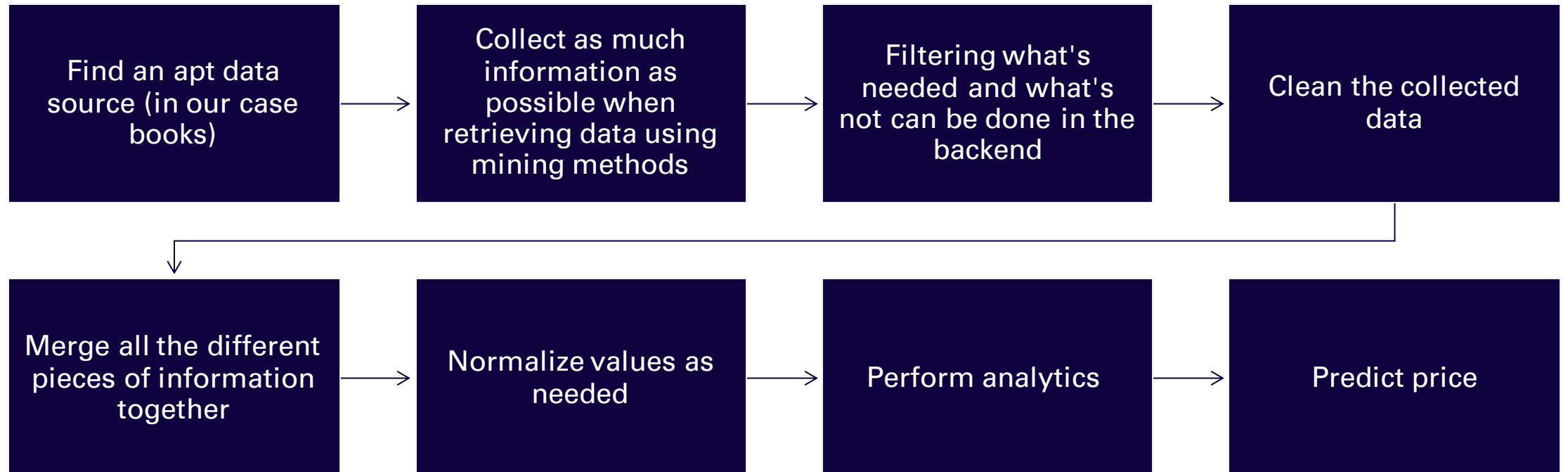


# BIG DATA ANALYTICS

Book price prediction




# Formulating the problem





# Data Sources and Mining Methods

- Data Source
    - Amazon (~1100 data points collected)
    - [Amazon search link](#)
  - Mining methods
    - Data Miner Pro
    - Beautiful Soup
- 




# Data Sources and Mining methods

- The data was scattered in multiple web pages.
- Since DataMiner was used to scrape websites, given the free tier limit, need to make sure that all data is scraped in the first attempt
- Scraped approx. 1300 pages (75 list pages and 1200 product description pages)
- Set the search keyword as "Big Data" on amazon search bar under the subcategory "Books" (along with filters to consider only paperback books) and collected the link to each book's description page
- Using product description links collected above, scraped individual book's information (no. of pages, publisher details, authors, etc.)
- Some data were not readily available, so, in some cases (e.g. Book specifications) scraped the entire html element and decided to retrieve necessary info in the next step (Data Cleaning).



# Data preprocessing and cleaning

- Amazon suggests 6 promoted books in every list page. Since there were 75 list pages, we had 450 promoted books. These books were either duplicates of the books we scrape or were irreverent to the search.
  - There was a pattern in which these promoted books appeared on each page. First and the last 2 books and 9th and the 10th books on the list page were promoted ones
  - By finding out the pattern above, the promoted books were removed
  - As mentioned in the previous slide, some of the book specs like no. of pages, publisher, authors had to be retrieved from the raw html objects that we scraped in the previous step and this was performed using beautiful soup
  - All the different pieces of information was then merged and stored in a csv file. ISBN number of book was used extensively for joining different pieces of information
- 

# Data Normalization



Now looking back at the consolidated data and after performing from preliminary analysis, some of the discrepancies in the data had been identified and removed



E.g. There was one book that had 10 authors and was priced exorbitantly



Less than 1% of the books had authors greater than or equal to 6. Considering those books as belonging to category '6'

