## A note about resubmission

Thank you for the feedback on the original submission. I was horrified to find out that my original charts were in fact stacked bar charts. I am new to ggplot, but I was surprised that this was the default behaviour since I would not consider this to be the usual desired outcome. In fact, I found the position='dodge' option did not work correctly in Python, so I have redone a number of the charts in R. Regardless, those original plots heavily coloured my approach to the remainder of my analysis (a lesson learned I hope). So I have ended up redoing a considerable amount of the project hence the longer than hoped turn around for this resubmission. Thanks again.

## 0. References

1. http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html
2. http://stats.stackexchange.com/questions/124995/how-do-i-interpret-the-p-value-returned-in-scipys-mann-whitney-u-test
3. http://statsmodels.sourceforge.net/devel/examples/notebooks/generated/ols.html
4. https://en.wikipedia.org/wiki/Coefficient_of_determination
5. http://blog.minitab.com/blog/adventures-in-statistics/how-high-should-r-squared-be-in-regression-analysis
6. http://www.r-bloggers.com/can-we-do-better-than-r-squared/
7. https://stevencarlislewalker.wordpress.com/2013/06/18/calculating-the-press-statistic-in-r/
8. http://support.minitab.com/en-us/minitab/17/topic-library/modeling-statistics/regression-and-correlation/goodness-of-fit-statistics/r-squared/#what-is-predicted-r-squared
9. https://en.wikipedia.org/wiki/Multicollinearity#Detection_of_multicollinearity
10. http://ggplot.yhathq.com/docs/geom_histogram.html
11. http://www.cookbook-r.com/Graphs/index.html

## 1. Statistical Test

1.1 We used a Mann-Whitney U-Test to analyse the NYC subway data. We want to perform a two-tailed test because our alternative hypothesis is that the distributions of ridership when it is raining and when it's not raining are different (we make no assumptions about the direction). The null hypothesis is that the ridership during rainy periods and non-rainy periods have the same distribution. The *p*-critical value is 0.05.

1.2 This statistical test is applicable because it can be used to test a null hypothesis that two distributions are the same against an alternative hypothesis they are different. Unlike Welch's T test which requires distributions be normal (which the ridership data is not) the Mann-Whitney U test is non-parametric, meaning is can be used when the distributions are unknown.

The Mann-Whitney U test also requires that the two distributions are independent and the response variables are continuous or ordinal. The ridership data set satisfies these constraints.

1.3 The mean entries during periods with rain was 1105.446, while the mean entries during periods without rain was 1090.279. The test returned a *U* value of 1924409167.0 and a *p*-value of 0.04999982. (Note the *p*-value returned by the function in the online exercise returns the **one** tailed value 0.02499991. To use it in a two tailed test as we wish to, it must be doubled.)

1.4 Since the two tailed *p*-value is less than our critical value of 0.05 (albeit narrowly), we reject the null hypothesis that the distributions of ridership during rainy periods and non-rainy periods are equal. Instead there is significant evidence to suggest they are different distributions. That is, we believe the occurrence of rain affects the volume of riders on the NYC subway.

## 2. Linear Regression

2.1 I used OLS from StatsModels.

2.2 The features I selected in my model were unit, weekend (1 = weekend, 0 = weekday), hour, mean temperature, rain and fog. Unit and hour were encoded as dummy variables.

2.3 I selected this subset of the available features based on intuition and data visualization. Some variables were not considered because of obvious correlation with others.

Unit is a proxy for the location of the station. It makes sense that some stations are regularly busier than others. Weekend and hour were selected because there are obvious variations in the ridership based on these variables which are exhibited in the graphs in the Section 3. Hour was encoded as a dummy variable since we don't expect a linear relationship between ridership and hour. i.e. We don't expect a constant increase or decrease throughout a day but rather there to be peak and non-peak times.

