Analysis of Hubway bike rental to predict hourly demand

Subhabrata Mukherjee

1 Introduction:

Hubway is a bicycle sharing system in the Boston, Massachusetts metro area. The system is owned by the cities of Boston, Cambridge, Somerville and Town of Brookline, and operated by Motivate and uses technology provided by 8D Technologies, as well as PBSC Urban Solutions bikes and docking stations. The bike share program officially launched on July 28, 2011 with 61 stations and 600 bicycles. As of May 2017, the system has deployed 180 stations with a fleet of over 1,600 bikes. Bike share systems have been growing in popularity across the United States. The concept is simple. There are racks of bikes set up around the city, and people can rent a bike for a short period of time, even if only to get from point A to point B. Obviously, there are many benefits to bike shares. However, that's not to say they don't come with its disadvantages. Some of the advantages are improved air quality, Convenience, Better bike Laws, Healthier people whereas some disadvantages are traffic accidents due to first time bikers with no helmet and slow traffic. There are also several reviews found online which is about the poor customer service, over charged rental fees, lack of docking station.

One of the problems with commercial bike sharing programs is unequal riding patterns that result in unequal bicycle distribution at the end of the day. This means that unless the bikes are redistributed at night, there will be insufficient bikes at certain locations and too many bikes at other locations for the number of riders who wish to use them.

For this project, my focuses is on Hubway rideshare issues in the City of Boston and analyze the root cause of different issues faced by riders every day and provided a detailed report on the following:

1. A summarized visualization of daily trend to know how the demand for bike is on weekdays as compared to weekend or holiday
2. Identifying the most prominent time of the day having highest and lowest demand for bike
3. Identifying the month with the most change in demand and potential reasons for the change

4. Building a predictive model of Hubway bike demand in each station using machine learning
5. Finally, after addressing #4, I wanted to identify the most salient features/variables used by the model for predicting Hubway bike demand, within the limitations of my dataset

2 Potential Clients

There are two different types of clients that could be interested in the findings from this project. The first type of clients would be the US online and print media that cover socioeconomic and urban issues. These clients are magazines that take an active interest in stories driven by socially relevant issues and are backed by data analytics, for creating awareness within the public while simultaneously enhancing the quality of their readership. For example, US online media such as US News and Analytics Vidhya would fall under this category. I also anticipate interest from Transportation department and private organization in Rideshare business who wants to invest in such program in other cities.

3 Datasets used, data wrangling, and data exploration

In 2012, Hubway and MAPC challenged the public to visualize half a million Hubway rides. The Challenge is now closed, but the data is still available for public use. Recently the organizers has posted new comprehensive trip history data. Data from Hubway's launch in 2011 through the end of the 2013 regular season are now available.

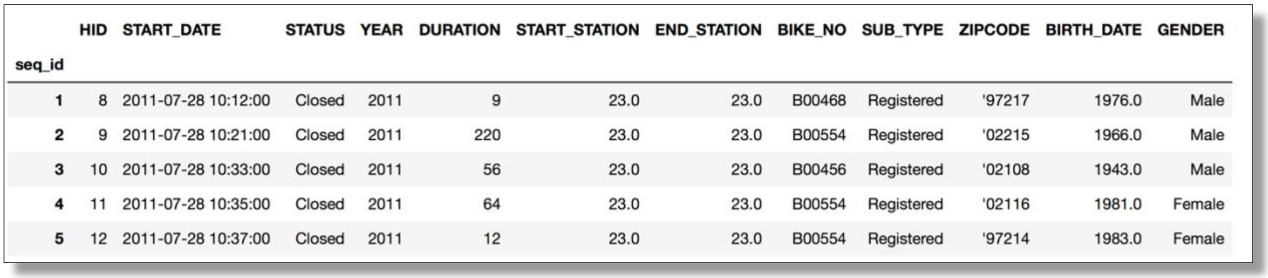
- The Hubway trip history data includes every trip taken through Nov 2013 – with date, time, origin and destination stations, plus the bike number and more. Related data (Census, neighborhoods, bike facilities, elevation, etc.) and Station status data, with information about available bikes and empty docks per station are also available.
- Since the demand for riding a bike is very much dependent on the weather, the daily temperature, humidity, wind speed details are collected by searching the historical weather details for Boston from Weather Underground website.

3.1 Reading in data

Each dataset was provided as a raw dataset in CSV format, which are imported as Pandas data frame. Each raw dataset resulted in two pandas data frames. After a preliminary look at the Hubway trip data I found that raw is having ride data from 2011 to 2013. Initially, I did not foresee any issue with the data but later I had to group the trip data by hour and merge with the weather dataset to get all related variables in one dataset.

3.2 Initial data exploration

After initial data exploration, I have found that hourly weather data is present from 10-01-2012 to 11-30-2013 whereas the trip data is present from 07-28-2011 to 11-30-2013. Hence, I had to continue my analysis on a reduce dataset after merging the trip data and weather data.



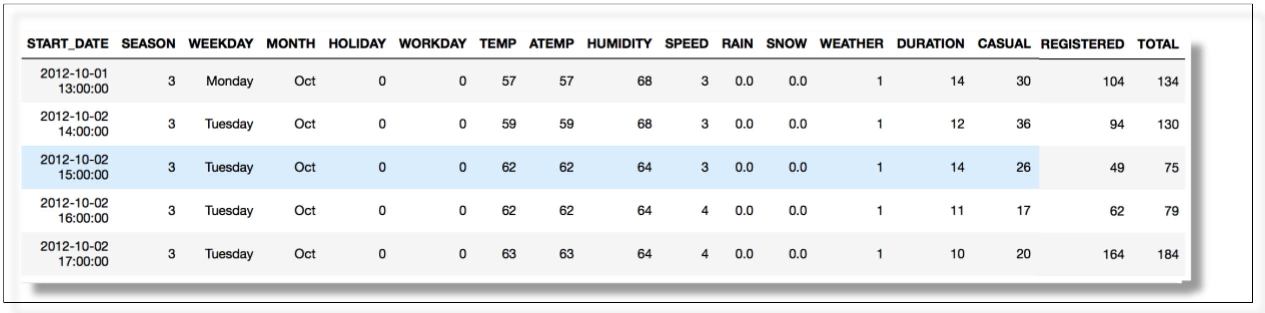
	HID	START_DATE	STATUS	YEAR	DURATION	START_STATION	END_STATION	BIKE_NO	SUB_TYPE	ZIPCODE	BIRTH_DATE	GENDER
seq_id												
1	8	2011-07-28 10:12:00	Closed	2011	9	23.0	23.0	B00468	Registered	'97217	1976.0	Male
2	9	2011-07-28 10:21:00	Closed	2011	220	23.0	23.0	B00554	Registered	'02215	1966.0	Male
3	10	2011-07-28 10:33:00	Closed	2011	56	23.0	23.0	B00456	Registered	'02108	1943.0	Male
4	11	2011-07-28 10:35:00	Closed	2011	64	23.0	23.0	B00554	Registered	'02116	1981.0	Female
5	12	2011-07-28 10:37:00	Closed	2011	12	23.0	23.0	B00554	Registered	'97214	1983.0	Female

3.3 Checking for missing values

After merging hourly trip data and weather data I verified if there is any missing value but could not find any missing data.

3.4 Normalizing Bike Rental data

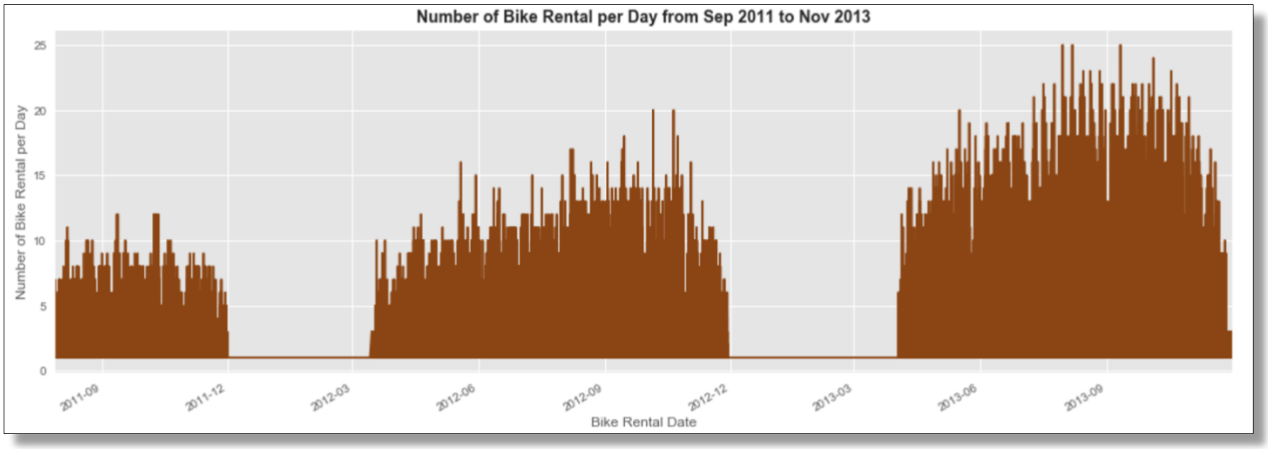
I realized while doing some initial analysis that the Hubway Trip dataframes does not have the hourly count of casual and registered user. Hence, I created another dataframe to represent the hourly trip data.



