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)

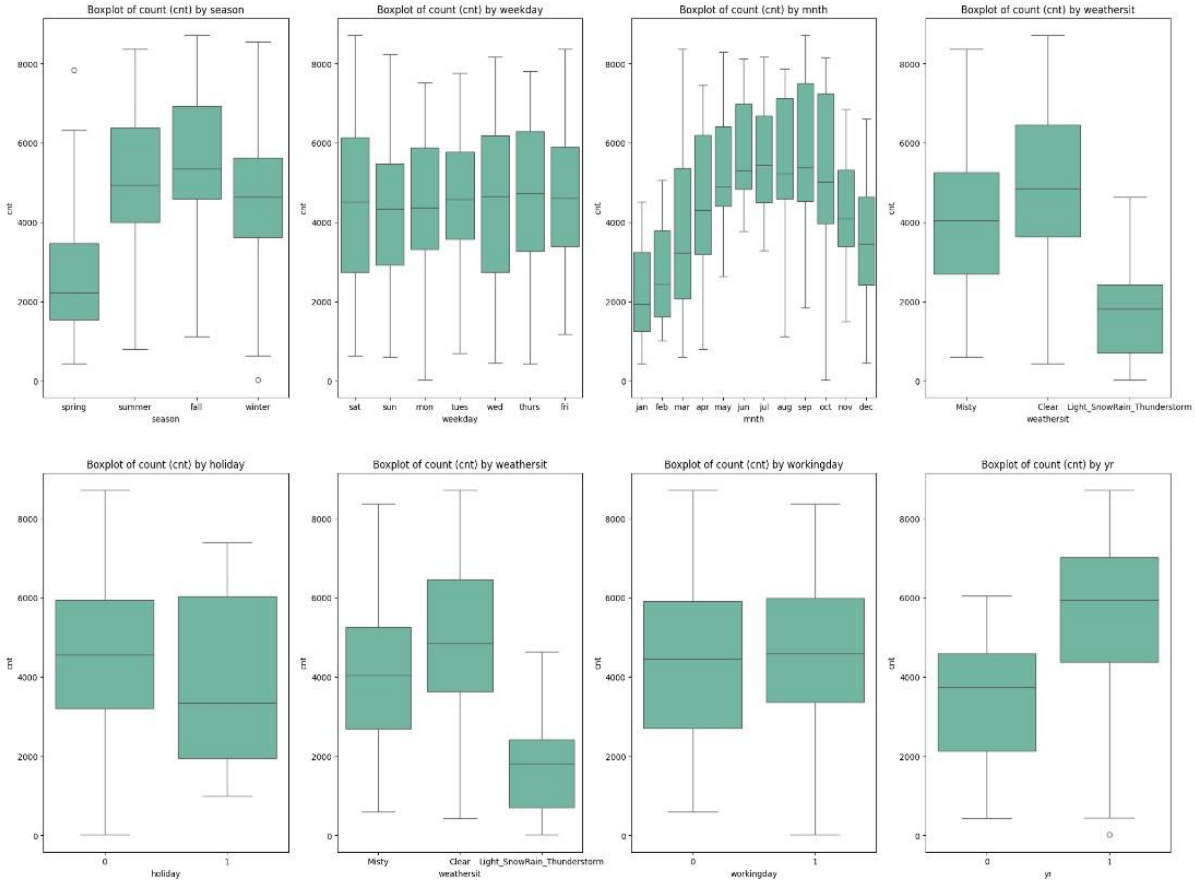
Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

Based on the analysis of categorical variables such as season, weathersit, yr, mnth, holiday and weekday the following points summarize their impact on the dependent variables:

- **Season: Bike demand is highest during the fall season, while the spring season sees the lowest demand.**
- **Weather Condition:**
- **"Clear, Few Clouds" weather is associated with the highest bike demand.**
- **"Light Snow, Light Rain" weather results in the lowest demand for bikes.**
- **"Mist-Cloudy" and "Mist-Broken" conditions lead to medium levels of demand.**
- **Year: In 2018, bike demand was relatively low, but demand saw a significant increase in 2019.**
- **Month: September records the highest demand, followed by October, August, and June. January, however, experiences the lowest demand.**
- **Holiday: Bike demand is notably higher on holidays compared to regular days.**
- **Weekday: There is little to no variation in bike demand across different weekdays.**

Below the boxplot to visually represent the data.



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 drops the first column during dummy variable creation.

This is important because it helps minimize the creation of redundant columns when generating dummy variables, thereby reducing the correlation that can arise between them

Suppose we have a categorical column with three distinct categories and we want to create dummy variables for them. If one category is labeled as "unfurnished" and another as "semi-furnished," it is clear that the third category must be "furnished" by elimination, so we don't need to include a dummy variable for the third category.

