ENME743
Haixiang Fang
116293242

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
   Holiday Schedule: http://dchr.dc.gov/page/holiday-schedule

2. 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.

	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	registered	count	date	hour	weekday	month
0	Spring	0	0	1	9.84	14.395	81	0.0	3	13	16	2011-01-01	00	Saturday	January
1	Spring	0	0	1	9.02	13.635	80	0.0	8	32	40	2011-01-01	01	Saturday	January
2	Spring	0	0	1	9.02	13.635	80	0.0	5	27	32	2011-01-01	02	Saturday	January
3	Spring	0	0	1	9.84	14.395	75	0.0	3	10	13	2011-01-01	03	Saturday	January
4	Spring	0	0	1	9.84	14.395	75	0.0	0	1	1	2011-01-01	04	Saturday	January

3. 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.

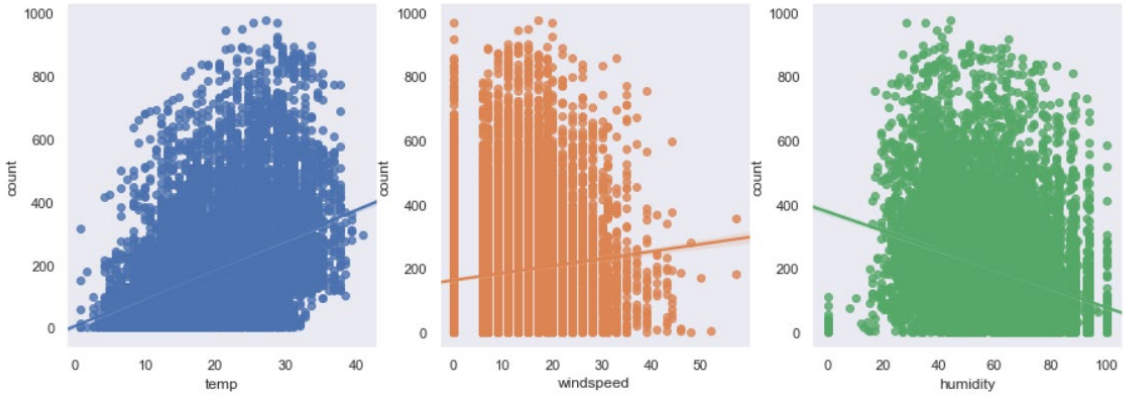
ENME743
Haixiang Fang
116293242

Data Visualization

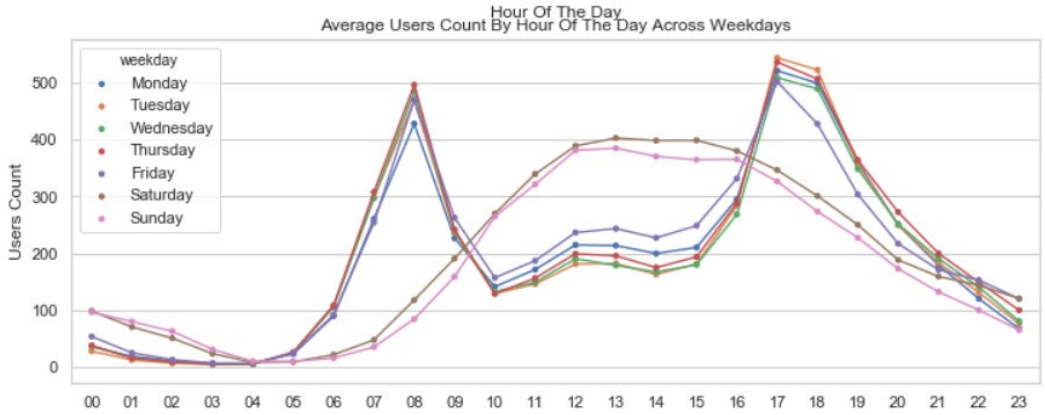
4. What type of visualizations did you use to understand your data? Why did you choose these particular visualizations over other choices?

    - 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.

5. 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.

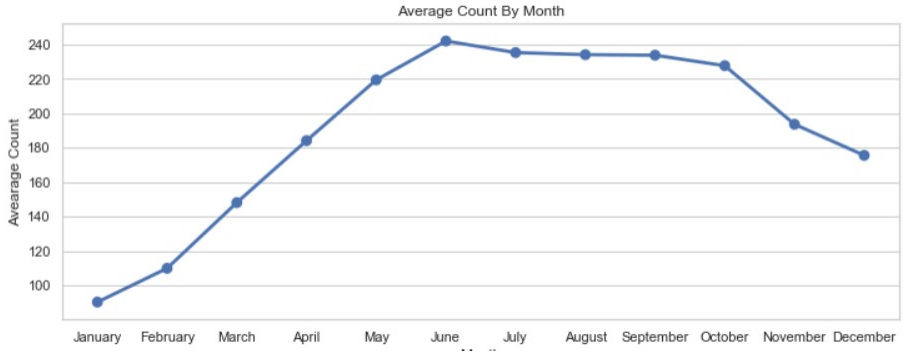
Data Interpretation

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

6. 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.

7. 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?

ENME743
Haixiang Fang
116293242
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.