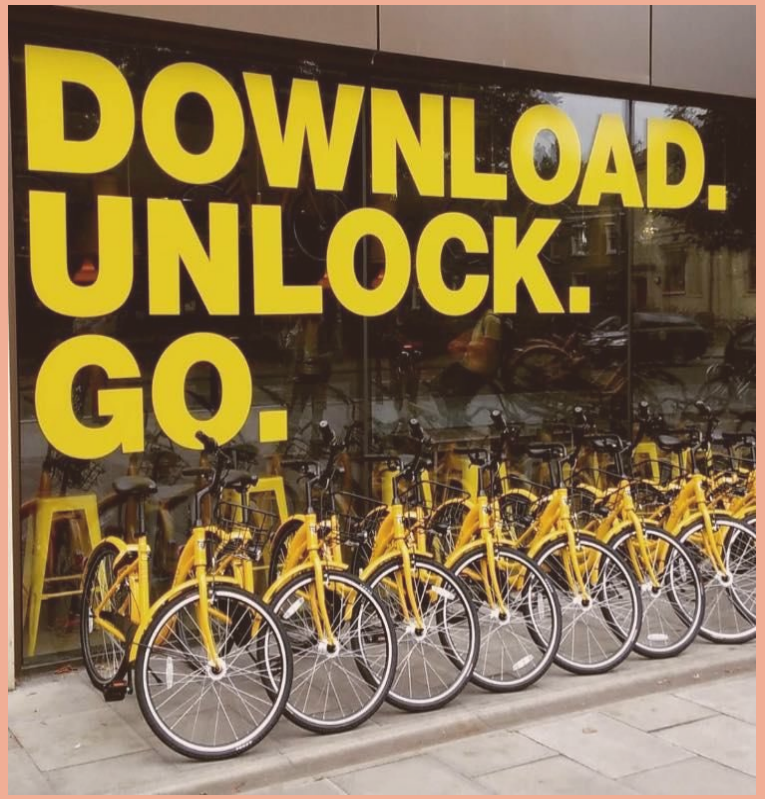
Data Analysis on Bike Sharing Dataset

Group No : 33



Contents

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

- Bike Sharing Systems are a new generation of traditional bike sharing rentals that has become automatic.
- Currently there are about 500 bike sharing programs.
- There is a great interest in these systems due to the important role in traffic, environmental and health issues.

Dataset

- The core dataset is related to the two-year historical log corresponding to years 2011 and 2012 from Capital Bikeshare system, Washington D.C., USA.

- The data is aggregated on two hourly and daily basis and then extracted and added the corresponding weather and seasonal information

Dataset Characteristics

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 weekday nor holiday is 1 otherwise is 0.

Dataset Characteristics

10 **weathersit :**

1) Clear, Few clouds, Partly cloudy
2) Mist + Cloudy, 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)

Dataset Characteristics

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 **windespeed**: 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

Dataset Description

- 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’.

	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

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

	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

Dataset Preprocessing

- dteday is changed to date removing month and year
- 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.

