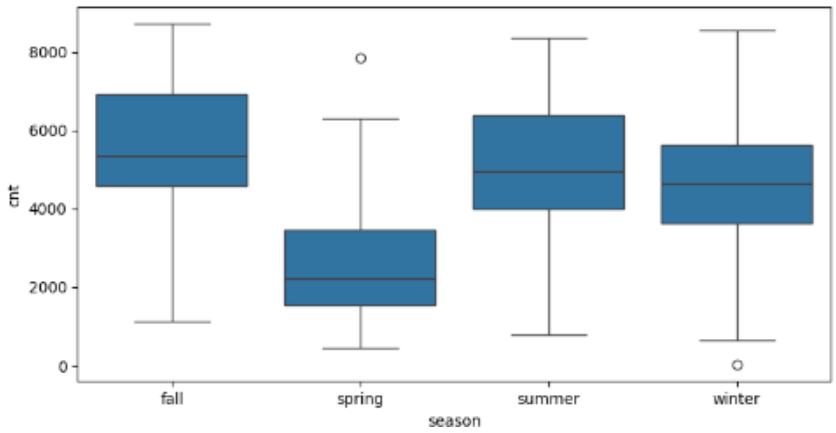
Assignment-based Subjective Questions

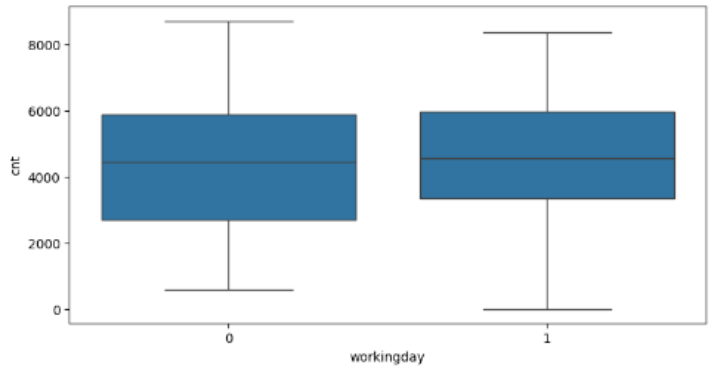
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

   Below are the categorical variables and their effect on the dependent variable cnt:

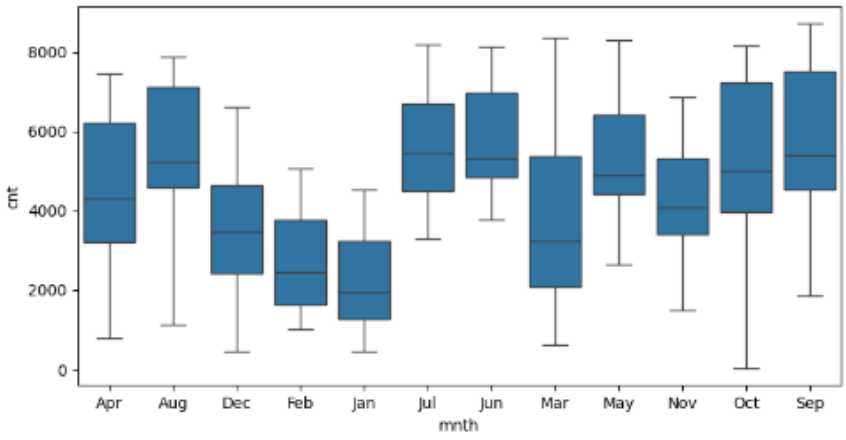
   - Season: Dependent variable 'cnt' is much higher in summer and fall compared to spring and winter.



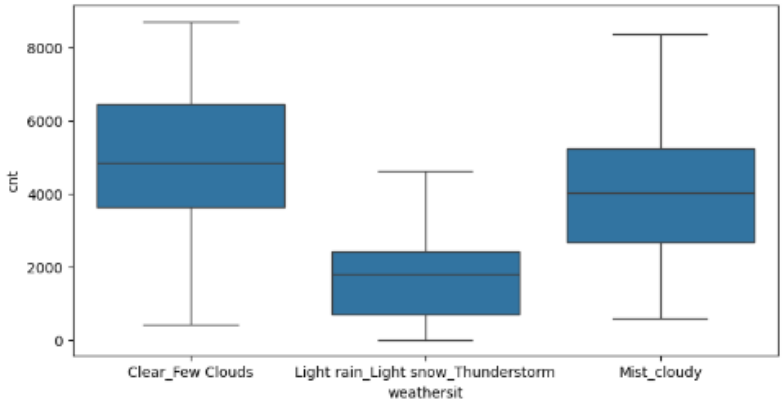
   - Working day: Registered users prefer renting bikes on working days.



   - Month: Rental demand is higher in June, July, August, September & October month.



- Weathersit: Most bikes are rented when sky is clear or with few clouds



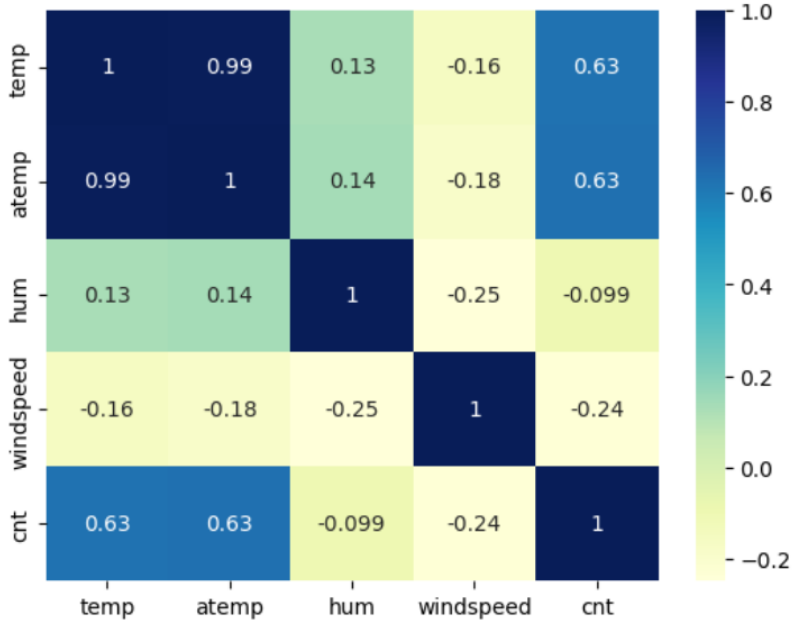
- Holiday: For casual users, more bike demands are seen on holidays.
- Working day: Slightly higher renting occurs on working day.
- Year: 2019 sees higher demand compared to 2018.

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

   By using drop_first=True, we drop the first column created by get_dummies. This avoids multicollinearity.
   For example, for month column, it will drop a column and create 11 columns instead of 12.

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

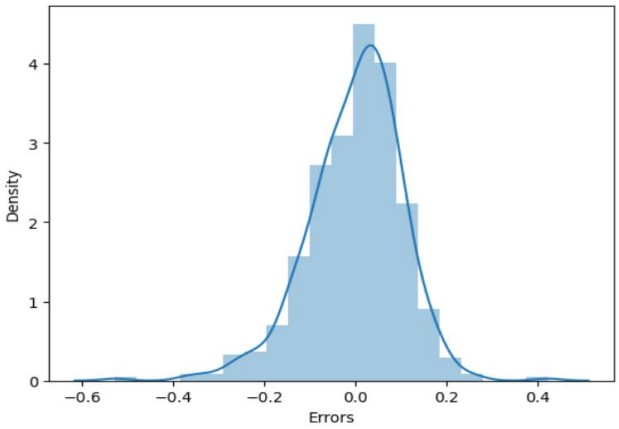
   Temp and atemp variable have highest correlation with cnt.



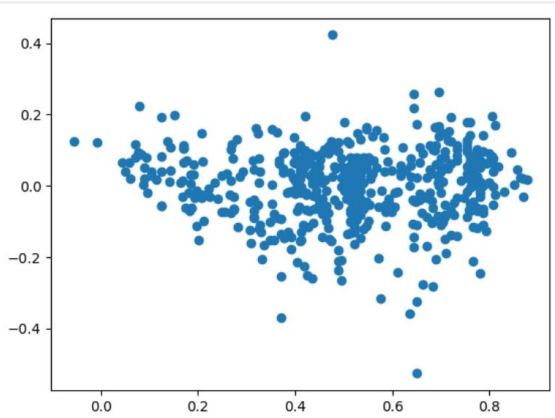
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

   Assumptions validated are as mentioned below:

   - Normalization: Errors are normally distributed around mean 0.



   - Independence: Error should not depict any clear pattern. Same can be seen in this scatter plot between predicted values and residual. X axis is residual. Y axis is predicted values.



   - Homoscedasticity: This implies that error terms have constant variance.

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

   According the model derived, top 3 significant features are:

   - weathersit_Light rain_Light snow_Thunderstorm (-0.311)
   - yr_1 (0.246)
   - season_spring (-0.220)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression is used to examine the relation between one or more independent variable with a dependent variable. It does this by fitting a linear equation to observed data points. Once the linear equation is derived referred as a model, it is used to predict the values of dependent variable based on provided independent variables.

It has 2 types:
Simple Linear Regression: When only one independent variable is uses to predict dependent variable.
Multiple Linear Regression: When more than one independent variables are used to predict dependent variable.

