Final Exam

Report on: Analyzing Amazon bestsellers through the 10-year period

(2012-2022).

Machine Learning

Presented by: Tejaswini Yeruva

Abstract

The majority of people still favor reading as a kind of leisure since it provides a special route to information and education. Books continue to be a significant and popular cultural product as a result. However, despite the fact that over 3 million books are created each year, only a small number are widely read, and fewer than 500 make it to the Amazon bestseller lists. Only a small number of authors can maintain control of the lists once they have been there for longer than a few weeks. In this research, we apply Machine Learning techniques to the study of book success by analyzing the characteristics and progression of best-seller sales over the last 10 years.

Our understanding of the book industry and, more broadly, of how we as a society interact with cultural products depends on the analysis of bestseller qualities and the identification of the universality of sales patterns with its driving forces.

Introduction

We are examining statistics about the top-selling books on Amazon for ten years span for this study which has the data with all the booksellers. Chris Kachmar, the data owner, made it accessible to the public, with no copyrights on the Kaggle platform.

50 best-selling books on Amazon were tracked in the dataset used for the analysis below from 2009 to 2022. It is challenging to derive conclusions from the raw data and compare them over time in this study because statistics on the best-selling books over the previous 10 years are gathered annually. The goal of this study is to use the collected data to cluster the population of best-selling books and make it easier to recognize these clusters and features from the data, allowing for more effective insight as we sample these best-selling books in the future. Price, the number of reviews or readings, popular genres, and so forth are examples of such features.

We can use this knowledge to identify patterns and predict future outcomes when we possess the analytical abilities to evaluate and analyze complex data in the modern day.

Overview

Our sample data contains 550 observations with around 7 variables. The goal of this project is to use machine learning techniques like transforming the data i.e. removing all the unwanted data, visualizing, and analyzing, finding a correlation between variables, and plotting them to retrieve the data.

```
str(bestsellers_data_09_22)
```

```
[1] "C:/Users/tejar/OneDrive/Desktop/ML Assignments"
[1] "Name"        "Author"      "User.Rating" "Reviews"     "Price"       "Year"        "Genre"
'data.frame':   550 obs. of  7 variables:
 $ Name       : chr  "10-Day Green Smoothie Cleanse" "11/22/63: A Novel" "12 Rules for Life: An Antidote to Chaos" "1984 (Signet Classics)" ...
 $ Author     : chr  "JJ Smith" "Stephen King" "Jordan B. Peterson" "George Orwell" ...
 $ User.Rating: num  4.7 4.6 4.7 4.7 4.8 4.4 4.7 4.7 4.7 4.6 ...
 $ Reviews    : int  17350 2052 18979 21424 7665 12643 19735 19699 5983 23848 ...
 $ Price      : int  8 22 15 6 12 11 30 15 3 8 ...
 $ Year       : int  2016 2011 2018 2017 2019 2011 2014 2017 2018 2016 ...
 $ Genre      : chr  "Non Fiction" "Fiction" "Non Fiction" "Fiction" ...
```

Data Cleaning

To make the data easier to reach, column names are being renamed, and removing all the unwanted data from the data set is since our goal is to analyze it for the last 10 years only.

```
```{r}
bestsellers_data_09_22 <- bestsellers_data_09_22 %>%
  rename(book_name = Name,
         author_name = Author,
         user_rating = User.Rating,
         reviews_number = Reviews,
         price = Price,
         year = Year,
         genre = Genre)

glimpse(bestsellers_data_09_22)
```
```

```
Rows: 550
Columns: 7
$ book_name      <chr> "10-Day Green Smoothie Cleanse", "11/22/63: A Novel", "12 Rules for Life: An Antidot…
$ author_name    <chr> "JJ Smith", "Stephen King", "Jordan B. Peterson", "George Orwell", "National Geograp…
$ user_rating    <dbl> 4.7, 4.6, 4.7, 4.7, 4.8, 4.4, 4.7, 4.7, 4.7, 4.6, 4.6, 4.6, 4.6, 4.5, 4.6, 4.5, 4.6,…
$ reviews_number <int> 17350, 2052, 18979, 21424, 7665, 12643, 19735, 19699, 5983, 23848, 23848, 460, 4149,…
$ price          <int> 8, 22, 15, 6, 12, 11, 30, 15, 3, 8, 8, 2, 32, 5, 17, 4, 6, 6, 8, 13, 14, 14, 13, 9, …
$ year           <int> 2016, 2011, 2018, 2017, 2019, 2011, 2014, 2017, 2018, 2016, 2017, 2010, 2011, 2018, …
$ genre          <chr> "Non Fiction", "Fiction", "Non Fiction", "Fiction", "Non Fiction", "Fiction", "Ficti…
```

Data Techniques and Insights

The insights are offered below for each investigation as we employ the machine learning approaches listed below.

