Online Bookstore Analysis – Amazon Bestsellers (2009-2019)

Overview:

A simple and effective bookstore analysis to help users buy the top bestsellers (2009-2019) on Amazon based on user ratings and reviews.

Data:

I used a dataset from Kaggle: <https://www.kaggle.com/datasets/sootersaalu/amazon-top-50-bestselling-books-2009-2019> (CSV file).

Data Cleaning:

I performed data cleaning on the above raw data using MS Excel. My clean-up tasks included the following steps:

* Performed a basic spellcheck to remove unwanted special characters, correct misspelled words, etc.
* Edited three of the column header names: ("Name" to "book\_name", "user rating" to "user\_rating", "price" to "price\_in\_$") to ensure compatibility with the database schema.
* Removed duplicate entries from name column - 200 rows of duplicate entries were removed.
* Researched and updated correct names for a few incorrect book and author names.

After the clean-up, I created a table schema in a PostgreSQL database and imported the cleaned-up CSV file to the database via pgAdmin 4 using the “Import/Export Data” tool. After the import, for a further clean-up, I checked for null values in each column in the table using SQL queries (Refer [File 1](#File1)).

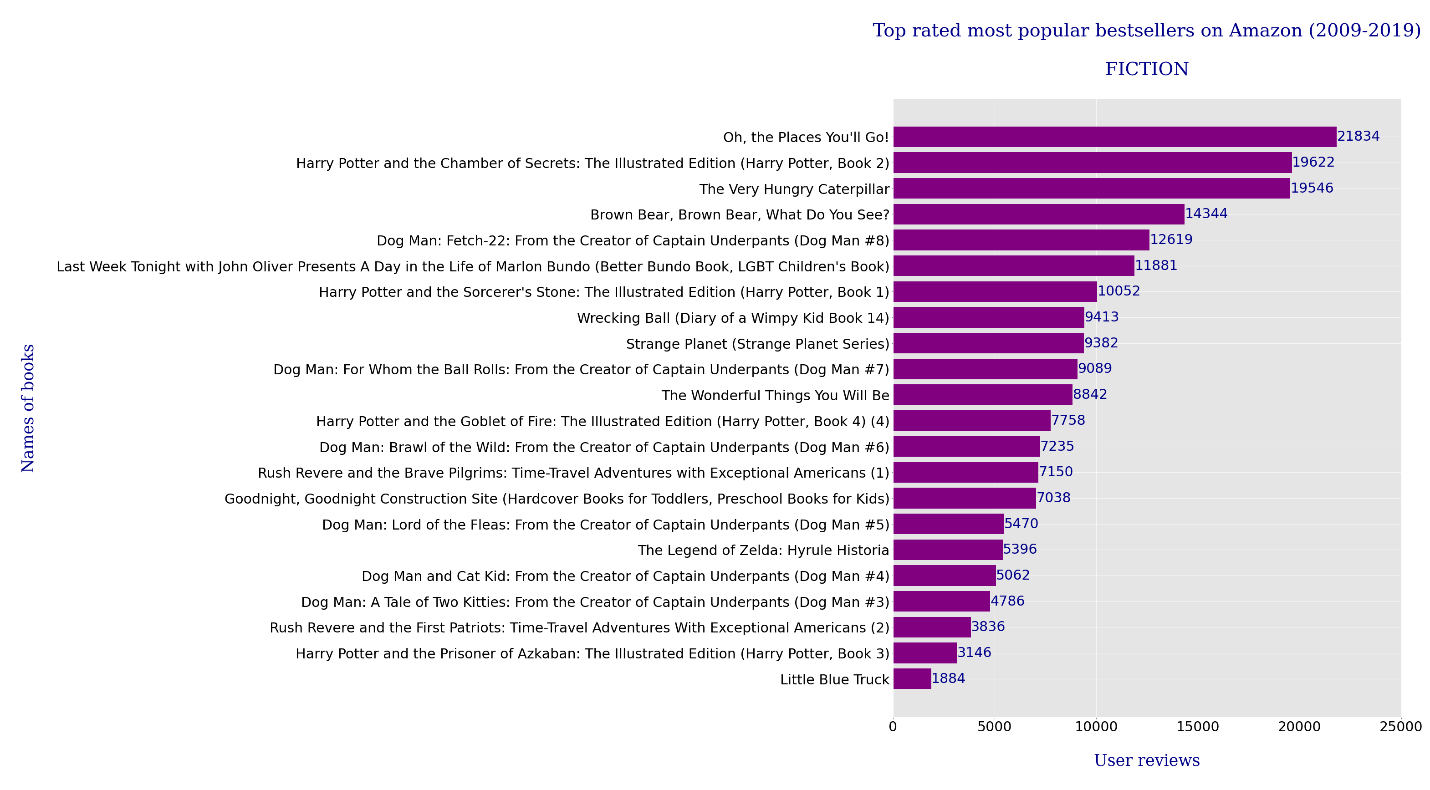
Univariate Exploratory Data Analysis:

I performed an exploratory data analysis in SQL to answer the following questions (Refer [File 2](#File2)):

