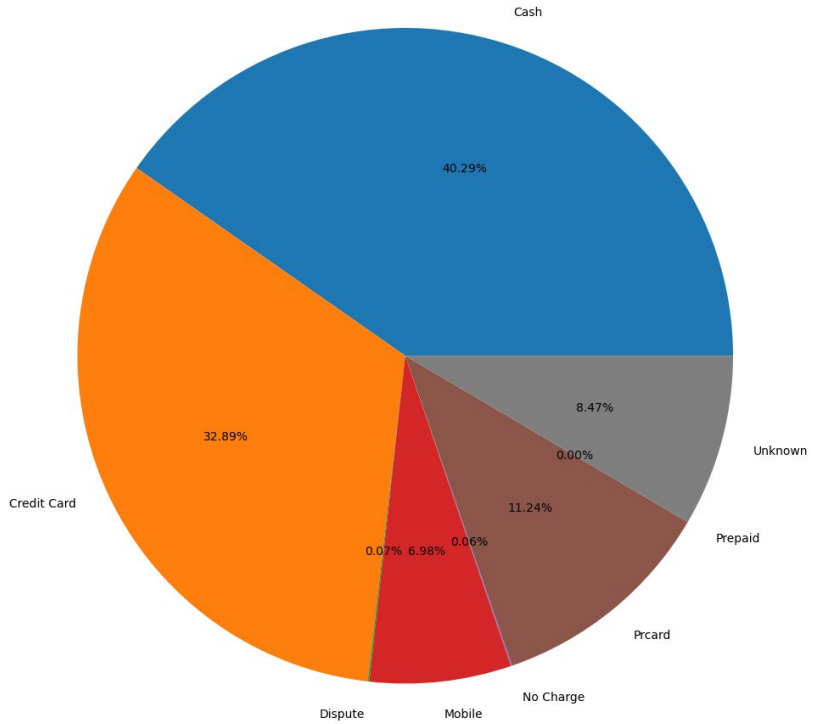
Chicago City Taxi Data: Ride Speed

- The city legal limit is 70 mph
- 0.12% of the total rides exceed this speed limit, so we exclude them from our analysis



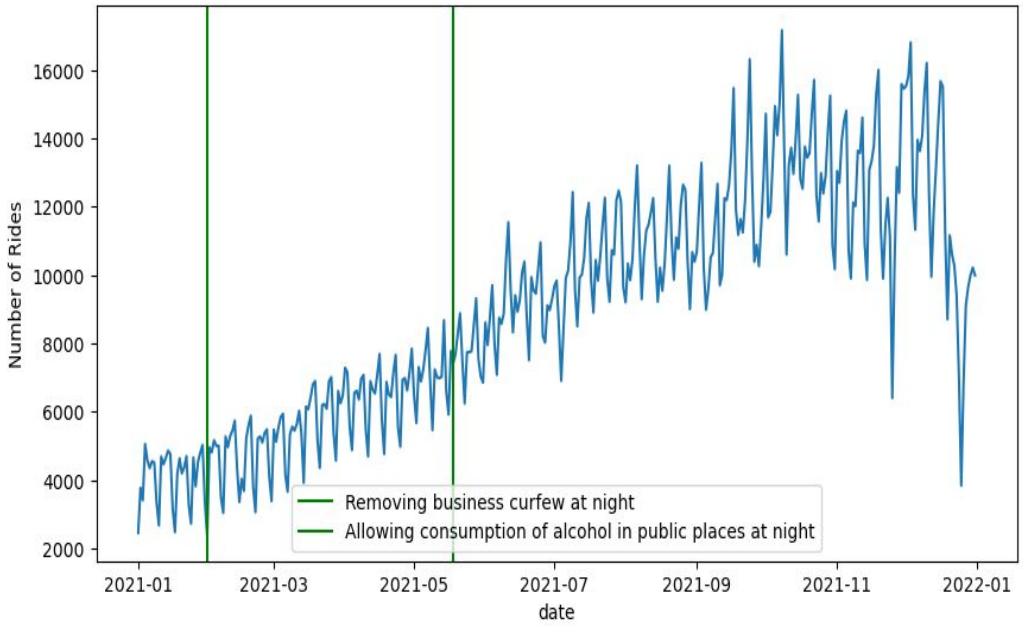
Payment Type

- Exclude disputed transactions or no charge
- They make up less than 0.2% of the dataset



