Python Movielens Project Source Code

import pandas as pd

import numpy as np

movies\_dataset = pd.read\_table('C:\\Data Science with Python\\Project\\movies.dat', sep='::',

names=['MovieID','Title', 'Genres'])

movies\_dataset.shape

movies\_dataset.head()

users\_dataset = pd.read\_table('C:\\Data Science with Python\\Project\\users.dat', sep='::',

names=['UserID', 'Gender','Age','Occupation', 'Zip-code'])

users\_dataset.head()

ratings\_dataset = pd.read\_table('C:\\Data Science with Python\\Project\\ratings.dat', sep='::',

names=['UserID', 'MovieID', 'Rating', 'Timestamp'])

ratings\_dataset.head()

import matplotlib.pyplot as plt

from matplotlib import style

%matplotlib inline

import seaborn as sns

# show age distribution

sns.set()

plt.hist(users\_dataset['Age'], bins=12)

plt.xlabel('Age')

plt.ylabel('Frequency')

plt.title('Age Distribution of Users')

Plt.show

# show rating distribution

plt.hist(ratings\_dataset['Rating'])

plt.xlabel('Rating')

plt.ylabel('Frequency')

plt.title('Overall Rating by Users')

Plt.show

df\_movieID=pd.merge(ratings\_dataset,movies\_dataset, on='MovieID')

df\_toystory = df\_movieID[df\_movieID['Title']=='Toy Story (1995)']

df\_toystory

# show user rating of toy story

plt.hist(df\_toystory['Rating'])

plt.xlabel('Rating')

plt.ylabel('Frequency')

plt.title('User Ratings of Toy Story')

Plt.show

df\_toystoryusers=pd.merge(df\_toystory,users\_dataset, on='UserID')

# show age viewership of toystory

plt.hist(df\_toystoryusers['Age'], bins=12)

plt.xlabel('Age')

plt.ylabel('Frequency')

plt.title('Toy Story viewership by Age')

Plt.show

df\_all = pd.merge(df\_movieID, users\_dataset)

# shows the top 25 movies by viewership rating

Movieviews = df\_movieID.groupby('Title').size().sort\_values(ascending=False).head(25)

Movieviews

usergrp = ratings\_dataset.groupby('UserID')

user2696 = usergrp.get\_group(2696)

User2696

# show rating data by user 2696

plt.hist(user2696['Rating'])

plt.xlabel('Rating')

plt.ylabel('Frequency')

plt.title('Rating data of User 2696')

Plt.show

df\_userandrating=pd.merge(ratings\_dataset,users\_dataset, on='UserID').head(500)

df\_userandrating

x\_feature = df\_userandrating[['MovieID','Occupation','Age']]

x\_feature.head()

y\_target = df\_userandrating[['Rating']]

x\_feature.shape

y\_target.shape

from sklearn.cross\_validation import train\_test\_split

x\_train, x\_test, y\_train, y\_test = train\_test\_split(x\_feature, y\_target,random\_state=1)

print x\_train.shape

print y\_train.shape

print x\_test.shape

print y\_test.shape

from sklearn.linear\_model import LinearRegression

linreg = LinearRegression()

linreg.fit(x\_train,y\_train)

print linreg.intercept\_

print linreg.coef\_

y\_pred = linreg.predict(x\_test)

y\_pred

from sklearn import metrics

print 'MSE value is %.2f '% np.sqrt(metrics.mean\_squared\_error(y\_test,y\_pred))

x\_test.hist()

plt.show()

x\_train.hist()

plt.show()