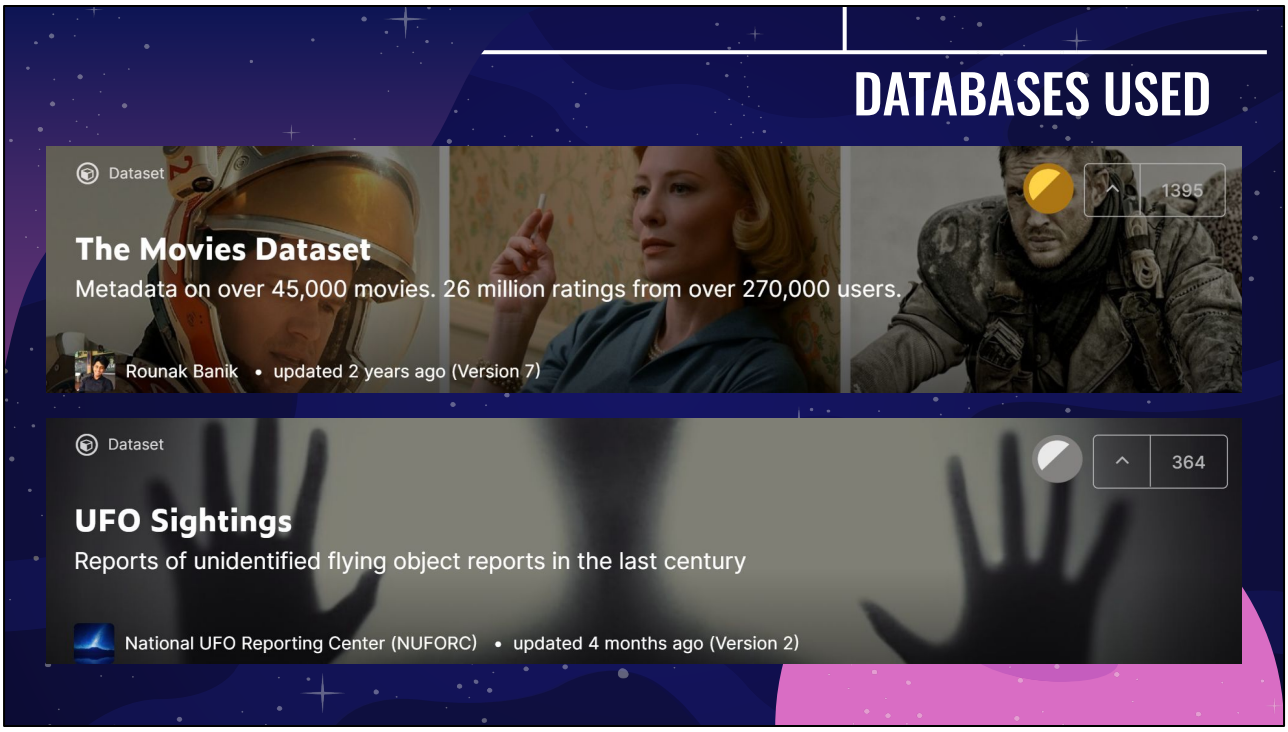
Welcome to our Distributed Networks final presentation of U-F-OH No by Anna Barone, Ethan Chan, Cameron Storton and Braden Wicker.



When looking through the abundance of interesting datasets on Kaggle, we found so many that we could use for this project. Although there were fascinating datasets on coronavirus cases, Golden Globe winners, common words used by presidential candidates, and more, we landed on these two datasets: The Movies and UFO Sightings.

Is there a correlation between science fiction movies and UFO sightings? What about other genres?

Using these two datasets, Braden and Ethan formulated the problem statement for this project:

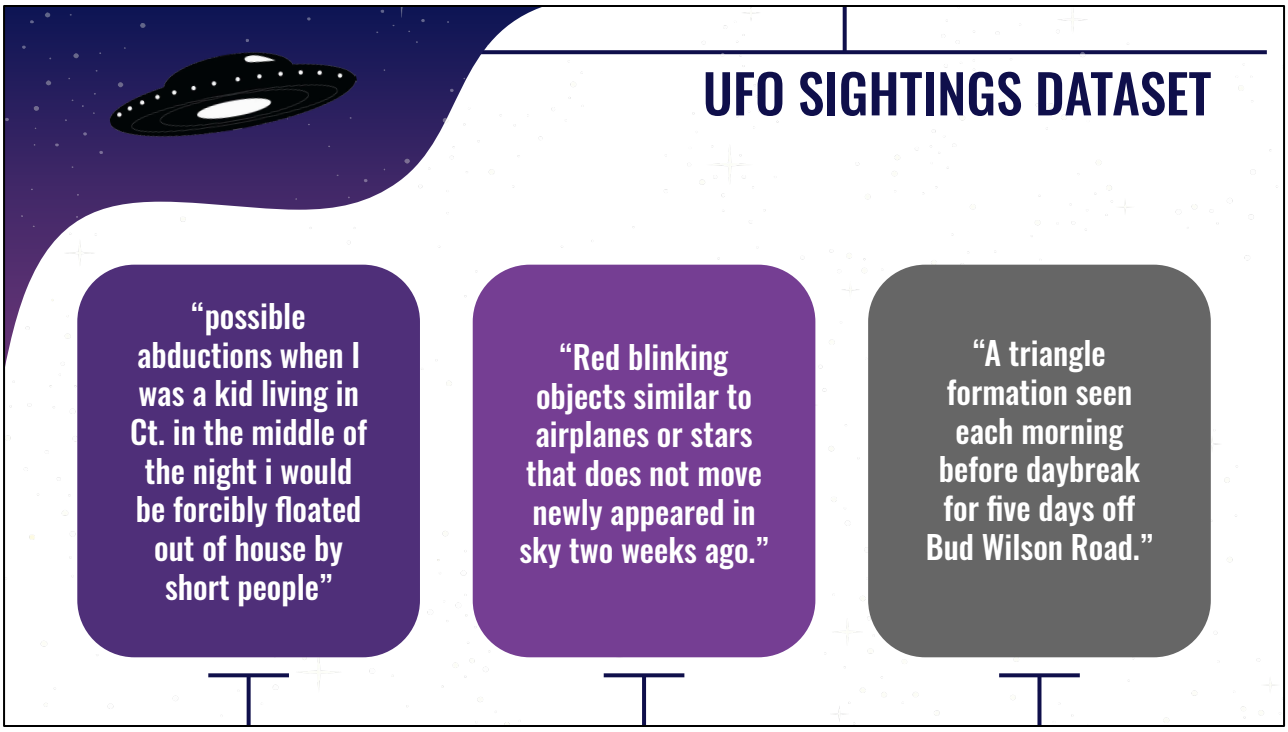
UFO SIGHTINGS DATASET

	A	B	C	D	E	F	G	H	I	J	K
1	10/10/1949 20:30	san marcos	tx	us	cylinder	2700	45 minutes	This event took place in early fall around 1949-50. It occurred	4/27/2004	29.8831	-97.9411
2	10/10/1949 21:00	lackland afb	tx		light	7200	1-2 hrs	1949 Lackland AFB&#44 TX. Lights racing across the sky &	12/16/2005	29.3842	-98.5811
3	10/10/1955 17:00	chester (uk/england)		gb	circle	20	20 seconds	Green/Orange circular disc over Chester&#44 England	1/21/2008	53.2	-2.91667
4	10/10/1956 21:00	edna	tx	us	circle	20	1/2 hour	My older brother and twin sister were leaving the only Edna t	1/17/2004	28.9783	-96.6458
5	10/10/1960 20:00	kaneohe	hi	us	light	900	15 minutes	AS a Marine 1st Lt. flying an FJ4B fighter/attack aircraft on a sc	1/22/2004	21.4181	-157.804
6	10/10/1961 19:00	bristol	tn	us	sphere	300	5 minutes	My father is now 89 my brother 52 the girl with us now 51 my	4/27/2007	36.595	-82.1889
7	10/10/1965 21:00	penarth (uk/wales)		gb	circle	180	about 3 mins	penarth uk circle 3mins stayed 30ft above me for 3 mins slov	2/14/2006	51.4347	-3.18
8	10/10/1965 23:45	norwalk	ct	us	disk	1200	20 minutes	A bright orange color changing to reddish color disk/saucer w	10/2/1999	41.1175	-73.4083
9	10/10/1966 20:00	pell city	al	us	disk	180	3 minutes	Strobe Lighted disk shape object observed close&#44 at low s	3/19/2009	33.5861	-86.2861
10	10/10/1966 21:00	live oak	fl	us	disk	120	several minutes	Saucer zaps energy from powerline as my pregnant mother re	5/11/2005	30.2947	-82.9842
11	10/10/1968 13:00	hawthorne	ca	us	circle	300	5 min.	ROUND &#44 ORANGE &#44 WITH WHAT I WOULD SAY WAS	10/31/2003	33.9164	-118.352
12	10/10/1968 19:00	brevard	nc	us	fireball	180	3 minutes	silent red /orange mass of energy floated by three of us in we	6/12/2008	35.2333	-82.7344
13	10/10/1970 16:00	bellmore	ny	us	disk	1800	30 min.	silver disc seen by family and neighbors	5/11/2000	40.6686	-73.5275
14	10/10/1970 19:00	manchester	ky	us	unknown	180	3 minutes	Slow moving&#44 silent craft accelerated at an unbelievable a	2/14/2008	37.1536	-83.7619
15	10/10/1971 21:00	lexington	nc	us	oval	30	30 seconds	green oval shaped light over my local church&#44power lines	2/14/2010	35.8239	-80.2536
16	10/10/1972 19:00	harlan county	ky	us	circle	1200	20minutes	On october 10&#44 1972 myself&#44my 5yrs.daughter&#442	9/15/2005	36.8431	-83.3219
17	10/10/1972 22:30	west bloomfield	mi	us	disk	120	2 minutes	The UFO was so close&#44 my battery in the car went to zero	8/14/2007	42.5378	-83.2331
18	10/10/1973 19:00	niantic	ct	us	disk	1800	20-30 min	Oh&#44 what a night &#33 Two (2) saucer-shaped&#44 glow	9/24/2003	41.3253	-72.1936
19	10/10/1973 23:00	bermuda nas			light	20	20 sec.	saw fast moving blip on the radar scope thin went outside anc	1/11/2002	32.3642	-64.6786
20	10/10/1974 19:30	hudson	ma	us	other	2700	45 minutes	Not sure of the eact month or year of this sighting but it was i	8/10/1999	42.3917	-71.5667

