Finding Success at the Box Office

An Unsupervised Learning Project

Annie Stanley // (LHL Oct 17/22 Cohort)

Where We're Going & How We're Getting There

- **Introduction**
  - Who Am I?
  - How Did I Get Here?
  - What Am I Even Talking About?
  - Why Should You Care?
  - Hopefully Not Too Many Philosophical Questions
- **Walkthrough**
  - What Did I Do?
  - How Did I Do It?
- **Results**
  - How Can This Be Used?
  - What Did I Find?
- **What Next?**
  - Where Will I Go From Here?

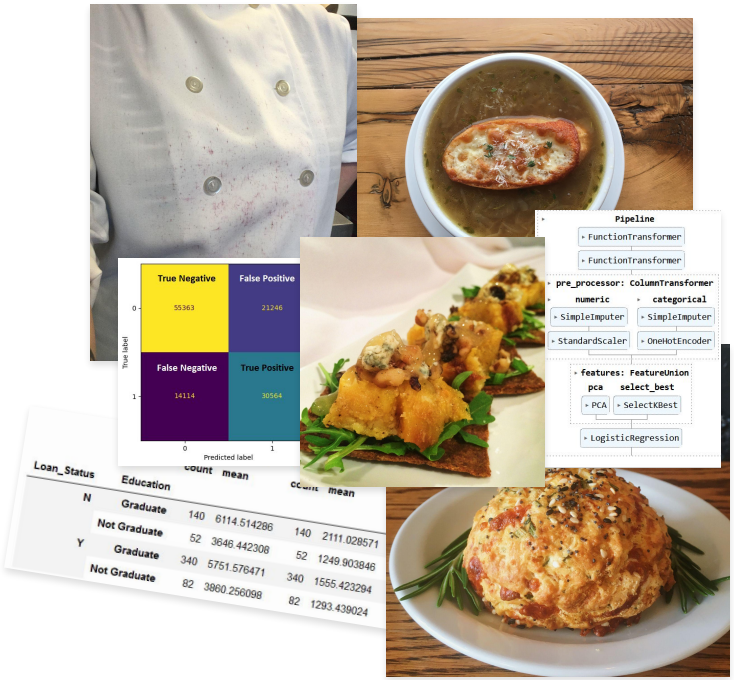
Introduction

```
print('Hello world…')
```

Hello!

From Farm to Table
To Data to ... Tables

- **Who Am I?**
  - Annie Stanley
  - Previous cook/baker/barista/payroll administrator
  - Future Data Analyst
- **How Did I Get Here?**
  - Too many years in kitchens
  - Incredibly interesting administrative work
  - A genuine & deep enjoyment of talking about graphs
- **What Am I Even Talking About?**
  - Movies!
  - What kinds make more money?
- **Why Should You Care?**
  - An industry of frequent change
  - Knowing what performs best can help with production & marketing decisions



Walkthrough

- Data Acquisition
- Cleaning
- EDA
- Modeling

Tech Stack

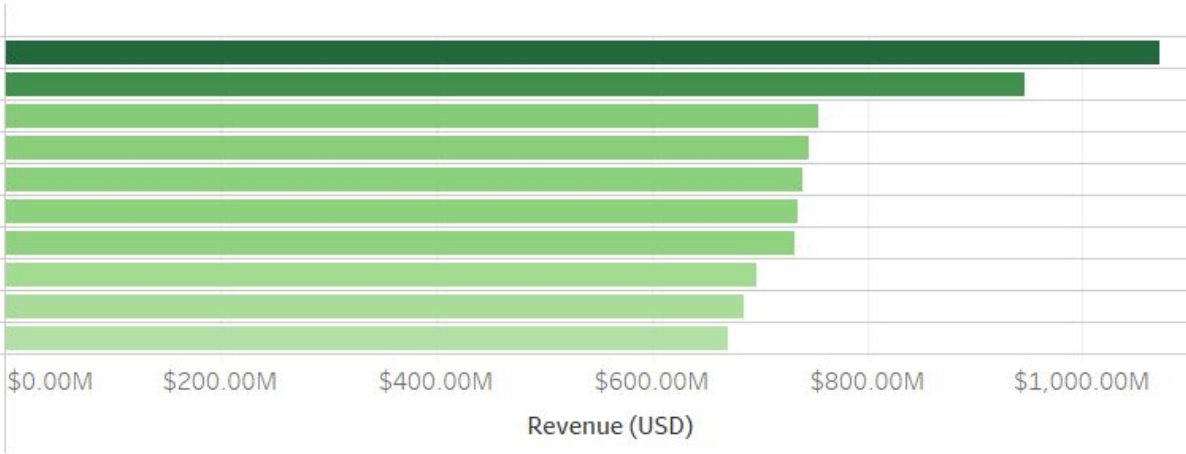
- **Python**
  - (Pandas, Matplotlib, Numpy, Seaborn, Scikit-Learn)
- **APIs & Web Scraping**
- **Tableau**

