FINDING THE BEST DATA SCIENCE BOOKS

Tanmay Grandhisiri

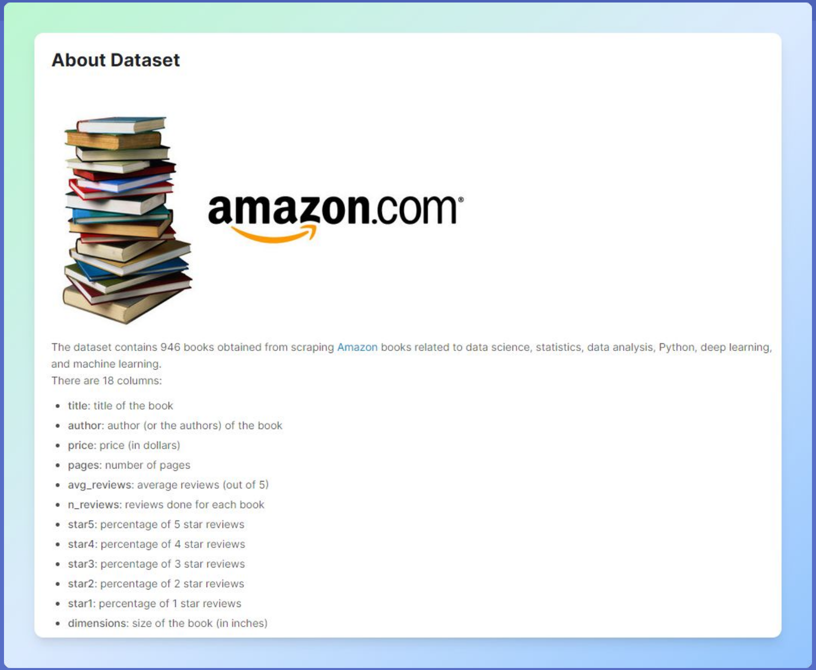
“Evaluating what kind of books should data science students choose from during their learning journey?”

QUESTIONS THAT I AM EXPLORING?

QUESTIONS

- Do more expensive books have better reviews?
- Is it always true that longer books are more expensive?
- What are the best Python books?
- What are the best Machine Learning books?
- What types of books should I look for in order to build my skills in Data Science?

DATASET USED



About Dataset

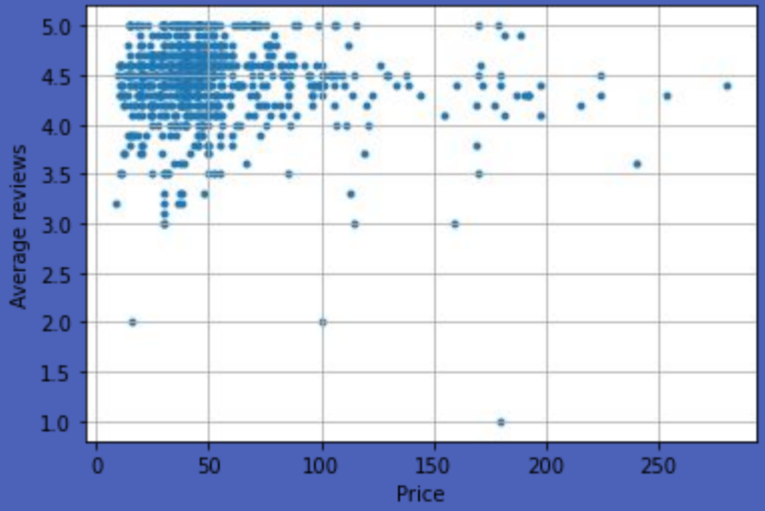
The dataset contains 946 books obtained from scraping Amazon books related to data science, statistics, data analysis, Python, deep learning, and machine learning.
There are 18 columns:

- **title**: title of the book
- **author**: author (or the authors) of the book
- **price**: price (in dollars)
- **pages**: number of pages
- **avg_reviews**: average reviews (out of 5)
- **n_reviews**: reviews done for each book
- **star5**: percentage of 5 star reviews
- **star4**: percentage of 4 star reviews
- **star3**: percentage of 3 star reviews
- **star2**: percentage of 2 star reviews
- **star1**: percentage of 1 star reviews
- **dimensions**: size of the book (in inches)

DO MORE EXPENSIVE BOOKS HAVE BETTER REVIEWS ?

NO !

