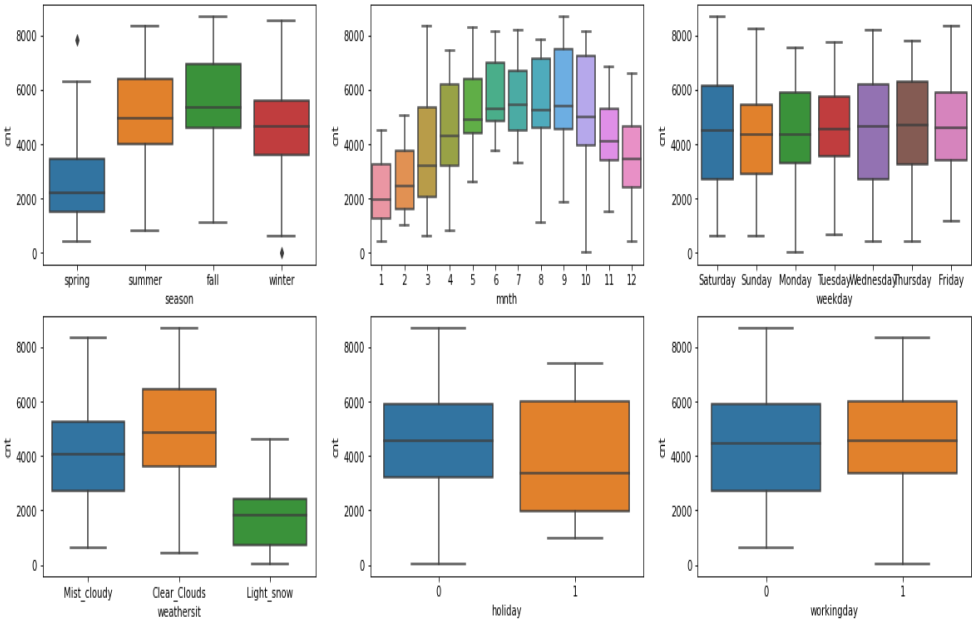
Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

A1. I used Boxplot to study the effect on the dependent variable ('cnt'). The inference is as below:

a. **season: Almost 26% of the bike booking were happening in fall season with a median of over 5000 booking (for the period of 2 years). This was followed by summer & winter with 25% & 24% of total booking. This indicates, season can be a good predictor for the dependent variable.**
b. **mnth: Almost 10% of the bike booking were happening in the months 5,6,7,8 & 9 with a median of over 4000 booking per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.**
c. **weathersit: Almost 67% of the bike booking were happening during "weathersit1: Clear, Few clouds, Partly cloudy, Partly cloudy" with a median of close to 5000 booking (for the period of 2 years). This was followed by "weathersit2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist" with 30% of total booking. This indicates, 'weathersit' does show some trend towards the bike bookings and can be a good predictor for the dependent variable.**
d. **holiday: Almost 97.6% of the bike booking were happening when it is not a holiday which means this data is clearly biased. This indicates, holiday CANNOT be a good predictor for the dependent variable.**
e. **weekday: weekday variable shows very close trend (between 13.5%-14.8% of total booking on all days of the week) having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor. I will let the model decide if this needs to be added or not.**
f. **workingday: Almost 69% of the bike booking were happening in 'workingday' with a median of close to 5000 booking (for the period of 2 years). This indicates, workingday can be a good predictor for the dependent variable**

