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. I have worked on a dataset with details on the car's pricing and its kind (compact, sedan, or SUV). The cost of the car serves as the dependent variable in this scenario, and the type of car serves as the categorical variable.
2. We can plot the price of each type of automobile as a bar graph or box graph to examine the relationship between the categorical variable (type of car) and the dependent variable (price).
3. Analyzing the effect of a categorical variable on a dependent variable involves visualizing and statistically testing for differences in the distribution of the dependent variable across the categories of the categorical variable.
4. This analysis can provide insights into how the categorical variable affects the dependent variable, which can be useful for making predictions and understanding patterns in the data.

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

Ans:

To prevent the dummy variable trap, which happens when one dummy variable can be anticipated from the others, it is crucial to use drop_first=True when creating dummy variables. We eliminate the redundant information and avoid the trap by deleting the first dummy variable.

the redundant information and avoid the trap by deleting the first dummy variable.

As a result, our model has fewer variables and it is simpler to understand the coefficients.

Furthermore, it can enhance the efficiency of several machine learning methods by reducing multicollinearity between the variables.

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

Ans: 'temp' variable has the highest correlation with the target variable in the case study

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

Ans:

1. Normality of error terms
   Error terms should be normally distributed
2. Multicollinearity check
   There should be insignificant multicollinearity among variables
3. Linear relationship validation
   Linearity should be visible among variables
4. Homoscedasticity
   There should be no visible pattern in residual values
5. Independence of residuals

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 contributing significantly towards explaining the demand of the shared bikes are:-

1. Temp
2. month_9 (September)
3. weathersit_3

General Subjective Questions

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

ans:

1. The relationship between a dependent variable and one or more independent variables can be modelled statistically using linear regression.
2. The linear equation that best fits the data is estimated by the algorithm. Y is the dependent variable, X1, X2,..., and Xn are the independent variables, and b0, b1, b2,..., bn are the coefficients that need to be estimated. The equation can be written as Y = b0 + b1X1 + b2X2 +... + bn*Xn.
3. The sum of the squared residuals between the predicted values and the actual values is minimised by the algorithm using a technique known as Ordinary Least Squares (OLS).
4. The R-squared statistic, which reflects the percentage of variance in the dependent variable that is explained by the independent variables, can be used to assess the validity of a linear regression model.
5. The independent variables must not be substantially connected with one another in order for linear regression to work.
6. Both basic linear regression and multiple linear regression, which both contain more than one independent variable, can be performed using linear regression.

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

Ans:

1. Anscombe's quartet is a set of four datasets that have nearly identical statistical properties, but appear very different when graphed
2. There are 11 X and Y values at each position in each dataset. The mean, variance, correlation coefficient, and linear regression line are identical across the datasets.
3. The datasets have comparable statistical characteristics, yet when plotted, they look substantially different. One dataset shows a linear relationship, another a non-linear relationship, a fourth has a strong link with the exception of an outlier that drastically alters the regression line.
4. The quartet emphasises the constraints of relying exclusively on summary statistics to draw inferences about the data as well as the significance of visualising data to discover patterns and relationships.

3. **What is Pearson's R? (3 marks)**

Ans:

1. A numerical evaluation of the strength of the linear connection between the variables is provided by Pearson's r.

2. The correlation coefficient will be positive if the variables tend to rise and fall together.
3. The correlation coefficient will be negative if the variables have a tendency to rise and fall in opposition, with low values of one variable correlated with high values of the other. Between +1 and -1 are the possible values for the Pearson correlation coefficient, or r.
4. There is no link between the two variables, as indicated by a value of 0. Positive associations have values greater than 0, meaning that if one variable's value rises, so does the value of the other.
5. A negative value is one with a value less than 0.

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

1. Feature scaling is a method for uniformly distributing the independent features in the data over a predetermined range.
2. It is done as part of the pre-processing of the data to deal with extremely variable magnitudes, values, or units.
3. In the absence of feature scaling, a machine learning algorithm would often prioritise larger values over smaller ones, regardless of the unit of measurement.
4. Example: If an algorithm does not use the feature scaling approach, it may assume that a value of 3000 metres is greater than a value of 5 kilometres, which is not the case and causes the algorithm to produce inaccurate predictions. To solve this problem, we employ feature scaling to equalise all values' magnitudes.

Normalized scaling and standardized scaling are both techniques used to scale data in machine learning. Here are the main differences between them:

1. Normalized scaling: Rescales the values of a feature between 0 and 1. It is appropriate when the distribution of the feature is unknown or skewed. The formula for normalized scaling is X_normalized = (X - X_min) / (X_max - X_min). Normalized scaling preserves the shape of the distribution, but compresses the range of values. It is useful when the absolute values of features are not important, only their relative values.

2. Standardized scaling: Rescales the values of a feature to have a mean of 0 and standard deviation of 1. It is appropriate when the distribution of the feature is Gaussian or approximately Gaussian. The formula for standardized scaling is X_standardized = (X - X_mean) / X_std. Standardized scaling shifts the distribution to have a mean of 0 and adjusts the spread of values. It is useful when comparing features with different units of measurement. Standardized scaling can be sensitive to outliers in the data.

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

1. VIF = infinity if there is perfect correlation. A high VIF score denotes a strong connection between the variables.
2. The presence of multicollinearity causes the variance of the model coefficient to be exaggerated by a factor of 4 if the VIF is 4.
3. VIF displays a complete correlation between two independent variables when its value is infinite.
4. If the correlation is perfect, we have R-squared (R2) = 1, which results in 1/ (1-R2) infinite. To fix this, we must remove the variable from the dataset that is the exact multicollinearity's cause.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
Ans:

1. A graphical method for assessing if two data sets originate from populations with a common distribution is the quantile-quantile (q-q) plot.
2. The quantiles of the first data set are plotted against the quantiles of the second dataset in a q-q figure. A quantile is the percentage of points that fall below the specified number. In other words, the 0.3 (or 30%) quantile is the value at which 30% of the data are below it and 70% are above it. Additionally, a 45-degree reference line is plotted
3. The points should roughly lie along this reference line if the two sets are drawn from a population with the same distribution.
4. Importance of the Q-Q plot: It is frequently desirable to determine whether the presumption of a common distribution is supported when there are two data samples. If so, location and scale estimators can combine the two sets of data to derive estimates for the shared position and scale. If two samples do differ, it is also helpful to comprehend the variations. More information about the nature of the difference can be gleaned from the q-q plot than from analytical techniques like the chi-square and Kolmogorov-Smirnov 2-sample tests.