CIS4560 Term Project Tutorial

Authors: Michael Miranda; Daniel Garrido; Giovanni Munoz; Sereyoudom Eab; Uriel Guijarro; Yana Pol

Instructor: Jongwook Woo

Date: 12/14/2023

Lab Tutorial

12/14/2023

Discogs Data Analysis using Hadoop

Objectives

List what your objectives are. In this hands-on lab, you will learn how to:

- Process data using Hadoop/HDFS
- Use SQL commands to perform the analysis.
- Create visualizations

Platform Spec

- CPU Speed: 1995.312 MHz
- # of CPU cores: 8
- # of nodes: 3
- Total Memory Size: 58 gb

Step 1: Download dataset and setup for Hadoop

This step will retrieve dataset from Kaggle, clean the dataset to ensure data integrity and accuracy, create environment for dataset placement, and finally create external table for data query and analysis.

1. Go to https://www.kaggle.com/datasets/ofurkancoban/discogs-releases-dataset and download dataset.

2. Using Git Bash, secure copy "archive.zip" folder to home directory in Linux Server using the following command:

```
scp C:/Users/Giovanni/Downloads/archive.zip
gmunoz58@129.146.90.117:/home/gmunoz58/
```

3. Secure shell into Linux Server:

```
ssh gmunoz58@129.146.90.117
```

4. Check for file, if exist then unzip

```
-bash-4.2$ ls -l
total 3465044
-rw-r--r-- 1 gmunoz58 gmunoz58 3548204891 Dec 16 04:32 archive.zip
-bash-4.2$ unzip archive.zip
Archive:  archive.zip
  inflating: discogs.csv
  inflating: discogs.sql
-bash-4.2$
-bash-4.2$ ls
archive.zip  discogs.csv  discogs.sql
-bash-4.2$
```

5. Remove "archive.zip" and SQL file to relieve space with the following command:

```
rm archive.zip discogs.sql
```

6. Run GREP command on “discogs.csv” file. The output will show the number of pipe characters in the file.

```
grep -c '|' discogs.csv
```

```
-bash-4.2$
-bash-4.2$ grep -c '|' discogs.csv
51659
```

7. Run SED command to remove all pipe characters from “discogs.csv” file. Then run GREP command to ensure that all pipe characters are deleted.

```
sed -i 's/|//g' discogs.csv
```

```
-bash-4.2$ sed -i 's/|//g' discogs.csv
-bash-4.2$ grep -c '|' discogs.csv
0
```

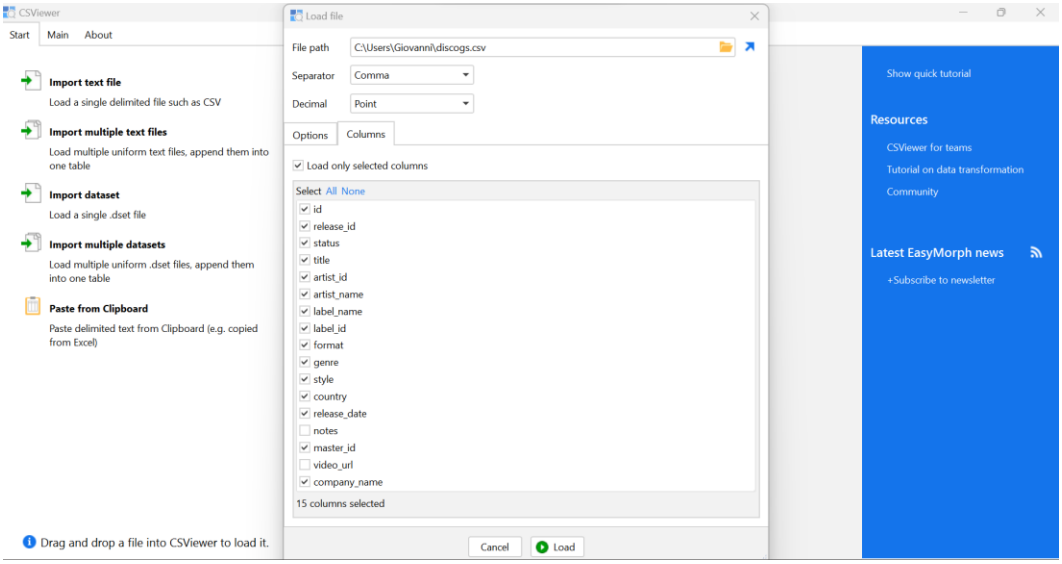
8. Exit Linux server, and secure copy “discogs.csv” file to local machine

```
scp gmunoz58@129.146.90.117:/home/gmunoz58/discogs.csv .
```

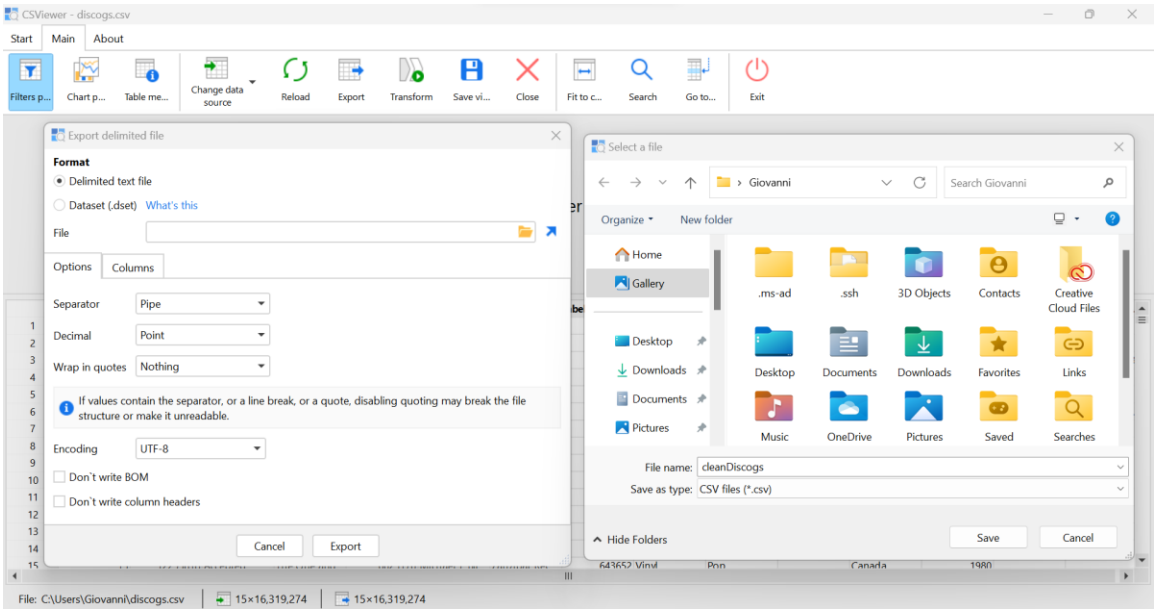
9. Go to https://csviewer.com/ and download CSViewer.
10. Open CSViewer, click **Import text file**, select **discogs.csv** file, click **Columns** tab, check Load only selected columns and check the following:

id	title	label_name	genre	release_date
release_id	artitst_id	label_id	style	master_id
status	artist_name	format	country	company_name

then click **Load**.



11. Click **Export**, change **Separator** to **Pipe**, change **Wrap in quotes** to **Nothing**, then export as **cleanDiscogs** csv file.



12. Using Git Bash, secure copy "cleanDiscogs.csv" file from local machine to Linux server using the following command.

```
scp C:/Users/Giovanni/cleanDiscogs.csv
gmunoz58@129.146.90.117:/home/gmunoz58/
```

13. Secure shell into Linux server and check for **cleanDiscogs.csv** file using **ls –l** command
14. Create DiscogsDataset directory in Hadoop File System, then confirm directory was created using the following commands:

```
hdfs dfs -mkdir DiscogsDataset

hdfs dfs -ls
```

```
-bash-4.2$ hdfs dfs -mkdir DiscogsDataset
-bash-4.2$ hdfs dfs -ls
Found 2 items
drwxr-xr-x   - gmunoz58 hdfs          0 2023-11-01 21:40 .hiveJars
drwxr-xr-x   - gmunoz58 hdfs          0 2023-12-16 05:45 DiscogsDataset
-bash-4.2$
```

15. Put cleanDiscogs.csv file into DiscogsDataset directory, then confirm file is in directory using the following commands:

```
hdfs dfs -put cleanDiscogs.csv DiscogsDataset/

hdfs dfs -ls DiscogsDataset/
```

```
-bash-4.2$ hdfs dfs -put cleanDiscogs.csv DiscogsDataset/
-bash-4.2$ hdfs dfs -ls DiscogsDataset/
Found 1 items
-rw-r--r--   3 gmunoz58 hdfs 2227795366 2023-12-16 05:47 DiscogsDataset/cleanDiscogs.csv
-bash-4.2$
```

16. Enter beeline, and create Group 5 database using the following commands:

```
beeline
```

```
CREATE DATABASE if not exists group5;
```

17. Confirm “group5” database was created.

```
show databases;
```

18. Make sure to use “group5” database.

```
use group5;
```

19. Create external table using dataset and confirm table creation.

```
CREATE EXTERNAL TABLE IF NOT EXISTS plzwork( id STRING, release_id
STRING, status STRING, title STRING, artist_id STRING, artist_name
STRING, label_name STRING, label_id STRING, format STRING, genre
STRING, style STRING, country STRING, release_date STRING, master_id
STRING, company_name STRING)
```

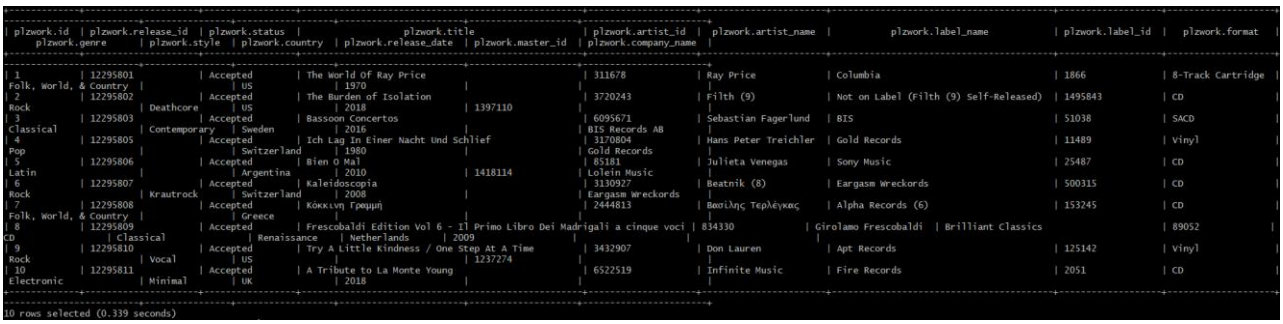
```
ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'
```

```
STORED AS TEXTFILE LOCATION '/user/gmunoz58/DiscogsDataset/'
```

```
TBLPROPERTIES ('skip.header.line.count'='1');
```

```
show tables;
```

```
SELECT * FROM plzwork LIMIT 10;
```



plzwork.id	plzwork.release_id	plzwork.status	plzwork.title	plzwork.artist_id	plzwork.artist_name	plzwork.label_name	plzwork.label_id	plzwork.format	plzwork.genre	plzwork.style	plzwork.country	plzwork.release_date	plzwork.master_id	plzwork.company_name
1	12295801	Accepted	The World Of Ray Price	311678	Ray Price	Columbia	1866	8-Track Cartridge	Folk, World, & Country		US	1970		
2	12295802	Accepted	The Burden of Isolation	3720243	Filth (9)	Not on Label (Filth (9) Self-Released)	1495843	CD	Rock	Deathcore	US	2018	1397110	
3	12295803	Accepted	Bassoon Concertos	6095671	Sebastian Fagerlund	BIS	51038	SACD	Classical	Contemporary	Sweden	2016		BIS Records AB
4	12295805	Accepted	Ich Lag In Einer Nacht Und Schlief	3170804	Hans Peter Treichler	Gold Records	11489	Vinyl	Pop		Switzerland	1980		Gold Records
5	12295806	Accepted	Bien O Mal	85181	Julieta Venegas	Sony Music	25487	CD	Latin		Argentina	2010	1418114	Lolein Music
6	12295807	Accepted	Kaleidoscopia	3130927	Beatnik (8)	Eargasm Wreckords	500315	CD	Rock	Krautrock	Switzerland	2008		Eargasm Wreckords
7	12295808	Accepted	Κόκκινη Γραμμή	2444813	Βασίλης Τερλέγκας	Alpha Records (6)	153245	CD	Folk, World, & Country		Greece			
8	12295809	Accepted	Frescobaldi Edition Vol 6 - Il Primo Libro Dei Madrigali a cinque voci	834330	Girolamo Frescobaldi	Brilliant Classics		89052	CD	Classical	Renaissance	Netherlands	2009	
9	12295810	Accepted	Try A Little Kindness / One Step At A Time	3432907	Don Lauren	Apt Records	125142	Vinyl	Rock	Vocal	US		1237274	
10	12295811	Accepted	A Tribute to La Monte Young	6522519	Infinite Music	Fire Records	2051	CD	Electronic	Minimal	UK	2018		

10 rows selected (0.339 seconds)

External Table **plzwork** will be the table used to make further queries for data analysis.

Step 2: Queries and Exporting

This step is to create the queries that we use to analyze and filter our data to be turned into visualizations and to export our analysis to be used for visualizations.

To create tables that will contain your query and be used to export...

```
CREATE TABLE example_table
ROW FORMAT DELIMITED FIELDS TERMINATED BY ‘/t’
STORED AS TEXTFILE LOCATION ‘/user/mmiran64/test/’
```

AS....

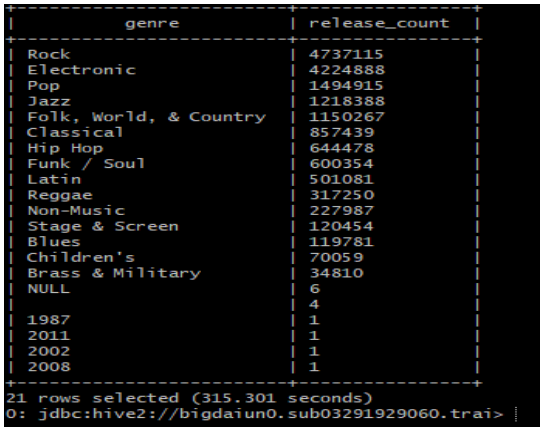
Highlighted above are the parameters specific to the user, so example_table will be the name of the table in your specific database in Beeline. Replace mmiran64 with your own username and test is the directory on HDFS where the output file will be stored.

Below are the queries used to analyze our dataset. The above CREATE TABLE template is used first and then after the AS, you will write the queries.

