Data science Bootcamp

Presenter:
Nora AlNashwan
@xnorax

Mentors:
Hissah Almuneef
Mead Alrshood
Reema Almashari

IBM History

IBM
CODE

Transform - CAMSS

IBM
CODE

IBM Cloud

IBM
CODE

Watson Studio

IBM CODE

Data Science

Introduction

IBM

CODE

Find Needle In Haystack

IBM CODE

Questions Data Science Answers

Is this dress or bag?

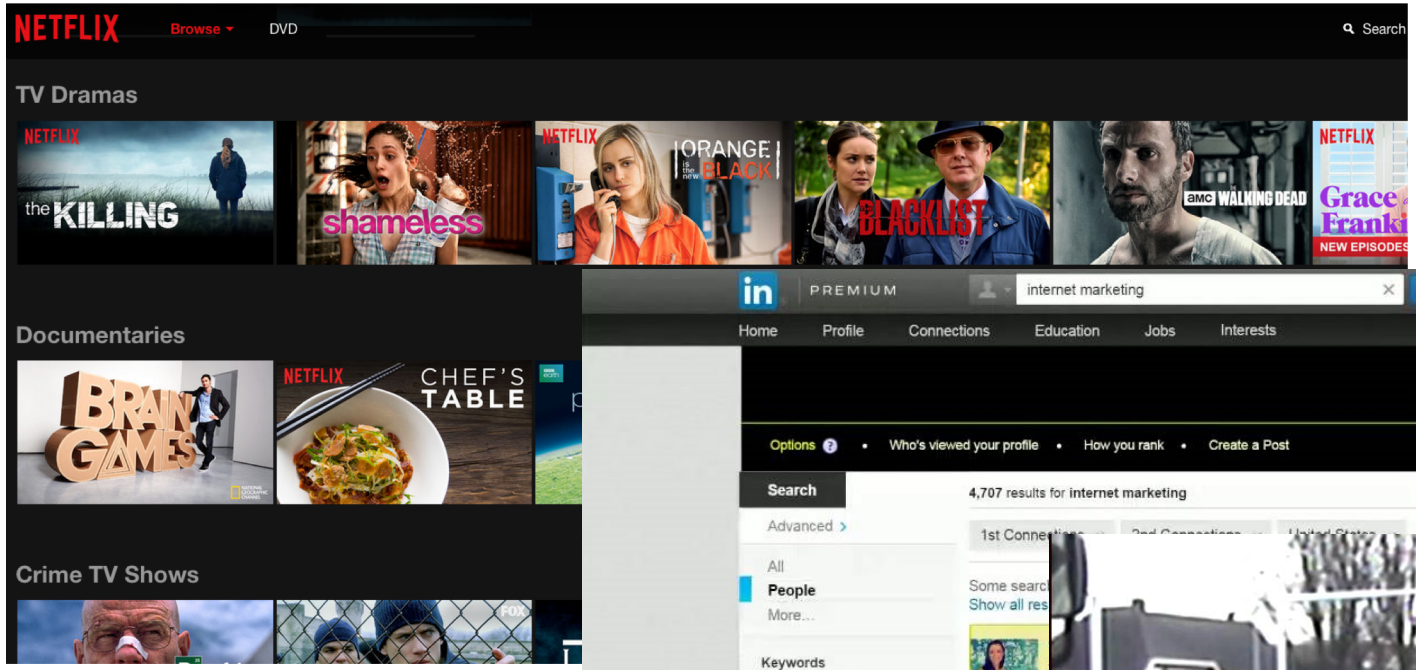
Is this weird?

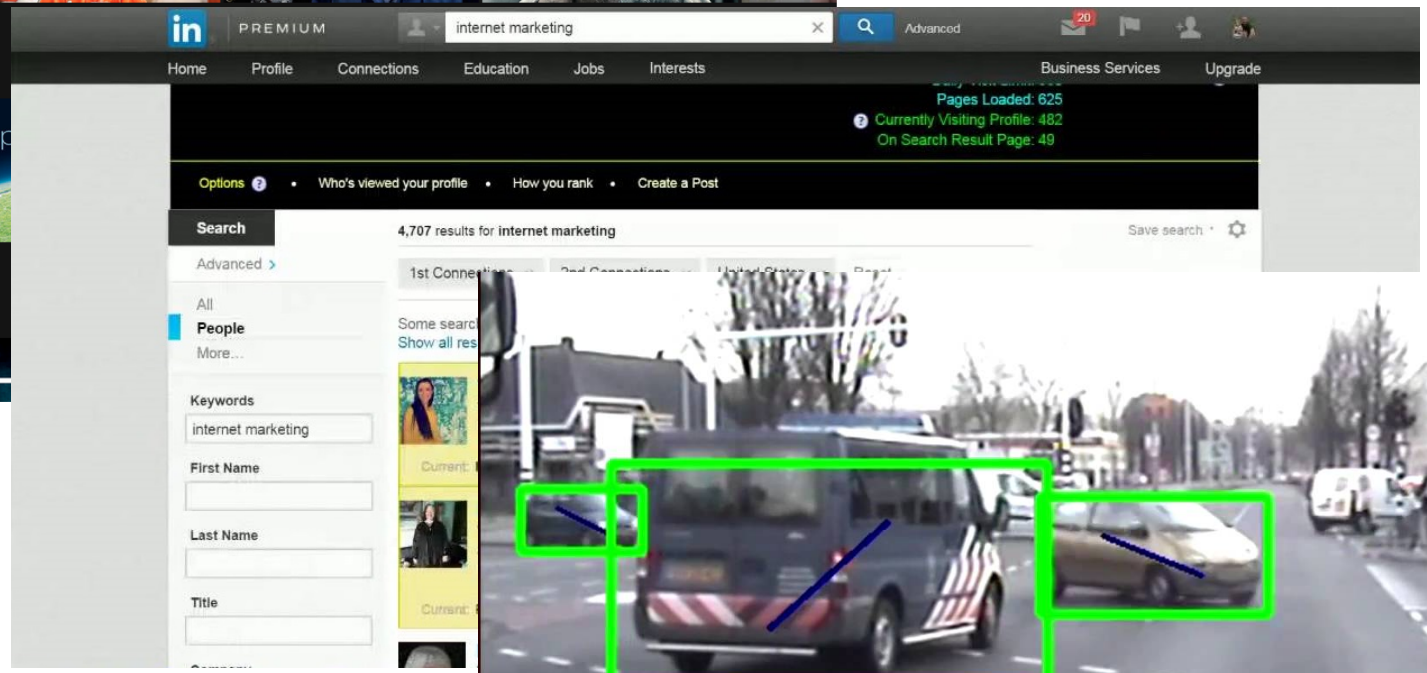
How much?

How is it segmented?

IBM CODE

Real Applications

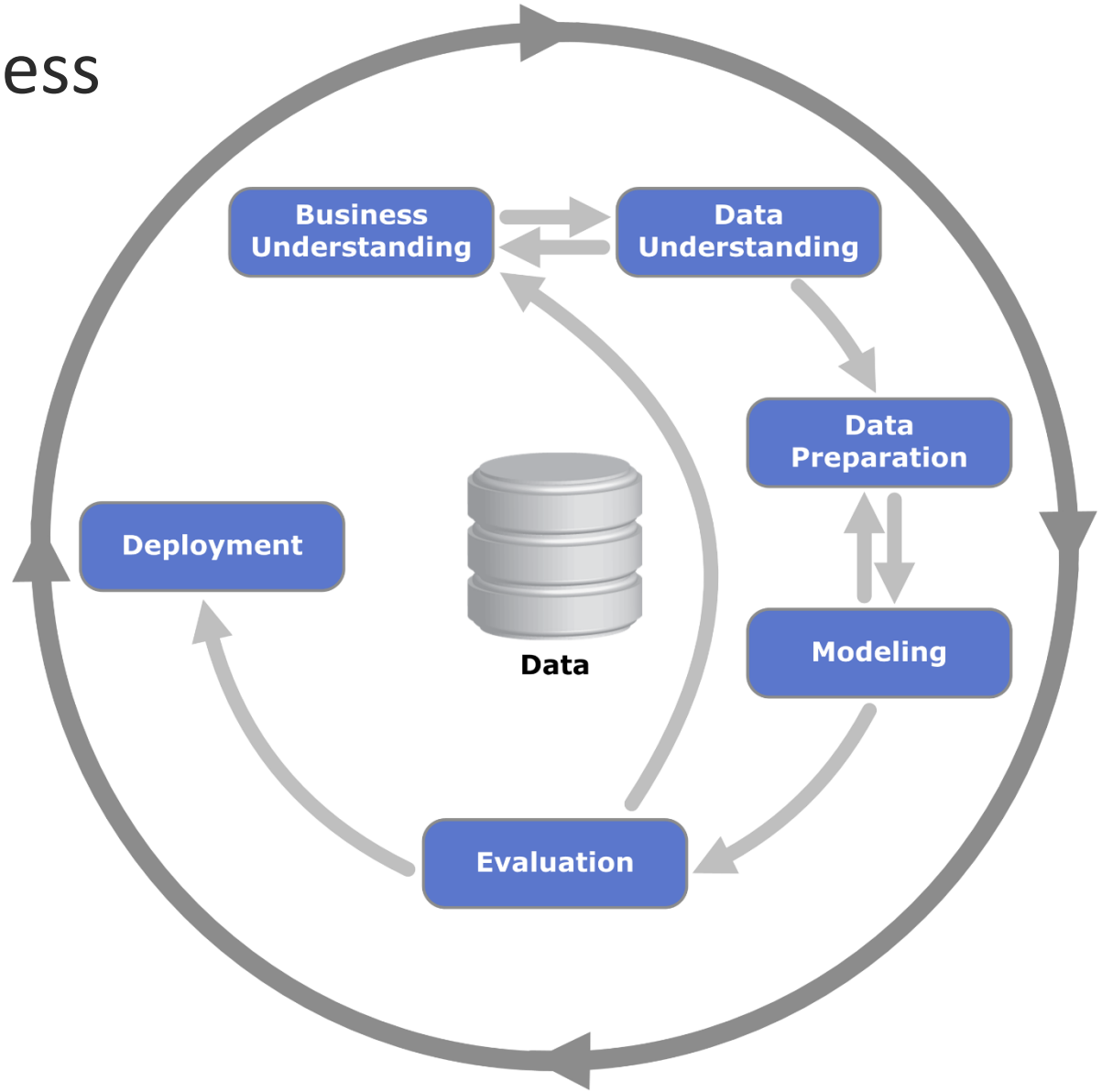




Do you know others?

