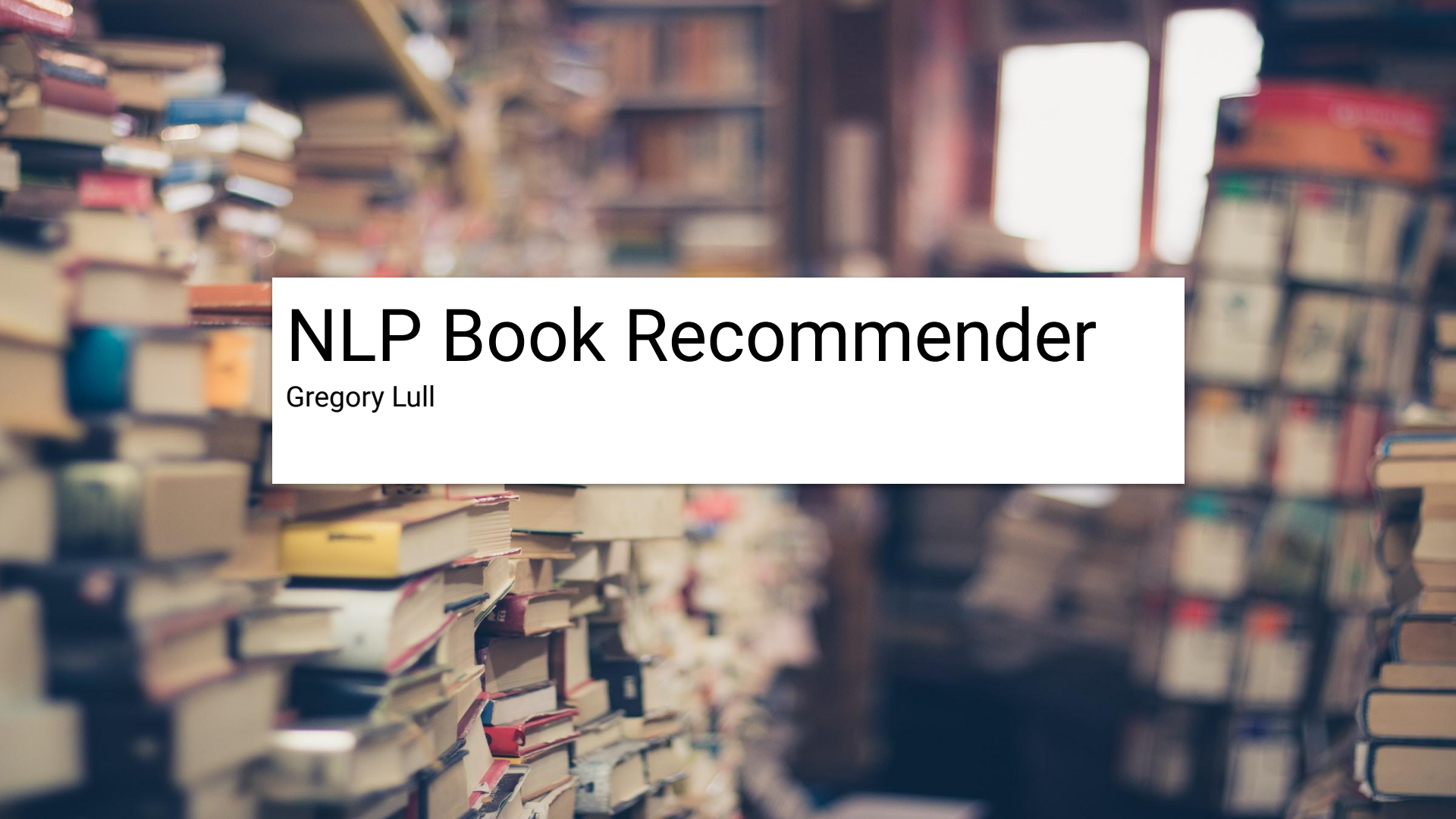


NLP Book Recommender

Gregory Lull

Reading Stats

- Adults read 4 to 13 books / year
- Lifetime: 100 - 1000 books

Data Collection

- 500 Kindle books
- 500 out of 50,000 Project Gutenberg books
- Cross referenced with Goodreads top 10k books

Methodology

- Clean text (lemmatize, stop words)
- → tokenize (bigrams, CountVectorizer | TF-IDF)
- → reduce dimensions (LSA | NMF)
- → cluster (KMeans)
- → cosine similarity for sorting results
- Final combo: TF-IDF → LSA → KMeans

Analysis

Similar books will cluster together, e.g. Harry Potter 1 should be in the same cluster as Harry Potter 2 - 7

Topic 1: pron, say, know, like, look, come, think, time, man, good

Topic 3: say, harry, mr, look, weasley, mrs, hermione, ron, little

But not always: Narnia, Dune, Ender’s Game

Results - Pride and Prejudice Top 10

1. JM Barrie - Peter Pan (1904)
2. **Jane Austen - Northanger Abbey (1803)**
3. EM Forster - A Room with a View (1908)
4. George Eliot - Daniel Deronda (1876)
5. Jhumpa Lahiri - Interpreter of Maladies (1999)
6. ME Braddon - Lady Audley's Secret (1862)
7. Thomas Hardy - The Woodlanders (1887)
8. **Jane Austen - Mansfield Park (1814)**
9. Edith Wharton - The Age of Innocence (1920)
10. Charles Dickens - Nicholas Nickleby (1838)

Conclusion

- Topics modeling words are related to each other, but can hijack clusters and “overfit”
- This can return a list of “similar books” but needs more guidance on how to form clusters.

Future Work

- Analyze dialogue sentiment and build character profiles.
- Use Goodreads user ratings to sort clusters.
- Get Kindle preview samples, need only the first 10%

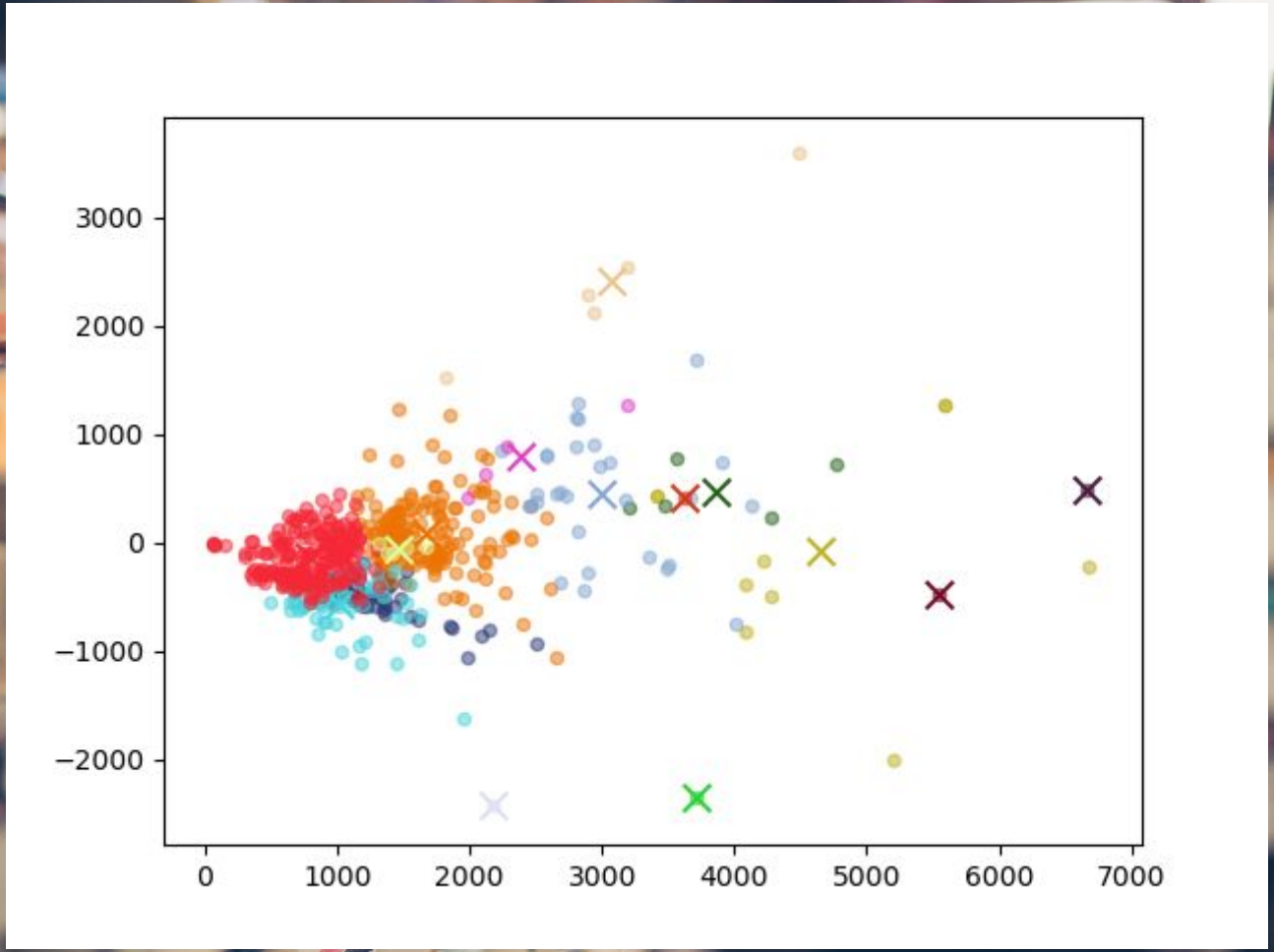
Appendix

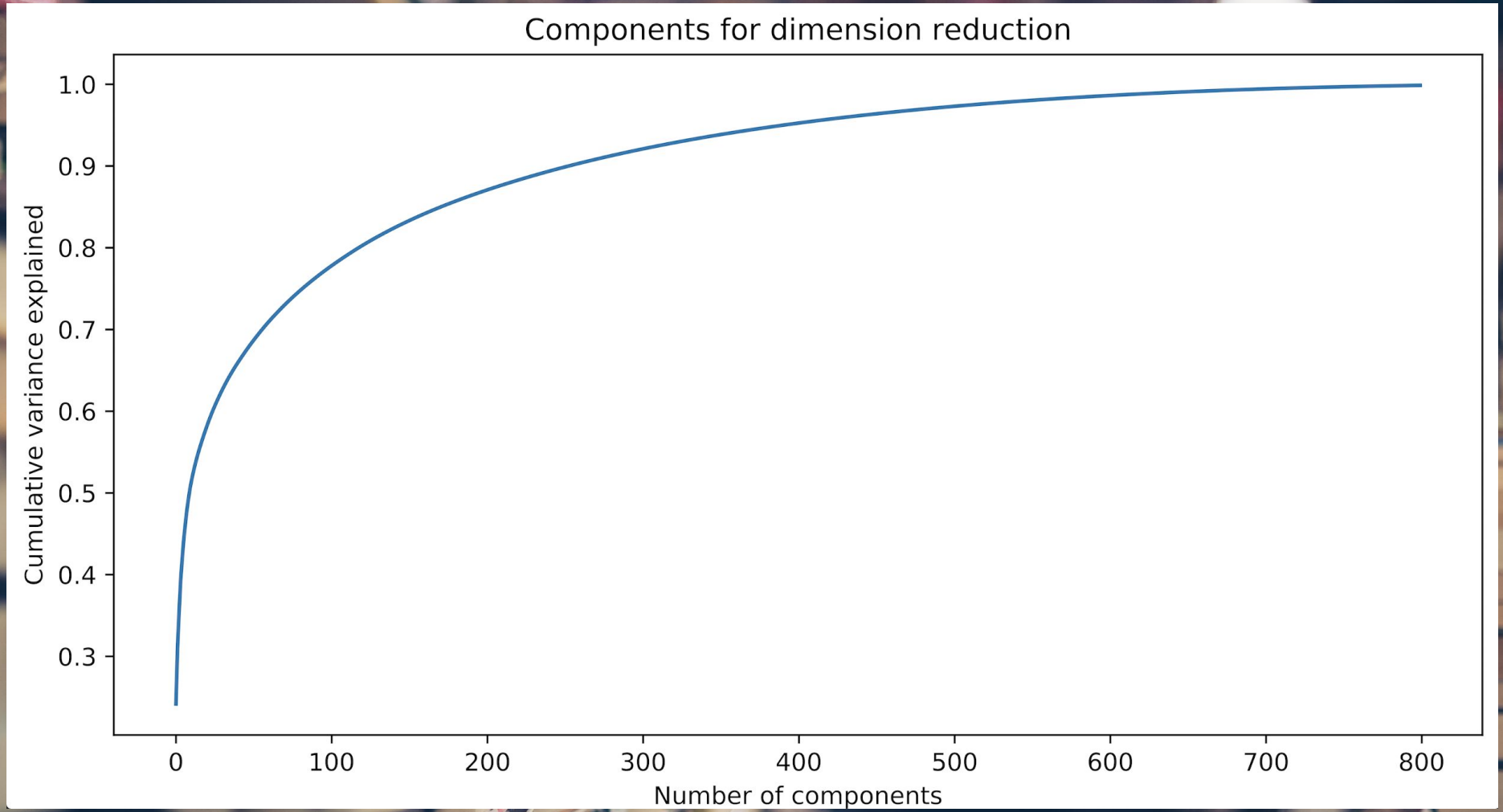
Topic 0 pron, say, like, know, look, come, think, time, man, good, tell, want, day, thing, way, hand, little, feel, pron know, people

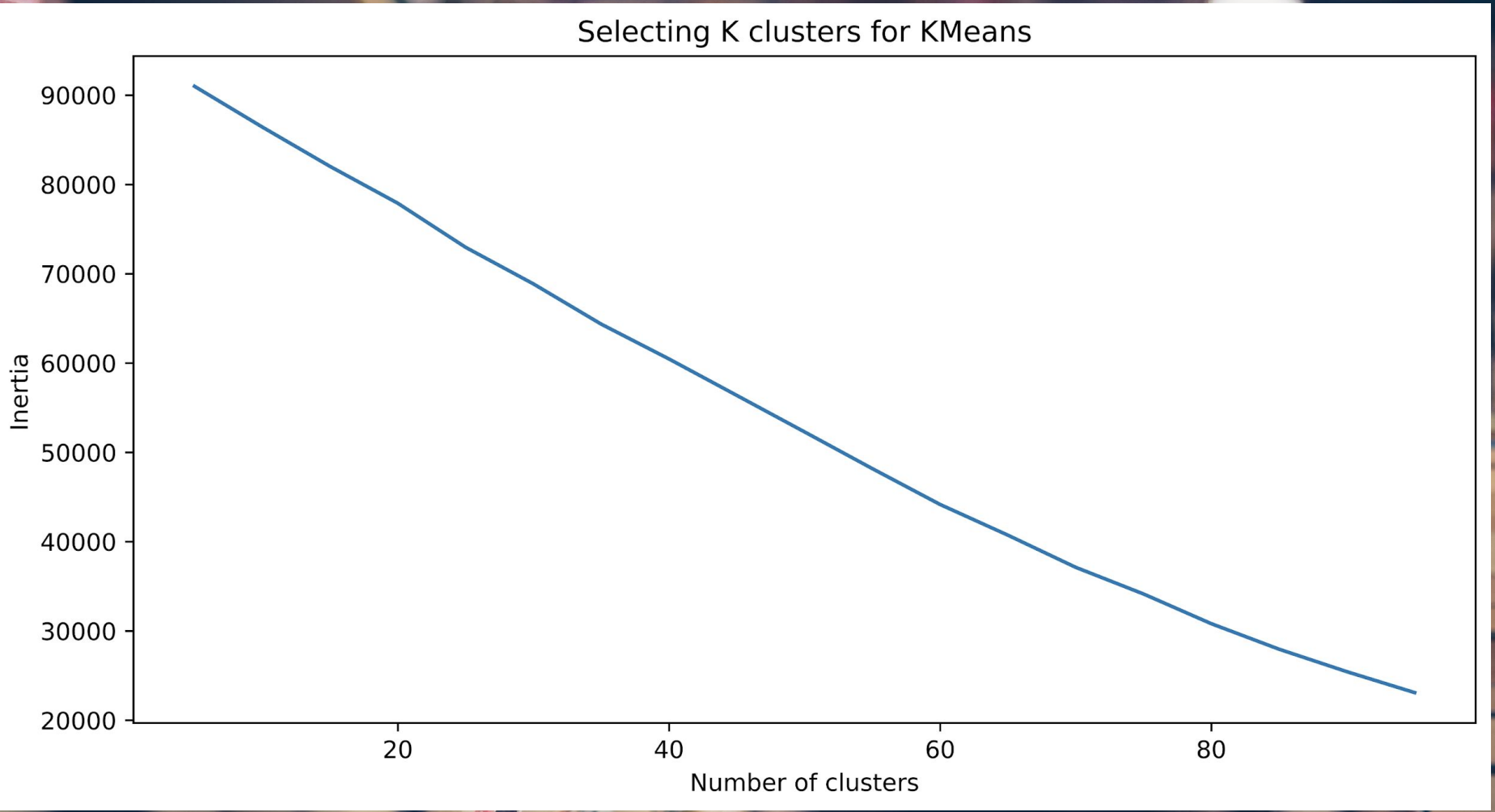
Topic 1 man, time, good, people, come, great, year, day, work, new, life, way, thou, know, world, little, old, long, thing, state

Topic 2 thou, thy, man, haue, say, thee, mr, sir, shall, come, lord, mrs, hath, king, pron haue, ye, th, lady, pg, little

Topic 3 say, harry, mr, look, weasley, mrs, hermione, ron, little, mr weasley, come, think, dumbledore, say pron, mrs weasley, know, door, say harry, say mr, ask


3000
2000
1000
0
−1000
−2000
0
1000
2000
3000
4000
5000
6000
7000


Components for dimension reduction
Cumulative variance explained
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0
100
200
300
400
500
600
700
800
Number of components


Selecting K clusters for KMeans
Inertia
90000
80000
70000
60000
50000
40000
30000
20000
20
40
60
80
Number of clusters