The three weather features were selected because it was thought that these may have some influence on the decision to use the subway versus another form of transport. For example, if it was raining people might prefer to drive rather than walk to a subway station. Conversely, if it is foggy, people may be hesitant to drive in low visibility. Temperature too may factor into the decision to catch a train. There were other weather variables available, but some of these are obviously correlated. For example, rain and precipitation *must* be correlated.

All possible models using every combination of these chosen regressors was tested to in order to select the best from among them. The Python code (SubwayOls.py) used to achieve this has been included in the submission. A table of results for each of these regressions in included in Section 6.1. The $R^2$ values alone were not used as the basis for determining the best model. This is not a good method of selection since the $R^2$ statistic will always grow as more features are added to the model. *Adjusted* $R^2$ is a related statistic which adjusts for the number of regressors so is better suited to this task. However, I have also included a few other measures to assist in model selection.

Firstly there is the $F$-statistic and its associated $p$-value. This is the statistic used in a test for the significance of the regression parameters in the model. The null hypothesis is that all coefficients equal zero. The results show all but one model (rain as the only feature) are significant at the 0.05 level.

I have also included the *condition number*. This is also an output of the StatsModels OLS package. This is a measure of multicollinearity which is another important consideration in model selection. The higher this number, the more collinearity exists in the model. Existence of multicollinearity can cause the coefficients on the features to be unstable under even small changes to the input data. Inspecting the table of results provided one can see that the addition of mean temperature to all models had a large impact on this measure. Hence models including mean temperature have been excluded from consideration.

In fact the majority of the models have questionable condition numbers. A general rule is that anything above 30 or so spells trouble. But for the purpose of this exercise, I tried to reach a balance between using this number and adjusted $R^2$ so we will continue to consider those models not including mean temperature. Turning our attention to adjusted $R^2$, we see the best valid model remaining includes the factors *unit, weekend, hour, rain* and *fog*.

2.4 The coefficients of the non-dummy variables are provided in the following table. Also provided are the *t*-statistic and *p*-values of tests for the significance of the coefficients. The results of these tests are provided automatically with the output from StatsModels OLS. The null hypothesis for these tests is that the coefficient equals zero. So we are able to reject the null hypothesis at the 0.05 level in all cases.

| **Factor** | **Coefficient** | ***t*-statistic** | ***p*-value** |
|---|---:|---:|---:|
| Intercept | 4169.019836 | 34.070911 | < 0.0001 |
| Weekend | -561.825249 | -56.108507 | < 0.0001 |
| Rain | -84.350520 | -7.855264 | < 0.0001 |
| Fog | 57.741254 | 4.283480 | < 0.0001 |

2.5 The $R^2$ value is 0.512706. (Note this differs from the value returned from the online exercise – which is likely due to the fact the exercise is only using 10% of the data)

2.6 The above $R^2$ value means that around 51% of the variability in the in the response variable (entry counts) is explained by the model with the chosen features.

