Data Management & SQL

FINAL DELIVERABLE

Shresth Sethi

Hult International Business School

Table of Contents

Comprehensive Final Deliverable

Section 1 (Theory):

Question 1:

What is a Data Moat? Why is it important to have one?

- A moat is a defense system build to protect against attacks. In the same way, companies are building various defense systems to maintain their competitive advantage. There are various economic moats that companies use which makes it difficult for competitors to copy. The same way there is data moat that a company has based on its proprietary data set. Competitors can copy your product, but they cannot copy your data which makes it a moat. It is important to build a strong data moat to protect from competitors stealing your market share.

Question 2:

What is the difference between OLAP and OLTP Databases? Why would you choose one over the other?

- OLTP is a transaction processing system that is online and current at the moment processing. This database is modeled using entity-relationship modeling. They are fast processing databases, so its effectiveness is measured by several transactions that it handled per second.
- Whereas, OLAP is an analytical processing system that is an online historical multidimensional data retrieval system. This database helps in retrieving data for analysis and decision making.
- Both databases serve different purposes, if the aim is to achieve high-speed processing like banking transactions then OLTP is preferred. If the aim is to analyze the historical data, then OLAP is preferred.

Question 3:

What are the 3 different roles in a modern data team? Which problems do each of them solve? How do they compare with each other?

- Three different roles in modern data team
  - Data Engineer
  - Data Analyst
  - Data Scientist
- Problems and comparison
  - Data Engineer: They solve the collection and data storage problems for the organization. They perform ETL functions to store unstructured data.
  - Data Analyst: On the data collected by data engineers a data analyst is responsible for cleaning data, and training it for analysis, they act as a bridge between data and business insights.
  - Data Scientist: They solve the problem of optimizing decision making through ML, and AI. They are responsible for future decision making and predictive outcomes.

Question 4:

Share your opinion on current data privacy laws. Are they doing enough to protect consumers? Why or why not?

- Current data privacy laws are doing just enough to protect customers. As the new law of CCPA California residents has the right to ask for information from the companies. This is just an example of how privacy laws are emerging and giving the power in the hands of consumers. The laws are doing enough to protect consumers it is the user that needs to be aware of these laws and make better use of it. A simple way to protect information is by disabling the cookies but that comes with a cost to convenience. There will always be a tradeoff between convenience and privacy.

Question 5:

How would you define the relationship between car owners and cars in the Entity Relationship (ER) model? Please provide an explanation why using real world examples.

- A car owner can have multiple cars.
- Each car belongs to one car owner.
- So, it is a 1:M relationship.
- There is Jake who owns three cars which are Mercedes, Audi, and BMW. The Mercedes has one owner i.e. Jake in the same way Audi has one owner i.e. Jake similarly BMW has one owner i.e. Jake. This in relational databases is called 1:M relationship between two entities that are Car Owners and Cars. The attributes for car owners will be name, address, amount and for cars will be model, year, and make.

Section 2 (Database Design):

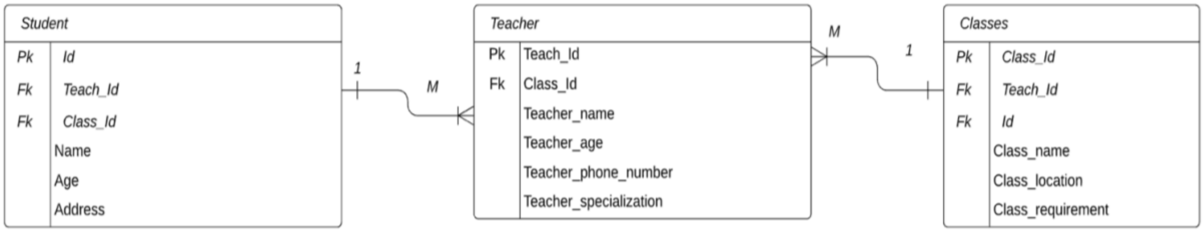
Question 1:

You are asked to model the many to many relationship between students and classes in a relational database.

- What changes do you need to make to support this relationship?
- Please create an ER diagram to show how these entities will relate to each after your changes.

STUDENTS M:M CLASSES

- Need to add another entity called teacher in between students and classes
- ER Diagram



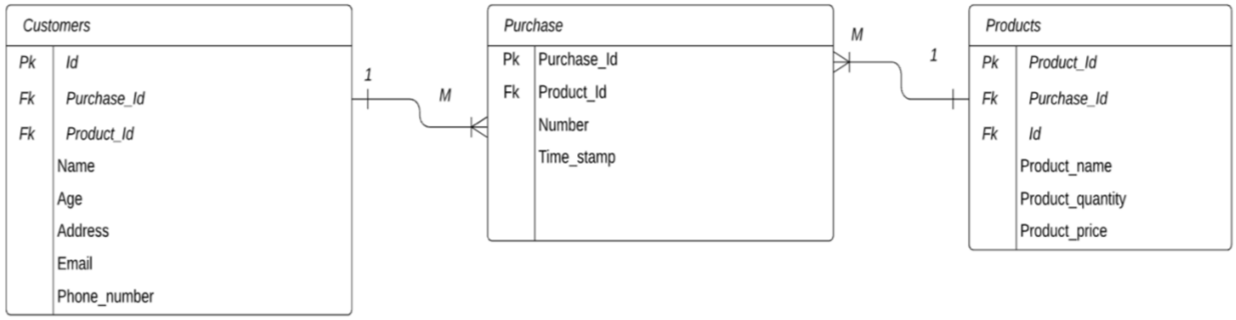
Question 2:

You are asked to model the many to many relationship between customers and products in a relational database.

- What changes do you need to make to support this relationship?
- Please create an ER diagram to show how these entities will relate to each after your changes.

