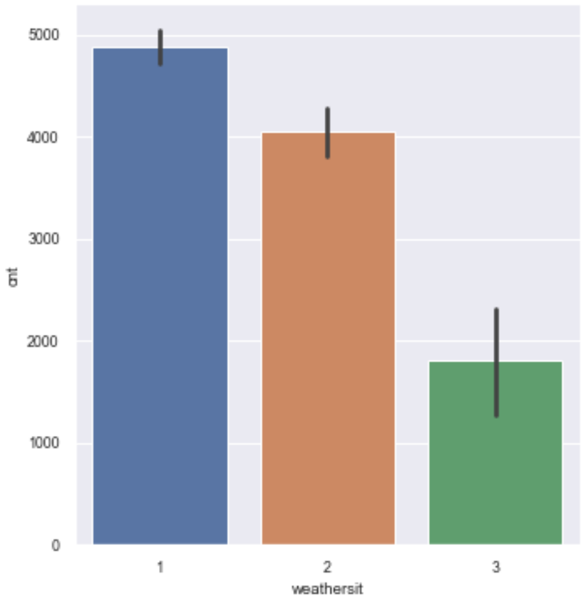
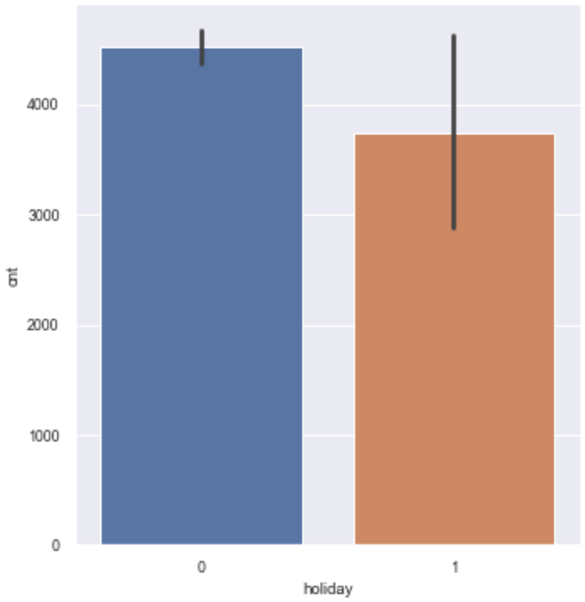
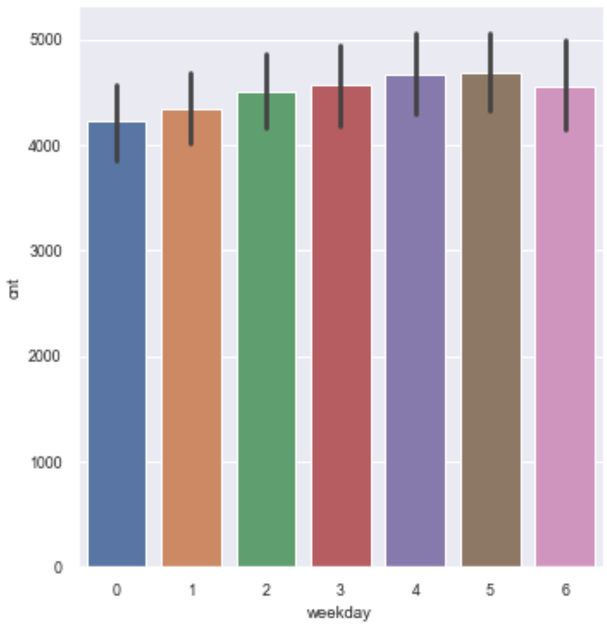
Assignment-based Subjective Questions

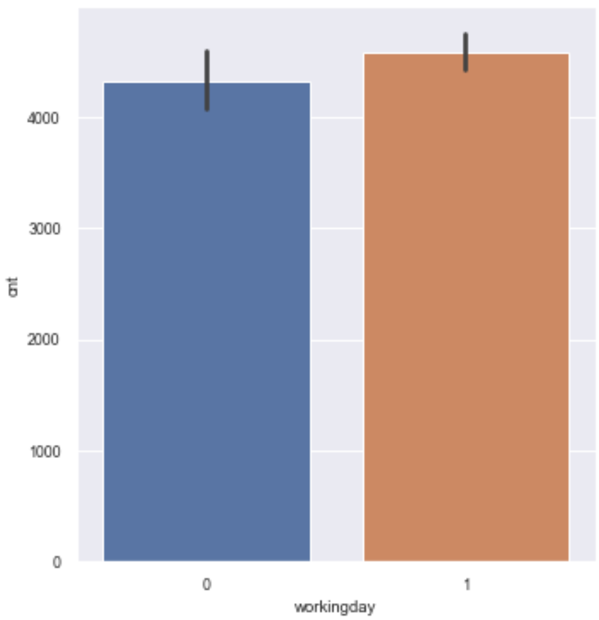
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

