MovieLens Dataset Exploratory Data Analysis using R

Devashish Bharadwaj

Subhendu Saha

Abstract

MovieLens dataset analysis using R is the aim of the project. We have used the dataset of 1 million records and done analysis on the data set using R statistical analysis. We have tried to show patterns in this data and completed the exploratory data analysis (EDA).

About the dataset:

- Users.csv – contains the information about the users.

  UserID::Gender::Age::Occupation::Zip-code

- Ratings.csv – contains the information about the ratings for the movies.

  UserID::MovieID::Rating::Timestamp

- Movies.csv – contains the information about the movies.

  MovieID::Title::Genres

We have made connections across the three files and tried to show which the most liked movies are and where in the United States they are liked. Also, we have information about the users so we can tell which type of users liked a particular genre of movies. We have then mapped the user occupation to the movie genres they have liked. The user's zip code is also mentioned so we can see how well the movie did in a particular region. We have read the CSV files in the R and made changes to the file and plotted the graphs. We have then given a short interpretation of the graphs.

I think this project is cool as we get to analyze our favorite movies and speculate why people liked them or why not? The movies have a genres associated with so we have categorize which are the popular genres among Americans. We have the timestamps from the time when users have rated the movie so we can tell how long the movie popular after its release. All in all the analysis can tell which movie has well liked and where. We can draw some interesting conclusion from this. We have tried to interpret the data in such a way that we can take away some meaning from the graphs and plots that come out from the data.

We have also worked on the examples given in the text book and build our knowledge of R. We have then applied this knowledge to the real world data from the MovieLens website. The examples in the text book given us enough knowledge about R and some startup code for our own project in R. We have used the code given in text book to start with the training data sets. And then applied the same logic to the real world data set form MovieLens.

Project Objectives

- Learn R programming language for statistical data analysis.
- Use R studio
- Solve the incomplete examples in the textbook.
- Derived graphs and plots from the examples data
- As the data is real world draw some conclusions from it
- Use the graphs and plots to find useful patterns in the data.
- Use the knowledge gained from solving examples to our own data
- Break the data into:
  i) Spatial
  ii) Temporal
  iii) Geospatial
- Analysis the data and plot graphs for single day/time.
- Extract the useful data and store it.
- Plot graphs over time and regions.
- Draw out meaningful conclusions
- How to apply these conclusions to the real world data?
- Have fun!

Project Approach

We have followed this approach in the project:

- Followed the examples in the textbook and understood the R code given.
- We have then completed the examples in the text book and written our own code in addition.
- Exploratory data analysis has been done on the data generated by our code.
- We have built our knowledge of R on these examples.
- Exploited patterns in the data and found out interesting facts about data.
- As the examples themselves are from real world data we have drawn accurate conclusions

New York Times Dataset:

- From the New York Times data we have inferred the reading habits of people of different ages.
- As the data has been collected over time we have found when visit the website the most.
- We have calculated the Click-through-rate to find out about how users are reacting to news.

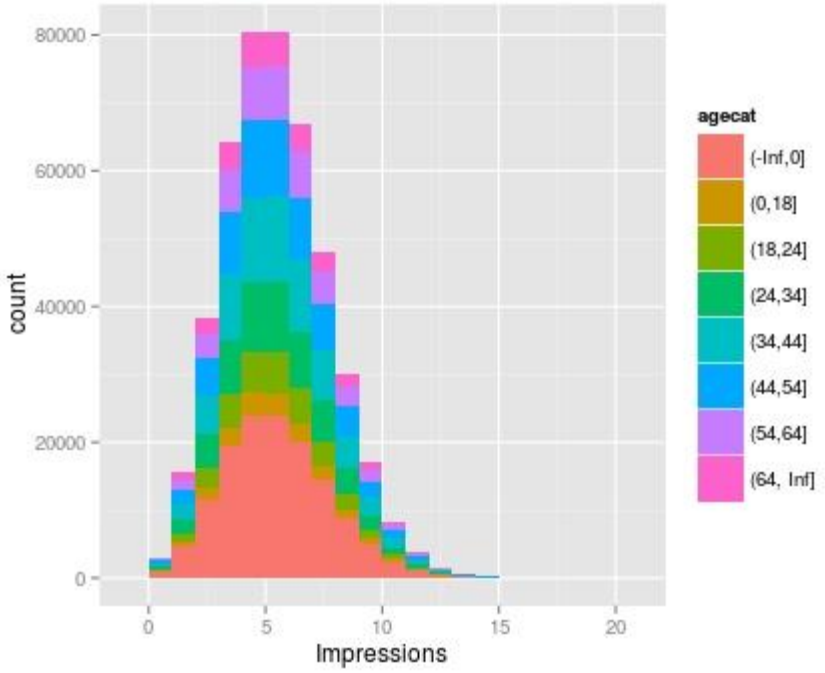
RealDirect Dataset:

- From the RealDirect data we have found about the real state market in New York city.
- The neighborhoods which have the cheapest real estate to the most expensive places in the city.
- We have shown which places were preferred by families and how much they paid for it.
- Some neighborhoods have tiny apartments while others have sprawling gardens.
- We have compared all the sales that were made by the company.

New York Times Dataset

Following the analysis of the New York Times data set starting 1st May 2012:

This part of analysis is for 1st May:

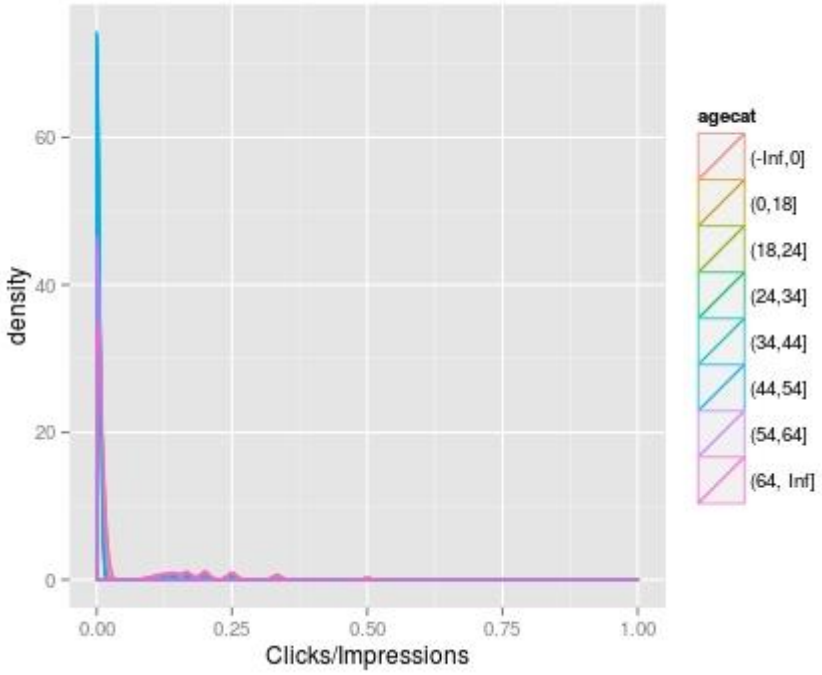


Conclusions:

- The older reader are more active.
- The Impressions count is decreases with age.
- The young people have little impression.
- We can say on the 1st of May the site has traffic from older people.

How is this good?

- Ads can be for older generation on this website.
- Better suited ads more ad clicks and more revenue.

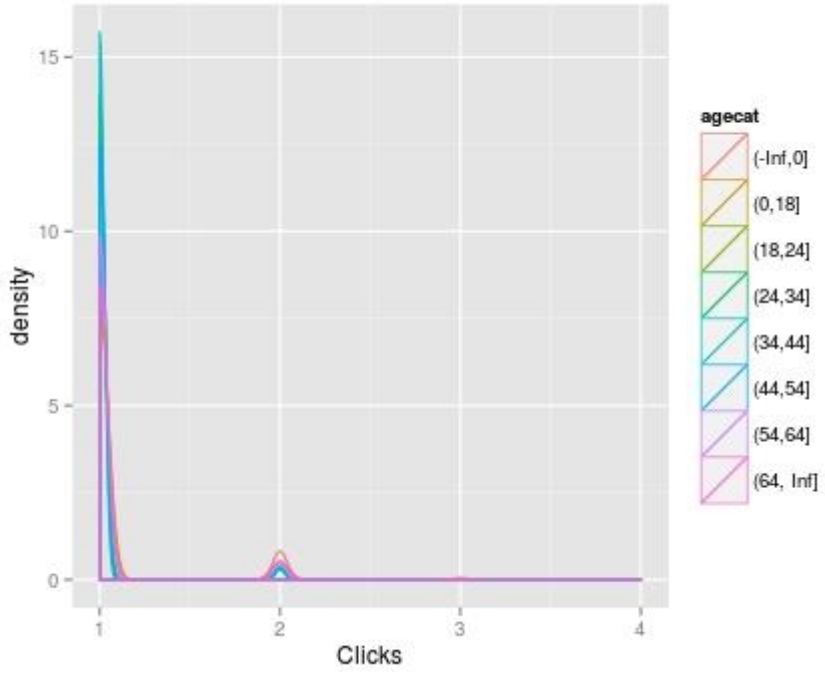


Conclusion:

- Middle age readers have the highest Click-through-rate.
- Middle age users are having a real impact on the website.

How is this good?

- We can target these users with specific ads.
- Front page news can be tilted toward this generation.



