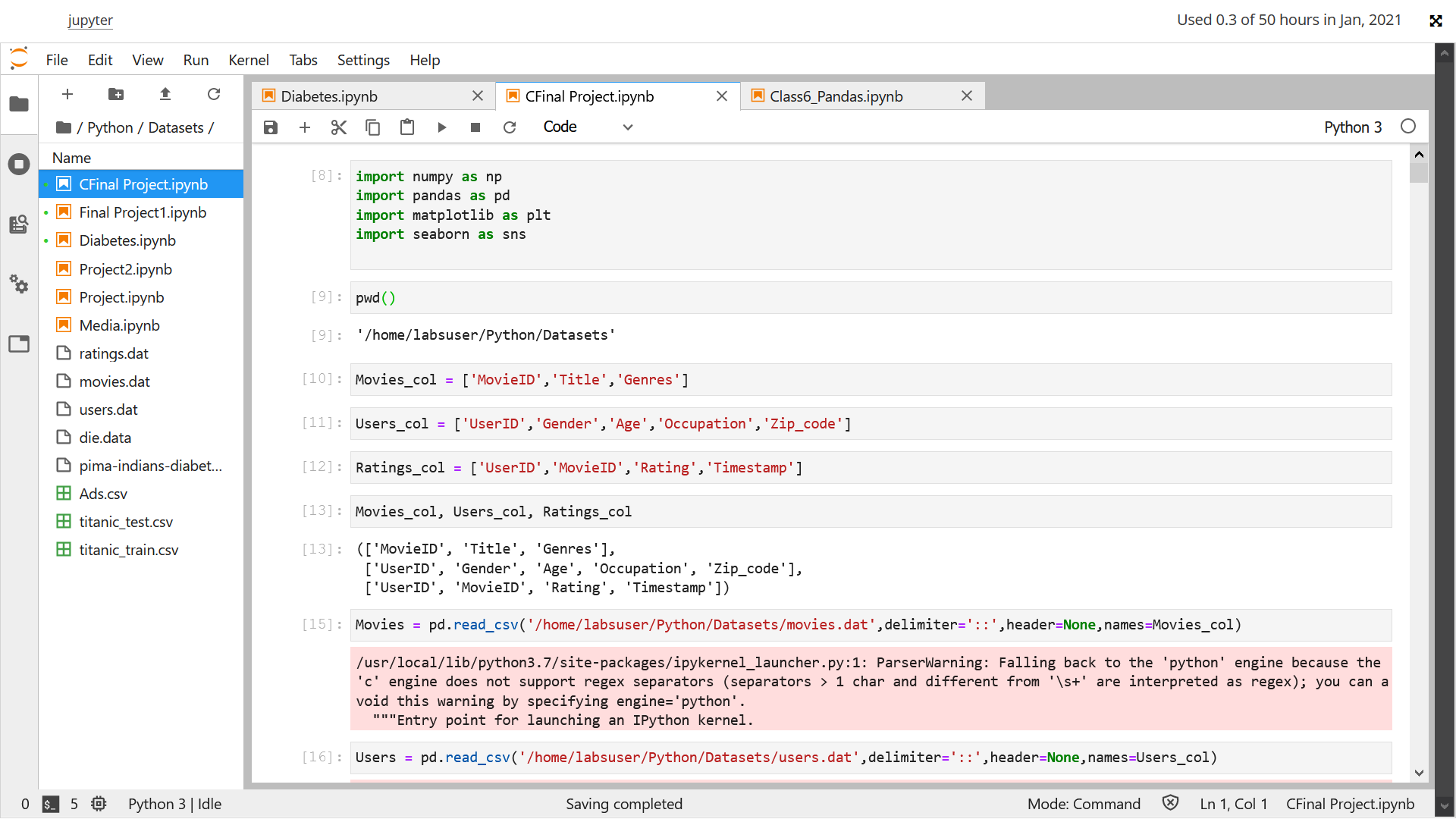
