Bike Share System: Data Analysis & Incentive Program Proposal

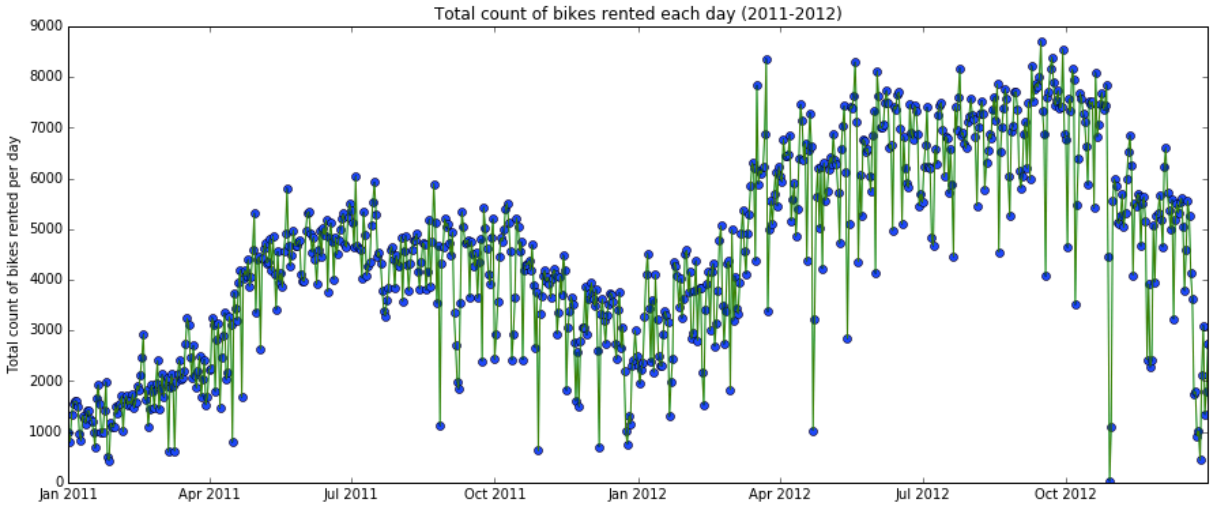
By Hannah Hagen
August 22, 2016

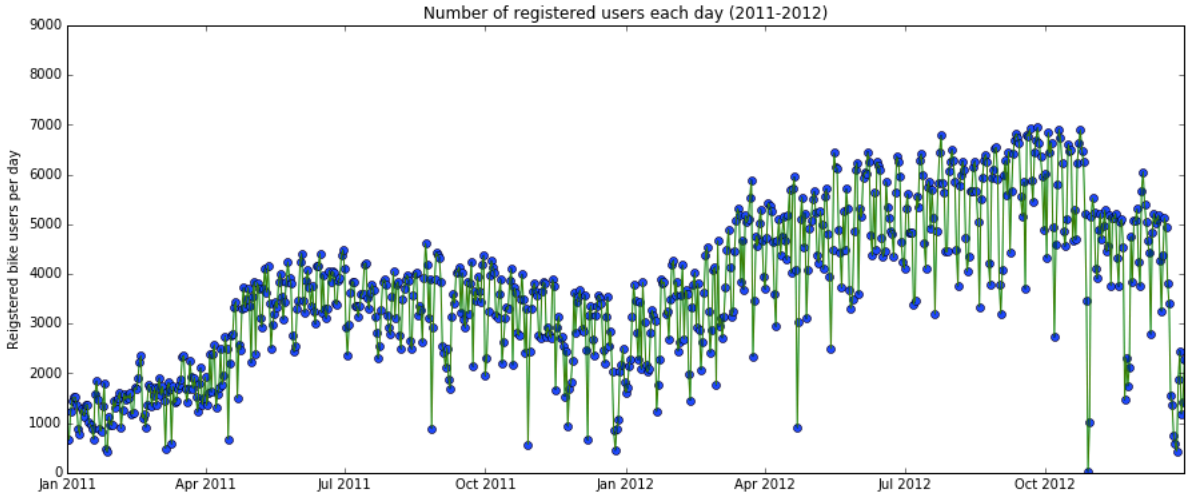
1. DATA ANALYSIS

a. Time Series Analysis

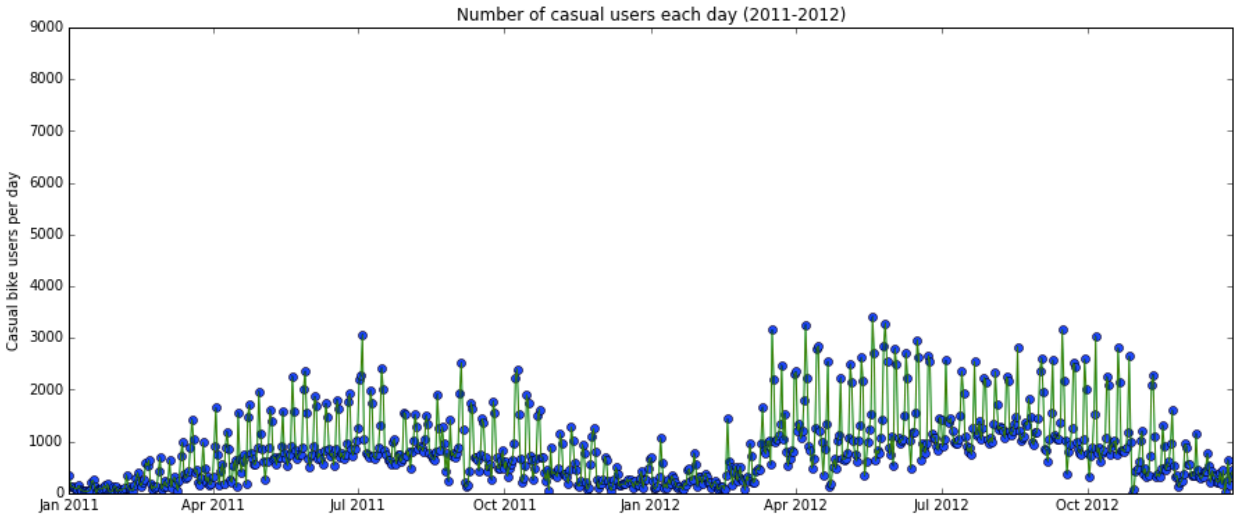
I created three time-series of:

- Total count
- Registered count
- Casual count





*In the future, I would leave out month from a multiple linear regression since it is not a linear variable but rather categorical. Therefore, the low calculate coefficient is misleading. In fact, month does have a strong impact on total count, but it is not linear. More on this later in the analysis.



From these time-series, it is evident that total count of bike rentals has increased with time. Registered users make up the majority of bike share rentals. Growth in registered users is largely responsible for the overall growth of total rentals over time.

Casual usage has grown modestly, with variability in day-to-day usage being larger in 2012 than 2011. The larger extremes seen in 2012 may suggest that more people were willing to try a bike share on certain days in 2012 than in 2011.

Takeaways:

- Registered users account for the bulk of bike share rentals (they also typically pay a membership fee) so incentive programs may want to target acquisition of registered users.
- On the flip side, the small number of casual users may mean we should work on appealing better to causal users.

b. Multiple Regression Analysis

I conducted a multiple regression analysis to obtain an overview of which independent variables (season, holiday, weather, etc.) have a significant impact on the dependent variables (number of bikes rented).

I conducted three individual multiple regression analyses using ALL independent variables (except dteday). The three predictive, or dependent variables, are total count, registered count and casual count. I chose to conduct a multiple regression analysis, as opposed to many single regression analyses, because I wanted to account for relationships between variables and compare coefficients and std error between variables when combined in a single model. In conducting a multiple regression analysis, I am making the following assumption (which may or may not be valid):

- the dependent variable really is a linear function of the independent variables, with independent and identically normally distributed errors--the coefficient estimates are expected to be unbiased and their errors are normally distributed.

*In the future, I would leave out month from a multiple linear regression since it is not a linear variable but rather categorical. Therefore, the low calculate coefficient is misleading. In fact, month does have a strong impact on total count, but it is not linear. More on this later in the analysis.

Because I am not certain this assumption is valid, a multiple regression analysis serves to indicate which variables will play a significant role in determining the predictive variable. Later, I will narrow in on key variables of interest and supplement this multiple regression analysis with tests/analyses for individual variables.

