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)**

Weather situation has a significant effect on dependent variable

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

When we create dummy variables from a categorical variable using one-hot encoding, we transform the categorical variable into a set of binary variables that represent the presence of each category. This is very useful because many analytical models require numerical input.

However, including a binary variable for every single category can introduce multicollinearity in our dataset, especially when these dummy variables are used as predictors in a regression model. Multicollinearity occurs when one predictor variable in a model can be linearly predicted from the others with a substantial degree of accuracy. This undermines the statistical significance of an independent variable.

By using `drop_first=True`, you omit the first level of the categorical variable, thereby creating k-1 dummies out of k categorical levels. This reduces the redundancy among the variables (since the dropped category can be inferred from the others being all zero) and eliminates multicollinearity stemming from the inclusion of all the dummy variables. Additionally, the interpretation of the regression coefficients becomes easier because they represent the effect with respect to the reference category (the one that was dropped).

In essence, `drop_first=True` keeps the dataset less complex and the model more stable and interpretable. However, remember that the decision to drop a dummy variable should fit the context of the analysis and the models being used. In certain models or frameworks, such as tree-based models, multicollinearity is less of an issue and keeping all dummy variables may be appropriate.

As a manager, it's crucial to work with your data team to ensure that they are preprocessing data in a way that is optimal for the particular analysis or modeling task at hand.

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

Temp has highest correlation with the Cnt variable.

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

By keeping fowling in mind:-

1. Linearity: The assumption of linearity implies that there is a linear relationship between the independent variables and the dependent variable. This can be checked by plotting the residuals versus the predicted values. If the residual plot shows a random scatter without distinct patterns, this assumption is likely to be true. Additionally, we could use scatter plots to visualize the relationship between the independent variables and the dependent variables to ensure they are linear.

2. Normality of Residuals: Linear regression assumes that the residuals (differences between observed and predicted values) are normally distributed. We can check this by using a Q-Q plot (quantile-quantile plot) to compare the distribution of residuals to a normal distribution. If the points fall roughly along a diagonal line, the residuals are normally distributed.

3. Homoscedasticity: This refers to the assumption that the residuals have constant variance at every level of the independent variables. If the variance of the residuals is not constant, we have a condition known as heteroscedasticity, which violates linear regression assumptions. This can be visually assessed by looking at a plot of residuals versus predicted values or the independent variables; the spread of the residuals should be roughly the same across all values of the independent variables. Tools like the Breusch-Pagan test or White's test can also be used to statistically check for homoscedasticity.

4. Independence of Residuals: The residuals should not be correlated with each other; that is, the presence of one value should not predict the next. To check this, we can calculate the Durbin-Watson statistic, which is a test statistic used to detect the presence of autocorrelation in the residuals. Values between 1.5 and 2.5 typically suggest that autocorrelation is not a concern.

5. No multicollinearity: Multicollinearity occurs when two or more independent variables are highly correlated with each other, which can affect the stability and interpretation of the coefficients. We can check for multicollinearity by looking at correlation matrices or by examining Variance Inflation Factors (VIF). Values of VIF exceeding 5 or 10 signify potentially problematic multicollinearity.

After validating these assumptions, if any violations are detected, they need to be addressed, as they could impact the predictive performance and interpretation of the regression model. Addressing such issues might require transformations of variables, adding interaction terms, or even considering different modeling approaches that are robust to violations of these assumptions.

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

Temperature, weather situation and year.

General Subjective Questions

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

Linear regression is a statistical method that is used to model the relationship between a dependent variable and one or more independent variables. The algorithm predicts the dependent variable value (y) based on the values of the independent variable(s) (x). This is done by finding the linear equation that best fits the observed data.

The linear equation for the simplest form of linear regression, which involves a single independent variable, is:

$$y = \beta0 + \beta1*x + \varepsilon$$

Where:

- y is the dependent variable.

- x is the independent variable.

- $\beta0$ is the intercept (the value of y when $x = 0$).

- $\beta1$ is the slope of the line (the change in y for a one-unit change in x).

- ε represents the error term (the difference between the predicted and actual values).

For multiple linear regression, where there are multiple independent variables, the equation is:

$$y = \beta0 + \beta1*x1 + \beta2*x2 + ... + \beta n*xn + \varepsilon$$

Here, $x1, x2, ..., xn$ represent the independent variables, and $\beta1, \beta2, ..., \beta n$ are the coefficients for each independent variable.

The linear regression algorithm involves the following steps:

1. **Model Specification**: Choose the form of the regression equation and identify the variables to be included in the model.

