EXPLORING AMAZON'S BESTSELLING BOOKS

Shreya Khettry

Abstract: My project utilizes a dataset that contains Amazon's Top 50 Bestselling Books from 2009-2019. The dataset contains 550 books and is categorized into fiction and nonfiction utilizing Goodreads. The aim of this project is to do an exploratory visualization based analysis of the dataset to find interesting information and trends.

Specs: I used Python 3 as my Programming Language and Jupyter Notebook as my programming environment. I utilized packages such as numpy, pandas, seaborn and matplotlib that are in-built in Python.

Contact: For more information please contact Shreya Khettry at **skhettry@umass.edu**

INTRODUCTION

For my Final Project I decided to look at a dataset that contains Amazon's Top 50 Bestselling Books from 2009-2019. The dataset contains 550 books and is categorized into fiction and nonfiction utilizing Goodreads. The dataset includes seven categories, such as Name of the Book, Author of the Book, User Rating, Number of Reviews, the Price of the Book, the Year(s) it ranked on the Bestseller list and whether the genre is Fiction or Nonfiction. (Note: I will be sending the csv file as part of the submission).

METHODOLOGY

I started off with loading my dataset into the environment and then performing basic preliminary analysis to learn and discover more about the dataset so that I was able to utilize it for my exploratory analysis. Looking at the dataset I was getting multiple questions in my mind and I decided to note them all down. I had a total of 17 questions and I thought that the best way to answer them was through visuals and hence decided to make this a primary visualization based project. I believe that when we look at or visualize things we tend to understand more and this was the primary reason I decided to do this.

My python notebook is divided into Two Major sections: Data Processing and Preliminary Exploration and Data Visualization Analysis. The first section consists of loading in the data and exploring the various columns and rows and their contents. The second section is structured by starting off with my question, followed by the code to

produce a graph to answer the question. Once the graph has been produced, I have written my analysis and conclusion right below it so that it becomes easier to view the question, code and conclusion all together. The following are the list of questions I have answered (in order):

1. What is the Rating Distribution for the Books?
2. What is the Genre Distribution for the Books?
3. What Is Price Distribution for the Books?
4. Is There a Price Difference Between the Two Genres?
5. What Is Review Distribution for the Books?
6. What Is the Number of Books Per Rating?
7. Which Year Has the Highest User Rating?
8. Which Year Has the Highest Reviews?
9. What is the Price Variation Through Time?
10. What were the Best Books over the Decade?
11. What Are the Highest Reviewed Books?
12. What Are the Lowest Reviewed Books?
13. What Are the Worst Rated Books?
14. What Are the Most Expensive Books?
15. What Are the Cheapest Books?
16. What Are the Best 10 Free Books?
17. Who Are the Most Popular Authors?

I hope you will enjoy going through my project.