Data Viz Final Project

Bingying Xia

Abstract

I use several visualization methods to study the spatial and temporal features of sharing bike system in Toronto.

The bike stations concentrate at downtown and spread out to the nearby wards. Stations located in the dense areas tend to have a high degree. There are some stations in Spadina-Fort York where the check-out number goes over the check-in number. Special cares should be paid to these stations.

Member riders enjoy the major amount of total rides, and tend to have trips with shorter duration. The overall usage of sharing bikes declines steadily over the three months. Though there is no evidence to prove that the total usage has a weekly period, usage in weekdays are more intensive than that in weekends.

The distribution of the routes# are right-tailed. 90% of the routes have less than 20 records. There are two loops in the top 10 most popular routes. The popularity of the loop route at Bay St/Queens Quay W (Ferry Terminal) is mainly contributed by casual riders. Also, casual riders tend to spend more time riding bikes in the evening in this route.

I cluster all the stations into three groups with cosine distance and complete linkage according to their temporal behavior: AVeraGe, Morning-Out-Night-In, and Morning-In-Night-Out. Each group has specific daily flow dynamic.

Introduction

Public Bicycles System Company (Bike Share) provides the City of Toronto with a network of bikes throughout the downtown core. Bike Share Docking stations allow casual and member users to pick up, and drop off bicycles from location to location, 24 hours a day, seven days a week.

The reason why I selected this topic as my final project is complicate. First, since now I'm studying near Great Toronto Area, I am eager to explore some local data. Second, I'm a fan of sharing bike system. I benefited from the sharing bike system a lot in Shanghai, China, because it narrowed the time distance between my home and the nearby metro station. Last but not least, all the datasets used in this project are published online. It's not hard to make bricks with straw. Though the dimensionality in this dataset is not high, there are still some interesting tasks to work with.

Interesting Tasks

- Explore the spatial features of stations. Find out how the location affects the check-in and check-out flow of the station.
- Identify temporal patterns of rides. Visualize the sharing bike dynamics of station states and trip circulation patterns. Study the time effect and user type influence on the temporal travel patterns.
- Represent flows between origin and destinations. Study features of the most popular route.
- Understand how station roles into origin and destination changes through the day. Cluster all the stations into several groups according to their flow dynamics.

Description of Dataset

The data used in this project is provided by the Toronto Parking Authority (https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/#343faeaa-c920-57d6-6a75-969181b6cbde) and publicbikesystem (https://tor.publicbikesystem.net/ube/gbfs/v1/en/station_information) and contains information about complete sharing bike rides in 2016 Q4. The data comes in .csv file with a total size of 29.7 MB and contains over 140 thousand trips. Each ride record consists of: Trip_id, trip_start_time, trip_stop_time, trip_duration_seconds, from_station_name, from_station_name_id, from_station_name_lat, from_station_name_lon, to_station_name, to_station_name_lat, to_station_name_lon, user_type, distance. Riders are anonymized to protect their privacy.

The following table illustates a small slice of the dataset.

	Trip_id	trip_start_time	trip_stop_time	trip_duration_seconds	from_station_name	from_station_name_id	from_station_name_lat	from_station_name_lo
1	462305	2016-10-01 00:00:25	2016-10-01 00:06:59	394	Queens Quay W / Dan Leckie Way	7075	43.63653	-79.3958
3	462307	2016-10-01 00:00:37	2016-10-01 00:06:59	383	Queens Quay W / Dan Leckie Way	7075	43.63653	-79.3958
4	462308	2016-10-01 00:01:03	2016-10-01 00:27:00	1557	Cherry St / Distillery Ln	7107	43.65028	-79.3568
5	462309	2016-10-01 00:01:12	2016-10-01 00:27:00	1547	Cherry St / Distillery Ln	7107	43.65028	-79.3568
7	462311	2016-10-01 00:01:58	2016-10-01 00:27:59	1562	Queen St W / York St (City Hall)	7202	43.65167	-79.3841
8	462312	2016-10-01 00:01:40	2016-10-01 00:08:59	440	Church St / Bloor St E	7034	43.67139	-79.3829

Spatial station feature

Station location and degree

First, all stations and associated wards are ploted on the map. Leaflet (https://rstudio.github.io/leaflet/) is used in this project. It's an open source JavaScript library for interactive maps. The shapefile of Toronto Ward is found from Open Data - City of Toronto (https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/#29b6fadf-0bd6-2af9-4a8c-8c41da285ad7). Here, 25-Ward Model is used instead of 47- Ward model, because it is more up-to-date.

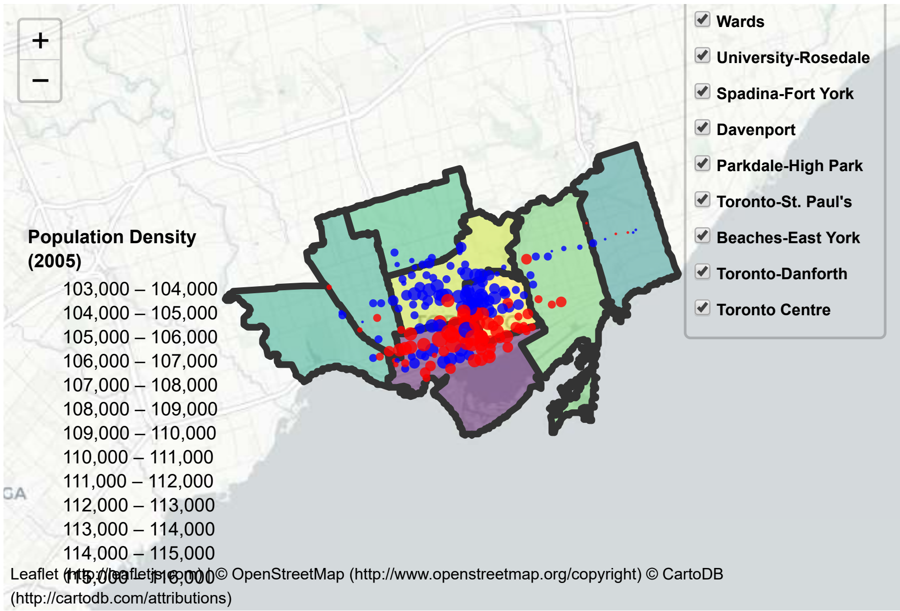
All the wards are colored by the total population of it. Ward population data are found from Individual Ward Maps 25 Ward Model (https://www.toronto.ca/city-government/elections/general-information/better-local-government-act-2018-bill-5/individual-ward-maps-25-ward-model/).

From the map, it is important to note that the sharing bike stations spread out from the city center to the surrounding wards. Spadina-Fort York, Toronto Center, and University and Rosedale enjoy the major station resources.

The dots areas are proportional to the total degree of each stations. Here, degree is the concept from network analysis. Because the sharing bike system is a directed graph, there are in-degree, out-degree, and total degree for each stations. By definition, degree indicates the number of lines that enter or go out from the nodes. In the sharing bike case, degree illustates the relationship between that station to the others. The station with a higher degree is associated with more stations. However, in my project, the location density of the stations affects the degree of the stations. Stations located in the ward concentrated with more stations are more likely to have a higher degree.

The color of the dots is determined by whether the total number of check-in bikes covers the one of check-out. Assume that there are unlimited docks in a station, so there will not be any space problem about how to check-in a bike. I focus on the limitation of the number of bikes offered in station. The red dots are the stations where the check-out number goes over the check-in number. These stations tend to locate in Spadina-Fort York. Special attention should be paid to these stations to make sure riders can find a bike as needed.

```
## OGR data source with driver: ESRI Shapefile
## Source: "/Users/apple/Documents/uwaterloo/stat842/final proj/WARD_WGS84.shp", layer: "WARD_WGS84"
## with 25 features
## It has 9 fields
## Integer64 fields read as strings:  AREA_ID
```



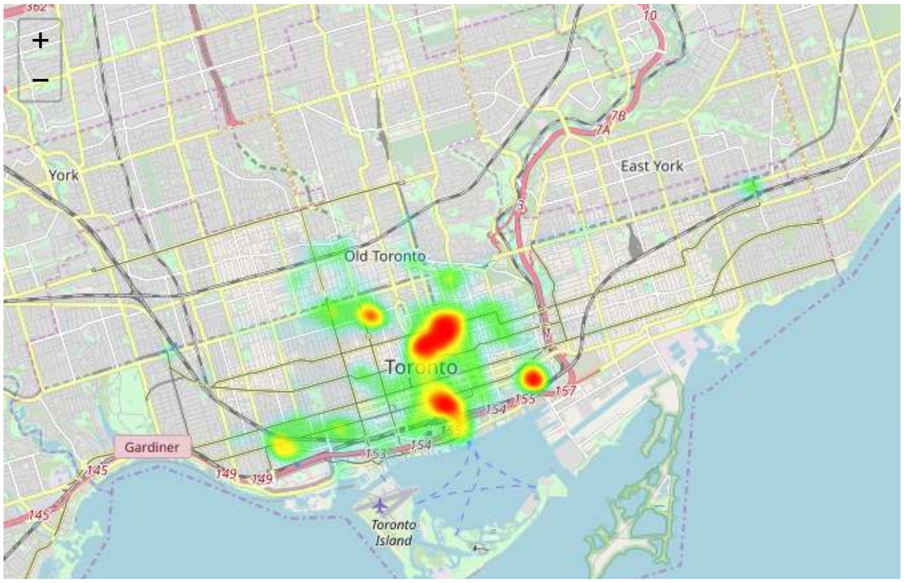
What stands out from the map above is that ward population is not significantly related to the number of bike stations in that ward. The number of stations in each ward varies a wide range. However, the population in every ward is similar to each other. University and Rosedale has 45 bike stations, but its population is not large. Meanwhile, Beaches and East York has relatively large population but few stations. We also count the bike sharing number in each ward.

Taking into consideration the huge difference in number of stations in each ward, it's not surprising to find that the ride spread very unevenly in each ward. Spadina-Fort York, Toronto Center, and University-Rosedale enjoy a major share of the total rides. However, the ride records per station (ride rate) is low in Beaches-East York. It's beyond my expectation that though there are only four stations in Parkdale-High Park, the ride rate is pretty much higher than Beaches-East York.

	Station#	Population(10k)	Ride#(100)	Ride/Station#
Spadina-Fort York	51	11.55	604.53	1185.35
University-Rosedale	45	10.95	382.34	849.64
Toronto Centre	37	10.69	342.29	925.11
Toronto-Danforth	11	10.79	32.98	299.82
Davenport	10	10.85	32.15	321.50
Beaches-East York	7	10.95	6.92	98.86
Toronto-St. Paul's	7	10.43	18.11	258.71
Parkdale-High Park	4	10.88	9.34	233.50

Station heat map

Heat map is used to visualize the spatial density of stations. The three red areas are highlighted because of the high density of bike stations. After checking the ward shape, I find that the three areas fall in University-Rosedale, Spadina-Fort York, and Toronto Center, which is consistent with the top three wards with most stations.

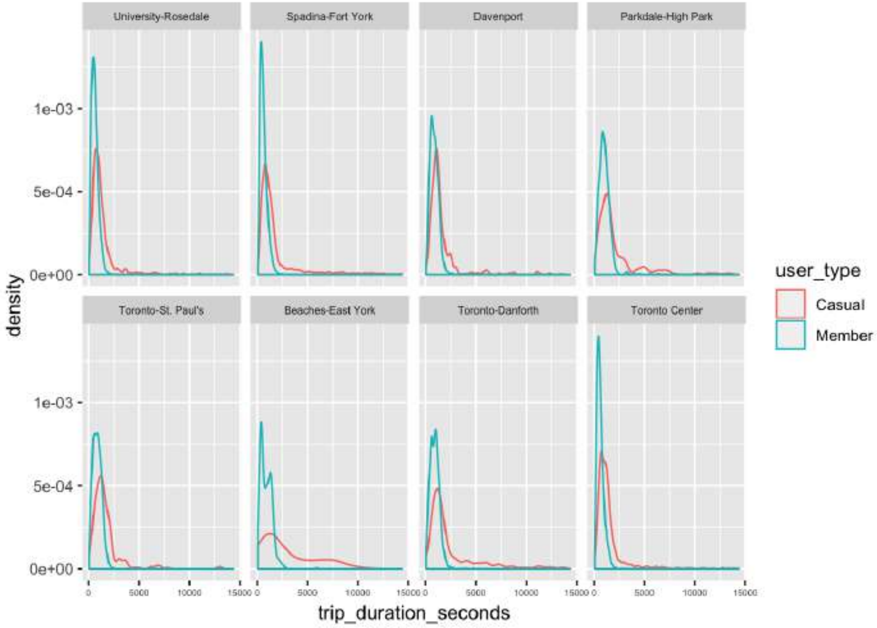


Leaflet (http://leafletjs.com) | © OpenStreetMap (http://openstreetmap.org) contributors, CC-BY-SA (http://creativecommons.org/licenses/by-sa/2.0/)

Density matrices – average trip duration

I aggregate all the rides according to their starting stations, and analyze the data in two groups: casual riders, and member riders. I check the density line chart of casual riders and member riders in each ward.

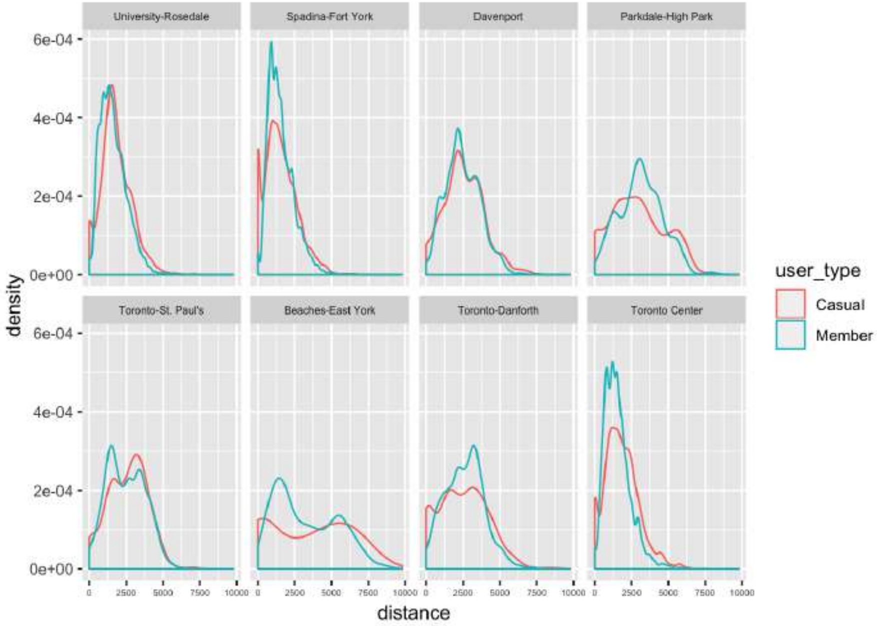
The average trip duration from casual riders is larger than that from member riders. The density shape of member riders is sharper than the casual ones. It goes up rapidly and drops suddenly. However, casual density does not have this feature. It usually goes up and down gently. There are also some small bumps in the right tail of the casual density.



Density matrices – average distance

I also check the density line chart of casual riders and member riders for each ward.

From these charts I know the density shape between casual and member riders are almost overlapped in University-Rosedale, and Davenport. However, the ones in Parkdale-High Park, Beaches-East York, and Toronto-Danforth are tremendously different. Also, the density shapes of casual riders in Spadina-Fort York, Parkdale-High Park, and Beach-East York show two peaks, which indicates some bimodality.



Temporal travel patterns

Descriptive Stats

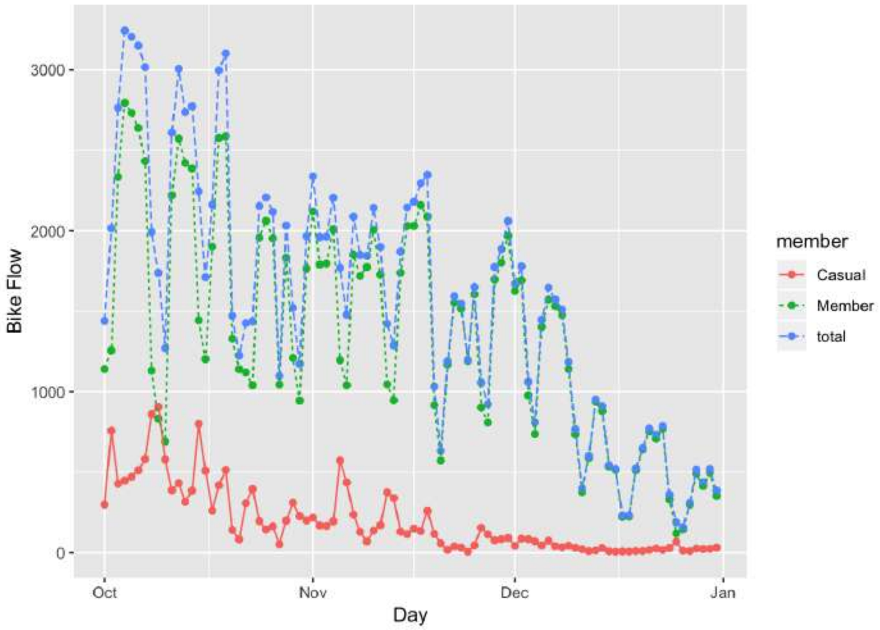
First, the overall trend of all stations all over the fourth quarter in 2016 is studied.

From Oct. 1st, 2016 to Dec. 31st, 2016, there were 142,866 sharing bike trips in Toronto. The average trip duration is 790 seconds (about 13 mins), and the average trip distance is 1656 meters, which implies that customers tend to riding sharing bike for short trip.

records	142866
station	172
start_time	2016-10-01 00:00:25
end_time	2016-12-31 23:39:00
avg_duration	790.8901
avg_distance	1656.4

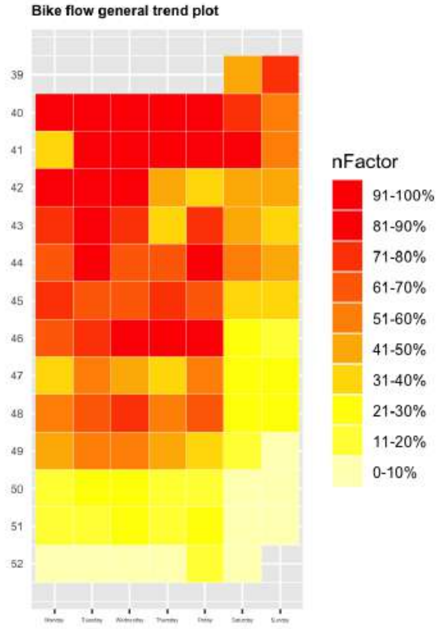
Overall trend

I also compare the travel patterns all over the three months between total riders, casual riders, and member riders. There is a decreasing trend in all the three lines. After December, the usage of casual riders dropped down to near zero. Meanwhile, the usage of member riders experienced a sharp decrease, and the size of member usage in December shrank to one third of the one in October. It might be related to the harsh winter climate in Toronto. Member usage significantly dominates the everyday bike flow. The member line is over two times higher than the casual line. Usage of members shows some periodic pattern, while usage of casual customers is more stochastic.



Calendar chart of total usage

In order to check whether the total usage data has a weekly period, I plot the day-by-day bike flow data in a calendar view. The darker red the cell is, the more total usage in that day. I can’t say that the total usage data has a weekly pattern. However, starting from November, the total usage is higher during weekdays than in weekends.



