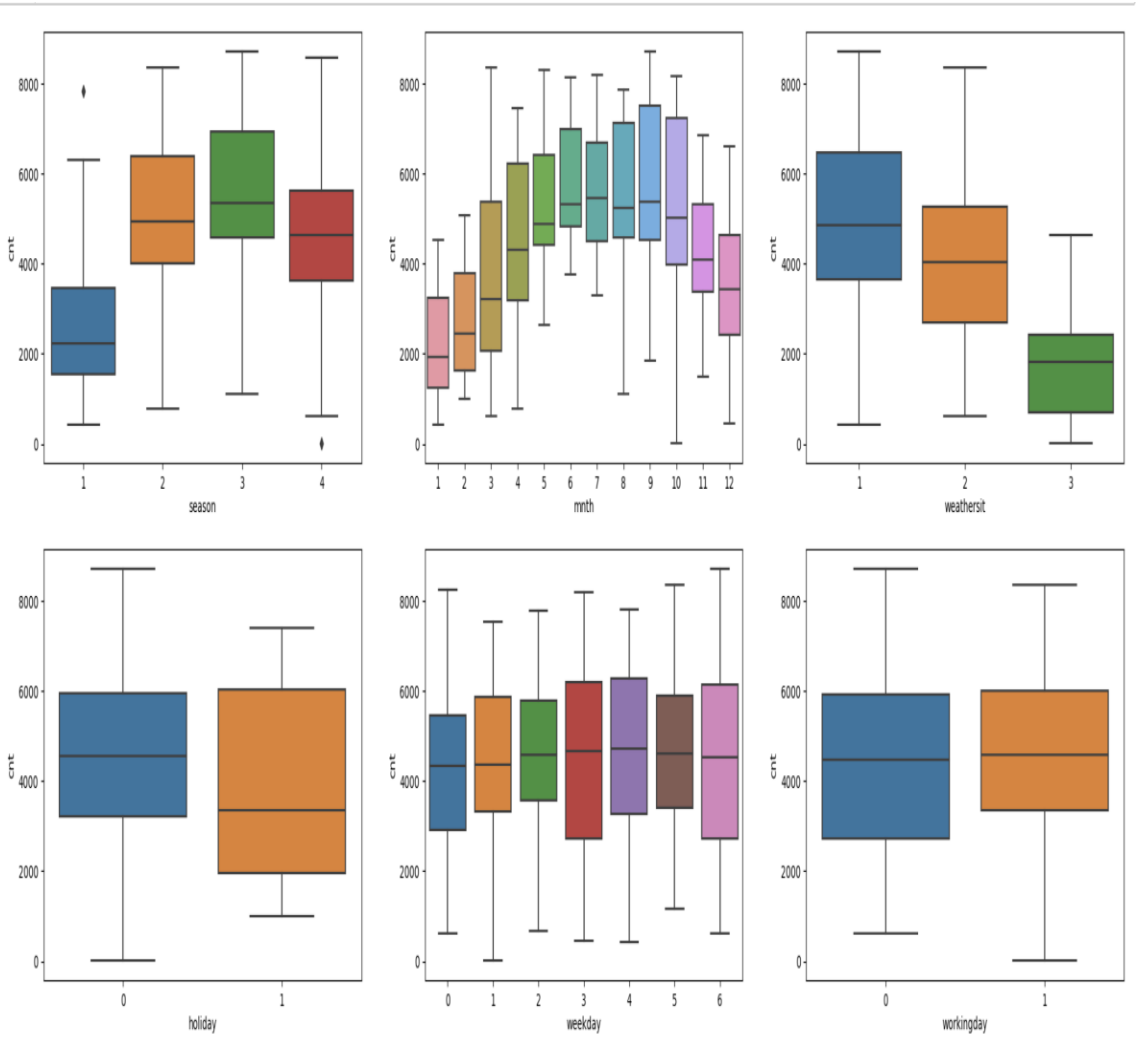
Assignment-based Subjective Questions

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



There were 6 categorical variables in the dataset. We used boxplot (refer the above picture) to study their effect on the dependent variable ('cnt'). The inferences that we could derive were -

season: Almost 32% of the bike booking were happening in season3 with a median of over 5000 booking (for the period of 2 years). This was followed by season2 & season4 with 27% & 25% of total booking. This indicates, season can be a good predictor for the dependent variable.

mnth: Almost 10% of the bike bookings were made in 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.

weathersit: Almost 67% of the bike booking were happening during weathersit1 with a median of close to 5000 booking (for the period of 2 years). This was followed by weathersit2 with 30% of total bookings. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.

holiday: Almost 97.6% of the bike bookings were made when it was not a holiday which means this data is clearly biased. This indicates that holiday cannot be a good predictor for the dependent variable.

