FIT5147 - Data Visualization and Exploration

Assignment: Data Exploration Project.

Subject: Exploring Bike Sharing in Austin, Texas, U.S.A from 2014 to 2017

MONASH University

Student name: Trung Kien Nguyen

Student ID: 29057957

Contents

List of Figures

1. Introduction

Bike share in the U.S has continued its brisk growth, with 35 million trips taken in 2017, 25% more than 2016, according to the report published by National Association of City Transportation Officials (2017). This growth is attributable to increasing ridership in existing systems as well as the launch of several major new bike share systems across the country.

This report is focusing on providing public bike sharing operators in Austin, the U.S. There are three datasets have been taken into consideration. The first dataset is bike share dataset which contains information on bike trips from 2013 to 2017, the second dataset is train station which is the place located the docks in the city. Both datasets are from Kaggle which was published by the City of Austin. The last dataset is also from Kaggle which contain weather information of each day in the City of Austin during the period of 3 years between 2013 and 2017. Three datasets will be combined in on single file in order to visualize.

2. Data Wrangling

This section is focusing on preparation for data visualization. Dataset will be checked and validated in term of value, format as well as imputing null data. Also, enriching attributes and selecting attributes are necessary to step to get the most relevant and important features for visualization. Data wrangling and data exploration in this report have been done by using R with a variety of libraries, including **dplyr**, **reshape2**, **tidyverse**, **ggplot2** and **maptools** along with **Tableau** which is a powerful software for visualization.

The given bike share datasets comprise of more than 649.000 rows, including 12 attributes of bike trip information. Information of datasets is following.

Datasets Name	Number of attributes	Number of rows
austin_bikeshare_stations	6	72
austin_bikeshare_trips	12	649.250
austin_weather	21	1320

Merging data is essential as all of three datasets need to be put in a single file. Therefore, the bike trip dataset will be merged with the train station dataset based on station id to get the latitude, longitude of the original station and destination station. The combination dataset will be merge to the weather data.

The main obstacle of the merging step is the huge number of data points and limited computer processing. Hence, in order to get through this difficulty, the bike trip dataset will be divided into two 7 subsets, each part encompasses of 100.000 rows except the last subset which has 49.250 rows. Each of these subsets will be separately merged to the station and weather data and then embrace 7 subsets into one single dataset.

The final dataset contains the data from Jun 2013 to Jun 2017. However, as the purpose of the report, the data of 2013 and 2017 will be eliminated out of the main dataset.

2.1.Fixing Null Value.

There are existing null data in the final dataset.

- Bikeid: 723 null rows
- End_station_id: 19842 null rows
- Month: 30752 null rows
- Start_station_id: 19041 null rows.
- Year: 30752 null rows

Bikeid will be eliminated as it is not affected by the visualization.

Month and Year will be imputed by using start time attribute as it contains day and time check-in of the trip.

To those end station id and start station id. Those value will be imputed by matching the station name in stations datasets to get the corresponding station id.

2.2.Feature Enriching

This part will create some new features from the attributes as it not only beneficial to visualize but also descriptive and inference statistics.

Following is new features which are created from given attributes.

Attribute	Attribute value	New Feature	New Value
start_time	2015-12-21 09:12:00	Date	2015-12-21
start_time	2015-12-21 09:12:00	Day	Mon
start_time	2015-12-21 09:12:00	Hour	09

Besides creating new features from given attributes, transform continues value to discrete categories could be advantageous in many ways. Following is new discrete categories which are created.

Attribute	Attribute value	New Feature	New Value
duration_minutes	58	duration_category	55-60
TempAvgF	78	tempavg_category	70-80

2.3.Feature Selection

Due to the merging of three huge data, the datasets comprise 37 attributes. Not all of these attributes will be put into visualization. Hence, selecting the main attributes are necessary. The following is a table of attributes are selected.

Name	Type	Value	Description
checkout_time	Factor	12:02:20	Checkout time
duration_minutes	Int	45	Self-explained
end_station_id	Num	2345	Self-explained
end_station_name	Factor	11th & San Jacinto	Self-explained
end_station_longitude	Num	-97.1231423	Self-explained
end_station_latitude	Num	30.02342342	Self-explained

start_station_id	Num	2345	Self-explained
start_station_name	Factor	11th & San Jacinto	Self-explained
start_station_longitude	Num	-97.1231423	Self-explained
start_station_latitude	Num	30.02342342	Self-explained
start_time	Factor	2014-01-01 00:12:00	Self-explained
subscriber_type	Factor	Walk-up	Self-explained
year	Num	2014	Self-explained
date	Factor	2016-12-12	Self-explained
TempAvgF	Int	88	Average temperature (F)
events	Factor	Thunder, Rain	Self-explained
day	Factor	Mon	Day of week
hour	Int	19	Check-in time hours
duration_category	Factor	55-60	Self-explained
tempavg_category	Factor	70-80	Self-explained

One of the main purposes of this section is to gain more understanding about the data as well as the domain of dataset, also, checking and validating potential error that could impact to the visualization.
After carefully checking the entry data, there is no error in term of format as well as the value of the data. Hence, the dataset can be put into the progress of exploration visualization.

3. Exploring and Visualization

3.1. Austin Bike Sharing Overview

The data is analyzed came from rides between Jan 2014 to Nov 2016.

In those 35 months, there were **550.225** rides, averaging about **20** minutes per trip.

Total riding time is **9.672.025 minutes**, or **161.200 hours** which equally to **1.089 weeks, 19 hours, and 20 seconds.**

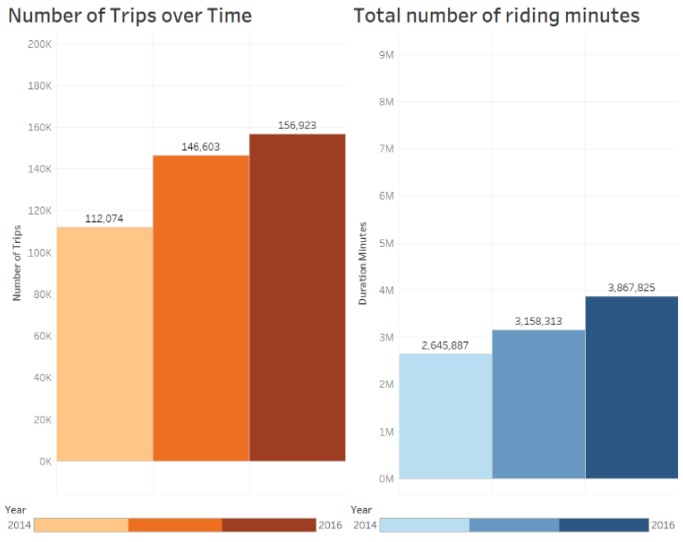

Figure 1: Number of trips over time and total number of riding minutes

3.2. Bikeshare Traffic System

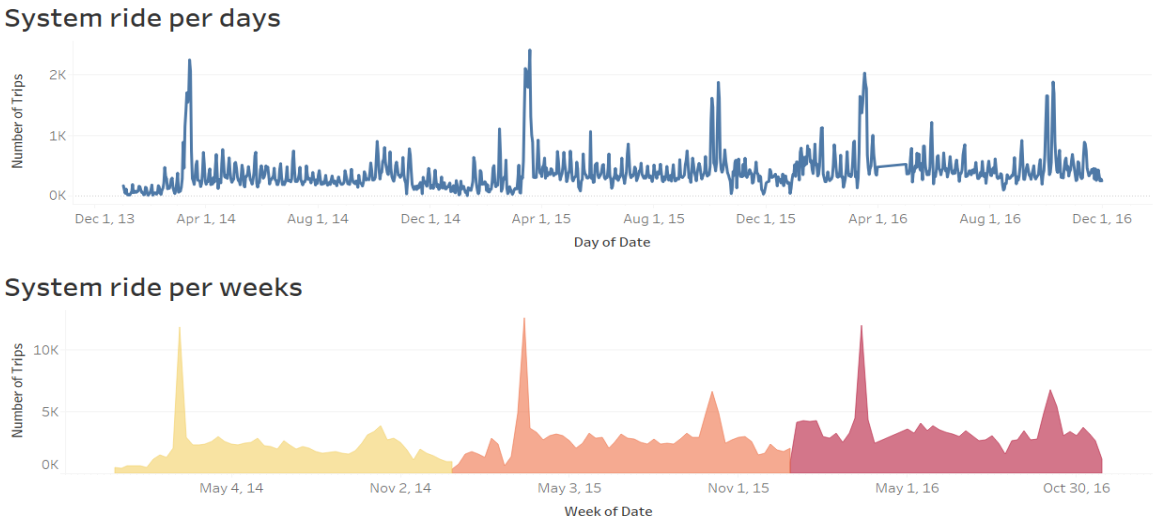

Figure 2: The usage of the system by days and by week over 3 years period

Over 3 years period, it is obvious people use the bike share the most in March, April, and October while the first month and the last month of those years witness less used.
The **business day:** 2015-03-19 with **2413** rides
The **calmest day:** 2015-01-10 with **10** rides

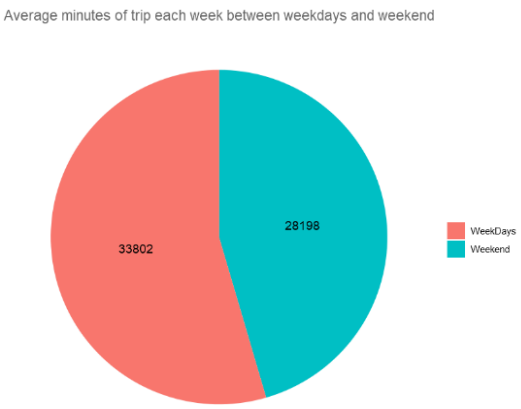

Figure 3: Average minutes of trip each between weekdays and weekends

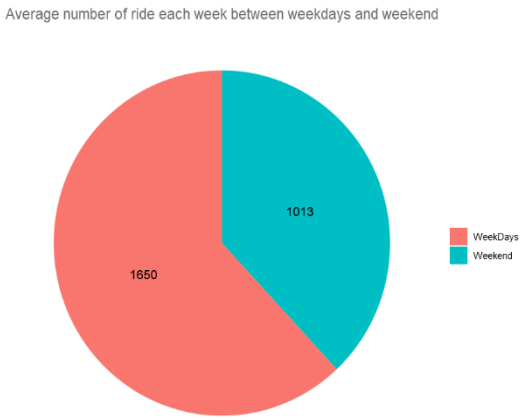

Figure 4: Average number of rides each week between weekdays and weekends

Also, look in the pie chart, it is clear that customers prefer to rides in the weekend more than weekdays and the long trip tend to happen in the weekend rather than during weekdays. That is the reason why the number of the trip in the weekend less than weekdays but the average trip duration is approximately $^{5}/_{6}$ of the average trip duration of weekdays

In term of trips traffic during hours of the day, there is a clear trend that riders start using bike share from the morning when they go out for work and continues to use the system more until 5 pm, then decrease significantly. The chart illustrates the system used per day of the month show how people are using the system during the month. As it depicts in the graph, the first 20 days of the month witnessed a large number of trips and then decreased steadily to the last day of the month since there

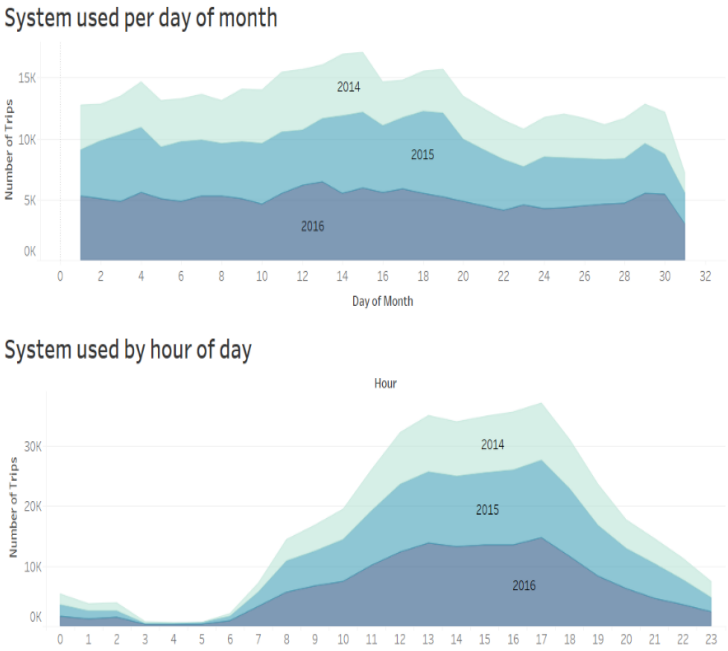

Figure 5: The used of system by days and hours

3.3. The Use of Customers

Understanding who is riding and how bike share is used can be difficult.
There is 62 types of subscribers in the system, but it seems like the top 3 subscribers dominate the most, which is Walk Up, Local365 and 24-hours-kiosk. All of the other types will be considered as the same subscriber.

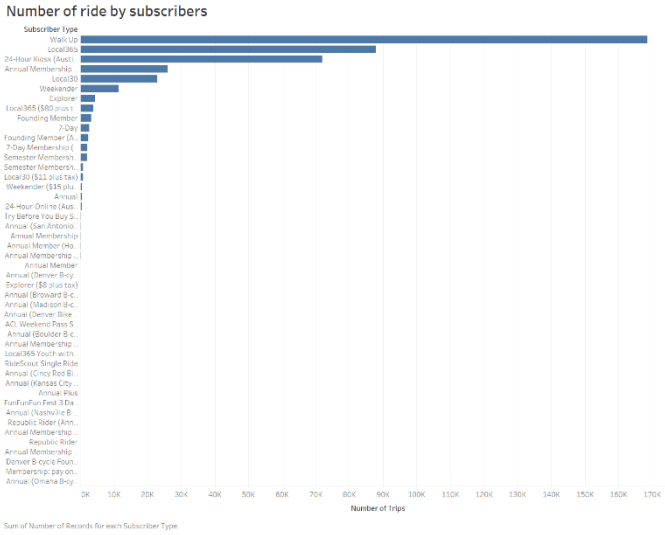

Figure 6: Number of Riders by subscribers

As it is illustrated in the fig, the number of Walk-Up group contributes the most to the uses of the system. However, an interesting about subscribers that back to 2014, there were no Walk Up and Local365 groups. However, since 2015, Austin bike share system cut down the 24-hours-kiosk ridership.

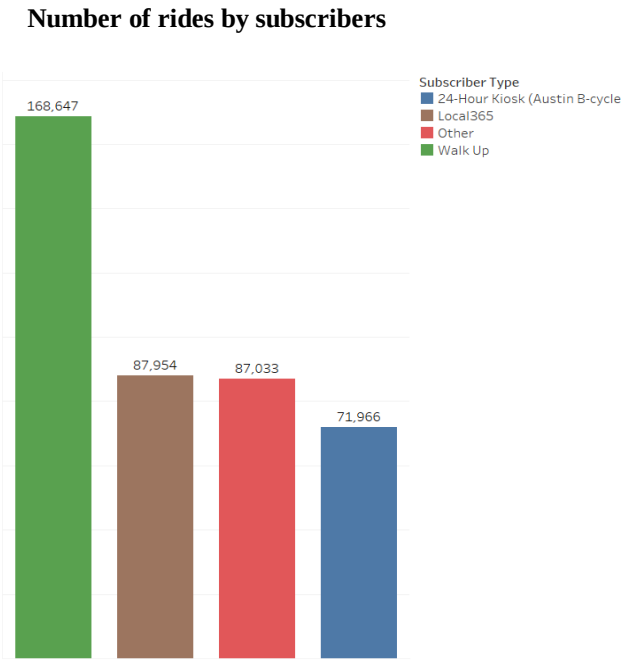

Figure 8: Number of rides by subscribers

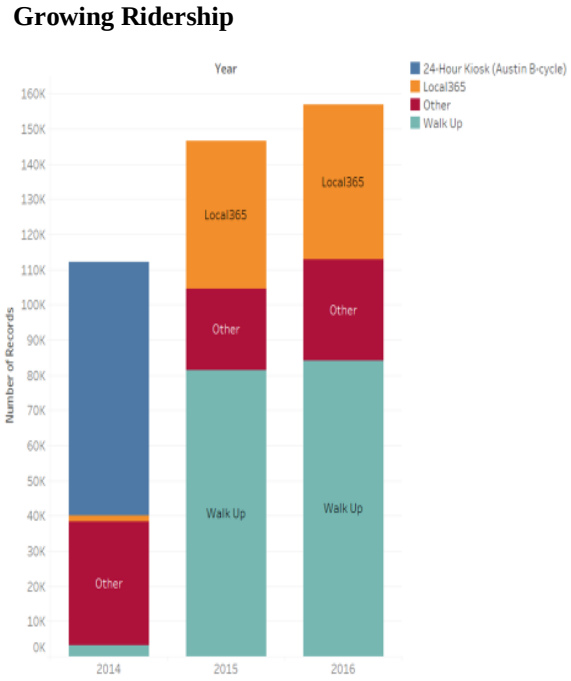

Figure 7: Growing Ridership

Apparently, Walk-Up is likely to be a commitment to ride, but how subscribers use the system? Look in figure 9, it appears that walk up riders are overwhelmingly other subscribers during the weekends, while there is a fluctuated of a number of rides of the subscribers during weekdays.

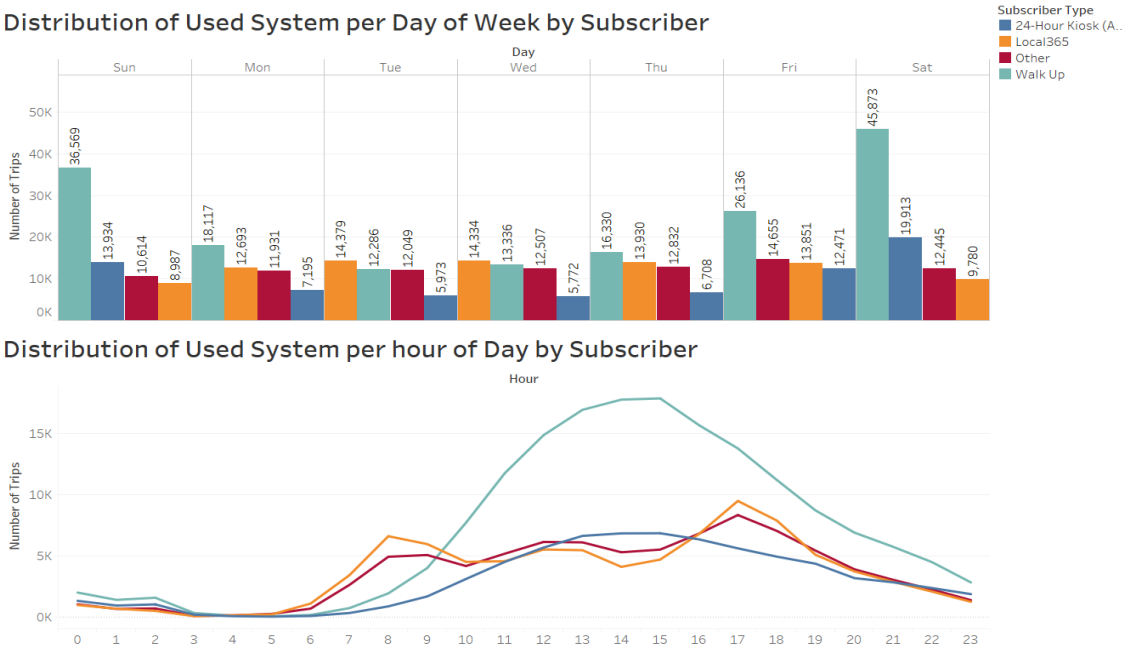

Figure 9: Distribution of used system of different measure by subscribers

Among subscribers that already spikes in use from 8 am to 5 pm with a bump from 1 pm noon to 15 pm in the afternoon. These users must be riding a bike to get to work, to go to lunch or head home. Customer hourly usage seems to fall along a bell-shaped distribution peaking at 14 pm to 15 pm, especially the customer signed up for 24 hours kiosk and walk up. On the other hand, "Local365" behave slightly different as it picks the top right in the 8 am in the morning and at 17 pm in the afternoon.

