Analyzing the New York Subway Dataset

Student Notes Project Review

Meets Specifications

Communication

SPECIFICATION
Analysis done using methods learned in the course is explained in a way that would be understandable to a student who has completed the class.

MEETS SPECIFICATION

Reviewer Comments
Vernacular reflects the teachings of the class.

SPECIFICATION
The answers are a well-formed summary of the analyses and do not leave out important information (i.e. fully answering the question).

MEETS SPECIFICATION

Reviewer Comments
I really appreciated the completeness of the project while being concise.

Quality of Visualizations

SPECIFICATION
Plots depict relationships between two or more variables.

MEETS SPECIFICATION

SPECIFICATION
All plots and data are of the appropriate type.

MEETS SPECIFICATION

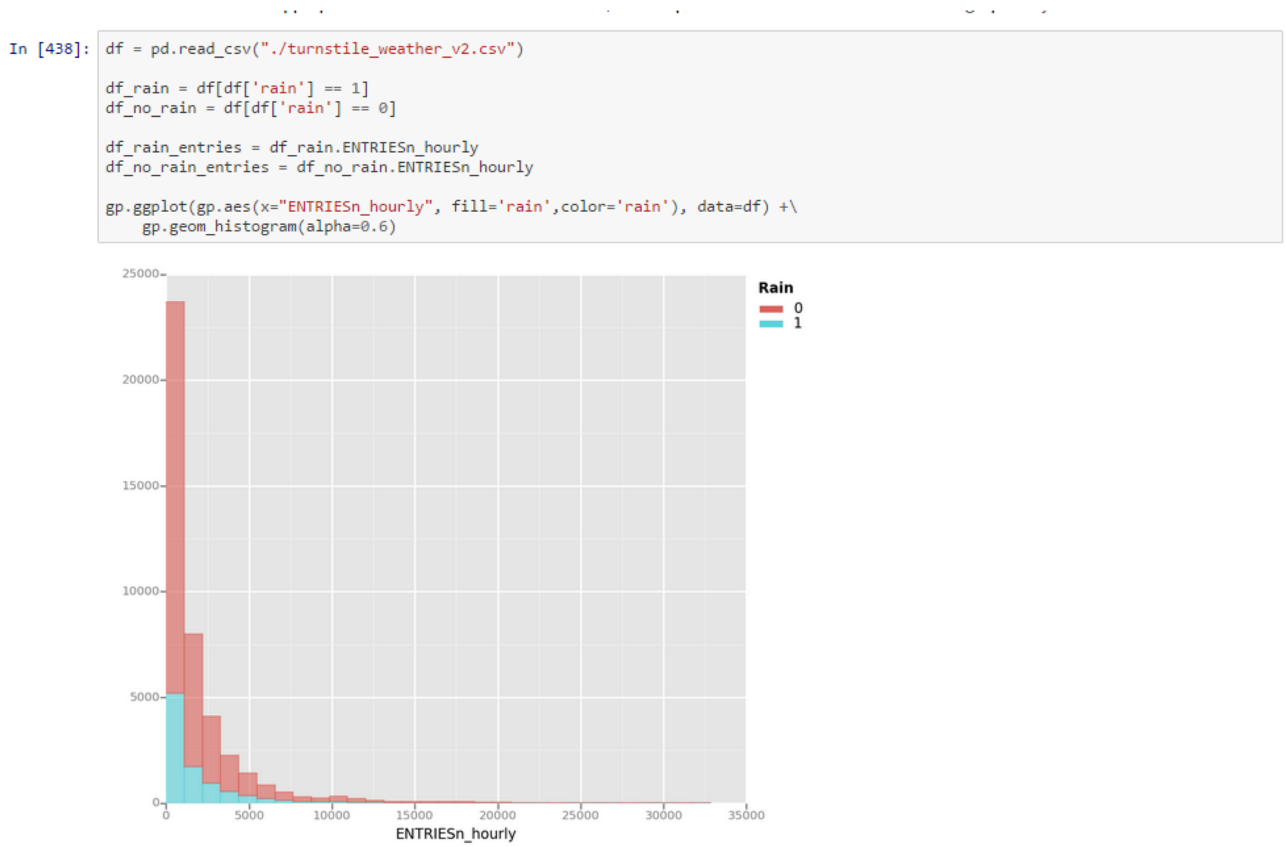
SPECIFICATION

All plots are appropriately labeled and titled. Plot is given an appropriate title. X-axis and y-axis are appropriately labeled. Visual cues (colors, size, etc) are easy to distinguish. It is clear what data are represented.

