Assignment based subjective questions

Question 1:

From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer 1:

Most categorical variables have p-value ~0 , meaning they are significant in the model and directly influence dependent variable.

Question 2:

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

Answer 2:

We create dummy variables for categorical variable. Categorical variable has fixed set of values, let's say some categorical variable has N values. Dummy variables are of binary in nature (0/1). So we can represent N different values with N-1 variables. As one of the value can always be mapped by 0's for rest of the dummy variables.

Question 3:

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

Answer3:

Target variable cnt has very high correlation (0.63) with temp and atemp.

Question 4:

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

Answer 4:

We validate the assumptions of Linear Regression by performing residual analysis after building the model on training set. Assumptions are error mean is centred around 0 and normally distributed.

Question 5:

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

Answer 5:

Apart from time-based growth (year on year from 2018 to 2019), Weathersit is contributing significantly towards explaining the demand of shared bikes. Next significant variable is temperature.

General Subjective Questions

Question 1:

Explain the linear regression algorithm in detail.

Answer 1:

- Linear regression algorithm is used to build a linear math equation to model the data.
- Generally, one variable is target variable and one or more variables are predictor variables. And it is assumed that target variable is linearly related to predictor variables.
- Difference in value for the target variable as per the straight line and as per actual value is called error. And Linear Regression algorithm tries to minimize this error using mean squared error function.
- The error mean is assumed to centred around 0 with normal distribution and constant variance.
- With help of model built by linear regression we can find out what variables are influencing target variables and also quantify their impact.

Question 2:

Explain the Anscombe's quartet in detail.

Answer 2:

- Anscombe's quartet talks about importance of data visualization over statistical properties of data like mean, standard deviation and correlation.
- It provides examples dataset for which all 3 properties are same but they are completely different datasets.

Question 3:

What is Pearson's R?

Answer 3:

- Pearson's R is correlation coefficient used in bi-variate analysis.
- Its typical values ranges between -1 and 1.
- Values from -1 to 0 indicates negative correlation for 2 variables, meaning increase in value for one variable results in decreased value for another variable.
- Values from 0 to 1 indicates positive correlation for 2 variables, meaning increase in value for one variable results in increased value for another variable.
- Value 0 indicates there is no correlation
- Irrespective of sign (+/-), higher value indicates stronger correlation.

Question 4:

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

Answer 4:

- In Linear Regression with multiple variables, generally each variable has values in different ranges. For example, in bike sharing dataset, year had value 2018,2019, temp has value between 2 to 35 and hum has value between 0 to 97. To bring all the variables to same range of values, scaling is performed on them.
- There are 2 type of scaling
    - Normalized scaling (MinMax Scaling)
        - In mean-max scaling first min (Xmin) and max (Xmax) is found out for each variable and then variable value is replaced with scaled value X'.
        - Formula for scaled value
            - X' = (X-Xmin)/(Xmax-Xmin)
        - The scaled values will always be between 0 to 1
        - Useful when distribution of data is not Gaussian or Unknown
    - StandardSclaing
        - In StandardScaling first mean (Xu) and Standard deviation (SD) is calculated and then variable value is replaced by scaled value X'.
        - Formula for scaled value
            - X' = (X-Xu)/SD
        - Useful when distribution of data is Gaussian or Unknown
        - Centres the data around mean and scales to a standard deviation of 1

Question 5:

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

Answer 5:

- VIF stands for Variance Inflation Factor and is a measure of correlation between variables.
- The formula for VIF = 1/(1-R2), where R2 = 1- RSS/TSS
- For VIF to be infinite, R2 has to be 1 and R2 to be 1 RSS has to be 0.
- Meaning there is not residual error and both variables are having exact same values

---

Question 6:

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

Answer 6:

- Q-Q plot is short name for quantile-quantile plot, where for any variable scatter plot is generated for quantile of sample distribution against theoretical quantile values.
- It provides a test for normal distribution, as for samples to be normally distributed it must follow its' theoretical quantiles values.
- If the scatter plot is straight line following y=x, then it confirms the normal distribution.
- For Linear regression, we assume that error is normally distributed and centred around 0. For validating this assumption Q-Q plot can be very useful.