	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

Categorical Variables

- Season
- Yr
- Mnth
- Holiday
- Weekday
- Workingday
- Weathersit
- dteday

Numerical Variables

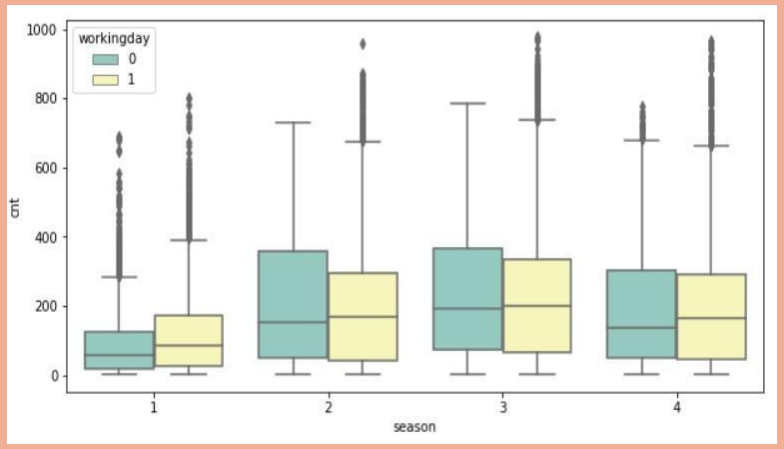
- Temp
- Atemp
- Hum
- Windspeed

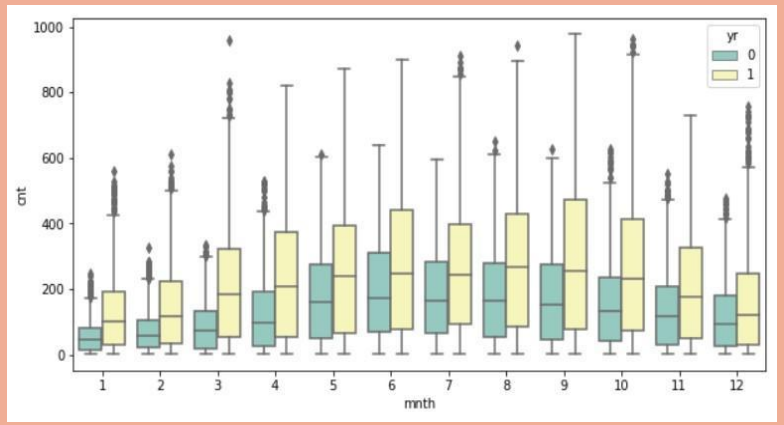
Target Variable : cnt

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)
```





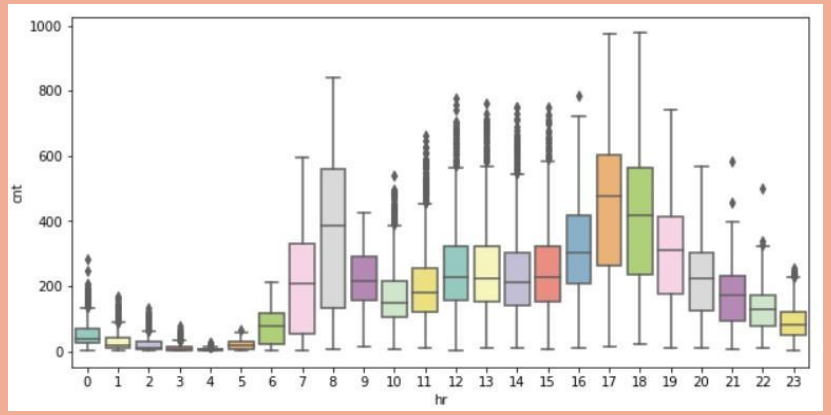
Data Analysis

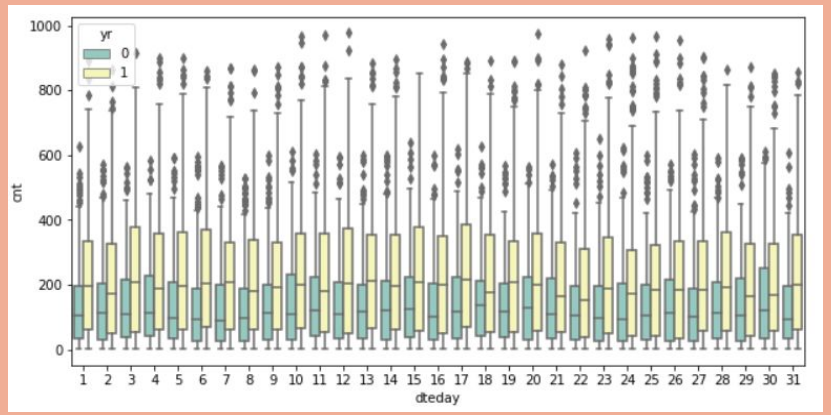
- Season 1 has least number of users and users come out for working days. In order seasons users prefer to travel during no working days.
- The number of users increased the next year whereas the interest to come out during clearer months and climate hasn't changed.

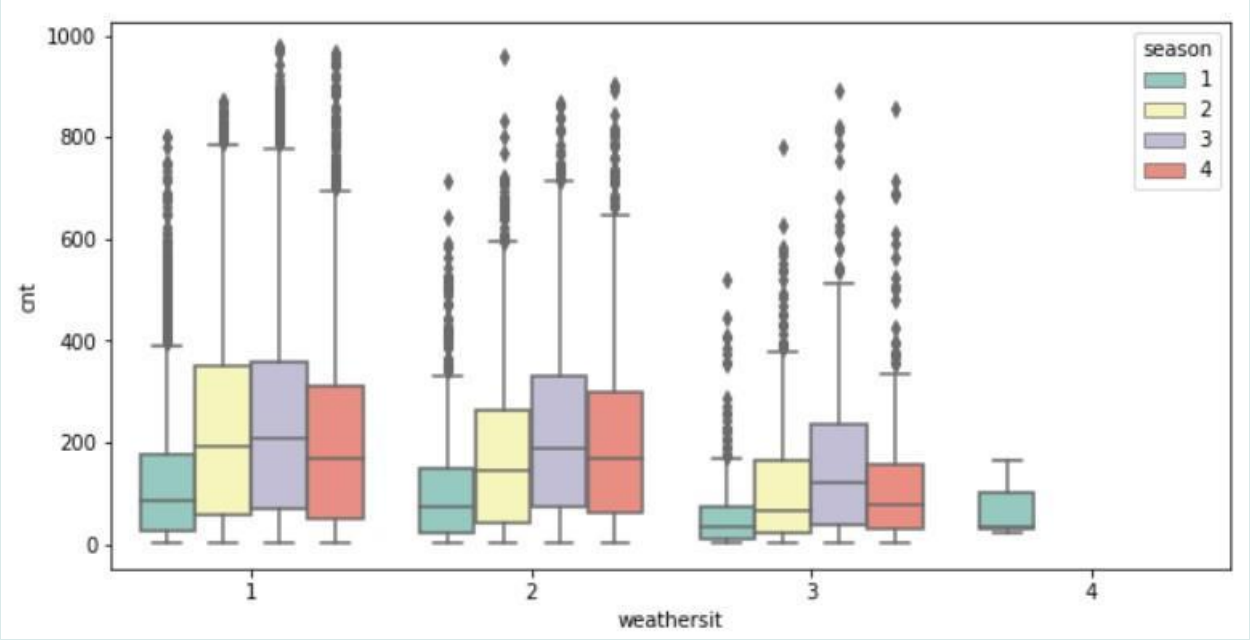
Data Analysis

- 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 4pm indicating weekdays or holidays.