1. To show Total Release Count by Genre

```
SELECT genre, COUNT(release_id) AS release_count FROM plzwork GROUP

BY genre ORDER BY release_count DESC;
```



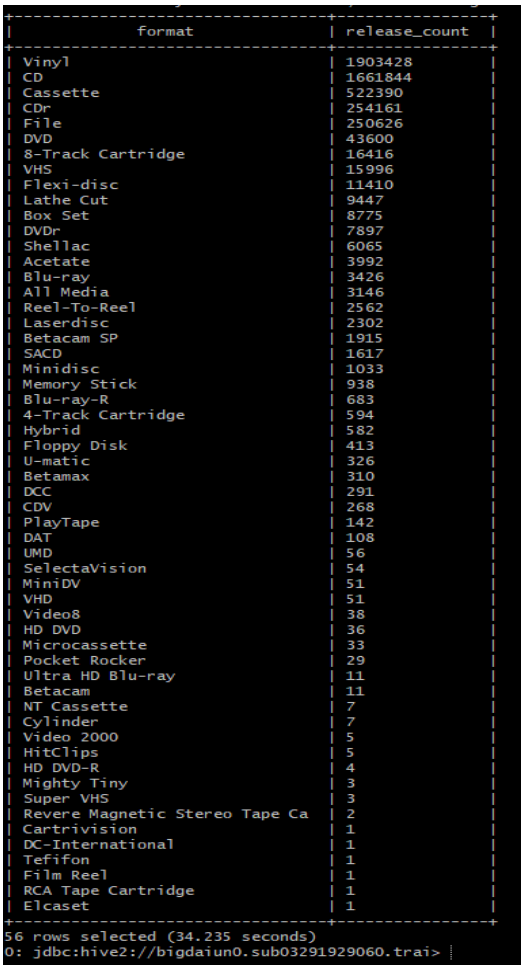
genre	release_count
Rock	4737115
Electronic	4224888
Pop	1494915
Jazz	1218388
Folk, World, & Country	1150267
Classical	857439
Hip Hop	644478
Funk / Soul	600354
Latin	501081
Reggae	317250
Non-Music	227987
Stage & Screen	120454
Blues	119781
Children's	70059
Brass & Military	34810
NULL	6
	4
1987	1
2011	1
2002	1
2008	1

21 rows selected (315.301 seconds)
0: jdbc:hive2://bigdaiun0.sub03291929060.trai>

2. To show Music Format by Rock Genre

```
SELECT format, COUNT(release_id) AS release_count FROM plzwork

WHERE genre= 'Rock' GROUP BY format ORDER BY release_count DESC;
```

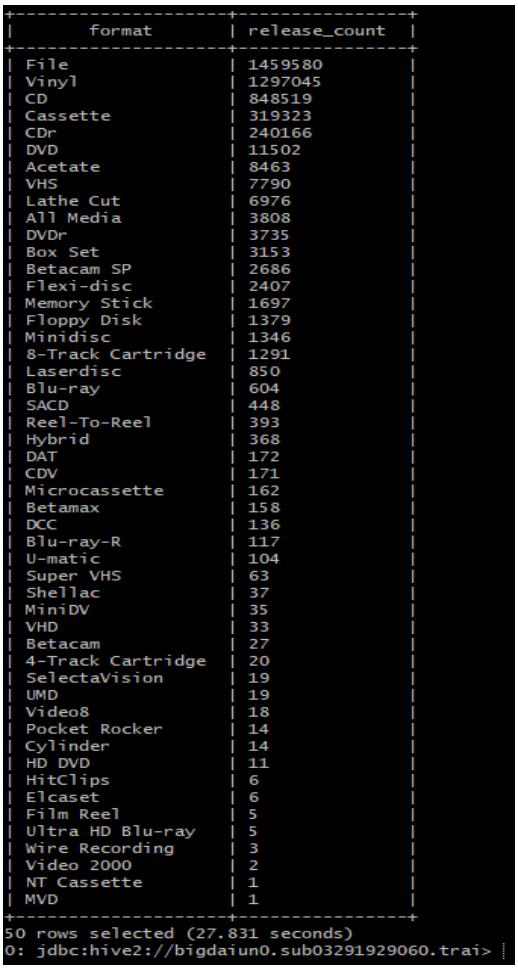


format	release_count
Vinyl	1903428
CD	1661844
Cassette	522390
CDr	254161
File	250626
DVD	43600
8-Track Cartridge	16416
VHS	15996
Flexi-disc	11410
Lathe Cut	9447
Box Set	8775
DVDr	7897
Shellac	6065
Acetate	3992
Blu-ray	3426
All Media	3146
Reel-To-Reel	2562
Laserdisc	2302
Betacam SP	1915
SACD	1617
Minidisc	1033
Memory Stick	938
Blu-ray-R	683
4-Track Cartridge	594
Hybrid	582
Floppy Disk	413
U-matic	326
Betamax	310
DCC	291
CDV	268
PlayTape	142
DAT	108
UMD	56
SelectaVision	54
MiniDV	51
VHD	51
Video8	38
HD DVD	36
Microcassette	33
Pocket Rocker	29
Ultra HD Blu-ray	11
Betacam	11
NT Cassette	7
Cylinder	7
Video 2000	5
HitClips	5
HD DVD-R	4
Mighty Tiny	3
Super VHS	3
Revere Magnetic Stereo Tape Ca	2
Cartrivision	1
DC-International	1
Tefifon	1
Film Reel	1
RCA Tape Cartridge	1
Elcaset	1

56 rows selected (34.235 seconds)
0: jdbc:hive2://bigdaiun0.sub03291929060.trai>

3. To show Music Format by Electronic Genre

```
SELECT format, COUNT(release_id) AS release_count FROM plzwork
WHERE genre= ‘Electronic’ GROUP BY format ORDER BY release_count
DESC;
```

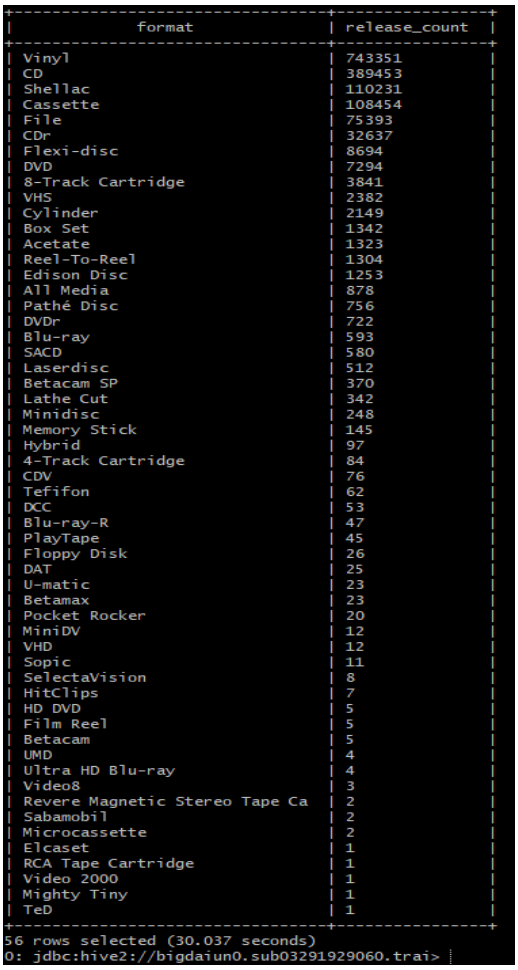


format	release_count
File	1459580
Vinyl	1297045
CD	848519
Cassette	319323
CDr	240166
DVD	11502
Acetate	8463
VHS	7790
Lathe Cut	6976
All Media	3808
DVDr	3735
Box Set	3153
Betacam SP	2686
Flexi-disc	2407
Memory Stick	1697
Floppy Disk	1379
Minidisc	1346
8-Track Cartridge	1291
Laserdisc	850
Blu-ray	604
SACD	448
Reel-To-Reel	393
Hybrid	368
DAT	172
CDV	171
Microcassette	162
Betamax	158
DCC	136
Blu-ray-R	117
U-matic	104
Super VHS	63
Shellac	37
MiniDV	35
VHD	33
Betacam	27
4-Track Cartridge	20
SelectaVision	19
UMD	19
Video8	18
Pocket Rocker	14
Cylinder	14
HD DVD	11
HitClips	6
Elcaset	6
Film Reel	5
Ultra HD Blu-ray	5
Wire Recording	3
Video 2000	2
NT Cassette	1
MVD	1

50 rows selected (27.831 seconds)
0: jdbc:hive2://bigdaiun0.sub03291929060.trai>

4. To show Music Format by Pop Genre

```
SELECT format, COUNT(release_id) AS release_count FROM plzwork
WHERE genre= 'Pop' GROUP BY format ORDER BY release_count DESC;
```

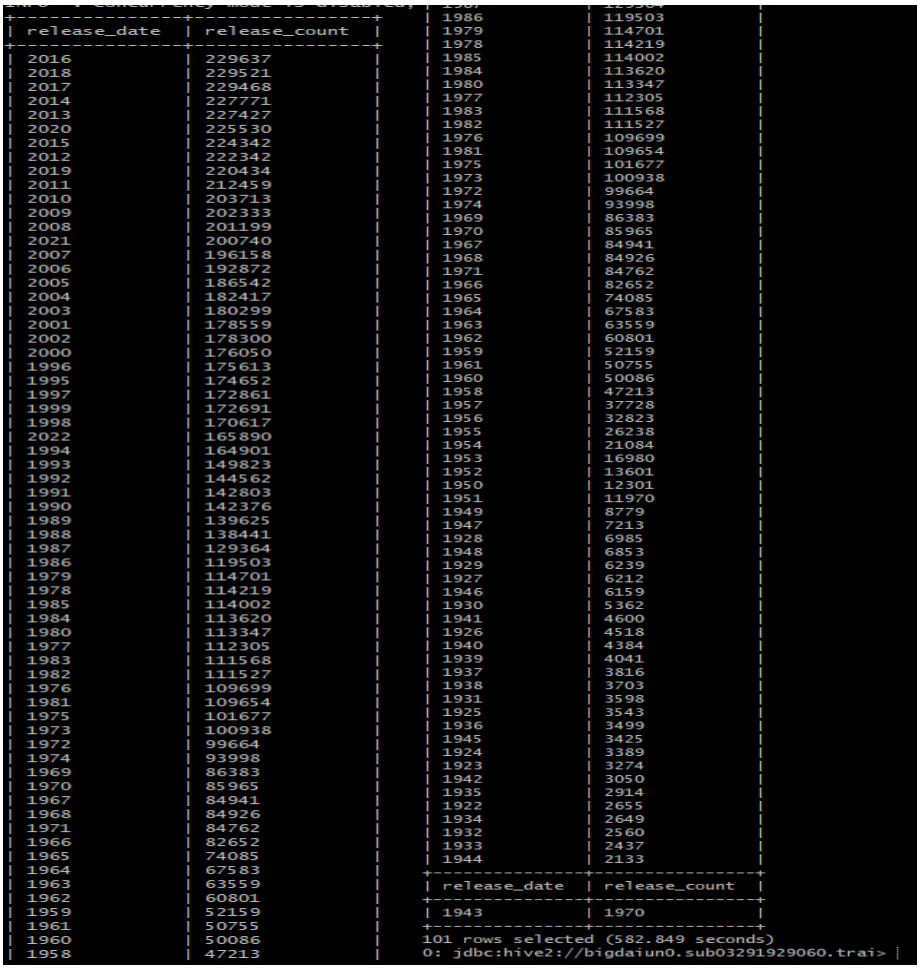


format	release_count
Vinyl	743351
CD	389453
Shellac	110231
Cassette	108454
File	75393
CDr	32637
Flexi-disc	8694
DVD	7294
8-Track Cartridge	3841
VHS	2382
Cylinder	2149
Box Set	1342
Acetate	1323
Reel-To-Reel	1304
Edison Disc	1253
All Media	878
Pathé Disc	756
DVDr	722
Blu-ray	593
SACD	580
Laserdisc	512
Betacam SP	370
Lathe Cut	342
Minidisc	248
Memory Stick	145
Hybrid	97
4-Track Cartridge	84
CDV	76
Tefifon	62
DCC	53
Blu-ray-R	47
PlayTape	45
Floppy Disk	26
DAT	25
U-matic	23
Betamax	23
Pocket Rocker	20
MiniDV	12
VHD	12
Sopic	11
SelectaVision	8
HitClips	7
HD DVD	5
Film Reel	5
Betacam	5
UMD	4
Ultra HD Blu-ray	4
Video8	3
Revere Magnetic Stereo Tape Ca	2
Sabamobil	2
Microcassette	2
Elcaset	1
RCA Tape Cartridge	1
Video 2000	1
Mighty Tiny	1
TeD	1

56 rows selected (30.037 seconds)
0: jdbc:hive2://bigdaiun0.sub03291929060.trai>

5. To show Release Count by Years (1922-2022)

```
SELECT release_date, COUNT(release_id) AS release_count
FROM plzwork WHERE release_date >= 1922 AND release_date <= 2022
AND country IN ('US', 'UK', 'Germany', 'Japan', 'France', 'Italy',
'Canada', 'Netherlands', 'Spain', 'Australia') GROUP BY
release_date ORDER BY release_count DESC;
```



release_date	release_count
2016	229637
2018	229521
2017	229468
2014	227771
2013	227427
2020	225530
2015	224342
2012	222342
2019	220434
2011	212459
2010	203713
2009	202333
2008	201199
2021	200740
2007	196158
2006	192872
2005	186542
2004	182417
2003	180299
2001	178559
2002	178300
2000	176050
1996	175613
1995	174652
1997	172861
1999	172691
1998	170617
2022	165890
1994	164901
1993	149823
1992	144562
1991	142803
1990	142376
1989	139625
1988	138441
1987	129364
1986	119503
1979	114701
1978	114219
1985	114002
1984	113620
1980	113347
1977	112305
1983	111568
1982	111527
1976	109699
1981	109654
1975	101677
1973	100938
1972	99664
1974	93998
1969	86383
1970	85965
1967	84941
1968	84926
1971	84762
1966	82652
1965	74085
1964	67583
1963	63559
1962	60801
1959	52159
1961	50755
1960	50086
1958	47213

1986	119503
1979	114701
1978	114219
1985	114002
1984	113620
1980	113347
1977	112305
1983	111568
1982	111527
1976	109699
1981	109654
1975	101677
1973	100938
1972	99664
1974	93998
1969	86383
1970	85965
1967	84941
1968	84926
1971	84762
1966	82652
1965	74085
1964	67583
1963	63559
1962	60801
1959	52159
1961	50755
1960	50086
1958	47213
1957	37728
1956	32823
1955	26238
1954	21084
1953	16980
1952	13601
1950	12301
1951	11970
1949	8779
1947	7213
1928	6985
1948	6853
1929	6239
1927	6212
1946	6159
1930	5362
1941	4600
1926	4518
1940	4384
1939	4041
1937	3816
1938	3703
1931	3598
1925	3543
1936	3499
1945	3425
1924	3389
1923	3274
1942	3050
1935	2914
1922	2655
1934	2649
1932	2560
1933	2437
1944	2133

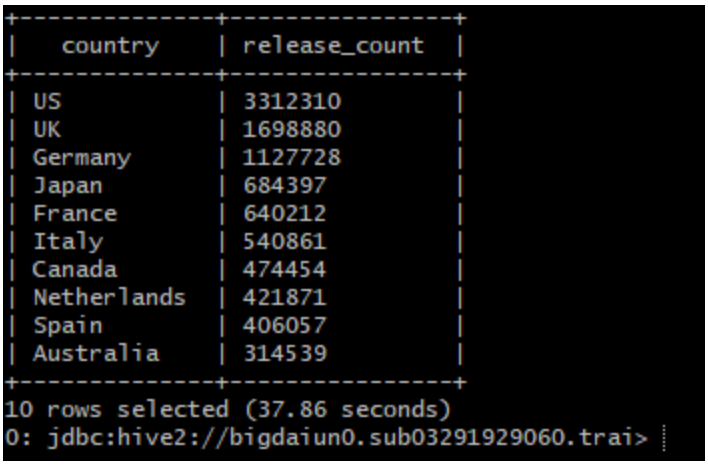
release_date	release_count
1943	1970

101 rows selected (582.849 seconds)

0: jdbc:hive2://bigdaiun0.sub03291929060.trai>

6. To show Release Count by Country (1922-2022)

```
SELECT country, COUNT(release_id) AS release_count
FROM plzwork WHERE release_date >= 1922 AND release_date <= 2022
AND country IN ('US', 'UK', 'Germany', 'Japan', 'France', 'Italy',
'Canada', 'Netherlands', 'Spain', 'Australia') GROUP BY country
ORDER BY release_count DESC;
```



country	release_count
US	3312310
UK	1698880
Germany	1127728
Japan	684397
France	640212
Italy	540861
Canada	474454
Netherlands	421871
Spain	406057
Australia	314539

10 rows selected (37.86 seconds)
0: jdbc:hive2://bigdaiun0.sub03291929060.trai>

After you are done with creating the tables for your queries, refer to the textfile location and in that directory, it should contain a 000000_0.txt output file. This is where your queries are located which will be used for visualizations.

