Project 1: Do more people ride the NYC subway when it is raining?

Resubmission based on grader feedback. For changed sections, look for **- UPDATED** in the heading.

1 Statistical Test - UPDATED

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

The statistical test was the Mann-Whitney U test, which produces a 1-tailed p-value.

Did you use a one-tail or a two-tail P value?

However, our test was non-directional: we wanted to see if `ENTRIESn_hourly` was *different* on days with rain. Perhaps rain makes people more inclined to stay home, or else take a taxi or bus to avoid getting soaked on the longer walk to a subway station.

On the other hand, maybe the denizens of NYC commuters who normally walk instead opt for the subway when it's wet outside.

To reflect this non-directionality, I multiplied the resulting 1-tail P value by 2.

What is the null hypothesis?

The null hypothesis is that the rainy day population (days where `rain` = 1) and non-rainy day population (days where `rain` = 0) are the same.

What is your p-critical value?

The p-critical value is +/- 0.05, the standard 2-tailed p-critical value.

1.2 Why is this statistical test applicable to the dataset?

As you can see in section 3.1, `ENTRIESn_hourly` does not follow a normal distribution on rainy nor on non-rainy days; it is positively skewed for both.

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.

```
1105.4463767458733, 1090.278780151855, 1924409167.0, 0.024999912793489721
```

1.4 What is the significance and interpretation of these results?

A p-value (two-tailed) of 0.04999982558698 indicates that there is a low likelihood of getting such a difference in means between the rain sample (mean = 1105.4463767458733) and the without rain sample (mean = 1090.278780151855) if they came from the same population.

2 Linear Regression

2.1 Approach used

My linear regression model used the Ordinary Least Squares (OLS) from statsmodel.

2.2 Features

Numeric variables

The model uses 1 numeric variable, `tempi`, from the improved CSV file

Dummy variables

There were 3 dummy variables:

- `remote unit`
- `hour`
- `day_week`

2.3 Why these variables?

Having spent 4 years commuting on the busiest line of the Paris metro (Line 1), I learned to anticipate how crowded my train would be just by looking at the clock and the day of the week, and knowing which stations were on my route.

On the other hand, finding a logical relationship between any weather-based variables and hourly entries was trial and error, helped a bit by data exploration using graphs. Often weather variables would increase r^2, but would have a high P-value and a confidence interval that crossed 0. Or else their coefficients had the opposite sign to what I would have logically expected.

It was only after I learned how to correct for multicollinearity as well as add multiple dummy variables to the model that coefficient for the `tempi` variable became negative (and barely increased the condition number), and so I included it in my final model.

2.4 parameters

- `constant`: 1886.59
- `tempi`: -119.8393

2.5 R-squared

- R-squared: 0.545
- Adjusted R-squared: 0.543

2.6. Interpretation - UPDATED

This R-squared value means that just over half of the variation in `ENTRIESn_hourly` (.545) can be explained by the values of the input variables. So based on R-squared, this model is fairly weak. Examining the residual distribution, and its relationship to the independent variables as well as predicted value also implies weakness in the model.

Compare predicted to actual

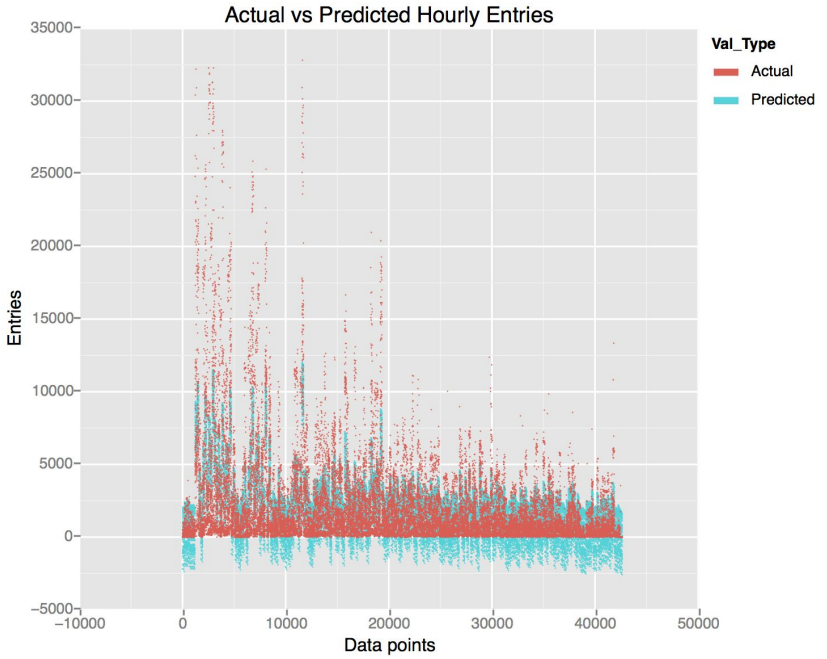

Figure 2-1: One problem with the model is it predicted a many negative entries!

Residual analysis

A test of the linear model's robustness is the randomness of the residuals. The deterministic part of the model are the coefficient & independent variables; the non-deterministic part of

the model is the errors, or residuals. If these errors are in fact **not** random, it means there is either a problem with variable selection, or a problem with the linear regression itself.

The randomness of residuals is demonstrated by:

- a mean of zero
- a normal distribution
- no discernible pattern of residuals with respect to the dependant or indepedent variables

So let's ask some questions about the model's residuals.

1. **Plot variables against the independent variables: are there any patterns?**

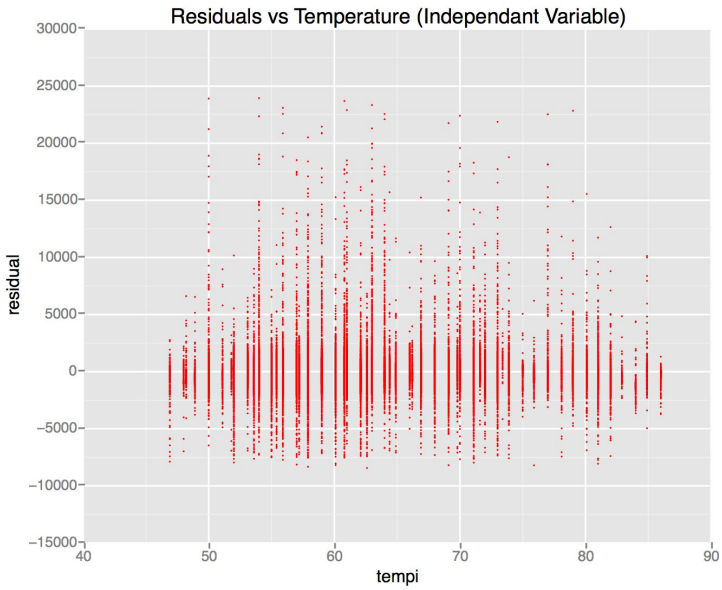

Figure 2-2: Residuals follow an accordian-like pattern: at either end of the x-axis (temperature) the variance of the residuals is more compressed. In the center (65℉), residual variance is even more compressed with virtually no outliers. This pattern is repeated to a lesser extent at 75℉.

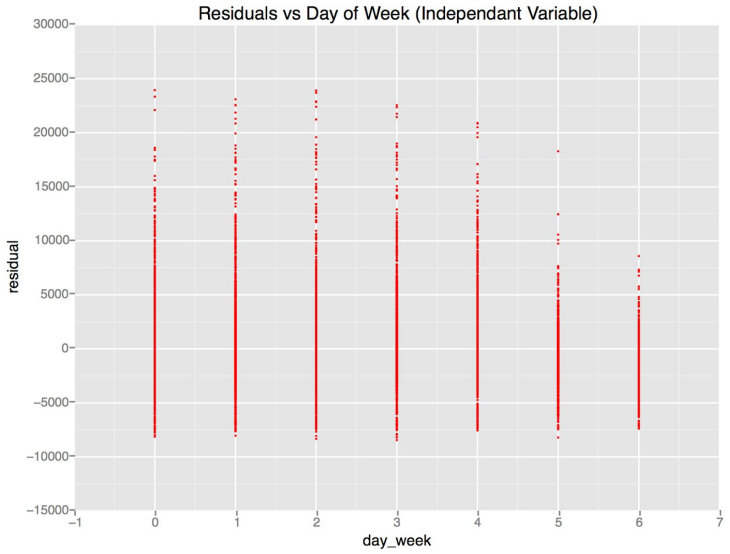

Figure 2-3: The pattern of variance for residuals by day of the week is similar to that of Figure 4-1, which shows daily averages of ENTRIESn_hourly: both are markedly reduced on weekends.

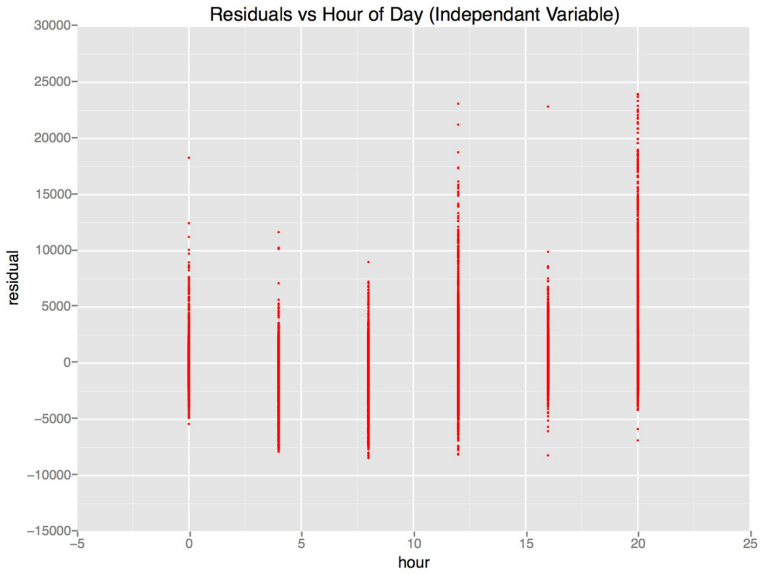

Figure 2-4: Similarly, the greatest residual variance falls at the peak traffic times of noon and 8 pm.

