Bike Share System: Data Analysis & Incentive Program Proposal

By Hannah Hagen
August 22, 2016

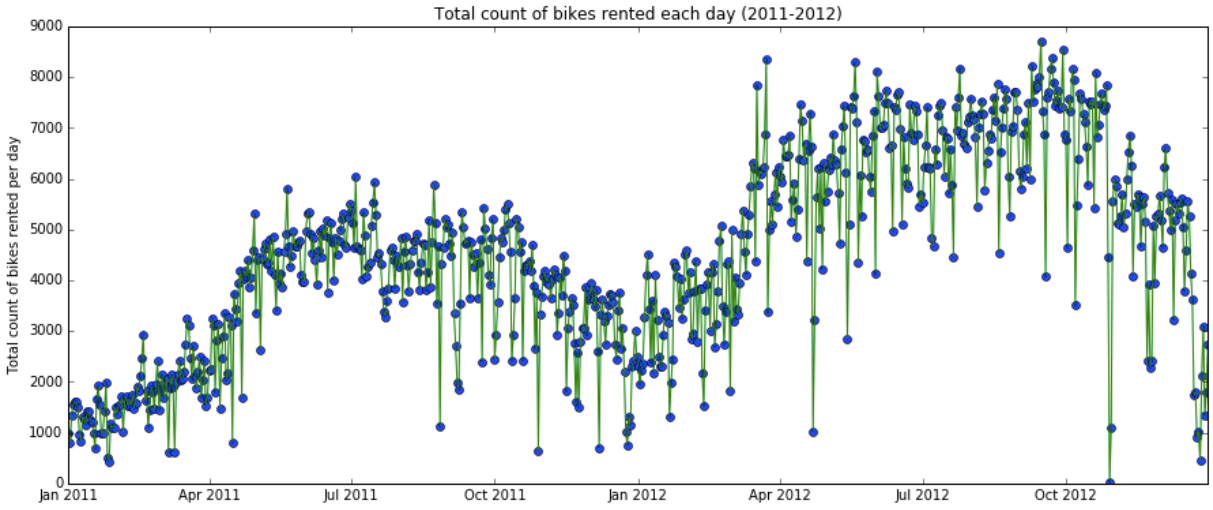
Overview: this report is organized into two parts: 1) data analysis, which includes visualizations, as well as discussions and notes on further tests/analyses I would conduct given additional time, and 2) overview of incentive program. The graphs are created using numpy and the code is included in the file analysis.ipynb.

