Name – Pritam Saha, SAP ID – I515969

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)

   Answer -

   Based on the visual analysis of the categorical variables against dependent variable cnt, following key inferences could be obtained –

Variable name	Inference
yr	There's a signi**fi**cant increase in bike rental counts (cnt) in 2019 (year 1) compared to 2018 (year 0). This suggests a **strong positive trend in bike usage over time.**
mnth	Bike rental demand is generally higher during the warmer months, specifically from May to October. Conversely, the winter months (like January and February) show lower counts.
season	Demand for bike rentals is significantly lower in the spring season compared to summer, fall, and winter. Summer and fall seasons tend to have the highest rental counts.
weathersit	The best weather conditions ('Clear/Partly Cloudy') correspond to the highest bike rental counts. As weather conditions worsen (e.g., 'Mist and Cloudy', 'Light Snow/Rain'), the bike rental counts decrease. It was also noted that there were no records for the worst weather condition ('Heavy Snow/Rain').
weekday	The median bike rental counts are almost similar across all weekdays.
holiday	Days designated as holidays generally show lower bike rental counts compared to non-holidays.
workingday	While not explicitly stated as a strong effect, working days often have higher registered user counts, and the overall cnt might show variations influenced by this.

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

   Answer –

   - It is important to use 'drop_first = True' such that for a categorical variable of k categories, it will create k-1 dummy variables.
   - The k-1 approach ensures that there's no perfect multicollinearity. The dr
   - Example, there were 4 seasons, sprint, summer, fall and winter, it was represented by 3 dummy variables, season_summer, season_fall, season_winter. The 'spring' category is now represented by all three of these dummy variables being 0.

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

   Answer –

   temp (temperature) has the highest correlation with the target variable cnt.

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

   Answer –

   Steps followed to validate the assumptions after building the model on the training data set –

   - We plotted a histogram of the error terms (residuals), which are the differences between the actual cnt values (y_train) and the predicted cnt values (y_train_pred) from the model.
   - We observed that the histogram of the error terms was approximately normally distributed (it showed a bell-shaped curve). This is a key assumption of linear regression, indicating that the errors are random and not systematically skewed.
   - The distribution of these errors was also centered around zero. This implies that the model is not consistently over-predicting or under-predicting the target variable. The positive and negative errors balance out, which is indicative of an unbiased model.

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

   Answer –

Here are the top 3 features contributing significantly towards explaining the demand for shared bikes (ordered by the absolute magnitude of their coefficients):

- temp (Normalized Temperature): With a coefficient of 0.549936, temperature has the largest positive impact on bike demand. As temperature increases, the demand for shared bikes significantly increases.
- weathersit_Light Snow/Rain (Light Snow/Rain Weather): With a coefficient of -0.288021, this weather condition has a very strong negative impact. When the weather is light snow or rain, the demand for shared bikes drastically decreases.
- yr (Year): With a coefficient of 0.233056, the year has a substantial positive effect. This indicates a significant increase in bike demand from 2018 (represented by 0) to 2019 (represented by 1), suggesting a growing popularity or awareness of the bike-sharing service.

General Subjective Questions

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

Answer –

Linear Regression is a statistical method used to model the linear relationship between a dependent variable (what you're trying to predict, often denoted as y) and one or more independent variables (predictors, denoted as x). The goal is to find the "best-fitting" straight line (or hyperplane in higher dimensions) that minimizes the sum of squared differences between the observed and predicted values.

1. Simple Linear Regression (SLR)

This models the relationship between a single dependent variable and a single independent variable.

- **Equation:** $y = \beta_0 + \beta_1 x + \varepsilon$
  - y: The dependent variable (predicted value).
  - β_0: The y-intercept (the value of y when x is 0).
  - β_1: The slope coefficient (the change in y for a one-unit change in x).
  - x: The independent variable (predictor).
  - ε: The error term (or residual), representing the unexplained variance or noise.

2. Multiple Linear Regression (MLR)

This extends SLR to model the relationship between a single dependent variable and two or more independent variables.

- **Equation:** $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon$
  - y: The dependent variable.
  - β_0: The y-intercept.
  - β_i: The coefficient for the i-th independent variable xi. It represents the change in y for a one-unit change in xi, holding all other predictors constant.
  - $x_1, x_2, ..., x_n$: The multiple independent variables.
  - ε: The error term.

Key Emphases:

- **Relationship between Variables (Linearity):** A fundamental assumption is that there's a linear relationship between the dependent variable and each independent variable. The coefficients (β values) quantify this relationship: a positive β means y increases with x, a negative β means y decreases with x, and the magnitude indicates the strength.
- **Multicollinearity:** This occurs when two or more independent variables in the MLR model are highly correlated with each other. It's problematic because it makes it difficult to ascertain the individual effect of each predictor on y, leads to unstable and unreliable coefficient estimates (large standard errors), and can make the model difficult to interpret. It's often detected using the Variance Inflation Factor (VIF).
- Normality of Errors (Residuals): Another crucial assumption is that the error terms (ε) are normally distributed with a mean of zero. This is vital for the validity of hypothesis tests (like t-tests for individual coefficients and the F-test for overall model significance) and for

constructing reliable confidence intervals. Violations can lead to incorrect conclusions about the statistical significance of predictors. This assumption is typically validated by examining histograms or Q-Q plots of the residuals.

Other important assumptions (briefly): errors are independent (no autocorrelation), and errors have constant variance (homoscedasticity).

In essence, Linear Regression aims to quantify relationships, but its reliability hinges on understanding and validating these underlying assumptions.

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

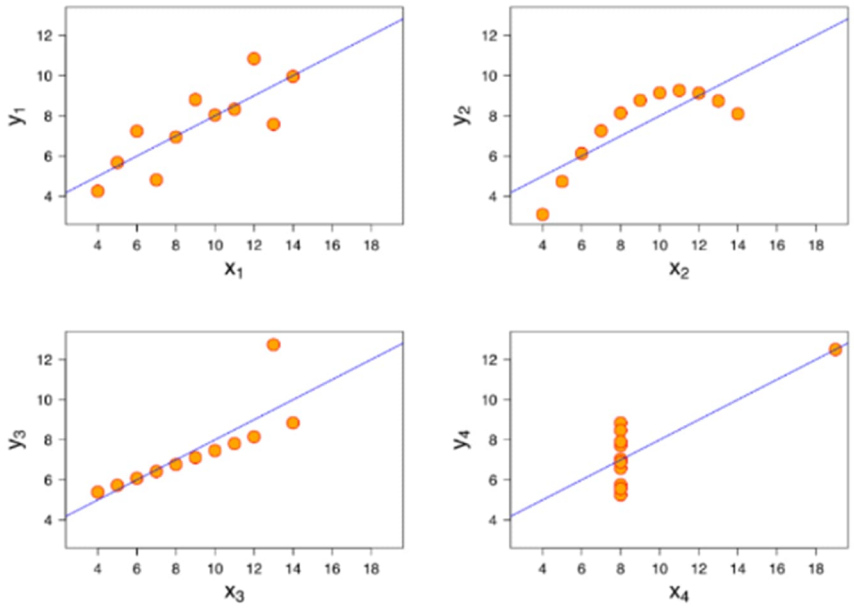
Answer –

Anscombe's Quartet is a famous collection of four distinct datasets, first published by statistician Francis Anscombe in 1973. Each dataset consists of eleven (x, y) points. The remarkable aspect of these datasets is that they share nearly identical simple descriptive statistics, yet when plotted graphically, they reveal strikingly different distributions and relationships between the variables.

Identical Statistical Properties: For all four datasets, these statistics are almost identical:

- Mean of x values: 9.0
- Variance of x values: 11.0
- Mean of y values: 7.50
- Variance of y values: 4.12
- **Correlation coefficient between x and y:** 0.816
- **Linear regression line equation:** y = 3.00 + 0.50x
- **Coefficient of determination (R-squared) for the linear regression:** 0.67

When we plot these four datasets on a x/y coordinate plane, we can observe that they show the same regression line, but each data set tells different story –



