Bike Sharing Demand

for

Capital BikeShare

Created By:

Prathamesh Parchure

Stevens Institute of Technology

Introduction

- Bike sharing system is a mean 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 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.

Objective

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

Data Source

- https://www.kaggle.com/c/bike-sharing-demand

About the Dataset

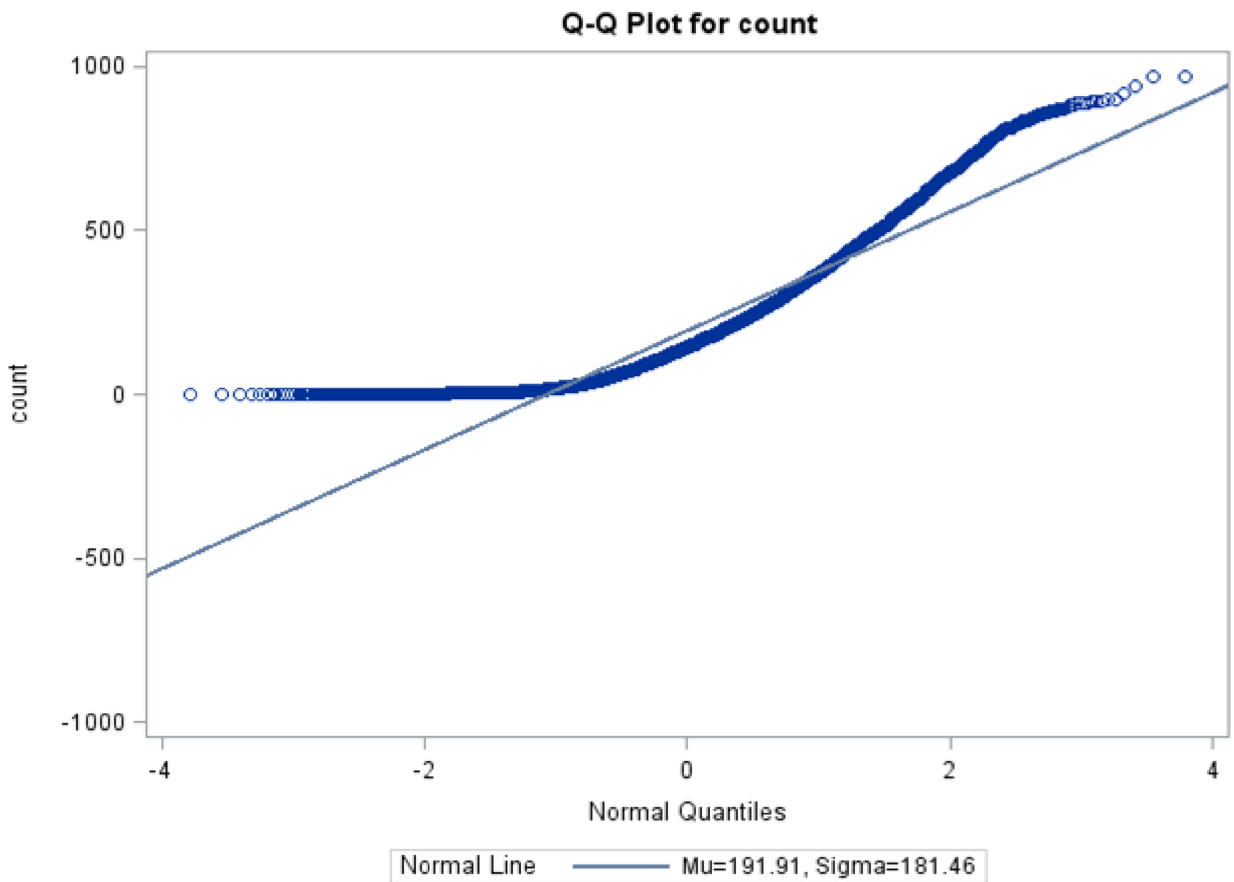
- The dataset represents 2 years of Capital bike demand in Washington D.C.
- Over 10,000 observations and 10 attributes.
- The data contains various attributes as:
  - datetime, season, holiday, working day, weather, temp, atemp, humidity, windspeed, count
- **Metadata**
  - **datetime**: hourly date and timestamp
  - **season**: categorical variable
    - 1 = Winter, 2 = Spring, 3 = Summer, 4 = Fall
  - **holiday**: whether the day is considered a holiday
  - **workingday**: whether the day is neither a weekend or holiday
  - **weather**: categorical variable

