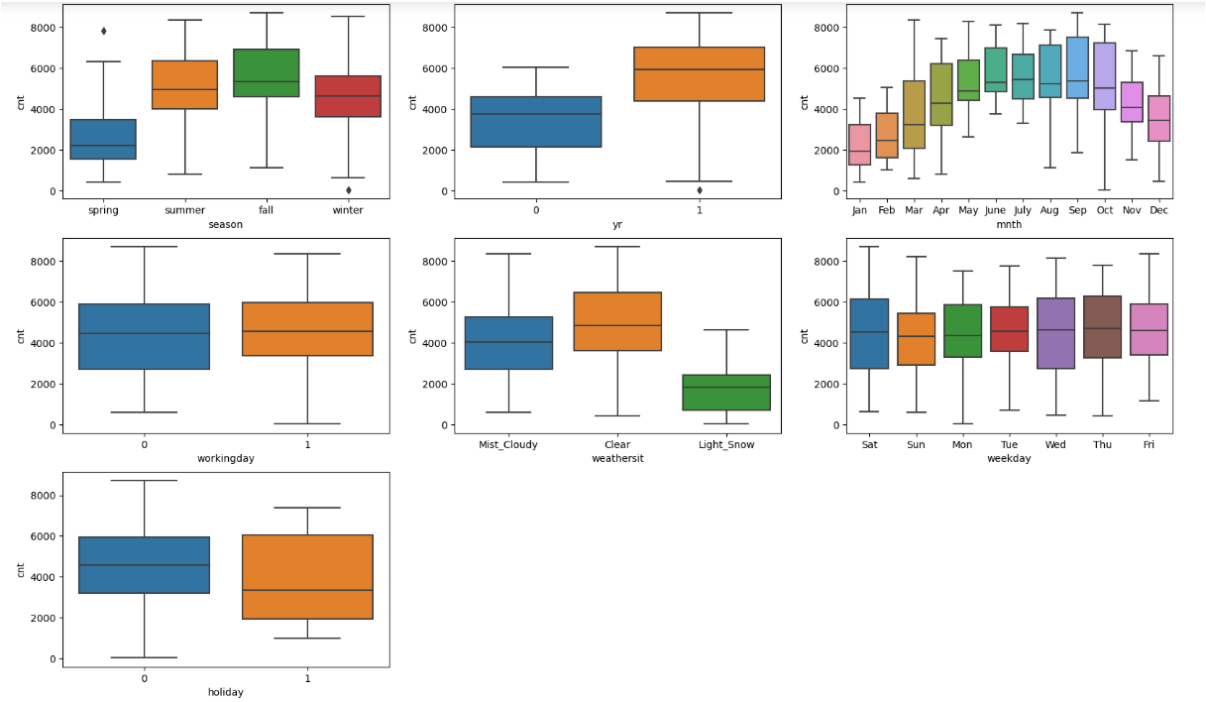
Assignment-based Subjective Questions

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

Ans:



From the analysis of the categorical variables—

- Bike rentals are more during the fall season following by the summer.
- In the year 2019 there were more bike rentals as compared to 2018.
- Bike rentals are more in clear weather as compared to others weather situations.
- On Wednesday, Thursday, and Saturday the bike rentals likely to be more.
- In the month of August, September and October bike rentals are more.
- On Working day, bike rentals are slightly more as compared to other days.