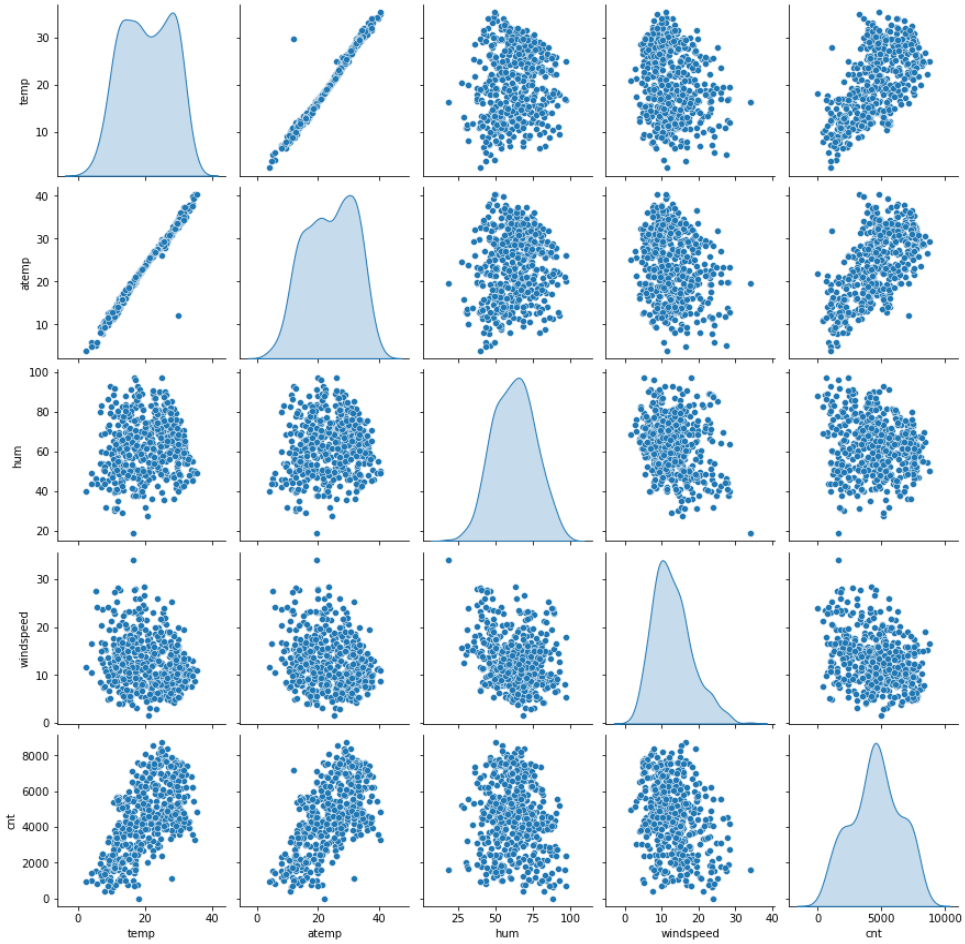


Q2. Why is it important to use **drop_first=True** during dummy variable creation?

A2. drop_first=True is used to drop the first dummy variable, created for each set of dummies. It is important as it helps in reducing the extra set of columns created during dummy variable creation. It reduces the correlation created among dummy variables.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

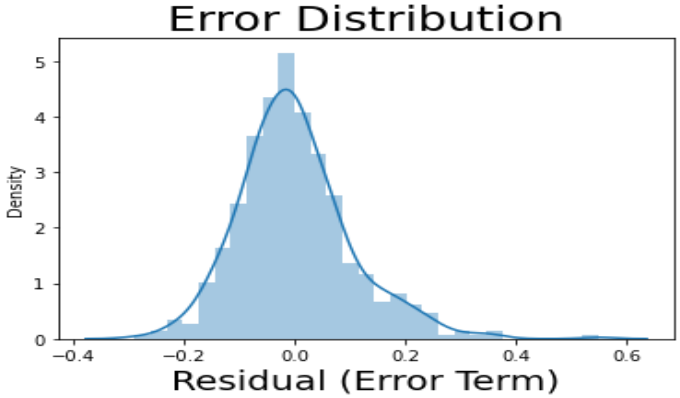
A3. Numerical variables ("temp", "atemp", "hum", "windspeed', "cnt") & Target Variable is 'cnt'. The highest correlation with "cnt" is "temp" & "atemp".