Data Transformation

Step 1:

Creating the variables for year-specific average values of price, rating, and number of reviews would display the median number of reviews, the mean price, and the mean rating of each year.

```
analysing_by_year <- bestsellers_data_12_22 %>%
  group_by(year) %>%
  summarize(average_rating = mean(user_rating),
            average_reviews_number = median(reviews_number), average_price = mean(price))

analysing_by_year
```

A tibble: 8 × 4

year <int>	average_rating <dbl>	average_reviews_number <dbl>	average_price <dbl>
2012	4.532	9333.5	15.30
2013	4.554	7094.0	14.60
2014	4.622	10514.0	14.64
2015	4.648	9144.0	10.42
2016	4.678	10545.0	13.18
2017	4.660	10560.5	11.38
2018	4.668	10456.0	10.52
2019	4.740	11185.0	10.08

Step 2:

Creating a variable to represent the overall number of books by genre published each year. This chart displays how many books are published a year in each genre.

year <int>	genre <chr>	books_number <int>
2012	Fiction	21
2012	Non Fiction	29
2013	Fiction	24
2013	Non Fiction	26
2014	Fiction	29
2014	Non Fiction	21
2015	Fiction	17
2015	Non Fiction	33
2016	Fiction	19
2016	Non Fiction	31

Step 3:

Creating a variable to represent the rating value by price. This displays the average rating and quantity of reviews for Price.

price <int>	average_rating <dbl>	average_reviews_number <dbl>
0	4.780000	6324.0
2	4.400000	6042.0
3	4.700000	5983.0
4	4.706897	10369.0
5	4.770588	10302.0
6	4.586667	11994.0
7	4.561111	9873.0
8	4.763636	19576.0
9	4.656000	8922.0
10	4.611111	8842.0

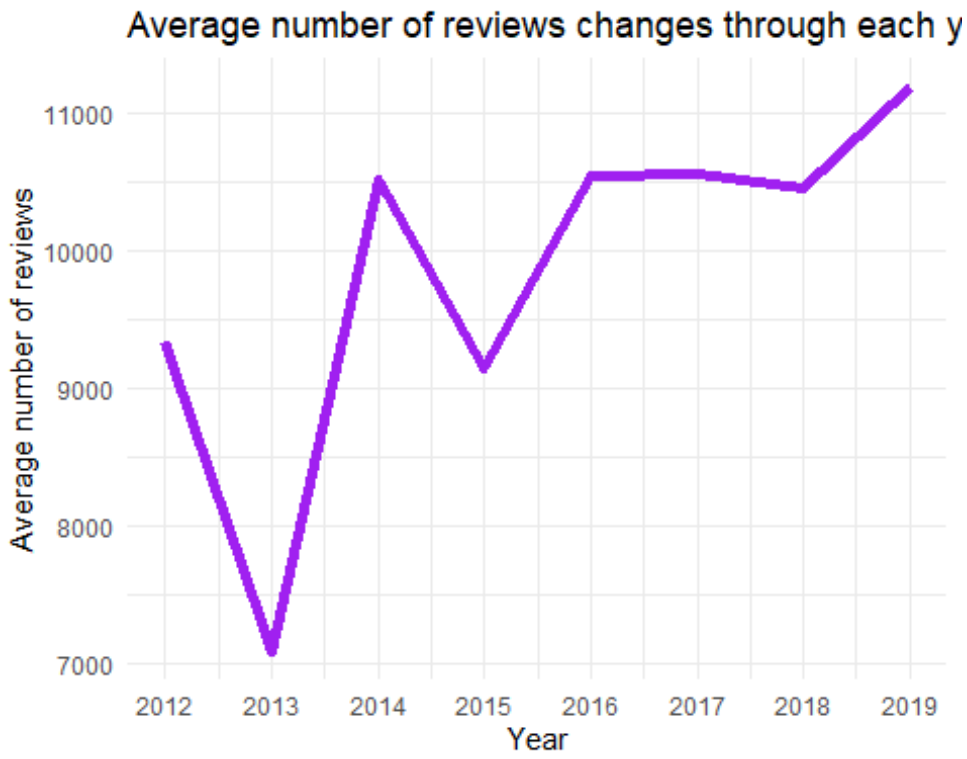
Note: The number of reviews for each book can be equivalent to the number of readings. The average number of reviews can be equivalent to the usual number of readings.

Data Visualization and Analysis

Plotting scientific data is done to show correlations between variables or to visualize variation.

