Assignment-Based Subjective Questions & Answers

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

Answer: From the boxplot Analysis that was done, the following were the observations regarding the effect on the dependent variable:

a. The value of dependent variable ‘cnt’ (demand for bicycle) was seen to be highest for the Fall season category and lowest for the spring season category
b. For the days when either it was a holiday or a weekend the ‘cnt’ variable value was seen to be the highest, compared to weekdays
c. For the weather conditions: Clear, Few clouds, Partly cloudy, Partly cloudy - the dependent variable value was seen to be highest. Whereas it was seen to be lowest for Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
d. For the year 2019, the dependent variable value was seen to be high as compared to 2018
e. For the months of Sep & Oct, the dependent variable values are seen to be high whereas for Jan & Feb, it is seen to be lowest

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

Answer: The drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

Answer: The feeling temperature in Celsius variable indicated by ‘atemp’ had the highest correlation with the target variable of ~ 0.63

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

Answer: The assumptions were validated by doing Residual Analysis to see if the error terms are normally distributed and by plotting a correlation plot between y_pred and y_test to see if the spread is not varying between the two terms

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

Answer: Temperature, windspeed and year were observed to be the top 3 features contributing significantly towards explaining the demand of the shared bikes.

General Subjective Questions & Answers

1. Explain Linear Regression model in detail.

Answer: Linear regression is an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. It is a statistical method used in data science and machine learning for predictive analysis.
The independent variable is also the predictor or explanatory variable that remains unchanged due to the change in other

variables. However, the dependent variable changes with fluctuations in the independent variable. The regression model predicts the value of the dependent variable, which is the response or outcome variable being analyzed or studied.

Thus, linear regression is a supervised learning algorithm that simulates a mathematical relationship between variables and makes predictions for continuous or numeric variables such as sales, salary, age, product price, etc.

This analysis method is advantageous when at least two variables are available in the data, as observed in stock market forecasting, portfolio management, scientific analysis, etc.

A sloped straight line represents the linear regression model.

It is represented mathematically by Y = m*X + b

Where X = dependent variable (target)

Y = independent variable

m = slope of the line (slope is defined as the ‘rise’ over the ‘run’)

The above process applies to simple linear regression having a single feature or independent variable. However, a regression model can be used for multiple features by extending the equation for the number of variables available within the dataset.

The equation for multiple linear regression is similar to the equation for a simple linear equation, i.e., y(x) = p0 + p1x1 plus the additional weights and inputs for the different features which are represented by p(n)x(n). The formula for multiple linear regression would look like,

$$y(x) = p0 + p1x1 + p2x2 + \ldots + p(n)x(n)$$

The machine learning model uses the above formula and different weight values to draw lines to fit. Moreover, to determine the line best fits the data, the model evaluates different weight combinations that best fit the data and establishes a strong relationship between the variables.

Furthermore, along with the prediction function, the regression model uses a cost function to optimize the weights (pi). The cost function of linear regression is the root mean squared error or mean squared error (MSE).

Fundamentally, MSE measures the average squared difference between the observation's actual and predicted values. The output is the cost or score associated with the current set of weights and is generally a single number. The objective here is to minimize MSE to boost the accuracy of the regression model.

2. Explain the Anscombe's quartet in detail

Answer : 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.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

3. What is Pearson's R?

Answer : 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.

The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

Although interpretations of the relationship strength (also known as effect size) vary between disciplines, the table below gives general rules of thumb:

Pearson correlation coefficient (r) value	Strength	Direction
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

The Pearson correlation coefficient is also an inferential statistic, meaning that it can be used to test statistical hypotheses. Specifically, we can test whether there is a significant relationship between two variables.

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

Answer: Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm. It is performed because 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.

Normalization/Min-Max Scaling: It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling: Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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

sklearn.preprocessing.scale helps to implement standardization in python.
One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

Answer: 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.

Answer: 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.

References:

1. Google
2. Wikepedia
3. Towards Data Science.com (Medium)
4. Programsbuzz.com