# Data Collection and Processing

1. (If applicable) For those who are doing independent projects, how and from what source did you collect the data? In what ways do you anticipate that this source might be similar or different from the data you expect to use by the end of the semester? Consider things like: noise, amount of data, will the data arrive online (streaming) or offline, in a different format, etc.

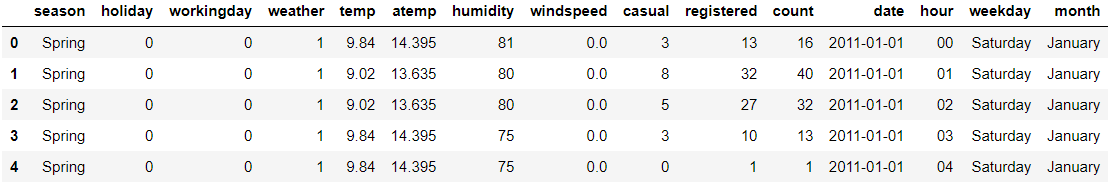
This dataset comprises of data from Bike Sharing Company related to Bike usage over the year 2011 & 2012 in Capital bikeshare system in Washington, DC with the corresponding weather and seasonal information. There is no missing data after evaluating. The complete data contains two sets. The training set is comprised of the first 19 days of each month, while the test set is the 20th to the end of the month.

UCI machine learning repository: <http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset>

Original Source: <http://capitalbikeshare.com/system-data>  
 Weather Information: [http://www.freemeteo.com](http://www.freemeteo.com/)  
 Holiday Schedule: <http://dchr.dc.gov/page/holiday-schedule>

1. What pre-processing or feature extraction, if any, did you have to perform on the data in order to make it amenable to your later visualizations? In what ways do you anticipate that this will be similar or different from what you expect to do at the end of the semester?

Classify the ‘season’ data from integer. Split the ‘datetime’ column up into ‘date’, ‘month’, ‘hour’, ‘weekday’ and then drop the original column. At the end of the semester, I might drop more other redundant data columns to build the linear regression model.



1. Given what you know now, what kind of system implementation choices might be more or less important going forward? For example, consider things like: Storage/Size requirements of the data, are there any platform restrictions (will this have to run on an embedded system or is a laptop/desktop fine), etc.

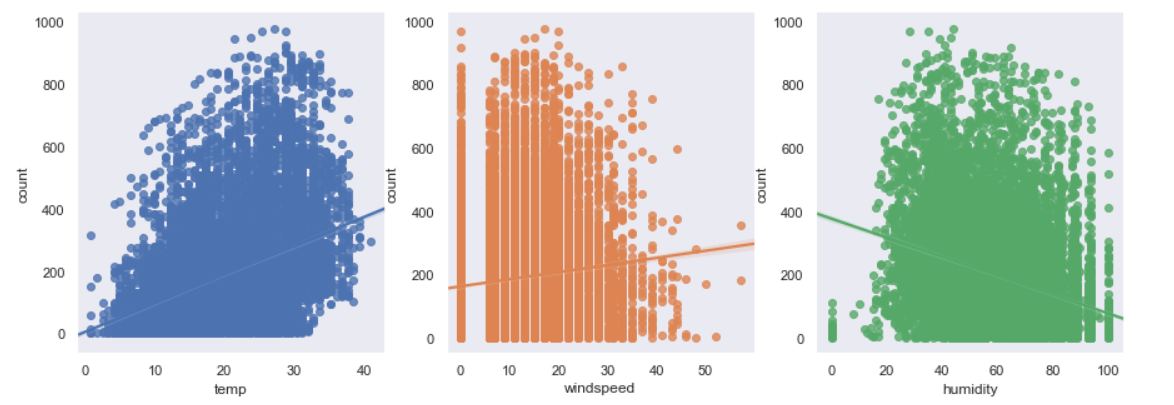
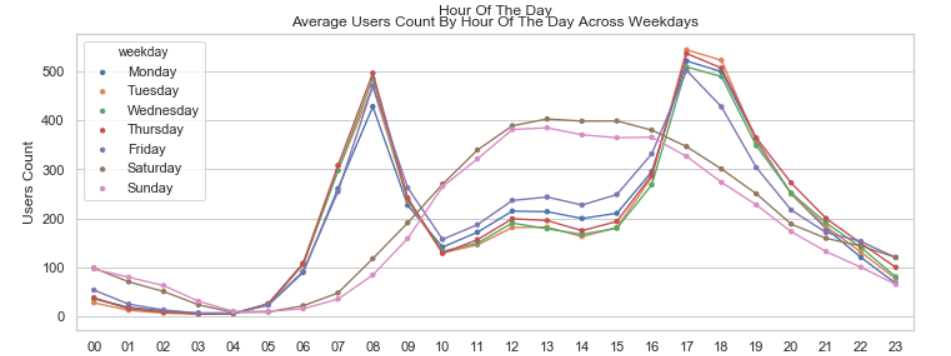
Memory usage of the data info after pre-processing is around 3MB. The data analysis is based on python 3.7.3 on the Jupyter Notebook. The operating system on the laptop/desktop is Windows 10. There is no platform restriction.

# Data Visualization

1. What type of visualizations did you use to understand your data? Why did you choose these particular visualizations over other choices?

* Line chart, bar chart and scatter plot through Matplotlib.
* Heatmap through Seaborn. A Heatmap is a graphical representation of data where the individual values contained in a matrix are represented as colors. It is perfect for exploring the correlation of features in a dataset.
* Regplot performs a simple linear regression model fit two simple variables.
* Lastly, I may use Seaborns pairplot and Pandas scatter\_matrix, which enable to plot a grid of pairwise relationships in a dataset.

