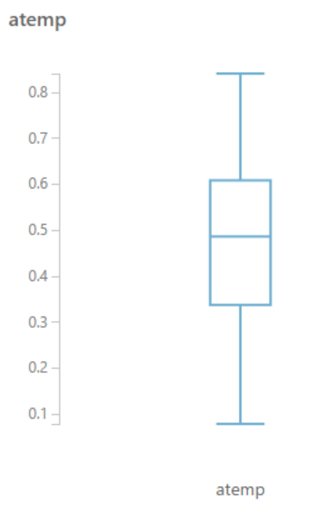
Report on SMART Data Analysis Test

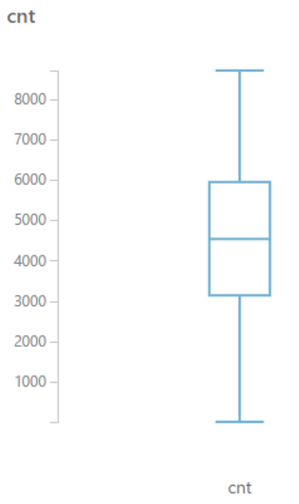
1. Exploration on Dataset

1.1 Check the level of cleanness of the data

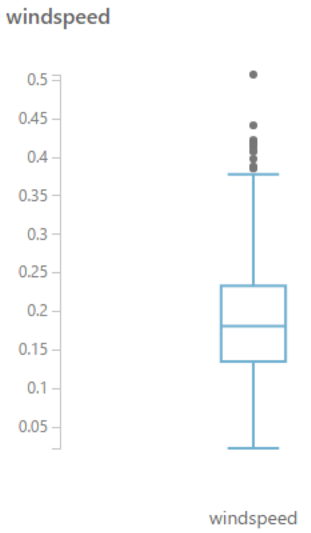
Platform: Microsoft Azure Machine Learning Studio

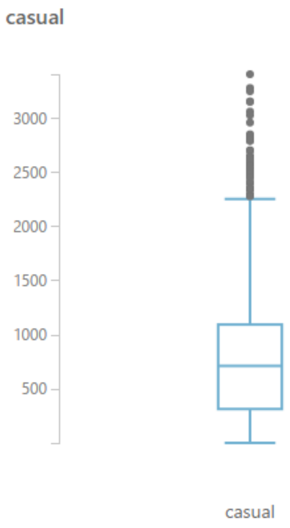
Observation on DAY file: Data generally clean, no missing values. Distributions of different columns (shown for atemp and cnt in box plot below as an example) suggest that the data sources collected are quite proportional with no certain bias towards certain type of attributes.



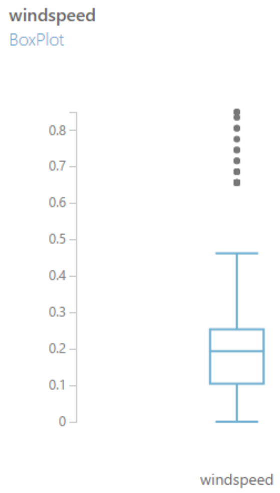


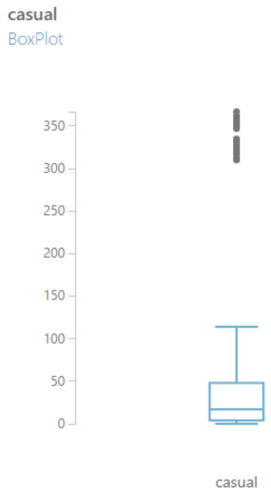
However, two columns – windspeed and casual, show a number of outliers, suggesting further processing for later stage modeling (perhaps in log scale).

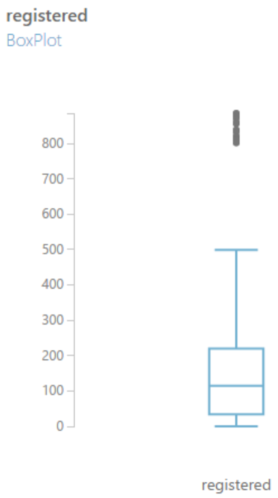


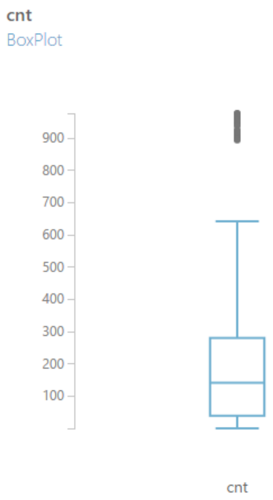


Observation on HOUR file: generally follow the trend of DAY file, however, windspeed and three types of counts (casual, registered, total) show greater variation as indicated by the number of outliers.









This wider variation is nonetheless reasonable as when this time series is brought one dimension down, the more details of the data will be revealed.