The results for each analysis are summarized below:

TOTAL COUNT

```
-------------------------Summary of Regression Analysis-------------------------

Formula: Y ~ <season> + <yr> + <mnth> + <holiday> + <weekday> + <workingday>
             + <weathersit> + <temp> + <atemp> + <hum> + <windspeed> + <intercept>

Number of Observations:         731
Number of Degrees of Freedom:   12

R-squared:         0.8002
Adj R-squared:     0.7972

Rmse:            872.4164

F-stat (11, 719):   261.8539, p-value:     0.0000

Degrees of Freedom: model 11, resid 719

-------------------------Summary of Estimated Coefficients-------------------------
      Variable       Coef    Std Err     t-stat    p-value    CI 2.5%   CI 97.5%
-------------------------------------------------------------------------------
        season   509.7752    54.7571       9.31     0.0000   402.4512   617.0992
            yr  2040.7034    65.1853      31.31     0.0000  1912.9403  2168.4665
          mnth   -38.9796    17.0791      -2.28     0.0228   -72.4545    -5.5046
       holiday  -518.9919   201.0403      -2.58     0.0100  -913.0310  -124.9529
       weekday    69.0622    16.2990       4.24     0.0000    37.1162   101.0082
-------------------------------------------------------------------------------
    workingday   120.3570    72.0073       1.67     0.0951   -20.7774   261.4914
    weathersit  -610.9870    78.3633      -7.80     0.0000  -764.5791  -457.3949
          temp  2028.9161  1403.6706       1.45     0.1488  -722.2782  4780.1104
         atemp  3573.2743  1589.3886       2.25     0.0249   458.0726  6688.4760
           hum -1018.8616   313.9952      -3.24     0.0012 -1634.2921  -403.4311
-------------------------------------------------------------------------------
     windspeed -2557.5691   456.2775      -5.61     0.0000 -3451.8731 -1663.2652
     intercept  1469.0031   240.2182       6.12     0.0000   998.1755  1939.8306
--------------------------------End of Summary---------------------------------
```

REGISTERED COUNT

*In the future, I would leave out month from a multiple linear regression since it is not a linear variable but rather categorical. Therefore, the low calculate coefficient is misleading. In fact, month does have a strong impact on total count, but it is not linear. More on this later in the analysis.

```
------------------------Summary of Regression Analysis------------------------

Formula: Y ~ <season> + <yr> + <mnth> + <holiday> + <weekday> + <workingday>
             + <weathersit> + <temp> + <atemp> + <hum> + <windspeed> + <intercept>

Number of Observations:         731
Number of Degrees of Freedom:   12

R-squared:         0.8163
Adj R-squared:     0.8135

Rmse:            673.8781

F-stat (11, 719):   290.3983, p-value:     0.0000

Degrees of Freedom: model 11, resid 719

----------------------Summary of Estimated Coefficients-----------------------
      Variable       Coef    Std Err     t-stat    p-value    CI 2.5%   CI 97.5%
--------------------------------------------------------------------------------
        season   447.9515    42.2959      10.59     0.0000   365.0515   530.8515
            yr  1754.0185    50.3509      34.84     0.0000  1655.3308  1852.7062
          mnth   -23.2353    13.1923      -1.76     0.0786   -49.0922     2.6217
       holiday  -244.7780   155.2890      -1.58     0.1154  -549.1445    59.5884
       weekday    42.6987    12.5898       3.39     0.0007    18.0227    67.3746
--------------------------------------------------------------------------------
    workingday   948.6076    55.6204      17.06     0.0000   839.5916  1057.6237
    weathersit  -497.9377    60.5299      -8.23     0.0000  -616.5764  -379.2990
          temp   834.0741  1084.2332       0.77     0.4420 -1291.0231  2959.1713
         atemp  2678.4190  1227.6869       2.18     0.0295   272.1526  5084.6853
           hum  -625.6313   242.5384      -2.58     0.0101 -1101.0066  -150.2561
--------------------------------------------------------------------------------
     windspeed -1695.5151   352.4412      -4.81     0.0000 -2386.2998 -1004.7304
     intercept   768.2098   185.5510       4.14     0.0000   404.5298  1131.8898
--------------------------------End of Summary--------------------------------
```

