Assignment-based Subjective Questions Name – Ritu Harivansh Prajapati

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

Ans - In the bike sharing dataset, let's consider the effect of the categorical variable 'weathersit' on the target variable 'cnt'. While performing EDA, I visualized the relationship between the categorical variables and the target variable. It was seen that during the weather situation 1 (Clear, few clouds, partly cloudy, a high number of bike rentals were made, with the median being 50,000 approximately. Similarly, certain inferences could be made 'season' and 'yr' as well. Also, during model building on inclusion of categorical features such as preseason etc we saw a significant growth in the value of R-squared and adjusted R-squared. This implies that the categorical features were helpful in explaining a greater proportion of variance in the dataset.

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

Ans - During dummy value creation (dummy encoding) it is advisable to use drop_first=True, otherwise we will get a redundant feature i.e. dummy variables might be correlated because the first column becomes a reference group during dummy encoding. For example, suppose we have a categorical feature 'is_male'. We use dummy encoding to get two features is_male_0 and is_male_1. After applying, get_dummies we get a table such as this.

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

Ans - The numerical variable 'registered' has the highest correlation with the target variable 'cnt', if we consider all the features. But after data preparation, when we drop registered due to multicollinearity the numerical variable 'atemp' has the highest correlation with the target variable 'cnt'.

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

Ans - Independent variables and the dependent variables could be transformed so that the relationship between them is linear. For instance, you could find that the relationship is linear between the log of the dependent variables and some of the independent variables squared (c.f. Polynomial Regression and Generalized Additive Models (GAM) for an interesting generalization of this.

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

Ans- Based on final model top three features contributing significantly towards explaining the demand are:

a- Temperature (0.552)
b- weathersit : Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds (-0.264)
c - year (0.256)

So it recommended to give these variables utmost importance while planning to achieve maximum demand

General Subjective Questions

1 - Explain the linear regression algorithm in detail.

Ans - Simple linear regression is a regression technique in which the independent variable has a linear relationship with the dependent variable. The straight line in the diagram is the best fit line. The main goal of the simple linear regression is to consider the given data points and plot the best fit line to fit the model in the best way possible.
Before moving on to how the linear regression algorithm works, let us take a look at a few important terminologies in simple linear regression.
Types Of Regression –
1 - **Simple Linear Regression** - One of the most interesting and common regression technique is simple linear regression. In this, we predict the outcome of a dependent variable based on the independent variables, the relationship between the variables is linear. Hence, the word linear regression.
2 **- Polynomial Regression** - In this regression technique, we transform the original features into polynomial features of a given degree and then perform regression on it.
3 **- Support Vector Regression** - For support vector machine regression or SVR, we identify a hyperplane with maximum margin such that the maximum number of data points are within those margins. It is quite similar to the support vector machine classification algorithm.
4 **- Decision Tree Regression** - A decision tree can be used for both regression and classification. In the case of regression, we use the ID3 algorithm (Iterative Dichotomise 3) to identify the splitting node by reducing the standard deviation.
5 **- Random Forest Regression** - In random forest regression, we ensemble the predictions of several decision tree regressions. Now that we know about different types of regression let us take a look at simple linear regression in detail.

2 - Explain the Anscombe's quartet in detail.

Ans - Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x,y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.
Quartet's Summary Stats
The summary statistics show that the means and the variances were identical for x and y across the groups:
1 - Mean of x is 9 and mean of y is 7.50 for each dataset.
2 - Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
3 - The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset
When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:
1 - Dataset I appear to have clean and well-fitting linear models.
2 - Dataset II is not distributed normally.
In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
Dataset IV shows that one outlier is enough to produce a high correlation coefficient.
This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

3 - What is Pearson's R?

Ans - In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus, it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.
The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association

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

Ans – Scaling - Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data pre-processing set.

Scaling Performed - Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc

Difference between normalized scaling and standardized scaling -

S.NO.	Normalisation	Standardisation
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called Min Max Scaler for Normalization.	Scikit-Learn provides a transformer called Standard Scaler for standardization.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.

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

Ans - If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Ans - Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q plots is to find out if two sets of data come from the same distribution. A 45-degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.