

Data Storytelling Final Project

Mitchell Beckner

3/28/2020

Author

Mitchell Beckner

Links

Final Project Markdown on GitHub: https://github.com/mitchb63/MovieLens_Analysis/blob/master/Beckner_Udacity_Storytelling_Final.pdf

Presentation Deck .pptx on GitHub: https://github.com/mitchb63/MovieLens_Analysis/blob/master/Beckner_MovieLens_Analysis_final.pptx

Data Cleaning and EDA in R: https://github.com/mitchb63/MovieLens_Analysis

Summary

This project was created as part of Udacity's Data Visualization Nanodegree program. The data used is from a Kaggle dataset that contains metadata for 45,000 movies listed in the MovieLens Dataset. The dataset consists of movies released on or before July 2017. Data points include cast, crew, plot keywords, budget, revenue, posters, release dates, languages, production companies, countries, TMDB vote counts and vote averages. The dataset also has 26 million ratings from 270,000 users for all 45,000 movies (Banik, 2017).

The objective of this project was to use the movies dataset and conduct the EDA necessary to understand the dataset as a whole. The output will be to discover if the dataset is balanced, if there are anomalies in the dataset that affect the applicability of the recommendation, and the final presentation itself that will be used for a mock recommendation to a management team (Udacity, 2020).

The problem statement for the project was defined as:

What are the key components that differentiate the top 20% of films from the remainder when considering profit margin?

Three hypotheses were constructed as the foundation of the study: * Story factors influence the film's profit margin * Production factors influence the profit margin * Public opinion factors affect the profit margin

The key findings of the analysis were that Story factors and budget appear to have the greatest influence on the profit margin of a film and the final recommendations were to focus on the production and promotion of films that are part of an ongoing collection and/or films with budgets less than approximately \$6 million. These movies were found to be more likely to have profit margins in the top 20% of those examined. The analytical process used, as well as the limitations and biases in the dataset, are described in the sections that follow.

Analysis Process

The primary dataset was downloaded from the Kaggle website (Banik, 2017). It consisted of 7 data files. Additional data was downloaded from the IMDB website that consisted of lists of the highest rated actors, producers and directors (nims-1975, 2015)(Smmsadrnezh, 2018). This information was used as reference to create binary variables for ‘use of top-rated cast’ and ‘use of top-rated crew’ for each film.

Some preliminary data preparation and cleaning was done in Excel before importing the data into R for further processing. Many of the “0” values were replaced with NA’s in order to accurately account for, and later impute the missing data. Duplicate rows were also discovered and removed. A summary of the distributions of the variables in the dataset is shown in Figure 1.

```
> summary(df_movies_init)
```

id	budget	original_language	popularity	release_date	revenue	runtime
2	1	Min. : 1	Min. : 0.000	Min. : 1874-12-09	Min. : 1	Min. : 1.0
3	1	1st Qu.: 2000000	1st Qu.: 0.386	1st Qu.: 1978-10-06	1st Qu.: 2400000	1st Qu.: 86.0
5	1	Median : 8000000	Median : 1.127	Median : 2001-08-30	Median : 16829545	Median : 95.0
6	1	Mean : 21614181	Mean : 2.921	Mean : 1992-05-16	Mean : 68489031	Mean : 97.5
11	1	3rd Qu.: 25000000	3rd Qu.: 3.678	3rd Qu.: 2010-12-17	3rd Qu.: 67308282	3rd Qu.: 107.0
(other):45428	Max. : 380000000	(other): 6781	Max. : 547.488	Max. : 2020-12-16	Max. : 2068223624	Max. : 1256.0
NA's : 1	NA's : 36554	NA's : 11	NA's : 4	NA's : 88	NA's : 38037	NA's : 1819
status	title	vote_average	vote_count	release_month	release_year	
Canceled : 2	Cinderella : 11	Min. : 0.50	Min. : 0	Jan : 5909	2014 : 1973	
In Production : 20	Alice in Wonderland : 9	1st Qu.: 5.30	1st Qu.: 3	Sep : 4834	2015 : 1904	
Planned : 15	Hamlet : 9	Median : 6.10	Median : 10	Oct : 4613	2013 : 1887	
Post Production: 98	Beauty and the Beast: 8	Mean : 6.01	Mean : 110	Dec : 3781	2012 : 1721	
Released : 44985	Les Misérables : 8	3rd Qu.: 6.90	3rd Qu.: 34	Nov : 3661	2011 : 1666	
Rumored : 229	(Other) : 45385	Max. : 10.00	Max. : 14075	(other): 22548	(other): 36195	
NA's : 85	NA's : 4	NA's : 2999	NA's : 4	NA's : 88	NA's : 88	

