Exploratory Data Analysis(EDA) on Bike Sharing in San Jose

EDA high level summary

- Description of variables and basic statistics
- Preliminary steps
  - Handle NaN in the dataset
- Calculations of summary statistics and metrics
  - User group analysis : Registered vs Unregistered user ratio in the dataset
  - Bike ride count per member and bike type
  - Geographical analysis on bike ride with map
  - Number of bike rides on weekdays/weekends for Member(registered) vs. Casual(unregistered)
  - Create new Metric: Daily Bike Occupancy Index
  - Bike usage trend for 1 month
  - Dailye bike ride demand trend per bike type
  - Hourly trip data for each user type
  - Ride time duration related data analysis
  - Ride distance related data analysis
- Conclusions and next step (Remaining questions)

Description of variables and basic statistics

The data is loaded to the dataframe based on Python pandas library

- 13 columns of variables in the bike ride data.
  - Categorical feature : 9 variables (features)
    - ride_id : Unique ride event id
    - rideable_type : types of bikes(electric, classic ,and docked bike)
    - start_station_name : name of station where bike ride starts
    - start_station_id : start station identification code
    - end_station_name : name of station where bike ride ends
    - end_station_id: start station identification code
    - started_at : bike ride start time in YYYY-MM-DD HH:MM:SS format
    - ended_at : bike ride end time in YYYY-MM-DD HH:MM:SS format
    - member_casual : type of bike user , member(registered) vs casual(unregistered)
  - Numerical feature : 4 variables(features)
    - start_lat : latitude of start station
    - start_lng : longitude of start station
    - end_lat : latitude of end station
    - end_lng : longitude of end station

Description of variables and basic statistics

- Categorical feature basic statistics

	count	unique	top	freq
ride_id	207023	207023	A67CB6CC130B48AB	1
rideable_type	207023	3	electric_bike	138598
started_at	207023	194952	2021-08-25 17:58:08	4
ended_at	207023	194495	2021-08-22 15:01:47	5
start_station_name	166811	465	Market St at 10th St	2423
start_station_id	166811	465	SF-J23-1	2423
end_station_name	162712	466	Market St at 10th St	2391
end_station_id	162712	466	SF-J23-1	2391
member_casual	207023	2	casual	121661

- NaN and Null in the dataset

```
ride_id                0
rideable_type          0
started_at             0
ended_at               0
start_station_name     40212
start_station_id       40212
end_station_name       44311
end_station_id         44311
start_lat              0
start_lng              0
end_lat                276
end_lng                276
member_casual          0
dtype: int64
```

- Numerical feature basic statistics

	count	mean	std	min	25%	50%	75%	max
start_lat	207023.0	37.749444	0.116645	37.280000	37.764285	37.776501	37.789620	37.880222
start_lng	207023.0	-122.371818	0.139639	-122.511282	-122.426964	-122.411306	-122.394625	-121.810000
end_lat	206747.0	37.749550	0.116624	37.240000	37.764277	37.776533	37.790000	37.900000
end_lng	206747.0	-122.371639	0.139574	-122.520000	-122.426630	-122.410887	-122.394586	-121.790000

Preliminary step: Handle NaN

- Rule 1: "start/end station name" and "start/end station id" have non-negligigle amount of NaN. They need to be properly handled.
    - Main idea : Some rows don't have Station name but they all have (latitude, longitude). If there exist other rows with similar (latitude, longitude) and valid station name/id, they can be used to fill the missing NaN station name/id

Fill NaN with station name/id with almost identical coordinate

Station name/id	Other feature	latitude	longitude
NaN	:	10.123xxx	10.123xxx
:	:	:	:
7th Ave at Cabrillo St	:	10.123xxx	10.123xxx

All NaN in station name/id can be filled in this method

- Rule 2: end_lat/lng columns have only 276 NaN out of 207023 elements, which is just 0.133%. Also if "end_lat" and "end_lng" are NaN, station name and id are also NaN. That makes these NaN datasets are not useful for data analysis. These NaN can be simply removed.

NaN count

```
ride_id                 0
rideable_type           0
started_at              0
ended_at                0
start_station_name  40212
start_station_id    40212
end_station_name    44311
end_station_id      44311
start_lat               0
start_lng               0
end_lat               276
end_lng               276
member_casual           0
dtype: int64
```

Preliminary step: Distance and Time duration

- Travel distance : The original data has latitude and longitude of bike ride start/end station. **So travel distance is computed from the coordinate of the two locations.**
    - **Coordinate is less intuitive** without converting it to better metric
    - Distance based on coordinate computation reference
      https://en.wikipedia.org/wiki/Haversine_formula
      https://towardsdatascience.com/heres-how-to-calculate-distance-between-2-geolocations-in-python-93ecab5bbba4
- Time duration : From the start and end time stamp of the bike ride, ride time duration is computed and added to the dataframe

