**Capstone Project for the Google Data Analytics Certificate**

**Case Study: Bestseller books in Amazon**

As a part of completing the google data analytics certificate, there is an optional capstone project which can be undertaken to make the knowledge gained throughout the course to stick. There are two different tracks and I have chosen the one where we can choose our own business case and the dataset to work on. Since, it is the first project, I wanted to have it simple and also flexible.

I have chosen the dataset – Amazon Top 50 Bestselling books which is an open dataset under a CC0 license and also designed the business case, the details of which are given below.

The dataset contains a list of top 50 books sold every year between 2009 and 2019 from Amazon. I am going to explore this dataset to arrive at some insights including but not limited to the below.

1. Which genre of books are the most sold?
2. Top 10 or 15 books
3. Top 10 or 15 authors
4. Which author sold the most single title books?
5. Which author sold the most books overall?
6. Which genre of books are most profitable?

Along with these insights, I will also be sharing the visualization and finally a presentation.

In the Google Data Analytics course, we approach any data analysis problem with six steps. They are,

1. Ask
2. Prepare
3. Process
4. Analyse
5. Share
6. Act

**Ask:**

Business task: To identify the trends in the sale of books and arrive at insights which could help a bookstore (online or physical) to stock up as necessary. The popular titles and authors will be shared as well as the most profitable genre.

Stakeholders: Anyone in the bookstore, the manager or the head of procurement who wants to know beforehand what to stock up.

Dataset: <https://www.kaggle.com/datasets/sootersaalu/amazon-top-50-bestselling-books-2009-2019>

Metrics: Revenue generated per genre, top 10 books and top 10 authors are a few metrics (kind of) which will help the client in making decisions.

**Prepare:**

The dataset from the link mentioned above is downloaded locally and stored in properly labelled file within a folder. It is a csv file. The raw data is retained and any analysis is made on a copy of this file. The dataset consists of 550 records (rows) and 7 fields.

Checking for credibility – does the data ROCCC? – The data is from a reliable source and original. We have the information from which we can arrive at some conclusion for our problem statement. However, it is not a current one as it is not updated regularly. But this is what we have and to my knowledge, for this project, it is sufficient. Also, I don not see any sort of bias in the data.

The data is of CC0 License, which means it is in the public domain and it can be used by anyone, with proper credits of course.

**Process:**

Tools to be used for analysis – Microsoft Excel, SQL (BigQuery), Tableau, Microsoft Powerpoint

The data as such is good to use and I didn’t have to do any cleaning. Except for a few values in the Price column being 0, which I am not sure if correct or a mistake and no way of ensuring, there is not any issues. Also, data manipulation was also not done as I didn’t see any need for it for the sake of our business case.

**Analyse:**

I have analysed the data in Microsoft Excel using Excel Functions as well as Pivot Tables. The spreadsheet is also attached. Please take a look at the analyses.

Below are some of the analyses done with Microsoft excel using functions

1. No of fiction books: =COUNTIF(Books!G:G,"Fiction")
2. No of non-fiction books: =COUNTIF(Books!G:G,"Non Fiction")
3. No of reviews greater than 4.0: =COUNTIF(Books!C:C,">4")
4. No of reviews less than 4: =COUNTIF(Books!C:C,"<=4")
5. Lowest rating: =MIN(Books!C:C)
6. Highest rating: =MAX(Books!C:C)
7. No of books with the lowest review (3.3): =COUNTIF(Books!C:C,"=3.3")
8. No of books with highest rating (4.9): =COUNTIF(Books!C:C,"=4.9")
9. Fiction books with rating 4.9: =COUNTIFS(Books!C:C,"=4.9",Books!G:G,"Fiction")
10. Non Fiction books with rating 4.9: =COUNTIFS(Books!C:C,"=4.9",Books!G:G,"Non Fiction")

Using Pivot table, a few important trends are captured on how the sale of fiction books differs from that of Non-fiction books, which is more profitable, the average, min and max prices and other summary statistics. There is also break up of the number of books per genre that topped every year, the top 20 authors and number of books with specific ratings.

It is found that the number of copies of non-fiction books sold are higher than the copies of fiction sold. Not just the numbers, but the revenue generated by sales of non-fiction book (63.86%) is much higher than the revenue by fiction books (36.14%).

SQL:

Some more analyses were also done using SQL in BigQuery DB. The dataset was imported to BigQuery Sandbox environment to analyze.

Query link: <https://console.cloud.google.com/bigquery?sq=844641499904:464632e821cc430e9701ff8a8c388747>

1. I tried to find the distinct number of book titles – 351 distinct titles were available.
2. To know which title was sold most, I tried to find the count of each title and grouped them using GROUP BY CLAUSE. Using ORDER BY in descending, I got the result of which title was sold the most. Finally, I also got a list of the top 10 most sold books. The top 2 books are of Non-fiction Genre.
3. Next, I want to find how many distinct authors are there and how many of their books are sold. This is to find the popular author. Similar to the above, we grouped the data by authors name and also found out the number of books sold for each author. I also found the top 15 authors that had sold maximum number of books of the same title. The top 2 authors are Non-Fiction writers.
4. However, it is interesting to see that there are a couple of fiction writers who had sold many books but they haven’t sold large amount of single title books. It is also surprising that popular writers (and my favorites) like J.K.Rowling, Dan Brown and Khaleed Hosseini didn’t make it to the top 10
5. The books that have the maximum rating and minimum rating are also listed
6. The books with maximum Price and minimum Price was also listed. It is surprising to know that the minimum price is 0$. I am not sure if it is a mistake with the data or is it a free edition book
7. Also, analysed which book has the greatest number of reviews and which the least.
8. It is observed that the number of non-fictions sold were slightly or markedly higher than the sale of fiction.

From all the above analyses, it is evident that the non-fiction books were very popular and also lucrative when compared to fiction.

**Share:**

As we complete the analysis, the insights and findings are visualized for easier understanding.

I have used Tableau to create visualizations and dashboard of some of these insights, which were key.

The tableau link to view the visualization is given below:

<https://public.tableau.com/app/profile/lavanya.g4679/viz/Bestsellingbooks_16754447001140/Dashboard1>

**Act:**

As evident from the analysis and visualizations, the non-fiction books are sold more than fiction and the revenue generated is also very high than fiction.

When it comes to stocking up books on particular authors, we need to check with authors had sold more number of books per title and should not simply go by the one who sold more number of books. Eg. Jeff Kinney has sold 12 books in total, but they are actually just one book per title. Whereas, American Psychological Association has sold only 10 books but it is 10 books of the same title. It would be beneficial to stock up such books than just going by author.

This is the end of this case study.