# Introduction

## Statement of the goal

In this report we analyze the Taxi Trips in Chicago 2020 dataset. The aim of the project is to define what variables may influence the revenue of the taxi cabs. We are specifically interested to understand how the cost of trip varies depending on the community area and which areas generate the highest revenue. We then study how the cost of the trip changes through time variables (date, month, weekday, hour).

## Brief summary of the approach

In this report we first present an exploratory data analysis which allows to familiarize with the data: study the distribution of continuous variables, the proportion of levels in categorical variables and the relation between those. We first present descriptive statistics analysis with respect to the community areas. We then present various interactive maps of Chicago which allows to see the change of the total price on average according to the time variables. Finally, we present estimates of revenue for taxi cabs per hour using bootstrap methods.

## Introduction to the problem

The analysis of similar datasets has been performed before but mostly with respect to the duration of the trip. Our analysis focuses mostly on the area (pickup and drop-off) as we find evidence that it's has an effect on the total cost of the trip. We think our analysis may be relevant for taxi companies to facilitate the task of assignment of taxi cabs around the city of Chicago.

What follows in this report is a commentary of the insights that we obtained from studying the data. Afterwards, we present the main tools used for our statistical analysis and discussions on the results obtained through their implementation. A conclusion at the end summarizes the most important take-out points.

# Exploratory Data Analysis

For the Exploratory data analysis we subsample 10 000 random rows to decrease computation time.

## Information about the continuous variable trip.total

The variable of interest in our analysis is called Trip.Total as it represents the total cost of the trip (which is the sum of other continuous variables available in our dataset: fare, tips, tolls, extras).

Chart

Description automatically generatedChart

Description automatically generated

By displaying the distribution of trip.total as a histogram and as a boxplot, we notice that the data is skewed to the right. On the histogram, we see that most of the data is situated on the left, meaning that the majority of the trips costs between 0 and 200$. As the price increase - we have fewer and fewer data points. Nevertheless, we notice that there are few extreme outliers with values close to 600$.

To respond to the skewness toward large values, let's use logarithmic scales in the following charts and graphs. By taking the logarithm of trip.total the data looks more evenly distributed.

Chart, histogram

Description automatically generatedChart, box and whisker chart

Description automatically generated

## Visualization of Categorical Variables - Frequency of Areas

As we are interested in the difference of Total cost of the trip per area, let's visualize the frequency of each area in the dataset. The following plots represent the most common areas for pick up and drop off of our dataset. The areas "Near North Side", "The Loop" and "Near West Side" account for more than 60% of the dataset.

Chart, pie chart

Description automatically generatedChart, pie chart

Description automatically generated

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

## Distribution of log.trip.total according to the areas

The following plots present the distribution of logarithm of trip.total (as the data looks more normalized and allows to see the difference in distributions by area) according to the pickup and drop-off areas. We notice a difference in total cost in the most common areas: for example, Garfield Ridge and O'Hare seem to have higher cost for pick up, for drop-off Morgan Park seems to have the highest cost.

Below, we also include the distributions by area of trip.total without taking the logarithm.

Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

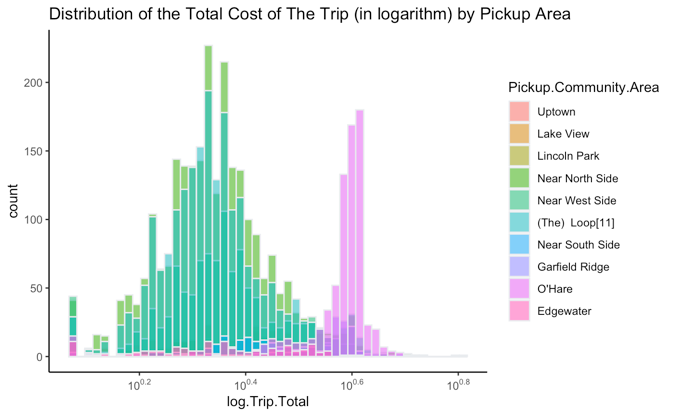
Description automatically generated

Chart, scatter chart

Description automatically generatedChart, waterfall chart

Description automatically generated

By exploring the histogram of log.Trip.Total according to the area, we notice that the distribution may be bimodal as we see 2 peaks. Again, for pickup areas O'Hare and Garfield Ridge see to have higher mean than the rest of the areas.

Chart, histogram

Description automatically generated

Chart

Description automatically generatedChart

Description automatically generated

## QQ plots

To explore the behaviour of the trip.total we plot QQ plots. As expected, trip.total doesn't follow a normal distribution. The logarithm of trip.total seems to have a better fit but still has multiple outliers that deviate from the theoretical quantiles.

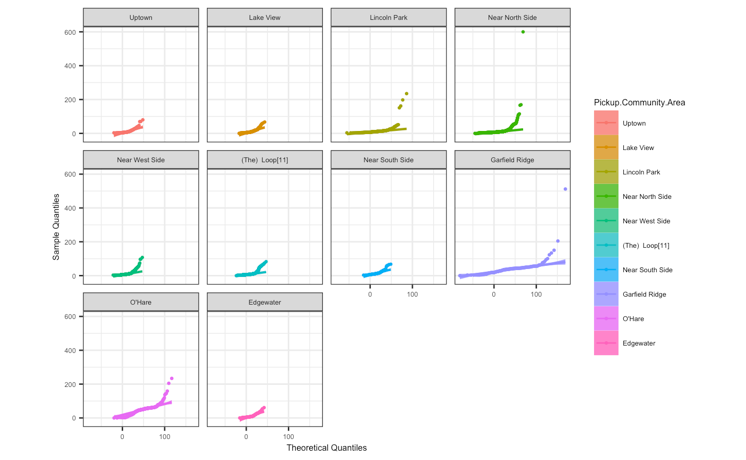
When we study the distribution of log.trip.total according to the most common areas, the distribution of log.trip.total looks better.

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

# Time series

As we have data for the year 2020, it makes sense to study the change in trip.toal (in logarithmic scale) by month. The following plots represent the change of price according to the area (pickup, dropoff) by month. For example, we see that price for pick up seems to be higher for all the month in Garfield Ridge and O'Hare. We also notice that almost all areas follow a similar trend in the time variable (for example, the cost of trips dropped in all areas in April).

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

We can also see how total cost of the trip varies according to the weekday.

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Here, we can observe the same graphs not in the logarithmic scale.

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated