Analyzing the NYC Subway Dataset

Answers

Answers to the short questionnaire of first module (Introduction to Data Science) of Udacity's nanodegree in data science.

Author: Javier Torrente.
E-mail: jtorrente84@gmail.com / jtorrente@e-ucm.es / contact@jtorrente.info

Section 0. References

The most relevant source for information and contents I have used is Udacity course materials. Most of the code needed to complete this project was provided by Udacity to help the student complete the different problems and exercises of the Intro to Data Science course. On top of that code base, I have produced new code and improved the existing one to complete the project. The main source code file is *nycsubway/module1/project1.py* and can be accessed on GitHub (https://github.com/jtorrente/nyc-data-analysis).

Apart from Udacity's materials, I have used additional sources to get deeper insight into Mann-Whitney's U test, especially how effect sizes should be reported for this test. It is often argued that when reporting statistical analyses inference tests should be accompanied not only by the value of the statistic used (e.g. 't' or 'U') and the p-value (probability of likelihood of the null hypothesis), but also by an estimator of the effect size. This has several benefits. First, it allows for discussing how important the relationship found between dependent and independent variable is. Second, it facilitates meta-review of research results in a particular topic.

In this regard, I have used the rank-biserial coefficient as an estimator of effect size. I have used the next three references about this topic:

http://yatani.jp/teaching/doku.php?id=hcistats:mannwhitney
https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test#Rank-biserial_correlation

Wendt, H. W. (1972). Dealing with a common problem in Social science: A simplified rank-biserial coefficient of correlation based on the U statistic. European Journal of Social Psychology, 2(4), 463-465. http://doi.org/10.1002/ejsp.2420020412

Section 1. Statistical Test

1.1 Which statistical test did you use to analyze the NYC subway data?

I ran three Mann-Whitney U tests to analyze the effect of dichotomic independent variables `'fog'`, `'rain'` and `'weekday'` on dependent variable `'ENTRIESn_hourly'`. The Mann-Whitney test is the non-parametric equivalent test to the two independent samples t-test, and it is used to compare differences between two groups when the assumptions of the t-test are not met. This is the case, as visual inspection of the histograms for these three variables showed clear signs of non-normality (see Figure 1).

Did you use a one-tail or a two-tail P value? What is the null hypothesis?

I used a two-tail p-value as direction of the difference in ridership caused by the presence or absence or rain cannot be predicted beforehand. Two-tail tests are also more conservative than one-tail tests, which reduces the probability of Type I error (rejecting a null hypothesis that is actually true). Given the large amount of

data in this data set, this is actually an interesting approach, as quite often even small differences can be found statistically significant if the sample size is large enough.

