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?

From the analysis of categorical variables such as season, year, month, weekday, holiday, workingday, and weather situation, we can infer the following effects on bike demand (cnt):

- Season:
  - Bike demand was highest in Fall and Winter, and lowest in Spring.
  - Warmer, pleasant seasons encourage outdoor activity, increasing rentals.
- Year (yr): Demand was significantly higher in 2019 compared to 2018, indicating rising popularity of bike-sharing systems over time.
- Month (mnth): Rentals generally increased from January to September, peaking in late summer and early fall, then declining in winter months.
- Weekday / Workingday / Holiday:
  - Rentals were slightly higher on weekends, and lower on holidays, suggesting that a large share of rides are for weekday commuting.
  - The difference was small, meaning people also use bikes for leisure.
- Weather Situation (weathersit):
  - Clear weather days had the highest demand,
  - Misty days showed moderate decline,
  - Rain or snow conditions caused a sharp drop in rentals.

  Categorical variables such as season, year, and weather condition have a significant influence on bike demand. Good weather and later months in the year increase usage, while poor weather or spring seasons reduce it.

Q2: Why is it important to use drop_first=True during dummy variable creation?
When we create dummy variables for a categorical feature (like season or weathersit), each unique category is converted into a new binary column (0 or 1). However, one of these columns is redundant because if we know the values of any three, we can perfectly infer the fourth.
This redundancy causes a problem called the Dummy Variable Trap.

- The Dummy Variable Trap occurs when there is perfect multicollinearity — i.e., one variable can be linearly predicted from others.
- It leads to unstable coefficients and makes the model matrix singular (non-

invertible).

To avoid this, we set drop_first=True while creating dummies. This drops one category automatically (usually the first one) and uses it as the baseline or reference category

Q3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
From the pair-plot and correlation heatmap among the numerical variables (temp, atemp, hum, windspeed, cnt), we observed the following relationships:

- Temperature (temp) has the highest positive correlation with the target variable cnt.
- This means that as temperature increases, the number of bike rentals also increases significantly.
- In contrast, higher humidity or windspeed tends to slightly reduce demand.

Q4: Hence, temp is the most influential numerical predictor of bike demand among the continuous features.
After fitting the model, the following diagnostic checks were performed on the training residuals:

- Linearity:
  - Checked the Residuals vs Fitted plot.
  - Residuals were randomly scattered around zero → linear relationship holds.
- Independence of Errors:
  - Verified via Durbin–Watson ≈ 2.0 → no autocorrelation detected.
- Homoscedasticity:
  - Residuals showed constant variance in the Residuals vs Fitted plot.
  - Also supported by the Breusch–Pagan test.
- Normality of Residuals:
  - Residual histogram was approximately bell-shaped.
  - Q–Q plot points lay close to the 45° line → residuals roughly normal.
- No Multicollinearity:
  - Checked VIF values → all < 10 after refinement.

  All key assumptions — linearity, independence, homoscedasticity, normality, and no multicollinearity — were reasonably satisfied

Q5: How did you validate the assumptions of Linear Regression after building the model on the training set?
After training the model, the following checks were performed to validate linear regression assumptions:

- Linearity:
  - Verified using the Residuals vs Fitted plot — residuals were randomly scattered around zero, confirming a linear relationship.
- Independence of Errors:
  - The Durbin–Watson statistic (~2.0) indicated no autocorrelation among residuals.
- Homoscedasticity:
  - The spread of residuals remained fairly constant across fitted values, satisfying constant variance.
- Normality of Residuals:
  - Residuals followed a bell-shaped curve in the histogram, and the Q–Q plot showed points close to the 45° line.
- No Multicollinearity:
  - Checked using VIF — all final features had VIF < 10.

  All key assumptions — linearity, independence, homoscedasticity, normality, and no multicollinearity — were reasonably satisfied in the final model.

Q6: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

From the final regression model, the top three significant predictors of bike demand (cnt) are:

- Temperature (temp)
  - Has the strongest positive coefficient (~+116).
  - Higher temperatures lead to more bike rentals, as people prefer biking in pleasant weather.
- Year (yr_1)
  - Positive coefficient (~+1970).
  - Rentals in 2019 were significantly higher than in 2018, indicating increasing popularity of bike-sharing systems.
- Weather Situation (weathersit_light_snow_rain)
  - Strong negative coefficient (~–2005).
  - Rain or snow drastically reduces the number of rentals.

Bike demand increases with temperature and year (trend), while bad weather conditions sharply reduce it. These three features contribute most significantly to explaining daily demand patterns.

General Subjective Questions

Q1: Explain the linear regression algorithm in detail.

Linear Regression is a supervised learning algorithm used to model the relationship between a dependent variable (target) and one or more independent variables (predictors) by fitting a linear equation to the observed data.

