Bixi Bikes Business Intelligence Report

By Chosen Canlas

For BrainStation Bixi Project Deliverable 1

October 4, 2020

Table of Contents

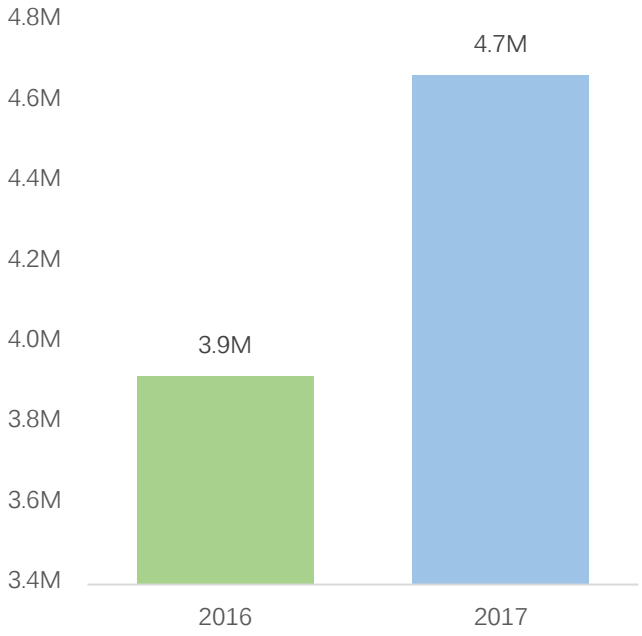
Summary

This report seeks to aid overall business growth by using MySQL and Excel to gain a high-level understanding into client usage of Bixi bikes for members and non-members, seasonal demand, and station popularity. This report will then make general recommendations on what actions can be taken to aid business growth.

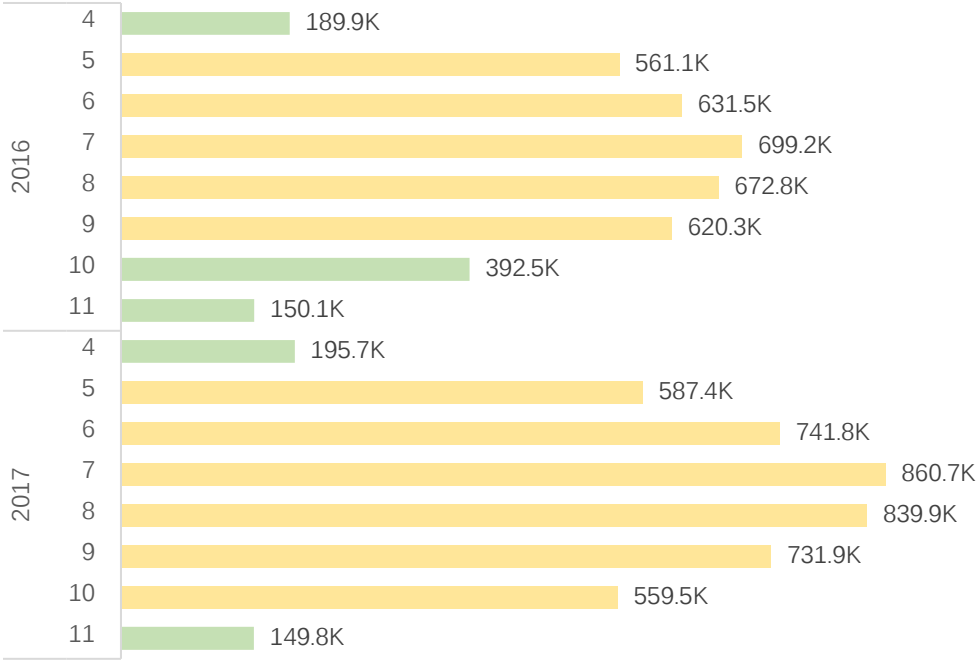
Methods and Findings

Using MySQL Workbench and Excel, I processed a dataset of 8.6M trips in 2016-2017 to find the following information:

