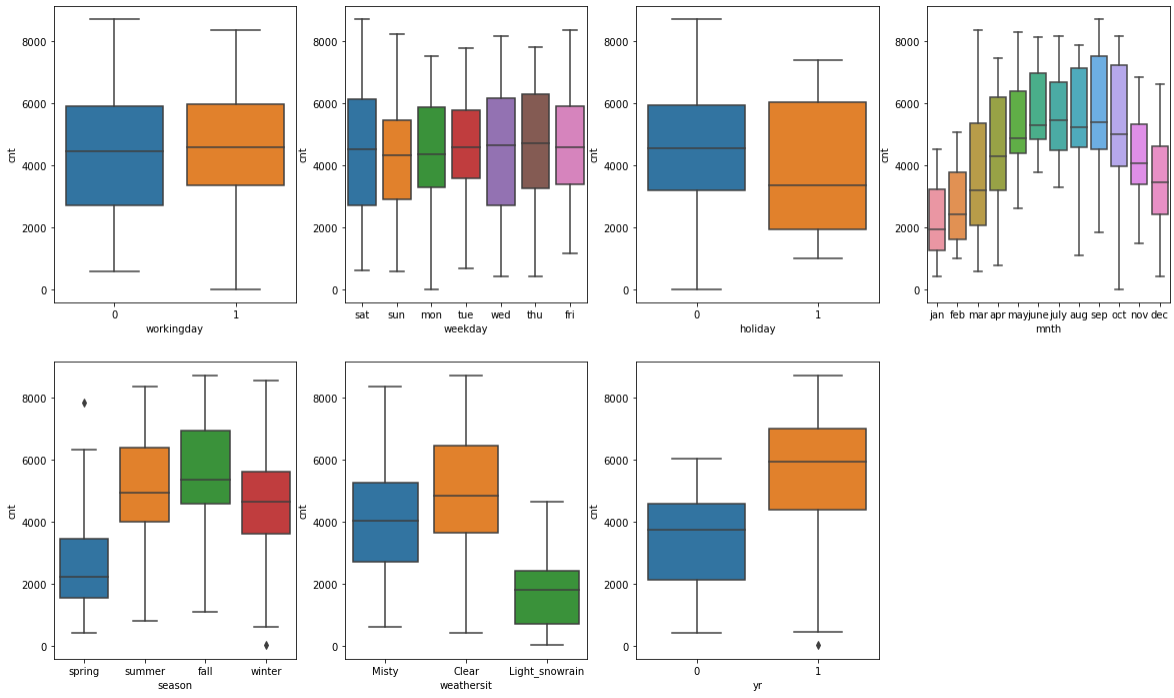
Assignment-based Subjective Questions

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)



As we can see above totally there are seven categorical variables in Bike Sharing Dataset,
workingday: Booking seemed to be almost equal either on working day or non-working day
weekday: Thu, Fir, Sat have more number of bookings as compared to the start of the week. But this is not very significant.
holiday: Bookings are more in when there is no holiday, the distribution is as biased.
mnth: Most of the bookings occurred between May and October, with a noticeable increase in bookings from the start of the year until mid-year. After that, the number of bookings began to decline as the year progressed towards its end.
Season: Fall season have more booking followed by summer, winter. Spring bookings are low
weathersit: When the weather is clear there are more booking which seems obvious
yr: Bookings have been increased drastically in 2019 compare with 2018

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)

Using drop_first=True is important as it helps eliminate the extra column created when generating dummy variables, thereby reducing the correlation among the dummy variables.

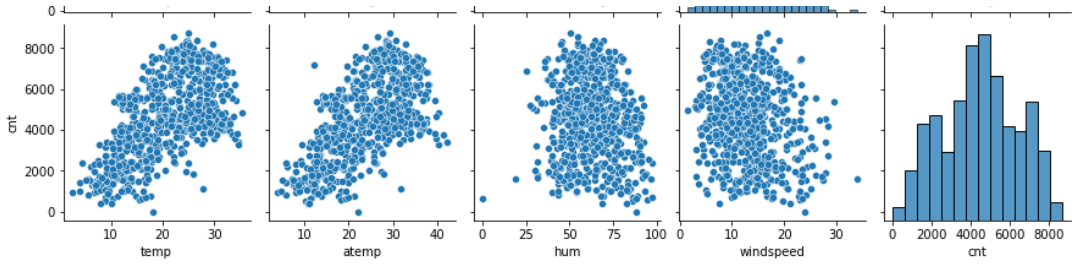
Syntax:

```
df = pd.get_dummies(df, drop_first=True)
```

- **drop_first**: bool, default is False. It indicates whether to create k-1 dummies from k categorical levels by removing the first level.

For example, if a categorical column has 3 values (A, B, and C) and we create dummy variables, if the value is not A or B, it must be C. Therefore, we don't need the third variable (C) to identify it, as the other two can implicitly indicate C.

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)



'temp', 'atemp' variable has the highest correlation with the target variable. Only 'temp' is used for model due to it is linear with 'atemp'

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)

I have validated the assumptions of the Linear Regression Model based on the following five criteria:

- **Normality of error terms**
  The error terms should follow a normal distribution.
- **Multicollinearity check**
  There should be minimal multicollinearity among the predictor variables.
- **Linearity validation**
  A linear relationship should be evident between the variables.
- **Homoscedasticity**
  Residual values should not show any discernible pattern.
- **Independence of residuals**
  The residuals should be free from autocorrelation.

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 explaining the demand of the shared bikes –

- **Temp**
- **Year**
- **Winter**

General Subjective Questions

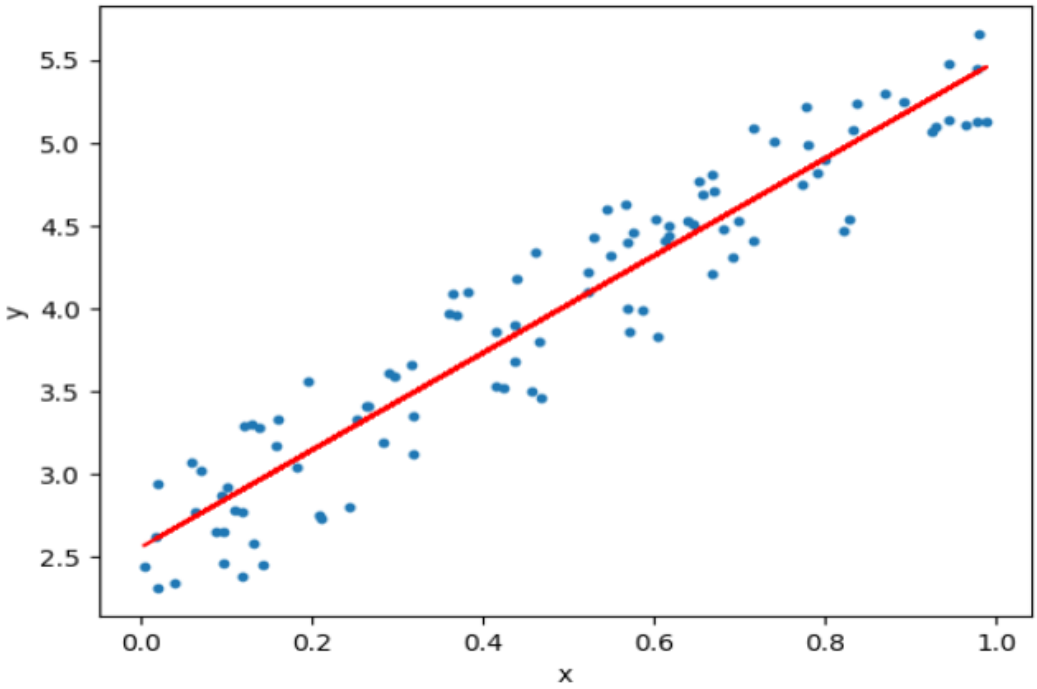
Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Linear Regression is a supervised machine learning algorithm used for regression tasks. It predicts the dependent variable based on the independent variables in the dataset.
Mathematically, Simple Linear Regression can be represented as shown below.
$y = C + \beta * X$
Multiple Linear Regression can be represented as shown below.
$y = C + \beta 0 * X0 + \beta 1 * X1 + \ldots\ldots + \beta n * Xn$

In linear regression, X and y represent two variables:

- X is the independent variable from the dataset.
- y is the dependent variable from the dataset.
- C is the intercept of the regression line.
- β is the coefficient of the independent variable.

The best-fit line helps us understand how changes in the independent variable(s) affect the dependent variable. The red line below is the best fit line for the dataset.



Hypothesis Testing in Linear Regression:

● Null Hypothesis: (H**0**): Coefficients of Linear Equation are equal to zero, i.e. coefficients are not significant

● Alternate Hypothesis: (H**1**): At Least one coefficient of Linear Equation should not be equal to zero

Assumptions associated with Linear Regression model:

1. **Linearity**: A straight-line relationship between independent (X) and dependent (Y) variables.
2. **Homoscedasticity**: Constant variance of residuals across all levels of X.
3. **Independence**: Observations are independent of each other.
4. **Normality**: Residuals are normally distributed for each value of X.