START_DATE	SEASON	WEEKDAY	MONTH	HOLIDAY	WORKDAY	TEMP	ATEMP	HUMIDITY	SPEED	RAIN	SNOW	WEATHER	DURATION	CASUAL	REGISTERED	TOTAL
2012-10-01 13:00:00	3	Monday	Oct	0	0	57	57	68	3	0.0	0.0	1	14	30	104	134
2012-10-02 14:00:00	3	Tuesday	Oct	0	0	59	59	68	3	0.0	0.0	1	12	36	94	130
2012-10-02 15:00:00	3	Tuesday	Oct	0	0	62	62	64	3	0.0	0.0	1	14	26	49	75
2012-10-02 16:00:00	3	Tuesday	Oct	0	0	62	62	64	4	0.0	0.0	1	11	17	62	79
2012-10-02 17:00:00	3	Tuesday	Oct	0	0	63	63	64	4	0.0	0.0	1	10	20	164	184

3.5 Bike Rental data exploration

After the initial exploration of the bike rental data I found that the rental data is distributed in a scatter way and there is no data for the January, February and March. After a quick enquiry about the missing data I came to know that due to weather condition the Hubway Bike rental remain closed for 3 months.

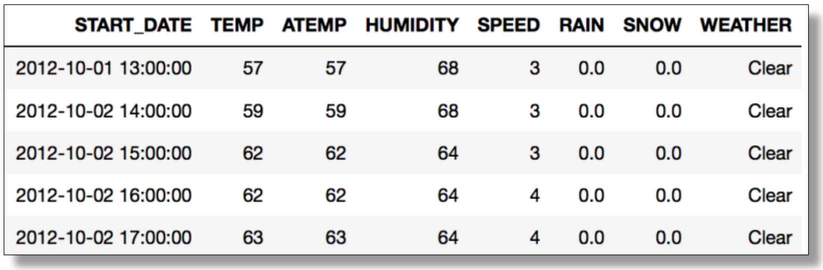


After exploring the daily rental data from 2011 to 2013 we can conclude that

1) There is an increase in bike rental over years and the program is gaining popularity
2) Weather is an important factor in determining the daily demand of bike rental

3.6 Weather data exploration

After downloading the hourly weather data from Weather Underground website, I cleaned the data, round the START_DATE column to the nearest hour and deleted additional columns.



START_DATE	TEMP	ATEMP	HUMIDITY	SPEED	RAIN	SNOW	WEATHER
2012-10-01 13:00:00	57	57	68	3	0.0	0.0	Clear
2012-10-02 14:00:00	59	59	68	3	0.0	0.0	Clear
2012-10-02 15:00:00	62	62	64	3	0.0	0.0	Clear
2012-10-02 16:00:00	62	62	64	4	0.0	0.0	Clear
2012-10-02 17:00:00	63	63	64	4	0.0	0.0	Clear

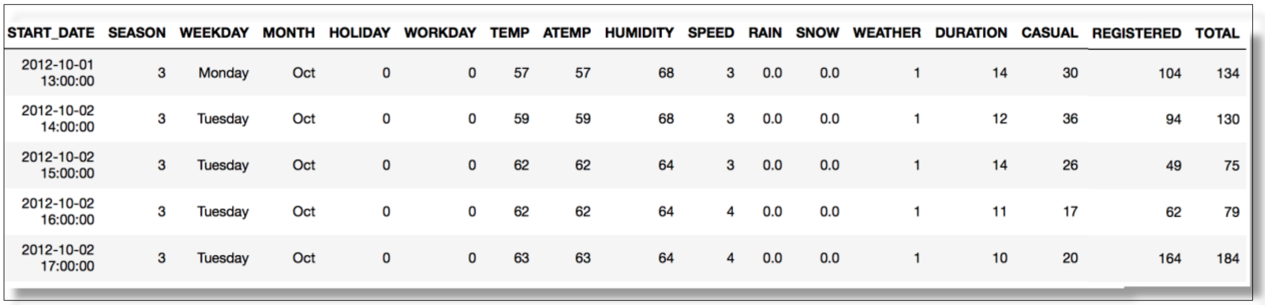
3.7 Exploration of combined dataset

I created another dataframe after normalizing the bike rental data by grouping the total number of dental per hour. The hourly rental dataframe and weather dataframe is then merged and I created a combined dataframe.

I also added few extra column which are calculated based on the combined dataframe.

Column Name	Column Description	Column Rule
HOLIDAY	To identify of the day is holiday	Holiday = 1 Otherwise 0
WORKDAY	To identify of the day is workday	Workday = 1 Otherwise 0
MONTH	The Month of the day	Jan - Dec
WEEKDAY	The day of the week	Monday - Sunday
SEASON	The season	1 = spring, 2 = summer, 3 = fall, 4 = winter
WEATHER	The Type of weather	1 = Clear, Clouds 2 = Mist, Drizzle, Fog 3 = Rain, Haze 4 = Snow, Thunderstorm, Squall

After merging the bike rental data and weather data the combined dataframe is shown below:

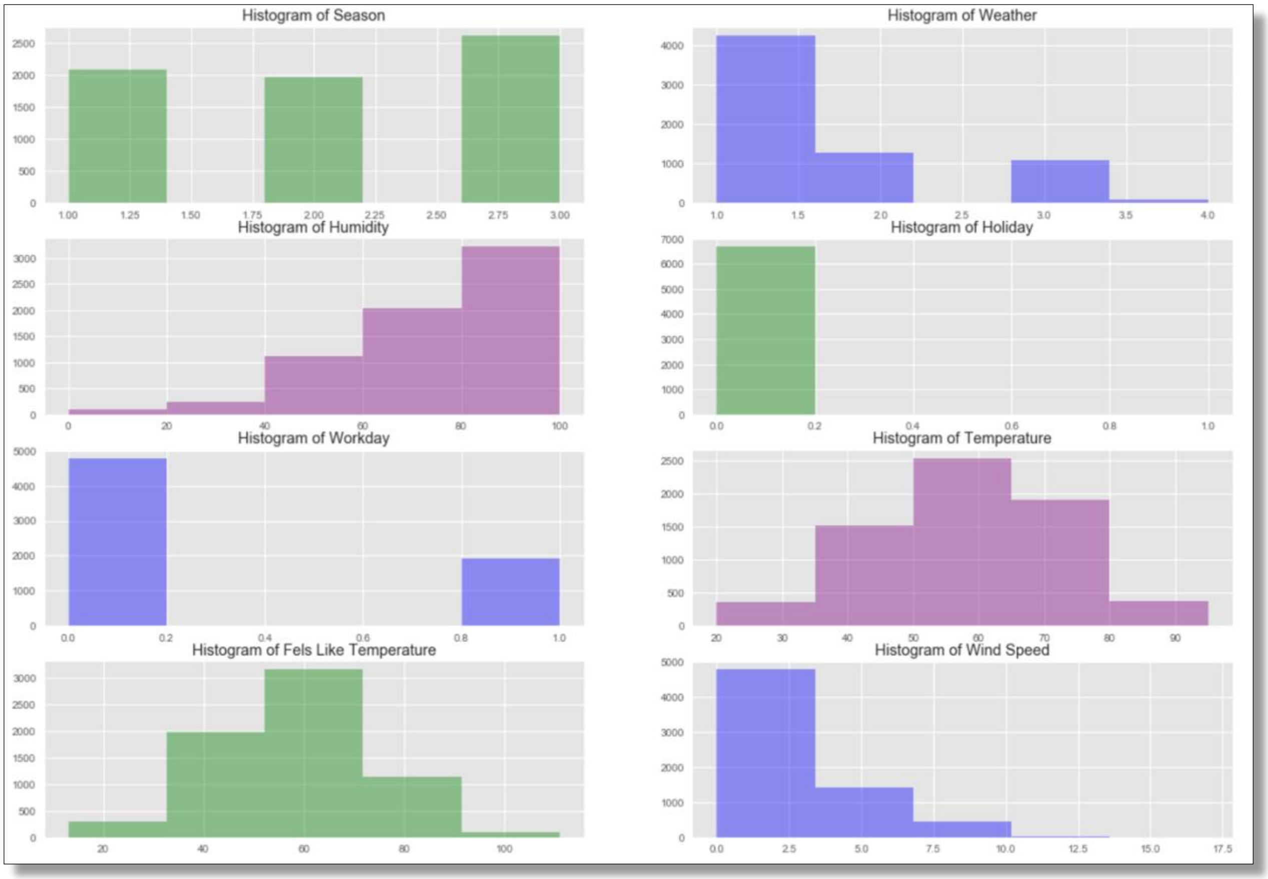


START_DATE	SEASON	WEEKDAY	MONTH	HOLIDAY	WORKDAY	TEMP	ATEMP	HUMIDITY	SPEED	RAIN	SNOW	WEATHER	DURATION	CASUAL	REGISTERED	TOTAL
2012-10-01 13:00:00	3	Monday	Oct	0	0	57	57	68	3	0.0	0.0	1	14	30	104	134
2012-10-02 14:00:00	3	Tuesday	Oct	0	0	59	59	68	3	0.0	0.0	1	12	36	94	130
2012-10-02 15:00:00	3	Tuesday	Oct	0	0	62	62	64	3	0.0	0.0	1	14	26	49	75
2012-10-02 16:00:00	3	Tuesday	Oct	0	0	62	62	64	4	0.0	0.0	1	11	17	62	79
2012-10-02 17:00:00	3	Tuesday	Oct	0	0	63	63	64	4	0.0	0.0	1	10	20	164	184

Now the dataframe has total 6696 records and 17 columns which are listed below:

3.8 Further exploration

During further exploration, I computed the distribution of numerical variables and generate a frequency table for numeric variables. The histogram plot for each numerical variable and analyze the distribution



Few inferences can be drawn by looking at these histograms:

1) Season has three categories and fall has higher contribution
2) Weather 1 has higher contribution i.e. mostly clear weather.
3) As expected, mostly working days and variable holiday is also showing a similar inference.
4) Temperature, feels-like-temperature, humidity and wind speed looks naturally distributed.

4 Prediction Using Machine Learning

4.1 Hypothesis Generation

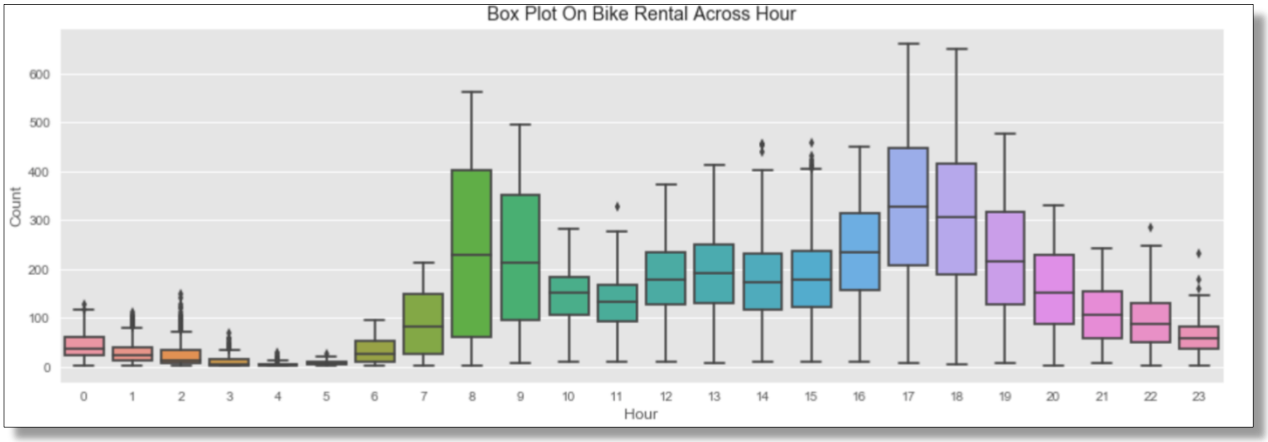
Before coming up with a Machine learning model I came up with the hypothesis which I thought could influence the demand of bikes:

1) **Hourly trend**: There must be high demand during office timings. Early morning and late evening can have different trend (cyclist) and low demand during 10:00 pm to 4:00 am.
2) **Daily Trend**: Registered users demand more bike on weekdays as compared to weekend or holiday.
3) **Rain**: The demand of bikes will be lower on a rainy day as compared to a sunny day. Similarly, higher humidity will cause to lower the demand and vice versa.
4) **Temperature**: In India, temperature has negative correlation with bike demand. But, after looking at Washington's temperature graph, I presume it may have positive correlation.
5) **Pollution**: If the pollution level in a city starts soaring, people may start using Bike (it may be influenced by government / company policies or increased awareness).
6) **Traffic**: It can be positively correlated with Bike demand. Higher traffic may force people to use bike as compared to other road transport medium like car, taxi etc.

4.2 Hypothesis Testing

4.2.1 Hourly Trend

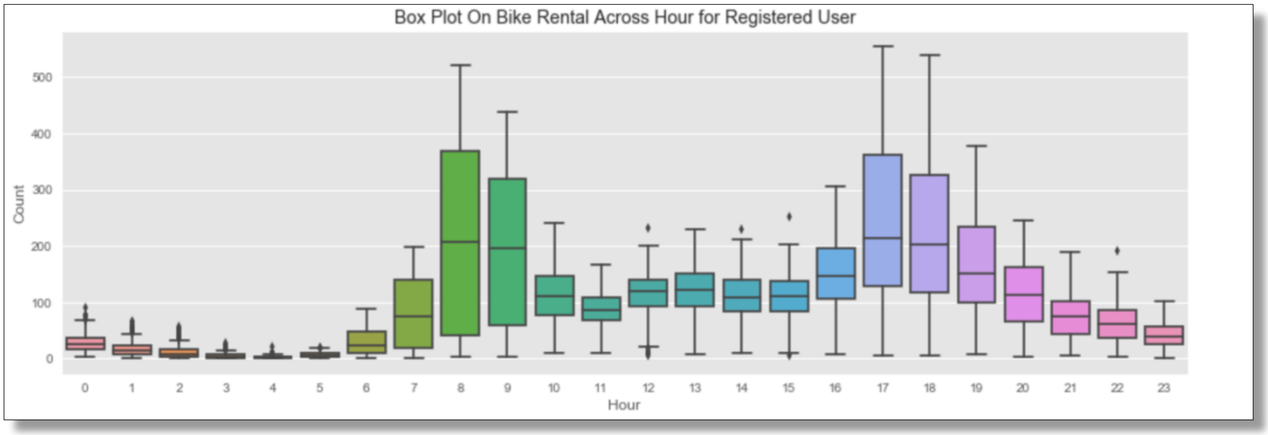
We don't have the variable 'hour' with us right now. But I can extract it using the date time column.

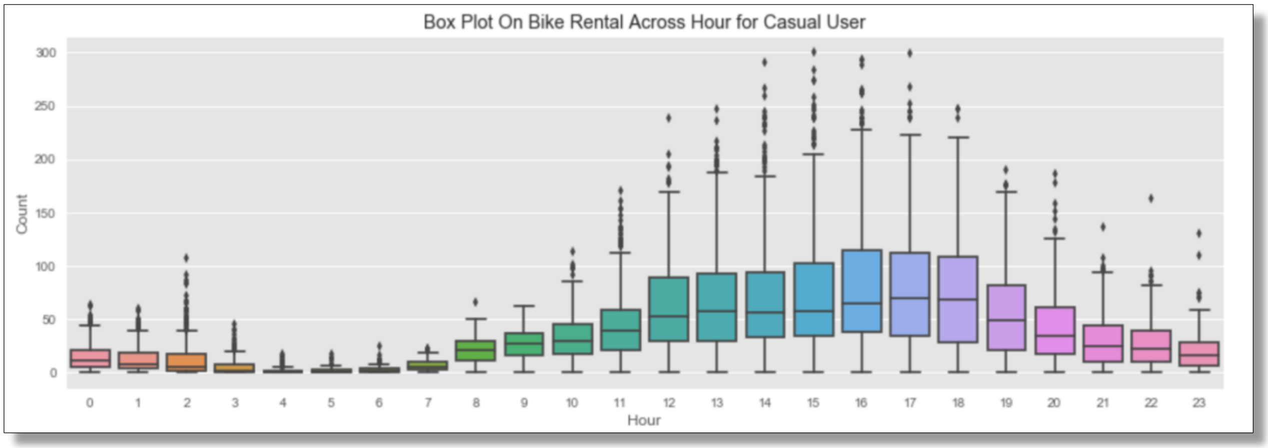


Above, we can see the trend of bike demand over hours and segregate the bike demand in three categories:

High: 7-9 and 17-20 hours	Average: 10-16 hours	Low: 0-6 and 21-24 hours

Here I have analyzed the distribution of total bike demand. Let's look at the distribution of registered and casual users separately.

