Project 2-2: Data Mining for a Book Company

By: Aaron Dzaboff

Presentation Layout

PROBLEM FORMULATION

DATA MINING

DATA PREPROCESSING /CLEANING

ANALYSIS

LIMITATIONS

1.
Problem Formulation

What Are We Solving For?

Analyze competitors to find trends and similarities

Book price comparable to competitors

2.

Data Mining

- **How I collected the information**
  - **What sources were used**

What was Mined?

Sources:


Books-A-Million


Barnes and Noble

Searches For Comparable Books Included:

- Big Data Analytics
- Hadoop Distributed File System
- Apache Spark
- Supervised Learning for Big Data
- Clustering for Big Data
- Deep Neural Networks
- Ensemble Learning

What was Mined?

Mining Method and Collection

- Utilized Google Chrome's Web Scrapper Add-On
- Collected information on 300 books
- Data Collected Includes:
  - Book Title
  - Author(s)
  - Price
  - Publisher
  - Publish Date
  - Page Count
  - Number of Previous Authored Books

3.
Data Preprocessing/Cleaning

Methods to clean data

Dropped Duplicates

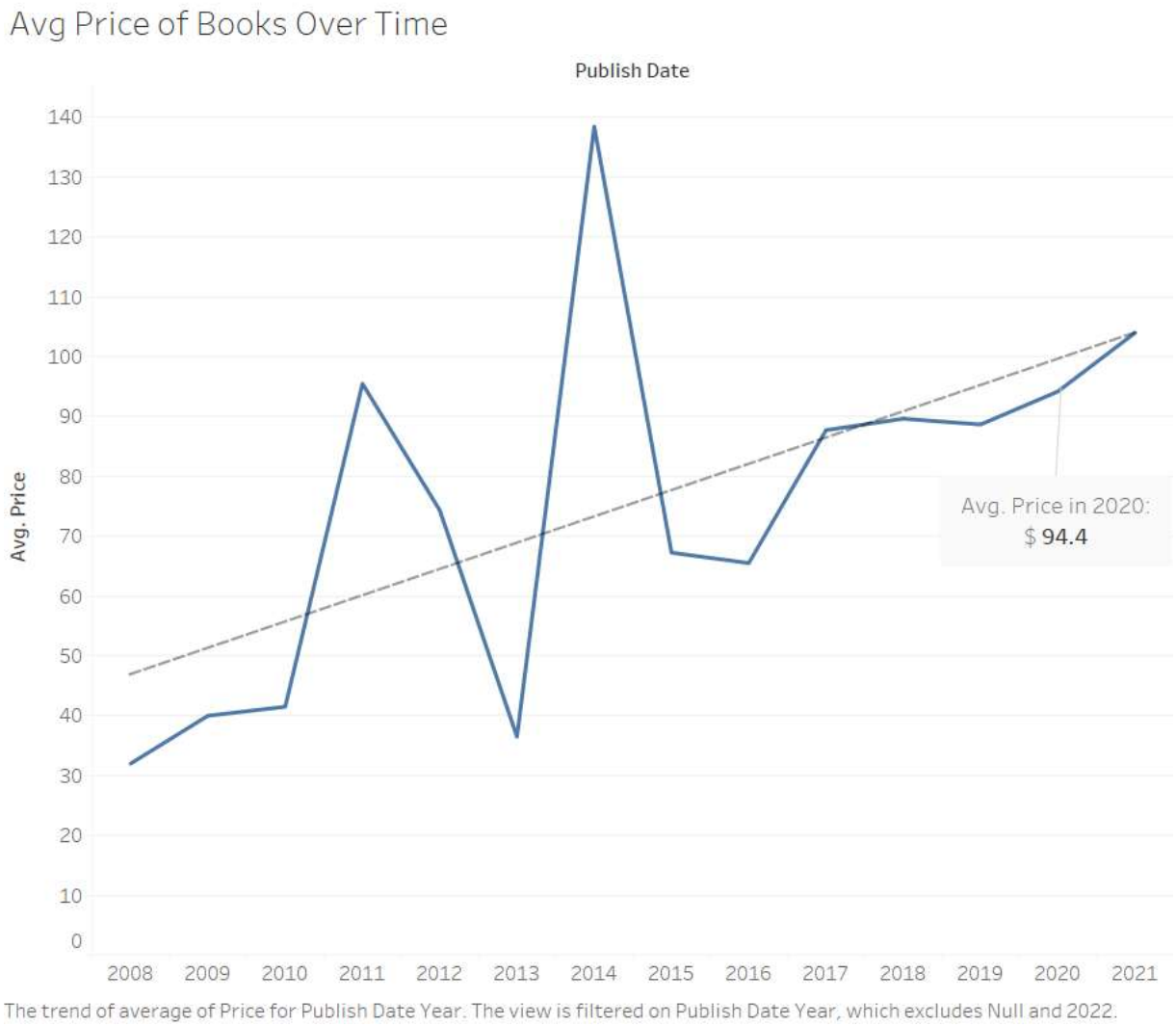
Filled NAs where applicable

Utilized splicing to format variables such as publish date, number of books previously authored, etc.