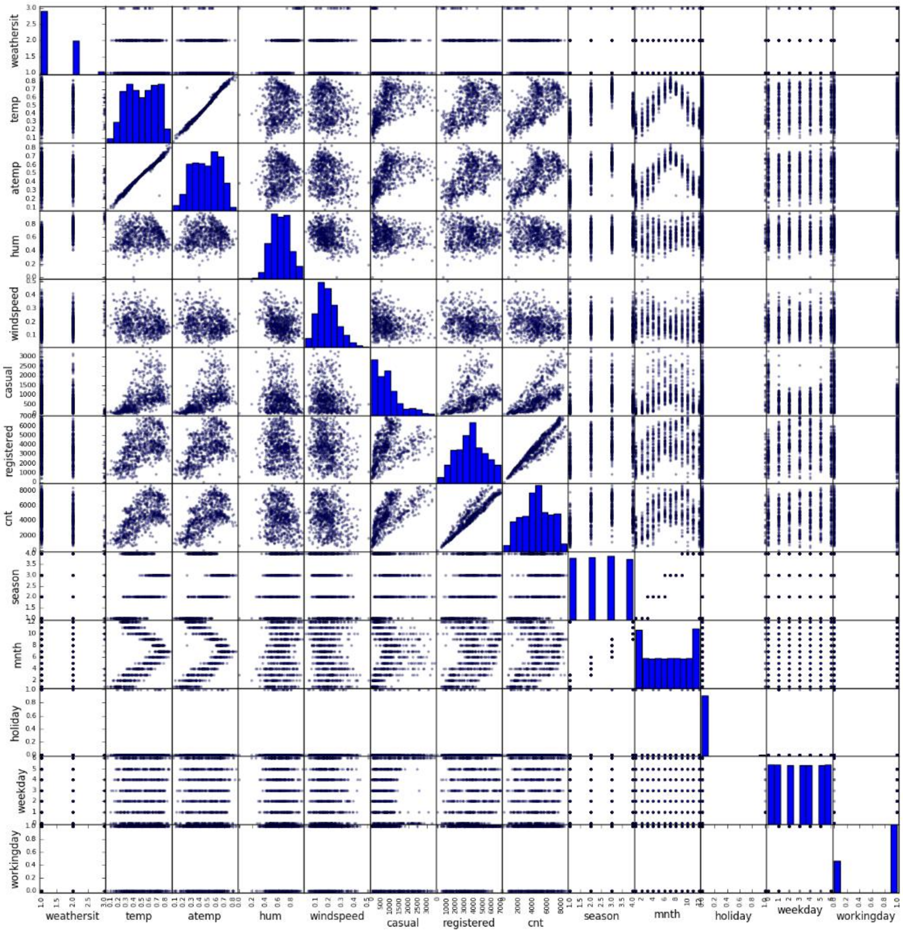
1.2 Relationship within attributes

It can be deduced that the number of customer (both casual and registered) renting the bike is a function depending on two categories of factors:

- Time: small-scaled: in hours of a day; and large-scaled: measured as months in a year
- Weather condition: weather, temperature, humidity, wind

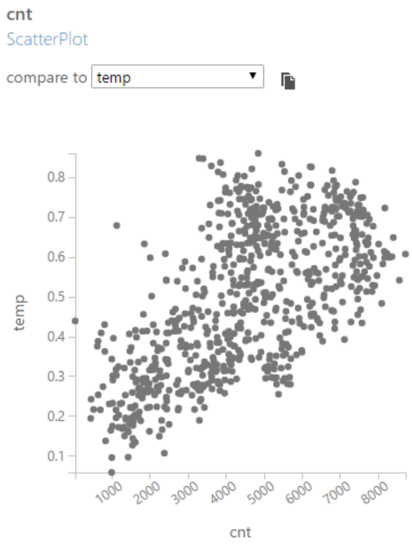
A scatter matrix to include all the feature (instant, year, and dtedate excluded).

Platform: JetBrains PyCharm



Weather:

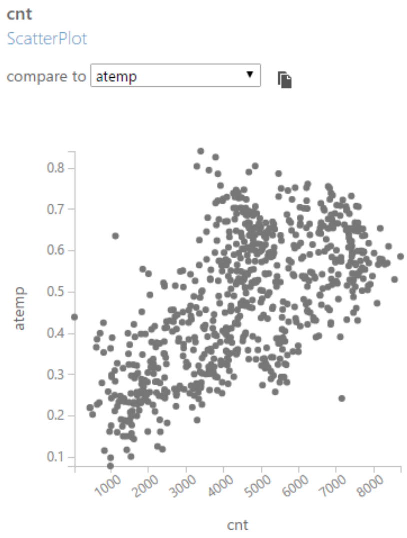
To take a closer look, scattered plots are screenshot below, and observation could be made that for weather condition category, temp/atemp show the greatest linear correlation with renting count, while humidity and wind do not display any obvious relationship.