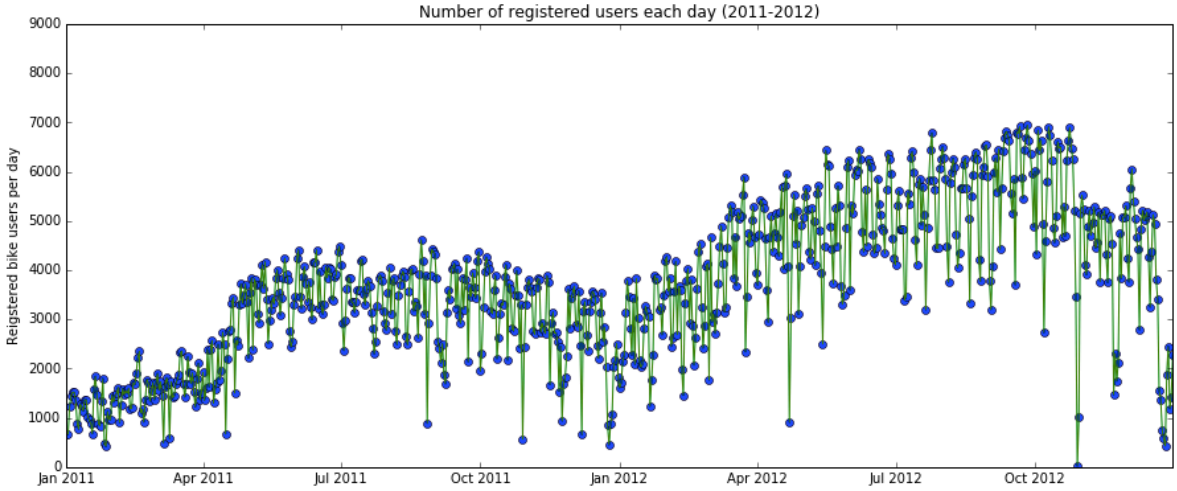
1. DATA ANALYSIS

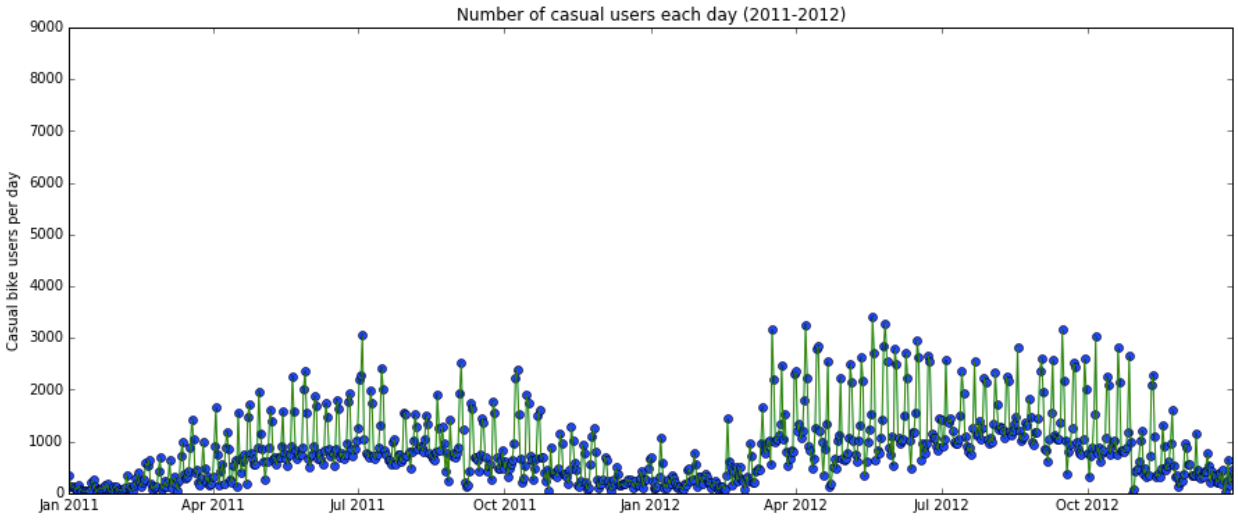
a. Time Series Analysis

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.

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 because I wanted to account for relationships between variables and compare coefficients and std error between variables when combined in a single, global 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. But the results (coef, std err) cannot be assumed to be correct. 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 analyses 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

```
------------------------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

```
------------------------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:

For the purpose of time, I focus my analysis on the results from the 'total count' regression analysis:

- The rank of variables in terms of magnitude of coefficient (largest to small) is:
    1) Feeling temp
    2) Wind speed
    3) Year
    4) Temp
    5) Humidity
    6) Weather
    7) Holiday
    8) Season
    9) Working day
    10) Week day
    11) Month
    - Note: If I did this again, I would leave out 'month' from a multiple linear regression since it is not a linear variable but rather categorical. Therefore, the low 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.

- 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. We could proceed with caution when basing conclusions on the coefficient measurements with high standard error.
    1) **Feeling temp (high std err)**
    2) Windspeed (medium std err)
    3) Year
    4) **Temp (high std err)**
    5) Humidity (medium std err)
    6) Weather
    7) Holiday (medium std err)
    8) Season
    9) Working day
    10) Week day
    11) Month *
    - Despite the high std err for feeling temp, its 95% confidence interval is 458 to 6688, which, despite being a large range, still places it above most other coefficients.

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, wind speed, 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.

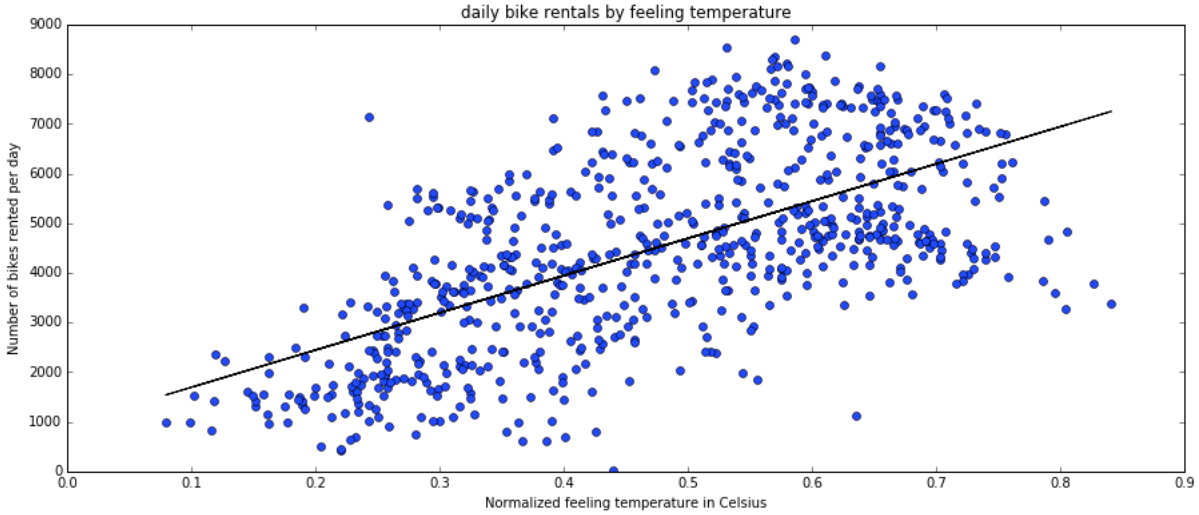
If I had more time, I would investigate the results for registered and casual users to better understand the factors that impact their behavior. For instance, for registered users it appears that rental rates are positively correlated with working day (more rentals on working days), whereas casual users are negatively correlated with working day (more casual rentals when it is not a working day).

I focused this analysis on total count since this is the most universal metric of bike share usage.

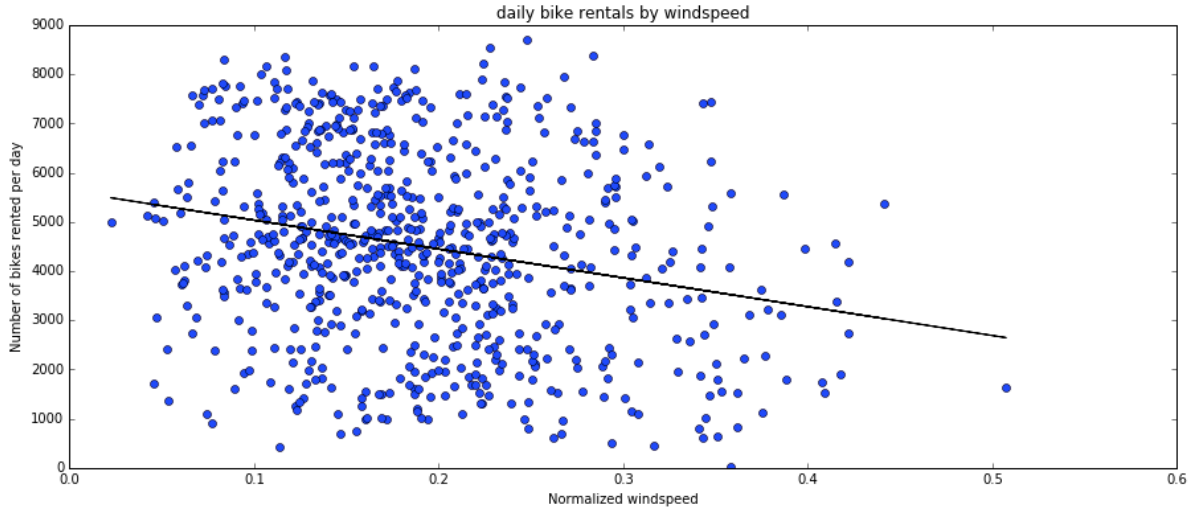
c. **Investigation into individual, influential variables: feeling temperature, wind speed, temperature, humidity and weather**

