Data Analysis on Bike Sharing Dataset

SDA Project - Group No : 33

5th December 2020

Team Details :

M Sai Amulya	S20170010099
G Sreeja	S20170010047
P Deepika Sowmya	S20170010110
B Gayatri Shivani	S20170010029

Abstract

Aim of the project is to analyse the data statistically of the given bike sharing dataset and predict the number of bike users using the features from the dataset and understand the features that help in increasing the number of bike users.

Introduction & Background

Bike sharing systems are a new generation of traditional bike rentals where the whole process from membership, rental and return back has become automatic. Through these systems, users are able to easily rent a bike from a particular position and return back at another position. Currently, there are about over 500 bike-sharing programs around the world which is composed of over 500 thousands bicycles. Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues.

About the dataset

The core data set is related to the two-year historical log corresponding to years 2011 and 2012 from Capital Bikeshare system, Washington D.C., USA which is publicly available in http://capitalbikeshare.com/system-data. We aggregated the data on two hourly and daily basis and then extracted and added the corresponding weather and seasonal information. Weather information is extracted from http://www.freemeteo.com.

Files associated:

- hour.csv : bike sharing counts aggregated on hourly basis. Records: 17379 hours
- day.csv - bike sharing counts aggregated on a daily basis. Records: 731 days

Dataset Characteristics

- Both hour.csv and day.csv have the following fields, except hr which is not available in day.csv

1. **Instant :** record index
2. **dteday :** date
3. **season :** season (1:springer, 2:summer, 3:fall, 4:winter)
4. **yr :** year (0: 2011, 1:2012)
5. **mnth :** month (1 to 12)
6. **hr :** hour (0 to 23)
7. **holiday :** weather day is holiday or not
8. **weekday :** day of the week
9. **workingday :** if day is neither weekend nor holiday is 1, otherwise is 0.
10. **weathersit :**
    1 : Clear, Few clouds, Partly cloudy, Partly cloudy
    2 : Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
    3 : Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
    4 : Heavy Rain + Ice Pellets + Thunderstorm + Mist, Snow + Fog
11. **temp :** Normalized temperature in Celsius. The values are divided to 41 (max)
12. **atemp :** Normalized feeling temperature in Celsius. The values are divided to 50 (max)
13. **hum :** Normalized humidity. The values are divided to 100 (max)
14. **windspeed :** Normalized wind speed. The values are divided to 67 (max)
15. **casual :** count of casual users
16. **registered :** count of registered users
17. **cnt :** count of total rental bikes including both casual and registered

Methodology

- Load and understand Data
- Data preprocessing
- Null value analysis
- Data analysis
- Outlier Analysis
- Normalisation of target
- Feature Selection
    - Univariate selection
    - Chi square, t , anova test
    - Correlation Analysis
- Test of assumptions
- Model Selection
- Feature importance

Load and Understand Data

- The **hour data** has 17379 observations with 17 characteristics, **day data** has 731 observations corresponding to a particular day with 16 characteristics other than 'hr'.

The hour data looks as below:

	instant	dteday	season	yr	mnth	hr	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
0	1	2011-01-01	1	0	1	0	0	6	0	1	0.24	0.2879	0.81	0.0	3	13	16
1	2	2011-01-01	1	0	1	1	0	6	0	1	0.22	0.2727	0.80	0.0	8	32	40
2	3	2011-01-01	1	0	1	2	0	6	0	1	0.22	0.2727	0.80	0.0	5	27	32
3	4	2011-01-01	1	0	1	3	0	6	0	1	0.24	0.2879	0.75	0.0	3	10	13
4	5	2011-01-01	1	0	1	4	0	6	0	1	0.24	0.2879	0.75	0.0	0	1	1

- Our **target** characteristic is **cnt** which is the sum of casual and registered users
- We understand that there are both categorical and numerical characteristics
- **Categorical data** : 'season', 'yr', 'mnth', 'hr', 'holiday', 'weekday', 'workingday', 'weathersit
- **Numerical data** : 'temp', 'atemp', 'hum', 'windspeed'
- **Dteday** is a date object and **instant** is the index of the observations

Numerical Data Summary :

	temp	atemp	hum	windspeed
count	17379.000000	17379.000000	17379.000000	17379.000000
mean	0.496987	0.475775	0.627229	0.190098
std	0.192556	0.171850	0.192930	0.122340
min	0.020000	0.000000	0.000000	0.000000
25%	0.340000	0.333300	0.480000	0.104500
50%	0.500000	0.484800	0.630000	0.194000
75%	0.660000	0.621200	0.780000	0.253700
max	1.000000	1.000000	1.000000	0.850700

Categorical Data Summary :

	season	holiday	mnth	hr	weekday	workingday	weathersit	yr
count	17379	17379	17379	17379	17379	17379	17379	17379
unique	4	2	12	24	7	2	4	2
top	3	0	7	17	6	1	1	1
freq	4496	16879	1488	730	2512	11865	11413	8734

Data Preprocessing

- In **dteday** we already have month and year in the dataset, so we have kept only the date info and included it in the categorial characteristics.
- Since we are only interested in the total count of the users, we have **dropped casual and registered** users count

- We also dropped the **instant** which is the index.
- After dropping the above mentioned characteristics we are left with **14 characteristics.**

After the changes, the data looks as follows:

	dteday	season	yr	mnth	hr	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	cnt
0	1	1	0	1	0	0	6	0	1	0.24	0.2879	0.81	0.0	16
1	1	1	0	1	1	0	6	0	1	0.22	0.2727	0.80	0.0	40
2	1	1	0	1	2	0	6	0	1	0.22	0.2727	0.80	0.0	32
3	1	1	0	1	3	0	6	0	1	0.24	0.2879	0.75	0.0	13
4	1	1	0	1	4	0	6	0	1	0.24	0.2879	0.75	0.0	1

Null Value Analysis

- There are **no null values or missing values** in the data

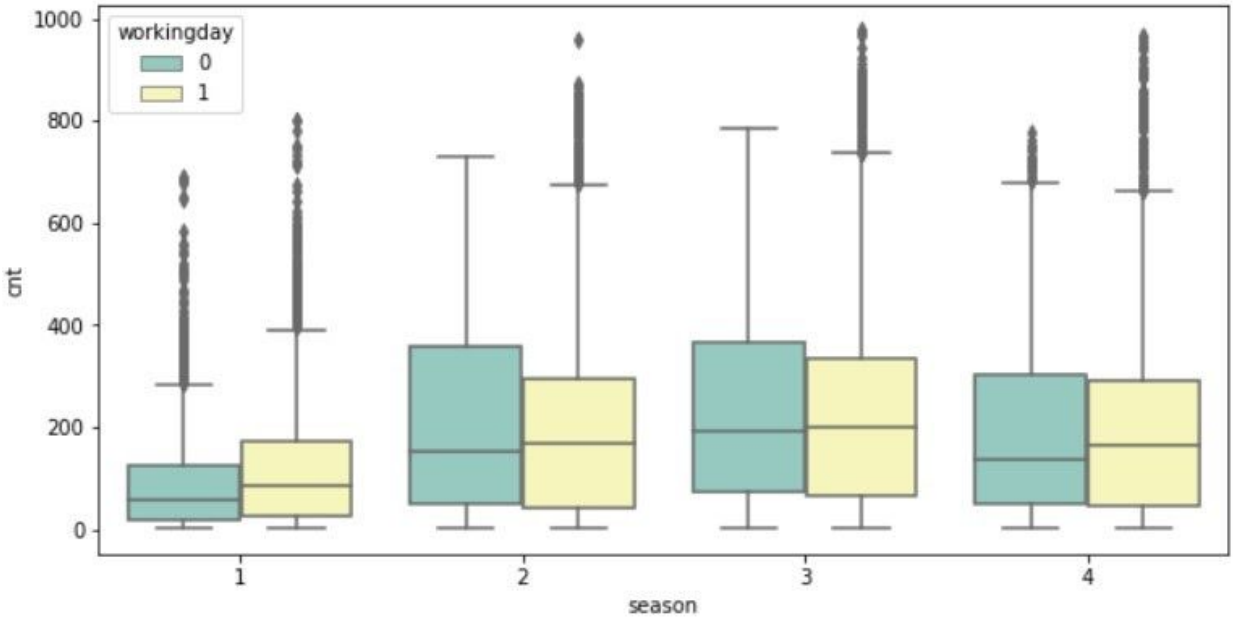
```
RangeIndex: 17379 entries, 0 to 17378
Data columns (total 14 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   dteday      17379 non-null  int64
 1   season      17379 non-null  int64
 2   yr          17379 non-null  int64
 3   mnth        17379 non-null  int64
 4   hr          17379 non-null  int64
 5   holiday     17379 non-null  int64
 6   weekday     17379 non-null  int64
 7   workingday  17379 non-null  int64
 8   weathersit  17379 non-null  int64
 9   temp        17379 non-null  float64
 10  atemp       17379 non-null  float64
 11  hum         17379 non-null  float64
 12  windspeed   17379 non-null  float64
 13  cnt         17379 non-null  int64
dtypes: float64(4), int64(10)
```

```
hour_data.isnull().sum()
```

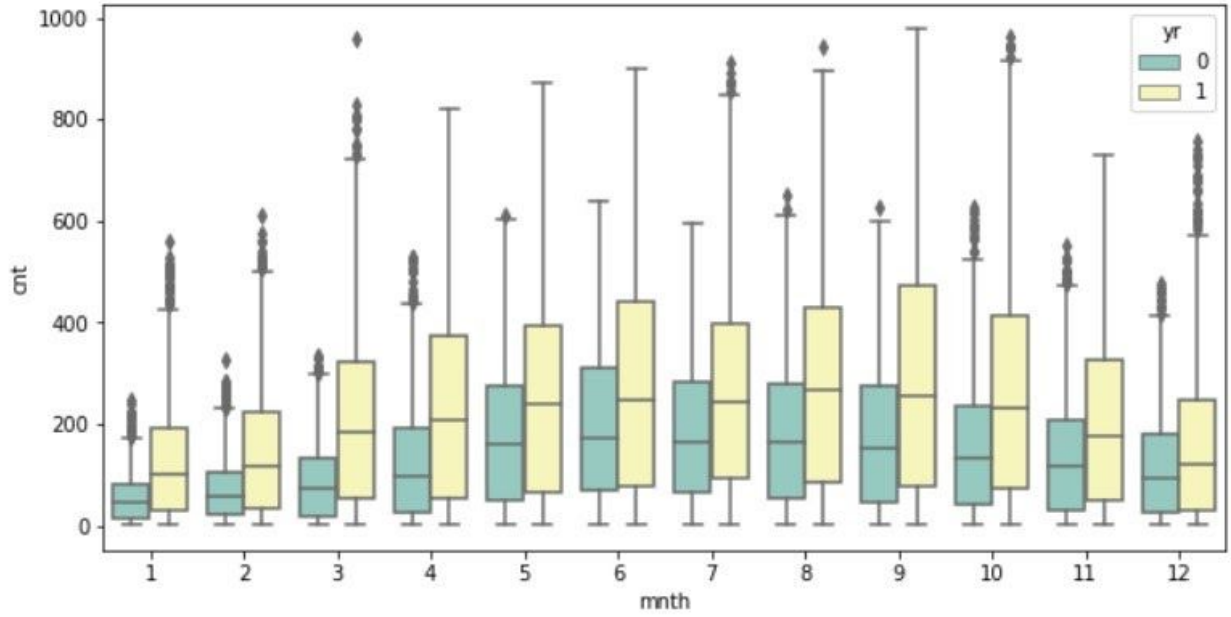
```
dteday        0
season        0
yr            0
mnth          0
hr            0
holiday       0
weekday       0
workingday    0
weathersit    0
temp          0
atemp         0
hum           0
windspeed     0
cnt           0
dtype: int64
```

Data Analysis

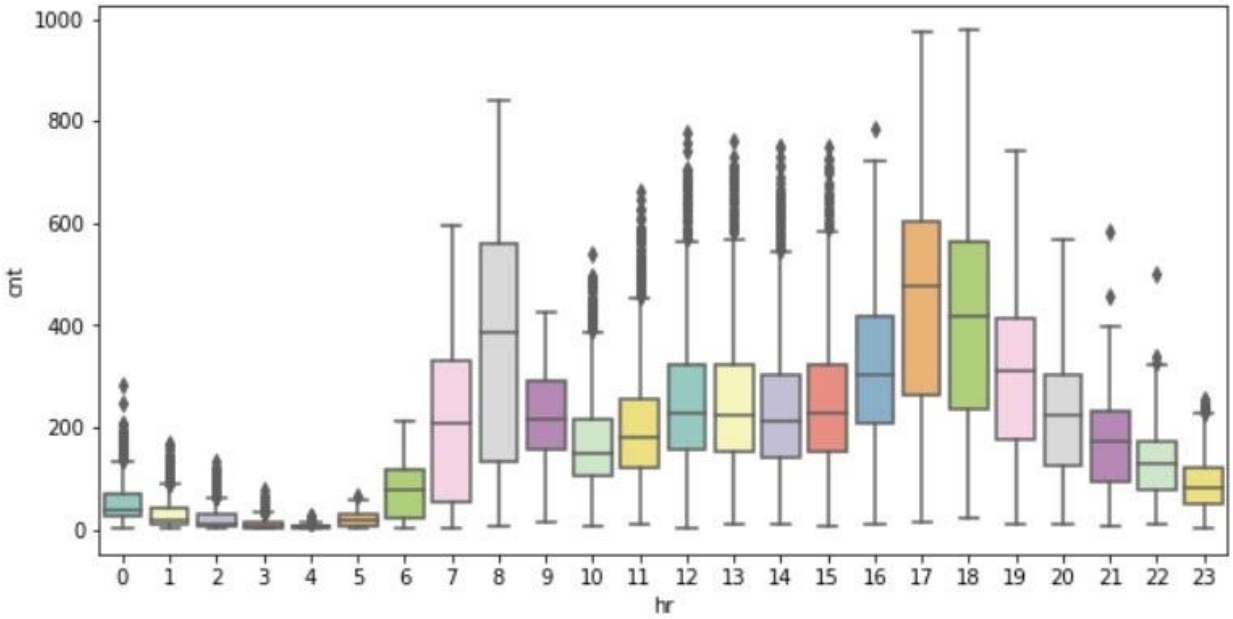
Season 1 has the least number of users and users come out for working days. In other seasons users prefer to travel during non working days.



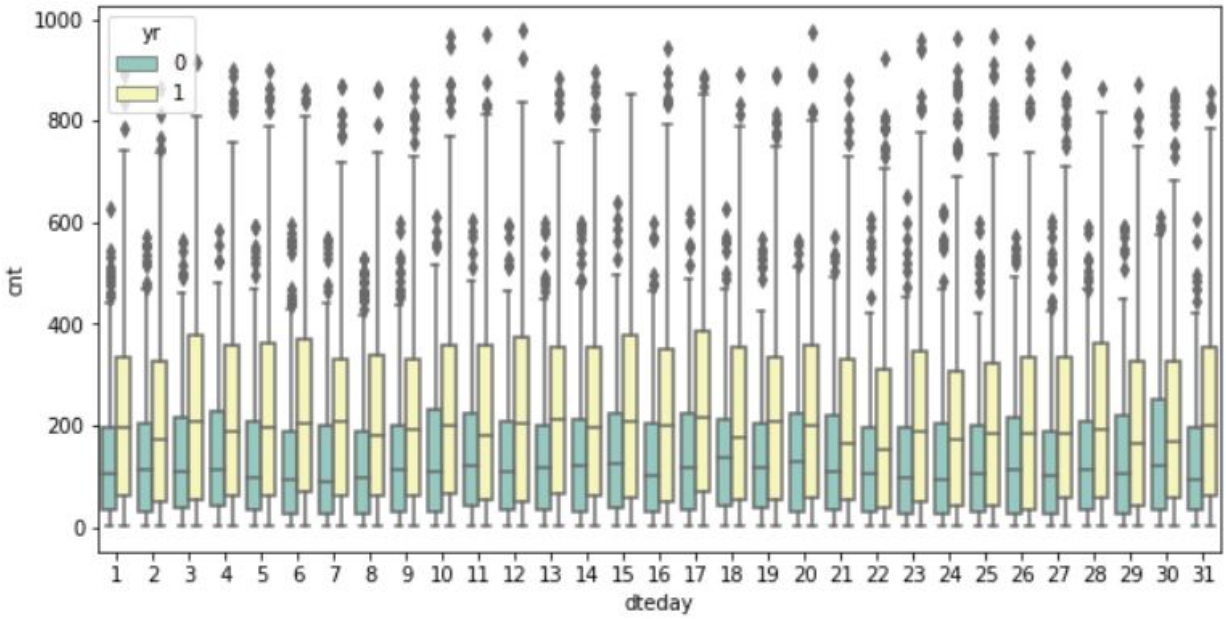
The number of users increased the next year whereas the interest to come out during clearer months and climate hasn't changed