CUSTOMERS M:M PRODUCTS

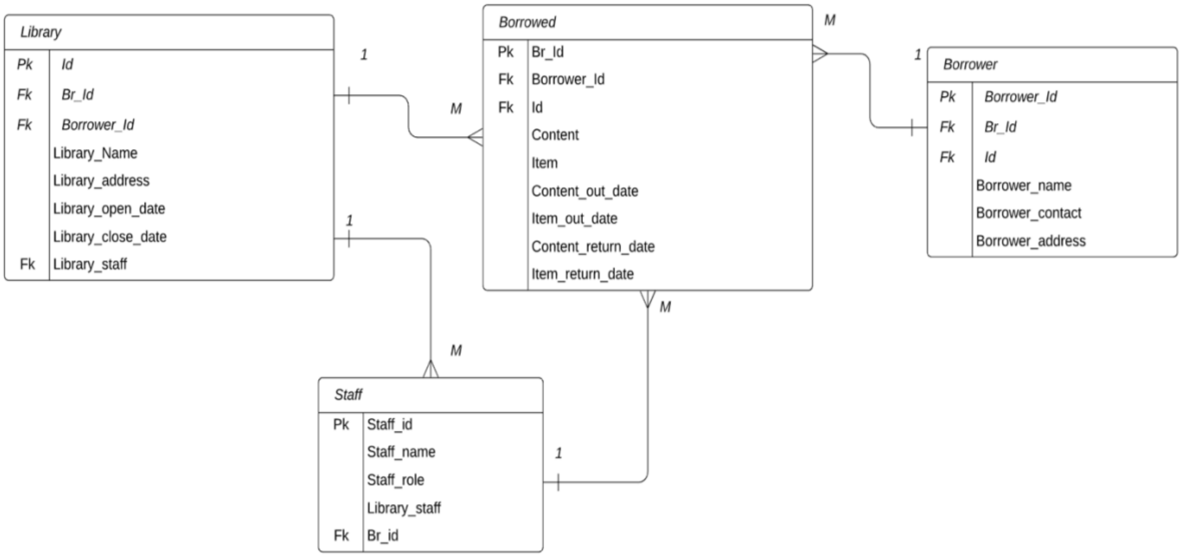
- Need to add another entity called teacher in between students and classes
- ER Diagram



Question 3:

Design an ER diagram for a library reservation system for a family of libraries based on the given characteristics.

- This system is for multiple libraries
- This system is for multiple borrowers
- There are multiple types of content that can be borrowed
- Borrowers can borrow multiple items at the same time
- Borrowers can borrow multiple types of content



Library	
Pk	Id
Fk	Br_Id
Fk	Borrower_Id
	Library_Name
	Library_address
	Library_open_date
	Library_close_date
Fk	Library_staff

Borrowed	
Pk	Br_Id
Fk	Borrower_Id
Fk	Id
	Content
	Item
	Content_out_date
	Item_out_date
	Content_return_date
	Item_return_date

Borrower	
Pk	Borrower_Id
Fk	Br_Id
Fk	Id
	Borrower_name
	Borrower_contact
	Borrower_address

Staff	
Pk	Staff_id
	Staff_name
	Staff_role
	Library_staff
Fk	Br_id

Section 3 (Data Analysis with SQL):

Data Set: bigquery-public-data -> london_bicycles

Problem: Empty bike stations

Problem Diagnoses:
I found traces of empty stations in the dataset where bike count is zero. The total number of stations is 781 out of the 8 stations are empty. Also found that there is one station that has bike count as 0 but the docking count is 21 and empty docs are 20 which means there should be a bike count of 1. Station name: St. Mary Axe, Aldgate. Considering this situation, I will assume that there is a bike present but that has not been reported. Total empty stations using this assumption is 8.

The problem is big as the number of people going to these stations is huge. The highest number of people 50331 visiting empty station Braham Street, Aldgate. The second highest is the Museum of London, Barbican with 48821 people renting bikes and so on. The footfall at these stations are reducing as people cannot find bikes and they move to another station.

Let’s dig deep into the problem, the most popular stations are Belgrove Street, King's Cross with 232221 users followed by Hyde Park Corner, Hyde Park with 213237 users then Waterloo Station 3, Waterloo with 199671 users.

The usage of rental bikes peaks in the morning and evening as this is normal to understand. Most of the office workers travel in the morning to reach their offices and leave by evening to reach their home. So, we saw the highest number of users at 08:00 am i.e. 2574531 rides and the next highest rides are at 05:00 pm i.e. 2525974. The numbers are almost similar for morning and evening rides which can be classified as daily office going population. Similarly, for other evening times at 06:00 pm is 2332198 and at 04:00 pm is 1662546. For the morning time, 09:00 am is 1474912. The afternoon is the time when there are not many riders.

