



# Filmline

10.27.2017

Chanchal Ravindra Kariwala, U1134986

Sheetal Krishna Mohanadas Janaki, U1135144

[GitHub Repository](#)

## Overview

### Participants

- Chanchal Ravindra Kariwala, u1134986, u1134986@utah.edu
- Sheetal Krishna Mohanadas Janaki, u1135144, u1135144@utah.edu

GitHub Repository - <https://github.com/chanchalkariwala/dataviscourse-pr-filmline>

## Background and Motivation

We started brainstorming about various project ideas and soon realized that every time we came close to finding an idea interesting, we couldn't gather appropriate data for it. For example, our first idea was on analyse how a country's tourism may be affected by aspects like passport strength - visa requirements, crime against tourists, GDP spent on tourism. The idea stemmed from [Passport Index](#).

Post this, we changed tactics and started looking for datasets that we found interesting. This led us to a ton of websites that provide data for a plethora of topics and we narrowed down to two topics, the one we have currently picked and another - a survey of young people ranking a variety of topics from 1-5, ranging from music genres to fear of spiders, alcohol addiction etc.

We decided to proceed with the movie database since it gives us plenty of scope to visualize data in various ways, and to draw relationships and contrast between a lot of facets of the movie industry.

## Project Objectives

Data we plan on visualizing:

1. Genre, language and overall popularity
2. Comparisons between:
  - a. Popular genre and year (To see if there's a trend to popular genres)
  - b. Language and popularity
  - c. Proportion of movies that fall under each genre
  - d. Production companies and revenue
  - e. Budget and revenue
  - f. Countries that produce the most number/most popular movies
3. Overlapping genres

There are many more comparisons and interesting relationships that can be drawn from the dataset.

## Data source and Processing

We retrieved our data from <https://www.kaggle.com/>, and plan to use a combination of a couple of datasets to implement the project.

Links: <https://www.kaggle.com/tmdb/tmdb-movie-metadata>

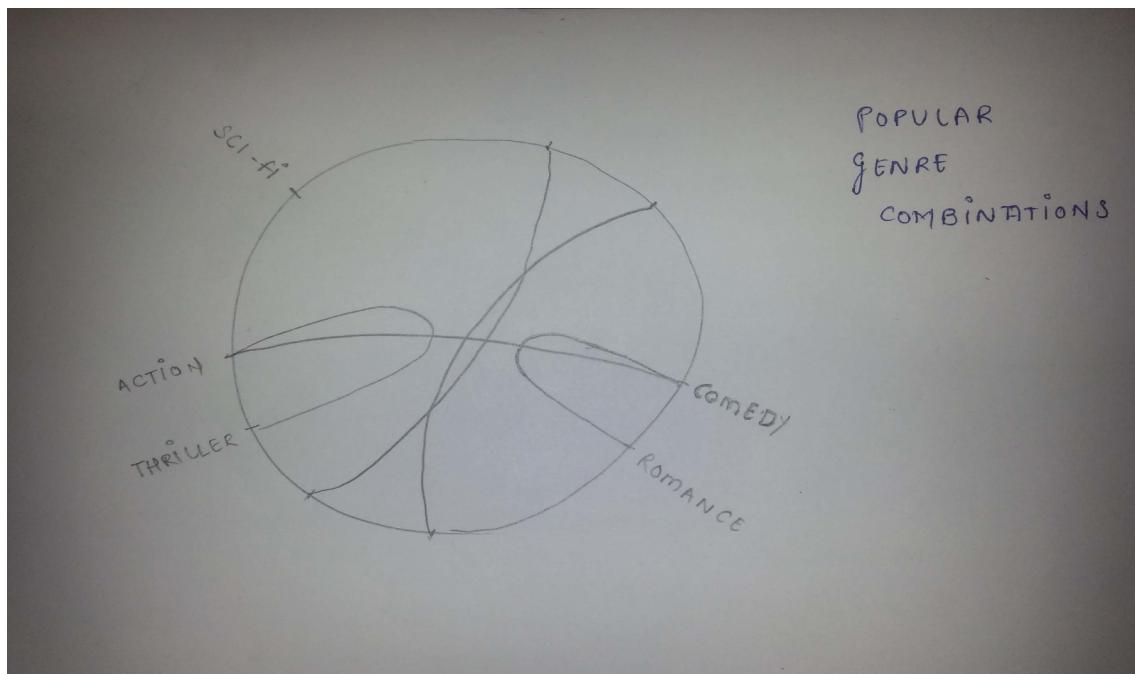
<https://www.kaggle.com/theacademy/academy-awards>

