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:

I have done analysis on categorical columns using the boxplot and bar plot. Below are the few points we can infer from the visualization –

- Most number of bookings are made in Fall season.
- September has highest number of bookings while median for bike rental count is highest for July. Trend of booking bikes increases from July and by October the booking trend decreases as winter approaches.
- Clear weather attracted more booking which seems obvious.
- Thu, Fir, Sat and Sun have more number of bookings as compared to the start of the week.
- When it's not holiday, booking seems to be less in number which seems reasonable as on holidays, people may want to spend time at home and enjoy with family.
- Booking seemed to be almost equal either on working day or non-working day.
- 2019 attracted more number of booking from the previous year, which shows good progress in terms of business.

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

Answer:

Using drop_first=True during dummy variable creation is important to avoid the issue of multicollinearity in regression models. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, leading to redundant information and difficulties in interpreting the model.

When creating dummy variables, the drop_first=True option removes one of the dummy variables for each categorical variable. By dropping the first dummy variable, we create a reference category, and the remaining dummy variables represent the presence or absence of other categories. This approach helps to ensure that there is no perfect correlation among the dummy variables.

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

Answer:

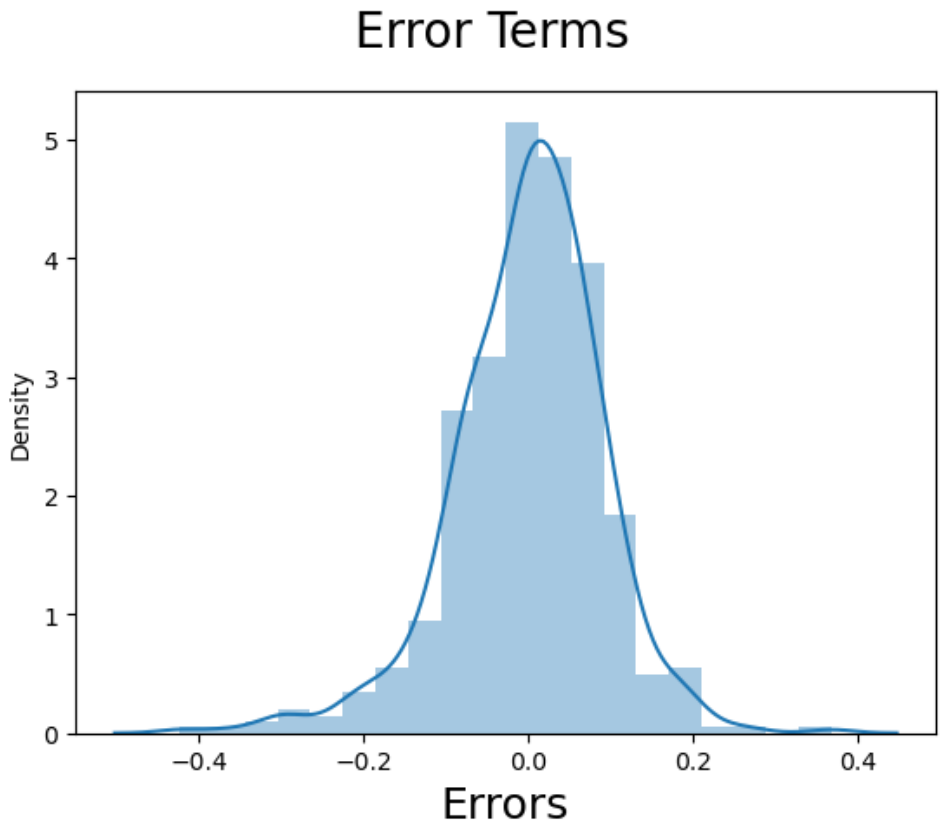
Looking at the pairplot before building the model, 'temp' and 'atemp' variable has the highest correlation with the target variable.

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

Answer:

To validate assumptions of the model, and hence the reliability for inference, we go with the following procedures:

- **Residual Analysis:** We need to check if the error terms are also normally distributed (which is in fact, one of the major assumptions of linear regression). I have plotted the histogram of the error terms and this is what it looks like:



- **Linear relationship between predictor variables and target variable:** This is happening because all the predictor variables are statistically significant (p-values are less than 0.05). Also, R-Squared value on training set is 0.832 and adjusted R-Squared value on training set is 0.8075. This means that variance in data is being explained by all these predictor variables.
- **Error terms are independent of each other:** Handled properly in the model. The predictor variables are independent of each other. Multicollinearity issue is not there because VIF (Variance Inflation Factor) for all predictor variables are below 5.