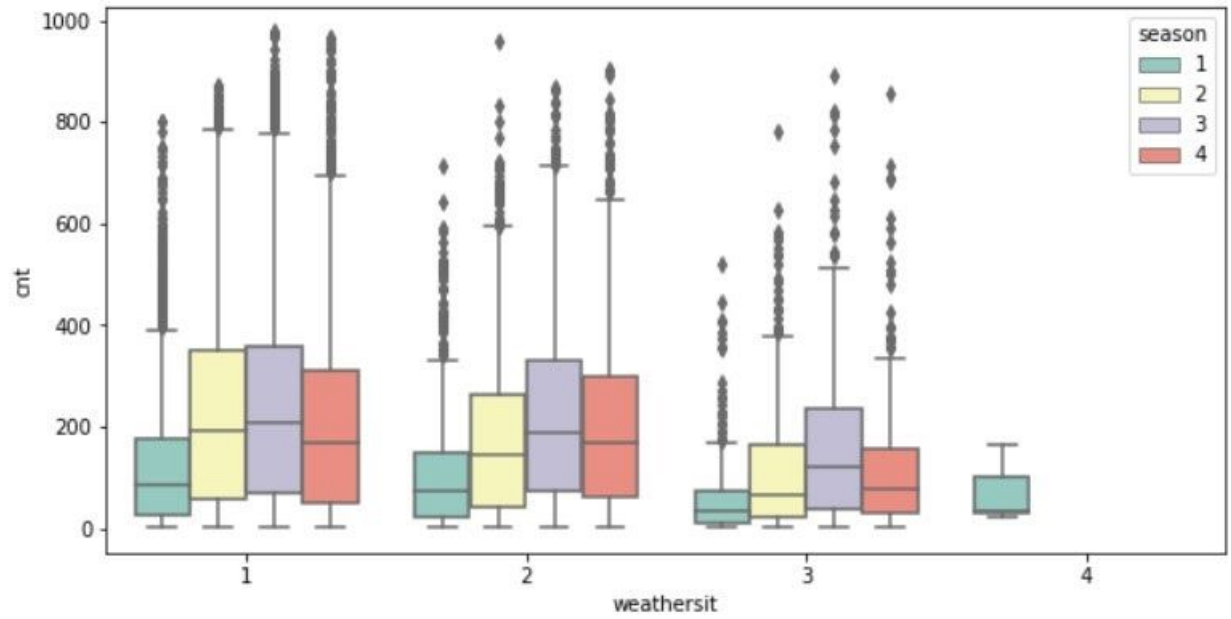
There are more users at 8am and 5pm i.e most users of the bicycle rental service use the bikes to get to work or school.
There are many outliers between 8am and 5pm indicating weekdays or holidays



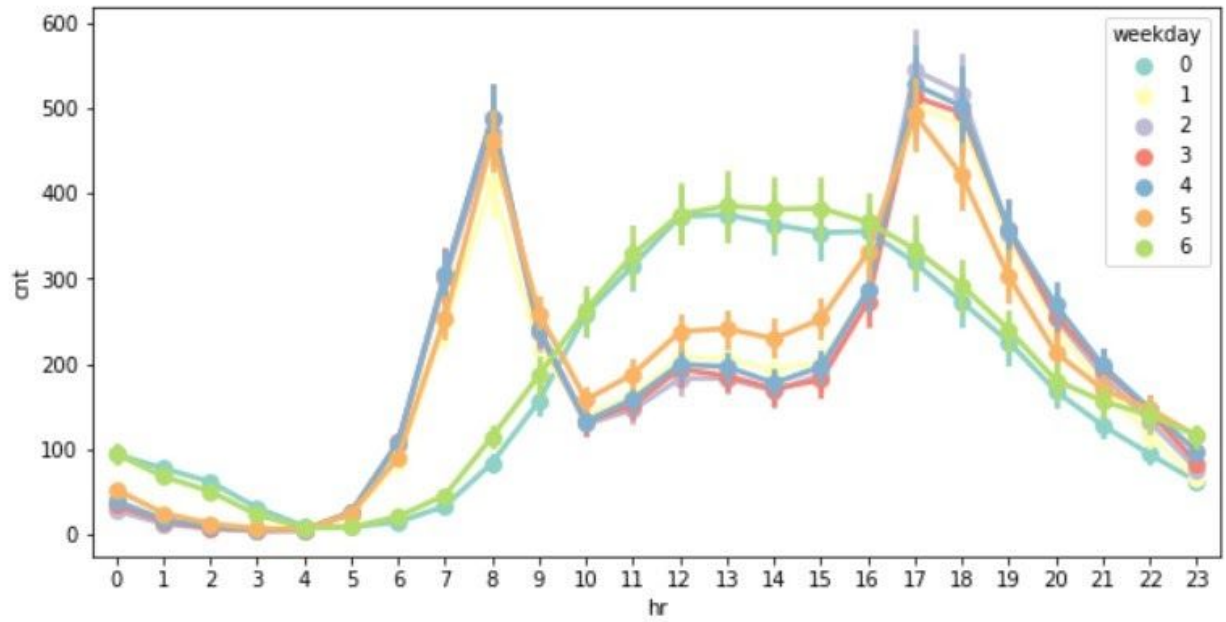
There isn't much difference with the date of the month and it doesn't effect the number of bike users but we can see a significant increase in count over the year



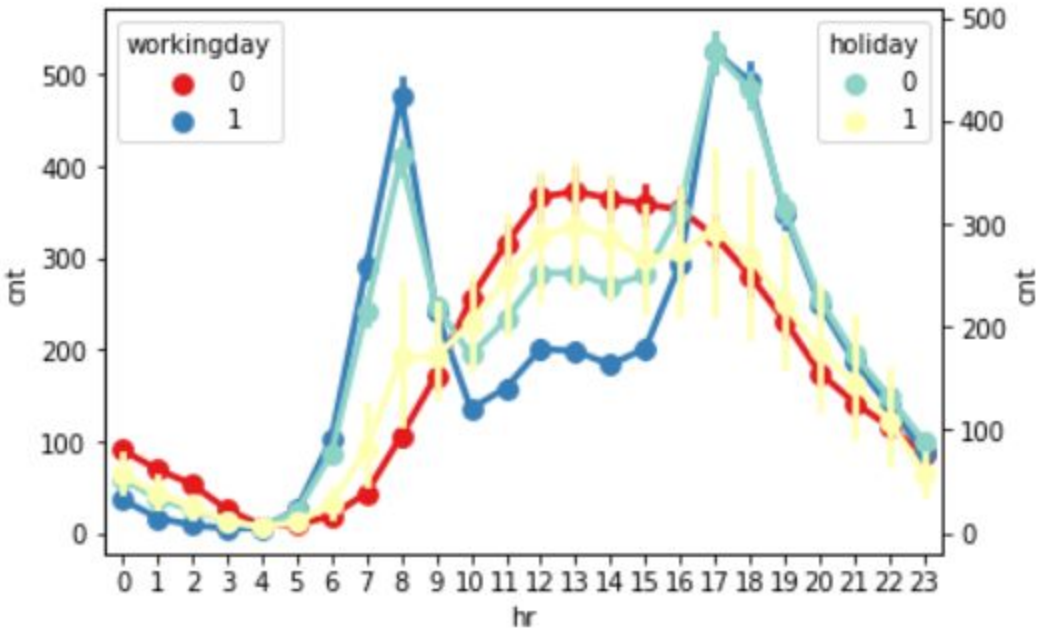
Only season 1 has weathersit of type 4 i.e Heavy Rain + Ice Pellets + Thunderstorm + Mist, Snow + Fog. It can also be inferred that more number of users are interested to use bikes at clearer weather situations and seasons 1 and 2



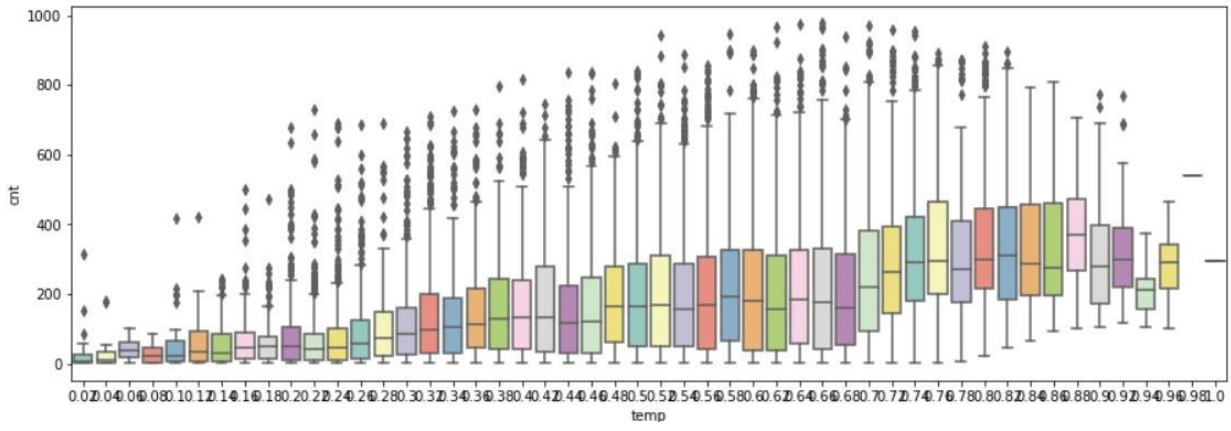
During week day i.e 0 and 6 the peak is observed around 2pm



