Introduction to Data Science Coursework

Using Washington DC Capital Bikeshare (CaBi) Data

Introduction

Across the globe, large cities are home to numerous modes of transport. These include trains, cars, buses, etc. Although many of these modes of transport are the daily norm for the commuting public, they can be very costly to the environment and to one's own pocket. For example, using personal vehicles (such as cars) adds to the already very congested roads and increases the ever-growing air pollution. In addition, using public transport (such as tubes or trains) can be very costly, especially in the UK!

For these reasons, public bike share (PBS) schemes have been implemented in countless large cities all over the world. They offer an alternative mode of transport around large cities, mainly targeting shorter-distance trips. PBS schemes have been very successful as they are a cost effective alternative, they allow users to engage in exercise and they are much more sustainable.

Therefore, studying PBS schemes is crucial as it allows local governing bodies to identify the efficacy of their scheme. Additionally, analysing the data collected from a scheme helps highlight any user trends. This allows improvements to be made that are tailored to the public and how they use the scheme.

This report focuses on data collected from a PBS scheme used in the Washington DC.

Data

The data to be analysed in this report is about bike rides undertaken by users of Capital Bikeshare (CaBi), a PBS scheme in Washington DC. The data consists of all the rides undertaken in 2019 and 2020 and was taken from this website: https://ride.capitalbikeshare.com/system-data (https://ride.capitalbikeshare.com/system-data)

Clean Data

Initially the data was cleaned into the following columns:

- **Duration** – type: duration, indicates the duration of ride (a calculated column: End Date – Start Date)
- **Start Date** – type: datetime, indicates the start date and time
- **End Date** – type: datetime, indicates the start date and time
- **Start Station Name** – type: factor, indicates the starting station name
- **End Station Name** – type: factor, indicates the ending station name
- **Member Type** – type: factor, indicates the member type (member or casual)
    - member (Annual Member, 30-Day Member or Day Key Member)
    - casual (Single Trip, 24-Hour Pass, 3-Day Pass or 5-Day Pass)
- **Start Latitude** – type: double, indicates the starting latitude
- **Start Longitude** – type: double, indicates the starting longitude
- **End Latitude** – type: double, indicates the ending latitude
- **End Longitude** – type: double, indicates the ending longitude

Excluded Data

- **Negative durations** - as these are impossible and therefore incorrect
- **Durations that were < 60 seconds** – as on the website these were stated as “potentially false starts or users trying to re-dock a bike to ensure it’s secure”
- **Durations that were > 30 minutes** – as until 2021 members could take unlimited rides of durations shorter than 30 minutes

The following variables were excluded from the data set as they were not used or considered in any analysis:

- **Start Station ID**
- **End Station ID**
- **Bike Number**
- **Ride ID**
- **Bike Type**
- **Is Equity**

Questions

In this report, we will investigate the following questions.

Question 1

Which stations are the busiest?

To operationalise this question, the variable we will focus on is the total number of departures. We will also specify the top five busiest starting stations. The departures included in this question will be from both years (2019 and 2020). Therefore the operationalised question becomes:

Which five CaBi stations in Washington DC had the highest number of departures between 2019 and 2020?

Question 2

When do CaBi users tend to ride?

To operationalise this question, the variable we will focus on is the average number of rides. The time frame we will specify is the hours of the day (0 - 23 hours). The rides included in this question will be from both years (2019 and 2020). Finally we will also distinguish between members and casual users. Therefore the operationalised question becomes:

Did the type of membership affect when CaBi users tended to ride, on average, across a day between 2019 & 2020?

Question 3

Does the month of the year affect how long people ride for?

To operationalise this question, the variable we will focus on is the average duration of a ride. The time frame we will specify is the hours of the day (0 - 23 hours). The rides included in this question will be from both years (2019 and 2020). Finally we will also distinguish between months of the year. Therefore the operationalised question becomes:

Did CaBi users, on average, increase the duration they rode for, across a day, depending on the month of the year between 2019 and 2020?

Question 4

Did the first COVID-19 lockdown lower the use of CaBi bikes?

To operationalise this question, the variable we will focus on is the total number of rides. The time frame we will specify is the months of the year. Finally we will also distinguish between each year as the first COVID-19 lockdown happened in 2020. Therefore the operationalised question becomes:

Did the first COVID-19 lockdown lower the use of CaBi bikes across the year 2020 compared with 2019?

Question 5

Does starting location affect the duration of the journey?

