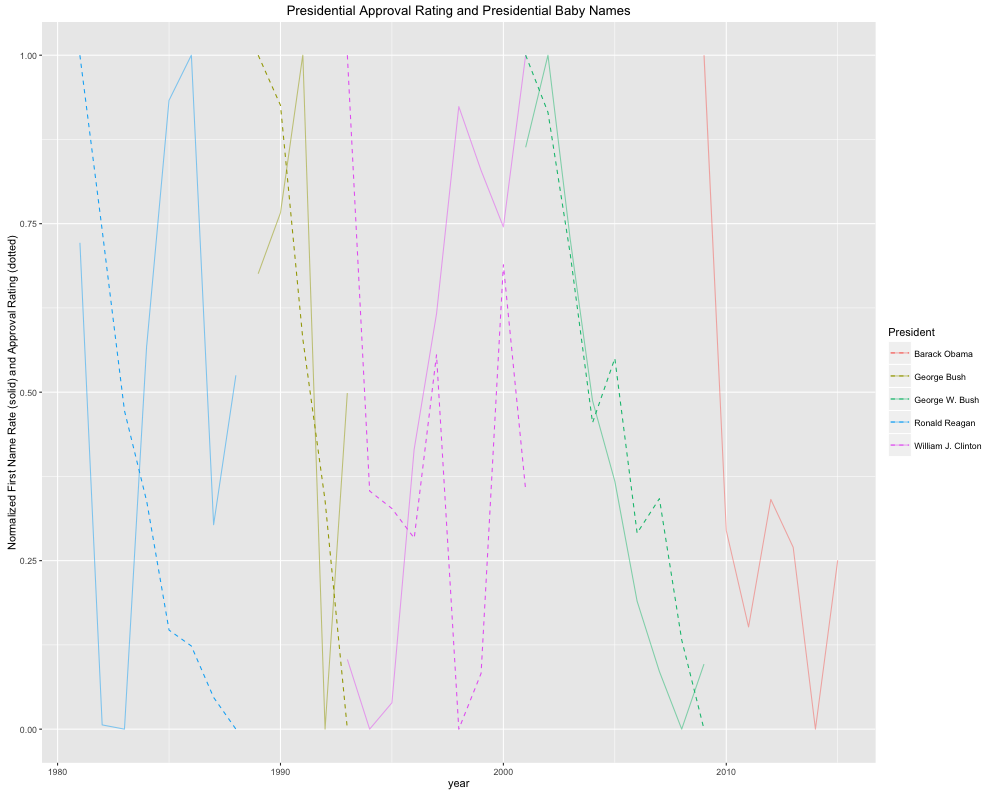
1. *“Instructions for executing your code”*
   1. **Presidential Approval**
      1. Ensure that the approval\_analysis.R script and the data/ directory is in the parent directory. The analysis draws from the data/presApprovalAverage.csv. Set your working directory to the parent directory of the R script and then run the code.
   2. **Movie Characters**
      1. Ensure that raw\_100\_films\_inflation\_adj.csv is in /data when executing movie\_data\_carpentry.ipynb. Please be aware that this is a python 2 notebook since the IMDbPY package only works under Python 2. Make sure IMDbPY has been installed before executing this file. After executing this notebook, movie\_data\_final.csv will be created. This file will be used in movie\_exploration.ipynb (Python 3) as input.
   3. **Disney Characters**
      1. For the (Rasha\_data\_coll.R) program:Ensure that the *(Disney\_Data.csv*) in the E: directory. After executing this program the (Rasha\_Data.csv) will be created.
      2. For the (*disney\_char.ipynb)*, ensure(Rasha\_Data.csv) and (StateNames.csv) in the output folder.
2. “*A 2-5 paragraph description of what the analysis from the modified script shows”*
   1. **Presidential Approval**
      1. The creator of the original Kaggle script (marked in our GitHub repo as original.ipynb) noticed that some Presidential names spiked during their respective Presidential campaigning years. The Presidential Approval analysis incorporates Presidential approval rating data over the Commander in Chief’s Presidency to see if it correlates to baby name rates. This particular analysis looks at the mean approval rating for a President during a year and see if it maps to the rate of their first name in babies for that respective year. Both approval rating and baby name rate or normalized using a min-max normalization method defined as function in the approval\_analysis.R script.
   2. **MVP of Sports**:
      1. We found a dataset that shows the MVPs of football in the NFL and then tried to find the peak of those names mentioned in the new dataset in the baby names and see whether the peak year is almost same with the year that they win the MVPs. Then we found that there is one name meet this requirement — Peyton.  
           
         Peyton won the MVPs in 2003,2004,2008 and 2009. This plot shows that the trend of “Peyton” and it shows that it increases from 2003 and reaches the peak in 2008 and 2009, which probably means that sports celebrities may have influence on babies’ name.
   3. **Disney Characters:**

In this analysis we want to know if there is an influence of the Disney characters on the baby names. The biggest challenge of this one is that collecting the data of Disney Characters along with the years and the movies names since there isn't any structured data in the internet. In order to collect the data, we used the data from IMDB websites(Disney\_Data.csv). However these data doesn’t contain the required information so I have written a program in R to collect the data from different pages in the IMDB using the links in the (Disney\_Data.csv).

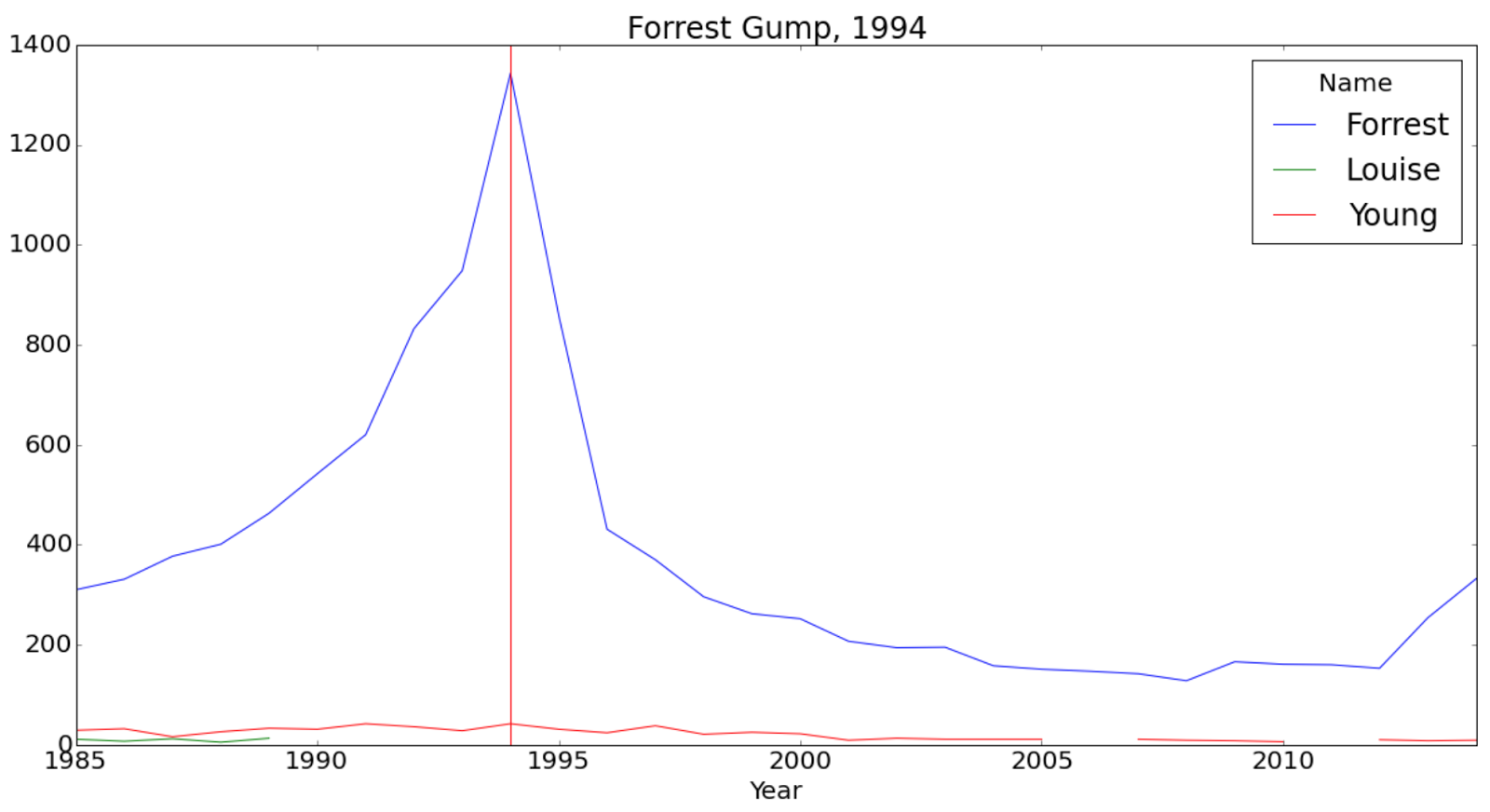
After collecting all the needed data, we analyzed the new data to see if there any correlation between the Disney characters and the baby names. In the data that collected there is a list of all characters of the movies. However, we just examined the first two characters since they are supposed to be the main characters in the movie.