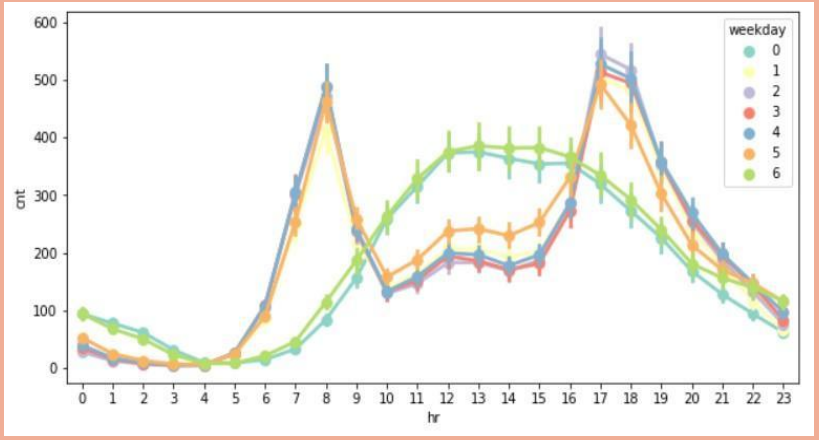
- There isn't much difference with the date of the month and it doesn't effect the number of bike users but we can see a significant increase in count over the year.

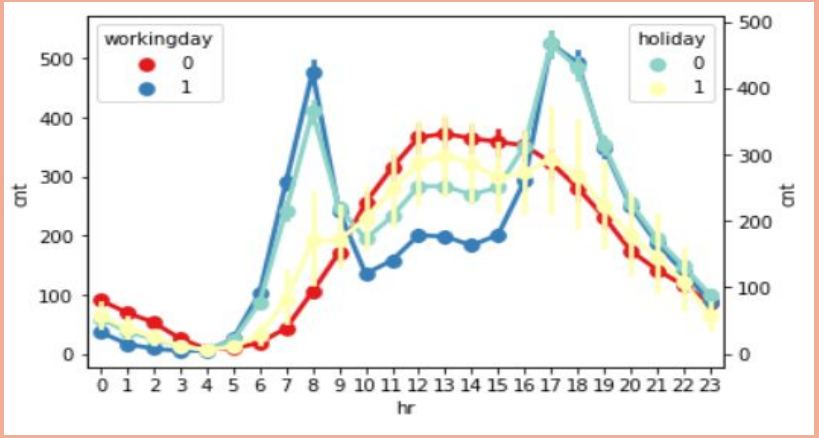






- Only season 1 has weathersit of type 4 i.e Heavy Rain + Ice Pellets + Thunderstorm + Mist, Snow + Fog. It can also be inferred that more number of users are interested to use bikes at clearer weather situations and season 1 and 2





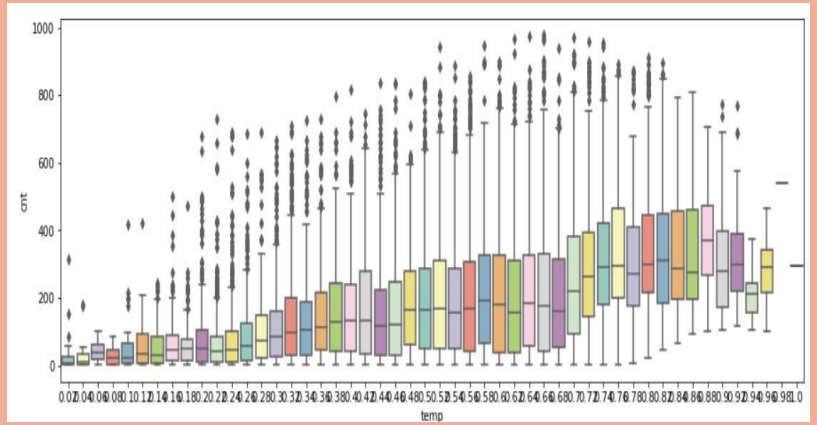
Data Analysis

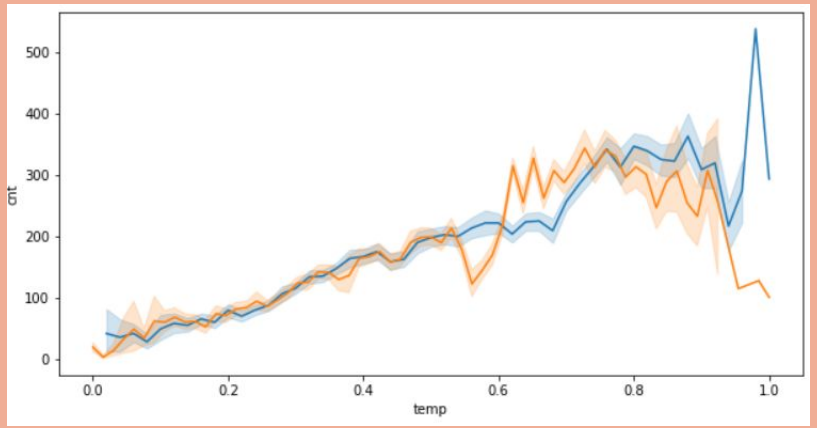
- During week day i.e 0 and 6 the peak is observed around 2pm
- For working day and holiday the cnt vs hour follow a similar pattern which is reasonable since non working days are holidays. The pattern is also similar to week the week days.