Question 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 –

- temp (Coef: 0.4498)
- year (Coef: 0.2342)
- sep (Coef: 0.0573)

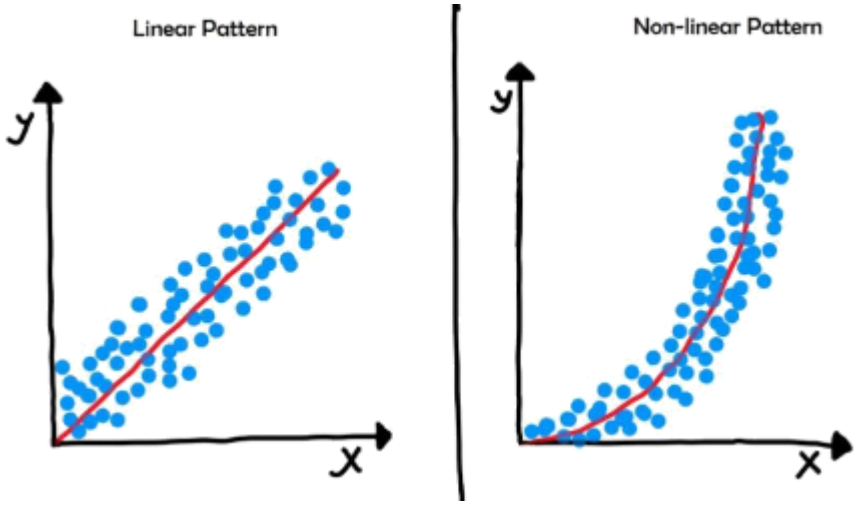
General Subjective Questions

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

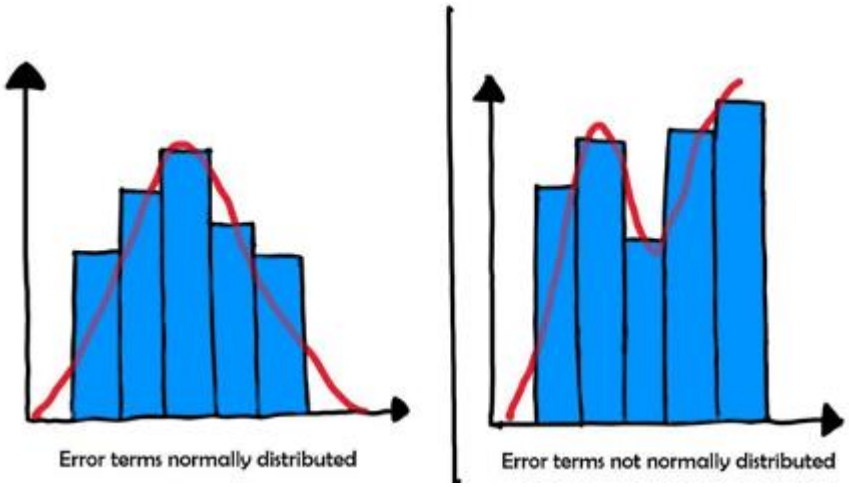
Answer:

Linear regression is a popular and widely used algorithm for predicting a continuous numeric output based on one or more input variables. It assumes a linear relationship between the input variables and the output variable. Here's a detailed explanation of the linear regression algorithm:

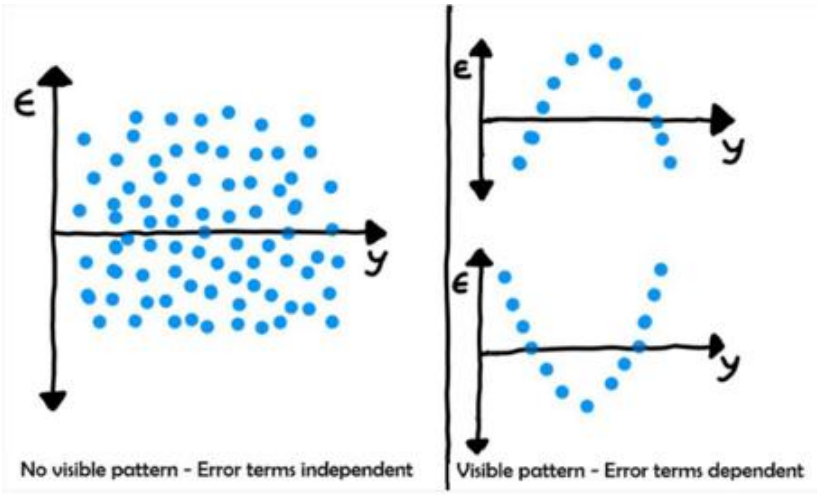
- Data Preparation:
  - Gather a dataset that contains the input variables (also called features or independent variables) and the corresponding output variable (also called the target or dependent variable).
  - Split the dataset into a training set and a test set. The training set is used to train the linear regression model, while the test set is used to evaluate its performance.
- Model Representation:
  - Linear regression aims to find the best-fitting linear equation that represents the relationship between the input variables and the output variable.
  - In its simplest form, linear regression is represented by the equation: $Y = \beta0 + \beta1*X1 + \beta2*X2 + ... + \betan*Xn$, where Y is the predicted output variable, β0 is the intercept, and β1 to βn are the coefficients corresponding to X1 to Xn (input variables).
- Model Training:
  - During training, the linear regression model adjusts the values of the coefficients (β0, β1, β2, ..., βn) to minimize the difference between the predicted values and the actual values in the training data.
  - The most common approach to training is called Ordinary Least Squares (OLS), which minimizes the sum of squared differences between the predicted values and the actual values.
- Model Evaluation:
  - Once the model is trained, it is evaluated using the test set to assess its performance on unseen data.
  - Common evaluation metrics for linear regression include mean squared error (MSE), root mean squared error (RMSE), and R-squared, which measures the proportion of variance in the target variable that is explained by the model.
- Making Predictions:
  - After the model is trained and evaluated, it can be used to make predictions on new or unseen data.
  - Given new input values, the model applies the learned coefficients to calculate the predicted output variable.
- Assumptions of Linear Regression:
  - The assumption about the form of the model: It is assumed that there is a linear relationship between the dependent and independent variables.

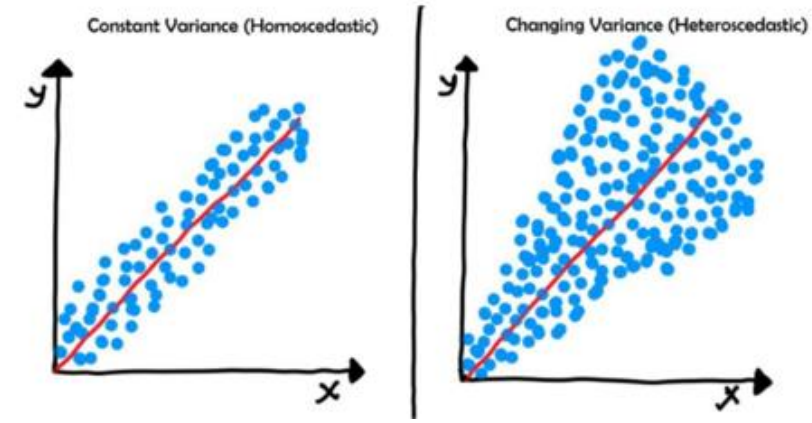


- Assumptions about the residuals:
  - Normality assumption: It is assumed that the error terms, ε(i), are normally distributed.
  - Zero mean assumption: It is assumed that the residuals have a mean value of zero, i.e., the error terms are normally distributed around zero.
  - Constant variance assumption: It is assumed that the residual terms have the same (but unknown) variance, sigma square. This assumption is also known as the assumption of homogeneity or homoscedasticity.
  - Independent error assumption: It is assumed that the residual terms are independent of each other, i.e., their pair-wise covariance is zero.



- Assumptions about the estimators:
  - The independent variables are measured without error.
  - The independent variables are linearly independent of each other, i.e., there is no multicollinearity in the data.





