MB4011 Business Intelligence and Analytics

FINAL PROJECT

STUDY CASE OF ARRO.INC

Cliff Richardo 19020293

Introduction

Background

As the third most populous city in the United States, people in Chicago commute almost every day using public transportation. Thus, convenience and ease of commuting are what people in Chicago considered the most before ordering a taxi. This orientation is met with the rising popularity of ride-sharing and/or ride-hailing platforms that accommodates customer demands.

Business Issue

The personal public transportation industry in Chicago is mastered by online ride-hailing taxi companies rather than local taxis provided by the government.

as part of public policymakers, researchers want **to deliver insight for ARRO to be able to compete with the famous ride-hailing company in Chicago**

Research Objective

OBJECTIVE 1

Generate business opportunities for ARRO from the Chicago taxi trip data

OBJECTIVE 2

Analyze the behavior of taxi users in Chicago to give recommendations in order to improve ARRO users' convenient

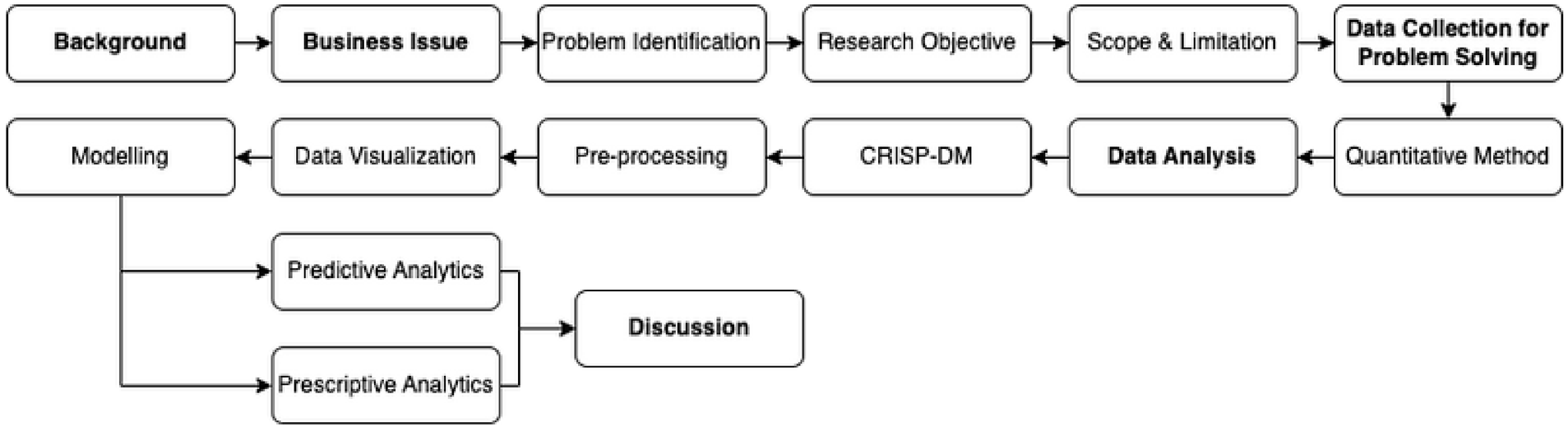
Project Scope and Limitations

- Geographic: Chicago
- Outcome: User behavior, Fare rate, Potential pick-up locations
- Problem Owner: ARRO Taxi Company

Research Design

Methodology

Research Design Flowchart



Research Design

Data Collection

BigQuery Public Dataset: Chicago Taxi Trips

Local taxis in Chicago below supervision of the local government

unique_key	taxi_id	trip_start_timestamp	trip_end_timestamp	trip_seconds	trip_miles
pickup_census_tract	dropoff_census_tract	pickup_community_area	dropoff_community_area	fare	tips
tolls	extras	trip_total	payment_type	company	pickup_latitude
pickup_longitude	pickup_location	dropoff_latitude	dropoff_longitude	dropoff_location	

23 Attributes

Research Design

Variable Identification

Correlation

- to generate analytics regarding fare on predictive and prescriptive analytics
- to generate analytics regarding fare and payment on prescriptive analytics
- to detect outliers in the dataset on data preprocessing
- provides the exact location of taxi pickup that will be useful to create data visualization (heatmap) on data analysis

Data Types

- string
- float
- integer
- timestamp

unique_key	trip_start_timestamp	trip_end_timestamp	trip_seconds	trip_miles
trip_total	payment_type	company	pickup_latitude	pickup_longitude

10 Attributes

Data Analysis

Data Pre-processing

DEFINING OUTLIERS

1. Records that contain **NULL values**: contains 117,247,322 records.
2. Records that contain **0 'trip_second'** values: contains 10,613,478 records.
3. Records that contain **0 'trip_miles'** values: contains 43,314,719 records.
4. Records that contain **0 'fare'** values: contains 281,684 records.
5. Records that contain **0 'trip_total'** values: contains 275,024 records.
6. Records that contain **unmatched values of timestamp differences with column trip_seconds**: contains 15,486,056 records.

RESULTS

There were **202,388,275 records** that we identified as an outlier, resulting in **1,715,438 records without outliers.** There were 172 unique companies. We decided to trim down the dataset to only records that came from the top 10 companies based on the overall trips from the dataset. We found that **84.43% of the clean dataset came from the top 10 companies.**

No.	Company	Count of Trip
1	Taxi Affiliation Services	665005
2	Dispatch Taxi Affiliation	187470
3	Choice Taxi Association	156089
4	Blue Ribbon Taxi Association	89579
5	Medallion Leasin	86548
6	Northwest Management LLC	67517
7	City Service	52240
8	KOAM Taxi Association	48697
9	Star North Management LLC	48515
10	Globe Taxi	46744

0.71%

data are being processed (1,448,404 total records)

Data Analysis

Data Visualization

Heat Map by Count of Trips

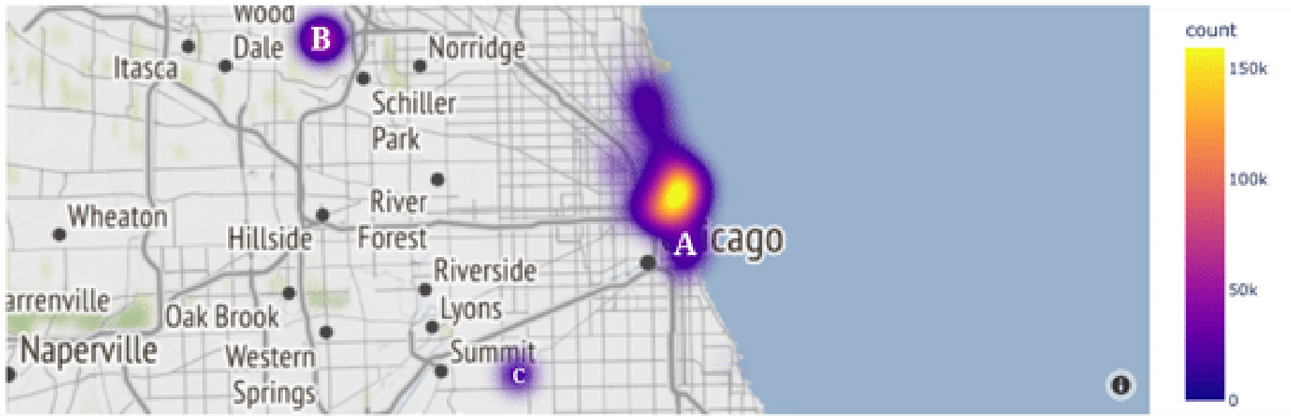


FIGURE D.1. The heat map is the count trip number of top 10 companies based on overall trip from the dataset.