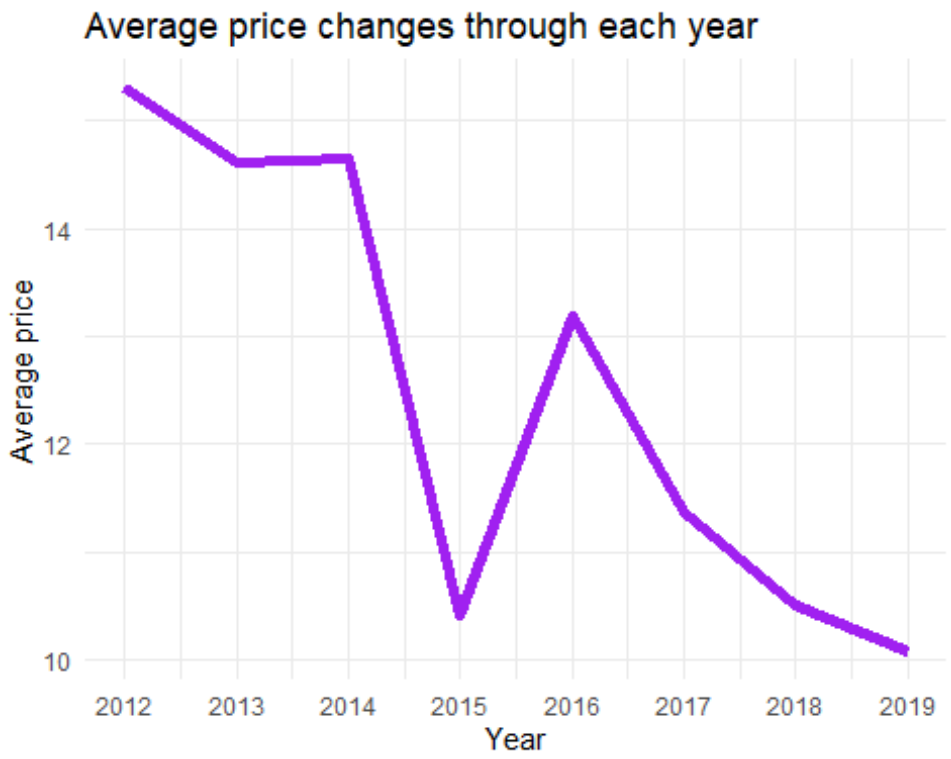
Step 1:

Examining the correlation between the year and the average amount of reviews (number of readings).

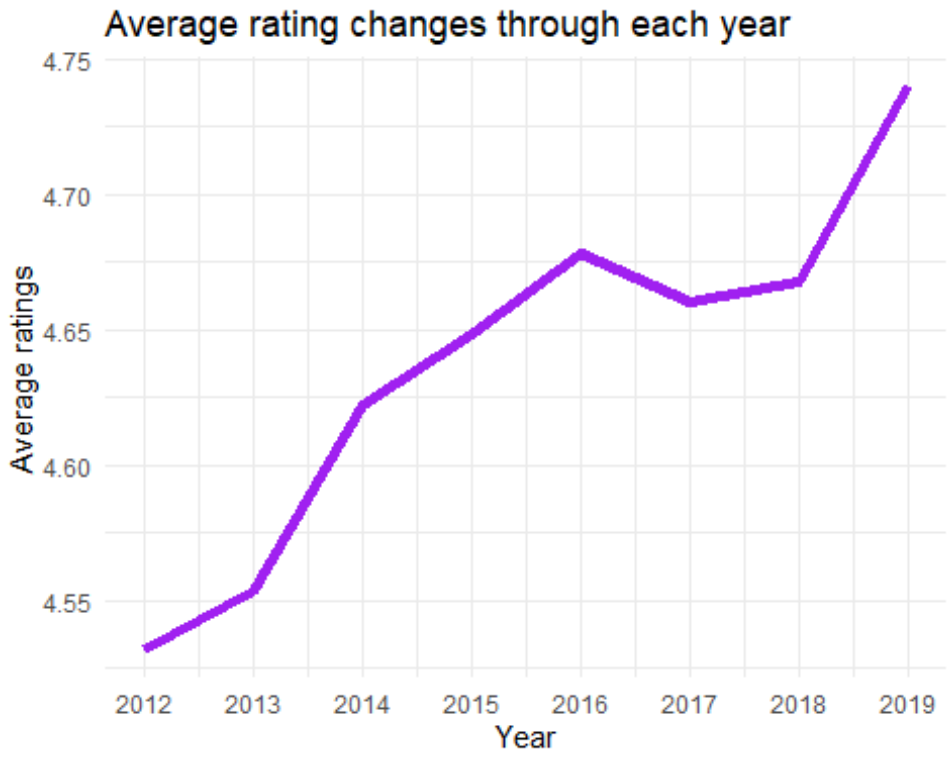


In 2020, there was a sharp increase in the number of reviews (also known as the number of book readings). We can infer that this was caused on by the COVID-19 pandemic, which began in 2020, logically.