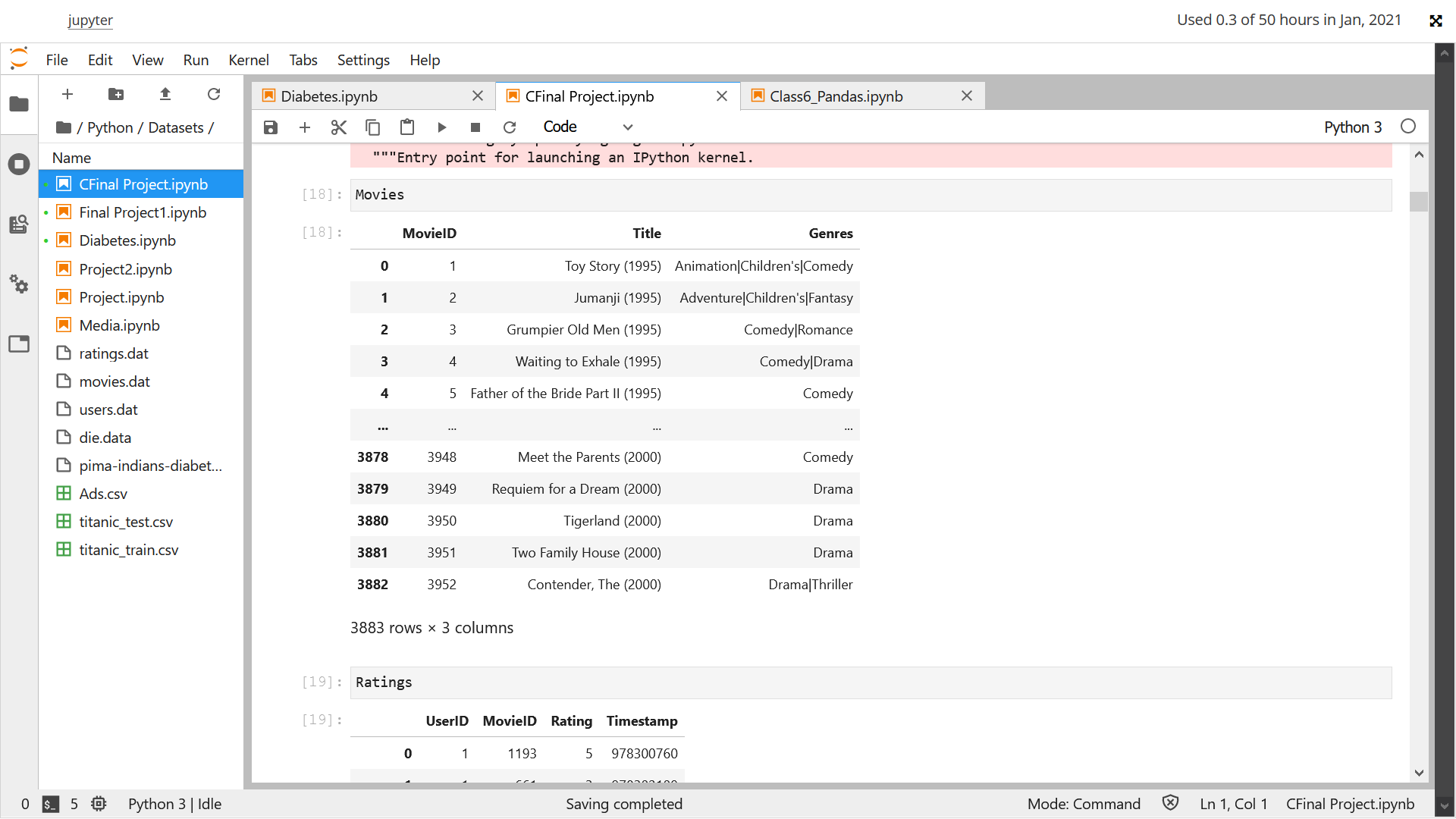
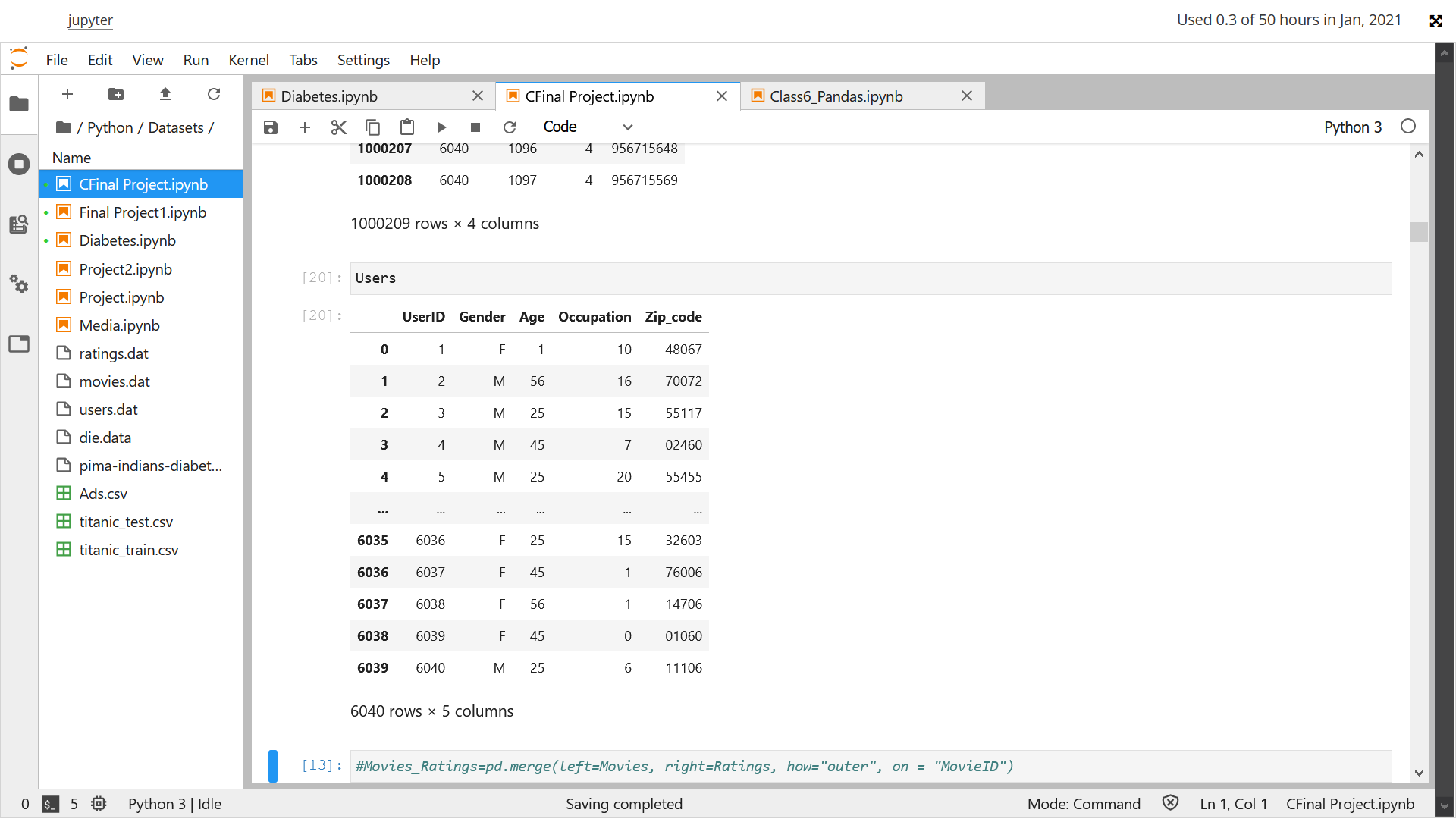
