

Data science Bootcamp

Presenter:
Nora AlNashwan
@xnorax

Mentors:
Hissah Almuneef
Mead Alrshood
Reema Almashari



IBM History



Transform - CAMSS

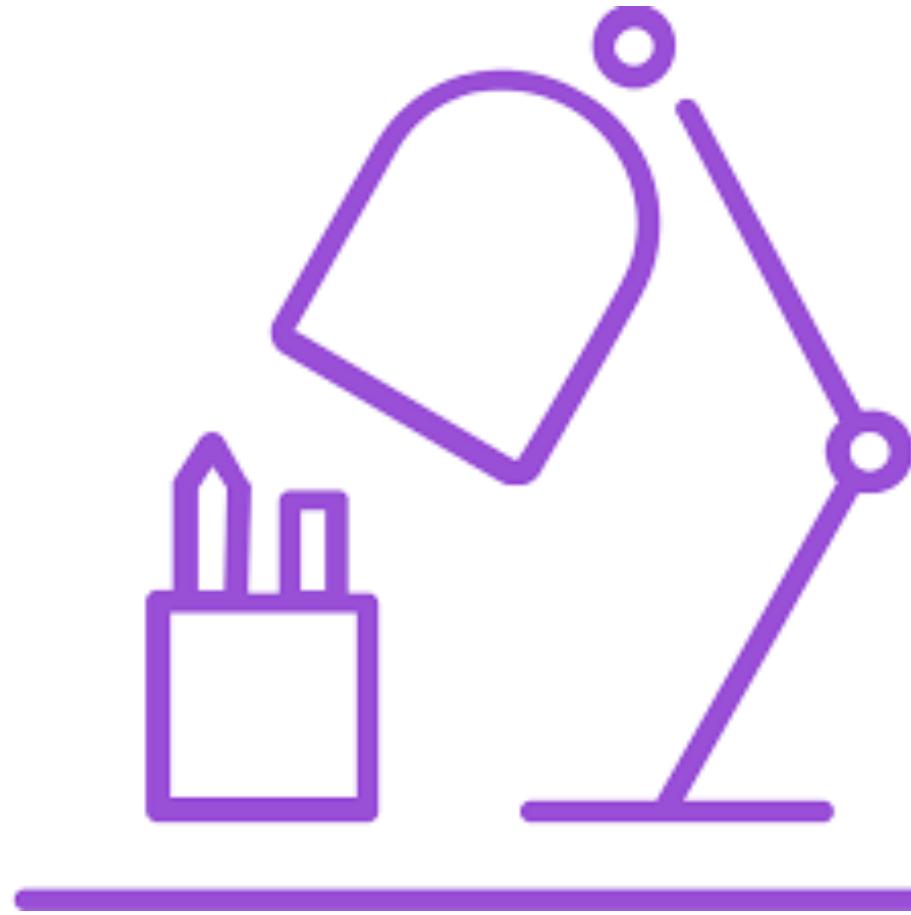


IBM Cloud



IBM cloud

Watson Studio



Data Science

Introduction

IBM
CODE
—

Find Needle In Haystack



Questions Data Science Answers



Is this dress or bag?



Is this weird?

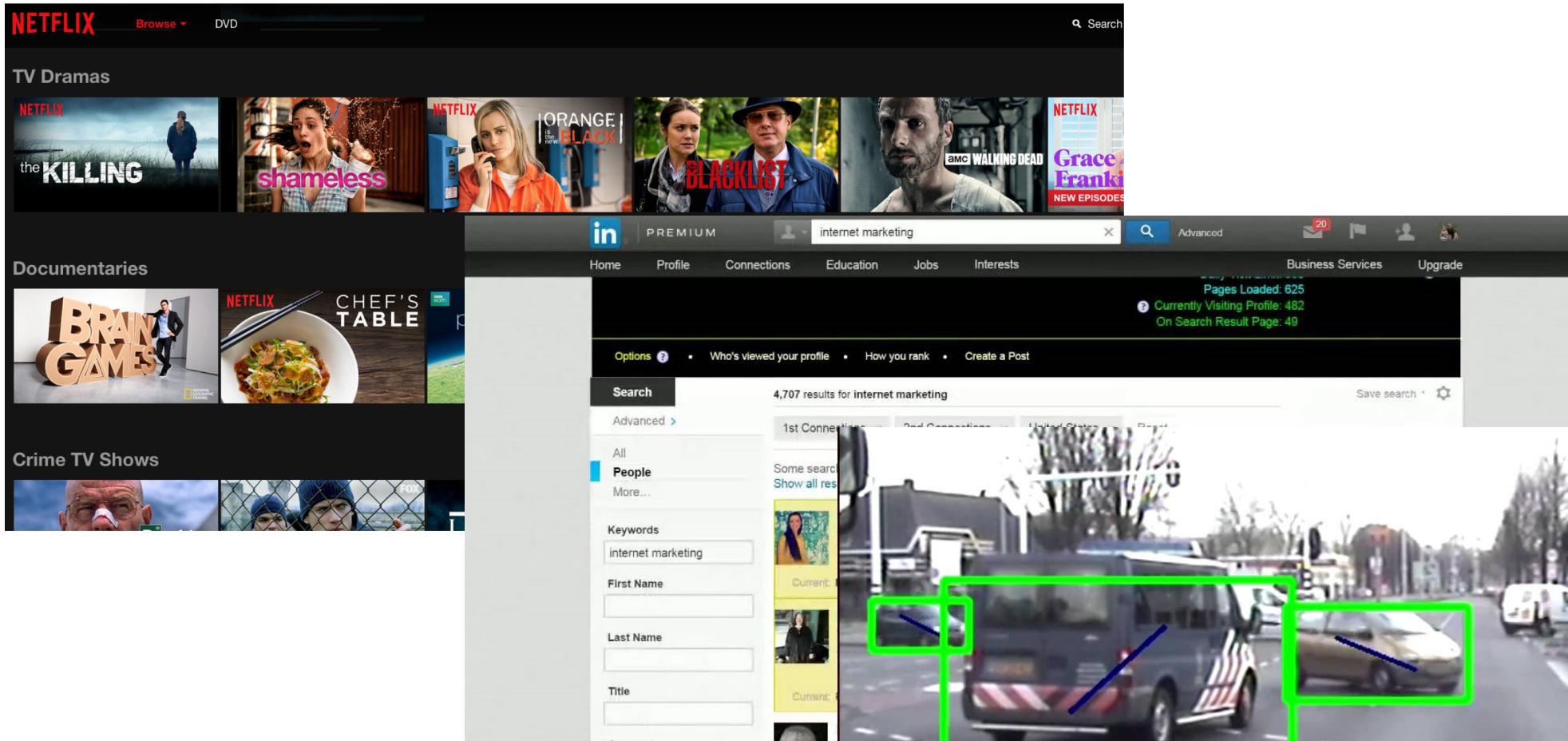


How much?



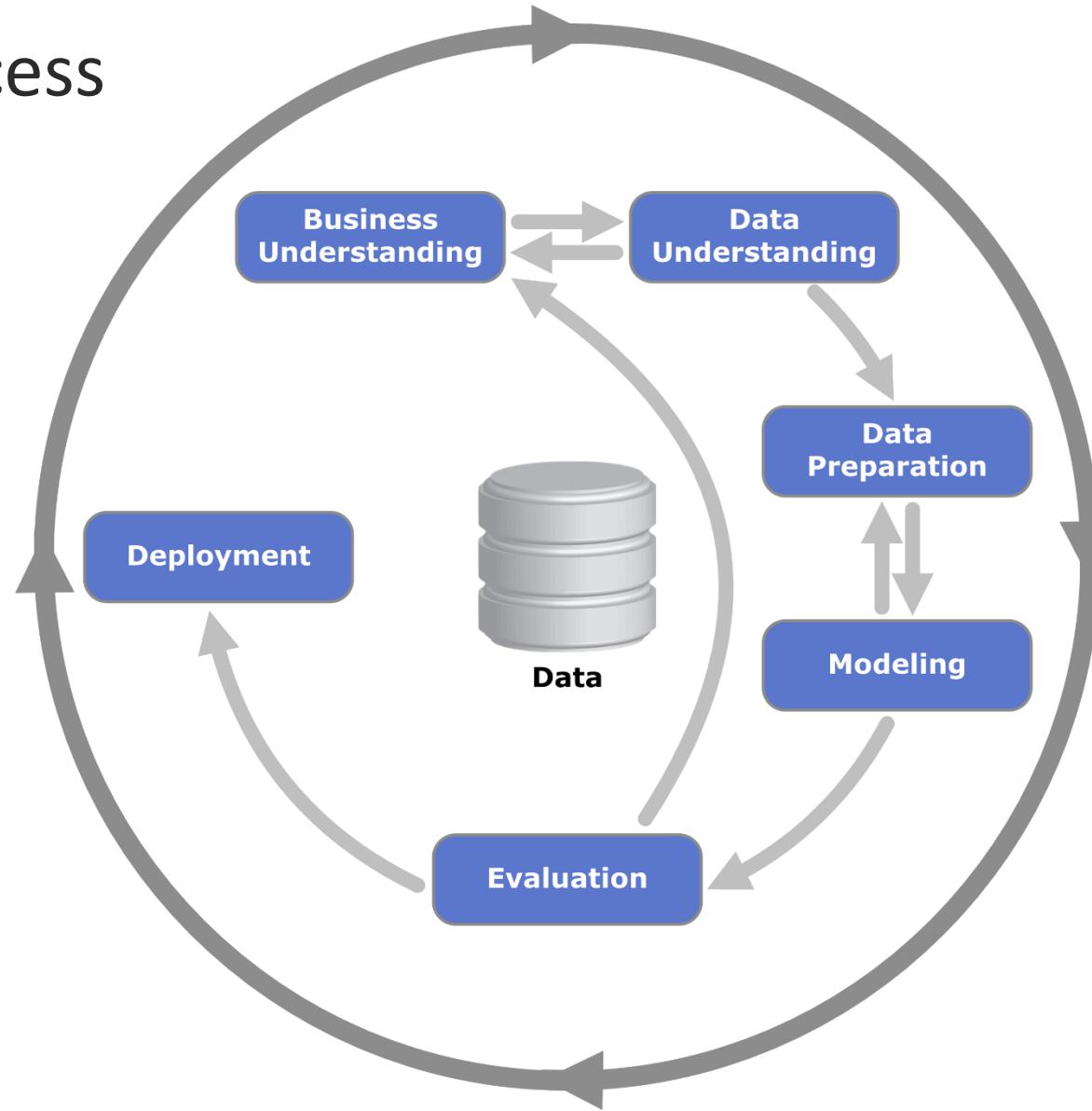
How is it segmented?

Real Applications



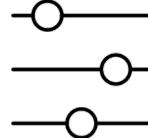
Do you know others?

Data Science Process



Data Science Tools

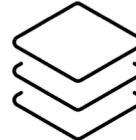
Filtering



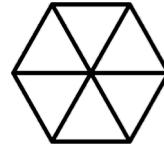
Visualization



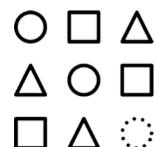
Preparing
the Data



Integration



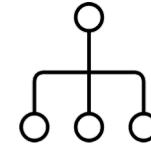
Analytics



Storage



Deploying the
ML Model

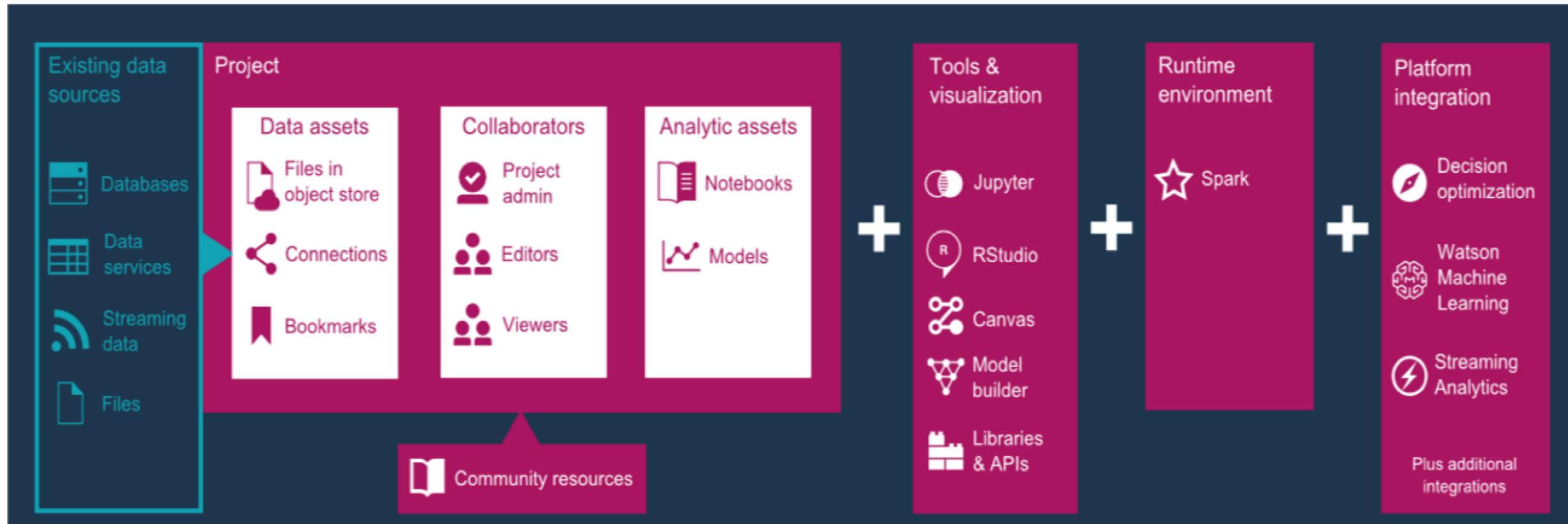


Watson Studio

Exploration

IBM
CODE
—

Watson Studio



Register on IBM Cloud

<https://ibm.biz/BdY4nQ>

IBM Cloud

Sign up for an IBMid and create your IBM Cloud account

Build on IBM Cloud for free with no time restrictions

Guaranteed free development with Lite plans

Develop worry-free and at no cost with cap based Lite plan services for as long as you like.

Start on your projects right away

Skip entering your credit card info and get working in just a few short steps.

Get a \$200 credit when you upgrade

After you upgrade to a Pay-As-You-Go account, you can use the credit to try new services or scale your projects. The credit is valid for 1 month and can be used with any of our IBM Cloud offerings.

Ready to get started? Sign up today!

IBM

CODE

Already have an IBM Cloud account? [Log in](#)

Email*



First Name*

Last Name*

Country or Region*

Choose [US](#) as a country

Password*



Search for Watson Studio Service

IBM Cloud Catalog Docs Support Manage Nora Sami's Account

Catalog

watson studio

All Categories (2) >

- Compute
- Containers
- Networking
- Storage
- AI (1)
- Analytics (1)
- Databases
- Developer Tools
- Integration
- Internet of Things
- Security and Identity
- Starter Kits

AI

Watson Studio
Lite • IBM

Embed AI and machine learning into your business. Create custom models using your own data.

Analytics

Analytics Engine
Lite • IBM

Flexible framework to deploy Hadoop and Spark analytics applications.

FEEDBACK

Create Instance

IBM Cloud Catalog Docs Support Manage

Nora Sami's Account

View all

Watson Studio

Lite • IBM

Watson Studio democratizes machine learning and deep learning to accelerate infusion of AI in your business to drive innovation. Watson Studio provides a suite of tools and a collaborative environment for data scientists, developers and domain experts.

Service name: Watson Studio-1q

Choose a region/location to deploy in: United Kingdom

Select a resource group: Default

View Docs Terms

AUTHOR IBM

PUBLISHED 08/01/2018

Features

- Use what you know. Learn what you don't
- Power on demand

Need Help? Contact IBM Cloud Support ↗

Estimate Monthly Cost
[Cost Calculator](#)

Create

Get Started



Watson Studio

Welcome to Watson Studio. Let's get started!

[Get Started](#)

Watson Studio Dashboard

IBM Watson Projects Tools Community Services Docs Support Manage Get sta

Welcome Jen!

Watson Studio is part of IBM Watson.

Try out other IBM Watson apps.

Get started with key tasks

New project Refine data New notebook Deep learning New Modeler flow New model

Recently updated projects View all (2) + New project

NAME	ROLE	COLLABORATORS	DATE CREATED	LAST UPDATED
------	------	---------------	--------------	--------------

The dashboard features a top navigation bar with links for IBM Watson, Projects, Tools, Community, Services, Docs, Support, Manage, and a notification bell. A large banner on the left says 'Welcome Jen!' and 'Watson Studio is part of IBM Watson.' Below it, there's a section titled 'Get started with key tasks' containing six cards: 'New project' (highlighted with a blue border), 'Refine data', 'New notebook', 'Deep learning', 'New Modeler flow', and 'New model'. At the bottom, there's a section for 'Recently updated projects' with a 'View all (2)' link and a '+ New project' button. A table header is shown below this section.

Create Project

IBM Watson Data Platform | Projects Tools Catalog Data Services Community US South

New project

Define project details

Name
Test Project 88

Description
Project description 3000

Choose project options

Restrict who can be a collaborator

Add a compute engine for data analysis

Define storage

Select storage type

Object Storage (Swift API) IBM Cloud Object Storage

Target Cloud Object Storage Instance
`cloud-object-storage-el`

Define compute engine

Select Spark service

`Spark™-dr`

If you associate the same Spark service with multiple projects, the Spark history server will display job history information for all the projects.

Cancel Create

Overview Page

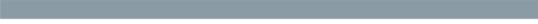
My Projects / my-new-project

Add to project     

[Overview](#) [Assets](#) [Environments](#) [Bookmarks](#) [Deployments](#) [Collaborators](#) [Settings](#)

my-new-project
Last Updated: May 03 2018

0 Assets **0** Bookmarks **1** Collaborators

Date created May 03 2018	Description No description available	Recent activity  Alerts related to this project will show here when the project is active.
Storage 0% of 25 GB used 		
Collaborators  Steve Martinelli Admin	View all (1)	
Bookmarks 	View all (0)	

Assets Page

The screenshot shows the IBM Watson Assets Page. At the top, there is a navigation bar with links for Projects, Tools, Community, Services, Docs, Support, Manage, and a notifications icon. Below the navigation bar, the page title is "My Projects / demo". The main content area has a header with tabs: Overview, Assets (which is highlighted with a blue box), Environments, Bookmarks, Deployments, Collaborators, and Settings. A search bar below the tabs contains the placeholder text "What assets are you looking for?".

