**CSCI 620/Section 03/Mior, Introduction to Big Data, Spring 2215 Assignment 6 – Data cleaning**

**Submitted by – Prakhar Gupta(pg9349)**

**Ans 1**

The extra JSON file contains IMDB id, Title label, Distributer Label, Currency to consider,cost of movie, box office revenue, and MPAA film rating label. The data is nested in value field of each JSON entity.

**Ans 2**

**(Refer to q2.py and q2mongo.txt)**

I joined the new data with the old movies data using \_id in both the tables. This way I was able to update (110690) movie ids with atleast one of the fields(Title label, Distributer Label, Currency to consider,cost of movie, box office revenue, and MPAA film rating label).

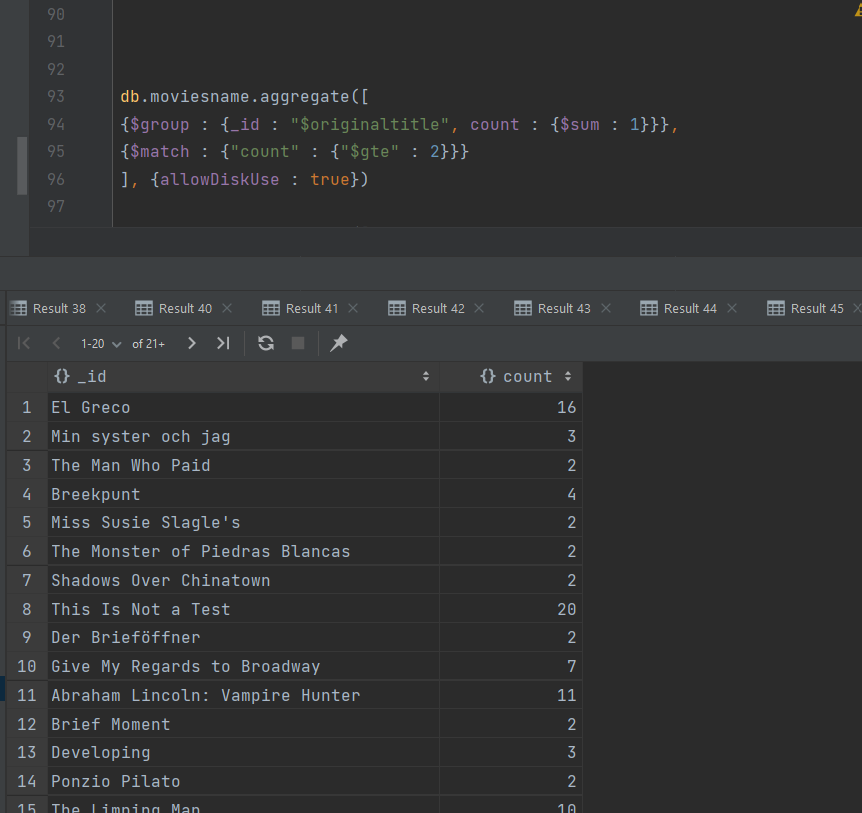
A new collection moviesnew is create with the updated data.

**Ans 3**

**(Refer to q3mongo.txt)**

I joined the new data with the old movies data using titlelabel in new json and orginaltitle in both the tables.

This result in only (76047) unique title matches. A lot of title had same name hence getting matched multiple time. A screenshot of small subset of the title which repeated is given below

-

In total there were 31329 titles which had more than one record after matching on title name. Complete list is given in Result46.csv where title name is followed by the number of time they get matched.