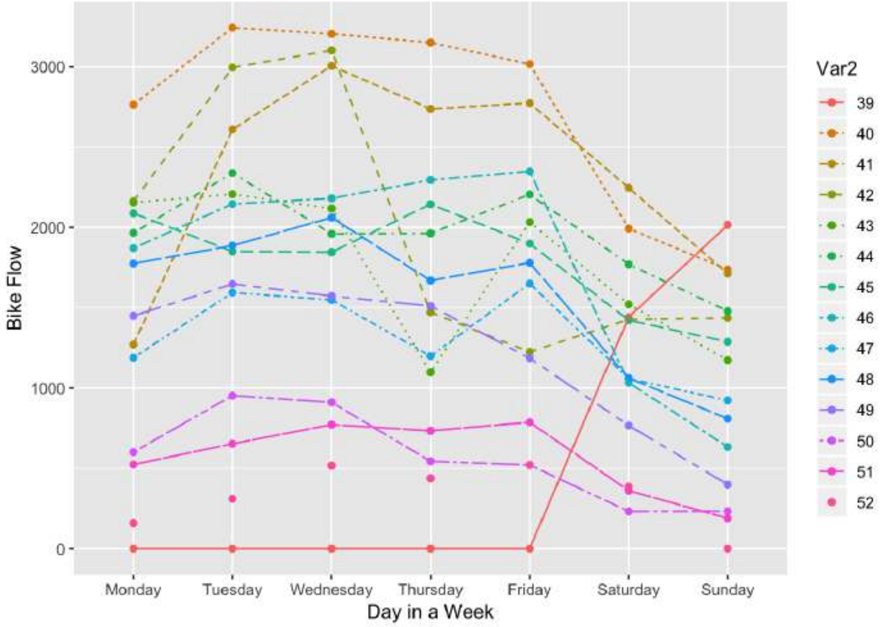
Weekly line chart

Though the total usage is not stationary over time, I still overlap the 14 weeks data in one line chart.

Most weeks hit their usage peak during Tuesday, Wednesday, or Thursday. However, there are some exceptions.

It's surprising to see that Week 39, the first two days of October reaches such a high level of usage during weekends. I checked online, but these two days are not holiday. I guess there are some special social events happened this weekend, because riding sharing bikes to attend social events is a convenient transportation mode, especially to avoid the traffic jam in the event locations.

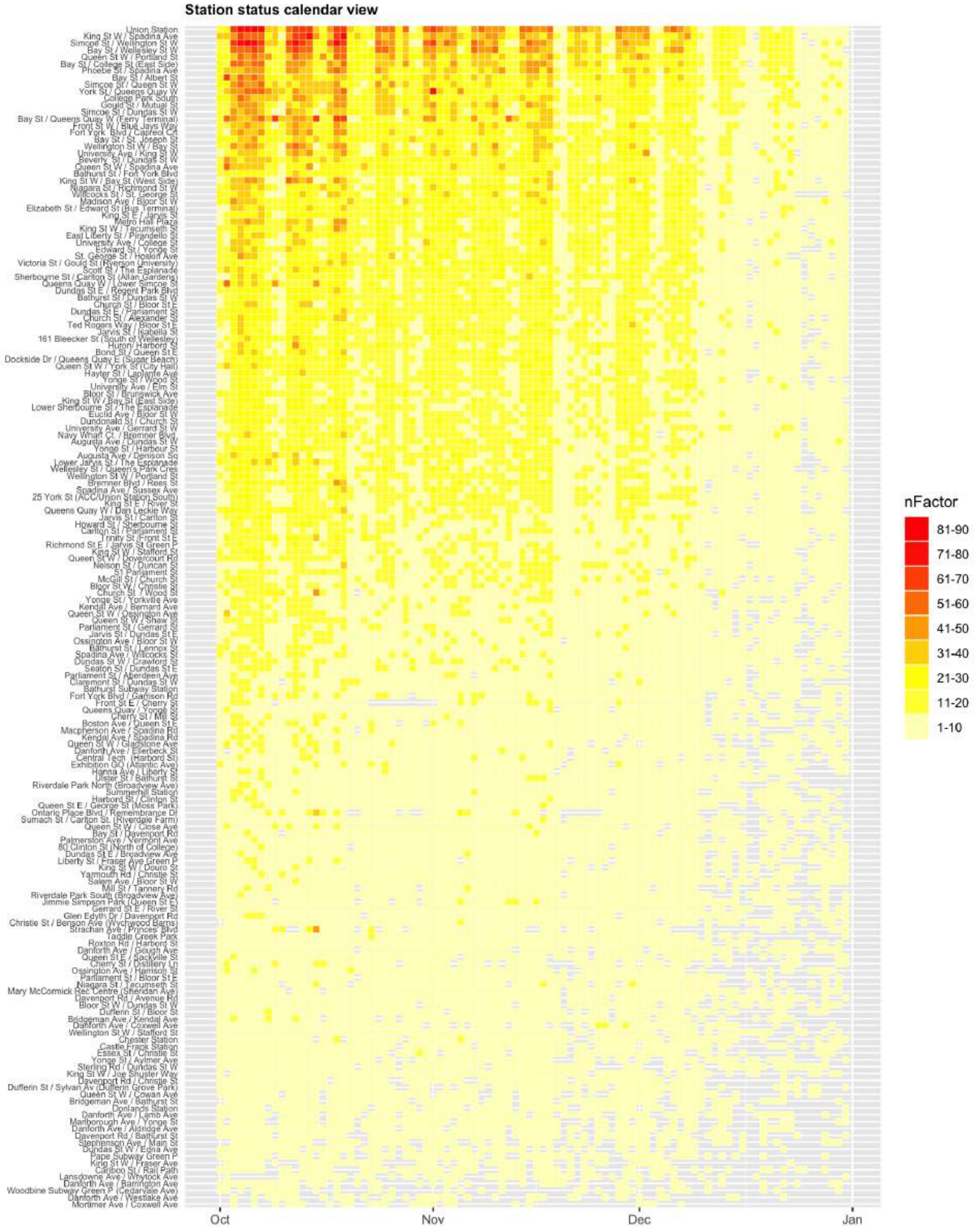
Week 42 is also an outlier. Total usage experienced an unexpected drop after Wednesday's peak. The weather didn't change dramatically in Week 42. Thus, it's hard to point out the reason for the mysterious decrease.



Calendar chart per station

Then I investigate the detailed temporal pattern for each station. Since the average over the stations are not informative, I aggregate the data for every station each day. I use a color-mapped bar to represent one station status. Each small cell in the bar stands for one entire day during these three months. Thus, the whole timeline for one station consists of one bar. The cells are colored based on the number of check-number of sharing bikes, with red associated with high level of usage, and yellow associated with low level of usage. In order to observe the patterns in the chart, I reorder the rows(stations) by the total number of usage of all three months.

I can see that there are about ten red-orange vertical stripes at the top of the chart. These vertical stripes mean that the sharing bike usage increases during these days. Meanwhile, a lot of station remain light usage status all over the time. Also, some unexpected days pop up in the chart, whose color is strikingly different from its neighbours.



Monthly daily chart per station — Weekdays and Weekends

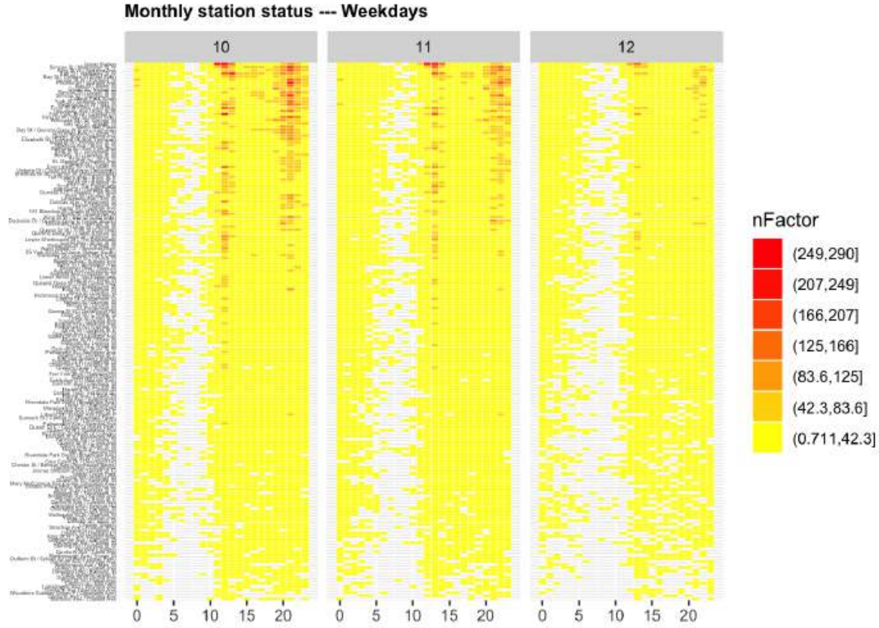
I aggregate the data into Weekdays versus Weekends to analyze the specific trend. I use daily timeline and divide the data into three months. Each line here in the chart is the total daily check-out bike number in a station in one month. Every single cell in the line represents for one-hour timewindow. The yellow-white vertical stripes at the bottom of the charts mean there is no usage during these period in the corresponding stations. And, the red-orange dots are representative for the high level of usage.

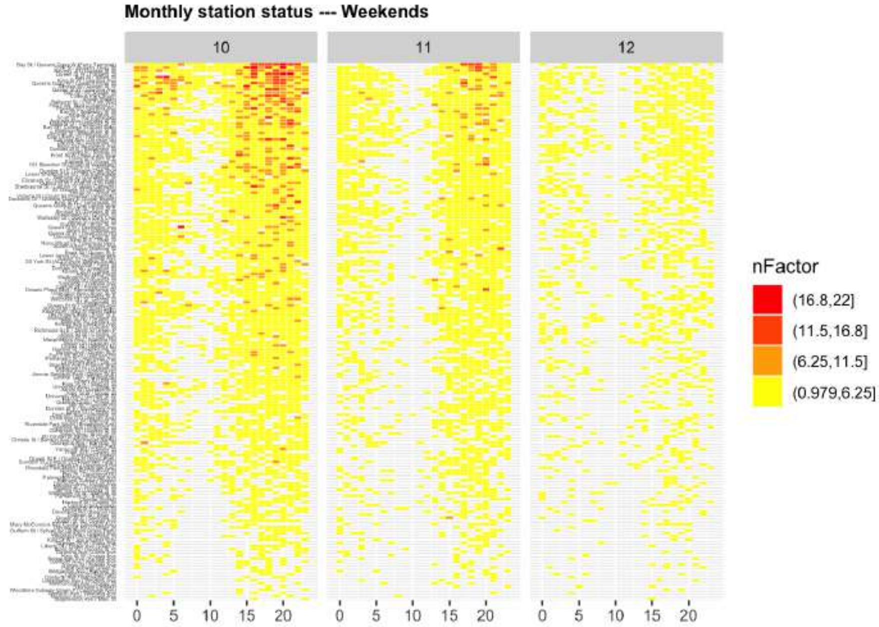
Comparing these six charts, I know weekdays and weekends are significantly different. Usage of sharing bike has two peaks in weekdays in October and November, one around 11:00, and the other around 21:00. The usage during the night is even higher than noon. I observe that the most intense activity happens in the nights of October. The duration of the weekday night peak is longer than the morning peak.

