Analyzing Gender Gap In Hollywood Movies

by Okey Ofili

Using Unsupervised and Supervised Models to Understand Gender Roles

Problem Statement

Applying **unsupervised** and **supervised** models on movie roles **scraped** from Wikipedia and IMDB to analyze gender bias in hollywood movies.

6,606

HOLLYWOOD MOVIES

8,266

UNIQUE ACTORS/ACTRESSES

21,250

SCRAPED MOVIE ROLES/CHARACTERS

The Process

1

Scrapping and Merging Data

Methodology for gathering the Data and merging different datasets across IMDB and Wikipedia

2

Exploratory Analysis and Modeling

Inference and Predictive analysis on the Data set.

3

Observations, Issues and Next Steps

Discuss observations in the data, issues faced during the project execution stage and next steps

“Without data you are just another person with an opinion.”

– W. Edwards Demmings

The Scraping Process

IMDB

This was pre-scrapped from kaggle data-set for movies with over 100 votes released 1st Quarter of 2020.

NNDB

Data was scraped with beautiful-soup to get cast members race, but the data-set was too unbalanced so was not used again.

Wikipedia

Utilized Wikipedia API to scrape website and get cast roles. This tasks took a whooping 40 hours! Then wrote code to dissect the string returned into usable parts.

The StopWord Process for Names

NLTK

Standard NLTK package, I identified all Nouns and then removed names by cross-checking against wordnet.synsets

IBM Watson

This was the most accurate prediction of names of Characters and people's, but I hit a API limit in the middle of the process, so had to abandon.

SpacyIO

This was a hybrid between IBM and NLTK, and performed better. It basically is able to extract Named Entity Extraction and identify what is PERSON or CITY. And it was FREE.