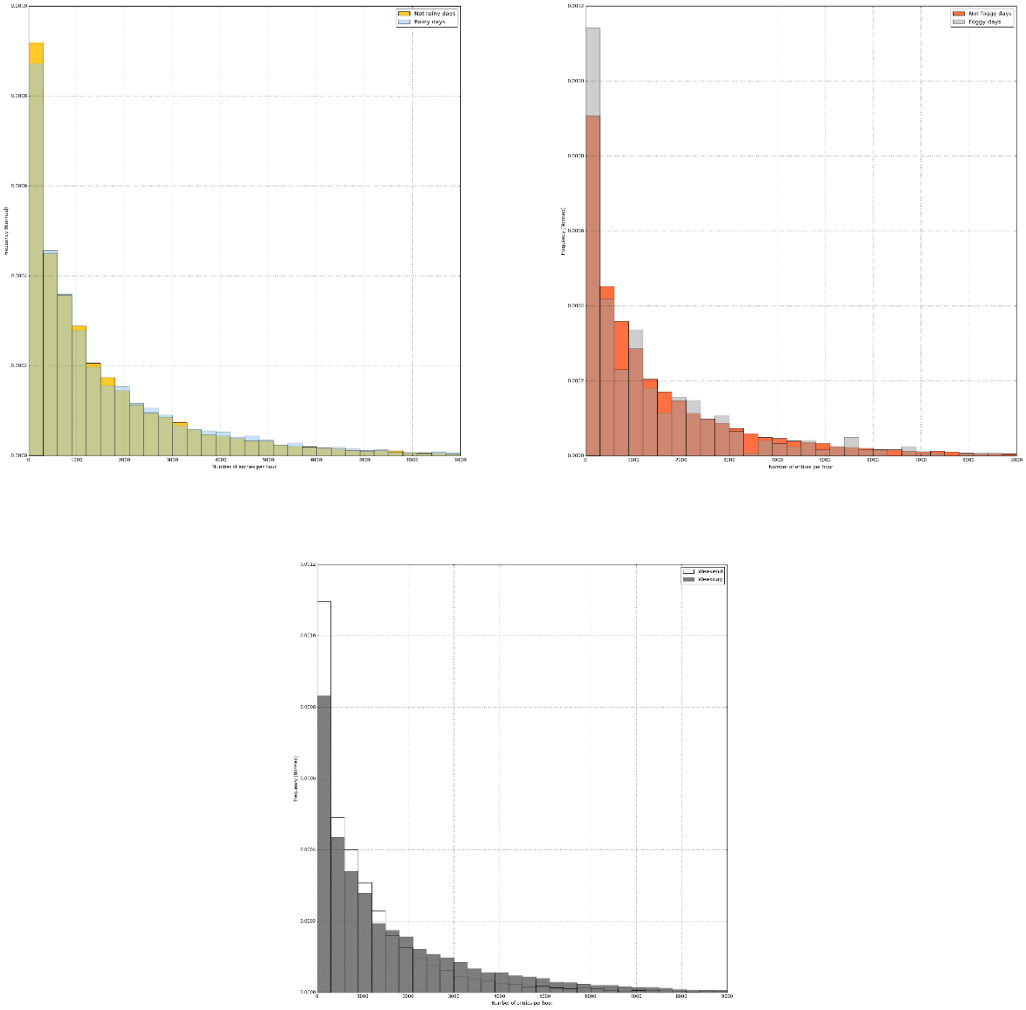

Figure 1. Histograms showing distribution of variable ENTRIESn_hourly for the different particular values of rain (top-left), fog (top-right) and weekday (bottom). Histograms are normalized to avoid differences in sample sizes between the groups. Distributions are right-skewed and cannot be considered normal.

Therefore, the 2-tail Mann-Whitney U test evaluates the null hypothesis that the two samples (in this case, readings when fog=1 and fog=0, for example) come from the same population. If that is the case, the probability of an element from sample A (e.g. not a foggy day) exceeding an element from sample B (e.g. foggy day) is 0.5. The alternative hypothesis is that the probability of elements in sample A exceeding elements in sample B or vice versa is higher than 0.5 (samples A and B come from different populations).

Example of null and alternative hypotheses:
Null hypothesis: "Fog does not affect ridership (ENTRIESn_hourly) – entries in foggy day samples and non-foggy days come from the same population".
Alternative hypothesis: "Ridership is different in foggy days than in non-foggy days".

What is your p-critical value?

Since I ran the tests against the whole dataset provided (`turnstile_weather_v2.csv`), which has 42,649 entries, I selected a p-critical value of 0.01 instead of the typical 0.05. The reason is that with such a large sample sizes, even small differences are likely to be identified as statistically significant. In these cases it is recommended to use a smaller critical p-value to limit the risk of having false positives.

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.

The Mann-Whitney U test is appropriate for all these three cases (fog, rain and weekday) because the next criteria are met:

- Observations are independent
- Sample sizes are greater than 20
- There are two groups to compare
- The dependent variable is measured at continuous or ordinal level

Also, this test is appropriate because the distributions of `ENTRIESn_hourly` split by `rain`, `fog` or `weekday` are not normal, as the next histograms illustrate (see Figure 1), and Mann-Whitney does not make the assumption of normality.

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.

