**Data Management and SQL Final**

Emily Gomez

**Section 1: Theory Questions**

1. A data moat is a competitive advantage based on a company’s data. To have a data moat factors to consider are having data that nobody else has the resources to access, is usually large, is protected (not open source), is up to date and ultimately meaningful. Technology is increasingly accessible and the ability to imitate is growing, but data is something that has to be accessed or collected and can separate businesses for longer and more efficiently. Bad data is harder to gain good insight or production from.
2. 3 roles in a data team are data scientist, data analyst, and data engineer. A data engineer ensures that data is in a usable and accessible format for others in the pipeline. A data scientist then takes that data, cleans it, and builds models. Data scientists are usually behind machine learning and models. The data analysts will then take those models and the data to gain insight. The data analyst links the data/ numbers and the business to drive action. Data analysts need to be efficient communicators. Hard technical skills decrease in that order as well.
3. My opinion on current data privacy laws is that the main issue is not in the laws themselves but in people’s knowledge of them. Many people do not realize what is collected about them or that they have power to request it be deleted or to limit collection even though the laws are in place for their protection. I think the laws are adequate for now seeing as I cannot think of what should change. It will be a constantly evolving field as new technologies are created and others broken down.
4. The difference between where and having in SQL is ‘where’ does not need a ‘group by’ clause but ‘having’ does. ‘Having’ allows for aggregations whereas ‘where’ does not. ‘Where’ goes row by row and ‘having’ takes a group.
5. There are many employees working in most offices. While usually not, there is a possibility that one employee may have multiple offices. When each employee has one office such as a local law firm it is employees M:1 office. When employees have more than one office such as doctors working in a network of hospitals, it would be M:M.

**Section 2: ER Diagrams and Design**

1. **Students to Classes**

A screenshot of a cell phone

Description automatically generated

* Add ‘Registration’ table to connect each student with every class taken and each class with every student registered.

1. **Customers to Products**

**A screenshot of a social media post

Description automatically generated**

* Add ‘Sales’ table to connect ‘Customers’ with ‘Products’.

1. **Library Reservation System**

A screenshot of a cell phone

Description automatically generated

* ‘Reservations’ links ‘Borrowers’, ‘Libraries’, and ‘Items’. ‘Inventories’ links ‘Items’, ‘Libraries’, and ‘Content’. ‘Items’ is also linked to ‘Content’.

**Section 3: Business Report**

Biking is a major mode of transportation in the Bay Area. It is an essential component to the lifestyle and culture in the Bay Area. Due to this, it is important to monitor trends and needs that arise.

While analyzing 1,947,419 trips about 1% were excluded due to extreme trip durations lasting between 3 to 200 days. These are considered outliers and are thought to come from complications or problems in the return process.

Since San Francisco is largely a commuter working city, some of the most popular stations to start and end trips at are stations around CalTrain, Embarcadero, and Market Street below the Financial District. Trip counts are highest within San Francisco, which aligns well with SF having the highest bike/trip ratio of 17.2. San Francisco also has the lowest average station emptiness (capacity/bikes available). East Bay, specifically Berkeley, Emeryville, and Oakland, has the highest average emptiness and some of the lowest number of bikes per trip. Excluding the stations not installed or not renting, there are 13 stations with a capacity of 0 or 0 bikes available in the Bay Area.

Trip demand during the week is higher than it is on weekends with Sunday usually having the lowest numbers. The peak trip start times during the week are between 7-10 AM and 4-7 PM as people travel to and from work. On weekends trips peak in the early afternoon with the highest volume being between 12-4 PM. The average number of trips per day over the last 30 logged days was 4,372 trips with a max of 6,377 and a minimum of 1,294 trips. Trip duration averages 12 minutes with some being as short as 1 minute and as long as nearly 3 hours (probably the leisurely riders).

This usage wears on bikes over time. Some have as many as 3,394 and as few as 1 trip logged on them. With nearly 4,000 bikes across the Bay Area, usage varies greatly. Total duration of all trips taken on a bike, excluding the 1% mentioned earlier, is as high as 5,464 hours and as low as 6 minutes.

With such variation bikes are used and moved at different rates. San Francisco’s bikes and docking stations are at risk for the most wear and could need the most maintenance. It also has the most disabled bikes. San Francisco has the highest number of bikes per trip and lowest emptiness though, so this is not an immediate problem but should be monitored. However, Berkeley has the most disabled docks. This may be a contributing factor to Berkeley having the fewer bikes per trip.

While emptiness could pose a problem in bike availability, there also needs to be a balance so that people can return bikes to that location. Fixing the disabled bikes and docks would help to increase product availability and access. Also, the only payment method logged was key/credit card. Usage peaks with riders who were born between 1980 and 1993. If another payment method was implemented that targeted middle aged adults, it may lead to more business.

Being that many workers will start and end around the same stations to and from work should help maintain demand. However, bikes need to start in those positions for this to be effective. This would help regulate ‘traffic patterns’ for the bikes on weekdays. On Sunday evenings, the Bay Area could use this data to reset the bikes to the most popular starting stations for Monday morning.

**Section 4: Dashboard**

A screenshot of a social media post

Description automatically generatedA screenshot of a social media post

Description automatically generated

**Appendix / SQL Queries:**

A screenshot of a computer

Description automatically generated

**Popular Trips:**

Select COUNT(trip\_id) as trips, start\_station\_name as origin, end\_station\_name as destination

from bigquery-public-data.san\_francisco\_bikeshare.bikeshare\_trips

GROUP BY start\_station\_name, end\_station\_name

order by trips desc

limit 3;

**Popular Start and End Stations:**

Select COUNT(trip\_id) as trips, start\_station\_name as origin

from bigquery-public-data.san\_francisco\_bikeshare.bikeshare\_trips

GROUP BY start\_station\_name

order by trips desc

limit 1;

SELECT trips.end\_station\_name, count(trip\_id) as Pop\_Start

FROM bigquery-public-data.san\_francisco\_bikeshare.bikeshare\_trips as trips

group by trips.end\_station\_name

order by Pop\_Start desc

limit 1;

**Fill Rates:**

SELECT trips.end\_station\_name, count(trip\_id) as Pop\_Start

FROM bigquery-public-data.san\_francisco\_bikeshare.bikeshare\_trips as trips

group by trips.end\_station\_name

order by Pop\_Start desc;

**Weekday Peak Usage:**

select hour, count(trip\_id) as trips

from (

SELECT

EXTRACT(DAYOFWEEK from start\_date) as weekday,

EXTRACT(HOUR from start\_date) as hour,

trip\_id

FROM bigquery-public-data.san\_francisco\_bikeshare.bikeshare\_trips

where EXTRACT(DAYOFWEEK from start\_date) in (2, 3, 4, 5, 6))

group by hour

order by trips desc;

**Weekend Peak Usage:**

select hour, count(trip\_id) as trips

from (

SELECT

EXTRACT(DAYOFWEEK from start\_date) as weekday,

EXTRACT(HOUR from start\_date) as hour,

trip\_id

FROM bigquery-public-data.san\_francisco\_bikeshare.bikeshare\_trips

where EXTRACT(DAYOFWEEK from start\_date) in (1,7))

group by hour

order by trips desc;

**Stations Without Capacity:**

SELECT info.name, is\_renting, capacity, num\_bikes\_available

FROM bigquery-public-data.san\_francisco\_bikeshare.bikeshare\_station\_status as status

full join bigquery-public-data.san\_francisco\_bikeshare.bikeshare\_station\_info as info

on status.station\_id = info.station\_id

WHERE is\_installed = TRUE and (num\_bikes\_available = 0 or capacity = 0 or is\_renting = FALSE)

**Trip Duration:**

Select Round(MIN(duration),2) as MIN, ROUND(AVG(duration),2) AS AVG, Round(MAX(duration),2) AS MAX

FROM (

Select TIMESTAMP\_DIFF(end\_date, start\_date, MINUTE) AS duration

from bigquery-public-data.san\_francisco\_bikeshare.bikeshare\_trips

order by duration asc

limit 1927945)

where duration >= 0;

**Bike Usage:**

SELECT bike\_number, ROUND(SUM(trip\_duration)/60,2) as total\_duration, ROUND(AVG(trip\_duration/60),2) as Avg\_duration, COUNT(trip\_id) as trips

FROM

(SELECT bike\_number , TIMESTAMP\_DIFF(end\_date , start\_date, MINUTE) AS trip\_duration, trip\_id

FROM bigquery-public-data.san\_francisco\_bikeshare.bikeshare\_trips

where TIMESTAMP\_DIFF(end\_date , start\_date, MINUTE) <= 1752677)

group by bike\_number

order by total\_duration desc;

**Ave Daily Trips:**

select ROUND(AVG(trips),0) AS AVG, MAX(trips) AS MAX, MIN(trips) AS MIN,

from

(SELECT EXTRACT(Date from start\_date) as date, count(\*) as trips

from bigquery-public-data.san\_francisco\_bikeshare.bikeshare\_trips

group by date

order by date desc

limit 30) as day\_trips;

**Daily Trips: Last 30 Logged Days:**

SELECT EXTRACT(Date from start\_date) as date, count(\*) as trips

from bigquery-public-data.san\_francisco\_bikeshare.bikeshare\_trips

group by date

order by date desc

limit 30;

**Trips per Station by Region:**

SELECT count(distinct(start\_station\_latitude)) as stations, count(trip\_id) as trips, round(count(trip\_id)/count(distinct(start\_station\_latitude)),2) as ratio, region.name

FROM bigquery-public-data.san\_francisco\_bikeshare.bikeshare\_trips as trips

full join bigquery-public-data.san\_francisco\_bikeshare.bikeshare\_station\_info as info

on trips.start\_station\_latitude = info.lat

full join bigquery-public-data.san\_francisco\_bikeshare.bikeshare\_regions as region

using(region\_id)

group by region.name

having count(start\_station\_latitude) >0

order by stations/trips asc

**Disabled Docks by Region:**

SELECT sum(num\_docks\_disabled), region.name

FROM bigquery-public-data.san\_francisco\_bikeshare.bikeshare\_station\_status as status

full join bigquery-public-data.san\_francisco\_bikeshare.bikeshare\_station\_info as info

using(station\_id)

full join bigquery-public-data.san\_francisco\_bikeshare.bikeshare\_regions as region

using(region\_id)

group by region.name

**Disabled Bikes by Region:**

SELECT sum( num\_bikes\_disabled ), region.name

FROM bigquery-public-data.san\_francisco\_bikeshare.bikeshare\_station\_status as status

full join bigquery-public-data.san\_francisco\_bikeshare.bikeshare\_station\_info as info

using(station\_id)

full join bigquery-public-data.san\_francisco\_bikeshare.bikeshare\_regions as region

using(region\_id)

group by region.name

**Rental Methods:**

select rental\_methods, count(trip\_id) as trip

FROM bigquery-public-data.san\_francisco\_bikeshare.bikeshare\_trips as trips

full join bigquery-public-data.san\_francisco\_bikeshare.bikeshare\_station\_info as info

on info.lat = trips. start\_station\_latitude

group by rental\_methods

order by trip desc

**Age Distribution:**

SELECT member\_birth\_year, count(member\_birth\_year) as trips

FROM bigquery-public-data.san\_francisco\_bikeshare.bikeshare\_trips

group by member\_birth\_year

order by member\_birth\_year desc;

**Station Avg Emptiness:**

select station, avg(emptiness) as avg\_emptiness, region

from(

select info.name as station, region.name as region, avg(capacity) as capacity, avg(num\_docks\_available) as docks\_avail, avg(num\_bikes\_available) as bikes\_avail, EXTRACT(HOUR from start\_date) as hour, round((avg(num\_docks\_available)/avg(capacity)),2) as emptiness

FROM bigquery-public-data.san\_francisco\_bikeshare.bikeshare\_station\_status as status

full join bigquery-public-data.san\_francisco\_bikeshare.bikeshare\_station\_info as info

on status.station\_id = info.station\_id

full join bigquery-public-data.san\_francisco\_bikeshare.bikeshare\_trips as trips

on trips.start\_station\_latitude = info.lat

full join bigquery-public-data.san\_francisco\_bikeshare.bikeshare\_regions as region

on info.region\_id = region.region\_id

where is\_installed = TRUE and is\_renting = TRUE AND capacity > 0

group by info.name, hour, is\_installed, region)

group by station, region

order by avg\_emptiness desc

**Top 1%:**

Select TIMESTAMP\_DIFF(end\_date, start\_date, MINUTE) AS duration

from bigquery-public-data.san\_francisco\_bikeshare.bikeshare\_trips

order by duration desc

limit 19000;

**Full Stations:**

SELECT info.name, capacity, num\_docks\_available, num\_bikes\_available

FROM bigquery-public-data.san\_francisco\_bikeshare.bikeshare\_station\_status as status

full join bigquery-public-data.san\_francisco\_bikeshare.bikeshare\_station\_info as info

on status.station\_id = info.station\_id

WHERE is\_installed = TRUE and is\_renting = TRUE AND capacity> 0 and (num\_docks\_available <= 0.2\*capacity)

order by num\_bikes\_available desc

**Bike availability by Region:**

select hour, count(trip\_id) as trips, sum(num\_bikes\_available) as bikes\_available, round(sum(num\_bikes\_available)/count(trip\_id),2) as bike\_trip, region

from (

SELECT

EXTRACT(DAYOFWEEK from start\_date) as weekday,

EXTRACT(HOUR from start\_date) as hour,

trip\_id,

num\_bikes\_available,

region.name as region

FROM bigquery-public-data.san\_francisco\_bikeshare.bikeshare\_trips as trips

full join bigquery-public-data.san\_francisco\_bikeshare.bikeshare\_station\_info as info

on trips. start\_station\_latitude = info.lat

full join bigquery-public-data.san\_francisco\_bikeshare.bikeshare\_station\_status as status

using (station\_id)

full join bigquery-public-data.san\_francisco\_bikeshare.bikeshare\_regions as region

using (region\_id)

where EXTRACT(DAYOFWEEK from start\_date) in (2, 3, 4, 5, 6))

where region is not null and num\_bikes\_available is not null

group by hour, region

order by bike\_trip asc;