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: Based on the analysis and visualization of the categorical variables, we can infer the following effects on the dependent variable, which represents the bike-sharing count:

Year (yr):The variable yr shows a strong positive influence on the bike-sharing count. This likely reflects a general increase in bike usage over the years, possibly due to growing popularity, infrastructure improvements, or an expanding user base.

Season (`season_winter`, `season_summer`, `season_spring`):

- **Winter (season_winter)**: This variable shows a negative coefficient, indicating that bike usage is lower during the winter months. This is expected as colder weather conditions typically deter outdoor activities.
- **Summer (season_summer)** and **Spring (season_spring)**: Both variables exhibit a positive effect on bike usage. Warmer weather and longer days during these seasons likely encourage more people to use bikes, thus increasing the count.

Weather Situation (weathersit_Light Snow/Rain, weathersit_Mist):

- **Light Snow/Rain (weathersit_Light Snow/Rain)**: This variable has a significant negative impact, suggesting that adverse weather conditions, such as light snow or rain, substantially decrease bike usage. These conditions can make biking less safe or enjoyable, thus reducing the count.
- **Mist (weathersit_Mist)**: This variable also negatively impacts bike usage, though to a lesser extent than Light Snow/Rain. Misty conditions may reduce visibility or create slick surfaces, which can deter biking.

These inferences align with common sense expectations about the factors influencing bike-sharing usage, where favorable weather and seasonal conditions generally boost activity, while adverse weather conditions reduce it.

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

Answer: When dealing with categorical variables, creating dummy variables transforms these into binary columns representing the presence or absence of each category. If there are k categories, this transformation will produce k

dummy variables. However, using all k dummy variables along with an intercept in a regression model will cause perfect multicollinearity (also known as the "dummy variable trap").

To avoid this, one category is dropped by setting `drop_first=True`. This action reduces the number of dummy variables to `k-1`, which removes the redundancy and allows the model to properly estimate the effects of each category compared to the dropped category (which serves as a reference group). This approach maintains the model's interpretability and statistical stability.

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

cnt	1.000000
atemp	0.629747
temp	0.626290
yr	0.569728
mnth	0.278191
season_summer	0.145325
weekday	0.067534
season_winter	0.064619
workingday	0.062542
holiday	-0.068764
hum	-0.098060
weathersit_Mist	-0.170686
windspeed	-0.233517
weathersit_Light Snow/Rain	-0.240602
season_spring	-0.561702

Based on the correlation values, the variable **`atemp` (feels-like temperature)** has the highest positive correlation with the target variable `cnt` (total bike rentals), with a correlation coefficient of **`0.629747`**. *This suggests that as the feels-like temperature increases, the number of bike rentals tends to increase as well, indicating a strong positive relationship between these variables.*

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 involves checking several key conditions to ensure the model is appropriate for the data. Here are the steps typically taken:

Linearity of the Relationship:

- The relationship between the independent variables and the dependent variable should be linear. This can be checked by plotting residuals vs. fitted values. There should be no obvious patterns, indicating a linear relationship.

Normality of Residuals:

- The residuals (differences between observed and predicted values) should be approximately normally distributed. This can be verified using a Q-Q plot (quantile-quantile plot) or by inspecting the histogram of residuals. Any significant deviations from normality might suggest issues with the model.

Homoscedasticity:

- The residuals should have constant variance (homoscedasticity). This can be checked by plotting the residuals against the fitted values. The spread of residuals should be consistent across all levels of the fitted values. A funnel shape indicates heteroscedasticity, which can affect the model's efficiency.

Absence of Multicollinearity:

- Multicollinearity occurs when independent variables are highly correlated with each other, which can inflate the variance of coefficient estimates. This can be checked using Variance Inflation Factor (VIF) values. Typically, a VIF value above 10 indicates significant multicollinearity.

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

Answer:

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

yr: This feature indicates the year and has the highest positive contribution, suggesting a significant increase in bike demand over the years.

season_summer: The summer season is associated with a higher demand for shared bikes, indicating more people use bikes during warmer weather.

temp: Temperature also positively correlates with bike demand, implying that better weather conditions encourage more bike usage.

General Subjective Questions

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

Answer: Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The goal is to find the best-fitting line that predicts the dependent variable based on the values of the independent variables. Here's a detailed explanation of the algorithm:

Basic Concept

Linear regression assumes a linear relationship between the dependent variable Y and one or more independent variables X. The general form of a linear regression model with one independent variable is:

$Y=\beta 0 +\beta 1 X+\epsilon$

Where:

- Y is the dependent variable (target variable).
- X is the independent variable (predictor).
- $\beta 0$ is the intercept (constant term).
- $\beta 1$ is the coefficient of the independent variable.
- ϵ is the error term, representing the difference between the observed and predicted values.

For multiple linear regression, with more than one independent variable, the model becomes:

$Y=\beta 0 +\beta 1 X1 +\beta 2 X2 +...+\beta n Xn +\epsilon$

Fitting the Model

The coefficients $\beta 0 ,\beta 1 ,...,\beta n$ are estimated using the method of least squares, which minimizes the sum of the squared differences between the observed values and the predicted values of the dependent variable. The objective is to minimize the cost function:$J(\beta)=2m1 \sum i=1m (Yi -Y^i)2$

Where:

- m is the number of observations.
- Yi is the actual value.
- Y^i is the predicted value.

This is solved by taking partial derivatives of the cost function with respect to each coefficient and setting them to zero, resulting in normal equations that can be solved to find the coefficients.

Assumptions of Linear Regression

To apply linear regression, several assumptions must be met:

- **Linearity**: The relationship between the independent and dependent variables is linear.
- **Independence**: The residuals (errors) are independent.
- **Homoscedasticity**: The residuals have constant variance at every level of the independent variable.
- **Normality**: The residuals are normally distributed.
- **No Multicollinearity**: Independent variables are not highly correlated with each other.

Evaluation Metrics

The performance of a linear regression model can be evaluated using several metrics:

- **R-squared (R2)**: Indicates the proportion of the variance in the dependent variable that is predictable from the independent variables.
- **Mean Squared Error (MSE)**: The average of the squared differences between the observed and predicted values.
- **Mean Absolute Error (MAE)**: The average of the absolute differences between the observed and predicted values.

Applications

Linear regression is widely used in various fields, including economics, biology, engineering, and social sciences, for tasks such as predicting outcomes, understanding relationships between variables, and making decisions based on data analysis.

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

Answer: Anscombe's Quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet they appear very different when graphed. The quartet was constructed by the statistician Francis Anscombe in 1973 to illustrate the importance of graphing data before analyzing it. Here's a detailed explanation of each dataset in the quartet:

Dataset 1:

- **Description:** This dataset represents a simple linear relationship. When you plot it, you see a clear linear trend.
- **Statistics:** The mean of the x-values and y-values, the correlation between them, and the regression line parameters (slope and intercept) are all quite similar to the other datasets.
- **Graph**: The scatter plot shows a linear pattern, with data points closely following a straight line.

Dataset 2:

- **Description**: This dataset also appears to have a linear relationship, but there's a single outlier which greatly affects the regression line.
- **Statistics:** The descriptive statistics (mean, variance, correlation, etc.) are similar to those of Dataset 1, yet the presence of the outlier makes the regression line different.
- **Graph:** The scatter plot shows a clear linear trend, but the outlier skews the regression line compared to Dataset 1.

Dataset 3:

- **Description:** This dataset exhibits a parabolic relationship rather than a linear one. The relationship between x and y is quadratic.
- **Statistics**: The mean, variance, and correlation are again similar to the other datasets, but these descriptive statistics do not capture the non-linear pattern.
- **Graph:** The scatter plot reveals a curve, showing a non-linear relationship that a simple linear regression line would not adequately describe.

Dataset 4:

- **Description:** This dataset has a linear relationship with a significant vertical variance, which means it has a large amount of variability in y-values that is not explained by x.

- **Statistics:** Despite having similar statistical properties, the graph shows a significant spread in y-values around the regression line.
- **Graph:** The scatter plot illustrates a linear trend, but there is a lot of vertical spread around the line.

Key Takeaways:

- **Statistical Summaries are Insufficient:** The datasets in Anscombe's Quartet have nearly identical means, variances, and correlations, yet their scatter plots show very different patterns. This demonstrates that relying solely on descriptive statistics can be misleading.

- **Importance of Data Visualization:** Visualizing data through scatter plots (or other types of plots) can reveal underlying patterns or anomalies that statistical summaries alone might miss.

- **Modeling Implications:** These examples show that different types of data relationships require different analytical approaches. For instance, a linear regression model might fit well for some datasets but not for others (e.g., a quadratic relationship).

Anscombe's Quartet underscores the necessity of exploring and understanding your data through visualization before diving into statistical analysis and modeling.

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

Answer: Pearson's R, also known as Pearson's correlation coefficient, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It's denoted as \(r \) and provides insights into how well the data points fit a straight line when plotted.

Key Points About Pearson's R:

- **Definition:**

- Pearson's R is calculated as the covariance of the two variables divided by the product of their standard deviations. It essentially measures the degree to which two variables change together.

- **Formula:**

  The formula for Pearson's R is:

  $$r = \mathrm{Cov}(X,Y) / \sigma X \, \sigma Y$$

  Where:

  - Cov(X, Y) is the covariance between variables X and Y.
  - σX and σY are the standard deviations of X and Y, respectively.
- **Range and Interpretation:**

  Range: r ranges from -1 to 1.
  - r =1 : Perfect positive linear relationship. As one variable increases, the other also increases in perfect proportion.
  - r = -1 : Perfect negative linear relationship. As one variable increases, the other decreases in perfect proportion.
  - r = 0 : No linear relationship. Changes in one variable do not predict changes in the other.

- **Strength of the Correlation:**
  - **0.70 to 1.00 or -0.70 to -1.00 :** Strong positive or negative linear relationship.
  - **0.40 to 0.69 or -0.40 to -0.69 :** Moderate positive or negative linear relationship.
  - **0.10 to 0.39 or -0.10 to -0.39 :** Weak positive or negative linear relationship.
  - **0.00 to 0.09 or -0.00 to -0.09 :** Very weak or no linear relationship.

- **Assumptions:**
  - **Linearity:** Pearson's R measures the strength of a linear relationship, so it assumes that the relationship between the variables is linear.
  - **Normality:** It assumes that the data follows a normal distribution.
  - **Homogeneity of Variance:** It assumes that the variance around the regression line is similar across values of the predictor variable.
- **Limitations:**
  - **Sensitivity to Outliers:** Pearson's R can be heavily influenced by outliers, which might skew the correlation coefficient.
  - **Does Not Imply Causation:** Even if Pearson's R indicates a strong correlation, it does not imply a causal relationship between the variables.

  In summary, Pearson's R is a widely used metric for evaluating the strength and direction of a linear relationship between two continuous variables, providing valuable insight into their association.

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

Answer: Scaling is a preprocessing technique used in data analysis and machine learning to adjust the range or distribution of numerical features. Scaling is essential because it helps to ensure that features contribute equally to the model and improves the performance and stability of many algorithms.

Why Scaling is Performed:

- **Uniformity:** Features measured on different scales can distort the results of algorithms that rely on distance calculations, such as k-nearest neighbors (KNN) or clustering algorithms.
- **Convergence:** Some optimization algorithms, like gradient descent, converge faster if the features are scaled to similar ranges.
- **Improved Performance:** Scaling can enhance the performance of algorithms by ensuring that all features are treated with equal importance

- **Types of Scaling:**

Normalized Scaling:

- **Definition:** Normalization adjusts the range of the feature values to a specific range, usually between 0 and 1.
- **Formula:**

x_{norm} = (x-min(x))/max(x)-min(x)

Where x is the original feature value, and min(x) and max(x) are the minimum and maximum values of the feature.

- **Purpose**: Normalization is used when the data needs to be scaled to a bounded range, which is useful for algorithms sensitive to the scale of features, like neural networks.

- **Standardized Scaling:**
- **Definition:** Standardization transforms the features to have a mean of 0 and a standard deviation of 1. This scaling method is also known as z-score normalization.
- **Formula:**

$$x_{std} = (x - \mu)/\sigma$$

Where x is the original feature value, μ is the mean of the feature,and σ is the standard deviation of the feature.

- **Purpose:** Standardization is useful when features have different units or when the data follows a normal distribution. It is often used in algorithms that assume normally distributed data or when the feature scaling needs to be less sensitive to outliers.

Key Differences:

- **Range**:
  - **Normalization:** Scales the feature values to a fixed range, typically [0, 1]. It is bounded and often used when the feature needs to be in a specific range.
  - **Standardization**: Scales the feature values to have a mean of 0 and a standard deviation of

  1. It is unbounded and maintains the original data distribution's shape.
- **Application Context:**
  - **Normalization:** More suitable for algorithms that assume bounded input, such as neural networks and some distance-based algorithms.
- **Standardization:** More suitable for algorithms that assume normally distributed data or when features have different units, such as linear regression and support vector machines.

In summary, scaling is crucial for ensuring features contribute equally and improving algorithm performance. Normalized scaling adjusts the range of feature values, while standardized scaling transforms feature values to have a mean of 0 and a standard deviation of 1. The choice between normalization and standardization depends on the algorithm used and the nature of the data.

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

Answer: Variance Inflation Factor (VIF) is a metric used to quantify the extent of multicollinearity in regression models. Multicollinearity occurs when two or more predictors in a regression model are highly correlated, which can lead to unreliable coefficient estimates and reduced interpretability.

Infinite VIF Values:

A VIF value can become infinite in specific cases, typically related to perfect multicollinearity. Here's why this happens:

- **Perfect Multicollinearity:**

- **Definition**: Perfect multicollinearity occurs when one predictor variable is a perfect linear combination of one or more other predictor variables in the model.
- **Cause of Infinite VIF:** When there is perfect multicollinearity, the matrix used to calculate the VIF becomes singular, meaning it cannot be inverted. Since the VIF is computed as:

$VIF_i = 1/(1 - R_i^2)$

where R_i^2 is the coefficient of determination of the regression of the i-th variable on all other variables, a perfect linear relationship between predictors causes R_i^2 to be 1. This results in:

$VIF_i = 1/(1-1) = 1/0 = \infty$

- **Redundant Predictors:**
  - **Definition**: Redundant predictors are those that provide no new information because they are linearly dependent on other predictors.
  - **Cause of Infinite VIF:** If a predictor is redundant (e.g., it is a perfect duplicate of another predictor), its VIF will be infinite because it cannot be separated from the other predictors' effects.

- **Implications**:
  - **Model Issues:** An infinite VIF indicates that the predictor is collinear to an extent that causes numerical instability in the model, making it impossible to estimate its effect reliably.
  - **Model Correction:** To address infinite VIF values, you typically need to remove or combine collinear predictors to reduce multicollinearity. Techniques such as principal component analysis (PCA) or regularization methods (e.g., Lasso or Ridge regression) can also help manage multicollinearity.

Summary

Infinite VIF values occur due to perfect multicollinearity, where one predictor is a perfect linear combination of others, leading to a singular matrix in the VIF calculation. This results in R^2_i, being 1, causing VIF to be undefined (infinite). Addressing such issues involves removing redundant predictors or using techniques to manage multicollinearity.

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

Answer: A Q-Q plot, or Quantile-Quantile plot, is a graphical tool used to assess if a dataset follows a particular theoretical distribution, commonly the normal distribution. It compares the quantiles of the dataset's distribution to the quantiles of a theoretical distribution.

What is a Q-Q Plot?

- **Definition:**

  A Q-Q plot is a scatter plot where the quantiles of the sample data are plotted against the quantiles of a theoretical distribution (e.g., normal distribution). If the sample data follows the theoretical distribution, the points will approximately lie on a straight line.

- **Construction:**
  - **Quantiles:** Quantiles are values that divide the dataset into intervals with equal probabilities. For example, in a normal Q-Q plot, you would compare the quantiles of the sample data to those of a normal distribution.
  - **Plotting:** The quantiles of the sample data are plotted on the y-axis, and the corresponding quantiles of the theoretical distribution are plotted on the x-axis.

Use and Importance in Linear Regression:

- **Assessing Normality of Residuals:**
  - **Purpose:** In linear regression, one of the key assumptions is that the residuals (the differences between observed and predicted values) should be normally distributed.
  - **Q-Q Plot Use:** By plotting the residuals' quantiles against the quantiles of a normal distribution, a Q-Q plot helps to visually assess whether this assumption holds. If the residuals are normally distributed, the points will roughly align with the straight line in the Q-Q plot.

- **Detecting Deviations from Normality:**

- **Purpose:** Deviations from normality in residuals can indicate potential issues with the model, such as missing variables, incorrect functional form, or heteroscedasticity.
- **Q-Q Plot Importance:** Significant deviations from the straight line (such as curvature or heavy tails) can suggest that the residuals are not normally distributed, which might impact the validity of statistical inferences and hypothesis tests based on the regression model.

- **Model Validation:**
  - **Purpose:** Validating model assumptions is crucial for ensuring the reliability of regression analysis results.
  - **Q-Q Plot Use:** The Q-Q plot is a diagnostic tool that helps in verifying the assumption of normality, which is necessary for valid hypothesis testing and confidence interval estimation in linear regression.

Interpretation:

- **Points on the Line:** If the points fall approximately along the reference line (often the 45-degree line), it indicates that the residuals are approximately normally distributed.
- **Points Deviating from the Line:** Systematic deviations from the line suggest non-normality. For example:
  - **Heavy Tails:** If the points curve away from the line at the ends, it may indicate that the residuals have heavier tails than the normal distribution.
  - **S-shaped Curve:** If the points form an S-shaped curve, it might suggest that the residuals have a skewed distribution.

Summary

A Q-Q plot is a visual diagnostic tool used to check if the residuals from a linear regression model follow a specified theoretical distribution, usually the normal distribution. Its importance lies in its ability to validate the normality assumption of residuals, which is critical for making valid statistical inferences and ensuring the reliability of the regression model.