######################## #######################################

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?

Ans)

To analyze the effect of categorical variables on the dependent variable, we can break it down into a few key steps. Typically, this is done using exploratory data analysis (EDA) and statistical techniques like cross-tabulation, chi-square tests, and visualizations (such as bar charts or box plots) to determine relationships between the categorical variables and the dependent variable.

Here's a general approach to analyzing the effect of categorical variables:

1. Cross-Tabulation

- Method: For each categorical variable, create a cross-tabulation or contingency table to observe how the values of the categorical variables are distributed across the dependent variable categories.
- Inference: This helps in identifying patterns or imbalances. For example, if you are looking at default status (dependent variable), you can see if certain categories (e.g., "employment type") have higher rates of default.

2. Chi-Square Test of Independence

- Method: Perform a chi-square test to assess if there is a statistically significant relationship between a categorical variable and the dependent variable.
- Inference: If the test shows a significant result (p-value < 0.05), this suggests that the categorical variable might influence the dependent variable. For example, loan type could significantly affect the likelihood of default.

3. Bar Charts and Count Plots

- Method: Use bar charts or count plots to visualize the distribution of the dependent variable across different categories.
- Inference: Visualizing the distribution can give a clear idea of the relationship. For instance, if a particular category (e.g., "marital status") leads to a significantly higher proportion of defaults, it indicates that the category may be a key driver of the dependent variable.

4. Categorical Encoding (One-Hot or Label Encoding)

- Method: Encode categorical variables and then use logistic regression or decision trees to assess the effect of these encoded variables on the dependent variable.
- Inference: The importance of each categorical feature can be gauged from the model's coefficients (in regression) or feature importance scores (in decision trees or random forests)

5. Box Plots (for ordinal categories)

- Method: For ordinal categorical variables, box plots can show how the dependent variable (if continuous) varies across different categories.
- Inference: If there is a clear trend or difference in the distribution across categories, it suggests an influence on the dependent variable. For example, income brackets might show a clear difference in loan default likelihood.

### Sample Inference:

- If you have education level as a categorical variable and loan default as the dependent variable, you might find that people with higher education levels tend to default less. This would indicate that education level has a negative correlation with loan default.

The exact inference depends on your dataset and the specific variables in question, but these are general approaches used to infer the effect of categorical variables on a dependent variable.

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

Using `drop_first=True` in `get_dummies()` is important in dummy variable creation to prevent multicollinearity when performing
regression analysis or machine learning algorithms. Heres why:

1. Multicollinearity:

- When you have multiple categories for a categorical variable (e.g., "red," "blue," "green" for color), `get_dummies()` will create a separate dummy variable for each category.
- If all categories are represented by dummy variables, one of the dummies can always be perfectly predicted by the others. This results in perfect multicollinearity, where one variable is a linear combination of others.
- Multicollinearity can distort statistical tests and make it difficult to interpret the coefficients in regression models.

2. Redundant Information:

- When you include all dummy variables, one is redundant. For example, if a categorical variable has three levels (A, B, C),
creating three dummies means if you know two of the dummy values, you can infer the third.
- Example: If a variable takes the values A, B, or C, and you create three dummy variables:
  - A: [1, 0, 0]
  - B: [0, 1, 0]
  - C: [0, 0, 1]
- The third column can be inferred if you have the first two, leading to redundancy.

3. drop_first=True:

- When `drop_first=True`, pandas drops the first dummy variable and only creates (k-1) dummy variables for k categories.
- This removes the redundancy and solves the multicollinearity issue.
- The dropped category is treated as a reference category, and the remaining dummies represent how the other categories differ
from that reference.

```
import pandas as pd
# Example DataFrame
df = pd.DataFrame({
   'color': ['red', 'blue', 'green', 'blue', 'green']
})

# Without drop_first
print(pd.get_dummies(df))

# With drop_first
print(pd.get_dummies(df, drop_first=True))

#Output without `drop_first=True`:
|   | color_blue | color_green | color_red |
|---|------------|-------------|-----------|
| 0 | 0          | 0           | 1         |
| 1 | 1          | 0           | 0         |
| 2 | 0          | 1           | 0         |
| 3 | 1          | 0           | 0         |
| 4 | 0          | 1           | 0         |

#Output with `drop_first=True`:

|   | color_blue | color_green |
|---|------------|-------------|
| 0 | 0          | 0           |
| 1 | 1          | 0           |
| 2 | 0          | 1           |
| 3 | 1          | 0           |
| 4 | 0          | 1           |
```

In this case, "red" is the reference category, and the remaining dummies represent how "blue" and "green" differ from "red."

### Conclusion:

Using `drop_first=True` simplifies your model by avoiding redundant information and prevents multicollinearity, improving model interpretability and efficiency.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Ans)
Based on the pair-plot you provided, the relationship between the variables can be observed visually. The target variable seems to be "cnt" (likely representing a count of some event). From the scatter plots:

- The variable "temp" (temperature) shows the strongest positive linear relationship with the target "cnt." This can be seen from the diagonal pattern in the scatter plot between "cnt" and "temp."

