Bixi Project

By Aline Santoso

1 Usage Volume Overview

First, we will attempt to gain an overall view of the volume of usage of Bixi Bikes and what factors influence it. To do so calculate:

1. The total number of trips for the years of 2016.

   3917401

2. The total number of trips for the years of 2017.

   4666765

3. The total number of trips for the years of 2016 broken-down by month.

Month	NumberOfTrips
4	189923
5	561077
6	631503
7	699248
8	672778
9	620263
10	392480
11	150129

4. The total number of trips for the years of 2017 broken-down by month.

Month	NumberOfTrips
4	195662
5	587447
6	741835
7	860732
8	839938
9	731851
10	559506
11	149794

5. The average number of trips a day for each year-month combination in the dataset.

Year	Month	AverageTripDaily
2016	4	11870.1875

2016	5	18099.2581
2016	6	21050.1
2016	7	22556.3871
2016	8	21702.5161
2016	9	20675.4333
2016	10	12660.6452
2016	11	10008.6
2017	4	12228.875
2017	5	18949.9032
2017	6	24727.8333
2017	7	27765.5484
2017	8	27094.7742
2017	9	24395.0333
2017	10	18048.5806
2017	11	9986.2667

Unsurprisingly, the number of trips varies greatly throughout the year. How about membership status? Should we expect member and non-member to behave differently? To start investigating that, calculate:

1. The total number of trips in the year 2017 broken-down by membership status (member/non-member).

Membership	NumberOfTrips
Member	3784682
Non-Member	882083

2. The fraction of total trips that were done by members for the year of 2017 broken-down by month.

   I feel that the fraction here can be interpreted in two different ways. I discussed this with Boris and he said it's the first one. For my own notes, I still put in the code and the result of the second case.

   First, it's the fraction of total trips done by member for all months, i.e. the total fraction of trips done by members and non-members for ALL months is 1, which is broken down by month in the table below:

Month	MemberFraction
4	0.035
5	0.1032
6	0.1285

7	0.141
8	0.1406
9	0.1295
10	0.1036
11	0.0297

Second, it's the fraction of total trips done by member for each month i.e. the total fraction of trips done by members and non-members for EACH month is 1, which is broken down by month in the table below:

Month	MemberFraction
4	0.8352
5	0.8197
6	0.8081
7	0.7643
8	0.7811
9	0.8258
10	0.8641
11	0.9246

Use the above queries to answer the questions:

1. Which time of the year the demand for Bixi bikes is at its peak?

   Using the query results from total number of trips in 2016 and 2017, as well as the daily average of trips broken down by year-month, we can see that **July** has always had the highest number of trips and daily average of trips.

   We can also run a query to find a maximum average trip daily by year:

Yr	Mth	Maximum
2016	7	22556.387
2017	7	27765.548

   As can be seen, the demand for Bixi bikes is at its peak on the month of July for both 2016 and 2017.

2. If you were to offer non-members a special promotion in an attempt to convert them to members, when would you do it?

   The best time to offer a non-member a special promotion is the month at which the fraction for member is the lowest, which is from the query result above is the month of **November.**

Month	MemberFraction
11	0.0297

2 Trip Characteristics

Given what we just learned about trip volume it seems the usage pattern of Bixi bikes in warmer and colder months is quite different. Let's take a closer look at the characteristics of trips and see what else we can uncover.

1. Calculate the average trip time across the entire dataset.

   AverageTripTime
   824.4291

2. Let's dig a bit deeper and slice the average trip time across a couple of interesting dimensions. Calculate the average trip time broken-down by:
   1. Membership status

Non-Member	1221.2917
Member	731.7721

   2. Month

Year	Month	AverageTripTime
2016	4	798.2132
2016	5	850.0497
2016	6	847.269
2016	7	886.6757
2016	8	868.6836
2016	9	814.9007
2016	10	729.368
2016	11	675.6693
2017	4	805.1808
2017	5	828.8954
2017	6	842.0157
2017	7	874.1786
2017	8	845.4499
2017	9	793.0245
2017	10	731.1515
2017	11	632.338

