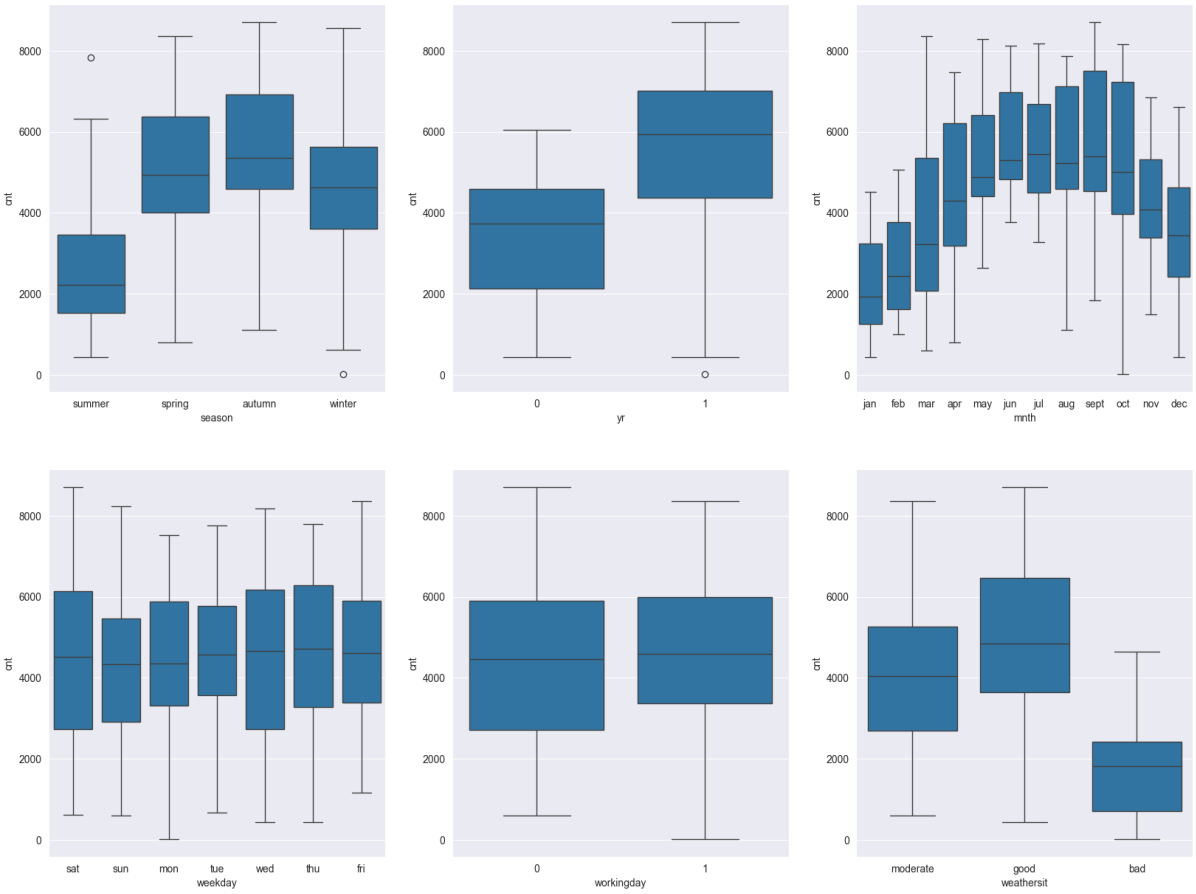
Assignment Questions and Answers

Assignment Questions

---

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

- The assignment contains the following categorical variables: "season," "workingday," "weathersit," "weekday," "yr," "holiday," and "mnth."
- “season” –
  - ❖ Summer and fall are the best seasons for biking, according to the data that is currently available.
  - ❖ Strategic advertising allows for the planning of higher targets for the summer and fall.
  - ❖ The springtime consumption ratio is noticeably low.
- “workingday” –
  - ❖ This refers to information about weekdays, weekends, and holidays.
  - ❖ While casual users prefer to hire bikes on non-working days, registered users rent bikes on workdays. When we consider the overall count, this effect is eliminated due to the inconsistent behavior of both registered and non-registered users.
- “weathersit” –
  - ❖ Clear or somewhat cloudy days are the most favorable weather conditions.
  - ❖ The number of registered users is relatively high, even on days with mild rain, suggesting that the bikes are being used for everyday commuting to work.
