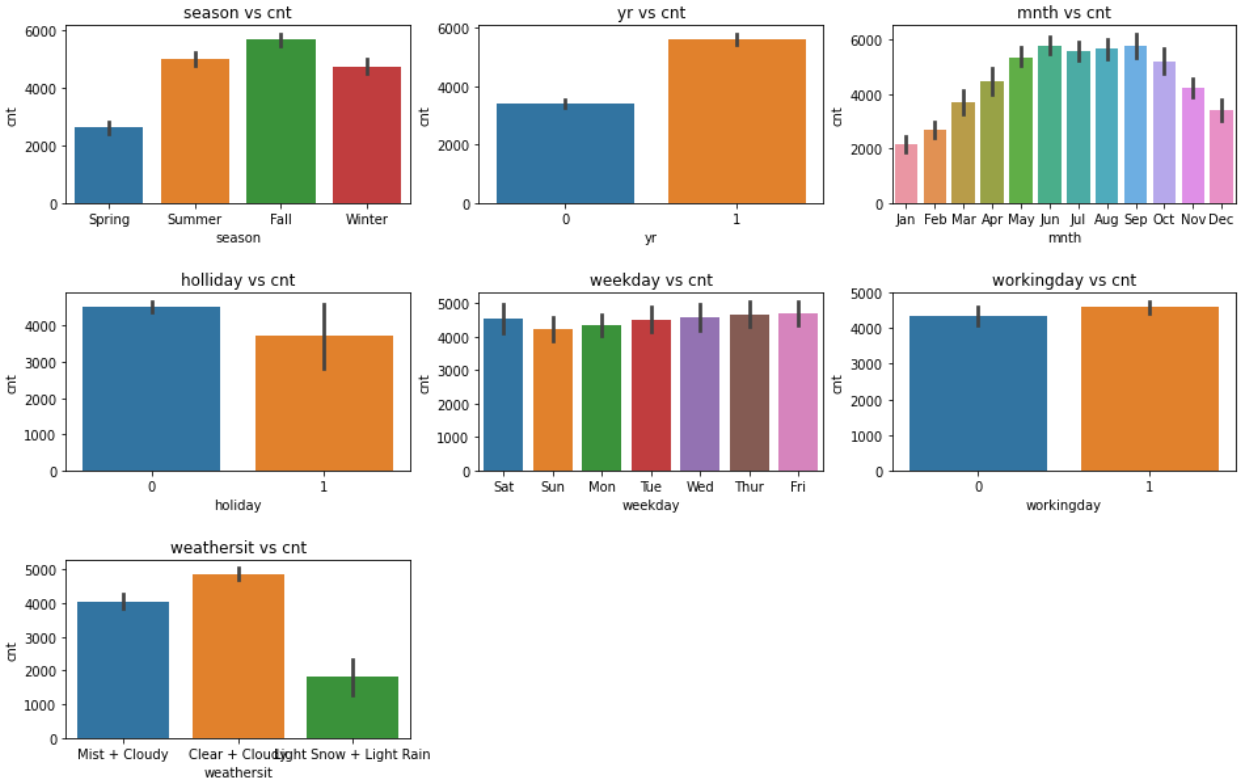
Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

After dropping unwanted variables (['instant','dteday','atemp','casual','registered']) there are total 7 categorical variables ([season','yr','mnth','holiday','weekday','workingday','weathersit]) in the dataset. After plotting each of the independent variables against the dependent variable (‘cnt’) below are the inferences of each.



1. Riders prefer fall seasons to go for riding and book least bikes in spring season.

2. In 2019, bikers booked more than 2018.

3. Bike booking increased from jan to jun then remained almost the same till sep and then started decreasing till dec. June and sept has the highest booking.

4. Non holiday days have more bike booking than holidays.

5. Almost all days have the same booking but sat has the highest booking.

6. Both working and non working days have same booking

7. Riders prefer clear and cloudy weather than snowy and rainy whether

Q2. Why is it important to use drop_first=True during dummy variable creation?