Acquisition & Cleaning

- MovieLens Dataset
- TMDB & OMDB APIs
- Box Office Mojo & Wikipedia Web Scraping
- Revenue Adjustments (Inflation)
- Genre & Actor Breakdowns
- Post-Streaming World (2007) Data Only

Top Domestic Box Office Revenue

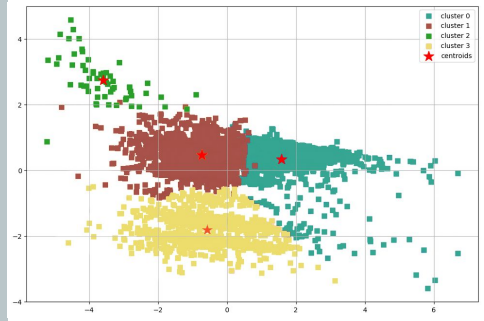
Title	Genre	Top Billed Actor
Star Wars: Episode ..	Action/Adventure	Daisy Ridley
Avatar	Action/Adventure	Sam Worthington
Black Panther	Action/Adventure	Chadwick Boseman
Jurassic World	Action/Adventure	Chris Pratt
Up	Comedy	Edward Asner
The Avengers	Action/Adventure	Robert Downey Jr.
Avengers: Infinity W..	Action/Adventure	Robert Downey Jr.
A Star Is Born	Drama	Lady Gaga
Star Wars: Episode ..	Action/Adventure	Daisy Ridley
The Dark Knight	Drama	Christian Bale



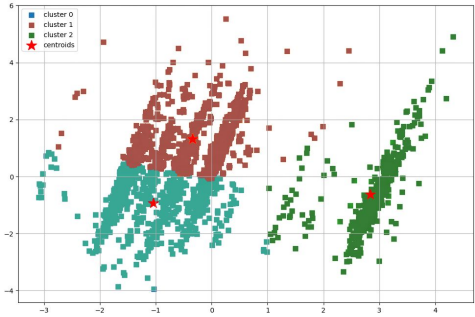
Modeling

- Clustered Genres
- Scaling (Robust Scaler)
- PCA
  - Component selection based on data loss
- K-Means & Hierarchical Clustering
  - Elbow graph
  - Silhouette score

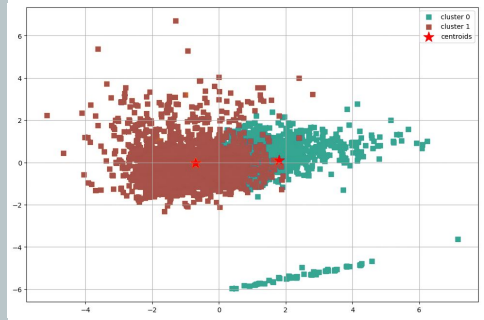
Drama Clusters



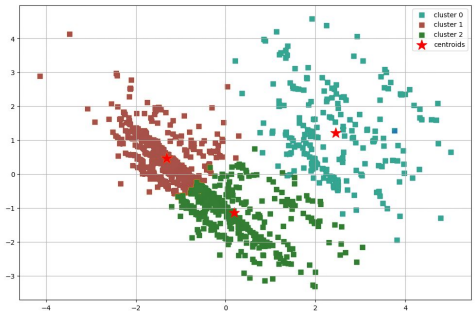
Action/Adventure Clusters



Comedy Clusters



Thriller/Horror Clusters



Modeling

- Clustered Genres
- Scaling (Robust Scaler)
- PCA
  - Component selection based on data loss
- K-Means & Hierarchical Clustering
  - Elbow graph
  - Silhouette score

