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?**

- Year : Demand for bikes increased in 2019 as compared to the demand in 2018.
- Holiday : Less demand on holidays.
- Seasons: More demand is seen in spring season than in summer and winter.
- Weathersit: Less demand when it snows or rains. Demand is more on a clear day.
- More demand is observed on Saturdays.
- Maximum demand of bikes was observed in the month of september.

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

In case we have K values, we only need to create k-1 variables. It is in order to drop an additional redundant column, that we use drop_first=True during dummy variable creation. E.g. for season variable, there can be four values which can be represented by three dummy variables as below:

- Winter: 000
- Summer:100
- Spring:010
- Fall: 001

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

temp and atemp are highly correlated to each other and also has highest correlation with cnt(target variable).

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

1. By doing residual analysis and plotting the error terms displaying normal distribution of errors with mean value as 0.

2. Compared R squared and adjusted R squared values of the test data and the trained data.

3. Plotted a graph to see the distribution of predicted and test data.

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

Top 3 features contributing significantly are:

Temp, weathersit(more demand on clear weather) and year.

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Linear regression is a statistical and machine learning algorithm used for modelling the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables and aims to find the best-fitting linear line through the data points. This line can then be used to predict the dependent variable's values for new or unseen data.

Linear regression models can be classified into two types depending upon the number of independent variables:

- Simple linear regression: When the number of independent variables is 1. The equation of a simple linear regression line is y = mx+c where m is the coefficient of variable indicating how much y changes for a unit change in x and c is the y-intercept i.e. value of y when x is 0. y is the dependent variable value for which is to be predicted and x is the independent variable's value.

- Multiple linear regression: When the number of independent variables is more than 1. The equation of Multiple linear regression line is y = b0 + b1*x1 + b2*x2 + ... + bn*xn

where: y is the dependent variable's value

b0 is the y-intercept.

b1, b2, ..., bn are the coefficients of the independent variables.

x1, x2, ..., xn are the values of the independent variables.

The goal of linear regression is to find the coefficients that minimize the difference between the predicted values and the actual values. This is done using a cost function, commonly the Mean Squared Error (MSE):

MSE = (1/N) * Σ(actual - predicted)^2

where N is the number of data points.

Steps to create linear regression model are :

1. Train-Test Datasets. Determine target variable and predictors
2. Pre-Processing Data. Apply pre-processing steps to your training and testing datasets separately in order to avoid data leakage.
3. Modeling. Instantiate regression algorithm.
4. Model Check. Model prediction and residuals.
5. Model Tuning.

Model Evaluation: To evaluate the performance of the linear regression model, it's important to assess how well it generalizes to new data. Common evaluation metrics include the coefficient of determination (R square), adjusted R square which measures the proportion of the variance in the dependent variable that's predictable from the independent variables. Other metrics is VIF which should be less than 5 and helps in identifying multicollinearity in the data.

Assumptions: Linear regression relies on several assumptions, including linearity (the relationship is linear), independence of errors (residuals), homoscedasticity (constant variance of residuals), and normality of residuals.

Q2. Explain the Anscombe's quartet in detail.

Anscombe's quartet was created by the statistician Francis Anscombe in 1973 to highlight the importance of data visualization. Anscombe's quartet is a set of four datasets that have nearly identical statistical properties when examined using summary statistics (such as mean, variance, correlation, and regression coefficients), but they display vastly different patterns when visualized graphically (to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone). This implies it is incorrect to rely only on summary statistics for understanding relationships within datasets and how important data visualization is to understand and draw completely different inferences from the same.

Q3. What is Pearson's R?

Pearson's correlation coefficient, often denoted as "r" or "Pearson's r," is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It assesses how well the relationship between the two variables can be described by a straight line. The coefficient ranges from -1 to 1, where:

- A value of 1 indicates a perfect positive linear relationship: As one variable increases, the other variable also increases proportionally.
- A value of -1 indicates a perfect negative linear relationship: As one variable increases, the other variable decreases proportionally.
- A value of 0 indicates no linear correlation: There is no linear relationship between the variables.