In a linear regression model, all variables should be numeric to make the model work. But many times we have categorical variables in the data set. Dummy coding provides one way through which we can use categorical variables in linear model estimation.
For each level in the categorical variable we create one column for each. And while encoding we put 1 where level value is present and 0 for all other levels in variable. For a variable, say, 'Relationship' with three levels, namely, 'Single', 'In a relationship', and 'Married', we create a dummy table like the following:

Variable	Dummy_1	Dummy_2	Dummy_3
Relationship Status	Single	In a Relationship	Married
Single	1	0	0
In a Relationship	0	1	0
Married	0	0	1

In the above table, it clearly indicates there is no need to define three different levels. Two levels are sufficient to explain the three levels. If we drop the single level then ‘in a Relationship’ and ‘Married’ are enough to explain.

Let's drop the dummy variable 'Single' from the columns and see what the table looks like:

Relationship Status	In a Relationship	Married
Single	0	0
In a Relationship	1	0
Married	0	1

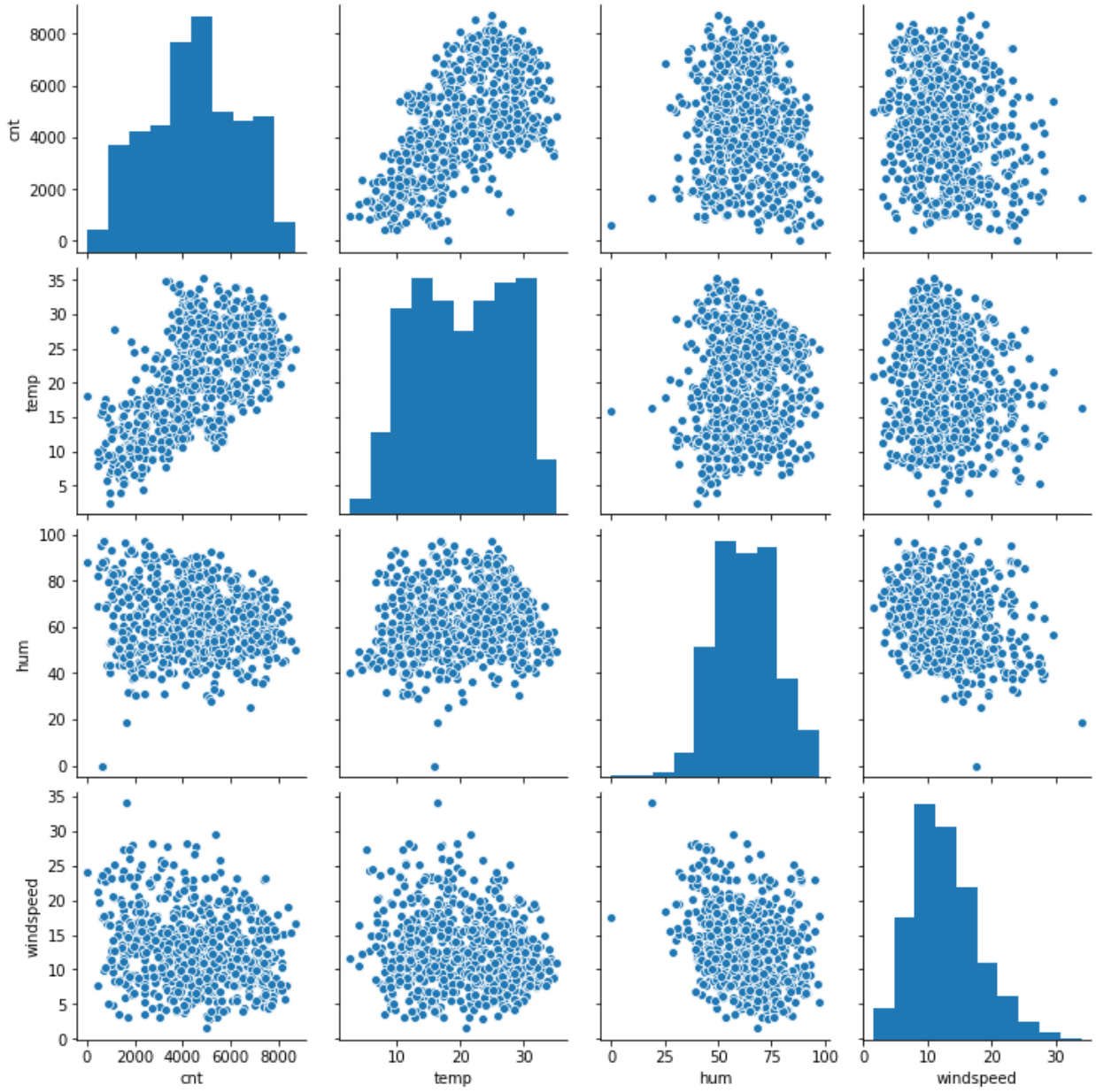
If both the dummy variables, i.e., 'In a relationship' and 'Married', are equal to zero, it means that the person is single. If 'In a relationship' is denoted by 1 and 'Married' by 0, it means that the person is in a relationship. Finally, if 'In a relationship' is denoted by 0 and 'Married' by 1, it means that the person is married.