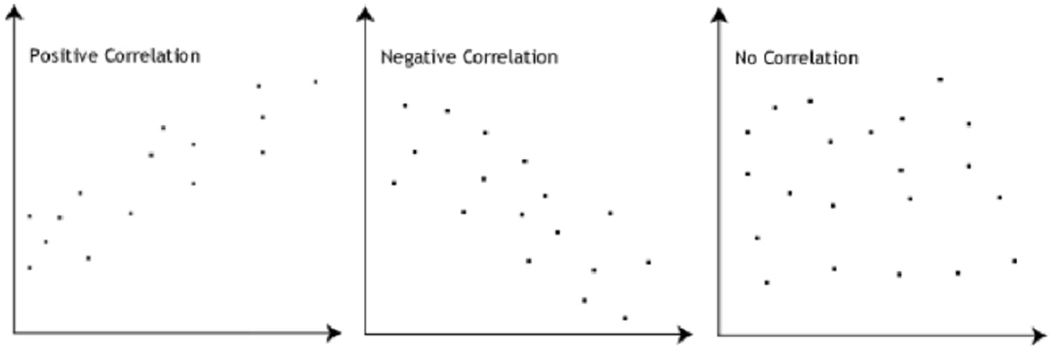
1. **Dataset I (Linear Relationship):** This dataset shows a straightforward, positive linear relationship between x and y with some scatter. This is what you would expect from a dataset that has a high positive correlation and is well-described by a simple linear regression model.
2. Dataset II (Non-**linear Relationship):** This dataset exhibits a clear curvilinear (parabolic) relationship between x and y. Although the linear regression line calculation produces the same line as Dataset I, it's evident from the plot that a linear model is entirely inappropriate for capturing the underlying pattern in the data.
3. Dataset III (Linear with Outlier In**fl**uence): In this dataset, there appears to be a strong linear relationship for most of the data points, like Dataset I. However, there's one significant outlier (an x value far from the others) that pulls the regression line towards itself. If this single outlier were removed, the regression line would change dramatically, highlighting the outlier's undue influence on the overall statistics.
4. **Dataset IV (Vertical Cluster with Influential Point):** This dataset is characterized by a cluster of data points that all have the same x value but varying y values. There is also a single data point with a very different x value that has a strong leverage effect on the regression line, dictating its slope. Without this one influential point, the slope of the regression line would be undefined or very different, as the remaining points only form a vertical line.

The Crucial Lesson of Anscombe's Quartet: Anscombe's Quartet serves as a powerful reminder that **summary statistics alone are insufficient for understanding the full nature of a** dataset. Relying solely on numerical summaries can be highly misleading, as vastly different data distributions can produce identical statistical measures. It strongly emphasizes the critical importance of visualizing data (performing Exploratory Data Analysis, or EDA) before drawing conclusions or fitting statistical models. Visual inspection can quickly reveal patterns, outliers, and relationships that numerical summaries completely obscure, guiding analysts toward appropriate modelling choices.

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

Answer –

Pearson's R, also known as the **Pearson correlation coefficient**, Pearson product-moment **correlation coefficient (PPMCC)**, or simply **correlation coefficient**, is a measure of the linear correlation between two sets of data. It is widely used in statistics to quantify the strength and direction of a linear relationship between two quantitative variables.



The value of Pearson's R always falls between -1 and +1, inclusive:

- +1: Indicates a perfect positive linear relationship. As one variable increases, the other increases proportionally.
- -1: Indicates a perfect negative linear relationship. As one variable increases, the other decreases proportionally.
- 0: Indicates no linear relationship between the two variables. This does *not* mean there is no relationship at all; there could be a strong non-linear relationship (e.g., a parabolic one).

Limitations: It only measures *linear* relationships. If the relationship is non-linear, Pearson's R might be close to zero even if there's a strong association (e.g., in Anscombe's Quartet, some datasets have a high R despite a non-linear pattern).

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

Answer –

Scaling is a data preprocessing technique used to transform numerical features in a dataset to a standard range or distribution. It involves adjusting the range or variance of the independent variables (features) so that they contribute equally to the distance metric used in many machine learning algorithms. Essentially, it brings all numerical values in columns to a comparable scale.

Scaling is performed for the following reasons –

- Many machine learning algorithms calculate distances between data points. If features have vastly different scales, the feature with the largest range will dominate the distance calculation, making the algorithm biased towards it. Scaling ensures all features contribute proportionally.
- Algorithms that rely on gradient descent (e.g., Linear Regression, Logistic Regression, Neural Networks) converge much faster when features are on a similar scale. Without scaling, the cost function can have elongated contours, making it difficult for gradient descent to find the optimal solution efficiently.
- Features with larger numerical values or ranges might implicitly be given more weight by certain algorithms, even if they are not inherently more important. Scaling prevents this unintended influence.

