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)

Ans -

Categorical variables Holiday, Spring, Light_Snow, Mist_Cloudy, Sunday and some selective months are improving the OLS and VIF reports.

- Demand is high in yr, holiday, spring, Light_Snow, Mist_Cloudy,3,5,6,7,8,9,10,Sunday.
- High demand in months ,3,5,6,7,8,9,10.
- High demand in weather condition - Light_Snow and Mist_Cloudy
- High demand season - spring

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

Ans –

a. It is important to use Drop_first=True, as it helps in reducing the extra column created during dummy variable creation.

b. Hence it reduces the correlations 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)

Ans –

a. The pair-plot among the numerical variables, yr and temp shows 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? (3 marks)

Ans –

a. I built the model on the training set by using RFE - Recursive feature elimination and checked results of OLS and VIF.

b. And then manually I added and removed different variables and validated whether our model improves or not. Checked Error term. Got final model.

c. then tried final model on test set . Ploted y_test Vs y_pred, the plotted points are overlapping and close to each other. Hence we can say that our model is can predict very well.

d. And then tried final model on test set and checked r2_score which was closed to train model score.

e. Calculated r_squared is close to our model r_squared value.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans - Based on the final model, Yr, holiday, spring, Mist_Cloudy are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

General Subjective Questions

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

Ans –

a. **Linear Regression** is a machine learning algorithm based on **supervised learning**.

b. It performs a **regression task**. Regression models a target prediction value based on independent variables.

c. It is mostly used for finding out the relationship between variables and forecasting.

d. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.

e. There are many names for a regression's dependent variable.

f. It may be called an outcome variable, criterion variable, endogenous variable, or regressand.

g. The independent variables can be called exogenous variables, predictor variables, or regressors.

h. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x)).

i. Hence, the name is Linear Regression. In the figure above, X (input) is the work experience and Y (output) is the salary of a person.

j. The regression line is the best fit line for our model. **Hypothesis function for Linear Regression :**

$y= a0+a1x+ \varepsilon$

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

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

Ans-

a. Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models.

b. Anscombe's quartet highlights the importance of plotting data to confirm the validity of the model fit.

c. In fact, we can check the different data sets are equal or not in terms of the mean and variance of the x and y values.

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

a. Pearson's r is a numerical summary of the strength of the linear association between the variables.
b. If the variables tend to go up and down together, the correlation coefficient will be positive.
c. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

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

a. What - it is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range.
b. It also helps in speeding up the calculations in an algorithm.
c. Why - Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

d. It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

e. Normalization/Min-Max Scaling:

   It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

f. Standardization Scaling:

   Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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

Ans –

a. If there is a perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables.
b. 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.
c. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans –

a. Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other.
b. 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.
c. 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.
d. If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.