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)

The nicer the weather, the more people rent bicycles.

With temperature, humidity, and wind speed, we can predict the weather. And we can predict number of people rent bike.!

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

A: Because a variable with n levels can be represented by n-1 dummy variables. It is important in order to achieve n-1 dummy variables as it can be used to delete extra column while creating dummy variables.

Example: Suppose you have a categorical feature "Color" with three categories: Red, Green, and Blue. If you create dummy variables without dropping the first one, you would have two dummy variables, say "Red" and "Green." In this case, the coefficient for "Red" would represent the change in the dependent variable compared to "Green." This might be less intuitive than having "Green" as the reference category, where the coefficient for "Green" would directly represent the change compared to "Red."

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

It adjusted temperature.!

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

We use VIF.

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

Temperature is the most impacts positively.

light snow/rain is the most impacts negatively.

Year impacts positively.

General Subjective Questions

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

Linear regression is typically used when you want to predict a continuous numeric outcome (the dependent variable) based on one or more input features (independent variables).

We need to build model with model equation. Tranning model with the goal isto estimate the coefficients (β_0, β_1, β_2, …) that minimize the sum of squared residuals (the vertical distances between the actual and predicted values).

After that we check the model and make predictions.

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

Suppose you are analyzing monthly sales data for a small online retail business that sells four different products: A, B, C, and D. You have recorded the number of units sold (Sales) for each product over the course of a year. Here are the summary statistics for each product:

	Mean Sales	Variance	Correlation with Month
Product A:	35 units	40	0.816
Product B:	35 units	40	0.816
Product C:	35 units	40	0.816
Product D:	35 units	40	0.816

At first glance, the summary statistics for all four products are identical. They all have the same mean sales, the same variance, and a high positive correlation with the month. Based on these summary statistics, one might conclude that the sales patterns for all four products are essentially the same.

However, when you visualize the data by creating line plots for each product's monthly sales over the year, you discover a striking difference. In this example, Anscombe's quartet principle applies because, despite having the same summary statistics, each product's sales pattern is significantly different. These differences have important implications for business strategy.

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

Pearson's correlation coefficient, often denoted as "r" or Pearson's "r," is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It's a widely used statistic in statistics and data analysis for assessing the degree to which two variables are associated.

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

Scaling in the context of data preprocessing refers to the process of transforming the values of features (variables) in a dataset to a specific range or distribution. Scaling is performed to ensure that different features are on a similar scale or have similar units, which can be crucial for various machine learning algorithms and data analysis techniques. Scaling helps to avoid problems related to the magnitude of features, which can otherwise lead to biased or inefficient model training and affect the performance of some algorithms.

Normalized scaling scales data to a specific range, typically between 0 and 1.

Standardized scaling centers data around a mean of 0 with a standard deviation of 1.

Normalized scaling is suitable when you want to maintain the relative relationships between data points and features while ensuring a consistent scale.

Standardized scaling is useful when you want to standardize features to have similar scales and facilitate comparisons between them.

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

In the context of categorical variables and dummy variable encoding, if you include all dummy variables for a categorical variable (i.e., without dropping one as a reference category), it can lead to perfect multicollinearity because the omitted category can be perfectly predicted from the others.

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

A Quantile-Quantile (Q-Q) plot is a graphical tool used in statistics and data analysis to assess whether a dataset follows a particular theoretical distribution, such as the normal distribution. It provides a visual comparison between the observed data and the expected values from a specified theoretical distribution.

A Q-Q plot can be used to visually assess whether the residuals (differences between observed and predicted values) are normally distributed. If the points on the Q-Q plot closely follow a straight line, it suggests that the residuals are normally distributed. Deviations from a straight line indicate departures from normality.