Student and Class Information

Dorothy Kessler
db4871@us.att.com

All Questions and quizzes were completed by Dorothy Kessler. My document is: DorothyKessler.UD359.FinalProject.pdf

Udacity Intro to Data Science, UD359: enrolled 1/27/15, coupon UMCCHJT8ARDHIFN
https://www.udacity.com/course/ud359

Final Project: Analyze NY Subway System; 4/8/15
I used the class interface; and the Anaconda 2.1.0 install of python 2.7 with the Spyder 2.3.1 IDE. All python code is available upon request.

Analyzing the NYC Subway Dataset

Data Set

For my exploratory analysis, I used turnstile_data_master_with_weather.csv, a regular csv file containing subway data from May 2011; with the data being read into a pandas DataFrame. I continued with that data set for my final Project: https://www.dropbox.com/s/meyki2wl9xfa7yk/turnstile_data_master_with_weather.csv

131,951 rows; columns A through V inclusive
44,104 with rain
87,847 without rain

Statistical Test to analyze NYC subway data

I looked at the data to examine the hourly entries in our NYC subway data to determine what distribution the data follows. I plotted two histograms to show hourly entries when raining vs not raining.

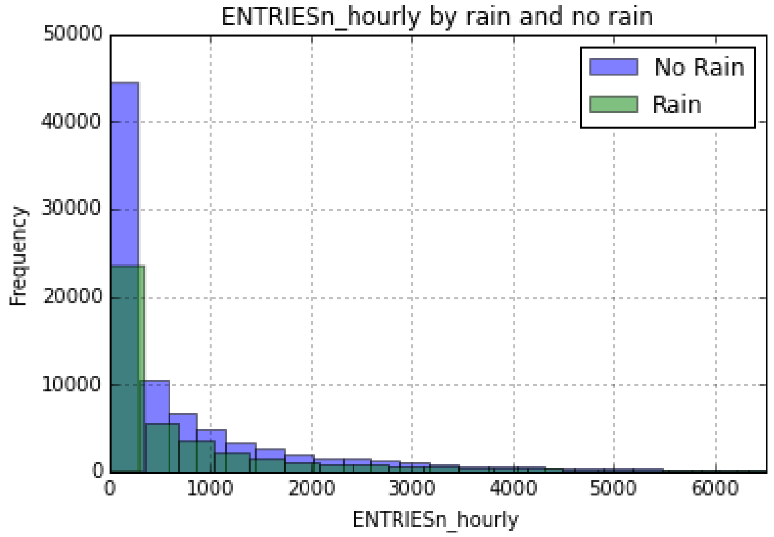


Figure 1: HISTOGRAM hourly entries rain and no rain

This graph indicates ridership is higher on days with no rain, however, in this data for May 2011, we have almost twice as many days in the month without rain as with rain. This graph also indicates the data does not follow a normal distribution.

Mann Whitney Wilcoxon U test

The null hypothesis

The null hypothesis is that distributions are identical; there is a 50% probability that an observation from a value randomly selected from a rainy day exceeds an observation randomly selected from a non-rainy day.

Conventional wisdom to formulate the Null Hypothesis H_0

H_0 = The ridership in the NY subway is not likely to be different on rainy days and non-rainy days. (no effect)

Alternative Hypotheses H_a

H_a = Ridership in the NY subway is impacted by rainy days (an effect)

2-sided alternative includes both > and < H_0 value

My p-critical value

My p-critical value is the standard value <= 0.05

α (significance level) conventionally at 0.05

5% change of observing a result as extreme

reject the null hypothesis even if p is 0.04999999

I ran Shapiro-Wilk Test to test if data was drawn from a normal distribution. This gave a warning that the p-value may not be accurate. So, I performed scipy normaltest which calculates a p-value that if low enough indicates a low chance that the distribution is normal. I concluded that the data was non-normal, so I could not use the Welch T-test.

I used the nonparametric, Mann-Whitney Wilcoxon U test. Non parametric tests apply to observations that are difficult to quantify when you have 2 independent samples. The p-critical value for my test was **0.05**.

