

# Forecasting Metro Ridership

Lena Nguyen | May 2, 2015 | GA DAT 6

# The Goal

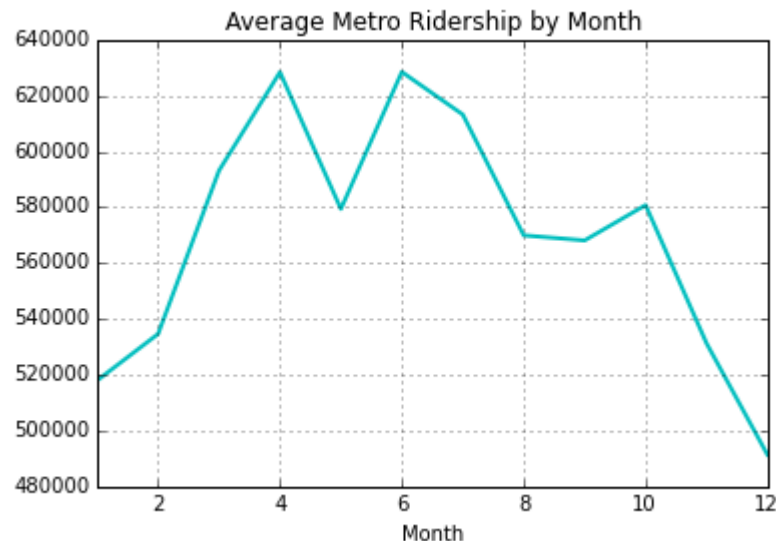
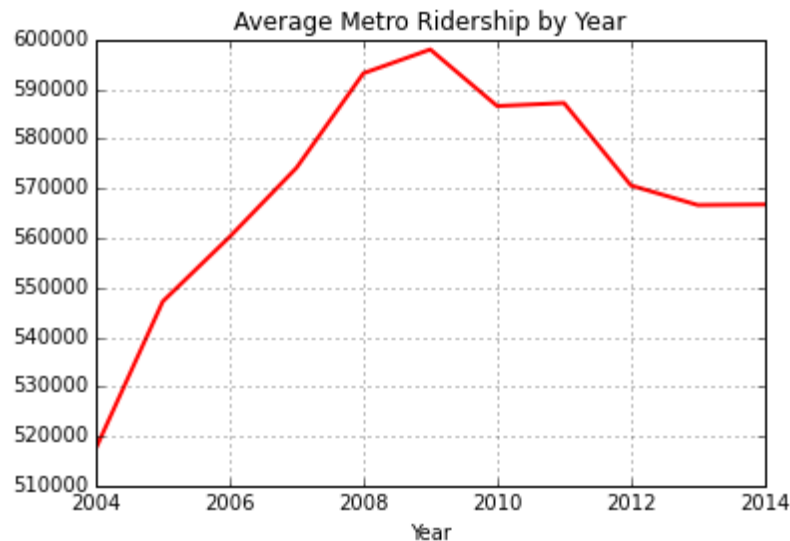
To forecast Metro ridership based on different input variables

