General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear Regression is a supervised learning algorithm used to model the relationship between a dependent variable (target) and one or more independent variables (features). The goal is to find a linear relationship between the variables by fitting a line (in the case of one feature) or a hyperplane (in the case of multiple features) that best predicts the target variable.

Steps Involved:

1. Assumptions:
    a. Linearity: The relationship between the independent and dependent variables is linear.
    b. Independence: Observations are independent of each other.
    c. Homoscedasticity: The variance of residuals (errors) is constant.
    d. Normality: The residuals of the model should be normally distributed.
2. Model Representation:
    a. Simple Linear Regression:

    $$y = \beta_0 + \beta_1 x + \epsilon$$

    y = \beta_0 + \beta_1x + \epsilon
    b. Multiple Linear Regression: \(y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_nx_n + \epsilon \)
    c. Here, y is the dependent variable, \(x_1, x_2, \dots, x_n \) are independent variables, \beta_0 is the intercept, \(\beta_1, \beta_2, \dots, \beta_n \) are coefficients, and \epsilon is the error term.
3. Fitting the Model:
    a. The coefficients \(\beta_0, \beta_1, \dots, \beta_n \) are estimated using methods like Ordinary Least Squares (OLS), which minimizes the sum of squared residuals \sum(\hat{y} - y)^2 , where \hat{y} is the predicted value.
4. Model Evaluation:
    a. The performance of the model is evaluated using metrics such as R-squared, Adjusted R-squared, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

Linear regression is widely used in predictive modeling, but its accuracy depends on the validity of the assumptions and the quality of the input data.

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's Quartet is a set of four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation, etc.) but appear very different when graphed. Created by statistician Francis Anscombe in 1973, the quartet illustrates the importance of graphing data before analyzing it and shows how different datasets can have similar statistical properties but very different distributions.

Datasets:

- The quartet consists of four sets of (x, y) pairs.
- All four datasets have the same mean, variance, and correlation for x and y, and the same regression line.

Importance:

- The first dataset shows a linear relationship that fits well with linear regression.
- The second dataset has a clear non-linear relationship, suggesting that linear regression is not appropriate.
- The third dataset is a linear relationship with an outlier, which significantly affects the regression results.
- The fourth dataset shows a vertical line where one point is an outlier in the x-direction, resulting in a misleading regression line.

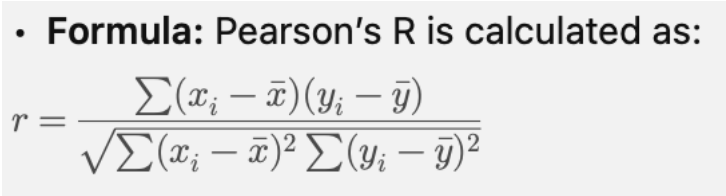
Conclusion:

Anscombe's quartet highlights the importance of visualizing data to identify patterns, outliers, and the appropriate statistical models for analysis.

3. What is Pearson's R? (3 marks)

Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear correlation between two variables. It quantifies the strength and direction of the linear relationship between them.

- Formula: Pearson's R is calculated as:



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 x_i and y_i are the individual data points, and \bar{x} and \bar{y} are the means of the variables.

- Range: The value of Pearson's R ranges from -1 to 1.
- r = 1 indicates a perfect positive linear relationship.
- r = -1 indicates a perfect negative linear relationship.
- r = 0 indicates no linear relationship.

Importance:

Pearson's R is widely used in statistics to understand the strength and direction of the linear relationship between two variables. However, it only measures linear relationships and can be misleading if the relationship is non-linear.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**

Scaling is the process of adjusting the range of independent variables or features of data so that they fit within a specific scale, such as 0-1 or have a mean of 0 and a standard deviation of 1. Scaling is important in algorithms that compute distances between data points, such as k-nearest neighbours or support vector machines, or when gradient descent optimization is used, as in linear regression.

Types of Scaling:

1. Normalized Scaling:
    - Definition: Normalization scales the features to a fixed range, typically 0 to 1.
    - Formula: $x' = \frac{x - \min(x)}{\max(x) - \min(x)}$
      x{\prime} = \frac{x - \min(x)}{\max(x) - \min(x)}
    - Use Case: Useful when the features have different units and you want to bring them to a common scale.
2. Standardized Scaling:
    - Definition: Standardization scales the features such that they have a mean of 0 and a standard deviation of 1.
    - Formula: x{\prime} = \frac{x - \mu}{\sigma}
    - Use Case: Useful when the features have different units and distributions but you want them to have the same statistical properties.

Difference:

- Normalization is bound to a specific range, which is particularly useful when you need to ensure that the features have the same scale.
- Standardization adjusts the data to a normal distribution with a mean of 0 and a standard deviation of 1, which is useful for algorithms that assume normally distributed data.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**

Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in regression models. It quantifies how much the variance of a regression coefficient is inflated due to collinearity with other predictors.

Infinite VIF:

- VIF becomes infinite when there is perfect multicollinearity between the features, meaning one predictor variable is a perfect linear combination of other predictors.
- In other words, if a feature can be exactly predicted by one or more other features in the dataset, the denominator in the VIF calculation, which is 1 - R^2 (where R^2 is the coefficient of determination), becomes zero, leading to an infinite VIF value.

Implications:

- An infinite VIF indicates that the regression model has redundant features, making it unstable. Removing or combining the collinear features can help stabilize the model.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**

Q-Q Plot (Quantile-Quantile Plot) is a graphical tool used to assess whether a dataset follows a particular distribution, most commonly the normal distribution. It plots the quantiles of the dataset against the quantiles of a theoretical distribution (like the normal distribution).

How to Interpret:

- If the points in a Q-Q plot fall along a straight line, it suggests that the data follows the theoretical distribution.
- Deviations from the straight line indicate departures from the theoretical distribution.

Use in Linear Regression:

- In linear regression, a key assumption is that the residuals (errors) are normally distributed. A Q-Q plot of the residuals can be used to check this assumption.
- If the residuals follow a straight line in the Q-Q plot, it confirms that they are approximately normally distributed, validating the assumption.
- If the residuals deviate significantly from the straight line, it suggests that the normality assumption may be violated, potentially affecting the validity of the regression model.
- Importance:
- The Q-Q plot is crucial for diagnosing issues with model assumptions, helping to ensure that the linear regression model is appropriate for the data at hand.

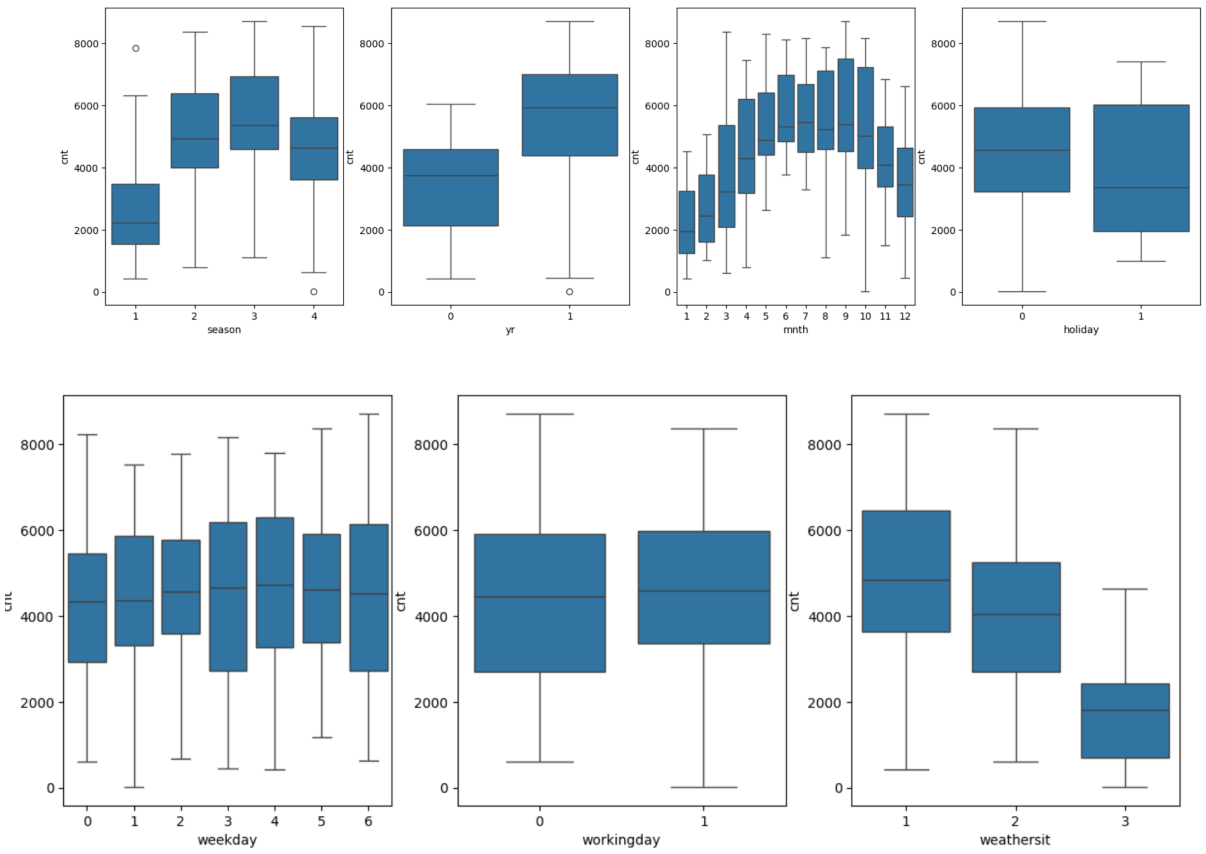
Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