Conclusion:

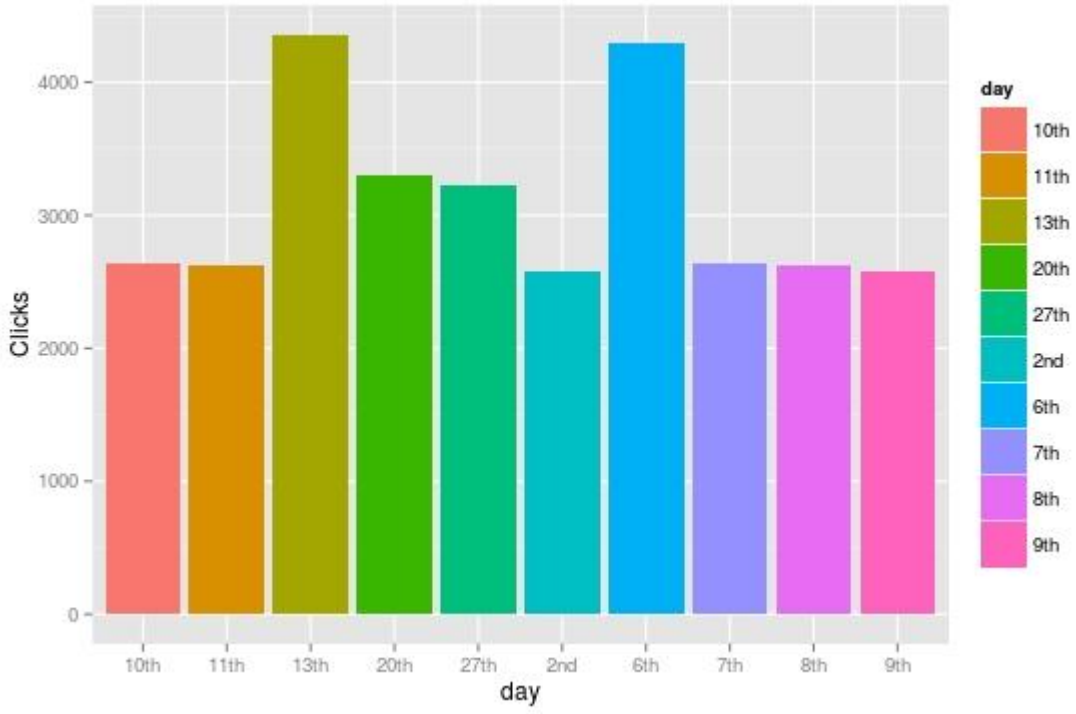
- Middle age readers have one clicks.
- These users are having a real impact on the website.
- Few users click more than once.
- They might be getting redirected to another site.
- Users don't find what they need in one click.

How is this good?

- We can target middle-age users with specific ads.
- We must stop users if they are leaving the site too soon.
- Unless they are clicking on ads, users are finding what they are looking for.

Sunday get the most traffic, shown in further analysis-

The following analysis is done for 30 days, 1st May 2012 to 31st May 2012:



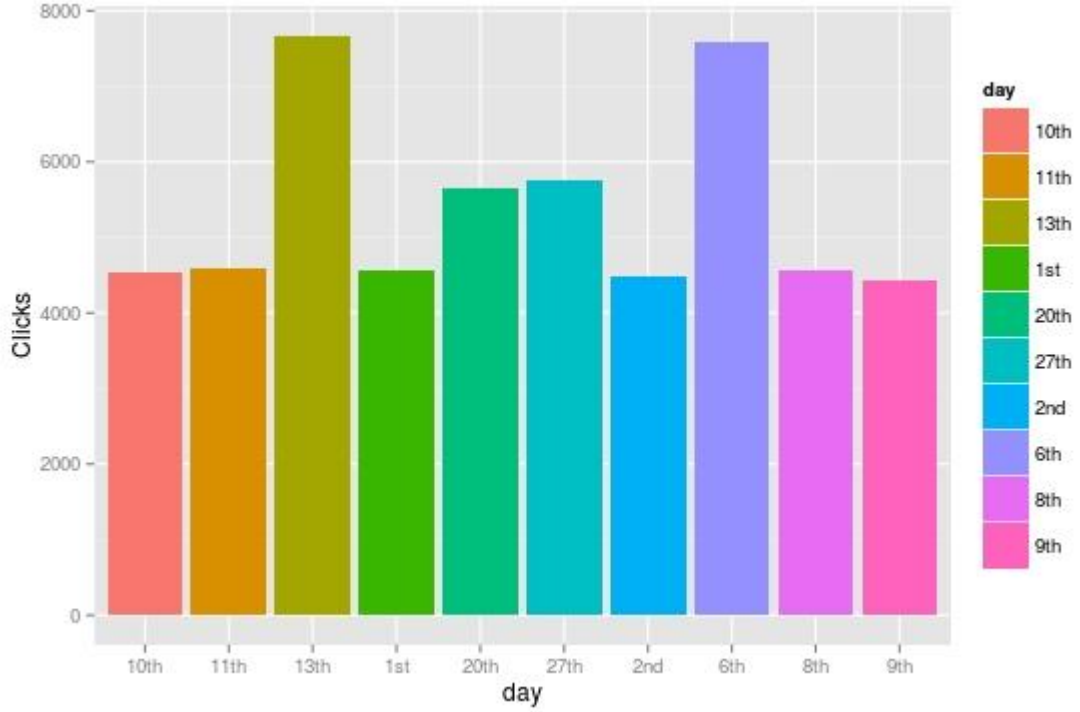
Conclusion:

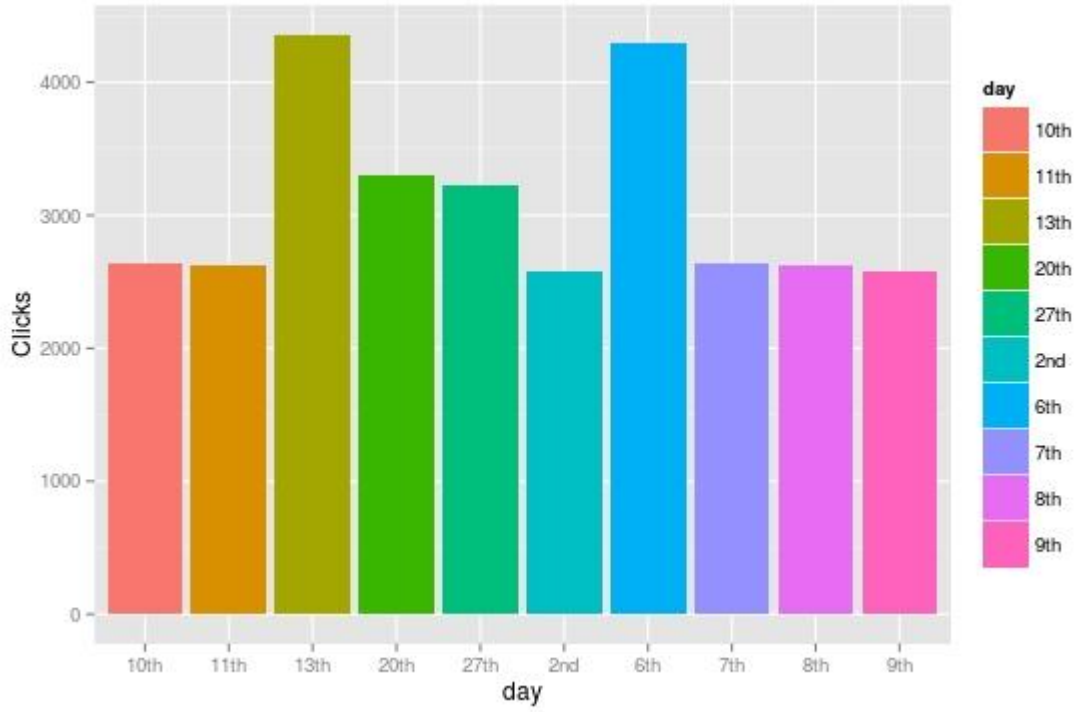
- On May 13th and May 6th we have highest traffic.
- These days are **Sundays**
- More people on site Sunday.
- Traffic goes down as the week progresses.

How is this good?

- We can say the traffic is more Sundays.
- Can make the site appealing on those days so users come back.

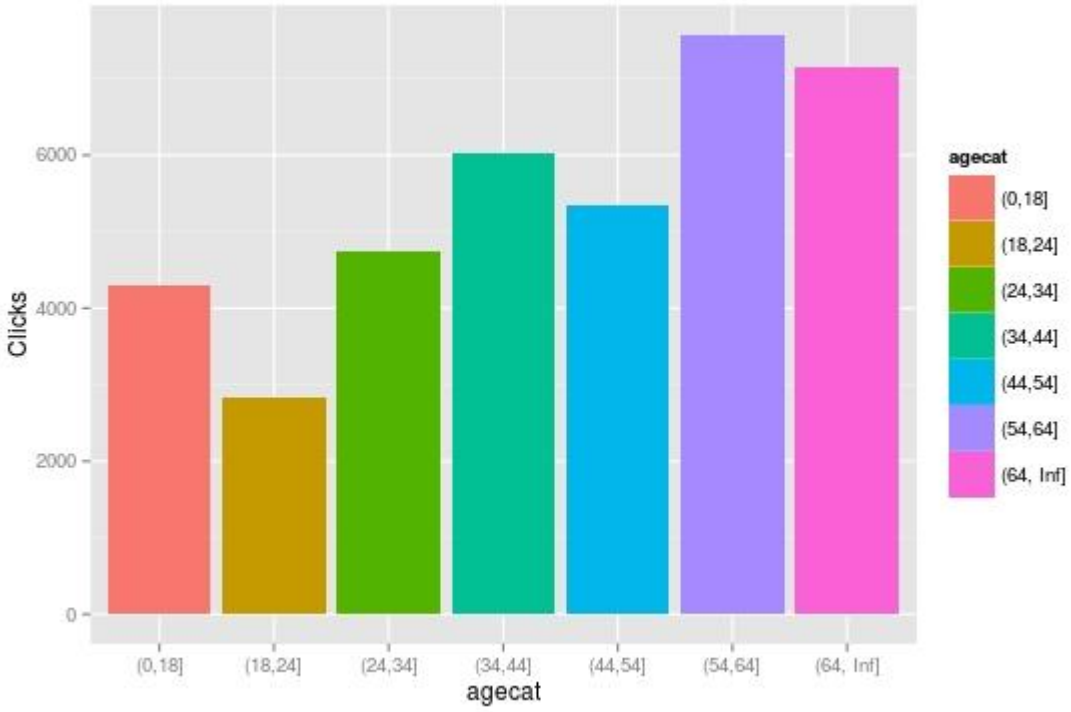
The site draws more traffic on Sundays as shown by all age groups:

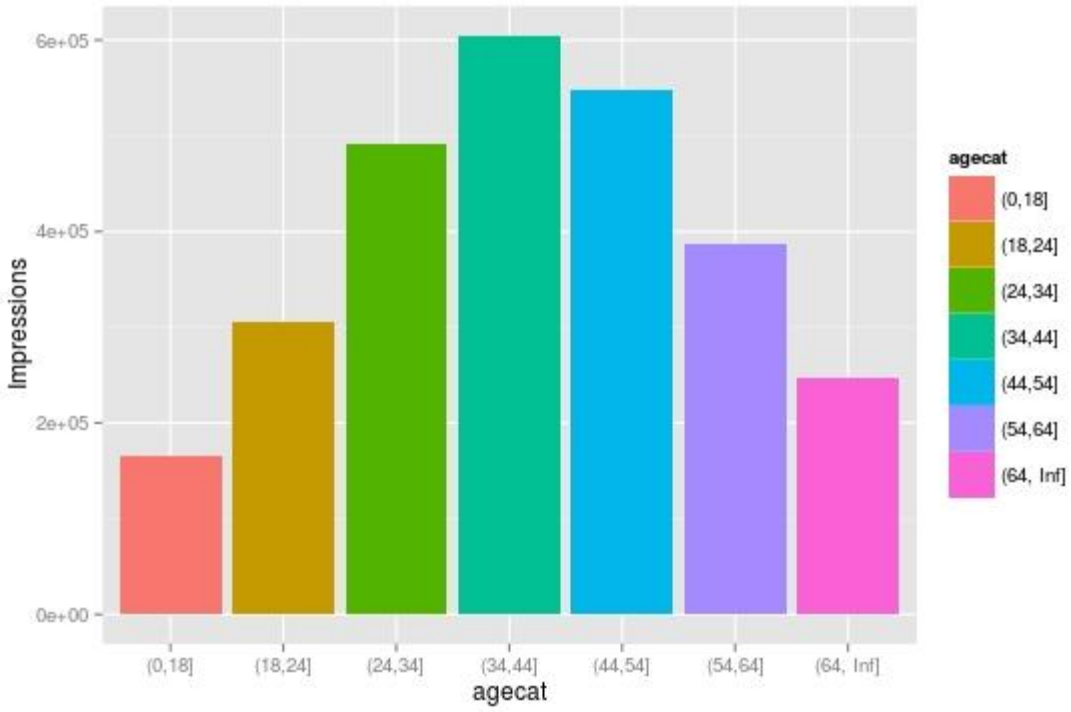




For Age-Group: 24-34 / Age-Group: 64+ (all groups show same behavior)

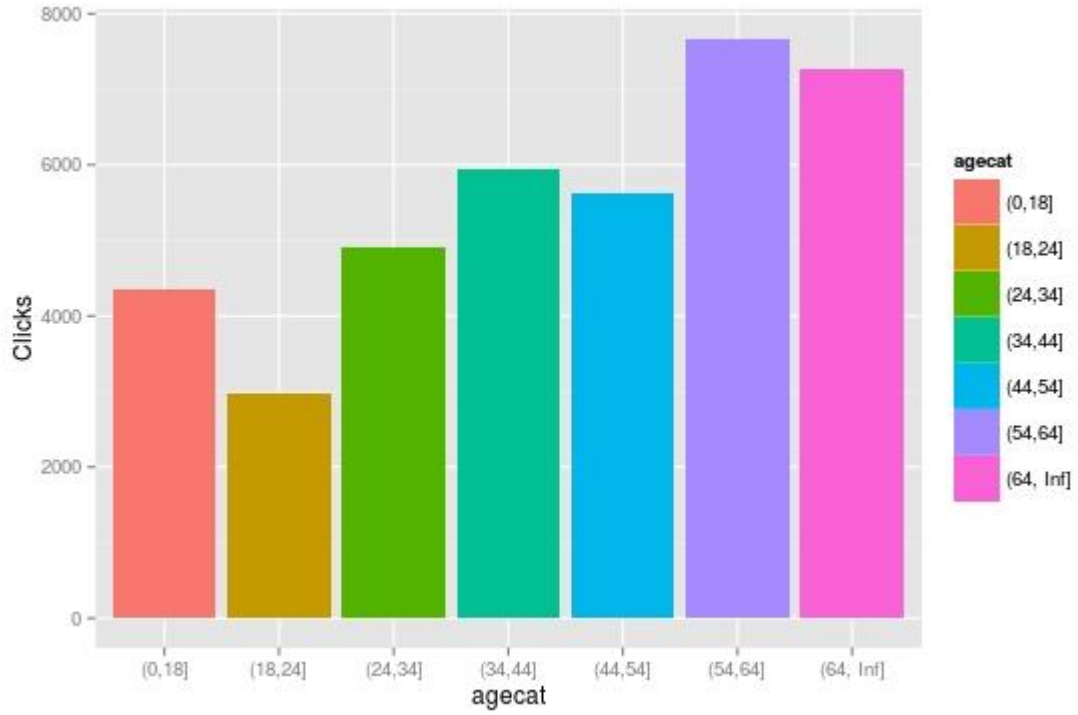
More Analysis 6th May:

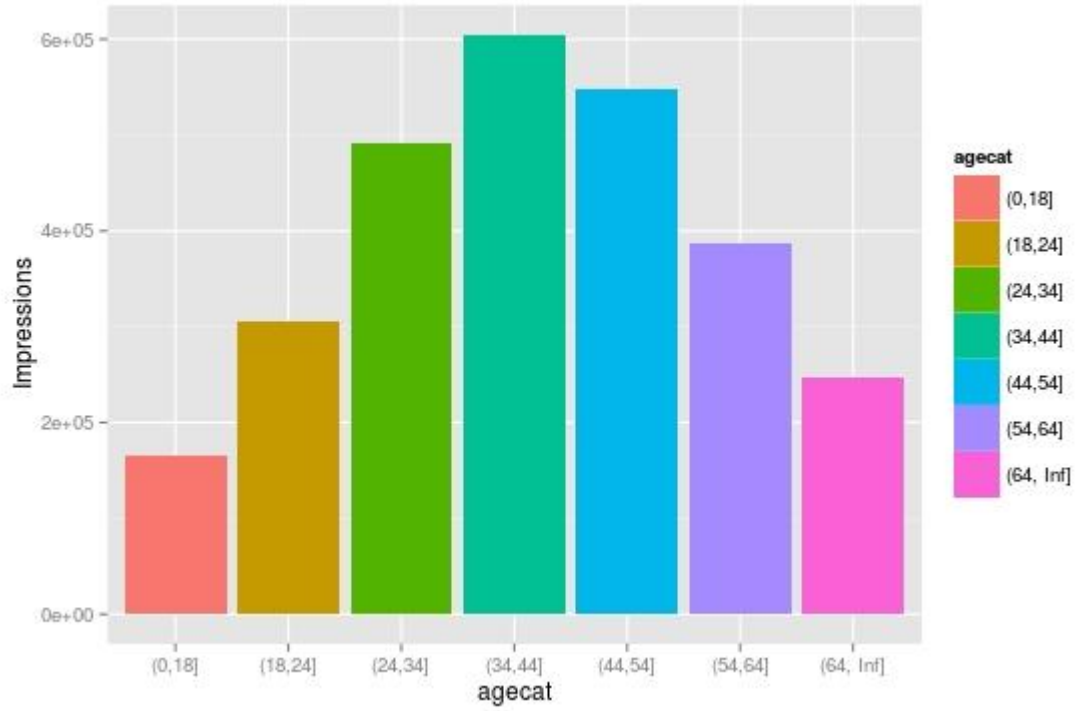




Day 6: Clicks and Impressions

More Analysis 13th May:



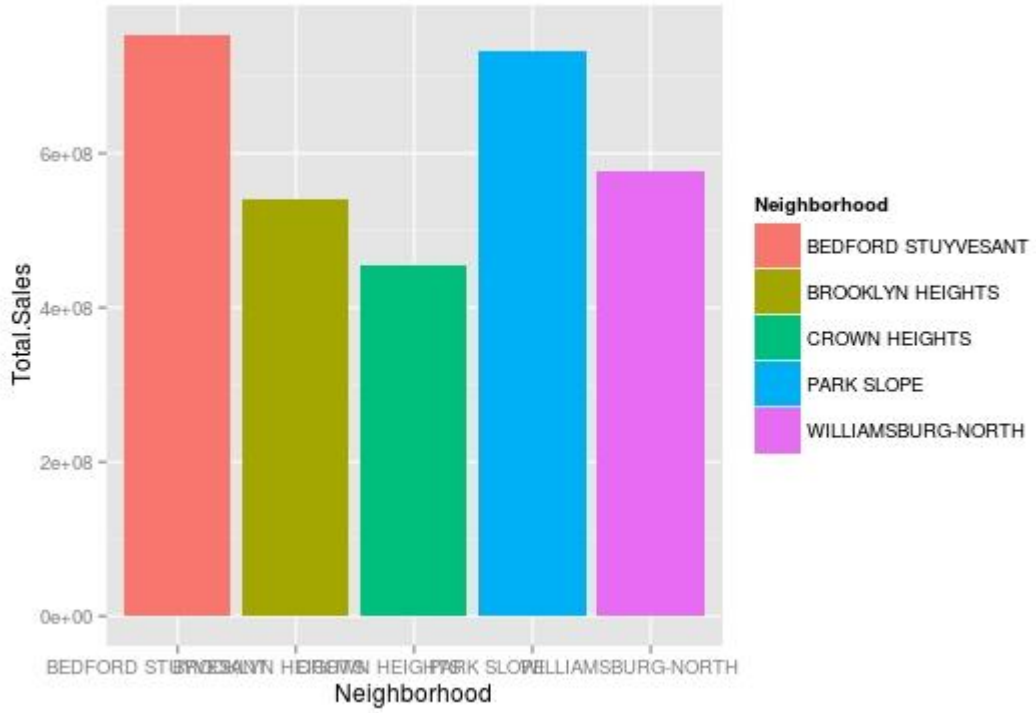


Day 13: Clicks and Impressions

- Some Sundays get highest traffic
- Middle-age and old people visit the site the most.
- Middle-age people have the highest impression.

RealDirect Dataset

EDA for RealDirect Dataset for Manhattan, New York.

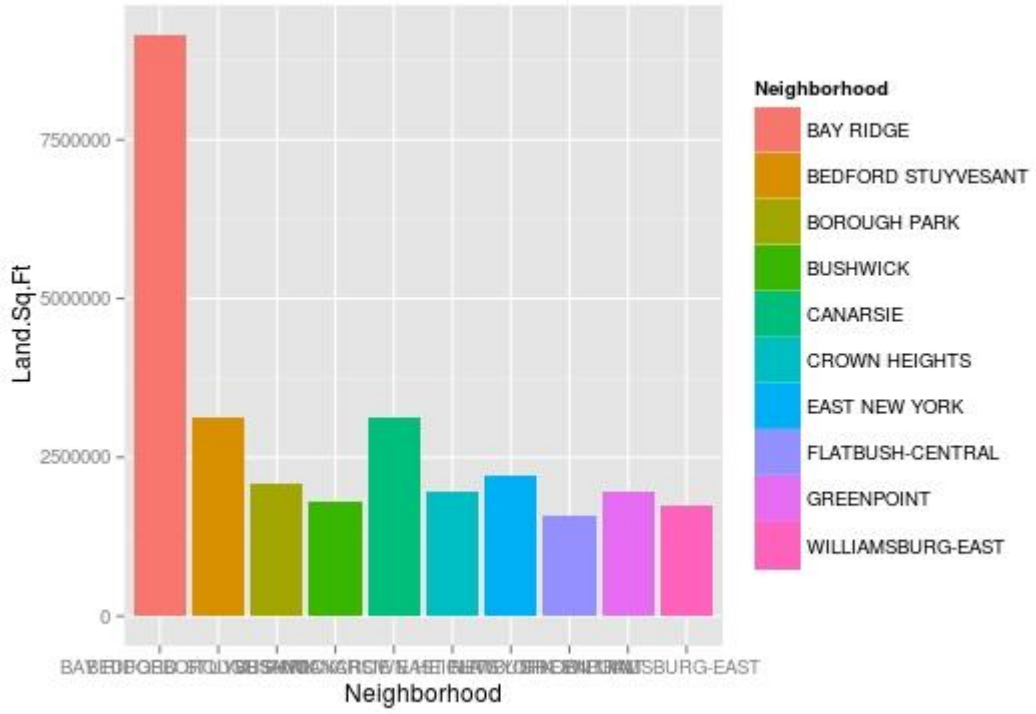


Conclusion:

- Highest sales in Bedford Stuyvesant
- These are the top 5 sales in Manhattan by the data.

How is this good?

- We know that which the most selling property in New York.
- The company can target these areas more.

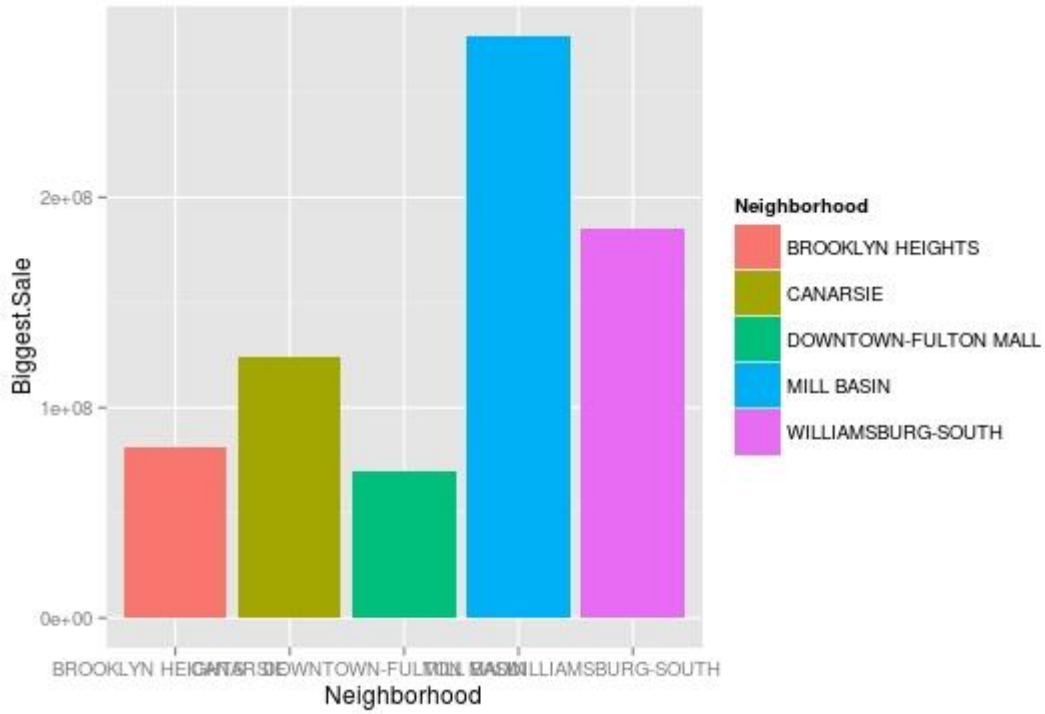


Conclusion:

- We know Bay Bridge has the biggest lands which were sold.
- Apartments in Bay Bridge are big.

How is this good?

- If customers need big apartments we can take them to Bay Bridge.
- We know which places have small apartments.

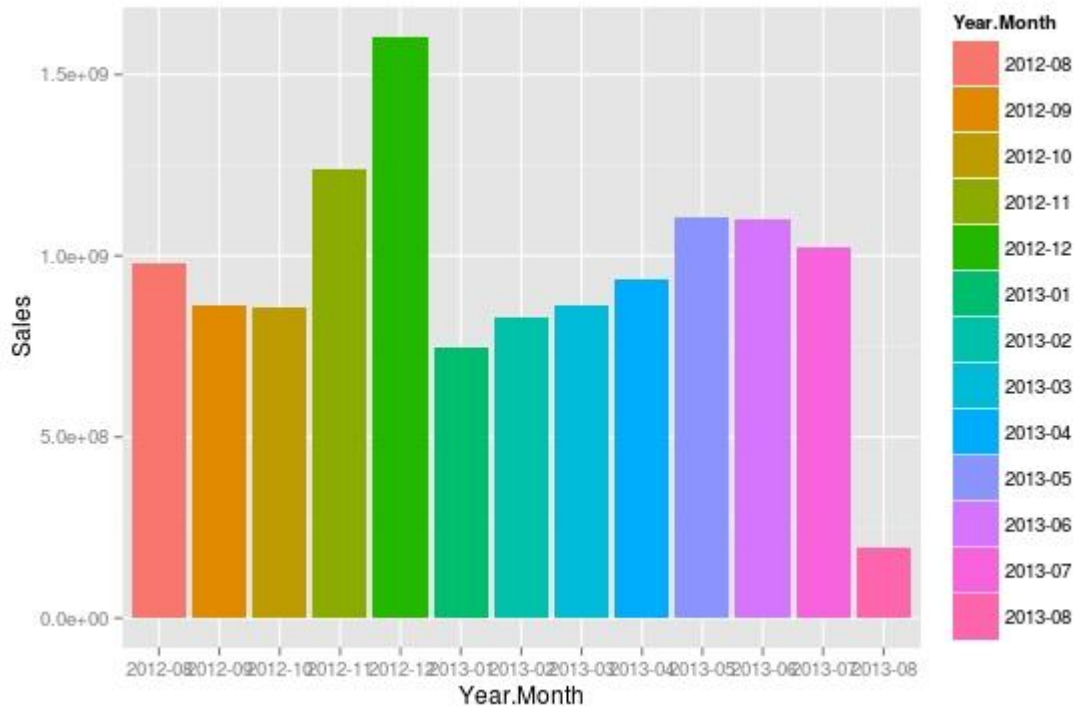


Conclusion:

- Biggest sale of the year was in Mill Basin.
- We can classify this area as a posh area.
- Not many sales but one-hit-wonders in this area.

How is this good?

- We only need one good sale in this area.
- Only rich customers can afford this place.



Conclusion:

- Most sales were done in Christmas time.
- We have a lot of sales during Christmas holidays.

How is this good?

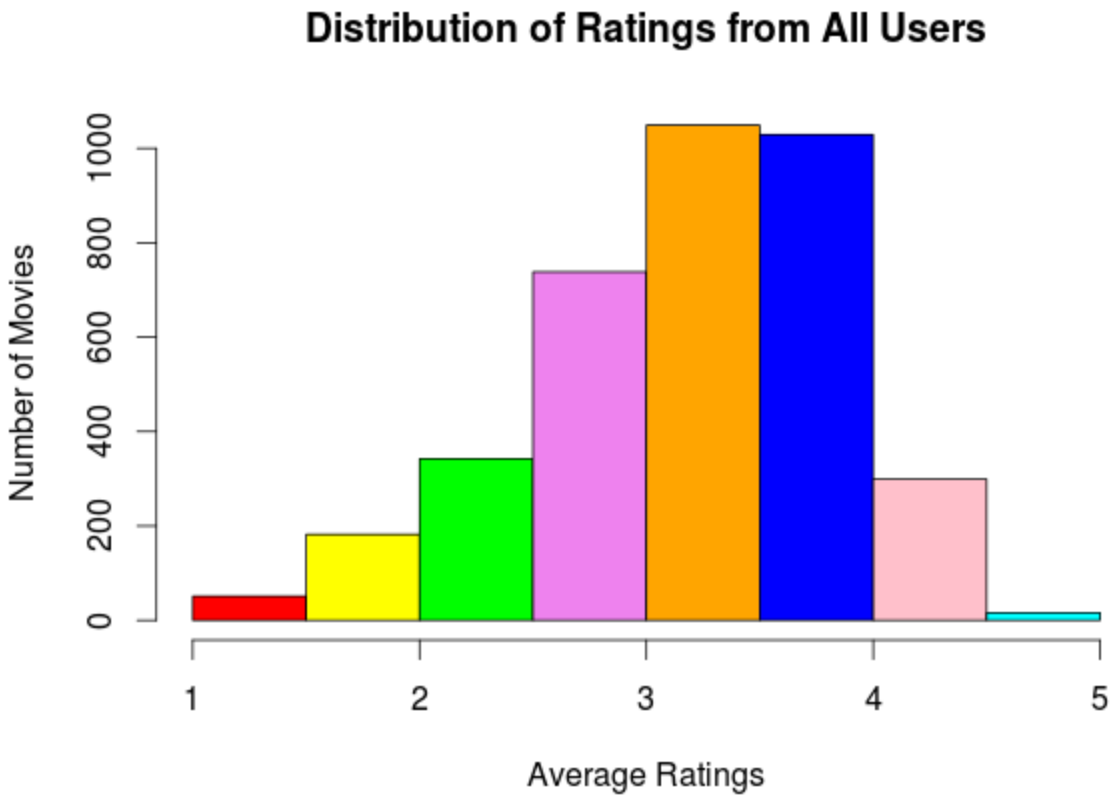
- We can focus on this time of the year and get good results.
- More advertising during Christmas.
- Christmas bonus for employees can also be done.
- Try to maximize the sales during the holiday period.

