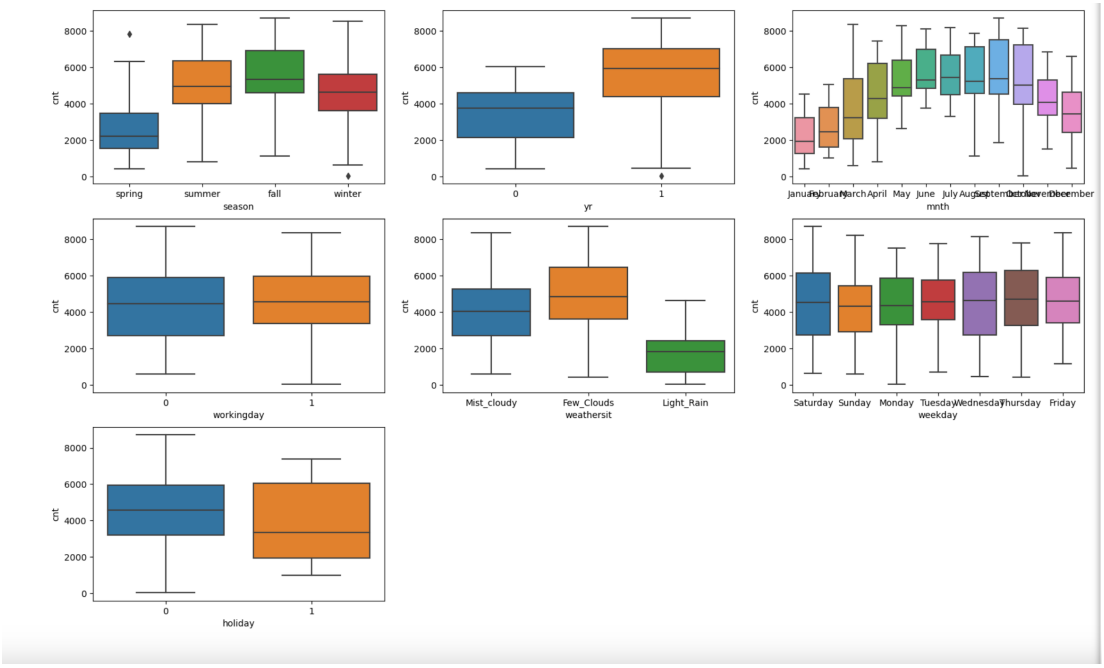
Assignment-based Subjective Questions

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)



- Most of bike bookings happening in fall sesion, and spring has less bookings

- Bookings increased from 2018 to 2019

- August, September and October had good amount of bookings

- Bookings are more on Few_clouns and mist_cloudy wether

- weekday looks similar on all days

- seems there are outlier in holidays

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)

When you create dummy variables, using drop_first=True helps prevent problems where some variables are too closely related. If variables are too similar, it can confuse the model and make it harder to understand.

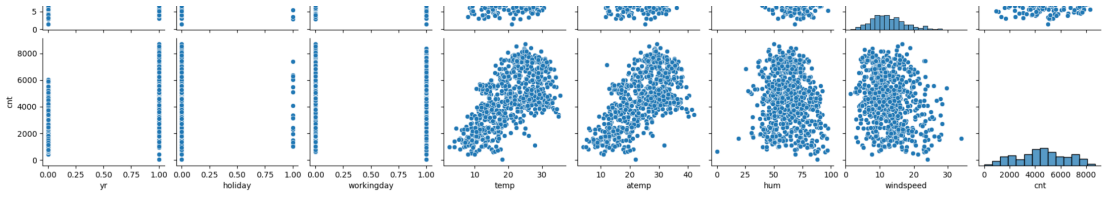
```
season_status = pd.get_dummies(bike_df['season'], drop_first = True)
season_status.head()
```

	spring	summer	winter
0	1	0	0
1	1	0	0
2	1	0	0
3	1	0	0
4	1	0	0

Notice how the fall season is not explicitly included. This is because when spring, summer, and winter are all 0, it implies that the season is fall.

