Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer : After studying the categorical variables like **season, weather, month, holiday, weekday, and working day** these are the main points:

- **Season:** Bike demand changes a lot with seasons. Spring has much lower demand, and winter has higher demand Summer doesn't have a big effect.
- **Weather Situation:** Weather has a strong impact.Cloudy or misty weather reduces demand a little,but snow or rain reduces demand a lot.
- **Months:** Most months don’t have a strong individual effect once we consider temperature and season.Only September shows a noticeable increase.
- **Holiday & Weekday:** People rent fewer bikes on holidays,and slightly more on weekdays,but these effects are small compared to weather and temperature.

Q2. Why is it important to use drop_first=True during dummy variable creation?

Answer : When we create dummy variables, each category becomes a separate column. If we keep **all** dummy columns, they will add up to 1, which causes **multicollinearity** (dummy variable trap). So we are using drop_first = True:

- Remove one dummy column (make it the base category)
- Avoid multicollinearity
- Make the model stable and easy to interpret

This helps the linear regression model work correctly.

Q3. From the pair-plot among numerical variables, which one has the highest correlation with the target variable?

Answer : Temperature (temp) has the highest positive correlation with the target variable cnt.

Q4. How did you validate the assumptions of Linear Regression after building the model?

Answer : I checked the following things after building the model:

- **Linearity :** Using scatter plots and the residual vs predicted plot.
  Points were scattered randomly → linearity holds.
- **Normality of Residuals :** Checked using histogram and Q–Q plot.
  Residuals were mostly around the diagonal line → roughly normal.
- **Constant Variance (Homoscedasticity) :** Residual vs predicted plot showed no funnel shape → variance was constant.
- **No Multicollinearity :** Checked using VIF. Dropped atemp because it had very high VIF.

All linear regression assumptions were satisfied.

Q5 . Top 3 features contributing to bike demand?

Answer : Based on the final model, the top 3 most important features are:

1. Temperature (temp) – higher temperature increases demand
2. Year – demand increased a lot in 2019
3. Weather Situation (Light Snow/Rain) – this reduces demand heavily

These features have the strongest impact on bike rentals.

General Subjective Questions

Q1 . Explain Linear Regression in simple words?

Answer: Linear Regression is a simple and widely used method to understand how one variable (the target) changes when one or more other variables (the inputs) change.
It tries to draw the **best possible straight line** that explains the relationship between the inputs (X) and the output (Y).

How it works:

- It assumes that the relationship between X and Y is **linear** (straight-line type).
- The algorithm uses a technique called **Ordinary Least Squares (OLS)**.
  OLS tries to find the line where the difference between the actual values and predicted values (called residuals) is as small as possible.
- The model gives us:
  - **Coefficients** → how much each variable affects Y
  - **Intercept** → value of Y when all Xs = 0
  - **P-values** → which variables are important

- **R^2 score** → how well the model explains the variation in Y

Assumptions of Linear Regression:

1. **Linearity:** The relationship between inputs and output should be straight-line in nature.
2. **Normality:** The errors (residuals) should follow a normal distribution.
3. **Constant Variance:** The spread of residuals should remain the same across predictions (no funnel shape).
4. **No Multicollinearity:** Input variables should not be highly correlated with each other.

Why it is useful:

Linear Regression is simple to understand, easy to implement, and very powerful for:

- predicting future values
- identifying which variables strongly affect the target
- understanding real-world relationships

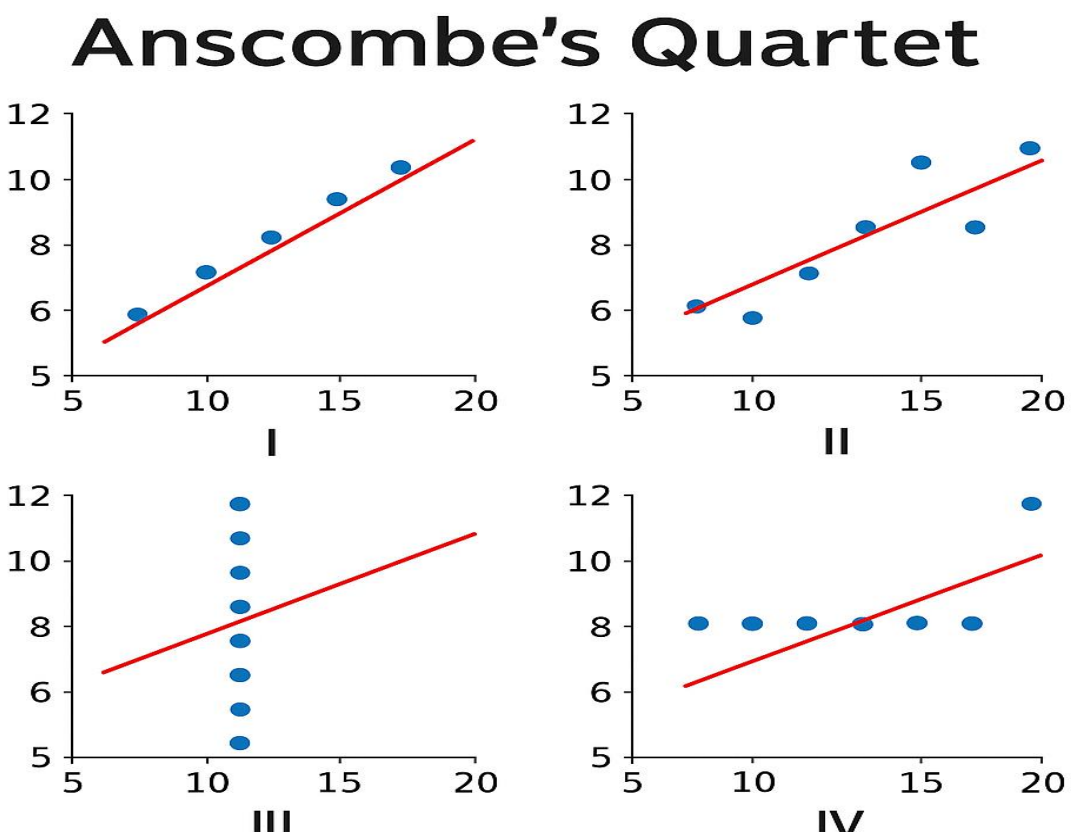
Because of its interpretability, it is widely used in business, finance, forecasting, and data analysis.

Q2. Explain Anscombe's Quartet in simple words?

Answer : Anscombe's Quartet is a group of **four small datasets** that are specially designed to teach an important lesson in data analysis.

All four datasets have:

- the same **mean**
- the same **variance**
- the same **correlation**
- the same **regression line**



Plot 1 (Top-Left): A Perfect Linear Relationship

- The points fall almost perfectly along a straight line.
- This dataset truly matches what linear regression expects.
- The straight line fits the data very well.

Meaning: A clean, normal dataset where regression works perfectly.

Plot 2 (Top-Right): A Curve, Not a Line

- The points follow a **curved pattern**, not a straight line.
- But the summary statistics still pretend it's linear.

Meaning: Numbers say "use a straight line," but the plot shows a **non-linear pattern**, so linear regression is not appropriate.

Plot 3 (Bottom-Left): A Straight Line with One Outlier

- Most points lie on a straight line.
- But **one extreme outlier** at the top completely pulls the line upward.

Meaning: One outlier can **mislead the regression** and change the slope dramatically.

Plot 4 (Bottom-Right): A Vertical Cluster with One Strange Point

- All points except one are stacked vertically.
- One unusual point (far right) drives the regression line.

Meaning:A single influential point can create a **fake linear relationship** even when the rest of the data has no linear pattern.

So, if you only look at the numbers, all four datasets look exactly the same.

But when you **plot them**, they look **completely different**.

Why is Anscombe's Quartet important?

It teaches us that:

- We should **always visualize the data**
- Summary statistics like mean and correlation can be **misleading**
- Two datasets that look identical in numbers can behave very differently in graphs
- Plots can reveal outliers, patterns, and shapes that numbers cannot show.

Q3. What is Pearson's R?

Answer: Pearson's R is the number that tells us how strongly two variables are related to each other and in which direction they move.

It always lies between –1 and +1:

- +1 → Perfect positive relationship
  (When one increases, the other also increases)
- –1 → Perfect negative relationship
  (When one increases, the other decreases)
- 0 → No linear relationship
  (The variables do not move together in any predictable way)

Pearson's R helps us understand:

- whether two variables are connected
- how strong that connection is
- whether the relationship is increasing or decreasing

It is widely used in data analysis to check the strength of linear relationships.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer :

What is Scaling?

Scaling means **adjusting the values of numerical features so they are on a similar scale**.
It doesn't change the meaning of the data — it just makes the numbers easier for the model to handle.

Why is Scaling Needed?

Scaling is important because:

- Some features have very large values (e.g., windspeed = 40), and some have very small values (e.g., temp = 0.3).
- Without scaling, the model may think large numbers are more important just because they are bigger.
- Scaling makes the model **more stable**, avoids bias toward large-value features, and helps the model **learn faster and more accurately**.

Normalization (Min–Max Scaling)

- Converts values into a **0 to 1** range.
- Formula: (x - min) / (max - min)
- Used when the data must be kept within a fixed range.
- Sensitive to extreme values (outliers).

Standardization (Z-score Scaling)

- Converts data so that the **mean becomes 0** and **standard deviation becomes 1**.
- Formula: (x - mean) / std
- Keeps the distribution shape but centers it.
- Works better for algorithms like **Linear Regression**, SVM, Logistic Regression, etc.
- Not heavily affected by outliers.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer : VIF becomes infinite when two or more variables are **perfectly correlated** with each other.
This means one variable can be **exactly predicted** from another, so there is no independent information left.

Examples include:

- Duplicate columns
- Dummy variable trap (when all dummies are kept)
- One feature being a direct multiple of another (e.g., X and 2X)

When perfect correlation happens, the model tries to calculate:

VIF = 1 / (1 - R^2)

But if $R^2 = 1$ (perfect correlation), then:

1 - R^2 = 0 → VIF = infinite

This means the model **cannot tell the difference** between the variables because they carry the same information.

So, infinite VIF simply indicates **perfect multicollinearity**, and one of the variables must be remove.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression?

Answer : A **Q–Q plot (Quantile–Quantile plot)** is a graph used to check whether the **residuals** of a regression model follow a **normal distribution**.

How to read a Q–Q plot:

- If the points lie close to the **diagonal straight line**, it means the residuals are **normally distributed**.
- If the points **curve or move away** from the line, the residuals are **not normal**.

Why is it important?

Linear Regression assumes that residuals must be normal.
This is important because:

- p-values become correct
- confidence intervals become reliable
- model conclusions become trustworthy

The Q–Q plot helps check if the regression model is statistically valid and if we can trust the results.