BIKE SHARING ASSIGNMENT

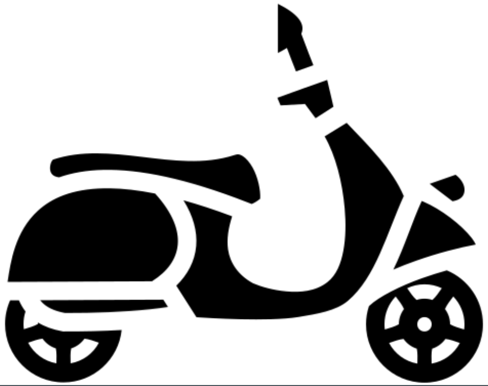
A LINEAR REGRESSION ASSIGNMENT

- SUBMITTED BY ANJALI VASHISTH

WHAT'S IN IT??

- AIM
- STEPS INVOLVED
- Assignment-based Subjective Questions
- General Subjective Questions
- CONCLUSION

AIM

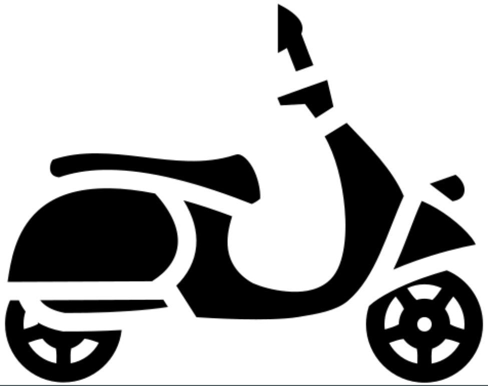


To Comprehend The Elements That Influence The Demand For These Shared Bikes.

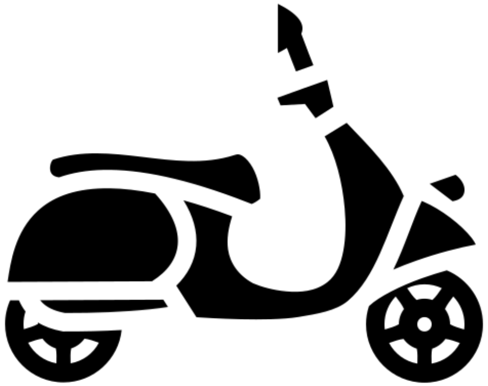
BoomBikes, In Particular, Is Interested In Learning More About The Factors That Influence Demand For These Shared Bikes In The United States. The Firm Is Curious About The Following:

- What Factors Play A Role In Estimating The Demand For Shared Bikes?
- To What Extent Do Those Factors Accurately Represent The Bike's Requirements?

STEPS INVOLVED…



1. *Reading and Understanding the Data (EDA)*
2. *Visualizing the Data*
3. *Data Preparation*
4. *Splitting the Data into Training and Testing Sets*
5. *Building a linear model*
6. *Residual Analysis of the train data*
7. *Making Predictions Using the Final Model*
8. *Model Evaluation*



Reading and Understanding the Data (EDA)

1. CHECKS AND TRANSFORMATIONS

CHECKS AND TRANSFORMATIONS

- ✓ Creating Data Frame from – "day.csv"
- ✓ Check for NULLS and DATA TYPE
- ✓ Check for the UNIQUE VALUE COLS
- ✓ Check for DUPLICACY

1. Drop "Instant", "dteday", "causal", "registered", "temp"

```
bikedata=bikedata.drop(["instant","dteday","casual", "registered","temp"], axis=1)
bikedata.sample()
```

	season	yr	mnth	holiday	weekday	workingday	weathersit	atemp	hum	windspeed	cnt
614	3	1	9	0	5	1	1	32.8602	73.625	11.500282	7504

- ✓ Dropping Columns which are not required:
- ✓ Drop "Instant", "dteday", "causal", "registered" – because these columns are not required as for "dteday" we have already month and year and we have the total of causal and registered in "cnt" variable. Also, atemp is the temperature which is temp we feel and accordingly its planned.
- ✓ Observed: year – 2018 = '0' and 2019 = '1'

VISUALIZING THE DATA

1. EASY VISUALIZATIONS

EASY VISUALIZATIONS –

Replaced digits with there real meanings

	season	yr	mnth	holiday	weekday	workingday	weathersit
0	1	0	1	0	6	0	2
1	1	0	1	0	0	0	2
2	1	0	1	0	1	1	1
3	1	0	1	0	2	1	1
4	1	0	1	0	3	1	1

```
bikevisual.head()
```

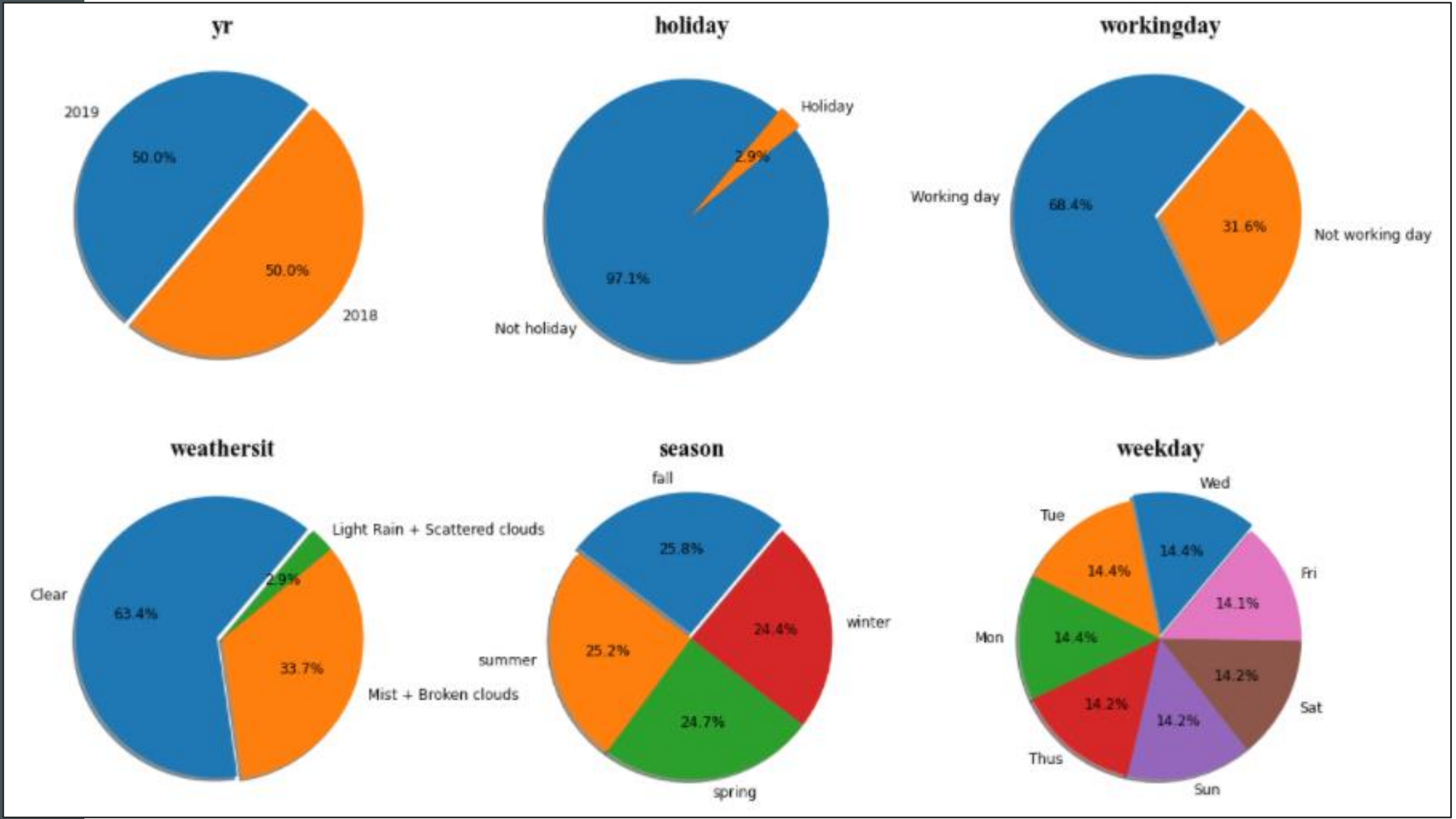
	season	yr	mnth	holiday	weekday	workingday	weathersit
0	spring	2018	Jan	Not holiday	Mon	Not working day	Mist + Broken clouds
1	spring	2018	Jan	Not holiday	Tue	Not working day	Mist + Broken clouds
2	spring	2018	Jan	Not holiday	Wed	Working day	Clear
3	spring	2018	Jan	Not holiday	Thus	Working day	Clear
4	spring	2018	Jan	Not holiday	Fri	Working day	Clear

