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:

- Categorical variables need to be recorded into a series of variables which can be entered into linear regression model.

- Regardless of the coding system you choose, the overall effect of the categorical variable will remain the same.

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

Answer: drop_first=True helps to reduce the extra column created during dummy variable creation. It reduces the co-relation created among dummy variables

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

Answer: registered

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

Answer:

a. *Linearity*: The relationship between X and the mean of Y is linear.

Pair-wise scatterplots are helpful in validating the linearity assumption as it is easy to visualize a linear relationship on a plot.

b. *Homoscedasticity*: The variance of residual is the same for any value of X.

From residual plot, it can be verified that they have constant variance which ensures Homoscedasticity.

c. *Independence*: Observations are independent of each other.

We can see from the vif that all the parameter having vif less than 3. It shows the parameters are independent of each other.

d. *Normality*: For any fixed value of X, Y is normally distributed.

To verify the normality of error, an easy way is to draw the distribution of residuals against levels of the dependent variable. The resulting curve is normal which indicates distribution is normal.

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: casual, season and weathersit

General Subjective Questions

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

Answer:

- Linear regression is a **linear model**, example : a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).

- When there is a single input variable (x), the method is referred to as **simple linear regression**. When there are **multiple input variables**, literature from statistics often refers to the method as multiple linear regression.

$Y = a0 + a1X + e$

Y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
a0= intercept of the line (Gives an additional degree of freedom)
a1 = Linear regression coefficient (scale factor to each input value).
ε = random error

Assumptions of linear regression :

a. Linearity: he relationship between X and the mean of Y is linear.

b. Homoscedasticity: The variance of residual is the same for any value of X.

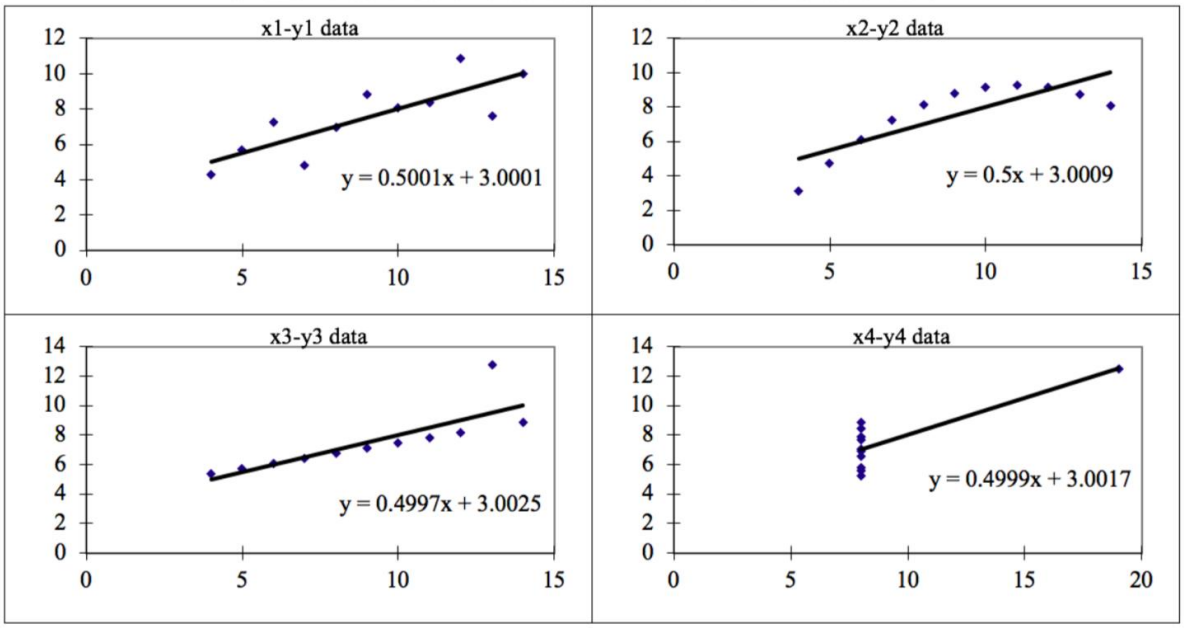
c. Independence: Observations are independent of each other.

d. Normality: For any fixed value of X, Y is normally distributed.

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

Answer:

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.



1. **Dataset 1:** this **fits** the linear regression model pretty well.

2. **Dataset 2:** this **could not fit** linear regression model on the data quite well as the data is non-linear.

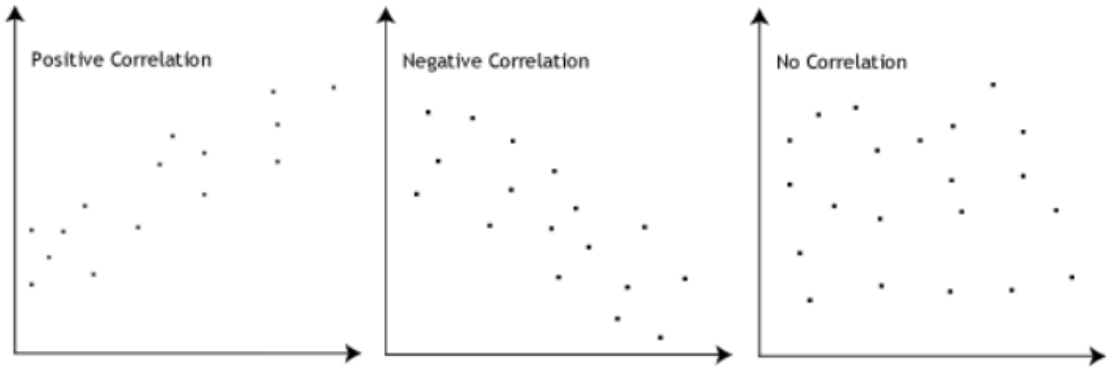
3. **Dataset 3:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

4. **Dataset 4:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

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

Answer:

The Pearson product-moment correlation coefficient (or Pearson correlation coefficient, for short) is a measure of the strength of a linear association between two variables and is denoted by *r*. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, *r*, indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).



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

Answer:

Scaling: It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Standardisation:

1. Mean and standard deviation is used for scaling.
2. It is used when we want to ensure zero mean and unit standard deviation.
3. It is not bounded to a certain range.
4. It is much less affected by outliers.
5. Scikit-Learn provides a transformer called StandardScaler for standardization.

Normalization:

1. Minimum and maximum value of features are used for scaling
2. It is used when features are of different scales.
3. Scales values between [0, 1] or [-1, 1].
4. It is really affected by outliers.
5. Scikit-Learn provides a transformer called MinMaxScaler for Normalization.

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

Answer: 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:

- Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.
- Also, it helps to determine if two data sets come from populations with a common distribution.
- 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.