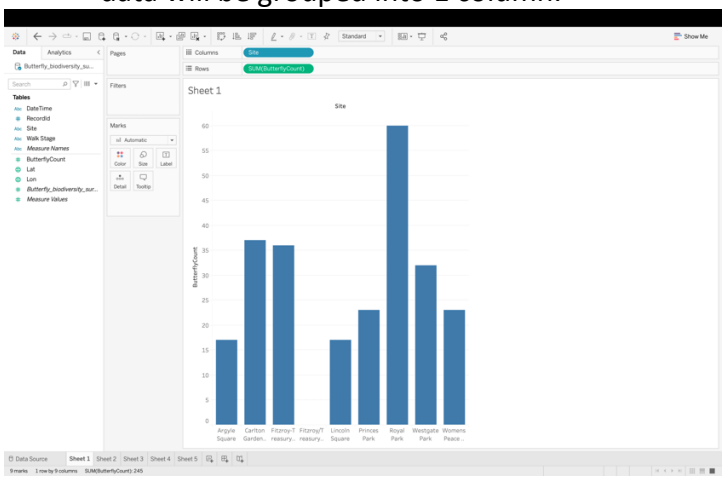
Cheng Ho Wang
32353391

FIT5147 Programming Exercise 1 Report

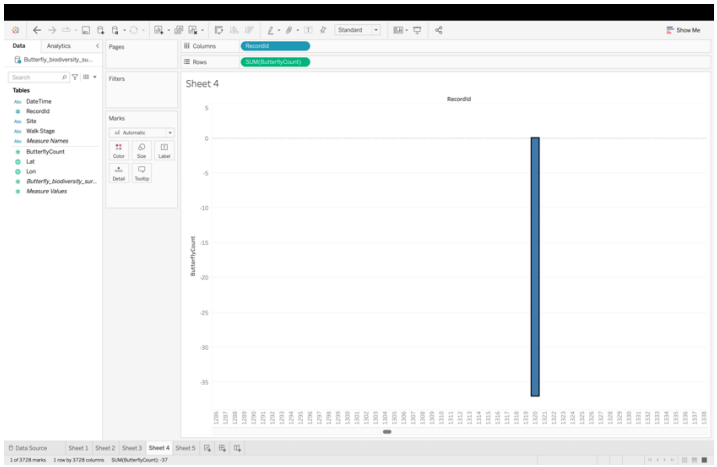
Data checking and cleaning:

3 Errors that I found in the dataset:

1. There is a difference in the naming format in the column of the site. “Fitzroy-Treasury Gardens” and “Fitzroy/Treasury Gardens” belong to the same thing. Since their formatting is different, Tableau categorizes them into a different column. Therefore, two different columns exist in the graph. To correct it, I have used R to rename all the “Fitzroy/Treasury Gardens” to “Fitzroy-Treasury Gardens” so that the data will be grouped into 1 column.

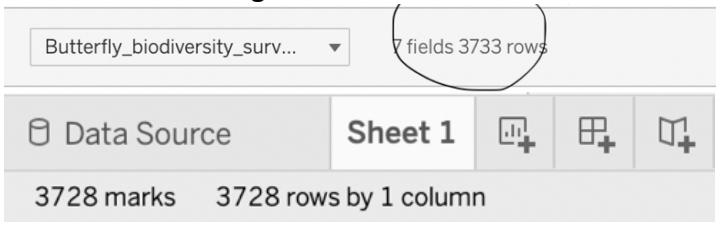


2. The count of butterflies should be always greater than or equal to 0. A negative number should not appear. The data 1320 is unusable since the count of the butterfly is -37. I have used the filter function in R to get rid of the negative number. The condition is set to count >= 0.



3. There is duplicated data from the original file. There are 3733 rows for the whole data but when we get the recorded into the sheet of tableau, it removed the

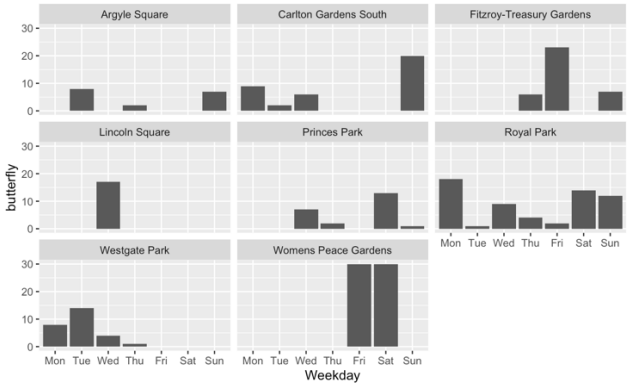
duplicated row automatically. As a result, there are only 3728 rows left. For removing the duplicated data, I used R to read the dataset and then use the unique function to get them cleaned dataset with no duplicated data.



