Data Analysis and Data Visualisation

—

- Saurav Poudel

Introduction to Data Science

—

Data Science

- What is Data?
- What is Data Science?
- Why learn Data Science?
- How you can get started?

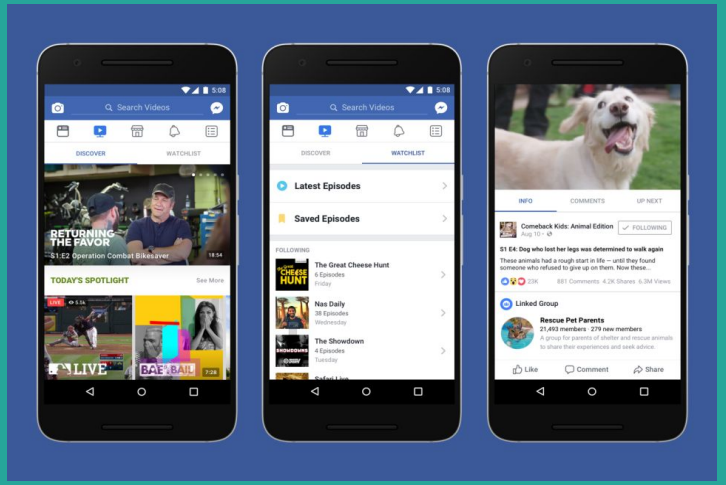
What is Data?

1. Data is simply a collection of facts.
2. We usually think of data as numbers.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Full Name	Hire Date	Location	State	Termination Date	Employment Type	Year	Base Salary	Bonus	Overtime	Commission	Total Compensation	Department	PTO Days	Sick Days	Performanc Score
2	Alison Johnson	5/13/2000	Boston	MA		Full-Time	2000	$89,000	$6,230	$0		$95,230	Marketing	14	2	
3	Corine M. 'Henderson	8/7/2000	Boston	MA		Full-Time	2000	$72,000	$7,920	$2,880		$82,800	R&D	14	1	
4	Julia Hegwood	8/24/2000	Bos"to"n	MA		Full-Time	2000	$45,000	$3,150	$0	$14,000	$62,150	Sales	14	2	
5	Jeremiah De Grazia	9/8/2000	Boston	MA		Full-Time	2000	$58,000	$4,640	$1,740		$64,380	Finance	12	1	
6	Willow Nevandro	10/18/2000	Boston	MA		Full-Time	2000	$34,000	$2,380	$1,020		$37,400	Administration	12	3	
7	Martie Elmasian	12/7/2000	Boston	MA		Full-Time	2000	$54,000	$6,480	$2,160		$62,640	R&D	13	6	
8	Alison Johnson	5/13/2000	Boston	MA		Full-Time	2001	$92,000	$7,360	$0		$99,360	Marketing	13	6	
9	Corine M. Henderson	8/7/2000	Boston	MA		Full-Time	2001	$75,000	$5,250	$3,750		$84,000	R&D	14	9	
10	Julia Hegwood	8/24/2000	Boston	MA		Full-Time	2001	$48,000	$2,880	$0	$44,000	$94,880	Sales	15	3	
11	Jeremiah De Grazia	9/8/2000	Boston	MA		Full-Time	2001	$62,000	$3,100	$3,100		$68,200	Finance	11	6	
12	Willow Nevandro	10/18/2000	Boston	MA		Full-Time	2001	$36,000	$2,520	$1,080		$39,600	Administration	14	3	
13	Martie Elmasian	12/7/2000	Boston	MA		Full-Time	2001	$60,000	$4,200	$2,400		$66,600	R&D	14	1	
14	Kelly Queen	1/13/2001	Los Angeles	CA		Full-Time	2001	$43,000	$5,160	$0	$42,000	$90,160	Sales	12	3	
15	Cristhian Roth	4/2/2001	Los Angeles	CA		Full-Time	2001	$54,000	$4,320	$0		$58,320	Administration	10	6	
16	Joeanne Melendez	5/4/2001	Los Angeles	CA		Full-Time	2001	$32,000	$2,560	$0		$34,560	R&D	14	9	
17	Brian M Stucki	5/10/2001	Chicago	IL		Full-Time	2001	$43,000	$3,440	$0	$22,000	$68,440	Sales	14	7	
18	Cindy Summerville	6/23/2001	Chicago	IL		Full-Time	2001	$63,000	$3,780	$0		$66,780	IT	12	9	
19	Tony Merrick	9/18/2001	Chicago	IL		Full-Time	2001	$105,000	$9,450	$3,150		$117,600	R&D	10	1	
20	Bryan Anderson	10/6/2001	Chicago	IL		Full-Time	2001	$66,000	$3,960	$3,300		$73,260	Accounting	13	3	
21	Janalee Fergleston	10/21/2001	Boston	MA		Full-Time	2001	$28,000	$1,680	$840		$30,520	R&D	11	3	


Video Downloader
professional
Web WebiWebWeb Shopping Pictures Videos Maps Translate Wikipedia
Suchen
Go to the original site
close
Deutschland vs. Argentinien 1:3 Fußball Testspi...
80:48
GER 0 3 ARG
Filter videos
Move videos