MovieLens Dataset

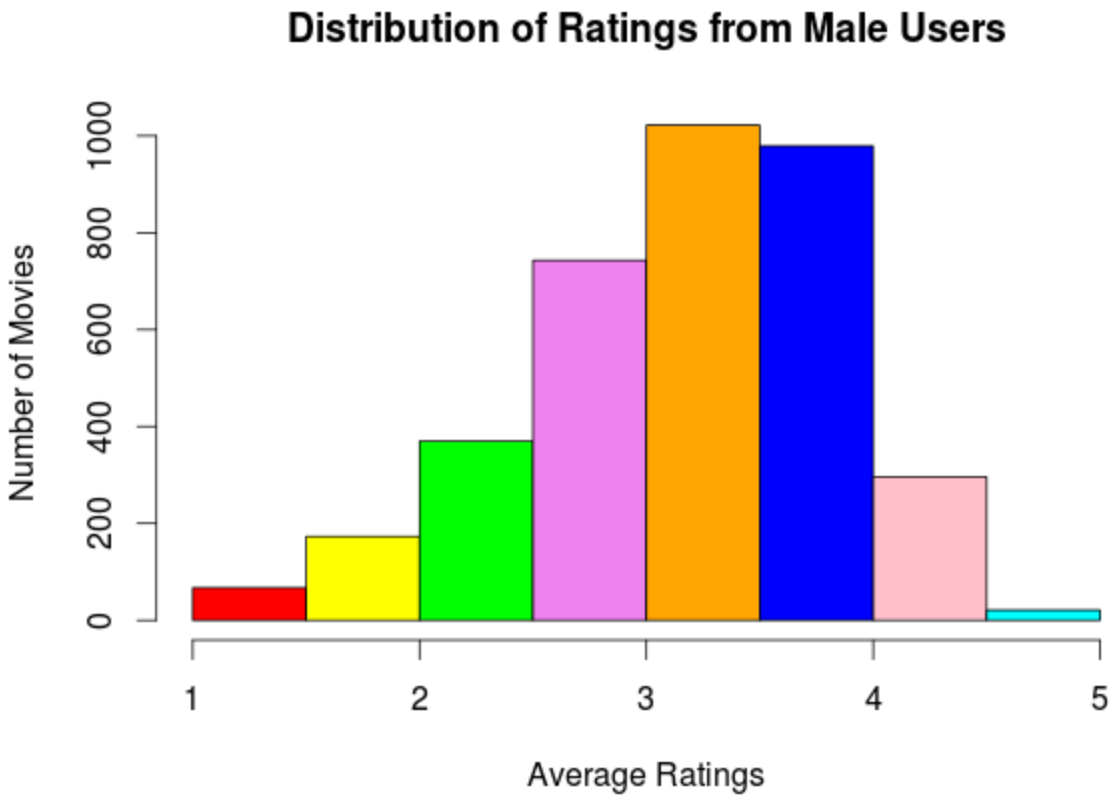
Our initial analysis for Exploratory Data Analysis for the above data set was to find out the overall ratings collected by the movies and compare them with the various demographic information presented by the users to find out how far the ratings varies.

Next we present a set of histograms which gives us an idea about the nature of the ratings as it varies with users from different demography.

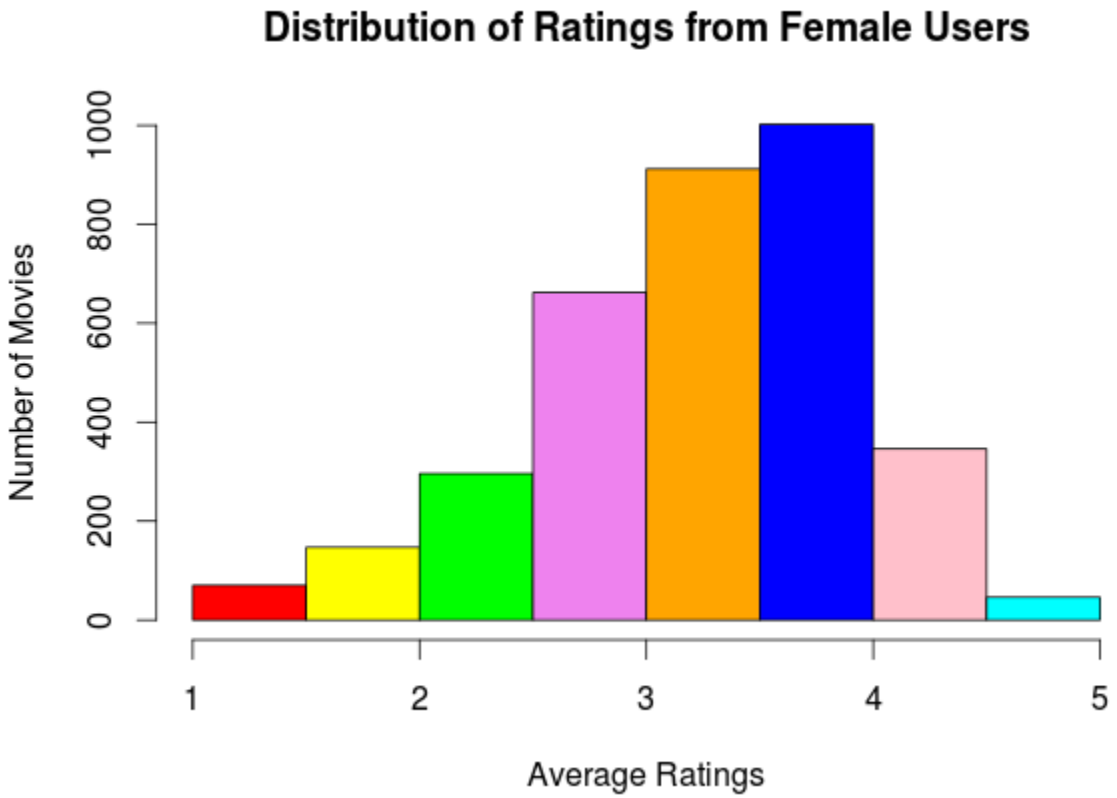
- Overall ratings distribution from all users