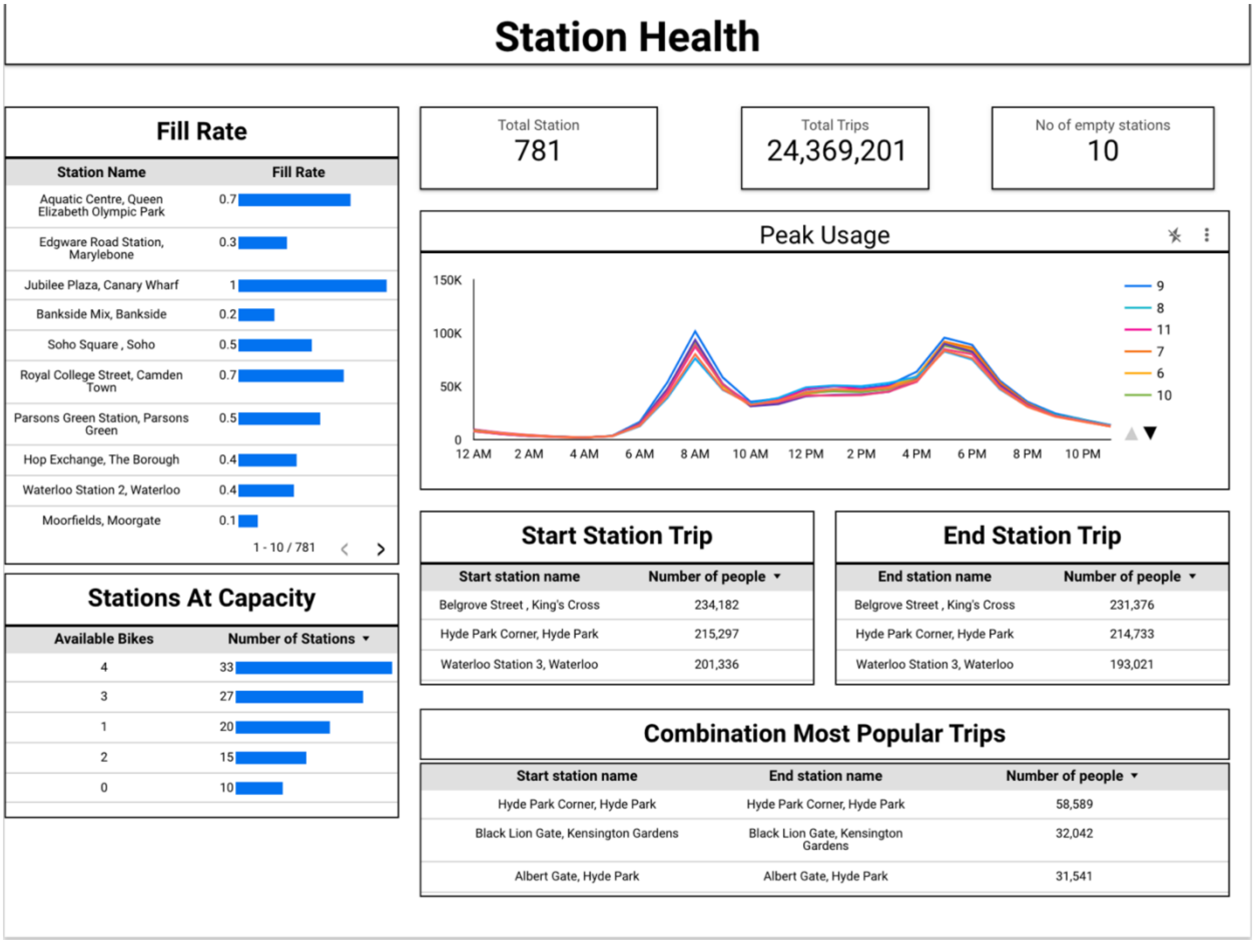
Hyde Park Corner to Hyde Park Corner is the most popular trip in the network, this makes sense Hyde Park Corner has a Hyde Park Corner tube station so people traveling for remote areas of London use this tube for their day to day commute. Another popular trip is Black Lion Gate, Kensington Gardens to Black Lion Gate, Kensington Gardens this is a famous spot in London with bars and tourist excursions, this park is immediate west to Hyde Park which makes the connection with Hyde Park Corner and Kensington garden have a high number of trips. This is followed by various other parks being the most popular trips.

There are six different types of rides that people take. The major difference between them is the ride that people take that starts at the same station and end at the same station and then another type of people are those who start at one station and end the ride at a different station. The number of people who have different stations as start and end is more than the people taking the same station rides. The max people ride time is less than 45 mins or less than 10mins there are fewer people whose ride time exceeds 45mins.

There is a pattern in the types of stations that are empty as all these stations have a ride time lower than 30 mins and when compared with stations that are not empty have a ride time usually greater than 30 mins. The number of people visiting these stations is less than 50k in a particular station. This could be the case as people see that there are no bikes in these stations, so they don't come here anymore except the people who require to be at these stations. The dock counts for these stations is low less than 20 docks in most of the cases only one station The Museum of London has 52 docs.

Section 4 (Data Visualization on top of SQL):

Station Health



Station Health

Fill Rate

Station Name	Fill Rate
Aquatic Centre, Queen Elizabeth Olympic Park	0.7
Edgware Road Station, Marylebone	0.3
Jubilee Plaza, Canary Wharf	1
Bankside Mix, Bankside	0.2
Soho Square , Soho	0.5
Royal College Street, Camden Town	0.7
Parsons Green Station, Parsons Green	0.5
Hop Exchange, The Borough	0.4
Waterloo Station 2, Waterloo	0.4
Moorfields, Moorgate	0.1

1 - 10 / 781

Total Station
781

Total Trips
24,369,201

No of empty stations
10

Peak Usage

Stations At Capacity

Available Bikes	Number of Stations ▾
4	33
3	27
1	20
2	15
0	10

Start Station Trip

Start station name	Number of people ▾
Belgrove Street , King's Cross	234,182
Hyde Park Corner, Hyde Park	215,297
Waterloo Station 3, Waterloo	201,336

