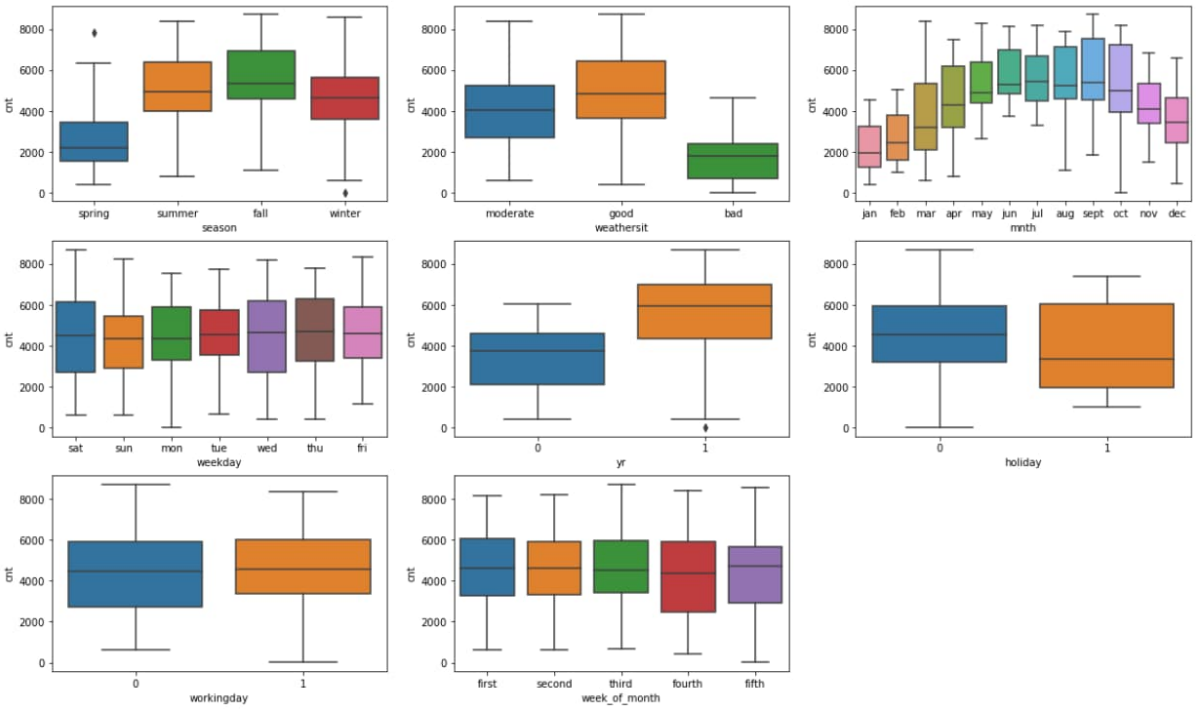
Assignment-based Subjective Questions

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

Answer:

I have done analysis on categorical columns using the boxplot and pair plot. Below are the few points we can infer from the visualization –

- There is a high demand in both years for the Fall season and very low demand in the Spring season.
- The demand for bikes gets reduced in bad weather and good weather attracts more demand.
- The demand for bikes gets increasing at the start of the year and is highest for the months of May, June, July, August, September, and October. And it starts decreasing at the end of the year.
- There is a very large number of bookings in 2019 as compared to 2018, which shows that this market is growing year by year and the company can make good progress out of it.
- When there is a holiday, demand for bikes seems to be low.
- The demand seemed to be almost equal either on the working day or non-working day.
- The demand seemed to be almost equal throughout the week of months.



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

Answer:

It helps in reducing the extra column created during dummy variable creation and it reduces the correlations created among dummy variables.

Syntax -

drop_first: bool, default False, which implies whether to get n-1 dummies out of n categorical levels by removing the first level.

As in the below example:

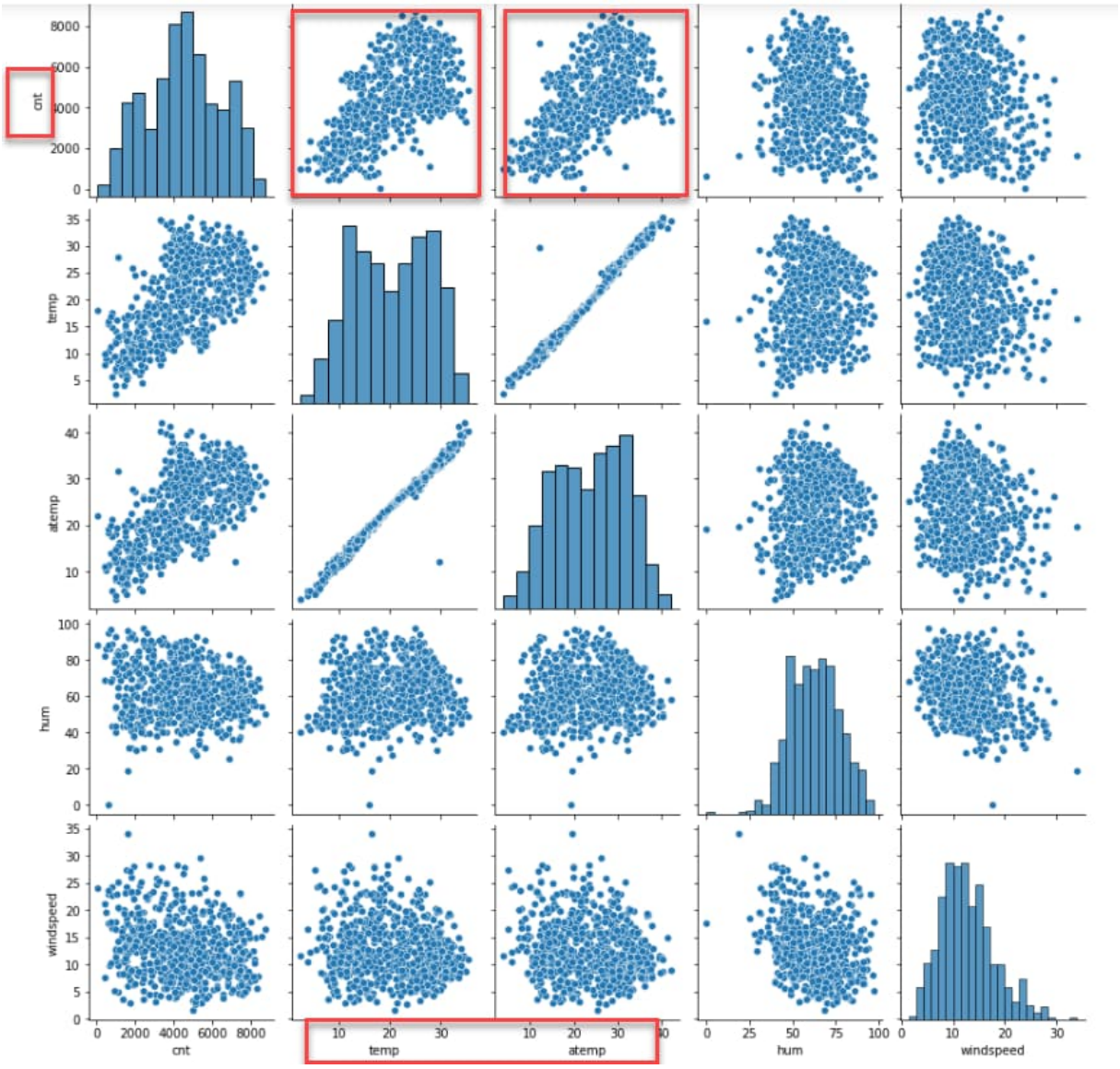
Season	Season_Fall	Season_Spring	Season_Summer	Season_Winter
Fall	1	0	0	0
Spring	0	1	0	0
Summer	0	0	1	0
Winter	0	0	0	1

When we drop Season_Fall then it can be calculated if all the other three categories like Spring, Summer, and Winter are zero, then it states that it's Season Fall. Hence there is no need to create an extra column and increase the correlation among them.

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

Answer:

Temp and atemp has the highest correlation with the target variable.