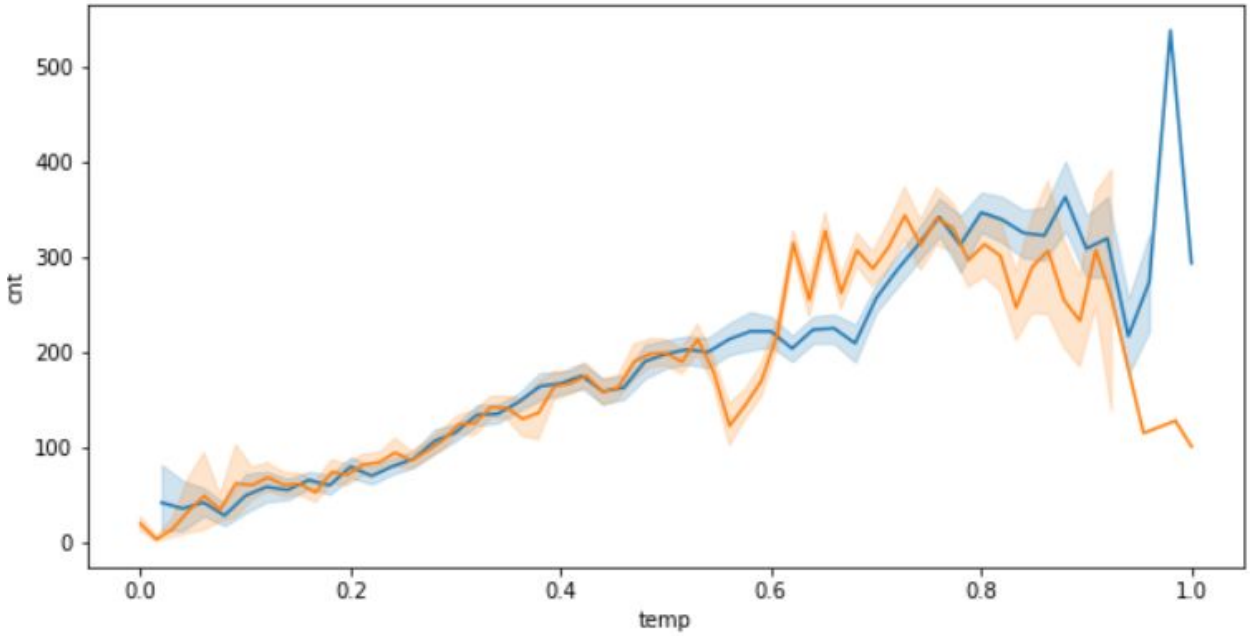
For working day and ~holiday the cnt vs hour follow a similar pattern which is reasonable since non working days are holidays. The pattern is also similar to the week days.



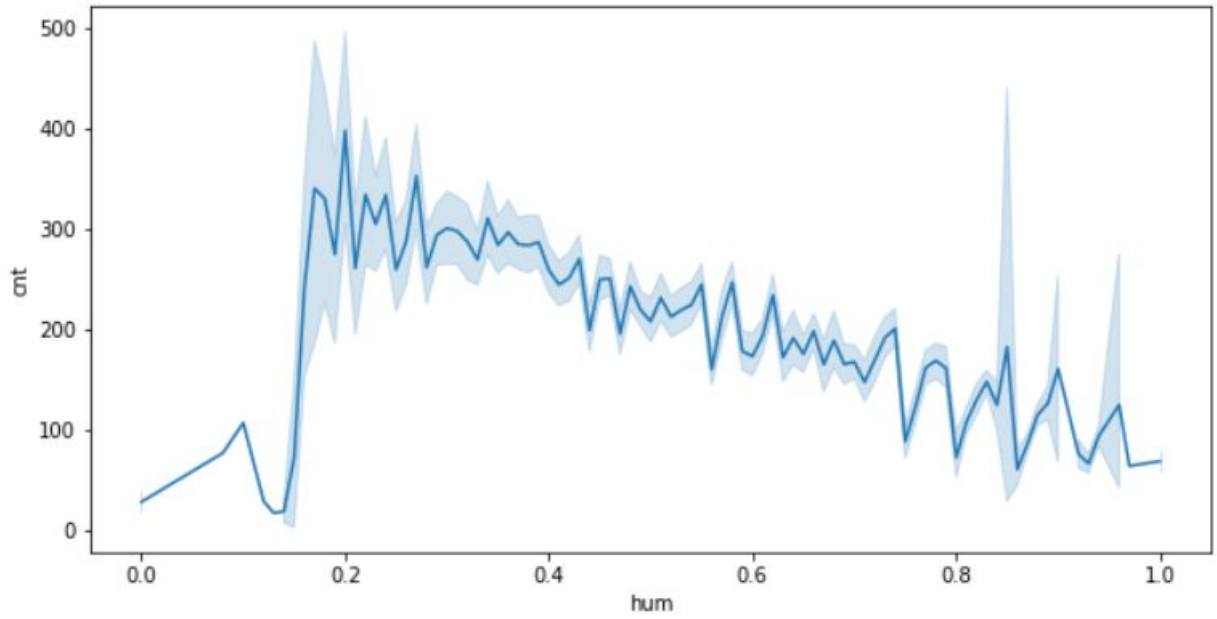
With the increase in temperature i.e as the climate gets warmer , the use of bikes has increased indicating the preference in warmer atmosphere.



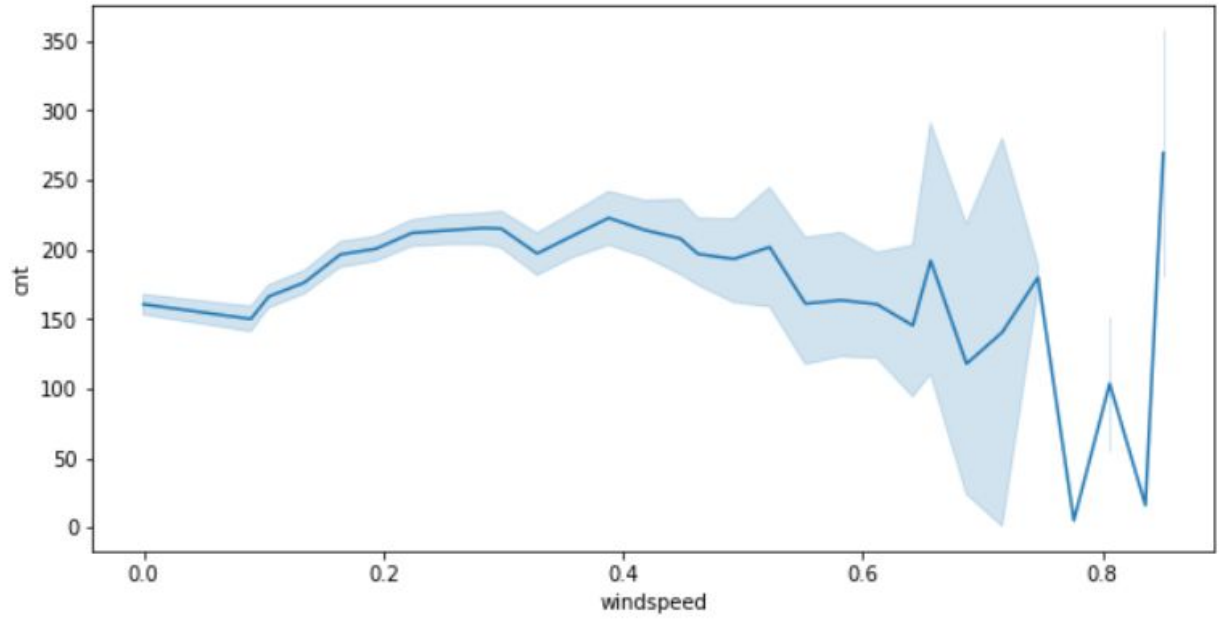
There is a difference between feeling temperature(atemp in orange color) and the actual temperature (temp in blue color) from 0.5 values.



We can see the variation in the count of bike users with humidity, the count decreases as humidity increases.

