1. **The names and emails of members in your group**
    a. Sneha Rudra (rudra@wisc.edu)
    b. Shantanu Singhal (singhal5@wisc.edu)
    c. Zirui Tao (ztao23@wisc.edu)

2. **List the schema of the two tables**

    This was the original schema of the two tables

TableA	TableB
SNo, Album, SongName, Artist, Time, Price, Genres, Released, Rights	Id,Album,Genres,ASIN,Label,Time,Rating ,TrackName,Price,Artist,Released

    The scheme of the two tables was transformed to be as described in the next step. The updated CSV files under Project Stage 1 on the Project Website.

3. **List the attributes in the set S**

    S = {'Id', 'Album', 'Genre', 'Label', 'Time', 'Track', 'Price','Artist', 'Released'}

    The common scheme has 9 attributes.

4. **For each attribute X in S, consider only Table A and discuss the following in the report**

    a. **The missing values as percentage and fractions, classifications of each attribute and the min, max and average length of the various textual attributes** is discussed in the table below

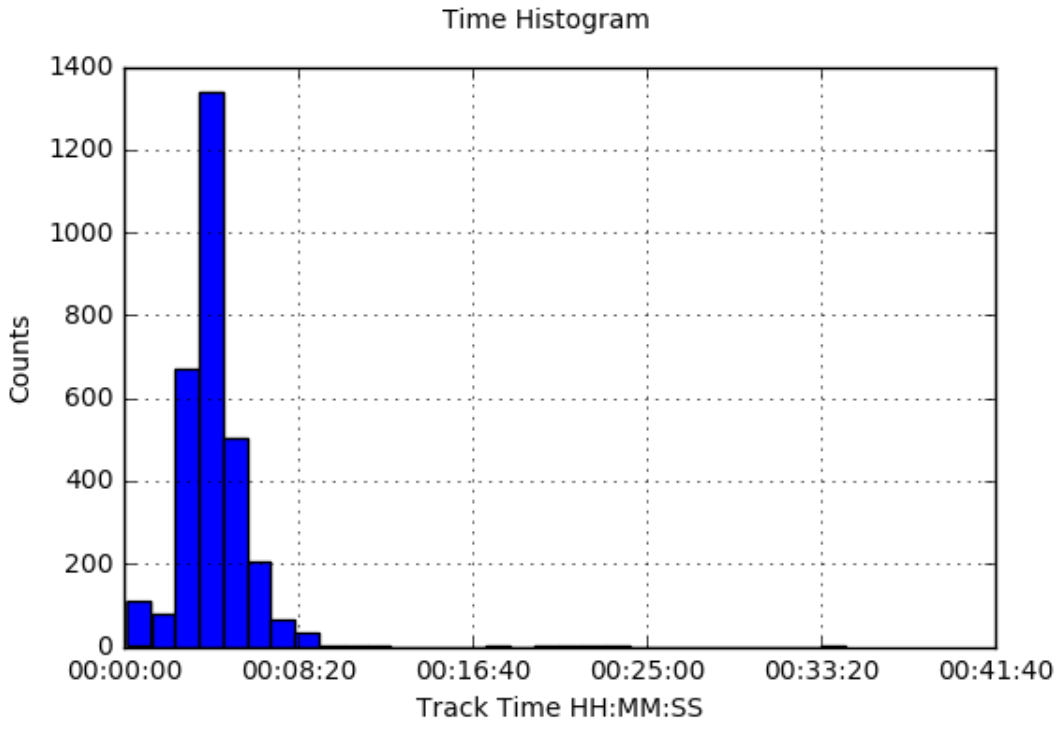
	Type	Missing Values		Length		
		Fraction	Percentage	Avg	Min	Max
Id	numeric		0			
Album	textual		0	31.86	1	90
Genres	categorical		0			
Label	textual	54 / 3019	1.79	31.26	13	141

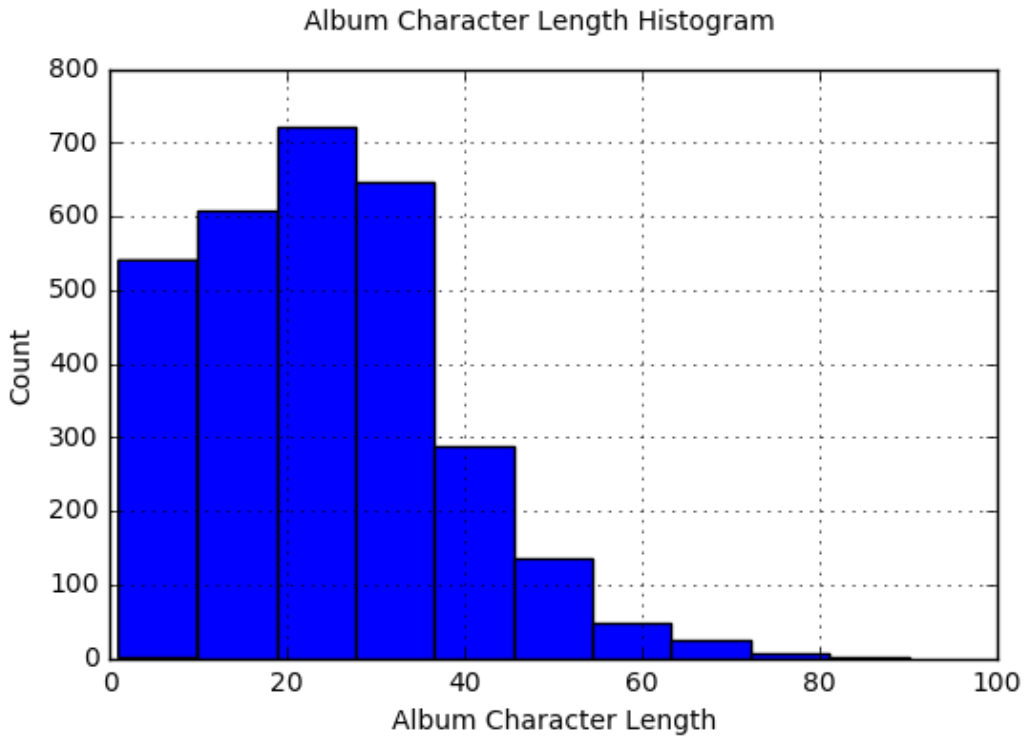
Time	Time		0			
Track	textual		0	26.59	1	104
Price	money	34 / 3019	1.13			
Artist	textual		0	20.15	4	85
Released	Date	1 / 3019	0.03			

b. **Proposed solutions for missing values**: matrix completion and/or replacing with flag values or average values.

c. **Show at least two histograms that your team has created. Find and report possible outliers and anomalies among the attribute values**

We picked two attributes to plot the histogram on our table: the “Time” attributes and “Album” attributes. Since Album belongs to the textual type. We plot its corresponding length distribution on histogram.





Based on the histograms that we created for track time and album name we've identified the following outliers for each

Track Duration Length	Album Name Length
1. 00:34:29 (id = 2786) 2. 00:34:25 (id = 2911)	1. 90 (id = 2980)

d. If the attribute value is supposed to follow a certain format (e.g., dates), then discuss if all values follow the same format, or if there is some problem with the format and we will have to standardize the formats later

	Expected Format
Genres	Comma separated list of Genre
Time	mm:ss

Price	$[0-9]+(\.[0-9]{1,2})
Released	The values in Released attribute should follow “dd-MMM-yy” but in some cases only the “yy” mentioned. Possible Solutions 1. We gather more information and update the tuple in the database 2. Split the “released ” attributes into “ released date” ,”released month”, “released year” and make the first two optional.

e. Are there synonyms among attribute values?

No, not in our datasets.

f. Sometimes attribute values are "sprinkled" all over the item. Do you have this problem with this attribute?

The *Artist* attribute is sprinkled in other attribute such as *Track* name. For example, in the track ‘Weight Of The World [feat. Blaque Keyz]’ by Jon Bellion, the name of another featuring artist Blaque Keyz appears but isn’t captured as a distinguished attribute.

g. Do you see any other data quality problems with this attribute?

Yes, the data we extracted (in Stage 1) was not encoded in UTF-8 which lead to IO errors while reading this data into the database for analysis. So we had to convert the file into UTF-8 encoding before feeding it to the database.

5. List any software tools that you have used to do the above data understanding and cleaning.

List of packages used: Pandas, Jupyter notebook, matplotlib, Numpy, Scipy