# Response:

## Number of metro riders per train

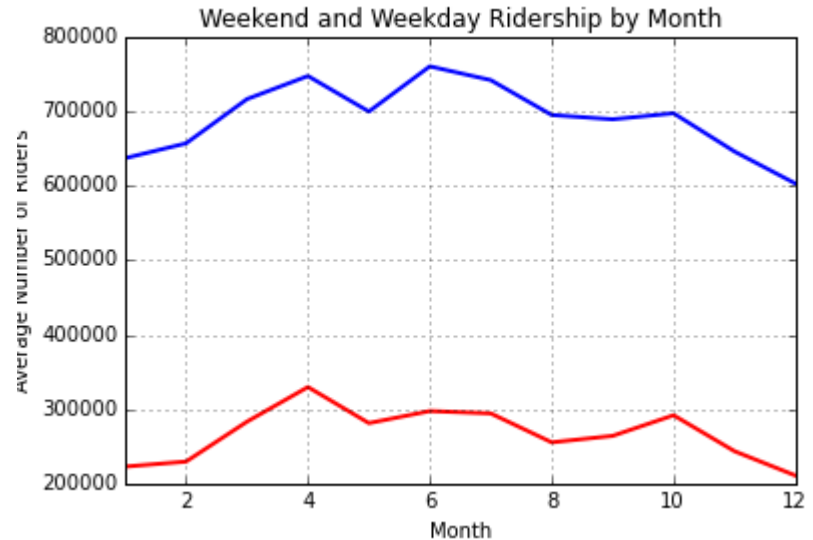
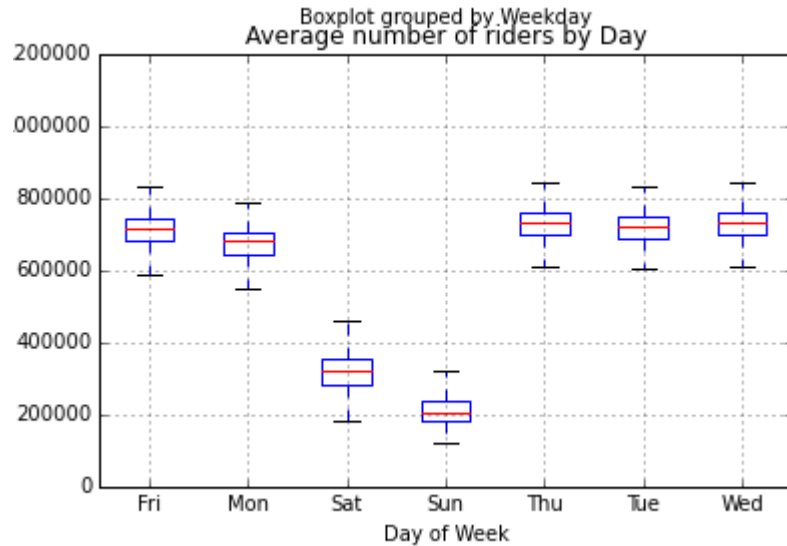
- Open Data DC: daily ridership data from 2004-2014, but not disaggregated by station
- Data has 4018 observations total
- Used simple math and data from WMATA website about train frequency to figure out how many trains run per day
- Using riders per train somewhat deals with the effect of weekday/weekend variations and holidays.

# Data Visualization



Using the metro ridership data from Open Data DC

# Data Visualization



In graph above, blue line is weekday and red line is weekend

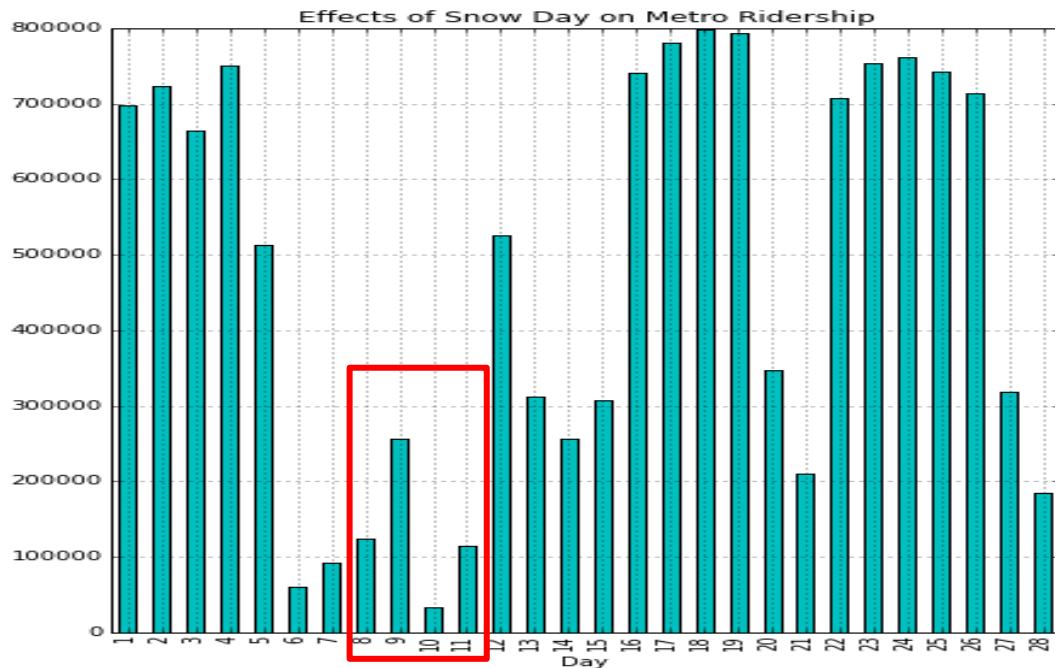
# Features

- Gas prices
- Weather (ie max temp, min temp, snow amount)
- Employment (ie total number of employed people)
- Holiday (if the federal government was closed)
- Capital Bikeshare: Number of registered/casual riders
- Binary variables for each month
- Binary variables for each day of the week

