1. From your analysis of the categorical variables from the above given dataset, what could you infer about their effect on the dependent variable?

A. Season

- **Spring (1):** ~2608 rentals (lowest)
- **Summer (3):** ~5644 rentals (highest)
- Rentals peak in summer and drop in winter, showing strong seasonal influence.

B. Year

- **2018 (0):** ~3406 rentals
- **2019 (1):** ~5610 rentals
  → Clear growth trend over time.

C. Month

- Rentals rise steadily from January (~2176) to September (~5766), then decline toward December (~3404).
  → Indicates strong monthly/seasonal variation.

D. Holiday

- **Holiday (1):** ~3735 rentals
- **Non-holiday (0):** ~4531 rentals
  →Fewer rentals on holidays compared to regular days.

E. Weekday

- Rentals are consistent across weekdays (~4200–4700), slightly higher on Fridays and Saturdays.

F. Working Day

- **Working day (1):** ~4590 rentals
- **Non-working day (0):** ~4330 rentals

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

Using drop_first=True when creating dummy variables is important to avoid the **dummy variable trap**, which occurs due to multicollinearity. By dropping the first category or column name 'Instant', we remove redundant information and ensure the model can uniquely interpret categories without perfect correlation.

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

atemp with corelation 0.621424

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

Assumptions to Validate for Lasso

I. Linearity: The relationship between predictors and target should be linear.

II. Independence of errors: Residuals should be independent.

III. Homoscedasticity: Residuals should have constant variance.

IV. Normality of residuals: Residuals should be approximately normal.

V. No perfect multicollinearity: Lasso helps reduce multicollinearity but doesn't eliminate the need to check.

VI. Feature scaling: Important for Lasso because of the penalty term (we already scaled features in the pipeline).

- Residual Analysis:
  - Plot residuals vs predicted values → should be randomly scattered.
  - Histogram or Q-Q plot of residuals → should look normal.
- Durbin-Watson Test for independence.
- Check Homoscedasticity → no funnel shape in residual plot.
- Normality Test → Shapiro-Wilk or KS test.
- Multicollinearity → VIF (though Lasso mitigates this)

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

- **atemp (feels-like temperature)**: Higher perceived temperature strongly increases bike demand.
- **expected_atemp**: Negative coefficient suggests that deviation from expected temperature reduces demand.
- **yr (year)**: Positive trend over time, indicating growing bike usage year-over-year.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression is one of the most fundamental algorithms in machine learning and statistics, used for predicting a continuous target variable based on one or more input features. Let's break it down step by step:

Linear Regression models the relationship between:

Dependent variable (Y): The outcome you want to predict.

Independent variable(s) (X): The input features.

It assumes this relationship is linear, meaning:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

Where:

β0: Intercept (value of Y when all X = 0)

β1: Coefficients (weights for each feature)

ϵ: Error term (difference between predicted and actual values)

Types of Linear Regression

Simple Linear Regression: One independent variable.

Multiple Linear Regression: Two or more independent variables.

Assumptions of Linear Regression

To work well, linear regression assumes:

Linearity: Relationship between X and Y is linear.

Independence: Observations are independent.

Homoscedasticity: Constant variance of errors.

Normality of errors: Residuals are normally distributed.

No multicollinearity: Features should not be highly correlated.

How Does It Work?

The algorithm finds the best-fitting line by minimizing the Sum of Squared Errors (SSE):

$$\text{SSE} = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

Where:

- y_i: Actual value
- $\hat{y}_i$: Predicted value

The most common method to minimize SSE is Ordinary Least Squares (OLS).

Steps in Linear Regression

Collect Data: Features (X) and target (Y).

Preprocess: Handle missing values, scale features if needed.

Fit Model: Estimate coefficients using OLS.

Predict: Use the learned coefficients to predict Y for new X.

Evaluate:

R^2 (Coefficient of Determination): How much variance in Y is explained by X.

RMSE (Root Mean Squared Error): Average prediction error.

Advantages

Simple and interpretable.

Works well for linearly related data.

