NYC BIKE SHARE

HARRISON LIN & CHRIS LYNCH

UNIVERSITY OF MICHIGAN SCHOOL OF INFORMATION

MOTIVATION

The New York CitiBike share scheme (NYC Bikes) is one of the world's largest, and data about bike usage is publicly available. Analysis of this data therefore represents a 'big data' challenge with 33 million rows of data.

We set out to explore the relationship with between Bike usage and the Weather in NYC over the first five months of 2018.

The questions we set out to answer were:

1. Which routes, as determined by pairs of bike hire stations are the busiest?
2. What is the relationship between weather conditions and total bike usage?
3. How does weather affect the duration of bike rides?

DATA SOURCES

01 NYC "CITI BIKE"

https://console.cloud.google.com/marketplace/product/city-of-new-york/nyc-citi-bike?project=translation-api-303322

Description The bike share scheme has 10,000 bikes and 600 bike stations. The data set started in 2013 and continues to receive daily updates. It is a public data set hosted in Google BigQuery.

Size The dataset had 33,319,019 records as at 6 April 2021. It has 16 columns related to bike hires such as start/stop time, location and user data.

Format GoogleBigQuery is a database that can be explored with SQL. The data is tabular with clear given data types (integer, string, datetime and so on).

Access method Query results can be exported as CSV/JSON and is also accessible via an API.

Description Weather data is available from 1901 to present day, provided by National Oceanic and Atmospheric Administration (NOAA) . The weather dataset can be accessed from the NOAA department of the National Center for Environmental Information archive. The key features of interest are the climate and temperature experienced during a given timeframe, in order to correlate it to the bike share data for user behavior analysis. (ex. Manhattan Station: GHCND:US1NYNY0074)

Size The dataset had 411,264 records from 2013-2020, when accessed. It contains 34 weather related features such as temperature, lat/long coordinates, precipitation.

Format The dataset is in CSV format consisting of integer, string, and datetime data types.

Access method We used an online query search is available for dataset requests, filtering features, location and date ranges. The results can be exported as CSV or Excel data and are also accessible via an API.

02 NOAA WEATHER FOR NYC

https://www.ncdc.noaa.gov/cdo-web/datasets

DATA MANIPULATION

Using Bigquery to extract bike ride information

The NYC Bike data is available as in a public Bigtable dataset on Google Cloud Platform (GCP). We explored this data set using SQL queries on the platform. We discovered that the most up to date data in the Bigtable was 2018, and only the first five months. So we extracted that data and created a dataset within the Bigtable area of GCP.

Accessing and cleaning the NOAA weather data

The weather data at different NYC stations can be requested from NOAA. The available datasets captures a range of time series weather data, from hourly to annual summaries. We requested for the shortest time interval hourly weather data to better capture the weather conditions experienced during a bike ride. We extracted data from 2018 to match the dates of the bike dataset.

Hosting data in a Postgres13 database on GCP

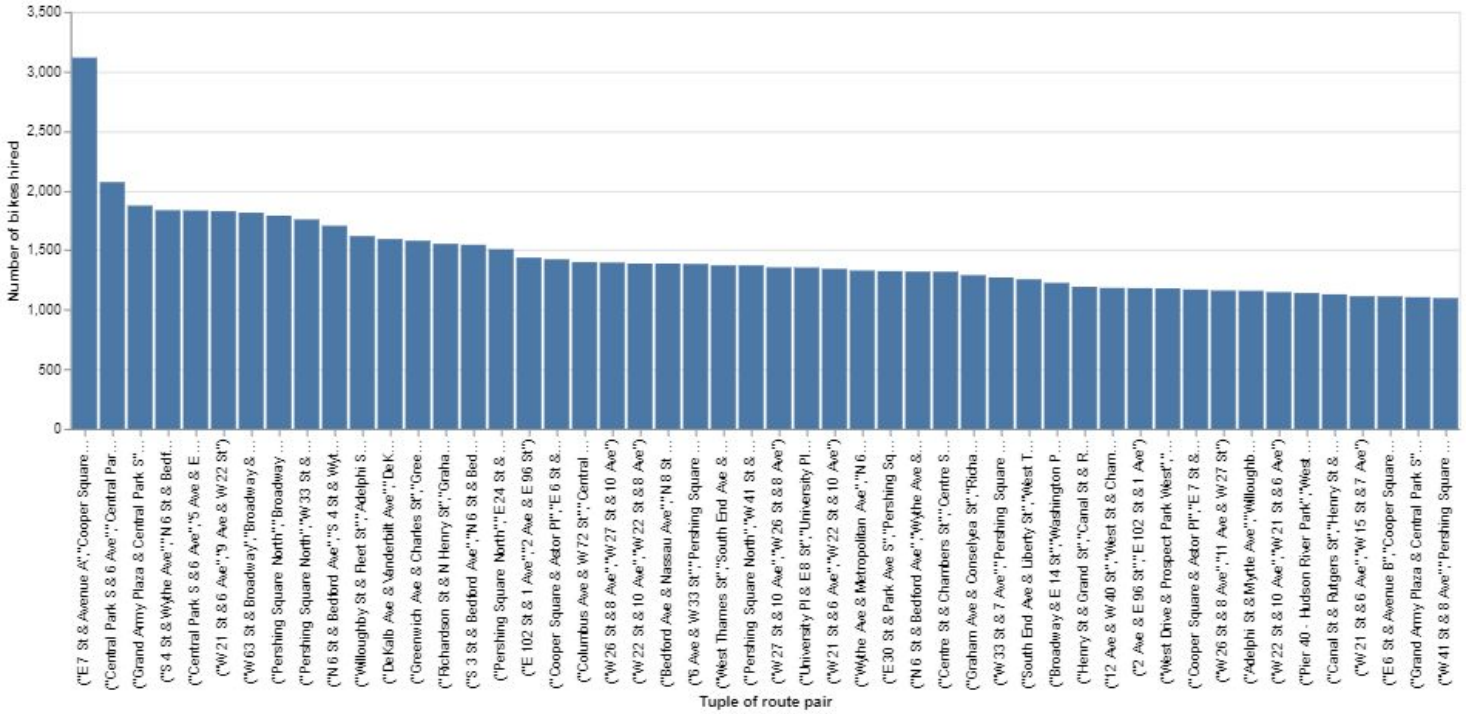
We created a specific project within GCP called "postgresql-nyc-bike-share-m1". This enabled us to create 'buckets' were we could import the BigTable datasets as CSVs. It also enabled us to instantiate a Postgres13 cloud hosted database into which we could import the CSV data from the 'buckets' in GCP storage. Finally we were able to join the datasets, and extract and manipulate data for analyses by running SQL queries against the postgresql database.

Manipulation of stored data for analyses

We were able to access our postgresql instance via psql, PGAdmin and python scripts. Most of the data manipulation and visualizations were performed via a script which connected through a shared Google Colab notebook. This enabled us to to 'pair programme' whilst working remotely in different time zones (Canada and UK). The final Colab notebook (minus the security login data) is available on our project repo at: https://github.com/imharrisonlin/NYC_BikeShare_Weather_Analysis .

Commuters spend less time on their bikes than tourists, and the tourists often don't aim to 'get' anywhere

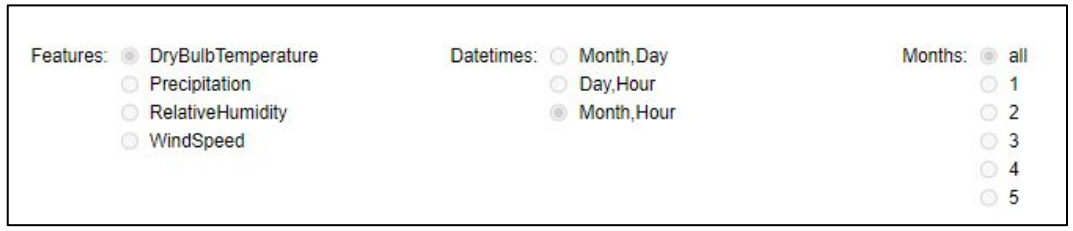
Below shows the top 50 busiest 'route pairs'. In the five months of data we investigated there are 245,068 such pairs. Analysis has shown that many of the most popular routes are of short average duration and represent commuter routes. Some of the busiest route pairs have both the same start and end point and are of longer duration, representing tourist rides - for example around Central park.

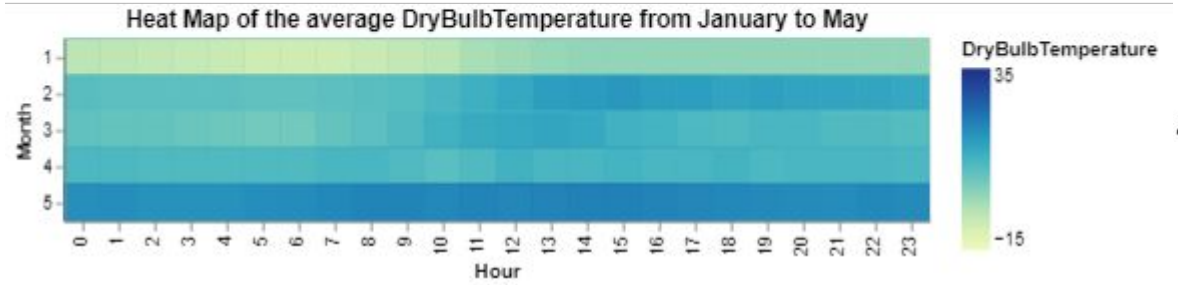


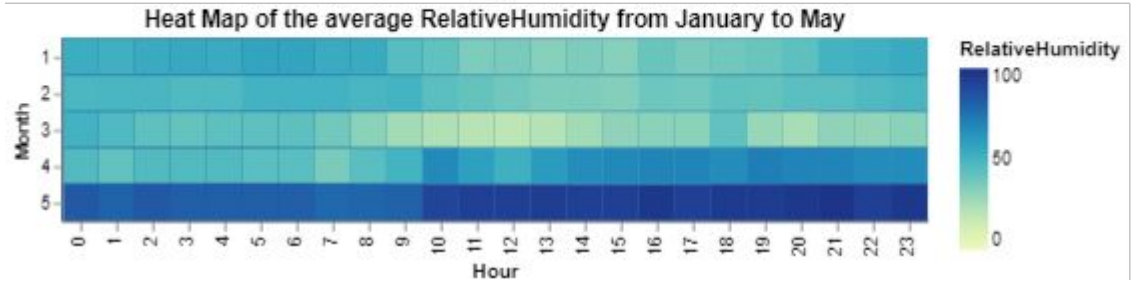
Above is an illustration of the start and end points for the 68th most popular route pair. Folium was used to create the image with the Green point being the start and red, the finish.

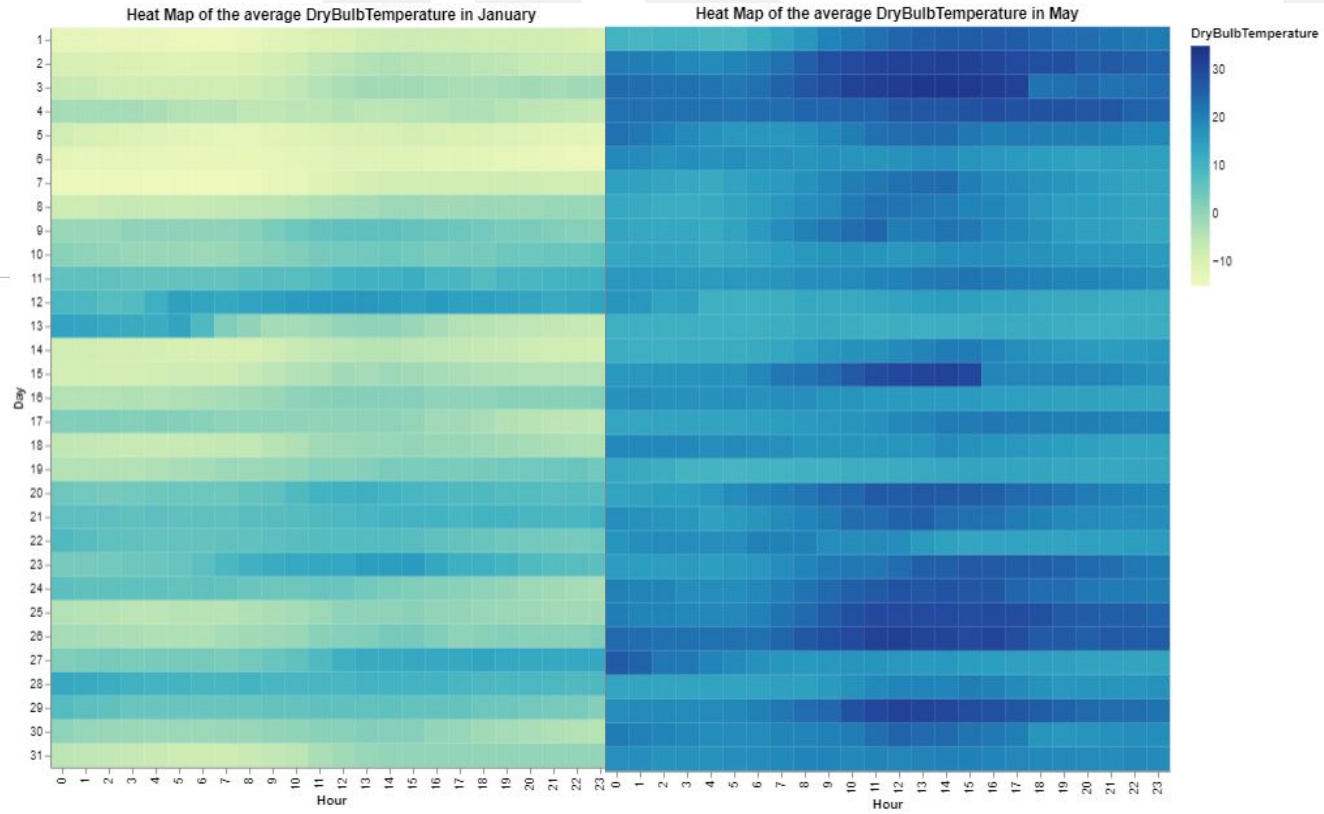
It was possible to illustrate all routes, and we have seen good examples of this such as: https://toddwschneider.com/posts/a-tale-of-twenty-two-million-citi-bikes-analyzing-the-nyc-bike-share-system/ however these 'overall' images are less interesting than specific routes which can be explored using the code on our project repo.

Summary of monthly, daily and hourly weather trends using a an interactive widget



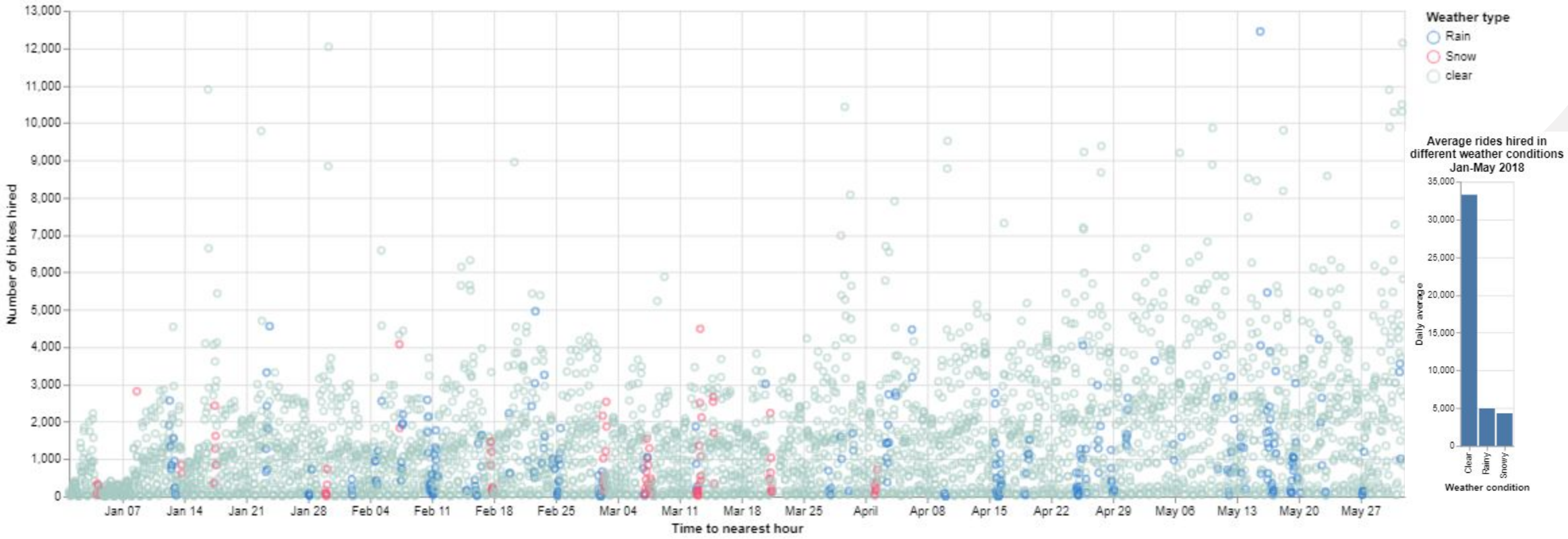






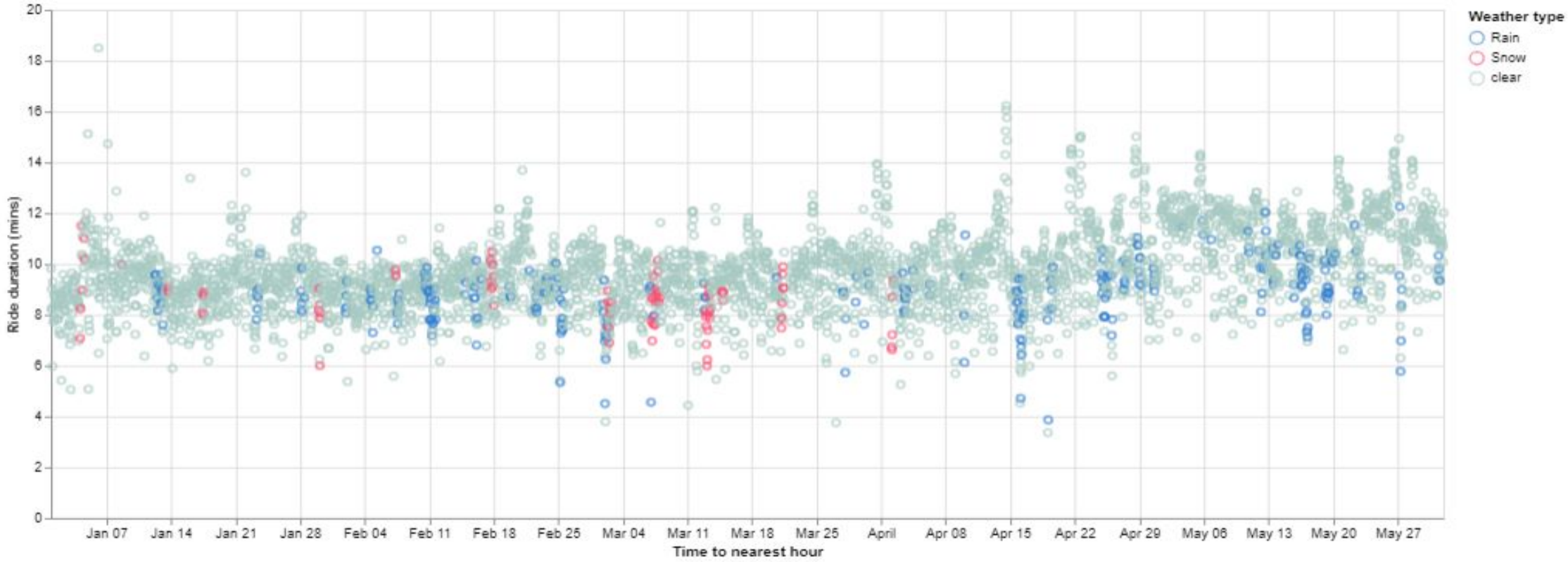
The interactive widget enabled us to efficiently visualize the individual weather conditions (i.e Temperature, precipitation, humidity and wind speeds) on different time scales (month, day, hour). On the top left chart, the month/hour temperature heat map shows the seasonality increase in average hourly temperature from below 0 in January to around 25 degrees in May. The individual months can be extrapolated to a day/hour temperature heat map to analyze daily fluctuations. Relative humidity showed a rapid change in May with average hourly humidity levels above 70% for the month.

More bikes are hired in better weather. Illustrated by hour in the main chart, and average numbers hired by day in smaller chart



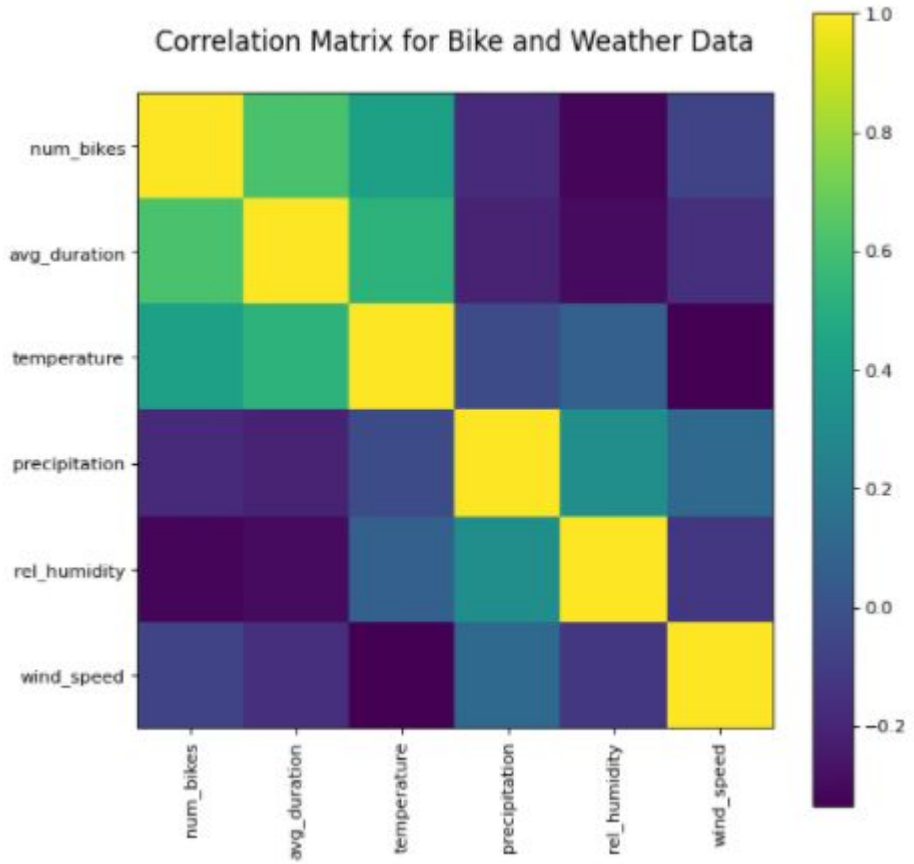
The main chart shows hourly bike data over Jan-May 2018. Each data point is the number of bikes hired in that hour, with a color to indicate if it was raining or snowing during the hour in question. The overall upwards trend reflects the seasonality of bike hires, but the main chart masks that fact that on days where it either snowed or rained - far fewer bikes got hired on average. This is illustrated on the chart to the right.

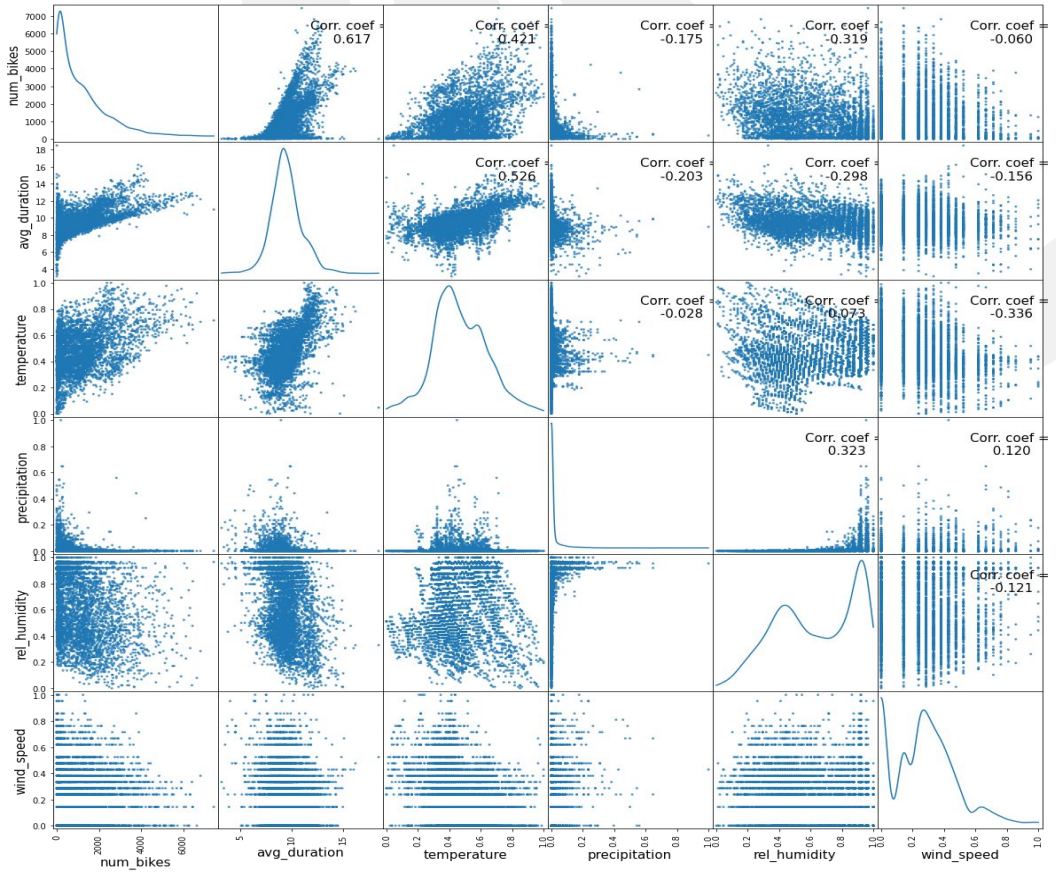
Average Bike rides durations are longer in clear weather compared rides in the rain and snow



The main chart shows hourly bike data over Jan-May 2018. Each data point is the number of bikes hired in that hour, with a color to indicate if it was raining or snowing during the hour in question. On days in which it snowed or rained, there were shorter ride durations as indicated in the chart. There is a higher ride duration trend observed in May, this is likely indicative of the seasonal trend in weather as the months of January to April were colder compared to the month of May as shown in the previous temperature heat map.

Bike hires and ride durations are correlated with changes in weather conditions





There is a positive correlation between bike hires and ride durations to temperature, with a correlation coefficient of 0.42 and 0.52 respectively. The positive correlation was reflective of the increase in ridership during the month of may when the temperature was higher. The precipitation and humidity is negatively correlated with a correlation of (-0.18, -0.20) and (-0.32, -0.30) respectively to bike hires and ride duration.

CONCLUSION & DEVELOPMENT

Conclusions

We were able to find a correlation between bike hires and positive weather, for both the number of hires and durations of rides. People clearly ride less in general and for shorter durations on days when there is either snow or rain.

People using citi bikes behave in two different ways, the first is a commuter profile of shorter rides getting from A to B and the second is a tourist profile riding around a given bike station and returning to it, taking longer on the bike than your typical commuter.

Developments

The time taken to download and manipulate the data was higher than expected, meaning we were not able to undertake the following analyses which would be interesting developments:

- training a classifier to predict bike usage and testing if the assumptions trained in NYC would work in another location.
- network analyses, looking for under/over utilised stations to suggest improvements in the layout of hire stations.

Image source: By Jim.henderson - Own work, CC0, https://commons.wikimedia.org/w/index.php?curid=26359195

STATEMENT OF WORK

CHRIS LYNCH

Chris focused initially on the bike data extract from BigQuery. Then after creating the postgresql instance, he looked at the sum of bike rides and route pairs. Chris initiated the project reporting.

HARRISON LIN

Harrison focussed initially on the weather data. He created the final join for the tables in postgresql and the interactive visualization widget. Then he led on the biking average duration analyses. Harrison set up the project github repo.

Both Chris and Harrison contributed at each stage of the project, discussing the approach to data extraction, exploration, manipulation and analysis. Using slack to communicate and Colab to share code, we were able to check one another's work and crack some of the bugs in our code. We both worked on the final report.