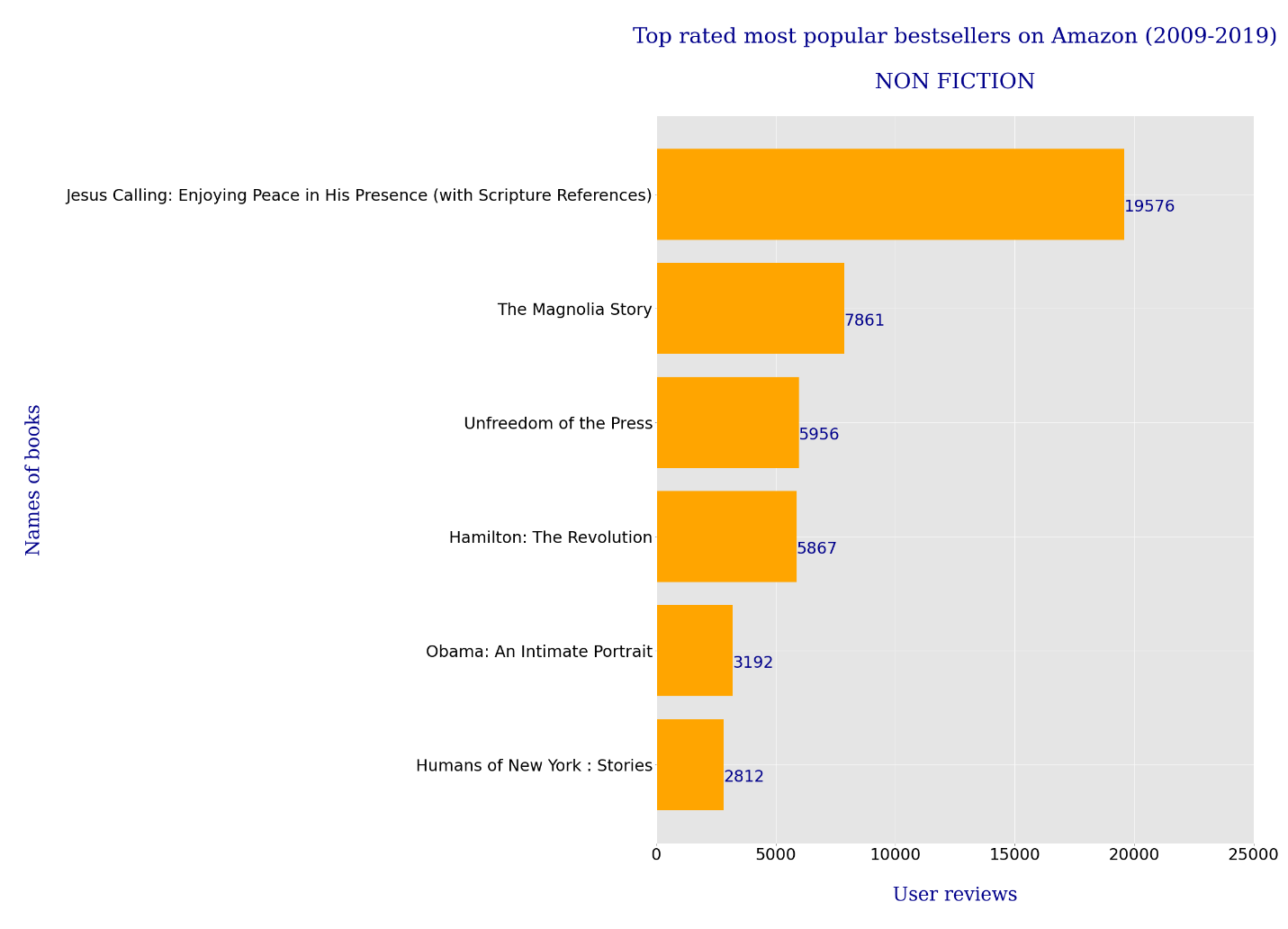
1. Which books have the highest user rating in both fiction and non-fiction genre?
2. From the results of the 1st question, what are the most popular books based on user reviews (in descending order) in both fiction and non-fiction genre?

Data Visualization (Results):

I used Python for data visualization (Refer [File 3](#File3) & [File 4](#File4)). Here are the results:









Conclusion:

In conclusion, there were a total of 22 bestsellers (2009-2019) with the highest user rating of 4.9 in the **fiction** genre. Among those 22, the most popular book based on user reviews was “Oh, the Places You’ll Go!” written by Dr. Seuss. It was an Amazon bestseller in the year 2012 priced at $8.

By contrast, there were only 6 bestsellers (2009-2019) with the highest user rating of 4.9 in the **non-fiction** genre. Among those 6, the most popular book based on user reviews was “Jesus Calling: Enjoying Peace in His Presence (with Scripture References)” written by Sarah Young. It was an Amazon bestseller in the year 2011priced at $8.

Software:

* MS Excel (version 2303)
* Database: PostgreSQL 15 (language: SQL, GUI: pgAdmin 4)
* IDE: Jupyter Notebook (language: Python 3.6.5)
* Python libraries: psycopg2, matplotlib’s pyplot, matplotlib’s PdfPages

Files attached:

1. Amazon Bestsellers CSV File
2. Data Analysis SQL File
3. Data Visualization Python Jupyter Notebook File
4. Results PDF File