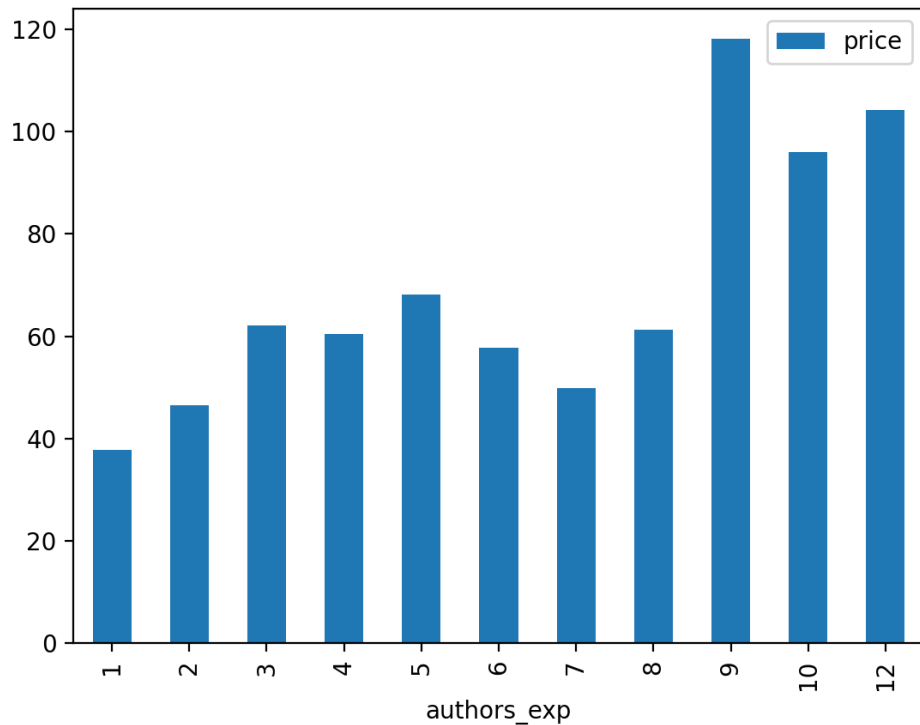


Again, there were relatively smaller number of books that had more than 950 pages. Considering all of them as belonging to 950 pages category



Since the number of pages has a wide range of values between 0 and 3000, we round off the number of pages to the nearest 100th value

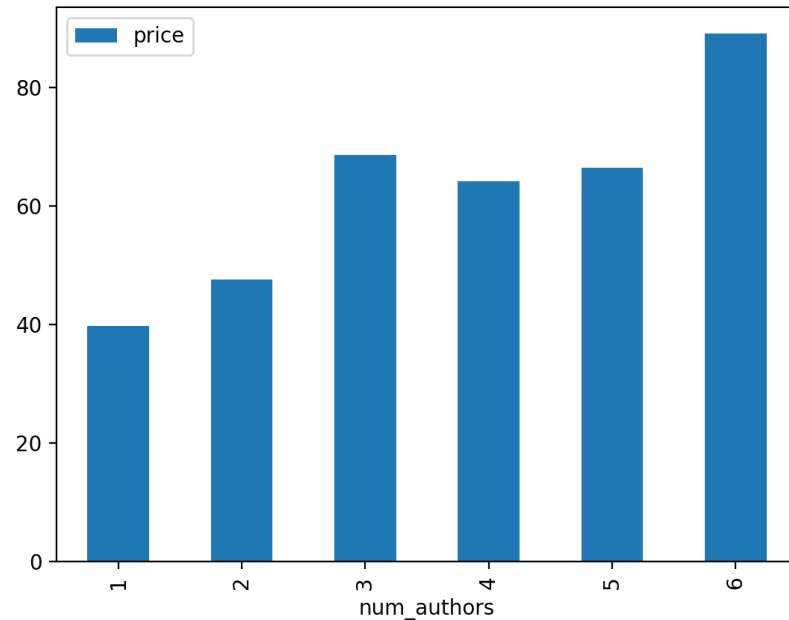
# Author Experience and Price



- As the number of books written by the authors of the book increases, the price increases until 5 and after that there is no real relationship between the author exp and the price
- Since in our case the number of books written by Jane Doe is 2 and this is the 3rd, the value to be considered is the avg price of a book written by authors who have written 3 books and that is 60\$