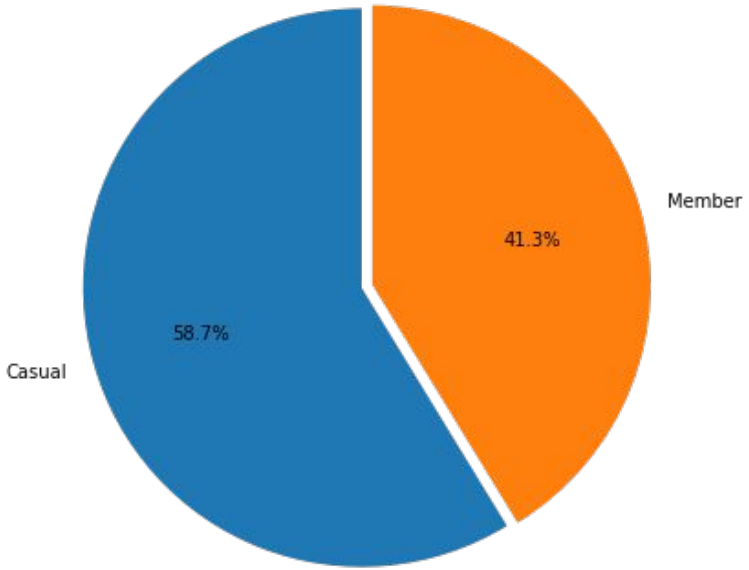
	ride_id	rideable_type	started_at	ended_at	start_station_name	end_lng	member_casual	time_delta	dist_km
0	A67CB6CC130B48AB	electric_bike	2021-08-07 13:30:11	2021-08-07 14:08:57	S Van Ness Ave at Market St	-122.388207	member	0 days 00:38:46	3.13
1	EA6D1C08FB8D1751	classic_bike	2021-08-16 18:34:12	2021-08-16 18:34:14	7th Ave at Cabrillo St	-122.464998	member	0 days 00:00:02	0.00
2	91E70C07BFA0BAED	electric_bike	2021-08-31 18:28:04	2021-08-31 19:12:40	7th Ave at Cabrillo St	-122.465001	member	0 days 00:44:36	0.00
3	1A5E792757C33356	electric_bike	2021-08-10 17:52:44	2021-08-10 18:02:46	23rd St at Tennessee St	-122.389884	member	0 days 00:10:02	2.10
4	B2EA7B7711640610	electric_bike	2021-08-11 18:01:21	2021-08-11 18:09:37	23rd St at Tennessee St	-122.389896	member	0 days 00:08:16	2.13

Summary statistics and metrics

User group analysis

- **Method**: As a first step of bike sharing data analysis, member vs casual user bike sharing service usage ratio is analyzed. Pie plot is the most frequently used visualization method for percentage of each group.

- **Observation**: As shown in the graph, there are more casual users then member. This is somewhat obvious and shows similar trend as other membership service such as Google ad-free subscriber.

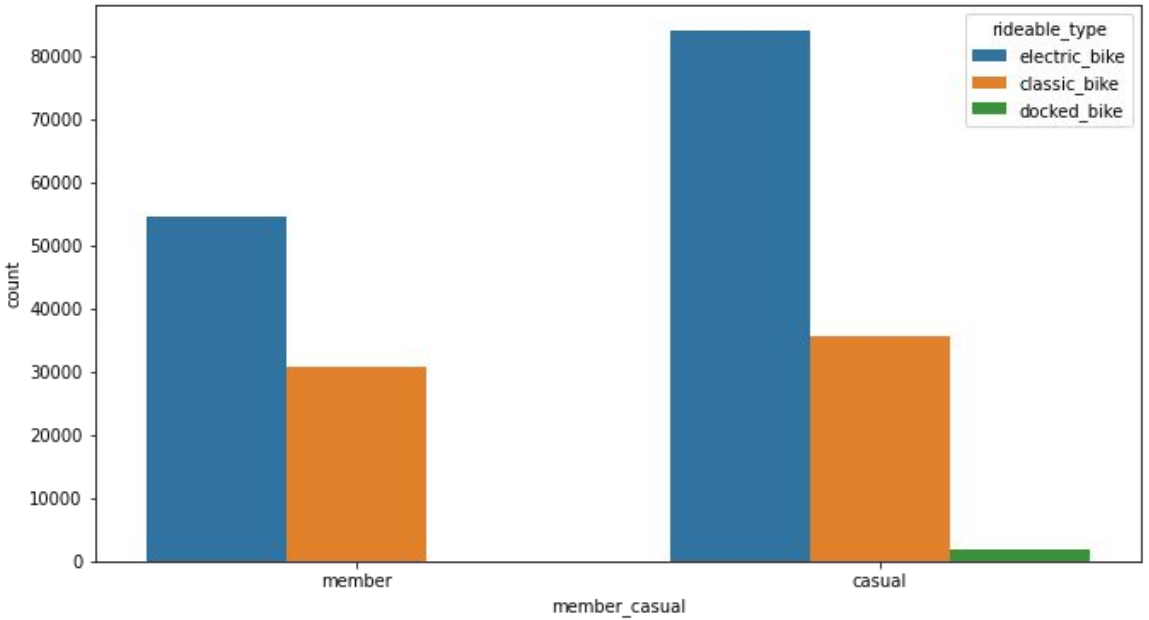


Summary statistics and metrics

Bike ride count per user and bike types

Observation

- In each user group, **electric bikes** are used more than other bike types
- **Casual users** tends to ride electric bike in comparison to member user.
- There are less than 3000 **docked bike** ride history which is **negligible** to other bike types.