CASUAL COUNT

*In the future, I would leave out month from a multiple linear regression since it is not a linear variable but rather categorical. Therefore, the low calculate coefficient is misleading. In fact, month does have a strong impact on total count, but it is not linear. More on this later in the analysis.

```
-------------------------Summary of Regression Analysis-------------------------

Formula: Y ~ <season> + <yr> + <mnth> + <holiday> + <weekday> + <workingday>
             + <weathersit> + <temp> + <atemp> + <hum> + <windspeed> + <intercept>

Number of Observations:         731
Number of Degrees of Freedom:   12

R-squared:         0.6883
Adj R-squared:     0.6835

Rmse:            386.2691

F-stat (11, 719):   144.3305, p-value:     0.0000

Degrees of Freedom: model 11, resid 719

-----------------------Summary of Estimated Coefficients-----------------------
      Variable       Coef    Std Err     t-stat    p-value    CI 2.5%   CI 97.5%
--------------------------------------------------------------------------------
        season    61.8237    24.2441       2.55     0.0110    14.3051   109.3422
            yr   286.6849    28.8613       9.93     0.0000   230.1168   343.2530
          mnth   -15.7443     7.5619      -2.08     0.0377   -30.5656    -0.9230
       holiday  -274.2139    89.0122      -3.08     0.0021  -448.6777   -99.7501
       weekday    26.3635     7.2165       3.65     0.0003    12.2192    40.5079
--------------------------------------------------------------------------------
    workingday  -828.2507    31.8818     -25.98     0.0000  -890.7390  -765.7623
    weathersit  -113.0493    34.6960      -3.26     0.0012  -181.0534   -45.0452
          temp  1194.8420   621.4859       1.92     0.0549   -23.2705  2412.9544
         atemp   894.8553   703.7140       1.27     0.2039  -484.4242  2274.1348
           hum  -393.2302   139.0238      -2.83     0.0048  -665.7168  -120.7436
--------------------------------------------------------------------------------
     windspeed  -862.0540   202.0204      -4.27     0.0000 -1258.0140  -466.0941
     intercept   700.7933   106.3584       6.59     0.0000   492.3308   909.2558
---------------------------------End of Summary---------------------------------
```

Takeaways:

- Total count:
  - The rank of variables in terms of magnitude of coefficient (largest to small) is:
    - Feeling temp
    - Windspeed
    - Year
    - Temp
    - Humidity
    - Weather
    - Holiday
    - Season
    - Working day
    - Week day
    - Month *
- The standard error for each variable is a measure of the precision of the measurement of the coefficient (similar to an inverse signal-to-noise ratio). It is important to take into account the precision because this indicates how much confidence we can have in our coefficient measurement. Below is the same list of

*In the future, I would leave out month from a multiple linear regression since it is not a linear variable but rather categorical. Therefore, the low calculate coefficient is misleading. In fact, month does have a strong impact on total count, but it is not linear. More on this later in the analysis.

variables ranked by coefficient magnitude, but those with notable std error are marked.

Feeling temp (high std err)
Windspeed (medium std err)
Year
Temp (high std err)
Humidity (medium std err)
Weather
Holiday (medium std err)
Season
Working day
Week day
Month *

- For instance, for feeling temp, the variable with the coefficient of largest magnitude, the 95% confidence interval for the coefficient is 458 to 6688, quite a large range but still placing that coefficient in the top half of the list ~95% of the time.

Conclusions:

- Based on this multiple regression analysis, which again is conducted on uncertain assumptions, it appears that lean demand periods are determined largely by weather and natural conditions, including feeling temperature, windspeed, temperature, humidity and weather.
- This was against my own expectation that human-based factors, such as holiday, working day, week day would have the greatest impact. (always interesting to be proven wrong).
- Year also plays a large role according to this analysis, but we already knew that from the time series.
- Based on this knowledge, I would recommend an incentives program that encourages users to take up bike sharing during weather conditions, in which they typically refrain from bike sharing.

