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:

   - When there is favorable weather, bike rentals are high.
   - On working days and non-holidays, bike rentals are high.

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

   Answer: Using drop_first=True when creating dummy variables is important because it helps prevent multicollinearity, which can confuse the model:

- **Avoids Redundancy**: Creating dummy variables results in each category of a categorical variable becoming its column. For example, we get 4 columns if we have 4 categories. Using drop_first=True we can drop one of these columns. Hence we will be left with 3 columns. Therefore, we can represent all categories without repeating the information.
- **Prevents Confusion**: Models get confused when we include all dummy columns. This is because having all dummy columns leads to a scenario where one column can be predicted from the others. Hence, to make the model more simple and stable, we should drop the first column.

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

   Answer: temp, shows the strongest correlation. In the pair plot, its visible temp and atemp show the same correlation, but we choose only one.

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

   Answer:
   - **Linearity**: There should be a linear relationship when we plot y_train against the y_train_pred.
   - **Independence of Residuals**: To ensure that no pattern exists, we need to plot residuals over time or another relevant sequence.

- **Homoscedasticity**: The spread needs to be constant across all predicted values when we plot residuals versus predicted values.
- **Normality of Residuals**: We will start with creating a histogram or a Q-Q plot of residuals. Residuals should either form a bell-shaped curve or fall along a straight line in a Q-Q plot to ensure the normality of residuals.
- **No Perfect Multicollinearity**: Variance Inflation Factor (VIF) can be calculated for each predictor. If the VIF value of the predictor is above 5, then this indicates high multicollinearity.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand for shared bikes?**
   Answer: The top 3 contributing features are yr, temp, and Winter

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

   Answer: Linear regression is something similar to finding the best-fit line from a bunch of scattered points on a graph. If you have a plot with data points scattered across, then drawing a straight line will represent the trend of these points.

   How Does It Work?

- **Data Points**: Consider each data point to be a pair of values. For example, if there is a list of houses, each point will represent the size and the price of the house.
- **Drawing a Line**: The line drawn through these points will best predict the house price based on its size. This line is called the "regression line."
- **Best-Fit Line**: The algorithm tries to minimize the distance between the actual points and the line to find the best-fit line. This distance is called "residuals" or "errors."
- **Formula of the Line**: The line is represented by the equation: y=mx+b
  - y is the predicted value (e.g., house price).
  - x is the input value (e.g., house size).
  - m is the slope of the line (how much y changes with x).
  - b is the y-intercept (the value of y when x is 0).
- **Finding the Best Line**: The slope (m) and intercept (b) is adjusted by the algorithm to ensure that the line fits the data points as closely as possible. This is done by minimizing the sum of the squared differences between the actual values and the predicted values. This method is called "least squares."

   Why Use Linear Regression?

Linear regression is a simple and easy way to interpret and understand the relationship between variables. For example, it shows how house prices increase with size.

Example

Let's say you have a few houses:

- House 1: 1000 sq ft, $300,000
- House 2: 1500 sq ft, $400,000
- House 3: 2000 sq ft, $500,000

Linear regression will draw a line that best represents the relationship between the values when plotted. If a new house is 1800 sq ft, you can use the line to predict its price.

Conclusion

Linear regression is an easy way to predict values and understand relationships between variables. This is similar to drawing the best straight line through data points to make predictions and insights.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a group of four datasets that look very different. However, they have almost identical statistical properties when plotted. These datasets were created to show the importance of visualizing data before analyzing it.

Why is Anscombe's Quartet Important?

The quartet shows that it can be misleading to rely only on summary statistics (like the average, variance, or correlation). To avoid arriving at wrong conclusions and to understand the true characteristics of data, it is important to plot and visualize data.

Let's Break It Down:

1. **Same Summary Statistics**: All four datasets have nearly identical summary statistics. They have the same:
    - Mean (average) of the x-values.
    - Mean (average) of the y-values.
    - Variance (spread) of the x-values and y-values.
    - Correlation between x and y.
    - Line of best fit (regression line).
2. **Different Plots**: Although they look similar, these datasets look different when plotted.
    - **Dataset 1**: A straight line is formed. This is the expectation from a linear relationship.

