Homework 2: Association Rule Learning

GROUP 7

GUERRA, HINOLAN, LASALA, LORENZO, ROCO

Groceries Dataset

Objective

- Calculate the association rules and find the significant/interesting items in this dataset.
- What would you recommend to the owner of a grocery store given these association rules?
- Is there any other grouping that could give us high confidence/interest?

Initial Preprocessing

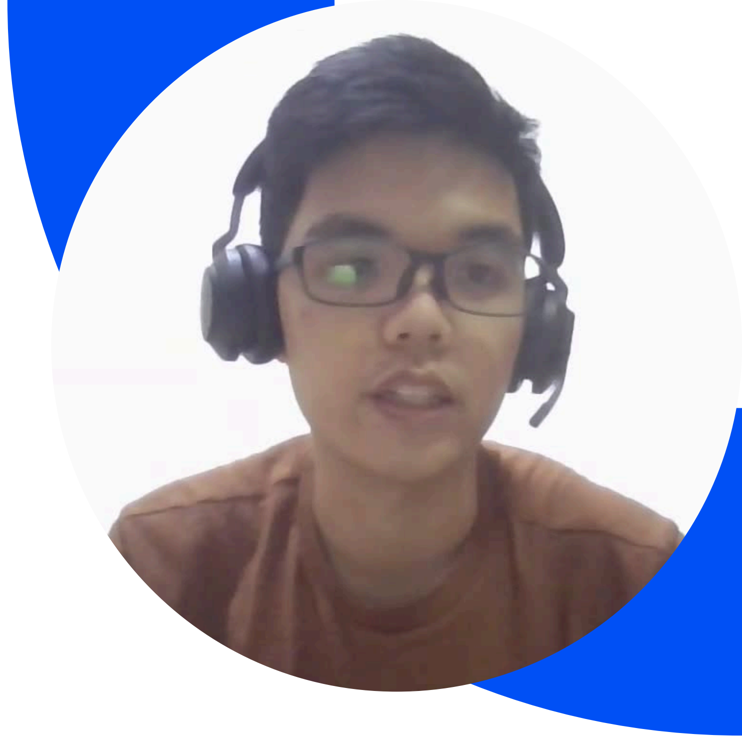
- Data type checking
- Date to datetime conversion
- Convert dataset to list data type
- Sorting items by **Member_number** and **Month**

Frequent Itemset Mining

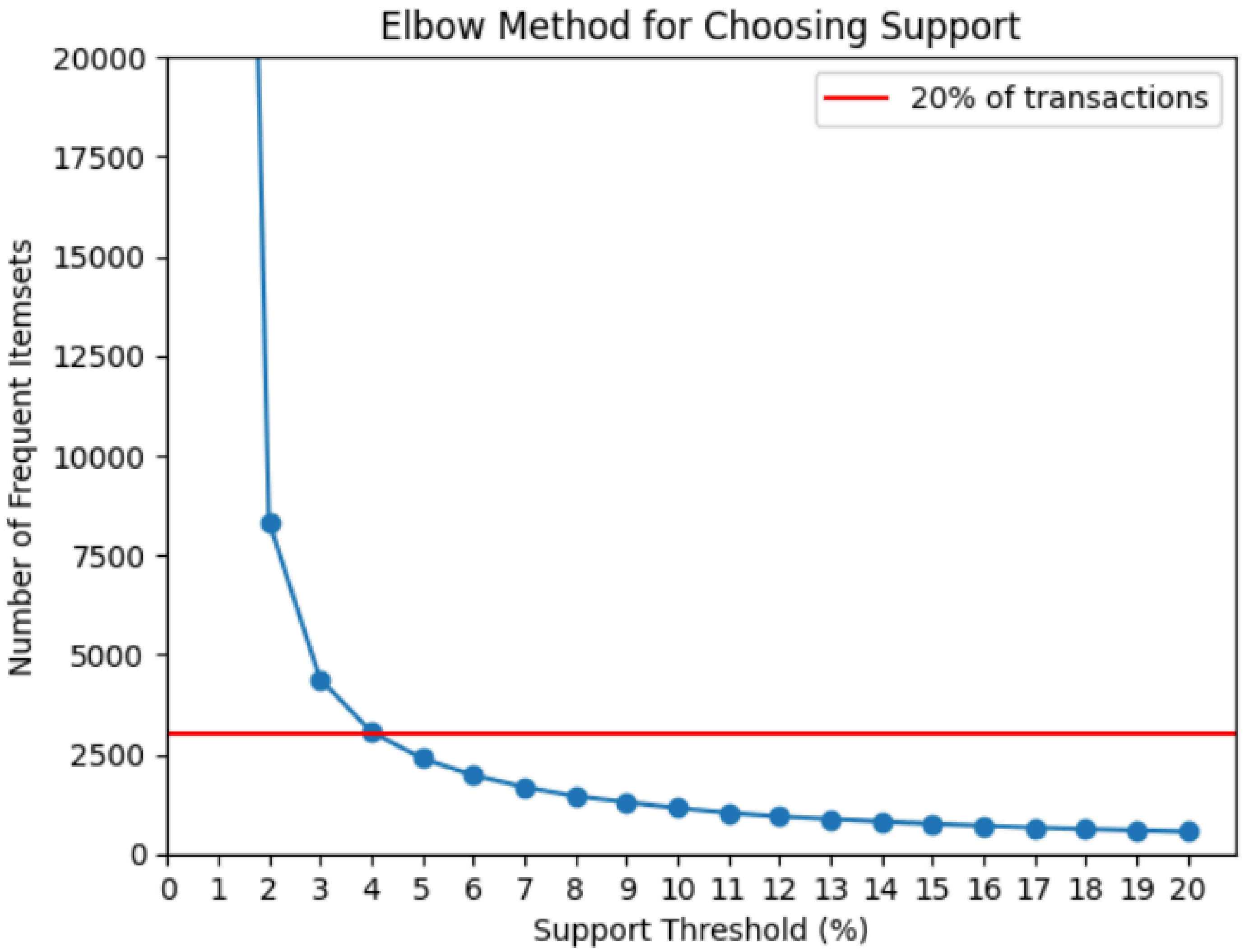
- Frequent Pattern Growth

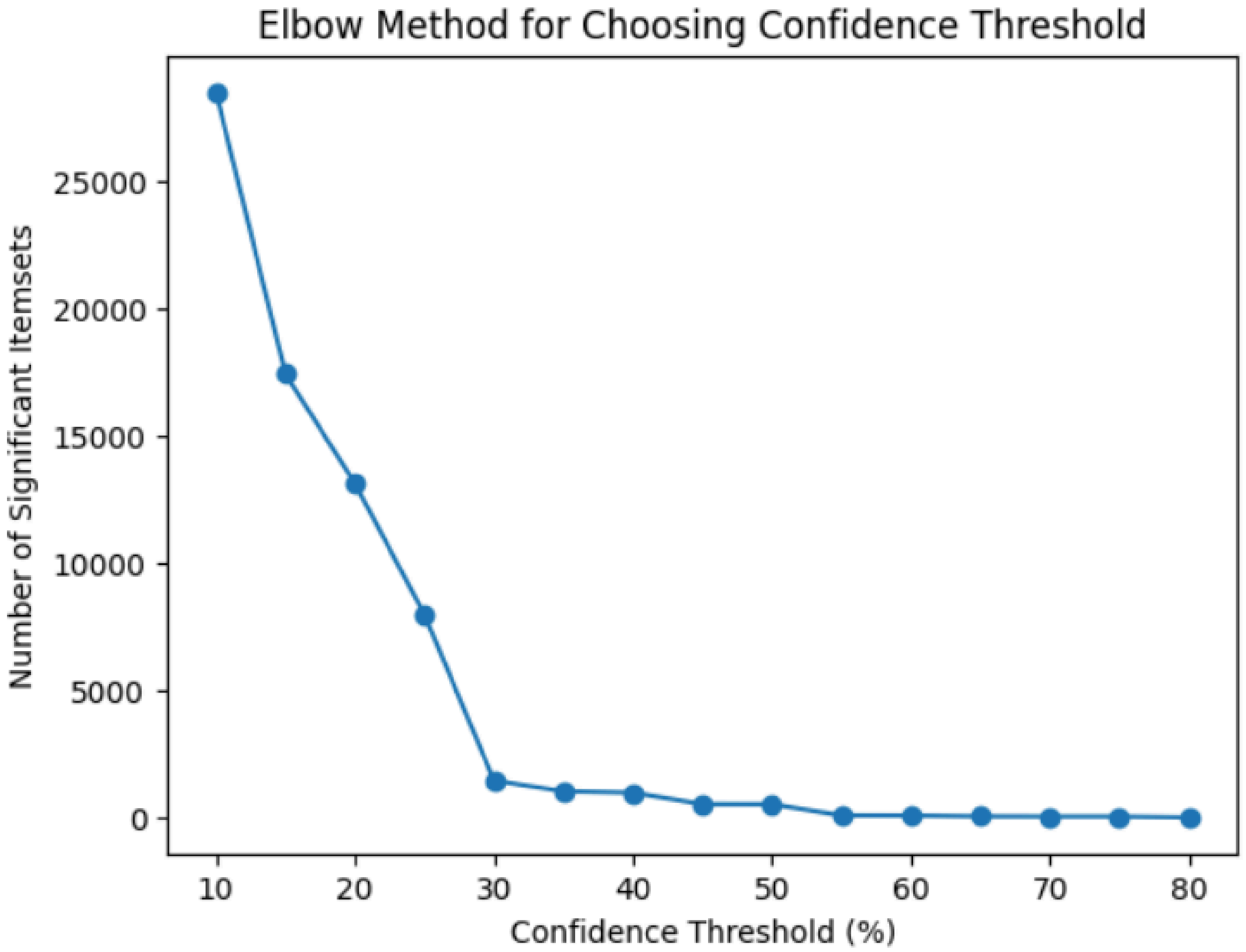
```python
# calculates the frequent itemset using fpgrowth algorithm
def get_fim(supp, transactions):
    result = fpgrowth(transactions, supp=supp, report='as')
    colnames = ['itemset'] + ['support_absolute', 'support_relative']
    df_result = pd.DataFrame(result, columns=colnames)
    df_result = df_result.sort_values('support_absolute', ascending=False)
    return df_result
```