Summary statistics and metrics

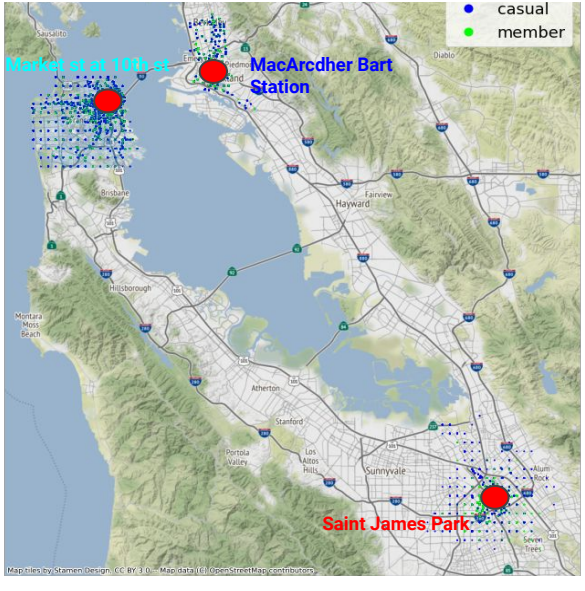
Geographical analysis on Bike ride

- **Method**: Start and End station locations are directly mapped to the bay area map to reveal the geospatial distribution of bike ride. (Ref https://medium.com/@sindhu.ravikumar/visualizing-spatial-data-with-geopandas-and-contextily-10e9b8e71e49)

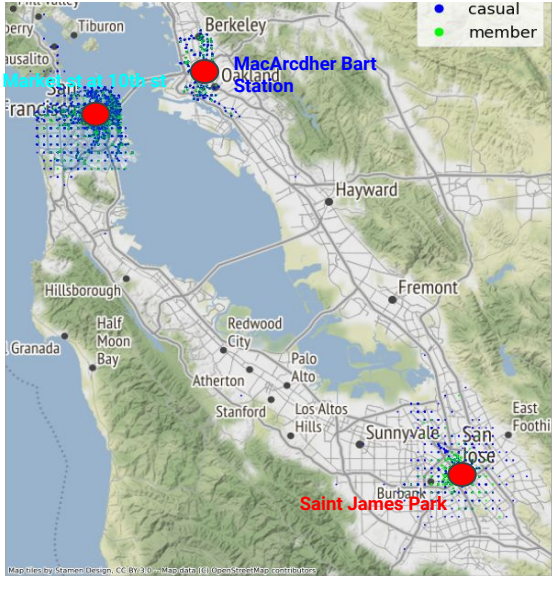
- **Observation**:
  - Most rides are concentrated on the most popular station : More bikes needed in this location

Most popular station on red circle

Start station distribution



End station distribution

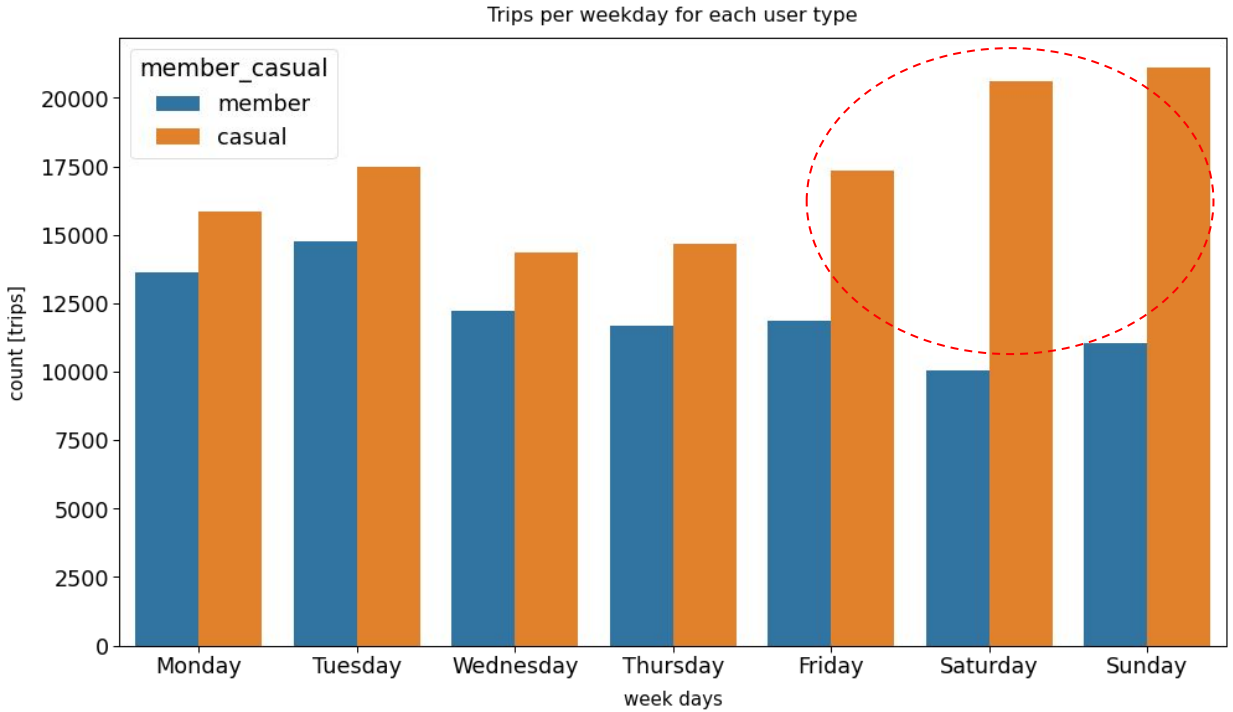


Summary statistics and metrics

Bike ride comparison : Weekend vs Weekdays per member type

Observation : Casual and Member have very different ride behavior. Bike rides of casual member significantly increases during the weekends

