CS 638 Project Stage 2 Report

Foo Zhi Yuan, zfoo@wisc.edu, Wayne Chew Ming Chan, mchew2@wisc.edu, Yi Ding, ding72@wisc.edu

10/25/2016

Schema:

In order to generate the same schema for both IMDB and rotten tomatoes, we have did some data cleaning on the csv files of both websites, eg:

1 .

In IMDB.csv, convert the duration from hour and minutes to only minutes, then remove the substring "min" in it so that it becomes integers for the ease of data analysis.

2.

In IMDB.csv, remove the substring of "(location)" in all the data under Release Date attribute, then change the release date from the format of "21 October 2016" to the format of "10/21/2016"

3.

In rottentomatoes.csv, convert the release date under date_in_theaters attribute from the format of "Sep 29, 2016" to the format of "09/29/2016

4.

In rottentomatoes.csv, remove the substring of "minutes" in all the data under duration attribute so that it becomes integers for the ease of data analysis

5.

In rottentomatoes.csv, change the format of all the data under audience_rating attribute from "4.1/5" to 4.1 by removing the substring "/5" so that it becomes integers for the ease of data analysis

.

.

.

.

.

.

.

.

.

Table schema for IMDB

id	INT	PRIMARY KEY
Title	VARCHAR	
Category	VARCHAR	
Duration	INT	
Rating	REAL	
Rating Count	INT	
Director	VARCHAR	
Writer	VARCHAR	
Release Date	DATE	
Description	TEXT	

Table schema for rottentomatoes

id	INT	PRIMARY KEY
title	VARCHAR	
genres	VARCHAR	
duration	INT	
audience_rating	REAL	
num_of_audience_rating	INT	
directors	VARCHAR	
writers	VARCHAR	
date_in_theaters	DATE	
movie_description	TEXT	

Attributes in the set:

The attributes are listed below:

```
##  [1] "id"                     "title"
##  [3] "genres"                 "directors"
##  [5] "writers"                "date_in_theaters"
##  [7] "duration"               "audience_rating"
##  [9] "num_of_audience_rating" "movie_description"
```

Solutions to missing value:

Since there are no missing value for atrribute id, title and num_of_audience_rating, we will focus on filling in the missing value for genres, directors, writers, date_in_theaters, duration, audience_rating and movie_description.

Possible solutions to fill in missing values:

1.

Compare the value with other databases. During stage 1 of this project, besides IMDB, we have also crawled another movie reviews website called rottentomatoes. We will compare the values against the data we crawled for rottentomatoes. If the two movies have the same name and date in theaters, we can safely assume that they are the same movie. Then, we can fill in the missing value using information from another table and vice versa.

2.

In addition to this, in the stage 1 of our project, our team has managed to write a generic crawler that can crawl any website, by simply modifying the xpath according to the attributes/ data that we want to crawl. We can use it on other movie reviews websites like Roger Ebert, Guardian, Yahoo! Movies and Meta Critic then compare the values against the data we crawled. If the two movies have the same name and duration, we will assume they are the same movies and fill in the missing value accordingly.

3.

For the missing values of attributes like duration, we will use regression to compute the value.

e.g. We can simply put some attributes as predictors like average audience rating and film categories into regression model (could be linear or polynomial or logistic) and train the model with the data that do not have missing values, then do prediction for the missing values based on other attributes it has in the same tuple.

4.

If the above steps don't work, for the missing values of some attributes like audience rating, we can take the mean, median or the rating with the highest frequency to fill in the space.