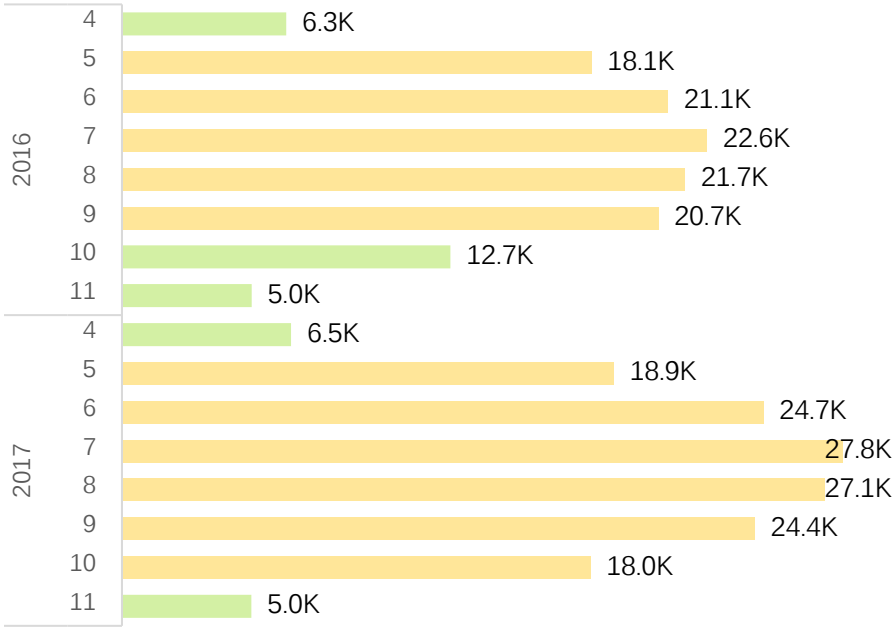
1. Number of trips per year



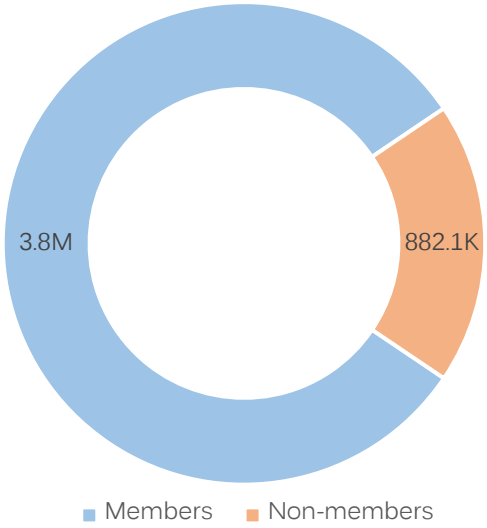
2. Number of trips per month (peak months in yellow)



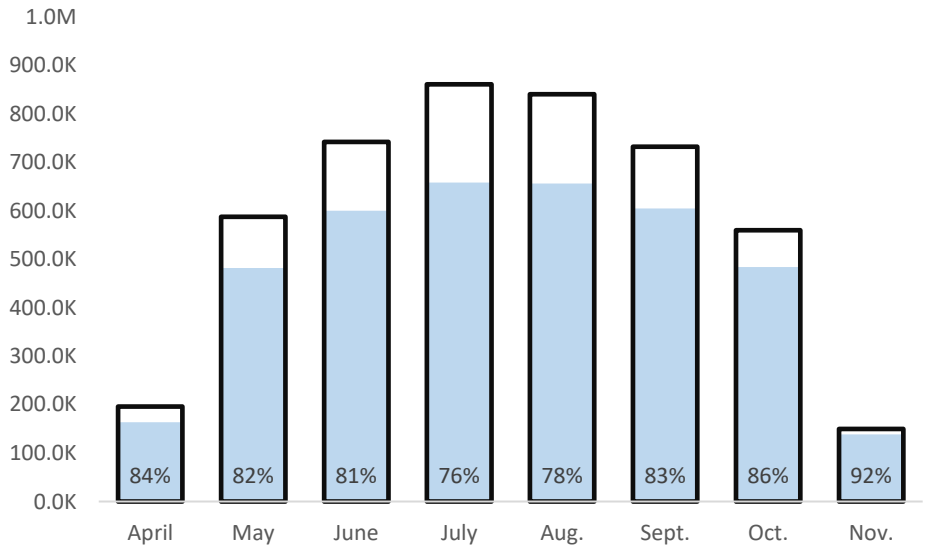
3. The average number of daily trips per month (peak months in yellow)



4. Number of trips according to membership status in 2017



5. Fraction/Percentage of total trips done by members month by month

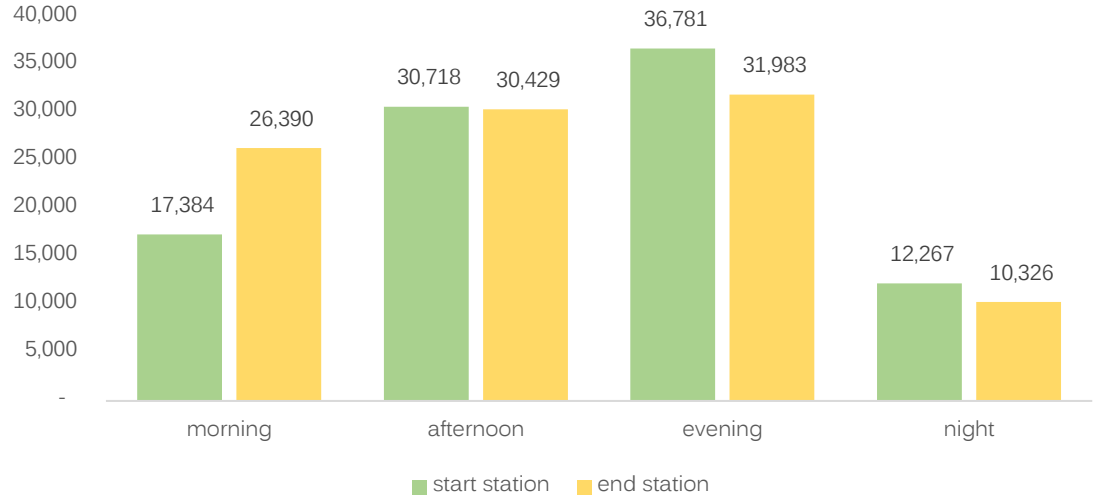


6. Most popular stations

Start Station Code	Name	Trips
6100	Mackay / de Maisonneuve	97,150
6184	Métro Mont-Royal (Rivard / du Mont-Royal)	81,279

6078	Métro Place-des-Arts (de Maisonneuve / de Bleury)	78,848
6136	Métro Laurier (Rivard / Laurier)	76,813
6064	Métro Peel (de Maisonneuve / Stanley)	72,298

7. Number of start and end trips at Mackay / de Maisonneuve station



8. Stations with a high fraction of round trips

Station	Roundtrips	Total_trips	PercentageOfTotalTrips
Métro Jean-Drapeau	8,658	28,672	30.20
Métro Angrignon	559	2,398	23.31
Berlioz / de l'Île des Soeurs	1,072	5,246	20.43
LaSalle / 4e avenue	600	2,991	20.06
Basile-Routhier / Gouin	330	1,708	19.32
Parc Plage	1,145	6,201	18.46
Gare Canora	437	2,439	17.92
LaSalle / Sénécal	464	3,151	14.73
Casino de Montréal	882	6,138	14.37

Quai de la navette fluviale	883	6,417	13.76
de la Commune / Place Jacques-Cartier	5,622	50,822	11.06
Jacques-Le Ber / de la Pointe Nord	300	2,719	11.03
Place du Commerce	927	8,569	10.82
Collège Édouard-Montpetit	144	1,439	10.01

Conclusions and Recommendations

Time-period and Seasonal Trends in Bixi Bike Usage

There is a significant trip increase with 2017 having 1.2M trips more than 2016. This shows an increase in demand for Bixi Bikes. Furthermore, there is a clear seasonal trend with bike usage. The peak months are from May to September, with the highest month being July in both years. This is also consistent with the daily average number of trips per month, for both years with July consistently being the highest month. May to September are the warmest months and are also peak season for tourism as it is the months when students are on vacation from school and families are more inclined to travel to Montreal. Given the significant increase from 2016 to 2017, I recommend increasing the number of bikes available to prepare for 2018 in order to anticipate demand and maximize revenue.

Members and Non-members comparison

Members usage consistently has a higher percentage of trips in comparison to non-members. However, the number of trips taken by non-members drops significantly during colder months. Part 5 of the previous section also indicates that membership trips drops during colder months, which could signify loss of revenue during those months too. To increase trips during off-peak months (October to April), given that most of the usage comes from members, I recommend that this be the period where special promotions are offered to non-members. There is no need to offer membership promotions to non-members during peak months because non-member fees are higher, and there is greater revenue potential from non-members per trip.

Station popularity

Mackay / de Maisonneuve is the top station in 2017. If we look closer, we can see geographical patterns in trips by comparing the number of trips that begin in Mackay / de Maisonneuve, and the number of trips that end there. In the morning, there is a difference of 9,006 trips indicating that users come from other stations to arrive at Mackay / de Maisonneuve. This may be because Mackay / de Maisonneuve is the closest station to many work sites and offices. This aligns with the afternoon, evening, and night comparisons where trips begin in Mackay / de Maisonneuve station are greater than the trips that end there. This time-period is in line with the end of the workday. This information is also important for logistics. If bikes are not transported from other stations to Mackay / de Maisonneuve station, there might not be enough bikes to meet the demand. An average of 165 bikes needs to be transported to the Mackay / de Maisonneuve station every month. Of course, this must vary according to seasonal bike usage trends.

Stations with high fraction of round trips

Stations with high fractions of round trips do not need as much bike transport management. Therefore, these locations do not require cost from the company. The top 5 stations are Métro Jean-Drapeau (30.20% of trips are round trips), Métro Angrignon (23.31%), Berlioz / de l'Île des Soeurs (20.43%), LaSalle / 4e avenue (20.06) , and Basile-Routhier / Gouin (19.32%). However, these percentages are quite low, and the company needs to be very active in keeping all stations stocked with bikes.

Appendix

This Appendix contains all the queries I used in MySQL Workbench to analyze the data.

Question 1

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

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

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

```
SELECT YEAR(start_date) AS Year, COUNT(start_date) AS NumberOfTrips

FROM trips

GROUP BY Year;
```

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

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

```
SELECT EXTRACT(YEAR_MONTH FROM start_date) AS MonthYear, DATE_FORMAT(start_date,
'%m') AS Month, DATE_FORMAT(start_date, '%Y') AS Year,

COUNT(start_date) AS NumberOfTrips

FROM trips

GROUP BY MonthYear;
```

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

```
SELECT EXTRACT(YEAR_MONTH FROM start_date) AS MonthYear, DATE_FORMAT(start_date,
'%m') AS Month, DATE_FORMAT(start_date, '%Y') AS Year,

COUNT(start_date) AS NumberOfTrips, Day(last_day(start_date)) AS NumberOfDaysPerMonth,
COUNT(start_date) /  Day(last_day(start_date)) AS DailyTripAvg

FROM trips

GROUP BY MonthYear;
```

Save your query results from the previous question (Q1.5) by creating a table called working_table1.

*I did this without writing a query since the question did not specify that we had to do it the querying way. I just right-clicked on the tables, and clicked 'Create a Table.'

Question 2

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).

```
SELECT is_member, YEAR(start_date) AS Year, COUNT(start_date) AS NumberOfTrips

FROM trips

WHERE YEAR(start_date) = 2017

GROUP BY is_member;
```

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

```
SELECT is_member, Month(start_date), COUNT(start_date) AS NumberOfTrips, COUNT(id) /
SUM(count(id))

FROM trips

WHERE YEAR(start_date) = 2017 AND is_member = 1

GROUP BY Month(start_date);
```

Question 4

It is clear now that average temperature 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? Solve this problem without using a subquery.

```
SELECT trips.start_station_code, stations.name, COUNT(trips.start_station_code) AS NumberofTrips

FROM trips

JOIN stations ON trips.start_station_code = stations.code

GROUP BY start_station_code

ORDER BY NumberofTrips DESC
```

```
LIMIT 5;
```

2. Solve the same question as Q4.1, but now use a subquery. Is there a difference in query run time between 4.1 and 4.2?

Yes, but it the run time changed on the day and time. In the beginning, run time in the previous query was 5 seconds. The next day, it increased to 12 seconds. The run time for the code with subquery was 9 seconds in the beginning and shortened to 6 seconds the next day. I am not sure what caused these changes in run time because I did not change the code.

```
SELECT
        NumberofTripsTbl.NumberofTrips,
   stations.code
FROM
        (
        SELECT
                COUNT(trips.start_station_code) AS NumberofTrips,
     trips.start_station_code
        FROM
                trips
        GROUP BY
                start_station_code
        ORDER BY
                NumberofTrips
        LIMIT 5
        ) AS NumberofTripsTbl
                JOIN stations ON
                        stations.code = NumberofTripsTbl.start_station_code;
```

