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 categorical variables we had were 'season' and 'weathersit'. These were converted to numerical values by using dummy variables.

   - If we take a look at their p-values, it seems to be 0.00 for 'summer', 'spring', 'winter', 'light_rain' and 'misty'. This shows us that the categorical variables had a statistically significant effect on the dependent variable.
   - Considering coefficient magnitude, all of them seem to have negative impact. The impact is more for 'spring' and 'light_rain', whereas the others also do have some negative impact. Note: light_rain here corresponds to weathersit value of 3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds).

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

   Using drop_first=True during dummy variable creation is important to avoid the "dummy variable trap," which is a situation of multicollinearity that can affect the performance and interpretability of regression models.

   The dummy variable trap arises when one variable can be perfectly predicted from the others. In the context of categorical variables, if you include dummy variables for all categories of a categorical variable, it creates perfect multicollinearity because the sum of the dummy variables will always be 1.

   By using drop_first=True, you omit one of the dummy variables for each categorical variable, effectively eliminating the multicollinearity issue. Dropping one dummy variable ensures that the variables are linearly independent, which is a requirement for regression models to work properly.

   Thus, using drop_first=True during dummy variable creation helps prevent multicollinearity and ensures that your regression model works effectively and provides meaningful insights.

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

   Based on the pair plot, it looks like 'yr', 'holiday', 'workingday' have higher correlation with target variable. Also, the same case for the categorical variables.

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

   There are various ways in order to validate the assumptions here:

   - Linearity: plotting the observed values vs predicted values
   - Normality of Residuals: Creating a histogram or a Q-Q plot of the model's residuals
   - Multicollinearity: Calculating the Variance Inflation Factor (VIF) for predictor variables. VIF values greater than around 5 suggests multicollinearity.

- Outliers: Analysing the residuals to identify outliers.
- R-squared and Adjusted R-squared: these would help to understand how well the model explains the variability in the data.

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

   - The top 3 factors here would be 'yr', 'weekday' and 'workingday'.

General Subjective Questions

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

   - Linear regression is a method to predict numerical outcomes using input features. It assumes a linear relationship between variables and finds a line that best fits data points.
   - Linear regression is a way to find a straight line that closely follows a set of data points. This line helps us understand how two things are connected and enables us to make predictions. The process involves these steps:
     - **Equation Setup**: We start by setting up a simple equation to predict one thing based on another. This equation is like a template that we'll adjust to fit our data. It's something like: Prediction = Starting Point + Change * Input + Unpredictable Part ($y = B0 + B1x + e$).
     - **Optimization**: Our goal is to make our equation predict as accurately as possible. To do this, we use some mathematical tricks to find the best values for the Starting Point and Change. These values make the line come as close as possible to our data points.
     - **Fitting the Line**: Imagine drawing a straight line through a scatterplot of dots. We move the line up or down and tilt it a bit to fit the dots well. This line represents our prediction equation and shows the relationship between the two things we're looking at.
     - **Making Predictions**: Now that we have the line, we can use it to make predictions. If we have a new value for the input, we can plug it into our equation and get an estimated prediction. It's like using the line to guess what might happen based on what we've seen before.
     - **Evaluation**: We want to know how good our line is at making predictions. R-squared is a measure that tells us how well the line fits the data. A higher R-squared means our line is doing a good job of capturing the relationship.
   - In summary, linear regression helps us find a straight line that helps us predict one thing based on another. We start with a basic equation, adjust it to fit the data, draw the line through our points, use the line to guess new outcomes, and check how well it works with a statistical measure. This process simplifies complex relationships and helps us make informed predictions.

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

- Anscombe's Quartet is a set of four small groups of data points that seem quite alike when we calculate simple things like averages and totals. However, the real eye-opener comes when we put them on graphs.
    - **Straight Line Story**: Imagine you have points that are in a perfect line when plotted on a graph. This means when one thing goes up, the other goes up too, following a simple pattern. It's like saying if you eat more ice cream, the temperature goes up – a clear link.
    - **Surprising Outlier**: Now, think about having similar points in a line, but then there's one point that's really different from the rest. It's like a person in a group doing something totally unexpected. This single point can mess up the whole line on the graph, changing how things seem to be connected.
    - **Troublesome Outlier**: In another set of points that almost match, there's one strange point that's way off. It's like everyone agrees except for that one person who has a totally different idea. This unusual point can pull the line on the graph in a different direction.
    - **No Line Here**: Lastly, imagine you have points that don't follow a line at all when you put them on a graph. It's like trying to draw a straight line through a crowd of people going in different directions – there's no simple pattern to see.
- The big lesson from all of this is that only looking at numbers can hide interesting and surprising stuff. Graphs help us see the real story hidden in the data, showing how things are connected in ways that numbers alone might not show.