Two-tail P value

Two unpaired, independent groups; Samples are large

Data measured on a continuous scale, every value is unique

"Distribution free", does not rely on mathematical normal distribution

A two-tailed test assigns half of the alpha to testing the significance in one direction and half of the alpha to testing significance in the other direction. So using the standard value of 0.05, this means .025 is in each tail of the distribution. I am testing for the possibility of the relationship in both directions.

```
if( p_value * 2 ) <= p-critical
   Reject the null hypothesis
```

Mann Whitney returns a 1-sided p value, and my test calculated.

```
0.019309634413792565
```

To get a two-sided p-value, multiply the returned p-value by 2

```
0.03861926882758513
```

np.mean with rain	np mean no rain	U	2-sided p
1105.4463767458733	1090.278780151855	1924409167.0	0.03861926882758513

The significance and interpretation of these results

If the P value is less than or equal to the p critical value, you can reject the null hypothesis. The smaller the P-value, the stronger the evidence against H_0. Because my two sided p_value is less than 0.05, the null hypothesis is rejected and I accept the Alternative Hypotheses H_a that ridership variations on rainy days are not due to chance. We have 95% confidence that the two populations are different.

Section 2. Linear Regression, Predicting the Ridership, Machine Learning

2.1 Approach to compute the coefficients theta and produce prediction for ENTRIESn_hourly

Using linear regression to predict how many total riders the New York City subway will have on a given day with a given a variety of factors. A linear relationship is used to predict the numerical value of *Y* for a given value of *X* using a straight line (*regression line*).If the slope of the line is known, and if we know the *y* intercept of that line, then we can put a value for *X* and predict the value for *Y*. In most cases, Y is the variable we are trying to predict. So we are trying to predict the number on entries per hour, our Y; using variables from our data set X.

The linear regression model is meant to predict ridership trends, not just when it rains, but in general.

2.2 Gradient descent: to compute coefficients theta used for the predictions

We are going to multiply features for X, input variables X_1 to X, by a set of coefficients, theta1 through thetaN. We do this to weight our model.

y = A*x_1 + B*x_2 + C*x_3 + ...

y is the **dependent** variable, the subway ridership estimate

x_1, x_2, and x_3 are the **independent** variables or features

A, B, and C are the coefficients

I tried a variety of features, including more weather related values. When I added maxpressurei my $\mathbf{R}^2$ value became the highest. The features I choose were weather related, time of day hour and ridership.

```
[['meantempi', 'Hour', 'rain', 'mintempi', 'maxtempi', 'meanwindspdi',
  'maxpressurei', 'EXITSn_hourly']]
```

Also, in my predictions function, I setup dummy Units for 465 columns, categorical data ’R001’ to ‘R552’, (no data for R026) which cannot be used in a mathematical formula.

My feature set is 131951 by 474 columns, 8 feature columns, 1 column for y intercept and 465 columns for the dummies units. Theta, one value per column, length of theta is 474. We use theta to understand how each individual feature affects our model. These are the theta values:

meantempi	**Hour**	**rain**	**mintempi**
-7.30923836	237.55016131	4.41460986	-46.35184054

maxtempi	**meanwindspdi**	**maxpressurei**	**EXITSn_hourly**
30.90659218	22.61111969	-23.22130564	1270.3962113

If the coefficient is positive, the dependent variable, ridership, increases as the independent variable, the feature, increases.

If the coefficient is negative, the dependent variable, ridership will decrease as the independent variable, the feature, increases.

You cannot say over time for these variables because hour is the only feature that moves with time. Only one feature is dependent on time. It is better to keep all other features constant so it is clear how each feature influences the estimate. To do a fair comparison of how rain influences ridership trends, you have to keep the other features constant.

2.3 Cost Function

Values of theta map to a single value. We perform gradient descent given data with an arbitrary set of features. My cost function is to measure how our thetas are doing at modeling the data. To minimize cost, I start with theta 0, and performed a gradient descent 75 times. We can see the cost is going down with each iteration, so we know we are heading toward minimum.

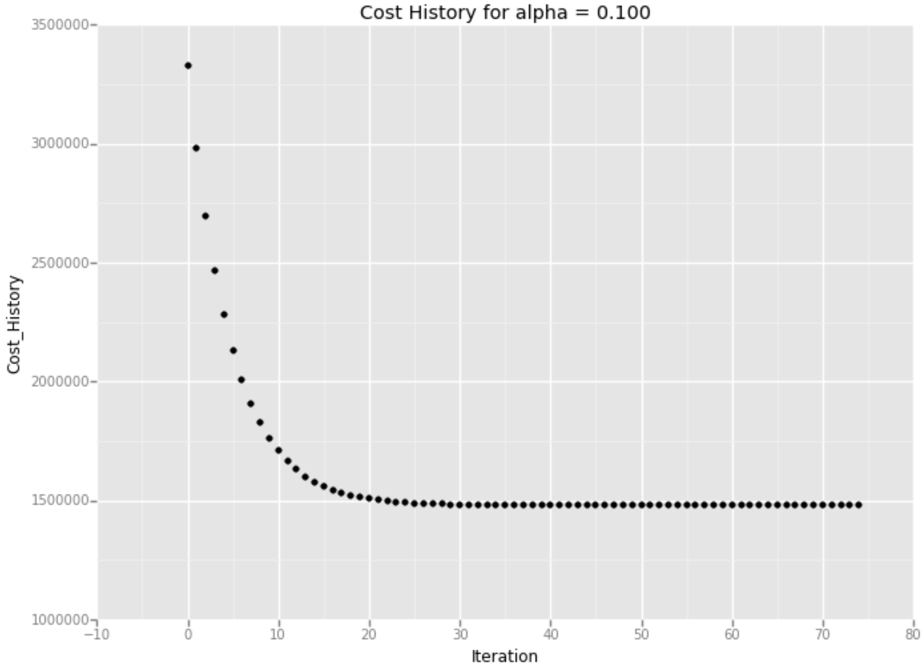


Figure 2: cost history linear regression with gradient descent, 75 iterations, showing decreasing cost

2.4 Scatter Plot

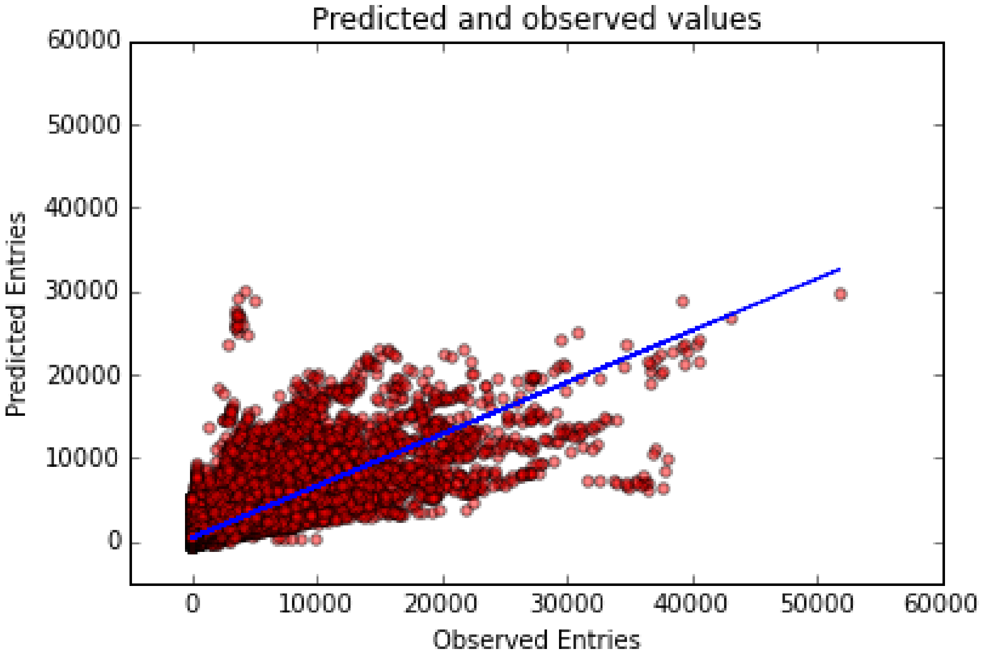


Figure 3: scatter plot, entries per hour actual and predicted. I created a plot with the predictor dependent variables on y, and the observed independent on x. I added alpha 0.5 to help get a sense of density. Shows a positive relationship (in terms of direction), both variables increase together.

2.4 Residuals

The differences between the predicted and the actual values. If you have a good model, you expect to see the residuals as a normal distribution.

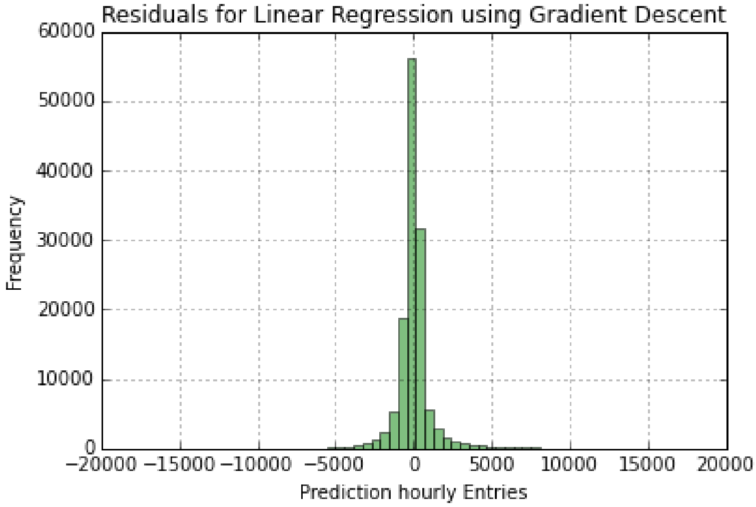
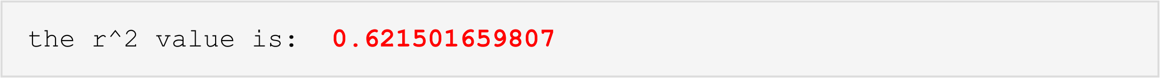


Figure 4: Residuals histogram for linear regression with gradient descent model.

Randomness and unpredictability are important parts of a regression model; we expect random errors to produce residuals that are normally distributed. Ideally, we want the residuals mean to be close to zero, mine was at 0.405266323897

2.5 My model's R^2 (coefficients of determination) value

The closer **R^2** is to 1, the better the model. It varies from 0 to 1 and bigger values are better. This gives us a way to evaluate how good our model is.



```
the r^2 value is:  0.621501659807
```

2.6 What does this R^2 value mean for the goodness of fit for the regression model

I believe the **R^2** value is telling me my model is a fair fit for predicting ridership given a variety of weather factors. It tells me the model could explain 62% of the variance and show a positive correlation. I do believe this linear model may be used predict ridership given this **R^2** value.

Section 3. Visualization

3.1 Two histograms: one of ENTRIESn_hourly for rainy days; one of ENTRIESn_hourly for non-rainy days.

Intervals representing volume of ridership (ENTRIESn_hourly) on x-axis and frequency of occurrence on y-axis. For example, each interval (along x-axis), the height of the bar for this interval represents the number of records (rows in data) with ENTRIESn_hourly falling into this interval.

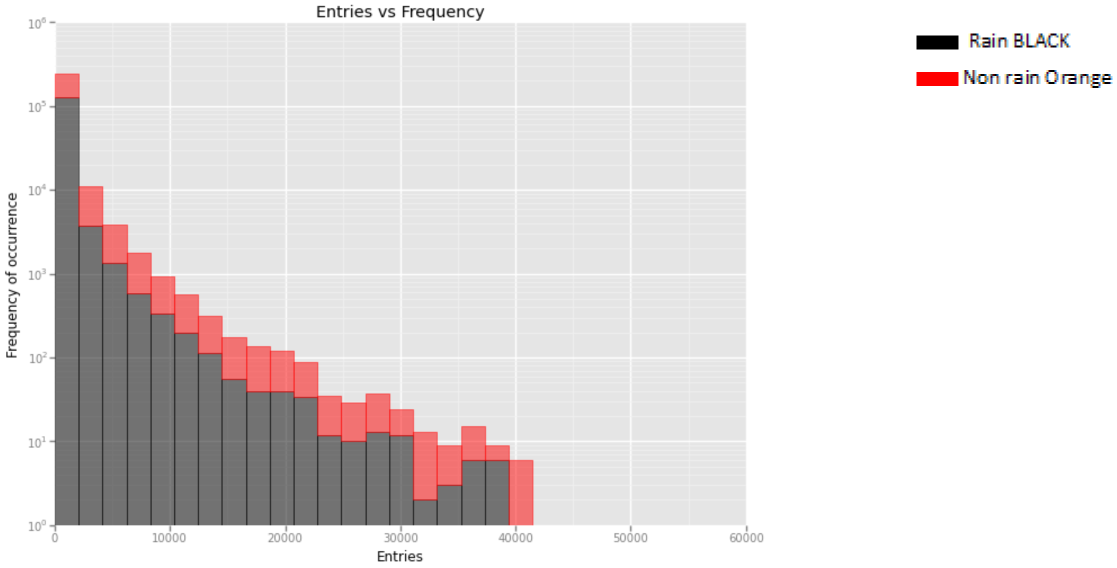


Figure 5: histogram entries on rainy days versus non rainy days

ridership decreases as weather events include rain, but the data sample contains twice as many non-rainy days as rainy days, and this graph does not illustrate a 2 to 1 relationship. Note the data is truncated.

3.2 One freeform visualization:

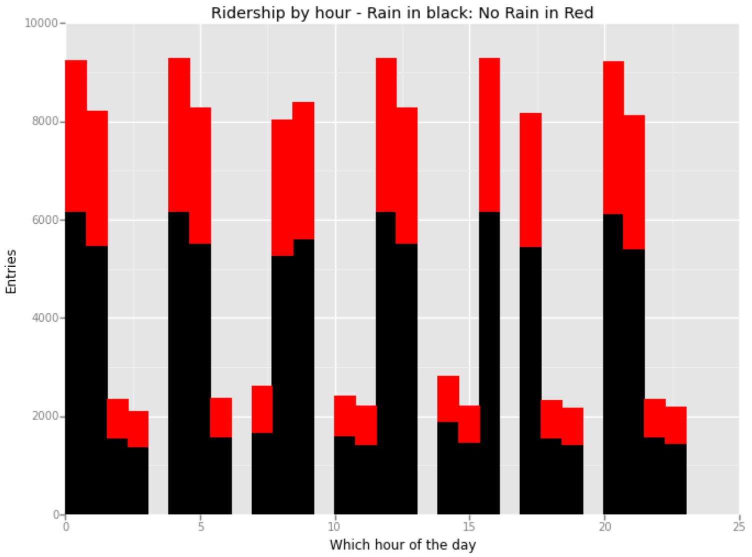


Figure 7: histogram entries on rainy days versus non rainy days:

Rain = black, no-rain = orange. Non normal with similar shapes

Showing ridership by hour of the day, with rain versus no rain. The data is not showing a normal curve, and appears with similar shapes much like the histogram on overall entries. The relationship continues to show the same slight increase for non-rainy days and while we certainly see ridership totals changing based on the hour, the relations between rainy and non-rainy totals does not appear to be affected by which hour it is.

Section 4 Conclusion

I believe more people ride the NYC subway when it is raining.

Looking at the Mann Whitney, and the average data, we see 95% confidence that the populations are different. It is important to note that the Mann-Whitney U-test can only conclude that the null hypothesis was rejected or not rejected. Any comparison of population parameters (ie. The averages) is independent of the statistical test. In this case, the comparison is only meant to qualify how the two populations are different.

The first analysis, plotting histograms of entries with rain versus no rain shows entries by hour is higher without rain. However, there is only one month of data and there were twice as many days without rain in that month. I accept the alternative hypothesis that ridership is impacted by rain.

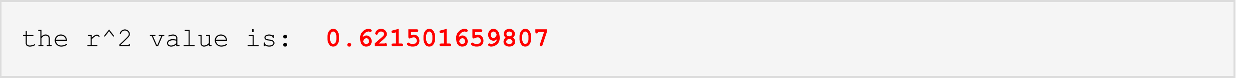
np.mean with rain	np mean no rain	U	2-sided p
1105.4463767458733	1090.278780151855	1924409167.0	0.03861926882758513

The **R**^{2} value for the linear regression was .62 indicating an ok correlation between the features selected and the ridership.

The scatter plot is showing us how the actual entries relate to the predicted entries, they are not related to the linear regression model, and they do not tell us how the features impact ridership. We see an upward trend, positive relationship on the scatter plot. If your prediction values are negative and your actual values are positive, that tells you your regression model is poor. The slope of the scatterplot trend line does not indicate that ridership increases as rain/weather events increase. It does however give a sense of whether or not the estimates tend to be larger or smaller than the actual values. The table below summarizes the possible scenarios (this assumes that the predicted values are not negative).

Slope Predicted vs Measured Value:
1.0 Predicted Value = Measured Value
> 1.0 Predicted Value > Measured Value
< 1.0 Predicted Value < Measured Value

The linear regression model, as detailed in section 2.2, tells us rain is a good predictor of increased ridership, as rain increases, so does ridership. Because rain is 1 or 0 in the data, we know based on the theta values that ridership increase by 4.4 riders hourly when there is rain.



```
the r^2 value is:  0.621501659807
```

Section 5. Reflection

1. **Shortcomings of the Dataset.**
   The datasets came from two different places: the Metropolitan Transportation Authority (MTA), which is responsible for transportation in New York, and the the National Climatic Data Center (NCDC), which is part of the National Oceanic and Atmospheric Administration. The weather data comes from measurements from three points in the city, JFK and LaGuardia airports, and Central Park. The subway data has turnstile information from 465 stations from all 6 Boroughs sampling on a 4 hour basis.

   The MTA Subway Turnstile data reports on cumulative number of entries and exits per row. We know the cumulative entry numbers are calculated as a count of entries since the last reading, i.e. the difference between

ENTRIESn of the current row and the previous row. What if the data in the previous row is NaN? How does the data being given in 4 hour intervals affect outcomes?

Furthermore, we know the subway data for entries is hourly, but it is combined to weather data which is on a daily basis, so for example, in the weather data, if it rained anytime during the day, all hours of the day are marked for rain.

The subway data we analyzed contained only one month of data, May 2011, which is not sufficient, and that month contains a holiday, Memorial Day. Additionally, Mother's Day falls in that time frame; as do baseball games at Yankee Stadium which draws many subway riders.

Of interest, the Daily News reported on 5/25/11, which is in our sample time, that there were many failures on Metro Cards, causing riders to swipe many time for a single admittance.

2. Shortcomings of the analysis

Linear regression looks at linear relationships among data that is dependent affected by independent parameters. One of the shortcomings is I may have a too complex model which becomes an issue if used just to increase the **R**^{2} value. I used an iterative process, choosing different predictor variables. I ran iterations with too few variables and too many, and I found my **R**^{2} increased as I added values leading to a more complex model.It might be a good idea to build the regression model gradually, to ensure that highly correlated variables are not included together. For more information about multicollinearity, including information about the condition number, see the following Wikipedia article: http://en.wikipedia.org/wiki/Multicollinearity.

Additionally, my regression model has features with a high correlation to each other; and furthermore, my inclusion of temperature data may not be a good choice when trying to decide ridership when raining. I ran several more tests, changing my feature set to be more simple, using 'Hour', 'meanwindspdi', 'maxpressurei', 'rain', 'EXITSn_hourly'. This gave me a similar scatter plot and an **R**^{2} value 0.621280086928, also very similar.

Inclusion of the EXITSn_hourly may not have been a good choice. When I removed this value from my feature set, my the **R**^{2} value dropped to 0.458424336672, and my scatter plot took on a more flat look, thus a non-linear model may improve the value of the predictions; and more than one month of data would be a good idea too.

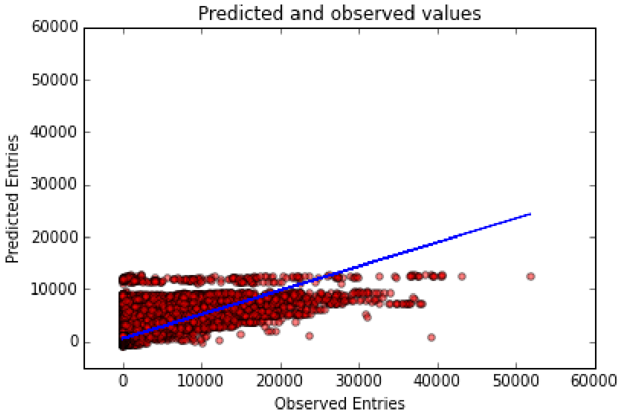


Figure 8: scatterplot removing EXITSn_hourly, using hour, meanwindspdi, maxpressurei and rain

3 (Optional) Do you have any other insight about the dataset that you would like to share with us?

I was interested in the result of entries by station. I did some analysis of ridership per station (not presented here) and noted there was a peak for UNIT R170; and a minimum of 0 from UNIT 447. I wondered where the stations were, and how this ridership affected items like nearby commerce, security, taxi ridership, and so forth.