- It is possible that **registered member use bikes mainly for commuting** so bike rides count reduces during the weekends.
- **Casual members use bike for non-commuting purpose** such as sightseeing and hiking during the weekend.



Summary statistics and metrics

New metric: Daily bike occupancy index

Motivation: In addition to the daily bike ride statistics, it is also important to understand how long bike users occupy once they start bike sharing service. The occupancy can be important factor in better bike allocation and purchase decision in addition to overall bike ride statistics

- $B_o = \frac{N_{bike}(day)}{N_{average}} \cdot T_{average}(day)$

  where $N_{bike}(day)$, $N_{average}$,and $T_{average}(day)$ are "the number of total bike rides each day", "Average bike ride a day" and "average bike ride time duration".

  - if $\frac{N_{bike}(day)}{N_{average}} < 1$, daily rides of the bike on that day is below average and vice versa. So multiplying this with averaged daily travel time duration $T_{average}(day)$ will give us bike occupancy of each day. This can be defined as new index called daily "Bike occupance index"
  - Bike occupancy index B_o can be used to apply different pricing each day resulting in more balanced bike usage distribution. That will help maximize the bike utilization.

Summary statistics and metrics

New metric: Daily bike occupancy index

Finding 1: The result shows that bike ride occupancy gradually increases from Monday to Thursday. Then it drops to Friday suddenly.

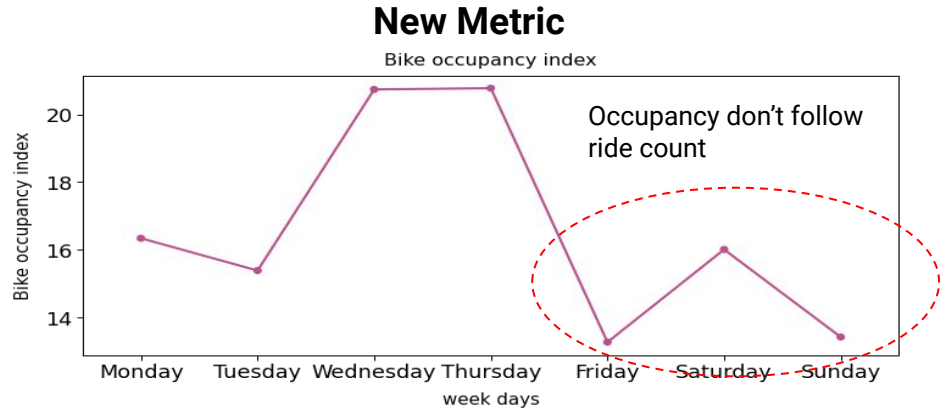
- This is quite opposite to Bike ride count. It means that **during the weekdays, many people don't return the bike and just occupies it.**

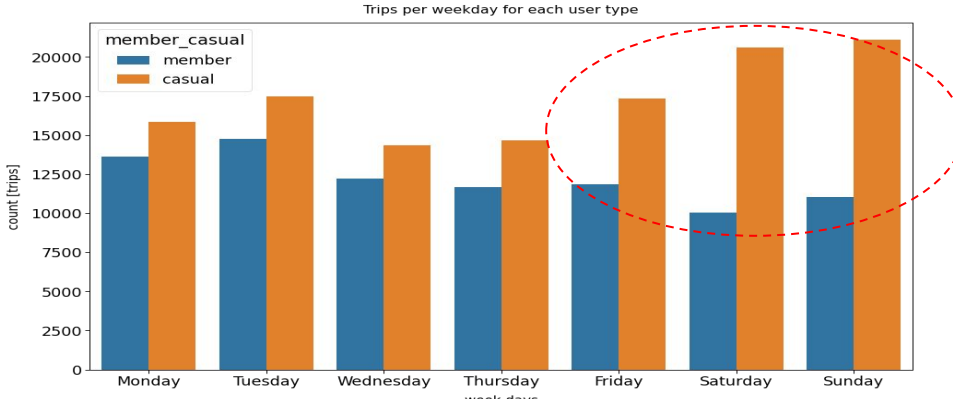
Finding 2: Even though more bikes are used during the weekend, bike occupance is not high. It means that many **people use bike for relatively shorter time during weekend** than during the weekdays

Suggestion:

To improve the bike utilization during weekdays, one can consider promotion or credit to early bike return users

Instead of purchasing more bikes to **handle the weekend demand, it might be better to increase the bike re-allocation staffs for better bike circulation**.





