How did you combine the two tables A and B to obtain E? Did you add any other table? When you did the combination, did you run into any issues? Discuss the combination process in detail, e.g., when you merge tuples, what are the merging functions (such as to merge two age values, always select the age value from the tuple from Table A, unless this value is missing in which case we select the value from the tuple in Table B).

We wrote a python script that parsed the Golden Data, and looked up the matching tuples in the two tables A and B by their unique identifiers. We resolved the data from the two tuples into the common schema for table E. There were often cases of narrowly inconsistent values while joining the attributes so we had to create strategies for various data types.

- We resolved conflicts in the textual attributes such as ‘Album Name’, ‘Genres’, ‘Label’ and ‘Artist’ by adopting the superset approach and picking the longer string out of the two.
- To resolve conflict in the ‘Time’ attribute, we picked the larger time duration from the two tuples.
- To resolve conflicts in the date attribute, ‘Released’, we picked the earlier date from between the two.
- To resolve conflicts in the ‘Price’ (money) attribute, we simply picked the value from, i.e. placed higher trust in, Table A.
- And we always picked the ‘Ratings’ attribute from the Amazon Music table (B) because it didn’t exist for table A.

We also added a table D at this stage to add some extra volume to the final dataset; So that we could perform better analytics on it. We generated table D by scraping additional music data from Amazon Music and ran it through the existing pipeline of cleaning, understanding and transforming.

Lastly, we also added an extra meta label to the final tuple in table E. Based on whether or not the term ‘Explicit’ appeared in the Album or Track name we set the Content Type as ‘Explicit’ or ‘Clean’ accordingly.

Statistics on Table E: specifically, what is the schema of Table E, how many tuples are in Table E? Give at least four sample tuples from Table E.

➔ The schema of table E is composed of the following parts

- Overlapping attributes from tables A and B. There were 8 attribute pairs between the two tables.
- We also included an attribute that was unique to table B
- We added a new attribute at this stage which is a Content Type label that has either of two values ‘Explicit’ or ‘Clean’.
- We added the unique identifiers from each table A, B and D so as to be able to do a reverse lookup later.
- We also added a new identifier for table E.

➔ These are all the attributes in table E

{ IdE, IdA, IdB, Album, Genres, Label, Time, Track Name, Price, Artist, Released, Rating, IdD, Type }

➔ There are a total of 94 tuples in table E

➔ Here is an example of a few randomly chosen tuples from the table

IdE	IdA	IdB	Album	Genres	Label	Time	Track Name	Price	Artist	Released	Rating	IdD	Type
9	12	s_938	1989	Pop, Music, Rock	2014 Big Machine Records, LLC.	0:03:15	I Know Places	$1.29	Taylor Swift	10/27/2014	0.666667		Clean
10	13	s_939	1989	Pop, Music, Rock	2014 Big Machine Records, LLC.	0:04:31	Clean	$1.29	Taylor Swift	10/27/2014 0	0.666667		Clean
11	224	s_912	Anti (Deluxe) [Explicit]	Pop, Music	2016 Westbury Road Entertainment.	0:01:12	James Joint [Explicit]	$1.29	Rihanna	1/29/2016	0.333333		Explicit
12	225	s_913	Anti (Deluxe) [Explicit]	Pop, Music	2016 Westbury Road Entertainment.	0:04:13	Kiss It Better [Explicit]	$1.29	Rihanna	5/2/2017	1.666667		Explicit
108			Fearless	Rap & Hip-Hop	Brainsick Muzik	0:01:52	Sweet Dreams [Explicit]	$0.99	Twisted Insane	10/31/2015 0:00	4	d_34	Explicit

What was the data analysis task that you wanted to do? (Example: we wanted to know if we can use the rest of the attributes to accurately predict the value of the attribute loan_repaid.) For that task, describe in detail the data analysis process that you went through.

We hypothesised a correlation between the Content Type and attributes such as Rating and Duration. We performed OLAP-style exploration and we aggregated the duration and rating metrics across ‘Explicit’ and ‘Clean’ tracks and mapped it on a scatter plot to visualize any underlying trends or patterns. Figure below shows the scatter plot that we created using Matplotlib and Plotly.

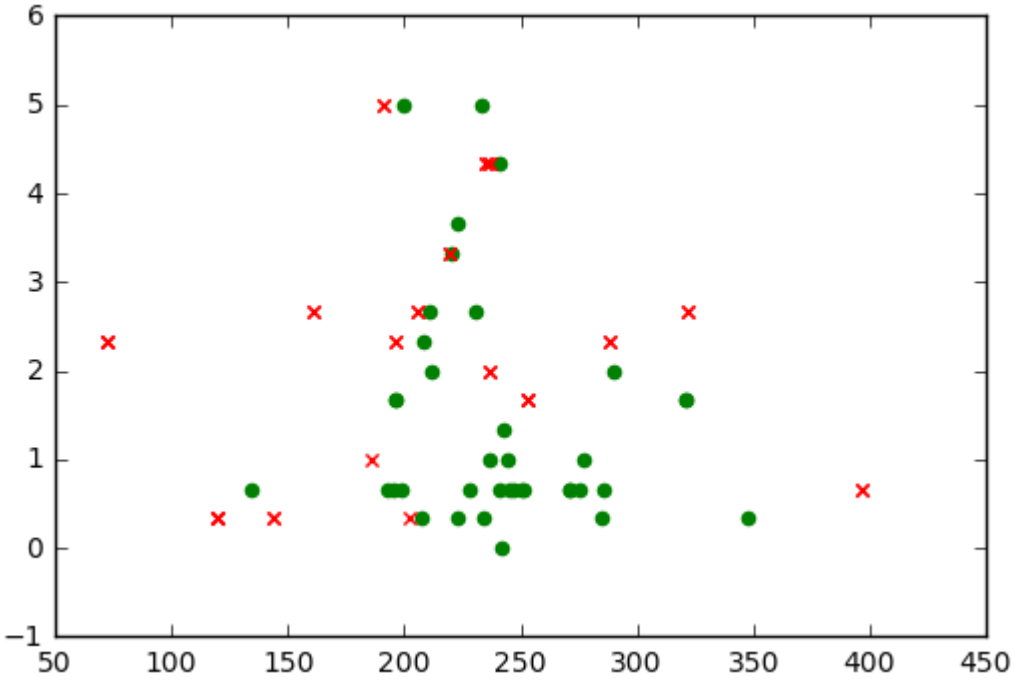

Figure: We’ve plotted the ‘Time’ attribute on the horizontal axis in seconds and ‘Rating’ in range 0 to 5 on the vertical axis. The Green dots are the ‘Clean’ tracks and the Red crosses correspond to the ‘Explicit’ tracks.

We also calculated the average Rating across the type dimension and while doing so we skipped over the unrated tracks to avoid then from skewing the final score.

Give any accuracy numbers that you have obtained (such as precision and recall for your classification scheme).

The average Rating scores for ‘Clean’ tracks is 1.486 and for ‘Explicit’ tracks it is 2.703.

What did you learn/conclude from your data analysis? Were there any problems with the analysis process and with the data?

After analyzing the scatter plot and the average Rating scores, we concluded that on an average the ‘Explicit’ are more highly rated than ‘Clean’ tracks.

However in our analysis we only looked at a small subset of songs spread across a narrow range of genres. To be able to conclusively verify our hypothesis we require a larger and more diverse dataset.

If you have more time, what would you propose you can do next?

We attempted as part of this Stage to write a classifier that would predict Rating of the song based on the Artist, Genre and Content Type of the tuple but due to the lack to good training data we weren’t able to train a classifier that would generalize well.

If we had more time, we would scrape or anyhow acquire more data to finish training our classifier.