The formula for calculating Pearson's correlation coefficient between two variables X and Y is:

scssCopy code

$$r = \Sigma((X - \bar{X}) * (Y - \bar{Y})) / \sqrt{(\Sigma(X - \bar{X})^2 * \Sigma(Y - \bar{Y})^2)}$$

Where:

- $\bar{X}$ is the mean of variable X.
- $\bar{Y}$ is the mean of variable Y.
- The summation Σ is performed over all data points.

Pearson's correlation coefficient has several key properties:

1. Scale Invariance: The coefficient is not affected by changes in the scale (multiplication or addition of constants) of the data.
2. Symmetry: The correlation between X and Y is the same as the correlation between Y and X.
3. Bounded Range: The coefficient always falls between -1 and 1.

4. Sensitive to Outliers: Pearson's r can be influenced by outliers, as they can disproportionately affect the means and standard deviations.

5. Only Measures Linear Relationships: Pearson's correlation measures only linear relationships between variables. If the relationship is nonlinear, the coefficient might not accurately capture the association.

6. Not Robust to Non-Normality: It assumes that the variables are normally distributed, and extreme deviations from normality can affect its reliability.

Pearson's correlation coefficient is commonly used in various fields, including statistics, data analysis, machine learning, and scientific research. It helps researchers and analysts understand the degree and direction of association between variables and provides insights into their potential relationships. However, it's important to remember that correlation does not imply causation—just because two variables are correlated does not mean that changes in one variable cause changes in the other.

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

Scaling is the process of transforming numerical features in a dataset to a consistent range. It involves adjusting the values of the features so that they are on a similar scale, which can be helpful in various machine learning algorithms and data analysis tasks. Scaling is performed to ensure that the features don't have vastly different magnitudes, which could lead to biased or incorrect model training, particularly in algorithms that are sensitive to the scale of input features.

The main reasons for performing scaling are:

1. Preventing Numerical Instability: Some algorithms, such as gradient descent in optimization, converge faster when features are on similar scales. Extreme differences in scales can cause numerical instability in these algorithms.

2. Improving Model Performance: Many machine learning algorithms, like k-nearest neighbors and support vector machines, rely on distance measures between data points. When features have different scales, a feature with a larger magnitude might dominate the distance calculation, leading to biased results.

3. Regularization Techniques: Algorithms that use regularization (such as Ridge and Lasso regression) penalize large coefficients. Scaling helps prevent features with large magnitudes from having disproportionately large effects on the regularization term.

Two common methods for scaling are normalized scaling and standardized scaling:

1. Normalized Scaling (Min-Max Scaling): Normalization scales features to a specified range, often between 0 and 1. The formula for normalized scaling is:

scssCopy code

```
X_normalized = (X - X_min) / (X_max - X_min)
```

Here, X is the original feature value, X_min is the minimum value in the feature, and X_max is the maximum value in the feature. This scaling technique preserves the relative distances between data points but might be sensitive to outliers.

2. Standardized Scaling (Z-Score Scaling): Standardization transforms features to have a mean of 0 and a standard deviation of 1. The formula for standardized scaling is:

makefileCopy code

X_standardized = (X - X_mean) / X_std

Here, X is the original feature value, X_mean is the mean of the feature, and X_std is the standard deviation of the feature. Standardization centers the data around zero and gives it a unit variance. It is less sensitive to outliers compared to normalization.

In summary:

- Normalized Scaling: Maps data to a specified range, often between 0 and 1. Preserves relative distances between data points. Sensitive to outliers.
- Standardized Scaling: Transforms data to have mean 0 and standard deviation 1. Centers the data and makes it unit variance. Less sensitive to outliers.