The page is organized into sections:

- Data assets**: Shows 0 assets selected. A table lists one asset: "Customer demographics and sales.csv" (Data Asset, Project, nailah altayyar, 24 Apr 2018, 4:56:26 pm). A "New data asset" button is available.
- Models**: Shows 0 models selected. A table header includes columns for NAME, STATUS, TYPE, RUNTIME, LAST MODIFIED (with a dropdown arrow), and ACTIONS. A message indicates "you currently have no models". A "New model" button is available.
- Experiments**: Shows 0 experiments selected. A table header includes columns for NAME, CREATED BY, LAST MODIFIED (with a dropdown arrow), and ACTIONS. A message indicates "you currently have no experiments". A "New experiment" button is available.
- Modeler flows**: Shows 1 flow selected: "Sales Promotion Study" (SPSS, nailah altayyar, 2 May 2018, 2:58:40 pm). A table header includes columns for NAME, TYPE, CREATED BY, LAST MODIFIED, and ACTIONS. A "New flow" button is available.
- Notebooks**: Shows 0 notebooks selected. A "New notebook" button is available.

Add Collaborators

demo

Add collaborators

Invite

Enter email address



Access level

Select an option



Add

Viewer

Editor

Admin

Cancel

Invite

Community

The screenshot shows the IBM Watson Community interface. At the top, there is a navigation bar with the following items: a menu icon, the "IBM Watson" logo, "Projects", "Tools", "Community" (which is highlighted with a blue box), and "Services". On the right side of the navigation bar are links for "Docs", "Support", and "Marketplace". Below the navigation bar, there is a search bar with the placeholder text "What are you looking for?" and a "All filters" button. Underneath the search bar, there are "Popular filters:" buttons for "Spark", "Deep Learning", and "Brunel".

Featured

Sort by: Featured

TUTORIAL Detecting Whisky brands with Core ML and IBM Watson <small>AUTHOR: Martin Mitevski DATE: Apr 25, 2018 LEVEL: Beginner TOPIC: Watson</small> <small>3 hearts, 1 bookmark</small>	TUTORIAL Build Deep Learning Architectures With... <small>AUTHOR: developerWorks TV DATE: Apr 02, 2018 LEVEL: Beginner TOPIC: Deep Learning +2</small> <small>9 hearts, 1 bookmark</small>	ARTICLE Introducing IBM Watson Studio <small>AUTHOR: Armand Ruiz DATE: Mar 20, 2018 TOPIC: Watson FORMAT: Web page</small> <small>12 hearts, 1 bookmark</small>	ARTICLE Apple, IBM add machine learning to... <small>AUTHOR: TechCrunch DATE: Mar 20, 2018 TOPIC: Watson FORMAT: Web page</small> <small>5 hearts, 1 bookmark</small>
--	--	---	---

All content

ARTICLE We Analyzed 1 Million Jupyter Notebooks —... <small>AUTHOR: Jupyter DATE: May 03, 2018 TOPIC: Notebook FORMAT: Web page</small> <small>0 hearts, 0 bookmarks</small>	ARTICLE Towards fairness in ML with adversarial... <small>AUTHOR: GoDataDrivenBlog DATE: May 02, 2018 TOPIC: Machine Learning FORMAT: Web page</small> <small>0 hearts, 0 bookmarks</small>	ARTICLE Using Machine Learning to Predict Outcomes... <small>AUTHOR: Inside Machine Learning DATE: May 01, 2018 TOPIC: Machine Learning FORMAT: Web page</small> <small>0 hearts, 0 bookmarks</small>	ARTICLE Creating multi-source and multi-target data... <small>AUTHOR: Wesley Williams DATE: Apr 30, 2018 TOPIC: Data Shaping +1 FORMAT: Web page</small> <small>0 hearts, 0 bookmarks</small>
ARTICLE Predict customer churn by building,... <small>AUTHOR DATE</small>	NOTEBOOK Create a multi source & target data flow <small>AUTHOR DATE</small>	NOTEBOOK Use Spark SQL to explore heating problems in... <small>AUTHOR DATE</small>	NOTEBOOK Use XGBoost to classify tumors <small>AUTHOR DATE</small>

Watson Studio

M&Ms activity

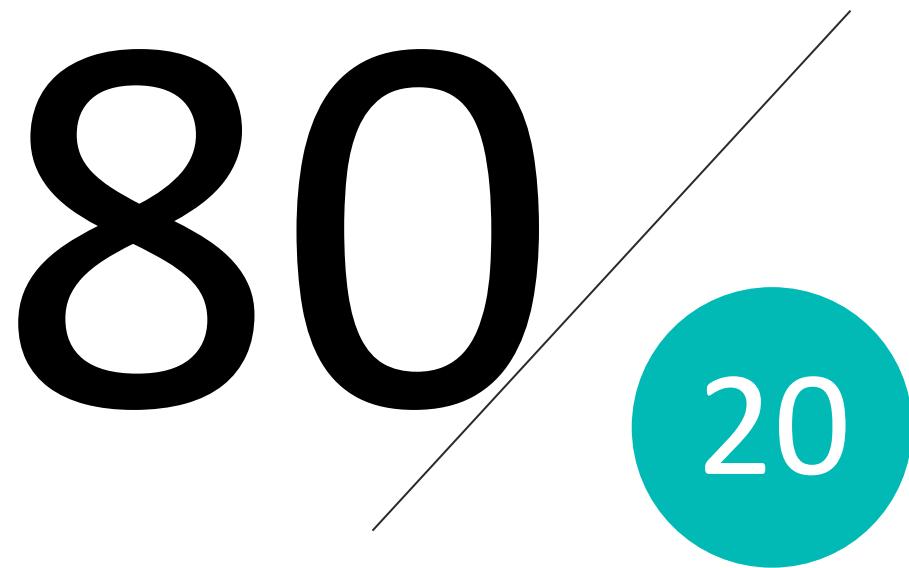
IBM
CODE
—

Data Exploration, Visualization, and Feature Engineering

Exploratory data analysis



Data cleaning, data exploration, feature engineering, pre-processing, etc...



Model building

Date Exploration Steps

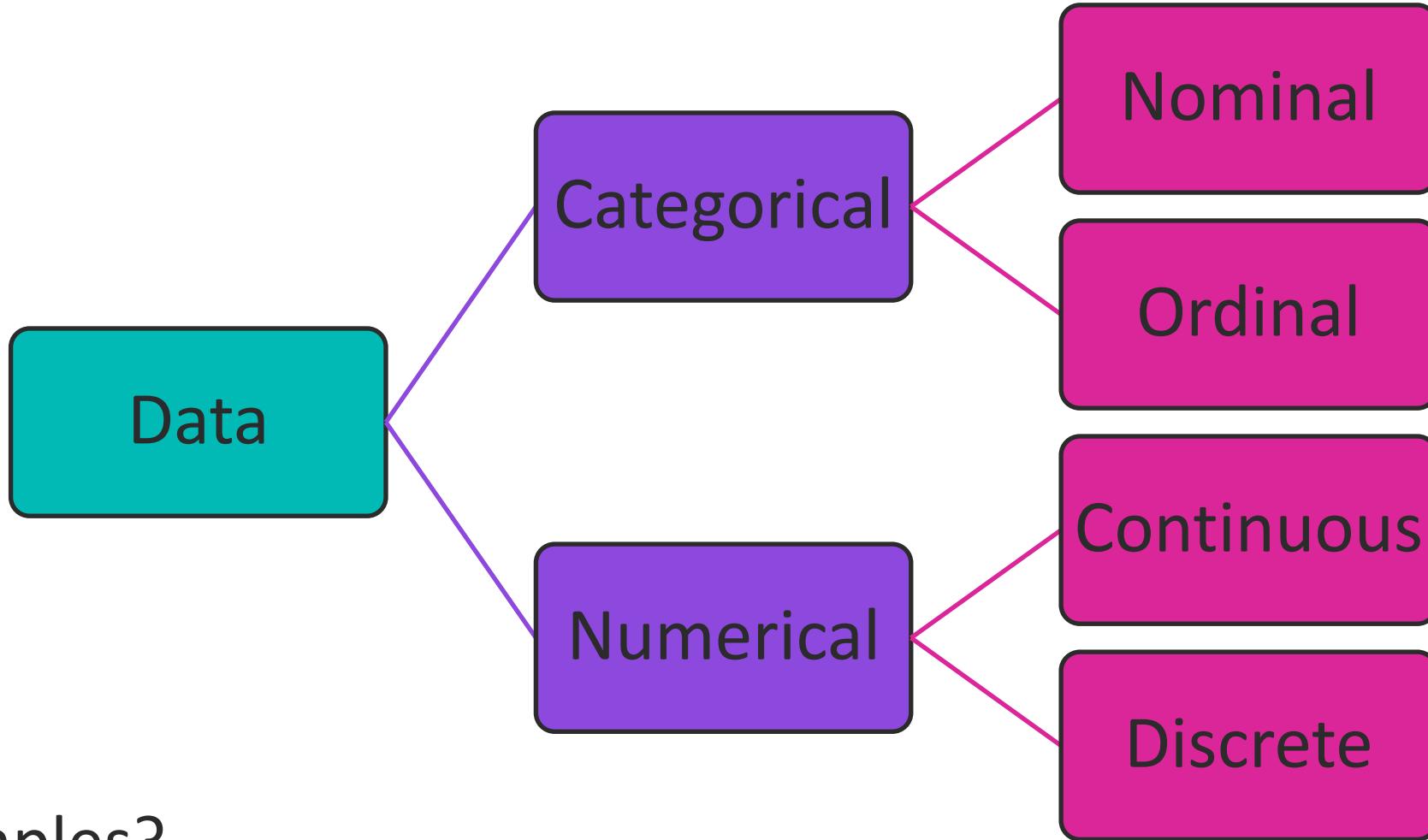
1. Identify Data Types
2. Explore Each Variable
3. Find Correlation
4. Missing Values Treatment
5. Outlier Treatment
6. Feature Engineering

Step 1

Identify data types

IBM
CODE

Data Types



Other examples?

