STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Analysis of factors affecting Bike Rentals in Washington D.C.

Prof. Khasha Dehnad
BIA 652-B

Team:

Bhagyashree Shende
Nithin Das
Purva Khopkar

TEAM

Purva Khopkar

Bhagyashree Shende

Nithin Das

Table of Contents

Introduction

What is Bike Sharing System?

- Means of renting bicycles via a network of kiosk locations throughout Washington D.C.
- People rent a bike from a one location and return it to a different place on an as-needed basis
- Bike sharing solution comes with some problems to be solved :
  - ❖ Vandalism
  - ❖ Parking problems
  - ❖ Logistic issues
- Increasing revenue for the company also one of the goals
- Thus, it is important to study the factors affecting the demand of bikes

Problem Statement

Analysis of the factors affecting the demand of bike rentals in Washington D.C.

Proposed Solution

Developing a classification model to determine the factors affecting demand of bike rentals

Data Source & Description

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

Features:

Name of Feature	Type	Value & Meaning
dteday	Categorical	Date
season	Categorical	1 = spring, 2 = summer, 3 = fall, 4 = winter
month	Categorical	1 to 12 for January to December
year	Categorical	0 = 2011, 1 = 2012
holiday	Categorical	1 = Holiday 0 = Not a Holiday
workingday	Categorical	1 = Not a weekend nor Holiday 0 = Holiday/Weekend
weather	Categorical	1 = Clear, 2 = Cloudy, 3 = Light Snow/Rain
temp	Numeric	normalized temperature in Celsius
atemp	Numeric	"feels like" normalized temperature in Celsius
humidity	Numeric	relative humidity
windspeed	Numeric	wind speed
casual	Numeric	number of non-registered user rentals
registered	Numeric	number of registered user rentals
cnt	Numeric	Total user rentals (Sum of casual and registered user rentals)

Dependent Variable : CNT

Data Pre-Processing

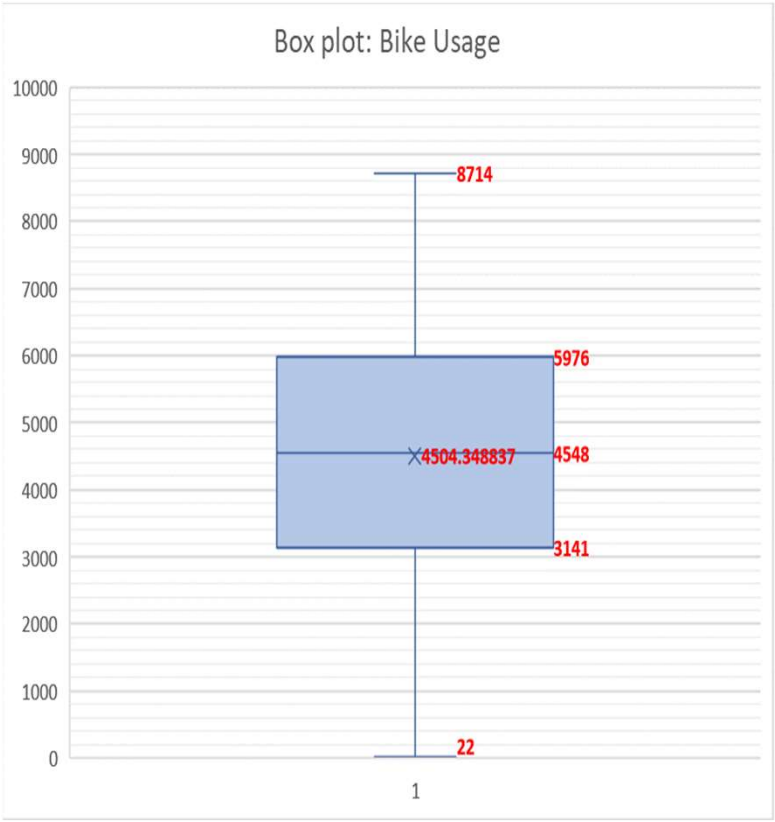
1. No missing values in dataset
2. Added a new column ‘Day’ by extracting the day value from date , range [1,31]
3. Removed Date column as it is redundant -> Year and Month is present

The MEANS Procedure

Variable	N Miss	N
instant	0	730
dteday	0	730
date	0	730
season	0	730
yr	0	730
mnth	0	730
holiday	0	730
weekday	0	730
workingday	0	730
weathersit	0	730
temp	0	730
atemp	0	730
hum	0	730
windspeed	0	730
casual	0	730
registered	0	730
cnt	0	730

- Outliers

Removed outlier of Bike Usage value of *22*

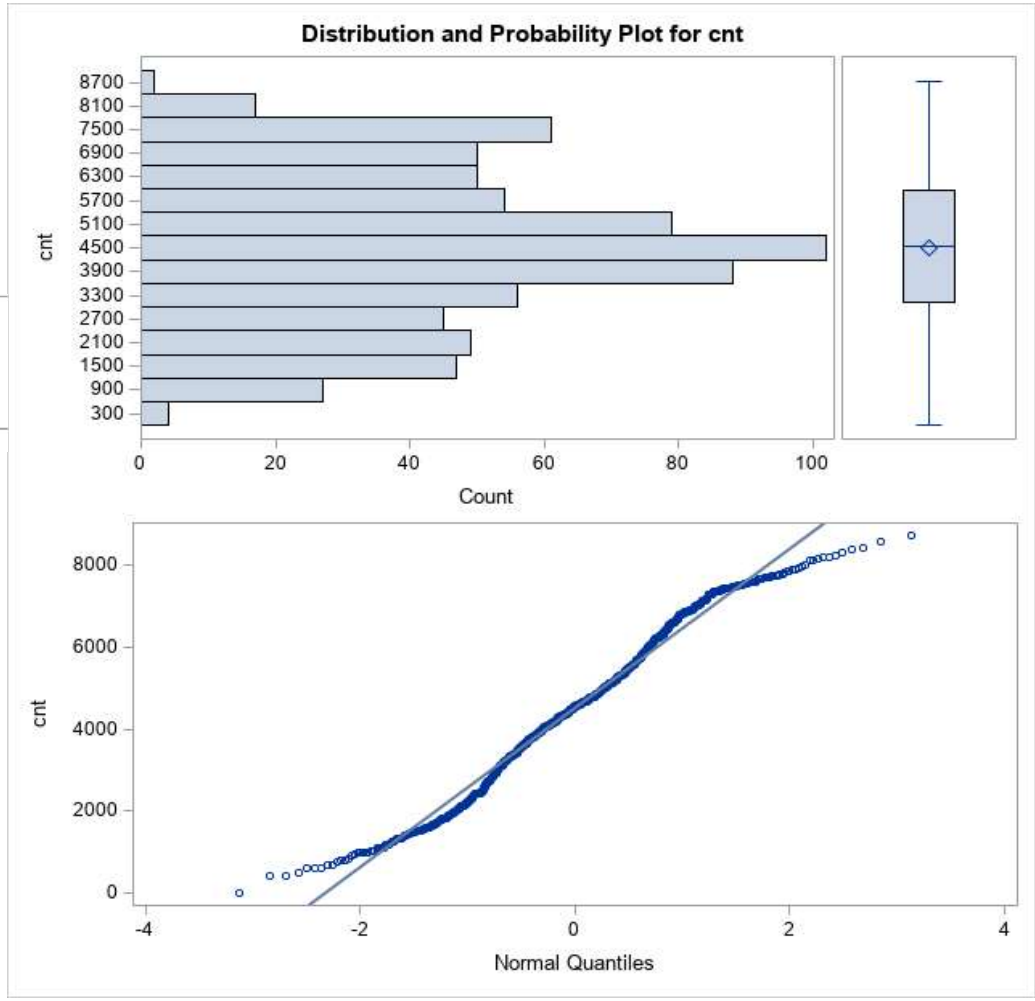


Exploratory Data Analysis

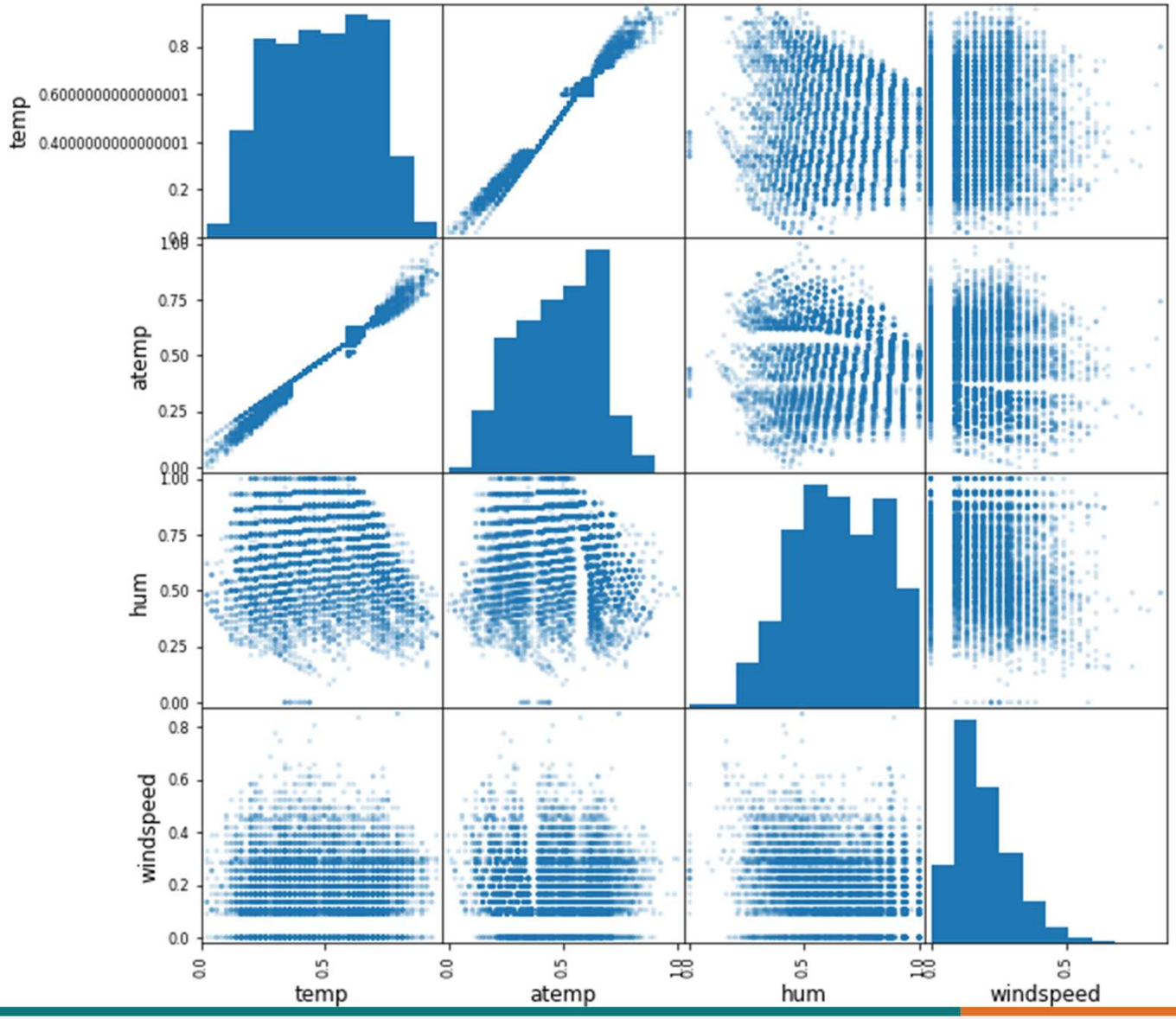
Checking for Normality of Dependent Variable

Using univariate analysis of dependent variable CNT, we find it is normally distributed

