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 in the dataset are:

- `season` (spring, summer, fall, winter)
- `weathersit` (clear, mist, light snow/rain, heavy rain/snow)

Inference:

- **Season** has a clear influence on demand. For example, **summer and fall** months typically see higher bike usage due to favorable weather conditions, while **winter and spring** months tend to have lower demand due to colder or unstable weather.
- **Weathersit** also impacts bike demand. As expected, **clear weather** shows the highest rental count, while **light snow or rain** and **heavy rain/snow** are associated with significantly lower demand. Bad weather discourages bike usage.

Thus, both `season` and `weathersit` show a **strong relationship with the target variable `cnt`** (bike demand), and must be included in the model as significant explanatory variables.

---

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

(2 marks)

Using `drop_first=True` when creating dummy variables helps to **avoid the "dummy variable trap"**, which occurs when one or more variables in a model are **highly collinear** (perfect multicollinearity).

For example, if we have four seasons and create four dummy variables (`spring`, `summer`, `fall`, `winter`), their sum will always be 1, making one variable redundant. This redundancy can:

- Lead to unstable or incorrect coefficient estimates.

- Make matrix inversion impossible during regression.

By setting `drop_first=True`, we drop one of the categories (e.g., spring), and treat it as a **baseline/reference**. The model then compares other categories against this baseline.

---

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

(1 mark)

From the pair-plot and correlation matrix, the numerical variable **`atemp`** (feeling temperature) shows the **highest positive correlation** with the target variable `cnt`.

This makes sense because warmer (but not too hot) temperatures generally encourage more people to use shared bikes. It better reflects perceived comfort than actual temperature (`temp`).

---

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

(3 marks)

The following steps were used to validate the assumptions of linear regression:

- **Linearity**: A scatter plot of predicted values vs. actual values (`y_pred` vs `y_test`) was used to check that the relationship is roughly linear.
- **Homoscedasticity (constant variance of residuals)**: A residuals plot (errors vs. predicted values) was used. The residuals should be randomly scattered around zero with no specific pattern.
- **Normality of residuals**: A histogram or Q-Q plot of residuals was plotted. For normality, the distribution should resemble a normal curve or fall approximately on a straight line in the Q-Q plot.
- **Multicollinearity**: The Variance Inflation Factor (VIF) can be computed for numerical variables. Variables with very high VIFs (e.g., >10) may be dropped to reduce multicollinearity.

Each of these checks helps ensure that the model's coefficients are reliable and the predictions are valid.

---

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

(3 marks)

Based on feature importance (coefficients after dummy encoding and fitting the model), the **top 3 contributing features** are typically:

1. **`atemp`** – Perceived temperature is the most positively correlated factor with bike rentals. As people feel more comfortable, they are more likely to rent bikes.

2. **`yr`** – The `yr` variable (0 for 2018, 1 for 2019) captures the time trend and shows that demand **increased** significantly from 2018 to 2019.

3. **`workingday`** – Days that are **not weekends or holidays** show higher demand, likely due to commuting behavior.

These features help explain both seasonal and long-term trends in shared bike demand.

1. Explain the linear regression algorithm in detail.

(4 marks)

Linear Regression is a **supervised learning algorithm** used for predicting a **continuous target variable** based on the values of one or more independent variables.

Linear regression tries to **model the relationship** between the dependent variable (`Y`) and one or more independent variables (`X`) by **fitting a linear equation**:

$Y=\beta0+\beta1X1+\beta2X2+...+\betanXn+\epsilon$

Where:

- Y: Target variable
- X1,X2,...,XnX_1, X_2, ..., X_nX1,X2,...,Xn: Independent variables
- β0\beta_0β0: Intercept

- β1,...,βn\beta_1, ..., \beta_nβ1,...,βn: Coefficients (slopes)
- ϵ\epsilonϵ: Error term

There are some key assumptions that we need to keep in mind or check before we apply linear regression algo on any data :

1.Linearity: Relationship between X and Y is linear.