VISUALIZING THE DATA

2. UNIVARIATE ANALYSIS

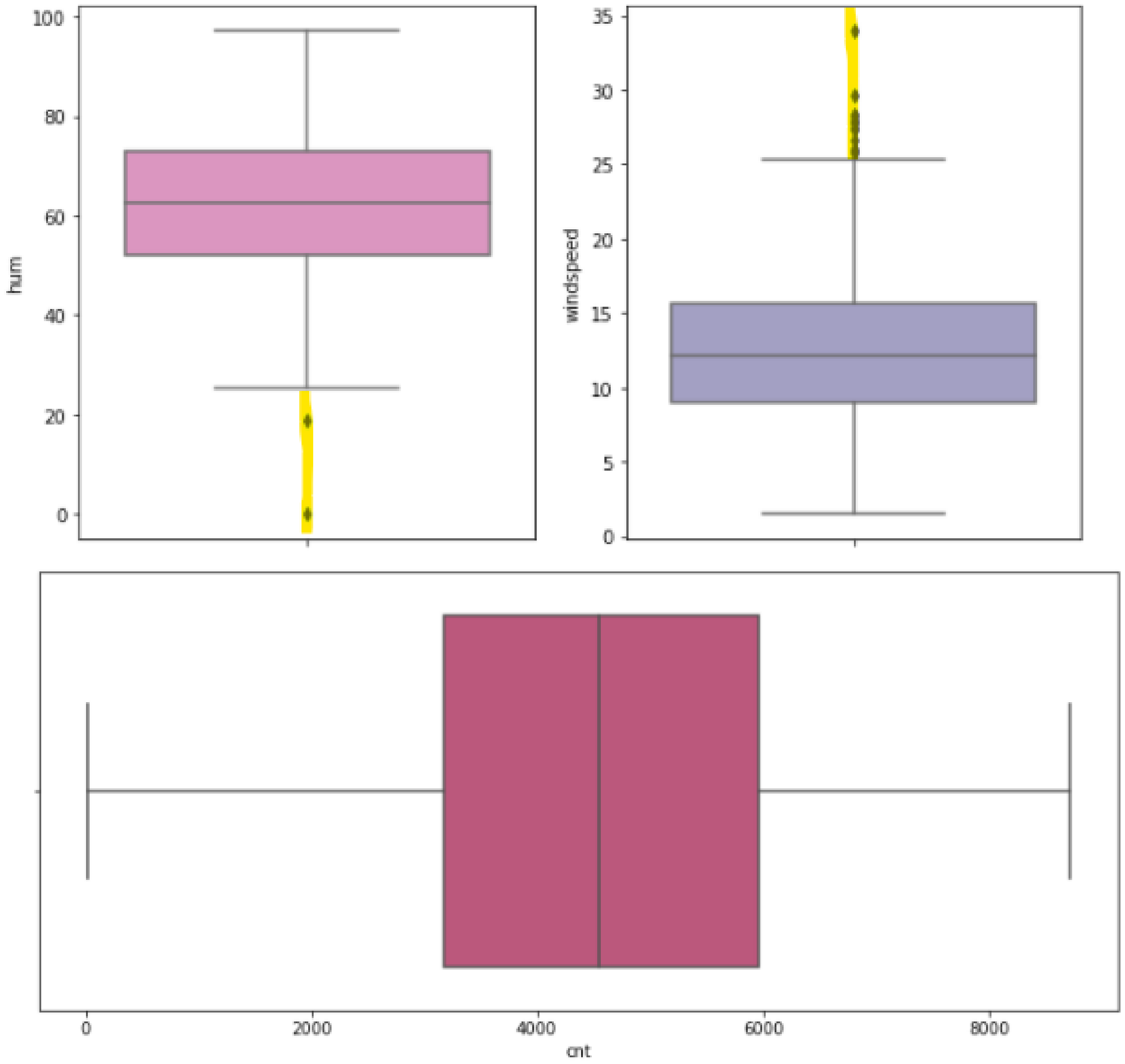
UNIVARIATE ANALYSIS – VISUALISING CATEGORICAL VARIABLES

- ✓ In two years span, there is only 2.9% holiday, and the rest are nonholiday.
- ✓ However, 68.4% is a working day, and 31.6% is a non-working day.
- ✓ Most of the time (63.4%), it's a clear sky, and rarely (2.9%) is a thunderstorm.



UNIVARIATE ANALYSIS – VISUALISING NUMERICAL VARIABLES

- ✓ WINDSPEED AND HUMIDITY SEEM TO HAVE OUTLIERS
- ✓ For "cnt" the value lies between 3000 to 6000.



VISUALIZING THE DATA

3. BIVARIATE ANALYSIS

BIVARIATE ANALYSIS – VISUALISING CATEGORICAL VARIABLES

“cnt” v/s {}:

1. {“yr”} –

   2019 had more cnt in range 4500 to 7000 than 2018

2. {“holiday”}

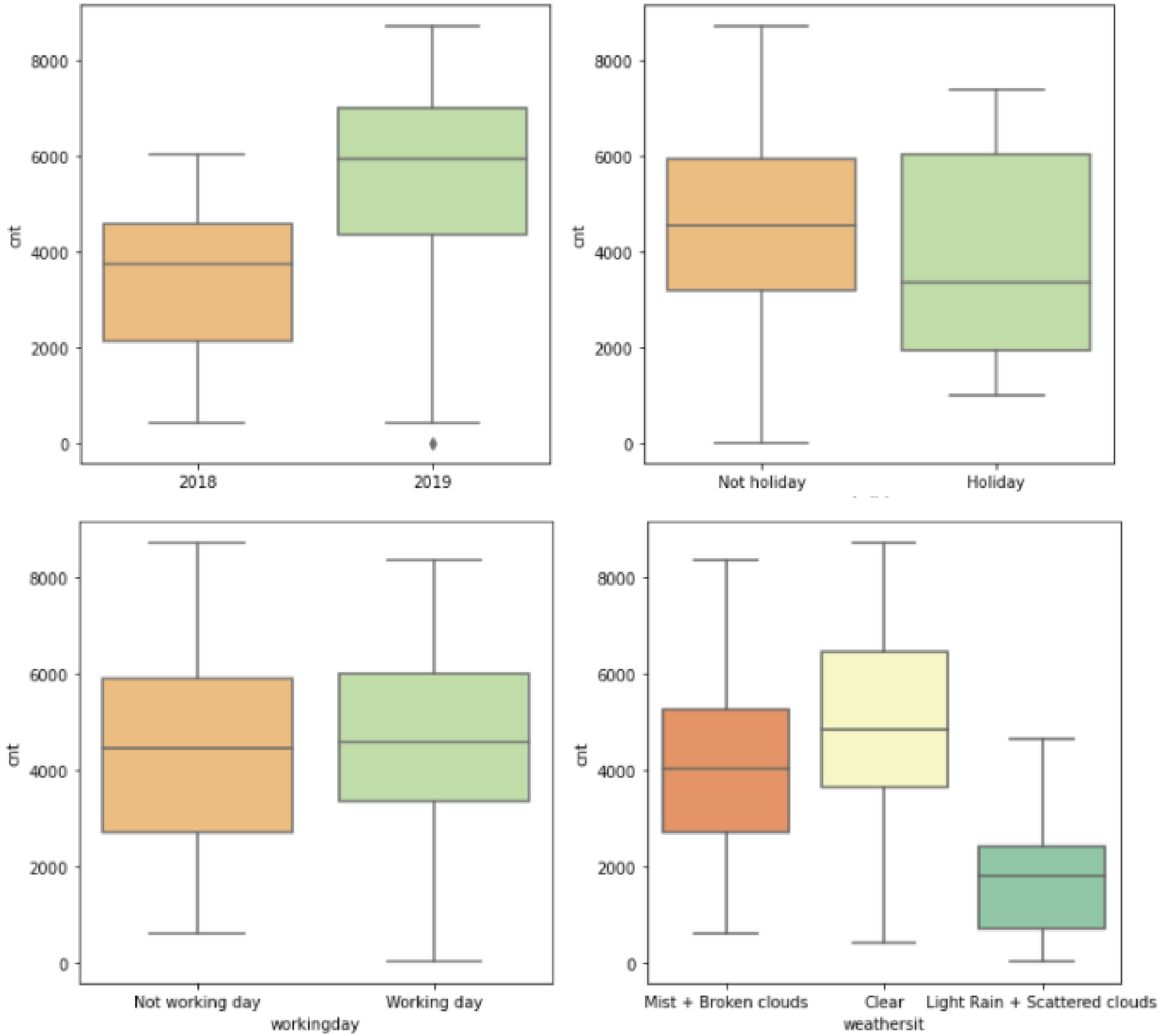
   More people rode bike on holidays

3. {“working day”}

   there is a slight chance that people will borrow bikes more on a nonworking day.

4. {“weathersit”}

   It is always best to ride bikes in clear weather followed by mist and broken clouds



BIVARIATE ANALYSIS – VISUALISING CATEGORICAL VARIABLES

"cnt" v/s {}:

5. {"season"}

   Most people tend to rent bikes in the fall and summer.

   And people prefer to book less in spring.

6. {"weekday"}

   Mean are the same for all weekdays.

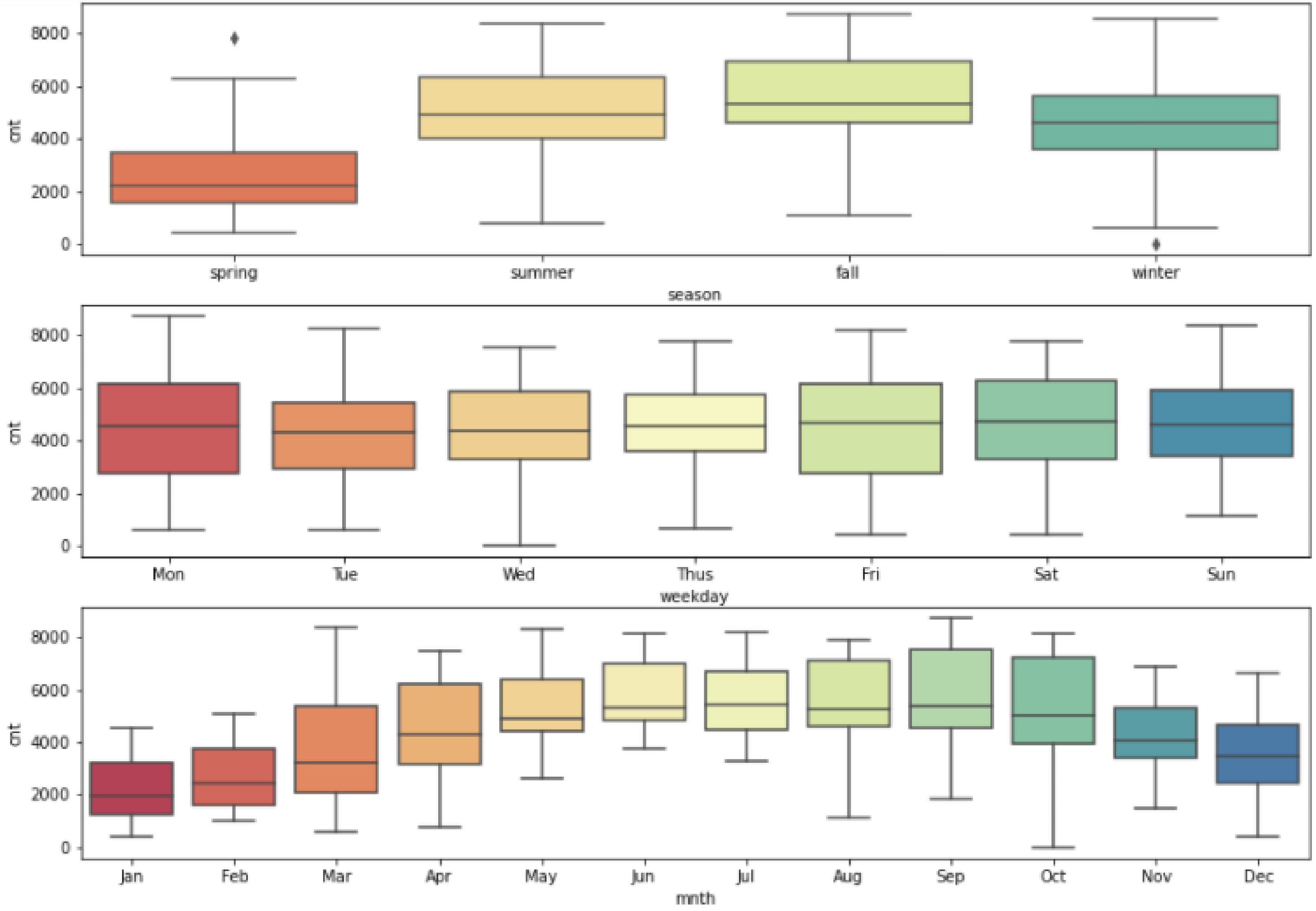
   But Tuesday has more chance of booking less the other weekday.