Limitations

Sensitive to outliers.

Assumes linearity.

Poor performance if assumptions are violated.

Core Linear Regression Methods

- ✓ Ordinary Least Squares (OLS)
    - What it is: The classic approach that minimizes the sum of squared errors.

- Use when: Data satisfies assumptions (linearity, homoscedasticity, no multicollinearity).
- Pros: Simple, interpretable.
- Cons: Sensitive to outliers and multicollinearity.

✓ Polynomial Regression

- What it is: Extends linear regression by adding polynomial terms.
- Use when: Relationship between predictors and target is non-linear but can be approximated by polynomials.
- Pros: Captures curvature.
- Cons: Risk of overfitting with high-degree polynomials.

✓ Stepwise Regression

- What it is: Automated variable selection (forward, backward, or both).
- Use when: Many predictors, need feature selection.
- Pros: Reduces complexity.
- Cons: Can lead to biased models if not validated.

Regularized Linear Regression Methods

These handle multicollinearity and overfitting: 4. Ridge Regression (L2 Regularization)

- Shrinks coefficients but keeps all features.
- Use when: Predictors are highly correlated.
- Latest trend: Often combined with cross-validation for tuning.

✓ Lasso Regression (L1 Regularization)

- Shrinks coefficients and can eliminate irrelevant features.
- Use when: Need feature selection + regularization.
- Latest trend: Used in high-dimensional datasets.

✓ Elastic Net

- Combines Ridge and Lasso.
- Use when: Many correlated features and need sparsity.
- Latest trend: Preferred for large datasets with mixed feature importance.

Robust & Advanced Variants

✓ Robust Regression (e.g., RANSAC)

Handles outliers by fitting on subsets.

Use when: Data has many outliers.

- ✓ Bayesian Linear Regression
  Incorporates prior knowledge and uncertainty.
  Use when: Probabilistic interpretation is needed.
  Latest trend: Popular in forecasting and uncertainty quantification.
- ✓ Quantile Regression
  Predicts conditional quantiles instead of mean.
  Use when: Interested in median or other quantiles (e.g., risk analysis).

Scenario	Recommended Method
Simple linear relationship	OLS
Non-linear trend	Polynomial Regression
Many predictors, multicollinearity	Ridge
Need feature selection	Lasso
Both multicollinearity & feature selection	Elastic Net
Outliers present	Robust Regression
Uncertainty matters	Bayesian Regression
Focus on median or extremes	Quantile Regression

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a famous example in statistics introduced by Francis Anscombe in 1973. It consists of four different datasets that have nearly identical summary statistics but very different distributions and visual patterns when graphed. The quartet illustrates the importance of data visualization and not relying solely on descriptive statistics.

You are analyzing sales performance for four different regions. You calculate:

- Average sales = $7,500
- Average marketing spend = $9,000
- Correlation between spend and sales = 0.82
- Regression line:
  Sales=3,000+0.5×Marketing Spend\text{Sales} = 3,000 + 0.5 \times \text{Marketing Spend}Sales=3,000+0.5×Marketing Spend

All four regions show identical summary stats. So, you might think:

“Great! Marketing spend strongly drives sales everywhere.”

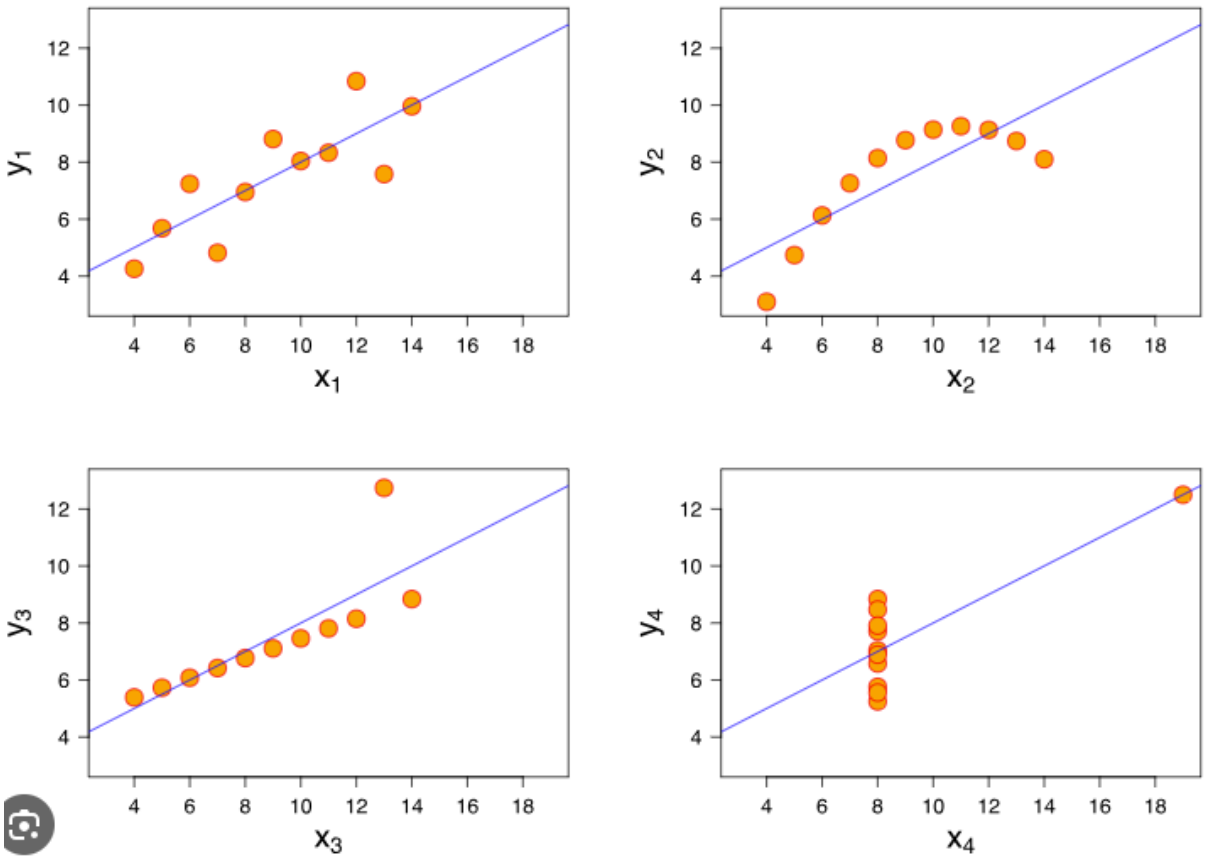
But When You Plot the Data...

Region A:Data points follow a nice straight line → Spend and sales truly correlate.

Region B:Data forms a curve → Spending more after a point doesn’t increase sales (diminishing returns).

Region C: Mostly linear, but one huge outlier (a campaign that spent $50,000 and got $100,000 sales) skews the stats.

Region D:All campaigns had similar spend and sales, except one extreme case that drives the correlation.



If you only look at averages and correlations:

- You might assume all regions behave the same.
- You could waste money by applying the same strategy everywhere.

Visualization reveals reality:

- Region B needs a cap on spend.
- Region C’s outlier should be investigated.

- Region D's correlation is misleading

3. What is Pearson's R

Pearson's R (also called the Pearson correlation coefficient) is a statistical measure that tells you how strongly two variables are related and in what direction.

Range:

$$-1 \leq r \leq 1$$

- r = I ⟶ Perfect positive correlation (as one increases, the other increases).
- r = -I ⟶ Perfect negative correlation (as one increases, the other decreases).
- r = 0 ⟶ No linear relationship.

It measures the strength and direction of a linear relationship between two continuous variables.

Formula:

$$r = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y}$$

Where:

- Cov(X,Y) = covariance between X and Y
- σX, σY = standard deviations of X and Y

Suppose you analyze:

- Marketing Spend (X) and Sales Revenue (Y) across campaigns.
- If r = 0.85, it means:
  - Strong positive correlation ⟶ Higher marketing spend tends to increase sales.
- If r = -0.60, it means:
  - Negative correlation ⟶ Higher discounts might reduce profit margins.