YELLOW AREA

Area with the high number of trips

PURPLE AREA

Area with the low trip number of taxi

Data Analysis

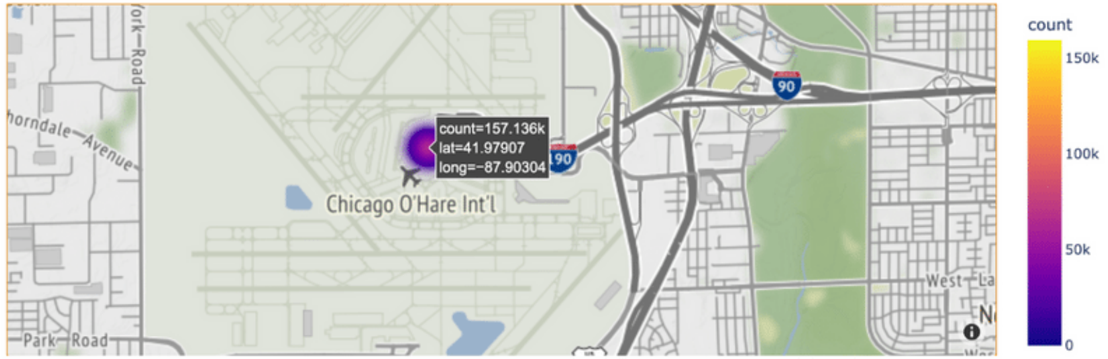


FIGURE D.2. Chicago O'Hare International Airport

The result of zooming the “B” point from Figure D.1. This area has a total of 157.136k trips.

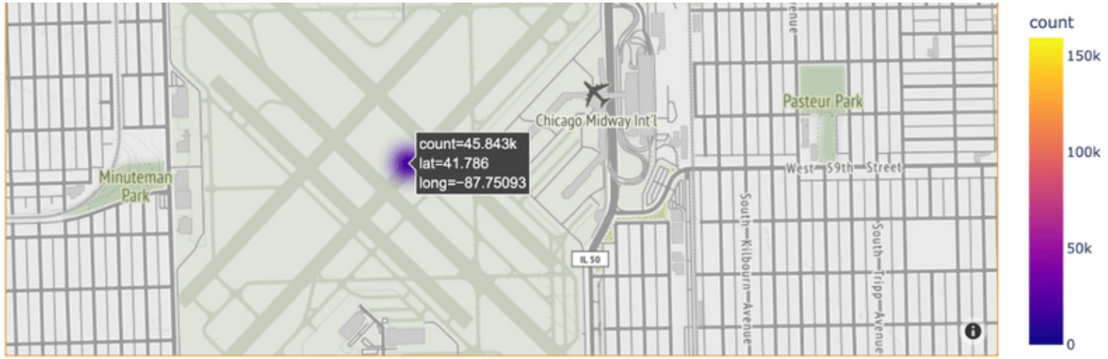


FIGURE D.3. Chicago Midway International Airport

The result of zooming the “C” point from Figure D.1 This area has a total of 45.843k trips.

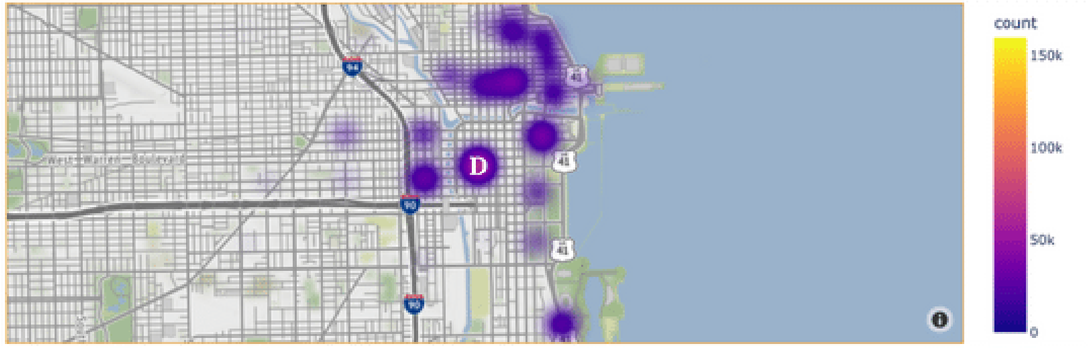


FIGURE D.4. West Warren Boulevard

West Warren Boulevard Area is the area which is the result of zooming the “A” point from Figure D.1

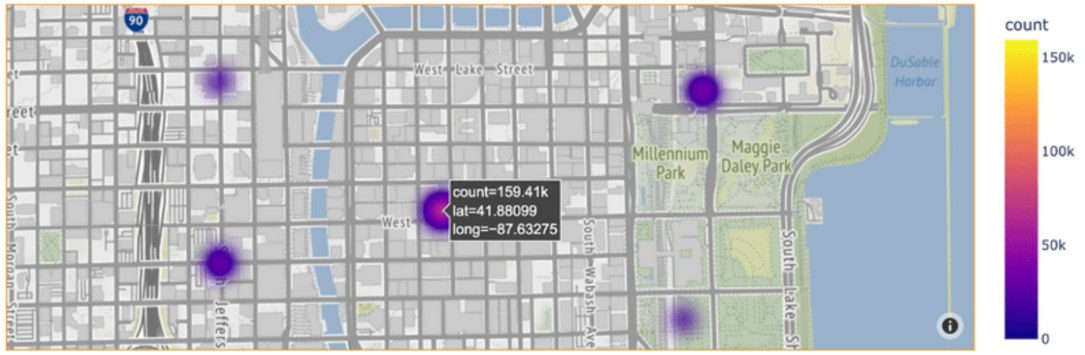


FIGURE D.5. West Monroe Street

West Monroe Street is the area which is the result of zooming the “D” point from Figure D.1.This area has a total of 159.41k trips.

Data Analysis

Modeling

Predictive Analysis

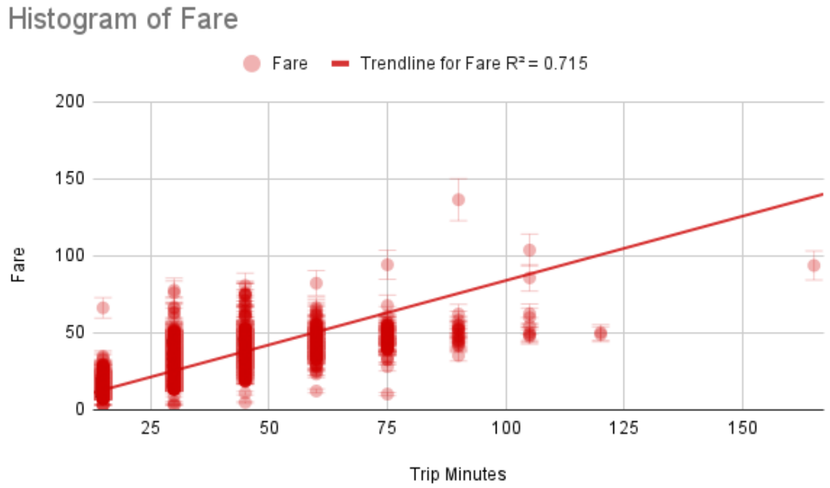
R^2	0.7154
Standard Deviation	6.7539
Coefficient (Model)	5.77 + 0.85x

R^2 Interpretation

71.54% of the variation in the fare can be explained by the number of minutes of trip taken.

Coefficient Interpretation

A fixed cost of 5.77 and an incremental of 0.85 dollars for every minute taken



Single Linear Regression (SLR)

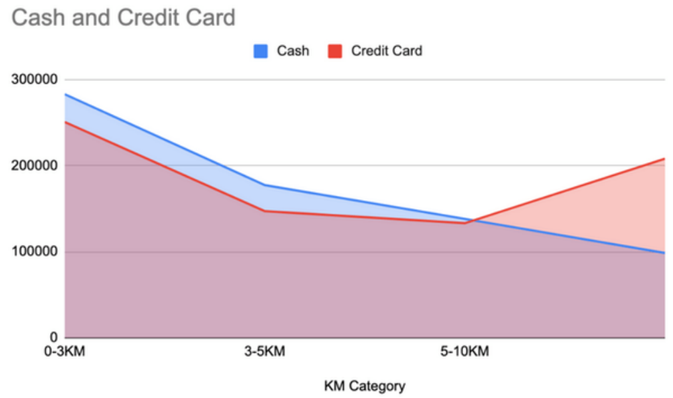
For the predictive analysis, we are using Single Linear Regression (SLR) to predict the fares as the dependent variable with **20,000 random samples** from clean data. The independent variable is average fare per minute.

Data Analysis

Modeling

Prescriptive Analysis

Fare and Payment Method



Based on the graph above, for the short trip distance (0-3 KM), people are likely to use **cash**. In contrast, for the long distance traveler, people are more likely to use their **credit card** as the graph rose sharply.

Average Fare per Minutes each Distance (KM) Range

The highest average fare per minute for distances of 0-10 KM is at **6 AM** with the average fare per minute of $1.39.

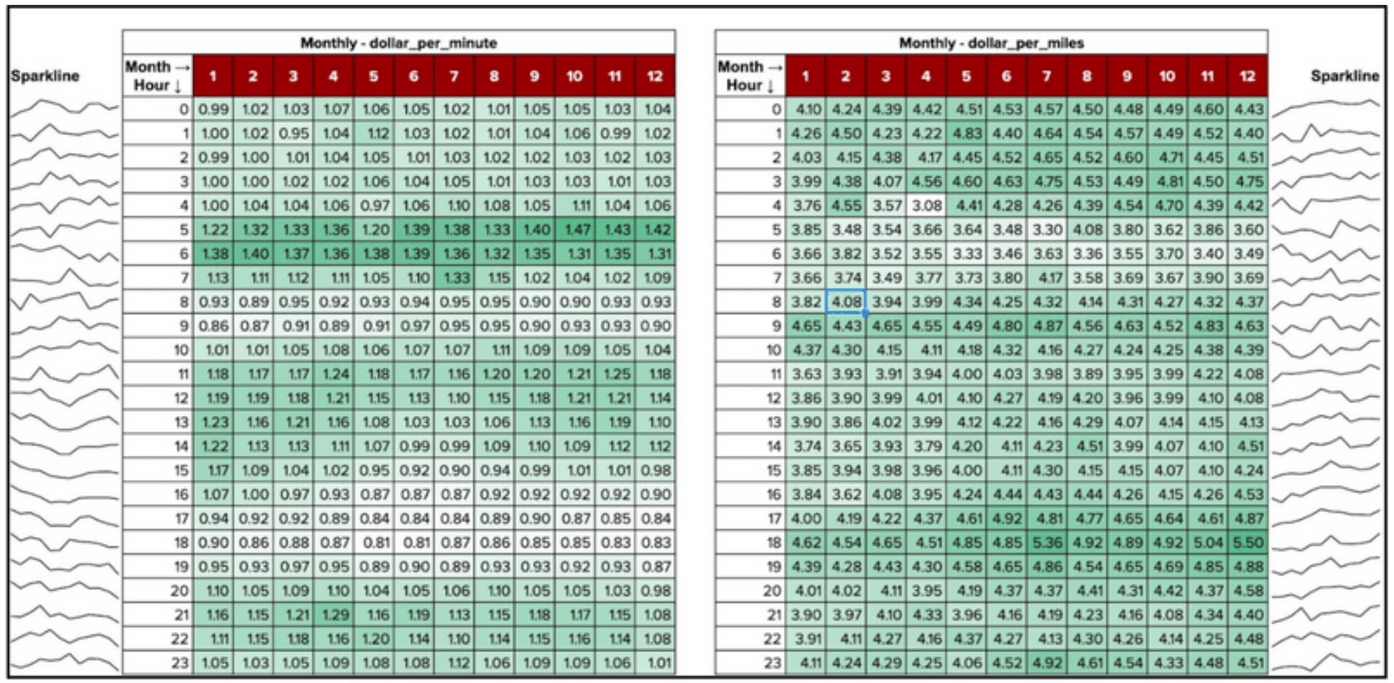
For the distance of >10 KM, the highest average fare per minute is at **22.00** with the average fare per minute of $1.87.