   And Monday has a high probability of renting a bike.

7. {"mnth"}

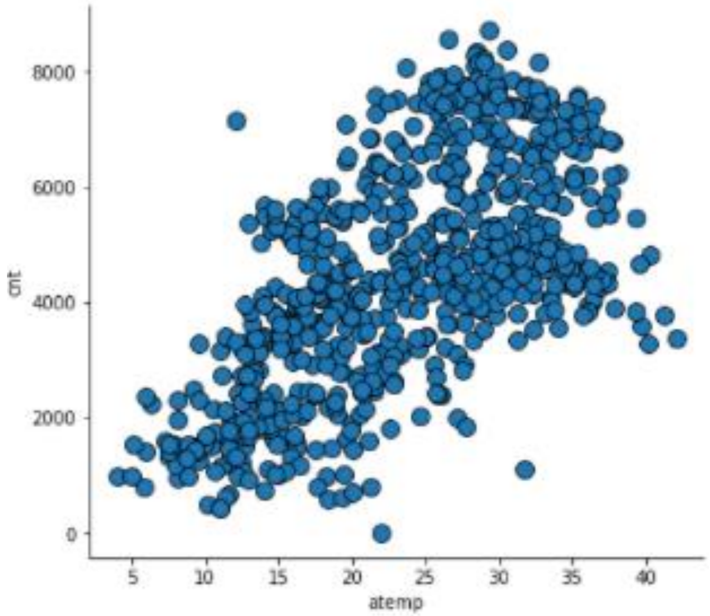
   People prefer to rent a bike between May to October.

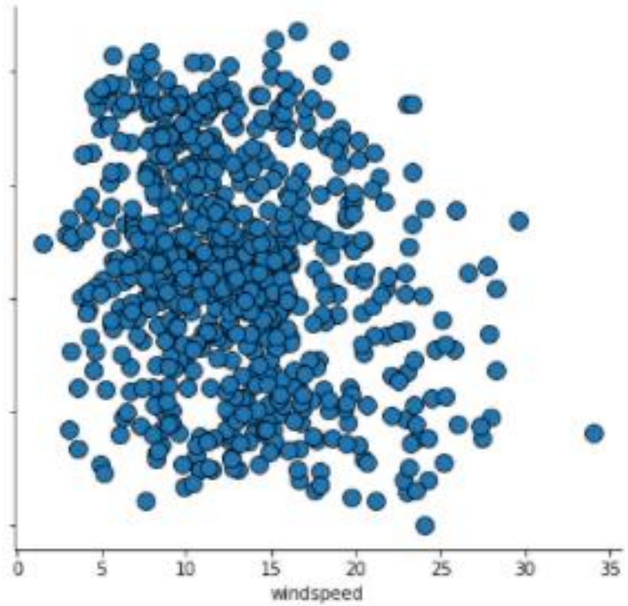
   And gradually decrease from November to Feb.

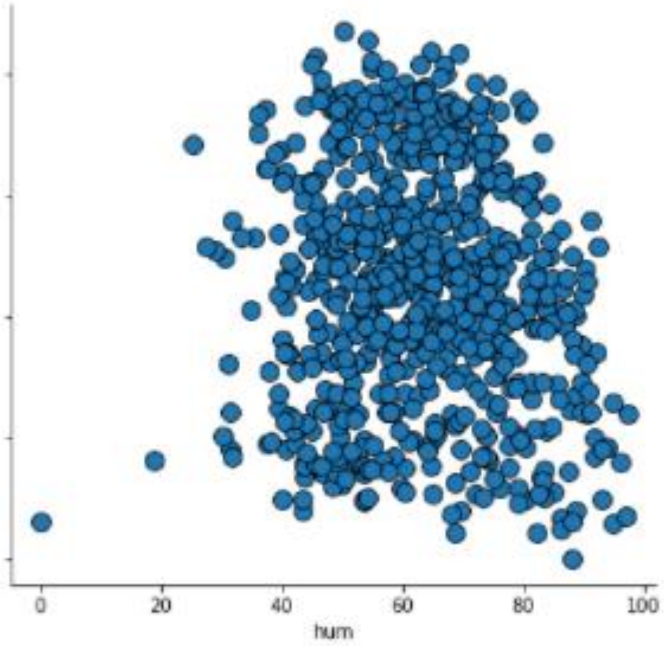


BIVARIATE ANALYSIS – VISUALISING NUMERICAL VARIABLES

- ✓ we can see when the “atemp” is around 36+ people tend to rent bikes less than temp between 20-33 people tend to book more.
- ✓ – When “hum” is between 40-85 people rent bikes more.
- ✓ – When “windspeed” is between 5-25 people rent bikes more.







VISUALIZING THE DATA

4. **MULTIVARIATE ANALYSIS**

MULTIVARIATE ANALYSIS – VISUALISING NUMERICAL VARIABLES

1. atemp vs. cnt vs. windspeed:

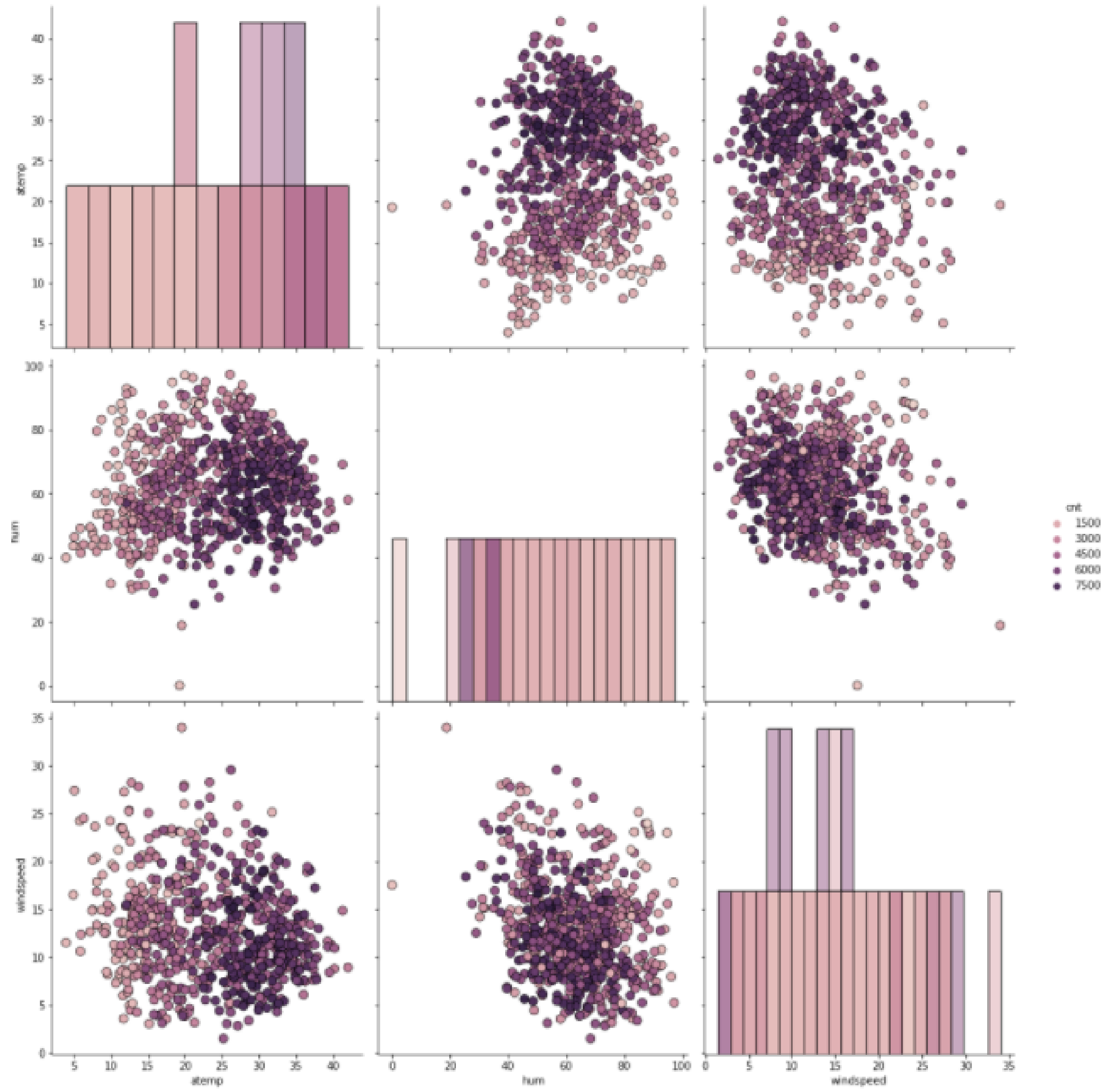
Most people rent bikes more when the atemp is between 20-35 with a wind speed of 5-20.

2. atemp vs. hum vs. cnt:

people rent bikes more when the atemp is between 20-35, with the hum between 40-80.

3. hum vs. cnt vs. windspeed:

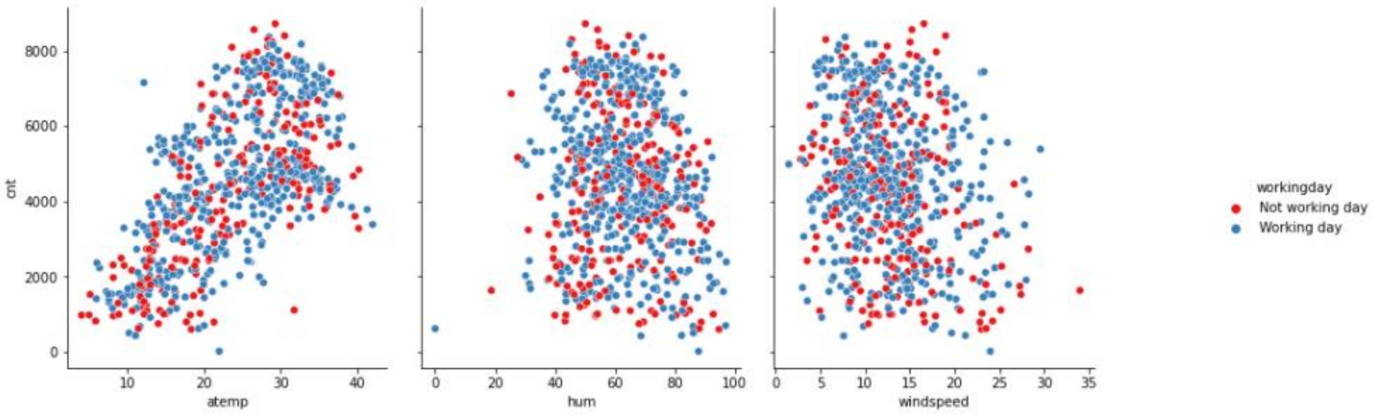
people rent bikes more when the hum is between 40-75, with windspeed between 5-20



