MULTIVARIATE DATA ANALYSIS

BIKE SHARING DEMAND

Abhilash Ugaonkar

Stevens Institute of Technology

INTRODUCTION

Bike sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city.

Using these systems, people are able to rent a bike from one location and return it to a different place on an as-needed basis. Currently, there are over 500 bike-sharing programs around the world.

OBJECTIVES

- To combine historical usage patterns with weather data in order to forecast hourly bike rental demand for the Capital Bikeshare program in Washington, D.C.
- To help Capital Bikeshare better understand demand and allocate bike resources accordingly.

DATASET INFORMATION

The dataset was obtained from Kaggle and it provides hourly rental data spanning two years for the Capital Bikeshare program in Washington, D.C.
It contains over 10,000 observations and 10 attributes.

The data contains various attributes such as:

- **datetime** - hourly date and timestamp
- **season** - season
    - 1: winter
    - 2: spring
    - 3: summer
    - 4: fall
- **holiday** - represents whether the day is considered a holiday
- **workingday** - represents whether the day is a working day
- **weather** - weather
    - 1: clear, few clouds, partly cloudy
    - 2: mist + cloudy, mist + broken clouds, mist + few clouds, mist
    - 3: light rain, light rain + thunderstorm + scattered clouds, light rain + scattered clouds
    - 4: heavy rain + ice pellets + thunderstorm + mist, snow + fog
- **temp** - temperature in Celsius
- **atemp** - feels-like temperature in Celsius
- **humidity** - relative humidity
- **windspeed** - wind speed
- **count** - number of total rentals

EXPLORATORY DATA ANALYSIS

VARIABLE IDENTIFICATION

The **Predictor** variables in the given dataset are,
Datetime, Season, Holiday, Workingday, Weather, Temp, Atemp, Humidity, Windspeed

and the **Target** variable is,
Count

Data Type and Variable Category of the variables are as below,

Variable Name	Data Type	Variable Category
season	Numeric	Categorical
holiday	Numeric	Categorical
workingday	Numeric	Categorical
temp	Numeric	Continuous
atemp	Numeric	Continuous
humidity	Numeric	Continuous
windspeed	Numeric	Continuous
count	Numeric	Continuous

UNIVARIATE ANALYSIS

Since the target variable is continuous, we need to understand the central tendency and spread of the variable. These are measured using statistical metrics and visualization methods as below.

Statistical Metrics:

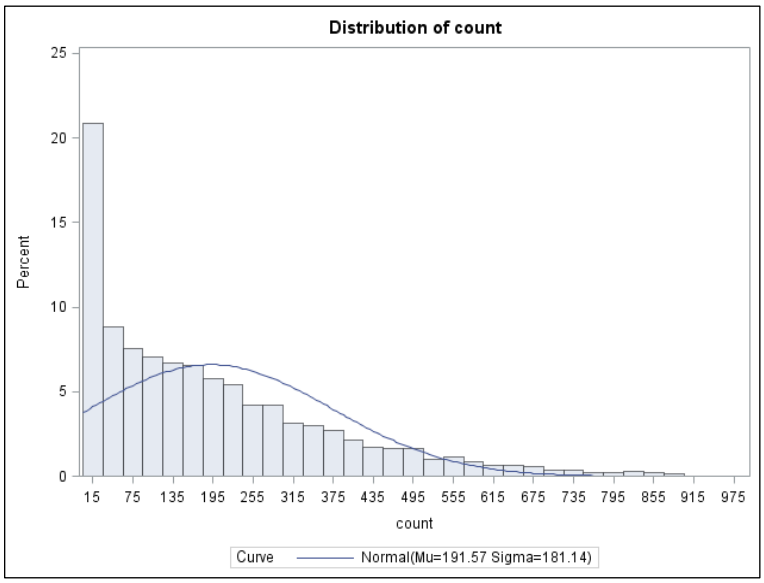
Moments			
N	10886	Sum Weights	10886
Mean	191.574132	Sum Observations	2085476
Std Deviation	181.144454	Variance	32813.3132
Skewness	1.24206621	Kurtosis	1.30009295
Uncorrected SS	756696168	Corrected SS	357172914
Coeff Variation	94.5558004	Std Error Mean	1.73616452

