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)

Inference from Bike Rental Distribution by categorical variable Season:

1. Fall season has highest Bike rentals.
2. spring season has lowest Bike rentals.

Inference from Bike Rental Distribution by categorical variable Weather Situation:

1. There are no Bike rentals in 'Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog' weather condition.
2. Highest Bike Rentals in 'Clear, Few clouds, Partly cloudy, Partly cloudy' weather condition.
3. Lowest Bike Rentals in 'Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds' weather condition.

Inference from Bike Rental Distribution by categorical variable Year:

1. Bike rentals have increased significantly in 2019.

Inference from Bike Rental Distribution by categorical variable Month:

1. September has highest Bike Rentals.
2. January has lowest Bike Rentals.

Inference from Bike Rental Distribution by categorical variable Weekday:

1. Friday has highest Bike Rentals.
2. Sunday has lowest Bike Rentals.

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)

drop_first=True is used to prevent multicollinearity. Keeping all dummy variables introduces redundancy, making regression models unstable.

Since one category is predictable from the others, removing it reduces complexity without losing information.

For example: The categorical variable season has 4 values: **spring, summer, fall and winter**. If four dummy variables are used, we have multicollinearity issue because the fourth can be determined if any three is known.

When we use drop_first=True, one dummy variable is dropped (**winter** in this case)

Season	spring	summer	fall	winter
spring	1	0	0	0
summer	0	1	0	0
fall	0	0	1	0
winter	0	0	0	1

After **drop_first=True:**

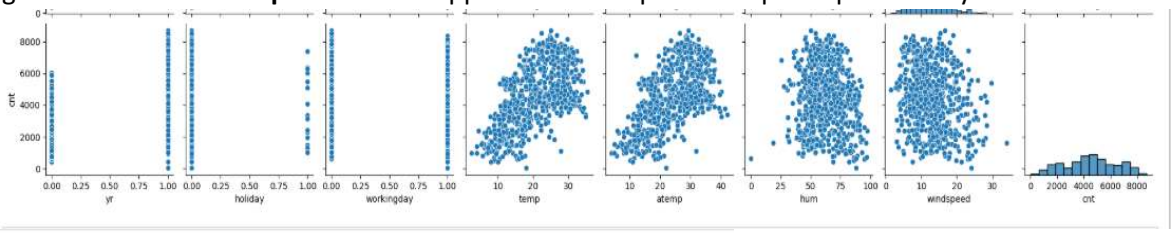
Season	spring	summer	fall
spring	1	0	0
summer	0	1	0
fall	0	0	1
winter	0	0	0

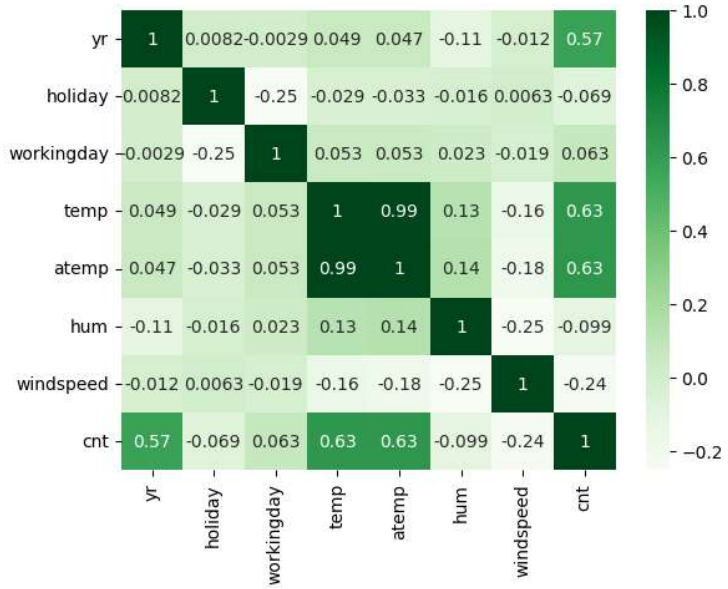
Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: <Your answer for Question 3 goes below this line> (Do not edit)

From the pair-plot **temp** and **atemp** numerical variables have highest correlation (0.63) with the target variable **cnt** . **atemp** has been dropped in next step and temp is kept for analysis.





