Section 1. Statistical Test

We can’t use “Welch t-test” for compare average of the two groups “rainy days” and “non-rainy days”, because neither the rain or no-rain histograms are normally-distributed. As such, a non-parametric test such as Mann-Whitney U is a good fit, while a test such as Welch’s two-sample t-test is not.

So we will use Mann-Whitney rank test (two-tailed) on samples “rainy days” and “non-rainy days”. We are using this test to determine if the two populations (“rainy days” and “non-rainy days”) have equal means based on the sample means.

H_0: Null hypothesis: muRain - muNoRain = 0

No significant difference in ridership on rainy vs non-rainy days.

H_1: Alternative hypothesis: muRain - muNoRain != 0

There is a statistically significant difference in ridership on rainy vs non-rainy days.

Metric	Value
Rain mean ridership	1105
No rain mean ridership	1090
Difference between means	15
P-value	**0.04999982558697944**

On average, not big difference between people who do ride the NYC subway on rainy days vs non rainy days. We multiplied p-value (0.024999912793489721) by two in Mann Whitney U test, because scipy returns p-value for one-sided test. The p-value = 0.04999982558697944. P-value is less than 0.05 - our significant level, so we should reject the null hypothesis (No significant difference in ridership on rainy vs non-rainy days.).

Probably in shiny days people more walk on the street to work enjoying weather. But in rainy days it is not comfortable.

Section 2. Linear Regression

I used Gradient descent approach to compute and produce predictions for ENTRIESn_hourly.

Features for model I decide try in model:
'rain', 'mintempi', Hour', ‘**fog**’
dummy variables - **UNIT**

‘UNIT’ was used as dummy variables because larger trafficked units would be expected to have more riders as opposed to smaller units.

‘Hour’ variable affects the most. Hour was chosen because it was highly linearly correlated with ENTRIESn_hourly. The R^2 based on ‘Hour’ and dummy variable was **0.46315**, correlation between “**Hour**” and “**ENTRIESn_hourly**” was **0.17799587** (by function np.corrcoef). And because ridership would vary based on the time of day as more people would probably ride the subway in rush hours (7 to 9:30 and 16 to 18). The theta value is **4.63508402e+02.**

‘rain’ were chosen based on the assumption that when bad weather has affect people’s trip behaviour. Adding this feature increased R^2 to **0.46322**. Correlation between “**rain**” and “ENTRIESn_hourly” was **0.01081.** The theta value is **-1.42740458e+01.**

‘**fog**’ were chosen based on the assumption that when bad weather has affect people’s trip behaviour. Adding this feature increased R^2 to **0.46356**. Correlation between “**fog**” and “ENTRIESn_hourly” was **0.02658.** The theta value is **5.92440705e+01.**