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

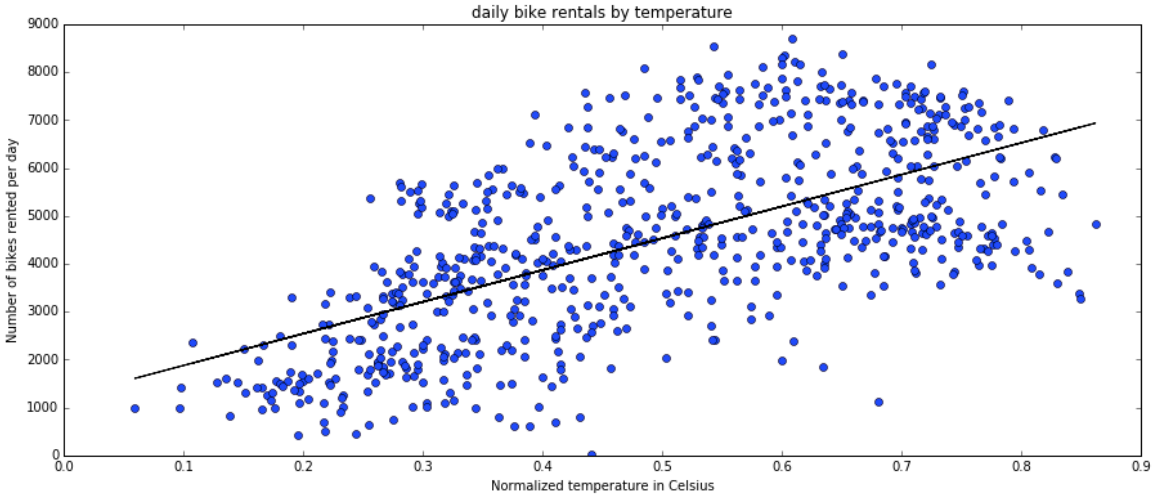
I conducted individual linear regressions for each variable.



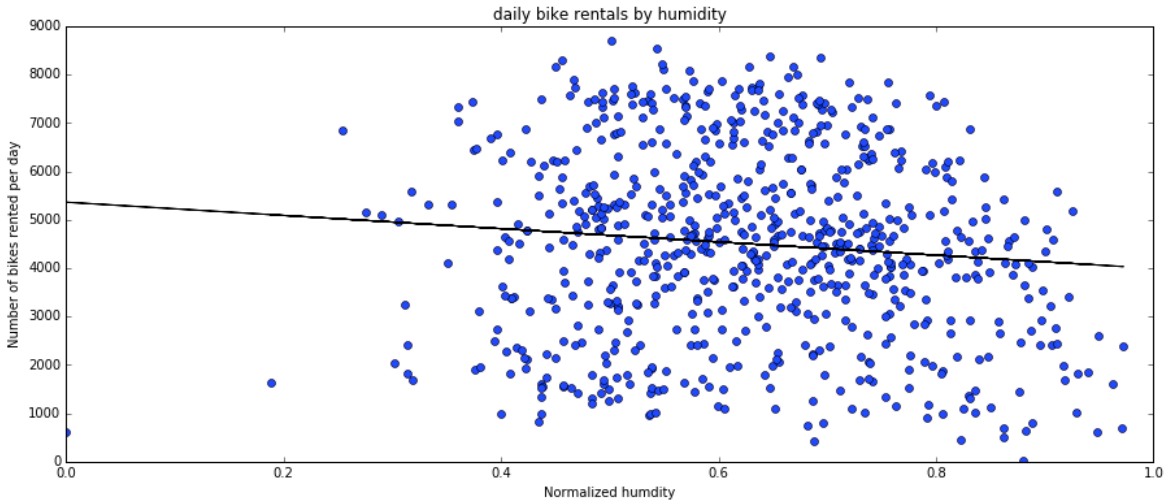
- Mostly linear, but drops off at high temperatures (convex)
- People seem more interested in bike during warmer weather. Cold weather has low rates for bike sharing.