So When we have a categorical variable with, say, 'n' levels, the idea of dummy variable creation is to build 'n-1' variables, indicating the levels. If we will not drop the first level then this column will have high correlation with other levels and result in multicollinearity in the data set.

So it is always advisable to use **drop_first=True** during dummy variable creation.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

For reference below is the pair plot between all numerical variables



Above plot clearly indicates the independent variable **'temp'** has the highest correlation with the target variable 'cnt'.

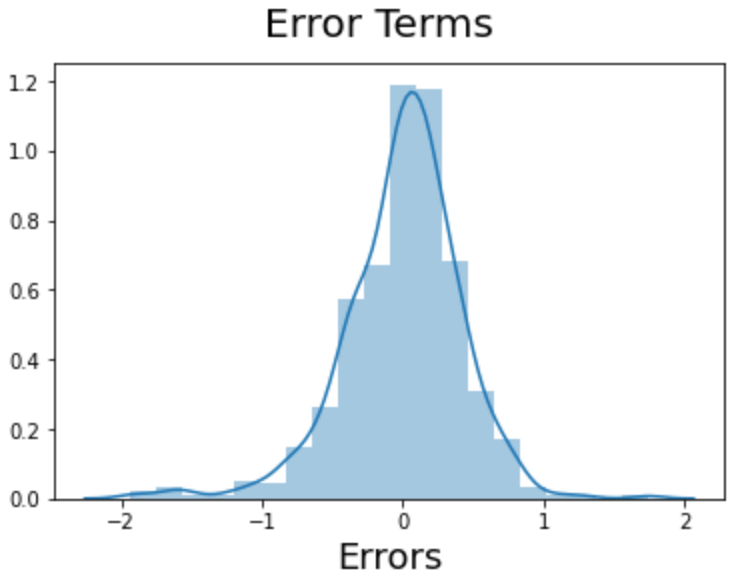
Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

After building the model on the train data set, we should check the linear regression assumptions to validate the model. There are three main linear regression assumptions. They are as follows:

1. Error terms are normally distributed with mean zero
2. Error terms have constant variance (homoscedasticity):
3. Error terms are independent of each other or no autocorrelation between residuals or error terms

We will check our model with the above assumptions and validate.