# No of authors and price



- As the number of authors increases, the price of the book also increases
- There is an unexplainable aberration to this rule when the number of books is 3
- Since in our case the number of authors is 1, we need not worry about the abnormality when the number of authors is 3 and therefore by this rule, the book can be priced around 40\$

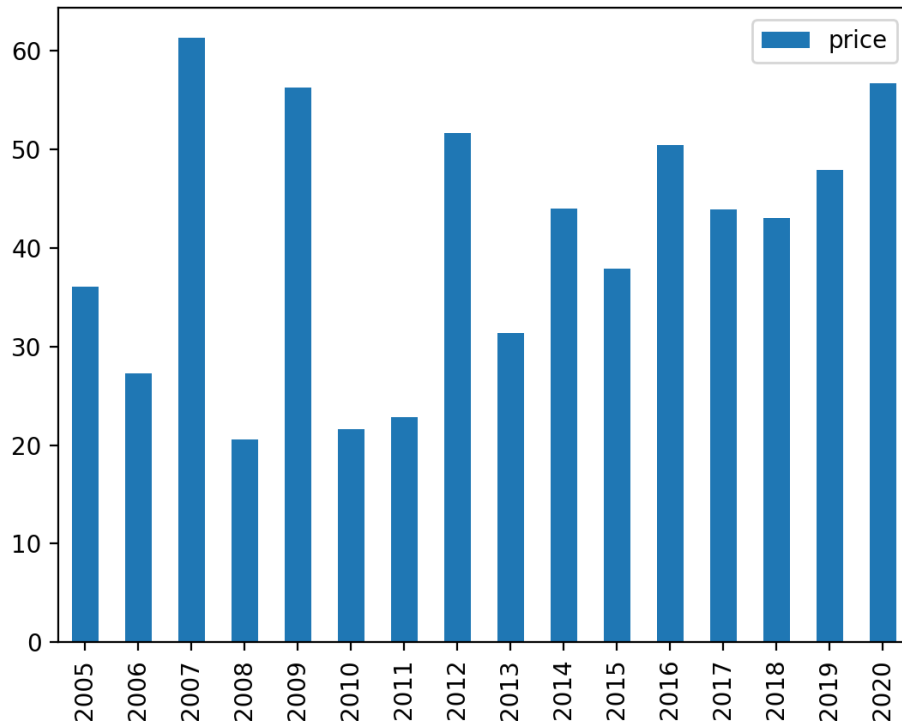


# No of pages and Book Price



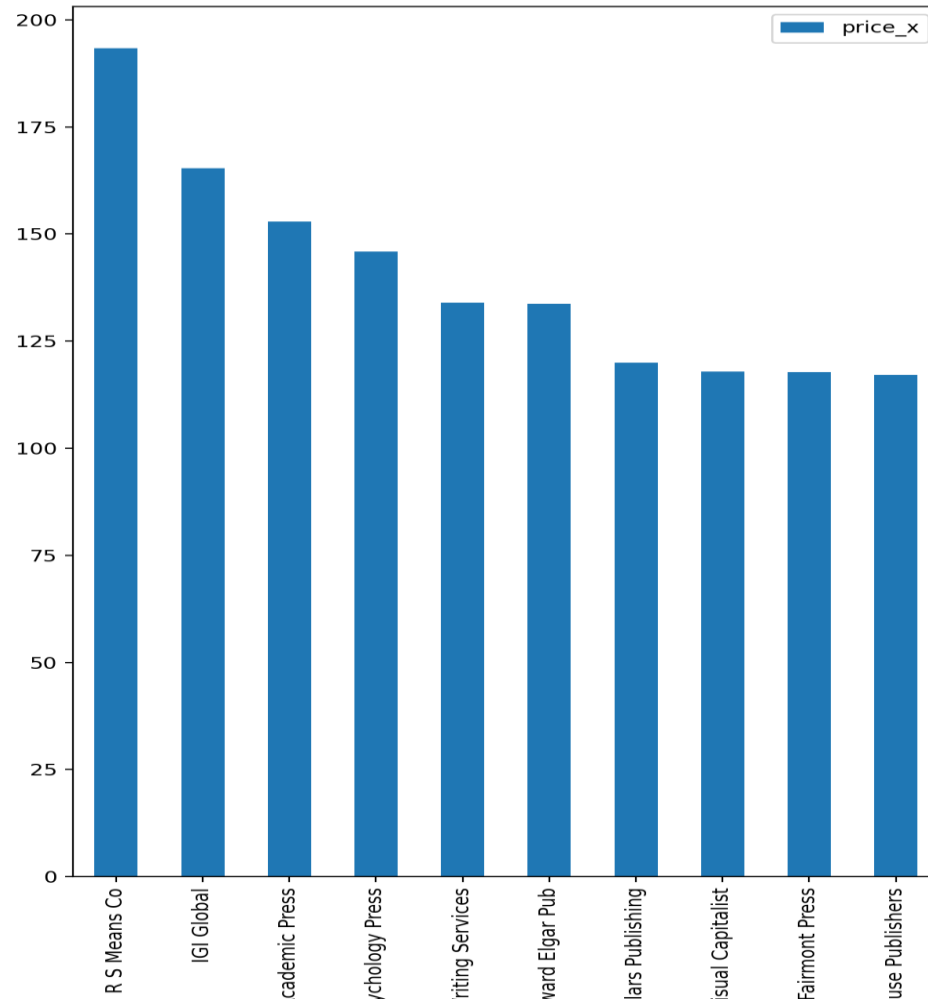
- As the number of pages increase, the price of the book increases
- There is a slight deviation to this rule when the number of pages are 400, 700 and 900
- Since in our case the number of pages is 372 and as per the round off strategy discussed in the data normalization slide, we should be considering the price of books that have 400 pages which is around 40\$

