Bay Area Bikeshare Analysis

October 3, 2019

W205 | Data Engineering: Project 1

Prepared by Navya Sandadi

1. Introduction

Bikeshare programs are an increasingly popular means for traffic congested cities to provide alternate transportation to their residents and visitors. They rent bikes to customers on a short term, thereby providing cheap, fast and on-demand service to users. However, not everyone is ready to switch from a car to a bike yet.

Ford GoBike, the company running Bay Area Bikeshare, is trying to increase ridership, and wants to offer deals through the mobile app to do so. This report looks at possible factors that would encourage more people to switch to bikeshare for commute.

Currently the company offers the following deals:

- a flat price for a single-way trip
- a day pass that allows unlimited 30-minute rides for 24 hours
- an annual membership

On studying the data, it was found that most Bay Area bikeshare usage falls into two categories:

- Weekday commuter trips during rush hour
- Mid-day and weekend leisure travel

After a thorough analysis, it was observed that there are a few opportunities to increase revenue and ridership. The report makes recommendations to management on capitalizing these opportunities by modifying existing offers to address the customers' needs more effectively.

2. Data Source

This report uses data from Google Big Query's public dataset 'san_francisco'. Data from the following tables was analyzed:

- bikeshare_trips
- bikeshare_stations
- bikeshare_status

bikeshare_trips has 983,648 rows, where each row provides details of a trip. This table contains data for trips that took place between 2013-08-29 and 2016-08-31.

bikeshare_stations contains data for 74 stations.

bikeshare_status has 107,501,619 rows. It provides the status of bikes available and docks available at different stations several times a day in the duration between 2013-08-29 and 2016-08-31.

3. Objective

This report specifically seeks to answer the following questions:

- What are the 5 most popular trips that you would call "commuter trips"?
- What are your recommendations for offers (justify based on your findings)?

3.1 What is a "commuter trip"?

A commuter trip can be characterized as one which:

- occurs on weekdays during rush hour between 7-9 am and 4-6 pm
- starts and ends at different stations (i.e, start station is not equal to end station)
- has start and end stations within the same city
- is ususally ridden by subscribers

4. Exploratory Data Analysis

On analyzing the rental activity of Subscribers and Customers (based on the day of the week labeled as weekday with values 1-7 for Sun-Sat respectively), it was observed that Subscribers are significantly more active on weekdays (Mon-Fri) than weekends, with Tuesday being the busiest and Friday being the slowest.

On the other hand, customers are less active on weekdays and more active on the weekends, more so on Saturdays. Sunday is the slowest day of the week. Rentals by month was also analyzed (labeled as values 1-12 for Jan-Dec). The months of December, January and February showed a sharp dip in bike rentals.

The duration for each trip was also analyzed. It was found that commuter trip are between 5-25 minutes and customer trips are around 50 minutes.

Overall, subscriber use bikesharing much more than customers.

[1]:
```
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
```

4.1 Riders: Subscribers vs. Customers

There are 313,731 subscribers and 110,336 customers. 74% of the riders are subscribers (i.e), those who hold an annual membership and 26% of them are customers (i.e.), those who hold a single trip pass or 3-day membership. Assuming that most subscribers are commuters, we can conclude that a majority of riders using bikeshare are commuters.

[46]:
```
```

```
! bq query --use_legacy_sql=FALSE --format=csv 'select subscriber_type,␣
↪count(*) as count from (select *, round(duration_sec / 60.0) as␣
↪duration_min, extract(DAYOFWEEK from start_date) as start_day, extract(HOUR␣
↪from start_date) as start_hour from `bigquery-public-data.san_francisco.
↪bikeshare_trips`) where duration_min >= 10 and duration_min < 60*20 group by␣
↪subscriber_type' > clients.csv
```

```
Waiting on bqjob_r64337c9dddee1302_0000016d66f83d0f_1 … (1s) Current status:
DONE
```

[47]:
```
client_base = pd.read_csv("clients.csv")
client_base
```

[47]:

	subscriber_type	count
0	Customer	110336
1	Subscriber	313731

4.2 Bike usage by day of the week

As seen in the bar chart and table below, bike usage is the greatest on Tuesdays, closely followed by Wednesday and Thursday. Monday and Friday are slower. Saturday and Sunday are the slowest days.

