Name: **Mark Daychman**

Data exploration project: Weather impact on bike usage in Paris

1 Motivation

I live in the 20e arrondissement of Paris and use public bicycles (Velib) nearly every day. In the morning, I often have to pass multiple Velib stations before finding a single available bike, whereas in the evening the bikes are abundant – I sometimes do not have space to park. I also noticed that when it rains or when it is cold there are more available bikes. This served as my personal motivation to explore the impact of weather conditions on the usage of public bikes in Paris.

All the code used in the project can be found here: `https://github.com/Pylyr/Big_Data-Infra_Proj/blob/main/test.ipynb`.

2 Data

Velib Metropole, as a government-funded public service, publishes all its data in open access. [3]. Unfortunately, it only publishes real-time, not historic data. However, gitHub user, Ophir Lojkine, used a custom script to collect this public data for several months, which he shared on his github page [2]. Specifically, I used his largest dataset spanning from the 26th November 2020 to the 9th April 2021, featuring nearly 10M data points. Every data point is associated to one of almost 1400 bike stations in Paris. It has the full capacity of the station as well as the total number of available electric and mechanical bikes at the time of the data point. Finally, each data point has a geolocation of the station.

I crossed this data with meteorological data I found on OpenMeteo [1]. It provides free hourly meteorological data for any point on the globe. Specifically, I collected the real and apparent temperatures, the wind speed, precipitation, cloud coverage, as well as whether the sun has set already or not.

3 Processing

As mentioned previously, the Velib data contained around 10M points and the weather dataset contained about 3200 points. To deal with such a large number of points and to make the project scalable for more data, I used PySpark for data processing.

The only available field on which I could join the two datasets is the timestamp. Unfortunately,

the Velib historic data is collected at irregular time intervals and OpenMeteo provides the data at the start of every hour. To join the two datasets, I converted their time to Unix timestamps, I crossed both tables together, calculated the absolute difference between the two timestamps and only kept the pairs with the smallest absolute differences.

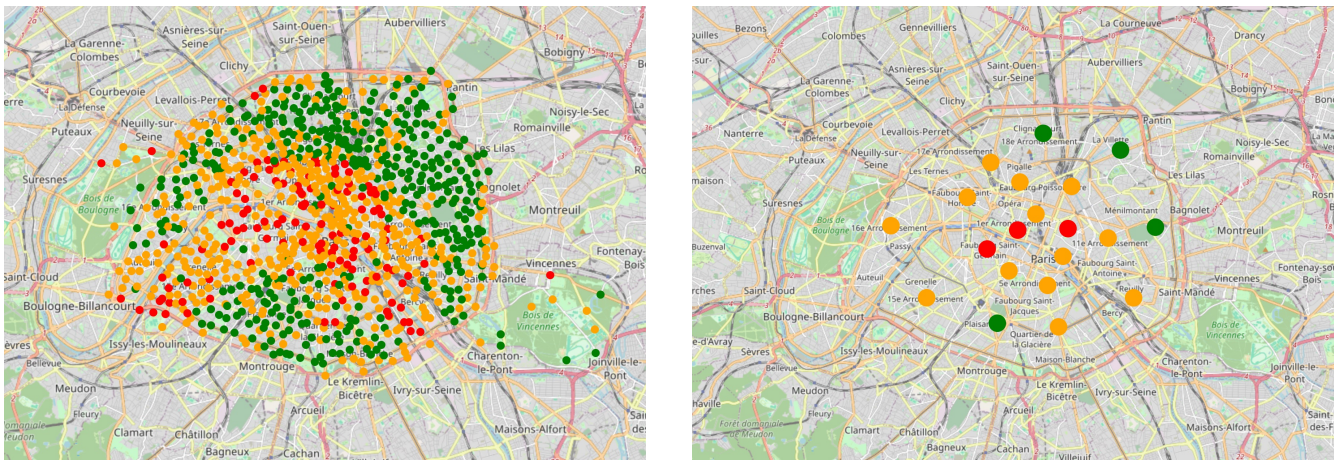
Then, I extracted all the distinct locations of Velib stations and reverse geocoded them using Google Maps API. Specifically, I was interested in grouping the data by city districts (also known as arrondissements). Google Maps does not explicitly provide the arrondissement number, but it can be extracted from the last 2 digits of the postal code. This also helps disgregard all Velibs that are slightly outside of Paris, because their post codes will not begin with a '75'.