Summary statistics and metrics

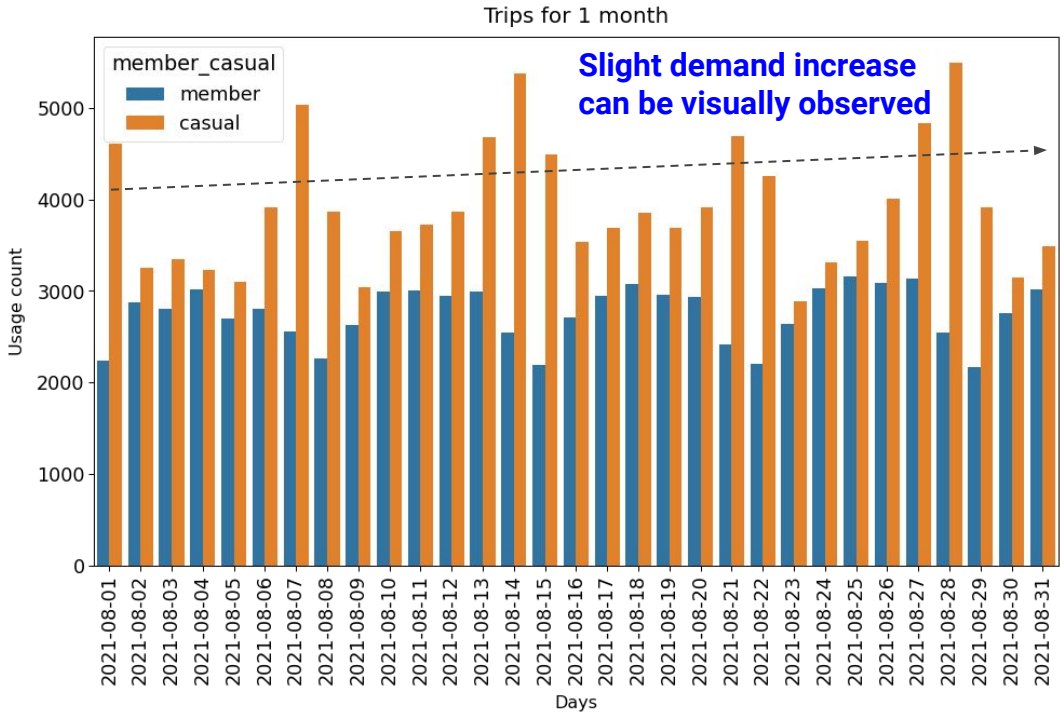
Bike ride trend for 1 month (range of dataset) per user type

Observation : Bike ride trend shows that demands changes periodically

- Member(registered user)'s ride count decreases during the weekend(8/1, 8/8 ~ 8/29 are Saturday) while casual(unregistered) users count increases during the weekend.
- This is aligned with initial observation and support the hypothesis that member mostly use bikes for commuting purpose

Interesting finding

- One can visually observe that **there is a growing demand**. This will be **further investigated in the next section**



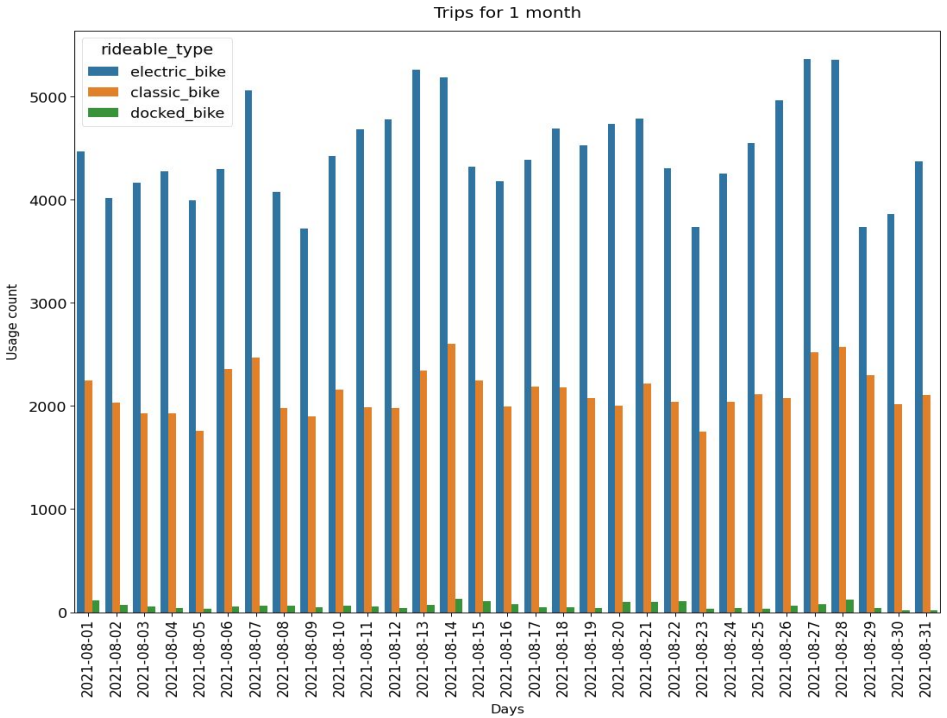
Summary statistics and metrics

Bike ride trend for 1 month (range of dataset) per bike type

Since the previous bike type analysis showed that electric bikes are more popular than other types of bikes, it is worth investigating demand for the given time duration

