Proposal of
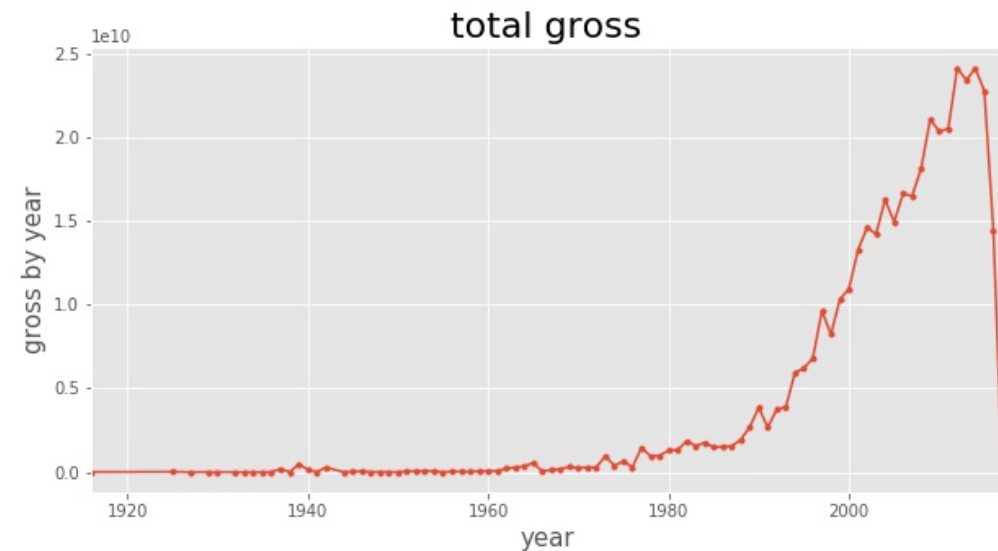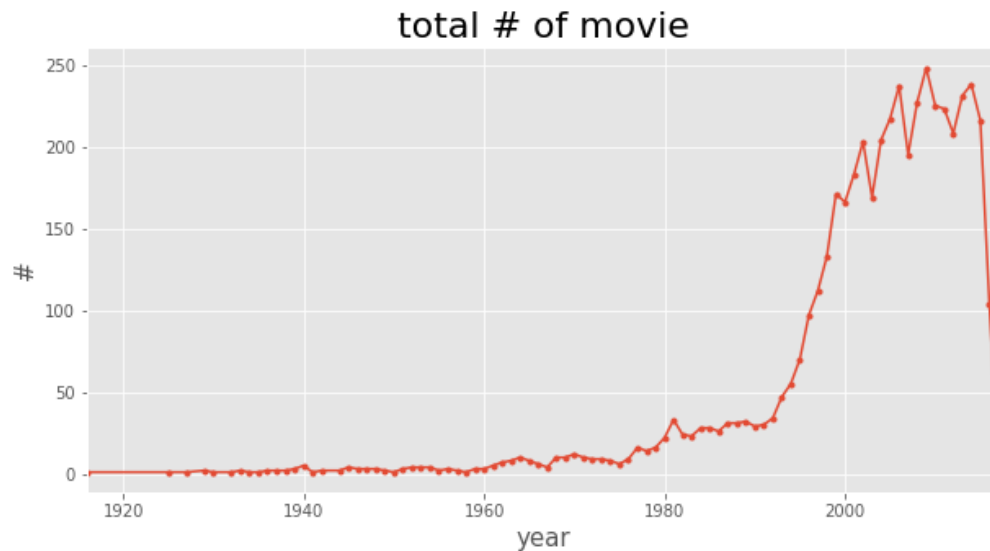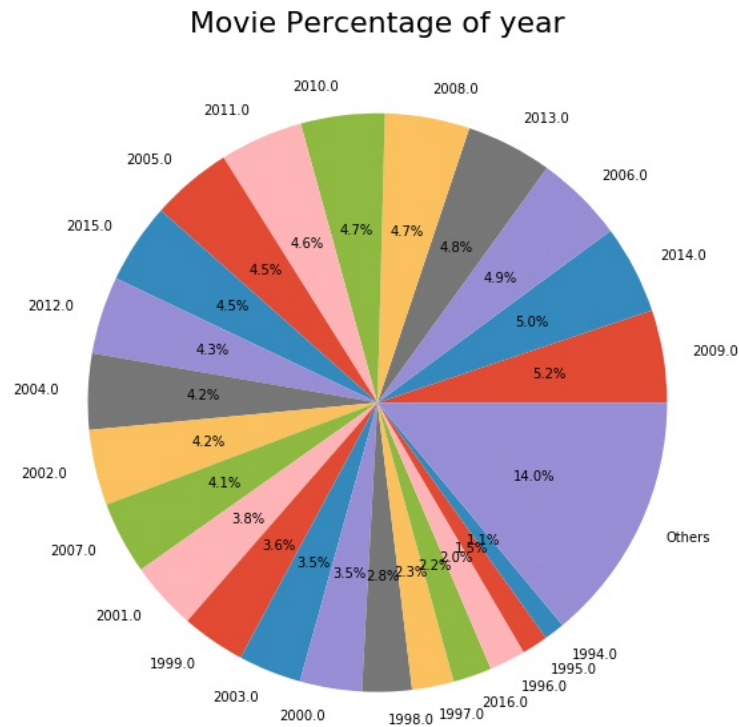# Analysis on TMDB data