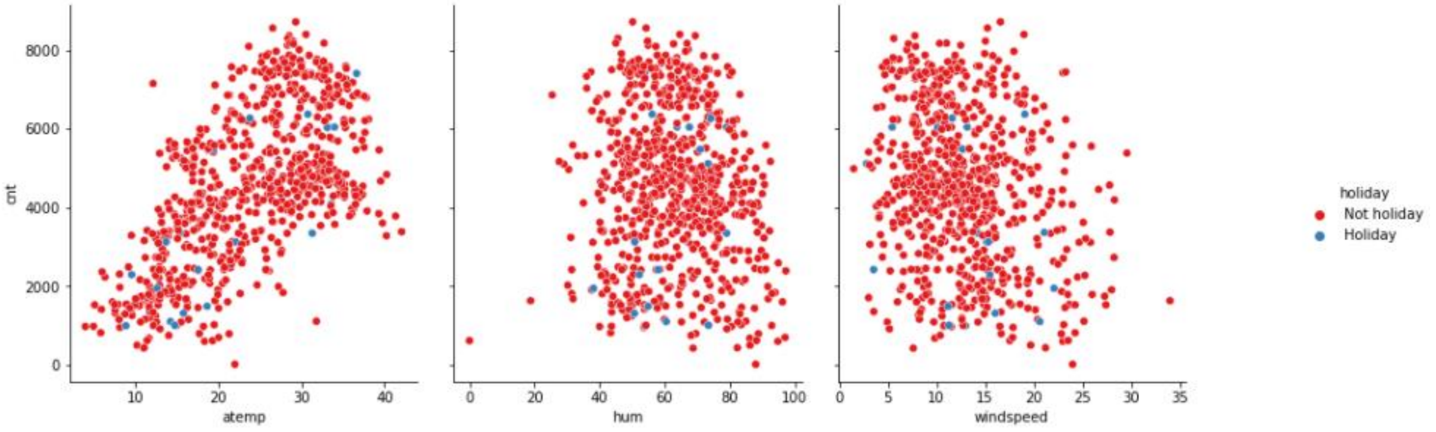
MULTIVARIATE ANALYSIS – VISUALISING CATEGORICAL VARIABLES

1. Distribution is kinda same for all – Working Day
2. For Holiday – we can explicitly see – Not holidays impact more

Working day

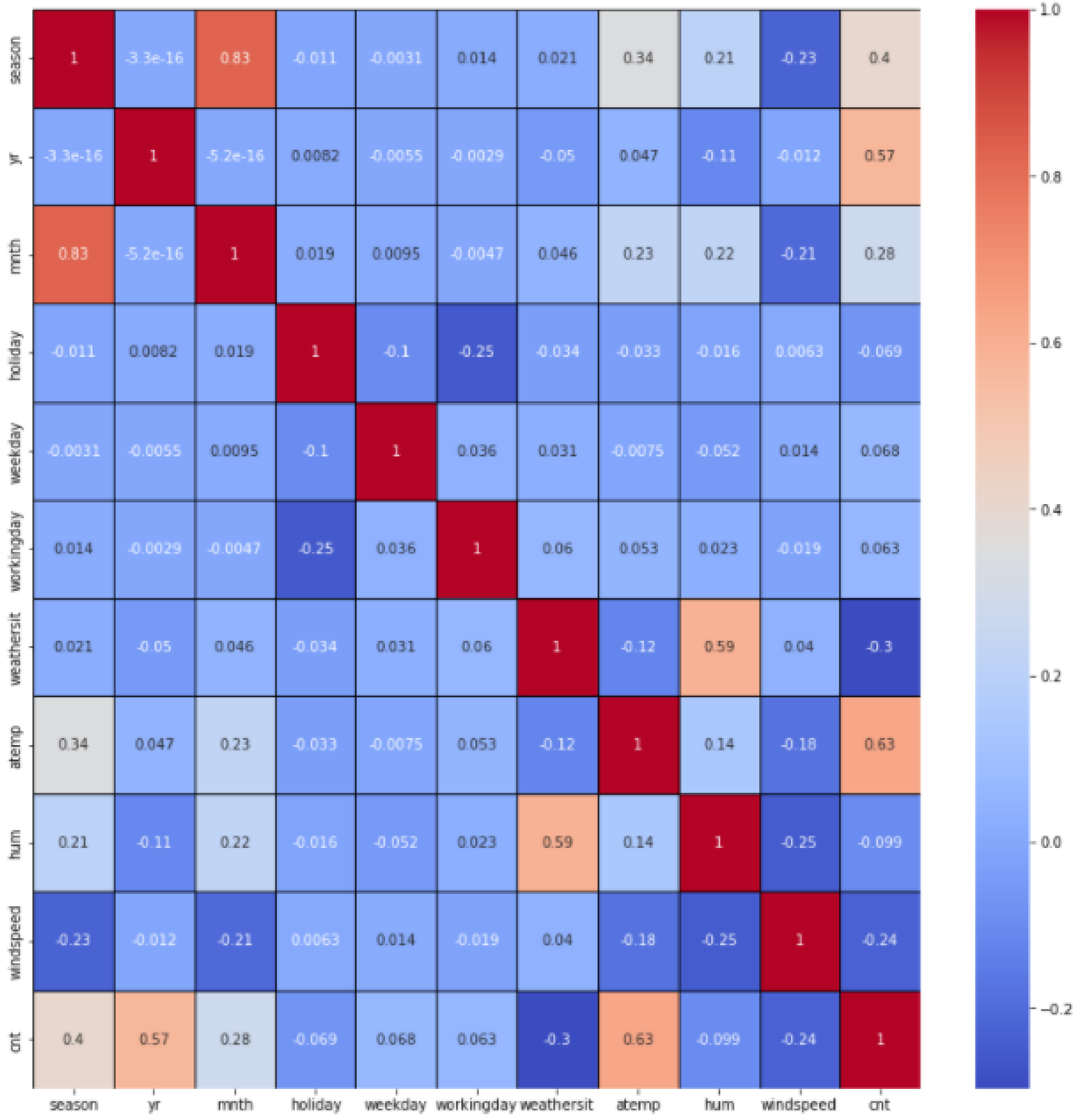


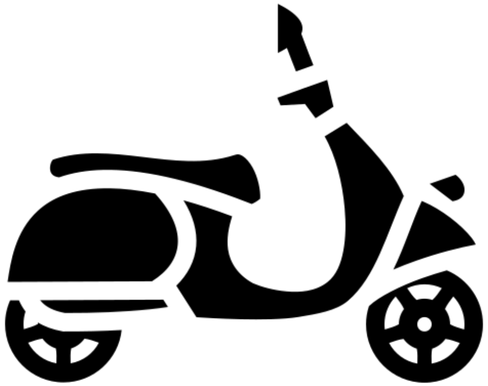
Holiday



MULTIVARIATE ANALYSIS – VISUALISING CATEGORICAL VARIABLES

1. season, mnth, atemp and yr shows high correlation with cnt
2. weathersit, hum, holiday and windspeed show high inverse correlation with cnt





DATA PREPARATION

1. **ENCODING**

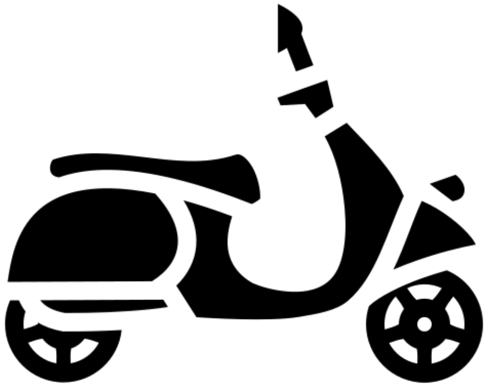
ENCODING

1. Create dummy variable for weathersit, season, weekday and mnth respectively
2. And Removing the columns which we have created a dummy variable
3. Converting year, holiday, and working day to 1 and 0s, so that we can use for model building
4. Merging the dataframe

```
   Light Rain + Scattered clouds  Mist + Broken clouds
0                              0                     1
1                              0                     1
2                              0                     0
3                              0                     0
4                              0                     0
   spring  summer  winter
0       1       0       0
1       1       0       0
2       1       0       0
3       1       0       0
4       1       0       0
   Mon  Sat  Sun  Thus  Tue  Wed
0    1    0    0     0    0    0
1    0    0    0     0    1    0
2    0    0    0     0    0    1
3    0    0    0     1    0    0
4    0    0    0     0    0    0
   Aug  Dec  Feb  Jan  Jul  Jun  Mar  May  Nov  Oct  Sep
0    0    0    0    1    0    0    0    0    0    0    0
1    0    0    0    1    0    0    0    0    0    0    0
2    0    0    0    1    0    0    0    0    0    0    0
3    0    0    0    1    0    0    0    0    0    0    0
4    0    0    0    1    0    0    0    0    0    0    0
```

	yr	holiday	workingday	atemp	hum	windspeed	cnt	Light Rain + Scattered clouds	Mist + Broken clouds	spring	...	Dec	Feb	Jan	Jul	Jun	Mar	May	Nov	Oct	Sep
0	0	0	0	18.18125	80.5833	10.749882	985	0	1	1	...	0	0	1	0	0	0	0	0	0	0
1	0	0	0	17.68695	69.6087	16.652113	801	0	1	1	...	0	0	1	0	0	0	0	0	0	0
2	0	0	1	9.47025	43.7273	16.636703	1349	0	0	1	...	0	0	1	0	0	0	0	0	0	0
3	0	0	1	10.60610	59.0435	10.739832	1562	0	0	1	...	0	0	1	0	0	0	0	0	0	0
4	0	0	1	11.46350	43.6957	12.522300	1600	0	0	1	...	0	0	1	0	0	0	0	0	0	0

