Final Analysis

First dataset:

- https://www.imdb.com/interfaces/

title.basics.tsv.gz - Contains the following information for titles:

- tconst (string) - alphanumeric unique identifier of the title
- titleType (string) – the type/format of the title (e.g. movie, short, tvseries, tvepisode, video, etc)
- primaryTitle (string) – the more popular title / the title used by the filmmakers on promotional materials at the point of release
- originalTitle (string) - original title, in the original language
- isAdult (boolean) - 0: non-adult title; 1: adult title
- startYear (YYYY) – represents the release year of a title. In the case of TV Series, it is the series start year
- endYear (YYYY) – TV Series end year. '\N' for all other title types
- runtimeMinutes – primary runtime of the title, in minutes
- genres (string array) – includes up to three genres associated with the title

Second dataset:

- https://www.kaggle.com/parsonsandrew1/nytimes-article-lead-paragraphs-18512017/data

NYTimes Article Lead Paragraphs 1851-2017

This tab-delimited file contains 8,013,073 dated lead paragraphs from articles published by The New York Times between September 1851 and July 2017.

Dataset minuses: *UPDATE: I've just learned that 1905-1930 are missing; I'm working to determine if this is my error or NYTimes'.*

We have movies data from 1890s, so we decided to use our history data starting from that period too. But because of lack of historical data, we are starting from 1930s.

Idea:

We want to analyse and maybe find a connection between movie genres coming out in cinemas and historical events.

For example, a possible correlation that we might see is that if there is a major historical event (a war, or 9/11), we might see films related to it a few years later.

Also, there is an interesting fact that when there was a revolution in the USSR in 1991 that led to changing USSR to Russia, all TV channels in Moscow showed "Swan Lake" (a ballet) during the revolution. So, we want to check if there is a connection between movies in the cinema and history.

Analysis Questions:

Our analysis consists of 3 parts:

- history analysis
- movie analysis
- correlation

For historical analysis:

We are building sentimental analysis of historical data. In this data we have: date and lead paragraph. We are building sentimental analysis of lead paragraph, so that we can see the positive and negative title for each year.

For film analysis:

We are counting the number for each genre for every year.

To find the correlation between genre production and event:

We are building multiple graphs over several decades.

Hypothesis:

Shortly after major historical events (wars, 9/11, etc.) we will be able to see a shift in genres as movies come out about those subjects.

We also may find a correlation when there are bad events happening, there are

Sentimental analysis news of 1930s:

```
neg          24686.0030
neu         491968.0250
pos          40662.6020
compound     84013.1581
dtype: float64
```

20 first lines of genres for 1930s ordered by count:

genres	count
Short	14642
Drama	13991
Comedy	13059
Documentary	6319
Romance	4304
Animation	3781
Musical	2995
Crime	2505
Music	2170
Action	2005
Family	1943
Adventure	1907
Western	1221
Mystery	1054
Sport	928
War	791
History	494
Fantasy	327
Thriller	315
Horror	259