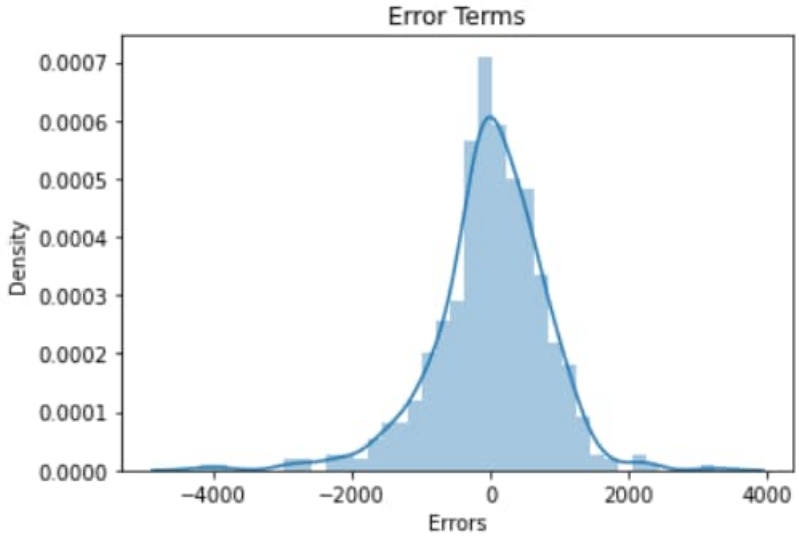


Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer:

I have validated the assumption of the Linear Regression Model based on below 5 assumptions -

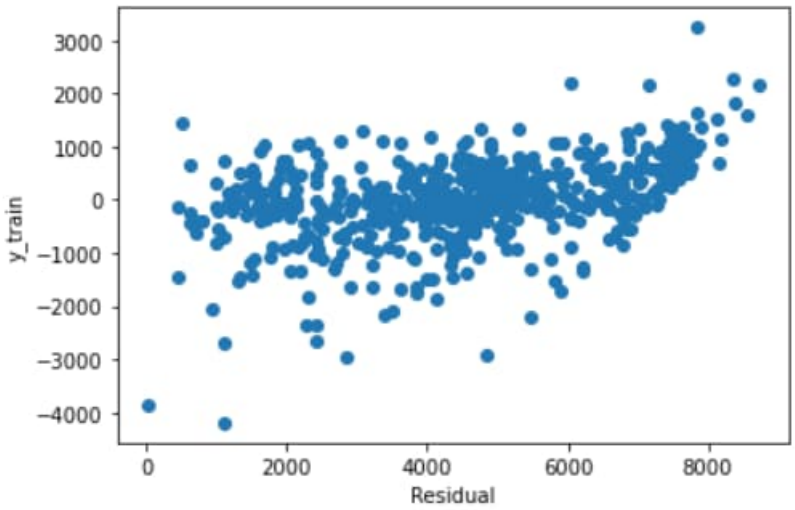
1. Normality of error terms
   Error terms are normally distributed



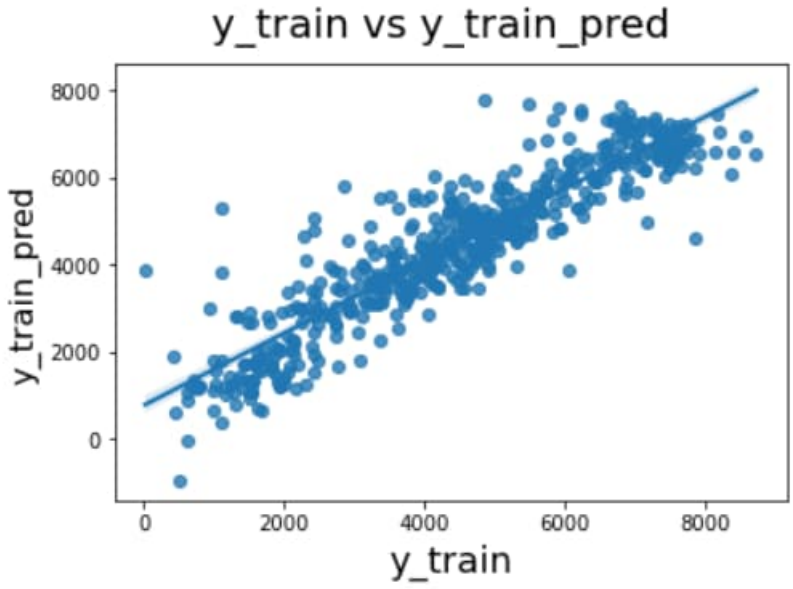
2. Multicollinearity check
   There is insignificant multicollinearity among variables as VIF is low for all selected features.

```
Notes:
[1] Standard Errors assume that
                Features   VIF
10             windspeed  3.99
1                  atemp  3.68
3          season_winter  2.50
0                     yr  1.96
5              mnth_nov  1.79
2          season_spring  1.72
8   weathersit_moderate  1.52
4              mnth_dec  1.45
6            weekday_sun  1.17
7         weathersit_bad  1.09
9                holiday  1.06
```

3. Linear relationship validation
   Linearity should be visible among variables

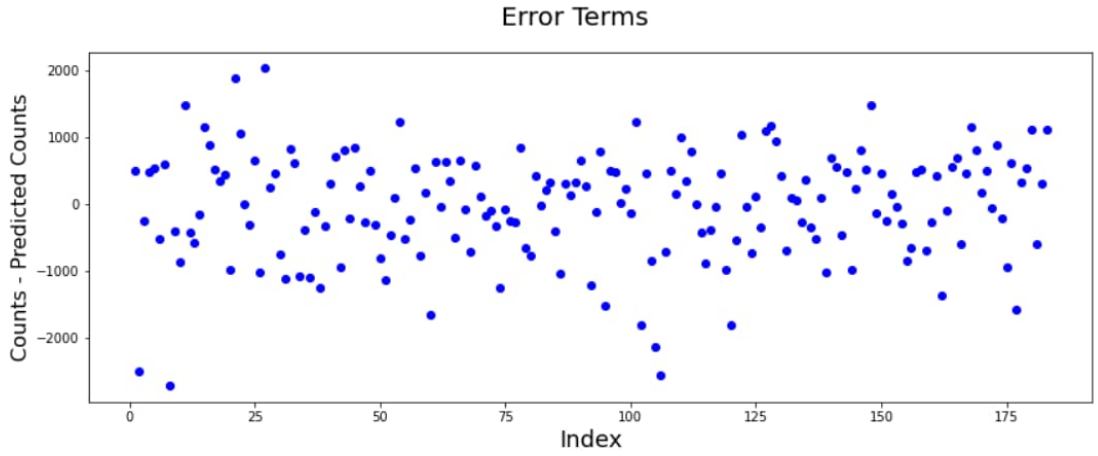


4. Homoscedasticity
   There should be no visible pattern in residual values.



;

5. Independence of residuals
   No autocorrelation

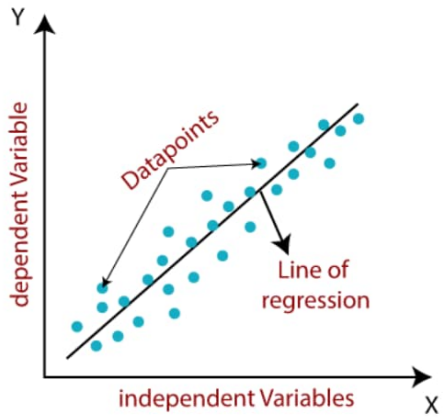


Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
Answer:
Below are the top 3 features contributing significantly towards explaining the demand for the shared bikes:

- atemp: Feeling temprature
- Season
- Weather

General Subjective Questions

Question 1. Explain the linear regression algorithm in detail. (4 marks)
Answer:
Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables.
Linear regression algorithm shows a linear relationship between a dependent and one or more independent variables. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.
The linear regression model provides a sloped straight line representing the relationship between the variables.



Equation to represent a linear regression is:

$$y= a_0+a_1x+ \varepsilon$$

- Y= Dependent Variable (Target Variable)
- X= Independent Variable (predictor Variable)
- a0= intercept of the line (Gives an additional degree of freedom)
- a1 = Linear regression coefficient (scale factor to each input value).
- ε = random error

Types of Linear Regression:
Linear regression can be further divided into two types of the algorithm:

Simple Linear Regression:

If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

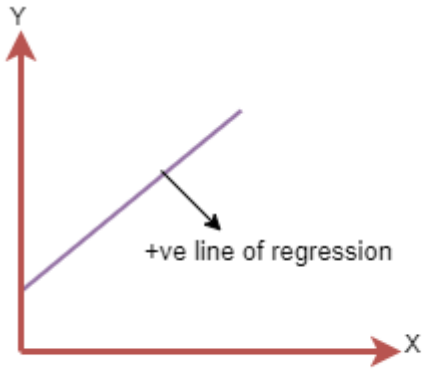
Multiple Linear regression:

If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Relationship in Linear Regression:

Positive Linear Relationship:

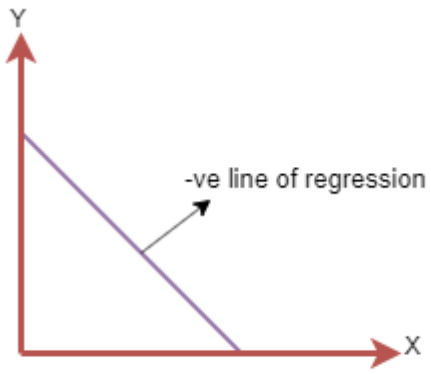
If the dependent variable increases on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship.



The line equation will be: $\mathbf{Y= a_0+a_1x}$

Negative Linear Relationship:

If the dependent variable decreases on the Y-axis and independent variable increases on the X-axis, then such a relationship is called a negative linear relationship.



The line of equation will be: $\mathbf{Y= -a_0+a_1x}$

Assumptions -
The following are some assumptions about the dataset that is made by the Linear Regression model –

1. Multi-collinearity –
   - Linear regression model assumes that it is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have a dependency in them.
2. Auto-correlation –
   - Another assumption the Linear regression model assumes is that it is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is a dependency between residual errors.
3. Relationship between variables –
   - Linear regression model assumes that the relationship between response and feature variables must be linear.

4. Normality of error terms –
   - Error terms should be normally distributed
5. Homoscedasticity –
   - There should be no visible pattern in residual values.

When working with linear regression, our main goal is to find the best-fit line which means the error between predicted values and actual values should be minimized. The best-fit line will have the least error.

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

Answer:

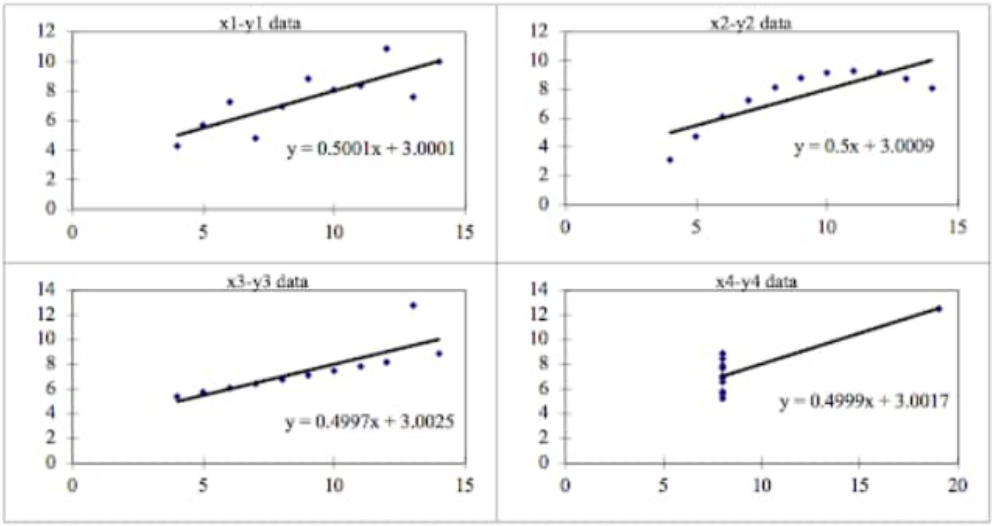
Anscombe's quartet was developed by Francis Anscombe to demonstrate both the importance of graphing data when analyzing it and the effect of outliers and other influential observations on statistical properties.

Anscombe's quartet is a group of four data sets, each containing eleven (x, y) pairs that are nearly identical in simple descriptive statistics, but there are peculiarities that fool the regression model once you plot each data set. But while visualizing them, Each graph tells a different story irrespective of their similar summary statistics.

Anscombe's Data								
Observation	x1	y1	x2	y2	x3	y3	x4	y4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89
			Summary Statistics					
N	11	11	11	11	11	11	11	11
mean	9.00	7.50	9.00	7.500909	9.00	7.50	9.00	7.50
SD	3.16	1.94	3.16	1.94	3.16	1.94	3.16	1.94
r	0.82		0.82		0.82		0.82	

The summary statistics show that the means and the variances were identical for x and y across the groups.

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:



Interpretation from visualization:

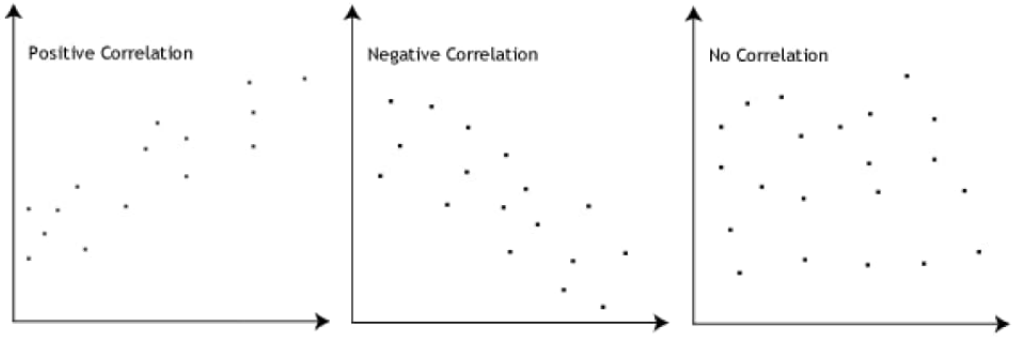
- X1-y1 data: fit the linear regression model pretty well.
- X2-y2 data: cannot fit the linear regression model because the data is non-linear.
- X3-y3 data: shows the outliers involved in the data set, which cannot be handled by the linear regression model.
- X4-y4 data: shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

Anscombe's quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

Question 3. What is Pearson's R? (3 marks)
Answer:

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition to the low values of one variable associated with the high values of the other, the correlation coefficient will be negative.

The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.



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

Answer:

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing.

Feature scaling is performed to handle highly varying values. If feature scaling is not done, then a machine learning model tends to consider greater values as higher and consider smaller values as lower values, regardless of the unit of the values, which is not true.

For example:

If an algorithm is not using the feature scaling method then it can consider the value 2000 grams to be greater than 3 kg but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to the same scale and tackle this issue.

Difference between Normalised Scaling and Standardized scaling:

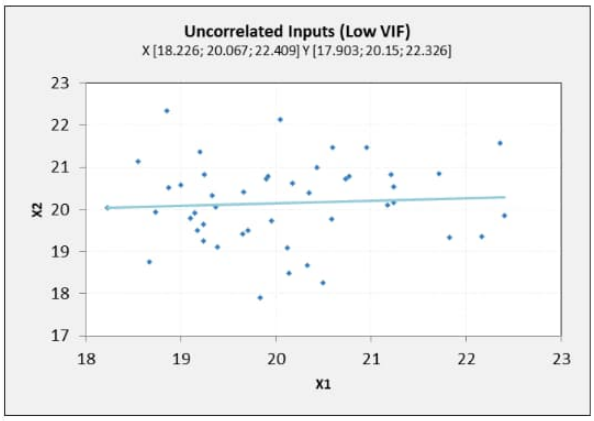
S.NO.	Normalization	Standardization
1	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3	Scales values between [0, 1] 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	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

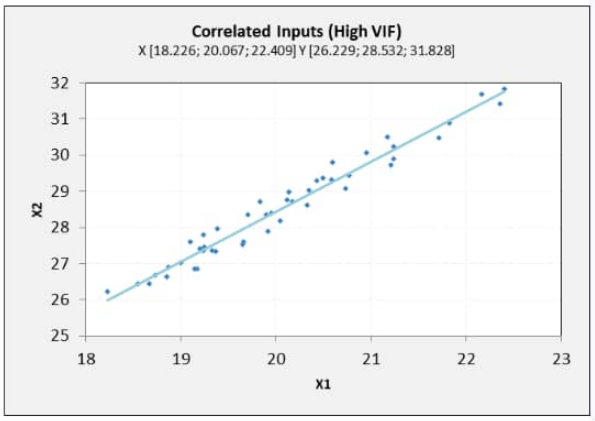
6	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

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

If there is a perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 7, this means that the variance of the model coefficient is inflated by a factor of 7 due to the presence of multicollinearity.

When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which leads to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.





Its impact on analysis:

- Model coefficients could not be computed due to singularity
- Model coefficients may be inaccurate
- Model coefficients may vary wildly when you add/subtract data
- Model coefficients may vary wildly when you add/drop terms
- Model coefficients may become statistically insignificant

Question 6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer:

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

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 falls below and 70% fall above that 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 of Q-Q plot:
When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. 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.