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?

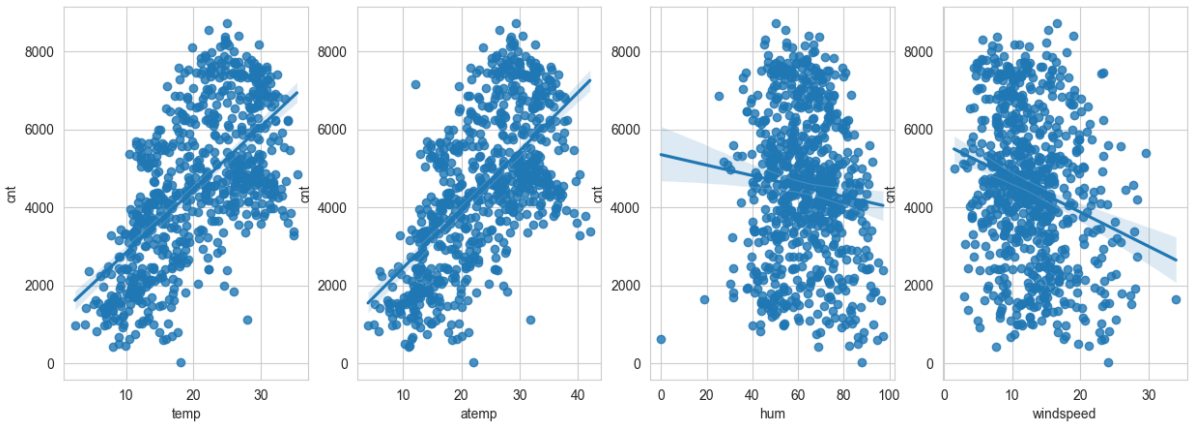
   (3 marks)

   Ans: Based on the analysis of the categorical variables, the following inferences can be made about their effect on bike usage:

   Bike usage is notably higher on working days, especially on Mondays. This suggests that bikes are used more frequently for commuting purposes, likely to and from workplaces or schools. Bike usage tends to increase in September and tends to fall from November to February, with better performance anticipated during the summer and winter months. The pattern indicates that bike usage is generally higher during the summer, likely due to more favorable weather conditions and increased outdoor activity. During high thunderstorms, rain, or snow, bike usage decreases significantly. Increases in humidity and windspeed are associated with a decrease in the dependent variable. High humidity can make cycling uncomfortable, while strong winds can make it physically challenging and less enjoyable. 2019 year correlates with an increase in the demand for bikes, indicating a strong positive trend over time.

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
   Ans : Using drop_first=True when creating dummy variables is essential for several reasons. It helps avoid multicollinearity by excluding one dummy variable, typically the first category, which serves as a reference. This approach prevents redundancy and ensures that the model remains stable and performs optimally. By dropping one dummy variable, you avoid including unnecessary variables that can distort model coefficients and impact performance, leading to clearer interpretation and more reliable results.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   Ans: ‘temp’ and ‘atemp’ variable has the highest correlation.



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

   Ans: Calculating the VIF for each independent variable. VIF values usually less than or equal to 5 suggest that multicollinearity is not a concern, indicating that the independent variables are not highly correlated with each other. Also, the p-values should be < 0.05. The R-squared and Adjusted R-Squared should be close to each other with not much difference. Also, the R-squared for training and test data sets should also be close. The residuals should follow a normal distribution with mean 0. Residuals should be randomly scattered around zero without forming any discernible patterns, indicating constant variance.

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

   Ans: Temperature, Humidity and Windspeed are the top 3 factors contributing significantly towards the demand of the shared bikes. Increases in humidity and windspeed are associated with a decrease in the demand. This suggests that higher humidity and windspeed negatively impact bike usage. A higher temperature has a substantial positive effect on bike usage. As temperatures rise, people are more inclined to use bikes, highlighting the importance of favorable weather for increasing bike demand.

General Subjective Questions

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

   Ans: Linear regression aims to model the relationship between a dependent variable Y and one or more independent variables (predictors) X.

   The simplest form is simple linear regression, where there is one independent variable.

   $Y=\beta0+\beta1X+\epsilon$

   When there are multiple independent variables, it's called multiple linear regression.

   $Y=\beta0+\beta1X1+\beta2X2+\cdots+\beta pXp+\epsilon$

   where X1,X2,...Xp are the independent variables, and $\beta1,\beta2,\ldots,\beta p$ are the corresponding coefficients.

   The objective of linear regression is to find the best-fitting line through the data points. This line is determined by estimating the coefficients such that the line

minimizes the difference between the observed values and the values predicted by the line.

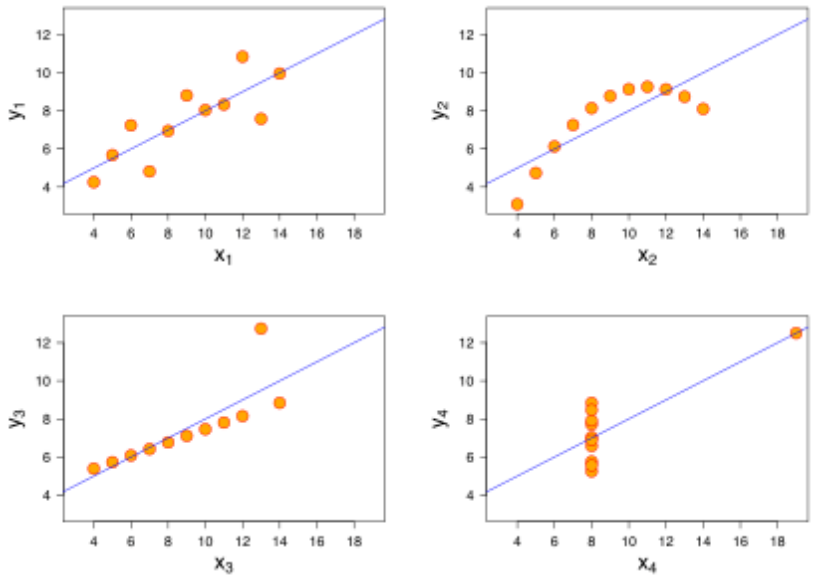
Assumptions of Linear Regression are:
- The relationship between the independent and dependent variables is linear.
- Error terms are independent of each other.
- Error terms have constant variance.
- The residuals (errors) are normally distributed with mean 0.