cnt

ScatterPlot

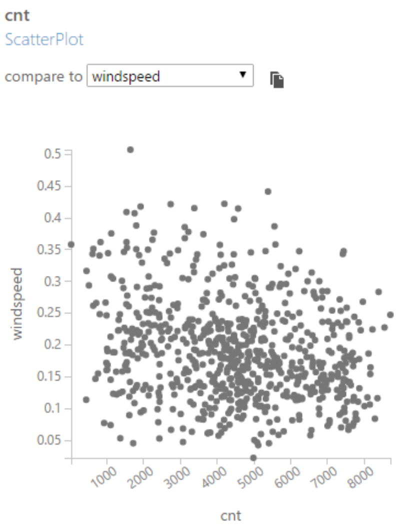
compare to temp



cnt

ScatterPlot

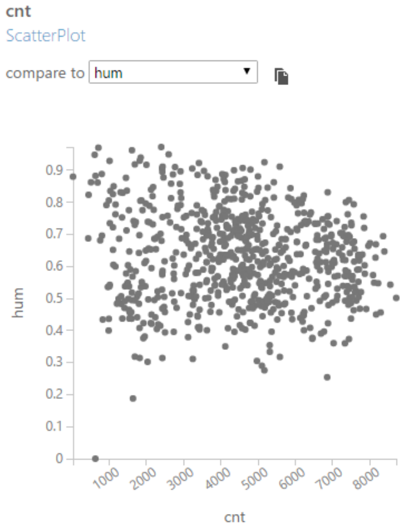
compare to atemp



cnt

ScatterPlot

compare to windspeed



cnt

ScatterPlot

compare to hum

And similar patterns show for the sub-categorical registered and casual count. Therefore, it can be deduced that temperature would yield a higher coeffient/weightage when machine learning is adopted at later stages.

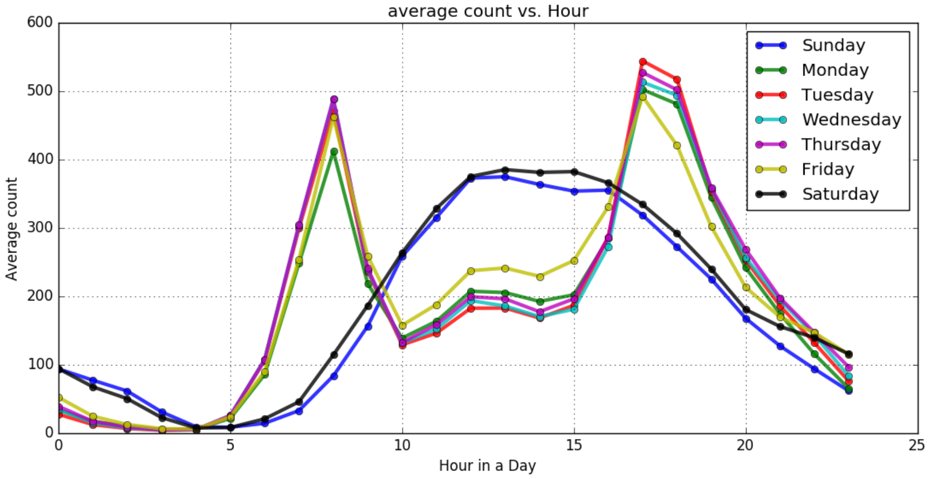
Time:

Platform: JetBrains PyCharm

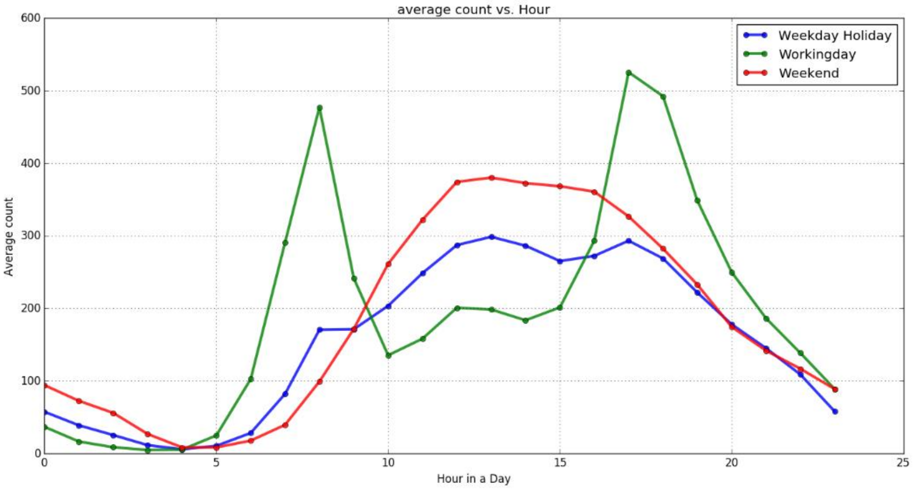
To take a closer look at how counts vary with respect to hour and day attribute:

On the **small-scale hourly-based** graph, there are some distinctive features noticeable:

- Weekend and weekday patterns have different shapes. Weekends lines have a smooth curve with climax on afternoon around 2pm, weekday lines are more steep with two climaxes corresponding to two rush hour commuting activities in a day.
- In general, weekday have greater counts of rental than weekends.
- As a minor point, weekday rental activities decrease in rush hours from Monday to Friday, but reversely increase in the afternoon.

