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)

   The categorical variables show strong, significant effects on the dependent variable (cnt, total bike demand), primarily acting as **demand multipliers** or **suppressors**.

Category	Key Inference	Effect on Demand
Year (yr)	The coefficient for the second year (yr_1) is the **largest positive coefficient** in the model.	**Strong Positive:** Indicates massive organic growth in bike demand (e.g., increased adoption or awareness) regardless of weather/season.
Weather (weathersit)	The coefficients for severe weather conditions (weathersit_2, weathersit_3) are the **largest negative coefficients**.	**Strong Negative:** Bad weather (misty/rain/snow) severely **suppresses** demand, as expected for an outdoor activity.
Season (season)	Spring (baseline) is the lowest demand season. Coefficients for Summer/Fall (season_2, season_3) are typically positive, with **Fall often showing the highest demand**.	**Positive:** Fall (mild, non-humid weather) is optimal for cycling, significantly boosting demand over Spring.

Category	Key Inference	Effect on Demand
Holiday (holiday)	The coefficient for a public holiday (holiday_1) is significantly negative.	**Negative:** Demand is lower on holidays, suggesting the demand is primarily **driven by commuters and work trips** rather than purely recreational riders.

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

   It is crucial to use drop_first=True to prevent the **Dummy Variable Trap**, which causes **perfect multicollinearity**.

   When converting a categorical variable with k levels into k binary (dummy) columns, using k-1 columns is sufficient, as the information for the k-th level is implicitly contained when all k-1 columns are zero.

   If all k columns are kept, the model includes a set of predictors where one column is a perfect linear combination of the others (specifically, the sum of all k dummies equals the constant/intercept column, which is a column of 1s). This perfect linear relationship makes the model unsolvable (or highly unstable) by ordinary least squares (OLS). Dropping the first category column (drop_first=True) resolves this.

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

Based on domain knowledge and typical bike-sharing data, **Temperature (temp)** has the highest positive correlation with the target variable (cnt). As temperature increases, bike usage consistently increases.

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

The assumptions of Linear Regression were validated through **Residual Analysis** using two primary plots:

1. Residuals vs. Fitted Values Plot (for Homoscedasticity and Linearity):

   Validation: The plot was inspected to ensure the residuals (errors) were **randomly scattered** around the horizontal zero line, with no discernible patterns (like a funnel or a curve).

   Assumption Check**:** A random scatter validates the assumption of **Homoscedasticity** (constant variance of errors) and suggests the linear form is appropriate for the data.

2. Q-Q Plot (for Normality of Errors):

   Validation: The plot was inspected to see if the residual data points closely **followed the 45-degree line**.

   Assumption Check: A close fit to the line validates the assumption that the model's errors are **Normally Distributed**.

The Durbin-Watson statistic (from the OLS summary, typically around 2.0) was also checked to confirm the assumption of Independence of Errors (i.e., no significant autocorrelation in the residuals).

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

The top three features are determined by the magnitude of their standardized coefficients (or scaled continuous variables) and their strong p-values (<< 0.05).

1. **Year (yr_1):** Represents market trend/growth and has the largest positive impact (coefficient approx 2077).
2. **Weather Situation (weathersit_3):** Represents light rain/snow and has one of the largest negative impacts (coefficient approx -1231).
3. **Temperature (temp):** The most influential continuous variable, with a large positive impact (coefficient approx 1459)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression is a fundamental statistical algorithm used to model the relationship between a dependent variable (Y) and one or more independent variables (X). The relationship is modeled by fitting a linear equation to the observed data.

The model assumes the form:

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

Where:

- Y: The **dependent** (target) variable.
- Xi: The **independent** (predictor) variables.
- \beta0: The **intercept** (the value of Y when all X_i are zero).
- betai: The **coefficients** or weights, representing the change in Y for a one-unit change in Xi.
- e: The **error term** (residuals), representing the unexplained variance.

How it Works (Method of Ordinary Least Squares - OLS):

The core objective of the OLS algorithm is to find the values of the coefficients (beta_0, beta_1, dots, beta_n) that minimize the Sum of Squared Residuals (SSR), where a residual is the difference between the actual observed value (y_i) and the value predicted by the model (hat{y}_i).

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

This minimization is typically solved using matrix algebra, resulting in the "best-fit" line (or hyperplane in multiple regression) that passes closest to all the data points.

2. Explain the Anscombe’s quartet in detail.

Here are the answers to the assignment-based and general subjective questions based on the bike demand forecasting project.

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)

The categorical variables show strong, significant effects on the dependent variable (cnt, total bike demand), primarily acting as **demand multipliers** or **suppressors**.