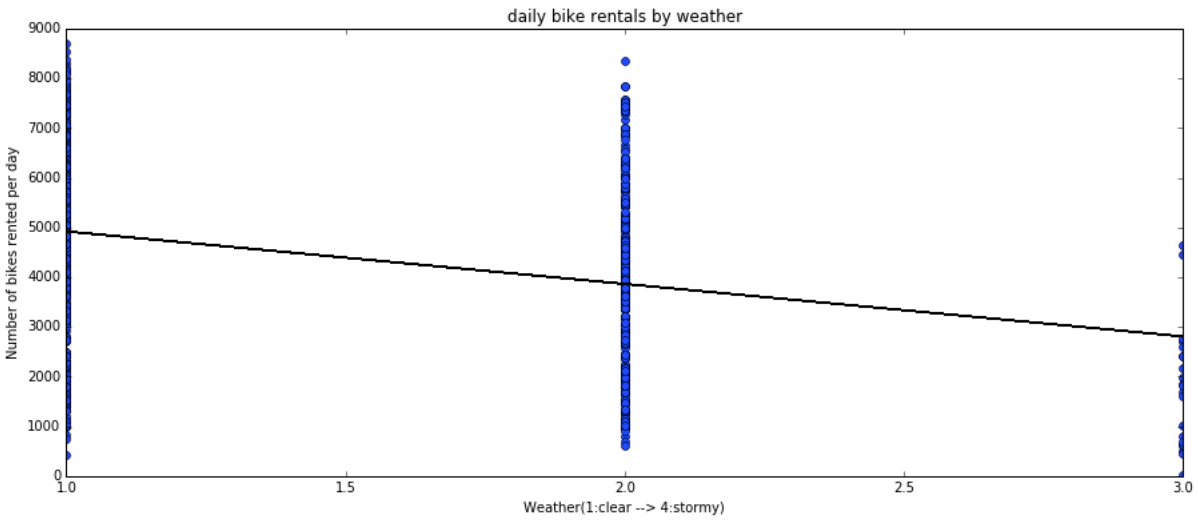
- Linear relationship: the higher the wind speed, fewer bike shares.



- 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.



- Stormier weather is correlated with fewer bike share rentals.

Takeaways

- High bike share rates are correlated with warm (but not hot) temperature, warm (but not hot) feeling temperature, low wind speed, low humidity and clear sky weather.

