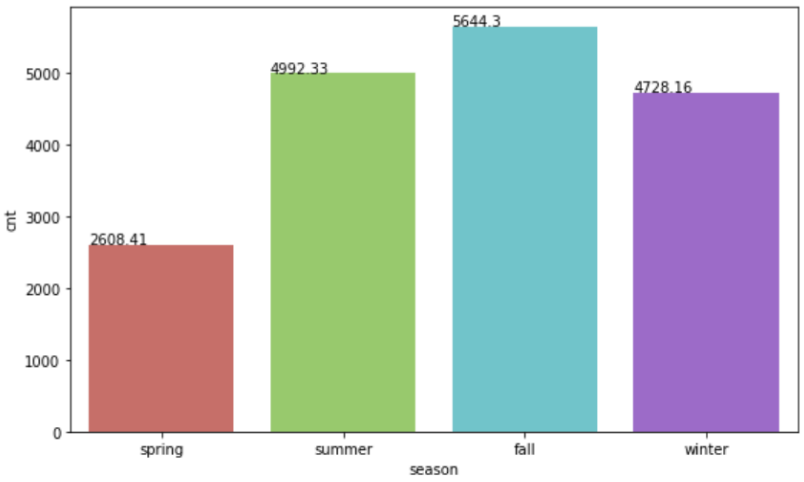
Assignment-based Subjective Questions

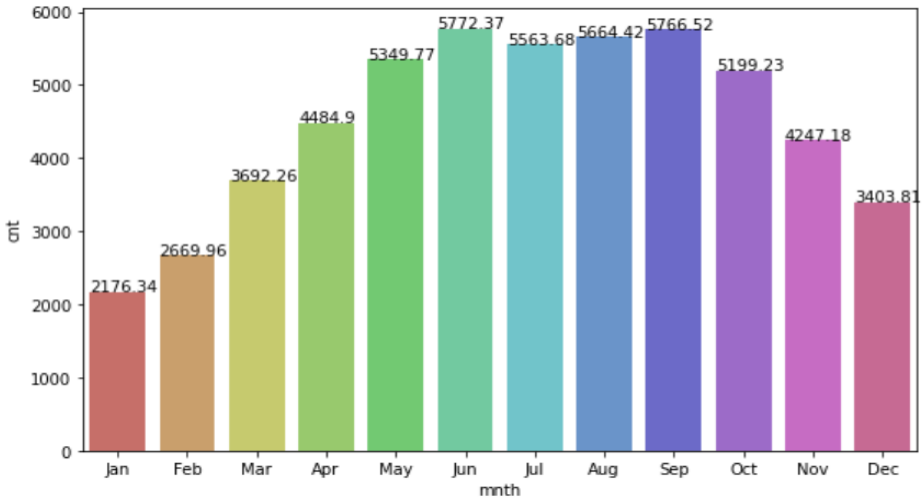
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

   Ans: There were 4 categorical variables in the dataset. The Bar plot is used to study their effect on the dependent variable (‘cnt’) .

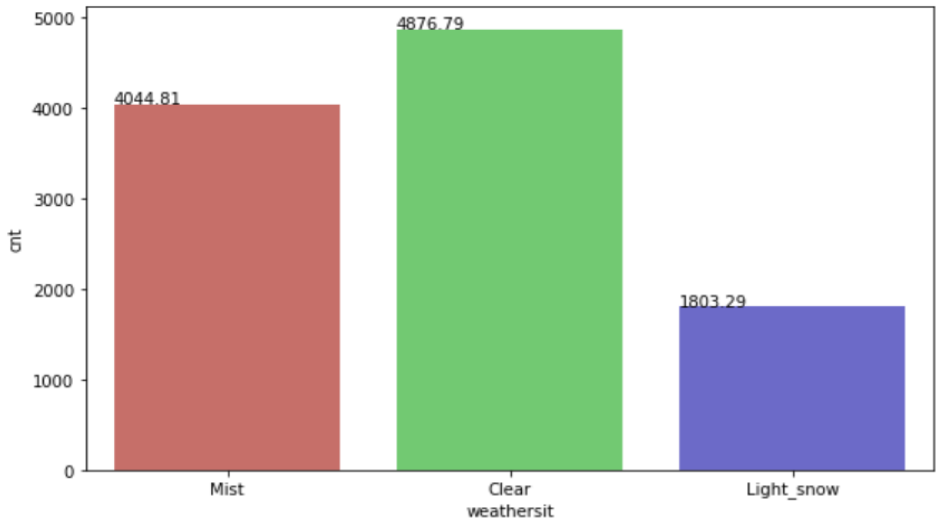
   season: Bike booking were happening in 'fall' season followed by 'summer' & 'winter'. This indicates, season can be a good predictor for the dependent variable.