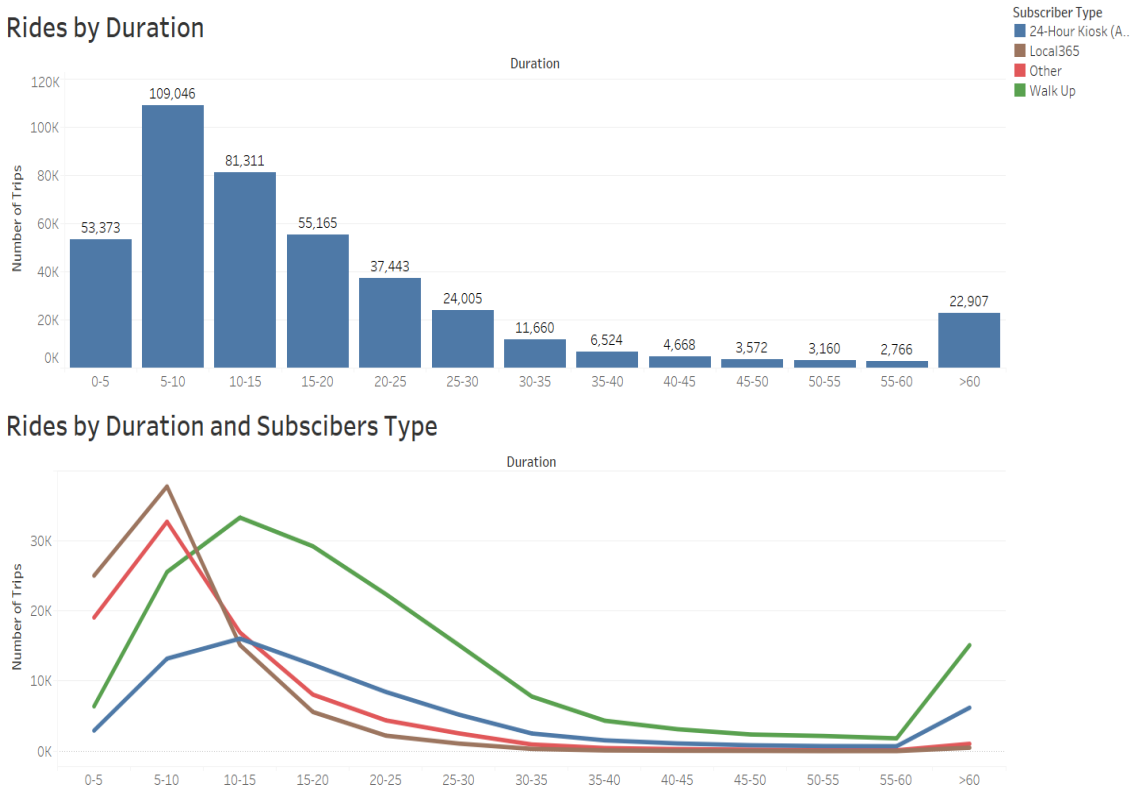

Figure 10: Distribution of rides by the duration of subscribers

The chart above also shows the number of rides by the duration of the trip and how users use the system in term of duration. The most common ride length is from 5 to 10 minutes, following by the length from 10-15 and 15-20. Interestingly, the long rides which are longer than 60 minutes are not rare, but in fact, there are quite a lot of users have a long ride. In the second chart illustrates the usage of different type of subscribers to the system. Turn out, Local365 is a group has the most ride from 5 to 10 minutes, even the number of rides is much less than the Walk Up group. Walk Up ridership, on the other hand, prefer to have a trip from 10 to 20 minutes, especially, this group is outstanding other groups because of customer's behaviors tend to have surprisingly long trips.

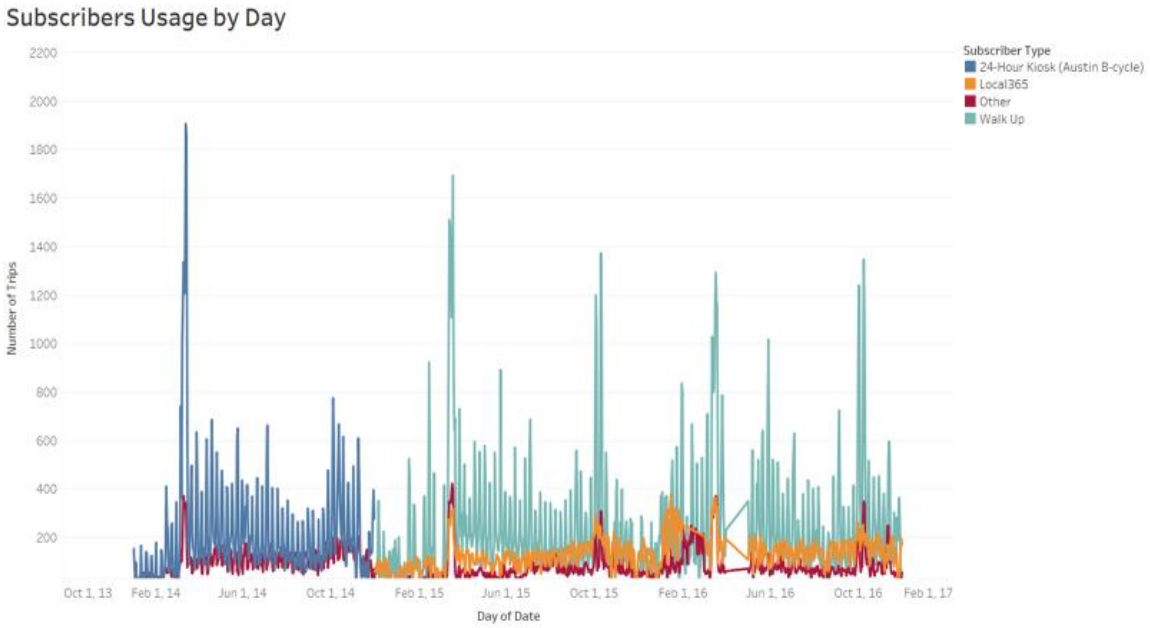

Figure 11: Traffic used by day from 2014 to 2017

It is also interesting to see how subscribers use the system by day. Turn out, 24-hour-kiosk only offer until Dec 2014, which mean there is no package 24-hour-kiosk for the customer to choose. Before 2015, there was no Local365 or Walk Up, these groups only available after 2015, however, none of them exceed the record that 24 hour-kiosk groups reach in March 2014. Again, ride by subscribers decreases significantly during Thanksgiving and between Christmas and New Year. Rides by customers increased noticeably on weekends

3.4.The Popular Original Station and Destination Station

Most Popular Starting Stations

Start Station Name	
5th & Bowie	19,942
City Hall / Lavaca & 2nd	19,592
4th & Congress	18,292
Convention Center / 4th St. @ MetroRail	17,421
2nd & Congress	17,309
Riverside @ S. Lamar	15,045
Davis at Rainey Street	14,732
Capitol Station / Congress & 11th	13,352
Rainey St @ Cummings	13,258
Pfluger Bridge @ W 2nd Street	12,845

Most Popular Ending Station

End Station Name	
City Hall / Lavaca & 2nd	23,452
4th & Congress	20,836
2nd & Congress	19,610
Convention Center / 4th St. @ MetroRail	19,520
5th & Bowie	18,548
Riverside @ S. Lamar	14,820
Davis at Rainey Street	14,257
Rainey St @ Cummings	12,714
Pfluger Bridge @ W 2nd Street	11,999
Barton Springs & Riverside	11,366

Figure 12: Most popular starting and ending stations

Here are the top 10 stations in the entire system to start or end a ride. Interestingly, while "5th & Bowie" hold the first position place where people start to ride but it was only in 5th position as ending stations. The highest number of destination station occupied by City Hall/ Lavaca & 2nd. Also, City Hall took the second place of top starting station with 19.592 rides, just about 400 behind 5th&Bowie with 19.942 rides. This poses an interesting question, do commuters use the Bike Share System to go to work the most if so, do company office normally located in City Hall?

