Github API Interrogation and Data Visualisation

For my project, the question I decided to explore was: Are there any trends in the number of commits to repositories on certain days or certain times? These differences could occur for many reasons and I intended to try find any trends or abnormalities in when programmers are committing to different repositories. I decided to focus on one repository for this project(https://github.com/pksunkara/octonode), however the code worked for any repositories that were tested.

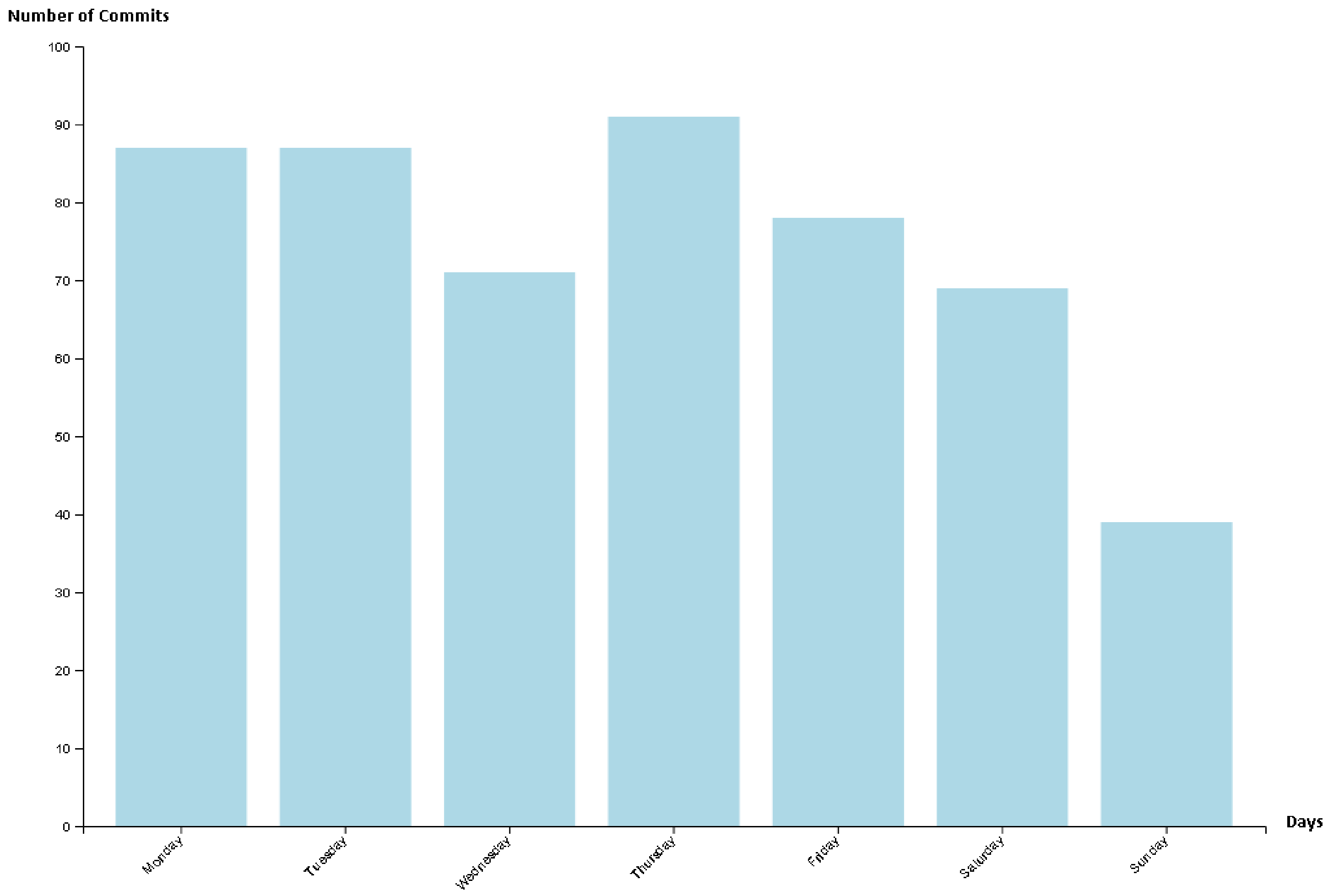
The project has three sections, the interrogation of the Github API and storing of the extracted data, the generating of CSV files for the data, and the visualisation of the data using HTMLs and the D3 library for Javascript.

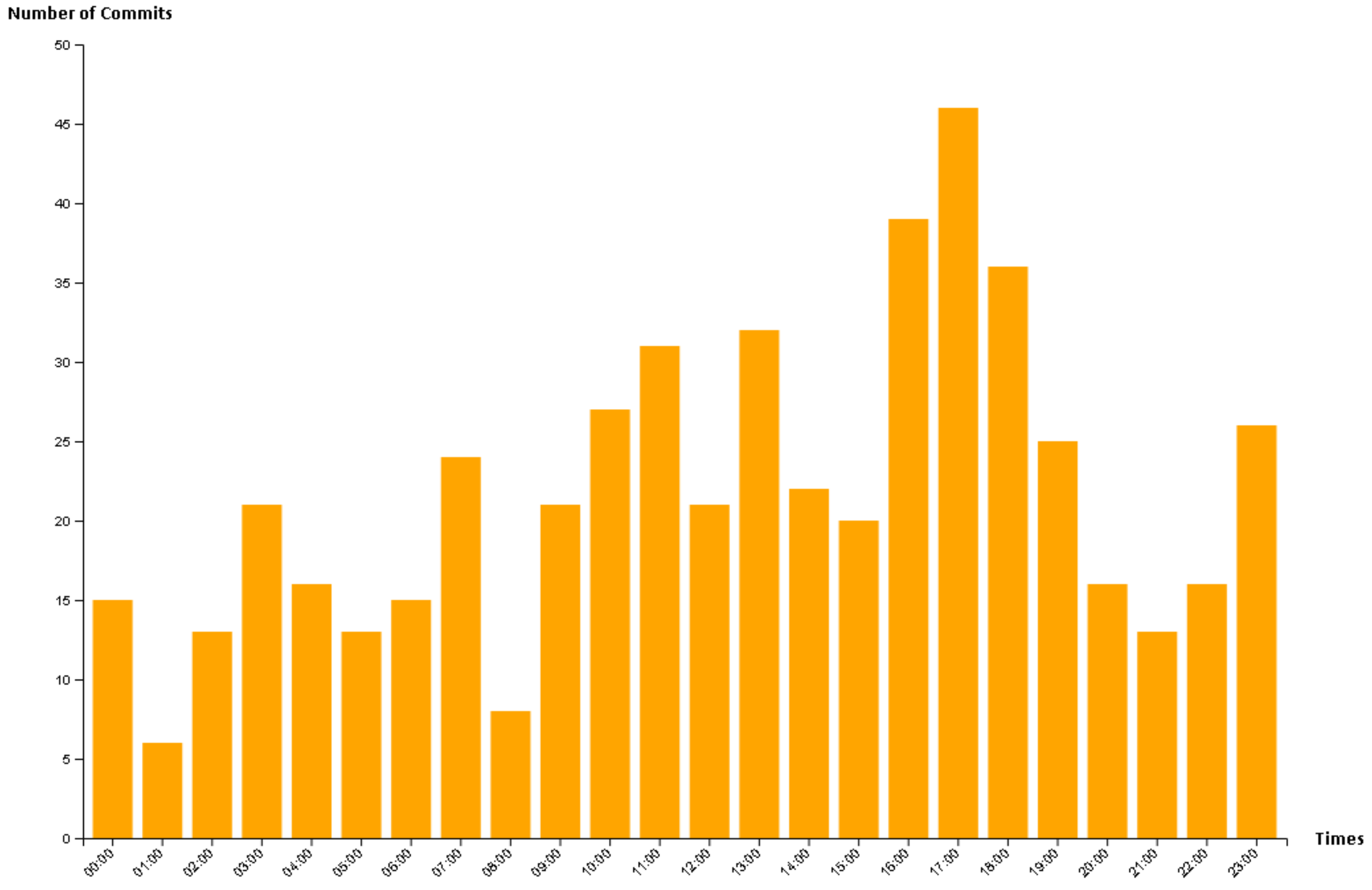
The first section of the project is the API_Interrogation.py file. Here, a SQLLite table is created with the required columns. Then, data is extracted from the ‘commits’ and ‘contributors’ JSONs for the repository from the Github API. This data is parsed to find the number of contributors, commits, and the date and time of each commit. The date, weekday and time are stored in the table.

In the CreateCSVs.py file, the data is put into CSV files to be read by the D3 HTMLs which generate visualisations of the data. CSV files are created for bar charts of the weekdays and times of commits, and for scatter plots comparing the amount of commits for a given weekday at each hour of the day.

The final section of the project takes the CSV files and generates graphs using the D3 Javascript library. There are 2 bar charts and 3 scatter plots. Two of the scatter plots are breakdowns of another scatter plot to view the data more clearly and notice some interesting trends.

The following bar charts show the amount of commits on each day of the week and the amount of for every hour of the day respectively.

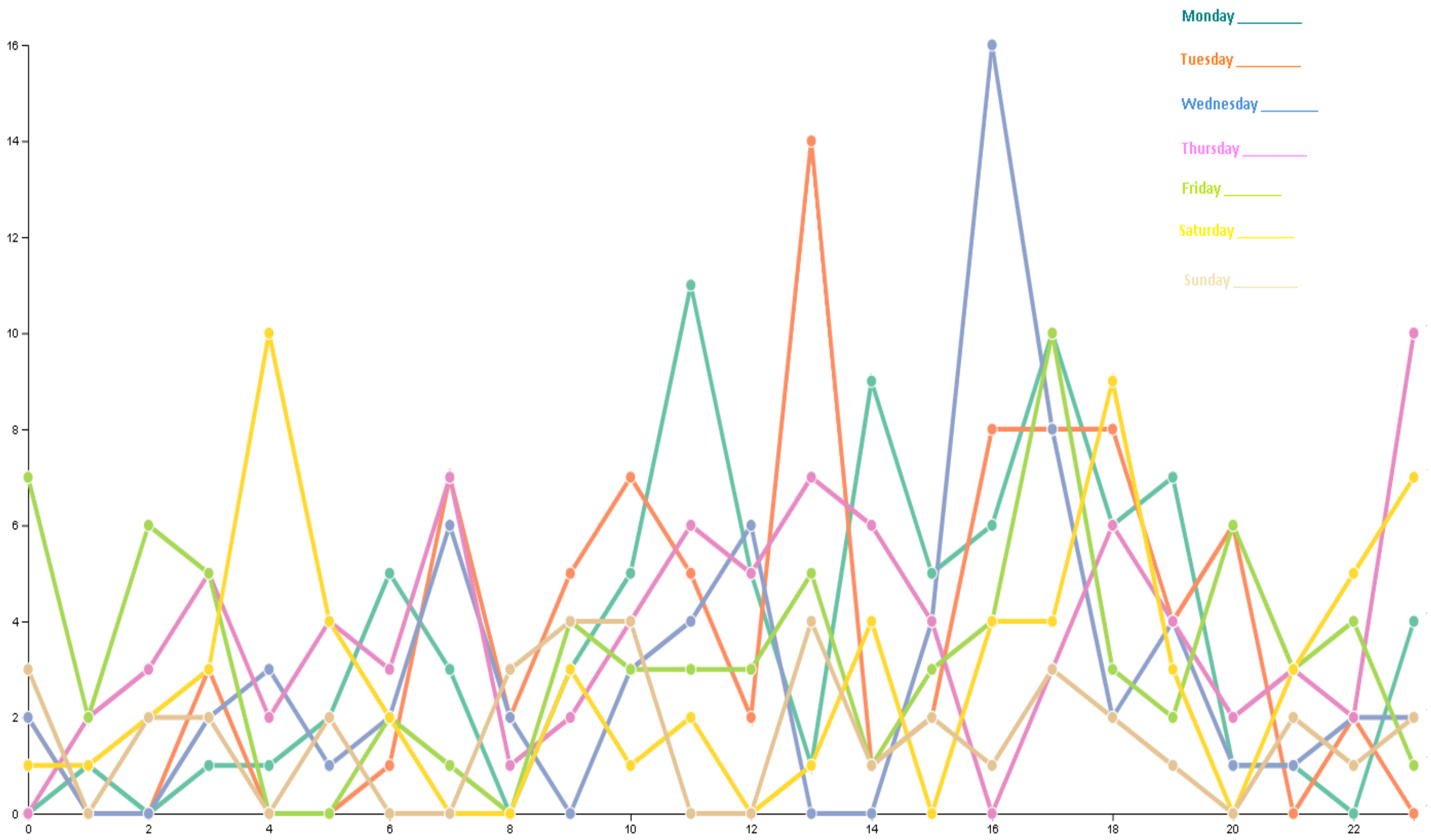

Number of Commits
100
90
80
70
60
50
40
30
20
10
0
Days
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday



The days of the week graph shows a trend of more commits near the beginning of the week which gradually decline as the week progresses. As expected, there are fewer commits over the weekends, however there are less commits than expected on Fridays when work would likely be finishing for the week. There is also a noticeable dip in the number during the middle of the week on Wednesday.

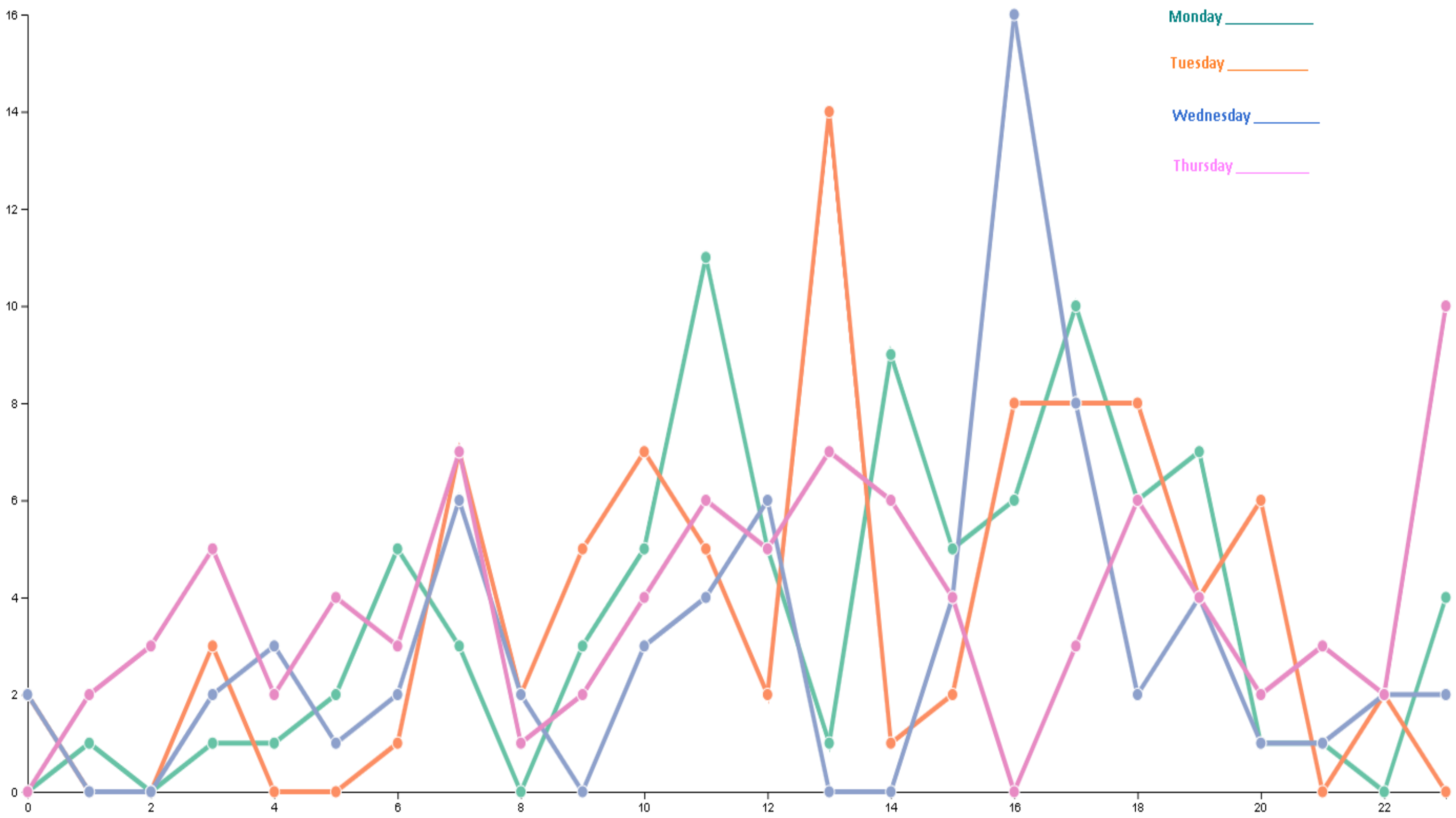
The hours of the day graph also shows some interesting trends. Most commits are between 16:00-18:00 at the end of the working day as usual. However, the graph is not as skewed to this point as expected, with many commits happening in the middle of the night. Interestingly, the distribution is fairly even overall besides the spike at the end of the workday. There is no noticeable drop in number of commits during the night.

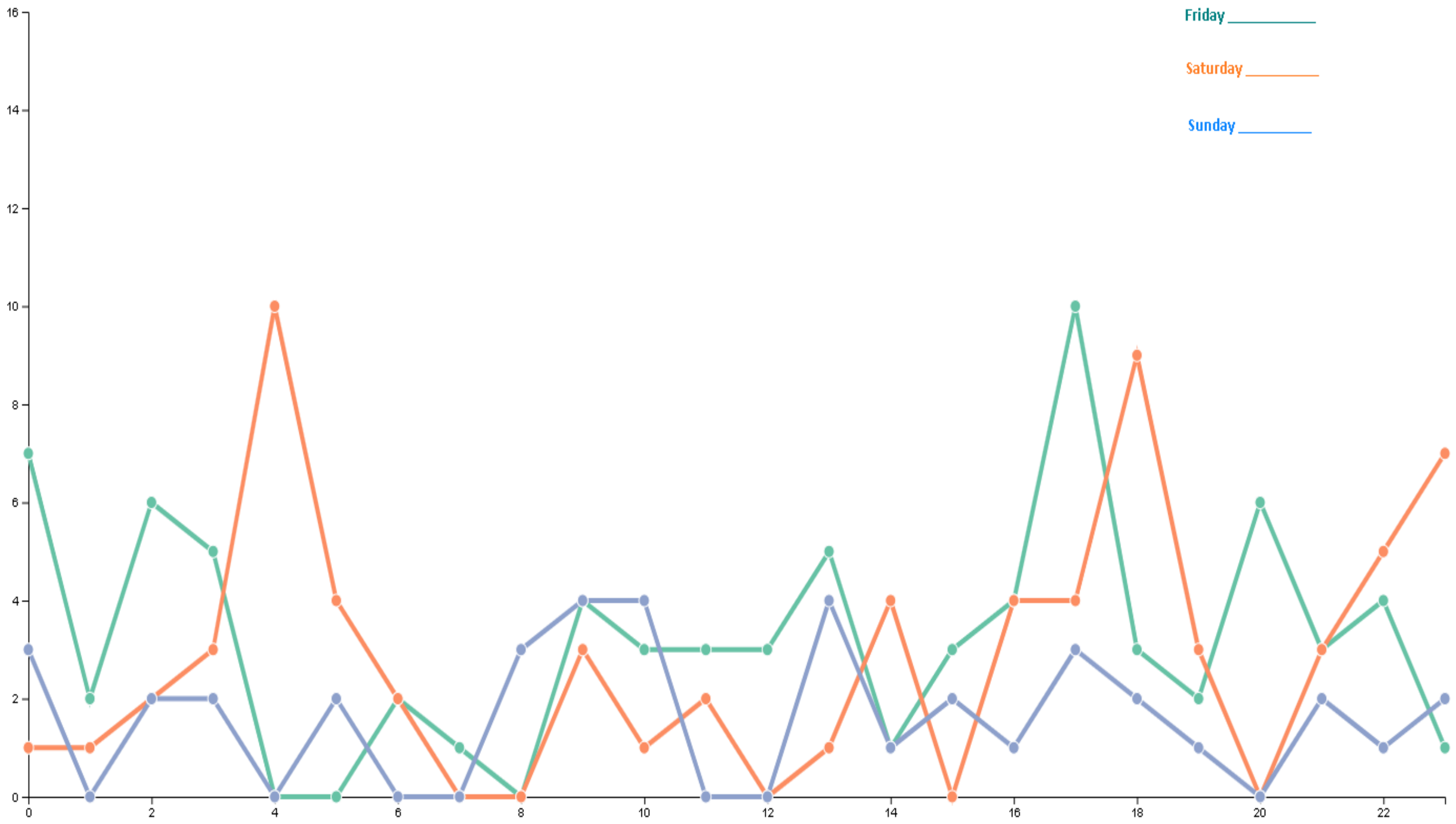
The scatter plots compare both these metrics on the same graph. The first plot shows every day of the week with all of the times.



This graph gives an overall view of the data extracted. It also displays an increase in commits at the end of the workday for the weekdays. As well as this, it shows an interesting outlier in the spike in commits at 4:00am on a Saturday. Unusually, this is actually when the most commits occur on Saturdays for this repository.

The following graphs separate this data into the start and the end of the week respectively.


Monday
Tuesday
Wednesday
Thursday
16
14
12
10
8
6
4
2
0
0
2
4
6
8
10
12
14
16
18
20
22


Friday
Saturday
Sunday
16
14
12
10
8
6
4
2
0
0
2
4
6
8
10
12
14
16
18
20
22

The original graph separated into two unique graphs gives a clearer view of the data and shows some more trends which would be more difficult to spot in the original plot. These graphs highlight many more commits occurring during the beginning of the week compared to the end of the week. Unsurprisingly, there are much less commits in mornings and during the nights of weekdays. Conversely, the end of the week has many more commits in the evening and during the night.

In conclusion, the visualisations of the interrogated data demonstrate some expected trends as well as some unique ones. The expected trends were common on other repositories tested as well. For example, the rise in the number of commits at the end of the workday was a common trait throughout most repositories. Overall, I found there are definitely certain days and times which tend to have more commits to repositories than others.