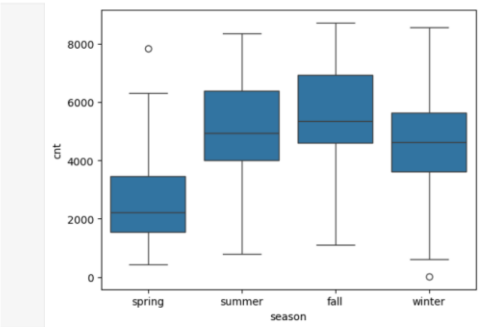
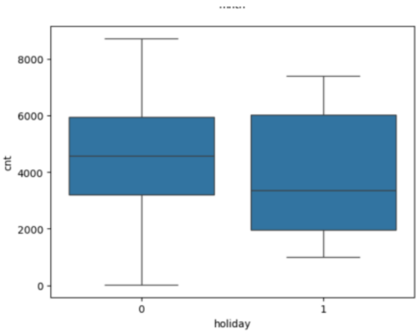
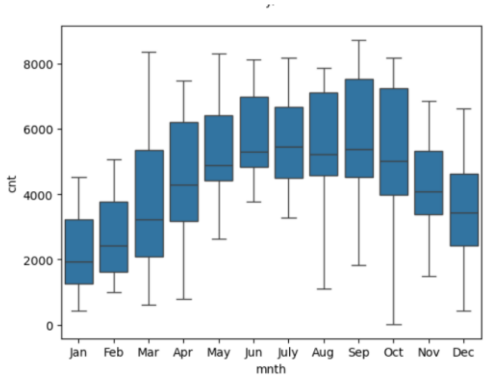
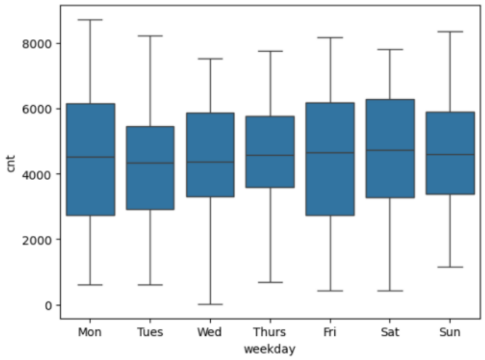
Assignment-based Subjective Questions

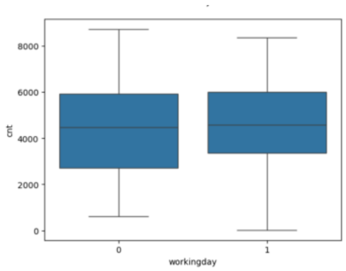
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