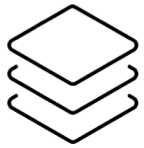
IBM CODE

Data Science Process

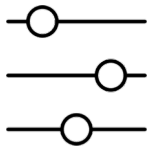


Data Science Tools

Preparing the Data



Filtering

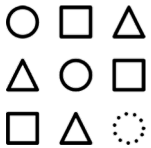


Integration

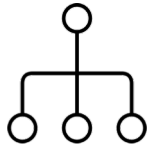
Storage

Visualization

Analytics



Deploying the ML Model



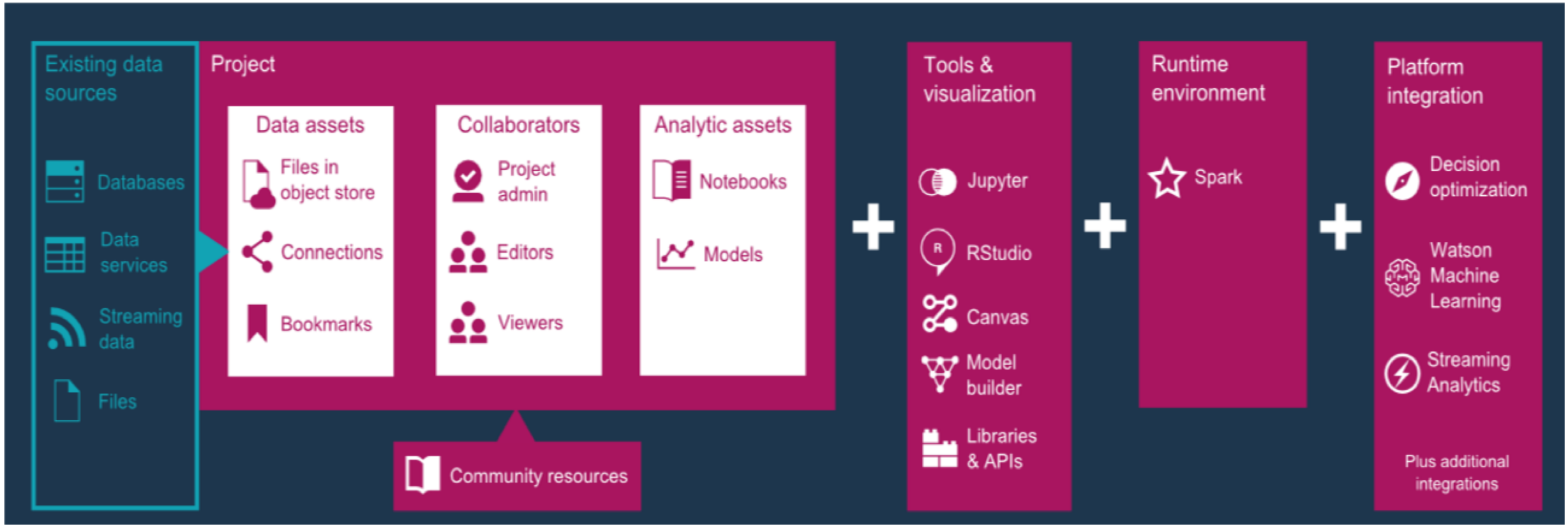
Watson Studio

Exploration

IBM

CODE

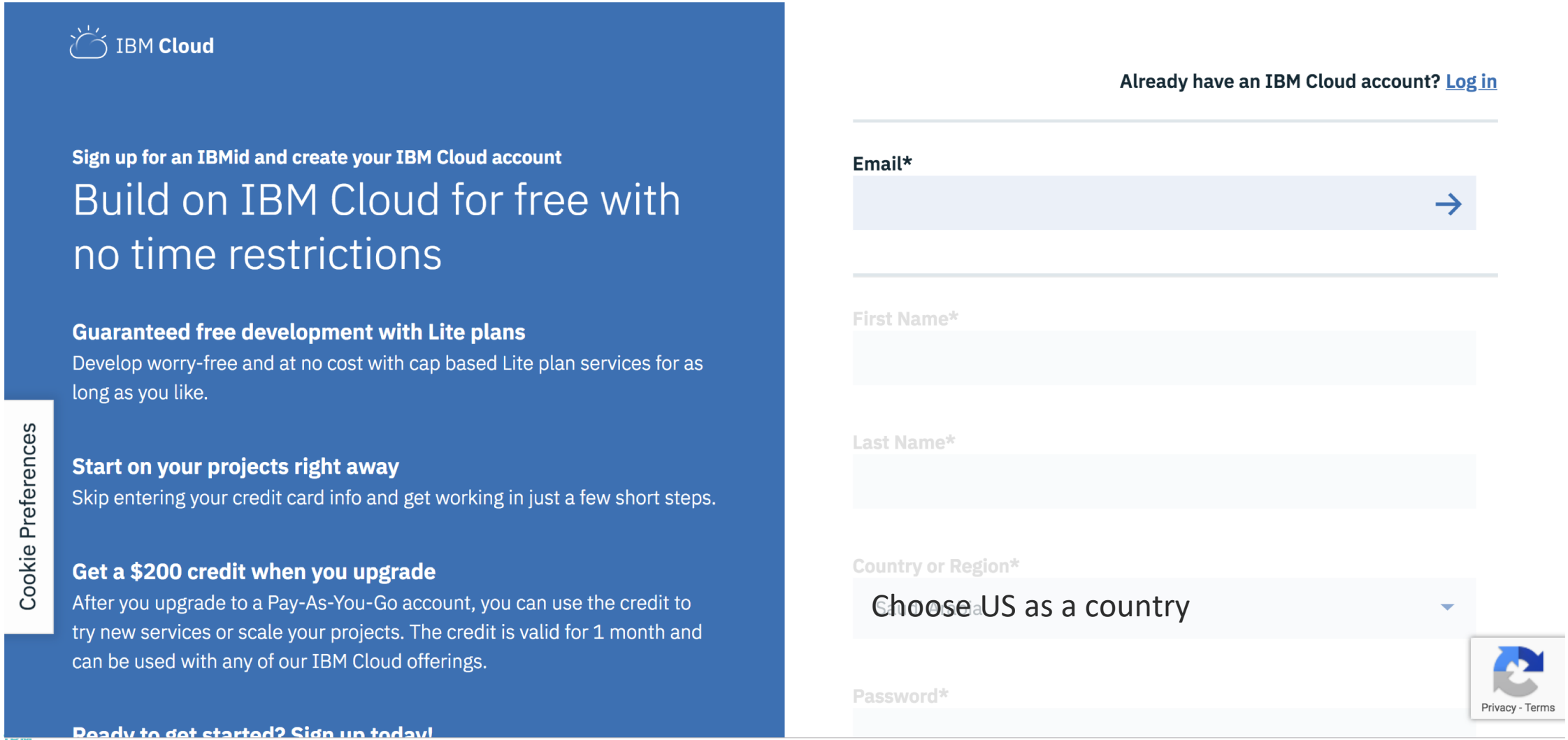
Watson Studio



IBM CODE

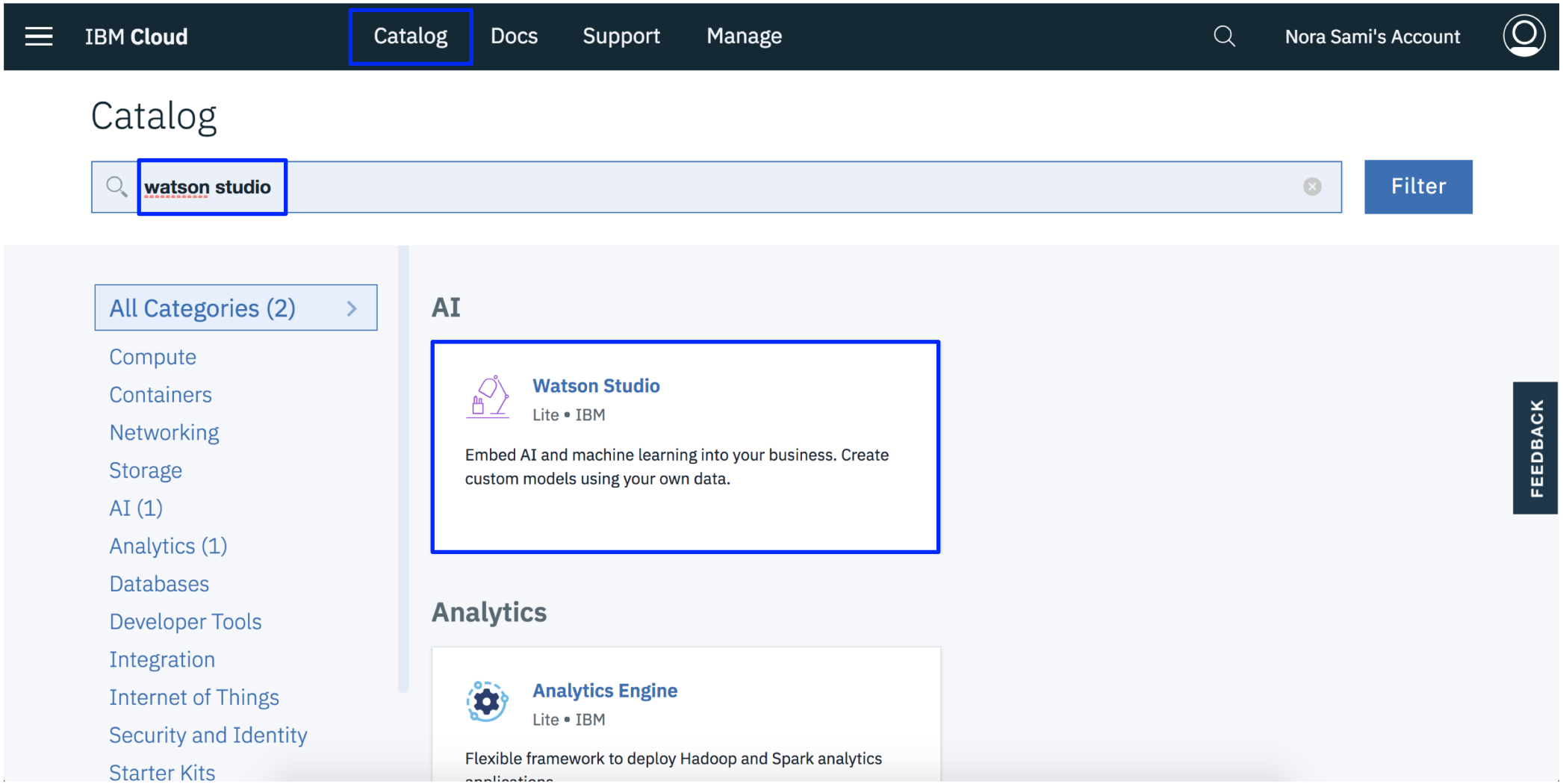
Register on IBM Cloud

https://ibm.biz/BdY4nQ



