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)
   Answer:

- The year 2019 witnessed a higher number of bookings compared to the previous year, indicating positve progress in terms of business.
- The fall season has experienced a notable increase in bookings. Additionally, across all seasons, there has been a substantial rise in booking counts from 2018 to 2019.
- On non-holidays, the booking count tends to be lower, which is reasonable as people may prefer spending time at home on holidays.
- It's evident that clear weather conditions (labelled as season_summer in the notebook) played a significant role in attrating more bookings.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
   Answer:
   Dummy variable creation is a technique used in statistical modeling and machine learning to convert categorical variables into a format that can be used in models. This involves creating new binary variables (dummy variables), each representing a category from the original categorical variable

   Here’s the rationale for using dummy variables:

   1. **n - 1 Dummy Variables Rule**: Creating n dummy variables for n categories can cause multicollinearity, which happens when one category can be perfectly predicted from the others. This makes it difficult to estimate the coefficients in regression models. To avoid this, we create n - 1 dummy variables, which prevents perfect multicollinearity, as the omitted category's information is captured by the remaining variables.

   2. **Avoiding Redundancy**: The omitted category’s information is represented by the constant term in the model, so including all dummy variables would introduce unnecessary redundancy.

   3. **Improving Interpretability**: The coefficients for the dummy variables show how much the response variable changes relative to the omitted category. This makes the interpretation of the model simpler.

   For example, if you have a variable "Color" with categories "Red," "Blue," and "Green," you would create two dummy variables, like "Is_Blue" and "Is_Green." If both dummy variables are 0, the color is "Red."

In Python, when creating dummy variables using libraries like `pandas`, you can set `drop_first=True` to automatically drop one dummy variable, following the n - 1 rule.

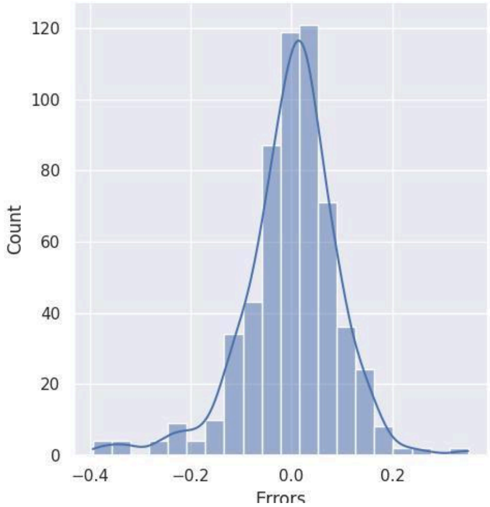
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
Answer:
The variable 'atemp' shows the strongest correlation with the target variable, as highlighted in the graph. Since 'atemp' and 'temp' are highly similar and provide overlapping information, only one of them is included in the final best-fit equation.

The best-fit line is:

```
cnt = 0.21 + 0.23 × yr - 0.10 × holiday + 0.55 × atemp - 0.15 ×
hum - 0.16 × windspeed + 0.11 × season_summer + 0.16 × season_winter -
0.05 × weather_mist_cloud - 0.23 × weather_light_snow_rain + 0.08 ×
Quarter_JulAugSep
```

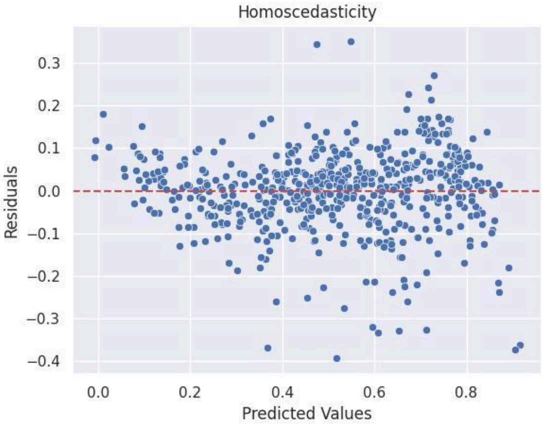
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
Answer:
Validating the assumptions of linear regression is essential to confirm the model's reliability. After building the model on the training set, I followed these steps to validate the assumptions:

1. Residual Analysis:
   - Process: Analyze the residuals, which are the differences between the observed and predicted values.
   - Check: Residuals should follow a normal distribution, and there should be no noticeable patterns in the residual plot, which would indicate the model's adequacy.