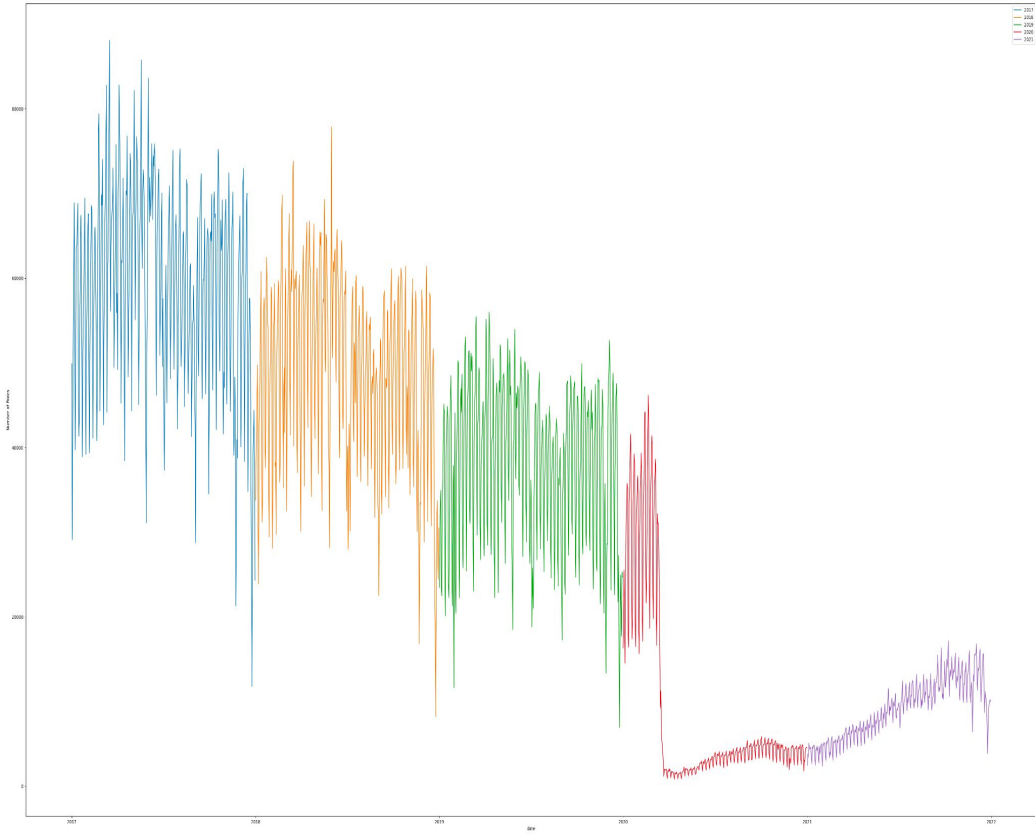
Chicago City Taxi Data: Rides throughout the year

- Generally number of rides increases throughout the year
- Seems to coincide with lifting of covid restrictions instead of general monthly trend



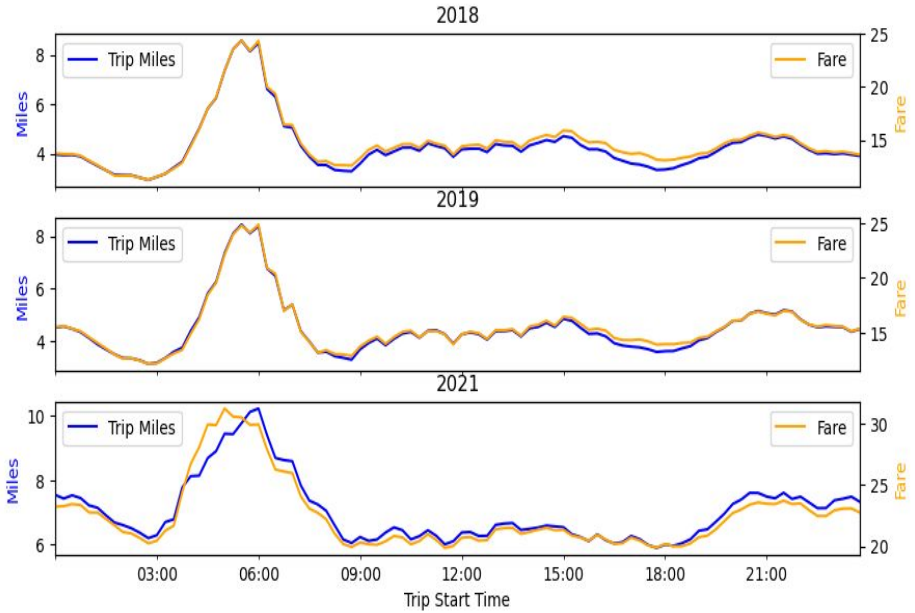
Chicago City Taxi Data: Rides throughout the years