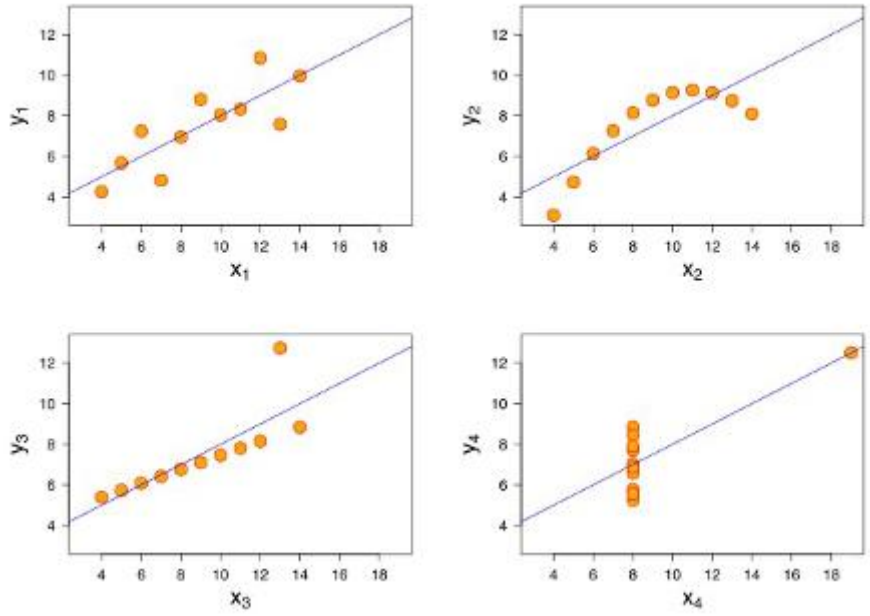
- Variations and Extensions:
    - There are variations and extensions to linear regression, such as multiple linear regression (with more than one input variable), polynomial regression (using polynomial terms), and regularization techniques like Ridge regression and Lasso regression to handle overfitting and improve generalization.

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

Answer:

Anscombe's quartet refers to a set of four datasets that have nearly identical simple descriptive statistics but exhibit significantly different patterns when visualized and analyzed. These datasets were created by the statistician Francis Anscombe in 1973 to highlight the importance of data visualization and the limitations of relying solely on summary statistics.

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed.

- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

Question 3: What is Pearson's R? (3 marks)

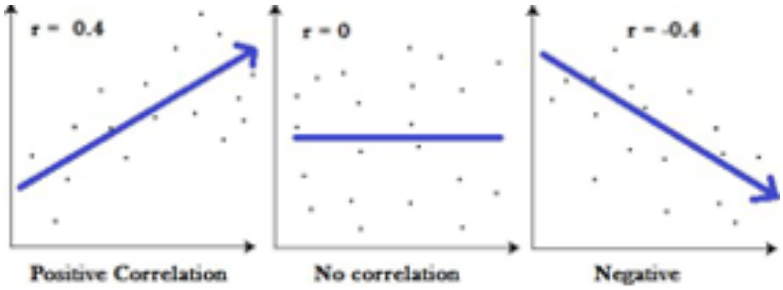
Answer:

Pearson's correlation coefficient, commonly referred to as Pearson's R or simply as the correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It is named after Karl Pearson, who developed this measure in the late 19th century.

Pearson's R is a value between -1 and +1, where:

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

- A value of -1 indicates a perfect negative linear relationship between the variables, meaning that as one variable increases, the other variable decreases proportionally.
- A value of 0 indicates no linear relationship between the variables, suggesting that there is no systematic change in one variable when the other variable changes.



Key features of Pearson's R:

- Linearity: Pearson's R measures the linear association between variables. It assumes that the relationship between the variables can be represented by a straight line.
- Scale-invariant: Pearson's R is unaffected by changes in the scale or units of measurement of the variables.
- Symmetry: The correlation coefficient is symmetric, meaning that the correlation between variable A and variable B is the same as the correlation between variable B and variable A.
- Affected by outliers: Extreme outliers can have a strong influence on the correlation coefficient, as it is sensitive to extreme values.

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

Answer:

Scaling, in the context of data preprocessing, refers to the process of transforming numerical variables to a specific range or distribution. It is performed to ensure that all variables are on a comparable scale and to avoid issues that may arise due to differences in the units or magnitude of the variables. Scaling is particularly important in machine learning algorithms that are sensitive to the scale of input features.

The primary reasons for performing scaling are:

- Avoiding Magnitude Bias: Variables with larger magnitudes can dominate the learning process and have a disproportionate impact on the results. Scaling helps to equalize the influence of different variables, preventing the bias caused by variables with larger scales.
- Enabling Comparisons: Scaling allows for meaningful comparisons between variables. Variables on different scales may have different ranges of values, making it challenging to compare their effects or interpret their relative importance.
- Enhancing Convergence: Scaling can aid the convergence of optimization algorithms used in machine learning models. When variables are on similar scales, optimization algorithms can reach the optimal solution more quickly.

