Bryce Rothschadl
Dr. Hien Nguyen
CS767: Big Data and Data Mining
2025-02-26

Assignment 2: Report

Schemas

Table A (Metacritic)
ID (Rank): Numeric
Title: Textual
Platform: Textual
Release Date: Textual
Metascore: Numeric

Table B (Opencritic)
ID (Rank): Numeric
Title: Textual
Platform: Textual
Release Date: Textual
Openscore: Numeric

The Metascore and Openscore attributes are both scores, however they are independent of one another so will not be included in the set S. Likewise, the rank attribute for each table is directly dependent on the respective score attributes, so it too will not be included in S. This leaves us with the Title, Platform, and Release Date attributes. An "s_ID" attribute will be introduced as the new key, it will be calculated by !!!

S
s_ID: Numeric
Title: Textual
Platform: Textual
Release Date: Textual

Missing Data Report

Missing Values in S considering Table A only:

Missing Values in S (Table A)		
Attribute (from A)	Fraction	Percentage
Title	0000/2500	0%
Platform	2500/2500	100%
Release Date	0000/2500	0%

Missing Values in S considering Table B only:

Missing Values in S (Table B)		
Attribute (from B)	Fraction	Percentage
Title	0000/2500	0%
Platform	0000/2500	0%
Release Date	0000/2500	0%

While the Platform attribute from Table A is completely empty (I had not realized this while working on assignment 1), it's worth noting that the information for this data for some games is sprinkled in the title instead. This may be useful for filling missing data later.

Possible Missing Data Solutions

One approach to filling the missing platform data from table A is to match as many games as possible in table B using the title and release data attributes. The lists are not identical, so this will not find every single game, however there is significant overlap between the two so this approach would fill in a large portion of the missing data. We would simply need to use some string matching technique(s) to find an entry from each table with matching titles.

However, some games share the same/very similar titles, such as:

Table	Rank	Title	Platform	Release Date	Score
A	581	Ratchet & Clank	N/A	"Nov 4, 2002"	88
A	1289	Ratchet & Clank (PS4)	N/A	"Apr 12, 2016"	85
B	394	Ratchet & Clank	"PS4, PS5"	"Apr 12, 2016"	86

A simple string matching technique might find the first and third rows to be the same game, as they share the exact same title across the two tables. However, we can use the release date to determine that this is not the case. The second and third rows are actually the same game (despite having slightly different strings for the title attributes across the two tables) and it is a sequel to the game that was released over a decade earlier from the first row. Keeping this in mind, it would be wise to also consider the release date attribute during this process. If we use a string matching technique to find similar games then we should also check that they were released at the same time.

Another approach is to check the title of games in table A for the platform. Using the same games from the example from the previous approach, we can safely assume that *"Ratchet & Clank (PS4)"* is available on the PS4. Likewise, we can probably assume that *"Mario Kart Wii"* is available on the Wii.

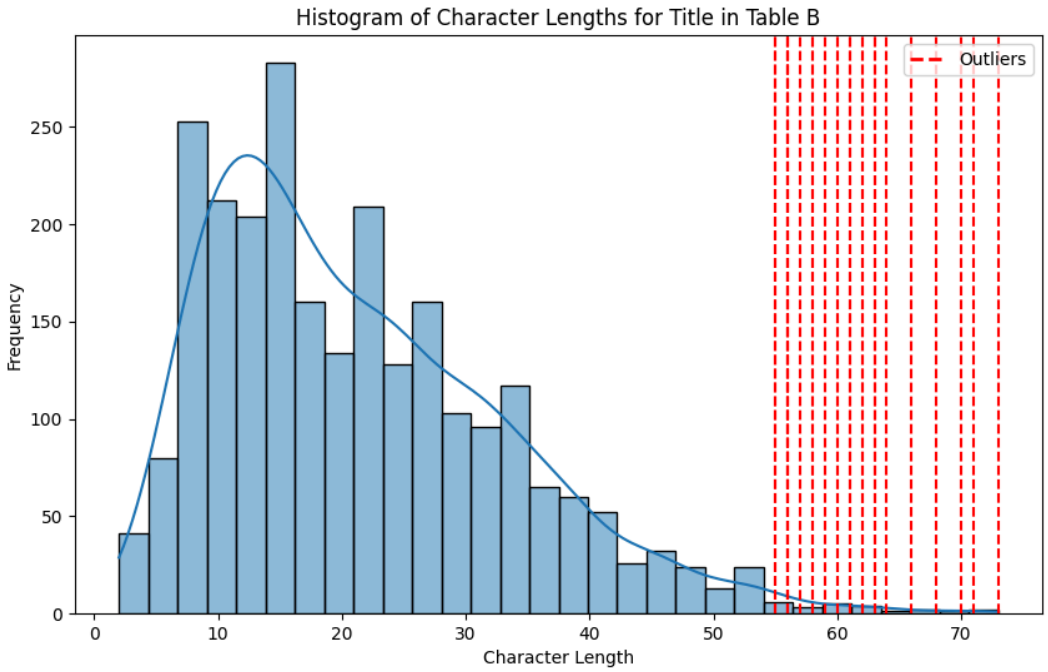
However, this is not a perfect solution. For example the game with ID 2224 in table A is titled *"The Sega Genesis Collection"*, however this game is not actually available on the Sega Genesis platform.

