Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer aboutttheir effect on the dependent variable? (3 marks)**

→ Based on the analysis of categorical variables in the dataset, we can infer their effect on the dependent variable. For instance, we can look at how the season affects the bike rental count.

→ From the data analysis, it was found that the highest bike rental count was during the fall season, followed by summer, spring, and winter.

→ This suggests that the season variable has a significant effect on the dependent variable (bike rental count).

→ Similarly, we can also examine other categorical variables like weather, holiday, and working day to infer their effect on the bike rental count.

→ For example, it was found that the bike rental count was higher on non-working days and during clear weather.

→ Therefore, we can infer that the working day and weather variables also have a significant effect on the bike rental count.

→ Overall, by analyzing the categorical variables and their effect on the dependent variable, we can gain insights into which factors contribute to higher or lower bike rental counts, which can help bike-sharing companies make data-driven decisions to improve their services and increase profits.

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

→ It is important to use the drop_first=True parameter during dummy variable creation to avoid the "dummy variable trap". The dummy variable trap is a scenario where the inclusion of all dummy variables in a model results in a perfect collinearity (linear dependency) among the variables. This can lead to incorrect estimation of coefficients and inflated standard errors, which can impact the accuracy and reliability of the model.

→ Therefore, dropping one of the dummy variables (i.e., drop_first=True) ensures that we only include a set of independent dummy variables in the model. This helps to avoid the issue of multicollinearity and provides more accurate estimates of the coefficients.

→ It also helps to avoid redundant information and reduces the number of variables, which can make the model more interpretable and computationally efficient and faster.

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

→ Looking at the pair-plot among the numerical variables, we can identify the variable that has the highest correlation with the target variable by examining the scatterplot between each numerical variable and the target variable.

→ Assuming the target variable is the variable we are trying to predict in the bike sharing dataset, we can look at the pair-plot and identify the numerical variable that has the strongest linear relationship with the target variable. In this case, the target variable is **"cnt"**, which represents the total number of bike rentals.

→ Based on the pair-plot, it appears that the variable "registered" has the strongest correlation with the target variable "cnt". This is indicated by the scatterplot between "registered" and "cnt", which shows a strong linear relationship with most of the data points falling on or near a straight line.

→ This suggests that the number of registered users has a strong positive correlation with the total number of bike rentals.

→ Therefore, we can conclude that the variable "registered" has the highest correlation with the target variable "cnt" based on the pair-plot among the numerical variables.

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

→ To validate the assumptions of linear regression after building the model on the training set many techniques can be used.

→ **Residual analysis:** Residuals are the differences between the predicted values and the actual values of the target variable.

→ **Normality of residuals:** Another assumption of linear regression is that the residuals follow a normal distribution. Method is to use a Q-Q plot, which plots the standardized residuals against the expected normal distribution. If the residuals are normally distributed, the points will fall along a straight line.

→ **Homoscedasticity:** Homoscedasticity is the assumption that the variance of the residuals is constant across all levels of the predictor variables.

→ **Multicollinearity:** Multicollinearity occurs when two or more predictor variables are highly correlated with each other.

→ **Outliers:** Outliers are extreme values that can affect the regression line and distort the results.

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

→ Based on the final model, the top 3 features contributing significantly towards explaining the demand of the shared bikes are:

→ **Temperature:** The temperature has the highest positive coefficient, indicating that as the temperature increases, the demand for shared bikes also increases.

→ **Humidity:** The humidity has a negative coefficient, indicating that as humidity increases, the demand for shared bikes decreases.

→ **Windspeed:** Windspeed has a negative coefficient as well, indicating that as windspeed increases, the demand for shared bikes decreases.

These features are important in predicting the demand for shared bikes and should be considered in any future decision-making related to bike sharing services.

General Subjective Questions

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

Answer:

→ Linear regression is a common statistical method used to model the relationship between a dependent variable (usually denoted Y) and one or more independent variables (usually denoted X).

→ The goal of linear regression is to find the line of best fit that represents the linear relationship between the variables.

→ In this algorithm, we assume that the relationship between the variables is linear, which means that as one variable increases, the other variable also increases at a constant rate.

→ The linear regression algorithm uses the method of least squares to find the line of best fit that minimizes the sum of the squared differences between the predicted and actual values.