Let's now examine how the average price and rating have changed over time.

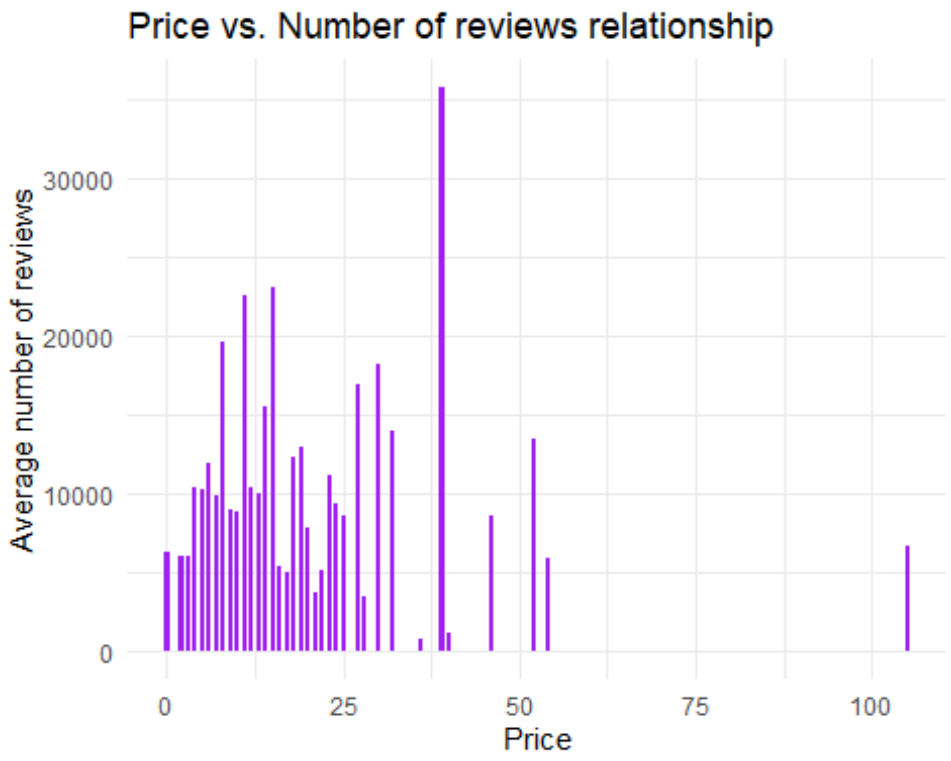


From 2012 to 2019, the average book price decreased, as can be shown. However, we can observe a significant average fall in prices in 2015. The average book price began to rise after 2019. According to my theory, the COVID-19 pandemic is also to cause. The previous visualization stated clearly that there would be a significant increase in demand for reading books in 2020, which, in turn, led to higher costs.