Differences were observed in ridership (ENTRIESn_hourly) between rainy days (Mean = 2028.196, SD = 3189.267) and non-rainy days (Mean = 1845.539, SD = 2878.727) has been found statistically significant (Mann-Whitney U = 153635120.5, p<.001) being the effect size small (Rank biserial correlation 0.03044 < .1). Similarly, differences observed between foggy days (Mean = 1631.981, SD = 2358.708) and non-foggy days (Mean = 1889.116, SD = 2957.534), and between workdays (Mean = 2158.043, SD = 3282.466) and weekends (Mean = 1207.457, SD = 1709.232) were statistically significant (Ufog = 8167089.0, Uweekday = 149803470.0, pfog, pweekday < .001) having effect sizes of 0.077 and 0.193 respectively.

1.4 What is the significance and interpretation of these results?

All these three tests found significant differences. That is, the likelihood of observing these data being the null hypotheses true is inferior to .001. However, the actual impact of fog and rain in ridership is very small, compared to the impact of 'weekday', which has the strongest influence.

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. OLS using Statsmodels or Scikit Learn
2. Gradient descent using Scikit Learn
3. Or something different?

I used (2) Gradient descent using Scikit Learn, 15 iterations. From the two alternatives demonstrated in the course, this was the most appropriate as I run the analysis using the entire dataset.

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

Features that produced a better prediction model (R2 closer to 1) were `'meanprecipi'`, `'meantempi'`, `'hour'`, `'weekday'`. I used a dummy `'UNIT'` variable as in the example analyzed in the course.

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.

First I tried `'fog'` and `'rain'`, but the R^2 was not very good. Then I thought that using two dichotomic (0|1) variables like fog and rain produced a model that was just too simple. I looked at the data and realized that continuous variables like temperature and precipitations could provide more information. For example, it may be reasonable to think that extreme temperature could influence ridership even when fog and rain are not present.

However, the results using just `'meanprecipi'` and `'meantempi'` as predictors were not particularly impressive, so I thought I was still missing part of the story. Then I started looking into 'weekday' and 'hour', as it is reasonable to think these variables affect ridership. 'Weekday' proved to have more influence on ridership than weather conditions like `'fog'` and `'rain'`.

2.4 What are the parameters (also known as "coefficients" or "weights") of the non-dummy features in your linear regression model?

'meanprecipi'	'meantempi'	'hour'	'weekday'	Intercept
-833.31631106	-7.88374986	35.8960692	452.19339996	1604.71433622

2.5 What is your model's R^2 (coefficients of determination) value?
0.427061451647

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?
It means that around 42.70% of the variation in the sample can be explained through the linear model built. The plot of residuals shows errors are concentrated around 0 (see Figure 2), which can be considered an indicator of goodness of fit. However, the histogram presents long tails, suggesting that there are cases where accuracy is low and justifying further analysis.

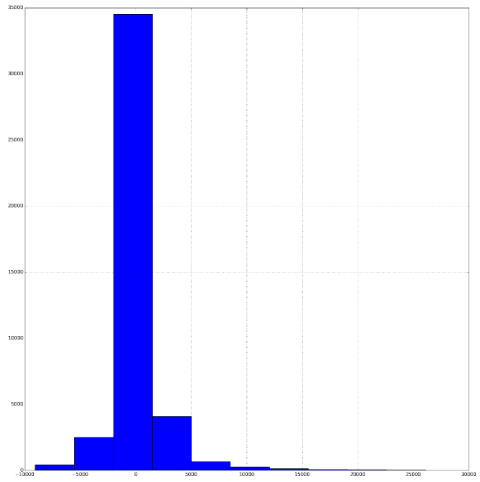

Figure 2. Histogram of residuals from the linear model

A probability plot of the residuals against the normal distribution shows that residuals are not normally distributed (see Figure 3). This is an indicator of poor goodness of fit of the linear model, as some predictions are extremely inaccurate, which reinforces the justification of further analysis.

