Project 2-2, Data Mining for a Book Company

Elliott Newman: Big Data Analytics:
ISE:4172

Presentation Outline

Data collection process

Summary Statistics

Visuals

Final Recommendation

Problem Statement

- I’m currently a Data Scientist working for a book company called ‘Book Company’. I’ve been asked to collect data on books from websites and use analytical techniques to estimate the price of our new book, ‘Big Data Analytics’ by Jane Doe.
- I’m giving a presentation on my data collection process, analysis of the data, and price recommendation.

Data Collection

This graphic outlines my overall data collection process at a high level.

Searched for Book Url Pages with Terms “Data Science”, “Hadoop” and “Spark” on the Barnes and Noble website.

Using Data Miner, scraped 180+ individual Book/Author Page URLs from Barnsandnoble.com

Data Miner Crawl Scrape- Parsed HTML from URLs to receive variables for analysis

Used Python/Excel to further clean and create a .csv file of the data

Data Cleaning Process

- Here's one example of a cleaning process. Throughout the entire process I used Regular Expressions, Pandas, SparkDFs, and Excel functions.

Product Details for the book "Beginning Data Science in R"

Product Details

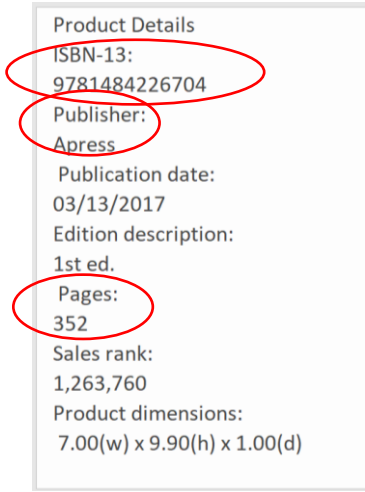
ISBN-13:	9781484226704
Publisher:	Apress
Publication date:	03/13/2017
Edition description:	1st ed.
Pages:	352
Sales rank:	1,263,760
Product dimensions:	7.00(w) x 9.90(h) x 1.00(d)

Data Miner is unable to parse each table value, so I save the entire table as one string variable, and parse the text in Python

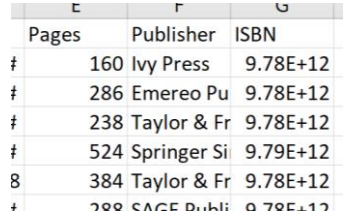
```
def getPages(df_col):
    pages = re.findall('Pages:\n(\d+)', df_col)
    if pages is None:
        return pages
    if pages:
        return int(pages[0])

def getPublisher(df_col):
    publisher = re.findall('Publisher:\n(.+)', df_col)
    if publisher is None:
        return publisher
    if publisher:
        return str(publisher[0])

def getIsbn(df_col):
    isbn = re.findall('ISBN-13:\n(\d+)', df_col)
    if isbn is None:
        return isbn
    if isbn:
        return str(isbn[0])
```

```
df['Pages1'] = df['AllInfo'].apply(getPages)
df['Publisher'] = df['AllInfo'].apply(getPublisher)
df['ISBN'] = df['AllInfo'].apply(getIsbn)
```

A string of details is parsed into several, new columns.



Product Details
ISBN-13:
9781484226704
Publisher:
Apress
Publication date:
03/13/2017
Edition description:
1st ed.
Pages:
352
Sales rank:
1,263,760
Product dimensions:
7.00(w) x 9.90(h) x 1.00(d)



	Pages	Publisher	ISBN
#	160	Ivy Press	9.78E+12
#	286	Emereo Pu	9.78E+12
#	238	Taylor & Fr	9.78E+12
#	524	Springer Si	9.79E+12
8	384	Taylor & Fr	9.78E+12

Final Data Format

Final Table Output, First 5 Rows

	ISBN	Title	Price	Year	Publisher	Author	Numb Authors	Author 1 Book Count	Pages
2	9.78E+12	Student's Solutions Manual for Statistics: The Art and Science of Learning from Data / Edition 4	$46.65	2016	Pearson Ed	Alan Agres	3	30	184
3	9.78E+12	Data Science in Practice	$149.99	2018	Springer In	Alan Said, V	2	34	195
4	9.78E+12	Hadoop in Practice: Includes 85 Techniques	$49.99	2012	Manning P	Alex Holme	1	13	536
5	9.78E+12	Apache Spark Machine Learning Blueprints	$39.99	2016	Packt Publi	Alex Liu	1	16	252
6	9.78E+12	Secrets of Statistical Data Analysis and Management Science!	$9.99	2018	Independei	Andrei Bes	1	64	52