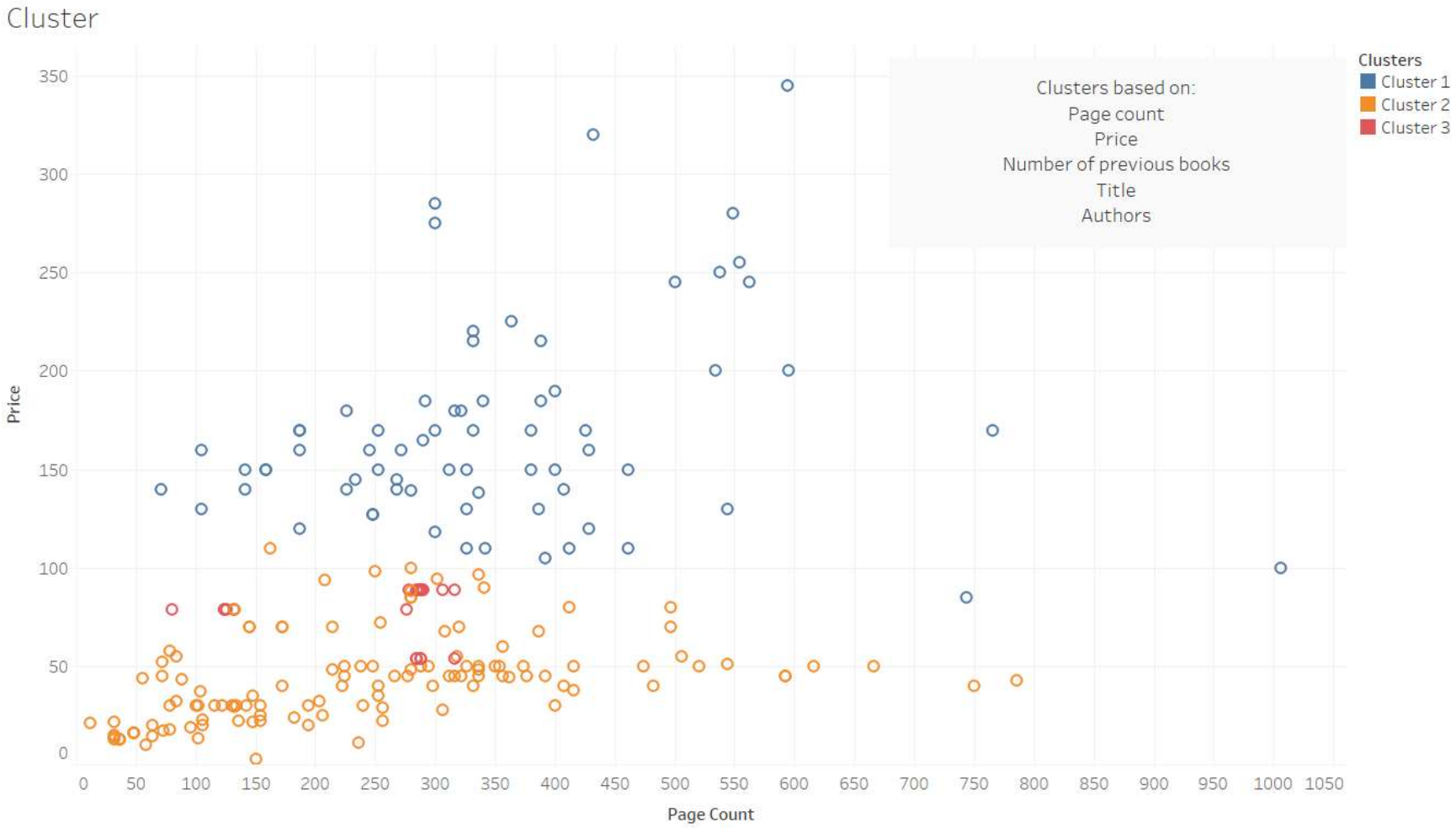
4.
Analysis

- **Methods used to find competitive price**
- **Visualizations**

Exploratory Analysis

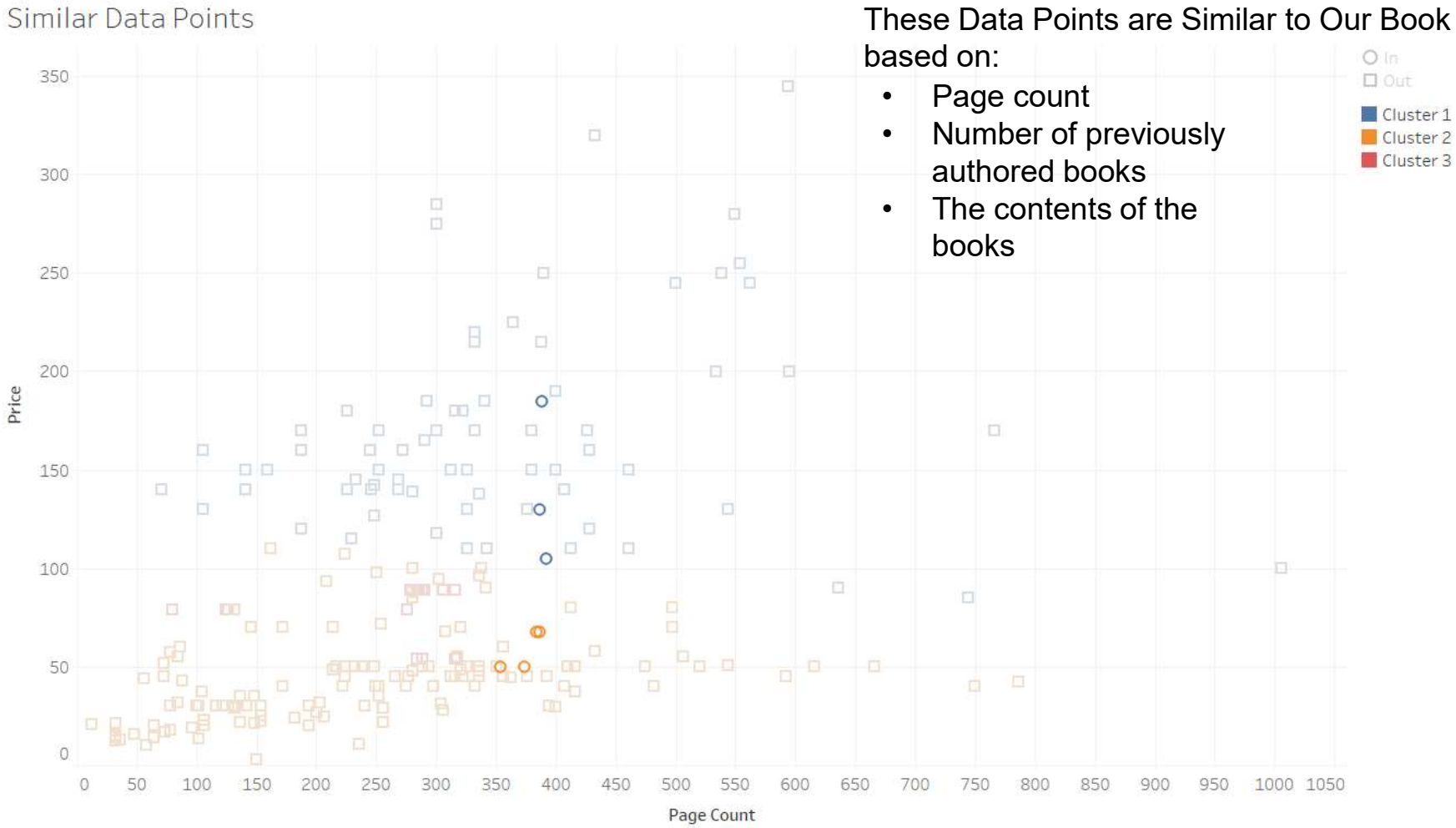


Clustering



Page Count vs. Price. Color shows details about Clusters (2). Details are shown for Author, Year of Publish Date and Title. The view is filtered on Clusters (2), which keeps Cluster 1, Cluster 2 and Cluster 3.

Where Does Our Book Fall?



These Data Points are Similar to Our Book based on:

- Page count
- Number of previously authored books
- The contents of the books

Page Count vs. Price. Color shows details about Clusters (3). Shape shows details about In / Out of Similar Data Point. Details are shown for Author, Year of Publish Date and Title. The view is filtered on Clusters (3) and In / Out of Similar Data