# Published Year and Book Price



- Though there is no consistent trend for price with respect to published year, over the last 3 years(2018 – 20), we see the price increases as year value increases indicating an increasing demand for big data books
- Since the book is scheduled to be released either in 2020 or 2021, as per the increasing trend, a price of 60\$ will be apt

# Publishers and Price



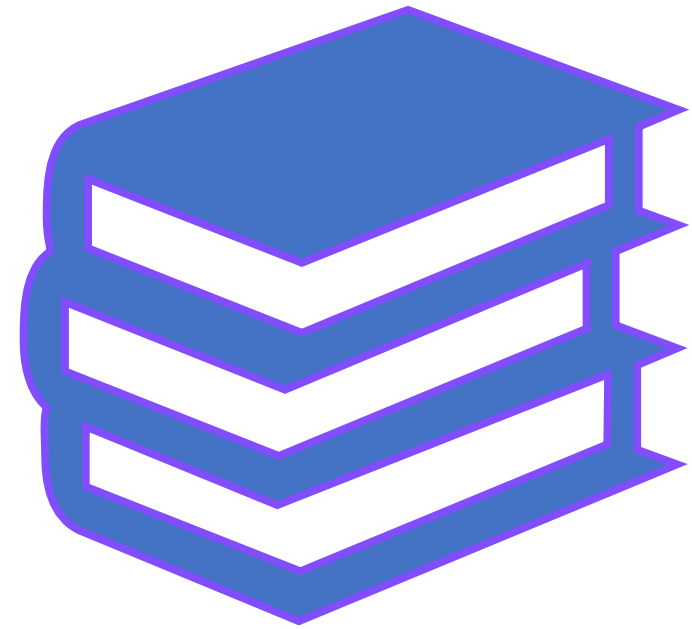
- Top 10 publishing companies who price the books highly are listed in the graph
- It's worth noticing the prices set by these companies for the book
- Since the company which is considering to publish the book is not known, it's not possible to glean much information about the price of the books set by the company or its nearest competitors
- But this graph will surely give an idea of the higher end of prices set by reputed companies and help the company to set the price based on it if needed

---

# Conclusion

- Book price predicted based on
  - Author Exp = 60\$
  - No. of authors = 40\$
  - No. of pages = 40\$
  - Published year = 60\$

Based on the above analysis, it would be apt to price the book between 47\$ and 53\$



# Limitations

- One limitation with the analysis is that we have considered only one source (Amazon).
- If the trend that we discovered in the analysis does in previous slides do not hold true for books listed in other websites, the predicted price might not be accurate
- To overcome this limitation, inclusion of datapoints from many other data sources is necessary
- Another limitation is the conclusion derived are based on the trends discovered through naked eyes. Therefore some other trends might have slipped pass.
- Incorporating some machine learning algorithms, for e.g. Linear Regression to predict target (Price) based on other dependent variable (no of pages, no of authors, author experience) could give us a more accurate estimate