1. “A 2-5 paragraph description of what the analysis from the two additional visualizations would (does) show”

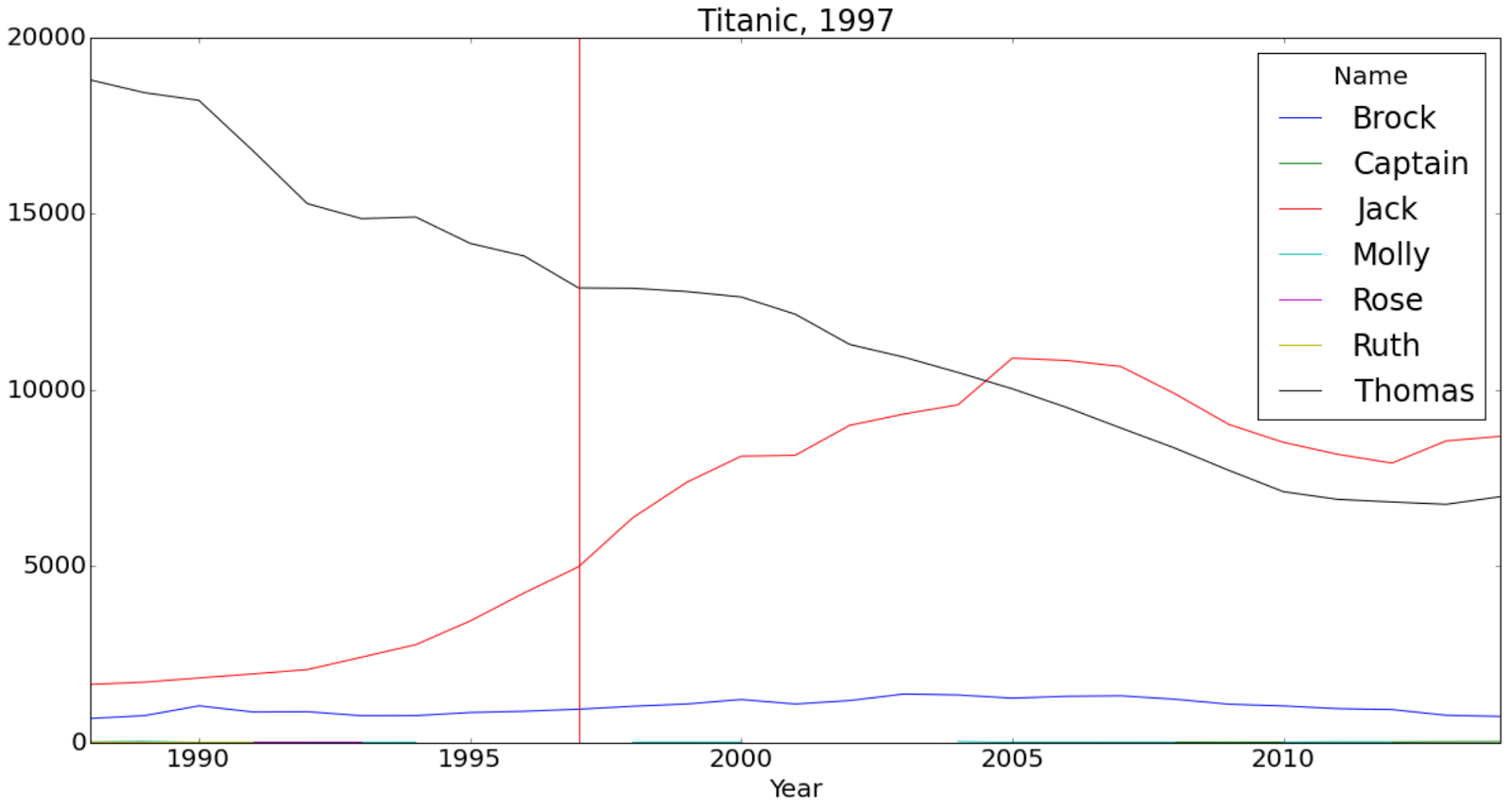
**Presidential Approval**

The graph above shows the trend of both the normalized first name rate (solid lines) and the normalized approval rating of that President. There seems to be some similarities in the trend lines for particular Presidents (ie. William Clinton), but overall it is hard to interpret whether or not approval rating is causal of baby name rates.

**Movie Characters**



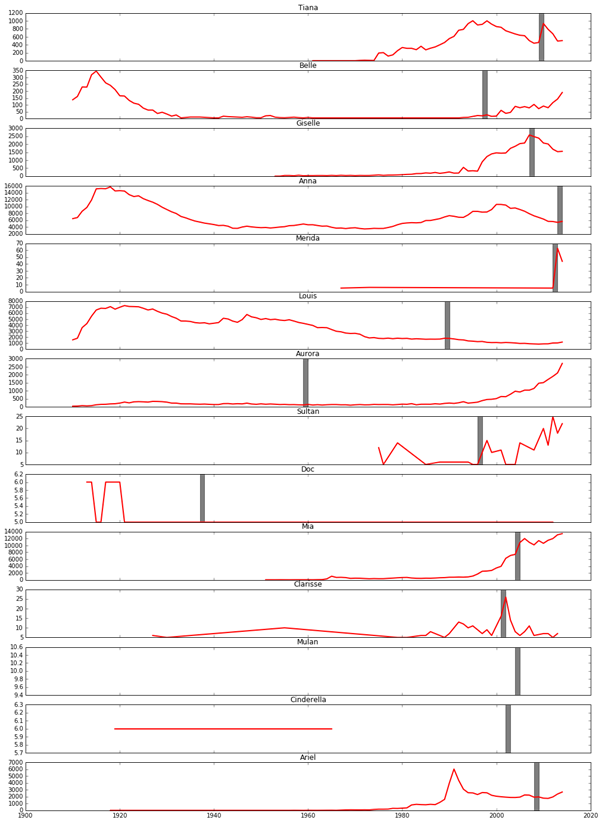