Basic Statistical Measures			
Location		Variability	
Mean	191.5741	Std Deviation	181.14445
Median	145.0000	Variance	32813
Mode	5.0000	Range	976.00000
		Interquartile Range	242.00000

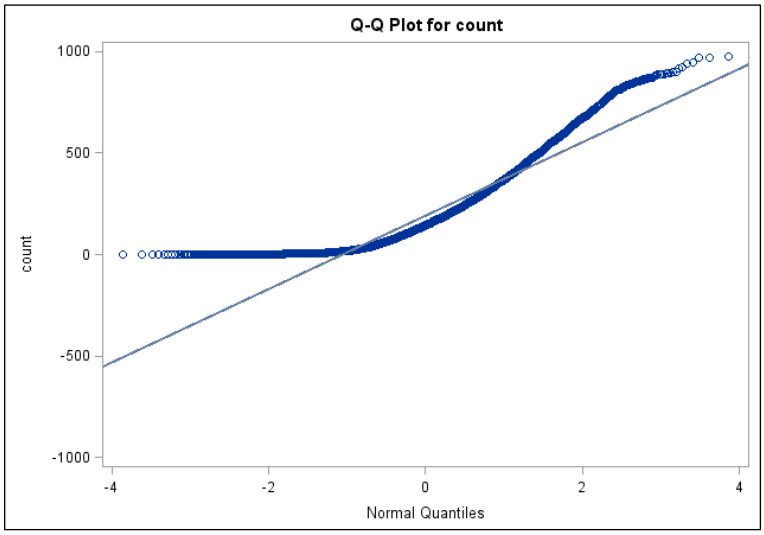
As evident from the above statistical measures, the mean is greater than the median and the measure of skewness is positive, indicating presence of right skewness.

Visualization:

We can verify the above by plotting a histogram,



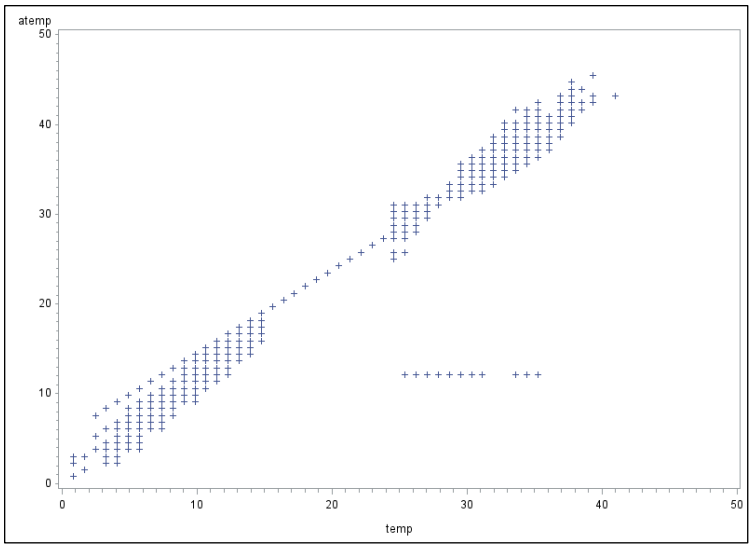
Hence, the distribution is indeed right skewed.
We continue univariate analysis of the target variable by graphing the quartile-quartile plot.



BI-VARIATE ANALYSIS

Scatter plots were graphed for every combination of continuous variables and it was found that the variables temp and atemp have a strong positive correlation.

Scatter plot between temp and atemp:



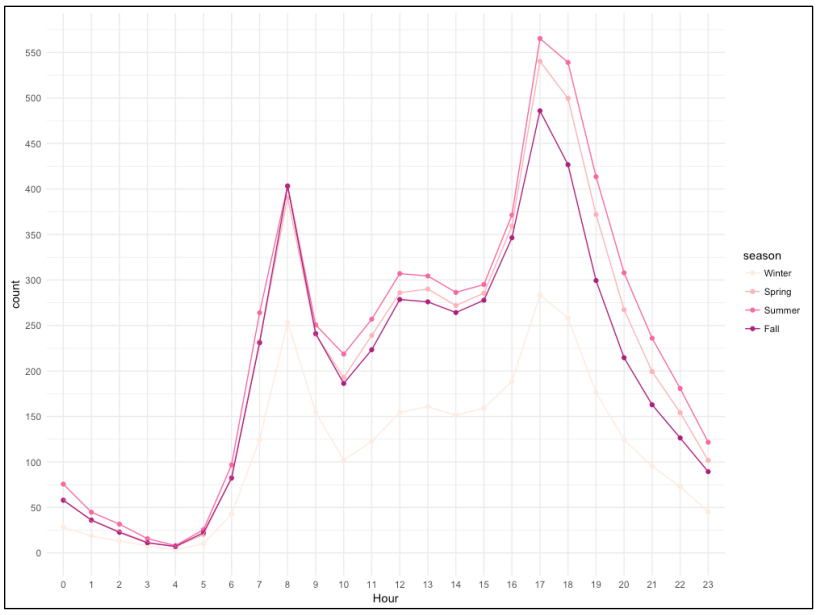
To verify this, the Pearson Correlation Coefficients between the variables were computed.

