Analysis of Chicago Divvy Bike Share Program Shows Unbalanced Demand Pattern

Jialu Chen, Yang Qu, Rouyi Ding

1. Introduction

Biking is beneficial to both of our health and environment. Many cities have launched public bike-sharing system to promote bicycle usage. In bike-sharing systems, people can borrow bikes from one station and return the bike back to any stations in the city, which can be used as a short-distance trip supplement for private vehicles as well as regular public transportation. Divvy program in Chicago is one of the mature systems that we would like to explore.

In bike-share system, the travel behaviors of people in different age and gender groups can be quite different. Meanwhile, for stations located at different places in the city, the bike usage can be quite skewed and unbalanced. Some stations that individuals like to borrow bikes from will lack enough bikes for people to check out, while some other stations that people normally return the bikes to will get jammed easily without enough docks for upcoming bikes. Therefore, we want to discover different usage patterns among various user groups and time periods, as well as flow patterns of stations.

The bicycle sharing system in Chicago can be separated into 4 segments according to flow patterns. In these four segments, the flow patterns of transition and business areas are opposite and obviously unbalanced. This unbalanced pattern is caused by commuters.

2. Method

2.1 Data Source

We investigated the dynamic bike flow patterns and customer types of bike sharing system in Chicago from January 2015 to December 2015. Divvy was launched in late June 2013. The system had 474 stations and 8274 bikes in 2015, and increased to 535 stations and 9214 bikes in 2016, which greatly improved its coverage and maximum working load. All bikes are available 24 hours, each station has a touch screen kiosk and docking system which support bikes check-in and check-out using a member key or ride code. Bike-sharing system allows people to borrow bikes with either “24-hour pass” or “annual subscribed membership”. “24-hour pass” is usually preferred by people for temporary usage, such as tourists and occasional riders, but charges per day are slightly higher. Meanwhile, “annual subscribed membership” is a great deal for people with frequent travel needs, such as local office workers and students.

The data we used in this project was obtained from Divvy website. (https://www.divvybikes.com) The datasets include trip, station and customer type information from January 2015 to December 2015. Trip dataset has 3183439 observations described by trip ID, start time and station, stop time and station and trip duration. Station dataset consists of station name, dock capacity, coordinates and created time. As for customer type, it is correlated with each trip and includes information about gender, age and user type.

2.2 Data Cleaning

Since the raw data we downloaded from Divvy website is already pretty tidy, our data cleaning is completed in only three steps. Firstly, we filtered and deleted trips with duration less than 1 minute, because these trips are too short and may not show a common use rule of the bike-sharing system. As a result, about 3.76 percent of trips were deleted. Secondly, in order to achieve standard data formats and make further statistical analysis easier, we transferred data to numeric type and unified the time format. Thirdly, we generated more variables, such as the date, time, exact hour and weekday variables from the original start time and stop time.

Since the raw data we downloaded from Divvy website is already pretty tidy, our data cleaning is completed in only three steps. Firstly, we filtered and deleted trips with duration less than 1 minute, because these trips are too short and may not show a common use rule of the bike-sharing system. As a result, about 3.76 percent of trips were deleted. Secondly, in order to achieve standard data formats and make further statistical analysis easier, we transferred data to numeric type and unified the time format. Thirdly, we generated more variables, such as the date, time, exact hour and weekday variables from the original start time and stop time.

2.3 Statistical Techniques

To investigate how flow patterns vary among stations, we first summarized demand features, namely, the amount of net check-in of each station, and then grouped stations with similar demand features. In this way, we could find out spatiotemporal over-demands distribution for bikes and docks in 2015 and further analyze the reasons for unbalance.

Specifically, the first step was to define the demand features for stations, we first divided a day into 12 2-hour time windows starting from 1:00am. Then for each station, we calculated the net check-in amount which was given by difference between total check-in amount and total check-out amount in each time window.

The second step was to group stations with similar demand features into same cluster. According to previous study, methods like K-means and hierarchical clustering are often used. Here we applied the hierarchical clustering method. Grouping the stations was an unsupervised problem without a true response and in hierarchical clustering, we could choose how many clusters we should have according to the dendrogram, while in K-Means method we need to set number of clusters K first, of which we didn't have any prior knowledge.