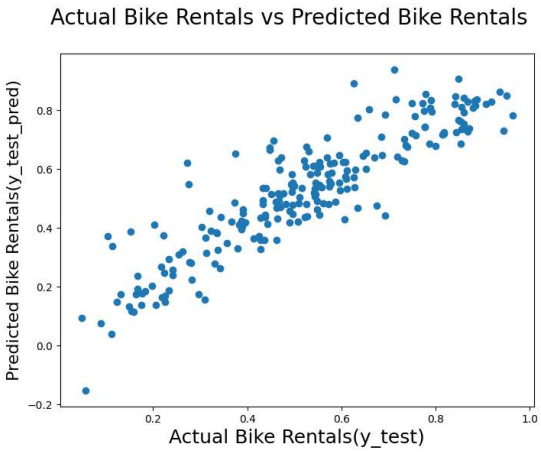
Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 4 goes below this line> (Do not edit)

1. Linearity:

Relationship is linear in the actual vs predicted bike rentals.



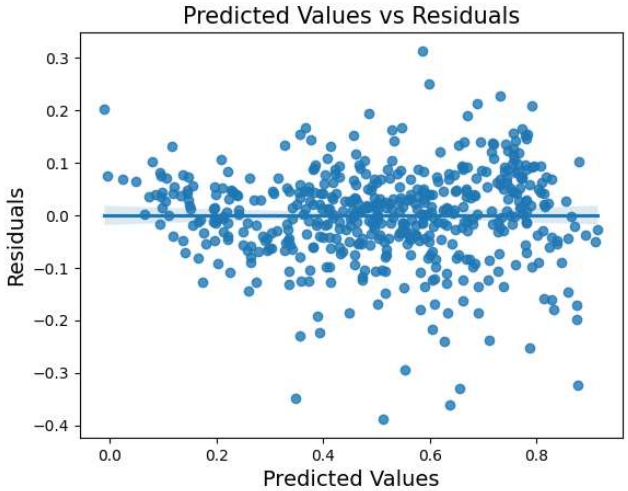
2. No Multicollinearity:

Features are not correlated (VIF < 5 for all variables except constant)

	Features	VIF
0	const	51.11
4	hum	1.88
2	workingday	1.65
10	Saturday	1.64
3	temp	1.60
11	Cloudy	1.56
8	July	1.43
6	summer	1.33
7	winter	1.29
12	Light Snow/Rain	1.24
9	September	1.19
5	windspeed	1.18
1	yr	1.03

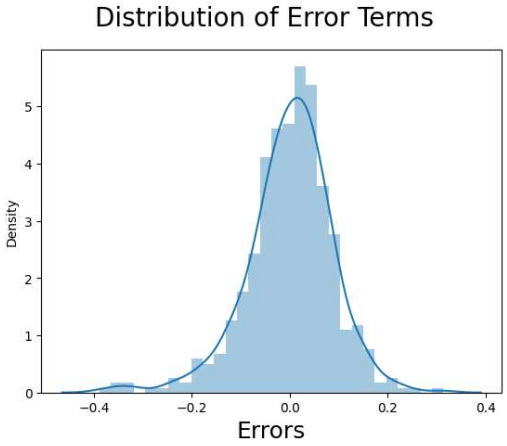
3. Homoscedasticity:

Variance of errors is constant and no pattern observed.



4. Normality of Errors:

Residuals are normally distributed.



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)

Top 3 features a contributing significantly towards explaining the demand of the shared bikes are (which having highest absolute coefficient value in the final model):