- 1 = clear, 2 = cloudy, 3 = light rain, 4 = heavy rain

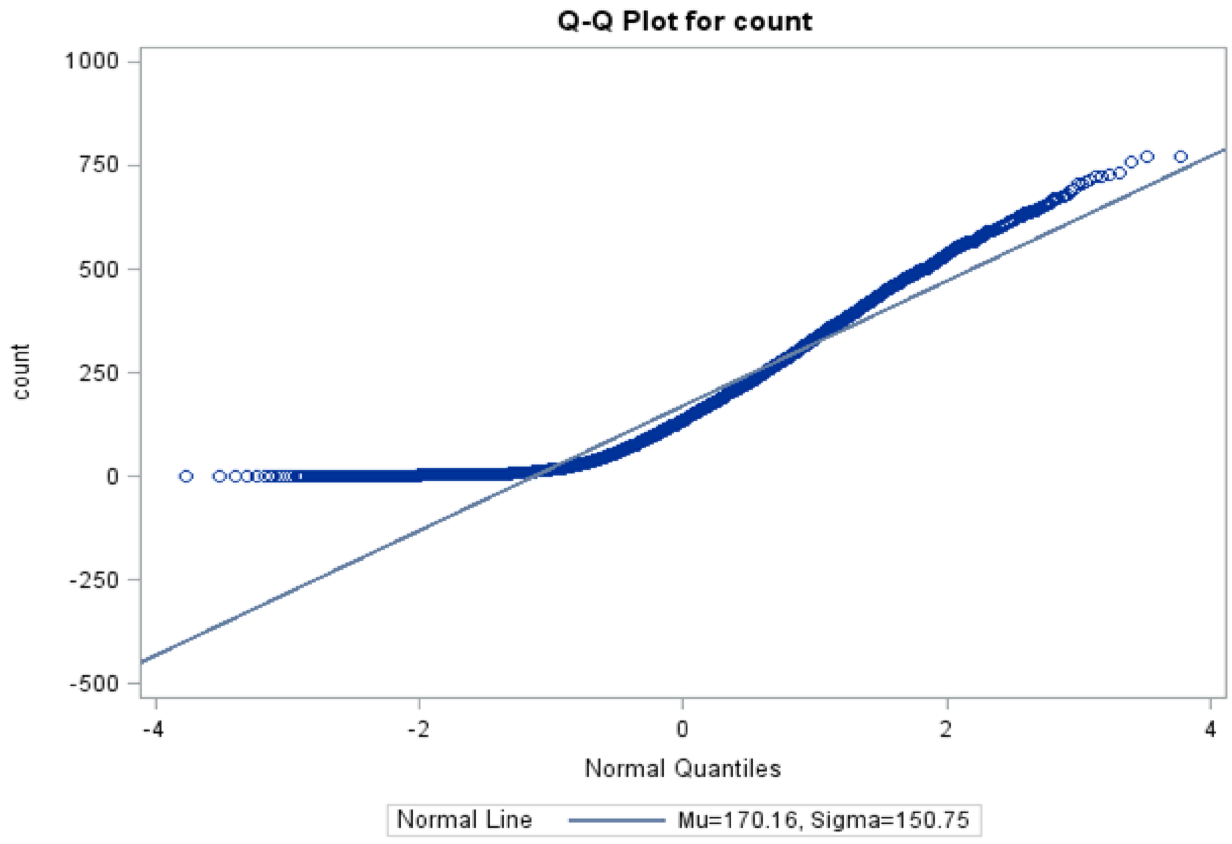
- **temp**: temperature in Celsius
- **atemp**: feels-like temperature in Celsius
- **humidity**: relative humidity
- **windspeed**: wind speed
- **count**: number of total rentals

Data Transformation

- Perform One-Hot-Encoding for categorical variables: Season, Weather and Datetime.
- Perform Cross Validation by splitting the original bikeshare dataset into train and test
- Remove the outliers using Studentized Residual method.



QQ-Plot Before Removing the Outliers



QQ-Plot After Removing the Outliers

- Removed outliers using studentized residuals.
- Studentized residual is the quotient resulting from the division of a residual by an estimate of its standard deviation.
- All records with studentized residual values greater than 2 or lower than -2 were deleted.