When transitioning from Simple to Multiple Linear Regression, additional considerations are necessary:

1. **Overfitting**: Adding more variables can make the model too complex, leading to overfitting. This results in high accuracy on the training set but poor generalization to the test set.
2. **Multicollinearity**: High correlation between predictor variables, measured using the Variance Inflation Factor (VIF), which can distort the model's estimates.
3. **Feature Selection**: Choosing the most relevant features from the dataset that have a significant impact on the model's performance to improve accuracy.

Model evaluation for a linear equation can be done using the following methods:

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)

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 quartet was created by the statistician **Francis Anscombe** in 1973 to demonstrate the importance of graphical analysis in statistics and how summary statistics alone can be misleading.

Dataset

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

Statistical summary of the data

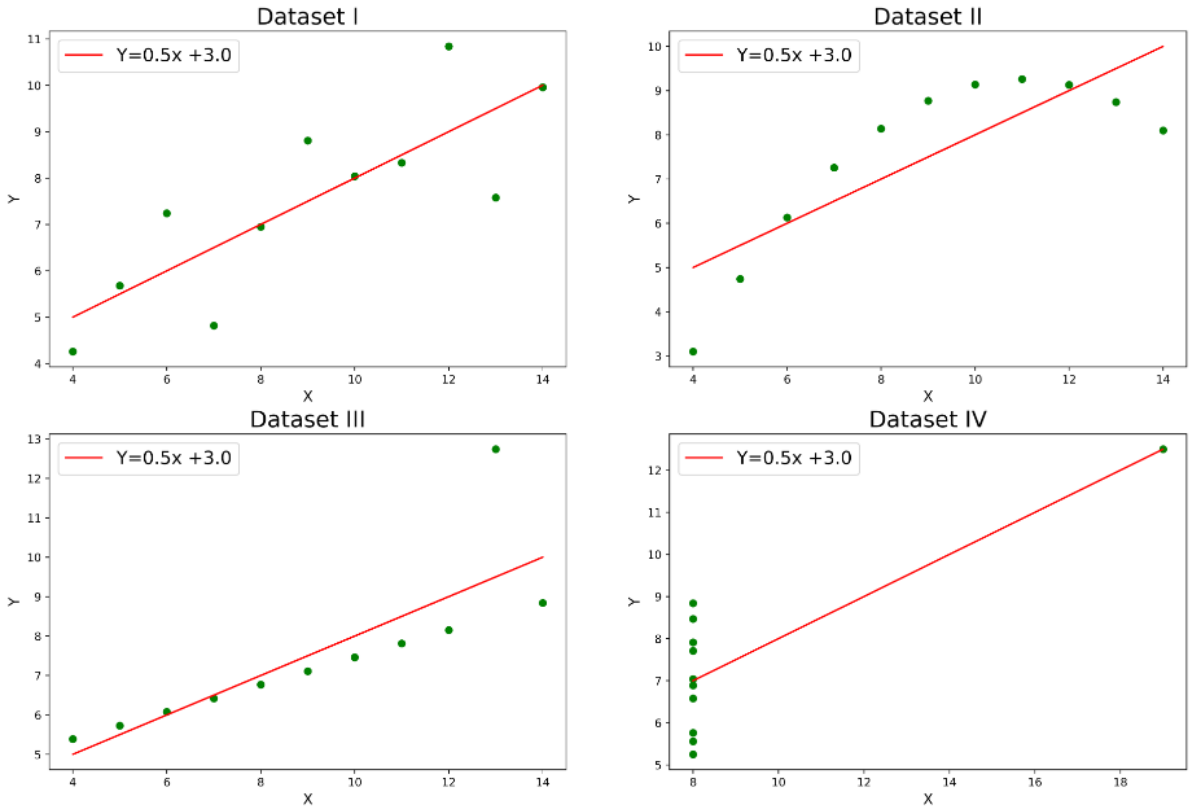
	I	II	III	IV
Mean_x	9.000000	9.000000	9.000000	9.000000
Variance_x	11.000000	11.000000	11.000000	11.000000
Mean_y	7.500909	7.500909	7.500000	7.500909
Variance_y	4.127269	4.127629	4.122620	4.123249
Correlation	0.816421	0.816237	0.816287	0.816521
Linear Regression slope	0.500091	0.500000	0.499727	0.499909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

The summary statistics show that the means and the variances were identical for x and y across the groups:
- 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 are 0.816 for each dataset

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

Below is one example.



Explanation of this output:

- In the first one (top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one (top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one (bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

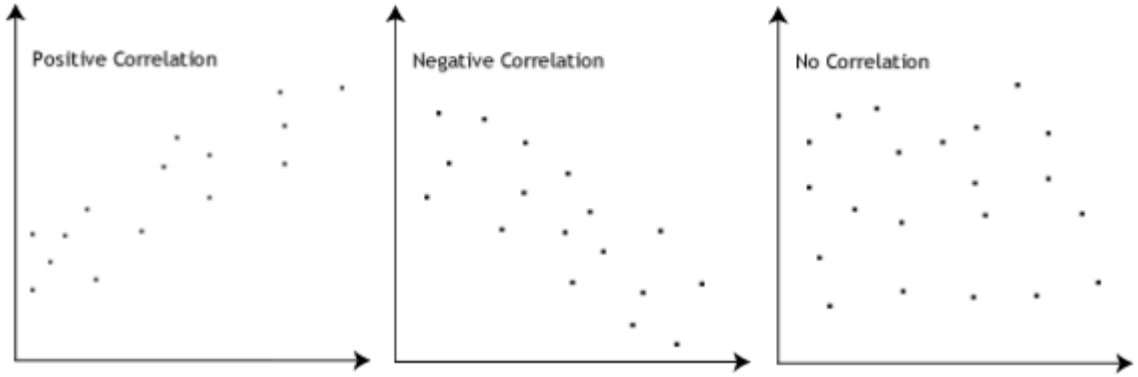
Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Pearson's r is a numerical measure that describes 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 increase while the other decreases, the correlation coefficient will be negative.

The Pearson correlation coefficient (r) ranges from +1 to -1:

- A value of 0 means no linear relationship between the variables.
- A positive value (greater than 0) indicates a positive relationship, meaning both variables increase or decrease together.
- A negative value (less than 0) indicates a negative relationship, meaning as one variable increases, the other decreases.

This is visually represented in a diagram, which illustrates the correlation.



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 standardize the independent features in a dataset to a fixed range. It is performed during data preprocessing to address the issue of features with varying magnitudes, values, or units. Without feature scaling, machine learning algorithms may give more weight to features with larger values and underestimate those with smaller values, regardless of their units.

For example, without feature scaling, an algorithm might incorrectly interpret 3000 meters as larger than 5 kilometers, even though this is not true. This can lead to incorrect predictions. Therefore, feature scaling is applied to bring all values to the same magnitude, ensuring more accurate results.

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)

The Variance Inflation Factor (VIF) can become infinite when there is perfect multicollinearity between two or more predictor variables in the dataset. This happens when one predictor variable is an exact linear function of one or more other predictors.

Here's why this happens:

1. Multicollinearity means that the independent variables in the model are highly correlated with each other. This makes it difficult to determine the individual effect of each predictor on the dependent variable.
2. When there is perfect multicollinearity, one of the predictors can be perfectly predicted from the others. In this case, the regression model cannot distinguish the unique contribution of that predictor, and the VIF for that predictor becomes infinitely large.

Example:

If you have two variables X1 and X2, and X2 = a * X1 (where "a" is a constant), then X2 is perfectly predictable from X1, leading to infinite VIF for X2.

How to Handle Infinite VIF:

- Remove the redundant variable(s) that are perfectly correlated.
- Combine highly correlated variables into a single variable (for example, through principal component analysis or other dimensionality reduction techniques).

Infinite VIF occurs due to perfect correlation or exact linear dependency among the predictors, which makes the regression model unable to estimate their individual effects accurately.

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 method used to assess whether two datasets originate from populations with the same distribution.

Use of Q-Q Plot:

A Q-Q plot compares the quantiles of two datasets by plotting the quantiles of the first dataset against the quantiles of the second dataset. A quantile represents the fraction (or percentage) of data points below a given value. For example, the 0.3 (or 30%) quantile is the value below which 30% of the data points fall, with 70% above it. A 45-degree reference line is also drawn on the plot. If both datasets come from populations with the same distribution, the points should align closely along this reference line. The more the points deviate from this line, the stronger the evidence that the datasets are from different distributions.

Importance of Q-Q Plot:

When comparing two data samples, it's important to determine if they come from populations with the same distribution. If the datasets share the same distribution, location and scale estimators can be combined to estimate the common location and scale. If the datasets differ, the Q-Q plot helps in understanding the differences more effectively than analytical methods like the chi-square or Kolmogorov-Smirnov tests. It provides deeper insights into the nature of these differences.