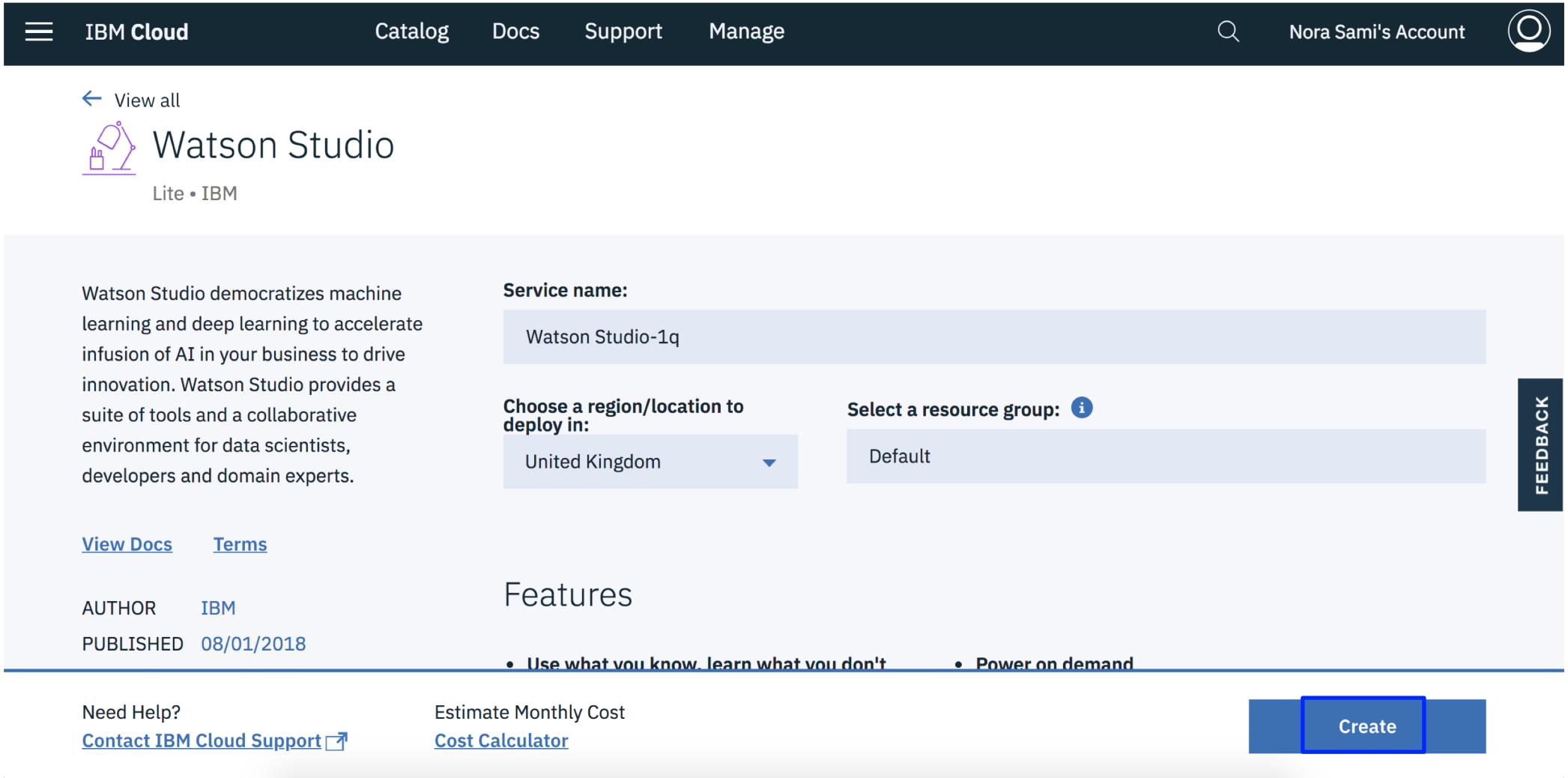
IBM CODE

Search for Watson Studio Service



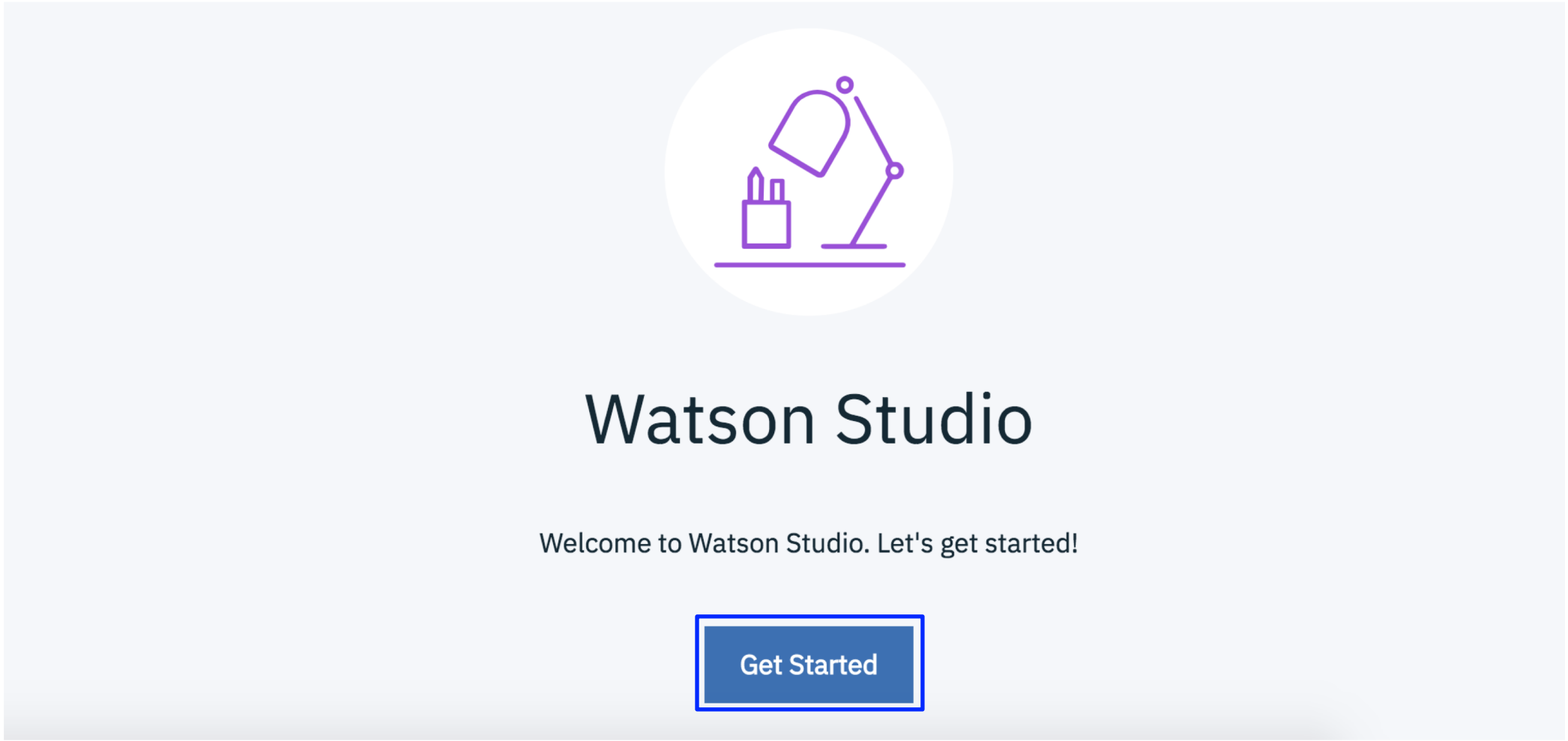
IBM CODE

Create Instance



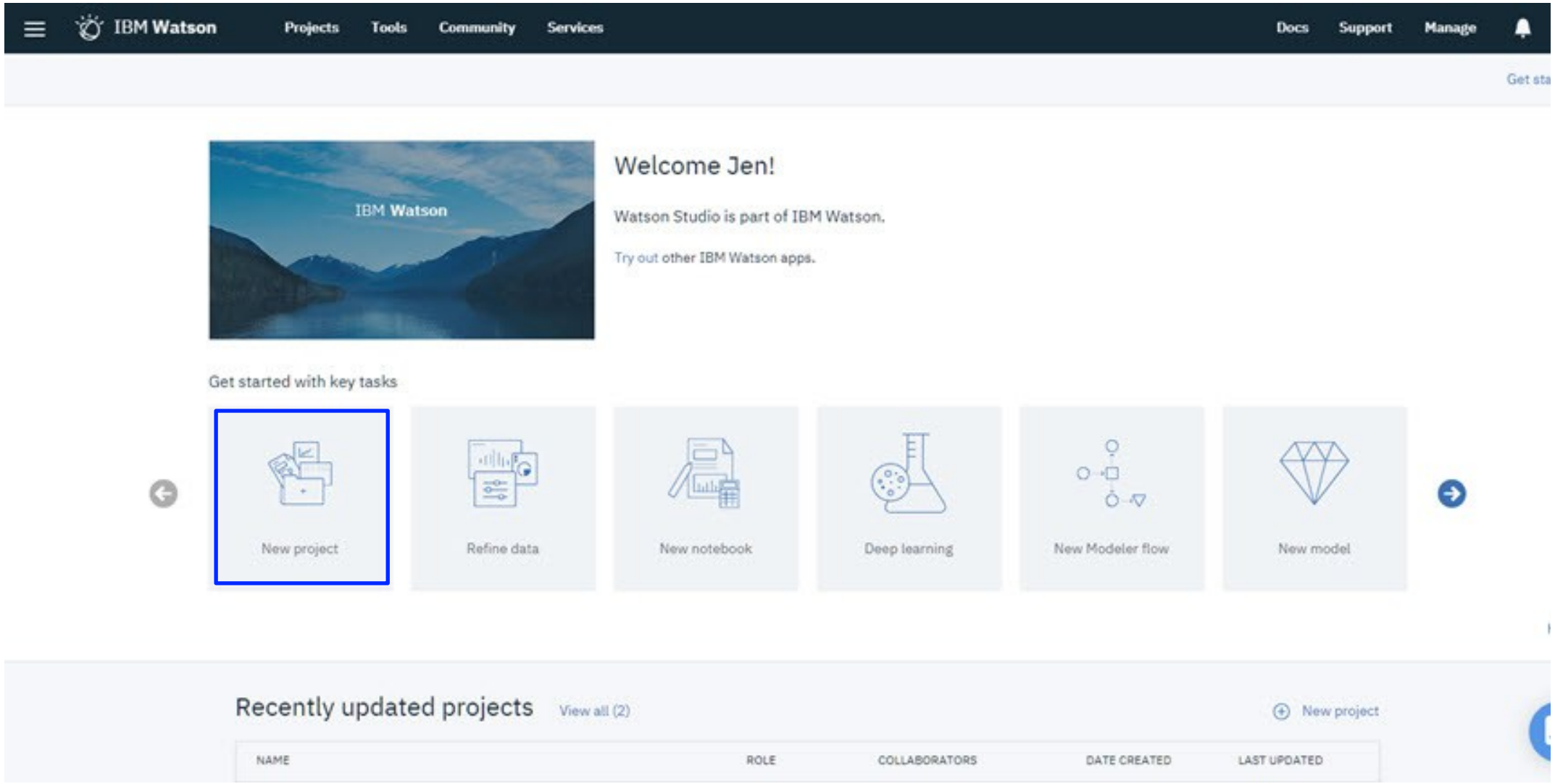
IBM CODE

Get Started



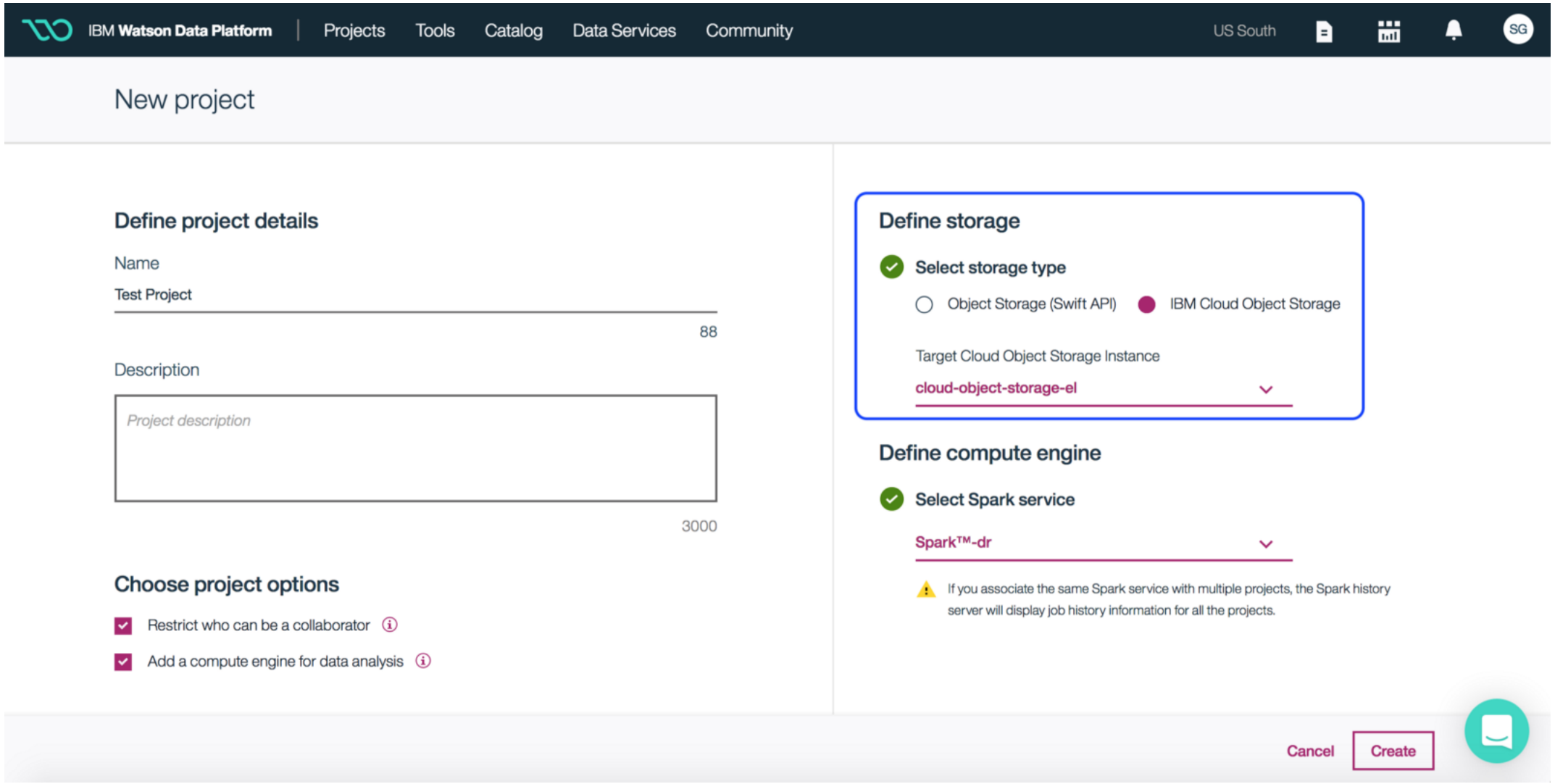
IBM CODE

Watson Studio Dashboard



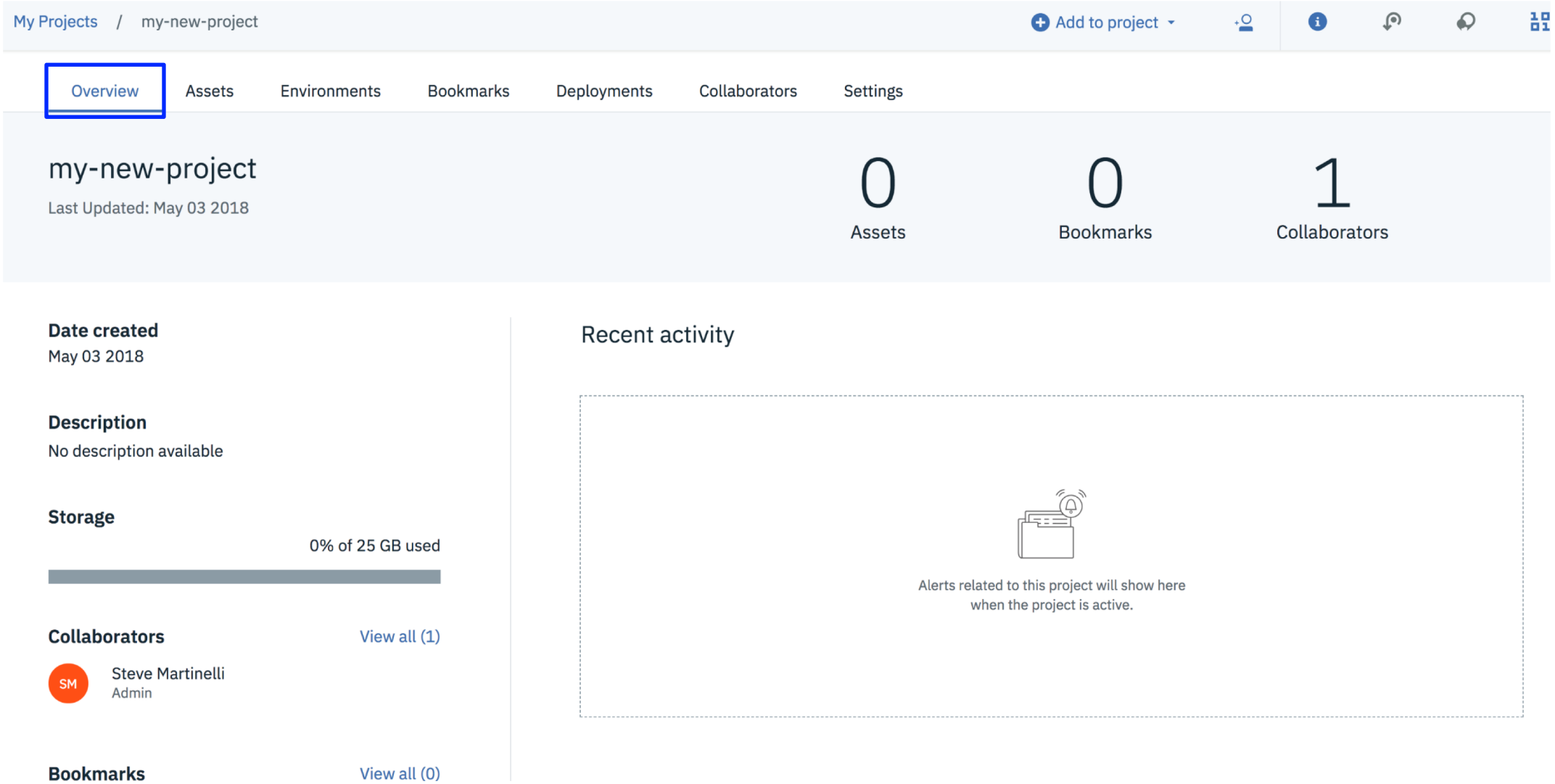
IBM CODE

Create Project



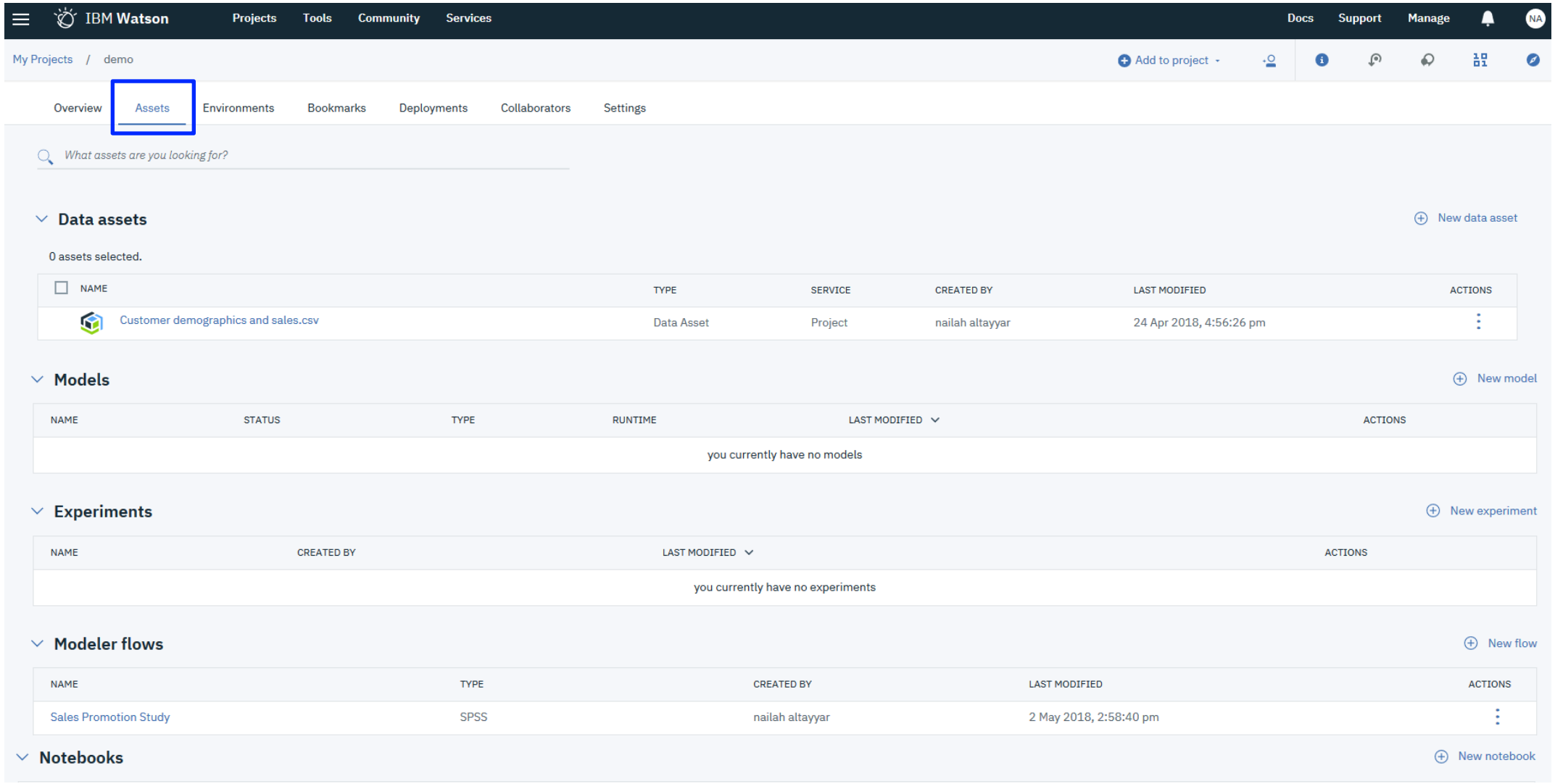
IBM CODE

Overview Page



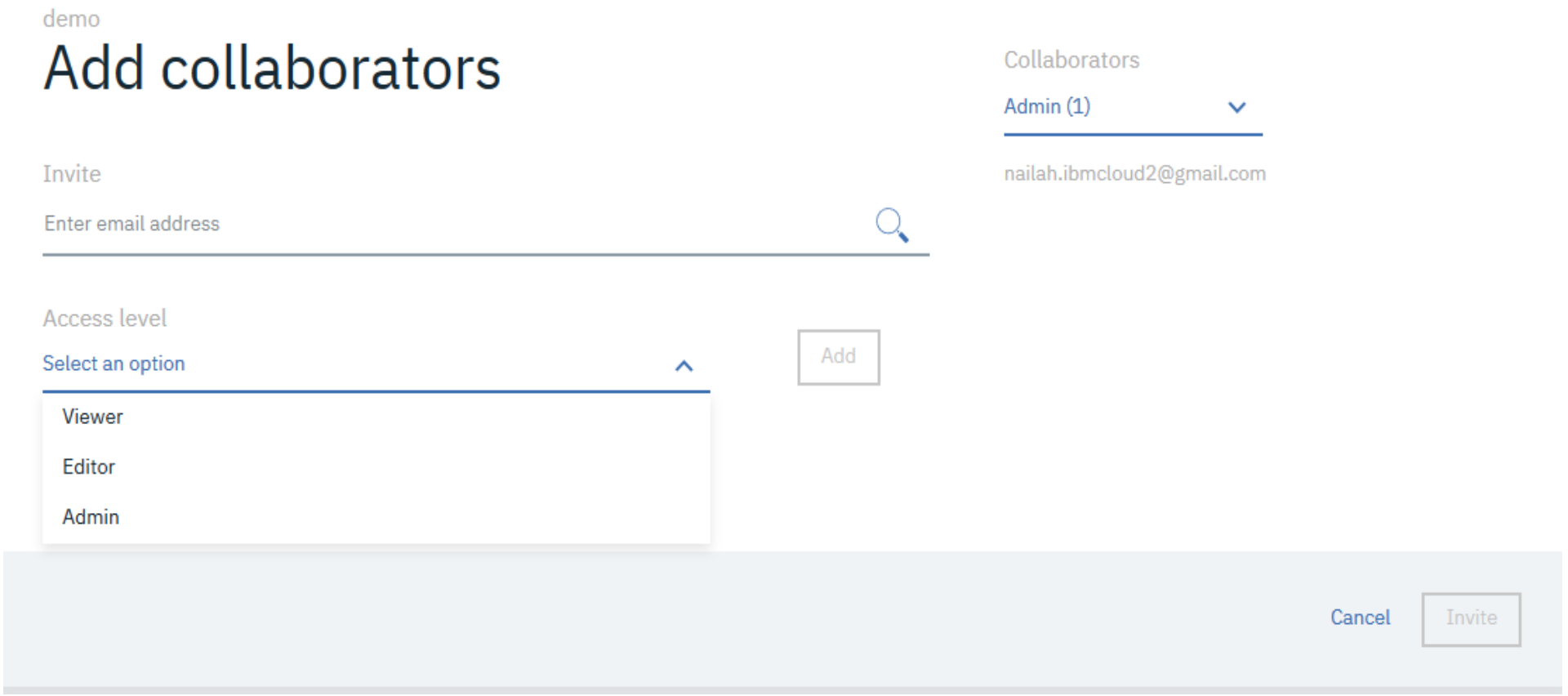
IBM CODE

Assets Page



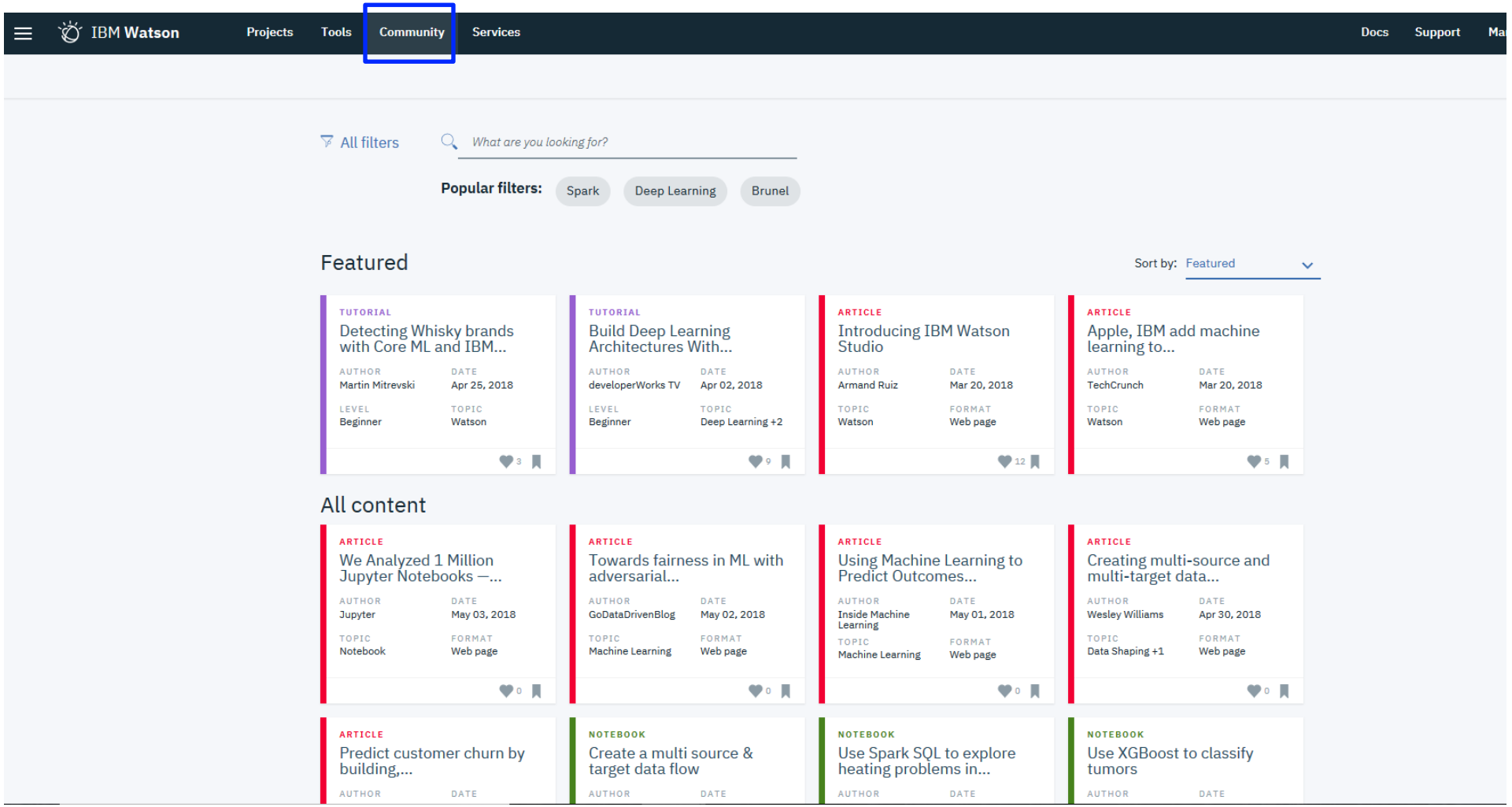
IBM CODE