Exit Beeline (CTRL+Z), and run the command to make sure output file is there...

```
Hdfs dfs –ls test
```

To transfer output file from HDFS to your Linux, run this command...

```
Hdfs dfs –get test/000000_0
```

Once the output file is in your Linux, you will have to run this command to download the output file from Linux to your local file system....

```
Scp mmiran64@129.146.90.117:/home/mmiran64/000000_0 test.csv
```

The output file or “test.csv” should be in your own computer and is in the Users directory.

7. For further analysis of the dataset, you can find 10 country’s yearly top 5 genres from 1922-2022. First create an empty external table, which will later be populated by a SELECT statement. Enter beeline, use group5, and enter the following command:

```
CREATE EXTERNAL TABLE IF NOT EXISTS
country_yearly_genre_analysis( year DATE, country STRING, genre
STRING, genre_count BIGINT)

ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'

STORED AS TEXTFILE LOCATION
'/user/gmunoz58/CountryYearlyGenreAnalysis/';
```

8. Next, we will populate the table. Overwrite **country_yearly_genre_analysis** table with the following query.

```
INSERT OVERWRITE TABLE country_yearly_genre_analysis

SELECT CONCAT(release_date, '-01-01') AS year,

       country,

       genre,

       genre_count

FROM ( SELECT release_date,

              country,

              genre,

              count(*) AS genre_count,

              RANK() OVER(PARTITION BY country, release_date

                          ORDER BY count(*) DESC) AS genre_ranks

       FROM plzwork

       WHERE release_date >= 1922

             AND release_date <= 2022

             AND country IN
('US','UK','Germany','Japan','France','Italy','Canada','Netherlands','
Spain','Australia')

       GROUP BY release_date, country, genre) AS ranked_genres_yearly

WHERE genre_ranks <= 5

ORDER BY release_date, country, genre_ranks;
```

Use the following SELECT statement to display 10 entries from **country_yearly_genre_analysis** table.

```
SELECT * FROM country_yearly_genre_analysis LIMIT 10;
```

country_yearly_genre_analysis.year	country_yearly_genre_analysis.country	country_yearly_genre_analysis.genre	country_yearly_genre_analysis.genre_count
1922-01-01	Australia	Folk, World, & Country	2
1922-01-01	Canada	Pop	41
1922-01-01	Canada	Jazz	36
1922-01-01	Canada	Classical	12
1922-01-01	Canada	Folk, World, & Country	10
1922-01-01	Canada	Brass & Military	2
1922-01-01	France	Jazz	11
1922-01-01	France	Classical	6
1922-01-01	France	Non-Music	1
1922-01-01	France	Pop	1

9. Exit beeline, then move the table as a **visAnalysis** file into Linux server with the following command:

```
hdfs dfs -get /user/gmunoz58/CountryYearlyGenreAnalysis/000000_0
visAnalysis
```

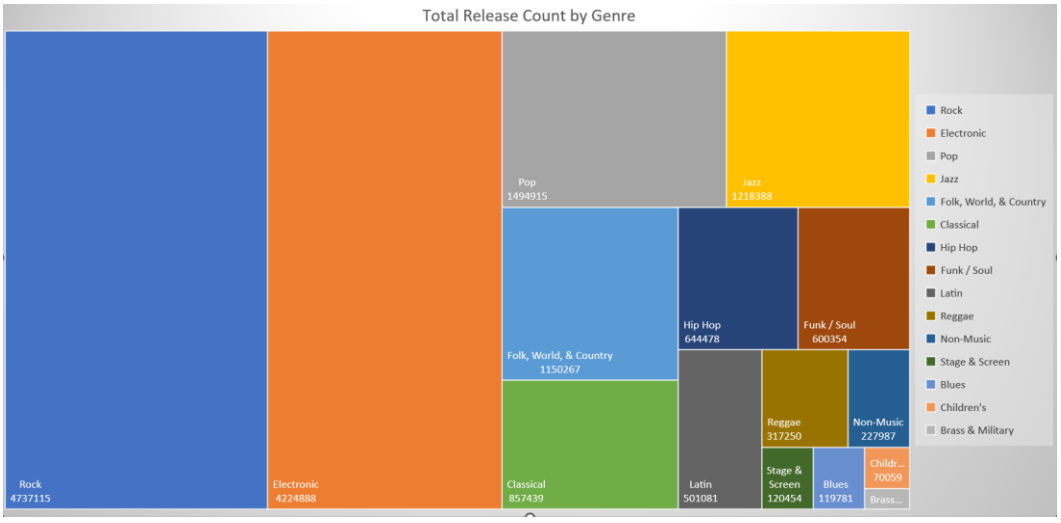
10. Exit Linux server, then secure copy **visAnalysis** file onto local machine.

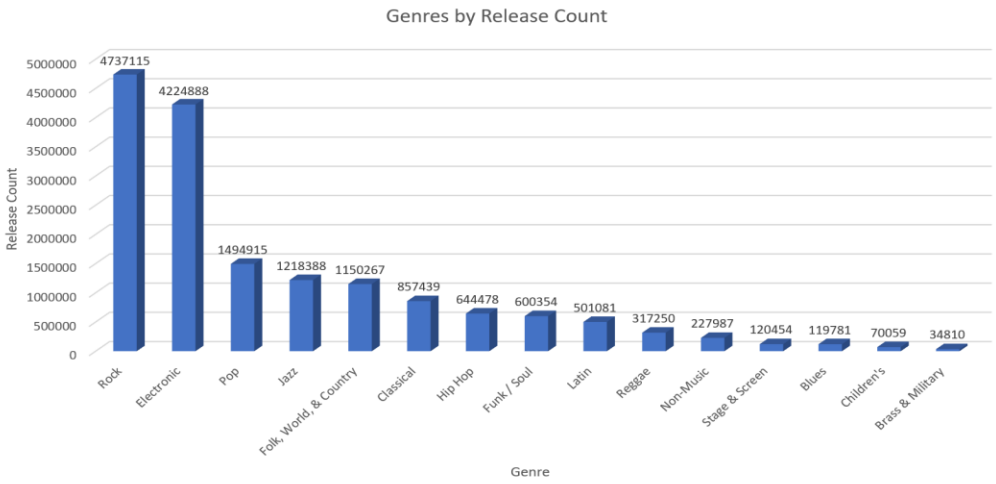
```
scp gmunoz58@129.146.90.117:/home/gmunoz58/visAnalysis .
```

Step 3: Visualization

This step is to use your output files/analysis into visualizations. Make sure to import the data in your CSVs into your worksheets. The visualizations match according to the order of the queries in Step 2.

