Catalog

1. Introduction

In this assignment, I will try to complete these tasks: data preprocessing, data analysis, visualization, and clustering with an engineering coding style, which may facilitate future improvements and enhancements.

When dealing with spatial-temporal data, it's necessary for us to consider many factors, such as the scale of data to determine which tools and packages to be used; the coordinate system of the data to ensure data consistency; mapping standards for accurate spatial perception. Hence, I would like to first introduce the architecture of my overall workflow to clarify how i manage and process the data.

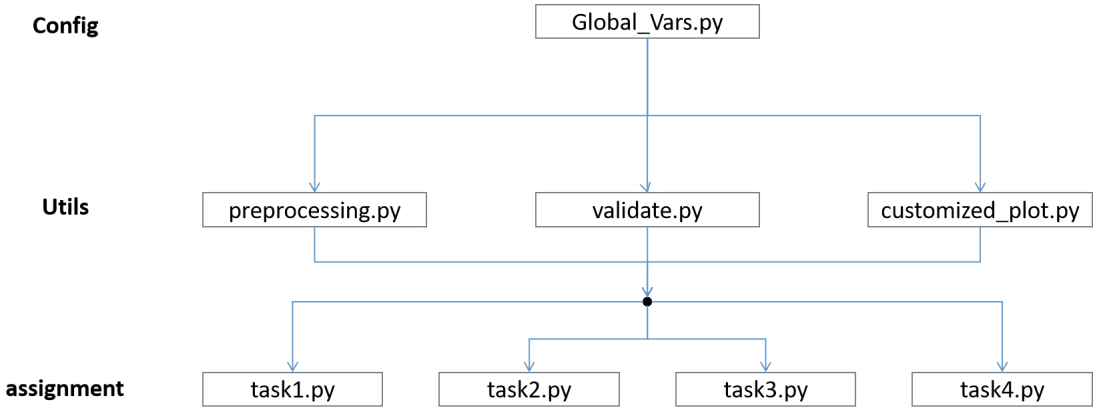

Fig 1. Architecture of this assignment

A three-level code structure(Config, Utils, assignment) is shown in Fig 1. Each section of the code corresponds to the content of different logical parts of managing the project, as Table 1 shown.

Table 1. Content of scirpts.

Logical Level	Script Name	Function
Config	Global_Vars.py	Store global static variables for the project, such as paths and projection coordinates.
Utils	Preprocessing.py	Filtering invalid data, coordinate transformation, and basic geographical calculations.
	Validate.py	Revalidating the spatiotemporal validity of the data before conducting further data analysis.
	Customized_plot.py	Customized plotting the data using different types of chart, for visualization and spatial perception.
Assignment	Task1-4.py	Completing the assignment task.

With the help of these scripts, we can efficiently complete the tasks in this assignment.

2. Task 1: Data Pre-processing

2.1 Data Cleaning for chicago_data.csv and station.csv

Corresponding scripts: (**utiils.preprocessing.py** and **0_task1.py**).

2.2 Answering Task 1

The answers of the questions in taks1 are shown in Table 2.

Table 2. Answers to Task 1.

Question	Answer
How many valid bicycle trips were documented on 25 July 2019?	**20187**
How many bike stations were used on that day?	**535**
How many unique bikes were used?	**3822**

3. Task 2: Coordinate Conversion and Dataset Analysis

3.1 Coordinate Conversion

Corresponding scripts: (**utiils.preprocessing.py** and **1_task2.py**).

3.2 Statistics on Trip Duration and Trip Distance

The answers of the questions in taks2 are shown in Table 3.

Table 3. Answers to Task 2.

Indicator	Trip duration	Trip distance
Max Value	31243.00	20083.35
Min Value	61.00	0.00 (departure s. == arrival s.)
Median	799.00	1753.74
Mean	1187.64	2352.77
25% percentile	465.00	1016.56
75% percentile	1403.00	3096.99
Standard deviation	1400.98	1972.31

4. Task 3: Data Visualization

Corresponding scripts: (**utiils.customized_plot.py** and **2_task3.py**).

4.1 Number of Departure Trips over 24 Hours

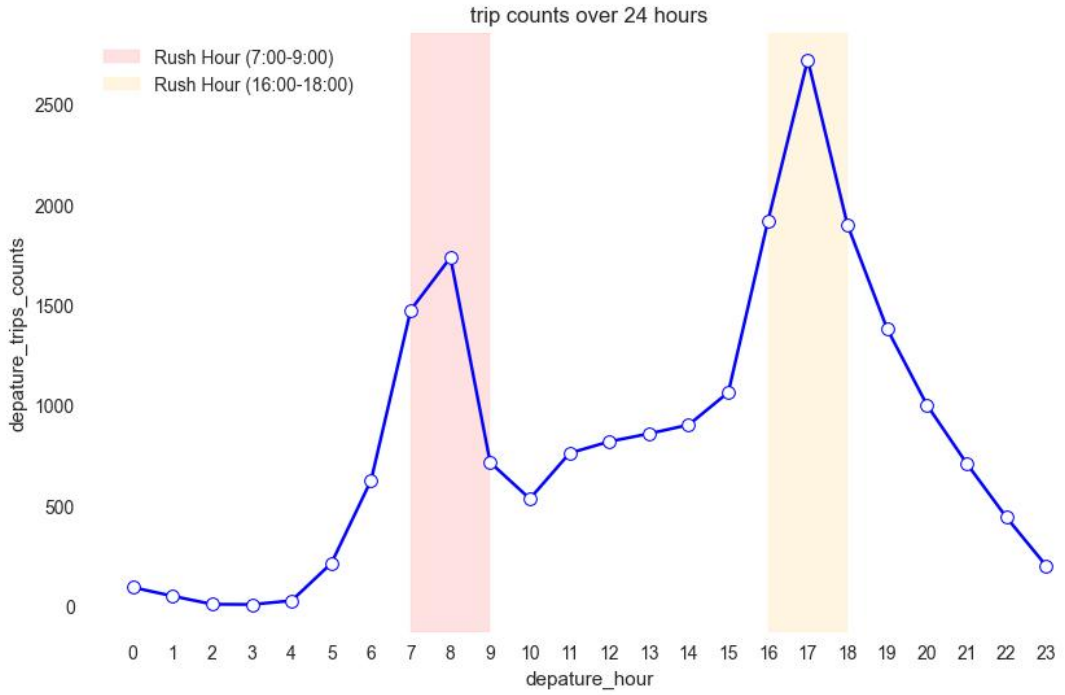

Fig 2. Trip counts by hour

The number of departure trips over 24hours exhibits a clear pattern and rhythm based on the plot of the data (Fig. 2). Prior to 6 AM, the trip count remains relatively low, typically below 500 trips. However, a steep increase occurred from about 6 AM, peeking at about 1700 counts at around 8 AM. Subsequently, from 9 AM onwards, the trip count decreases slightly, stabilizing at around 800 trips, and this level is maintained until about 3 PM. Then, from 3 PM onwards, the trip count experiences another significant surge, reaching its daily peak at around 5 PM, with 2700 trips. After this peak, a gradual decline followed throughout the evening, eventually reaching around 200 trips in the end of the day.

From this distribution of trip numbers, it can be inferred that the trip count for these Divvy shared bikes is greatly related to the commuting intensity on workdays. The morning and afternoon peaks correspond to the rush hour when people commute to and from work, while during the non-working hours(eg. Late night), the trip count remains relatively low.

So my conclusion is that there is a strong rhythm in the flow of the number of trips over 24 hours a day.

4.2 Distribution of Trips at Different Stations

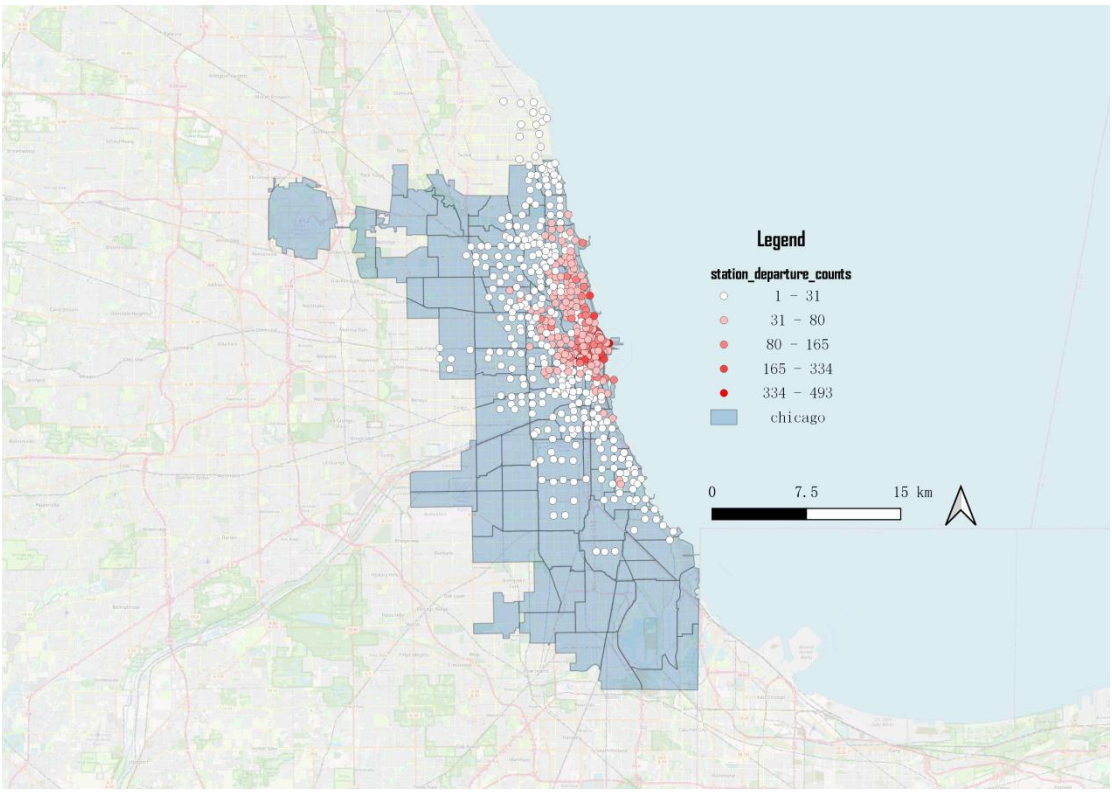

Fig 3. Spatial distribution of departure trips aggregated by stations.

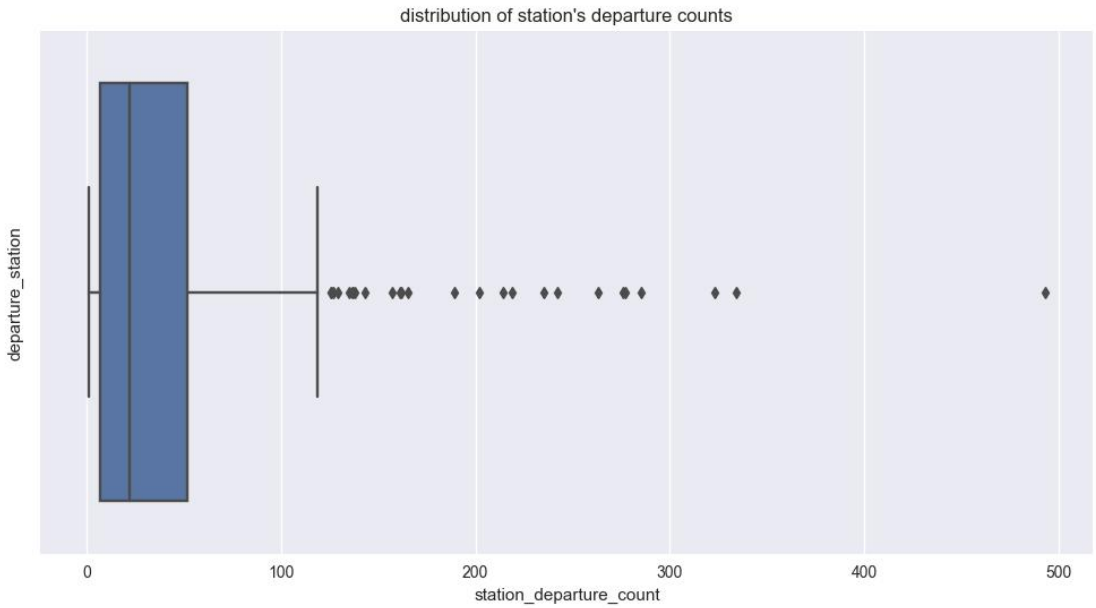

Fig 4. Numeric distribution of departure trips aggregated by stations.