Therefore, when dealing with categorical variables with 'n' distinct categories, we only need to create 'n - 1' dummy variables to fully represent the data, as the last category can be inferred.

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)

Both 'temp' and 'atemp' show a strong correlation with the 'cnt' variable, as demonstrated in the pair plot.

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)

- **Below are the assumptions for simple linear regression:**
- **Linear Relationship: There should be a linear relationship between the independent (X) and dependent (Y) variables.**
- **Normality of Error Terms: The error terms should follow a normal distribution (note: this assumption applies to the errors, not X or Y).**
- **Independence of Errors: The error terms must be independent of one another.**
- **Homoscedasticity: The variance of error terms should remain constant (no heteroscedasticity).**

With these assumptions, we can confidently make inferences about the model. Without them, we would not be able to draw valid conclusions. Additionally, there is no assumption regarding the distribution of X and Y

themselves; the focus is solely on the normality of the error terms.

To ensure these assumptions were met, the model was built using the training set with the following checks:

- **Coefficient Values: Non-zero coefficients suggest that there is a meaningful relationship between the independent and dependent variables.**
- **p-Values: p-values less than 0.05 indicate that the variables are statistically significant in the model.**
- **(VIF): VIF values below 5 suggest that there is no multicollinearity between the predictor variables.**
- **F-Statistic and Prob(F-Statistic): A high F-statistic and a low p-value for the F-statistic show that the model's overall fit is statistically significant and not due to random chance or solely due to individual predictors.**

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 top 3 features contributing significantly towards the demand of the shared bikes:

1. **atemp**
2. **yr**
3. **weathersit**

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 is a statistical method used to model the relationship between a dependent variable and one or more independent variables. A linear relationship implies that any change in the value of the independent variable(s) (either an increase or decrease) will result in a corresponding change in the dependent variable. This relationship can be mathematically expressed using the equation:

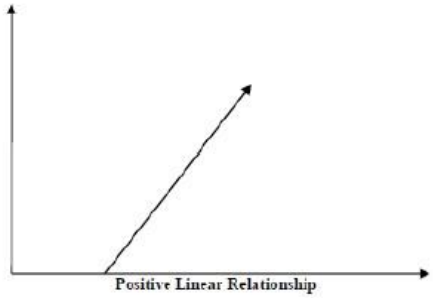
Y = mX + c

Where:

- **Y is the dependent variable being predicted,**
- **X is the independent variable used for prediction,**
- **m represents the slope of the regression line, indicating how changes in X affect Y,**
- **c is the constant or the Y-intercept, which defines the value of Y when X equals zero.**
- **Additionally, the linear relationship between variables can be either positive or negative, as outlined below:**

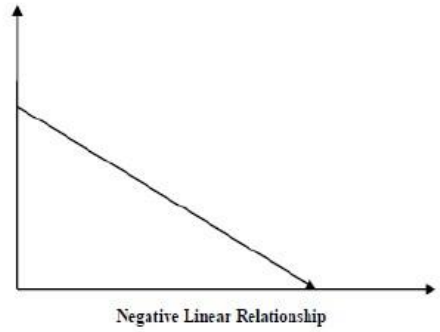
Positive Linear Relationship:

A positive linear relationship occurs when both the independent and dependent variables increase together. This means that as the independent variable rises, the dependent variable also rises. This relationship can be visualized with an upward-sloping line on a graph.



Negative Linear Relationship:

A negative linear relationship refers to a scenario where an increase in the independent variable leads to a decrease in the dependent variable. In other words, as the independent variable rises, the dependent variable falls. This type of relationship is depicted by a downward-sloping line on a graph.



Linear regression can be classified into two main types:

Simple Linear Regression
Multiple Linear Regression
Assumptions of Linear Regression:
The linear regression model is based on the following key assumptions about the dataset:

a. Multicollinearity
Linear regression assumes that there is little to no multicollinearity among the independent

variables. Multicollinearity occurs when two or more predictors are highly correlated with each other, which can affect the reliability of the model's estimates.

b. Autocorrelation

The model assumes that there is minimal or no autocorrelation in the residuals. Autocorrelation arises when the residuals (errors) are correlated with each other, violating the assumption of independent errors.

c. Linearity

Linear regression assumes that the relationship between the independent (predictor) variables and the dependent (response) variable is linear. This means that changes in the predictors are expected to result in proportional changes in the dependent variable.

d. Normality of Error Terms

The error terms (residuals) should be normally distributed, which allows for valid statistical inferences about the model parameters.

e. Homoscedasticity

There should be constant variance of the residuals across all levels of the independent variables. This means the spread of residuals should remain consistent and not show any patterns when plotted.