- Other pre-processing steps included, but weren't limited to:
  - Separating Years (MM/DD/YYYY) in a Book's publishing Datetime field
  - Getting Author Info
  - Converting strings to floats (Book price '$80.50' to 80.50)
  - Removing Duplicates

Data Dimensions:
9 columns
184 rows

Summary Statistics

- To find an estimated price for the book, I first looked at the data's overall summary statistics for quantitative variables.

Independent Quantitative Variables

summary	Year	Numb Authors	Author 1 Book Count	Pages
count	184	184	102	181
mean	2016.8641304347825	1.5543478260869565	4641.823529411765	310.01657458563534
stddev	2.9118218255156694	0.8539734249942099	12719.43361060116	186.4037873702654
min	2005	1	2	22
25%	2016	1	3	170
50%	2018	1	5	288
75%	2019	2	30	400
max	2021	5	39304	1408

Target Variable, Book Price

summary	Price
count	182
mean	59.051483516483394
stddev	43.60810522148074
min	2.99
25%	32.49
50%	46.65
75%	64.99
max	219.99

Note: There are some missing data, especially for Author Book Count. I will discuss this more in the "limitations and Considerations" slide.

Subsetting our Data

- The summary stats give general info for all data cases, but I wanted to further investigate. Using data filtering, I subset the data into books with similar features. I was given 5 results.

Price	Published Year	Publisher	Author	Numb. Authors	Author Publ. Book Count	Page Count
$44.99	2018	Apress	Hien Luu	1	1	393
$39.95	2020	SAS Institute	James D. Miller	1	37	380
$169.99	2018	Springer Int.	Shu-Heng Chan	1	25	388
$34.99	2019	Packt. Publ.	Stephen Klosterman	1	2	374
$50.99	2019	CRC Press	Yu Ding	1	1	400

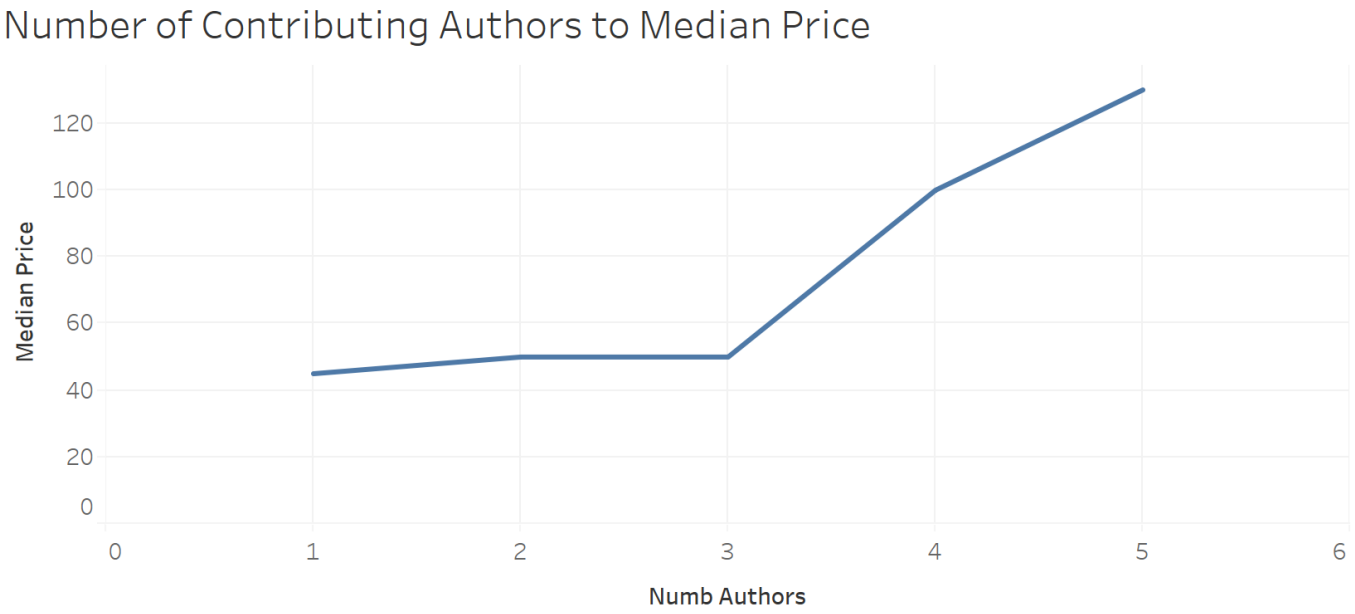
Average Price for similar books: $42.73
Median Price for similar books: $42.47