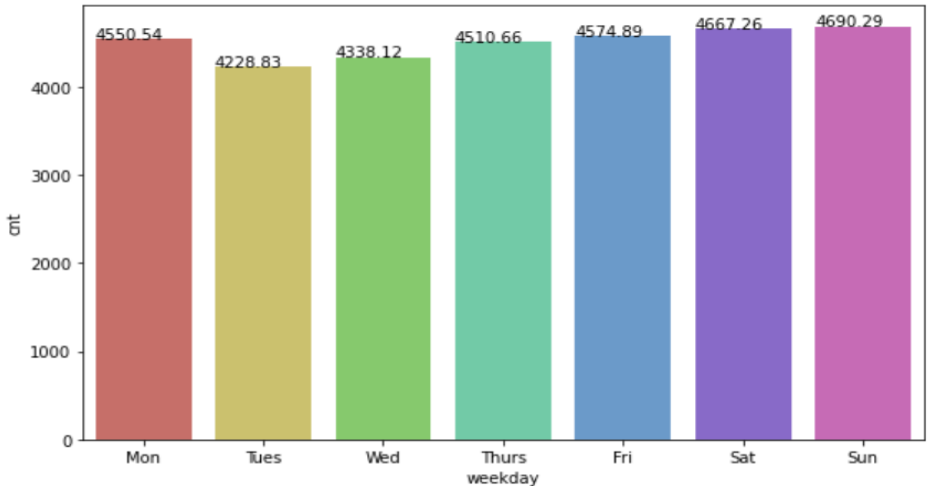
   mnth: Maximum Bike booking happened over the months from May to September with a peak in June. This indicates, mnth has some trend for bookings and can be a good predictor for the depend



   weathersit: Over 45% of the bike booking were happening during Clear weather followed by . This was followed by Mist weather with more than 37%. This indicates that weathersit shows some trend towards bike bookings, which can be a good predictor for the dependent variable.



weekday: It's not clear if weekday variable a good predictor for the target variable. But, Saturday and Sunday most of the people avail of bike rental services.



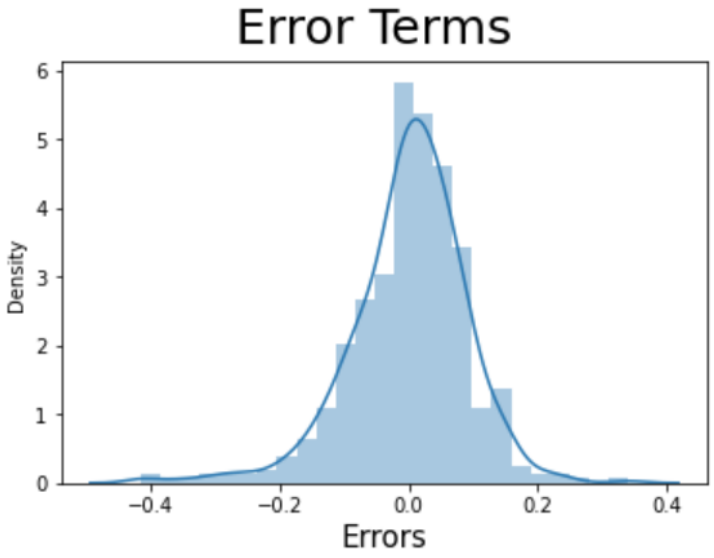
2. Why is it important to use **drop_first=True** during dummy variable creation?

   Ans: drop_first=True we use after dummy variable being created, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   Ans: Variable 'temp' has the highest correlation with the target variable 'cnt'.

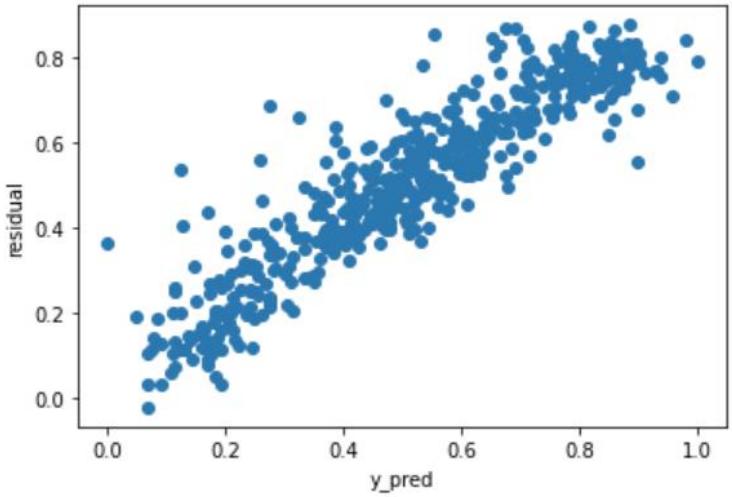
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

   a) Error is normally distributed with mean value 0.



b) Linear Relationship:

Linearity between residuals and y-predicted values



c) Multi-collinearity:

There are no variables which have high VIF values in the model which means no multi-collinearity.

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

Ans: As per the final model, if we consider the co-efficient of variables then temp, yr and Light_snow are contributing more.

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Ans: Machine learning models are of two categories namely Supervised learning and Unsupervised learning. The regression model is a target prediction value based on the independent variable(s).

