Steve's Presentation

Exploratory Data Analysis

IMDB Dataset (Kaggle)

The movies dataset includes 85,855 movies with attributes such as movie description, average rating, number of votes, genre, etc.

The ratings dataset includes 85,855 rating details from demographic perspective.

The names dataset includes 297,705 cast members with personal attributes such as birth details, death details, height, spouses, children, etc.

The title principals dataset includes 835,513 cast members roles in movies with attributes such as IMDb title id, IMDb name id, order of importance in the movie, role, and characters played.

'imdb_title_id', 'title', 'original_title', 'year', 'date_published', 'genre', 'duration', 'country', 'language', 'director', 'writer', 'production_company', 'actors', 'description', 'avg_vote', 'votes', 'budget', 'usa_gross_income', 'worlwide_gross_income', 'metascore', 'reviews_from_users', 'reviews_from_critics'

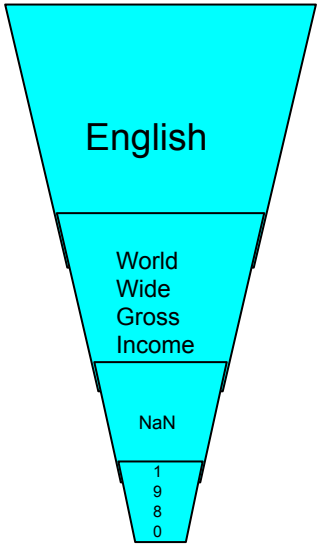
85K rows of data - matched on 'imdb_title_id'

Where To Search In This Vast Dataset

Movies from years 1894 - 2020

:

	imdb_title_id	title	original_title	year	date_published	genre	duration	country	language	director	...	females_30age_avg_vote	females_
0	tt0035423	Kate & Leopold	Kate & Leopold	2001	2002-03-01	Comedy, Fantasy, Romance	118	USA	English, French	James Mangold	...	6.5	
1	tt0079285	Saturn 3	Saturn 3	1980	1980-05-08	Adventure, Horror, Sci-Fi	96	UK	English	Stanley Donen, John Barry	...	5.5	
2	tt0080319	Dalle 9 alle 5... orario continuato	Nine to Five	1980	1981-03-26	Comedy	109	USA	English, French	Colin Higgins	...	7.1	
3	tt0080339	L'aereo più pazzo del mondo	Airplane!	1980	1980-10-30	Comedy	88	USA	English	Jim Abrahams, David Zucker	...	7.4	
4	tt0080360	Stati di allucinazione	Altered States	1980	1981-11-20	Horror, Sci-Fi, Thriller	102	USA	English, Spanish	Ken Russell	...	6.6	



Notionally...

- which movies did users generally find similar
- did types of movies separate by voter type
- interesting findings within

Data Cleaning

Trim Rows, Nan's, Columns

1980+

```
In [127]: # something was throwing off my conversation of years to ints
          years = set()
          for y in movie_data.year:
              try:
                  years.add(int(y))
              except:
                  print(y) #found it
          print(years)

          TV Movie 2019
          {1894, 1906, 1911, 1912, 1913, 1914, 1915, 1916, 1917, 1918, 1919, 1920, 1921, 1922, 1923, 1924, 1925, 1926, 1927,
          1928, 1929, 1930, 1931, 1932, 1933, 1934, 1935, 1936, 1937, 1938, 1939, 1940, 1941, 1942, 1943, 1944, 1945, 1946,
          1947, 1948, 1949, 1950, 1951, 1952, 1953, 1954, 1955, 1956, 1957, 1958, 1959, 1960, 1961, 1962, 1963, 1964, 1965,
          1966, 1967, 1968, 1969, 1970, 1971, 1972, 1973, 1974, 1975, 1976, 1977, 1978, 1979, 1980, 1981, 1982, 1983, 1984,
          1985, 1986, 1987, 1988, 1989, 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003,
          2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020}

In [128]: movie_data = movie_data[movie_data.year != 'TV Movie 2019']

In [129]: len(movie_data) #check

Out[129]: 85854

In [130]: movie_data['year'] = movie_data['year'].astype(int)

In [263]: #movie_data.year[:10] #verify

In [132]: # now get rid of anything before 1980
          movie_data = movie_data[movie_data.year >=1980]

In [133]: 'English' in movie_data.language.iloc[2]

Out[133]: True

In [135]: movie_data = movie_data[movie_data['language'].str.contains('English', na=False)]

In [136]: movie_data['worlwide_gross_income'].isna().sum() # is there enough data to do my analysis?

Out[136]: 18130

In [137]: # ok get rid of the movie-data with no worldwide gross income
          movie_data = movie_data[movie_data['worlwide_gross_income'].notna()]
```