1.72 x

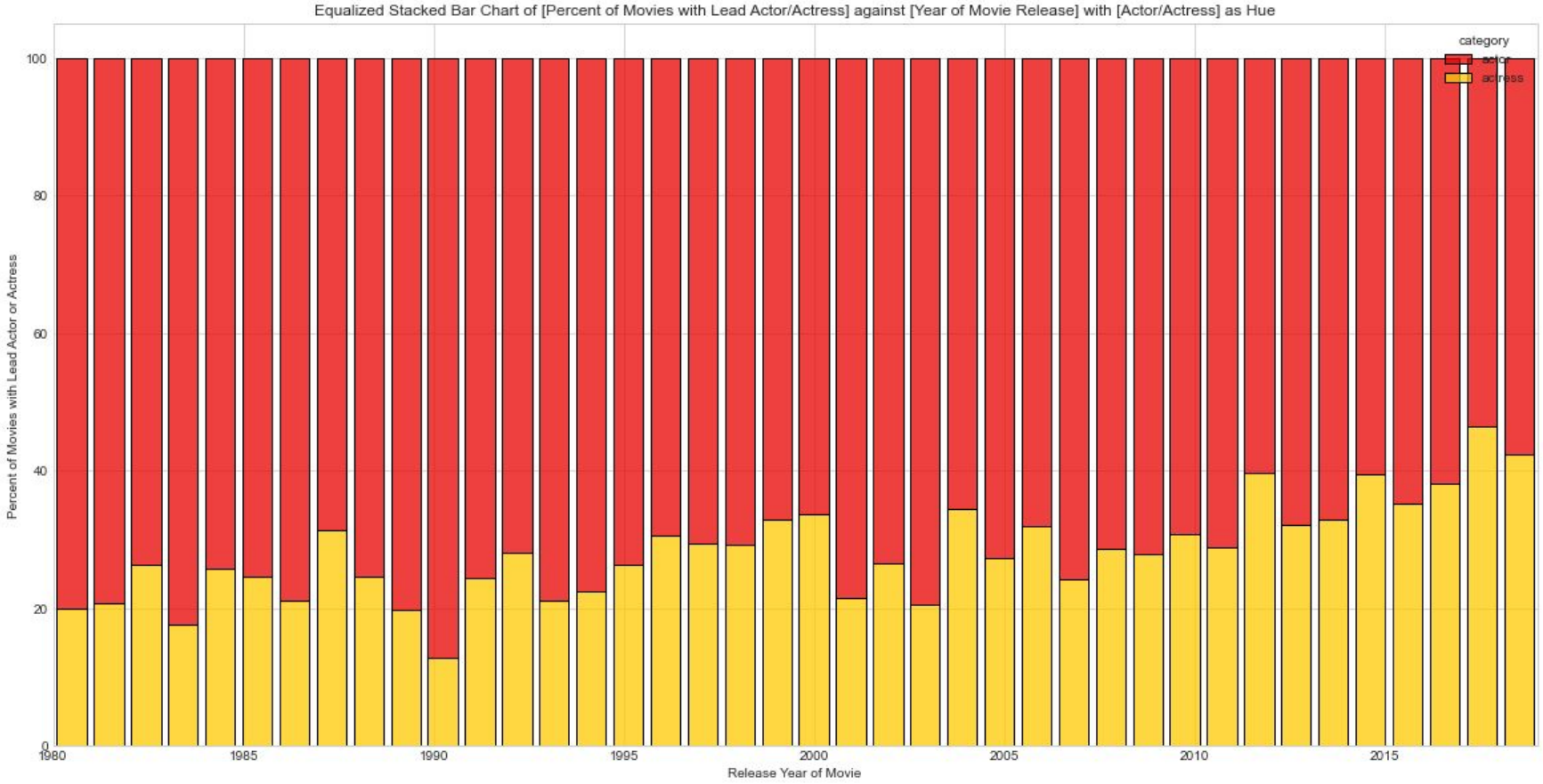
RATIO OF MALE ACTORS CHILDREN TO FEMALE ACTORS

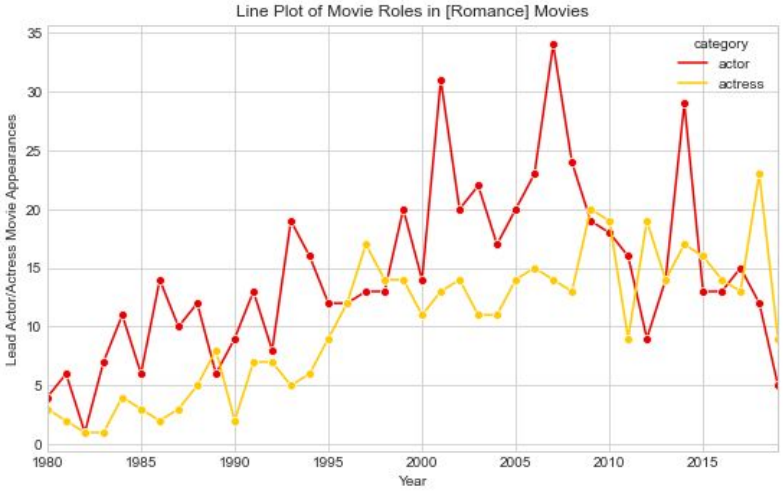
1.57 x

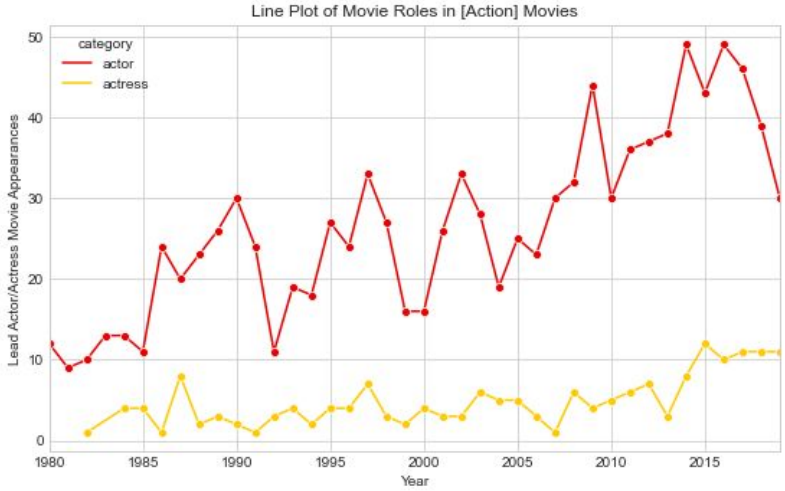
BUDGET RATIO OF MALE LED MOVIES TO FEMALE LED MOVIES

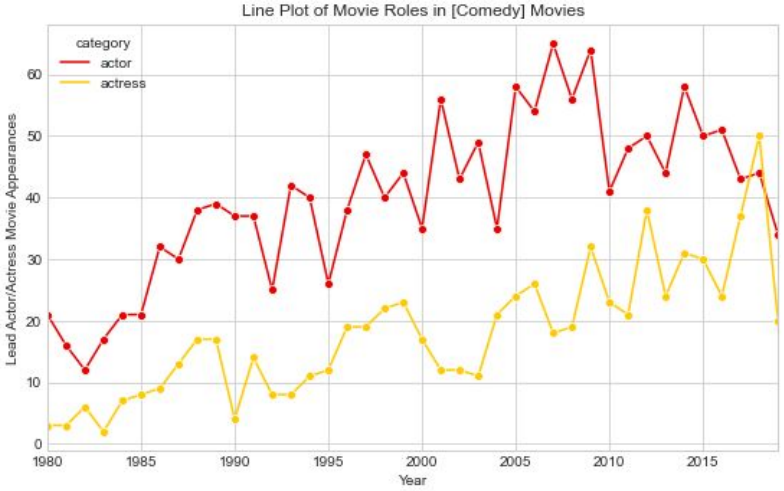
1.00 x

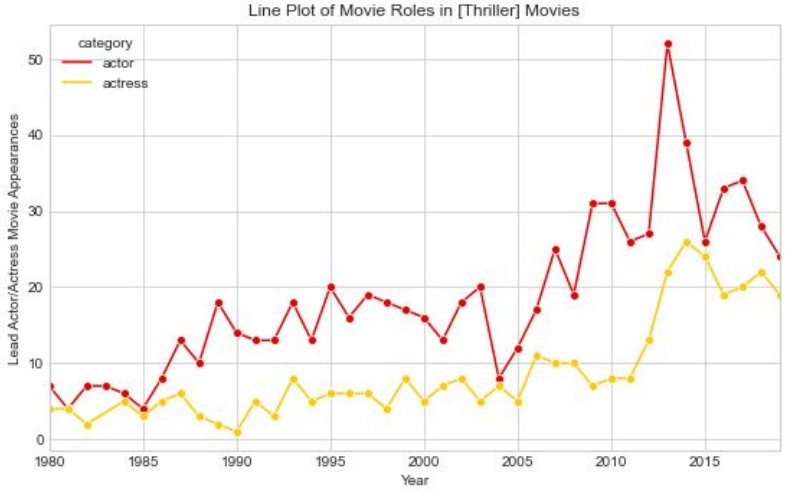
METASCORE RATIO OF MOVIES

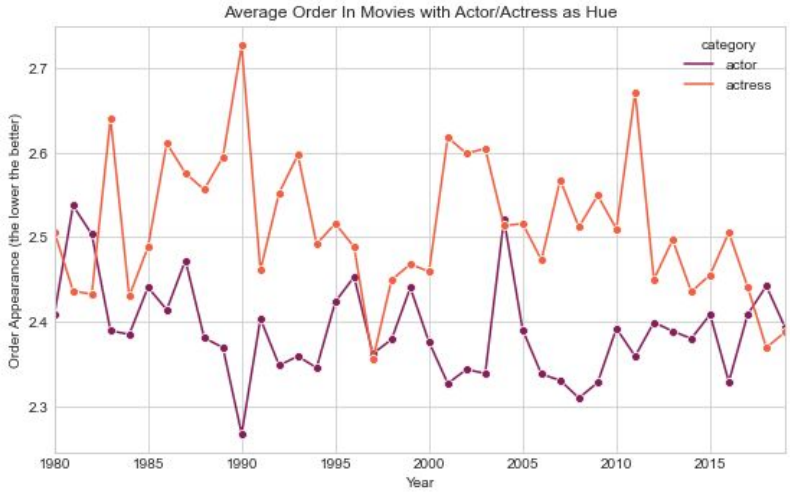

