Project 1 Presentation Notes

It would be helpful to talk a little about your reasons for joining, what the experience has been like for you so far, and how the project went for you.

Reasons for joining:

I recently separated from the military, where I was a physicist for 7 years. I have both a bachelors and masters in physics, with most of my programming and data analysis experience in Matlab and an introductory course in Java. I first discovered the Data Scientist world when looking for jobs (since many postings inlcude a Physics degree as desiriable), and the more I learned about it the more it seemed like a great fit, since my favorite parts of my physics experience have been programming and data exploration/analysis.

I'm currently looking for a job as a Data Scientist, so I looked into a number of different online courses to improve my skills and make me more marketable. I ultimately chose Udacity because of the Nanodegree concept of having a community like this along with a "portfolio" at the end of the process.

Experience so far:

I've really enjoyed it so far, and I've been learning a lot.

How the project went for me:

I wish I had looked ahead to the project description to see how the problem sets fit into the Project, otherwise I probably would have recorded some of my answers better along the way. Otherwise, I thought the lessons and problem sets did a good job of preparing me for what I need to know for the project.

After that, discussing the methodology seems like a good idea. How did you explore the data?

I mostly did everything solely using the built-in grader. I did this mainly to streamline the coding process and make things a little bit more frictionless. Since I had to use the browser IDE anyway, it was easier to just tweak some of the code as necessary to find the extra bits of information I was looking for in the project. When I first started

the Nanodegree my python skills were pretty lacking, so that was the easiest way to dive in. My workflow looked like this: modify the downloadable scripts using Sublime Text (my eyes much prefer a darker color-scheme; I absolutely love Sublime Text but any code editor would be fine), copy over to the Udacity IDE to check for errors/accuracy, and then continue on. I relied heavily on the documentation and sites like StackOverflow to help when I was stuck or trying to figure out a particular way to do something in python or Pandas.

IPython was also really helpful for trying little bits of code to see if a certain method or function was performing as I expected.

As an aside, for the current project I've been doing the same thing, but I recently discovered the "Build" command in Sublime Text, which runs the code in the editor itself and displays any output. It's great for checking for syntax errors in your code, and for actually running the code if you have the data files set up locally.

What kind of plots did you try before settling on the one included in your project?

I chose the plots for my project partially from the prompts/examples given, and partially from what interested me.

Since I've never lived in a city with a subway system, I thought it would be interesting to examine how ridership changed the time of day on weekends versus weekdays. I did this because I assumed that ridership patterns would change during the weekend (compared to weekdays) because most people don't work on the weekends and probably socialize more then.

For my other plot, I knew I wanted to tie in the rain data, so I decided to plot ridership by day of the week and rain status. This would both show the relationship between ridership and specific day of the week, as well as the impact of rain on ridership on those days. I ulitimately decided to inlcude the plot of ridership by hour and day type in my project, since so much of the rest of the questions were related to the impact of rain.

Looking at the weekday data, reported ridership entries peak during the hours of 8:00–11:59 PM. One possible explanation for this is people either going out to dinner or coming home from dinner and work at the end of the day, again since the readings taken during those time could include entries from the previous 3:59 hours. Weekday reported ridership is at its lowest during the hours of 4:00–

7:59 AM, as expected (since that is probably when most people are sleeping). According to the NYC MTA website, the subway does actually run 24 hours a day. Reported ridership then increases with each time bin, when it dips slightly during the hours of 4:00–7:59 PM, likely corresponding to a slight dip during mid-afternoon (factoring in the temporal intervals in the readings), before the evening rush. Overall, reported ridership is higher during the evening hours than during the daytime. Some possible explanations for this are people going out to dinner or running errands after work further away from their normal commute area, or the fact that people may be less likely to walk instead of taking the subway when it is dark outside due to safety concerns.

While weekend reported ridership is lower than weekday reported ridership, much of the same trends hold. Reported weekend ridership is lowest during 4:00–7:59 AM, as on weekdays, and increases throughout the day. However, the peak average reported ridership occurs during the hours of 4:00–7:59 PM, compared to 8:00–11:59 PM on weekdays. This is likely due to the fact that most people don't work on weekends, so they can go out to eat or run errands earlier in the day. One other interesting insight is that average reported ridership is higher during the hours of 12:00–3:59 AM on weekends than on weekdays, likely due to an increase in nightlife and late-night activities on the weekend (particularly Saturday).

How did you go about choosing features for your regression?

Like most of the other students, I relied on a mix of intuition and improvement in results.

I used rain because that is part of the main relationship that we're trying to test, and I suspected that rain would affect the amount of people riding the subway.

I used Hour because I thought that the time of day would have a large impact on the ridership totals (e.g., lower ridership during the late evening when people are mostly sleeping).

I used mintempi because I suspected ridership would increase (or at least be impacted) if it was especially cold outside, since people would probably be less likely to walk outside instead. This of course depends on the assumptions made on the order of preferred modes of transit; depending on availability and cost concerns, I could see taxi ridership increasing on cold days.

I used maxtempi for most of the same reasons I used mintempi, except assuming the opposite: that ridership would increase if it was

warm outside.

I used meanwindspdi because I thought that if it was especially windy outside, people would be less likely to walk outside.

I used fog because I assumed if it was foggy outside people would be more likely to take the subway than walk.

I used precipi because I thought that the amount of rain might have an impact on how many people decided to take the subway (e.g., people might be less inclined to "tough it out" with an umbrella or raincoat if it was raining especially hard, as indicated by a large precipi value).

I used UNIT because it was included in the default code and because it made sense that the unit collecting the turnstile information (and the station, by proxy) could have an impact on the ridership. This was included as a dummy variable because it is a categorical feature, and we wanted to denote in our model if an entry belonged to a certain category or not.

In the end, all of the features I selected were ultimately included because they also resulted in a larger R2 value than the R2 value calculated without them, and they could be included without

exceeding the 30 second server calculation limit.

Using StatsModels OLS, the coefficients are:

```
rain              10.046302
Hour              62.239983
mintempi         -15.976535
maxtempi           3.794069
meanwindspdi      33.752174
fog              274.059647
precipi          -81.110211
```

My calculated R2 value using StatsModels OLS was 0.485414650479.

What about the data surprised you?

What surprised me most was that the impact of rain was different than I expected (although admittedly I never lived in a big city with public transportation). I'm sure part of the reason this question was answered was because it required some additional thought/consideration. It was a good reminder to go in with open eyes and no bias, since intuition alone can let you down.

I was also surprised that the data was as coarse and messy as it was. It would have been great to have hourly readings to be able to get a

better understanding of the relationship between rain and subway ridership, as well as general ridership trends throughout the day.

What about the project did you find particularly difficult?

Trying to determine the best approach and the best way to analyze the data, since there were so many directions you could go with so many features in the data.