Observation & finding : Bike ride trend shows that demands changes periodically

- This is aligned with initial observation and support the hypothesis that member mostly use bikes for commuting purpose



Summary statistics and metrics

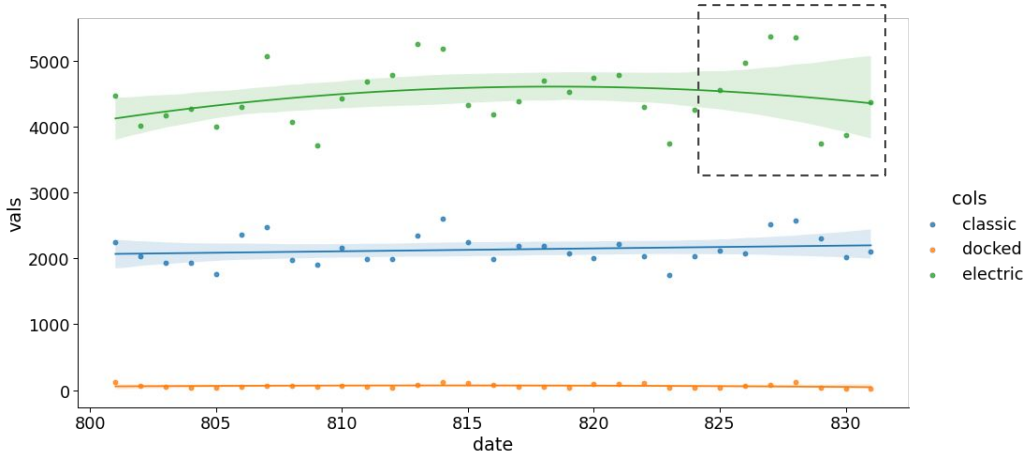
Bike ride trend per bike type with regression line

Observation: The **second order regression** plot for bike ride per bike type shows that **there is a growing demand for electric bike ride while other bike types don't change**

- Due to short time frame data, regression line got distorted at the boundary of the data.
- If there were more monthly ride data, more clear bike ride statistics could be identified.

This could be important statistics that can help purchase department plan the future bike purchase plan with limited budget

Due to insufficient data points, regression line showed slight decrease at the end. More data point can help improve this phenomenon



Date : converted to integer for regression: i.e 805 → 8/5

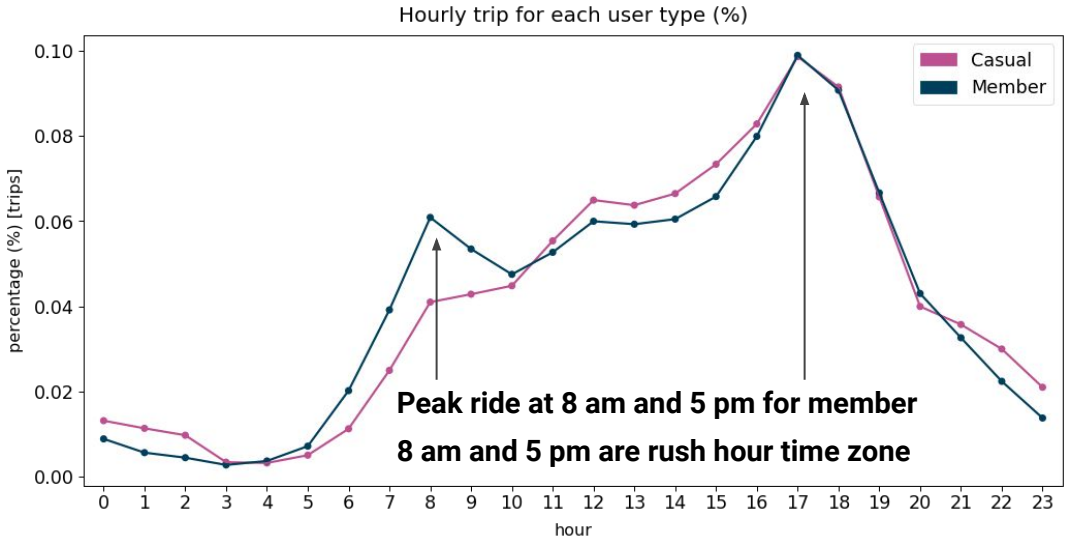
Summary statistics and metrics

Hourly trip data per user type during a day

Observation

- Hourly data shows that **member rides bike mostly at 5 pm and 8 am.**
- Hourly data shows that **casual user tend to ride bikes more in the afternoon**
- **Both user types rides bikes most at 5 pm** when most people start to return home

This is also aligned with initial hypothesis that member uses bike for commuting since 5 pm and 8 am are the rush hour time.