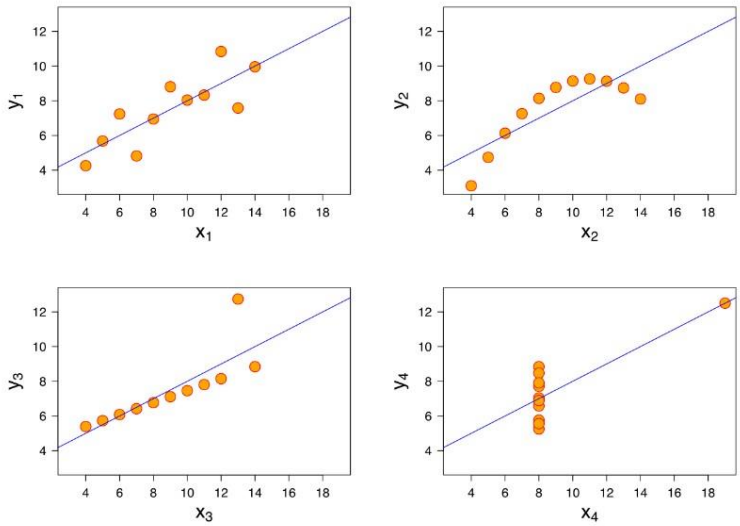
Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Anscombe's Quartet was created by statistician Francis Anscombe, consists of four distinct datasets, each containing eleven (x, y) data points. Despite having identical summary statistics, the behavior of these datasets is drastically different when plotted. The key takeaway is that the graphical representation of the data reveals vastly different patterns, highlighting how summary statistics alone can be misleading in understanding the underlying data.

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	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	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

- **Mean of x is 9 and mean of y is 7.50 for each dataset.**
- **Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset**
- **The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset**

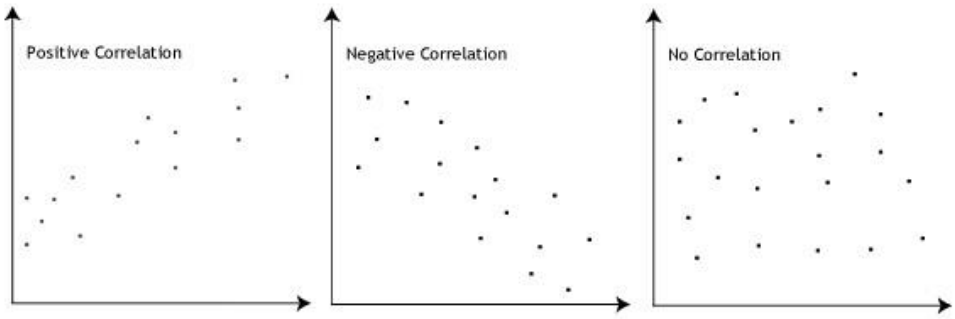
When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:



Dataset I seems to have a well-defined and accurate linear relationship.
Dataset II does not follow a normal distribution.
Dataset III demonstrates a linear trend, but the presence of an outlier distorts the regression results.
Dataset IV illustrates how a single outlier can lead to a misleadingly high correlation coefficient.
This set of datasets highlights the critical role that data visualization plays in analysis. By visualizing the data, we can gain valuable insights into its structure and get a clearer understanding of the underlying patterns.

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, denoted as r, quantifies the strength and direction of the linear relationship between two variables. If the variables tend to increase or decrease together, the correlation coefficient will be positive. Conversely, if one variable increases while the other decreases, the correlation coefficient will be negative. The value of r ranges from +1 to -1, where a value of 0 indicates no linear relationship between the variables. A positive value of r signifies a positive correlation, meaning that as one variable increases, the other tends to increase as well. A negative value of r indicates a negative correlation, where an increase in one variable is associated with a decrease in the other. This relationship is visually represented in the diagram below:



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)

Feature scaling is a technique used to normalize the independent variables in a dataset to a consistent range. This step is typically performed during data preprocessing to address issues related to varying magnitudes, values, or units across features. Without feature scaling, machine learning algorithms may place more importance on features with larger values or scales, leading to biased results. For example, without scaling, an algorithm might incorrectly interpret a value of 3000 meters as being greater than 5 kilometers, even though 5 km is actually larger. Feature scaling resolves this by standardizing the values, ensuring that all features contribute equally to the model and preventing the algorithm from being misled by differences in magnitude or units.