Figure 1: Raw Data Distribution Summary

Next the dataset was filtered to include only films that had been released to the public and those with budgets less than \$59,500,000. This eliminated the statistical outliers found in the budget variable. New variables for *profit* and *profit_margin* were created and then used to create a *profit_group* categorical variable that placed each movie into either the top 20% or bottom 80% based on profit margin. Finally, R code was written to process the separate files for ratings, keywords, cast and crew data, country and language data, genre, and movie collection information. The full code is available at the link posted above.

Various R packages were used to explore the missingness present within the data. Figure 2 shows a plot of the pattern of missing data and Figure 3 shows the patterns of missingness. It should be noted that while the original dataset contained information on over 45,000 movies, removing duplicates and filtering for ‘Released’ movies with a stated budget greater than 0 resulted in only 7925 films. These films were used as the basis for the remaining investigation.

The ‘MICE’ package was then used to impute the missing values in the dataset (Noghrehchi, 2015). Five imputed datasets were created using 50 iterations of the chained equation process. The results were examined for convergence and found to be satisfactory. A single imputed dataset was then randomly selected and used for the remainder of the analysis.

Following the imputation of missing values, the variables for profit and profit_margin were re-generated and again used to create the profit_group variable that placed each movie into either the top 20% or bottom 80% based on profit margin. A summary of the variable distributions for the imputed dataset are shown in Figure 4. This data was then standardized and correlation analysis was run. An excerpt of the results of this analysis is shown in Figure 5 with the corresponding p-values shown in Figure 6. These results showed significant correlations betw.

Following the correlation analysis, additional EDA was performed in R to confirm the results and lay the groundwork for the final visualizations. Final visualizations were then produced in Tableau and R with some additional polishing done in Adobe Illustrator. The final presentation was then assembled based on the PowerPoint ghost deck created in a prior project.