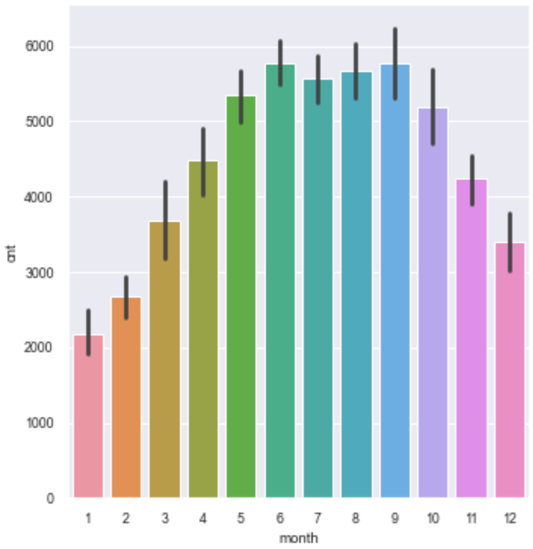
- Fall Season has more bookings.
- Year 2019 has more bookings than 2018.
- Bookings are higher in the middle of the year.
- Bookings are more during non-holidays.
- “Weathersit” - when 1 i.e. clear there are more bookings.
- Thu, Fir, Sat, and Sun have a slightly greater number of bookings as compared to the start of the week.
- Not much of deviation on working day vs non-working day.

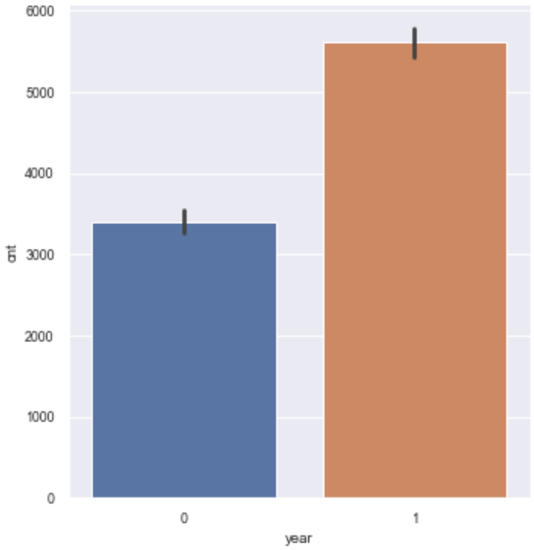


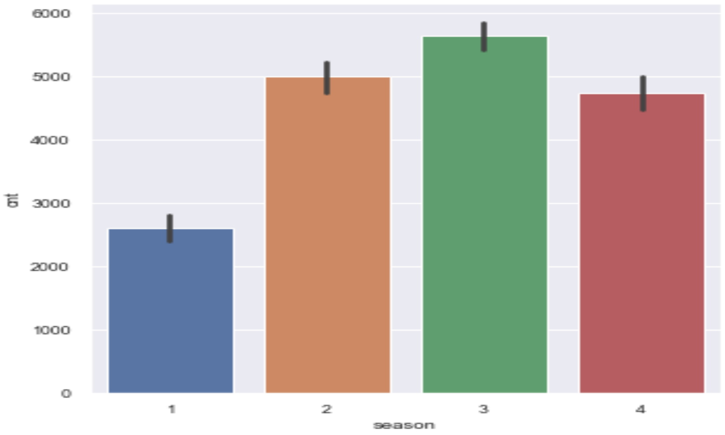



5000
4000
3000
2000
1000
0
cnt
0
1
2
3
4
5
6
weekday


4000
3000
2000
1000
0
cnt
0
1
workingday


6000
5000
4000
3000
2000
1000
0
cnt
1
2
3
4
5
6
7
8
9
10
11
12
month


6000
5000
4000
3000
2000
1000
0
cnt
0
1
year


6000
5000
4000
3000
2000
1000
0
cnt
1
2
3
4
season

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

When creating dummy variables from categorical features, the drop_first=True parameter is used to address the issue of multicollinearity in regression models. Multicollinearity occurs when two or more predictor variables are highly correlated, leading to instability in the model and making it difficult to interpret the individual effects of each variable.

To understand why drop_first=True is important, let's consider an example with a binary categorical variable (i.e., a variable with only two categories).

Suppose we have a categorical variable "Color" with two categories: "Red" and "Blue." If we create dummy variables without dropping the first category (using drop_first=False), we would end up with two dummy variables: "Red" and "Blue."

- When "Color" is "Red," both dummy variables will be set to 1, and the intercept term in the regression model will represent the effect of "Red."
- When "Color" is "Blue," the dummy variable "Red" will be 0, and the effect of "Blue" is represented by the intercept term plus the coefficient of the "Blue" dummy variable.