# Ridership and Weather

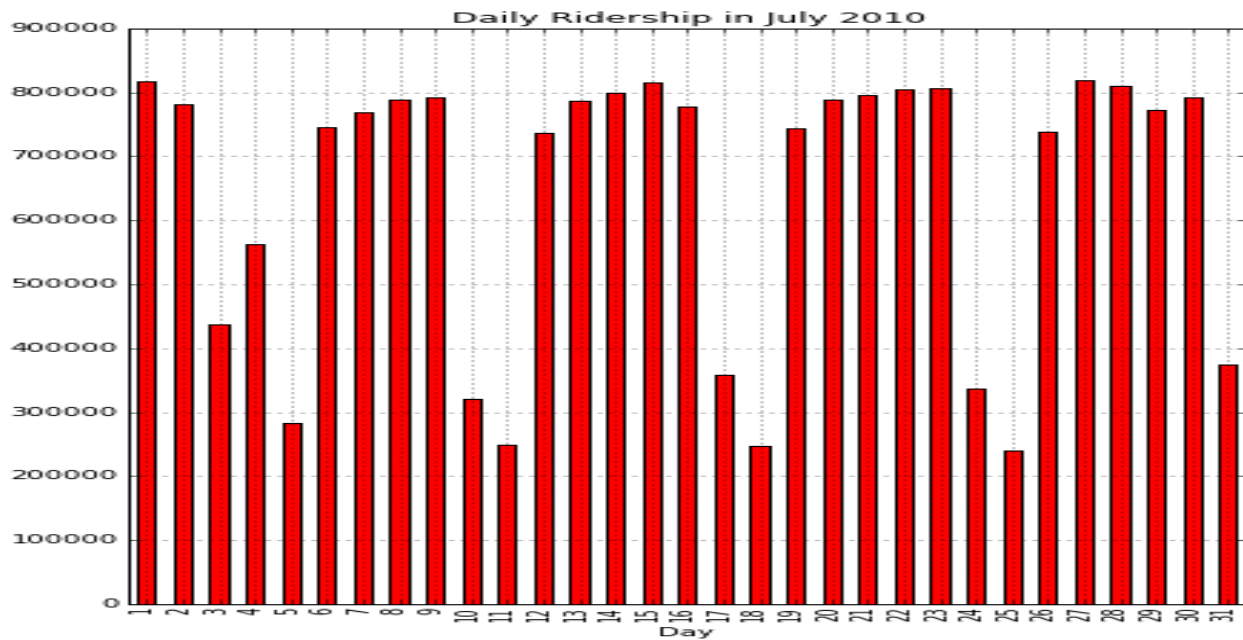
- Tourists less likely to visit during cold weather
- People are less likely to go places during bad weather
- Snow day means fewer people are going anywhere so much lower ridership