   3. Day of the week

DayOfWeek	DayName	AverageTripTime
1	Sunday	914.1739
2	Monday	798.6486

3	Tuesday	794.618
4	Wednesday	792.4604
5	Thursday	790.7546
6	Friday	798.8752
7	Saturday	908.984

4. Station name

   To break it down by station name, we first need to join the *station* table with the *trips* table. However, the *station* table has 1 column with the station code in it while the *trips* table has 2 columns with the station code: the start_station_code and the end_station_code. So, I JOIN the *station.code* with *trips.start_station_code* and JOIN the *station.code* with *trips.end_start_code*. I then use UNION to join these two. Limiting the result to 10.

code	name	AverageTripTime
5002	St-Charles / Montarville	1088.9547
5003	Place Longueuil	1117.3171
5004	St-Charles / Charlotte	958.8347
5005	St-Charles / St-Sylvestre	992.902
5006	Collège Édouard-Montpetit	1505.1572
5007	Métro Longueuil - Université de Sherbrooke	1304.7749
6001	Hôtel-de-Ville 2 (du Champs-de-Mars / Gosford)	962.7091
6002	Ste-Catherine / Dézéry	885.9231
6003	Clark / Evans	718.0312
6004	Hôtel-de-Ville (du Champs-de-Mars / Gosford)	1049.7147

- Which station has the longest trips on average?

  Using a query to organize the result in a descending order by the duration, **Métro Jean-Drapeau** has the longest trips on average.

code	name	AverageTripTime
6501	Métro Jean-Drapeau	1889.95

- Which station has the shortest trips on average?

  Similarly, using a query to organize the result in the ascending order by the duration, **Métro Georges-Vanier (St-Antoine / Canning)** has the shortest trips on average.

code	name	AverageTripTime
6408	Métro Georges-Vanier (St-Antoine / Canning)	514.1979

- Extremely long / short trips can skew your results. How would avoid that?

Using the average and the standard deviation, I would exclude the outliers from the results and just include those with 95% confidence interval i.e. data within 2 standard deviations from the mean (code written in the query).
However, from the basic descriptive statistic:

Average	SD	Maximum	Minimum
824.43	652.8	7199	61

The result seems to indicate a skewed result, probably right-skewed since the average is closer to the minimum than the maximum.

3. Let's call trips that start and end in the same station "round trips". Calculate the fraction of trips that were round trips and break it down by:
    1. Membership status

Membership	RTFraction
Member	0.0115
Non-Member	0.0092

    2. Day of the week

DayOfWeek	DayName	RT_Fraction
1	Sunday	0.0042
2	Monday	0.0026
3	Tuesday	0.0025
4	Wednesday	0.0026
5	Thursday	0.0025
6	Friday	0.0027
7	Saturday	0.0038

4. Discuss the differences you observed and come up with possible explanations.
From the previous two query results, we can conclude that round trips are done more by members than non-members and that there are more round trips on weekends than on weekdays.

This result makes sense since members are more likely to use the Bixi bikes to run quick errands around their home or work etc. which would start and end at the same station. Non-members such as tourists, for example, would usually begin at one station, to one point of interest and another i.e. another different station. Since errands are usually done on weekends, the increase in round trips on weekends can be explained to this as well. Furthermore, non-members are more likely to use the bikes on weekends too, for a quick leisurely round trip or for running errands, hence the increase of round trips on weekends. For example, comparing Montreal to

Vancouver, members and non-members are more likely to use bike-sharing around Stanley Park on weekends, which could be done as a round-trip.

3 Popular Stations

It is clear now that average temperature, weekends and membership status are intertwined and influence greatly how people use Bixi bikes. Let's try to bring this knowledge with us and learn something about station popularity.

1. What are the names of the 5 most popular starting stations?
   Most popular starting station means the starting station with the most trips.

Station Name	start_station_code	NumberOfTrips
Mackay / de Maisonneuve	6100	97150
Métro Mont-Royal (Rivard / du Mont-Royal)	6184	81279
Métro Place-des-Arts (de Maisonneuve / de Bleury)	6078	78848
Métro Laurier (Rivard / Laurier)	6136	76813
Métro Peel (de Maisonneuve / Stanley)	6064	72298

2. What are the names of the 5 most popular ending stations?
   Most popular ending station means the ending station with the most trips.