Add Collaborators



IBM CODE

Community



IBM CODE

Watson Studio

M&Ms activity

IBM CODE

Data Exploration, Visualization, and Feature Engineering

Exploratory data analysis

IBM CODE

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

Model building

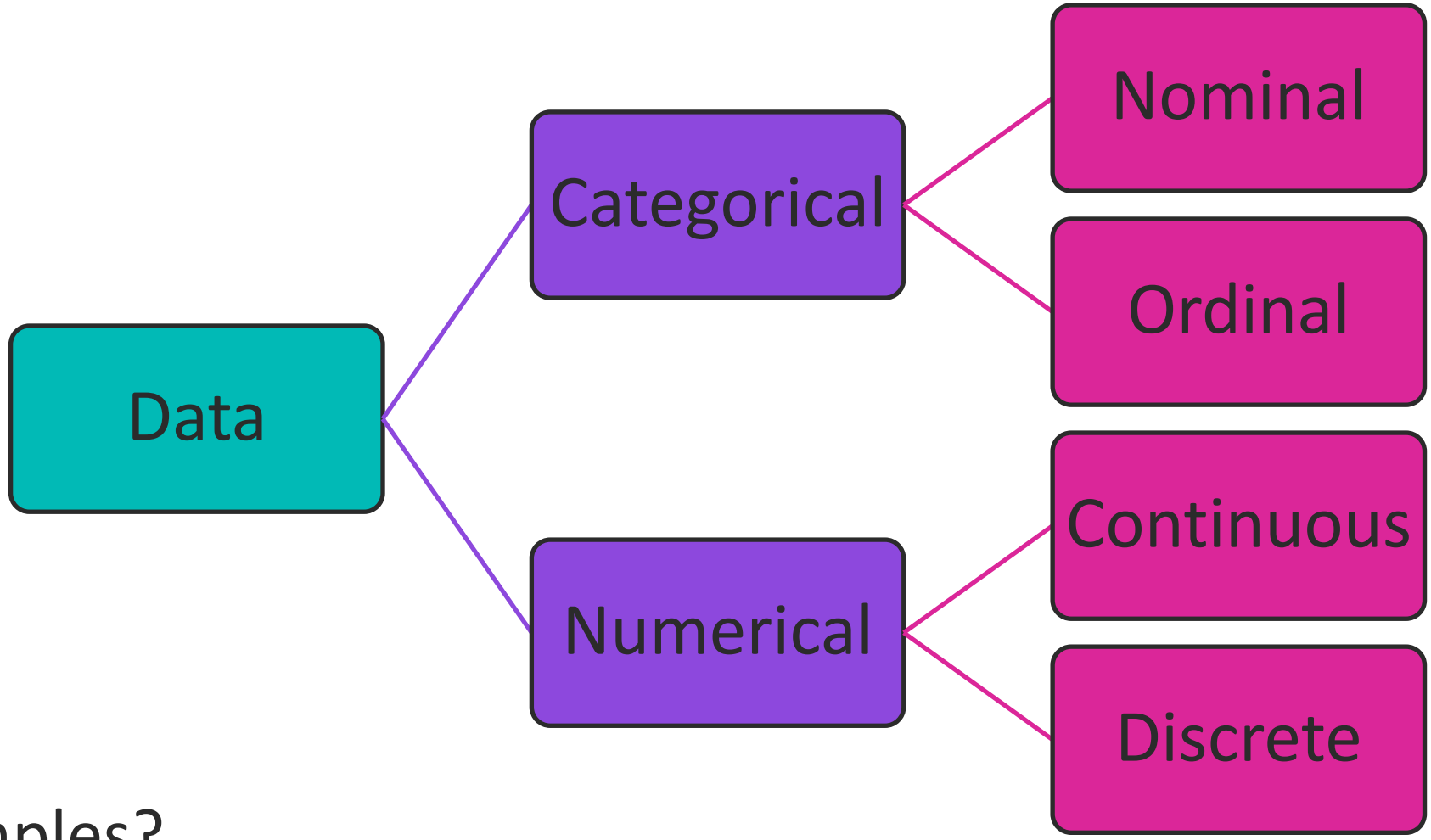
Date Exploration Steps

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

Data Types



Other examples?

IBM CODE

Identify data types

Categorical

Categorical

	JobID	Agency	PostingType	#OfPositions	BusinessTitle	CivilServiceTitle	TitleCodeNo	Level	JobCategory	FullOrPartTime	...	AdditionalInformation	ToApply	Hours	WorkLocation1