Conclusions

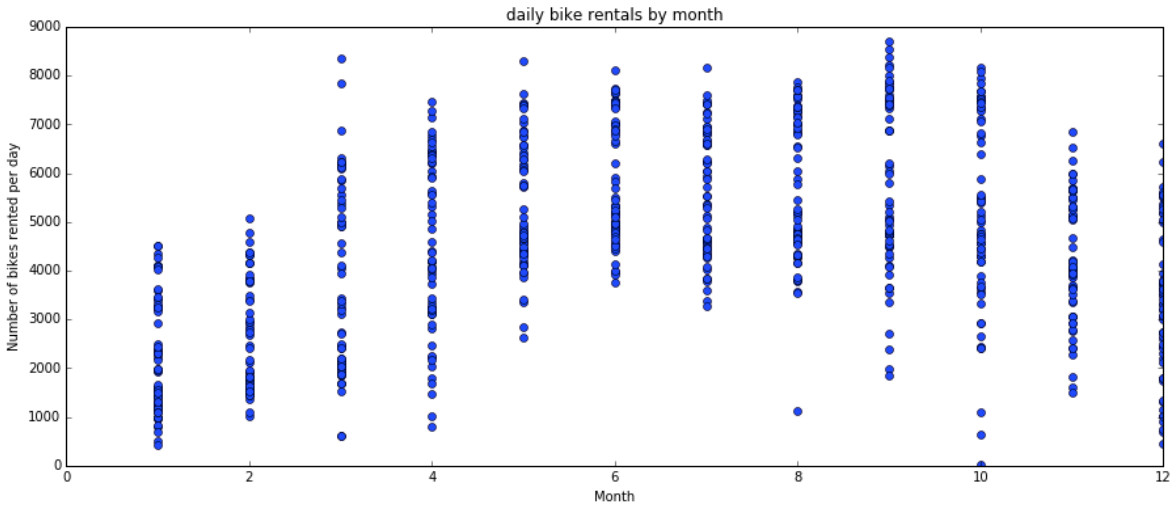
- I would create a program that incentivizes bike sharing during the conditions of low temperature, high wind speed, high humidity or stormy weather.

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.

d. Effect of Month

Since the linear regression was an inappropriate way to assess the impact of month, I conduct an analysis here, creating a scatter plot of rental rates by month.

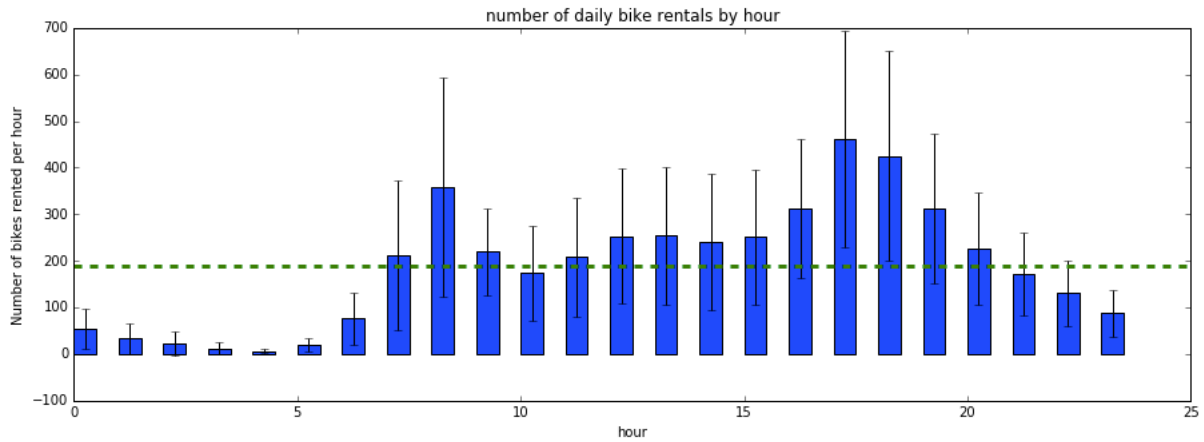


Here you can see that bike share rates are highest May-August. This corresponds with higher bike rates correlated with warm, clear weather as we might expect from the previous analysis (assuming this is a location in the Northern hemisphere). Based on this data, I would consider creating an incentive plan for the fall and winter months (Sep – April). Perhaps the membership has a 10% lower monthly cost in Sep-April than May-July to encourage more bikers in these under-utilized months. Additionally, the agency could sign up individuals for year-round memberships to encourage users to stay active biking all year. This could be incentivized by having a reduced yearly membership rate, as compared to the monthly rate, which the bike share program in DC currently offers. This may encourage members to continue using their membership throughout the year, even in the colder months.

Some non-monetary incentives could include free friend/guest passes during Sept-April. This means that members could take a friend on a bike trip with them for free, in order to expose more people to the bike share program, while taking advantage of under-utilized bikes in these months.