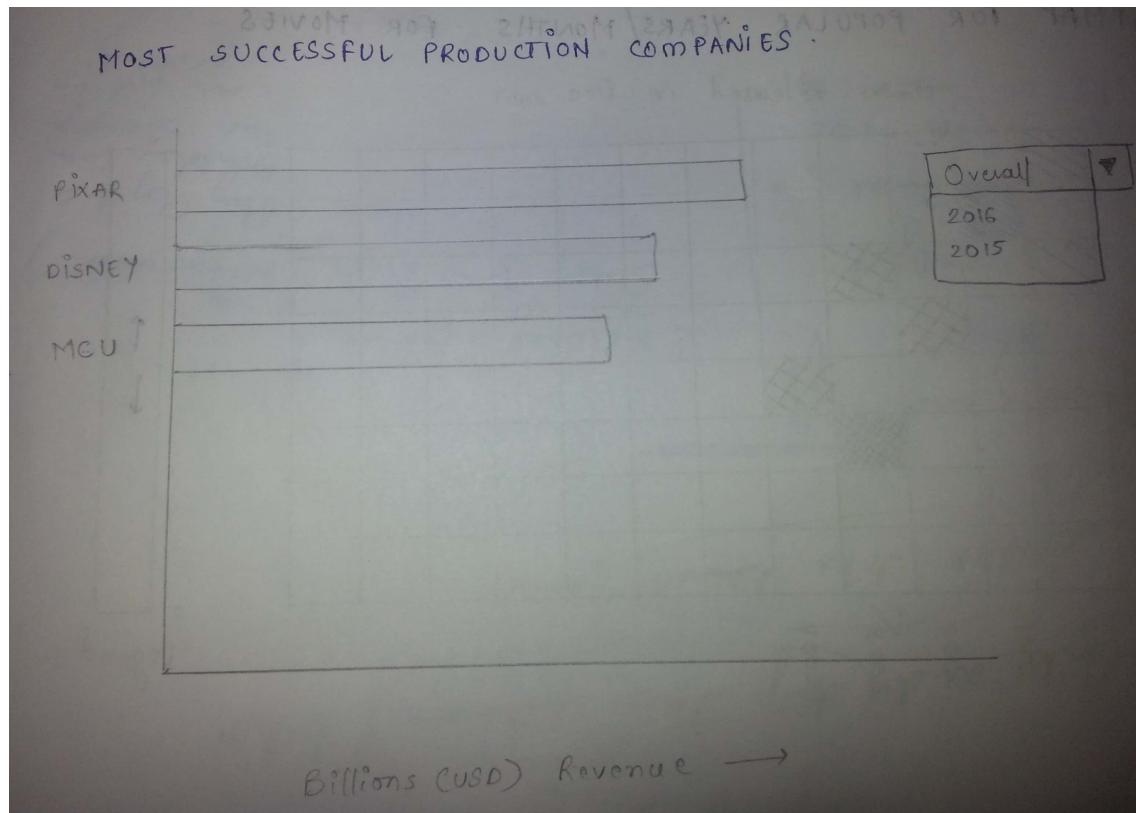
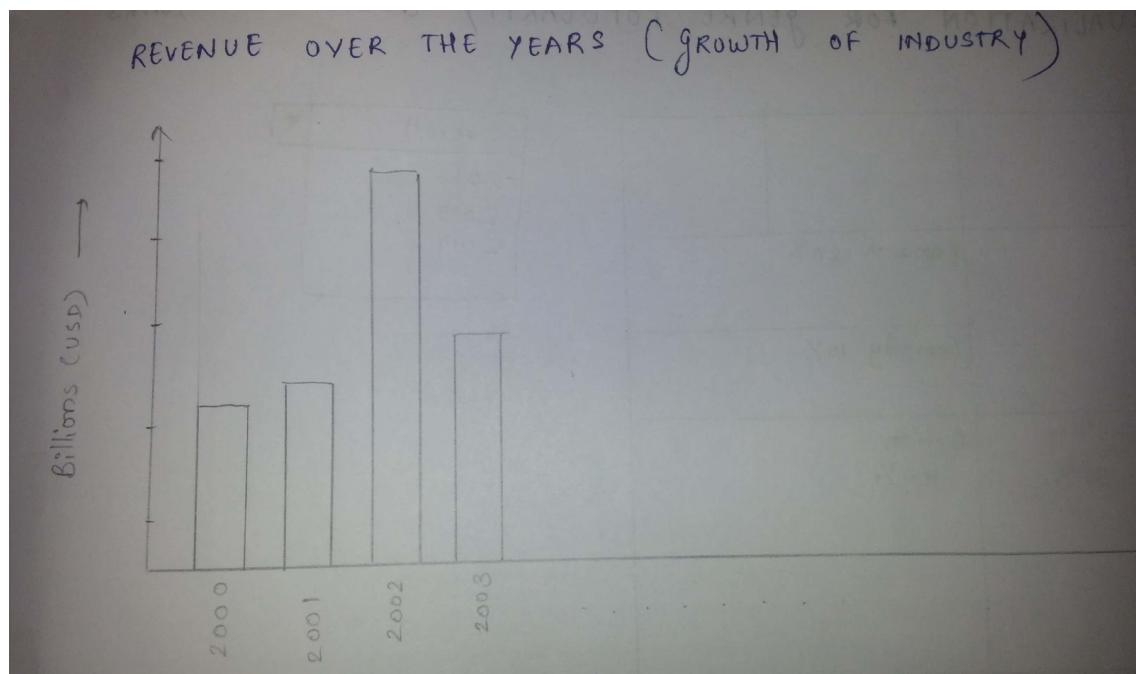
<https://www.kaggle.com/rounakbanik/the-movies-dataset>