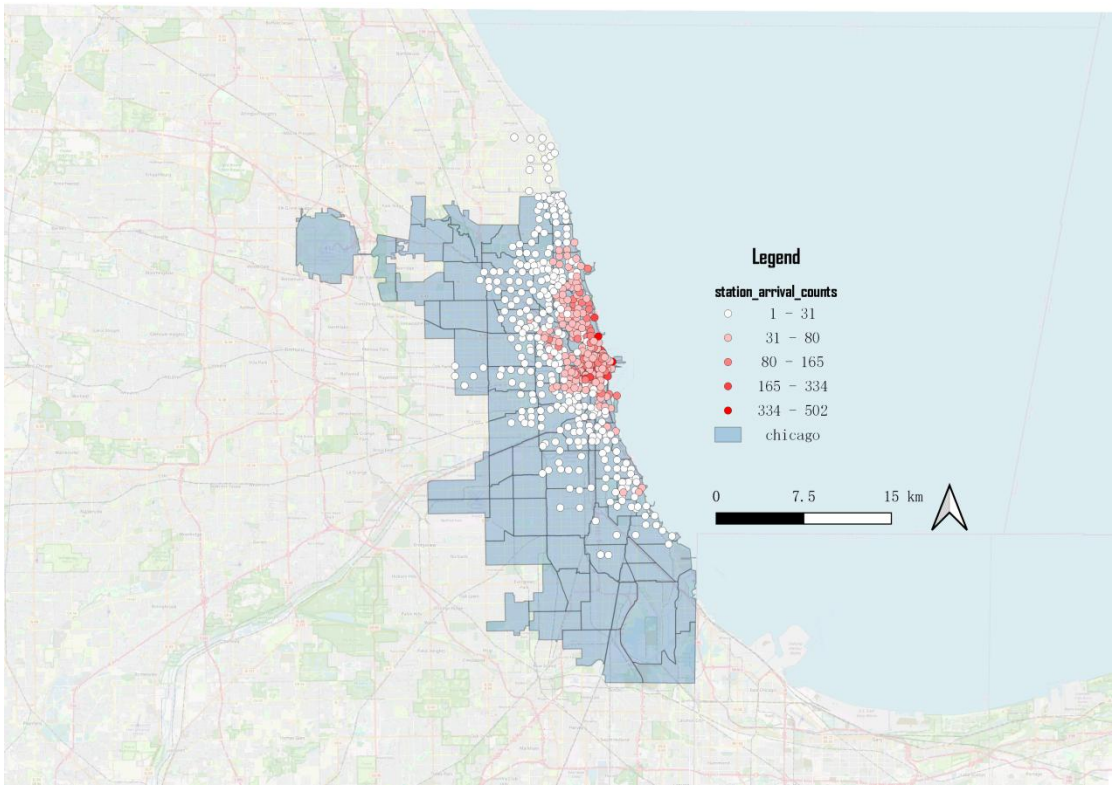

Fig 5. Spatial distribution of arrival trips aggregated by stations

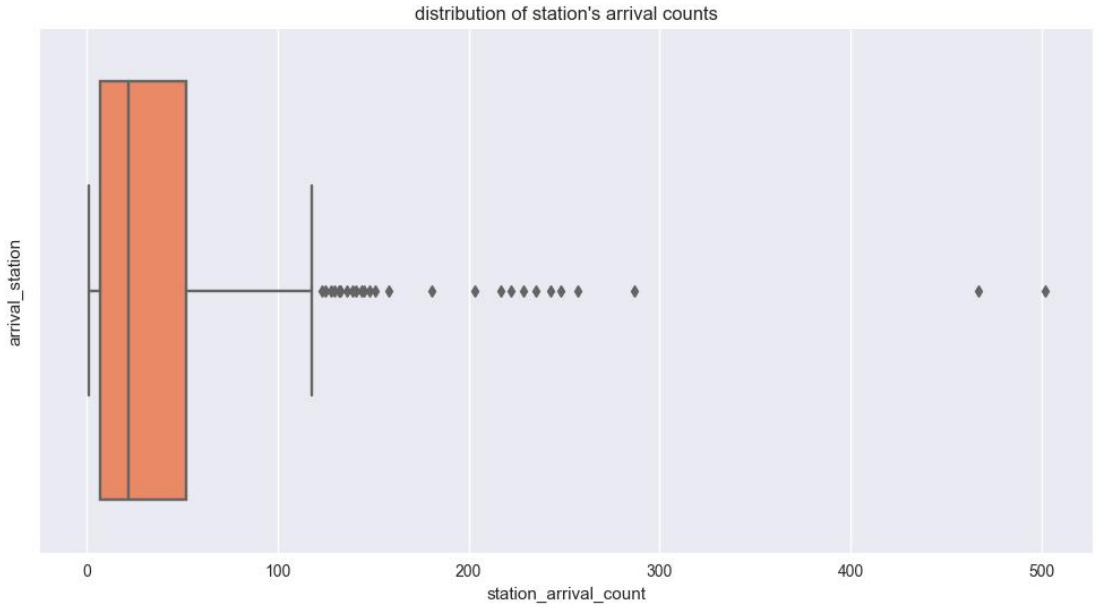

Fig 6. Numeric distribution of arrival trips aggregated by stations

The distribution of departure and arrival trips at different stations were analyzed and two sets of visualizations have been created to illustrate this:

From Spatial Perspective: The spatial distribution can be clearly observed in Fig. 3 and Fig 5. Station located in downtown Chicago consistently exhibited significantly higher trip counts, whether it was for departures or arrivals. In contrast, station outside the city center typically had trip counts ranging from 1 to 31. The downtown area stations consistently recorded counts between 31 and 165, with a few more central stations even reaching counts between 334 and 493. This indicates a clear centralization effect, with the majority of trips occurring in the downtown area and often used for short-distance travel.