- “weekday” –
  - ❖ There is no discernible trend with the weekday when looking at the "cnt" column.
  - ❖ But when the relationship is plotted using "registered" users, we see that working days see a greater bike utilization rate. And it's the opposite with "casual" users.
- “yr” –
  - ❖ Two years of data are available, and during 2018 and 2019, the number of bikes grew.
- “holiday” –
  - ❖ The amount of motorcycles consumed on holidays when "casual" and "registered" users are contrasted, it is shown that the former ride bikes more frequently while on vacation.
- “mnth” –
  - ❖ The months of June, July, August, September, and October have a higher bike rental ratio. During the months indicated in point 1, the 75 quantile grows.



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

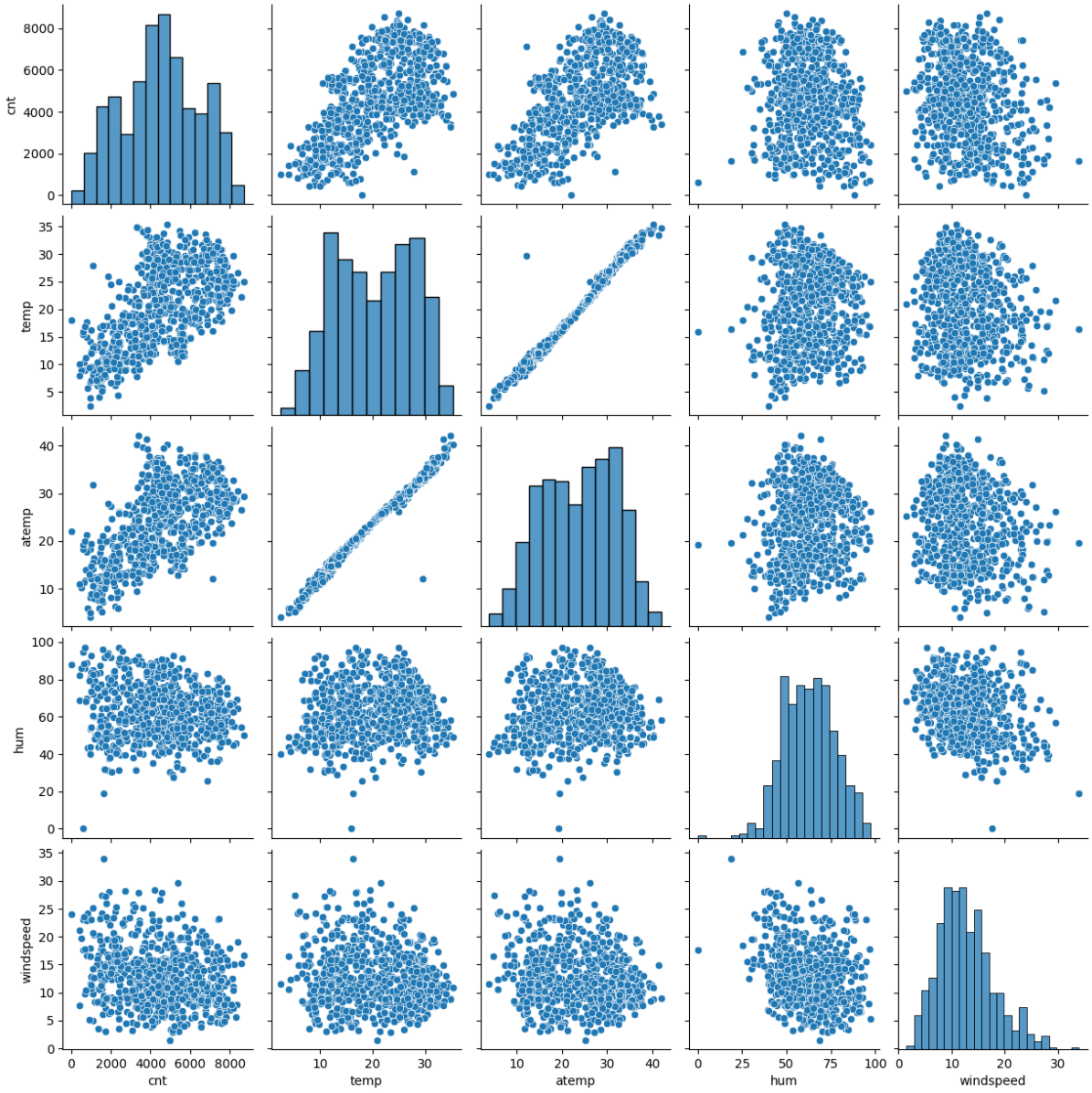
The dummy variables are made to span the range of values of the categorical variable using one-hot encoding. The values of each dummy variable are 1 and 0. The category's presence is represented by a 1 and its absence by a 0. This indicates that there will be three dummy variables if the category variable contains three categories.