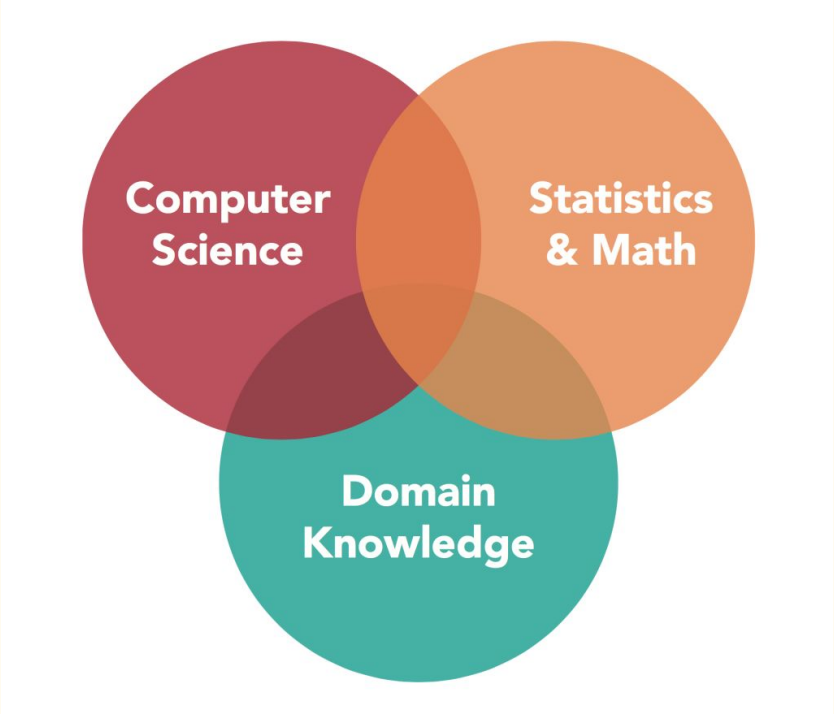

LEONARDO DICAPRIO
MATT DAMON
JACK NICHOLSON
and MARK WAHLBERG
THE DEPARTE
TIM ROBBINS
MORGAN FR
SHAWSHA
REDEMPTI
MEMENT
QUENTIN
TARANTINO
WALTZ
DICAPRIO
WASHINGTON
JACKSON
DJANGO


Search Videos
DISCOVER
WATCHLIST
RETURNING THE FAVOR
S1:E2 Operation Combat Bikesaver
TODAY'S SPOTLIGHT
See More
LIVE
BAE BAIL
Latest Episodes
Saved Episodes
FOLLOWING
The Great Cheese Hunt
Nas Daily
The Showdown
INFO
COMMENTS
UP NEXT
Comeback Kids: Animal Edition
Linked Group
Rescue Pet Parents
Like
Comment
Share

Can we make sense of this data?

What is Data Science?

1. Data Science is the science of making sense of data.
2. It is the combination of Statistics, Computer Science, and Domain Knowledge.
3. With the goal to extract knowledge of Data.



Business Intelligence vs Data Science

- Business Intelligence usually answers ‘What happened’ and ‘What is happening’.
- Data Science goes one step ahead.
- It also answers ‘What will happen in future’.
- Data Science is an evolution of Business Intelligence.

