****

**Members**

Alyan Khowaja

Muhammad Hasnain

Abdul Rehman

Muzamil Shah

Jawad Hussain

Afaf Ghani

Hanzla

Muib Ur Rehman

Hooria Gaba

Contents

[**Introduction:** 3](#_Toc170942751)

[**Data Collection, Cleaning and Pre-processing:** 3](#_Toc170942752)

[**Exploratory Data Analysis (EDA):** 3](#_Toc170942753)

[**Visualization:** 4](#_Toc170942754)

[**Statistical Analysis:** 5](#_Toc170942755)

[**Conclusion and Recommendations:** 6](#_Toc170942756)

**Introduction**

The Amazon Bestselling Books Analysis Model project aims to analyse the data of bestselling books on Amazon to uncover trends, patterns, and insights. By utilizing Python for data analysis, we seek to understand the factors contributing to a book's success, providing valuable information for authors, publishers, and marketers.

**Data Collection:** Muhammad Hasnain and Abdul Rehman

* Data was sourced from [specific source: Kaggle], including key attributes like Title, Author, Price, Rating, Number of Reviews, Genre, Publication Date, and Sales Rank.
* <https://www.kaggle.com/datasets/abdulhamidadavize/top-100-best-selling-books-on-amazon-20092021>

**Data Cleaning and Pre-processing:** Abdul Rehman and Muhammad Hasnain

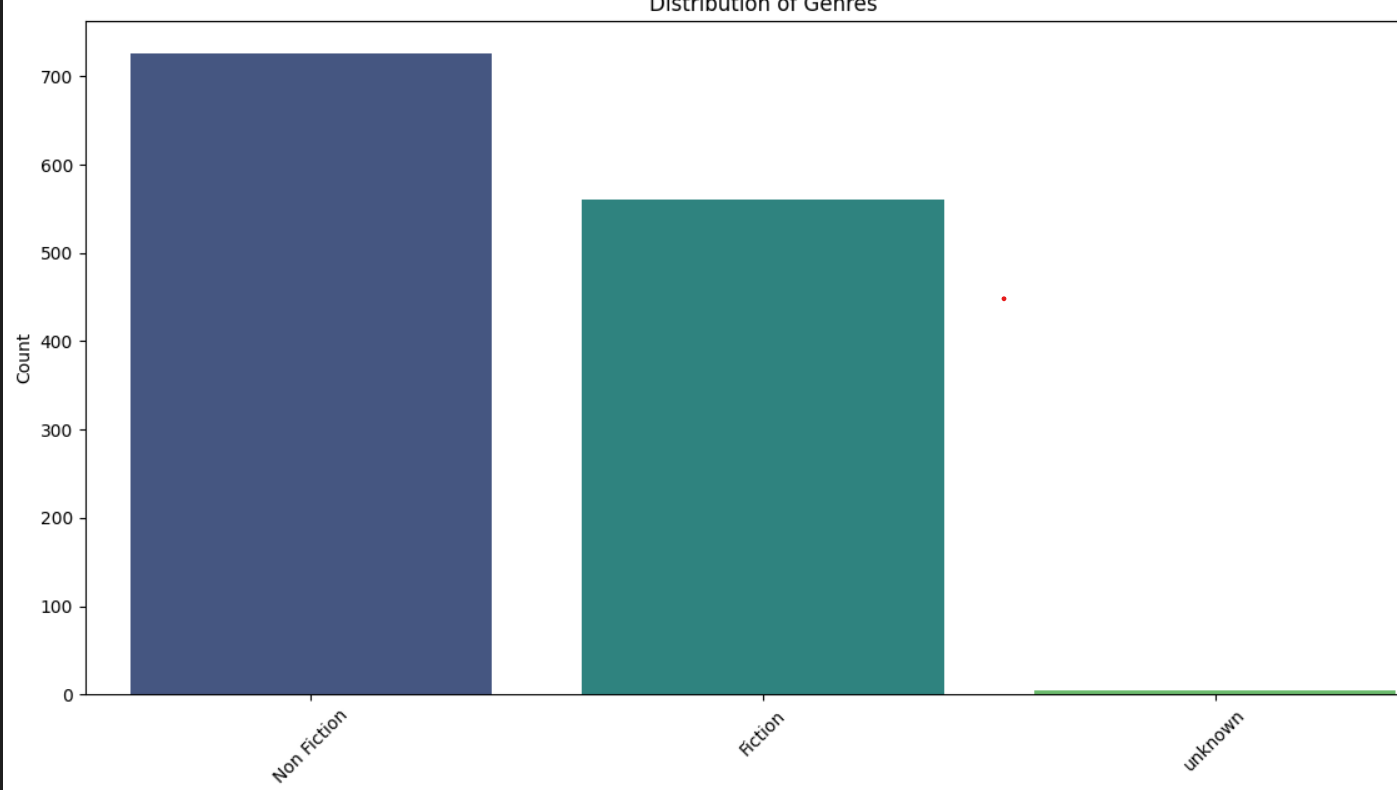
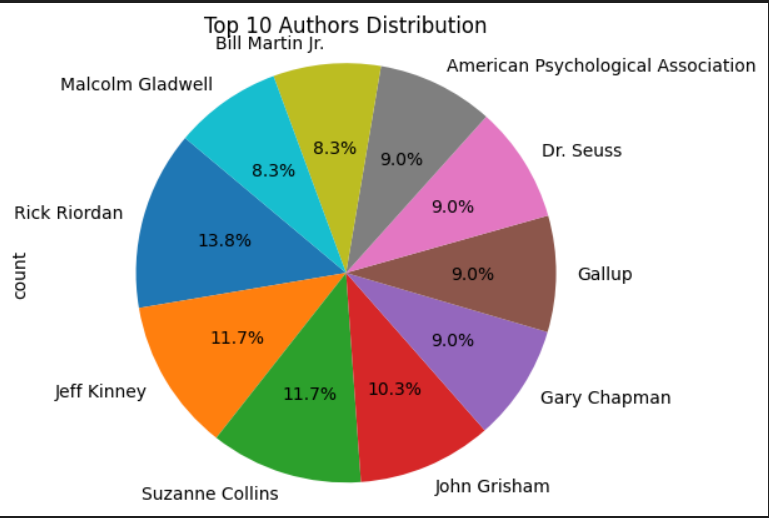
* Missing values were handled by [method used, e.g., imputation or removal].
* Duplicates and outliers were removed to ensure data integrity through removing and IQR.
* Categorical data were standardized, and numerical data were normalized.