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	

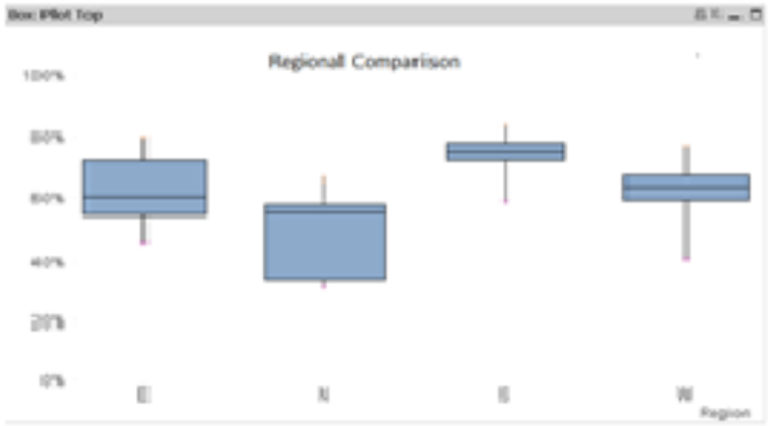
IBM CODE

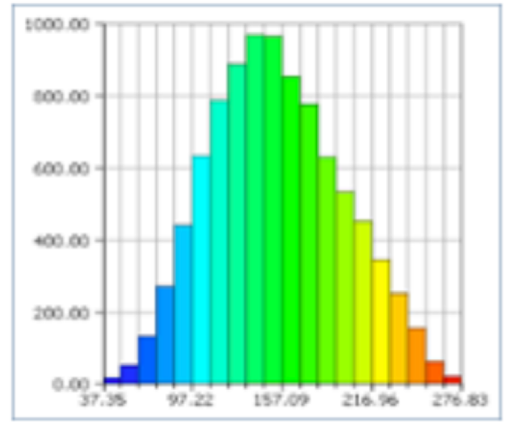
Step 2

Explore variables one by one

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	





IBM
CODE

A. Central Tendency

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

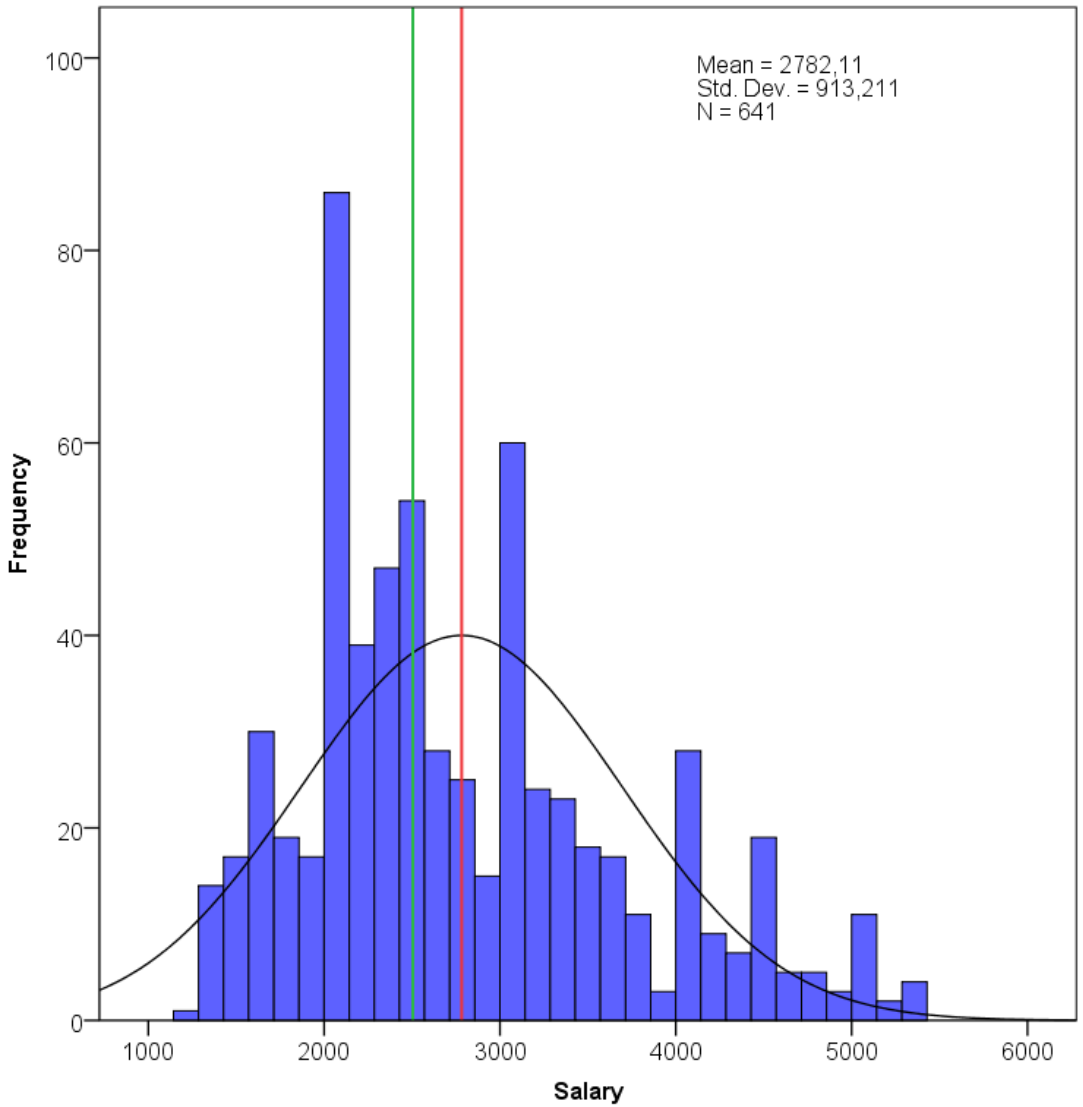
B. Variance and Standard Deviation

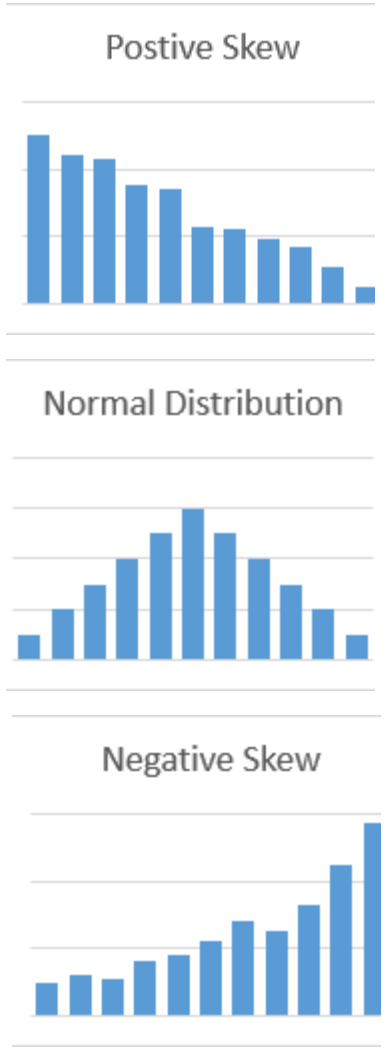
Standard deviation is the square root of the Variance