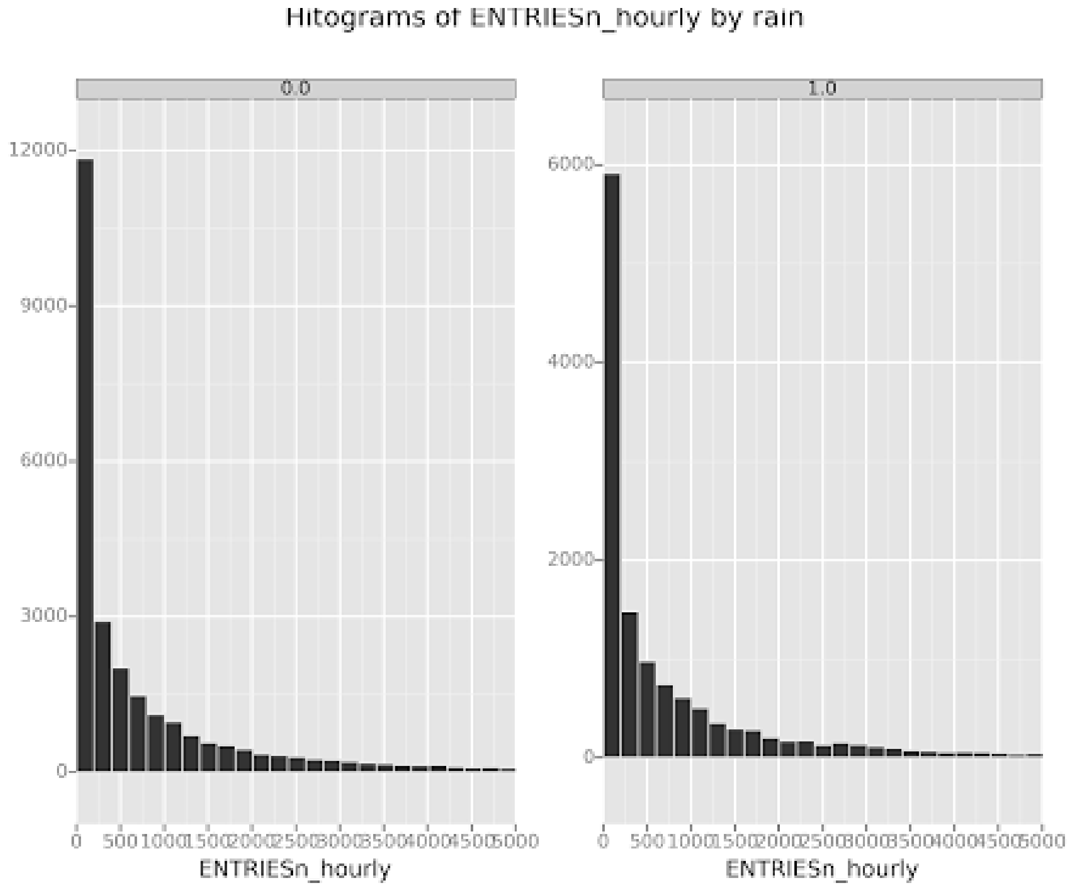
“mintempi” were chosen based on the assumption that when bad weather came with traffic jams and more people prefer the metro instead of walking under rain or when weather cold. Adding this feature increased R^2 to **0.46486**. Correlation between “**mintempi**” and “ENTRIESn_hourly” was **0.02979.** The theta value is **-6.14152407e+01.**

Model: The chosen features were **‘UNIT’** and **‘Hour’.**

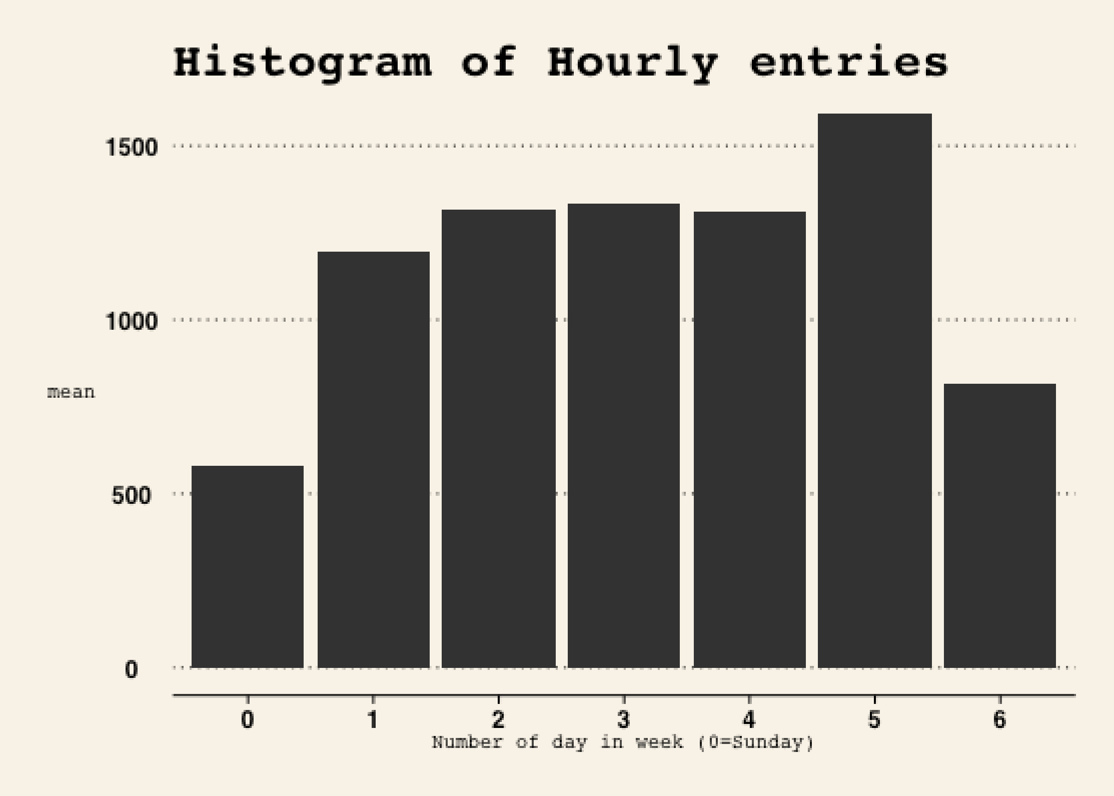
The coefficients (or weights) of the non-dummy features (Hour) in linear regression model is 4.63382098e+02.

Typical range for R^2 is between 0 and 1. Our final R^2 value is **0.4631**. We have an R-square of 0.4631 then we know that the variability of the Y values around the regression line is 1-0.4631 times the original variance. In other words we have explained ~46% of the original variability, and are left with ~54% residual variability. This does not represent a strong linear relationship between the dependent variables and the independent variable; the **linear model is not appropriate for this data set**.

Section 3. Visualization



Both histograms have similar shape of the distribution for rain and non-rainy days (with long tail skewed to the right). We see that our data have much more information for non rain days probably because was collected on non rainy days (we not have 50%/50% rain/non-rain days in month), so we can't do conclusion about big difference between rainy and non-rainy days.



This second visualization was created using R language (RStudio) with ggplot2, ggthemes, dplyr packages based on file turnstile_data_master_with_weather.csv. It shows the daily median 'ENTRIESn_hourly'. Some interesting insights from this visualization are:

1) As expected working days more loaded (Mon-Fri).
2) Friday tho most loaded day. Probably a lot of people happy finish working week and go to home and after that go to find some fun.

Section 4. Conclusion

As I found by analysis of the data, more people on average do ride the subway when it is raining vs when it is not raining. That is approve by the Mann-Whitney U Test results which show that the 15 difference in averages and that is statistically significant. But difference is not enough to say there is any usefull difference to predict higher or lower ridership based on the presence of rain. Supporting this conclusion, that adding rain to the predictive model not increased R2.

Comparative histograms show the distributions are nearly identical in shape except for the y-axis height (frequency) which can be attributed to the fact that there were not equal rainy/non rainy days in May month.

It is obvious that the ridership changes over the day (Hour), so probably for predict ridership more likely will be better polynomial model.

Section 5. Reflection

I really enjoyed working with these data, it was a great challenge. I live in the city of Krasnodar, Russia and work with real data of NYC subway for me is very inspiring. However will be great have data for whole year or for seasons where weather can have significant effect on riderships as example autumn or winter.

I am sure that the model (linear regression) will not perform well with data from other months, because as I wrote above in autumn or winter riderships can have different behaviour. May be random days of year for the sample can improve this.

References

1. **Intro to Descriptive Statistics** https://www.udacity.com/course/ud827
2. Introduction to Data Science with R http://shop.oreilly.com/product/0636920034834.do
3. Welch's t test http://en.wikipedia.org/wiki/Welch's_t_test
4. “One-tail vs. two-tail P values” http://graphpad.com/guides/prism/6/statistics/index.htm?one-tail_vs__two-tail_p_values.htm
5. “Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit?” http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit