Bike-Share Usage in London

Group 5: Meilin Li, Yunhe Jia, You Wu, Yixuan Zeng

Data Description

- **Source:** Kaggle "Bike-Share Usage in London and Taipei Network"
- **Data size:** 5.07 GB
- **Data files:**
  - **london.csv:** Usage based, includes the duration, start time & station, end time & station of each usage.
  - **london_station.csv:** Station based, includes all the bike-share stations and their coordinates.

Data Analytic Goal

Analytical goal:

1. Understand the time pattern of bike users
2. Bike rental duration distribution
3. Analyze relationship between startstation and duration
4. Bike rental spatial distribution

Bike Rental Time Pattern

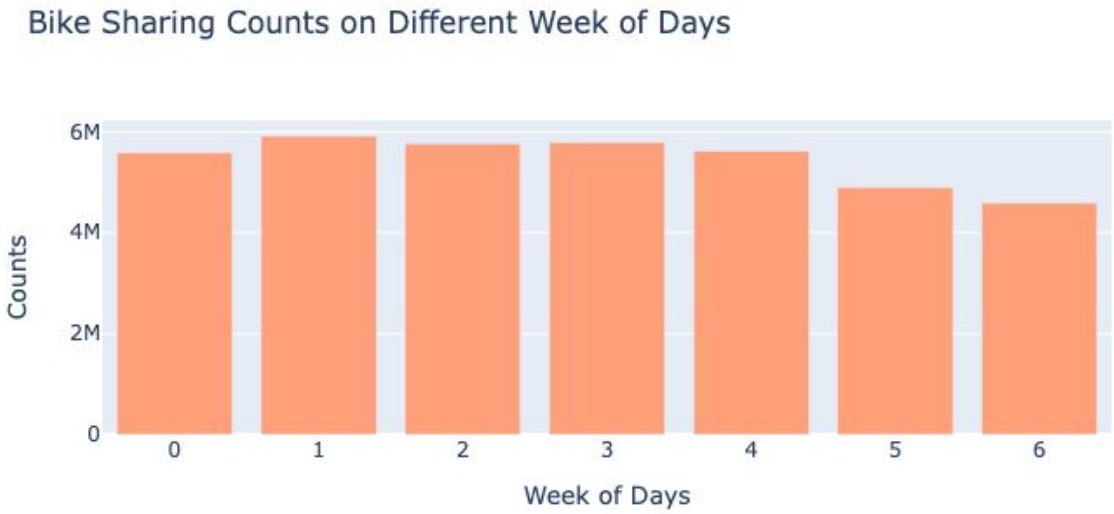
Analytical goal: Understand the time pattern of bike users

- Bike rentals on different month, hour and week of days
- Heatmap of time pattern