However, the usage is more stationary in Weekdays in December. The usage in December is relatively more inactive, as well.

On the other hand, sharing bike usage during weekend is at a much lower level compared to the weekdays. There are large blank areas at the bottom and the middle of the chart, which means a large part of the stations are inactive in the weekends, and almost all the stations are inactive around the noon. What's more, the decline of the usage during weekends is more rapidly over the months.

It is important to note that the color scales are different in the following two charts. Because of the difference in the demand of sharing bike between weekdays and weekends, if I ploted the charts in the same scale, patterns during weekends would be undiscernable. Thus, I choose two color scales to plot these Monthly daily charts.





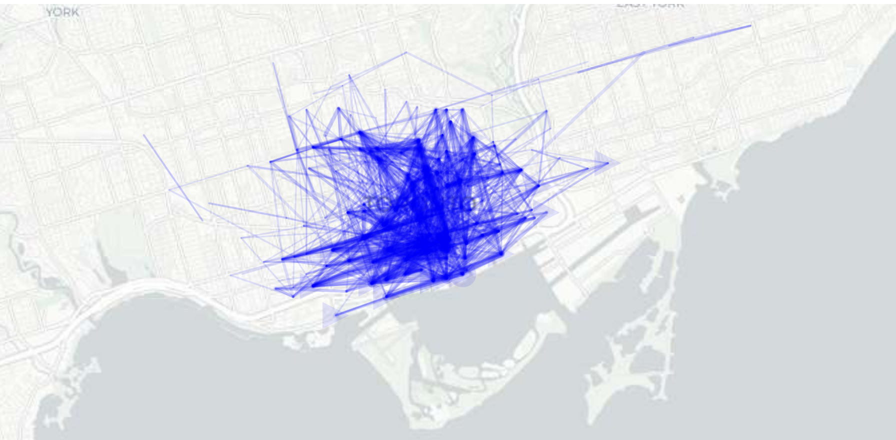
Flow study

Flow threshold selection

The threshold of flow d influences the flow structure and computation intensity. If d is too small, I will have difficulty in recognizing the structure in the trip flow. If d is too large, it requires more computational power. I investigate the quantiles of the bike flow n and find that it has a right-tailed density. I select the 20 to be the threshold. 20 is near the 90% percentile. I abandon the irrelevant trips and concentrate to analyze the top 10% welcomed routes.

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
1	1	1	2	3	4	5	8	12	21	405

The flow map shows the Euclidean travel routes between two stations. Each line stands for one specific route. The thickness of the route is proportional to the total trip number in each route.



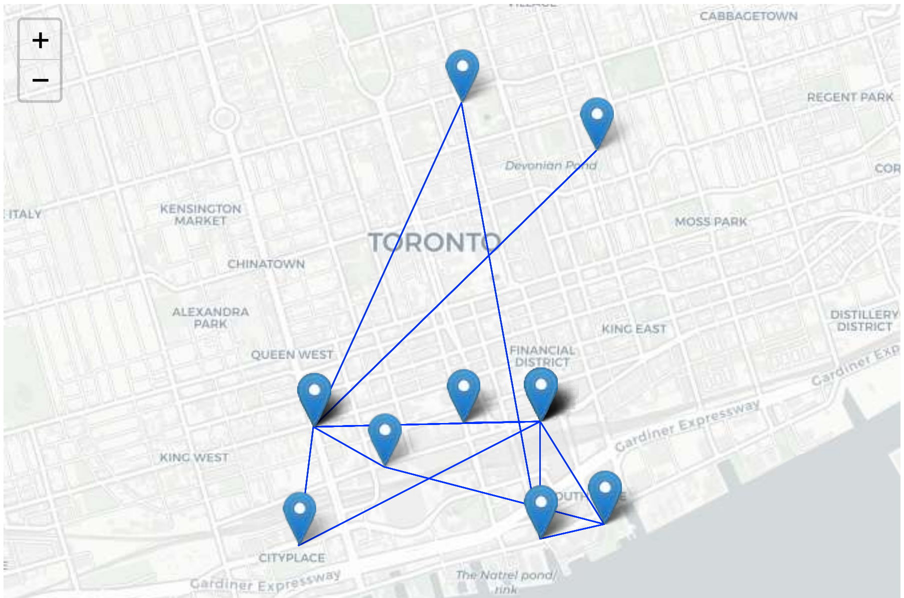
Leaflet (http://leafletjs.com) | © OpenStreetMap (http://www.openstreetmap.org/copyright) © CartoDB (http://cartodb.com/attributions)

What are the top 10 most popular routes?

The top ten most popular route are displayed below.

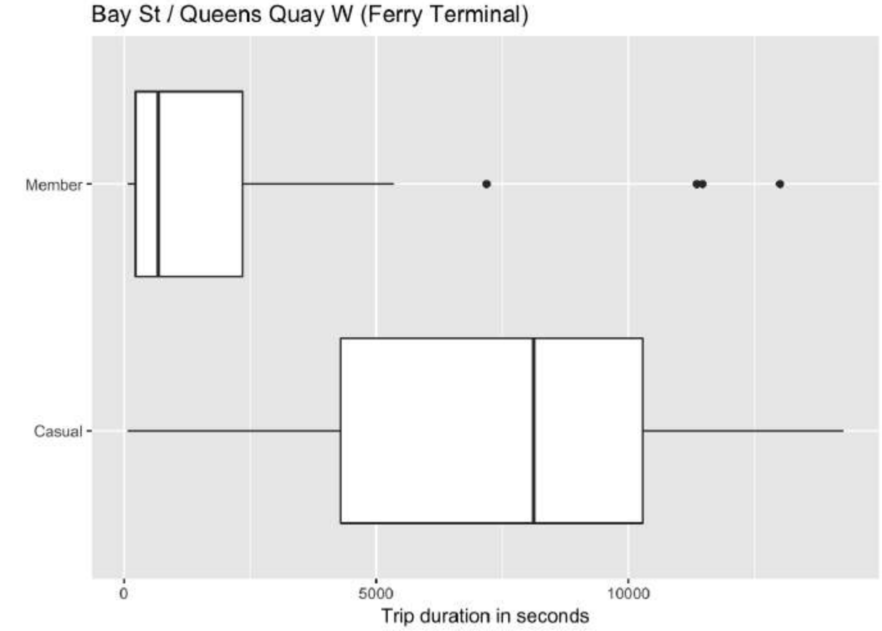
from_station	to_station	n
Bay St / Queens Quay W (Ferry Terminal)	Bay St / Queens Quay W (Ferry Terminal)	405
Front St W / Blue Jays Way	Union Station	322
King St W / Spadina Ave	Union Station	293
Simcoe St / Wellington St W	King St W / Spadina Ave	213
Union Station	Gould St / Mutual St	205
Union Station	King St W / Spadina Ave	201
Fort York Blvd / Capreol Crt	Union Station	193
King St W / Spadina Ave	Simcoe St / Wellington St W	188
Bay St / College St (East Side)	Union Station	186
York St / Queens Quay W	York St / Queens Quay W	177

It’s interesting to find that the most popular route is a loop, starting from and ending at Bay St / Queens Quay W (Ferry Terminal). Union Station is also a popular station since it appears in 6 routes. The tenth most popular route is another loop route, starting from and ending at York St / Queens Quay W.

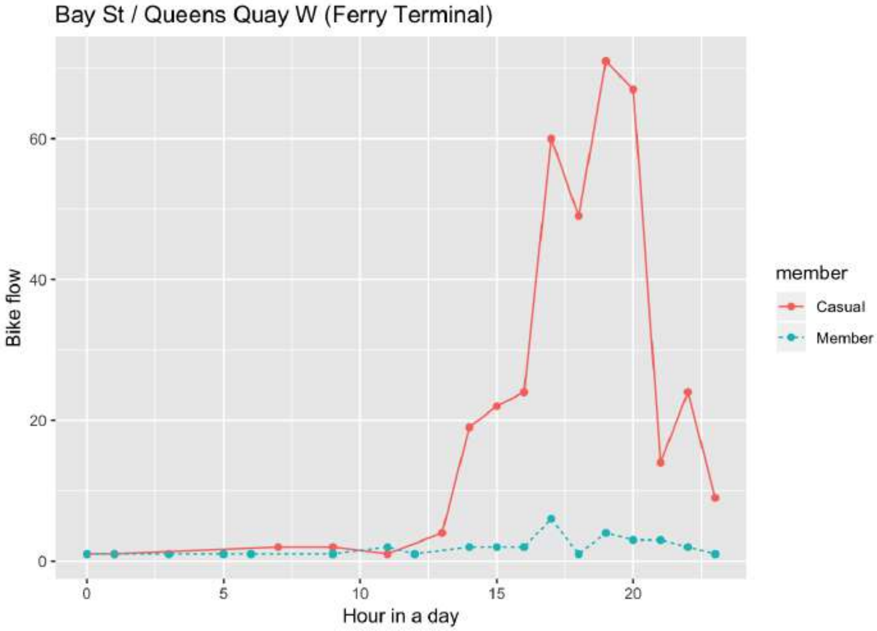


Leaflet (http://leafletjs.com) | © OpenStreetMap (http://www.openstreetmap.org/copyright) © CartoDB (http://cartodb.com/attributions)

I check the details with the loop at Bay St / Queens Quay W (Ferry Terminal). The trip duration is even ten times longer than the average trip duration in this dataset. On average, riders will spend over 2 hours in this loop route, which implies that they might go around a big circle before they go back and return the bikes at the start point. This data seems to suggest that tourists tend to rent a sharing bike as a Casual rider to enjoy the beautiful view around the city.



I also plot the line chart of the average usage of casual riders and member riders during a day. I define the usage as the number of check-out bikes. The casual riders make up the majority of the customers in this loop route. The casual usage starts to increase from the noon 12:00, and remains at a high level from afternoon 16:00 till evening 21:00.



