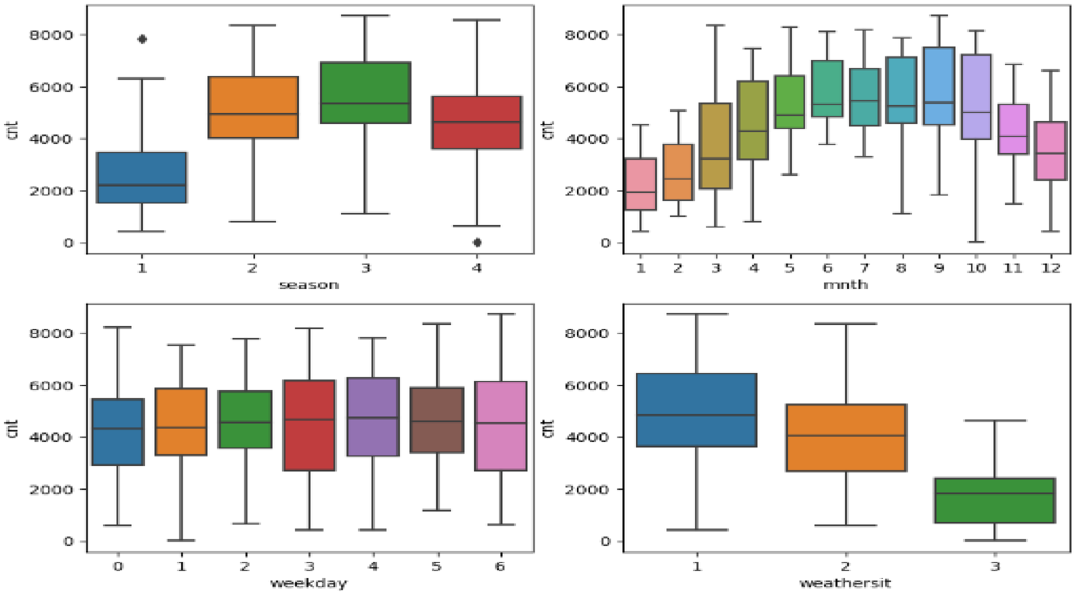
Assignment-based Subjective Questions

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

Ans:



Based on the above box plot:

a. {1: 'spring', 2: 'summer', 3: 'fall', 4: 'winter'}
   Fall season has more bookings as per the above box plot.

b. {1:'Jan', 2:'Feb', 3:'Mar', 4:'Apr', 5:'May', 6:'Jun', 7:'Jul', 8:'Aug', 9:'Sep', 10:'Oct', 11:'Nov', 12:'Dec'}
   Most of the bookings gradually starts rising from the beginning of the year, highest bookings are observed in May, June, July, August, and September and then gradually starts to decline.

c. {1:'Clear', 2:'Mist', 3:'Light Snow', 4:'snow_fog'}
   Clear weather attracts more booking than any other weather.

d. For holiday demand seems to drop, which seems reasonable, people might want to spend some time with family.

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

Ans:

When creating dummy variables from categorical data, drop_first=True is an essential parameter used to prevent multicollinearity in statistical models, particularly in regression analyses.

Using drop_first=True in dummy variable creation helps to enhance the reliability and stability of the statistical model by mitigating multicollinearity issues that might arise due to the interdependence among dummy variables representing categorical data.

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

Ans:

Looking at the pair plot among numerical variables temperature (**temp**) has the highest correlation with the target variable ‘cnt’.

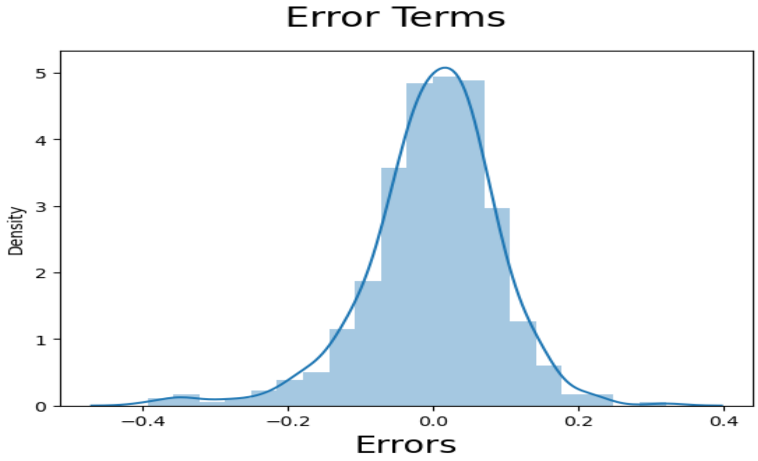
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

Ans:

Validating the assumptions of linear regression after building the model on the training set involves several diagnostic checks to ensure that the model satisfies the underlying assumptions.

