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 like ‘season’, ‘weathersit’ are not ordinal i.e. there is no particular natural order to leverage. They are nominal categorical variables, so there is no need to use special encoding methods like – dummy or one-hot encoding. Current independent values provided are good enough to use (1,2,3,4 etc.). These variables certainly impact target dependent variable which is observed through the corelation matrix. Similarly other binary variables for identifying working and public holiday have impact on target variable. Datetime variable ‘dteday’ is not categorical, however that needs to be converted as day of the year.

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

Ans:

This is the most important parameter. This takes a boolean value, True or False. If False (default), this will perform one-hot encoding. If True, this will drop the first category of each categorical variable, create k-1 dummy variables for each categorical variable and perform dummy encoding. Since we didn’t need dummy encoding , drop_first=True should be used.

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

with the target variable? (1 mark)

Ans: - ‘atemp’ i.e. actual temperature variable has highest corelation with target variable

4. How did you validate the assumptions of Linear Regression after building the model on the

training set? (3 marks)

Ans:

After choosing right features based on Variance Inflation factor and corelation analysis, model was built using both methods – sklearn and statsmodel. Sklearn predictions were reviewed for MSE ie.. mean square error to confirm model fitness. Using stats model , coefficient values and intercept were reviewed to be same as earlier fit. Adjusted R-square was not penalised at all, F-statistic value was very low , similarly P-values were reviewed and they were not large, which indicated best choice of features. R-square is 1 which means, all features are explaining the model well.

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:

‘Windspeed’, ‘day of the year’ i.e. date and ‘actual temperature’ features are contributing significantly towards explaining the demand.

General Subjective Questions

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

   Ans:

Regression is a method of modelling a target value based on independent predictors. Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable
Y = a_0 + a_1 * X

Similarly, if there are multiple predictors, algorithm can be extended to find best intercept and coefficient values.

Y = a_0 + a_1 * X1 + a_2 * X_2 + a_3 * X_3 + ….

The motive of the linear regression algorithm is to find the best values for a_0 , a_1 etc. We convert this search problem for a_0 , a_1 into a minimization problem where we would like to minimize the error between the predicted value and the actual value which is further measured by R2 value. R2 value determines the proportion of variance in the dependent variables that can be explained by independent variables i.e. predictors.

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

Ans: Anscombe's quartet is a group of four data sets that are nearly identical in simple descriptive statistics. Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

Data Set 1: fits the linear regression model pretty well.
Data Set 2: cannot fit the linear regression model because the data is non-linear.
Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model.
Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

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

Ans: Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

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

Ans: Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data pre-processing step.

In normalised scaling , features is scaled between 0 and 1 using min-max scaling . In standardization, values of each feature in the data have zero mean and unit variance.

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

Ans:
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.

5. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression
Ans: Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.