- Rides tend to peak during summer
- Total number of rides are increasing as the market is recovering from covid



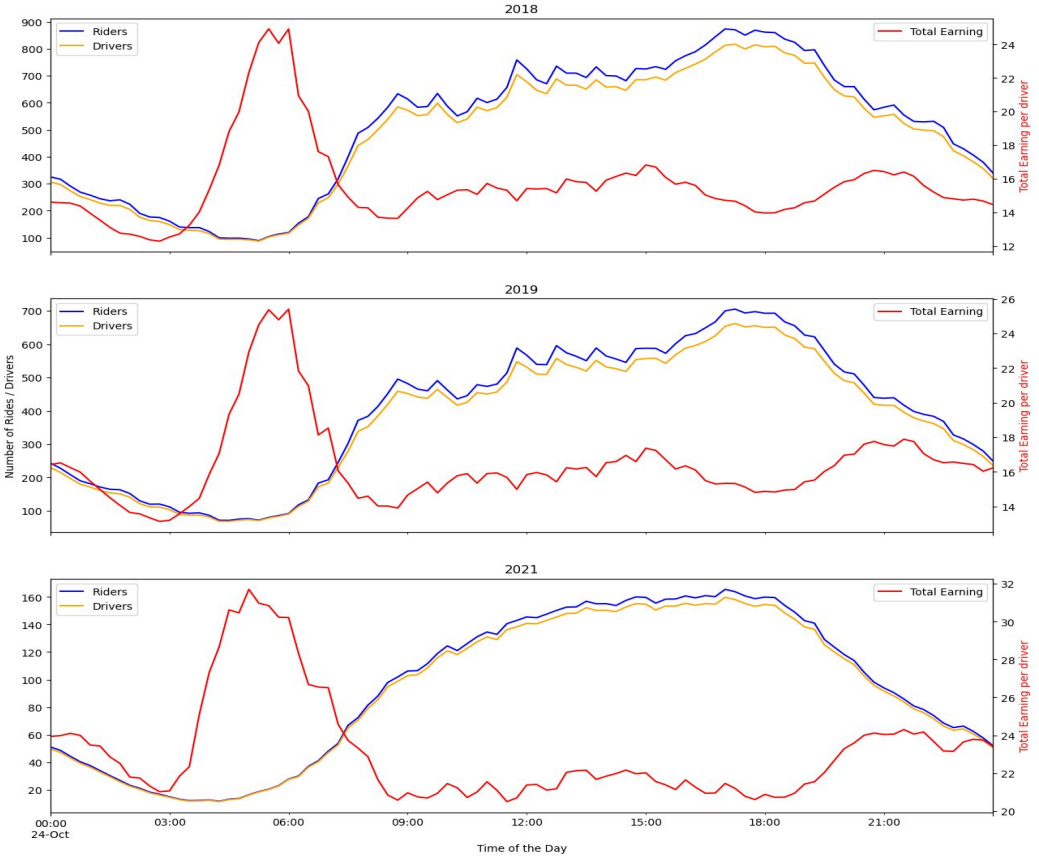
Chicago City Taxi Data: Daily Ride Patterns

- Trip fare and trip miles peak in the morning
- This translates to higher total earning per hour in the morning and smaller gap in the number of riders to drivers



