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. Weather situation has a major role in determining the target variable. Ex- normal, cloudy.
2. Season has a major role in determining the target variable. Ex- Spring, Winter.
3. Year has a major role in determining the target variable. Ex- 2018, 2019.
4. Month also has some role in determining the target variable. Ex- Sep

---

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)

1. Prevents multicollinearity by avoiding the perfect multicollinearity. Ex – if there are 2 dummy columns without drop_first=Ture, then when one column has 1 then other will be zero & this has a colinearly of -1. It can be avoided by dropping 1 column.
2. Increases model efficiency by reducing the number of estimated parameters.
3. Improves interpretation by setting a reference category, which allows you to interpret the coefficients of the other categories relative to this baseline.

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)

1. Registered has highest correlation with the target variable (considering registered & casual).
2. atemp has highest correlation with target variable (when registered & casual are removed).

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. Linear relationship - Coefficient of independent variables mean there is linear relationship between independent & target variable. Also, p-value < 0.05 means that the coefficient is likely not to equal zero.
2. No Multicollinearity – by removing the variables that are highly correlated to other variables. We have used VIF values for finding the variables that are highly correlated & dropping those variables.

3. Normality of residuals - it is assumed that the error term is normally distributed, we can see it by plotting histplot of residuals.
4. Homoscedacity - The variance of the residuals is constant across all values of the independent variable. Its can be observed by plotting residuals vs y predicted value.

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)

1. Season –> Spring – negative effect
2. Weather situation –> Normal – positive effect
3. Year – positive effect

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)

a) Linear regression – it is a method used to model the linear relationship between a dependent (target) variable and one or more independent (predictor) variables. linear relationship means the relationship between target variable (y) & independent variable (x) can be represented by a straight line (i.e., y=mx+c).

b) The general equation for linear regression is:

$$y = b0 + b1*x1 + b2*x2 + ... + bn*xn$$

where:
y: The dependent variable (the value we want to predict)
b0: The intercept (y value when xi = 0)
b1, b2, ..., bn: The coefficients (the weights or importance of each independent variable)
x1, x2, ..., xn: The independent variables

c) Assumptions of Linear Regression
1. Linearity - The relationship between the dependent and independent variables is linear.
2. Independence - The observations are independent of each other (i.e. each row is independent of other row).
3. Homoscedasticity - The variance of the errors is constant across the independent variables.
4. Normality - The errors follow a normal distribution, with mean at zero.
5. No multicollinearity - The independent variables are not highly correlated with each other.
6. No endogeneity - There is no relationship between the errors and the independent variables.

d) Objective of Linear Regression
The objective of linear regression is to find the best-fitting line (optimum values of b0,b1, …bn) that minimizes the difference between the observed values and the predicted values & this is called the residual (error or cost function).

e) Gradient Descent Approach

we use gradient descent to find the optimal coefficients for the independent variables. Gradient descent iteratively updates the coefficients by moving in the direction of the gradient that reduces the cost function.

Cost function = $[\sum \boldsymbol{y_i} - \widehat{\boldsymbol{y}}_{\boldsymbol{i}}]^{\mathbf{2}}$

Where:

- is the cost function (Residual sum of squares).
- α is the learning rate (controls how large the steps are).
- bi+1 = bi - α*∇ are the coefficients.

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 comprises a set of four datasets, having identical statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when plotted on a graph.