Why is this useful?

- If r is close to +I, you know spending more strongly relates to higher sales.
- If r is close to 0, spending more doesn't really affect sales.
- If r is negative, spending more might reduce sales (rare, but possible in some cases).

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

Scaling means changing the range of your data so that all features are on a similar scale.
For example:

- One column = Age (range: 18–70)
- Another column = Income (range: $20,000–$200,000)

If you use these directly in a model, Income dominates Age because its numbers are much bigger.
Scaling fixes this by bringing both to a comparable range.

Why is Scaling Performed?

- Many algorithms (like Linear Regression, KNN, SVM) depend on distance or magnitude.
- If features have very different scales, the model becomes biased toward large numbers.
- Scaling makes training faster and more accurate.

Types of Scaling

❖ Normalization (Min-Max Scaling)

- Rescales data to 0–1 range.
- Formula:
  Formula:
  $$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$
- Example:
  If Age = 50 (range 18–70):
  50 – 18 / 70 – 18 = 0.61
- Use when you know min and max values matter (e.g., image pixels).

❖ Standardization (Z-score Scaling)

Rescales data to mean = 0, standard deviation = 1.

Formula:

$$X' = \frac{X - \mu}{\sigma}$$

μ (mu) = mean of the feature (average value).

σ (sigma) = standard deviation of the feature (how spread out the values are around the mean).
It measures variability.

If Age = 50, mean = 30, std = 10

X’ = 50 – 30 / 10 = 2

Use when data has outliers or unknown range.

Calculations of Mu and Sigma

Example: If ages are [20, 30, 40], then

μ (mu) = (20+30+40) divided by 3 = 30

σ (sigma)

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

$$\sigma = \sqrt{\frac{(20-30)^2 + (30-30)^2 + (40-30)^2}{3}} = \sqrt{\frac{100+0+100}{3}} \approx 8.16$$

Use when data has outliers or unknown range.

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

VIF (Variance Inflation Factor) measures how much the variance of a regression coefficient is inflated due to multicollinearity (when predictors are highly correlated).

The formula for VIF is:

$$\text{VIF}_i = \frac{1}{1 - R_i^2}$$

Where: R_i^2 = R-squared value when predictor iii is regressed on all other predictors.

If $R_i^2 = 1$, then:

$$\text{VIF} = \frac{1}{1-1} = \frac{1}{0} \to \infty$$

This happens when:

- One predictor is a perfect linear combination of other predictors.
- Example:
- If you have: $X_3 = X_1 + X_2$

You want to predict house price using these features:

- Square Footage
- Number of Rooms
- Total Area in Square Feet

But notice:

$$\text{Total Area in Square Feet} = \text{Square Footage}$$

This means Total Area is a perfect duplicate of Square Footage.

When you regress Total Area on other predictors, $R^2 = 1$ because it's perfectly explained by them.

Why does this matter?

- Perfect multicollinearity makes regression unstable.
- Coefficients become meaningless because predictors are redundant.

Real-world scenario:
If your dataset has derived features (like "Total Spend" and "Marketing Spend"), or dummy variables that sum to 1, you can easily hit infinite VIF.

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

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to compare the distribution of your data to a theoretical distribution—most commonly the normal distribution.

It plots quantiles of your sample data against quantiles of a theoretical distribution (e.g., normal).

If your data follows that distribution, the points will lie roughly on a straight diagonal line.

Linear regression assumes Residuals (errors) are normally distributed. This is important for valid hypothesis tests and confidence intervals. A Q-Q plot helps check this assumption: If residuals are normal → Points align along the diagonal. If residuals deviate → You see curves or outliers.

Detects non-normality (skewness, heavy tails).

Helps decide if you need:

- Transformations (e.g., log, Box-Cox)

- Robust regression methods

Imagine you fit a regression model for sales vs marketing spend:

- After fitting, you plot residuals on a Q-Q plot.
- If points curve upward → Residuals have heavy tails → Normality assumption fails.