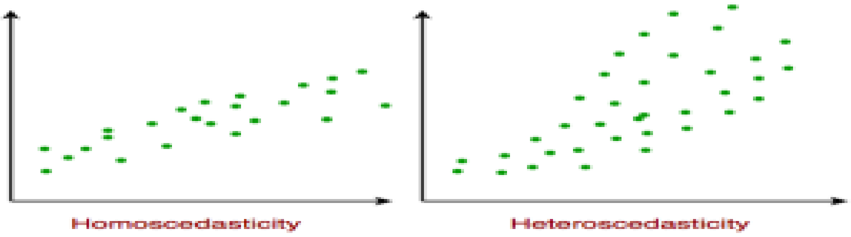
1. Residual Analysis:

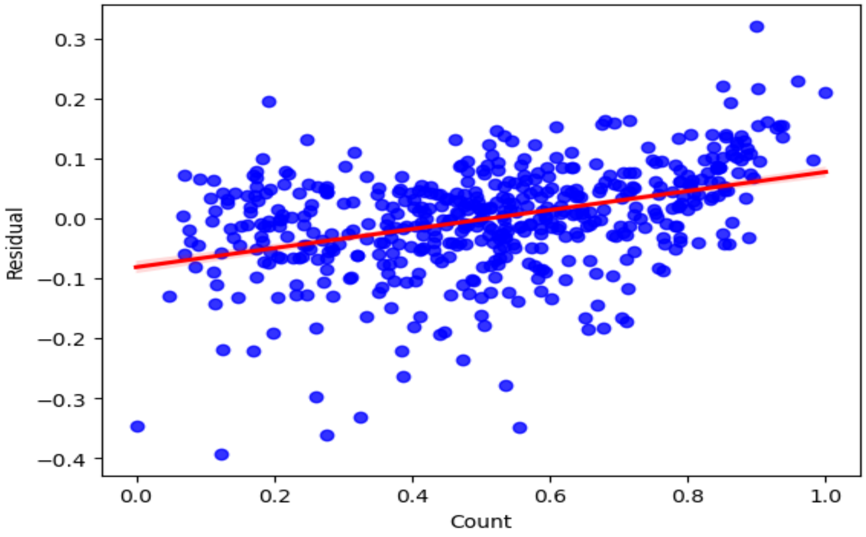
Residual analysis is a fundamental technique used in statistics to assess the goodness of fit of a statistical model to a set of data.



2. Hompscedasticity(Constant Variance):

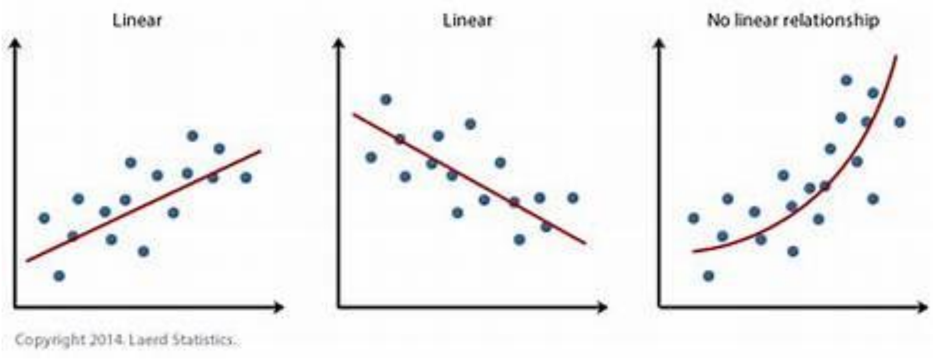
Assumes that the variance of the errors is constant across all levels of the independent variable.