→ The steps in linear regression algorithm are:

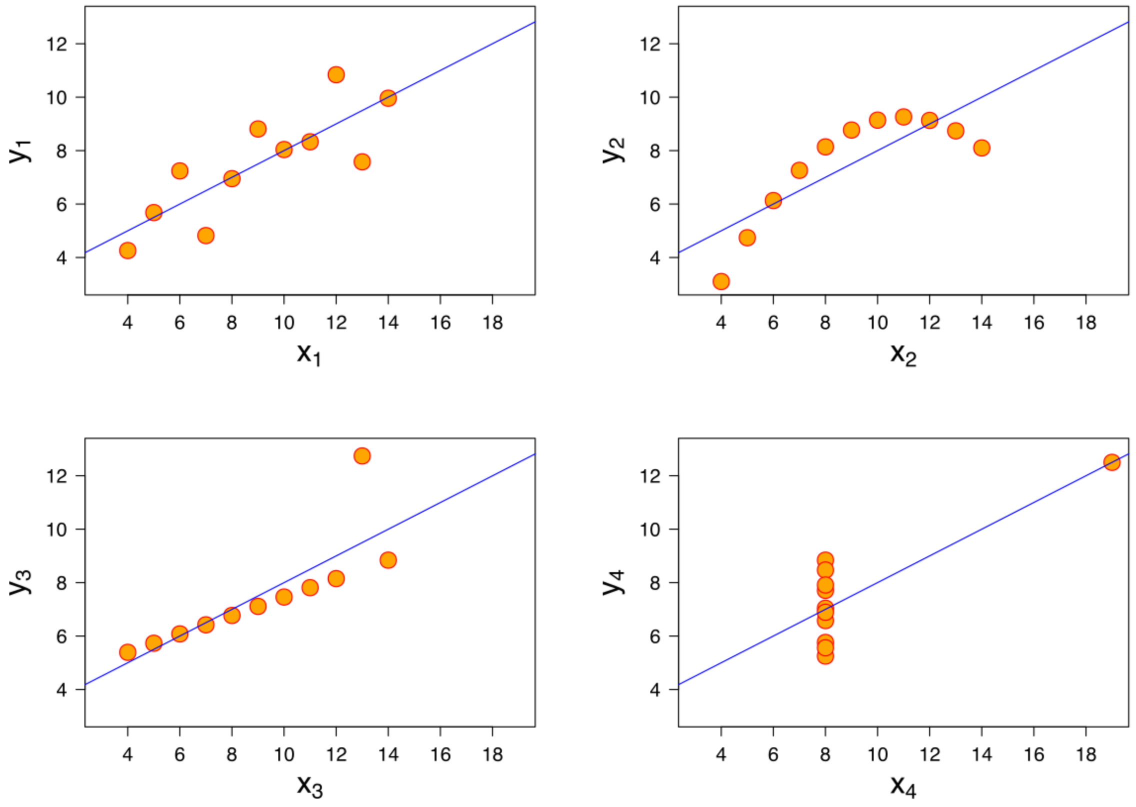
- **Data Collection -** Collection of data on dependent and independent variables.
- **Data preprocessing -** Cleans data by removing outliers, missing values, or inconsistencies.
- **Split Data -** Split data into training and test sets.
- **Scale Features -** Normalizes data by scaling features to a common scale.
- **Model Training -** Trains a linear regression model on the training data.
- **Predict -** Use the trained model to predict the value of the dependent variable.
- **Model Evaluation -** Evaluate model performance by comparing predicted values to actual values using metrics such as root mean square error (RMSE), mean absolute error (MAE), or R-squared.

→ $y= mx +c$

where , m – slope and c is constant

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

→ Anscombe's quartet is a group of four sets of data that look very different when drawn on a graph, but have similar statistical properties like the average and the spread.
→ The quartet was created to show that we should not blindly trust numbers and formulas when analyzing data.
→ Instead, we should look at visual representations of data to better understand any patterns and details that are not immediately obvious from just looking at the numbers.
→ This is important because analyzing data in this way can provide us with a better understanding of the relationships between different variables.



3. What is Pearson’s R? (3 marks)

Answer:

→ Pearson's r, also known as Pearson's correlation coefficient or simply correlation coefficient, is a statistical measure that indicates the strength and direction of a linear relationship between two continuous variables.

→ It is named after British mathematician and biostatistician Carl Pearson.

→ Pearson's r ranges from -1 to 1, where -1 means a perfect negative correlation (one variable increases, the other variable decreases), 0 means no correlation, and 1 means a perfect positive correlation (one variable increases, l other variable increase) also increases).