MEETS SPECIFICATION

Reviewer Comments

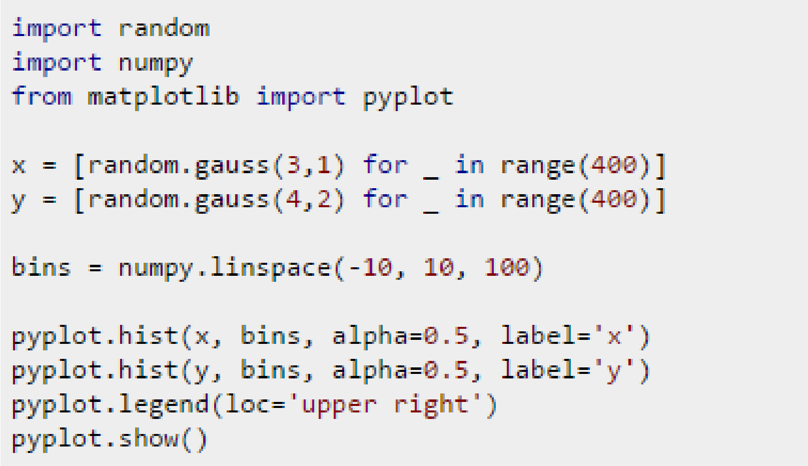
Stacked Histogram



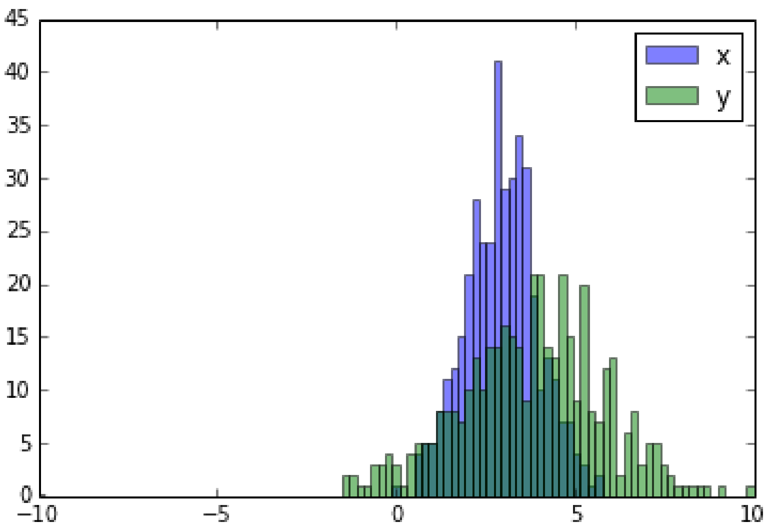
```
In [438]: df = pd.read_csv("./turnstile_weather_v2.csv")

          df_rain = df[df['rain'] == 1]
          df_no_rain = df[df['rain'] == 0]

          df_rain_entries = df_rain.ENTRIESn_hourly
          df_no_rain_entries = df_no_rain.ENTRIESn_hourly

          gp.ggplot(gp.aes(x="ENTRIESn_hourly", fill='rain',color='rain'), data=df) +\
              gp.geom_histogram(alpha=0.6)
```

Currently, the conditions represented in this plot are stacked - meaning the no-rain bar starts are the end of the rain bar. This is the default mode of the ggplot library for python and, currently, I am unaware of any methods around it. Stacked histograms are not recommended because it is **(1)** difficult to compare the conditions being represented because the specific value of each are added, **(2)** it the distribution of the top condition will seem more skewed because it is being pushed up by the bottom, and **(3)** without proper labeling (as seen here) it is extremely easy to misinterpret the results for being "unstacked". I would recommend using `matplotlib` library. I have attached some examples of bi-conditional histograms below.