Summary statistics and metrics

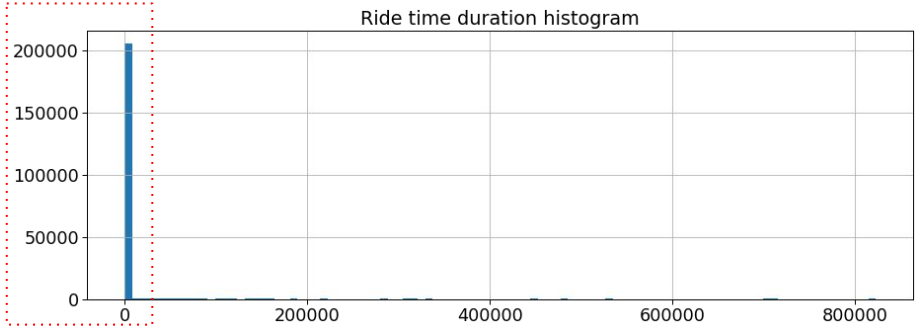
Ride time duration

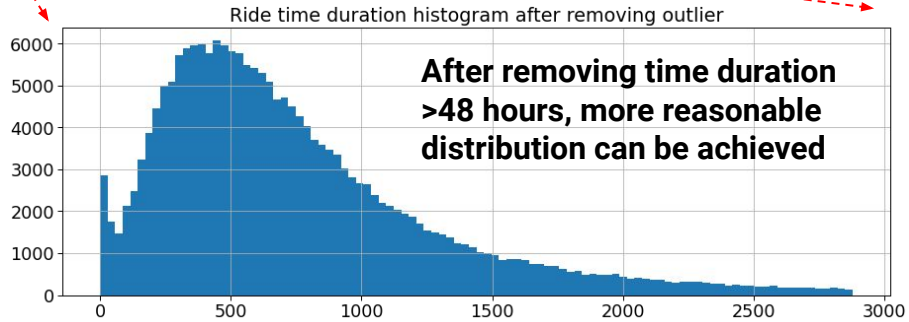
Observation

- To check the ride time duration, the time duration in time delta format is converted to second format. Then histogram of the data was analyzed in the first histogram "ride time duration histogram"
  - The result shows that **some users didn't bike more than days**, which makes **histogram severely right-skewed**.
  - Data with ride time duration > 48 hours are removed as outlier

- After removing the outlier, the data shows clear trend. **It's still right-skewed overall and around 500 minutes(8.5 hours) are the maximum ride frequency data point**. **This time matches the most common work hour 8.5 hours**

Severely right skewed ride time duration data



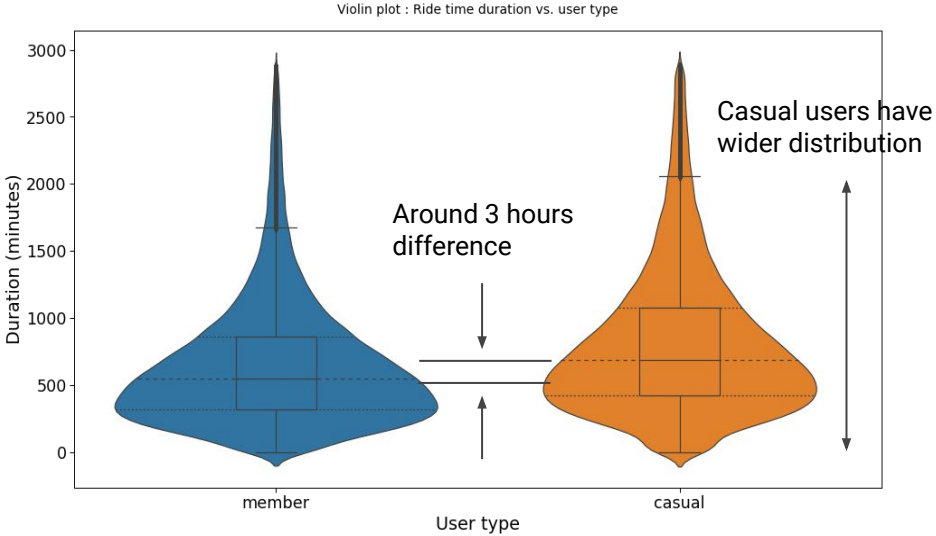


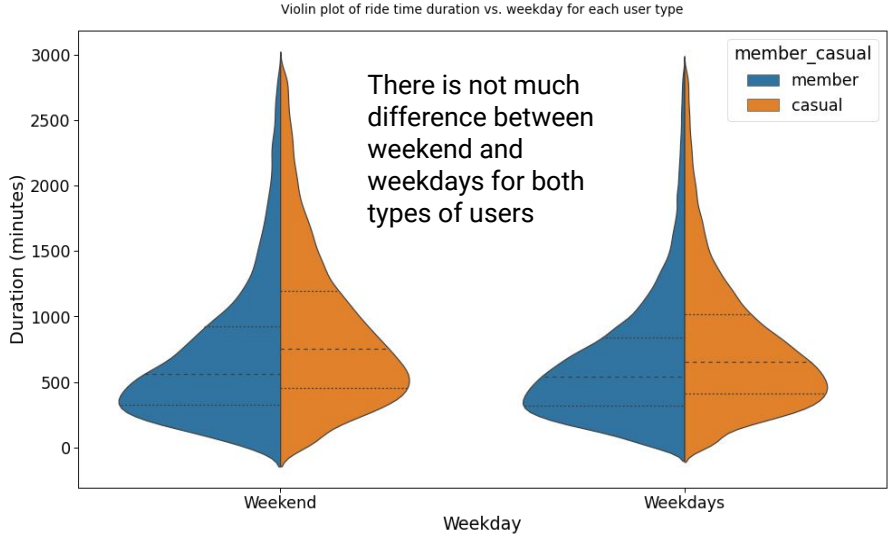
Summary statistics and metrics

Violin plot of ride time duration per member and weekend/weekdays

Observation :