Drama

Clustering Method:
- K-Means

Silhouette Score:
- 0.353

Data Loss:
- 25%

Cluster Sizes:
- C0: 1,474
- C1: 2,169
- C2: 59
- C3: 922

Action/Adventure

Clustering Method:
- K-Means

Silhouette Score:
- 0.392

Data Loss:
- 29%

Cluster Sizes:
- C0: 762
- C1: 709
- C2: 368

Comedy

Clustering Method:
- Hierarchical

Silhouette Score:
- 0.292

Data Loss:
- 23%

Cluster Sizes:
- C0: 787
- C1: 2,032

Thriller/Horror

Clustering Method:
- K-Means

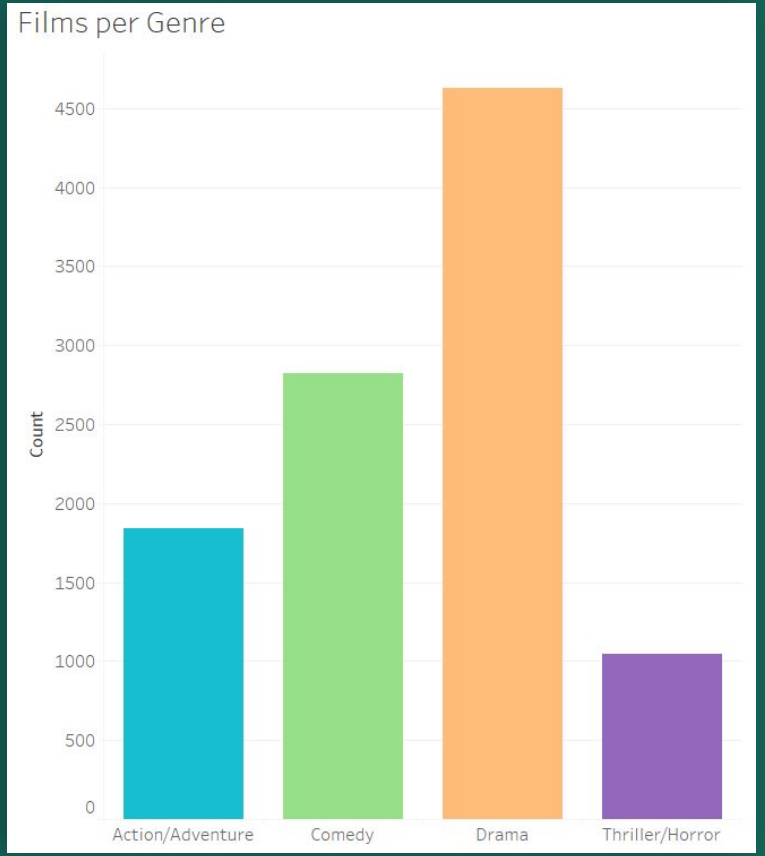
Silhouette Score:
- 0.284

Data Loss:
- 18%

Cluster Sizes:
- C0: 206
- C1: 443
- C2: 339

Usage Example/ Results

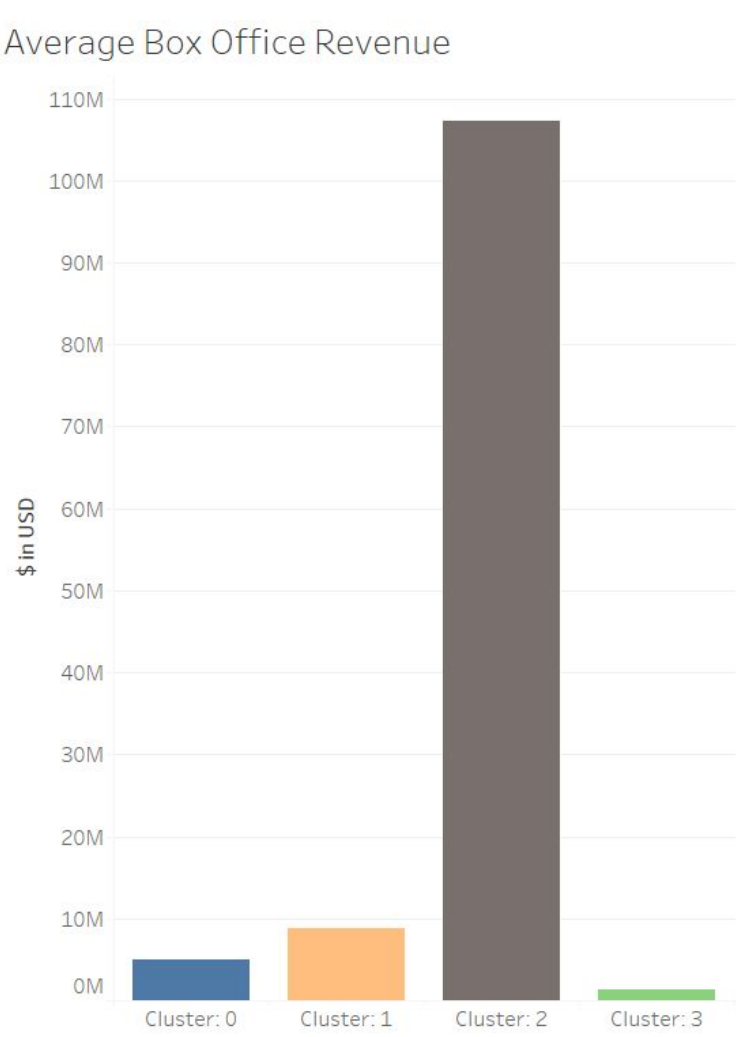
- Box Office Performance
- Critical/Audience Performance



Drama

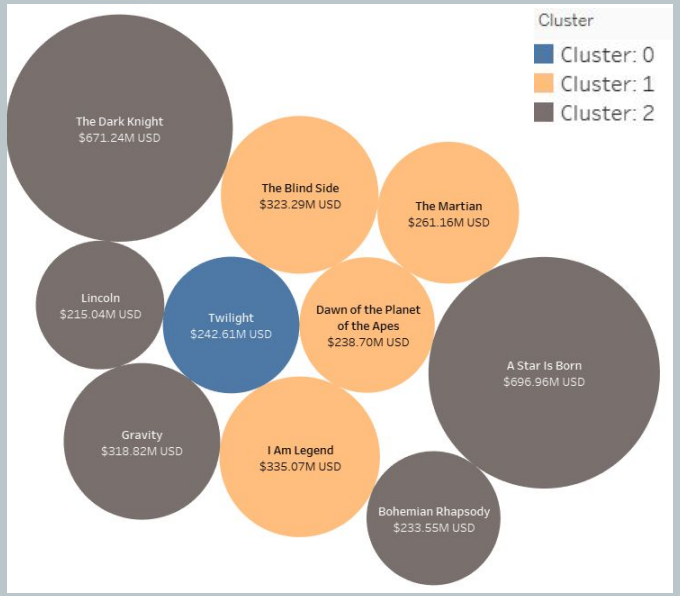
Drama

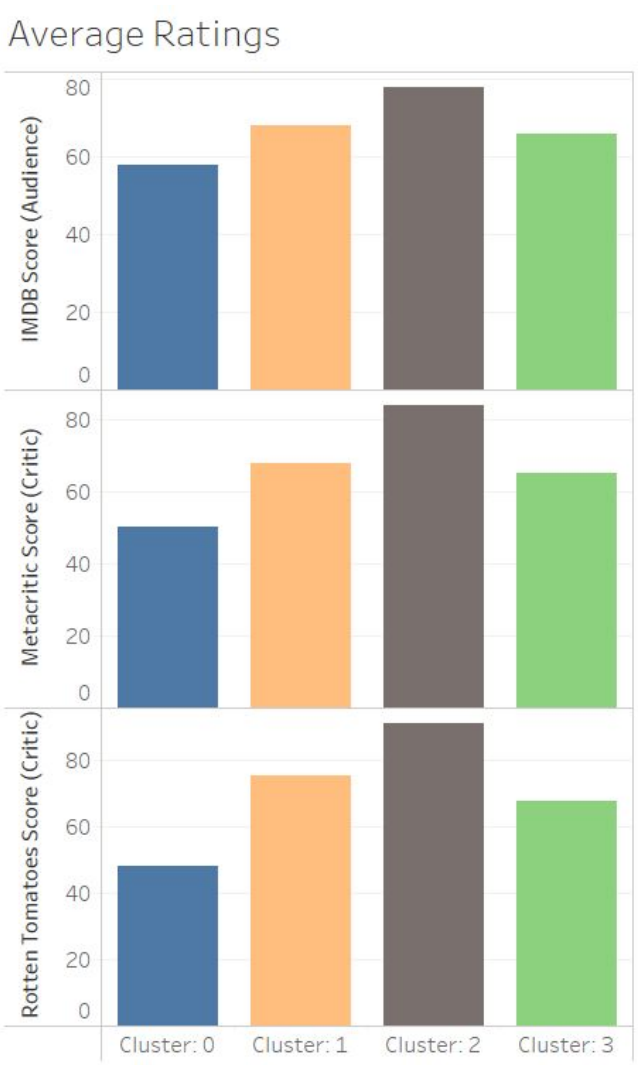
- **Avg Box Office Revenue:**
  - C0: $4.99M
  - C1: $8.74M
  - C2: $107.21M
  - C3: $1.26M
- **Clear Leader in Cluster 2**



