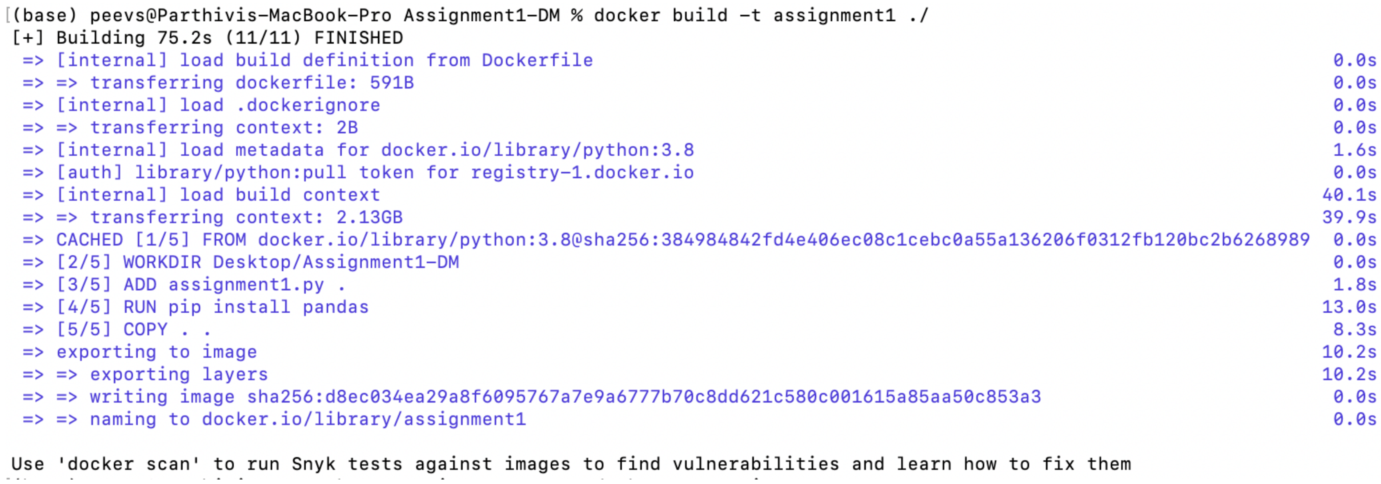
Data Mining - Assignment 1

Student Parthivi Varshney
Professor Karl Ni
Course Data Minig

Part 1

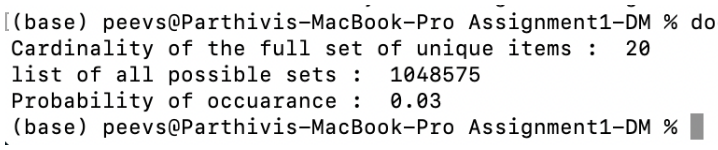
Command to build Docker Image :
docker build -t assignment1 ./
Output :



```
(base) peevs@Parthivis-MacBook-Pro Assignment1-DM % docker build -t assignment1 ./
[+] Building 75.2s (11/11) FINISHED
 => [internal] load build definition from Dockerfile                                                 0.0s
 => => transferring dockerfile: 591B                                                                 0.0s
 => [internal] load .dockerignore                                                                    0.0s
 => => transferring context: 2B                                                                      0.0s
 => [internal] load metadata for docker.io/library/python:3.8                                        1.6s
 => [auth] library/python:pull token for registry-1.docker.io                                        0.0s
 => [internal] load build context                                                                   40.1s
 => => transferring context: 2.13GB                                                                 39.9s
 => CACHED [1/5] FROM docker.io/library/python:3.8@sha256:384984842fd4e406ec08c1cebc0a55a136206f0312fb120bc2b6268989  0.0s
 => [2/5] WORKDIR Desktop/Assignment1-DM                                                             0.0s
 => [3/5] ADD assignment1.py .                                                                       1.8s
 => [4/5] RUN pip install pandas                                                                    13.0s
 => [5/5] COPY . .                                                                                   8.3s
 => exporting to image                                                                              10.2s
 => => exporting layers                                                                             10.2s
 => => writing image sha256:d8ec034ea29a8f6095767a7e9a6777b70c8dd621c580c001615a85aa50c853a3         0.0s
 => => naming to docker.io/library/assignment1                                                       0.0s

Use 'docker scan' to run Snyk tests against images to find vulnerabilities and learn how to fix them
```

Command to run Docker Container :
docker run assignment1
Output : (When the docker is run, the result displayed is of questions asked in Part 2 of the assignment)



```
(base) peevs@Parthivis-MacBook-Pro Assignment1-DM % docker run assignment1
Cardinality of the full set of unique items :  20
list of all possible sets :  1048575
Probability of occuarance :  0.03
(base) peevs@Parthivis-MacBook-Pro Assignment1-DM %
```

Part 2

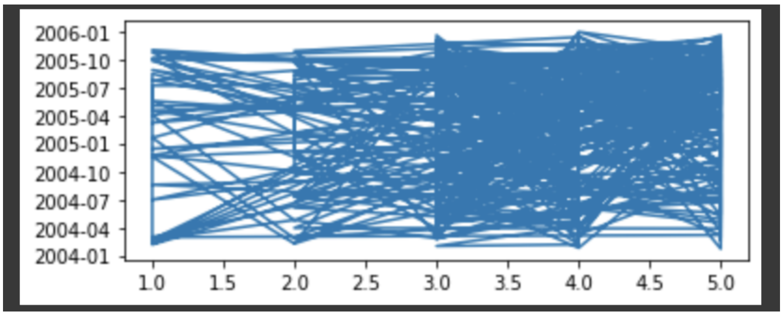
Question 2 :

