Module 1 | Technique Practice

Trang Tran

CPS, Northeastern University

ALY6040 | Data Mining

Winnie Li

Apr 23, 2023

Introduction

We are looking at a dataset from IMDb that contains information for titles of the history genre only. The dataset has 14 variables with 30,054 observations. The variables are "tconst" (title ID), "averageRating" (weighted average of all the individual user ratings), "numVotes" (number of votes the title has received), "titleType" (the type/format of the title), "primaryTitle" (the more popular title), "originalTitle" (original title, in the original language), "isAdult" (0: non-adult title; 1: adult title), "startYear" (represents the release year of a title), "endYear" (TV Series end year. '\N' for all other title types), "runtimeMinutes" (primary runtime of the title), "genres", "parentTconst" (alphanumeric identifier of the parent TV Series), "seasonNumber" (season number the episode belongs to), and "episodeNumber" (episode number of the tconst in the TV series).

Descriptive Statistics

```
──────────────────────
Column type frequency:
  character               11
  numeric                 3
──────────────────────
Group variables           None

── Variable type: character ──────────────────────────────────────────────────────────────────────
   skim_variable  n_missing complete_rate min max empty n_unique whitespace
 1 tconst                 0         1       9  10     0    30054          0
 2 titleType              0         1       5  12     0       10          0
 3 primaryTitle           0         1       1 150     0    26492          0
 4 originalTitle          0         1       1 150     0    26573          0
 5 startYear              0         1       2   4     0      125          0
 6 endYear                0         1       2   4     0       67          0
 7 runtimeMinutes         0         1       1   4     0      404          0
 8 genres                 0         1       7  30     0      218          0
 9 parentTconst       14172         0.528   9  10     0     1388          0
10 seasonNumber       14172         0.528   1   4     0       45          0
11 episodeNumber      14172         0.528   1   3     0      141          0

── Variable type: numeric ────────────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate       mean         sd p0 p25  p50  p75    p100 hist
1 averageRating         0             1    7.35        1.10    1 6.7  7.4  8.1      10 ▁▁▂▇▅
2 numVotes              0             1 1139.      15770.      5 9   21   87   1244496 ▇▁▁▁▁
3 isAdult               0             1    0.000765    0.0277  0 0    0    0        1 ▇▁▁▁▁
```

Figure 1: Summary statistics using [skim] function

I utilized the [skim] function (Figure 1) to obtain an overview of the dataset. Figure 1 shows that the dataset comprises 11 character variables and 3 numeric variables. However, the variable

"isAdult" is actually categorical data disguised as a numeric variable. On the other hand, the variable "runtimeMinutes" is a numeric variable that was mistakenly classified as a character variable, so I converted them to the right classes later. The table also reveals that three variables ("parentTconst", "seasonNumber", and "episodeNumber") have 14,172 missing values each, accounting for almost 50% of the total number of rows. Yet, this level of missing data is expected and reasonable given the nature of these variables. The dataset does not contain any duplicates, as confirmed by the [anyDuplicated] function. I will conduct further investigation on the "isAdult", "endYear", and "startYear" variables.

```
> # frequency of isAdult
> table(df$isAdult)

    0     1
30031    23
```

The number of adult titles (1) is insignificant compared to the number of observations in this data file. We can either ignore or remove it from the dataset.

```
> # frequency of endYear
> table(df$endYear)

  \\N  1950  1951  1957  1958  1960  1961  1963  1964  1965  1966  1967  1968  1969  1970  1971  1972
28839     1     1     1     1     2     3     2     2     2     2     1     2     2     4     3     6
 1973  1974  1975  1976  1977  1978  1979  1980  1981  1982  1983  1984  1985  1986  1987  1988  1989
    9     7     4     9     7    14    10     9    12    11     6    19    14    10     8     9     7
 1990  1991  1992  1993  1994  1995  1996  1997  1998  1999  2000  2001  2002  2003  2004  2005  2006
    7     5     8     6    11     8    13    12    16     9    15    15    12    23    16    24    19
 2007  2008  2009  2010  2011  2012  2013  2014  2015  2016  2017  2018  2019  2020  2021  2022
   30    33    36    21    36    31    48    41    58    59    91    96   101    79    35     1
```

As seen in the frequency table of the variable “endYear” above, there are 28,839 NAs (\\N) that belong to non-TV Series. We can have choices to deeper analyze sub-groups of this dataset. In the

```
> # frequency of startYear
> table(df$startYear)

 \\N 1895 1897 1898 1899 1900 1901 1902 1904 1906 1907 1908 1909 1910 1911 1912 1913
   1    2    3    1    5    3    3    2    4    3    1    9    9    9    7    6   12
1914 1915 1916 1917 1918 1919 1920 1921 1922 1923 1924 1925 1926 1927 1928 1929 1930
  10   13   12   10    8    6    4    7   10    6   11    6   11   14   20   11   13
1931 1932 1933 1934 1935 1936 1937 1938 1939 1940 1941 1942 1943 1944 1945 1946 1947
   5   10   13   22   28   39   41   41   42   31   30   37   35   29   24   20   26
1948 1949 1950 1951 1952 1953 1954 1955 1956 1957 1958 1959 1960 1961 1962 1963 1964
  22   32   33   46   41   58   41   72   51   54   47   49   59   59   45   48   94
1965 1966 1967 1968 1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981
  65   60   58   63  124  114  112  144  135  148  122  134  170  195  136  142  143
1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998
 196  170  173  123  113  128  246  247  218  199  167  139  169  251  207  299  219
1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
 362  373  382  374  477  602  599  741  792  936 1074 1302 1172 1512 1418 1598 1598
2016 2017 2018 2019 2020 2021
1552 1556 1596 1511 1125  497
```

frequency table of the variable “startYear”, there is only 1 row showing \\N. Besides, this variable should be changed into a numeric variable.

I have a further analysis on histograms of the “averageRating”, “numVotes”, and “titleType” variables. Figure 2 shows a left-skewed distribution of the average of all the individual user ratings. While Figure 3 reveals that the majority of the votes the title has received is under 50,000.

Figure 2: Histogram of averageRating

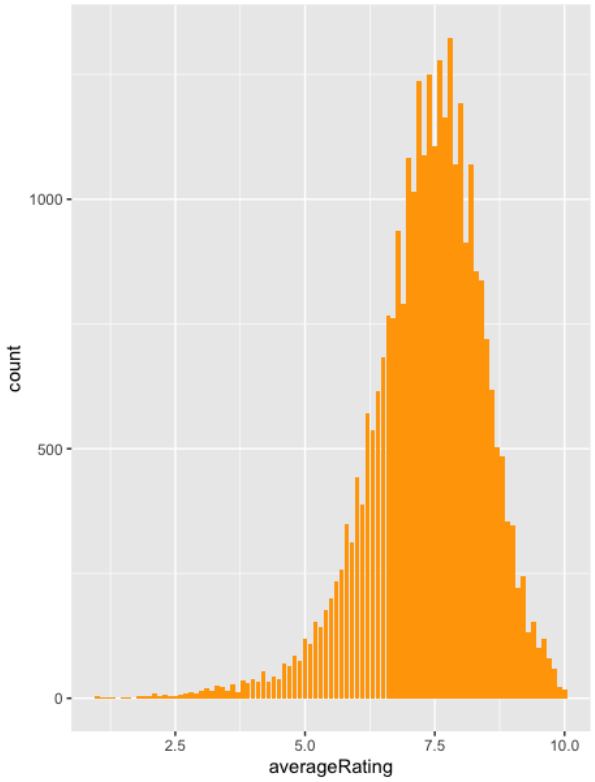
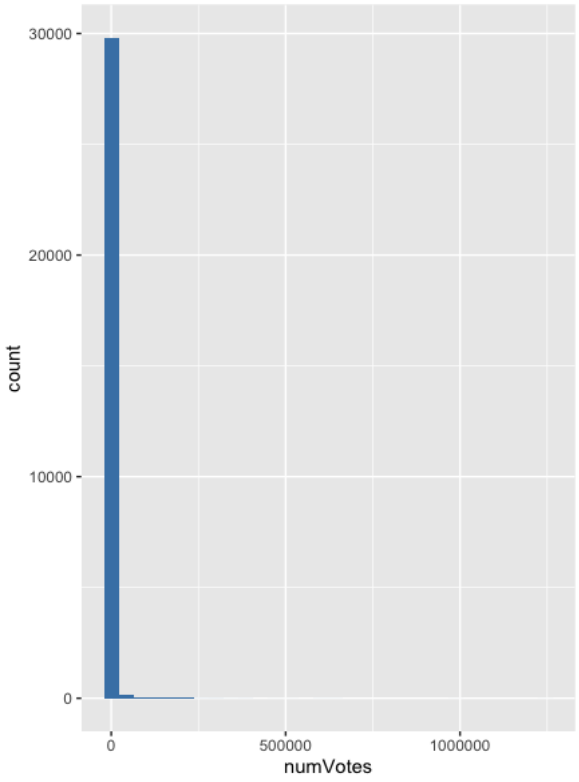


Figure 3: Histogram of numVotes



Data Pre-processing

After converting “isAdult” into the character type, also “startYear” and “runtimeMinutes” into the numeric type, I imputed the mean in all numeric columns. As we started to understand the

data, I built the correlation matrix to see the strength and direction of the linear relationship between numeric variables (Figure 4). The relationships between these variables are not robust.

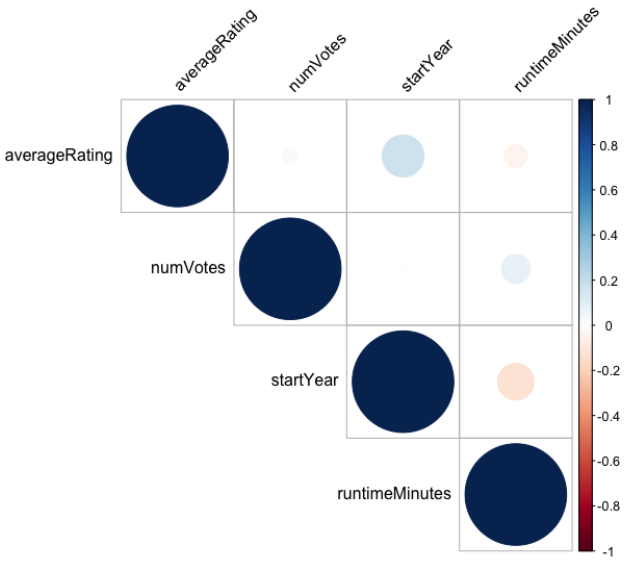


Figure 4: Correlation Matrix

Proposal for next steps

1. Consider sub-group analysis on the "endYear" variable, focusing on the 28,839 NAs that belong to non-TV series titles.
2. Consider conducting a deeper analysis of the "titleType" variable, potentially using clustering or classification techniques.
3. Explore any potential relationships between the variables using other statistical methods, such as regression.
4. Explore the trends in average ratings, number of votes, and runtime for titles of this history genre.
5. Identify the top-rated or most popular titles in this dataset, and further investigate the factors that may have contributed to their success.
6. Cleansing the data would be to address the missing data in the "parentTconst", "seasonNumber", and "episodeNumber" variables. Since these variables represent information specific to TV series, it may be appropriate to impute missing values with 0 or "not applicable" rather than the mean.

References

1. IMDb Datasets. Retrieved April 23, 2023 from. https://www.imdb.com/interfaces/