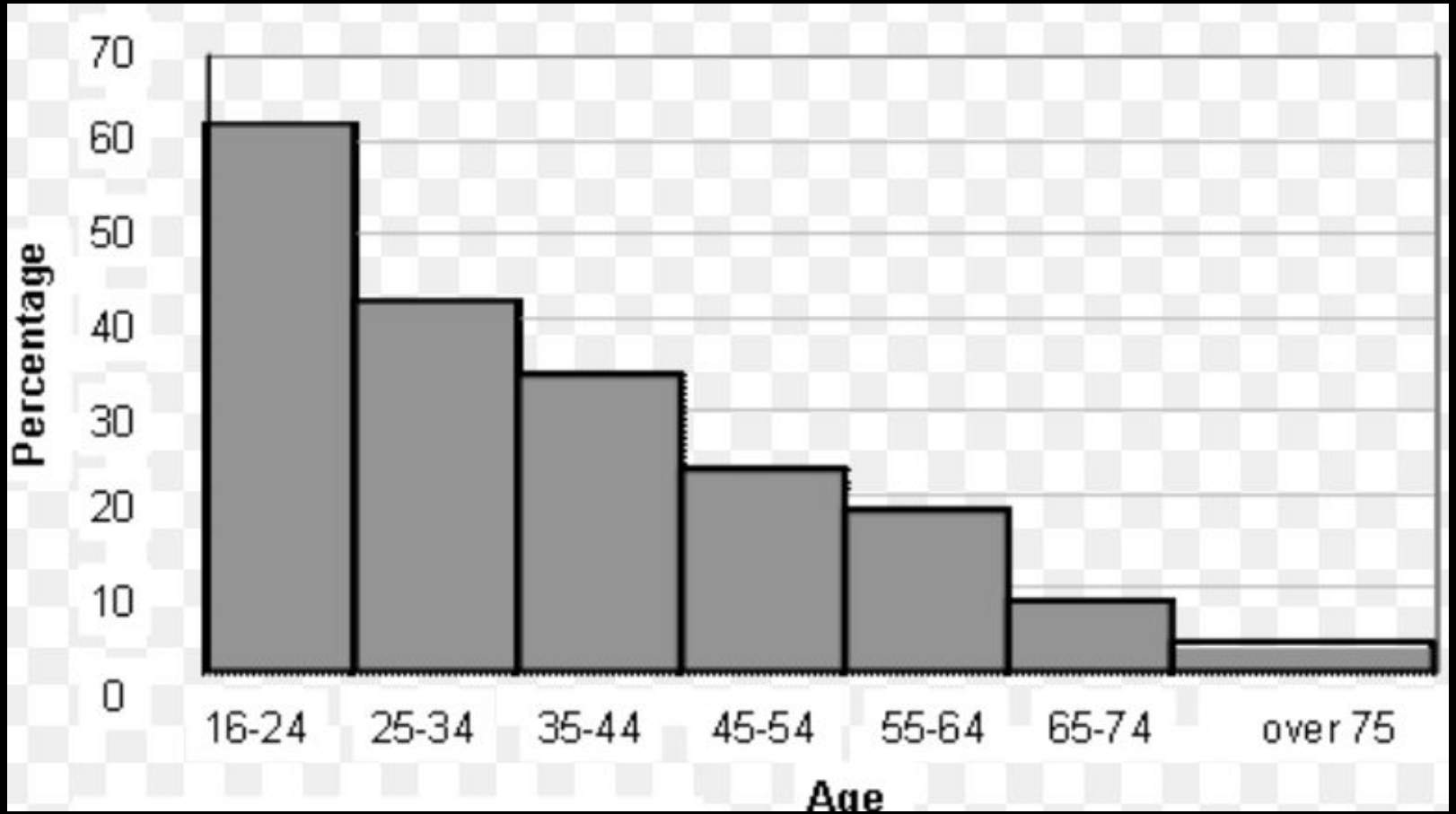
Data and Statistics:

Simple Example: Age

—

Age

- Note the data.
- Store the data.
- Summarize.
- Estimate.


70
60
50
40
30
20
10
0
Percentage
16-24
25-34
35-44
45-54
55-64
65-74
over 75
Age

Machine Learning

- The word Machine Learning comes from Computer Science.
- Use of algorithms to learn from data.
- The learning part in Data Science is Machine Learning.

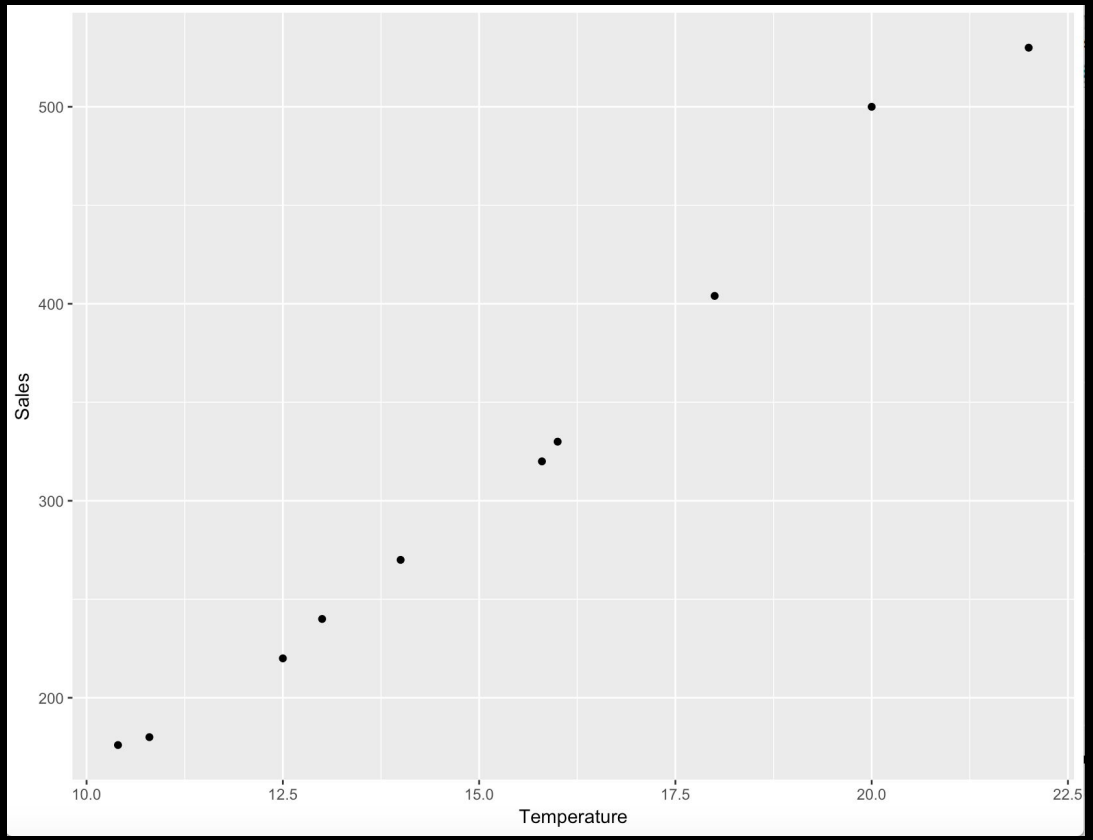
Time for some real examples!