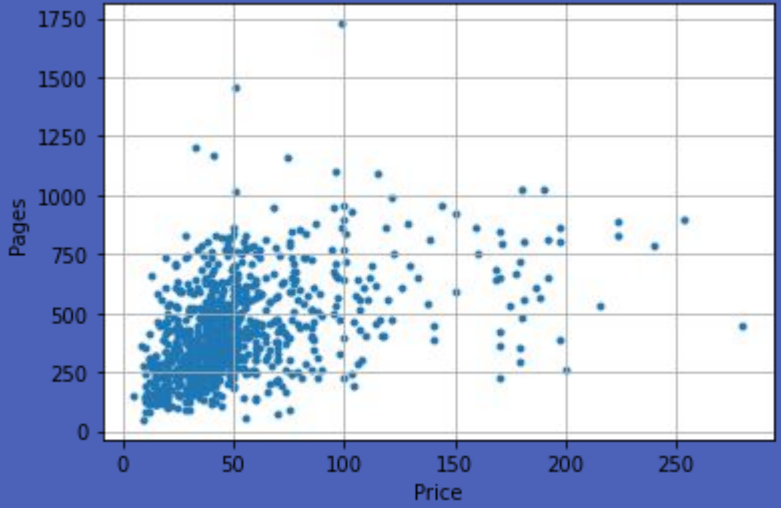
- Lower Price -> More affordable
- More affordable -> More reviews
- More reviews -> Higher chance of getting good reviews
- Affordable books have better reviews



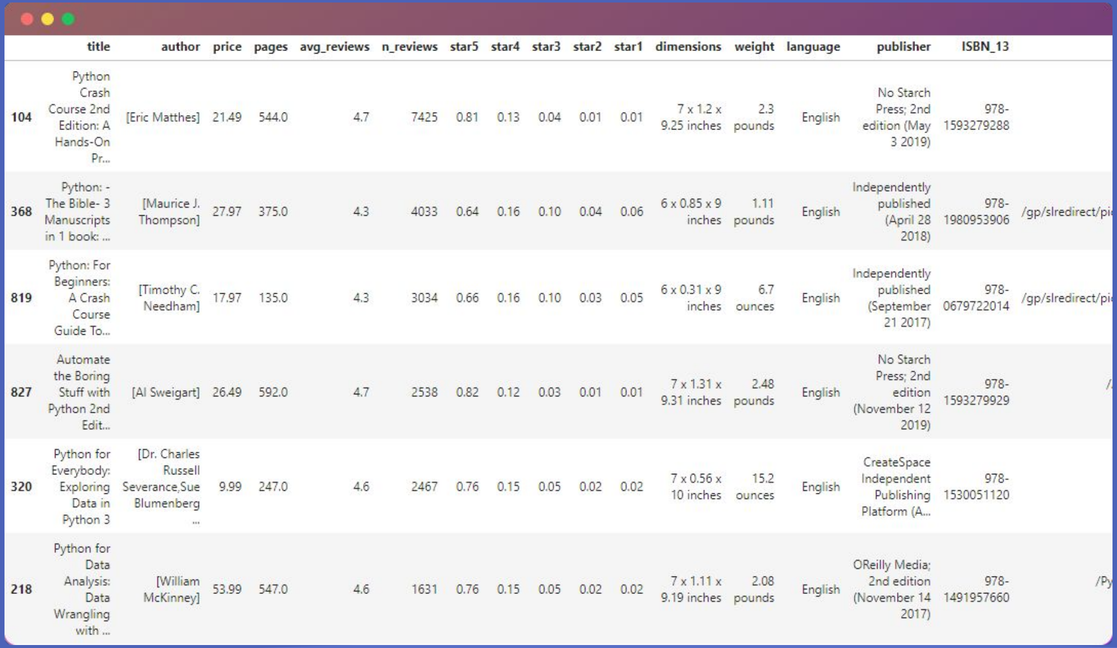
IS IT ALWAYS TRUE THAT LONGER BOOKS ARE MORE EXPENSIVE?

YES !

- Positive correlation
- Longer book = Higher Price
- Longer books can take more time to develop
- Higher price