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.

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.

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

If we include all dummy variables derived from a categorical variable in the model, it can lead to multicollinearity. Multicollinearity occurs when two or more independent variables are highly correlated, which can cause problems in regression analysis. Including all dummy variables creates perfect multicollinearity because the values of one dummy variable can be perfectly predicted from the others.

By setting drop_first=True, we avoid the dummy variable trap. It drops one of the dummy variables, and the information from that dropped variable is still captured by the intercept term in the model. This eliminates perfect multicollinearity.

If we have categorical variables with n-levels, then we need to use n-1 columns to represent the dummy variables.

Consider a Categorical column with 3 types of values, we want to create dummy variable for that column. If one variable is neither furnished nor semi_furnished, then it is obvious unfurnished. So, we do not need a 3rd variable to identify the unfurnished.

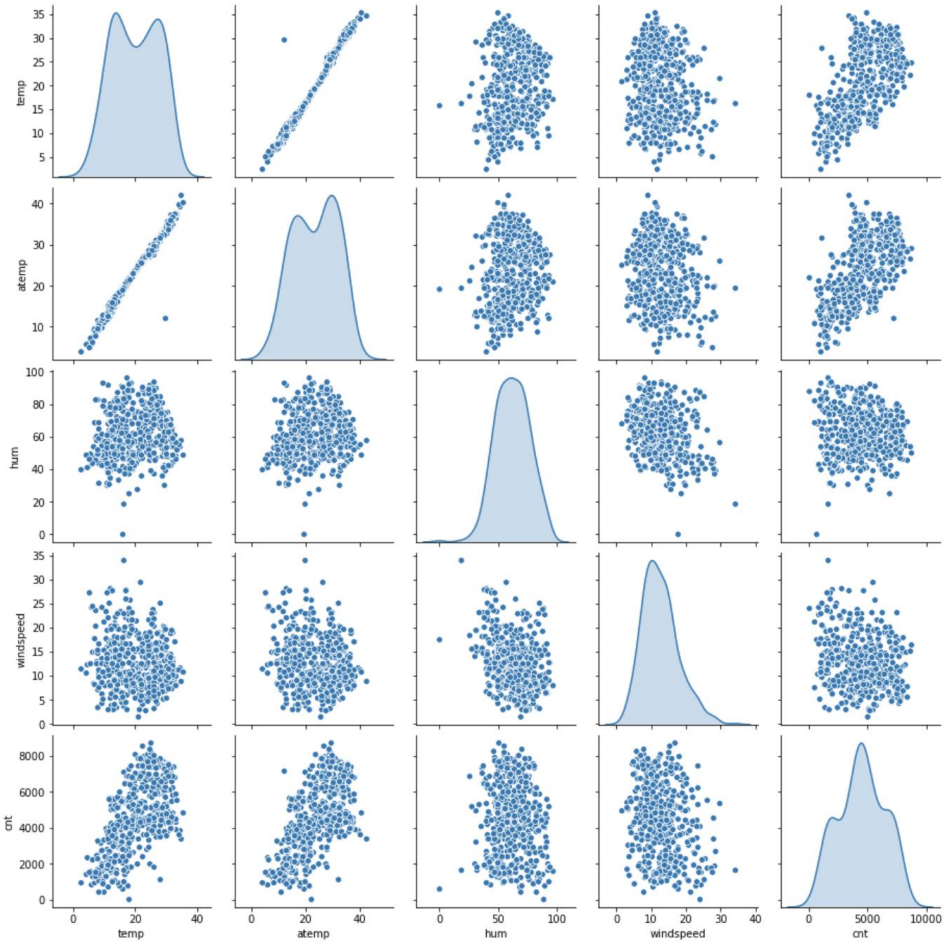
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

"temp" and "atemp" are the two numerical variables which are highly correlated with the target variable (cnt).