Temperature vs Ice Cream Sales

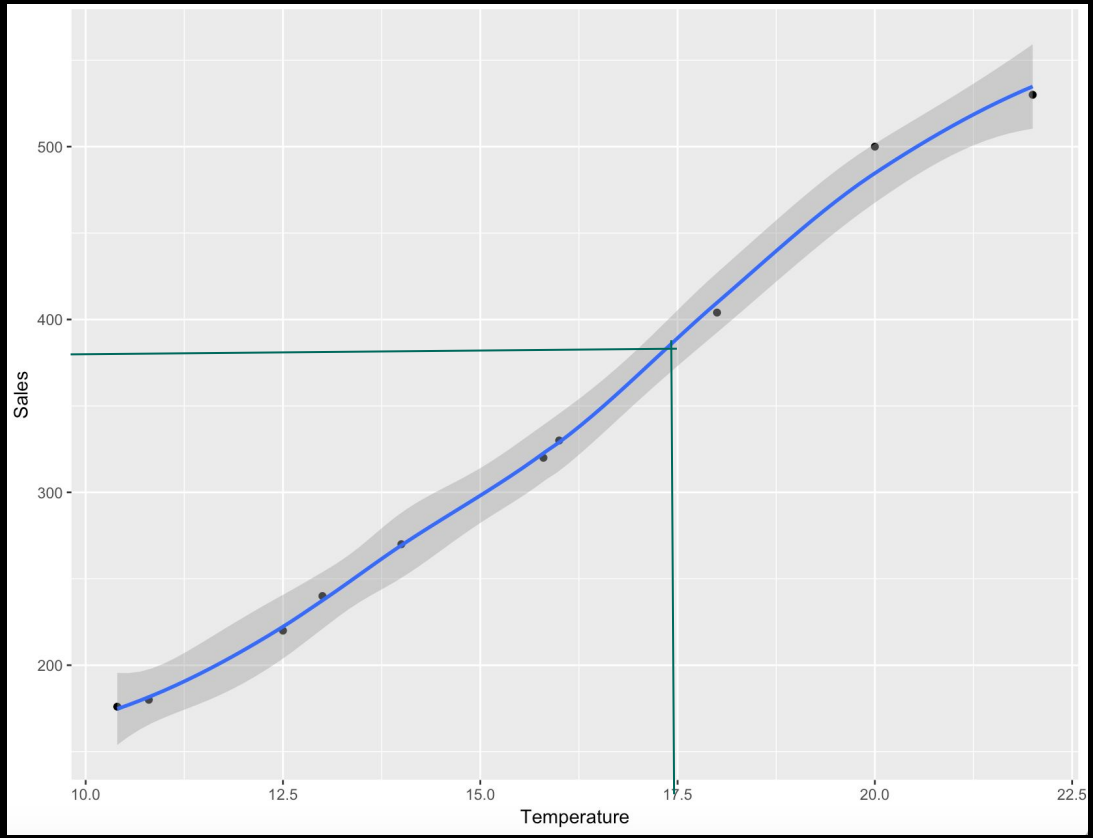
Temperature	Sales
12.5°	$220
15.8°	$320
10.8°	$180
10.4°	$176
18°	$404
16°	$325
20°	$500
13°	$240
14°	$260
22°	$530

Temperature	Sales
10.4°	$176
10.8°	$180
12.5°	$220
13°	$240
14°	$260
15.8°	$320
16°	$325
18°	$404
20°	$500
22°	$530