Equalized Stacked Bar Chart of [Percent of Movies with Lead Actor/Actress] against [Year of Movie Release] with [Actor/Actress] as Hue
category
actor
actress
Percent of Movies with Lead Actor or Actress
0
20
40
60
80
100
1980
1985
1990
1995
2000
2005
2010
2015
Release Year of Movie

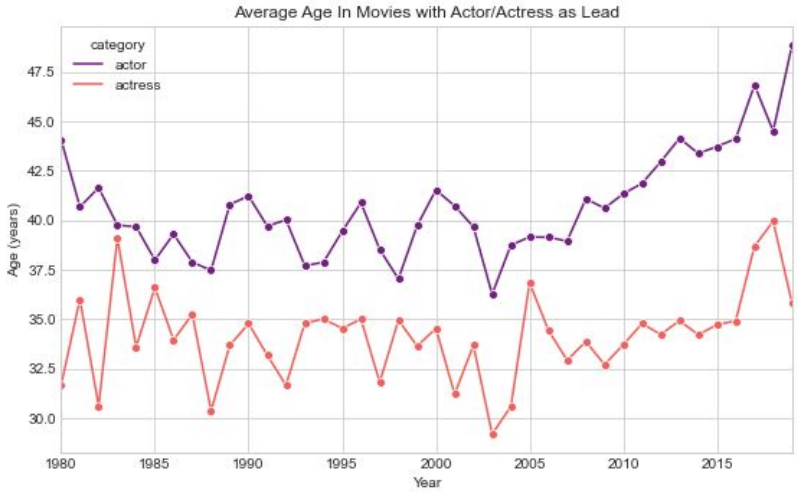

Line Plot of Movie Roles in [Romance] Movies
category
actor
actress
Lead Actor/Actress Movie Appearances
Year

