Scott Burns
Udacity – Data Analyst Nanodegree
Project 1: Analyzing the NYC Subway Dataset

Section 0: References

Section 1	Intro to Data Science – Udacity Course, Cross Validated: Post 9573 Scipy Shapiro-Wilk Test documentation, "Is normality testing 'essentially useless'?" – Cross Validated
Section 2	Pandas/Statsmodel/Scikits-learn: Cross Validated – Post 47913, Intro to Data Science Problem Set 3, Stack Overflow – how to print estimated coefficients…., Scikit Learn – Linear Regression Example, StackOverflow: python: get list from pandas dataframe, Linear Regression with Python, StackOverflow: Run an OLS Regression with Pandas Dataframe, Statsmodels documentation: Ordinary Least Squares, StackOverflow: Why do I get only one parameter from a statsmodels OLS fit?
Section 3	Yhat ggplot documentation, Pandas Pivot Table documentation, StackOverflow: Convert Series of strings to float in Python
Section 4	Intro to Data Science – Udacity Course
Section 5	Intro to Data Science – Udacity Course

Section 1: Statistical Test

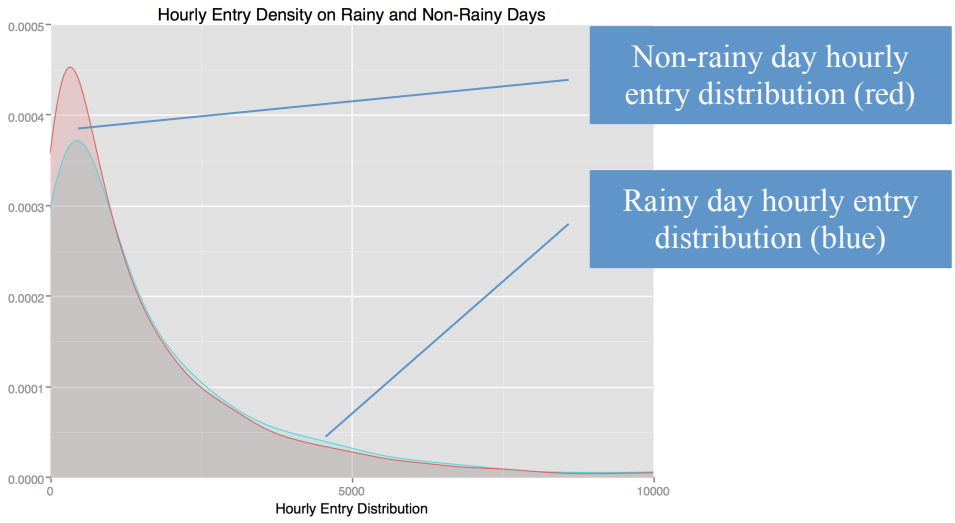
1.1

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

I used a Mann-Whitney U-test to assess whether the evidence suggests that we can reject the null hypothesis that rainfall during a day isn't associated with an increase in the number of people riding the New York City subway.

I chose this test because hourly ridership on rainy and non-rainy days does not appear to follow normal distributions, as demonstrated visually in the probability density chart below.

Figure 1: Overlaid Hourly Entry Distribution for the New York City Subway System (rainy / non-rainy days)



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

I used a one-tail P value to assess whether we can reject the null hypothesis, because the question posed for the project is "*figure out if more people ride the subway when it is raining versus when it is not raining".* This implies that we must test whether the sample distribution of ridership is shifted higher on rainy days than non-rainy days, not just whether it seems to be different.

1.1.3 What is the null hypothesis?

The null hypothesis I adopted was that rainfall is not associated with subway ridership being larger.

1.1.4 What is your p-critical value?

To test the null hypothesis, I used a p-critical value of 0.025, because this is a fairly common benchmark in statistics for assessing whether evidence is sufficient to reject a null hypothesis in one-tailed tests, and I feel comfortable with the risk of Type I errors implied by this threshold.

1.2

1.2.1 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 test is more appropriate to this case than a standard Welch's t-test, because the sample distributions are clearly not normal, and the validity of a Welch's test depends on the assumption that the samples are normally distributed.

While I understand from background reading that normality testing is not useful for large datasets, running Shapiro-Wilk tests on both samples suggest that they are also not normally distributed, as my test statistics yield p-values under 1%. However, it seems that for most large samples, the test also yields statistics that support rejecting the null hypothesis of normal distribution.

Based on visual assessments (as shown in Figure 1), and the results of Shapiro-Wilk tests, I determined that it would be more prudent to test the null hypothesis that rain doesn't affect hourly entries using a Mann-Whitney test, which doesn't rely on an assumption of normally distributed samples.

1.3

1.3.1 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.

Table 1: Summary Statistics for Entries at New York City Subway Stations

	Non-rainy days ('Rain' = 0)	Rainy days ('Rain' = 1)
Mean	1,846	2,028
Median	893	939
Standard error	2,879	2,028
Sample size	33,064	9,585
u-statistic	153,635,121	
p-value (one-tailed)	2.74E-06	

1.4

1.4.1 What is the significance and interpretation of these results

If the null hypothesis were true, then we would expect to see a u-statistic of at least this magnitude in approximately ~3 cases out of 1,000,000 (p-value: ~2.7e-6). Given that we set a p-critical value of 2.5% as our threshold, under this criterion we can reject the null hypothesis that rain in a day has no effect on subway ridership

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?*

To determine coefficients and predict ENTRIESn_hourly I used Scikit Learn's Stochastic Gradient Descent function, as well as the Statsmodels OLS function to 'sanity check' my Scikit Learn results and produce a summary table with coefficient values and t-scores for each coefficient (I didn't see similar functionality in Scikit Learn).

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

For the regression estimate, I used the following features in my model, with Category Dummies created for each instance of the second column elements in Table 2

Table 2: Features used in linear regression model

Features	Category Dummies
rain	UNIT
precipi	Hour
meantempi	conds
meanwspdi	
weekday	

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.

I selected the features listed in Table 2 because I could intuitively justify that each might have a significant influence on ridership. Other potential features available in the dataset seemed to me as having a more tenuous theoretical association with 'ENTRIESn_hourly', or would be already sufficiently proxied by the included features – e.g. meantempi for maxtempi to be left out of the final specification.

The resulting model R-squared of .52 (as calculated in Scikit Learn) suggested that this specification was substantially explanatory, and about 8 percentage points above the linear regressions on the same data I performed as part of a similar exercise during the *'Intro to Data Science'* Udacity course. Based on this, I feel comfortable that this is a reasonably good model to predict ridership.

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

A summary table of model coefficients (excluding 'UNIT' dummy variables) as calculated with StatsModels OLS is below in Table 4. Preceding that, in Table 4 is a table of coefficient values as calculated with my latest run with Scikit Learn.

Table 3: Coefficient Values as calculated with Scikit Learn Stochastic Gradient Descent function

Feature	Coefficient
rain	90.91
precipi	-1,910.25
meantempi	-25.27
meanwspdi	-9.12
weekday	883.58
hour_0	-403.19
hour_4	-1,510.14
hour_8	-1,203.25
hour_12	1,196.92
hour_16	393.39

Feature	Coefficient
hour_20	1,341.98
conds_Clear	90.32
conds_Fog	4,118.67
conds_Haze	324.26
conds_Heavy Rain	639.34
conds_Light Drizzle	-419.27
conds_Light Rain	264.93
conds_Mist	1,621.93
conds_Mostly Cloudy	-214.53
conds_Overcast	-145.54
conds_Partly Cloudy	-138.81
conds_Rain	292.02
conds_Scattered Clouds	137.71

Note that the coefficients differ slightly, but non-dummy coefficients largely share the same signs and magnitudes where coefficients are statistically significant (P-critical < 0.05). The R-squared for the regression model calculated with Scikit Learn is also slightly lower than that for the model calculated with StatsModels OLS (0.52 vs 0.55), which I anticipated, as the Gradient Descent algorithm may stop at a solution somewhat short of an optimal value.

Table 4: Summary statistics for model calculated by StatsModels OLS function

OLS Regression Results

Dep.Variable:	ENTRIESn_hourly	R-squared:	0.545
Model:	OLS	Adj. R-squared:	0.542
Method:	Least Squares	F-statistic:	195.1
		Prob (F-statistic):	0.00
Time:	10:31:05	Log-Likelihood:	-3.85E+05
No. Observations:	42649	AIC:	7.70E+05
Df Residuals:	42388	BIC:	7.72E+05
Df Model:	260		

	coef	std err	t	P>|t|	[95.0%	Conf. Int]
const	2,182.69	96.72	22.57	0.00	1,993.12	2,372.25
rain	47.65	30.58	1.56	0.12	-12.28	107.59
precipi	-2,795.40	760.08	-3.68	0.00	-4,285.16	-1,305.64
meantempi	-21.29	1.63	-13.06	0.00	-24.48	-18.09
meanwspdi	-8.03	4.30	-1.87	0.06	-16.45	0.39
weekday	996.68	22.66	43.99	0.00	952.27	1,041.08
hour_0	-39.26	27.17	-1.45	0.15	-92.52	14.00
hour_4	-1,129.66	27.13	-41.63	0.00	-1,182.85	-1,076.48
hour_8	-873.92	28.36	-30.82	0.00	-929.49	-818.34
hour_12	1,571.56	27.14	57.91	0.00	1,518.37	1,624.75
hour_16	861.93	26.80	32.16	0.00	809.40	914.47
hour_20	1,792.03	27.27	65.71	0.00	1,738.58	1,845.49
conds_Clear	61.71	51.57	1.20	0.23	-39.38	162.79
conds_Fog	1,097.11	273.64	4.01	0.00	560.77	1,633.45
conds_Haze	274.74	71.19	3.86	0.00	135.20	414.28
conds_Heavy Rain	174.59	135.25	1.29	0.20	-90.50	439.67
conds_Light Drizzle	-727.61	110.77	-6.57	0.00	-944.73	-510.49
conds_Light Rain	72.67	63.77	1.14	0.25	-52.31	197.66
conds_Mist	1,501.34	379.65	3.96	0.00	757.21	2,245.46
conds_Mostly Cloudy	-308.26	50.88	-6.06	0.00	-407.99	-208.53
conds_Overcast	-197.95	48.80	-4.06	0.00	-293.60	-102.30
conds_Partly Cloudy	33.08	61.29	0.54	0.59	-87.04	153.20

2.5 What is your model's R2 (coefficients of determination) value?

The model's R-squared value is 0.55 as calculated using StatsModels OLS function and 0.52 as calculated using Scikit SGD Regressor function.

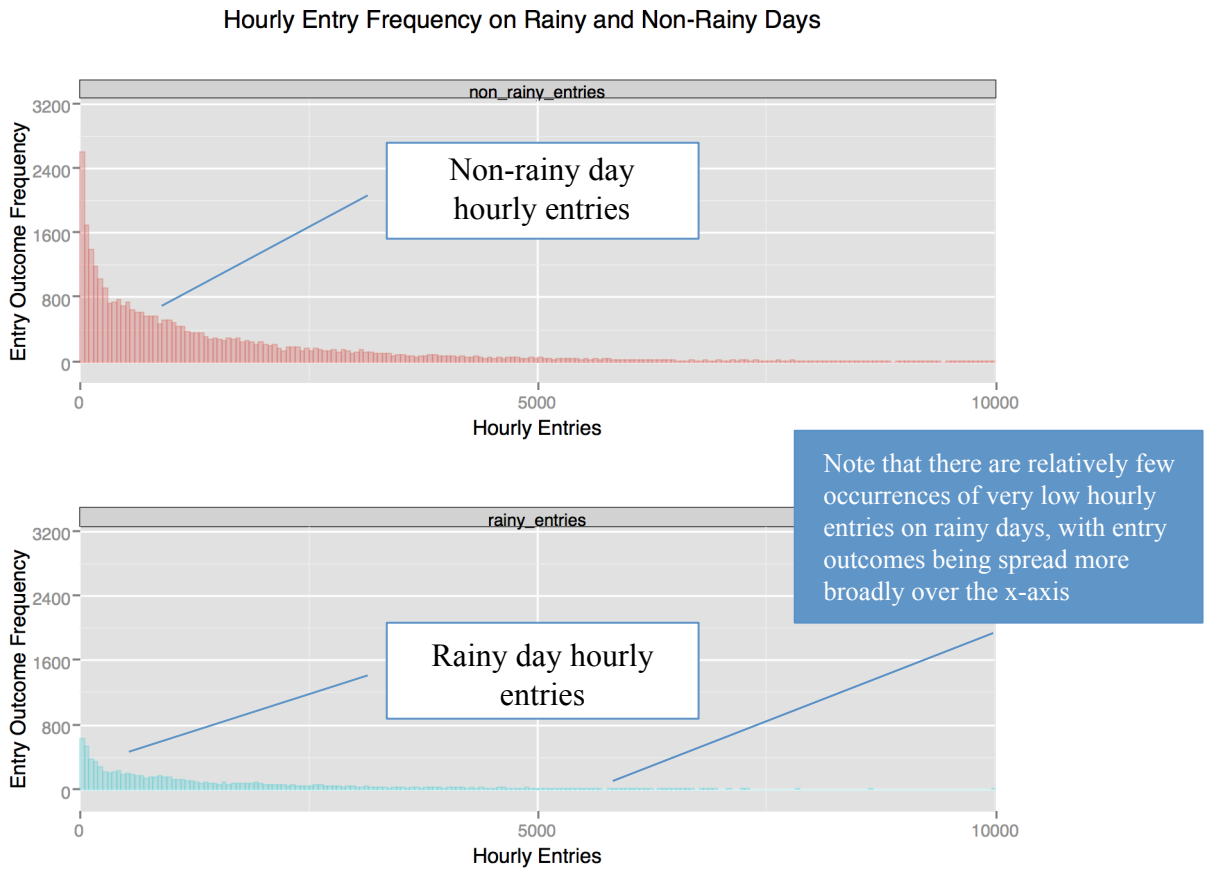
2.6 What does this R2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R2 value?

This R-squared value suggests that the features used in the model can explain up to 55% of variation in subway ridership. Based on experience with linear regression in other contexts, this seems like a strongly predictive model, and one that is appropriate for predicting ridership.

Section 3: Visualization

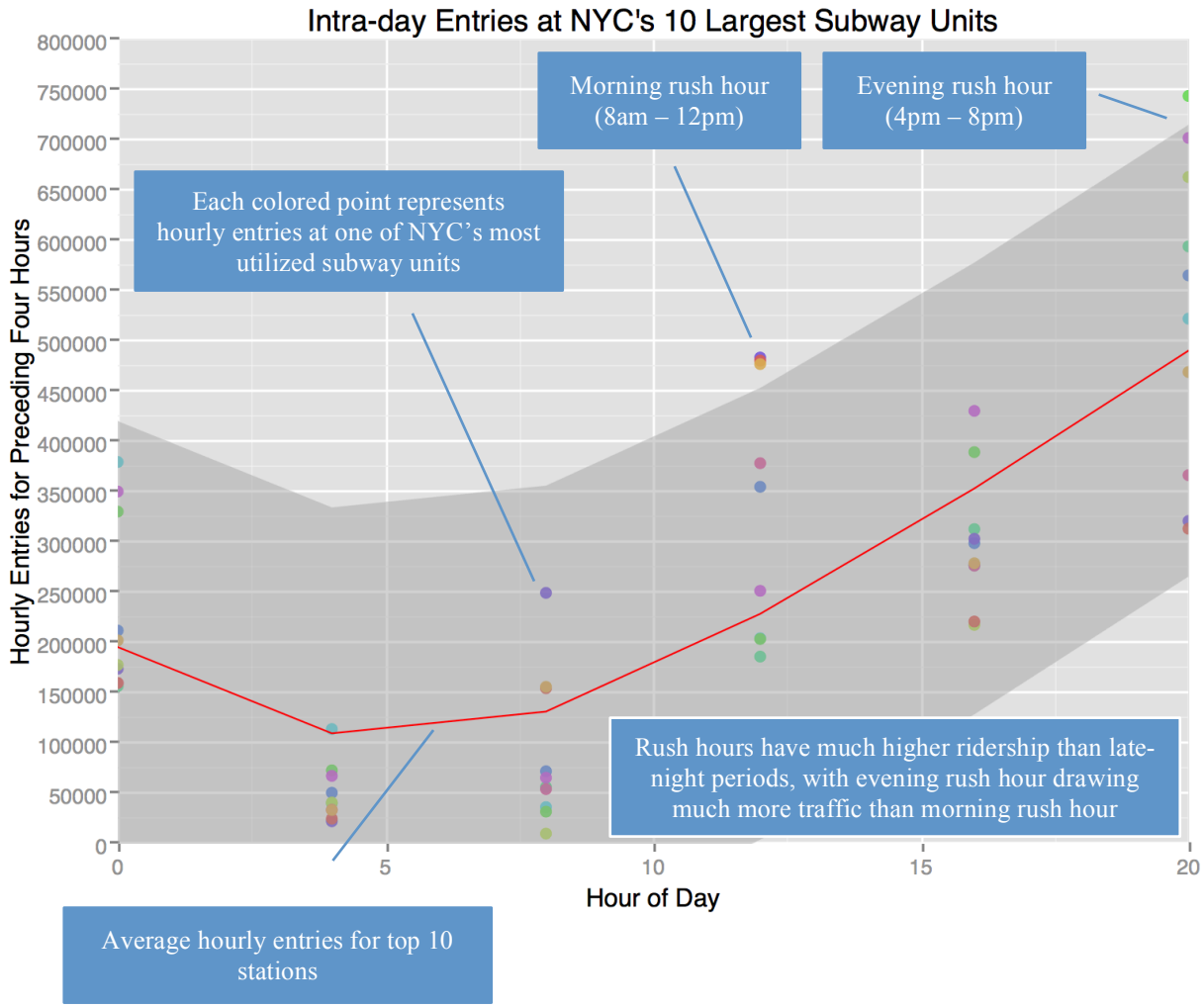
4.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

Figure 2: Hourly histogram for the New York City subway system (rainy / non-rainy days)



4.2 *One visualization can be more freeform. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots) or attempt to implement something more advanced if you'd like.*

Figure 3 Hourly Entries at New York City Subway Stations (by Station / Unit)



Section 4: Conclusion

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?*

Based on my analysis in Sections 1 to 3, and my interpretation of the data, I think we can say that more people ride the subway when it is rainy. Findings from our investigation in

Section 1 suggest that we can reject the null hypothesis that rainy weather is not associated with increased entries in the subway.

4.2 What analyses lead you to this conclusion? You should use results from both your statistical

I base my answer in question 4.1 on the following. As we saw in Section 1, average and median levels of hourly ridership per subway unit on rainy days were 2,028 and 939 versus 1,846 and 893, respectively, for days that were not rainy. Furthermore, a Mann-Whitney test of the distributions from the two samples indicates that if rain didn't affect ridership, the differences between the samples seen in our overview would only be observed about 3 times out of a million.

This evidence is sufficient to allow us to reject the null hypothesis that ridership doesn't increase on rainy days. Instead I would feel confident hypothesizing that ridership on rainy days is typically higher than on non rainy days.

Findings in Section 2 provide further evidence that rainy-type weather is positively associated with subway ridership.

In this section, we constructed a model to predict subway entries that seemed to provide substantial explanatory power, with an R-squared value of 0.54 and F-test statistic P-value of under 0.005.

Several dummy coefficients related to rainy weather in the model were large in magnitude, positively signed, and highly statistically significant, including 'Fog' and 'Mist' conditions, where the presence of each was associated with increases in hourly ridership of 1,097 and 1,501 per Unit, respectively and both were highly statistically significant (with t-stats showing P>|t| under 0.005).

Strangely, our precipitation feature was negatively signed, relatively large and highly statistically significant, seeming to suggest that heavier rainfall on rainy days may actually lead to decreasing ridership, all other features being equal (or that some rainy variables are multicollinear). While in general ridership falls on rainy days, the relationship between weather and hourly entries could be more complex than I first expected.

Section 5: Reflection

5.1 *Please discuss potential shortcomings of the methods of your analysis*

A major shortcoming of my analysis may be my inclusion of many variables related rainy weather in the regression model built for Section 2. This may have created

multicollinearity that hinders our ability to correctly interpret the impact of each included feature on subway ridership. In the model I included the 'rain' dummy variable, as well as 'precipi' and the conditions dummies, several of which are correlated with rain. Even though this is the case, I see how each of the variables could uniquely impact ridership and do not wish to remove them from the model.

Another potential issue in the dataset is that other factors related to rain and more directly associated with changes in subway ridership may have been omitted, potentially making my interpretations in the previous sections somewhat invalid. For example, it may be the case that if another variable, such as an indicator of traffic density around a UNIT, were to be included in the dataset, the significance of the rain-related features would diminish. This could happen if rain itself wasn't associated with lower ridership, but through its impact on traffic gridlock was causing more people not to drive, but ride the subway.

Omitted variables could have some impact on the interpretation of our model in Section 2, but shouldn't affect the conclusion from Section 1 that rain does seem to be associated with higher subway ridership, though the causal channel for this is not clear.