Question 5

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

```
SELECT 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?

Code to see whole table:

```
SELECT name, start_date, end_date,
        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 "start_time_of_day",
        CASE
    WHEN HOUR(end_date) BETWEEN 7 AND 11 THEN "morning"
    WHEN HOUR(end_date) BETWEEN 12 AND 16 THEN "afternoon"
    WHEN HOUR(end_date) BETWEEN 17 AND 21 THEN "evening"
    ELSE "night"
    END AS "end_time_of_day"
FROM
(
SELECT * FROM stations
WHERE name LIKE '%Mackay / de Maisonneuve%'
) AS M
JOIN trips
ON M.code = trips.start_station_code;

#no. of starts in morning, afternoon and evening

SELECT COUNT(name) AS trips_start,
```

```
        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 "start_time_of_day"
FROM
(
SELECT * FROM stations
WHERE name LIKE '%Mackay / de Maisonneuve%'
) AS M
JOIN trips
ON M.code = trips.start_station_code
GROUP BY start_time_of_day;
```

Code to see the no. of starts in morning, afternoon and evening for trips that end with M station:

```
SELECT COUNT(name) AS trips_start,
        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 " start _time_of_day"
FROM
(
SELECT * FROM stations
WHERE name LIKE '%Mackay / de Maisonneuve%'
) AS M
JOIN trips
ON M.code = trips. start _station_code
GROUP BY start_time_of_day;
```

Code to see the no. of ends in morning, afternoon and evening for trips that end with M station:

```
SELECT COUNT(name) AS trips_end,
        CASE
    WHEN HOUR(end_date) BETWEEN 7 AND 11 THEN "morning"
    WHEN HOUR(end_date) BETWEEN 12 AND 16 THEN "afternoon"
    WHEN HOUR(end_date) BETWEEN 17 AND 21 THEN "evening"
    ELSE "night"
    END AS "end_time_of_day"
FROM
(
SELECT * FROM stations
WHERE name LIKE '%Mackay / de Maisonneuve%'
) AS M
JOIN trips
ON M.code = trips.end_station_code
GROUP BY end_time_of_day;
```

Question 6

List all stations for which at least 10% of trips are round trips. Round trips are those that start and end in the same station.

This time we will only consider stations with at least 500 starting trips. (Please include answers for all steps outlined here)

First, write a query that counts the number of starting trips per station.

```
SELECT
        trips.start_station_code,
        COUNT(trips.id) AS NumberofTrips
FROM
        trips
GROUP BY trips.start_station_code;
```

Second, write a query that counts, for each station, the number of round trips.

```
SELECT
        start_station_code,
   COUNT(start_station_code) AS NumberofRoundTrips

        FROM
            trips
        WHERE start_station_code = end_station_code
   GROUP BY start_station_code;
```

Combine the above queries and calculate the fraction of round trips to the total number of starting trips for each station.

```
SELECT
        stations.name AS Station,
   COUNT(IF(trips.start_station_code = trips.end_station_code, 1, NULL)) AS Roundtrips,
        COUNT(trips.id) AS Total_trips,
   100*(COUNT(IF(trips.start_station_code = trips.end_station_code , 1, NULL))/COUNT(trips.id)) AS
PercentageOfTotalTrips
FROM trips
JOIN stations
        ON trips.start_station_code = stations.code
GROUP BY stations.name
ORDER BY PercentageOfTotalTrips DESC;
```

Filter down to stations with at least 500 trips originating from them and having at least 10% of their trips as round trips.

```
SELECT
        stations.name AS Station,
```

```
   COUNT(IF(trips.start_station_code = trips.end_station_code , 1, NULL)) AS Roundtrips,

        COUNT(trips.id) AS Total_trips,

   100*(COUNT(IF(trips.start_station_code = trips.end_station_code , 1, NULL))/COUNT(trips.id)) AS
PercentageOfTotalTrips

FROM trips

JOIN stations

        ON trips.start_station_code = stations.code

GROUP BY stations.name

HAVING PercentageOfTotalTrips >= 10

AND Total_trips >= 500

ORDER BY PercentageOfTotalTrips DESC;
```