Tyler Wang

Feb 26, 2020.

Introduction and Outlines

- Dataset(link):
  - TMDB tmdb_5000_credits.csv
    - [movie_id,title,cast,crew]
  - TMDB tmdb_5000_movies.csv
    - [budget,genre,shomepage,id,keywords,original_language,original_title,overview,popularity,production_companies,production_countries,release_date,revenue,runtime,spoken_languages,status,tagline,title,vote_average,vote_count]

- Insight into dataset,
  - See the distribution of gross and genre, showing situation of Movie market
  - Predict the relation between gross and vote / budget

- Target people/group
  - Audiences / Screenwriters / Directors / Production companies / investors
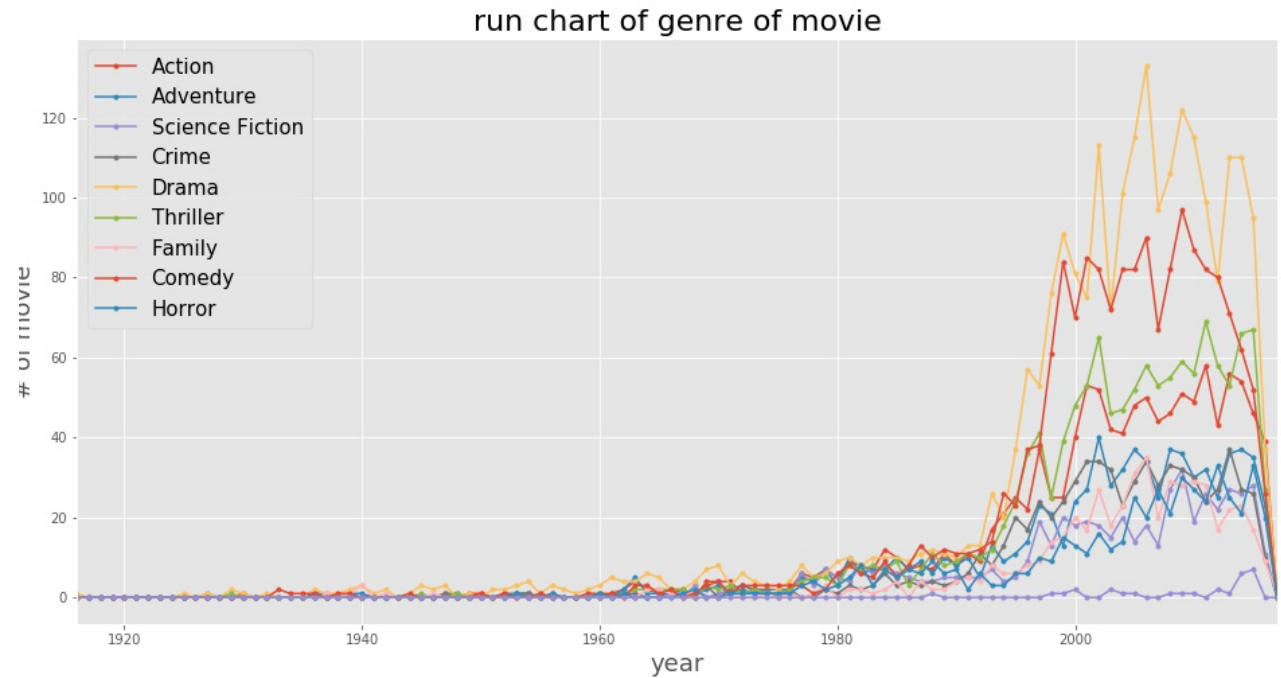
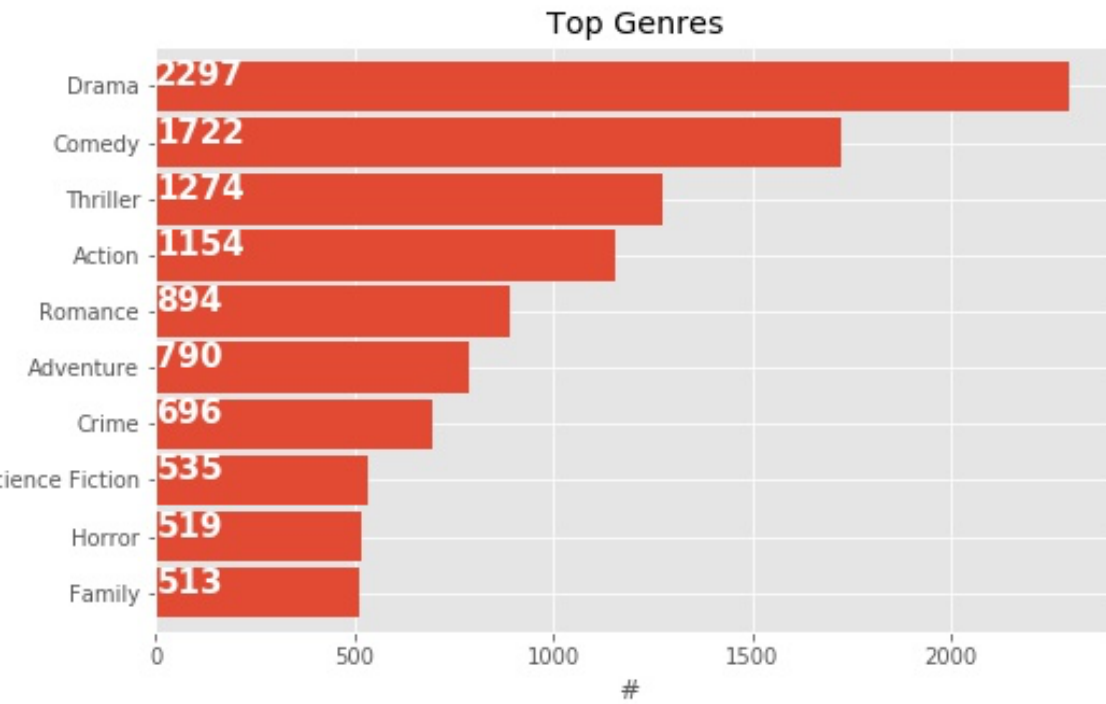# Total # of movie and gross of each year

Both total # of movie and total gross increase dramatically in the last decade, of 20th century. But # of movies increase slowly after 2000 and become stable after 2010, but the total gross doesn't because stable until the last 3 years. Maybe the ticket is more expensive.

Movie Percentage of year

- You may see that through the largest few sections, they are all around 4.7%.
- The movie market is become Saturate. It's just floating around.

What genre does people like?