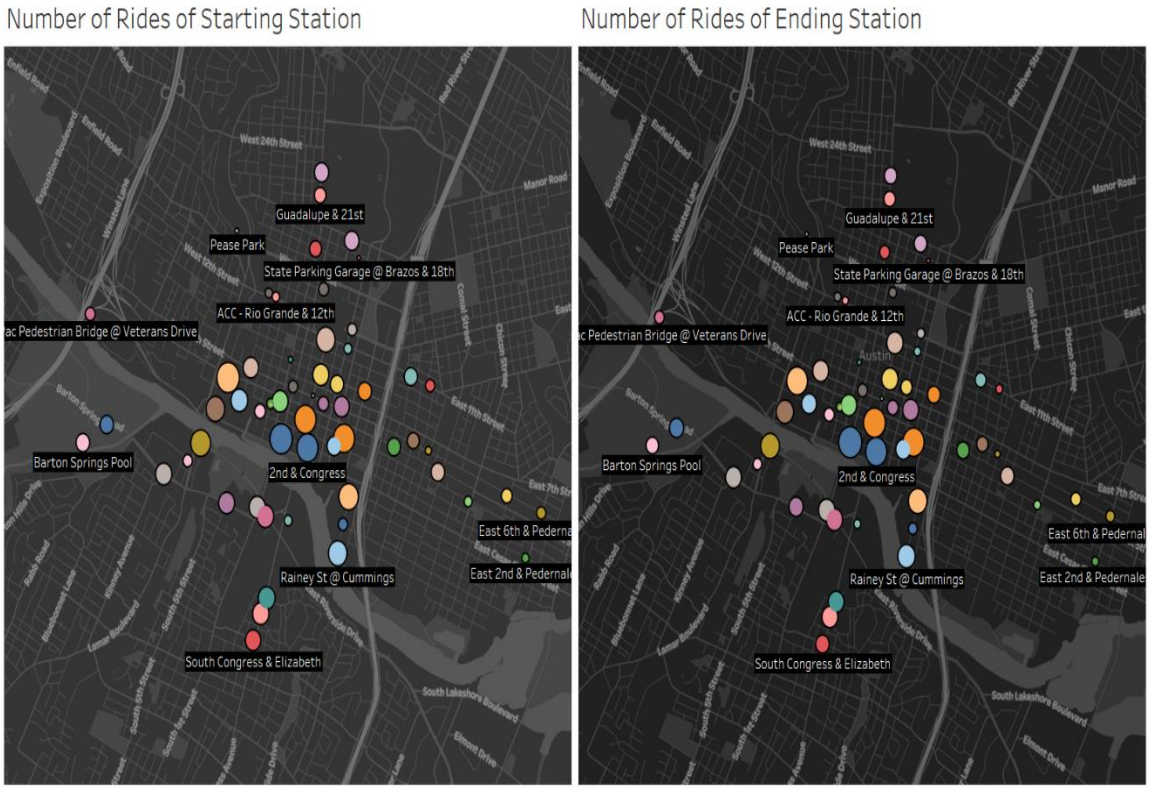

Figure 13:Number of rides by station on Map

The map above is an intuitive way to see the location and the number of rides of each station in the City of Austin. It seems like customer use bike to commute between place inner the city more than in the suburb. The biggest blue circle right in the center are City Hall station and 2nd&Congress station.