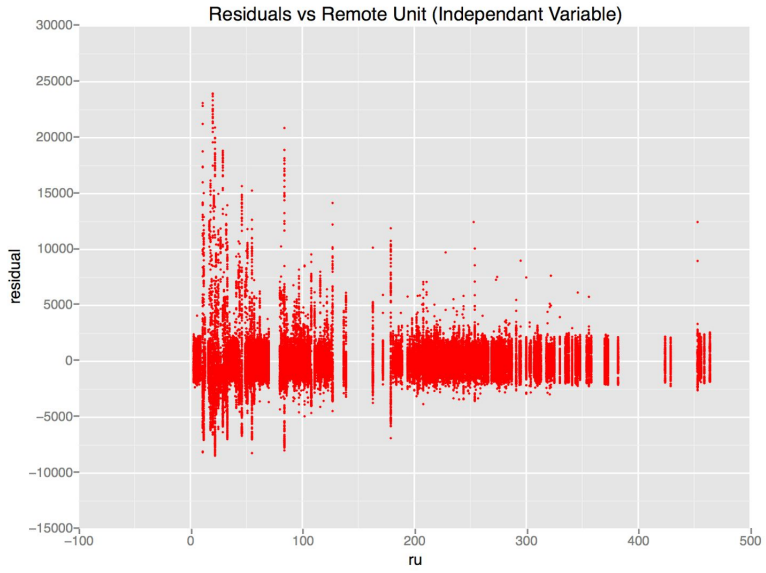

Figure 2-5: This looks extremely similar to Figure 2-6, which plots residuals along the data points:

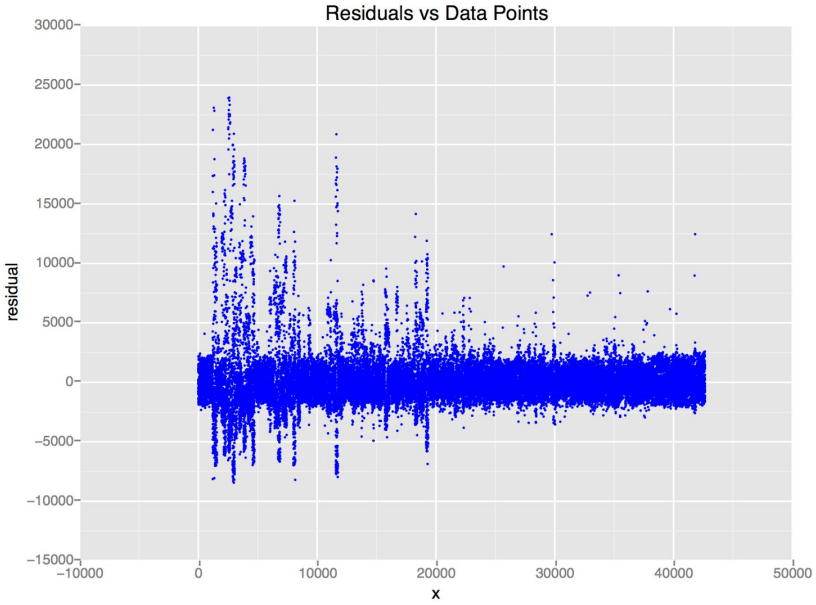

Figure 2-6: Plot of actual data points and their residuals.

2. Examine residual plots for outliers.

While the mean is near zero and the distribution appears to be normal, there is a very long positive tail extending to nearly 25000.

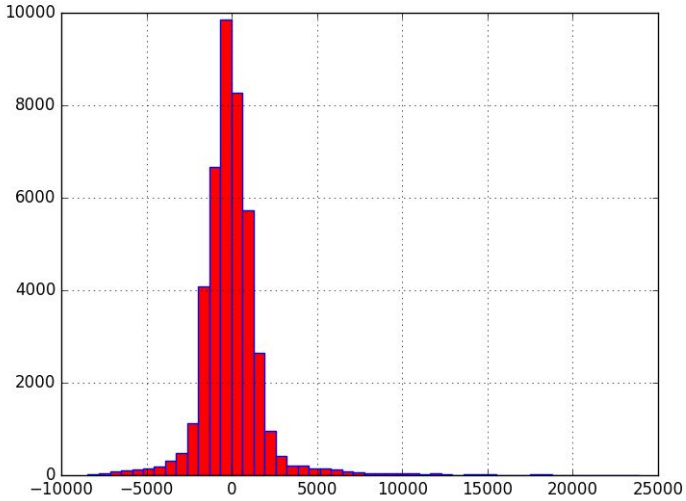

Figure 2-7: A mostly-normal distribution (with a mean of of approximately zero), but with a very long tail.

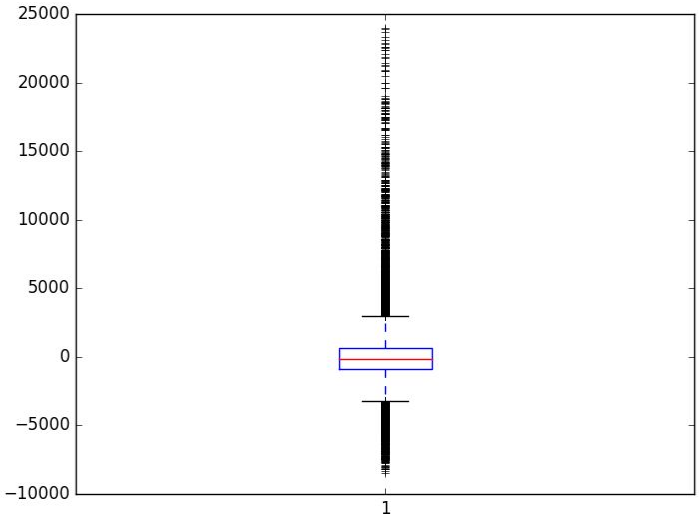

Figure 2-8: This boxplot of residuals shows many negative & positive outliers.

3. Plot residuals on a probability plot

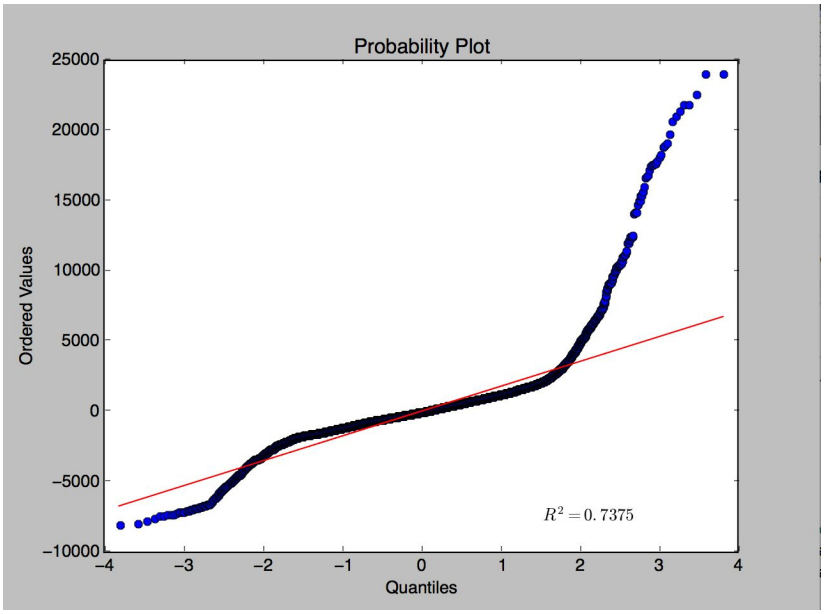

Figure 2-9: The residuals seem to follow a normal distribution until +2 standard deviations from the mean.

As the histogram in figure 2-8 implied, the distribution as shown in figure 2-9 seems to be "mostly normal" in that for +/- 2 standard deviations it follows the normal distribution line (the red line). Where the negative tail moves away from the normal distribution line, it still follows the line which implies a linear relationship. On the other hand, the positive tail veers steeply away from the line.

4. Plot residuals against fitted values

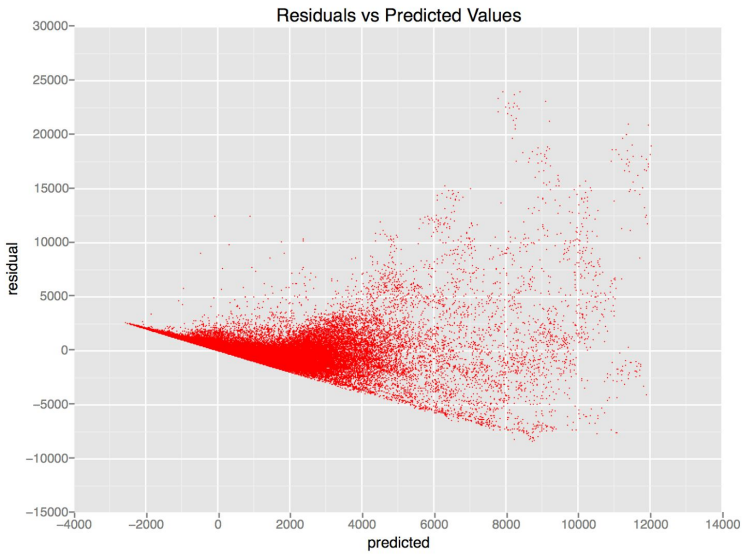

Figure 2-10: As the predicted value of subway entries increases, the variance of residuals dramatically increases. This changing variance along the x-axis suggests that the errors in the model are not random.