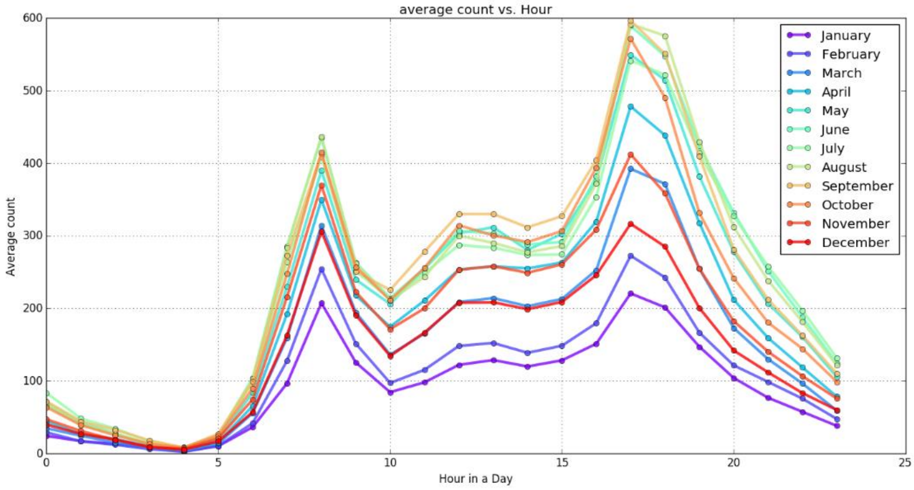


From the working day-holiday (non-working day) point of view in addition, holidays have even lower utilization rates.

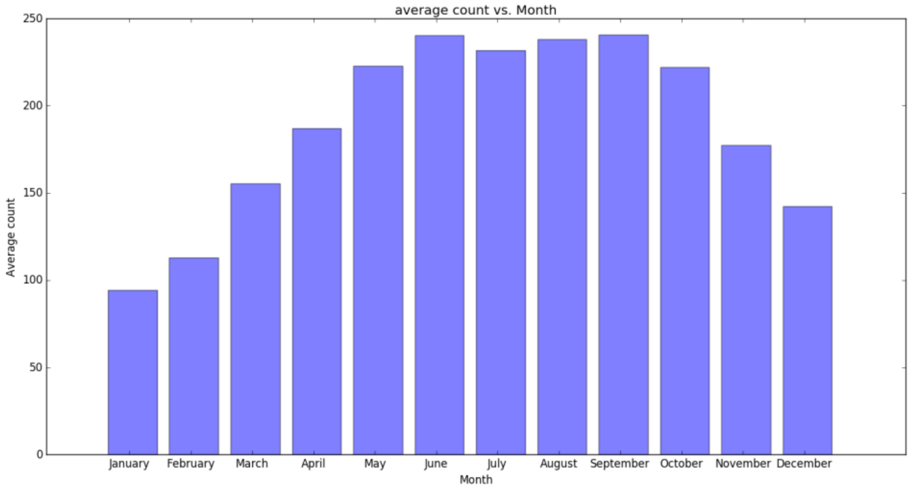


On the **large-scale monthly-based** graph, temperature trends can be obviously observed.

- The coldest months January and February witness the lowest rental activity (this also suggests that the data is collected from a region of temperate climate), and in warm months particularly from May to September, the rental activity is quite active.
- A minor trend: user show inclination to avoid extreme temperatures in summer. As the graph representing the activity in August shows that the number peaks among all candidates at morning and evening periods, but dies down in afternoon, when the weather is presumably very hot.



It can be even more obviously displayed as in the form of a bar plot.



Therefore, this presumably temperature-related trend requires us to further look into the weather condition attributes in the next section. It also means that in the machine learning algorithm, 'month' should not be included.

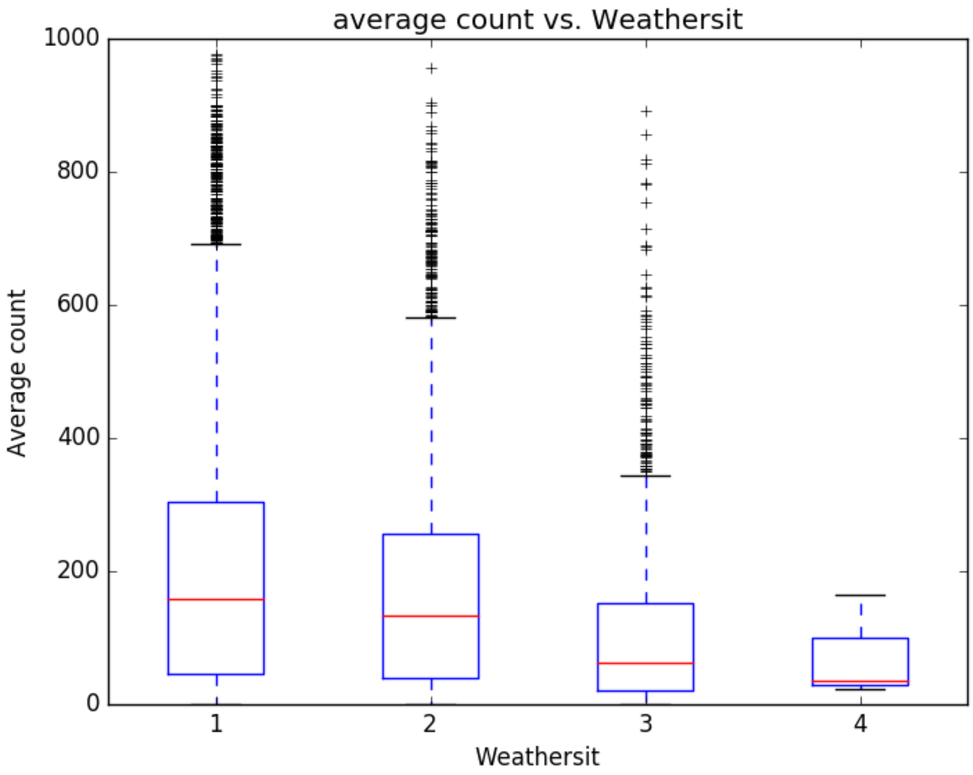
Weather:

Firstly noticeably there are 5 attributes: **weathersit** (categorical), **temp** and **atemp** (normalized continuous), **hum** (normalized continuous), **windspeed** (normalized continuous).

However, temp and atemp show great correlation to each other, based on the assumption that atemp (feeling temperature) is more related to people's subjective feelings, for the sake of simplicity, only atemp will be used for subsequent analysis.

Therefore, there are only 1 categorical value and 3 continuous values related to weather condition: weathersit, atemp, humidity and windspeed. Intuitively, rental activity decreases with the severity of weather condition deteriorates, atemp being at extreme, and windspeed and humidity too high.

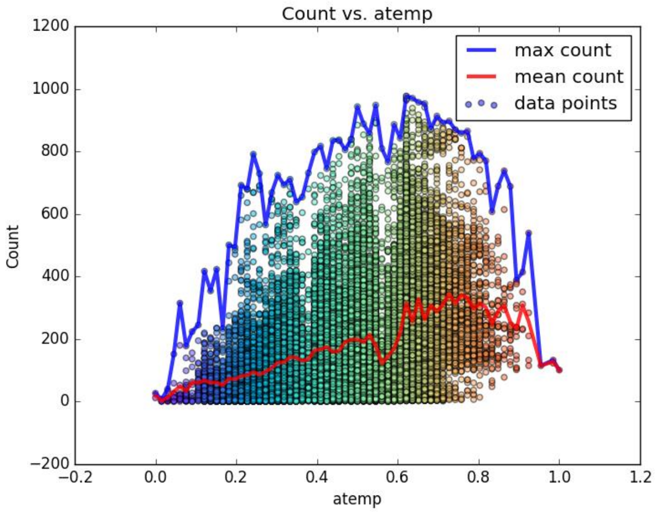
To verify such hypothesis, the following charts are plotted:

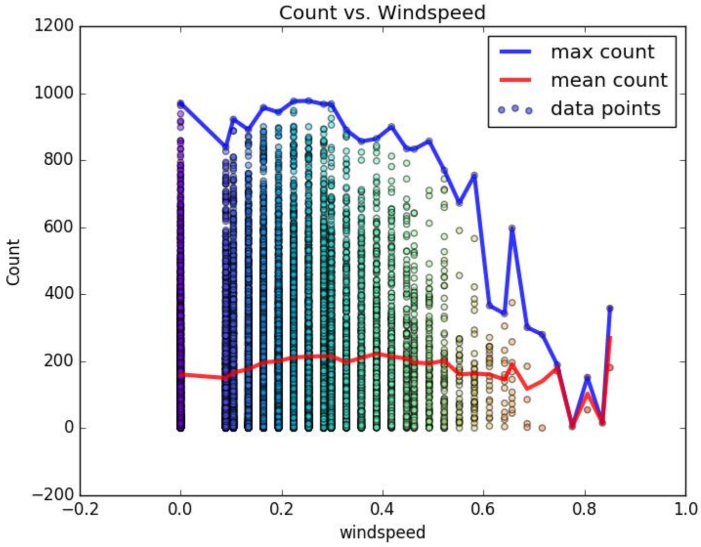


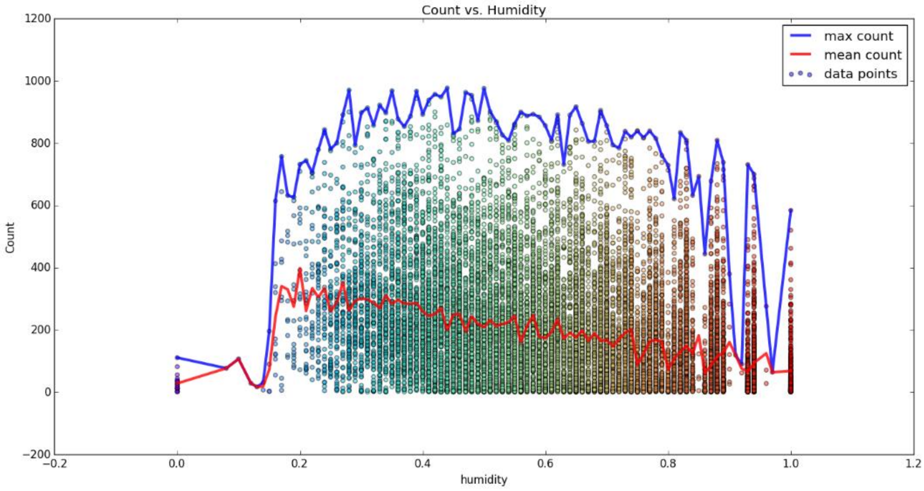
The box plot shows that when weather is relatively good (1 & 2) the rental is not affected, but at extreme weather (3 & 4) the activity decreases, albeit it is noticeable that there exist quite number of outliers suggesting that weather condition cannot capture the feature very well.