Analysis

After performing Regression Analysis on the data, it was observed that the Variation Inflation Factor for the variables 'temp' and 'atemp' is greater than 30.

On further analysis and looking at the Pearson Correlation analysis, it was observed that 'temp' and 'atemp' are are highly correlated.

The CORR Procedure

2 Variables:	temp atemp

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
temp	7600	19.97304	7.86291	151795	2.46000	41.00000
atemp	7600	23.35647	8.54134	177509	2.27500	45.45500

Pearson Correlation Coefficients, N = 7600
Prob > |r| under H0: Rho=0

	temp	atemp
temp	1.00000	0.98281 <.0001
atemp	0.98281 <.0001	1.00000

To reduce the multicollinearity and to perform dimensionality reduction, principal components are computed. 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

The PRINCOMP Procedure

Observations	7600
Variables	2

Simple Statistics

	temp	atemp
Mean	19.97304211	23.35646645
StD	7.86290987	8.54133509

Correlation Matrix

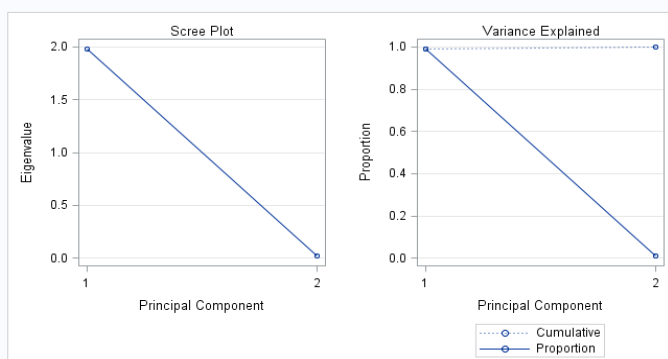
	temp	atemp
temp	1.0000	0.9828
atemp	0.9828	1.0000

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

Eigenvectors

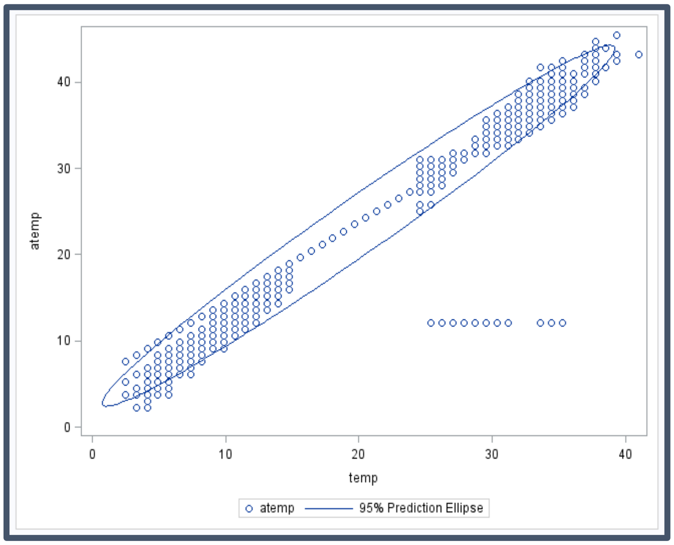
	Prin1	Prin2
temp	0.707107	0.707107
atemp	0.707107	-.707107

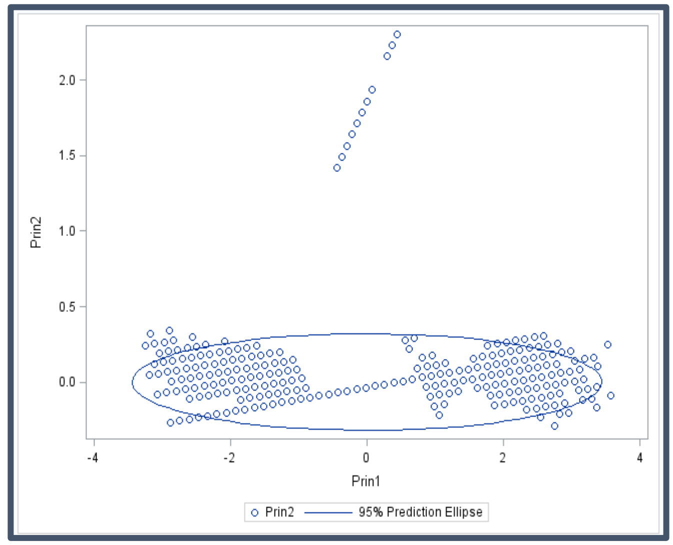


The first principal component explains about 99.14% of the total variance, providing a good summary of data, whereas the second principal component explains only about 0.86% of the total variance. Since first component explains 99.14% of the total variance, it provides a good summary of the data.