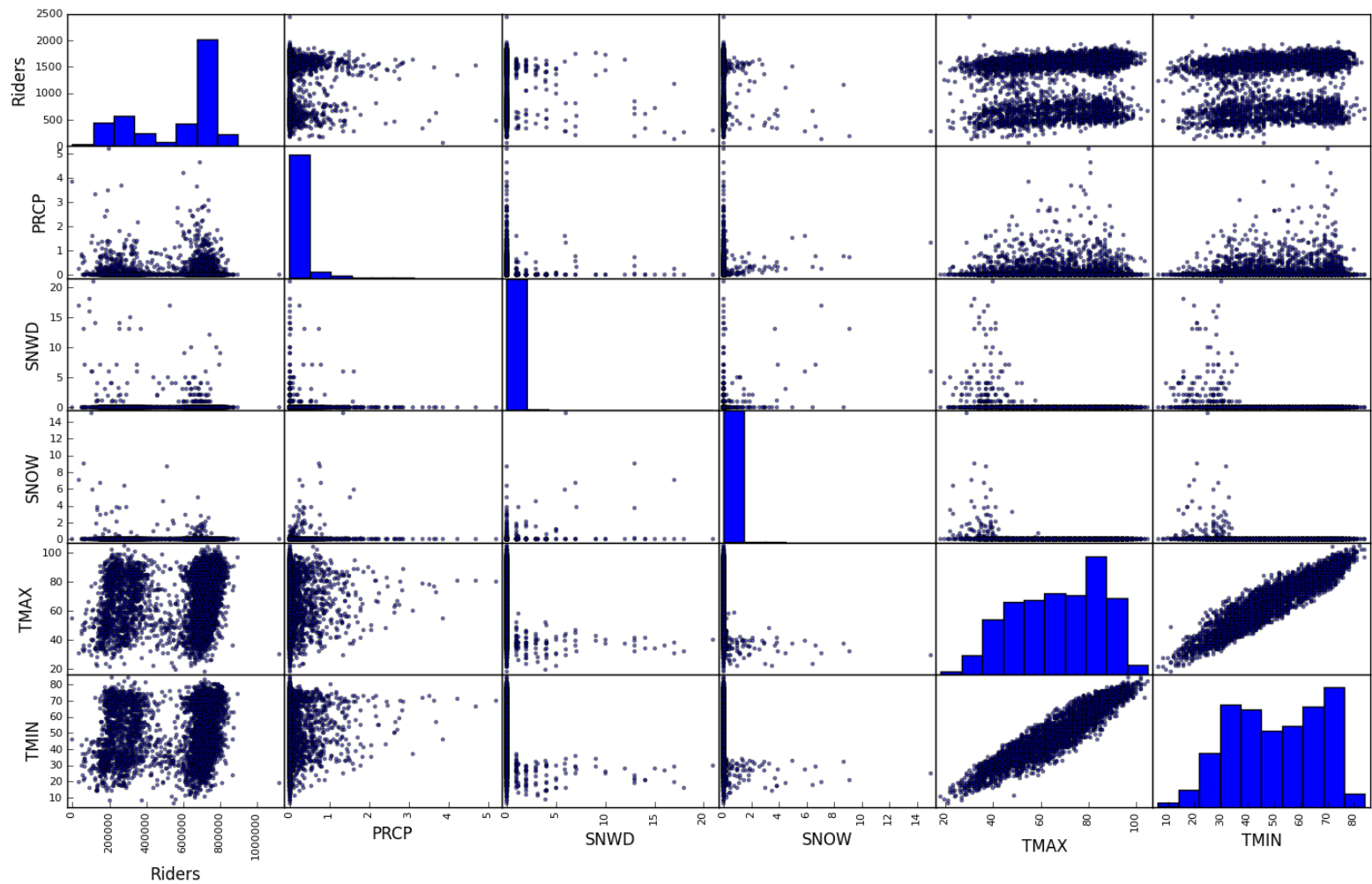
# Winter Ridership (February 2010)





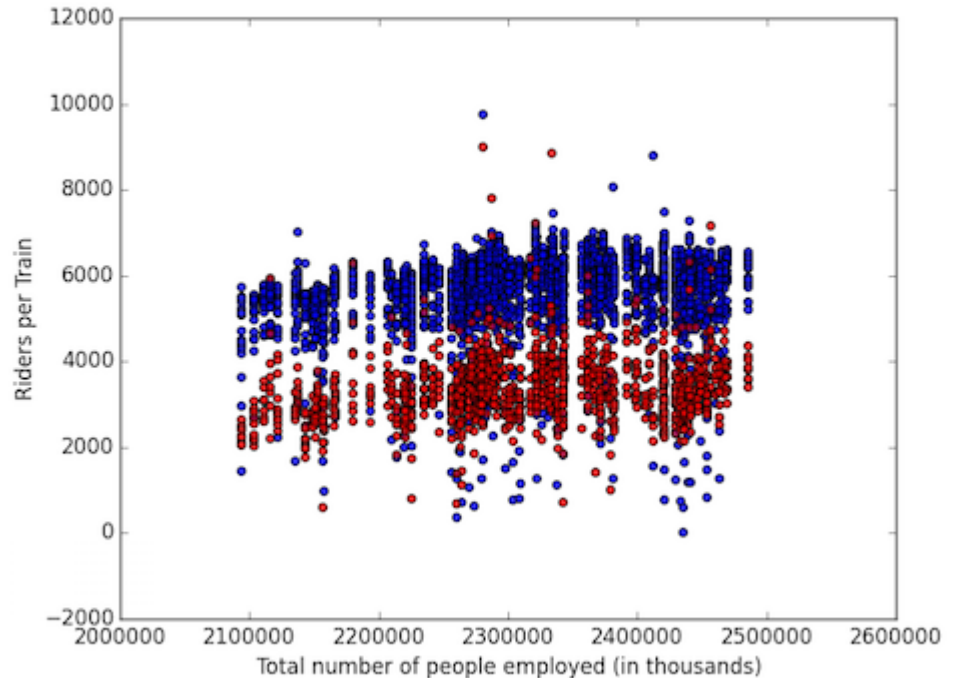
# Summer Ridership (July 2010)