- The lean demand times appear to occur based on weather (not on holiday). So some kind of incentive program to encourage users to bike when sun is not out, or the windspeed is high.
- Improve the bikes for these conditions
    - Safety is a big concern in stormy conditions: tricycles, equip mirrors on bikes, lounging bikes that is more comfortable, less exposed (warmer).
    - Offer rental of helmets (with windshield) or gloves
    - For hot conditions, encourage office to have a shower at work. Partner with gym for users to shower before work.

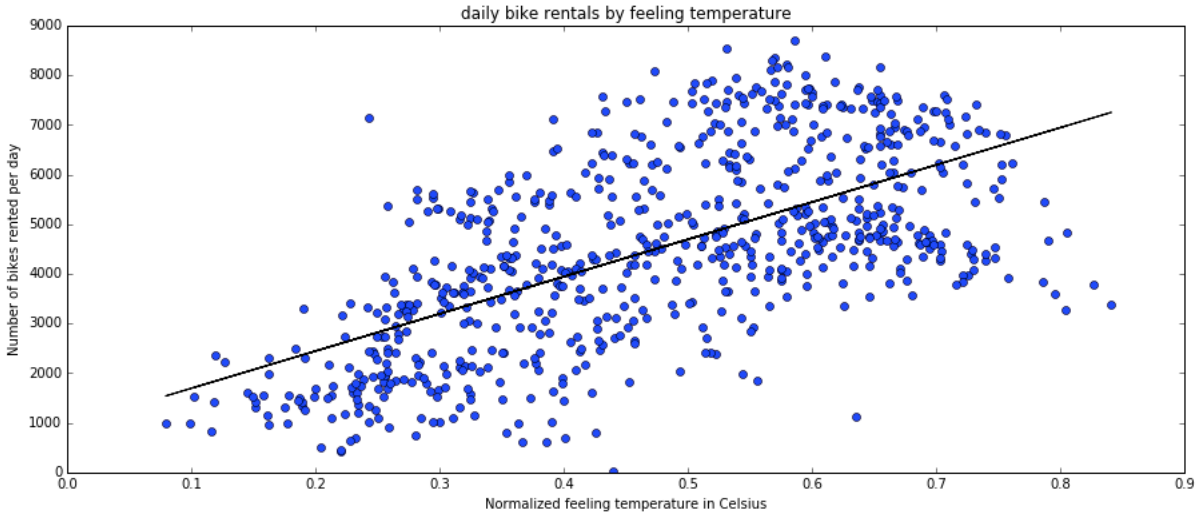
If I had more time, I would conduct a similar analysis for registered and casual users to better understand the factors that impact their behavior. For the majority of this analysis and the purposes of time, I focus on total count since this is the most universal metric of bike share usage.

c. **Investigation into individual variables: feeling temperature, wind speed, temperature, humidity and weather**

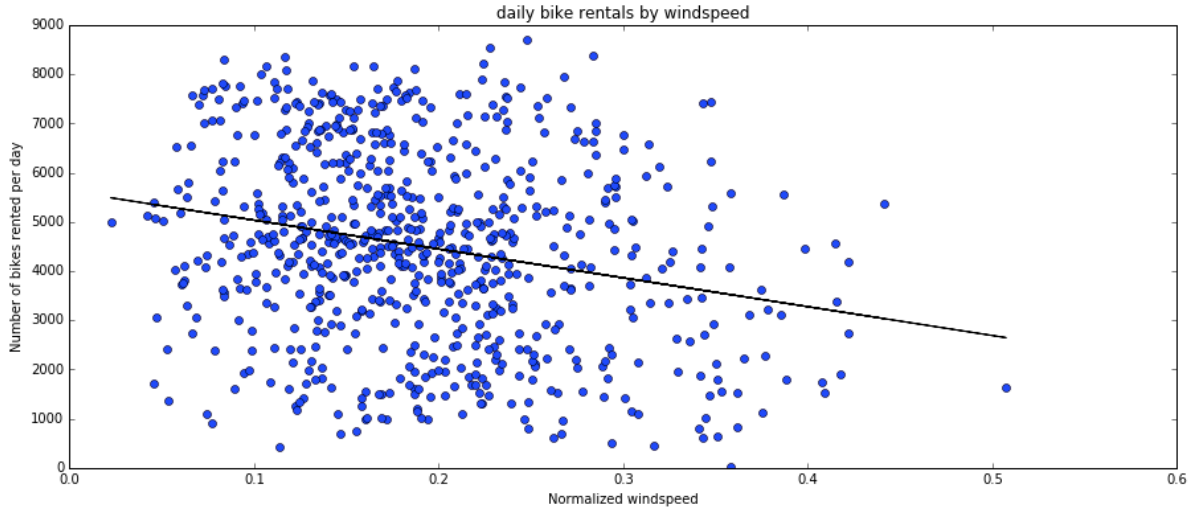
*In the future, I would leave out month from a multiple linear regression since it is not a linear variable but rather categorical. Therefore, the low calculate coefficient is misleading. In fact, month does have a strong impact on total count, but it is not linear. More on this later in the analysis.