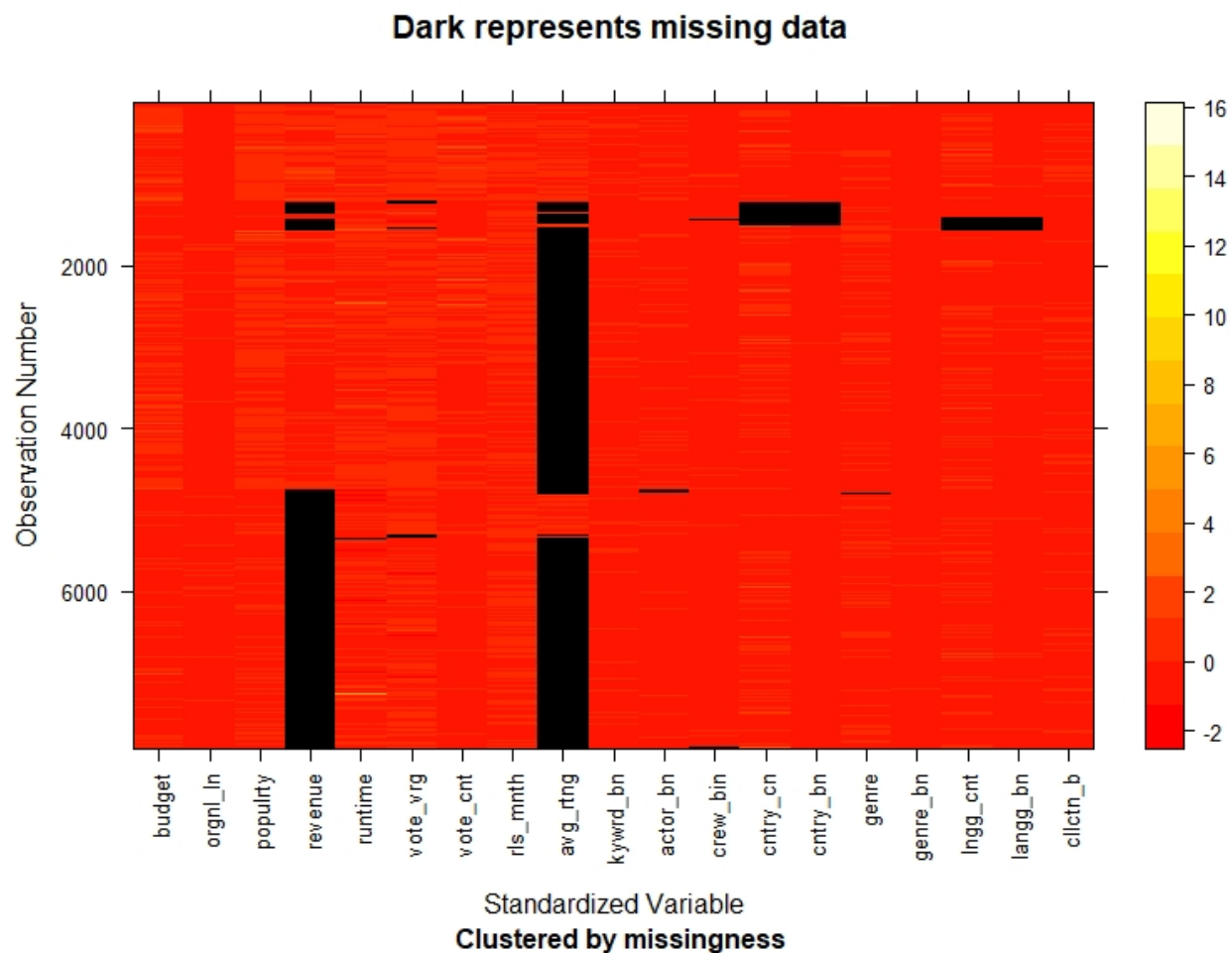


Figure 2: Missing Data as Shown in MI

```
> show(missing_bins)
Object of class missing_data.frame with 7925 observations on 20 variables

There are 95 missing data patterns

Append '@patterns' to this missing_data.frame to access the corresponding pattern for every observation or perhaps use table()

  id      type missing method model
budget      continuous      0 <NA> <NA>
original_language unordered-categorical 0 <NA> <NA>
popularity    continuous      0 <NA> <NA>
revenue      continuous 3435  ppd linear
runtime      continuous   66  ppd linear
vote_average  continuous   90  ppd linear
vote_count    continuous    0 <NA> <NA>
release_month unordered-categorical    4  ppd mlogit
avg_rating    continuous 6151  ppd linear
keyword_bin   binary       0 <NA> <NA>
actor_bin     binary      77  ppd logit
crew_bin      binary      37  ppd logit
country_count continuous   286  ppd linear
country_bin   binary    286  ppd logit
genre         unordered-categorical   57  ppd mlogit
genre_bin     binary      0 <NA> <NA>
language_count continuous  165  ppd linear
language_bin  binary    165  ppd logit
collection_bin binary      0 <NA> <NA>
```