Ideal Support and Confidence Threshold



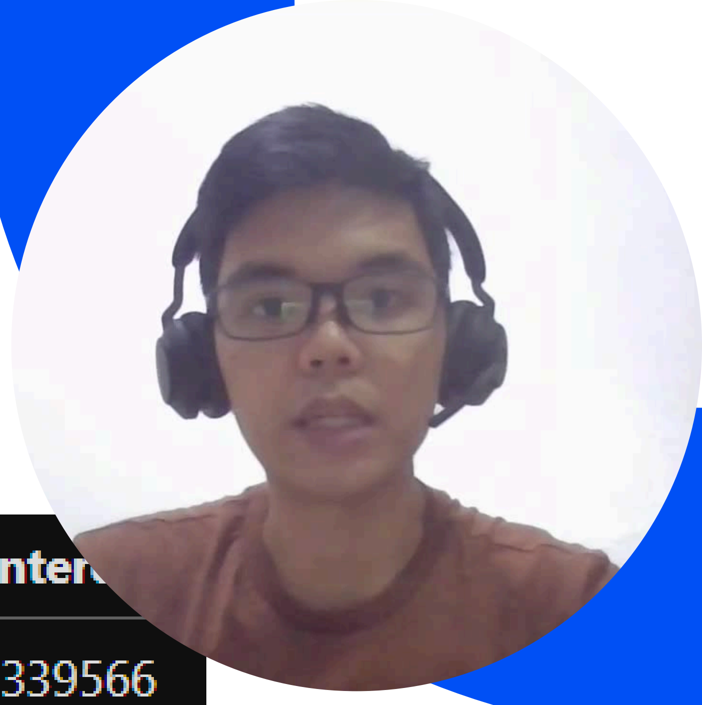
- Elbow method





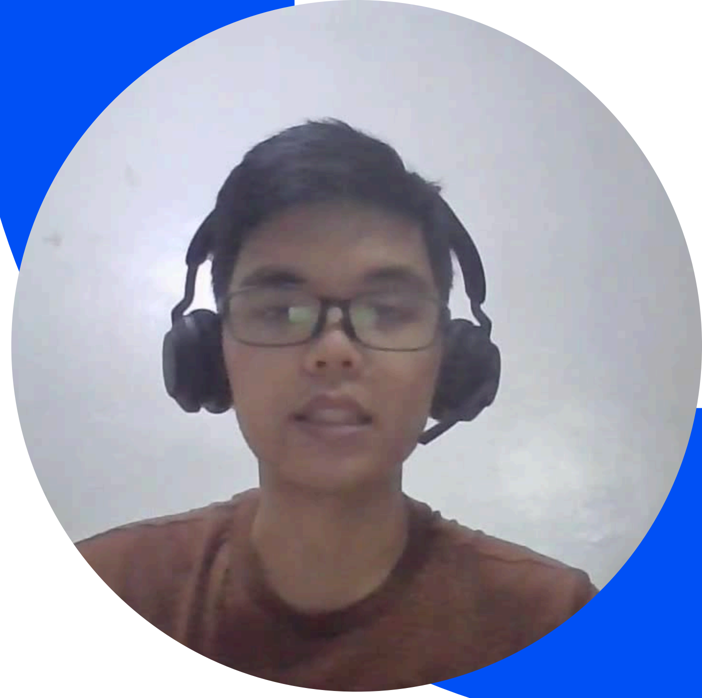
Interesting Itemsets

	consequent	antecedent	support_absolute	support_relative	confidence	inter
1418	whole milk	(brandy,)	38	0.002540	0.342105	0.339566
230	whole milk	(pork, sausage)	23	0.001537	0.391304	0.389767
348	whole milk	(beef, whipped/sour cream)	21	0.001403	0.333333	0.331930
249	other vegetables	(pork, citrus fruit)	20	0.001337	0.350000	0.348663
255	yogurt	(pork, citrus fruit)	20	0.001337	0.300000	0.298663
...	...	...	...	...	...	...
746	pastry	(specialty chocolate, curd, citrus fruit)	4	0.000267	0.500000	0.499733
766	other vegetables	(specialty chocolate, hamburger meat)	4	0.000267	0.500000	0.499733
804	bottled beer	(misc. beverages, waffles)	4	0.000267	0.500000	0.499733
692	other vegetables	(ham, chicken)	4	0.000267	0.500000	0.499733
1058	herbs	(hard cheese, sugar)	4	0.000267	0.500000	0.499733



Recommendations

Grouped by Date & Member

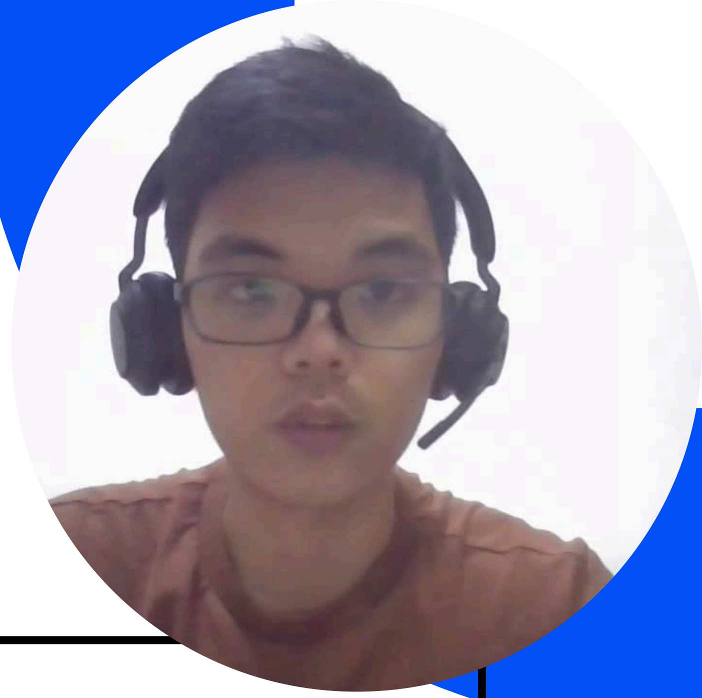


- The owner of the grocery store can use the association rules to determine which items are frequently bought together. In this case, we recommend that the store place the following items near each other to encourage customers to buy them together (top 5 item combinations):
  - brandy with whole milk
  - pork and sausage with whole milk
  - beef and whipped/sour cream with whole milk
  - pork and citrus fruit with other vegetables
  - pork and citrus fruit with yogurt
- Based on the top 5 recommendations, all the association rules are mostly attributed with items that are seen together inside the refrigerator.

Other Groupings

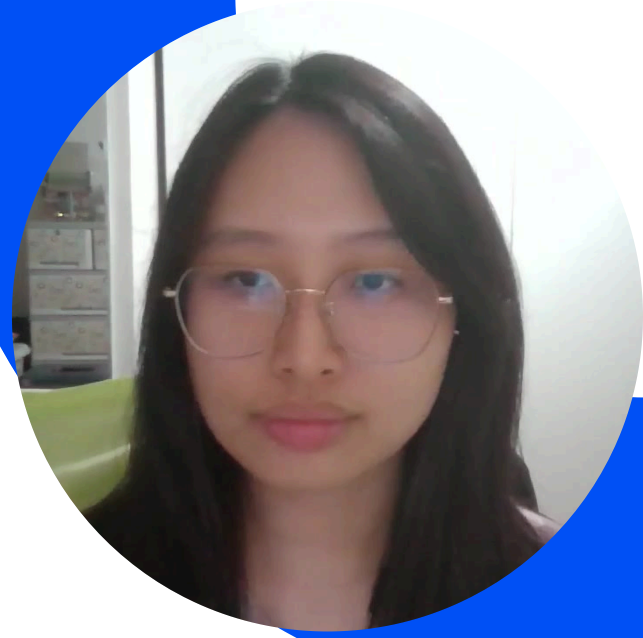
- **Grouped by Month and Member**
  - Support threshold is set to 7 and confidence threshold is set to 30
- **Grouped by Quarter and Member**
  - Support threshold is set to 15 and confidence threshold is set to 30
- **Grouped by Weekday and Member**
  - Support threshold is set to 9 and confidence threshold is set to 30

Recommendations for Other Groupings



Per Month	Per Quarter	Per Weekday
• sausage and yogurt with whole milk • pastry and soda with whole milk • coffee and rolls/buns with other vegetables • chocolate and other vegetables with whole milk • whipped/sour cream and citrus fruit with whole milk	• rolls/buns, and other vegetables with whole milk • yogurt, and other vegetables with whole milk • yogurt, and roll/sbuns with whole milk • tropical fruit, other vegetables with whole milk • yogurt, and soda with whole milk	• sausage, and yogurt with whole milk • pastry, and other vegetables with whole milk • canned beer, and soda with whole milk • pastry, and rolls/buns with whole milk • pastry, and yogurt with whole milk

Spotify Tracks Dataset



Description

- Contains Spotify tracks across 125 genres, along with their audio features and metadata
    - Track details: track name, popularity, duration, etc.
    - Musical attributes: danceability, energy, tempo, etc.

Objective

- Identify which combinations of audio features and genres are most associated with different levels of popularity

Initial Preprocessing

Initial

114,000 rows

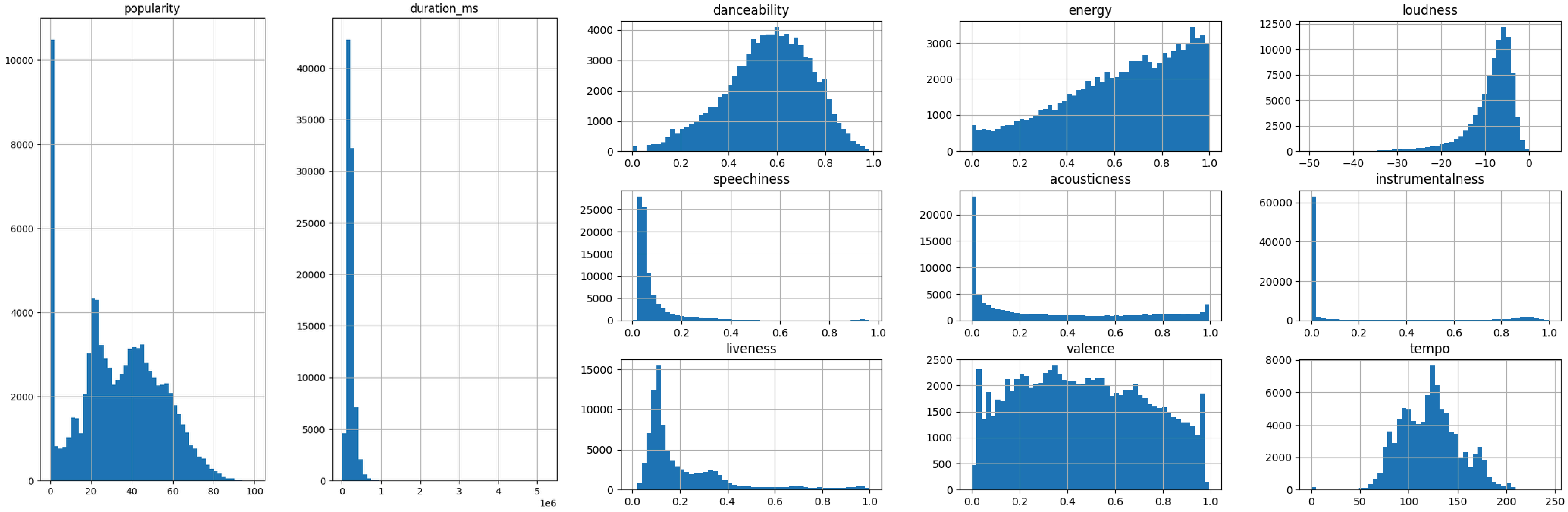
→

Preprocessed

89,740 rows

- Drop Null Values (1)
- Drop Duplicate Songs and Retain the Most Popular Version (24,359)