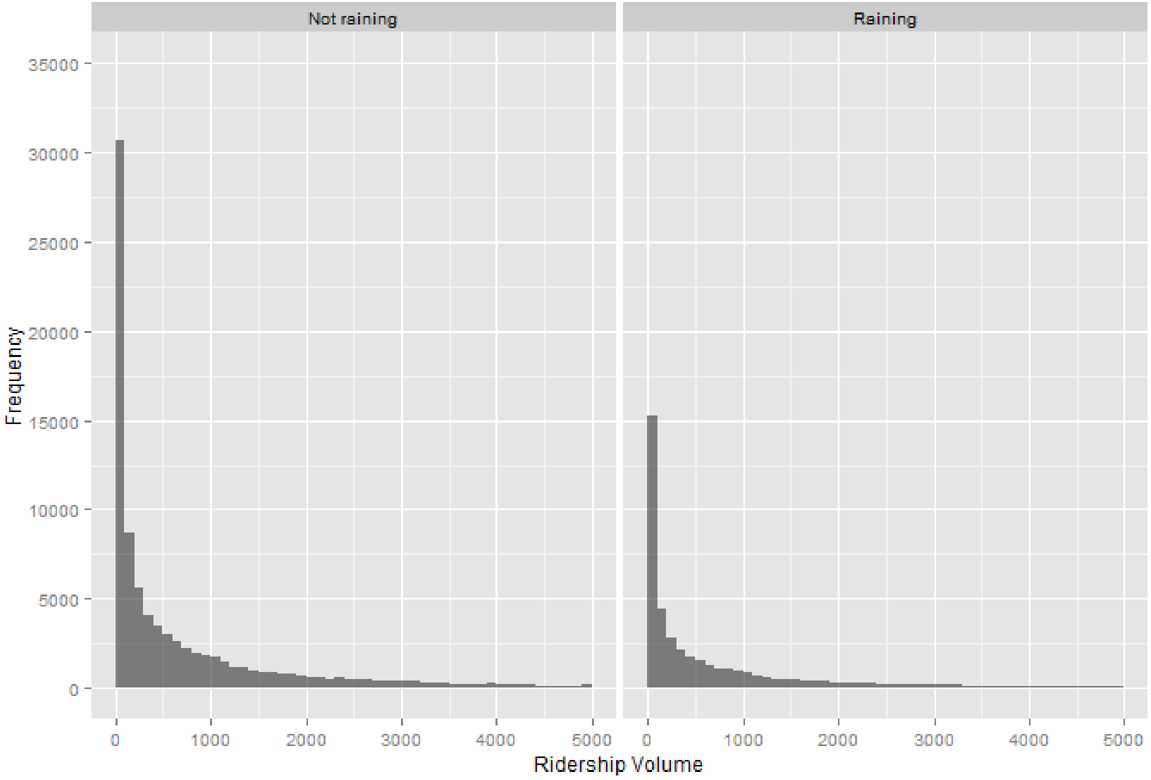
In response to the question: “Do you think this linear model to predict ridership is appropriate for this dataset, given this $R^2$ value?” – I don’t think this is a good question. Firstly, $R^2$ or even adjusted $R^2$ should not be used as a measure of predictive power. The predicted $R^2$ (derived from *PRESS*) can be used to choose among models if prediction is the goal. It has been included as the last column in the table in Section 6.1.

Using this statistic, we do find the same model, i.e. the one with all regressors is the best for predictive purposes (pred. $R^2$ = 0.507898). This measure is slightly smaller than both $R^2$ and adjusted $R^2$ which is to be expected. However, this statistic like $R^2$ doesn’t tell us if we *should* use the model for prediction. To answer this, we need more context. For what purpose are we predicting the ridership? How accurate do we require our predictions to be? If we know this, we can perhaps use the *prediction intervals* of the model to determine suitability.

However, if we are to base a decision on predicted $R^2$ alone, I would say no, the model is not suitable for prediction, because the predicted error sum of squares still accounts for less than half (1 - 0.507898 = 0.492102) of the total sum of squares.