Data Merging

Combining Data
(Left) Genre, Filters
+
(Right) Voter Groups

```
ratings_data.columns
```

```
Index(['imdb_title_id', 'weighted_average_vote', 'total_votes', 'mean_vote',
       'median_vote', 'votes_10', 'votes_9', 'votes_8', 'votes_7', 'votes_6',
       'votes_5', 'votes_4', 'votes_3', 'votes_2', 'votes_1',
       'allgenders_0age_avg_vote', 'allgenders_0age_votes',
       'allgenders_18age_avg_vote', 'allgenders_18age_votes',
       'allgenders_30age_avg_vote', 'allgenders_30age_votes',
       'allgenders_45age_avg_vote', 'allgenders_45age_votes',
       'males_allages_avg_vote', 'males_allages_votes', 'males_0age_avg_vote',
       'males_0age_votes', 'males_18age_avg_vote', 'males_18age_votes',
       'males_30age_avg_vote', 'males_30age_votes', 'males_45age_avg_vote',
       'males_45age_votes', 'females_allages_avg_vote',
       'females_allages_votes', 'females_0age_avg_vote', 'females_0age_votes',
       'females_18age_avg_vote', 'females_18age_votes',
       'females_30age_avg_vote', 'females_30age_votes',
       'females_45age_avg_vote', 'females_45age_votes',
       'top1000_voters_rating', 'top1000_voters_votes', 'us_voters_rating',
       'us_voters_votes', 'non_us_voters_rating', 'non_us_voters_votes'],
      dtype='object')
```

```
#merge the two
movie_data = movie_data.merge(ratings_data, left_on='imdb_title_id', right_on='imdb_title_id')
```

```
movie_data.head()
```

	imdb_title_id	title	original_title	year	date_published	genre	duration	country	language	director	...	females_30age_avg_vote	females_
0	tt0035423	Kate & Leopold	Kate & Leopold	2001	2002-03-01	Comedy, Fantasy, Romance	118	USA	English, French	James Mangold	...	6.5	
1	tt0079285	Saturn 3	Saturn 3	1980	1980-05-08	Adventure, Horror, Sci-Fi	96	UK	English	Stanley Donen, John Barry	...	5.5	
2	tt0080319	Dalle 9 alle 5... orario continuato	Nine to Five	1980	1981-03-26	Comedy	109	USA	English, French	Colin Higgins	...	7.1	
3	tt0080339	L'aereo più pazzo del mondo	Airplane!	1980	1980-10-30	Comedy	88	USA	English	Jim Abrahams, David Zucker	...	7.4	
4	tt0080360	Stati di allucinazione	Altered States	1980	1981-11-20	Horror, Sci-Fi, Thriller	102	USA	English, Spanish	Ken Russell	...	6.6	

5 rows × 70 columns

Focusing on Metrics

What might lead to ‘like’ movies
Define ‘like’

```
#obligatory tops and bottoms
movie_data.sort_values('weighted_average_vote', ascending=False)[['title', 'genre', 'weighted_average_vote']][:20]
```