Sparkline	Monthly - dollar_per_minute				
	Range → Hour ↓	0-3KM	3-5KM	5-10KM	>10KM
	0	0.99	0.82	0.99	1.81
	1	1.01	0.81	0.99	1.80
	2	0.99	0.84	1.01	1.85
	3	1.00	0.84	1.01	1.85
	4	0.98	0.81	1.01	1.82
	5	1.20	0.82	1.02	1.84
	6	1.39	0.85	1.01	1.86
	7	1.12	0.85	0.99	1.84
	8	0.87	0.80	0.95	1.75
	9	0.81	0.80	0.91	1.71
	10	0.92	0.79	0.94	1.72
	11	1.06	0.82	0.95	1.77
	12	1.05	0.80	0.94	1.75
	13	1.00	0.80	0.93	1.73
	14	0.99	0.84	0.94	1.78
	15	0.92	0.80	0.92	1.72
	16	0.87	0.81	0.91	1.72
	17	0.82	0.78	0.86	1.65
	18	0.78	0.78	0.82	1.60
	19	0.85	0.78	0.88	1.66
	20	0.98	0.81	0.97	1.78
	21	1.13	0.82	1.01	1.83
	22	1.10	0.85	1.02	1.87
	23	1.04	0.82	1.00	1.82

Data Analysis

Modeling

Prescriptive Analysis



Monthly - dollar_per_minute

Sparkline	Month → Hour ↓	1	2	3	4	5	6	7	8	9	10	11	12
	0	0.99	1.02	1.03	1.07	1.06	1.05	1.02	1.01	1.05	1.05	1.03	1.04
	1	1.00	1.02	0.95	1.04	1.12	1.03	1.02	1.01	1.04	1.06	0.99	1.02
	2	0.99	1.00	1.01	1.04	1.05	1.01	1.03	1.02	1.02	1.03	1.02	1.03
	3	1.00	1.00	1.02	1.02	1.06	1.04	1.05	1.01	1.03	1.03	1.01	1.03
	4	1.00	1.04	1.04	1.06	0.97	1.06	1.10	1.08	1.05	1.11	1.04	1.06
	5	1.22	1.32	1.33	1.36	1.20	1.39	1.38	1.33	1.40	1.47	1.43	1.42
	6	1.38	1.40	1.37	1.36	1.38	1.39	1.36	1.32	1.35	1.31	1.35	1.31
	7	1.13	1.11	1.12	1.11	1.05	1.10	1.33	1.15	1.02	1.04	1.02	1.09
	8	0.93	0.89	0.95	0.92	0.93	0.94	0.95	0.95	0.90	0.90	0.93	0.93
	9	0.86	0.87	0.91	0.89	0.91	0.97	0.95	0.95	0.90	0.93	0.93	0.90
	10	1.01	1.01	1.05	1.08	1.06	1.07	1.07	1.11	1.09	1.09	1.05	1.04
	11	1.18	1.17	1.17	1.24	1.18	1.17	1.16	1.20	1.20	1.21	1.25	1.18
	12	1.19	1.19	1.18	1.21	1.15	1.13	1.10	1.15	1.18	1.21	1.21	1.14
	13	1.23	1.16	1.21	1.16	1.08	1.03	1.03	1.06	1.13	1.16	1.19	1.10
	14	1.22	1.13	1.13	1.11	1.07	0.99	0.99	1.09	1.10	1.09	1.12	1.12
	15	1.17	1.09	1.04	1.02	0.95	0.92	0.90	0.94	0.99	1.01	1.01	0.98
	16	1.07	1.00	0.97	0.93	0.87	0.87	0.87	0.92	0.92	0.92	0.92	0.90
	17	0.94	0.92	0.92	0.89	0.84	0.84	0.84	0.89	0.90	0.87	0.85	0.84
	18	0.90	0.86	0.88	0.87	0.81	0.81	0.87	0.86	0.85	0.85	0.83	0.83
	19	0.95	0.93	0.97	0.95	0.89	0.90	0.89	0.93	0.93	0.92	0.93	0.87
	20	1.10	1.05	1.09	1.10	1.04	1.05	1.06	1.10	1.05	1.05	1.03	0.98
	21	1.16	1.15	1.21	1.29	1.16	1.19	1.13	1.15	1.18	1.17	1.15	1.08
	22	1.11	1.15	1.18	1.16	1.20	1.14	1.10	1.14	1.15	1.16	1.14	1.08
	23	1.05	1.03	1.05	1.09	1.08	1.08	1.12	1.06	1.09	1.09	1.06	1.01