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

Pearson's R is a number that tells us how two sets of numbers are related. It shows if they move together (positive relationship), move in opposite directions (negative relationship), or don't seem to move together at all (no relationship). It's like a measure of how close friends two sets of numbers are. If R is close to 1 or -1, they're really good friends in terms of moving together, and if it's close to 0, they're not so close.

- A positive R (close to 1) indicates a strong positive correlation: As one set of numbers increases, the other tends to increase as well.
- A negative R (close to -1) indicates a strong negative correlation: As one set of numbers increases, the other tends to decrease.
- An R close to 0 suggests a weak or no linear correlation: The two sets of numbers don't show a clear linear relationship.

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

Scaling is a process in data preprocessing where we adjust the range or distribution of our data. It's like putting our data on a common scale to make comparisons and calculations easier.

Why Scaling is Performed:

Equal Treatment: Scaling ensures all variables have the same influence when analysing data. Variables with larger ranges might dominate others in calculations.

Algorithm Sensitivity: Many machine learning algorithms work better when data is scaled. Scaling helps them converge faster and gives equal importance to all features.

Interpretability: Scaled data is easier to visualize and interpret, as it's on a common scale.

Normalized Scaling vs. Standardized Scaling:

Normalized Scaling:

- Here, we scale the data to a specific range, often between 0 and 1.
- It's useful when we want all data to fit within a certain interval, making it easier to compare.
- Formula: x_normalized = (x - min) / (max - min)

Standardized Scaling:

- In this method, we transform the data to have a mean of 0 and a standard deviation of 1.
- It's beneficial for algorithms that assume normal distribution or for features with different units.
- Formula: x_standardized = (x - mean) / standard deviation

In summary, scaling is the process of adjusting data to a common scale for better analysis. Normalized scaling rescales data to a specific range (often 0 to 1), while standardized scaling centers data around mean 0 and standard deviation 1. Both techniques make data more suitable for analysis and modelling.

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

VIF, or Variance Inflation Factor, is a measure used to assess multicollinearity in a multiple regression analysis. It quantifies how much the variance of the estimated regression coefficient is increased due to collinearity among predictor variables. A high VIF indicates that the predictor variables are highly correlated, which can affect the reliability of the regression model's results.

The situation where the VIF value becomes infinite can occur due to a specific type of multicollinearity known as "perfect multicollinearity." This happens when one predictor variable in the regression model can be perfectly predicted from a combination of other predictor variables. In other words, there is an exact linear relationship among some of the variables.

When perfect multicollinearity occurs, one predictor variable can be expressed as a linear combination of others. This creates an issue in the calculations for VIF, leading to division by

zero in the formula used to compute VIF. As a result, the VIF for the predictor variable involved in the linear relationship becomes infinite.

In practical terms, perfect multicollinearity can arise due to various reasons, such as:

- **Data Errors**: Incorrect data entry or measurement errors can introduce perfect multicollinearity.
- **Redundant Variables**: Including redundant variables in the regression model, such as using the same variable in different units (e.g., using both inches and centimetres for height).
- **Dummy Variable Trap**: When creating dummy variables, one category may be perfectly predicted from the other categories, leading to perfect multicollinearity.
- **Data Manipulation**: Aggregating or transforming variables can accidentally create perfect multicollinearity.

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

(3 marks)

A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess whether a dataset follows a specific theoretical distribution, such as a normal distribution. It compares the quantiles of the dataset's values with the quantiles of the chosen theoretical distribution. If the points on the Q-Q plot roughly follow a straight line, it suggests that the data follows the chosen distribution.

A Q-Q plot plays a significant role in linear regression:

- **Checking Assumptions**: Linear regression assumes that the errors (residuals) follow a normal distribution. By creating a Q-Q plot of the residuals, we can quickly assess whether this assumption holds. If the points on the Q-Q plot deviate from a straight line, it may indicate that the assumption of normality is violated.
- **Identifying Outliers**: Outliers can affect the normality of residuals. A Q-Q plot can help identify outliers by showing if the data points deviate from the theoretical distribution's line. Outliers might cause the line to bend or tilt.
- **Model Evaluation**: A well-behaved Q-Q plot indicates that the linear regression model is a good fit for the data. If the Q-Q plot deviates significantly from a straight line, it suggests that the model might not capture the underlying distribution of the data accurately.
- **Adjusting the Model**: If the Q-Q plot reveals deviations from normality, it signals the need for further investigation. You might consider transforming the data or exploring more complex models to improve the fit and meet the assumptions.

In summary, a Q-Q plot is a powerful visual tool that helps assess the fit of the linear regression model by examining the distribution of residuals. It aids in checking assumptions, identifying outliers, evaluating model performance, and guiding potential adjustments to enhance the model's accuracy and reliability.