# PROJECT – ROUGH DRAFT

11 April 2015

All the materials for my project can be found here: https://github.com/l2nguyen/metro

What I have done so far:

- Collected datasets for Metro rail ridership, weather, unemployment, gas prices
- Visualized the data for those variables
- Merged those dataset into one dataset to use for modeling
- Made some very simple linear regression model for weekday ridership and evaluated them using RMSE. They were not so good
    - First model: Features were precipitation related variables (precipitation, snow depth, snow fall). RMSE was very high (105,196) and it appears that it always predicted about the same number for most cases.
    - Second model: Features were precipitation related variables and unemployment rate. RMSE was slightly better than the first model (105,110). From graphing the residuals, I saw that it predicted most of the weekday ridership correctly but there were cases where it was way off. That drove the RMSE up.

What remains to be done:

- Transform the capital bikeshare system data into the format that I want it. I want to use it as a proxy for tourism but also, I think the capital bikeshare has also given people an alternative form of transportation from the Metro rail
- Visualize the feature variables and metro ridership together
- Standardize the data by number of cars or flatten the data into weekly ridership numbers so that I do not have to make a weekday/weekend model
- More modeling because the ones I have made so far are not very good at predicting metro ridership from looking at the RMSE
- Make the documentation more informative. It is currently only the data dictionary of the dataset that I used for modeling