Pearson Correlation Coefficients, N = 10886 Prob > |r| under H0: Rho=0		
	atemp	**temp**
atemp	1.00000	0.98495 <.0001
temp	0.98495 <.0001	1.00000

As evident from the above table, temp and atemp are linearly correlated. We need to transform these variables to linearly uncorrelated variables to remove multicollinearity.

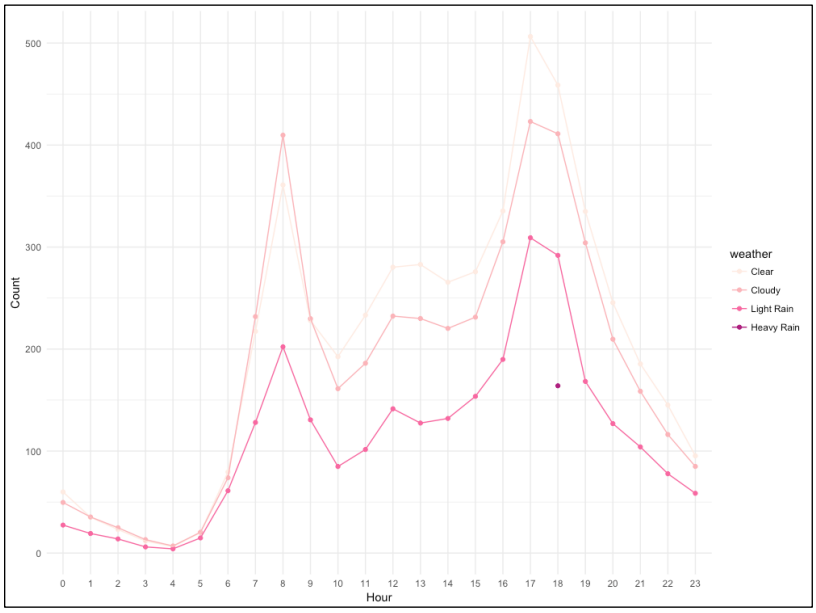
MULTI-VARIATE ANALYSIS

We perform some multivariate analysis to get an idea about the distribution of demand.
First, we plot a line graph between hour and count by season to get an idea of hourly and seasonal demands.



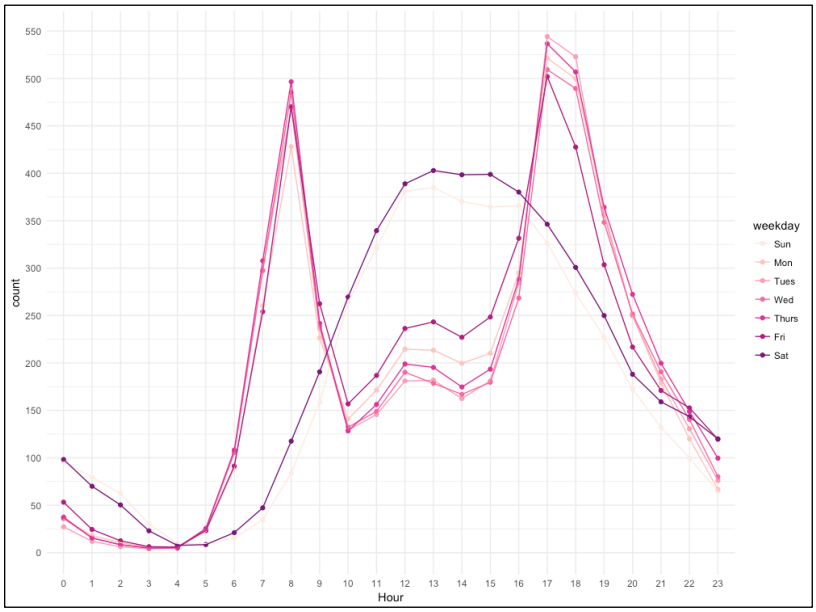
From the above graph, it's clear that the demand peaks at 8 in the morning and 5 in the evening. Also, there's more demand in summer when compared to that in winter.

Since weather could also influence the demand, we plot a line graph between hour and count by weather to visualize the trend.



From the graph, weather appears to have influenced the demand as the count by hour for clear weather is significantly higher when compared to when it's raining.

The day of the week could also influence when bikes are rented. Visualizing this using a line graph,



We see that during weekdays, the demand peaks at 8 am and 5 pm whereas during the weekends, the distribution is smoother and peaks mid-day.

OUTLIER DETECTION & TREATMENT

Since there were no missing values in the dataset, we proceed directly to the outlier detection process.

There are two types of outliers:

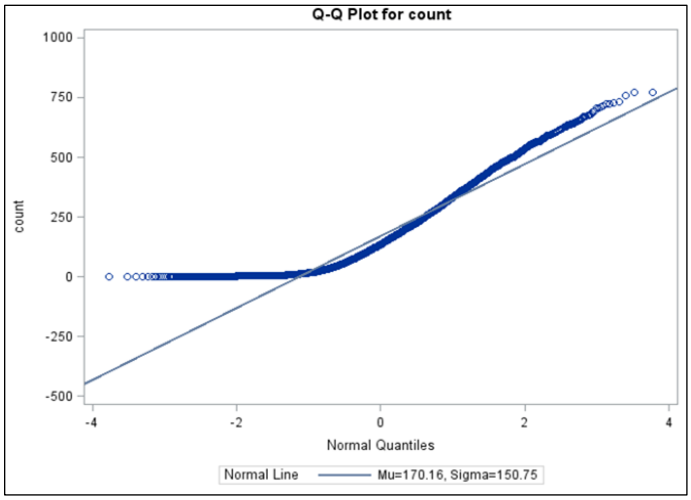
- Univariate outliers
- Multivariate outliers

Declaring an observation as an outlier based on just one feature (univariate) could lead to unrealistic inferences and since the dataset is small, it seems more realistic to perform multivariate outlier analysis.

One of the methods to do multivariate outlier analysis is by using studentized residuals, which is the quotient resulting from the division of a residual by an estimate of its standard deviation.

Treatment of Outliers:
All the records with studentized residual values greater than an absolute value of 2 were deleted from the dataset.

To visualize the effect of removing outliers, we plot the quantile-quantile plot for the variable count after removal of outliers.



As evident from the above plot, the target variable now has nearly the same characteristics as that of the ideal theoretical distribution.

VARIABLE TRANSFORMATION

As mentioned earlier, variables temp and atemp are linearly correlated. Since one of the assumptions of linear regression is no presence of multicollinearity, we need to transform these variables into linearly uncorrelated variables, which can be done by performing Principal Component Analysis (PCA).

Principal Component Analysis converts a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.

The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance, i.e. it accounts for as much of the variability in the data as possible.

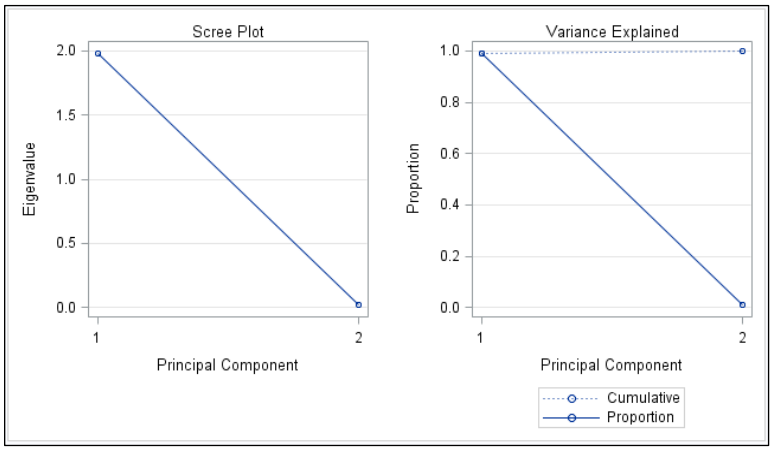
Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	1.98281432	1.96562864	0.9914	0.9914
2	0.01718568		0.0086	1.0000

The above result displays eigenvalues of the correlation matrix.
The first principal component explains about 99.14% of the total variance, whereas the second principal component explains only about 0.86% of the total variance.
Since first component explains 99.14% of the total variance, it provides a good summary of the data.

Eigenvectors		
	Prin1	Prin2
temp	0.707107	0.707107
atemp	0.707107	-.707107

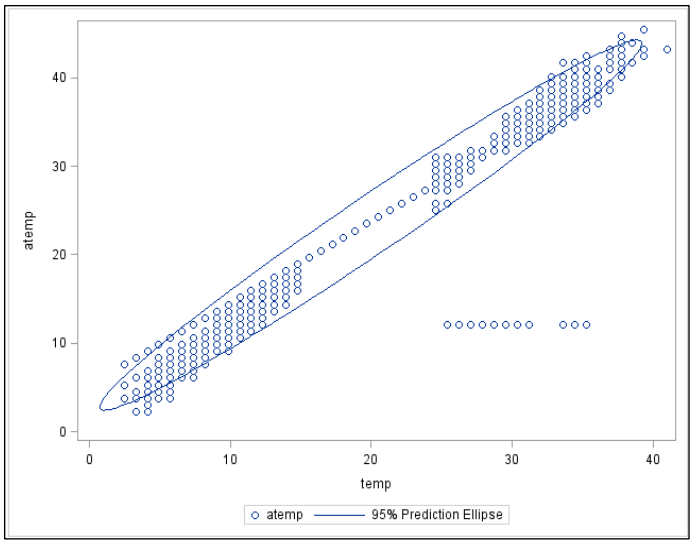
From the eigenvectors matrix, we could represent the first principal component Prin1 as a linear combination of the original variables temp and atemp:

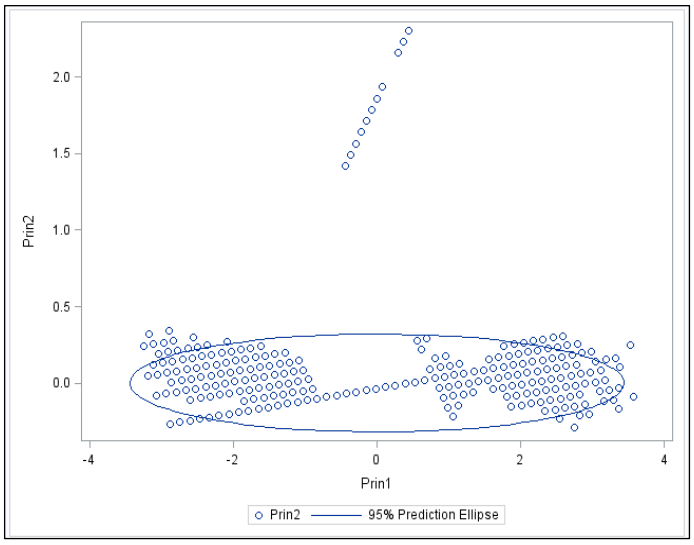
$$Prin1 = 0.707107 \times temp + 0.707107 \times atemp$$



The scree plot shows that Prin1 explains almost 100% of the total variance. So, we use Prin1 as a predictor variable and Prin2 is dropped.

To explain this further, we plot two scatter plots. One with temp and atemp variables, and the other with Prin1 and Prin2 components.