5 rows × 29 columns



SPLITTING THE DATA INTO TRAINING AND TESTING SETS

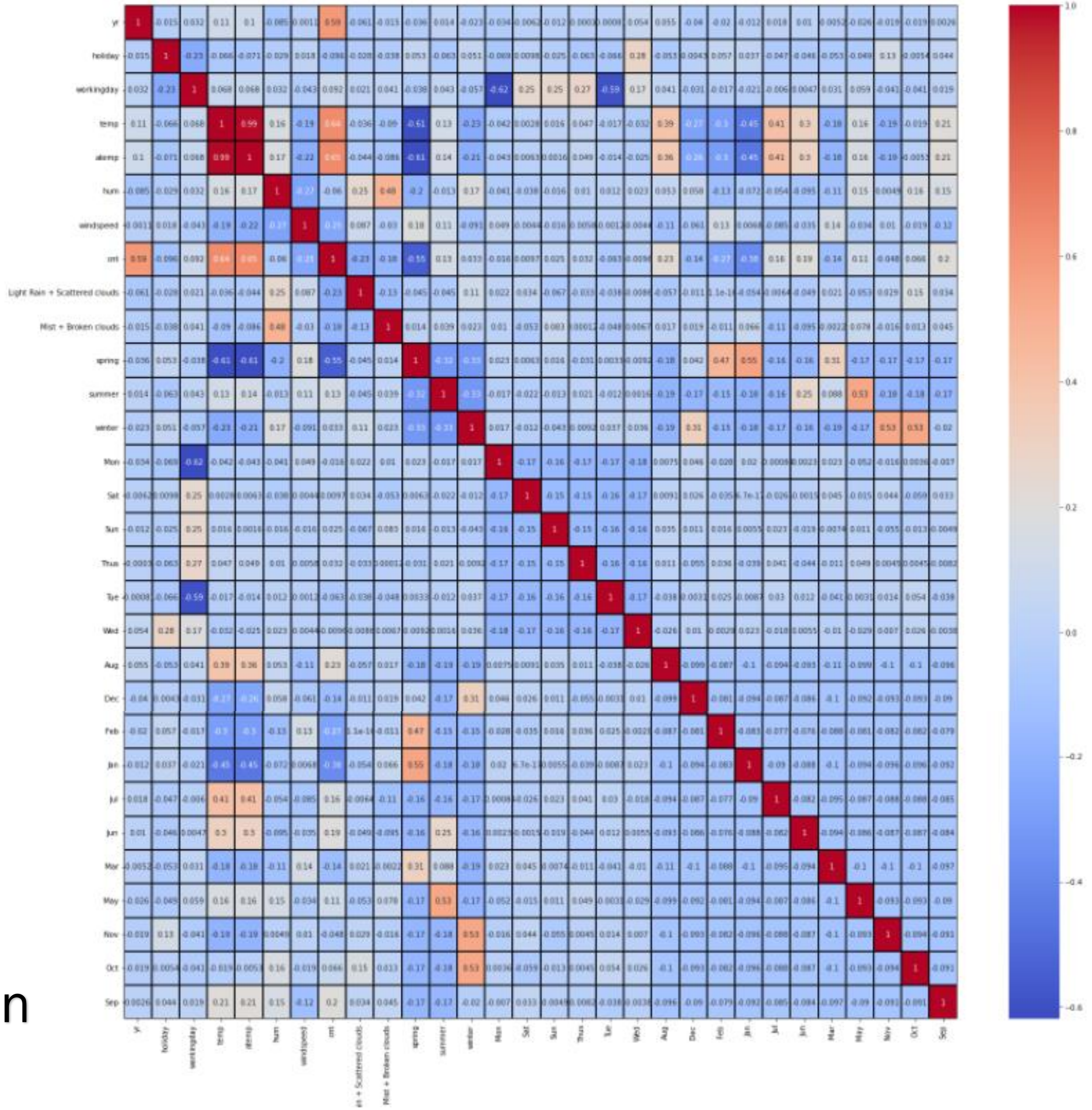
1. TRAIN AND TEST ()

TRAIN TEST

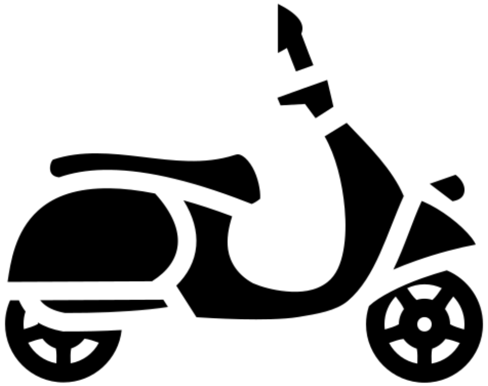
Importing libraries for training and testing data and dividing them into 70 -30 ratio

Rescaling feature of bike_train data frame - Apply *scaler* to all the columns except the '1s and 0s' and 'dummy' variables and fit-transforming it using **Min-Max scaling**

We find that There is multi-collinearity between the variables, but while building model will use RFE to auto-remove it.



Train



BUILDING A LINEAR MODEL

1. RFE (RECURSIVE FEATURE ELIMINATION)

RFE (RECURSIVE FEATURE ELIMINATION)

applying RFE to auto eliminate till 18 variable

And making a list of column names, support and ranking of RFE

Finally collecting the supporting columns with rank 1.

	name	support	rank
0	yr	True	1
26	Nov	True	1
22	Jul	True	1
21	Jan	True	1
20	Feb	True	1
19	Dec	True	1
12	Mon	True	1
11	winter	True	1
10	summer	True	1
28	Sep	True	1
8	Mist + Broken clouds	True	1
1	holiday	True	1
2	workingday	True	1
3	temp	True	1
9	spring	True	1
5	hum	True	1
6	windspeed	True	1
7	Light Rain + Scattered clouds	True	1

BUILDING A LINEAR MODEL

2. BUILDING MODEL USING STATS MODEL, FOR THE DETAILED STATISTICS

1ST MODEL SUMMARY

Turns out Feb has High P Value

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.852
Model:                            OLS   Adj. R-squared:                  0.846
Method:                 Least Squares   F-statistic:                     156.6
Date:                Sun, 07 Nov 2021   Prob (F-statistic):          3.55e-190
Time:                        16:20:03   Log-Likelihood:                 525.15
No. Observations:                 510   AIC:                            -1012.
Df Residuals:                     491   BIC:                            -931.8
Df Model:                          18
Covariance Type:            nonrobust
===================================================================================================
                                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------------
const                             0.2873      0.038      7.630      0.000       0.213       0.361
yr                                0.2311      0.008     29.082      0.000       0.215       0.247
holiday                          -0.0499      0.027     -1.857      0.064      -0.103       0.003
workingday                        0.0443      0.011      3.878      0.000       0.022       0.067
temp                              0.4598      0.038     12.233      0.000       0.386       0.534
hum                              -0.1456      0.037     -3.904      0.000      -0.219      -0.072
windspeed                        -0.1887      0.025     -7.440      0.000      -0.239      -0.139
Light Rain + Scattered clouds    -0.2583      0.026     -9.924      0.000      -0.309      -0.207
Mist + Broken clouds             -0.0600      0.010     -5.813      0.000      -0.080      -0.040
spring                           -0.0518      0.022     -2.390      0.017      -0.094      -0.009
summer                            0.0377      0.015      2.483      0.013       0.008       0.068
winter                            0.1035      0.018      5.852      0.000       0.069       0.138
Mon                               0.0542      0.014      3.761      0.000       0.026       0.083
Dec                              -0.0491      0.018     -2.727      0.007      -0.085      -0.014
Feb                              -0.0339      0.021     -1.591      0.112      -0.076       0.008
Jan                              -0.0640      0.021     -3.025      0.003      -0.106      -0.022
Jul                              -0.0517      0.018     -2.883      0.004      -0.087      -0.016
Nov                              -0.0465      0.019     -2.499      0.013      -0.083      -0.010
Sep                               0.0718      0.017      4.319      0.000       0.039       0.104
==============================================================================
Omnibus:                       81.478   Durbin-Watson:                   2.041
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              219.245
Skew:                          -0.787   Prob(JB):                     2.46e-48
Kurtosis:                       5.800   Cond. No.                         23.7
==============================================================================
```

1ST VIF RULES & VALUES

But our rule for removing variable is:

1. Both VIF(5+) and P-value(0.05+) is high:

   - If p-value is high and VIF value is high then remove it.

2. High-low:

   - If p-value is high and VIF value is low then remove it.

   - If p-value is low and VIF value is high then remove it.

3. Both VIF(5<) and P-value(0.05<) is Low:

   - We keep the variable.

	Features	VIF
4	hum	27.37
3	atemp	18.36
2	workingday	5.25
8	spring	4.28
5	windspeed	4.15
9	winter	3.12
13	Jan	2.39
7	Mist + Broken clouds	2.27
0	yr	2.07
10	Mon	1.97
12	Feb	1.92
16	Nov	1.85
11	Dec	1.67
14	Jul	1.47
15	May	1.33
6	Light Rain + Scattered clouds	1.27
17	Sep	1.26
1	holiday	1.20

However, `Feb` is having high p-value and low VIF, so we will remove it.

2ND MODEL SUMMARY

Feb col was dropped;

Turns out Holiday has High P Value Now its VIF value will be Checked

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.848
Model:                            OLS   Adj. R-squared:                  0.843
Method:                 Least Squares   F-statistic:                     161.5
Date:                Tue, 15 Mar 2022   Prob (F-statistic):          1.08e-188
Time:                        23:47:50   Log-Likelihood:                 518.91
No. Observations:                 510   AIC:                            -1002.
Df Residuals:                     492   BIC:                            -925.6
Df Model:                          17
Covariance Type:            nonrobust
===================================================================================================
                                      coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------------
const                               0.3243      0.033      9.776      0.000       0.259       0.389
yr                                  0.2331      0.008     29.047      0.000       0.217       0.249
holiday                            -0.0467      0.027     -1.718      0.086      -0.100       0.007
workingday                          0.0448      0.012      3.877      0.000       0.022       0.068
atemp                               0.4446      0.034     12.968      0.000       0.377       0.512
hum                                -0.1509      0.038     -3.972      0.000      -0.226      -0.076
windspeed                          -0.1671      0.026     -6.530      0.000      -0.217      -0.117
Light Rain + Scattered clouds      -0.2535      0.026     -9.607      0.000      -0.305      -0.202
Mist + Broken clouds               -0.0609      0.010     -5.833      0.000      -0.081      -0.040
spring                             -0.0990      0.015     -6.414      0.000      -0.129      -0.069
winter                              0.0757      0.014      5.320      0.000       0.048       0.104
Mon                                 0.0561      0.015      3.853      0.000       0.028       0.085
Dec                                -0.0494      0.017     -2.892      0.004      -0.083      -0.016
Jan                                -0.0521      0.018     -2.918      0.004      -0.087      -0.017
Jul                                -0.0581      0.017     -3.369      0.001      -0.092      -0.024
May                                 0.0339      0.016      2.161      0.031       0.003       0.065
Nov                                -0.0509      0.018     -2.773      0.006      -0.087      -0.015
Sep                                 0.0637      0.016      4.070      0.000       0.033       0.094
==============================================================================
Omnibus:                       79.323   Durbin-Watson:                   2.019
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              251.889
Skew:                          -0.713   Prob(JB):                     2.01e-55
Kurtosis:                       6.134   Cond. No.                         20.9
==============================================================================
```

2^{ND} VIF RULES & VALUES

But our rule for removing variable is:

1. Both VIF(5+) and P-value(0.05+) is high:

   - If p-value is high and VIF value is high then remove it.

