**M2 Practical Challenge:**

**Creating Your Own Alteryx Workflow**

The objective is to get a list of books base on “Goodreads Books”, for that I’m going to use two datasets, and combine them to obtain a bigger one.

Once that’s done, I’m going to make use of a third dataset in order to get a Bestseller Ranking for over 1 million books.

The goal is to obtain a list of the top 10 best “Goodreads Books” based on the Bestseller Ranking. As well as make some Scatter plots render, comparing the BestSeller Ranking to the Average Rating, Rating Count, and Number of pages. That way we can see the influence in the ranking of these variables.

Table of Contents

[1. Importing the data 2](#_Toc83480038)

[2. Create the WorkFlow: 4](#_Toc83480039)

[2.1. Creating the Books dataset 4](#_Toc83480040)

[2.2. Creating the BestSeller Ranking map 7](#_Toc83480041)

[2.3. Transforming the data 8](#_Toc83480042)

[3. Outputting the data 10](#_Toc83480043)

[4. Review 14](#_Toc83480044)

# Importing the data

First, we need to get 3 data sources. For that I have use Kaggle to obtain 3 .csv files related with books:

<https://www.kaggle.com/jealousleopard/goodreadsbooks/version/2>

<https://www.kaggle.com/dylanjcastillo/7k-books-with-metadata>

<https://www.kaggle.com/sp1thas/book-depository-dataset?select=dataset.csv>

The first .csv file include 3500 books obtain from the app “[goodreads books](https://www.goodreads.com/)”.

The second .csv file include 6810 books, made by “DylanCastillo”, based on the “Soumik's Goodreads-books dataset” and the “Google Books API”

The third .csv file contains 1,109,383 books, and its bestseller rank.

The input from the first dataset is:

Imagen de la pantalla de un computador

Descripción generada automáticamente con confianza baja

The input from the second dataset is:

Pantalla de computadora

Descripción generada automáticamente con confianza media

And the input from the third dataset is:

Imagen que contiene interior, computadora, tabla, grande

Descripción generada automáticamente

# Create the WorkFlow:

Interfaz de usuario gráfica, Diagrama

Descripción generada automáticamente

## Creating the Books dataset

On the upper part, the two Goodreads Books are imported. Before joining them, we need to make sure the have the same name and number of columns, as well as the same type of data.

Diagrama

Descripción generada automáticamente

I make use of the select tool to change the types of columns, change the names, and select only the needed columns.

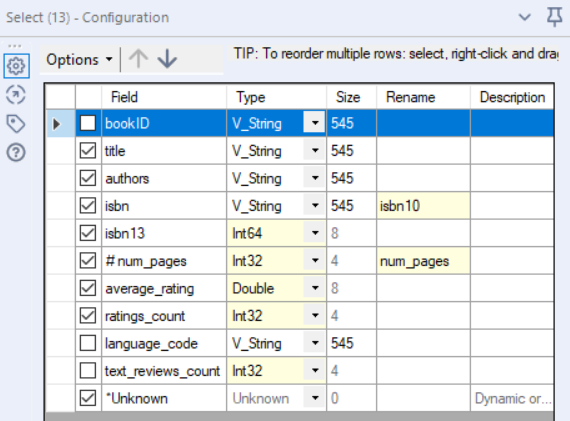


Figure 2 - books.csv select

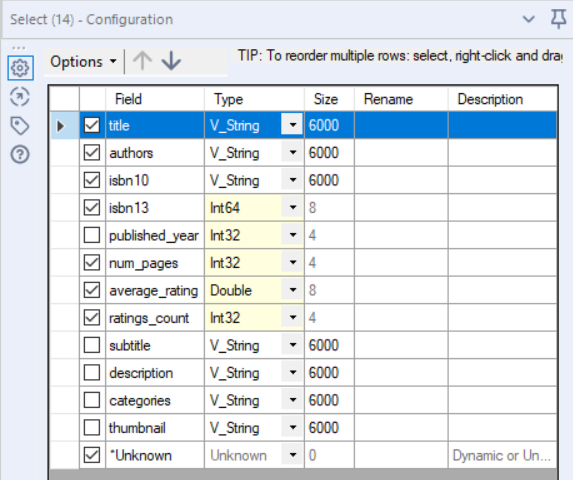


Figure 2 – 7k\_books.csv select

Once the two data source are modified, we can combine them. For that I made use of the Union tool. The main problem is that there are some books that have come from both datasets, for that we need to delete the duplicates books.

But first lest clean the Title and Authors columns, for that I used the Data Cleansing tool, I decided to replace with blanks the nulls values, as the main identifier of the book is the ISBN, so a missing title or authors is not a big problem. Also deleted some Whitespaces and Tabs.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Lastly, I made use of the Unique tool to remove duplicated books base on the ISBN 13 as the key.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Doing the previous steps, I now have a dataset of 8.550 unique books, that have the next information:

Interfaz de usuario gráfica, Aplicación, Tabla, Excel

Descripción generada automáticamente

## Creating the BestSeller Ranking map

On the lower part, we can see the creation of the map between the books ISBN and the BestSeller Ranking.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

The first thing I did was eliminate the rows that contains an empty value on the bestseller rank column. For that I made use of the Filter tool, to get the non-empty rows.

After that with the use of the Select tool, I selected the needed columns, such as title, bestsellers-rank, isbn10, isbn13 and lang.