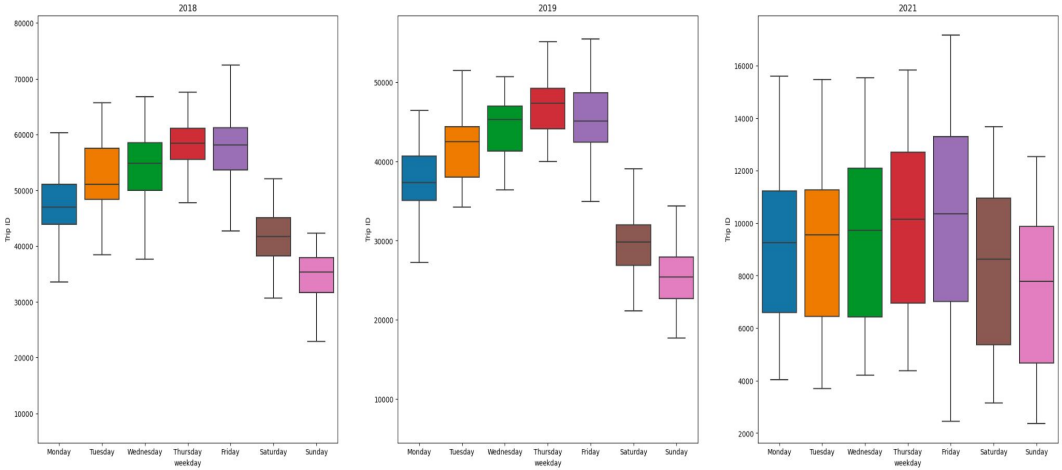
Chicago City Taxi Data: Daily Ride Patterns

- Trip fare and trip miles peak in the morning
- This translates to higher total earning per hour in the morning and smaller gap in the number of riders to drivers

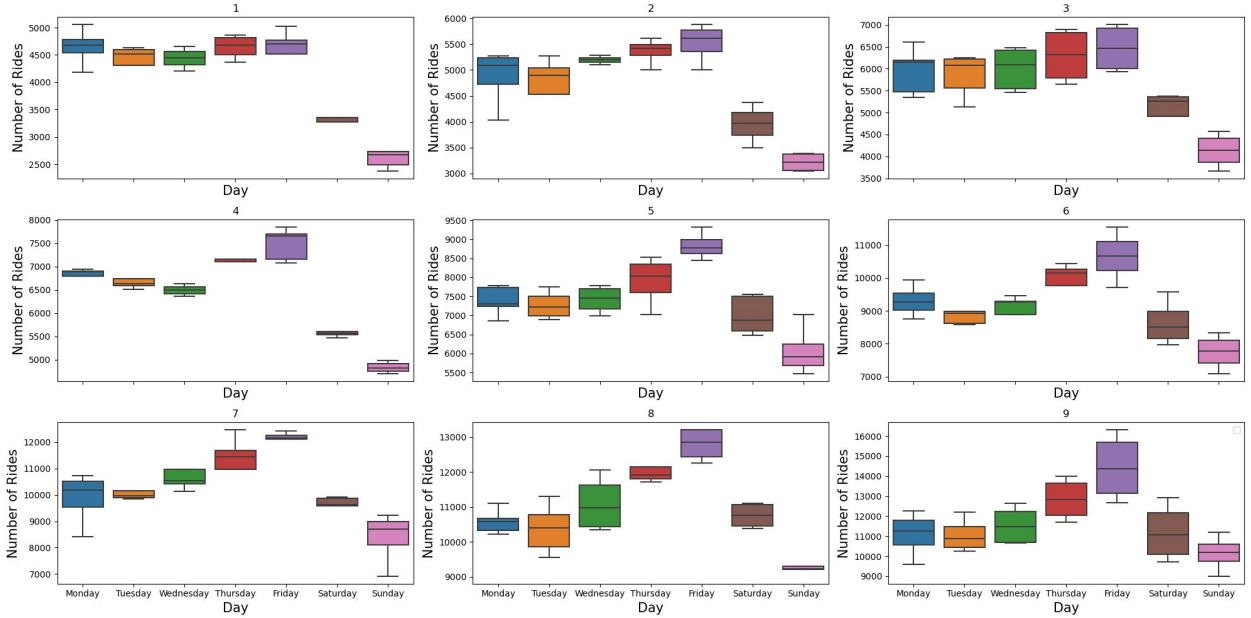


Chicago City Taxi Data: Weekly Ride Patterns

- Number of rides peak on Friday and fall dramatically during the weekend
- The boxplots overlap a lot for 2021 data but this is confounded by the lifting of covid restrictions