From Numeric Perspective: The box-plots for both departure and arrival trip counts aggregated by stations displayed similar distributions. The 25th percentile and 75th percentile were both approximately centered around 20 trips and 50 trips, respectively, with the median around 30 trips. The maximum trip count of the box-plot(75th percentile + 1.5*IQR) was approximately 120,but numerous outliers with counts exceeding the maximum value were observed. The highest outlier values reached up to 500 trips in both departures and arrivals stations. This suggests a substantial variation in station usage intensity, with the majority experiencing low usage in non-central area.

In summary, the visualizations reveal that a great variation in stations usage, which is highly related to where their spatial location.

4.3 Distribution of Trip Distance

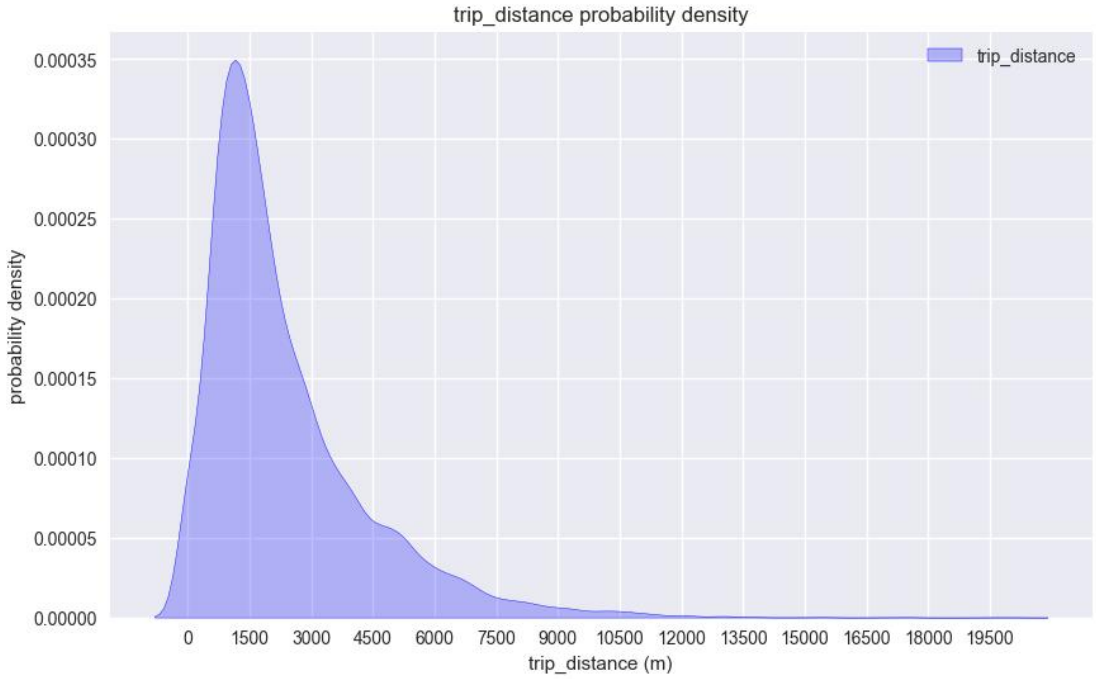

Fig 7. Probability density curve of trip distance

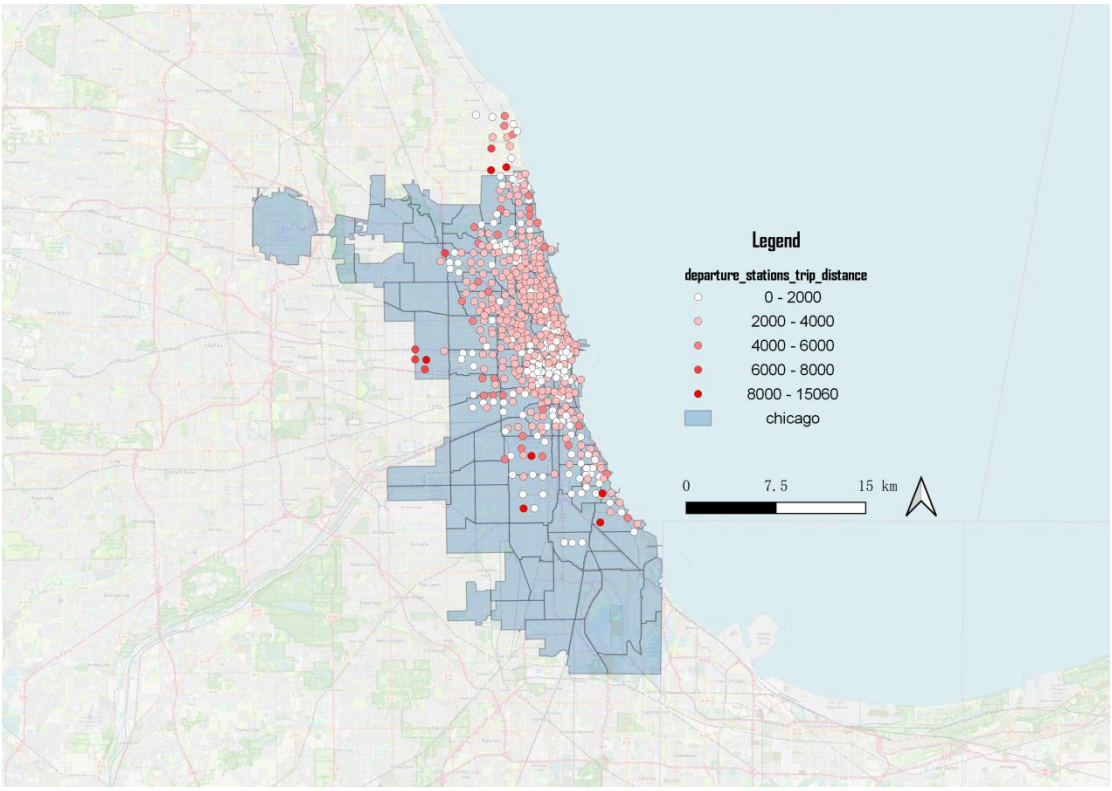

Fig 8. average trip distance aggregated by departure stations

Regarding the distribution of trip distance (measured as straight line Euclidean distance), I have analyzed and drawn the following conclusion:

From Probability Density Distribution Perspective: The probability density distribution (Fig. 7) of trip distance exhibits a long-tail distribution. The majority of trips have distances within 4500 meters, and a relatively high percentage of people ride for about 1500 meters in a trip. However, there are a few exceptional cases where trip distances approach nearly 20,000 meters. These outliers indicate that while most trips are short-distance journeys, there are occasional significantly longer trips.

From Geospatial Perspective: From the geospatial view(Fig. 8), we aggregated all the trip distance by average to their departure stations, it is evident that stations located in the city center tend to have lower trip distance in average. Trip distances in the city center predominantly fall within the 0-2000 meter range, while stations slightly further from the city center have trip distance in the 2-4km range. In contrast, stations located on the outskirts of Chicago usually experience trip distances of 6km or more. This suggests that suburban users are more likely to engage in long-distance rides, while downtown users typically opt for short distance trips, often within a 2 kilometer radius. We have also create a map that focus on the average trip distance aggregated by arrival stations, it shows a similar distribution to Fig. 8.

In summary, the distribution of trip distances indicates that the majority of trips are relatively short in the whole city, though there area occasional longer journey which is outliers. Additionally, the geospatial mapping indicates that those who rides in city center are tend to ride a short-distance trip while those who live in suburban have a relatively higher probability to ride a long-distance trip.

However, the average value we used is easily affected by outliers and causes large deviations. When we conduct in-depth research on more specific topics, **we should fully consider the impact of outliers on the statistical information of the overall sample**.

4.4 Distribution of Trip Duration

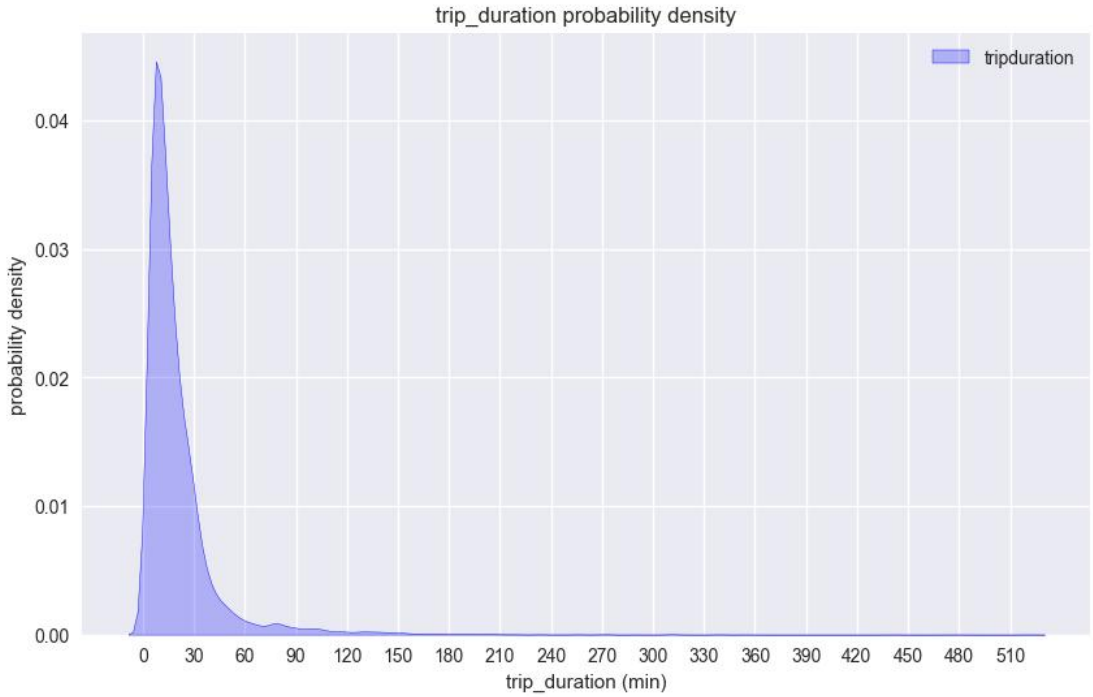

Fig 9. Probability density of trip duration

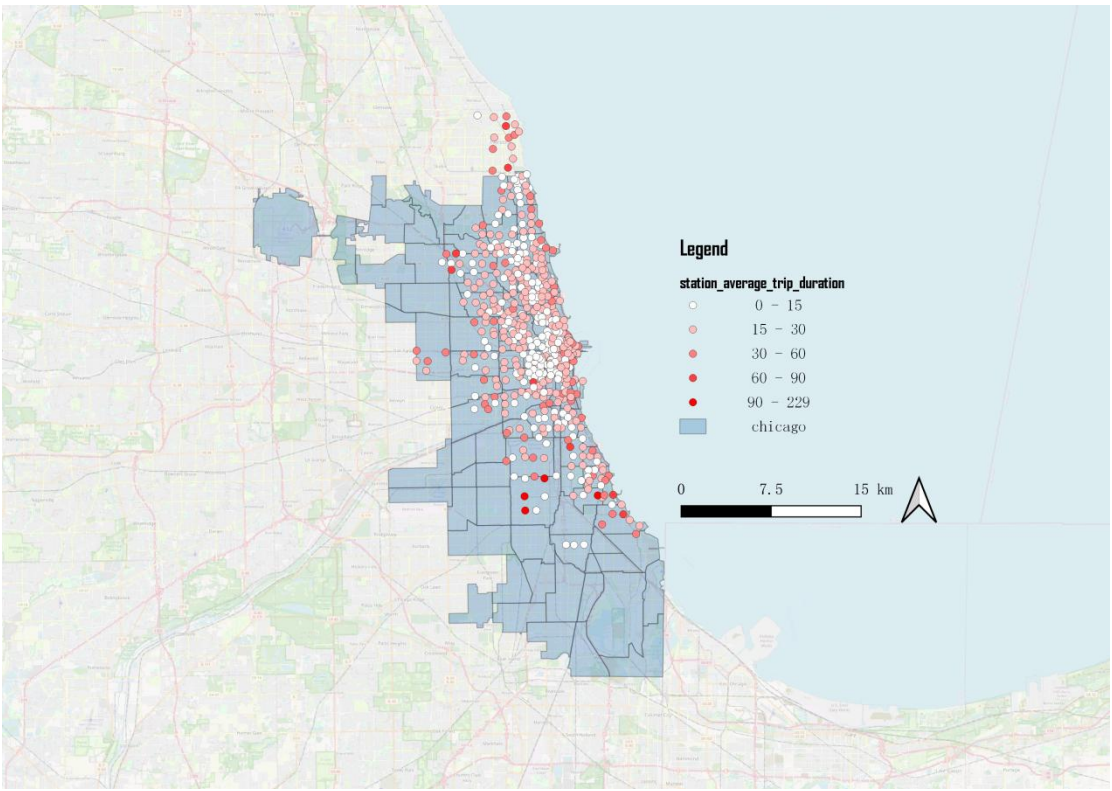

Fig 10. average trip duration for departure stations

Trip duration exhibits a probability density distribution(Fig. 9) shows a similar distribution to the distribution of departure distance(Fig. 7). What's more, more than 75% of trips have duration within 30minutes, and over 90% of trips are completed within 60 minutes. Similarly, there are outliers in trip duration, as evident from the figure 9., which even reaches over 500 minutes in one trip.

Generally, longer trip distances tend to require more time to complete, so there is **a strong correlation between trip duration and trip distance**. As observed in the graph(Fig. 10), many departure stations located in the city center have trip duration predominantly falling with in the 0-30 minute range, with very few trips extending beyond 60minutes. In contrast, stations located on the outskirts of the city tend to have a more common occurrence of trips exceeding 60 minutes, **strongly recalling the task4.3 (Fig. 8)**.

