Assignment-based Subjective Questions

Q1 From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer

categorical variable inference

- Season: Most booking happens in season 3 followed by season 2 and 4.Season 4 winter is positively correlated ,season 1 spring is negatively correlated
- weathersit: Weathersit_2 (Mist + Cloudy) and weathersit_3(Lightsnow+light) rain is negatively correlated. weather sit does show some trend towards the bike bookings and is a good predictor for the dependent variable.
- mnth: month of September count is maximum and is positively correlated. Mean count is higher in the month of June till October compared to the remaining month
- Holiday: most of the bike booking when it is not holiday so holiday is a good predictor for dependent variable
- year – count has increased in the month of 2019 higher than previous year
- weekday: Shows very close trend , This is not a strong predictor

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

Answer

drop_first=True Instructs the dummy variable creation process to drop the first category, resulting in one fewer dummy variable compared to the total number of categories.

Q3Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer

Temp,atmp,cnt variables have the highest correlation

Q4How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer

Assumptions of Linear Regression after building model on training set is handled by

1 .Residual analysis: This analysis is such that the distribution of error is centered around zero

Res is calculated as : y_train-y_train_pred

2 There is No Multicollinearity between the predictor variables. This can be checked by the VIF value higher the VIF more is collinearity

3 Checking the p value which should be less than the 0.5

4. Linearity: Plot the residuals against the predicted values to check for any patterns or nonlinearity.

Q5 Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer

- Temperature (temp)
- Weathersit(Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)
- year

General Subjective Questions

Q1Explain the linear regression algorithm in detail.

Answer

Linear regression is a supervised machine learning algorithm. Output variable to be predicted is a **continuous variable**, example scores of a student. Linear regression algorithm helps to explain the relationship between a dependent and one or more independent variable using a straight line.

Independent variable is known predictor and dependent variable is known as output variables

Steps followed:

- Data preparation – cleaning and analyzing the data imputing missing values, removing duplicates.

- Model training – Dividing the data in training and test set

Training data is used for the model to learn during modelling

Testing data is used by the trained model for prediction and model evaluation

Linear regression assumes the linear relationship between input and output variable .(y = c + m1x1+m2x2+m3x3..+mixi)

If one independent variable it is called Simple linear regression

If more than one dependent variable it is called as Multiple Linear regression

- Model evaluation: Evaluation is done to find the best fit regression line and can be found by minimizing the cost function (RSS) using the methods:

1. Gradient
2. Gradient descent

- The strength of a linear regression model is mainly explained by R^2, where $R^2 = 1 - (RSS / TSS)$

1. RSS: Residual Sum of Squares
2. TSS: Total Sum of Squares

Q2 Explain the Anscombe's quartet in detail.