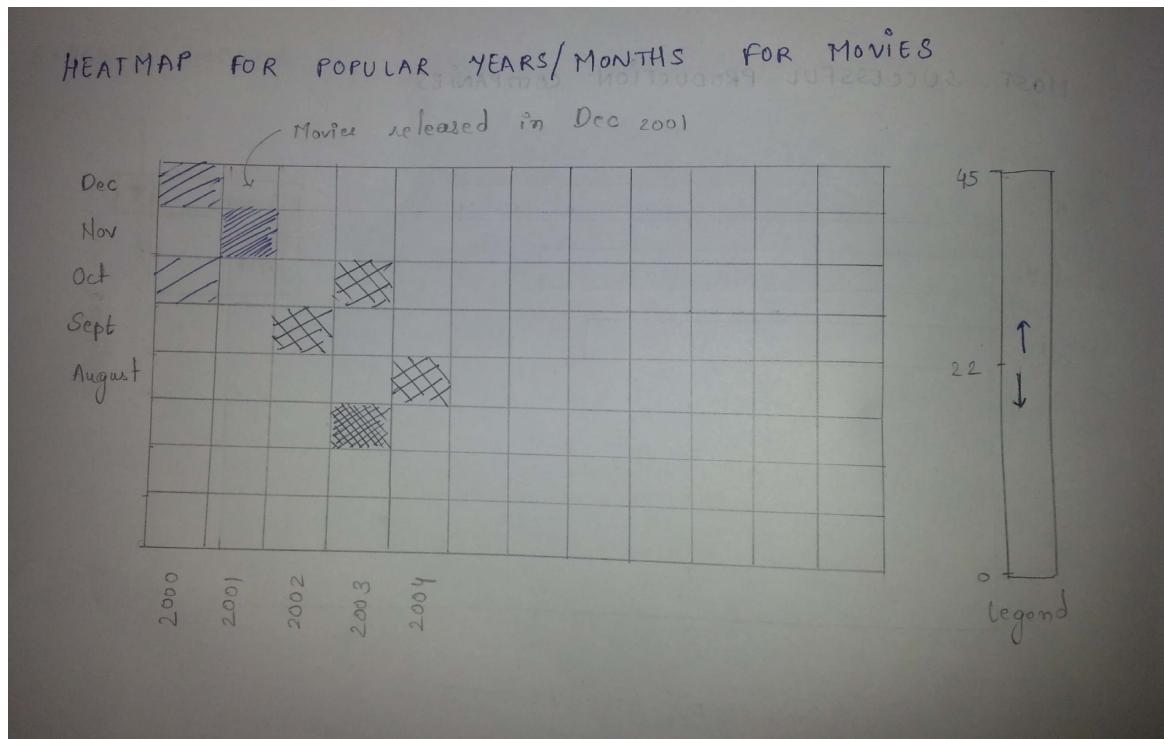
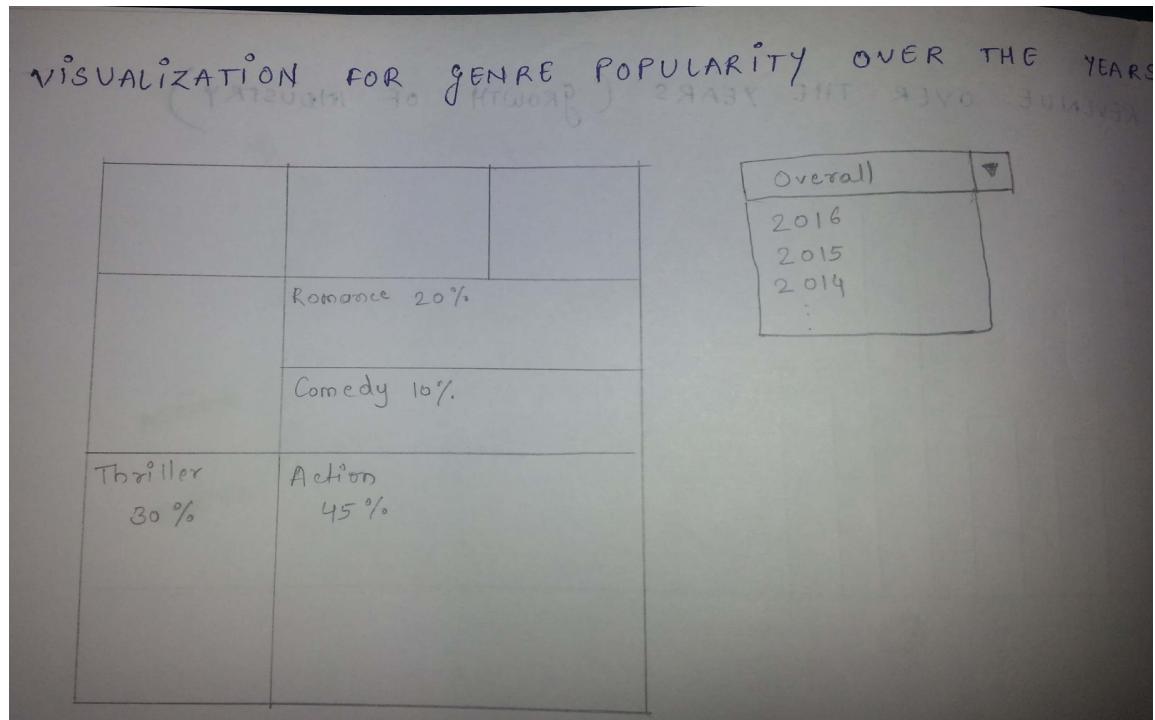
The overall dataset is too large, primarily because of the vast number of rows but also because it contains too much information, much of which we do not plan on using, so a considerable amount of data munging will be required to combine the datasets, remove unnecessary columns and also split (if it comes to that) the file into parts, containing data required for each visualization. We might use javascript or python to process the data.

## Visualization Design

We have thought about using barcharts, chord diagrams, scatter plots, venn diagrams and line charts to visualize the relationships we have drawn from the data so far. We plan to implement a combination of some of them to give it a more unique touch.









## Project Features

### I. Must have:

The primary objectives listed earlier are a must, as the objective of the project is to tell a story about the movie industry.

### II. Optional:

We also wish to extend the project to include more data on actors, their relationship with each other (as in the movies they've shared, etc), the number of awards they've won and so on.

## Project Schedule:

### Week 1

1. Functioning GitHub website
2. Trace out the specifications of the planned visualizations.
3. Process the data required for most of the above visualizations.

### Week 2

1. Complete pending data processing.
2. Have at least half of the basic form of visualizations up and ready.

### Week 3

1. Enhance existing visualizations and add interactions and animations.

### Week 4

1. Add remaining visualization and complete their functionality.

### Week 5

1. Wrap up work, and if time permits, enhance website with additional features and information on the project.