Assignment-based Subjective Questions

Q1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

A: We can infer the following from the model prediction of categorical columns:

- **Year**: A significant positive coefficient (2001.89, p-value 0.000) indicates that bike demand has increased over the years.
- **Season**: Demand varies with seasons.
  - Spring: Decreased demand (-1170.13)
  - Winter: Increased demand (494.77), statistically significant.
- **Month**: Specific months show distinct demand patterns.
  - July: Notable decrease in demand (-482.89).
  - September: Significant increase in demand (483.00).
- **Weekday**: Day of the week impacts demand.
  - Sunday: Demand decreases compared to other weekdays (-335.83).
- **Weather**: Weather conditions strongly influence demand.
  - Clear: Demand rises (665.38).
  - Light Snow/Rain: Significant drop in demand (-1639.94).

Conclusion: Bike demand is strongly influenced by year, seasons, specific months, weekdays, and weather conditions.

Q2. Why is it important to use drop_first=True during dummy variable creation?

A. **Using drop_first=True to Avoid the Dummy Variable Trap:**

When creating dummy variables for categorical data, it's crucial to use drop_first=True to avoid the dummy variable trap, which occurs when the variables are highly collinear and lead to redundancy in the model. This redundancy can cause issues with model interpretation and multicollinearity.

For example, for the categorical variable "Season" with categories Light_snowrain, Clear,Heavy_Snowrain and Misty ,creating dummy variables would result in four columns. Including all four introduces redundancy, as knowing any three columns determines the fourth, leading to multicollinearity.

By using drop_first=True, you drop one dummy variable (e.g., Clear_dummy), leaving you with other three. This ensures the remaining variables are independent and prevents multicollinearity, as the dropped variable is implicitly accounted for when the other three variables are 0.

Now, the model can use these without multicollinearity issues.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

A. 'temp' and 'atemp' have the highest correlation and linear relationship with the 'cnt' target variable

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

A. Performed the below steps:

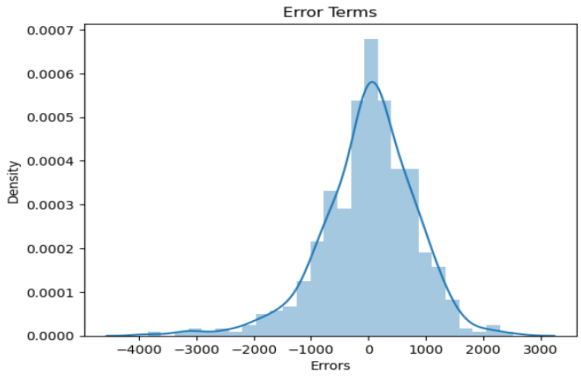
1. **Normality of Residuals**: Plotted a histogram of the residuals to check if they follow a normal distribution. The residuals form a bell-shaped distribution in the histogram and can see that the errors are normally distributed having a mean of 0.

Residual Analysis

```
y_train_pred = lr.predict(X_train[cols])
```

```
#Plot a histogram of the error terms
def plot_res_dist(act, pred):
    sns.distplot(act-pred)
    plt.title('Error Terms')
    plt.xlabel('Errors')
```

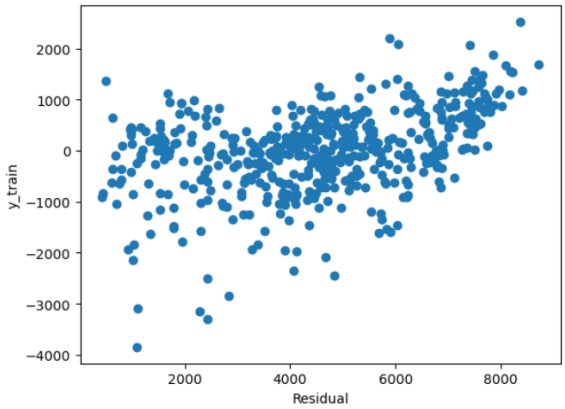
```
plot_res_dist(y_train, y_train_pred)
```



2. **Linearity of the Relationship**: Plotted the residuals (errors) against the predicted values to check for linearity. Verified that the residuals are randomly scattered around zero.This indicates that the model's assumption of a linear relationship between the independent variables and the dependent variable is valid.

Linearity Check

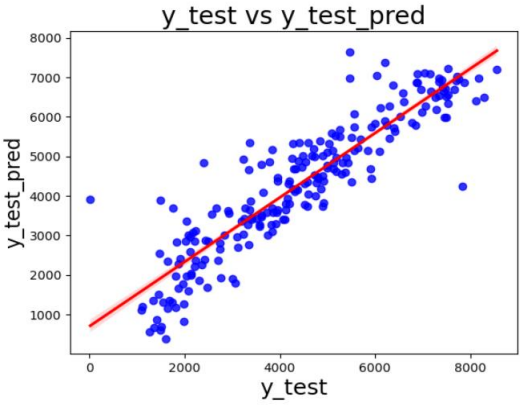
```
# scatter plot for the check
residual = (y_train - y_train_pred)
plt.scatter(y_train,residual)
plt.ylabel("y_train")
plt.xlabel("Residual")
plt.show()
```



3. **Homoscedasticity (Constant Variance of Errors)**: Verify that the error term, or residual, is the same across all values of the independent variables

Homoscedacity ¶

```
plt.figure()
sns.regplot(x=y_test, y=y_test_pred, ci=68, fit_reg=True,scatter_kws={"color": "blue"}, line_kws={"color": "red"})
plt.title('y_test vs y_test_pred', fontsize=20)
plt.xlabel('y_test', fontsize=18)
plt.ylabel('y_test_pred', fontsize=16)
plt.show()
```



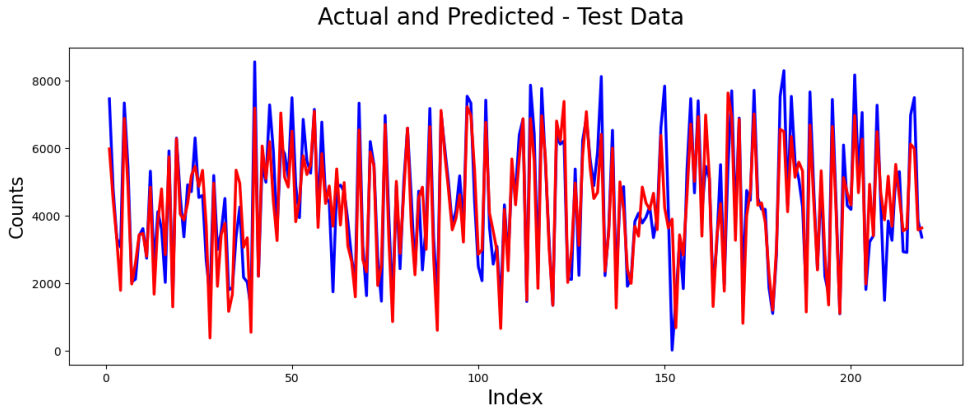
4. **No Multicollinearity**: Variance Inflation Factor (VIF) - Modelled until the VIF for the predictor variables are < 5

	Features	VIF
8	temp	3.86
6	weathersit_Clear	2.73
0	yr	2.06
1	season_spring	1.38
2	mnth_july	1.38
3	season_winter	1.37
4	mnth_sep	1.20
5	weekday_sun	1.18
7	weathersit_Light_snowrain	1.07

5. **Plot test vs Predicted Test Values:** Plotted actual vs Predicted values for the test data and can see that the test data is close to the actuals

```
#Function to plot Actual vs Predicted
#Takes Actual and PRedicted values as input along with the scale and Title to indicate which data
def plot_act_pred(act,pred,scale,dataname):
    c = [i for i in range(1,scale,1)]
    fig = plt.figure(figsize=(14,5))
    plt.plot(c,act, color="blue", linewidth=2.5, linestyle="-")
    plt.plot(c,pred, color="red",  linewidth=2.5, linestyle="-")
    fig.suptitle('Actual and Predicted - '+dataname, fontsize=20)              # Plot heading
    plt.xlabel('Index', fontsize=18)                               # X-Label
    plt.ylabel('Counts', fontsize=16)                              # Y-Label
```

```
#Plot Actual vs Predicted for Test Data
plot_act_pred(y_test,y_test_pred,len(y_test)+1,'Test Data')
```



Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

A. From the model predictions the below features were determined as critical factors for determining demand:

- **Year**: Bike demand increases each year.
- **Season**: Bike demand is generally lower in spring compared to other seasons.
- **Weather**: Clear weather increases bike demand, while light snow or rain decreases it significantly.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

A. Linear regression is a basic statistical method used to model the relationship between a dependent variable (outcome) and one or more independent variables (predictors), aiming to find the best-fitting linear relationship between them.