1. Total Release Count by Genre

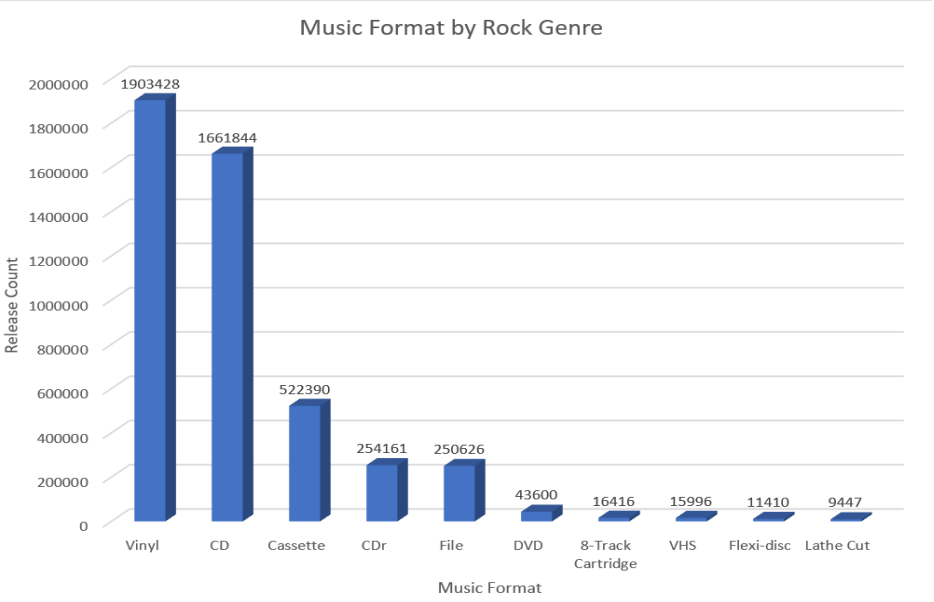


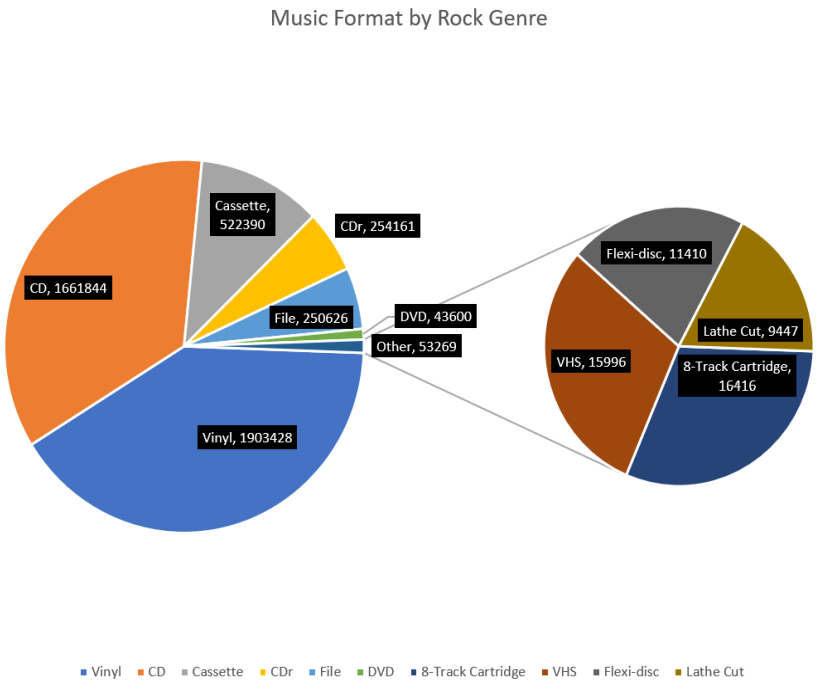


To create the treemap, go to the Insert tab of Excel, look in the Charts category and look for the dropdown icon that says Tree map. Click on your Tree map and look a plus icon, it shows the chart elements. Check the boxes for chart titles, data labels and legends.

To create the bar chart, follow the first few steps of the previous instruction but instead of looking for the tree map icon, search for the bar chart icon, a drop down will be presented for you and make sure to choose the 3-D clustered column option. Choose all chart elements except data label and legend.

2. Music Format by Rock Genre





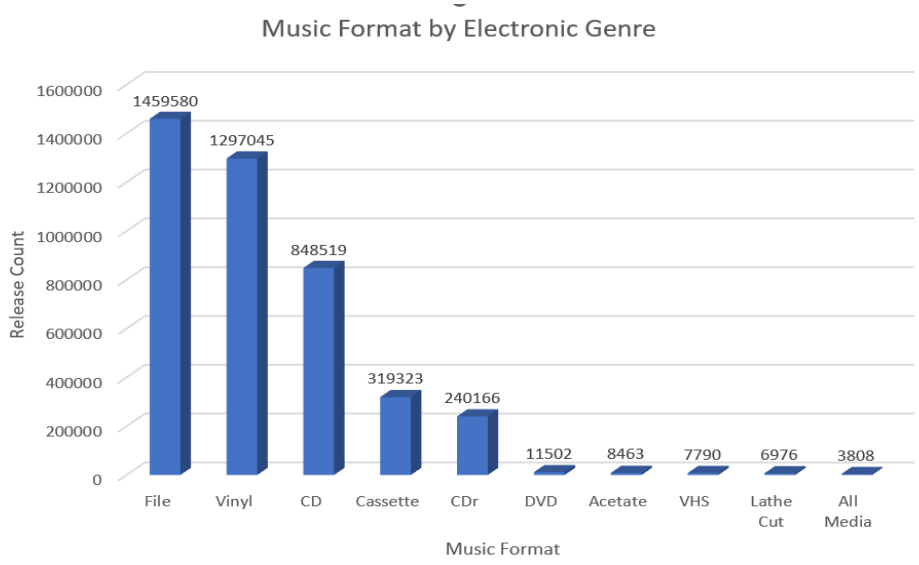
Similar to previous instructions, to create the bar chart, Insert tab > Look for the charts category>

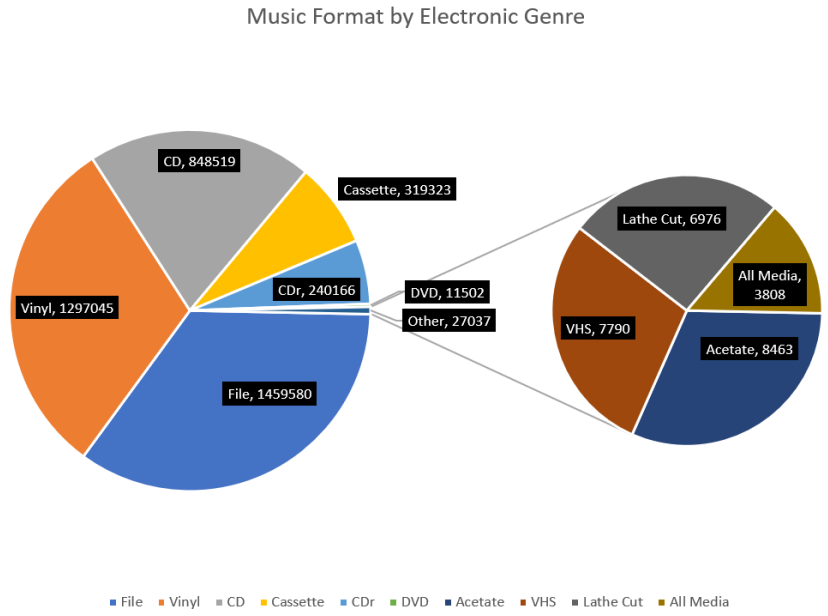
Choose the bar chart option > Choose every chart element except data tables and legends

To create the pie chart, Insert tab > Look for the charts category > Pie Chart Icon > 2-D Pie > Pie of Pie

Option > All chart elements

3. Music Format by Electronic Genre





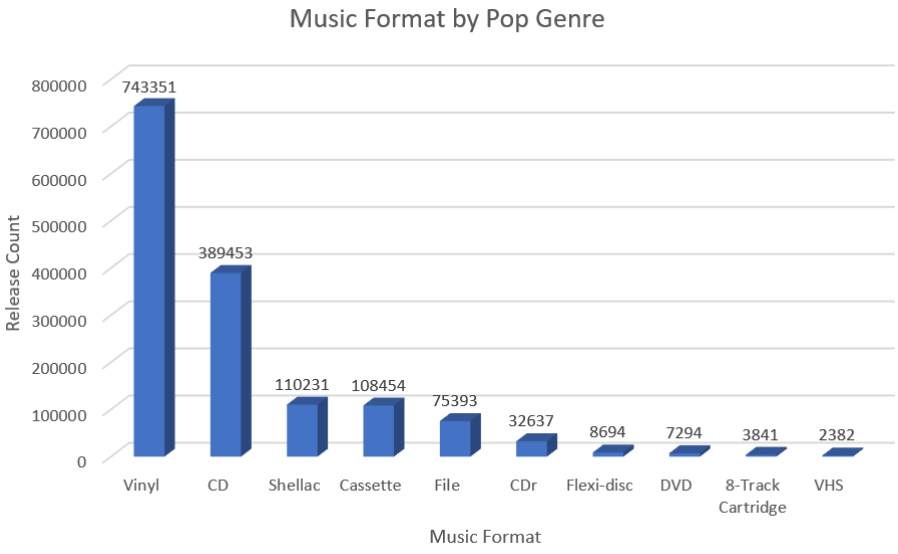
Similar to previous instructions, to create the bar chart, Insert tab > Look for the charts category>

