**Report on data analysis using python**

**About dataset:**

The dataset provided contains movie reviews given by Amazon customers. Reviews were given between May 1996 and July 2014.

**Data Dictionary:**

UserID – 4848 customers who provided a rating for each movie

Movie 1 to Movie 206 – 206 movies for which ratings are provided by 4848 distinct users

**Data Considerations:**

All the users have not watched all the movies and therefore, all movies are not rated. These missing values are represented by NA.

Ratings are on a scale of -1 to 10 where -1 is the least rating and 10 is the best.

We use python libraries called pandas for exploratory analysis then matplotlib for visualizations and sklearn for regression analysis.

First we prepare data by analyzing it in all possible ways.

After analysis the prepared dataset is then used to answer the questions in problem statement.

The prepared dataset is then divided into test and train datasets and a model is trained and evaluated to check the performance of model.

**Results and outcomes of project:**

After completing this project we can answer the following questions.

Which movies have maximum views/ratings?

What is the average rating for each movie? Define the top 5 movies with the maximum ratings.

Define the top 5 movies with the least audience.

**Model building:**

Some of the movies hadn’t been watched and therefore, are not rated by the users. Netflix would like to take this as an opportunity and build a machine learning recommendation algorithm which provides the ratings for each of the users.

Divide the data into training and test data

Build a recommendation model on training data

Make predictions on the test data

After building a recommendation model we can now get the ratings of movies.

**Screenshots of project work done:**

The jupyter notebook itself is very descriptive about the project with proper markdown





