**Bike Share System: Data Analysis & Incentive Program Proposal**

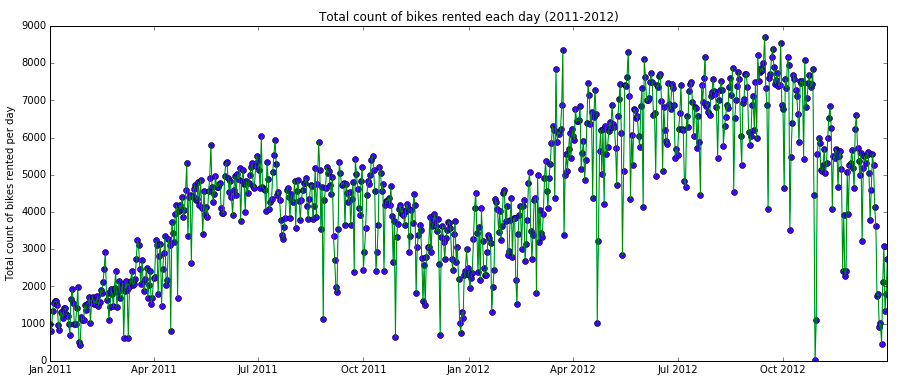
By Hannah Hagen

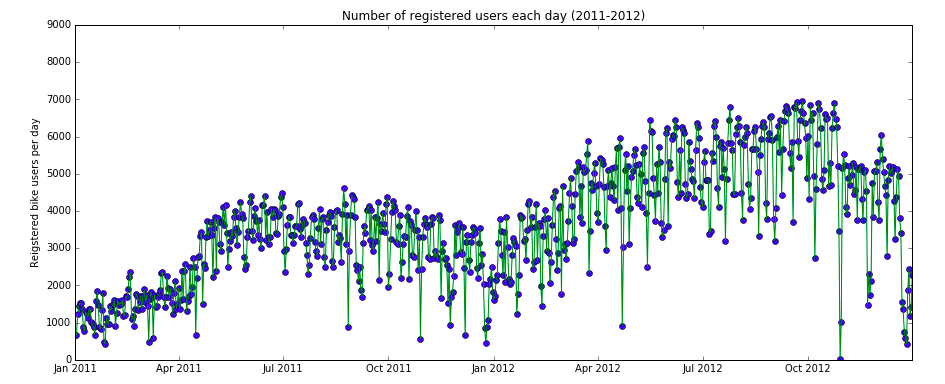
August 22, 2016

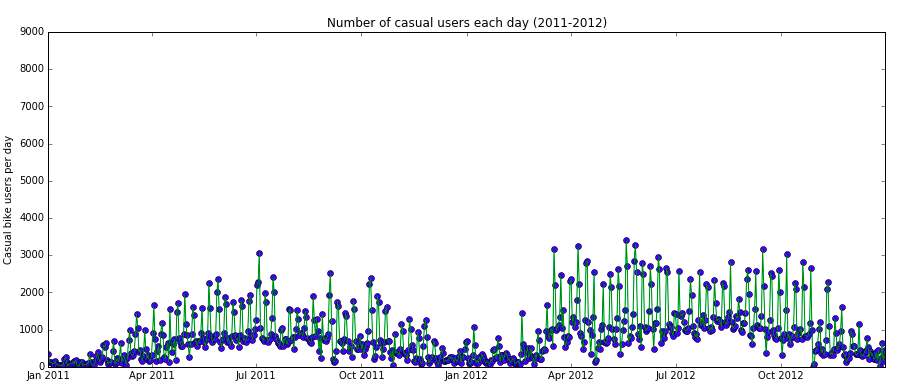
1. **DATA ANALYSIS**
   1. **Time Series Analysis**

I created three time-series of:

* + - Total count
    - Registered count
    - Casual count







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.
  1. **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).