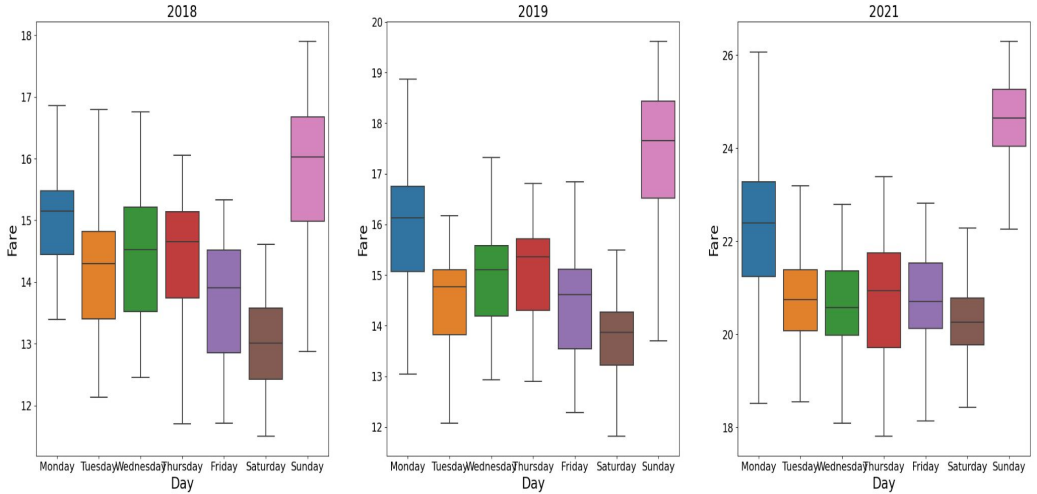
Chicago City Taxi Data: Weekly Ride Patterns



- Visualizing the ride patterns during each month of the year results in a ride pattern similar to that of the previous years

Chicago City Taxi Data: Weekly Fare Patterns

- Median fare is highest on Sunday, suggesting that people might travel more / longer distance on this day
- The median fare decreases from Monday to Saturday in general despite people taking more rides during the weekdays