[23]:
```
! bq query --use_legacy_sql=FALSE --format=csv 'select day_of_week, count(*) as␣
↪num_of_trips from (select *, round(duration_sec / 60.0) as duration_min,␣
↪extract(DAYOFWEEK from start_date) as day_of_week from `bigquery-public-data.
↪san_francisco.bikeshare_trips`) where duration_min >= 10 and duration_min <␣
↪60*20 group by day_of_week order by num_of_trips desc' > trip_freq_by_day.csv
```

```
Waiting on bqjob_rdc7db2d04f58e24_0000016d649842cb_1 … (1s) Current status:
DONE
```

[31]:
```
trip_freq_week = pd.read_csv('trip_freq_by_day.csv', index_col = "day_of_week")
```

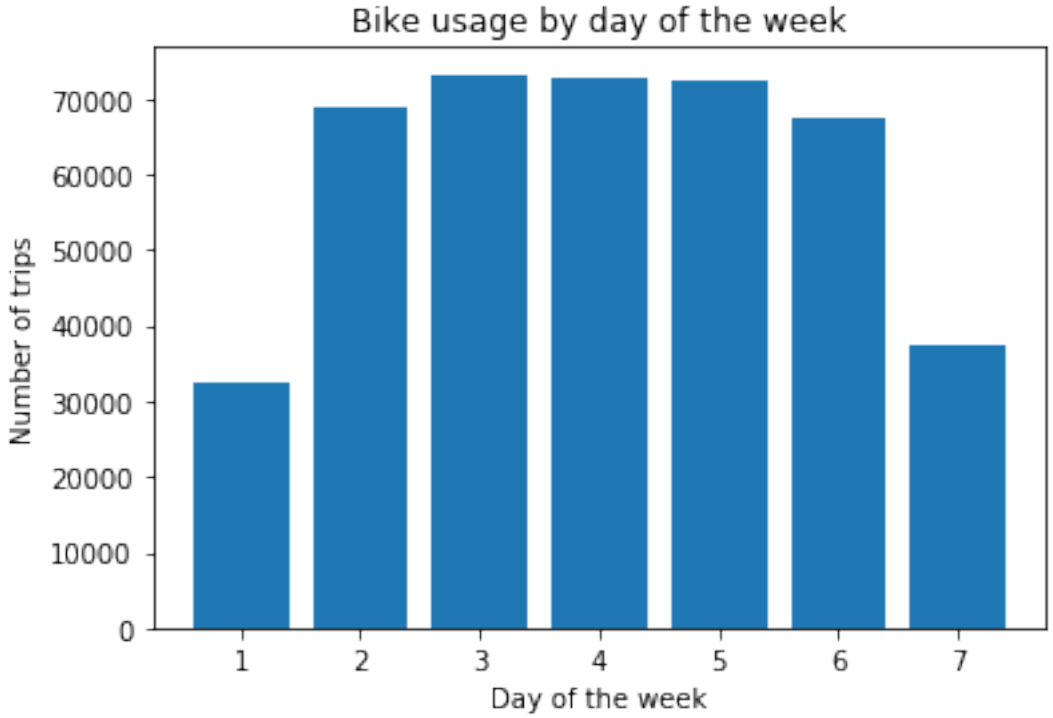
[32]:
```
# 1=Sunday, 2=Monday, 3=Tuesday, 4=Wednesday, 5=Thursday, 6=Friday, 7=Saturday
trip_freq_week
```

[32]:

day_of_week	num_of_trips
3	73180
4	72783
5	72237
2	68750
6	67482
7	37255
1	32380

[61]:
```
# Distribution of trips by day of the week
plt.bar(trip_freq_week.index, trip_freq_week.num_of_trips)
plt.xlabel('Day of the week')
plt.ylabel('Number of trips')
plt.title('Bike usage by day of the week')
```

[61]:
```
Text(0.5, 1.0, 'Bike usage by day of the week')
```



4.3 Bike usage by hour on weekdays

As seen in the line chart below, the demand for bikesharing peaks at 8 am and 5 pm. 7-9 am in the morning and 4-6 pm in the evening are hours with greatest demand.