The following 3 plots shows the trending with respect to atemp, humidity and windspeed respectively. In atemp diagram, two curves from max and mean counts all show a bell-curved distribution, suggesting the behavior that people's avoiding of extreme temperature. In windspeed and humidity diagram however, the trend is largely monotonically decreasing (the final upward motion in windspeed and initial jitter in humidity is due to the lack of available data points).

That being said, atemp, humidity and windspeed are good indicators for the average intensity of rental activity, but we have to admit that due to the large fluctuations, it is not helpful to predict the exact number for a given set of parameters.


Count vs. atemp
max count
mean count
data points
Count
atemp
1200
1000
800
600
400
200
0
−200
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2


Count vs. Windspeed
max count
mean count
data points
Count
windspeed
1200
1000
800
600
400
200
0
−200
−0.2
0.0
0.2
0.4
0.6
0.8
1.0

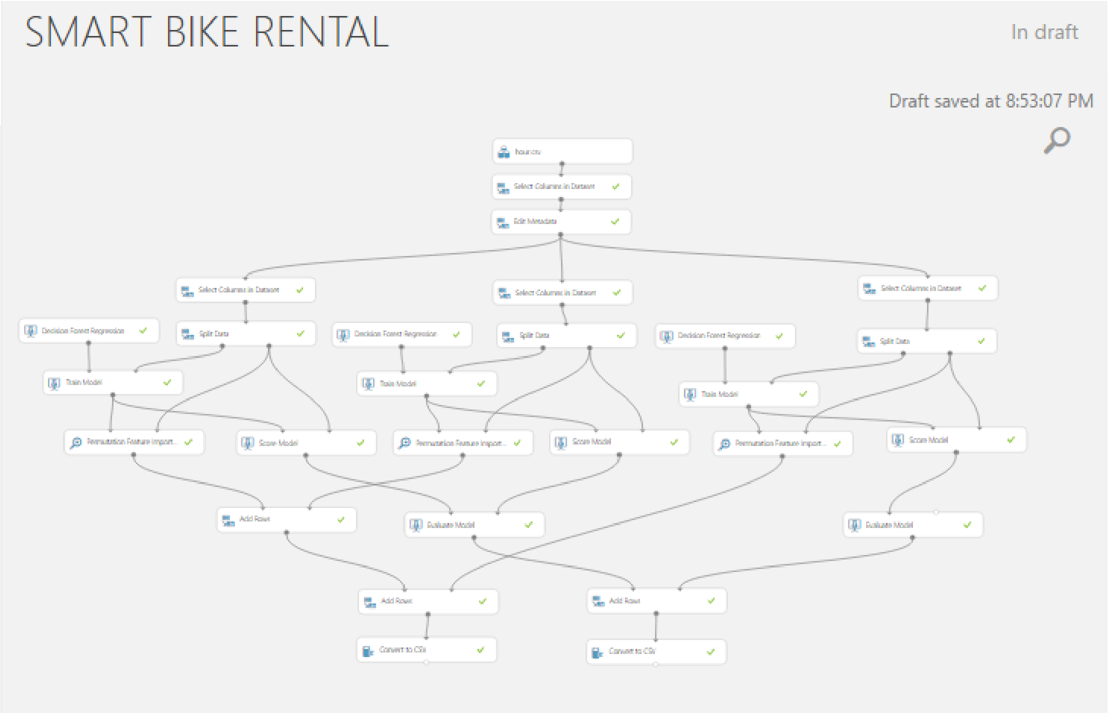

Count vs. Humidity
max count
mean count
data points
Count
humidity
1200
1000
800
600
400
200
0
−200
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2

2. Machine Learning

Platform: Microsoft Azure Machine Learning Studio

2.1 Data Flow

A screenshot:


SMART BIKE RENTAL
In draft
Draft saved at 8:53:07 PM
hour.csv
Select Columns in Dataset
Edit Metadata
Select Columns in Dataset
Select Columns in Dataset
Select Columns in Dataset
Decision Forest Regression
Split Data
Decision Forest Regression
Split Data
Decision Forest Regression
Split Data
Train Model
Train Model
Train Model
Permutation Feature Import...
Score Model
Permutation Feature Import...
Score Model
Permutation Feature Import...
Score Model
Add Rows
Evaluate Model
Evaluate Model
Add Rows
Add Rows
Convert to CSV
Convert to CSV

