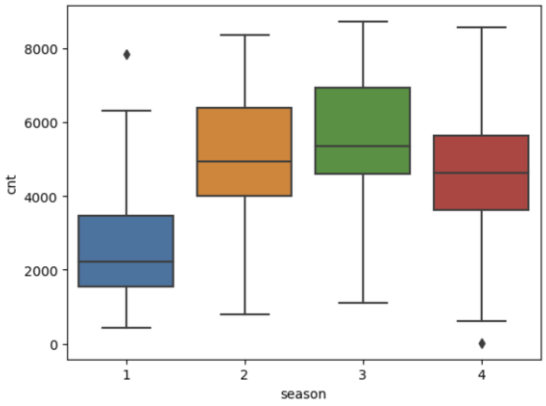
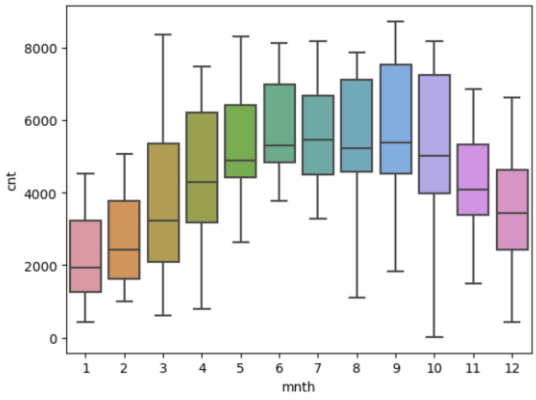
Name - Neelam Mahapatro

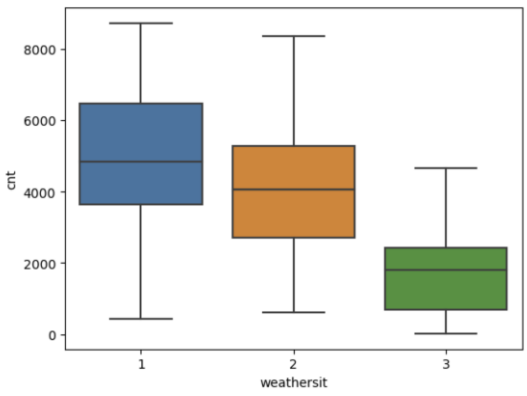
Batch - C46

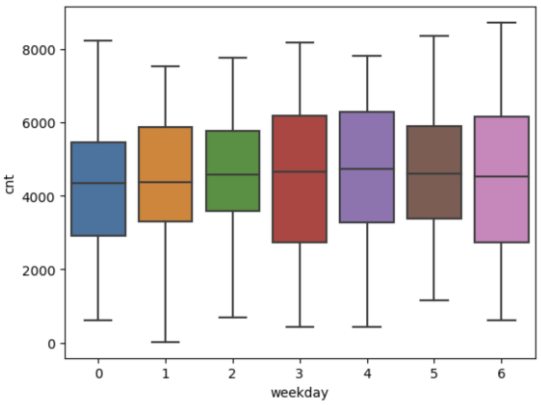
Assignment-based Subjective Questions

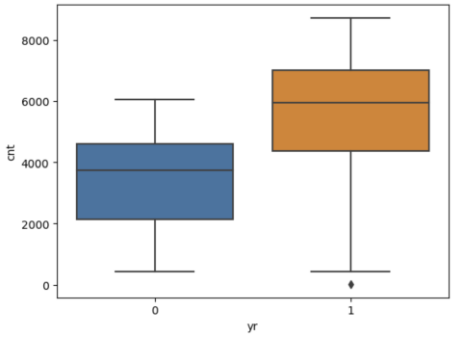
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
   Ans – The boxplot results for different categorical variables obtained during the analysis are as below.











   a. Season , year, month and weathersit categorical variables have significant impacts on the bike hire counts.

