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?
Ans : Categorical variable year 2019 and weathersit-light rain, snow has highest effect on target variable

2. Why is it important to use drop_first=True during dummy variable creation?
Ans : We have to use drop_first=True to avoid creating unnecessarily extra dummy variables. E.g. if there are 3 categories then we dont need to create 3 dummy variables, only 2 dummy variables are sufficient to capture the information.

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

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
Ans : The assumptions of linear regression are validated by checking

1. If error terms are normally distributed in distribution plot
2. Error terms have constant variance
3. Error terms are independent of each other
4. Linear relationship can be checked by using R-square

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

A. Temp
B. Year 2019
C. Weathersit : Light rain and light snow

General Subjective Questions

1. Explain the linear regression algorithm in detail
Ans: Steps to be followed in Linear Regression Algorithm:

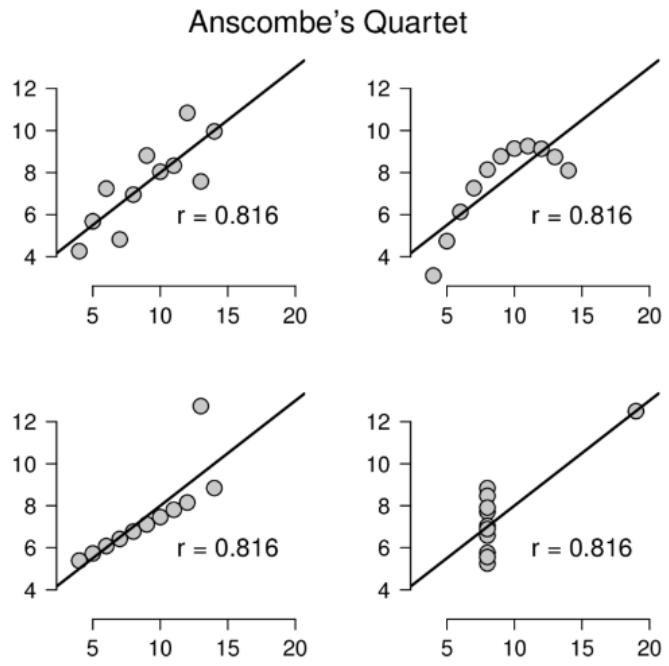
a) Reading and understanding the data : includes importing libraries and cleaning and manipulating data to make it up to the standards that exploratory data analysis can be performed by treating null values, changing data types and removing unwanted rows and columns
b) Visualizing the data : Visualizing numerical variables using scatter or pairplots in order to interpret data.
c) Data preparation: includes converting categorical variables with varying degrees of levels into dummy variables (numerical in nature) so that these variables can be

represented during model building in order to contribute to the best fitted line for the purpose of better prediction.

d) Splitting the data into training and test sets : train-test split ratio is 70:30 or 80:20.
e) Rescaling: It is a method used to normalize the range of numerical variables with varying degrees of magnitude.
f) Building a linear model
    i) Forward Selection: First we select the one, which has highest correlation and then we move on to the second highest and so on.
    ii) Backward Selection: We add all the variables at once and then eliminate variables based on high multicollinearity (VIF>5) or insignificance (high p-values).
    iii) RFE or Recursive Feature Elimination : automated version where we select that we need “m” variables out of “n” variables and machine provides rankings. A rank 1 means that feature is important for the model, while a rank 4 implies that we are better off, if we don’t consider the feature.
g) Residual analysis of the train data: errors (y_actual — y_pred) are distributed across the model. A good residual analysis will signify that the mean is centred around 0.
h) Prediction: predict the test dataset by transforming it onto the trained dataset and calculate r2_score of the test set.

2. Explain the Anscombe’s quartet in detail.

Anscombe's quartet highlights the importance of plotting data to confirm the validity of the model fit. In each chart, the Pearson correlation between the x and y values is the same, r = .816. In fact, the four different data sets are also equal in terms of the mean and variance of the x and y values. But the graphical displays reveal that the patterns are very different from one another, and that the Pearson correlation (a linear measure of association) is only valid for the data set from the top left panel.



3. What is Pearson's R?

Ans: Pearson R is used to measure linear correlation. It varies from -1 to 1. It determines the strength and direction of correlation.

a) If Pearson R is between 0 and 1 then indicates positive correlation
b) If Pearson R is between 0 and -1 then indicates negative correlation
c) Pearson R equal to zero, then no correlation
d) Formaula:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

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

Ans : If one variable has values in the range of 10 to 20 and another has a range of 1000 to 10000, then for interpretability we use scaling to make both in the same range. There are two methods of scaling : normalized and standardized

a) Normalized scaling: to set values between 0 and 1
b) Standardized scaling : to bring all of the data into a standard normal distribution with mean zero and standard deviation one

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Ans : Infinite VIF 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.
Ans: Q-Q plots are Quantile-Quantile plots. They plot the quantiles of a sample distribution against quantiles of a theoretical distribution. It helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.

a) For normal distribution: it gives roughly straight line
b) For uniform distribution : both the right and left tails are small and the extreme values in the above plot are falling close to the center
c) Skewed data : left side of the plot deviating from the line, it is left-skewed. When the right side of the plot deviates, it's right-skewed