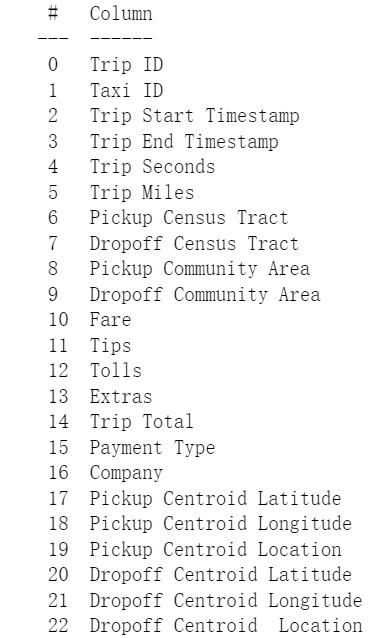
Chicago taxi trips EDA

Yingyi Luo yluo52@uic.edu

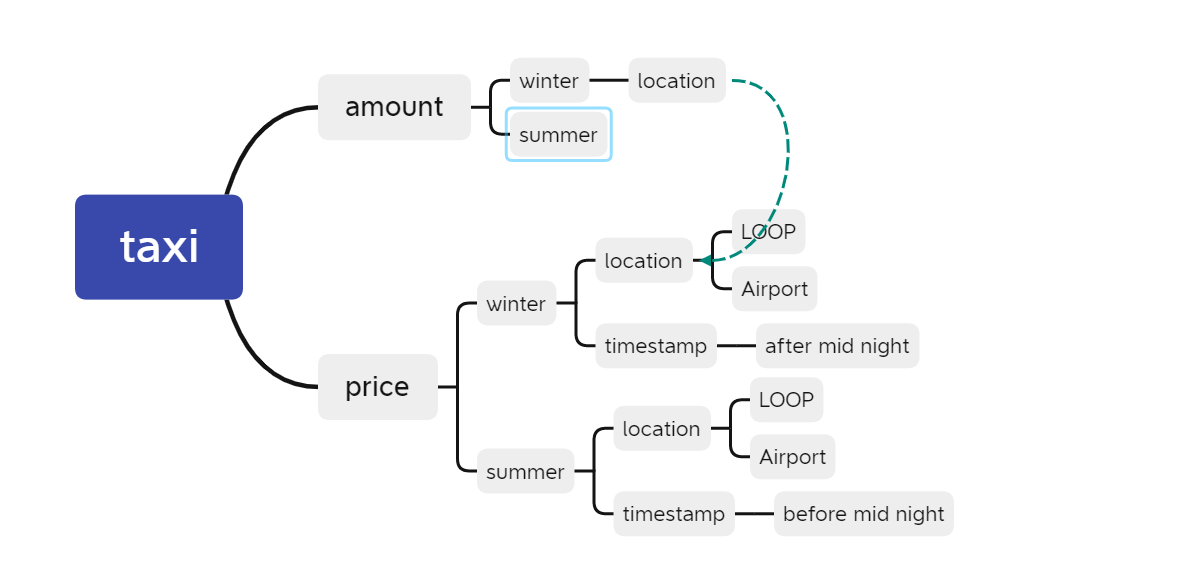
Haoxuan Zeng hzeng20@uic.edu

# 1 Data Sets and prepossessing

**We obtained our data from a website called data, city of Chicago, focusing on Taxi Trips. This database contains 200 million pieces of records, with a lot of information included such as time stamp, pick up and drop off area, tolls and tips, ext. In our project, we picked up 2.32 million pieces of records as our dataset and we divided the dataset to 2 parts regarding the seasons, summer and winter. The summer dataset contains data from 1st June, 2021 to 30 Sept, while the winter dataset contains data from 1st Oct. to 31 Dec. We also obtained the Boundaries/community areas of Chicago and we mainly concentrated on the airport and loop area and use zip code to make a data clip for this specific area.**



