Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 1 goes below this line> (Do not edit)

To infer the effects of categorical variables on the dependent variable (cnt for bike demand), we would typically analyse the distribution of bike demand across each category. Here's how each categorical variable might affect bike demand:

1. **Season** (season):Different seasons are expected to show varying levels of demand for bike rentals. For instance, the demand could be higher in summer or fall when weather conditions are more favourable for outdoor activities and lower in winter due to cold temperatures.
2. **Year** (yr):This variable indicates whether the data is from 2018 (0) or 2019 (1). As bike-sharing services grow in popularity, the demand in 2019 might be higher than in 2018, indicating a positive trend over time.
3. **Month** (mnth):Monthly variations might also affect bike demand. Summer months like June, July, and August might see higher demand than winter months, especially in areas with distinct seasonal weather patterns.
4. **Holiday** (holiday):Holidays might see lower demand if fewer people commute to work, or higher demand if people tend to rent bikes for leisure activities.
5. **Weekday** (weekday):Weekday variations could show differences in demand, with weekends possibly seeing more leisure-related rentals, while weekdays might have higher demand related to commuting.
6. **Working Day** (workingday):Bike demand might be higher on working days as people use bikes for commuting purposes. Weekends and holidays might show different trends in demand, leaning more towards leisure rides.
7. **Weather Situation** (weathersit):Weather conditions play a significant role in outdoor activities. For example, clear or partly cloudy days might have higher bike demand, while days with mist, rain, or snow might see a reduction in rentals.

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 2 goes below this line> (Do not edit)

Using drop_first=True during dummy variable creation is important to avoid the **dummy variable trap** and ensure that our regression model remains interpretable and does not encounter multicollinearity issues. Here's why it matters:

1. **Avoiding Multicollinearity**:

    - When we create dummy variables, each category of a categorical feature is represented by a separate binary (0 or 1) column. For example, if a feature has 3 categories, we end up with 3 columns.
    - However, this introduces **redundancy**: knowing the values of the first two columns automatically tells us the value of the third column. Including all dummy variables would cause multicollinearity, where predictor variables are highly correlated with each other.
    - To avoid this issue, we drop one dummy column (using drop_first=True), which acts as a **reference category**. The remaining columns will contain enough information to represent all categories.
2. **Simplifying Model Interpretation**:
    - By dropping the first dummy variable, the model interprets the coefficients of the remaining dummy variables relative to the omitted (dropped) category. This makes the model easier to interpret, as each coefficient represents the effect of being in that category versus the reference category.
3. **Preventing Overfitting in Linear Models**:
    - Including all dummy variables might lead to an overfitted model because the regression algorithm might assign weights to redundant features, leading to over-reliance on specific categories.
    - By dropping one dummy, we eliminate the redundancy and allow the model to generalize better.

In summary, drop_first=True helps in:

- Avoiding multicollinearity
- Simplifying model interpretation
- Reducing the risk of overfitting

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)

From analysing bike-sharing datasets, we often find that **temperature (temp)** tends to have the highest positive correlation with cnt (bike demand). This is because warmer temperatures generally encourage outdoor activities, which can increase bike rentals.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

From analysing bike-sharing datasets, we often find that **temperature (temp)** tends to have the highest positive correlation with cnt (bike demand). This is because warmer temperatures generally encourage outdoor activities, which can increase bike rentals.

Steps to Verify Using Pair-Plot or Correlation Matrix:

1. **Pair-Plot**:
    - By creating a pair-plot, we can visualize scatter plots of each numerical variable against cnt. The strength of the linear relationship (how tightly the points cluster around a line) gives a visual indication of correlation.
    - In the pair-plot, if temp vs. cnt shows a strong upward trend, it suggests a high positive correlation.
2. **Correlation Matrix**:
    - Calculate the correlation matrix to get the exact correlation values between cnt and other numerical variables.
    - The correlation with the highest absolute value (closest to 1) indicates the strongest relationship with cnt.