Temporal Behavior Clustering

In order to find whether there is possible undetected structure between stations according to pickup and return activity, I try to use cluster technique to the dataset. Here I define the time window of analysis. Longer time window in the calendar plot has the advantage to show the long-term trend. However, it may lose some detailed features for the temporal behavior clustering. Some prior studies used more homogeneous time windows to aggregate biking behavior @zhou2015understanding, while other studies used some specific time window to summarize it @froehlich2009sensing. In my project, I choose to follow the homogenous division, and introduce the time window with more detailed information: every hour being a window, staring from 00:00 to 24:00. During each time window, I calculate the numbers of rent and return in each station.

$$Rent_{tk} = \sum \#bike$$

$$Return_{tk} = \sum \#bike$$

where t is one of the 24 time windows, and k is the station number.

Cosine distance

Hierarchical clustering methods are used to find the station with similar temporal demands. When clustering, I use cosine dissimilarity to calculate the distance between each station.

$$cos(Station_{ki}, Station_{kj}) = \frac{\sum_{t=1}^{24} Rent_{tki} \times Rent_{tkj} + \sum_{t=1}^{24} Return_{tki} \times Return_{tkj}}{\sqrt{\sum_{t=1}^{24}(Rent_{tki}^2 + Return_{tki}^2)}\sqrt{\sum_{t=1}^{24}(Rent_{tkj}^2 + Return_{tkj}^2)}}$$

$$d_{cos}(Station_{ki}, Station_{kj}) = 1 - cos(Station_{ki}, Station_{kj})$$

where $cos(Station_{ki}, Station_{kj})$ suggests the cosine similarity between $station_{ki}$ and $station_{kj}$. $d_{cos}(Station_{ki}, Station_{kj})$ suggests the cosine distance between $station_{ki}$ and $station_{kj}$.

I get the inspiration of using $cos(Station_{ki}, Station_{kj})$ from text clustering. In text clustering, each element X contains the count of some word for that document. In the sharing bike case, each element $Station_{ki}$ contains the count of check-out and check-in bike during certain time window. If $cos(Station_{ki}, Station_{kj})$ is large, the temporal pattern of check-in and check-out bike activity between two stations is similar. Hence, I will assign them into the same cluster.

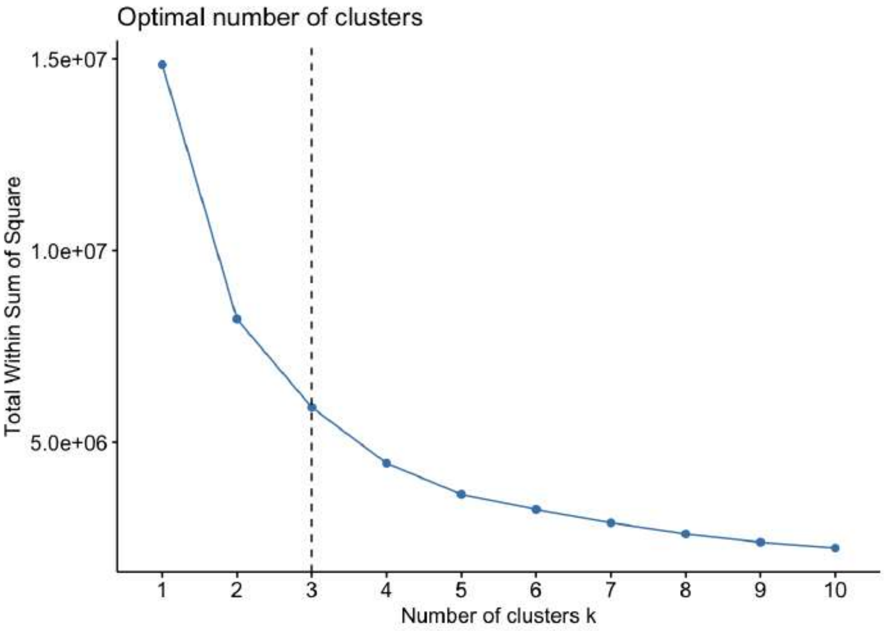
Complete linkage

Complete linkage is applied to calculate the distance between two clusters A and B. I prefer compact small diameter cluster over stringy ones.

$$d(A, B) = max(d_{cos}(x, y) : x \in A, y \in B)$$

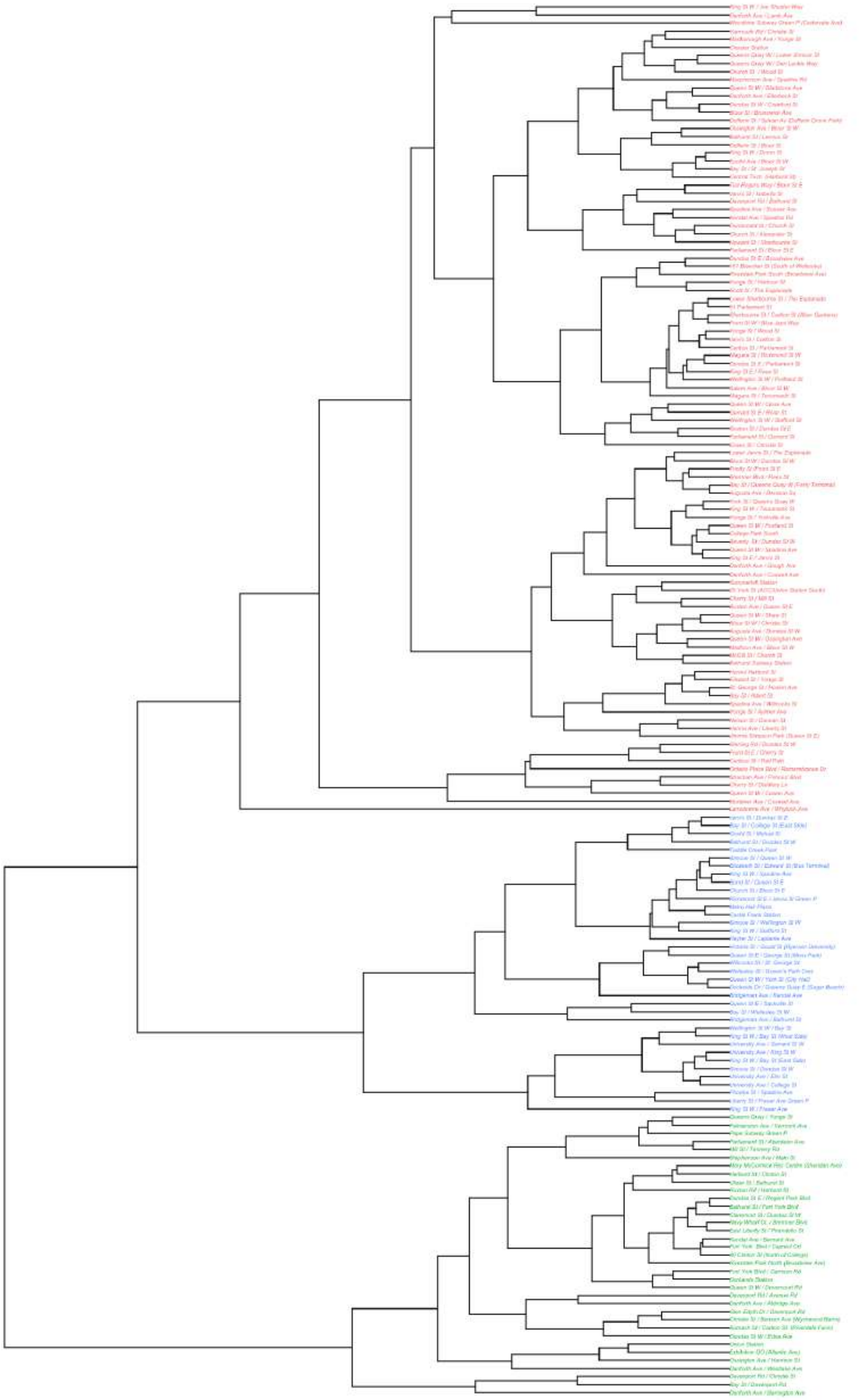
Choose proper k

In agglomerative hierarchical clustering, I need to define the number of clusters in the clustering algorithm. From the elbow plot, I select $k = 3$.



Hierarchical clustering result

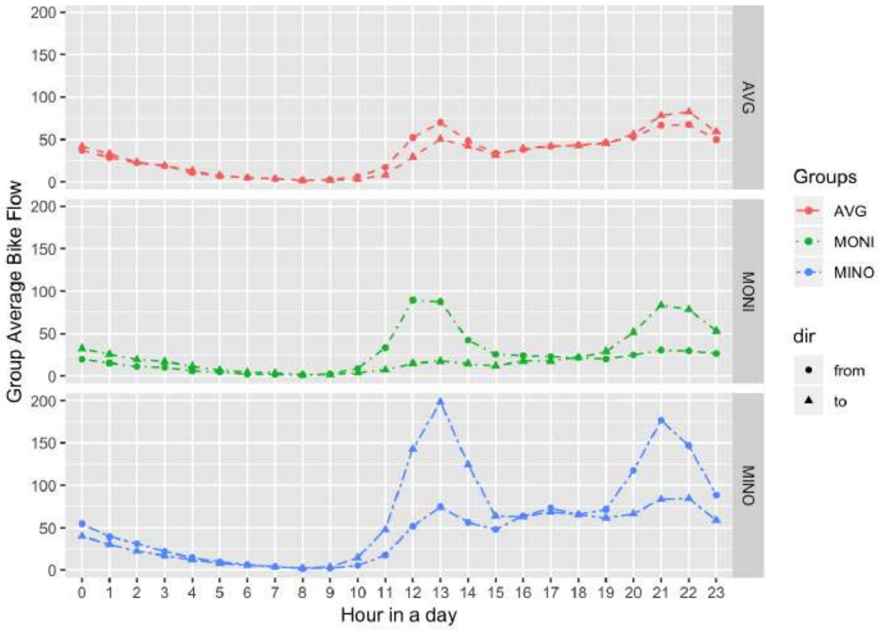
The clustering result is displayed below.


