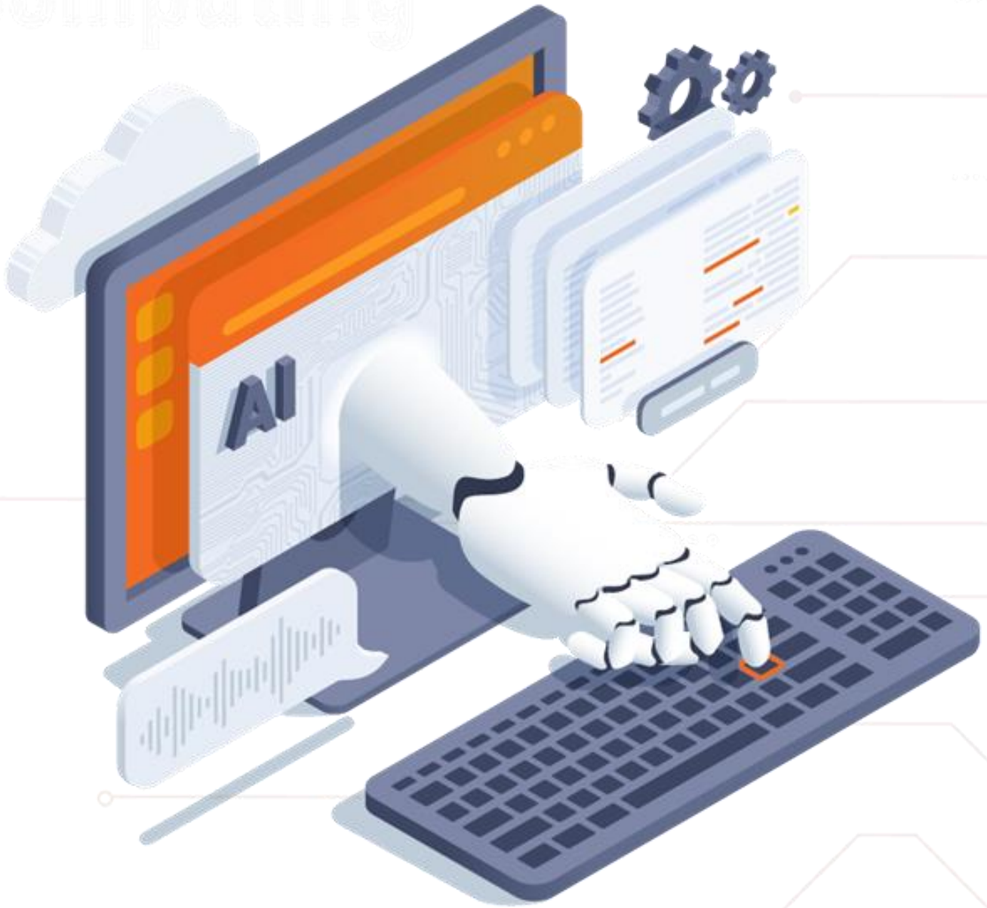
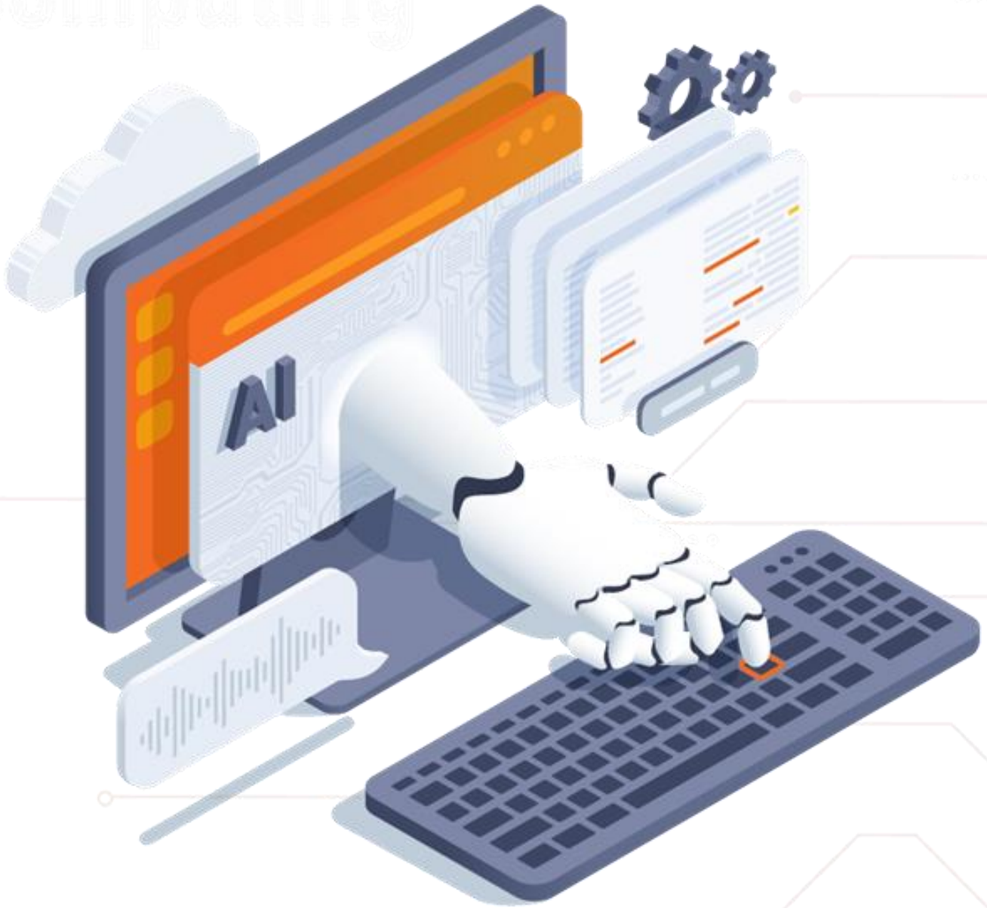


Caltech | Center for Technology & Management Education

Introduction to Artificial Intelligence

Powerd by simplilearn



Caltech | Center for Technology & Management Education

Machine Learning Workflow

Powerd by simplilearn

Learning Objectives

By the end of this lesson, you will be able to:

- Describe the seven steps of machine learning workflow

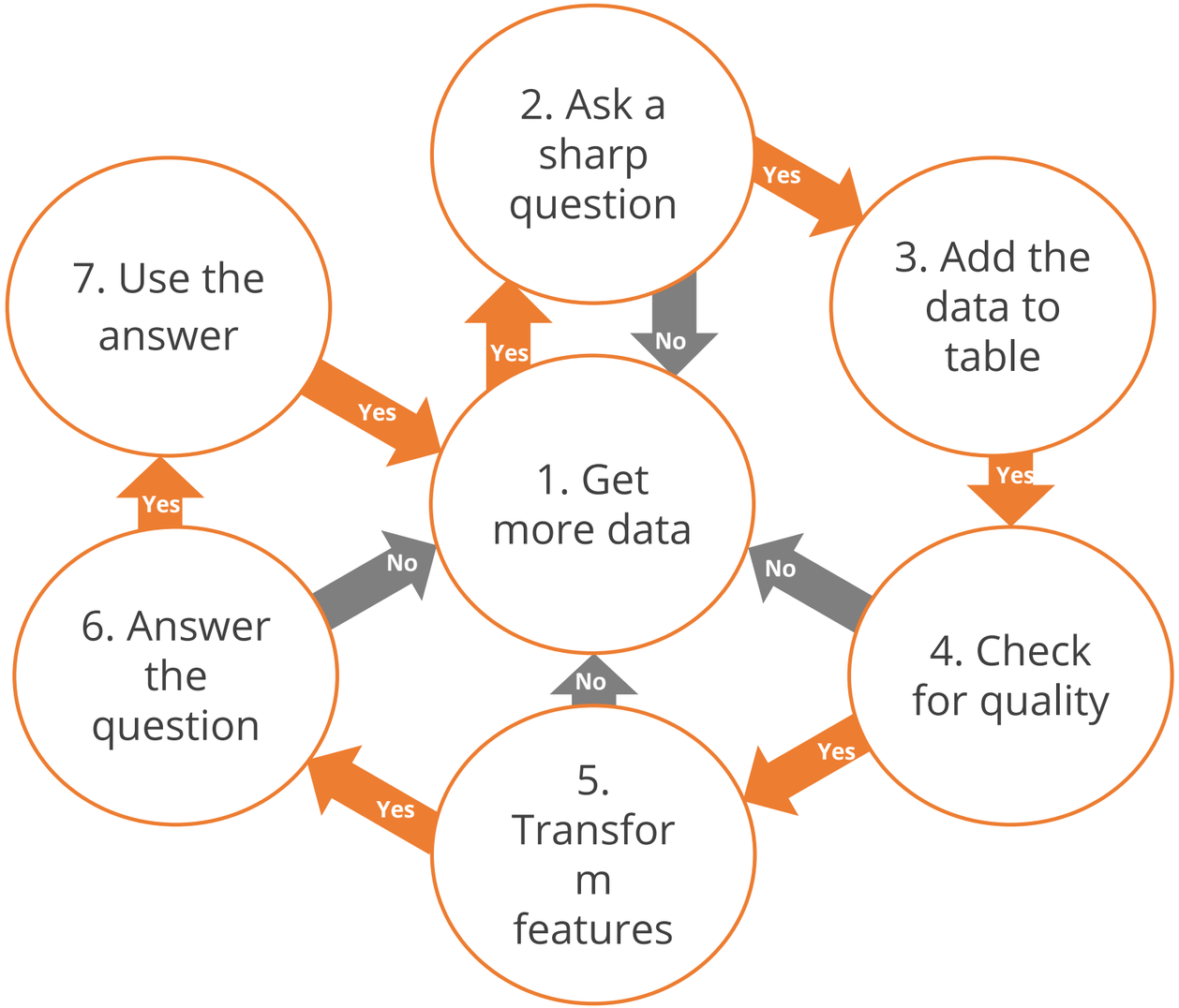
Machine Learning Process

The Machine Learning Workflow

It is essential for all technical and non-technical stakeholders to understand machine learning workflow to understand:

- The job of data scientist
- The processes a data scientist follows to provide feedback to decision-makers
- The machine learning process in a business environment

The Machine Learning Process



Step 1: Get More Data

- The data collected is used to investigate a business challenge.
- The quality of the predictive model depends upon the quantity and quality of the data gathered.
- The data can be collected in different formats.

Data Format Examples

Names

Type	Sonia Tran
Variety	Caramel latte
Id	Air Force One
Model number	R2-D2
Category	Chocolate
Text	Best. Show. Ever.

Numbers

Money	$300m
Count	69 pizzas
Pixel brightness	232/255
Temperature	30 degree F
Sound intensity	0.64

Names that look like numbers

Zip code	95126
Social security number	602-47-1899
Serial number	100000023987
Credit card number	5467-3345-2122-5508
Sound intensity	0.64

Names that look like numbers and can be turned into numbers

Place	First, second, third
Time zone	Pacific, mountain, central, eastern
Train stops	Diridon, San Francisco, Sunnyvale, Menlo Park
Side	Left, middle, right
Sound intensity	0.64

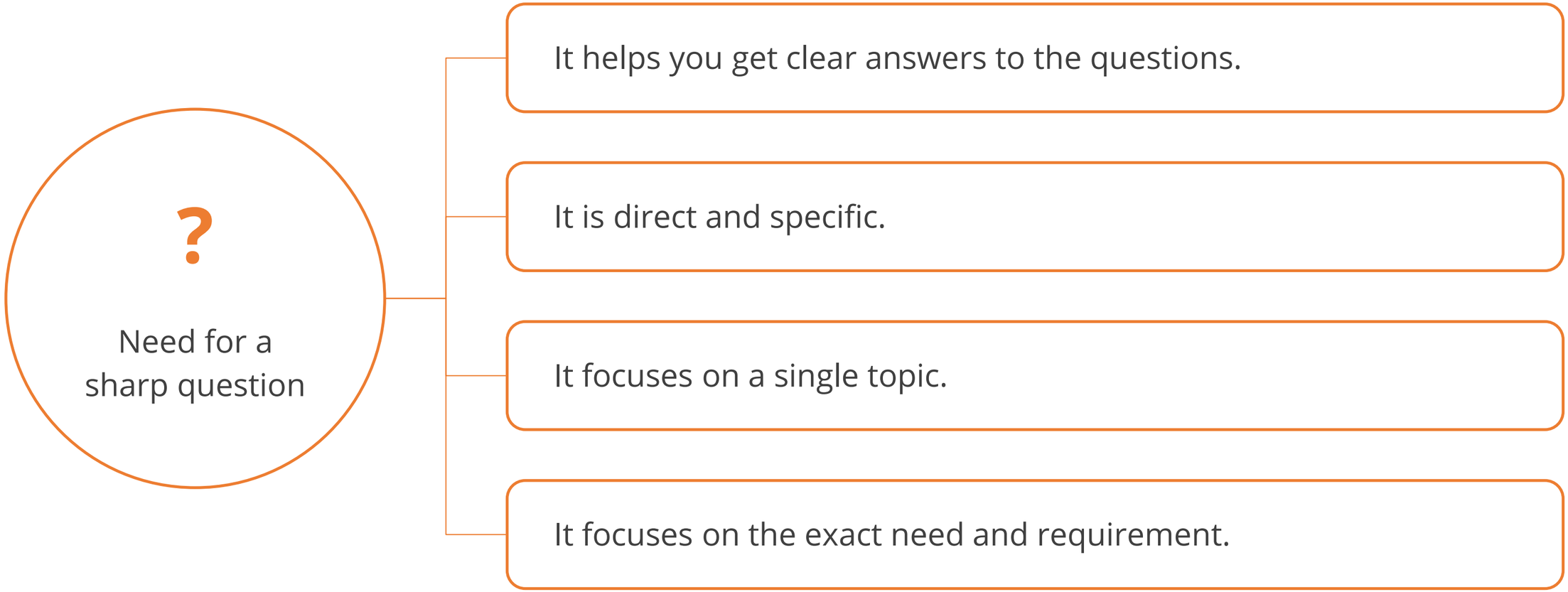
Goals of Machine Learning Workflow

The goals of data science and machine learning is to:

- Derive answers to business challenges
- Derive meaningful conclusions from complicated issues
- Identify actionable steps given a wide set of variables

To be able to do so, you need to ask a sharp question as opposed to a vague one.

Step 2: Ask a Sharp Question



Vague vs. Sharp Questions

Vague questions

1. What should you do?
2. How should you live my life?
3. Which career path should you take?
4. Which data can tell you about your business?

Sharp questions

1. Which route will get you to work fastest?
2. How many times a user will use the new product features?
3. How can the revenue be maximized?

Sharp Question Example

Goal

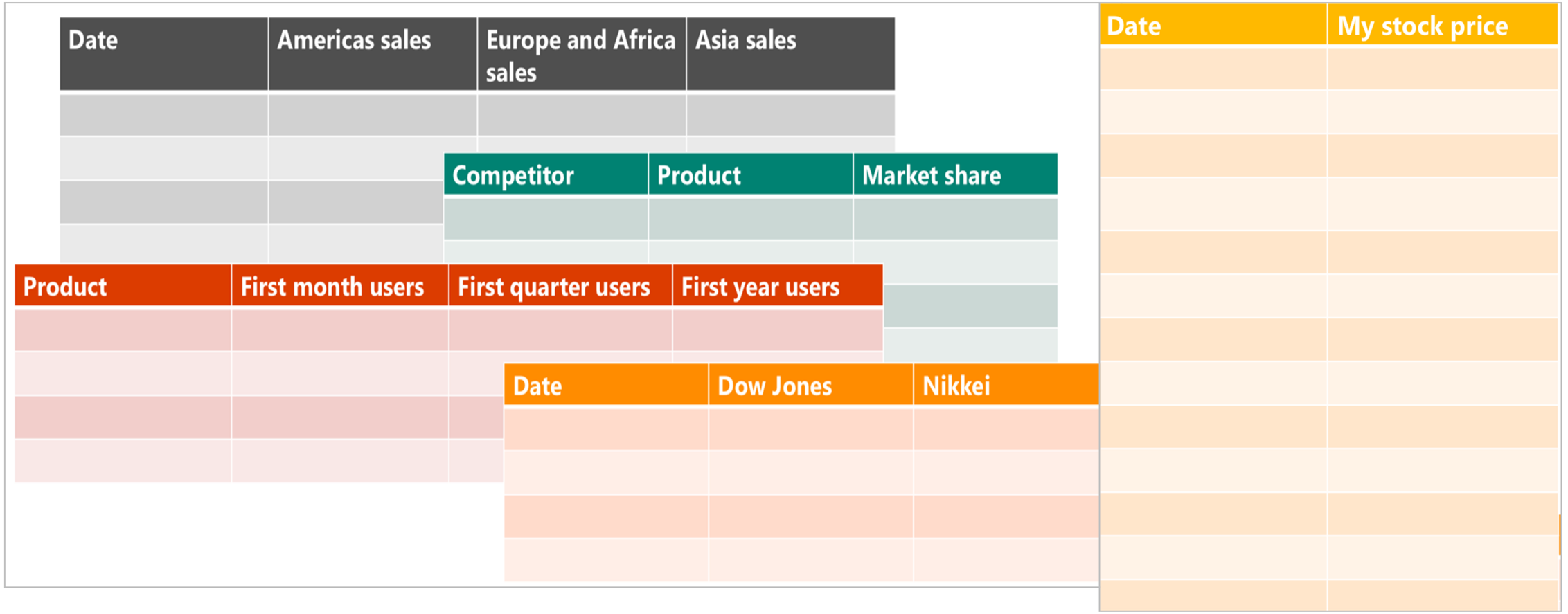
- Your ultimate goal is to be able to analyze your historical data and predict the stock price at a future date.
- You have to study all different tables of data in your database and analyze how your company is doing month by month in terms of sales.
- This will ultimately lead you to understand how the company is doing in terms of it's market share.

Sharp Question

- What will be your company's stock price next week?

Sharp Question Example

The company's database is divided into following tables:



Step 3: Add Data to the Table

01 Data analyst arranges data in database tables in a systematic manner.

02 Systematic arrangement of data helps in detailed analysis.

03 Data is stored in the table in the form of columns and rows.

04 Table columns represent data of a single type and rows represent records pertaining to one entity.

05 The final step is to aggregate, distribute, compute, or measure to derive a data analysis.

Data Analysis in Machine Learning

- Data analysis is the process of deriving new findings from the historical data.
- It mainly focuses on aggregating table data to find the answers to various business problems.
- It is one of the essential steps performed by data analysts to build machine learning algorithm.

Example: Add Data to the Table

- Each table row represents observations across given attributes.
- The stock price column shows the stock value across different dates.

Stock price	Date	Day of week	Dow Jones	Last month sales	Last quarter sales	Market share	New users last month	New users last quarter	Days since press release	Days since product release	Total users
57.3	5/21	Tue	17,245	68.8M	211.2M	23.1%	63,522	195,322	3	96	2.49M
58.8	5/22	Wed	17,289	68.8M	211.2M	23.1%	63,522	195,322	4	97	2.49M
56.9	5/23	Thu	17,115	68.8M	211.2M	23.1%	63,522	195,322	5	98	2.49M
57.4	5/24	Fri	17,278	68.8M	211.2M	23.1%	63,522	195,322	6	99	2.49M

Example: Data Analysis

Aggregate and distribute the data as shown here:

Quarter	Total Sales
2015Q4	119.2M
2016Q1	221.0M
2016Q2	215.9M
2016Q3	189.3M
2016Q4	211.2M
…	…

Month	Total Sales
2016/01	43.0M
2016/02	60.1M
2016/03	55.5M

Example: Aggregate

- You can aggregate the data in the table to derive answers.
- This process is called data analysis and involves counting total observations in a table or combining data from multiple tables.