1. Documentation
   http://matplotlib.org/api/pyplot_api.html
2. Histogram Example



```
import random
import numpy
from matplotlib import pyplot

x = [random.gauss(3,1) for _ in range(400)]
y = [random.gauss(4,2) for _ in range(400)]

bins = numpy.linspace(-10, 10, 100)

pyplot.hist(x, bins, alpha=0.5, label='x')
pyplot.hist(y, bins, alpha=0.5, label='y')
pyplot.legend(loc='upper right')
pyplot.show()
```



Quality of Analysis

SPECIFICATION

When using statistical tests and linear regression models, the choice of test type and features are always well justified based on the characteristics of the data.

MEETS SPECIFICATION

Reviewer Comments

Great work assigning proper Categorical Variables

I have noticed you have added more dummy variables than the default UNIT variable. This is great intuition since the the majority of the R-squared can be attributed to categorical variables rather than continuous variables and that most of our features are, in fact, categorical. For examples, let's take the *hour* feature. Time-series is a tricky variable. In situations where one is concerned about how the passage of time affects the outcome variable, time-series variables can be treated as regular non-dummy variables. In this instance, however, we are not concerned about the passage of time as much as we are concerned about **what** time it is and how *that* affects the outcome variable. In this case, the time-series variable is, in fact, a categorical variable. This means the *hour* variable should technically be used as a dummy-variable with each instance in time having its own column and values indicating if it is that time or not.

Great work!

SPECIFICATION

Statistical tests and linear regression models are described thoroughly, and the reasons for choosing them are articulated clearly.

MEETS SPECIFICATION

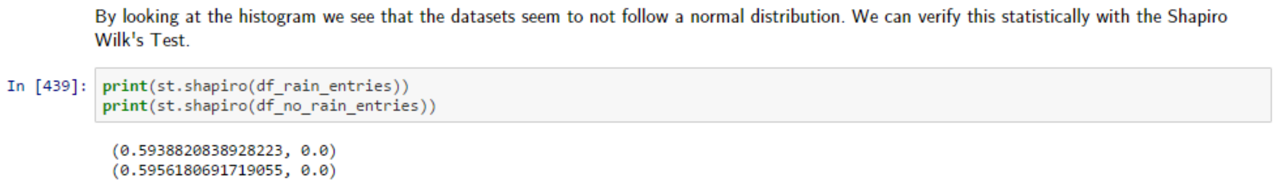
SPECIFICATION

The use and interpretation of statistical techniques are correct.

MEETS SPECIFICATION

Reviewer Comments

Use of a normality test

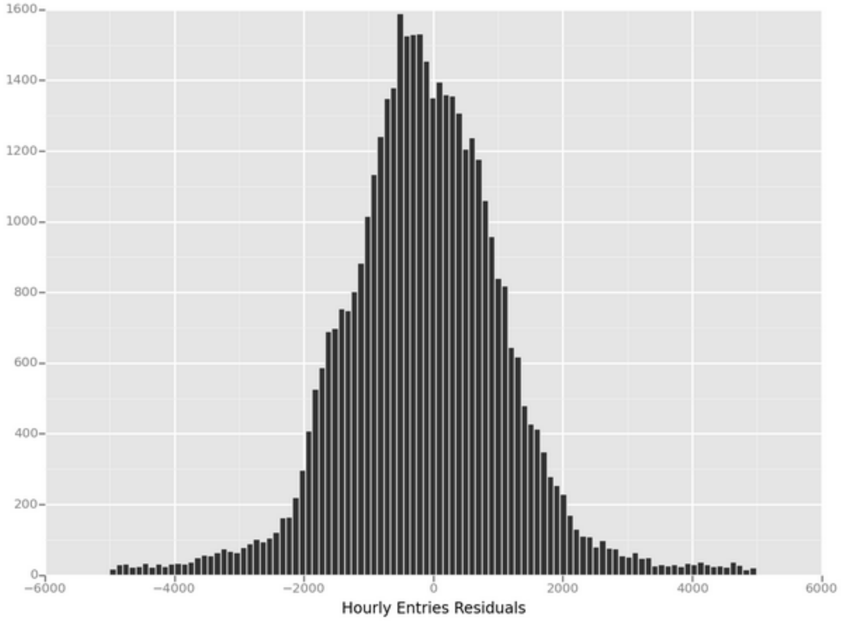


By looking at the histogram we see that the datasets seem to not follow a normal distribution. We can verify this statistically with the Shapiro Wilk's Test.