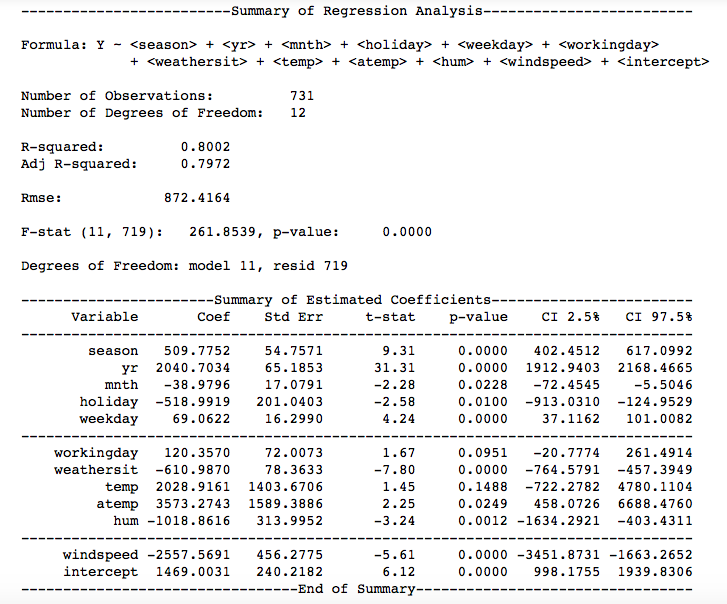
I conducted three individual multiple regression analyses using ALL independent variables (except dteday). The three predictive, or dependent variables, are total count, registered count and casual count. I chose to conduct a multiple regression analysis, as opposed to many single regression analyses, 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.

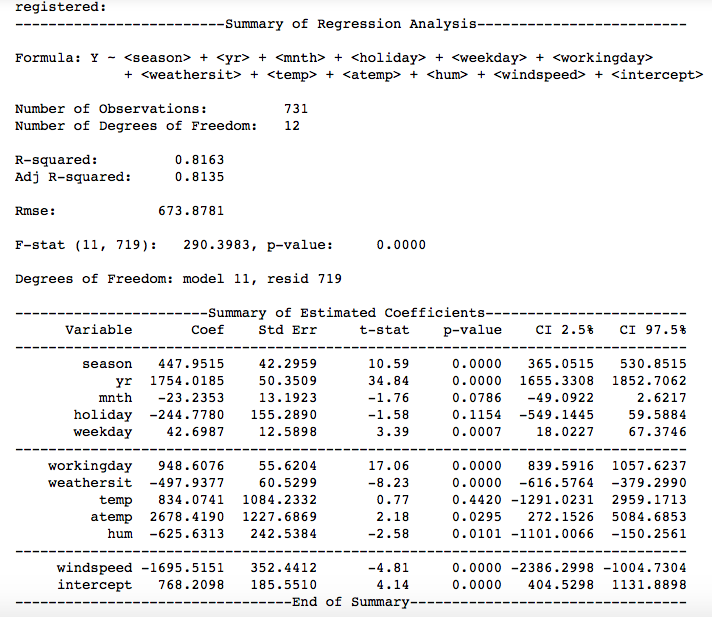
Because I am not certain this assumption is valid, a multiple regression analysis serves to indicate which variables will play a significant role in determining the predictive variable. Later, I will narrow in on key variables of interest and supplement this multiple regression analysis with tests/analyses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

The results for each analysis are summarized below:

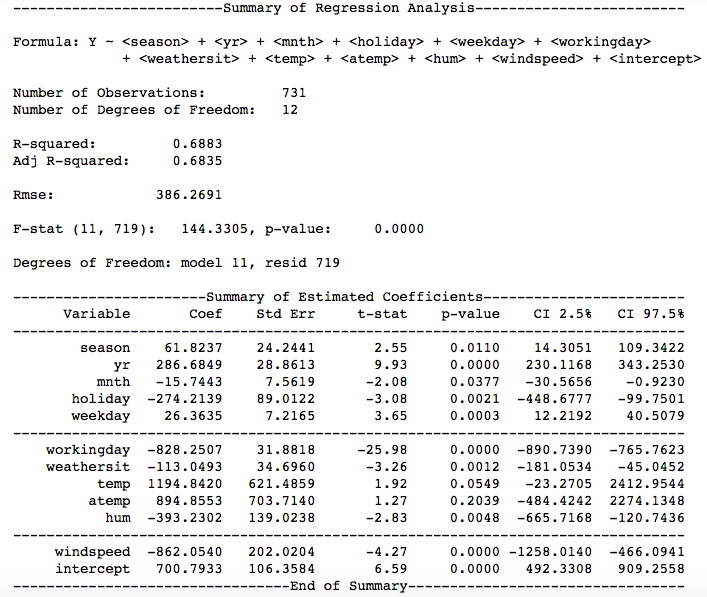
TOTAL COUNT



REGISTERED COUNT



CASUAL COUNT



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 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:

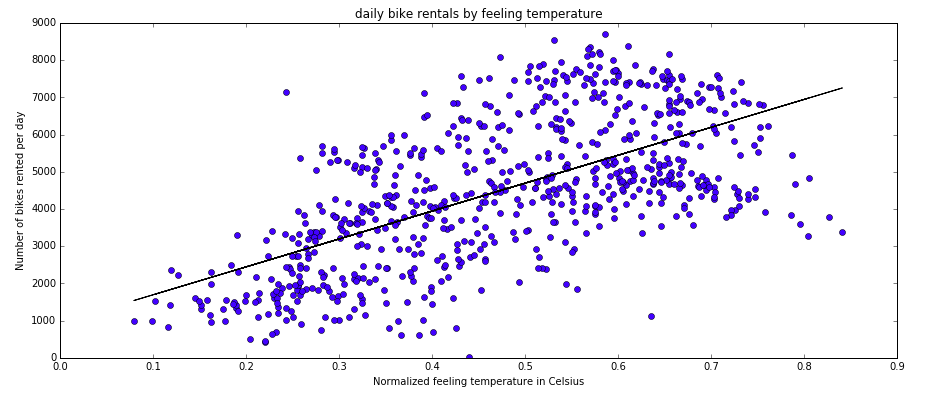
* Based on this multiple regression analysis, which again is conducted on uncertain assumptions, it appears that lean demand periods are determined largely by weather and natural conditions, including feeling temperature, windspeed, temperature, humidity and weather.
* This was against my own expectation 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.

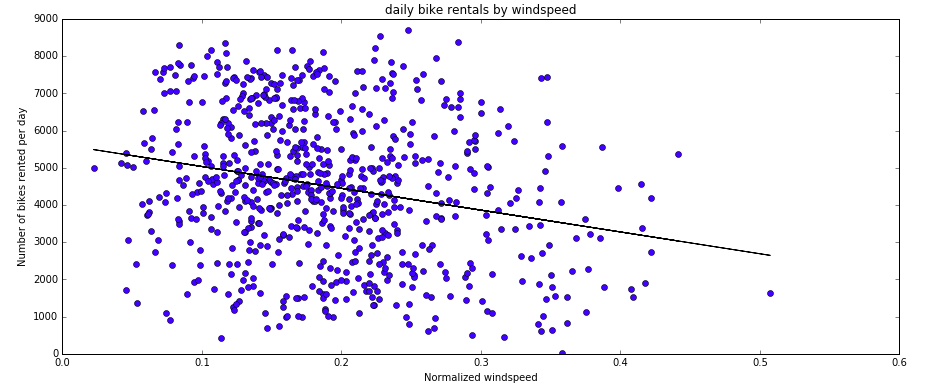
* 1. **Investigation into individual variables: feeling temperature, wind speed, temperature, humidity and weather**