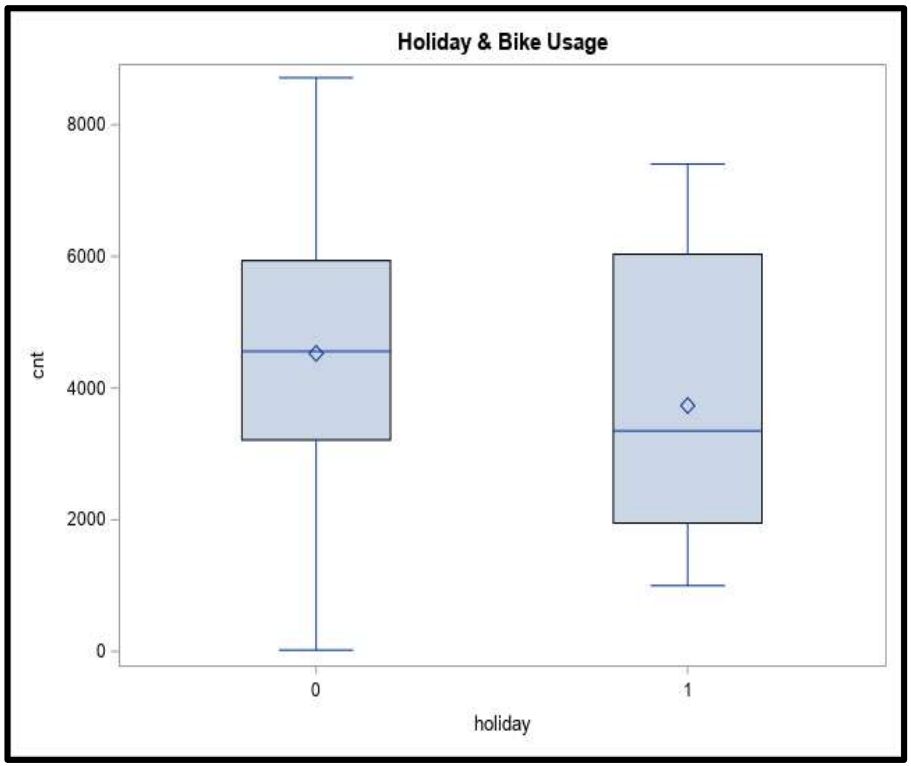
```
29
30 proc univariate data=daydata normal plot;
31 var  cnt;
32 run;
33
```

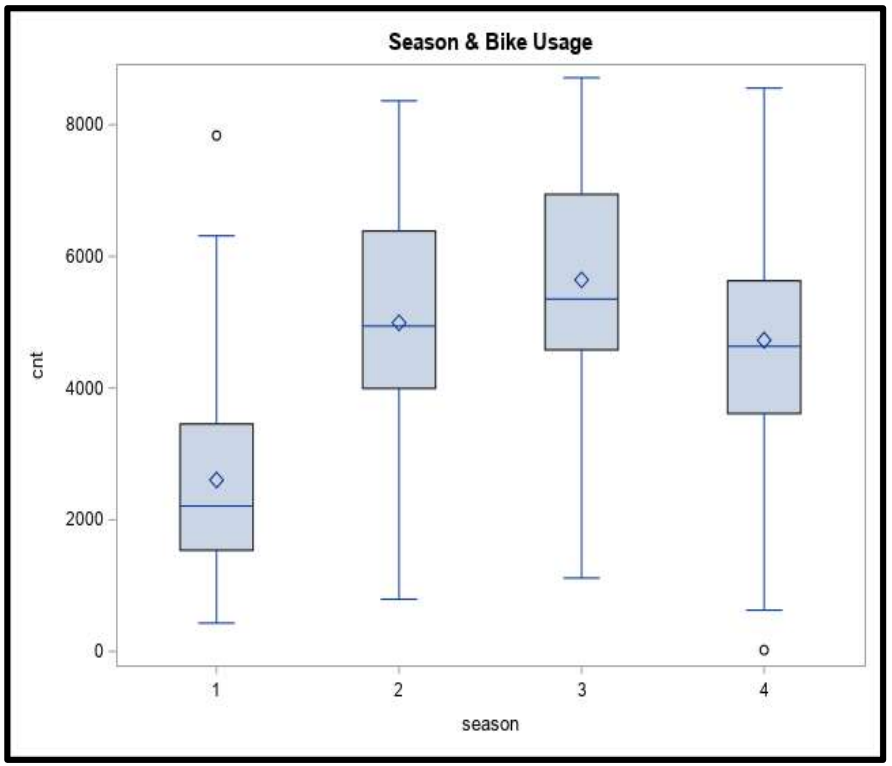


Relationship between Numeric Variables



Relationship of Categorical Variables with Dependent Variables



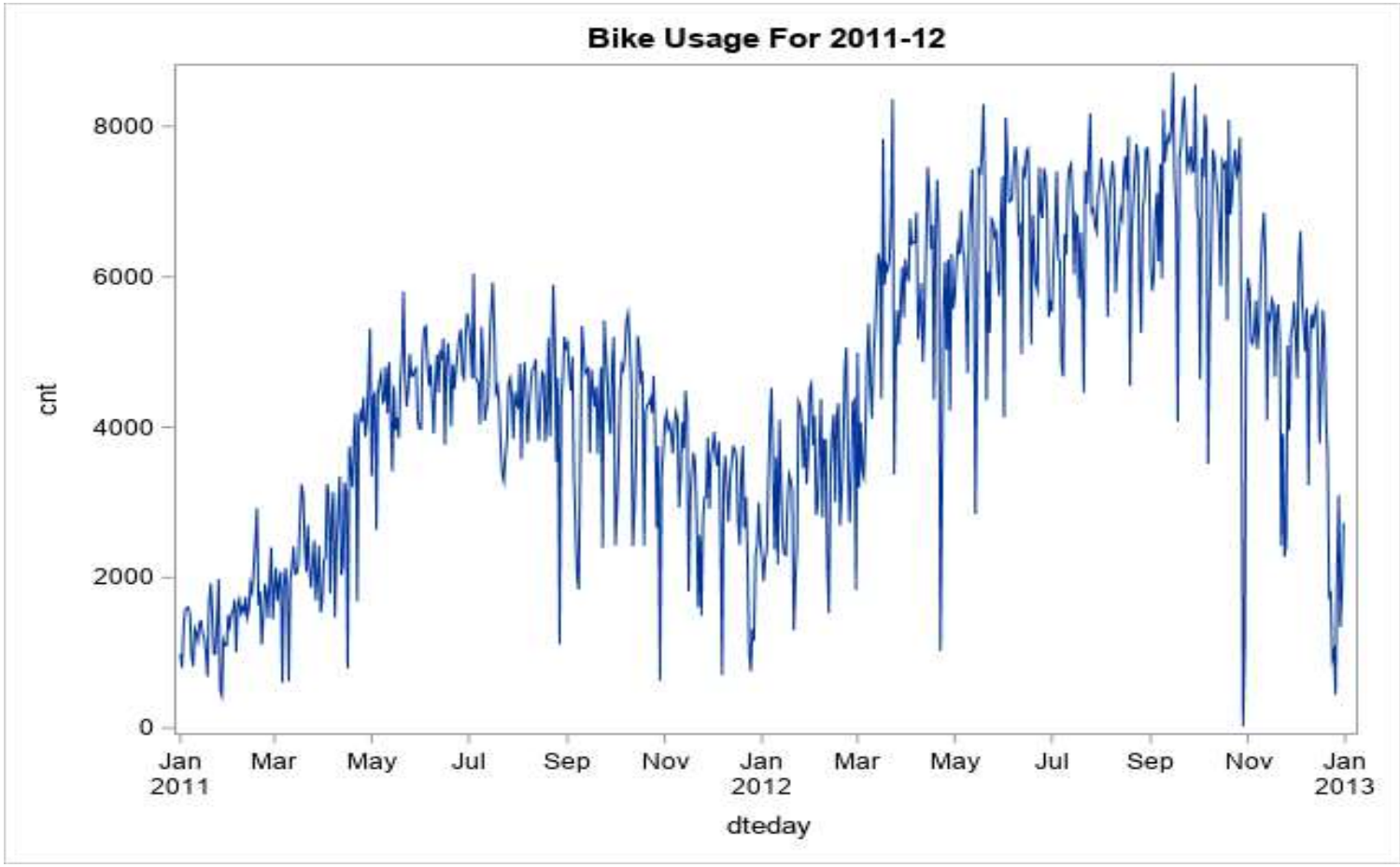


- Average demand is more on a non-holiday
- Count is normally distributed on a non-holiday
- May be there are a fixed group of people using bikes daily?

