Matthew Wolff

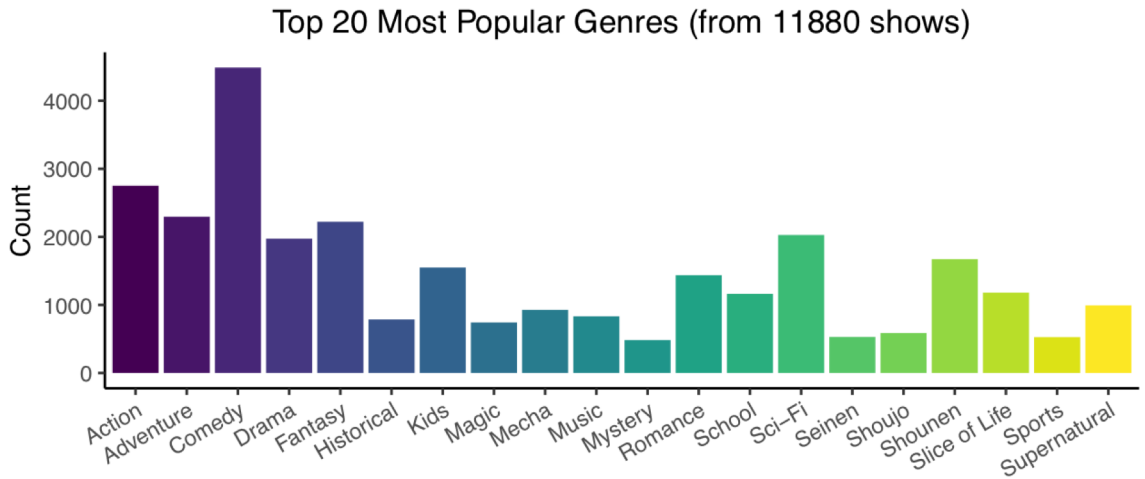
Making Recommendations from Show Rating Data Using PCA & Clustering

Cleaning

- Read in the CSV files for **shows** (1MB) and **ratings** (106MB)
- **Show** variables:
  - "show_id" "name" "genre" "type" "episodes" "rating" "members"
- **Rating** variables: "user_id" "show_id" "rating"
- Do generic cleaning (converting types, splitting text lists into R lists)
  - Eliminate or fill in missing values (*only for a few of the most popular*)

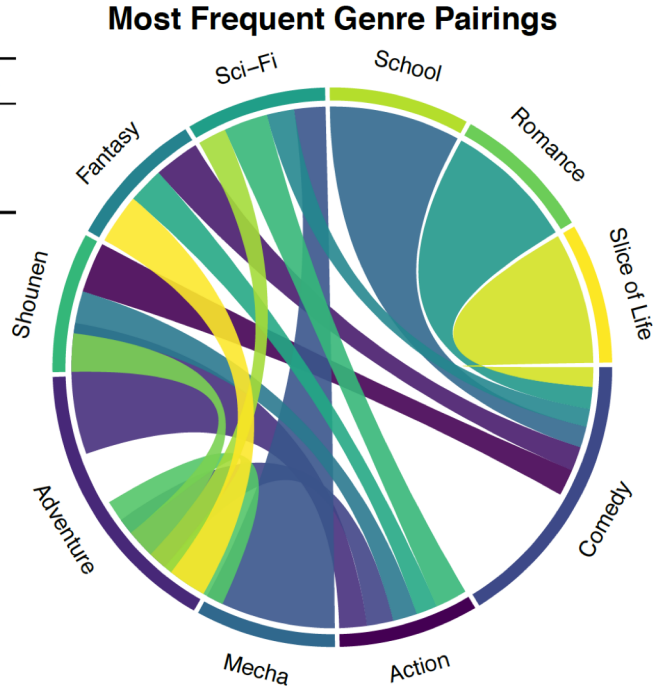
Table 1: Summary of Show Data and Missing Values

	Minimum	1st Quartile	Median	Mean	3rd Quartile	Max	Missing Vals
episodes	1.00	1.00	2.00	12.38	12.00	1818.00	340
rating	1.670	5.880	6.570	6.474	7.180	10.000	230
members	5	225	1550	18071	9437	1013917	0



Exploratory Analysis

- **Most common** genres (comedy)
- **Most frequently paired** genres (*notice*: comedy is often a standalone genre!)
- Certain genre pairings more common than others
- Genre preference could be useful
  - If a user likes certain shows, they might like similarly themed shows



Frequent Pairing Counts

Genre1	Genre2	Count
Action	Sci-Fi	1008
Adventure	Fantasy	921
Comedy	Shounen	918
Adventure	Comedy	895
Action	Adventure	867
Comedy	Fantasy	839
Action	Comedy	838
Action	Shounen	766
Comedy	Romance	753
Comedy	School	729
Mecha	Sci-Fi	712
Adventure	Shounen	692
Comedy	Slice of Life	672
Action	Fantasy	663
Adventure	Sci-Fi	646
Comedy	Sci-Fi	624

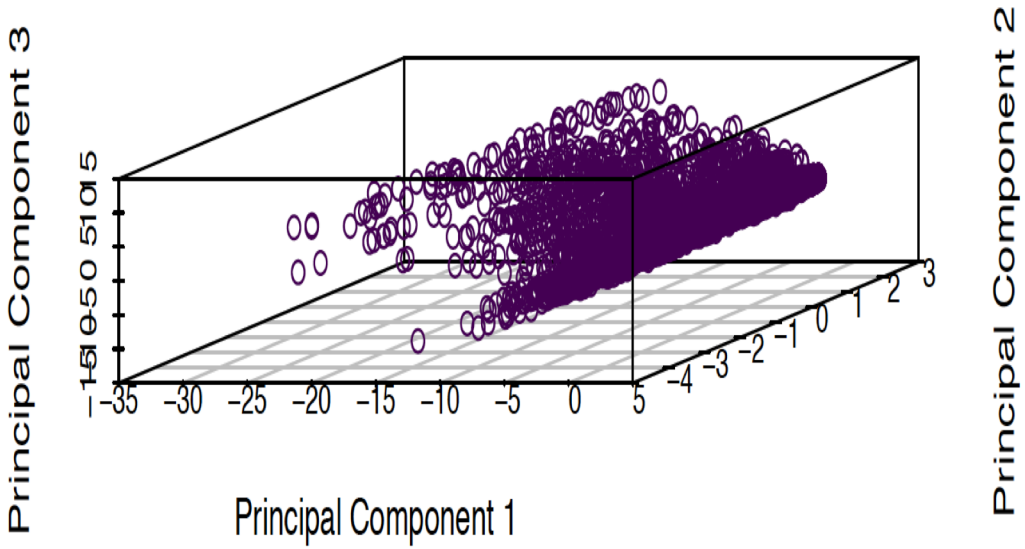
Wrangling Data

- **Join the normalized data sets (on "show_id")**
  - Rename shared **rating** column to reflect **user ratings** *vs* **show ratings**
- **Determine which ratings to treat as "good"**
  - Compute the mean rating for every user based on their rated shows
  - If they rate a show higher than their mean rating, they relatively liked it!
- **Take a subset of the data**
  - Filter out *users* whose ID is greater than 50,000 **(keep 70% of users)**
  - Filter out *entries* where the user rated lower than their average rating **(50%)**
  - Filter out *shows* whose number of "members" (AKA ratings) is less than 100,000 **(keep 5% of shows, which contain 55% of all ratings)**
- **Cross-tabulate *user_id* and *show_id***
  - This will create a contingency table we can use for PCA
  - Join with our data subset on "user_id" to create our PCA input
    - Keep crosstab columns + "type", "episodes", "mean_rating", "members", and "num_genres" *(a created column w/number of genres)*

Table 2: Principal Component Analysis: Top Eigenvalues

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	176.25673	32.340685	32.34069
comp 2	35.27865	6.473147	38.81383
comp 3	33.00825	6.056560	44.87039
comp 4	28.20805	5.175789	50.04618
comp 5	17.47536	3.206489	53.25267

Post–PCA: Three Principal Components, Unclustered

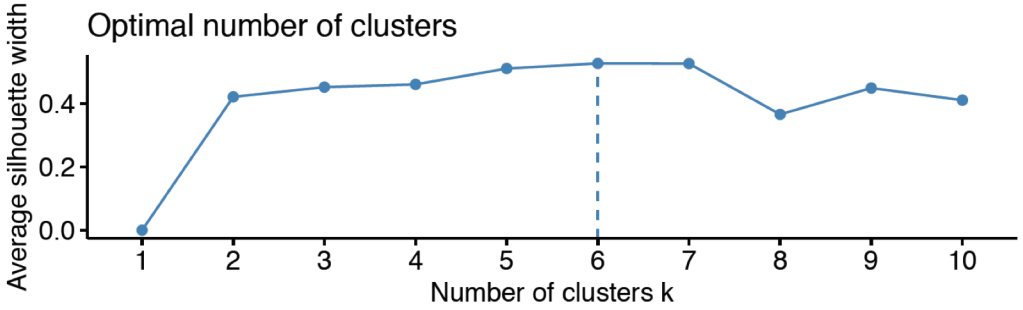


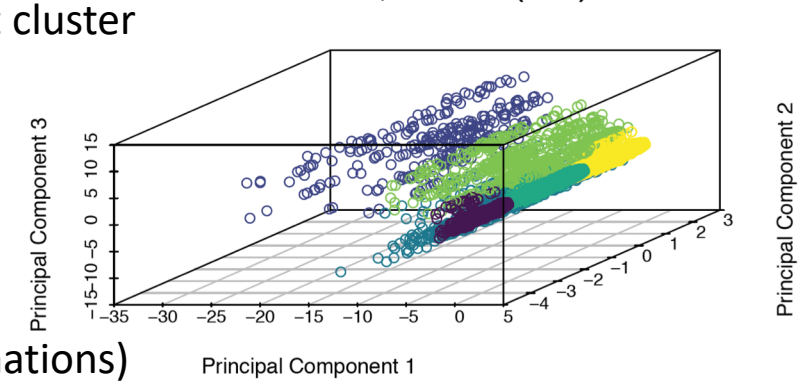
Principal Component Analysis

- **Perform PCA on the data**
- **Capture 45% of variance with first 3 components**
- **Visualize!**

Clustering

- **Do k-means clustering with *k* = { 1 .. 10 }**
  - Use silhouette analysis to evaluate *k* clusters
  - *Hard* clustering, so points are only assigned nearest cluster
- **Not shown: Clustering using Gaussian Mixture Models**
  - Not very easy to show.....
  - *Soft* clustering allows for recommendations outside of closest cluster
  - Also allows for evaluating how good of a match a cluster is!





Generating Recommendations

- **Introduce two new users**
  - Each has their own history of show ratings
- **Determine which cluster new user belongs to**
  - Extract three principal components (use known linear combinations)
  - Plot, determine nearest cluster
- **Isolate nearest cluster as a group of points**
  - Sort by rating –> recommendation!
  - Extract all genres & analyze most common cluster genres *(on right)*

Table 3: Cluster 1 Recommendations

Name	Genres
Fullmetal Alchemist: Brotherhood	Action, Adventure, Drama, Fantasy, Magic, Military, Shounen
Naruto: Shippuuden	Action, Comedy, Martial Arts, Shounen, Super Power
Darker than Black: Ryuusei no Gemini	Action, Mystery, Sci-Fi, Super Power

Table 4: Cluster 2 Recommendations

Name	Genres
Clannad: After Story	Drama, Fantasy, Romance, Slice of Life, Supernatural
Shigatsu wa Kimi no Uso	Drama, Music, Romance, School, Shounen
Kimi ni Todoke 2nd Season	Romance, School, Shoujo, Slice of Life