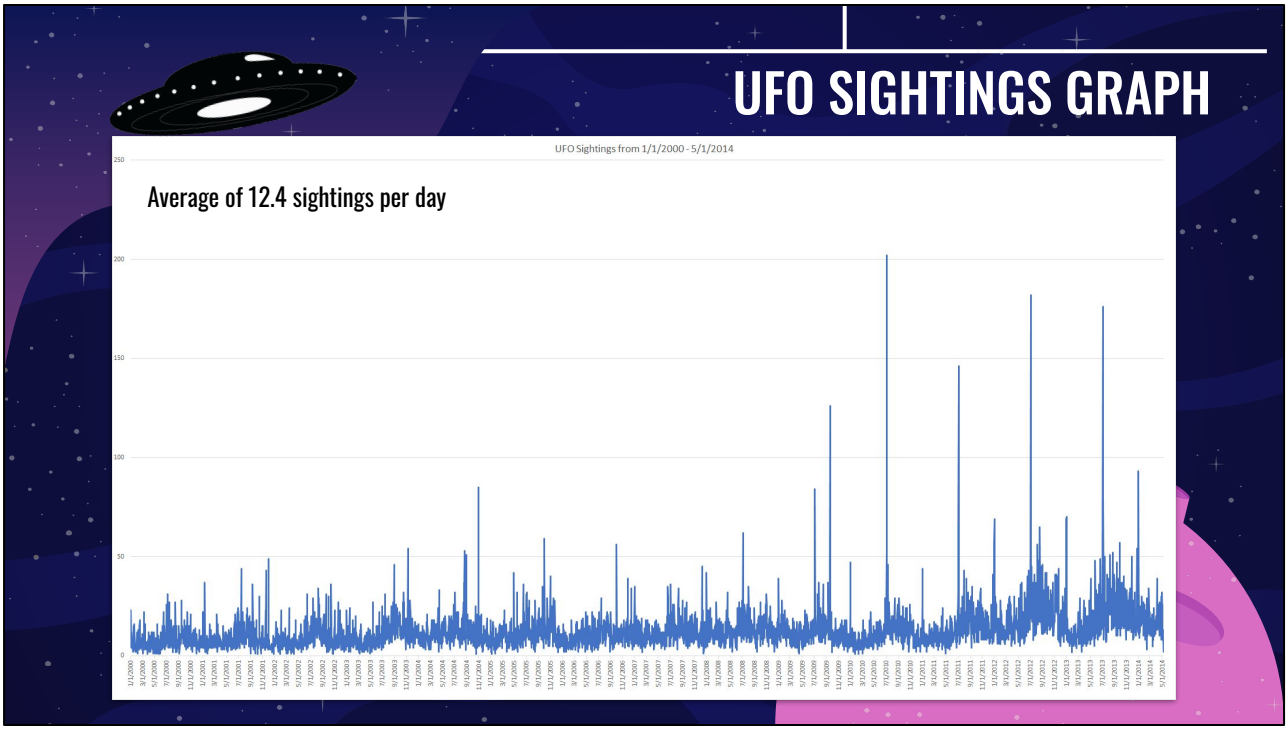
Data Provided by NUFORC

All of the data is taken from the National UFO Reporting Center. The National UFO Reporting Center (NUFORC) is an organization in the United States that investigates UFO sightings and/or alien contacts. NUFORC has been in continuous operation since 1974. This dataset contains the 80,000 reported and NUFORC-catalogued UFO sightings over its history, most of which were in the United States.

The NUFORC has a website with an online reporting form that those who spot a UFO can fill out. Through a note on the website, the administration of NUFORC hopes to filter out obviously mistaken UFO sightings, such as the new “train” or Starlink micro satellites, the planet Venus, and the star Sirius.



Here are some examples of the **Comments** column that most UFO reporters used to explain their UFO sighting. As you can see, some may not be reliable but others could be.



Before starting, we decided to plot the UFO sightings from 2000 to 2014. To help better understand the information we present later on, it is important to note that the average UFO sightings per day is 12.4.

An entertaining statistic we found was that the maximum UFO sightings in one day - 202 - was on July 4th (coincidence?).

THE MOVIES DATABASE

	A	B	C
1	genres	popularity	release_date
2	[{'id': 16, 'name': 'Animation'}, {'id': 35, 'name': 'Comedy'}, {'id': 10751, 'name': 'Family'}]	21.946943	10/30/1995
3	[{'id': 12, 'name': 'Adventure'}, {'id': 14, 'name': 'Fantasy'}, {'id': 10751, 'name': 'Family'}]	17.015539	12/15/1995
4	[{'id': 10749, 'name': 'Romance'}, {'id': 35, 'name': 'Comedy'}]	11.7129	12/22/1995
5	[{'id': 35, 'name': 'Comedy'}, {'id': 18, 'name': 'Drama'}, {'id': 10749, 'name': 'Romance'}]	3.859495	12/22/1995
6	[{'id': 35, 'name': 'Comedy'}]	8.387519	2/10/1995
7	[{'id': 28, 'name': 'Action'}, {'id': 80, 'name': 'Crime'}, {'id': 18, 'name': 'Drama'}, {'id': 53, 'name': 'Thriller'}]	17.924927	12/15/1995
8	[{'id': 35, 'name': 'Comedy'}, {'id': 10749, 'name': 'Romance'}]	6.677277	12/15/1995
9	[{'id': 28, 'name': 'Action'}, {'id': 12, 'name': 'Adventure'}, {'id': 18, 'name': 'Drama'}, {'id': 10751, 'name': 'Family'}]	2.561161	12/22/1995
10	[{'id': 28, 'name': 'Action'}, {'id': 12, 'name': 'Adventure'}, {'id': 53, 'name': 'Thriller'}]	5.23158	12/22/1995
11	[{'id': 12, 'name': 'Adventure'}, {'id': 28, 'name': 'Action'}, {'id': 53, 'name': 'Thriller'}]	14.686036	11/16/1995
12	[{'id': 35, 'name': 'Comedy'}, {'id': 18, 'name': 'Drama'}, {'id': 10749, 'name': 'Romance'}]	6.318445	11/17/1995
13	[{'id': 35, 'name': 'Comedy'}, {'id': 27, 'name': 'Horror'}]	5.430331	12/22/1995
14	[{'id': 10751, 'name': 'Family'}, {'id': 16, 'name': 'Animation'}, {'id': 12, 'name': 'Adventure'}]	12.140733	12/22/1995
15	[{'id': 36, 'name': 'History'}, {'id': 18, 'name': 'Drama'}]	5.092	12/22/1995
16	[{'id': 28, 'name': 'Action'}, {'id': 12, 'name': 'Adventure'}]	7.284477	12/22/1995

As stated before, the second dataset we used was “The Movies” which contains 32.85 MB of metadata for over 45,000.The metadata categories are: {adult, belongs_to_collection, budget, genres, homepage, id, imdb_id, original_language, original_title, overview, popularity, poster_path, production_companies, production_countries, release_date, revenue, runtime, spoken_languages, status, tagline, title, video, vote_average, vote_count}.

