Analysing Popular US TV shows by viewers per million and have them compared.

Kevin Carmody
School Of Computing
National College Of Ireland
Dublin
x15468652@student.ncirl.ie

Abstract— **This paper presents the comparison between three of the biggest TV shows in the United States right now; Suits, NCIS:LA and the original NCIS. Our TV study will explore various information such as, views (per million), date released and the rating it was given provided by an online website.**

Suits is a drama series about lawyers and what they do daily and how they perform while moving up the ranks at the firm. It was first aired in 2011 and its currently entering its ninth season. NCIS & NCIS:LA is an action tv series about tackling terrorism before it happens and about investigating crimes associated with the navy. NCIS is the original show but in 2009 NCIS:LA was created as a spin off. Both shows are still going today as NCIS is onto its seventeenth season and NCIS:LA is entering its eleventh season

From my study of all these TV shows, I will analyze the patterns within the dataset I have created in R for each TV show and see if I can find any interesting patterns within the shows Data. From all this data, I will be able to see if the shows had views dropped or increased; or if ratings had dropped or increased as the shows went by. A top twenty MapReduce pattern will be used to see various patterns such as top 20 episode regarding views & top 20 episodes regarding ratings.

Keywords—component, formatting, style, styling, insert (key words)

I. Introduction (*Heading 1*)

Both NCIS shows and Suits are one of the best TV shows in America in the last 15 years with critics raving about them ever since their release date. NCIS was introduced to viewers in the early 2000s and is still going to this day. With a loyal fanbase, the show might keep going for the next few years. In 2008/9, the producers of NCIS thought it would be great to do a Spin-Off Series of the show, therefore created NCIS:LA. The show was a major hit in the US and became one of the go-to shows to watch on a weekly basis according to TV critics. Suits was created in 2011 and became an instant hit with Fox producers instantly. The show gives us an idea what happens within a lawyer's firm and what they do with a few extra incentives added in..

The idea of this project was to download an already created dataset of NCIS:LA from the internet. This dataset contained already featured information from Wikipedia but lacked information about the rating system of each episode. I worked through the data and cleaned the data to make sure I had only the information I needed to give and create patterns

for the project. To create and produce a dataset for NCIS and Suits, I web scrapped data from 2 various websites such as Wikipedia for all episodic information such as episode name, number of series & number of the episode within the season; and iMDB for the ratings for each episode. The two datasets were then combined into one dataset. Once all information was gathered and cleaned, it was then stored as a CSV file that was then to be used for further analyses.

The objective of this project was to find in-depth information in relation to various top-rated TV shows about consistency of ratings towards the show and views (per million) while also searching for drops or increases in ratings and views. MapReduce patterns are then inserted to find specific results that I wanted to see prior to analyzing.

II. Literature review

The project consists of three datasets that were obtained through various methods. One Dataset, called NCISLA was obtained through a website called Kaggle. The Dataset obtained information from what looked like to be from Wikipedia. The Dataset had variables such as Name of Show, Season, Viewers per million and number per episode. The dataset had 5 seasons of data which was about 120 rows within the dataset. This dataset would allow me to compare various variables against other shows such as viewers. I felt like the project could have used an extra variable, so I went searching for a location to find ratings per episode that was reviewed by critics and viewers of IMDB.

The second and third dataset for this study is being web scrapped via an API. Data for the dataset is to be extracted from various outlets such as Wikipedia and IMDB. NCIS and Suits have their own dedicated Wikipedia pages per season with all episodes including variables such as viewers per million, air date and number of episodes, weekly rank and more. The data can be used for extensive analysis. Suits and NCIS also had their own IMDB rating pages for each season. The ratings can be used to see if the quality of the episodes overtime started to get better or started to decline. Same goes for the viewers. We can use that variable to see if there is a increase or a decline in viewers of that certain TV Show.

III. Data

The project consists of three datasets that were obtained by either web scrapping or manual download through Kaggle. The First Dataset was NCIS:LA data that was downloaded via Kaggle main website. The Dataset well fully web scrapped and cleaned contained various variables such

as number in series, viewers and ratings. and the following image is the entire summary of the data.

Fig.1 Summary NCISLA

This is the dataset for NCIS:LA after been cleaned. The process to gather this was download the information from Kaggle and then clean the datasets within RStudio. Of course, the dataset is never perfect, so we have to make sure it suits our requirements along with future datasets. As you can see from above, The Dataset consists of 8 variables, many of which were created in RStudio or Web scrapped from another website such as IMDB.

Our next and second dataset comes from NCIS:LA big brother, the original NCIS. This dataset is web scrapped from an online resource via an API created in RStudio by me. 5 random seasons (in a row e.g Season 12-16) were scrapped via the API and then cleaned by me using RStudio. This Dataset consisted of 8 variables, some which were removed due to unnecessary requirements such as Weekly Rating or DVR Viewers. Once these variables were removed, I could then add in my own unique variables for the project.

Fig.2 Summary of NCIS

The above image is a summary of the NCIS Dataset after cleaning and adding in my own unique variables. This dataset is then created as a CSV file for further analysis including MapReduce.

The final dataset comes from a TV show called Suits. The same techniques are done on this dataset but with the addition that there is less episodes per season than the other 2 datasets we previously spoke about above. So, it's likely that future analysis could be skewed if and when compared against the NCIS datasets.

Fig.3 Summary Suits

IV. METHODOLOGY

The methodology used within this study is the Knowledge Discovery in Databases methodology, otherwise known as the KDD methodology. KDD is known as the **"process of finding knowledge in data, and emphasizes the "high-level" application of particular data mining methods"**

The system follows the process off: Selection, Pre-Processing, Transforming, Data Mining, Evaluation and Knowledge.
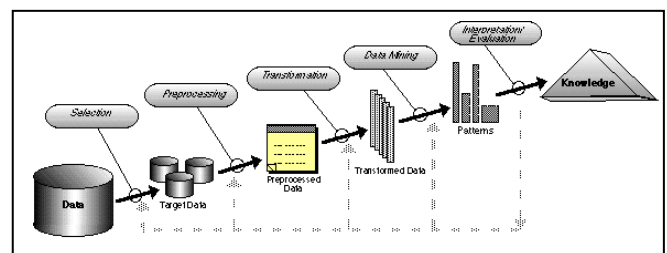

Fig.4 KDD Methodology

The image above is the KDD methodology and the following with the definition of all KDD processes.

Selection: This stage of the KDD process is vital to having a successful project as its utmost key to have the most precise data for your study. For the Project, three datasets were handpicked to go up against each other. NCIS:LA was downloaded through a resource website while NCIS and Suits were scrapped via a API created by myself in RStudio.

Pre-Processing: The NCIS:LA dataset was imported to RStudio after downloading the CSV file. The pre-processing part of the project starts when I removed unnecessary columns within the file. I then cleaned the dataset to make sure that there was no missing data from any of the columns and that all variables were named correctly. Numeric columns were either changed to integers or doubles for the project.

Transformation: When all datasets have been selected and processed successfully, the datasets will then need to be saved to a local database. For this study, I've chosen to save the datasets as a CSV file for easy access. From there, the data can be easily accessed by MapReduce Patterns for further analysis of the datasets that I see fit. For example, a Top 20 MapReduce Pattern.

Data Mining: The part of the methodology is where further analysis is required and commence. Soon as all datasets are created as CSV files, we can then commence various analysis on the datasets such as MapReduce Patterns as previously mentioned before. For this study, I will use a Top 20 MapReduce Pattern on all 3 datasets to see if I can see any interesting patterns. I will look mostly towards ratings and viewers (per million) in this analysis as well as season numbers.

Evaluation: Once our analysis is done on all datasets, we can know look at all results and see what our analysis have given us. Our evaluation should show us what we were looking for, such as Top 20 viewed episodes and Top 20 rated episodes.

Knowledge: From applying the KDD Methodology, we gained great knowledge by comparing results against other shows to see which was more superior by looking at the highest number in views.

V. IMPLEMENTATION AND ARCHITECTURE

To draw this project to a close, I have executed a MapReduce pattern on all three datasets that were created by using the KDD Methodology as previously spoken about in the previous chapter. For this implementation to come to fruition, three datasets were extracted as CSV files and then used in a MapReduce Pattern to find the Top 20 episodes based on viewers. The MapReduce was saved into a text file which was then turned into a python file by saving it as a .py file. The file was saved in the same location as all other CSV files for easier access to analyze the data. Once our coding is done, we can then go ahead and run our python files on our selected and cleaned datasets that were created previously in our RStudio.

Our python file gathers our Top 20 episodes from each show that is based on views as per million and them sorts them out from number one to number twenty in order. Once completed we can then look at all datasets and search for some interesting insights within the data.

The following is various software used to gather all information required for this study.

MS Excel: Excel is used to create better looking graphs for the project. Certain graphs will be used to create interesting graphical insights to the study.

Python: Python is where all coding patterns are created for future analysis on various datasets of my selection. The MapReduce used for this study is a Top 20 episodes based on views over 5 seasons. The patterns will be used on all three datasets.

Command Prompt: Command Prompt was used to run the patterns on the datasets and then sort them out in order of number one to number twenty. It's also used to create new CSV files with each show, showing its top 20 episodes.

RStudio: This is where most of our KDD methodology is done. The data is scrapped from online resources where it is then cleaned and transformed into data that is applicable for our study. Three R Script were created for each show where it is cleaned, transformed and exported as CSV files.

VI. RESULTS

As for the results of our completed study, I found quite a few interesting insights within the datasets pre and post MapReduce.

For our first dataset, our NCIS:LA dataset showed us interesting insights.
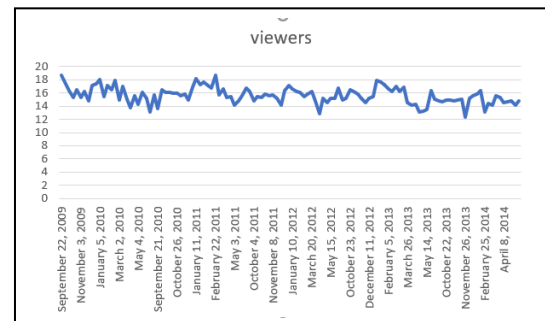

Fig.5 Viewers per Million NCIS:LA

The above image created presents to us two major variables within the dataset for NCIS: LA. These variables are viewers per million and original air date. The graph here is created in MS Excel to show us the viewers per million over a 5-year span.

We can see over five seasons of the show, that viewers stay at a consistent level, even though there seems to be a drop-in viewership, the viewership does tend to rise again and again. Reasons behind this could be down to multiple theories. Reasons such as new viewers giving it a go at the start and maybe not finding it that interesting throughout the years and would only come back from time to time or if maybe there was a special guest previewed for the shows in advertisements. From the third season to the fifth season there seems to be a consistent viewership ranging from 12m to 14m views per episode.

The results from the MapReduce pattern created by me, we also see some stimulating insights into NCIS:LA viewership.
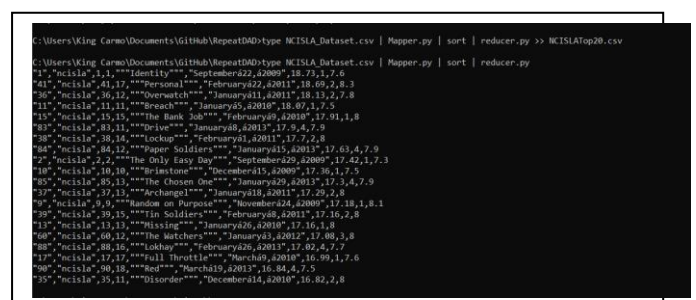

Fig.6 MapReduce Results NCISLA

The top twenty episodes range between 16.82m to 18.73.

Out of those twenty episodes, many of the episodes come within the first two seasons. Reasons for this is because NCIS:LA was originally envisioned to be a spin-off season to the original NCIS show that we will comes across later. It's quite possible that NCIS:LA just wasn't for some people who loved the original TV show.
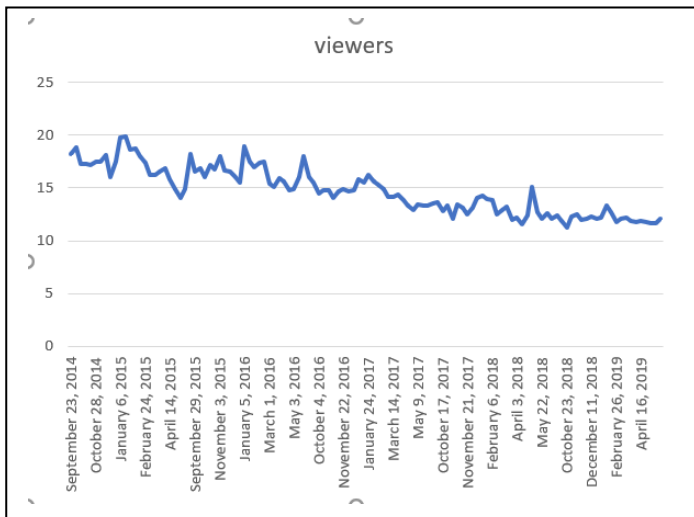
Fig.7 Viewers per Million NCIS

Our second dataset shows us many stimulating insights into the original NCIS over the last five seasons since 2014. It's important to know that this show came to TV in the early 2000s and has kept up with viewership on a consistent basis since then.

However, the graph above shows us that over the last five seasons, we are seeing a drop-in viewer per million. As before, many reasons can cause this but as I'm an avid viewer of the show, I know why these drop-in views happened. Two major cast members left the show around 2013/2014 that were a huge mainstay in the shows history and new characters brought in to replace them weren't connecting with the viewers. Despite that, the views are still above 10m+ which is quite good for a show that has been going over seventeen years. These views can be contributed to the show having such a loyal fanbase over the years with characters like Gibbs keeping loyal fans watching the show.
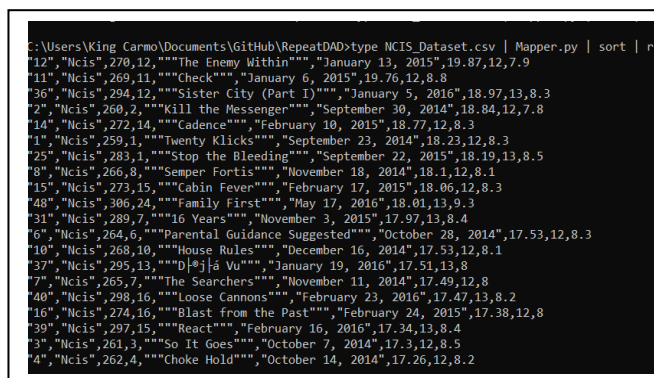

Fig.8 MapReduce Results for NCIS

The above image is of our MapReduce Pattern for NCIS showing us it's Top 20 episodes for the last five seasons.

As expected, the majority of the Top20 shows came within the Season 12-13 as shown by the line chart as previously shown.
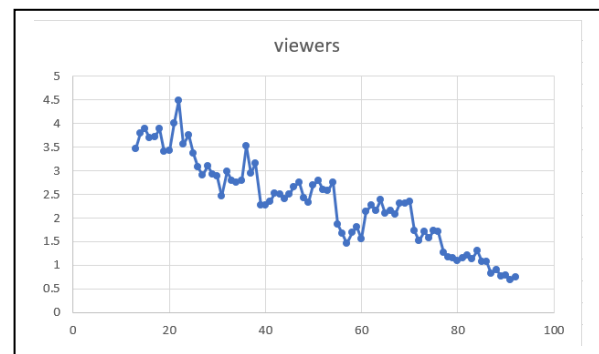

Fig.9 Viewers Per Million, Suits

Finally, Suits. This show started in 2011 and became a mainstay on the USA Network and Netflix which brought many viewers to the show. But let's see what happened to the show over time:

The above chart is such an attracting insight into the show's viewership. The show dropped massive in viewers around the 25th episode of the show, slightly rises but decrease at a consistent level for the next 60 episodes. This is a striking image; the show went from 4.5m million to less than 1m views in the matter of 5 seasons which I find shocking as the show does have great ratings on IMDB.
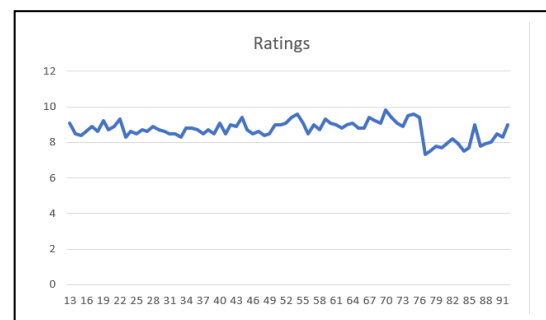

Fig.10 Ratings for Suits

As you can see from the image above, ratings have always stayed at a consistent level at between 8 stars to 10 stars which is what I find quite interesting for a show that is decreasing in viewership.
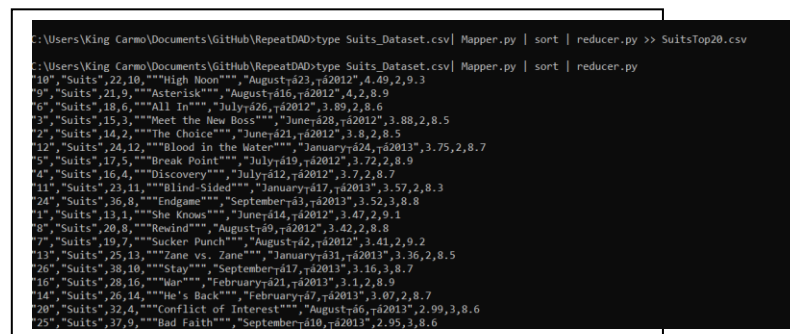

Fig. 11 MapReduce Results for Suits

The above image is our MapReduce Pattern and it confirms what we have seem the other images provided for the show called Suits. Majority of its viewers came around

Season 2, Season 3. With Viewership ranging from 2.95m to 4.49m on an episodic basis. Rating are very high as well, all ranging between 8 to about 9.3 stars.

For a show to have its viewers decrease so rapidly but have there ratings still very high, an internal problem must have happened such as Netflix being a pay to watch program that many people wouldn't buy into that.

VII. Conclusion and future work

From my understanding and after completing my study on all three datasets, I was able to discover many exciting insights within the data. All three datasets have one thing in common that is that there all decreasing in viewership overtime. NCIS and NCIS:LA were slowly and a consistent rate losing viewers, but Suits lost over 4m viewers within 60 episodes whereas NCIS LA have lost about 4m viewers in maybe 10 years. All shows tend to have there best 20 episodes around the same time as each other as well.

For future work, I would have loved to do more season of all shows to gather more precise insights within the numbers of the show. Everything I did for this case study, I would have done the same thing but with a wider range of data such as who directed the show, who wrote the script and so on. This could give me some exciting information to see who the most influential person behind the scenes is.

VIII. References

[1] en.wikipedia.org, "Suits Season 2," [Online]. Available: https://en.wikipedia.org/wiki/Suits_(season_2). [Accessed July 2019].

[2] en.wikipedia.org, "Suits Season 3," [Online]. Available: https://en.wikipedia.org/wiki/Suits_(season_3). [Accessed July 2019].

[3] en.wikipedia.org, "Suits Season 6," [Online]. Available: https://en.wikipedia.org/wiki/Suits_(season_6). [Accessed July 2019].

[4] en.wikipedia.org, "Suits Season 5," [Online]. Available: https://en.wikipedia.org/wiki/Suits_(season_5). [Accessed July 2019].

[5] en.wikipedia.org, "Suits Season 4," [Online]. Available: https://en.wikipedia.org/wiki/Suits_(season_4). [Accessed July 2019].

[6] en.wikipedia.org, "NCIS Season 12," [Online]. Available: https://en.wikipedia.org/wiki/NCIS_(season_12). [Accessed June 2019].

[7] en.wikipedia.org, "NCIS Season 13," [Online]. Available: https://en.wikipedia.org/wiki/NCIS_(season_13). [Accessed June 2019].

[8] en.wikipedia.org, "NCIS Season 14," [Online]. Available: https://en.wikipedia.org/wiki/NCIS_(season_14). [Accessed June 2019].

[9] en.wikipedia.org, "NCIS Season 15," [Online]. Available: https://en.wikipedia.org/wiki/NCIS_(season_15). [Accessed June 2019].

[10] en.wikipedia.org, "NCIS Season 16," [Online]. Available: https://en.wikipedia.org/wiki/NCIS_(season_16). [Accessed June 2019].