	title	genre	weighted_average_vote
2765	Le ali della libertà	Drama	9.3
14545	The Transcendents	Music, Mystery, Thriller	9.2
7328	Il cavaliere oscuro	Action, Crime, Drama	9.0
2743	Pulp Fiction	Crime, Drama	8.9
4198	Il Signore degli Anelli - Il ritorno del re	Action, Adventure, Drama	8.9
13489	The Crucible	Drama	8.9
2519	Schindler's List	Biography, Drama, History	8.9
15327	Distant Sky: Nick Cave & The Bad Seeds Live in...	Music	8.8
3937	Fight Club	Drama	8.8
11541	The Phantom of the Opera at the Royal Albert Hall	Drama, Music, Musical	8.8
8537	NT Live: Cyrano de Bergerac	Drama, Romance	8.8
3733	Il Signore degli Anelli - La compagnia dell'An...	Action, Adventure, Drama	8.8
8760	Metallica and San Francisco Symphony S&M2	Music	8.8
2644	Forrest Gump	Drama, Romance	8.8
10852	Les Misérables in Concert: The 25th Anniversary	Drama, Music, Musical	8.8
9772	Inception	Action, Adventure, Sci-Fi	8.8
10965	National Theatre Live: Frankenstein	Drama, Sci-Fi	8.7
4199	Il Signore degli Anelli - Le due torri	Action, Adventure, Drama	8.7
3913	Matrix	Action, Sci-Fi	8.7
14668	National Theatre Live: A View from the Bridge	Drama	8.7

Focusing on Metrics

What might lead to ‘like’ movies
Define ‘like’

```
#obligatory tops and bottoms
movie_data.sort_values('weighted_average_vote', ascending=False)[['original_ti
```

	original_title	genre	weighted_average_vote
2765	The Shawshank Redemption	Drama	9.3
14545	The Transcendents	Music, Mystery, Thriller	9.2
7328	The Dark Knight	Action, Crime, Drama	9.0
2743	Pulp Fiction	Crime, Drama	8.9
4198	The Lord of the Rings: The Return of the King	Action, Adventure, Drama	8.9
13489	The Crucible	Drama	8.9
2519	Schindler's List	Biography, Drama, History	8.9
15327	Distant Sky - Nick Cave & The Bad Seeds Live i...	Music	8.8
3937	Fight Club	Drama	8.8
11541	The Phantom of the Opera at the Royal Albert Hall	Drama, Music, Musical	8.8
8537	NT Live: Cyrano de Bergerac	Drama, Romance	8.8
3733	The Lord of the Rings: The Fellowship of the Ring	Action, Adventure, Drama	8.8
8760	Metallica & San Francisco Symphony - S&M2	Music	8.8
2644	Forrest Gump	Drama, Romance	8.8
10852	Les Misérables in Concert: The 25th Anniversary	Drama, Music, Musical	8.8
9772	Inception	Action, Adventure, Sci-Fi	8.8
10965	National Theatre Live: Frankenstein	Drama, Sci-Fi	8.7
4199	The Lord of the Rings: The Two Towers	Action, Adventure, Drama	8.7
3913	The Matrix	Action, Sci-Fi	8.7
14668	National Theatre Live: A View from the Bridge	Drama	8.7

Focusing on Metrics

What might lead to ‘like’ movies
Define ‘like’

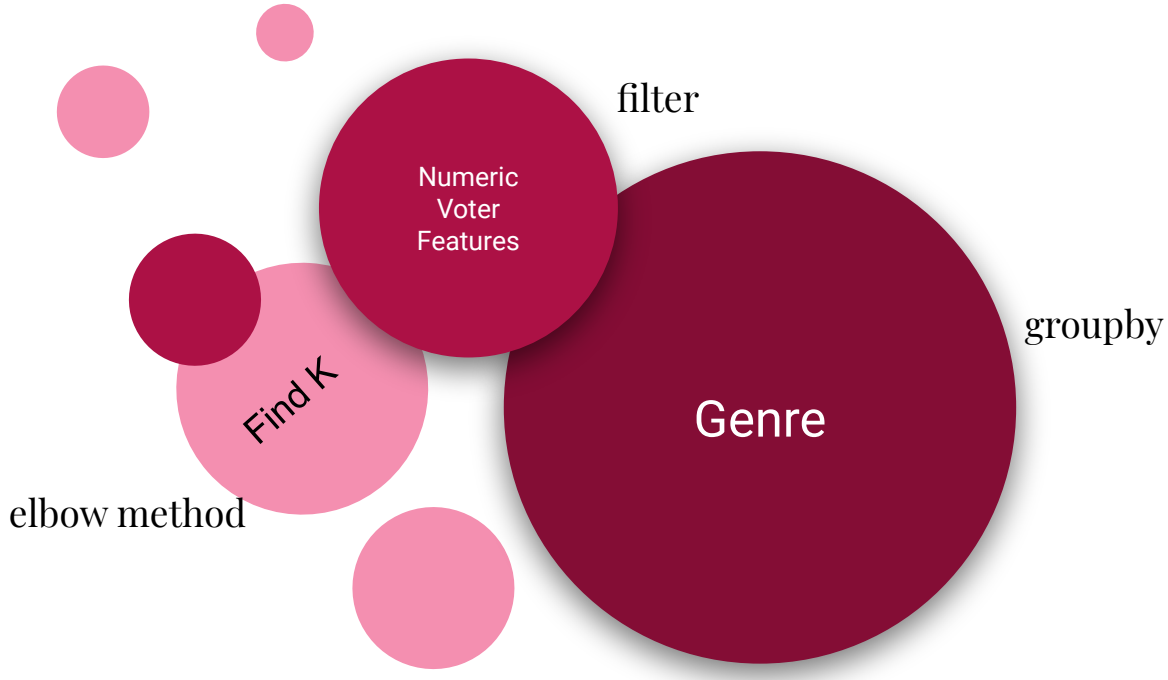
```
: data.sort_values('weighted_average_vote', ascending=True)[['original_title', '
```

