Scoot Employee Rides Data Analysis Report
By Gurpreet Singh

**Objective:**

The objective of this assignment is to analyze scoot employee rides data to quantify free utilization by employees.

**Dataset description:**

The data set provided had 78,160 records in csv format and have the following columns:

| 1 | id | scooter_id | vehicle_type | user_id | scoot_plan_i | ride_type_id | start_time_l | end_time_lo | price | scoot_move | start_odome | end_odomet | start_range | end_range | start_loca | start_lat | start_lon | end_locatior | end_lat | end_lon |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 305028 | 90 | 1 | 6026 | 5 | 8 | 29:42.7 | 26:08.7 | 0 | t | 3692.6 | 3696 | 18 | 13 | 97 | 37.804907 | -122.44199 | 113 | 37.796233 | -122.43475 |

**How I approached this problem:**

- Read the assignment requirements and observed the data set to get a general sense of what is being asked
- Conducted background research on Scoot and how it operates.
- Accumulated a list of tools best for analyzing and visualizing data

**Tools Used:**

**Virtual Environment: Anaconda (for handling python dependencies)**
**For Data Processing: Pandas, NumPy**
**For Visualization: Matplotlib, Plotly (sign up required for use), Jupyter Notebook**

**Cleaning up data:**

- One of the first steps before processing data is validating consistency of data (check if there is any inconsistencies in data, or noise in data), as in this case some of the rows were missing values for some columns.

- Since most of the metrics were based on number of rides, I decided to divide the data into two categories: rides where scoot moved (**scoot_moved == t**) and rides where scoot did not move.

- **Hence for most part, the metrics are calculated based on the attribute if scoot_moved is True.**

# PART 1:

- **Ride Count by day:**

  Ride count by day could be further classified into:

  - Average Ride count **per hour**
  - Average Ride count **every weekday**
  - Average Ride count **per day for the whole duration**

I decided to break down into these three categories to analyze how the ride activity changes over the period of 24 hours, over the period of a week, and over the period of a whole year.

**Average Ride Count per hour:**

The following is the daily average ride count from **Jan 1, 2016 to May 12, 2017**

| Hour | Avg. Ride Count |
|---|---|
| 00:00 | 3 |
| 01:00 | 1 |
| 02:00 | 0 |
| 03:00 | 0 |
| 04:00 | 2 |
| 05:00 | 4 |
| 06:00 | 8 |
| 07:00 | 28 |
| 08:00 | 69 |
| 09:00 | 124 |
| 10:00 | 174 |
| 11:00 | 196 |
| 12:00 | 208 |
| 13:00 | 196 |
| 14:00 | 195 |
| 15:00 | 204 |
| 16:00 | 165 |
| 17:00 | 134 |
| 18:00 | 118 |
| 19:00 | 72 |
| 20:00 | 43 |
| 21:00 | 18 |
| 22:00 | 12 |
| 23:00 | 6 |

fig 1.1

While these numbers might not seem appealing, visualizing the results into a histogram gives a better sense of ride counts for every hour in a day.

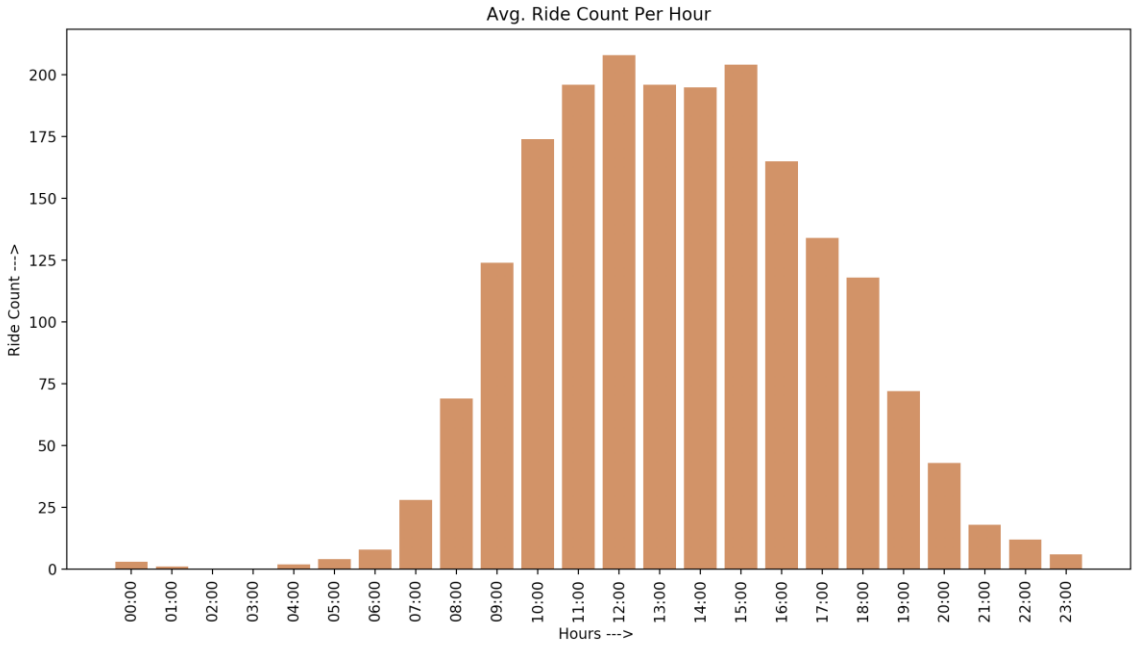


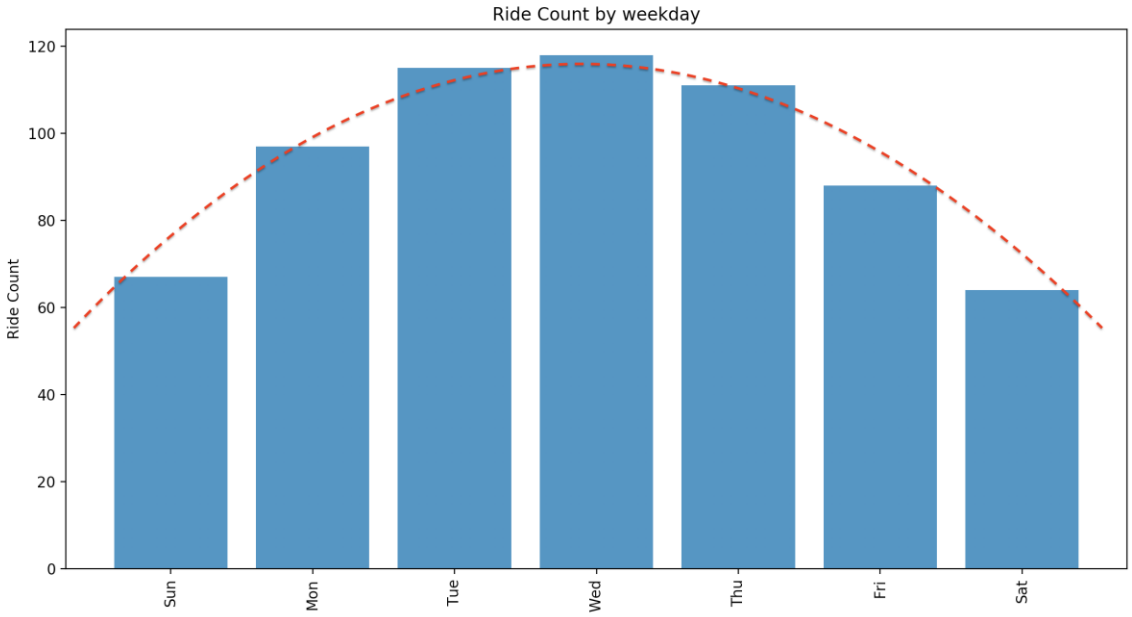

fig 1.2

**Findings**:

- Highest average rides for a day are around peak hours (lunch hours 12PM – 3PM)
- It can be inferred that employees use scoot mostly during lunch time

**Average Ride count every weekday:**

fig 1.3



**Findings**:

- Average ride count by weekday gives a visual interpretation of scoot usage every day of the week.
- There are more average rides on weekdays (Tuesday, Wednesday, and Thursday specifically) than there are on weekends.