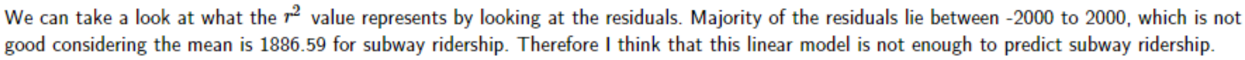
```
In [439]: print(st.shapiro(df_rain_entries))
          print(st.shapiro(df_no_rain_entries))

          (0.5938820838928223, 0.0)
          (0.5956180691719055, 0.0)
```

Outstanding use of a a measurable statistical test in order to determine normality.

Great use of Residuals



```
<ggplot: (350391517)>
```



We can take a look at what the r^2 value represents by looking at the residuals. Majority of the residuals lie between -2000 to 2000, which is not good considering the mean is 1886.59 for subway ridership. Therefore I think that this linear model is not enough to predict subway ridership.

Although out of the scope of the class, you have correctly utilized residuals. Great work here. Later on I will be talking about other ways to judge residuals. It is optional, but you may find interest in it.

SPECIFICATION

All conclusions are correctly justified with data.

MEETS SPECIFICATION

Reviewer Comments

RESIDUAL (MODEL ERROR) ANALYSIS IS A VITAL SKILL TO HAVE IN JUDGING THE PERFORMANCE AND APPROPRIATENESS OF A MODEL. THE FOLLOWING IS OPTIONAL BUT I HIGHLY RECOMMEND FAMILIARIZING YOURSELF WITH THIS MATERIAL AS IT IS NOT ONLY APPLICABLE TO MANY OF THE QUESTIONS IN THE PROJECT, BUT IS CRUCIAL TO YOUR FUTURE AS A DATA SCIENTIST.

Interpreting the Residuals (Optional)

Examining the residuals (errors) of a model can give us some great insights to the behaviour of the dataset. In the following I have given a simple rundown on how to examine them. As you will see, it can be used to justify many of the sections in the project.

THE FOLLOWING MATERIAL IS A GREAT WAY TO JUSTIFY THE CONCLUSIONS MADE IN **SECTION 2.6**, **SECTION 4**, AND **SECTION 5** AND IS HIGHLY RECOMMENDED TO FAMILIARIZE YOURSELF WITH IT.