Temperature vs Ice Cream Sales

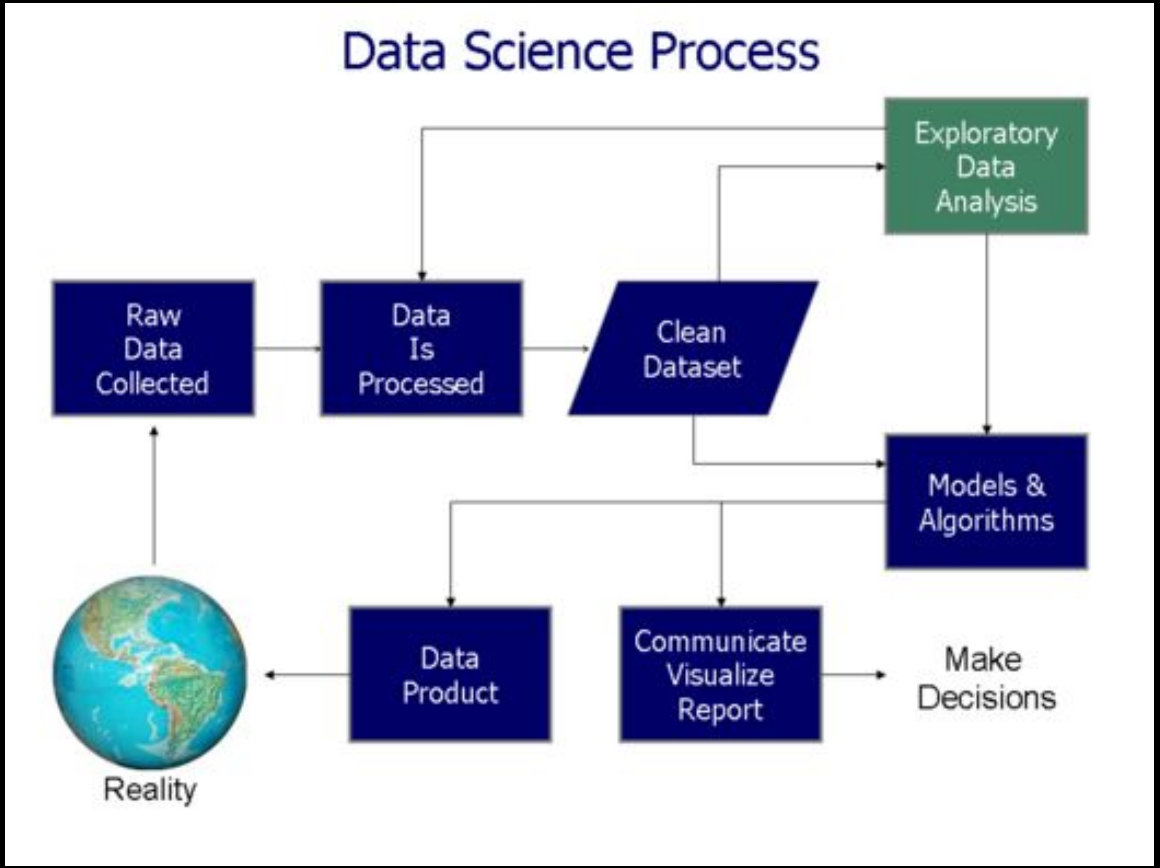


Temperature vs Ice Cream Sales



Data Science Process

- Get Data.
- Clean Data.
- Model Data.
- Visualize it.


Data Science Process
Exploratory Data Analysis
Raw Data Collected
Data Is Processed
Clean Dataset
Models & Algorithms
Reality
Data Product
Communicate Visualize Report
Make Decisions

Data Science = Little bit of (Math + Coding + Common Sense)

But wait, why do we need Math anyway?

Because Math is super boring!