1. **temp**
2. **Light Snow/Rain**
3. **yr**

OLS Regression Results

Dep. Variable:	cnt	**R-squared:**	0.843
Model:	OLS	**Adj. R-squared:**	0.839
Method:	Least Squares	**F-statistic:**	222.7
Date:	Tue, 25 Feb 2025	**Prob (F-statistic):**	4.14e-191
Time:	23:28:29	**Log-Likelihood:**	511.32
No. Observations:	510	**AIC:**	-996.6
Df Residuals:	497	**BIC:**	-941.6
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1712	0.028	6.014	0.000	0.115	0.227
yr	0.2286	0.008	28.267	0.000	0.213	0.244
workingday	0.0524	0.011	4.791	0.000	0.031	0.074
temp	0.5960	0.022	26.667	0.000	0.552	0.640
hum	-0.1709	0.037	-4.558	0.000	-0.245	-0.097
windspeed	-0.1888	0.026	-7.393	0.000	-0.239	-0.139
summer	0.0827	0.011	7.770	0.000	0.062	0.104
winter	0.1355	0.010	12.930	0.000	0.115	0.156
July	-0.0439	0.018	-2.450	0.015	-0.079	-0.009
September	0.0928	0.016	5.816	0.000	0.061	0.124
Saturday	0.0625	0.014	4.429	0.000	0.035	0.090
Cloudy	-0.0536	0.010	-5.129	0.000	-0.074	-0.033
Light Snow/Rain	-0.2391	0.026	-9.100	0.000	-0.291	-0.188

Omnibus:	65.304	**Durbin-Watson:**	2.069
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	148.523
Skew:	-0.689	**Prob(JB):**	5.61e-33
Kurtosis:	5.257	**Cond. No.**	20.5

Variable	absolute coef value
temp	0.596
Light Snow/Rain	0.2391
yr	0.2286
windspeed	0.1888
hum	0.1709
winter	0.1355
September	0.0928
summer	0.0827
Saturday	0.0625
Cloudy	0.0536
workingday	0.0524
July	0.0439

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 Algorithm:

1. lntroduction to Linear Regression:

Linear Regression is a supervised learning algorithm used for predicting a continuous dependent variable (Y) based on one or more independent variables (X). It assumes a linear relationship between the variables.

2. Equation of Linear Regression:

The equation of a simple linear regression model is: $Y=\beta_0+ \beta_1X+\epsilon$

Where:

Y = Dependent variable (Target)

X = Independent variable (Feature)

β_0= Intercept (Value of Y when X = 0)

β_1 = Slope (Change in Y for a unit change in X)

ϵ = Error term (Difference between actual and predicted values)

For multiple linear regression, where there are multiple independent variables

$X_1, X_2,...,X_n$, the equation extends to:

$Y= \beta_0+ \beta_1X_1+ \beta_2X_2+...+ \beta_nXn+\epsilon$

3. How Linear Regression Works:

Fit a Line to the Data:

The algorithm finds the best-fitting straight line (in case of simple regression) or hyperplane (for multiple regression).

Minimizing the Error:

The line is chosen by minimizing the Sum of Squared Errors (SSE), which is done using Ordinary Least Squares (OLS) method.

Finding the Best Coefficients:

The coefficients are calculated to minimize the difference between actual and predicted values.

4. Assumptions of Linear Regression:

For a linear regression model to be valid, it must satisfy these assumptions:

Linearity: The relationship between X and Y must be linear.

Independence: Observations should be independent of each other.

Homoscedasticity: Constant variance of residuals (errors).

Normality: Residuals should be normally distributed.

No Multicollinearity: Independent variables should not be highly correlated with each other.

5. **Evaluation Metrics for Linear Regression**

To measure how well the model fits the data, we use:

Mean Squared Error (MSE): Measures average squared differences between actual and predicted values.

Root Mean Squared Error (RMSE): Square root of MSE, making errors interpretable in original units.

R^2 Score (Coefficient of Determination): Explains how much variance in Y is explained by X.

6. **Applications of Linear Regression**

Predicting house prices
Sales forecasting
Stock price prediction
Analyzing trends in economics and finance

Conclusion:

Linear Regression is a powerful yet simple model for predicting continuous variables. However, it works best when its assumptions are met and is sensitive to outliers and multicollinearity.