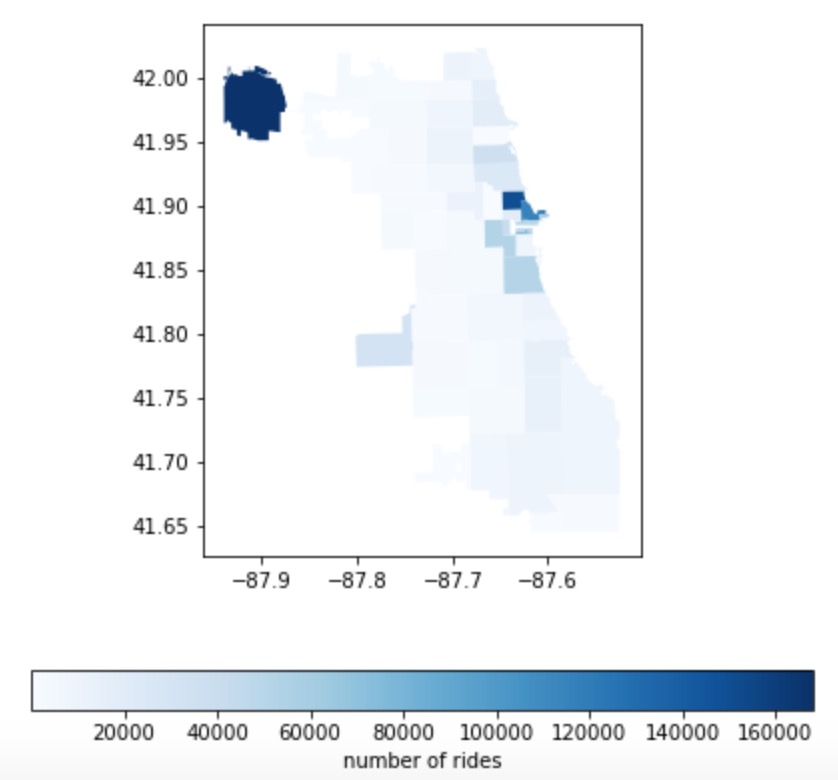
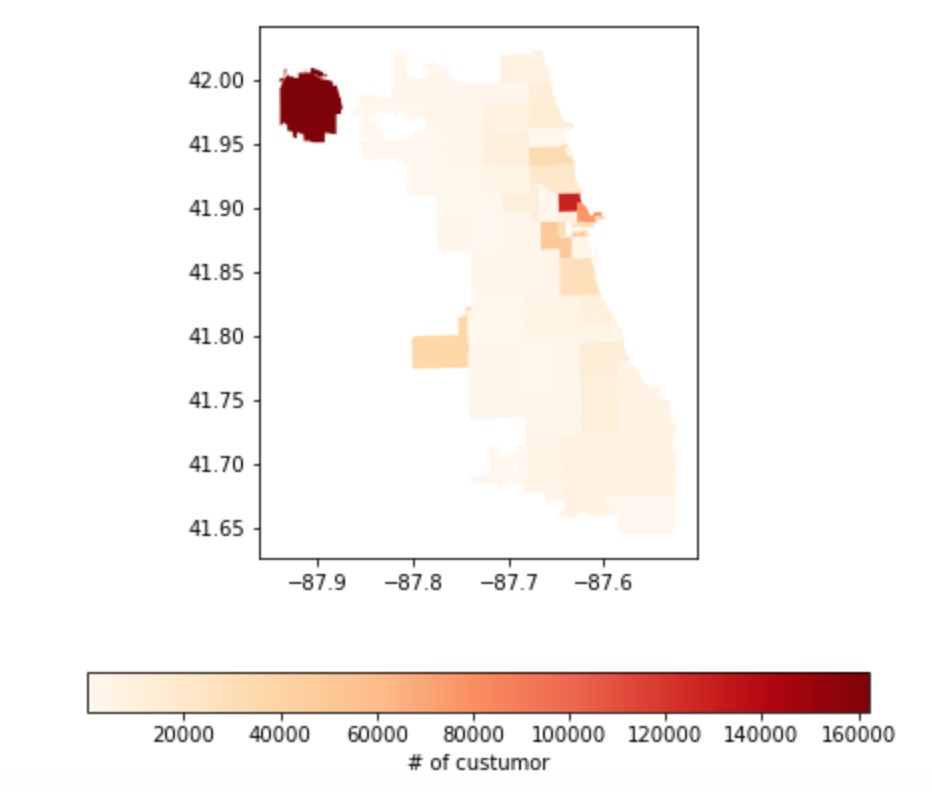
## 2 Exploratory Data Anlysis(EDA)



