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?

Categorical variables in the dataset are - season, yr, mnth, weekday, holiday (binary), workingday (binary), weatherist. Target variable – cnt

season: median close to 55000, season 3 takes more booking, followed by season 2 and season 4. Out of all season 1, we have very less demand.
mnth: more booking in the months of 6, 7, 8, 9, 10 with almost the same median value. Of all considerable booking from April to November, higher demand
holiday: 3000 to 6000 bikes booked. Mean shows little above to 4000 bikes. And even on holidays, there is more booking. But median stands less than 4000.
weekday: the median lies between 4000 to 4500 and more booking on wednesdays. Very close trend each day as compared all the days in the week.
workingday: the median stands same for working day and NOT a working a day. But working day spread more even & perfect. Seems to be a good variable
weathersit: very look demand in weathersit-3. More demand in weathersit-1 then comes weather-2. Varying median in all the weathersit 1,2,3

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

When we are encoding the categorical variable, we use the method pd.get_dummies(dataframe)
Assume we have a categorical variable as gender with 2 categories Male and Female.
Using get_dummies() method it creates two columns as male and female and marks “1” in the data relevantly for each member record. If drop_first is set to True, it deletes one column and still the data holds good specify “1” or “0”. So for “n” categories, it is “n-1” columns. This helps in reduction of features making the model easier and does not lead to multi-collinearity. Again, keeping drop_first=True is a still an option based on the data.

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

It is very clear from the pair-plot, that the highest correlation is with atemp, temp with the target variable cnt. The dots show very close and progressive. They are not that scattered for these two features.

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

There few different ways with which we can validate the assumptions of a Linear Regression model.

a. **Homoscedasticity**: all the residuals will have constant variance. Plotting these residuals and checking the spread. If the spread increases or decreases systematically, there may be an issue
b. **Multicollinearity**: Calculate Variance Inflation Factors (VIF) for each variable. VIF values of the independent variables should not be high or are not highly correlated. VIF values above to 10 or 5 (depends on how much we take), may indicate multi-collinearity.
c. **Linearity**: Independent variables and the dependent variable is always linear and this can be viewed using scatter plots. There should not be a pattern.
d. **Residuals**: There should not exist any pattern when these residuals are marked in the scatter plot.

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

In general, lower multicollinearity is generally desirable, VIF values measure the extent of multicollinearity for each feature. Lower VIF values generally indicate lower multicollinearity. From the model that I have worked (Bike_Sharing_MLR.ipynb)

a. weathersit_3: 1.05
b. weathersit_2: 1.06
c. workingday: 1.01

General Subjective Questions

1	Explain the linear regression algorithm in detail.

Linear Regression is a statistical method used for modelling the relationship between a dependent variable (target) and one or more independent variables (features or predictors). It assumes that the relationship between the variables is linear, represented by a straight line. The goal of linear regression is to find the best-fitting line that minimizes the difference between the predicted and actual values of the dependent variable.

Simple Linear Regression:
Works with one independent variable

Mathematical equation: y = mx + b
Where, y: dependent variable (target), x: Independent variable (feature/column),
m: Slope of the line and b: Y-intercept

Multiple Linear Regression:
Extension of Simple Linear Regression to multiple independent variables.

Mathematical equation:
y = b0 + b1 * X1 + b2 * X2 + b3 * X3 + …… + bn * xn
b0: Y-intercept, y: dependent variable (target),
x1 to xn: Independent variables (features/columns)
b1 to bn: coefficients of x1, x2,….,xn

Assumptions of Linear Regression model to validate: There few different ways with which we can validate the assumptions of a Linear Regression model.

a. **Homoscedasticity**: all the residuals will have constant variance. Plotting these residuals and checking the spread. If the spread increases or decreases systematically, there may be an issue
b. **Multicollinearity**: Calculate Variance Inflation Factors (VIF) for each variable. VIF values of the independent variables should not be high or are not highly correlated. VIF values above to 10 or 5 (depends on how much we take), may indicate multi-collinearity.
c. **Linearity**: Independent variables and the dependent variable is always linear and this can be viewed using scatter plots. There should not be a pattern.
d. **Residuals**: There should not exist any pattern when these residuals are marked in the scatter plot.

