
# Introduction

A popular stereotype of British people is that we are a nation obsessed with the weather. I chose to use the London Transport data from the Summer Data challenge (http://summerdatachallenge.com) for this coursework, and I decided to look into relationships between public transport usage and weather conditions.

**Data domain**

The README file provided with the data states that the files include bus, tube and cycle hire data for 2012. The data files are structured differently for each transport type –

- For bus data hourly boarding figures are provided by bus route for each day in 2012
- For tubes the entry and exit figures are provided for each 15 minute interval by station for each day in 2012. For stations with more than one entry/exit gate the data is shown in total and by gate location.
- For cycle hire the total hires per day is provided from mid-2010 to mid-2014. Also provided is the number of hires per month, the average hire time per month and the annual totals.
- In addition I sourced daily weather data for mean daily temperature, rainfall and humidity during 2012-2014 from http://nw3weather.co.uk/ created and maintained by Ben Lee-Rodgers, used with his permission.

**Analytical questions**

I wanted to investigate the effect that weather patterns have on transport usage. Personally I use Waterloo as my main route into London, and when the weather is pleasant I prefer to walk from Waterloo if my onward destination is only a mile or two away. Walking at a reasonable pace I can cover two miles in half an hour and I have found this is often as quick as taking a tube or a bus. This surprised me at first but it is a habit which has many advantages. Similarly, the London Cycle Hire scheme offers a self-powered, open-air alternative to the bus or tube.

The hypotheses I wanted to investigate were -

- Tube and bus usage will decrease in pleasant weather
- Levels of cycle hire will increase in pleasant weather

**Objectives**

Understanding how demand is likely to be affected by the weather would be useful for the purposes of resource planning, for instance scheduling dips in service at times when demand is likely to be lower anyway.

Another objective is a social one. It is good to build exercise into daily life and if variations in propensity to walk or cycle can be understood in terms of weather conditions then campaigns and initiatives to encourage physical activity can be planned to coincide with periods when people are more likely to be receptive.

**Problems and considerations**

- Sourcing detailed weather data was a problem. After some hunting I found http://nw3weather.co.uk/ which provides meteorological observations from a site located near Hampstead which is close enough geographically to the sample data sites to be highly relevant. Rich weather data for temperature, humidity, rainfall and wind speed (and more) is available.

- There is a lot of variation in daily transport in London generally. Especially during 2012 which was the year of the Queens Jubilee and the Olympics.

- The bus and tube data covers a large area of the city, and there are many different geographies to consider. In central London where distances between destinations are shorter it is possible to consider walking or cycling as an alternative to public transport, but for longer journeys this is not realistic. For the purposes of my project I decided to work with a subset of the bus and tube data from an area where there were viable alternatives. I chose to use the tube data for Waterloo station and data for buses that go over Waterloo Bridge. Cycle hire data was given as totals only so I had no choice for that analysis. I disregarded cycle hire data before 2012 on the basis that the initial period of the scheme would have less typical patterns of usage.

- The bus and tube network serves different requirements at different times of the day, and on weekday and weekends. On weekdays the well-known morning rush is composed of workers coming into the city for 9 or 9.30. Tourists and visitors will come in later in the day. There will be an evening peak of people finishing work, and then evening use by people going out to restaurants and theatres. London is not a 24 hour city like New York, but there are night buses and late trains used by shift workers and revellers. As cycle hire data is only available daily it is not possible to look at these different usage patterns for that data.

**Strategy**

For bus and tube data I planned to split the data into different periods of the day to look more closely at different types of transport user. For cycle hire data I didn't have as much detail available so I thought I would examine working days vs weekends and bank holidays on the assumption that this would somewhat reflect commuters vs recreational cyclists.

My strategy was to visualise the data in the first instance, looking for general trends. Then to investigate correlations that would highlight areas worthy of more in-depth investigation.

As I had several measures of weather available I anticipated using a feature reduction method of analysis. I expected I would need to transform some weather features into bins rather than using a strict numerical scale, for instance to look at hot days vs cold days.

