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)

Categorical variables yr, atemp, season_winter, weathersit, windspeed, month have an impact on the target variable. yr, atemp, season_winter has positive impact while weathersit and windspeed have negative impact.

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)

If there are k categorical variables, we need k-1 dummy variables which can be achieved using drop_first=True

---

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)
atemp, temp has highest correlation value of 0.63 with target variable cnt.

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. Error terms are normally distributed
2. Error terms have constant variance

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. yr
2. atemp
3. season_winter

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>

Modelling uses machine learning algorithms, in which machine learns from data like humans learn from their experience. Objective of regression is to find a linear equation that can best determine the value of dependent variable Y for different values of independent variables X.

The function takes the form –

$$\boldsymbol{y = \beta 0 + \beta 1 X 1 + \beta 2 X 2 + \cdots + \beta n X n + \epsilon}$$

$\beta 0 - Intercept, \epsilon - Error\ Term,\ \ Function\ can\ be\ be\ predicted\ as$ value of Y increases by $\beta 1\ for$ unit increase in X1, given other X values remaining constant.

Assumptions of Simple linear regression –

1. Linear relationship between X and y.
2. Error terms are normally distributed.
3. Errors terms are independent of each other.
4. Error terms have constant variance.

For optimizing the best fitting line following techniques are used –

1. Least squared method
2. Gradient descent

R-squared and Adjusted R-squared are used to evaluate the performance of models.

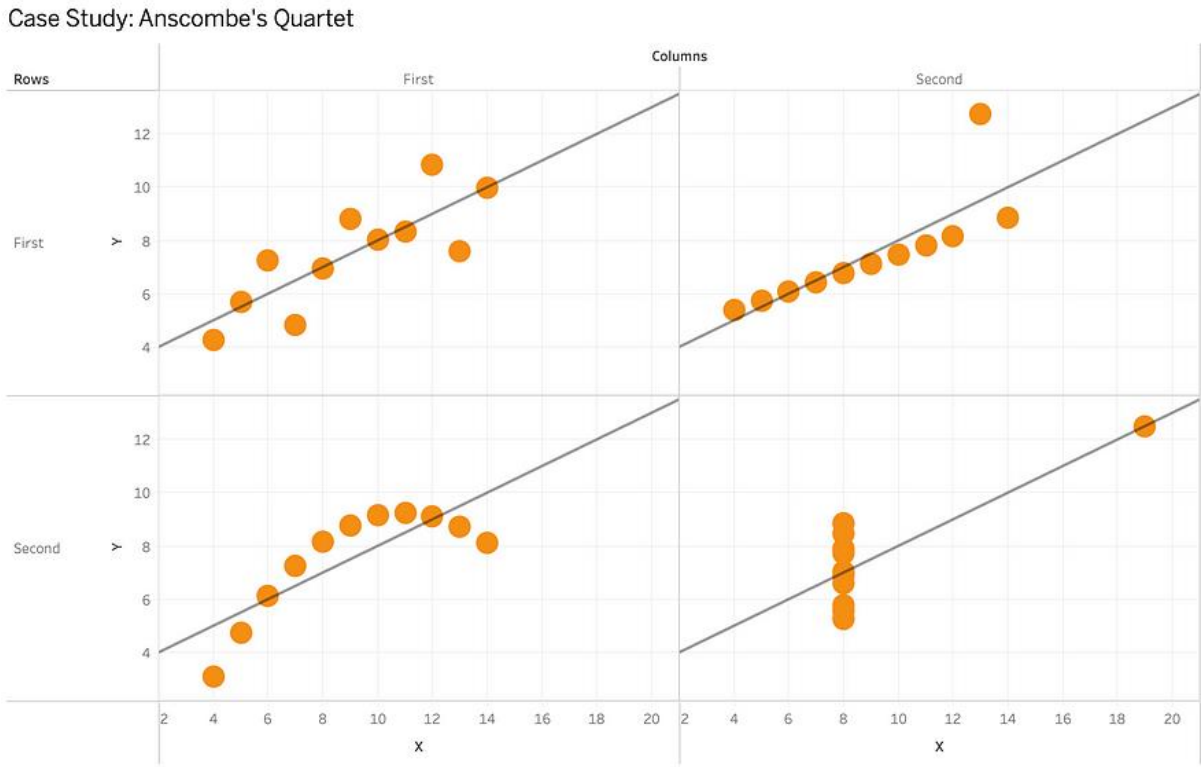
Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>