King St W / Joe Shuster Way
Danforth Ave / Lamb Ave
Woodbine Subway Green P (Cedarvale Ave)
Yarmouth Rd / Christie St
Marlborough Ave / Yonge St
Chester Station
Queens Quay W / Lower Simcoe St
Queens Quay W / Dan Leckie Way
Church St / Wood St
Macpherson Ave / Spadina Rd
Queen St W / Gladstone Ave
Danforth Ave / Ellerbeck St
Dundas St W / Crawford St
Bloor St / Brunswick Ave
Dufferin St / Sylvan Av (Dufferin Grove Park)
Ossington Ave / Bloor St W
Bathurst St / Lennox St
Dufferin St / Bloor St
King St W / Dunn St
Euclid Ave / Bloor St W
Bay St / St. Joseph St
Central Tech (Harbord St)
Ted Rogers Way / Bloor St E
Jarvis St / Isabella St
Davenport Rd / Bathurst St
Spadina Ave / Sussex Ave
Kendal Ave / Spadina Rd
Dundonald St / Church St
Church St / Alexander St
Howard St / Sherbourne St
Parliament St / Bloor St E
Dundas St E / Broadview Ave
157 Bleecker St (South of Wellesley)
Riverdale Park South (Broadview Ave)
Yonge St / Harbour St
Scott St / The Esplanade
Lower Sherbourne St / The Esplanade
51 Parliament St
Sherbourne St / Carlton St (Allan Gardens)
Front St W / Blue Jays Way
Yonge St / Wood St
Jarvis St / Carlton St
Carlton St / Parliament St
Niagara St / Richmond St W
Dundas St E / Parliament St
King St E / Power St
Wellington St W / Portland St
Salem Ave / Bloor St W
Niagara St / Tecumseth St
Queen St W / Close Ave
Gerrard St E / River St
Wellington St W / Stafford St
Sumach St / Dundas St E
Parliament St / Gerrard St
Grace St / College St
Lower Jarvis St / The Esplanade
Bloor St W / Dundas St W
Trinity St / Front St E
Bremner Blvd / Rees St
Bay St / Queens Quay W (Ferry Terminal)
Augusta Ave / Denison Sq
York St / Queens Quay W
King St W / Tecumseth St
Yonge St / Yorkville Ave
Queen St W / Portland St
College Park South
Beverly St / Dundas St W
Queen St W / Spadina Ave
King St E / Jarvis St
Danforth Ave / Gough Ave
Danforth Ave / Coxwell Ave
Summerhill Station
York St (ACC/Union Station South)
Cherry St / Mill St
Boston Ave / Queen St E
Queen St W / Shaw St
Bloor St W / Christie St
Augusta Ave / Dundas St W
Queen St W / Ossington Ave
Madison Ave / Bloor St W
McGill St / Church St
Bathurst Subway Station
Borden St / Harbord St
Charles St / Yonge St
St. George St / Hoskin Ave
Bay St / Albert St
Spadina Ave / Willcocks St
Yonge St / Aylmer Ave
Nelson St / Duncan St
Hanna Ave / Liberty St
Jimmie Simpson Park (Queen St E)
Sterling Rd / Dundas St W
Front St E / Cherry St
Cariboo St / Rail Path
Ontario Place Blvd / Remembrance Dr
Strachan Ave / Princes' Blvd
Cherry St / Distillery Ln
Queen St W / Cowan Ave
Mortimer Ave / Coxwell Ave
Lansdowne Ave / Whytock Ave
Jarvis St / Dundas St E
Bay St / College St (East Side)
Gould St / Mutual St
Bathurst St / Dundas St W
Taddle Creek Park
Simcoe St / Queen St W
Elizabeth St / Edward St (Bus Terminal)
King St W / Spadina Ave
Bond St / Queen St E
Church St / Bloor St E
Richmond St E / Jarvis St Green P
Metro Hall Plaza
Castle Frank Station
Simcoe St / Wellington St W
King St W / Bathurst St
Hayter St / Laplante Ave
Victoria St / Gould St (Ryerson University)
Queen St E / George St (Moss Park)
Willcocks St / St. George St
Wellesley St / Queen's Park Cres
Queen St W / York St (City Hall)
Dockside Dr / Queens Quay E (Sugar Beach)
Bridgman Ave / Kendal Ave
Queen St E / Sackville St
Bay St / Wellesley St W
Bridgman Ave / Bathurst St
Wellington St W / Bay St
King St W / Bay St (West Side)
University Ave / Gerrard St W
University Ave / King St W
King St W / Bay St (East Side)
Simcoe St / Dundas St W
University Ave / Elm St
University Ave / College St
Phoebe St / Spadina Ave
Liberty St / Fraser Ave Green P
King St W / Fraser Ave
Queens Quay / Yonge St
Palmerston Ave / Vermont Ave
Pape Subway Green P
Parliament St / Aberdeen Ave
Mill St / Tannery Rd
Stephenson Ave / Main St
Mary McCormick Rec Centre (Sheridan Ave)
Harbord St / Clinton St
Ulster St / Bathurst St
Huron St / Harbord St
Dundas St E / Regent Park Blvd
Bathurst St / Fort York Blvd
Claremont St / Dundas St W
Navy Wharf Ct / Bremner Blvd
East Liberty St / Pirandello St
Kendal Ave / Bernard Ave
Fort York Blvd / Capreol Crt
80 Clinton St (North of College)
Riverdale Park North (Broadview Ave)
Fort York Blvd / Garrison Rd
Donlands Station
Queen St W / Dovercourt Rd
Davenport Rd / Avenue Rd
Danforth Ave / Aldridge Ave
Glen Edyth Dr / Davenport Rd
Christie St / Benson Ave (Wychwood Barns)
Sumach St / Carlton St (Riverdale Farm)
Dundas St W / Edna Ave
Union Station
Exhibition GO (Atlantic Ave)
Ossington Ave / Harrison St
Danforth Ave / Westlake Ave
Davenport Rd / Christie St
Bay St / Davenport Rd
Danforth Ave / Barrington Ave

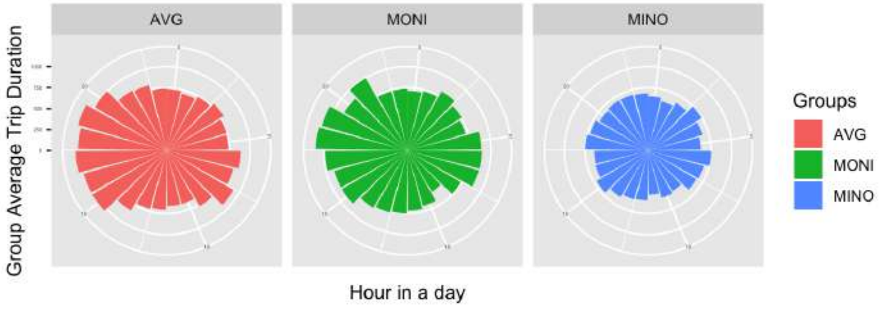
Three different patterns in the temporal pick and return activities are distinguished. From the line chart for each group, I can label the groups according to their daily check-in and check-out patterns.

- 100 stations are assigned to group 1. These stations show a slightly higher level of both check-in and check-out activity from 11:00 to 15:00 and 21:00 to 24:00. There is no significant difference between check-in and check-out activity. I can call group 1 an *AVG* group.
- 35 stations are assigned to group 2. These stations show a much more active check-out rate in the morning, from 11:00 to 15:00, while the check-in rate is about two times higher than the check-out one during night, from 20:00 to 23:00. I can call group 2 a *Morning-Out-Night-In* group.
- 37 stations are assigned to group 3. These stations share the opposite pattern compared to the Morning-Out-Night-In group. Also, this group show an active dynamism from noon to midnight. Moreover, the rate of check-in bike reached the peak (about 200) at 13:00. On the other hand, the rate of check-out bike hit about 175 at 21:00. I can call group 3 a *Morning-In-Night-Out* group.

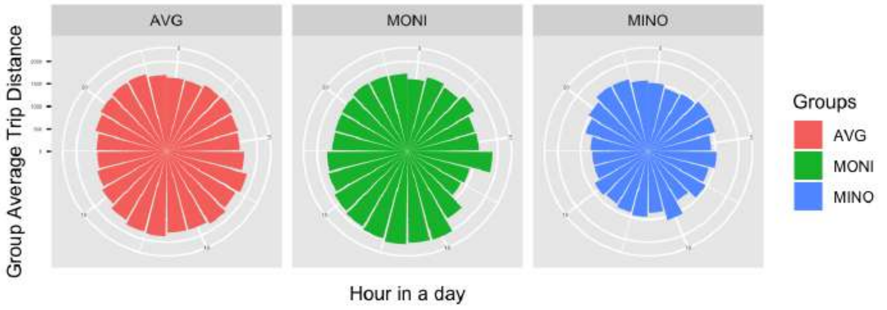
When I look across the three groups, I can conclude that in the morning AVG and MONI stations are major origins of check-out while MINO stations are major destinations of check-in However, in the night, the situation is totally opposite. MINO stations are major origins of check-out while AVG and MONI stations are major destinations of check-in.



I can also use rose diagram to display the pattern of daily check-in and check-out bike flow between three clusters. The height of the bar is proportional to the trip duration in each timewindow. What interesting about the data is that trips starting from MINO stations are shorter than the other two. In AVG and MONI group, riders tend to have longer trips during evening.



I also use rose diagram of average distance in each timewindow. It should be noted that the average distance from AVG stabilizes at a high level in the morning and fluctuate slightly after noon. On the other hand, the one from MONI is short during the morning, but grows rapidly before noon and also remains at a high level in the night. The average distance from MONI flutters gently all day long.



I summarise some trip features for the three groups. From the table I know that AVG has the longest average trip duration, and AVG group seems to be most welcomed by casual riders. Take the large group size, it's not surprised to see that AVG has the most trip records.

