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)

Answer: Below are my inferences about the categorical variables on the dependent variables:

i. Weathersit: Demand for shared bikes increases during clear weather which has the highest median. Demand decreases when the weather becomes cloudy, misty and starts with light snow. There is no demand during heavy snow.

ii. Season: Maximum shared bike demand is seen during the fall season (which has the highest median) followed by summer and winter and least demand in spring season.

iii. Months: Higher demand is seen during the month of June till October with highest demand in the month of September. There is sluggish demand during the month of December, January, and February.

iv. Weekday: There isn't much deviation in median between different weekdays. Saturday sees a little higher demand compared to all other weekdays.

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

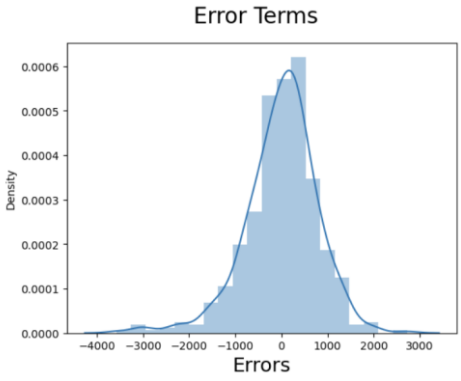
Answer: A variable with n levels can be represented by n-1 dummy variables. What it means is, even if we remove the first column, it can still represent the data. If for example n = 3 and we remove the 1st column. If column 2 & 3 are zeros, it automatically implies that 1st column is 1. It thus, removes redundancy.

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

Answer: "temp" or temperature has the highest correlation with the target variable.

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

Answer: We validate the assumptions of Linear Regression after building the model on the training set by plotting a distplot of the residual and analysing it to see if it is a normal distribution or not and if it has a mean of zero. Below is the diagram from the Bike sharing assignment which shows a normal distribution with mean = 0.



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

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

i. Temp

ii. Year

iii. Winter Season

General Subjective Questions

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

Answer: Linear regression is a fundamental statistical technique used for predicting a dependent variable based on one or more independent variables. Below is the detailed explanation of the algorithm:

1. Model

The relationship between the dependent variable y and the independent variables X is based on below formula:

$y=\beta0+\beta1x1+\beta2x2+\ldots+\beta nxn+\epsilon$

where:

- y = dependent/target variable.
- x1,x2,…,xn are the independent variables/features.
- β0 is the intercept term.
- β1,β2,…,βn are the coefficients for the independent variables.
- ϵ is the error term (residual), representing the difference between the observed and predicted values.

2. Assumptions

For linear regression to provide reliable results, the following assumptions must hold:

- **Linearity**: The relationship between the dependent and independent variables should be linear.
- **Independence**: The residuals (errors) should be independent.
- **Homoscedasticity**: The residuals should have constant variance (homoscedasticity).
- **Normality**: The residuals should be normally distributed (for hypothesis testing).
- **No multicollinearity**: Independent variables should not be highly correlated.

3. Estimation of Coefficients

The goal of linear regression is to estimate the coefficients β0,β1,…,βn such that the error term ϵ is minimized. This is typically done using the **Ordinary Least Squares (OLS)** method, which minimizes the sum of the squared residuals. OLS function is available in python and can be called through Statsmodels.

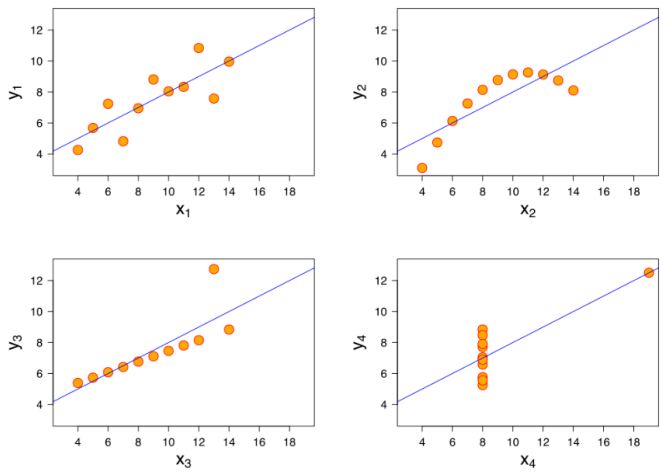
4. Evaluation of the Model

After estimating the coefficients, it is crucial to evaluate the model's performance. Common metrics include:

- **R-squared**: Represents the proportion of the variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1, with higher values indicating a better fit.
- **Adjusted R-squared**: Adjusts R2R^2R2 for the number of predictors in the model, providing a more accurate measure for multiple regression.
- **Root Mean Squared Error (RMSE)**: The square root of MSE, giving an indication of the average error in the same units as the dependent variable.
- **Residual Analysis**: Involves examining the residuals to check if the assumptions of linear regression hold, such as plotting residuals vs. fitted values to check for homoscedasticity and normality.

