Assignment-based Subjective Questions
Solution by Ayaz Aslam | January 29, 2025

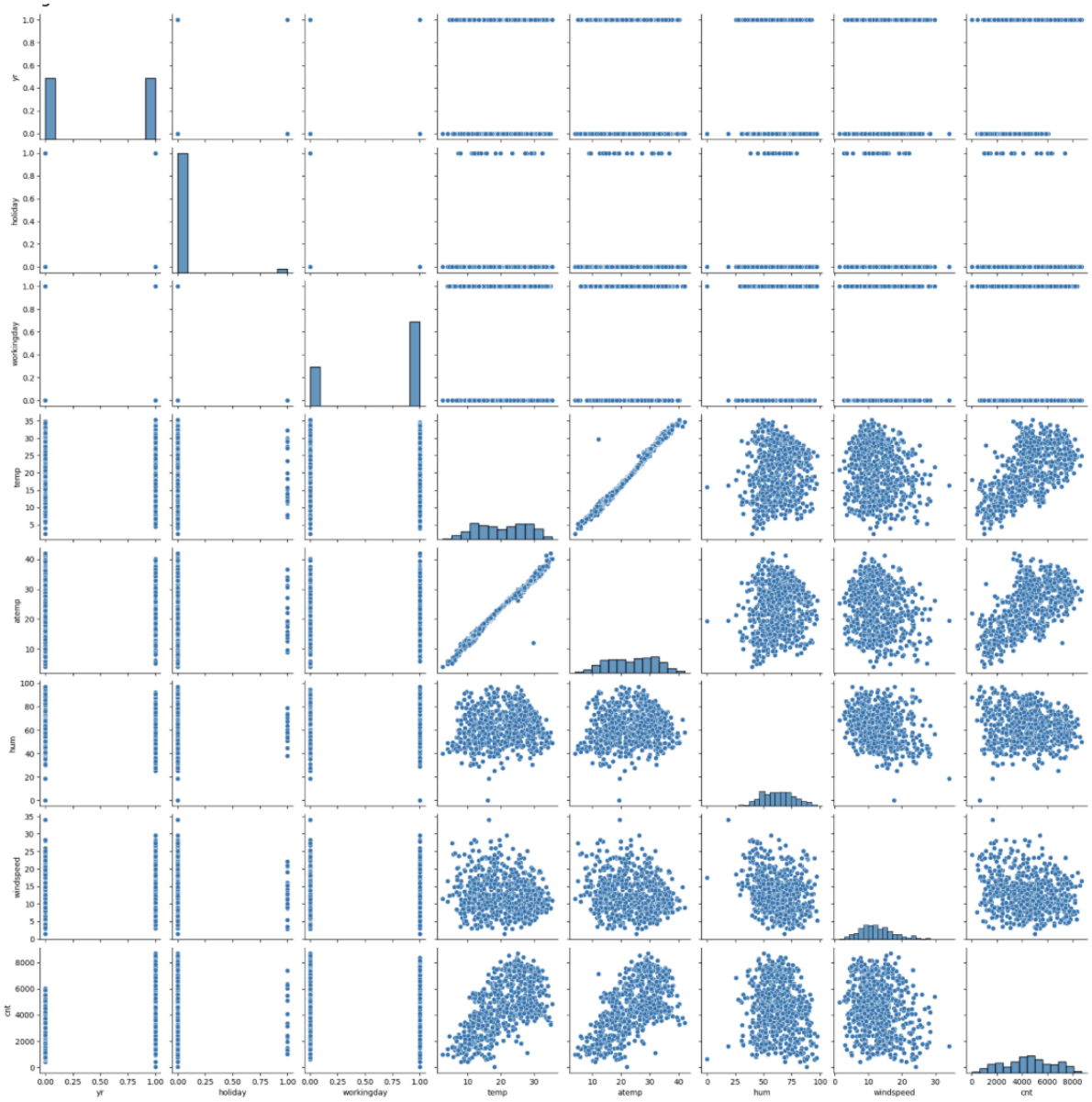
Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 1 goes below this line> (Do not edit)

Ans. Following are the effects for categorical variables on the dependent variable:
a. Company should focus on expanding business during Fall, Summer and Winter
b. September has shown great demand.
c. It has been observed that the demand for bike rentals had gone up from 2018 to 2019. So we can say that it will go up once the situation gets normal post Covid
d. There would be less bookings during Bad and no demand in severe weather conditions.
e. There is no much demand during the holidays

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)

The drop_first =True is important as it helps in reducing the extra column created during creation of the dummy variable. Dropping the column is important because the importance or value of that left over variables can be found by remaining variables. So to avoid redundancy we are dropping a column. This helps the column to become linearly independent.