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

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

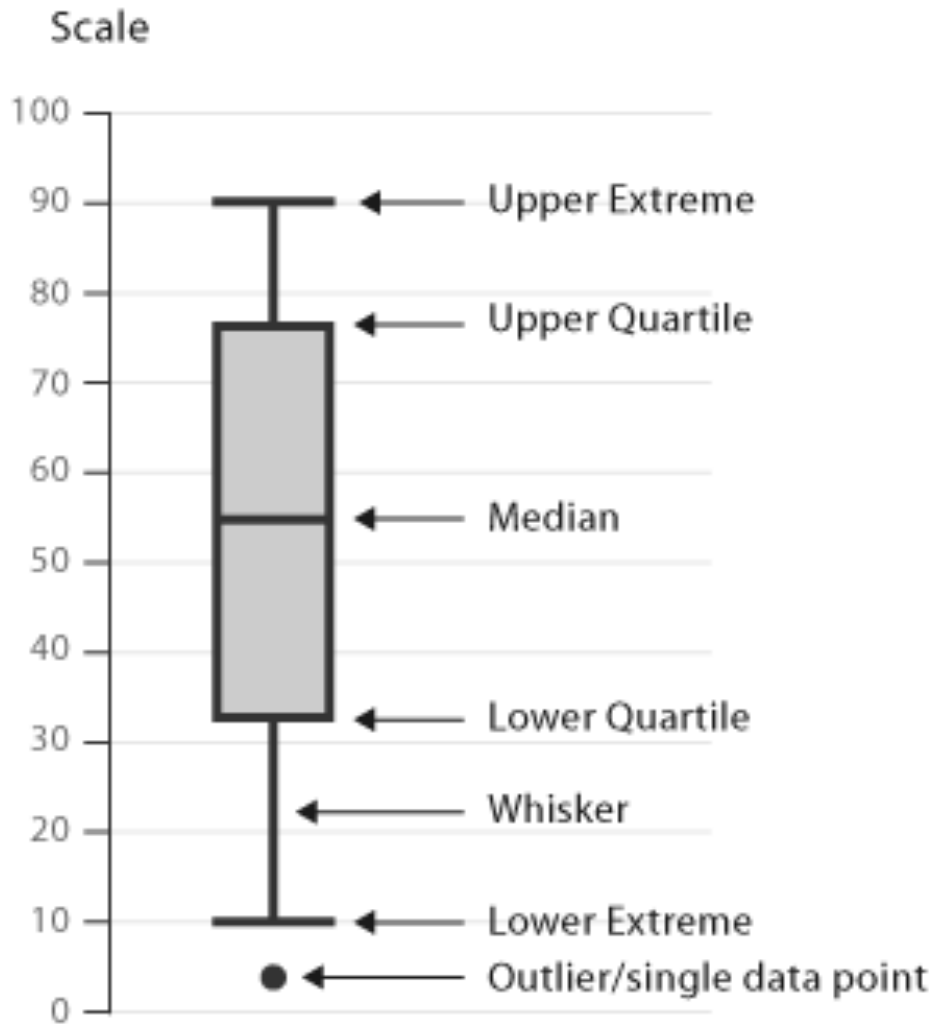
C. Histogram & Boxplot

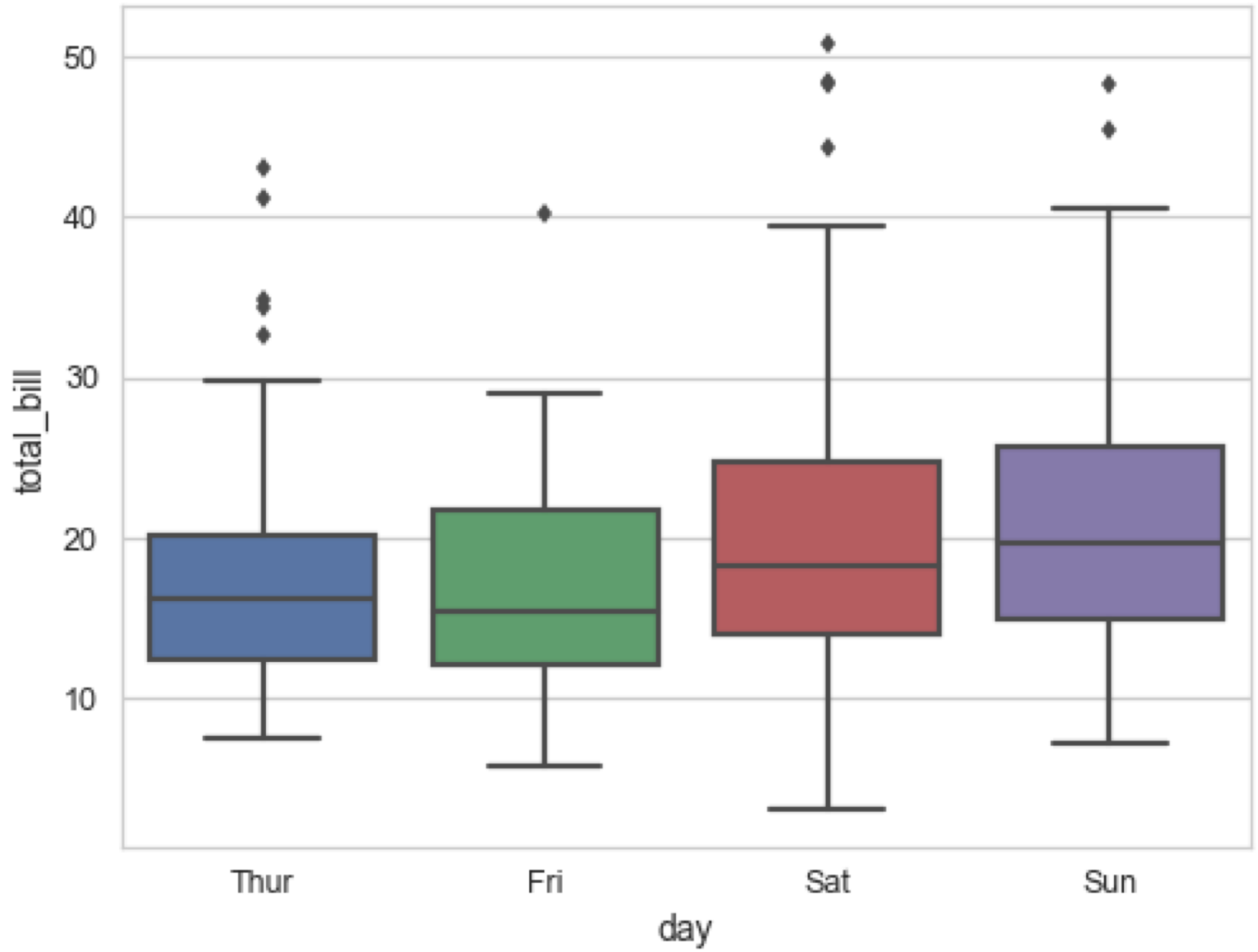




IBM CODE

C. Histogram & Boxplot

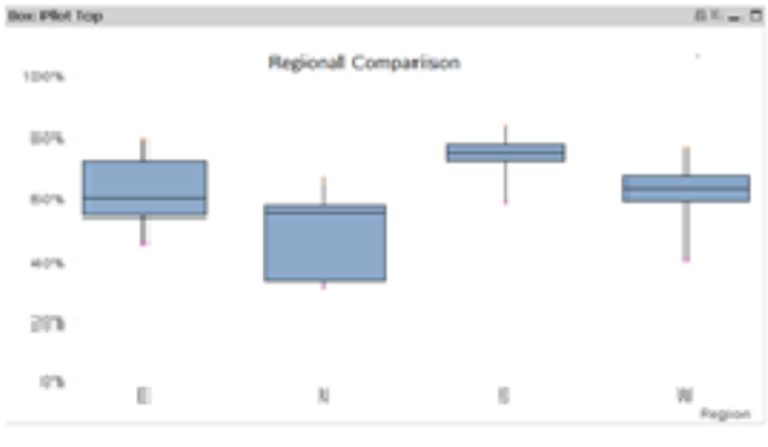


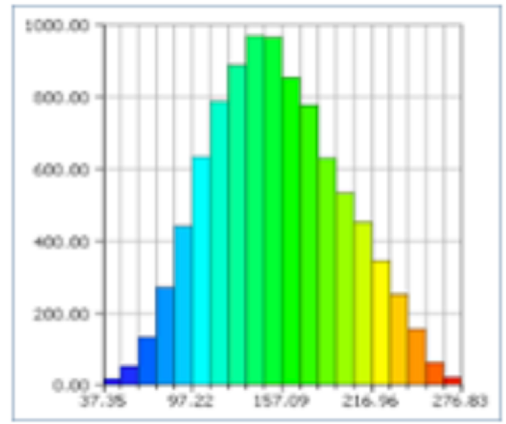


CODE

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	





IBM
CODE

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

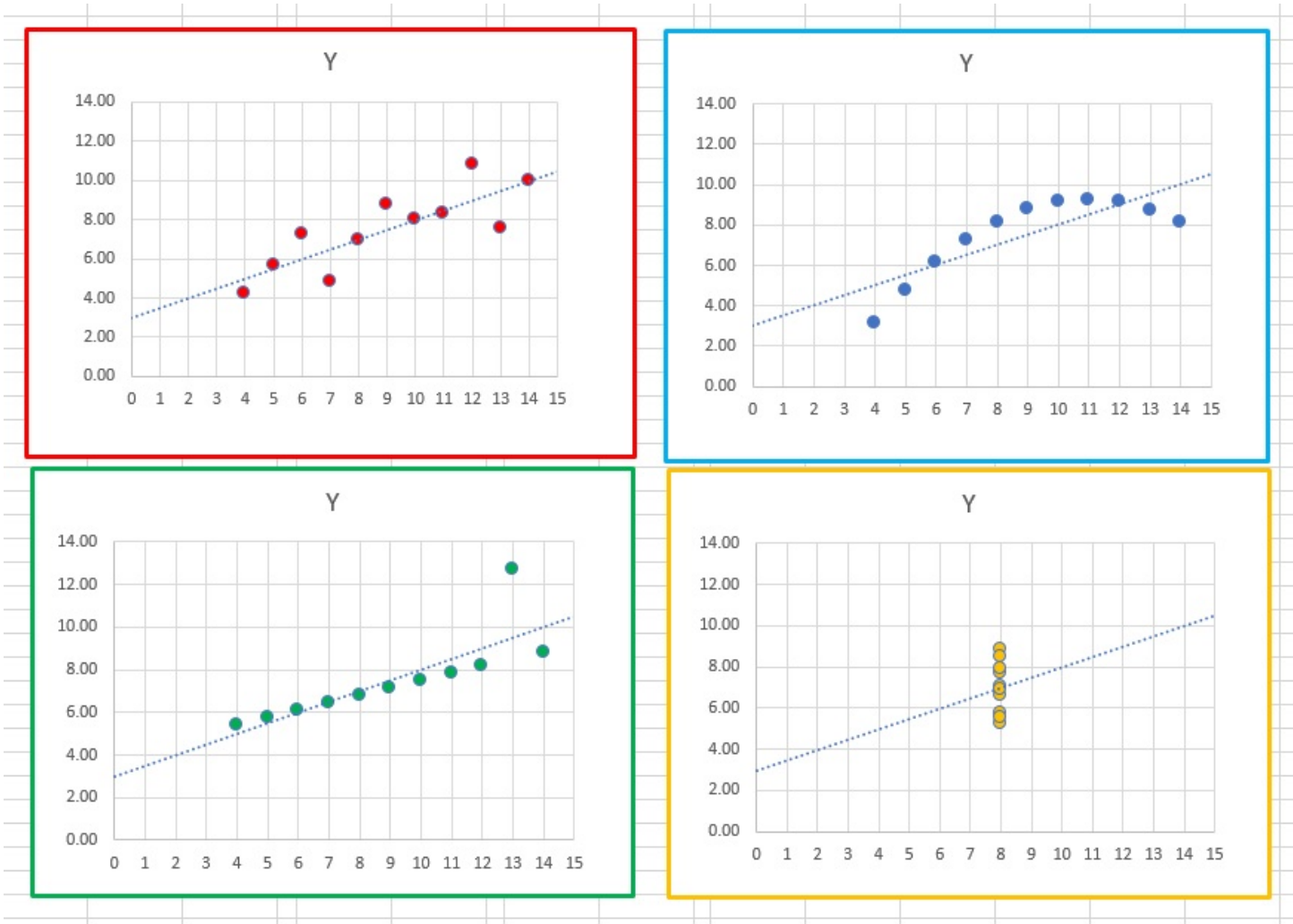
	XMean	YMean	XVar	YVar
Red	9.00	7.50	11.00	4.125