Identify data types

JobID	Agency	PostingType	#OfPositions	Categorical				Categorical				AdditionalInformation	ToApply	Hours	WorkLocation1
				BusinessTitle	CivilServiceTitle	TitleCodeNo	Level	JobCategory	FullOrPartTime	...					
0 87990	DEPARTMENT OF BUSINESS SERV.	Internal	1	Account Manager	CONTRACT REVIEWER (OFFICE OF L	40563	1			...	Salary range for this position is: 42,405–...				
1 97899	DEPARTMENT OF BUSINESS SERV.	Internal	1	EXECUTIVE DIRECTOR, BUSINESS DEVELOPMENT	ADMINISTRATIVE BUSINESS PROMOT	10009	M3			F ...		In addition to applying through this website, ...			
2 102221	DEPT OF ENVIRONMENT PROTECTION	External	1	Project Specialist	ENVIRONMENTAL ENGINEERING INTE	20616	0			F ...	Appointments are subject to OMB approval	click the apply now button	35 hours per week/day		
3 102221	DEPT OF ENVIRONMENT PROTECTION	Internal	1	Project Specialist	ENVIRONMENTAL ENGINEERING INTE	20616	0			F ...	Appointments are subject to OMB approval	click the apply now button	35 hours per week/day		

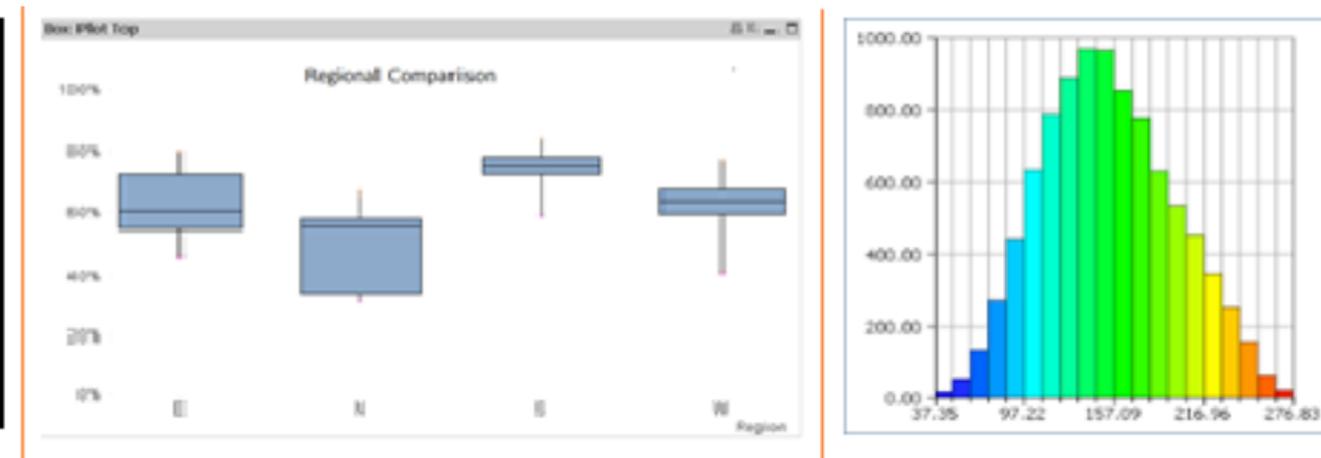
Step 2

Explore variables one by one

IBM
CODE

Numerical Variable

Central Tendency	Measure of Dispersion	Visualization Methods
Mean	Range	Histogram
Median	Quartile	Box Plot
Mode	IQR	
Min	Variance	
Max	Standard Deviation	
	Skewness and Kurtosis	



A. Central Tendency

- **Mode:** most frequent measure
- **Median:** mid-point of an array of measures
- **Mean:** arithmetic average (Sum/N)

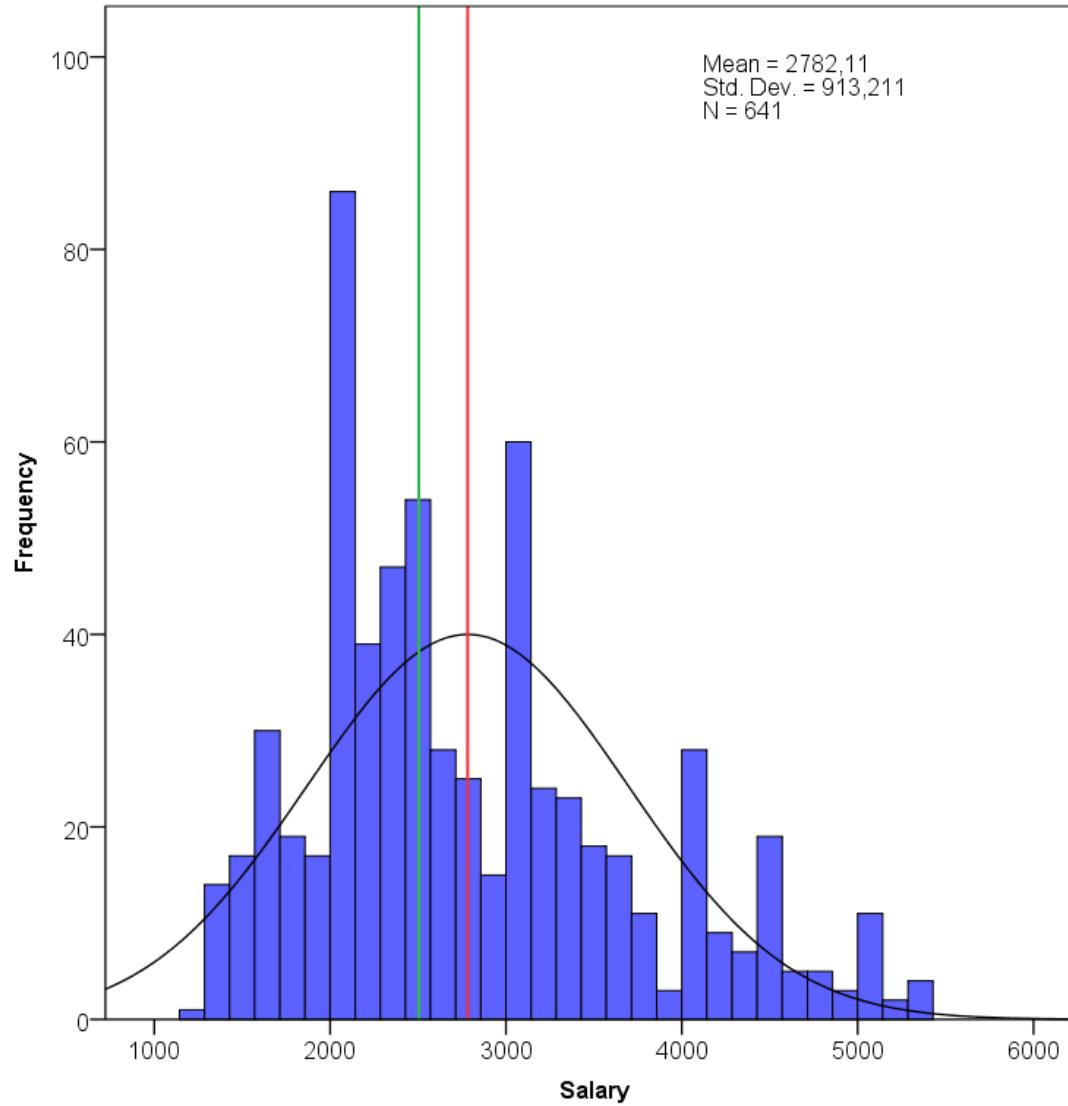
B. Variance and Standard Deviation

Standard deviation is the square root of the Variance

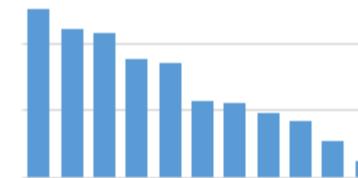
$$SD = \sqrt{\frac{\sum |x - \bar{x}|^2}{n}}$$