The demand falls sharply during mid-day from 10 am to 3 pm. Since unrented bikes are a lost opportunity to make revenue, it is recommended to offer discounts to customers during mid-day on weekdays from 10 am to 3 pm as an incentive to encourage tourists to rent bikes. Also, since Monday and Friday are slower days for commuters, as seen above, offering an all day pass for customers on these days could increase ridership.

[81]:

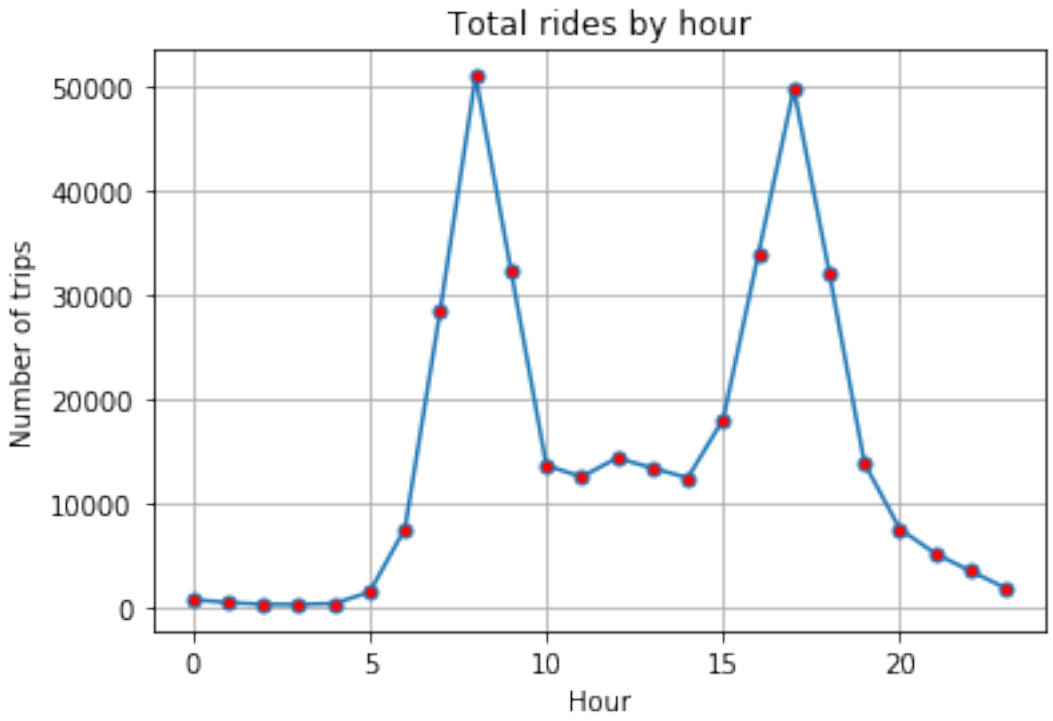
```
! bq query --use_legacy_sql=FALSE --format=csv 'select start_hour, count(*) as␣
↪trip_freq from (select *, round(duration_sec / 60.0) as duration_min,␣
↪extract(DAYOFWEEK from start_date) as start_day, extract(HOUR from␣
↪start_date) as start_hour from `bigquery-public-data.san_francisco.
↪bikeshare_trips`) where duration_min >= 10 and duration_min < 60*20 and␣
↪(start_day <> 1 and start_day <> 7) group by start_hour order by start_hour'␣
↪> rental_by_hour.csv
```

```
Waiting on bqjob_r7049ab8cbc6a01a7_0000016d64d168a9_1 … (1s) Current status:
DONE
```

[82]:
```
bike_hour = pd.read_csv('rental_by_hour.csv', index_col = 'start_hour')
```

[94]:
```
plt.plot(bike_hour.index, bike_hour.trip_freq, marker='o',␣
↪markerfacecolor='red', markersize=5)
plt.grid()
plt.xlabel('Hour')
plt.ylabel('Number of trips')
plt.title('Total rides by hour')
```

[94]:
```
Text(0.5, 1.0, 'Total rides by hour')
```



4.4 Average duration of trips across the week

As dislplayed in the table and bar chart below, the average duration for trips on a weekday are 21-25 minutes and on weekends are 48-50 minutes. Since most of the weekday trips are commuter trips, they tend to be shorter (less than 30 minutes) and since most of the trips on weekends are for leisure, they tend to be longer (closer to an hour). We can conclude that commuter/subscriber trips last less than 30 minutes and leisure/customer trips last longer. However, currently subscribers are being offered unlimited 45-minute trips and customers are being offered 30-minute trips.