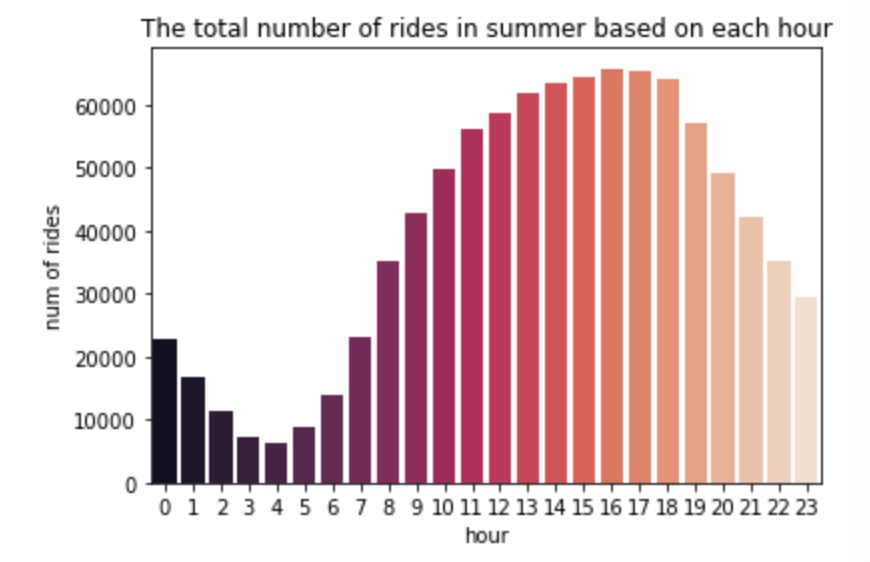
Here is the mind mapping through our data analyzing. Firstly, we separate our data into two dimension: the amount and the price. Secondly, we explore our data in different seasons and different places.

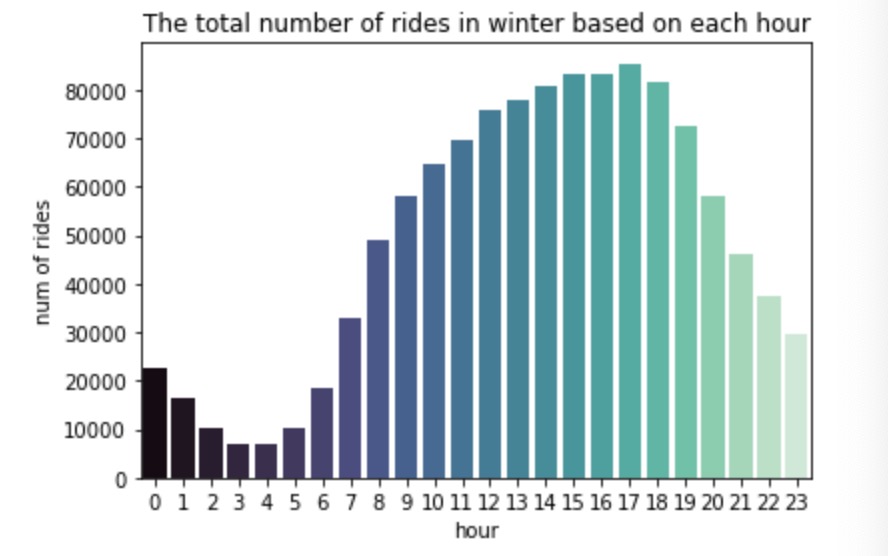
### 2.1 Different numbers in seasons

Firstly, we investigated the number of customers in summer and winter and we found that way more people would take taxi in the winter which might due to the extremely cold weather. Some patterns were witnessed, for example, people traveling from the airport and within the loop area are more likely to call a taxi.



