Data Wrangling report

Three main datasets:

Book Sales (called publishers) from Corgis database. Luckily, the data is quite clean and clearly labeled. Nevertheless, some cleaning steps are performed as followed:

1. Change columns name with dot/space and uppercase to snake case and under case.
2. Remove redundant category name in front of columns names
    a. Ex: statistics.sale price to sale_price
3. Drop 3 revenue columns: amazon revenue, author revenue, gross sales to focus on daily average unit sales only. All of these columns are strongly correlated.
4. Drop sale rank column with the same reasons as above.

There are no missing values in this set.

After exploring multiple aspects from price to sales, there are no outliers visible. Most data points stay within reasonable range.

Book Reviews and Rating – this data is compiled from 2 different data sets: BX-Books and BX-Book-Rating. Both requires these same cleaning steps

1. Read csv file with separator of ";" and encoding ISO-8859-1
2. Change columns name to snake case format.

For BX-Book-Rating, data is group by ISBN to calculate the mean review score for each ISBN. Afterward, joining to BX-Books data set with key of ISBN.

As these data sets are from a nice source, they both contains no missing or mis labeled data. There are also no outliers clearly visible.

Goodreads dataset – this data set is from Kaggle. However, since it is acquired through scraping goodreads website, there are some cleaning steps:

1. First problem is extremely troublesome. Since this is a csv file, there are about 5 rows where authors name is concatenated with a comma instead of the dot comma. These rows are not the norm with this data set and interfere with reading the file. The easiest fix is manually editing the csv file according to the error message. Again, this is only viable since there are less than 5 of these errors or else a more systematic approach is much preferred.

2. Columns names are converted to snake case with lowercase.
3. Publication date columns are converted to date time object.

There are a very small percent of missing values presented in this dataset. Thus, a simple dropna from pandas is used here.

There are no visible outliers.