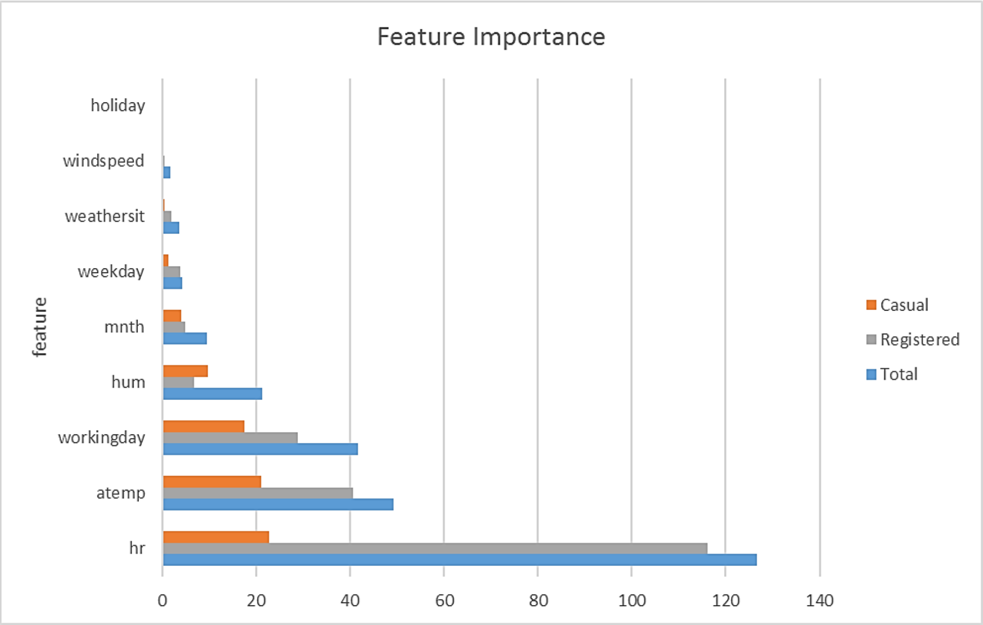
The published version could be viewed at:

https://gallery.cortanaintelligence.com/Experiment/SMART-BIKE-RENTAL-Predictive-Exp-1

2.2 Feature Importance

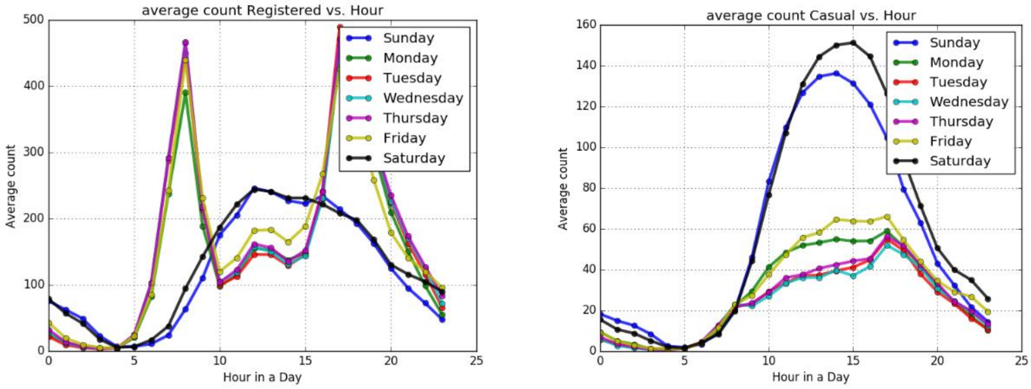
Platform: Office Excel

The 3 most important features overall are: hour, atemp, and humidity. However, there shows great variance between the influencers on casual and registered users. Noticeably, the 3 most important features in particular affect registered users much more than casual users. This phenomenon indicates that casual users are much less concerned about the weather conditions, and it also shows that registered users are out of question have a strong preference for a certain time slot such that 'hour' showing very great weightage.



To further look into the mentioned possibility, the following diagram shows the breakdown of users with respect to hour, workingday and weather conditions (weathersot. atemp and humidity).

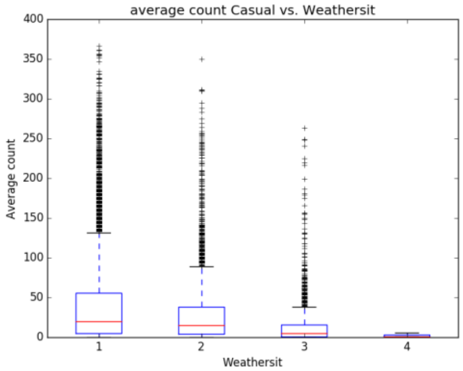
2.2.1 Difference in Casual and Registered: Hour