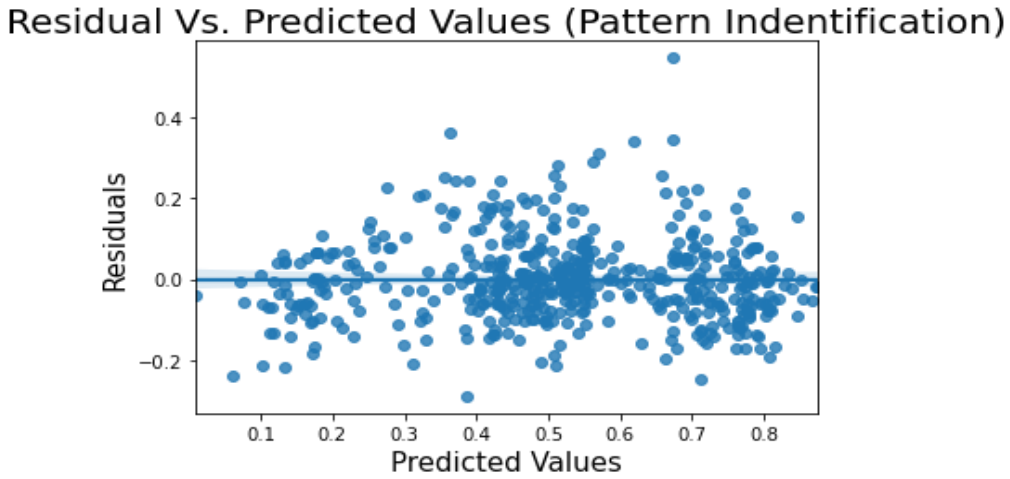
Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

A4. For validating the assumptions of linear regression after building the model on the training set, we use three below methods to check that there are no more significant variables left to add significant to the outcome, thus we check randomness of error (between predicted and actual values.

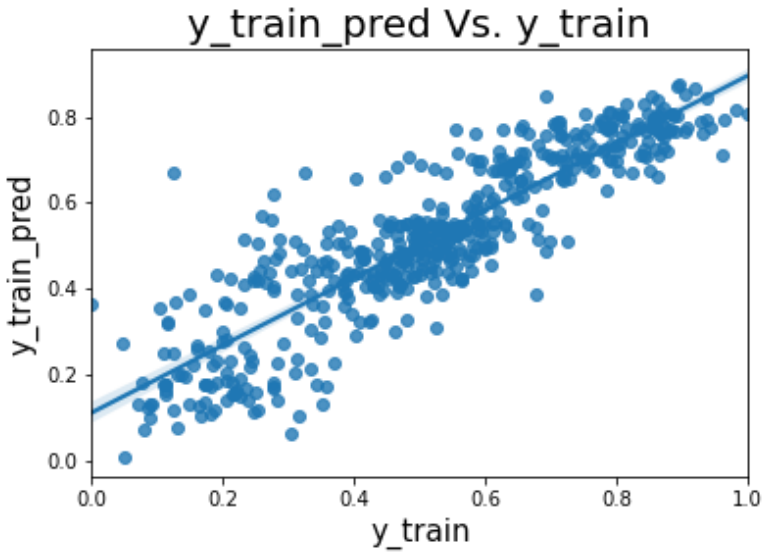
a. **Normal Distributed Error Terms (to check whether the residual analysis on predicted and trained variables are distributed near Zero)**



b. **Residual Error terms for being Independent (to check whether the errors are distributed normally near Zero or we can say that there is no relation between residual and predicted values)**



c. **Homoscedasticity (to check that the residuals are equal across the regression line, constant variance)**



Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

A5. The top 3 features contributing towards explaining the demand of the shared bikes.

a. **"yr"- A coefficient value of '0.2362' indicated that a unit increase in 'yr' variable increases the bike hire numbers by '0.2362' units.**
b. **"mnth_Sep"- A coefficient value of '0.0718' indicated that a unit increase in 'mnth_Sep' variable increases the bike hire numbers by '0.0718' units.**
c. **"weathersit_Light_snow"- A coefficient value of '-0.3281' indicates that a unit increase in "weathersit_Light_snow" variable decreases the bike hire by '-0.3281' units.**

General Subjective Questions

Q1. Explain the linear regression algorithm in detail?

A1. Linear Regression is a machine learning algorithm based on supervised learning and is one of the very basic forms of ML where we train model to predict the behaviour of the data based on some variables. It performs a regression task to compute the regression coefficients. Regression models a target prediction based on the independent variables.