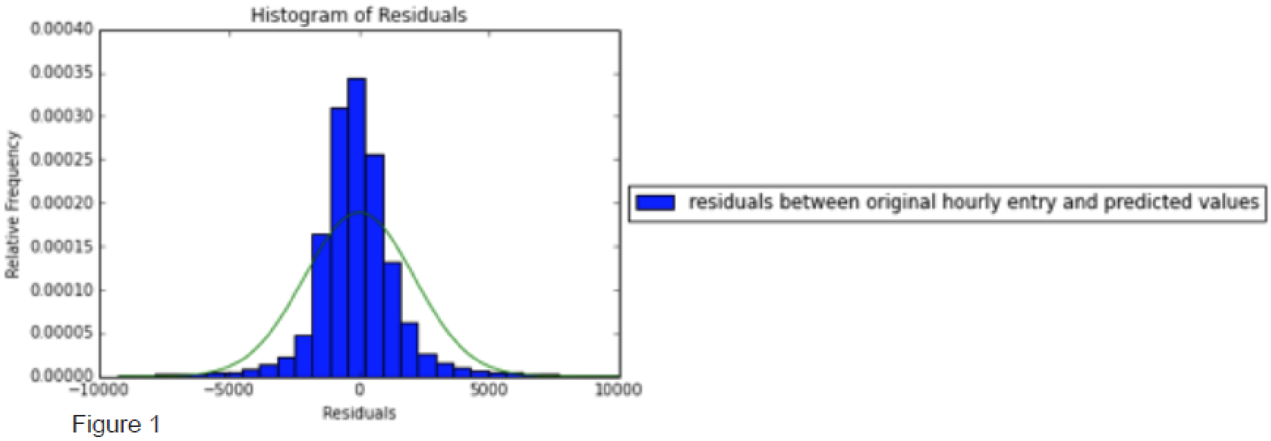


Figure 1

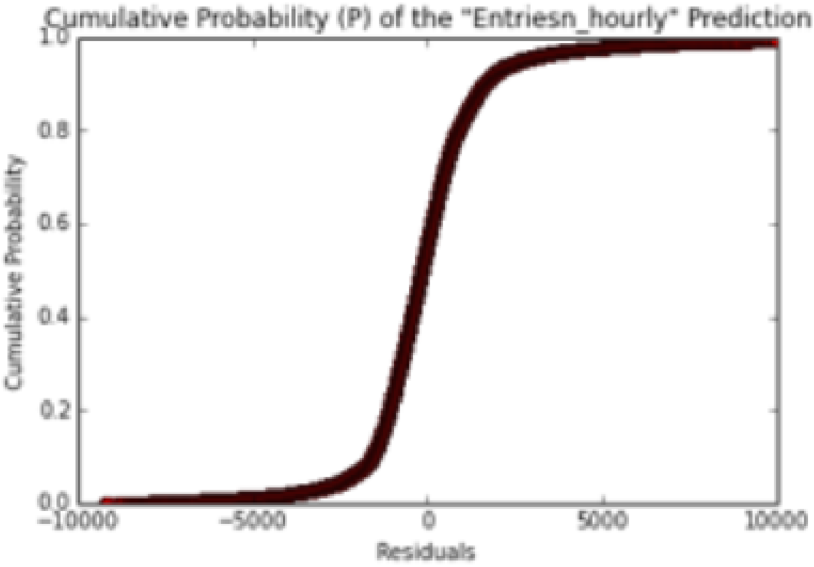
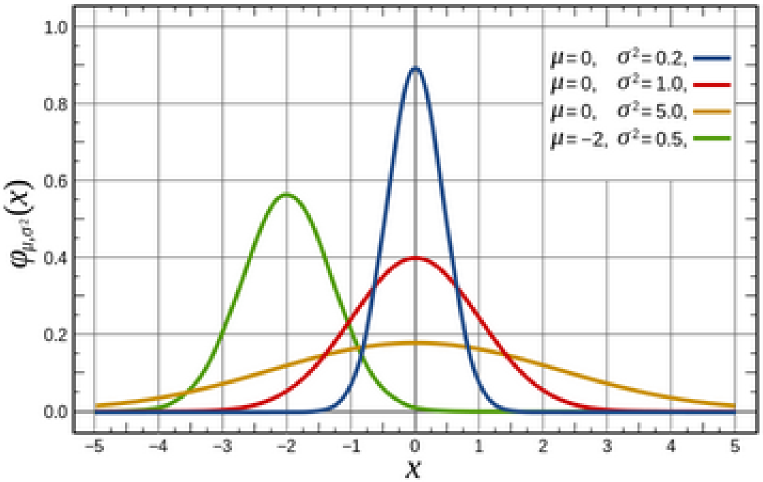


Figure 2

- Above, in *Figure 1*, I have included a version of our residuals with an overlay of the normal distribution, also known as the **P**robablity **D**ensity **F**unction **(PDF)**. Below that, in *Figure 2*, is the **C**umulative **D**ensity **F**unction **(CDF)** of our residuals. Both are great ways to visually determine distributions.
- Below are examples of both these functions with different **means** and **standard deviations**.

This is the **Probability Density Function**.

$$f(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



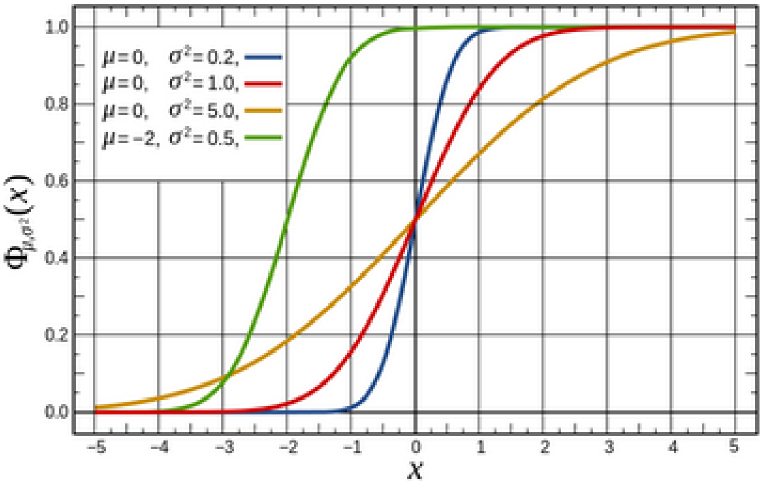
"Normal Distribution PDF" by Inductiveload - self-made, Mathematica, Inkscape. Licensed under Public Domain via Commons