In hierarchical clustering, we got a distance matrix from R function `dist()` to evaluate the dissimilarity of stations according to 12 net check-ins. A dendrogram was built by this distance matrix using `hclust()`. This dendrogram was an upside-down tree starting from treating each station as one cluster and then fusing the most similar clusters until all stations belong to one single cluster. According to the shape of the tree, we cut the dendrogram at a reasonable height, which resulted in 4 clusters. `cutree()` gave us the final clusters of stations. Here we chose manhattan distance in `dist()` to calculate dissimilarities between stations and complete linkage in `hclust()` to define

dissimilarities between clusters.

3. Result

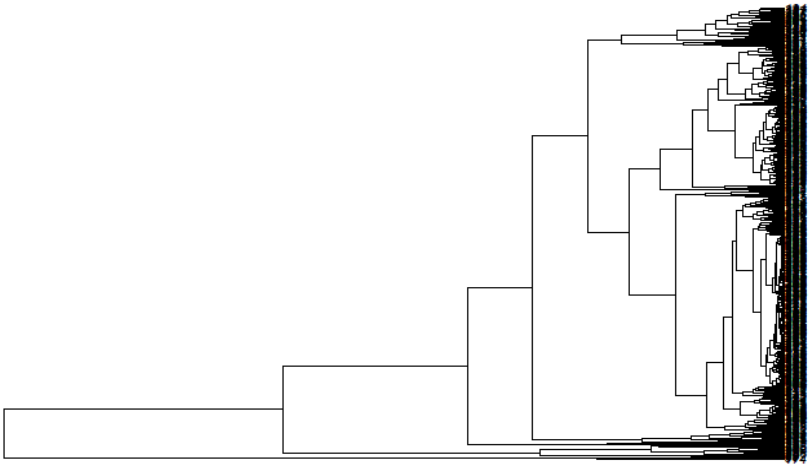

Figure 1: Cluster Dendrogram for Net Check-in Amounts

In hierarchical clustering, we got dendrogram shown in Figure1. According to the dendrogram, we cut the tree at k = 4 and got 4 clusters. The resulting clusters are shown in Figure 2. Then for each cluster, we investigated different demand features by comparing check-in and check-out amount within each hour.

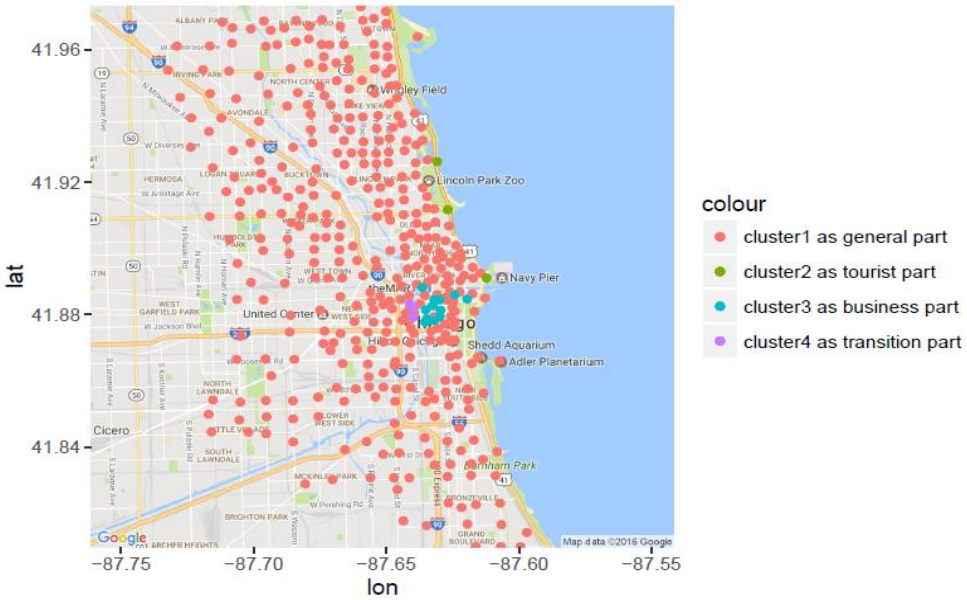

Figure 2: Flow Pattern Clusters for Stations in Chicago

Figure3 shows the flow patterns of the four clusters. We could clearly see four different flow patterns and spatiotemporal distributions of clusters.