Linear Regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). A linear relationship between x (input) and y (output) variables occur. The linear equation for univariate linear regression is

$$Y = \beta_0 + \beta_1 X$$

β_0= intercept/model coefficient
β_1= coefficient for x/model coefficient
Y = output/dependent/target
X= input/independent/predictor

Simple Linear Regression: When there is only single independent/feature variable (X)
Multiple Linear Regression: When there are multiple independent/feature variables (Xi)

$Y = \beta_0 + \beta_1 X$ (SLR)

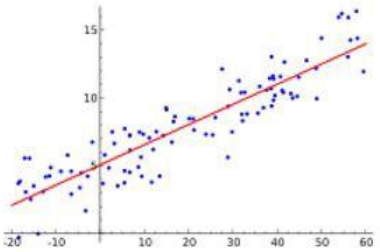
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon$ (MLR)

Where:

Y = how far up ↑ and X = how far along →

$\beta_1, \beta_2 .. B_p$ = Slope or Gradient (how steep the line is)

β_0 = value of Y when X=0 (Y-intercept)



Assumptions for Linear Regression algorithm:

1. **Target variables and input variables are linearly dependent.**
2. **Error terms are normally distributed.**
3. **Error terms are independent of each other.**
4. **Error term have constant variance (Homoscedasticity).**

Libraries used are:
statsmodels allows users to fit statistical models by importing OLS and sklearn are the two machine learning libraries.

The advantage and disadvantage of linear regression algorithm:

1. **Linear regression provides a powerful statistical method to find the relationship between variables. It hardly needs further tuning. However, it's only limited to linear relationships.**
2. **Linear regression produces the best predictive accuracy for linear relationship whereas its little sensitive to outliers and only looks at the mean of the dependent variable.**

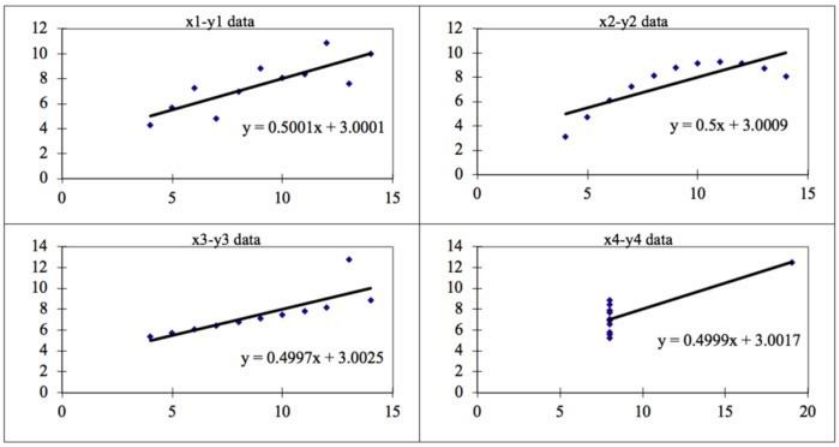
Q2. Explain the Anscombe's quartet in detail?

A2. Anscombe's quartet can be defined as a group of four datasets which are identical in simple descriptive statistics, but there is distinct issue in the dataset that fools the regression model when built. They have very different distributions and appear differently when plotted on the scatter plot. Anscombe's quartet is a classic example of the drawback to just reporting correlation.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analysing and model building, and the effect of other observations on statistical

properties. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x, y points in all four datasets.

Anscombe's Data										
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	



The four datasets can be described as:

- **Dataset 1: this fits the linear regression model pretty well.**
- **Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.**
- **Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model.**
- **Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model.**

Application:

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyse according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets. Hence, all the important features in the dataset must be visualised before implementing any machine learning algorithm on them which will help to make a good fit model.

Q3. What is Pearson's R?

Pearson's R or Pearson correlation coefficient (PCC) or the bivariate correlation is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviation, thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation.

The Pearson coefficient correlation has a high statistical significance. It looks at the relationship between two variables. It seeks to draw a line through the data of two variables to show their relationship. This linear relationship can be positive or negative.

