How is people's riding time with shared bikes assciated with where they rent the bike and their subscription status in San Francisco?

Introduction

Nowadays low carbon travel caught more and more attention, with public shared bikes, people are more willing to ride bikes instead of driving, and it is more convient to park and the cost is also lower. Today we are going to find out the relationships between people's duration time, their home zip code and if they subscript to the membership of the system or not.

Our dataset is from Aug 29, 2013 to Sept 9, 2013 in San Francisco, the U.S. In San Francisco, there are a lot of bike sharing stations just like in downtown Toronto area. There are also many bike sharing companies in SF that people can choose. Normally,there are two types of payment that people can choose, one is through a subscription, another one is purchasing single trips. For example for Lyft,if people purchase a subscription, it cost them $5 for 30 min for each trip, for trips longer than 30 min, it cost them $0.13 for each additional minute, and for people who do not have a subscription, Lyft charges them by minute but a lot more than $0.13 per minute. However, every bike sharing brand has its own station, and the subscription is not universal used, therefore, depending on which brand people choose, cost can be vary, thus, number of trips, the duration time, connection with zip code can be varied.

Therefore, we will look at the data and insert some plots to see if there are relationships and analysis the basic pattern in these relationship.

Importing Data

In [89]:
```
import pandas as pd
import numpy as np
import qeds
%matplotlib inline
```

Summary Statistics

In [70]:
```
#Read the trip file as DataFrame use np.funtions.
trip = pd.read_csv("trip.csv")

#Select zip_code, duration time and subscription status columns from the trip file.
Xs_and_Y = trip[['zip_code', 'duration', 'subscription_type']]
#Looking at the types of each columns, prep for the summary statistics.
Xs_and_Y.info()

#Print summary statistics for Xs and Y by their type.
print(Xs_and_Y.describe(include=[object])  )
print (Xs_and_Y.describe(include=[np.number]))
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 669959 entries, 0 to 669958
Data columns (total 3 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   zip_code           663340 non-null  object
 1   duration           669959 non-null  int64
 2   subscription_type  669959 non-null  object
dtypes: int64(1), object(2)
memory usage: 15.3+ MB
       zip_code subscription_type
count    663340            669959
unique     7439                 2
top       94107        Subscriber
freq       78704            566746
            duration
count  6.699590e+05
mean   1.107950e+03
std    2.225544e+04
min    6.000000e+01
25%    3.440000e+02
50%    5.170000e+02
75%    7.550000e+02
max    1.727040e+07
```

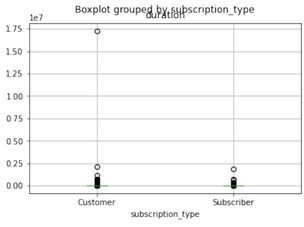
From summary statistics, there are two types of data:

Duration contains duration time for every trips in unit of second, so its a "number" type of data, and the summary statisitc presents how many observations there are, the mean, std, max and so on.

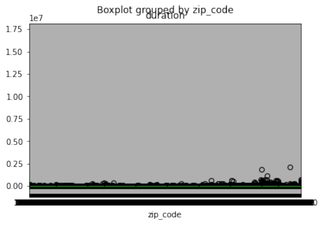
Zip-code and subscription_type are object type of data. Although zip-code is a collection of numbers in order, it means nothing to calculate the mean and the std, it only presents where do people live. Therefore we only count them and we can see there are 663340 people registered their home address in the system and there are total 7439 unique zip-code in the system. The one has been registered the most of zip-code in 94107, maybe there are more bike station near that location, or maybe just people around there like biking. Subscription_type is a dummy variable, like mentioned in the introduction, it only says if the person subscript or not in the past trips and we can see that in this data set, there are more people subscript in the system.

Graphs with Analysis

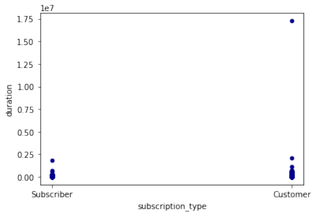
In [82]:
```
boxplot_st = Xs_and_Y.boxplot(column='duration', by = 'subscription_type')
```



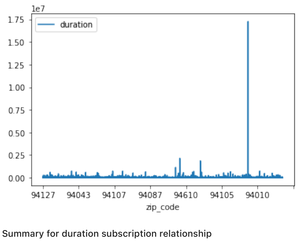
In [83]:
```
boxplot_zip = Xs_and_Y.boxplot(column='duration', by = 'zip_code')
```



In [88]:
```
ax1 = Xs_and_Y.plot.scatter(x='subscription_type',
...                         y='duration',
...                         c='DarkBlue')
```



In [87]:
```
lines = Xs_and_Y.plot.line(x='zip_code', y='duration')
```



Summary for duration subscription relationship

From the box plot, we can see the average biking time for customers or subscribers using the public bikes in SF is very similar, which is kind of something that I expected but still superised, since with subscription, you can basically ride for free for every 30 minutes to 45 minutes trips but when you are a customer, you need to pay for your every single trip within 30 minetes, therefore, everyone was trying to keep the trip within 30 minutes to aviod the second cost, we can see without the outlier in the customer dataset, every trip is less than 25 min. The fact I was superised is there are actually this many customers will purchase single trip since they are relatively expensive, most of the single trips are about $3 for 30 min. But still, the number is users in both types are similar, customers are a little less than the subscriber (according to the summary stat). And since subscription is a dummy variable, which means it will only contains 0 (customer) or 1 (Subscript), therefore, the plot graph looks similar to the boxplot and the information is quiet similar.

Summary for duration zip_code relationship

While I want to see the relationship between people's houses' zipcode and their duration time, it is hard to analyze that without category the range of the zip code. But we can still see something with the two graphs we get now. First, we can see from the boxplot graph that average duration time comparing all zip_code area is pretty much the same. Next, we know that the average population density is different across the city, it is just like in dt Toronto, there is always more people than Markham, therefore, we can see at the right part of the boxplot, there are more trips and longer duration time, and maybe there has more stations, more people, more events. But at the left side of the graph, there is not so many trips and duration time. By looking at the line plot in the second graph with the same data and summary statisitc, we can visually see how popular the bike sharing is in zip-code 94107, the people live in that area used bike at least 30 times more than people live in other area.

Why choose these two variables with duration time?

It interests me of how bike sharing companies are manage their business, zip-code is no doubt one of the most important elements when considering operate the business, therefore, by looking at the data of zip-code, we may guess where does this company want to foucs the most (there might has more stations, maybe at the center of the city...) or maybe there is the only place they can operate because of competition. By looking at the relationship of the duration and subscription, it tells us how people behave when it comes to buy the pass or not, so the company may consider to switch their strategies if there are more people using the one-time purchase. It is interest to see above results because that is not exactly I would expect.