Conclusion of residual analysis

While the mean is zero and the distribution is roughly normal, there are

- many outliers, especially on the positive side
- there is a clear pattern between the residuals and the independent variables
- there is an even more distinct pattern between residuals and the predicted values.

All of this call into doubt the validity of the model.

3. Visualization

3.1 Histogram

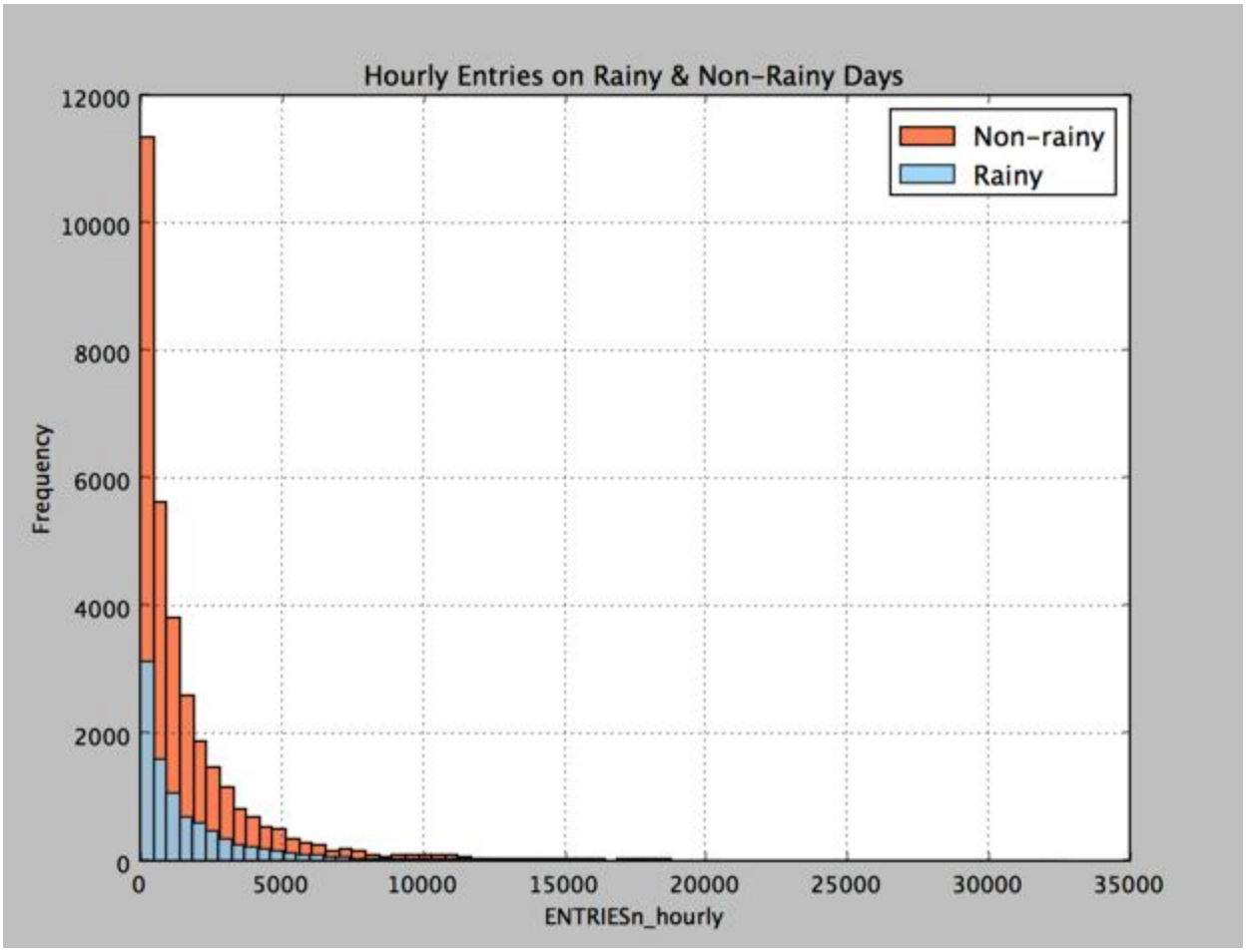

Figure 3-1: Both non-rainy and rainy samples are positively skewed.