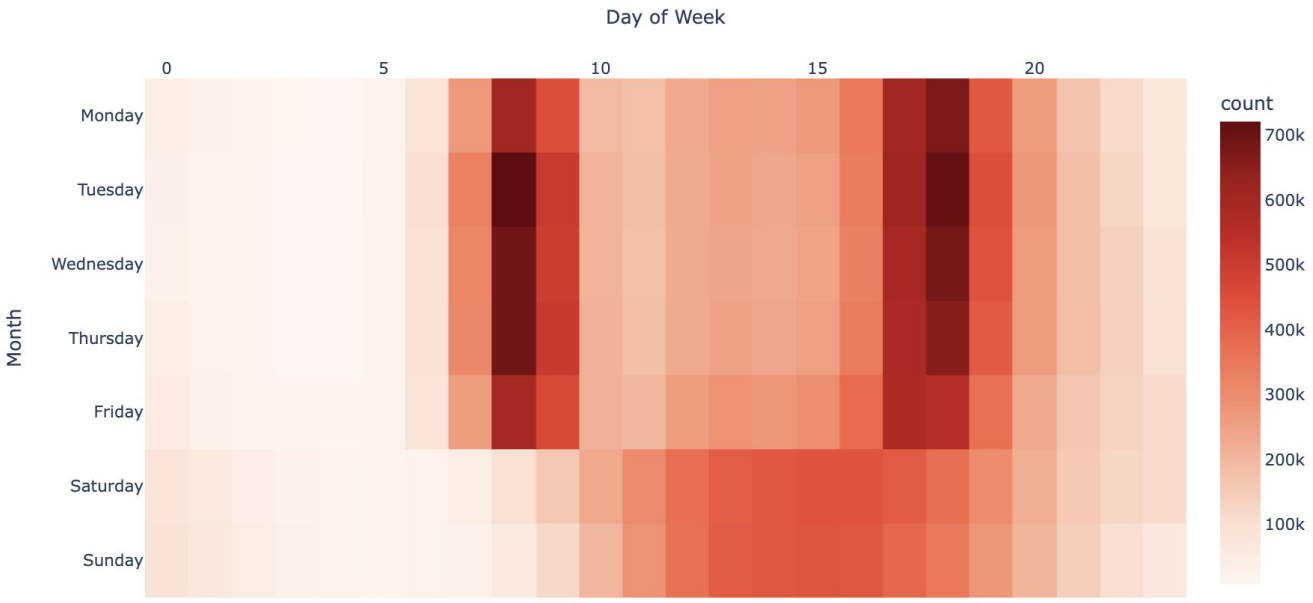
Month Distribution

Hours Distribution

Week of Days Distribution



Heatmap on Week of Days and Hours

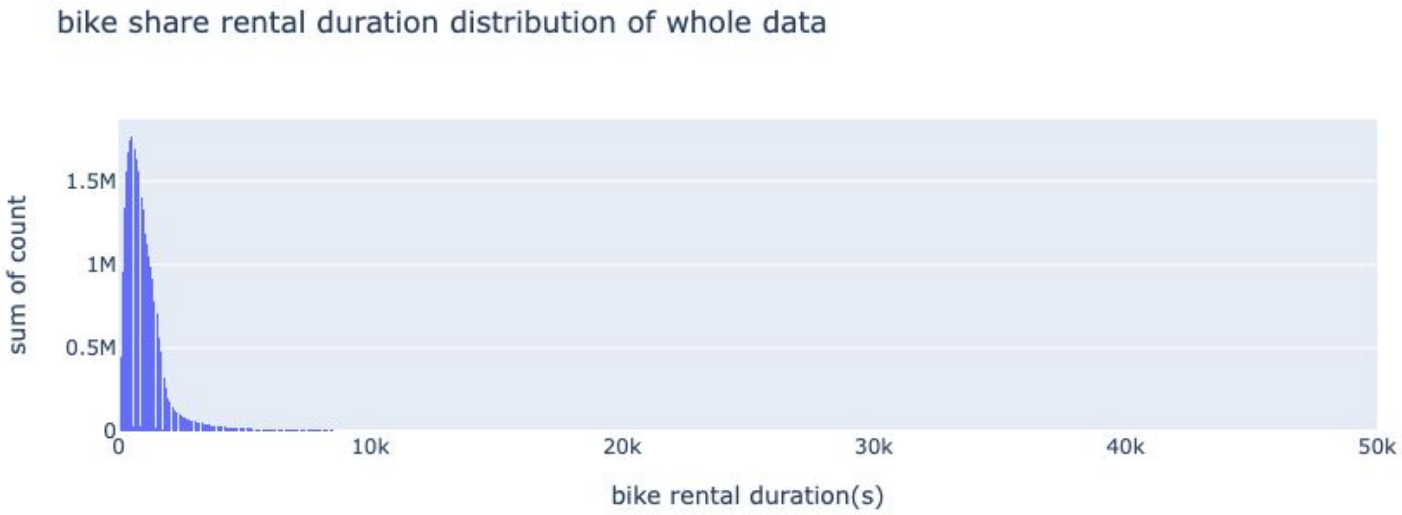


Bike Rental Duration Distribution

Analytical goal: can be used to make tiered price decision.