Other incentive ideas could include using games like Pokémon Go to encourage people to bike around a neighborhood and quickly catch Pokémon. Perhaps you could have a "gym" positioned at a bike station. Pokémon Go currently does this with business establishments to attract customers to business establishments.

e. Effect of time of day

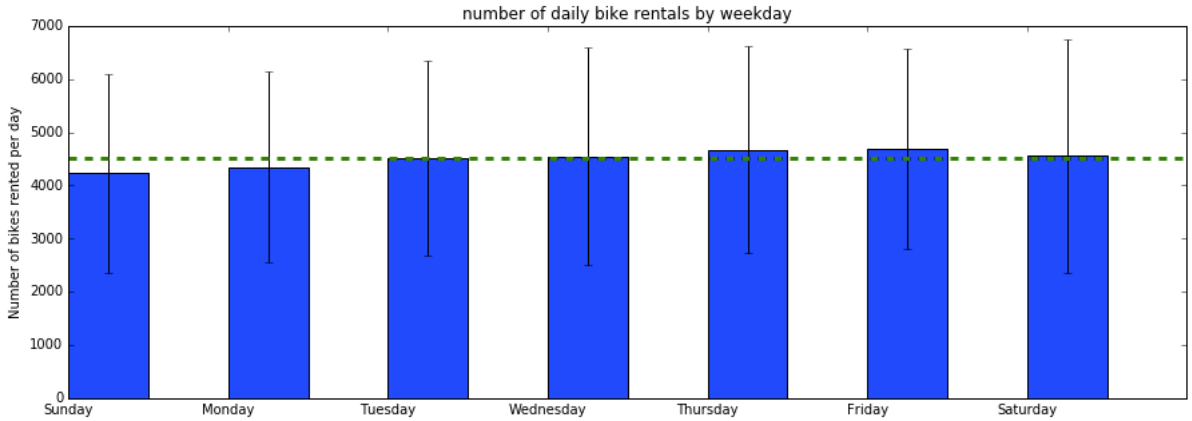


Note: dashed line signifies count averaged over all hours.

- Rentals are highest at 8am and 4pm – 7pm. This corresponds with regular work commute hours.
- Based on this data, an incentive could be established that reduces the cost of bike rental during non-peak hours:
    - 15% discount during low demand hours: 8pm – 6am and 10am.
    - 5% discount during mid demand hours: 9am – 3pm
    - 5% upcharge during high demand? 8am and 4pm – 7pm

f. **Investigation into less influential variables: holiday, weekday, working day**

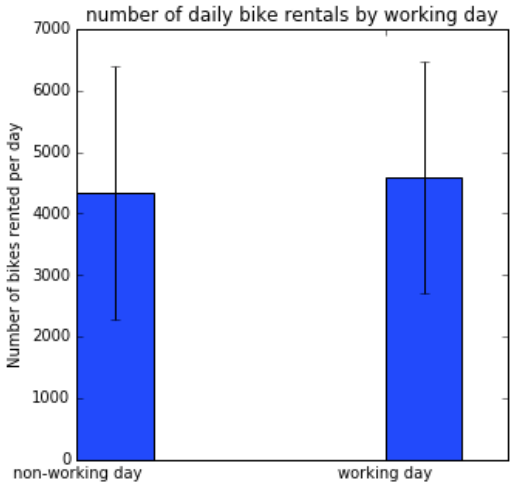
Surprisingly (to me), these variables had the smallest impact on bike share usage. Below is my analysis of bike rental rates by weekday and by working day.



Note: the error bars signify the standard deviation of the count for each weekday. Dashed line signifies count averaged over all weekdays.

It is evident from the graph that bike rentals do not vary significantly by day of the week.

- Sunday and Monday have lowest ridership
- Thursday and Friday show largest ridership
- Wednesday and Saturday have largest variability (std)
- The maximum variation between any two weekday averages (max - min / min) is 10.9%.