It is recommended that subscribers be offered unlimited 30-minute trips on weekdays and 45-minute trips on weekends at a reduced annual membership fee. This would encourage more commuters to sign up for membership. It is also recommended to increase the customer's trip to at least 45 minutes. When people are on a leisure trip, especially in a hilly terrain like San Francisco, they could feel stressed by the 30 minute limit. A 45 to 60 minute limit for the same or slightly higher price would allow them more flexibility and freedom to enjoy their ride and have a positive experience.

[67]:
```
# Average duration of bike trips
! bq query --use_legacy_sql=FALSE --format=csv 'select day_of_week,␣
↪avg(duration_min) as avg_min from (select *, round(duration_sec / 60.0) as␣
↪duration_min, extract(DAYOFWEEK from start_date) as day_of_week from␣
↪`bigquery-public-data.san_francisco.bikeshare_trips`) where duration_min >=␣
↪10 and duration_min < 60*20 group by day_of_week order by day_of_week' >␣
↪avg_trip_dur.csv
```

```
Waiting on bqjob_r1188d640780d8066_0000016d64c0eed4_1 … (0s) Current status:
DONE
```

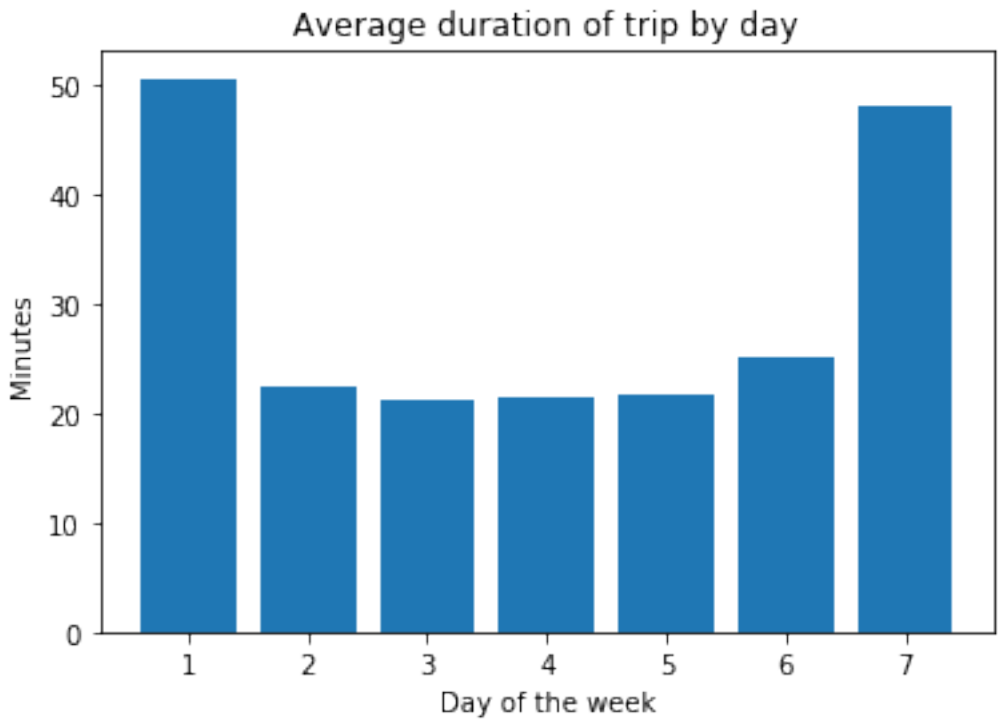
[69]:
```
avg_duration = pd.read_csv('avg_trip_dur.csv', index_col = 'day_of_week')
avg_duration
```

[69]:

day_of_week	avg_min
1	50.653860
2	22.504189
3	21.252077
4	21.420002
5	21.825311
6	25.199713
7	48.064045

[96]:
```
plt.bar(avg_duration.index, avg_duration.avg_min)
plt.xlabel('Day of the week')
plt.ylabel('Minutes')
plt.title('Average duration of trip by day')
```

