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)

Response:

1. For 'Season': '2:summer', '3:fall' - there has been increase in total number of bike rentals
2. For Year 2019 - there has been increase in total number of bike rentals
3. On holidays - count of bike rentals have increased
4. For Month 4-10 - count of bike rentals have increased
5. Bike rental count has increased for weathersit 'Clear, Few clouds, Partly cloudy', 'Partly cloudy'

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

Response:

1. It is considered that we need to crate dummy variables for n-1 levels (where n is total levels)
2. This is done to create more efficient models and avoid multicollinearity

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

Response:

1. Observation from pair-plot among the numerical variable:
   'atemp' (feeling temperature in Celsius) has the highest correlation with the target variable 'cnt'

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Response:

Below are the assumptions of linear regression:

1. Linear relationship between X and y: validated using coef values for each variable used in model building. Each variable X has linear relation to the y value
2. Normal distribution of error terms: validated using Residual analysis by plotting a distplot
3. Independence of error terms: validated using Residual analysis by plotting a distplot
4. Constant variance of error terms: validated using Residual analysis by plotting a distplot

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

Response:

3 features contributing significantly towards explaining the demand are:

1. atemp
2. yr
3. season

General Subjective Questions

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

Response:

Linear regression is a fundamental supervised learning algorithm used for continuous variable prediction. It assumes a linear relationship between the independent variables (X) and the dependent variable (Y) you want to predict. Here's a breakdown of the algorithm:

1. **Data Preparation:**

Assemble your data: You'll need a dataset containing independent variables (features) and a dependent variable (target) you want to predict.

Preprocess the data**:** This may involve handling missing values, scaling numerical features (often recommended for better model performance), and one-hot encoding categorical variables (using pandas.get_dummies with drop_first=True to avoid multicollinearity).

2. Model Representation:

The core of linear regression lies in the linear equation:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \varepsilon$

where:

Y: Dependent variable (what you want to predict)

X_i: Independent variables (features used for prediction)

β_0: Intercept (constant term)

β_i: Coefficients (weights for each independent variable)

ε: Error term (represents the difference between the actual Y and the predicted Y)

3. Finding the Best Fit Line:

The goal is to find the values for the coefficients (β) that minimize the sum of squared errors (SSE). SSE represents the total squared difference between the actual Y values and the predicted Y values based on the current estimates of β.

4. Model Evaluation:

Once you have the coefficients, you can use the model to predict Y for new data points. However, it's crucial to evaluate the model's performance:

R-squared: Represents the proportion of variance in the dependent variable explained by the independent variables. It ranges from 0 to 1, with higher values indicating a stronger relationship.

Residual Analysis: Examining the distribution and patterns in the residuals (errors) can reveal potential issues like heteroscedasticity or model misspecification.

5. Model Refinement (Optional):

Based on the evaluation results, you may need to refine the model:

Addressing Assumptions: If assumptions like linearity or homoscedasticity are violated, data transformation or robust regression methods may be necessary.

Feature Selection: Identifying and removing irrelevant or redundant features can improve performance.

Regularization techniques: Techniques like L1 or L2 regularization can help prevent overfitting and improve model generalizability.

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

Response:

Anscombe's quartet is a set of four data visualizations created by statistician Francis Anscombe in 1973. All four datasets have the following surprising characteristic:

They share the same basic descriptive statistics - mean, variance, standard deviation, correlation coefficient, and linear regression line.

However, when plotted as scatter plots, they reveal vastly different underlying data distributions and relationships. This serves as a powerful reminder of the importance of data visualization in statistical analysis.

Datasets:

Each dataset consists of 11 data points (X, Y values). While the exact values may vary slightly depending on the source, they all share the same key statistical properties.

Importance of Visualization:

Anscombe's quartet highlights how relying solely on basic descriptive statistics can be misleading. Despite having identical summary statistics, the scatter plots reveal very different stories:

Dataset 1: Appears to show a clear positive linear relationship.

Dataset 2: Shows a seemingly random scatter with no discernible pattern.

Dataset 3: Has a strong outlier that significantly affects the linear regression line.

Dataset 4: Has a curved, non-linear relationship that the linear regression line fails to capture.

These visualizations emphasize the importance of looking beyond summary statistics and visually inspecting the data to understand its true distribution and potential relationships.

Why is it Important?

Anscombe's quartet serves as a cautionary tale for data analysts. It highlights the following key points:

1. Visualization is Crucial: Summary statistics alone can be deceptive. Data visualization through scatter plots and other techniques helps reveal patterns, outliers, and non-linearities that might be missed by just looking at numbers.
2. Underlying Assumptions: Statistical methods like linear regression often rely on assumptions about the data, such as linearity and normality of errors. Visualization helps assess if these assumptions are met and if the chosen model is appropriate.
3. Model Selection: Depending on the data distribution, different statistical models might be more suitable. Visualization can guide you towards the most appropriate modeling approach.

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

Response:

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that reflects the strength and direction of the linear relationship between two continuous variables. It represents the degree to which two variables tend to change together in a linear fashion.

Values: It ranges from -1 to +1.

Positive R (0 to +1): Indicates a positive linear relationship. As the value of one variable increases, the value of the other variable also tends to increase. The closer R is to 1, the stronger the positive correlation.

Negative R (-1 to 0): Indicates a negative linear relationship. As the value of one variable increases, the value of the other variable tends to decrease. The closer R is to -1, the stronger the negative correlation.

Zero R: Indicates no linear relationship between the variables. The changes in one variable are not associated with predictable changes in the other.

Interpretation: The absolute value of R tells you the strength of the relationship, but not the direction (positive or negative). You need to consider the sign of R to understand the direction.

Limitations: It only measures linear relationships. If the relationship between the variables is not linear, Pearson's R may not be a good indicator of their association.

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

Response:

Scaling transforms the value of features(variables) to a common scale range. This helps bring different sets of values to a common range improving in stability of the model.

There are two popular rescaling methods;

Normalization: Normalization scales features to a range between 0 and 1

Standardisation: Standardization scales features to have a mean of 0 and a standard deviation of 1

The advantage of Standardisation over normalization is that it doesn't compress the data between a particular range as in 0-1 scaling. This is useful, especially if there are extreme data point (outlier).

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

(3 marks)

Response:

This happens when there is perfect multicollinearity among the independent variables, where one variable can be perfectly predicted by others. Here the regression of the variable on the others will have an R2 value as exactly 1.

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

(3 marks)

Response:

A Q-Q plot also called as quantile-quantile plot is a plot of the quantiles(distribution) of the first data set against the quantiles(distribution) of the second data set. it's primarily used to assess whether the errors (residuals) of your model follow a normal distribution, which is one of the key assumptions of linear regression.

It can have below variations:

1. Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis
2. Y-values < X-values: If y-quantiles are lower than the x-quantiles.
3. X-values < Y-values: If x-quantiles are lower than the y-quantiles.
4. Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis