Assignment-based Subjective Questions – Solutions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans. Following are the effects for categorical variables on the dependent variable:

a. Company should focus on expanding business during Fall, Summer and Winter

b. September month has shown great demand.

c. It has been observed that the demand for bike rentals had gone up from 2018 to 2019. So we can say that it will go up once the situation gets normal post Covid

d. There would be less bookings during Bad and no demand in severe weather conditions.

e. There is no much demand during the holidays

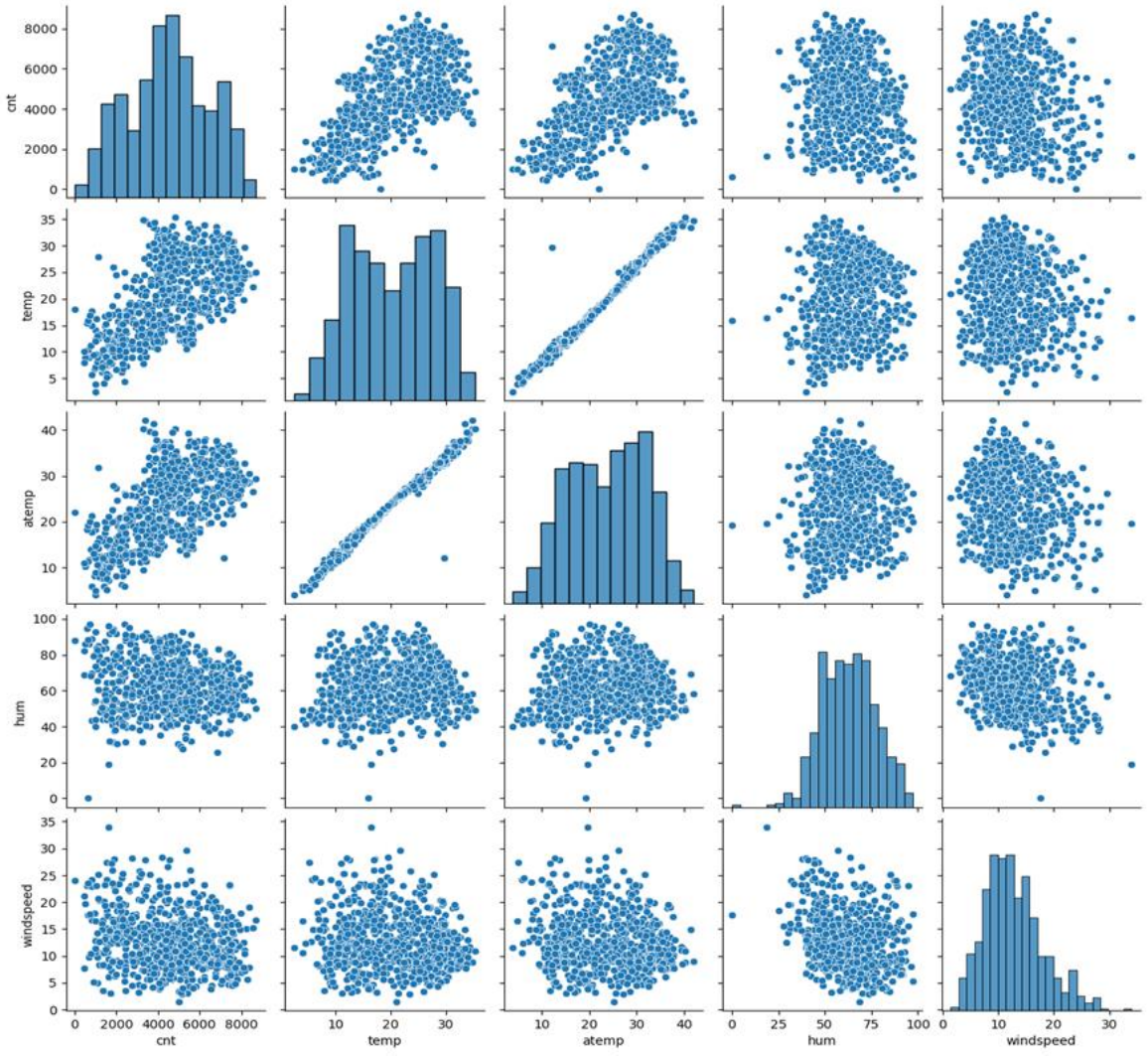
2. Why is it important to use drop_first=True during dummy variable creation?

Ans. When we create dummy variables from categorical data, we're essentially turning categories into numbers (0s and 1s). However, if we keep all the dummy variables, it can lead to a problem called multicollinearity, where some variables become dependent on others.

To avoid this, we use drop_first=True. This means we drop one of the dummy variables for each category. By doing this, we make sure that the information from the dropped variable can be inferred from the others, and we avoid redundancy. This also helps keep our model straightforward and ensures that our variables are truly independent from each other, which is important for accurate predictions.

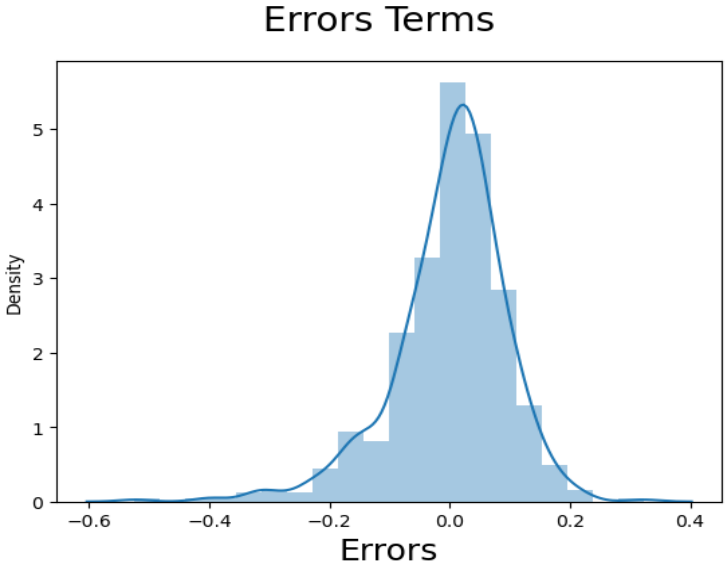
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans. Temp and atemp are the numerical variables which are showing the highest correlation with the target variable. The variable with the highest correlation coefficient is the one that has the strongest linear relationship with the target variable.



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

Ans.



In linear regression, one of our assumptions is that the "errors" or "residuals" (the differences between our predicted values and the actual values) should behave like a normal distribution. Think of it as a bell-shaped curve.

To check if this is happening, we plot a distribution of the residuals. If the curve looks like a bell and is centred around zero, it means our assumption is met. In other words, it shows that, on average, our predictions are correct (mean = 0), and the errors have a typical pattern like a normal distribution. This is important because it helps us trust the reliability of our model.

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

Ans. Based on the final model temp, weather, year are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

1. Temp – coefficient = 0.405878
2. Year – coefficient = 0.225011
3. Hum -coefficient = -0.342461

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans:

- Linear regression is a supervised machine learning algorithm used for predictive modelling and statistical analysis.
- It aims to model the relationship between a dependent variable (target) and one or more independent variables (features or predictors) by fitting a linear equation to observed data.
- The primary goal is to find the best-fitting line (or hyperplane in the case of multiple features) that minimizes the difference between the predicted and actual values of the target variable.

Types of Linear Regression:

1.Simple Linear Regression:

Simple linear regression deals with a single independent variable and a single dependent variable.

The relationship between the independent and dependent variables is represented by a straight line.

$y = mx + b$

where:

y is the dependent variable (target).

x is the independent variable (predictor).

m is the slope of the line, representing the relationship between x and y.

b is the intercept, indicating the point where the line intersects the y-axis.

2.Multiple Linear Regression:

Multiple linear regression extends the concept of linear regression to multiple independent variables.

It models the relationship between the dependent variable and multiple predictors using a linear equation in a multi-dimensional space.

The model equation for multiple linear regression is:

$y = b0 + b1*x1 + b2*x2 + ... + bn*xn$

where:

y is the dependent variable.

b0 is the intercept.

b1, b2, ..., bn are the coefficients for each independent variable x1, x2, ..., xn.

Key Concepts:

Coefficients (Weights): The coefficients (b values) in the linear equation represent the strength and direction of the relationship between each predictor and the target variable. These coefficients are learned during the model training process.

Residuals: Residuals are the differences between the predicted values and the actual values of the target variable. The goal is to minimize the sum of squared residuals, which is the basis for the "least squares" method used in linear regression.

Assumptions:

Linearity: The relationship between the predictors and the target is linear.

Independence: Observations are independent of each other.

Homoscedasticity: The variance of residuals is constant across all levels of predictors.

Normality: The residuals follow a normal distribution.

Model Training:

In simple linear regression, finding the best-fit line involves calculating the slope (m) and intercept (b) that minimize the sum of squared residuals (ordinary least squares).

In multiple linear regression, the model aims to find the coefficients (b values) that minimize the sum of squared residuals, taking into account multiple predictors

Model Evaluation:

Common metrics for evaluating linear regression models include:

Mean Squared Error (MSE): Measures the average squared difference between predicted and actual values.

R-squared (R^2) Score: Indicates the proportion of variance explained by the model.

2. Explain the Anscombe's quartet in detail

Ans:

Anscombe's quartet is a fascinating statistical paradox that consists of four distinct datasets, each of which has nearly identical simple descriptive statistics (mean, variance, correlation, regression line). However, when you visualize and explore these datasets, you'll discover that they have significantly different underlying patterns. Anscombe's quartet was created by the statistician Francis Anscombe in 1973 to emphasize the importance of data visualization in understanding data and detecting outliers or unusual observations.

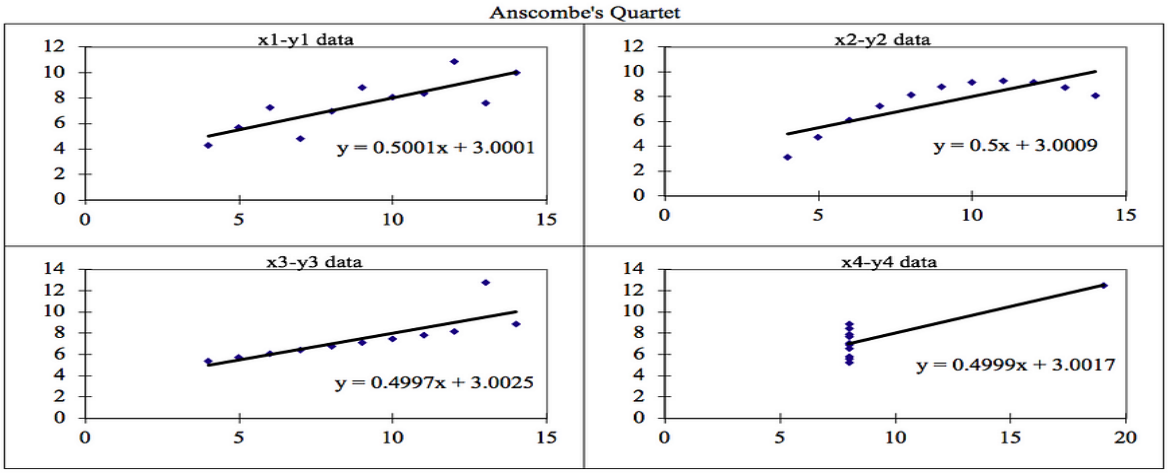
The four datasets can be described as:

1. The first scatter plot (top left) appears to be a simple linear relationship,

2. The second graph (top right); cannot fit the linear regression model because the data is non-linear

3. In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line. It shows the outliers involved in the dataset which cannot be handled by linear regression model

4. Finally, the fourth graph (bottom right) illustrates the impact of a high-leverage outlier on correlation, even though most data points don't exhibit a meaningful relationship.



3. What is Pearson's R?

Ans:

The Pearson correlation method is the most common method used for numerical variables. It assigns a value between − 1 and 1, where 0 is no correlation, 1 is total positive correlation, and −

1 is total negative correlation. Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product- moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations.

Pearson's R Formula is as follows:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Here,

1. r = correlation coefficient
2. xi = values of the x-variable in a sample
3. $\bar{x}$ = mean of the values of the x-variable
4. yi = values of the y-variable in a sample
5. $\bar{y}$ = mean of the values of the y-variable

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

Ans:

Scaling in the context of data preprocessing and machine learning refers to the process of transforming the numerical features of a dataset to a common scale or range. Scaling is performed to ensure that all features contribute equally to the analysis and modeling processes. It is particularly important when working with machine learning algorithms that are sensitive to the scale of the input data. Two common scaling techniques are normalized scaling and standardized scaling, which have distinct purposes and characteristics:

1. Normalized Scaling (Min-Max Scaling):

Purpose: The goal of normalized scaling is to rescale the features to a specific range, typically between 0 and 1. This ensures that all values fall within the same interval.

Formula: The formula for min-max scaling is as follows:

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

where:

- X_normalized is the scaled value.
- X is the original value.
- X_min is the minimum value of the feature in the dataset.
- X_max is the maximum value of the feature in the dataset.

Range: After scaling, all feature values are within the range [0, 1].

2. Standardized Scaling (Z-Score Scaling or Standardization):

Purpose: Standardized scaling transforms the features to have a mean of 0 and a standard deviation of 1. It centers the data around 0 and scales it based on its spread.

Formula: The formula for standardization is as follows:

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

where:

- X_standardized is the standardized value.
- X is the original value.
- X_mean is the mean of the feature in the dataset.
- X_std is the standard deviation of the feature in the dataset.

Mean and Standard Deviation: After standardization, the mean of the feature becomes 0, and the standard deviation becomes 1.

Key Differences:

Range:

- Normalized Scaling: The range of normalized data is typically between 0 and 1.
- Standardized Scaling: The range of standardized data has no specific bounds and can include positive and negative values.

Centering:

- Normalized Scaling: Data is not centered around a specific mean value.

- Standardized Scaling: Data is centered around a mean of 0.

Spread:

- Normalized Scaling: The spread (range) of data is adjusted based on the minimum and maximum values of the original feature.
- Standardized Scaling: The spread is adjusted based on the standard deviation of the original feature.

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

Ans:

- VIF (Variance Inflation Factor) measures multicollinearity in a regression model.
- VIF quantifies how much the variance of a regression coefficient is increased due to multicollinearity.
- Sometimes, VIF becomes infinite.
- This happens when a variable can be perfectly predicted by other variables (perfect multicollinearity).
- Perfect multicollinearity means there is an exact linear relationship between variables.
- VIF becomes infinite because the denominator in the VIF formula becomes zero (1 - 1 = 0) when R-squared (R^2) is exactly 1.
- Perfect multicollinearity can distort regression results and hinder interpretation.
- To address it, consider removing one correlated variable, creating composite variables, or using regularization techniques like Ridge or Lasso regression.

  The value of VIF is calculated by the below formula:

$$VIF_i = \frac{1}{1-R_i^2}$$

  Where, 'i' refers to the ith variable.

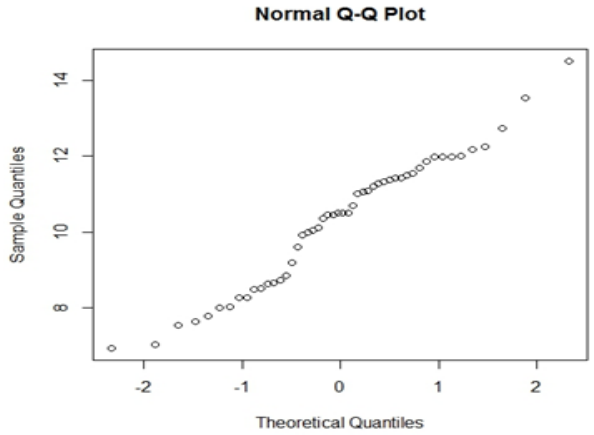
  If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables**.**

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

Ans:

The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.



Use of Q-Q plot in Linear Regression: The Q-Q plot is used to see if the points lie approximately on the line. If they don't, it means, our residuals aren't Gaussian (Normal) and thus, our errors are also not Gaussian.

Importance of Q-Q plot: Below are the points:

I. The sample sizes do not need to be equal.

II. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.

III. The q-q plot can provide more insight into the nature of the difference than analytical methods.