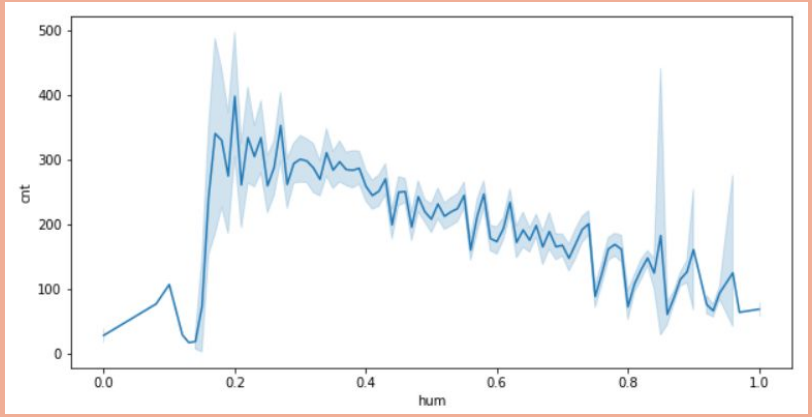
Data Analysis

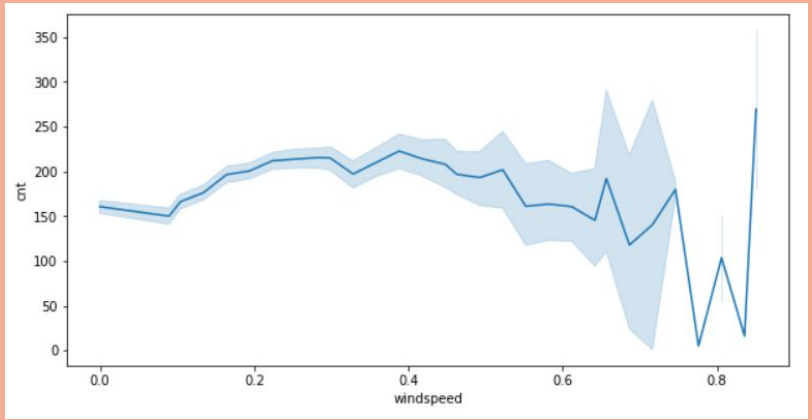
- With the increase in temperature i.e as as the climate gets warmer , the use of bikes has increased indicating the preference in warmer atmosphere.

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









Data Analysis

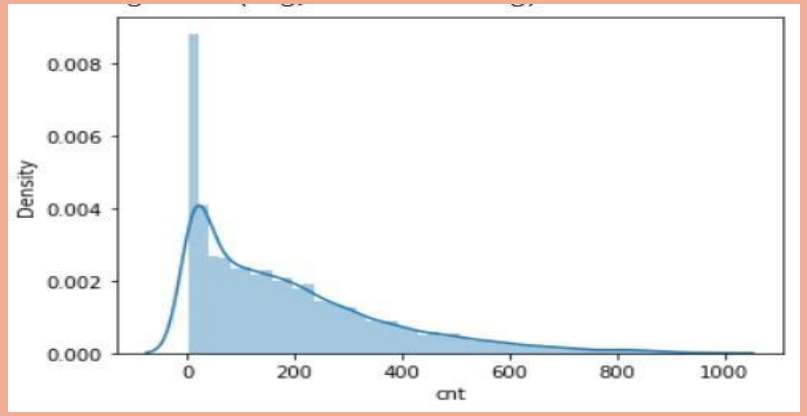
- We can see the variation in the count of bikes users with humidity, the count decreases as humidity increases.
- Windseed doesn't look like it has a greater effect on count because till 0.8 the graph is almost linear.

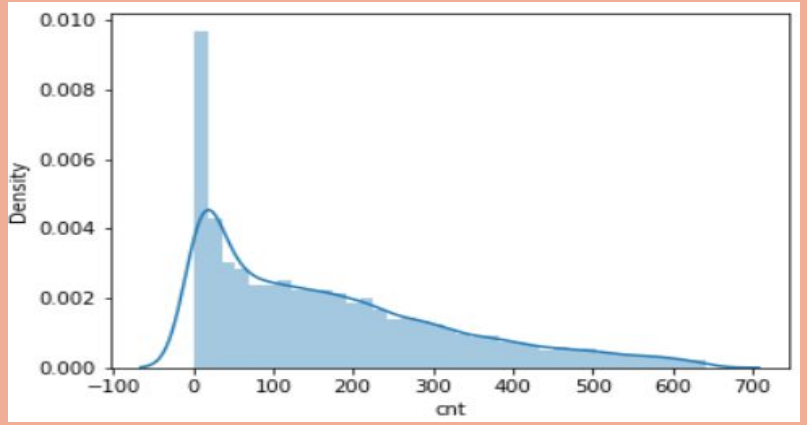
Removal of outliers

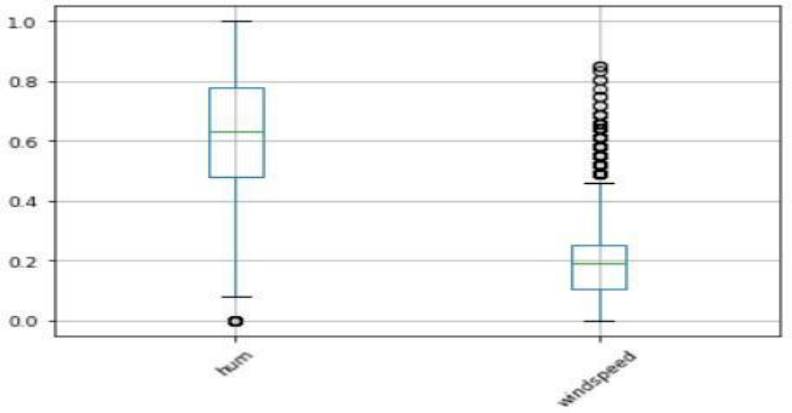
The outliers are removed by finding IQR Range and have removed the points that doesn't fall under the IQR Range

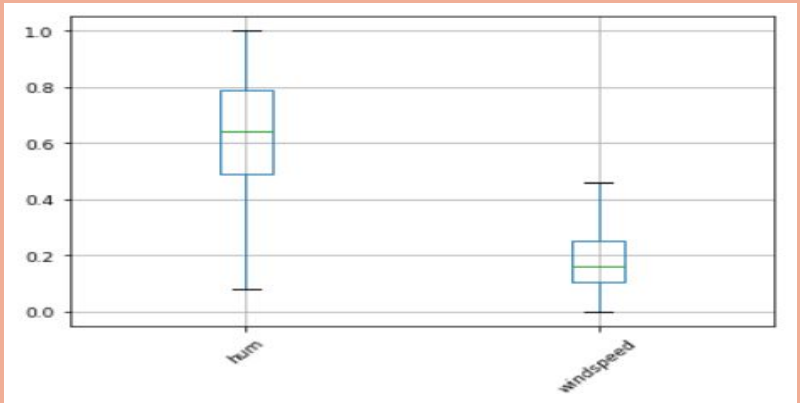