## Data Preparation

The .CSV files containing the bus and tube data were in different formats, and I needed to manipulate them to be in a common format for my analysis. First I got to know both sets of data by visual inspection, and then looked at the data distribution of frequency counts for the columns. The data were quite clean with no missing values. There were a small number of strange times in the bus data (28:00 for instance) and I chose to delete these records as I had plenty of good data. Once I had an idea of structure and oddities I could make a plan for reformatting the data to give me what I needed.

## Bus Data

```
Route_id,Date,Time_start,Time_end,Boardings
1,01/01/2012,05:00,06:00,18
1,01/01/2012,06:00,07:00,224
1,01/01/2012,07:00,08:00,184
1,01/01/2012,08:00,09:00,192
1,01/01/2012,09:00,10:00,231
1,01/01/2012,10:00,11:00,247
1,01/01/2012,11:00,12:00,322
1,01/01/2012,12:00,13:00,361
```

| | A | B | C | D | E |
|---|---|---|---|---|---|
| | Date | Total | Morning Rush | Morning | Afternc |
| 1 | 01/01/2012 | 81962 | 4542 | 17298 | 25 |
| 2 | 02/01/2012 | 110567 | 5300 | 24182 | 37 |
| 3 | 03/01/2012 | 178220 | 26088 | 36035 | 46 |
| 4 | 04/01/2012 | 202703 | 29672 | 42784 | 54 |
| 5 | 05/01/2012 | 208133 | 32481 | 42005 | 54 |
| 6 | 06/01/2012 | 215481 | 30614 | 43925 | 58 |
| 7 | 07/01/2012 | 150632 | 7827 | 35793 | 47 |

*Fig 1a. Bus Data manipulation*

For the bus data I needed to extract by route number to give me just the buses going over Waterloo Bridge, I also dropped the records with odd time stamps at this stage. Then I accumulated by route, date and time to give me 6 periods per day. I output the data as an intermediate .CSV at this stage to check totals, then accumulated this file to give me daily totals, stripping off the route information.

## Tube Data

```
Station,Gate_location,Date,Day_type,Flag,Entry_or_exit,1
5-0200,0200-0300,0300-0400,0400-0500,0500-0600,0600-070
Acton Town,,2012-01-01,3,,Entry,2637,3,2,1,3,0,3,0,1,1,1
Acton Town,,2012-01-01,3,,Exit,2813,5,1,2,3,11,8,9,3,8,
Aldgate,,2012-01-01,3,,Entry,1369,16,5,3,6,12,7,7,4,7,0,
Aldgate,,2012-01-01,3,,Exit,1285,2,2,2,2,1,0,3,3,1,4,0,
Aldgate East,TOTAL,2012-01-01,3,,Entry,6678,3,5,12,7,11,
Aldgate East,TOTAL,2012-01-01
Aldgate East,East Gates,2012-
Aldgate East,East Gates,2012-
Aldgate East,West Gates,2012-
Aldgate East,West Gates,2012-
Alperton,,2012-01-01,3,,Entry
Alperton,,2012-01-01,3,,Exit,
```

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Date | Total | Morning I | Morning | Afternoo | Evening F |
| 2 | 2012-01-01 | 1357378 | 3.92 | 20.21 | 34.40 | 15.94 |
| 3 | 2012-01-02 | 1850744 | 4.03 | 23.09 | 34.79 | 16.88 |
| 4 | 2012-01-03 | 2930241 | 18.54 | 17.47 | 21.58 | 23.97 |
| 5 | 2012-01-04 | 3219655 | 18.04 | 18.05 | 22.02 | 23.03 |
| 6 | 2012-01-05 | 3349238 | 18.04 | 18.17 | 21.57 | 22.43 |
| 7 | 2012-01-06 | 3417753 | 16.89 | 17.74 | 22.32 | 21.28 |
| 8 | 2012-01-07 | 2260135 | 4.91 | 22.44 | 30.71 | 16.58 |
| 9 | 2012-01-08 | 1475718 | 4.49 | 24.65 | 34.20 | 17.81 |
| 10 | 2012-01-09 | 3471344 | 19.24 | 17.60 | 21.12 | 22.82 |
| 11 | 2012-01-10 | 3607583 | 19.00 | 17.84 | 21.16 | 22.11 |

*Fig 1b. Tube data manipulation*

The tube data presented different challenges. Although I eventually decided to work with Waterloo, I began by looking at total stations covered by the data, which was up to Zone 3 as far as I could tell. Some stations (like Aldgate shown left) had what was essentially duplicate per-gate and Total data and I needed to strip out some records for these stations. Both Entry and Exit figures were given and I needed to consider this. I chose to use just Entry figures for my purposes, on the assumption that this would be better for estimating whether people chose a different method of journeying on from Waterloo.

## Cycle Hire Data/Weather Data

| Date | DayOfWeek | DayType | Cycle# | Temperature | FeelsLike | Rainfall | WindSpeed | Humidity | RainType |
|---|---|---|---|---|---|---|---|---|---|
| 2012-01-01 | Sun | Hols | 5,171 | 11.1 | 11.8 | 7.8 | 6.4 | 90 | Higher |
| 2012-01-02 | Mon | Hols | 9,091 | 5.8 | 3.3 | 0 | 6.4 | 81 | None |
| 2012-01-03 | Tue | Work | 10,094 | 8.4 | 6.4 | 10.6 | 11.4 | 82 | Higher |
| 2012-01-04 | Wed | Work | 13,936 | 6.9 | 3.7 | 3.4 | 9.3 | 75 | Lower |
| 2012-01-05 | Thu | Work | 14,191 | 8.9 | 6.8 | 4.6 | 11.1 | 70 | Higher |
| 2012-01-06 | Fri | Work | 17,713 | 6.6 | 4.6 | 0 | 5.4 | 76 | None |
| 2012-01-07 | Sat | Hols | 12,556 | 9.2 | 7.8 | 0 | 6.4 | 68 | None |
| 2012-01-08 | Sun | Hols | 10,487 | 8.7 | 7.6 | 0.1 | 4.5 | 70 | Lower |
| 2012-01-09 | Mon | Work | 19,472 | 10.1 | 9.8 | 0 | 4.9 | 83 | None |
| 2012-01-10 | Tue | Work | 20,548 | 10.1 | 9.7 | 0 | 5.2 | 85 | None |
| 2012-01-11 | Wed | Work | 20,958 | 9.1 | 7.9 | 0 | 6.4 | 83 | None |
| 2012-01-12 | Thu | Work | 20,879 | 8.4 | 7 | 0 | 6.4 | 79 | None |
| 2012-01-13 | Fri | Work | 19,691 | 2.8 | 1.6 | 0 | 2.9 | 84 | None |
| 2012-01-14 | Sat | Hols | 13,843 | 0.7 | 0.5 | 0 | 0.8 | 86 | None |
| 2012-01-15 | Sun | Hols | 12,149 | 2.4 | 1.1 | 0 | 3.1 | 79 | None |
| 2012-01-16 | Mon | Work | 19,136 | 2.1 | 0.8 | 0 | 2.7 | 76 | None |
| 2012-01-17 | Tue | Work | 20,170 | 2.1 | 1.3 | 0 | 1.9 | 82 | None |

*Fig 1c. Cycle Hire and Weather data manipulation*

The Cycle Hire required very little reformatting, as it was given as total hires per day.

The daily figures for temperature, humidity, rainfall and wind speed were easy to manually cut and paste from the NW3Weather website. I spot checked that year start and end figures matched the website after I had finished in case of human error. As I had cycle hire data for 2012-mid 2014 and matching weather data, I used a larger time period for the cycle data analysis than for buses and tubes.

Once I had the data files in a common per-day format I merged them into a spreadsheet and created some further variables using excel functions to give me data such as classification of days into weekdays and weekends. Rainfall data required some thought as about half the sample days had no rainfall, and the remaining days had a numerical value for amount of rainfall. I imputed a column with 3 bins giving me no rainfall, less than the mean rainfall, and greater than the mean rainfall. I added in school holidays by hand.

# Analysis

## Waterloo Bus and Tube usage

I used Tableau Public to make time series visualisations of bus and tube usage data for various periods of the day, also showing weather and school holidays. There was a lot of information, for the purposes of this report I have just presented total use, morning rush hour, temperature and school holidays. Looking at tube entries for Waterloo gives the chart below. There's a lot going on here. The important patterns are that absolute usage figures for morning rush hour decline when there are school holidays. The increased total usage during the Olympics (shown in green) is clearly visible.

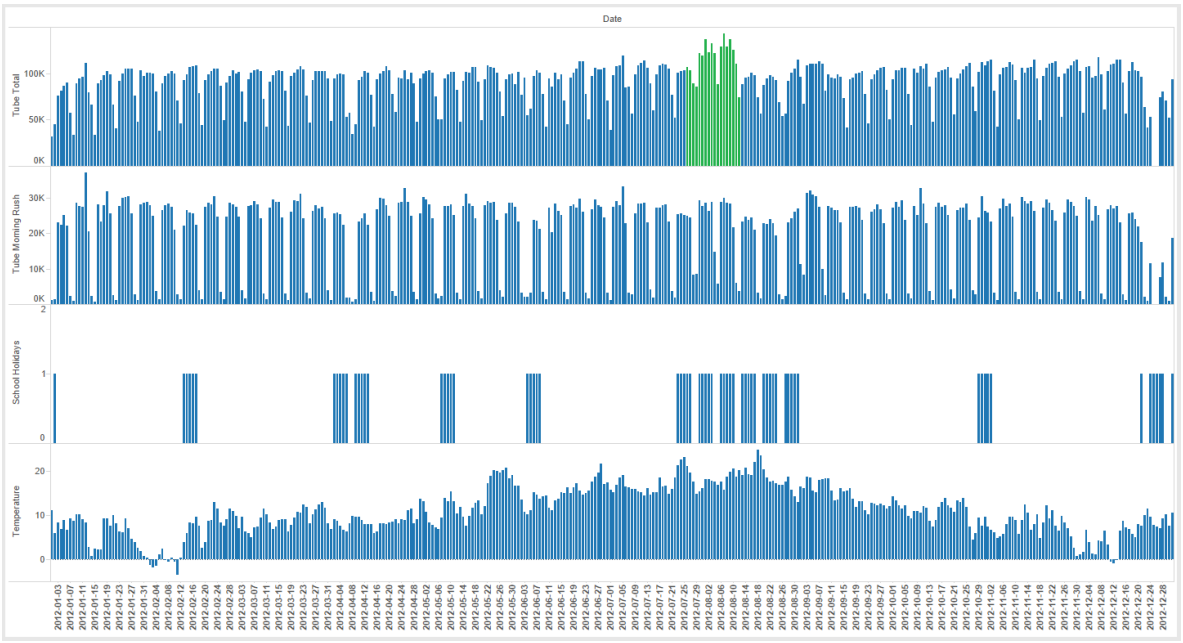



*Figure 2a – Daily Waterloo Tube usage during 2012, split by time periods and shown with average daily temperature and school holidays.*

The equivalent figures for buses going over Waterloo Bridge show a different pattern of usage, but still one which varies strongly with school holidays. There was no obvious Olympic effect on this set of buses.

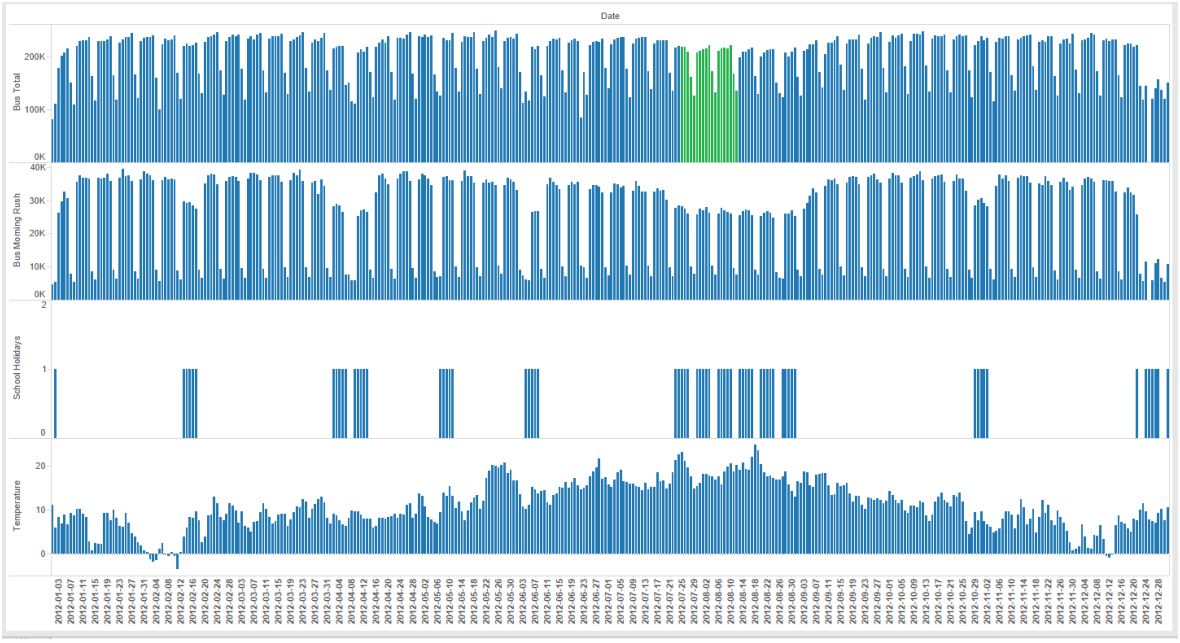



*Figure 2b – Daily Waterloo Bus usage during 2012, split by time periods and shown with average daily temperature and school holidays.*

There are definite patterns visible in this data and visualisations of the full range of daily time periods were interesting but not shown here due to space constraints. However, for the purposes of this analysis, I think there are too many variables involved to attribute any behaviour to weather related patterns. For instance, school holiday periods is clearly a confounder. For the remainder of this analysis I decided to concentrate on the cycle hire data.

**Cycle Hire Data**

The cycle hire data did not show the same decline during school holidays as the bus and tube data shown so far, so I progressed to looking at cycle hire figures in relation to weather patterns. I began by considering the weather data I had available. I had mean daily figures for - temperature, "feelslike" temperature (taking wind chill into account), wind speed and humidity. I wanted to know if any of these conditions were strongly correlated with any others, and I used Python to create a cross-correlation of the data.

| | School Holidays | Cycle# | Temperature | FeelsLike | Rainfall | WindSpeed | Humidity |
|---|---|---|---|---|---|---|---|
| School Holidays | 1.000000 | 0.121859 | 0.201199 | 0.197277 | -0.117856 | -0.012793 | -0.152039 |
| Cycle# | 0.121859 | 1.000000 | 0.681747 | 0.661759 | -0.401330 | -0.248702 | -0.528375 |
| Temperature | 0.201199 | 0.681747 | 1.000000 | 0.990571 | -0.012987 | -0.158934 | -0.319935 |
| FeelsLike | 0.197277 | 0.661759 | 0.990571 | 1.000000 | 0.008649 | -0.226908 | -0.238538 |
| Rainfall | -0.117856 | -0.401330 | -0.012987 | 0.008649 | 1.000000 | 0.022085 | 0.337746 |
| WindSpeed | -0.012793 | -0.248702 | -0.158934 | -0.226908 | 0.022085 | 1.000000 | -0.194971 |
| Humidity | -0.152039 | -0.528375 | -0.319935 | -0.238538 | 0.337746 | -0.194971 | 1.000000 |

*Table 1. Spearman correlations for weather data for 2012 to mid-2014*

Temperature and "feelslike" temperature are (obviously) strongly related (r=0.99), and I chose to use just temperature, reasoning that it was a directly measured value. Interestingly there was a low correlation between temperature and rainfall (r=-0.01).

Rainfall is an unusual data type. For about half the days in the sample there was no rain at all. For days when it rained there was a spectrum from heavy to light rainfall. As I was working with mean daily rainfall, days where there was light rain all day were indistinguishable from days when there was a short heavy downpour. Each situation might give the same mean rainfall but have a different effect on cycling behaviour. I created a data column with three bins for days when there was no rainfall, days when there was less than the mean rainfall and days with greater than the mean rainfall. I then looked at the mean number of bike hires among these groups when cross tabulated with holidays (weekends and bank holidays) and workdays.

| | Cycle# | | | |
|---|---|---|---|---|
| RainType | Higher | Lower | None | All |
| DayType | | | | |
| Hols | 15387.285714 | 18232.226190 | 25503.005882 | 21705.503226 |
| Work | 19821.605769 | 25426.583691 | 29890.179272 | 26882.759366 |
| All | 18269.593750 | 23520.192429 | 28474.962049 | 25284.204183 |

*Table 2. Mean Cycle Hires amongst Holidays and Workday by rainfall amounts*

Expressing the data in Figure 2 in terms of % distance from the mean usage on all days gives –

| | Higher Rainfall | Lower Rainfall | None |
|---|---|---|---|
| Hols | -29% | -16% | 17% |
| Work | -26% | -5% | 11% |
| All | -28% | -7% | 13% |

*Table 3. Expressed as % from the mean*

This suggests that heavy rainfall makes a big difference to usage, whereas light rainfall doesn't seem to be such a deterrent, especially on workdays. It would be interesting to look at this analysis broken down to a more granular level, for instance by hour as for tube and bus data.

Given the low correlation between temperature and rainfall, the mixed Boolean/integer range nature of the rainfall data, and the fact that I only had daily rainfall totals, I decided this was the limit of the useful analysis that could be done in this area. I narrowed my sample to dry workdays only, so I could discount the factors of rainfall and recreational use (as far as possible given the data).

A graph of wind speed vs cycle hire showed few meaningful patterns. There seemed to be a decrease in usage at very high wind speeds, but there were not enough extremely windy days in the sample (and therefore in reality) to make this a useful line of enquiry. Temperature and humidity proved more interesting.

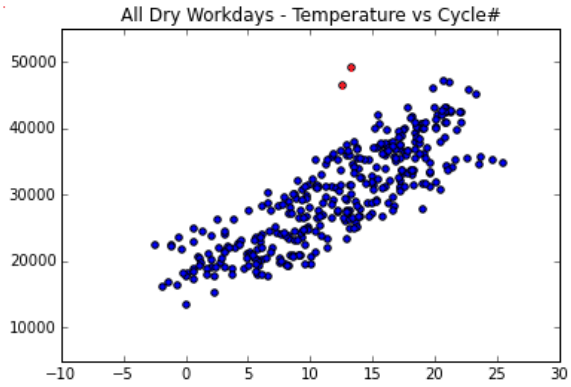



*Fig 3a. Temperature vs Cycle Hire, Dry Workdays*

There is a strong (r=0.68) correlation between temperature and cycle usage. The two outliers in red at the top of the chart are for 2 days when there was a tube strike. I decided to leave these in the chart because they were a measure of a real effect.

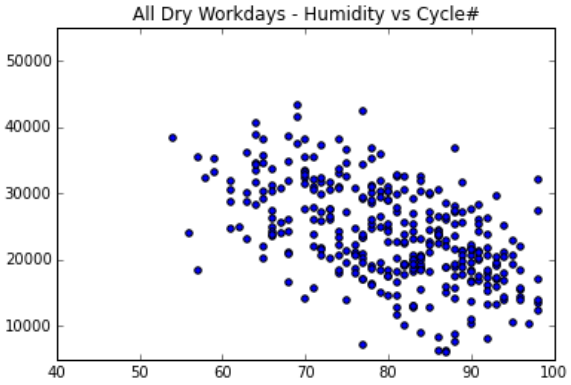


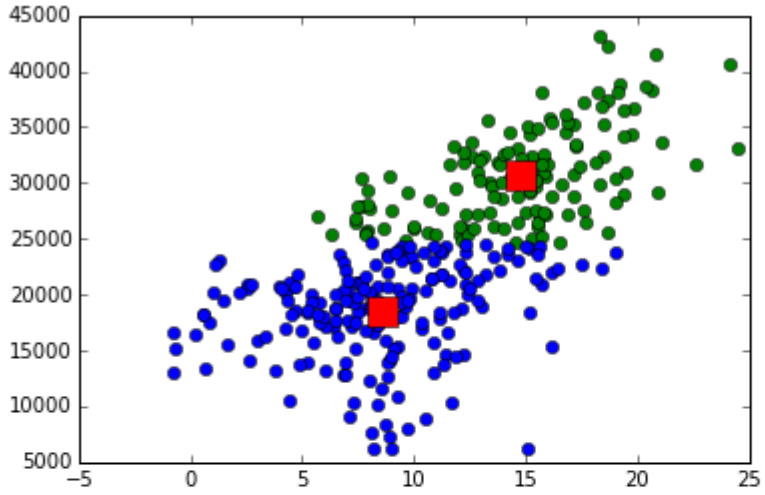

*Fig 3b. Humidity vs Cycle Hire, Dry Workdays*

There was a weaker (r= -0.53) but not insignificant correlation between humidity and cycle usage.