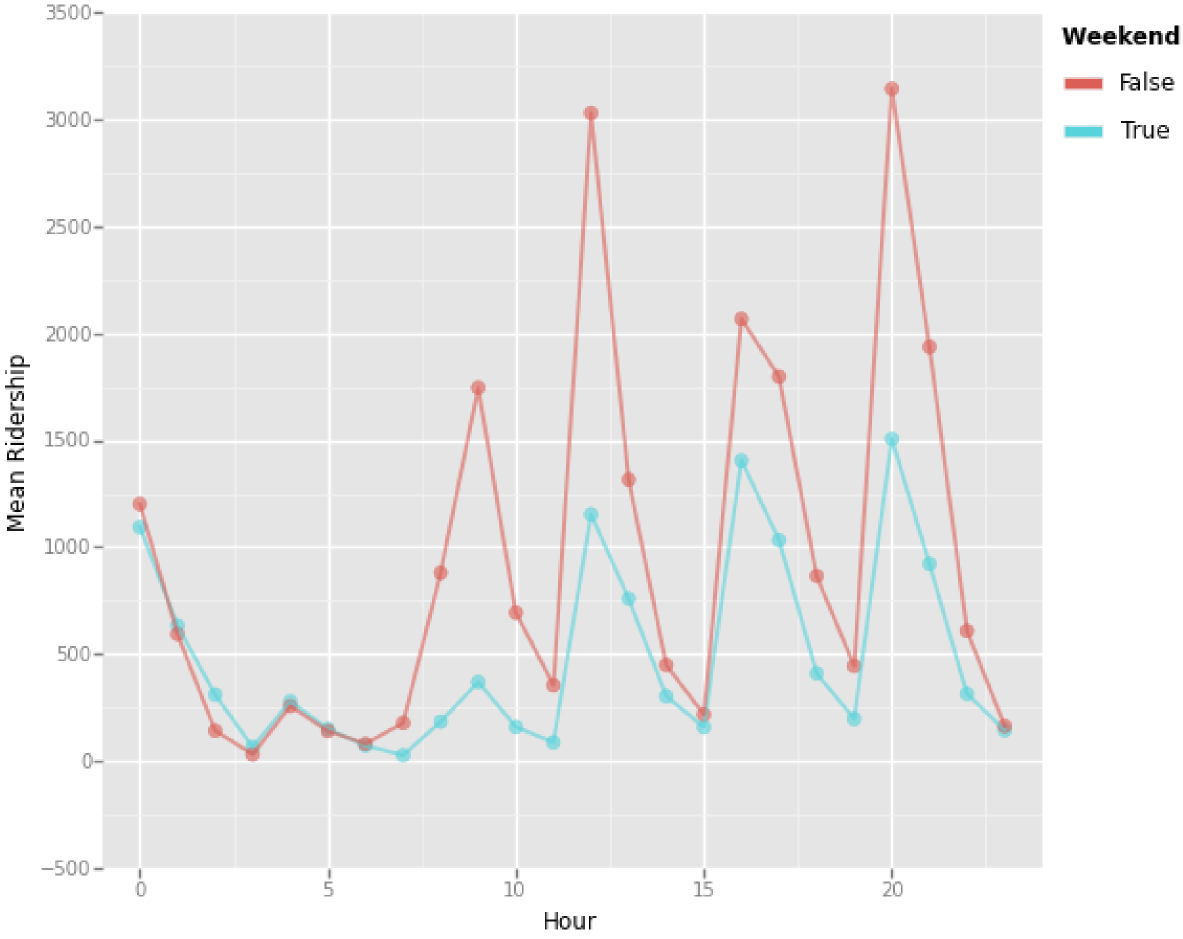
# 3. Visualization

## 3.1 Histograms of hourly entries for periods with rain and no rain



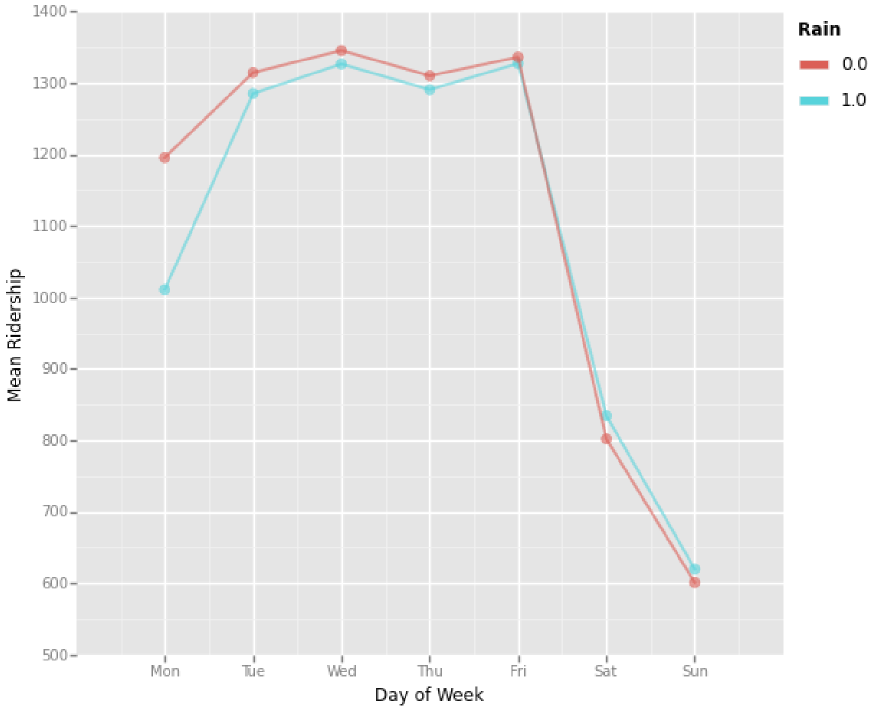
The horizontal axes of the histograms have been cut off at 5,000 entries to better show their form. There are very long tails on these distributions which make including them detrimental to its appearance. For information, the proportion of observations with hourly entries greater than 5,000 is only 4.64%. The figure shows that the distribution of ridership when there is rain is not as steep at the lower end of hourly entries as the distribution when the weather is fine. As we noted in the previous section, this results in the mean ridership during rainy periods being slightly greater than during non-rainy periods.

3.2 Mean ridership for weekdays and weekends by hour



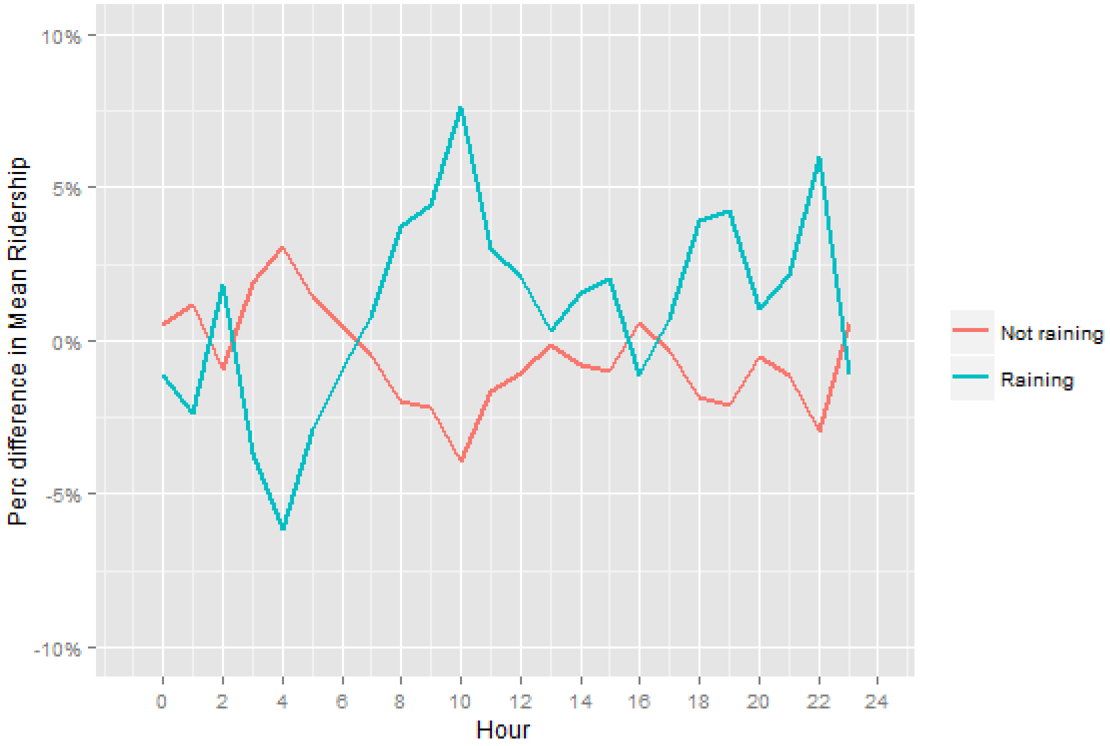
This chart shows there is a distinct difference between ridership on weekdays and weekends. At the same time we can also observe the variation in entries depending on the time of the day. There appears to be several peak times which happen every few hours. We can see "rushes" at 9am, midday, 4pm, 8pm, and midnight. Also, there is an understandable lull in ridership during the late night/early morning hours. This graph influenced the choice of regressors tested in the models in the previous section.