After removing outlier data from **cnt**

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









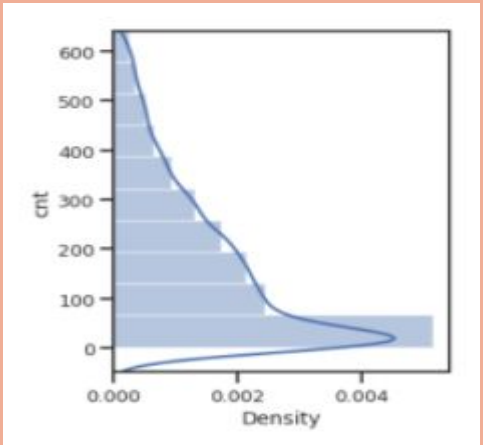
Outlier Analysis

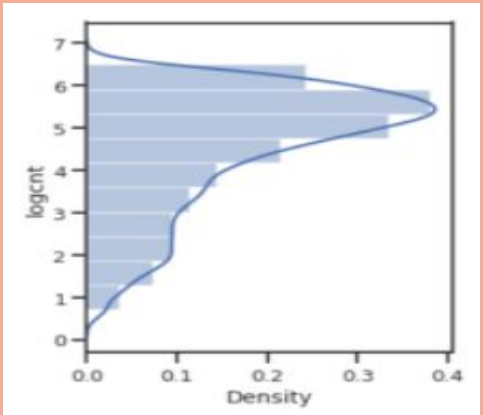
After removing outlier data from **hum**

- 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





Normalisation of target

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

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

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

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

Feature Selection

- We used **SelectKbest** with ANOVA test assessing whether the averages of more than two groups are statistically different from each other.
- The values of the scores for each variables is shown here.

	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

- We have selected 6 features and dropped weekday, holiday, dteday

Chi Square and T-test

Chi-Square test on weekday and working day:

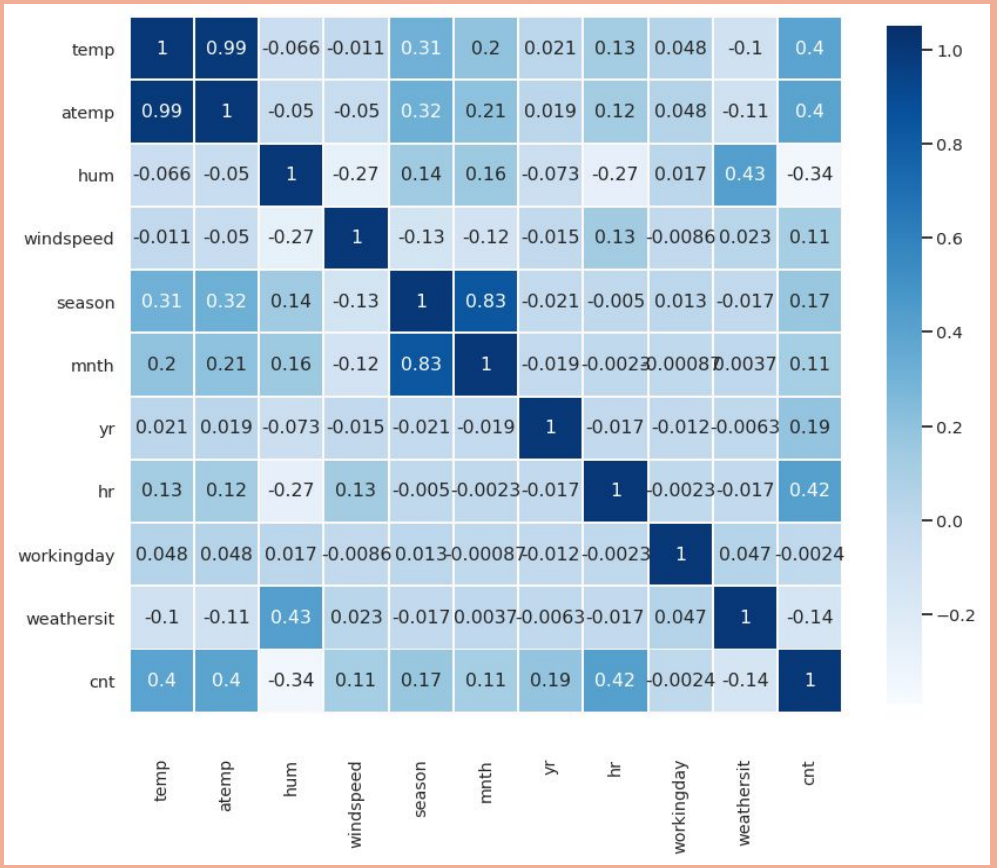
- P value is 0.0
- Dependent(reject H0)

Chi-Square Test on holiday and working day:

- P value is 4.1179732
- Dependent(reject H0)

T-test on atemp and temp :

- p-value 0.0
- t_value: 75.7829177
- We are rejecting null hypothesis.
- Thus there is a variation in atemp and temp



Correlation Analysis

- We have dropped mnth, windspeed and atemp from our feature variables.
- The final feature list contains the following after feature selection.

‘Temp’, ‘hum’, ‘season’, ‘hr’, ‘workingday’, ‘weathersit’, ‘yr’

Test of Assumptions

- Homoscedasticity
- No Multicorrelation
- No Autocorrelation
- Normality

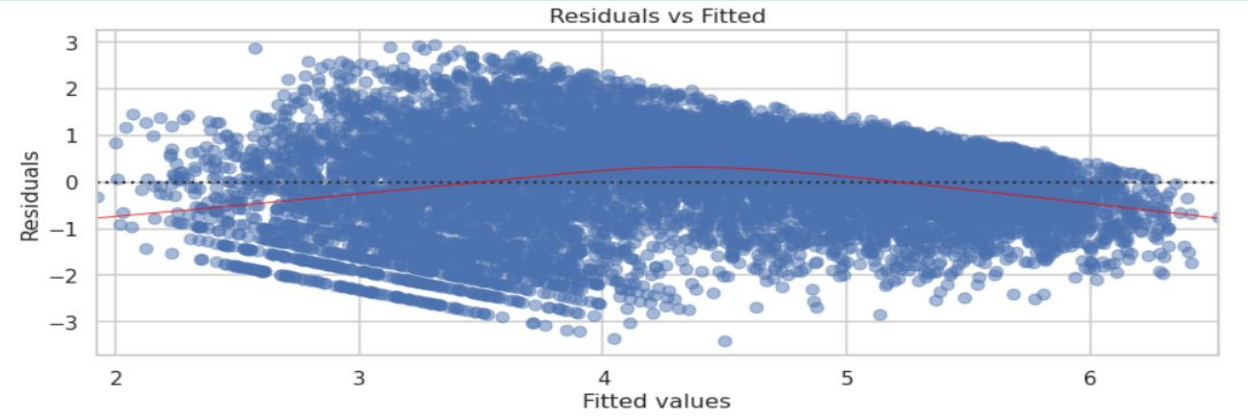
OLS Regression Results