The mongo db queries and program highlights the issue which one would face matching on title label-

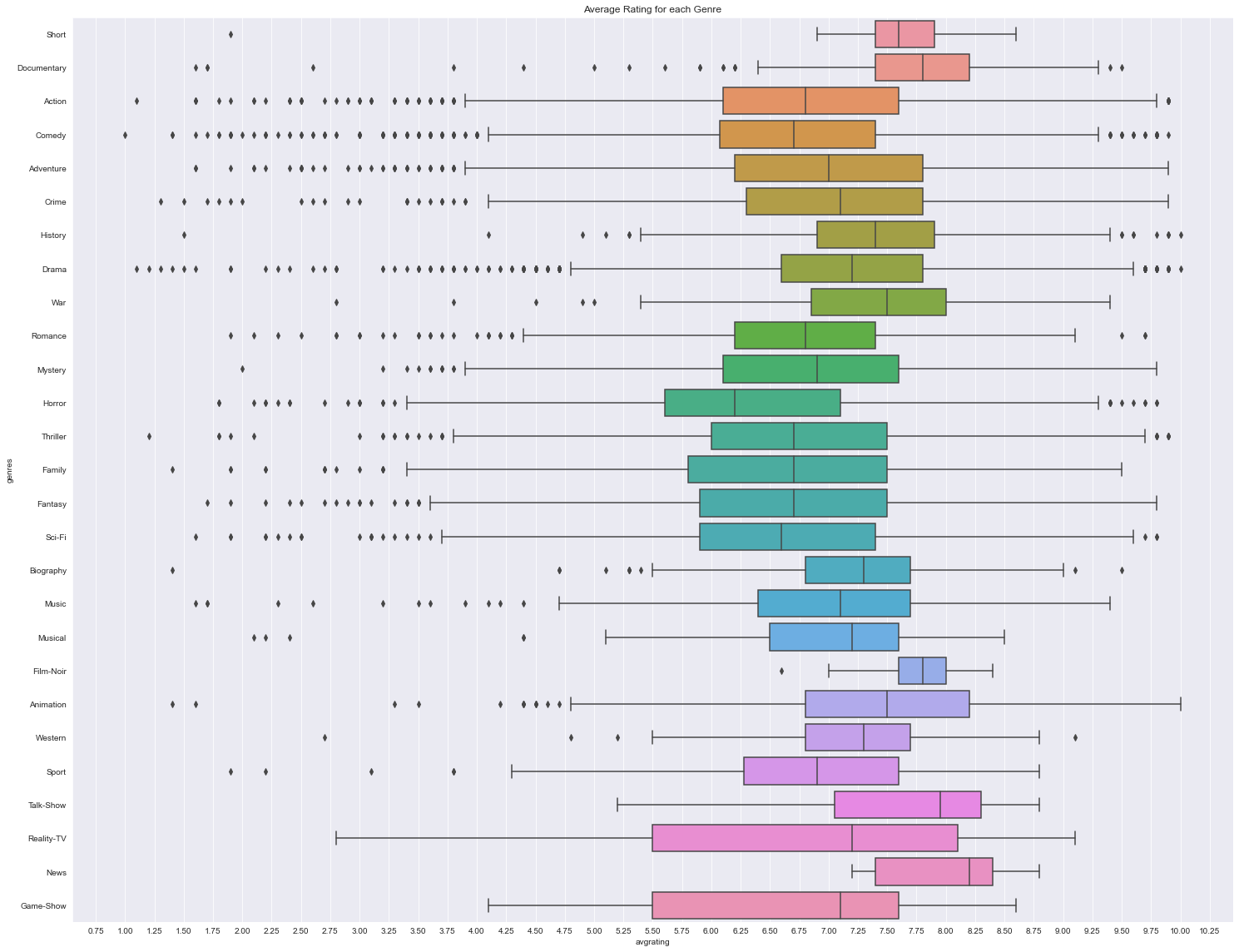
* The match rate is lower compared to matching on ids 76047 unique vs 110690. Hence leading to loss of information.
* It also create ambiguity in the data since movies can have same title matching them by title label only causes a lot of contradictory information to be added in the documents. For example (“This is Not a Test”) matched 20 times in total.
* Similar issue will be faced when trying to match on (title)primarytitle with title label with an lower match rate. Causing more ambiguity.

**Ans 4**

**(refer to VizNotebook.ipynb for detailed code+plot report)**

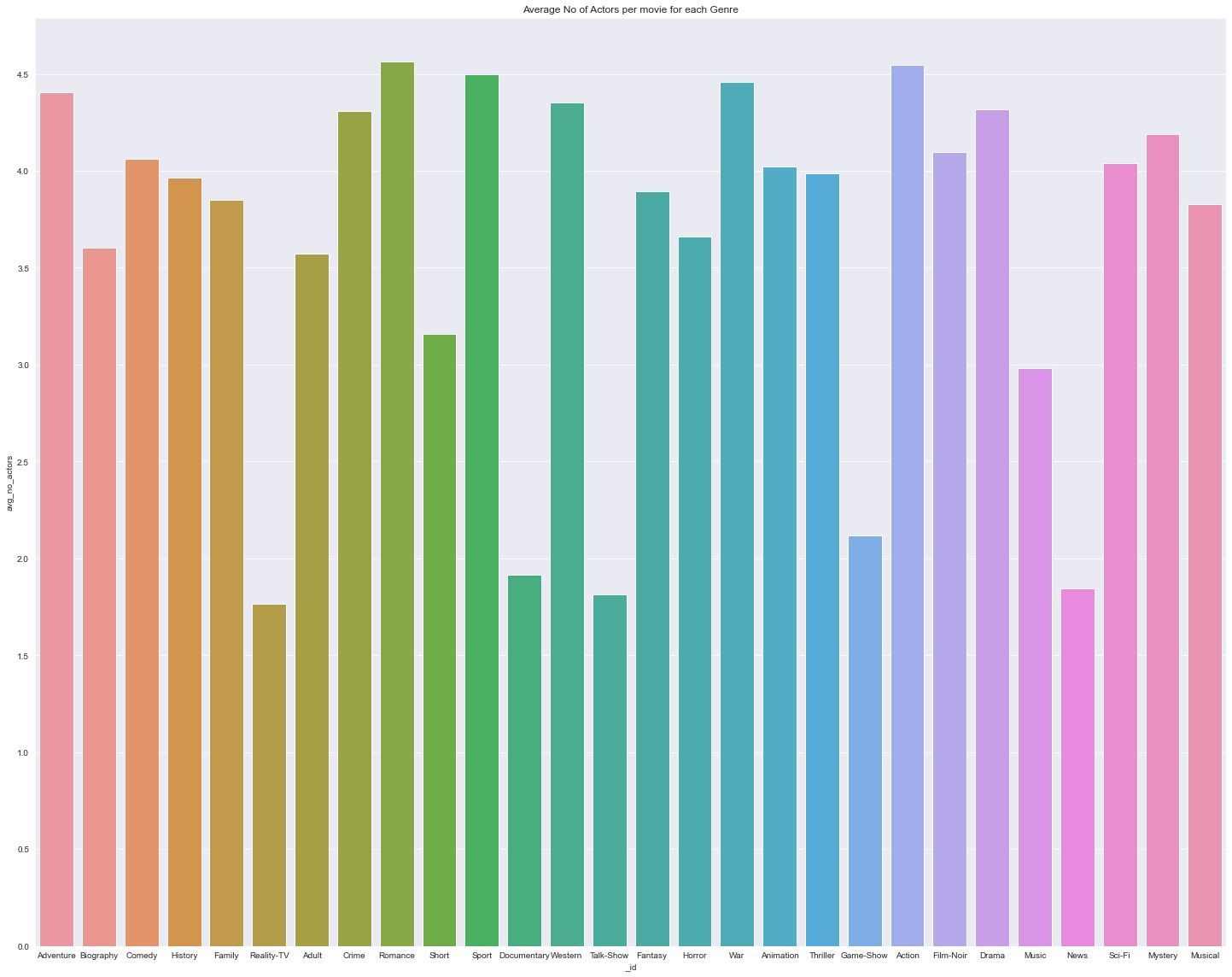
**(code only is available VizNotebook.py)**

