Assignment-based Subjective Questions

Q1 : From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable ?

- Firstly, we use the df.nunique() function to check the distinct rows for a variable.
- Here, in the bike dataset, we found the following categorical variables.

season	4
yr	2
mnth	12
holiday	2
weekday	7
workingday	2
weathersit	3

- Bike demand is highest in the Fall season, while lowest in the spring season. Summer and Winter were in between.
- For weather situation, the first condition where the weather is clear has highest cnt value and is most favorable, and when there is heavy snow/rainfall there are no users indicating a non-favorable event.
- Rentals are lower during holiday
- For months, ‘September’ had the most rentals, while ‘December’ had the fewest rentals.
- Weekends had more bookings as compared to weekdays.
- Working day had nominal impact on the dependent variable.

Q2 : Why is it important to use **drop_first=True** during dummy variable creation ?

- During EDA (Exploratory Data Analysis), to analyze categorical variables, we create dummy variables based on the number of unique values in a particular variable.
- All these dummy variable columns are correlated to each other which we call as multicollinearity.
- Multicollinearity isn’t a good sign while processing categorical data as it makes the coefficients unstable and unreliable.
- Thus, when we use the **pandas.get_dummies** function, we mention an argument as drop_first = True.
- To generalize, if we have ‘n’ dummy variables, ‘n-1’ dummy variables will be able to predict the value of the nth dummy variable.

- ➔ For example, if we have a column called Payment mode with 3 unique values viz. UPI, Card and Net banking. Thus, will take only 2 variables let's say UPI will be 1-0 and Card will be 0-1, so we don't need a separate column for Net banking because 0-0 will indicate Net banking. Thus, will drop the Net banking variable.

Q3 : Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable ?

- ➔ 'temp' and 'atemp' both variables have higher correlation with the target variable 'cnt'.

Q4 : How did you validate the assumptions of Linear Regression after building the model on the training set ?

- ➔ Validation of assumptions after model building were done on following aspects :
  - i. Linearity of relationship between the predictor variables and output variable.
  - ii. Normality of the error distribution
  - iii. Constant variance of the errors (Homoscedasticity).
  - iv. By calculating the VIF (Variance Inflation Factor) between features (Less Multicollinearity)

Q5 : Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes ?

- ➔ After implementing the final model, 'temperature' , 'weather situation' and the year variable have a significant impact on the target variable.

General Subjective Questions

Q1 : Explain the linear regression algorithm in detail.

- ➔ Machine learning techniques are classified in three different categories. They are mainly regression classification and clustering. Linear regression is one of the most popular regression method algorithms used in ML.
- ➔ As the name suggests, linear regression assumes linear relationship between predictor variable and independent variable. The primary formula for linear regression algorithm is y = mx+c, where 'm' indicates the slope of the line and 'c' indicates the intercept.
- ➔ Linear regression algorithm is used to predict the output of a continuous numerical variable. It is classified in two types of viz. Simple Linear regression(SLR) and Multiple Linear regression (MLR).
- ➔ The method to implement Linear regression is to divide the i/p dataset into Train and test sections in a suitable ratio of 70-30 or 80-20.
- ➔ Then we use the training part of the dataset to visualize the relationship between the 'x' and 'y' variables.
- ➔ For linear regression, we use the statsmodel or the scikit-learn libraries to train the model on the training dataset.
- ➔ Thereafter we get a straight line with minimum mean square error. Thus, we calculate the M&C values off the straight-line equation.
- ➔ Then we apply the same formula to the test data set and calculate the output value of 'y'.

Q2 : Explain the Anscombe's quartet in detail.

