Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

From my analysis of the categorical variables in the bike-sharing dataset, I observed the following key effects on the dependent variable (cnt):

- ✓ **Season** - Fall (season_4) showed the highest rentals, possibly because of comfortable temperatures and minimal rainfall. Winter (season_1) had the lowest rentals, as colder temperatures discourage cycling.
- ✓ **Year (yr)** - The positive coefficient indicates that bike rentals in 2019 were significantly higher than in 2018. This suggests that the bike-sharing program gained more users over time.
- ✓ **Weather Situation (weathersit)** - Poor weather conditions (rain, fog, snow) resulted in a drop in bike rentals
- ✓ **Month (mnth)** - Summer and Fall months (May–September) had peak rentals. This trend is closely related to temperature, where higher temperatures encourage biking.
- ✓ **Weekday** - Minimal impact on bike rentals, indicating relatively consistent usage throughout the week.
- ✓ **Holiday** - Bike demand was lower on holidays compared to non-holidays. We can assume that since bike-sharing is often used for commuting, fewer people rent bikes on holidays.

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

When creating dummy variables (one-hot encoding) for categorical features in a dataset like Bike Sharing, we often use drop_first=True. This helps to avoid the dummy variable trap and prevent multicollinearity in regression models.

The dummy variable trap occurs when we create N dummy variables for an N-category feature, leading to perfect multicollinearity in regression models.

Example: - Consider the season column with 4 categories:

1- Spring
2- Summer
3- Fall
4- Winter

When we do one-hot encode for season, we include all four dummy variables which leads to perfect multicollinearity, making the regression model unstable.

Setting drop_first=True removes one dummy variable (e.g., season_1), using it as a reference category.

The model now implicitly assumes that when all dummy variables are 0, the observation belongs to the dropped category (Spring).

Thus, setting drop_first=True removes multicollinearity, Keeps the model stable.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)

'temp' variable has the highest correlation with the target variable.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

I have validated the assumption of Linear Regression Model based on the following:

- ✓ **Linearity Assumption** (Target variable is linearly related to predictors).
  - Scatterplots of features vs. target - The plot showed a general linear trend between predictors (temp, hum, etc.) and cnt.
  - Residuals vs. Fitted (Predicted) Values Plot - showed a random scatter, confirming no strong non-linearity.
- ✓ **No Multicollinearity** (Predictors should not be highly correlated)
  - Correlation Heatmap - temp and atemp had a high correlation, so we dropped atemp to avoid multicollinearity
  - Variance Inflation Factor (VIF) Analysis - VIF scores were computed, and we removed high-VIF features to ensure independent predictors
- ✓ **Homoscedasticity** (Constant Variance of Residuals)
  - Residuals vs. Predicted Values Plot - showed a fairly even spread of residuals without a clear funnel shape. No strong heteroscedasticity was observed, meaning variance remained roughly constant
- ✓ **Normality of Residuals** (Residuals should follow a normal distribution)
  - Histogram of Residuals - The histogram of residuals was approximately bell-shaped, indicating near-normality
  - Q-Q Plot - showed residuals mostly aligning with the diagonal, confirming near-normal distribution.
- ✓ **No Autocorrelation of Residuals** (Residuals should not be correlated)
  - Durbin-Watson Test - was close to 2, indicating no serious autocorrelation in residuals

---

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

Based on the final OLS model summary and feature importance (coefficients & p-values), the top 3 significant features, driving bike demand (cnt) are:

- ✓ Temperature is the biggest driver of bike demand.
- ✓ Bike-sharing usage increased significantly in 2019 compared to 2018
- ✓ Bad weather negatively impacts rentals, discouraging people from cycling

---

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Linear regression is a fundamental supervised learning algorithm which is used to predict continuous values based on input features. It assumes a linear relationship between the dependent variable (target) and one or more independent variables (known as predictors).

Linear Regression is a powerful and widely used technique for predictive modeling, but it works best when its assumptions hold. If relationships are non-linear or multicollinearity exists, then it's best to look for alternative models.

There are two types of Linear Regression:
Simple Linear Regression – one independent variable
Multiple Linear Regression – Multiple independent variables

Mathematical expression is – Y = B0 + B1X1 + B2X2 + …+ BnXn + error

Assumptions of Linear Regression are –
Linearity – There should be a linear relationship between the predictors and the target variable
No Multicollinearity – Predictors should not be highly correlated. If the Variance Inflation Factor (VIF) is greater than 5, then that variable should be removed
Homoscedasticity – Residuals should have constant variance across all predicted values. This can be checked through Residual vs. Predicted Plot
No Autocorrelation – Residual should be independent. This can be checked through Durbin-Watson test
Normality of Residual – Residuals should be normally distributed. This can be checked through Q-Q plots

---

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Anscombe's Quartet was introduced by statistician Francis Anscombe in 1973 which consists of four

different datasets that have identical summary statistics but exhibit very different distributions when plotted graphically.
It highlights the importance of **data visualization** in statistical analysis rather than relying solely on summary statistics like mean, variance, correlation, or regression lines.

