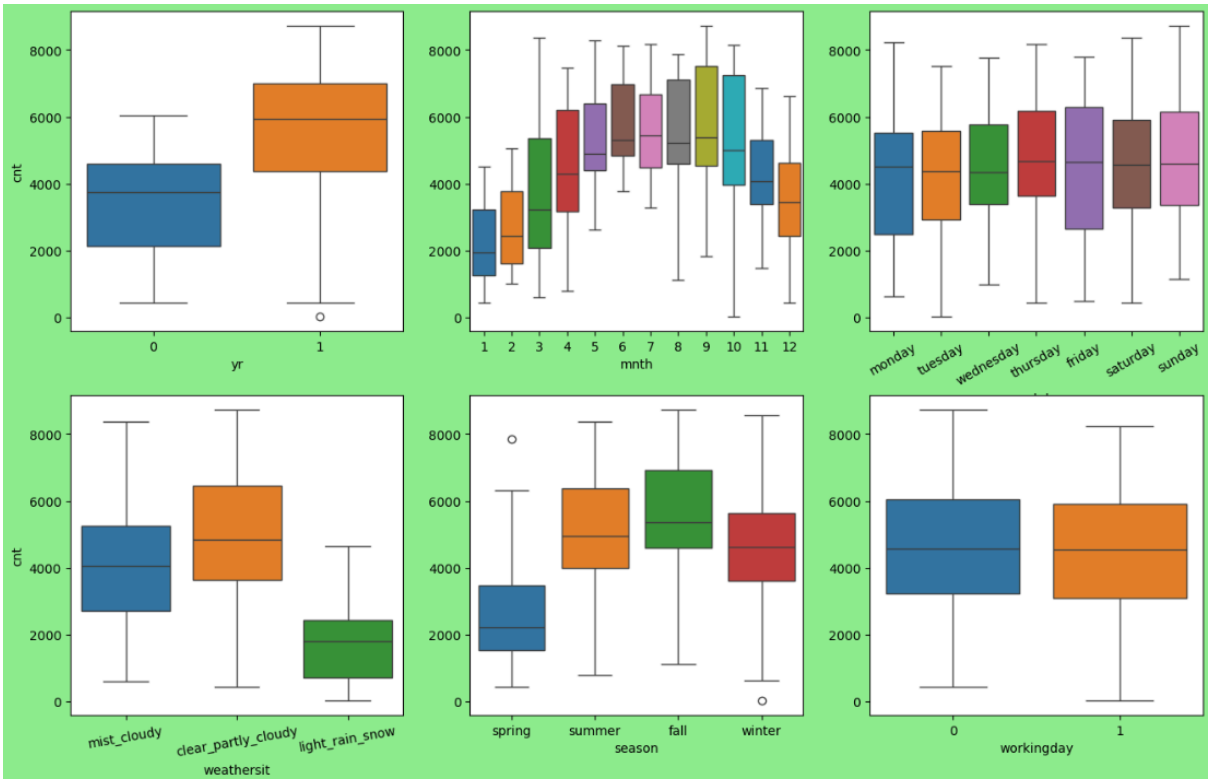
Assignment-based Subjective Questions

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

Ans.1 : Several categorical variables we have such as 'yr'(year), 'mnth'(month), 'weekday', 'weathersit'(weather situation), 'season', 'workingday', significantly influence the dependent variable 'cnt'. The following figure illustrates the correlation between these variables and 'cnt'.