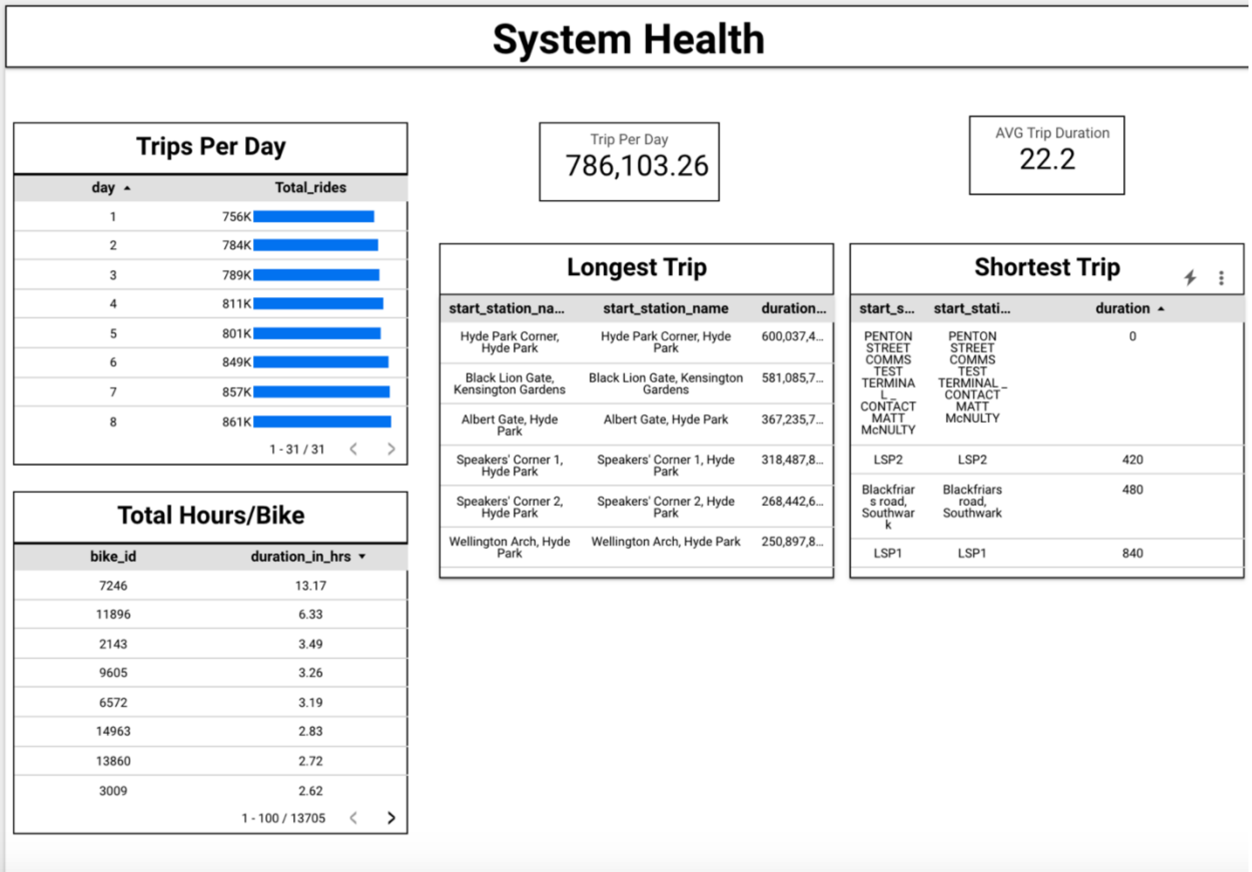
End Station Trip

End station name	Number of people ▾
Belgrove Street , King's Cross	231,376
Hyde Park Corner, Hyde Park	214,733
Waterloo Station 3, Waterloo	193,021

Combination Most Popular Trips

Start station name	End station name	Number of people ▾
Hyde Park Corner, Hyde Park	Hyde Park Corner, Hyde Park	58,589
Black Lion Gate, Kensington Gardens	Black Lion Gate, Kensington Gardens	32,042
Albert Gate, Hyde Park	Albert Gate, Hyde Park	31,541

System Health



System Health

Trips Per Day

day ▲	Total_rides
1	756K
2	784K
3	789K
4	811K
5	801K
6	849K
7	857K
8	861K

1 - 31 / 31

Trip Per Day
786,103.26

AVG Trip Duration
22.2

Longest Trip

start_station_na...	start_station_name	duration...
Hyde Park Corner, Hyde Park	Hyde Park Corner, Hyde Park	600,037,4...
Black Lion Gate, Kensington Gardens	Black Lion Gate, Kensington Gardens	581,085,7...
Albert Gate, Hyde Park	Albert Gate, Hyde Park	367,235,7...
Speakers' Corner 1, Hyde Park	Speakers' Corner 1, Hyde Park	318,487,8...
Speakers' Corner 2, Hyde Park	Speakers' Corner 2, Hyde Park	268,442,6...
Wellington Arch, Hyde Park	Wellington Arch, Hyde Park	250,897,8...

Shortest Trip

start_s...	start_stati...	duration ▲
PENTON STREET COMMS TEST TERMINAL _ CONTACT MATT McNULTY	PENTON STREET COMMS TEST TERMINAL _ CONTACT MATT McNULTY	0
LSP2	LSP2	420
Blackfriars road, Southwark	Blackfriars road, Southwark	480
LSP1	LSP1	840

Total Hours/Bike

bike_id	duration_in_hrs ▼
7246	13.17
11896	6.33
2143	3.49
9605	3.26
6572	3.19
14963	2.83
13860	2.72
3009	2.62

1 - 100 / 13705

Appendix:

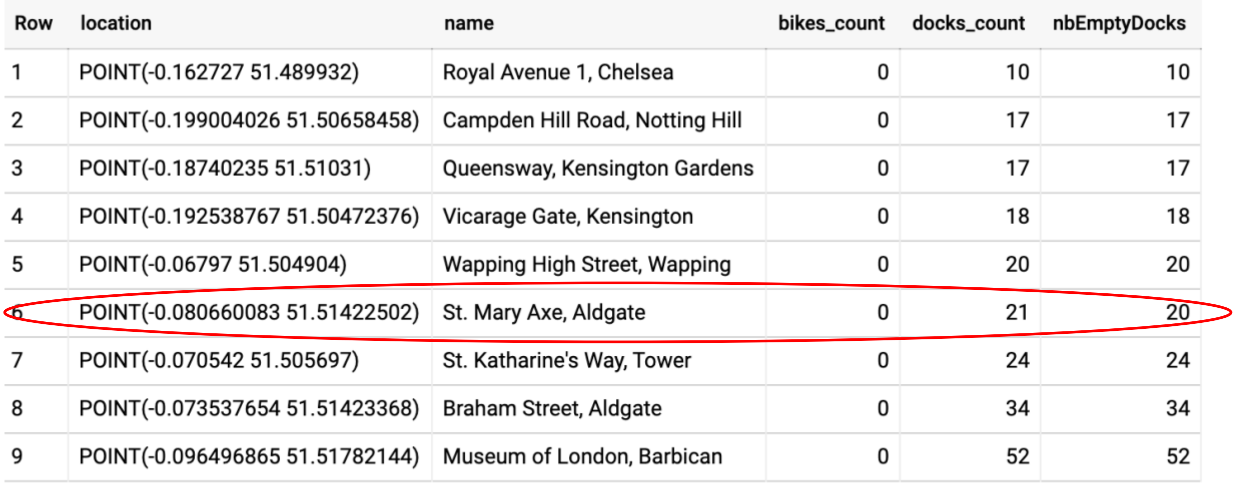
```
SELECT
  count(Distinct( name))
FROM
  `bigquery-public-data.london_bicycles.cycle_stations`

SELECT
  nbEmptyDocks,
  docks_count,
  bikes_count,
  name
FROM
  `bigquery-public-data.london_bicycles.cycle_stations`
WHERE
  nbEmptyDocks > 0 ;


  -- Can you find any traces of empty stations?
SELECT
  ST_GeogPoint(longitude,
    latitude) AS location,
  name,
  bikes_count,
  docks_count,
  nbEmptyDocks
FROM
  `bigquery-public-data.london_bicycles.cycle_stations`
WHERE
  docks_count = nbEmptyDocks;
```

Result: This shows the stations that equal number or empty docs and total docs

```
  ---------------------------------------------------
SELECT
  ST_GeogPoint(longitude,
    latitude) AS location,
  name,
  bikes_count,
  docks_count,
  nbEmptyDocks
FROM
  `bigquery-public-data.london_bicycles.cycle_stations`
WHERE
  bikes_count = 0;
```

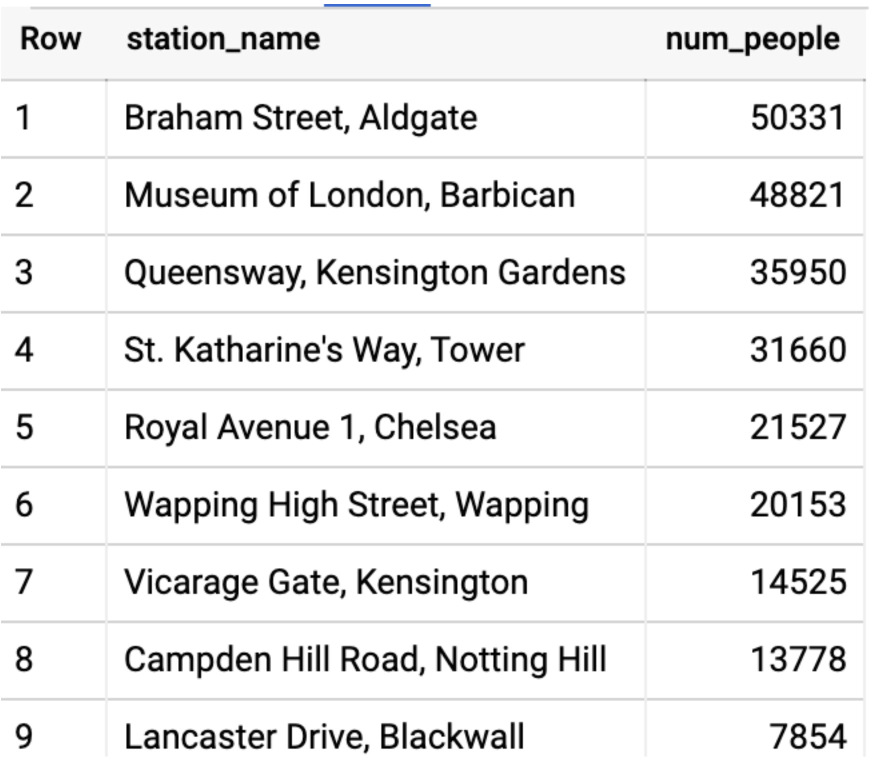
Result: This shows the stations that have bike count as 0 i.e there is no bike at that station

Found that there is one station that has bike count as 0 but the dock count is 21 and empty docs is 20 that means there should be a bike count of 1. Station name: St. Mary Axe, Aldgate. Considering this situation I will assume that there is a bike that has not been reported and the total number of empty stations will be 8.