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

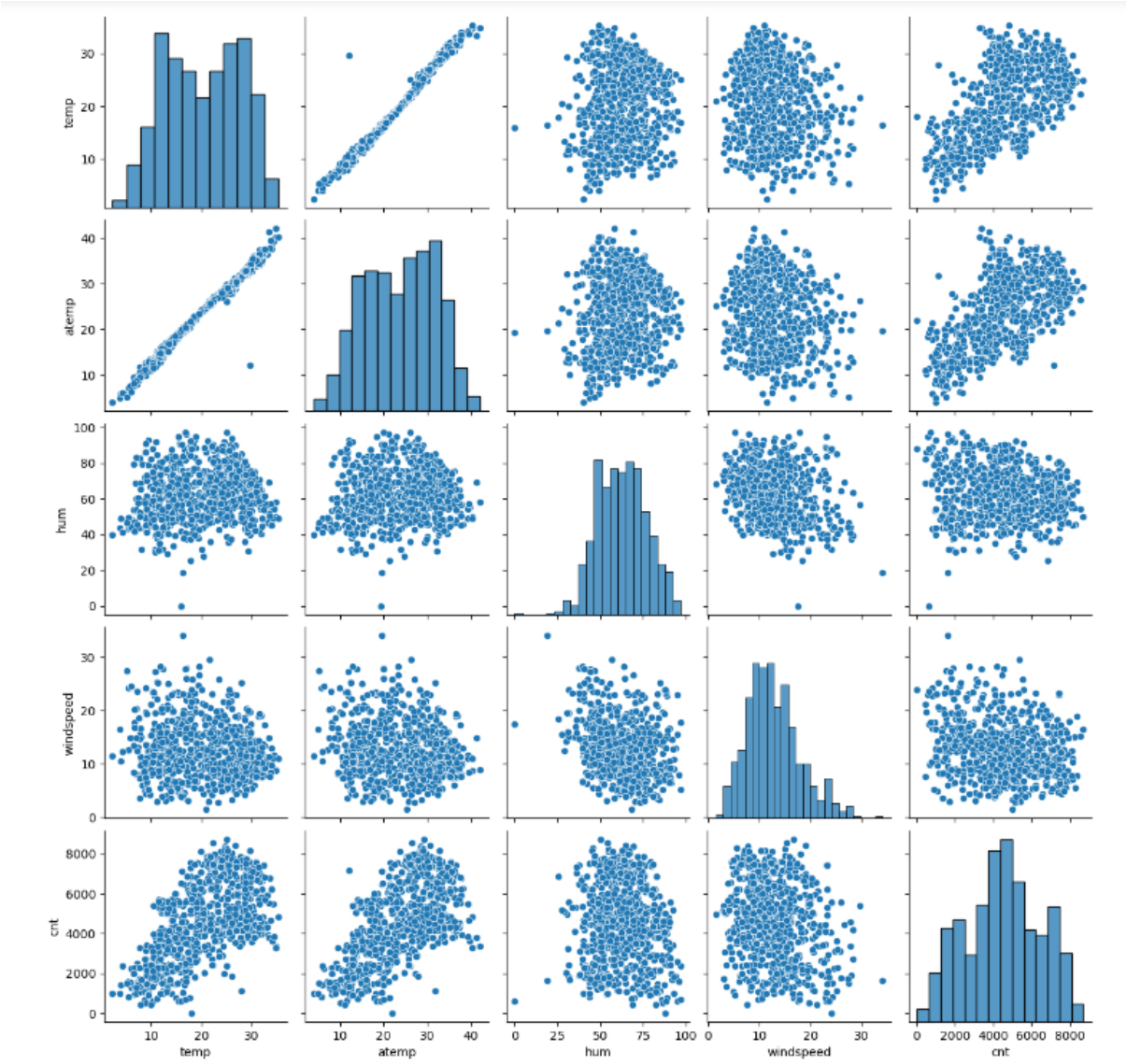
Ans: Whether to get k-1 dummies out of k categorical levels by removing the first level. Basically, drop_first = True helps in reducing the extra column created during the dummy variable creation and hence avoid redundancy of any kind.

For Example, consider a Categorical column with 3 types of values, we want to create dummy variable for that column. If one variable is neither furnished nor semi_furnished, then it is obvious unfurnished. So, we do not need 3rd variable to identify the unfurnished.

Value	Indicator Variable	
Furnishing Status	furnished	semi-furnished
furnished	1	0
semi-furnished	0	1
unfurnished	0	0

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

   Ans: The temp and atemp variable has the highest correlation with the target variable.



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

   Ans: Validated the assumptions by checking—

   - There is no multicollinearity by checking VIF. Multicollinearity occurs when the independent variables are too highly correlated with each other. We calculated the VIF (Variance Inflation Factor) to quantify how strongly the feature variables in the new model are associated with one another.

- Residuals distribution should follow normal distribution and centred around 0 (mean = 0). We validated this assumption about residuals by plotting a distribution plot of residuals and saw if residuals are following normal distribution or not.

- There should be linear relationship between independent and dependent variables. We visualised the numeric variables using a pairplot to see if the variables are linearly related or not.

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

   Ans: the top 3 features contributing significantly are light_snow, spring and year towards the demand of shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Ans: Linear regression is a quiet and simple statistical regression method used for predictive analysis and shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression. If there is a single input variable (x), such linear regression is called **simple linear regression**. And if there is more than one input variable, such linear regression is called **multiple linear regression.**

   Linear regression equation => $Y = mX + c$

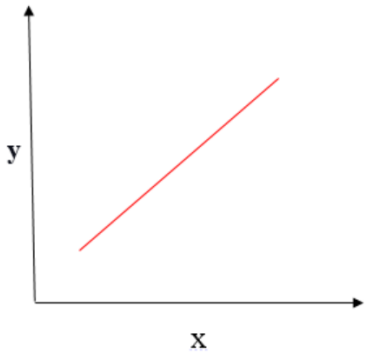
   Y = dependent variable

   X = independent variable

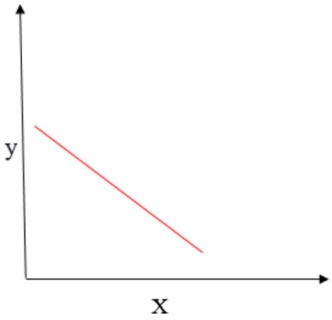
   M = coefficient (slope)

   C = intercept of the line

   If the dependent variable increases on the Y-axis and the independent variable progress on X-axis, then such a relationship is termed a Positive linear relationship. As shown in figure below.