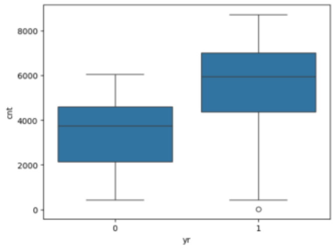


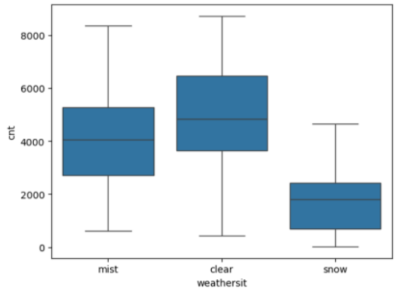












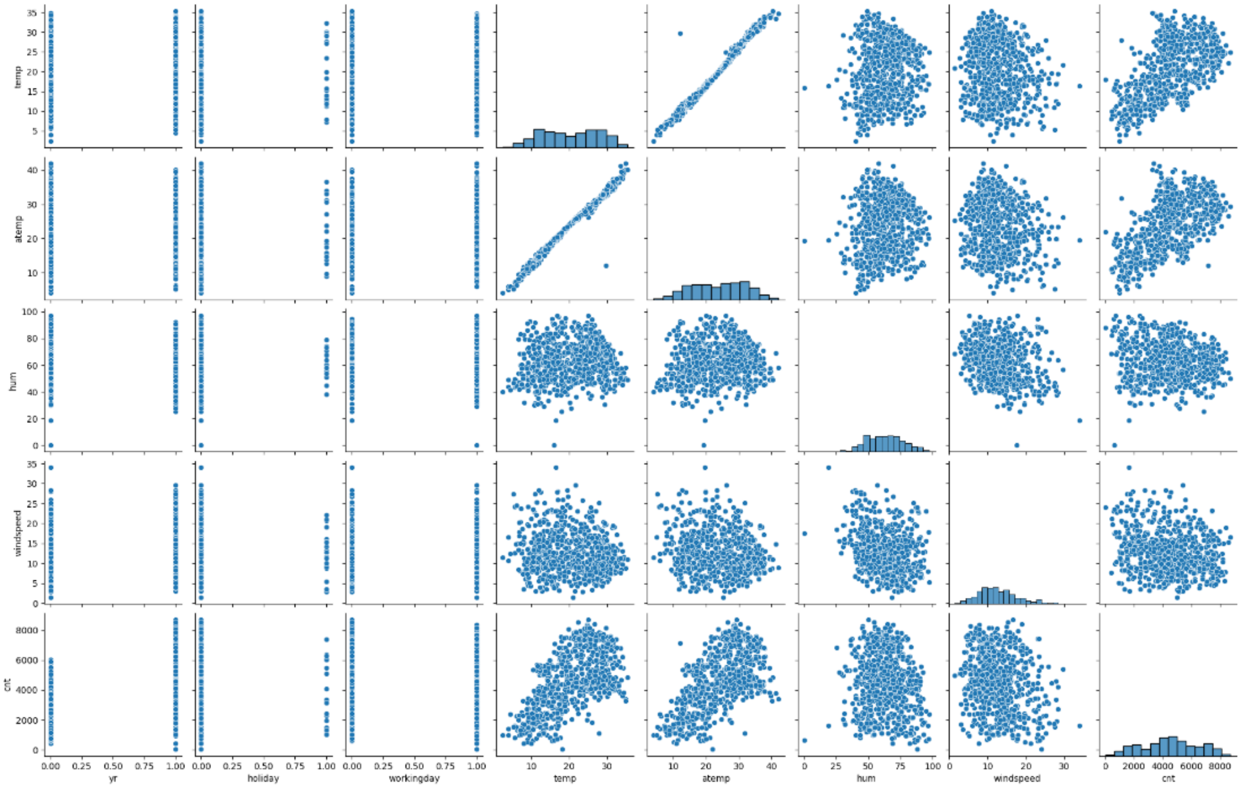
- Demand for bikes reduces on holiday
- Demand for bikes peaked in September and reduce in December.
- Demand for bikes is more in clear weather and reduce on snow and heavy rainfall.
- More bikes were on demand in 2019 than 2018
- Demand for bikes are almost same on working and non-working day. There is not much difference in the booking on weekend or weekdays
- Demand for bikes is almost constant during entire week.
- Demand for bikes are more in fall, followed by summer , winter and spring.

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

   Answer: drop_first is important during creating dummy variables to avoid multicollinearity in regression model.If Including all k dummy variables in regression model without dropping once, they can be perfectly correlated with each other. Which can lead to issue in interpreting the model.

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

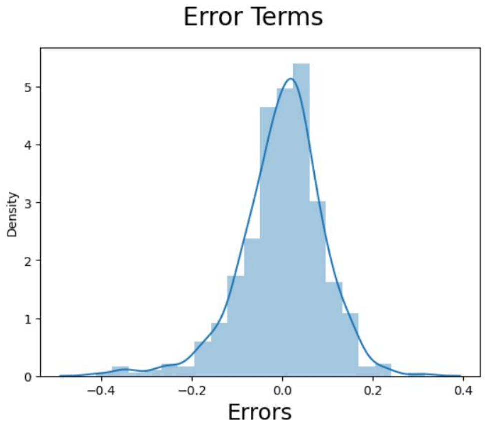
As can be seen from the below snapshot of pair plot **temp** and **atemp** are highly correlated with each other.



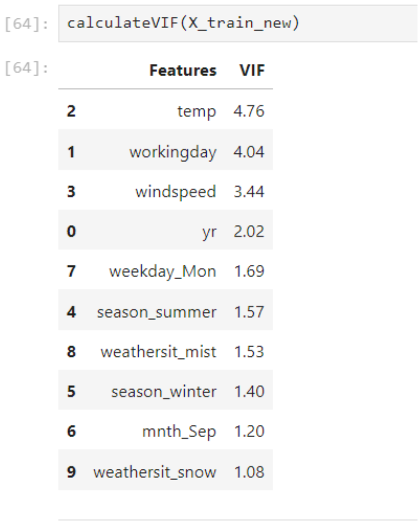
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Validated the assumption of Linear Regression Model based on below 5 assumptions -