When constructing dummy variables, drop the base/reference category by using drop_first = True. This is done in order to prevent the addition of multi-collinearity to the model that would result from include all dummy variables. When all other dummy variables in a given category have 0 in a single row, it is easy to determine which category is the reference..

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

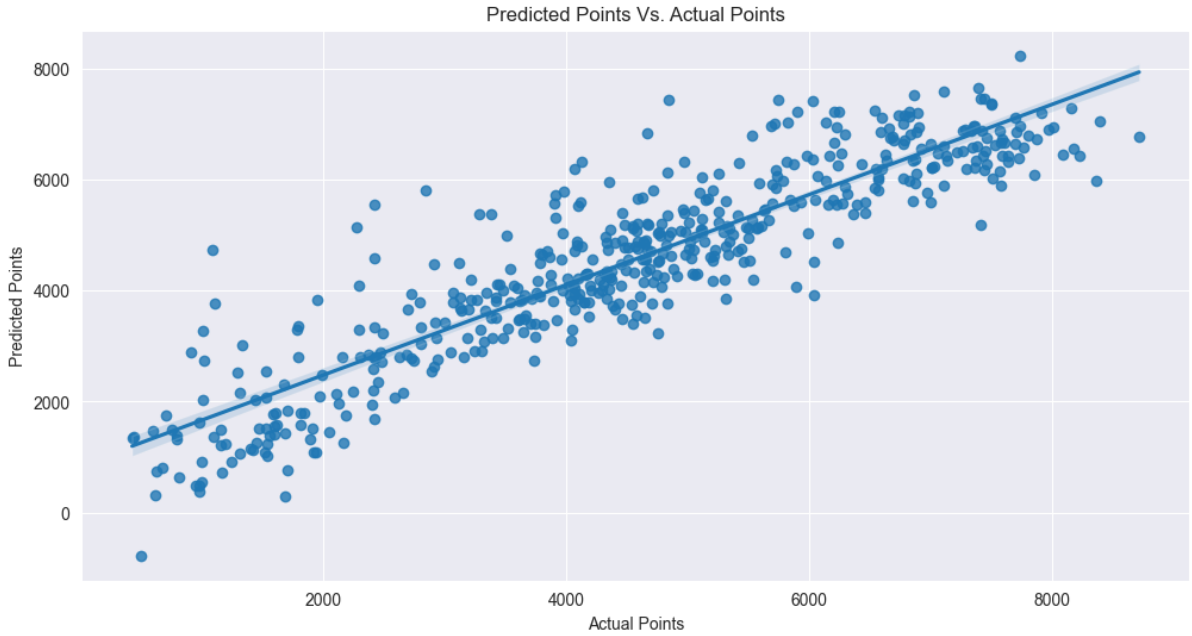
- The variable "temp" has the highest correlation (0.63) with the target variable.
- Ignoring the association between these two variables, the casual and registered variables are essentially a portion of the goal variable since the values of these columns add up to the target variable.

- Since "atemp" is derived from temperature, humidity, and wind speed and is removed during model development, it is not taken into consideration.

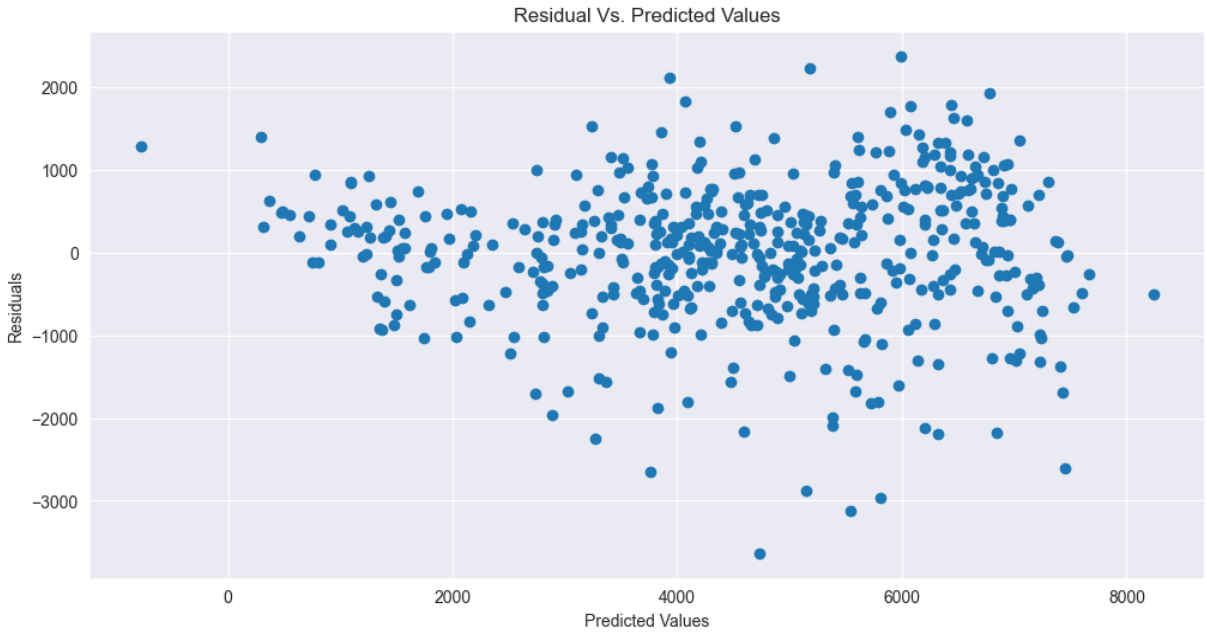


4. How did you validate the assumptions of Linear Regression after building the model on the training set?

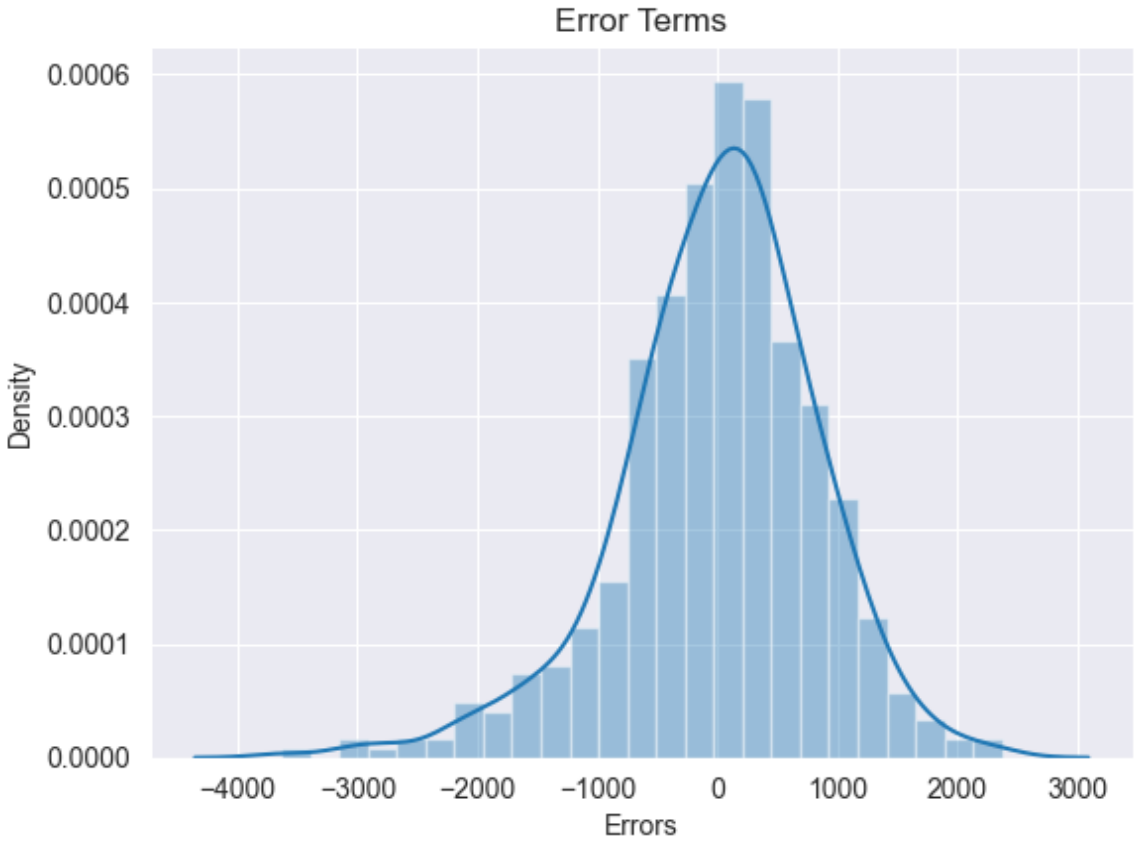
1. Linear relationship between independent and dependent variables: As seen in the next graphic, the linearity is verified by examining the points that are symmetrically distributed around the diagonal line of the actual vs. predicted plot.



2. Error terms are independent of each other – Since there is no discernible pattern in the error terms concerning prediction, we can conclude that the error terms are unrelated to one another.

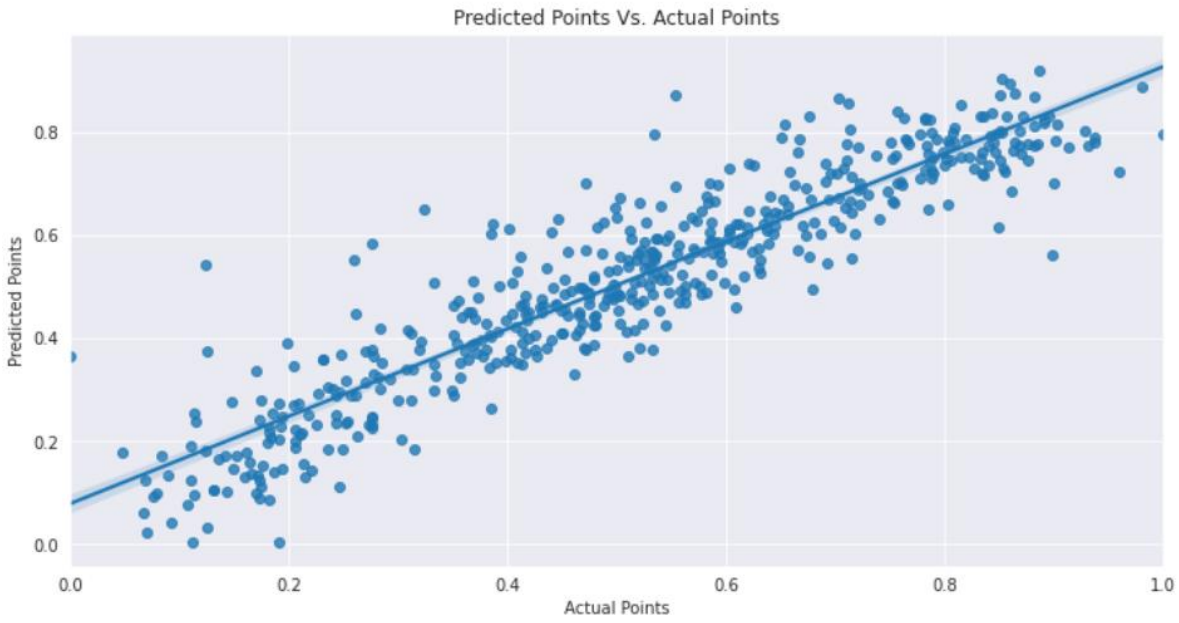


3. Error terms are normally distributed: . Understanding the normal distribution of error terms and the mean of 0 is aided by the histogram and distribution plot. The same is amply illustrated in the illustration below.



4. **Error terms have constant variance (homoscedasticity)**:

As we can see, error terms roughly follow a constant variance, which is consistent with the hypothesis of homoscedasticity.



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

The first three factors are::

- **weathersit** :
  The single most important factor that has a beneficial impact on business is temperature. Other environmental factors, such as humidity, wind speed, rain, and cloud cover, have a negative impact on business.
- **'Yr':**
  Considering the geological characteristics, the growth appears organic year after year.
- **'season':**
  The need for shared bikes is mostly influenced by the winter months.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

One type of predictive modelling technique that shows us the link between the independent variables (predictors) and the dependent variable (goal variable) is called linear regression. Given that linear regression illustrates a linear relationship, it determines how the dependent variable's value changes in response to the independent variable's value. Such linear regression is referred to as simple linear regression if there is only one input variable (x). Additionally, this type of linear regression is known as multiple linear regression if there are many input variables. The relationship between the variables is described as a slanted straight line by the linear regression model.

Either a positive or negative linear relationship can be represented by a regression line.

Finding the optimal values for a0 and a1 is the aim of the linear regression algorithm, which seeks to identify the best fit line with the least amount of error.

The Mean Squared Error (MSE), cost function, or RFE are used in linear regression to determine the optimal values for a0 and a1, which yield the best fit line for the data points.

- The process of determining the optimal linear relationship between the independent and dependent variables is known as linear regression.The algorithm maps the relationship between independent variables and dependent variable using the best fitting line.
- There are 2 types of linear regression algorithms
  - Simple Linear Regression
    - $Y = \beta_0 + \beta_1 X.$
  - Multiple Linear Regression
    - $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \in$
  - $\beta_0 = value\ of\ the\ Y\ when\ X = 0\ (Y\ intercept)$
  - $\beta_1, \beta_2, \ldots, \beta_p = Slope\ or\ the\ gradient.$

- Cost functions: These aid in determining the optimal values for β0,β0,β0,…, and βp, which in turn aid in predicting the target variable's probability. To find the best fitting line for predicting the dependent variable, the minimization strategy is utilized to lower the cost functions. There are two different techniques to cost function minimization. – **Unconstrained and constrained**.
- We discover that there are mistakes in the mapping of the real numbers to the line when searching for the best fit line. The residuals are all that these errors are. OLS (Ordinary least square) is used to reduce the error squares..
    - $e_i = y_i - y_{pred}$ - > error for each of the data point.
    - OLS is used to minimize the total e^2 which is called as Residual sum of squares.
    - $\text{RSS} = = \sum_{i=1}^{n} (y_i - y_{pred})^2$
- To estimate beta coefficients and minimize the residual sum of squares, utilize the Ordinary Lease Squares approach.

---

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet is a collection of four data sets that, when analyzed using basic descriptive statistics, are almost identical to one another; yet, the datasets have some quirks that, if a regression model is constructed, would trick the model. When shown on scatter plots, they display differently and have rather different distributions. It was designed to highlight the significance of plotting the graphs prior to analysis and model construction, as well as the impact of additional observations on statistical features.The statistical observations in all four data set plots are almost identical, providing identical statistical information on variance and mean of all x, y points throughout the datasets..

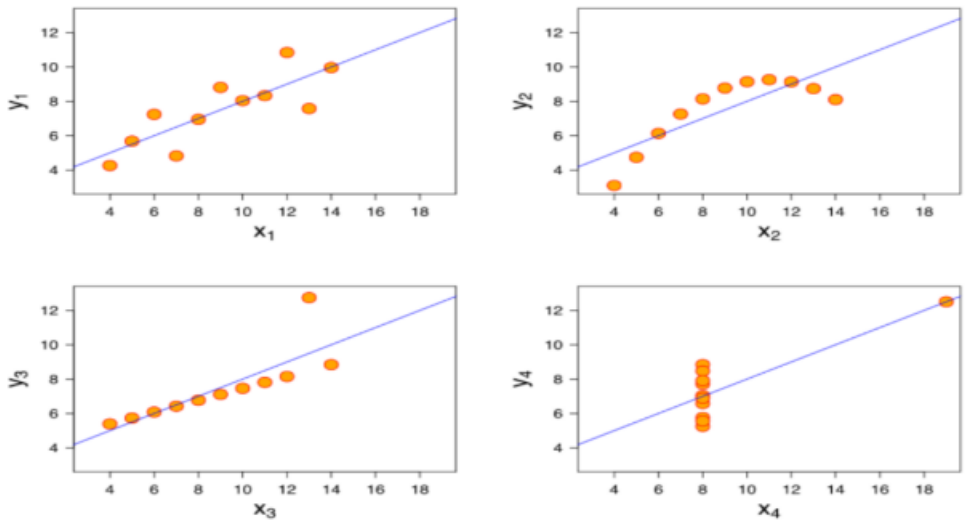
- Illustrations
    - One of the data set is as follows:

	x1	x2	x3	x4	y1	y2	y3	y4
0	10	10	10	8	8.040000	9.140000	7.460000	6.580000
1	8	8	8	8	6.950000	8.140000	6.770000	5.760000
2	13	13	13	8	7.580000	8.740000	12.740000	7.710000
3	9	9	9	8	8.810000	8.770000	7.110000	8.840000
4	11	11	11	8	8.330000	9.260000	7.810000	8.470000
5	14	14	14	8	9.960000	8.100000	8.840000	7.040000
6	6	6	6	8	7.240000	6.130000	6.080000	5.250000
7	4	4	4	19	4.260000	3.100000	5.390000	12.500000
8	12	12	12	8	10.840000	9.130000	8.150000	5.560000
9	7	7	7	8	4.820000	7.260000	6.420000	7.910000
10	5	5	5	8	5.680000	4.740000	5.730000	6.890000

    - When the a rementioned data set's descriptive statistics are examined, everything appears to be the same:

	x1	x2	x3	x4	y1	y2	y3	y4
count	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000
mean	9.000000	9.000000	9.000000	9.000000	7.500909	7.500909	7.500000	7.500909
std	3.316625	3.316625	3.316625	3.316625	2.031568	2.031657	2.030424	2.030579
min	4.000000	4.000000	4.000000	8.000000	4.260000	3.100000	5.390000	5.250000
25%	6.500000	6.500000	6.500000	8.000000	6.315000	6.695000	6.250000	6.170000
50%	9.000000	9.000000	9.000000	8.000000	7.580000	8.140000	7.110000	7.040000
75%	11.500000	11.500000	11.500000	8.000000	8.570000	8.950000	7.980000	8.190000
max	14.000000	14.000000	14.000000	19.000000	10.840000	9.260000	12.740000	12.500000

- But when these points are plotted, the relationship appears quite different, as seen below.



- Anscombe's Quartet denotes the possibility that, when plotted, several data sets with a great deal of statistical similarity could nevertheless differ from one another.
- The foursome warns of the risks associated with outliers in data sets. View the two graphs at the bottom. The descriptive statistics would have been entirely different in that scenario if those outliers had not existed.
- Vital information.
  - ❖ Before data analysis, plotting the data is a good idea and a crucial step.
  - ❖ Eliminating outliers is important while analyzing the data.
- The data collection as a whole is not completely represented by descriptive statistics.

---

3. What is Pearson's R?

The Pearson product-moment correlation coefficient (PPMCC), bivariate correlation, Pearson's r, and Pearson's correlation coefficient are other names for the Pearson's correlation coefficient in statistics. It is a statistical measure of how well two variables are correlated linearly.:

- *-1 coefficient -> strong inversely proportional relationship.*
- *0 coefficient -> no relationship.*
- *1 coefficient -> strong proportional relationship.*

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

---

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

Scaling is the process of changing data to make it fit a predetermined range. This kind of data pre-processing step involves fitting data to a specified scale in order to expedite algorithmic computations. Features with different ranges, magnitudes, and units are present in the collected data. Inaccurate modeling will occur from the algorithm's tendency to weigh large values and overlook other factors if scaling is not done.

1. Normalizing scaling differs from standardizing scaling. The minimum and maximum values of the features are utilized in normalized scaling, whereas the mean and standard deviation are used in standardize scaling.
2. Standardized scaling is employed to guarantee zero mean and unit standard deviation, while normalized scaling is utilized when features are of different scales.
3. In contrast to standardized scaling, which lacks or is not limited in a certain range, normalized scaling scales values between (0,1) and (-1,1).
4. Outliers have an impact on normalized scaling, but they have no influence on standardized scaling.
5. When we don't know anything about the distribution, we use normalized scaling; when the distribution is normal, we use standardized scaling.
6. Standardized scaling is referred to as Z Score Normalization, while normalized scaling is called scaling normalization.

$$MinMaxScaling: x = \frac{x - \min(x)}{\max(x) - \min(x)}$$

$$Standardization: x = \frac{x - mean(x)}{sd(x)}$$

---

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

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

When the VIF will be limitless is indicated plainly by the VIF formula. The VIF is infinite if the R^2 is 1. The perfect correlation between the two independent variables accounts for R^2 value of 1.

---

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

Plotting the quantiles of two probability distributions against one another graphically allows for the comparison of the results using a Q–Q plot, also known as a probability plot.
A graphical tool called a quantile-quantile (Q-Q) plot can be used to determine if a set of data is likely to have originated from a normal, exponential, or uniform distribution.
The QQ plot can also be used to assess the similarity or dissimilarity of two distributions.
You can anticipate a more linear QQ plot if they are somewhat comparable. Scatter plots are the most useful tool for testing the linearity assumption. Second, all variables must be multivariate normal in order to perform a linear regression analysis. The best way to verify this assumption is with a Q-Q-Plot or a histogram.
Q-Q plots are important for Linear Regression because they allow us to verify whether or not the train and test datasets are from the same population. This is particularly useful when we have two datasets for the analysis.
Benefits:

- It can be used to larger sample sizes as well;
- This plot allows for the detection of numerous distributional features, such as changes in scale, position, symmetry, and outlier existence.

- Using a Q-Q plot on two datasets, determine whether they both originated from populations with similar distributions.
- If the location and scale of both datasets are same
- If the distribution shapes of the two datasets are identical, or if tail behavior is present in both datasets