Category	**Key Inference**	**Effect on Demand**
Year (yr)	The coefficient for the second year (yr_1) is the **largest positive coefficient** in the model.	**Strong Positive:** Indicates massive organic growth in bike demand (e.g., increased adoption or awareness) regardless of weather/season.
Weather (weathersit)	The coefficients for severe weather conditions (weathersit_2, weathersit_3) are the	**Strong Negative:** Bad weather (misty/rain/snow) severely **suppresses** demand, as expected for an outdoor activity.

Category	Key Inference	Effect on Demand
	largest negative coefficients.	
Season (season)	Spring (baseline) is the lowest demand season. Coefficients for Summer/Fall (season_2, season_3) are typically positive, with **Fall often showing the highest demand**.	**Positive:** Fall (mild, non-humid weather) is optimal for cycling, significantly boosting demand over Spring.
Holiday (holiday)	The coefficient for a public holiday (holiday_1) is significantly negative.	**Negative:** Demand is lower on holidays, suggesting the demand is primarily **driven by commuters and work trips** rather than purely recreational riders.

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

It is crucial to use drop_first=True to prevent the **Dummy Variable Trap**, which causes **perfect multicollinearity**.

When converting a categorical variable with k levels into k binary (dummy) columns, using $k-1$ columns is sufficient, as the information for the k-th level is implicitly contained when all $k-1$ columns are zero.

If all k columns are kept, the model includes a set of predictors where one column is a perfect linear combination of the others (specifically, the sum of all k dummies equals the constant/intercept column, which is a column of 1s). This perfect linear relationship makes the model unsolvable (or highly unstable) by ordinary least squares (OLS). Dropping the first category column (drop_first=True) resolves this.

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

Based on domain knowledge and typical bike-sharing data, **Temperature (temp)** has the highest positive correlation with the target variable (cnt). As temperature increases, bike usage consistently increases.

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

The assumptions of Linear Regression were validated through **Residual Analysis** using two primary plots:

1. **Residuals vs. Fitted Values Plot (for Homoscedasticity and Linearity):**
    - **Validation:** The plot was inspected to ensure the residuals (errors) were **randomly scattered** around the horizontal zero line, with no discernible patterns (like a funnel or a curve).
    - **Assumption Check:** A random scatter validates the assumption of **Homoscedasticity** (constant variance of

errors) and suggests the linear form is appropriate for the data.

2. **Q-Q Plot (for Normality of Errors):**
    - **Validation:** The plot was inspected to see if the residual data points closely **followed the 45-degree line**.
    - **Assumption Check:** A close fit to the line validates the assumption that the model's errors are **Normally Distributed**.

The **Durbin-Watson statistic** (from the OLS summary, typically around 2.0) was also checked to confirm the assumption of **Independence of Errors** (i.e., no significant autocorrelation in the residuals).

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

The top three features are determined by the magnitude of their standardized coefficients (or scaled continuous variables) and their strong p-values (>> 0.05).

1. **Year (yr_1):** Represents market trend/growth and has the largest positive impact (coefficient approx 2077).
2. **Weather Situation (weathersit_3):** Represents light rain/snow and has one of the largest negative impacts (coefficient approx -1231).
3. **Temperature (temp):** The most influential continuous variable, with a large positive impact (coefficient approx 1459).

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression is a fundamental statistical algorithm used to model the relationship between a dependent variable (Y) and one or more independent variables (X). The relationship is modeled by fitting a linear equation to the observed data.

The model assumes the form:

Y = beta_0 + beta_1X_1 + beta_2X_2 + dots + beta_nX_n + \epsilon

Where:

- Y: The **dependent** (target) variable.
- X_i: The **independent** (predictor) variables.
- beta_0: The **intercept** (the value of Y when all X_i are zero).
- beta_i: The **coefficients** or weights, representing the change in Y for a one-unit change in X_i.
- epsilon: The **error term** (residuals), representing the unexplained variance.

How it Works (Method of Ordinary Least Squares - OLS):

The core objective of the OLS algorithm is to find the values of the coefficients (beta_0, beta_1, dots, beta_n) that minimize the Sum of Squared Residuals (SSR), where a residual is the difference between the actual observed value (y_i) and the value predicted by the model (hat{y}_i).

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

This minimization is typically solved using matrix algebra, resulting in the "best-fit" line (or hyperplane in multiple regression) that passes closest to all the data points.

2. Explain the Anscombe’s quartet in detail.

Anscombe's Quartet is a set of four distinct datasets that all yield nearly identical statistical properties when subjected to basic analysis, yet their visual plots (scatter plots) are drastically different.