**(pdf of plots is in viz.pdf)**

**Figure 1**

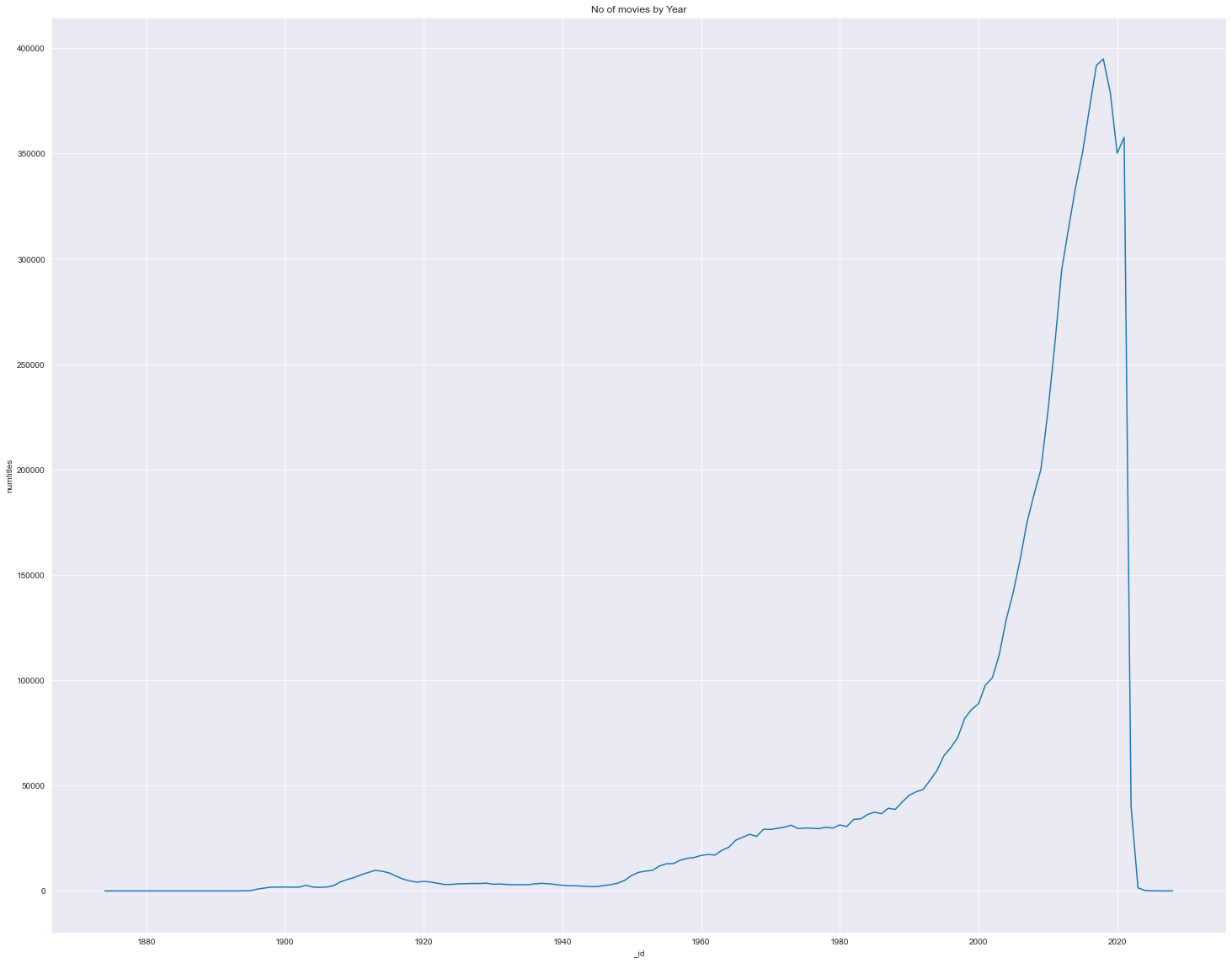
We can see a lot of outliers for each genres. Also news genre has the highest median among them and documentary has the highest third quartile among all genres.

**Figure 2**



Crime has the highest number of average actors per genre followed by action.

**Figure 3**



As expected we see the number of movie releases increase over time and drop as we have very few movies declaring their start year