These variables appear to be the most significant according to the multiple linear regression.

I conducted individual linear regressions for each variable.

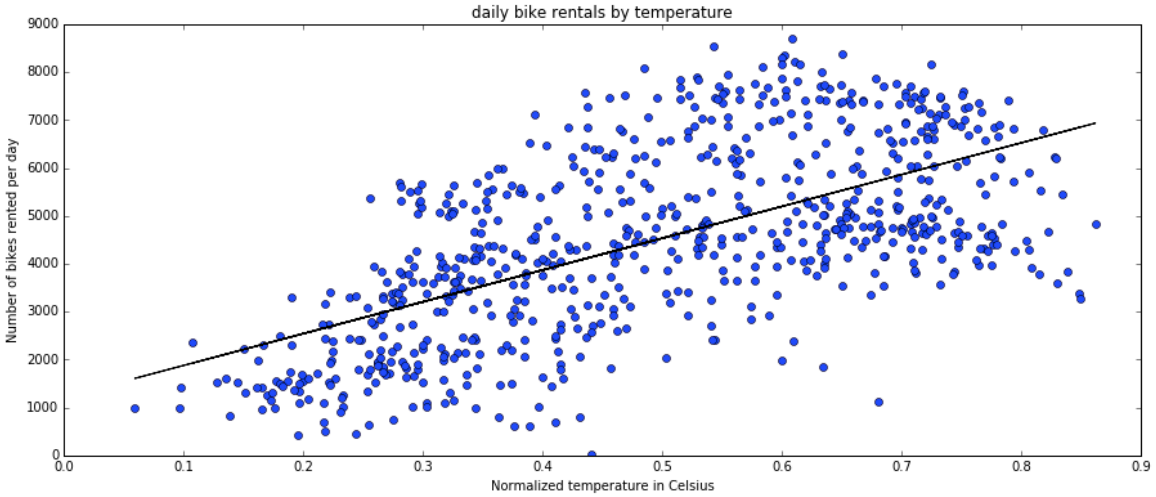


- Mostly linear, but drops off at high temperatures (convex)
- People seem more interested in bike during warmer weather. Cold weather has low rates for bike sharing.

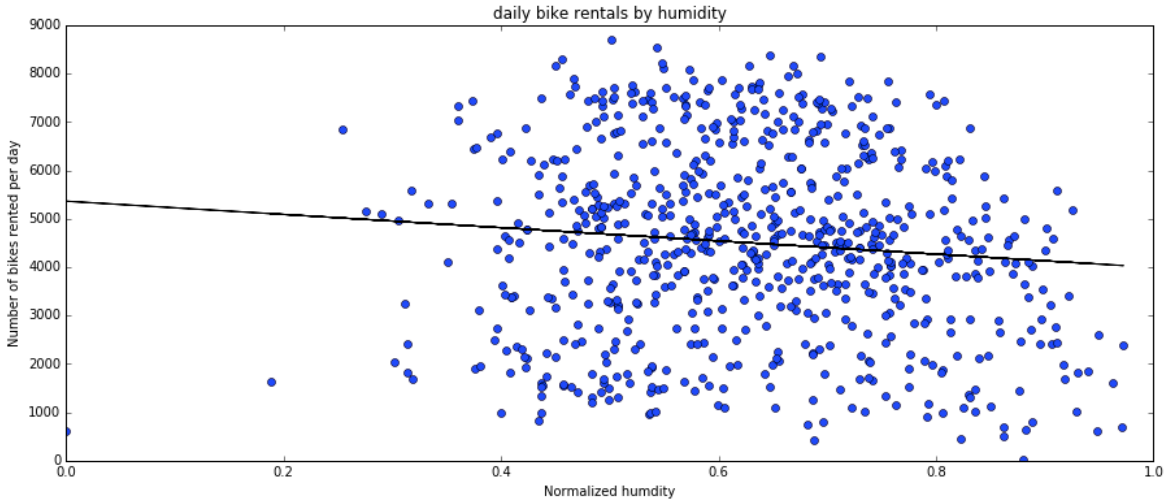


- Linear relationship here, the higher the wind speed, fewer bike shares.