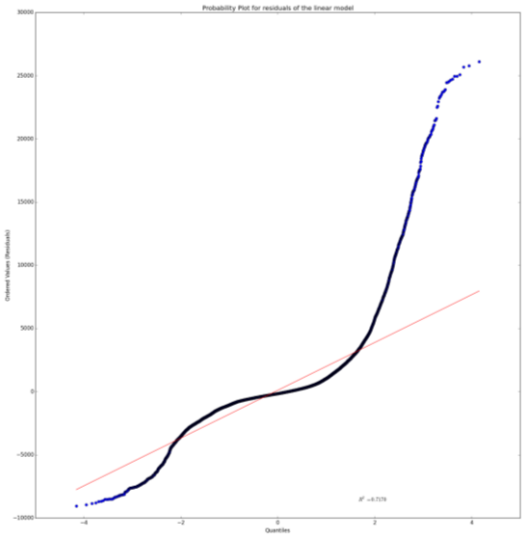

Figure 3. Probability plot of the linear model residuals against the normal distribution.

Section 3. Visualizations

Please include two visualizations that show the relationships between two or more variables in the NYC subway data.

3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

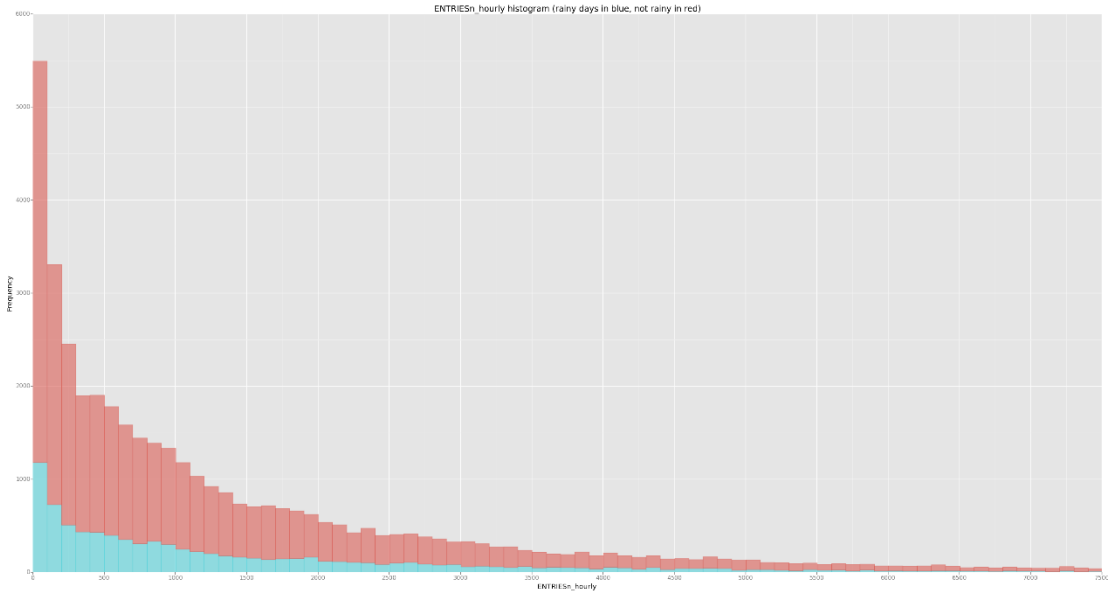

Figure 4.Histogram showing frequencies of values in ENTRIESn_hourly grouped by value of variable 'rain' (rainy days in blue, not rainy days in red). Cut-off point set up to 7500. Bin width set to 100.

3.2 One visualization can be more freeform.

I tried out different visualizations. One of the most interesting is displayed in Figure 5.