1. Types of Linear Regression

- **Simple Linear Regression**: Models the relationship between one independent variable (X) and the dependent variable (Y) with a straight line.
- $$Y=\beta0+\beta1X+\epsilon$$

  Where,
  - Y is the dependent variable.
  - X is the independent variable.
  - β0 is the intercept (the value of Y when X=0).
  - β1 is the slope (the change in Y for a one-unit change in X).
  - ϵ is the error term (the difference between the observed and predicted values).
- **Multiple Linear Regression**: Models the relationship between two or more independent variables (X1,X2,…,Xn) and the dependent variable (Y) with a hyperplane in multi-dimensional space.

  $$Y=\beta0+\beta1X1+\beta2X2+…+\beta nXn+\epsilon$$

  Where:
  - X1,X2,…,Xn are the independent variables.
  - β1,β2,…,βnare the coefficients representing the impact of each independent variable on Y.

2. Assumptions of Linear Regression

- **Linearity**: The relationship between the independent and dependent variables is linear.
- **Independence**: Observations are independent of each other.
- **Homoscedasticity**: The variance of the error terms is constant across all levels of the independent variables.
- **Normality**: The residuals (differences between observed and predicted values) are normally distributed.
- **No multicollinearity**: In multiple linear regression, the independent variables should not be too highly correlated with each other.

3. **The Objective Function** The objective of linear regression is to minimize the difference between the actual and predicted values. This difference is measured using a loss function, typically the Mean Squared Error (MSE), defined as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Where:

- yi is the actual value.
- y^i is the predicted value.
- n is the number of observations

Minimizing the MSE involves finding the optimal values of β0,β1,…,βn that minimize this error.

4. **Finding the Coefficients (Ordinary Least Squares - OLS)**

The most common method for finding the coefficients in linear regression is the Ordinary Least Squares (OLS) method. OLS estimates the coefficients by minimizing the sum of the squared differences between the observed and predicted values.

For simple linear regression, the formulas to calculate the coefficients are:

$$\beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\beta_0 = \bar{Y} - \beta_1\bar{X}$$

Where:

$\bar{X}$ and $\bar{Y}$ are the means of the independent and dependent variables, respectively.

5. Evaluating the Model

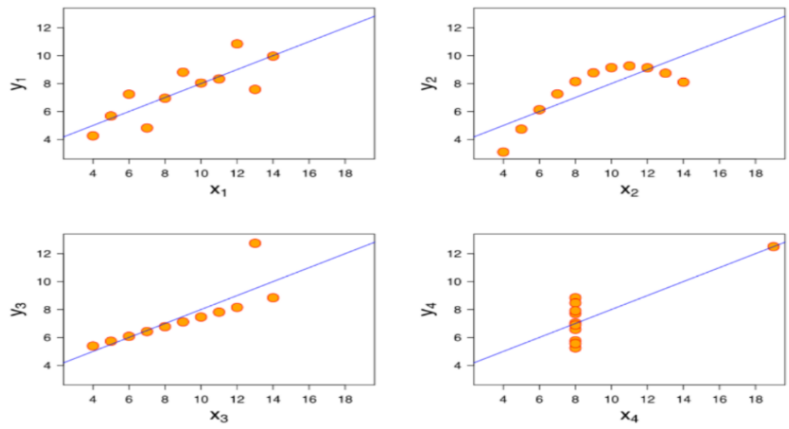
After fitting the model, evaluating its performance is crucial. Common metrics include:

- **R-squared (R^2)**: Measures the proportion of variance in the dependent variable explained by the independent variables, ranging from 0 to 1, with higher values indicating a better fit.
- **Adjusted R-squared**: Adjusts R^2 for the number of predictors, offering a more accurate measure for models with multiple independent variables.
- **Residual Analysis**: Involves checking the residuals (errors) to ensure they are randomly distributed and satisfy the assumptions of linear regression.

2. Explain the Anscombe's quartet in detail.

A. Anscombe's Quartet is a set of four datasets which was designed to demonstrate the importance of data visualization by showing how different datasets can have the same statistical properties but very different distributions when plotted.

For example,



Each dataset contains two variables, x and y, with identical statistical properties:

- Mean and variance of x and y
- Correlation coefficient between x and y
- Slope and intercept of the regression line

Statistical Properties: All datasets have:

- Identical means and variances for x and y.
- Same correlation coefficient.
- Similar regression line slopes and intercepts.