For example:

Positive linear relationship: In most cases, universally, the income of a person increases as his/her age increases.

Negative linear relationship: If the vehicle increases its speed, the time taken to travel decreases, and vice versa.

The Pearson correlation coefficient is symmetric: corr(X,Y) = corr(Y,X).

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

A4. Scaling: is a step of pre-processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

There are 2 types of scaling in liner regression.

1. **Normalized (Min-max Scaling)**
2. **Standardized Scaling**

Why scaling is performed:

Due to high variation in the magnitudes, units and range of feature variables. The model will behave inaccurate due to this variation. If Scaling is not performed, then algorithm only consider magnitude in account and not units, hence it will lead to inaccurate modelling. To solve this, we have to perform scaling to bring all the variables to the same level of magnitude. In short,

a. **Helps in interpretation,**
b. **Faster convergence of gradient descent.**
c. **p-value, model accuracy, R2, F-statistics, T-statistics will not change**
d. **Only coefficients get changed**

Normalized (Min-Max Scaling) Vs Standardized:

Normalized (Min-Max Scaling)

1. **It brings all of the data in the range of 0 and 1**
2. **sklearn.preprocessing.MinMaxScaler helps in implement normalization in python.**
3. **MinMaxScaler take care of all the outliers during implementation.**

$$x = \frac{x - \min(x))}{\max(x) - \min(x)}$$

Standardized Scaling

1. **This replaces the values by their Z scores. It brings all of the data into standard normal distribution which has mean (μ) zero and standard deviation one (sigma)**
2. **sklearn.preprocessing.scale helps in implement standardization in python.**

3. **It doesn't compress the data between the particular range as Min-max scaling. This is useful, if there are extreme data outliers.**

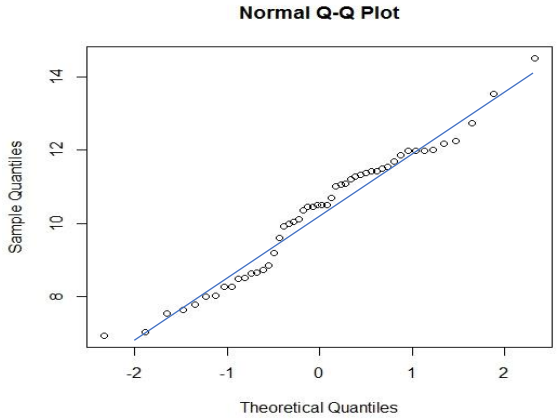
$$x = \frac{x - \text{mean}(x))}{\text{sd}\ (x)}$$

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
A5. When the value of R^2 is equals to 1, which means there is a serious/perfect correlation between two independent variables, this leads to the value of VIF=1/(1-R^2) to becomes infinite. This is also one of the reasons of Multicollinearity and to solve this issue we have to drop one of the variables from the dataset.
The correlated variable will also have infinite R^2 due to the correlation with other variables, same can be corrected after dropping one of the highly correlated variables.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression?
A6. Q-Q plot (Quantile-Quantile Plots) is a kind of graphical plot. The data included in this plotting is of two sets of quantiles and are plotted against each other (x-y axis)

Examples: Q-Q Plot comparing the distribution of standardised daily maximum temperature at 2 states in India for two different months (May/June).

This helps in linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions. If both sets of quantiles comes from the same distribution, we suppose to get a normal straight line along the points.



We can implement Q-Q plot based on following scenarios:

1. **If two data sets come from populations with a common distribution**
2. **If two data sets have common location and scale**
3. **If two data sets have similar distribution shapes**
4. **If two data sets have similar tail behaviour**

Importance of Q-Q Plot

1. **It can be used with sample sizes**
2. **Many distribution aspects like shifts in scale/location, changes in symmetry, and the presence of outliers can also be detected from this plot.**

Uses of Q-Q Plot

1. **Q-Q plots can be used to compare the shape of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.**
2. **Q-Q plots can be used to compare collections of data, or theoretical distributions.**
3. **Q-Q plots containing two sample data to view non-parametric approach to compare underlying distributions.**