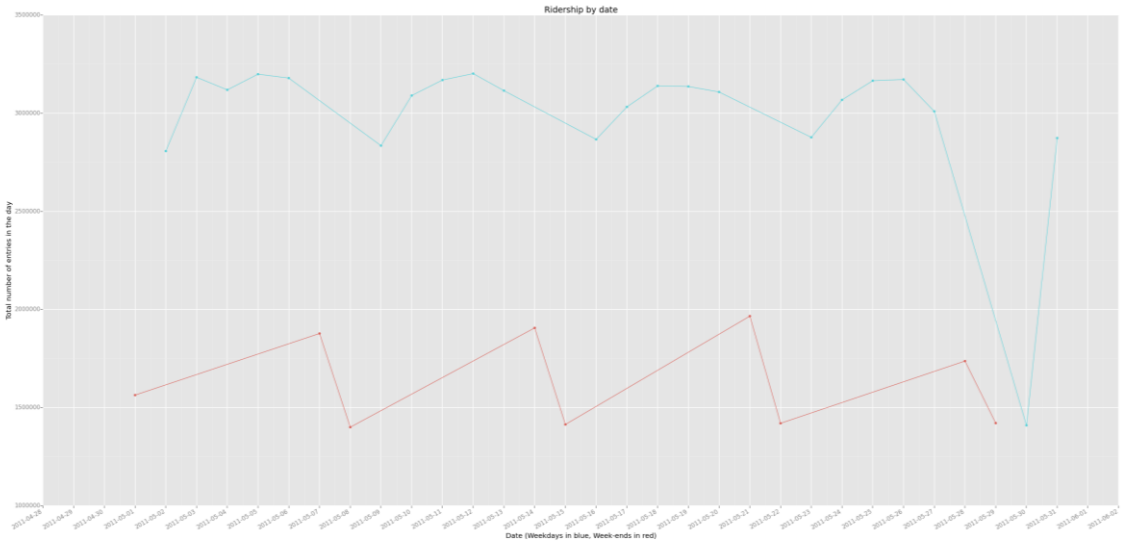

Figure 5. Ridership (accumulated value of ENTRIESn_hourly) per day. Weekends are shown in red and weekdays in blue. The chart shows ridership is higher in working days. Also, it can be observed that on the 30/05/2011 ridership was suspiciously low. This may be either because it was a bank holiday or caused by an error in the data.

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?
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.

In absolute figures, there are less people riding the subway when it is raining because there are less rainy days in the data set provided than rainy days. As a result, the not-rainy series of the histogram shown in Figure 4 is taller than the rainy days series.

However, relative ridership frequency increases when it is raining, as proved by the results of the Mann-Whitney U test. This test shows a difference of 182.65 hourly entries (in average) in favor of rainy days, being this difference statistically significant ($p<.001$).

Also, R^2 value obtained for the linear model decreases if variable `'meanprecipi'` is removed, which suggests that there's a relationship between precipitations (e.g. rain) and ridership.

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:

1. Dataset,
2. Analysis, such as the linear regression model or statistical test.

One of the main shortcomings I have found in the data set is the limited number of dates considered. The dataset contains entries corresponding to the period from 28th April 2011 to 2nd June 2011, both days included (36 days in total). All dates are concentrated in spring, when precipitations and temperature are different than in other seasons (winter, autumn and summer). Also, not including data from other years may be a source of bias (2011 could have been a particularly sunny or rainy year). To reach solid conclusions, data entries should be distributed uniformly across different years and the 12 months of the year.

The inference test used (Mann-Whitney) is appropriate for the analysis of the effect that rain has on ridership, as the conditions required by the test were met.

The linear model, however, showed some limitations. The non-normality of the residuals discussed in section 2 suggested that the model produces extremely inaccurate predictions in some cases. This motivated further analysis to understand under what circumstances the model produces poor predictions. Figure 6 shows a simple plot of all the residuals in the sample. A suspicious concentration of large residuals can be observed within the first 5,000-10,000 entries. A downward tendency in the size of the residuals can also be observed.