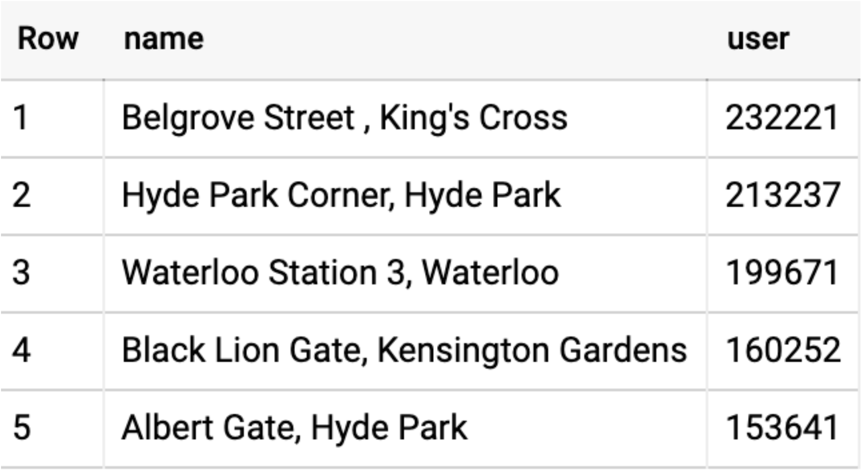
Row	location	name	bikes_count	docks_count	nbEmptyDocks
1	POINT(-0.162727 51.489932)	Royal Avenue 1, Chelsea	0	10	10
2	POINT(-0.199004026 51.50658458)	Campden Hill Road, Notting Hill	0	17	17
3	POINT(-0.18740235 51.51031)	Queensway, Kensington Gardens	0	17	17
4	POINT(-0.192538767 51.50472376)	Vicarage Gate, Kensington	0	18	18
5	POINT(-0.06797 51.504904)	Wapping High Street, Wapping	0	20	20
6	POINT(-0.080660083 51.51422502)	St. Mary Axe, Aldgate	0	21	20
7	POINT(-0.070542 51.505697)	St. Katharine's Way, Tower	0	24	24
8	POINT(-0.073537654 51.51423368)	Braham Street, Aldgate	0	34	34
9	POINT(-0.096496865 51.51782144)	Museum of London, Barbican	0	52	52

```
  -- If yes, how big is this problem?
SELECT
  station.name AS station_name,
  count(distinct(rental_id)) AS num_people
FROM
  `bigquery-public-data.london_bicycles.cycle_stations` AS station
INNER JOIN
  `bigquery-public-data.london_bicycles.cycle_hire` AS hire
ON
  station.id = hire.start_station_id
WHERE
  station.docks_count = station.nbEmptyDocks
GROUP BY
  station_name
ORDER BY
  num_people DESC;
```



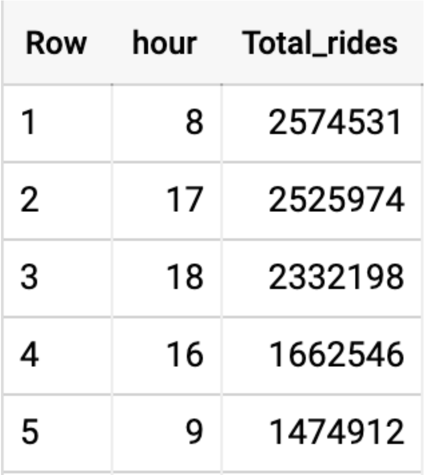
Row	station_name	num_people
1	Braham Street, Aldgate	50331
2	Museum of London, Barbican	48821
3	Queensway, Kensington Gardens	35950
4	St. Katharine's Way, Tower	31660
5	Royal Avenue 1, Chelsea	21527
6	Wapping High Street, Wapping	20153
7	Vicarage Gate, Kensington	14525
8	Campden Hill Road, Notting Hill	13778
9	Lancaster Drive, Blackwall	7854

```
  -- What are the most popular stations in the network?

SELECT
  name,
  COUNT(DISTINCT (rental_id)) AS user
FROM
  `bigquery-public-data.london_bicycles.cycle_stations` AS station
INNER JOIN
  `bigquery-public-data.london_bicycles.cycle_hire` AS hire
ON
  station.id = hire.start_station_id
WHERE
  hire.duration > 60
GROUP BY
  name
ORDER BY
  user DESC
LIMIT
  5;
```



Row	name	user
1	Belgrove Street , King's Cross	232221
2	Hyde Park Corner, Hyde Park	213237
3	Waterloo Station 3, Waterloo	199671
4	Black Lion Gate, Kensington Gardens	160252
5	Albert Gate, Hyde Park	153641

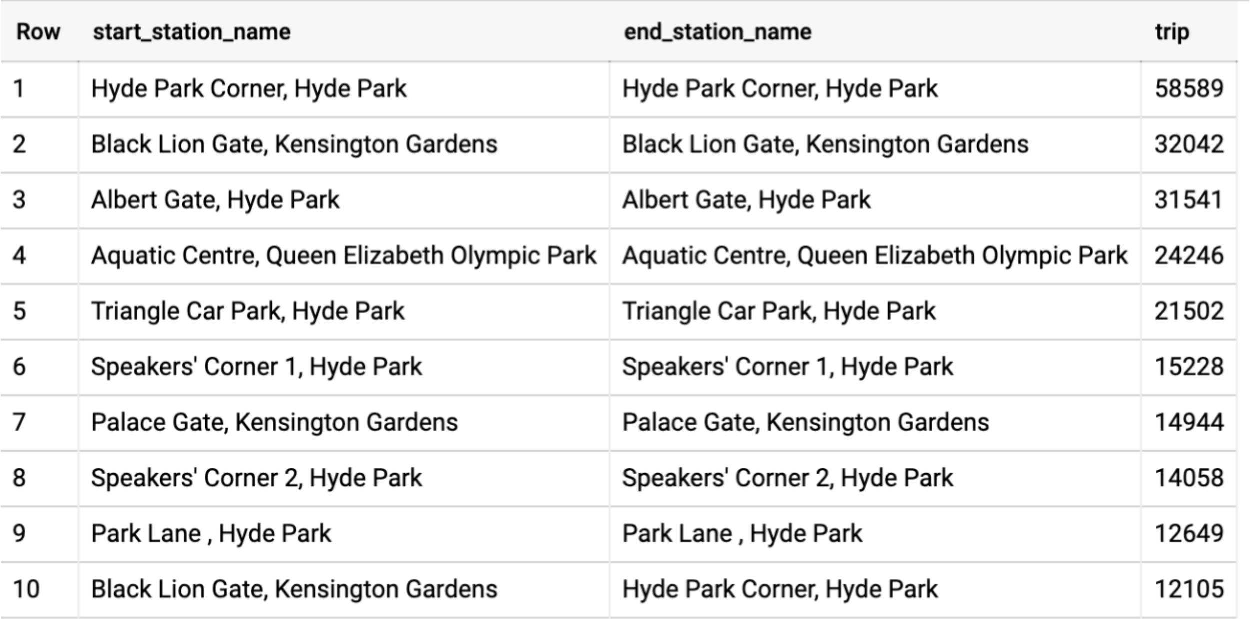
```
-- When does their usage peak?
SELECT
  --EXTRACT(DAYOFWEEK FROM start_date) AS DAY,
  EXTRACT(HOUR FROM start_date) AS hour,
  COUNT(rental_id) AS Total_rides
FROM `bigquery-public-data.london_bicycles.cycle_hire`
WHERE duration > 60
GROUP BY hour
ORDER BY Total_rides DESC;
```