2. **Parameter Estimation**: Estimate the coefficients (β0, β1, ..., βn) using a method such as Ordinary Least Squares (OLS). OLS finds the values of the coefficients that minimize the sum of squared residuals, where a residual is the difference between an observed value and the predicted value provided by the model.

3. **Model Fitting**: Assess how well the model fits the data by looking at metrics such as R-squared, which measures the proportion of variance in the dependent variable that can be explained by the independent variables.

4. **Validation**: Check the validity of the model by verifying assumptions (linearity, independence, homoscedasticity, normality of residuals) and using techniques such as cross-validation, where the model is tested on a subset of the data not used in estimating the parameters.

5. **Prediction**: Use the model to predict the dependent variable from new observations of the independent variables.

6. **Diagnosis and Refinement**: Investigate the performance and accuracy of the model. Possible issues might include underfitting (the model is too simple and doesn't capture the underlying data patterns) or overfitting (the model is too complex and captures noise instead of the underlying data pattern). Refine the model as needed by adding or removing variables, or by using different forms of regularization.

Linear regression has several assumptions that must hold for the model to be valid. Violations of these assumptions can lead to incorrect conclusions. These assumptions include:

- Linearity: The relationship between the independent and dependent variables is linear.

- Independence: The observations are independent of each other.

- Homoscedasticity: The variance of the residual is the same for any value of the independent variable.

- Normality: The residuals are normally distributed.

Linear regression is widely used due to its simplicity and interpretability. It is a foundational algorithm for many statistical analysis and predictive modelling tasks.

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

Anscombe's quartet is a set of four different datasets, each with eleven x,y pairs, that were constructed by the statistician Francis Anscombe in 1973. The datasets are a prime example of why graphical representation of data is essential before analysing it. Here's why:

1. **Summary Statistics Can Be Misleading**: Each of the four datasets in the quartet has nearly the same mean and variance for both the x and y variables, the same correlation coefficient (approximately 0.816), and linear regression lines with nearly identical slopes and intercepts. However, when graphed, each dataset looks very different. This showcases that statistical properties can appear identical despite the underlying data being fundamentally different.

2. **Visual Analysis Is Crucial**: When plotted, the datasets reveal four distinct patterns:

- The first dataset appears to be a simple scatterplot of points that seem to follow a linear relationship, with some scatter around the regression line. This is a scenario most would expect when the statistical parameters align as mentioned.

- The second dataset shows a clear non-linear relationship (parabolic). Here, using a linear model would be inappropriate, despite the statistics suggesting that a linear model fits the data well.

- The third dataset appears to be linear but with a distinct outlier which, if not identified, could greatly influence any interpretation of the data.

- The last dataset is perhaps the most unusual; it looks like a straight line except for one outlier point. All points fall on a horizontal line except for the outlier, which suggests that the correlation driven by the outlier is illusory.

3. **Implications for Analysis**: The lesson from Anscombe's quartet is that analysts must look at the data both numerically and graphically. Statistics alone often fail to reveal the underlying structure, trends, or patterns that can be critical for making informed decisions.

In summary, Anscombe’s quartet teaches us that while summary statistics are useful, they do not replace the need for a more comprehensive approach to data analysis, which includes visual inspection of the datasets. As a manager, applying this diligence ensures that decisions are not made based on misleading statistics, but rather on a thorough understanding of the actual data at hand.

3. What is Pearson’s R? (3 marks)

Pearson's r, also known as Pearson Product-Moment Correlation Coefficient (PPMCC) or simply Pearson's correlation coefficient, is a statistic that measures the linear correlation between two variables, X and Y. It has a value between +1 and -1, where:

- 1 indicates a perfect positive linear correlation,

- 0 indicates no linear correlation, and

- -1 indicates a perfect negative linear correlation.

The formula to calculate Pearson's r is:

$$r = \Sigma[(Xi - \bar{X})(Yi - \bar{Y})] / [\surd\Sigma(Xi - \bar{X})^2 * \surd\Sigma(Yi - \bar{Y})^2]$$

where:

- Xi and Yi are the individual sample points indexed with i,

- $\bar{X}$ is the mean of the X samples, and

- $\bar{Y}$ is the mean of the Y samples.

In simple terms, Pearson's r assesses how well a linear equation can describe the relationship between two variables. If r is close to +1 or -1, it indicates a strong relationship where changes in one variable are closely related to changes in the other. If r is near zero, it suggests that there is little to no linear relationship between the variables. Pearson's r is commonly used in statistical analysis to determine the strength and direction of a linear relationship.

As a manager, you could use Pearson's r to analyze correlations between various business metrics. For example, you might examine the correlation between marketing spend and sales revenue or between employee satisfaction scores and productivity levels. Understanding the strength and direction of these relationships could guide strategic decisions and resource allocation.

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

Scaling in a business or technical context generally refers to the process of adjusting the size or capability of a company, system, or process to handle different levels of demand. Scaling can be executed in various contexts, including infrastructure, operations, staffing, or production capacity. Companies perform scaling to ensure that they can efficiently handle an increase in workloads, user traffic, or market demand without compromising on performance or service quality.

In data analysis and machine learning, scaling refers to the process of adjusting the range of independent variable or feature values. The primary objectives of feature scaling are to normalize or standardize the range of independent variables or features of data. It is an important pre-processing step for many machine learning algorithms because it can dramatically impact the performance of the models.

The two common types of scaling are normalized scaling and standardized scaling:

1. Normalized scaling, also known as Min-Max scaling, rescales the feature to a range usually between 0 and 1. The formula for calculating the normalized value of an element x is:

`x_normalized = (x - min(x)) / (max(x) - min(x))`

where min(x) is the minimum value in the feature column x and max(x) is the maximum value. Normalization ensures that each feature contributes approximately proportionately to the final distance between data points.

2. Standardized scaling, or Z-score normalization, involves rescaling the features so that they have the properties of a standard normal distribution with a mean of 0 and a standard deviation of 1. The formula for calculating the standardized value of an element x is:

`x_standardized = (x - mean(x)) / std(x)`

where mean(x) is the average value and std(x) is the standard deviation of the feature column x. Standardization is useful especially for algorithms that assume a Gaussian distribution in the features and weigh inputs based on their variance accordingly.

In summary, scaling is performed to prepare data for use in machine learning models and to ensure that variables are on comparable scales. Normalized scaling brings values into a bounded interval, useful for algorithms that are sensitive to input magnitude, such as neural networks and those that use distance measures, like k-NN. Standardized scaling, on the other hand, is useful when the data needs to have a normal distribution, often a prerequisite for optimization algorithms in machine learning such as gradient descent.

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

(3 marks)

Certainly. VIF stands for Variance Inflation Factor, which is a measure used to detect the presence and intensity of multicollinearity in regression analyses. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated with each other, which can affect the reliability and interpretability of the coefficients of the model.

The VIF for an independent variable is calculated as:

\[VIF_i = \frac{1}{1 - R^2_i} \]

where \(R^2_i \) is the coefficient of determination of a regression of independent variable \(i \) on all the other independent variables.

The VIF value becomes infinite when the \(R^2_i \) is exactly 1. This occurs when the independent variable \(i \) is a perfect linear combination of the other independent variables in the model, meaning it can be exactly predicted from the others with no error. In other words, there is perfect multicollinearity.

When the VIF is infinite, it indicates that one or more of the variables should be removed from the model to ensure that the regression coefficients and standard errors are reliable and the model is correctly specified. It is not possible to estimate the effects of predictors on the dependent variable separately when there is perfect multicollinearity.

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

(3 marks

A Q-Q plot, or Quantile-Quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, Exponential, or Uniform distribution. It plots the quantiles of the data set against the quantiles of the chosen theoretical distribution. If the points lie on the 45-degree reference line, it suggests that the data follows the selected theoretical distribution.

In the context of linear regression, the Q-Q plot is often used to validate one of the key assumptions—the normality of residuals. The residuals of a regression model are the differences between observed and predicted values by the model. For the best performance and validity of statistical tests associated with regression, it is assumed that the residuals are normally distributed.

By using a Q-Q plot to compare the empirical quantiles of the residuals to the theoretical quantiles of a normal distribution, we can visually assess this normality assumption. If the residuals are normally distributed, the points should fall approximately along a straight line.

The importance of a Q-Q plot in linear regression includes:

1. Model Validation: It helps in validating the normality of residuals, which is a key assumption in linear regression analysis.

2. Diagnosing Issues: It can help identify problems with the distribution of the residuals that may suggest model misspecification, such as skewness, outliers, or a need for transformation of variables.

3. Informing Transformations: If the residuals do not follow a normal distribution, a Q-Q plot can suggest the type of transformation that might bring the residuals closer to normality.

The Q-Q plot is an essential diagnostic tool because if the linear regression assumptions are violated, the reliability of the inference based on the regression model, like confidence intervals and hypothesis tests, may be compromised.