3.2 Other graph

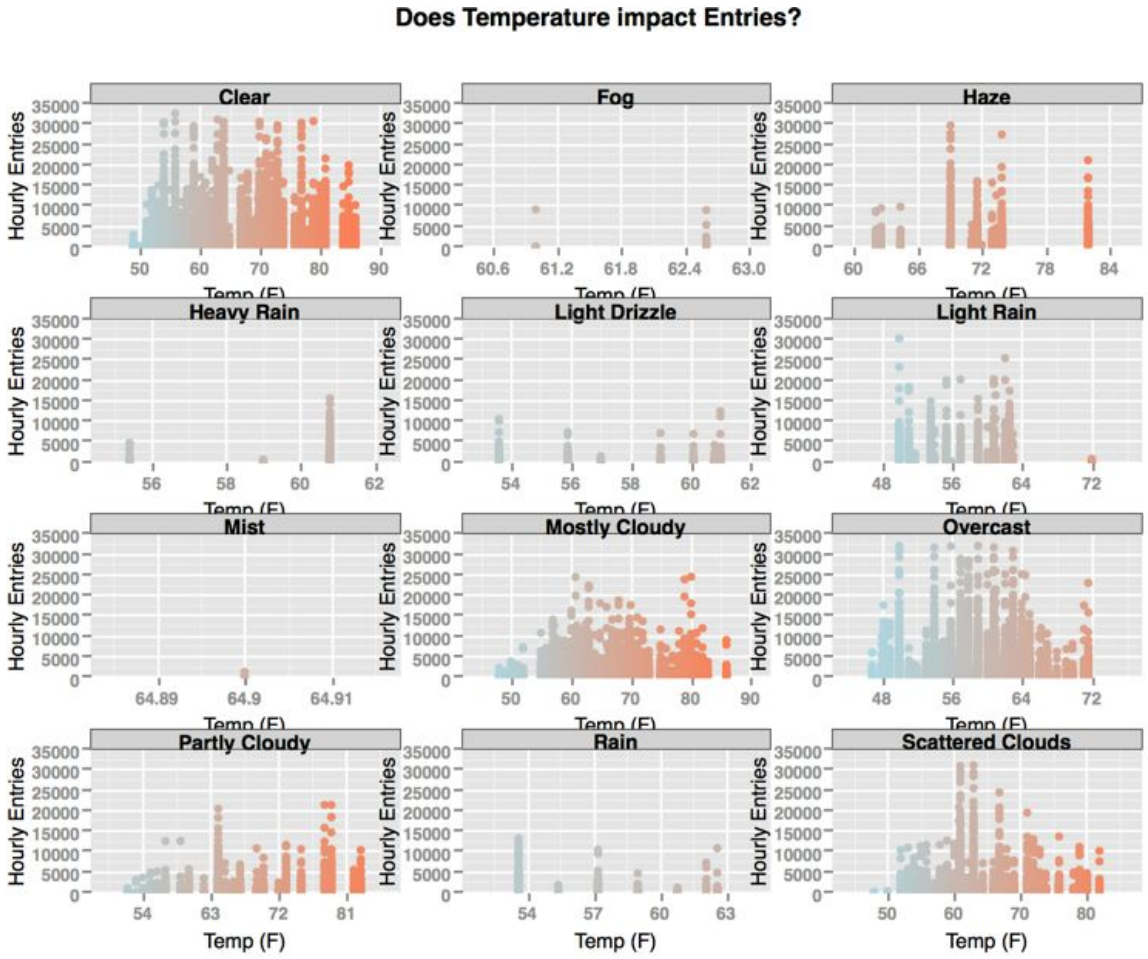

Figure 3-2: There's a lot going on in this facet graph, but I created it mainly to look at the patterns for Heavy Rain, Light Drizzle, Light Rain, and Rain conditions, and if temperature makes a difference. With the exception of Light Rain (and to a lesser extent, Heavy Rain), none of these 'rain' conditions led to any dramatic spikes in hourly entries. This might also be because there weren't many data points for them.

Interestingly, what does seem to change with weather conditions is the positive or negative correlation between temperature and hourly entries

4. Conclusion - UPDATED

4.1 Do more people ride the NYC subway when it is raining or when it is not raining?

In May 2011, it was significantly more likely that the average hourly entries would be higher on days with rain than those without, based on the results of the Mann-Whitney U test.

This is also born out when you look at a boxplot and statistical description of the two samples.