StationName	end_station_code	NumberOfTrips
Berri / de Maisonneuve	6015	103720
Mackay / de Maisonneuve	6100	99128
Métro Place-des-Arts (de Maisonneuve / de Bleury)	6078	95343
Métro St-Laurent (de Maisonneuve / St-Laurent)	6012	86886
Métro Peel (de Maisonneuve / Stanley)	6064	76551

3. If we break-up the hours of the day as follows:

```SQL
CASE
    WHEN HOUR(start_date) BETWEEN 7 AND 11 THEN "morning"
    WHEN HOUR(start_date) BETWEEN 12 AND 16 THEN "afternoon"
    WHEN HOUR(start_date) BETWEEN 17 AND 21 THEN "evening"
    ELSE "night"
END AS "time_of_day"
```

   1. How is the number of starts and ends distributed for the station *Mackay / de Maisonneuve* throughout the day?
      Let's find out the station code of the Mackay / de Maisonneuve station first.
      code name
      6100 Mackay / de Maisonneuve

Now, we can find out the number of starts throughout the day at Mackay / de Maisonneuve station:

time_of_day	NumberOfStarts
night	12267
morning	17384
afternoon	30718
evening	36781

And the number of ends at Mackay / de Maisonneuve station:

time_of_day	NumberOfEnds
night	9949
morning	27351
afternoon	30817
evening	31011

2. Explain the differences you see and discuss why the numbers are the way they are.
   From the previous question, we can see that the distribution for the number of ends and starts at the station is similar, in that it's lowest at night and then increases in the morning followed by afternoon and then highest in the evening.

   From previous queries, we see that Mackay / de Maisonneuve is the most popular starting station and the second most popular ending station, this explains why the afternoon numbers are similar between 'starts' and 'ends'. If Mackay / de Maisonneuve is in a central area, hours between 12:00 to 16:00 are when people, whether locals or tourist, members or non-members are most active. In the morning, the number of ends is significantly higher than the number of starts, which might indicate that Mackay / de Maisonneuve is in the business/working area where people would end up when they commuted to work. This is also supported by higher number of starts in the evening and at night, since people would be leaving from work around these times.

4. Which station has proportionally the least number of member trips? How about the most? To damper variance, consider only stations for which there were at least 10 trips starting and ending from it.
   To find out which station has the least number of member trips, we have to break down the number of trips by its membership. Since I made a new table earlier on that combined the stations based on their codes, I'm going to use that table for this query.

The stations that have the proportionally **least** number of member trips are these three below: CHSLD Benjamin-Victor-Rousselot (Dickson / Sherbrooke), Place Longueuil and CHSLD Éloria-Lepage (de la Pépinière / de Marseille) as shown below:

code	StationName	Member Fraction	TotalStart	TotalEnd
7009	CHSLD Benjamin-Victor-Rousselot (Dickson / Sherbrooke)	0.0001	954	954
5003	Place Longueuil	0.0001	953	953
7075	CHSLD Éloria-Lepage (de la Pépinière / de Marseille)	0.0001	933	933

The stations that have the proportionally **most** number of member trips is Mackay / de Maisonneuve.

code	StationName	MemberFraction	TotalStart	TotalEnd
6100	Mackay / de Maisonneuve	0.0188	160991	160991

5. List all stations for which at least 10% of trips are round trips. Recall round trips are those that start and end in the same station. This time we will only consider stations with at least 50 starting trips.

The step by step queries below are all in the .sql file. The final result of this query is:

code	StationName	NumberStartTrips	RT_Start_Fraction
6501	Métro Jean-Drapeau	28672	0.302
7048	Métro Angrignon	2398	0.2331
6428	Berlioz / de l'Île des Soeurs	5246	0.2043
7015	LaSalle / 4e avenue	2991	0.2006
6736	Basile-Routhier / Gouin	1708	0.1932
6359	Parc Plage	6201	0.1846
7007	Gare Canora	2439	0.1792
6714	LaSalle / Sénécal	3151	0.1473
6502	Casino de Montréal	6138	0.1437
6109	Quai de la navette fluviale	6417	0.1376
7075	CHSLD Éloria-Lepage (de la Pépinière / de Marseille)	475	0.1263
6026	de la Commune / Place Jacques-Cartier	50822	0.1106
6016	Jacques-Le Ber / de la Pointe Nord	2719	0.1103
6429	Place du Commerce	8569	0.1082

5006	Collège Édouard-Montpetit	1439	0.1001

1. First, write a query that counts the number of starting trips per station.
   Here are the first 5:

code	StationName	NumberStartTrips
6100	Mackay / de Maisonneuve	97150
6184	Métro Mont-Royal (Rivard / du Mont-Royal)	81279
6078	Métro Place-des-Arts (de Maisonneuve / de Bleury)	78848
6136	Métro Laurier (Rivard / Laurier)	76813
6064	Métro Peel (de Maisonneuve / Stanley)	72298

2. Second, write a query that counts, for each station, the number of round trips.
   Here are the first 5:

code	StationName	NumberRoundTrips
6501	Métro Jean-Drapeau	8658
6026	de la Commune / Place Jacques-Cartier	5622
6036	de la Commune / St-Sulpice	4123
6023	de la Commune / Berri	2591
6050	de la Commune / McGill	2182

3. Combine the above queries and calculate the fraction of round trips to the total number of starting trips for each station.
   Here are the first 5:

code	StationName	RT_Start_Fraction
6501	Métro Jean-Drapeau	0.302
7048	Métro Angrignon	0.2331
6428	Berlioz / de l'Île des Soeurs	0.2043
7015	LaSalle / 4e avenue	0.2006
6736	Basile-Routhier / Gouin	0.1932

4. Filter down to stations with at least 50 trips originating from them.
   Ordering the number of start trips so that we can see minimum value to make sure that the filter is applied. Here are the first 5:

code	StationName	NumberStartTrips	RT_Start_Fraction
7075	CHSLD Éloria-Lepage (de la Pépinière / de Marseille)	475	0.1263
7009	CHSLD Benjamin-Victor-Rousselot (Dickson / Sherbrooke)	534	0.0449

5003	Place Longueuil	618	0.0502
7023	CHSLD St-Michel (Jarry / 8e avenue)	855	0.0234
7016	Métro Langelier (Sherbrooke / Langelier)	909	0.0781

5. Given what we learned above about the relation between round trips, membership status, and day of the week, where would you expect to find stations with a high fraction of round trips?

   Copying the final table where the result is filtered to only contain fraction of round trips greater than 10% i.e. 0.1 fraction:

code	StationName	NumberStartTrips	RT_Start_Fraction
6501	Métro Jean-Drapeau	28672	0.302
7048	Métro Angrignon	2398	0.2331
6428	Berlioz / de l'Île des Soeurs	5246	0.2043
7015	LaSalle / 4e avenue	2991	0.2006
6736	Basile-Routhier / Gouin	1708	0.1932
6359	Parc Plage	6201	0.1846
7007	Gare Canora	2439	0.1792
6714	LaSalle / Sénécal	3151	0.1473
6502	Casino de Montréal	6138	0.1437
6109	Quai de la navette fluviale	6417	0.1376
7075	CHSLD Éloria-Lepage (de la Pépinière / de Marseille)	475	0.1263
6026	de la Commune / Place Jacques-Cartier	50822	0.1106
6016	Jacques-Le Ber / de la Pointe Nord	2719	0.1103
6429	Place du Commerce	8569	0.1082
5006	Collège Édouard-Montpetit	1439	0.1001

   I would expect to find stations with a high fraction of round trips in areas where there are a lot of members since members do round trips more often than non-members. Members are more likely to be locals so I would expect the area to be more residential.

   However, from the table, we can see that Métro Jean-Drapeau, which has the highest fraction of round trips, also has a much higher number of starting trips than majority of other stations in the table. I would expect this area to be also a tourist area based on the number of starting trips. It might contain some points of interest in the area, which can be done in a round trip.

Once again, comparing it to Vancouver, this area might be close to a park like Stanley park, where users can bike around in a round trip.