DATA COLLECTION & PRE-PROCESSING GROUP PROJECT

GROUP 25

Venkata Soma Sundaram Pappu – 12110110

Joydeep Vasudeva – 12010029

Phani Ayyalasomayajula – 12110007

Executive Summary

The objective of our **Data Collection & Pre-processing** group project is to create an end-to-end data collection and pre-processing pipeline for **"Grammy Awards"**. Our input data comprises of the "Grammy awards" domain seed for which the data collection from various external sources and pre-processing steps are carried out to convert into a structured source. One of the key objectives set in our data collection is to extract the data from various sources – structured and unstructured and ensure the reliability of the data.

We have started with the seed datafile and started enriching the source by extracting new fields from related "Grammy awards" sources across the world wide web. We have enhanced the seed datafile by adding new information rich and reliable attributes and increased the number of attributes to **40**.

We have started searching for our data sources through – Wikipedia artist pages, Grammy awards site, and leveraged Google search to extract the artist related key attributes. From these sources, we have retrieved related data with key attributes to enhance the datafile. For data extraction, we have utilized the following scraping python libraries which enable to access and parse data from the external data sources:

- **Beautifulsoup** – parsing HTML & XML sources
- **Pandas** – Data manipulation & analysis
- **Wikipedia** – Data analysis and parsing from Wikipedia. Get the snippets of URLs from Wikipedia artist pages, build artist URLs, and the related artist attributes.
- **Google search engine** – Retrieve additional relevant attributes through Google search appending the artist names using the "requests" python library.
- **Sweetviz** - High-density visualizations to perform EDA (Exploratory Data Analysis)
- **Tableau** – For merge/joins of the data from multiple sources via primary key

As the data collection was performed from multiple sources, we have created a unified knowledge database by merging the datasets based on the primary key field. The final unified dataset is stored in a CSV format. After the completion of data collection, we started to clean the data before putting it to use for further analysis. In this preliminary step, we classify the data into Missing data & Noisy data. For missing data, our approach is to ignore the entire tuple of data, and remove the records with missing values from the dataset. For noisy data, we have segmented available data into categories based on their characteristics and then processed it accordingly. For transforming the data, we have used the attribute selection from a pre-constructed range and use them accordingly. For data reduction and accuracy, we have only used the attributes that are highly relevant and discarded unnecessary data.

Key challenges which we encountered in our Data collection & pre-processing are:

- Missing data.
- Data inconsistencies.
- Human error leading with different data types.
- Data manipulation.
- Advance programming.

Domain and Seed sources:

- We have started with the seed datafile and started enriching the source by extracting new fields from related “Grammy awards”. Total number of relevant attributes **40**. We have extracted rich information across multiple data sources both structured and unstructured to prepare the most cleaned and accurate dataset.

Data Sources:

- Data sources crawled: Wikipedia artist pages, Grammy awards site, and Google search to extract the artist related key attributes. From these sources, we have retrieved related data with key attributes to enhance the datafile.

Data Crawling & Conversion methods:

- For data extraction, we have utilized the following scraping python libraries which enable to access and parse data from the external sources

    - **Beautifulsoup** – parsing HTML & XML sources
    - **Pandas** – Data manipulation & analysis
    - **Wikipedia** – Data analysis and parsing from Wikipedia. Get the snippets of URLs from Wikipedia artist pages, build artist URLs, and the related artist attributes.
    - **Google search engine** – Retrieve additional relevant attributes through Google search appending the artist names using the “requests” python library.
    - **Sweetviz** - High-density visualizations to perform EDA (Exploratory Data Analysis)
    - **Tableau** – For merge/joins of the data from multiple sources via primary key

Key Challenges encountered:

- Missing data either due to manual input or a random behavior.
- Data inconsistencies due to the variations created in the data sets and when multiple values are stored in the same field, which could lead to Data integrity.
- When storing information under different data types, human error could lead to incorrect representations of the data.
- Data manipulation – When working with software for datafiles, sometimes the data is broken during the regular file open/read/write operations.
- Advance programming needed to convert the desired data sources into a structured data.

Data cleaning/pre-processing methods:

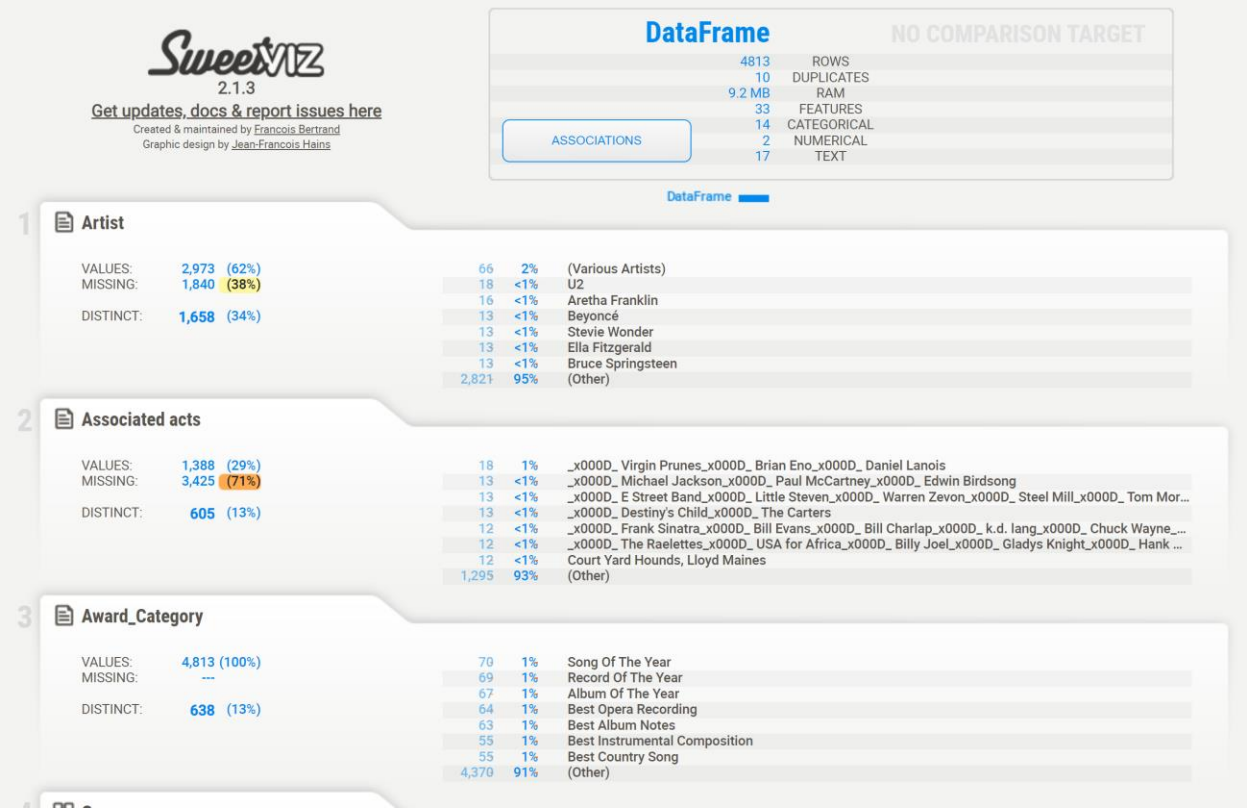
- Created a unified knowledge database by merging the datasets based on the primary key field.
- Clean the data before putting it to use for further analysis.

- Classify the data into Missing data & Noisy data. For missing data, our approach is to ignore the entire tuple of data, and remove the records with missing values from the dataset. For noisy data, we have segmented available data into categories based on their characteristics and then processed it accordingly.
- For transforming the data, we have used the attribute selection from a pre-constructed range and use them accordingly.
- For data reduction and accuracy, we have only used the attributes that are highly relevant and discarded unnecessary data.

Any Observations/ Insights and Analysis on the data collected?

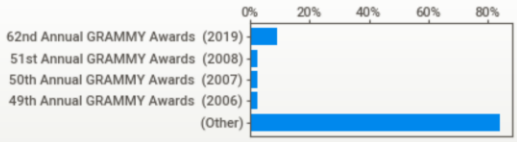
- We have performed an Exploratory Data Analysis (EDA), by using an open-source python library "Sweetviz".

Visualizations/Insights shared below:



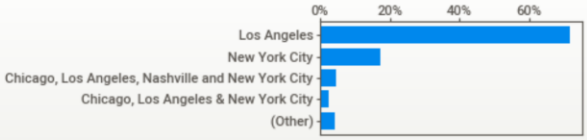
4 **Ceremony**

VALUES:	4,813	(100%)
MISSING:	---	
DISTINCT:	62	(1%)



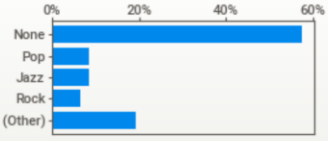
5 **City_Grammy_Held**

VALUES:	4,813	(100%)
MISSING:	---	
DISTINCT:	7	(<1%)



6 **Main_Genre**

VALUES:	2,741	(57%)
MISSING:	2,072	(43%)
DISTINCT:	28	(<1%)



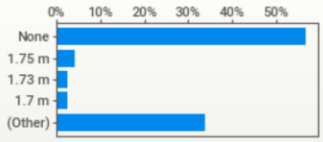
7 **Genres**

VALUES:	1,972	(41%)
MISSING:	2,841	(59%)
DISTINCT:	818	(17%)

44	2%	Jazz
27	1%	Country
25	1%	Hip hop
18	<1%	_x000D_ Rock_x000D_ alternative rock_x000D_ pop rock_x000D_ post-punk
16	<1%	_x000D_ Soul_x000D_ R&B[1]_x000D_ gospel_x000D_ jazz_x000D_ pop
16	<1%	Countrycountry pop
15	<1%	_x000D_ Polka_x000D_ rock_x000D_ country
1,811	92%	(Other)

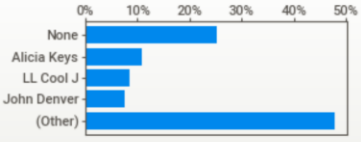
8 **Height_of_artist**

VALUES:	2,741	(57%)
MISSING:	2,072	(43%)
DISTINCT:	56	(1%)



9 **Host_of_grammy**

VALUES:	4,813	(100%)
MISSING:	---	
DISTINCT:	21	(<1%)



10 **Img_of_artist**

VALUES:	3,446	(72%)
MISSING:	1,367	(28%)
DISTINCT:	1,463	(30%)

26	<1%	https://www.grammy.com/sites/com/files/styles/artist_circle/public/muzooka/John%2BWilliams/John...
22	<1%	https://www.grammy.com/sites/com/files/styles/artist_circle/public/muzooka/U2/U2_1_1_157838523...
21	<1%	https://www.grammy.com/sites/com/files/styles/artist_circle/public/muzooka/Vladimir%2BHorowitz/...
20	<1%	https://www.grammy.com/sites/com/files/styles/artist_circle/public/muzooka/Henry%2BMancini/Hen...
20	<1%	https://www.grammy.com/sites/com/files/styles/artist_circle/public/georgsolti-spotlight-78824961.pn...
19	<1%	https://www.grammy.com/sites/com/files/styles/artist_circle/public/muzooka/Bruce%2BSpringsteen/...
19	<1%	https://www.grammy.com/sites/com/files/styles/artist_circle/public/muzooka/Pierre%2BBoulez/Pierr...
3,299	96%	(Other)

11 **Instruments**

VALUES:	1,219	(25%)
MISSING:	3,594	(75%)
DISTINCT:	257	(5%)

216	18%	Vocals
62	5%	Vocalsguitar
51	4%	_x000D_ Vocals_x000D_ guitar
45	4%	Vocals, guitar
41	3%	_x000D_ Vocals_x000D_ piano
35	3%	Piano
23	2%	Vocalspiano
746	61%	(Other)

12 **Labels**

VALUES:	1,784	(37%)
MISSING:	3,029	(63%)
DISTINCT:	775	(16%)

29	2%	Columbia
18	1%	_x000D_ Island_x000D_ Interscope_x000D_ Mercury_x000D_ CBS Ireland
16	<1%	_x000D_ J.V.B_x000D_ Columbia_x000D_ Atlantic_x000D_ Arista_x000D_ RCA
15	<1%	Rounder, Vanguard, Starr
13	<1%	_x000D_ Tamla_x000D_ Motown_x000D_ So What the Fuss Records
13	<1%	_x000D_ Decca_x000D_ Verve_x000D_ Capitol_x000D_ Reprise_x000D_ Pablo
13	<1%	_x000D_ Parkwood_x000D_ Columbia_x000D_ Music World
1,667	93%	(Other)

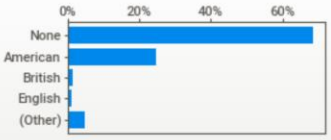
13 **Members**

VALUES:	373	(8%)
MISSING:	4,440	(92%)
DISTINCT:	180	(4%)

18	5%	_x000D_ Bono_x000D_ Adam Clayton_x000D_ The Edge_x000D_ Larry Mullen Jr.
12	3%	_x000D_ Natalie Maines_x000D_ Emily Strayer_x000D_ Martie Maguire
10	3%	_x000D_ Dave Grohl_x000D_ Nate Mendel_x000D_ Pat Smear_x000D_ Taylor Hawkins_x000D_ Chris Shi...
8	2%	_x000D_ Alan Paul_x000D_ Janis Siegel_x000D_ Cheryl Bentyne_x000D_ Trist Curless
7	2%	Jorge HernÃ¡ndez1968-present _x000D_ HernÃ¡n HernÃ¡ndez 1968-present _x000D_ Eduardo HernÃ¡nd...
7	2%	_x000D_ Alvin Chea_x000D_ Khristian Dentley_x000D_ Joey Kibble_x000D_ Mark Kibble_x000D_ Claude...
7	2%	_x000D_ James Hetfield_x000D_ Lars Ulrich_x000D_ Kirk Hammett_x000D_ Robert Trujillo
304	82%	(Other)

14 **Nationality**

VALUES:	2,735	(57%)
MISSING:	2,078	(43%)
DISTINCT:	23	(<1%)



15 **Notable works**

VALUES:	13	(<1%)
MISSING:	4,800	(>99%)
DISTINCT:	11	(<1%)

3	23%	I Know Why the Caged Bird Sings"On the Pulse of Morning"
1	8%	Original cast member of Saturday Night Live
1	8%	The Andy Griffith ShowMatlock
1	8%	Tom Haverford in Parks and RecreationChet in 30 Minutes or LessDev in Master of None
1	8%	Chicago PoemsThe People, YesAbraham Lincoln: The Prairie Years and The War YearsRootabaga Stories
1	8%	The Flip Wilson Show
1	8%	The Carol Burnett ShowMiss Agatha Hannigan in Annie Eunice Harper Higgins on Mama's Family
4	31%	(Other)

16 **Occupation**

VALUES:	610	(13%)
MISSING:	4,203	(87%)
DISTINCT:	177	(4%)

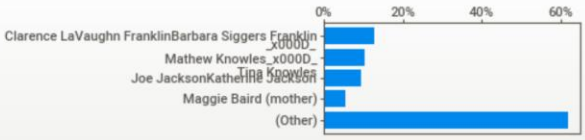
23	4%	Singersongwriter
19	3%	_x000D_ Singer_x000D_ songwriter
19	3%	_x000D_ Singer_x000D_ songwriter_x000D_ musician
18	3%	Singer
17	3%	_x000D_ Musician_x000D_ singer_x000D_ songwriter_x000D_ record producer
16	3%	_x000D_ Singer_x000D_ songwriter_x000D_ actress_x000D_ pianist_x000D_ civil rights activist_x000D_...
14	2%	_x000D_ Singer_x000D_ songwriter_x000D_ musician_x000D_ record producer
484	79%	(Other)

17 **Official Page URL**

VALUES:	4,813	(100%)
MISSING:	---	
DISTINCT:	62	(1%)

18 **Parent(s)**

VALUES:	126	(3%)
MISSING:	4,687	(97%)
DISTINCT:	35	(<1%)



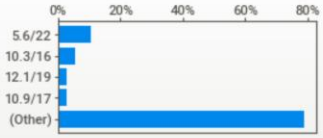
19 **Place of Birth**

VALUES:	2,732	(57%)
MISSING:	2,081	(43%)
DISTINCT:	295	(6%)

1,686	62%	None
66	2%	Long Beach, California, United States
30	1%	American
29	1%	Los Angeles, California, United States
23	<1%	New York, New York, United States
23	<1%	Detroit, Michigan, United States
23	<1%	Philadelphia, Pennsylvania, United States
852	31%	(Other)

20 Rating/Share (Households)_TRP

VALUES:	4,137	(86%)
MISSING:	676	(14%)
DISTINCT:	45	(<1%)



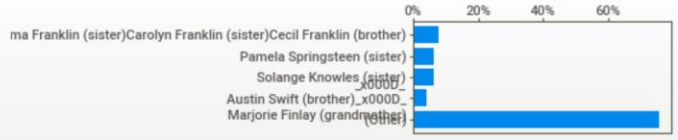
21 Real Name

VALUES:	2,741	(57%)
MISSING:	2,072	(43%)
DISTINCT:	638	(13%)

1,100	40%	None
71	3%	Prince Rogers Nelson
18	<1%	Ray Charles Robinson
17	<1%	Aretha Louise Franklin
17	<1%	Anthony Dominick Benedetto
15	<1%	Alison Maria Krauss
15	<1%	James William Sturr Jr.
1,488	54%	(Other)

22 Relatives

VALUES:	204	(4%)
MISSING:	4,609	(96%)
DISTINCT:	58	(1%)



23 Relavant wiki URL

VALUES:	2,741	(57%)
MISSING:	2,072	(43%)
DISTINCT:	1,401	(29%)

66	2%	https://en.wikipedia.org/wiki/Compilation_album
18	<1%	https://en.wikipedia.org/wiki/Jimmy_Sturr
18	<1%	https://en.wikipedia.org/wiki/U2
16	<1%	https://en.wikipedia.org/wiki/Aretha_Franklin
13	<1%	https://en.wikipedia.org/wiki/Beyonc%C3%A9
13	<1%	https://en.wikipedia.org/wiki/Ella_Fitzgerald
13	<1%	https://en.wikipedia.org/wiki/Bruce_Springsteen
2,584	94%	(Other)

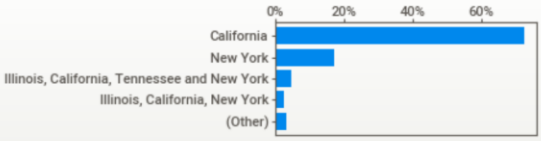
24 Spouse Name

VALUES:	2,738	(57%)
MISSING:	2,075	(43%)
DISTINCT:	360	(7%)

1,915	70%	None
18	<1%	Lesley Jones
17	<1%	Jay Z
13	<1%	Jessica Biel
13	<1%	Michelle Obama
12	<1%	Jodi Stewart
12	<1%	Leona Johnson
738	27%	(Other)

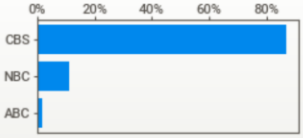
25 State

VALUES:	4,813	(100%)
MISSING:	---	
DISTINCT:	6	(<1%)



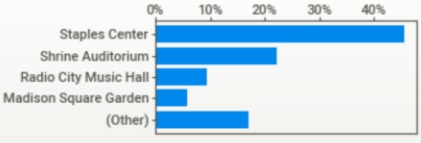
26 Television/Radio Coverage

VALUES:	4,813	(100%)
MISSING:	---	
DISTINCT:	3	(<1%)

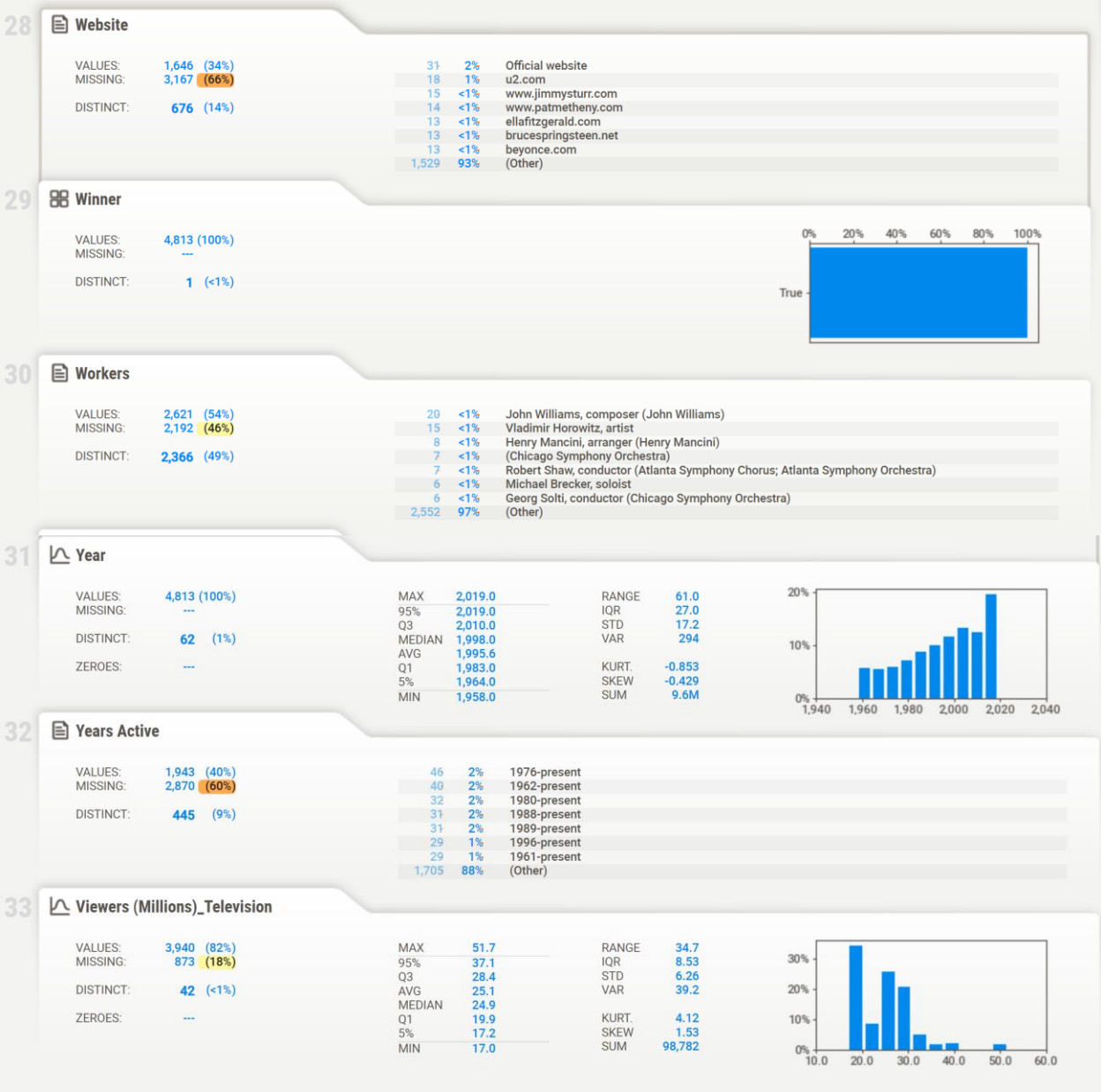


27 Venue

VALUES:	4,813	(100%)
MISSING:	---	
DISTINCT:	11	(<1%)



28 Website



Strategy to enhance the data with crowd sourcing methods:

- The Grammy awards is a past data and does not apply to the crowd sourcing strategy.

References and Sources used for this Assignment:

- Wikipedia
- Discogs.com
- Grammy.com
- Python.org
- https://stackoverflow.com