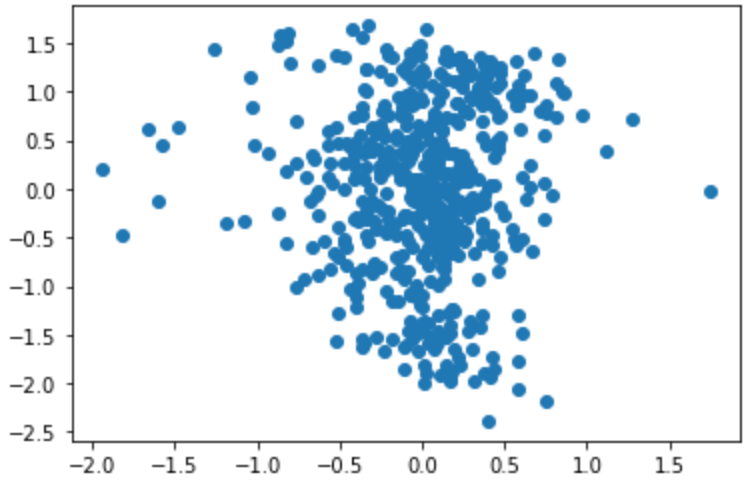
1. We plot the distribution plot of residuals and find the plot below.



We also calculated the mean of residuals : np.mean(res) -
mean: 4.0621101253932197e-16

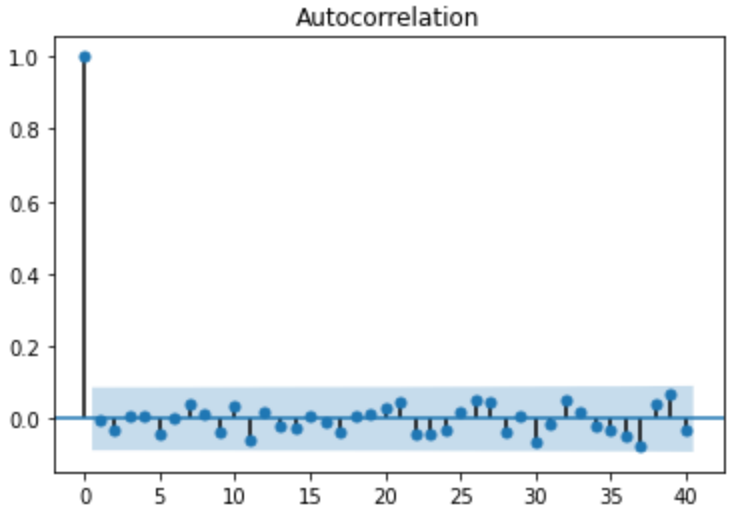
Above plot is centered around zero and mean is almost zero so Error terms are normally distributed

2. We plot the residuals with y_train predicted and find the plot below



Above graph clearly indicates there is no pattern and most of the values are centered around zero so there is no homoscedasticity or having constant variance.

3. We plot the Autocorrelation of residuals and find the plot below



All auto correlation values are in significance limit and there is no autocorrelation between residuals.

Our model clearly validates all assumptions of linear regression.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

We can see that the equation of our best fitted line is:

cnt = 0.2553 + 0.5207 * yr + 0.4723 * temp - 0.5113 * season_Spring + 0.2336 * season_Winter - 0.2933 * mnth_Jul + 0.2688 * mnth_Sep - 0.1981 * weekday_Sun - 1.3369 * weathersit_Light Snow + Light Rain - 0.3483 * weathersit_Mist + Cloudy

Below are the top 3 features in bike sharing model:

1. Light Snow & Light Rain weather (weathersit_Light Snow + Light Rain)
2. Year (yr)
3. Temperature (temp)

General Subjective Questions

Q 1. Explain the linear regression algorithm in detail.

Linear regression is a statistical model through which we analyze the linear relationship between a dependent variable and independent variables. Linear relationship means value of dependent variable changes(increase or decrease) accordingly when value of independent variables changes(increase or decrease)

Linear regression model can be represented with the help of following equation −

$$Y=mX+b$$

Here, Y is the dependent variable

X is the independent variable we are using to make predictions.