Anscombe's Quartet comprises four datasets with nearly identical statistical properties but vastly different distributions when visualized. Each dataset contains eleven (x, y) points. Created by Francis Anscombe in 1973, it highlights the importance of data visualization in statistical analysis. The quartet demonstrates how outliers and specific data points can significantly impact results. Anscombe aimed to challenge the belief that numerical calculations are precise while graphs offer only rough approximations.

	x1	y1	x2	y2	x3	y3	x4	y4
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.95	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.74	13	12.74	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.10	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.10	4	5.39	8	5.56
	12	10.84	12	9.13	12	8.15	19	12.50
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
	x1	y1	x2	y2	x3	y3	x4	y4
Mean of x	9		9		9		9	
Variance of x	11		11		11		11	
Mean of y		7.5		7.5		7.5		7.5
Variance of y		4.122		4.122		4.122		4.122
Correlation between x & y	0.816		0.816		0.816		0.816	
linear regression line	y1 = 3 + 0.5x1		y2 = 3 + 0.5x2		y3 = 3 + 0.5x3		y4 = 3 + 0.5x4	

Dataset used by Francis Anscombe, all of the x values are identical and changes happen only in y values. Mean and variance of both x and y, and correlation between x and y and linear regression line are all identical.



Originally, the data sets look identical but obviously they are not as evidenced through visualization and not through summary statistics.

Anscombe's Quartet remains highly relevant today, especially in data science, machine learning, and statistical analysis. In modern applications, datasets are often large and complex, and many analysts rely on automated statistical summaries.

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>

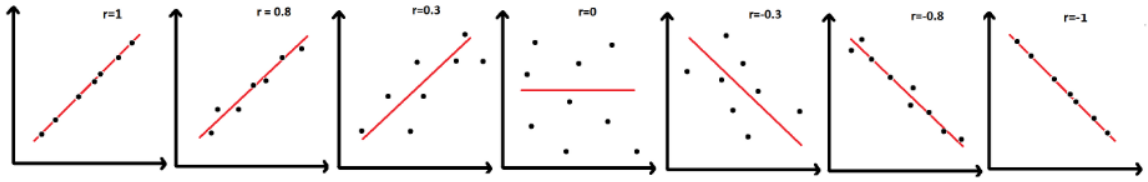
Pearson's R is a measure of linear correlation between two variables. It quantifies how strongly and in what direction two continuous variables are related.

Formula for Pearson's R –

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

We need to check if the relation between variables is significant, to check the lineal correlation between variables we can use the Person's r, or Pearson correlation coefficient.

The range of the possible results of this coefficient denoted by ρ is in the range of (-1,1).



Example of scatter plots with different values of correlation coefficient(ρ).

A value close to 1 indicates a strong positive correlation, meaning both variables increase together, while a value close to -1 indicates a strong negative correlation, meaning one increases as the other decreases. A value near 0 suggests little to no linear relationship between the variables.

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>

When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret.

So we need to scale features because of two reasons:

1. Ease of interpretation
2. Faster convergence for gradient descent methods

Difference between normalized and standardized scaling –

1. Standardize scaling – The variables are scaled in such a way that their mean is zero and standard deviation is one.

$$x = \frac{x - mean(x)}{sd(x)}$$

2. Normalized scaling (Min-Max Scaling) - The variables are scaled in such a way that all the values lie between 0-1 using the maximum and the minimum values in the data.

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

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>

Formula for VIF is given as follows -

$$\mathrm{VIF_i} = \frac{1}{1 - R_i^2}$$

If the value of Ri is 1, VIF becomes infinite. Infinite value of VIF means data has high multicollinearity with other variables.

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)

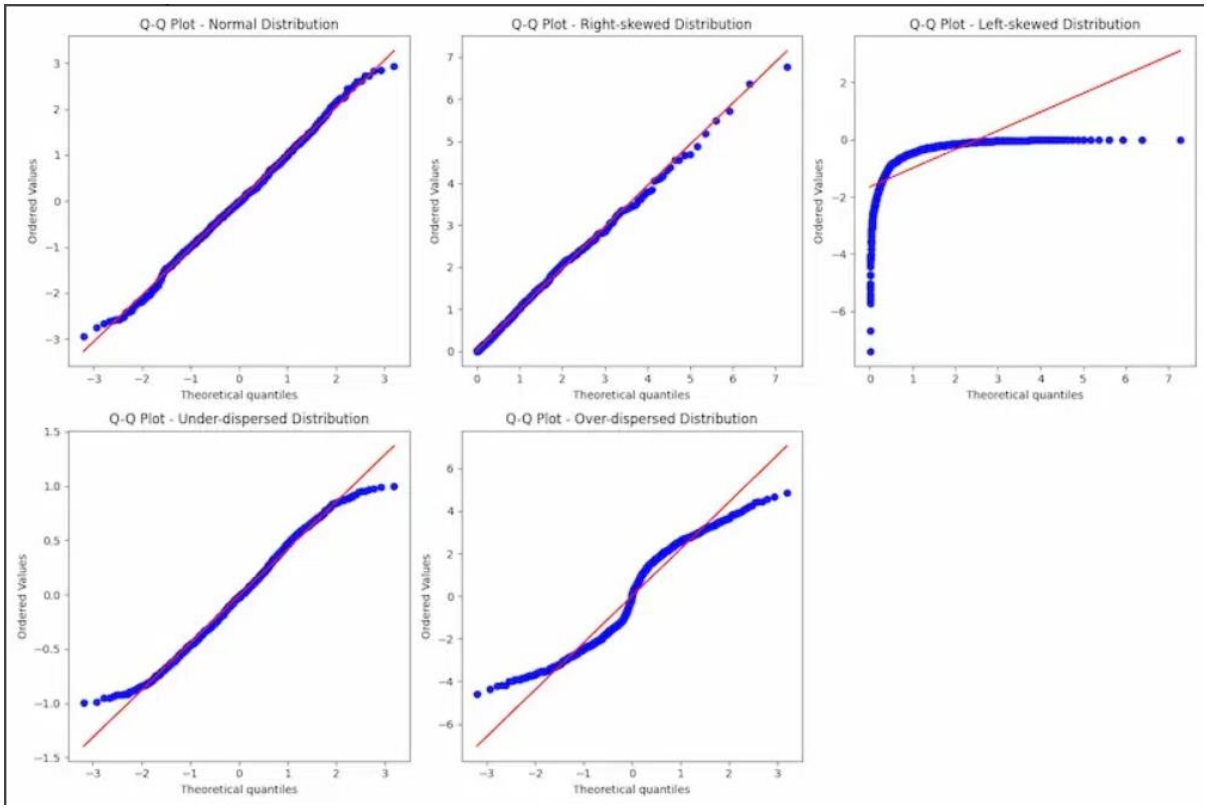
<Your answer for Question 11 goes here>

A Q-Q (Quantile-Quantile) plot is a graphical tool used to compare the distribution of a dataset against a theoretical distribution (usually normal distribution). It helps visualize how well the data follows a specific distribution by plotting quantiles of the actual data against quantiles of the theoretical distribution.

In linear regression, a Q-Q plot is used to check whether the residuals (errors) follow a normal distribution, which is a key assumption for OLS regression. The residuals should be normally distributed for valid statistical inference (like hypothesis testing and confidence intervals)

Importance of Q-Q plot in Linear regression –

1. **Checks Normality of Residuals** – Ensure that residuals follow normal distribution which is key assumption in linear regression.
2. **Identify outliers** – Extreme deviations from straight line indicates outliers.
3. **Detect Skewness** – If points curve upward or downward, the residuals are skewed.
4. **Validated model assumptions** – If residuals deviate significantly from normality, standard regression assumptions might be violated, affecting p-values and confidence intervals.



Q-Q plot for different distributions.