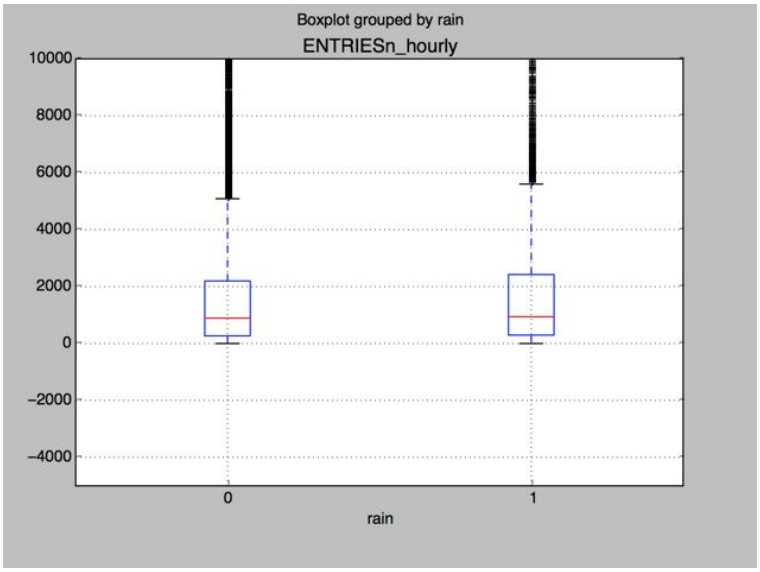

Figure 4-1: Comparing boxplots of `ENTRIESn_hourly` for rain and non-rain groups

```
In [17]: turnstile_weather.ENTRIESn_hourly[turnstile_weather.rain==0].describe()

Out[17]: count    33064.000000
         mean      1845.539439
         std       2878.770848
         min          0.000000
         25%        269.000000
         50%        893.000000
         75%       2197.000000
         max      32814.000000
         Name: ENTRIESn_hourly, dtype: float64

In [10]: turnstile_weather.ENTRIESn_hourly[turnstile_weather.rain==1].describe()

Out[10]: count     9585.000000
         mean      2028.196035
         std       3189.433373
         min          0.000000
         25%        295.000000
         50%        939.000000
         75%       2424.000000
         max      32289.000000
         Name: ENTRIESn_hourly, dtype: float64
```

Figure 4-2: Descriptive statistics for both groups. Both median and IQR are larger for the 'rain' group.

However, my OLS linear regression model showed that `rain` was not a significant input variable for predicting hourly entries, and that there were more significant variables, with better confidence intervals, that improved the R-squared value more noticeably.

This contradiction leads me to conclude that there are lurking variables behind the difference in hourly entries for May 2011. It cannot be attributed to rain alone.

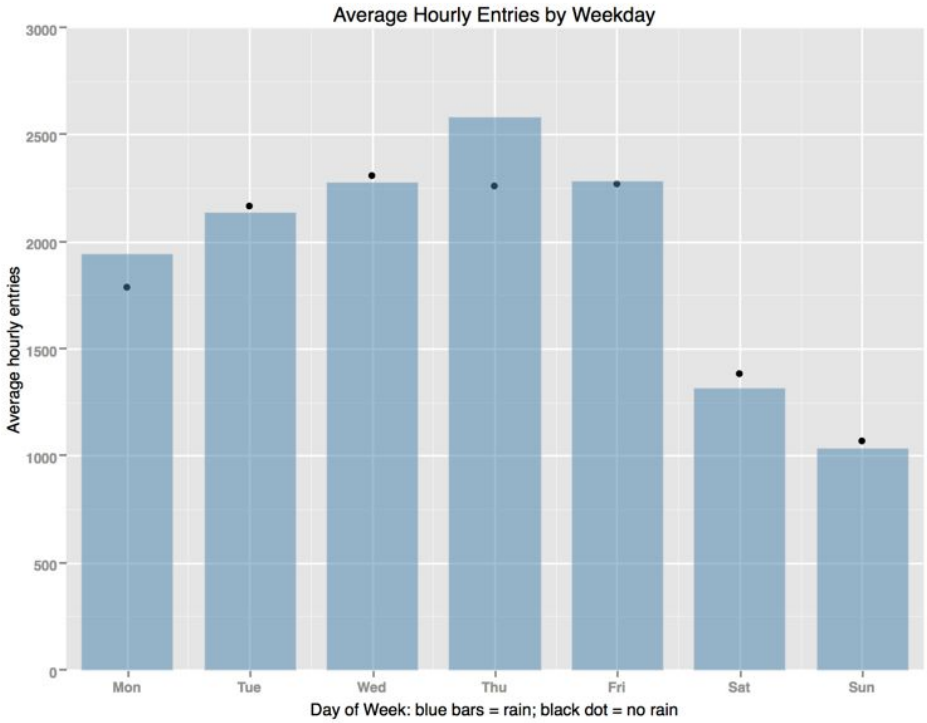

Figure 4-3: While ridership is noticeably higher when there is rain vs. no rain on Monday and Thursday, the difference is not as great as the variation in ridership on different days of the week, particularly weekends.