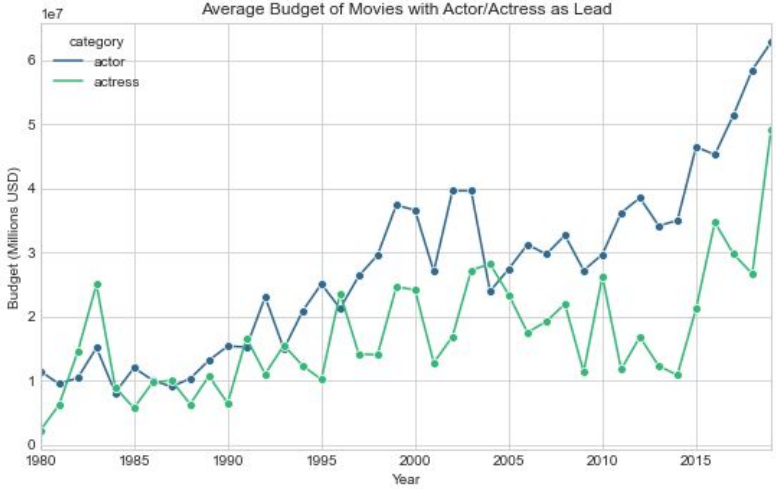

Line Plot of Movie Roles in [Action] Movies
category
actor
actress
Lead Actor/Actress Movie Appearances
Year

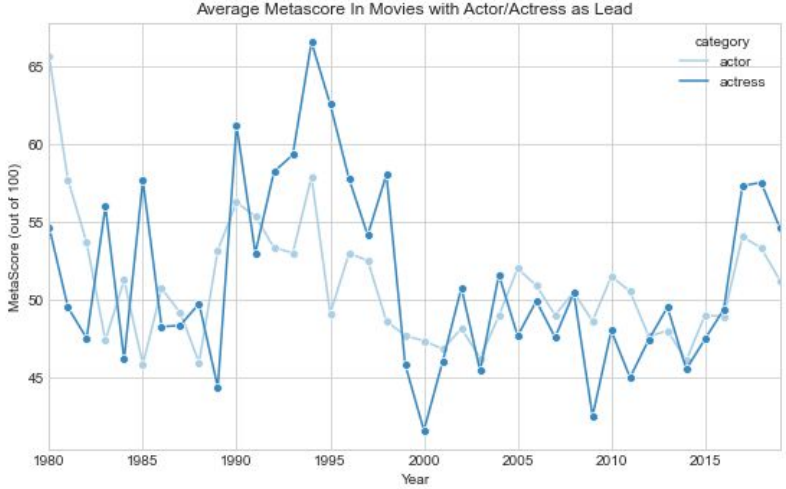

Line Plot of Movie Roles in [Comedy] Movies
category
actor
actress
Lead Actor/Actress Movie Appearances
Year