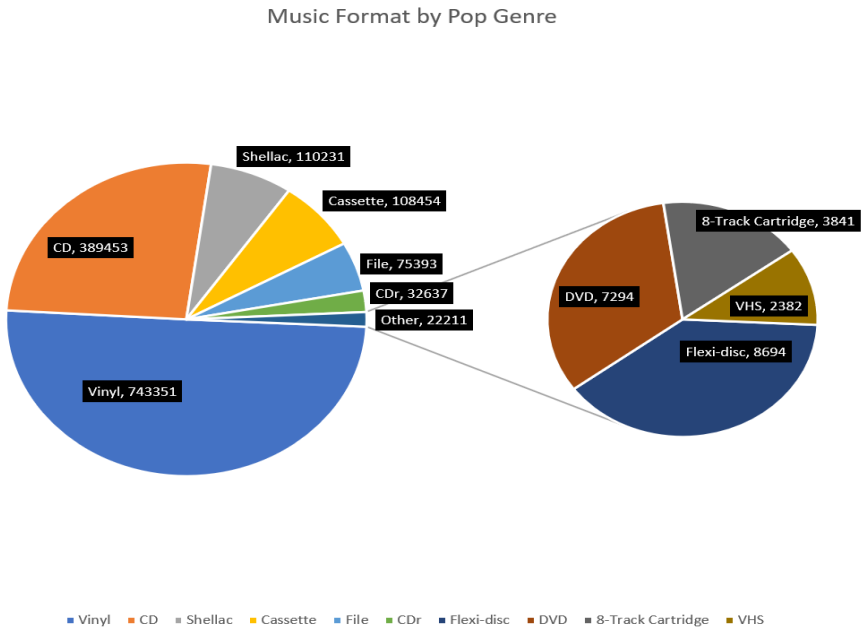
Choose the bar chart option > Choose every chart element except data tables and legends

To create the pie chart, Insert tab > Look for the charts category > Pie Chart Icon > 2-D Pie > Pie of Pie

Option > All chart elements

4. Music Format by Pop Genre





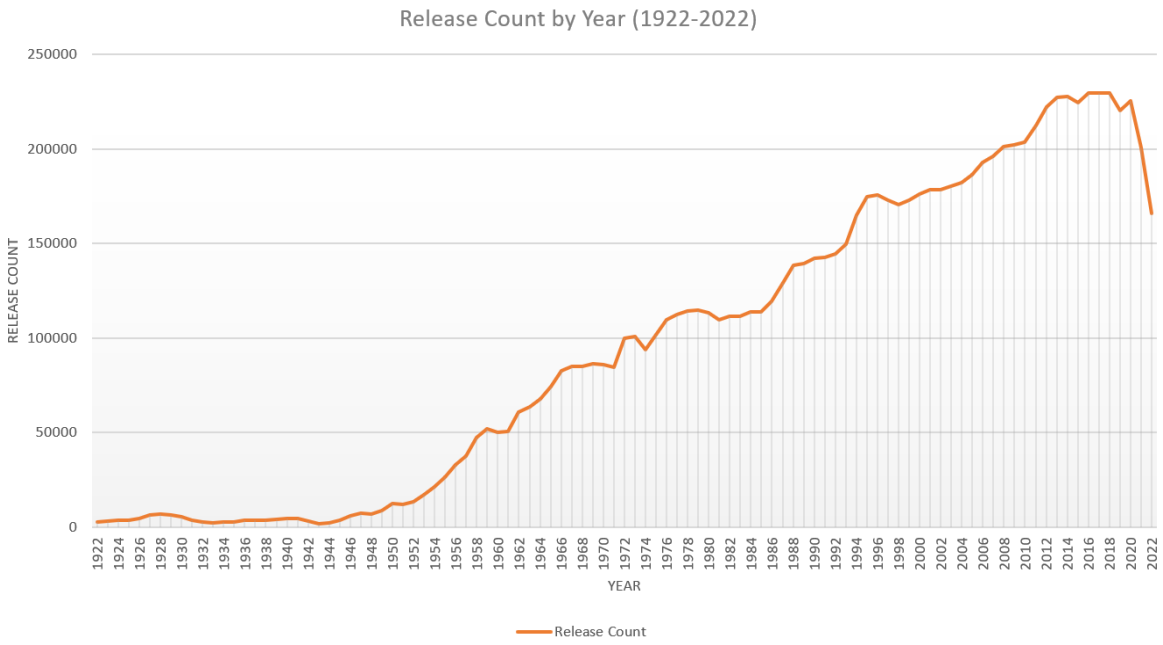
Similar to previous instructions, to create the bar chart, Insert tab > Look for the charts category>

Choose the bar chart option > Choose every chart element except data tables and legends

To create the pie chart, Insert tab > Look for the charts category > Pie Chart Icon > 2-D Pie > Pie of Pie

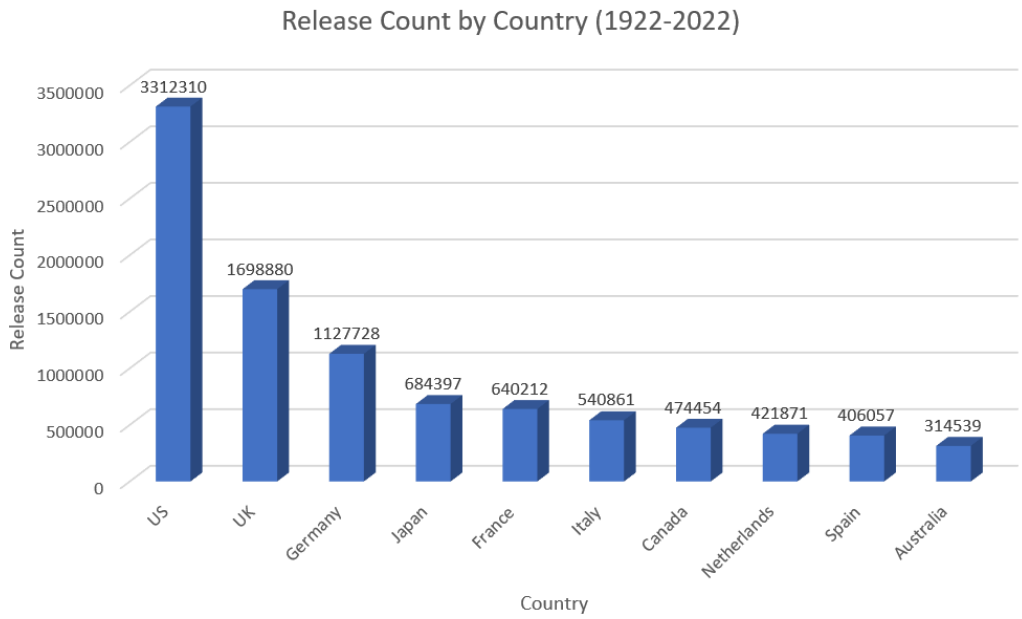
Option > All chart elements

5. Release Count by Year



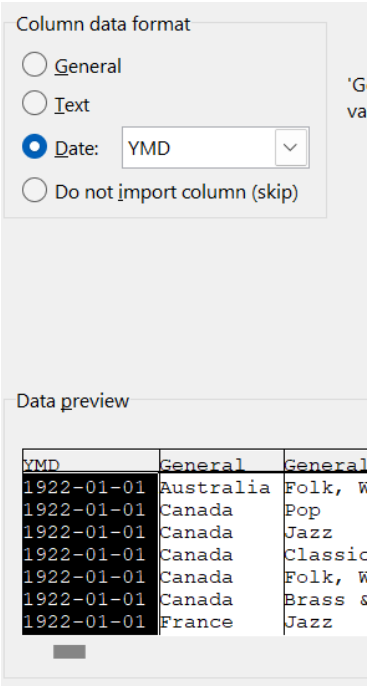
To create this line graph, Insert tab > Look for the charts category > Line Graph Icon > 2-D Line > Line Option > Choose chart elements Axes, Axis titles, Chart title, gridline, and legend > Make sure the years are in the X-axis and release count is in the Y-axis