**Average Ride count by day from Jan 1, 2016 – May 12, 2017:**

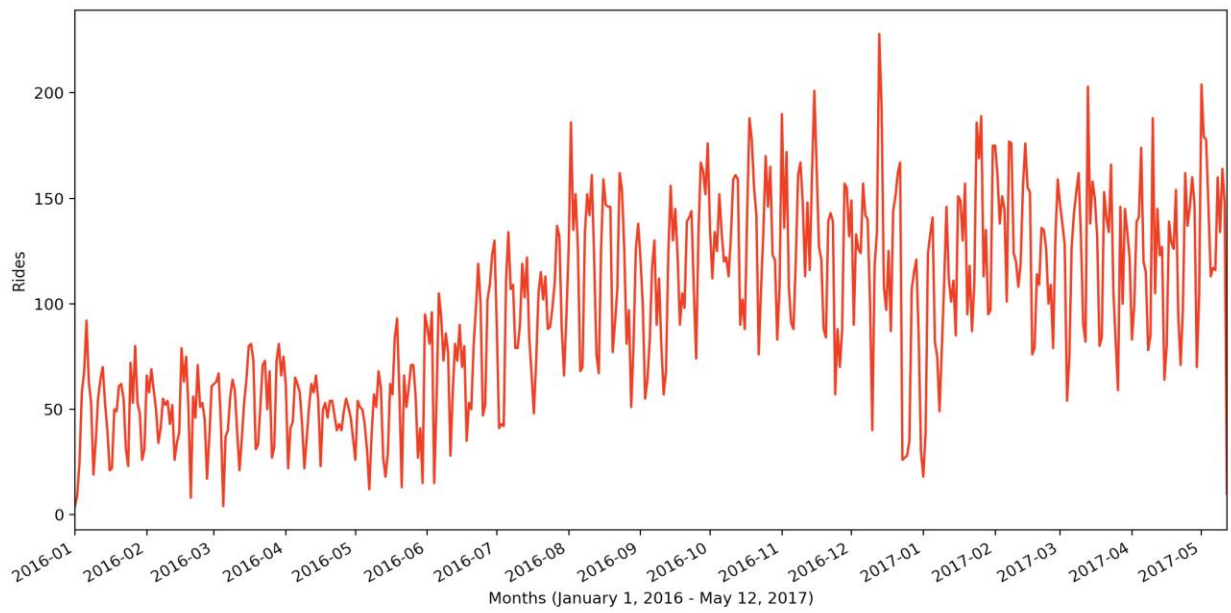



fig 1.4

This line graph is the most interesting one as it shows the scoot usage by employees for the whole duration of dataset.

- The average ride count increased towards the end of the year 2016 (which shows the increase in scoot usage)
- Starting from year 2017, the ride count has been consistent and from analyzing previous year's trend, **it is probable that more employees will start using scoot over time**.

- **Ride count by User**

| User Id. | Ride Count |
|---|---|
| 5089 | 3487 |
| 15542 | 2514 |
| 5719 | 2467 |
| 12032 | 2451 |
| 19782 | 2204 |
| 17453 | 1997 |
| 24084 | 1716 |
| 26805 | 1665 |
| 28940 | 1576 |
| 24763 | 1563 |
| 6024 | 1418 |
| 25386 | 1387 |

Ride Count by user counts users (user_id) with most rides throughout the duration of data set. This information does not tell why employees with most rides tend to use scoot while others do not, hence, to analyze the pattern I cross checked user ride count against starting locations and ending locations.
Therefore, I plotted starting and ending points for both top 20 users and lowest 20 users to notice any peculiarities.

The following is the map for top 20 users with their starting locations (blue) and ending locations (red)

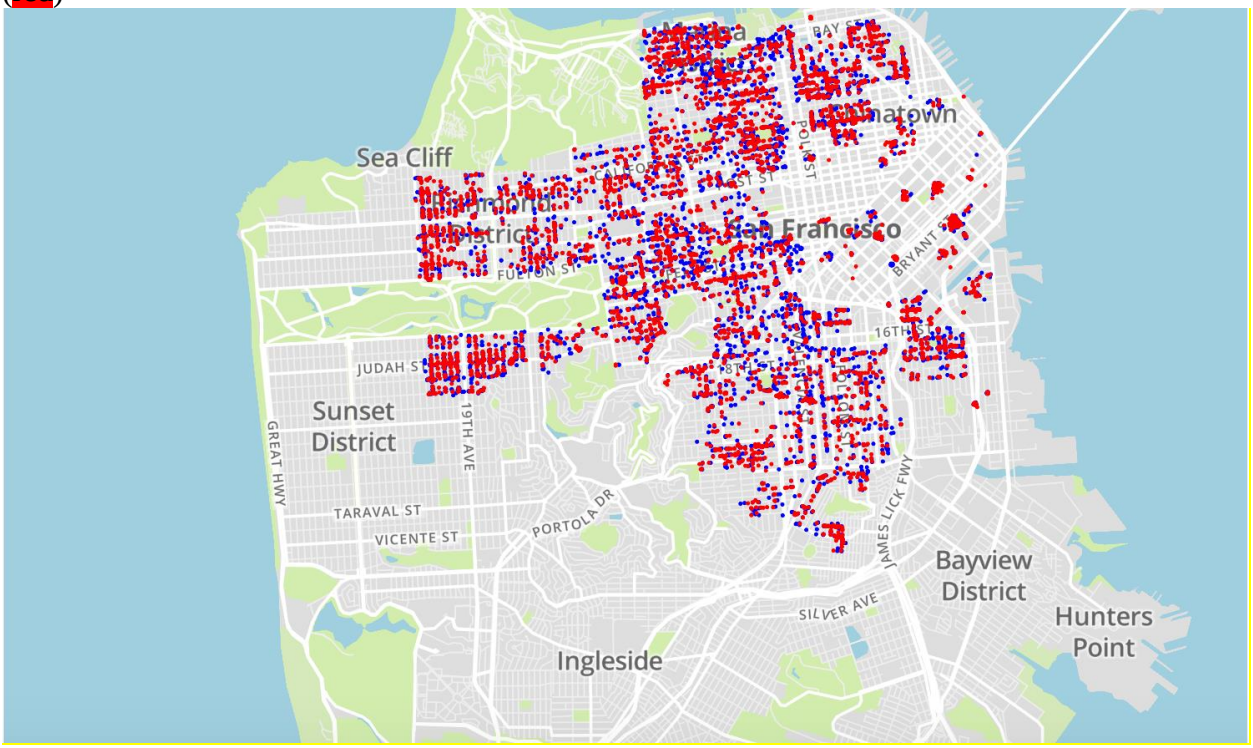



fig 1.5

For top 20 users, the starting points and ending points are dense and fairly distributed throughout the downtown.

However, the distribution for lowest 20 users is sparse and limited to south Mission street and south market street. (Cyan – lowest starting points, Magenta = lowest ending points)

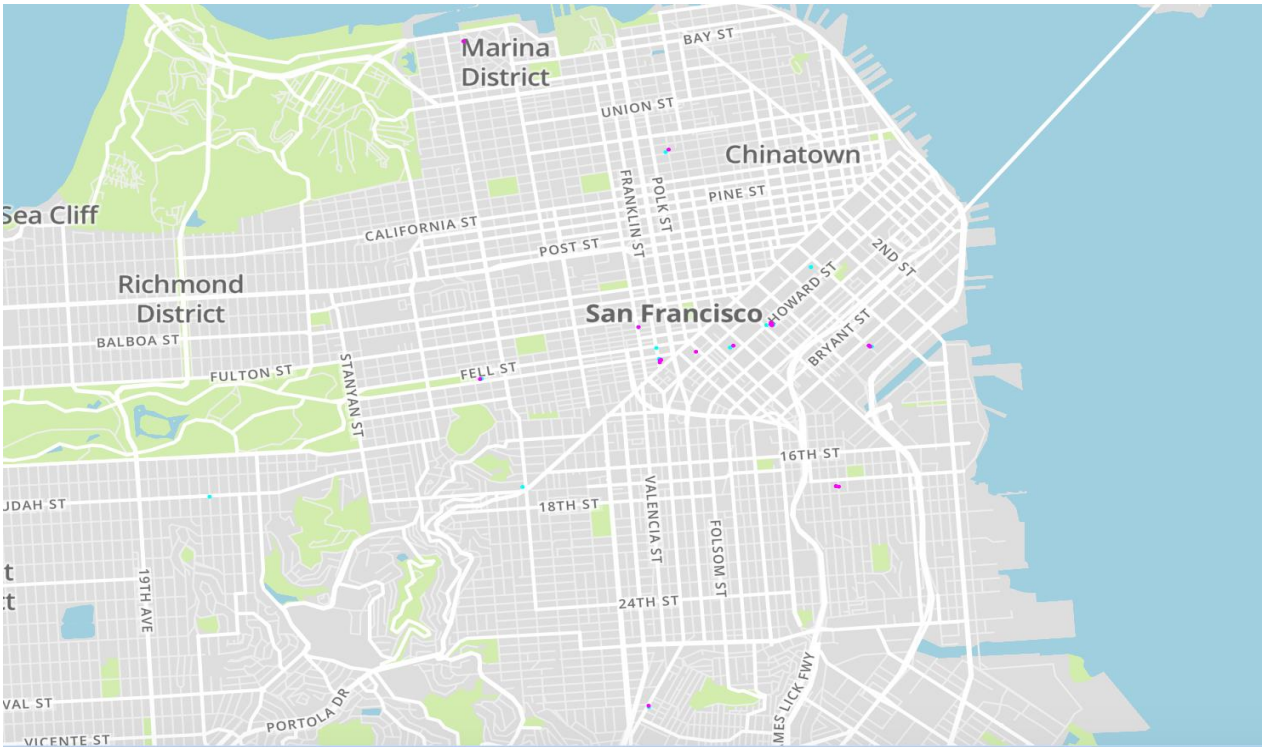



fig 1.6

- **Ride count and total mileage by vehicle type**

  **Ride Count by Vehicle type:**

| Id. | Ride Count |
|---|---|
| 6 | 37184 |
| 1 | 26128 |
| 3 | 12409 |
| 5 | 2281 |
| 4 | 107 |
| 7 | 50 |

The following is the graph visualization for the ride count by vehicle type:

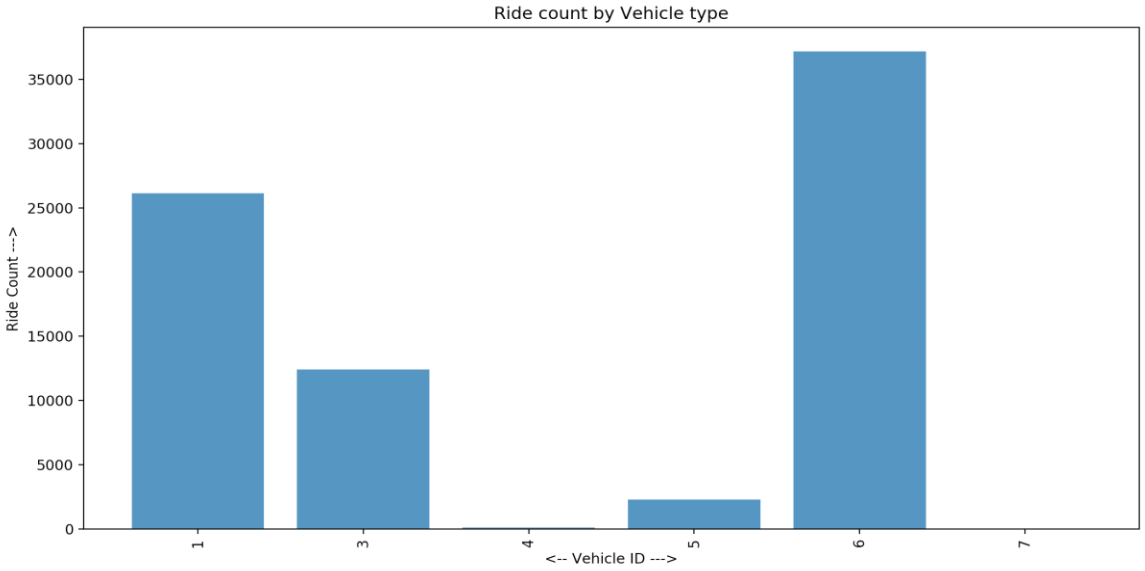



fig 1.7

From the graph above **Vehicle type 6** is preferred the most followed by vehicle type 1, 3, 5, 4, and 7.

**Total Mileage by Vehicle type:**

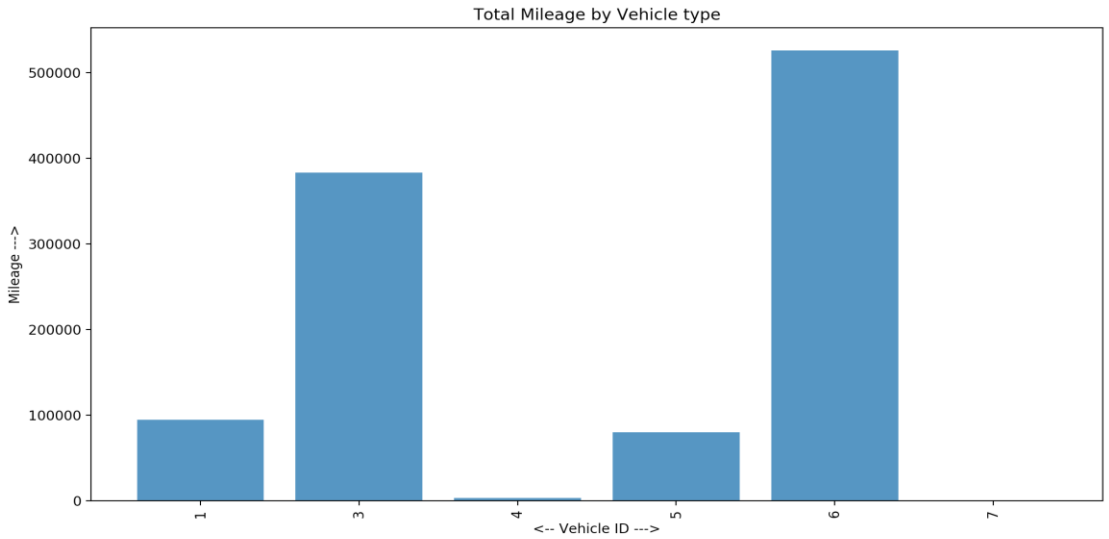



fig 1.8

Comparing most used vehicle type against its mileage gives some interesting insights:

- **Vehicle type 6 has most mileage (as anticipated since it is used most)**
- Surprisingly, Vehicle type 3 has more mileage than vehicle type 1 (I am assuming that vehicle type 3 might be more efficient or comfortable than vehicle type 1, although it might not be the case)
- If vehicle type 3 turns out to be more comfortable and efficient than vehicle type 1 (based on prior assumption), employees using vehicle type1 could be encouraged to use vehicle type 3 or 6 to increase ride usage by employees.

- **Distribution of ride mileage**

  For the distribution of ride mileage, I calculated number of rides for ride mileage ranging from 0 to 10 miles.