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.

Homoscedasticity



- Also used Goldfeld-Quandt Test is used to test for heteroscedasticity.

Multicorrelation

- Tested using variation Inflation Factor (VIF)

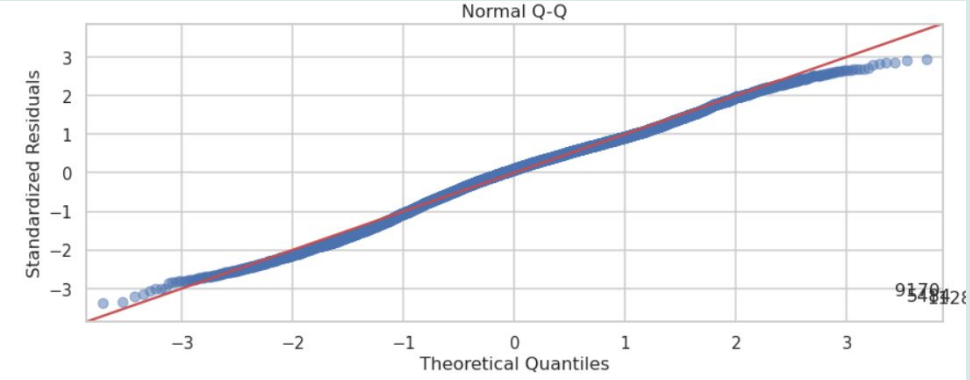
	feature	VIF
0	temp	7.507405
1	hum	11.495971
2	season	6.741054
3	hr	3.390411
4	workingday	2.966929
5	weathersit	7.408545
6	yr	1.864638

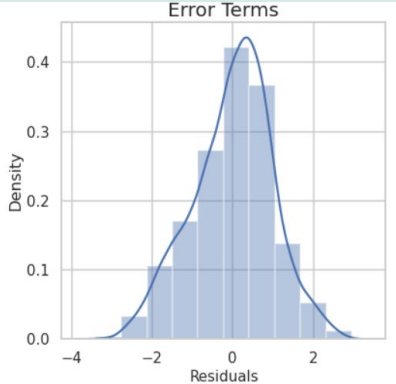
- Found humidity failed the test as its value is greater than 10

Autocorrelation

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

Normality





- Tested using Q-Q plot found straight line graph when theoretical quantiles are plotted against residuals.

Model Selection

We have considered 4 models for our problem and R^2 and MSE errors for validation set are listed below.

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