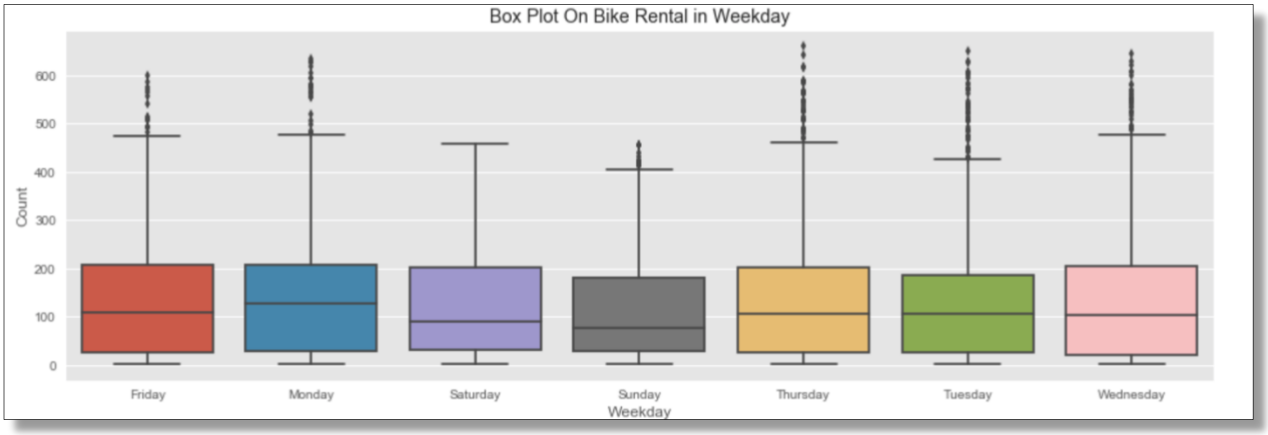




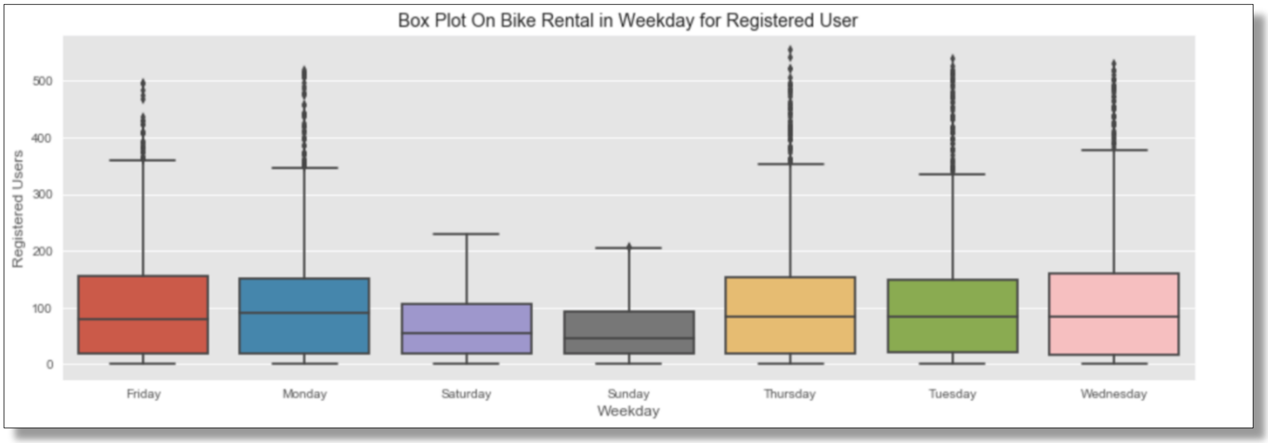
Above we can see that registered users have similar trend as count. Whereas, casual users have different trend. Thus, we can say that 'hour' is significant variable and our hypothesis is 'true'.

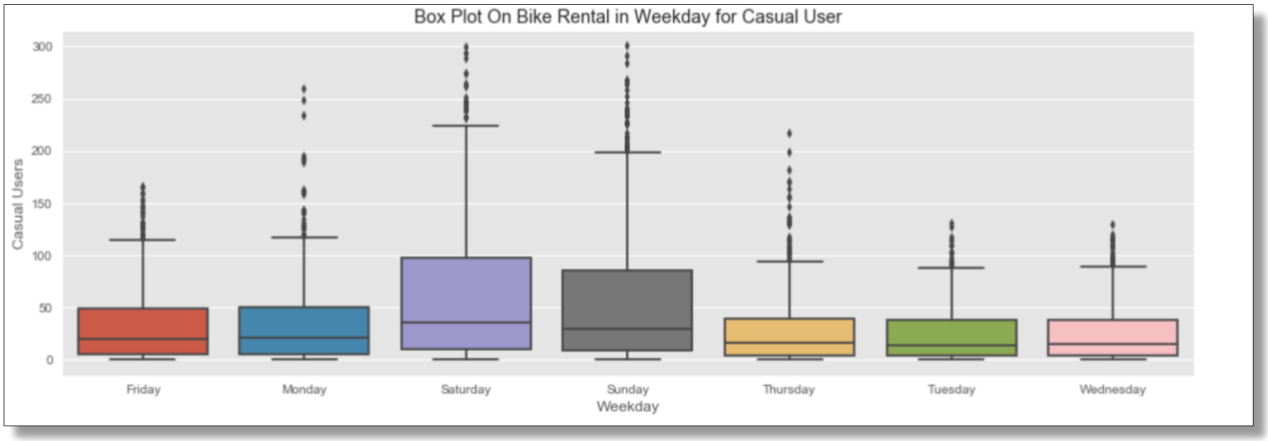
4.2.2 Daily Trend

Like Hour, I generate a variable for day from date time variable and then plot it.



Above, we can see the trend of bike demand over weekend is slightly less the demand over weekday. Let us see how the bike demand looks between registered and casual users.