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 descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.
The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.
The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

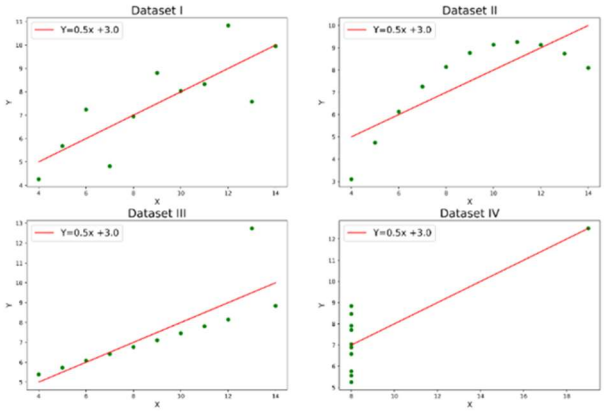
Purpose of Anscombe's Quartet

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

Anscombe's Quartet Dataset

The four datasets of Anscombe's quartet.

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



Anscombe's quartet Plot

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 (Correlation Coefficient):

Pearson's R is commonly used in statistics and machine learning to assess relationships between variables. Pearson's correlation coefficient (r) measures the linear relationship between two continuous variables.

It ranges from -1 to +1:

+1 indicates Perfect positive correlation (as one increases, the other increases).

-1 indicates Perfect negative correlation (as one increases, the other decreases).

0 indicates No correlation (no linear relationship).

It is calculated as:

$r= (\Sigma (X_i - \bar{X})(Y_i - \bar{Y}))/\sqrt{(\Sigma (X_i - \bar{X})^2 \Sigma (Y_i - \bar{Y})^2)}$

Where:

X and Y are the two variables

$\bar{X}$ and $\bar{Y}$ are their means

Numerator measures how X and Y vary together.

Denominator normalizes by the individual variations of X and Y.

This formula helps determine how strongly X and Y are linearly related.

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 technique used in data preprocessing to adjust the range of numerical features so they have a consistent scale. It is crucial in machine learning algorithms that rely on distances (example: regression, clustering, SVM, and neural networks).

Scaling is performed

1. To Improve Model Performance

Many ML models (example: gradient descent-based algorithms) converge faster when features are scaled properly.

2. To Handle Features with Different Ranges

Some features may have a large scale (example: income in thousands), while others may have a small scale (example: age in years). Scaling ensures fair treatment of all features.

3. To Prevent Dominance of Larger Magnitude Features

Without scaling, variables with larger numerical ranges dominate the learning process, making models biased.

Difference Between Normalized Scaling and Standardized Scaling:

Feature	Normalization (Min-Max Scaling)	Standardization (Z-Score Scaling)
Formula	$X'=(X-X_{min})/(X_{max}-X_{min})$	$X'=(X-\mu)/\sigma$
Range	Scales data between [0,1] (or [-1,1])	Mean = 0, Std Dev = 1
Sensitive to Outliers	Yes (Outliers affect min/max values)	No (Less sensitive due to standard deviation)
Use Case	When features have bounded values (example: pixel values, percentages)	When features follow a normal distribution (example: Gaussian data)
Effect on Shape	Preserves the original distribution shape	Changes the distribution to have 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 **VIF** value becomes **infinite** when there is perfect multicollinearity between variables. This happens in two main cases:

1. Exact Linear Dependence Between Two or More Predictors:

If one independent variable is a perfect linear combination of others, the denominator in the VIF formula becomes zero, causing VIF to be infinite.

2. Presence of Duplicate or Redundant Features:

If a feature is exactly duplicated in the dataset, VIF will be infinite.

This often happens when:

A categorical variable is not properly one-hot encoded.

A feature is accidentally duplicated in the dataset.

Mathematical way:

The VIF formula for a variable X_k is:

$VIF=1/(1-R_k^2)$

Where:

R_k^2 is the R-squared value from regressing X_k on all other independent variables.

If $R_k^2=1$ (i.e., perfect correlation), then:

$1- R_k^2=0$

$VIF=1/0=\infty$

This means X_k is completely explained by other predictors, making VIF infinite.

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 used to compare the distribution of a dataset to a theoretical distribution, typically a normal distribution. It helps in checking whether the data follows a specific distribution.

Use of a Q-Q Plot in Linear Regression:

1**. Checking Normality of Residuals**

One key assumption of linear regression is that the residuals (errors) are normally distributed. A Q-Q plot helps visualize whether residuals follow a normal distribution.

If residuals form a straight line, then Normality assumption holds.

If residuals deviate (for example: S-shape, tails away) then Residuals are not normally distributed.