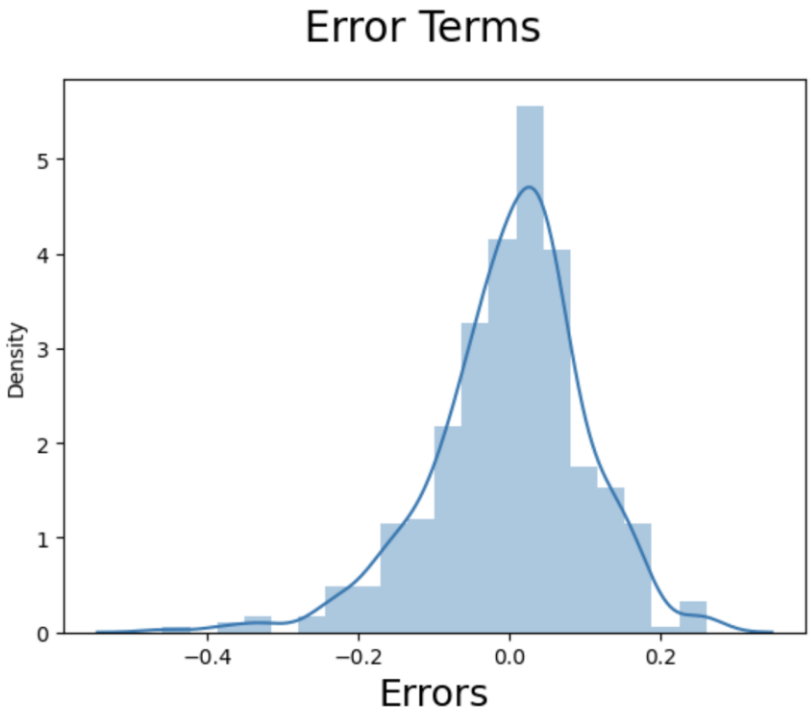
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

We have done following tests to validate assumptions of Linear Regression:

a. There should be a linear relationship between independent and dependent variables. We visualized the numeric variables using a pairplot to see if the variables are linearly related or not. Using the pair plot, we could see there is a linear relation between temp and atemp variable with the predictor ‘cnt’



b. Residuals distribution should follow normal distribution and centered around 0 (mean = 0). We validated this assumption about residuals by plotting a histogram of residuals and saw if residuals are following normal distribution or not. From the histogram below, we can see that the Residuals are normally distributed. Hence our assumption for Linear Regression is valid.



c. linear regression assumes that there is little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other. We calculated the VIF (Variance Inflation Factor) to quantify how strongly the feature variables in the new model are associated with one another.

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

The top 3 noteworthy features are:

1. temp - coefficient: 0.581783
2. yr - coefficient: 0.232798
3. weathersit_3 - coefficient: -0.256657

General Subjective Questions

6. Explain the linear regression algorithm in detail.

Linear regression is a machine learning algorithm based on supervised learning. It performs a regression task, which means it predicts a continuous output variable (y) based on one or more input variables (x). It is mostly used for finding out the linear relationship between variables and forecasting.
The basic idea of linear regression is to find a line that best fits the data points, such that the distance between the line and the data points is minimized. The line can be represented by an equation of the form:
$y = \theta 0 + \theta 1x$
where θ0 is the intercept (the value of y when x is zero) and θ1 is the slope (the change in y for a unit change in x). These are called the parameters or coefficients of the linear model.

To find the best values of θ0 and θ1, we need to define a cost function that measures how well the line fits the data. A common choice is the mean squared error (MSE), which is the average of the squared differences between the actual y values and the predicted y values:

$MSE = (1/n) * \sum(y - y')^2$

where n is the number of data points, y is the actual value, and y' is the predicted value.

The goal is to minimize the MSE by adjusting θ0 and θ1. There are different methods to do this, such as gradient descent, normal equation, or using libraries like scikit-learn.

Linear regression can also be extended to multiple input variables (x1, x2, …, xn), in which case the equation becomes:

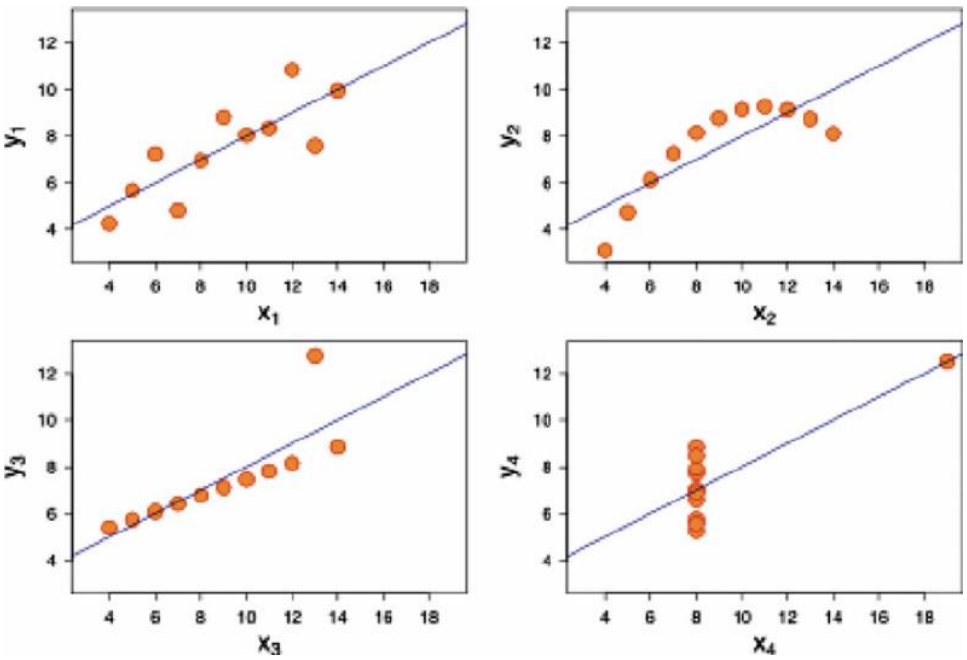
$y = \theta 0 + \theta 1x1 + \theta 2x2 + \ldots + \theta nxn$