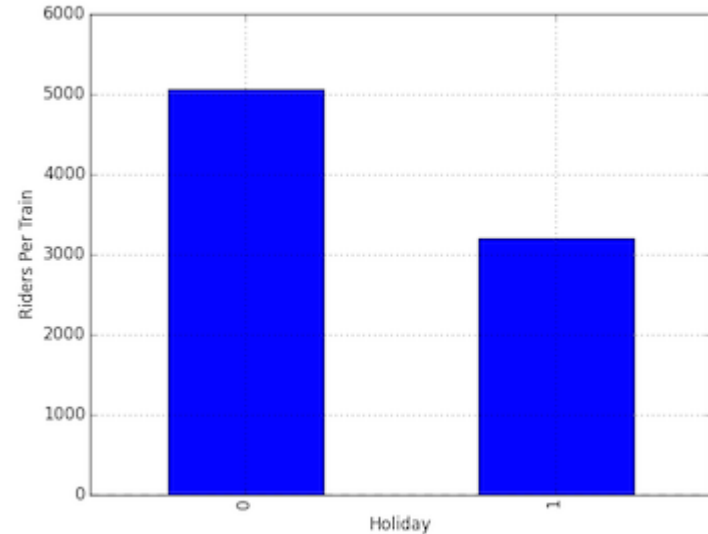
# Ridership and Employment

- Weekend = Red dots; Weekday = Blue dots
- Employment does not seem to have much of an obvious relationship with number of riders



# Ridership and Holidays

- It being a holiday has a very obvious relationship with metro ridership
- Number of riders per train is 40% less on holidays than on regular days



# Dealing with outliers

- Linear regression models are very sensitive to outliers
- Standardized response variable with z-score
- Look at characteristics of outliers. Should they be removed?
- Ended up removing outliers greater than  $\pm 3.5$  SDs away (Only two data points)

# First set of models

- All models use trimmed dataset
- No parameter tuning
- Poor performance by all models

	CV RMSE Score	Train Set R Squared	Test Set R Squared
Linear Regression (w/o feature selection)	1180.55	0.146	0.136
Linear Regression (w/ feature selection)	1256.35	0.253	0.253
Random Forest	1146.71	0.868	0.267
Gradient Boosting	1074.53	0.454	0.356

# Weekday/Weekend Double Model

- Hypothesis: Features have different effect on weekend and weekday ridership.
- Split the original dataset into a weekend dataset and weekday dataset
- Train a separate model for each dataset

# Weekday/Weekend Double Model

Features	Weekday (p-values)	Weekend (p-values)
<b>WT05 (Hail)</b>	0.00004	0.033893
<b>WT16 (Rain)</b>	0.69943	0.000155
<b>WT03 (Thunder)</b>	0.00077	0.100543
<b>March</b>	0.00220	0.016964
<b>May</b>	0.26629	0.009586
<b>August</b>	0.91672	0.081480
<b>September</b>	0.59677	0.731114



# Weekday/Weekend Double Model

Model	RMSE		R squared	
	Weekday	Weekend	Weekday	Weekend
Linear Regression (w/o feat selection)	721.20	608.55	0.455	0.366
Linear Regression (w/ feat selection)	749.19	662.37	0.475	0.444
Gradient Boosting	617.17	604.65	0.546	0.412

# Adding more features

- Added days of the week as a set of binary variables
- No parameter tuning
- Significantly better performance by all models!