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

Ans: Anscombe's quartet is a set of four datasets that illustrate the importance of graphing data before analyzing it. Despite having nearly identical statistical properties, each dataset reveals a different pattern when visualized, highlighting how summary statistics alone can be misleading.

The Four Datasets

Dataset I: Shows a clear linear relationship.
Dataset II: Displays a linear trend with an influential outlier.
Dataset III: Features a non-linear, parabolic relationship.
Dataset IV: Contains a vertical cluster of points with a single outlier.



Anscombe's quartet serves as a powerful reminder that while statistical measures are useful, they do not always capture the underlying structure of the data. Visualization is a critical step in the data analysis process, allowing analysts to uncover patterns, anomalies, and insights that summary statistics alone might miss.

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

Ans: Pearson's R, also known as the Pearson correlation coefficient, is a measure of

the linear relationship between two continuous variables. It quantifies the strength and direction of the association between the variables.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

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

Strength:
0.1 to 0.3: Weak positive correlation.
0.3 to 0.5: Moderate positive correlation.
0.5 to 1: Strong positive correlation.
-0.1 to -0.3: Weak negative correlation.
-0.3 to -0.5: Moderate negative correlation.
-0.5 to -1: Strong negative correlation.

Direction: Indicates whether the relationship is positive (as one variable increases, the other does too) or negative (as one variable increases, the other decreases). Useful for determining how strongly two variables are related.

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

Ans: Scaling refers to the process of transforming data into a specific range or distribution. It is crucial in data preprocessing for machine learning and statistical analysis, ensuring that different features or variables contribute equally to the analysis or model. Scaling adjusts the range or distribution of data, often making it easier to compare or analyze.

Two common types of scaling are **normalization** and **standardization**:
Normalization (Min-Max Scaling):
Normalization is sensitive to outliers because the range is determined by the minimum and maximum values.

Transforms the data to a fixed range, usually [0, 1] or [-1, 1].

Xnorm = X−Xmin / Xmax−Xmin where X is the original value, Xmin is the minimum value in the feature, and Xmax is the maximum value.

Standardization (Z-Score Scaling):

Standardization is less sensitive to outliers, but the transformed data can have values beyond a specific range (not bounded).

Transforms the data to have a mean of 0 and a standard deviation of 1.

Xstd = X−μ / σ, where X is the original value, μ is the mean of the feature, and σ is the standard deviation.

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

(3 marks)

Ans: Variance Inflation Factor (VIF) measures how much the variance of an estimated regression coefficient is increased due to multicollinearity with other predictors. A VIF value quantifies how much the variance of a coefficient estimate is inflated compared to when predictors are uncorrelated.

An infinite VIF typically indicates perfect multicollinearity among the predictor variables. This means that one or more predictors are perfectly or nearly perfectly correlated with each other, making it impossible to distinguish their individual effects on the dependent variable.

How to Address Infinite VIF:

Determine which predictors are perfectly collinear and remove or combine them.

PCA can transform the predictors into a set of uncorrelated components, effectively addressing multicollinearity issues.

Techniques like Ridge Regression add a penalty term to the regression model, which can help mitigate the effects of multicollinearity.

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

(3 marks)

Ans: A **Q-Q plot** (Quantile-Quantile plot) is a graphical tool used to assess whether a dataset follows a theoretical distribution, typically the normal distribution. It compares the quantiles of the dataset against the quantiles of a specified theoretical distribution.

Quantiles: Quantiles are values that divide a dataset into intervals with equal probabilities.

Plotting:

X-axis: Quantiles of the theoretical distribution.

Y-axis: Quantiles of the dataset.

In a Q-Q plot, if the points lie approximately along a straight line (typically the 45-degree line), it indicates that the dataset follows the theoretical distribution closely.

Deviations from this line suggest departures from the theoretical distribution.

Importance of a Q-Q plot in linear regression:

In the context of linear regression, a Q-Q plot is primarily used to assess whether the residuals (errors) of the regression model are normally distributed

Many linear regression models assume that the residuals are normally distributed. This assumption is crucial for hypothesis testing and for constructing confidence intervals around the regression coefficients.

By plotting the quantiles of the residuals against the quantiles of a normal distribution, you can visually check if the residuals conform to a normal distribution.

If residuals are not normally distributed, it might signal issues such as omitted variables, incorrect functional form, or the need for a different model.

Example of Interpretation:

Straight Line: If the residuals' quantiles closely follow the 45-degree line, it indicates that the residuals are approximately normally distributed.

S-shaped Curve: A Q-Q plot with an S-shaped pattern suggests that the residuals might have heavier or lighter tails than a normal distribution.

Other Deviations: If the plot shows systematic deviations (e.g., bending away from the line), it suggests that the residuals do not follow the normal distribution, which could indicate model specification issues.