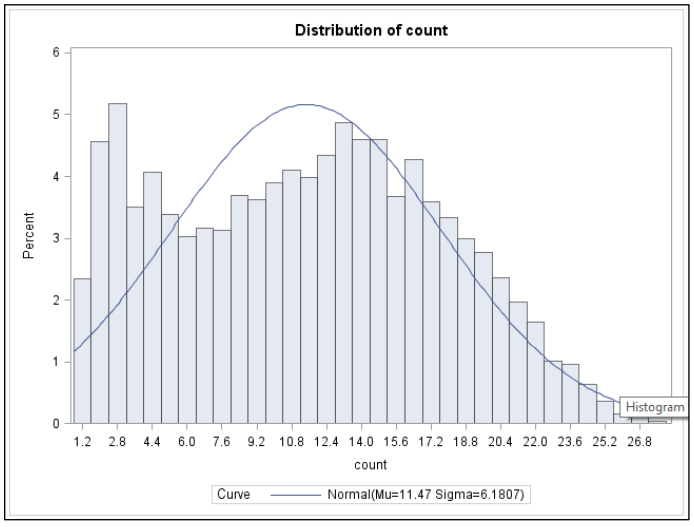


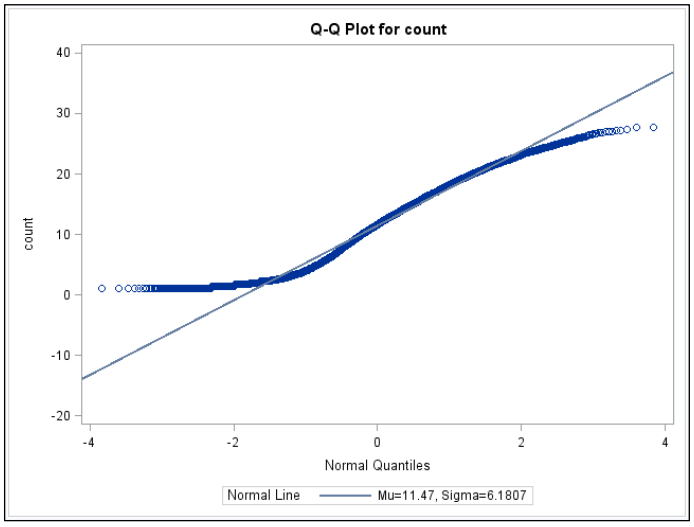
The scatter plot on the left indicates that temp and atemp are highly correlated, as most of the points lie along the line from the bottom left to the upper right of the graph.
Whereas in the scatter plot on the right, most of the points lie along the Prin1 axis, indicating that Prin1 explains most of the variance.

SQUARE ROOT TRANSFORMATION

Since the distribution of the target variable is right skewed, we perform square root transformation to normalize it.

To visualize the effect of square root transformation, we plot the histogram and quantile-quantile plots after performing the transformation on the variable count.





As we can see, the distribution is almost normal now and we can proceed further.

CREATING DUMMY VARIABLES

To include a categorical variable with more than two levels in a multiple regression prediction model, additional steps are needed to ensure that the results are interpretable. These steps include recoding the categorical variable into a number of separate, dichotomous variables. This recoding is called creation of dummy variables.

We create dummy variables for the following variables,

Season

The variable season is categorical containing 4 levels representing winter, spring, summer and fall.

In order to make sure the results were interpretable; we recode the variable into 4 separate variables.

Weather

Like season, the variable weather is categorical containing 4 levels representing weather types like clear, cloudy, light rain and heavy rain. We recode this variable as well into 4 separate variables to represent the aforementioned types.

Datetime

Unlike the above two variables, the variable datetime is not categorical. However, this variable need to be re-coded to make sure the results are interpretable. We recode the variable into 24 separate variables with each variable representing hour of a day, and 7 separate variables with each variable representing days of a week.

MODELING

The next step is to build a model for predicting the hourly demand. We do this using the Multiple Linear Regression algorithm.

MULTIPLE REGRESSION

Multiple regression is a generalization of linear regression by considering more than one independent variable, and a specific case of general linear models formed by restricting the number of dependent variables to one.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_\rho X_{i\rho} + \in_i$$

We select the variables for the model by performing stepwise regression, which involves a series of alternating forward selection and backward elimination steps to add and remove variables to the model and find all significant variables.

After performing regression with stepwise selection, we obtained 33 variables for the regression model.