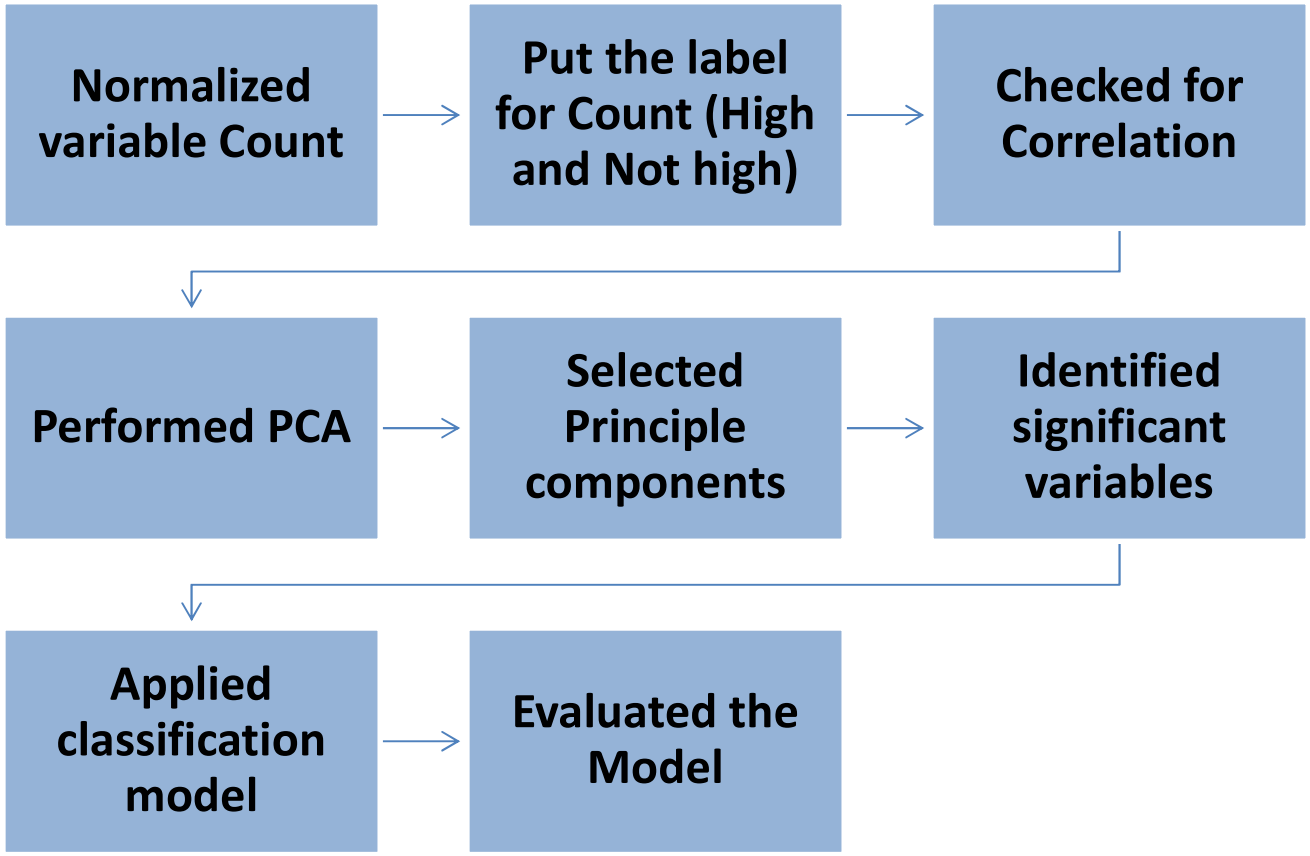
- Average demand is more in season 3 (Summer)

Relationship of Categorical Variables with Dependent Variables

(Contd.)



Workflow


Normalized variable Count
Put the label for Count (High and Not high)
Checked for Correlation
Performed PCA
Selected Principle components
Identified significant variables
Applied classification model
Evaluated the Model

Data Preparation

- **Dependent variable – cnt (Count of total rental bikes)**

Normalized the cnt variable using min-max normalization function

- **Labelling cnt variable as High and Not High (Mean + Standard dev)**

Mean -> 0.492513 S.D -> 0.233018

High > **0.725531**

Not High < **0.725531**

Prediction	Frequency	Percent	Cumulative Frequency	Cumulative Percent
HIGH	143	19.59	143	19.59
NOT HIGH	587	80.41	730	100.00

- **Checking the correlation between the variables**

Numeric values in the dataset :
temp, atemp, humidity , windspeed

Pearson Correlation Coefficients, N = 730 Prob > |r| under H0: Rho=0				
	temp	atemp	hum	windspeed
temp	1.00000	0.99171 <.0001	0.12798 0.0005	-0.15756 <.0001
atemp	0.99171 <.0001	1.00000	0.14082 0.0001	-0.18360 <.0001
hum	0.12798 0.0005	0.14082 0.0001	1.00000	-0.25511 <.0001
windspeed	-0.15756 <.0001	-0.18360 <.0001	-0.25511 <.0001	1.00000

Performing PCA

- Standardized the variables with mean = 0 and Standard Deviation =1
- Conducted the PCA to get the principal components

Eigenvectors				
	Prin1	Prin2	Prin3	Prin4
temp	0.663490	0.248244	0.031327	-.705106
atemp	0.667997	0.225726	0.018479	0.708863
hum	0.219157	-.688287	0.691522	-.005376
windspeed	-.255990	0.643183	0.721439	0.017614

