Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Ans 1. The demand for bikes is less in the month of spring when compared with other seasons and also the demand is less in the months of Jan and Feb. Also rain also affects the demand for bikes.

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

Ans 2. If we skip the use of drop_first = True, then **n** dummy variables will be presented instead of **n-1**, and these predictors(**n** dummy variables) are themselves correlated which is known as multicollinearity and it, in turn, leads to Dummy Variable Trap.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans 3. Variables atemp and temp have the highest correlation among all numerical variables.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans 4. Assumption 1:- The Dependent variable and Independent variable must have a linear relationship - A simple pairplot of the dataframe helped me see if the Independent variables exhibit linear relationship with the Dependent Variable.

Assumption 2:- Residuals must be normally distributed - Used a Distribution plot on the residuals and see if it is normally distributed.

Assumption 2:- There is a linear relationship between the features and target - Linear regression captures only linear relationships. This can be validated by plotting a scatter plot between the features and the target.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans 3. The Top 3 features contributing significantly towards the demands of share bikes are:

- temp and atemp
- weathersit(light_Snow/light_rain).
- yr

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Ans 1. Linear regression finds a linear equation that best describes the correlation of the explanatory variables with the dependent variable. This is achieved by fitting a line to the data. The line tries to minimize the sum of the squares of the residuals. The residual is the distance between the line and the actual value of the explanatory variable. Finding the line of best fit is an iterative process.

The following is an example of a resulting linear regression equation:

y = b + b1x1 + b2x2 + ….

In the example above, y is the dependent variable(or the target variable), and x1, x2, and so on, are the explanatory variables. The coefficients (b1, b2, and so on) explain the correlation of the explanatory variables with the dependent variable. The sign of the coefficients (+/-) designates whether the variable is positively or negatively correlated. b0 is the intercept that indicates the value of the dependent variable assuming all explanatory variables are 0.

A linear regression model helps in predicting the value of a dependent variable, and it can also help explain how accurate the prediction is. This is denoted by the R-squared and p-value values. The R-squared value indicates how much of the variation in the dependent variable can be explained by the explanatory variable and the p-value explains how reliable that explanation is. The R-squared values range between 0 and 1. A value of 0.8 means that the explanatory variable can explain 80 percent of the variation in the observed values of the dependent variable. A value of 1 means that a perfect prediction can be made, which is rare in practice. A value of 0 means the explanatory variable doesn't help at all in predicting the dependent variable. Using a p-value, you can test whether the explanatory variable's effect on the dependent variable is significantly different from 0.

There is one more value called adjusted R-squared value which takes in the factor of awarding penalties based on the number of variables taken , the more the variables(features) the more the penalty. If the difference between R-squared and adjusted R-squared is more than 5% then it makes the model overfitting.

2. Explain the Anscombe’s quartet in detail. (3 marks)

Ans 2. Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots. Anscombe’s quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

Data Set 1: fits the linear regression model pretty well.
Data Set 2: cannot fit the linear regression model because the data is non-linear.
Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model.
Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

3. What is Pearson’s R? (3 marks)

Ans 3: The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. Pearson's r is a numerical summary of the strength of the linear association between the variables.

The Pearson's correlation coefficient varies between -1 and +1 where:

1. r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
2. r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
3. r = 0 means there is no linear association
4. r > 0.5 means there is a weak association
5. r > 0.5 < 0.8 means there is a moderate association
6. r > 0.8 means there is a strong association

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans 4: Scaling is a part of data preprocessing which helps to bring the data into a certain range. Scaling is performed so that no variable overpowers or suppresses the effect of other variables. For example let's say we have two variables one in the range of 200-300 and one in the range of 10-15 so the 200 one will be overpowering the other as its coefficient will be larger this will lead to a biased model. So to avoid this we use scaling. There are two types of scaling.

Normalization typically rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

Normalisation

- Minimum and maximum value of features are used for scaling
- It is used when features are of different scales.
- Scales values between [0, 1] or [-1, 1].
- It is really affected by outliers.
- Scikit-Learn provides a transformer called MinMaxScaler for Normalization.
- This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.
- It is useful when we don’t know about the distribution
- It is a often called as Scaling Normalization

Standardization

- Mean and standard deviation is used for scaling.
- It is used when we want to ensure zero mean and unit standard deviation.
- It is not bound to a certain range.
- It is much less affected by outliers.
- Scikit-Learn provides a transformer called StandardScaler for standardization.
- It translates the data to the mean vector of original data to the origin and squishes or expands.
- It is useful when the feature distribution is Normal or Gaussian.
- It is often called Z-Score Normalization.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans 5: 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 leads 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. (3 marks)

Ans 6: Q-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 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.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

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