Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 1 goes below this line> (Do not edit)

1. Bike demand in year 2019 is higher as compared to 2018.
2. Bike demand is high in the months from May to October.
3. Bike demand in the fall is the highest.
4. Bike demand takes a dip in spring.
5. Bike demand is high if weather is clear or with mist cloudy while it is low when
6. there is light rain or light snow.
7. The demand of bike is almost similar throughout the weekdays.
8. Bike demand doesn't change whether day is working day or not.

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 2 goes below this line> (Do not edit)

drop_first=True helps in reducing the extra column created during the dummy variable creation and hence avoid redundancy of any kind.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)

The temp variable has the highest correlation with the target variable of 0.63 which is the highest among all the numerical variables.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

1. Linearity of relationship between response and predictor variables.
2. Normality of the error distribution (Normal distribution of error terms).
3. Constant variance of the errors or Homoscedasticity.
4. Less Multi-collinearity between features (Low VIF)

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

The top 3 features contributing significantly towards the demand of the shared bikes are the

1. Temp
2. Year
3. Light Snow

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Linear Regression is an ML algorithm used for supervised learning. It helps in predicting a dependent variable(target) based on the given independent variable(s). The regression technique tends to establish a linear relationship between a dependent variable and the other given independent variables. There are two types of linear regression- simple linear regression and multiple linear regression. Simple linear regression is used when a single independent variable is used to predict the value of the target variable. Multiple Linear Regression is when multiple independent variables are used to predict the numerical value of the target variable. A linear line showing the relationship between the dependent and independent variables is called a regression line. A positive linear relationship is when the dependent variable on the Y-axis along with the independent variable in the X-axis. However, if dependent variables value decreases with increase in independent variable value increase in X-axis, it is a negative linear relationship.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Anscombe's quartet consists of four data sets that have nearly identical simple descriptive statistics but have very different distributions and appear very different when presented graphically. Each dataset consists of eleven points. The primary purpose of Anscombe's quartet is to illustrate the importance of looking at a set of data graphically before beginning the analysis process as the statistics merely does not give the an accurate representation of two datasets being compared.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Pearson's R, also known as the **Pearson correlation coefficient**, is a statistical measure that evaluates the linear relationship between two continuous variables. It indicates both the strength and the direction of the correlation.

1. Understanding the relationship between variables, e.g., height and weight, test scores and study hours.
2. Feature selection in machine learning (checking correlations between variables).
3. Hypothesis testing to determine if two variables are significantly correlated.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Scaling is a preprocessing technique in data analysis and machine learning where the values of a dataset are adjusted to fit within a specific range or distribution. This is typically done to ensure that all features contribute equally to the model's performance, avoiding dominance by features with larger ranges.

Difference Between Normalized Scaling and Standardized Scaling

Feature	Normalization	Standardization
Definition	Rescales the values to a specific range, typically [0, 1].	Rescales values to have a mean of 0 and a standard deviation of 1.
Formula	$X_{norm} = (X - X_{min})/(X_{max} - X_{min})$	$X_{std} = (X - \mu)/\sigma$
Result	Values are within a fixed range (e.g., [0, 1]).	Values follow a standard normal distribution (mean = 0, std = 1).
When to Use	When features are not normally distributed and bound ranges are required (e.g., image processing).	When features are normally distributed or required for algorithms assuming Gaussian distribution.
Impact	Sensitive to outliers since it depends on min/max values.	Less sensitive to outliers but doesn't guarantee bounds.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

The **Variance Inflation Factor (VIF)** measures the extent of multicollinearity in a regression model. It quantifies how much the variance of a regression coefficient is inflated due to the correlation between independent variables.

VIF becomes infinite when there is **perfect multicollinearity**, meaning one predictor variable is a perfect linear combination of one or more other predictors in the model. This leads to the denominator of the VIF formula becoming zero, causing the value to approach infinity.

Formula for VIF:

$$\mathbf{VIF_i = 1/(1-R_i^2)}$$

Where:

1. R_i^2: Coefficient of determination of the regression of the iii-th predictor on all other predictors.
2. If $R_i^2=1$: Perfect multicollinearity exists (predictor can be exactly predicted by others), and $1-R_i^2=0$, making VIF_i infinite.

Causes of Infinite VIF:

1. **Duplicate Variables**: Exact copies of variables or near duplicates in the dataset.
2. **Linear Dependence**: One variable is a perfect linear combination of others
3. **Feature Encoding Issues**: Dummy variable traps in one-hot encoding, where categories are not properly reduced (e.g., including all categories without dropping one).

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

A **Q-Q plot** (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a dataset to a theoretical distribution (commonly a normal distribution). It helps assess whether the data follows a particular distribution by plotting the quantiles of the data against the quantiles of the theoretical distribution.

Structure of a Q-Q Plot

1. **X-axis**: Theoretical quantiles (from the specified theoretical distribution, e.g., normal distribution).
2. **Y-axis**: Sample quantiles (from the observed data).

If the data aligns closely with the theoretical distribution, the points in the Q-Q plot will roughly form a straight diagonal line. Deviations from this line indicate departures from the specified distribution.

Use and Importance of Q-Q Plots in Linear Regression

In linear regression, Q-Q plots are essential for verifying the **normality assumption** of the residuals. Many statistical tests and model performance metrics in linear regression rely on this assumption.

1. **Assumption Testing**:
    - **Normality of Residuals**: The residuals of the regression model should be approximately normally distributed. The Q-Q plot helps visualize whether this assumption holds.
    - Points forming a straight line indicate normality.
    - Deviations, such as an S-shaped curve, indicate skewness or heavy tails (non-normal distribution).
2. **Outlier Detection**: Points far from the diagonal line in the Q-Q plot may represent outliers or extreme values that could affect the regression model.
3. **Model Diagnostics**: If residuals deviate from normality, the model may not perform well in terms of inference (e.g., p-values, confidence intervals) or predictions.
4. **Improving Model**: Based on the Q-Q plot, transformations (e.g., log, square root) or alternative models (e.g., generalized linear models) can be used to better fit the data.