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?

Ans. After analyzing categorical variables demand of bikes is less in winter months like Nov, Dec, Jan and spring. Weather like mist and light snow has negative impact on the demand of bikes. Working day like Mon has positive impact.

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

Ans. Its important to use drop_first =True to handle dummy variable trap due to multicollinearity and any one value's column can be identified with other remaining value's column.

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

Ans. Highest correlation is with temp with target vairiable

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

Ans. I did the validation by doing residual analysis and plotting distribution graph for error to see if they are distributed normally or not.

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

Ans. Top 3 features are temp, light snow and spring

General Subjective Questions

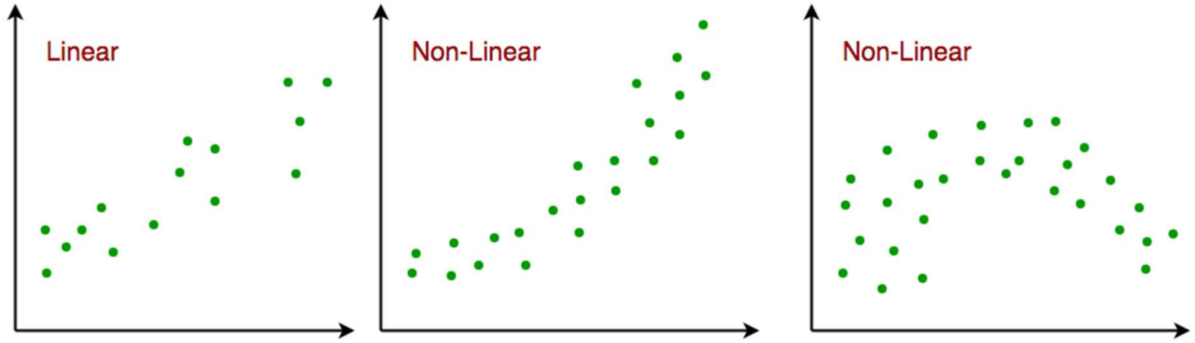
1. **Explain the linear regression algorithm in detail**

Ans. Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features. When the number of the independent feature, is 1 then it is known as Univariate Linear regression, and in the case of more than one feature, it is known as multivariate linear regression.

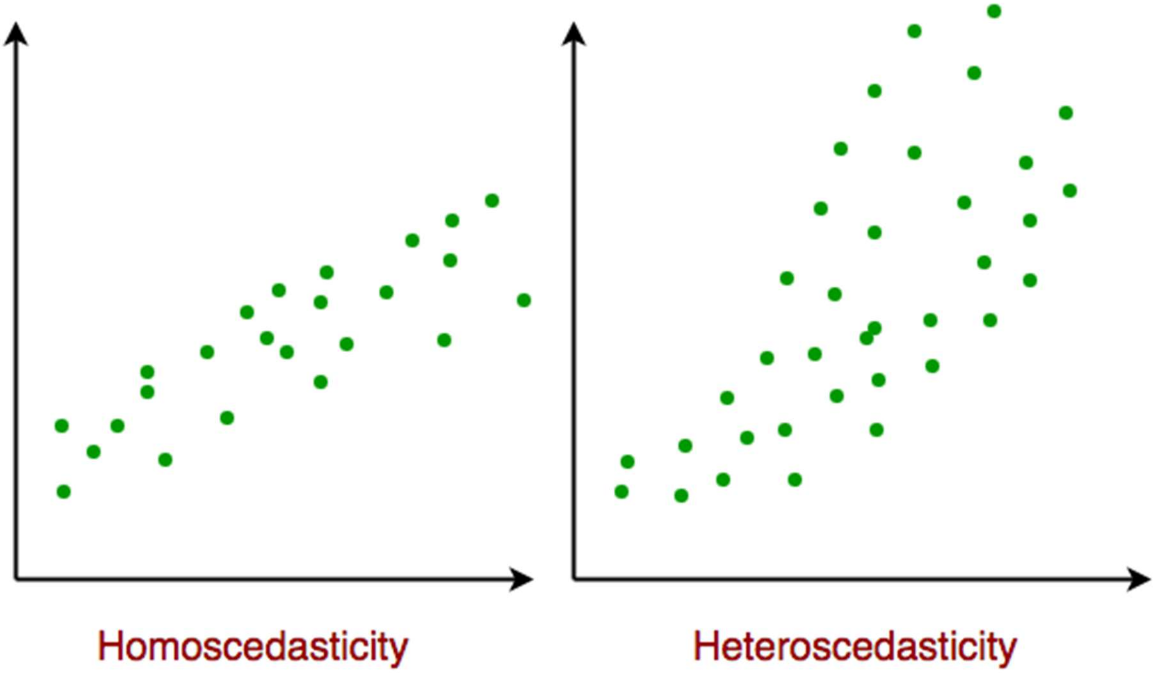
Assumption for Linear Regression Model
Linear regression is a powerful tool for understanding and predicting the behaviour of a variable, however, it needs to meet a few conditions in order to be accurate and dependable solutions.

1. Linearity: The independent and dependent variables have a linear relationship with one another. This implies that changes in the dependent variable follow those in the independent variable(s) in a linear fashion. This means that there should be a straight line that can be drawn through the data points. If the relationship is not linear, then

linear regression will not be an accurate model.



2. Independence: The observations in the dataset are independent of each other. This means that the value of the dependent variable for one observation does not depend on the value of the dependent variable for another observation. If the observations are not independent, then linear regression will not be an accurate model.
3. Homoscedasticity: Across all levels of the independent variable(s), the variance of the errors is constant. This indicates that the amount of the independent variable(s) has no impact on the variance of the errors. If the variance of the residuals is not constant, then linear regression will not be an accurate model.



Homoscedasticity in Linear Regression

4. Normality: The residuals should be normally distributed. This means that the residuals should follow a bell-shaped curve. If the residuals are not normally distributed, then linear regression will not be an accurate model.

5. No multicollinearity: There is no high correlation between the independent variables. This indicates that there is little or no correlation between the independent variables. Multicollinearity occurs when two or more independent variables are highly correlated with each other, which can make it difficult to determine the individual effect of each variable on the dependent variable. If there is multicollinearity, then linear regression will not be an accurate model.

Types of Linear Regression

There are two main types of linear regression:

- Simple Linear Regression: This is the simplest form of linear regression, and it involves only one independent variable and one dependent variable. The equation for simple linear regression is:

  where:
  - Y is the dependent variable
  - X is the independent variable
  - β0 is the intercept
  - β1 is the slope
- Multiple Linear Regression: This involves more than one independent variable and one dependent variable. The equation for multiple linear regression is:

  where:
  - Y is the dependent variable
  - X1, X2, ..., Xp are the independent variables
  - β0 is the intercept
  - β1, β2, ..., βn are the slopes

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation, and linear regression) but appear very different when graphed. This collection of datasets was created by the statistician Francis Anscombe in 1973 to illustrate the importance of visualizing data and the limitations of relying solely on summary statistics. The quartet highlights the concept that datasets with similar statistical properties can exhibit vastly different patterns when graphically represented.

Here are the details of Anscombe's quartet:

Dataset 1:

Summary Statistics:

Mean of x: 9.0

Mean of y: 7.5

Variance of x: 11.0

Variance of y: 4.12

Correlation between x and y: 0.816

Linear regression line:

y=3.0+0.5x

Dataset 2:

Summary Statistics:

Mean of x: 9.0

Mean of y: 7.5

Variance of x: 11.0

Variance of y: 4.12

Correlation between x and y: 0.816

Linear regression line: y=3.0+0.5x

Dataset 3:

Summary Statistics:

Mean of x: 9.0

Mean of y: 7.5

Variance of x: 11.0

Variance of y: 4.12

Correlation between x and y: 0.816

Linear regression line:

y=3.0+0.5x

Dataset 4:

Summary Statistics:

Mean of x: 9.0

Mean of y: 7.5

Variance of x: 11.0

Variance of y: 4.12

Correlation between x and y: 0.816

Linear regression line:

$y=3.0+0.5x$

Key Observations:

Despite having identical summary statistics, the datasets exhibit different patterns when graphed. Dataset 1 shows a linear relationship, Dataset 2 a non-linear relationship, Dataset 3 a perfect linear relationship with an outlier, and Dataset 4 a relationship heavily influenced by a single data point. The importance of visual inspection is highlighted by the fact that relying solely on summary statistics might mask important aspects of the data.

Implications:

Anscombe's quartet emphasizes the need for exploratory data analysis and the use of graphical representations to understand the underlying patterns in data.It cautions against blindly relying on summary statistics, as datasets with different structures can produce similar numerical summaries. In essence, Anscombe's quartet serves as a reminder that while statistical measures provide valuable insights, visualizing data is crucial for gaining a comprehensive understanding of its characteristics and relationships.

3. **What is Pearson's R?**

Ans. Pearson's correlation coefficient, often denoted by r, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It is named after Karl Pearson, who introduced the concept in the late 19th century. Pearson's r ranges from -1 to 1, where:

$r=1$ indicates a perfect positive linear relationship.

$r=-1$ indicates a perfect negative linear relationship.

$r=0$ indicates no linear relationship.

The Pearson's correlation coefficient formula is

$r = [n(\Sigma xy) - \Sigma x\Sigma y]/$Square root of$\sqrt{}[n(\Sigma x2) - (\Sigma x)2][n(\Sigma y2) - (\Sigma y)2]$

In this formula, *x* is the independent variable, *y* is the dependent variable, *n* is the sample size, and Σ represents a summation of all values.