Textual Report

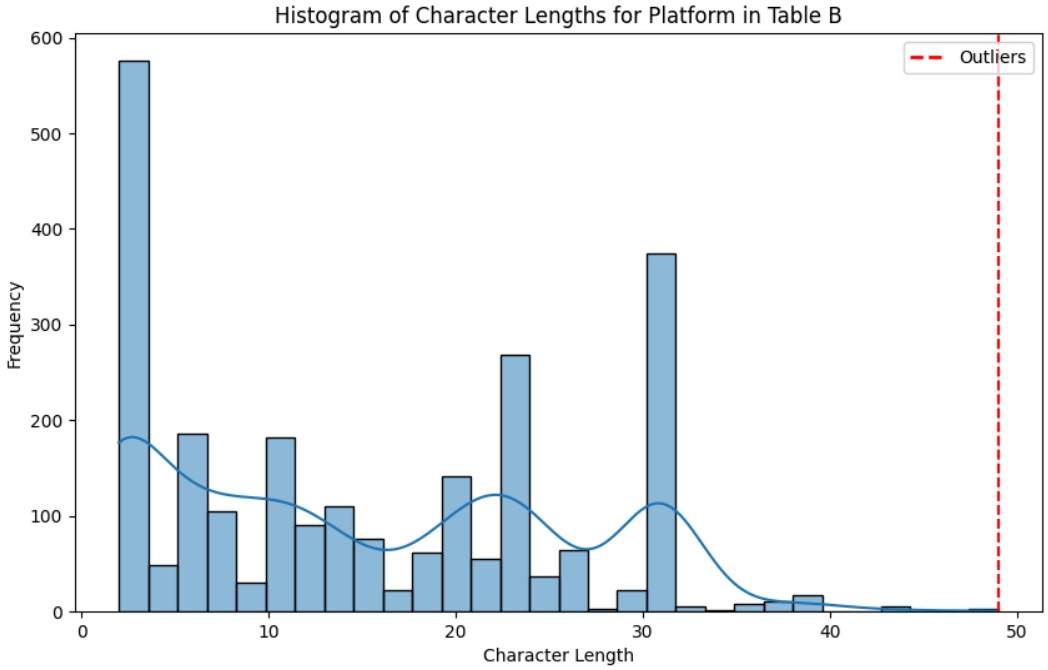
Table A			
Attribute	Avg Length	Min Length	Max Length
Title	21.43	2.00	73.00
Platform	0.00	0.00	0.00
Release Date	11.75	11.00	12.00

Table B			
Attribute	Avg Length	Min Length	Max Length
Title	21.24	2.00	73.00
Platform	15.05	2.00	49.00
Release Date	11.75	11.00	12.00

Histograms (Table B)



Outliers detected in Title (character lengths):
[55, 55, 56, 56, 56, 56, 57, 58, 58, 59, 60, 60, 61, 61, 62, 62, 63, 63, 64, 66, 68, 70, 71, 73]



Outliers detected in Platform (character lengths):
[49, 49]

As far as I am aware, all attributes correctly follow their specified format, besides the missing values in Table A.

Software Tools Used

Python packages:

- Matplotlib
- pandas
- Seaborn