- Distribution of whole data
- Distribution of duration within 2h

Whole Data


bike share rental duration distribution of whole data
sum of count
1.5M
1M
0.5M
0
0
10k
20k
30k
40k
50k
bike rental duration(s)

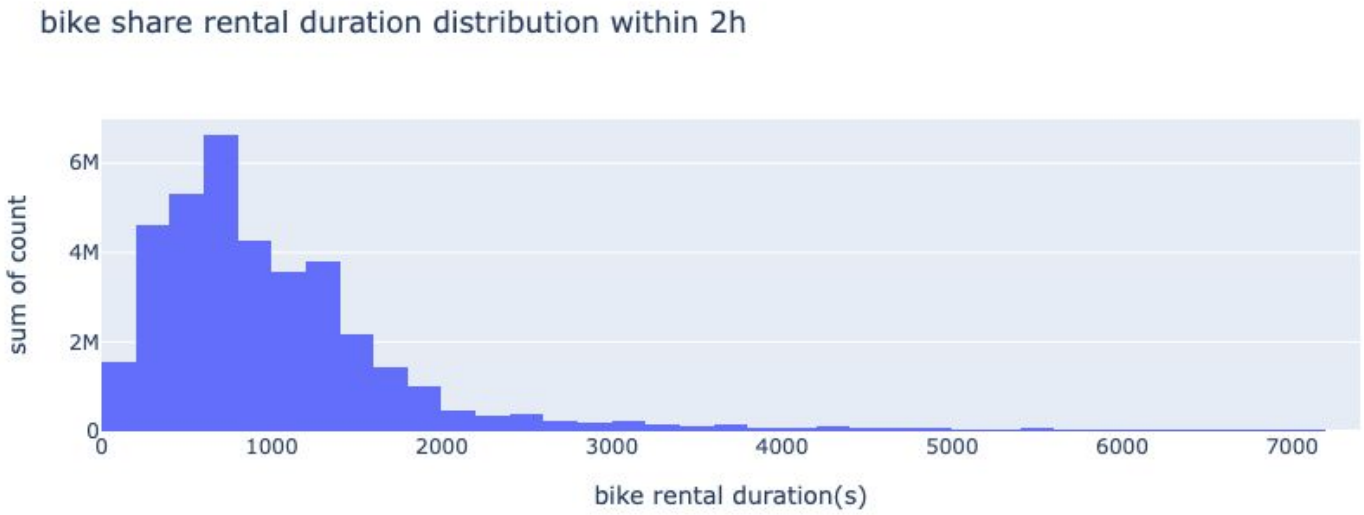
Duration within 2 hours

```
[19]: print (time.time() - start_time)
```