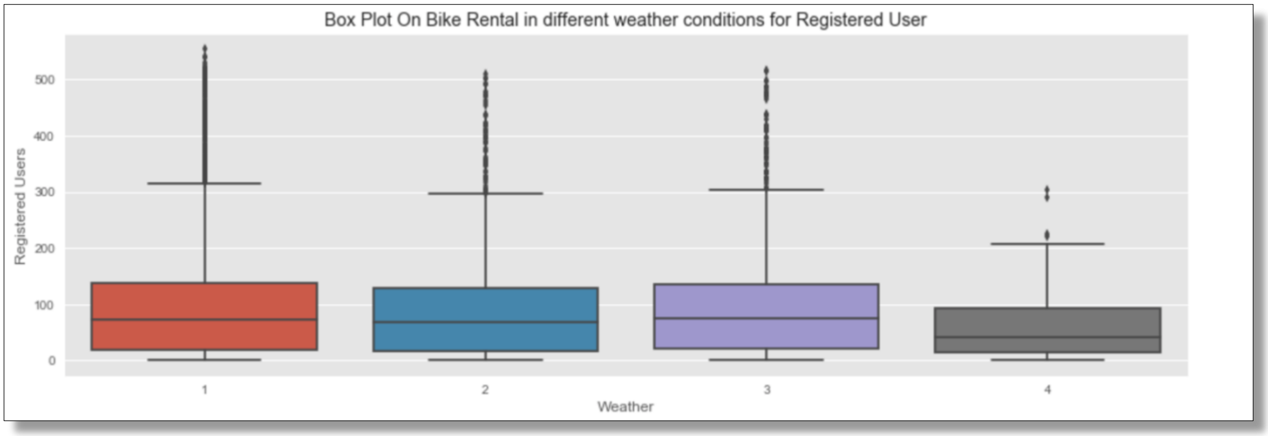


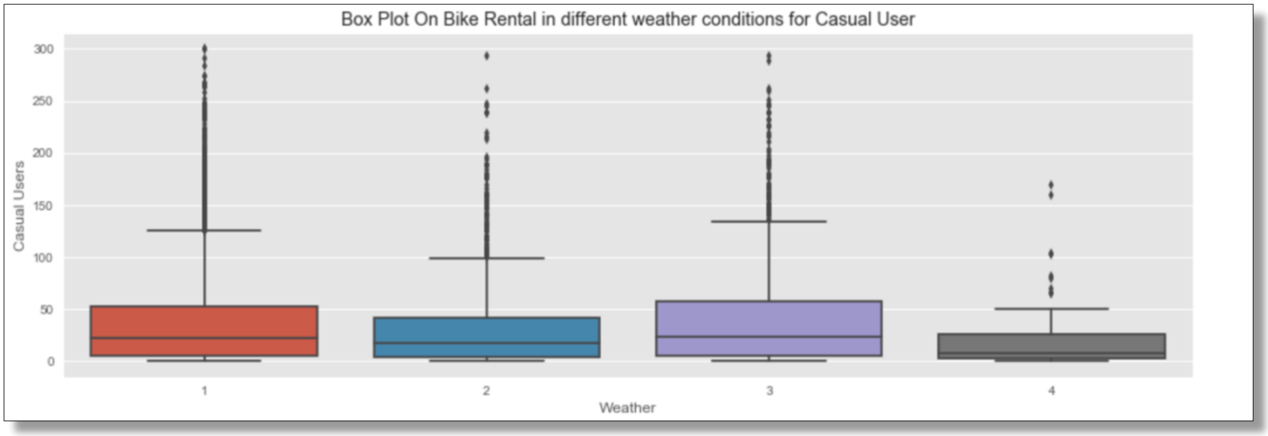


Above we can see the demand for bike rental is more during weekend for casual users but less during weekend for registered users. Thus, we can say that 'Weekday' is significant variable and our hypothesis is 'true'.

4.2.3 Rain

Since we don't have the 'rain' variable in the dataframe but have 'weather' which is sufficient to test our hypothesis. As per variable description, weather 1 represents clear weather, 2 represents drizzle, mist, fog, weather 3 represents heavy rain, haze and 4 represents snow, thunderstorm, squall. Look at the plot:

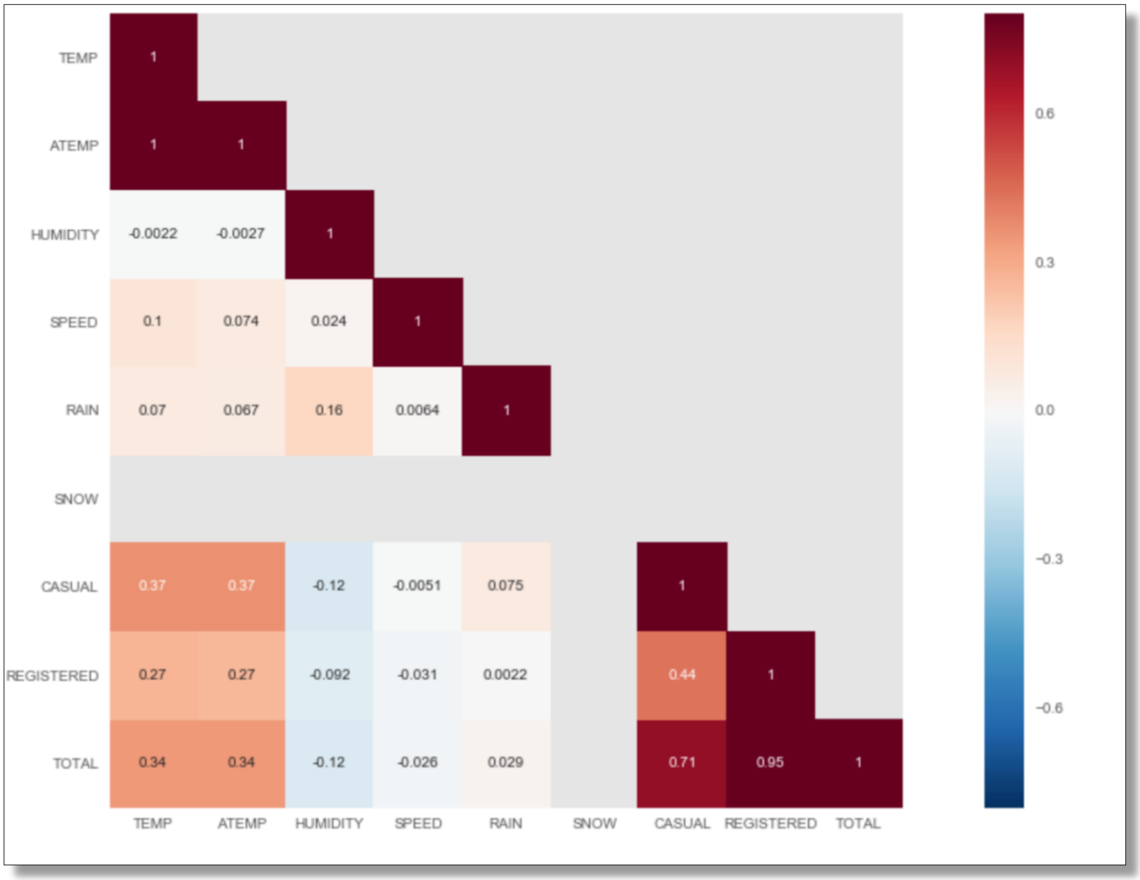




Above we can see the demand for bike rental is more when the weather is clear for both registered user and casual users. Surprisingly we can also see that the demand for bike rental is more even when there is heavy rain. Thus, we can say that 'Rain' is significant variable and our hypothesis is 'true'.

4.2.4 Temperature, Wind Speed and Humidity

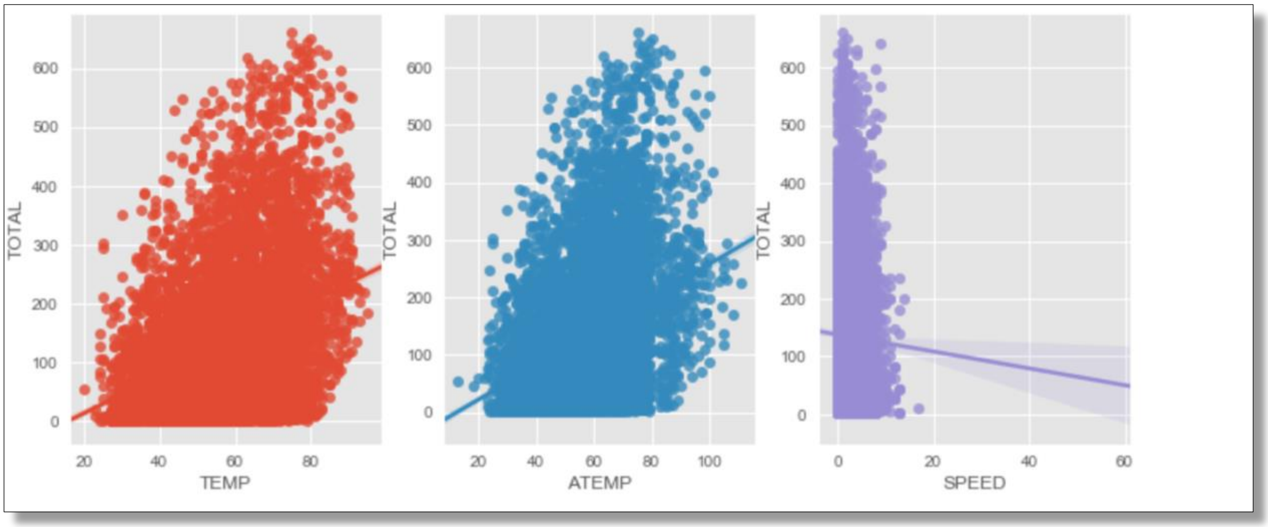
Now we want to see how temperature, Wind speed and Humidity can impact the demand for bike rental.



Here are a few inferences you can draw by looking at the above histograms:

1) Variable temp is positively correlated with dependent variables (casual is more compare to registered)
2) Wind speed has lower correlation as compared to temp and humidity

The following scatter plot also indicate the same observation.



4.3 Linear Regression Analysis

We'll next explore how this data can be modeled to predict demand for bike rental using trip data.

4.3.1 Feature selection

Thirteen columns were originally chosen as possible features of interest that might impact bike rental; their names and descriptions are listed below.

Full Name	Type	Description
SEASON	Category	Four categories-> 1 = spring, 2 = summer, 3 = fall, 4 = winter
HOUR	int64	The Hour of rental
WEEKDAY	Category	The day of the week of rental
MONTH	Category	The month of rental
HOLIDAY	Category	Whether the day is a holiday or not (1/0)
WORKDAY	Category	Whether the day is neither a weekend nor holiday (1/0)
TEMP	int64	Hourly temperature in Fahrenheit
ATEMP	int64	Feels like temperature in Fahrenheit
HUMIDITY	int64	Relative humidity
SPEED	int64	Wind speed
RAIN	float64	Amount of rain
SNOW	float64	Amount of snow
WEATHER	Category	1. Clear, Clouds 2. Mist, Drizzle, Fog 3. Rain, Haze 4. Snow, Thunderstorm, Squall

Using Pandas library, I converted the following categorical variables into dummy/indicator variables.

1) SEASON
2) HOUR
3) WEEKDAY
4) MONTH
5) HOLIDAY
6) WORKDAY
7) WEATHER

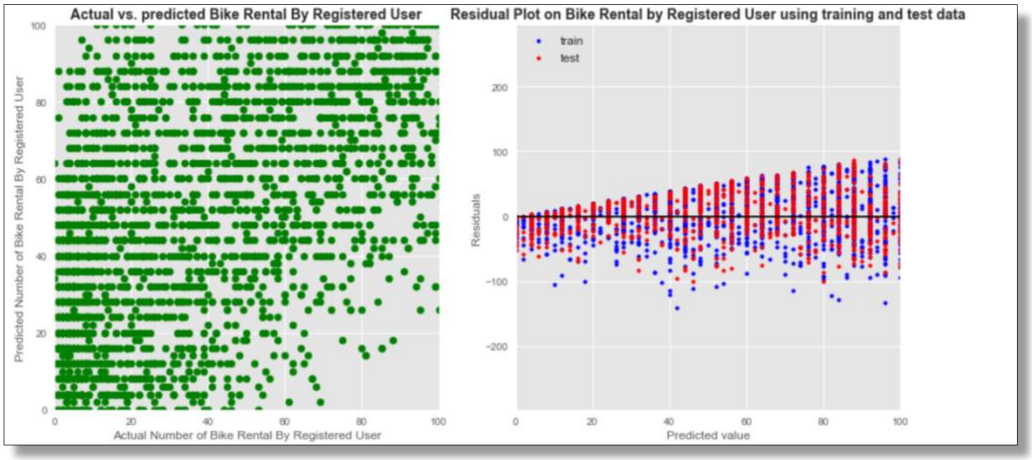
There are three dependent variables identified as follows:

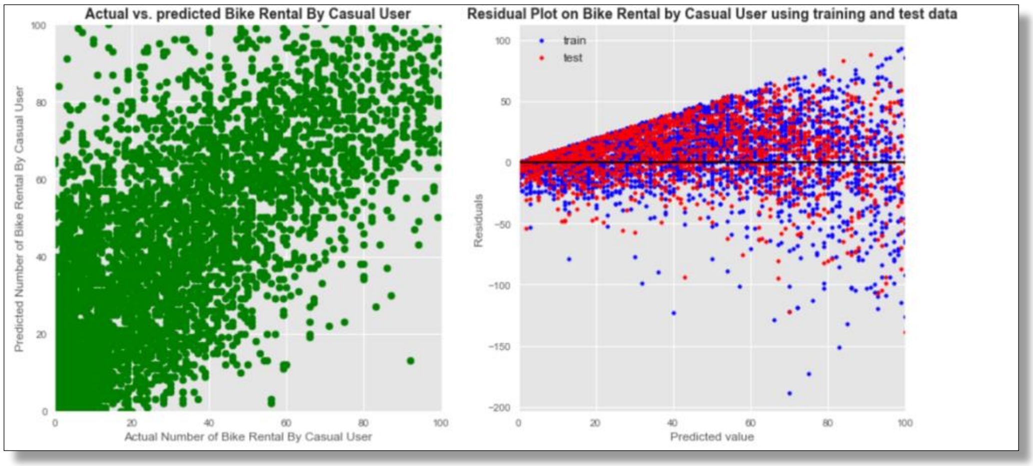
Full Name	Type	Description
Registered	int64	Number of registered user
Casual	int64	Number of non-registered user
Casual	int64	Number of total rentals (registered + casual)

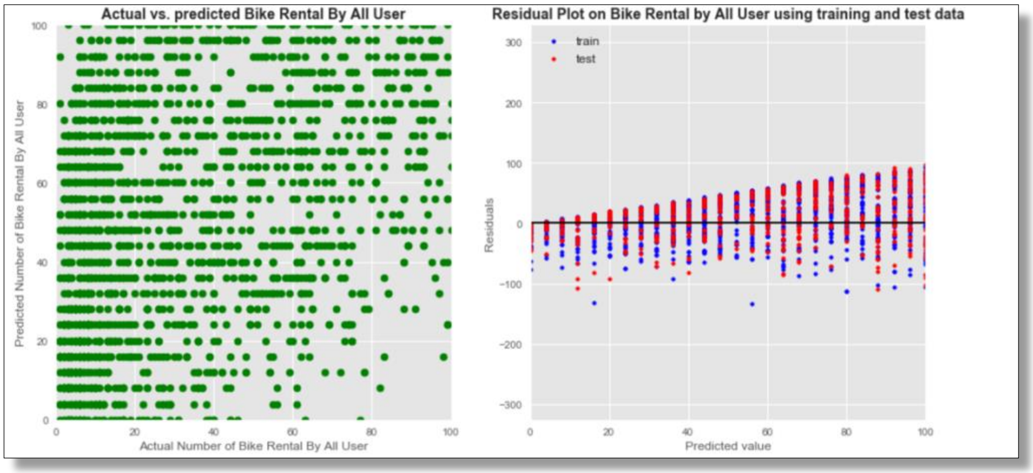
4.3.2 Linear Regression Results on Bike Rental

The LinearRegression class from sklearn was used to fit the model with ordinary least squares (OLS) linear regression. The data showing demand for bike rental data from 07-28-2011 to 11-30-

2013 and the rental data vector were split into a training (70%) and a test (30%) set, and the model was fit on the training data after normalizing the features.



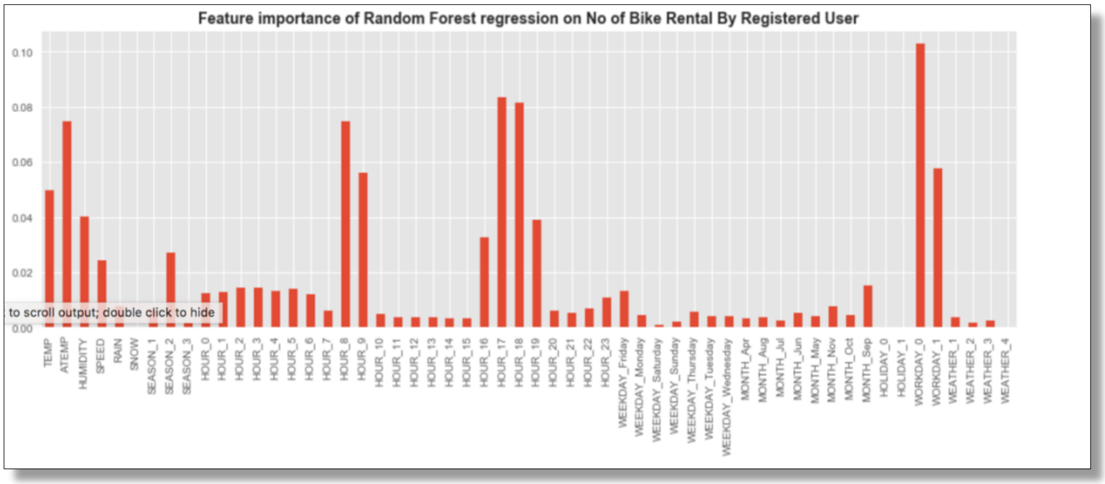




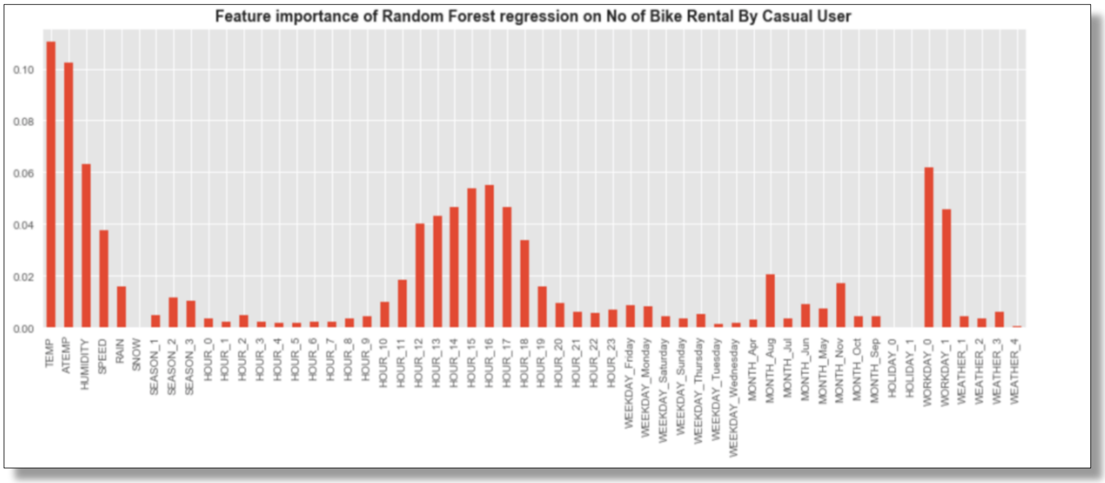
4.4 Random Forest Regression

After linear regression produced a subpar model for the demand for bike rental, we can turn to random forest regression using the same training and test sets from linear regression. The rental data vector were split into a training (70%) and a test (30%) set, and the model was fit on the training data after normalizing the features.

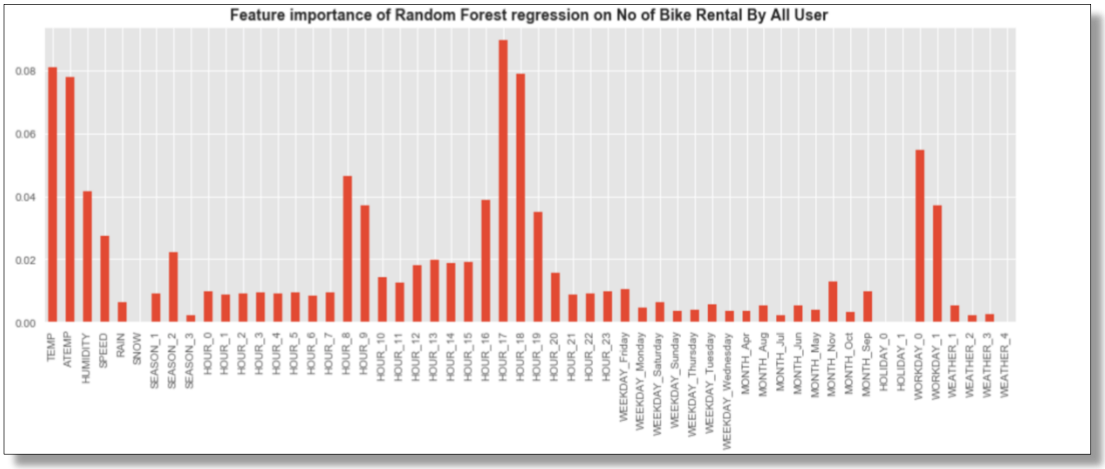
4.4.1 Random Forest Regression Results on demand for Bike rental



Feature importance of Random Forest regression on No of Bike Rental By Registered User



Feature importance of Random Forest regression on No of Bike Rental By Casual User

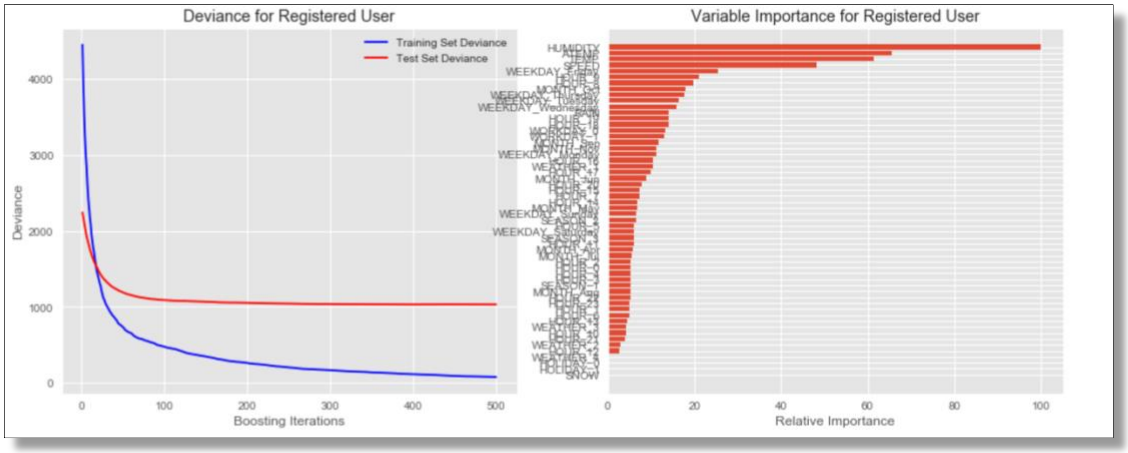


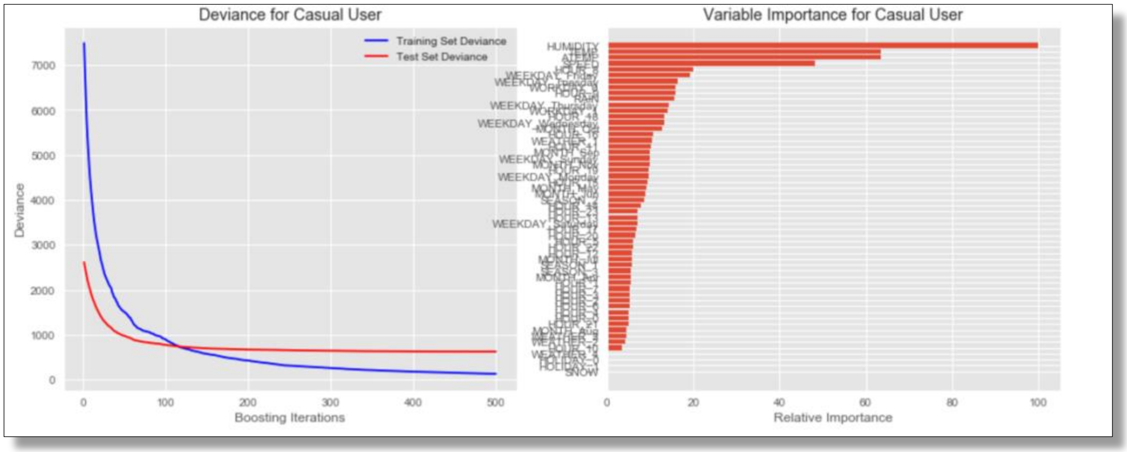
Feature importance of Random Forest regression on No of Bike Rental By All User

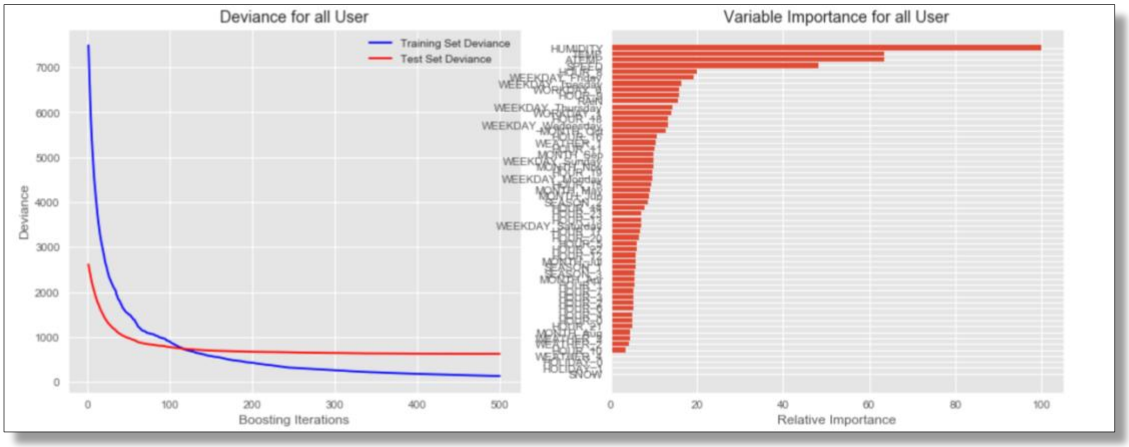
4.5 Gradient Boosting Regression

After random forest regression produced a subpar model for the demand for bike rental, we can turn to Gradient Boosting regression using the same training and test sets from linear regression. The rental data vector was split into a training (70%) and a test (30%) set, and the model was fit on the training data after normalizing the features.