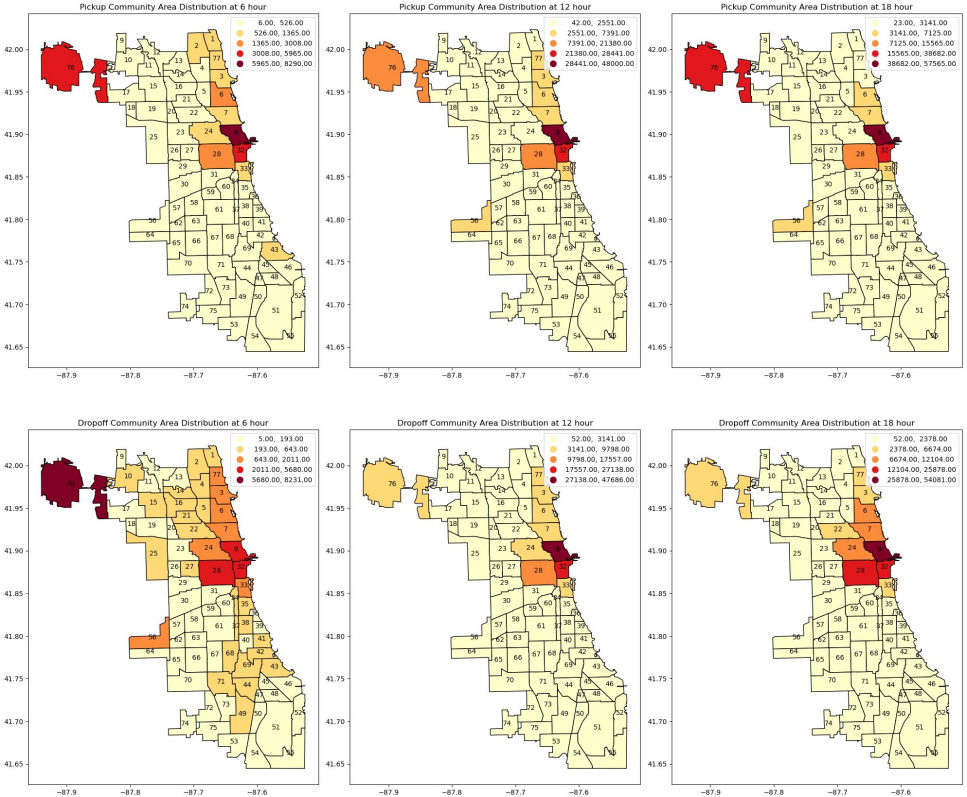


Chicago City Taxi Data: Ride Patterns by Community Area

- Divided into 77 community areas
- Some community area information is missing but can be imputed based on latitude and longitude information

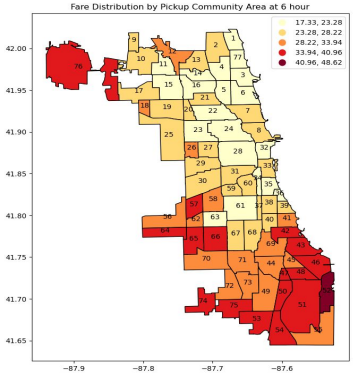
Chicago City Taxi Data: Ride Patterns by Community Area

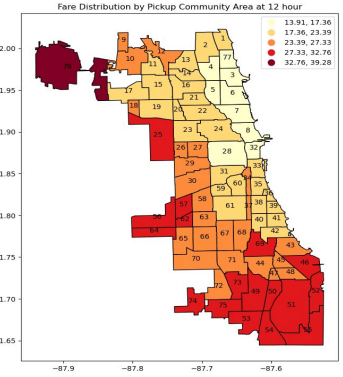
- Highest volume of ride start and ride end at the downtown area
- Followed by the O Hare Airport

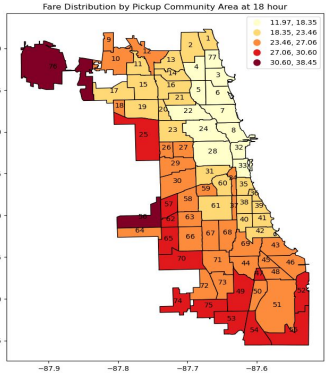


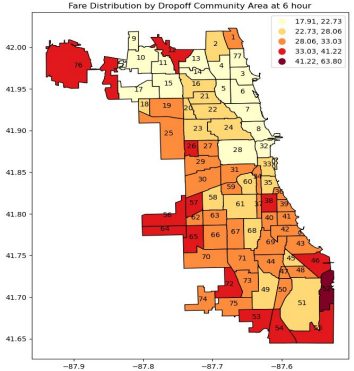
Chicago City Taxi Data: Fare Patterns by Community Area

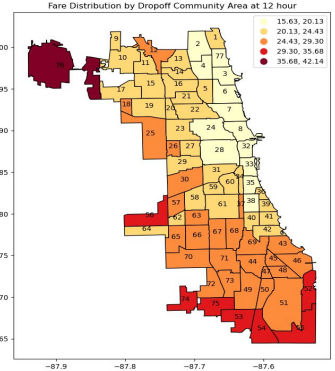
- Longest pattern of commute from people who live in the southern part of the city and the O Hare Airport

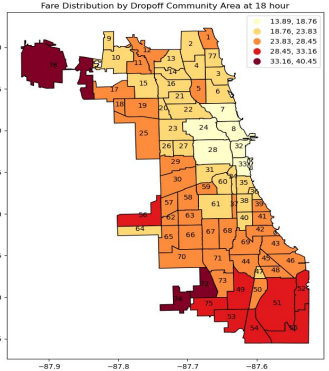






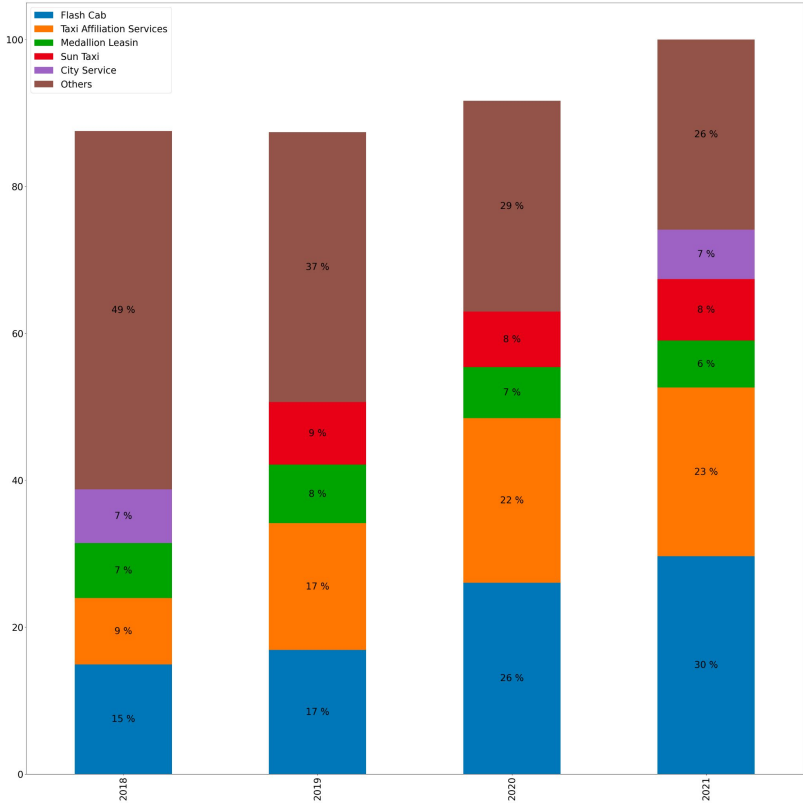






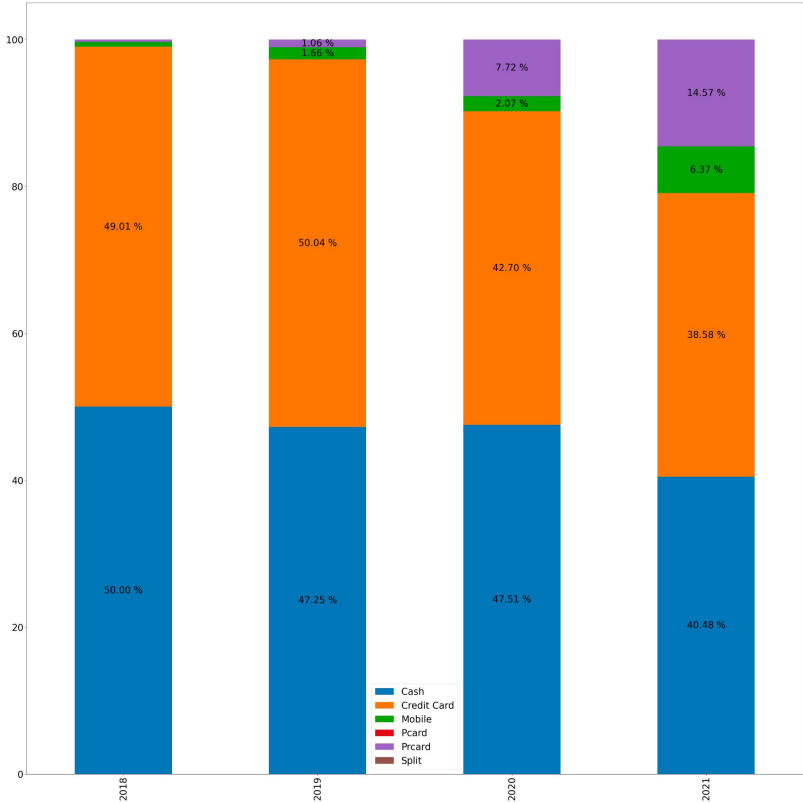
Chicago City Taxi Data: Market Share

- Generally, the pre-covid market shares in 2018 are distributed more evenly
- Market is consolidating through the pandemic with “Others” falling from 49% to 26%
- Flash Cab is gaining the most market share



