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:

- Bike usage was higher in 2019.
- There was an increase in bike usage from May through September, followed by a decrease.
- Bikes usage is higher in the summer and fall seasons.
- Bike usage is particularly high when the weather is clear.

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

Ans: By dropping one of the dummy variables effectively removes the redundancy prevents collinearity. The dropped dummy variable acts as a reference category. The coefficients of the remaining dummy variables represent the change in the outcome variable with respect to the reference category.

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

Ans: "temp" has the 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:

- Normal Distribution of Error terms – Error terms are normally distributed with mean zero.
- Error terms are independent of each other – The error terms should not be dependent on one another.
- Linear relationship – X and Y should display some linear relationship.

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

Ans:

- Temp
- yr_2019
- weathersit_snow

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Linear Regression is a method of finding the best straight line fitting to the given data , that is finding the best linear relationship between the independent and dependent variables. 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 done by Sum of Squares Residuals Method. Linear regression is a statistical method that is used for predictive analysis. It makes predictions for continuous /real or numeric variables such as sales, salary, age etc.

Types of Linear Regression

Simple Linear Regression: A single independent variable is used to predict the value of a numerical dependent variable.

Multiple Linear Regression: When more than one independent variable is used to predict the value of a numerical dependent variable.

Assumptions of Linear Regression:

Linearity Assumption: It is assumed that there is a linear relationship between the dependent and independent variables.
Normality Assumption: It is assumed that the error terms are normally distributed.
Zero mean assumption: It is assumed that the error terms are normally distributed around zero.
Constand Variance assumption: It is assumed that the residual terms have the same variance.
Independent error assumption: It is assumed that the residual terms are independent of each other.

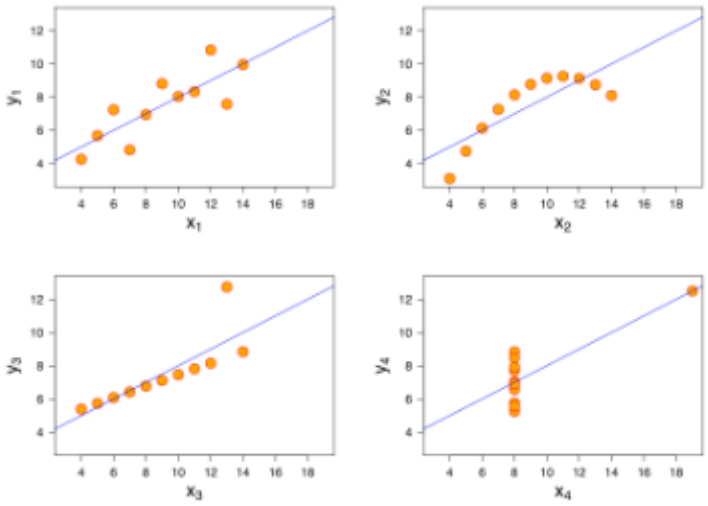
To check of linear regression is suitable for any given data, a scatter plot is used. If the relationship is linear, we can go for linear model. But if it is not linear, we must apply some transformations to make the relationship linear.

In case of univariate linear regression plotting a scatter, plot is easy. In multi variate analysis, two dimensional pairwise scatter plots can be plotted.

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet is a modal example to demonstrate the importance of data visualization and plotting the data before analysing it with statistical properties.

When we plot the sample datasets having similar statistical analysis, across x & y coordinate plane, we get the following results,



As we can see all the four linear regression are exactly the same. But there are some peculiarities in the datasets which have fooled the regression line.

Dataset1: This plot shows that there is a linear relationship between x and y.
Dataset2: This plot shows that there is no linear relationship.
Dataset3: This plot looks like a linear relationship except for one large outlier.
Dataset4: This plot looks like of x remains constant, except for one outlier as well.

3. What is Pearson's R?

Ans: Pearson's R or Pearson's correlation coefficient, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to 1, where

- 1 indicates the perfect positive linear relationship: as one variable increases, the other variable also increases.

- -1 indicates a perfect negative linear relationship: as one variable increases, the other variable decreases.
- 0 indicates no relationship between variables.

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

Ans: Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then an algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of values.

Normalization	Standardization
This method scales the model using minimum and maximum values.	This method scales the model using the mean and standard deviation.
Values on the scale fall between [0, 1] and [-1, 1].	Values on a scale are not constrained to a particular range.
Additionally known as scaling normalization.	This process is called Z-score normalization.
When the feature distribution is unclear, it is helpful.	When the feature distribution is consistent, it is helpful.

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

Ans: If there is a perfect correlation then VIF value is infinity. A large value of VIF indicates that there is a correlation between the variables. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

VIF value as infinity shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2=1, which leads to VIF as infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Ans: The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

Use of Q-Q plot:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.