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)**

From the coefficients provided in the regression analysis output, we can interpret the impact of the variable's "summer" and "winter" on the response variable (dependent variable), assuming all other variables in the model are held constant.

Summer (summer):

Coefficient: 0.0772

Interpretation: Holding all other variables constant, being in the summer season is associated with an increase in the response variable by 0.0772 units.

Confidence Interval: The 95% confidence interval for the coefficient is [0.056, 0.098], indicating that we are 95% confident that the true effect of being in the summer season lies between 0.056 and 0.098 units.

Winter (winter):

Coefficient: 0.1264

Interpretation: Holding all other variables constant, being in the winter season is associated with an increase in the response variable by 0.1264 units.

Confidence Interval: The 95% confidence interval for the coefficient is [0.105, 0.147], indicating that we are 95% confident that the true effect of being in the winter season lies between 0.105 and 0.147 units.

These interpretations suggest that both the summer and winter seasons have a positive impact on the response variable. Specifically, being in the summer season is associated with a smaller increase compared to being in the winter season. However, it's important to note that the interpretation should

consider the context of the data and the specific variables included in the regression model. Additionally, other factors such as interactions between variables or confounding variables may influence the interpretation of the coefficients.

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

Using drop_first=True during dummy variable creation is important for avoiding multicollinearity issues in regression analysis. When creating dummy variables from categorical variables, multicollinearity can arise when one of the dummy variables can be perfectly predicted from the other dummy variables. This is known as the "dummy variable trap.

dropping the first dummy variable, you ensure that there is no perfect linear relationship among the dummy variables. This helps prevent perfect multicollinearity, which can cause problems in regression analysis, such as unstable coefficients and inflated standard errors.

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

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

1. Check the residuals (the differences between observed and predicted values) to ensure they are normally distributed with constant variance (homoscedasticity). Plotting the residuals against the predicted values can help identify patterns or heteroscedasticity.

2. Check the linearity assumption by plotting the observed values against the predicted values. The relationship should be approximately linear.

3. Assess multicollinearity among the predictor variables using variance inflation factors (VIFs) or correlation matrices. VIF values greater than 5 or 10 may indicate multicollinearity issues that need to be addressed.

4 Evaluate model fit statistics, such as R2 adjusted Adjusted R2 *and root mean squared error (RMSE), on both the training and validation sets to assess the overall performance of the model.*

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

Workingday, Summar and Temp are the feature help the count

General Subjective Questions

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

Linear regression is a statistical method used to model the relationship between a dependent variable (Y) and one or more independent variables (X). The basic idea is to find the best-fitting linear equation that describes the relationship between the variables. This equation can then be used for making predictions or understanding the strength and nature of the relationships.

Linear regression categorized into Simple linear regression and Multiple linear regression

Simple Linear Regression:

1. Model Representation: Y=MX+B -> Y is the dependent variable and X is the independent variable. M is the slope and b is the intercept

The ultimate objective of the linear regression is to reduce the sum of square of difference of actual value and predicated value and the value predicated by the linear equation.

Multiple Linear regression

Multiple Linear Regression is an extension of simple linear regression, allowing for the modeling of the relationship between a dependent variable and multiple independent variables. In the multiple linear regression model, the relationship is expressed as a linear equation involving several predictors.

Y=b0+b1x1+b2x2...bnxn+e0

Here x1, x2, ... are the independent variable and Y is the dependent variable. here also formulate a cost function which finds the difference between the predicated and actual values.

The cost function measures the difference between predicated and actual values.

Linear regression is a supervised learning, this can use only for labeled data set

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

Anscombe's quartet serves as a valuable tool to promote statistical literacy and to remind analysts that a thorough examination of data through visualizations can provide insights that may not be apparent through numerical summaries alone. It encourages a balanced and comprehensive approach to data analysis, combining statistical rigor with graphical exploration for a more understanding of the underlying patterns in the data

The purpose of Anscombe's quartet is to highlight the limitations of relying solely on summary statistics and to emphasize the importance of visual exploration in understanding datasets. Specifically. This helps to identify even if two datasets have the same mean, variance, and correlation, their underlying structures can be distinct. It demonstrates how outliers or influential points can significantly impact statistical measures and regression models

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

Pearson's correlation coefficient, often denoted as r or Pearson's r, is a measure that quantifies the strength and direction of a linear relationship between two continuous variables. Pearson's r assumes a linear relationship between the variables. If the relationship is non-linear, r may not accurately represent the strength of the association. The formula for calculating Pearson's correlation coefficient between two variables

X and Y is:

$r = \sum(Xi - \bar{X})(Yi - \bar{Y}) / \text{Square root}(\sum(Xi - \bar{X})(Yi - \bar{Y}))$

Xi and Yi are individual data point and X-bar, and Y-bar are the mean of X and Y respectively.

Consideration

1. Pearson's r assumes a linear relationship between the variables. If the relationship is non-linear, r may not accurately represent the strength of the association.
2. r can be influenced by outliers. Outliers may disproportionately affect the correlation coefficient, leading to potentially misleading interpretations
3. r does not assume normality in the data, but it may be sensitive to extreme values.

Python function for the r

```
import numpy as np

np.corrcoef(X, Y)[0, 1]
```

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

Scaling is a data preprocessing technique used in machine learning to standardize the range of independent variables or features of the dataset. The goal of scaling is to ensure that all variables contribute equally to the analysis and prevent variables with larger scales from dominating the modeling process. Scaling is essential when working with algorithms that are sensitive to the scale of input features, such as distance-based methods and gradient-based optimization algorithms.

Advantages

Scaling makes it easier to compare the coefficients or importance of different features within a model. This enhances the interpretability of the model, allowing practitioners to understand the relative impact of each feature on the target variable.

Scaling helps stabilize the covariance and correlation values between features. This is particularly important when performing certain analyses.

Regularization techniques, such as L1 or L2 regularization in linear models, are more stable and effective when features are on a similar scale. Scaling prevents some features from having disproportionately large regularization penalties

Scaling ensures consistency when evaluating models using metrics that are sensitive to the scale of the predictions, such as mean absolute error or root mean squared error.

Normalized Scaling: transforms values to a specific range (e.g., [0, 1]), Preserves the relative relationships between values. Sensitive to outliers.

X normalized=X-Min(X)/Max(x)-Min(X)

Python Code:

```
from sklearn.preprocessing import MinMaxScaler
```

```
scaler = MinMaxScaler()

X_normalized = scaler.fit_transform(X)
```

Standardized Scaling centers data around zero with a standard deviation of 1, making it less sensitive to outliers and suitable for algorithms assuming normally distributed features.

Xstandardized= X-Mean(X)/standard(X)

Python Code:

```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

X_standardized = scaler.fit_transform(X)
```

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

The Variance Inflation Factor (VIF) measures the extent to which the variance of an estimated regression coefficient increases due to collinearity in the predictor variables. A high VIF indicates that a predictor variable is highly correlated with other variables, making it challenging to isolate the individual effect of that variable on the dependent variable. While the VIF is a valuable diagnostic tool for detecting multicollinearity, it may sometimes produce infinite values. This occurs when perfect multicollinearity is present in the dataset

Perfect multicollinearity happens when one or more independent variables in a regression model can be exactly predicted by a linear combination of other variables. Mathematically, this is represented by a perfect correlation or a situation where one variable is a constant multiple of another.

1. One or more variables can be expressed as a linear combination of other variables.
2. In regression models with categorical variables represented as dummy variables, perfect multicollinearity can occur when one dummy variable is a perfect linear combination of others. This is known as the dummy variable trap.
3. Addressing perfect multicollinearity is essential to ensure the reliability and interpretability of regression analysis results. Strategies for dealing with perfect multicollinearity include:

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

A Quantile-Quantile plot is a graphical tool used to assess whether a given dataset follows a specified probability distribution, typically the normal distribution. It compares the quantiles of the empirical data to the quantiles of a theoretical distribution, such as the normal distribution. The Q-Q plot displays the relationship between the quantiles of the two distributions on a scatter plot, allowing visual inspection of how well the empirical data fit the theoretical distribution

In linear regression, it is assumed that the residuals (the differences between observed and predicted values) are normally distributed. Q-Q plots allow us to visually inspect whether the residuals conform to this assumption.

If the residuals follow a normal distribution, the points in the Q-Q plot should fall approximately along the diagonal line.

Departures from the diagonal line indicate deviations from normality, which may suggest violations of the assumption and may require further investigation or data transformation.