Point. The Clusters (3) filter keeps Cluster 1, Cluster 2 and Cluster 3. The In / Out of Similar Data Point filter keeps Out and In. The view is highlighted where IN/OUT(Similar Data Point) contains "In".

Comparing the Similar Books

Similar Books Comparison to Find Average Price

Title	Author	Publish Date	Num Of Books	Page Count	Price
Deep Learning Techniques and Optimization Strategies in Big Data Analytics	J. Joshua Thomas (Editor)	October, 2019	2.0	388.0	185.0
Big Data Analytics Using Multiple Criteria Decision-Making Models	Ramakrishnan Ramanatha..	June, 2017	2.0	386.0	130.0
Big Data Analytics and Intelligence : A Perspective for Health Care	Poonam Tanwar	September, 2020	2.0	392.0	105.0
Big Data Analytics Using Multiple Criteria Decision-Making Models / Edition 1	Ramakrishnan Ramanathan	June, 2017	2.0	386.0	68.0
Harness Oil and Gas Big Data with Analytics: Optimize Exploration and Prod..	Keith R. Holdaway	May, 2014	2.0	384.0	67.5
Mastering Hadoop	Sandeep Karanth	December, 2014	2.0	374.0	50.0
Mastering Apache Spark 2.x	Romeo Kienzler	July, 2017	2.0	354.0	50.0
Grand Total					93.6

Num Of Books, Page Count and Price broken down by Title, Author and Publish Date. The data is filtered on Similar Data Point, which keeps 7 members.

The average price for the most similar books is $93.60

Should set the price of the book in the range of:

$93-$100

5.
Limitations

How Can This Project be Improved?

- Not every multiple authored books collected due to complications with web scrapping
- Collect more data (300 samples isn't a lot)
- Could collect more variables to help aid in clustering analysis

Thanks!

Any questions?