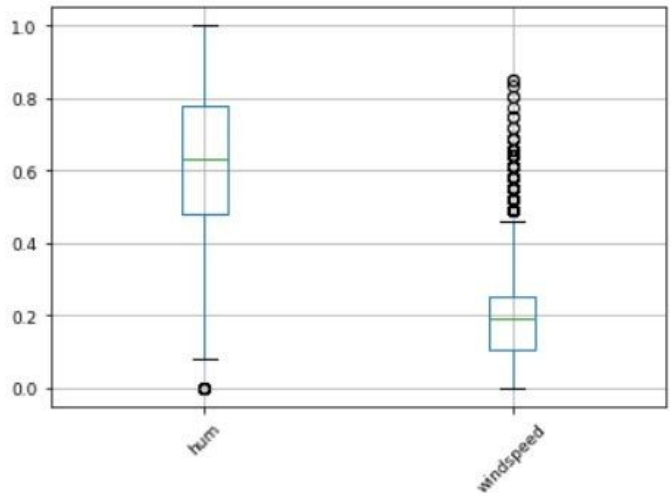


Windseed doesn't look like it has a greater effect on count because till 0.8 the graph is almost linear.

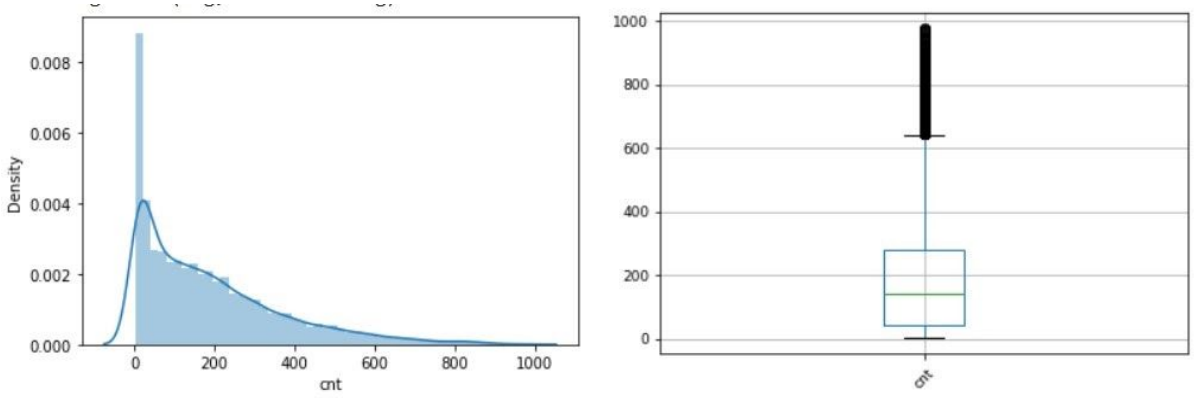


Outlier Analysis

By doing the box plot we have identified the outliers in hum and windspeed and can be seen in the below plot



We also identified the outliers in the target ‘cnt’ characteristics and can be observed in the below plots

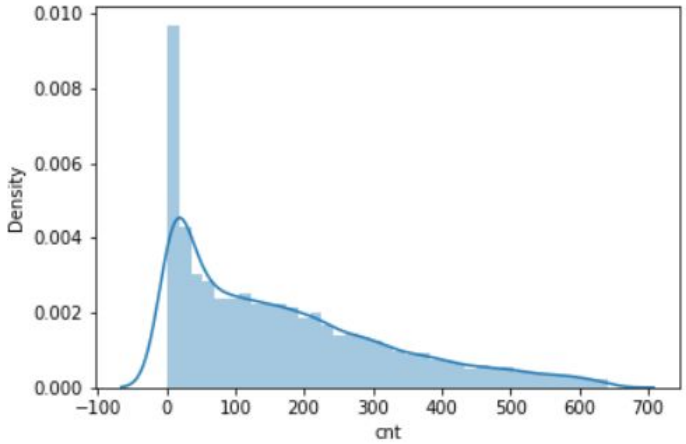


Removal of outliers

- The outliers are removed by finding IQR Range and have removed the points that doesn't fall under the IQR Range

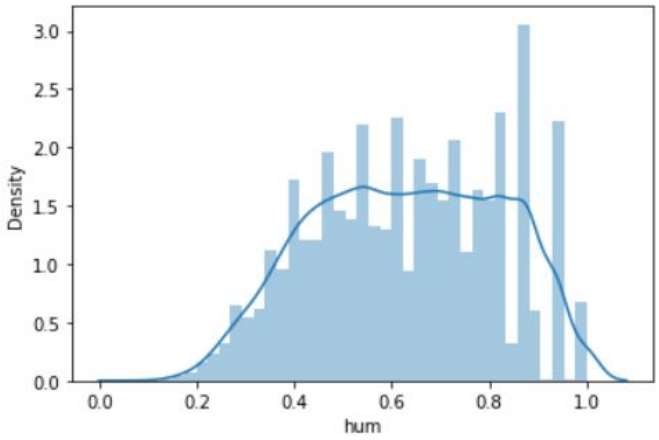
After removing outlier data from **cnt**

- Samples in the data with outliers : 17379
- Samples in the data without outliers for cnt : 16874



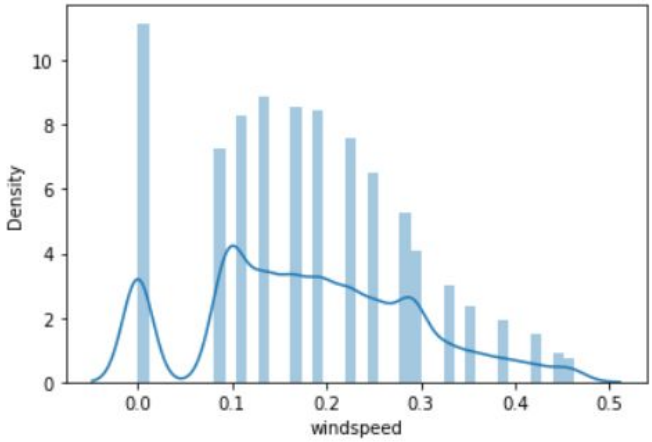
After removing outlier data from **hum**

- Samples in the data with outliers : 16874
- Samples in the data without outliers for cnt : 16852

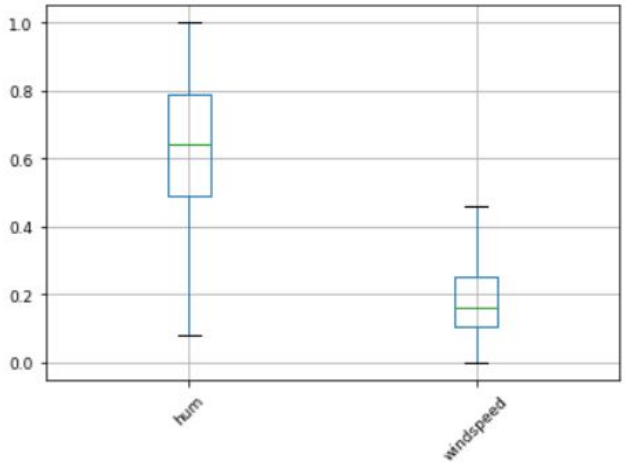


After removing outlier data from **windspeed**

- Samples in the data with outliers : 16852
- Samples in the data without outliers for cnt : 16822



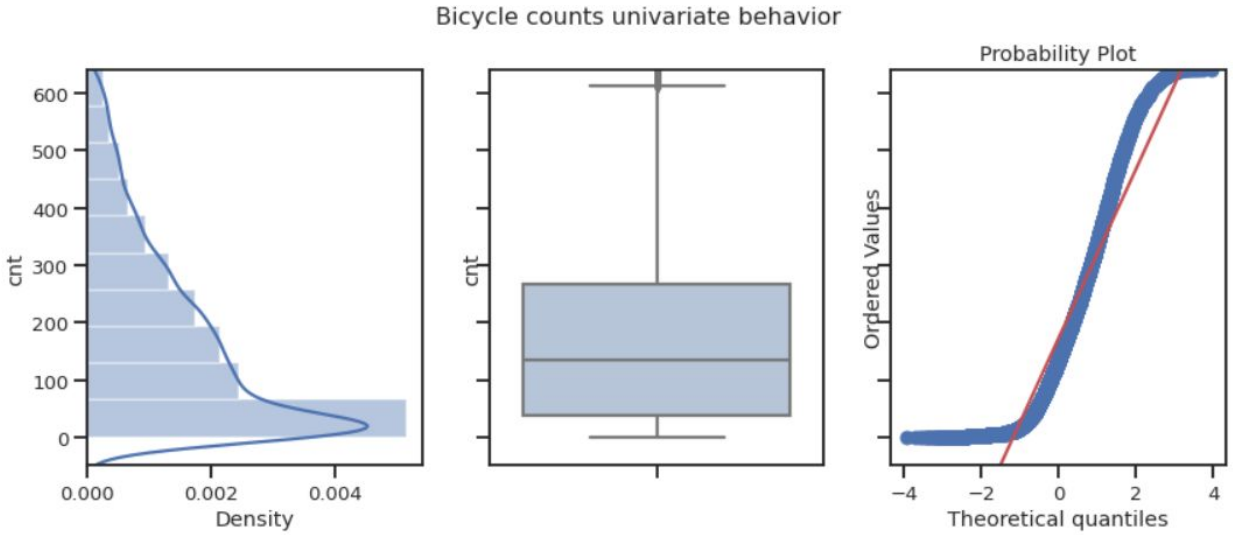
The below box plot represents data of hum and windspeed without outliers



Normalisation of target

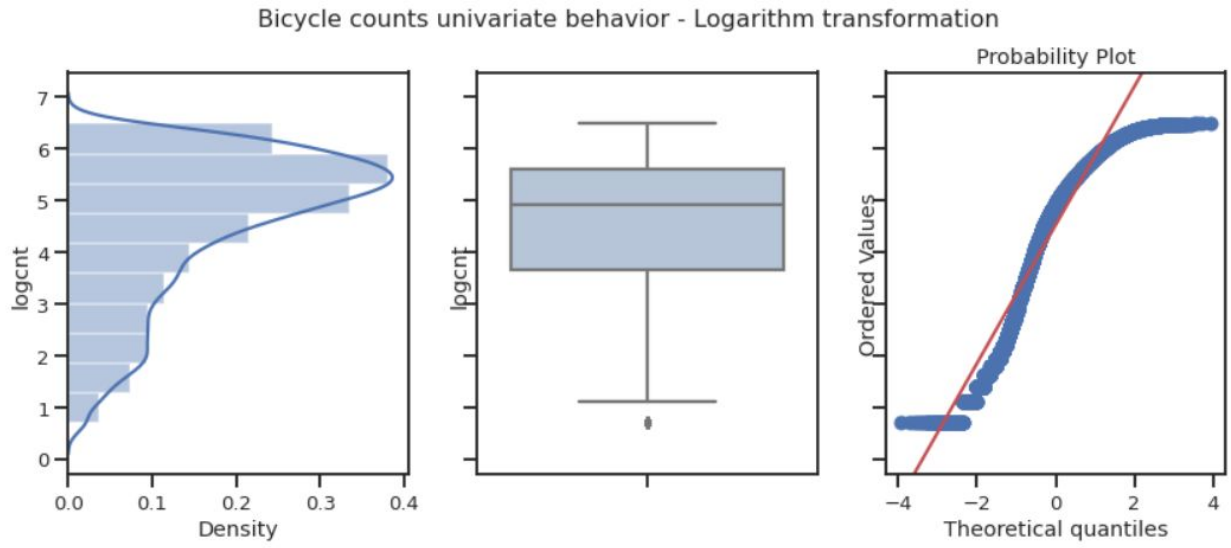
Before normalization of target , the plots of **cnt**

- Clearly Q-Q plot is not linear and the distplot is +vely skewed.



After normalisation, the plots of transformed cnt i.e log (1+cnt)

- We can observe that the +ve skewness of the graph is reduced.



Feature Selection

Including more features in the model makes the model more complex, and the model may be overfitting the data. Some features can be the noise and potentially damage the model. By removing those unimportant features, the model may generalize better.

Univariate Selection

- Statistical tests can be used to select those features that have the strongest relationship with the output variable.
- The scikit-learn library provides the SelectKBest class that can be used with a suite of different statistical tests to select a specific number of features.
- The scikit-learn machine library provides an implementation of the ANOVA f-test in the f_classif() function.
- The ANOVA test assesses whether the averages of more than two groups are statistically different from each other. This analysis is appropriate for comparing the averages of a numerical variable for more than two categories of a categorical variable
- The values of the scores for each variable is given below

	Specs	Score
4	hr	16.038324
1	season	2.352069
2	yr	2.135567
3	mnth	1.688107
7	workingday	1.591190
8	weathersit	1.300866
6	weekday	1.124277
5	holiday	0.975642
0	dteday	0.955519

After univariate selection we have dropped weekday, dteday and holiday

Chi- Square Test

The Chi-Squared test is a statistical hypothesis test that assumes (the null hypothesis) that the observed frequencies for a categorical variable match the expected frequencies for the categorical variable

Chi-Square test on weekday and workingday :

- p value is 0.0
- Dependent (reject H0)

Chi-Square test on holiday and workingday

- p value is 4.117973248471021e-233
- Dependent (reject H0)

workingday alone is enough and this justifies the dropping of holiday and weekday

T- Test

They assess whether the averages of two groups are statistically different from each other. This analysis is appropriate for comparing the averages of a numerical variable

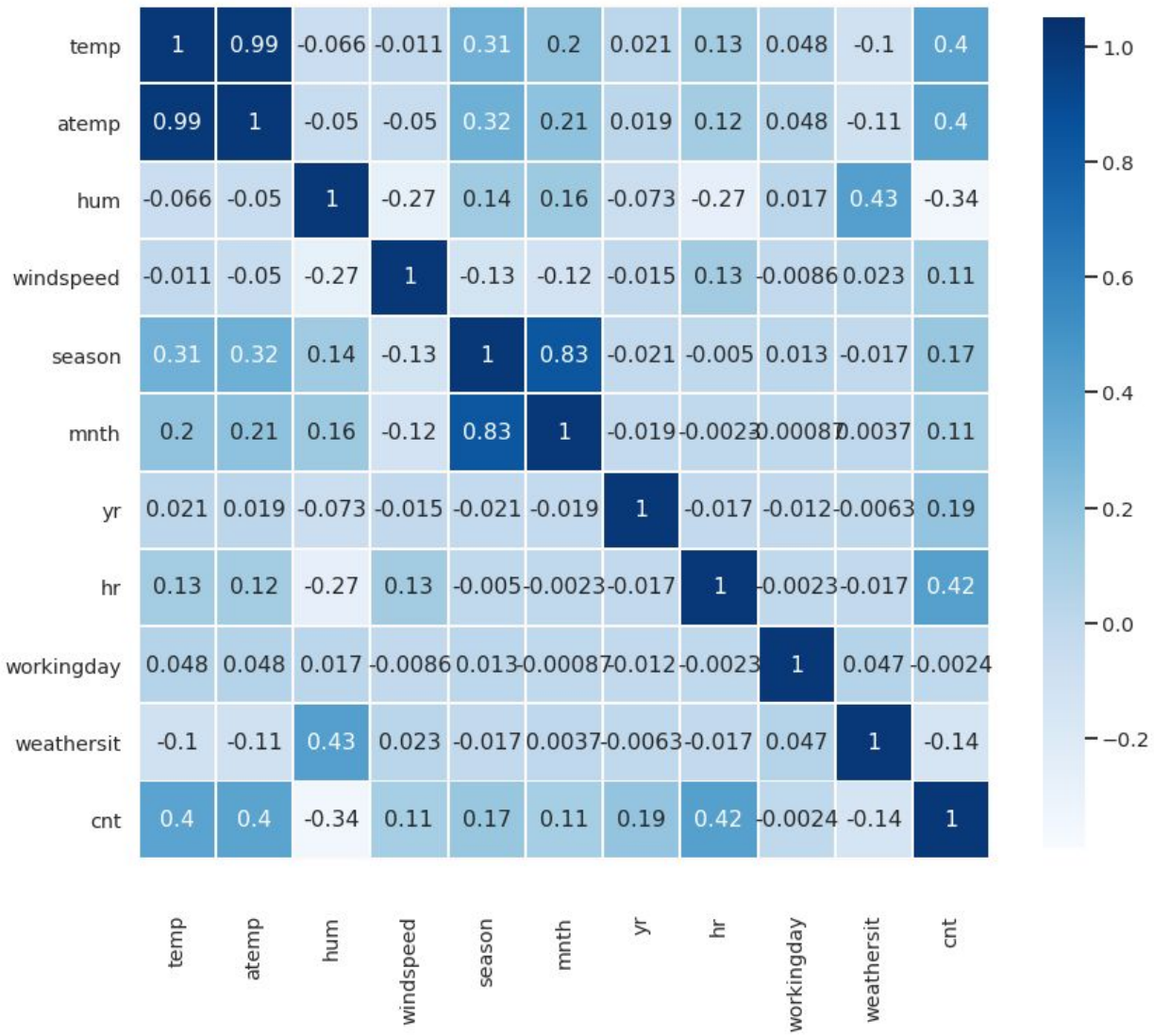
T-test on atemp and temp

- Null hypothesis : mean of the two variables is same
- p-value: 0.0
- t_value: 75.78292177802633
- we are rejecting null hypothesis
- Thus there is a variation in atemp and temp

Correlation Analysis

Correlation analysis is a statistical method used to evaluate the strength of relationship between two quantitative variables

Below is the Correlation heatmap for our features and target



Observations from heatmap

- There is a high correlation with hr, temp, atemp for target 'cnt'
- There is a negative correlation with weathersit, working day and humidity for target 'cnt'
- Mnth and windspeed are least correlated
- Temp and atemp are highly correlated with each other
- Similarly month and season are highly correlated with each other

Thus we have dropped mnth, windspeed and atemp from our feature variables

The final features list contains the following after feature selection :

```
'temp', 'hum', 'season', 'hr', 'workingday', 'weathersit','yr'
```

Test of Assumptions :

OLS :

OLS is a commonly used regression method and simple method to understand the relationship between dependent and independent attributes. Though this is a simple method which makes certain assumptions, yet its most used method to understand the effect of independent attributes on dependent. We can understand the assumptions that are made for a linear regression method from OLS.