4.5.1 Gradient Boosting Regression Results on demand for Bike rental







4.6 Conclusions

After evaluating three different regression model let's compare the statistics as per the following table:

Model	Registered User	Casual User	All User
Linear Regression	R2 value: 0.6 MSE Training: 3720.20 MSE Test: 3776.83 MSE K-Fold: 3813.78 SE K-Fold: 206.69	R2 value: 0.58 MSE Training: 897.87 MSE Test: 830.70 MSE K-Fold: 918.24 SE K-Fold: 100.55	R2 value: 0.63 MSE Training: 5756.49 MSE Test: 5714.19 MSE K-Fold: 5877.10 SE K-Fold: 341.43
Random Forest	R2 value: 0.81 MSE Training: 426.87 MSE Test: 2138.91 MSE K-Fold: 2313.64 SE K-Fold: 298.94	R2 value: 0.76 MSE Training: 105.17 MSE Test: 536.22 MSE K-Fold: 620.00 SE K-Fold: 56.71	R2 value: 0.81 MSE Training: 659.69 MSE Test: 3338.04 MSE K-Fold: 3711.93 SE K-Fold: 330.84
Gradient Boosting	R2 value: 0.85 MSE Training: 242.86 MSE Test: 1642.80 MSE K-Fold: 1890.12 SE K-Fold: 164.75	R2 value: 0.84 MSE Training: 59.26 MSE Test: 361.22 MSE K-Fold: 499.91 SE K-Fold: 51.03	R2 value: 0.86 MSE Training: 342.64 MSE Test: 2411.35 MSE K-Fold: 2900.61 SE K-Fold: 207.09

4.6.1 Preferred Model for Registered User

The adjusted R-squared value for the demand of bike rental for registered user using Linear model is only 0.60, a moderate number that may indicate that a linear relationship may not the right fit for this dataset.

We next look at the mean squared error (MSE), which is almost similar on the training data (MSE = 3720.20) and for the test data (MSE = 3776.83). To limit the effects of overfitting, we can use k-fold cross-validation (k = 5), but this procedure calculates a high MSE of 3813.78, indicating that the demand of bike rental for registered user is not an excellent candidate for linear regression.

The R-squared value from the fitted random forest regression on the response data is much higher than the linear model is 0.81. However, while the MSE on the training set is 426.87, the MSE for the test set jumped significantly to 2138.91. The regression for the response data is over fit. Performing a k-fold cross validation with 5 folds still showed a higher MSE of 2313.64. We can conclude that neither linear or random forest regression is a good fit to model this data.

However, from the feature importance we can have a sense of which variables the model found to have the largest impact on mean decrease node impurity during the training. We can see that WORKDAY_0 [indicate whether the day is neither a weekend nor holiday] and HOUR_17 [indicate that time is 5 PM] were more important than the other variables in decreasing node impurity.

The R-squared value from the fitted gradient boosting regression on the response data is slightly higher than the random forest model, is 0.85. However, while the MSE on the training set is 242.86, the MSE for the test set jumped significantly to 1642.80. The regression for the response data is over fit. Performing a k-fold cross validation with 5 folds still showed a higher MSE of 1890.12.

However, from the variable importance are shown below to give a sense of which variables the model found to have the largest impact. We can see that Humidity and Temperature were more important than the other variables.

We can conclude that gradient boosting regression is a good fit to model the demand for bike rental for registered user.

4.6.2 Preferred Model for Casual User

The adjusted R-squared value for the demand of bike rental for casual user using Linear model is only 0.58, a moderate number that may indicate that a linear relationship may not the right fit for this dataset.

We next look at the mean squared error (MSE), which is almost similar on the training data (MSE = 897.87) and for the test data (MSE = 830.70). To limit the effects of overfitting, we can use k-fold cross-validation (k = 5), but this procedure calculates a high MSE of 918.24, indicating that the demand of bike rental for registered user is not an excellent candidate for linear regression.

The R-squared value from the fitted random forest regression on the response data is much higher than the linear model is 0.76. However, while the MSE on the training set is 105.17, the MSE for the test set jumped significantly to 536.22. The regression for the response data is over fit.

Performing a k-fold cross validation with 5 folds still showed a higher MSE of 620.00. We can conclude that neither linear or random forest regression is a good fit to model this data.

However, from the feature importance we can have a sense of which variables the model found to have the largest impact on mean decrease node impurity during the training. We can see that Temperature and Humidity were more important than the other variables in decreasing node impurity.

The R-squared value from the fitted gradient boosting regression on the response data is slightly higher than the random forest model, is 0.84. However, while the MSE on the training set is 59.26, the MSE for the test set jumped significantly to 361.22. The regression for the response data is over fit. Performing a k-fold cross validation with 5 folds still showed a higher MSE of 499.91.

However, from the variable importance are shown below to give a sense of which variables the model found to have the largest impact. We can see that Humidity and Temperature were more important than the other variables.

We can conclude that gradient boosting regression is a good fit to model the demand for bike rental for casual user.

4.6.3 Preferred Model for All User

The adjusted R-squared value for the demand of bike rental for all user using Linear model is only 0.63, a moderate number that may indicate that a linear relationship may not the right fit for this dataset.

We next look at the mean squared error (MSE), which is almost similar on the training data (MSE = 5756.49) and for the test data (MSE = 5714.19). To limit the effects of overfitting, we can use k-fold cross-validation (k = 5), but this procedure calculates a high MSE of 5877.10, indicating that the demand of bike rental for registered user is not an excellent candidate for linear regression.

The R-squared value from the fitted random forest regression on the response data is much higher than the linear model is 0.81. However, while the MSE on the training set is 659.69, the MSE for the test set jumped significantly to 3338.04. The regression for the response data is over fit. Performing a k-fold cross validation with 5 folds still showed a higher MSE of 3711.93. We can conclude that neither linear or random forest regression is a good fit to model this data.

However, from the feature importance we can have a sense of which variables the model found to have the largest impact on mean decrease node impurity during the training. We can see that Temperature and HOUR_17 [indicate that time is 5 PM] were more important than the other variables in decreasing node impurity.

The R-squared value from the fitted gradient boosting regression on the response data is slightly higher than the random forest model, is 0.86. However, while the MSE on the training set is 342.64, the MSE for the test set jumped significantly to 2411.35. The regression for the response data is over fit. Performing a k-fold cross validation with 5 folds still showed a higher MSE of 2900.61.

However, from the variable importance are shown below to give a sense of which variables the model found to have the largest impact. We can see that Humidity and Temperature were more important than the other variables.

We can conclude that gradient boosting regression is a good fit to model the demand for bike rental for casual user.

4.7 Future Work

While the correlations shown in this paper are certainly intriguing, the real value of this analysis lies in the predictive models' application to new data points. Any city government can use this report on demand for bike rental data in similar socio-economic condition and predict the demand of a bike share model.

Recommendations for further action based on this analysis:

• For Registered User:

Conduct a study to determine the root cause if there is any relationship between a stating station and ending station and then predict the demand for such combination.

• For Casual User:

Investigate if the number of casual user can be converted in to registered user so that the demand can be prediction with more accuracy.

Future opportunities to continue this work include further training of the regression model on similar city with more accurate dataset to determine if the demand for bike rental holds good.

Incorporating income data may also add accuracy to the regressions, allowing Department of transport to pinpoint the demand for any station. We can also determine if these correlations are standardized across station by applying the regression models to similar station.