*In the future, I would leave out month from a multiple linear regression since it is not a linear variable but rather categorical. Therefore, the low calculate coefficient is misleading. In fact, month does have a strong impact on total count, but it is not linear. More on this later in the analysis.

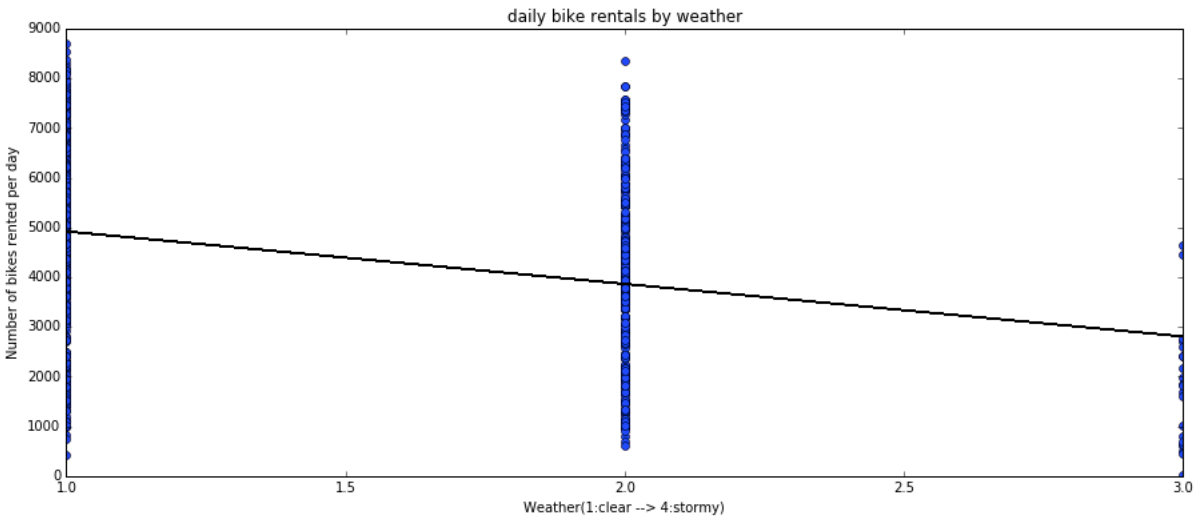


- Again convex relationship. Linear except at high temperatures (drops off)
- Warm weather sees higher bike share rates.



- Linear relationship, but not very strong. Lots of variability (std error). At high humidity, the bike share usage is slightly lower.

*In the future, I would leave out month from a multiple linear regression since it is not a linear variable but rather categorical. Therefore, the low calculate coefficient is misleading. In fact, month does have a strong impact on total count, but it is not linear. More on this later in the analysis.



Takeaways

-

Analyses I would conduct if I had more time:

- Non-linear regressions for each variable, since many of them do not have linear behavior.
- A multiple (non-linear) regression using only feeling temperature, wind speed, temperature, humidity and weather as independent variables and test the strength of this model as compared to the global, linear model.
- Create several different models (using different variables for multiple regression). I would train the model on 80% of the data and use the other 20% of the data as a validation data set to measure the model's ability to predict new data points, as well as measure the model's error.

Lastly, I want to circle back around on my investigation of the impact of month on bike share usage since the multiple linear regression was an inappropriate way to assess the impact of month. First I create a scatter plot. Here you can see that

*In the future, I would leave out month from a multiple linear regression since it is not a linear variable but rather categorical. Therefore, the low calculate coefficient is misleading. In fact, month does have a strong impact on total count, but it is not linear. More on this later in the analysis.