Company A

Company B

Very Good Salary

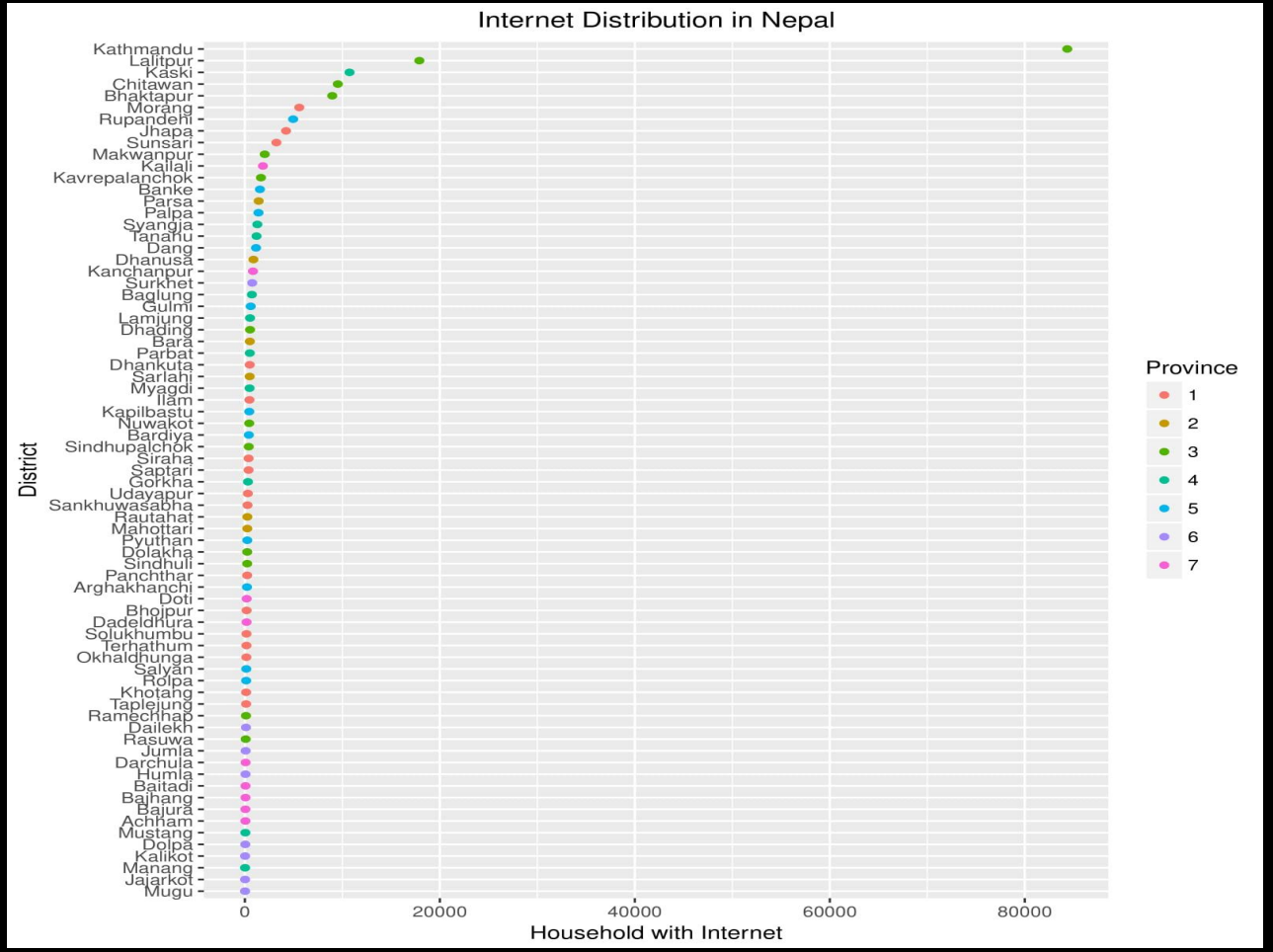
Okayish Salary

Company A	Company B
Average Salary : 50,000	Average Salary : 35,000

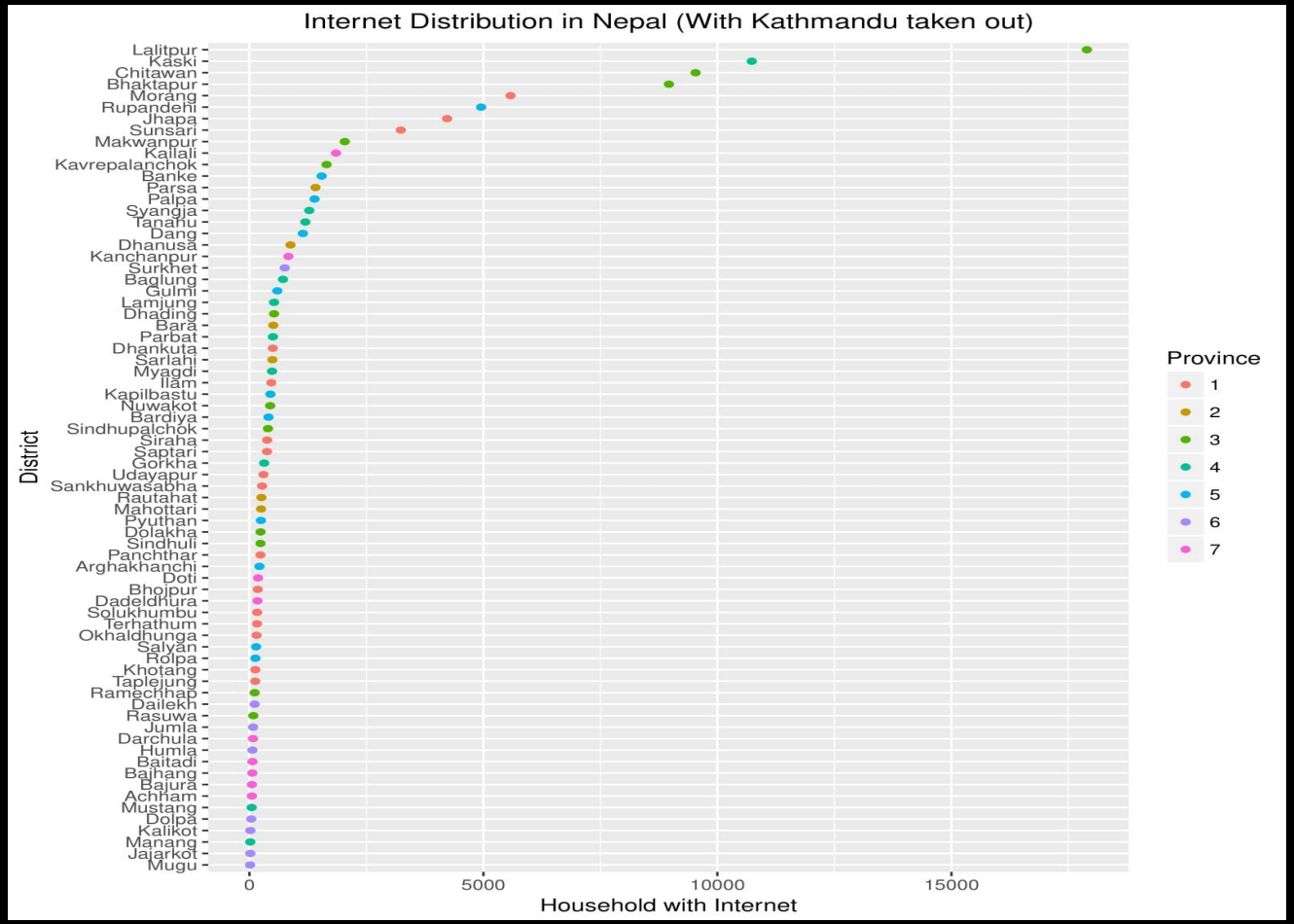
Company A		Company B
10,000		25,000
10,000		25,000
10,000		30,000
20,000		30,000
20,000		30,000
15,000		30,000
15,000		40,000
10,000		40,000
2,00,000		40,000
2,00,000		50,000

That's when Statistics will save you!

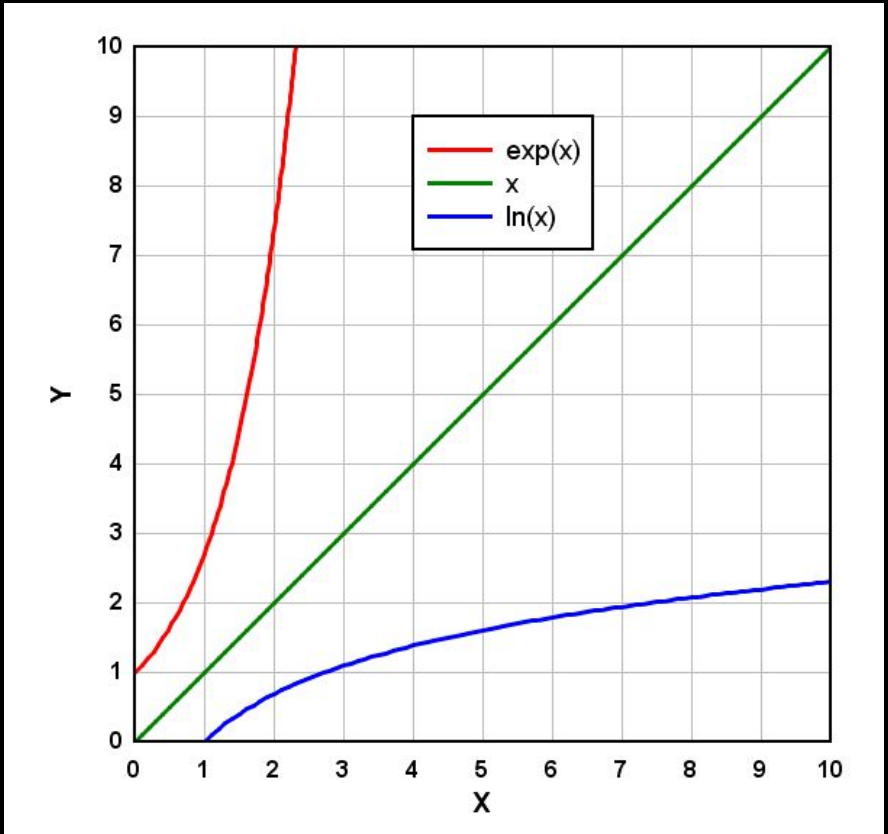
Internet Distribution in Nepal



Internet Distribution in Nepal

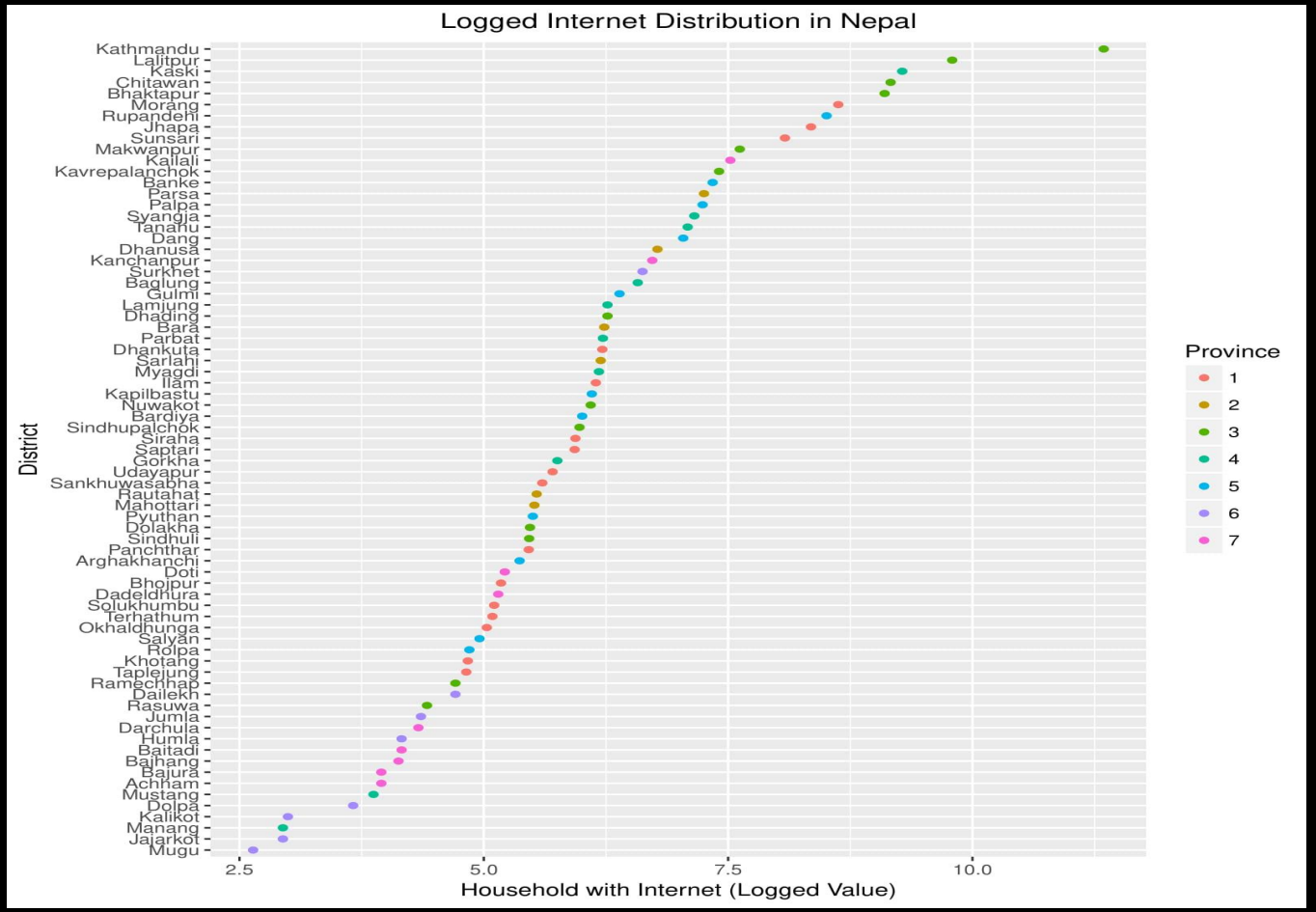


LOGARITHMIC SCALE



Taking log scale shows dispersion in smaller values and compression in larger values.

Internet Distribution in Nepal



If you have lung cancer, the probability of you smoking is 80%.

The probability of you having a lung cancer if you smoke.

10%

And here comes Probability as your friend!

Statistics + Probability + Linear Algebra + Calculus

Now comes the Programming part!

—

Tools for Data Science

- R
- Python
- SPSS
- SAS

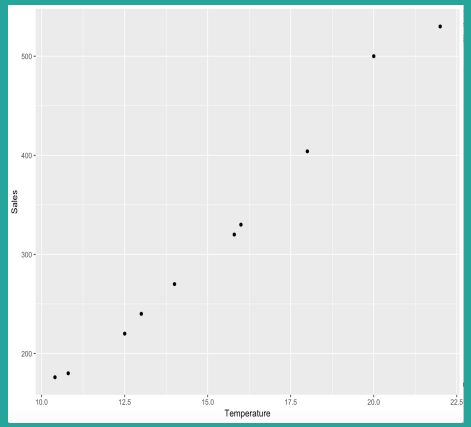
A Programming Language known as R.

Why R?

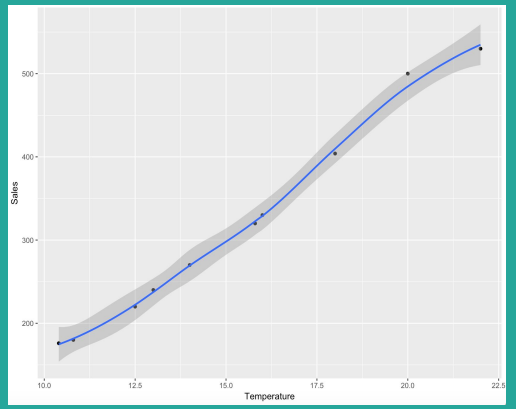
- Programming language for Data Science and Statistics.
- It is the easiest language ever.
- Perfect for non programmers.
- Strong community.
- It is free.

Temperature	Sales
12.5°	$220
15.8°	$320
10.8°	$180
10.4°	$176
18°	$404
16°	$325
20°	$500
13°	$240
14°	$260
22°	$530

1 line code!



Temperature	Sales
10.4°	$176
10.8°	$180
12.5°	$220
13°	$240
14°	$260
15.8°	$320
16°	$325
18°	$404
20°	$500
22°	$530



Where do I fit in?

- Computer Programmer
- Business Analyst
- Social Sciences (NGO/INGO)