- Mathematical Representation

  For Multiple Linear Regression, the model is expressed as:

  $y = \beta0 + \beta1x1 + \beta2x2 + \cdots + \beta nxn + \epsilon$

  Where:

  y → dependent (target) variable

  xi → independent variables

  β0 → intercept (value of y when all xi =0)

  βi → coefficients showing how each predictor affects y

  ϵ → random error term

- Objective: The goal is to find coefficient values (β_i) that **minimize the difference** between predicted and actual values, measured using the **Residual Sum of Squares (RSS):**

  $$RSS = \sum(y_i - \widehat{y_i})^2$$

The best-fit line minimizes this error.

- Steps Involved
  - Collect and preprocess the data (handle missing values, encode categoricals).
  - Split into training and test sets.
  - Fit the model by estimating coefficients via Ordinary Least Squares (OLS).
  - Evaluate the model using metrics like R^2, Adjusted R^2, RMSE.
  - Validate assumptions (linearity, independence, normality, homoscedasticity).
  - Interpret coefficients and refine the model if required.
- Advantages
  - Simple to implement and interpret.
  - Works well when relationships are approximately linear.
  - Provides clear coefficient-based insights.
- Limitations
  - Assumes linear relationships between variables.
  - Sensitive to outliers and multicollinearity.
  - Poor performance on nonlinear data.

  Linear Regression fits the best possible straight line through the data by minimizing the sum of squared residuals, helping explain and predict how independent variables influence the target variable.

Q2: Explain the Anscombe's quartet in detail.

Anscombe's Quartet is a set of four datasets created by statistician Francis Anscombe (1973) to highlight the importance of visualizing data and not relying solely on numerical summaries like mean, variance, or correlation.

- What It Contains
  Each dataset (I, II, III, IV) has:
  - Nearly identical statistical properties (mean, variance, correlation ≈ 0.816, and regression line).
  - Yet, when plotted, they look completely different and reveal very different relationships.
- Key Observations
  - Dataset I: Points form a clear linear relationship — linear regression works well.
  - Dataset II: Relationship is clearly non-linear — a straight line is not a good fit.
  - Dataset III: Contains an outlier that heavily distorts the regression line and correlation.
  - Dataset IV: All points lie on a vertical line except one extreme value, creating an artificially high correlation.
- Anscombe's Quartet demonstrates that:
  - Identical summary statistics can hide very different data patterns.
  - Visualization (scatterplots, histograms, etc.) is essential before modeling.
  - Outliers and non-linearity can mislead regression and correlation analysis.

Q3: What is Pearson's R?

Pearson's R, also known as the Pearson correlation coefficient, measures the strength and direction of the linear relationship between two continuous variables.

- Formula

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

Where:

- r→ Pearson correlation coefficient
- x_i, y_i→ individual paired observations
- $\bar{x}, \bar{y}$→ means of X and Y

- Interpretation
  - When r is positive, the variables move in the same direction — as X increases, Y also increases.
  - When r is negative, the variables move in opposite directions — as X increases, Y decreases.
  - An r close to +1 or –1 indicates a strong linear relationship.
  - An r near 0 means there is little or no linear correlation.

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

Scaling is the process of transforming numerical features so that they are within a specific range or distribution. It ensures that all variables contribute equally to the model rather than being dominated by features with large numeric values.

- Why is Scaling Performed?
  - Different features often have different units and magnitudes (e.g., temperature in °C vs windspeed in m/s).
  - Algorithms that depend on distance (like KNN, SVM, or gradient-based models) or matrix computations (like linear regression) can be biased toward larger-valued features.
  - Scaling helps models converge faster and maintains numerical stability in optimization.
- Normalization scales values between 0 and 1, while Standardization transforms features to have zero mean and unit variance, improving model stability and performance.

Q5: 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 — that is, when predictors are correlated with each other. Sometimes, the value of VIF becomes infinite or extremely large.
This happens when perfect multicollinearity exists among the independent variables.

- Reason:
  - **Perfect multicollinearity** means one variable can be **exactly predicted** by a combination of other variables.
  - In such cases, the denominator in the VIF formula becomes zero or very close to zero:
  - $VIF = \frac{1}{1-R^2}$

  If $R^2 = 1$(perfect correlation), then $1 - R^2 = 0$, making VIF infinite.

- Solution:
  - Remove one redundant dummy variable (use drop_first=True).
  - Avoid including variables that are linear combinations of others.

Q6: 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 a dataset (usually residuals) with a theoretical distribution — most commonly, the normal distribution.
It plots the quantiles of the sample data against the quantiles of a theoretical normal distribution.

- If the points fall approximately along the 45° reference line, the data follows a normal distribution.
- Deviations from the line indicate departures from normality (e.g., skewness or heavy tails).

Use in Linear Regression: In linear regression, a Q–Q plot is used to check the normality of residuals, which is one of the key assumptions of the OLS method.

- Purpose: Ensure that residuals (errors) are normally distributed.
- Why it matters:
  - Normality of residuals ensures valid hypothesis testing and reliable confidence intervals for coefficients.
  - Violations of this assumption may lead to biased p-values or unreliable inferences.
- Interpretation
  - Points along 45° line: residuals are approximately normal.
  - Curved pattern (S-shape): skewness (left or right).
  - Extreme deviations at ends: heavy tails (outliers).