Drama

- Cluster 2 = Leader
  - Strongest ratings across the board
  - Star power (Lady Gaga)
  - Franchise power (Batman)

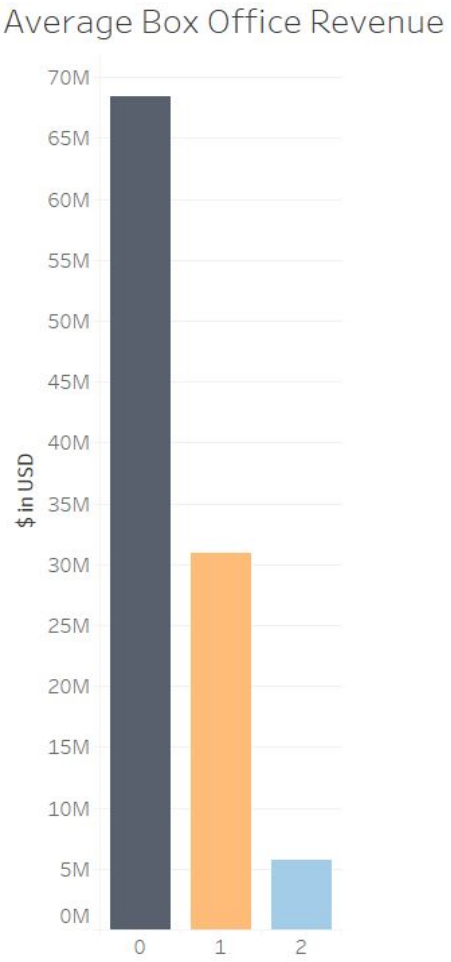




Action/Adventure

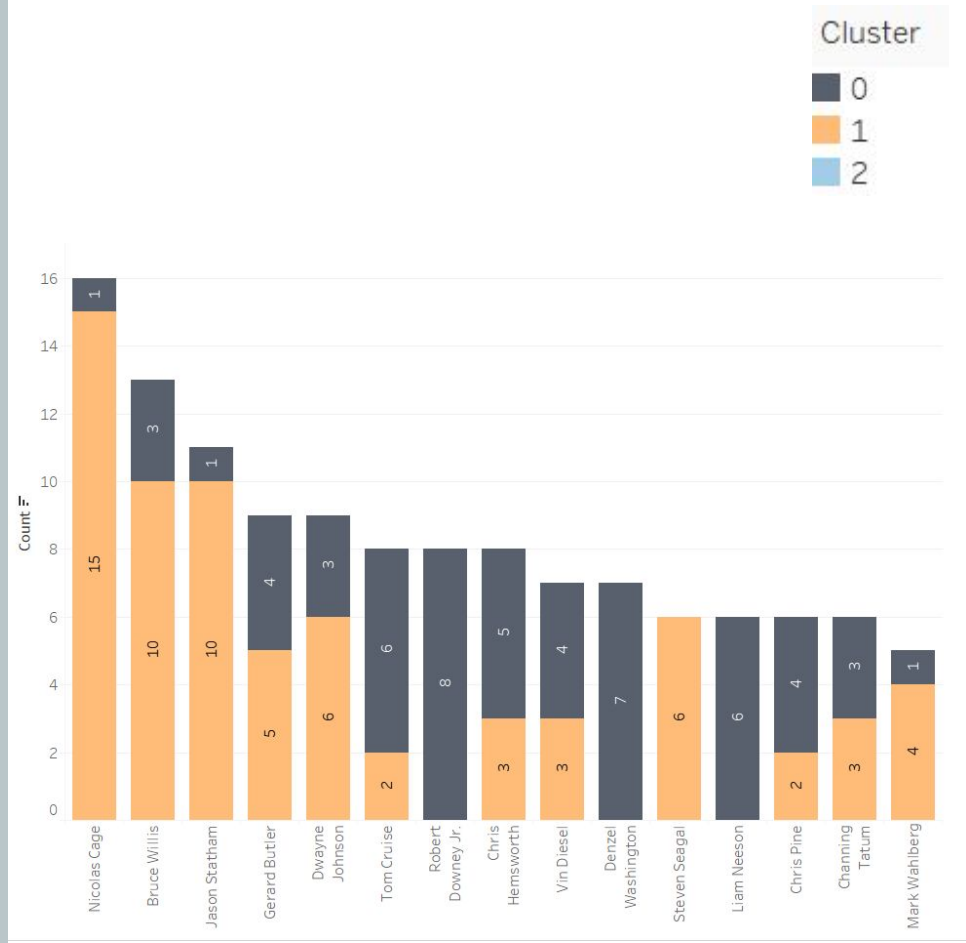
Action/Adventure

- **Avg Box Office Revenue:**
  - C0: $68.40M
  - C1: $30.88M
  - C2: $5.68M
- **Clear Leader in Cluster 0**
  - Strongest ratings across the board
  - Highest revenue



Action/Adventure

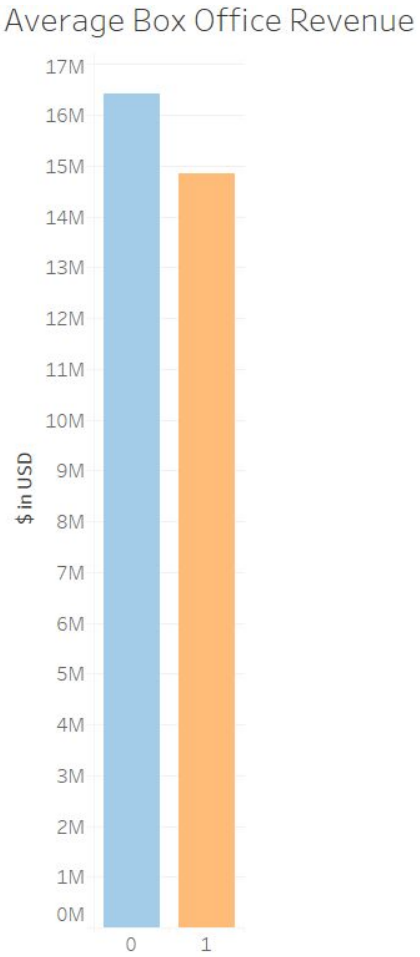
- Few (if any) big name/repeat actors in Cluster 2
- Cluster 0 has lots of big name actors in repeat roles
    - More dramatic actors
- Cluster 1 is focused more on straight up action

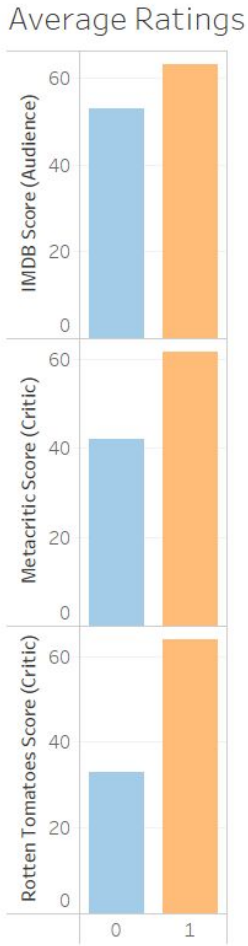


Comedy

Comedy

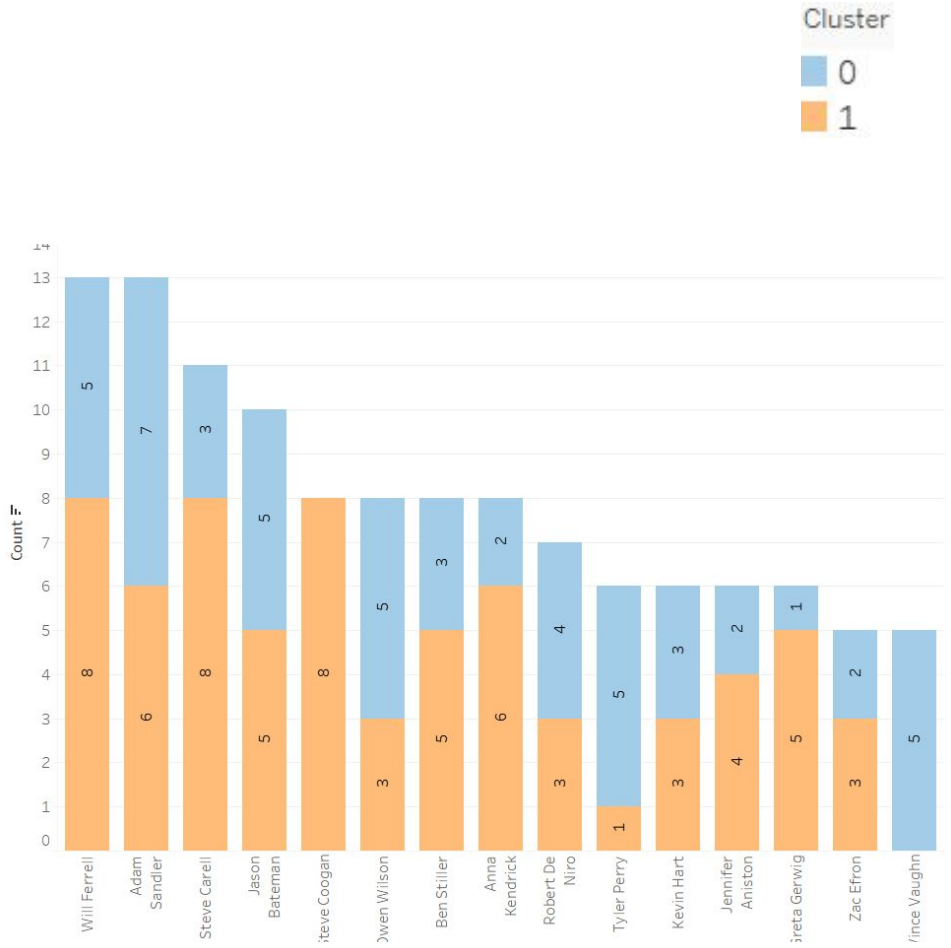
- **Avg Box Office Revenue:**
  - C0: $16.41M
  - C1: $14.84M
- **No clear ‘leader’**





Comedy

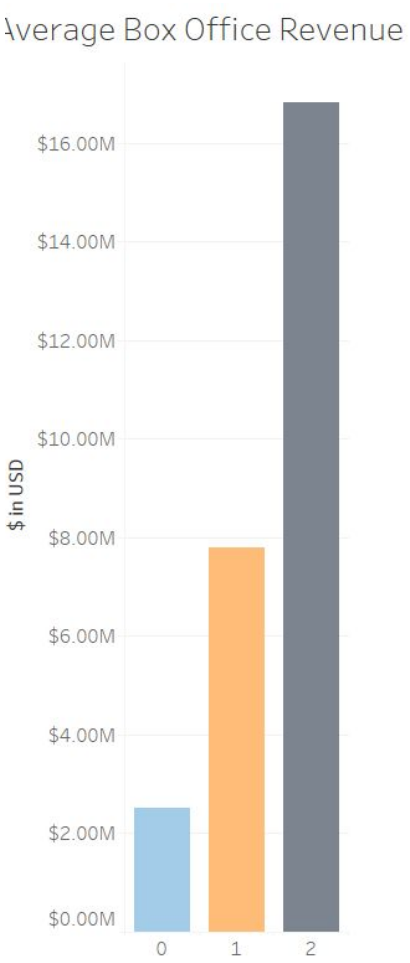
- Some actors = more $$$
- Some actors = better ratings
- Humour is incredibly subjective

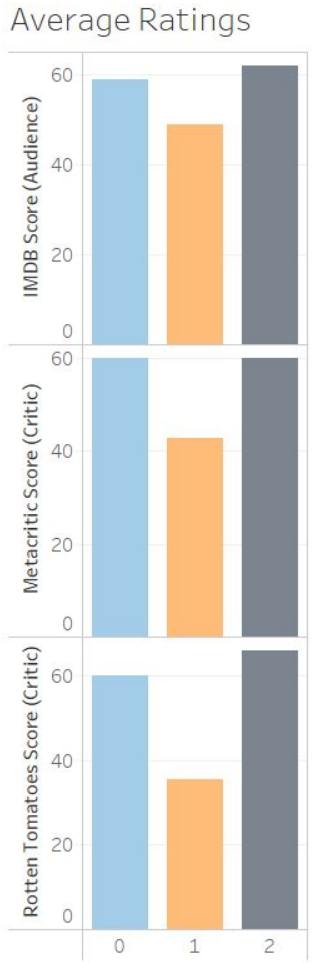


Thriller/Horror

Thriller/Horror

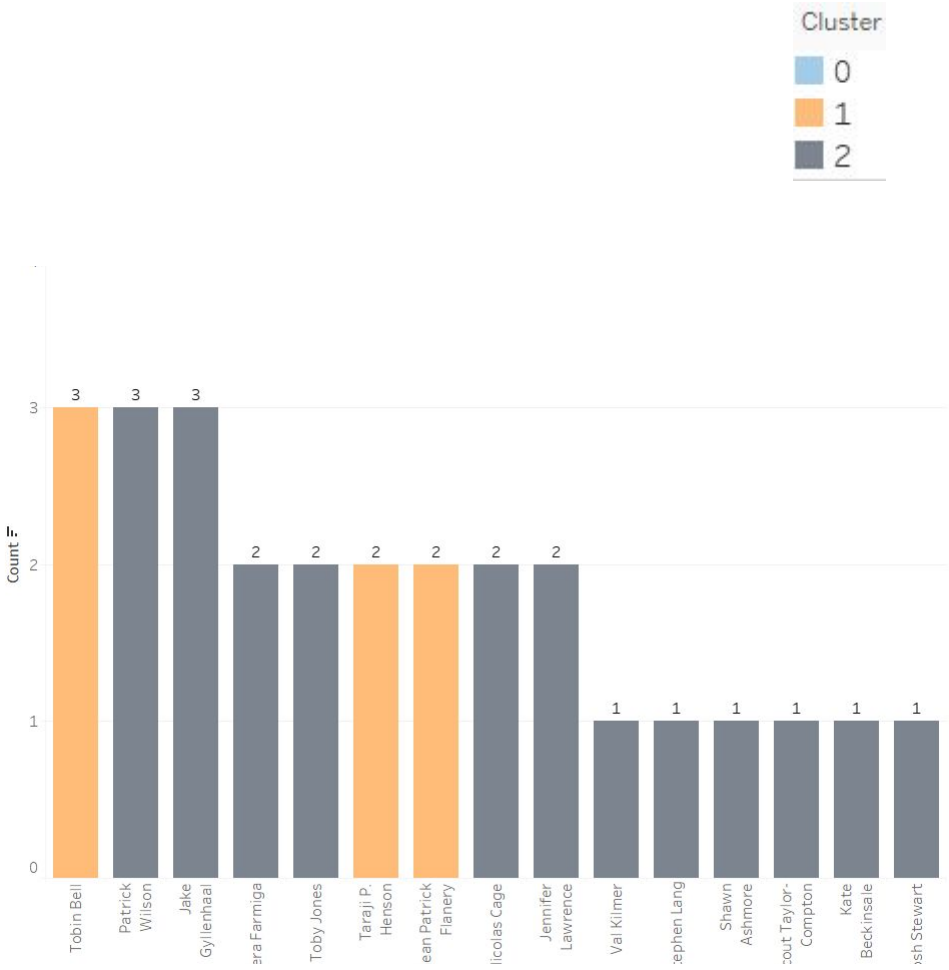
- **Avg Box Office Revenue:**
  - C0: $2.50M
  - C1: $7.79M
  - C2: $16.81M
- **Cluster 2 = Leader**





Thriller/Horror

- No repeat actors in Cluster 0
- Cluster 2 = most repeat actors
- Horror movies tend to be more niche
    - Fewer repeat big-name actors
- Thriller could have a cross-over with Drama



What's Next?

- Use a Larger Sample
- Include More Recent Data
  - Data only went to 2018
  - It would be interesting/useful to see pandemic effects
- Investigate Finer Details
  - Sub-Genre
  - Directors
  - Academy Awards & Nominations
  - Other Awards & Nominations
- NLP on Reviews
- Predictive Model?
  - Actors + genres + review keywords = [x] revenue and/or [y] # of awards/nominations
  - Director + actor + genre + sub-genre + plot keywords = [x] critic rating and/or [y] audience rating

Thank You

Find Me:

- **LinkedIn/GitHub/Creddle:**
  - http://annie-stanley.ca

Data Sources:

- **MovieLens:**
  - https://grouplens.org/datasets/movielens/
- **The Movie Database (TMDB):**
  - https://developers.themoviedb.org/3/getting-started/introduction
- **The Open Movie Database (OMDB):**
  - https://www.omdbapi.com/
- **Wikipedia - Academy Awards & Nominations:**
  - https://en.wikipedia.org/wiki/List_of_Academy_Award-winning_films
- **Box Office Mojo:**
  - https://www.boxofficemojo.com/