:

	original_title	genre	weighted_average_vote
8515	Proud American	Drama	1.1
13394	Saving Christmas	Comedy, Family	1.4
14572	Smolensk	Drama, Thriller	1.4
4764	Foodfight!	Animation, Action, Adventure	1.5
8520	2177: The San Francisco Love Hacker Crimes	Sci-Fi	1.5
14356	Izzie's Way Home	Animation, Adventure	1.5
14196	Aerials	Drama, Sci-Fi, Thriller	1.6
9780	Kung Fu Joe	Comedy	1.7
14440	The Adventures of Panda Warrior	Animation, Action, Adventure	1.8
6651	Pledge This!	Comedy	1.8
5754	Exorcism	Horror, Thriller	1.8
11772	The Obama Effect	Comedy, Drama	1.8
15410	Trolled	Animation, Adventure, Comedy	1.8
15169	Sex and the Future	Comedy	1.9
7877	The Hottie & the Nottie	Comedy, Romance	1.9
9258	Disaster Movie	Comedy	1.9
9444	Hari Puttar: A Comedy of Terrors	Comedy, Drama, Family	1.9
4983	Superbabies: Baby Geniuses 2	Comedy, Family, Sci-Fi	1.9
10068	The Prodigy	Animation	1.9
12638	Amy	Horror	2.0

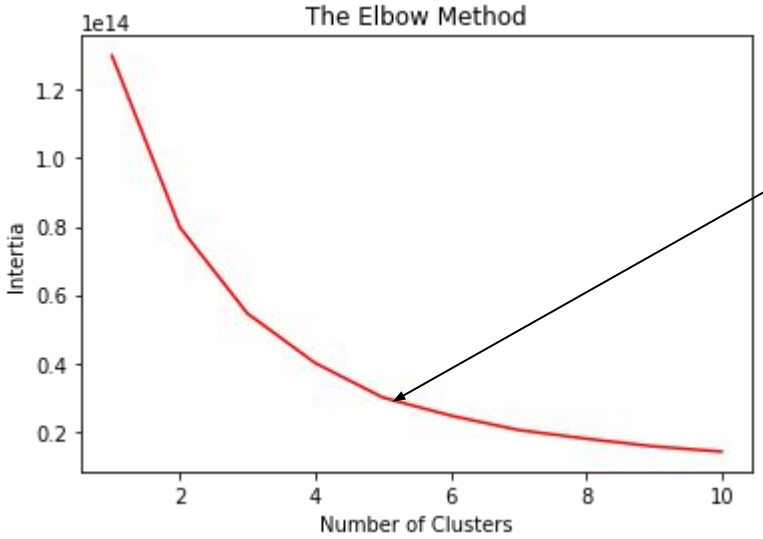
Genres and Weighted Average Score - A K Means Study