Last executed at 2021-12-08 21:21:03 in 73ms

Spark Job Progress

450.34646940231323

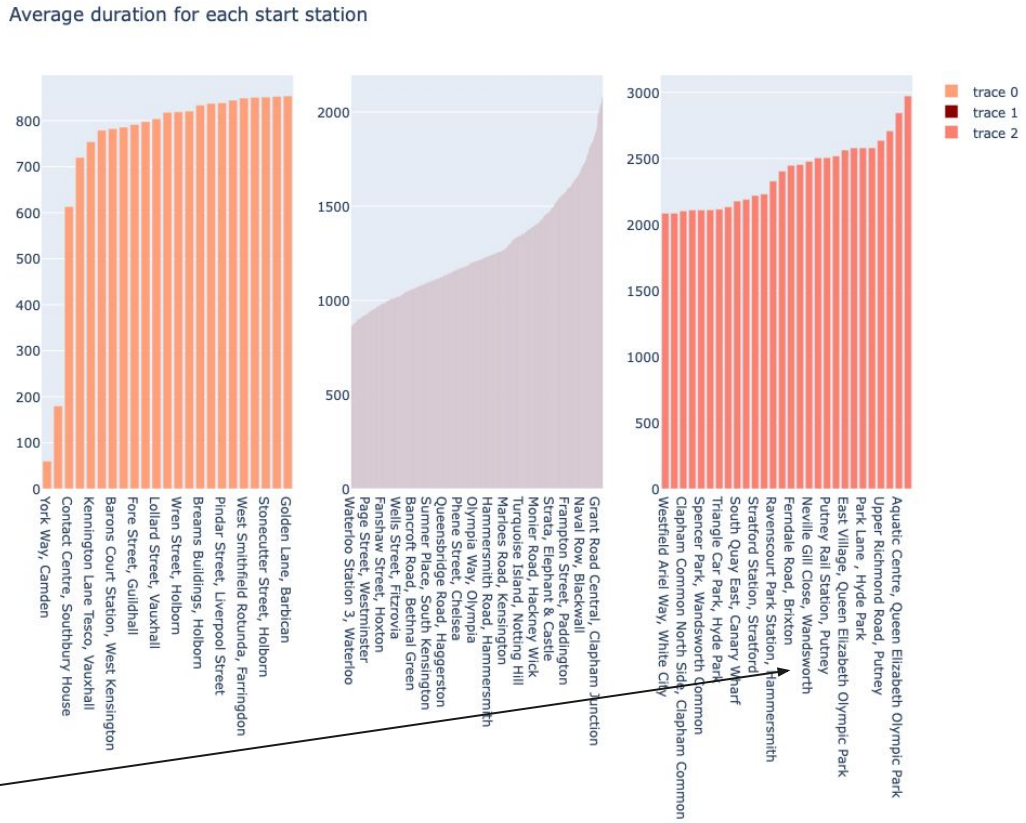


Bike Rental Duration and Start Station Location Distribution

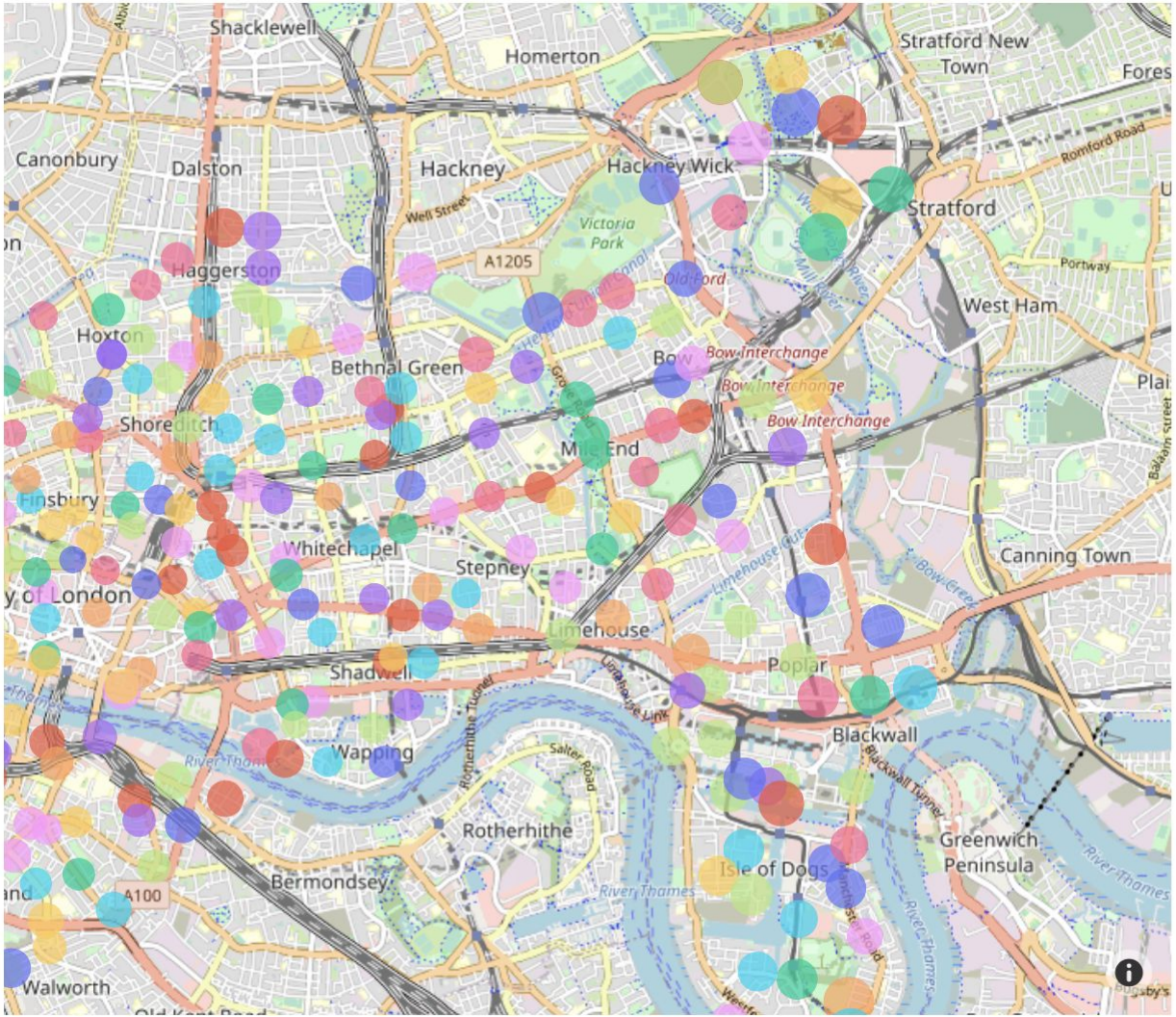
Average of duration represents how far that people go with bike sharing from the start station position

Help us planning the facilities like distribution of water fountains spot and charger stations, etc.

For the start location which spend long avg duration means they may travel for longer distance, like weekends they may spend time to hanging around the parks.



Parks HERE!!



color

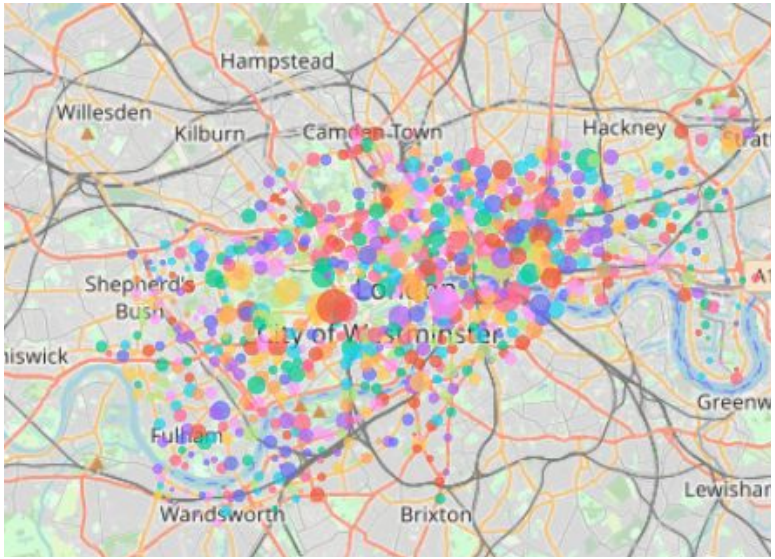
- Everington Street, Fulham
- Farringdon Street, Holborn
- Ansell House, Stepney
- Harrowby Street, Marylebone
- Westminster Bridge Road, Elephant & Castle
- Lansdowne Road, Ladbroke Grove
- Old Ford Road, Bethnal Green
- Thornfield House, Poplar
- Crinan Street, King's Cross
- Caldwell Street, Stockwell
- The Tennis Courts, The Regent's Park
- Star Road, West Kensington
- Taviton Street, Bloomsbury
- Sadlers Sports Centre, Finsbury
- Peterborough Road, Sands End
- Knaresborough Place, Earl's Court
- Mexfield Road, East Putney
- Windsor Terrace, Hoxton
- Furze Green, Bow
- Lisson Grove, St. John's Wood
- Stockwell Roundabout, Stockwell
- St. John's Road, Clapham Junction
- King Edward Walk, Waterloo
- Queensdale Road, Shepherd's Bush
- Imperial Wharf Station, Sands End
- Putney Bridge Road, East Putney
- Teversham Lane, Stockwell