3. **Linearity :**

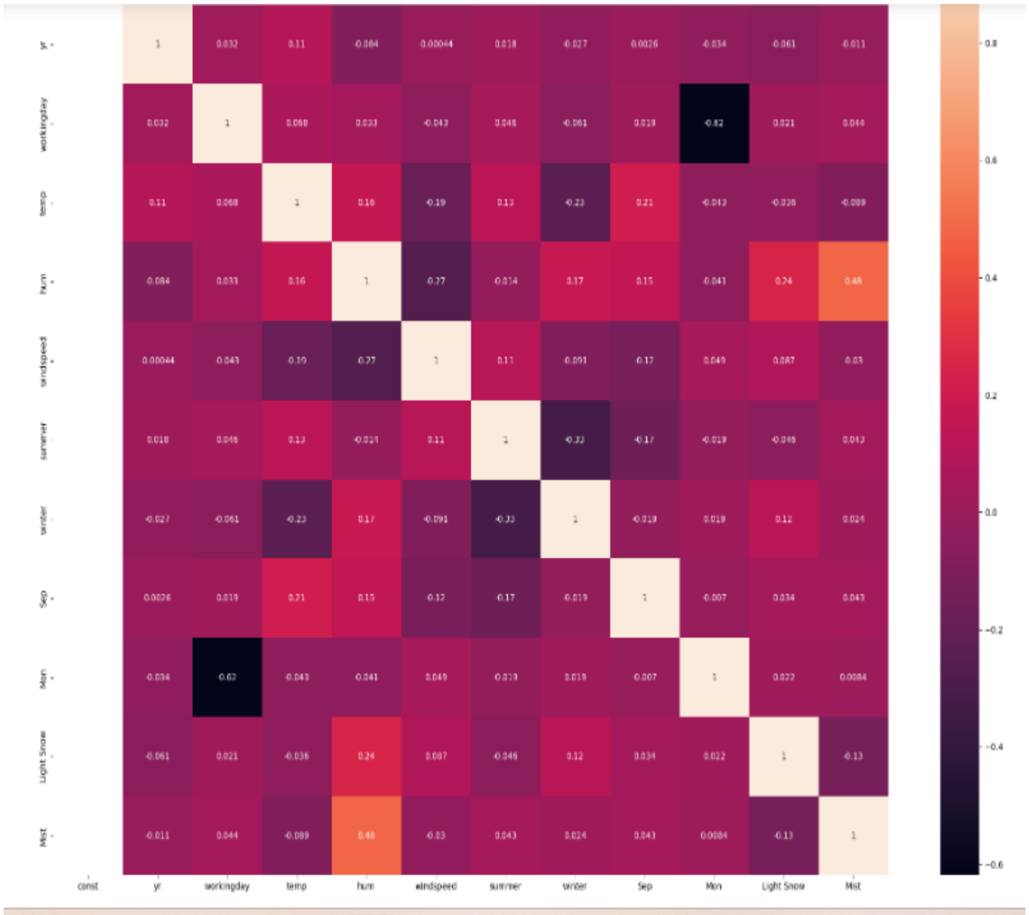
   Plot the residuals against each predictor variable. If there's a pattern or curve in these plots, it suggests a violation of the linearity assumption.



Copyright 2014 Laerd Statistics.

4. **Multicollinearity :**

   Calculating VIF for each predictor variable to detect multicollinearity. VIF values above a certain threshold (usually 5 or10) indicate high multicollinearity.



5. **Cross-Validation :**

Performing cross-validation on the training set to evaluate how well the model generalizes to new data. This helps assess whether the model is overfitting or underfitting the training data.

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

Ans:

Based on the final model the top 3 features contributing significantly towards explaining the demand of the shared bikes are:

1. **Year :**

   The demand for the bike will rise in the upcoming year based on the previous data.

2. **Temperature :**

   Temperature is one of the parameter in which if unit increases demand increase.

3. **Working day :**

   On the working days the demand of the bikes will get increase as the no of working professional's needs to travel.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Ans:

Linear regression is a fundamental statistical method used for modeling the relationship between a dependent variable (target) and one or more independent variables (predictors).

It assumes a linear relationship between the predictor variables and the target variable.

The goal of linear regression is to find the best-fitting linear equation that describes the relationship between the variables.