x1	1	1	1
x2	1	2	2
x3	1	2	3
std	0	0.57735027	1

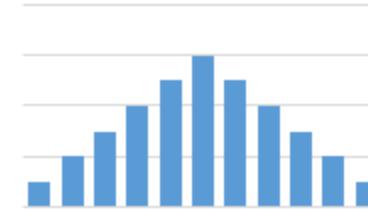
C. Histogram & Boxplot



Positive Skew



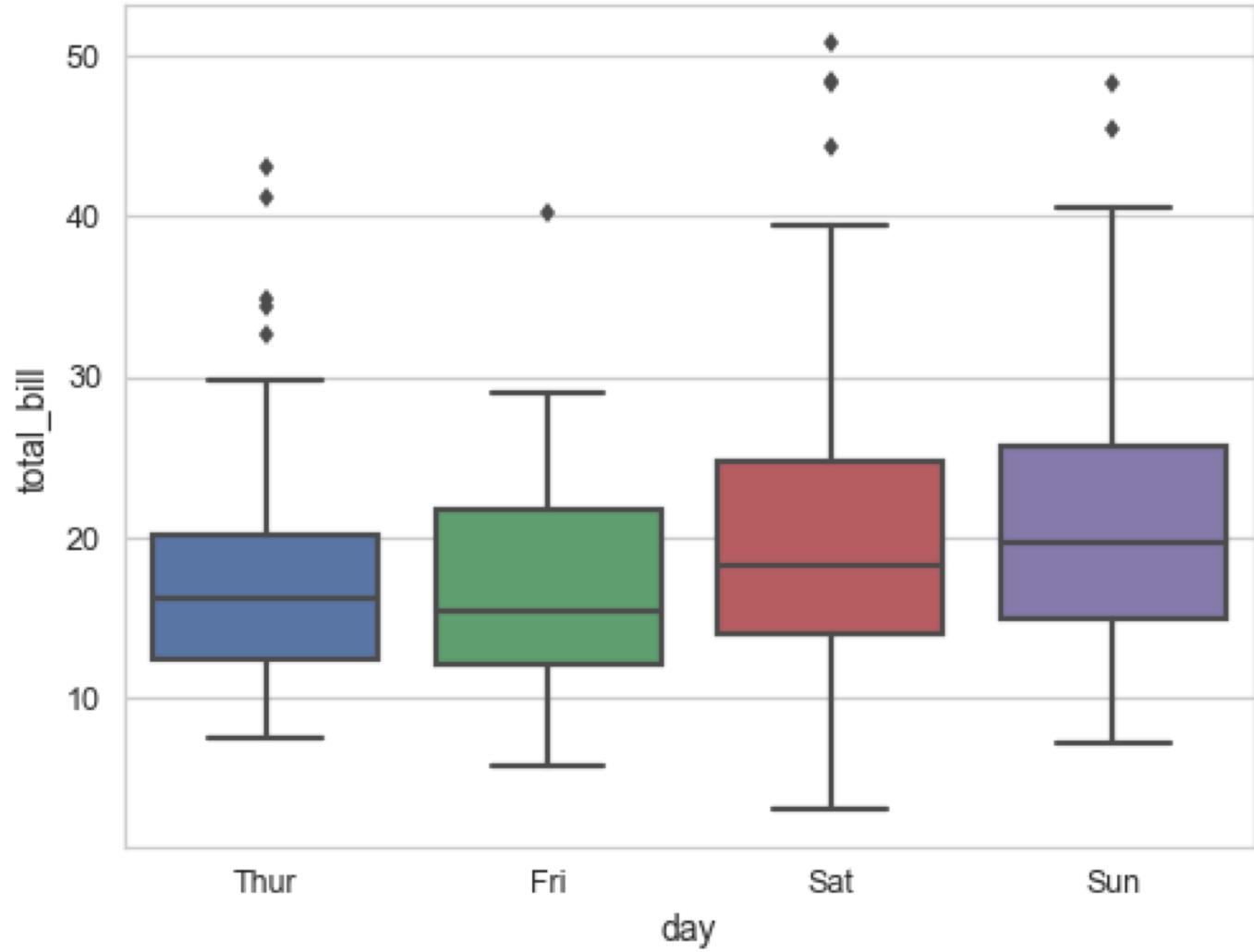
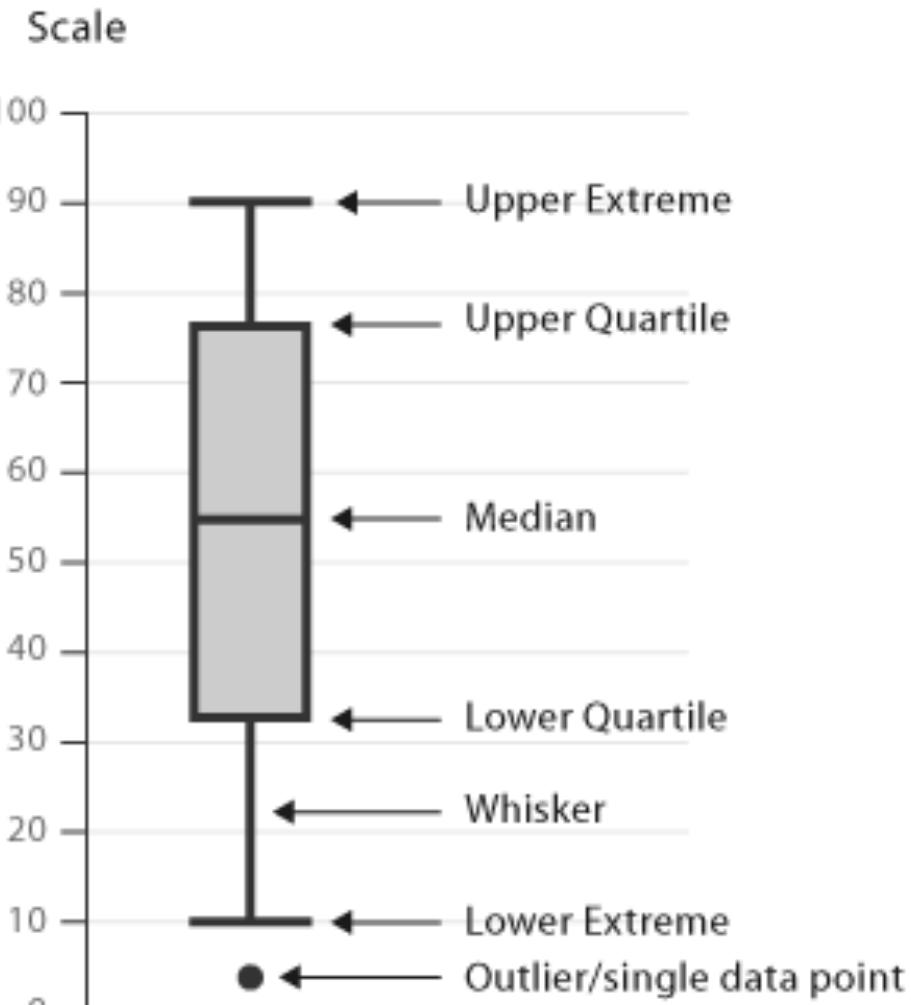
Normal Distribution



Negative Skew



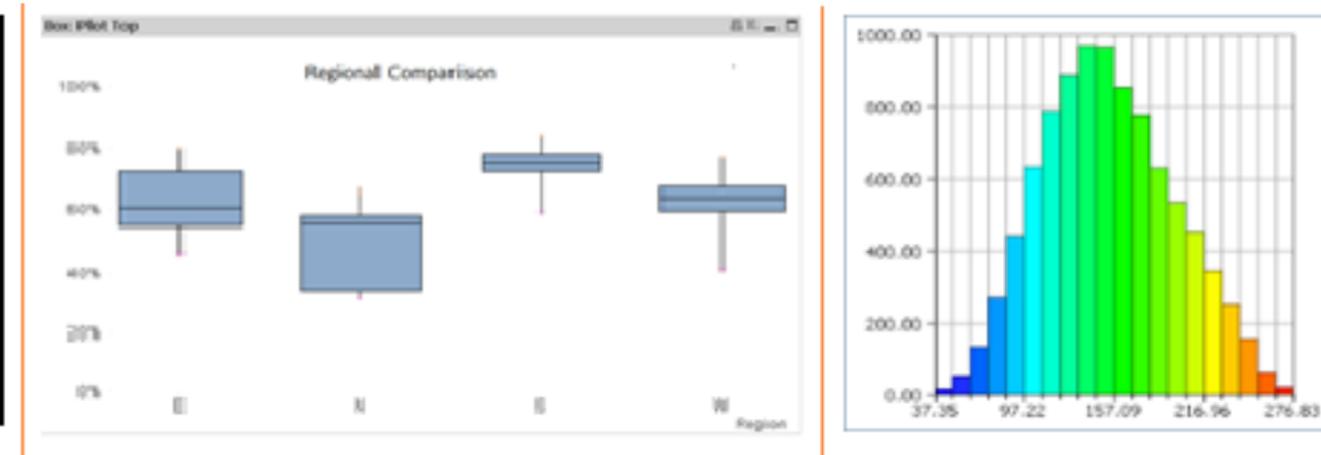
C. Histogram & Boxplot



CODE

Do you think this is enough?

Central Tendency	Measure of Dispersion	Visualization Methods
Mean	Range	Histogram
Median	Quartile	Box Plot
Mode	IQR	
Min	Variance	
Max	Standard Deviation	
	Skewness and Kurtosis	



Anscombe's quartet

X	Y	X	Y
10.00	8.04	10.00	9.14
8.00	6.95	8.00	8.14
13.00	7.58	13.00	8.74
9.00	8.81	9.00	8.77
11.00	8.33	11.00	9.26
14.00	9.96	14.00	8.10
6.00	7.24	6.00	6.13
4.00	4.26	4.00	3.10
12.00	10.84	12.00	9.13
7.00	4.82	7.00	7.26
5.00	5.68	5.00	4.74

