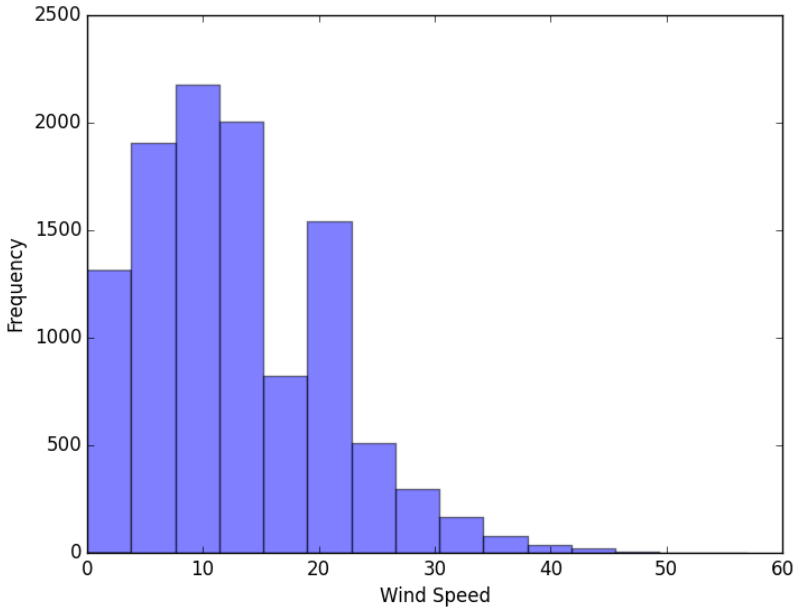
CMPSC 448 PROJECT

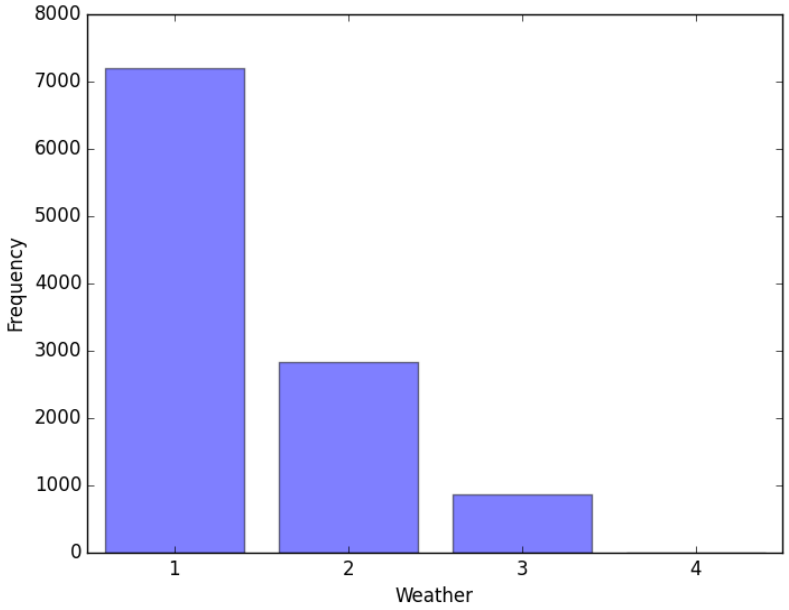
BIKE SHARING DEMAND

Jie Zheng
Matthew Wood
Rashmi Shukla
Kevin Sheeran
Kevin Chiang

Analyzing the Data

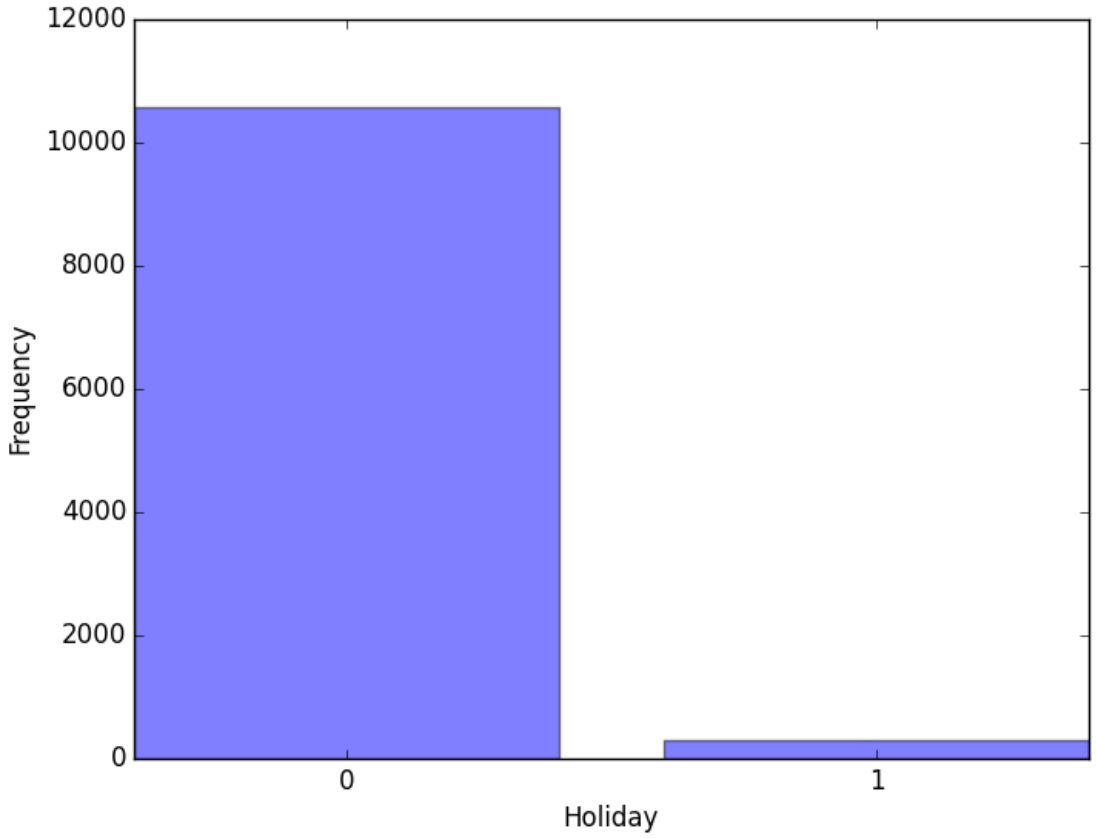
Lopsided frequencies





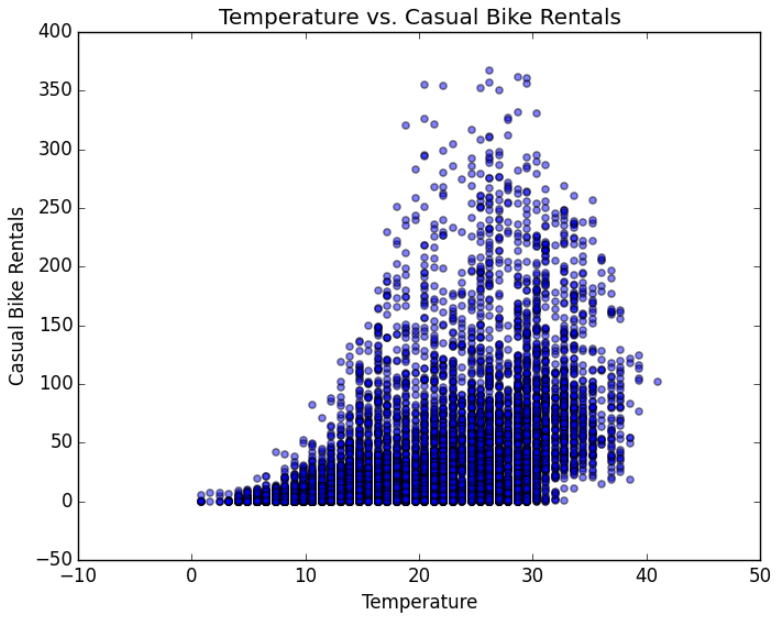
Analyzing the Data

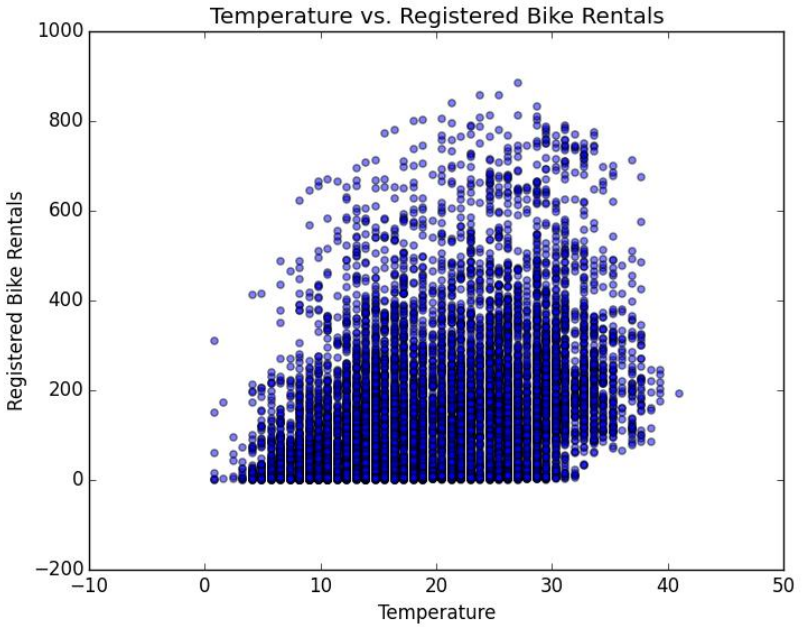
Lopsided frequencies



Analyzing the Data

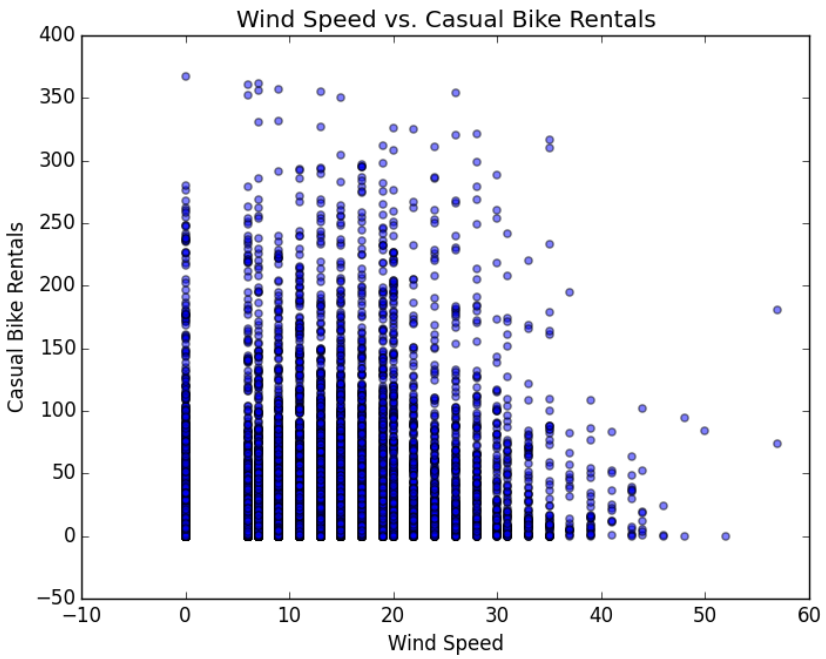
- Scatter plots of each variable vs. target variable
- Rentals by casual riders vs. registered riders

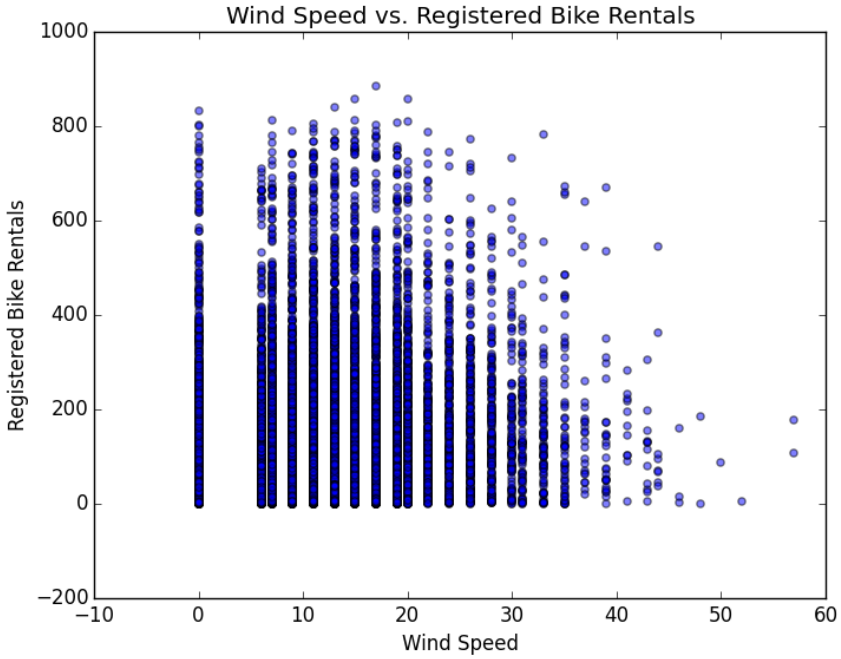




Analyzing the Data

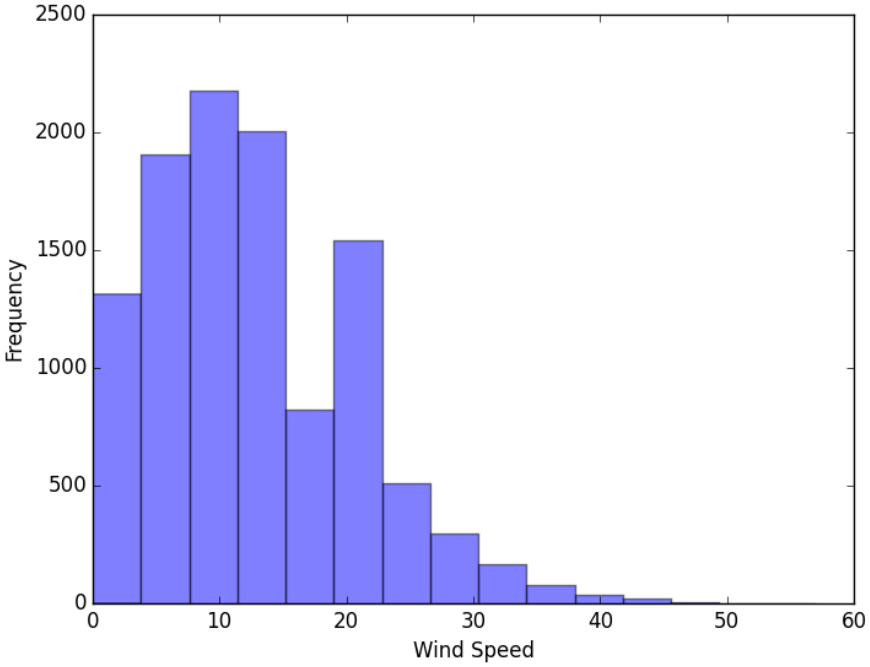
At first, it appears that people rent less frequently on very windy days, but…

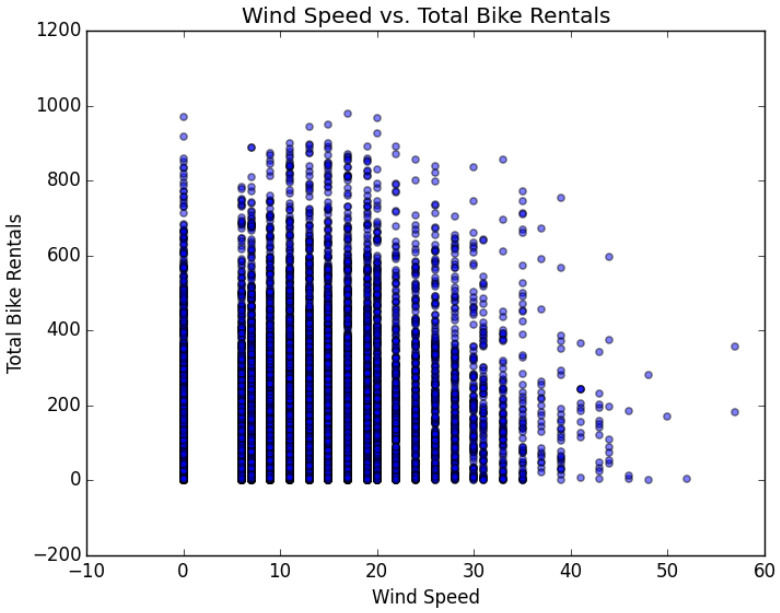




Analyzing the Data

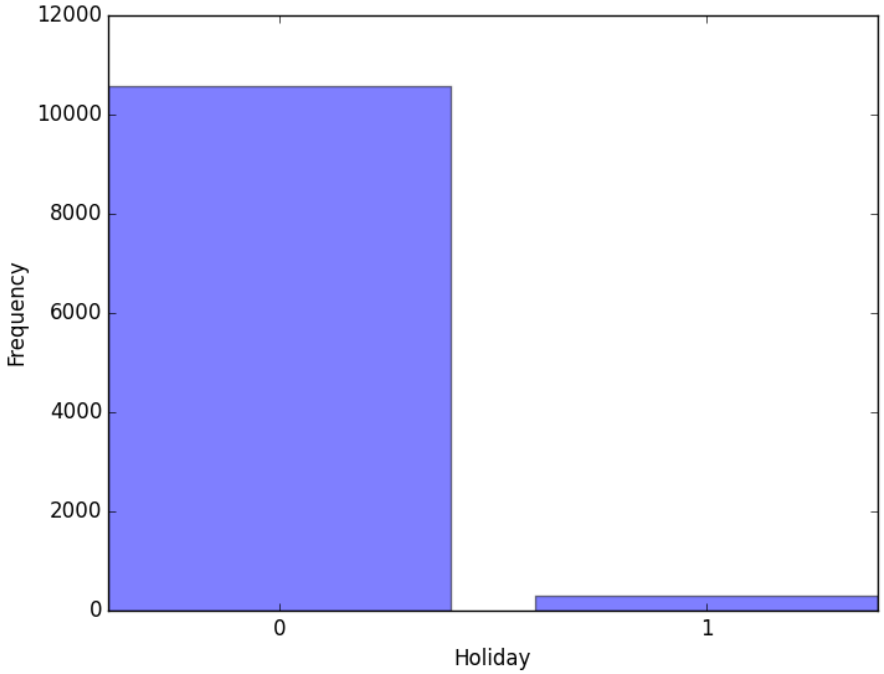
…it's actually because wind speed frequencies are heavily skewed to the right