Evaluation Metrics:

a. Mean Squared Error (MSE)
b. Root Mean Squared Error (RMSE)
c. Mean Absolute Error (MAE) and (coefficient of determination).

Cost Function:
Cost function measures the performance of a machine learning model for a data set. Cost function quantifies the error between predicted and expected values and presents that error in the form of a single real number

$$\text{Cost Function (J)} = \frac{1}{n}\sum_{i=0}^{n}\left(h_{\theta}\left(x^{i}\right) - y^{i}\right)^{2}$$

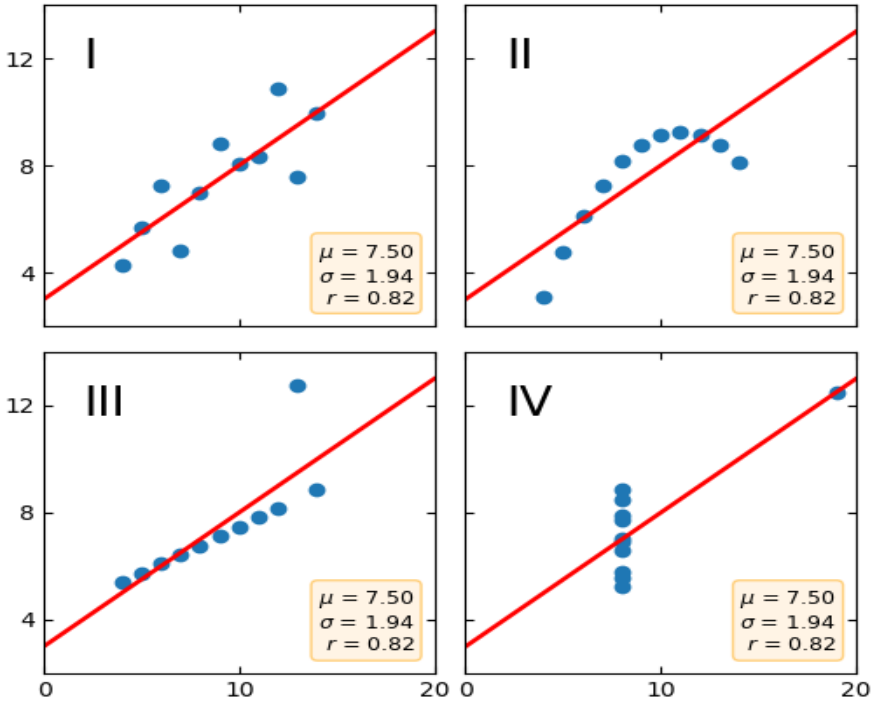
n is the number of training events.

h(theta) is the hypothesis function of the predicted value
y is the predicted value of the dependent variable

Gradient Descent: You have to find the direction in which the error decreases constantly. This can be done by finding the difference between errors. The small difference between errors can be obtained by differentiating the cost function and subtracting it from the previous gradient descent to move down the slope

2	Explain the Anscombe's quartet in detail.

It is a set of four datasets (x, y) that have nearly identical simple descriptive statistics (mean, variance, correlation, and linear regression) but differ significantly when visualized. To illustrate the importance of data visualization in understanding the underlying patterns in a dataset and understand the patterns and relationship.
Outliers and influential points can have a significant impact on regression analysis. Even if datasets have similar summary statistics, they may exhibit different structures when visualized. Anscombe's quartet serves as a powerful reminder that relying solely on summary statistics can be misleading, and exploring data through visualization is essential for a comprehensive understanding. It highlights the importance of data exploration and graphical representation in statistical analysis.
This picture gives the patterns on how the datasets visualisation differs with the 4 metrics.