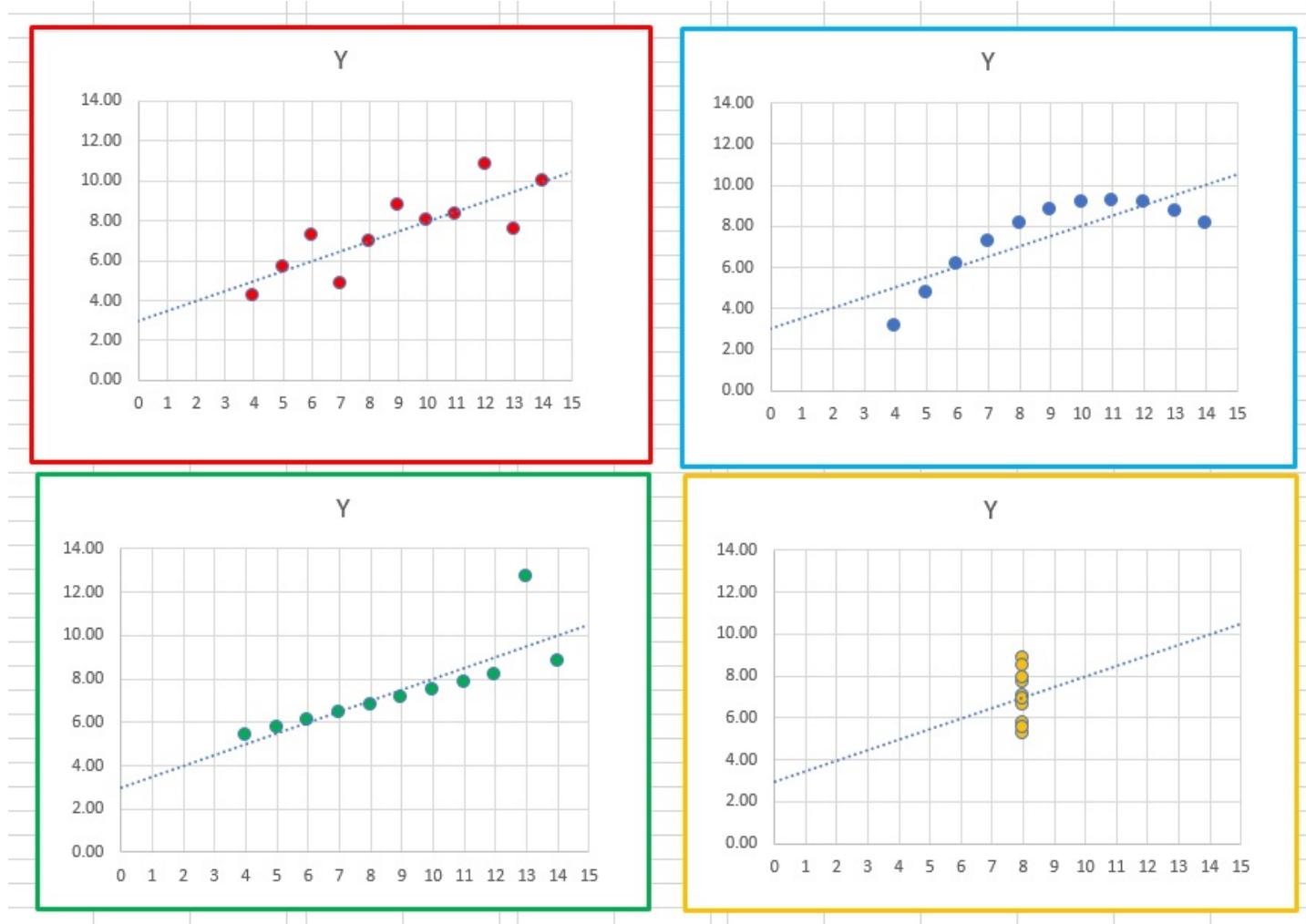
X	Y	X	Y
10.00	7.46	8.00	6.58
8.00	6.77	8.00	5.76
13.00	12.74	8.00	7.71
9.00	7.11	8.00	8.84
11.00	7.81	8.00	8.47
14.00	8.84	8.00	7.04
6.00	6.08	8.00	5.25
4.00	5.39	19.00	12.50
12.00	8.15	8.00	5.56
7.00	6.42	8.00	7.91
5.00	5.73	8.00	6.89

	XMean	YMean	XVar	YVar
Red	9.00	7.50	11.00	4.125
Blue	9.00	7.50	11.00	4.125
Green	9.00	7.50	11.00	4.125
Orange	9.00	7.50	11.00	4.125

Anscombe's quartet

X	Y	X	Y
10.00	8.04	10.00	9.14
8.00	6.95	8.00	8.14
13.00	7.58	13.00	8.74
9.00	8.81	9.00	8.77
11.00	8.33	11.00	9.26
14.00	9.96	14.00	8.10
6.00	7.24	6.00	6.13
4.00	4.26	4.00	3.10
12.00	10.84	12.00	9.13
7.00	4.82	7.00	7.26
5.00	5.68	5.00	4.74

X	Y	X	Y
10.00	7.46	8.00	6.58
8.00	6.77	8.00	5.76
13.00	12.74	8.00	7.71
9.00	7.11	8.00	8.84
11.00	7.81	8.00	8.47
14.00	8.84	8.00	7.04
6.00	6.08	8.00	5.25
4.00	5.39	19.00	12.50
12.00	8.15	8.00	5.56
7.00	6.42	8.00	7.91
5.00	5.73	8.00	6.89



Anscombe's quartet demonstrates both the importance of **graphing data** before analyzing it and the effect of **outliers** on statistical properties.

Step 3

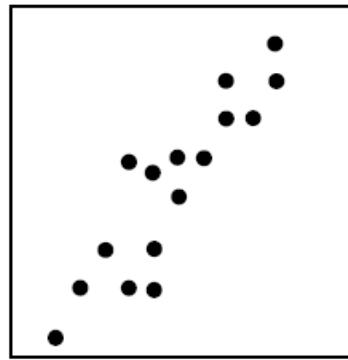
Find association and disassociation between variables