If the dependent variable decreases on the Y-axis and the independent variable increases on the X-axis, such a relationship is called a negative linear relationship. As shown in figure below.



2. Explain the Anscombe's quartet in detail.

Ans: 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.

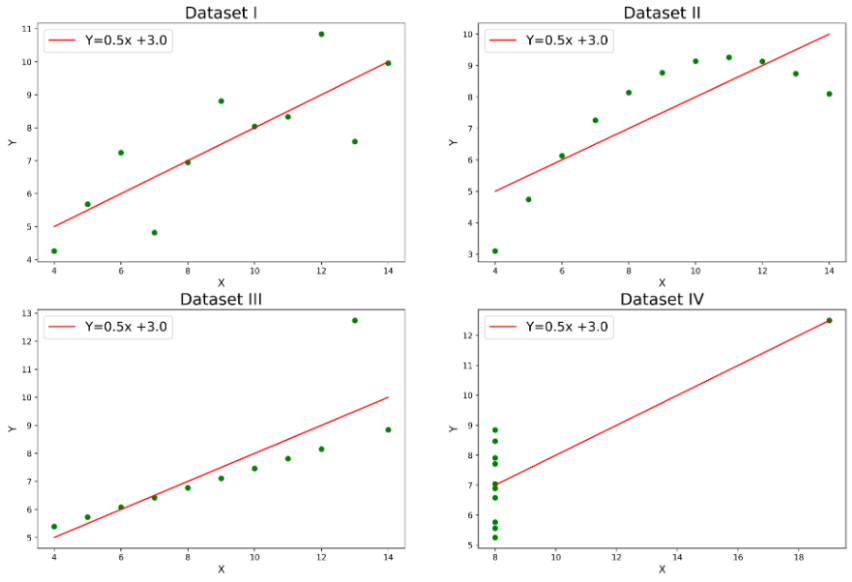
The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, 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.

Though the below descriptive statistical value looks same.

	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

But on plotting the scatter plot and linear regression it looks totally different.



Anscombe's quartet Plot

3. What is Pearson's R?

Ans: The **Pearson correlation coefficient (r)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Pearson correlation coefficient (r)	Correlation type	Interpretation	Example
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the **same direction**.	Baby length & weight: The longer the baby, the heavier their weight.
0	No correlation	There is **no relationship** between the variables.	Car price & width of windshield wipers: The price of a car is not related to the width of its windshield wipers.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the **opposite direction**.	Elevation & air pressure: The higher the elevation, the lower the air pressure.

Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

   Ans: Scaling is a technique performed in pre-processing during building a machine learning model to standardize the independent feature variables in the dataset in a fixed range.

   The dataset could have several features which are highly ranging between high magnitudes and units. If there is no scaling performed on this data, it leads to incorrect modelling as there will be some mismatch in the units of all the features involved in the model.

   The difference between normalization and standardization is that while normalization brings all the data points in a range between 0 and 1, standardization replaces the values with their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ)** zero and standard deviation one (**σ**).

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

   Ans: The value of VIF is infinite when there is a perfect correlation between the two independent variables. The R-squared value is 1 in this case. This leads to VIF infinity as VIF equals to 1/(1-R2). This concept suggests that is there is a problem of multi-collinearity and one of these variables need to be dropped to define a working model for regression.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

   Ans: Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any probability distribution like normal, uniform, exponential.

   The power of Q-Q plots lies in their ability to summarize any distribution visually.

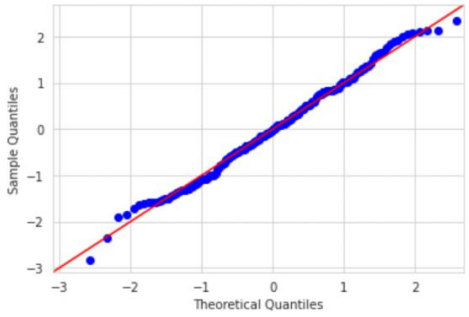
   QQ plots is very useful to determine.

   If two populations are of the same distribution

   - If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it's met using this.
   - Skewness of distribution

   In Q-Q plots, we plot the theoretical Quantile values with the sample Quantile values. Quantiles are obtained by sorting the data. It determines how many values in a distribution are above or below a certain limit.

   If the datasets we are comparing are of the same type of distribution type, we would get a roughly straight line. Here is an example of normal distribution.



   We can see that since we are plotting the data with the theoretical quantiles of a normal distribution, we are getting almost a straight line.