Limitations are it assumes a linear relationship between the input variables and the output variable, which may not always be the case. Another limitation is that it may be sensitive to outliers or multicollinearity.

7. Explain the Anscombe's quartet in detail.

Anscombe's Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have quite a different distribution and look different when plotted on a graph. It was developed to emphasize both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties.

- The first scatter plot (top left) is a simple linear relationship.
- The second graph (top right) is not distributed normally; while there is a relation between them is not linear.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line the calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) 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.



8. What is Pearson's R?

Pearson's r is a numerical summary of the strength of the linear association between the variables. Its value ranges between -1 to +1. It shows the linear relationship between two sets of data. In simple terms, it tells us "Can we draw a line graph to represent the data?"

Formula

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

r = correlation coefficient

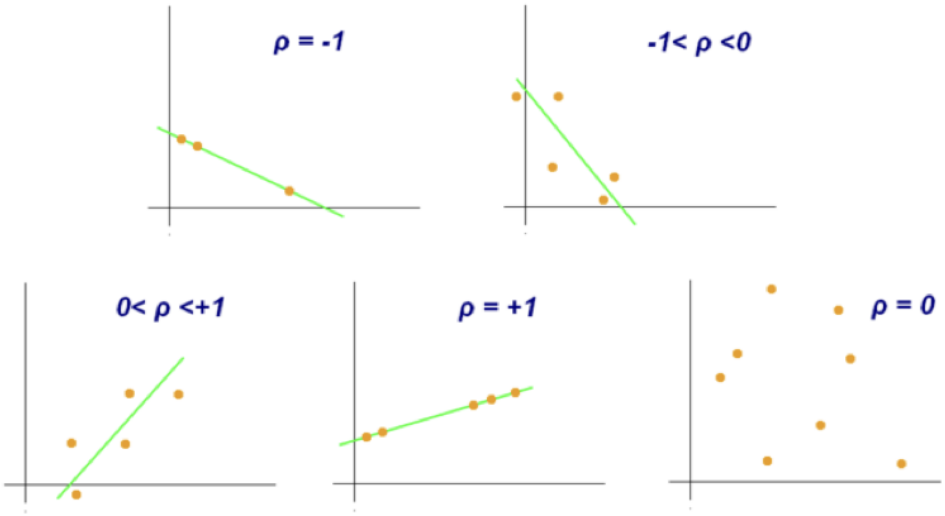
x_i = values of the x-variable in a sample

$\bar{x}$ = 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

As can be seen from the graph below, r = 1 means the data is perfectly linear with a positive slope r = -1 means the data is perfectly linear with a negative slope r = 0 means there is no linear association



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

Feature scaling is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data pre-processing stage to deal with varying values in the dataset. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values.

- Normalization is used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbours and Neural Networks.
- Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

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

The VIF (Variance Inflation Factor) gives how much the variance of the coefficient estimate is being inflated by collinearity. If there is perfect correlation, then VIF = infinity. It gives a basic quantitative idea about how much the feature variables are correlated with each other. It is an extremely important parameter to test our linear model.

$$VIF = \frac{1}{1 - R^2}$$

Where R-1 is the R-square value of that independent variable, we want to check how well this independent variable is explained well by other independent variables. If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and its R-squared value will be equal to 1. So, VIF = 1/(1-1) which gives VIF = 1/0 which results in “infinity” The numerical value for VIF tells you (in decimal form) what percentage the variance (i.e., the standard error squared) is inflated for each coefficient. For example, a VIF of 1.9 tells you that the variance of a particular coefficient is 90% bigger than what you would expect if there was no multicollinearity — if there was no correlation with other predictors.

A rule of thumb for interpreting the variance inflation factor:

- 1 = not correlated.
- Between 1 and 5 = moderately correlated.
- Greater than 5 = highly correlated.

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

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. It is used to compare the shapes of distributions. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that is straight. The q-q plot is used to answer the following questions:

● Do two data sets come from populations with a common distribution?
● Do two data sets have common location and scale?
● Do two data sets have similar distributional shapes?
● Do two data sets have similar tail behavior?