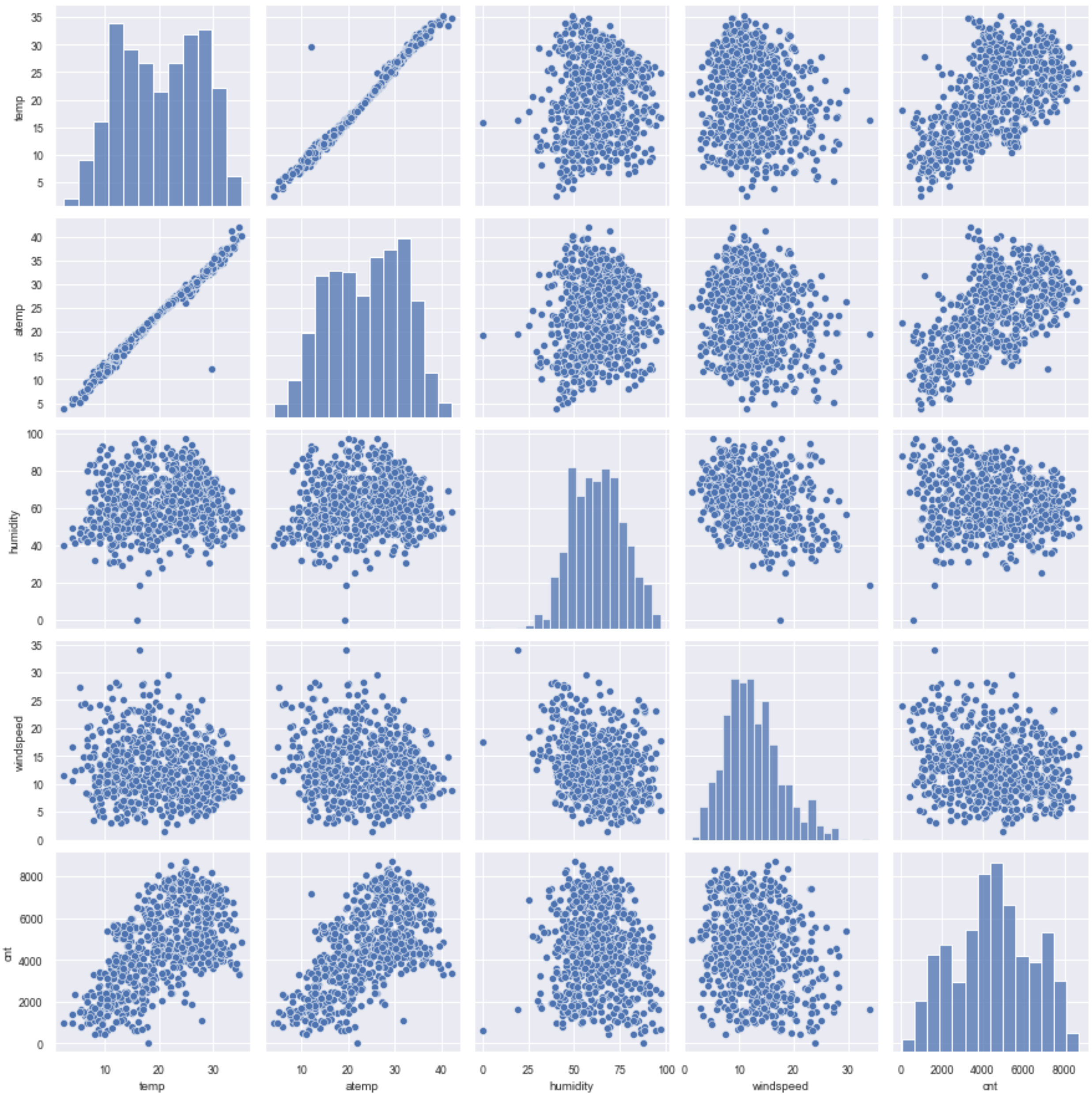
The problem with this approach is that the two dummy variables are not independent; their values are perfectly negatively correlated. If one of them is 1, the other must be 0, which introduces multicollinearity. This correlation can cause numerical instability and inflated standard errors, making it challenging to interpret the individual effects of the variables accurately.

To avoid multicollinearity, we use drop_first=True, which means we drop one of the dummy variables. In our example, if we drop "Red" (the first category), we'll only have one dummy variable: "Blue." Now, the presence of "1" in the "Blue" dummy variable indicates "Blue," and "0" represents "Red." This eliminates the perfect negative correlation and the multicollinearity issue.

In general, if you have a categorical variable with "n" categories, creating "n-1" dummy variables with drop_first=True ensures independence among the dummy variables and helps in the accurate interpretation of coefficients in regression models.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

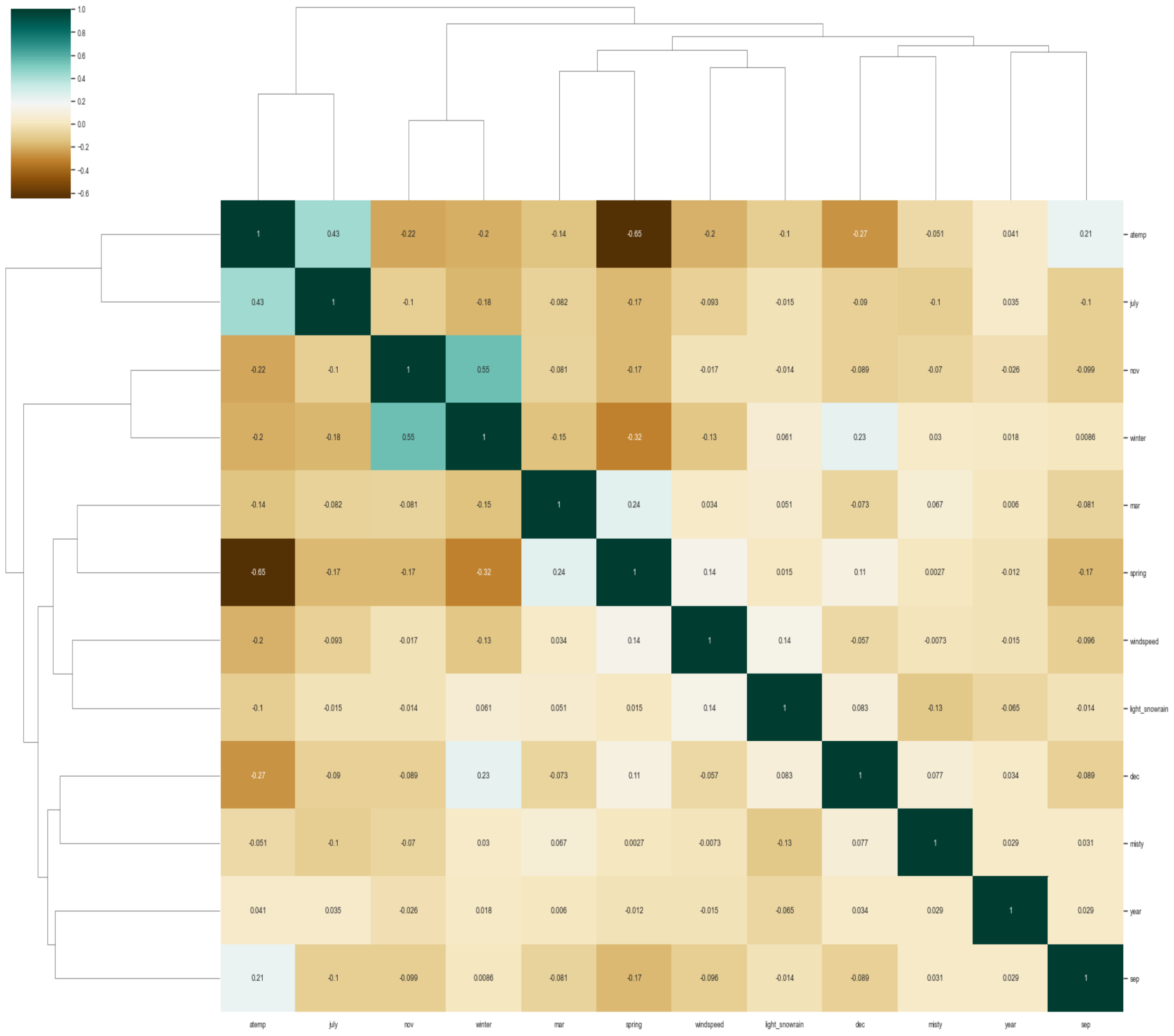
In the below pair-plot, temp and atemp seem to have high correlation with the target variable cnt.



4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

I performed the below checks to validate the assumptions of Linear Regression.

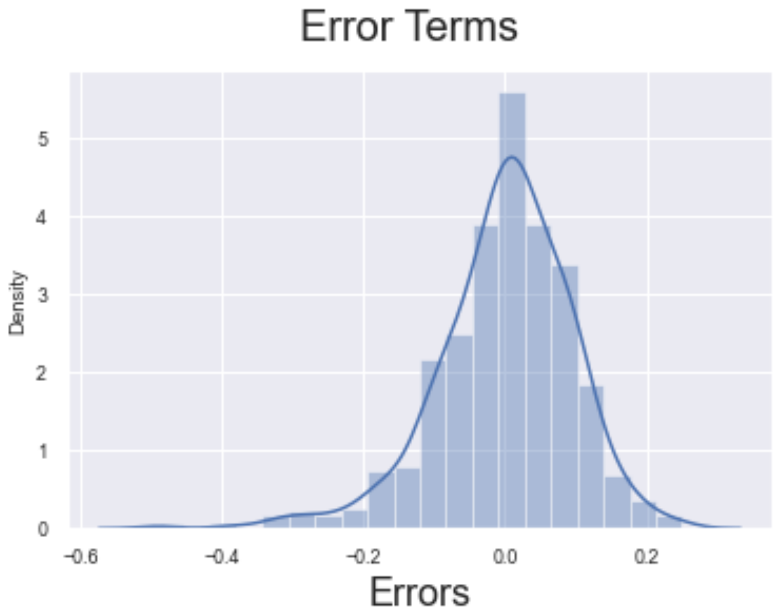
Multi Collinearity: No two variables are highly correlated (correlation factor > 0.8 means highly correlated)