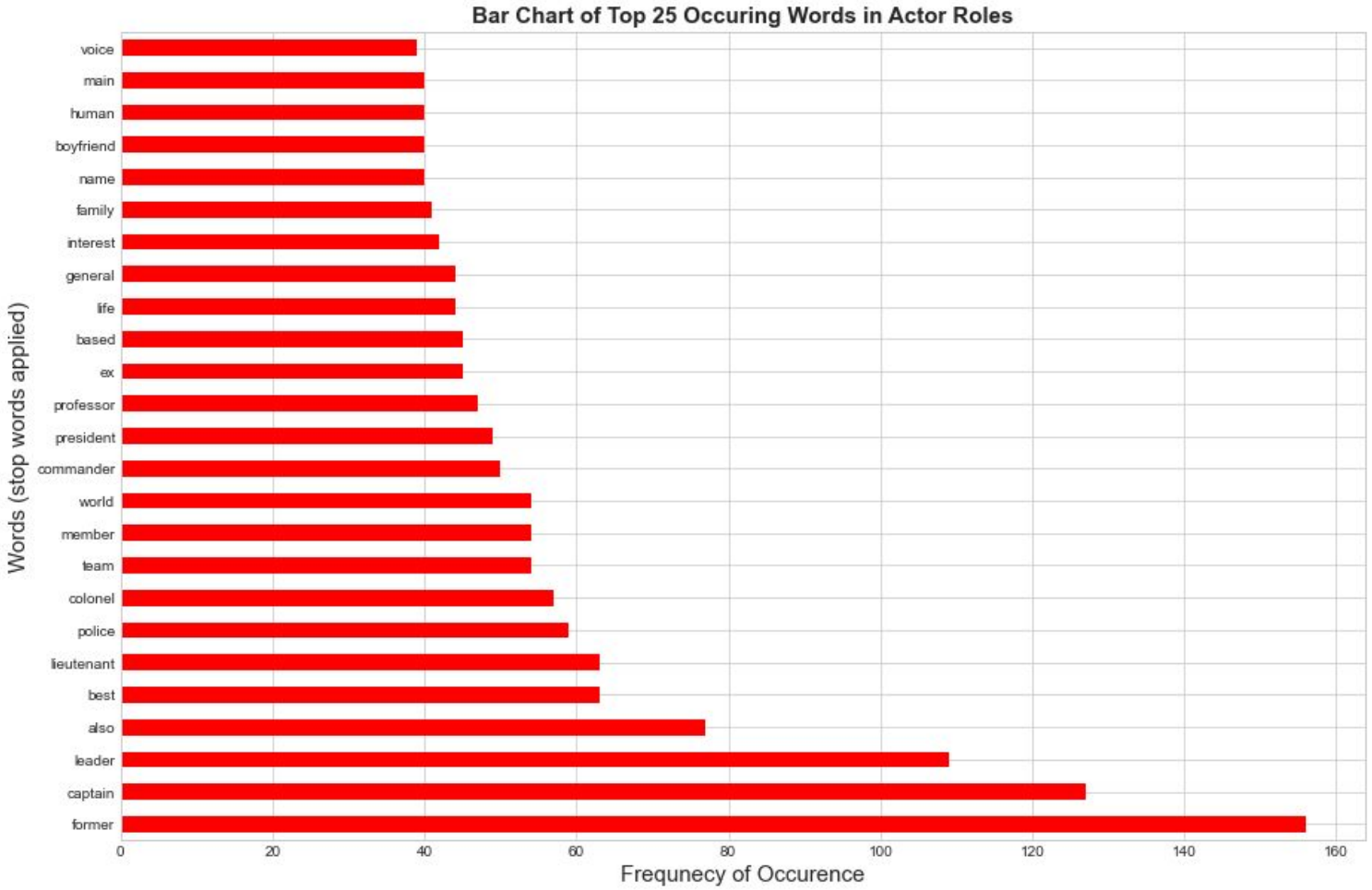

Line Plot of Movie Roles in [Thriller] Movies
category
actor
actress
Lead Actor/Actress Movie Appearances
Year

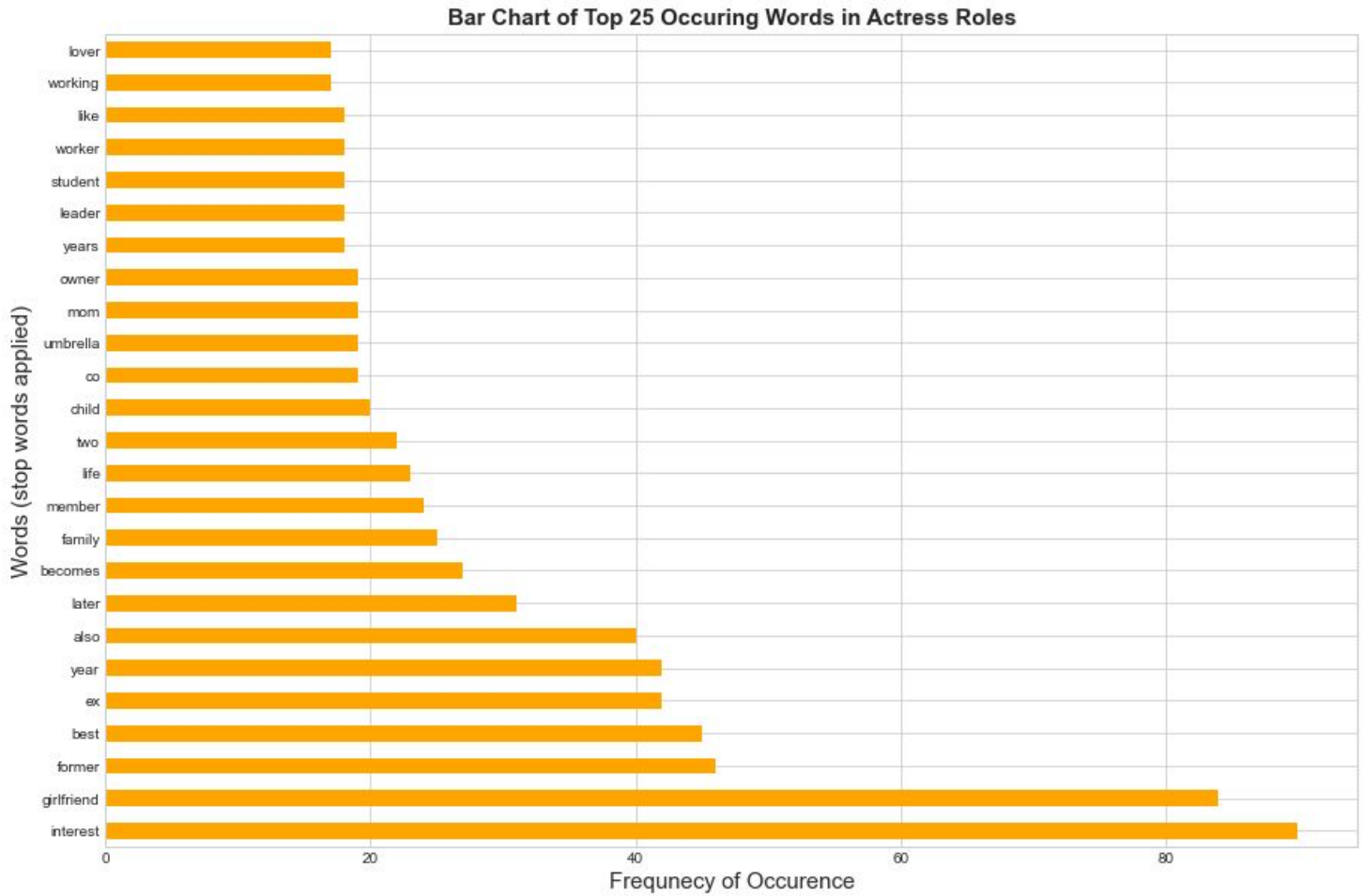

Average Order In Movies with Actor/Actress as Hue
category
actor
actress
Order Appearance (the lower the better)
Year

