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:

Categorical variables in the dataset are season, weathersit, yr, mnth, holiday, weekday. From the analysis, it can be inferred as below:

- Good weather has good demand for bike sharing.
- Fall has the highest bike demand.
- The demand is increasing year-on-year (2018 vs 2019).
- Demand is continuously growing each month till June. September month has highest demand. After September, demand is gradually decreasing.
- Some days of the week has moderately higher demand than the other days but it is not clearly giving demand.
- Bike demand has decreased on a holiday.

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

Ans:

We use parameter drop_first = True to drop the first dummy variable, thus it will give n-1 dummies out of n discrete categorical levels by removing the first level. Hence it reduces the correlations created among dummy variables by reducing the extra column.

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

Ans:

temp and atemp has the highest correlation with the target variable ‘cnt’ based on the pairplot.

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

Ans:

- Relationship is linear between target and predictor variables.
- Error terms are normally distributed with mean at 0.0.
- Error terms does not vary much as the predictor variable changes. Which means constant variance of the errors or Homoscedasticity.
- Low Variance Inflation Factor (VIF) or Less Multi-collinearity between features.

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:

The top 3 features directly influencing the count are the features with highest coefficients. These are: temp, September month and year.

General Subjective Questions

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

Ans:

Linear regression is a supervised machine learning technique. It analyses past/historic data to establish relationship/pattern between a set of input and one output, such that it can be represented by a formula y= mx + c that can be used to predict future values of output based on new inputs. This is achieved by fitting a line to the data using least squares. 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 input variable. Finding the line of best fit is an iterative process and a model is finalized. The following is an example of a resulting linear regression equation:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots$$

In the example above, y is the dependent or output variable, and x1, x2, and so on, are the independent or input variables. The coefficients (b1, b2, and so on) explain the correlation of the independent 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 independent 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 independent 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 independent 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 independent variable doesn't help at all in predicting the dependent variable. Using a p-value, you can test whether the independent variable's effect on the dependent variable is significantly different from 0.

According to the conditions of linear regression which states that the error curve must be normally distributed, we proceed to testing the model with the test dataset. The conclusion hence drawn on the model would be used to provide valuable insights/predictions on datapoints in the range of the model.

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

Ans:

Anscombe’s quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm. 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.

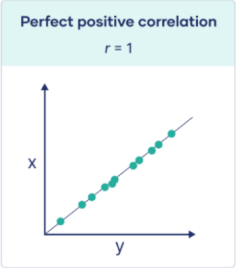
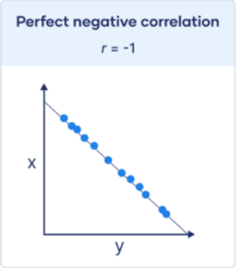
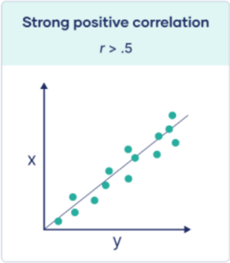
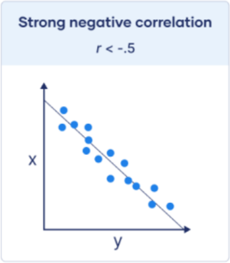
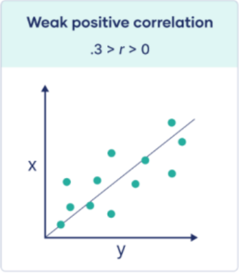
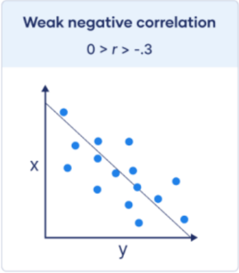
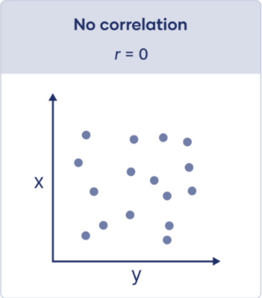
So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

3. What is Pearson’s R? (3 marks)

Ans:

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. When one variable changes, the other variable changes in the same direction.

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

When r is 1 or -1, all the points fall exactly on the line of best fit:	When r is greater than .5 or less than –.5, the points are close to the line of best fit:
 	 
When r is between 0 and .3 or between 0 and –.3, the points are far from the line of best fit:	When r is 0, a line of best fit is not helpful in describing the relationship between the variables:
 	

Formula for calculating the Pearson correlation coefficient (r):

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

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

Ans:

Feature scaling is a data pre-processing technique used to transform the values of features or variables in a dataset to a similar scale. The purpose is to ensure that all features contribute equally to the model and to avoid the domination of features with larger values.

Feature scaling becomes necessary when dealing with datasets containing features that have different ranges, units of measurement, or orders of magnitude. In such cases, the variation in feature values can lead to biased model performance or difficulties during the learning process.

There are several common techniques for feature scaling, including standardization, normalization, and min-max scaling. These methods adjust the feature values while preserving their relative relationships and distributions.

By applying feature scaling, the dataset's features can be transformed to a more consistent scale, making it easier to build accurate and effective machine learning models. Scaling facilitates meaningful comparisons between features, improves model convergence, and prevents certain features from overshadowing others based solely on their magnitude.

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Here's the formula for normalization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Here, Xmax and Xmin are the maximum and the minimum values of the feature, respectively.

Standardization is another Feature scaling method where the values are centred around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero, and the resultant distribution has a unit standard deviation.

Here's the formula for standardization:

$$X' = \frac{X - \mu}{\sigma}$$

μ: mean of the feature values

σ: the standard deviation of the feature values

Difference between Normalization and Standardization:

Normalization	**Standardization**
Rescales values to a range between 0 and 1	Centres data around the mean and scales to a standard deviation of 1

Useful when the distribution of the data is unknown or not Gaussian	Useful when the distribution of the data is Gaussian or unknown
Sensitive to outliers	Less sensitive to outliers
Retains the shape of the original distribution	Changes the shape of the original distribution
May not preserve the relationships between the data points	Preserves the relationships between the data points
Equation: (x – min)/(max – min)	Equation: (x – mean)/standard deviation

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

(3 marks)
Ans:
The greater the VIF, the higher the degree of multicollinearity. If there is perfect correlation, multicollinearity is perfect (i.e., the regressor is equal to a linear combination of other regressors), then VIF tends to infinity.
The R-squared value will be 1. Hence VIF, which is (1/(1-R^2)) turns out to approach infinity.

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

(3 marks)

Ans:

Q-Q plot, or quantile-quantile plot, is a graphical tool that help us assess the validity of some assumptions in regression models, such as normality, linearity, and homoscedasticity. It is used to assess if sets of data come from the same statistical distribution. It is particularly helpful in linear regression when we are given testing and training datasets differently. In this scenario, it becomes important to check whether both sets of the data come from the same background, in order to maintain the sanity of the model.