Assignment-based Subjective Questions

Question 1.

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

Answer:-

There Are Few Categorical Variables Season, yr, mnth, weekday, workingday, weathersit,. These categorical variables are Dependent on cnt ..

Question 2.

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

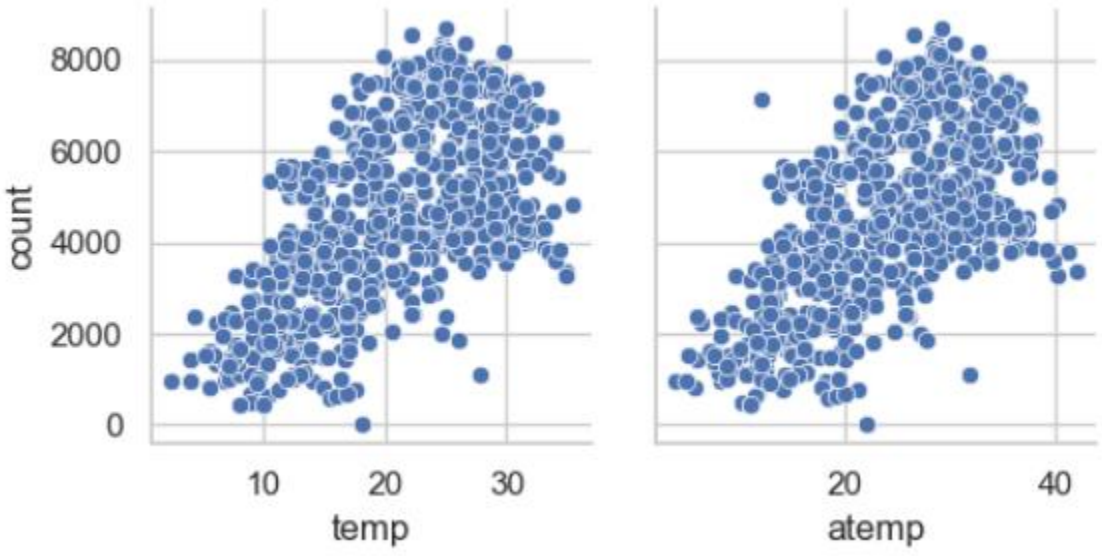
Answer:-

The Reason For Dummy Variables is For categorical variables with Having ‘n’ values,We Going To create ‘n-1’ columns. Each Indicating weather that level exist or not by using ‘0’,’1’.hence drop_first=True is used for resultant can match upto ‘n-1’.

Question 3.

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

Answer:-



The tempt and attempt variables have highest correlation than compare to cnt

Question 4.

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

Answer:-

Linear regression model can be validate based on **Linearity of Relationships** , Independence of Errors, Homoscedasticity (Constant Variance of Errors), Normality of Residuals, No Multicollinearity (for Multiple Linear Regression), Absence of Outliers or Influential Points

Question 5.

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

Answer:-

The demand of the shared bikes are temperature, year and season

General Subjective Questions

Question 6.
Explain the linear regression algorithm in detail?

Answer:-
Linear regression is a supervised learning algorithm used to model the relationship between a dependent variable y (target) and one or more independent variables X (predictors). Its objective is to find a linear equation that best describes this relationship.

1. **Objective of Linear Regression**

The goal of linear regression is to find the best-fit line that minimizes the error between predicted and actual values

The equation of a linear regression model is:

$y=\beta 0+\beta 1x+\epsilon$

Where:-

Y=dependent variable

β0=intercepet

β1=slope of line

ϵ=Error

For Multiple Linear Regression

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

Where:-

x1,x2,…,xn Independent variables.
β1,β2,…,βn: Coefficients corresponding to each independent variable

2. Assumptions of Linear Regression

For the model to work well and give reliable results, the following assumptions must hold:

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

Independence of Errors: Residuals (errors) are independent of each other.

Homoscedasticity: Residuals have constant variance across all levels of predictors.

Normality **of** Residuals: Errors are normally distributed.

No Multicollinearity: Predictors are not highly correlated with each other.