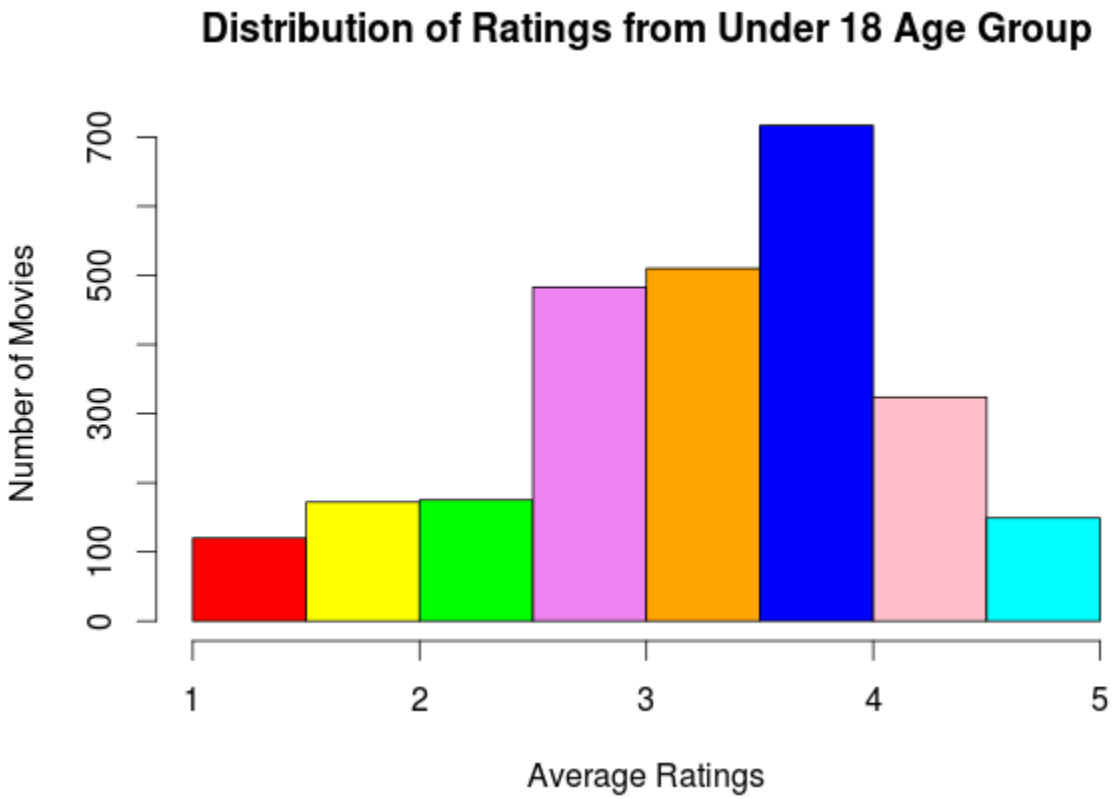
- Ratings from Male users only

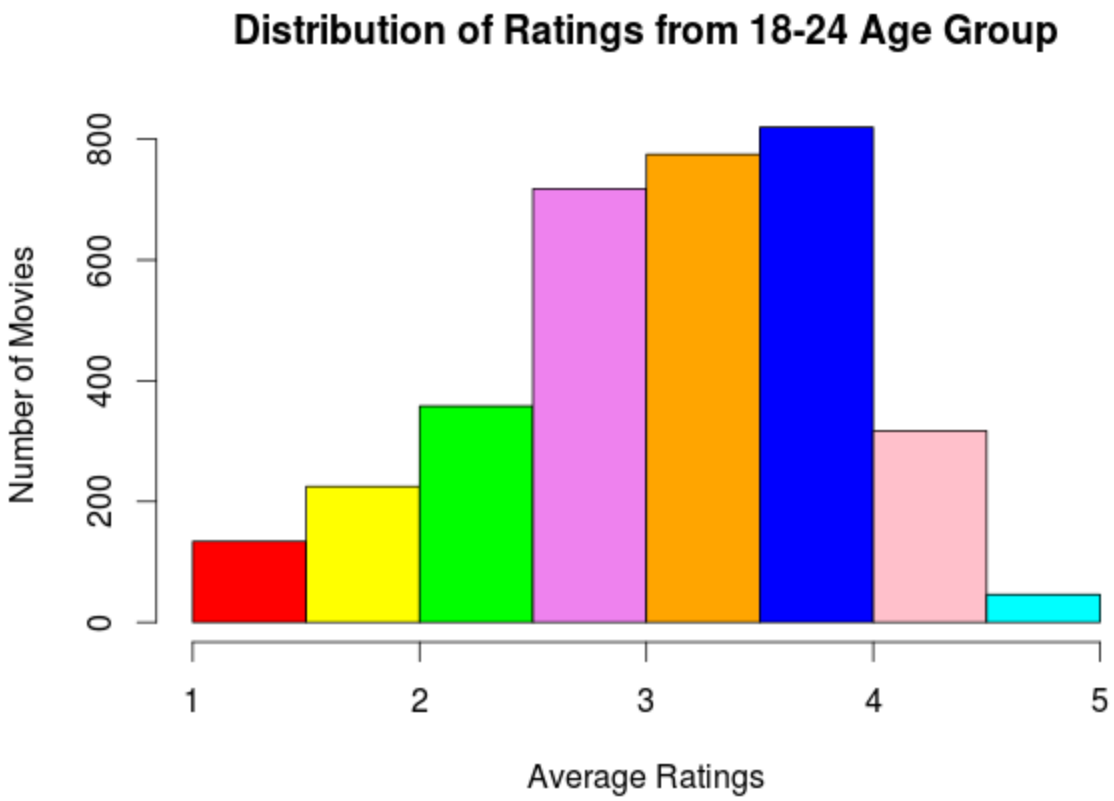


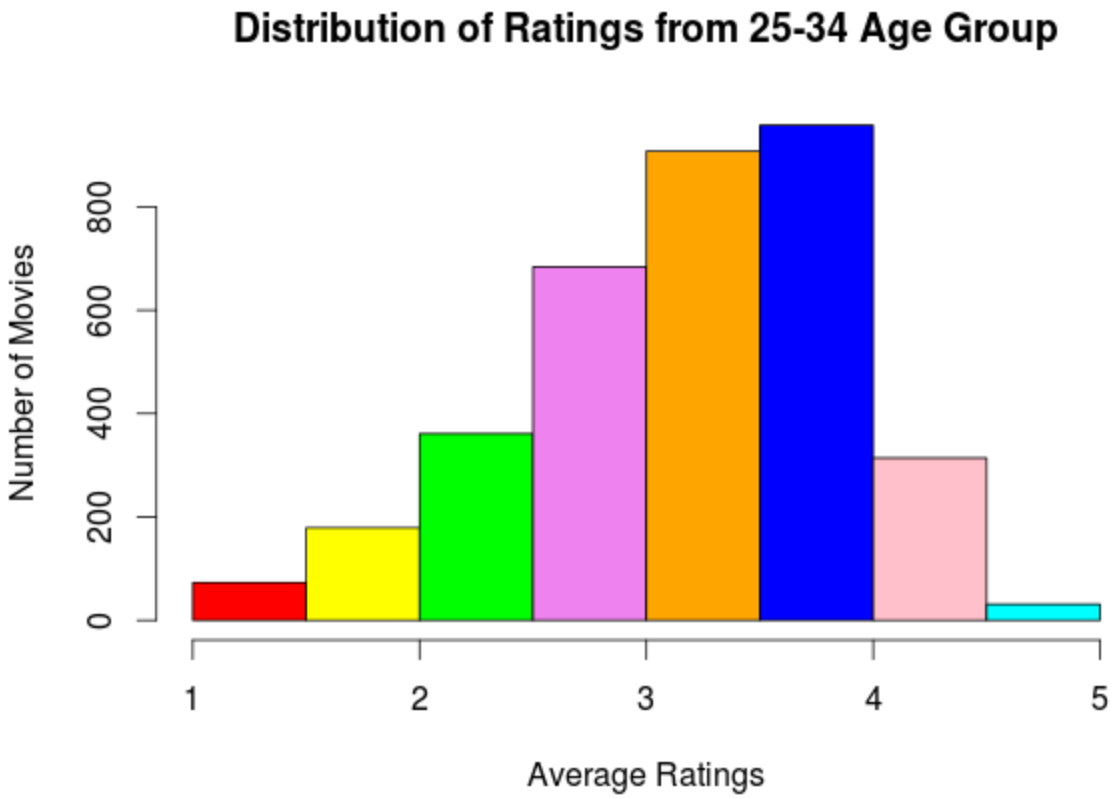
- Ratings from female users only

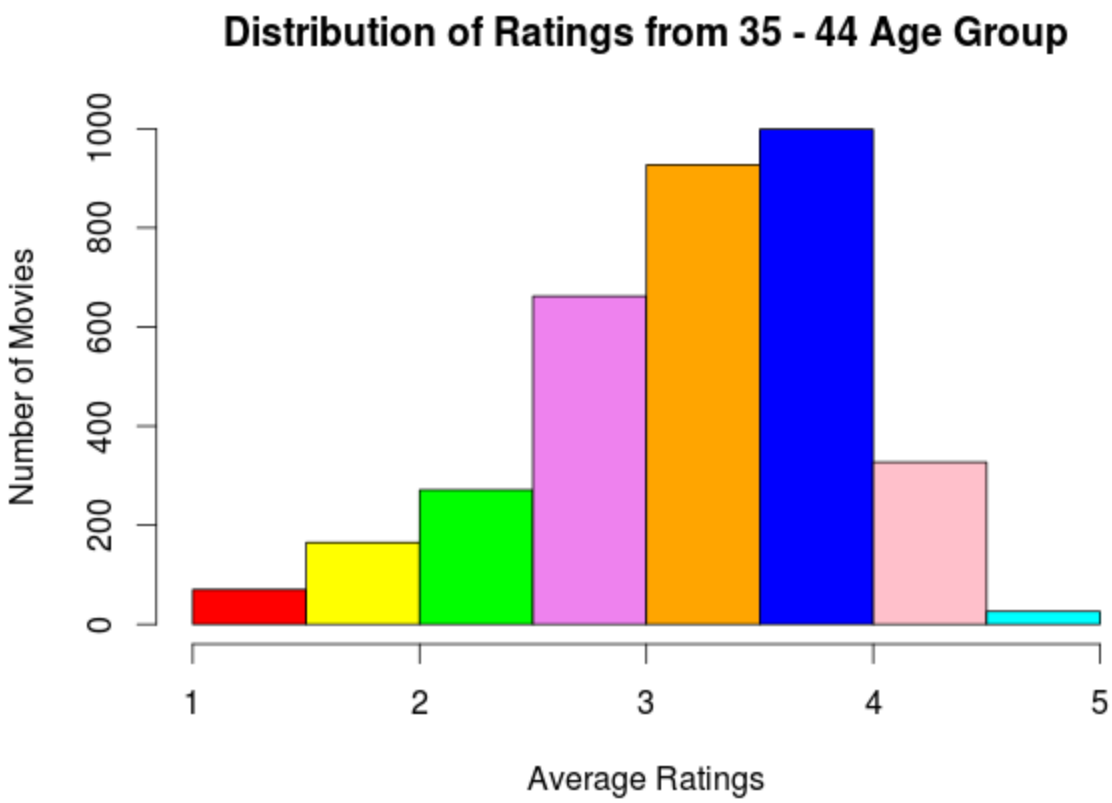


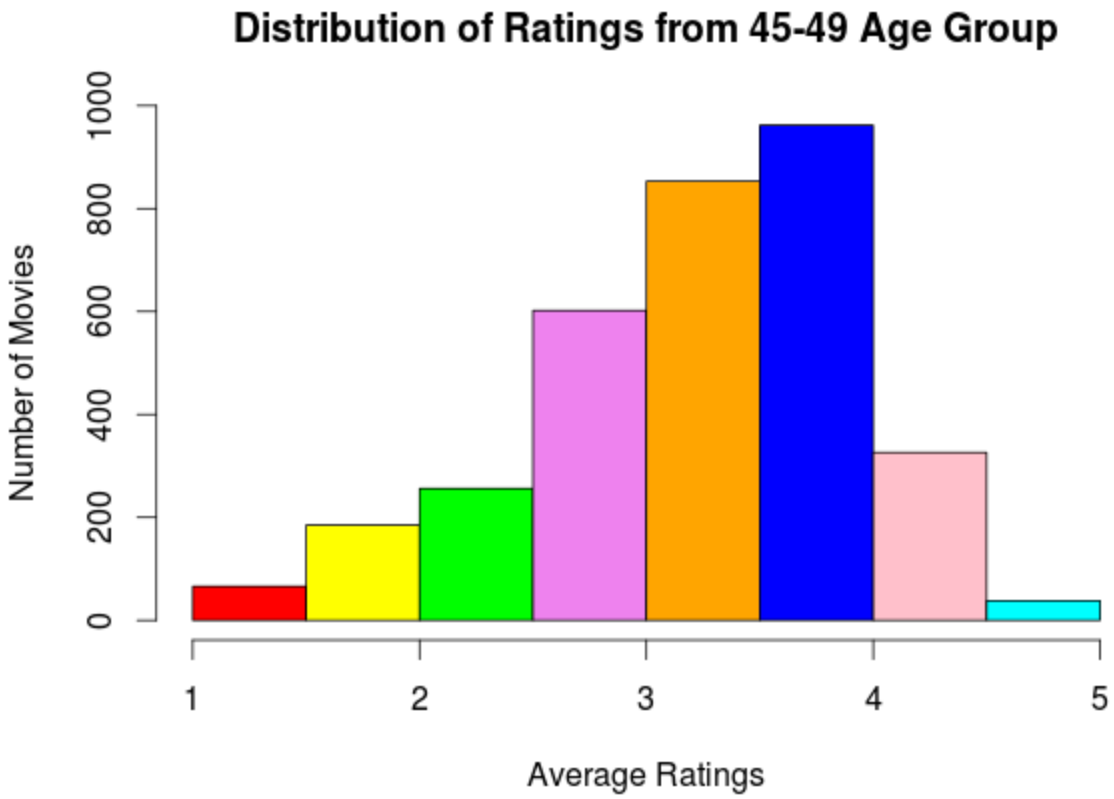
- Rating distribution from different age groups