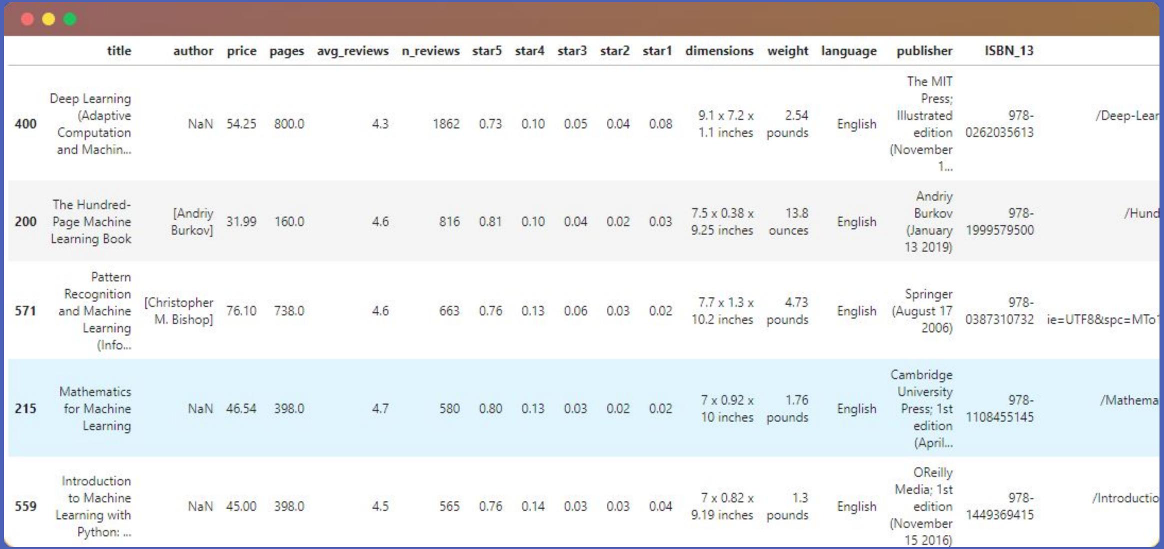


WHAT ARE THE BEST PYTHON BOOKS ?



	title	author	price	pages	avg_reviews	n_reviews	star5	star4	star3	star2	star1	dimensions	weight	language	publisher	ISBN_13	
104	Python Crash Course 2nd Edition: A Hands-On Pr...	[Eric Matthes]	21.49	544.0	4.7	7425	0.81	0.13	0.04	0.01	0.01	7 x 1.2 x 9.25 inches	2.3 pounds	English	No Starch Press; 2nd edition (May 3 2019)	978-1593279288	
368	Python: - The Bible- 3 Manuscripts in 1 book: ...	[Maurice J. Thompson]	27.97	375.0	4.3	4033	0.64	0.16	0.10	0.04	0.06	6 x 0.85 x 9 inches	1.11 pounds	English	Independently published (April 28 2018)	978-1980953906	/gp/slredirect/pi
819	Python: For Beginners: A Crash Course Guide To...	[Timothy C. Needham]	17.97	135.0	4.3	3034	0.66	0.16	0.10	0.03	0.05	6 x 0.31 x 9 inches	6.7 ounces	English	Independently published (September 21 2017)	978-0679722014	/gp/slredirect/pi
827	Automate the Boring Stuff with Python 2nd Edit...	[Al Sweigart]	26.49	592.0	4.7	2538	0.82	0.12	0.03	0.01	0.01	7 x 1.31 x 9.31 inches	2.48 pounds	English	No Starch Press; 2nd edition (November 12 2019)	978-1593279929	/
320	Python for Everybody: Exploring Data in Python 3	[Dr. Charles Russell Severance,Sue Blumenberg ...	9.99	247.0	4.6	2467	0.76	0.15	0.05	0.02	0.02	7 x 0.56 x 10 inches	15.2 ounces	English	CreateSpace Independent Publishing Platform (A...	978-1530051120	
218	Python for Data Analysis: Data Wrangling with ...	[William McKinney]	53.99	547.0	4.6	1631	0.76	0.15	0.05	0.02	0.02	7 x 1.11 x 9.19 inches	2.08 pounds	English	OReilly Media; 2nd edition (November 14 2017)	978-1491957660	/Py

WHAT ARE THE BEST MACHINE LEARNING BOOKS ?



	title	author	price	pages	avg_reviews	n_reviews	star5	star4	star3	star2	star1	dimensions	weight	language	publisher	ISBN_13	
400	Deep Learning (Adaptive Computation and Machin...	NaN	54.25	800.0	4.3	1862	0.73	0.10	0.05	0.04	0.08	9.1 x 7.2 x 1.1 inches	2.54 pounds	English	The MIT Press; Illustrated edition (November 1...	978-0262035613	/Deep-Lear
200	The Hundred-Page Machine Learning Book	[Andriy Burkov]	31.99	160.0	4.6	816	0.81	0.10	0.04	0.02	0.03	7.5 x 0.38 x 9.25 inches	13.8 ounces	English	Andriy Burkov (January 13 2019)	978-1999579500	/Hund
571	Pattern Recognition and Machine Learning (Info...	[Christopher M. Bishop]	76.10	738.0	4.6	663	0.76	0.13	0.06	0.03	0.02	7.7 x 1.3 x 10.2 inches	4.73 pounds	English	Springer (August 17 2006)	978-0387310732	ie=UTF8&spc=MTo1
215	Mathematics for Machine Learning	NaN	46.54	398.0	4.7	580	0.80	0.13	0.03	0.02	0.02	7 x 0.92 x 10 inches	1.76 pounds	English	Cambridge University Press; 1st edition (April...	978-1108455145	/Mathema
559	Introduction to Machine Learning with Python: ...	NaN	45.00	398.0	4.5	565	0.76	0.14	0.03	0.03	0.04	7 x 0.82 x 9.19 inches	1.3 pounds	English	OReilly Media; 1st edition (November 15 2016)	978-1449369415	/Introductio