[96]:
```
Text(0.5, 1.0, 'Average duration of trip by day')
```



4.5 Most common duration of commuter trips

From the graph and table below, it is evident that most of the commuter trips are 5 to 15 minutes long. Thus, management can create 2 kinds of annual memberships, one with unlimited 15-minute trips and one with unlimited 30-minute trips, priced accordingly based on the minutes. This would target individual commuter needs more effectively. However, it is recommended to keep unlimited 45-minute trips on weekends in both memberships as there is very low ridership during the weekends anyways. It could serve an incentive for loyal subscribers to be more active by biking in the weekends.

[16]:
```
! bq query --use_legacy_sql=FALSE --format=csv 'select duration_min, count(*)␣
↪as trip_freq from (select *, round(duration_sec / 60.0) as duration_min,␣
↪extract(DAYOFWEEK from start_date) as start_day, extract(HOUR from␣
↪start_date) as start_hour from `bigquery-public-data.san_francisco.
↪bikeshare_trips`) where duration_min >= 5 and duration_min < 60*20 and␣
↪((start_hour >= 7 and start_hour <= 9) or (start_hour >= 16 and start_hour␣
↪<= 18)) and start_station_id <> end_station_id and start_day <> 1 and␣
↪start_day <> 7 group by duration_min order by duration_min' >␣
↪commuter_duration.csv
```

```
Waiting on bqjob_r506124ea08586558_0000016d66ded772_1 … (0s) Current status:
DONE
```

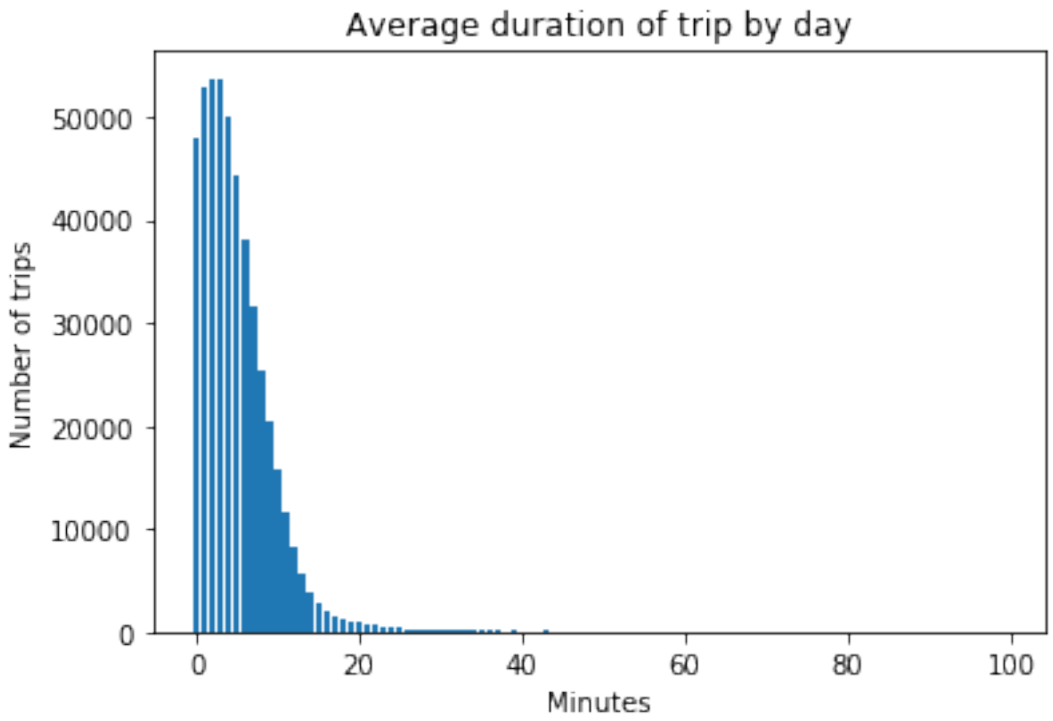
[44]:
```
commuter_min = pd.read_csv("commuter_duration.csv")
commuter_min.head(16)
```