In conclusion, **the average trip duration of station-based is closely related to the spatial location of the station**. The average travel time of stations located in the city center is relatively short, while the average travel time of stations located on the outskirts of the city is longer.

5. Task 4: Bike Station Clustering and Visualization

5.1 Density-based Spatial Clustering using DBSCAN

Corresponding scripts: (**3_task4.py**).

5.2 Number of Clusters and Station ID

The number of clusters is : 19

(Notice: The requirement in Assignment is : **the number of samples in a neighborhood for a point to be considered as a core point is 2 stations.** In my clustering, **I excluded the current point itself**. In other words, the rule to determine whether the current point is a core is that, its neighborhood, at least, **existing 3 points including itself,** and 2 points excluding itself.)

Due to the large number of stations,the report of station ids in each cluster is in **cluster_groups.csv in the same folder.**

5.3 [Bonus] Visualization of Clusters

Packages used: matplotlib & geopandas.

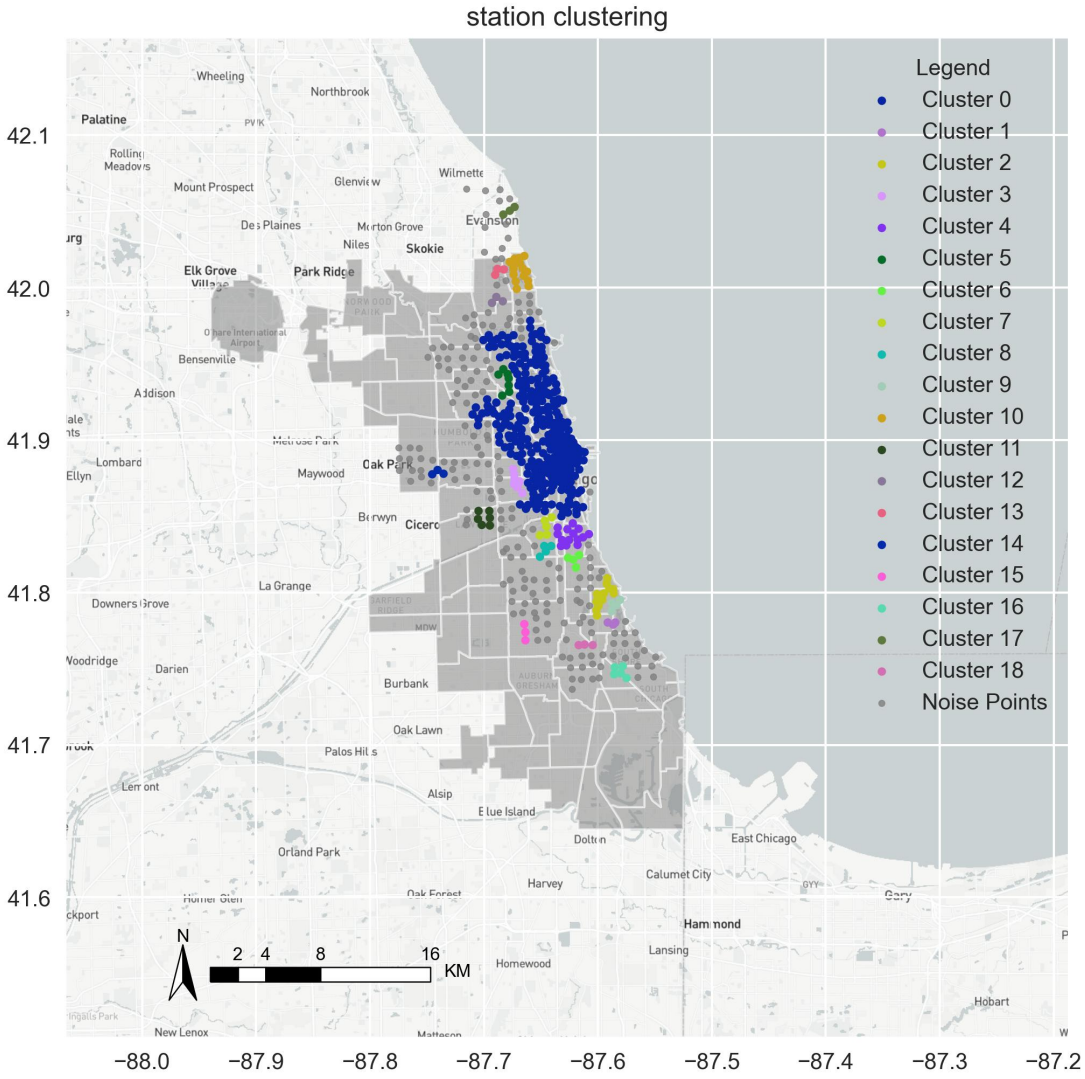

Fig 11. spatial distribution of stations clustering