Monthly - dollar_per_miles

Month → Hour ↓	1	2	3	4	5	6	7	8	9	10	11	12	Sparkline
0	4.10	4.24	4.39	4.42	4.51	4.53	4.57	4.50	4.48	4.49	4.60	4.43	
1	4.26	4.50	4.23	4.22	4.83	4.40	4.64	4.54	4.57	4.49	4.52	4.40	
2	4.03	4.15	4.38	4.17	4.45	4.52	4.65	4.52	4.60	4.71	4.45	4.51	
3	3.99	4.38	4.07	4.56	4.60	4.63	4.75	4.53	4.49	4.81	4.50	4.75	
4	3.76	4.55	3.57	3.08	4.41	4.28	4.26	4.39	4.54	4.70	4.39	4.42	
5	3.85	3.48	3.54	3.66	3.64	3.48	3.30	4.08	3.80	3.62	3.86	3.60	
6	3.66	3.82	3.52	3.55	3.33	3.46	3.63	3.36	3.55	3.70	3.40	3.49	
7	3.66	3.74	3.49	3.77	3.73	3.80	4.17	3.58	3.69	3.67	3.90	3.69	
8	3.82	4.08	3.94	3.99	4.34	4.25	4.32	4.14	4.31	4.27	4.32	4.37	
9	4.65	4.43	4.65	4.55	4.49	4.80	4.87	4.56	4.63	4.52	4.83	4.63	
10	4.37	4.30	4.15	4.11	4.18	4.32	4.16	4.27	4.24	4.25	4.38	4.39	
11	3.63	3.93	3.91	3.94	4.00	4.03	3.98	3.89	3.95	3.99	4.22	4.08	
12	3.86	3.90	3.99	4.01	4.10	4.27	4.19	4.20	3.96	3.99	4.10	4.08	
13	3.90	3.86	4.02	3.99	4.12	4.22	4.16	4.29	4.07	4.14	4.15	4.13	
14	3.74	3.65	3.93	3.79	4.20	4.11	4.23	4.51	3.99	4.07	4.10	4.51	
15	3.85	3.94	3.98	3.96	4.00	4.11	4.30	4.15	4.15	4.07	4.10	4.24	
16	3.84	3.62	4.08	3.95	4.24	4.44	4.43	4.44	4.26	4.15	4.26	4.53	
17	4.00	4.19	4.22	4.37	4.61	4.92	4.81	4.77	4.65	4.64	4.61	4.87	
18	4.62	4.54	4.65	4.51	4.85	4.85	5.36	4.92	4.89	4.92	5.04	5.50	
19	4.39	4.28	4.43	4.30	4.58	4.65	4.86	4.54	4.65	4.69	4.85	4.88	
20	4.01	4.02	4.11	3.95	4.19	4.37	4.37	4.41	4.31	4.42	4.37	4.58	
21	3.90	3.97	4.10	4.33	3.96	4.16	4.19	4.23	4.16	4.08	4.34	4.40	
22	3.91	4.11	4.27	4.16	4.37	4.27	4.13	4.30	4.26	4.14	4.25	4.48	
23	4.11	4.24	4.29	4.25	4.06	4.52	4.92	4.61	4.54	4.33	4.48	4.51	