3	What is Pearson's R?

Denoted by r, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. The degree to which two variables tend to increase or decrease together, in a linear way. r is sensitive to outliers and assumes that the relationship between variables is linear.

It ranges from -1 to 1.

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

The formula for Pearson's correlation coefficient between variables X and Y in a dataset is given by:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \cdot \sum (Y_i - \bar{Y})^2}}$$

Where:

- X_i and Y_i are the individual data points in the variables X and Y.
- $\bar{X}$ and $\bar{Y}$ are the means of variables X and Y, respectively.

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

Scaling: Standardizing the numerical values of variables in a dataset to a common scale. The goal of scaling is to ensure that all features contribute equally to the computation of distances, similarities, and model parameters, especially in algorithms that are sensitive to the scale of the input features.

Min-Max Scaling (Normalized): Transforms data to a specific range, usually between 0 and 1. Useful when the data distribution does not follow a normal distribution

Formula: $X_{\text{scaled}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$

Standardized scaling (Z-score Normalization): Transforms data to have a mean of 0 and a standard deviation of 1. Preserves the shape of the original distribution and is suitable for algorithms that assume normally distributed features.

Formula: $X_{\text{scaled}} = \frac{X - \text{mean}}{\text{standard deviation}}$

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

VIF is infinite:
Yes, if you see my Bike_Share_MLR.ipynb, there are bunch of features that I have removed who VIF values are seen as infinite 'inf'. I had removed them after the first model run.
Normally, a high VIF is typically a concern, an infinite VIF can occur when there is perfect multicollinearity among the predictor variables. Perfect multicollinearity arises when one predictor variable in the model can be exactly predicted by a linear combination of other predictor variables. A good mathematical way for perfect multicollinearity is that the determinant of the matrix of predictor variables is zero, where the inverse of the matrix cannot be computed. When this happens, VIF calculations involve dividing by zero, resulting in an infinite VIF value.

Why does this happen?
Linear dependency between two independent variables
Dummy variables, perfect multicollinearity can occur when one dummy variable can be exactly predicted from other dummy variables. This often happens when including all levels of a categorical variable without omitting one reference category (and this exactly happened to me in the LR assignment, with the dummy columns)

How to eradicate?
Verify the data issues, removal of redundant variables, variable transformation (combining variables into one, to lessen the number of features, a good technique), and using Regularization eliminating multi-collinearity.

6	What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression

Q-Q (Quantile-Quantile) plot is a graphical tool used in statistics to assess whether a dataset follows a specific theoretical distribution, such as the normal distribution

This is used to –
Generate Quantiles, Plot Quantiles, Pattern or Distribution Analysis and Interpretation.

If points in the Q-Q plot follow a straight line, it suggests that the data follows the theoretical distribution
S-shaped curves or bends in the Q-Q plot may indicate non-normality or other departures from the theoretical distribution
If the points deviate upward or downward at the tails, it may indicate heavy or light tails compared to the normal distribution

Shapes of Q-Q plot:
Straight Line: Indicates a good fit between the observed data and the theoretical distribution.
S-Shaped Curve: Indicates non-normality or a departure from the assumed distribution.

Points at Tails: Deviations at the tails can provide insights into the distribution's tails compared to the expected normal distribution

In Linear Regression Q-Q plot:

a. For assessing the normality assumption of the residuals. The residuals are the differences between the observed values and the values predicted by the linear regression model.
b. If the residuals are normally distributed, it implies that the statistical inferences (such as confidence intervals and hypothesis tests) based on the linear regression model are valid
c. Deviations from a straight line may indicate non-normality, skewness, or other distributional issues with the residuals.
d. Helps identify outliers in the residuals. Outliers may manifest as points deviating significantly from the expected normal distribution
e. Q-Q plot reveals substantial departures from normality, it may prompt a re-evaluation of the linear regression model