For cluster 1, it included most stations and nearly covered the whole Chicago except center part, thus we referred to cluster 1 as general area in following analysis. Looking at the flow pattern shown

in Figure3, we found that the check-in and check-out behaviors were quite similar during a day. The check-out amount exceeded check-in amount slightly in morning peak hours and opposite in afternoon peak hours. As for noon time, there were still a large amount of trips taking place in cluster 1. In general, there were no significant peaks in the morning and afternoon.

For cluster 2, there were only 3 stations included, which were near the Great Lake, thus we referred to cluster 2 as tourist area. According to figure3, Check-in and check-out patterns of cluster 2 were quite similar with the number of check-in slightly greater than check-out. Trips from these stations mainly happened between 10:00 am and 8:00 pm and reached peak at about 3:00 pm.

For cluster 3, all stations were near the center of Chicago. In the morning between 5:00am and 10:00am, a large amount of bikes were checked in with only a little checked out. While in the afternoon after 3pm, the amount of check-in decreased a lot while number of check-out went up dramatically. With this opposite demand for check-in and check-out, we speculated that cluster 3 was a business area and people came to work in the morning and left in the afternoon. Thus, we referred to this cluster as business area.

For cluster 4,its flow pattern is almost opposite from cluster 3, with people mostly checked out in the morning and checked in in the afternoon, which suggested that people might ride a bike from cluster 4 area to work. From the map, we found that these stations were near the Union Station and a few other big stations. It is likely that people took subways to these subway stations and then switched to a bike to go to work. We will refer to this cluster as transition area.