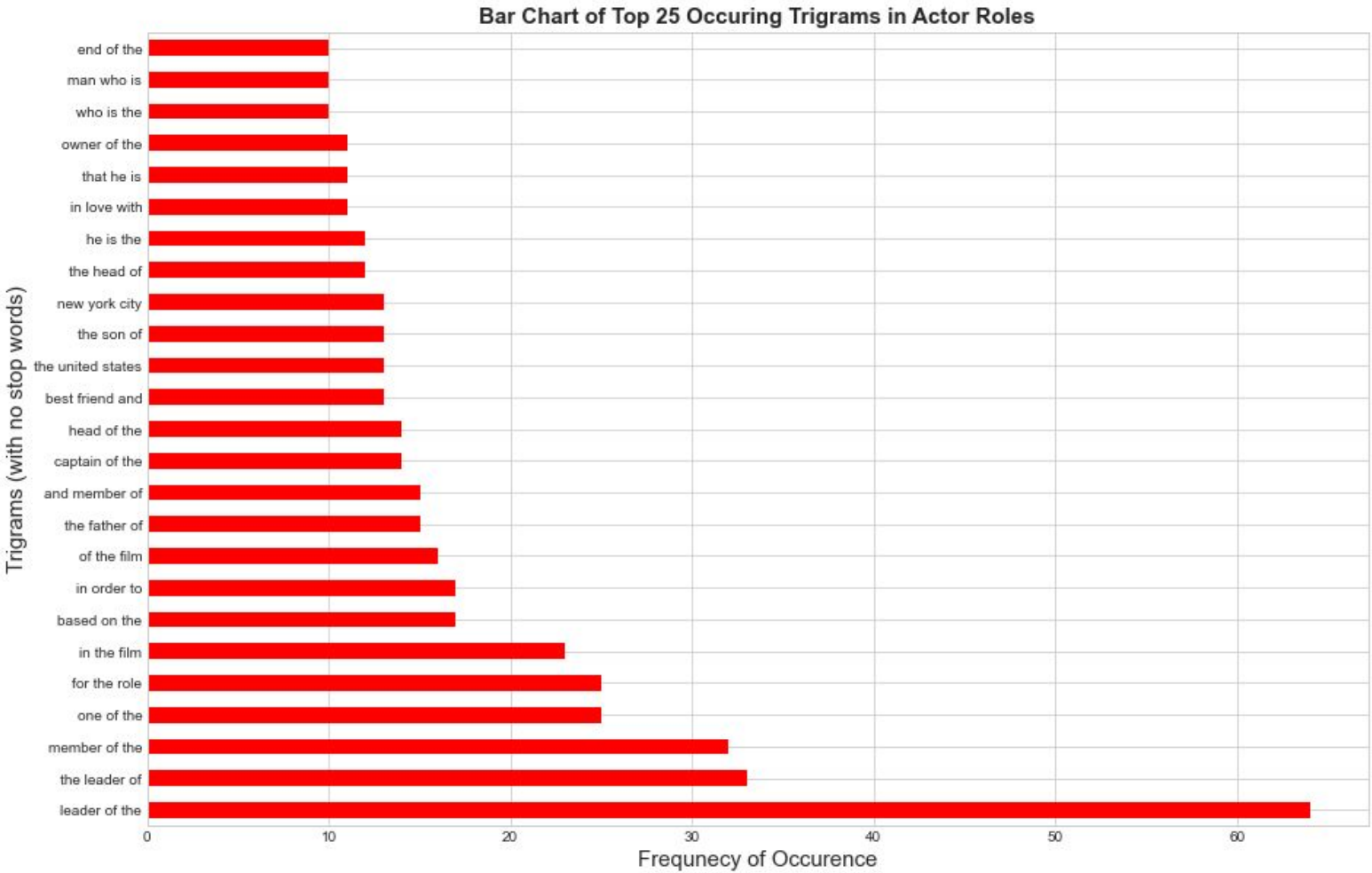

Average Age In Movies with Actor/Actress as Lead
category
actor
actress
Age (years)
Year