If $\mu = 0$ and $\sigma = 1$, the distribution is called the **standard normal distribution** or the **unit normal distribution** denoted by $N(0,1)$ and a random variable with that distribution is a **standard normal deviate**.

- The curve with the **red line** represents the normal distribution
- One can visually tell obviously non-normal distributions, but is there are way to test it measurably?
    - **Shapiro-Wilk**
      https://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test
      The Shapiro-Wilk Test is a common statistical test for checking normality they same way we test our hypothesis for T-tests and Mann-Whitney U-tests. The **Null** is that the population is normally distributed and we test to reject or fail to reject that null.

This is the **Cumulative Distribution Function**

$$\Phi(x) \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt$$



"Normal Distribution CDF" by Inductiveload - self-made, Mathematica, Inkscape. Licensed under Public Domain via Commons

- The **red line**, again represents the normal distribution.

The Key idea I want to convey is that if we were to examine our residuals and compare their distributions to the Normal Distributions above, we would see that the residuals are not normally distributed. That, coupled with the fact that the residual histogram has very long-tails with extremely high absolute errors, means that our model is actually NOT a good fit for the dataset.

SPECIFICATION

No incorrect conclusions are drawn from the data.

MEETS SPECIFICATION

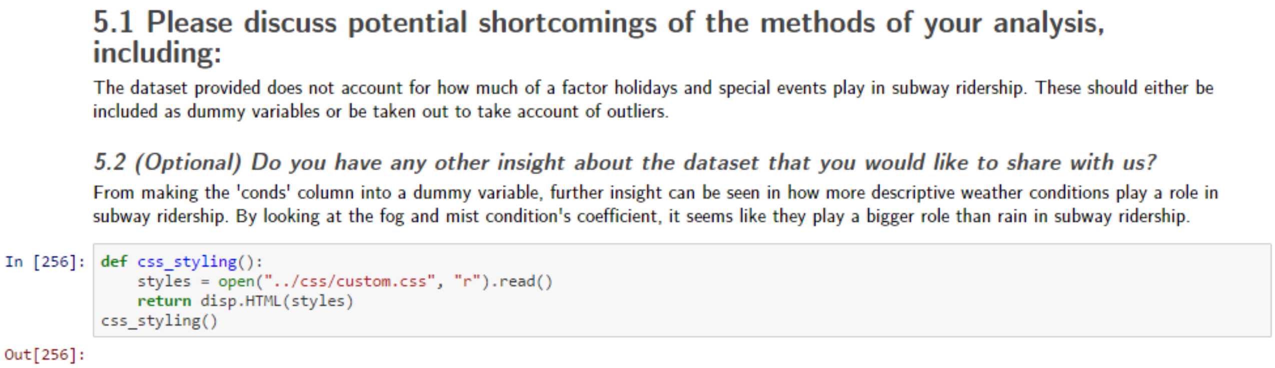
SPECIFICATION

Some shortcomings of the dataset and statistical tests or regression techniques used are appropriately acknowledged.

MEETS SPECIFICATION

Reviewer Comments

Shortcomings (OPTIONAL)



5.1 Please discuss potential shortcomings of the methods of your analysis, including:

The dataset provided does not account for how much of a factor holidays and special events play in subway ridership. These should either be included as dummy variables or be taken out to take account of outliers.