From the eigenvectors matrix, we could represent the first principal component Prin1 as a linear combination of the original variables temp and atemp: $Prin1 = 0.707107 \times temp + 0.707107 \times atem$

The 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.

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.

$$Yi = \beta 0 + \beta 1Xi1 + \beta 2Xi2 + \cdots + \beta \rho Xi\rho + \in i$$

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.

Stepwise Variable Selection Method is used to select significant variables, while removing the insignificant ones.

Summary of Stepwise Selection								
Step	Variable Entered	Variabl Remove d	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Prin1		1	0.1648	0.1648	15456	1499.2	<.0001
2	humidity		2	0.1152	0.28	12278	1215.6	<.0001
3	h17		3	0.0735	0.3535	10250	864.13	<.0001
4	h18		4	0.057	0.4106	8678	735.03	<.0001
5	h8		5	0.055	0.4656	7161	782.08	<.0001
6	h19		6	0.0317	0.4973	6288.5	478.57	<.0001
7	fall		7	0.0297	0.527	5471.5	476.27	<.0001
8	h16		8	0.0222	0.5492	4861.5	373.35	<.0001
9	h7		9	0.0225	0.5717	4241.4	399.46	<.0001
10	h9		10	0.022	0.5937	3636.6	410.63	<.0001
11	h20		11	0.0168	0.6105	3173.9	328.01	<.0001
12	h12		12	0.0111	0.6217	2868.2	223.57	<.0001
13	h13		13	0.01	0.6317	2593.7	206.33	<.0001
14	h15		14	0.0098	0.6415	2324.8	207.63	<.0001
15	h21		15	0.0102	0.6517	2045.6	221.88	<.0001
16	h11		16	0.0119	0.6635	1720.4	267.14	<.0001
17	h14		17	0.0129	0.6765	1365.4	303.17	<.0001
18	h10		18	0.0151	0.6916	951.13	370.65	<.0001

Summary of Stepwise Selection								
Step	Variable Entered	Variabl Remove d	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
19	h22		19	0.0101	0.7016	674.79	256.21	<.0001
20	lt_rain		20	0.0065	0.7082	496.37	169.77	<.0001
21	h6		21	0.0052	0.7134	355.5	136.85	<.0001
22	h23		22	0.0051	0.7185	216.54	137.45	<.0001
23	spring		23	0.0026	0.721	148.15	69.25	<.0001
24	h0		24	0.0011	0.7222	118.56	31.21	<.0001
25	fri		25	0.0009	0.723	96.775	23.56	<.0001
26	workingd ay		26	0.0008	0.7238	77.579	21.06	<.0001
27	clear		27	0.0006	0.7244	61.652	17.85	<.0001
28	windspee d		28	0.0006	0.725	47.319	16.29	<.0001
29	h1		29	0.0003	0.7253	41.573	7.73	0.0054
30	h5		30	0.0003	0.7256	36.648	6.92	0.0085
31	sun		31	0.0002	0.7257	33.577	5.07	0.0244
32	winter		32	0.0001	0.7259	32.314	3.26	0.0709
33	h2		33	0.0001	0.726	31.154	3.16	0.0755

Evaluation

Prediction of Demand for Test Dataset

Using the multiple regression model, we predict the demand for the records in the test dataset.

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.

Conclusion

According to parameter estimates,

- Demand increases significantly during rush hours (7a.m. - 9a.m. & 5p.m. - 7p.m.)
- There's very low demand during rainy days.
- People tend to use more bikes during fall and spring.
- There's high demand during rush hours irrespective of season.

Application

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

- Capital Bike Share. Bike Sharing Demand in Washington D. C. [DB/OL]https://www.kaggle.com/c/bikesharing-demand
- Afifi, A. & S. May & V. A. Clark Practical Multivariate Analysis Fifth Edition[M]. NW: CRC Press, 2012: 119- 154 & 357-376