Mathematically, we can express the regression equation as Y=mX+C

Where **Y** = Dependent Variable
X = Independent Variable
m = Slope
C = Intercept

Once we find the best m and c values, we get the best fit line. So, when finally we are using our model for prediction, it will predict the value of Y for the input value of X.

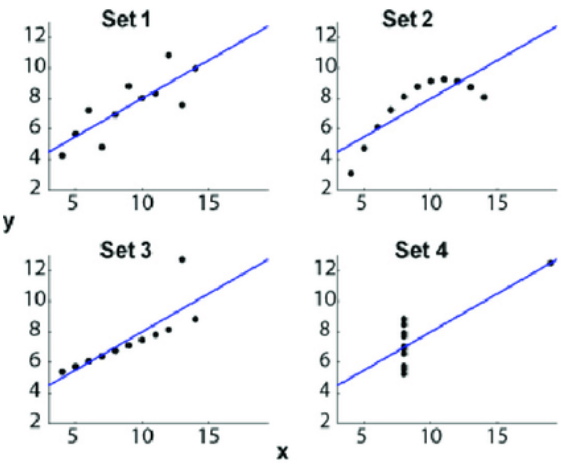
Cost Function(J)

By achieving the best-fit regression line, the model aims to predict the Y value such that the error difference between the predicted value and the true value of the Y point is minimum. So, it is very important to find the optimum values for C and m, to reach the best value that minimizes the error between the predicted y value (pred) and true y value (y).

Cost function(J) of Linear Regression is the Root Mean Squared Error (RMSE) between predicted Y value (pred) and true Y value (Y). To update c and m values in order to reduce Cost function (minimizing RMSE value). The idea is to start with random c and m values and then iteratively updating the values, reaching minimum cost.

Q2. Explain the Anscombers quartet in detail.

Ans: Anscomber's quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built.



Mean of Y	**7.50 in all 4 XY plots**
Sample variance of Y	**4.122 or 4.127 in all 4 XY plots**
Correlation (r)	**0.816 in all 4 XY plots**
Linear regression	**y = 3.00 + (0.500 x) in all 4 XY plots**

Data sets for the 4 XY plots

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

Q3. What is Pearson's R?

Ans: The correlation coefficient in statistics is also known as Pearson's R measures the strength of the linear relationship between two variables. Pearson's R is always between -1 and 1.r = 1 depicts positive linear relationship i.e., as the values in X axis increase the values in V axis increases. r = -1 depicts negative linear relationship i.e. as the values in X axis increases the values in Y axis decreases. r = 0

depicts that there is no correlation between two variables.

The formula to calculate the r is mentioned below.

r = covariance / (standard deviation of one variable i.e. X standard deviation of one variable i.e. Y)

Example 1:

If there is a situation where the lady has to pick strawberries from the garden and the owner has to pay DKK 1 per strawberry then

A. correlation coefficient between the number of strawberries the lady picks and the earnings for that one day will be 1.

B. correlation coefficient between the number of strawberries the lady picks and the amount of money the owner has for that one day will be -1.

Example 2:

In another scenario, An HR executive of a company is asked to create a table in MS excel with two columns A and B wherein Column A represents age and column B earnings. The earnings increase steadily with age. So, in this case, the correlation coefficient between age and earnings will be 1.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans: Scaling or feature scaling is converting all variables in the same scale. It is a way of normalizing the range of independent variables or features in the available data set.

Example: If we have different weights ranging from 10 lb to 100 lb, then instead of representing the weight in lb we may choose to represent the values between 0 and 1, where 0 is the lowest weight and 1 is the highest weight available.

Gradient descent will struggle to converge without proper scaling. Cluster algorithm will be skewed by differing scales as they rely on distance metrics.

Example :
If we have a data set which has features like height and weight and we don't scale it and try to plot the data. Then the plotting will be wide spread and the distance will be huge, which will be a wrong indicative to our analysis.

In normalized scaling the range remains between 0 and 1.
Formula-> x (normalized value) = [x-min(x)] / [max(x) - min(x)]

In standardized scaling, the values are represented in standard deviation from the mean, likely in the range between -3σ to +3σ.

Formula-> x (normalized value) =[x — average (x)] / standard deviation

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans: VIF stands for Variance Inflation Factor. If VIF = ∞(Infinity) then there is a perfect correlation between two independent variables. i.e if $R^2 = 1$, $VIF = 1/(1-R^2) = \infty$

Where R^2 is the coefficient of determination in linear regression. R^2 value lies between 0 and 1.

A general rule of thumb is that if

- VIF >= 10, then there is high multi-collinearity between the predictor variables.
- 5 <= VIF >= 10, then there is moderate collinearity between predictor variables. But, in some cases we might choose to go with moderately high VIF values if it does not affect the model results.
- VIF<= 5, then there is mild collinearity between predictor variables and this is very much acceptable.
- VIF=1, No Multi-collinearity

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans: Q—Q (quantile-quartile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.

Use and importance of a Q-Q plot in linear regression:

To check homoscedasticity and heteroscedasticity error terms after building the training model, we use Q-C), plots. Below example is the generic Q-Q plot in linear regression.