The movie “Forrest Gump” had spiked baby name “Forrest” in 1994.

But the movie “Titanic” didn’t spike “Jack” in 1997 although there is an increasing trend. One possible reason is that Jack is a more common name and the spikes were diluted by the large population base. For the movie series “Star Wars”, they spiked “Luke” and “Anakin” in different episode releases. The plots can be found in the “movie\_exploration.csv” notebook or in the slides. 

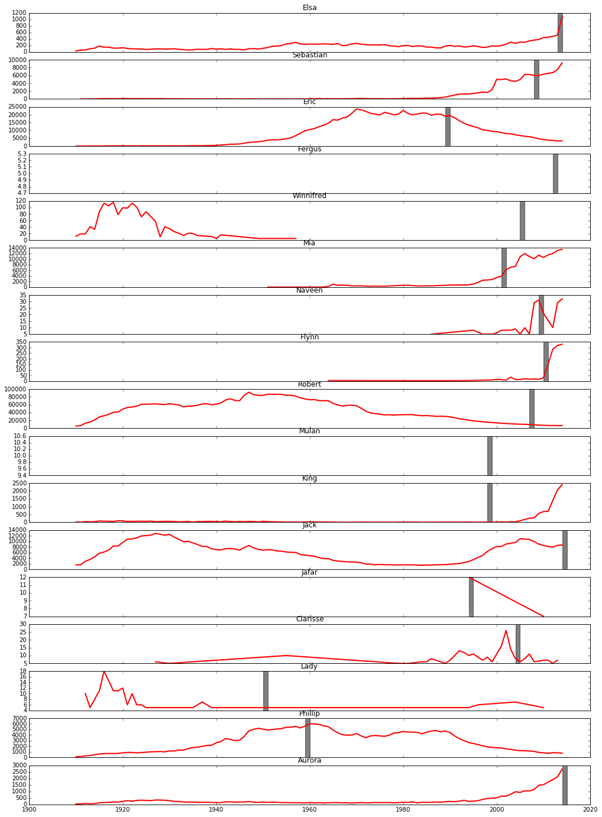
Movie characters do have influences on baby names. It’s easier to observe spikes for uncommon names than common names.