Normality of error terms : Error terms should be normally distributed



Multicollinearity check: There should be insignificant multicollinearity among variables.



```
[64]: calculateVIF(X_train_new)
```

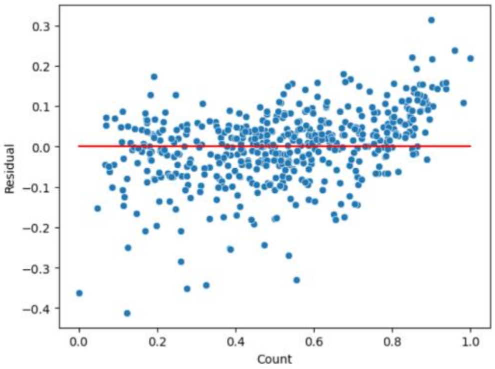
[64]:

	Features	VIF
2	temp	4.76
1	workingday	4.04
3	windspeed	3.44
0	yr	2.02
7	weekday_Mon	1.69
4	season_summer	1.57
8	weathersit_mist	1.53
5	season_winter	1.40
6	mnth_Sep	1.20
9	weathersit_snow	1.08

Linear relationship validation: Linearity should be visible among variables.

Used pair plot to check if the variables are linearly related

Homoscedasticity: There should be no visible pattern in residual values.



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

The top 3 significant features are:

1. Temp – coefficient 0.5499 **temperature in Celsius**
2. season winter – coefficient 0.1307 **(Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog)** that is determined by weathersit
3. yr – coefficient 0.2331 **year**

General Subjective Questions

Question 1: Explain the linear regression algorithm in detail.

Answer: Linear regression is a statistical model that examines the straight-line relationship between a dependent variable and a set of independent variables. When the values of these independent variables change (go up or down), the dependent variable also changes correspondingly.

Mathematically, this relationship is represented by the equation:

$$Y = mX + c$$

Here's what each part of the equation means:

- Y is the variable we want to predict (the dependent variable).
- X is the variable we use to make predictions (the independent variable).
- m is the slope of the line, showing how much Y changes when X changes.
- c is the intercept, meaning the value of Y when X is zero.

The relationship between X and Y can be either positive or negative:

- Positive relationship: Both variables increase together.
- Negative relationship: One variable increase while the other decreases.

Linear regression comes in two types:

- Simple Linear Regression: Uses one independent variable to predict the dependent variable.
- Multiple Linear Regression: Uses multiple independent variables to predict the dependent variable.

Assumptions of linear regression include:

- Little to no multicollinearity: Independent variables should not be highly correlated.
- Little to no autocorrelation: Residual errors (the differences between predicted and actual values) should not be correlated.
- Linear relationship: The relationship between independent and dependent variables should be linear.
- Normally distributed error terms: Errors should follow a normal distribution.
- Homoscedasticity: Residuals should have constant variance (no pattern in the residuals).

These assumptions help ensure that the linear regression model accurately reflects the relationship between variables and produces reliable predictions.

Question 2: Explain the Anscombe’s quartet in detail.

The Anscombe Quartet consists of four datasets, all sharing identical descriptive statistics like means, variances, R-squared values, correlations, and linear regression lines. However, when these datasets are plotted as scatter plots on a graph, they reveal distinct and diverse patterns.

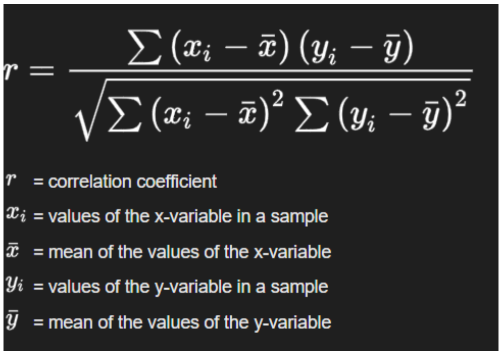
Created by statistician Francis Anscombe in 1973, these datasets were designed to underscore the significance of visualizing data and to caution against relying solely on summary statistics, which can sometimes be misleading.

Each dataset in Anscombe's quartet comprises 11 pairs of x-y data points. When visualized, each dataset exhibits its own unique relationship between x and y, displaying varied patterns of variability and differing strengths of correlation. Despite these visual differences, all four datasets maintain identical summary statistics, including mean and variance for both x and y, correlation coefficient between x and y, and the equation of the linear regression line.

Question 3. What is Pearson’s R?

Answer: Pearson's correlation coefficient is a number that tells us how strong and in what direction two variables are related in a straight-line manner. Its value can range from -1 to 1. A value of 1 means there's a perfect positive relationship: when one variable goes up, the other always goes up. A value of -1 means there's a perfect negative relationship: when one variable goes up, the other always goes down. A value of 0 means there's no straight-line relationship between the variables.

To calculate Pearson's correlation coefficient, we look at how the two variables vary together compared to their individual variations. This helps us understand how closely connected the two variables are in statistics.



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

r = correlation coefficient

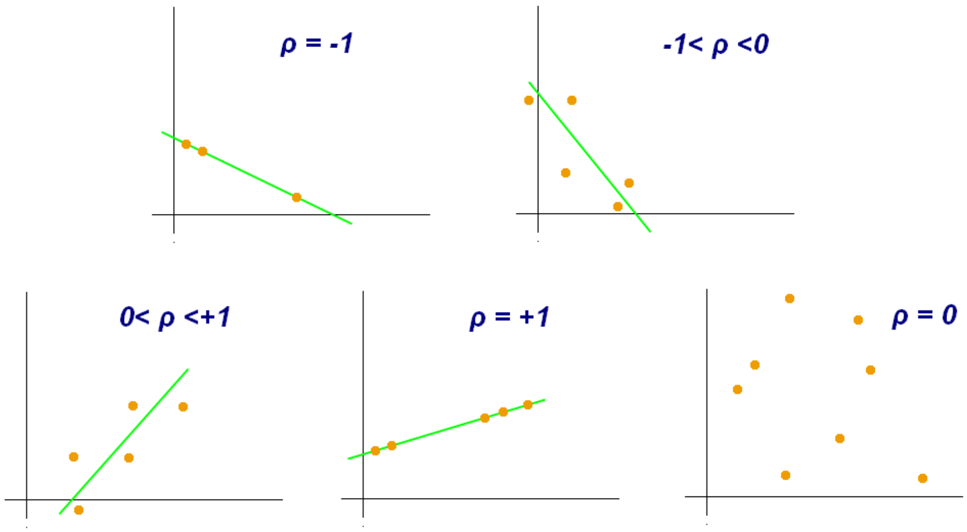
x_i = values of the x-variable in a sample

$\bar{x}$ = 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

As can be seen in the graph, when r = 1 , it means the data forms a perfectly straight line with a positive slope. When r = -1 , the data forms a perfectly straight line with a negative slope. And when r = 0 , it means there's no straight-line relationship between the data points.



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

Answer: Feature Scaling is a crucial data preprocessing technique which is aimed at standardizing independent features within a fixed range. This adjustment is necessary to handle disparities in magnitudes, values, or units among different features. Without feature scaling, machine learning algorithms may disproportionately prioritize larger values and neglect smaller ones, irrespective of their significance.

For instance, consider an algorithm that doesn't utilize feature scaling. It might mistakenly perceive a value like 300 ml as greater than 4 liters, which is incorrect. This disparity can lead to inaccurate predictions. Feature Scaling resolves this issue by normalizing all values to a consistent magnitude.

S.NO.	Normalized scaling	Standardized scaling
1.	Minimum and maximum value of the feature is used for scaling	Mean and standard deviation of feature value is used for scaling.
2.	It is used when features are on 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.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.

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

Answer: Variance Inflammation Factor (VIF) of an independent variable represents how well the variable is explained by other independent variables. In the order words we can say that it VIF determines the strength of the correlation between the independent variables.

Formula to calculate VIF is :

$$\text{VIF} = \frac{1}{1-R^2}$$

Where, R^2 is determined to find out how well an independent variable is described by the other independent variables. It's values ranges from 0 to 1. Where 1 means that the independent variables can perfectly be explained by other independent variable. In this case, VIF becomes infinity which means that there is perfect collinearity between the independent variables.

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

Answer: The quantile-quantile (q-q) plot serves as a visual tool to assess whether two datasets originate from populations with a shared distribution. It plots the quantiles of one dataset against those of another, where a quantile represents the fraction (or percent) of values below a specified point. A 45-degree reference line on the plot indicates where points should ideally align if the datasets share the same distribution. Deviations from this line suggest differences in distribution between the datasets, with greater deviations indicating larger disparities.

The q-q plot is crucial for evaluating the assumption of a common distribution when comparing two datasets. If the datasets share a distribution, combining them allows for more accurate estimation of location and scale parameters. Conversely, if the datasets differ, the q-q plot offers insights into the nature of these differences, surpassing the depth provided by methods such as the chi-square and Kolmogorov-Smirnov tests.