1. Goal of the mini-project

What are the zones that generate the most revenue? What are the reasons?

1. Brief summary of the approach
2. Brief introduction to the problem
3. Descriptive analysis of the data
4. Presentation of results

First of all, we do the basic analysis on the dataset and look at each variable. For the continues variables, they were visualised using ggpair function in GGally package, a powerful tool that gives scatterplot for each pair of the variables and a histogram for each single variable, as well as the correlation between each pair of variables. For instance, we expect to see a correlation between trip\_total and trip\_miles. In the scatter plot, we can spot an almost linear line and also the outliners. In the correlation cell, we get a correlation coefficient of 0.467, which indicates a moderate positive association.

Calendar

Description automatically generated

To further visualise the variable trip\_total, we use a geom\_histogram in ggplot2 and carefully select the binwidth.

Chart, histogram

Description automatically generated

With regards to discrete variables, we have 77 zones in Chicago and it is too much to plot out. However, since we are interested in the zones the generate most revenue, we plot out the top 10 zones. To visualise the distribution of the data as well as give a comprehensive comparison among zones, we use boxplots in parallel. From the plots below, it is clear that zone 76 and zone 56 generate the most revenue, whereas they are also the two zones where the average avenue reaches the highest, hence they might be where drivers like most.

Chart, box and whisker chart

Description automatically generated

As for the dropoff area, it is obvious that zone 76 produced most revenue.

Chart, box and whisker chart

Description automatically generated

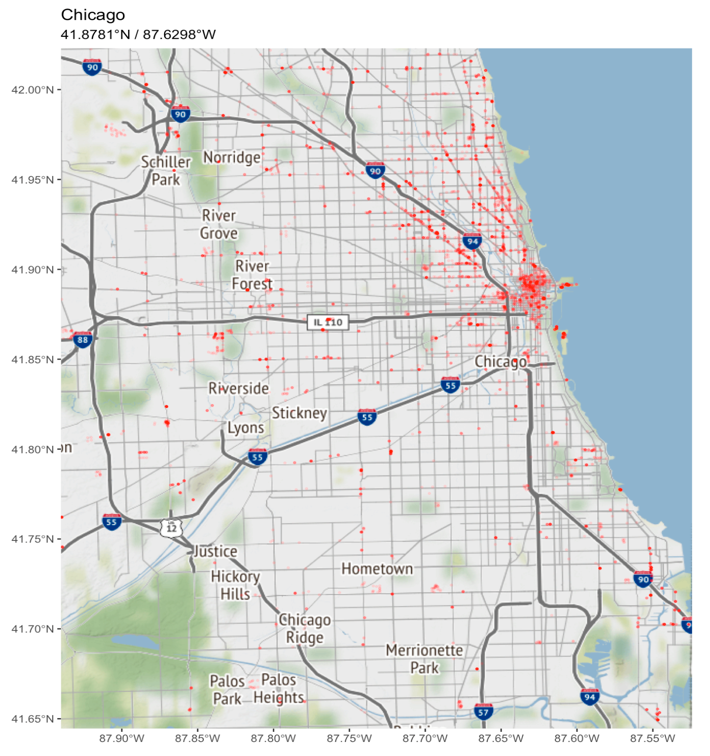
However, we prefer using pickup\_community\_area over dropoff\_comunity\_area to answer our question since we see the pickup zone as the location where new revenue are generated and are of driver’s interest.

After looking at the revenue map, we wonder why some of places generate more revenue than others? To further explore the reason, we introduce outside dataset about the amenities as well as the population data from Wikipedia.

Map

Description automatically generated by Olav

We extract amenity information including theatre, restaurant and hospital from Open Street Map Data (osmdata) package and plot it out using the same package. Due to the limit of the dataset, we cannot get all the places where people are likely to take a taxi. But we can still identify the places with most clustered red dots are popular area for taking taxi, also they might be in the centre of Chicago.



To make it easier to compare, we plot out all roads for cars and add the zone border into the map, as well as the amenities on top of them.

Map

Description automatically generated

Now, it becomes clearer that the reason for higher revenue in certain zone might be related to the density of certain amenities.

The only exception is O’Hare, where not many amenities are gathered but a hot spot for taxis. It turned out to be that O’Hare is the location of Chicago O'Hare International Airport, which explains why it attracts most of the taxis and produces such a high revenue.

Moreover, we think population might be another main factor that affects the total avenue of each zone. Hence we extract the population data from Wikipedia page and join it into our original dataset.

Map

Description automatically generated (By modifying Olav’s code)

We apply logarithmic function on population to make the map more sensible. After comparing the two maps, we think the population might be one of the exploratory variables to the total revenue.