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

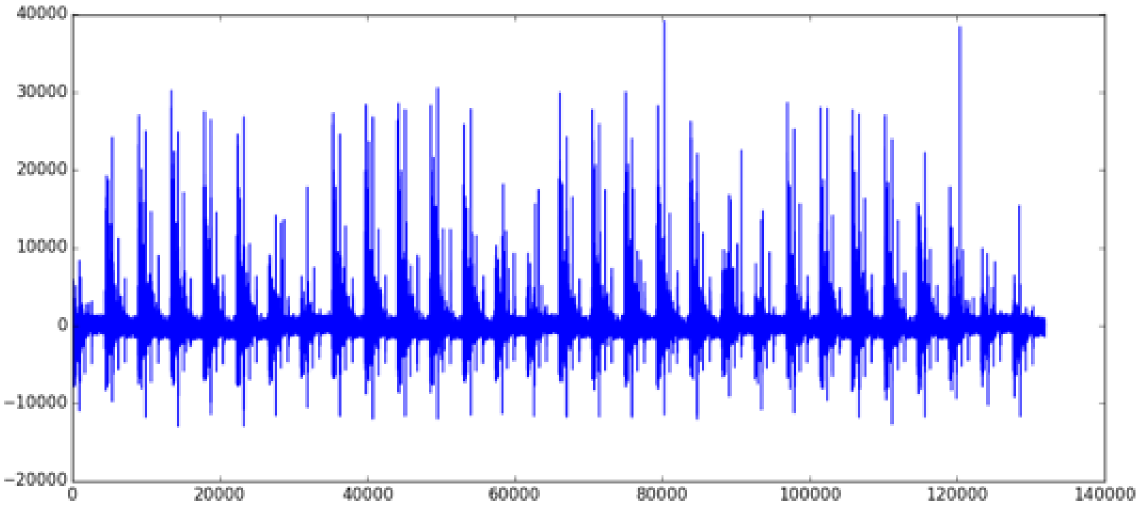
From making the 'conds' column into a dummy variable, further insight can be seen in how more descriptive weather conditions play a role in subway ridership. By looking at the fog and mist condition's coefficient, it seems like they play a bigger role than rain in subway ridership.

```
In [256]: def css_styling():
              styles = open("../css/custom.css", "r").read()
              return disp.HTML(styles)
          css_styling()
Out[256]:
```

Thank you for your discussion on shortcomings. Consider the following ideas as well. Some you may have already mentioned but I like to keep to maintain continuity.

1. **Dataset**
   - Consider the scope of the data. It spans a month. Do you believe that is long enough?
   - Because there are many variables included in the dataset that might be very closely related, such as minimum, mean and maximum temperature, it may be difficult to disentangle the effects of such similar features and we may run the risk of problems with collinearity, which can cause some linear regression algorithms to give incorrect results. http://en.wikipedia.org/wiki/Multicollinearity
2. **Statistical Test**
   - You can talk about the fact that the Statistical Test is only comparing the differences between the conditions of one feature; rain, when clearly other are other variables that seem to affect the ridership even more.
3. **Regression Model**

Do you think that a linear model is appropriate in this context? Why? You could simply extend to 5.1 the reasoning regarding the residuals I proposed earlier. In addition it might be interesting to plot the residual per data point, some interesting patterns might emerge to help understand why a linear model might not be the best choice for this problem.
As I have given hints in the prior section, examining the residuals of a regression model is one of the best ways to determine the effectiveness of the model. You can use any of the techniques above. By merely plotting the difference between predictions and actual values (residuals) you will, most likely, see that the residuals follow a cyclical pattern. If so, that might prove that some non-linearity in the data should be addressed by designing a non linear model. The code is really simple and looks like this: import matplotlib.pyplot as plt
plt.plot(data - predictions) plt.show().
The following is an example of what you might see (randomly created).



You may see a plot similar to this that maybe shows more of a cyclical/higher degree shape of the residuals. This could indicate that the outcome does not respond linearly to the features and a linear model might not be the best fit.

How satisfied are you with this feedback?

? Have a question about your review? Email us at review-support@udacity.com and include the link to this review.

NANODEGREE PROGRAMS

Front-End Web Developer
Full Stack Web Developer
Data Analyst

iOS Developer

Android Developer

Intro to Programming

Tech Entrepreneur

STUDENT RESOURCES

Blog

Help & FAQ

Catalog

Veteran Programs

PARTNERS & EMPLOYERS

Georgia Tech Program

Udacity for Business