**Disney characters:**

The following graph shows the first character in each movie. there are some names appear more than once since some movies has many versions. I filtered out those names to be able to draw them. We can see some characters have influence on the baby names such as ('Tiana', 'Belle','Giselle', 'Sultan', 'Mia', 'Merida', 'Clarisse'). Some of them doesn’t appear in the original data set. other don’t affect the baby names such as 'Cinderella', 'Aurora').



The next graph shows the second characters in each movie as stated in the IMDB also we can see some of those characters are influence the baby names such as (Elsa, Sebastian, Mia, Hynn, Jafar) where other don’t have such influence such as (Winnfred) some other influenced negatively such as (Naveen) which was an animal in the beginning. Other don’t appear in the original data set such as (Fergus).



1. *Individual team member reflections on the process.*

**James Bain**

The process was actually fairly straightforward once we agreed upon a starting data set. Once that was in place, it was easy to pick a particular question to answer involving Presidential approval ratings. The beauty of this mini-analysis is that it involved scraping data from an online source and then figuring out how to aggregate that data to join them to the starting data set. This analysis was in no way exhausted. Future directions could lead to a more robust analysis between approval rating and baby name rates such as accounting for lag times, or looking at last names as well. We could also incorporate state level name rates as well as political affiliation during that election year for that state to see if that plays a role in how babies are named.

**Andrew Pistole**

The real challenge of this process was choosing the datasets. Our first process was thinking of questions based on one data set (the baby names), thinking of questions, and then finally finding the other datasets. By trial of error we were finally able to come up with questions to ask between two datasets and find the data required to back it up. Once having both sets of data the process was quite easy and straightforward. We are no strangers to the second part of the process. Next time I would wonder if actually choosing two datasets first and then thinking of question would make the process that much easier or not.

**Keevey Song**

Before the professor mentioned this mini project, I never thought that we could combine two datasets in order to dig deeper and get to know more. Because of too few attributes in the dataset of San Francisco salary, we changed the dataset to baby's name. And I combined the baby's name with MVPs of popular sports such as football and basketball. And I did find something interesting that popular players may have influence on baby's name. Next time, if I have a dataset, I will think deeper and try to find something else and combine them in order to get more interesting information.

**Peng Zhao**

The first challenge for the movie characters part was getting the data of characters. The original movie dataset only contains the rank, title, year and simple description of the movies. In order to get a structured data, I used the third-party package IMDbPY to query the IMDb’s web server based on movie titles. This was very slow and for each movie it took about 6-7 second to get the characters. I limited the characters to be the top 10 characters for each movie to save time. The other challenge was there are 10 characters in one movie and about 1-2 popular movies each year. Thus, unlike the president data, there are too many characters each year and it’s hard to directly correlate them to baby names due to the limitation of visualization methods. Then I decided to explore baby names by different movies and got some interesting results which indicated the characters in popular movies do have influences on baby names. The future work could be analyzing if names of cast also have influences and seeing if the scandals of some of the cast can influence the trend.

**Rasha Gargees**

I have made a lot of research to find a data that can be integrated with the baby names. However, finding a structured data isn't easy. Finally, I decided to write a program in R to scrap a data from the website IMDB and that was another challenge. The data that I was interested in, is Disney characters data. I successfully retrieved this data. As the next step, I analyzed this data to see if there is any correlation between the baby names and the Disney characters. I selected the first two characters, which are supposed to be the main characters. Some of the finding in this analysis is that, the baby names trend affected by the Disney characters positively. On the other hand, there are some character the do not affect the baby names or affected them negatively. For the future work, I will analysis the data with each state separately to see if any state follow this trend more than other states.