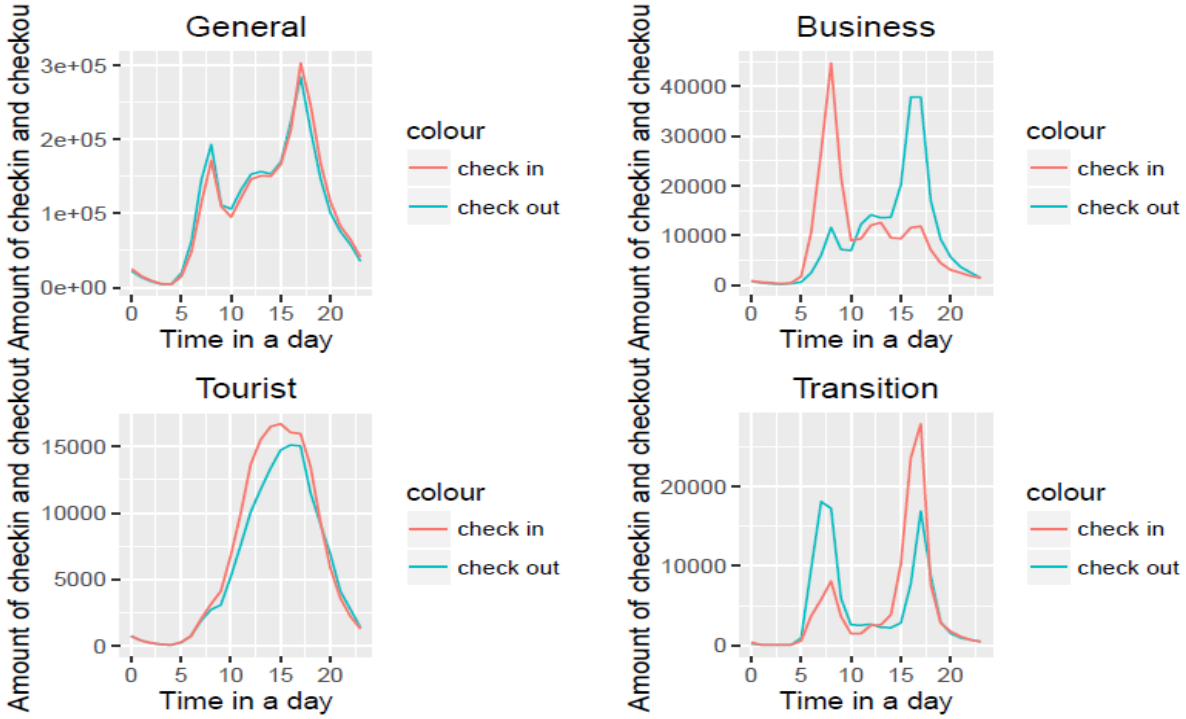

Figure 3: Check-in and Check-out Demand for General Stations

According to the analysis above, we put our focus on two clusters: business area and transition area, because these two areas have especially unbalanced check-in and check-out patterns during morning peak hours and afternoon peak hours, which reflects serious over-demand for bikes and docks.

In order to find the main source of over-demand in different time periods, we used pie graph to show how each cluster contributed to the trips associated with business and transition clusters respectively and compared performance of three different time periods. First, we divided time into three periods: morning peak hours (5:00 am - 10:00 am), noon hours (10: 00 am – 3:00 pm) and afternoon peak hours (3:00 pm – 6:00 pm). Secondly, we calculated the relative number of trips checked out from transition cluster. Since the number of stations in each cluster varies a lot, the cluster with more stations is likely to have more trips. Therefore, we divided the number of trips by the number of stations to get relative number of trips, which eliminated the effect of clusters' difference in the amount of stations. Results about trips went to transition area was got in similar way.

The result of trips checked-out from transition area were shown in Figure4. We found that in morning peak hours, more than half trips went to business area. While in other two time periods, the number of trips went to business area decreased obviously. Together with the flow pattern of transition area, these pie graphs might be an evidence that the over-demand for bikes during morning peak hours was caused by people going to work.

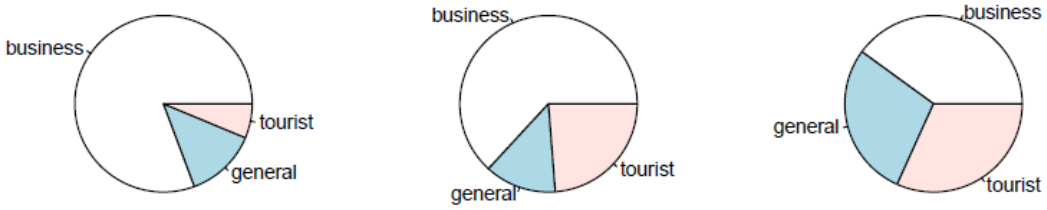

Figure 4: Proportion of Check-out from Residence in 5:00 - 10:00, 10:00 - 15:00 and 15:00 - 20:00

As for trips went to transition area, we got Figure 5, which had opposite proportion changes in three time periods. In three pie plots, trips from business area were more than all other areas, especially in afternoon peak hours, which again might be caused by commuters.

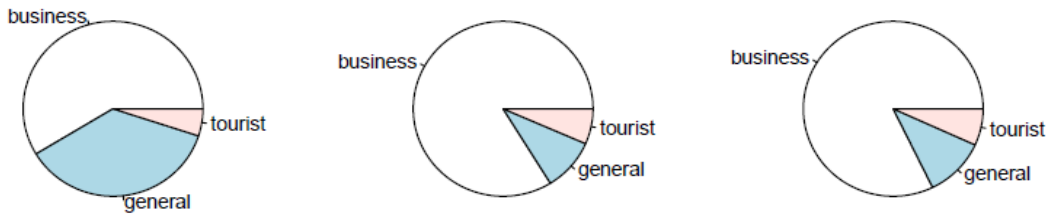

Figure 5: Proportion of Check-in to Residence in 5:00 - 10:00, 10:00 - 15:00 and 15:00 - 20:00

To further ensure our guess that the unbalanced pattern in transition and business area were mainly caused by people going to work, we studied the check-in and check-out patterns separately in weekdays and weekends and found there were different flow patterns between trips in weekdays and weekends.

From the Figure 6, in weekends, the original unbalanced flow pattern disappeared. The figures were almost unimodal in every cluster in weekend which means the unbalanced pattern comes from the weekdays.

