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?**

   Inferences suggest that when the weather is snowy or misty, the demand for bike-sharing decreases.
   This implies that unfavorable weather conditions, such as snow or mist, may deter individuals from opting for shared bikes, leading to a decline in demand.
   Clear Weather is the Best for High Demands:
   This suggests that when the weather is clear, people are more inclined to use shared bikes, potentially due to the pleasant conditions and increased willingness to engage in outdoor activities.
   The inference mentions the absence of recorded data for rainy weather.
   This could be due to the unavailability of data during rainy conditions or a lack of bike-sharing activity during such weather.

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

   Using drop_first=True during dummy variable creation in pandas is crucial because it helps avoid multicollinearity issues in regression models. By dropping the first level of each categorical variable turned into dummy variables, you prevent perfect multicollinearity, which can cause problems in interpreting regression coefficients and model performance. In short, it enhances the accuracy and reliability of your regression analysis.

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

   There is a strong correlation or dependence between the temp (actual temperature) and atemp (feeling-like temperature) columns.
   The two temperature-related variables seem to exhibit a close relationship, indicating that changes in one variable are associated with corresponding changes in the other.

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

   I will do Residual Analysis to examine the residuals (the differences between predicted and actual values) and also will verify whether the residuals are normally distributed using histogram.
   Scatter plots of residuals against predicted values or independent variables is use to Ensure homoscedasticity and linearity.
   Multicolinearity is verified using variance inflation factors (VIF).

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

   Year (yr):

Coefficient: 0.24
Positive coefficient suggests that as the year increases, the demand for shared bikes tends to increase.

Summer Season (season_summer):
Coefficient: 0.04
Positive coefficient indicates that during the summer season, there is an increase in demand for shared bikes.

Winter Season (season_winter):
Coefficient: 0.08
Positive coefficient suggests that during the winter season, there is an increase in demand for shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression is a supervised machine learning algorithm used for predicting a continuous outcome variable (also called the dependent variable) based on one or more predictor variables (independent variables) that are assumed to have a linear relationship with the outcome. The goal of Linear Regression is to find the best-fitting straight line (hyperplane in higher dimensions) that minimizes the sum of squared differences between the observed and predicted values. Here's a detailed explanation of the Linear Regression algorithm:

Assumptions:
Linearity: Assumes a linear relationship between independent and dependent variables.
Independence: Assumes independence of observations.
Homoscedasticity: Assumes constant variance of errors.
Normality of Residuals: Assumes residuals are normally distributed.
No Multicollinearity: Assumes little or no multicollinearity among independent variables.

Simple Linear Regression:
In the case of a single predictor variable (X), the simple linear regression equation is

$y=\beta 0+\beta 1 \cdot X+\varepsilon$,

where:
y is the dependent variable,
$\beta 0$is the intercept,
$\beta 1$ is the slope, and
ε is the error term.
3. Multiple Linear Regression:
Extends simple linear regression to multiple predictors:
$y=\beta 0+\beta 1 \cdot x 1+\beta 2 \cdot x 2+\ldots+\beta n \cdot x n+\varepsilon$, where

$x 1, x 2, \ldots, x n$ are the predictor variables.

4. Model Training:
The model is trained using a dataset with known values of the dependent variable.
The coefficients ($\beta 0, \beta 1, \ldots, \beta n$) are estimated using methods like Ordinary Least Squares (OLS).

Cost Function:
The goal is to minimize the cost function, which is the sum of squared differences between the predicted and actual values.
Gradient Descent :
An optimization algorithm that iteratively adjusts the model parameters to minimize the cost function. Optional for simple linear regression, more commonly used for multiple linear regression with a large dataset.
Model Evaluation:
Model performance is assessed using metrics like R-squared, Mean Squared Error (MSE), or Mean Absolute Error (MAE).
Interpretation of Coefficients:
Coefficients ($\beta 0, \beta 1, \ldots, \beta n$) represent the change in the dependent variable for a one-unit change in the corresponding predictor variable, holding other variables constant.
Prediction:
Once trained, the model can be used to make predictions on new, unseen data.
Regularization :
In cases of overfitting, regularization techniques like Ridge or Lasso regression can be applied to penalize large coefficients.
Applications:
Commonly used in economics, finance, biology, and various other fields for predictive modeling and analysis.
Limitations:
Assumes a linear relationship, may not capture complex nonlinear patterns.
Sensitive to outliers.
Extensions:
Polynomial Regression: Allows modeling nonlinear relationships by introducing polynomial terms.
Multiple Algorithms: Other regression algorithms like Ridge, Lasso, and Elastic Net address some limitations of simple linear regression.
Linear Regression, while simple, is a powerful tool for modeling and predicting quantitative relationships in various domains. Its simplicity and interpretability make it a popular choice for regression tasks.

2. **Explain the Anscombe's quartet in detail.**

   Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet they have very different distributions and appear quite distinct when graphed. This quartet was created by the statistician Francis Anscombe in 1973 to illustrate the importance of visualizing data and the limitations of relying solely on summary statistics.

   The four datasets in Anscombe's quartet have the same means, variances, correlations, and linear regression lines, but they differ significantly in their distributions. Here are the characteristics of each dataset:

   Dataset I:

Mean of x: 9.0
Mean of y: 7.5
Variance of x: 11.0
Variance of y: 4.12
Correlation between x and y: 0.816
Linear regression line: =3.00+0.50
y=3.00+0.50x

Dataset II:

Mean of x: 9.0
Mean of y: 7.5
Variance of x: 11.0
Variance of y: 4.12
Correlation between x and y: 0.816
Linear regression line: =3.00+0.50
y=3.00+0.50x

Dataset III:

Mean of x: 9.0
Mean of y: 7.5
Variance of x: 11.0
Variance of y: 4.12
Correlation between x and y: 0.816
Linear regression line: =3.00+0.50
y=3.00+0.50x

Dataset IV:

Mean of x: 8.0
Mean of y: 7.5
Variance of x: 11.0
Variance of y: 4.12
Correlation between x and y: 0.817
Linear regression line: =3.00+0.50
y=3.00+0.50x

The lesson from Anscombe's quartet is that while summary statistics may provide a quick overview of the data, they can be insufficient to understand the underlying patterns or relationships. Visualizing data through graphs and plots is crucial for gaining insights and making informed interpretations. It emphasizes the importance of exploratory data analysis and graphical representation in addition to numerical summaries.

3. What is Pearson’s R?

Pearson's correlation coefficient, denoted as r, is a statistical measure that quantifies the degree to which two variables are linearly related.

r ranges from -1 to 1.
If =1 r=1: Perfect positive linear relationship.
If =−1 r=−1: Perfect negative linear relationship.
If =0
r=0: No linear relationship.

- The sign of r indicates the direction of the relationship:
- Positive r: As one variable increases, the other tends to increase.
- Negative r: As one variable increases, the other tends to decrease.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

   Scaling, in the context of data analysis, refers to the process of transforming numerical variables to a standard range or distribution. The purpose of scaling is to ensure that variables with different scales and units are comparable and do not disproportionately influence the analysis. Scaling is particularly important in algorithms that are sensitive to the magnitude of variables, such as distance-based algorithms or optimization algorithms.

   Why Scaling is Performed:

   Comparison: Scaling allows for a fair comparison between variables with different units. Without scaling, variables with larger magnitudes might dominate the analysis.
   Convergence: In optimization algorithms, scaling can help improve the convergence speed, preventing the algorithm from taking longer to converge due to differences in variable scales.
   Distance Metrics: Algorithms that rely on distance measures, such as k-nearest neighbors or k-means clustering, can be affected by the scale of variables. Scaling helps in creating a level playing field for these algorithms.

   Normalized Scaling vs. Standardized Scaling:

Normalized Scaling	Standardized Scaling
Normalized scaling scales values between 0 and 1.	Standardized scaling centers values around 0 with a standard deviation of 1.
Normalized scaling can be sensitive to outliers.	Standardized scaling is more robust to outliers because it relies on the mean and standard deviation.
Normalized scaling is often used when the distribution of the data is not assumed to be Gaussian.	Standardized scaling is preferred when the data is expected to be normally distributed or when the algorithm requires standardized input.

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

   In practical terms, an infinite VIF indicates that the variance of the estimated regression coefficient for that particular variable is extremely large, making it difficult to assess its individual contribution to the model. An infinite VIF is a clear sign of severe multicollinearity.

When dealing with infinite VIF values, it's crucial to address the multicollinearity issue by either removing one or more of the highly correlated variables or by employing other techniques such as variable transformation or regularization methods (like ridge regression) to mitigate the multicollinearity problem. Addressing multicollinearity is important for obtaining stable and reliable regression coefficients and avoiding inflated standard errors in the model.

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

   A Q-Q (Quantile-Quantile) plot is a graphical tool used in statistics to assess whether a dataset follows a particular theoretical distribution, such as the normal distribution. It compares the quantiles of the observed data against the quantiles of the expected distribution. In the context of linear regression, Q-Q plots are often used to check the assumption of normality of the residuals.