Visualisation of Top 1000 IMDB Movies

Jirat Manpadungkit

Abstract --

I. Introduction

```
> glimpse(imdb1000)
Rows: 1,000
Columns: 16
$ Poster_Link   <chr> "https://m.media-amazon.com/images/M/MV5BMDFkYTc0MGEtZmNhMC00ZDIzLWFmNTEtODM1ZmRlYWMwM…
$ Series_Title  <chr> "The Shawshank Redemption", "The Godfather", "The Dark Knight", "The Godfather: Part I…
$ Released_Year <chr> "1994", "1972", "2008", "1974", "1957", "2003", "1994", "1993", "2010", "1999", "2001"…
$ Certificate   <chr> "A", "A", "UA", "A", "U", "U", "A", "A", "UA", "A", "U", "UA", "A", "UA", "A", "A", "U…
$ Runtime       <chr> "142 min", "175 min", "152 min", "202 min", "96 min", "201 min", "154 min", "195 min",…
$ Genre         <chr> "Drama", "Crime, Drama", "Action, Crime, Drama", "Crime, Drama", "Crime, Drama", "Acti…
$ IMDB_Rating   <dbl> 9.3, 9.2, 9.0, 9.0, 9.0, 8.9, 8.9, 8.9, 8.8, 8.8, 8.8, 8.8, 8.8, 8.7, 8.7, 8.7, 8.7, 8…
$ Overview      <chr> "Two imprisoned men bond over a number of years, finding solace and eventual redemptio…
$ Meta_score    <int> 80, 100, 84, 90, 96, 94, 94, 94, 74, 66, 92, 82, 90, 87, 73, 90, 82, 83, 90, 96, NA, 7…
$ Director      <chr> "Frank Darabont", "Francis Ford Coppola", "Christopher Nolan", "Francis Ford Coppola",…
$ Star1         <chr> "Tim Robbins", "Marlon Brando", "Christian Bale", "Al Pacino", "Henry Fonda", "Elijah …
$ Star2         <chr> "Morgan Freeman", "Al Pacino", "Heath Ledger", "Robert De Niro", "Lee J. Cobb", "Viggo…
$ Star3         <chr> "Bob Gunton", "James Caan", "Aaron Eckhart", "Robert Duvall", "Martin Balsam", "Ian Mc…
$ Star4         <chr> "William Sadler", "Diane Keaton", "Michael Caine", "Diane Keaton", "John Fiedler", "Or…
$ No_of_Votes   <int> 2343110, 1620367, 2303232, 1129952, 689845, 1642758, 1826188, 1213505, 2067042, 185474…
$ Gross         <chr> "28,341,469", "134,966,411", "534,858,444", "57,300,000", "4,360,000", "377,845,905", …
```

Fig. 1. Description of the "imdb1000" data frame showing number of rows, number of columns, column names, datatypes in each column and the first few rows of values.

II. Data

A. *Description of Data*

The main data source in this visualisation project is the data frame called "imdb1000" loaded from a file named "imdb_top_1000.csv" [1]. This data frame contains exactly 1000 rows and 16 features, namely *Poster_Link, Series_Title, Released_Year, Certificate, Runtime, Genre, IMDB_Rating, Overview, Meta_score, Director, Star1, Star2, Star3, Star4, No_of_Votes, Gross*, as shown in Figure 1. All columns except "Meta_score" and "Gross" have no null values. The "Meta_score" column contains 157 NAs and the "Gross" column contains 169 blank strings.

B. *Data Cleaning and Transformation*

As illustrated in Figure 1, the datatype of *Released_Year, Runtime* and *Gross* are characters. Additionally, *Runtime* and *Gross* contain non-numerical characters, which are " min" and "," respectively. Thus, these non-numerical characters were first removed, and their datatypes were then changed to integer.

Through this process, one value in the *Released_Year* column was converted into NA as the value was initially a string. This row belongs to the movie *Apollo 13*, hence the missing value was simply imputed via a google search of the released year, which was 1995. [2]

To deal with the missing data in *Meta_score* and *Gross* columns, we assumed that these data are missing at random, thus with approximately 15% incompleteness, data imputation was performed for both columns.

III. Image

In 2006, Nicholas U. Mayall Kitt Peak 4m telescope with a MOSAIC camera captured a deep optical charged-coupled device (CCD) image of an extragalactic sky.The data from the CCD image was then extracted as a nested array with the same structure as the image size (2570 4611) px. Each value in the array represents the brightness of that pixel, with 1 unit of pixel value (ADU) generated by 3.1 counts of electrons in that pixel and read noise of 5.6 electrons, or 1.8 digital count.

The CCD used in this experiment is a 16-bit counter, meaning any counts beyond 2^{16} will cause "overfilling" and "blooming", increasing counts in neighbouring pixels. These artefacts can be seen in

our CCD image (Fig.1a), where vertical streaks represent erroneous pixel values. Furthermore, the detector has a saturation point of around 36,000 ADU where the linearity begins to diverge. Another feature of this image is that the edges and corners are more noisy since they have fewer sub-exposures. [4].

IV. METHOD

A. *Data Pre-processing*

Due to the image's unequal sub-exposures at the corners, it was cropped by 200 px all around, decreasing the data shape to (2170 4211) px. The 1.8 pixel read noise was then eliminated from each pixel. A distribution plot of the cropped data's pixel values counts revealed that the background signal ranged between 3000 and 4000. Local background was then determined within each (100 100) px section of the image by fitting a gaussian function to each local pixel value distribution between 3000 and 4000. The gaussian means represent the local background signals, whereas the sigma, or standard deviation, represents the camera noise. The variation of local backgrounds is depicted in Fig.1b) after the image had been masked, which we will be discussed in the next paragraph.

The "blooming" and saturated artifacts were identified by their streaks and by checking pixel values above the saturation point, respectively. They were then masked by setting all pixel values that were above 5 sigmas of the local background to zero.

B. *Galaxies Detection*

The galaxies were detected and counted using a SciPy machine learning approach. The hierarchical clustering model is used in this method, and the specific function is *scipy.cluster.hierachy.fclusterdata*[8]. Due to the complexity of this function, data was divided into 7 sections, each having the entire width of the cropped image and a height of 600 px, with the exception of the seventh segment, which has a height of 611 px. The coordinates of all pixels with pixel values larger than 5 sigmas were collected for each section so that they could be provided into the function.

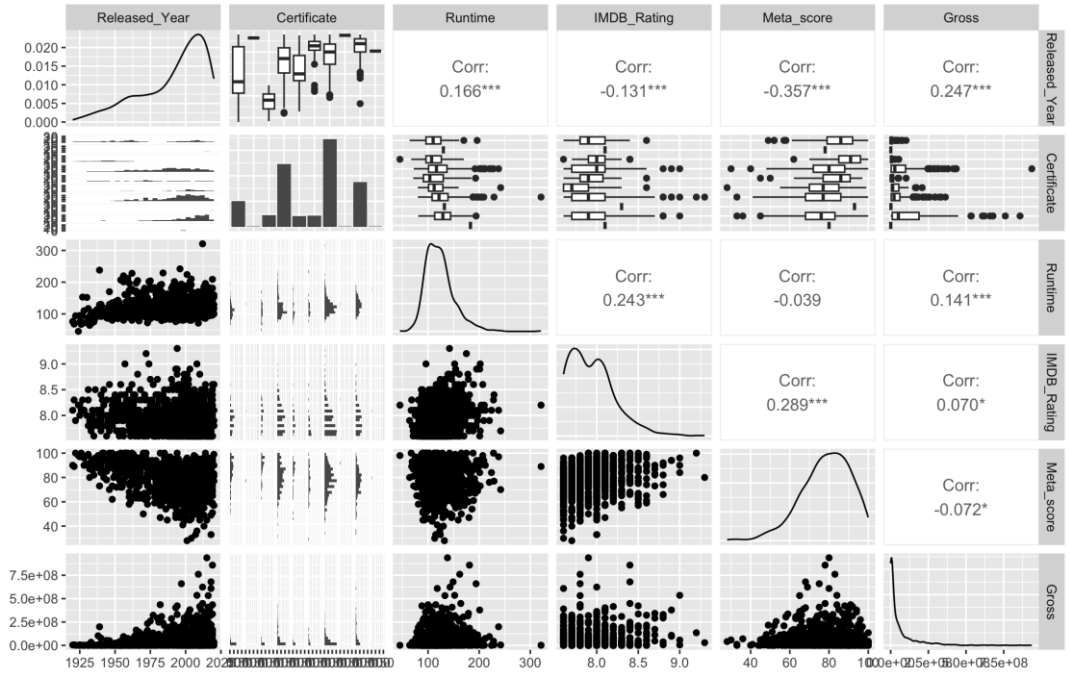

Fig. 1. A) Optical CCD image produced by SAOImageDS9 software b) Variation of backgrounds in different region

Note that 5 sigmas will be set to be the lower limit of the pixel value to be considered as a galaxy, here this relatively small value was used to mask bright artifacts because these artifacts have effects on neighbouring pixels, but not significantly enough to increase their pixel values to the saturation level. If these neighbouring pixels were left unmasked, they will be detected as galaxies later on.

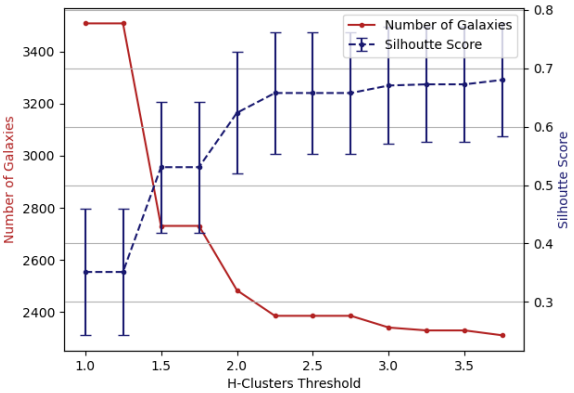

Fig. 2. Mean silhouette score of the 7 data sections shown in blue with standard deviation as the errors. Red plot shows clusters number including clusters with size of 1 and 2 pixels

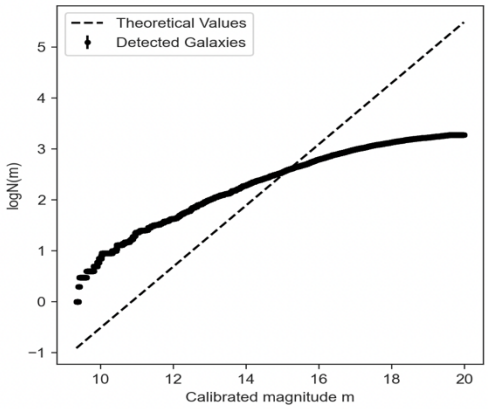

Fig. 3. Plot of counts against magnitude with estimated expectation values

This function utilises a hierarchical clustering model that is agglomerative, this means that it starts by considering each data point (each pixel) as its own separate cluster. At each step, the function identifies the two closest clusters which is determined by the Euclidean distance between two closest points of the two separate clusters (single linkage method). These two clusters then merges. The process of identifying and merging then repeats itself, with the single linkage distance note down at each step, until the entire dataset becomes one cluster, hence the name 'agglomerative'. The differentiation of galaxies occurs when the threshold distance parameter is considered, this distance indicates the upper bound of the single-linkage distance for data points to be considered as one cluster, which was chosen to be 2.25 px in for the final result. The function then returned a list of cluster labels corresponding to each coordinate that was initially passed in. For a better understanding, fig.4 in Appendix A shows a dendrogram produced to represent which clusters were grouped together at what distances apart, with the final clusters determined by the threshold [4,7].

Before the threshold of 2.25 px was finalised, two evaluation methods were used, first by Silhouette score (*sklearn.metrics.silhouette_score*[9]*)* and second by overserving the cluster numbers. Silhouette score is defined by $\frac{a(i)-b(i)}{\max(a(i),b(i))}$, where $a(i)$ is the mean distance of point i to all other points in its cluster, and $b(i)$ is the mean distance of i to all the points in the nearest cluster, which ranges from -1 to 1, where a score close to 1 means that i is more similar to its cluster than the others. The total silhouette score is the average score of all sample i. [5,7] Silhouette scores of different clustered datasets produced with different thresholds were calculated and shown in Fig.3. The score rises and plateaus at 2.25 px indicating that further clustering models do not improve the clustering results much. To prevent misidentifying multiple galaxies as one, threshold 2.25 px makes the best candidate. Another method was done by plotting the number of clusters against threshold. Decreasing the threshold from 2.25 px shows a drastic increase in the number of clusters indicating that there are many misidentification of points as being its own galaxy. This is true if we assume the most galaxies are more than 0.5 px apart, thus making 2.25 px the chosen threshold.

V. Results and Discussion

After running the clustering model, cluster labels of each section were combined and the total number of galaxies was found , where galaxies of sizes 1 and 2 pixels discarded (Table I). The galaxy magnitudes of all galaxies were also calculated, then the local background was subtracted from the galaxies magnitudes. Using Eqn.2, magnitude were calibrated, and the number of galaxies with magnitude brighter than m were counted as a function of m. Log base 10 of the counts was taken and plotted against the calibrated magnitude, illustrated by Fig.3, with a dotted line showing a fitted linear curve with gradient of 0.6, Eqn.1. Three straight lines with no fixed gradient were then fitted to the Log plot, and the values of the estimated gradients are summarised in Table I.

The first section of count plot gives a gradient value 49.3% larger than 0.6, whilst the third section

TABLE I
UNCERTAINTIES

	Value	% Difference from Expectation
Total number of galaxies	1915	-
Gradient in each magnitude range		
9.33 – 20.0 (whole)	0.304	49.3
9.33 – 10.3	0.860	- 43.3
10.3 – 16.0	0.309	48.5
16.0 – 20.0	0.121	79.8

TABLE I. Table summarising the number of galaxies detected with size above 2 pixels and with distance threshold of 2.25, and the gradient values estimated from curve fitting Eqn.1 to the logN(m) plot for difference magnitude range. Errors are in the order of 10^{-4} to 10^{-7}.

of the plot has a slope of 79.8% difference smaller than 0.6. These two deviations may be explained by the fact that distant galaxies are observed to be brighter and denser than they presently are, due to the finite speed of light and the evolution of galaxies [2,3,4]. This goes against the assumption made about the uniform universe. The tailing-off characteristic of the curve may also be due to the fact that dimmer galaxies are harder to be detected, and they may have values below 5 sigmas, thus they were not considered in the data processing.

VI. CONCLUSIONS

This experiment involves the processing of a CCD image taken from an extragalactic field and resulted in a relation between counts of galaxies and their magnitudes. The relationship was found to have some deviation from the theoretical values, specifically the gradients for different parts of the plot. These deviations explained by the fault the assumption made about the uniform universe, and also the experimental limitations of detection dim galaxies.

Note that 5 sigmas will be set to be the lower limit of the pixel value to be considered as a galaxy, here this relatively small value was used to mask bright artifacts because these artifacts have effects on neighbouring pixels, but not significantly enough to increase their pixel values to the saturation level. If these neighbouring pixels were left unmasked, they will be detected as galaxies later on.

REFERENCES

[1] Shankhdhar, H. (2021) *IMDB movies dataset*, *Kaggle*. Available at: https://www.kaggle.com/datasets/harshitshankhdhar/imdb-dataset-of-top-1000-movies-and-tv-shows (Accessed: March 20, 2023).

[2] *Apollo 13* (1995) *IMDb*. IMDb.com. Available at: https://www.imdb.com/title/tt0112384/ (Accessed: April 20, 2023).

[3] Tyson, J.A. and Jarvis, J.F., 1979. Evolution of galaxies-Automated faint object counts to 24th magnitude. *The Astrophysical Journal*, *230*, pp.L153-L156.

[4] D.L. Clements, N. Skrzypek, Y Unruh, 3rd Year Lab Module, Version 1.5: Revised September 2021, Imperial College

[5] Alpaydin, E., 2004. Introduction to Machine Learning (Adaptive Computation and.

[6] Saleh, H., 2020. *The The Machine Learning Workshop: Get ready to develop your own high-performance machine learning algorithms with scikit-learn*. Packt Publishing Ltd.

[7] V. Pauli, G. Ralf, et al**.,** SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17(3), 261-272

[8] Eads, D. (ed.) (2007) *Scipy.cluster.hierarchy.fclusterdata#*, *scipy.cluster.hierarchy.fclusterdata - SciPy v1.9.3 Manual*. Available at: https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.fclusterdata.html#scipy.cluster.hierarchy.fclusterdata (Accessed: November 15, 2022).

[9] Layton, R., Fouchet, A. and Guillemot, T. (no date) *Sklearn.metrics.silhouette_score*, *scikit*. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html (Accessed: November 20, 2022).

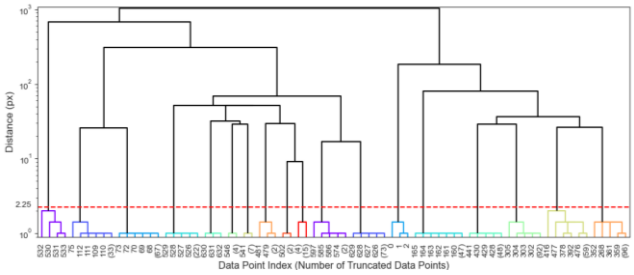

Fig. 4. Dendrogram showing an example of the hierarchical clustering of a small section of data, where the y-axis represents the single-linkage distance, x-axis represents the data points indices, the numbers in brackets are the additional data points in the same cluster truncated for simplicity.

APPENDIX A

APPENDIX B

Feedback from previous assessment: Good pace of presentation and good formatting of slides. They do explain well the context and the motivation of the work. Extended work is undertaken in data analysis, students investigating various ways to obtain the desired results. Introducing assumptions before experiment is unusual, better to mention that it is about a plucked string at the beginning, if not the entire set-up There is a discrepency with the model in your results, which can be explained easily with the quality of Lorentzian fitiing procedure, instead of questioning the model. Slide 13 would be more convincing with some fits that show that the model holds

Note that 5 sigmas will be set to be the lower limit of the pixel value to be considered as a galaxy, here this relatively small value was used to mask bright artifacts because these artifacts have effects on neighbouring pixels, but not significantly enough to increase their pixel values to the saturation level. If these neighbouring pixels were left unmasked, they will be detected as galaxies later on.