[44]:

	duration_min	trip_freq
0	5.0	47854
1	6.0	52767
2	7.0	53596
3	8.0	53737
4	9.0	49934
5	10.0	44259
6	11.0	38096
7	12.0	31615
8	13.0	25378
9	14.0	20450
10	15.0	15725
11	16.0	11509
12	17.0	8180
13	18.0	5703
14	19.0	3935
15	20.0	2874

[42]:
```
plt.bar(commuter_min.index, commuter_min.trip_freq)
plt.xlabel('Minutes')
plt.ylabel('Number of trips')
plt.title('Average duration of trip by day')
```

[42]:
```
Text(0.5, 1.0, 'Average duration of trip by day')
```



4.6 Comparison of usage between customers and subscribers (weekday vs. weekend)

Most subscribers ride during the weekdays (82% subscribers vs. 18% customers) and most customers ride during the weekend (67.5% customers vs. 32.5% subscribers).

[60]:
```
! bq query --use_legacy_sql=FALSE --format=csv 'select subscriber_type,␣
↪count(*) as weekday_trips from (select *, round(duration_sec / 60.0) as␣
↪duration_min, extract(DAYOFWEEK from start_date) as start_day from␣
↪`bigquery-public-data.san_francisco.bikeshare_trips`) where duration_min >=␣
↪10 and duration_min < 60*20 and start_day <> 1 and start_day <> 7 group by␣
↪subscriber_type' > weekday.csv

! bq query --use_legacy_sql=FALSE --format=csv 'select subscriber_type,␣
↪count(*) as weekday_trips from (select *, round(duration_sec / 60.0) as␣
↪duration_min, extract(DAYOFWEEK from start_date) as start_day from␣
↪`bigquery-public-data.san_francisco.bikeshare_trips`) where duration_min >=␣
↪10 and duration_min < 60*20 and (start_day = 1 or start_day = 7) group by␣
↪subscriber_type' > weekend.csv
```

```
Waiting on bqjob_r163092f1e23be0f6_0000016d671c3945_1 … (1s) Current status:
DONE
```

[61]:
```
weekday = pd.read_csv("weekday.csv")
weekday
```

[61]:
	subscriber_type	weekday_trips
0	Customer	63315
1	Subscriber	291117

[62]:
```
weekend = pd.read_csv("weekend.csv")
weekend
```

[62]:
	subscriber_type	weekday_trips
0	Customer	47021
1	Subscriber	22614

4.7 Commuters who are not subscribers

Out of 481,239 commuters, 94% of them are subscribers. There is opportunity to convert the remaining 6% of the commuters to subscribers. However, offering only an annual membership may hinder conversion as people tend to shy away from a year long commitment. Offering short term memberships such as weekly, monthly and quarterly passes will encourage such commuters to sign up for a membership. Once they have a positive experience, they might possibly upgrade to a longer membership.

Moreover, since ridership is very sparse is winter (refer to graph below), offering riders short term options will make them feel more valued as they don't have to lose money during those months, or during summer vacations. It is recommended to price the memberships accordingly based on the duration. For example, a monthly membership will be more expensive than an annual membership but cheaper than a weekly membership.

[54]:
```
! bq query --use_legacy_sql=FALSE --format=csv 'select subscriber_type,␣
↪count(*) as trip_freq from (select *, round(duration_sec / 60.0) as␣
↪duration_min, extract(DAYOFWEEK from start_date) as start_day, extract(HOUR␣
↪from start_date) as start_hour from `bigquery-public-data.san_francisco.
↪bikeshare_trips`) where duration_min >= 5 and duration_min < 60*20 and␣
↪((start_hour >= 7 and start_hour <= 9) or (start_hour >= 16 and start_hour␣
↪<= 18)) and start_station_id <> end_station_id and start_day <> 1 and␣
↪start_day <> 7 group by subscriber_type' > commuter_membership.csv
```

```
Waiting on bqjob_r13c0fe9fcc96353d_0000016d670459e1_1 … (0s) Current status:
DONE
```

[55]:
```
commuter_type = pd.read_csv("commuter_membership.csv")
commuter_type
```

[55]:
	subscriber_type	trip_freq
0	Subscriber	454793
1	Customer	26446

4.8 Bike usage by month

There is a sharp fall in ridership from December to February, possibly due to the cold weather and holiday season. Management should consider offering a heavily discounted winter pass to encourage riders to use bikesharing during these months.

