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)

1) Fall season have more bike users compared to others
2) Users are more in year 2019 compared to 2018
3) Mid of the year have users peaked. Start and end of year have less count for users compared to mid-year
4) Bookings are more in non-holidays
5) Wed, Thur, Fri and Sat shows good number of bike users. Sun shows a lesser count
6) Working day have more user count as expected
7) Clear weather has more user count compared to others

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)

During Dummy variable for example if the categorical column has values as A, B and C. It can be represented with two values like if it is not A or B, it is C and also if it A or B, it won't be C. So, the value C is not required and can be dropped ideally

---

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 have the highest correlation with the target cnt 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)

1) After building the model verified that all the VIF values for the feature selected have a value below 5

2) Validated the P value for all the feature variable selected have a value less than 0.01

3) Validated the R^2 value is good

4) Validated F static value is greater than 1 and also prob of F-static is almost equal to zero

5) Residual analysis was performed and it was validated that distribution is normal and mean value for error is zero

6) Plotted heatmap to confirm and then validate that no multi collinearity is observed

7) Value of 2.1 for Durbin Watson value indicate that there is little to no auto correlation in the residuals which is a good sign.

8) Residual values are not displaying any specific patterns

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)

Below are the top3 features significantly contributing towards Bike demand

1) temp
2) Light_SnowRain
3) yr

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)

<Your answer for Question 6 goes here>

Linear Regression is one of the simplest and most widely used supervised learning algorithms for predicting a continuous dependent variable based on one or more independent variables.
It establishes a linear relationship between the dependent variable Y and independent variable X

Linear Regression is mainly used in predictive modeling and statistical analysis, such as:

- Predicting sales based on advertisement spending
- Estimating house prices based on features like area, number of bedrooms, etc.
- Forecasting stock prices, revenue growth, etc.

1. Types of Linear Regression

A. Simple Linear Regression
In Simple Linear Regression, there is one independent variable X and one dependent variable Y

The relationship is expressed as:

$Y= \beta 0X + \beta 1 + e$

Y = Target variable
X = independent variable
β0 = intercept (Value of Y when X= 0)
β1 = slope (Shows how Y changes when X changes by 1 unit)
e =Error term

B. Multiple Linear Regression
In Multiple Linear Regression, there are multiple independent variables

Used when multiple factors affect the outcome.
Y= β0X + β1X1 + β2X2+....+ βnXn +e

2. How Linear Regression Works

Step 1: Assumptions of Linear Regression
Before applying linear regression, ensure the following assumptions hold:

a) Linearity – The relationship between dependent and independent variables should be linear.
b) Independence – Observations should be independent of each other.
c) Homoscedasticity – The variance of residuals should be constant across all levels of X
d) Normality of Residuals – The residuals should be normally distributed.
e) No Multicollinearity – Independent variables should not be highly correlated.

Step 2: Finding the Best-Fit Line
The goal of Linear Regression is to find the best-fit line that minimizes the error between actual values and predicted values.
This is done using the Ordinary Least Squares (OLS) method, which minimizes the sum of squared errors (SSE):

Step 3: Evaluating Model Performance
Once the model is trained, its performance is assessed using various metrics:

a) R^2 value - Measures how much variance in Y is explained by X-axis (The value range 0 to 1)
b) Mean squared Error - Measures the average squared difference between actual and predicted values
c)Durbin-Watson Test: Used to detect autocorrelation in residuals.
Ideal value = 2 (no autocorrelation).

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)

<Your answer for Question 7 goes here>

Anscombe's Quartet is a set of four different datasets that have nearly identical summary statistics (mean, variance, correlation, regression line) but vastly different distributions when plotted.
It was created by Francis Anscombe in 1973 to highlight the importance of visualizing data rather than relying solely on summary statistics.

The Four Datasets and Their Behavior

Dataset 1 (Linear Relationship)
Forms a perfect linear relationship with some natural scatter.
Regression line is a good fit.

Dataset 2 (Non-linear Relationship)

The data follows a curved (quadratic) relationship, not a straight line.
A linear regression model is inappropriate.

Dataset 3 (Influence of an Outlier)
Most data points follow a linear trend, but a single outlier significantly affects the regression line.
Shows how outliers can distort analysis.

Dataset 4 (Vertical Outlier)
Almost all points are constant in x, except for one extreme outlier.
The correlation coefficient remains the same, but the regression line is misleading.

Anscombe’s Quartet is Important because of following reason

1)Summary statistics alone can be misleading:
If you only look at means, variances, and correlations, you might conclude that all four datasets are similar.

2)Data visualization is crucial:
Simply plotting the data reveals patterns, relationships, and outliers that summary statistics fail to capture.

3)Linear regression is not always the right choice:
If data is non-linear or has outliers, a linear regression model may not work well.