Monthly Average Fare per Minutes each Hour

The high average fare per minute lay between 5-6 AM with the highest of $1.40 in February. The low average fare per minute lay between 17.00-18.00 especially in the last three quarters of the year

Monthly Average Fare per Miles each Hour

The high average fare per miles lay between **18.00-19.00** with the highest of $5.36 in July. In contrast, the lightest color on the table is around **6 AM** for every month.

Correlation to Research Objective

Pinpoint what aspects that ARRO could improve in company application that can strengthen their positioning within the Chicago transportation market.

Bias and Outlier Analysis

There was 99.29% data from 204,103,713 records that we identified as an outlier, resulting in 1,715,438 records without outliers. Data with a null value and zero value (for the variable that should be filled by a number such as trip_seconds, trip_times, trip_total, etc.) need to be removed to result in valid trip data.

Method Analysis

Predictive Analysis

The research use of Single Linear Regression (SLR) Predictive Model with 20,000 random samples from clean data to predict the taxi's fare rate rather for every trip mile or trip minutes.

Prescriptive Analysis

The prescriptive analysis in this research will be implemented to give ARRO recommendations about the suitable payment method for certain trip distances, determine user and fare rate behavior in form of visualization, and give recommendations according to it.

Managerial Implications & Recommendations

Arro

Operation Recommendations

1. Arro provide a newer and more convenient method of digital paymeny (venmo, apple pay, and paypal)

2. New driver-side application feature that detect crowded area to be accomodate with

3. Enhance their pricing algorithm which highly impacts the pricing strategy in the marketing functions

Arro

Human Capital Recommendations

1. Increase Arro capacity and convenience by continuous recruitment and training

2. Recruitment process align with the rise of taxi trip fare to create a cross-subsidization

Managerial Implications & Recommendations

Arro

Marketing Recommendations

1. Advertisements in highly concentrated area shown in the heat map

2. Competitive pricing strategy based on the average fare per minute in conventional taxi

Conclusion

1. Single Linear Regression model to predict the fare by a fixed cost of 5.77 and an incremental of 0.85 dollars for every minute taken.
2. For **operation** areas the suggestions are to accept widely used digital payment, New driver-side application feature that detect crowded area to be accommodate with, and enhance their pricing algorithm by using the average price per minute in a distance traveled
3. For **human capital** areas the suggestion is to increase their capacity and convenience as a business by recruiting new driver partners
4. For **marketing** areas the suggestions are to focus the advertisement on the customer concentration area, create a competitive pricing strategy based on the average fare per minute in conventional taxi

Recommendation for Further Study

Examine in-depth analysis using the trip **data of the online ride-hailing platform** in Chicago, which are Lyft and Uber.

MB4011 Business Intelligence and Analytics

THANK YOU

STUDY CASE OF ARRO.INC

Hana Aqila	19020143
Samuel Mauliate	19020174
Loudiva Pramudya	19020209
Fauzan Khalil	19020217
Cliff Richardo	19020293