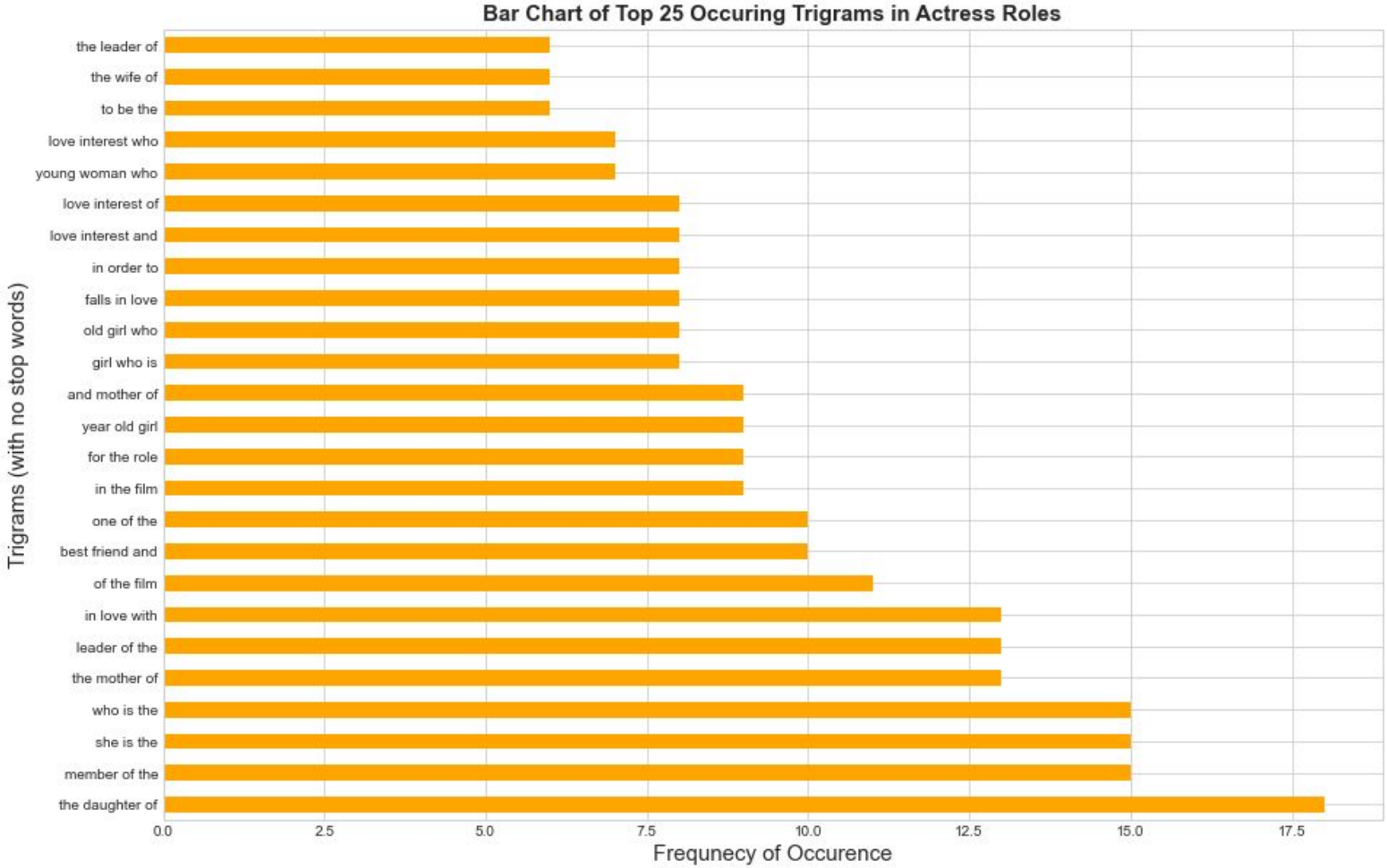

Average Budget of Movies with Actor/Actress as Lead
1e7
category
actor
actress
Budget (Millions USD)
Year


Average Metascore In Movies with Actor/Actress as Lead
category
actor
actress
MetaScore (out of 100)
Year


Bar Chart of Top 25 Occuring Words in Actor Roles
Words (stop words applied)
voice
main
human
boyfriend
name
family
interest
general
life
based
ex
professor
president
commander
world
member
team
colonel
police
lieutenant
best
also
leader
captain
former
0
20
40
60
80
100
120
140
160
Frequnecy of Occurence


Bar Chart of Top 25 Occuring Words in Actress Roles
Words (stop words applied)
lover
working
like
worker
student
leader
years
owner
mom
umbrella
co
child
two
life
member
family
becomes
later
also
year
ex
best
former
girlfriend
interest
0
20
40
60
80
Frequnecy of Occurence


Bar Chart of Top 25 Occuring Trigrams in Actor Roles
Trigrams (with no stop words)
end of the
man who is
who is the
owner of the
that he is
in love with
he is the
the head of
new york city
the son of
the united states
best friend and
head of the
captain of the
and member of
the father of
of the film
in order to
based on the
in the film
for the role
one of the
member of the
the leader of
leader of the
0
10
20
30
40
50
60
Frequnecy of Occurence


Bar Chart of Top 25 Occuring Trigrams in Actress Roles
Trigrams (with no stop words)
the leader of
the wife of
to be the
love interest who
young woman who
love interest of
love interest and
in order to
falls in love
old girl who
girl who is
and mother of
year old girl
for the role
in the film
one of the
best friend and
of the film
in love with
leader of the
the mother of
who is the
she is the
member of the
the daughter of
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Frequnecy of Occurence

Prediction of Gender Based on Role?