K Means Study



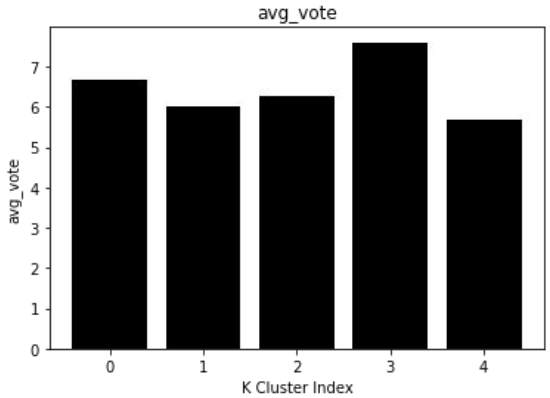
Inspect what ‘natural’ clusters have in common and who is in them

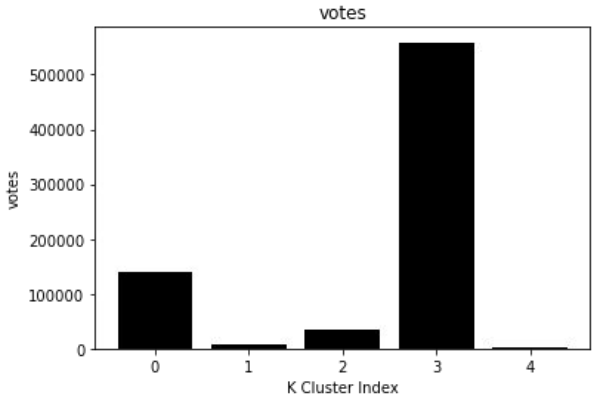
K Means Analysis

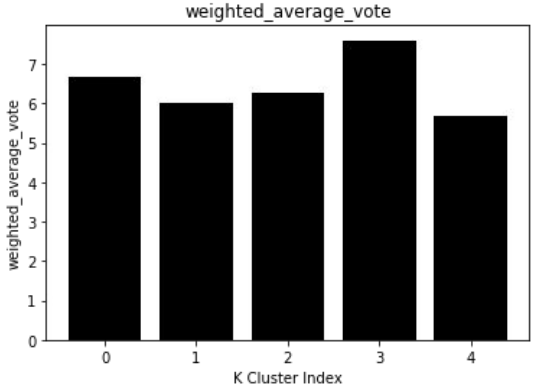


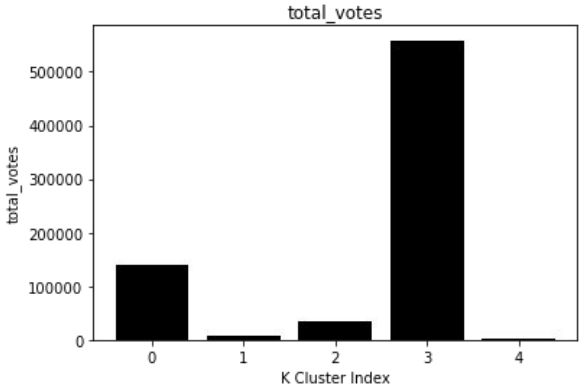
Calling this due to 'hopefully' better explainability & separation

Inspecting the clusters...some interesting findings...