2. High-low:

   - If p-value is high and VIF value is low then remove it.

   - If p-value is low and VIF value is high then remove it.

3. Both VIF(5<) and P-value(0.05<) is Low:

   - We keep the variable.

	Features	VIF
4	hum	26.90
3	atemp	17.83
2	workingday	5.24
5	windspeed	4.14
8	spring	3.11
9	winter	3.11
7	Mist + Broken clouds	2.27
0	yr	2.07
10	Mon	1.97
15	Nov	1.83
12	Jan	1.76
11	Dec	1.55
13	Jul	1.46
14	May	1.33
6	Light Rain + Scattered clouds	1.26
16	Sep	1.26
1	holiday	1.19

\`Holiday\` is having high p-value and low VIF, so we will remove it.

3RD MODEL SUMMARY

Feb and holiday col was dropped;

Turns out May has High P Value as its val is not less than .030
Now its VIF value will be Checked

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.847
Model:                            OLS   Adj. R-squared:                  0.842
Method:                 Least Squares   F-statistic:                     170.7
Date:                Tue, 15 Mar 2022   Prob (F-statistic):          3.52e-189
Time:                        23:55:41   Log-Likelihood:                 517.39
No. Observations:                 510   AIC:                            -1001.
Df Residuals:                     493   BIC:                            -928.8
Df Model:                          16
Covariance Type:            nonrobust
=================================================================================================
                                    coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
const                             0.3178      0.033      9.624      0.000       0.253       0.383
yr                                0.2333      0.008     29.014      0.000       0.217       0.249
workingday                        0.0517      0.011      4.757      0.000       0.030       0.073
atemp                             0.4437      0.034     12.916      0.000       0.376       0.511
hum                              -0.1500      0.038     -3.939      0.000      -0.225      -0.075
windspeed                        -0.1675      0.026     -6.533      0.000      -0.218      -0.117
Light Rain + Scattered clouds    -0.2531      0.026     -9.573      0.000      -0.305      -0.201
Mist + Broken clouds             -0.0607      0.010     -5.802      0.000      -0.081      -0.040
spring                           -0.1001      0.015     -6.472      0.000      -0.130      -0.070
winter                            0.0761      0.014      5.338      0.000       0.048       0.104
Mon                               0.0631      0.014      4.496      0.000       0.036       0.091
Dec                              -0.0502      0.017     -2.931      0.004      -0.084      -0.017
Jan                              -0.0528      0.018     -2.953      0.003      -0.088      -0.018
Jul                              -0.0571      0.017     -3.309      0.001      -0.091      -0.023
May                               0.0343      0.016      2.181      0.030       0.003       0.065
Nov                              -0.0545      0.018     -2.987      0.003      -0.090      -0.019
Sep                               0.0617      0.016      3.948      0.000       0.031       0.092
==============================================================================
Omnibus:                       85.041   Durbin-Watson:                   2.004
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              279.769
Skew:                          -0.754   Prob(JB):                     1.77e-61
Kurtosis:                       6.300   Cond. No.                         20.9
==============================================================================
```

3RD VIF RULES & VALUES

But our rule for removing variable is:

1. Both VIF(5+) and P-value(0.05+) is high:

    - If p-value is high and VIF value is high then remove it.

2. High-low:

    - If p-value is high and VIF value is low then remove it.

    - If p-value is low and VIF value is high then remove it.

3. Both VIF(5<) and P-value(0.05<) is Low:

    - We keep the variable.

	Features	VIF
3	hum	26.79
2	atemp	17.71
1	workingday	4.67
4	windspeed	4.12
8	winter	3.11
7	spring	3.08
6	Mist + Broken clouds	2.26
0	yr	2.07
9	Mon	1.84
14	Nov	1.80
11	Jan	1.76
10	Dec	1.55
12	Jul	1.46
13	May	1.33
5	Light Rain + Scattered clouds	1.26
15	Sep	1.26

`May` is having high p-value and low VIF, so we will remove it.

4^{TH} MODEL SUMMARY

Feb, May and holiday col was dropped;

Every thing seems fine here – all variables have low P value now

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.846
Model:                            OLS   Adj. R-squared:                  0.841
Method:                 Least Squares   F-statistic:                     180.4
Date:                Wed, 16 Mar 2022   Prob (F-statistic):          2.71e-189
Time:                        00:11:56   Log-Likelihood:                 514.94
No. Observations:                 510   AIC:                            -997.9
Df Residuals:                     494   BIC:                            -930.1
Df Model:                          15
Covariance Type:            nonrobust
===============================================================================================
                                  coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
const                           0.3191      0.033      9.628      0.000       0.254       0.384
yr                              0.2327      0.008     28.852      0.000       0.217       0.249
workingday                      0.0520      0.011      4.767      0.000       0.031       0.073
atemp                           0.4409      0.034     12.796      0.000       0.373       0.509
hum                            -0.1370      0.038     -3.630      0.000      -0.211      -0.063
windspeed                      -0.1674      0.026     -6.504      0.000      -0.218      -0.117
Light Rain + Scattered clouds  -0.2577      0.026     -9.743      0.000      -0.310      -0.206
Mist + Broken clouds           -0.0613      0.010     -5.845      0.000      -0.082      -0.041
spring                         -0.1074      0.015     -7.094      0.000      -0.137      -0.078
winter                          0.0688      0.014      4.944      0.000       0.041       0.096
Mon                             0.0625      0.014      4.437      0.000       0.035       0.090
Dec                            -0.0515      0.017     -2.999      0.003      -0.085      -0.018
Jan                            -0.0538      0.018     -2.994      0.003      -0.089      -0.018
Jul                            -0.0639      0.017     -3.749      0.000      -0.097      -0.030
Nov                            -0.0553      0.018     -3.019      0.003      -0.091      -0.019
Sep                             0.0551      0.015      3.578      0.000       0.025       0.085
==============================================================================
Omnibus:                       82.147   Durbin-Watson:                   2.003
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              251.610
Skew:                          -0.751   Prob(JB):                     2.31e-55
Kurtosis:                       6.096   Cond. No.                         20.8
==============================================================================
```

4TH VIF RULES & VALUES

But our rule for removing variable is:

1. Both VIF(5+) and P-value(0.05+) is high:

   - If p-value is high and VIF value is high then remove it.

2. High-low:

   - If p-value is high and VIF value is low then remove it.

   - If p-value is low and VIF value is high then remove it.

3. Both VIF(5<) and P-value(0.05<) is Low:

   - We keep the variable.

	Features	VIF
3	hum	25.65
2	atemp	17.69
1	workingday	4.67
4	windspeed	4.12
8	winter	2.93
7	spring	2.92
6	Mist + Broken clouds	2.26
0	yr	2.07
9	Mon	1.83
13	Nov	1.80
11	Jan	1.76
10	Dec	1.55
12	Jul	1.41
5	Light Rain + Scattered clouds	1.25
14	Sep	1.21

Hum and atemp have high VIF values hence they shall be removed but…

…season, mnth, atemp and yr had high correlation with cnt hence *we will not remove atemp*

After this model two models were followed removing workingday and Monday columns from the dataframe finally

BUILDING A LINEAR MODEL

3. HYPOTHESIS TESTING

HYPOTHESIS

Here, we can see that all the Betha is not 0, hence proved null hypothesis is false

```
2  model.params

const                            0.292594
yr                               0.236070
atemp                            0.410930
windspeed                       -0.143023
Light Rain + Scattered clouds   -0.288052
Mist + Broken clouds            -0.080351
spring                          -0.112230
winter                           0.057974
Dec                             -0.053943
Jan                             -0.057243
Jul                             -0.058775
Nov                             -0.057552
Sep                              0.051825
dtype: float64
```

R-square value:

R-squared: **0.832**

Adj. R-squared: **0.828**

F-Staitsics:

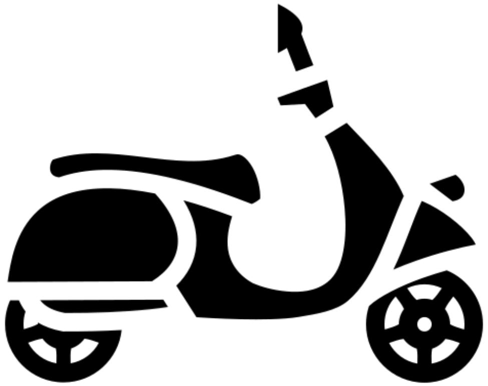
F-Statistics is used for testing the overall significance of the Model. The higher the F-Statistics, the more significant the Model is.

F-statistic: **205.5**

Prob (F-statistic): **6.75e-184**

The equation of best fitted surface based on final model is:

$$cnt = (0.292594 + 0.23607\ x\ yr) + (0.292594 + 0.41093\ x\ atemp) + (0.292594 - 0.143023\ x\ windspeed) + (0.292594 - 0.288052\ x\ Light\ Rain\ +\ Scattered\ clouds) + (0.292594 - 0.080351\ x\ Mist\ +\ Broken\ clouds) + (0.292594 - 0.11223\ x\ spring) + (0.292594 + 0.057974\ x\ winter) + (0.292594 - 0.053943\ x\ Dec) + (0.292594 - 0.057243\ x\ Jan) + (0.292594 - 0.058775\ x\ Jul) + (0.292594 - 0.057552\ x\ Nov) + (0.292594 + 0.051825\ x\ Sep)$$



RESIDUAL ANALYSIS

1. NORMALITY OF ERRORS

NORMALITY OF ERRORS

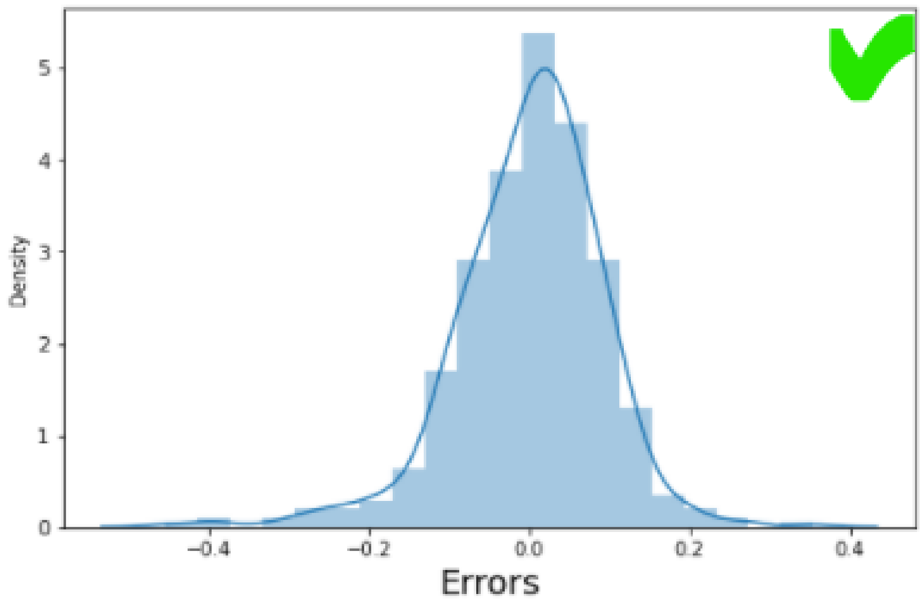
The calculated cnt value is subtracted from the real cnt value to formulize a normality graph using histogram

```
# Predicting the X_train_rfe data to get y_train_predict
y_train_predict = model.predict(X_train_rfe)
```

```
##### NORMALITY OF ERRORS:
```

```
# Plot the histogram of the error terms
fig = plt.figure()
plt.figure(figsize = (8,5))
sns.distplot((y_train - y_train_predict), bins = 20)
fig.suptitle('Error Terms', fontsize = 20)                  # Plot heading
plt.xlabel('Errors', fontsize = 18)                         # X-label
plt.show()
```

```
<Figure size 432x288 with 0 Axes>
```