We are having 2 types of linear regression models:

a. **Simple Linear Regression :**
In simple linear regression, there is one independent variable (predictor) and one dependent variable (target). The linear relationship between them can be represented as:

y = $\beta0+\beta1\cdot x+\epsilon$

y - represents dependent variable.
X - represents independent variable.
$\beta 0$ - intercept (constant)
$\beta 1$ - slope (coefficient)
ϵ - is the error term, representing the difference between the actual and predicted values

b. **Multiple Linear Regression :**
In multiple linear regressions, there are multiple independent variables:

y = β0 + β1.x1 + β2.x2+……+ βn.xn+ ϵ

y - represents dependent variable.
x1, x2…xn- represents independent variable.

$\beta 0$ - intercept (constant)
$\beta 1$ - slope (coefficient)
ϵ - is the error term.

Steps in Linear Regression:

a. Data Collection.

b. Data Preprocessing.

c. EDA.

d. Model Building.

e. Model Evaluation.

f. Prediction.

Assumption in Linear regression:

a. Linearity
b. Independence
c. Hompscedasticity
d. Normality
e. No Multicollinearity

2 . Explain the Anscombe's quartet in detail.

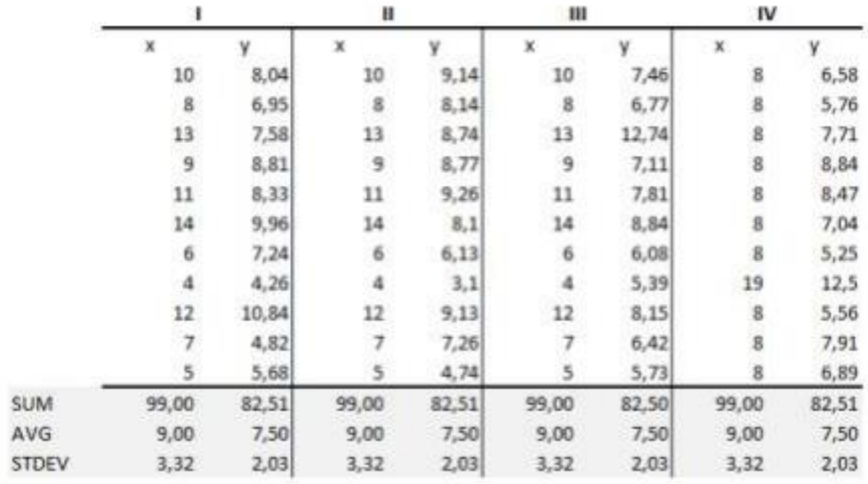
Ans:

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 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. All four sets are identical when examined using simple summary statistics, but vary considerably



	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

As per the above table we can the following details:

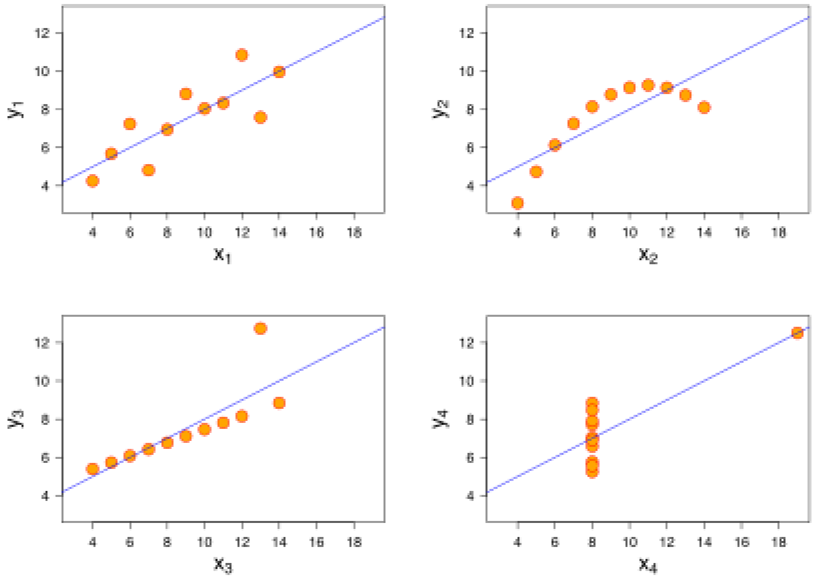
Mean (x) = 9

Mean (y) = 7.50

Standard Deviation (x) = 3.32

Standard Deviation (y) = 2.03

Even if the Statistical data is same for all data set the graph representation is different for all.



3. What is Pearson’s R?

Ans:

Pearson’s R is a statistical measure that assesses the linear relationship between two continuous variables.

The value of Pearson's r ranges from -1 to 1.

1 indicates a perfect positive linear relationship: as one variable increases, the other variable also increases proportionally.

0 indicates no linear relationship: the variables are not correlated.

-1 indicates a perfect negative linear relationship: as one variable increases, the other variable decreases proportionally.

The formula for Pearson's correlation coefficient (r) is:

$$r = \frac{n(\sum xy)-(\sum x)(\sum y)}{\sqrt{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2]}}$$

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

Ans:

Scaling is a preprocessing technique used in machine learning to normalize the range of independent variables or features in a dataset. It transforms the data so that all features have a similar scale or range.

As you saw in the demonstration for Simple Linear Regression, scaling doesn't impact your model. Here we can see that except for area, all the columns have small integer values. So it is extremely important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. This might become very annoying at the time of model evaluation. So it is advised to use standardization or normalization so that the units of the coefficients obtained are all on the same scale. As you know, there are two common ways of rescaling:

- **Min-Max scaling (normalization)** between 0 and 1.

  Normalization typically scales the data within a specific range, often between 0 and 1.

  Formula:

$$X_{\text{normalized}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

- **Standardization** (mean-0, sigma-1).

Standardization transforms the data to have a mean of 0 and a standard deviation of 1.

Formula:

$$X_{\text{standardized}} = \frac{X - \text{mean}(X)}{\text{std}(X)}$$

Key differences:

Normalization scales data between 0 and 1, while standardization centers data around 0 with a standard deviation of 1.

Normalization compresses the data within a specific range, while standardization preserves the shape of the distribution but changes its scale.

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

The Variance Inflation Factor (VIF) is a measure used in regression analysis to assess the extent of multicollinearity among predictor variables.

Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, which can lead to issues in estimating the individual contribution of each variable to the dependent variable.

The formula for VIF is:
ViF = 1 / 1 – Ri^2

If the Ri^2 value is very close to 1, it means that the ith variable can be almost perfectly predicted by the other variables in the model, indicating high multicollinearity.

In such cases, the denominator in the VIF formula becomes very close to zero, and therefore, the VIF can become very large or, in theory, infinite.

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

Ans:

A QQ plot is created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

Use of Q-Q plot:

A Q-Q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value.

That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above value. A 45-degree reference line is also plotted.

If the two sets come from a population with the same distribution, the points should fall approximately along this reference line.

The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance:

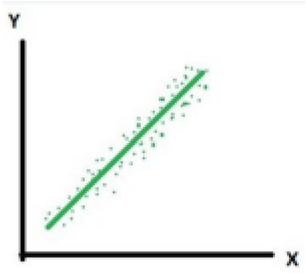
The Quantile-Quantile (Q-Q) plot holds significant importance in various aspects of statistical analysis due to its ability to visually assess the distribution of a dataset and compare it with a theoretical distribution.

The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.

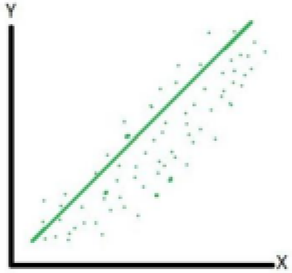
Interpretation:

1. If the points on the Q-Q plot form a straight line, it suggests that the data closely follow the assumed distribution.

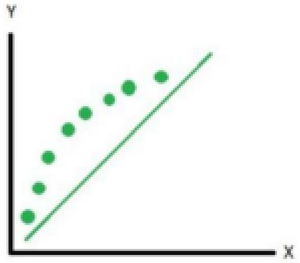
   In the case of linear regression, if the residuals' Q-Q plot forms a straight line, it indicates that the assumption of normally distributed residuals is met.



2. If the y – quantiles are lower than the x – quantiles. It indicates y values have a tendency to be lower than x values.



3. If the x – quantiles are lower than the y – quantiles. It indicates x values have a tendency to be lower than y values.



4. If the y – quantiles are lower than the x – quantiles and after that point the y – quantiles are higher than the x – quantiles.