Row	hour	Total_rides
1	8	2574531
2	17	2525974
3	18	2332198
4	16	1662546
5	9	1474912

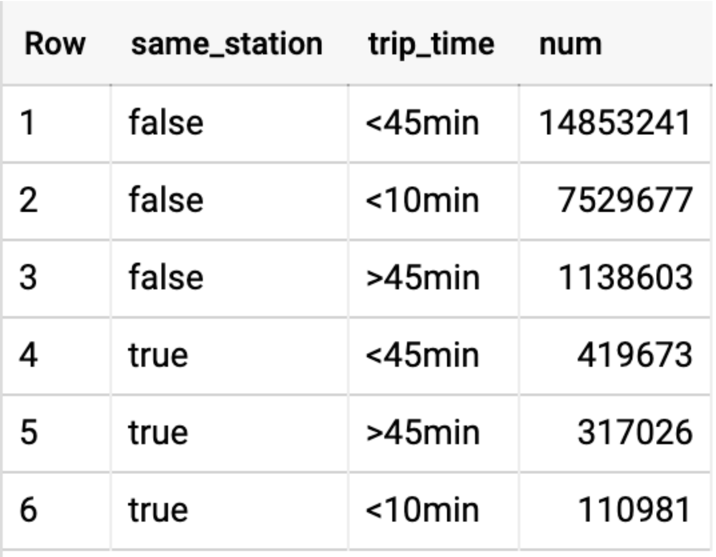
```
-- What are the most popular trips in the network?
SELECT
  start_station_name,
  end_station_name,
  COUNT(*) AS trip
FROM
  `bigquery-public-data.london_bicycles.cycle_hire`
WHERE duration > 60
GROUP BY
```

```
  start_station_name,
  end_station_name
ORDER BY
  trip DESC
LIMIT
  10;
```



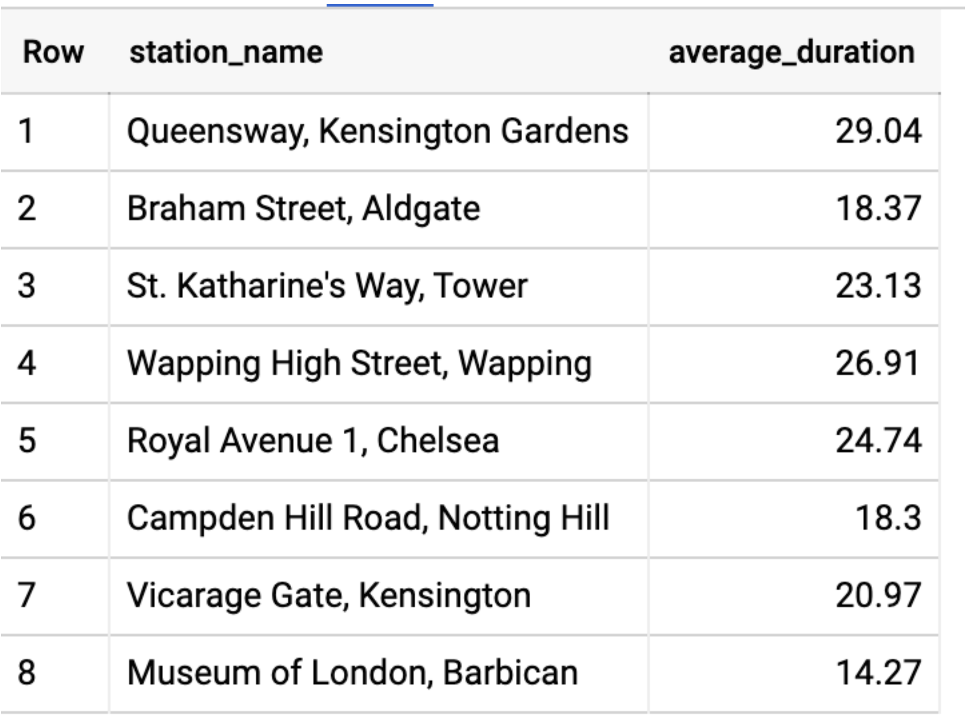
Row	start_station_name	end_station_name	trip
1	Hyde Park Corner, Hyde Park	Hyde Park Corner, Hyde Park	58589
2	Black Lion Gate, Kensington Gardens	Black Lion Gate, Kensington Gardens	32042
3	Albert Gate, Hyde Park	Albert Gate, Hyde Park	31541
4	Aquatic Centre, Queen Elizabeth Olympic Park	Aquatic Centre, Queen Elizabeth Olympic Park	24246
5	Triangle Car Park, Hyde Park	Triangle Car Park, Hyde Park	21502
6	Speakers' Corner 1, Hyde Park	Speakers' Corner 1, Hyde Park	15228
7	Palace Gate, Kensington Gardens	Palace Gate, Kensington Gardens	14944
8	Speakers' Corner 2, Hyde Park	Speakers' Corner 2, Hyde Park	14058
9	Park Lane , Hyde Park	Park Lane , Hyde Park	12649
10	Black Lion Gate, Kensington Gardens	Hyde Park Corner, Hyde Park	12105

```
  -- Are there differences in the types of rides that people take?
SELECT
  CASE
    WHEN start_station_id = end_station_id THEN TRUE
  ELSE
  FALSE
END
  AS same_station,
  CASE
    WHEN duration<600 THEN '<10min'
    WHEN duration<2700 THEN '<45min'
    WHEN duration>=2700 THEN ">45min"
END
  AS trip_time,
  COUNT(*) AS num
FROM
  `bigquery-public-data.london_bicycles.cycle_hire`
GROUP BY
  same_station,
  trip_time
ORDER BY
  num DESC;
```