Visual Differences: The datasets exhibit distinct patterns:

- Dataset I: Linear relationship with no outliers.
- Dataset II: Linear relationship with an outlier affecting the regression line.
- Dataset III: Non-linear relationship, curve fit.
- Dataset IV: Linear relationship with an influential outlier distorting the regression line.

In conclusion,

- **Importance of Visualization**: The quartet highlights that summary statistics alone can be misleading. Visualization is essential to understand the true nature of the data.
- **Model Fitting**: Different datasets with the same statistical properties might require different modelling approaches, underscoring the need for graphical analysis before drawing conclusions.

3.What is Pearson's R?

A. Pearson's R, or Pearson correlation coefficient, measures the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to +1:

- **+1:** Perfect positive linear relationship
- **-1:** Perfect negative linear relationship
- **0:** No linear relationship

Key Points:

- **Positive Values** indicate that as one variable increases, the other also tends to increase.
- **Negative Values** indicate that as one variable increases, the other tends to decrease.
- **Magnitude** shows the strength of the correlation: values closer to +1 or -1 represent a stronger relationship, while values near 0 indicate a weaker relationship.

Example:

Suppose you have data on hours studied and exam scores:

Hours Studied	Exam Score
1	50
2	55
3	60
4	65
5	70

To calculate Pearson's R, you follow these steps:

1. Compute the means of both variables:
    - Mean of Hours Studied ($\bar{x}$) = (1 + 2 + 3 + 4 + 5) / 5 = 3
    - Mean of Exam Score ($\bar{y}$) = (50 + 55 + 60 + 65 + 70) / 5 = 60
2. Calculate the covariance between Hours Studied and Exam Score.
3. Calculate the standard deviations of Hours Studied and Exam Score.
4. Apply the Pearson's R formula:

$$r = \frac{\text{Cov}(x, y)}{s_x \cdot s_y}$$

where Cov(x, y) is the covariance between x and y, and sx and sy are the standard deviations of x and y, respectively.

Pearson's R measures the linear relationship:

Positive Value: In this case, Pearson's R is close to +1, indicating a strong positive correlation—more hours studied generally leads to higher exam scores.

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

- Scaling is a data preprocessing technique used to adjust the range and distribution of numerical features in a dataset. It involves transforming the values of features so that they fit within a specific range or follow a particular distribution, which can improve the performance of machine learning algorithms.
- Scaling is performed for several reasons:
  - **Algorithm Performance**: Many machine learning algorithms, especially those that rely on distance calculations (e.g., k-Nearest Neighbour's, Support Vector Machines, Gradient Descent-based methods), are sensitive to the scale of the input data. Features with larger ranges can dominate the calculation, leading to biased models.
  - **Model Convergence**: In optimization algorithms like Gradient Descent, scaling can lead to faster convergence by ensuring that all features contribute equally to the error function.
  - **Improved Interpretability**: Scaling makes coefficients in linear models more interpretable by placing all features on the same scale.
  - **Avoiding Numerical Instability**: Scaling can prevent numerical issues that arise when working with very large or very small values.
- **Difference Between Normalized Scaling and Standardized Scaling:**
  - **Range**: Normalization scales data to a specific range (typically 0 to 1), while standardization scales data based on its mean and standard deviation.
  - **Distribution Assumption**: Normalization does not assume any specific distribution of data, whereas standardization assumes that data follows a normal distribution.
  - **Impact on Data**: Normalization compresses the data to a fixed interval, whereas standardization centers the data around 0 and scales it based on the spread of the data.

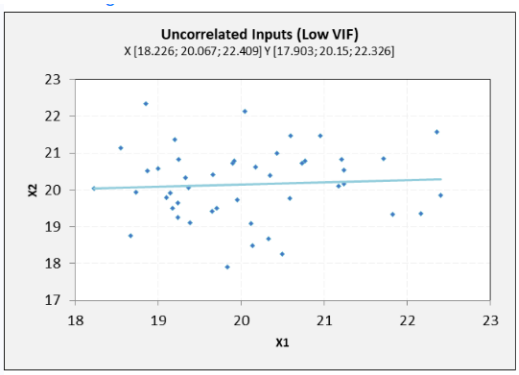
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