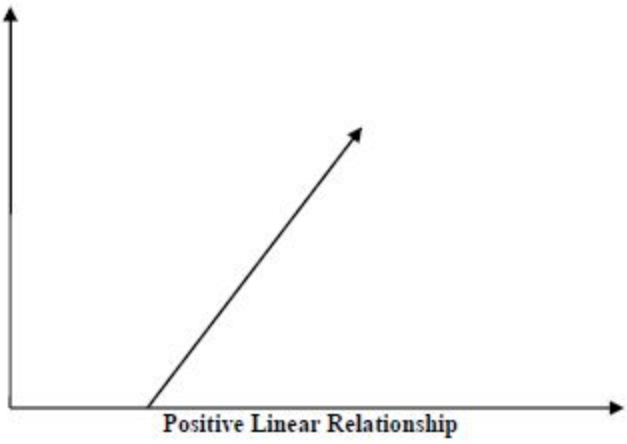
m is the slope of the regression line

b is a constant, known as the Y-intercept.

The linear relationship can be positive or negative as explained below −

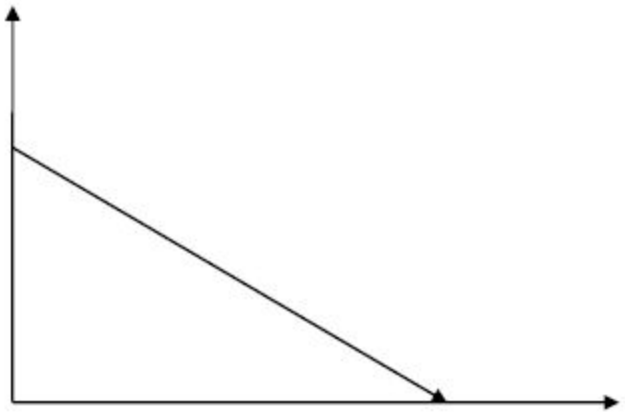
Positive Linear Relationship

A linear relationship will be called positive if both independent and dependent variables increase. It can be understood with the help of following graph −



Negative Linear relationship

A linear relationship will be called negative if independent increases and dependent variables decrease. It can be understood with the help of following graph −



Negative Linear Relationship

Types of Linear Regression

Linear regression can be of two types:

- **Simple Linear Regression:**
  If a single independent variable is used to predict the value of a numerical dependent variable, then this 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 this is called Multiple Linear Regression.

Finding the best fit line:

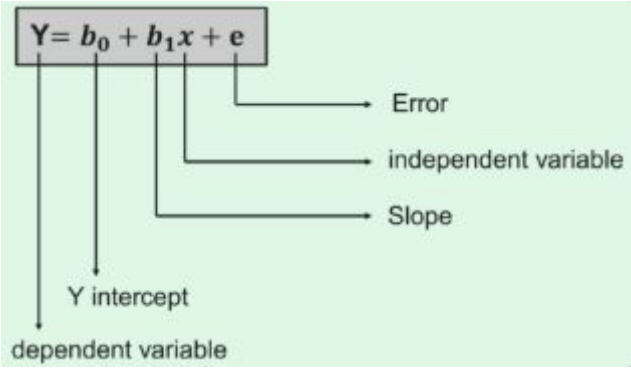
Whenever we build any model there is always an error between predicted values and actual values. In linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized. To find the best fitted line we use the cost function.

Cost function-

Residuals: The distance between the actual value and predicted values is called residual. If the observed points are far from the regression line, then the residual

will be high, and so the cost function will be high. If the scatter points are close to the regression line, then the residual will be small and hence the cost function.

The best fit line can be based on the linear equation given below.



- The dependent variable is denoted by Y.
- A line that touches the y-axis is denoted by the intercept b0.
- b1 is the slope of the line, x represents the independent variables
- The error in the resultant prediction is denoted by e.

The **cost function** provides the best possible values for b0 and b1 to make the best fit line for the data points.

$$minimize \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

$$J = \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

We choose the function above to minimize the error. We square the error difference and sum the error over all data points, the division between the total number of data points. Then, the produced value provides the averaged square error over all data points. It is also known as MSE(Mean Squared Error)

Model Performance:

The process of finding the best model out of various models is called optimization. We use R-squared method to find the best model:

R-squared method:

- R-squared is a statistical method that determines the goodness of fit.
- It measures the strength of the relationship between the dependent and independent variables on a scale of 0-100%.
- The high value of R-square determines the less difference between the predicted values and actual values and hence represents a good model.
- It can be calculated from the below formula:

$$\text{R-Squared} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

RSS = Residual sum of square
TSS = Total sum of square

Assumptions of Linear Regression

Below are some important assumptions of Linear Regression.