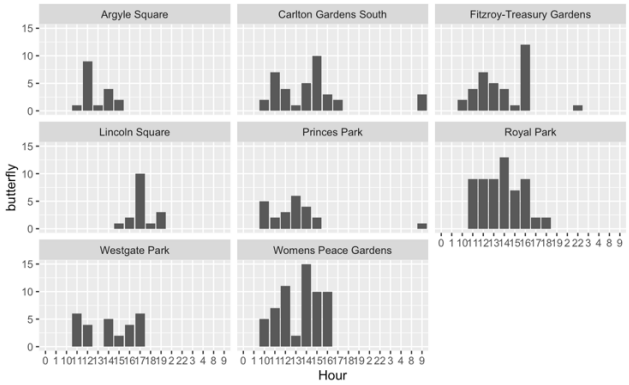
Data exploration:

Q1 Compare and contrast the number of butterflies observed in each area (Site). In what ways does this show that the sites are similar and/or different from each other? Consider this on both an hourly and day-of-the-week timescale.

For the day-of-the-week timescale, the data is randomly distributed. At Carlton Gardens South, Fitzroy-Treasure Gardens, Princes Park, and Womens Place Gardens, more butterflies are appearing on the weekend. However, in Lincoln Square, Argyle Square, and Westgate Park, there are more butterflies appear on weekdays.

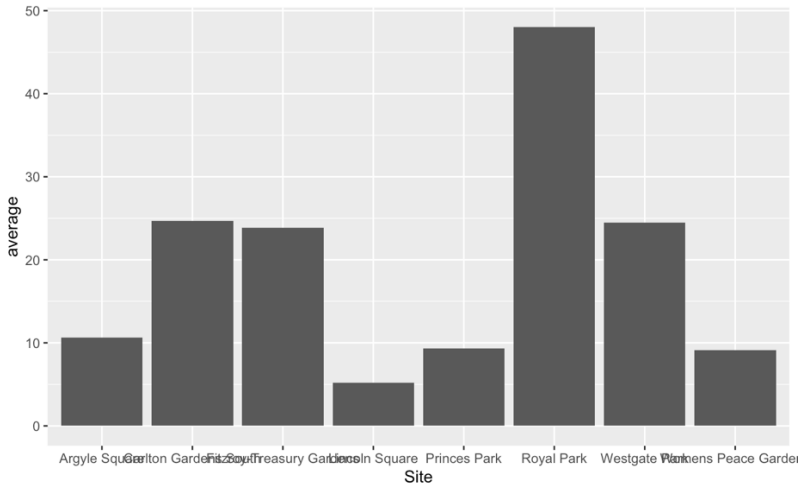


For the hourly time scale, butterflies appear only in the daytime, and there are more butterflies, especially at around 1 pm to 2 pm. For Argyle Square and Princes Park, the number of butterflies in the morning is the largest.



By using a bar chart, we can compare the number easily with a glance. Then, I have put them into a facet wrap where we can group some of the data into a small graph, for example, I have used Site to group the data here. After that, we can compare the data between different sites.

Q2: Compare and contrast the number of records for each area (Site) per hour. What does this tell you about the use of the sites for the study? In what ways does this support, challenge, or change your conclusions for the first question?



For the number of records per hour at each site, there is a really big difference. There are most records at Royal Park and least at Lincoln Square. It means that there is a bias created at the stage of data collection. The number of data collected is a huge difference. Or we could say the frequency of recording data is not the same.

From the graph above, we can see Lincoln square got the least number of data while Royal Park got the most. The maximum and minimum value has a difference of 40 which is not a good thing.

Therefore, when we look back to question 1, the data in Lincoln square are different from others. For example, there are more butterflies found on Tuesday but not on other days. Therefore, a bias exists in this case because there are not enough data to get a reasonable distribution for the whole dataset.

On the other hand, Royal Park got the most data which can give us an objective statistic. We can do the analysis based on the distribution.