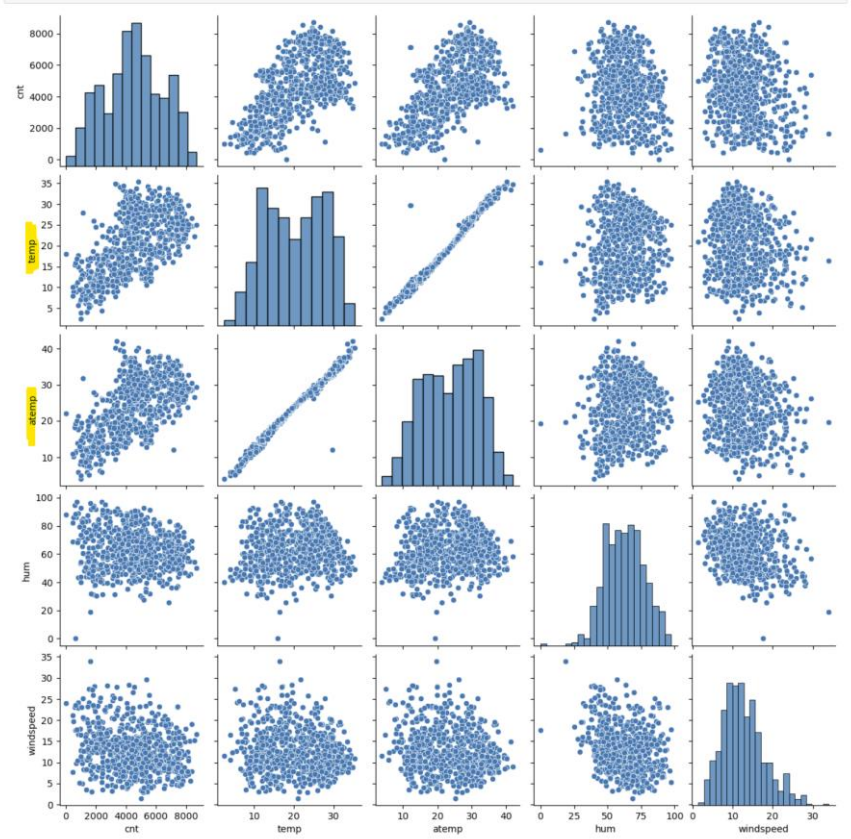
Example – Year 2019 has much bike hire counts then 2018 year.
Spring season has much less bike hire counts than any other season

b. However, variables like weekday doesnot have any impact on the output variables.

2. Why is it important to use drop_first=True during dummy variable creation?
Ans - If a category variable has N levels then during dummy variable creation it creates N number of variables. But with respect to multi-collinearity, one variable will be redundant i.e when other variables are zero, this will be one, When any one of other variable it 1, this is 0. **Hence,** it is essential to use drop_first to True during creation of dummy variables and p-value will also not get affected by multicollinearity.

Example – Year variable – As per the data set whether it is 2018 or 2019 based on the based on 0 and 1 value respectively. If we create dummy variable on this category, it will create 2 variables (is_2018 and is_2019). Here, when is_2018 becomes 1, is_2019 automatically becomes 0. Hence, we can get rid of is_2018, only one variable (is_2019) is sufficient and required.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Ans – By pair-plot among numerical variables, temp and atemp has the highest correlation with the bike rent counts.