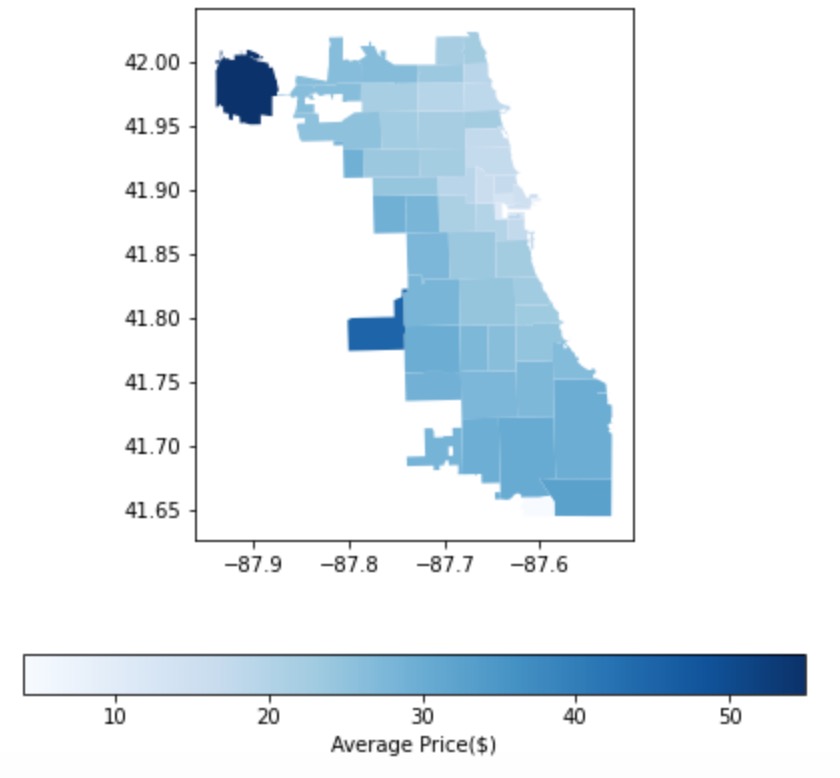
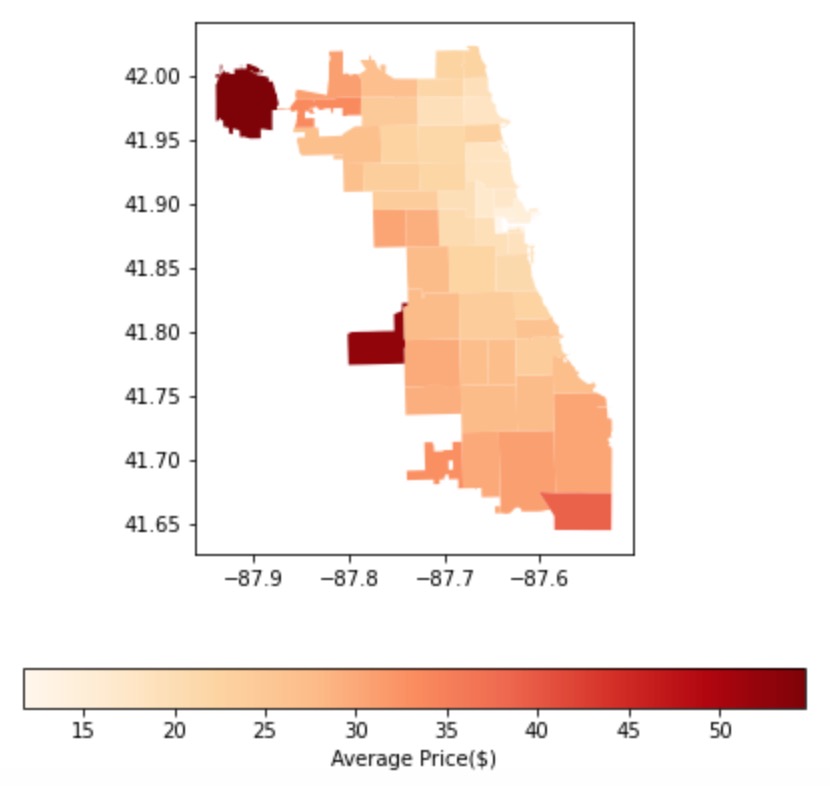
This is the time distribution of riders in summer and winter. The patterns are similar for summer and winter, as you can see, the busy hours in summer and winter are 9am to 11 pm which is very intuitive, but remember the number of total customers was higher in winter comparing to that in the summer.