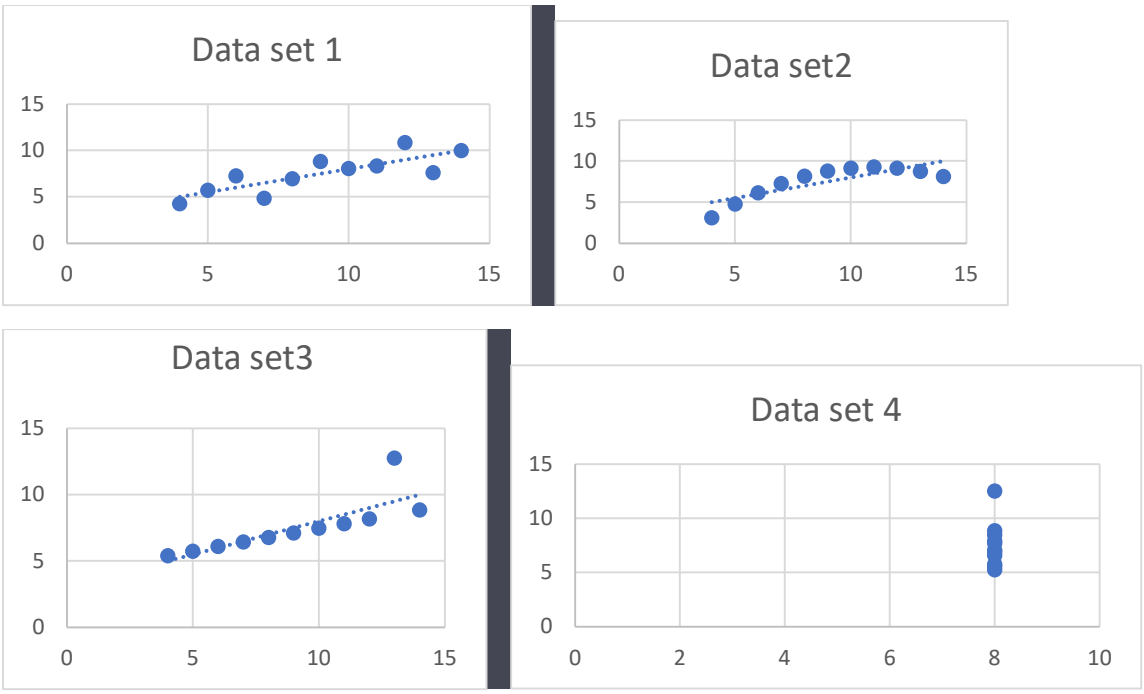
Answer

Anscombe's quartet illustrates the importance of visualization using plots rather than depending on the summary statistics.

It is the four data sets which have identical statistical data in terms of the mean, variance, correlation coefficient but when plotted on graph they show different characteristics

	dataset 1		dataset 2		dataset 3		dataset 4
	10	8.04	10	9.14	10	7.46	8
	8	6.95	8	8.14	8	6.77	8
	13	7.58	13	8.74	13	12.74	8
	9	8.81	9	8.77	9	7.11	8
	11	8.33	11	9.26	11	7.81	8
	14	9.96	14	8.1	14	8.84	8
	6	7.24	6	6.13	6	6.08	8
	4	4.26	4	3.1	4	5.39	8
	12	10.84	12	9.13	12	8.15	8
	7	4.82	7	7.26	7	6.42	8
	5	5.68	5	4.74	5	5.73	8
mean	9	7.5	9	7.5	9	7.5	9
SD	3.16	1.94	3.16	1.94	3.16	1.94	3.16
r	0.82		0.82		0.82		0.82

When plotted in graph we see that all four datasets behave differently



Q3 What is Pearson's R?

Answer

- Pearson's R helps to find the relationship between two continuous variables. It can be used in linear relationship.
- The value of R lies between -1 and 1

  -1 = a perfect linear relationship when one variable increases the other decreases

  1 = a perfect linear relationship when one variable increases the other also increases

  0 = There is no linear relationship between the two variables

- Pearson does not help in determining the cause-and-effect relationship

- Formula for calculating R

  $R = (\Sigma((X - \mu X)(Y - \mu Y))) / (n\sigma X\sigma Y)$

  where:
  Σ represents the summation symbol, which indicates that the formula is applied to each pair of corresponding X and Y values.
  X and Y are the two variables being correlated.
  μX and μY are the means (averages) of X and Y, respectively.
  σX and σY are the standard deviations of X and Y, respectively.
  n is the number of data points (observations).

Q4 What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer

Scaling is used to transform data which are in different units, range to a common scale so that a fair comparison can be done. example comparing humidity and temperature.

Scaling is performed for

- Comparison and Interpretation:
- Algorithm Performance

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

x_standardized = (x - mean(x)) / standard_deviation(x)

Normalization brings the values within the 0-1 range, preserving the relative order of the values. Standardization centers the data around 0 with a standard deviation of 1,

Q5 You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer

VIF is infinity when there isa perfect correlation between two variables. There is a multicollinearity in the data.

Formula of VIF is given by

VIF = 1/1-R^2 .

When the correlation between variables is 1 or -1, R^2 becomes equal to 1, and thus the denominator of the VIF formula becomes zero. Consequently, the VIF value becomes infinite.

Q6 What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer

Q-Q plot is also called as Quantile-Quantile plot. It helps to identify if a set of data came from some theoretical distribution such as a Normal, exponential or Uniform distribution

It is a tool which helps in assessing the normality assumption in Linear Regression. If the data points in the Q-Q plot fall approximately on or near a straight line, it indicates that the residuals follow a normal distribution. Deviation from a straight line suggests the departure from normality.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.