⊛ Assignment - Based Subjective Questions:

① From your Analysis of Categorical variables from the dataset, what could you infer about their effect on dependent variable?

Ans: In the given assignment, I first Identified categorical variables ; · season · yr · mnth · holiday · week day · working day · weathersit against the target variable · cnt · and Exploratory Data Analysis with Visualizations was done. It was very clear that · weather situation median was around 50,000 approximately and similar things were observed in · season · yr.

And the final model building shows a significant growth of $R^2$ and adjusted $R^2$ for · yr · season etc.

∴ Yes, categorical variables had a major & significant impact on the dependent variable.

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

Ans: Yes, it is highly advisable to use drop_first = True, because it drops the first column of dummies which helps in reducing the extra column.
therefore it reduces co-relations created among dummy variables.

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

Ans: Looking at the pairplot, the variable having highest correlation with the target variable 'cnt' is 'temp' ('atemp') because 'registered' variable was removed in Data Preparation.

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

Ans: following steps are carried out;

a) Test for Normal Distribution of Errors (Error terms / Residuals) by visualizing them on a distribution plot.

b) Add and drop variables based on each model VIF, and p-values to avoid multicollinearity.

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

Ans: ① 'temp' ; temperature feels in °C

② 'yr' ; year of the records made

③ 'winter' ; sub-category of season (4)

● General Subjective Questions:

① Explain Linear Regression in detail.

It is a Machine Learning algorithm. It is a Supervised Learning algorithm.

this performs regression tasks, linearly, on models based on independent and dependent variables.

This model is used to deduce relationship b/w different variables and their predicted (forecasting) values.

It is the process of fitting a straight line b/w the independent and dependent variables in the available dataset and predicting the future value of variables.

A simple Linear Regression model explains the relationship between a dependent and independent variable using a straight line.

Generally denoted by equation: $y = mx + c$

m; slope of the line

c: intercept

also denoted as $y = \beta_0 + \beta_1 x$

Residuals: defined as difference b/w y-coordinates of actual value and predicted value.

$$RSS = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2$$

for multiple linear Regression: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$

y: predicted value, $\beta_0$: constant ; $\beta_1 \dots \beta_n$: model parameters.

② Anscombe's Quartet:

It compreises of four Data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.

Each dataset contains of eleven $(x, y)$ points. It helps to demonstrate both the importance of graphing data when analyzing it, and effects of outliers and other influential observations.

For all 4 data sets:

Mean of x, Sample Varience of x, Mean of y,

Sample Varience of y, Correlation of x & y, Linear Regression line, $R^2$ is calculated.

③ Pearson's R ?

It is also known as Pearson correlation co-efficient, Bivariate correlation. It is a measure of linear correlation between two sets of data.

It is the ratio between the covariance of two variables and the product of their standard deviations;

thus it is essentially a normalized measurement of covariance such that the result always has a value b/w -1 and +1.

The measure can only reflect a linear correlation of variables, and ignores many other types of relations/correlation

④ What is Scaling? why is it performed? Normalized v/s Standardized.

Scaling, also known as Feature Scaling basically is putting the feature values into the same range so that computations can be done on different variables fairly

① Ease of Interpretation

② Faster convergence for Gradient methods.

Model Accuracy is not affected by scaling but only coefficients are altered.

Scaling methods never change the shape or distribution of the original variable, It only scales and shifts them.

Standardization brings all data into standar Normal distribution

$$x = \frac{x - mean(x)}{sd(x)}$$

Normalization brings all data withing range of 0 to 1.

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

⑤ Sometimes value of VIF infinite. Why?

If perfect correlation ; VIF = infinite. this shows perfect correlation b/w 2 independent variables.

In case of perfect corr, we get $R^2 = 1$, leading $1/R^2$ to infinity

An infinite VIF value indicats that the corresponding variable may be expressed exactly by a linear combination of other variables

⑥ What is Q-Q plot? Use & importance in Linear Regression.

Quantile - Quantile plots are plots of two quantiles against each other. A Quantile is a fraction where certain values fall below that quantile.

Purpose of Q-Q plots is to find out if two sets of data come from the same distribution.

A Q-Q plot is used to compare the shapes of distributions providing a graphical view of how properties such as location, scale, skewness are similar or different in two distributions.