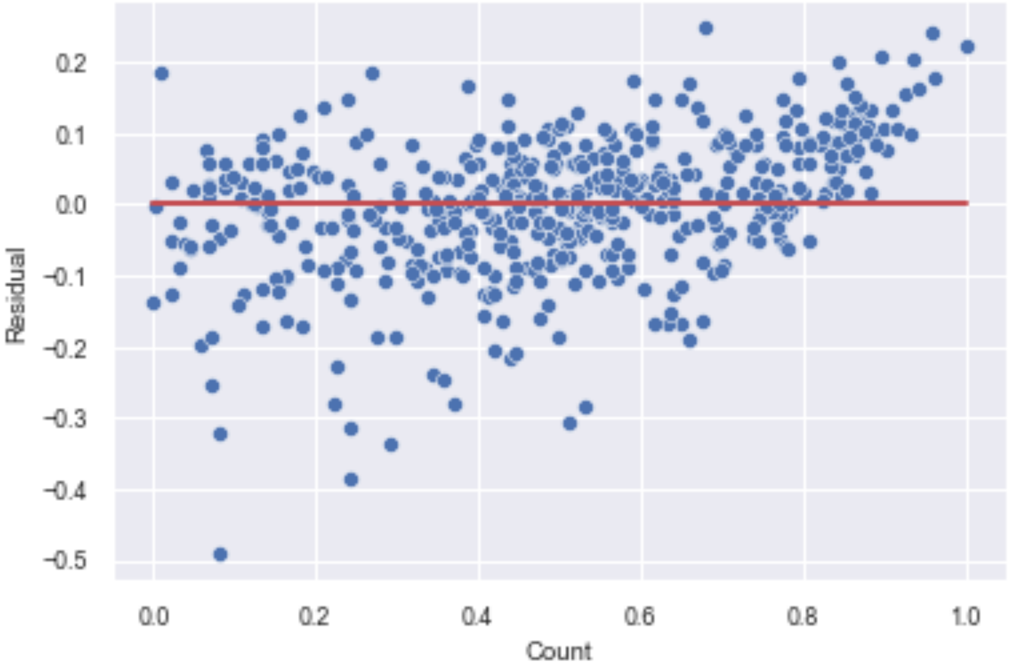
	atemp	july	nov	winter	mar	spring	windspeed	light_snowrain	dec	misty	year	sep
atemp	1	0.43	-0.22	-0.2	-0.14	-0.65	-0.2	-0.1	-0.27	-0.051	0.041	0.21
july	0.43	1	-0.1	-0.18	-0.082	-0.17	-0.093	-0.015	-0.09	-0.1	0.035	-0.1
nov	-0.22	-0.1	1	0.55	-0.081	-0.17	-0.017	-0.014	-0.089	-0.07	-0.026	-0.099
winter	-0.2	-0.18	0.55	1	-0.15	-0.32	-0.13	0.061	0.23	0.03	0.018	0.0086
mar	-0.14	-0.082	-0.081	-0.15	1	0.24	0.034	0.051	-0.073	0.067	0.006	-0.081
spring	-0.65	-0.17	-0.17	-0.32	0.24	1	0.14	0.015	0.11	0.0027	-0.012	-0.17
windspeed	-0.2	-0.093	-0.017	-0.13	0.034	0.14	1	0.14	-0.057	-0.0073	-0.015	-0.096
light_snowrain	-0.1	-0.015	-0.014	0.061	0.051	0.015	0.14	1	0.083	-0.13	-0.065	-0.014
dec	-0.27	-0.09	-0.089	0.23	-0.073	0.11	-0.057	0.083	1	0.077	0.034	-0.089
misty	-0.051	-0.1	-0.07	0.03	0.067	0.0027	-0.0073	-0.13	0.077	1	0.029	0.031
year	0.041	0.035	-0.026	0.018	0.006	-0.012	-0.015	-0.065	0.034	0.029	1	0.029
sep	0.21	-0.1	-0.099	0.0086	-0.081	-0.17	-0.096	-0.014	-0.089	0.031	0.029	1

Linearity: Linearity among variables was checked

Normality of error terms: It should follow a normal distribution.



Homoscedasticity: No visible pattern observed from above plot for residuals.



Independence Of Residuals: Durbin-Watson value of final model lr_6 is 2.066, which signifies there is no autocorrelation.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Top 3 features are:
“year” – coefficient: 0.2416
“atemp” – coefficient: 0.4418
“light_snowrain”: -0.2659

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression is a popular supervised machine learning algorithm used for predictive modeling and regression tasks. It aims to establish a linear relationship between a dependent variable (also known as the target) and one or more independent variables (predictors or features). The algorithm's goal is to find the best-fitting line that minimizes the difference between the predicted values and the actual values of the target variable.

Steps and concepts involved in the linear regression algorithm:

Data Preparation:
Split the data into two sets: the training set and the test set. The training set is used to build the model, while the test set evaluates its performance.
Linear Equation:

Linear regression represents the relationship between the target variable and predictors using a linear equation in the form:
$Y = \beta0 + \beta1X1 + \beta2X2 + ... + \beta n*Xn$

In this equation, Y is the target variable, β0 is the intercept (the value of Y when all predictors are zero), β1, β2, ..., βn are the coefficients for each predictor (representing the slope of the line), and X1, X2, ..., Xn are the values of the corresponding predictors.

Cost Function:
The model's goal is to find the best values for the coefficients (β0, β1, β2, ..., βn) to minimize the difference between the predicted and actual target values.
The cost function, often referred to as the "mean squared error" (MSE) or "mean squared residual," calculates the average squared difference between the predicted and actual target values over the entire training dataset.
The objective is to minimize this cost function to get the best-fitting line.

Gradient Descent (Optimization):
Gradient descent is an optimization algorithm used to find the optimal coefficients that minimize the cost function.
It works iteratively by adjusting the coefficients in the opposite direction of the gradient of the cost function until it reaches a minimum.
The learning rate determines the step size in each iteration, ensuring the algorithm converges to the minimum without overshooting.

Training the Model:
During the training phase, the linear regression algorithm uses the training dataset to adjust the coefficients (β0, β1, β2, ..., βn) using gradient descent.
The process continues until the model reaches the minimum cost (i.e., the best-fitting line).
Making Predictions:

Once the model is trained and the coefficients are determined, it can be used to make predictions on new data by plugging in the predictor values into the linear equation.

Model Evaluation:
The performance of the linear regression model is evaluated using the test dataset.
Common evaluation metrics include the mean squared error (MSE), mean absolute error (MAE), R-squared (coefficient of determination), and others.

Interpretation:
After obtaining the final model, the coefficients (β0, β1, β2, ..., βn) can be interpreted to understand the strength and direction of the relationship between the target variable and each predictor.

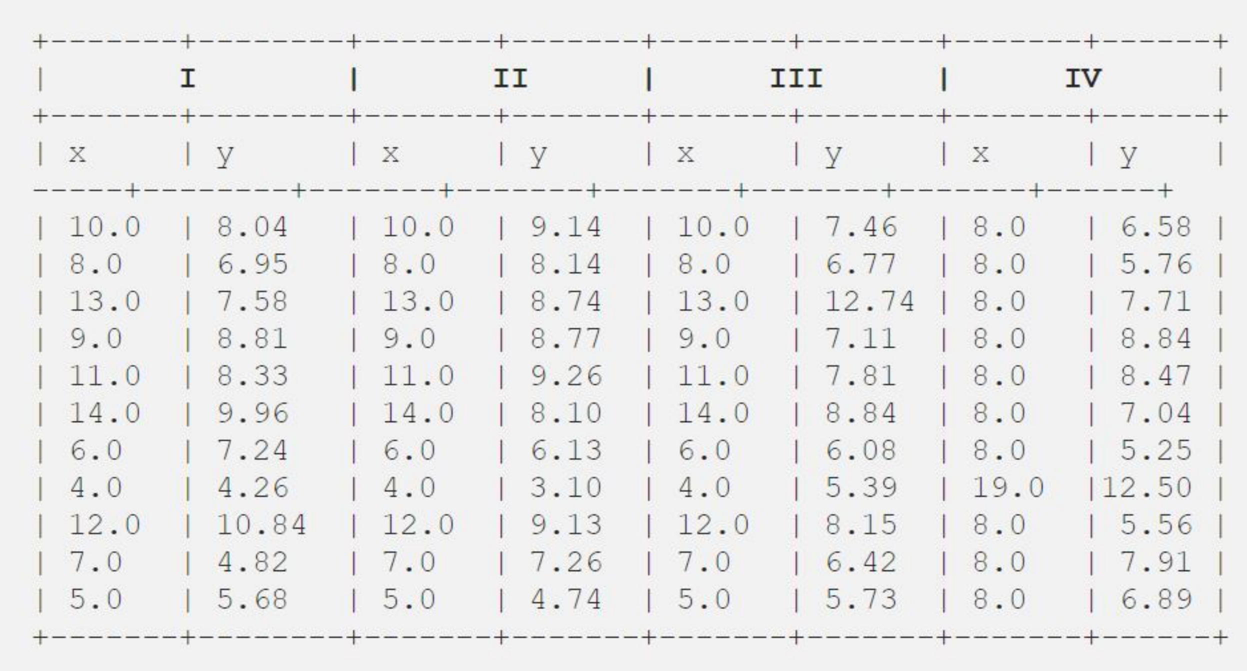
Linear regression is a fundamental and straightforward algorithm, and it serves as a basis for more complex regression models in machine learning. Note that it assumes a linear relationship between the target and predictors, which may not always hold true in practice.