    - **Dataset 2**: A curve is formed. This shows a nonlinear relationship.
    - **Dataset 3**: There is a single outlier that greatly influences the data. The remaining points are almost constant.
    - **Dataset 4**: Forms a vertical line of points with one outlier. This outlier affects the line of best fit.

The Key Takeaways:

- **Dataset 1**: The regression line fits properly as we can see the strong linear relationship.
- **Dataset 2**: Despite the summary statistics, the relationship is not linear as a circle is formed.
- **Dataset 3**: One outlier impacts the regression line drastically.
- **Dataset 4**: The regression line is influenced by the single outlier.

Visual Example

Imagine you have four different scatter plots:

1. **First Plot**: The dots are all close to a straight line.
2. **Second Plot**: The dots form a curve.
3. **Third Plot**: Most dots are in a horizontal line, but one dot is far away.
4. **Fourth Plot**: The dots form a vertical line, with one dot far away.

These plots, although different patterns would all show the same average, variance, and correlation if we focus only on the numbers.

Conclusion

Anscombe's quartet will help us understand that numbers alone will not give the full picture. Visualizing data is crucial to learning the real patterns and relationships in data. When it comes to data analysis, appearances matter a lot.

3. What is Pearson's R?

Answer: Pearson correlation coefficient is a number that tells how two things are related to each other and in what direction.

Think of it this way:

1. **What Does It Measure?**:
    - It measures the strength and direction of the relationship between two variables.
    - For example, it can show how test scores are related to the amount of time spent studying.

    The Scale:

    - The value of Pearson's R ranges from -1 to 1.
        - **1**: Perfect positive relationship. As one variable increases, the other one also increases in a perfectly straight line.
        - **-1**: Perfect negative relationship. As one variable increases, the other one decreases in a perfectly straight line.
        - **0**: No relationship. Changes in one variable do not predict changes in the other.
2. **Positive vs. Negative**:
    - **Positive Relationship** (0 to 1): If Pearson's R is positive, it means that as one variable goes up, the other one also tends to go up.
        - Example: The more hours you study, the higher your test score tends to be.
    - **Negative Relationship** (-1 to 0): If Pearson's R is negative, it means that as one variable goes up, the other one tends to go down.
        - Example: The more hours you spend doing other chores, the lower your test score tends to be.
3. **Strength of the Relationship**:
    - **Close to 1 or -1**: There is a strong relationship wherein the variables move closely together in a predictable way.
    - **Close to 0**: There is a weak relationship wherein the variables don't move together predictably.

Example:

Let us look at the relationship between the number of hours someone exercise and their level of happiness. The data can be collected from different people:

- If Pearson's R is **0.80** - there's a strong positive relationship. More exercise is associated with higher happiness.
- If Pearson's R is **-0.80** - there's a strong negative relationship. More exercise is associated with lower happiness.
- If Pearson's R is **0.04** - there's almost no relationship. That means people exercising and their happiness level doesn't seem to be related.

Why It's Useful:

Pearson's R helps us learn about relationships in data and make predictions. We are also able to see if our assumptions about relationships are true or not. It's like a quick check to see if two things are related and how strong that relationship is.

To conclude, Pearson's R is useful to see if two variables are related. That is if they move together, and how strong their relationship is.

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

This is a data preprocessing technique used to adjust the range and distribution of features (variables) in your data. This is to ensure that analysis and model training is contributed equally by all features.

Why is Scaling Performed?

1. **Improves Model Performance**: It is observed that when features have similar scales, a lot of machine learning algorithms perform better. There are a lot of algorithms like k-nearest neighbors (KNN), support vector machines (SVM), gradient descent-based methods (like linear regression), etc. that are sensitive to the scale of the data.
2. **Speeds Up Convergence**: Scaling can speed up convergence in algorithms that have gradient descent. Features that are on different scales will lead to a slow and inefficient optimization process.
3. **Ensures Equal Contribution**: Scaled features will contribute equally to the model. Therefore features with larger values will not be dominating the learning process.

Difference Between Normalized Scaling and Standardized Scaling

Normalized Scaling (Min-Max Scaling):

- **What It Does**: Data is transformed to fit within a specific range, generally[0, 1].
- **When to Use**: This method is useful when the data distribution does not follow a Gaussian distribution (bell curve). Can also be used when the values have to be bonded to a specific range.
- **Example**:
    - Original data: [10, 20, 30, 40, 50]
    - Normalized data: [0, 0.25, 0.5, 0.75, 1]

Standardized Scaling (Z-Score Standardization):

- **What It Does**: Data is transformed to have a mean of 0 and a standard deviation of 1.
- **When to Use**: This method is useful when the data follows a Gaussian distribution. It's useful for algorithms where data is assumed to be normally distributed or is sensitive to the scale of the features.
- **Example**:
    - Original data: [10, 20, 30, 40, 50]
    - Mean: 30
    - Standard deviation: 15.81 (approx.)
    - Standardized data: [-1.26, -0.63, 0, 0.63, 1.26]

Summary

- **Scaling**: Helps to adjust the range and distribute data features.
- **Why**: Helps to improve model performance, fastens convergence, and leads to equal feature contribution.

- **Normalized Scaling**: Transforms data to a specific range, typically [0, 1].
- **Standardized Scaling**: Transforms data to have a mean of 0 and a standard deviation of 1.

More accurate and efficient machine learning models can be built when features are on a comparable scale. Scaling helps to achieve this.

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

Answer: VIF helps us to understand how much the variance of a regression coefficient is inflated due to collinearity with other predictors. In short, it shows how much one feature (variable) in your dataset can be explained by the other features.

Why Can VIF Be Infinite?

1. **Perfect Collinearity**:
    - If one feature is a perfect linear combination of one or more other features, we call it an infinite VIF
    - For example, if you have two features, X and Y, and Y is exactly twice X for every data point (Y = 2X), then they are perfectly collinear.
2. **Division by Zero**:
    - VIF is calculated using a formula divided by a very small number (R^2). We get infinite VIF, when features are perfectly collinear and the denominator in this formula becomes zero.

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

Answer: The Q-Q plot expands into the "quantile-quantile" plot and helps you to compare the distribution of a dataset to a theoretical distribution(generally the normal distribution). This is a tool to see if your data follows a certain distribution.

How Does a Q-Q Plot Work?

1. **Plotting Points**: The data points are displayed as dots on a graph. Each dot represents a quantile of your dataset (i.e., a data point at a certain position in the sorted data).
2. **Theoretical Line**: The plot contains a straight line. This line shows the theoretical distribution you are comparing your data to. For example, if you're comparing your data to a normal distribution, the line represents a perfectly normal distribution.
3. **Comparison**: You can understand the match between data and theoretical distribution by comparing where the dots lie relative to the straight line.
    - If the dots fall along the line, your data follows the theoretical distribution.
    - If the dots deviate from the line, your data does not follow the theoretical distribution.

Use and Importance of a Q-Q Plot in Linear Regression

When building a linear regression model, there are several assumptions you need to check to ensure your model is valid. One key assumption is that the residuals (the differences between observed and predicted values) are normally distributed. The Q-Q plot helps you check this assumption.

Checking Normality of Residuals

1. **Residuals**: Residuals (which is the difference between actual data points and the value predicted in the model) are obtained after fitting a linear regression model.
2. **Create Q-Q Plot**: A Q-Q plot of residuals can be created to see if they follow a normal distribution.
3. **Interpretation**:
    - **Dots on the Line**: The dots will lie approximately along a straight line if the residuals are normally distributed. This means your model's predictions are valid and reliable.
    - **Dots off the Line**: The dots will deviate significantly from the line when the residuals are not normally distributed. This means that the model may not be the best fit for the data.

Conclusion

A Q-Q plot is a simple but powerful tool to check if your data follows a specific distribution. Linear regression helps to verify if the residuals are normally distributed. This is important to validate the model. The model’s predictions may not be reliable if the residuals do not follow a normal distribution. This will lead to refining themodel or considering different approaches.