By dropping the first category (fall), we avoid the issue of multicollinearity and keep our model simpler.

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 against cnt values, temp and atemp numeric values are highly corelated

---

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. The relationship between the independent variables and the dependent variable should be linear.
2. Residual distribution should follow normal distribution and mean should be '0', below is the distribution for our model

3. Multicollinearity should not be there in data, if independent variables are highly corelated with each other we will have Multicollinearity we validated Multicollinearity using Variance Inflation Factor

---

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. The coefficient for temperature is positive (3354.1194), indicating that higher temperatures are associated with higher cnt values.
2. The positive coefficient for the year (2051.1595) implies that as time progresses, the cnt values tend to increase.
3. Winter, September and Saturday have positive Coefficients

```
Intercept : [2609.07303236]
```

20]:

	MLR Coefficients
temp	3354.119377
yr	2051.159533
winter	567.398062
September	431.475735
Saturday	191.973208
February	-407.281541
July	-528.433026
November	-531.082388
December	-557.684535
Mist_cloudy	-695.077396
January	-709.939106
spring	-774.444149
windspeed	-1394.401671
Light_Rain	-2549.746736

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>

Linear regression tries to find the best straight line that fits through the data points on a graph. This line used to predict the value of one variable with another and the line is called "regression line"

Simple Linear Regression (SLR):

Simple Linear Regression uses just one independent variable to predict the value of a dependent variable.

Regression line formula: $Y = mX + c$

Y : Dependent variable, that value that we want to predict
X : Independent variable, that values which are used to predict the dependent variable
m : slope of the line, how Y changes on changing of X
c: The intercept, the value of Y when X is 0

Multiple Linear Regression (MLR):

Multiple Linear Regression is a statistical method that uses two or more independent variables to predict the value of a single dependent variable.

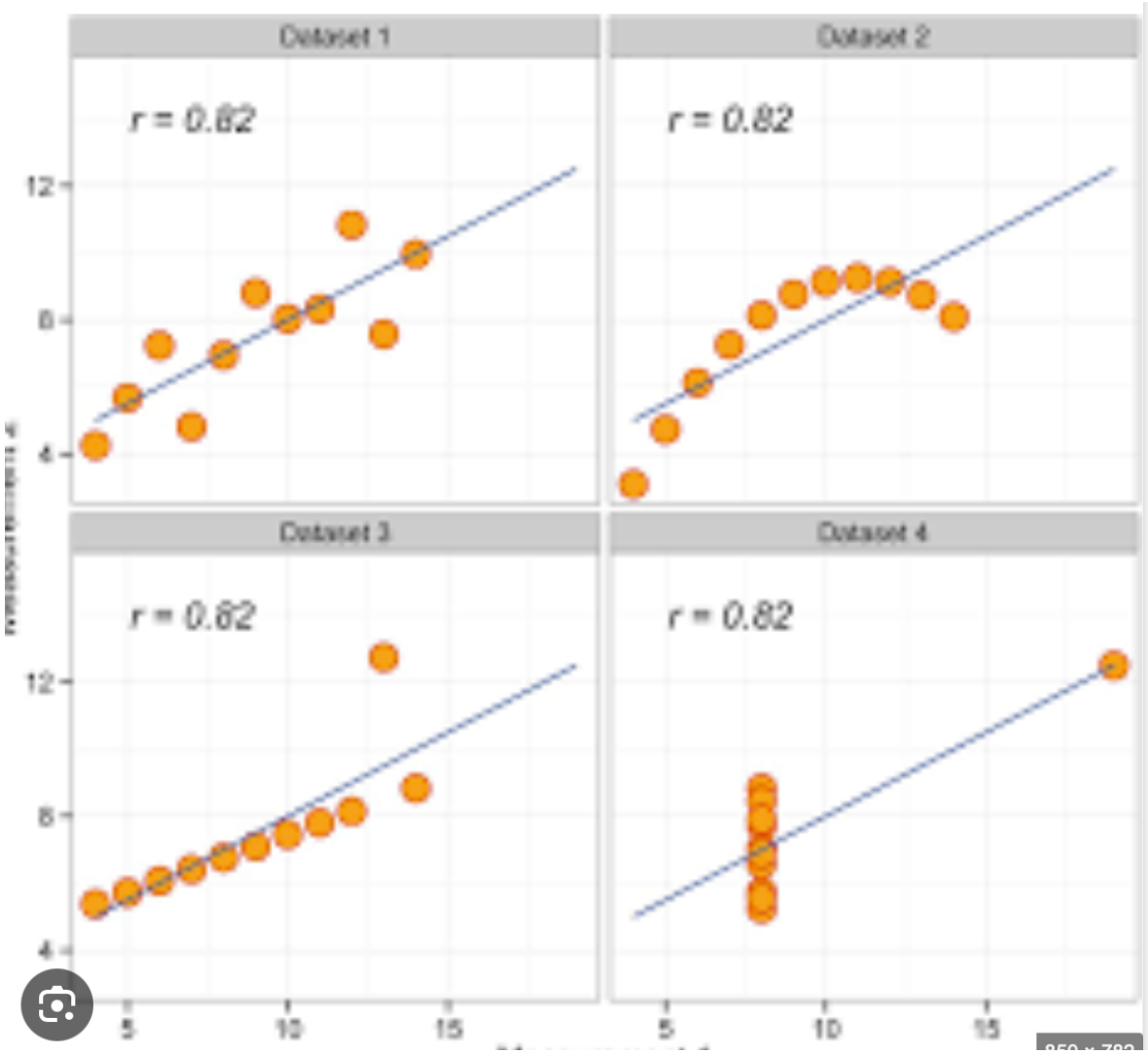
Regression line formula: $Y = m1X1 + m2X2 + \ldots + mnXn + c$