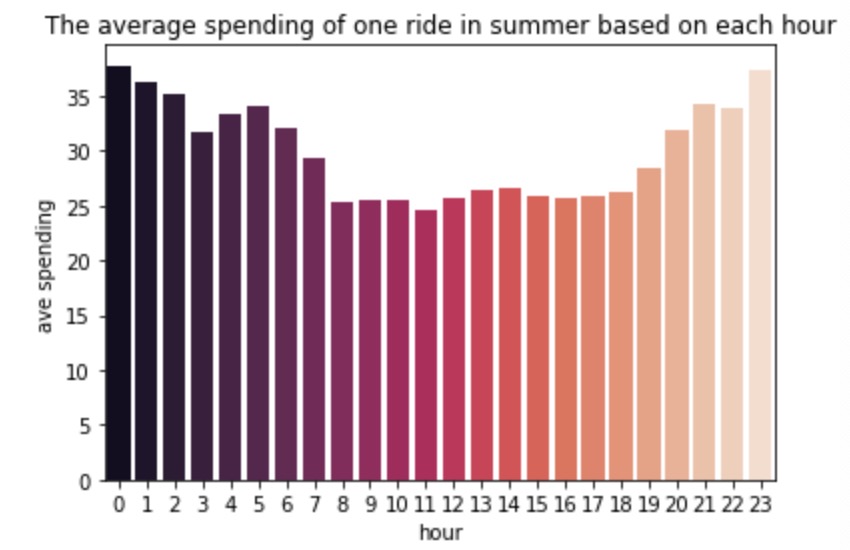


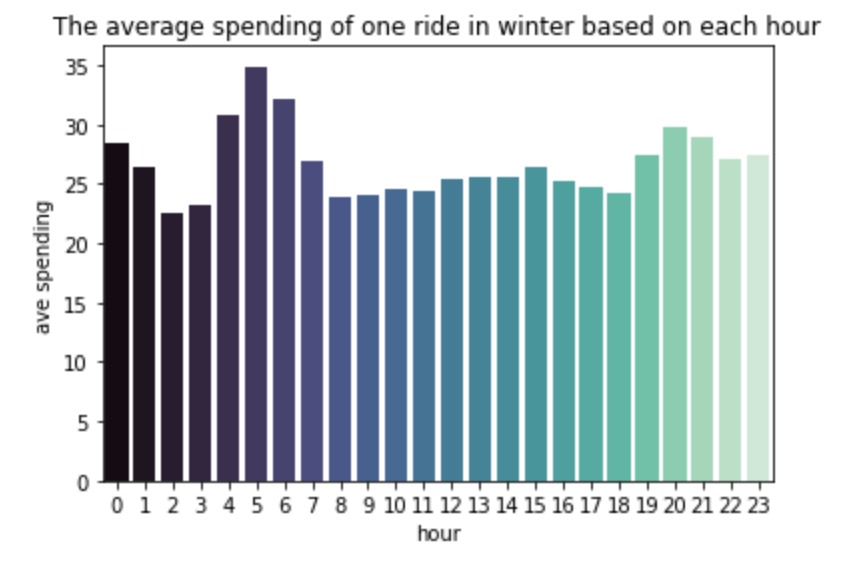
### 2.2 Average price in seasons

This two panels show the average tolls in summer and winter. The darker the color the higher the price. According to the price range scale and the color distribution, there is no significant difference regarding the price in different seasons, but the price distribution did showed a specific pattern, for example, if you called a taxi to go to the O’hare airport, you are probably paying more than 50 dollars with tips. However if your destination is loop area, you are more likely to pay lower than 30 dollars.