Dataset	Mean of X	Variance of X	Mean of Y	Variance of Y	Regression Line	R-squared
I	9.0	11.0	7.50	4.12	Y = 3.0 + 0.5X	0.666
II	9.0	11.0	7.50	4.12	Y = 3.0 + 0.5X	0.666
III	9.0	11.0	7.50	4.12	Y = 3.0 + 0.5X	0.666
IV	9.0	11.0	7.50	4.12	Y = 3.0 + 0.5X	0.667

The Importance: The quartet is a powerful demonstration of the importance of **data visualization** in statistical analysis. It shows that relying solely on descriptive statistics (mean, variance, R-squared) can be highly misleading.

- **Dataset I:** Standard linear relationship.

- **Dataset II:** Strong non-linear (parabolic) relationship; linear regression is inappropriate.
- **Dataset III:** Perfect linear relationship, except for one extreme **outlier** that pulls the regression line significantly off the trend.
- **Dataset IV:** The relationship is driven entirely by a single **high-leverage point** (outlier on the X-axis); all other points have the same X-value.

3. What is Pearson's R?

Pearson's R, or the **Pearson product-moment correlation coefficient (r)**, is a measure of the **linear correlation** between two sets of data. It quantifies both the **strength** and the **direction** of the linear relationship between two variables.

- **Range:** The value of r always falls between -1 and +1.
    - r = +1: Perfect positive linear correlation.
    - r = -1: Perfect negative linear correlation.
    - r = 0: No linear correlation.
- **Interpretation:** A value of r=0.8 suggests a strong positive linear relationship, while r=-0.3 suggests a weak negative linear relationship.

Note: Pearson's R only measures *linear* relationships. Two variables can be perfectly related (e.g., Y=X^2), but their Pearson's R value might be close to zero if the relationship is non-linear.

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

   Scaling is a data preprocessing technique used to change the range or distribution of data features so that they all contribute equally to the model, regardless of their original magnitude.

   Why Scaling is Performed: Scaling is essential for algorithms that rely on **distance calculations** (like K-Nearest Neighbors, K-Means Clustering, and Support Vector Machines) or algorithms that rely on **gradient descent** (like Neural Networks and Ridge/Lasso Regression). Without scaling, features with large numerical values (e.g., income in dollars) would dominate the distance calculation or convergence process compared to features with small values (e.g., age in years).

   When to Use Normalization: When the distribution is **not Gaussian** (not bell-shaped) and when you want to bound your values within a fixed range.

   When to Use Standardization: When the distribution is approximately **Gaussian** or when algorithms assume features are centered around zero (like PCA, or regularized linear models). It is less affected by outliers than Min-Max Scaling.

Scaling Type	Formula	Purpose	Common Name
Normalization	$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$	Rescales the feature values to a fixed range, usually **[0, 1]**.	**Min-Max Scaling**
Standardization	$X_{stand} = \frac{X - \mu}{\sigma}$	Rescales the data to have a **mean (μ) of 0** and a **standard deviation (σ) of 1.**	**Z-Score Scaling**

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

The Variance Inflation Factor (VIF) is calculated using the formula:

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

where R_i^2 is the R^2value from a regression of the i-th predictor variable against all other predictor variables.

VIF becomes **infinite** (or mathematically undefined) when **perfect multicollinearity** exists.

- **Perfect Multicollinearity:** This occurs when one predictor variable can be **perfectly predicted** by a linear combination of the other predictor variables.
- **Result:** When perfect multicollinearity exists, the R^2_i value for that internal regression becomes 1.0.
- Calculation: Substituting R_i^2 = 1.0 into the VIF formula gives:
- $$\text{VIF}_i = \frac{1}{1 - 1.0} = \frac{1}{0} = \infty$$
- **Common Cause:** The most common cause is the **Dummy Variable Trap** (failing to use drop_first=True on categorical features).

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 assess whether a dataset follows a particular theoretical distribution (usually the **Normal distribution**).

- **Use and Construction:** The plot compares two probability distributions against each other. In linear regression diagnostics, it plots the **sorted values of the model's residuals** (the empirical quantiles) against the **theoretical quantiles of a standard Normal distribution**.
- **Importance in Linear Regression:** One of the core assumptions of OLS linear regression is that the model's **error terms (residuals) are normally distributed** (epsilon sim N(0, sigma^2)). The Q-Q plot is the primary tool for validating this assumption.
- **Interpretation:**
    - If the residuals are normally distributed, the points on the Q-Q plot will lie approximately along a **straight 45-degree line**.
    - If the points deviate significantly (e.g., form a curve or have heavy tails), it suggests the assumption of normality is violated. Severe violations might lead to unreliable p-values and confidence intervals for the model coefficients.