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: In our final Multi Linear Regression model, we have finalised below categorical variables and their respective coefficients are shown :

Working day : 0.0561

Season :

- Summer : 0.0886
- Winter : 0.1307

Weekday :

- Sat : 0.0675

Weathersit :

- Light_Rain : - 0.2871
- Misty : - 0.0800*

Interpretation : We can use coefficients mentioned above to infer about the dependant variables.

A unit increase in x results in an increase/decrease in average y by coefficents

As per the magnitude of coefficients, y can be increased or decreased. For eg: Light rain and Misty weather have negative coefficients ,so it may decrease the value of dependant variable.

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

Ans: In general drop_first is used to drop extra column which is created during dummy variable creation. For example, we have Region field which is having values North, East, West and South. When we create dummy variable for Region then it will actually creates 4 columns for each region values. But even if we create 3 dummy columns for North, East and West then it is obvious missing is South.

Actually, when we create dummy variable and we don't remove 1 column then it creates multi - collinearity in the data and it create the problem of dummy variable trap that can mess up our machine learning model.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans: In raw data, temp columns has the highest correlation.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans: To validate the assumptions of Linear Regression we can do the following :

a) The dependent variable and independent variable must have a linear relationship.
   - We can use pairplot to visualize the dataframe and see if the linear relationship exists between dependant and independent variables.

b) Multicollinearity should not exist.
   - We can use heatmap to visualize the correlation among independent variables. In case number of independent variables are large then we can use Variance Inflation Factor (VIF).
     i. If VIF =1, less multicollinearity
     ii. If VIF<5, Moderate multicollinearity
     iii. If VIF>5, High multicollinearity and we need to drop those columns.

c) Residuals must be normally distributed.
   - We can draw the residuals using Distribution plot and see residuals behaviour

d) No heteroscedasticity.
   - We can draw a scatter plot against Residuals vs Fitted values, if funnel shape pattern exists which means heteroscedasticity exists.

e) No autocorrelation in residuals.
   - We can use Durwin-Watson test.
   - DW =2 is the ideal case which says no autocorrelation exists
   - 0<DW<2 -> Positive autocorrelation
   - 2<DW<4 -> Negative autocorrelation.

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: Temp, Windspeed and weather

General Subjective Questions

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

Ans: Linear regression is one of the basic form of machine learning which comes under supervised machine learning. We train a model to predict the behaviour of your data based on some independent variables.As name suggests Linear which means the two variables which are on the x-axis and y-axis should be linearly correlated.

Linear regression is used to predict a quantitative response Y from the independent variable X.

We can write linear regression equation as:

$Y = mx + c$

Here, m = slope of the line

C = y-intercept of the line

X = Independent variable from dataset

Y = Dependant variable from dataset

Below are the use cases of Linear Regression :

1) Customer churn
2) Price prediction
3) GDP growth prediction
4) Employee attrition

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

Ans : Anscombe's quartet can be defined as group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that canhelp you identify the various anomalies present in the data like outliers, linear separability of the data etc.

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

Ans : Pearson's R means a statistical relationship between two variables , either positive or negative.
In other words, kts a measurement of how dependant two variables are on one another.

If correlation coefficient is -1, it indicates a strong negative relationship.

If correlation coefficient is 0, it indicates no relationship.

If correlation coefficient is 1, it indicates a strong positive relationship between the variables.

A higher absolute value of the correlation coefficient indicates a stronger relationshipbetween variables.

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

Ans: Scaling is a step of pre-processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. For eg: For house Price Prediction, we had features like number of bedrooms, number of bathrooms, area in sqft etc. Here the magnitude of "area in sqft" and "number of bedrooms" are entirely different or it is not comparable. In order to make these predictors on same or comparable magnitude we need to use scaling. I f scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling.

Normalized : It brings all of the data in the range of 0 and 1.

Sklearn.preprocessing.MinMaxScaler helps to implement normalization in python
MinMaxScaling : **x = x – Min(x) / Max(x) – Min(X)**

Standardized scaling : Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean zero and standard deviation one.

Standaisation : **x = x – mean(x) / stddev(x)**

In case of Normalization, it loses some information in the data especially about outliers.

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

(3 marks)

Ans : When there is a value of VIF = Infinite, it means there is a perfect correlation between two independent variables. AS we know the formula of VIF :

VIF = 1 / 1 - R2

In case of perfect correlation, where model covers all variations in the data then R2 becomes 1.
As per above VIF formula it gives 1/0 which is infinity.

To solve this problem we need to drop one of the highly correlated columns.
Eg: If we have 2 columns in our dataset, Temp in C and Temp in K. We know as one of the variable increases other will decrease or vice versa. In this case we can drop any one variable.

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

(3 marks)

Ans : To asses if the given data

is normally distributed for ease of inferring useful information, we can use Q-Q plot. It is a graphical tool to validate if two datasets are coming from populations with common distribution.

Quantiles are the breakpoints that divide the ordered numerical data into equal sized bins.

Percentiles are a type of quantiles that divide the data into 100 equal bins, quartile divide the data into 4 equal parts.

Importance of Q-Q plot :

- Two datasets/samples can of different size.
- Q-Q plot can detect outliers.
- We can also validate one of the assumption of Linear Regression i.e. residual of the model is normally distributed.