To operationalise this question, the variables we will focus on are start latitude and start longitude. The rides included in this question will be from both years (2019 and 2020) and will be focused on members as this ensures that ride duration is limited between 1 and 30 minutes. Finally we will distinguish by the duration of the ride. Therefore the operationalised question becomes:

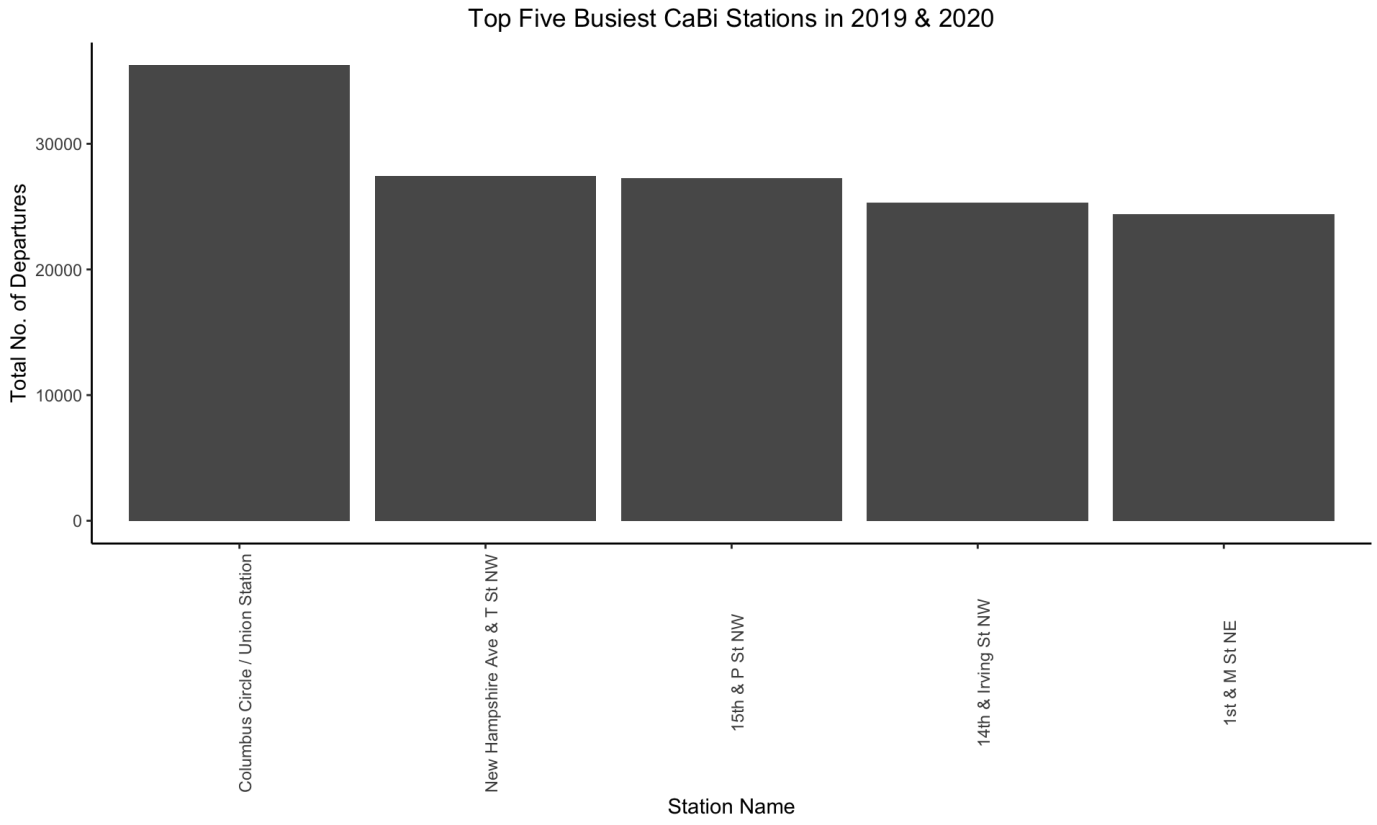
Did the starting location (starting latitude and longitude) of CaBi users affect the duration of their journeys between 2019 and 2020?

Analysis

Answer to Question 1

Which five CaBi stations in Washington DC had the highest number of departures between 2019 and 2020?

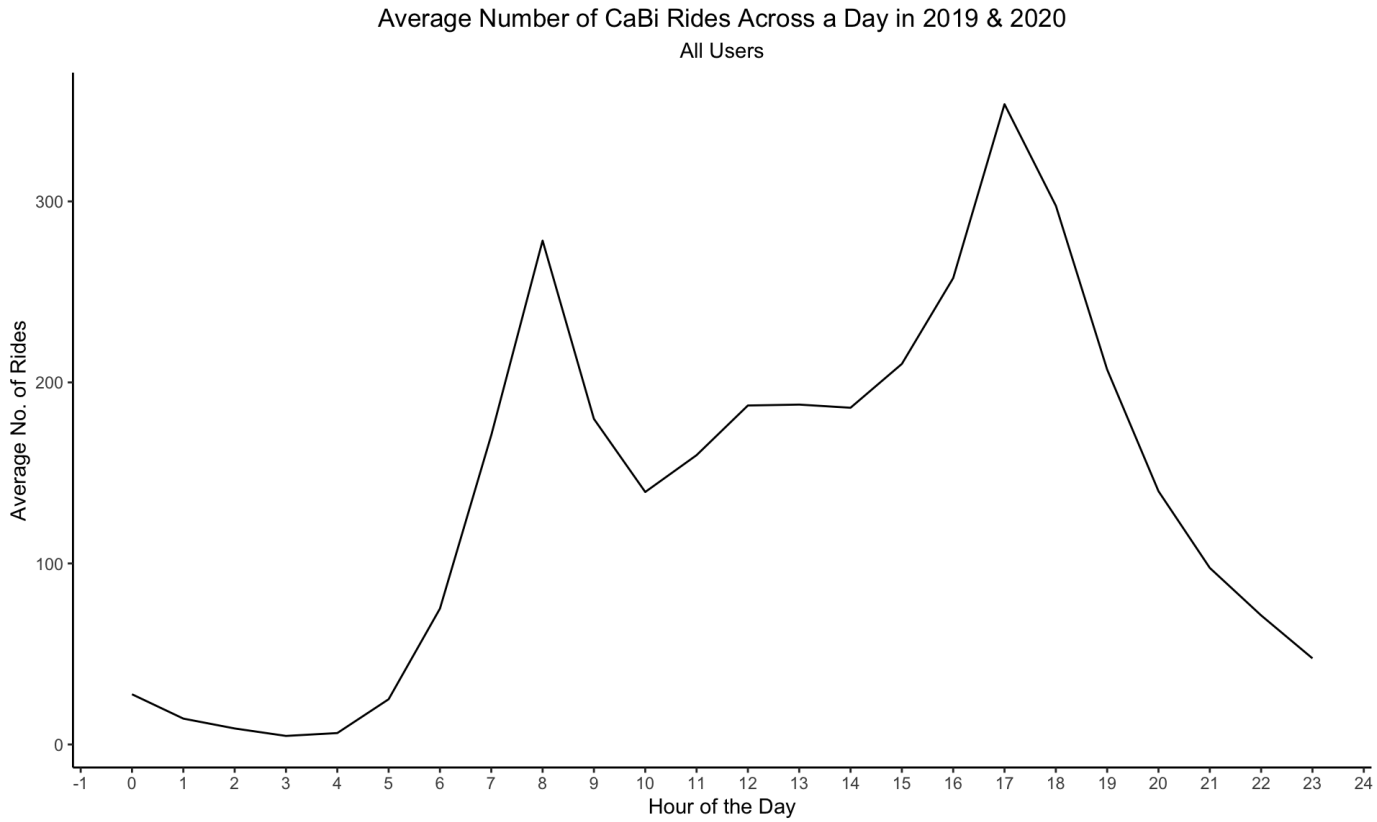
The figure below identifies the five busiest departure stations. This insight could be used by the local governing body to inform them where to deploy more bikes and also where to install more bike docking stations. It also indicates where the highest bike congestion areas could be. This suggests that the infrastructure for bike riding in and around these stations should reflect this.

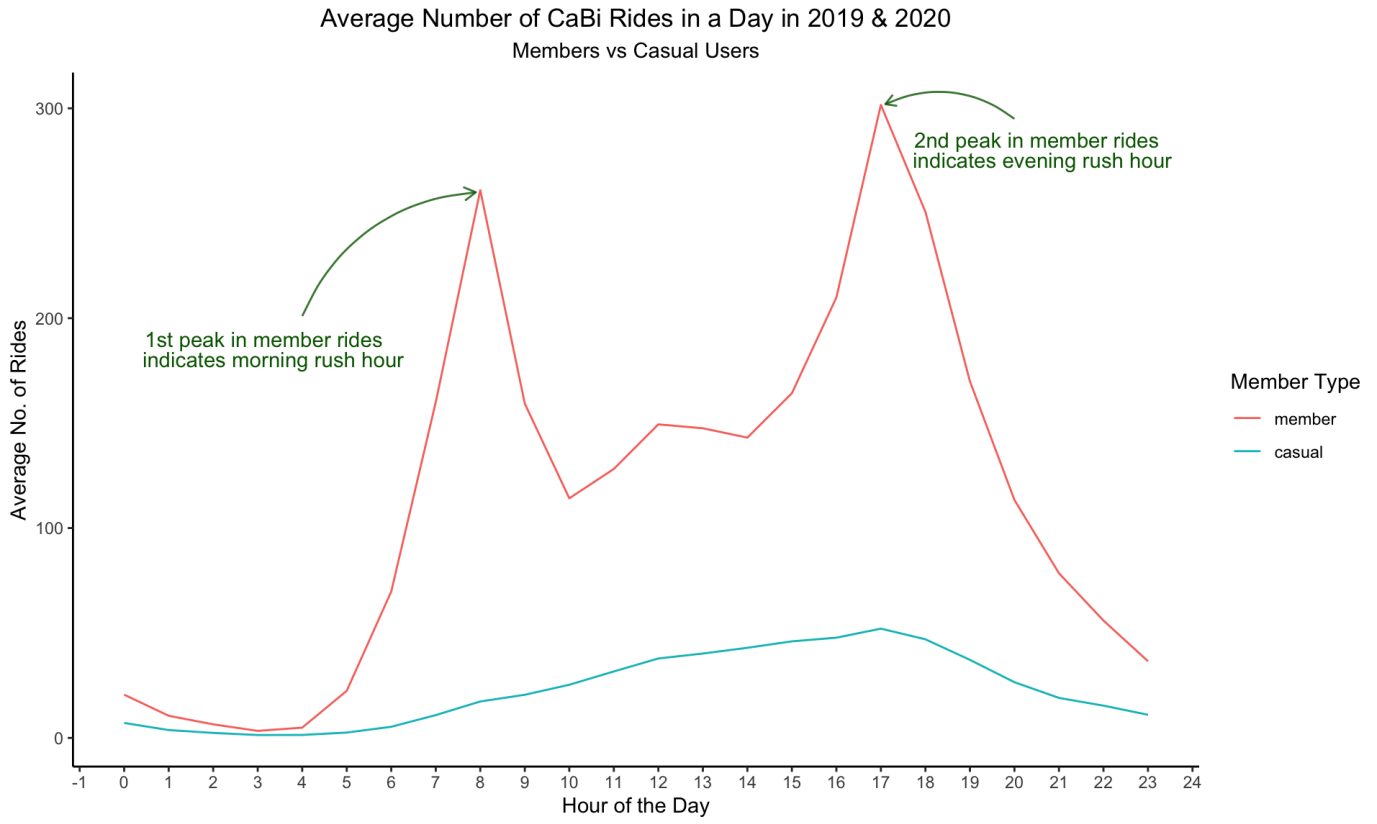


Answer to Question 2

Did the type of membership affect when CaBi users tended to ride, on average, across a day between 2019 & 2020?

The first figure below displays the trend for rides per hour, for all users. The second figure below displays the trend for rides per hour, divided into members and casual users. As you can see the second figure displays a huge difference in the trends for members and casual users. In the first plot this insight is completely hidden. This reinforces the need for the different membership schemes. The members trend tells us that they used the bikes to commute to work and to get around the city on a daily basis. This is because there are two peaks (one for each rush hour) and also their average number of rides is clearly much greater than the casual users. The casual users trend tells us that their use increased as the day went on and peaked in the early evening. Overall this figure has also identified when the busiest periods of the day were and this could inform when more policing should happen.



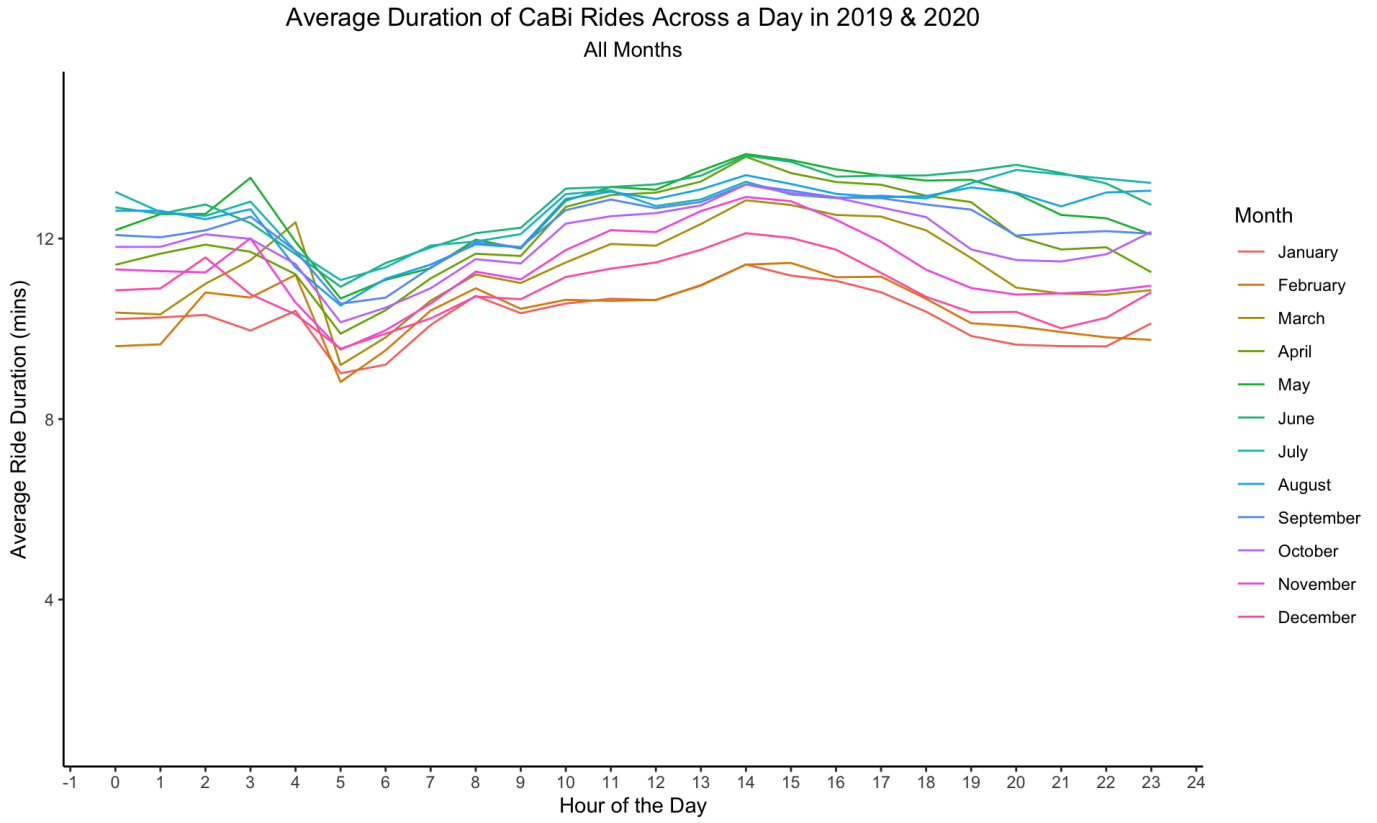


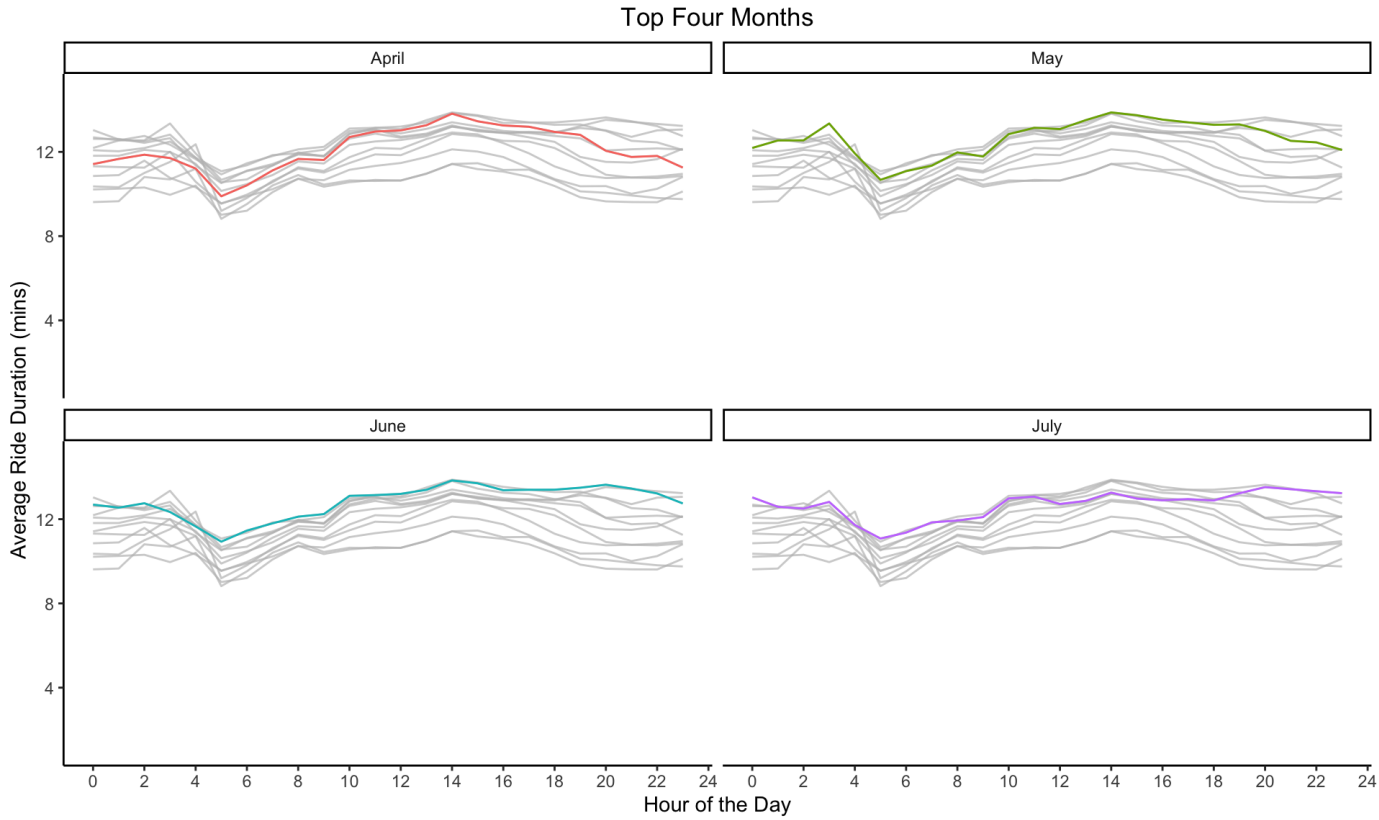
Answer to Question 3

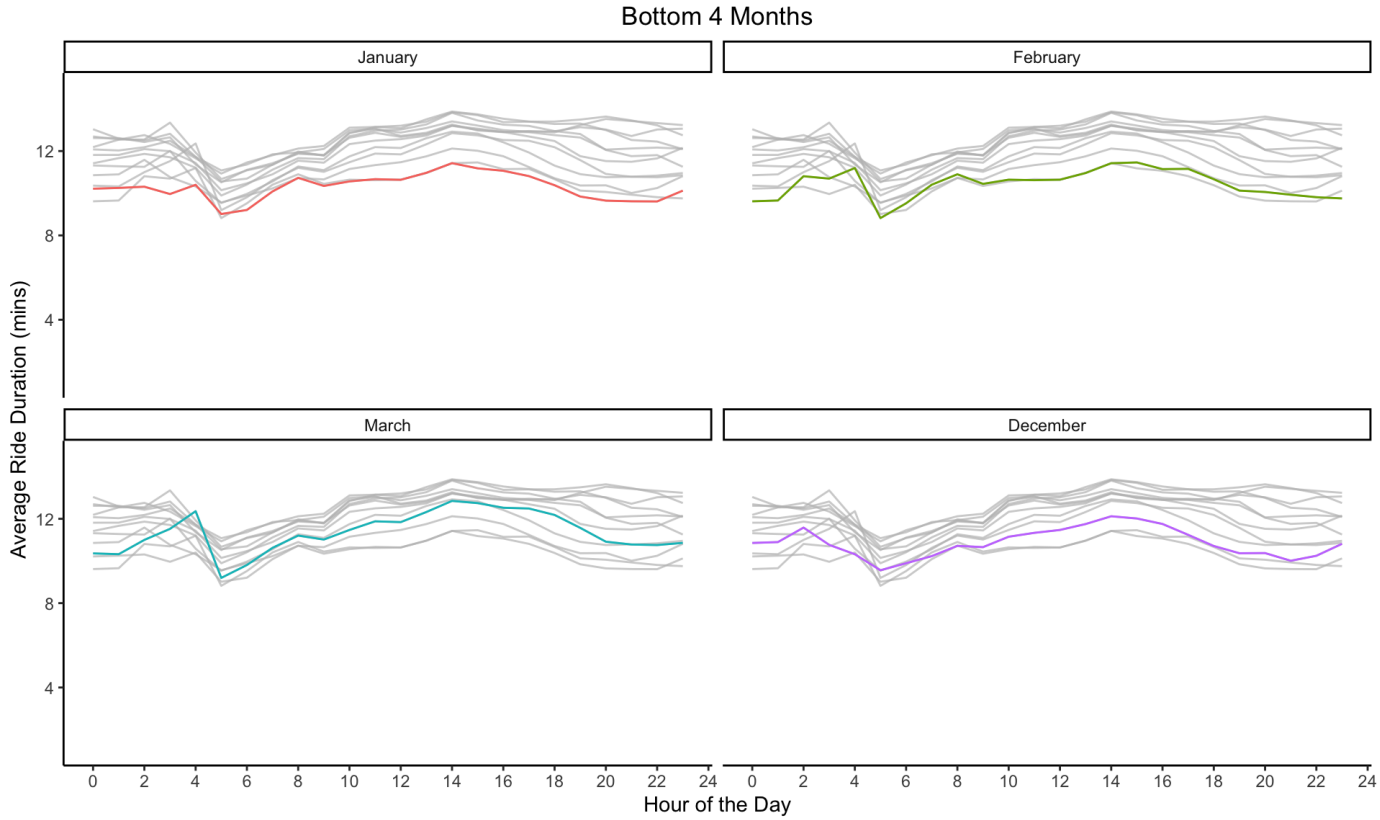
Did CaBi users, on average, increase the duration they rode for, across a day, depending on the month of the year between 2019 and 2020?

