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?*
   - That winter and summer season has positive coefficient whereas spring has a negative coefficient.
   - Both challenging and favourable (derived features) have negative coefficient.
     - challenging - Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
     - Favourable - Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist.
   - Year can be seen having a positive coefficient hence positive correlation with dependent variable.
   - Holiday has a negative correlation with dependent variable.

2. *Why is it important to use drop_first=True during dummy variable creation?*
   - For a feature if you have 'N' categories, you only need 'N-1' dummy variables to represent them. By setting ‘drop_first=True’ during dummy variable creation, one dummy variable is omitted from each set of dummy variables created for a categorical variable. This eliminates perfect multicollinearity among the dummy variables.

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

4. *How did you validate the assumptions of Linear Regression after building the model on the training set?*
   - The residuals are assumed to be normally distributed , by plotting a distplot on residuals.
   - The relationship between the independent variables and the dependent variable is linear as observed by model on test data.

- All residuals are independent of each other , no pattern is observed and is constant variance as observed by plotting a Scatter plot on residuals.
- No multicollinearity as observed by VIF < 5.

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

- temp
- yr
- Challenging_weather - Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

Contd….

General Subjective Questions :

1. *Explain the linear regression algorithm in detail.*

Linear regression is a supervised learning algorithm used for predicting a continuous outcome variable (also called the dependent variable) based on one or mor predictor variables (independent variables). It assumes a linear relationship between the predictor variables and the target variable. The goal of linear regression to find the best-fit line that minimizes the sum of squared differences between the predicted and actual values.

Here's a detailed explanation of the linear regression algorithm:

Model Representation:

1. Simple Linear Regression:
   - In simple linear regression, there is one independent variable (X) and one dependent variable (Y). The relationship between them is modeled as a straight line:

   $Y=\beta0+\beta1X+\varepsilon$

   where:

   - Y is the dependent variable.
   - X is the independent variable.
   - β0 is the intercept (y-intercept), representing the value of Y when X is 0.
   - β1 is the slope, representing the change in Y for a one-unit change in X.
   - ε is the error term representing the deviation of the actual Y from the predicted Y.

2. Multiple Linear Regression:

- In multiple linear regression, there are multiple independent variables X1,X2,....,Xn and one dependent variable Y:

  $Y=\beta0 + \beta1X1 + \beta2X2 + \beta3X3 + \betanXn + \varepsilon$

  The model has n predictors with associated coefficients β0,β1, β2, β3 ... βn.

$$Y=\beta0 + \beta1X1 + \beta2X2 + \beta3X3 + \betanXn + \varepsilon$$

Objective Function:

The objective in linear regression is to find the values of β1, β2, β3 ... βn that minimize the sum of squared differences between the predicted and actual values. This is known as the Ordinary Least Squares (OLS) method.

Minimize: $\sum (m,i=1) (Yi - (\beta0 + \beta1Xi1 + \beta2Xi2 + \beta3Xi3 + ...+ \betanXin)) ^ 2$

Optimization:

The minimization problem is often solved using optimization algorithms, such as gradient descent. The goal is to update the coefficients iteratively to reach the values that minimize the cost function.

Assumptions:

Linear regression relies on several assumptions:

1. Linearity: The relationship between predictors and the response variable is linear.
2. Independence of Errors: Residuals (errors) are independent of each other.
3. Homoscedasticity: Residuals have constant variance across all levels of predictors.
4. Normality of Errors: Residuals are normally distributed.
5. No Perfect Multicollinearity: Independent variables are not perfectly correlated.

Evaluation:

The performance of the linear regression model is often evaluated using metrics such as Mean Squared Error (MSE), R-squared (\(R^2\)), and others, depending on the problem.

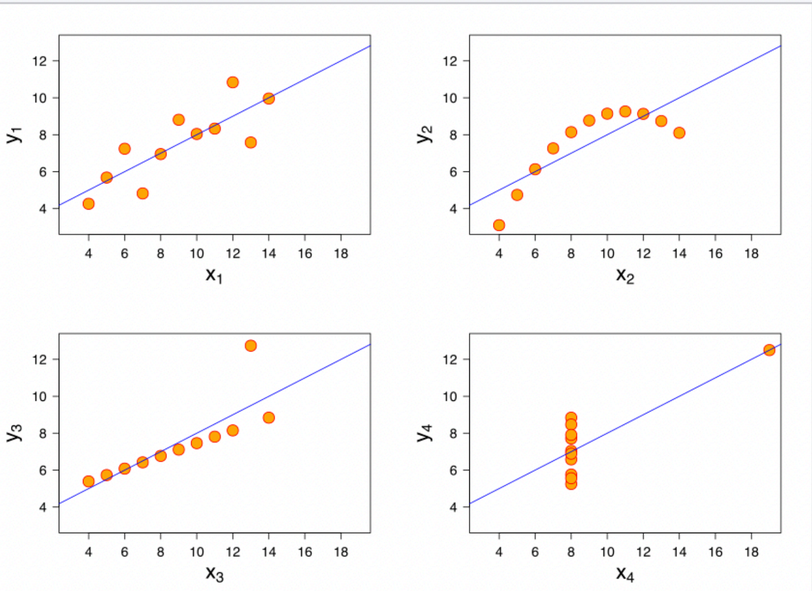
Implementation:

In Python, this is implemented using scikit-learn and statsmodels.api

This is a high-level overview of linear regression.

2. Explain the Anscombe’s quartet in detail.

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very differen distributions and appear very different when graphed. Each dataset consists of eleven (x, y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough”



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

For all 4 data sets :

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x: s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y: s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression: R^2	0.67	to 2 decimal places

- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two correlated variables, where y could be modelled as gaussian with mean linearly dependent on x.
- For the second graph (top right), while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier, which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

3. What is Pearson’s R?

Pearson's correlation coefficient, often denoted as r, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to 1, where:

- $r = 1$ indicates a perfect positive linear relationship.
- $r = -1$ indicates a perfect negative linear relationship.
- $r = 0$ indicates no linear relationship.

The formula for Pearson's correlation coefficient between two variables X and Y in a dataset of n pairs of observations is given by:

$$r = (\sum (Xi - X')(Y - Y')) / ((\sum (Xi - X')^2 * \sum (Yi - Y')^2)^{0.5})$$

where:

- Xi and Yi are the individual data points.
- X' and Y' are the means of X and Y, respectively.

Pearson's correlation coefficient measures the linear relationship between two variables. A positive r indicates a positive correlation (as one variable increases, the other tends to increase) , while a negative r indicates a negative correlation (as one variable increases, the other tends to decrease).

It's important to note that Pearson's correlation coefficient assumes that the relationship between the variables is linear and that the data is normally distributed. Additionally, outlier can strongly influence the correlation coefficient, so it's always a good practice to complement correlation analysis with visual inspection of the data and consider other types of correlation coefficients or non-linear relationships if needed.

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

Scaling in the context of data preprocessing refers to the process of transforming variables to a standardized range. This is done to ensure that no variable has more influence than another in the analysis. Scaling is especially important when dealing with algorithms that are sensitive to the scale of input features, such a distance-based methods or optimization algorithms.

Why is Scaling Performed?

1. Algorithm Sensitivity:
   - Some machine learning algorithms are sensitive to the scale of input features. For example, distance-based algorithms like k-nearest neighbors (KNN) can b heavily influenced by the scale of variables.

2. Convergence of Optimization Algorithms:
   - Optimization algorithms, such as gradient descent, converge faster when variables are on similar scales. This helps in achieving better performance and fas convergence.

3. Interpretability:
   - Scaling can improve the interpretability of coefficients in linear models. It ensures that coefficients represent the change in the dependent variable for a one- unit change in the respective independent variable.

Types of Scaling:

Range:

1. Normalized Scaling (Min-Max Scaling):
   - Formula:
     X normalised = (X - Xmin) / (Xmax - Xmin)
   - Range:
     - Scales the variable to a range between 0 and 1.
   - Advantages:
     - Preserves the shape of the original distribution.
   - Considerations:
     - Sensitive to outliers.

2. Standardized Scaling (Z-Score Scaling):
   - Formula:
     X standardised = (X - μ) / σ
   - Range:
     - Scales the variable to have a mean (μ) of 0 and a standard deviation (σ) of 1.
   - Advantages:
     - Less sensitive to outliers compared to Min-Max Scaling.
   - Considerations:
     - Preserves the shape of the original distribution but doesn't bound the variable within a specific range.

Key Differences:

1. Range:
   - Normalized Scaling: Scales variables to a specific range (e.g., 0 to 1).
   - Standardized Scaling: Centers variables around a mean of 0 with a standard deviation of 1.

2. Sensitivity to Outliers:
   - Normalized Scaling: Sensitive to outliers.
   - Standardized Scaling: Less sensitive to outliers.

3. Interpretability:
   - Normalized Scaling: Preserves the original distribution and the interpretability of original values.
   - Standardized Scaling: Preserves the shape of the distribution but doesn't maintain original values' interpretability.

The choice between normalized and standardized scaling depends on the specific requirements of the analysis and the characteristics of the data. Each has its advantages and considerations, and the decision may also be influenced by the characteristics of the machine learning algorithm being used.

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

Reasons for Infinite VIF:

1. Perfect Multicollinearity:

- Infinite VIF occurs when one predictor can be expressed as a perfect linear combination of other predictors. This means th
is a mathematical relationship (exact correlation) among the predictors, making the design matrix singular and causing issues
estimating coefficients.

2. Data Redundancy:

- If one predictor is a constant multiple of another (e.g., \(X_1 = 2X_2\)), there is a perfect linear relationship, leading to infinit
VIF.

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

A Quantile-Quantile (Q-Q) plot is a graphical tool used in statistics to assess whether a dataset follows a particular theoretical
distribution. It is particularly useful for checking the assumption of normality in linear regression or other statistical analyses.

Use and Importance in Linear Regression:

1.Normality Assessment:

- In linear regression, the normality assumption is crucial for valid statistical inferences, hypothesis testing, and constructing accurate confidence intervals. Q-Q plots help assess whether the residuals (differences between observed and predicted values) are normally distributed.

2.Residual Analysis:

- A common use of Q-Q plots is to examine the distribution of residuals. If the residuals are normally distributed, the Q-Q plot should exhibit a straight line. Deviations from a straight line may indicate non-normality of residuals.

3.Identification of Outliers:

- Q-Q plots can help identify outliers and deviations from normality more effectively than histograms or other graphical methods.

4.Comparison of Distributions:

- Q-Q plots allow a direct visual comparison between the observed distribution and a theoretical distribution. This can be useful when assessing whether a non-parametric distribution (e.g., the t-distribution) might be a better fit for the data.