IBM
CODE

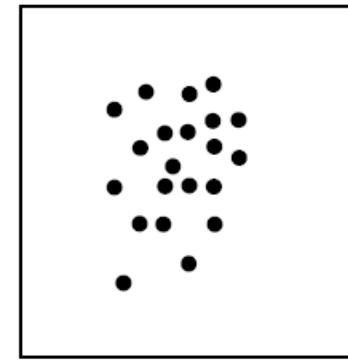
Continuous Variables



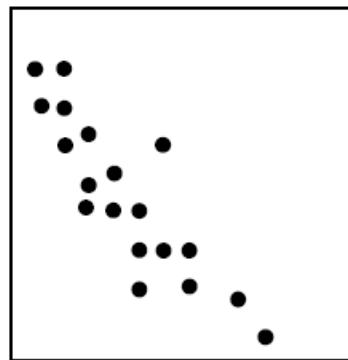
Strong positive correlation



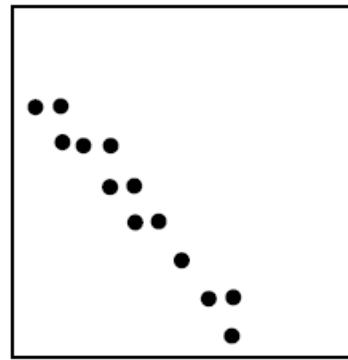
Moderate positive correlation



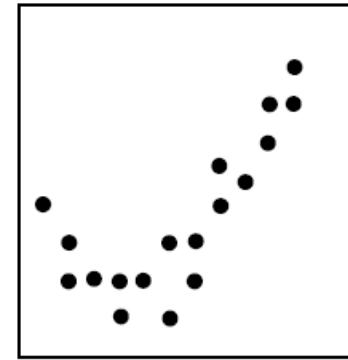
No correlation



Moderate negative correlation



Strong negative correlation



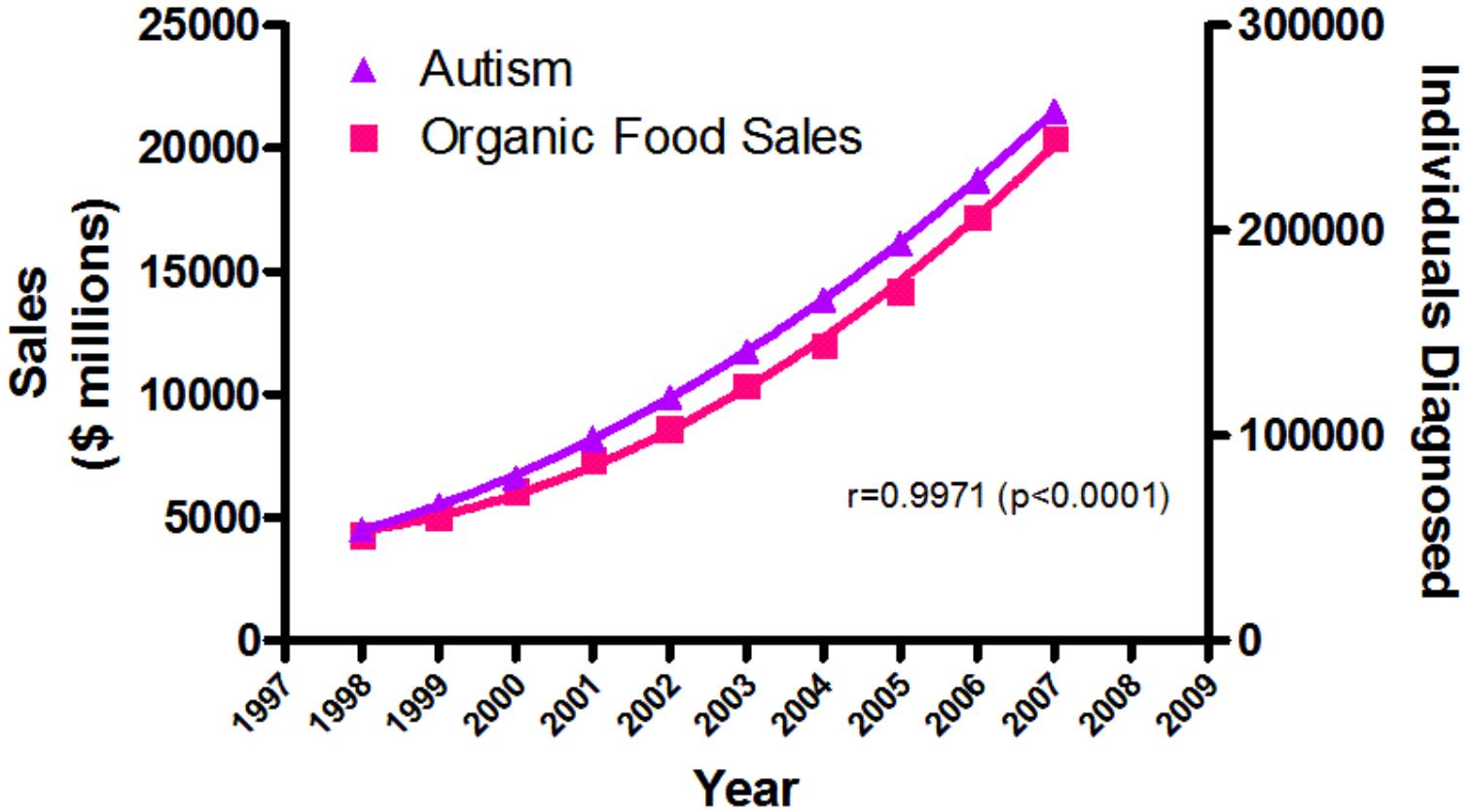
Curvilinear relationship

Categorical Variables

**Whole Table
Relative Frequencies -
Divide all cells by 240.**

MathBits.com	Sport Utility Vehicle (SUV)	Sports Car	Totals
male	$\frac{21}{240} = 0.09$	$\frac{39}{240} = 0.16$	$\frac{60}{240} = 0.25$
female	$\frac{135}{240} = 0.56$	$\frac{45}{240} = 0.19$	$\frac{180}{240} = 0.75$
Totals	$\frac{156}{240} = 0.65$	$\frac{84}{240} = 0.35$	$\frac{240}{240} = 1.00$

Correlation Does Not Imply Causation



Sources: Organic Trade Association, 2011 Organic Industry Survey; U.S. Department of Education, Office of Special Education Programs, Data Analysis System (DANS), OMB# 1820-0043: "Children with Disabilities Receiving Special Education Under Part B of the Individuals with Disabilities Education Act"

Step 4

Date cleaning



Missing Data Impact

Name	Weight	Gender	Play Cricket/ Not
Mr. Amit	58	M	Y
Mr. Anil	61	M	Y
Miss Swati	58	F	N
Miss Richa	55		Y
Mr. Steve	55	M	N
Miss Reena	64	F	Y
Miss Rashmi	57		Y
Mr. Kunal	57	M	N

Gender	#Students	#Play Cricket	%Play Cricket
F	2	1	50%
M	4	2	50%
Missing	2	2	100%

Name	Weight	Gender	Play Cricket/ Not
Mr. Amit	58	M	Y
Mr. Anil	61	M	Y
Miss Swati	58	F	N
Miss Richa	55	F	Y
Mr. Steve	55	M	N
Miss Reena	64	F	Y
Miss Rashmi	57	F	Y
Mr. Kunal	57	M	N

Gender	#Students	#Play Cricket	%Play Cricket
F	4	3	75%
M	4	2	50%

How to treat missing values?

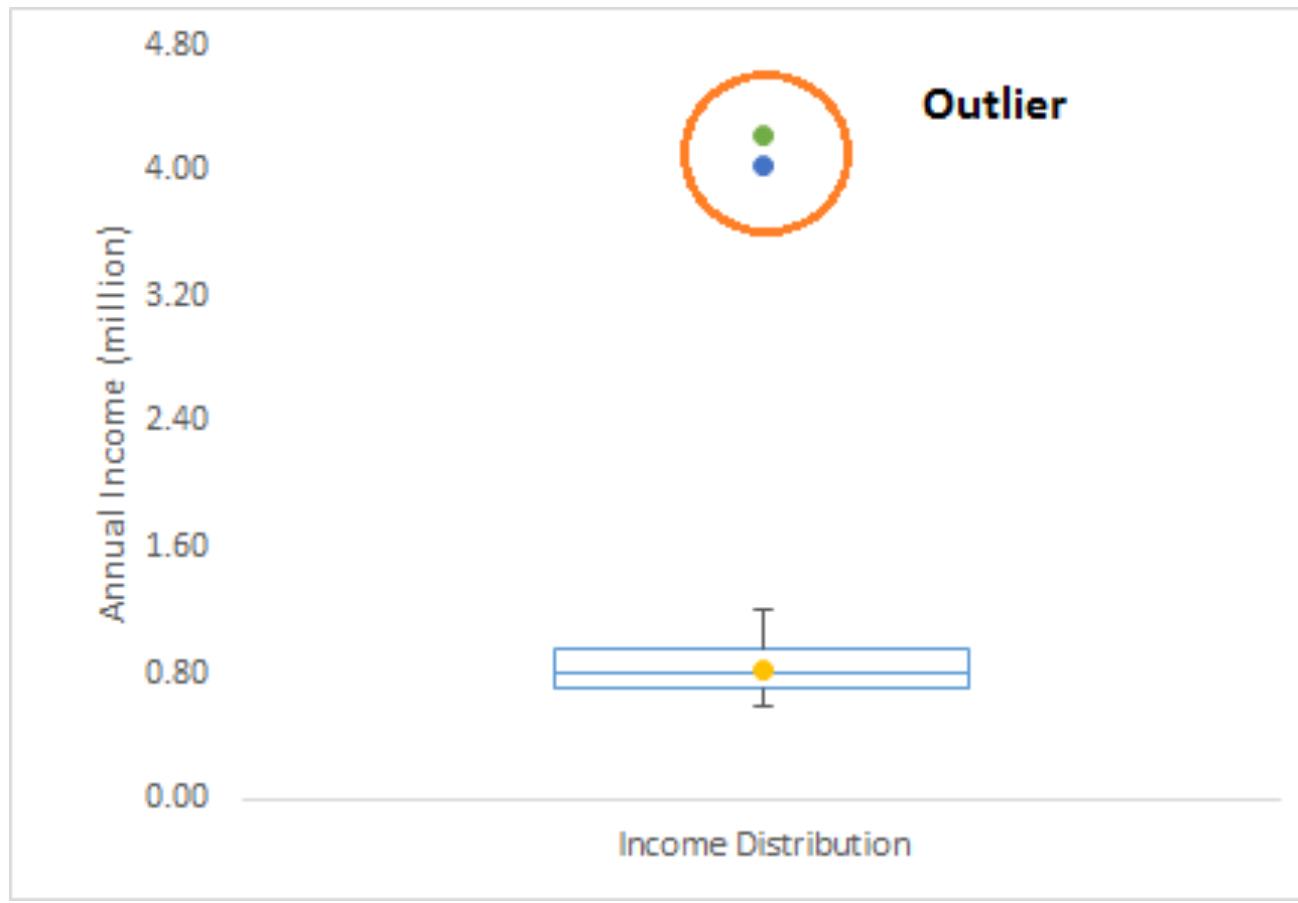
1. Deletion
2. Mean/Mode/Median Imputation
3. Prediction Model

Step 5

Outlier detection and treatment

IBM
CODE

Outliers



Name	Salary
Tom	5,000
Susan	7,000
Sam	12,000
Ahmed	15,000
Bob	19,000
Bill Gates	1,000,000,000
Mean without Bill salary	11,600
Mean with Bill salary	200,011,600

How to deal with outliers?

1. Deletion
2. Transformation
3. Imputing

Step 6

Feature engineering



Feature Engineering

Extracting more information from existing data.



Date
7/8/18
8/8/18
9/8/18
10/8/18
11/8/18
12/8/18
1/8/19
2/8/19
3/8/19



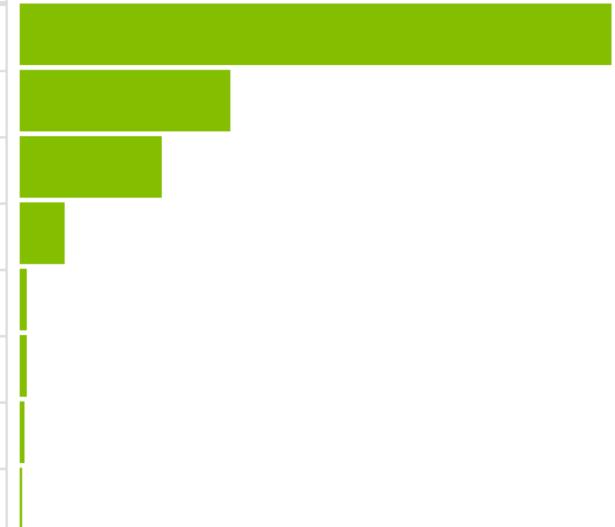
Day
Tue
Wed
Thu
Fri
Sat
Sun
Mon
Tue
Wed

Why Feature Engineering Is Important?



Most frequent items from `<SArray>`

Value	Value	Count	Percent
Mr. Mohamad	Mr.	517	58.025%
Miss. Nora	Miss.	185	20.763%
Mrs. Rasha	Mrs.	125	14.029%
Master. Ahmd	Master.	40	4.489%
Dr. Saad	Dr.	7	0.786%
Rev. Joe	Rev.	6	0.673%
Sir. Saed	Sir.	5	0.561%
Col. Eead	Col.	2	0.224%
Jonkheer. Slm	Jonkheer.	1	0.112%
Lady. Sara	Lady.	1	0.112%
the Countess. S	the Countess.	1	0.112%
Ms. Niyarrah	Ms.	1	0.112%



Build a Machine Learning Model

For Classification

IBM
CODE

Model

=

Data

+

Algorithm

Data or Algorithms?