Figure 3: Missing Data Summary

```

> summary(df_imp_1)
  id      budget original_language popularity revenue runtime vote_average vote_count
5   : 1    Min.   : 1             en      :6497    Min.   : 0.00    Min.   : 1    Min.   : 4    Min.   : 0.50    Min.   : 0
11  : 1    1st Qu.: 1525000        fr      : 234    1st Qu.: 1.84    1st Qu.: 1477676 1st Qu.: 91    1st Qu.: 5.40    1st Qu.: 17
13  : 1    Median : 6229577        ru      : 136    Median : 5.62    Median : 9366227 Median :100    Median : 6.20    Median : 71
14  : 1    Mean   :12048609        hi      : 130    Mean   : 6.33    Mean   : 32164871 Mean  :105    Mean   : 6.06    Mean   : 298
15  : 1    3rd Qu.:19000000        es      : 113    3rd Qu.: 9.19    3rd Qu.: 33697647 3rd Qu.:114    3rd Qu.: 6.80    3rd Qu.: 285
16  : 1    Max.   :59000000        de      : 94     Max.   :228.03    Max.   :792965326 Max.   :523    Max.   :10.00    Max.   :11444
(other):7919          (other): 721
release_month avg_rating keyword_bin actor_bin crew_bin country_count country_bin genre genre_bin language_count
Sep   : 958    Min.   :0.50    0:7139    0:6920    0:7787    Min.   :1.00    0:7825    Drama  :2092    0:7887    Min.   :1.0
Oct   : 830    1st Qu.:2.83    1: 786    1:1005    1: 138    1st Qu.:1.00    1: 100    Comedy :1590    1: 38    1st Qu.:1.0
Dec   : 714    Median :3.24          Median :1.00          Median :1.00    Action  :1174          Median :1.0
Jan   : 705    Mean   :3.15          Mean   :1.34          Mean   :1.34    Horror  : 710          Mean   :1.4
Aug   : 674    3rd Qu.:3.56          3rd Qu.:1.00          3rd Qu.:1.00    Crime   : 383          3rd Qu.:2.0
Apr   : 618    Max.   :4.50          Max.   :5.00          Max.   :5.00    Adventure: 375          Max.   :5.0
(other):3426
language_bin collection_bin profit profit_margin profit_group
0:7587    0:6689    Min.   :-53223667    Min.   :~1999999    Bottom_80:6340
1: 338    1:1236    1st Qu.: -1862635    1st Qu.: -1    Top_20 :1585
          Median : 1923000    Median : 0
          Mean   : 20116262    Mean   : -4719
          3rd Qu.: 18907422    3rd Qu.: 1
          Max.   : 782465326    Max.   : 1

```

Figure 4: Imputed Data Distribution Summary

```

> corr <- round(cor(df_imp1_std, use="pairwise.complete.obs"), 6)
> corr

```

	profit_group	profit_margin	budget	popularity	revenue	runtime
profit_group	1.000000	0.037199	-0.238984	0.045164	0.278838	-0.1010
profit_margin	0.037199	1.000000	0.044505	0.042649	0.038395	0.0133
budget	-0.238984	0.044505	1.000000	0.352454	0.482830	0.2048
popularity	0.045164	0.042649	0.352454	1.000000	0.422242	0.1111
revenue	0.278838	0.038395	0.482830	0.422242	1.000000	0.1409
runtime	-0.101066	0.013334	0.204837	0.111162	0.140948	1.0000
vote_average	0.101696	0.030254	0.041441	0.233059	0.184336	0.2569
vote_count	0.150940	0.031343	0.354916	0.592744	0.650708	0.1130
avg_rating	0.022407	-0.011500	-0.023558	0.016004	0.033491	0.0344
keyword_bin	0.057835	0.002307	0.027240	0.080756	0.088348	0.0060
actor_bin	-0.048348	0.026336	0.277909	0.182051	0.171968	0.1397
crew_bin	0.017846	0.009899	0.077829	0.093790	0.134254	0.0731
country_count	-0.092008	0.003703	0.114902	0.069720	0.004748	0.0940
country_bin	0.005652	0.002186	-0.038108	-0.030632	-0.029949	0.0140
genre_bin	-0.007307	-0.000057	-0.037219	-0.031921	-0.019363	-0.0032
language_count	-0.051013	-0.003741	0.116051	0.094269	0.065350	0.1637
language_bin	0.005308	-0.016623	0.003511	0.013190	0.012173	0.0702
collection_bin	0.120681	0.023916	0.089118	0.170178	0.255433	-0.0508
profit	0.362190	0.031314	0.288230	0.377213	0.977716	0.1050

Figure 5: Data Correlation Summary


```

> p.mat <- round(cor_pmat(df_imp1_std), 6)
> p.mat

