**Lab: Bikeshare data using SQL and Excel**

**Data source selection**

As different countries and cities have different factors impacting business performance, it’s hard to make comparison across business using count of trips and growth rate only.

In this instance, I choose Citybike - New York because it has sufficient fields of data including gender, age, bike for each trip etc.

**Trends and descriptive analytics**

1. **How many trips were there in each month of each year? How many different bikes were there in each month of each year?**

WITH summary AS

(SELECT \* FROM citibike\_2016

UNION

SELECT \* FROM citibike\_2017

UNION

SELECT \* FROM citibike\_2018

UNION

SELECT \* FROM citibike\_2019)

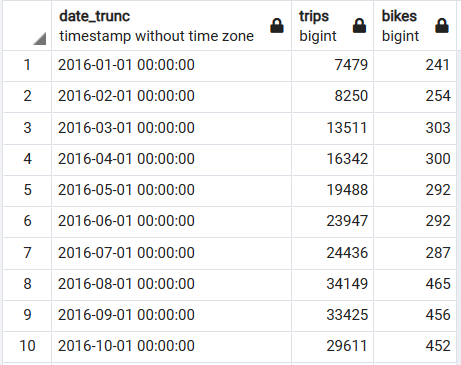
SELECT DATE\_TRUNC('month',start\_time),

COUNT(\*) AS trips,

COUNT(DISTINCT bike\_id) AS bikes

FROM summary

GROUP BY DATE\_TRUNC('month',start\_time);



1. **Is the subscription side of these businesses growing? (In absolute terms; and also as a proportion of the sharing).**

**SQL:**

WITH summary AS

(SELECT \* FROM citibike\_2016

UNION

SELECT \* FROM citibike\_2017

UNION

SELECT \* FROM citibike\_2018

UNION

SELECT \* FROM citibike\_2019)

SELECT DATE\_TRUNC('month',start\_time),

user\_type,

COUNT(\*)

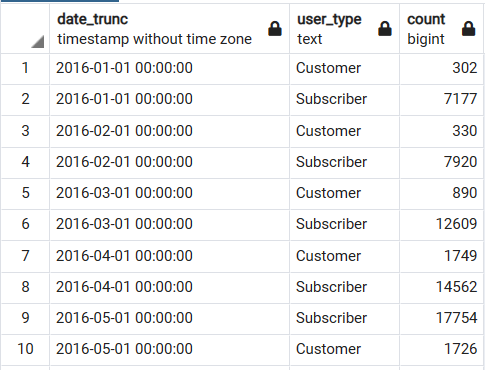
FROM summary

WHERE user\_type IS NOT NULL

GROUP BY DATE\_TRUNC('month',start\_time),

user\_type

ORDER BY DATE\_TRUNC('month',start\_time);



1. **If 2020 hadn’t been, well, 2020, what would you expect to have seen?**

Workings in excel file.

**Geospatial**

1. **What was the longest journey? What do we know about it?**

The longest journal was from Bergen Ave to Amsterdam Ave of 14.03km.

WITH summary AS

(SELECT \* FROM citibike\_2016

UNION

SELECT \* FROM citibike\_2017

UNION

SELECT \* FROM citibike\_2018

UNION

SELECT \* FROM citibike\_2019)

SELECT c1.name AS "From", c2.name AS "To",

CALCULATE\_DISTANCE(c2.latitude, c2.longitude, c1.latitude, c1.longitude, 'K')

FROM summary

JOIN citibike\_stations AS c1

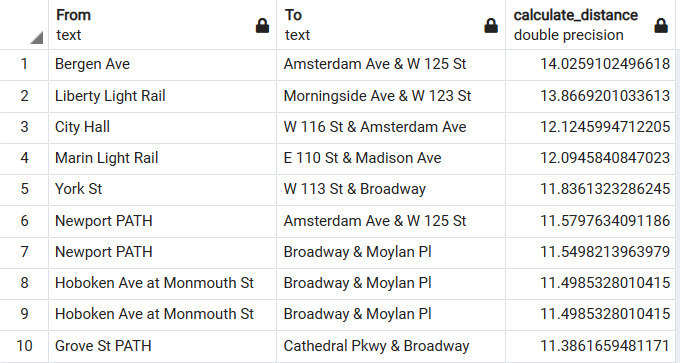
ON start\_station\_id=c1.id

JOIN citibike\_stations AS c2

ON end\_station\_id=c2.id

ORDER BY CALCULATE\_DISTANCE(c2.latitude, c2.longitude, c1.latitude, c1.longitude, 'K') DESC

LIMIT 10;



1. **What was the furthest relocation?**

The relocation from Amsterdam Ave to McGinley St was the longest of 13.54km.

WITH summary AS

(SELECT \* FROM citibike\_2016

UNION

SELECT \* FROM citibike\_2017

UNION

SELECT \* FROM citibike\_2018

UNION

SELECT \* FROM citibike\_2019)

SELECT c2.name AS "From",

c1.name AS "To",

CALCULATE\_DISTANCE(c2.latitude, c2.longitude, c1.latitude, c1.longitude, 'K')

FROM

(SELECT \*, LAG(end\_station\_id,1)OVER(PARTITION BY bike\_id ORDER BY start\_time) AS prev\_id

FROM summary) AS window\_table

JOIN citibike\_stations AS c1

ON start\_station\_id=c1.id

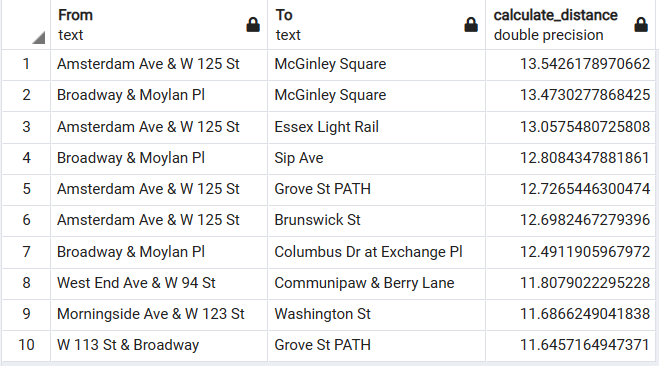
JOIN citibike\_stations AS c2

ON prev\_id=c2.id

WHERE window\_table.prev\_id!=start\_station\_id

ORDER BY CALCULATE\_DISTANCE(c2.latitude, c2.longitude, c1.latitude, c1.longitude, 'K') DESC

LIMIT 10



1. **How far is a typical journey?**

Typical journey is 0.97km.

WITH summary AS

(SELECT \* FROM citibike\_2016

UNION

SELECT \* FROM citibike\_2017

UNION

SELECT \* FROM citibike\_2018

UNION

SELECT \* FROM citibike\_2019)

SELECT AVG(CALCULATE\_DISTANCE(c2.latitude, c2.longitude, c1.latitude, c1.longitude, 'K'))

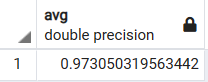
FROM summary

JOIN citibike\_stations AS c1

ON start\_station\_id=c1.id

JOIN citibike\_stations AS c2

ON end\_station\_id=c2.id



**Business and Commercial**

1. **What sort of people use these bikes, and when do they use them?**

**Trip by age groups:**

WITH summary AS

(SELECT \* FROM citibike\_2016

UNION

SELECT \* FROM citibike\_2017

UNION

SELECT \* FROM citibike\_2018

UNION

SELECT \* FROM citibike\_2019)

SELECT CASE WHEN birth\_year<=2003 AND birth\_year>1993 THEN '1994-2003'

WHEN birth\_year<=1993 AND birth\_year>1983 THEN '1984-1993'

WHEN birth\_year<=1983 AND birth\_year>1973 THEN '1974-1983'

WHEN birth\_year<=1973 AND birth\_year>1963 THEN '1964-1973'

WHEN birth\_year<=1963 AND birth\_year>1953 THEN '1954-1963'

WHEN birth\_year<=1953 AND birth\_year>1943 THEN '1944-1953'

WHEN birth\_year<=1943 AND birth\_year>1933 THEN '1934-1943'

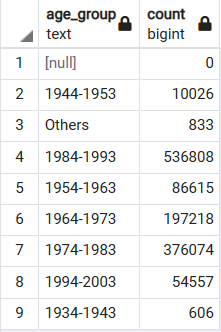
WHEN birth\_year<=1933 THEN 'Others'

END AS age\_group,

COUNT(birth\_year)

FROM summary

GROUP BY age\_group



**Trip by time period:**

WITH summary AS

(SELECT \* FROM citibike\_2016

UNION

SELECT \* FROM citibike\_2017

UNION

SELECT \* FROM citibike\_2018

UNION

SELECT \* FROM citibike\_2019)

SELECT CASE WHEN DATE\_PART('hour',start\_time)>0 AND DATE\_PART('hour',start\_time)<=2 THEN '00-02'

WHEN DATE\_PART('hour',start\_time)>2 AND DATE\_PART('hour',start\_time)<=4 THEN '02-04'

WHEN DATE\_PART('hour',start\_time)>4 AND DATE\_PART('hour',start\_time)<=6 THEN '04-06'

WHEN DATE\_PART('hour',start\_time)>6 AND DATE\_PART('hour',start\_time)<=8 THEN '06-08'

WHEN DATE\_PART('hour',start\_time)>8 AND DATE\_PART('hour',start\_time)<=10 THEN '08-10'

WHEN DATE\_PART('hour',start\_time)>10 AND DATE\_PART('hour',start\_time)<=12 THEN '10-12'

WHEN DATE\_PART('hour',start\_time)>12 AND DATE\_PART('hour',start\_time)<=14 THEN '12-14'

WHEN DATE\_PART('hour',start\_time)>14 AND DATE\_PART('hour',start\_time)<=16 THEN '14-16'

WHEN DATE\_PART('hour',start\_time)>16 AND DATE\_PART('hour',start\_time)<=18 THEN '16-18'

WHEN DATE\_PART('hour',start\_time)>18 AND DATE\_PART('hour',start\_time)<=20 THEN '18-20'

WHEN DATE\_PART('hour',start\_time)>20 AND DATE\_PART('hour',start\_time)<=22 THEN '20-22'

WHEN DATE\_PART('hour',start\_time)>22 AND DATE\_PART('hour',start\_time)<=24 THEN '22-24'

END AS time\_group,

COUNT(start\_time)

FROM summary

GROUP BY time\_group