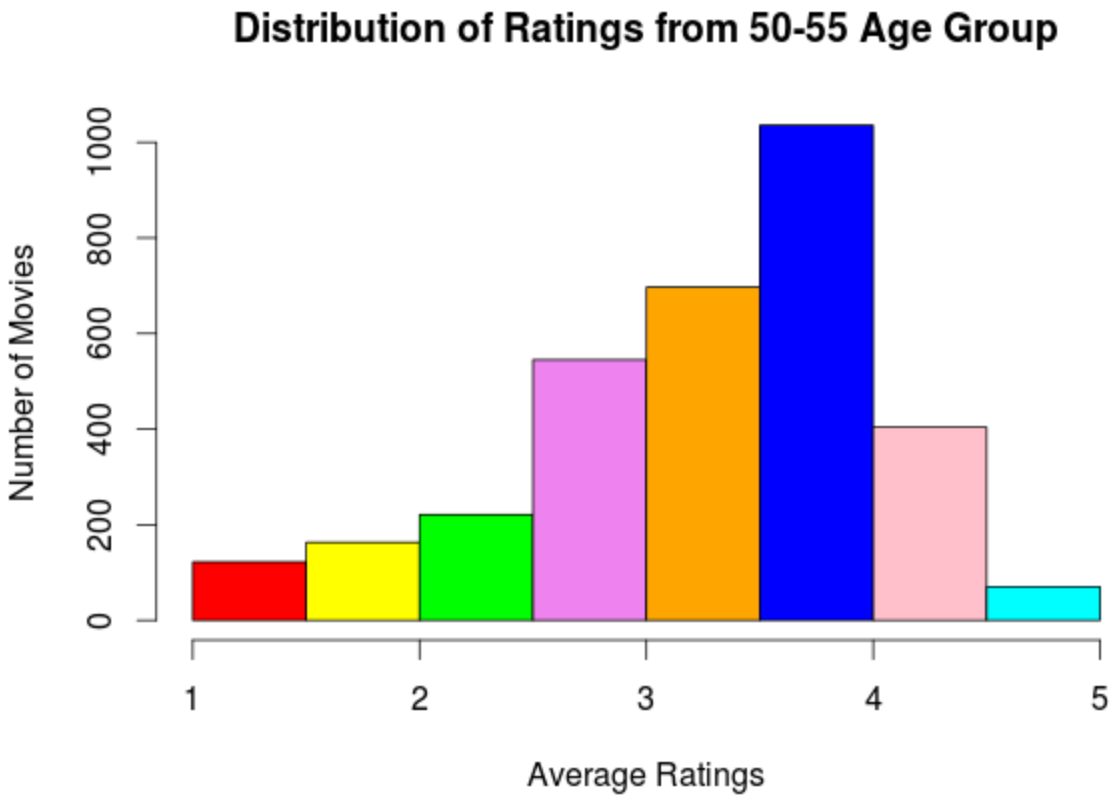

Distribution of Ratings from Under 18 Age Group
Number of Movies
0
100
300
500
700
1
2
3
4
5
Average Ratings

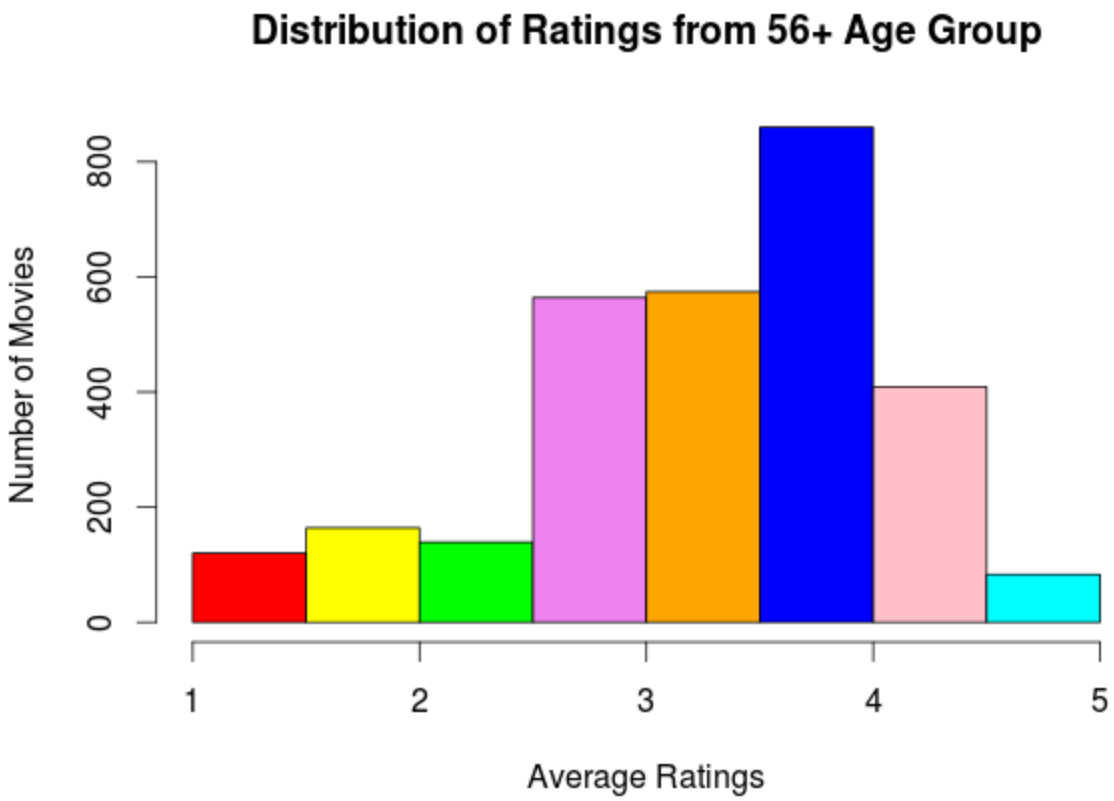

Distribution of Ratings from 18-24 Age Group
Number of Movies
0
200
400
600
800
1
2
3
4
5
Average Ratings


Distribution of Ratings from 25-34 Age Group
Number of Movies
0
200
400
600
800
1
2
3
4
5
Average Ratings

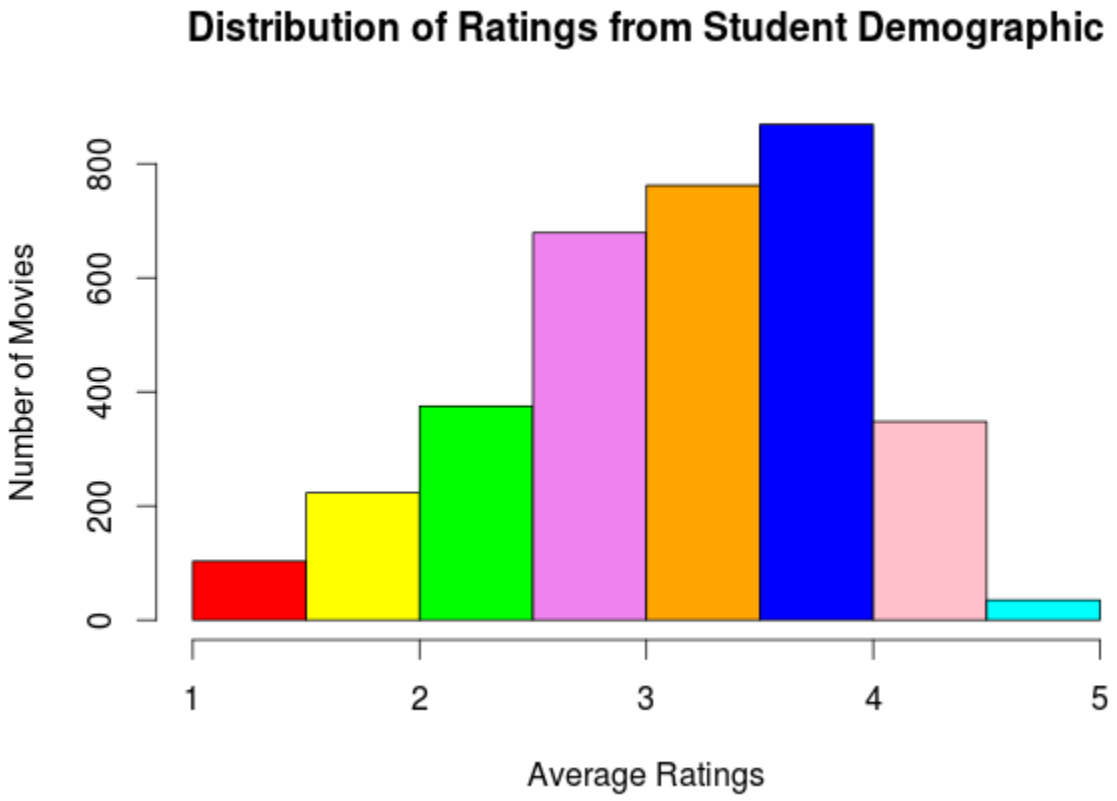

Distribution of Ratings from 35 - 44 Age Group
Number of Movies
0
200
400
600
800
1000
1
2
3
4
5
Average Ratings