The correlation for humidity with temperature for the dry workdays sample set is (r=-0.32)

I performed a set of K-means analyses in Python with combinations of temperature, humidity and hires. These analyses gave a good fit of two clusters distinguished by temperature. Removing the humidity data from the analysis had very little effect on the clusters so I removed it from the analysis.

The result of the cluster analysis is shown in Fig 4 below. Analysing the original data by the cluster membership shows Cluster 0 (which I'll call 'warmer days') accounts for 4,358,198 total hires on 143 dry workdays where the mean temperature was 14.6 degrees. Cluster 1 (which I'll call 'cooler days') accounts for 3,627,643 total hires on 194 dry workdays where the mean temperature was 8.5 degrees.




CLUSTER 0
(142, 3)

| | Temperature | Cycle# |
|---|---|---|
| count | 142.000000 | 142.000000 |
| mean | 14.697887 | 30691.535211 |
| std | 3.737304 | 4111.601285 |
| min | 5.700000 | 24723.000000 |
| 25% | 12.600000 | 27297.000000 |
| 50% | 15.200000 | 30352.000000 |
| 75% | 17.100000 | 32944.000000 |
| max | 24.500000 | 43183.000000 |

Cycle# 4358198.0

CLUSTER 1
(195, 3)

| | Temperature | Cycle# |
|---|---|---|
| count | 195.000000 | 195.000000 |
| mean | 8.456923 | 18603.297436 |
| std | 3.891981 | 4195.269375 |
| min | -0.800000 | 6175.000000 |
| 25% | 6.450000 | 16595.000000 |
| 50% | 8.400000 | 19272.000000 |
| 75% | 10.900000 | 21512.500000 |
| max | 19.000000 | 24701.000000 |

Cycle# 3627643.0

*Fig 4.Cluster Analysis of Temperature and Cycle Hire for Dry Workdays, 2012 to mid-2014*

## Conclusion

Given the bus and tube data provided, it is not possible to investigate relationships with weather patterns.

For the cycle hire data however, there does seem to be a relationship with both rainfall and temperature. On workdays, and therefore I assume amongst commuters, cycling increases with warmer weather. Commuters are not discouraged by lighter rain, but heavy rainfall is a disincentive to cycle. Recreational users are less likely to use bikes on rainy days.

These findings could be used to improve the planning of cycle availability. If there is a model of how usage will vary with the weather, bike shortages can be minimised. It would also be useful to understand when requirements will be lower so that bike maintenance can be planned with less disruption.

## Tools and Software Choices

I used a variety of tools during this project, to reflect the different types of data and output I wanted. Mainly I used Python to prepare the CSV files provided. The bus and tube data were very large files which could not be read by Excel in the first instance. Python was a good tool because it gave me the option of reading the files either as a pandas dataframe or as just a raw text file which I could split and accumulate as I wanted using iterative loops.

I moved to using Excel once I had the data reformatted into a smaller, more manageable day-per-record format. In Excel it was easier to merge data from different sources, and Excel was a better visual tool for inspecting the data. I cut and pasted the weather data into Excel and reformatted to match the tube, bus and cycle data. Adding imputed data such as “Day of Week”, “School holidays” and creating bins for “Rainfall level” was quicker and easier in Excel than in Python.

Tableau Public offers some very nice methods of visualising data, especially for time series plots, and it was simple to look at bus and tube data against holidays and temperature with this software. The output was very visually attractive for little or no programming effort.

I used Python for statistical analysis such as correlations and clustering. I took screenshots from the Python console and reformatted them as necessary using windows paint to give me tables, graphs and figures for my report.

## Declaration

By submitting this work, I declare that this work is entirely my own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the coursework instructions and any other relevant programme and module documentation. In submitting this work I acknowledge that I have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. I also acknowledge that this work will be subject to a variety of checks for academic misconduct. Marks are provisional and subject to change in response to moderation, assessment board decisions and any ongoing investigations of suspected academic misconduct.