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?

Based on the analysis of the categorical variables, we can infer the following:

1. Season: The season variable shows that the demand for shared bikes varies across different seasons. In particular, the demand tends to be higher during the spring and the summer seasons compared to fall and winter.

2. Year (yr): The year variable indicates that the demand for shared bikes has increased from 2018 to 2019. This suggests that the bike-sharing system has gained popularity over time, leading to a higher demand for shared bikes.

3. Weather Situation (weathersit): The weather situation variable reveals that the demand for shared bikes is influenced by weather conditions. Specifically, clear or slightly cloudy weather (weathersit = 1) corresponds to higher demand, while unfavorable weather conditions such as mist, rain, snow, and fog (weathersit = 2, 3, 4) result in lower demand.

Overall, the categorical variables in the dataset provide valuable insights into how different factors impact the demand for shared bikes.

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

By setting drop_first = True, we drop one of the dummy variables and avoid the problem of multicollinearity.
By dropping the first dummy variable, we ensure that there is no perfect linear relationship between the independent variables, and we can obtain more reliable and interpretable results from our regression model.

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

A-temp has the highest correlation, assuming casual and registered are removed while calculating the correlation between ‘cnt’ and other numerical variables

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

After building the linear regression model on the training set, I validated the assumptions with:

1. Residual analysis: Analyze the residuals (the differences between the predicted values and actual values). You can plot the residuals against the predicted values and look for any patterns or trends.

2. Multicollinearity: Assess the presence of multicollinearity among the independent variables. Calculate the variance factor (VIF) for each independent variable and check if there are any variables with high VIF values, indicating high correlation with other variables.

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

The top 3 features contributing significantly towards explaining the demand of the shared bikes are:

1. Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

2. Season(summer)

3. atemp

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a supervised machine learning algorithm used for predicting a continuous numerical value based on the relationship between independent variables (features) and a dependent variable (target). It assumes a linear relationship between the features and the target variable.

Here's a step-by-step explanation of how linear regression works:

Data Preparation: The first step is to gather and prepare the data. This involves collecting the relevant dataset, cleaning the data by handling missing values, outliers, and preprocessing the data if required.

Model Representation: In linear regression, the relationship between the features and the target variable is represented by a linear equation of the form:

y = b0 + b1*x1 + b2*x2 + ... bn*xn

where:

y is the target variable(dependent variable)

b0 is the intercept(y-axis intercept)

b1,b2,...bn are the coefficients (slopes) associated with each feature x1,x2, ...xn

Model Training: The goal is to estimate the values of the coefficients (b0, b1, ..., bn) that best fit the given data. This is done through a process called "model training" or "model fitting." The most common approach is to use the method of least squares, where the objective is to minimize the sum of squared differences between the predicted values and the actual values.

Cost Function: The cost function (also known as the loss function) quantifies the error between the predicted values and the actual values. In linear regression, the most commonly used cost function is the mean squared error (MSE), which calculates the average of the squared differences between the predicted and actual values.

Gradient Descent: To minimize the cost function and find the optimal values of the coefficients, an optimization algorithm like gradient descent is used. Gradient descent iteratively adjusts the coefficients by taking steps proportional to the negative gradient of the cost function until convergence is achieved.

Model Evaluation: Once the model is trained, it needs to be evaluated to assess its performance. Common evaluation metrics for linear regression include the coefficient of determination (R-squared), root mean squared error (RMSE), mean absolute error (MAE), etc.

Prediction: After the model is trained and evaluated, it can be used to make predictions on new, unseen data. By plugging in the values of the features into the linear equation, the model predicts the corresponding target variable.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a set of four datasets that have nearly identical statistical properties but exhibit different patterns when visualized. To highlight the importance of graphical data analysis and to demonstrate the limitations of relying solely on summary statistics.

The quartet consists of four sets of x-y coordinate data pairs, labeled I, II, III, and IV. Each set contains 11 data points, and the x and y values are presented as pairs. While the summary statistics (such as mean, variance, and correlation) for all four datasets are the same, the datasets differ significantly in terms of their distribution and relationship between variables.

Dataset I:
x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5

y: 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68

Dataset I appears to have a linear relationship between x and y, with a slight positive slope. It follows a relatively normal distribution, with some scatter around the line.

Dataset II:
x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5

y: 9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74

Dataset II also shows a linear relationship between x and y, but with a different slope. It includes an outlier at the far right, which significantly affects the regression line.

Dataset III:
x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
y: 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73

Dataset III does not follow a linear relationship between x and y. It contains a clear nonlinear pattern, where the relationship is better represented by a quadratic curve.

Dataset IV:
x: 8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8

y: 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89

Dataset IV has a single outlier that heavily influences the linear regression line. The majority of the points are clustered around x = 8, except for the outlier at x = 19.

3. What is Pearson’s R?

Pearson's R, also known as Pearson's correlation coefficient or simply correlation coefficient, is a statistical measure that quantifies the linear relationship between two variables.

It is widely used to assess the strength and direction of the linear association between variables.

Pearson's R is a value that ranges from -1 to +1, where:

A value of +1 indicates a perfect positive linear relationship, meaning that as one variable increases, the other variable also increases proportionally.

A value of -1 indicates a perfect negative linear relationship, meaning that as one variable increases, the other variable decreases proportionally.

A value of 0 indicates no linear relationship or correlation between the variables.

The formula for calculating Pearson's R is as follows:

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

where:

X and Y are the respective values of the two variables.
$\bar{X}$ and $\bar{Y}$ are the means of the X and Y variables, respectively.
Σ denotes the summation symbol.

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

Scaling, in the context of data preprocessing, refers to the process of transforming the numerical features of a dataset to a consistent scale. It involves adjusting the values of the features so that they fall within a specific range or have a specific distribution. Scaling is performed to ensure that all features contribute equally to the analysis and modeling processes and to avoid any bias or dominance of certain features based on their original scales.

The main reasons for performing scaling are as follows:

Comparable Magnitudes: Features in a dataset often have different units and scales. Scaling brings all features to a similar magnitude, preventing features with larger values from dominating those with smaller values. This ensures that the impact of each feature on the analysis or model is based on their actual relevance rather than their scale.

Improved Model Performance: Many machine learning algorithms, such as gradient descent-based algorithms, are sensitive to the scale of the input features. Features with larger scales can have a disproportionate impact on the model's behavior and can lead to slower convergence or biased results. Scaling helps in achieving faster convergence and better model performance.

Interpretability: Scaling makes it easier to interpret the coefficients or weights assigned to different features in a model. When features are on the same scale, the magnitude of the coefficients reflects their relative importance in predicting the target variable.

Normalized Scaling (Min-Max Scaling):

In normalized scaling, also known as min-max scaling, the values of the features are scaled to a fixed range, typically between 0 and 1.

The formula for normalized scaling is: scaled_value = (value - min_value) / (max_value – min_value)

Normalized scaling preserves the relative relationships between the data points and ensures that the minimum value maps to 0 and the maximum value maps to 1. However, it can be sensitive to outliers.

Standardized Scaling (Z-Score Scaling):

In standardized scaling, also known as z-score scaling or standardization, the values of the features are transformed to have a mean of 0 and a standard deviation of 1.

The formula for standardized scaling is: scaled_value = (value - mean) / standard_deviation

Standardized scaling centers the data around the mean and scales it based on the standard deviation. It ensures that the transformed values have a mean of 0 and a standard deviation of 1. Standardization is less affected by outliers compared to normalization.

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

The occurrence of infinite VIF (Variance Inflation Factor) values is typically a result of perfect multicollinearity within the dataset. Perfect multicollinearity exists when there is a linear relationship between two or more predictor variables in a regression model. This means that one or more variables can be expressed as a perfect linear combination of other variables.

When perfect multicollinearity is present, it leads to issues in the estimation of the regression coefficients, which in turn affects the calculation of VIF. VIF is calculated as the ratio of the variance of the estimated coefficient of a predictor variable to the variance of that coefficient if that variable were uncorrelated with the other predictor variables. Mathematically, VIF is calculated as 1 / (1 - R^2), where R^2 is the coefficient of determination of the predictor variable with the other predictor variables.

In the presence of perfect multicollinearity, the coefficient of determination (R^2) becomes equal to 1, resulting in an infinite value for VIF. This occurs because the estimated coefficient's variance is zero, making it impossible to calculate the ratio.

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

A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess the distributional similarity between two datasets. In the context of linear regression, a Q-Q plot is commonly used to evaluate the assumption of normality for the residuals of the regression model.

The Q-Q plot compares the quantiles of the observed residuals to the quantiles of a theoretical distribution, typically the standard normal distribution (mean = 0, standard deviation = 1). The observed residuals are plotted on the y-axis, and the theoretical quantiles are plotted on the x-axis.

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

Assessing Normality Assumption: The Q-Q plot allows us to visually inspect whether the residuals follow a normal distribution. If the observed residuals fall approximately along a straight line in the Q-Q plot, it suggests that the residuals are normally distributed. Departures from a straight line indicate deviations from normality.

Detecting Skewness or Outliers: Deviations from the straight line in the Q-Q plot can indicate the presence of skewness or outliers in the residuals. If the points deviate significantly from the line, it suggests that the residuals may not be normally distributed, indicating potential issues with the linear regression assumptions.

Guiding Model Improvements: The Q-Q plot provides insights into the shape and distribution of the residuals. If the plot reveals systematic departures from normality, it suggests that the model may need improvement, such as by considering non-linear transformations of variables or including additional predictors to better capture the underlying relationship.

Validating Inference: Normality of the residuals is a crucial assumption for valid statistical inference in linear regression. If the residuals violate the normality assumption, the p-values, confidence intervals, and hypothesis tests associated with the regression coefficients may be unreliable. The Q-Q plot helps in validating this assumption and provides confidence in the regression analysis.