2. Explain the Anscombe's quartet in detail. (3 marks)

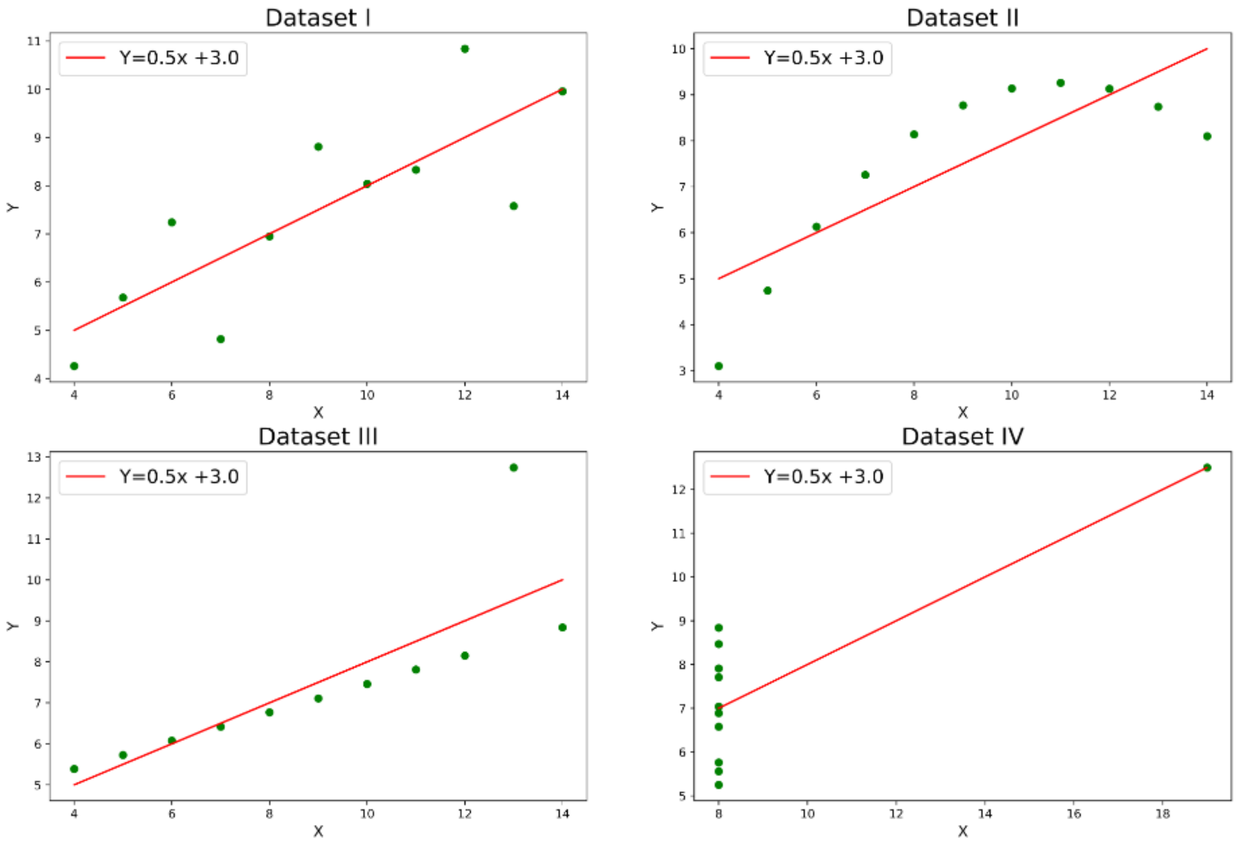
Anscombe's quartet comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph. The datasets were created by the statistician Francis 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.

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.



```
+-------+--------+-------+--------+-------+--------+-------+------+
|       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 |
+-------+--------+-------+--------+-------+--------+-------+------+
```



Explanation of the 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.

3. What is Pearson’s R? (3 marks)

Pearson's R is a statistical measure of the linear correlation between two variables. It is a number between -1 and 1, where 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation.