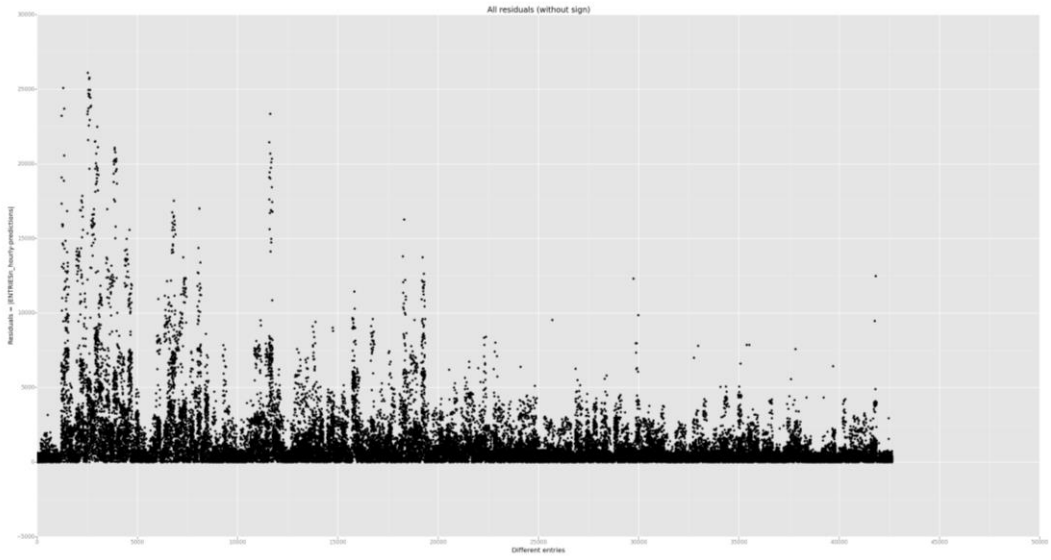

Figure 6. Size of the residuals (unsigned) for each data entry. Large residuals are concentrated on the left side.

A detailed analysis of the first 5,000 elements showed recurrent large differences between hourly entries and exits. The difference between hourly entries and exits was calculated and plotted, showing a similar pattern to the residuals' (see Figure 7). This unexplained effect could be introducing noise into the model.

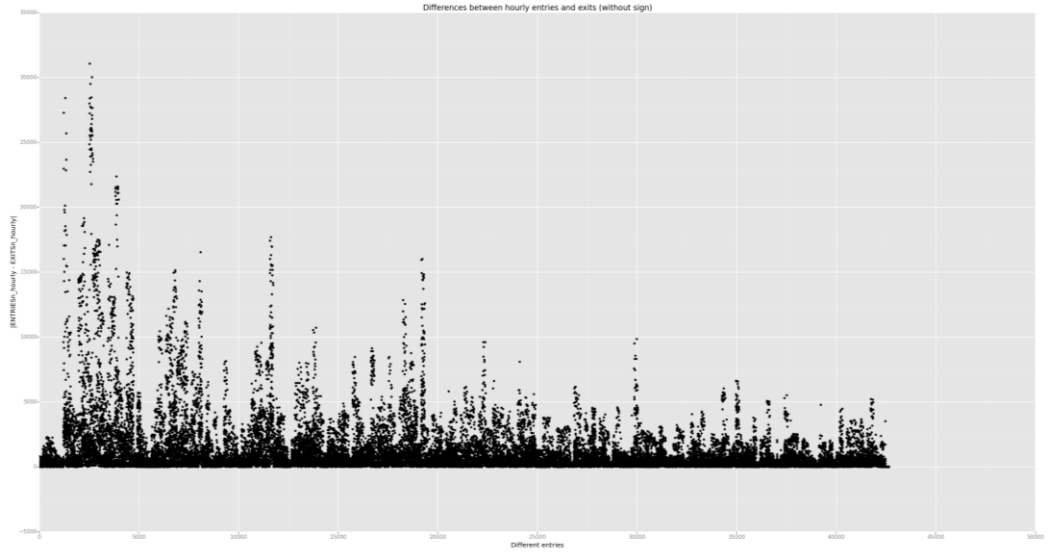

Figure 7. Difference between hourly entries and exits, unsigned, plotted for each data entry. Unusually large differences are also concentrated on the left side.

Co-linearity could be affecting the accuracy of the model. To evaluate this potential issue, correlations between features and hourly entries were analyzed. Data are provided in table 1:

	meanprecipi	**meantempi**	**hour**	**weekday**	**ENTRIESn_hourly**
meanprecipi	1.000000	-0.229034	-0.001771	0.112940	0.035649
meantempi	-0.229034	1.000000	0.001370	-0.014609	-0.026693
hour	-0.001771	0.001370	1.000000	-0.005271	0.286798
weekday	0.112940	-0.014609	-0.005271	1.000000	0.145431

Table 1. Correlation matrix for features of the linear model and values (ENTRIESn_hourly).

These correlations are not particularly strong. Also, factors did not apparently had unexpected values, which leads to think that collinearity was not a problem for this model.