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?

Answer:
Inference drawn from categorical variables with their effect on dependent variable cnt:

1. Season fall has highest influence on rental demands.
2. 2019 had more demands as compared to previous , might follow similar trend in coming years
3. Demand is increasing each month till june.In month of Sep, it shows hihest demands then it decreases.
4. If there is holiday demand decreases.
5. Weekdays shows similar pattern, doesnot imply much about demand.
6. Weathersit - clear wethersit has higher demands for rental bikes.

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

Answer:
To avoid dummy variable trap.Dummy variable trap will lead to multi collinearity issue among Categorical variables.Also n-1 categories can explain n categories.If all the values are 0 it represents first column.

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

Answer:
temp and atemp variable has highest correlation with cnt

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

Answer:

1. Perform residual analysis on training set, where we can confirm errors are normally distributed with mean 0.
2. Plotting scatter plot that confirms linear relationship between predictor variable and actual target variable.
3. Plot of residuals (actual-predicted)which shows Error terms are independent of each other.
4. Error terms have constant variance(Homoscedacity) can be derieved from scatter plot of predicted and actual target variable.

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: 3 features contributing significantly: Temp,yr,mnth_sep

General Subjective Questions

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

Answer:

Linear Regression finds the best linear relationship between the independent and dependent variables. It is a method of finding the best straight-line fitting to the given data. In technical terms, linear regression is a machine learning algorithm that finds the best linear-fit relationship on any given data, between independent and dependent variables. It is mostly done by the Sum of Squared Residuals Method.

Assumptions about the residuals:

1. Normality assumption: It is assumed that the error terms, ε(i), are normally distributed.
2. 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.
3. 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.
4. 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 independent variables:

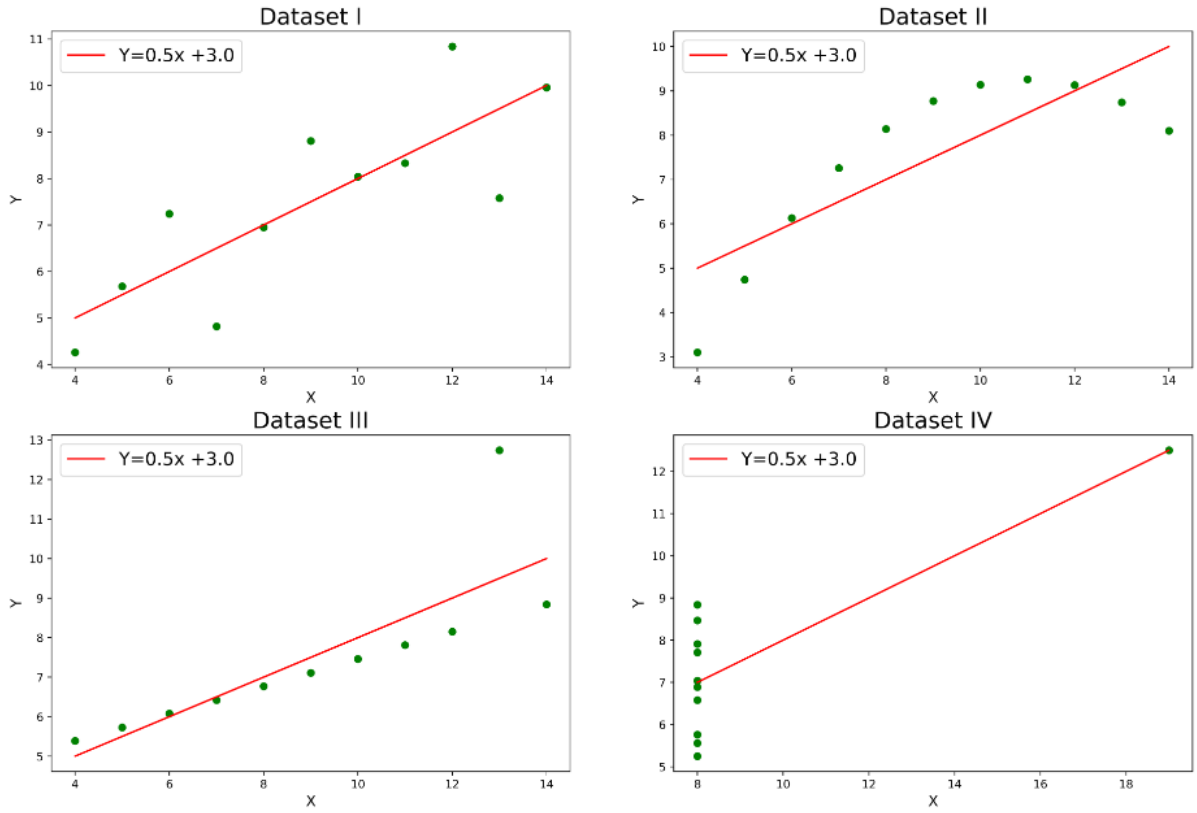
1. The independent variables are linearly independent of each other, i.e., there is no multicollinearity in the data.
2. Independent variables are measured without error.