Preprocessing Music Features



Binning Criteria

3 Bins

Range < 3 * standard deviation

Low, Medium, High

5 Bins

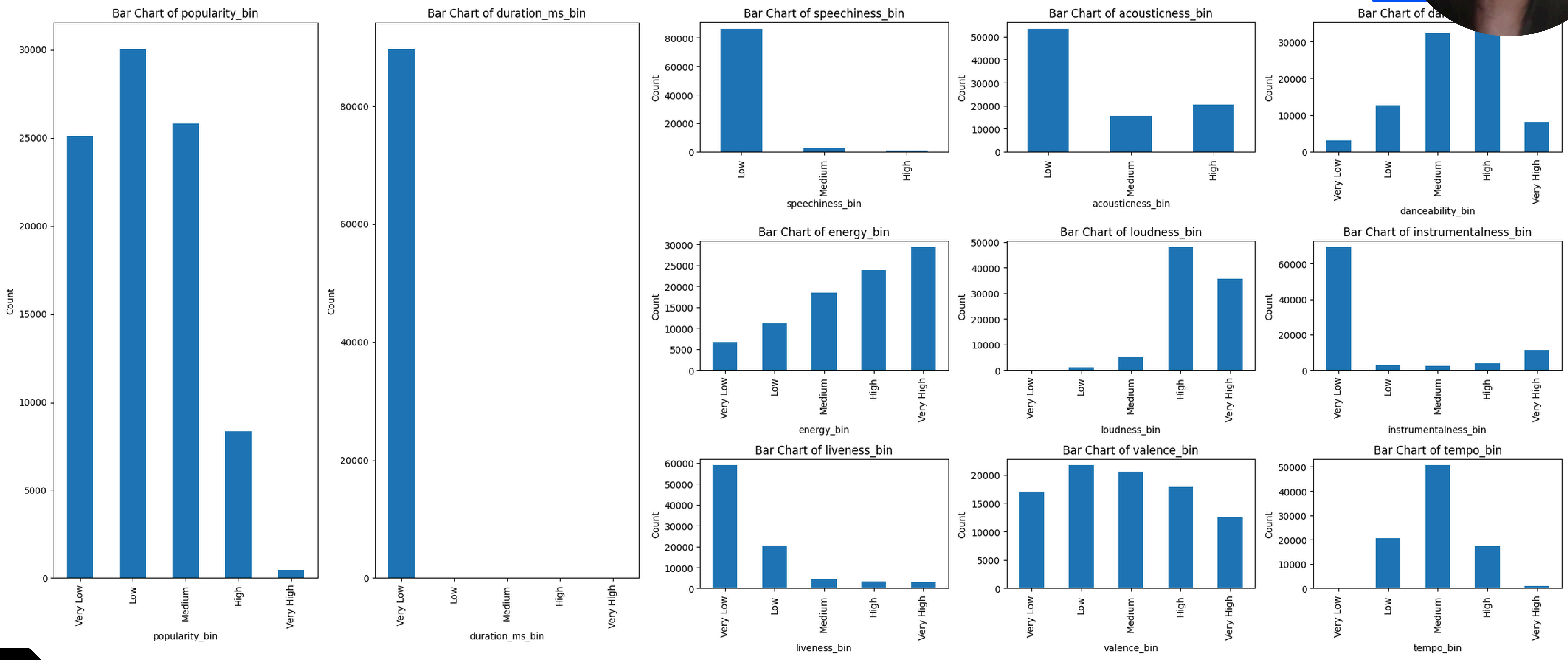
Range >= 3 * standard deviation

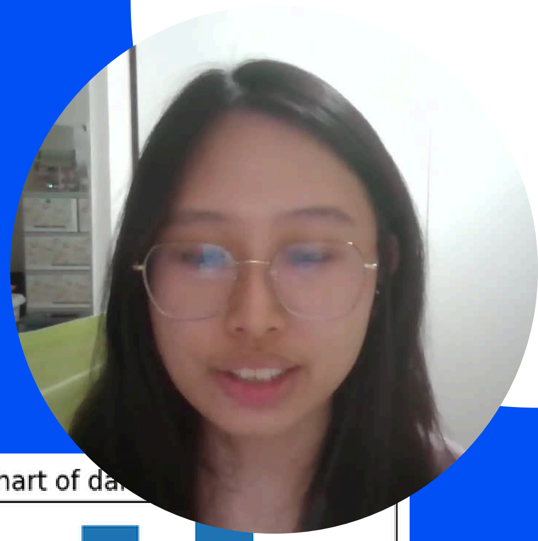
Very Low, Low, Medium, High, Very High

<u>Logic:</u>

If data is more dispersed, assign more bins to capture data granularity.