The parameter estimates of which are below,

Parameter Estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**VIF**
Intercept	1	16.00617	0.19061	83.97	<.0001	0
workingday	1	-0.37386	0.08001	-4.67	<.0001	1.76308
Prin1	1	1.2076	0.03416	35.35	<.0001	3.02065
humidity	1	-0.02757	0.00191	-14.47	<.0001	1.74502
windspeed	1	-0.01977	0.00367	-5.39	<.0001	1.17708
summer	1	-0.49285	0.08725	-5.65	<.0001	1.8422
fall	1	1.11768	0.08728	12.81	<.0001	1.87015
winter	1	-1.29779	0.10214	-12.71	<.0001	2.58136
clear	1	0.18378	0.06792	2.71	0.0068	1.35602
lt_rain	1	-2.17885	0.11533	-18.89	<.0001	1.27773
sat	1	0.36051	0.10125	3.56	0.0004	1.57284
thu	1	0.3111	0.08497	3.66	0.0003	1.16636
fri	1	0.7224	0.08396	8.6	<.0001	1.14979
h0	1	-6.59468	0.15518	-42.5	<.0001	1.34303
h1	1	-8.12707	0.15562	-52.22	<.0001	1.35077
h2	1	-9.18162	0.15694	-58.5	<.0001	1.35638
h3	1	-10.25341	0.15942	-64.32	<.0001	1.35485
h4	1	-10.69778	0.15873	-67.4	<.0001	1.36984
h5	1	-8.84319	0.15778	-56.05	<.0001	1.3827
h6	1	-4.90472	0.15749	-31.14	<.0001	1.38625
h8	1	5.33086	0.20334	26.22	<.0001	1.18519
h9	1	0.958	0.154	6.22	<.0001	1.32547
h10	1	-1.31988	0.15426	-8.56	<.0001	1.30487
h11	1	-0.82541	0.15666	-5.27	<.0001	1.28817
h12	1	0.3811	0.15741	2.42	0.0155	1.28878
h14	1	-0.63893	0.1596	-4	<.0001	1.29479
h16	1	2.25645	0.15646	14.42	<.0001	1.31076
h17	1	6.04675	0.17077	35.41	<.0001	1.2429
h18	1	5.14569	0.17022	30.23	<.0001	1.23842
h19	1	2.56229	0.15578	16.45	<.0001	1.29364
h20	1	0.52576	0.15326	3.43	0.0006	1.31007
h21	1	-1.19528	0.15365	-7.78	<.0001	1.31672
h22	1	-2.58595	0.15398	-16.79	<.0001	1.32515
h23	1	-4.46084	0.15438	-28.9	<.0001	1.33478

EVALUATION

PREDICTION OF DEMAND FOR TEST DATASET

Using the multiple regression model, we predict the demand for the records in the test dataset.
After obtaining the predicted values, we square them to account for the square root transformation done earlier.

ACCURACY OF MODEL

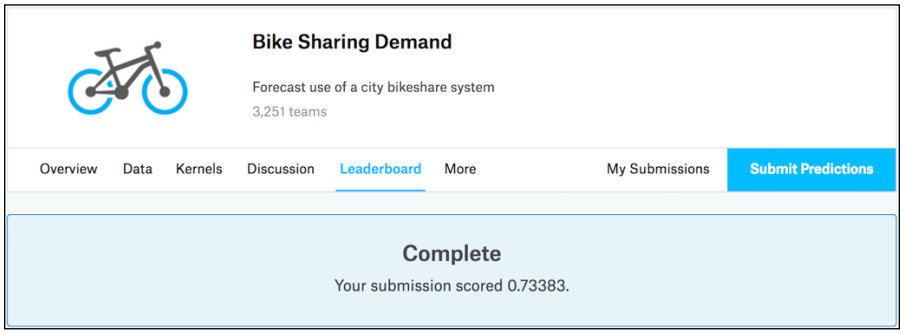
To calculate the accuracy of the model, we use Root Mean Square Logarithmic Error (RMSLE).

We calculate RMSLE using the following formula,

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i+1)-\log(a_i+1))^2}$$

where,
n is the number of records in test set
p is the predicted count
a is the actual count

The RMSLE was computed to be 0.73383.



INFERENCE

According to the parameter estimates computed using multiple regression,

- The demand for bikes was found to increase significantly during rush hours on weekdays. (7 am – 8 am and 4 pm – 5 pm).
- The rainy weather had a negative impact on the bike demand.
- Also, there was relatively higher demand for bikes in fall.

DEPLOYMENT

Since our model can predict the demand, we can use it to,

- Allocate bikes and plan operational activities.
- Plan promotional activities during winter to stimulate more demand.

REFERENCES

[1] Capital Bike Share. Bike Sharing Demand in Washington D. C.. [DB/OL]https://www.kaggle.com/c/bike-sharing-demand

[2] Afifi, A. & S. May & V. A. Clark *Practical Multivariate Analysis Fifth Edition*[M]. NW: CRC Press, 2012: 119-154 & 357-376