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?

Considering demand to be the dependent variable, the following observations are made:

- Season has an effect on demand – demand is higher during Summer and Fall
- Demand has increased from 2018 to 2019
- Demand is higher on non-holidays and working days
- Demand is higher when the weather is either clear or misty/cloudy

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

The get_dummies method in Pandas creates a column for each value that the categorical variable can take. This, effectively, creates an extra row.

For example, if the categorical variable has values X, Y, and Z, the value being Z could be inferred when the flags for X and Y are not set, meaning all the values are represented using only 2 dummy variables instead of 3.

So, we drop the first of these columns using the drop_first parameter in this method.

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

The “temp” variable has the highest correlation with the target variable. “atemp” also has a high correlation with the target variable. However, it should be noted that “temp” and “atemp” have high correlation with each other.

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

The error terms must be normally distributed. To ensure this, a distplot of the error terms was plotted. Also, temp has high correlation with the target variable.

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

The top 3 variables are temperature, humidity, and year.

Note: Year is an important variable for the model, but since weather is also shown to be a significant predictor, windspeed can also be considered as it has a high coefficient.

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Linear Regression is a supervised learning algorithm that tries to model the relationship between two or more variables using a linear equation.

The variable being predicted is called the target variable, and the other variables are called the predictor variables. If there is exactly one predictor variable, Linear Regression fits a straight line through the data points. If there are more than one, then a "hyperplane" is fit.

The crux of the Linear Regression algorithm is finding **the best line** to fit the data. The best line is defined as the line that results in the smallest difference between the predicted target values and the actual target values (the difference is called a Residual). The most common metric for finding this difference is the method of least squares. This method minimizes the sum of the squares of the residuals.

It should be noted that Linear Regression only works under certain conditions. A linear relationship must exist between the variables of interest. Also, the residuals must be normally distributed with a mean of 0 and a constant variance.

Q2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a demonstration of the danger of using only summary statistics (mean, variance, correlation, etc.) when analyzing a dataset. It was constructed by English statistician Frank Anscombe.

The quartet is made up of 4 datasets. Each dataset contains 11 (x, y) pairs. The datasets have nearly identical summary statistics as follows:

- Mean of x: 9
- Variance of x: 11
- Mean of y: 7.50
- Variance of y: 4.125
- Correlation between x and y: 0.816
- Linear regression line: y = 0.500x + 3.00

When plotted, however, the datasets are drastically different from each other. The first is nearly linear, the second is polynomial, the third is linear but for an outlier, and the fourth has constant x values but for an outlier.

The main takeaway of the Anscombe quartet is that summary statistics, while important, may be misleading on their own; it is important to use them as only one part of a much larger data analysis process.

Q3. What is Pearson's R?

Pearson's R is a correlation coefficient. A correlation coefficient is a measure of the strength of the relationship between two variables. Pearson's R is the most commonly used correlation coefficient in Linear Regression. It shows the linear relationship between two sets of data.

Like any correlation coefficient, Pearson's R ranges from –1 to 1.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

In Machine Learning, Feature Scaling is a part of the data preprocessing step. During scaling, the values of all the independent variables are adjusted to have the same range. This is done so that the model assigns weights to the variables correctly. Also, if scaling is not performed, then the modeling algorithm performs slower than it would otherwise.

For example, consider a dataset where the independent variables are Age in Years (18 – 70), Years of Experience (0 – 30), Height in feet (4.5 - 6.5), and Weight in kg (60 – 100). After scaling, all these variables would have the same range.

Depending upon the scaling technique used, the range could be

- between 0 and 1 – the scaling performed is called Normalization, also called Min-Max scaling
- centered at 0 with a standard deviation of 1 – the scaling method is called Standardization.

Standardization is preferred over Normalization when it is preferable to retain the spread in the original data in the scaled data.

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

Variance Inflation Factor (VIF) is calculated as:

$$\text{VIF} = 1 \div (1 - R^2)$$

When the R-squared value is 1, then the VIF is infinite. If there is perfect correlation between two independent variables, then the R-squared value becomes 1. To solve this problem, we need to drop one of these variables from our model.

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

A quantile-quantile (Q-Q) plot is used to compare the quantiles of given data against the quantiles of a desired distribution (normal distribution by default). This means that Q-Q plot is a way of checking whether a given set of data is normally distributed.

In the case of Linear Regression, it is a requirement that the error terms be normally distributed. So, a Q-Q plot would be one way of verifying this, in addition to a histogram/distplot.