[8]:
```
# August is busiest and December is slowest
! bq query --use_legacy_sql=FALSE --format=csv 'select month, count(*) as␣
 ↪trip_freq from (select *, round(duration_sec / 60.0) as duration_min,␣
 ↪extract(DAYOFWEEK from start_date) as start_day, extract(HOUR from␣
 ↪start_date) as start_hour, extract(MONTH from start_date) as month from␣
 ↪`bigquery-public-data.san_francisco.bikeshare_trips`) where duration_min >=␣
 ↪10 and duration_min < 60*20 group by month order by month' > month_usage.csv
```

```
Waiting on bqjob_r7e0df57ac36381f0_0000016d66b242ed_1 … (0s) Current status:
DONE
```

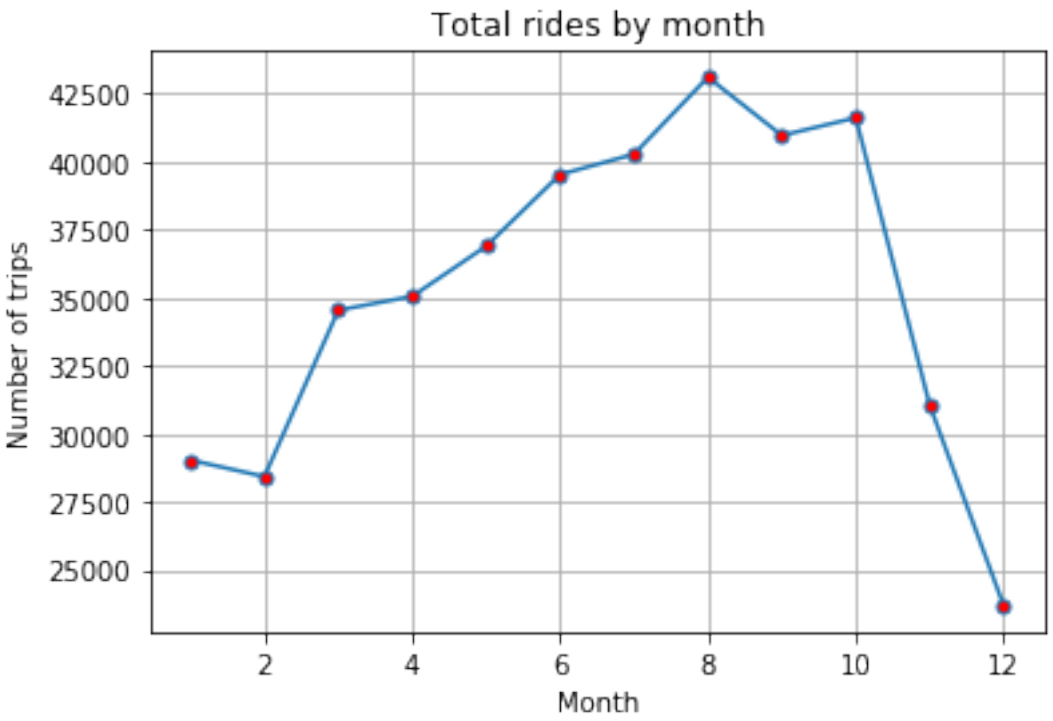
[9]:
```
usage_by_month = pd.read_csv("month_usage.csv", index_col = "month")
usage_by_month
```

[9]:

month	trip_freq
1	29033
2	28429
3	34534
4	35042
5	36884
6	39512
7	40258
8	43084
9	40946
10	41595
11	31051
12	23699

[10]:
```
plt.plot(usage_by_month.index, usage_by_month.trip_freq, marker='o',␣
 ↪markerfacecolor='red', markersize=5)
plt.grid()
plt.xlabel('Month')
plt.ylabel('Number of trips')
plt.title('Total rides by month')
```

[10]:
```
Text(0.5, 1.0, 'Total rides by month')
```



4.9 Five most popular commuter trips

The following table gives the 5 most popular commuter trips. All the stations are within San Francisco city.