Arrondissement	**Total Capacity**
15	3361
12	2630
13	2298
11	2252
16	2152
17	2131
20	1918
19	1860
18	1856
14	1849
8	1617
10	1573
5	1204
7	1135
9	1103
6	1003
2	742
1	712
4	706
3	390

Table 1: The total capacity of all Velib stations per arrondissement

Since different arrondissements have a different number of stations, instead of dealing with absolute numbers, I found the total number of available capacity across all stations in an arrondissement. This data is shown in Table 1. Then I grouped the data by timestamp and arrondissement to get the total number of available bikes at that time in the whole arrondissement. Divided by the total capacity of the arrondissement, you get the current usage of bikes at that time, called a **capacity ratio**. You can see the result of capacity ratio aggregation in Figures 1a and 1b. The figures show the number of available bikes per station and per arrondissment, respectively. The data points are color-coded, where green shows that the station/arrondissment is less than 20% full, orange shows that the station is less than 70% full and red shows that it is over 70% full.

Finally I used the `corr` function available in Spark to calculate the Pearson's correlation coefficient between the capacity ratio and the 5 available weather parameters. The correlation coefficients are shown in Figure 2.



(a) Color-coded capacity ratio per station

(b) Color-coded capacity ratio per city district

Figure 1: The two maps represent how full the stations/arrondissements are at 13:59 on 26 Nov 2020. Green = <20% bikes are available, Orange = <70% bikes are available, Red = >70% bikes are avaialble.

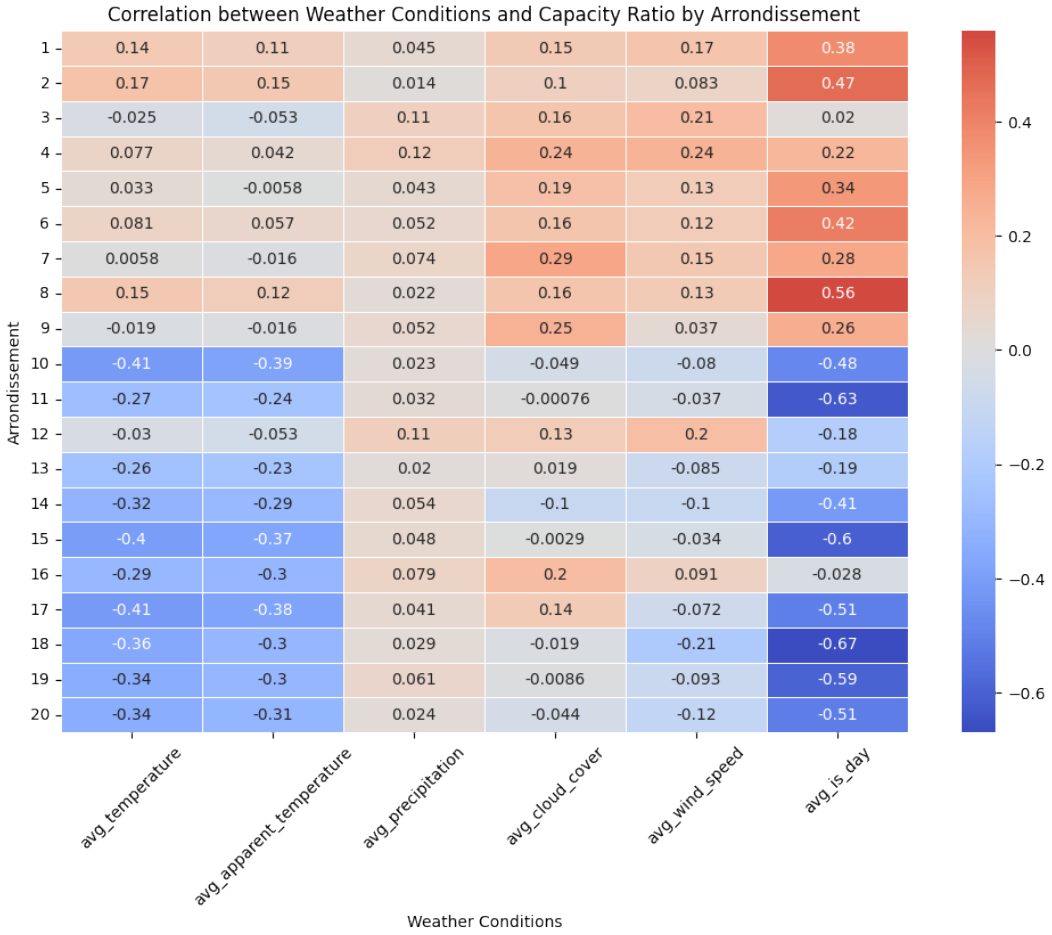


Figure 2: The heatmap of Pearson correlation coefficients between the weather conditions and the capacity ratio by arrondissement.

4 Analysis

Figure 2 provides a range of interesting insights regarding the correlation between weather conditions and bike availability in different arrondissements.

Outer arrondissements commute to the center during the day and back in the evening. The strong positive correlation between `avg_is_day` and `capacity_ratio` in central arrondissements (e.g., 1st, 2nd, 5th, 7th, and 8th) suggests that bikes are more available during the day in these areas. In contrast, in the outer arrondissements (e.g., 15th, 18th or my 20th arrondissement). This trend implies that people from these districts commute to the city center in the morning, reducing bike availability early in the day, and return in the evening, increasing availability as the day progresses.

People take significantly fewer bikes when it is cold. The negative correlation between `avg_temperature` and `capacity_ratio` across most arrondissements confirms that colder weather reduces bike usage. This trend is especially noticeable in outer districts, where bike availability increases significantly in lower temperatures, indicating that fewer people are using bikes when it is cold.

People take fewer bikes when it is windy or cloudy in the center. In central arrondissements (e.g., 10th, 15th, and 16th), `avg_wind_speed` and `avg_cloud_cover` exhibit a noticeable negative correlation with bike availability, suggesting that windy and cloudy conditions discourage biking in dense urban areas. This could be due to a combination of lower visibility, discomfort, and overall reduced enthusiasm for outdoor commuting.

People take slightly fewer bikes when it is raining or snowing. Precipitation has a mild positive correlation with `capacity_ratio`, suggesting that rain and snow have a moderate effect on bike availability. This means that while some people still opt for biking in light rain, significant precipitation levels likely lead to a decrease in overall bike usage.

5 Conclusion and Further Work

Future work could expand by using a larger time interval of at least several years. Our dataset is only limited to fall and winter months, so the analysis does not refect the impact of warm summer weather. Given how central arrondisements exhibit different behavior from the outer ones, incorporating socioeconomic data (e.g., population density, work hubs) to explain regional differences in bike usage could also be interesting. It would also be interesting to use predictive modeling to forecast bike availability based on upcoming weather conditions. Using these insights, the government and bike operators can optimize station distribution and encourage a more sustainable and equal use of bikes across the city without creating bottlenecks.

References

[1] Open-Meteo. *Open-Meteo: Free Weather API for Historical and Forecast Data*. Accessed: 2024-02-02. 2024. URL: `https://open-meteo.com/`.

[2] Loïc Vasselin. *Historique Vélib' Open Data*. Accessed: 2024-02-02. 2024. URL: https://github.com/lovasoa/historique-velib-opendata.

[3] Vélib' Métropole. *Vélib' Métropole Open Data - GBFS Service*. Accessed: 2024-02-02. 2024. URL: https://www.velib-metropole.fr/donnees-open-data-gbfs-du-service-velib-metropole.