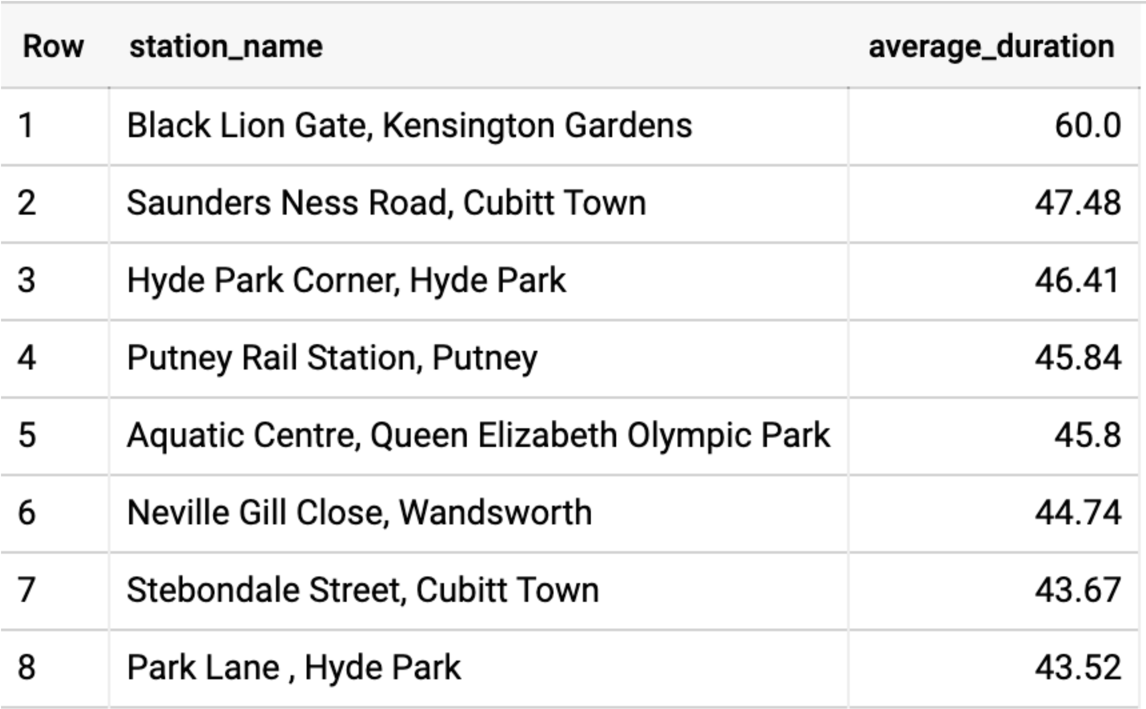
Row	same_station	trip_time	num
1	false	<45min	14853241
2	false	<10min	7529677
3	false	>45min	1138603
4	true	<45min	419673
5	true	>45min	317026
6	true	<10min	110981

```
  -- Is there a pattern in the types of stations that are empty?

SELECT
  station.name AS station_name,
  ROUND(AVG(hire.duration)/60,2) AS average_duration
FROM
  `bigquery-public-data.london_bicycles.cycle_stations` AS station
INNER JOIN
  `bigquery-public-data.london_bicycles.cycle_hire` AS hire
ON
  station.id = hire.start_station_id
WHERE
  station.docks_count = station.nbEmptyDocks
GROUP BY
  station_name;
```



Row	station_name	average_duration
1	Queensway, Kensington Gardens	29.04
2	Braham Street, Aldgate	18.37
3	St. Katharine's Way, Tower	23.13
4	Wapping High Street, Wapping	26.91
5	Royal Avenue 1, Chelsea	24.74
6	Campden Hill Road, Notting Hill	18.3
7	Vicarage Gate, Kensington	20.97
8	Museum of London, Barbican	14.27

```
---------------------------------------
-- NOW CHECKING FOR STATIONS THAT ARE NOT EMPTY
---------------------------------------
SELECT
  station.name AS station_name,
  ROUND(AVG(hire.duration)/60,2) AS average_duration
FROM
  `bigquery-public-data.london_bicycles.cycle_stations` AS station
INNER JOIN
  `bigquery-public-data.london_bicycles.cycle_hire` AS hire
ON
  station.id = hire.start_station_id
WHERE
  station.docks_count > station.nbEmptyDocks
GROUP BY
  station_name
ORDER BY
  average_duration desc
Limit 8;
```



Row	station_name	average_duration
1	Black Lion Gate, Kensington Gardens	60.0
2	Saunders Ness Road, Cubitt Town	47.48
3	Hyde Park Corner, Hyde Park	46.41
4	Putney Rail Station, Putney	45.84
5	Aquatic Centre, Queen Elizabeth Olympic Park	45.8
6	Neville Gill Close, Wandsworth	44.74
7	Stebondale Street, Cubitt Town	43.67
8	Park Lane , Hyde Park	43.52