## 3.3 Mean ridership for rainy and non-rainy periods by day



When we break down the data by day, this chart suggests that during the week, more people ride the subway when it is not raining, but on the weekends, it is the opposite.

## 3.4 Difference from mean ridership for rainy and non-rainy periods by hour



This chart shows the relative difference between the means of ridership when it is raining and not raining to the overall mean, by hour. In contrast to the previous figure, it suggests that for the majority of the hours in a day, more people ride the subway when it's raining.

## 4. Conclusion

4.1 I believe more people ride the subway when it is *not* raining.

4.2 The charts provided in Section 3 show that it is easy to draw a different conclusion about the effect of rain depending on how the data is viewed. If we look at it by day (3.3), ridership appears generally higher when it is not raining. When we look at it by hour (3.4) it is the opposite. This disparity illustrates that we should not look at the effect of rain in isolation. That is, there a many other possible confounding variables which may have an effect on subway ridership.

One approach to accounting for these factors is to use a regression model. The model we selected in Section 2 includes rain as a factor but also other variables we considered to have an effect on subway ridership. The result of the *t*-test for the significance of the rain coefficient in the regression model was provided in the table in Section 2.4. The *t*-statistic for the rain coefficient is -7.855 which has a *p*-value < 0.0001. Therefore we can reject the null hypothesis that the coefficient equals zero at the 5% level. The rain feature has a coefficient of -84.351, which means that in the event of rain (rain = 1), this feature has a negative effect on the volume of ridership when all other factors in the model are kept constant.

Lastly, the Mann-Whitney U test performed in Section 1 rejected the null hypothesis that the distributions of hour entries were equal for rainy and non-rainy periods. However, this test was performed over all time periods and locations which we have seen were significant indicators of ridership.

## 5. Reflection

5.1 As mentioned in the last paragraph, I believe performing the statistical test in Section 1.1 across all time periods and locations is a short coming. The day of week, time of day and unit have been clearly shown to have an impact on the ridership volume, so better statistical tests would account for this by excluding these factors. For example, we could perform separate Mann-Whitney U tests for each level of these factors or alternatively, we could randomly select samples from the data set partitioned into rainy and non-rainy observations to control for these factors. Significance tests for the coefficients in the regression model also overcome this problem as the influence of all factors in the model are accounted for.

The dataset while being quite large, is restricted to only one month of observations (May 2011). To draw conclusions about the effect of the provided variables on the volume of ridership it would be better to have data for other times of the year as well.