```

	profit_group	profit_margin	budget	popularity	revenue	runtime
profit_group	0.000000	0.000926	0.000000	0.000058	0.000000	0.000000
profit_margin	0.000926	0.000000	0.000074	0.000146	0.000629	0.235288
budget	0.000000	0.000074	0.000000	0.000000	0.000000	0.000000
popularity	0.000058	0.000146	0.000000	0.000000	0.000000	0.000000
revenue	0.000000	0.000629	0.000000	0.000000	0.000000	0.000000
runtime	0.000000	0.235288	0.000000	0.000000	0.000000	0.000000
vote_average	0.000000	0.007071	0.000224	0.000000	0.000000	0.000000
vote_count	0.000000	0.005263	0.000000	0.000000	0.000000	0.000000
avg_rating	0.046082	0.305994	0.035983	0.154271	0.002865	0.002865
keyword_bin	0.000000	0.837279	0.015306	0.000000	0.000000	0.588279
actor_bin	0.000017	0.019051	0.000000	0.000000	0.000000	0.000000
crew_bin	0.112152	0.378264	0.000000	0.000000	0.000000	0.000000
country_count	0.000000	0.741735	0.000000	0.000000	0.672550	0.000000
country_bin	0.614887	0.845717	0.000691	0.006388	0.007669	0.209691
genre_bin	0.515464	0.995930	0.000920	0.004484	0.084773	0.774920
language_count	0.000006	0.739171	0.000000	0.000000	0.000000	0.000000
language_bin	0.636603	0.138957	0.754675	0.240353	0.278564	0.000000
collection_bin	0.000000	0.033252	0.000000	0.000000	0.000000	0.000000
profit	0.000000	0.005305	0.000000	0.000000	0.000000	0.000000

Figure 6: p-Values for Correlated Data

Limitations and Biases

###Data Collection Phase

Many problems were found in the provided dataset. As noted above, the data contained duplicate rows, and a large amount of missing data. In addition, some the data that was present was obviously erroneous and/or suspect. For example, some movies had negative values listed as their budget. Without more information on the methods of data collection and their sources, it is impossible to determine the reasons for these errors and omissions, therefore, assumptions that were made regarding the data being missing completely at random are essentially indefensible. It is also likely that a great deal of response bias is present in the popularity, vote average and rating data since this information was presumably collected at various lengths of time after a film was released.

Data Processing Phase

Again, as noted above, the MICE package was used to impute missing data in R. While every attempt was made to select the best available algorithm for each variable, and results were visually examined to ensure that the imputed values were ‘logically reasonable, there is the potential for a great deal of bias being introduced at this point in the process. In addition, in an attempt to use the cast and crew data that was provided, additional lists of “top-rated” actors, producers, and directors were obtained from the IMDB website. Each movies cast and crew was checked against these lists and movies were flagged if they used actors or crew members on these outside lists. However, there is no standard criteria that was used to establish the “top-rated” status of these people so again, bias was likely introduced at this point in the analysis.

Insight Phase

While every attempt was made to verify results and recommendations by multiple methods including correlation analysis, creation of visualizations, and careful analysis of potential biases, it is possible that confirmation

bias is present to a degree in the findings. In addition, were a model to be created in order to predict profit margin based on the variables determined to be significant, it is possible that the model would be over or under fitted given the relatively small percentage of the dataset that began with complete information.

References

- Anonymous. (1970, January 1). McKinsey Presentations - How to Apply Ghost (aka Shell and Skeleton) Decks and Pages. Retrieved March 7, 2020, from <http://workingwithmckinsey.blogspot.com/2013/07/McKinsey-presentations-ghost-decks.html>
- Banik, R. (2017, November 10). The Movies Dataset. Retrieved March 1, 2020, from https://www.kaggle.com/rounakbanik/the-movies-dataset#movies_metadata.csv
- nims-1975. (2015, July 2). Top 12 Best Film Producers/Directors. Retrieved from <https://www.imdb.com/list/ls051116454/>
- Noghrehchi, F. (2015, May 29). Retrieved from <https://web.maths.unsw.edu.au/~dwarton/missingDataLab.html>
- Smmsadrnezh. (2018, March 5). Best and Worst Actors Rating (Sorted). Retrieved from <https://www.imdb.com/list/ls070682726/>
- Udacity. (2020). Build a Data Story Final Project. Retrieved from <https://classroom.udacity.com/nanodegrees/nd197/parts/c26016f6-de16-42b7-98f8-9de7374f8247/modules/2c1ae2e0-f442-4268-9931-c8b2d0c07ea2/lessons/060138e7-74ee-4391-8ced-2f1fbcca002c/concepts/1e50f104-b96f-462a-a408-a155ea4a2467>