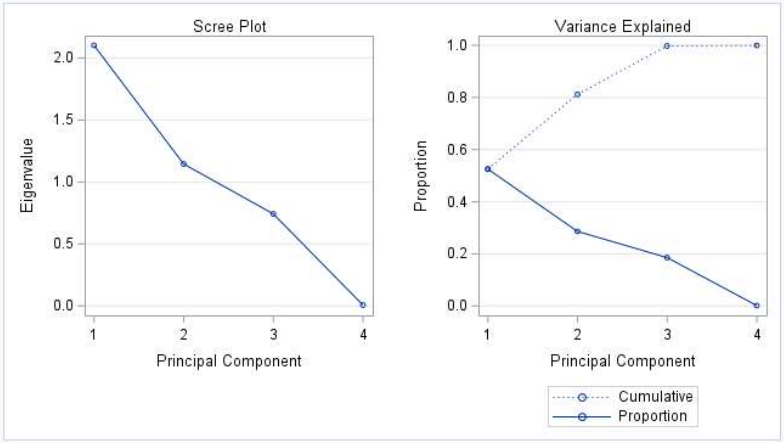
Evaluating the PCA results

- Selected the principal component 1 and 2

due to following reasons:

- Eigenvalue is greater than 1.
- The cumulative proportion is **0.8105**

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	2.10131375	0.96080938	0.5253	0.5253
2	1.14050437	0.39025129	0.2851	0.8105
3	0.75025308	0.74232429	0.1876	0.9980
4	0.00792879		0.0020	1.0000



Approach 1- Logistic Regression – Variable selection

- Variable selection using Logistic regression on data set
- We used forward selection method.
- Important variables from the regression model are : Month, Holiday, Weathersit, Prin1

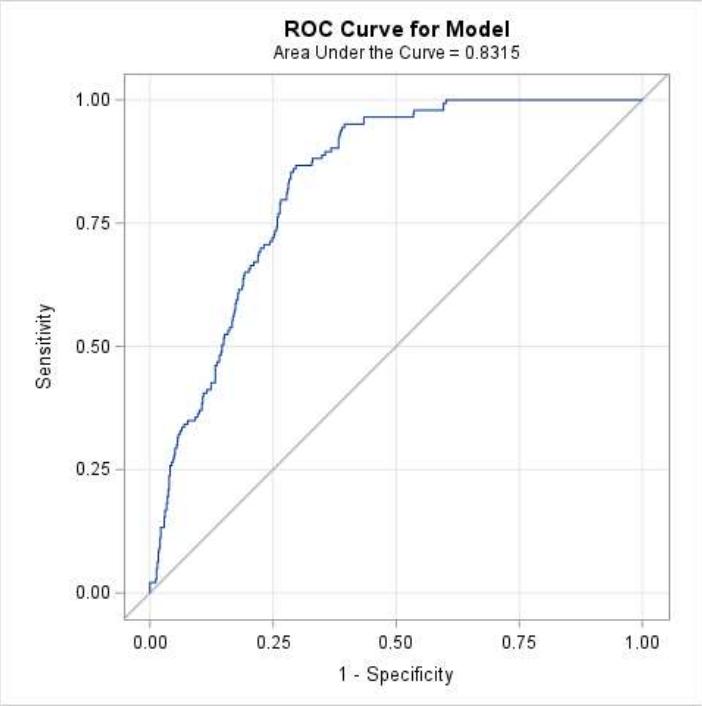
```
proc logistic data=princomp;
  class prediction date mnth yr season holiday weekday workingday weathersit;
  model prediction(event="High")=  date mnth yr season holiday weekday workingday weathersit prin1 prin2 yr
                / selection=forward;
          output out = pred p = phat upper=ucl lower=lcl;
run;
```

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
mnth	11	33.0616	0.0005
yr	1	0.0305	0.8613
season	3	7.3626	0.0612
holiday	1	5.1296	0.0235
weathersit	2	34.8781	<.0001
Prin1	1	3.9656	0.0464

Logistic Regression on selected variables

- We performed Regression model on selected variables

```
proc logistic data=princomp;
 class prediction mnth weathersit;
 model prediction(event="High")= mnth weathersit prinl holiday/lackfit ctable outroc =rocl;
          output out = pred p = phat upper=ucl lower=lcl;
run;
```



Classification Table								
Correct		Incorrect		Percentages				
Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
116	405	182	27	71.4	81.1	69.0	61.1	6.3