[57]:
```
! bq query --use_legacy_sql=FALSE --format=csv 'select start_station_name,␣
↪end_station_name, count(*) as trip_count from (select *, round(duration_sec /
↪ 60.0) as duration_min, extract(DAYOFWEEK from start_date) as start_day,␣
↪extract(HOUR from start_date) as start_hour from `bigquery-public-data.
↪san_francisco.bikeshare_trips`) where duration_min >= 10 and duration_min <␣
↪60*20 and ((start_hour >= 7 and start_hour <= 9) or (start_hour >= 16 and␣
↪start_hour <= 18))and start_station_id <> end_station_id and start_day <> 1␣
↪and start_day <> 7 group by start_station_name, end_station_name order by␣
↪trip_count desc LIMIT 5' > pop_trips.csv
```

```
Waiting on bqjob_r1bc71902e65669f4_0000016d671067db_1 … (1s) Current status:
DONE
```

[58]:
```
popular_trips = pd.read_csv("pop_trips.csv")
popular_trips
```

[58]:
```
                          start_station_name  \
0  San Francisco Caltrain (Townsend at 4th)
1                          Steuart at Market
```

```
2      Harry Bridges Plaza (Ferry Building)
3  San Francisco Caltrain (Townsend at 4th)
4  San Francisco Caltrain (Townsend at 4th)

                              end_station_name  trip_count
0           Harry Bridges Plaza (Ferry Building)        4074
1       San Francisco Caltrain (Townsend at 4th)        3987
2       San Francisco Caltrain (Townsend at 4th)        3516
3                            Steuart at Market        3030
4  Temporary Transbay Terminal (Howard at Beale)        2885
```

5. Summary of Findings and Recommendations

- What are the 5 most popular trips that you would call "commuter trips"?
  - Commuter trips are characterised as:
    - occurs on weekdays during rush hour between 7-9 am and 4-6 pm
    - starts and ends at different stations (i.e, start station is not equal to end station)
    - has start and end stations within the same city
  - The 5 most popular commuter trips are as follows:

[59]:
```
popular_trips
```

[59]:
```
                         start_station_name  \
0  San Francisco Caltrain (Townsend at 4th)
1                         Steuart at Market
2      Harry Bridges Plaza (Ferry Building)
3  San Francisco Caltrain (Townsend at 4th)
4  San Francisco Caltrain (Townsend at 4th)

                              end_station_name  trip_count
0           Harry Bridges Plaza (Ferry Building)        4074
1       San Francisco Caltrain (Townsend at 4th)        3987
2       San Francisco Caltrain (Townsend at 4th)        3516
3                            Steuart at Market        3030
4  Temporary Transbay Terminal (Howard at Beale)        2885
```

- What are your recommendations for offers (justify based on your findings)?
  - The demand for bikesharing falls sharply during mid-day from 10 am to 3 pm on weekdays. Since unrented bikes is lost opportunity to make revenue, it is recommended to offer discounts during 10 am to 3 pm on weekdays as an incentive to encourage tourists to rent bikes.
  - Monday and Friday are slower days for commuters/subscribers, thus, offering an all day pass for customers on these days could increase ridership. It will encourage recreational riding when the commuters are less likely to be using the bikes.
  - Most commuter trips are less than 30 minutes and most of the leisure trips by visitors are closer to 60 minutes, it is recommended to offer 2 types of annual memberships for 15 and 30 minute unlimited rides and offer a cheaper annual membership while increasing the time limit for 24-hour or 3-day members to 45 minutes, to attract more ridership.

When people are on a leisure trip, especially in a hilly terrain like San Francisco, they can feel stressed by the 30 minute limit. A 45 to 60 minute limit for the same or slightly higher price will give them more flexibility and freedom to enjoy their ride.

- There are some riders who take trips characteristic of a commuter trip but do not have an annual membership. They use a day pass or the 3-day membership. To help convert such riders from customers to subscribers, it is recommended to introduce a more flexible and short term memberships such as a weekly, monthly pass or quarterly pass, so that people do not shy away from becoming a subscriber just because of the year long commitment of an annual membership.
- Moreover, there are fewer commuters from December to February because of the holidays and the cold weather in Bay Area. The ones who choose not to ride in winter will not have to lose out on money by subscribing to an annual membership. Such flexibility will make riders feel more valued and the company will in turn earn their loyalty.