Top Genres

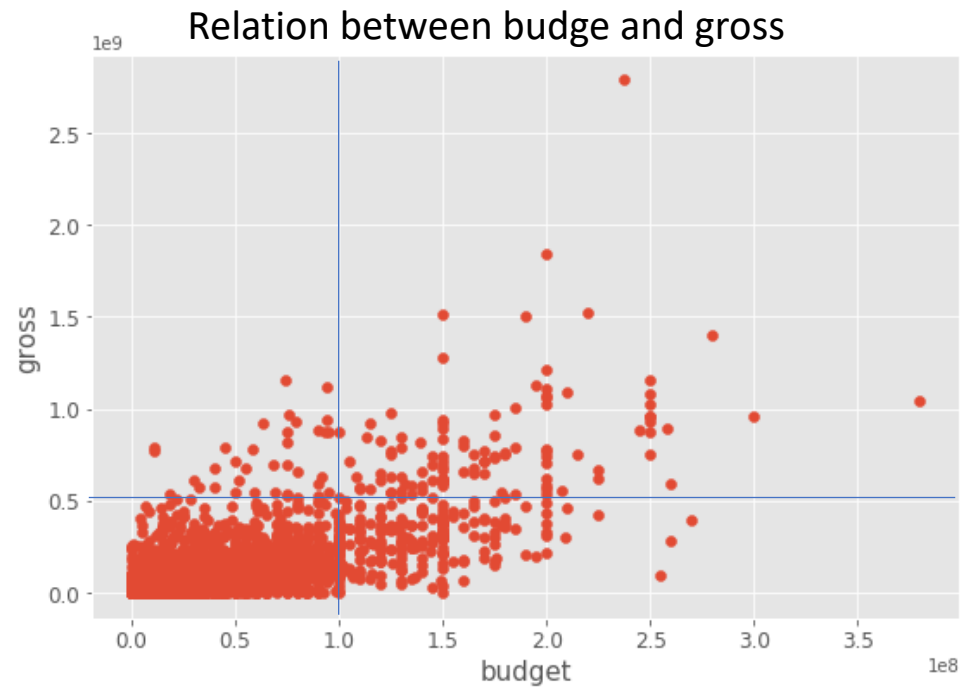| Genre	#
Drama	2297
Comedy	1722
Thriller	1274
Action	1154
Romance	894
Adventure	790
Crime	696
Science Fiction	535
Horror	519
Family	513

run chart of genre of movie

Clearly, Drama and comedy movies have the top number among all genres. Even thought the number becomes stable in the resent several years. But they still are the most popular genres.
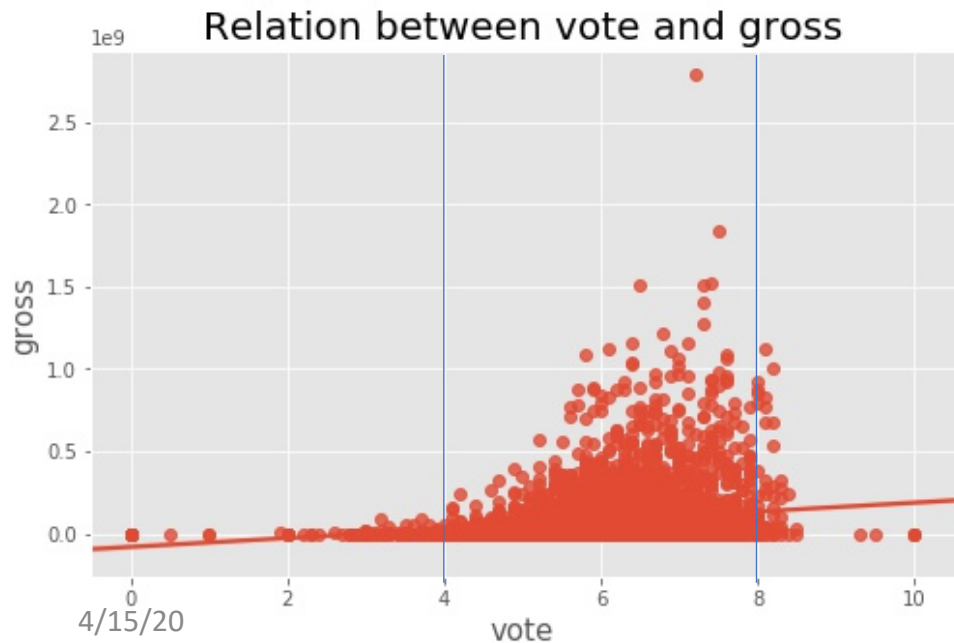
Comedy and Drama movies are Popular now!

Relation between budge and gross

- Roughly speaking, the more budget the movie had, the more gross it can have, but it's not guaranteed. We can also see that budget of most movie is lower than $100M and gross is lower than $50M

Relation between vote and gross

- We can also see that the movie with higher gross always has high vote. Most vote is between 4~8

Plans and proposals

- We can analyze the average ratio gross/budget of of each year, country, genre see which one has highest reward

- We can also analyze if some actor/directors/screen writer will improve the gross/budget

- We can also analyze what's the favorite genre/actor/director of each country/language.

- Then we can clean up the outliers, select features and use the data to predict the gross/budget of a movie, before it's on screen

- We can scrape data form IMDB and other movie rating website, in order to work on anything more.

- We can also try to use NLP to analyze the comments of vote and un-vote of movies. To evaluate the relation between votes and comments.

- We can also use the same alg. for the comments before the movies on screen, in order to predict the gross.