The first figure below shows us that the average duration of a rides across the day sat within a range of around 8 minutes to 14 minutes. It also shows us that the duration of rides tended to be longer in the summer months. This is likely due to the increase in temperature and sunny weather. This contrasts to the lower duration of rides, on average, in the winter months, likely due to the colder weather and less light hours. This

finding is emphasized in the second and third figures. The second figure highlights the 4 months with the highest peak ride duration (April, May, June and July). Whilst the third figure highlights the 4 months with the lowest trough of ride duration (January, February, March, December).



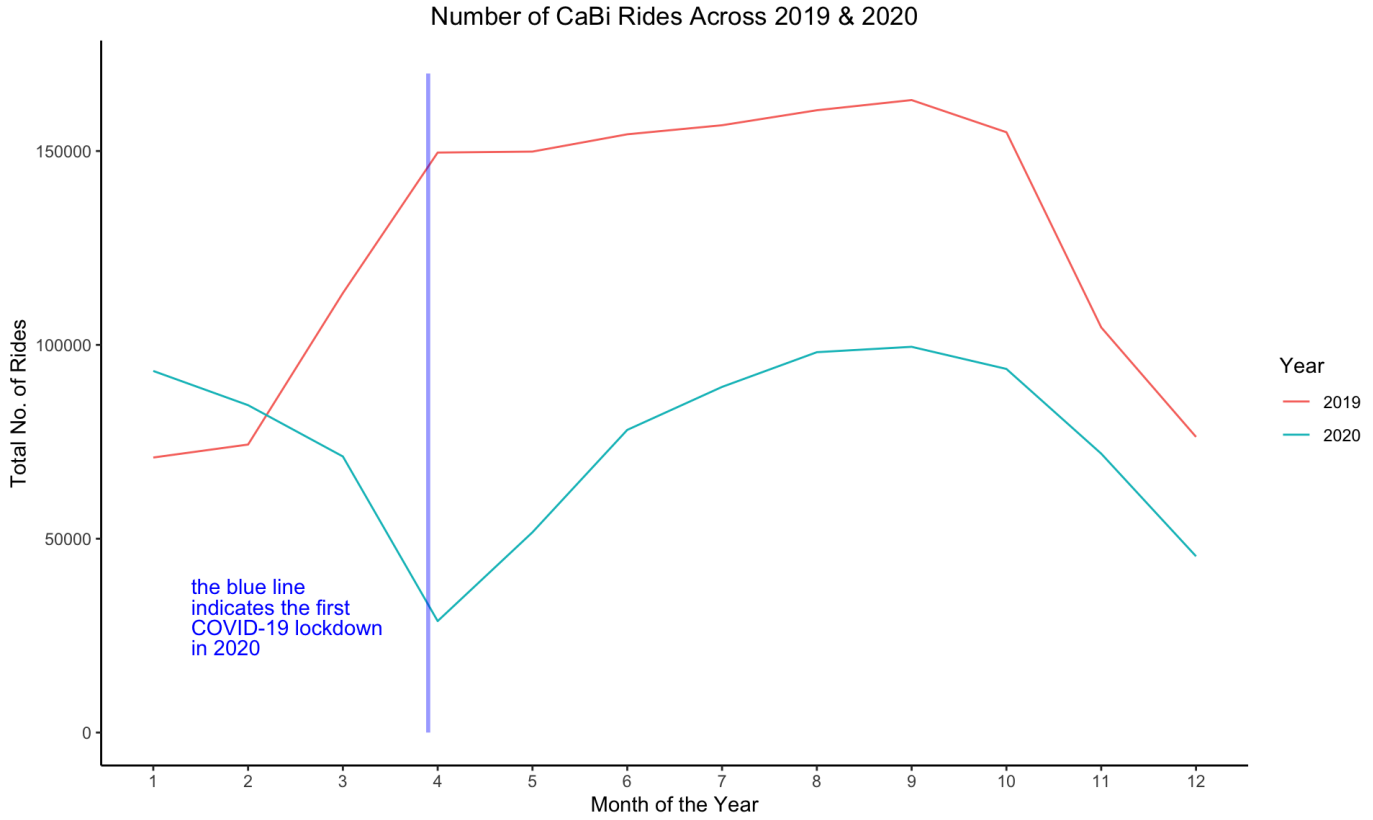




Answer to Question 4

Did the first COVID-19 lockdown lower the use of CaBi bikes across the year 2020 compared with 2019?

The figure below displays a large reduction in the number of rides taken in 2020 compared with 2019. The key difference in the two lines (2019 and 2020) is the period of time before and after the first COVID-19 lockdown (30th March 2020, indicated by the blue line). As you can see in 2019 the number of rides taken increased a lot from February to April. In 2020 over this same time period the number of rides taken decreased a lot. A reason for the decrease in rides starting before the lockdown (January to March) may be due to people already beginning to stay at home because COVID-19 outbreaks were spreading across the globe at this point. The reason for the increase in rides very shortly after the the first lockdown could be a mixture of lockdown rules being lifted quickly and people becoming fed up of staying inside and disobeying the rules.

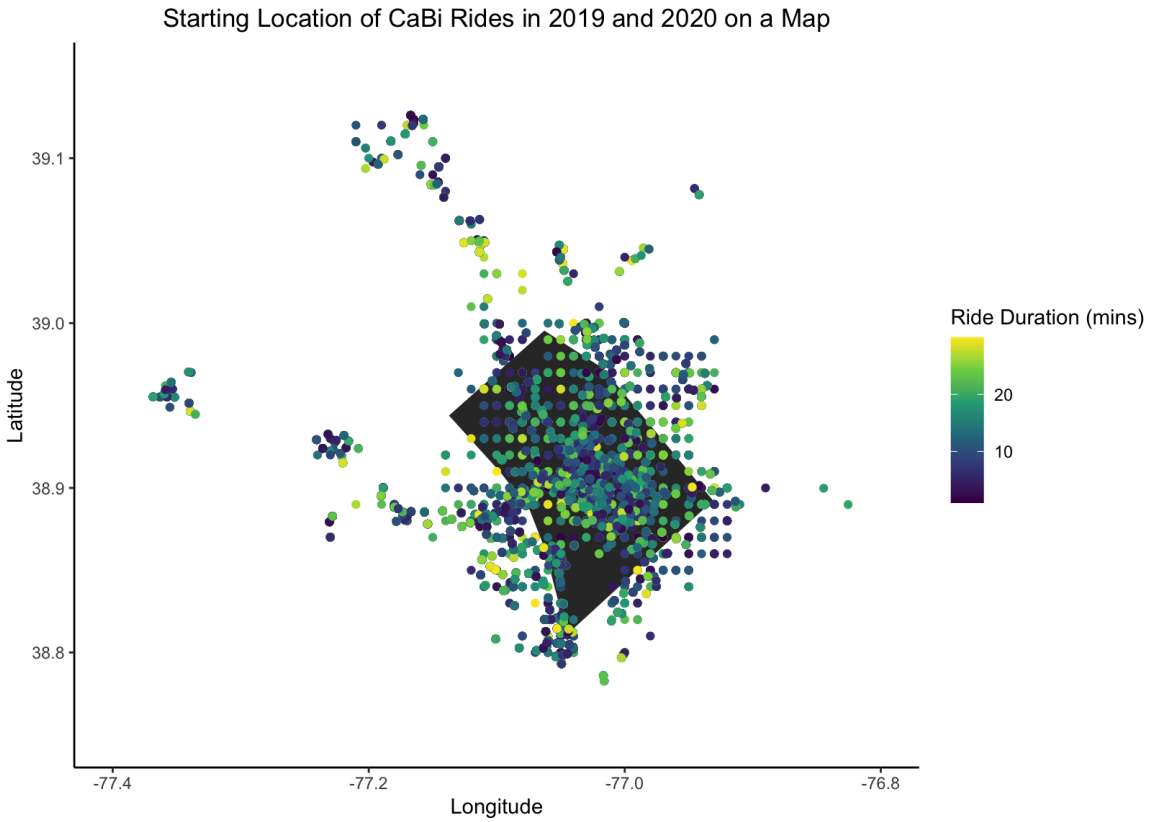


Answer to Question 5

Did the starting location (starting latitude and longitude) of CaBi users affect the duration of their journeys between 2019 and 2020?

The figure below displays the starting location of all the rides as if you were viewing it on a map. The shape underneath the points is Washington DC (aka the District of Columbia). The figure shows how CaBi bikes were not contained only within the city itself but they also expanded to the neighbouring states and districts. It is also clear that the duration of a ride did not depend on the starting location as the colour of the points are scattered randomly around the figure.

For this analysis there were many missing latitude and longitude values. These rides were chosen to be discarded as the latitude and longitude information is not there so they could not be plotted in the figure.



Conclusion

In summary, the results give a great deal of insight into the trends of CaBi bike users in 2019 & 2020. Firstly they highlighted the busiest stations, the busiest times of the day and also the busiest periods of the year. This information could allow local governing bodies to tailor specific infrastructure and traffic control to improve the CaBi service. The results also highlight how the CaBi scheme was affected by COVID-19 and that it stretches far across Washington DC and the neighbouring districts.

This analysis has also ignited many more questions about CaBi bike users in 2019 & 2020. For example, why did the number of rides start to drop before the first COVID-19 lockdown? and why did the number of rides increase rapidly after it? In addition, to find out if the differences between groups were meaningful further statistical analysis could be performed. For example, was the average duration of rides in the summer months significantly higher than in the winter months?