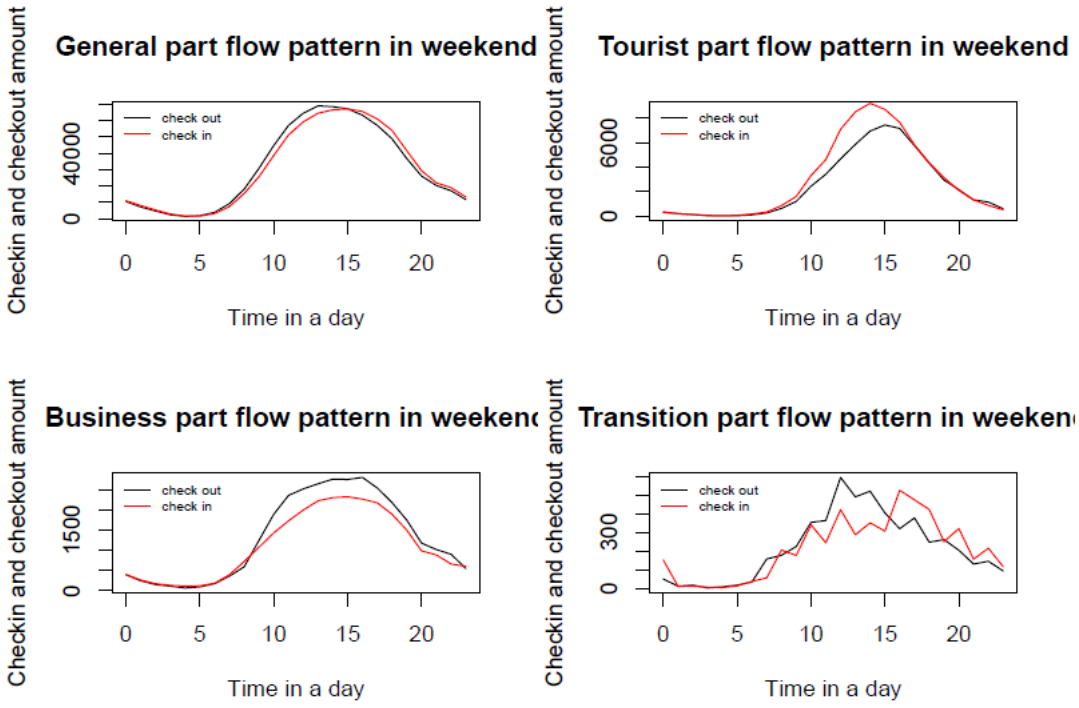

Figure 6: Flow Pattern in Weekend

In addition, we did user type analysis and gender analysis. We found that the proportion of subscribers between transition and business area was about 20 percent larger than average level. And the proportion of male between transition and business area was about 15 percent larger than average level. We used two-tailed test of population proportion to test if the two proportions are significantly different. The null hypothesis of the two-tailed test about population proportion can be expressed as follows:

$$p = p_0$$

where p_0 is hypothesis value of the true population proportion p. Define the test statistic z in terms of the sample proportion and the sample size:

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0\,(1 - p_0)/n}}$$

Then the null hypothesis of the two-tailed test is to be rejected if $z \leq z_{\alpha/2}$ or $z \geq z_{\alpha/2}$, where $z_{\alpha/2}$ is the $100(1-\alpha)$ percentile of the standard normal distribution.

The output for the user type test:

From Part	To Part	Time	proportion	p-value
Transition	Business	5:00-10:00	0.9885693	<2.2e-16
Business	Transition	5:00-10:00	0.935719	<2.2e-17
total	total	5:00-10:00	0.7238193	

From the above test, the portion of subscribers in the center of the city was much larger than that in the suburbs of the city. Around 90% of people were subscribers in the trips from transition part to business part, while there was about 70% of subscriber in all trips. This high percentage showed that there were more regular users in the center of the city.

The output for the gender test is:

From Part	To Part	Time	proportion	p-value
Transition	Business	5:00-10:00	0.9298129	<2.2e-16
Business	Transition	5:00-10:00	0.8888102	<2.2e-16
total	total	5:00-10:00	0.7490187	

Besides, by analyzing gender of users, we found the proportion of males is larger in trips from transition area to business area.

