Data wrangling report

The following report describes the steps to clean the publisher data set from Corgis which includes 27000 e-book sales records from Amazon. Since this comes from a standard source, it requires minimal data wrangling. However, one must address multiple redundant columns along with many outliers in the data.

Impressions on the data and steps taken are as followed:

1. There is no missing value in the data.
2. The data is clearly labeled with correct format but the column name contains dot/space with mixing capital letters. Thus, columns' names are changed to lowercase and converted into snake case. Suffixes are also dropped.
    a. Ex: Statistics.Sale.Price => sale_price
3. Multiple columns with sales statistic including author revenue, total revenue, gross sales heavily correlate with each other. Thus, we remove all of them except for daily average units sold as the sole dependent variable. This step will ensure no collinearity exits when building the models.
4. Drop sale rank column with the same reasons as above.
5. 'sold by' column includes different names for the same company. They are adjusted to the same company.
    a. Ex: Harper Collins Christian Publishing, Publishers and Publishing become HarperCollins
6. Outliers appear in many columns such as prices, sales and total number of reviews. Removing them will take away the usefulness of the data. Therefore, all outliers are left as is for exploratory analysis. But when building model, those columns will be log transformed to take away the long tail.
7. Lastly, the cleaned data set is renamed to book_sale.csv and saved.