SRH HOCHSCHULE HEIDELBERG
Intelligence in Learning

Project Report
for
Data Management

Under the supervision of
Prof. Dr. Ajinkya Prabhune
and
Mr. Ashish Chouhan

Topic: Data Uncleaning

Submitted by –

Anurag Singh

Introduction

The document contains an overview of the process of uncleaning of a dataset by considering various data quality dimensions. The uncleaning process was performed under the guidance of Prof. Dr. Ajinkya Prabhune and Mr. Ashish Chouhan.

Dataset

The dataset used for the uncleaning process is Spotify Tracks dataset which contains a list of spotify tracks along with their details. There are 17 columns and 20,000 rows in the dataset. The source of the dataset is Kaggle.com.

Column_name	Data_type
Id	Char
name	String
album	String
album_id	Char
artists	String
artist_ids	Char
track_number	Int
explicit	Boolean
loudness	Float
speechiness	Float
instrumentalness	Float
duration_ms	Int
time_signature	int
year	Int
release_date	date

Tools

The tool used for data uncleaning is Open-refine. The language used within the tool to perform different operations is GREL and Python.

Selection of rows for Uncleaning

The rows were randomly selected by using Python code in open-refine.

Code-> *list = [random.randint(1, 20000) for i in range(0,850)]*

A separate column is added to document the rows related to the data quality dimension so that it would be easier to verify the uncleaned rows and respective columns.

Uncleaning

1. **Uniqueness**

➔ I duplicated 801 rows to perform uncleaning for uniqueness quality dimension.

2. **Validity**

➔ I changed the date format for 733 rows of the "release_date" column to perform uncleaning for Validity quality dimension.

3. **Consistency**

➔ I transformed the 760 rows of the "name" column to uppercase to perform uncleaning for Consistency quality dimension.

4. **Accuracy**

➔ I added special characters to the 784 rows of the "album" column to perform uncleaning for Accuracy quality dimension.

5. **Completeness**

➔ I blanked down 851 rows of the "year" column to perform uncleaning for Completeness quality dimension.

6. **Conformity**

➔ I used different a term “0” for same concept “False” for 782 rows of the “explicit” column to perform uncleaning for Conformity quality dimension.

7. **Timliness**

➔ I changed the date of 888 rows of the “last_updated” column to perform uncleaning for Timeliness quality dimension.

As the last_updated column is manually added to meet the timeliness quality dimension, so I have to add the dates randomly.

Conclusion

The goal of cleaning 25% of the 20,000 records was achieved by performing above operations.

25% of 20,000 = 25*20000/100 = 5000

Uncleaned no. of rows = 801+733+760+784+851+782+888

= 5599

So, the no. of uncleaned records is greater than 25%.

Bibliography

References:-

i) Dataset Source.
https://www.kaggle.com/rodolfofigueroa/spotify-12m-songs