Stock price	Date	Day of week	Dow Jones	Last month sales	Last quarter sales	Market share	New users last month	New users last quarter	Days since press release	Days since product release	Total users
57.3	5/21	Tue	17,245	68.8M	211.2M	23.1%	63,522	195,322	3	96	2.49M
58.8	5/22	Wed	17,289	68.8M	211.2M	23.1%	63,522	195,322	4	97	2.49M
56.9	5/23	Thu	17,115	68.8M	211.2M	23.1%	63,522	195,322	5	98	2.49M
57.4	5/24	Fri	17,278	68.8M	211.2M	23.1%	63,522	195,322	6	99	2.49M

Example: Aggregate and Distribute

You can focus on all observations for a particular column or feature and total it.

Month	Total sales
2016/01	43.0M
2016/02	60.1M
2016/03	55.5M
2016/04	41.7M
2016/05	68.8M
...	...

Quarter	Total sales
2015Q4	119.2M
2016Q1	221.0M
2016Q2	215.9M
2016Q3	189.3M
2016Q4	211.2M
...	...

Example: Distribute, Compute, and Measure

- This is an example of performing aggregate, distribute, compute and measure operations on data in tables.
- Each feature and their observations are distributed across the table and then combined.

Stock price	Date	Day of week	Dow Jones	Last month sales	Last quarter sales	Market share	New users last month	New users last quarter	Days since press release	Days since product release	Total users
57.3	5/21	Tue	17,245	68.8M	211.2M	23.1%	63,522	195,322	3	96	2.49M
58.8	5/22	Wed	17,289	68.8M	211.2M	23.1%	63,522	195,322	4	97	2.49M
56.9	5/23	Thu	17,115	68.8M	211.2M	23.1%	63,522	195,322	5	98	2.49M
57.4	5/24	Fri	17,278	68.8M	211.2M	23.1%	63,522	195,322	6	99	2.49M

Example: Estimate

The market share column shows the estimated stock price values of the company that are derived from the previous steps.

Stock price	Date	Day of week	Dow Jones	Last month sales	Last quarter sales	Market share	New users last month	New users last quarter	Days since press release	Days since product release	Total users
57.3	5/21	Tue	17,245	68.8M	211.2M	23.1%	63,522	195,322	3	96	2.49M
58.8	5/22	Wed	17,289	68.8M	211.2M	23.1%	63,522	195,322	4	97	2.49M
56.9	5/23	Thu	17,115	68.8M	211.2M	23.1%	63,522	195,322	5	98	2.49M
57.4	5/24	Fri	17,278	68.8M	211.2M	23.1%	63,522	195,322	6	99	2.49M

Step 4: Check for Quality

Quality check determines if the data is acceptable for further investigation.

For an algorithm to work, the data in a column should be in a consistent format.

It involves computation and analysis of the data derived from previous steps.

Check for Quality: Example

- The Birth year column in the table has data format inconsistencies.
- The date in this column needs to be converted to a consistent format to make it readable for the ML algorithm.

ID	First name	Last name	Birth year	Height	Birthplace	Identity is secret	Can fly	Alignment	Wears cape
7435	Bruce	Wayne	1969*	6′ 2″	Gotham	Y	3	anti-villain	black
0958	Ororo	Munroe	--1979--	5′ 11″	Manhattan		9	good	long
9471	Diana	Trevor	1618	5′ 8″	Paradise Island	Y	Jet	truth	rarely
9483	Janet	Van Dyne	19.42	5′ 4″	Cresskill		tiny	Good	Not really
0696	Peter	Parker	1111983	5′ 10″	Queens	Y	Fall	right	never
5531	Harleen	Quinzell	1981	5′ 2″	Gotham	Y	-	evil	no
4734	Erik	Lehnsherr	1-9-3-2	6′ 0″	Hamburg		Lev.	mutants	Absolutely
7757	Natasha	Romanova	1983	5′ 7″	St. Petersburg		jet	depends	No way
0323	Jean	Grey	"1977"	5′ 6″	Annandale		No	good	Mostly not
3980	Clark	Kent	"1954"	6′ 4″	Krypton	Y	12	Truth	always
3057	Victor	Von Doom	"1943"	6′ 2″	Latveria		1	Bad	yes
0573	Stephen	Strange	1968	6′ 2″	Philidelphia		not	light	Y
7452	Thor	Odinson	2287 BC	6′ 6″	Norway		10	Good	Of course
1437	Selina	Kyle	1998	5′ 7″	Gotham	Y	NA	Neutral	It clashes
1883	Raven	Darkholme	..1911..	5′ 10″	unknown	Y	no	mostly bad	Not really
5830	Kara	Zor-el	1961	5′ 7″	Krypton	Y	fast	G	Yes

Check for Quality: Example

The Birth year column in the table has inconsistent format.

ID	First name	Last name	Birth year	Height	Birthplace	Identity is secret	Can fly	Alignment	Wears cape
7435	Bruce	Wayne	1969	6′ 2″	Gotham	Y	3	anti-villain	black
0958	Ororo	Munroe	1979	5′ 11″	Manhattan		9	good	long
9471	Diana	Trevor	1618	5′ 8″	Paradise Island	Y	Jet	truth	rarely
9483	Janet	Van Dyne	1942	5′ 4″	Cresskill		tiny	Good	Not really
0696	Peter	Parker	1983	5′ 10″	Queens	Y	Fall	right	never
5531	Harleen	Quinzell	1981	5′ 2″	Gotham	Y	-	evil	no
4734	Erik	Lehnsherr	1932	6′ 0″	Hamburg		Lev.	mutants	Absolutely
7757	Natasha	Romanova	1983	5′ 7″	St. Petersburg		jet	depends	No way
0323	Jean	Grey	1977	5′ 6″	Annandale		No	good	Mostly not
3980	Clark	Kent	1954	6′ 4″	Krypton	Y	12	Truth	always
3057	Victor	Von Doom	1943	6′ 2″	Latveria		1	Bad	yes
0573	Stephen	Strange	1968	6′ 2″	Philidelphia		not	light	Y
7452	Thor	Odinson	-2287	6′ 6″	Norway		10	Good	Of course
1437	Selina	Kyle	1998	5′ 7″	Gotham	Y	NA	Neutral	It clashes
1883	Raven	Darkholme	1911	5′ 10″	unknown	Y	no	mostly bad	Not really
5830	Kara	Zor-el	1961	5′ 7″	Krypton	Y	fast	G	Yes

Step 5: Transform Features

- This step includes **Feature Engineering**.
- Each characteristic of a data element is known as a **feature**.
- Feature engineering enables you to make sense out of the data, especially when there are multiple features.
- Some features may not give useful information for the model, whereas some features may be combined to derive meaningful information.
- Feature engineering helps you overcome such challenges.