Linear relationship between the independent and target variables:

Linear regression assumes the linear relationship between the dependent and independent variables.

Small or no multicollinearity between the features:

Multicollinearity means high-correlation between the independent variables. Due to multicollinearity, it may be difficult to find the true relationship between the predictors and target variables. So, the model assumes either little or no multicollinearity between the features or independent variables.

Homoscedasticity Assumption:

Homoscedasticity is a situation when the error term is the same for all the values of independent variables. With homoscedasticity, there should be no clear pattern distribution of data in the scatter plot.

Normal distribution of error terms:

Linear regression assumes that the error term should follow the normal distribution pattern with mean zero.

No autocorrelations:

The linear regression model assumes no autocorrelation in error terms. If there will be any correlation in the error term.

Linear Regression Use Cases

- Sales Forecasting
- Risk Analysis
- Housing Applications To Predict the prices and other factors
- Finance Applications To Predict Stock prices, investment evaluation, etc.

Q 2: Explain the Anscombe's quartet in detail.

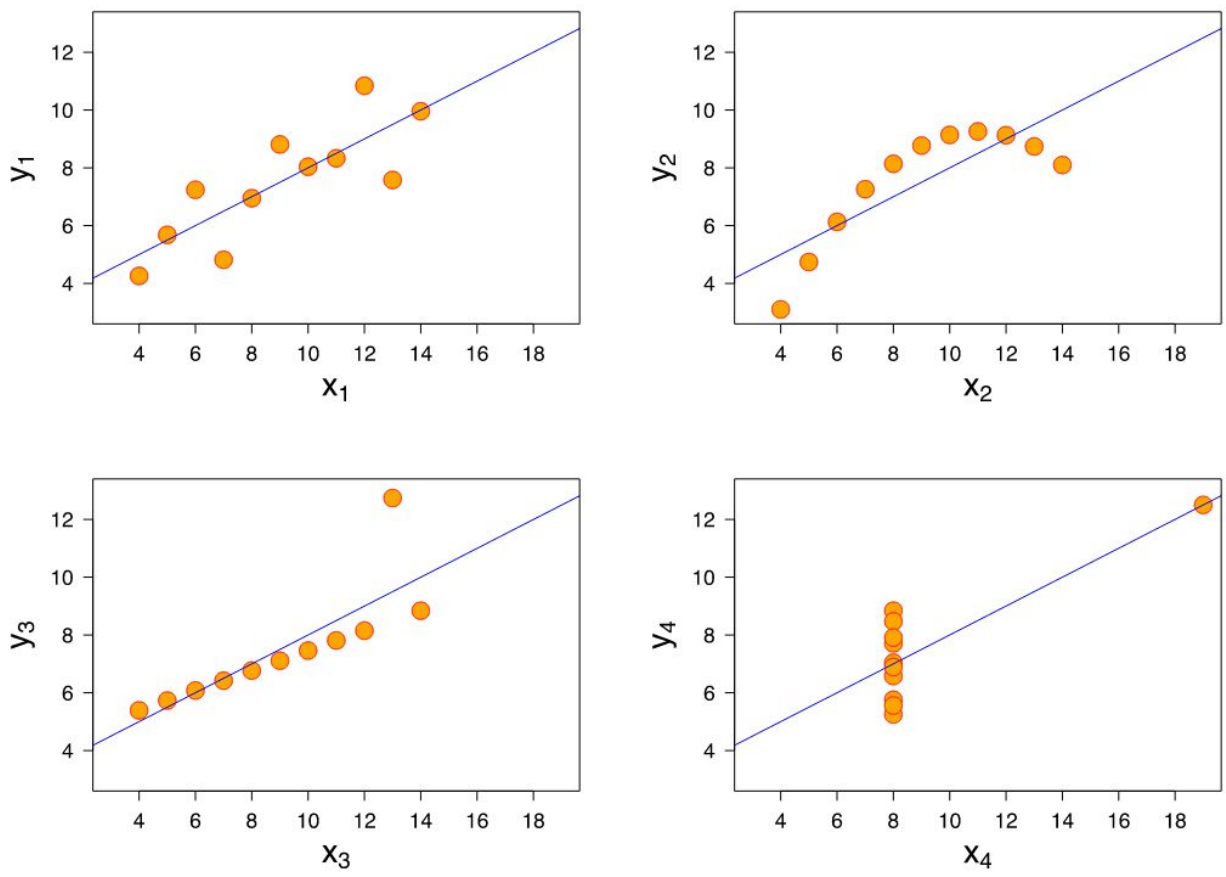
Statistics are great for describing general trends and aspects of data, but statistics alone can't fully depict any data set. **Francis Anscombe** realized this in 1973 and created a group of four datasets that appear to be similar when using typical summary statistics, yet tell four different stories when graphed. Each dataset consists of eleven (x,y) pairs as follows:

I		**II**		**III**		**IV**	
x1	y1	x2	y2	x3	y3	x4	y4
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

All the summary statistics you'd think to compute are close to identical:

- The average x value is 9 for each dataset
- The average y value is 7.50 for each dataset
- The variance for x is 11 and the variance for y is 4.12
- The correlation between x and y is 0.816 for each dataset
- A linear regression (line of best fit) for each dataset follows the equation y = 0.5x + 3

So far these four datasets appear to be pretty similar. But when we plot these four data sets on an x/y coordinate plane, we get the following results:



He saw that the **graphs were completely different even though the summary was exactly similar**.

- The first scatter plot (top left) appeared to be a simple linear relationship, corresponding to two variables correlated and following the assumption of normality.
- The second graph (top right) was not distributed normally; while a relationship between the two variables is obvious
- In the third graph (bottom left), the distribution is linear, but should have a different regression line

- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

Anscombe's Quartet is a great demonstration of the importance of graphing data to analyze it.
Additionally, Anscombe's Quartet warns of the dangers of outliers in data sets. If the bottom two graphs didn't have that one point that strayed so far from all the other points, their statistical properties would no longer be identical to the two top graphs.

Q 3. What is Pearson's R?

Pearson's correlation coefficient is a statistical measure of the strength of a linear relationship between paired data. it is denoted by *r* and varies from -1 to 1

$$-1 <= r <= +1$$

Furthermore:

- Positive values denote positive linear correlation;
- Negative values denote negative linear correlation;
- A value of 0 denotes no linear correlation;
- The closer the value is to 1 or –1, the stronger the linear correlation

We can describe the strength of the correlation using the below table for the absolute value of *r*:

- .00-.19 "very weak"
- .20-.39 "weak"
- .40-.59 "moderate"
- .60-.79 "strong"
- .80-1.0 "very strong"

Assumptions

The calculation of Pearson's correlation coefficient and subsequent significance testing of it requires the following data assumptions to hold:

- interval or ratio level;
- linearly related;
- bivariate normally distributed.

Q4 : What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is a technique to standardize the independent variable present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying values.

Why is scaling performed?

Most of the time, the collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then the algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

Example: If an algorithm is not using a feature scaling method then it can consider the value 3000 meter to be greater than 5 km 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 magnitudes.

It is important to note that **scaling just affects the coefficients** and none of the other parameters like **t-statistic, F-statistic, p-values, R-squared**, etc.

Normalized scaling Vs standardized scaling

Normalization:

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Here's the formula for normalization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively.

- When the value of X is the minimum value in the column, the numerator will be 0, and hence X' is 0
- On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X' is 1
- If the value of X is between the minimum and the maximum value, then the value of X' is between 0 and 1

Standardization:

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

Here's the formula for standardization:

$$X' = \frac{X - \mu}{\sigma}$$

μ is the mean of the feature values and σ is the standard deviation of the feature values.

Normalization is a good technique to use when you do not know the distribution of your data or when you know the distribution is not Gaussian (a bell curve). Normalization is useful when your data has varying scales and the algorithm you are using does not make assumptions about the distribution of your data.

Standardization assumes that your data has a Gaussian (bell curve) distribution. This does not strictly have to be true, but the technique is more effective if your attribute distribution is Gaussian. Standardization is useful when your data has varying scales and the algorithm you are using does make assumptions about your data having a Gaussian distribution, such as linear regression, logistic regression etc.

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

The variance inflation factor (VIF) measures the extent of correlation between two independent variables in a model. It is used for diagnosing multicollinearity. Higher values signify that independent variables are correlated and it is difficult to assess accurately the contribution of predictors to a model.

How the VIF is computed

The extent to which a predictor is correlated with the other predictor variables in a linear regression can be calculated by the *R-squared* statistic of the regression. The *variance inflation* for a variable is then computed as:

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

How to interpret the VIF

A VIF can be computed for each predictor in a predictive model. A value of 1 means that the predictor is not correlated with other variables. The higher the value, the greater the correlation of the variable with other variables. Values of more than 5 are regarded as being high, with values of 10 or more being regarded as very high.

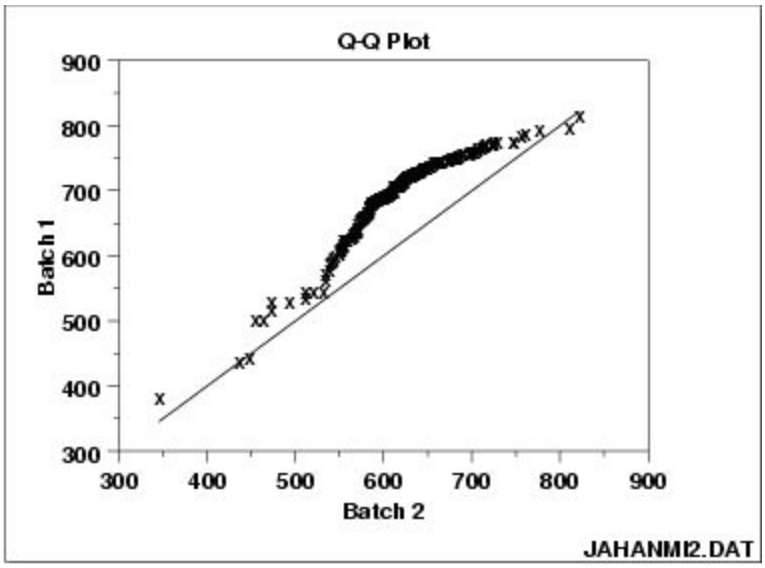
If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

An **infinite VIF value indicates** that the corresponding variable may be expressed exactly by a linear combination of other variables.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Q–Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. Also, it helps to determine if two data sets come from populations with a common distribution.

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.



This q-q plot shows that

1. These 2 batches do not appear to have come from populations with a common distribution.

2. The batch 1 values are significantly higher than the corresponding batch 2 values.
3. The differences are increasing from values 525 to 625. Then the values for the 2 batches get closer again.

The q-q plot is formed by:

Vertical axis: Estimated quantiles from data set 1
Horizontal axis: Estimated quantiles from data set 2

Both axes are in units of their respective data sets. That is, the actual quantile level is not plotted. For a given point on the q-q plot, we know that the quantile level is the same for both points, but not what that quantile level actually is.

If the data sets have the same size, the q-q plot is essentially a plot of sorted data set 1 against sorted data set 2. If the data sets are not of equal size, the quantiles are usually picked to correspond to the sorted values from the smaller data set and then the quantiles for the larger data set are interpolated.

The q-q plot is used to answer the following questions in linear regression:

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behavior?