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 demand in the fall is the highest.

•Bike demand takes a dip in spring.

•Bike demand in year 2019 is higher as compared to 2018.

•Bike demand is high in the months from May to October.

•Bike demand is high if weather is clear or with mist cloudy while it is low when there is light rain or light snow.

•The demand of bike is almost similar throughout the weekdays.

•Bike demand doesn’t change whether day is working day or not.

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

Ans .

•It is important in order to achieve k-1 dummy variables as it can be used to delete extra column while creating dummy variables.

•For Example: We have three variables: Furnished, Semi-furnished and un-furnished. We can only take 2 variables as furnished will be 1-0, semi-furnished will be 0-1, so we don’t need unfurnished as we know 0-0 will indicate un-furnished. So we can remove it

•It is also used to reduce the collinearity between dummy variables .

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

Ans .

• atemp and temp both have same correlation with target variable of 0.63 which is the highest among all numerical variables.

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

Ans .

•Linearity of relationship between response and predictor variables.

•Normality of the error distribution (Normal distribution of error terms).
•Constant variance of the errors or Homoscedasticity.

•Less Multi-collinearity between features (Low VIF)

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

Ans .

•temp

•light_rain_snow

•Sep

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: It is a form of regression, where the target variable is continuous. It estimates the relationship between a target variable and one or more predictor variables.

The Equation of linear Regression is

$y= m1x1 + m2x2 + m3x3 + \ldots\ldots\ldots + m(n)\, x(n) + c$.

Where y is target variable and x1, x2 , x3 …… xn are predictor variables . And we have two unknowns, m, and c, and we need to choose those values of m and c, which provides us with the minimum error. We need to get the best fit line which is the line that has the minimum error. In linear regression, when the error is calculated using the sum of squared error, this type of regression is known as OLS, i.e., Ordinary Least Squared Error Regression.

Error function is explained by ‘e = - y’, and error depends on the values of ‘m’ and ‘c’. Our aim is to build an algorithm which can minimize the error.

And in order to do so we use cost function of Linear Regression, Which is:

$J (mi, c)= (1/2n)\Sigma(yi - yp)^2$

Where yi and yp are expected values and predicted values.

Our main aim is to minimize J by changing m and c and it can be done using Gradient Descent Algorithm.

Cost function measures the performance of a Machine Learning model for given data.

2. Explain the Anscombe’s quartet in detail.

Ans: Anscombe's quartet consists of four data sets that have nearly identical simple descriptive statistics but have very different distributions and appear very different when presented graphically. Each dataset consists of eleven points. The primary purpose of Anscombe’s quartet is to illustrate the importance of looking at a set of data graphically before beginning the analysis process as the statistics merely does not give the an accurate representation of two datasets being compared.

3. What is Pearson’s R?

Ans: Pearson's Correlation Coefficient is used to establish a linear relationship between two quantities. It gives an indication of the measure of strength between two variables and the value of the coefficient can be between -1 and +1.

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

Ans: Scaling is a technique performed in pre-processing during building a machine learning model to standardize the independent feature variables in the dataset in a fixed range. The dataset could have several features which are highly ranging between high magnitudes and units. If there is no scaling performed on this data, it leads to incorrect modelling as there will be some mismatch in the units of all the features involved in the model. The difference between normalization and standardization is that while normalization brings all the data points in a range between 0 and 1, standardization replaces the values with their Z scores.

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

Ans: The value of VIF is infinite when there is a perfect correlation between the two independent variables. The R�squared value is 1 in this case. This leads to VIF infinity as VIF equals to 1/(1-R2). This concept suggests that is there is a problem of multi-collinearity and one of these variables need to be dropped in order to define a working model for regression.

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 are used to plot quantiles of a sample distribution with a theoretical distribution to determine if any dataset concerned follows any distribution such as normal, uniform or exponential distribution. It helps us determine if two datasets follow the same kind of distribution. It also helps to find out if the errors in dataset are normal in nature or not.