# 6. Appendix

## 6.1 Results from performing all possible regressions on the variables chosen in Section 2.3

| Model | Cond. No. | *F*-stat | *p*-value | $R^2$ | Adj. $R^2$ | Pred. $R^2$ |
|---|---:|---:|---:|---:|---:|---:|
| unit, weekend, hour, rain, fog, mean temp | 37273.013 | 282.051 | 0.000000 | 0.513017 | 0.511198 | 0.508207 |
| unit, weekend, hour, rain, mean temp | 37272.804 | 282.484 | 0.000000 | 0.512890 | 0.511074 | 0.508086 |
| unit, weekend, hour, rain, fog | 667.765 | 282.276 | 0.000000 | 0.512706 | 0.510889 | 0.507898 |
| unit, weekend, hour, fog, mean temp | 37272.522 | 282.244 | 0.000000 | 0.512678 | 0.510861 | 0.507870 |
| unit, weekend, hour, mean temp | 37272.331 | 282.812 | 0.000000 | 0.512667 | 0.510854 | 0.507868 |
| unit, weekend, hour, rain | 658.341 | 282.779 | 0.000000 | 0.512638 | 0.510825 | 0.507837 |
| unit, weekend, hour | 625.860 | 283.174 | 0.000000 | 0.512474 | 0.510664 | 0.507679 |
| unit, weekend, hour, fog | 633.556 | 282.597 | 0.000000 | 0.512477 | 0.510663 | 0.507674 |
| unit, hour, rain, fog, mean temp | 37272.595 | 270.066 | 0.000000 | 0.501653 | 0.499796 | 0.496697 |
| unit, hour, fog, mean temp | 37272.099 | 270.567 | 0.000000 | 0.501605 | 0.499751 | 0.496655 |
| unit, hour, mean temp | 37271.916 | 270.896 | 0.000000 | 0.501394 | 0.499543 | 0.496450 |
| unit, hour, rain, mean temp | 37272.389 | 270.340 | 0.000000 | 0.501395 | 0.499540 | 0.496443 |
| unit, hour, fog | 605.531 | 270.505 | 0.000000 | 0.501033 | 0.499181 | 0.496083 |
| unit, hour, rain, fog | 644.366 | 269.953 | 0.000000 | 0.501036 | 0.499180 | 0.496078 |
| unit, hour, rain | 633.706 | 270.356 | 0.000000 | 0.500895 | 0.499042 | 0.495944 |
| unit, hour | 596.695 | 270.895 | 0.000000 | 0.500879 | 0.499030 | 0.495935 |
| unit, weekend, rain, fog, mean temp | 37150.674 | 211.490 | 0.000000 | 0.429478 | 0.427447 | 0.423760 |
| unit, weekend, rain, mean temp | 37150.535 | 211.858 | 0.000000 | 0.429378 | 0.427351 | 0.423668 |
| unit, weekend, rain, fog | 653.288 | 211.711 | 0.000000 | 0.429209 | 0.427181 | 0.423494 |
| unit, weekend, fog, mean temp | 37150.202 | 211.674 | 0.000000 | 0.429166 | 0.427138 | 0.423452 |
| unit, weekend, mean temp | 37150.064 | 212.126 | 0.000000 | 0.429161 | 0.427138 | 0.423457 |
| unit, weekend, rain | 643.605 | 212.123 | 0.000000 | 0.429157 | 0.427134 | 0.423452 |
| unit, weekend | 610.421 | 212.439 | 0.000000 | 0.428994 | 0.426975 | 0.423296 |
| unit, weekend, fog | 618.262 | 211.982 | 0.000000 | 0.428995 | 0.426971 | 0.423287 |
| unit, rain, fog, mean temp | 37150.276 | 203.051 | 0.000000 | 0.419008 | 0.416944 | 0.413160 |
| unit, fog, mean temp | 37149.804 | 203.451 | 0.000000 | 0.418964 | 0.416904 | 0.413123 |
| unit, mean temp | 37149.666 | 203.750 | 0.000000 | 0.418796 | 0.416740 | 0.412963 |
| unit, rain, mean temp | 37150.137 | 203.311 | 0.000000 | 0.418796 | 0.416736 | 0.412955 |
| unit, fog | 589.533 | 203.470 | 0.000000 | 0.418461 | 0.416404 | 0.412622 |
| unit, rain, fog | 629.577 | 203.034 | 0.000000 | 0.418464 | 0.416403 | 0.412616 |
| unit, rain | 618.555 | 203.378 | 0.000000 | 0.418351 | 0.416294 | 0.412512 |
| unit | 580.451 | 203.808 | 0.000000 | 0.418340 | 0.416288 | 0.412510 |
| weekend, hour, rain, fog, mean temp | 1266.087 | 753.306 | 0.000000 | 0.133580 | 0.133403 | 0.133304 |
| weekend, hour, rain, mean temp | 1265.841 | 781.424 | 0.000000 | 0.133453 | 0.133282 | 0.133191 |
| weekend, hour, rain, fog | 22.391 | 779.821 | 0.000000 | 0.133216 | 0.133045 | 0.132950 |
| weekend, hour, rain | 22.069 | 810.572 | 0.000000 | 0.133152 | 0.132988 | 0.132901 |
| weekend, hour, mean temp | 1265.257 | 810.511 | 0.000000 | 0.133143 | 0.132979 | 0.132894 |
| weekend, hour, fog, mean temp | 1265.267 | 779.358 | 0.000000 | 0.133147 | 0.132976 | 0.132884 |
| weekend, hour | 20.962 | 842.638 | 0.000000 | 0.132917 | 0.132760 | 0.132680 |
| weekend, hour, fog | 21.224 | 808.928 | 0.000000 | 0.132918 | 0.132753 | 0.132666 |
| hour, rain, fog, mean temp | 1265.778 | 687.182 | 0.000000 | 0.119278 | 0.119104 | 0.119008 |
| hour, fog, mean temp | 1265.051 | 714.241 | 0.000000 | 0.119214 | 0.119047 | 0.118957 |
| hour, mean temp | 1265.030 | 742.487 | 0.000000 | 0.119000 | 0.118839 | 0.118757 |