Ranking the popular starting and ending station

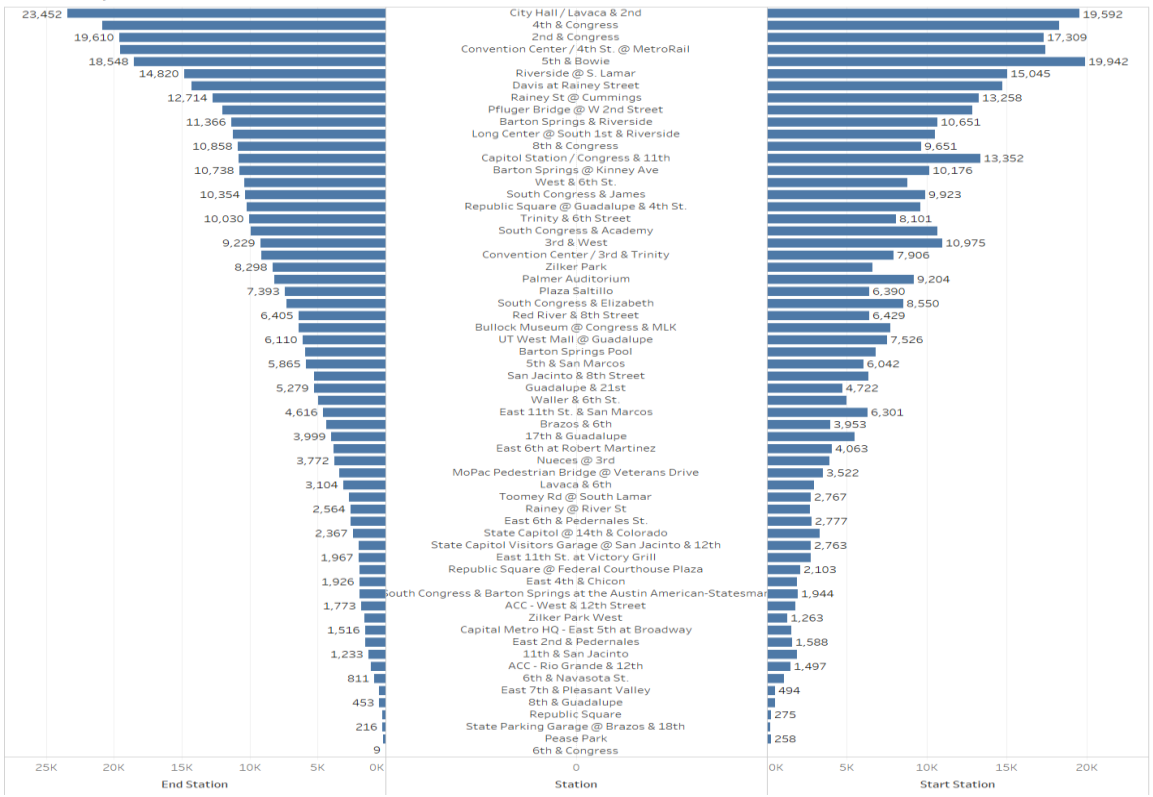

Figure 14:Ranking the popular starting and ending station

Most Traveled Routes

Starting Station Name	Ending Station Name	Number of Rides
5th & Bowie	4th & Congress	2,079
4th & Congress	5th & Bowie	1,896
5th & Bowie	City Hall / Lavaca & 2nd	1,835
Republic Square @ Guadalupe & 4th St.	5th & Bowie	1,664
3rd & West	City Hall / Lavaca & 2nd	1,503
City Hall / Lavaca & 2nd	5th & Bowie	1,471
Convention Center / 4th St MetroRail	City Hall / Lavaca & 2nd	1,466
South Congress & Academy	2nd & Congress	1,434
5th & Bowie	Convention Center / 4th St.	1,377
Rainey St @ Cummings	City Hall / Lavaca & 2nd	1,306

3.5.The Impact of Weather Conditions

The weather could be the main factor that has influences on riders' decision of using bike share, this section will examine the correlation between weather and the number of rides. Over a vast amount number attributes, only average temperature (Fahrenheit measure) of the day and event of the day such as Rain, Thunder, Frog… are selected.

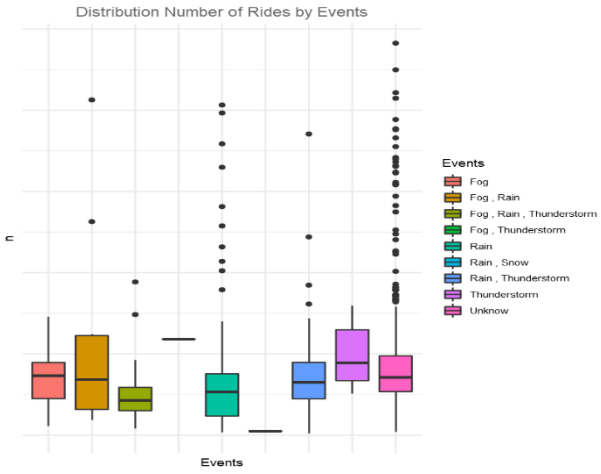

Figure 15:Distribution number of rides by Events

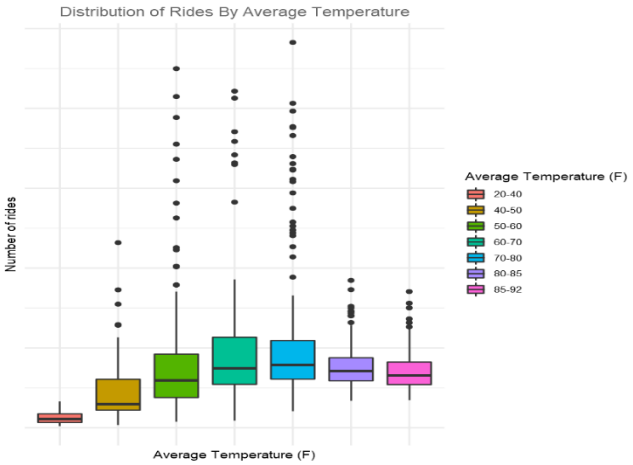

Figure 16:: Distribution of riders by average temperature

Apparently, the average temperature is quite significantly affected to customers used as riders prefer to ride when the average temperature is around 50F-80F. It is plausible as riders do not use that much of bike share system frequently if the temperature is too low (less than 40F) or two high (greater than 85F). In term of weather condition, it seems like customers still ride no matter how the weather condition. Interestingly, even in bad condition such as Thunderstorm or Rain or heavy Snow, the number of rides even more than normal. It is reasonable as customers normally use bike share for a short trip from 5 to 20 minutes.

The heatmap illustrates the phenomenon that stated weather condition does not affect much to the traffic of the system. Most of the rides witness in duration from 5 to 20 minutes no matter weather conditions. Another interesting bump up that there were no trips that have durations longer than 25 minutes if there were rain and snow at the same time.

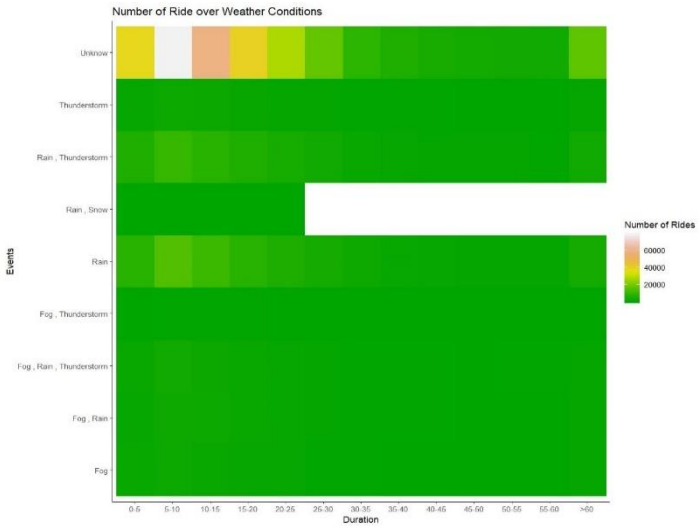

Figure 17:Number of rides in term of duration over weather conditions by heatmap representation