Model Selection

NuSVR has the least MSE and high R^2 values. The MSE, RMSLE, R^2 errors for training, validation and test sets are mentioned below

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

Model Selection

We have considered Random Forest Regressor that provide with feature ranking function

The MSE, RMSLE, R^2 errors are mentioned below

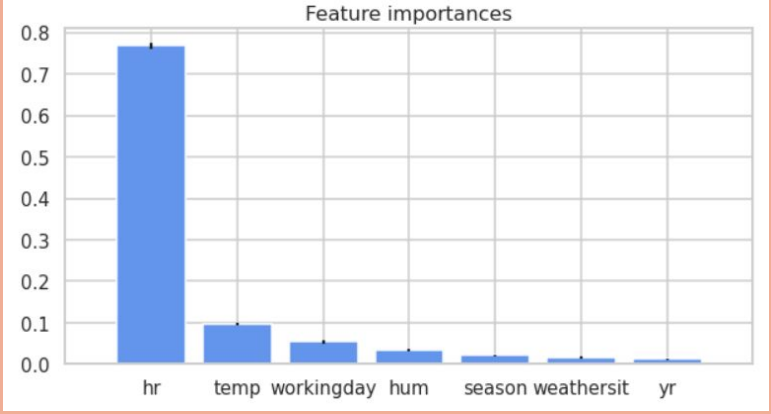
Model	Dataset	MSE	RMSLE	R^2 score
RandomForestRegressor	training	0.02	0.04	0.99
RandomForestRegressor	validation	0.24	0.10	0.87
RandomForestRegressor	test	0.30	0.11	0.84

Feature Importance

For the features we have considered for training :

- The feature ranking and the feature importance plot are given

```
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)
```



Thanks!

Team Details :

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