Section 1: Statistical test

> ***1.1*** *Which statistical test did you use to analyse the NYC subway data? Did you use a one-tail or a two-tail P value? What is the null hypothesis? What is your p-critical value?*

The non-parametric Mann-Whitney U test in this section was implemented in order to investigate whether hourly entries differed as to whether it rained. The null hypothesis examined was therefore $\mu_1 = \mu_2$. As per common convention, it was defined $p - critical = 0.05$.

> ***1.2*** *Why is this statistical test applicable to the dataset? In particular, consider the assumptions that the test is making about the distribution of ridership in the two samples.*

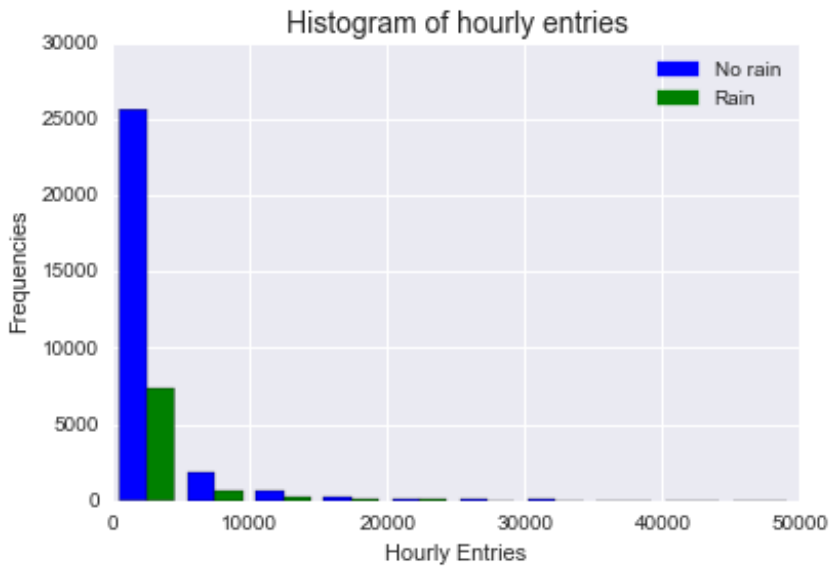

Figure 1: Histograms of hourly entries

From the histogram it can be inferred that the distributions of both subsets of the data are non-normally distributed; hence a non-parametric test for comparison of means is more appropriate.

> ***1.3*** *What results did you get from this statistical test? These should include the following numerical values: p-values, as well as the means for each of the two samples under test.*

The sample means for ENTRIESn_hourly for the two cases (rain and no rain) are $\bar{x}_1 \approx 1105$ ($n_1 = 44104$) and $\bar{x}_2 \approx 1090$ ($n_2 = 87847$).

The test (two-tailed as equality is being examined) returns a p-value of approximately $0.04999 < p - critical = 0.05$

> ***1.4*** *What is the significance and interpretation of these results?*

The calculated p-value is less than the chosen p-critical and the null hypothesis is rejected. Therefore the Mann-Whitney indicates that the population means are unequal.

This result implies that more people entered subway stations when it rained; however this finding does not yet constitute evidence that there is a causal relationship between ridership and rain.

Section 2: Linear Regression

> ***2.1*** *What approach did you use to compute the coefficients theta and produce prediction for ENTRIESn_hourly in your regression model:*
>
> 1. *Gradient descent (as implemented in exercise 3.5)*
> 2. *OLS using Statsmodels*
> 3. *Or something different?*

I used a multiple regression approach implemented by OLS{statsmodels} in Python.

> ***2.2*** *What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?*

Firstly, I saw that some stations had more than one turnstiles measuring entries. I therefore decided to accumulate data "per station" for analysis.

I initially devised a regression model incorporating non-meteorological features in order to provide a base for the analysis, namely

`'ENTRIESn_hourly ~ C(hour) + station + C(hour):station + weekend'`, where

ENTRIESn_hourly: hourly entries,
C(hour): hour of day (as a nominal variable) of entry
station: station of entry
weekend: binary value indicating whether entries are on a week day or the weekend
C(hour):station: interaction term modelling the association between hour of day and station

I subsequently tested meteorological variables such as rain, precipitation; I also calculated the Wind Chill Index (a unit measuring "*the perceived decrease in air temperature felt by the body on exposed skin due to the flow of air*") based on the following formula:

$$T_{\mathrm{wc}} = 35.74 + 0.6215T_{\mathrm{a}} - 35.75V^{+0.16} + 0.4275T_{\mathrm{a}}V^{+0.16}$$

The dummy variables involved are hour, station, weekend and rain.

> ***2.3*** *Why did you select these features in your model? We are looking for specific reasons that lead you to believe that the selected features will contribute to the predictive power of your model.*

- *Your reasons might be based on intuition. For example, response for fog might be: "I decided to use fog because I thought that when it is very foggy outside people might decide to use the subway more often."*
- *Your reasons might also be based on data exploration and experimentation, for example: "I used feature X because as soon as I included it in my model, it drastically improved my R^2 value."*

It is fairly intuitive that factors primarily affecting subway ridership are non-meteorological; it is clear that the station, the time of day and whether it is a weekday or weekend are more significant to a person's choice for using the subway.

The interaction of station and hour is also of interest. An explanation of the interaction term is the following:

In a purely additive or main effect model, the predictive effect of an independent variable is constant and invariant. We can change the values of all of other independent variables in all ways and combinations, but the predictive effects will remain unchanged. In an interactive model, the predictive effect of a variable in an interaction changes and varies with the values of the other independent variable(s) in the interaction.

> ***2.4*** *What are the coefficients (or weights) of the non-dummy features in your linear regression model?*

I eventually only used qualitative variables for the linear model.

> ***2.5*** *What is your model's R^2 (coefficients of determination) value?*

My final model is the following:

```
'ENTRIESn_hourly ~ C(hour) + station + C(hour):station + weekend'
```

It does not contain meteorological variables and returns an R^2 of 0.822 and an adjusted R^2 of 0.816.

The incorporation of wind chill index, rain and precipitation does not improve the adjusted R^2.

> ***2.6*** *What does this R^2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R^2 value?*

The calculated R^2 is quite high and it seems that ridership can be quite accurately predicted using the specified model. Residual analysis can be seen below:

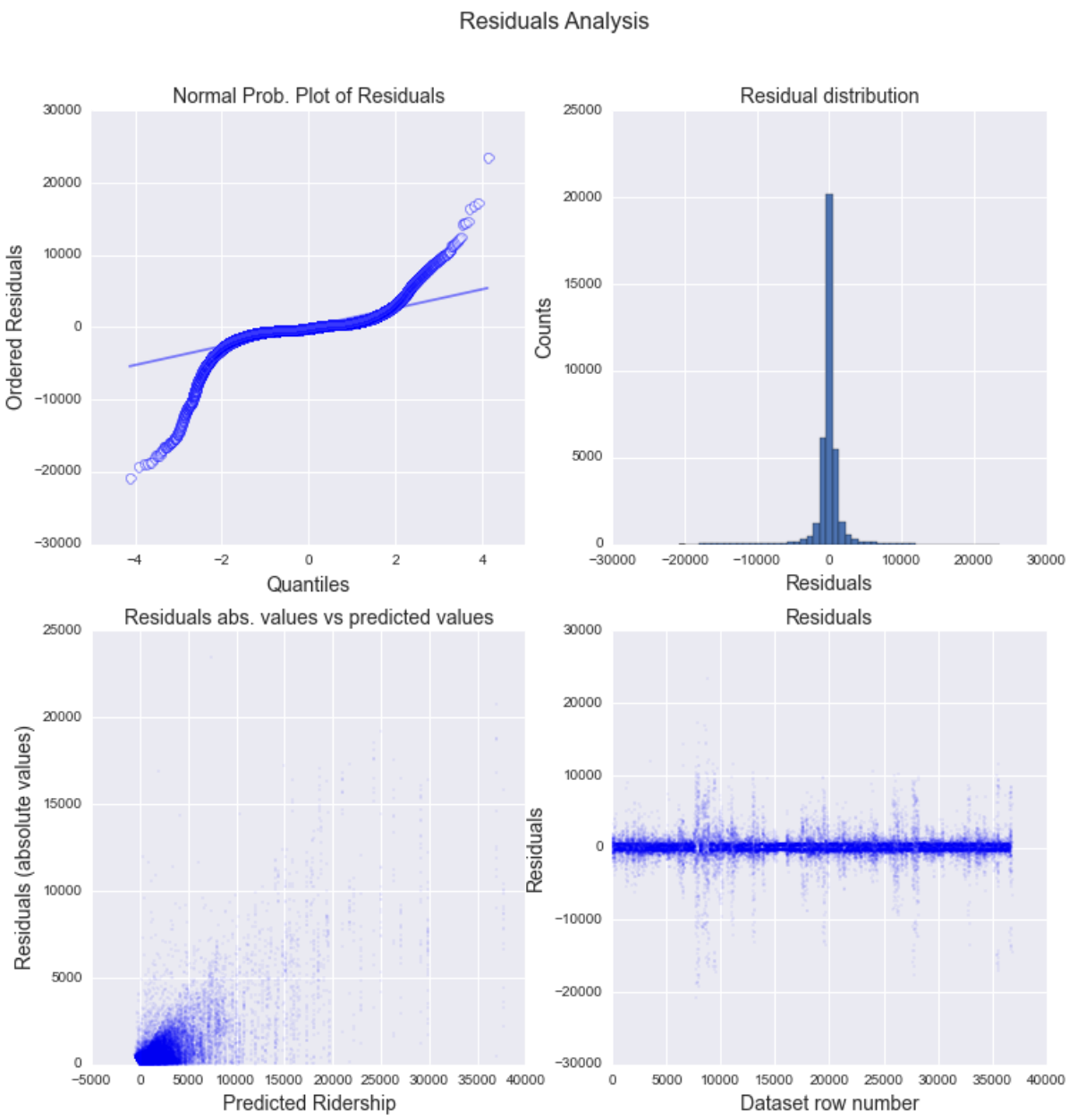

Figure 2: Residuals Analysis

All subplots show that residuals are random and symmetrically distributed around 0; a good sign for the model fit. However from the normal probability plot and the histogram it can be inferred that the distribution of the errors is not Gaussian – it is longer tailed. This implies that perhaps a more complex (polynomial?) model may provide better predictive strength.

The high value for R^2 raises suspicions for overfitting; however the fairly intuitive and logical derivation of the linear model implies that it probably constitutes a model effective in prediction in a global sense.

Section 3: Visualisation

Please include two visualizations that show the relationships between two or more variables in the NYC subway data. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots, or histograms) or attempt to implement something more advanced if you'd like.

Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

> ***3.1*** *One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.*
>
> *You can combine the two histograms in a single plot or you can use two separate plots.*
>
> *If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.*
>
> *For the histograms, you should have intervals representing the volume of ridership (value of ENTRIESn_hourly) on the x-axis and the frequency of occurrence on the y-axis. For example, each interval (along the x-axis), the height of the bar for this interval will represent the number of records (rows in our data) that have ENTRIESn_hourly that falls in this interval.*
>
> *Remember to increase the number of bins in the histogram (by having larger number of bars). The default bin width is not sufficient to capture the variability in the two samples.*

In the following two figures we can see

- The normalised histograms of hourly entries for rainy and non-rainy weather conditions.
- The Kernel Density Estimates of the distributions of hourly entries for rainy and non-rainy weather conditions.

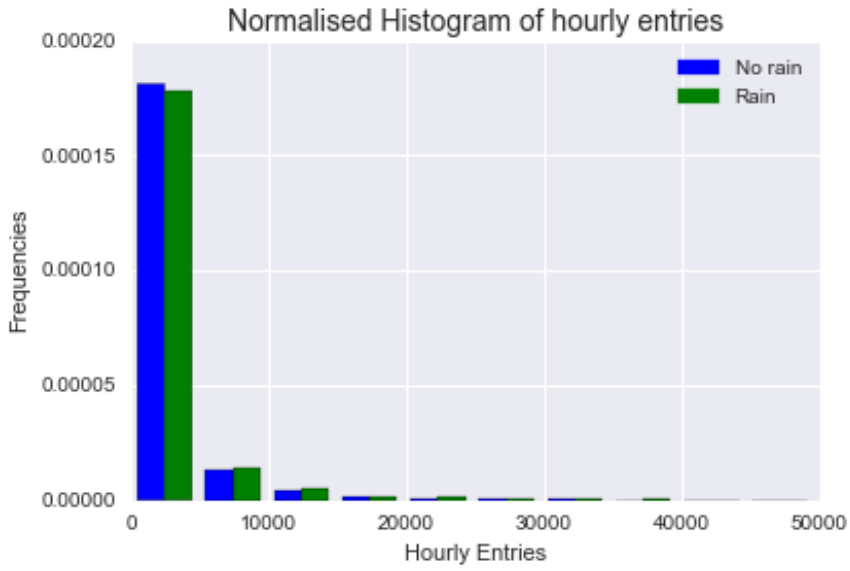

Figure 3: Normalised Histogram of hourly entries

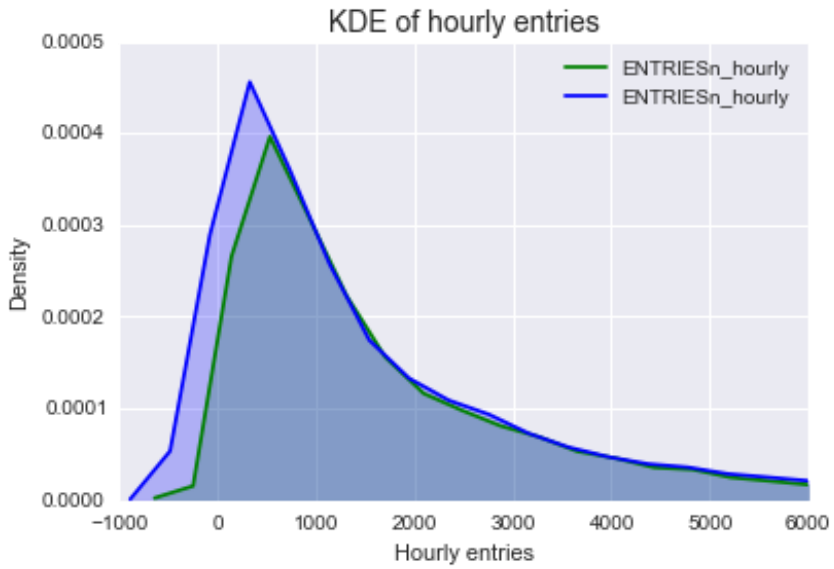

Figure 4: Kernel Density Estimates for hourly entries

> ***3.2*** *One visualisation can be more freeform. Some suggestions are:*
>
> *Ridership by time-of-day*
>
> *Ridership by day-of-week*

The following visualisation uses as background a rudimentary map of NYC (using Python module “Basemap”). Overlaid are the location of stations; sizes of points correspond to the total ridership per station and the colour corresponds to the average “rain” variable for that specific station; i.e. the proportion of rainy instances in the data set for that station.

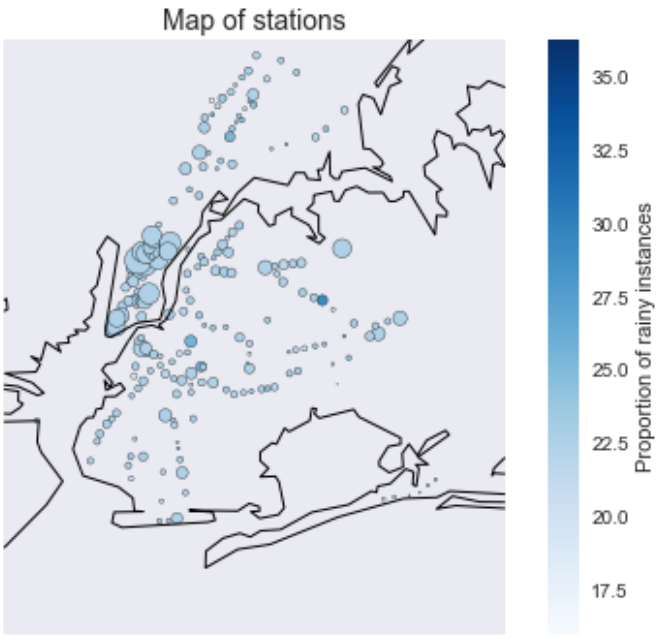

Figure 5: Map of stations; size corresponds to total entries per station and colour to average rainy instances

Section 4: Conclusion

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.

> ***4.1*** *From your analysis and interpretation of the data, do more people ride the NYC subway when it is raining or when it is not raining?*

From Figure 1, it can be inferred that more people rode the NYC subway when it did not rain during the month of May of 2011. Below is a figure showing the daily rain incidence (as recorded):

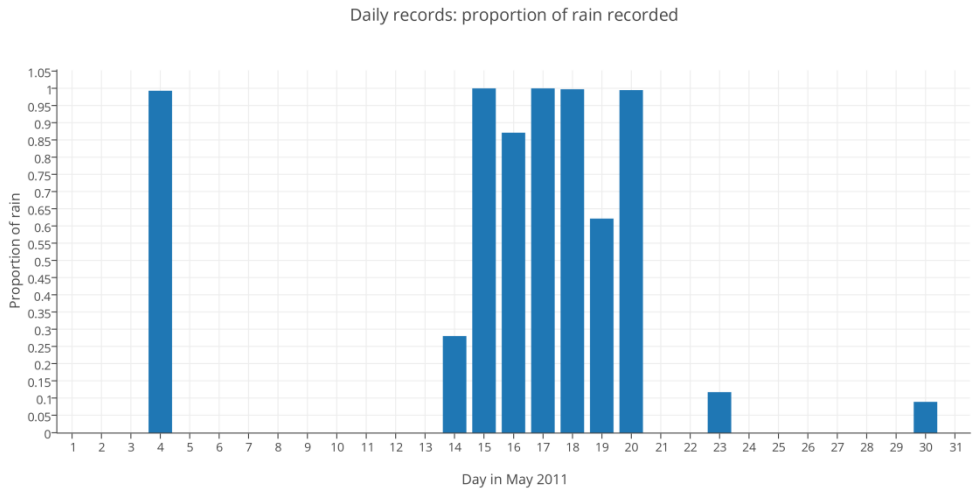

Figure 6: Proportion of daily rain incidence

From Figure 6 we can see that it only rained for 10 days in May, therefore it is to be expected that more people rode the subway when it did not rain.

However, the comparison of ridership means of Section 1 (when it rained vs not) showed in a statistically significant manner that the rate of entry of people was higher when it rained.

The above observations do not reveal much about the causal relationship between rain and ridership though. More appropriate visualisations for this are Figures 3 and 4 (normalised histograms and KDEs); these imply that the probability density functions for ridership when it rains and when not are very similar. Therefore it is suggested that rain may not have a significant impact upon ridership.

> ***4.2*** *What analyses lead you to this conclusion? You should use results from both your statistical tests and your linear regression to support your analysis.*

The suspicion that meteorological variables do not have a significant effect upon ridership is confirmed by the results of statistical modelling.

The aforementioned model ,

```
'ENTRIESn_hourly ~ C(hour) + station + C(hour):station + weekend',
```

does not incorporate any meteorological variables, yet accounts for approximately 82% of the variability. Incrementally adding rain, fog, Wind Chill Index and temperature variables did not improve the predictive capacity of the model as demonstrated by the adjusted R^2 values returned.

Section 5: Reflection

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.

> ***5.1*** *Please discuss potential shortcomings of the methods of your analysis, including:*
>
> *Dataset,*
>
> *Analysis, such as the linear regression model or statistical test.*

The dataset was provided in .csv format and was structured, i.e. columns represented variables. The checks for data set integrity showed absence of missing values and significant outliers. However, it was noted that in some cases data was not recorded continuously, as there was no recordings for some stations for some time intervals. This is expected to incur bias in the statistical analyses.

It was also observed that meteorological variables are not very reliable, as in some cases recorded weather data takes place from weather stations situated quite remotely from the subway stations (up to 5 miles distance –see Figures 7 and 8).

Figure 7: Map of subway and weather stations

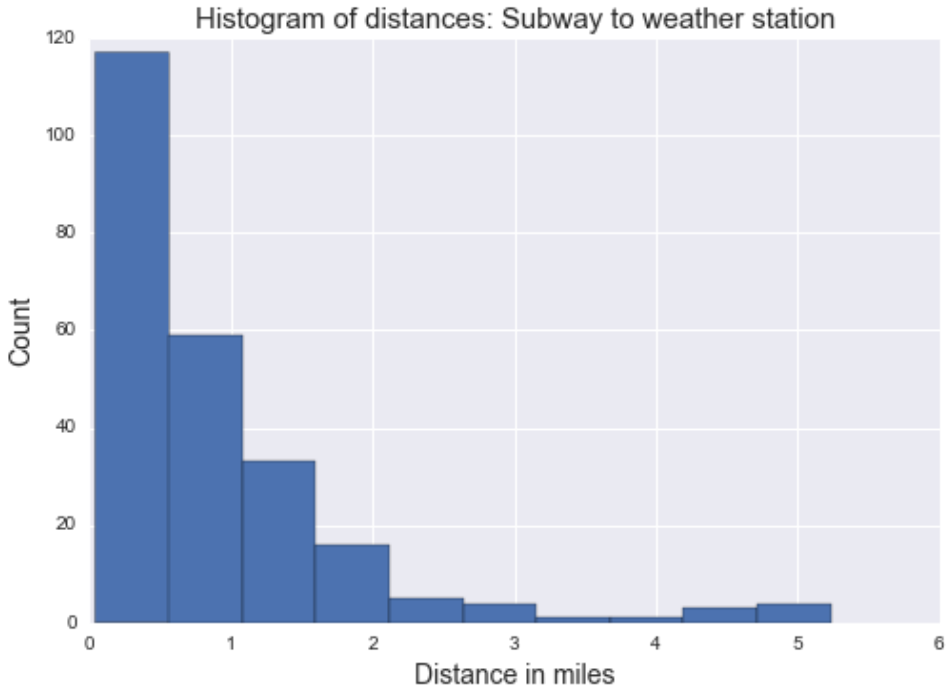

Figure 8: Histogram of distances between stations

> ***5.2*** *(Optional) Do you have any other insight about the dataset that you would like to share with us?*

I examined graphically the hypothesis that the completeness of recording is associated with distance from the central stations. By specifying the 42 Street subway station as the central point, I calculated

the distance of all stations from this point and plotted against number of recordings. Interpretation shows that apart from few outliers, there is no strong evidence supporting this hypothesis.

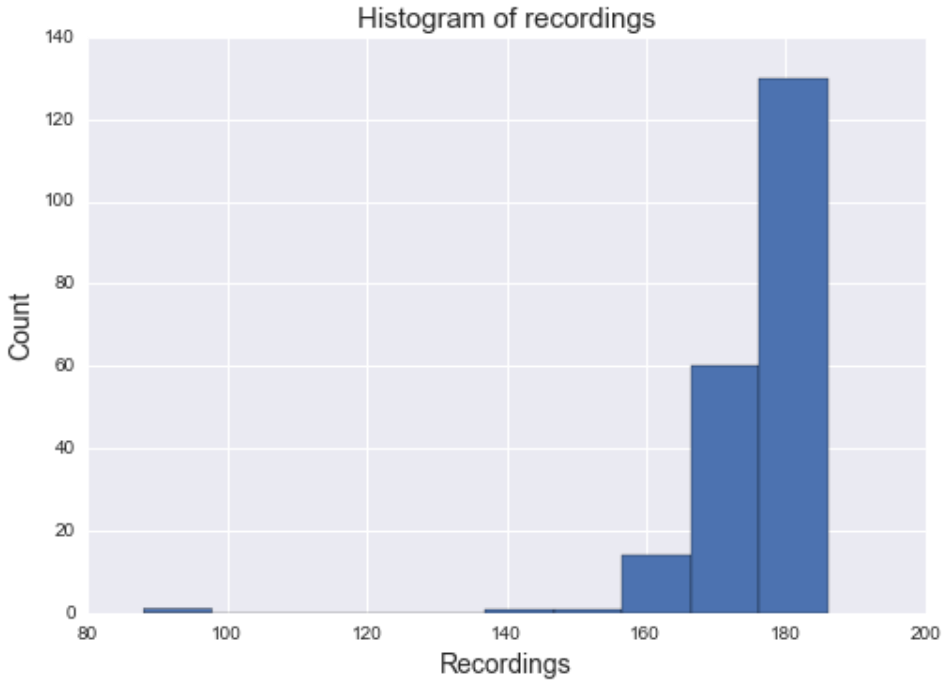

Figure 9: Histogram showing recording completeness

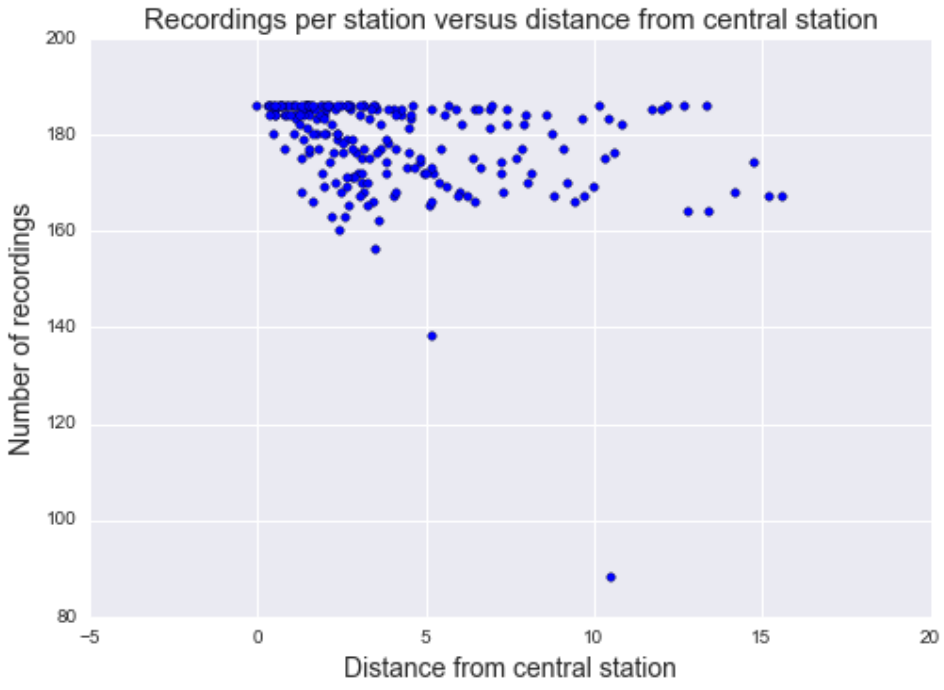

Figure 10: Scatter plot of distance from central station vs. number of recordings

Additionally, there were cases where the precipitation variable was zero and at the same time rain was equal to one; I believe this is impossible. For those cases, I replaced the zero precipitation values with the global average precipitation when it rained. This rudimentary imputation may have a small effect upon accuracy of modelling.