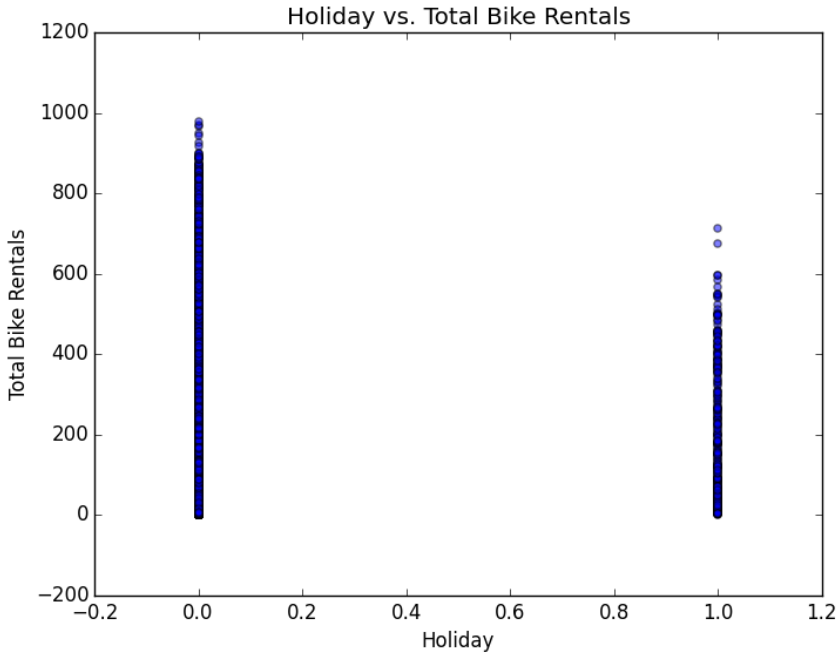




Analyzing the Data

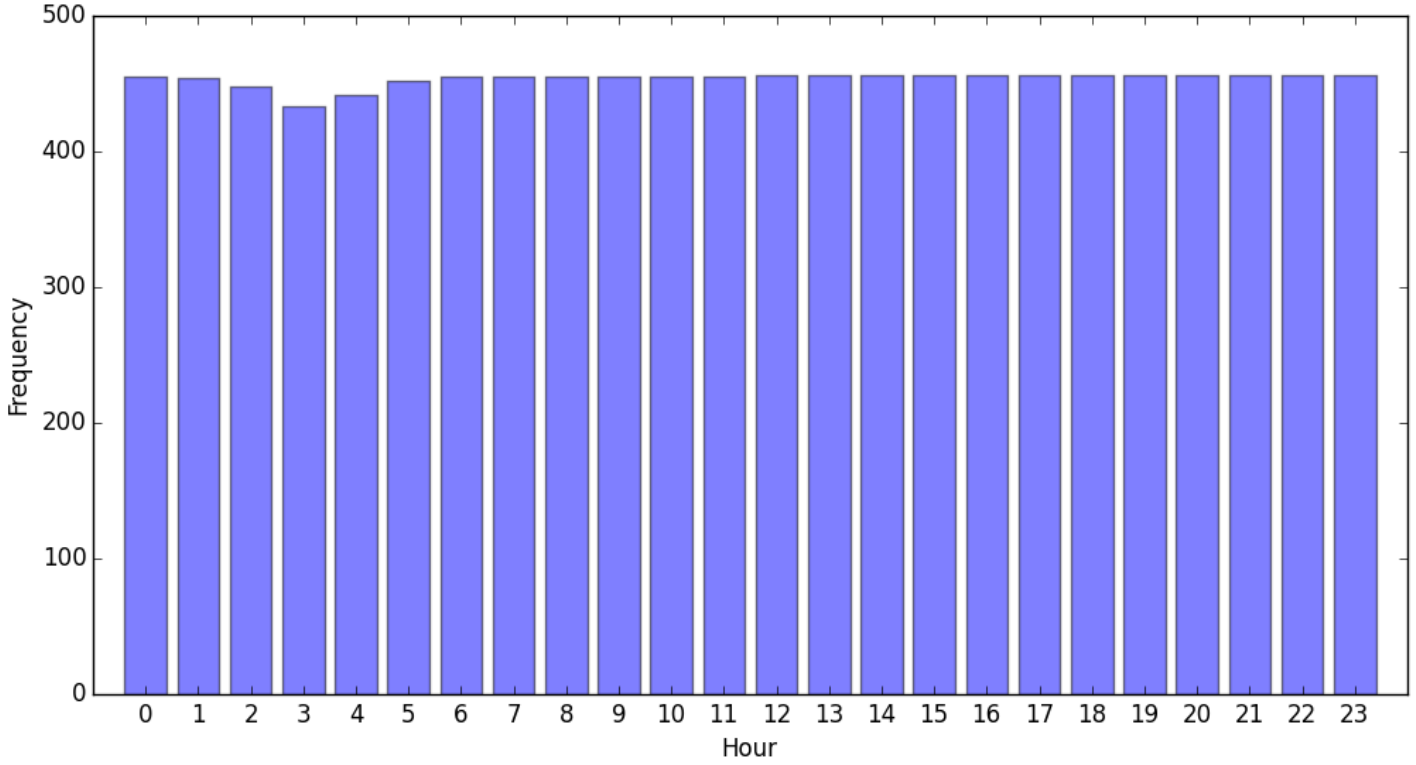
Similar issue with bike rentals during holidays





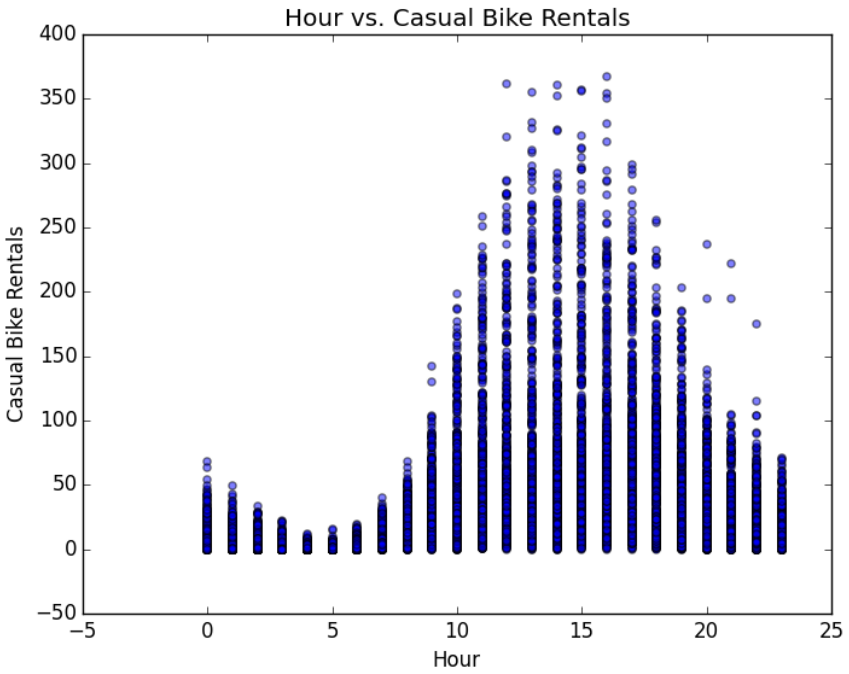
Analyzing the Data

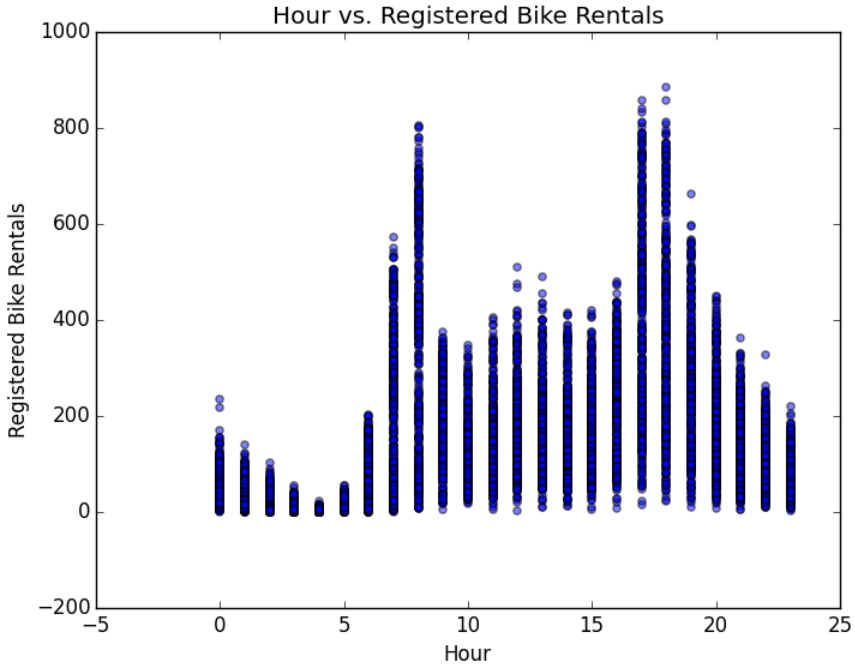
Plotted histograms of each variable

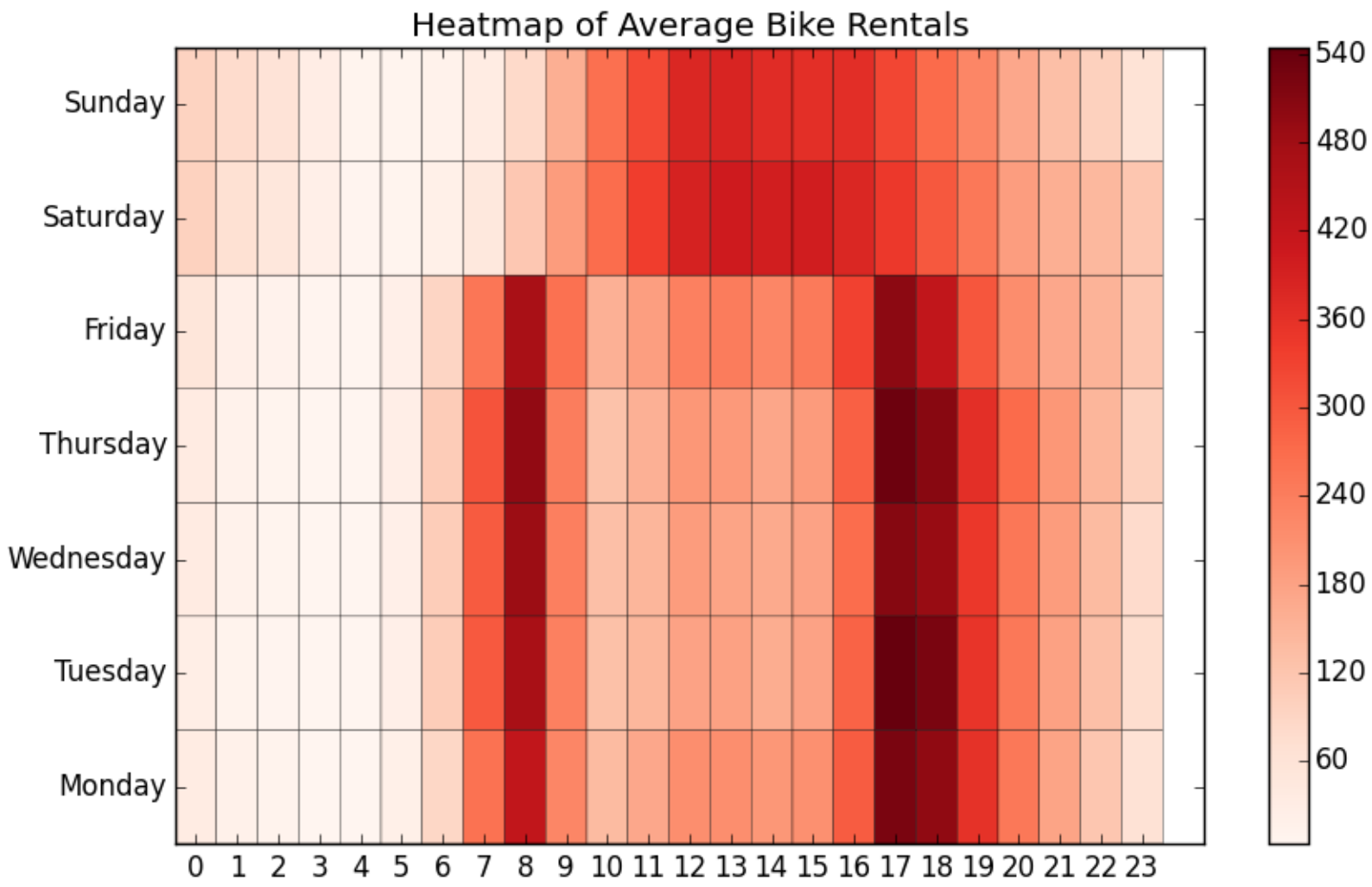


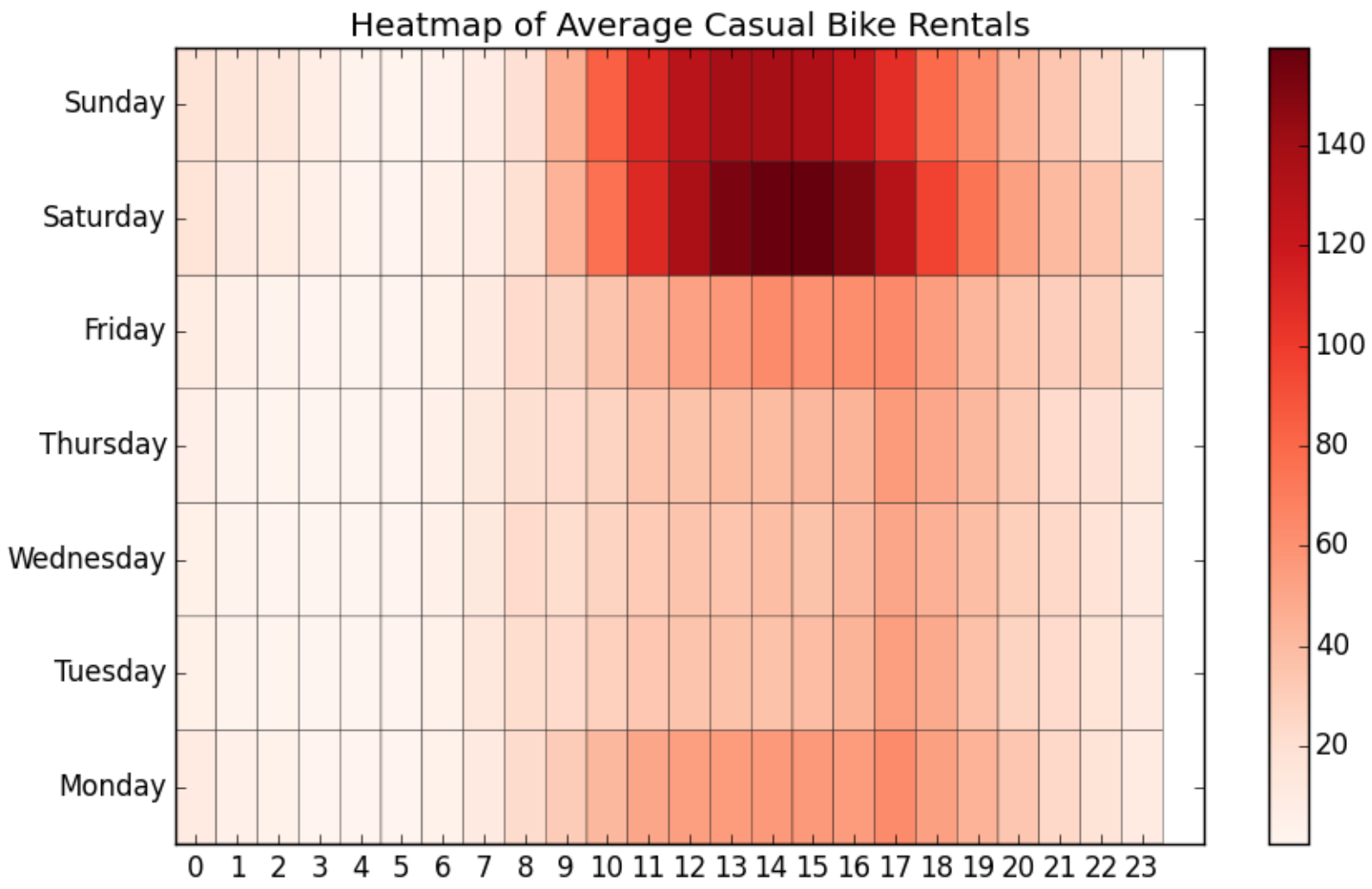
Analyzing the Data

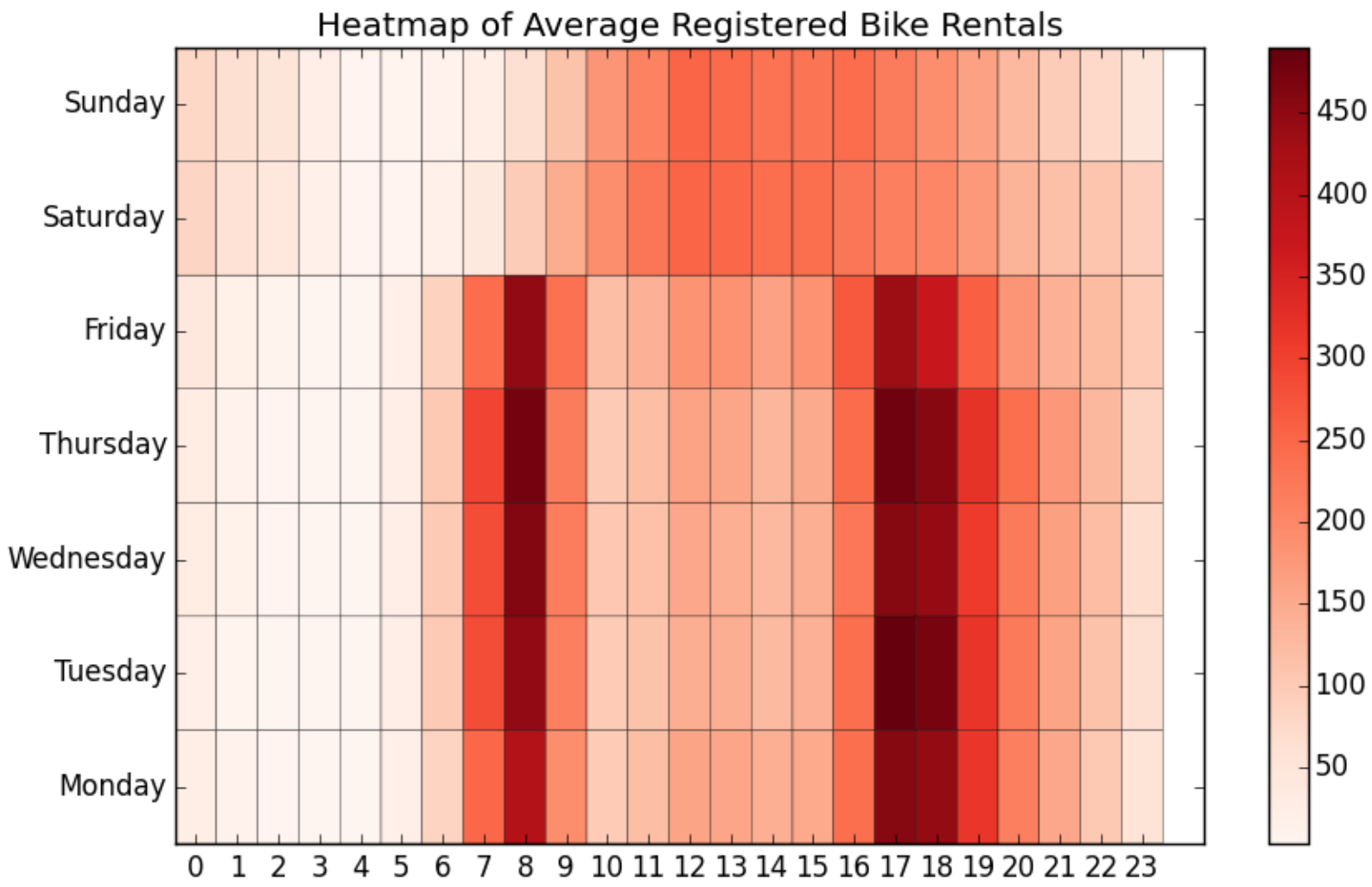
Pretty good correlation between
the time of day (hour) and bike rentals






Heatmap of Average Bike Rentals
Sunday
Saturday
Friday
Thursday
Wednesday
Tuesday
Monday
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
540
480
420
360
300
240
180
120
60


Heatmap of Average Casual Bike Rentals
Sunday
Saturday
Friday
Thursday
Wednesday
Tuesday
Monday
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
140
120
100
80
60
40
20


Heatmap of Average Registered Bike Rentals
Sunday
Saturday
Friday
Thursday
Wednesday
Tuesday
Monday
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
450
400
350
300
250
200
150
100
50

Models

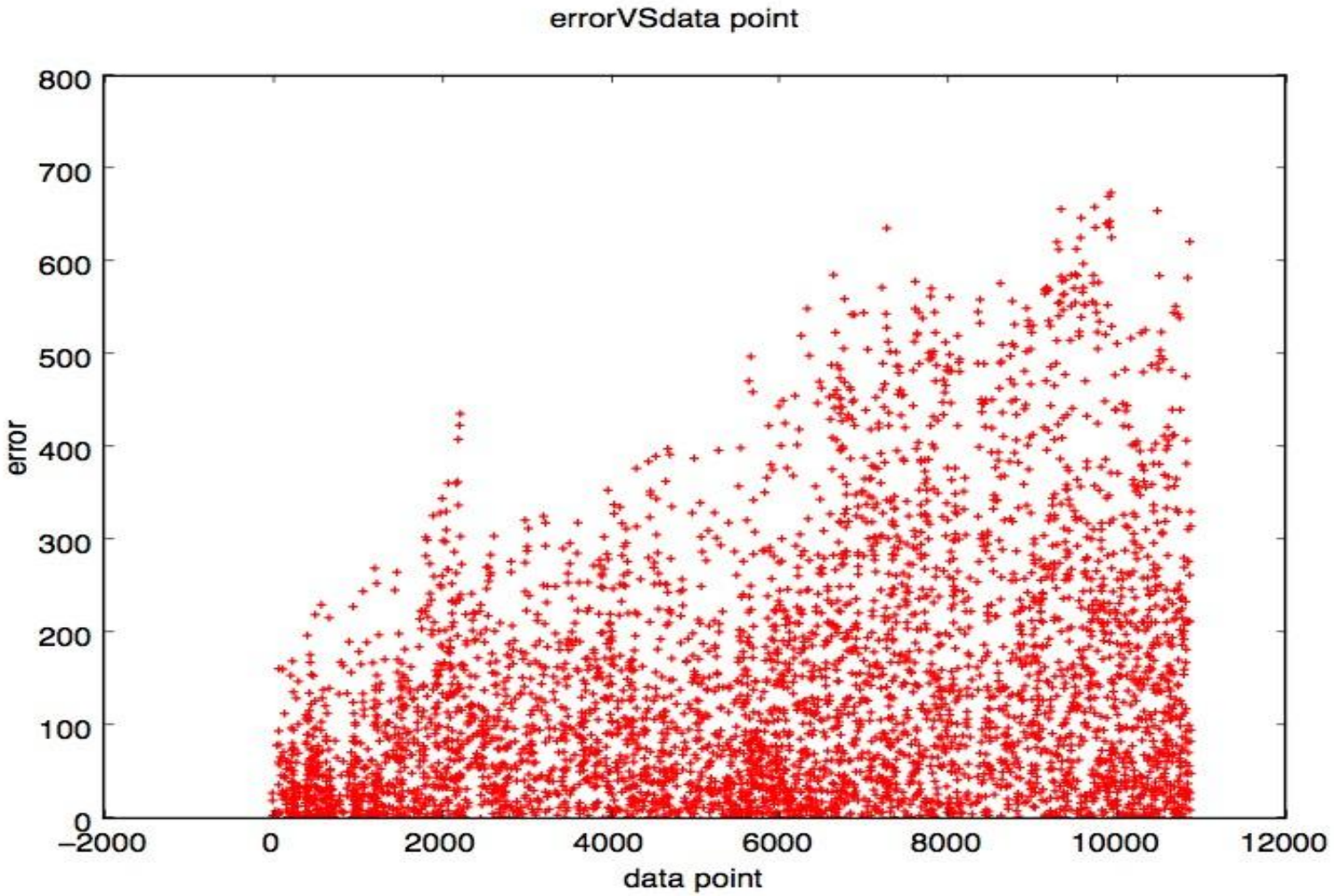
- Stochastic Gradient Descent (with standard, extracted features, e.g. hour, temperature, working day, etc.)
- Decision tree (mostly based on hour of day)
- Decision tree (using DecisionTreeRegressor)

Stochastic Gradient Descent

(with existing features)

- Standardized scaling on features for SGD application
- Learning rate used: constant and eta is specified as 0.01
- Weights calculated:[-17.95317062 27.28805465 -1.88975591 -2.03379066 1.56579281 16.34534945 45.56873864 -53.84663567 4.54432701] for attributes Time, season, holiday, working day, weather, temp, abs temp, humidity and wind speed
- Kaggle score of about 1.5