Missing Values:

```
## id:   fraction:0/4149,      percentage:0missing
## title:   fraction:0/4149,      percentage:0missing
## genres:   fraction:64/4149,      percentage:0.02missing
## directors:   fraction:72/4149,      percentage:0.02missing
## writers:   fraction:314/4149,      percentage:0.08missing
## date_in_theaters:   fraction:468/4149,      percentage:0.11missing
## duration:   fraction:78/4149,      percentage:0.02missing
## audience_rating:   fraction:417/4149,      percentage:0.1missing
## num_of_audience_rating:   fraction:0/4149,      percentage:0missing
## movie_description:   fraction:69/4149,      percentage:0.02missing
```

Classify attibutes:

id : numeric

title : textual

genres : categorical

duration: numeric

audience_rating: numeric

num_of_audience_rating: numeric

directors : textual

writers: textual

date_in_theaters: numeric

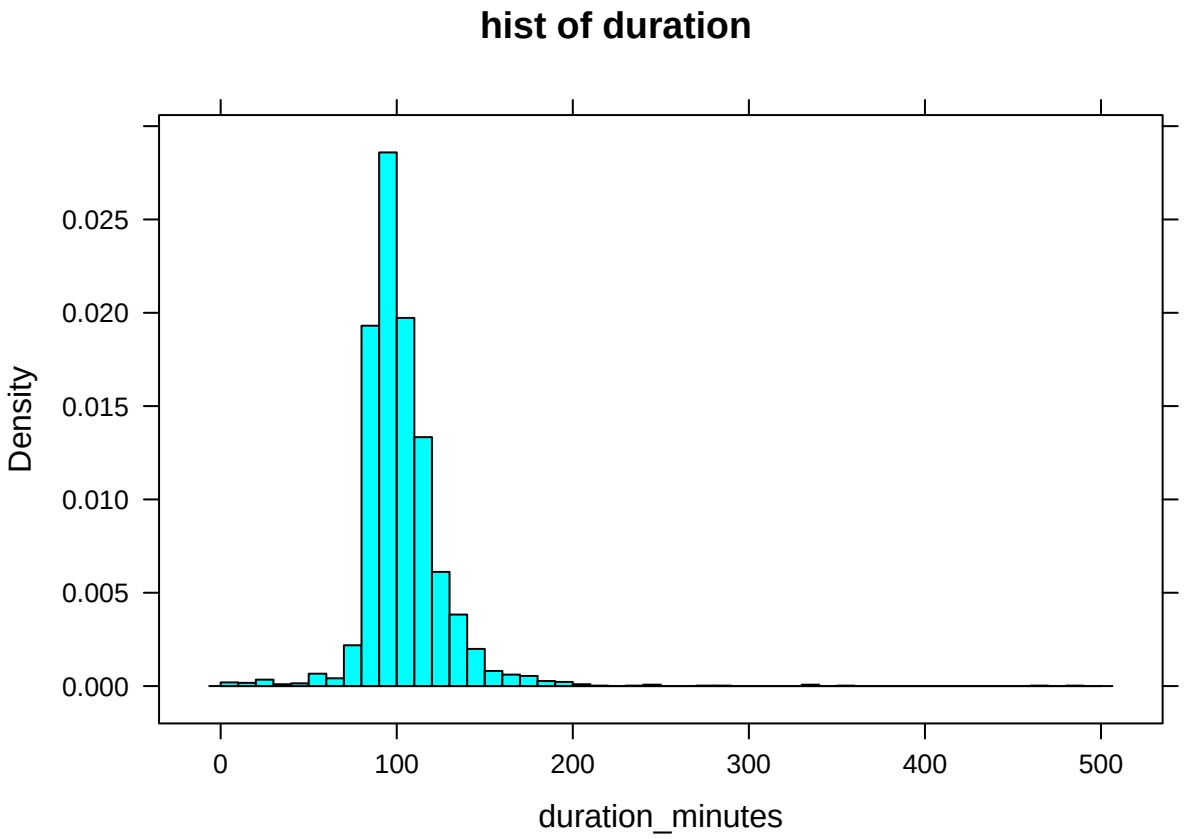
movie_description: textual

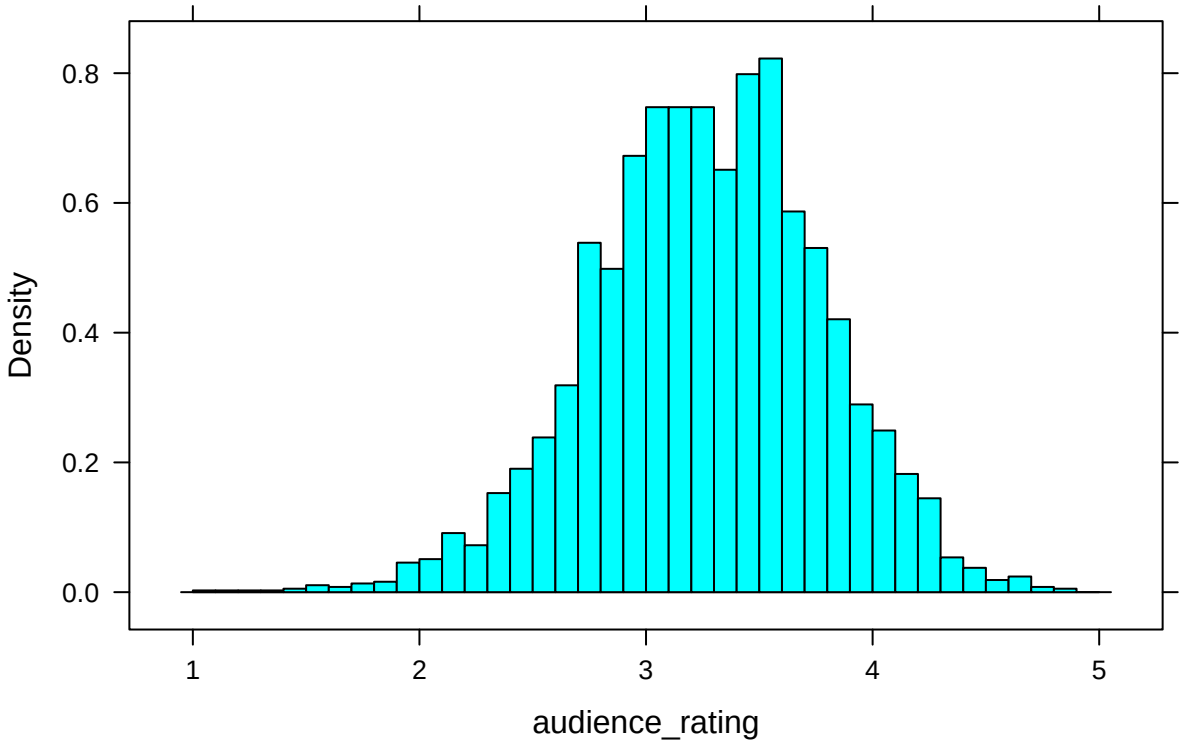
Average length, min and max of textual attributes:

```
## title:  average length: 15.91853,   max: 83,   min: 1
## genres:  average length: 24.51701,   max: 91,   min: 6
## directors:  average length: 15.77802,   max: 155,   min: 6
## writers:  average length: 25.02947,   max: 184,   min: 5

## movie_description:  average length: 673.7872,   max: 3068,   min: 6
```

Outlier detecting:





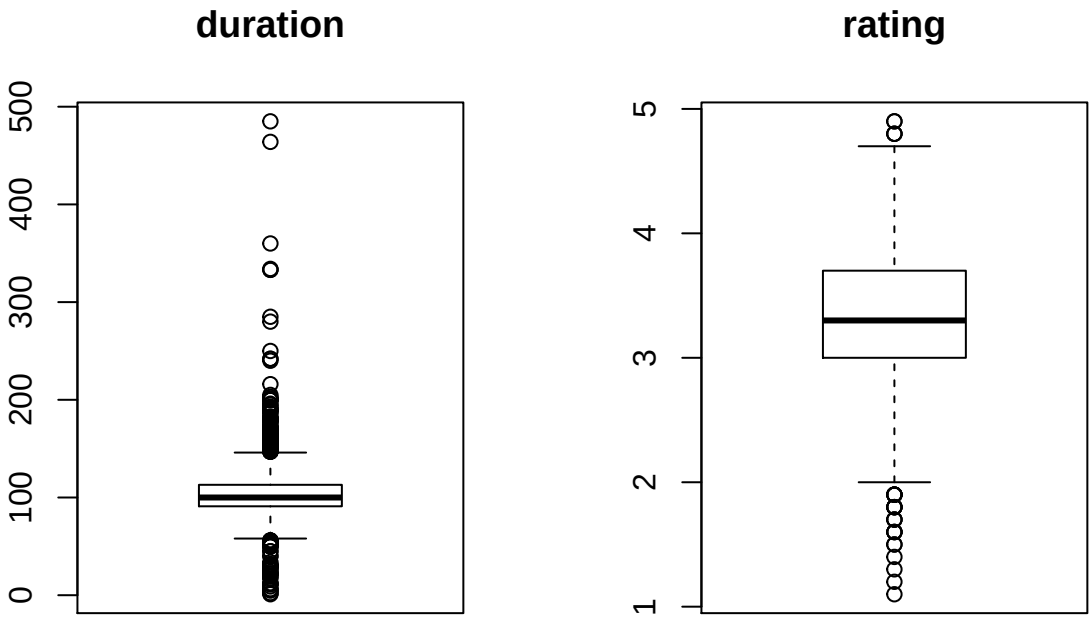
we can see from the above 2 graphs that **HISTOGRAM is really not the tool you want to choose for outlier detecting. It works pretty well on showing you how the distribution looks like though**

I strongly recommend that we use:

boxplot

to show outliers:

hint:those points that are far away from the quartile ranges can be considered potential outliers,and the boundaries cound be adjusted *[as it depends]*



#If all values follow the same format:

The data *under date_in_theaters* attribute is following the same format **MM/DD/YYYY**

The *id* is following the same format as well, which is sequential, from 1 to n, where n is the total number of movies, since the id is being assigned by us when crawling the data.

For some movies, there are multiple genres, directors and writers. Each of these are separated by comma.

Synonyms?

When crawling IMDB, there are two attributes called **directors** and **creators**.

However, upon analyzing the data, we quickly found out that there's a very significant trend/ correlation between these two attributes, eg: *Given a movie, when there's data for directors, there will not be data for creator; when there's data for creators, there will not be data for directors.*

We then found out that directors and creators are essentially the same, so we merge both of the attributes and the data together.

Below shows the attributes in **rottentomatoes.csv** and its corresponding synonymous attributes in **imdb.csv**,

eg: the **release date** in *IMDB.csv* and the **date_in_theater** in *rottentomatoes.csv* are the same

Rottentomatoes.csv	IMDB.csv
title	Title
genres	Category
duration	Duration
audience_rating	Rating
num_of_audience_rating	Rating Count
directors	Director
writers	Writer
date_in_theaters	Release Date
movie_description	Description

Attributes value that sprinkled all over?

In the csv data that we submitted during the stage 1 of our project, there do have some attribute values that are "sprinkled" all over the item. However, after that, we rewrite our crawler from scratch and changes the way we extract data.

eg: instead of extracting by position, we extract by selectors, then compare if the element in the selector is the attribute that we want to extract, then only put it into dictionary.

Therefore, there's **no** (known) attribute values that are "sprinkled" all over the item for the latest csv files that we generated.

Data quality problem:

we have found some duplicates exsiting in our IMDB.csv:

```
##                                          All the Way
##                                                    4
##                                 When Dinosaurs Ruled
##                                                    4
##                                           Bridesmaids
##                                                    3
##                                         Casino Royale
##                                                    3
##                                            Guinevere
##                                                    3
##                                               Hamlet
##                                                    3
##                           I Can Do Bad All By Myself
##                                                    3
##                                           Inside Out
##                                                    3
##                                            Neighbors
##                                                    3
##                                           Nine Lives
##                                                    3
##                                                 NOVA
##                                                    3
##                                               Stolen
##                                                    3
##                                 10 Cloverfield Lane
##                                                    2
##                           10 Things I Hate About You
##                                                    2
##                                         10,000 B.C.
##                                                    2
##                                               11:14
##                                                    2
##                                            13 Sins
##                                                    2
##                                             2 Guns
##                                                    2
##                                                 21
```

```
##                                                 2
##                                          21 Grams
##                                                 2
##                                     21 Jump Street
##                                                 2
##                                          29 Palms
##                                                 2
##                                 3 Hearts (3 coeurs)
##                                                 2
##                                             50/50
##                                                 2
##                          A Birder's Guide To Everything
##                                                 2
##                                 A Hologram for the King
##                                                 2
##                                      A Knight's Tale
##                                                 2
##                                  A Little Bit Of Heaven
##                                                 2
##                                      A Little Chaos
##                                                 2
##                    A Man Called Ove (En man som heter Ove)
##                                                 2
##                                    A Midnight Clear
##                                                 2
##                                  A Room With a View
##                                                 2
##                                   A Royal Night Out
##                                                 2
##                                    A Thousand Words
##                                                 2
##                                          About Adam
##                                                 2
##                                          About Alex
##                                                 2
##                                         About Sunny
##                                                 2
##                        Ace the Case: Manhattan Mystery
##                                                 2
##                                         Adaptation
##                                                 2
##                                           Addicted
##                                                 2
##                                  Ae Dil Hai Mushkil
##                                                 2
##                                           Aftermath
##                                                 2
##                                   Afternoon Delight
##                                                 2
##                                Ain't Them Bodies Saints
##                                                 2
##                                          Air America
##                                                 2
##                                      Alan Partridge
```

```
##                                                          2
## Alexander and the Terrible, Horrible, No Good, Very Bad Day
##                                                          2
##                                        Alice in Wonderland
##                                                          2
##                              Alice Through the Looking Glass
##                                                          2
##                                       Alice Upside Down
##                                                          2
##                                                    Alien
##                                                          2
##                                                   Aliens
##                                                          2
##                                     All You Need Is Cash
##                                                          2
##                                            Almost Famous
##                                                          2
##                                          Alpha and Omega
##                                                          2
##                                         Always Woodstock
##                                                          2
##                                        American Graffiti
##                                                          2
##                                       American History X
##                                                          2
##                                           American Honey
##                                                          2
##                                        American Pastoral
##                                                          2
##                                           American Ultra
##                                                          2
##                                            Amos & Andrew
##                                                          2
##                                        An American Carol
##                                                          2
##                                       An Invisible Sign
##                                                          2
##                                                Anastasia
##                                                          2
##                                              Anesthesia
##                                                          2
##                                           Another Year
##                                                          2
##                                                   Antz
##                                                          2
##                                             Arbitrage
##                                                          2
##                                            Armageddon
##                                                          2
##                                         Around The Block
##                                                          2
##                                                 Arrival
##                                                          2
##                                                  Arthur
```

```
##                                                 2
##                                       Arthur Newman
##                                                 2
##                                        At Any Price
##                                                 2
##                           Austin Powers in Goldmember
##                                                 2
##                                         Awakenings
##                                                 2
##                  Bad Lieutenant: Port of Call New Orleans
##                                                 2
##                                           Bad Moms
##                                                 2
##                                        Bad Teacher
##                                                 2
##                                       Barbary Coast
##                                                 2
##                                       Barely Lethal
##                                                 2
##                        Batman v Superman: Dawn of Justice
##                                                 2
##                                         Battleship
##                                                 2
##                                       Beacon Point
##                                                 2
##                             Beasts of the Southern Wild
##                                                 2
##                                       Beautiful Boy
##                                                 2
##                                     Beautiful Kate
##                                                 2
##                                   Beautiful People
##                                                 2
##                                Beauty and the Beast
##                                                 2
##                                             Bessie
##                                                 2
##                                     Best in Show
##                                                 2
##                                        Big Bully
##                                                 2
##                                        Big Night
##                                                 2
##                                          Big Sky
##                                                 2
##                                       Big Trouble
##                                                 2
##                                        Billy Budd
##                                                 2
##                          Billy Lynn's Long Halftime Walk
##                                                 2
##                                          Birdman
##                                                 2
##                                          (Other)
```

```
##                                                                3937
```

Software Used:

R,RStudio

Python Packages: CSV,re