Y : Dependent variable, that value that we want to predict
X1, X2… Xn : Independent variable, that values which are used to predict the dependent variable
m1, m2….mn : slope of the line, for respective independent variable
c: The intercept

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 is a collection of four datasets that have nearly identical simple descriptive statistics (such as mean, variance, correlation, and linear regression line), yet they have very different distributions and appear very different when graphed.



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, also known as the Pearson correlation coefficient, is a measure of the strength and direction of the linear relationship between two variables. It ranges from -1 to 1, where:

- **+1** indicates a perfect positive linear relationship: as one variable increases, the other variable also increases.
- **0** indicates no linear relationship: the variables do not have any predictable linear association.
- **-1** indicates a perfect negative linear relationship: as one variable increases, the other variable decreases.

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 the process of transforming features in your data so they have similar ranges. This helps algorithms perform better and faster. Here's a breakdown of the concepts:

1. **Normalized Scaling:** Also known as Min-Max Scaling, it transforms the features to a fixed range, typically [0, 1].
2. **Standardized Scaling:** Also known as Z-score normalization, it transforms the features to have a mean of 0 and a standard deviation of 1.

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 sometimes be infinite due to perfect multicollinearity. This occurs when one predictor variable in your model can be perfectly explained by one or more of the other predictor variables.

- **VIF < 1**: This indicates a negative correlation, which is rare.
- **1 ≤ VIF < 5**: This suggests a low to moderate correlation and is usually okay.
- **VIF ≥ 5**: This indicates high correlation, which could cause problems; you should consider re moving or combining these predictors.

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 (Quantile-Quantile) plot is used to see if a dataset follows a particular theoretical distribution. It works by comparing the quantiles of the observed data to the quantiles of this other distribution. I said 'theoretical distribution' to be exact, but often when we create a Q-Q plot

Use and Importance of a Q-Q Plot in Linear Regression:

- **Visualizing Residual Distribution**: Compares the residuals to a normal distribution; points lie on a straight line if normally distributed.

- **Assessing Normality**: Checks if residuals are normally distributed, which is a key assumption in linear regression.
- **Identifying Outliers**: Points deviating from the line can show outliers in the data.
- **Model Diagnostics**: Ensures residuals meet the normality assumption for reliable regression results.