Tricks of Feature Engineering

Data-specific	• Scale Invariant Feature Transform (SIFT): Images • Term Frequency-Inverse Document Frequency (TF-IDF): Text
Domain-specific	• Econometric, technological, agricultural, and sociological data engineering
Deep Learning	• Images, text, and audio data engineering

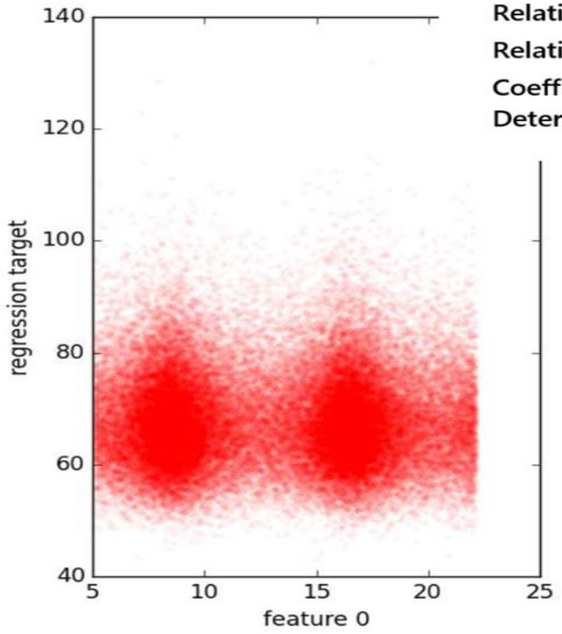
Transform Features: Example

- There are 3 columns and 65,670 rows.
- Features 0 and 1 have similar values.
- The numbers are meaningless and scattered.

rows 65670 | columns 3

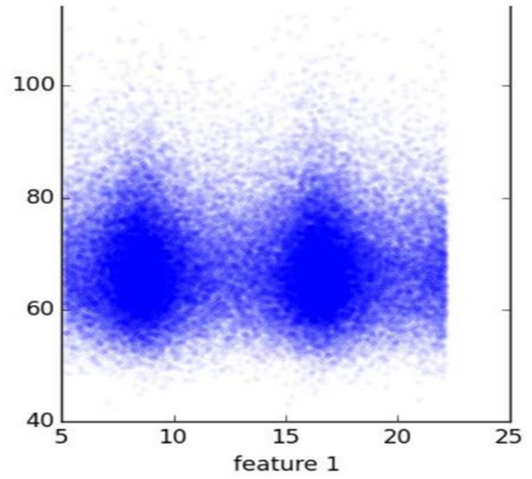
view as

0	1	2
5.107477	5.135881	60.479023
5.113939	5.141432	61.419001
5.117143	5.13772	82.774271
5.118805	5.145063	62.552338
5.119299	5.144294	66.799533
5.11949	5.140815	77.870507
5.120502	5.147892	64.326006
5.121868	5.14889	61.743756
5.121949	5.149292	64.493967
5.123392	5.148504	69.140338
5.124216	5.148921	69.449809



Metrics

Mean Absolute Error	6.485434
Root Mean Squared Error	8.280206
Relative Absolute Error	0.991422
Relative Squared Error	0.983903
Coefficient of Determination	0.016097



Transform Features: Example (contd.)

- Values of feature column 0 is multiplied with every observation in feature column 1.
- These values are plotted in image 2.

columns
4

0	1	2	Multiply(1_0)
5.107477	5.135881	60.479023	26.231395
5.113939	5.141432	61.419001	26.292971
5.117143	5.13772	82.774271	26.290449
5.118805	5.145063	62.552338	26.336574
5.119299	5.144294	66.799533	26.335178
5.11949	5.140815	77.870507	26.318351
5.120502	5.147892	64.326006	26.359789
5.121868	5.14889	61.743756	26.371937
5.121949	5.149292	64.493967	26.374413
5.123392	5.148504	69.140338	26.3778
5.124216	5.148921	69.449809	26.384186
5.126409	5.154655	62.028089	26.42487