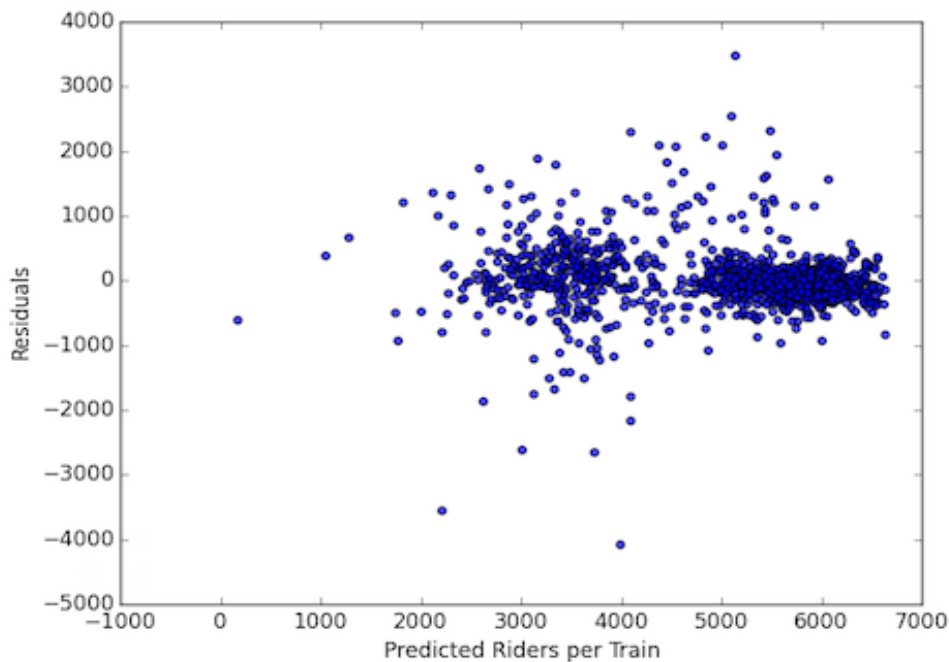
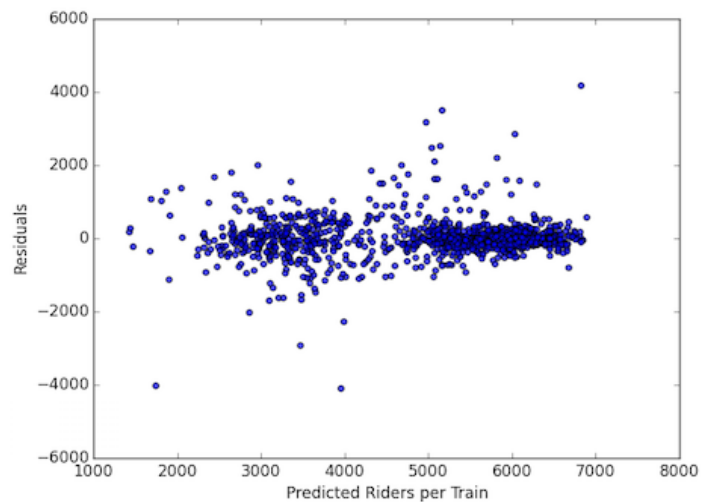
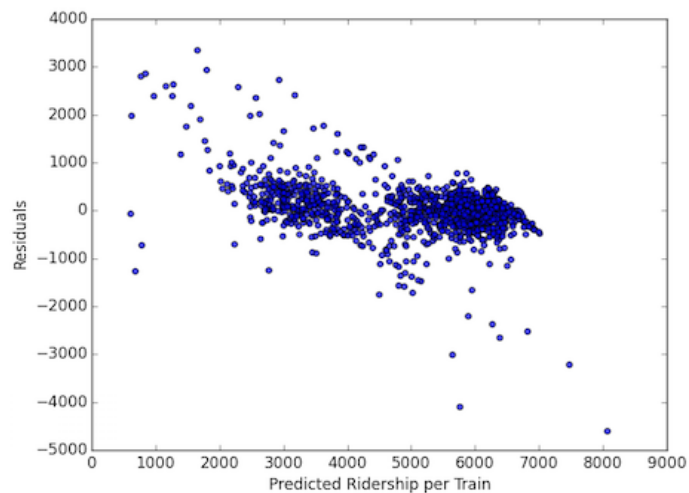
	CV RMSE Score	Train Set R Squared	Test Set R Squared
Linear Regression (w/o feature selection)	576.64	0.817	0.818
Linear Regression (w/ feature selection)	575.56	0.820	0.821
Random Forest	529.16	0.973	0.844
Gradient Boosting	480.98	0.899	0.871

## Residuals Plots

Lower left: Linear Regression

Lower right: Random Forest Regressor

Below: Gradient Boosting Regressor



# Challenges and Lessons Learned

- I could not find data for all the features I wanted (ie Uber/Lyft)
- Data processing takes a long time
- People are really hard to predict
- If you do not have good features in your data, even the finest model will not save you

# Future work

- Look at the larger residuals and see if they have anything in common (maybe even model it)
- Add more features, such as dates for sports games and cultural/political events
- Study the characteristics of long/short haul trips and see how they differ (if at all)
- Study how people are moving through the system and when certain stops/routes get more use
- Look at how the new Silver Line is affecting metrorail ridership