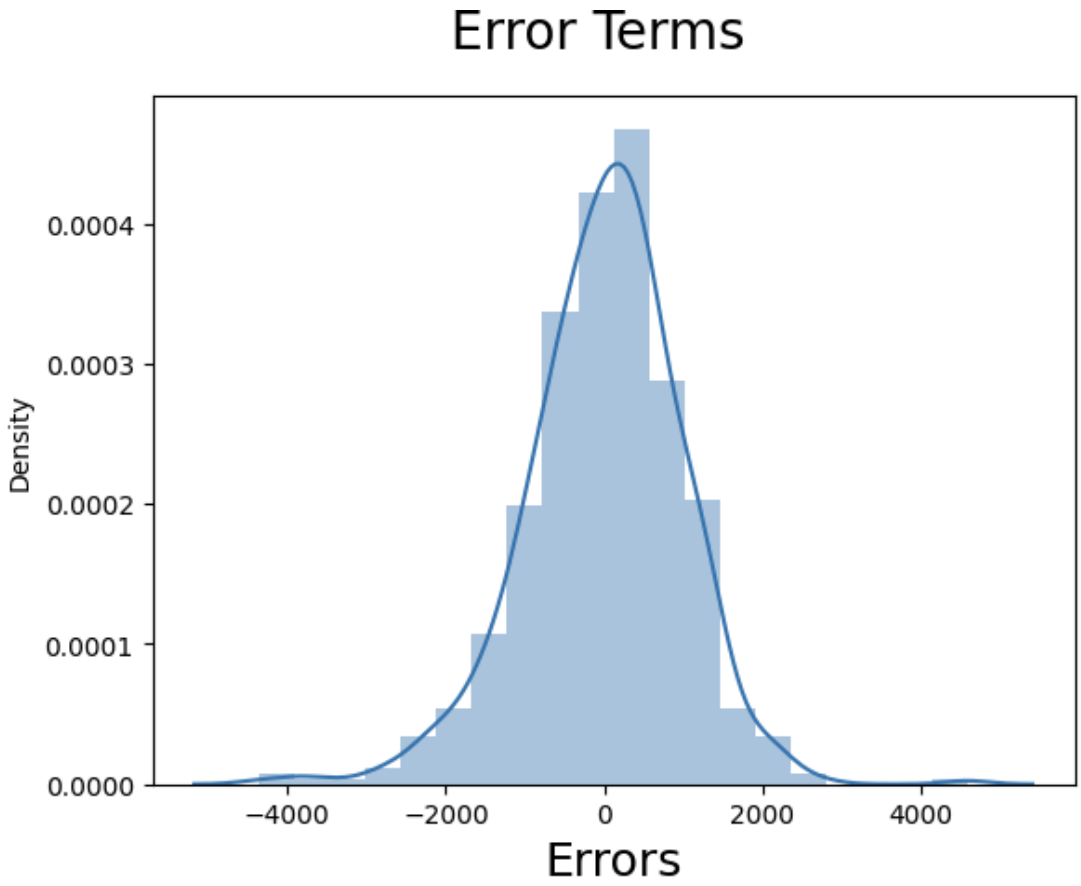
Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)



Temp and atemp are the numerical variables which are showing the highest correlation with the target variable.

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)

Residuals distribution should follow normal distribution and centered around 0.(mean = 0). We validate this assumption about residuals by plotting a distplot of residuals and see if residuals are following normal distribution or not. The above diagram shows that the residuals are distributed about mean = 0.



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)

Based on the final model yr, temp and weather are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

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)

<Your answer for Question 6 goes here>

Answer:

Linear regression is a statistical method used to model the relationship between a dependent variable (also known as the target or output variable) and one or more independent variables (also known as features or predictors). The goal of linear regression is to find the best-fitting line (or hyperplane in the case of multiple predictors) that minimizes the difference between the actual values and the predicted values.

Here's a step-by-step breakdown of the linear regression algorithm:

1. **Assumptions:**
    - There is a linear relationship between the dependent variable and the independent variables.
    - The residuals (the differences between the observed and predicted values) are normally distributed.
    - The variance of the residuals is constant (homoscedasticity).
    - The observations are independent of each other.
2. **Equation of a Line:** In simple linear regression (with one independent variable), the relationship is represented by the equation:
   $Y=\beta0+\beta1X+\epsilon$
   Where:
    - Y is the dependent variable (target),
    - X is the independent variable (predictor),
    - $\beta0$ is the intercept (the value of Y when X=0),
    - $\beta1$ is the slope (the rate of change in Y for a unit change in X),
    - ϵ is the error term (residuals).
3. **Objective:** The objective is to find the optimal values of $\beta0$ and $\beta1$ that minimize the residual sum of squares (RSS):
   $RSS=\sum i=1n(Yi-(\beta0+\beta1Xi))2$
   Where n is the number of data points, and Yi and Xi are the observed values.
4. **Finding the Best Fit Line:** To find the best-fit line, we use optimization techniques, typically gradient descent or the closed-form solution (normal equation):
   $\beta\hat{}=(XTX)-1XTY$
   Where:
    - $\beta\hat{}$ is the vector of estimated coefficients (intercept and slope),
    - X is the matrix of independent variables (with a column of ones for the intercept),
    - Y is the vector of observed dependent variable values.
5. **Interpretation:**
    - The slope $\beta1$ represents the change in the dependent variable for a one-unit change in the independent variable.
    - The intercept $\beta0$ represents the value of the dependent variable when the independent variable is zero.
6. **Prediction:** Once the coefficients are determined, predictions can be made for new values of X using the equation:
   $Y\hat{}=\beta0+\beta1X$
7. **Evaluation:** The model's performance can be evaluated using metrics such as:
    - **R-squared (R2)**: Measures the proportion of the variance in the dependent variable that is predictable from the independent variable(s).
    - **Mean Squared Error (MSE)**: Measures the average squared difference between the observed and predicted values.

Linear regression is a simple yet powerful technique used for prediction and trend analysis.

However, it assumes a linear relationship, and may not work well if the data shows complex, non-linear patterns.

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)

<Your answer for Question 7 goes here>

Answer:

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics, such as mean, variance, correlation, and linear regression results, yet appear very different when graphed. The purpose of Anscombe's quartet is to demonstrate the importance of visualizing data before analyzing it using statistical methods, as relying on summary statistics alone can lead to misleading conclusions.

The quartet consists of four datasets, labeled **I**, **II**, **III**, and **IV**, each containing eleven data points with two variables, Xand Y. For each dataset, the following statistics are identical:

- 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 line: Y=3+0.5X

However, when you plot these datasets, you observe stark differences:

1. **Dataset I**: Displays a typical linear relationship between X and Y.
2. **Dataset II**: Consists of a perfect linear relationship with all points lying along the regression line, except for one outlier.
3. **Dataset III**: Shows a non-linear relationship, with X and Y forming a parabolic pattern.
4. **Dataset IV**: Demonstrates a case where all but one data point lie in a linear fashion, while one point is far removed, distorting the analysis.

The quartet highlights how the same statistical measures can be obtained from different data patterns. It emphasizes the importance of graphical analysis and suggests that statistical techniques may not capture the full complexity of data, leading to different interpretations depending on the underlying distribution.

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)

<Your answer for Question 8 goes here>

Answer:

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the linear relationship between two variables. It ranges from -1 to 1 and indicates the strength and direction of the correlation:

- **r=1**: Perfect positive linear correlation, meaning that as one variable increases, the other variable increases proportionally.
- **r=−1**: Perfect negative linear correlation, meaning that as one variable increases, the other decreases proportionally.
- **r=0**: No linear correlation, meaning that changes in one variable do not have any predictable relationship with changes in the other variable.

The formula for Pearson's R is given by:

$$r=\sum(Xi-\bar{X})(Yi-\bar{Y})\sum(Xi-\bar{X})2\sum(Yi-\bar{Y})2$$

Where:

- Xi and Yi are the individual data points of the variables X and Y,
- $\bar{X}$ and $\bar{Y}$ are the means of X and Y, respectively.