The algorithm has various steps mentioned as below:

- Analyzing the data: Analyzing the correlation between variables. This is done using scatter plots. Identifying relevant variables which can help formulate the model.
- Estimating the model: Finding the line that best fits the data.
  Generally, this is done using minimizing the expression of RSS (Residual Sum of Squares).

  The linear regression model can be represented mathematically as:
  Y=β0+β1X1+β2X2+⋯+βnXn+ϵY = \beta_0 + \beta_1X_1 + \beta_2X_2 + \cdots + \beta_nX_n + \epsilonY=β0+β1X1+β2X2+⋯+βnXn+ϵ

  Here, Y is the predicted value of the dependent variable (e.g., bike rentals).
  And X_1, X_2, ..., X_n are the independent variables (e.g., temperature, weather condition, year).
  β_0 is the intercept (the predicted value when all independent variables are zero).
  β_1, β_2, ..., β_n are the coefficients for each independent variable.
  ε is the error term or residual (the difference between the observed and predicted values).

- Evaluating the model: The main metric to measure the extent of estimated fit of linear line is R2 (R square) which is defined as total variance/ explained variance.

  Significance of a model fit is measured using F statistics.

Assumptions of Linear Regression:

- There isa linear relationship between variables x & y.
- Normality: Error terms are normally distributed around mean 0.
- Independence: Residuals are independent.
- Homoscedasticity: This implies that error terms have constant variance.

2. Explain the Anscombe's quartet in detail.
   It is a set of four datasets, each with 11 data points and they have similar statistical properties but show a very different visual behavior. The four datasets are as below:

	Description	Statistical Properties	Visual Pattern
Dataset 1	Data points are randomly scattered around a straight line.	Same mean, variance, correlation. Regression line is also same.	Linear Relationship. Regression line is a good fit.
Dataset 2	Data points are randomly scattered around a straight line but with a non linear relationship between the variables.		Non-linear relationship. Data does not fit the straight line well.
Dataset 3	Data points are randomly scattered around a straight line but with one outlier that is far away from rest of the data.		Contains an outlier that skews the regression line.
Dataset 4	Data points are randomly scattered around a non linear curve.		Vertical distribution of data points and a misleading regression line.

   Learnings:
   a. Data visualization is essential to identify patterns and outliers.
   b. Outliers can be important.
   c. Descriptive statistics does not fully depict the dataset.

3. What is Pearson's R?

   Pearson's Correlation Coefficient is used to measure how strongly two continuous variables are related to each other in a straight line fashion. It ranges from -1 to 1 and is denoted by symbol 'r'.

   - r = 1 implies perfect positive relationship
   - r = 0 means no relationship
   - r = -1 implies perfect negative relationship

   It is calculated using below formula:

   $$r = \Sigma[(xi - \bar{x})(yi - \bar{y})] / (\sqrt{[\Sigma(xi - \bar{x})^2]} * \sqrt{[\Sigma(yi - \bar{y})^2]})$$

   xi, yi being individual data points for 2 variables.
   $\bar{x}$, $\bar{y}$ are mean of those variables.

It is used to measure only straight-line relationships. It can be affected by outliers.

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

   Scaling means to adjust the range of feature values.
   When we have a lot of independent variables and a lot of them are on different scales, which lead to a model with very weird coefficients that are difficult to interpret. So scaling is needed for 2 reasons:
   1. Ease of interpretation
   2. Faster convergence of gradient descent methods.

   Normalized scaling: Variables are scales such that all values lie between 0 and 1.

   It is given by:

   X' = (x – xmin) / (xmax - xmin)

   Standardized Scaling: Variables are scaled in a way that their mean is zero and standard deviation is one.
   It is given by:
   X' = (x – mean(x)) / sd(x)

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

   VIF is defined as 1/ (1 – R*R)

   VIF of infinity indicates Rsquared = 1, which implies perfect correlation between two features. This implies,

   1. Perfect Multicollinearity: When one feature is a perfect linear combination of another.
   2. Dataset has duplicate or redundant features.

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

   Q-Q Plots are quantile quantile plots. It is tool to assess if 2 datasets are from common distribution. Theoretical distributions can be normal, exponential or uniform.

   They are used in linear regression to check if the train and test data sets are from the population with same distribution.

   Interpretations:

a. Similar distribution – If all the data points of quantile are lying around the straight line at an angle of 45 degrees from x axis.
b. If y-value quantiles are lower than x value quantiles.
c. If x value quantiles are lower than y value quantiles.
d. If data points are lying away from straight lines.

Advantages:
Distribution aspects like loc, scale shifts, outliers and symmetry changes can be deduced from single plot.
In short, if any dataset follows normal, uniform or exponential distribution, QQ plot can be used to plot quantiles of sample distribution. It also helps to know if errors in dataset are normal in nature or not.