- Similarly, the variable "atemp" (which may represent apparent temperature) also shows a strong correlation with "cnt." Among these two, "temp" appears to have the strongest correlation with the target, based on visual inspection.

Q4) Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Ans)
variables year , season/ weather situation and month are significant in predicting the demand for shared bikes .

Q5) How did you validate the assumptions of Linear Regression after building the model on the training set?
Ans)
After building a Linear Regression model, its crucial to validate the assumptions of the model to ensure its accuracy and
generalizability. The following are the key assumptions of Linear Regression and common techniques to validate them

1. Linearity of the relationship between features and target:
- Assumption: The dependent variable (target) should have a linear relationship with each independent variable (features).
- How to validate:
  - Residual Plot: Plot residuals (difference between observed and predicted values) versus the predicted values. The residuals should be randomly scattered around zero, without any distinct patterns (e.g., curved or funnel-shaped).
- Scatter Plot: Plot the features against the target variable to visually check for linear relationships.
  - Partial Regression Plots: These help to visualize the effect of each predictor on the target while keeping other variables constant.Corrective Action: Apply transformations (e.g., log, polynomial features) if the relationships are non-linear.

2. Homoscedasticity (constant variance of errors):
- Assumption: The variance of the residuals should remain constant across all levels of predicted values.
- How to validate:
  - Residual Plot: Look for patterns in the residual plot. If the residuals show a "funnel" shape (i.e., the variance increases or decreases as the predicted values increase), it indicates heteroscedasticity.

Corrective Action: Apply transformations to the dependent variable (e.g., log transformation) or use models that can handle heteroscedasticity (e.g., Generalized Least Squares).

3. Independence of errors:
- Assumption: The residuals (errors) should be independent of each other, meaning there is no autocorrelation.

- How to validate:
  - Durbin-Watson Test: This statistical test detects the presence of autocorrelation in the residuals. A value close to 2 indicates no autocorrelation, while values close to 0 or 4 suggest positive or negative autocorrelation, respectively.
  - Plot Residuals over Time: If your data is time-based (e.g., time series), plot the residuals over time to detect any patterns (which indicate dependence).

Corrective Action: If autocorrelation is present, you might need to use time-series-specific techniques such as ARIMA models.

4. Normality of residuals:

- Assumption: The residuals should be normally distributed.
- How to validate:
  - Histogram or Q-Q Plot: Plot a histogram of the residuals or use a Q-Q plot (Quantile-Quantile plot) to check if the residuals follow a normal distribution. In a Q-Q plot, the points should fall along the 45-degree reference line if the residuals are normally distributed.
  - Shapiro-Wilk Test: A formal statistical test for normality. However, this test can be overly sensitive in large datasets, so visual inspections are often more practical.

Corrective Action: If the residuals are not normally distributed, you may need to apply transformations to the target variable (e.g., log transformation).

###
###### General Subjective Questions ################
###

Q1). Explain the linear regression algorithm in detail. (4 marks)
Ans)

Linear Regression is a simple yet powerful algorithm used to model the relationship between a dependent variable (target) and one or more independent variables (predictors). The goal of linear regression is to find a linear relationship between the input variables
and the output. Let’s break down the concept and details of Linear Regression:

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x).
So, this regression technique finds out a linear relationship between x (input) and y (output).
Hence, the name is Linear Regression.
Linear Regression may further divided into

1. Simple Linear Regression/ Univariate Linear regression
2. Multivariate Linear Regression

Q2. Explain the Anscombe’s quartet in detail. (3 marks)
Ans)

Anscombe’s Quartet is a group of four datasets that have nearly identical summary statistics but exhibit strikingly different distributions and visual patterns when plotted. It was created by statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data before analyzing it and relying solely on summary statistics.

### Key Concepts Illustrated by Anscombe’s Quartet:

1. Importance of Data Visualization: Anscombe’s Quartet shows that similar summary statistics (like mean, variance, correlation) can hide significant differences in data patterns. This emphasizes that relying solely on statistics can be misleading, and visualizations like scatter plots reveal important insights.
2. Limitations of Summary Statistics: The quartet highlights how datasets can have the same statistical measures (e.g., mean, standard deviation, correlation) but differ in structure. This demonstrates that statistics alone may not capture the underlying patterns in the data.

### Lessons from Anscombe’s Quartet:

1. Visual Analysis is Essential: Even though summary statistics are the same, visualizations like scatter plots reveal the true nature of the relationships between variables.

2. Impact of Outliers: Outliers can heavily influence regression models and summary statistics. Detecting and handling them properly is crucial.

3. Non-Linearity: Regression lines and correlation measures assume linearity, but real-world data often follow non-linear
patterns, which can only be identified through plotting.

Q3. What is Pearson's R?(3 marks)
Ans)
Pearson's R, also known as the Pearson correlation coefficient (denoted as \(r \)), is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It was developed by Karl Pearson in the early 20th century.
The Pearson correlation coefficient (r) is the most widely used correlation coefficient and is known by many names

- Pearson's r
- Bivariate correlation
- Pearson product-moment correlation coefficient (PPMCC)
- The correlation coefficient

The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset.
Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

Although interpretations of the relationship strength (also known as effect size) vary between disciplines, the table below gives general rules of thumb:

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

Scaling is the process of transforming the features of your data to a similar range or standard so that they can be compared on the same scale. In machine learning, many algorithms are sensitive to the magnitude of the features. Therefore, scaling ensures that the data is treated uniformly and that no feature dominates due to its range of values.

### Why is Scaling Performed?

1. Improves Model Performance:
   - Some machine learning algorithms like k-nearest neighbors (KNN), support vector machines (SVM), and gradient descent-based algorithms (e.g., linear regression, logistic regression) are sensitive to the scale of features.
   - When features are on different scales, models that use distance-based metrics may give undue importance to variables with larger ranges.

2. Ensures Faster Convergence:
   - Algorithms that involve gradient descent optimization converge faster when features are scaled since large differences in feature magnitudes can cause the model to take longer to optimize.

3. Prevents Dominance of Certain Features:
   - Without scaling, features with a large range may dominate the model, leading to biased predictions.

Difference Between Normalized Scaling and Standardized Scaling

#### 1. Normalized Scaling (Min-Max Scaling):

- Definition: Normalization transforms the data into a fixed range, usually between 0 and 1. It is calculated as:
- Range: The resulting values lie between 0 and 1 (or any predefined range like [-1, 1]).

- When to Use:
   - Normalization is preferred when the distribution of the data is not Gaussian (non-normal) or when the algorithm does not make any assumption about the distribution of data (e.g., KNN, neural networks).
   - It is useful when you need to preserve the relative relationships between data points (e.g., the ratio of values).

#### 2. Standardized Scaling (Z-score Standardization):

- Definition: Standardization scales data to have a mean of 0 and a standard deviation of 1. It is calculated as:
- Range: The values are not bounded within any specific range. Standardized values typically range between -3 and +3, but they can
exceed this based on the distribution.

- When to Use:
   - Standardization is preferred when the data follows a normal (Gaussian) distribution and when algorithms assume a standard distribution of the input data (e.g., linear regression, logistic regression, SVM, K-means).
   - It is often used in cases where the algorithms assumption includes a normal distribution, or when the spread of data across different features is important.

Q5) You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
Ans)

The Variance Inflation Factor (VIF) measures how much the variance of a regression coefficient is inflated due to multicollinearity between the independent variables. In cases where VIF becomes infinite, it indicates a perfect multicollinearity between one or more independent variables. This means that one predictor variable is a perfect linear combination of the other predictor(s), making it impossible to compute a unique solution for the regression coefficients.

### Why VIF Becomes Infinite:

1. Perfect Multicollinearity:

- If one independent variable is a perfect linear function of another (or a combination of others), the VIF for that variable will be infinite. For example, if:

2. Singular Matrix in Regression:

- When perfect multicollinearity exists, the design matrix (matrix of predictor variables) in a regression model becomes singular (non-invertible). This leads to an inability to compute the inverse of the matrix, which is required to estimate regression coefficients. As a result, the VIF calculation breaks down, and the value is effectively infinite.

### How to Handle Infinite VIF:

1. Remove Redundant Variables: Identify and remove variables that are perfect linear combinations of other variables.
2. Principal Component Analysis (PCA): Use dimensionality reduction techniques like PCA to transform correlated variables into a set of uncorrelated components.
3. Regularization: Techniques like Ridge Regression can help mitigate the effects of multicollinearity by adding a penalty for large coefficients.

Q6) What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
Ans)

A Q-Q (Quantile-Quantile) Plot is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, typically the normal distribution. It helps assess whether a dataset follows a particular distribution by plotting the quantiles of the dataset against the quantiles of the theoretical distribution.

- X-axis: Theoretical quantiles (from the theoretical distribution, usually normal).
- Y-axis: Sample quantiles (from the observed data).

If the data follows the theoretical distribution, the points in the Q-Q plot will roughly lie along a 45-degree line.
Deviations from this line indicate departures from the expected distribution.

#### Importance of Q-Q Plot in Linear Regression:

1. Assess Normality of Residuals:

- The most crucial assumption in linear regression is that the residuals (differences between observed and predicted values) are normally distributed. By plotting the residuals in a Q-Q plot, you can visually assess whether they deviate from normality.
- Why is this important?: If residuals are not normally distributed, the confidence intervals, hypothesis tests (like t-tests),and p-values for the regression coefficients may become invalid.

2. Detect Skewness and Kurtosis:

- A Q-Q plot helps you detect skewness (asymmetry) or heavy/light tails in the residuals. Non-normally distributed residuals might suggest that a different model is more appropriate or that transformations (e.g., logarithmic or square root) should be applied to the data

3. Diagnose Model Fit:

- If the residuals are not normally distributed, this could indicate that the linear model does not fit the data well, and alternative models (such as polynomial regression or non-linear models) may need to be considered.

4. Identifying Outliers:

- A Q-Q plot can reveal outliers that do not follow the distribution of the rest of the data. If a few points deviate significantly from the line, it may indicate influential outliers that could skew the regression model.