Everyone is talking about
advances in **algorithms**.

But it is and was the data, or
more specifically how we **store**
and **process** the data that was
the single most important factor
in the explosion of data science
over the last decade.

Supervised Learning

outlook	windy	play
sunny	false	no
sunny	true	no
rainy	false	yes
rainy	true	no
sunny	true	yes
rainy	true	no

outlook	windy	play
sunny	false	no
sunny	true	no
rainy	false	yes

Train Model

outlook	windy	play
rainy	true	?
sunny	true	?
rainy	true	?

Predictor

Naive Bayes



outlook	windy	play
sunny	false	no
sunny	true	no
rainy	false	yes
rainy	true	no
sunny	true	yes
rainy	true	no

Probability that we can play the game

$$P(\text{Outlook}=\text{Sunny} \mid \text{Play}=\text{Yes}) = 1/2$$

$$P(\text{Windy}=\text{true} \mid \text{Play}=\text{Yes}) = 1/2$$

$$P(\text{Play}=\text{Yes}) = 2/6$$

Probability that we can not play the game

$$P(\text{Outlook}=\text{Sunny} \mid \text{Play}=\text{No}) = 2/4$$

$$P(\text{Windy}=\text{true} \mid \text{Play}=\text{No}) = 3/4$$

$$P(\text{Play}=\text{No}) = 4/6$$

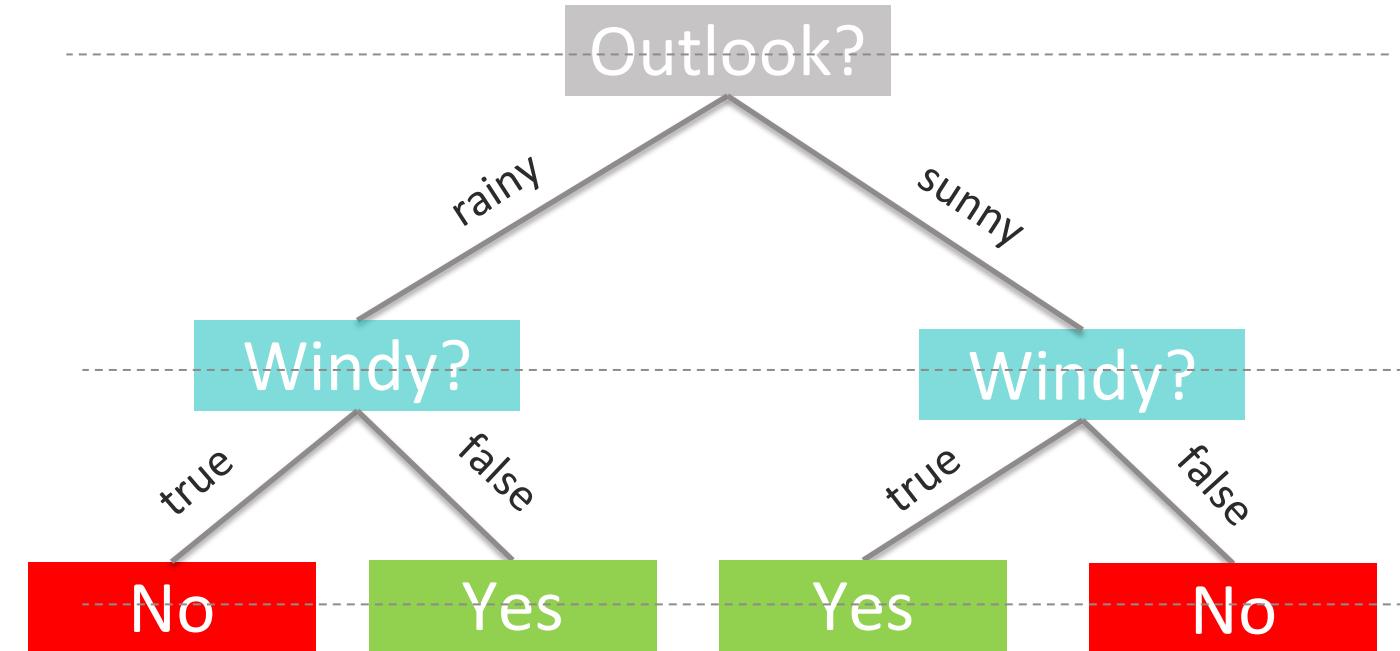
$$P(X \mid \text{Play}=\text{Yes}) = 1/3 * 1/4 * 2/6 = 0.08333$$

$$P(X \mid \text{Play}=\text{No}) = 2/3 * 3/4 * 4/6 = 0.25$$

Classification Tree

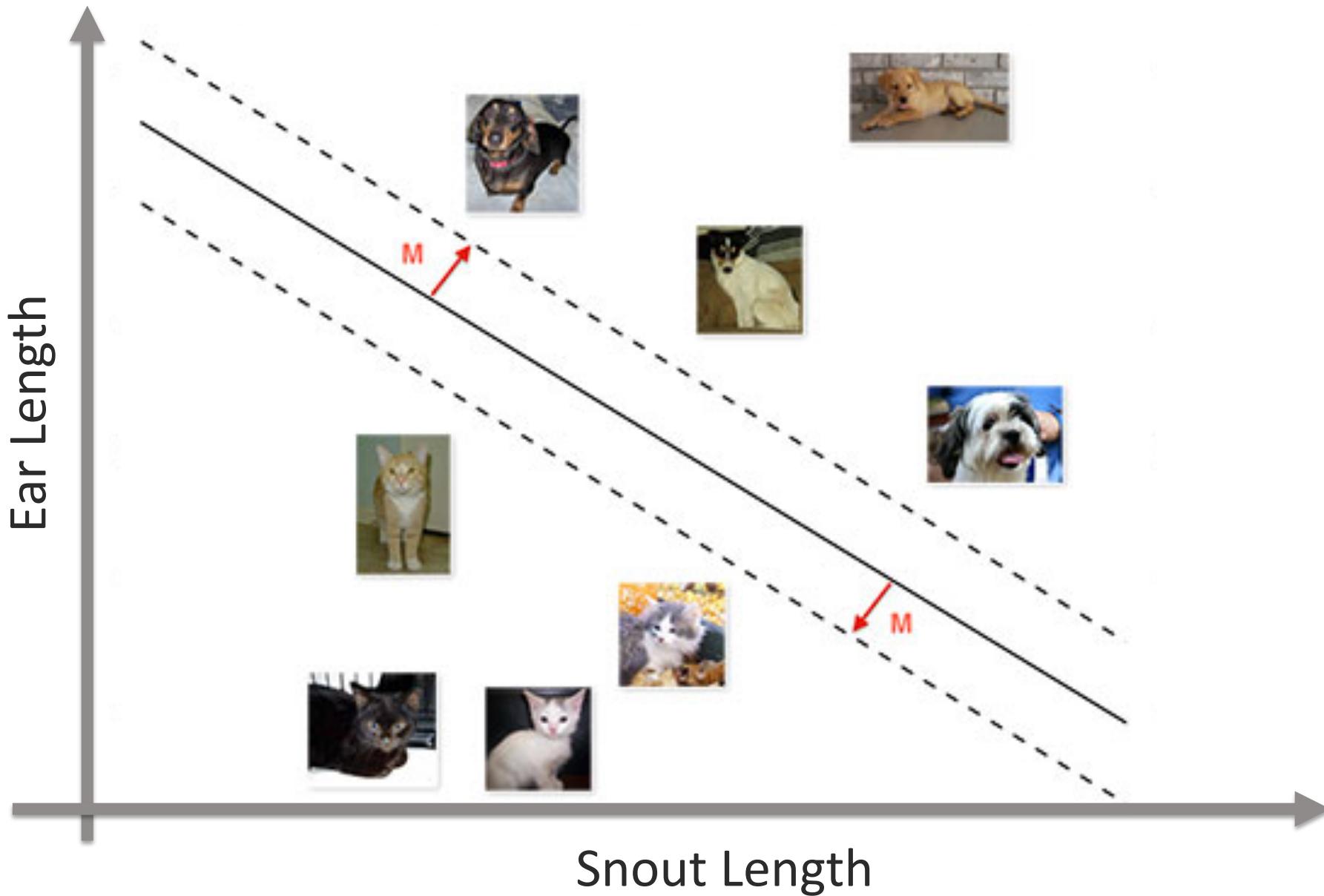


outlook	windy	play
sunny	false	no
sunny	true	yes
rainy	false	yes
rainy	true	no
sunny	true	yes
rainy	true	no



Decision Rule: IF (Outlook=sunny) AND (Windy=true) THEN Play=no

Support Vector Machine



Model Evaluation

Accuracy, Precision and Recall

IBM

CODE

Confusion Matrix

	p' (Predicted)	n' (Predicted)
P (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

P = Nike Bag



N = Nike Bag



Key Metrics

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$



FP: It is NIKE



FN: It is not NIKE

$$\uparrow \text{Precision} = \frac{TP}{TP+FP} \downarrow$$

$$\uparrow \text{Recall} = \frac{TP}{TP+FN} \downarrow$$

Predict Adults Income with SPSS Modeler

In Watson Studio

IBM
CODE

Create Visual Recognition Model

In Watson Studio

IBM
CODE