Visualising Categorical Variables

As you might have noticed, there are a few categorical variables as well. Let's make a boxplot for some of these variables.

```
[593]: plt.figure(figsize=(20, 12))
plt.subplot(2,4,1)
sns.boxplot(x = 'season', y = 'cnt', data = bike_df)
plt.subplot(2,4,2)
sns.boxplot(x = 'yr', y = 'cnt', data = bike_df)
plt.subplot(2,4,3)
sns.boxplot(x = 'mnth', y = 'cnt', data = bike_df)
plt.subplot(2,4,4)
sns.boxplot(x = 'holiday', y = 'cnt', data = bike_df)
plt.subplot(2,4,5)
sns.boxplot(x = 'weekday', y = 'cnt', data = bike_df)
plt.subplot(2,4,6)
sns.boxplot(x = 'workingday', y = 'cnt', data = bike_df)
plt.subplot(2,4,7)
sns.boxplot(x = 'weathersit', y = 'cnt', data = bike_df)
plt.show()
```



Few observations from above heat map:

1. Median of bike rental count is highest for fall season. Need to check the statistical significance of this.
2. Also maximum bike rental count is for fall season.
3. Median of bike rental count is highest for year 2019. Also maximum bike rental count is for year 2019.
4. Highest bike rental count is for month 9 i.e. September.
5. Median of bike rental count is highest for month 7 i.e. July.

6. Lowest bike rental count is for month 10 i.e. October.
7. Median and highest bike rental count is for no holiday days marked with 0.
8. Bike rental median for all days are almost close to each other.
9. Bike rental median for working/non-working days are almost close to each other.
10. Bike rental median is highest for weathersit 1. Also maximum bike rental count is for weathersit 1 and lowest is for weathersit 3

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

When creating dummy variables from categorical data, using the drop_first=True parameter in pandas' get_dummies() function is important to avoid the "dummy variable trap."

Understanding the Dummy Variable Trap

The dummy variable trap occurs when the dummy variables are perfectly collinear, meaning one dummy variable can be predicted from the others. This happens because the sum of all dummy variables for a categorical feature equals 1, creating multicollinearity, which can lead to issues in regression models.

Example

Consider a categorical feature Color with three categories: Red, Blue, and Green.
If you create dummy variables without dropping the first one, you'd get:

Color_Red	Color_Blue	Color_Green
1	0	0
0	1	0
0	0	1

Notice how knowing the values of two dummy variables allows you to deduce the value of the third, creating multicollinearity.

Why Use drop_first=True

Using drop_first=True drops the first dummy variable (e.g., Color_Red), resulting in:

Color_Blue	Color_Green
0	0
1	0
0	1

Here, the first category (Red) becomes the baseline, and the model interprets the dropped variable as the reference category. Now, the other two dummies indicate whether the observation belongs to Blue or Green. This avoids multicollinearity and ensures the model coefficients are more interpretable.

Conclusion

Using drop_first=True during dummy variable creation:

1. **Prevents Multicollinearity**: Ensures that no dummy variable can be linearly predicted from the others, which is essential for accurate and stable regression models.
2. **Simplifies Model Interpretation**: Provides a baseline (reference category) against which the other categories are compared.

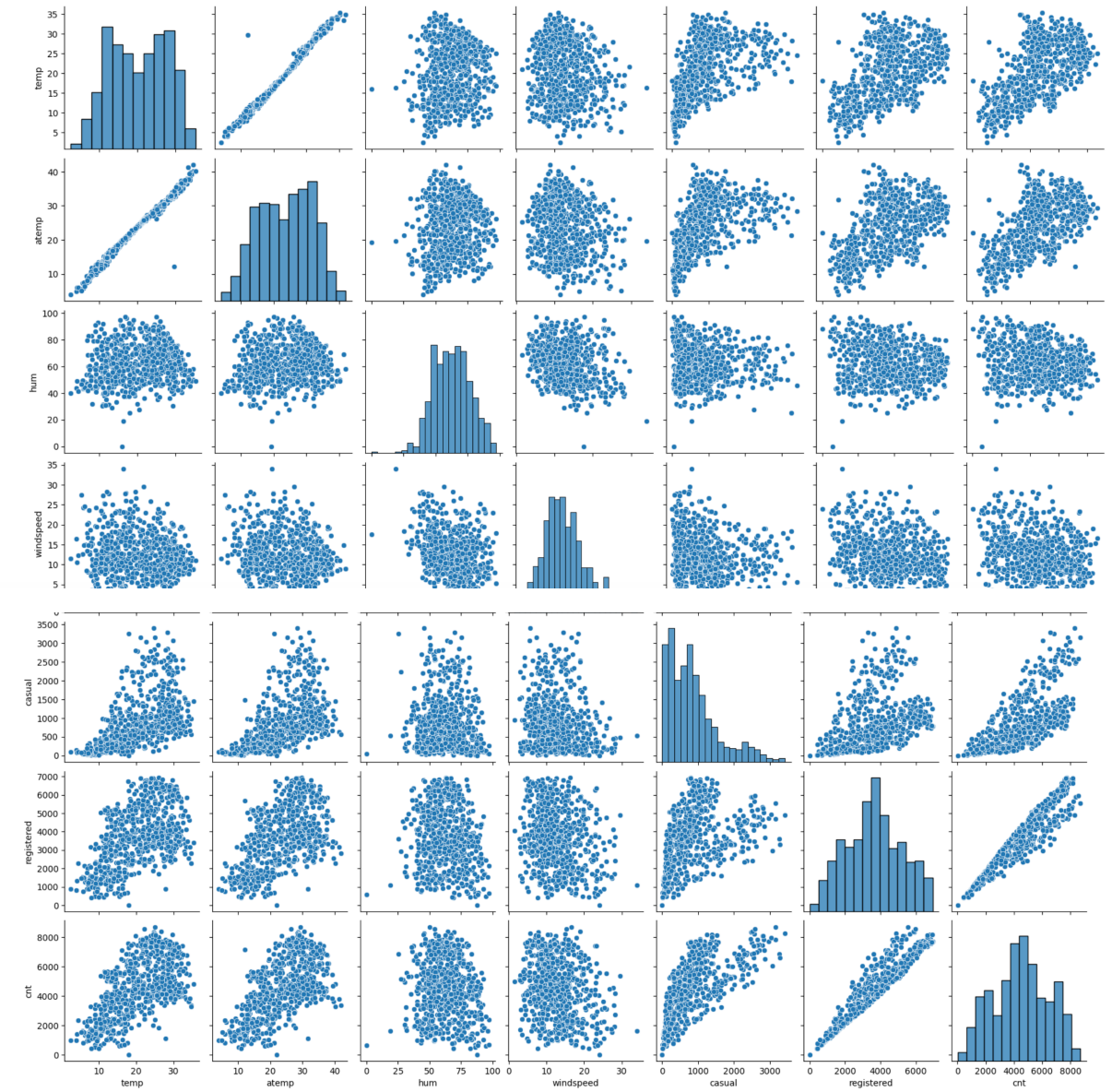
In summary, it is crucial to use **drop_first=True** when creating dummy variables to avoid the dummy variable trap, which can distort the results of your regression models.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