Bike Rental Spatial Distribution

- Where and when people are more likely to use sharing bikes?
  - Count and hour mode of start station on map
  - Count and hour mode of end station on map

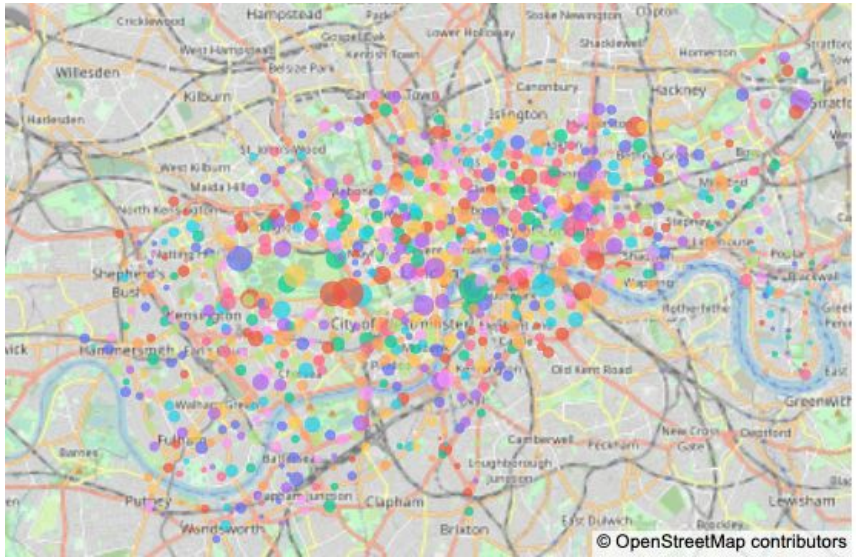
Start Station

Start_Station_HTML



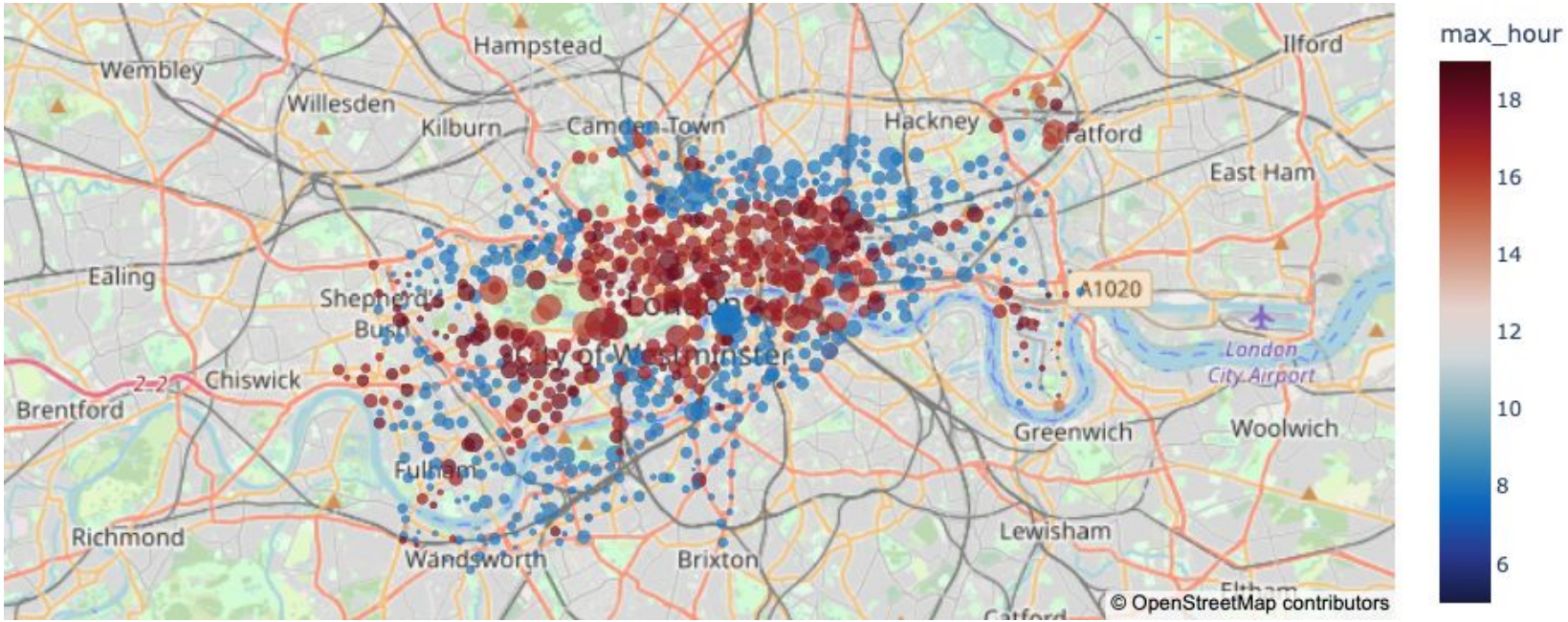
End Station

End_Station_HTML



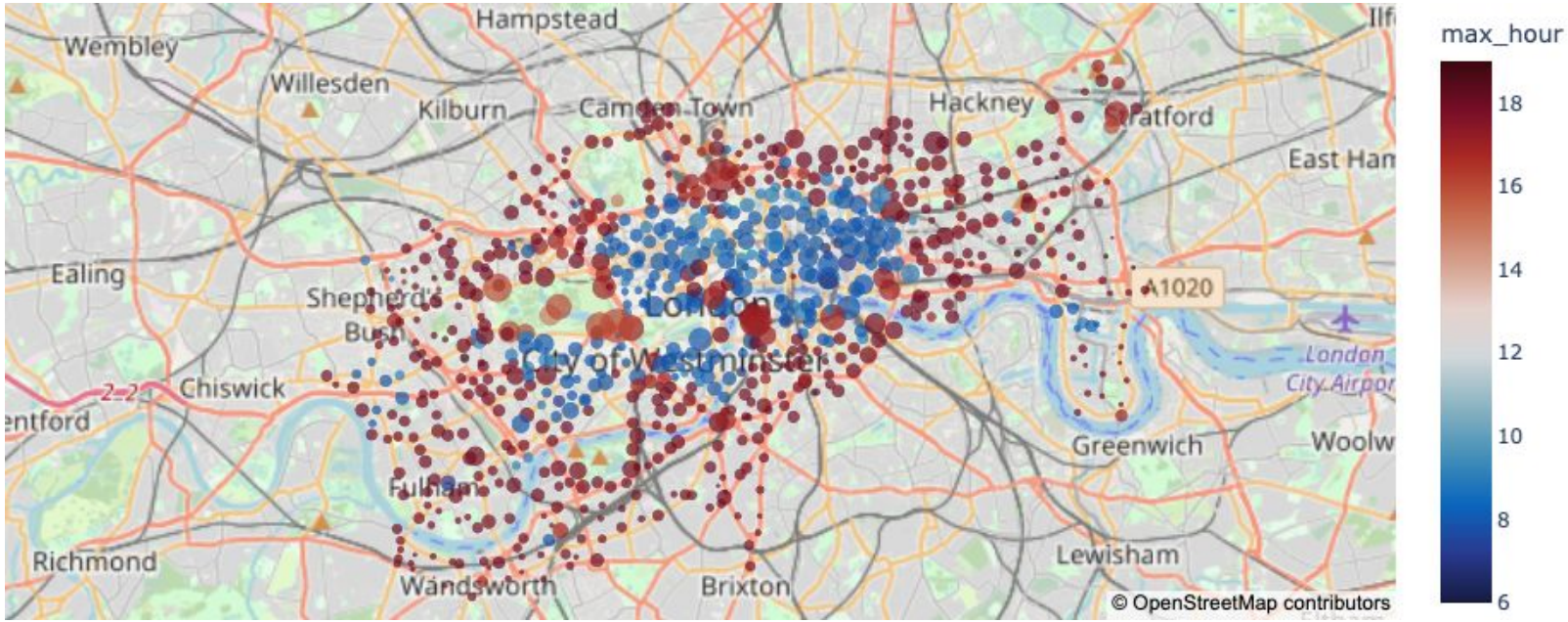
Start Station with Mode Hours

Start_Hour_HTML



End Station with Max Hour

End_Hour_HTML



Execution Time-Four plots in total

[22]:
```
endtime = time.time()
endtime- startime
```

Last executed at 2021-12-08 13:22:02 in 65ms

```
699.091676235199
```

Efficiency Improvement - Experiment

```
## Testing efficiency
start_group1 = time.time()