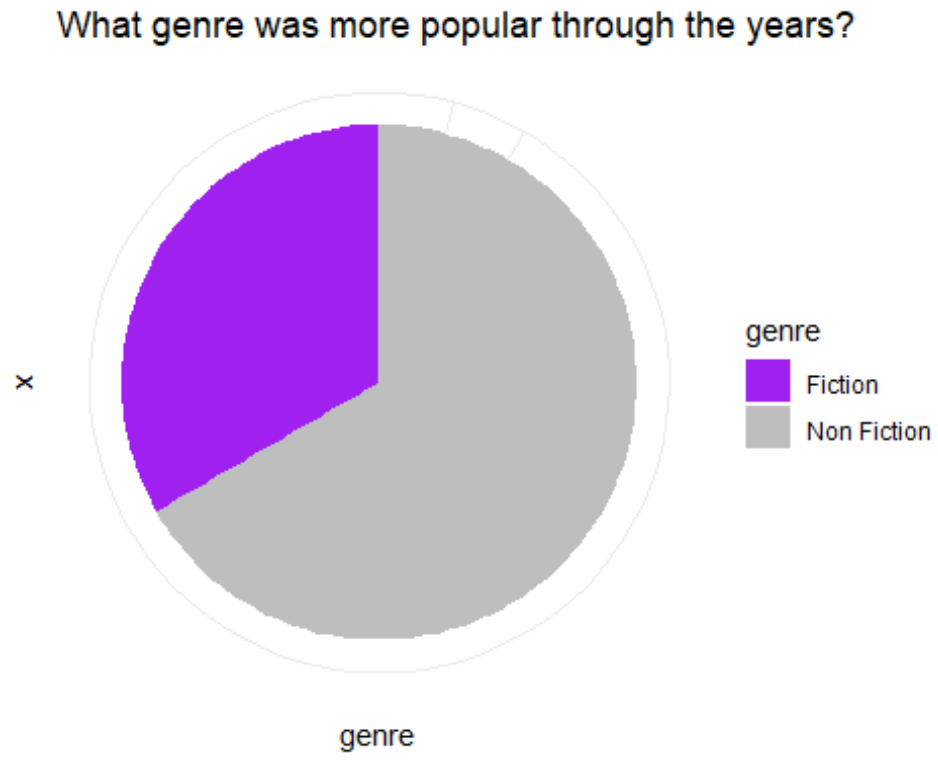


The average ratings are exceptionally high (starting at around 4.5 out of 5), which is not strange considering that this is a bestseller data set. When I compare this figure to the “Average number of reviews fluctuates across each year” plot, I can see that although the average number of reviews nearly remained constant from 2012 to 2019, the average rating increased.

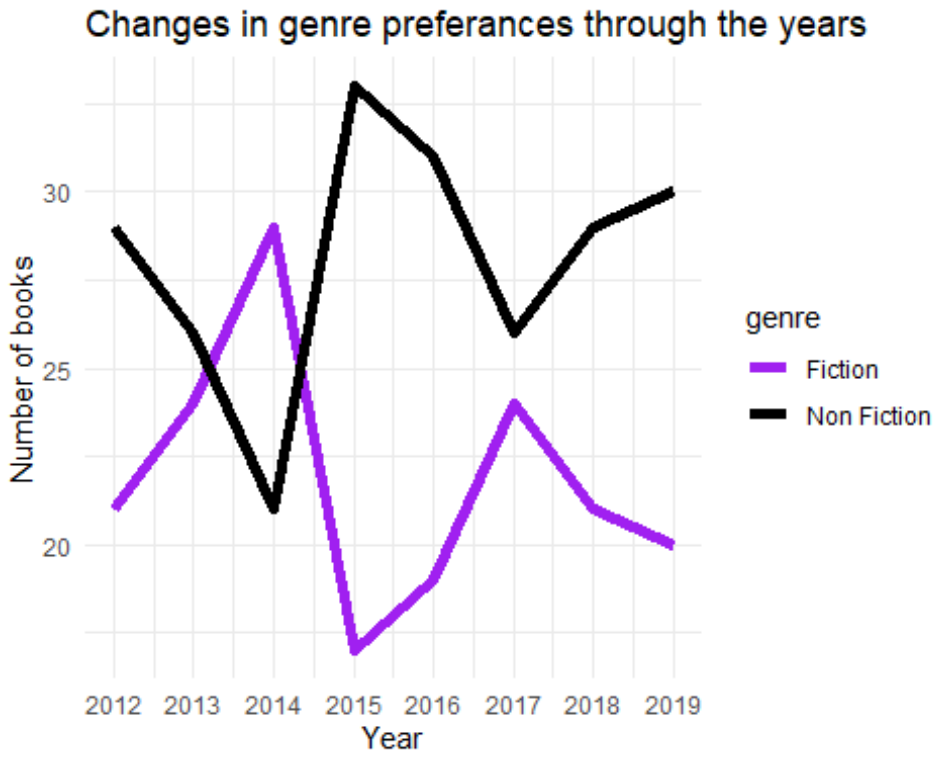
Let’s look at how the costs and reviews relate to each other (number of readings)



Let's now examine which genre received greater attention.



We can observe that nonfiction books were far more popular. And now let's analyze a more specific plot - the number of books published in each category over time.



Non-fiction books were preferred throughout the whole decade, with the exception of 2014. Additionally, both genres became intertwined in 2021, and as a result, the popularity of the fiction genre has surpassed that of non-fiction by 2022.

Conclusion:

Below are the few conclusions that were determined from analyzing the data.

Analysis conclusions:

1. Readers prefer Non-fiction books to Fiction ones.
2. In 2020 the number of reviews (considered the number of readings) significantly increased.
3. The book prices were constantly decreasing, until 2020. From 2020 we can see book prices going up.
4. Possibility: The second and third clauses might be connected with the COVID-19 pandemic that started in 2020.

Note: All the techniques were sourced with the references from Kaggle to analyze the data and other open sources which were remodeled for our study.