Code Example:

Here's how you can check the correlation directly:

```
# Calculate correlation matrix
correlation_matrix = data.corr()

# Extract correlations with the target variable 'cnt'
correlation_with_cnt = correlation_matrix['cnt'].sort_values(ascending=False)
print(correlation_with_cnt)
```

Expected Outcome:

In most cases, temp or atemp (feeling temperature) typically shows the highest positive correlation with cnt. Humidity (hum) and windspeed (windspeed) might have weaker negative correlations, as extreme humidity or high winds can reduce outdoor activities.

Validating the assumptions of Linear Regression is essential to ensure that the model is appropriate for the data and that its predictions are reliable. Here's a common approach to validate these assumptions after building the model on the training set:

1. Linearity Assumption

- **Goal**: Ensure that the relationship between each predictor and the target variable is approximately linear.
- **Validation**:
    - Plot the residuals (errors) vs. the predicted values. If the residuals are randomly scattered around zero without any clear pattern, it indicates that the linearity assumption holds.
    - Alternatively, plot each independent variable against the target variable cnt to see if there's a linear trend.

```
# Residuals vs Predicted values plot
plt.scatter(y_train_pred, residuals)
plt.axhline(0, color='red', linestyle='--')
plt.xlabel('Predicted Values')
plt.ylabel('Residuals')
plt.title('Residuals vs Predicted Values')
plt.show()
```

2. Independence of Errors

- **Goal**: The residuals (errors) should be independent of each other.
- **Validation**:
    - Check the Durbin-Watson statistic (close to 2 indicates no significant autocorrelation). This is particularly important in time-series data but is often considered in non-time-series as well.

```
from statsmodels.stats.stattools import durbin_watson
durbin_watson_value = durbin_watson(residuals)
print("Durbin-Watson statistic:", durbin_watson_value)
```

3. Homoscedasticity (Constant Variance of Errors)

- **Goal**: Ensure that the variance of errors remains constant across all levels of the predicted values.
- **Validation**:
    - Plot the residuals vs. predicted values. If the spread of residuals remains constant across the range of predictions, homoscedasticity holds. If there's a funnel shape (increasing or decreasing spread), it indicates heteroscedasticity.

```
plt.scatter(y_train_pred, residuals)
plt.axhline(0, color='red', linestyle='--')
plt.xlabel('Predicted Values')
plt.ylabel('Residuals')
plt.title('Residuals vs Predicted Values (Checking Homoscedasticity)')
plt.show()
```

4. Normality of Errors

- **Goal**: The residuals should be approximately normally distributed, which ensures reliable confidence intervals and hypothesis tests.
- **Validation**:
    - Plot a histogram or Q-Q plot of the residuals. In a Q-Q plot, if the residuals follow a straight line, they are approximately normally distributed.
    - Perform the Shapiro-Wilk or Kolmogorov-Smirnov test if a formal test is needed (though in large samples, slight deviations from normality are acceptable).

```
# Q-Q plot
import scipy.stats as stats
stats.probplot(residuals, dist="norm", plot=plt)
plt.title('Normal Q-Q Plot')
plt.show()

# Histogram of residuals
plt.hist(residuals, bins=30)
plt.xlabel('Residuals')
plt.title('Histogram of Residuals')
plt.show()
```

5. Multicollinearity

- **Goal**: Ensure that predictors are not highly correlated with each other, which can destabilize the model and make coefficient estimates unreliable.
- **Validation**:
    - Calculate the Variance Inflation Factor (VIF) for each predictor. A VIF value above 5 (or sometimes 10) indicates high multicollinearity.

```
from statsmodels.stats.outliers_influence import variance_inflation_factor

# Calculate VIF for each feature
vif_data = pd.DataFrame()
vif_data["feature"] = X_train.columns
vif_data["VIF"] = [variance_inflation_factor(X_train.values, i) for i in range(X_train.shape[1])]

print(vif_data)
```

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

In a typical bike-sharing demand dataset, here are the features that often show strong influence:

1. Temperature (temp or atemp)

- Temperature is usually one of the strongest predictors of bike demand, as warmer weather encourages outdoor activities. Higher temperatures generally correlate positively with an increase in bike rentals.

2. Year (yr)

- Since bike-sharing systems tend to grow in popularity over time, the year feature (with values indicating 0 for 2018 and 1 for 2019) often has a strong impact on demand. The coefficient for this variable typically shows an increase in bike rentals in the more recent year, reflecting a growing user base.

3. Season (season)

- The season can significantly affect demand patterns, with certain seasons (like summer and fall) generally seeing higher demand compared to others (like winter). This feature helps capture these seasonal variations in bike rentals.

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Linear Regression is a statistical technique used to explore the relationship between one dependent variable (what we want to predict) and one or more independent variables (the predictors). It aims to find a straight line (or a linear relationship) that best describes how changes in the predictors affect the target variable.

The primary goal of Linear Regression is to create a model that can accurately predict the target variable based on the input features. It does this by finding the optimal line (or plane, in the case of multiple variables) that minimizes the distance between the actual values and the predicted values. This line represents the best approximation of the relationship in the data.

Linear Regression uses a method called **Ordinary Least Squares (OLS)** to find the best-fit line. The OLS method works by looking for the line that has the smallest total of squared differences between the actual data points and the points on the line. These squared differences are known as "errors" or "residuals." By minimizing these errors, Linear Regression finds the line that best captures the trend in the data.

Linear Regression relies on several assumptions to ensure that the model is reliable and interpretable:

- Linearity: The relationship between each predictor and the target variable should be linear (i.e., can be represented with a straight line).

- Independence of Errors: The errors (differences between actual and predicted values) should be independent of each other.

- Constant Variance (Homoscedasticity): The spread of the errors should be consistent across all values of the predictors.

- Normality of Errors: The errors should follow a normal distribution.

- No Multicollinearity: In the case of multiple predictors, they should not be highly correlated with each other to avoid redundancy and instability in the model.

The output of a Linear Regression model includes coefficients for each predictor. These coefficients tell us the direction and strength of the relationship between each predictor and the target variable. For example, a positive coefficient means that as the predictor increases, the target variable also tends to increase. A larger absolute value of the coefficient indicates a stronger effect on the target.

Linear Regression is widely used due to its simplicity and interpretability, especially as a starting point in predictive modelling. However, it has limitations, such as assuming a linear relationship and requiring the assumptions mentioned above. When these assumptions don't hold, other techniques may be more appropriate.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Anscombe's Quartet is a set of four different datasets created by British statistician **Francis Anscombe** in 1973. These datasets demonstrate the importance of visualizing data before interpreting it based solely on summary statistics. Each of the four datasets in the quartet has nearly identical statistical properties, including the mean, variance, correlation coefficient, and linear regression line, but they look vastly different when graphed. This highlights how relying only on numerical summaries can be misleading and reinforces the value of data visualization.

Key Characteristics of Anscombe's Quartet

Each dataset in Anscombe's Quartet has:

1. The same **mean** for the x and y values.
2. The same **variance** for the x and y values.
3. The same **correlation coefficient** between x and y (around 0.816).
4. The same **linear regression line** ($y = 3 + 0.5x$).

Despite having these identical summary statistics, the four datasets are visually very different and each illustrates different kinds of data patterns and potential issues in data interpretation.

The Four Datasets in Anscombe's Quartet

Let's break down each dataset and understand what makes it unique:

1. **Dataset 1**: A typical linear relationship
    - In the first dataset, the points are spread around a straight line. This dataset looks like what you might expect for a typical linear regression, where the relationship between x and y can be reasonably modeled with a straight line.
    - Visualizing Dataset 1 would show a strong linear trend with points scattered relatively close to the regression line.
2. **Dataset 2**: A nonlinear relationship
    - In the second dataset, the points form a curve rather than a straight line. However, because the regression line calculation assumes a linear relationship, it still provides a line as the best-fit model.
    - In reality, a curved or polynomial regression line would better represent this data. This dataset shows that summary statistics can suggest a linear relationship even when the true relationship is nonlinear.

3. **Dataset 3**: An outlier that heavily influences the line
    - The third dataset has most of its points aligned vertically, with one extreme outlier that influences the regression line.
    - This outlier dramatically affects the slope and position of the line, misleading the interpretation of the relationship between x and y. This dataset demonstrates how outliers can skew results, even when the summary statistics appear consistent.
4. **Dataset 4**: Vertical line with no real relationship
    - In the fourth dataset, all data points except one are nearly aligned in a vertical column, with one point far away from the others.
    - Here, there is essentially no meaningful linear relationship between x and y, but the summary statistics still match those of the other datasets. This dataset highlights how summary statistics like correlation can be meaningless if the data pattern doesn't support it.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Pearson's R, also known as the **Pearson correlation coefficient** or simply the **correlation coefficient**, is a statistical measure that describes the **strength and direction of the linear relationship between two continuous variables**. Developed by **Karl Pearson**, it's widely used to quantify the association between variables in various fields, such as social sciences, finance, and natural sciences.

Key Characteristics of Pearson's R

1. **Range**:
    - Pearson's R ranges from -1 to +1.
    - An R value of **+1** indicates a perfect positive linear relationship (as one variable increases, the other also increases in a perfectly predictable way).
    - An R value of **-1** indicates a perfect negative linear relationship (as one variable increases, the other decreases in a perfectly predictable way).
    - An R value of **0** indicates no linear relationship between the variables (although they might still have a nonlinear relationship).
2. **Interpretation**:
    - **Positive values** (e.g., 0.5, 0.7) mean that as one variable increases, the other also tends to increase, suggesting a positive relationship.
    - **Negative values** (e.g., -0.5, -0.7) mean that as one variable increases, the other tends to decrease, suggesting a negative relationship.
    - The closer the value of R is to ±1, the **stronger** the linear relationship between the two variables.
3. **Symmetric Measure**:
    - Pearson's R is symmetric, meaning that the correlation between variable X and variable Y is the same as the correlation between Y and X.
4. **Linear Relationships**:

- Pearson's R only measures **linear** relationships. It does not capture **nonlinear** associations between variables. Therefore, if two variables have a curved relationship, Pearson's R may be close to 0, even if they are strongly related in a non-linear way.

Calculation (Conceptually)

The calculation of Pearson's R is based on the **covariance** between the two variables, normalized by the product of their standard deviations. This normalization allows Pearson's R to be **dimensionless** (meaning it has no units) and makes it comparable across different datasets.

Example Interpretation

- **R = +0.9**: Strong positive relationship. As one variable increases, the other tends to increase as well.
- **R = -0.8**: Strong negative relationship. As one variable increases, the other tends to decrease.
- **R = +0.2**: Weak positive relationship, indicating a slight tendency for one variable to increase as the other does.
- **R = 0**: No linear relationship. Changes in one variable are not associated with predictable changes in the other.

Assumptions for Using Pearson's R

1. **Linearity**: The relationship between the variables should be linear.
2. **Interval or Ratio Scale**: The variables should be continuous and measured on an interval or ratio scale.
3. **Normality**: The data should ideally be normally distributed, especially for small sample sizes.
4. **Homogeneity of Variances**: Variability in scores should be similar across the range of values for each variable.

Importance of Pearson's R

Pearson's R is essential in statistics as it provides a simple, interpretable measure of the relationship between two variables, helping researchers and analysts understand patterns, make predictions, and test hypotheses. However, because it only measures linear relationships, it's often used in combination with visualizations (like scatter plots) to understand the true nature of the data.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Scaling refers to the process of adjusting the range and distribution of data features in a dataset. This is often done to prepare data for various machine learning algorithms that perform better or converge faster when features are on a similar scale and close to a normal distribution.

Why is Scaling Performed?

1. **Improves Algorithm Performance**: Many machine learning algorithms, such as k-nearest neighbors (KNN) and gradient descent-based algorithms, are sensitive to the scale of input features. Scaling helps improve their performance and accuracy.
2. **Enhances Convergence Speed**: In optimization algorithms, scaling can help speed up the convergence process, leading to faster training times.
3. **Prevents Dominance of Features**: If features have different units or scales, those with larger ranges can dominate the distance calculations or optimization processes, skewing the results.

Types of Scaling

1. **Normalized Scaling (Min-Max Scaling)**:
    - This method rescales the feature to a fixed range, typically [0, 1].
    - It is sensitive to outliers; if there are extreme values, they can significantly affect the scaled values.
2. **Standardized Scaling (Z-score Scaling)**:
    - This method rescales the feature to have a mean of 0 and a standard deviation of 1, effectively creating a standard normal distribution.
    - It is less sensitive to outliers than min-max scaling and is often preferred when the data is normally distributed or when the distribution is unknown.

Key Differences

- **Range**: Normalized scaling adjusts values to a specific range (e.g., 0 to 1), while standardized scaling adjusts values based on the mean and standard deviation.
- **Outlier Sensitivity**: Normalization can be heavily influenced by outliers, whereas standardization is more robust against them.
- **Distribution**: Normalized data will always be bounded within the range, while standardized data can take on any value, depending on the original distribution.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

The Variance Inflation Factor (VIF) measures how much the variance of an estimated regression coefficient increases when your predictors are correlated. A VIF value is

calculated for each predictor variable, indicating how much the variance of a regression coefficient is inflated due to multicollinearity.

Infinite VIF: Causes

An infinite VIF occurs under specific conditions:

1. **Perfect Multicollinearity**:
    - This is the most common reason for an infinite VIF. It happens when one predictor variable is an exact linear combination of one or more other predictor variables.
2. **Insufficient Data**:
    - If the dataset has very few observations relative to the number of predictors, it can lead to perfect multicollinearity, making it impossible to estimate coefficients for all predictors accurately.
3. **Dummy Variables**:
    - In categorical data, creating dummy variables can sometimes lead to multicollinearity if one category can be perfectly predicted from others. For example, if you have a categorical variable with three categories and you create two dummy variables, the third category can be perfectly inferred (i.e., it acts as a linear combination of the first two).

Implications of Infinite VIF

When a variable has an infinite VIF, it indicates that it provides no unique information beyond what is already provided by other variables in the model. This can cause instability in the estimation of regression coefficients and lead to unreliable statistical inferences.

To address infinite VIF values, you can:

- Remove one of the perfectly collinear variables.
- Combine collinear variables into a single predictor through techniques like PCA (Principal Component Analysis).
- Ensure proper data pre-processing and variable selection to avoid introducing multicollinearity.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
(Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

A **Q-Q (Quantile-Quantile) plot** is a graphical tool used to assess if a dataset follows a particular theoretical distribution, most commonly the normal distribution. It compares the quantiles of the dataset against the quantiles of the expected distribution.

Construction of a Q-Q Plot

- The x-axis represents the quantiles of the theoretical distribution (e.g., normal distribution).
- The y-axis represents the quantiles of the observed data.
- If the points on the Q-Q plot fall approximately along a straight line (usually the 45-degree line), it indicates that the observed data distribution closely matches the theoretical distribution.

Uses of a Q-Q Plot in Linear Regression

1. **Assessing Normality of Residuals**:
    - In linear regression, one of the key assumptions is that the residuals (the differences between observed and predicted values) should be normally distributed. A Q-Q plot can visually assess this assumption. If the residuals appear linear in the Q-Q plot, this suggests that they are normally distributed.
2. **Identifying Deviations from Normality**:
    - The Q-Q plot helps identify any deviations from normality, such as skewness or kurtosis. For instance:
        - If the points curve upwards or downwards at the ends, it indicates that the data may have heavier or lighter tails than the normal distribution (indicative of outliers or extreme values).
        - If there is a systematic departure from the line, it suggests a violation of the normality assumption.
3. **Model Diagnostics**:
    - By examining the Q-Q plot of residuals, you can diagnose potential issues in the model. Non-normal residuals can indicate problems like model misspecification or the need for transformation of the response variable.

Importance of a Q-Q Plot in Linear Regression

- **Validation of Assumptions**: The normality of residuals is critical for hypothesis testing and confidence intervals in regression analysis. A Q-Q plot provides a straightforward way to validate this assumption.
- **Improving Model Quality**: Identifying deviations from normality allows for model refinement, such as applying transformations (e.g., logarithmic, square root) or considering non-linear models, thus enhancing model performance.
- **Visualization**: Q-Q plots offer a visual representation that can be easier to interpret than statistical tests for normality, providing a quick assessment of how well the data adheres to the normal distribution.