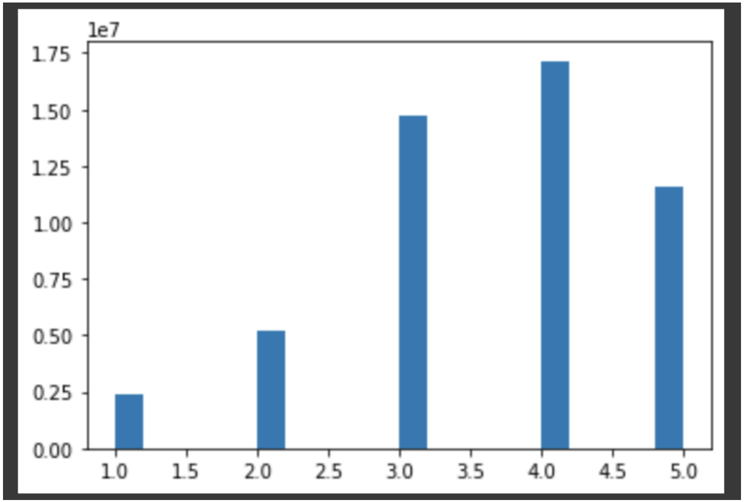
$$2^N - 1$$

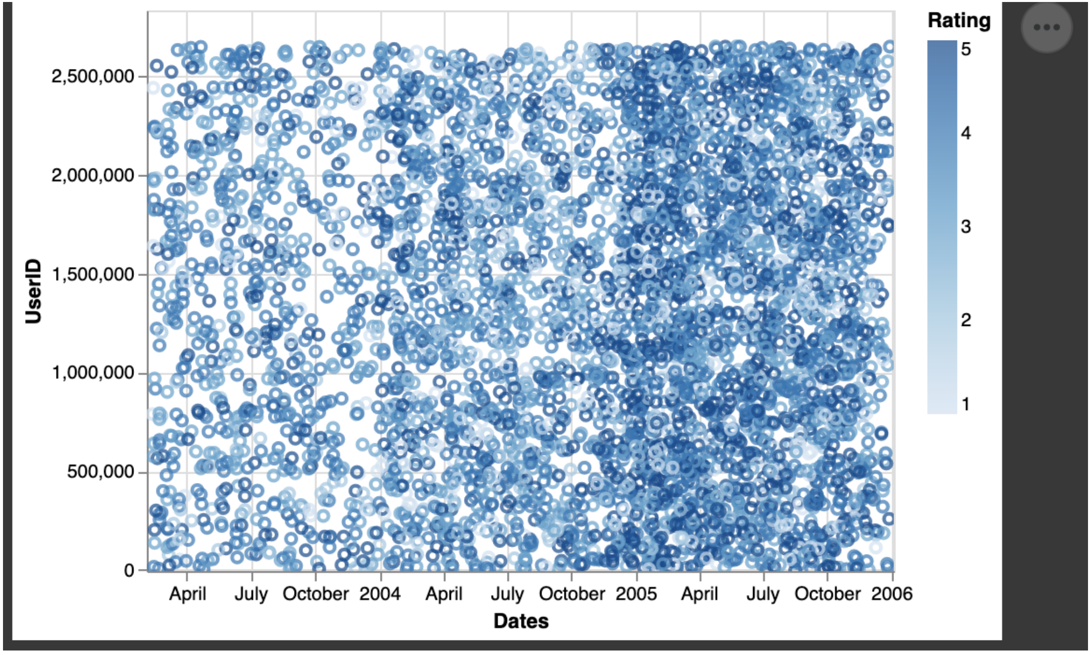
Part 3

Question 1 :
Total Movie Records : 17770

Question 2 :
The plot of star ratings over user and time depict that most movies are given an average rating of 3 and 4, while only a handful movies get rated 1. Below are some plots to depict the same :



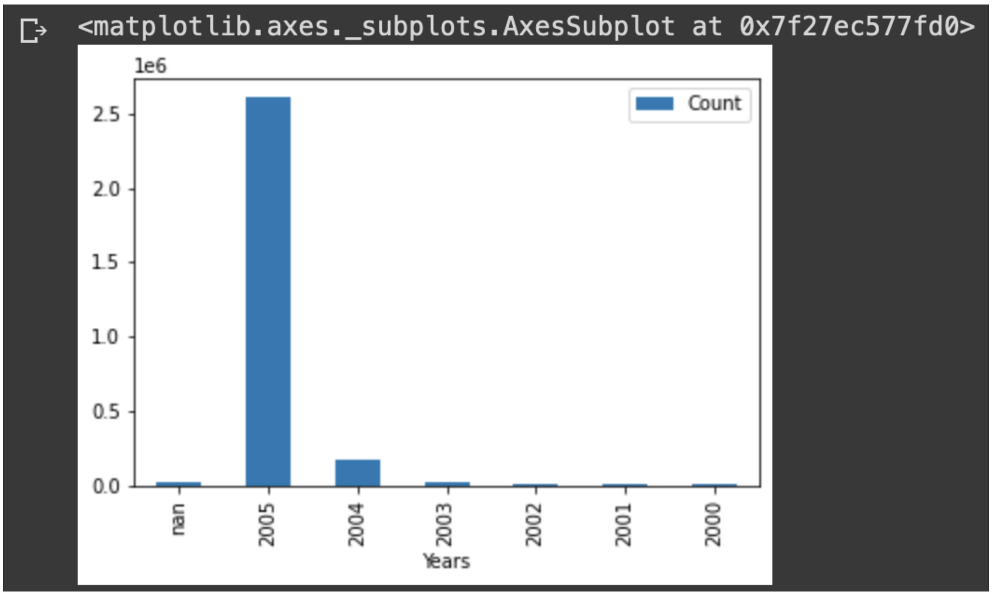




Question 3 :
Looking into Qualify.txt, we can see that most movies were watched by the user during the time period of 2000 to 2005. A total of 17470 movies were watched during this span. Calculating the percentage of movies watched during this span results in 98% of movies getting more popular over time.

Question 4 :
Assuming the fact that a movie is re-released with the same name, there are about 401 movies that have been re-released.

Question 5 :
We can also try to extract in which year were the users more active, or in other word if the usage rate of the platform has gone up. Since the data is till year 2005, from the graph below it is clear that the rate of use has definitely gone up.



```
<matplotlib.axes._subplots.AxesSubplot at 0x7f27ec577fd0>
```

Question 6 :
One interesting problem that we can solve with this provided that is displaying every users all time favourite movie in their watch-list. Collection of user's favourite movies can also give us a list of all-time favourites.