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?

Answer:

Analyzing the categorical variables in the dataset will allow us understand how these features influence the demand for shared bikes. Below are the key inferences based on the categorical variables in the dataset:

1. **Season:**
    - Fall which shows the highest demand due to favorable weather.
    - Winter shows the lowest demand due to colder conditions.
2. **Weather Situation:**
    - Clear/Partly Cloudy: Highest bike rentals.
    - Mist/Cloudy: Moderate demand.
    - Rain/Snow: Sharp decline in rentals, especially during heavy conditions.
3. **Holiday:**
    - Holidays can either increase demand or reduce it.
4. **Working Day:**
    - Higher rentals on working days due to commuting.
    - Weekends see demand -oriented usage patterns.

Insight:

Seasonality, weather and day type significantly influence bike demand. Favorable conditions drive usage, while adverse weather and non-working days reduce the demand.

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

Answer:

When all categories of a variable are encoded as dummy variables, the final set of dummy variables adds up to 1. This introduces redundancy and perfect multicollinearity which can distort regression analysis.

Below is the reason why drop_first=True is Important while building model:

- **Reduces Multicollinearity:** By dropping one dummy variable the model avoids redundancy which ensure that the dummy variables are not perfectly correlated.
- **Improves Interpretability:** The dropped category serves as the reference against which other categories are compared.
- **Prevents Errors in Regression Models:** Linear regression models assume no multicollinearity among predictors; violating this can make coefficients unstable and unreliable.

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

Answer:

From the pair-plot of numerical variables (temp, atemp, hum, windspeedand cnt), **temp** emerges as the variable with the strongest positive correlation to the target variable cnt (bike rentals). This indicates that higher temperatures are associated with increased bike demand, likely because warmer weather conditions encourage outdoor activities.

Detailed Observations:

1. **temp:**
    - Shows a strong positive correlation with cnt.
    - As the temperature rises, bike rentals increase, likely due to more comfortable and appealing weather for biking.
    - This makes temperature a key driver of demand in the dataset.
2. **atemp:**
    - Exhibits a similar positive trend as temp, since it reflects perceived temperature.
    - High atemp values correlate strongly with increased bike rentals, reinforcing the impact of weather comfort on demand.

3. **hum:**
    - Displays a moderate negative correlation with cnt.
    - Higher humidity levels might make biking uncomfortable, slightly reducing demand on more humid days.
4. **windspeed:**
    - Weak negative correlation with cnt.
    - Strong winds are likely a deterrent for outdoor activities like biking, though the effect is less pronounced compared to other variables.

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

Answer:

In linear regression analysis, several key assumptions must be satisfied to ensure the validity and reliability of the model's estimates. Here's an expanded explanation of these assumptions, along with methods for their validation and potential remedies when violations occur:

1. Linearity

- **Assumption**: The relationship between the independent variables and the dependent variable is linear. This ensures the model captures the true relationship between predictors and the target variable.
- **Validation**:
    - Plot the **predicted values vs. actual values** or **residuals vs. predicted values**.
    - A random scatter of residuals indicates linearity, while patterns (e.g., curved or systematic structures) suggest non-linearity.
    - **Addressing Non-Linearity**:
        - Apply transformations to the features or target variable (e.g., logarithmic, square root).
        - Add polynomial terms to capture non-linear relationships.

2. Independence of Errors

- **Assumption**: Residuals should be independent of each other, meaning the error terms should not display any systematic relationship (no autocorrelation).
- **Validation**:
    - Perform the **Durbin-Watson test**:
        - Values close to 2 suggest no autocorrelation.
        - Values < 2 indicate positive autocorrelation and values > 2 suggest negative autocorrelation.
    - Particularly relevant for **time-series data**, where successive observations might be correlated.
    - **Addressing Autocorrelation**:
        - Use time-series regression models or lag variables.
        - Consider ARIMA models for sequential data.

3. Homoscedasticity

- **Assumption**: The variance of residuals should remain constant across all levels of the independent variables (homoscedasticity).
- **Validation**:
    - Plot **residuals vs. predicted values**. The spread of residuals should be uniform across the predicted values.
    - A funnel-like shape (residuals spreading wider or narrower) indicates **heteroscedasticity** (non-constant variance).
    - **Addressing Heteroscedasticity**:
        - Apply transformations to stabilize variance (e.g., log or square root).
        - Use **weighted least squares regression**, which assigns weights to minimize the effect of non-constant variance.

4. Normality of Errors

- **Assumption**: Residuals should be normally distributed to ensure valid statistical inference, particularly for hypothesis testing and confidence interval estimation.
- **Validation**:
    - Plot a **histogram** or a **Q-Q plot** of residuals:
        - Histogram: Residuals should form a bell-shaped curve.
        - Q-Q Plot: Residuals should align closely with the diagonal line.
    - Perform statistical tests:
        - **Shapiro-Wilk Test** or **Kolmogorov-Smirnov Test**: A p-value > 0.05 suggests normality.
    - **Addressing Non-Normality**:
        - Apply transformations to the dependent variable or residuals (e.g., log or Box-Cox transformation).
        - Use non-parametric models if normality cannot be achieved.

5. No Multicollinearity

- **Assumption**: Independent variables should not be highly correlated with each other, as multicollinearity distorts coefficient estimates, making them unreliable.
- **Validation**:
    - Calculate the **Variance Inflation Factor (VIF)** for each predictor:
        - VIF > 5 (or 10) indicates problematic multicollinearity.
    - Use a **correlation matrix** to identify pairs of highly correlated predictors.
    - **Addressing Multicollinearity**:
        - Drop one of the correlated variables.
        - Combine highly correlated variables using techniques like **Principal Component Analysis (PCA)**.

        - Regularize the model using techniques like **Ridge Regression** or **Lasso Regression**.

6. Model Fit

- **Validation**:
    - Evaluate performance using the following metrics:
        - **R-squared**: Proportion of variance explained by the model.
        - **Adjusted R-squared**: Adjusted for the number

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

Answer:

To identify the top 3 features significantly contributing to the demand for shared bikes in the final linear regression model, follow these steps:

1. Analyze Coefficients and Statistical Significance

- Look at the **coefficients** of the independent variables in the model. Higher absolute values of coefficients indicate a stronger influence on bike demand.
- Check the **p-values** to determine significance. Variables with p-values < 0.05 are considered statistically significant.

2. Rank Features by Importance

- After identifying significant variables, rank them based on the magnitude of their **standardized coefficients** (beta coefficients). Standardization ensures comparability across features with different units or scales.

Example Interpretation:

From the final model, the top 3 contributing features might be:

1. **Temperature (temp)**: Indicates that demand increases with warmer temperatures.
2. **Year (yr)**: Shows an upward trend in demand from 2018 to 2019.
3. **Working Day (workingday)**: Suggests higher demand on non-holidays or weekdays.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

Linear regression is a **supervised learning algorithm** used to model the relationship between a dependent variable (target) and one or more independent variables (predictors). It assumes a linear relationship between the predictors and the target variable.

Key Concepts of Linear Regression

1. **Model Equation**:
    - For a single predictor (x): $y=\beta0+\beta1x+\epsilon$
    - For multiple predictors (x1,x2,...,xn): $y=\beta0+\beta1x1+\beta2x2+...+\beta nxn+\epsilon$

    Where:

    - y: Dependent variable (target).
    - x1,x2,...,xn : Independent variables (features).
    - β0: Intercept (value of yyy when all predictors are 0).
    - β1,β2,...,βn : Coefficients (weights) representing the contribution of each predictor.
    - ϵ: Error term (captures unexplained variance).
2. **Linear Assumptions**:
    - Linearity: The relationship between predictors and the target is linear.
    - Independence of Errors: Residuals are independent of each other.
    - Homoscedasticity: Constant variance of residuals.
    - Normality of Errors: Residuals are normally distributed.
    - No Multicollinearity: Predictors are not highly correlated.

Steps in the Linear Regression Algorithm

- Initialize the Model: Start with a random guess for the coefficients
- Compute Predictions: Use the initial coefficients to predict y for all data points:

- Calculate the Cost Function: Measure the model's performance using the MSE or RSS (Residual Sum of Squares)
- Optimize the Coefficients: Adjust the coefficients to minimize the cost function.
- Evaluate the Model: Use performance metrics like R-squared, Adjusted R-squared and RMSE and MAE

Types of Linear Regression

1. **Simple Linear Regression**:
    - A single independent variable (xxx).
    - Example: Predicting house prices based on area.
2. **Multiple Linear Regression**:
    - Multiple independent variables (x1,x2,...,xn).
    - Example: Predicting house prices based on area, locationand number of bedrooms.
3. **Regularized Linear Regression**:
    - Adds penalties to the coefficients to prevent overfitting.
        - **Ridge Regression**: Penalizes the sum of squared coefficients (L2 regularization).
        - **Lasso Regression**: Penalizes the sum of absolute coefficients (L1 regularization).

Strengths of Linear Regression

- Simplicity: Easy to implement and interpret.
- Speed: Efficient for small to medium-sized datasets.
- Baseline Model: Serves as a good starting point for understanding relationships in data.

Limitations of Linear Regression

- Assumption-Dependent: Relies on strong assumptions (e.g., linearity, homoscedasticity).
- Sensitive to Outliers: Outliers can disproportionately affect the model.
- Limited Expressiveness: Cannot model complex, non-linear relationships without transformations.

- Multicollinearity Issues: High correlation among predictors can distort coefficient estimates.

2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's quartet is a collection of four distinct data sets that have nearly identical summary statistics, yet their visual patterns differ dramatically. This example was created to show the limitations of relying solely on numerical summaries and emphasize the importance of visualizing data. Despite having the same mean, variance, correlation and linear regression results, each data set demonstrates a unique relationship between the variables. Anscombe's quartet serves as a reminder that statistical analysis should always be complemented with graphical examination to gain a deeper understanding of the data.

Components of Anscombe's Quartet

The quartet consists of four datasets, each containing 11 data points (x, y). All four datasets share the following identical properties:

- **Mean of x**: 9
- **Mean of y**: 7.5
- **Variance of x**: 11
- **Variance of y**: 4.12
- **Correlation between x and y**: 0.82
- **Linear regression equation**: $y=3+0.5xy = 3 + 0.5xy=3+0.5x$

Although these summary statistics are the same for all datasets, they exhibit very different patterns when plotted. This underscores the fact that relying on numerical metrics alone can be misleading. Anscombe's quartet is a powerful illustration of why statistical analysis should be accompanied by graphical exploration. While numerical measures like means, variances and correlations are useful, they can obscure the true nature of the data. By visually inspecting the data through scatter plots or other graphical methods, analysts can gain a more complete understanding of the data, identify trends and detect problems like outliers or non-linear relationships. The key takeaway is clear: "Visualize your data before drawing conclusions from statistical summaries."

3. What is Pearson's R?

Answer:

Pearson's R, also known as **Pearson's correlation coefficient**, is a measure of the strength and direction of the linear relationship between two continuous variables. It quantifies how much one variable changes in relation to another and its value ranges from -1 to +1.

Formula for Pearson's R

The formula to calculate Pearson's correlation coefficient is:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Where:

- r is the Pearson correlation coefficient.
- n is the number of data points.
- x and y are the individual data points for the two variables.
- $\sum x$ is the sum of all values in x.
- $\sum y$ is the sum of all values in y.
- $\sum xy$ is the sum of the products of corresponding values of x and y.
- $\sum x^2$ is the sum of the squares of the values of x.
- $\sum y^2$ is the sum of the squares of the values of y.

Interpretation of Pearson's R

- **r=1**: Perfect positive linear relationship. As one variable increases, the other also increases in a perfectly linear fashion.
- **r=−1**: Perfect negative linear relationship. As one variable increases, the other decreases in a perfectly linear fashion.
- **r=0**: No linear relationship between the variables. There may still be some relationship, but it is non-linear or due to other factors.
- **0<r<1**: A positive linear relationship, with values closer to 1 indicating a stronger positive relationship.

- **$-1<r<0$**: A negative linear relationship, with values closer to -1 indicating a stronger negative relationship.

Strength of the Correlation

- **0.1 to 0.3** (or -0.1 to -0.3): Weak positive (or negative) correlation.
- **0.3 to 0.5** (or -0.3 to -0.5): Moderate positive (or negative) correlation.
- **0.5 to 1** (or -0.5 to -1): Strong positive (or negative) correlation.

Use Cases for Pearson's R

- **Investigating Relationships**: Pearson's R is commonly used in statistics to test the strength and direction of the relationship between two continuous variables.
- **Data Analysis**: It is widely used in fields such as economics, psychology and biology to explore correlations between variables.
- **Predictive Modeling**: While Pearson's R itself is not used for prediction, it can inform decisions in selecting predictor variables when building linear regression models.

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

Answer:

What is Scaling?

Scaling is the process of transforming numerical data into a specific range or distribution to ensure all features are comparable. This step is crucial in machine learning models, especially for algorithms that rely on the magnitude of the features, such as distance-based algorithms (e.g., K-Nearest Neighbors, Support Vector Machines and gradient-based optimization techniques like linear regression). Scaling ensures that no single feature disproportionately affects the model due to its larger magnitude.

Why is Scaling Performed?

Scaling is done for the following reasons:

1. **Improved Model Performance**: Many algorithms perform better when the features are on a similar scale. This prevents features with larger values from dominating the model's learning process.

2. **Faster Convergence**: In optimization algorithms like gradient descent, scaling ensures more efficient convergence, as all features contribute equally to the gradient update.
3. **Equal Weight to Features**: Some models, such as linear regression or support vector machines, assume that all features contribute equally. Scaling prevents bias due to features with larger numerical values.
4. **Algorithm Assumptions**: Certain algorithms, like K-Nearest Neighbors and PCA, assume that features are scaled before calculating distances or similarities.

Difference Between Normalized Scaling and Standardized Scaling

Normalization and standardization are the two main types of scaling and they differ in how they transform the data.

1. Normalization (Min-Max Scaling) : Normalization rescales the data so that all values fall within a fixed range, typically [0, 1] or [-1, 1]. The transformed values are bounded within a specific range. It is sensitive to outliers, which can distort the range of the data and affect the scaling. Commonly used when the data must be constrained within a certain range, or when the distribution is skewed or unknown.

2. Standardization (Z-Score Scaling) : Standardization transforms the data by centering it around a mean of 0 and scaling it according to the standard deviation. The data is centered around 0, with values distributed according to the standard deviation. It does not bound the values within a fixed range. It is less sensitive to outliers compared to normalization because it uses the standard deviation, which is less influenced by extreme values. Commonly used when data follows a normal distribution or when working with algorithms like linear regression, PCA and SVM

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

Answer:
A Variance Inflation Factor (VIF) measures how much the variance of a regression coefficient is inflated due to multicollinearity with other predictors. A high VIF indicates a high correlation between a predictor and other predictors in the model, which can make the coefficient estimates unstable and difficult to interpret.

The **VIF** can become infinite in the Perfect Multicollinearity and Linear Dependence Between Predictors cases. When VIF is infinite, the model becomes unstable because it cannot distinguish the individual effects of the correlated predictors. This leads to unreliable coefficient estimates. Infinite VIF is a clear

indication of multicollinearity, which should be addressed either by removing one of the correlated predictors or by using techniques like Principal Component Analysis (PCA) to reduce dimensionality.

Perfect Multicollinearity:

- This occurs when one of the independent variables is a perfect linear function of another predictor.

Linear Dependence Between Predictors:

- If one predictor is a linear combination of other predictors, the model cannot uniquely estimate the coefficients of the variables.

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

Answer:

A **Quantile-Quantile (Q-Q) plot** is a graphical tool used to compare the distribution of a dataset with a theoretical distribution, often the **normal distribution**. The plot displays the quantiles of the observed data against the quantiles of the theoretical distribution.

In a Q-Q plot:

- The **x-axis** shows the quantiles of the reference distribution (typically normal).
- The **y-axis** shows the quantiles of the observed data.

If the points lie along a straight line, this suggests that the observed data follows the reference distribution. Any deviation from this straight line indicates that the data does not follow the reference distribution.

Use of a Q-Q Plot in Linear Regression

In linear regression, one of the key assumptions is that the **residuals (errors)** are normally distributed. The Q-Q plot is a helpful visual tool to check this assumption by plotting the residuals of the regression model against a normal distribution.

How to Use a Q-Q Plot in Linear Regression:

1. **Fit the linear regression model**: First, perform regression analysis on the data.

2. **Obtain the residuals**: Calculate the residuals, which are the differences between the observed and predicted values.
3. **Create the Q-Q plot**: Plot the residuals against the quantiles of a normal distribution.
4. **Interpret the plot**:
    - If the points lie along the straight line, the residuals are normally distributed and the assumption is validated.
    - If the points deviate from the line, particularly at the extremes, the residuals may not be normally distributed, which violates the assumption.

Importance of a Q-Q Plot in Linear Regression

1. **Checking Normality of Residuals**:
    - A Q-Q plot visually checks if the residuals follow a normal distribution, which is crucial for linear regression. Normally distributed residuals indicate the model is appropriate and that statistical inference (e.g., confidence intervals, hypothesis tests) is valid.
2. **Validating Assumptions**:
    - Normality of residuals is a critical assumption for accurate statistical tests in regression, such as t-tests and F-tests. A Q-Q plot helps confirm that this assumption holds, ensuring the reliability of the results.
3. **Identifying Outliers**:
    - Deviations from the normal distribution, especially in the tails of the plot, may indicate outliers or influential data points. This can prompt further investigation into the data.
4. **Model Diagnostics**:
    - The Q-Q plot is an important diagnostic tool in regression analysis. It helps identify potential issues with the residuals, such as skewness, heavy tails, or non-linearity, which can affect model accuracy. Addressing these issues improves model performance.
5. **Improving Model Assumptions**:
    - If the Q-Q plot shows significant deviation from normality, transformations (such as log transformation) on the data may help achieve normality, improving the model's validity.