RESIDUAL ANALYSIS

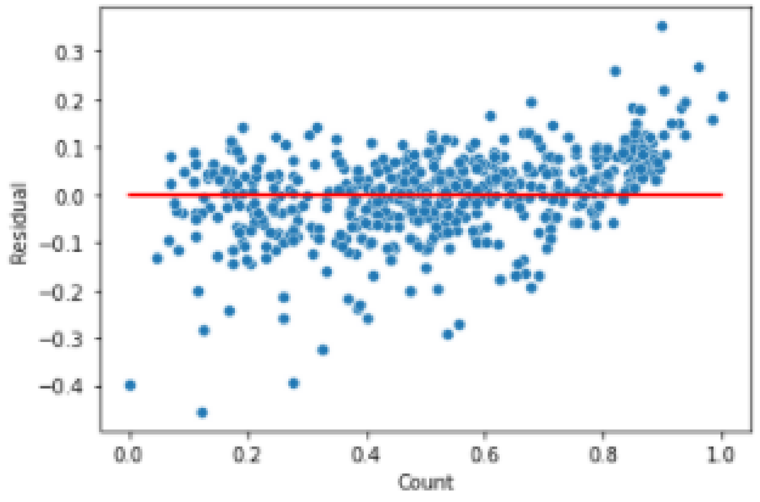
2. DEPENDENCY OF ERRORS

ERROR DEPENDENCY?

The errors are independent from each other as we can observe…

Error are independent of each other

```
residual = y_train - y_train_predict #getting residual
sns.scatterplot(y_train,residual) # ploting y_train vs residual
plt.plot(y_train,(y_train - y_train), '-r') # ploting a stright line on 0th of y-axis
plt.xlabel('Count')
plt.ylabel('Residual')
plt.show()
```



RESIDUAL ANALYSIS

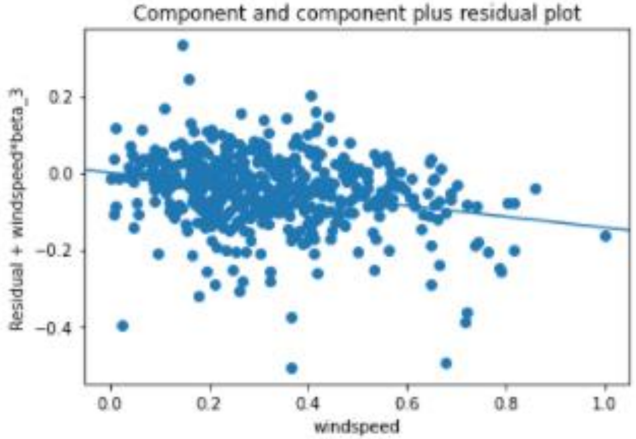
3. LINEAR RELATIONSHIP BETWEEN X AND Y

LINEAR RELATIONSHIP

We'll check the graph with our numerical values…

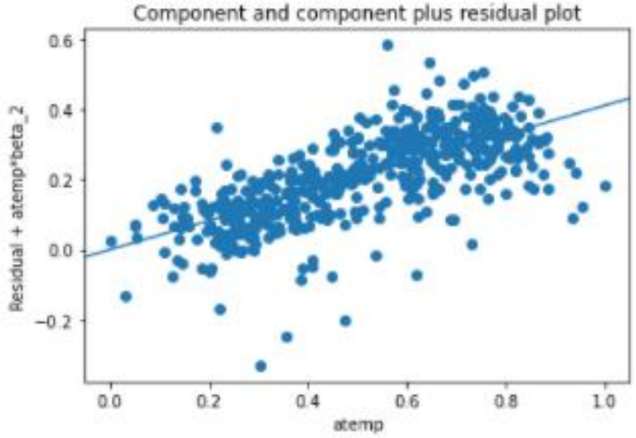
for windspeed

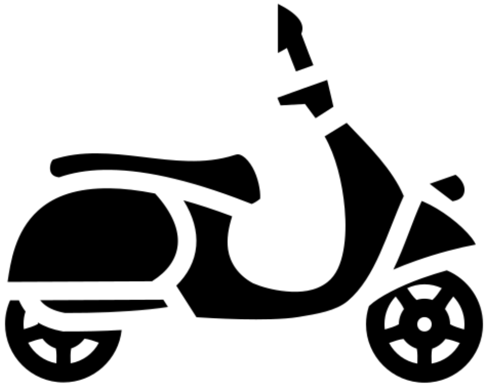
```
sm.graphics.plot_ccpr(model, 'windspeed')
plt.show()
```



for atemp

```
sm.graphics.plot_ccpr(model, 'atemp')
plt.show()
```





MAKING PREDICTIONS USING THE FINAL MODEL

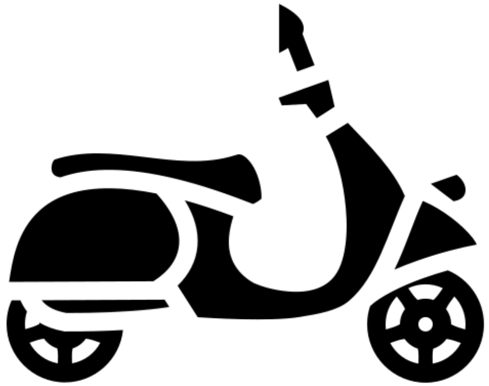
PREDICTIONS VIA TEST

PREDICTIONS VIA TEST

We’ll test the data on the basis of the final columns we finally picked which supported the decision making.

And Predicting X_test using the prevouse model trained using X_train

	const	yr	atemp	windspeed	Light Rain + Scattered clouds	Mist + Broken clouds	spring	winter	Dec	Jan	Jul	Nov	Sep
count	219.0	219.000000	219.000000	219.000000	219.000000	219.000000	219.000000	219.000000	219.000000	219.000000	219.000000	219.000000	219.000000
mean	1.0	0.479452	0.532991	0.313350	0.027397	0.319635	0.255708	0.232877	0.086758	0.077626	0.105023	0.073059	0.086758
std	0.0	0.500722	0.217888	0.159947	0.163612	0.467403	0.437258	0.423633	0.282125	0.268194	0.307285	0.260830	0.282125
min	1.0	0.000000	0.025950	-0.042808	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	1.0	0.000000	0.344751	0.198517	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	1.0	0.000000	0.549198	0.299459	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	1.0	1.000000	0.714132	0.403048	0.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
max	1.0	1.000000	0.980934	0.807474	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000



MODEL EVALUATION

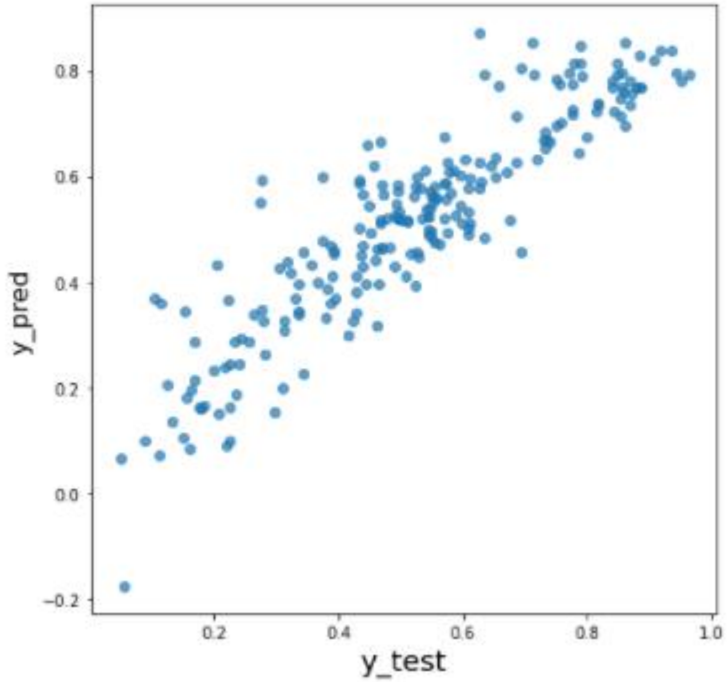
1. REAL V/S PREDICTED

REAL V/S PREDICTED

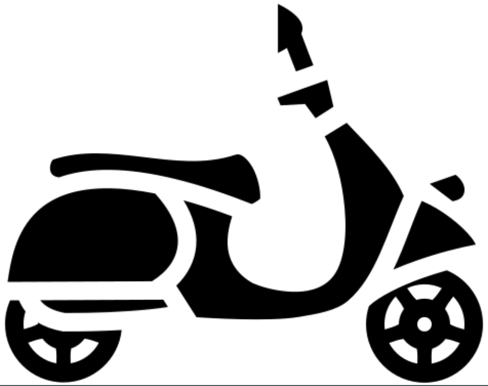
R2 Value Calculation for Test data dataframe: 0.818

And Adjusted R2 Value comes out to be: 0.806

- the graph for actual versus predicted values.



ASSIGNMENT-BASED SUBJECTIVE QUESTIONS



Answers to following questions provided:

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

WE HAVE CATEGORICAL VARIABLES SUCH AS yr(year), holiday, workingday, weathersit, season, weekday, and mnth(month).

"CNT" V/S {}:

1. {"YR"} –

2019 HAD MORE CNT IN RANGE 4500 TO 7000 THAN 2018

2. {"HOLIDAY"}

MORE PEOPLE RODE BIKE ON HOLIDAYS

3. {"WORKING DAY"}

THERE IS A SLIGHT CHANCE THAT PEOPLE WILL BORROW BIKES MORE ON A NONWORKING DAY.