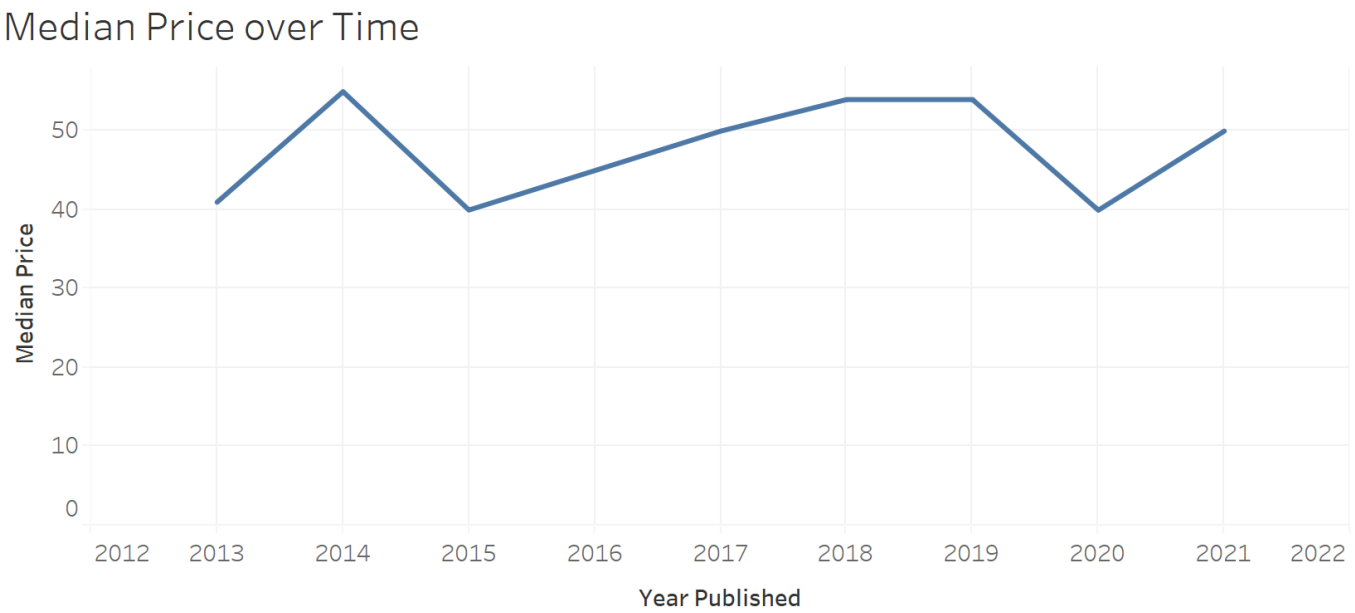
I excluded Shu-Heng Chan's book priced at $169.99. Doing some qualitative research on him, he is a Taiwanese Professor with great research experience, and founder of a research center. I assumed our author was younger and less experienced.

After my similar feature analysis I went back to the full dataset and visualized variables in comparison to book prices.

Looking at trends for Price, the median book price stays in the $40.00-$50.00 range for 1-3 contributing authors.