Pearson's R assumes that the relationship between the variables is linear, and both variables should be continuous and approximately normally distributed for the correlation to be meaningful. The closer the absolute value of r is to 1, the stronger the linear relationship between the variables.

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)

<Your answer for Question 9 goes here>

You can easily copy the text below and paste it directly into Google Docs:

Answer:

Scaling is the process of transforming the features or variables in a dataset to a common scale, without distorting differences in the ranges of values. It is particularly important in machine learning algorithms that rely on the distance between data points (e.g., k-nearest neighbors, support vector machines) or algorithms that assume certain data distributions (e.g., linear regression, principal component analysis).

Why is scaling performed?
Scaling is performed for several reasons:

1. **Improved model performance: Many machine learning algorithms, especially those that use distance or gradient-based optimization, perform better when the features are on the same scale.**
2. **Convergence speed: For algorithms that use gradient descent (e.g., linear regression, neural networks), scaling ensures faster convergence to the optimal solution, as features with large values do not dominate the optimization process.**
3. **Equal importance: Without scaling, features with larger numeric ranges may influence the model more than features with smaller ranges, leading to biased results.**

Difference between normalized scaling and standardized scaling:

1. **Normalized scaling (Min-Max scaling):**
    - **Involves rescaling the data so that it falls within a specified range, usually [0, 1].**
    - **Formula:**

    $X_{norm} = (X - X_{min}) / (X_{max} - X_{min})$
    Where X_{min} and X_{max} are the minimum and maximum values of the feature.

    - **Use case: Suitable when the data needs to be constrained to a specific range, especially for algorithms sensitive to the absolute values of features (e.g., neural networks).**
2. **Standardized scaling (Z-score scaling):**
    - **Involves rescaling the data to have a mean of 0 and a standard deviation of 1. This is done by subtracting the mean and dividing by the standard deviation of each feature.**
    - **Formula:**

    $X_{std} = (X - \mu) / \sigma$
    Where μ is the mean and σ is the standard deviation of the feature.

    - **Use case: Useful when the data has outliers or when the scale of the features is significantly different, and you want to ensure that each feature contributes equally regardless of the original scale.**

In summary, normalized scaling is typically used when you want data in a fixed range, while standardized scaling is used when you want to center the data around zero with unit variance, which can be more robust to outliers.

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)

<Your answer for Question 10 goes here>

Variance Inflation Factor (VIF) is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. It is calculated by taking the ratio of the variance of all a given model's betas divide by the variance of a single beta if it were fit alone. The higher the VIF value, the greater the correlation of the variable with other variables. Values of more than 4 or 5 are sometimes regarded as being moderate to high, with values of 10 or more being regarded as very high. If there is perfect correlation, then VIF = infinity. An infinite VIF value means that the variable is exactly linear combination of other variable. If the independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it's R-squared value will be equal to 1. So, VIF = 1/(1-1) which gives VIF = 1/0 which results in "infinity"

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)

<Your answer for Question 11 goes here>

Answer:

A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess if a dataset follows a specific theoretical distribution, commonly the normal distribution. It compares the quantiles of the data with the quantiles of a theoretical distribution. If the data follows the distribution closely, the points will lie approximately along a straight line. Deviations from this line indicate departures from the expected distribution.

In a Q-Q plot:

- The **x-axis** represents the theoretical quantiles (e.g., for the normal distribution).

- The **y-axis** represents the observed quantiles from the data.

Use and importance of a Q-Q plot in linear regression:

- **Checking normality of residuals:** In linear regression, one of the assumptions is that the residuals (the differences between the observed and predicted values) are normally distributed. A Q-Q plot can visually help assess this assumption. If the points in the Q-Q plot deviate significantly from a straight line, it suggests that the residuals are not normally distributed, which could invalidate statistical tests based on this assumption.
- **Identifying outliers:** Q-Q plots can also reveal outliers in the data. Points that deviate far from the line indicate observations that do not fit the assumed distribution, which may be important to investigate further.
- **Model diagnostics:** A Q-Q plot is a useful diagnostic tool for ensuring that the assumptions of linear regression are met. If the normality assumption is violated, it might indicate that a transformation of the dependent variable is needed, or a different regression model might be more appropriate.

In summary, the Q-Q plot helps assess the validity of the normality assumption for residuals, which is crucial for the reliability of hypothesis tests and confidence intervals in linear regression.