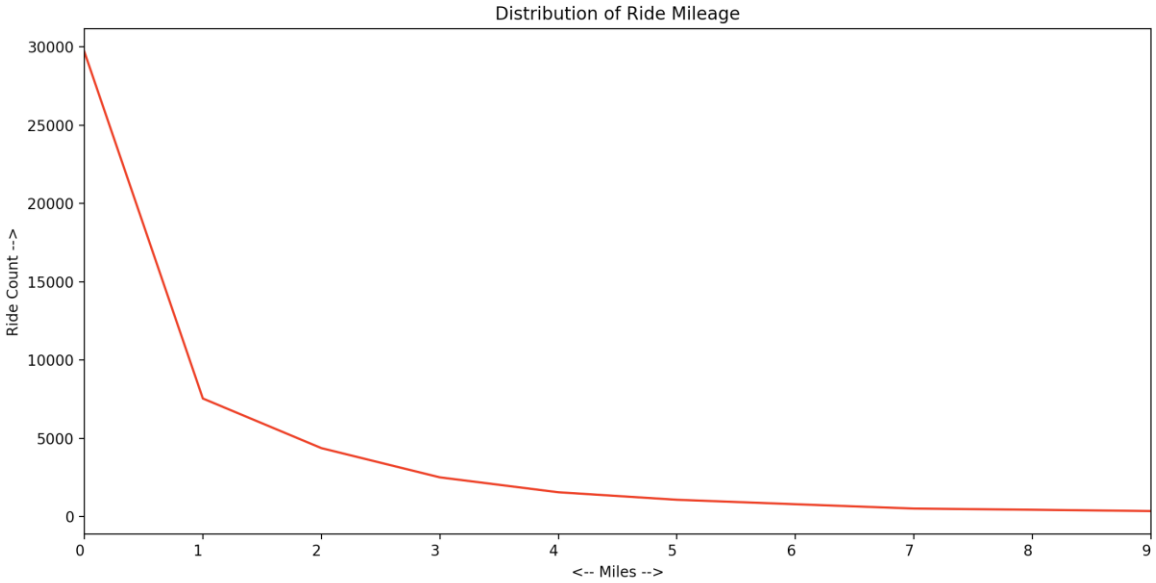



fig 1.9

  - From the graph, there are about **29698 rides that range between 0 – 1 mile**
  - It can be inferred that **60%** of the employees drive less than a mile (1 or 2 blocks)

- **Top 5 locations by total volume (ride starts & ride ends)**

```
Top 5 starting locations:
Loc.ID   Count
108.0    10013
72.0      1206
1.0       1061
58.0      1036
109.0      816
Name: start_location_id, dtype: int64
Top 5 ending locations:
Loc.ID   Count
108.0    9345
152.0    1334
72.0     1186
1.0      1074
58.0      998
```

  These are top 5 starting and ending points in the SF downtown. Since these number do not tell what these locations are, I decided to visualize these hot spots to get a better sense of hot spot locations:

The maps below could be accessed from this link

Distribution of top 5 starting points on a map:

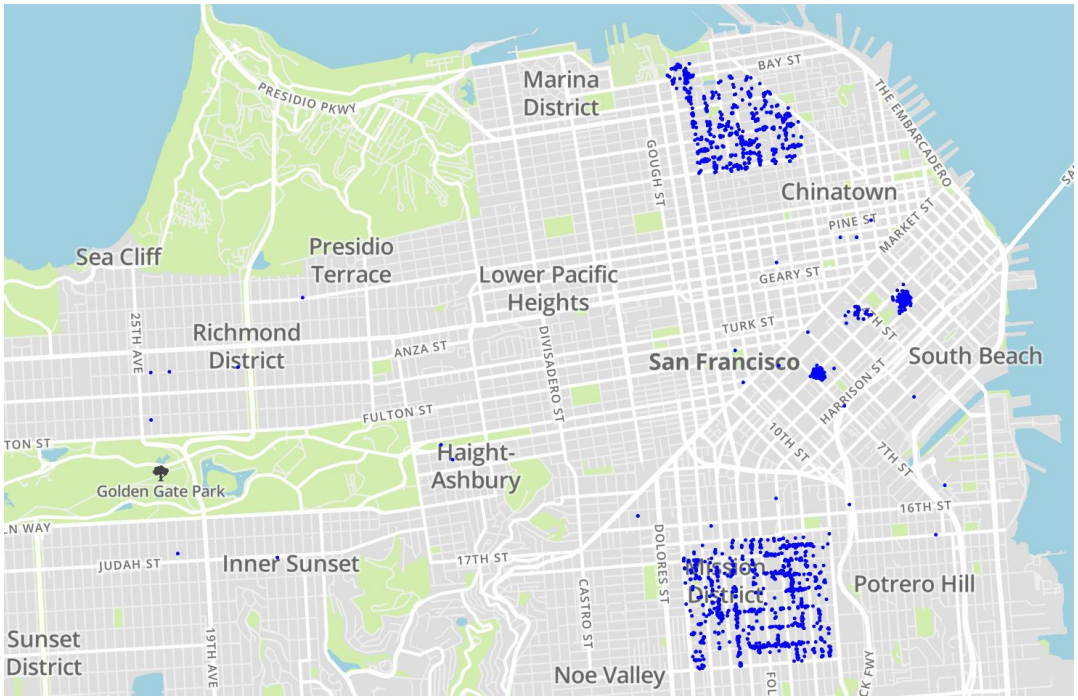



fig 1.10

Distribution of top 5 ending points:

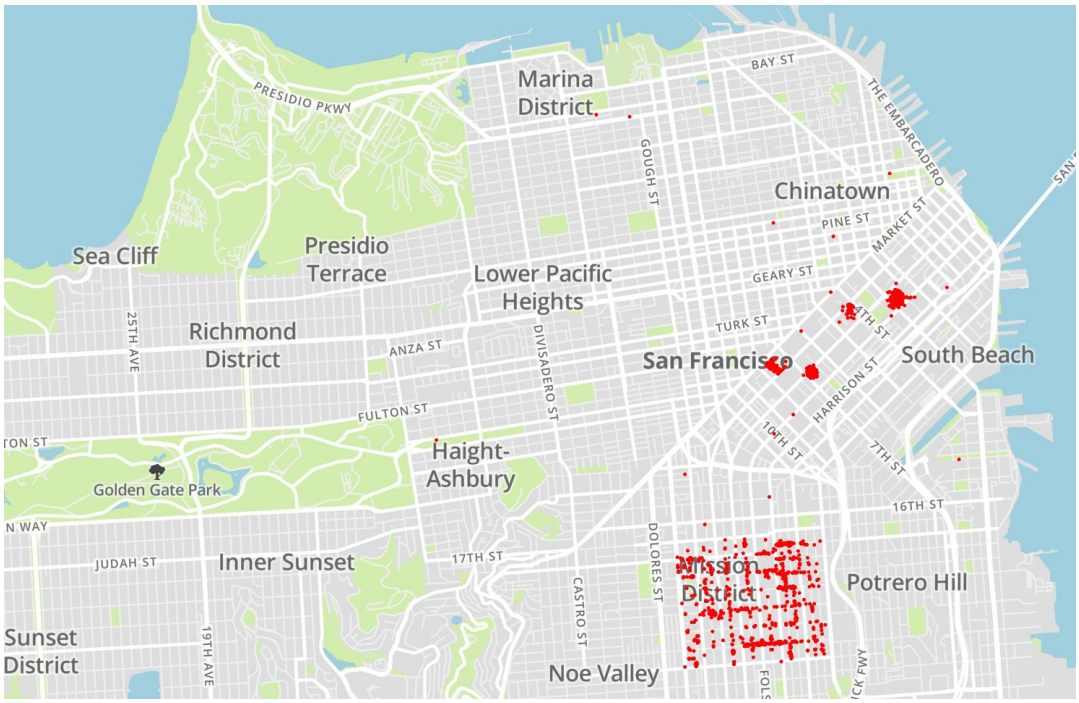



fig 1.1

Starting and Ending points together:

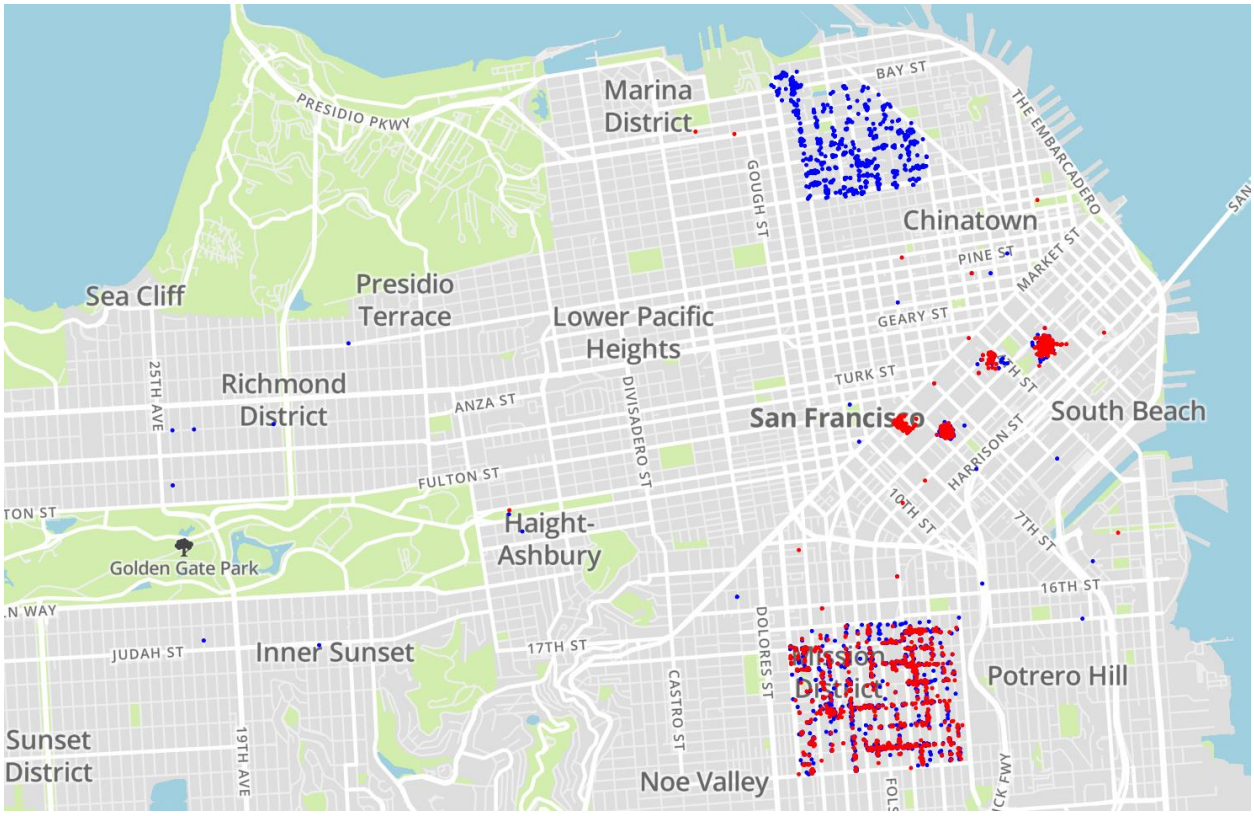



fig 1.12

# Part 2:

1. Based on these metrics, can you distinguish certain locations that see significantly more activity? How does this change by time of day and day of week?

