**Changes and Info**

* Size of data set: 7 cols, 551rows
* ColA header: Name changed to Book Title
* Data sorted by year (asc) and by user rating (dsc)
* Used ‘len’ on colC for user rating, found 3 chars. Using conditional formatting

found some cells with 1 char. There were the ‘4’ ratings.Changed the values to numeric and reduced to 1 decimal.

* ColD with average Reviews:’11953’ found some very low values as ‘37’ ,

Maybe in the analysis process, these low values should NOT be taken into consideration.

* Checking colE(Price) , using the ‘min’ function found the value 0(zero) and some

other low values like 1,2 etc in some books. Checking amazon for the current prices,

found many inconsistencies and decided to delete completely the ColE(prices) as invalid.Our task does NOT need the prices , if it did we should have made a new collection.

* New Cols E, F (Year,Genre) checked using the “column stats” function and they are clear.
* Cols A, B changed data to “plain text”, Col E (Year) had problems changing to “Date” format and decided to leave it automatic.
* Copy created and named Bestsellers NO Duplicates. Duplicates removed “200 duplicates removed”, 351 unique rows remained. The Duplicates are there because some books made the top 50 more than one time. We have 2 copies now, one with all duplicates (Bestsellers with Duplicates) and one without any (Bestsellers NO Duplicates.
* The Cleaning Log is checked and our Data is clean and ready for analysis!

**Changes during analysis process**

* Found a duplicate at the same year (dups at different years make sense, but in the same is a NO NO) Book Title: The help (1 dup deleted)
* Year 2014, 2 more duplicates were deleted. “Unbroken: A World War II Story of Survival, Resilience, and Redemption” and “The Fault in Our Stars”.
* The sheet with the duplicates is invalid and decided to drop and work

only with the No Duplicates one!