**Exploratory Data Analysis (EDA):** Muzamil Shah and Jawad Hussain

* Descriptive statistics provided insights into the dataset's basic characteristics including mean, median, mode and standard deviation.
* Distribution analysis using histograms and box plots revealed [specific insights, e.g., common price ranges or rating distributions].
* Correlation analysis identified significant relationships between variables using heat map.

**Visualization:** Afaf Ghani

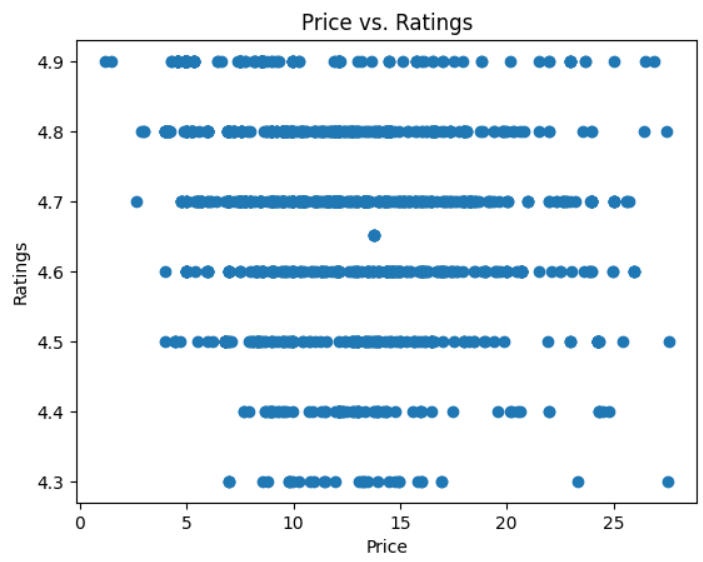
* Bar charts and pie charts illustrated the distribution of genres and authors.