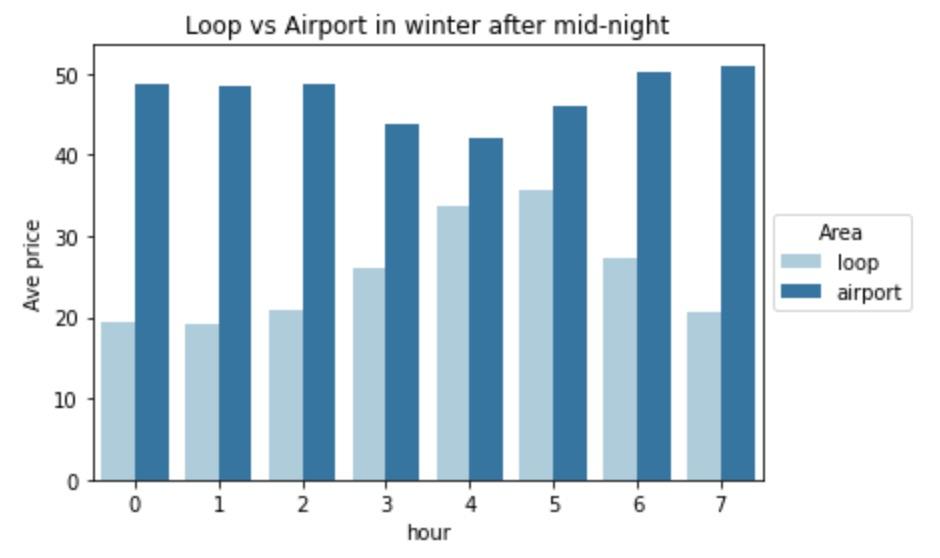


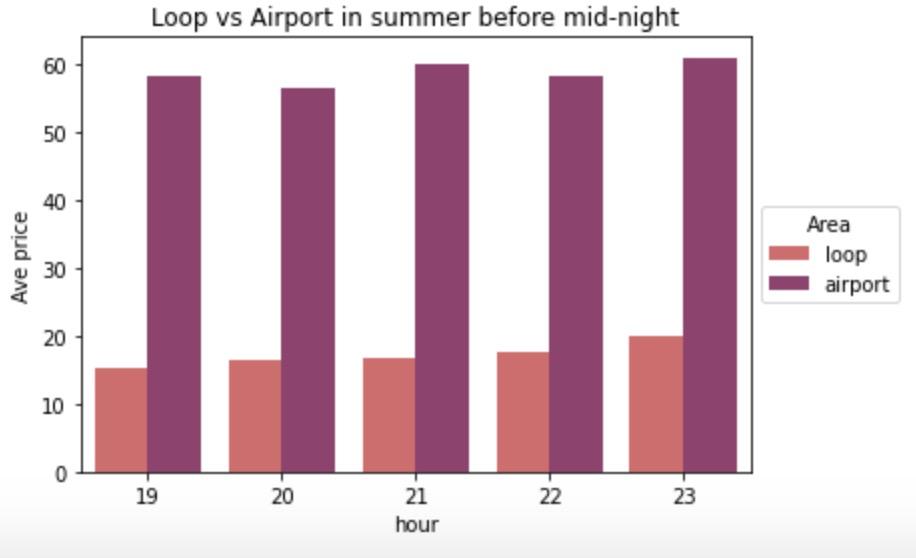
We further investigated the price distribution within a day, which is shown in this two panels. Interestingly, we found that the price peaks around 8am and 6pm, and people will pay more in the early morning in a winter day while pay more in the late night during summer as well.





Let’s move on to the more specific figures. In this two panels, we can see that, thought the price is increasing during these two specific time periods, we can still tell that the most tremendous increase happened after the mid-night in the loop area and happened before the midnight in the airport area.





Our next goal is to make Price prediction, time spend prediction and cheapest time prediction. They are about How much will a person pay for his trip in one places to another of Chicago. How many time will a person spend when they start a taxi ride from one place to another. And finally we would Find the most recommendation time for a person to take a taxi.