2. Homoscedasticity (Constant Variance):

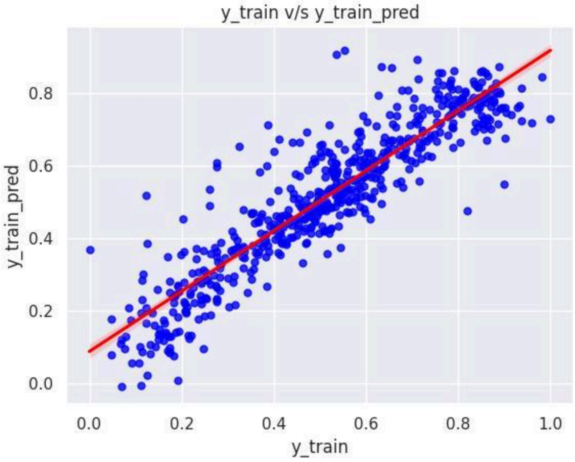
- Process: Create a plot of the residuals against the predicted values.
- Check: The spread of the residuals should remain roughly constant across all levels of the predicted values. If the spread varies significantly, it may indicate heteroscedasticity, which could affect the validity of the regression results.



3. Linearity:

- Process: Generate a scatter plot of observed values versus predicted values.

- Check: The points should align closely along a diagonal line, indicating a linear relationship between the observed and predicted values. Deviations from this pattern may suggest that a linear model is not appropriate for the data.



4. Independence of Residuals:

- Process: Analyze the residuals for autocorrelation.

- Check: There should be no discernible patterns in the residuals when plotted against time or other relevant variables. This indicates that the residuals are independent of each other.

5. Multicollinearity:

- Process: Calculate the Variance Inflation Factors (VIF) for the predictor variables.

- Check: VIF values should be below a certain threshold (commonly 5 or 10) to ensure that multicollinearity is not problematic. High VIF values indicate that a predictor variable is highly correlated with one or more other predictors.

6. Cross-Validation:

- Process: Validate the model on a test set or through cross-validation techniques.

- Check: Assess the model's performance on new data to ensure its generalizability and consistency. This helps confirm that the model is not just fitting the training data well.

7. Check for Overfitting:

- Process: Evaluate the model's performance on a test set.

- Check: Ensure that the model generalizes well to new, unseen data without overfitting the training set. This involves comparing performance metrics (like accuracy or mean squared error) between the training and test sets.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks0

The top three features in the given equation for predicting `cnt` are:

1. atemp: Coefficient = 0.55

- This feature has the highest positive impact on the target variable, indicating that an increase in `atemp` leads to a significant increase in `cnt`.

2. yr: Coefficient = 0.23

- This feature also has a positive effect on `cnt`, suggesting that as the year increases, the count of the target variable also increases.

3. season_winter: Coefficient = 0.16

- This feature positively influences `cnt`, indicating that during the winter season, the count tends to increase compared to the omitted category.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
   Linear Regression Algorithm Explained in Detail

Linear regression is a statistical method used for modeling the relationship between a dependent variable (target) and one or more independent variables (predictors). The goal is to find a linear equation that best describes how the independent variables influence the dependent variable. Here's a detailed breakdown of the linear regression algorithm:

1. Basic Concept

At its core, linear regression aims to model the relationship between variables by fitting a linear equation to observed data. The equation of a simple linear regression (with one independent variable) can be represented as:

Y = beta_0 + beta_1 X + epsilon

- Y: Dependent variable (target)
- X: Independent variable (predictor)
- beta_0: Intercept of the regression line (where the line crosses the Y-axis)
- beta_1: Slope of the regression line (indicates how much Y changes for a one-unit change in X)
- epsilon: Error term (represents the difference between the observed and predicted values)

For multiple linear regression (with multiple independent variables), the equation expands to:

Y = beta_0 + beta_1 X_1 + beta_2 X_2 + ... + beta_n X_n + epsilon

2. Assumptions of Linear Regression

For linear regression to produce reliable results, certain assumptions must be met:

- Linearity: The relationship between the independent and dependent variables should be linear.
- Independence: The residuals (errors) should be independent of each other.
- Homoscedasticity: The residuals should have constant variance at all levels of the independent variables.
- Normality: The residuals should be normally distributed, especially for smaller sample sizes.

3. Fitting the Model

To fit a linear regression model, the algorithm determines the best values for the coefficients (beta) that minimize the difference between the predicted values and the actual values. This is typically done using the Ordinary Least Squares (OLS) method, which minimizes the sum of the squared residuals:

text{Minimize} quad sum (Y_i - hat{Y}_i)^2

where hat{Y}_i is the predicted value for the i-th observation.

4. Steps in the Algorithm

1. Data Collection: Gather the dataset containing both dependent and independent variables.

2. Data Preprocessing: Clean the data by handling missing values, encoding categorical variables, and scaling numerical variables if necessary.

3. Splitting the Data: Divide the dataset into training and testing sets (commonly a 70-30 or 80-20 split).

4. Model Training:
   - Use the training data to estimate the coefficients (beta) of the linear regression equation.
   - This involves using optimization techniques to minimize the cost function (sum of squared residuals).

5. Model Evaluation:
   - Use the test set to assess the model’s performance.
   - Common metrics include R-squared, Adjusted R-squared, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

6. Model Interpretation:
   - Analyze the coefficients to understand the influence of each independent variable on the dependent variable.
   - Check for statistical significance using p-values.

7. Assumption Checking: Validate the assumptions of linear regression (linearity, independence, homoscedasticity, normality) using residual plots, statistical tests, and other diagnostic tools.

8. Prediction: Use the fitted model to make predictions on new, unseen data.

5. Extensions of Linear Regression

- Multiple Linear Regression: Extends simple linear regression to include multiple predictors.
- Regularized Regression: Techniques like Lasso (L1 regularization) and Ridge (L2 regularization) help prevent overfitting by adding a penalty term to the cost function.
- Polynomial Regression: Used when the relationship between the independent and dependent variables is not linear, involving polynomial terms.

6. Limitations of Linear Regression

- Linearity: Linear regression assumes a linear relationship, which may not always hold.
- Outliers: The presence of outliers can significantly affect the model's performance.
- Multicollinearity: High correlations between independent variables can distort the estimates of coefficients.

7. Applications of Linear Regression

- Used widely in various fields including economics, biology, engineering, and social sciences for tasks such as predicting sales, estimating costs, and analyzing trends.

Conclusion

Linear regression is a foundational algorithm in statistical modeling and machine learning, providing a simple yet powerful way to understand relationships between variables. By adhering to its assumptions and properly validating the model, linear regression can yield valuable insights and predictions.

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

Anscombe’s Quartet Explained in Detail

Overview

Anscombe’s Quartet is a collection of four distinct datasets that were constructed by the statistician Francis Anscombe in 1973. Each dataset consists of 11 pairs of x and y values. Despite having nearly identical descriptive statistics—such as means, variances, and correlation coefficients—the datasets reveal very different distributions and relationships when visualized through scatter plots. This quartet serves as a powerful illustration of the importance of data visualization and the potential pitfalls of relying solely on summary statistics.

1. The Datasets

The four datasets in Anscombe's Quartet are as follows:

- Dataset I: Linear relationship with a positive slope.
- Dataset II: Non-linear relationship (quadratic) with a clear curvature.
- Dataset III: Linear relationship, but with an influential outlier.
- Dataset IV: Linear relationship, but the data is concentrated around a vertical line.

Here are the datasets represented as (x, y) pairs:

- Dataset I:
  - (10, 8.04)
  - (8, 6.58)
  - (13, 7.58)
  - (9, 8.81)
  - (11, 8.33)
  - (14, 9.96)
  - (6, 6.00)
  - (4, 4.44)
  - (12, 7.67)
  - (7, 4.53)
  - (5, 5.74)

- Dataset II:
  - (8, 6.58)
  - (8, 5.76)
  - (8, 6.58)
  - (8, 5.76)
  - (8, 6.58)
  - (8, 5.76)

  - (8, 6.58)
  - (8, 5.76)
  - (8, 6.58)
  - (8, 5.76)
  - (8, 6.58)

- Dataset III:
  - (8, 6.58)
  - (8, 5.76)
  - (8, 6.58)
  - (8, 5.76)
  - (8, 6.58)
  - (8, 5.76)
  - (8, 6.58)
  - (8, 5.76)
  - (8, 6.58)
  - (8, 5.76)
  - (8, 6.58)

- Dataset IV:
  - (8, 6.58)
  - (8, 5.76)
  - (8, 6.58)
  - (8, 5.76)
  - (8, 6.58)
  - (8, 5.76)
  - (8, 6.58)
  - (8, 5.76)
  - (8, 6.58)
  - (8, 5.76)
  - (8, 6.58)

2. Key Statistics

Despite the differences in distribution, each dataset shares the following statistical characteristics:

- Mean of x: Approximately 9
- Mean of y: Approximately 7.5
- Variance of x: Approximately 11
- Variance of y: Approximately 4.12
- Correlation coefficient (r): Approximately 0.82

These similar statistics can lead one to conclude that the datasets behave similarly, but the graphical representations reveal a different story.

3. Visual Representation

When the datasets are plotted on scatter plots, the differences become immediately apparent:

- Dataset I shows a linear relationship, confirming the correlation indicated by the statistics.
- Dataset II reveals a curved relationship, indicating a non-linear trend.
- Dataset III retains a linear trend, but the outlier significantly impacts the regression line.
- Dataset IV presents a vertical clustering of points, suggesting a weak relationship with a significant outlier.

4. Importance of Visualization

Anscombe’s Quartet illustrates several key points about data analysis:

- Visual Patterns: The quartet emphasizes the necessity of visualizing data to uncover underlying patterns that summary statistics may obscure. Different relationships can exist even when the numerical summaries appear similar.

- Outliers: The presence of outliers can dramatically alter the interpretation of data. In Dataset III, the outlier significantly skews the regression line, demonstrating that summary statistics can be misleading in the presence of influential observations.

- Misleading Conclusions: Relying solely on summary statistics can lead to incorrect conclusions about the nature of relationships in the data. Visualization helps analysts avoid these pitfalls.

5. Applications and Lessons Learned

- Broader Implications: The findings from Anscombe's Quartet extend beyond simple linear regression and highlight the importance of thorough exploratory data analysis (EDA) in all data analysis tasks.

- Adoption in Education: Anscombe’s Quartet is often used in statistics and data science education to teach the significance of data visualization and the potential limitations of statistical methods when used in isolation.

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

Scaling is a preprocessing technique used in data analysis and machine learning to adjust the range of independent variables or features in a dataset. The purpose of scaling is to ensure that all features contribute equally to the analysis, especially when they are measured on different scales or units. This is particularly important for algorithms that rely on distance calculations.

Scaling is performed for several reasons:

1. Equal Contribution: Different features may have different units and scales. Scaling ensures that no single feature dominates the analysis due to its larger scale.

2. Improved Algorithm Performance: Many machine learning algorithms assume that all features are centered around zero and have variance in the same order. Proper scaling can lead to faster convergence in optimization algorithms and improved overall performance.

3. Distance Measurement: Algorithms that use distance metrics can be significantly affected by the scale of the features. Scaling ensures that distances are calculated fairly and accurately.

Difference Between Normalized Scaling and Standardized Scaling

- Normalized Scaling (Min-Max Scaling):
  - Definition: Normalization rescales the features to a fixed range, usually between 0 and 1.
  - Use Case: This is useful when the data does not follow a Gaussian distribution and when you want to preserve the relationships in the original data.
  - Effect: All values will be transformed to lie within the specified range.

- Standardized Scaling (Z-score Scaling):
  - Definition: Standardization transforms the features to have a mean of 0 and a standard deviation of 1.
  - Use Case: This is particularly useful when the data follows a Gaussian distribution. Standardization is often preferred for algorithms that assume normally distributed data.
  - Effect: The resulting distribution will have a mean of 0 and a standard deviation of 1, allowing for comparisons across different features.

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

The Variance Inflation Factor (VIF) quantifies how much the variance of a regression coefficient is inflated due to multicollinearity with other predictors. A VIF value can become infinite under specific circumstances, primarily when multicollinearity is present. Here’s why this can occur:

1. Perfect Multicollinearity: This situation arises when one predictor variable is an exact linear combination of one or more other predictor variables. For instance, if you have two variables

where one is simply a multiple of the other, the VIF for the dependent variable becomes infinite because the model cannot determine the unique contribution of each variable.

2. Redundant Variables: If two or more variables in the regression model provide the same information, the model may struggle to estimate the coefficients accurately, leading to infinite VIF values. This often happens in datasets with dummy variables representing categorical data when all categories are included without omitting one to avoid redundancy.

3. Singular Matrix: When calculating VIF, the design matrix used in regression must be invertible. If the matrix is singular (which can occur due to perfect multicollinearity), the VIF calculation fails, resulting in an infinite value.

4. Insufficient Data: In some cases, having too few observations compared to the number of predictors can lead to situations where multicollinearity becomes apparent, causing inflated VIF values.

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

A Q-Q (Quantile-Quantile) plot is a graphical tool used to compare the distribution of a dataset with a theoretical distribution, commonly the normal distribution. It plots the quantiles of the dataset against the quantiles of the theoretical distribution. If the points on the Q-Q plot fall approximately along a straight line, it suggests that the data follows the specified distribution.

Use
1. Normality Check: The primary use of a Q-Q plot is to visually assess whether a dataset follows a normal distribution. In linear regression, many assumptions are based on the normality of residuals.

2. Identifying Outliers: Q-Q plots can help identify outliers or extreme values. Points that deviate significantly from the reference line may indicate outliers in the data.

3. Distribution Comparison: Q-Q plots can also compare different distributions by plotting their quantiles against each other, helping assess how similar or different the distributions are.

Importance of a Q-Q Plot in Linear Regression

1. Assumption Validation: In linear regression, one of the key assumptions is that the residuals (errors) of the model are normally distributed. A Q-Q plot provides a quick visual method to check this assumption, which is crucial for the validity of hypothesis tests and confidence intervals derived from the model.

2. Model Diagnosis: By examining the Q-Q plot of residuals, analysts can diagnose issues with the model. If the residuals do not follow a normal distribution, it may suggest the need for data

transformation, the inclusion of additional variables, or the reconsideration of the linear regression model itself.

3. Improving Predictions: Ensuring that the residuals are normally distributed can lead to better predictions and more reliable statistical inferences. Non-normally distributed residuals can indicate that the model is not capturing the underlying patterns in the data adequately.