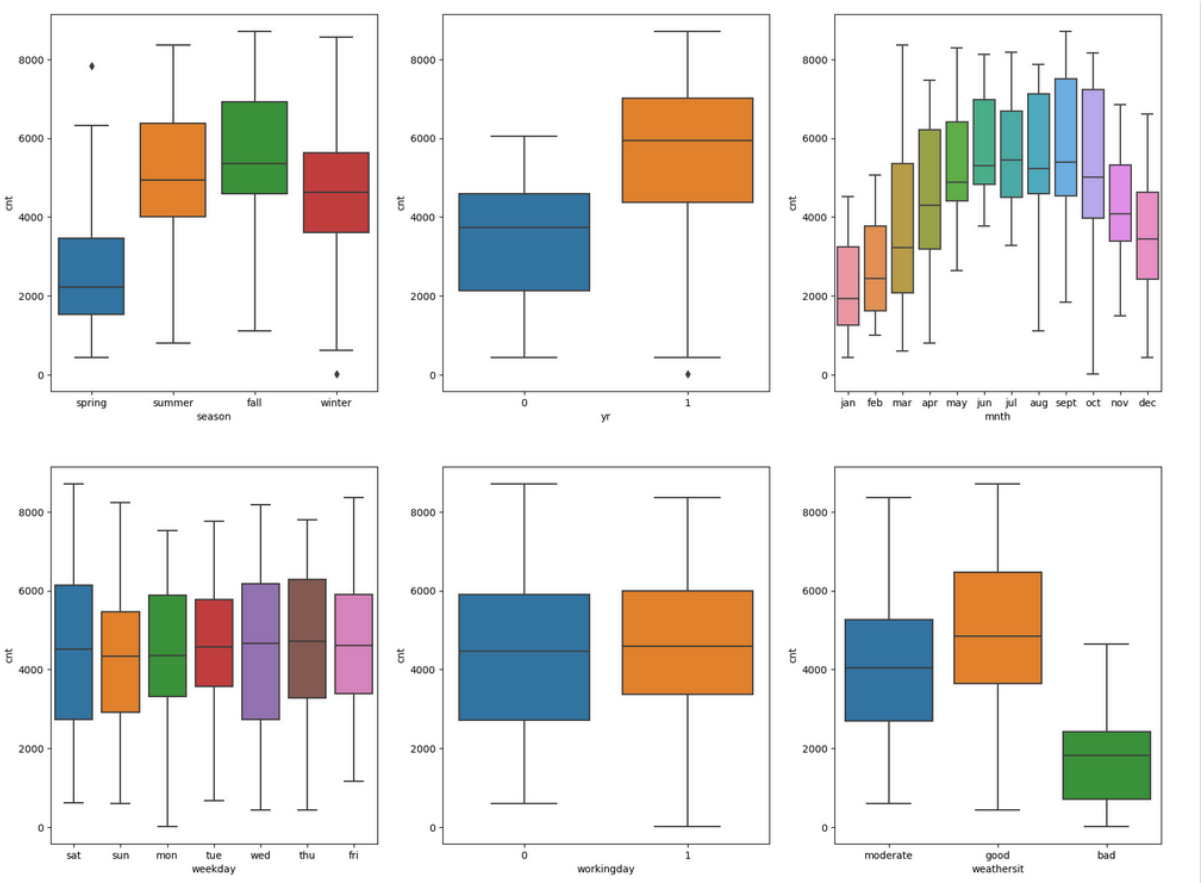
Assignment-based Subjective Questions

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

Several categorical variables, including season, month, yr, weekday, working day, and weathersit, exert significant influence on the dependent variable 'cnt.' The following figure illustrates the correlation among these variables.

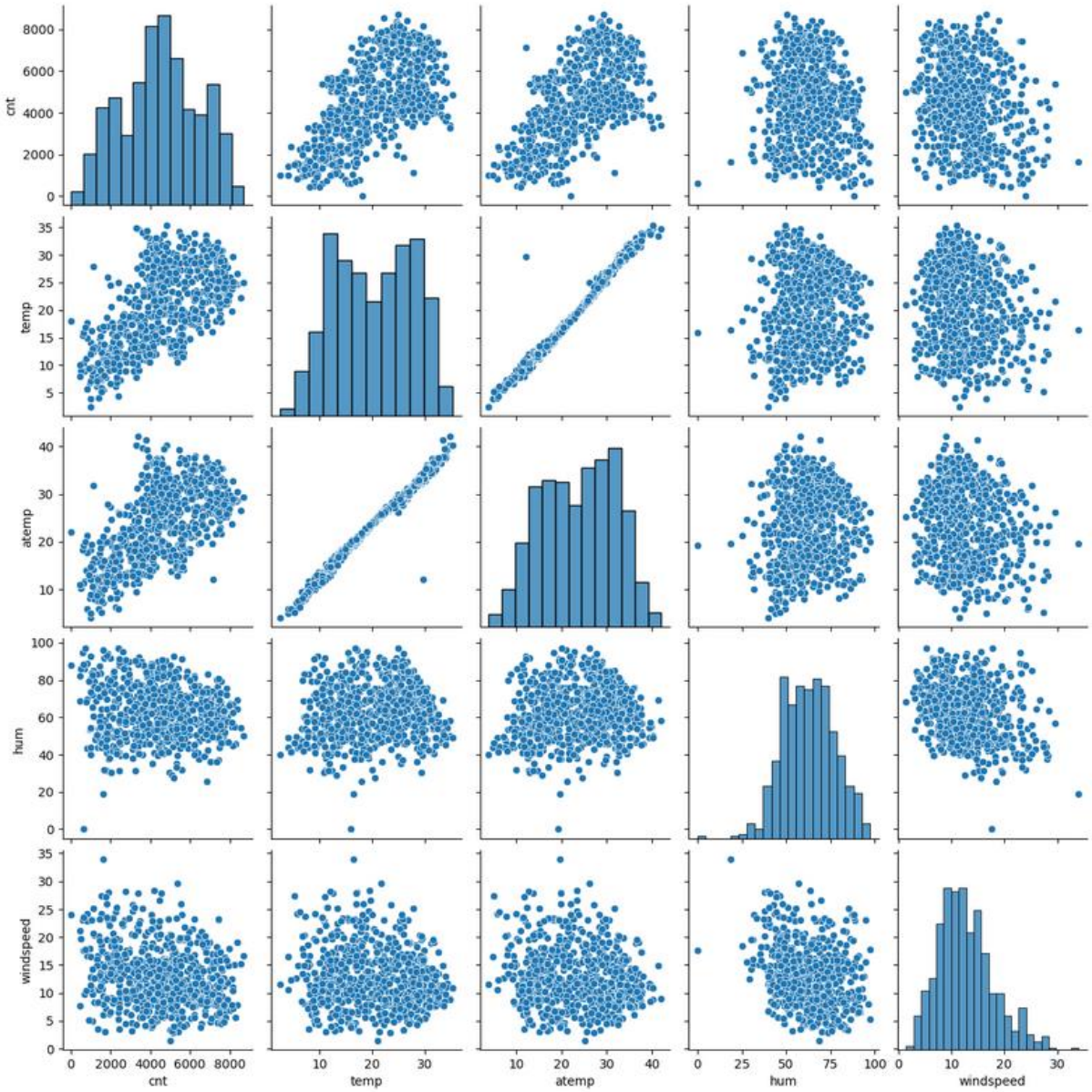


2. Why is it important to use drop_first=True during dummy variable creation?

The purpose of creating dummy variables is to represent categorical variables with 'n' levels by generating 'n-1' new columns, each indicating the presence or absence of a particular level using binary values (0 or 1). Therefore, the parameter 'drop_first=True' is employed to ensure that the resulting dummy variables represent 'n-1' levels, reducing correlation among them.

For example, if there are 3 levels, setting 'drop_first=True' will exclude the first column, effectively encoding the information about the presence of the first level through the absence of the other two levels.

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



The 'temp' and 'atemp' variables have highest correlation when compared to the rest with target variable as 'cnt'.

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

Linear Regression models are validated based on Linearity, No auto-correlation, Normality of error, Homoscedasticity, Multicollinearity. Ensuring that these assumptions hold or addressing violations of these assumptions is crucial for building reliable and valid linear regression models. Violations of these assumptions may require transformations of variables, data pre-processing, or choosing alternative modelling techniques.

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

Top 3 features that has significant impact towards explaining the demand of the shared bikes are temperature, year and season

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a fundamental statistical and machine learning algorithm used for modelling the relationship between a dependent variable (often denoted as "Y") and one or more independent variables (often denoted as "X"). Its primary goal is to find the best linear relationship that predicts the dependent variable based on the independent variables. If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression. The linear regression model gives asloped straight line describing the relationship within the variables. A regression line can be a Positive Linear Relationship or a Negative Linear Relationship. The goal of the linear regression algorithm is to get the best values for a0 and a1 to find the best fit line and the best fit line should have the least error. In Linear Regression, RFE or Mean Squared Error (MSE) or cost function is used, which helps to figure out the best possible values for a0 and a1, which provides the best fit line for the data points

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyze it and build your model. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another. Anscombe's quartet intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough.