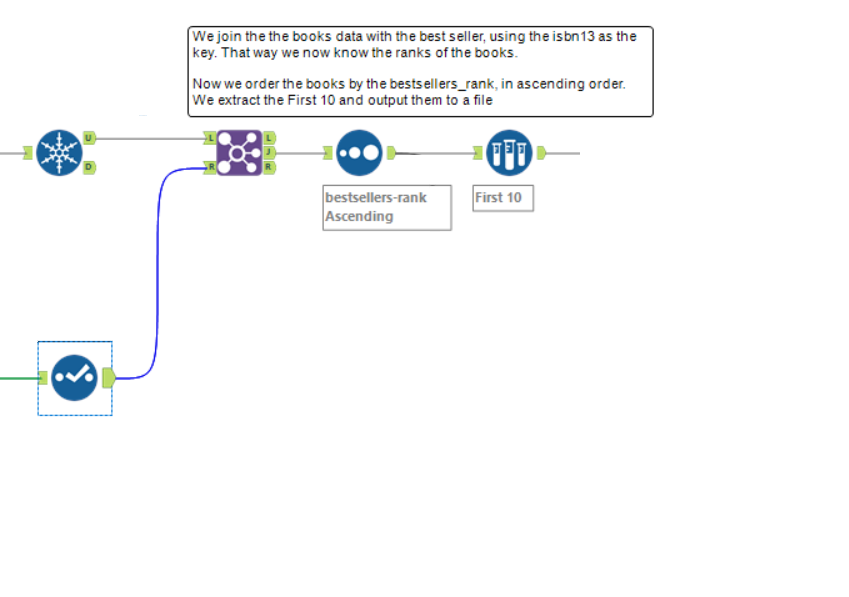
Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente Tabla

Descripción generada automáticamente

## Transforming the data

For this part, the next part of the workflow is used.



First the output of the previous sections, are join, using the Join tool, and using the isbn13 field as the key. That way the output will contain the books from the merge dataset from “Goodreads Books”, and the associated bestseller ranking, from the other input.

Interfaz de usuario gráfica, Tabla

Descripción generada automáticamente

After doing this step, I sample order the books by ascending order of the bestseller rank, by using the Sort tool. After that I used the Sample tool to get the first 10 books.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

This way the output data from the sample is the following:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

# Outputting the data

For that I just connected the output from the Sample tool directly into an Output Data tool, and save it as “top\_10\_bestsellers\_from\_goodreads\_books.yxdb”

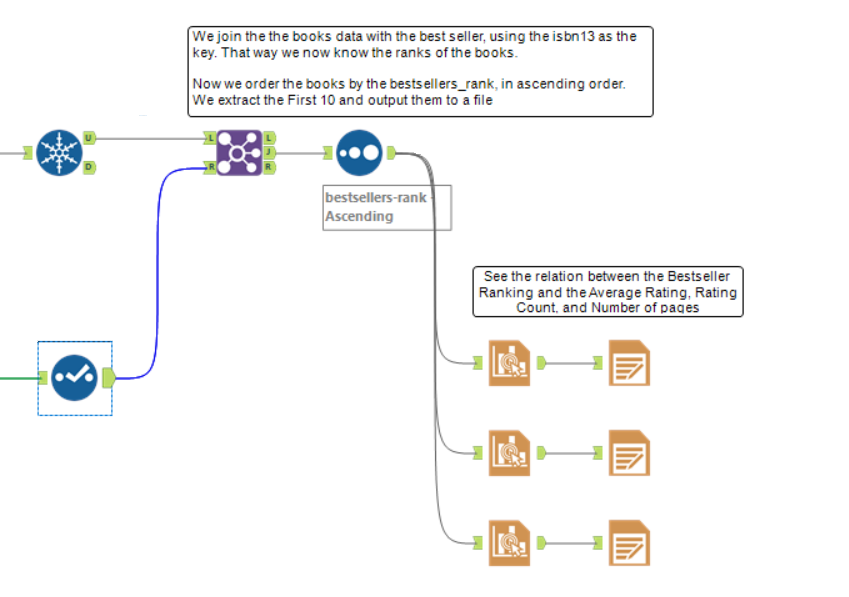


The output data will we the following:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Another Output data that is presented is some Scatter plats renders.



For that I connected the output of the Sort tool into 3 Interactive Chart tool, in order to make 3 graphs. Each one compare 3 thing to the Bestseller Rank. 1. Average Rating, 2. Rating Count, 3. Number of Pages.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente Interfaz de usuario gráfica

Descripción generada automáticamenteInterfaz de usuario gráfica

Descripción generada automáticamente

From than each graph is converted into a render an outputted as a temporal HTML file to visualize, using the Render tool.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

The 3 Scatter plots render are the following:

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

# Review

Interfaz de usuario gráfica, Diagrama

Descripción generada automáticamente

As we can see from the workflow, there are 10 different tools used:

1. In/Out Category
   1. Input Data (1)
   2. Output Data (2)
2. Preparation Category
   1. Select (3)
   2. Data Cleansing (4) (NEW)
   3. Unique (5) (NEW)
   4. Filter (6)
   5. Sort (7)
   6. Sample (8)
3. Join Category
   1. Union (9)
   2. Join (10)

Also 2 new tools have been used, “Data Cleansing” and “Unique”, from the Preparation Category.

Apart from these tools, other tools have been used:

1. In/Out Category
   1. Browse
2. Reporting Category
   1. Interactive Chart
   2. Render
3. Documentation Category
   1. Comment