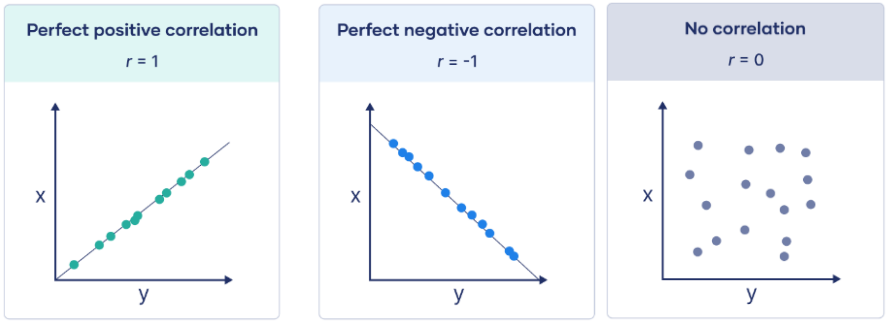
2. Explain the Anscombe's quartet in detail. (3 marks)

Answer: Anscombe's quartet consists of four datasets that have nearly identical simple descriptive statistics, such as mean, variance, correlation, and linear regression line. Despite these similarities, when graphed, they reveal strikingly different distributions and relationships between variables. This illustrates the importance of visualizing data rather than just relying solely on summary statistics. The quartet demonstrates that datasets with the same statistical properties can exhibit vastly different patterns, highlighting the necessity of visual analysis in identifying underlying structures, anomalies, and insights that summary statistics alone might not easily provide.



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

Answer: Pearson's R, also known as the Pearson correlation coefficient, measures the linear relationship between two continuous variables. It ranges from -1 to 1, where 1 indicates a perfect positive linear correlation, -1 indicates a perfect negative linear correlation, and 0 indicates no linear correlation. Pearson's R is calculated as the covariance of the two variables divided by the product of their standard deviations. It provides insights into the strength and direction of the relationship, helping to identify how one variable changes in relation to another. It is commonly used in statistical analysis and data exploration.



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

Answer: Scaling is the process of adjusting the range of features in a dataset to a common scale without distorting differences in the ranges of values. Scaling is performed to improve the performance and training stability of machine learning algorithms, particularly those sensitive to feature magnitudes like gradient descent-based methods and distance-based algorithms.

Normalized scaling (Min-Max scaling) transforms data to a specific range, typically [0, 1]:

$x' = (x - \min(x)) / (\max(x) - \min(x))$

Standardized scaling (Z-score scaling) transforms data to have a mean of 0 and a standard deviation of 1:

$x' = (x - \mu) / \sigma$

Normalization rescales data, while standardization centers and scales data based on statistical properties.

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

Answer: When there is a perfect multicollinearity, we have VIF = infinity.

Perfect multicollinearity occurs when one predictor variable is an exact linear combination of one or more other predictors. This situation makes it impossible to isolate the individual contribution of any one predictor, leading to instability in the regression coefficients.

In the case of perfect correlation, we get R squared = 1, which lead to 1/(1 – R2) = infinity.

For solving this, one must remove or combine collinear predictors.

6.What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer: A Q-Q (Quantile-Quantile) plot is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, typically the normal distribution. In a Q-Q plot, the quantiles of the sample data are plotted against the quantiles of the theoretical distribution.

Use in Linear Regression:

Assess Normality: In linear regression, one of the assumptions is that the residuals (errors) are normally distributed. A Q-Q plot helps assess this assumption by showing if the residuals follow a straight line (indicative of normality).

Identify Outliers: Deviations from the straight line can indicate outliers or heavy-tailed distributions.

Importance: Ensuring that residuals are normally distributed validates the statistical tests used in regression analysis, such as confidence intervals and hypothesis tests, making the model's predictions and inferences more reliable.