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
7.8012	7	0.3504

Approach 2- Random Forest – Variable selection

- Use Random Forest to verify important features

RandomForestClassifier (n_estimators=1000, random_state=0,min_samples_split=5, n_jobs=-1)

```
('date', 0.060550855883866966)
('season', 0.06045778370715398)
('yr', 0.3330848375770469)
('mnth', 0.10895788660044785)
('holiday', 0.008899895754418094)
('weekday', 0.03819861948918749)
('weathersit', 0.054368931074843085)
('workingday', 0.009528392123165396)
('Prin1', 0.23782489474990706)
('Prin2', 0.08812790303996342)
```

- *Setting threshold value to 0.1 gives 'yr', 'mnth', 'Prin1'*

- *" We have seen an increasing trend in the Bike Usages from 2011 to 2012"*
- *" Data shows decrease in bike usage shows decreasing trend between July and December each year"*

Logistic Regression with variable selection

- Divided dataset to training-test with 70%-30% proportion
- LogisticRegression(Bike Usage ~ Month+ Year +Prin1)

Model Evaluation

➢ Confusion Matrix

	Not High	High
Not High	167	12
High	17	24

➢ Sensitivity

Sensitivity= TP/(TP+FN) ~ approx. *93.29%*

➢ Specificity

Specificity = TN/(TN+FP) ~ approx. *58.5%*

➢ Model Accuracy

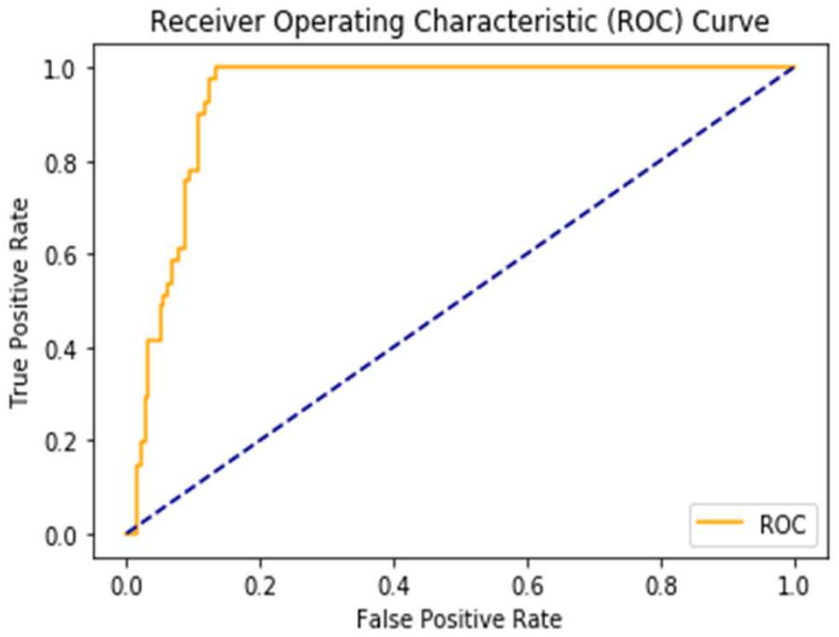
Accuracy= TP+ TN/(TP+TN+FP+FN)
~ approx. *86.81%*

ROC Curve

(Contd.)

- Area Under Curve

AUC : *0.94*



Highlights:

- *As the data contains 587 instances of "Not High" usage and 143 instances of "High" usage, there is a potential chance of class imbalance problem*
- *Therefore, we focus more on the Specificity of the model i.e. model performance in classifying "High" usage cases*
- *Now, we will compare our model with Null model*

Model Comparison and Inference

Model	Accuracy	Event = High	Event = Not High	AUC
Approach 1	71.4%	81.1%	69.0%	0.83
Approach 2	86.81%	58.5%	93.29%	0.94

- Model selection is depending upon the business requirement of the company if “High“ usage is the event of concern then Approach 1 can be accepted otherwise if “Not high” usage is the event of concern then Approach 2 can be accepted
- Factors significantly affecting the Bike Sharing trend in Washington DC by comparing 2 models are: Month, Prin1 (Temperature, feels like temperature , Humidity and Windspeed)
- We can observe that there is increase in bike rentals between **January – June** for both the years
- It can be also observed that Weather situation and holiday also plays an important role in bike rentals usage
- The analysis of these factors will help the company in efficient demand forecasting of the bikes

THANK YOU