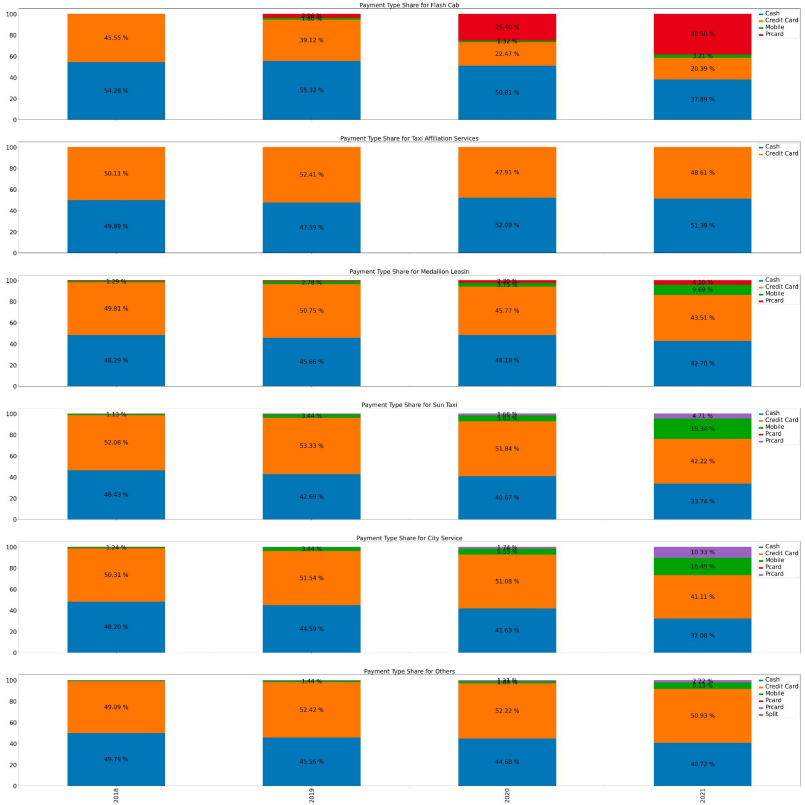
Chicago City Taxi Data: Payment Type

- Payment type is a mixture of cash and non-cash payment
- Cash payment is slowly declining over the years from 2018 to 2021 in favor of Prcard



Chicago City Taxi Data: Payment Type

- Payment type by the market leaders are similar
- Flash Cab, the one that gains the most market share, has a high, growing percentage of Prcard usage as it claims market leadership



Fare Prediction: Metrics

- Given the trip information, predict the fare for the trip
- Define two metrics:
  - Mean Absolute Error:
    - $\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$
    - Measures average error in prediction
  - Root Mean Squared Error:
    - $RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}}$
    - Measures the standard deviation of the error

Fare Prediction: Performance and Benchmarking

- Benchmark based on Chicago city guidelines:
  - Fare = 3.25 + miles * 2.25 + (trip seconds / 36) * 0.20 + 0.50 * non cash payment + 4 * airport
- Split the data 30% as held-out test set and validate on the other years datasets
- Split remaining 70% into 75% for training and 25% for validation set

Fare Prediction: Linear Model

- Try out linear model to see if it can perform better than benchmark
- Coefficients seem to be quite close to the chicago city fare guidelines

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

Fare Prediction: Linear Model

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

- Linear model can outperform benchmark quite a bit
- This indicates a slight difference in the actual amount being charged compared against the city guidelines
- Including non-cash payment and airport trip improves RMSE by a bit
- Including trip per mile for different company does not lead to significant performance improvement

Fare Prediction: Tree-based Ensemble Regressor

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

- Much better validation performance by tree-based ensemble regressors vs linear model
- This means that on average, our prediction is off by ~ $1.1 with a standard deviation of $3 in the error
- Try to incorporate location information!

Fare Prediction: Tree-based Ensemble Regressor

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

- Including pickup and dropoff area (including missing information) as one-hot vectors seem to improve the MAE and the RMSE of the predictions
- We decided to use these models for final validation

Fare Prediction

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

- RandomForest model generally performs better than XGBoost model but has slightly higher RMSE
- Sklearn implementation takes up a lot of storage space
- In general our prediction is only off by $1.00 on average but RMSE is still rather large

Fare Prediction: Improvement

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**

- We use Neural Network embedding as features for tree-based models
- We manage to improve the performance on validation dataset significantly

Fare Prediction: Improvement

Dataset	XGBoost Regressor with NN Embedding MAE	XGBoost Regressor with NN EmbeddingRMSE	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

- Test set performance is much better than the previous RandomForest model and model is significantly smaller (3.6 MB vs 8.5 GB)