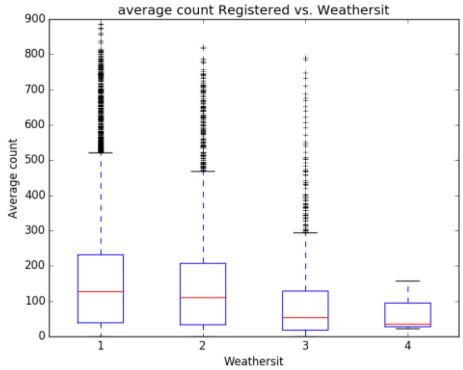


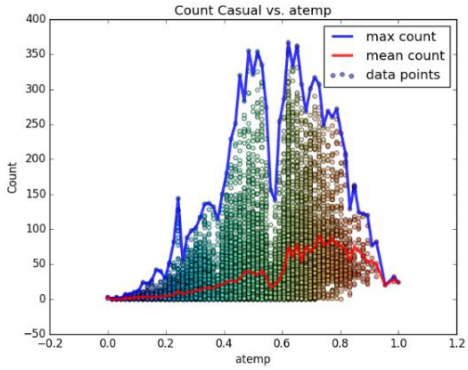
This two plots clearly shows that registered user rent heavily on weekdays, particularly rush hours, casual users frequent on Weekends, and all climaxes are located near early afternoon.

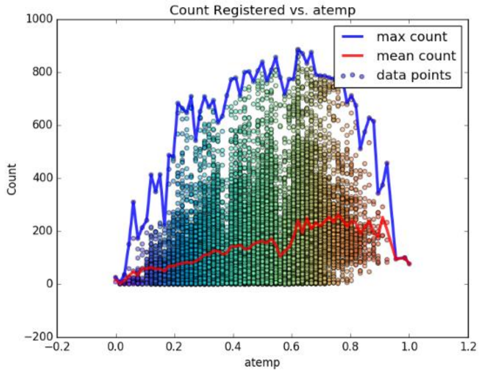
2.2.2 Difference in Casual and Registered: Weather Condition

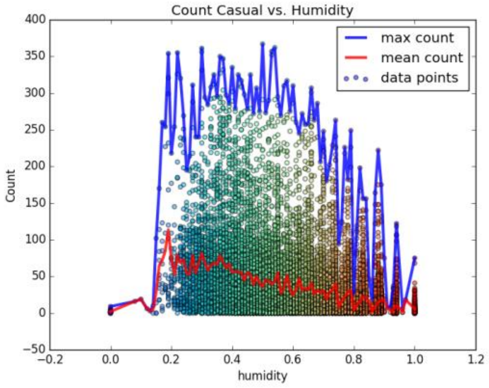
Among 3 factors, atemp shows the greatest variance in graph, this is consistent with the result generated from the feature importance diagram. Hence it could be concluded that registered users are less affected by weather conditions.

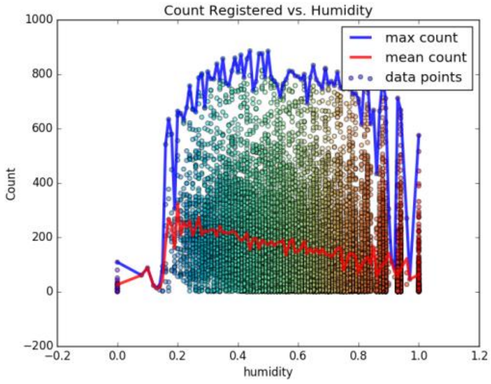












2.3 Evaluation on Machine Learning Algorithm

Using Decision Forest Regression in Azure as the machine learning model, the following table shows the results from evaluating the model by splitting 60% of the data as training set.

Count_Type	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error
Total	51.07385113	78.18628172	0.365108922
Casual	12.27701014	21.61456122	0.36276539
Registered	42.76339094	66.80195991	0.382422449

From the table, it could be observed that most errors in the total count prediction are contributed by the registered user, which means that casual user's distribution is more uniform, or in other words, change accordingly with different variables.

3. Conclusions and Suggestions on Incentive Scheme

As KDD process is definitely iterative; therefore, I would like to sum up the key findings from the research down above:

- Users:
    - There are approximately 2.5 times the number of registered users than casual users.
- Time:
    - Registered user favors strongly on weekday use for commuting while casual users are more for weekend/holiday afternoon chill-out.
    - Weekday holidays are less popular than weekends.
    - Warmer months have more rental than cold ones.
- Weather Conditions:
    - In general, Registered users are less likely to be affected by weather conditions.
    - Temperature stands for the single most important factor. Monthly trend shows that rental activity peaks at early summer and early Autumn.
    - Humidity has some level of importance, and is negatively related to rental activities.
    - Weathersit, windspeed bring some influences, but much of it is relatively unimportant unless the condition is extremely unsuitable for biking.

Based on the observations, the following suggestions may be made for the incentive scheme for better performance:

- Promotion Campaign to attract more casual users to become registered users. And of course, try to increase the pool of users in general.
- Allocate more bikes at dense residential region in the morning and more bike at CBD region when the day off for better utilization, as many populations are using rental services as a mean of commuting.
- Inventory could be reduced during winter and early spring to source for alternative income and alleviate seasonal low demand.
- Expecting more scattered rental activity in the weekends as casual users will dominate. The allocation should therefore be sparsely distributed among the region.