Assumptions of linear regression :

1. Linearity: there is a linear relationship between our features and responses. This is required for our estimator and predictions to be unbiased.
2. No multicollinearity: features are not correlated. If this is not satisfied, our estimator will suffer from high variance.
3. Gaussian (Normal Distributed) errors: our errors are Gaussian distributed with mean 0.
4. Homoscedasticity: errors have equal variance. If this is not satisfied, there will be other linear estimators with lower variance.
5. No or little autocorrelation.

Summary from OLS:

OLS Regression Results

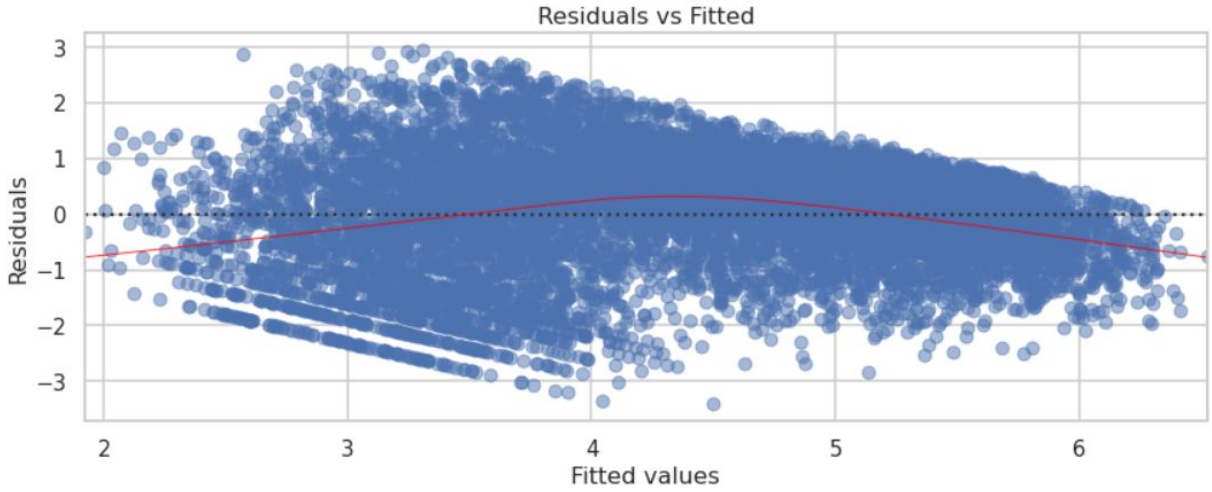
Dep. Variable:	y	**R-squared:**	0.469
Model:	OLS	**Adj. R-squared:**	0.469
Method:	Least Squares	**F-statistic:**	1250.
Date:	Sun, 06 Dec 2020	**Prob (F-statistic):**	0.00
Time:	06:29:07	**Log-Likelihood:**	-14120.
No. Observations:	9913	**AIC:**	2.826e+04
Df Residuals:	9905	**BIC:**	2.831e+04
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.6937	0.054	49.794	0.000	2.588	2.800
x1	2.0806	0.058	35.745	0.000	1.966	2.195
x2	-1.4353	0.062	-23.037	0.000	-1.557	-1.313
x3	0.1611	0.011	14.968	0.000	0.140	0.182
x4	0.0960	0.002	62.771	0.000	0.093	0.099
x5	-0.0105	0.022	-0.482	0.630	-0.053	0.032
x6	0.0318	0.018	1.801	0.072	-0.003	0.066
x7	0.5261	0.033	16.037	0.000	0.462	0.590

Omnibus:	100.258	**Durbin-Watson:**	0.531
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	102.695
Skew:	-0.245	**Prob(JB):**	5.01e-23
Kurtosis:	2.905	**Cond. No.**	103.

1. Linearity

The first assumption we check is linearity. We can visually check this by fitting ordinary least squares (OLS) and use that model for predicting. We then plot the residuals vs predictions. The thumb rule to look at this plot is there should not be any patterns and this plot should appear like a random plot for linear assumption to be true.



- From the graph , since we have a curved pattern we can say that the linearity doesn't hold here.

2. Multicollinearity

This can be verified with Variation Inflation Factor (VIF).

VIF Test :

Variance Inflation Factor (VIF) is used to detect the presence of multicollinearity. It measures how much the variance of the estimated regression coefficients are inflated as compared to when the predictor variables are not linearly related. It is obtained by regressing each independent variable, say X on the remaining independent variables (say Y and Z) and checking how much of it (of X) is explained by these variables.

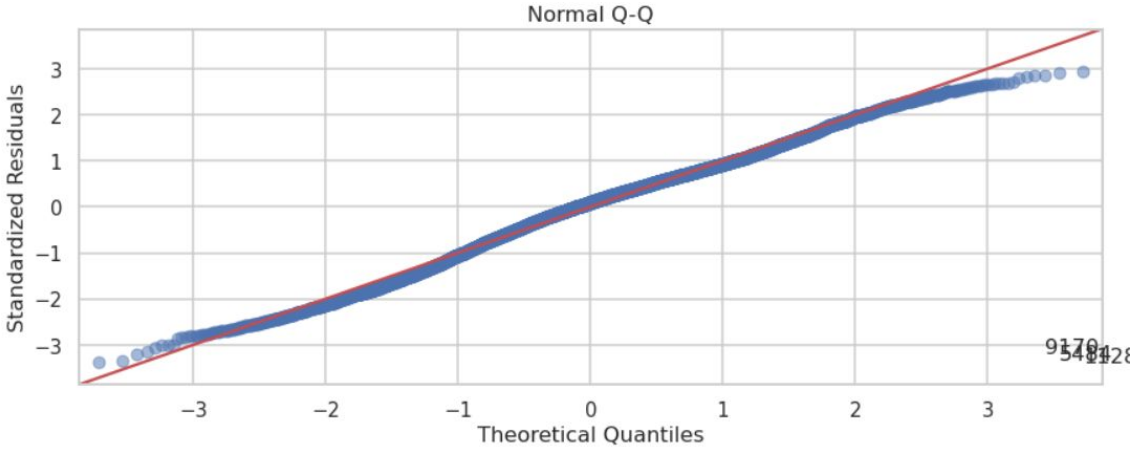
$$VIF = \frac{1}{(1-R^2)}$$

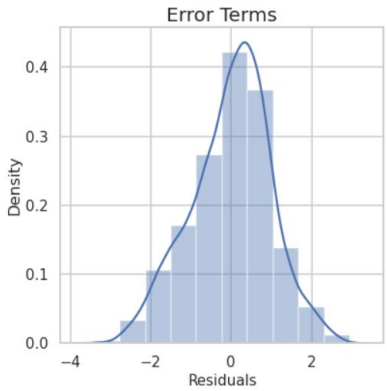
- With VIF > 5 there is an indication that multicollinearity may be present; with VIF > 10 there is certainly multicollinearity among the variables.
- From the table we can see that humidity has high VIF value > 10 indicating certain presence of multicollinearity indicating a certain multicollinearity among the variables.

	feature	VIF
0	temp	7.507405
1	hum	11.495971
2	season	6.741054
3	hr	3.390411
4	workingday	2.966929
5	weathersit	7.408545
6	yr	1.864638

3 .Normality

The error terms should be normally distributed. This can be verified using the QQ plot of residual terms. The errors should follow the red line for this assumption to hold true.