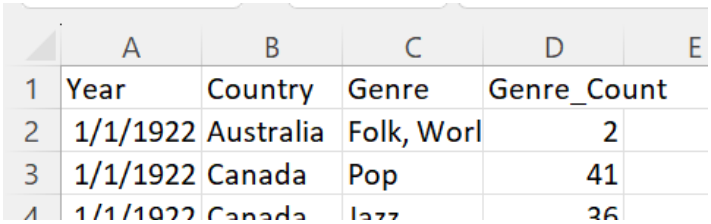
6. Release Count by Country



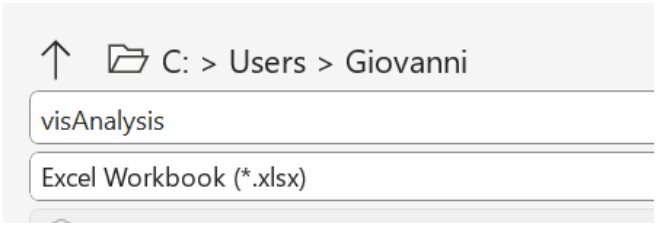
To create this bar chart, Insert tab > Chart Category > Bar Chart Icon > 3-D Column > 3-D Clustered Column > All chart elements except data tables and legends

7. For **visAnalysis** file: Open Excel, load visAnalysis file using pipe as delimiter, make sure to format year column as **Date: YMD**, and insert headers as **Year**, **Country**, **Genre**, and **Genre_Count**, then save as **.xlsx** file.



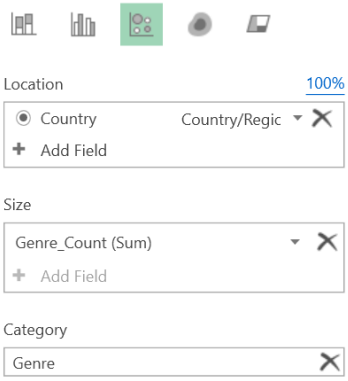


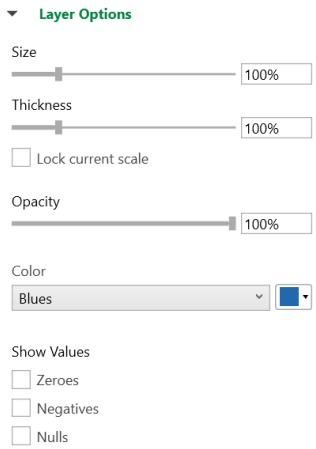
	A	B	C	D	E
1	Year	Country	Genre	Genre_Count	
2	1/1/1922	Australia	Folk, Worl	2	
3	1/1/1922	Canada	Pop	41	
4	1/1/1922	Canada	Jazz	36	

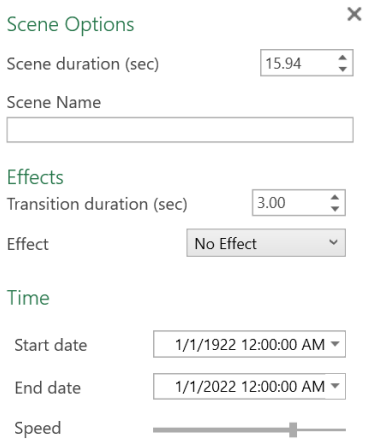


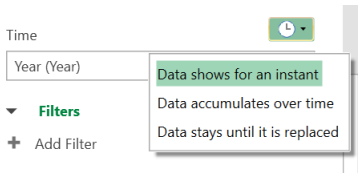
8. **Insert** > **3D Map**

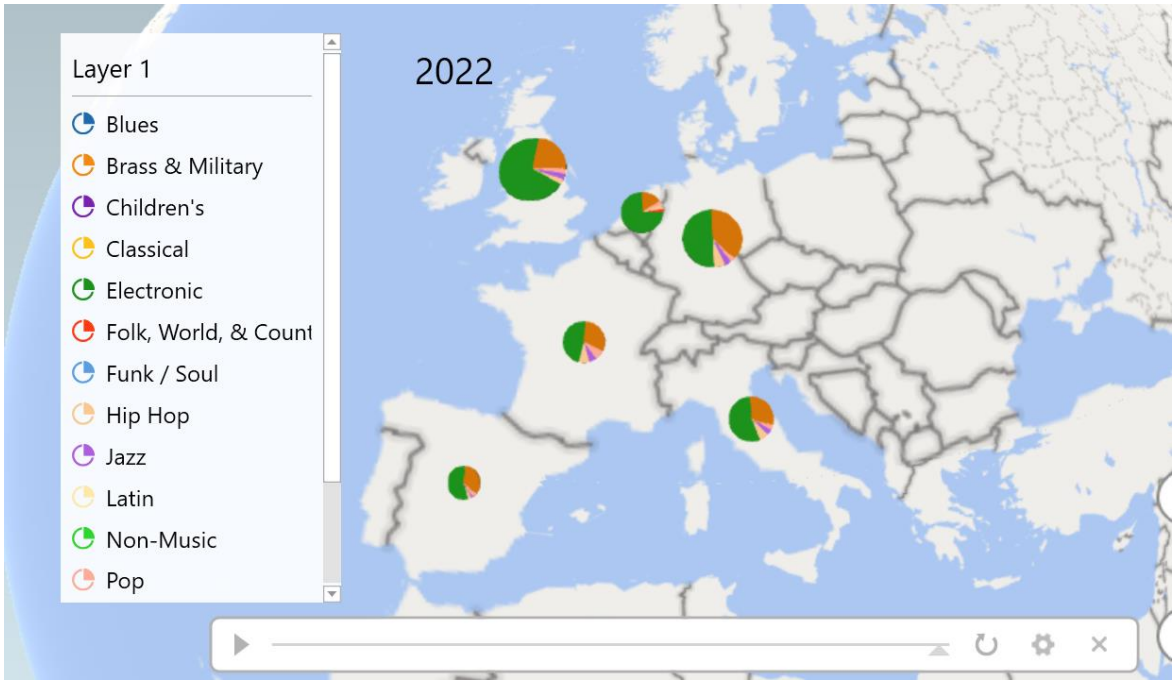
9. Change visualization to **Bubble**. Set Location to **Country (Country/Region),** Size to **Genre_Count(Sum)**, Category to **Genre**, Time to **Year(Year)** and click on Clock Icon next to Time and change to **Data shows for an instant**. Select dropdown Layer Options: under *Show Values* uncheck **Zeroes** and **Negatives**. Adjust Scene Options as needed.











References

1. URL of Data Source: https://www.kaggle.com/datasets/ofurkancoban/discogs-releases-dataset
2. URL of your Github: https://github.com/danielgarrido1/Term-Project-Abstract-Discogs
3. URL of CSViewer: https://csviewer.com/
4. URLs for Linux commands (SED & GREP:)
    - https://www.howtogeek.com/666395/how-to-use-the-sed-command-on-linux/

- https://www.geeksforgeeks.org/grep-command-in-unixlinux/