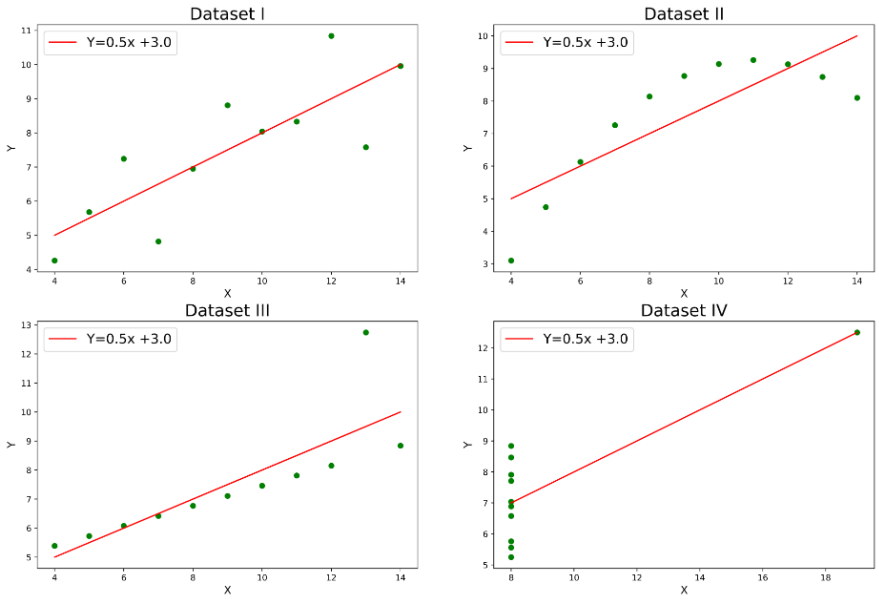
Ex) consider the follow dataset -

	x1	x2	x3	x4	y1	y2	y3	y4
0	10	10	10	8	8.04	9.14	7.46	6.58
1	8	8	8	8	6.95	8.14	6.77	5.76
2	13	13	13	8	7.58	8.74	12.74	7.71
3	9	9	9	8	8.81	8.77	7.11	8.84
4	11	11	11	8	8.33	9.26	7.81	8.47
5	14	14	14	8	9.96	8.10	8.84	7.04
6	6	6	6	8	7.24	6.13	6.08	5.25
7	4	4	4	19	4.26	3.10	5.39	12.50
8	12	12	12	8	10.84	9.13	8.15	5.56
9	7	7	7	8	4.82	7.26	6.42	7.91
10	5	5	5	8	5.68	4.74	5.73	6.89

When we find the statical parameters of the above data set we find results as follows & we feel that all the data set are almost the same –

	I	II	III	IV
Mean_x	9.000000	9.000000	9.000000	9.000000
Variance_x	11.000000	11.000000	11.000000	11.000000
Mean_y	7.500909	7.500909	7.500000	7.500909
Variance_y	4.127269	4.127629	4.122620	4.123249
Correlation	0.816421	0.816237	0.816287	0.816521
Linear Regression slope	0.500091	0.500000	0.499727	0.499909
Linear Regression intercept	3.000091	3.000909	3.002455	3.001727

When we plot a scatter plot of the above data set, then we observe data set are different –



Data-set 1 - shows a linear relationship between x & y with some variance.
Data-set 2 - shows a nonlinear relationship between x & y.
Data-set 3 - looks like a linear relationship between x and y, except for one large outlier.
Data-set 4 - looks like the value of x remains constant, except for one outlier.

Hence, we can conclude that just the statical summary of the dataset is not enough, we need to use graphical representation in combination of statical summary to make 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 correlation coefficient (r) measures the strength and direction of the linear relationship between two variables. It ranges from -1 to 1:

- -1 <= r < 0: negative correlation
- 0: No correlation
- 0< r <= +1: positive correlation

$$r = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2}\sqrt{(Y-\bar{Y})^2}} \qquad r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2]\,[n\sum y^2 - (\sum y)^2]}}$$

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 a process of transforming the data within a specific range. It is a critical step in data preprocessing, as many algorithms perform better when the data is scaled properly. It also helps to ensure that features are not weighted improperly.
Scaling Important –

- **Uniform Feature Importance:** If features have different scales, the algorithm may give more importance to features with larger values, causing bias.
- **Gradient-Based Optimization**: Algorithms such as linear regression, logistic regression, and neural networks use gradient descent, if features are on different scales, the gradient descent may converge slowly & requiring more iterations.
- **Outliers treatment:** Scaling can help in handling the outliers better. For example, standardization is less affected by outliers.

S.NO.	Normalization	Standardization
1	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2	used when features are of different scales.	used when we want to ensure zero mean and unit standard deviation.
3	Scales values between [0, 1].	It is not bounded to a certain range.
4	It is really affected by outliers.	It is much less affected by outliers.
7	Retains the shape of the original distribution	Changes the shape of the original distribution
8	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.
9	Xnormal = (x – min)/(max – min)	Zscore = (x – mean)/standard deviation

---

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)

$$VIF = \frac{1}{1 - R_j^2}$$

Where R_j^2 - the Pearson's correlation for the jth variable.

When the R^2 value is 1 (i.e. when an independent variables are perfectly determine all other independent variables), the VIF for that variable tends to infinite. We can say that the independent variable is redundant & it is explained by all other variables.

Hence, we need to remove this variable else the machine learning model will assign a higher coefficient to jth variable & then model seems to be dependent on jth variable.

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)

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data probably came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.

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

1. When we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.
2. Normal distribution of errors - We can check If residuals follow a normal distribution.
3. Homoscedasticity - Q-Q plot to check if the residuals have a constant variance, which is an assumption for the

---