The choice between normalized and standardized scaling depends on the characteristics of your data and the requirements of your specific machine learning algorithm. Some algorithms, like neural networks, might benefit from standardized scaling, while others, like decision trees, may not require scaling at all.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

The Variance Inflation Factor (VIF) is a measure used in regression analysis to assess multicollinearity among predictor variables. Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated, which can lead to issues in interpreting the individual effects of these variables and can affect the stability of regression coefficient estimates.

The formula for calculating the VIF for a predictor variable is:

$VIF = 1 / (1 - R^2)$

Where R^2 is the coefficient of determination obtained by regressing the predictor variable against all other predictor variables in the model.

In the context of calculating VIF, an infinite value occurs when R^2 is equal to 1, resulting in a denominator of 0 in the VIF formula. This happens when one predictor variable is perfectly predicted by a linear combination of other predictor variables. In other words, one predictor is a perfect linear combination of the others, indicating extreme multicollinearity.

This perfect multicollinearity can occur for a few reasons:

1. Duplication of Predictor Variables: If you inadvertently include the same predictor variable more than once in the regression model, it will be perfectly correlated with itself, leading to a perfect R^2 and thus an infinite VIF.

2. Linear Dependency: If one predictor variable is a constant multiple of another predictor variable (linear dependency), this can also lead to perfect multicollinearity and an infinite VIF.

3. Data Issues: In some cases, data cleaning or preprocessing issues might cause variables to be perfectly correlated, leading to infinite VIF values.

It's important to identify and address multicollinearity issues before proceeding with regression analysis. To handle multicollinearity and prevent infinite VIF values, you can take the following steps:

- Carefully review data and the predictor variables used in the model. Ensure that there are no duplicate or linearly dependent variables.
- Consider dropping one of the highly correlated variables to eliminate the perfect multicollinearity.

Identifying multicollinearity and addressing it appropriately can help ensure the reliability and interpretability of your regression results.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A Quantile-Quantile (Q-Q) plot is a graphical tool used to assess whether a dataset follows a specific theoretical distribution, such as the normal distribution. It compares the quantiles of the dataset to the quantiles of the theoretical distribution, allowing us to visually determine whether the two distributions are similar.

Here's how a Q-Q plot works:

1. Theoretical Quantiles: Theoretical quantiles are the expected values that data points would take on if they followed a specific distribution, such as the normal distribution. These quantiles are calculated based on the cumulative distribution function of the theoretical distribution.

2. Sample Quantiles: Sample quantiles are the actual data points sorted in ascending order. These quantiles are plotted against the corresponding theoretical quantiles.

3. Plotting: In a Q-Q plot, each data point's quantile is plotted on the y-axis against the corresponding theoretical quantile on the x-axis. If the data follows the theoretical distribution closely, the points on the plot will roughly align along a straight line. Deviations from a straight line can indicate departures from the assumed distribution.

The use and importance of a Q-Q plot in linear regression:

1. Assumption Checking: Linear regression assumes that the errors (residuals) follow a normal distribution. A Q-Q plot of the residuals can help assess the validity of this assumption. If the residuals deviate significantly from the expected diagonal line, it suggests that the normality assumption might not hold, indicating that the regression results might be unreliable.

2. Detecting Outliers: Outliers are data points that significantly deviate from the majority of the data. In a Q-Q plot, outliers can appear as points that deviate from the expected line. Identifying these outliers is crucial because they can disproportionately influence the regression model's coefficients and predictions.

3. Model Validation: A Q-Q plot is a visual tool to validate the distributional assumptions of the residuals. If the residuals exhibit a pattern or significant deviations from normality, it suggests that the linear regression model might not be appropriate for the data.

4. Guidance for Transformations: If you observe a clear pattern or non-linearity in the Q-Q plot, it might indicate that a transformation of the response variable or predictor variables could improve the model's fit.

In summary, a Q-Q plot is a valuable diagnostic tool in linear regression. It helps us assess the normality assumption, detect outliers, and identify potential issues with the regression model. If the Q-Q plot indicates departures from the assumed distribution, it might be necessary to consider alternative modeling techniques or data transformations to ensure the validity of the regression analysis.