Preprocessing Music Features





Association Rule Learning

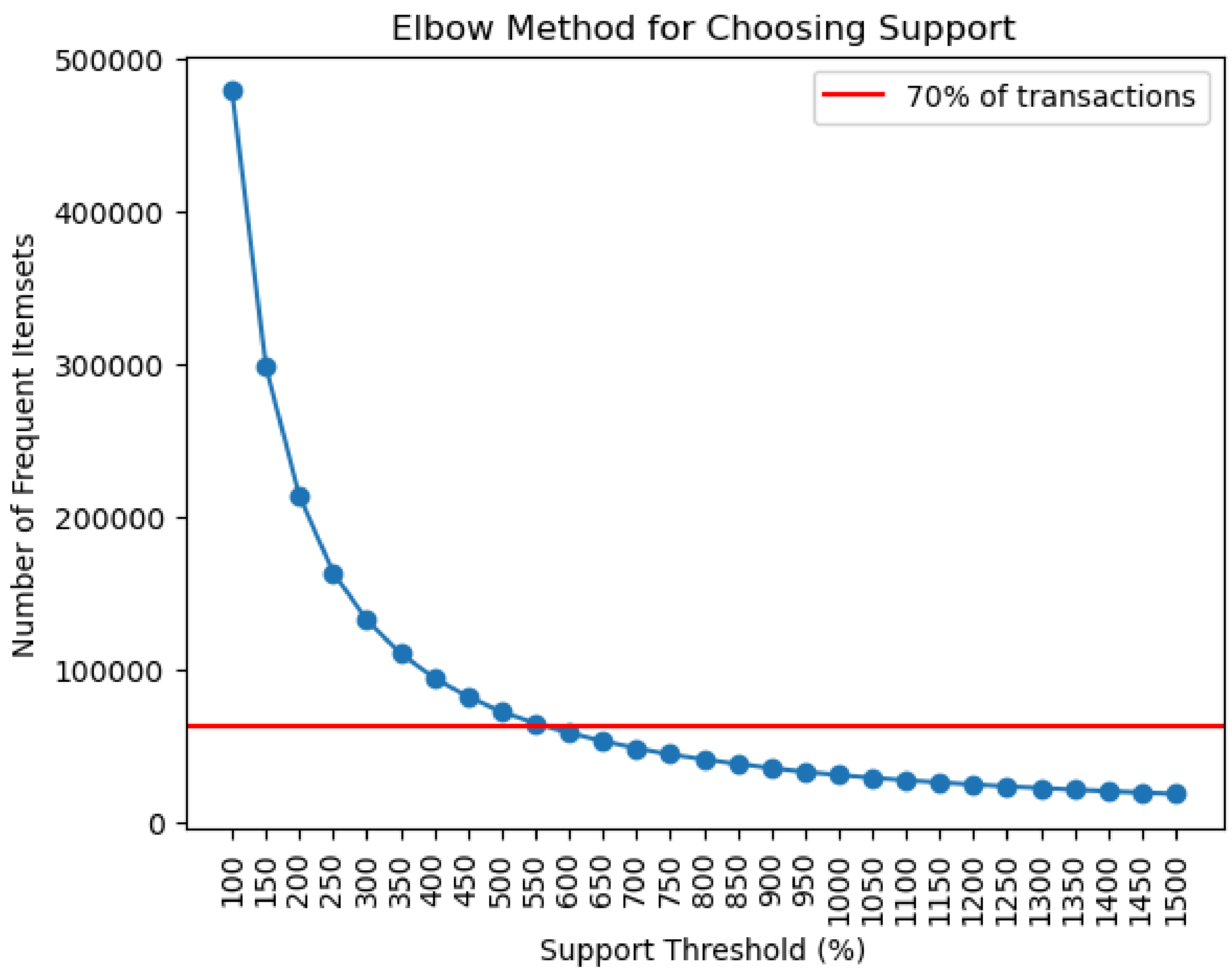
CHOOSING SUPPORT THRESHOLD

Support Levels Tested

100 - 1500 in intervals of 50

Observations

very gradual slope
no clear elbow point



Association Rule Learning

CHOOSING SUPPORT THRESHOLD

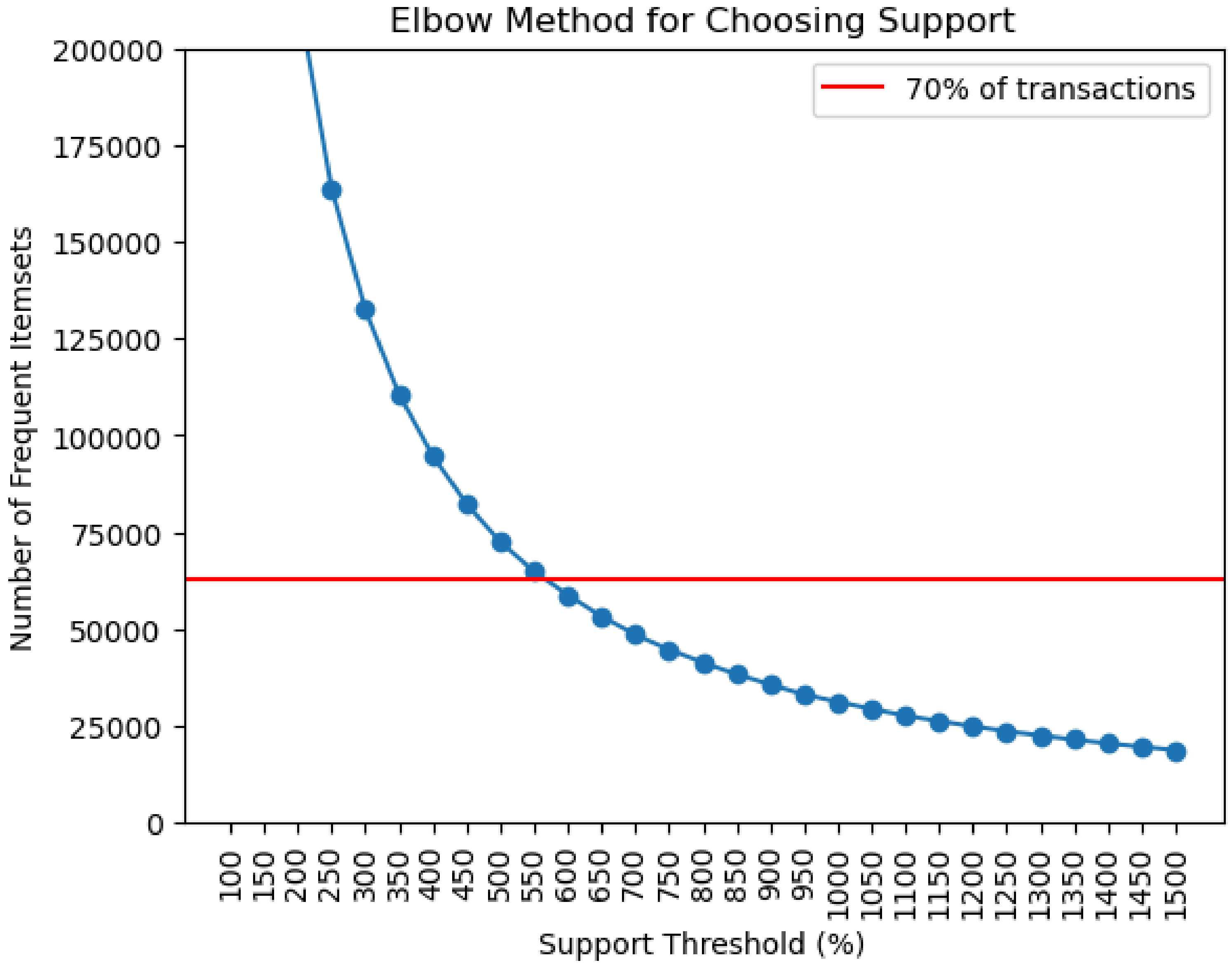
Support Levels Tested

100 - 1500 in intervals of 50

Support Threshold Chosen

550 -- nearest support threshold below horizontal line (70% of transactions)

After further analysis, this level ensures that significant patterns in track popularity is captured



zoomed in

Association Rule Learning

CHOOSING CONFIDENCE THRESHOLD

Confidence Levels Tested

10 - 100 in intervals of 5

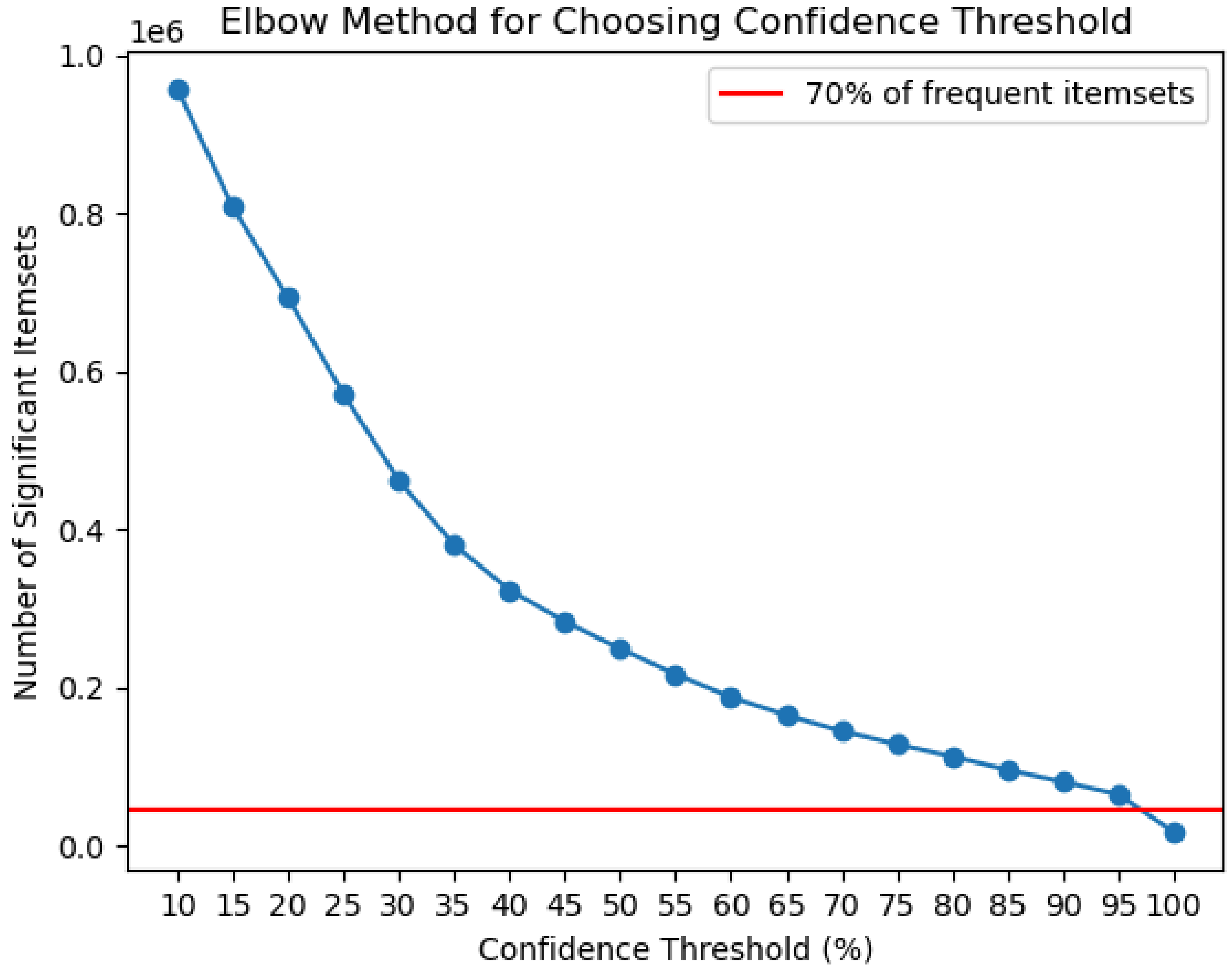
Confidence Threshold Chosen

10

Why 10?

higher threshold levels **eliminate rules** with **popularity as a consequent**

popularity may be **influenced by many factors** rather than a **single dominant pattern/itemset**



Association Rule Learning

INTEREST CALCULATION

In [251…

```
rules_filtered['interest'] = rules_filtered['confidence_pct']-rules_filtered['support_relative']
rules_filtered = rules_filtered.sort_values('interest', ascending=False)
rules_filtered
```

Out[251…

	consequent	antecedent	support_absolute	support_relative	confidence_pct	interest
929457	popularity_bin_Medium	(sertanejo, loudness_bin_Very High, instrument...	643	0.007165	0.993818	0.986652
929444	popularity_bin_Medium	(sertanejo, loudness_bin_Very High, speechines...	643	0.007165	0.993818	0.986652
929453	popularity_bin_Medium	(sertanejo, loudness_bin_Very High, instrument...	643	0.007165	0.993818	0.986652
929441	popularity_bin_Medium	(sertanejo, loudness_bin_Very High, speechines...	643	0.007165	0.993818	0.986652
929437	popularity_bin_Medium	(sertanejo, loudness_bin_Very High, duration_m...	648	0.007221	0.993865	0.986644
...	...	...	...	...	...	...
11961	popularity_bin_Medium	(duration_ms_bin_Very Low,)	25778	0.287252	0.287553	0.000301
18149	popularity_bin_Very Low	(duration_ms_bin_Very Low,)	25044	0.279073	0.279366	0.000293
18147	popularity_bin_Medium	()	25792	0.287408	0.287408	0.000000
24333	popularity_bin_Very Low	()	25081	0.279485	0.279485	0.000000
5567	popularity_bin_Low	()	30030	0.334633	0.334633	0.000000

119536 rows × 6 columns

Association Rule Learning

INTERESTING ITEMSETS

Generation of Results

obtained the longest antecedent of top 5 appearing genres in each popularity bin

Why?

popularity is influenced by many factors and not a single dominant itemset

to obtain a broader perspective on possible factors associated with different popularity levels

('iranian', 'liveness_bin_Very Low', 'speechiness_bin_Low', 'duration_ms_bin_Very Low') ⇒ Very Low popularity

('romance', 'acousticness_bin_High', 'loudness_bin_High', 'instrumentalness_bin_Very Low', 'duration_ms_bin_Very Low') ⇒ Very Low popularity

Sample Antecedents for Very Low Popularity

Association Rule Learning

INTERESTING ITEMSETS

Very Low & Low Popularity

Genres: Iranian, Romance, Grindcore, Chicago House

Tracks tend to be
- short
- have low speech content
- high energy
- loud but not acoustic

Medium Popularity

Genres: Sertanejo, Pagode, Mandopop, Acoustic

Tracks tend to be/have
- low instrumentals
- diverse genres and acoustic elements

High Popularity

Genres: K-pop, Pop

Tracks tend to be/have
- higher emphasis on vocals, short durations, and digital production
- low instrumentals --> indicator of clearer vocal presence

Association Rule Learning

KEY RECOMMENDATIONS

Song popularity is influenced by:

- Genre, loudness, energy, acousticness, speechiness, liveness, instrumentalness, tempo, and duration.

For high popularity (e.g., K-pop, Pop):

- **Low acousticness, low instrumentalness, low speechiness:** more digital & vocal-driven
- **Very short duration**: Increases appeal

For less popular genres (e.g., Iranian, Romance, Grindcore, Chicago-House):

- **Increase loudness, slightly reduce duration:** Aligns with popular song characteristics
- **Lower acousticness & instrumentalness:** Could enhance appeal.

High-energy genres (e.g., Party, Sertanejo):

- **Boosts energy levels, possibly increasing popularity:** Tend to engage listeners more effectively

Association Rule Learning

LEARNINGS

Goal of Analysis

spotify dataset had a more possibilities in terms of what could be analyzed compared to the groceries dataset

important that the goal of analysis was decided at the beginning as this directed the flow of the association rule learning

Heavy Preprocesing

unlike the groceries dataset, the spotify dataset required more preprocessing (especially for the numerical data)

Correlation Does Not Imply Causality

Thank you!