4. {"WEATHERSIT"}

IT IS ALWAYS BEST TO RIDE BIKES IN CLEAR WEATHER FOLLOWED BY MIST AND BROKEN CLOUDS

5. {"SEASON"}

MOST PEOPLE TEND TO RENT BIKES IN THE FALL AND SUMMER.

AND PEOPLE PREFER TO BOOK LESS IN SPRING.

6. {"WEEKDAY"}

MEAN ARE THE SAME FOR ALL WEEKDAYS.

BUT TUESDAY HAS MORE CHANCE OF BOOKING LESS THE OTHER WEEKDAY.

AND MONDAY HAS A HIGH PROBABILITY OF RENTING A BIKE.

7. {"MNTH"}

PEOPLE PREFER TO RENT A BIKE BETWEEN MAY TO OCTOBER.

AND GRADUALLY DECREASE FROM NOVEMBER TO FEB.

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

1. When we create a dummy variable for a categorical variable which have more than two values as n, what it does is that it make n new columns for that categorical variable in the form of 1s and 0s.

2. FOR EXAMPLE: LOW, MEDIUM, AND HIGH.
   L M H
   1 0 0
   0 1 0
   0 0 1

3. It Will Create A Three-variable Low, Medium, And High. But It's Evident That If One Is Medium And The Second Is High, Then The Third Will Be Low. So Dropping The First Column Will Make Low As 00, Medium 10, And High Will Be 01. And Helps To Reduce The Correlations Created Among Dummy Variables.
   M H
   0 0
   1 0
   0 1

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

1. If we talk about the logical variables; The Numerical variables with highest correlation are: atemp and since atemp values were mostly aligned with temp variable so we can include temp variable to the answer too.

	atemp	hum	windspeed	cnt
cnt	0.630685	-0.098543	-0.235132	1.000000

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

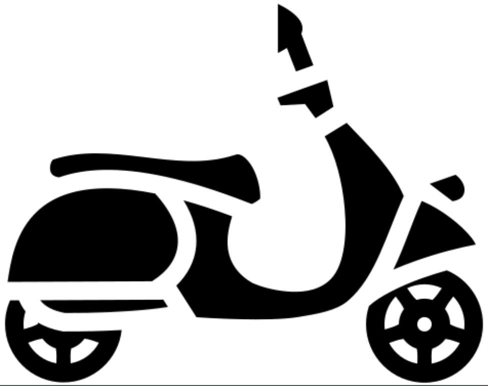
1. For building the model, I use p-value, VIF, R-square, and Adj. R-square.
2. For validation, I check whether the error is normally distributed or not. If it's not, then the p-value we have obtained during the model building will become unreliable.
3. Error are independent of each other cause errors follow no pattern and are independent of each other.
4. There is a linear relationship between X and y, as you may notice from previous slides

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

1. Temperature shows the highest coefficient of around 0.410930, which means when the temp is increased by one unit, the rental bike increases by 0.410930
2. Light Rain + Scattered clouds shows a negative coefficient of around 0.288052; this means when this variable increases by one unit, the rental bike will decrease by 0.288052.
3. yr shows the 2nd highest coefficient of around 0.23607, which means when the yr is increased by one unit, the rental bike increases by 0.23607
4. windspeed shows a second highest negative coefficient of around 0.143023; this means when windspeed variable increases by one unit, the rental bike will decrease by 0.143023.
5. spring shows a third highest negative coefficient of around 0.11223; this means when spring variable increases by one unit, the rental bike will decrease by 0.11223.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

GENERAL SUBJECTIVE QUESTIONS



Answers to following questions provided:

1. Explain the linear regression algorithm in detail. (4 marks)
2. Explain the Anscombe’s quartet in detail. (3 marks)
3. What is Pearson’s R? (3 marks)
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.(3 marks)

1. Linear Regression is a machine learning model which solves a regression task. And it is based on supervised learning. It is used for predictive analysis and shows the relation between a continuous variable. It establishes the relationship between the independent variable (X) and the dependent variable(y).

2. The best fit line is described as:
   y = B0 + B1X + E
   where y is the dependent variable
   x is the independent variable
   B0 is slop
   B1 is the coefficient of X
   E is the error

3. If the independent variable (X) is just one dimension/variable, it's called simple linear regression. And the equation is y = B0 + B1X +E

4. If the independent variable(X) is more than two dimensions/variable, then it is called multiple linear regression. And the equation is
   y = B0 + B1X1 + B2X2+…+BnXn + E

5. To check the best fit line, we have to minimize the error, and to prevent the error; we use the R-square for simple LR and Adj. R-square for multi LR.
   R-square = 1- {Residual Sum of Squares (RSS) / Total Sum of Squares(TSS)}
   n is the sample size
   k is a number of the independent variable.

$$R^2_{adj} = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1}\right]$$

1. Explain the linear regression algorithm in detail. (4 marks)

1. Anscombe’s quartet comprises four nearly identical datasets in simple descriptive statistics like mean, standard deviation, and count. But have a very different distribution and appears very different when plotted on the graph.
2. Francis Anscombe discovered it in 1973 to illustrate the importance of plotting the graphs before analyzing and building the models. And to show the effects of observation on statistical properties and the impact of outliers on statistical properties.So four datasets with 11 data-point have identical statistical statements like mean, variance, correlation, and count.
3. After that, he told the council to analyze using only descriptive statistics. And found out all have almost the same statistics.
4. So this tells that before applying any algorithm or building model, it's essential to visualize it first. And tell us that the data feature must be plotted to see the distribution, which can help us see the outliers, linear separation of the data, diversity of the data, and more. Also, to use Linear Regression, we have to know that the data has a linear relationship.
5. When the dataset is plotted by scatter plot, they realize that some of the datasets cannot be plotted using linear regression, which had fooled them.

2. Explain the Anscombe’s quartet in detail. (3 marks)

1. So this tells that before applying any algorithm or building model, it's essential to visualize it first. And tell us that the data feature must be plotted to see the distribution,
2. which can help us see the outliers, linear separation of the data, diversity of the data, and more. Also, to use Linear Regression, we have to know that the data has a linear relationship.
3. When the dataset is plotted by scatter plot, they realize that some of the datasets cannot be plotted using linear regression, which had fooled them.

3. What is Pearson's R? (3 marks)

1. What is:
   - Scaling is a part of data preprocessing applied on the dataset to normalized in a particular range, and it helps to make the calculation faster for the algorithm.
2. Why is:
   - Dataset is a collection of numbers in general, and it comes with different units, magnitude, and range. So higher/larger the number, the more superiority it has in some sort in the model building. Because the machine learning algorithm works on the number and doesn't know what represents then numbers if the scale is not done, it takes magnitude, not the unit, so it is incorrect.
   - To solve this issue, we used scaling to bring variables to the same level. However, scaling affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.
3. Normalized scaling/ Min-Max scaling:
   - It brings all the data in the range of 1 and 0. And it is sensitive to outliers.
   - sklearn.preprocessing.MinMaxScaler is used to apply Min-Max scaling.

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

4. Standardized scaling:
   - It brings all the data into its standard normal distribution center around its mean(0) and one standard deviation.
   - If the data is not normally distributed, then it's not a good scale to use.
   - sklearn.preprocessing.scale is used to implement standardization

$$x_{new} = \frac{x - \mu}{\sigma}$$

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

As we can see from the formula, R2 is the only variable that affects VIF, and if R2=1, it means it's perfectly correlated, and it causes the VIF to be infinity. So Basically, it means the two variables have a perfect correlation. If the VIF is high, then it indicates that there is a correlation. If the VIF=5, the variance of the model coefficient is inflated by a factor of five due to the presence of multicollinearity.

If the VIF is greater than ten, then there is multicollinearity. And between 5-10 is moderate, less than five will be less chance, and one will be no multicollinearity.

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

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Quantile-Quantile plot is a graphical tool to help us evaluate if the one row is plausibly came from some theoretical distribution such as a Normal, exponential, or Uniform distribution. It is used to compare the shape of the distribution by checking the graph plots properties such as scale,
distribution, location, and skewness the same or different from the two distributions.

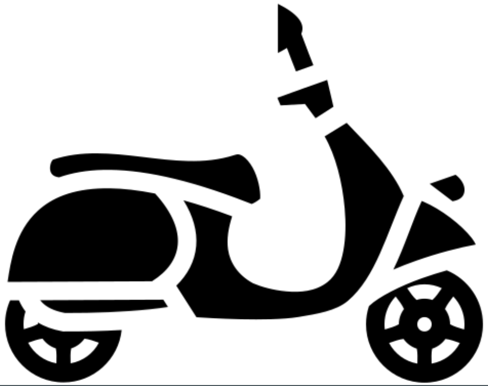
If all point of quantiles lies on or close to a straight line at an angle of 45 degrees from x -axis, then it means the distribution is similar. If all point of quantiles lies away from the straight line at an angle of 45 degrees from the x-axis, the distribution is different.

This can be used to check whether:

- The training and test dataset is from the same distribution or not.
- The dataset follows theoretical distribution such as a Normal, exponential, or Uniform distribution

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.(3 marks)

CONCLUSION



If the temperature feel is highly correlated and hiving a high coefficient with cnt, but we know from EDA that if the temperature goes higher than 35 , we see fewer bikes rented. So they should know that when the temperature feels like between 20-35, there will be more people renting bikes.

If the weather is like Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds, then people tend to rent bikes less.

The windspeed shows a negative coefficient with cnt, but we need to know when wind speed is between 5-20 then people rent more bikes through EDA, but we have to know that wind speed above will always affect.

spring shows a negative coefficient with cnt, so we should offer more in spring.

The business has increased from 2018 to 2019 so that's why it show high coefficient with cnt

R-square value: a. For train:

- R-squared: 0.832
- Adj. R-squared: 0.828

b. For test:

- R-squared: 0.818
- Adj. R-squared: 0.806

c. Difference between train and test R-square value is 0.832-0.818 = 0.014 (1.4%)

d. Difference between train and test Adj. R-squared value is 0.828-0.806 = 0.022 (2.2%)

The equation of best fitted surface based on final model is:¶

$$cnt = (0.292594 + 0.23607\ x\ yr) + (0.292594 + 0.41093\ x\ atemp) + (0.292594 - 0.143023\ x\ windspeed) + (0.292594 - 0.288052\ x\ Light\_Rain\_ + Scattered\_clouds) + (0.292594 - 0.080351\ x\ Mist + \_Broken\_clouds) + (0.292594 - 0.11223\ x\ spring) + (0.292594 + 0.057974\ x\ winter) + (0.292594 - 0.053943\ x\ Dec) + (0.292594 - 0.057243\ x\ Jan) + (0.292594 - 0.058775\ x\ Jul) + (0.292594 - 0.057552\ x\ Nov) + (0.292594 + 0.051825\ x\ Sep)$$

THANK YOU ☺