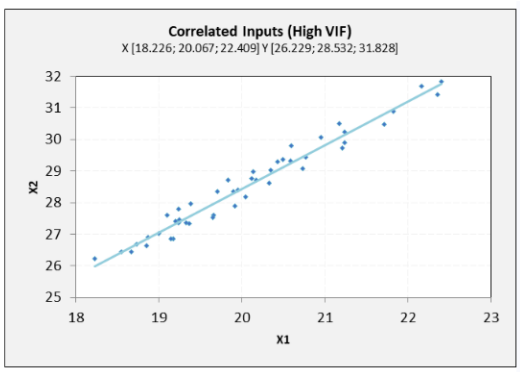
A. A Variance Inflation Factor (VIF) value can become infinite when there is perfect multicollinearity among the predictor variables in a regression model. This means that one predictor variable can be expressed as a perfect linear combination of other predictor variables.

Reasons for Infinite VIF:

1. **Perfect Multicollinearity:**
   - When one predictor variable is a linear combination of others, the correlation between predictors becomes perfect (correlation coefficient of ±1), leading to infinite VIF values for those predictors.
2. **Dummy Variable Trap:**
   - In categorical data, including all dummy variables (one-hot encoding) without dropping one category can cause perfect multicollinearity. This is known as the dummy variable trap.
3. **Redundant Variables:**

- Including variables that are highly redundant or calculated from others (e.g., including both "total_income" and "base_salary" + "bonus" if the latter sum is equal to the former) can also lead to high or infinite VIF values.





Impact and Resolution:

- **Impact:** Infinite VIF values indicate that the model is overfitted due to multicollinearity, which can cause unstable estimates and unreliable statistical inferences.
- **Resolution:** To resolve issues with infinite VIF values, you can:
  - Remove or combine redundant variables.
  - Perform feature selection to reduce multicollinearity.
  - Use dimensionality reduction techniques like Principal Component Analysis (PCA).

In summary, infinite VIF values occur due to perfect multicollinearity, where predictors are linearly dependent on each other, making it impossible to separate their individual effects in the regression model.

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

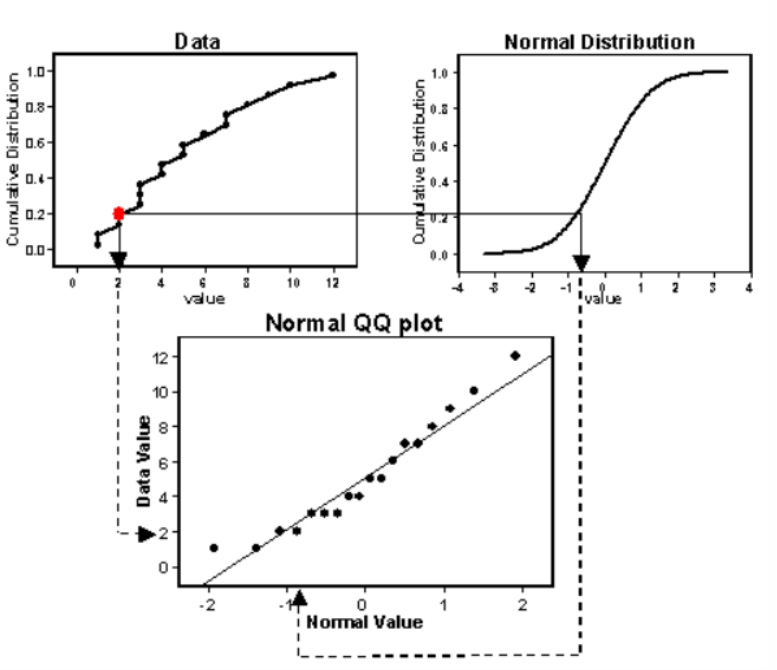
A. A Q-Q plot compares the quantiles of a dataset with the quantiles of a theoretical distribution (often normal).

Use in Linear Regression:

1. **Check Residual Normality:** Assesses if residuals are normally distributed, an important assumption for valid inferences and hypothesis testing.
2. **Model Diagnostics:** Identifies deviations from normality, which can indicate issues like non-linearity or heteroscedasticity.
3. **Goodness-of-Fit:** Evaluates how well residuals fit the normal distribution, enhancing model reliability.

Example:

- **Normal Distribution:** Points lie on a straight line, indicating normality.
- **Non-Normal Distribution:** Points deviate from the line, suggesting departures from normality.


Data
Cumulative Distribution
value
Normal Distribution
Cumulative Distribution
value
Normal QQ plot
Data Value
Normal Value