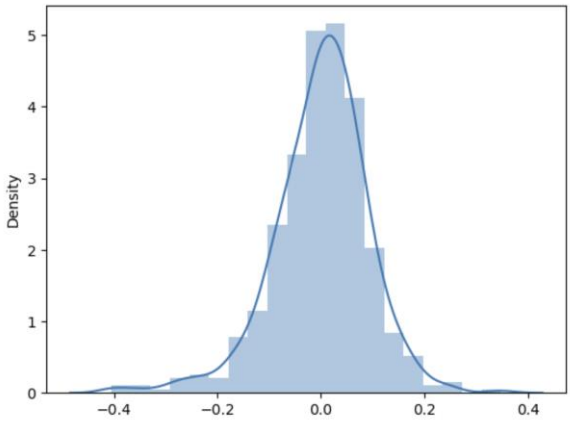


4. How did you validate the assumptions of Linear Regression after building the model on the training set?

   Ans –

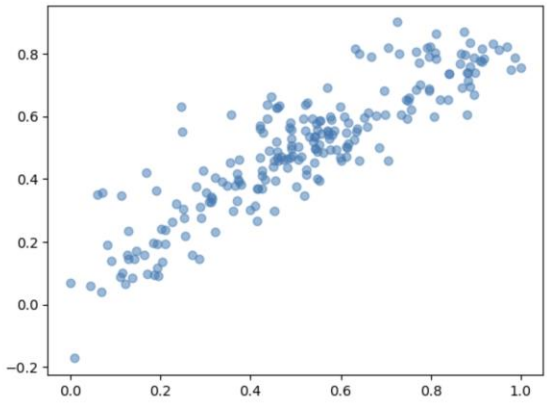
   a. By plotting Error Terms –

      - Error Terms has the normal distribution
      - Error Terms has the mean of 0
      - It has a constant variance



   b. Plot between y_test_pred and y_test

      - This has a linear pattern and y_test_pred nearly equal to y _test which is expected



   c. R2 and Adjusted R2 score for Train and Test data set

      - R2 score is higher and hence the model is good
      - R2 score of Train set is higher then that of Test set

- Train R^2 :0.830
- Train Adjusted R^2 :0.827
- Test R^2 :0.7824
- Test Adjusted R^2 :0.7731

Summary of the Linear Regression –

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.830
Model:	OLS	Adj. R-squared:	0.827
Method:	Least Squares	F-statistic:	270.8
Date:	Sun, 08 Jan 2023	Prob (F-statistic):	6.57e-186
Time:	22:04:24	Log-Likelihood:	490.01
No. Observations:	510	AIC:	-960.0
Df Residuals:	500	BIC:	-917.7
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1264	0.017	7.511	0.000	0.093	0.160
yr	0.2328	0.008	27.886	0.000	0.216	0.249
temp	0.5211	0.022	23.460	0.000	0.477	0.565
windspeed	-0.1516	0.025	-5.954	0.000	-0.202	-0.102
season_2	0.1016	0.011	9.038	0.000	0.080	0.124
season_4	0.1374	0.011	12.758	0.000	0.116	0.159
weathersit_2	-0.0809	0.009	-9.069	0.000	-0.098	-0.063
weathersit_3	-0.2786	0.025	-11.109	0.000	-0.328	-0.229
mnth_8	0.0557	0.017	3.304	0.001	0.023	0.089
mnth_9	0.1133	0.017	6.791	0.000	0.081	0.146

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   Ans – Based on the Final model, the top 3 features contributing significantly for demand of shared bikes are –
   a. Year : 0.2328
   b. Temperate : 0.5211
   c. Weathersit_3 : -0.2786

   The summary of the model is also attached above.

General Subjective Questions

1. Explain the linear regression algorithm in detail.
   Ans - In Linear regression model, we try to predict the output as a linear combination of input predictor while minimizing the cost functions. Here we assume certain things to be satisfied.
   - Linear dependency between output and each input variable
   - Homoscedaticity. (variance is constant)
   - Error term is normally distributed with mean zero
   - No multi-collinearity between input variables

   To train the model, we try to minimize the cost function. It can be done in many ways.
   a. By differentiation method
   b. By gradient descent method

c. R-square term
d. F-statistics

In general, linear regression is divided into 2 types.

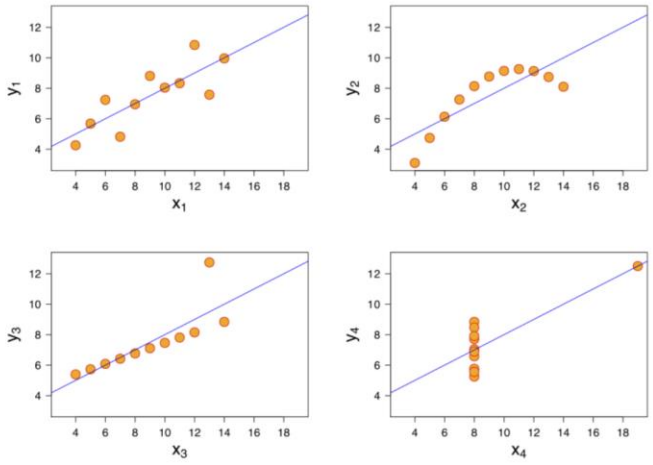
i. Simple linear regression - with only one independent Variable

ii. Multiple linear regression - with multiple independent Variable

2. Explain the Anscombe's quartet in detail.

Ans - Anscombe's quartet tells that sometimes statistical summary is not sufficient to infer the relationship. We have to visualize the distribution of data instead of relying on the data