- ➔ Anscombe's quartet is an illustration of the fact that analyzing the dataset through visualization gives a better understanding of the data rather than numerical statistics.
- ➔ Anscombe's quartet comprises a set of 4 data sets having identical descriptive statistical properties in terms of mean, variance, R-squared, correlations and linear regression lines but having different representations.
- ➔ This method is named after the statistician Francis Anscombe in 1973 to demonstrate the importance of visualising data and to show that summary statistics alone can be misleading.
- ➔ In this method, we find mean, STD, correlations, slope and intercept of 'x' and 'y' for all four datasets and create a statistical summary of all.
- ➔ However, when we examine the data by drawing scatter plots for the four datasets, we can observe the inherent difference even if the statistical values appear to be uniform.

Q3 : What is Pearson's R ?

- ➔ Pearson correlation coefficient, denoted as r, is a statistical measure that calculates the strength and direction of the linear relationship between two variables on a scatter plot.

- ➔ Value of r ranges between -1 and 1.
    - o 1 indicates a perfect positive linear relationship
    - o -1 indicates a perfect negative relationship
    - o Zero indicates no linear relationship between the variables
- ➔ The Pearson correlation coefficient essentially captures how closely the data points tend to follow a straight line when plotted together it's important to remember that correlation does not imply causation - just because 2 variables are related does not mean one causes the change in the other.
- ➔ Types of Pearson correlation coefficient :-
    - o adjusted correlation coefficient
    - o weighted correlation coefficient
    - o reflective correlation coefficient

  etc.
- ➔ For Pearson correlation coefficient to work, there are certain assumptions :
    - o **Linear relationship**:- it assumes a linear relationship between the 2 variables under consideration.
    - o **Normality** :- the variables should follow a normal distribution.
    - o **Homoscedasticity** :- this assumption suggests that the variability in one variable should be consistent across all levels of other variable.
    - o **Independence** :- the observations used to compute the correlation coefficient should be independent of each other. Independence ensures that each data point contributes uniquely to the analysis without being influenced by other observations.

Q4 : What is scaling ? Why is scaling performed ? What is the difference between normalised scaling and standardised scaling ?

- ➔ Scaling and normalization are important aspects of data pre-processing while building machine learning models
- ➔ Scaling is applied to independent variables to normalize the data within a particular range to speedup calculations.
- ➔ Many times, dataset features data with highly varying magnitudes, units and range. Without scaling, algorithm only takes magnitude in account and not units.
- ➔ To cope with this, we perform scaling to bring all the variables to the same level o magnitude.
- ➔ **Normalization scaling** brings all the data in the range of 0 to 1 while **standardized scaling** brings all the data into a standard normal distribution which has mean and standard deviation.

- ➔ Normalization has a disadvantage as compared to standardization that it loses some information in the data, especially about outliers.

Q5 : You might have observed that sometimes the value of VIF is infinite. Why does this happen?

- ➔ VIF stands for Variance Inflation Factor is a statistical measure to know the correlation among predictor variables.
- ➔ The formula of VIF is 1/1-R^2. Thus, for VIF to become infinite, the denominator term should be 0. For denominator to be 0, the value of R^2 should be 1.
- ➔ R^2 = 1 indicates that there is perfect correlation among two variables.
- ➔ To generalize, high value of VIF indicates that that there is a high correlation between variables.
- ➔ Thus, infinite VIF means perfect correlation between two independent variables.
- ➔ To solve this problem, we drop one of the variables in the dataset which is causing perfect multicollinearity.

Q6 : What is a Q-Q plot ? Explain the use and importance of a Q-Q plot in linear regression

- ➔ The "quantile-quantile" plot is a graphical method for determining if a data set follows a certain probability distribution or whether 2 samples of data came from the same population or not.
- ➔ Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution. They are commonly used in statistics, data analysis and quality control to check assumptions and identify departures from expected distribution.
- ➔ From linear regression perspective, Q-Q plot is a useful visualization tool with several different advantages
  - ○ **Detecting outliers**
  - ○ **Flexible comparison of distribution**
  - ○ **Visual interpretation**
  - ○ **Deviation sensitivity**
  - ○ **Assessing normality**

➔ Different types of Q-Q plots are Normal distribution, right-skewed distribution, left-skewed distribution, under dispersed distribution, over dispersed distribution etc.