   - From fig 1.12, **Russian Hill-Macondray Lane Historic District** and **Mission district** are the two locations with most activity (from top 5 locations by volume).

   - Further zooming into the map, there are very **dense hot spots along Market st, Howard st, and Mission st. (fig 1.13)**

   - **The activity is highest during peak hours of a day (fig 1.2) and gradually decreases towards the end of the day.**

   - **During weekday, the activity is highest on Tuesdays, Wednesdays, and Thursdays.**

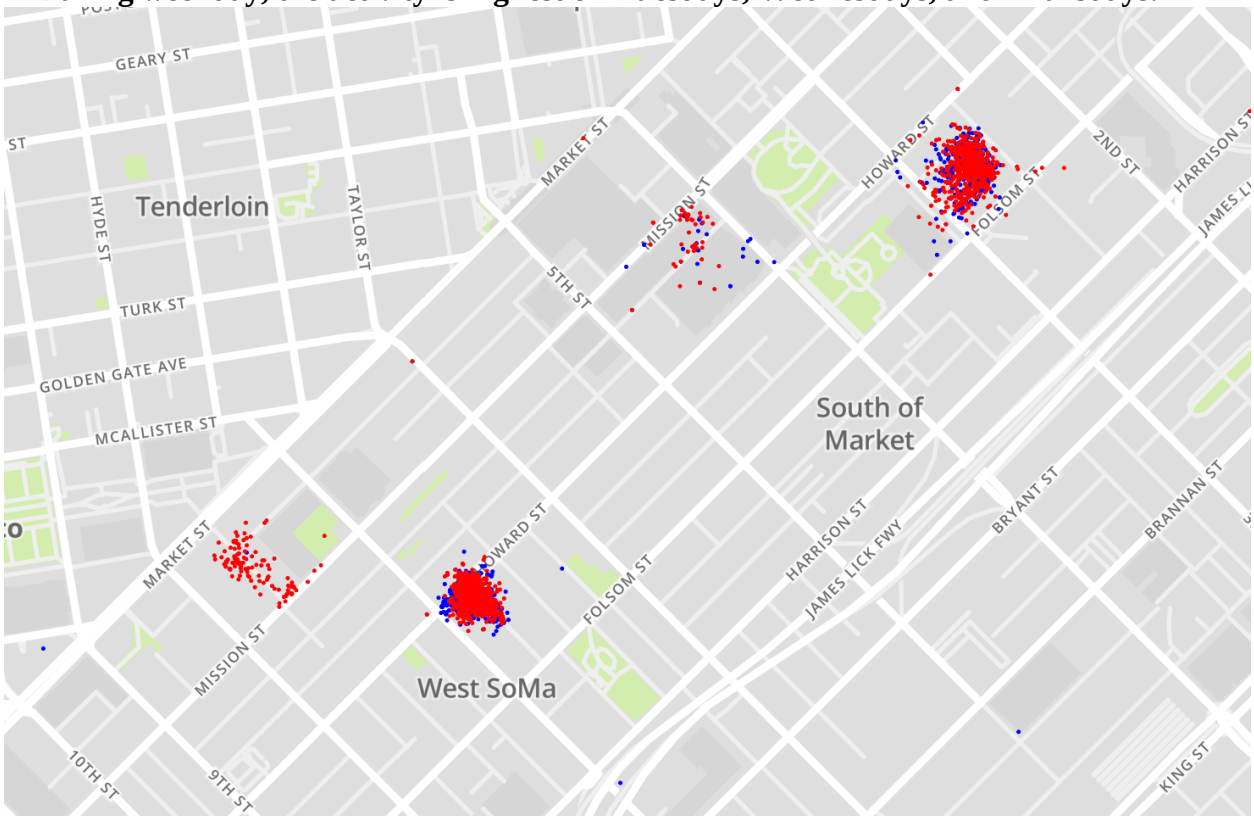



fig 1.13

2. Are there other noticeable patterns, or peculiarities with the ride data? What might explain these?

   **Findings:**

   - **Russian Hill-Macondray Lane Historic District has mostly starting points and no ending points as opposed to other hot spot locations. (Fig. 1.12)**

   - **Hence, it can be inferred that most rides flow from north San Francisco to south San Francisco.**

   - **As noticed from fig 1.9 (Distribution of Ride mileage) and fig 1.13, 60% of the employees stay within a block or two and drive less than a mile per ride.**