The Pearson formula r is:

r = (NΣxy - ΣxΣy) / sqrt((NΣx^2 - (Σx)^2) (NΣy^2 - (Σy)^2))

where:

N = number of observations

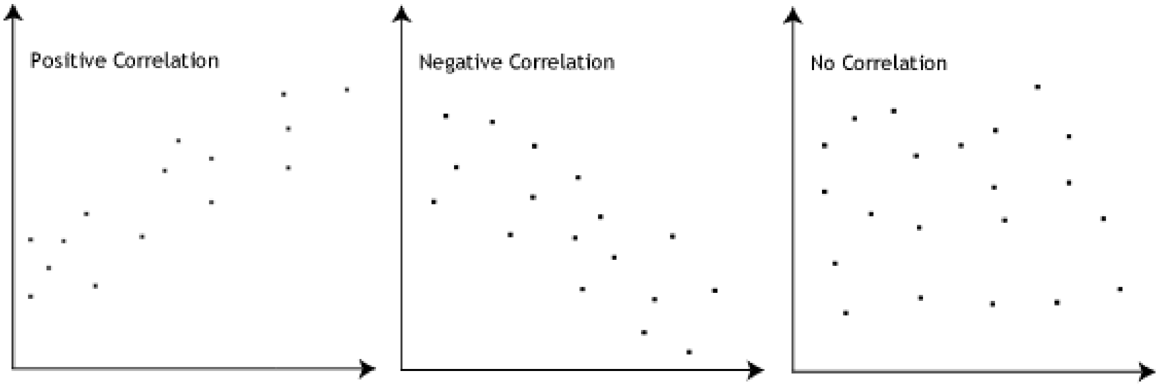
Σ = the sum of

x = value of first variable

y = value of second variable

→ Pearson's correlation coefficient is widely used in various fields such as psychology, sociology and science. finance.

→ It can be used to determine a relationship between two variables, test hypotheses about that relationship, and predict the value of one variable relative to another.



4. **What is scaling? Why is scaling performed? What is the difference between normalized scalingand standardized scaling?** (3 marks)
Answer:

→ Scaling is a preprocessing step in data analysis that involves converting the values of features (variables) in a dataset to a similar scale.
→ This is done to ensure that all features are treated equally by the analysis algorithm and to prevent features with larger value ranges from dominating features with smaller ranges.
→ To improve the performance of machine learning algorithms: some algorithms are sensitive to feature scale and may perform poorly if features have different scales.
→ Improve interpretability: When features are at the same scale, it is easier to compare their relative importance and explain model coefficients.
→ Accelerate optimization: Scale can help optimization algorithms converge.

→ There are two common methods of scaling data: normalized scaling and normalized scaling.

→ Normalized scaling transforms data into a range between 0 and 1. The formula for normalized scaling is:

x_normalized = (x - min(x)) / (max (x)) - min(x))

where x is a single observation and min(x) and max(x) are features The minimum and maximum values of x for all observations.

Normalized scaling (also called z-score scaling) transforms the data to have a mean of 0 and a standard deviation of 1.
→ The formula for the standardized scale is:

x_standardized = (x - mean(x)) / std(x)

where x is a single observation, mean(x) is the mean of characteristic x over all observations, std(x) is the standard deviation of the characteristic x for all observations.

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

→ The variance inflation factor (VIF) is a measure used to detect multicollinearity in regression analysis .

→ That determines how much the variance of estimated regression coefficients increases due to correlations between predictor variables.

→ A high VIF value indicates that the predictor is highly correlated with other predictors, indicating multicollinearity in the model.

→ Infinite VIF values occur when one of the predictors in the model is a linear combination of the other predictors.

→ In this case, the regression model cannot be estimated correctly, resulting in an infinite VIF value.

→ To avoid problems with infinite VIFs, it is important to check for multicollinearity between predictors before building a regression model.

→ One approach is to calculate correlation coefficients between predictor variables, and if two or more variables are highly correlated, one of them can be dropped from the model.

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 technique used to compare the distribution of a sample of data to a theoretical distribution such as the normal distribution.

→ It plots the quantiles of the sample data on the x-axis and the quantiles of the theoretical distribution on the y-axis.

→ In linear regression, Q-Q plots are often used to test whether the residuals of the regression model obey the normal distribution.

→ The residuals represent the difference between the observed and predicted values of the dependent variable, and they must follow a normal distribution with zero mean and constant variance for a linear regression model to be valid.

→ To create a Q-Q plot for the residuals, first sort the residuals in ascending order, then calculate the quantile of the theoretical normal distribution for each residual.

→ The resulting pairs of observed and expected quantiles are then plotted on a Q-Q plot.

→ If the residuals are normally distributed, the points on the Q-Q plot should form a line with slope 1.

→ Any deviation from this line indicates that the residuals do not follow a normal distribution.

→ In summary, Q-Q plots are useful tools for testing assumptions of normality and identifying outliers in linear regression models.

→ They can help ensure the validity and reliability of the model and guide analysis improvements