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

Answer:

Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics.



Explanation of this output:

1. In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
2. In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
3. In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
4. Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

3. **What is Pearson's R?** (3 marks)

Answer:

Pearson's R or correlation coefficient is a measure of linear correlation between two sets of data.It is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

- -1 indicates a strong negative relationship.If one parameter increases other decreases.

- 1 indicates strong positive relationships.If once increases , other parameter also increases.
- And a result of zero indicates no relationship at all.

5. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer:

Feature scaling is a data preprocessing technique used to transform the values of features or variables in a dataset to a similar scale. The purpose is to ensure that all features contribute equally to the model and to avoid the domination of features with larger values.. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionate. Another reason why feature scaling is applied is that gradient descent converges much faster with feature scaling than without it.

Normalization: Also known as min-max scaling or min-max normalization, it is the simplest method and consists of rescaling the range of features to scale the range in [0, 1].Formula for normalization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Here, Xmax and Xmin are the maximum and the minimum values of feature,respectively.

Standardization: Feature standardization makes the values of each feature in the data have zero mean and unit variance. The general method of calculation is to determine the distribution mean and standard deviation for each feature and calculate the new data point By below formula

$$X' = \frac{X - \mu}{\sigma}$$

μ is the mean of the feature values and σ is the standard deviation of the feature

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. This 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. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables

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

(3 marks)

Answer:

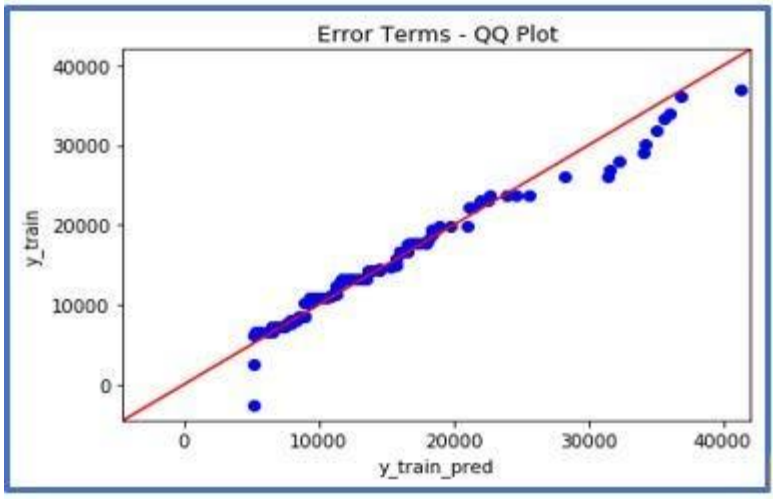
Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45-degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

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.

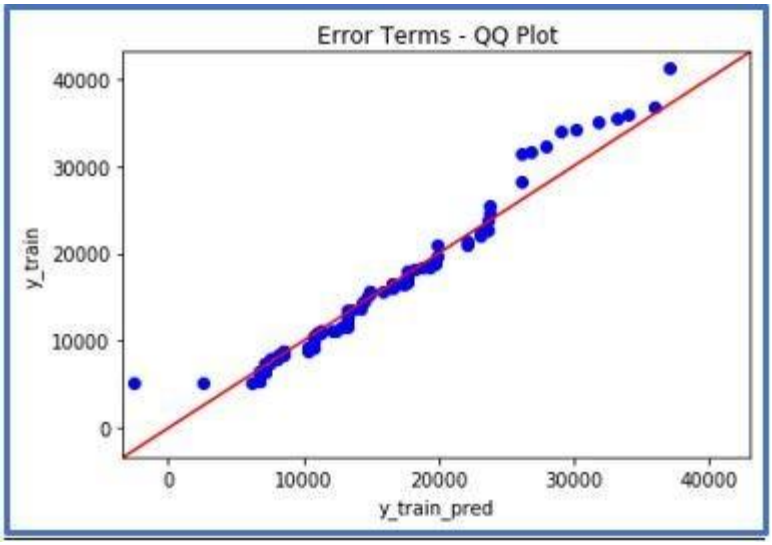
Below are the possible interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

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



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



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