2.Independence: Observations are independent.

3.Homoscedasticity: Constant variance of errors.

4.Normality: Residuals are normally distributed.

5.No multicollinearity among predictors.

Objective of linear regression is to minimize the **sum of squared residuals** (errors) between predicted and actual values using **Ordinary Least Squares (OLS)**:

There are 2 types of algorithm as below,

- **Simple Linear Regression** (1 feature)
- **Multiple Linear Regression** (multiple features)

Linear regression is widely used due to its **simplicity, interpretability**, and solid statistical foundation.

2. Explain the Anscombe's quartet in detail.

(3 marks)

Anscombe's quartet is a group of **four datasets** created by statistician **Francis Anscombe** to illustrate the importance of **visualizing data** before analyzing it statistically.

All four datasets have:

- Nearly identical **means**, **variances**, **correlations**, and **regression lines**
- But they have **very different distributions and patterns** when plotted

- Despite similar statistical summaries, each dataset shows a **different story** when visualized:
    - One follows a linear trend
    - One is curved
    - One has an outlier
    - One has a vertical line of identical Xs

Hence we understood that ,

Summary statistics can be misleading. Always visualize your data. Scatter plots can reveal anomalies, outliers, and incorrect assumptions.

3. What is Pearson's R?

(3 marks)

Pearson's R (correlation coefficient) measures the **linear relationship** between two continuous variables.

$r=(Cov(X,Y))/ \sigma X \sigma Y$

Where:

- Cov(X, Y): Covariance between X and Y
- $\sigma X, \sigma Y$\sigma_X, \sigma_Y$\sigma X, \sigma Y$: Standard deviations of X and Y

📈 **Interpretation:**

- $r=1$r = 1$r=1$: Perfect positive linear correlation
- $r=-1$r = -1$r=−1$: Perfect negative linear correlation
- $r=0$r = 0$r=0$: No linear correlation

Pearson's R is sensitive to outliers and assumes both variables are normally distributed.

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

(3 marks)

Scaling is the process of **transforming features to a similar range or distribution** to ensure fair contribution in a model (especially those that use distance or gradient descent).

Scaling is done to ,

- To improve **model convergence** (e.g., in gradient descent)
- To ensure features are **comparable**
- To avoid **bias** toward features with large numeric ranges

There are 2 types of scaling as below :

1.**Normalization** is useful for algorithms where bounded input is required (e.g., neural nets).

It uses min max scaling where it converts all values in the range of (0,1) or (-1,1)

2. Standardization is preferred when data is normally distributed or when dealing with outliers.

It uses z-score method for convergence where mean is 0 and standard deviation is 1.

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

(3 marks)

VIF (Variance Inflation Factor) measures how much a feature is **correlated with other features** (i.e., multicollinearity).

VIF(Xi)=11−Ri2\text{VIF}(X_i) = \frac{1}{1 - R_i^2}VIF(Xi)=1−Ri21

Where Ri2R_i^2Ri2 is the R-squared value from regressing feature XiX_iXi on all other features.

VIF Becomes Infinite When:

- Ri2=1R_i^2 = 1Ri2=1: The feature is **perfectly linearly dependent** on other features.
- This leads to **division by zero**, and the VIF becomes infinite.

Infinite VIF implies that ,

- **Severe multicollinearity** exists.
- The model will be unstable, coefficients unreliable.
- Solution: Remove or combine collinear features.

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

(3 marks)

A **Q-Q plot (Quantile-Quantile plot)** is a graphical tool to compare the **distribution of a dataset** with a theoretical distribution, usually the **normal distribution**.

In Linear Regression:

- Used to **check if residuals are normally distributed**, which is a key assumption of linear regression.

Interpretation:

- If points lie on the 45° line, residuals are normally distributed.
- Systematic deviations (curves or bends) indicate **non-normality**, suggesting potential model issues.

Importance:

- Validates assumption for **hypothesis testing, confidence intervals, and p-values**.
- Helps assess whether linear regression is an appropriate model.