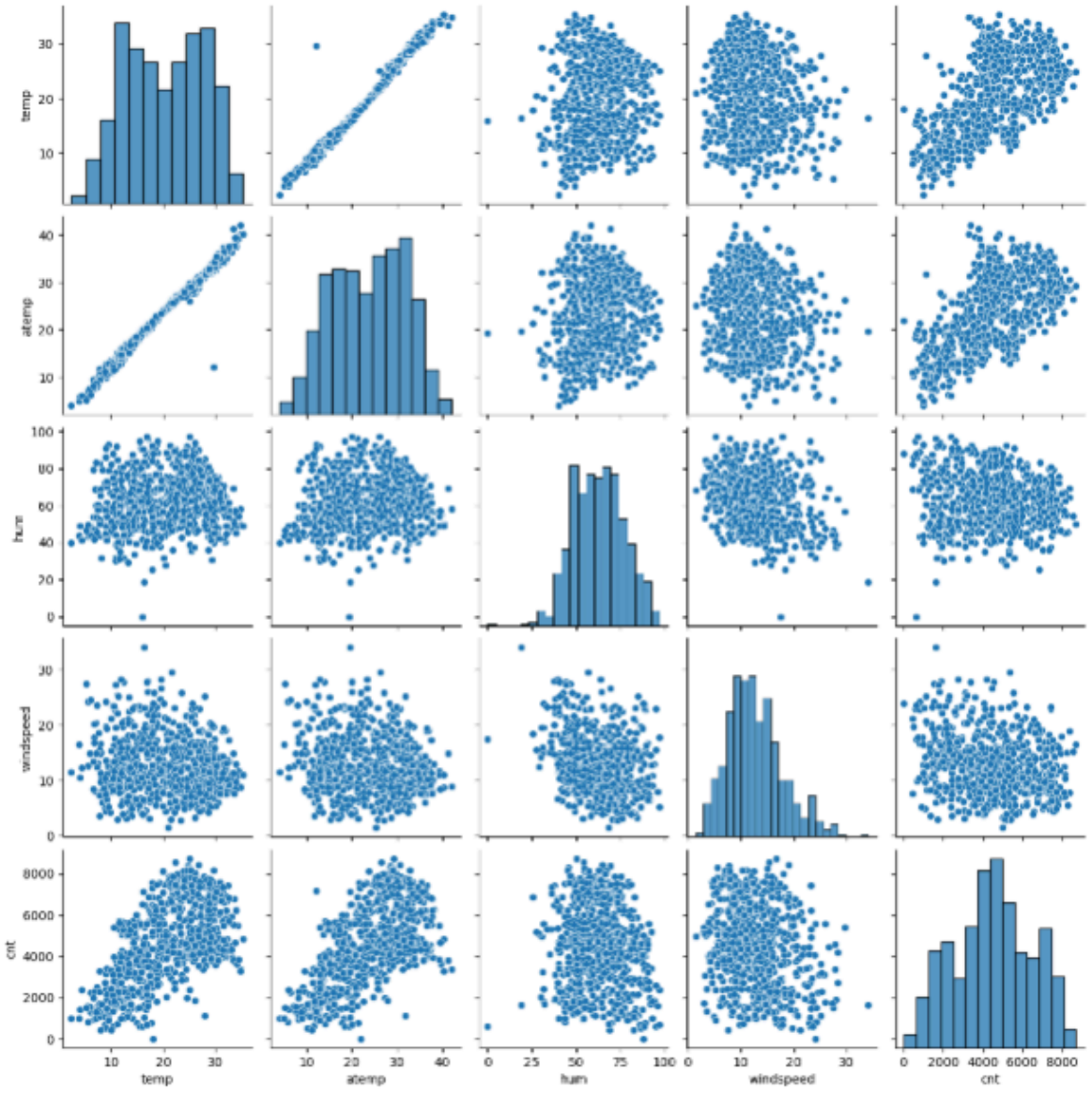


Q.2. Why is it important to use drop_first=True during dummy variable creation ?

Ans.2 : While creating dummy variables, using "drop_first = True" is important to avoid the dummy variable trap which refers to the situation where multicollinearity occurs in a regression model.
Multicollinearity happens when one or more predictor variables are highly correlated, making the model difficult to distinguish between them.
Hence, "drop_first = True" prevents multicollinearity, ensures the model is well-specified and simplifies the interpretation of the result.

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

<u>Ans.3</u> : By looking at the following pairplot we found that **"temp"** and **"atemp"** have the highest correlation with the target variable "cnt" as compared to the other numerical variables.



<u>Q.4</u>. How did you validate the assumptions of Linear Regression after building the model on the training set ?

<u>Ans.4</u> : Linear regression model validated based on Linearity, No-autocorrelation, Normality of Residuals, Homoscedasticity, Multicollinearity.

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

Ans.5 : Top 3 features contributing significantly towards explaining the demand of the shared bikes are Month, weather and season.

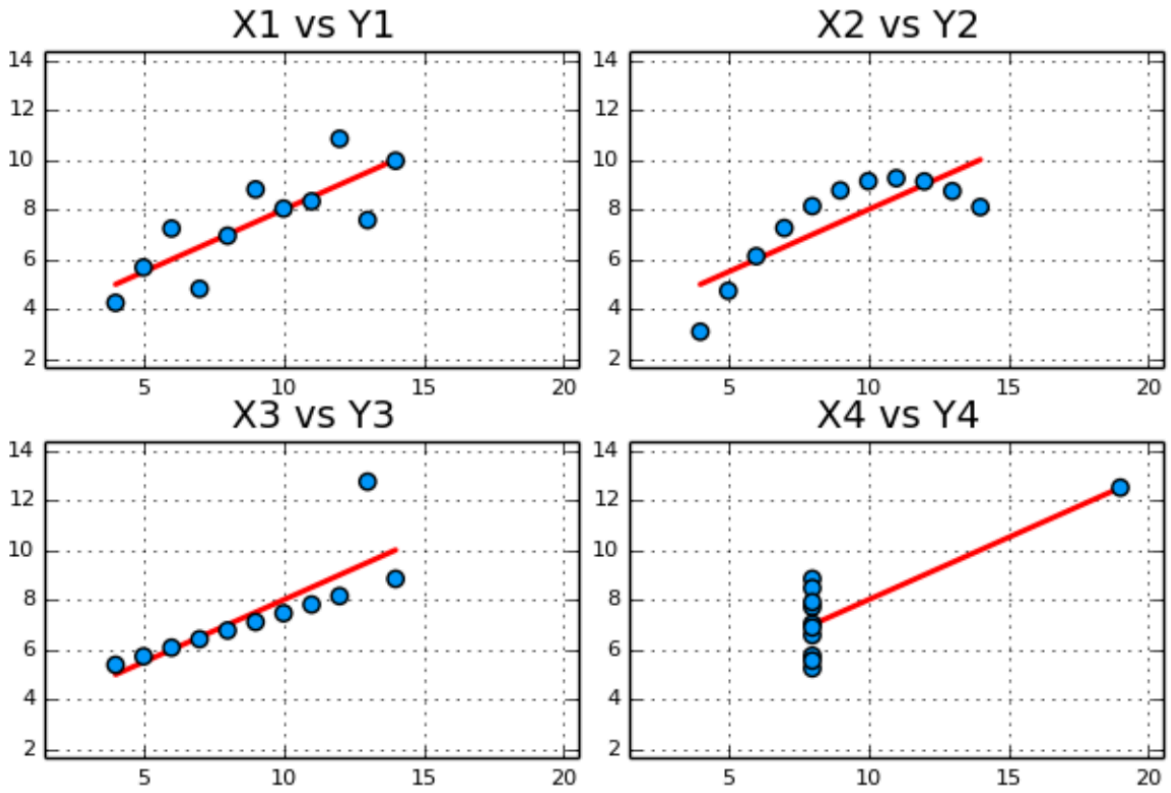
General Subjective Questions

Q.1. Explain the linear regression algorithm in detail.

Ans.1 : Linear regression is a fundamental statistical method used to model the relationship between dependent variable (target variable) and independent variable (predictors). Since linear regression shows the linear relationship, which means it finds that the value of dependent variable is changing according to the value of independent variable. If there is a single input variable or single independent variable such linear regression is called Simple Linear Regression. And if there is more than one input variable or independent variable such linear regression is called Multiple Linear Regression. This model produces a straight line with a slope describing the relationship between the variables. A regression line can be a Positive linear relationship or a Negative linear relationship.The goal of a linear regression algorithm is to find the best value for the coefficient $\beta 0$ and $\beta 1$ to find the best fit line with minimal errors. To achieve this, methods like RFE (Recursive Feature Elimination) or MSE (Mean Squared Error) are deployed. These methods help us to find the best values for the coefficient $\beta 0$ and $\beta 1$ and a best fit line with minimal errors.