Models



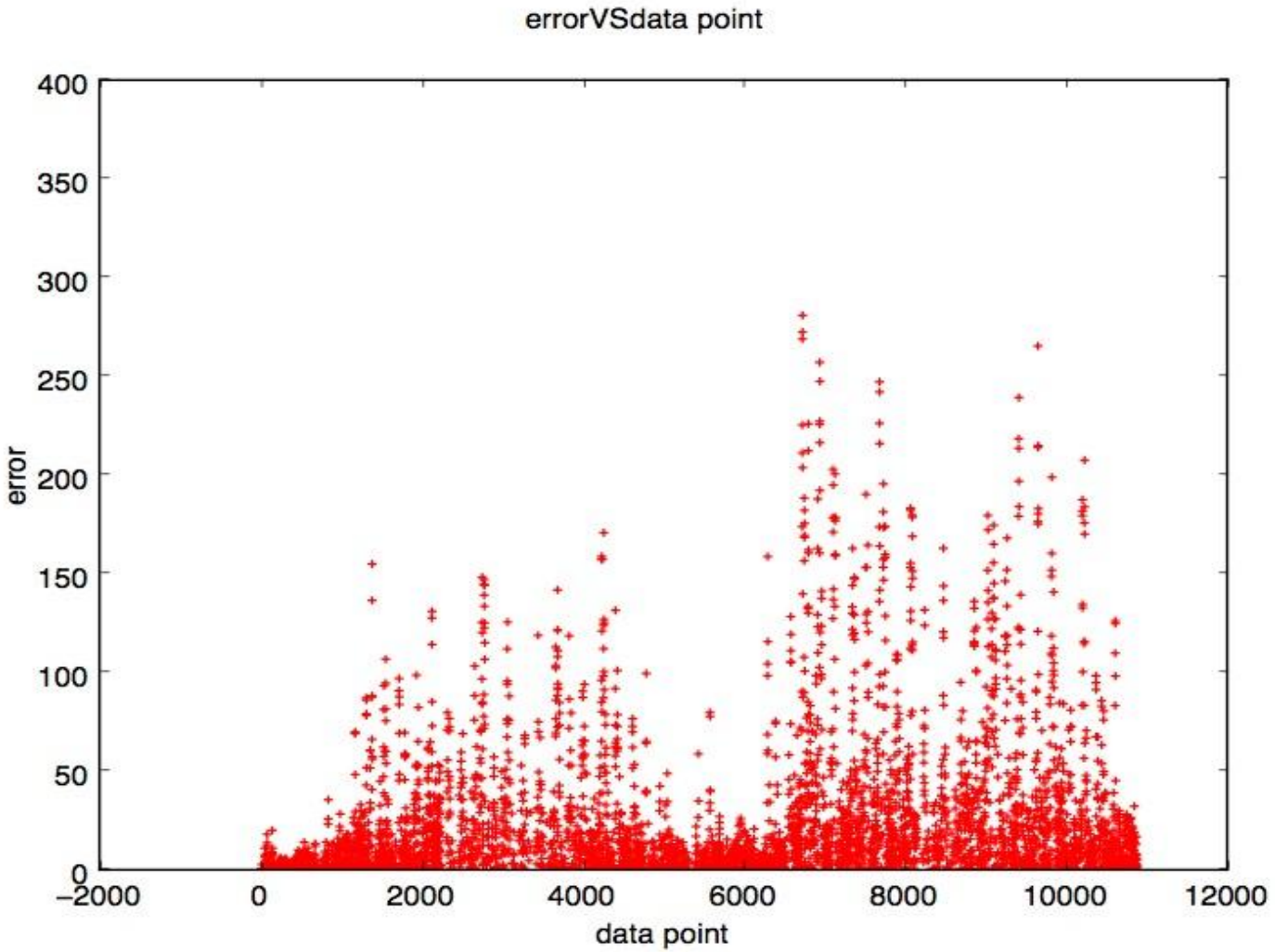
mean squared error (MSE) for SGD=23701.73795801

Stochastic Gradient Descent

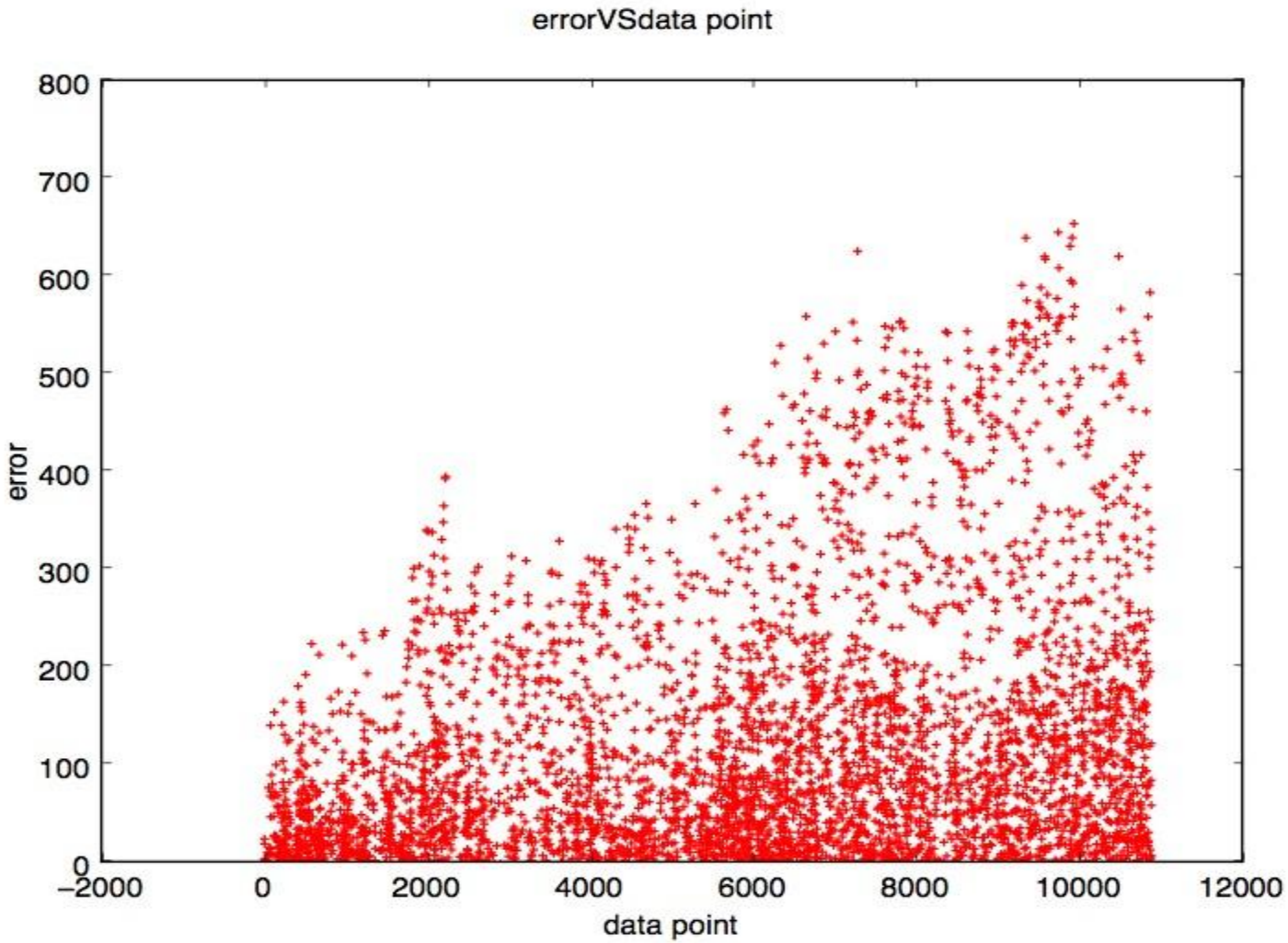
(with existing features, continued)

- If calculated separately for causal riders, Weights calculated :[-1.57859042 2.8888112 -2.55565872 -18.17153219 2.55648632 10.6053987 11.75334161 -17.1178336 0.9962883] for attributes Time, season, holiday, working day, weather, temp, abs temp, humidity and wind speed
- If calculated separately for registered riders, Weights calculated :[-1.57859042 2.8888112 -2.55565872 -18.17153219 2.55648632 10.6053987 11.75334161 -17.1178336 0.9962883] for attributes Time, season, holiday, working day, weather, temp, abs temp, humidity and wind speed
- All the three cases conclude that the dependence on weather is least
- Kaggle score of about 1.5

Models (SGD-casual)