- Bike rentals are higher on working days. This may indicate that there are more regular, commute-oriented users than recreational users.

2. INCENTIVE PLAN OVERVIEW

Based on the data, it appears that the lean-demand periods are correlated most strongly with the following factors (not ranked in any particular order, since this has not been rigorously determined):

Cold or hot Feeling temp
High Wind speed
Cold or hot Temp <----- bundled as “weather and natural conditions”
High Humidity
Stormy Weather

&

Month (sep. – april)
Hour (8pm – 6am)

The incentive program outlined below aims to increase bike share usage during these lean demand periods.

First, I will state that the most responsive incentive plan would collect real-time rental data from bike stations and use the real-time demand to set prices (similar to Uber surge pricing). However, I would cap the surge prices for members. In the case that real-time rental data is unavailable, the following program(s) should incentivize bike usage during the conditions most correlated with lean-demand.

1) Financial incentives
    a. Apply discount during poor weather conditions (cold or hot temperature, high humidity, stormy weather, high wind speed). This requires bike stations with IoT capability to update prices. If stations had this capability, then they would likely just transmit real-time demand data to generate demand-based pricing. So, in essence, this scheme seems unlikely.
    b. Apply discount for low-demand months
        i. May-July have highest rental rates

        ii. 10% discount on Sep. – April
    c. Apply discount for low-demand hours
        i. 15% discount during low demand hours: 8pm – 6am and 10am.
        ii. 5% discount during mid demand hours: 9am – 3pm
        iii. 5% upcharge during high demand? 8am and 4pm – 7pm

2) Non-monetary incentives
    a. Run promotions during low-demand months
        i. Free friend/guest passes to get people to introduce the service to more people
        ii. Free month trail periods during low-demand months, the objective being to secure members for the rest of the year. A free trail during these months would have the least impact on other paying members, since bikes should be amply available at most bike stations and under-utilized.
        iii. Free bike day all over a city—anyone can rent a bike for free on this day. The idea is to get people to download the bike share app and create an account so that they may use it later.
        iv. Partner with a festival that occurs across a city.
    b. Incentives during poor weather conditions
        i. The bike station could have a roof that keeps the bikes dry during rainy conditions
        ii. Run competitions/challenges at businesses for employees to compete on who bikes to work the most days (even in rough weather conditions)
    c. Incentives during low-demand hours
        i. The low demand hours are at night (8pm-6am) and midday (10am-3pm).
        ii. I would incentive those individuals that use bikes for pleasure rather than for work commute. This could include:
            1. Placing bike stations in recreational areas (along a bike river path, along a nature trail area)
            2. Place bike stations in areas where individuals will use the bikes at non-commute times, such as universities, college dorms, gyms, parks, etc.

3) Actions/ improvements to the system that would increase demand during lean periods
    a. Create options that are more appealing during poor weather conditions.
        i. Provide tricycles at bike rental stations. Many individuals (whether old, disabled or unexperienced on bike) feel safer on tricycles. In particular, during uncertain weather conditions (wet or ice-y roads) people feel safer on tricycles than two wheeled bikes. The slippage is lower and chance for accident is also reduced.
            1. Reclining bikes are more protected from wind and elements and warmer.
        ii. Helmets with wind shield
        iii. Mirrors on bike improves safety
    b. Partner with employers to incentivize employees to bike to work
        i. One guess for the cause of low ridership during high humidity / hot temperature is that 1) its uncomfortable and 2) people don't want to come to work sweaty and seem unprofessional
        ii. The agency could encourage employers to have a shower in their building or bike share stations at their office to encourage employees that it will be acceptable for them to arrive at work a bit sweaty in the morning.
        iii. Employers should offer bike share memberships for their employees for free (the employer pays the cost)
    c. Partner with health insurance agencies to provide a discount for health insurance if a payer bikes regularly. It may encourage people to bike to work even when conditions are less than ideal (weather-wise).
    d. Partner with games like Pokémon Go.