Characteristics	Normalized Scaling	Standardized Scaling
Formula	X_scaled = (X - X_min) / (X_max - X_min)	X_scaled = (X - μ) / σ • μ (mu): The mean of the feature. • σ (sigma): The standard deviation of the feature.
Transformation	This method scales and translates each feature individually such that it falls within a specific range, typically [0, 1] or [-1, 1].	This method scales features such that they have a mean (**μ**) of 0 and a **standard deviation (σ**) of 1. This is also known as Z-score normalization.
Characteristics	It directly adjusts the range of the data. The minimum value of the feature becomes 0 (or	It transforms the data into a standard normal distribution (or at least brings it closer to one if the original data is not

	-1) and the maximum value becomes 1.	normally distributed). The values are not bounded to a specific range, meaning outliers are preserved (they will have large absolute Z-scores).
When to Use	It's useful when you need features to be within a strict bounded range. It's less affected by outliers than standardization if the outliers are within the min-max range, but a single extreme outlier can compress most of the data into a very small range.	It's preferred for algorithms that assume a Gaussian distribution of features (e.g., Linear Regression, Logistic Regression, Gaussian Naive Bayes, PCA, SVMs with RBF kernel). It's robust to outliers because they do not compress the rest of the data as much as min-max scaling might. However, the exact range of values after standardization is not guaranteed.

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

Answer –

An infinite VIF (Variance Inflation Factor) value occurs when there is perfect **multicollinearity** among the predictor variables in your regression model.

To understand why, let's look at the formula for VIF for a predictor Xi:

$VIF_i = 1 / (1 - R_i^2)$

Where:

- R_i^2 is the R-squared value obtained from regressing the predictor Xi against all the *other* independent variables in the model.

Now, consider what happens when there is perfect multicollinearity:

- **Perfect Multicollinearity**: This means that one predictor variable can be perfectly predicted (or explained) by a linear combination of the other predictor variables. In simpler terms, one X_i is a direct linear function of other X's.

- Impact on R_i^2: If Xi can be perfectly predicted by the other predictors, then the regression of Xi on those other predictors will result in an R_i^2 value of 1 (or very, very close to 1).
- In**fi**nite VIF: If R_i^2 = 1, then the denominator of the VIF formula becomes (1 - 1) = 0. Dividing by zero results in an in**fi**nite VIF.

Common scenarios that lead to infinite VIF (perfect multicollinearity):

- If you create k dummy variables for a categorical variable with k categories and include all of them along with an intercept term in the model. One dummy variable can be perfectly predicted by the others.
- If you accidentally include the same variable twice, or include one variable that is a direct linear combination of others (e.g., total sales, and then sales by product A and sales by product B, where total = A + B).

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

(3 marks)

Answer –

A Q-Q plot compares the quantiles of your sample data against the quantiles of a theoretical distribution. If the two distributions match, the points on the Q-Q plot will approximately lie on a straight line.

How it's Constructed (Conceptually):

1. Order Data: Your sample data points are sorted in ascending order.
2. **Calculate Quantiles**: For each data point in your sorted sample, its empirical quantile is determined.
3. **Theoretical Quantiles**: Corresponding quantiles are calculated from the theoretical distribution (e.g., standard normal distribution).
4. Plot: Each data point's empirical quantile is plotted against its corresponding theoretical quantile.

Interpretation:

- Straight Line: If the data points fall approximately along a straight line (often a 45-degree line), it suggests that your sample data follows the theoretical distribution.
- **Deviations from the Line**: Any significant departure from the straight line indicates that your data does not follow the theoretical distribution. For example:
    - S-shape: Suggests the data has lighter or heavier tails than the theoretical distribution.
    - Curvature: Indicates skewness in the data.
    - **Points at Ends Deviating**: Points only at the ends deviating from the line usually mean outliers or heavier/lighter tails.

Use and Importance in Linear Regression

In linear regression, one of the crucial assumptions for valid inference (hypothesis testing, confidence intervals) is that the error terms (residuals) are normally distributed. The Q-Q plot is a primary tool used to visually check this assumption.

1. Checking Normality of Residuals: After fitting a linear regression model, you extract the residuals (the differences between the observed and predicted values). You then create a Q-Q plot of these residuals against a theoretical normal distribution.
2. **Validation of Assumptions**: If the residuals fall along a straight line on the Q-Q plot, it provides strong visual evidence that the normality assumption is met. This is important because:
    - **Reliable Hypothesis Testing**: The p-values and confidence intervals for your model's coefficients are more reliable when residuals are normally distributed. If this assumption is violated, your conclusions about the statistical significance of predictors might be inaccurate.
    - Model Validity: Meeting the normality assumption helps confirm the overall statistical validity of your linear regression model and the generalizability of its findings.

While other methods like histograms can give a general idea of distribution, a Q-Q plot is often more precise for identifying deviations from normality, especially in the tails of the distribution. It allows you to see not just *if* data is normal, but *how* it deviates from normality (e.g., skewness, heavy/light tails).