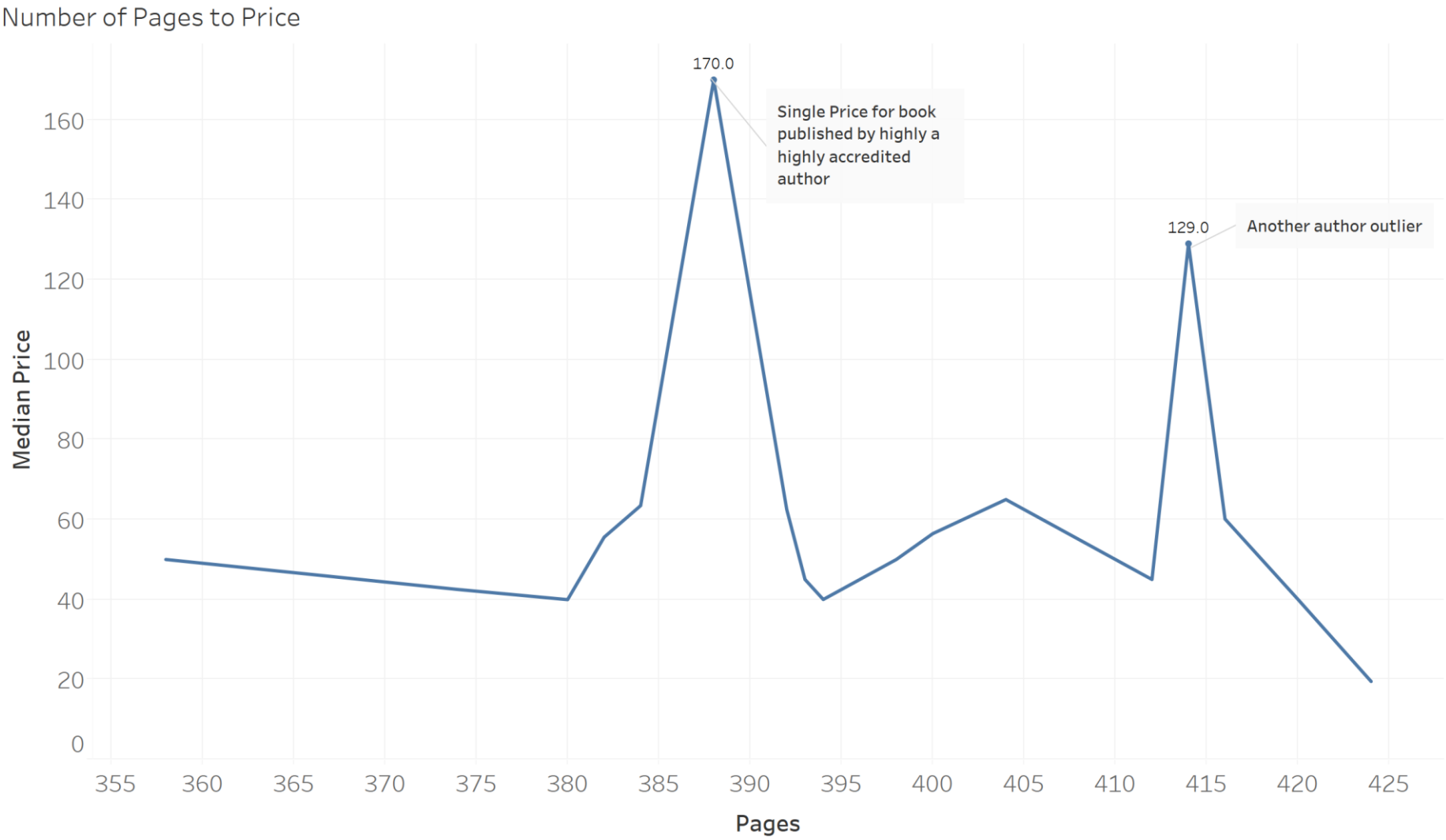
For price over time, the price of books have stayed between $40.00 to $55.00 from 8 years ago to pre-orders for 2021.





Number of Pages Compared to Price

Looking at the costs of books with similar page numbers, excluding the outliers, most median prices fall in the $40.00-$60.00 range.



Final Recommendation

- Based off my data analysis, I concluded a book pricing of $39.99. This matches a price range to 4 similar books, and the trends among the whole dataset. The data analysis suggested a price range of $40.00 - $60.00
- We have a fairly new author, and we don’t want to over-price our textbook.

Final Price recommendation: $39.99

Limitations and Considerations

- Some authors might have higher priced books due to their accreditation.
- I would use BeautifulSoup and Python instead of Data Miner to automate more of the HTML parsing process.

References

- Barnesandnoble.com
- Data Miner Chrome Plug-in
- In-Class Lectures
- Pandas Documentation
- PySpark Documentation
- Regexr.com
- Stack Overflow