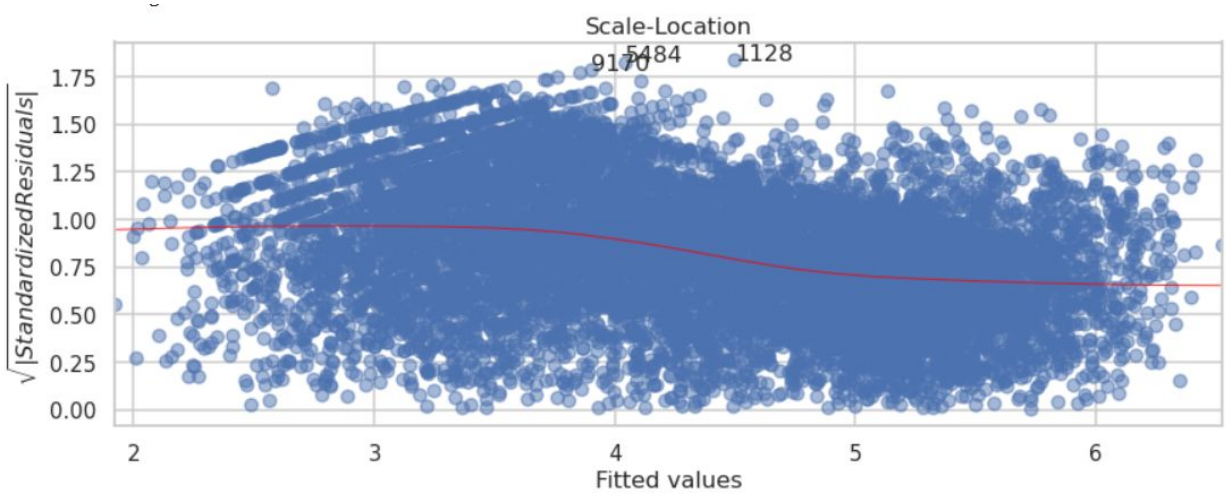




- The residuals follow normality and its mean is approximately equal to 0 and the plot is slightly -ve skewed.
- The qq plot of the residuals is a linear indicating the normality of errors.

4.Homoscedasticity.

The error terms variance should be constant. This can be verified by plotting standardised residuals against fitted values.



The Goldfeld-Quandt Test is used to test for heteroscedasticity. The test splits the data into two groups and tests to see if the variances of the residuals are similar across the groups.

The Null hypothesis is that the variance in the two sub-samples are the same i.e there is no heteroscedasticity.
If p < 0.05 : we reject H0

Values from test :

- [('F statistic', 1.1288972553295835), ('p-value', 1.0889571099175361e-05)]
- Here p value is less than 0.05 thus we are rejecting the null hypothesis and its heteroscedasticity.

5. Durbin Watson Test :

Durbin-Watson test. Durbin-Watson's d tests the null hypothesis that the residuals are not linearly auto-correlated. While d can assume values between 0 and 4, values around 2

indicate no autocorrelation. As a rule of thumb values of 1.5 < d < 2.5 show that there is no auto-correlation in the data.

- The Durbin Watson value is 0.531 indicating a positive autocorrelation i.e the residuals are not independent from each other.

Conclusion:

- Thus Linear regression can't be a good fit for our problem.

Model Selection:

We have considered 4 models for our problem and are as follows :

Decision Tree Model:

It is a type of supervised learning algorithm. In this algorithm, we split the population into two or more homogeneous sets. This is done based on most significant attributes/ independent variables to make as distinct groups as possible.

Random Forest Model:

Random Forest is an ensemble of decision trees. In Random Forest, we have a collection of decision trees (so known as "Forest"). To classify a new object based on attributes, each tree gives a classification and we say the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest).

SVM Model:

A support vector machine (SVM) is a supervised learning algorithm that analyzes data for classification and regression analysis. SVM is a supervised learning method that looks at data and sorts it into one of two categories. An SVM outputs a map of the sorted data with the margins between the two as far apart as possible.

NU-SVM Model:

Nu-SVM has the advantage of using a parameter nu(bounded by 0 and 1) for controlling the number of support vectors. The parameter nu represents the lower and upper bound on the number of examples that are support vectors and that lie on the wrong side of the hyperplane, respectively.

Overview Metrics Used:

Mean Square Error :

$$MSE(X,Y) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2}$$

Root Mean Square Log Error:

$$\text{RMSLE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log(x_i) - \log(y_i))^2}$$

R Square Error:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- The MSE, R^2 errors for training, validation sets for the models are shown in the below table.

Model	Mean Squared Error	R^2 score
DecisionTreeRegressor	0.38	0.80
SVR	1.24	0.35
NuSVR	0.12	0.94
RandomForestRegressor	0.24	0.87

From the above table it is clear that the the NuSVR has the least MSE and high R^2 and we have selected this model for testing

The MSE, RMSLE, R^2 errors for training, validation and test datasets are mentioned in the below table.

Model	Dataset	MSE	RMSLE	R^2 score
NuSVR	training	0.12	0.09	0.94
NuSVR	validation	0.12	0.08	0.94
NuSVR	test	0.24	0.10	0.87

Feature importance :

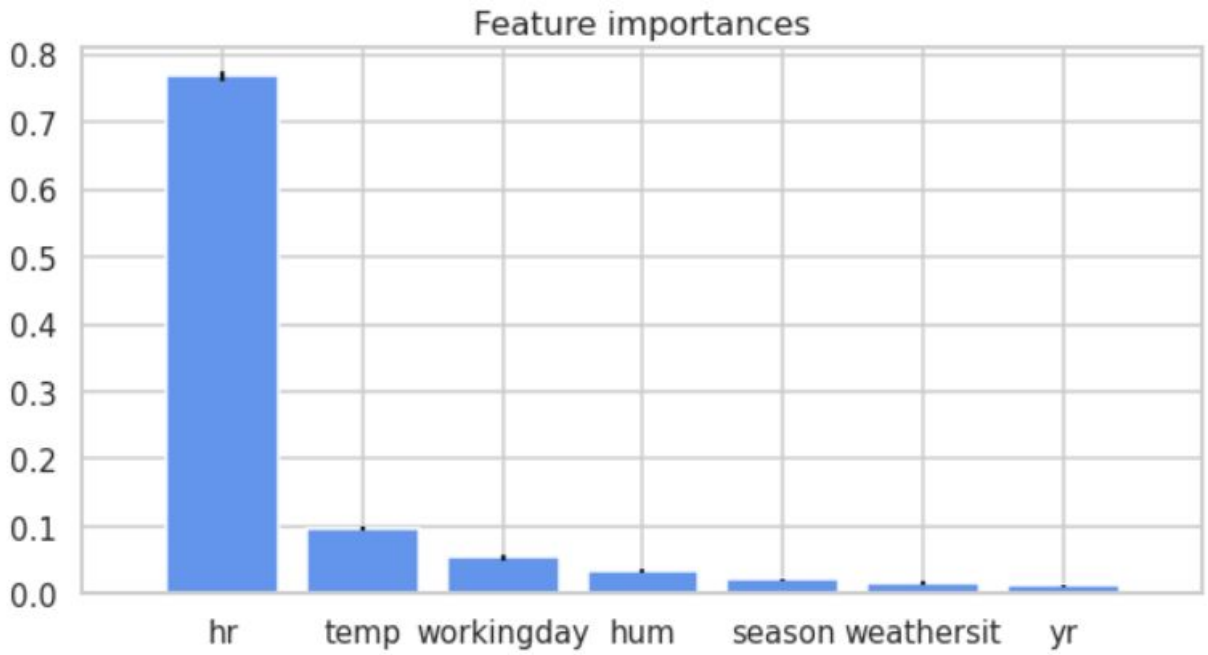
- For understanding the feature importance we have considered the random forest regressor which provides a feature ranking function.
- The MSE, RMSLE, R^2 errors for training, validation and test datasets are mentioned in the below table.

Model	Dataset	MSE	RMSLE	R^2 score
RandomForestRegressor	training	0.02	0.04	0.99
RandomForestRegressor	validation	0.24	0.10	0.87
RandomForestRegressor	test	0.30	0.11	0.84

- The feature ranking for the features we considered in the training set is mentioned below.

```
Feature ranking:
1. feature hr (0.767785)
2. feature temp (0.095660)
3. feature workingday (0.054043)
4. feature hum (0.034058)
5. feature season (0.020429)
6. feature weathersit (0.015838)
7. feature yr (0.012187)
```

- The feature importance plot for the features we considered in the training set is given below.



Conclusion:

- From the above data analysis we have a good understanding of the dataset and the features.
- From the above feature importance it is evident that the hr and temp play a good role to increase the number of bike users.
- It is also significant that the bike users have increased compared to the previous year.
- Peak hours of the day, warmer temperatures and clearer weather situations makes the users use the bikes
- Of the models used NuSVM is the good predictor.