Q.2. Explain the Anscombe's quartet in detail.

Ans.2 : Anscombe's quartet is a collection of four datasets to illustrate the importance of visualising data before analysing it. These datasets are nearly identical in many statistical properties, such as mean, variance and correlation but they have very different distributions when plotted on scatter plots.



Dataset 1: The points form a straight line when plotted and fits linear regression model as it seems to be a linear relationship between X and Y.

Dataset 2: The points form a curve, not a straight line. It does not show a linear relationship between X and Y , which means it does not fit the linear regression model.

Dataset 3: Most points form a straight line, but one point is way off (an outlier).
It shows outliers present in the dataset which can't be handled by a linear regression model.

Dataset 4: Almost all points are stacked vertically except for one point that's far away. It has a high leverage point means it produces a high correlation coefficient.

These four data set plots which have nearly the same statistical observations, which provides the same statistical information that involves variance, and mean of all x,y points in all four datasets. But when you plot these plots on scatter plots they look completely different.

This shows you can't always trust summary numbers alone, you need to visualize the data to understand the full picture.

Q.3. What is Pearson's R?

Ans.3 : Pearson's R is also known as Pearson's correlation coefficient. It is the most common way of measuring linear relationship between two variables.
Formula to find Pearson's R is given below:

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

Where,

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

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

Ans.4 :

What is scaling -
It is a step of data pre-processing which is applied on independent variables to normalise the data within a particular range. It also helps in speeding the calculations in an algorithm.

Why is scaling performed -
Most of the times collected dataset contains variables highly varying in magnitude, units and range. If scaling is not done then the algorithm only takes magnitude in account and not units which leads to incorrect modelling. To solve this problem we have to do scaling to bring all the variables to the same level of magnitude.
It is important to note that scaling only affects the coefficients and none other parameters like p-value, R-squared etc.

Difference between Normalized and Standardized scaling -

1. In normalized scaling minimum and maximum value of features are used whereas in Standardize scaling mean and standard deviation is used for scaling.

2. Normalized scaling brings all the values in the range of 0 and 1 whereas standardized scaling is not bounded in a certain range.

3. Normalized scaling is affected by outliers whereas standardized scaling is not affected by outliers.

4. Normalized scaling is used when we don't know about the distribution whereas standardized scaling is used when distribution is normal.

5. Normalized scaling is called as scaling normalization whereas standardized scaling is called as Z Score Normalization.

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

Ans.5 : VIF(Variance Inflation Factor) helps in explaining the relationship between one independent variable with all the other independent variables.
The formulation of VIF is given below:
VIF is greater than 10 is considered very high VIF, we should eliminate the variable,
VIF is greater than 5 is considered to be ok but it is worth inspecting,
VIF less than 5 is considered a good VIF, there is no need to eliminate variable.

VIF is calculated by the given formula :

$$VIF_i = \frac{1}{1 - R_i^2}$$

A very high VIF shows the perfect correlation between two independent variables. In such cases of perfect correlation, we get R-squared value 1 then we calculate the VIF as 1/1-1^2 becomes infinite, this causes perfect Multicollinearity. In this case we drop the variable which causes multicollinearity.

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

Ans.6 :

What is Q-Q plot :
Q-Q plot (Quantile - Quantile plot) is a graphical tool used to compare two probability distributions by plotting their quantiles against each other.
It helps to assess whether a data set follows a normal distribution.

Uses of Q-Q plot :
Q-Q plot (Quantile - Quantile plot) can also be used to determine whether the two distributions are similar or not. If they are quite similar we can expect the Q-Q plot to be more linear. The linearity assumption can be tested with scatter plots.

Importance of Q-Q plot :
In Linear regression, when we have a train and test dataset we can create a Q-Q plot by which we can confirm that both the train and test dataset are from the population with the same distribution or not.
Advantages :
1) It can be used with sample size also.
2) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.
Q-Q plot use on two datasets to check:
1) If both datasets came from population with common distribution.
2) If both datasets have common location and common scale.
3) If both datasets have a similar type of distribution shape.
4) If both datasets have tail behaviour.