3. Mathematics Behind Linear Regression

Finding the Best-Fit Line

The best-fit line is determined by minimizing the **sum of squared errors (SSE)**. This is called the **Ordinary Least Squares (OLS)** method. The error (residual) for a data point is:

$ei=yi-y\hat{}i$

where:

yi: Actual value.

y^i : Predicted value ($y\hat{}i=\beta0+\beta1xi$ in simple linear regression).

The SSE is:

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Linear regression is a form of predictive modeling technique which tells us the relationship between the dependent (target variable) and independent variables (predictors). Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable. If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression. The linear regression model gives a sloped straight line describing the relationship within the variables. A regression line can be a Positive Linear Relationship or a Negative Linear Relationship. The goal of the linear regression algorithm is to get the best values for β0 and β1 to find the best fit line and the best fit line should have the least error. In Linear Regression, RFE or Mean Squared Error (MSE) or cost function is used, which helps to figure out the best possible values for a0 and a1, which provides the best fit line for the data points

Question 7. Explain the Anscombe's quartet in detail.?

Anscombe's quartet is a set of four datasets that have nearly identical statistical properties but exhibit very different distributions and relationships when plotted. It was introduced by statistician **Francis Anscombe** in 1973 to emphasize the importance of visualizing data before analyzing it.

Anscombe's quartet consists of **four datasets** (labeled I,II,III, IV) that have the following identical statistical properties:

- **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 equation:** y=3+0.5x

Despite these identical numerical summaries, the datasets have **very different relationships and structures**, which are only evident when plotted.

2. The Four Datasets

Each dataset has 11 data points and shows a different type of behavior:

Dataset I (Linear Relationship)

- This dataset closely follows a linear trend.
- The linear regression model (y=3+0.5x) is a good fit.
- When plotted, the data points lie approximately along the regression line with minor random deviations.

Dataset II (Nonlinear Relationship*)*

- This dataset forms a parabolic (curved) shape when plotted.
- The linear regression model (y=3+0.5xy = 3 + 0.5xy=3+0.5x) is inappropriate here because the relationship is not linear.

Dataset III (Outlier-Driven Relationship)

- Most points in this dataset form a vertical line, except for one influential outlier.
- The linear regression line is skewed due to this single outlier.
- Without the outlier, the relationship would not resemble the regression line.

Dataset IV (Vertical Line with an Outlier)

- Most of the data points lie on a single vertical line.
- A single extreme outlier drives the apparent correlation and regression line.
- This dataset highlights how a spurious relationship can arise due to a single influential point.

3. Key Takeaways

Anscombe's quartet highlights several critical lessons in data analysis:

a. The Importance of Visualization

- Summary statistics (like mean, variance, and correlation) do not always tell the full story.
- Visualizing data (e.g., with scatter plots) is essential to understand the underlying relationships and patterns.

b. Context Matters

- Identical statistical properties can correspond to very different data distributions and relationships.
- Proper interpretation requires understanding the context and structure of the data.

c. Outliers and Influential Points

- Outliers and influential data points can significantly distort results.
- Always check for these anomalies and consider their impact on your analysis.

d. Linear Models Have Limits

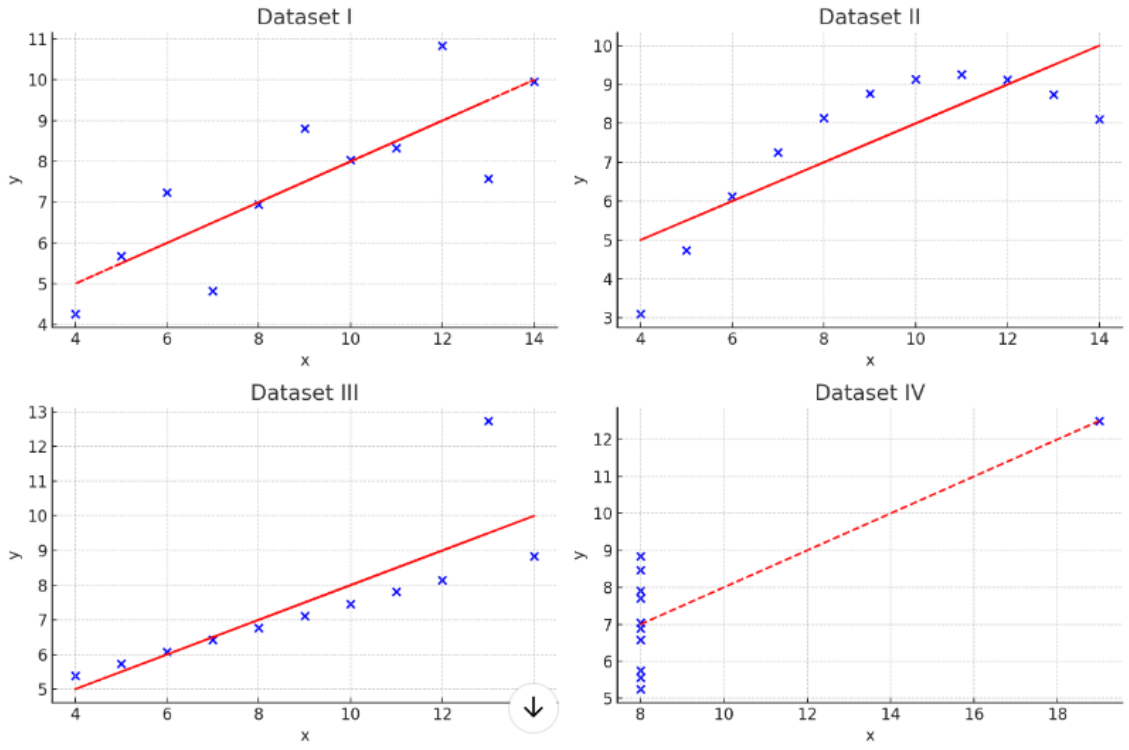
- Linear regression assumes a linear relationship, but this is not always appropriate.
- Data exploration helps determine whether a linear model is suitable or if transformations or non-linear models are needed.

---

4. Visual Representation

The stark differences between the datasets become clear when visualized. Each scatter plot of x vs. y tells a unique story:

1. **Dataset I:** Straightforward linear relationship.
2. **Dataset II:** Clear curve, not linear.
3. **Dataset III:** Outlier distorts the relationship.
4. **Dataset IV:** A single outlier dominates the data.



8 Question . What is Pearson's R?

Answer:-

Pearson's R, also called the **Pearson correlation coefficient**, is a statistical measure used to quantify the strength and direction of the linear relationship between two continuous variables. It is a widely used metric in statistics and machine learning for understanding the association between variables.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Question 9.

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

Answer:-

Scaling is the process of transforming the values of features in a dataset so that they are on a similar scale or within a specific range. It ensures that all features contribute equally to the model's learning process, especially in algorithms that are sensitive to the magnitude of the input features.

1.In normalized scaling minimum and maximum value of features being used whereas in Standardize scaling mean and standard deviation is used for scaling.

2. Normalized scaling is used when features are of different scales whereas standardized scaling is used to ensure zero mean and unit standard deviation.

3. Normalized scaling scales values between (0,1) or (-1,1) whereas standardized scaling is not having or is not bounded in a certain range.

4. Normalized scaling is affected by outliers whereas standardized scaling is not having any effect by outliers.

5. Normalized scaling is used when we don't know about the distribution whereas standardized scaling is used when distribution is normal.

6. Normalized scaling is called as scaling normalization whereas standardized scaling is called as Z Score Normalization.

Question 10.

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

Answer:-

The **Variance Inflation Factor (VIF)** is a metric used to measure the degree of multicollinearity in a regression model. Specifically, it quantifies how much the variance of a regression coefficient is inflated due to multicollinearity with other features.

A VIF value of greater than 10 is definitely high, a VIF of greater than 5 should also not be ignored and inspected appropriately. A very high VIF value shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

Question 11. What is a Q-Q plot?

A **Q-Q plot** (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a dataset to a theoretical distribution (e.g., a normal distribution). It helps determine whether the data follows a specific distribution by plotting the quantiles of the dataset against the quantiles of the theoretical distribution

Importance of QQ Plot in Linear Regression : In Linear Regression when we have a train and test dataset then we can create Q-Q plot by which we can confirm that both the data train and test data set are from the population with the same distribution or not.

Advantages:

- ✓ It can be used with sample size also
- ✓ Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot Q-Q plot use on two datasets to check
- ✓ If both datasets came from population with common distribution
- ✓ If both datasets have common location and common scale
- ✓ If both datasets have similar type of distribution shape
- ✓ If both datasets have tail behavior