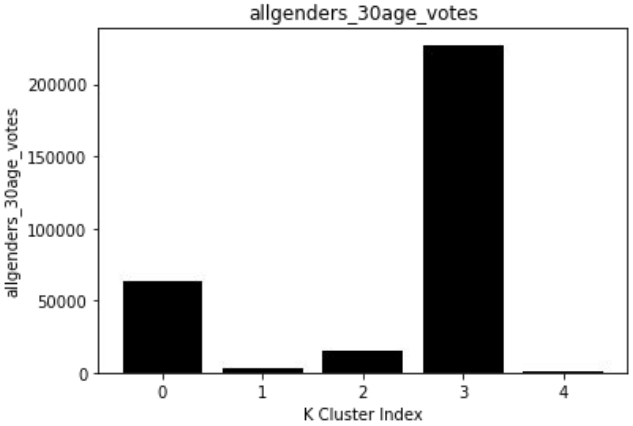


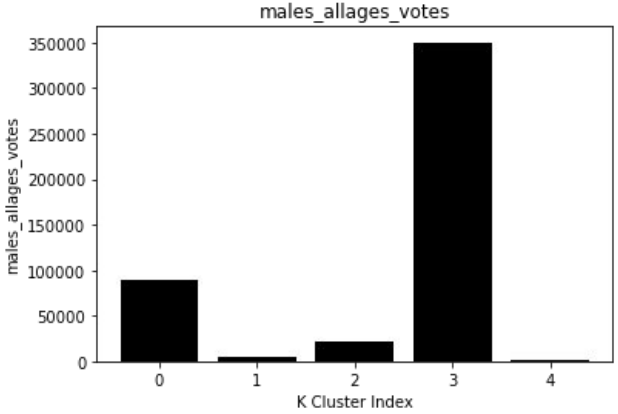


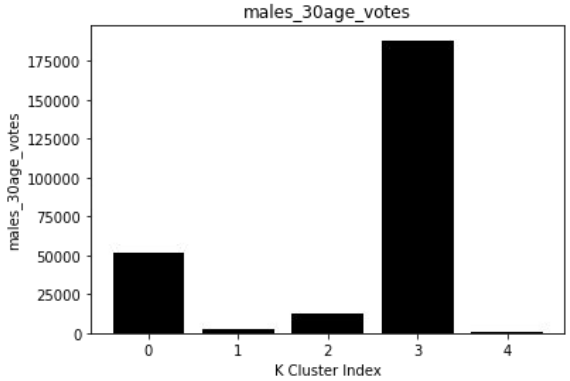


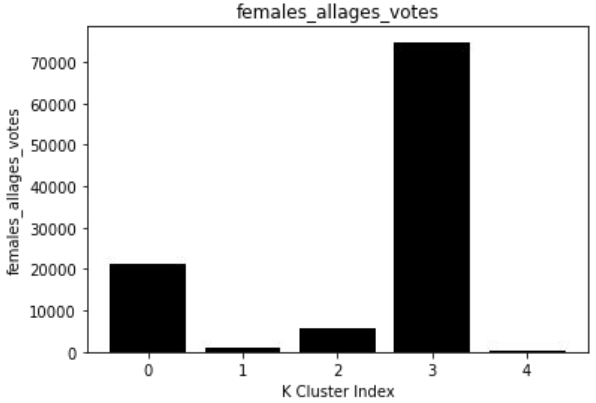


Inspecting the clusters...some interesting findings...









Let’s pivot on this thread a little bit

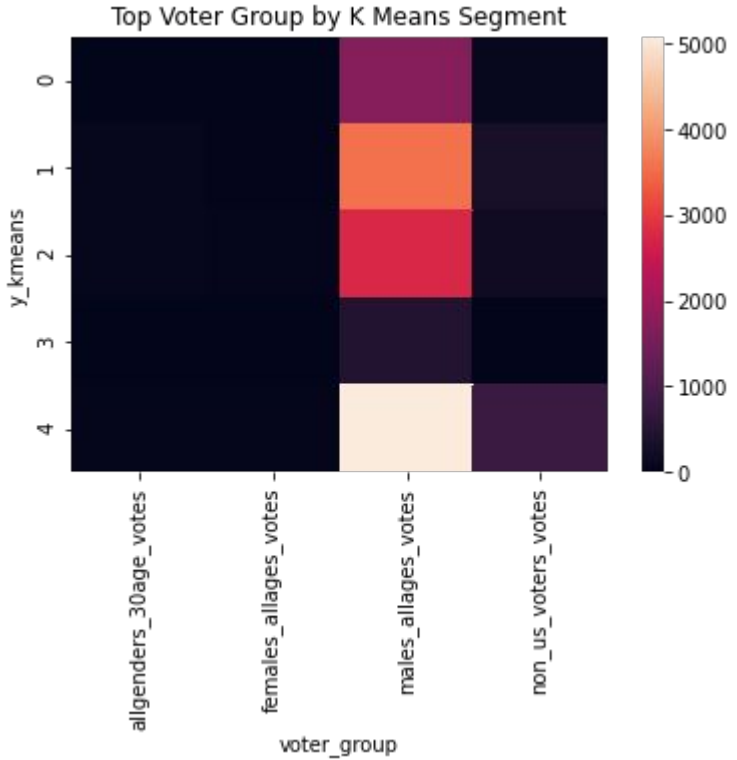
Voter Info

What is in cluster 3:

2007-20012

Lot’s of blockbuster hits

Not much initial variance amongst voters



What to watch...No Surprise Here!

Just...The Best

If you remove the best

Winner for The Female Vote

Just...The Worst