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

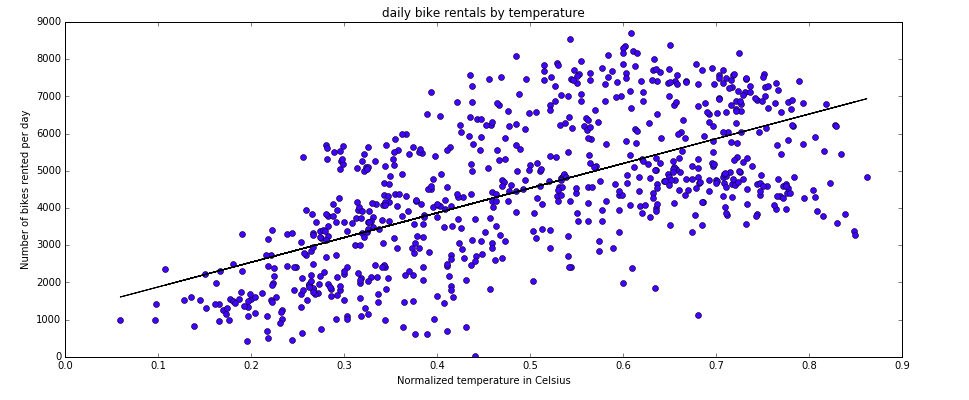
I conducted individual linear regressions for each variable.



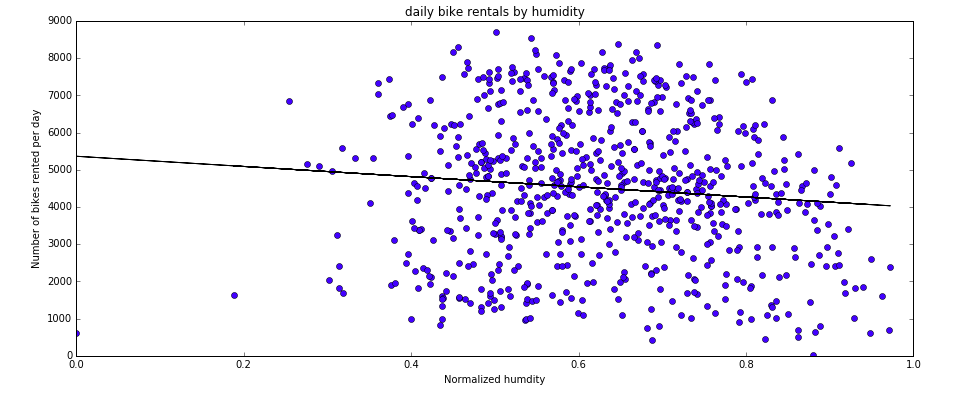
* Mostly linear, but drops off at high temperatures (convex)
* People seem more interested in bike during warmer weather. Cold weather has low rates for bike sharing.



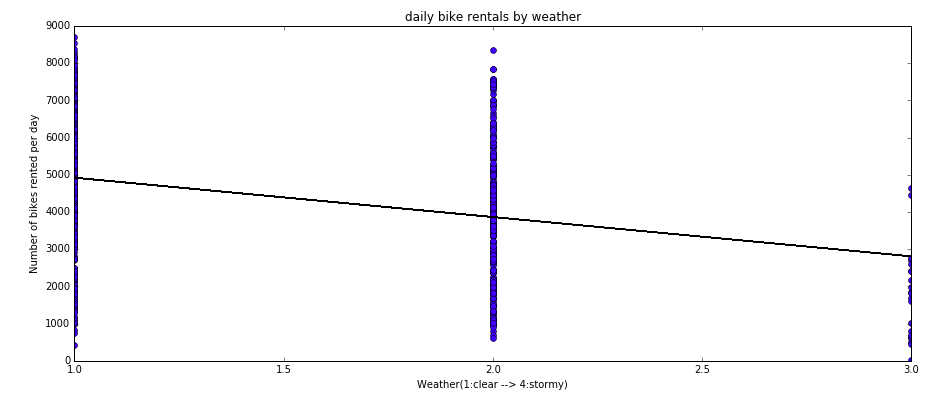
* Linear relationship here, the higher the wind speed, fewer bike shares.



* 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.



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