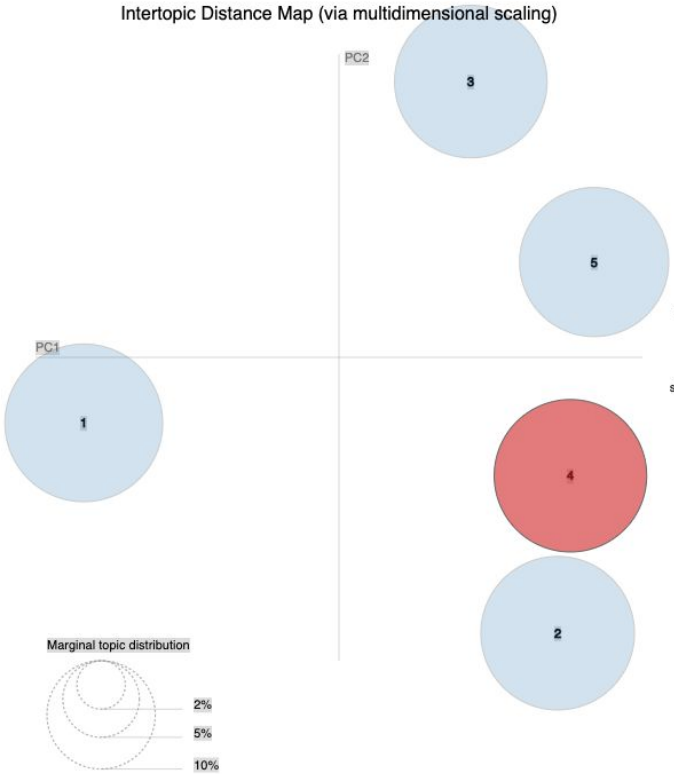
89%

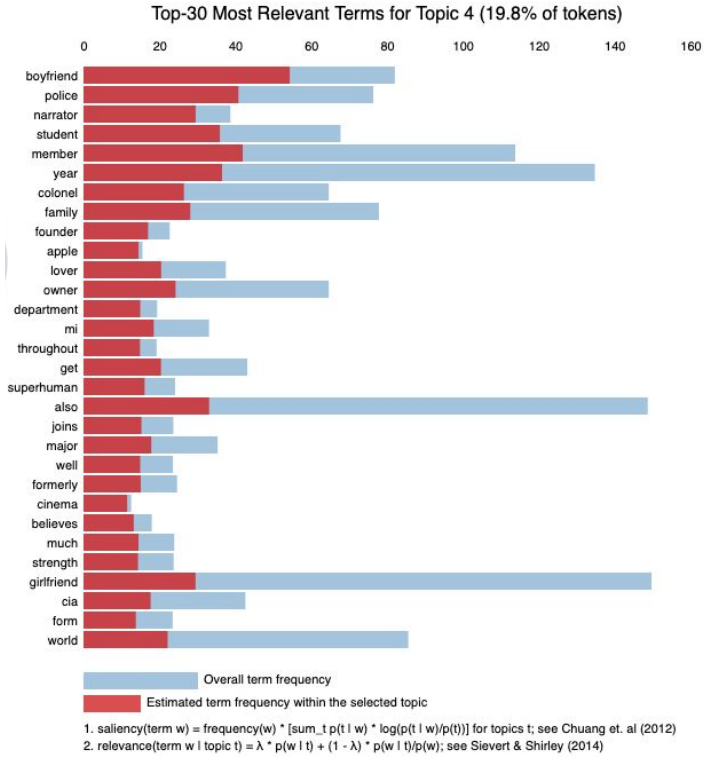
ACCURACY SCORE (NO STOPWORDS)

63.1%

ACCURACY SCORE (WITH STOP WORDS)

Latent Dirichlet Allocation





Issues

Too little text as the ROLES scrapped from wikipedia were mostly stub descriptions.

Syncing scraped data across different datasets leads to data loss. A better approach would be to **focus on one thing. Wikipedia or IMDB.**

Stopwords to remove names ended up removing important words like ‘Nurse’

Key Observations

Genre has a clear influence on roles on a Gender basis. For action movies there is a significant gap versus Romance movies.

Movies with male leads tends to have **1.5 times** more budgeted to it, even though average metacritic scores are same across genres.

Next Steps is to get richer Data on the role descriptions or leverage movie transcripts to infer this. Also to get cast race added to dataset and sell it!

Thanks

email **oaofili@gmail.com** for further questions