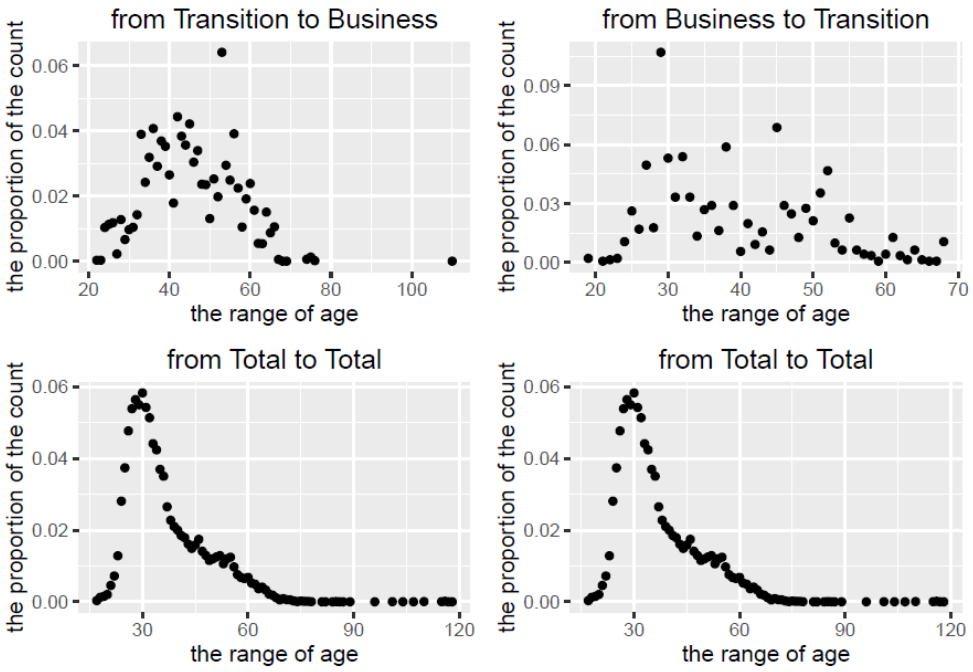

Figure 7: Range of age

Also, we did some age analysis of the users and found that the users in the center of the city tended to be older than the general users as shown in Figure 7. The users' age distribution in the center of the city was different from the age distribution in the suburbs of the city. The users' ages tended to be 30s to 50s in the center of the city, while the general age was much younger than this. Considering all of the analyses above, we can give a profile of users taking trips from transition to business areas in the center of the city. They are mainly male subscribers in their 30s to 50s, which suggests that they might have stable jobs considering their ages and regular needs of bicycle trips. These people are thus likely to be commuters riding bikes to work.

4. Conclusion

This study investigated bike-sharing system from station perspective. We developed an effective approach to extract four meaningful clusters from massive amounts of travel data. Among these four clusters, transition area and business area have opposite unbalanced flow patterns.

To figure out the cause of the unbalanced problem, we further analyzed the trip sources and user features of these two areas. For trip sources, we found that in peak morning hours, more than half trips that started from transition area ended at business area. While in peak afternoon hours, trip directions became opposite. As for users associated with these trips, we found that they were mostly male subscribers of 30s to 50s. Since these users were likely to have a stable jobs and they were moving in the center of Chicago in morning and afternoon peak hours, we concluded that the unbalanced pattern was caused by commuters.

The derived flow patterns and cause analysis were valuable to provide reference and evidence for sustainable transportation plan. For system operators, they may need to consider additional redistribution approaches to solve over-demand for bikes and docks in the transition and business area. For instance, the system operators may propose incentives to encourage people to return bikes to the transition area during peak morning hours. For government, they may get better understanding to traffic demands and make improvement to public transportation, which may help ease the traffic problem in Chicago. In addition, the general flow pattern of Chicago may apply to cities with similar downtown structures. It is a good reference for bike-sharing system design.

5. Limitation and future work

By now we have found out that an unbalanced flow pattern exists, but this result cannot be directly used. From our plots and maps we might suggest that new stations are needed in certain areas, such as the business area, but to figure out where and how many new stations should be built, we need to have a quantified measure of effects of newly online stations on surrounding stations.