For this project, we utilized these elements of the Kaggle “The Movies” Dataset. Each row signifies a movie that was released on **release_date** with a **popularity** of the second column and that fit into every single genre listed in the first column. As you can see, there parsing for this file proved very difficult.

THE INNER-WORKINGS OF OUR CODE

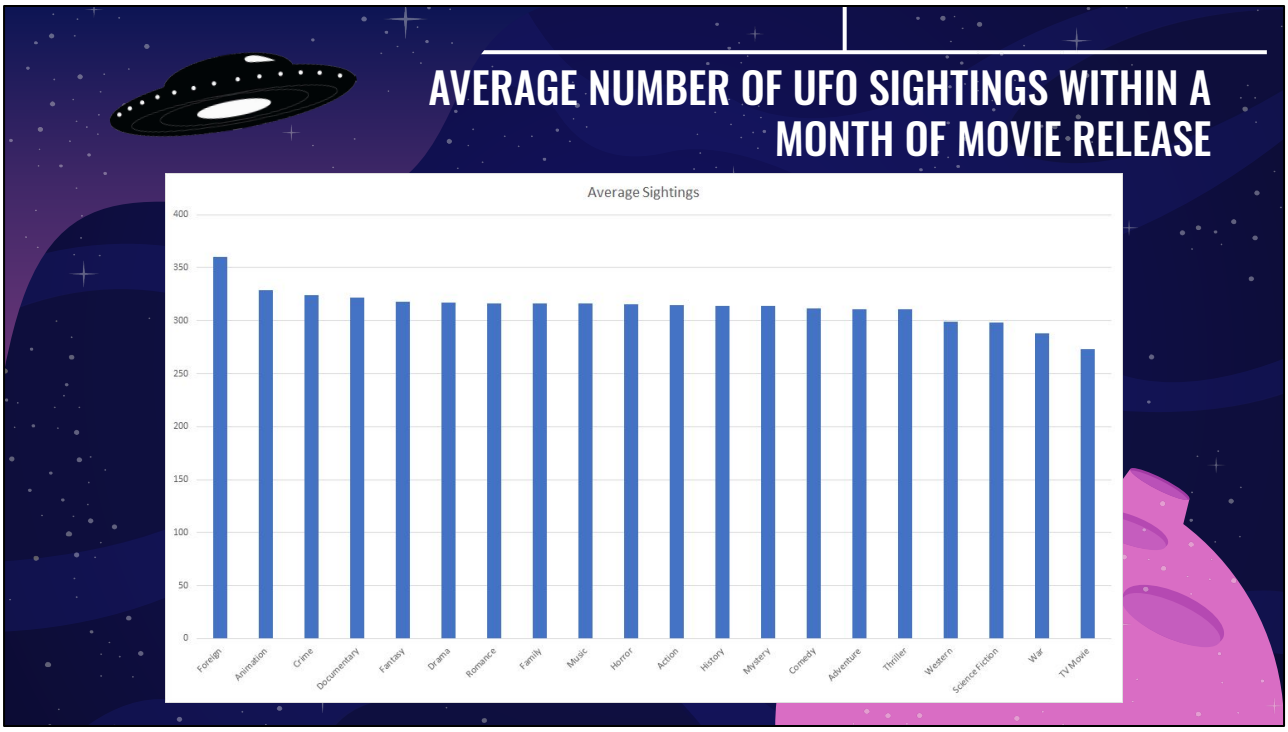
Finding the # of UFO sightings within 10 days of a movie release:

```
def getNumSightingsWithinMonth(startDate: LocalDate,
                               sightings: List[(LocalDate, Long)]): Int = {
 sightings.filter({case(date, count) => ((date.toEpochDay() - startDate.toEpochDay)
       <= DAYS_IN_MONTH && (date.toEpochDay - startDate.toEpochDay) >= 0)})
   .map({case(date, count) => count.toInt})
   .fold(0)((acc, count) => acc + count)
```