4)Outliers can drastically affect regression and correlation:
One extreme point can change the entire regression line, leading to incorrect interpretations.

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)

<Your answer for Question 8 goes here>

Pearson’s R, also known as the Pearson correlation coefficient (r), is a statistical measure that quantifies the linear relationship between two continuous variables.
It determines how strongly and in what direction the two variables are related.

Values for coefficient r and its indication

r =1 - positive correlation
0.7<=r <1 - Strong positive correlation
0.3 <=r <0.7 - moderate positive correlation
0 <=r <0.3 - weak positive correlation
r = 0 - no correlation
-0.3 < r < =0 - weak negative correlation
-0.7 < r < = -0.3 - Moderate negative correlation
-1 < r < = -0.7 - Strong negative correlation

r = -1 - perfect negative correlation

Positive - As X increases, Y increases
Negative - As X increases, Y decreases
r =0 No linear relationship between X and Y

Pearson R can be used in below cases

1)Both variables are continuous (numeric).
2)The relationship is linear. (Check using scatter plots.)
3)Data is normally distributed.

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)

<Your answer for Question 9 goes here>

Scaling is a data preprocessing technique used to transform numerical features so they have a similar range.
This is important because many machine learning algorithms perform better when features are on the same scale.

Scaling is important for below reason

a) Prevents bias towards larger numbers: Algorithms like Gradient Descent, K-Nearest Neighbors (KNN), and Support Vector Machines (SVM) work better when features have a uniform scale
b) Improves model convergence: Optimization algorithms like Stochastic Gradient Descent (SGD) converge faster when features are scaled
c) Avoids dominance of certain features: Some models (like Linear Regression) may assign higher importance to larger features, distorting results.

Difference Between Normalization and Standardization

There are two main techniques for scaling:

a) Normalization (Min-Max Scaling)
b) Standardization (Z-score Scaling)

Normalization

a) Formula:
Normalized value = (X -Xmin)/ (Xmax-Xmin)

b) Rescales the values between 0 and 1.
c) Maintains the original distribution but shifts it to a smaller range.

---

d) Sensitive to outliers since it depends on Xmax and Xmin

Standardization (Z-score Scaling)

a) Formula
Standard Value = (X -Mean)/ (standard deviation)
b) Transforms data so that mean = 0 and standard deviation = 1.
c) Works well even when data has outliers.
d) Useful when the data follows a Gaussian (Normal) distribution.

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)

<Your answer for Question 10 goes here>

The Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in a regression model.
When VIF becomes infinite (or extremely large), it indicates a serious multicollinearity issue.
This usually happens due to one of the following reasons:

1. Perfect Multicollinearity (Exact Linear Dependency)
If one independent variable is a perfect linear combination of other variables, then VIF will be infinite.

2. Dummy Variable Trap (One-Hot Encoding Issues)
When performing one-hot encoding (converting categorical variables into binary columns), including all categories can cause perfect multicollinearity.

3. Highly Correlated Variables (Severe Multicollinearity)
If two or more independent variables are very strongly correlated (not necessarily perfectly correlated), VIF values can become extremely large, approaching infinity.

4. Insufficient Data (More Features than Samples)
When the number of independent variables exceeds or is close to the number of observations, the matrix inversion in regression calculations fails, leading to infinite VIF.

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)

<Your answer for Question 11 goes here>

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

It helps to visually assess whether a variable follows a given distribution by plotting the quantiles of the sample data against the quantiles of the theoretical distribution.

Reading a Q plot

X-axis: Theoretical quantiles (expected values if data were normally distributed).

Y-axis: Observed quantiles (actual values from your dataset).

Interpretation:

Data points lie along the 45-degree diagonal line → The data follows a normal distribution.

Data points deviate significantly from the line → The data does not follow a normal distribution.

S-shaped curve → Possible skewness (right or left).

Upward/downward bending at ends → Heavy-tailed (leptokurtic) or light-tailed (platykurtic) distribution.

Importance of Q-Q Plot in Linear Regression

In linear regression, one key assumption is that residuals (errors) should be normally distributed.

A Q-Q plot is used to check this assumption.

It is important for below reason

a) Validates statistical tests:

Many hypothesis tests (e.g., t-tests, F-tests, confidence intervals) assume normality of residuals.

b) If residuals are not normally distributed, p-values might be unreliable.

c) Ensures accurate model predictions:

If residuals are normal, predictions and confidence intervals are more reliable.

d) Detects outliers & skewness:

If points deviate at the ends, it indicates heavy tails (outliers).

If points curve, the data might be skewed.

Way to use Q-Q plot in Linear Regression

After fitting a linear regression model, extract the residuals.

Plot the Q-Q plot to check if residuals follow a normal distribution.

If residuals are normally distributed, regression assumptions hold, and inference is reliable.