Each dataset consists of 11 (x, y) points, and all four share the following identical statistical properties:

Statistic	Value (for all 4 datasets)
Mean of x	9.0
Mean of y	7.5
Variance of x	10.0
Variance of y	3.75
Correlation (x, y)	0.816
Linear Regression Equation	y = 3.00 + 0.50x

At first glance, these datasets appear to be statistically identical. However, plotting them reveals completely different distributions.

When graphical analysis was done, different distributions were observed as mentioned below

- ✓ **Dataset 1**: Linear Relationship - A classic linear relationship between x and y
- ✓ **Dataset 2**: Non-Linear Relationship - Curved relationship between x and y (quadratic).
- ✓ **Dataset 3**: Strong Influence of Outliers - Data follows a linear trend except for one extreme outlier.
- ✓ **Dataset 4**: Vertical Outlier (Leverage Point) - Most data points have the same x value.

So, key take away is - Summary statistics are not enough – Always visualize your data before drawing conclusions.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Pearson's R (also known as Pearson Correlation Coefficient) is a statistical measure that quantifies the linear relationship between two continuous variables. It is denoted by r and ranges from -1 to +1.
Pearson's r helps us understand the strength and direction of relationships between numerical variables. In our Bike Sharing dataset, it helped us:

- ✓ Drop highly correlated features (atemp due to multicollinearity).
- ✓ Identify key predictors for bike rentals (temp, humidity, windspeed).
- ✓ Improve our regression model.

How to interpret Pearson's R

Pearson's r Value	Relationship Strength	Interpretation
+1	Perfect Positive Correlation	As one variable increases, the other increases perfectly.
+0.7 to +0.9	Strong Positive Correlation	Variables are strongly related
+0.3 to +0.7	Moderate Positive Correlation	Some relationship exists.
0	No Correlation	No linear relationship between variables.

-0.3 to -0.7	Moderate Negative Correlation	Some inverse relationship exists.
-0.7 to -0.9	Strong Negative Correlation	A strong inverse relationship exists.
-1	Perfect Negative Correlation	As one variable increases, the other decreases perfectly.

Pearson's R works well for linear relationships; however, it has some limitations like it does not capture non-linear relationships. It is affected by outliers and only measures correlation, not causation.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Scaling is the process of transforming numerical features into a dataset to ensure that they are on a similar scale, preventing certain features from dominating the model due to their large values.
For example, If a dataset has (different scaling values)

- Age ranging from 0 to 100
- Salary ranging from Rs 10000 to Rs 1000000

A model would assign more importance to Salary simply because of its larger magnitude, even if Age is equally important. Scaling ensures that all features contribute equally to the model.

Why scaling is performed –

- ✓ Improves Model Performance - Prevents certain features from dominating the learning process.
- ✓ It Enhances Numerical Stability - Reduces Computational Errors
- ✓ Speeds Up Training - Many ML algorithms converge faster with scaled data.
- ✓ Prevents Bias Due to Feature Magnitude - Ensures fair weight assignment in models like regression and neural networks.

Difference between normalized scaling and standardized scaling

Normalized scaling	Standardized scaling
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are different values	It is used when we want to insure zero mean and unit standard deviation
Scales values between [0,1] or [-1,1].	It is not bound to a certain range
It is really affected by outliers	It is much less affected by outliers
Scikit-Learn provides a transformer called MinMaxScaler for Normalization	Scikit-Learn provides a transformer called StandardScaler for standardization

Use Normalization if data is not normally distributed and needs to be bound (0-1 range).
Use Standardization if data follows a Gaussian distribution and you need zero mean and unit variance

---

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

The main job of Variance Inflation Factor (VIF) is to measure how much a predictor variable is correlated with other predictor variables in a regression model.

A VIF value becomes infinite when the denominator of the formula ($1 - R^2$) becomes zero, meaning that $R^2=1$. This happens when:

- One predictor variable is a perfect linear combination of other predictor(s)
- when one-hot encoding is applied to a categorical variable without dropping one category
- If a dataset already has an intercept column (a constant column of 1s), adding another constant manually in statsmodels will create redundancy
- If two or more columns are exact duplicates of each other, their correlation is 100%, making VIF infinite (like temp and atemp)

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

A Q-Q Plot (Quantile-Quantile Plot) is a graphical tool which is used to assess whether a dataset follows a particular distribution, mostly normal distribution. It compares the quantiles of a dataset against the quantiles of a theoretical normal distribution and if the data follows a normal distribution, the points will align closely along the diagonal reference line.

To interpret the Q-Q plot, it consists of X-axis (Theoretical quantiles - expected values from a normal distribution) and Y-axis (Sample quantiles - actual data points)

Observation	Interpretation
Points fall along the 45° line	Data is normally distributed
S-shaped curve	Data is skewed (left/right)
Downward curve at ends	Data has light tails (less variance)
Upward curve at ends	Data has heavy tails (outliers)

Use and importance of a Q-Q plot in linear regression –

In linear regression, one of the key assumptions is that the residuals (errors) follow a normal distribution. A Q-Q plot helps check this assumption. In linear regression, Normality is very important because if residuals are normally distributed, regression inferences (p-values, confidence intervals) are reliable. But, If the Q-Q plot shows deviations, transformations (log, square root) may be needed to improve model validity

---