data_tidy = data_rdd.map(lambda x:parse_csv(x)).filter(lambda x:x[1]!='').filter(lambda x:int(float(x[1]))>0)
f = data_tidy.first()
c = data_tidy.count()
end_group1 = time.time()
```

Last executed at 2021-12-08 16:51:11 in 5m 38.14s

```
start_group2 = time.time()
data_tidy = data_rdd.map(lambda x:parse_csv(x)).filter(lambda x:x[1]!='').filter(lambda x:int(float(x[1]))>0)
data_tidy.cache()
f = data_tidy.first()
c = data_tidy.count()
end_group2 = time.time()
```

Last executed at 2021-12-08 16:53:39 in 2m 27.63s

Efficiency Improvement - Result

```
print(f"The execution time for Group 1 is {end_group1-start_group1:.2f} seconds.")
print(f"The execution time for Group 2 is {end_group2-start_group2:.2f} seconds.")
print(f"We save {(end_group1-start_group1)-(end_group2-start_group2):.2f} seconds every t
```

Last executed at 2021-12-08 17:00:13 in 61ms

▸ Spark Job Progress

```
The execution time for Group 1 is 336.19 seconds.
The execution time for Group 2 is 146.50 seconds.
We save 189.69 seconds every time we use data_tidy starting from the second action.
```

Member Cluster settings

Clusters:

msds694-emr	j-30EXQQ2BTN5EX	Starting
msds694-emr	j-1KP5LBR05M46B	Starting
msds694-emr	j-1QWEDHKR59UXK	Waiting Cluster ready

Notebooks:

msds694-nn	Ready	j-1KP5LBR05M46B
notebook-yunhe	Stopped	j-1QWEDHKR59UXK
meilin2333	Ready	j-1QWEDHKR59UXK
msds694-yoli	Stopped	j-3MYUG5P1HL4QN

Lesson Learned

- Spark is really helpful for big data analysis and each team member can work together on EMR.
- Building a data pipeline in the beginning of the project can save a lot of time.