- Research (Further Studies)
- Hobbyist

Flexibility is the reason why I pursued Data Science.

Course Structure

To get a clear picture of our whole semester!

Course Structure:

- Total 48 hours (24 Lectures)
- Theory : 32 hours
- Programming : 16 hours

Theory Structure:

- Statistics : 4 Lectures
- Machine Learning : 8 Lectures
- Graph and Networks : 2 Lectures
- Text Mining & NLP : 2 Lectures

Statistics:

- Basic Statistics
- Distributions
- Causality
- Hypothesis Testing

Machine Learning:

- Linear Algebra & Calculus
- Linear Regression
- Gradient Descent
- Logistic Regression
- Clustering (Unsupervised Learning)

Programming Structure:

- Basic R
- Data Manipulation in R
- Data Visualisation in R
- Machine Learning in R
- Graph & Network Analysis in R
- Text Mining in R

Topics we won’t cover!

Topics we won’t cover :

- Big Data
- Deep Learning
- Econometric Models

Project!

Project Structure:

- Use of topic and a real dataset
- All Data Science pipeline
- At least 3 Machine Learning algorithms or Statistical Modeling or Graph Analysis or Text Analysis

Project Structure:

- Report
- Presentation
- 3 Milestones : 20% - 20% - 60%

Approach

Concepts vs Formulae

End Note!

Thank You!