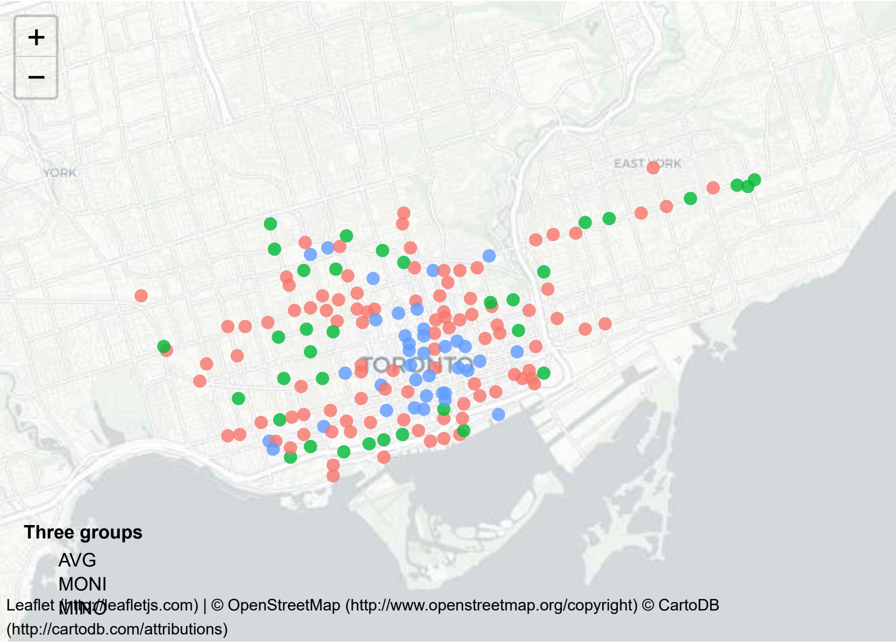
MONI group has the longest average trip distance. The member ratio is similar to the MINO one. The average trip duration is slightly smaller than the AVG one.

MINO has the shortest average trip distance, and it enjoys the highest member ration. It's reasonable to infer that stations in MINO is more popular in member riders.

Group	Casual(1000#)	Member(1000#)	Member Ratio	Average Duration(mins)	Average Distance(km)
AVG	11.8	65.4	84.72	42.5	3.47
MONI	2.1	17.5	89.29	40.8	3.73
MINO	4.2	41.8	90.87	33.1	3.08

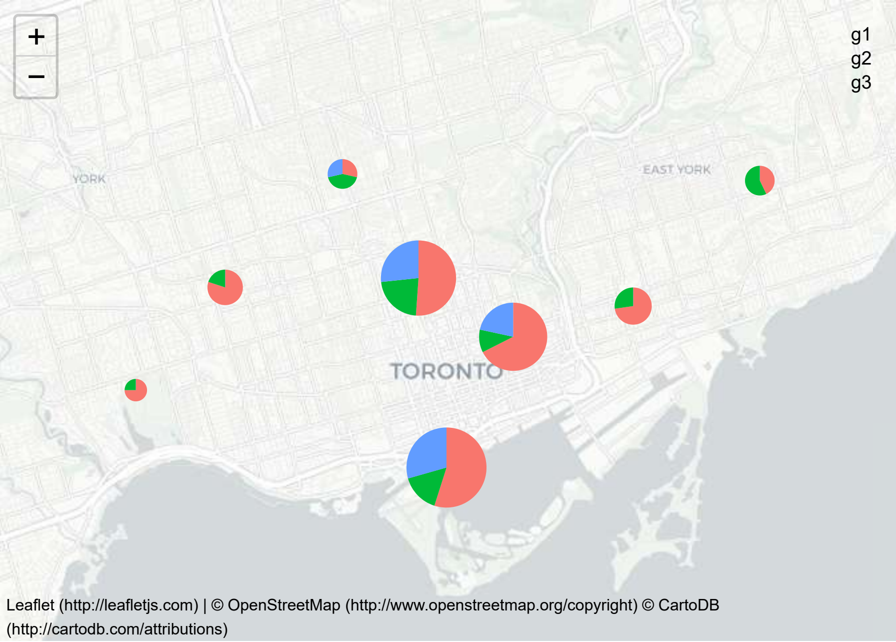
The geographical distribution of clusters is displayed below. I can use this plot to examine whether the temporal activities are related to the spatial factors. Station within the same cluster tend to locate in Toronto with certain patterns.

- AVG: Stations in AVG group is colored by red dots. These red dots spread all over the city.
- MONI: Station in MONI group is colored with green dots. MONI stations seem more likely to lie along the lake and lie around the Toronto downtown. The proportion of MONI stations amounts to nearly half in Toronto-St. Paul's, and even exceeds half in Beaches-East York.
- MINO: Station in MINO group is colored with blue dots. MINO stations are more concentrated in the center of Toronto. These stations are only located in University-Rosedale, Spadina-Fort York, and Toronto-St. Paul's. Since these three wards are main working and study areas, it's reasonable to speculate that office workers and students are the major users of MINO.


+
−
YORK
EAST YORK
TORONTO
Three groups
AVG
MONI
MINO
Leaflet (http://leafletjs.com) | © OpenStreetMap (http://www.openstreetmap.org/copyright) © CartoDB (http://cartodb.com/attributions)

Wards and clusters

I use glyphs to show the group status in each ward. Areas of circles are proportional to the number of stations. Pie slices proportional to station number in each group.



Leaflet (http://leafletjs.com) | © OpenStreetMap (http://www.openstreetmap.org/copyright) © CartoDB (http://cartodb.com/attributions)

Conclusion

I am interested in understanding the spatial and temporal behavious of sharing bike system in Toronto. With the help of several visualization methods, the hidden structure behind the system is unveiled. Actions should be conducted to deal with the unbalance of bike storage. More bikes should be transported to the stations which lack bikes. Also, there could be some deals during weekends to encourage the usage of sharing bikes.

It must be remembered that this project is limited by my understanding of sharing bike system in Toronto and the evaluation of the data. Further studies can be carried out to understand the real-time sharing bike system, the best location of stations, and the allocation of bikes.

Reference

```
@article{zhou2015understanding,
  title={Understanding spatiotemporal patterns of biking behavior by analyzing massive bike sharing data in Chicago},
  author={Zhou, Xiaolu},
  journal={PloS one},
  volume={10},
  number={10},
  pages={e0137922},
  year={2015},
  publisher={Public Library of Science}
}

@inproceedings{froehlich2009sensing,
  title={Sensing and predicting the pulse of the city through shared bicycling.},
  author={Froehlich, Jon and Neumann, Joachim and Oliver, Nuria and others},
  booktitle={IJCAI},
  volume={9},
  pages={1420--1426},
  year={2009}
}
```

Appendix

How I gather this dataset?

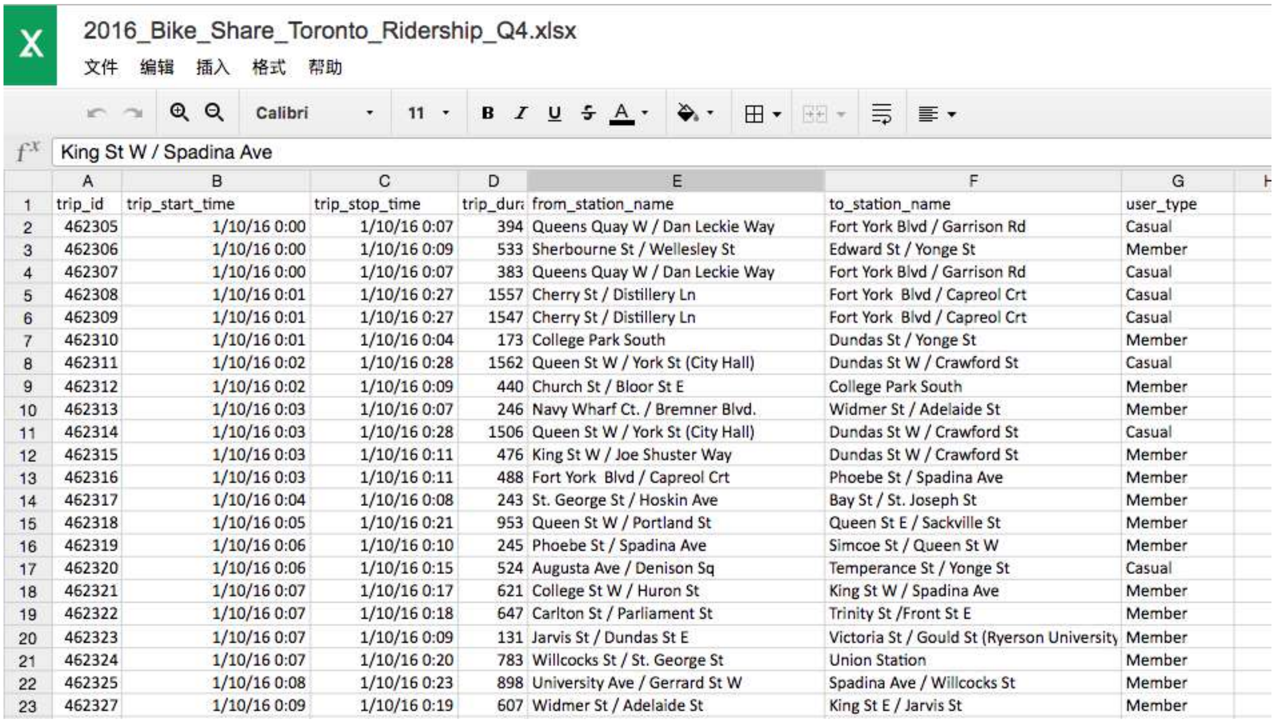
There are two major resources for this dataset.

One is from Bike Share Toronto Rideship Data (https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/#343faeaa-c920-57d6-6a75-969181b6cbde). All the trip in this dataset is anonymized. Due to my computational limit, I only select 2016 Q4 (https://www.toronto.ca/ext/open_data/catalog/data_set_files/2016_Bike_Share_Toronto_Ridership_Q4.xlsx) to study. This dataset contains trip data, including:

- Trip start day and time
- Trip end day and time
- Trip duration

- Trip start station
- Trip end station
- User type

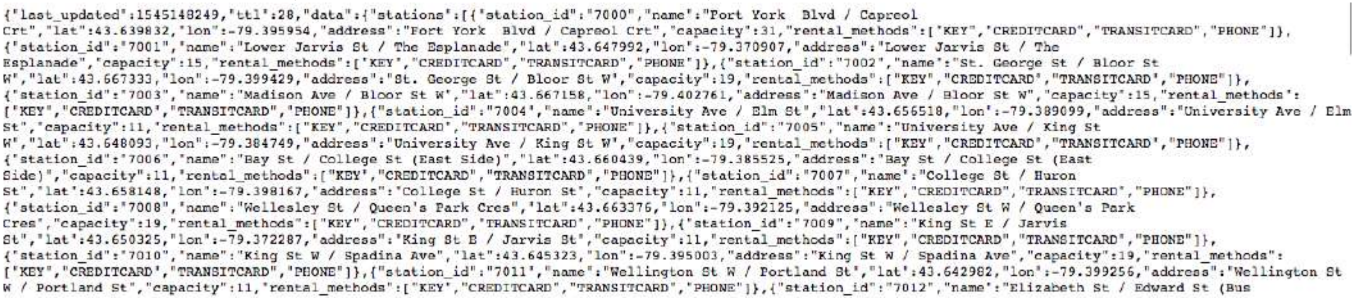
I download the dataset and use Excel to modify the date format.



	A	B	C	D	E	F	G
1	trip_id	trip_start_time	trip_stop_time	trip_dura	from_station_name	to_station_name	user_type
2	462305	1/10/16 0:00	1/10/16 0:07	394	Queens Quay W / Dan Leckie Way	Fort York Blvd / Garrison Rd	Casual
3	462306	1/10/16 0:00	1/10/16 0:09	533	Sherbourne St / Wellesley St	Edward St / Yonge St	Member
4	462307	1/10/16 0:00	1/10/16 0:07	383	Queens Quay W / Dan Leckie Way	Fort York Blvd / Garrison Rd	Casual
5	462308	1/10/16 0:01	1/10/16 0:27	1557	Cherry St / Distillery Ln	Fort York Blvd / Capreol Crt	Casual
6	462309	1/10/16 0:01	1/10/16 0:27	1547	Cherry St / Distillery Ln	Fort York Blvd / Capreol Crt	Casual
7	462310	1/10/16 0:01	1/10/16 0:04	173	College Park South	Dundas St / Yonge St	Member
8	462311	1/10/16 0:02	1/10/16 0:28	1562	Queen St W / York St (City Hall)	Dundas St W / Crawford St	Casual
9	462312	1/10/16 0:02	1/10/16 0:09	440	Church St / Bloor St E	College Park South	Member
10	462313	1/10/16 0:03	1/10/16 0:07	246	Navy Wharf Ct. / Bremner Blvd.	Widmer St / Adelaide St	Member
11	462314	1/10/16 0:03	1/10/16 0:28	1506	Queen St W / York St (City Hall)	Dundas St W / Crawford St	Casual
12	462315	1/10/16 0:03	1/10/16 0:11	476	King St W / Joe Shuster Way	Dundas St W / Crawford St	Member
13	462316	1/10/16 0:03	1/10/16 0:11	488	Fort York Blvd / Capreol Crt	Phoebe St / Spadina Ave	Member
14	462317	1/10/16 0:04	1/10/16 0:08	243	St. George St / Hoskin Ave	Bay St / St. Joseph St	Member
15	462318	1/10/16 0:05	1/10/16 0:21	953	Queen St W / Portland St	Queen St E / Sackville St	Member
16	462319	1/10/16 0:06	1/10/16 0:10	245	Phoebe St / Spadina Ave	Simcoe St / Queen St W	Member
17	462320	1/10/16 0:06	1/10/16 0:15	524	Augusta Ave / Denison Sq	Temperance St / Yonge St	Casual
18	462321	1/10/16 0:07	1/10/16 0:17	621	College St W / Huron St	King St W / Spadina Ave	Member
19	462322	1/10/16 0:07	1/10/16 0:18	647	Carlton St / Parliament St	Trinity St /Front St E	Member
20	462323	1/10/16 0:07	1/10/16 0:09	131	Jarvis St / Dundas St E	Victoria St / Gould St (Ryerson University	Member
21	462324	1/10/16 0:07	1/10/16 0:20	783	Willcocks St / St. George St	Union Station	Member
22	462325	1/10/16 0:08	1/10/16 0:23	898	University Ave / Gerrard St W	Spadina Ave / Willcocks St	Member
23	462327	1/10/16 0:09	1/10/16 0:19	607	Widmer St / Adelaide St	King St E / Jarvis St	Member

screen shot of 2016 Q4 sharing bike trip

The information of station coordinates is found from publicbikesystem (https://tor.publicbikesystem.net/ube/gbfs/v1/en/station_information). I transform the .json file into .csv, and use Excel-vlookup to join the station location file with the trip file with key word station name.



```
{"last_updated":1545148249,"ttl":28,"data":{"stations":[{"station_id":"7000","name":"Fort York  Blvd / Capreol
Crt","lat":43.639832,"lon":-79.395954,"address":"Fort York  Blvd / Capreol Crt","capacity":31,"rental_methods":["KEY","CREDITCARD","TRANSITCARD","PHONE"]},
{"station_id":"7001","name":"Lower Jarvis St / The Esplanade","lat":43.647992,"lon":-79.370907,"address":"Lower Jarvis St / The
Esplanade","capacity":15,"rental_methods":["KEY","CREDITCARD","TRANSITCARD","PHONE"]},{"station_id":"7002","name":"St. George St / Bloor St
W","lat":43.667333,"lon":-79.399429,"address":"St. George St / Bloor St W","capacity":19,"rental_methods":["KEY","CREDITCARD","TRANSITCARD","PHONE"]},
{"station_id":"7003","name":"Madison Ave / Bloor St W","lat":43.667158,"lon":-79.402761,"address":"Madison Ave / Bloor St W","capacity":15,"rental_methods":
["KEY","CREDITCARD","TRANSITCARD","PHONE"]},{"station_id":"7004","name":"University Ave / Elm St","lat":43.656518,"lon":-79.389099,"address":"University Ave / Elm
St","capacity":11,"rental_methods":["KEY","CREDITCARD","TRANSITCARD","PHONE"]},{"station_id":"7005","name":"University Ave / King St
W","lat":43.648093,"lon":-79.384749,"address":"University Ave / King St W","capacity":19,"rental_methods":["KEY","CREDITCARD","TRANSITCARD","PHONE"]},
{"station_id":"7006","name":"Bay St / College St (East Side)","lat":43.660439,"lon":-79.385525,"address":"Bay St / College St (East
Side)","capacity":11,"rental_methods":["KEY","CREDITCARD","TRANSITCARD","PHONE"]},{"station_id":"7007","name":"College St / Huron
St","lat":43.658148,"lon":-79.398167,"address":"College St / Huron St","capacity":11,"rental_methods":["KEY","CREDITCARD","TRANSITCARD","PHONE"]},
{"station_id":"7008","name":"Wellesley St / Queen's Park Cres","lat":43.663376,"lon":-79.392125,"address":"Wellesley St W / Queen's Park
Cres","capacity":19,"rental_methods":["KEY","CREDITCARD","TRANSITCARD","PHONE"]},{"station_id":"7009","name":"King St E / Jarvis
St","lat":43.650325,"lon":-79.372287,"address":"King St E / Jarvis St","capacity":11,"rental_methods":["KEY","CREDITCARD","TRANSITCARD","PHONE"]},
{"station_id":"7010","name":"King St W / Spadina Ave","lat":43.645323,"lon":-79.395003,"address":"King St W / Spadina Ave","capacity":19,"rental_methods":
["KEY","CREDITCARD","TRANSITCARD","PHONE"]},{"station_id":"7011","name":"Wellington St W / Portland St","lat":43.642982,"lon":-79.399256,"address":"Wellington St
W / Portland St","capacity":11,"rental_methods":["KEY","CREDITCARD","TRANSITCARD","PHONE"]},{"station_id":"7012","name":"Elizabeth St / Edward St (Bus
```

Because of the development of sharing bike system in Toronto, location of some stations can't be found nowadays. Thus, I just abandon trips from or to these stations. The total number of ride records drops from 217569 to 142866.

I also calculate the trip distance according to the coordinates with `{r eval=F} distHaversine()` function in `{r eval=F} library(geosphere)` .

System Information

```
## R version 3.5.1 (2018-07-02)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS Sierra 10.12.6
## 
## Matrix products: default
## BLAS: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRlapack.dylib
## 
## locale:
## [1] C
## 
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base     
## 
## other attached packages:
##  [1] ape_5.2                  factoextra_1.0.5        
##  [3] proxy_0.4-22             leaflet.minicharts_0.5.4
##  [5] leaflet.extras_1.0.0     bindrcpp_0.2.2          
##  [7] rgdal_1.3-6              sp_1.3-1                
##  [9] geosphere_1.5-7          ggplot2_3.1.0           
## [11] reshape2_1.4.3           dplyr_0.7.8             
## [13] leaflet_2.0.2            lubridate_1.7.4         
## 
## loaded via a namespace (and not attached):
##  [1] tidyselect_0.2.5  purrr_0.2.5       lattice_0.20-35  
##  [4] colorspace_1.3-2  htmltools_0.3.6   viridisLite_0.3.0
##  [7] yaml_2.2.0        rlang_0.3.0.1     ggpubr_0.2       
## [10] later_0.7.5       pillar_1.3.0      glue_1.3.0       
## [13] withr_2.1.2       bindr_0.1.1       plyr_1.8.4       
## [16] stringr_1.3.1     munsell_0.5.0     gtable_0.2.0     
## [19] htmlwidgets_1.3   codetools_0.2-15  evaluate_0.12    
## [22] labeling_0.3      knitr_1.20        httpuv_1.4.5     
## [25] crosstalk_1.0.0   parallel_3.5.1    highr_0.7        
## [28] Rcpp_1.0.0        xtable_1.8-3      promises_1.0.1   
## [31] scales_1.0.0      jsonlite_1.6      mime_0.6         
## [34] gridExtra_2.3     digest_0.6.18     stringi_1.2.4    
## [37] ggrepel_0.8.0     shiny_1.2.0       grid_3.5.1       
## [40] tools_3.5.1       magrittr_1.5      lazyeval_0.2.1   
## [43] tibble_1.4.2      cluster_2.0.7-1   crayon_1.3.4     
## [46] tidyr_0.8.2       pkgconfig_2.0.2   assertthat_0.2.0 
## [49] rmarkdown_1.11    viridis_0.5.1     R6_2.3.0         
## [52] nlme_3.1-137      compiler_3.5.1
```