1. Include copies of the visualizations (at least two). What is each visualization plotting, and what is the main thing you want someone to take away from the visualization? In what ways is the visualization limited or unable to capture the full extent of the phenomena you are interested in?

*  From the plotting, we can obtain that the customers tend to ride under a comfortable weather condition, which includes a reasonable range of temperature, wind speed and humidity. Extreme weather will inhibit bike rental demand. 
* The figure examines usage pattern on each day of the week. We clearly see that on weekdays, usage spikes around 8am and 5pm. This likely corresponds to people using the bike-sharing service as commute transportation. On weekends, the usage has a broad peak from 12pm to 4pm. People are likely using these bikes for leisure.

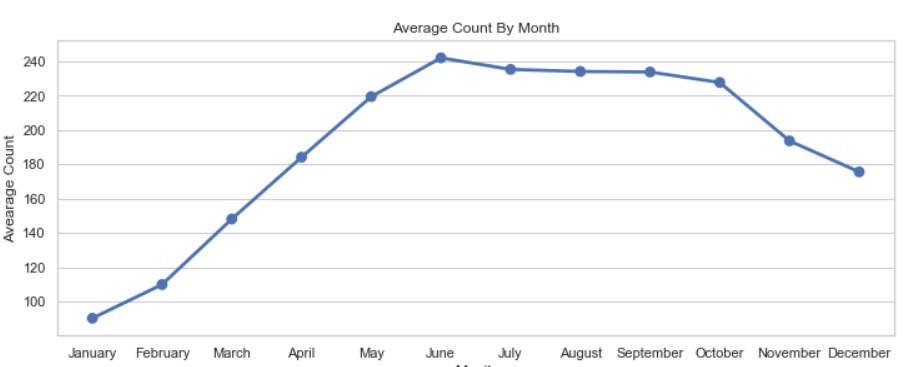
# Data Interpretation

Refer to your above visualizations (or reproduce selected portions of them below) to answer the following questions:

1. Looking at your visualizations, what can you say regarding whether or not the following factors are present/not-present in your data:

* Linearity
* Sparsity
* Smoothness
* Heteroscedasticity
* Class Imbalance (if applicable)
* Intrinsic Dimensionality
* Outliers

From the chart of the count related to weather conditions, we can see the linearity and sparsity. As humidity goes higher, people tend to not use the bike.



The chart of average rental count by month shows the smoothness of the demand change.

Other factors are not presented yet.

1. Given the above, what kind of trade-offs do you think you will have to make along each of the following dimensions:

* amount of available data vs. model complexity
* noise vs. model complexity (i.e., possible overfitting)
* need to model or predict one thing well vs. balancing performance across a range of objectives

Are there any trade-offs beyond those just mentioned that you anticipate dealing with?

Since the amount of data is limited. I will add some complexity by implementing polynomial regression in linear model, like using several degree of polynomials range(1,10) and get different MSE. I actually use GridsearchCV method to find the best parameters for polynomial.

# Problem Formulation

1. Which of the following types of machine learning problems does your project fit under and why (if multiple, explain each):

Regression (Supervised)

K Nearest Neighbors Regression and Linear (and Polynomial) Regression are used to model the relationship between time and rental counts data.

# Model Choice

1. For the type of problem you identified above, what would be an appropriate "Baseline" model that we learned about in class (or via the readings) that you could use as a performance benchmark? Put another way, what is an easy-to-implement benchmark approach (whether naive or state-of-the-art) that you can use as a "first pass" at solving the problem, and why did you pick it? Your final model at the end of the semester will try to improve upon this baseline.

First, we start by prediction the outcome variable with its mean to set the baseline. Later, Linear Regression, KNN, and decision tree will be implemented as benchmark to improve upon the baseline.

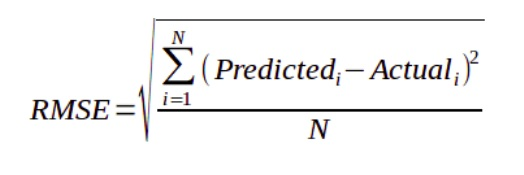
1. Look back at your data visualization: which assumptions from that (if any) are reflected in either the modeling choices or performance of the baseline model? Which assumptions are not reflected in the baseline model? What kind of things might you try for your final "improved" model at the end of the semester that might account for those assumptions?

The average rental counts by datetime fluctuation assumption are reflected in the modeling choices. The usage pattern of each day within the week will not be modeled since its high dimensionality.

# Testing Methodology

1. What constitutes “success” for your model? How will you know (i.e., quantitatively measure) if the model has done a good job? Put another way, what specific loss or score function will you use to compare your baseline model and your final model (either write it out and show how you would calculate it)?

Root Mean Squared Error (RSME) and R2 (R Squared) metric will be used to determine the accuracy of the model in predicting the target values. K-fold Cross Validation will be used to evaluate the RSME and R2 and avoid over-fitting.



1. Implement (using either an existing library or your own code) your above choice of baseline model on your currently available data, and assess its performance using the "success" criteria (i.e., cost function) you selected above. Where applicable, use the appropriate type of cross-validation to differentiate performance on training data versus held-out test data, and report how well your baseline model does on your problem, making sure to describe how you do the cross-validation. This train/test performance will become "the bar" that your improved model at the end of the semester will try to exceed. Include either tables or figures of the performance where appropriate.