WHAT TYPES OF BOOKS SHOULD I LOOK FOR IN ORDER TO BUILD MY SKILLS IN DATA SCIENCE?

Steps to answer the question:

Categories	The categories of books available have to be found in order to find the types of books
Model	A model to be found to categorize the books
Features	Features must be selected based on which the books will be categorized
Libraries	Appropriate libraries for the model have to be found
Results	The results obtained from the model have to make sense and should be presentable

THE MODEL I WILL BE USING IS ________________

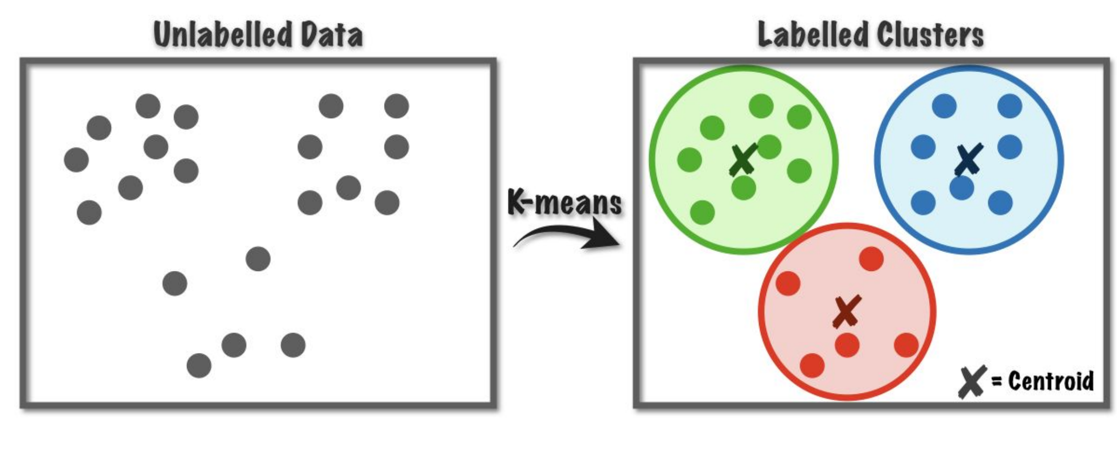
K-MEANS CLUSTERING

Objective of clustering is to find interesting patterns within the data not to make any predictions

K-means is an iterative algorithm that just randomly initializes centroids/centers for the clusters in the dataset

Clustering is an unsupervised machine learning technique that divides the entire data into groups of data such that each data points are similar to the other data points

K-MEANS CLUSTERING



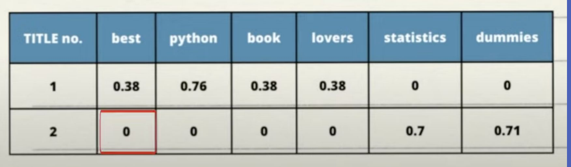
Obstacles with K-Means clustering:

- Book titles have to be converted into Numeric features
- The optimal number of clusters have to be found

SOLUTIONS

TEXT VECTORIZATION

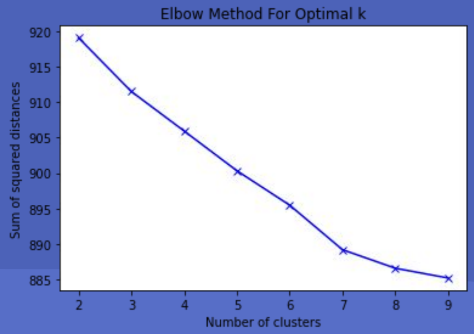
Convert book title into array of numbers



TITLE no.	best	python	book	lovers	statistics	dummies
1	0.38	0.76	0.38	0.38	0	0
2	0	0	0	0	0.7	0.71

ELBOW METHOD

This is for finding optimal number of clusters



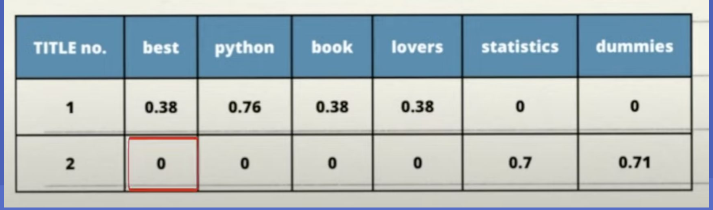
TF-IDF VECTORIZER

Frequency of **x** in **y**

Total **number** of documents

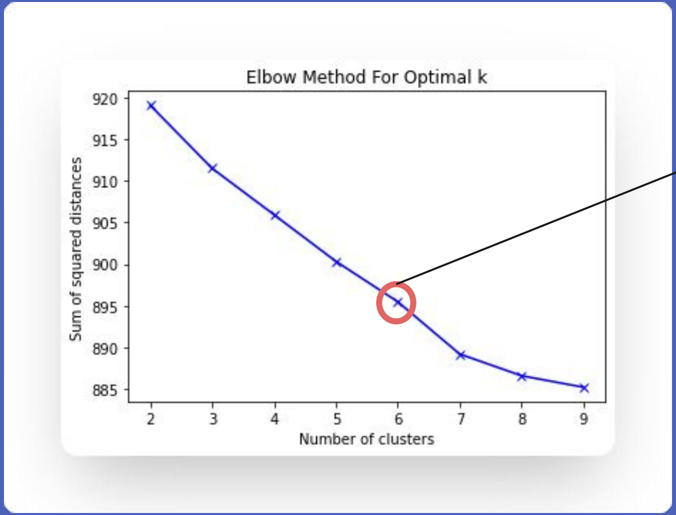
$$W_{(x,y)} = tf_{(x,y)} \times \log\left(\frac{N}{df_x}\right)$$

Number of documents containing **x**



TITLE no.	best	python	book	lovers	statistics	dummies
1	0.38	0.76	0.38	0.38	0	0
2	0	0	0	0	0.7	0.71

ELBOW METHOD

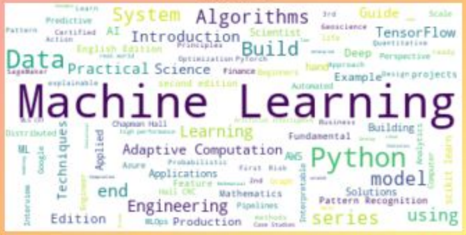


Point after which the curve starts to **flatten**

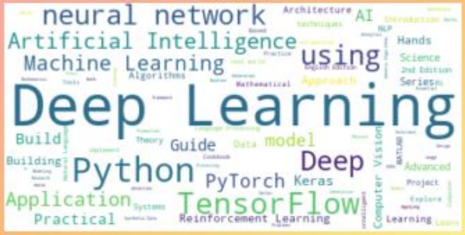
This is the optimal number of **clusters (6)**

CLUSTERS VISUALIZATION

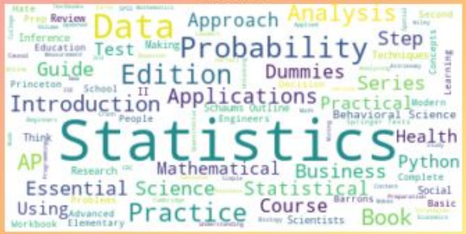
Cluster 1



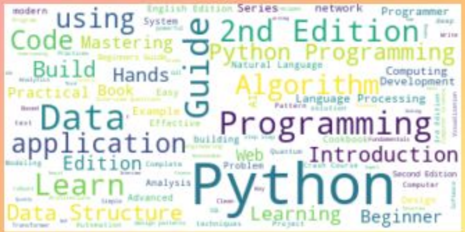
Cluster 4



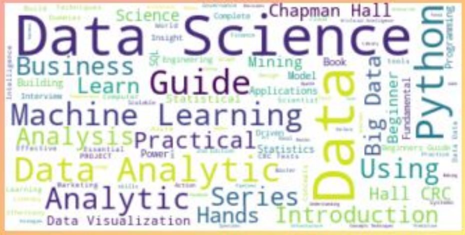
Cluster 2



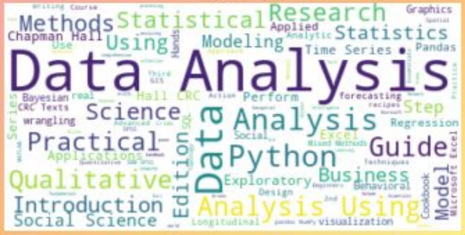
Cluster 5



Cluster 3



Cluster 6



THANK YOU