Two commonly used scaling techniques are normalized scaling (also known as MinMax scaling) and standardized scaling (also known as z-score scaling):

1. Normalized Scaling (MinMax Scaling):
    a. Normalized scaling transforms variables to a specific range, typically between 0 and 1.
    b. It calculates the scaled value for each data point by subtracting the minimum value of the variable and dividing it by the range (maximum value minus minimum value).
    c. Normalized scaling preserves the shape of the original distribution and ensures that the minimum value is transformed to 0 and the maximum value to 1.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

2. Standardized Scaling (Z-score Scaling):
    a. Standardized scaling standardizes variables to have a mean of 0 and a standard deviation of 1.
    b. It calculates the scaled value for each data point by subtracting the mean of the variable and dividing it by the standard deviation.
    c. Standardized scaling transforms the variable distribution to have a mean of 0 and a standard deviation of 1, making it more suitable for certain statistical analyses and algorithms that assume standard normal distribution.

$$x' = \frac{x - \bar{x}}{\sigma}$$

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

Answer:

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

Question 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, also known as a normal probability plot, is a graphical tool used to assess the distributional similarity between a given dataset and a theoretical distribution, typically the normal distribution. It is a useful diagnostic tool in linear regression and other statistical analyses to check the assumption of normality of residuals or other variables.

A Q-Q plot in linear regression is used for the following reasons:

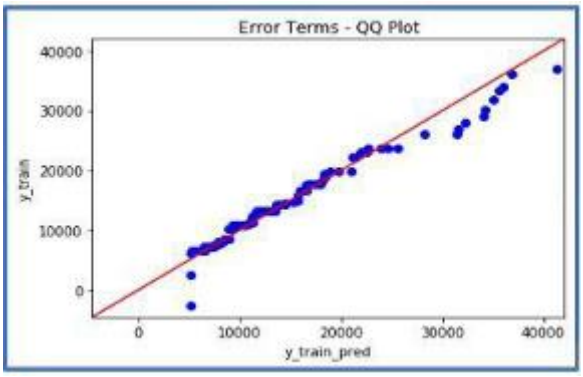
1. Assessing Normality:
   - In linear regression, one of the key assumptions is that the residuals (the differences between observed and predicted values) follow a normal distribution.
   - The Q-Q plot allows us to visually examine whether the residuals conform to a normal distribution or deviate significantly from it.
   - By plotting the quantiles of the residuals against the quantiles of the standard normal distribution, we can identify any departures from normality.

2. Interpretation of Q-Q Plot:

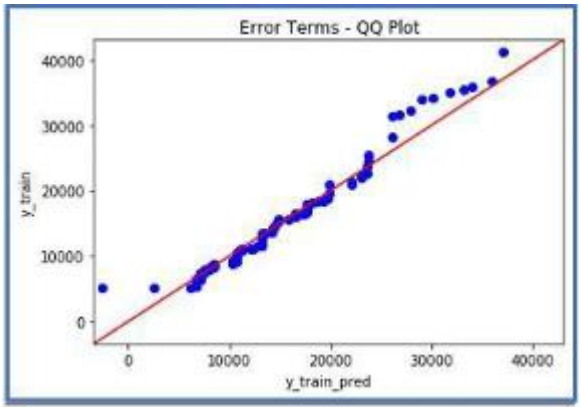
   A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

   Below are the possible interpretations for two data sets.

   - Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x –axis
   - Y-values < X-values: If y-quantiles are lower than the x-quantiles.



   - X-values < Y-values: If x-quantiles are lower than the y-quantiles.



- Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x –axis

3. Decision-Making:
    - The Q-Q plot allows researchers to make informed decisions about the normality assumption and the appropriateness of linear regression.
    - If the Q-Q plot indicates that the residuals significantly depart from normality, it may be necessary to explore alternative regression models or consider transforming the variables to achieve normality.
4. Robust Inference:
    - If the assumptions of normality are met, it enables the use of various statistical tests and inference procedures that rely on the assumption of normality, such as hypothesis testing and confidence intervals.
    - Assessing normality through the Q-Q plot ensures the validity and reliability of the statistical inferences made from the linear regression analysis.