Image 1

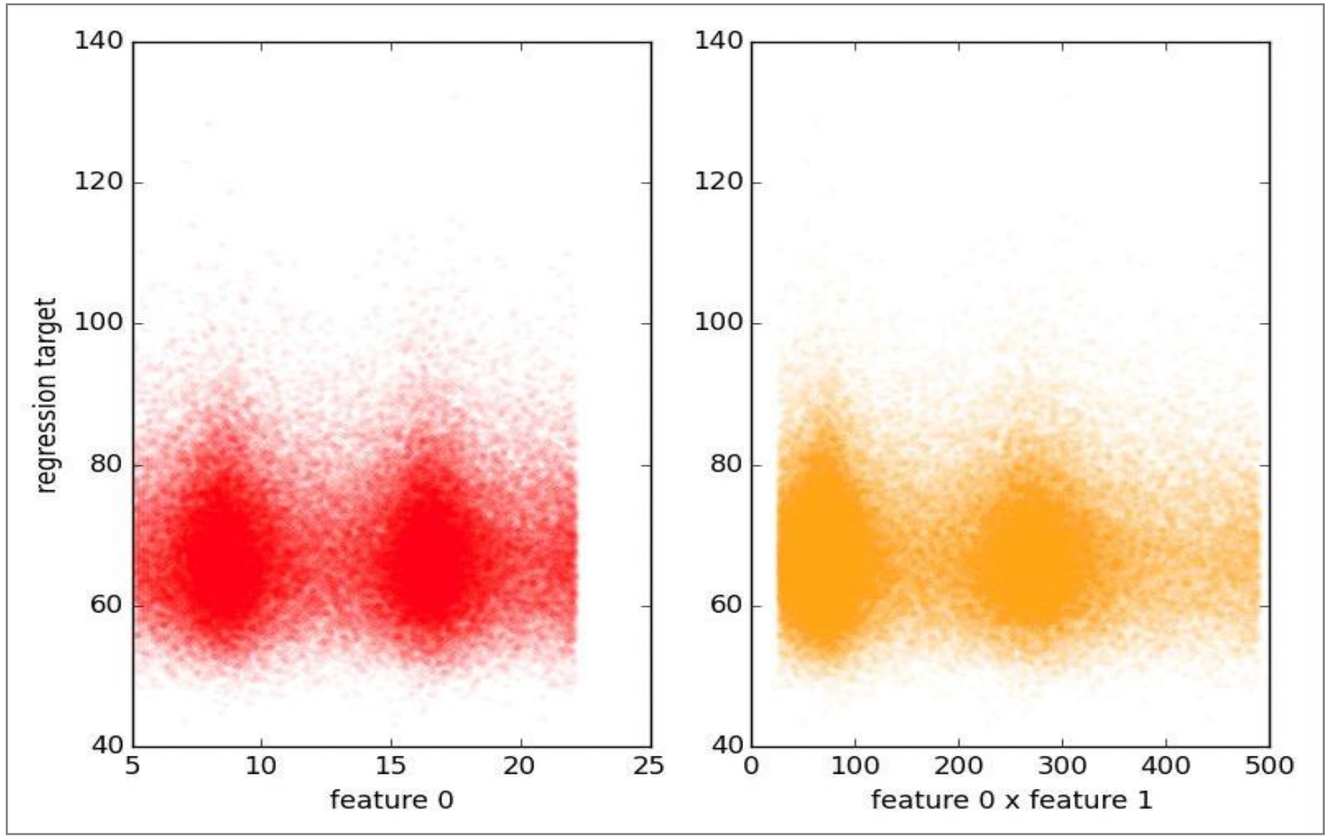


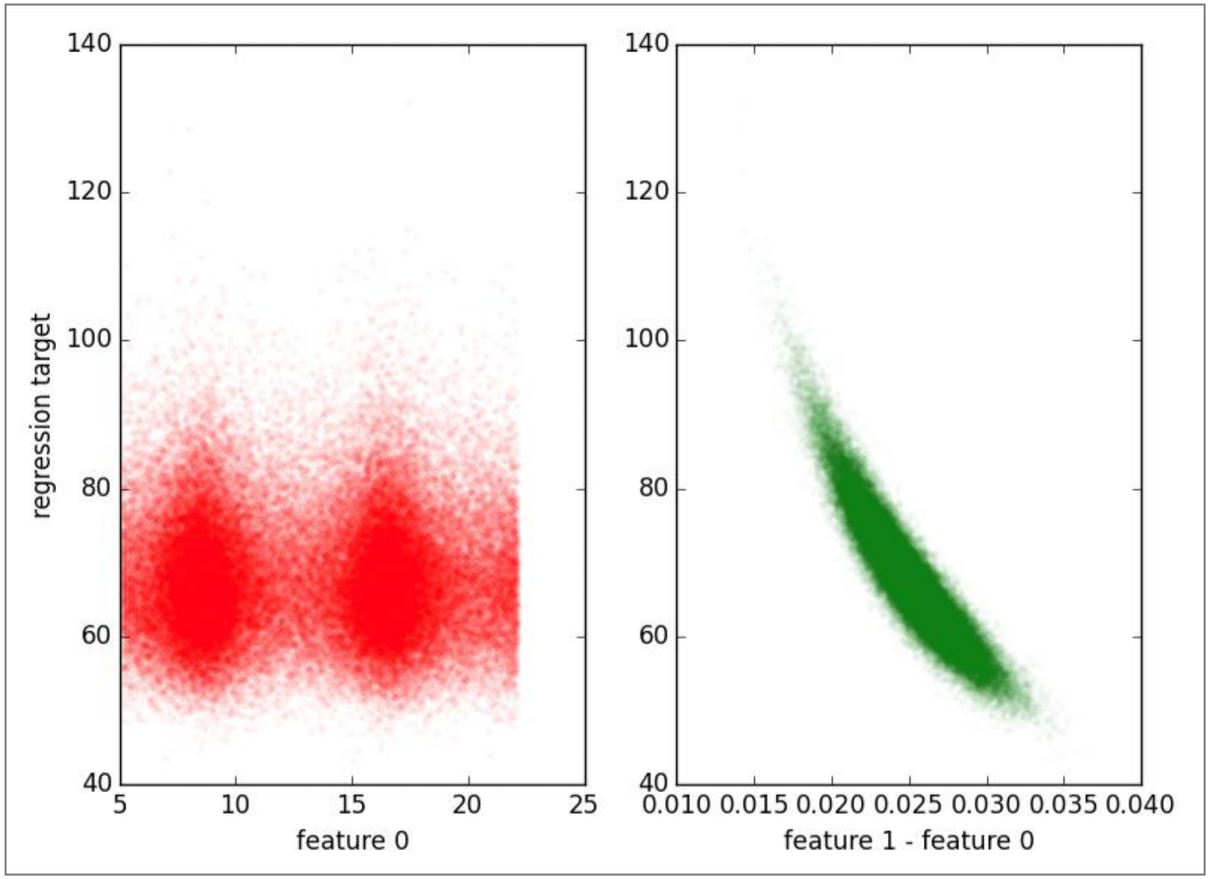
Image 2

Transform Features: Example (contd.)

- When we subtract feature 0 from feature 1 and plot it, we get a curve.
- This curve is normal or Gaussian distribution or bell-shaped curve.

columns
4

0	1	2	Subtract(1_0)
5.107477	5.135881	60.479023	0.028404
5.113939	5.141432	61.419001	0.027493
5.117143	5.13772	82.774271	0.020578
5.118805	5.145063	62.552338	0.026258
5.119299	5.144294	66.799533	0.024995
5.11949	5.140815	77.870507	0.021325
5.120502	5.147892	64.326006	0.02739
5.121868	5.14889	61.743756	0.027022
5.121949	5.149292	64.493967	0.027343
5.123392	5.148504	69.140338	0.025112
5.124216	5.148921	69.449809	0.024705
5.126409	5.154655	62.028089	0.028246



Step 6: Answer the Questions

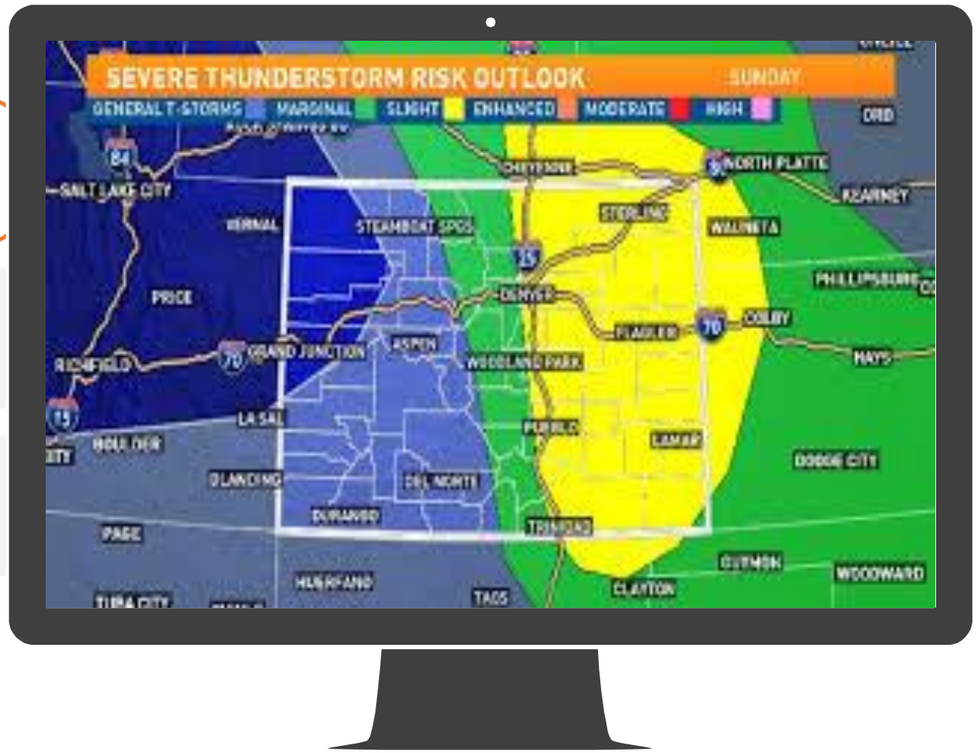
- This step helps analyze if the obtained answers are clear.
- These questions include:

1. How much or how many?
2. Which category?
3. Which group?
4. Does this look strange?
5. Which action?

Answer the Questions: Type 1

How much or how many?