Blue	9.00	7.50	11.00	4.125
Green	9.00	7.50	11.00	4.125
Orange	9.00	7.50	11.00	4.125

IBM
CODE

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



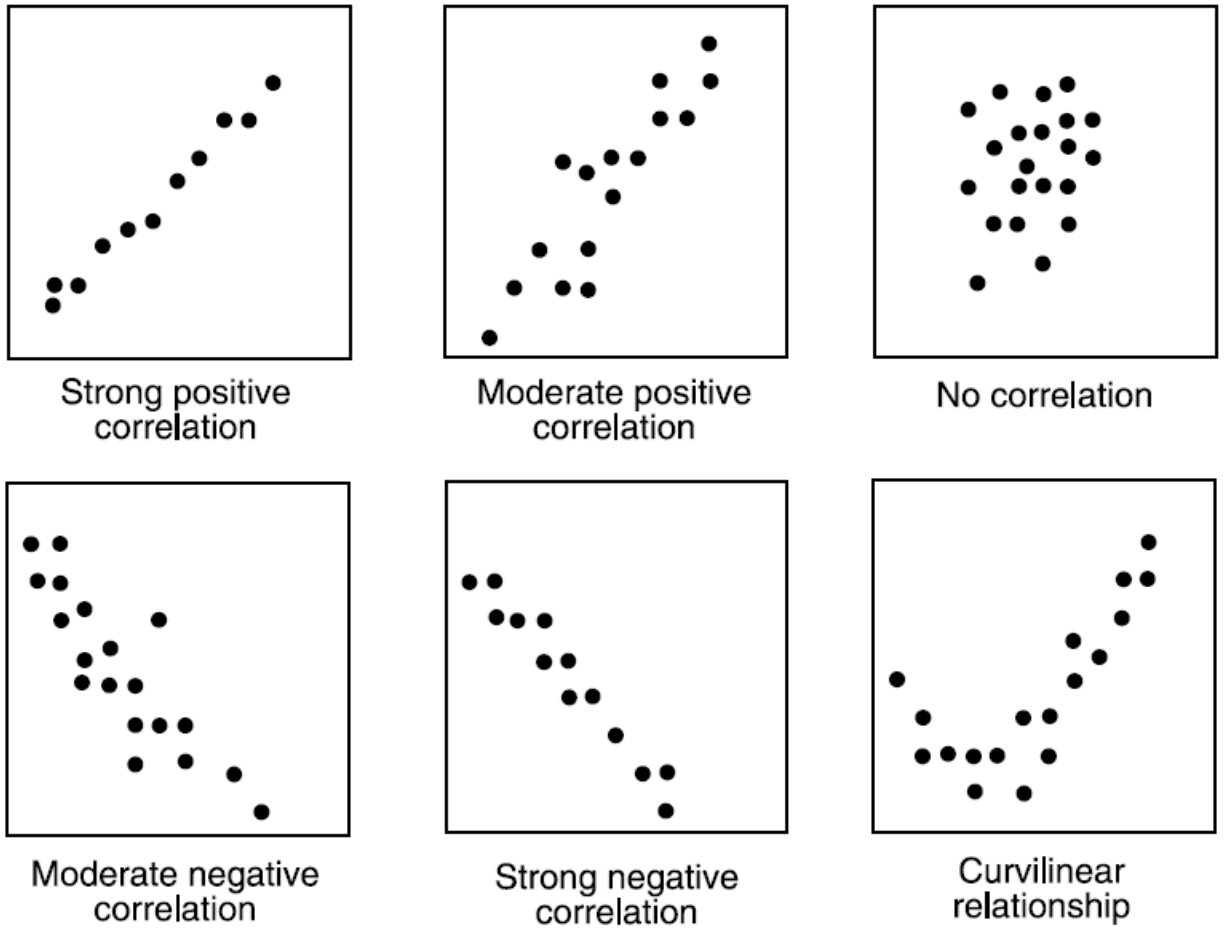
IBM
CODE

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

Continuous Variables



Categorical Variables

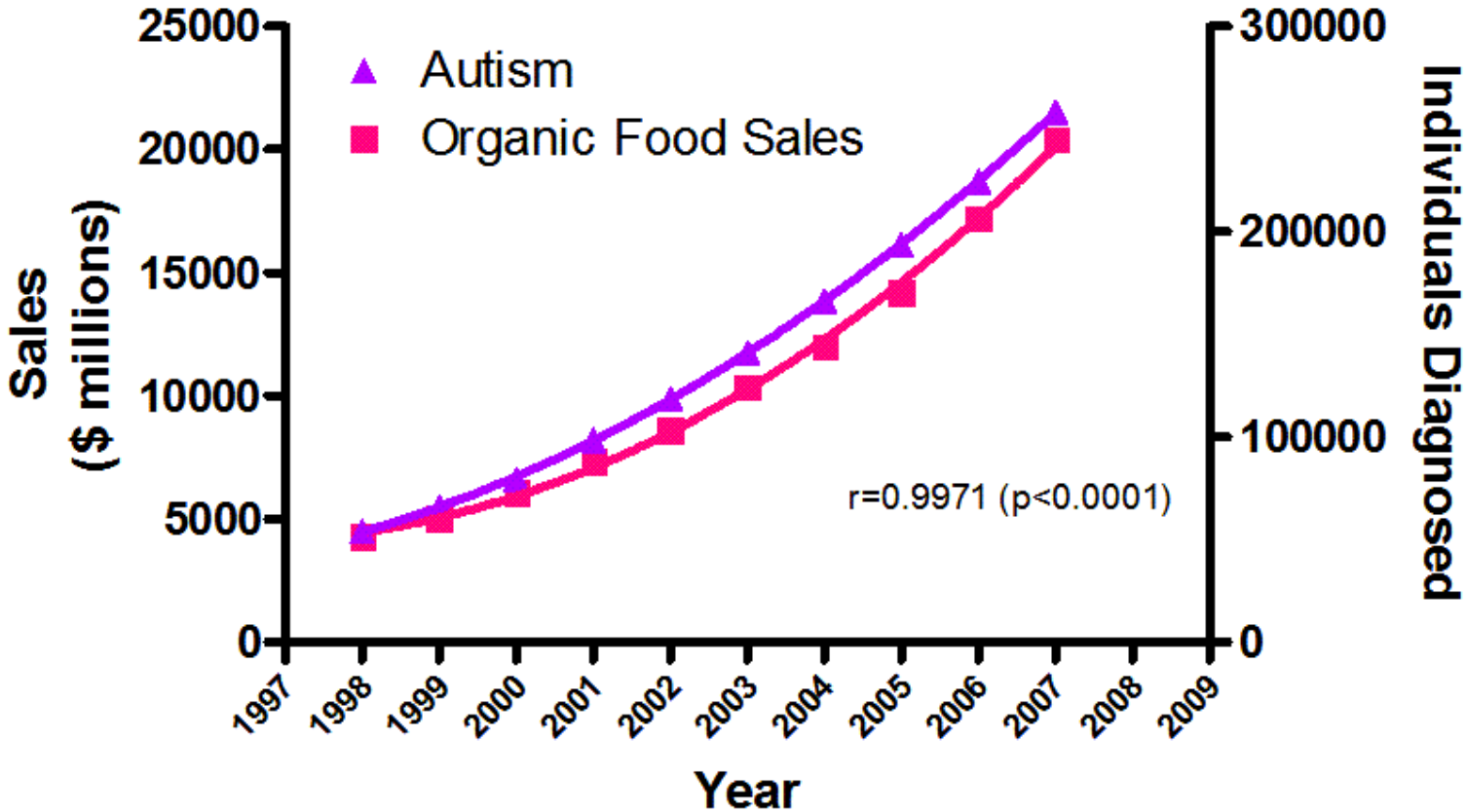
Whole Table

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

IBM CODE

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%

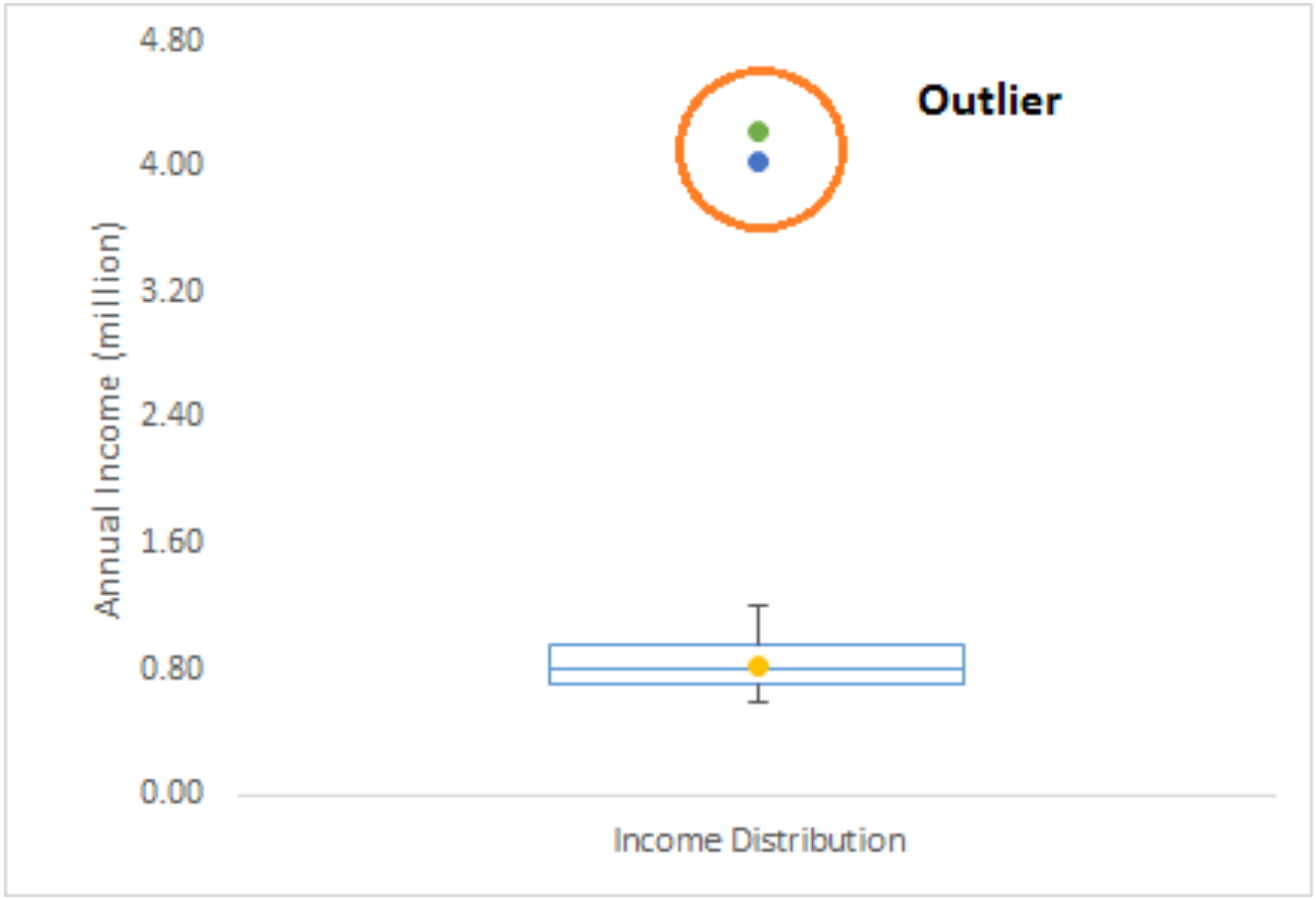
How to treat missing values?