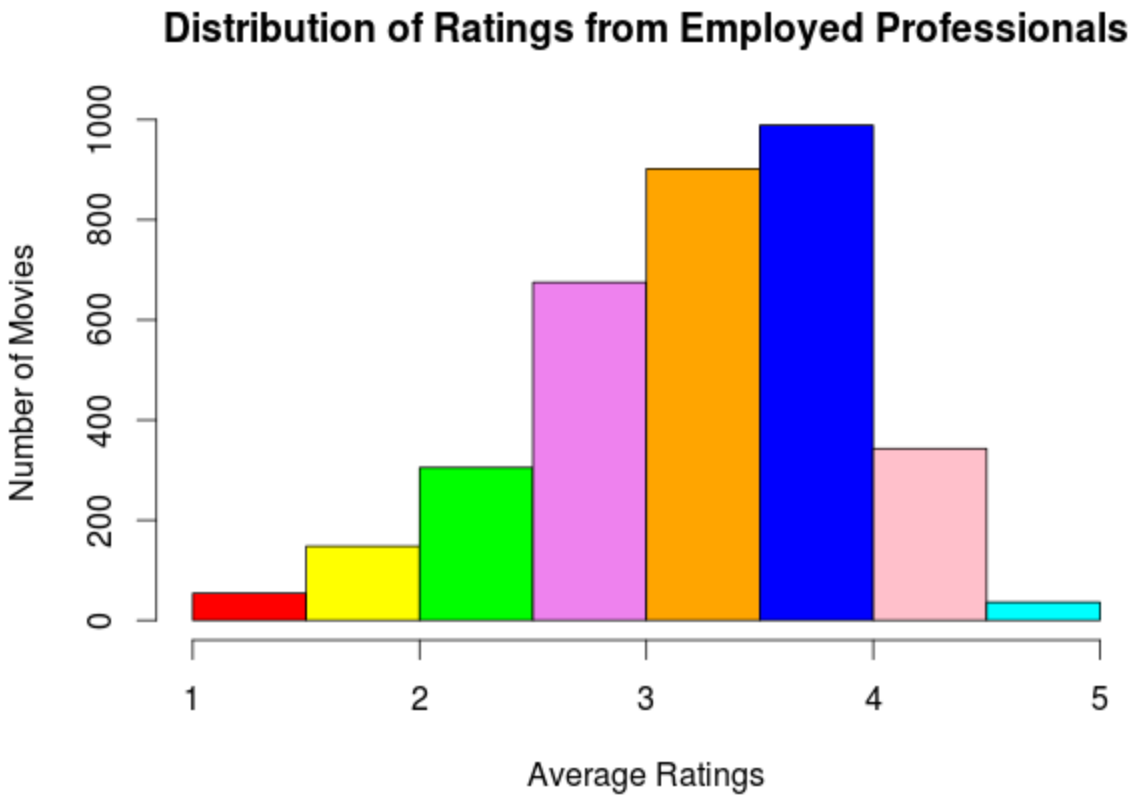

Distribution of Ratings from 45-49 Age Group
Number of Movies
0
200
400
600
800
1000
1
2
3
4
5
Average Ratings

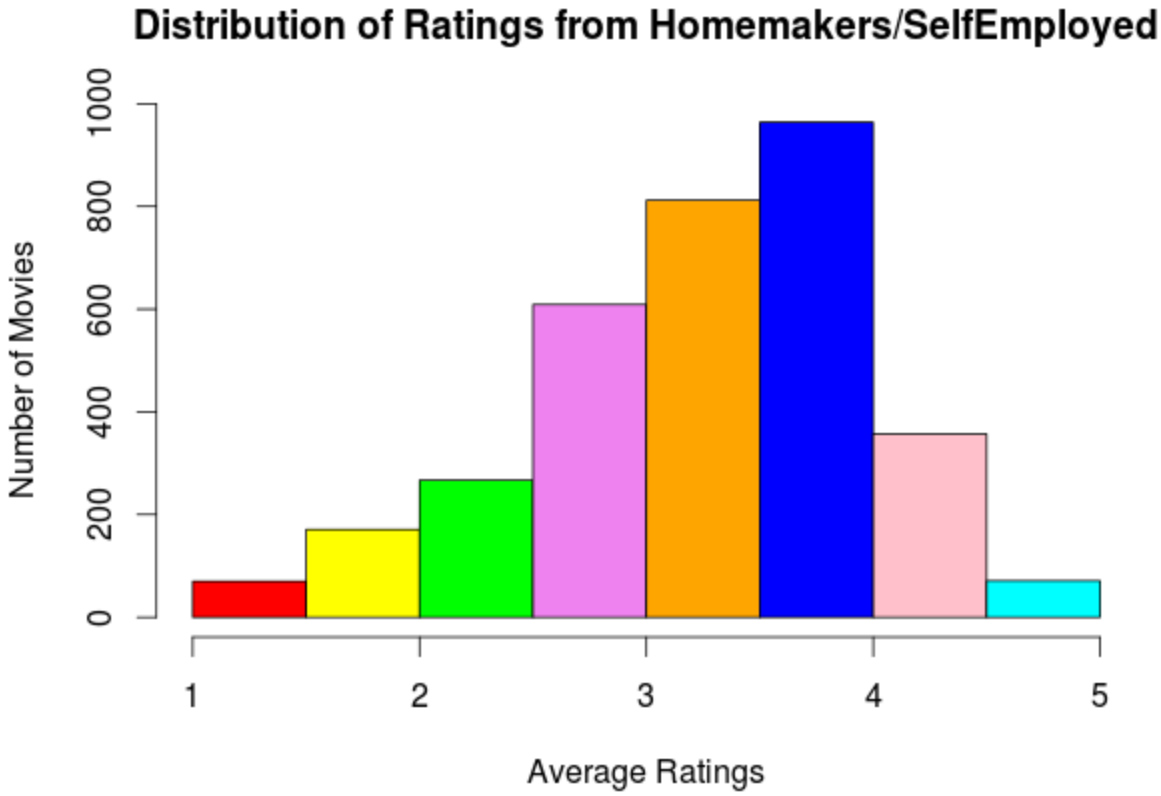

Distribution of Ratings from 50-55 Age Group
Number of Movies
0
200
400
600
800
1000
1
2
3
4
5
Average Ratings

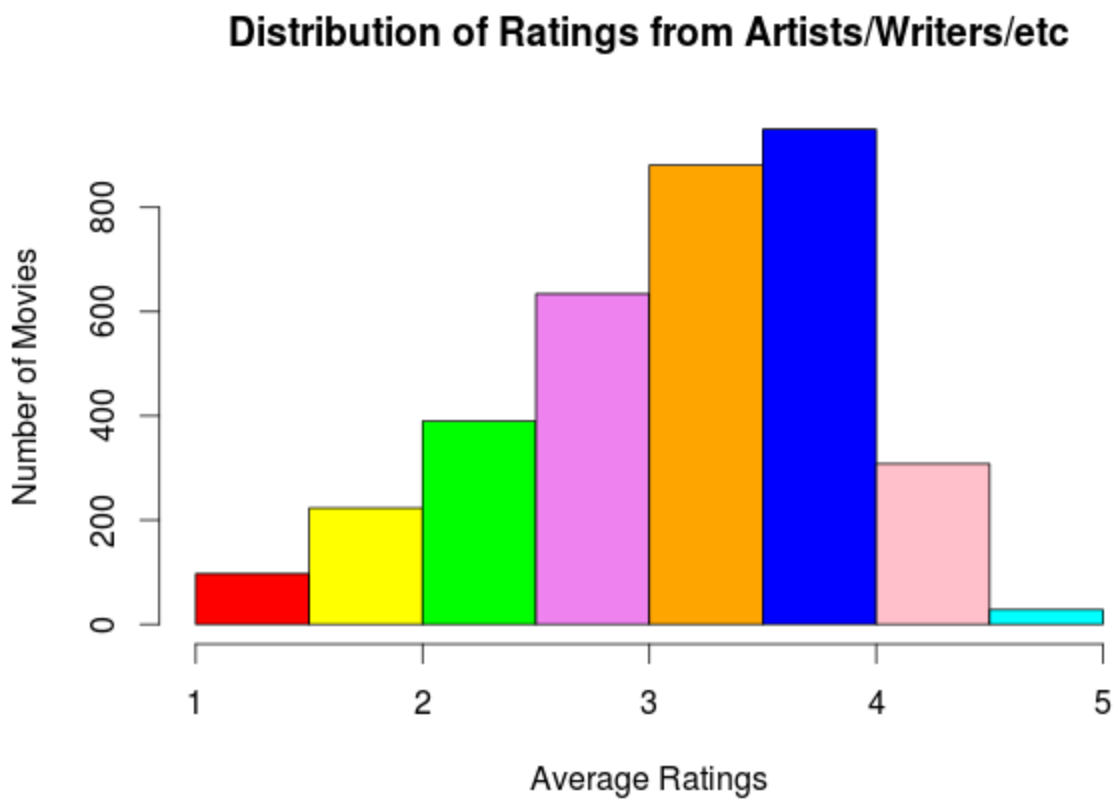

Distribution of Ratings from 56+ Age Group
Number of Movies
0
200
400
600
800
1
2
3
4
5
Average Ratings

- Ratings Distribution from Users belonging to different occupation




Distribution of Ratings from Employed Professionals
Number of Movies
0
200
400
600
800
1000
1
2
3
4
5
Average Ratings


Distribution of Ratings from Homemakers/SelfEmployed
Number of Movies
0
200
400
600
800
1000
1
2
3
4
5
Average Ratings



4. Conclusion

From the above graphs we can infer that the majority of ratings received by the movies falls in the 3.0 – 4.0 range which is understandable. What is interesting however is how the higher that average ratings are given. Younger users tend to be liberal with giving higher ratings to movies and so are the those above 50. Users falling in the middle age group tend to give no so many 4.0/5.0 ratings. When we see the occupation based segments, clearly people belonging to intellectual pursuits are very stingy ratings. They tend to give mostly 2.5 - 4.0 ratings compared to the rest.

Lessons Learned

New York Times Dataset

- New York Times has mature reader.
- The NYT site gets most traffic on Sunday.
- Mostly, female readers.

RealDirect Dataset

- Bay Bridge has biggest apartments for sale.
- Bedford Stuyvesant neighborhood had most sales
- Christmas time is the best for sales.

MovieLens Dataset

- Young users give good ratings.
- Older users give lower ratings.