Visualising Numeric Variables

Let's make a pairplot of all the numeric variables

```
]: sns.pairplot(bike_df[["temp","atemp","hum","windspeed","casual","registered","cnt"]])
   plt.show()
```



Few observations from above pairplot:

1. **temp and atemp are highly correlated.**
2. temp is also correlated with casual, registered and cnt.
3. atemp is also correlated with casual, registered and cnt.
4. casual is correlated with registered and cnt. And thats obvious because casual + registered = cnt

We need to take care of these correlations among possible predictor variables to avoid multicollinearity once we start building our Multiple Linear ##### Regression Model.

Also we will not use casual and registered in our preprocessed data for model training. This might leak out the crucial information and model might get ##### overfit.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Validating the assumptions of Linear Regression after building the model on the training set is a crucial step to ensure the model is reliable and the inferences drawn from it are valid. Here's how you can validate the assumptions:

1. Linearity of the Relationship
   - Check: The relationship between the independent variables and the dependent variable should be linear.
   - Validation:
     - Plot the observed vs. predicted values (a scatter plot). The points should lie along a straight diagonal line.
     - Plot residuals (errors) vs. predicted values. The residuals should be randomly scattered around zero, with no obvious patterns.

2. Independence of Errors
   - Check: The residuals (errors) should be independent.
   - Validation:
     - Use the Durbin-Watson test to check for autocorrelation in the residuals. A value close to 2 indicates no autocorrelation.
     - Alternatively, you can plot residuals over time or order and look for patterns. Randomly scattered residuals indicate independence.

3. Homoscedasticity (Constant Variance of Errors)
   - Check: The variance of the residuals should remain constant across all levels of the independent variables.
   - Validation:
     - Plot the residuals vs. predicted values. The spread of residuals should be constant (i.e., no funnel or cone shape). If there's a funnel shape, it indicates heteroscedasticity.
     - You can also perform the **Breusch-Pagan test or White test** to statistically test for heteroscedasticity.

4. Normality of Residuals
   - Check: The residuals should be approximately normally distributed.
   - Validation:
     - Histogram of residuals: Plot a histogram of the residuals. It should resemble a bell-shaped curve.

- Q-Q plot: Plot a quantile-quantile (Q-Q) plot of the residuals. The points should fall approximately along the diagonal line.
- Shapiro-Wilk test: This statistical test checks for normality. A high p-value (> 0.05) indicates that the residuals are normally distributed.

5. No Multicollinearity

- Check: The independent variables should not be highly correlated with each other.
- Validation:
  - Calculate the Variance Inflation Factor (VIF) for each independent variable. A VIF value > 10 indicates high multicollinearity.
  - You can also look at the correlation matrix of the independent variables. High correlation coefficients (near ±1) suggests multicollinearity.

6. No Endogeneity

- Check: There should be no correlation between the independent variables and the error term.
- Validation:
  - This is a more advanced assumption, often checked using techniques like instrumental variables if you suspect the presence of endogeneity

Conclusion

By validating these assumptions, you can ensure that the linear regression model you've built on your training set is appropriate and robust. If any assumptions are violated, you may need to consider alternative models or techniques to address the issues, such as transforming variables, adding interaction terms, or using a different modeling approach altogether.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

The top 3 features contributing significantly towards explaining the demand for shared bikes are:

1. Year (yr):
    - Coefficient: 0.2311
    - This feature has the highest t-value (28.757), indicating a strong positive relationship with bike demand. As the year advances, bike demand increases significantly.
2. Temperature (temp):
    - Coefficient: 0.4599
    - Temperature is another highly significant factor with a large positive coefficient, meaning that higher temperatures lead to higher bike demand.
3. Weather Situation (weathersit_3):
    - Coefficient: -0.2491
    - This feature has a significant negative coefficient, indicating that certain adverse weather conditions (category 3) greatly reduce bike demand.

These features have strong statistical significance and large coefficient magnitudes, making them the top contributors to explaining the variation in bike demand in the dataset.