| | | | | | | |
|---|---|---|---|---|---|---|
| hour, rain, mean temp | 1265.506 | 712.783 | 0.000000 | 0.119000 | 0.118833 | 0.118744 |
| hour, fog | 20.261 | 739.192 | 0.000000 | 0.118534 | 0.118374 | 0.118289 |
| hour, rain, fog | 21.588 | 709.652 | 0.000000 | 0.118539 | 0.118372 | 0.118280 |
| hour, rain | 21.223 | 738.199 | 0.000000 | 0.118394 | 0.118233 | 0.118149 |
| hour | 19.959 | 770.209 | 0.000000 | 0.118381 | 0.118228 | 0.118151 |
| weekend, rain, fog, mean temp | 662.501 | 431.321 | 0.000000 | 0.012907 | 0.012877 | 0.012838 |
| weekend, rain, mean temp | 652.348 | 570.130 | 0.000000 | 0.012797 | 0.012774 | 0.012744 |
| weekend, rain, fog | 3.722 | 561.260 | 0.000000 | 0.012600 | 0.012578 | 0.012543 |
| weekend, rain | 2.893 | 838.122 | 0.000000 | 0.012544 | 0.012529 | 0.012504 |
| weekend, mean temp | 635.681 | 837.004 | 0.000000 | 0.012528 | 0.012513 | 0.012490 |
| weekend, fog, mean temp | 636.402 | 558.148 | 0.000000 | 0.012531 | 0.012509 | 0.012477 |
| weekend | 2.427 | 1648.565 | 0.000000 | 0.012340 | 0.012332 | 0.012315 |
| weekend, fog | 2.995 | 824.291 | 0.000000 | 0.012340 | 0.012325 | 0.012299 |
| rain, fog, mean temp | 662.459 | 33.342 | 0.000000 | 0.000757 | 0.000735 | 0.000698 |
| fog, mean temp | 636.249 | 46.274 | 0.000000 | 0.000701 | 0.000686 | 0.000656 |
| mean temp | 635.591 | 68.604 | 0.000000 | 0.000520 | 0.000512 | 0.000491 |
| rain, mean temp | 652.230 | 34.312 | 0.000000 | 0.000520 | 0.000505 | 0.000476 |
| fog | 2.767 | 17.053 | 0.000036 | 0.000129 | 0.000122 | 0.000097 |
| rain, fog | 3.581 | 8.854 | 0.000143 | 0.000134 | 0.000119 | 0.000087 |
| rain | 2.414 | 1.237 | 0.266089 | 0.000009 | 0.000002 | -0.000021 |