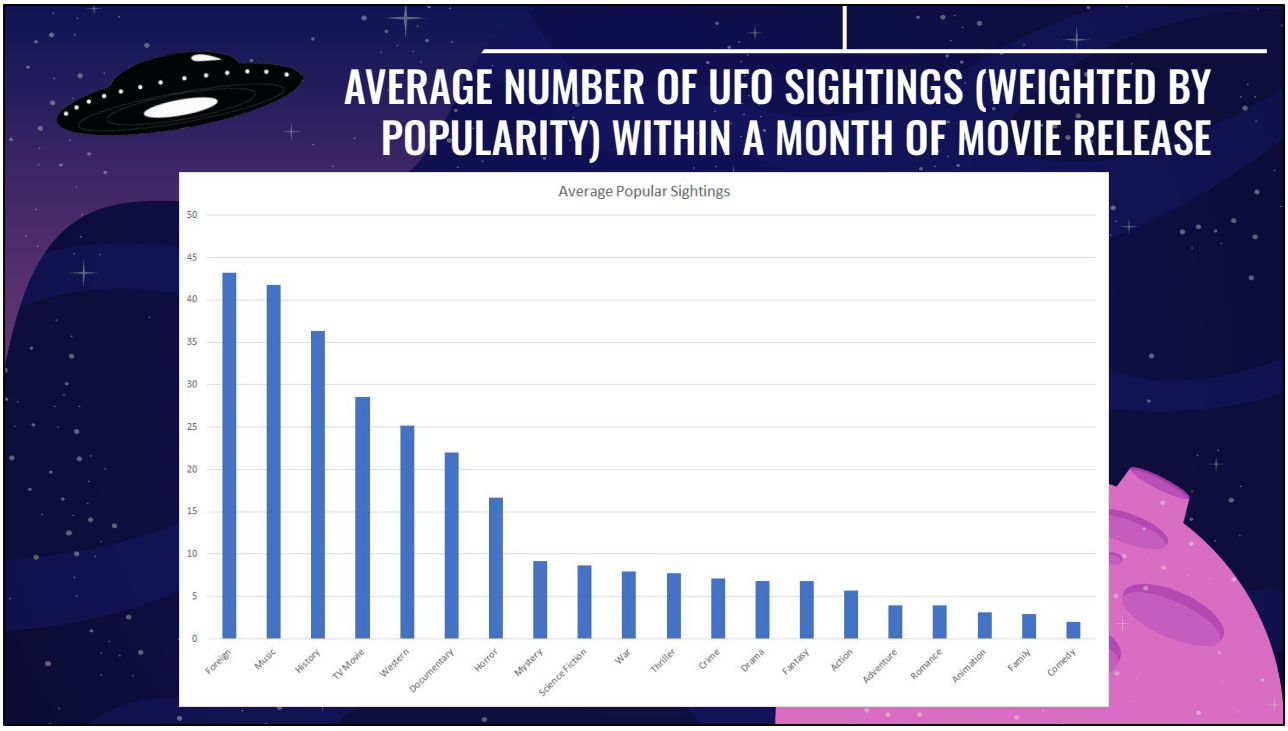
Formula used to normalize the # of sightings by the movie popularity

$$\frac{\sum \frac{\textit{movie popularity}}{\textit{highest popularity in the genre}} * (\#\ \textit{of sightings})}{\textit{total number of movies in genre}}$$

Once we parsed the two dataset we decided to use, we needed to figure out the number of UFO sightings within a month (30 days) of each movie's release date. We used this code and the Hadoop file system to do that for us - this task was most likely the most timely since the movies dataset has over 45k entries (at 4.5 MB) and the sightings dataset has over 80k entries (at 13 MB).



The first result we gathered from our search for a correlation is the average number of sightings within a month of release per movie for every genre listed in the Movies dataset. If there was a correlation between science fiction and UFO sightings, we would expect the number of UFO sightings in the science fiction column to be much higher than all other genres. As you can see, this is not true - foreign films have the highest average of UFO sightings within a month at 356. To verify that an average within the 300 range was reasonable, we looked at the previously stated average UFO sightings per day (12.4), so the average within a month - 370 sightings - checks out.



Since there was no correlation in just the average sightings, we decided to see if the popularity of each movie changes anything correlation. However, we wanted to make sure each genre was normalized since there's a chance that another genre's films could have been more popular than science fiction films.

The total time to perform the data analytics task was 16.810s.

Overall, we found there is no correlation between movie genres and UFO sightings.

In the movies database, most movies had multiple subgenres. For example, many movies were also under Music, and there were many Foreign movies as well.

WHAT WE LEARNED

- The relationship between UFO sightings and movie releases
- How frequently UFOs are sighted (around 12 a day)
- Data Normalization
- CSV Parsing
- Spark Optimization

MAJOR OBSTACLES

- Determining proper normalization metric
  - Normalize popularity by most popular movie in each genre, or
  - Normalize popularity by sum of popularities within a genre, or
  - Normalize popularity by overall most popular movie?
- Correlating movie popularity with number of sightings for a given movie
- Creating regular expressions to match complex datasets