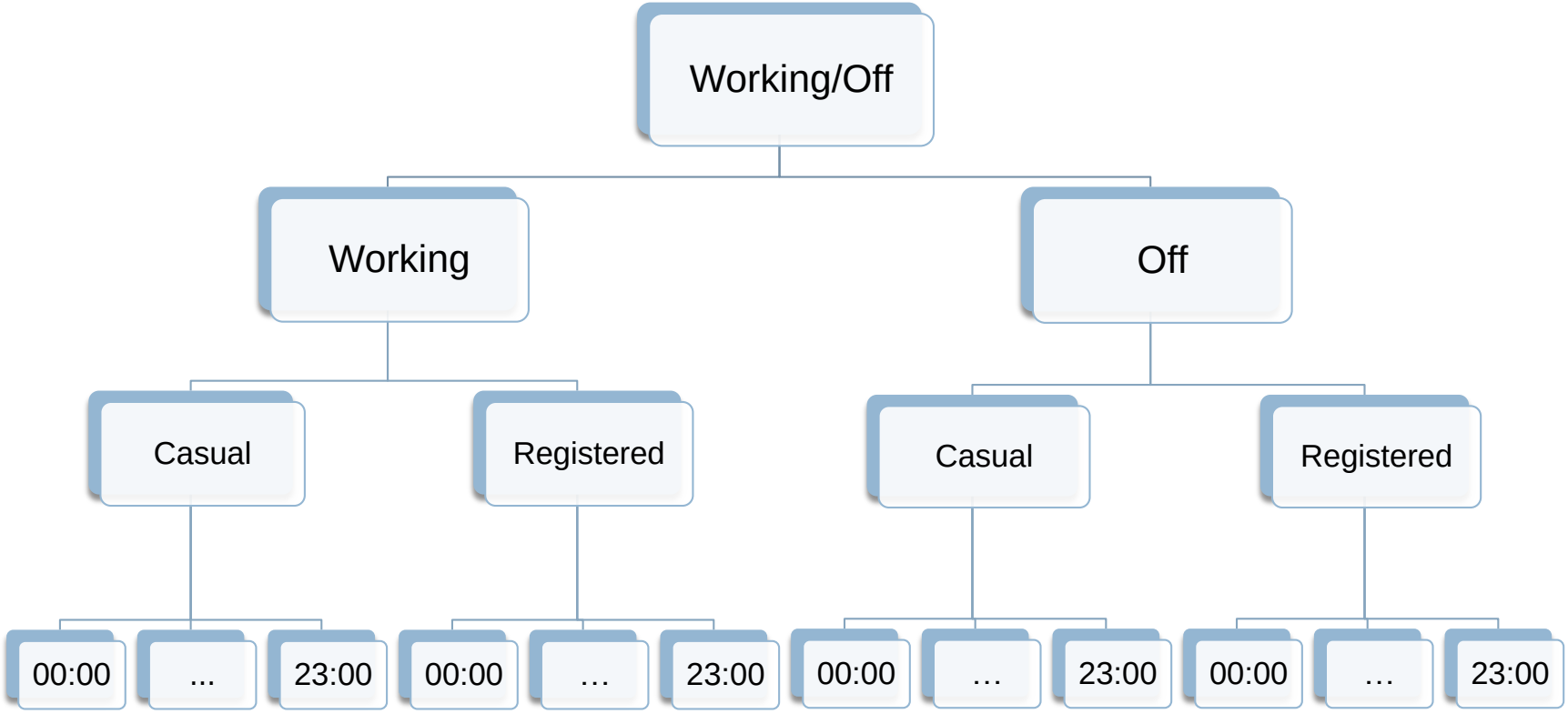


Models (SGD-registered)



mean squared error (MSE) for SGD=18105.59068184

Decision Tree (without scikit)



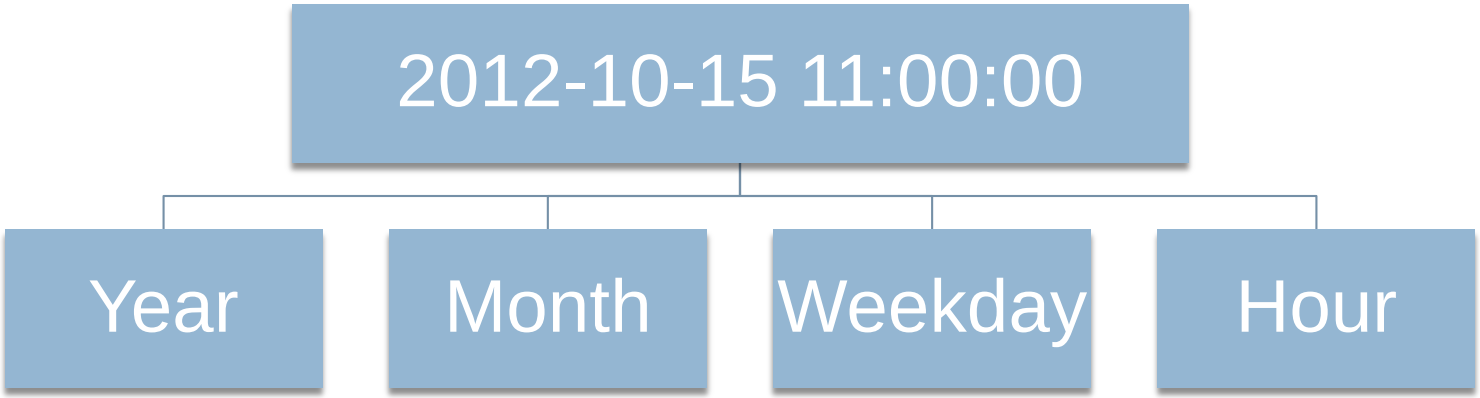
Decision Tree (without scikit)

- Work day/off day
- Casual/Registered
- Hour of the day
- Predictions based on averages (overfitting)
- Weather was not a factor
- Each new feature adds significant complexity
- Score of .69313 on Kaggle

New Features

- Initial thought process: feed multiple features into a function to generate a new feature that is more linearly correlated with the target variables
- Year
- Month
- Weekday
- Hour