S.NO.	Normalized scaling	Standardized scaling
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

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)

IF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables. For example, we would fit the following models to estimate the coefficient of determination R1 and use this value to estimate the VIF:

$$X_1=C+ \alpha_2 X_2+\alpha_3 X_3+$$

$$〖VIF〗_1=1/(1-R_1^2)$$

Next, we fit the model between X2 and the other independent variables to estimate the

coefficient of determination R2:

$X_2=C+ \alpha_1 X_1+\alpha_3 X_3+$

〖If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. orthogonality can refer to the independence of predictors: Independent Variables: If two independent variables are orthogonal, it means they do not share any variance. This can reduce multicollinearity, leading to more stable and interpretable models. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. This would mean that that standard error of this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation). The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity. A general rule of thumb is that if VIF > 10 then there is multicollinearity. Note that this is a rough rule of thumb, in some cases we might choose to live with high VIF values if it does not affect our model results such as when we are fitting a quadratic or cubic model or depending on the sample size a large value of VIF may not necessarily indicate a poor model.

VIF	Conclusion
1	No Multicollinearity
4-5	Moderate
10 or greater	Severe

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 plausibly 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. This helps in a scenario of linear regression 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.

Few advantages:

- **It can be used with sample sizes also**
- **Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.**

It is used to check following scenarios:

If two data sets —

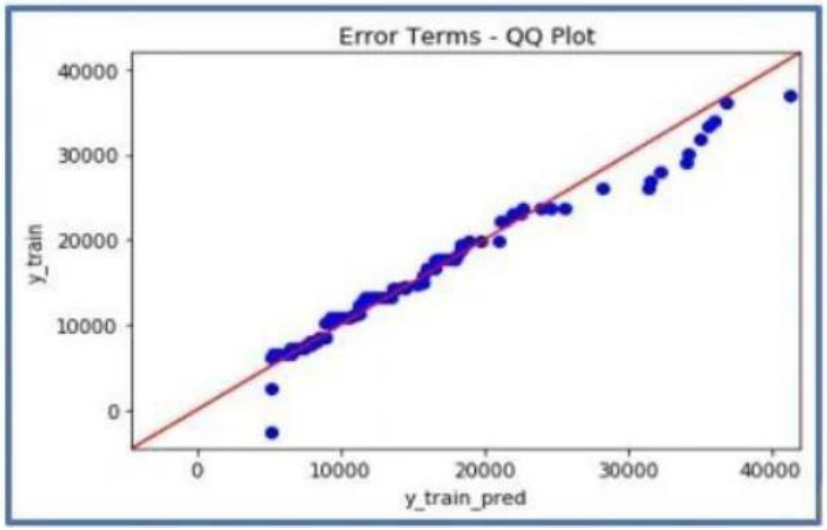
i. **come from populations with a common distribution**
ii. **have common location and scale**
iii. **have similar distributional shapes**

iv. have similar tail behavior Interpretation:

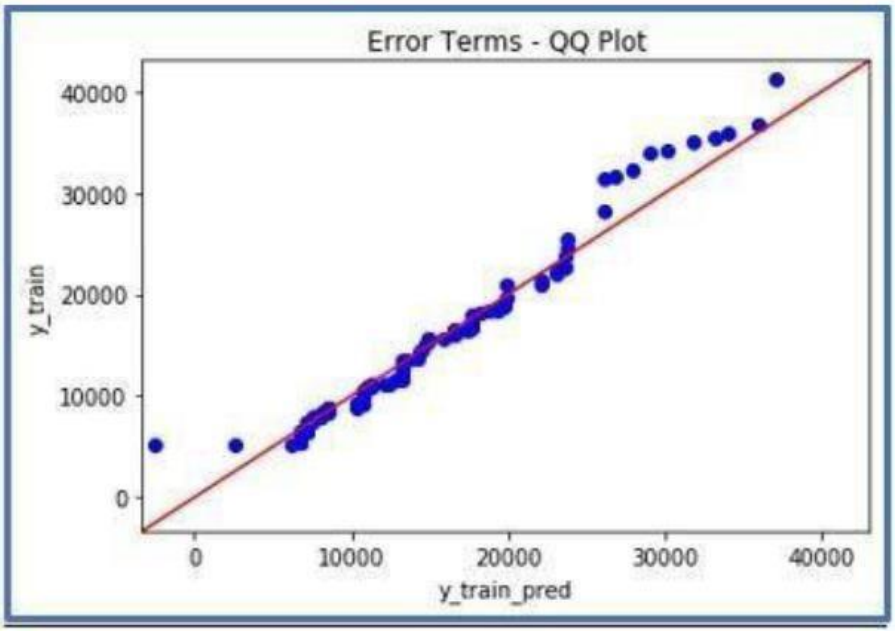
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

Below are the possible interpretations for two data sets.

- **Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis**
- **Y-values < X-values: If y-quantiles are lower than the x-quantiles.**



- **X-values < Y-values: If x-quantiles are lower than the y-quantiles.**



- **Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis**

Python: statsmodels.api provide qqplot and qqplot_2samples to plot Q-Q graph for single and two different data sets respectively.