Below given exaplme consists of 4 datasets that have same simple statistical properties like mean, SD, Correlation, , but are completely different from one another in distribution. Every dataset consists of eleven (x,y) points. These were constructed by the statistician Francis Anscombe in 1973 to showcase that both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties.



I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

All above data points have almost identical X mean, Y mean, Standard deviation of x, Correlation coefficient between X and y, Fitted Linear regression line.
However, all of them have different distribution of points.

3. What is Pearson's R?

Ans - It measures linear relationship strength between 2 variables.
It is also called as correlation coefficient between 2 variables.
The expression can be represented as –

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}$$

X , Y are the two variables for which we want to count the correlation coefficient. The value of this lies between -1 and 1.
- If the value is positive, it means X and Y have a positive correlation which means if one increases other will increase and if one will decrease, other will decrease.
- If the value is negative, , it means X and Y have a negative correlation which means if one increases other will decrease and vice versa.
- If the value is 0, it means X and Y are not having any correlation. They are independent on each other

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans - Scaling is a technique to bring a wide range of values into a particular range.

• In Linear Regression, scaling is performed for 2 purposes. First is it is better to compare the effect 2 input predictor on output variable. Second the gradient descent algorithm can converge to solution faster if inputs are in same scale.

• Without Scaling also, our model will have same goodness of prediction. Scaling just gives us a better way to look and feel the impact of each input variable on output. There are 2 types of popular scaling technique used.

• Normalisation or Min-Max scaling

• Standardisation In Normalisation, we map the value to between 0 and 1 using below relation Normalized_value= $value–min$ max – min

In standardisation, we make the mean of the dataset 0 and standard deviation as 1. Standardized_value= $value–mea$

4. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

VIF = 1 1–Ri^2 Where Ri = good ness of fit of one input variable in terms of other input variable.

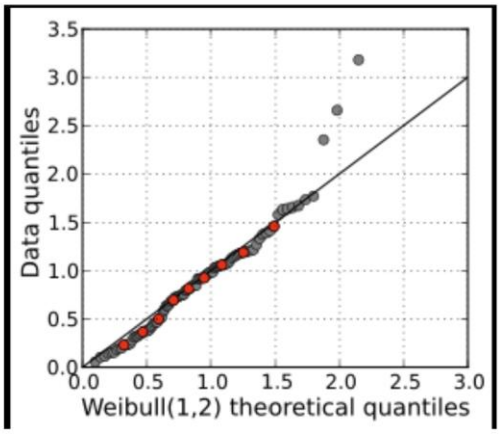
- If VIF=1, it tells zero correlation between one input variable and all other input variables in the model.
- If VIF=1 to 5, it tells some correlation between one input variable and all other input variables in the model. But it not very serious.
- If it is greater than 5 indicates severe correlation between one input variable and all other input variables in the model. In this case, the coefficient estimates and p-values in the regression output are likely unreliable.
- If Ri^2=1 i.e. (Ri= 1 or-1) the input variable is fully linearly dependent on all other variables, VIF will be infinite. We must drop that variable in our prediction.

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

Ans - This Q-Q plot is used to determine if 2 datasets are derived from the same population or not.

UseCase of Q-Q plot Linear Regression- When we receive train and test dataset separately for a linear regression model, we can verify that both dataset have come from same population using this technique.

Example of QQ Plot –



While building machine learning model, we need to check the distribution of the error terms or prediction error using a Q-Q plot. If there is a significant deviation from the mean, we might have to check the distribution of the feature variables and consider transforming them into a normal shape.