1. What will be the temperature this Friday?
2. How many people will like my post?
3. What will be my product sales next month?



Answer the Questions: Type 1 (contd.)

How much or how many?

1. What will be the temperature this Friday?
2. How many people will like my post?
3. What will be my product sales next month?

Answer the Questions: Type 1 (contd.)

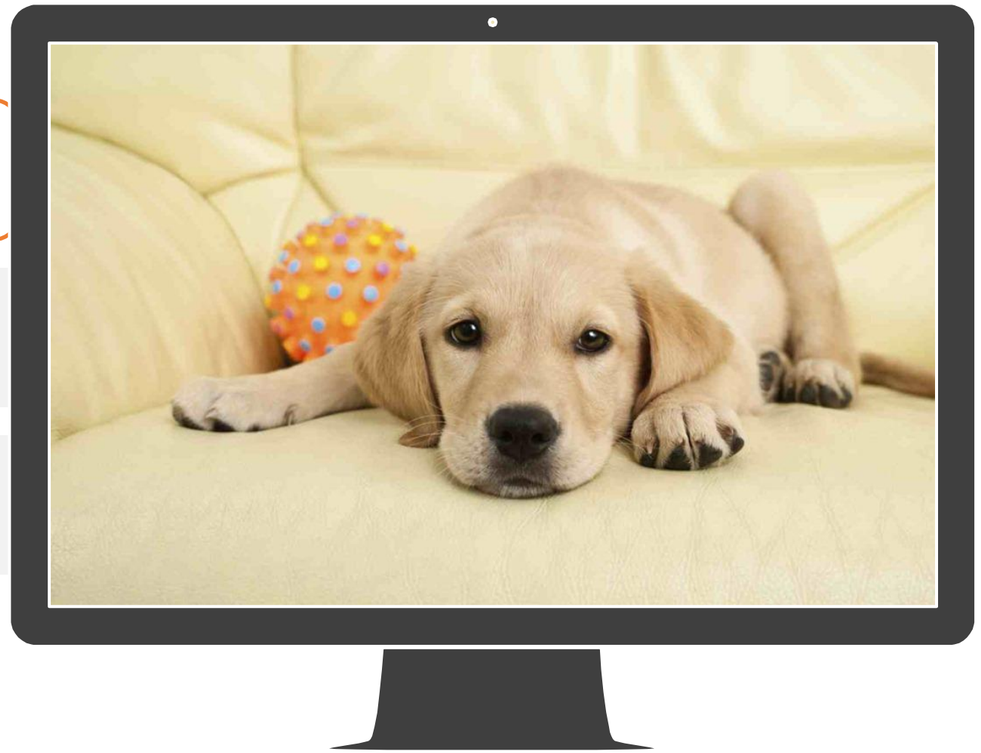
How much or how many?

1. What will be the temperature this Friday?
2. How many people will like my post?
3. What will be my product sales next month?

Answer the Questions: Type 2

Which category?

1. Is this an image of a dog?
2. What is the topic of this news article?
3. Which hotel in my area offers free Wi-Fi?



Answer the Questions: Type 2 (contd.)

Which category?

1 Is this an image of a dog?

2 What is the topic of this news article?

3 Which hotel in my area offers free Wi-Fi?

Answer the Questions: Type 2 (contd.)

Which category?

1. Is this an image of a dog?
2. What is the topic of this news article?
3. Which hotel in my area offers free Wi-Fi?

Answer the Questions: Type 3

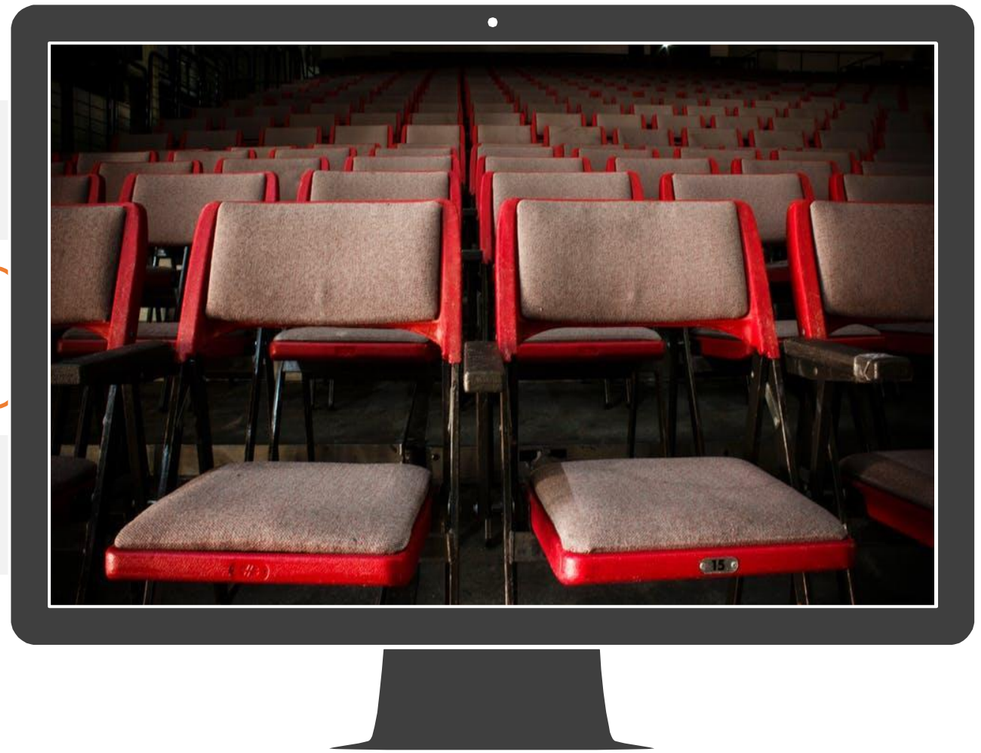
Which group?

1. Which shoppers purchase similar products?
2. Which group of viewers like horror movies?
3. How best can you divide this book into ten topics?

Answer the Questions: Type 3 (contd.)

Which group?

1. Which shoppers purchase similar products?
2. Which group of viewers like horror movies?
3. How best can you divide this book into ten topics?



Answer the Questions: Type 3 (contd.)

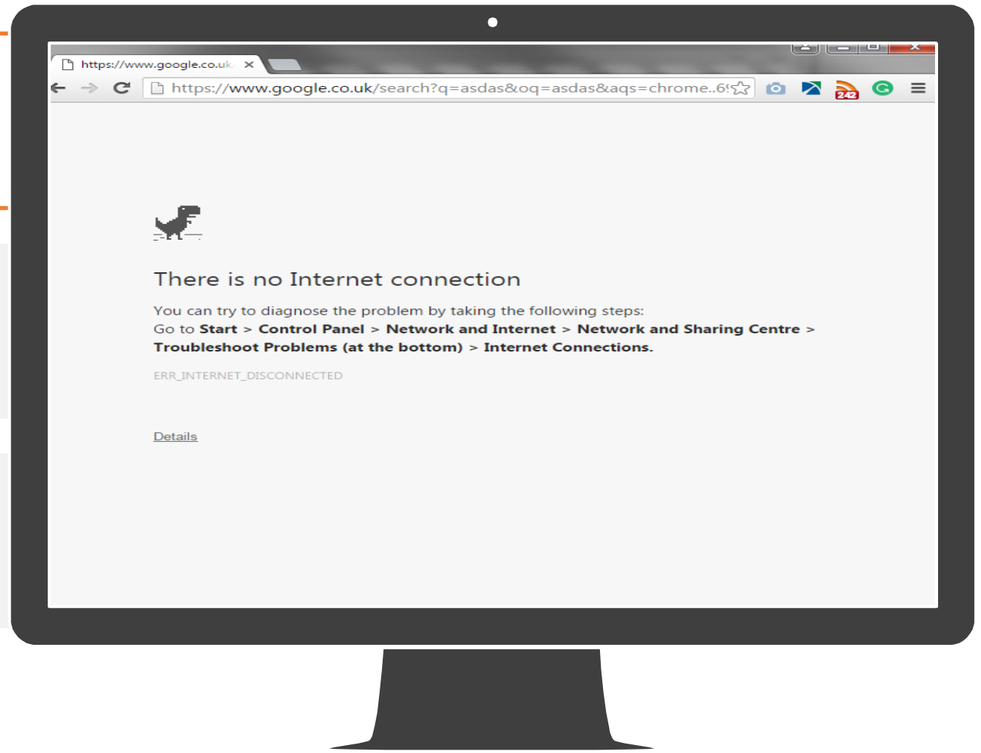
Which group?

1. Which shoppers purchase similar products?
2. Which group of viewers like horror movies?
3. How best can you divide this book into ten topics?

Answer the Questions: Type 4

Does this look strange?

1. Is this internet message typical?
2. Is this heart beat reading abnormal?
3. Does these transactions look unusual as opposed to customer's usual credit card transactions ?



Answer the Questions: Type 4 (contd.)

Does this look strange?

1. Is this internet message typical?
2. Is this heartbeat reading abnormal?
3. Does these transactions look unusual as opposed to customer's usual credit card transactions ?

Answer the Questions: Type 4 (contd.)

Does this look strange?

1. Is this internet message typical?
2. Is this heart beat reading abnormal?
3. **Do these transactions look unusual as opposed to customer's usual credit card transactions ?**

Answer the Questions: Type 5

Which action?

1. Should I vacuum again or should I not?
2. Should I run the red light?
3. Should I raise or lower the temperature ?

Answer the Questions: Type 5 (contd.)

Which action?

1. Should I vacuum again or should I not?
2. Should I run the red light?
3. Should I raise or lower the temperature ?

Answer the Questions: Type 5 (contd.)

Which action?

1. Should I vacuum again or should I not?
2. Should I run the red light?
3. **Should I raise or lower the temperature ?**

Step 7: Use the Answer

There are plenty of ways to use the answer derived from the previous step.

1. For making up a decision
2. For proposing the price of an item
3. For publishing the results obtained as a part of research paper
4. For constructing a dashboard on power BI
5. For making changes to product features

Note: Power BI is a business analytics tool by Microsoft.

Demo: Machine Learning Workflow

A demo on how a buyer decides which property he can purchase.

COURSE-END PROJECT

Key Takeaways

- Machine learning workflow involves seven steps.
- Step one involves getting more data, which is the process of deriving relevant data to answer business questions.
- The next step is to always ask sharp questions and avoid using vague ones to get the desired response for a question.
- Third step is to arrange the raw data in tables to analyze the data better.

Key Takeaways

- In the fourth step, data quality is checked to ensure data consistency.
- In the fifth step, transform features help you in making the machine learning model more efficient.
- In the sixth step, answers are derived from the data model to help you answer the business questions.
- In the seventh step, this answer is used to implement in production or ML algorithm.

Knowledge Check

Knowledge Check
1

What are the different kinds of data?

A. Data as numbers only

B. Data can only be names that can be changed into numbers

C. Data includes names, numbers, and names that can be turned into numbers. But, it excludes names that look like numbers

D. Data includes names, numbers, names that can be turned into numbers, and names that look like numbers

Knowledge Check
1

What are the different kinds of data?

A. Data as numbers only

B. Data can only be names that can be changed into numbers

C. Data includes names, numbers, and names that can be turned into numbers. But, it excludes names that look like numbers

D. Data includes names, numbers, names that can be turned into numbers, and names that look like numbers

The correct answer is **D**

Data can be names, numbers, names that look like numbers, and names that can be turned into numbers

Knowledge Check

2

What are the different ways to ensure data quality?

A. Data quality is due to business unit malfunction or due to providing incomplete data

B. Data quality can be handled through communicating with business unit(s), handling missing numbers, removing outliers, plotting the values in a column, and fitting to a distribution

C. Once missing values in a column are removed, every column has value/observations and data quality reaches close to 100%

D. Data quality is the job of data analysts and Database Administrators (DBA)

Knowledge Check

2

What are the different ways to ensure data quality?

A. Data quality is due to business unit malfunction or due to providing incomplete data

B. Data quality can be handled through communicating with business unit(s), handling missing numbers, removing outliers, plotting the values in a column, and fitting to a distribution

C. Once missing values in a column are removed, every column has value/observations and data quality reaches close to 100%

D. Data quality is the job of data analysts and Database Administrators (DBA)

The correct answer is **B**

Data can be made consistent by handling missing numbers, plotting the column values, fitting them to distributions, and removing outliers.