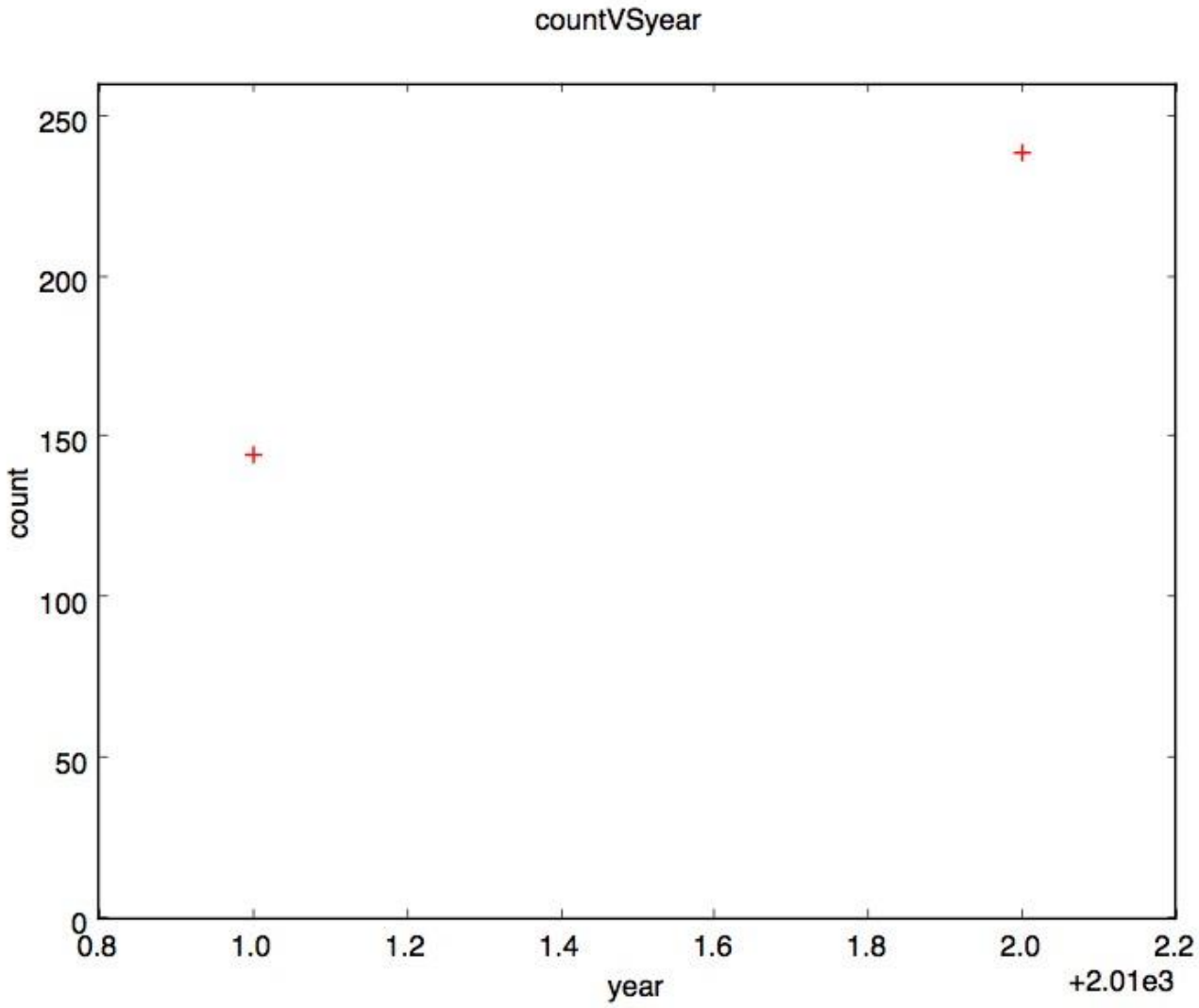
Feature Generation



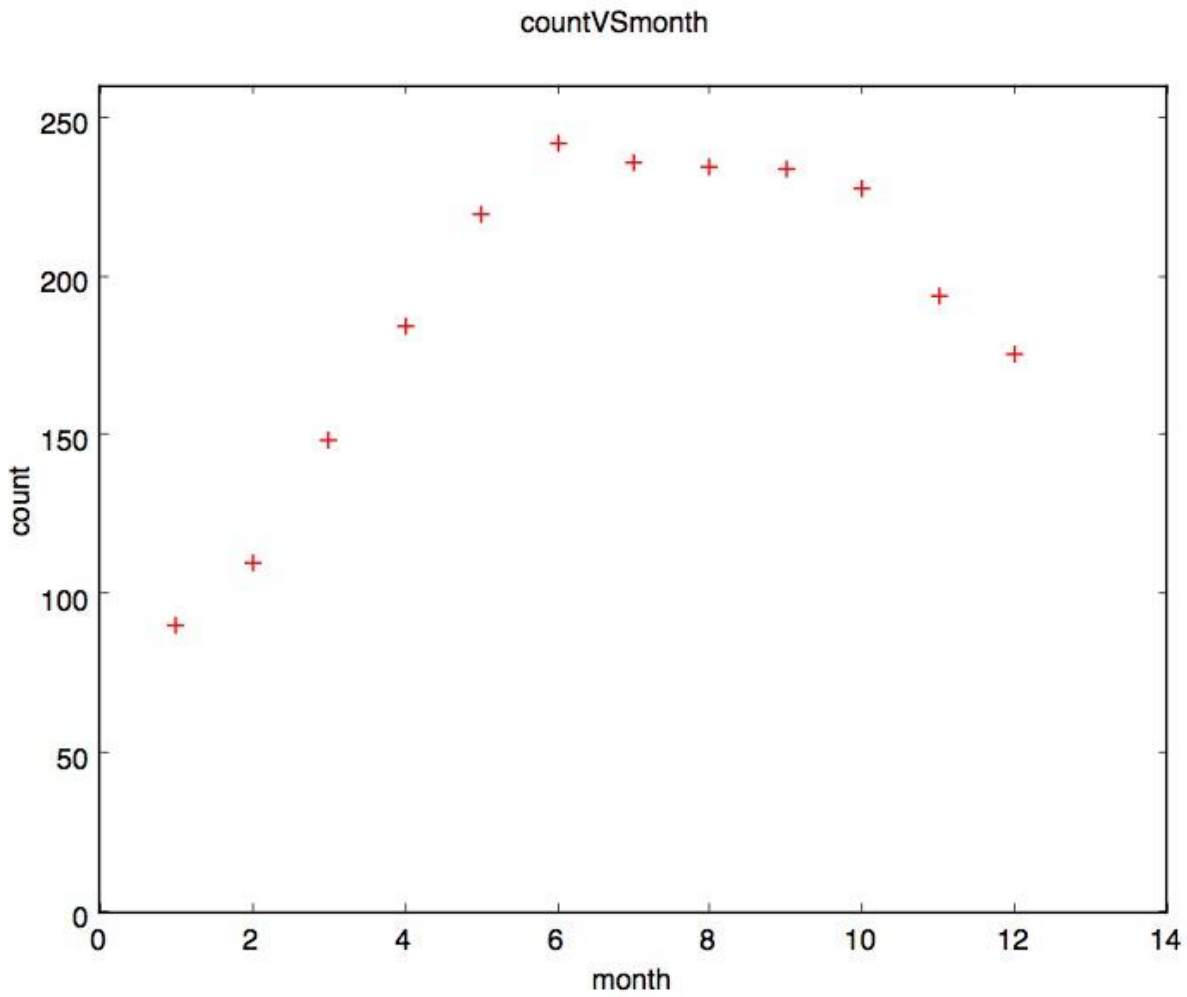
- Use Weekday instead of date, from first 20 days

we have more than 2 weeks' data to predict last 10 days' bike rent

Year feature



Month Feature



mean squared error (MSE)=0.130350909425

Decision Tree (using scikit-learn)

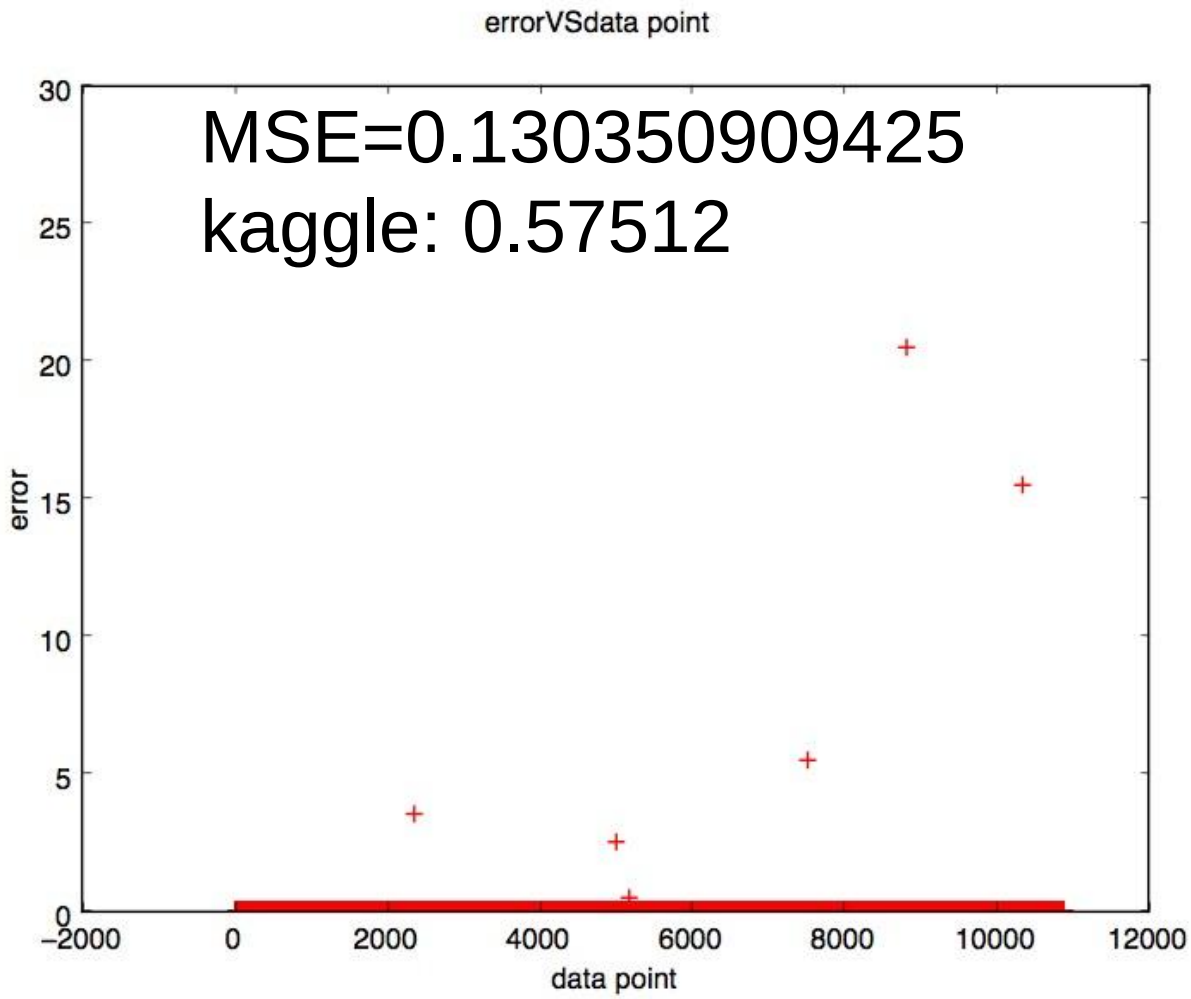
- Decision tree with 12 features

  MSE=0.130350909425

  Hour has strong correlation with count
- Decision tree without hour feature

  MSE=1006.42257334

Decision Tree (using scikit-learn)



Conclusions

- Using scikit-learn’s decision tree API allowed us to use more features
- Decision tree performed better than SGD
- SGD relies on linear dependence