- The casual users have more ride time duration : Higher median value in the box plot
  - The ride time duration of most of the members are 480 min (8 hours): largest width around 480
  - The ride time duration of most of the unregistered casual user : 3 hours longer than member's ride time.
- The casual user has wider time duration distribution : Max point of box plot is much higher than member's max point
  - Casual user's ride pattern is not similar to each other



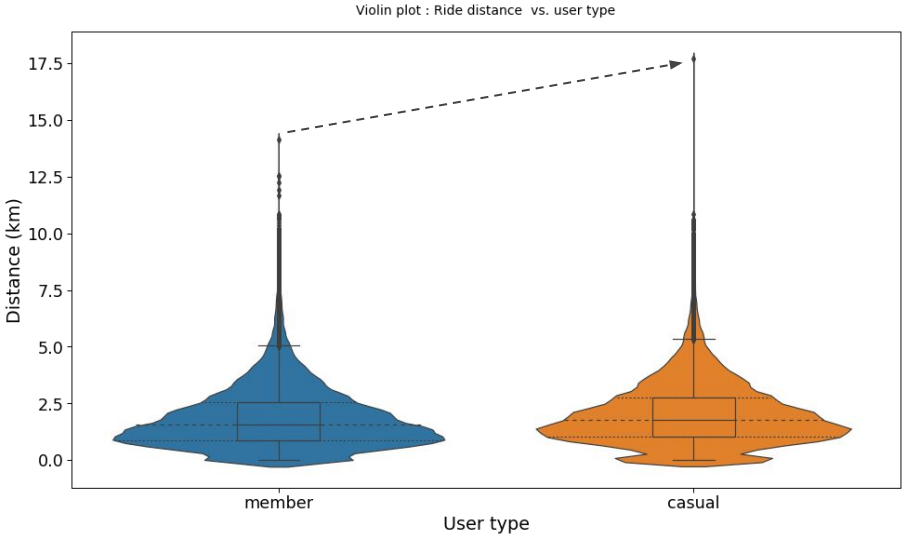


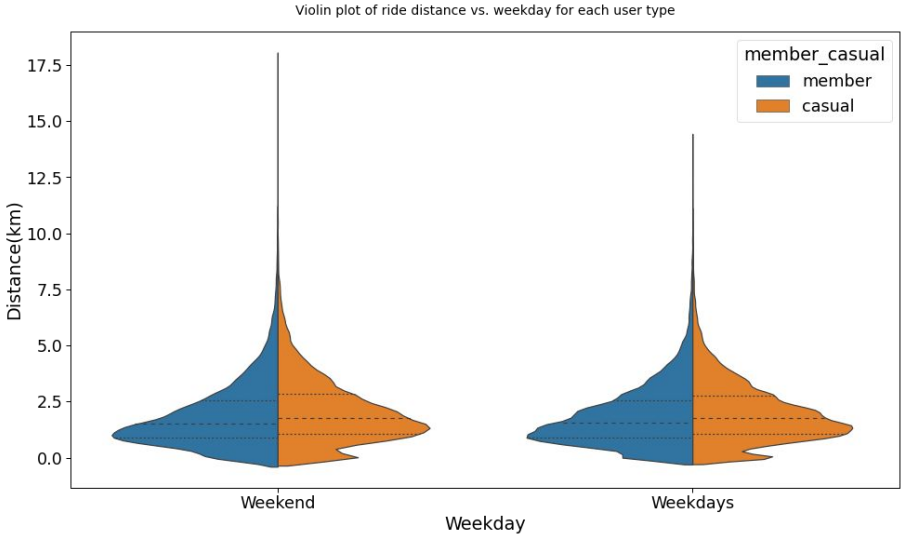
Summary statistics and metrics

Violin plot of ride time distance per member and weekend/weekdays

Observation :

- The median bike ride distance of member is around 1.7 km which is 10 ~20 min distance from the station location
- The median bike ride distance of casual user is around 2 km which is slightly higher than member user's travel distance and much wider distribution





Conclusion

- Throughout the data analysis, 2 simple metrics were additionally created which are time duration and travel distance. Then 1 new dimension feature called “bike occupancy index” was created
- Geographical bike ride location distribution shows that most bike rides are concentrated on few popular station. More bike need to be prepared in those locations
- The summary statistics give us the following information
  - It is highly possible that member uses bike for commuting while casual user use bike for other purpose such as sight seeing
    - Member user’s bike usage pattern is more predictable
    - Casual users ride bikes mostly during the weekends while members ride bikes mostly during weekdays
  - Casual user occupies bike for longer time and return late during weekdays
    - By providing promotion or credit to early return user, bike utilization rate can be improved during weekdays
  - The regression plot shows that there is a growing demand for electric bikes compared to the other bike types
    - It is recommended to purchase more electric bikes to boost bike sharing service