2**. Detecting Outliers and Skewness:**

Outliers: appear as points deviating far from the straight line.

Skewness: If points curve upward (right-skewed) or downward (left-skewed).

3. **Identifying Heavy or Light Tails:**

Heavy tails (fat tails, high kurtosis): Points at the ends fall above the line.

Light tails (low kurtosis): Points at the ends fall below the line.

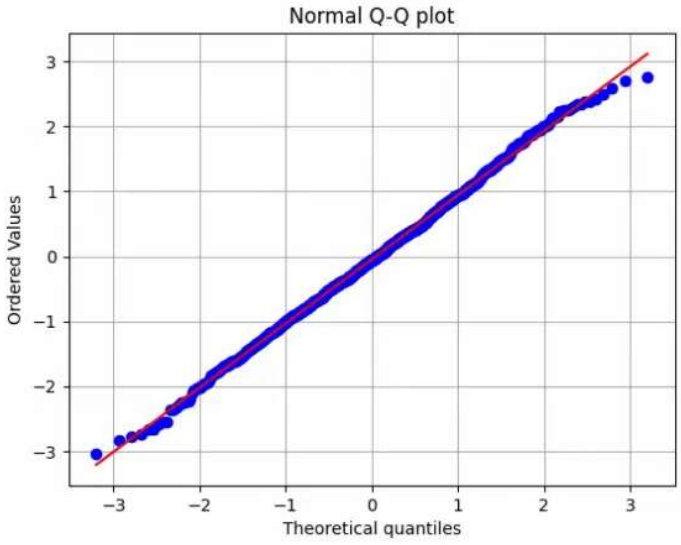
Importance of Q-Q Plot in Linear Regression:

Validates the normality assumption of residuals.

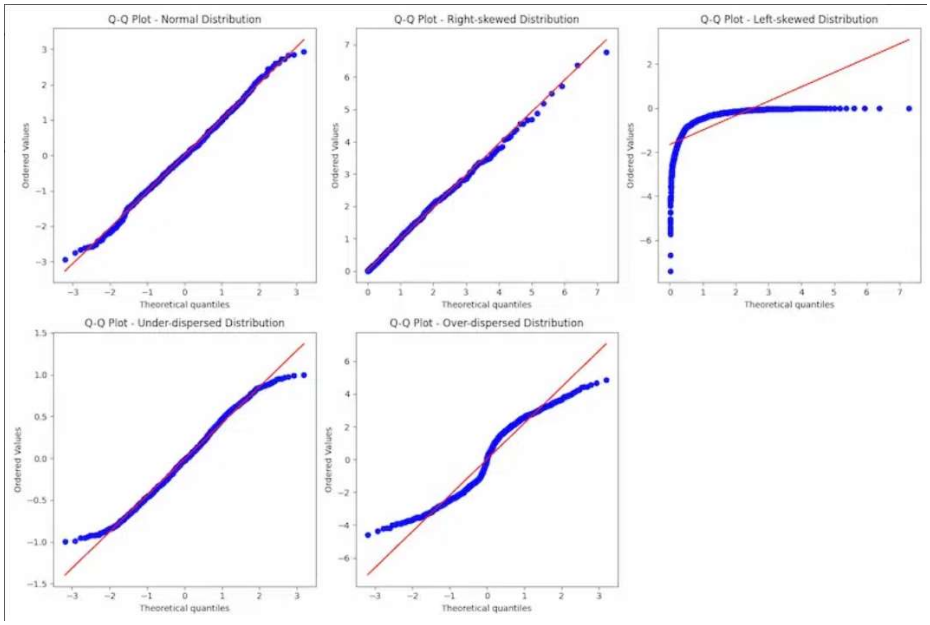
Helps detect skewness, kurtosis, and outliers.

Ensures the accuracy of confidence intervals and p-values.

Prevents misleading regression results due to non-normal errors



Q-Q plot


Q-Q Plot - Normal Distribution
Q-Q Plot - Right-skewed Distribution
Q-Q Plot - Left-skewed Distribution
Q-Q Plot - Under-dispersed Distribution
Q-Q Plot - Over-dispersed Distribution
Ordered Values
Theoretical quantiles