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

Ans. In the context of linear regression, scaling refers to the process of transforming the input features (independent variables) to a similar scale. This is typically done to ensure that all features contribute equally to the computation of the regression coefficients and to improve the numerical stability and convergence of the optimization algorithm used to estimate those coefficients.

Here are the reasons why scaling is performed in linear regression:

1. Equal Contribution of Features:
   - When features have different scales, the impact of one feature on the regression coefficients can dominate over others. Scaling ensures that all features contribute more evenly to the model, preventing bias towards variables with larger magnitudes.
2. Numerical Stability and Convergence:
   - The optimization algorithms used to estimate the coefficients (such as gradient descent) often converge faster and more reliably when features are on a similar scale. Large differences in scale can result in slow convergence or prevent the algorithm from converging altogether.
3. Interpretability:
   - Scaling does not affect the relationship between the features and the dependent variable. It only standardizes the scale of the features. However, scaling can make the interpretation of the coefficients more straightforward, as the coefficients represent the change in the dependent variable associated with a one-unit change in the corresponding standardized feature.

Both normalized scaling and standardized scaling aim to bring the features to a similar scale, they differ in terms of the specific transformation applied.

Normalized Scaling (Min-Max Scaling):

1. Range:
   - Normalized scaling, also known as Min-Max scaling, scales the features to a specific range, typically between 0 and 1.
2. Formula:
   - The formula for normalized scaling is: Scaled value=Original value−MinMax−MinScaled value=Max−MinOriginal value−Min
   - Here, Min and Max are the minimum and maximum values of the feature, respectively.
3. Scale Interpretation:
   - After normalization, the minimum value of the feature becomes 0, and the maximum becomes 1.
4. Use Case:
   - Normalized scaling is often suitable when the features have a bounded range, and the goal is to bring all features to a uniform scale while preserving the relative relationships between data points.

Standardized Scaling (Z-score Normalization):

1. Distribution:
   - Standardized scaling, also known as Z-score normalization, scales the features to have a mean of 0 and a standard deviation of 1.
2. Formula:

- The formula for standardized scaling is:
  Scaled value=Original value−MeanStandard DeviationScaled value=Standard DeviationOriginal value−Mean
- Here, Mean is the mean of the feature, and Standard Deviation is its standard deviation.

3. Scale Interpretation:
   - After standardization, the mean of the feature becomes 0, and its standard deviation becomes 1.
4. Use Case:
   - Standardized scaling is often preferred when the distribution of the features is approximately normal or when working with algorithms that assume a normal distribution. It is less sensitive to outliers compared to Min-Max scaling.

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

Ans. The VIF quantifies how much the variance of the estimated regression coefficients is inflated due to multicollinearity.

The formula for VIF for an independent variable Xi is given by:

VIF=1 /1−2VIFi=1−Ri21

where Ri2 is the R2 value obtained by regressing Xi against all other independent variables.

Now, if the Ri2 is very close to 1 (or exactly 1), it leads to a denominator close to zero, resulting in a situation where the VIF approaches infinity. This occurs when one or more independent variables in the model can be perfectly predicted by a linear combination of the other independent variables.

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

Ans. A Q-Q plot, short for quantile-quantile plot, is a graphical tool used in statistics to assess whether a dataset follows a specific theoretical distribution. It compares the quantiles of the observed data against the quantiles of a theoretical distribution, typically a normal distribution. Q-Q plots are useful for checking the assumption of normality and identifying departures from normality in a dataset.

Structure of a Q-Q Plot:

- The x-axis represents the quantiles of a theoretical distribution (often the standard normal distribution).
- The y-axis represents the quantiles of the observed data.

Use and Importance in Linear Regression:

1. Assessing Normality:
    - Q-Q plots are commonly used to check whether the residuals of a linear regression model are normally distributed. Normality of residuals is a key assumption in linear regression. If the points on the Q-Q plot roughly fall along a straight line, it suggests that the residuals follow a normal distribution.
2. Identifying Departures from Normality:
    - Departures from normality are revealed by deviations from a straight line in the Q-Q plot. For example, if the points deviate upward or downward at the tails, it indicates that the residuals have heavier or lighter tails than a normal distribution.
3. Detecting Outliers:
    - Q-Q plots can help in identifying outliers or extreme values. Outliers may manifest as points that deviate significantly from the expected straight line pattern.
4. Model Diagnostics:
    - Q-Q plots are a valuable tool in model diagnostics. They provide a visual indication of how well the assumed distribution (normality) aligns with the actual distribution of the residuals.
5. Comparing Distributions:
    - Q-Q plots can be used to compare the distribution of one dataset to another. This is particularly useful when comparing empirical data to a theoretical distribution or when comparing residuals from different models.