The Pearson's R coefficient is calculated using the following formula:

r = (n(∑xy) - (∑x)(∑y)) / (√[n(∑x^2) - (∑x)^2][n(∑y^2) - (∑y)^2])
where:

n is the number of observations
∑xy is the sum of the products of the x and y values
∑x is the sum of the x values
∑y is the sum of the y values
∑x^2 is the sum of the squares of the x values
∑y^2 is the sum of the squares of the y values
Pearson's R is a useful tool for understanding the relationship between two variables. It can be used to determine whether there is a significant correlation between two variables, and to measure the strength of the correlation.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling is the process of transforming the features in a dataset so that their values share a similar scale. This is done to ensure that all features contribute equally to the learning process and to prevent features with larger scales from dominating the model.

There are two main types of scaling:

Normalized scaling (also known as min-max scaling) scales the features so that they lie between a minimum and maximum value, typically 0 and 1. This is done by subtracting the minimum value from each feature and then dividing by the difference between the minimum and maximum values.

Standardized scaling (also known as z-score normalization) scales the features so that they have a mean of 0 and a standard deviation of 1. This is done by subtracting the mean from each feature and then dividing by the standard deviation.

Scaling is performed for several reasons, including:

To improve the accuracy of machine learning models. Scaling can help to ensure that all features contribute equally to the learning process, which can lead to more accurate models.
To improve the performance of machine learning models. Scaling can help to speed up the training process and to improve the generalization performance of the models.
To make features more comparable. Scaling can help to make features with different scales more comparable, which can make it easier to interpret the results of machine learning models.
The main difference between normalized scaling and standardized scaling is that normalized scaling scales the features so that they lie between a minimum and maximum value, while standardized scaling scales the features so that they have a mean of 0 and a standard deviation of 1.

Normalized scaling is often used when the features have different scales, and you want to make them more comparable. Standardized scaling is often used when the features are normally distributed, and you want to improve the accuracy of machine learning models.

In general, it is a good idea to try both normalized scaling and standardized scaling and see which one works better for a particular dataset.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Yes, the value of Variance Inflation Factor (VIF) can sometimes be infinite. This happens when two or more independent variables are perfectly correlated, meaning that they are perfectly predictable from each other. In this case, the R-squared of the regression of one variable on the other is 1, which leads to a VIF of infinity.

There are a few reasons why two variables might be perfectly correlated. One reason is that they are measuring the same thing. For example, if you have two variables that measure the height of a person, they will be perfectly correlated. Another reason is that they are caused by the same underlying factor. For example, if you have two variables that measure the income of a person's parents, they will be perfectly correlated.

A VIF of infinity is a sign of severe multicollinearity. Multicollinearity can cause problems with statistical inference because it can make it difficult to determine which variable is actually driving the results. In general, it is a good idea to avoid VIFs that are greater than 10.

If you have a VIF that is infinite, there are a few things you can do to address the problem. One thing you can do is to remove one of the variables that is perfectly correlated. Another thing you can do is to combine the two variables into a single variable. Finally, you can use a technique called ridge regression, which is designed to deal with multicollinearity.
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
A Q-Q plot, or quantile-quantile plot, is a graphical tool used to assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. It is a scatter plot of the quantiles of the data set against the quantiles of a theoretical distribution. If the data set comes from the theoretical distribution, the points in the Q-Q plot will fall along a straight line.

In linear regression, a Q-Q plot can be used to assess the assumption that the residuals are normally distributed. If the residuals are normally distributed, the points in the Q-Q plot will fall along a straight line. If the points in the Q-Q plot deviate from a straight line, it may indicate that the residuals are not normally distributed. This could be due to a number of factors, such as outliers, non-linearity, or heteroscedasticity.

Here are some of the uses and importance of a Q-Q plot in linear regression:

To assess the assumption of normality: A Q-Q plot can be used to assess the assumption that the residuals are normally distributed. If the residuals are normally distributed, the points in the Q-Q plot will fall along a straight line. If the points in the Q-Q plot deviate from a straight line, it may indicate that the residuals are not normally distributed. This could be due to a number of factors, such as outliers, non-linearity, or heteroscedasticity.
To identify outliers: Q-Q plots can also be used to identify outliers. Outliers are data points that fall far away from the rest of the data. Outliers can have a significant impact on the results of a linear regression model. By identifying outliers, you can take steps to remove them from the dataset or to adjust your model to account for their presence.
To check for non-linearity: Q-Q plots can also be used to check for non-linearity in the data. Non-linearity occurs when the relationship between the independent and dependent variables is not linear. If the points in the Q-Q plot do not fall along a straight line, it may indicate that the relationship between the independent and dependent variables is not linear. This could be due to a number of factors, such as a quadratic relationship or a logarithmic relationship.

Overall, Q-Q plots are a useful tool for assessing the assumptions of linear regression. By using Q-Q plots, you can identify potential problems with your model and take steps to correct them.