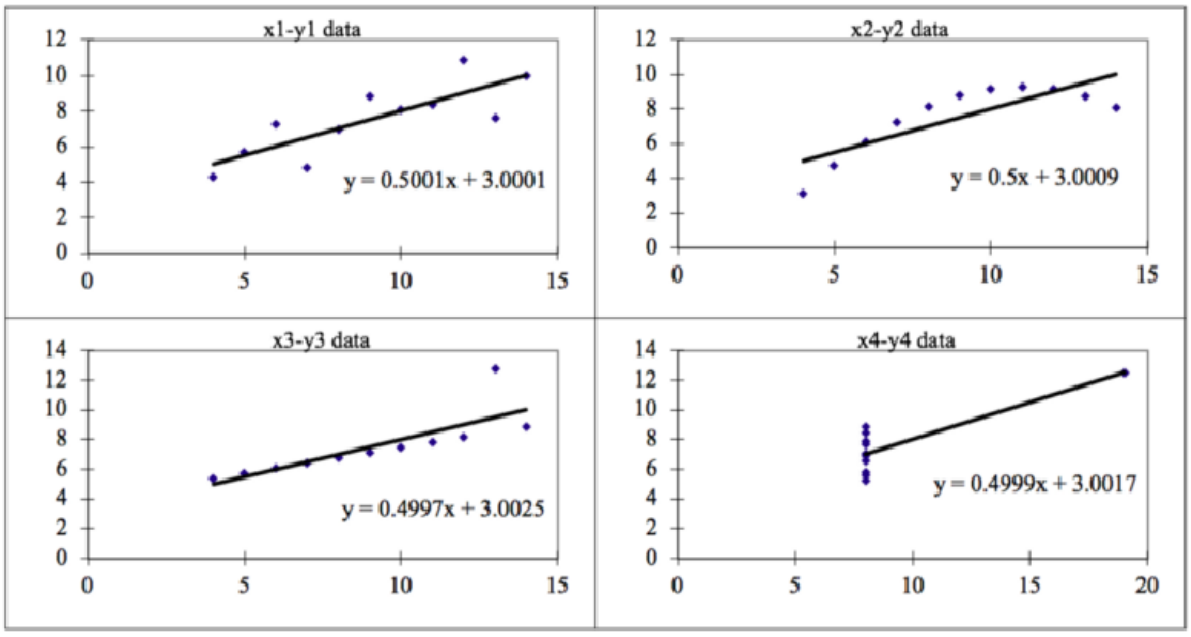
These four plots can be defined as follows:

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

The statistical information for all these four datasets is approximately similar and can be computed as follows:

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	

When these models are plotted on a scatter plot, all datasets generate a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:



The four datasets can be described as:

1. The first scatter plot (top left) appears to be a simple linear relationship,

2. The second graph (top right); cannot fit the linear regression model because the data is non-linear

3. In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line. It shows the outliers involved in the dataset which cannot be handled by linear regression model

4. Finally, the fourth graph (bottom right) shows the outliers involved in the dataset which cannot be handled by linear regression model. It shows an example when one high-leverage point is enough to

produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables. It shows the outliers involved in the dataset which cannot be handled by linear regression model

3. What is Pearson's R?

The Pearson correlation method is the most common method used for numerical variables. It assigns

a value between − 1 and 1, where 0 is no correlation, 1 is total positive correlation, and − 1 is total negative correlation. Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), 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 deviations.

Pearson's R Formula is as follows:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Here,

a. r = correlation coefficient
b. xi = values of the x-variable in a sample
c. $\overline{x}$ = mean of the values of the x-variable
d. yi = values of the y-variable in a sample
e. $\overline{y}$ = mean of the values of the y-variable

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

Scaling means you're transforming your data so that it fits within a specific scale. It is one type of data pre-processing step where we will fit data in specific scale and speed up the calculations in an algorithm. Collected data contains features varying in magnitudes, units and range. If scaling is not performed than algorithm tends to weigh high values magnitudes and ignore other parameters which will result in incorrect modelling.

Difference between Normalizing Scaling and Standardize Scaling:

a. In normalized scaling minimum and maximum value of features being used whereas in Standardize scaling mean and standard deviation is used for scaling.
b. Normalized scaling is used when features are of different scales whereas standardized scaling is used to ensure zero mean and unit standard deviation.
c. Normalized scaling scales values between (0,1) or (-1,1) whereas standardized scaling is not having or is not bounded in a certain range.
d. Normalized scaling is affected by outliers whereas standardized scaling is not having any effect by outliers.
e. Normalized scaling is used when we don't know about the distribution whereas standardized scaling is used when distribution is normal.
f. Normalized scaling is called as scaling normalization whereas standardized scaling is called as Z Score Normalization

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

Variance Inflation Factor (VIF) is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. It is calculated by taking the ratio of the variance of all a given model's betas divide by the variance of a single beta if it were fit alone.

The higher the VIF value, the greater the correlation of the variable with other variables. Values of more than 4 or 5 are sometimes regarded as being moderate to high, with values of 10 or more being regarded as very high. If there is perfect correlation, then VIF = infinity. An infinite VIF value means that the variable is exactly linear combination of other variable. If the independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it's R-squared value will be equal to 1. So, VIF = 1/(1-1) which gives VIF = 1/0 which results in "infinity"

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

Quantile - Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, Exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution. A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions. The power of Q-Q plots lies in their ability to summarize any distribution visually.

The advantages of the Q-Q plot are:

a. The sample sizes do not need to be equal.
b. Many distributional aspects can be simultaneously tested.

Q-Q plot is very useful to determine:

a. If two populations are of the same distribution
b. If residuals follow a normal distribution. Having a normal error term is an assumption in regression
c. and we can verify if it's met using this.
d. Skewness of distribution