4.2 What analyses led you to this conclusion?

The Mann-Whitney U test showed that mean ridership was significantly greater with rain (mean = 1105.4463767458733) than without rain (mean = 1090.278780151855), U = 1924409167.0, p =0.024999912793489721 one-tailed.

To analyze the predictions, I modified the linear_regression function slightly to return this: `model.fit().summary()`. My explanations below are based on this summary.

When I made a simple model with `rain` as the only variable, the R-squared was only 0.075. By contrast, replacing `rain` with `tempi` returned an R-squared of 0.296.

Adding `UNIT` as a dummy variable, adding a constant, and dropping `unit_R022` (for 34 St-Herald Square station) and normalizing increased R-squared to 0.382.

- The unit-coefficients' signs make sense: since St-Herald Square is in midtown Manhattan, close to Times Square and Grand Central Station, all other units have negative coefficients. The one exception, R084, is 59 St-Columbus station, at the south-west corner of Central Park, serving 4 subway lines, according to Google Maps.
- However, the `tempi` coefficient is positive, which is counter-intuitive. When it's colder outside, one would expect more people to take the subway instead of walking.

Adding `hour` as an additional dummy variable and dropping `hour_0` (midnight) increased R-squared to 0.518.

- Now the `tempi` coefficient is negative.
- The `hour` coefficients make sense: those for 4 am and 8 am are negative, while lunch trips and returning commuters can explain the positive coefficients for noon, 4 pm and 8 pm.

Finally, adding `day_week` as the third dummy variable and dropping `day_0` (Monday) improved R-squared to 0.545.

- The `tempi` coefficient remains unchanged.
- Again, the `day_week` coefficients' signs are logical: positive for Tuesday-Friday, negative for Saturday and Sunday.

What happens if we now replace `tempi` with `rain`?

- R-squared goes down slightly, to 0.544.
- The `rain` coefficient is positive, but has a very small t-value (1.628) and a relatively high p-value (0.104) indicating that this is not a significant variable, reinforced by a confidence interval that spans 0 (-3.355 36.230).

5 Reflection - UPDATED

5.1 Issues with the dataset and the analysis

5.1.1. Issue with dataset

Limited date-range: The data was only for May 2011, and ridership in May is probably not representative of ridership patterns throughout the year. For example, in the springtime people may be more eager to walk, even if it's raining. Perhaps not so much in February.

Multiple weather variables could lead to collinearity: For example, `tempi`, `mintempi`, `meantempi` and `maxtempi` values must by definition be highly correlated. Since we are measuring *hourly* ridership, it would be more logical to only include `tempi` since it is the only temperature variable for that same time of day at which as the `ENTRIESn_hourly` value.

As another example, since barometric pressure changes is used to forecast changes in weather, `pressurei` will likely be correlated with `precipi` and `wspdi`. Since the question is whether rain impacts ridership, `precipi` is the most relevant variable of the three.

Ridership & rain are measured at different intervals: Ridership spikes at certain times during the day but we do not know when exactly the rain fell during the day, yet we are examining rain's impact on *hourly* ridership.

There is a higher frequency of 0 values on non-rainy than rainy days: There 897 rows in total where `ENTRIESn_hourly` is 0. This occurs at various times of the day. Were these turnstiles broken? Or simply reset? If the latter this will understate ridership.

Of these 897 "0" rows

- 764 belonged to the non-rain sample. The total non-rain sample size is 33064.
    - 764/33064 = 2.31% of non-rain ENTRIESn_hourly have a value of 0.

- 133 belonged to the rain sample. The total sample size for rain is 9585.
    - 133/9585 = 1.39% of rain ENTRIESn_hourly have a value of 0.

5.1.2. issues with analysis or statistical interpretation

Statistical interpretation

There are limitations to the Mann-Whitney U-test for accepting or rejecting a null hypothesis. It only indicates whether there is a greater chance of obtaining a higher value from one sample over another.

The rain group was much smaller compared to non-rain (9585 vs. 33064 observations). While both groups had many large outliers for ENTRIESn_hourly, I wonder if these outliers had much more weight in the smaller rain group and are the reason behind the higher mean.

In the interest of re-submitting this project sooner rather than later, I'll need to leave this question unanswered.

The other question, which is unanswerable, is what did these outliers represent: legitimate exceptions, or errors?

Model analysis

As discussed in section 2, this model had a low R^2 value of .54, meaning barely over half of the predicted values can be explained by the independent variables.

Moreover, plotting residuals against the independant & dependant variables repeatedly showed a non-constant variance along the x-axis. In other words, the errors in the model do not appear to be random ones. Meanwhile, a histogram and boxplot of residuals revealed many outliers.

All of this calls into doubt the validity of the variables chosen, or with the linear model itself.

Considering that this model uses more dummy categorical variables than continuous numeric variables, I suspect it would make more sense to use a logistic regression rather than a linear one.