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

Step 5

Outlier detection and treatment

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**

IBM CODE

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

→

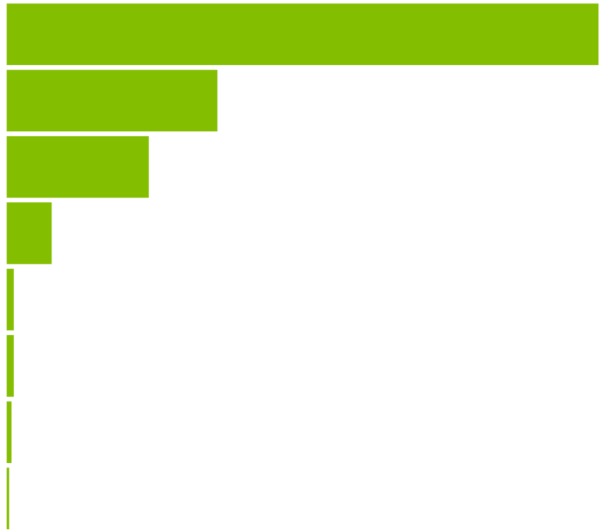
Day
Tue
Wed
Thu
Fri
Sat
Sun
Mon
Tue
Wed

Why Feature Engineering Is Important?

Value
Mr. Mohamad
Miss. Nora
Mrs. Rasha
Master. Ahmd
Dr. Saad
Rev. Joe
Sir. Saed
Col. Eead
Jonkheer. Slm
Lady. Sara
the Countess. S
Ms. Niyarrah

Most frequent items from <*SArray*>

Value	Count	Percent
Mr.	517	58.025%
Miss.	185	20.763%
Mrs.	125	14.029%
Master.	40	4.489%
Dr.	7	0.786%
Rev.	6	0.673%
Sir.	5	0.561%
Col.	2	0.224%
Jonkheer.	1	0.112%
Lady.	1	0.112%
the Countess.	1	0.112%
Ms.	1	0.112%



IBM CODE

Build a Machine Learning Model

For Classification

IBM CODE

Model

=

Data

+

Algorithm

IBM CODE

Data or Algorithms?

Everyone is talking about advances in algorithms.

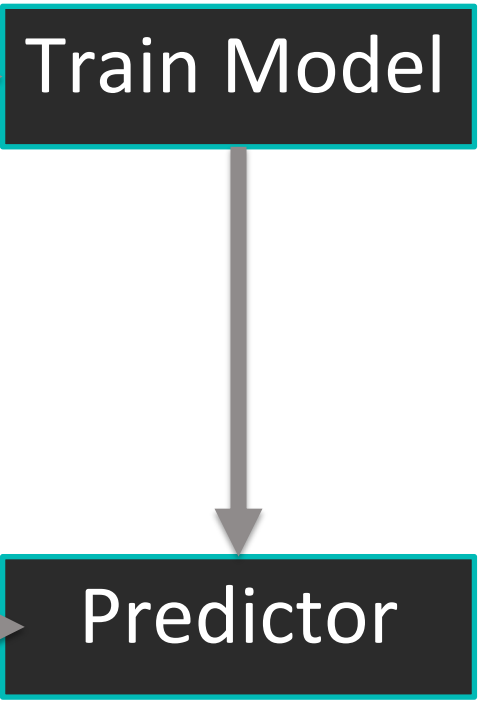
But it is and was the data, or more specifically how we store and process the data that was the single most important factor in the explosion of data science over the last decade.

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

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



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(Outlook=Sunny | Play=Yes) = 1/2
P(Windy=true | Play=Yes) = 1/2
P(Play=Yes) = 2/6

Probability that we can not play the game
P(Outlook=Sunny | Play=No) = 2/4
P(Windy=true | Play=No) = 3/4
P(Play=No) = 4/6

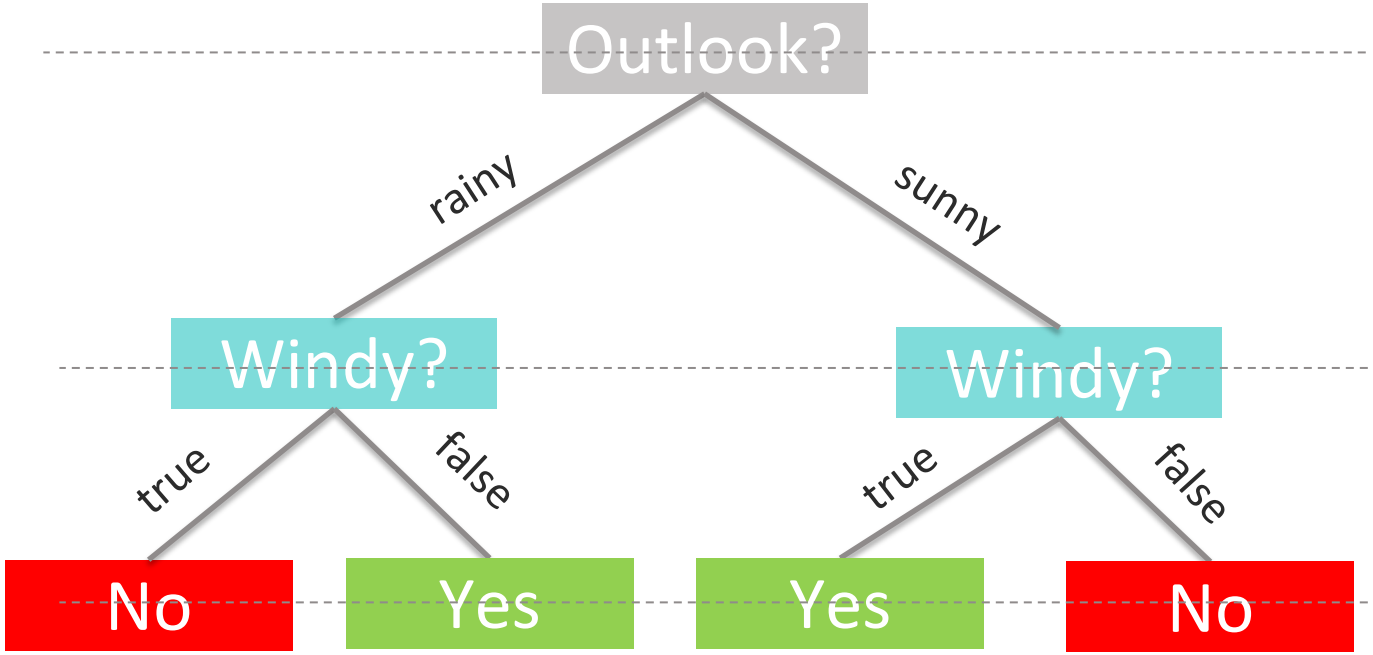
P(X | Play=Yes) = 1/3 * 1/4 * 2/6 = 0.08333

P(X | Play=No) = 2/3 * 3/4 * 4/6 = 0.25

Classification Tree

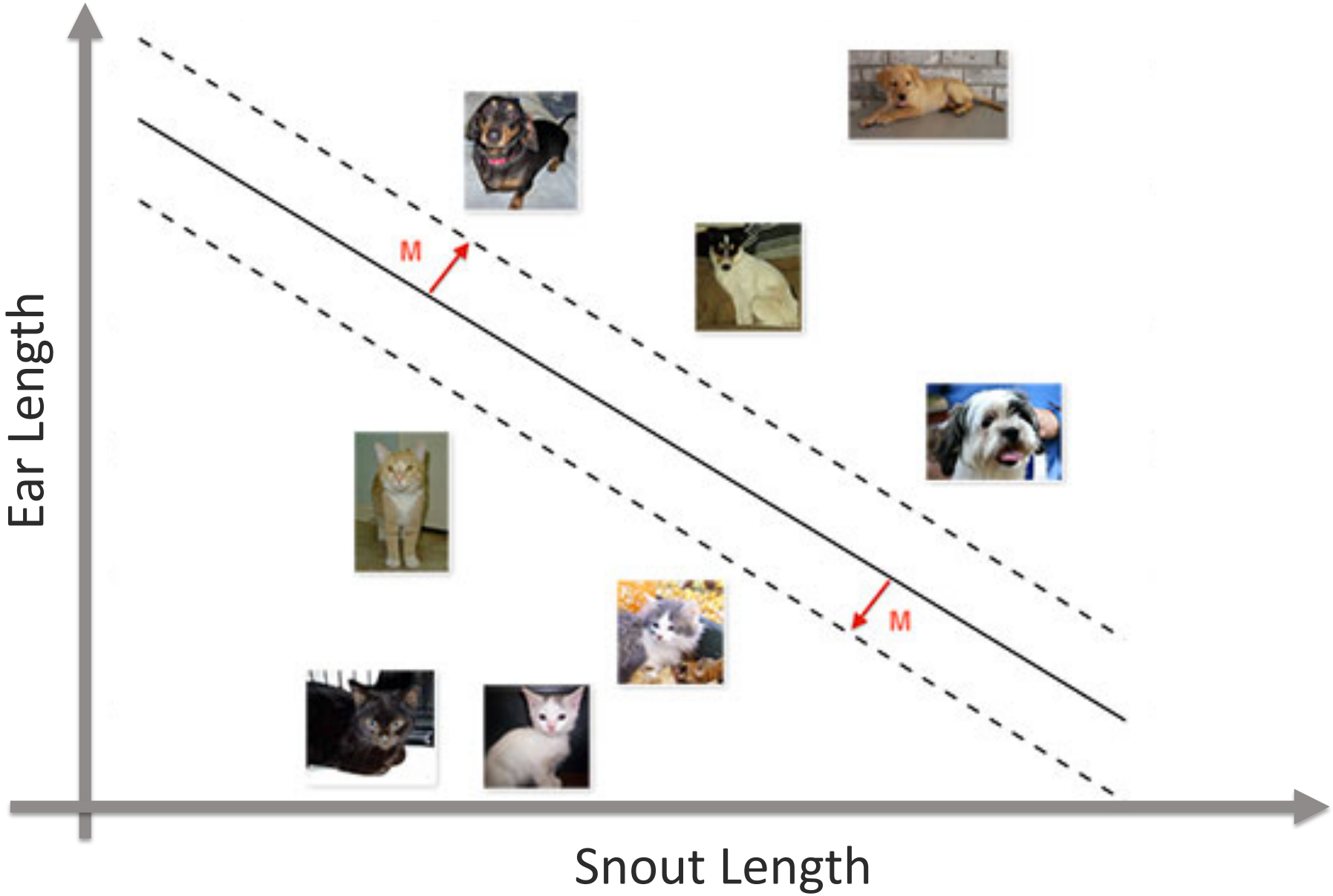


outlook	windy	play
sunny	false	no
sunny	true	yes
rainy	false	yes
rainy	true	no
sunny	true	yes
rainy	true	no



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

Support Vector Machine



Model Evaluation

Accuracy, Precision and Recall

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

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

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

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

FP: It is NIKE

FN: It is not NIKE

Predict Adults Income with SPSS Modeler

In Watson Studio

Create Visual Recognition Model

In Watson Studio

IBM CODE