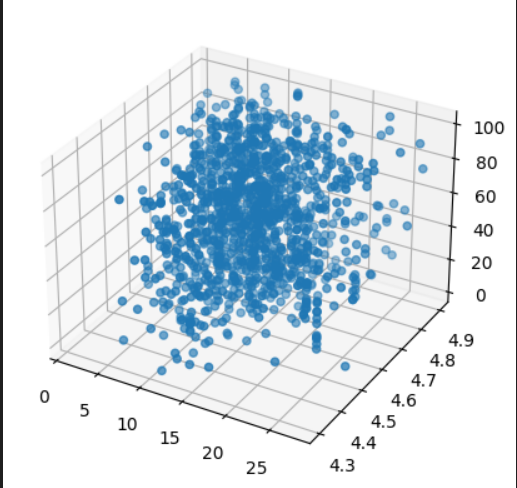
* Line graphs showed trends over time, such as changes in average ratings.



* Scatter plots visualized relationships, such as price versus rating.



* 3d projection of scatter plot to visualized relationships, such as price, rating and sales rank.



**Statistical Analysis:** Alyan, Hanzla and Munib ur Rehman

* Hypothesis testing confirmed the significance of various factors.

**The output of your T-test indicates the following:**

T-statistic: -0.58 (approximately), suggesting that the means of the two groups (Fiction and Non-Fiction) are very close to each other. P-value: 0.55, which is much higher than the typical significance level (e.g., 0.05).

Interpretation**:** Fail to Reject the Null Hypothesis: This means there is no statistically significant difference between the average sales ranks of Fiction and Non-Fiction genres in your dataset. The high P-value suggests that any observed difference in means is likely due to random chance. You can conclude that, based on this analysis, the sales ranks for the Fiction and Non-Fiction genres are similar.

**The output of your Chi-square test indicates:**

In this case, the p-value is 0.5807, which is greater than the significance level of 0.05. Therefore, we fail to reject the null hypothesis. This means there is no significant association between the genre of the products and their sales rank (whether they are in the Top 50 or not).

In summary, the Chi-square test results suggest that the genre of the products does not have a significant impact on whether the products achieve a top sales rank.

* Regression analysis identified key predictors of sales rank e.g.: Price, Number of Reviews, Ratings, Genre

To identify the factors influencing the sales rank of books, a linear regression model was developed. This model aimed to predict the sales rank based on features such as price, number of reviews, ratings, and genre.

The categorical variable Genre was converted into numerical form using OneHotEncoder. This encoder transformed the Genre column into multiple binary columns, each representing a unique genre. The first category was dropped to avoid multicollinearity.

The performance of the linear regression model was evaluated using the Root Mean Squared Error (RMSE). This metric provides an understanding of the model's prediction error, with a lower RMSE indicating better model performance.

* Clustering grouped books with similar characteristics, revealing [specific insights, e.g., common traits among top-selling books].

To identify groups of books with similar characteristics, k-means clustering was employed. This method helps in understanding patterns and segmenting the data into clusters based on features such as price, ratings, and the number of reviews.

To determine the optimal number of clusters, the elbow method was used. This method involves plotting the inertia (sum of squared distances of samples to their closest cluster center) for a range of cluster numbers and identifying the point where the rate of decrease sharply slows, indicating the optimal number of clusters.

A scatter plot was created to visualize the clusters. Each point represents a book, colored by its cluster assignment.

**Actionable Insights**

* **Authors:** Focus on genres with higher average ratings and review counts.
* **Publishers:** Invest in marketing strategies that boost ratings and reviews.
* **Marketers:** Target promotional efforts based on identified key factors influencing sales.

### Conclusion

This project provides valuable insights into the factors contributing to a book's success on Amazon. By leveraging Python for data analysis, we have uncovered patterns and trends that can inform authors, publishers, and marketers in their strategies. The findings from this analysis offer actionable recommendations for improving book sales and achieving greater success in the competitive book market.