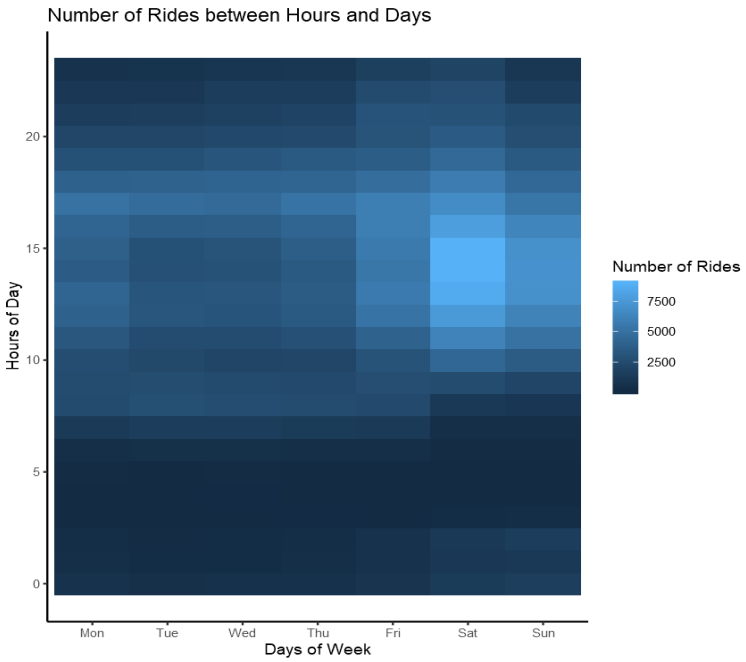

Figure 18:Number of rides between hours and days

Also, it is motivating to see how riders use the system during the hours of the day of the week.

Obviously, the number of rides increased significantly during the weekends. The busiest time of the day was from 8 am to 5 pm. The number of rides decreased dramatically after 5 pm except for Friday and Saturday. It is surprising that the number of rides during the night on Friday and Saturday are more than the other day. Perhaps, the customer used bike to go out late on Friday and Saturday as they do not have to go to work on the following day. Sunday on the other hand, the number of rides after 5 pm was the same with weekdays (except Friday). It is reasonable as customers have to go to work on Monday.

Finally, seeing the correlation between average temperature and the number of rides is necessary to test the assumption of temperature effects riders.

Look in the graph of figure 19, it is quite nice as the fitness line demonstrates the perfect relationship of those values.

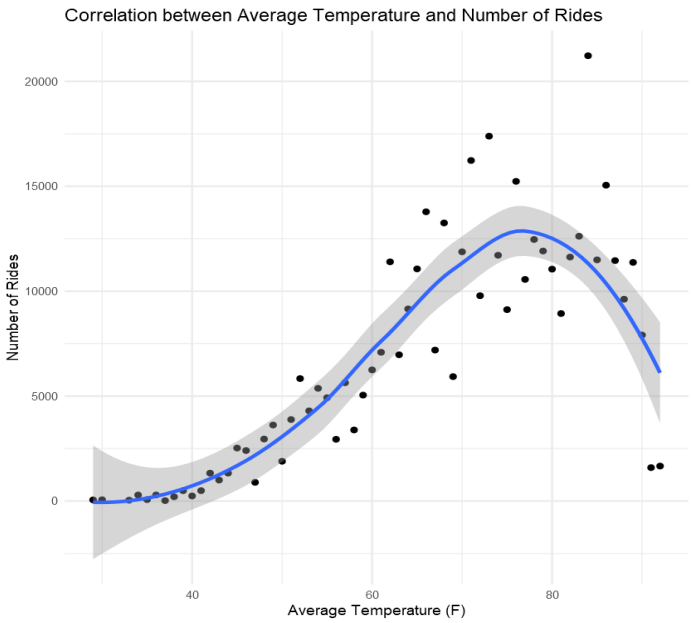

Figure 19:Correlation between average temperature and number of rides

4. Conclusion

The data of bike share system in the City of Austin has been collected from 2014 to 2016 come along with weather data from the same period. Overall, the number of riders increase year by year and it seems like Walks Up ridership and Local365 ridership was dominant of the system from 2015 to end of 2016 while 24-hour-kiosk was the most popular back to 2014. Also, it looks like weather condition does not affect much to customer's decision to ride, however, the average temperature has an influence on riders as they do not prefer to ride when the temperature is too low as well as too high.

Among of various reasons to ride, it seems like commuters used bike share system for going to work, heading home or having lunch as the traffic is quite high during the working hour from 8 am to 5 pm and normally reach the top around 1 pm to 5 pm. Also, the used of weekends were overwhelmingly the weekdays.

One dimension not explored in this analysis was the traffic of each starting station and ending station. We saw the most popular routes, however, it does not illustrate the subscriber's type factor as we would know what is the popular routes by subscribers.

One recommendation that can be made is for station growth in City Hall. Users are already shown to be commuters heading to work, and City Hall is high density of business.

5. Reflection

After finishing this assignment, I kept asking to myself if Melbourne bike share system has good data to analyze as I am living and studying in Melbourne. Unfortunately, turn out the City of Melbourne does not have that good dataset. The data is published just comprise the number of docks in the station across the city which does not provide much information in order to analyze.

Which shows that the number of customers used bike share in the City of Austin, I expect the same phenomenon witness in Melbourne as well. As air pollution, global warming is the red alert to the world, if people use more public transport or bike or electric vehicle, the burden in the transportation system will be down and reduce contamination as well.

References:

National Association of City Transportation Officials (NACTO), 2017. NACTO's new reports, Shared Micromobility in the U.S.:2018. Retrieved from https://nacto.org/shared-micromobility-2018/