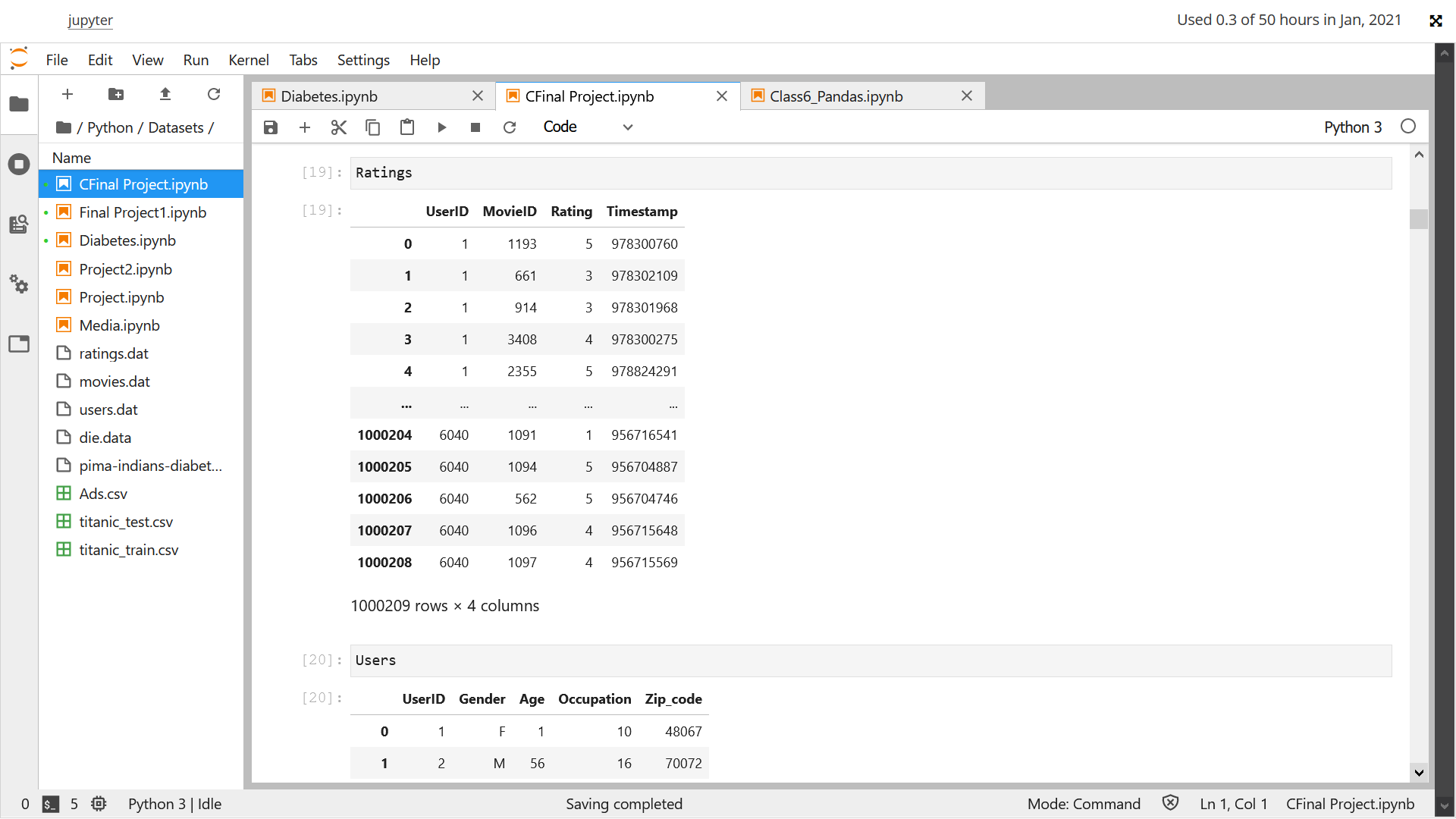
Project 2 - Movie lens case

Analysis task

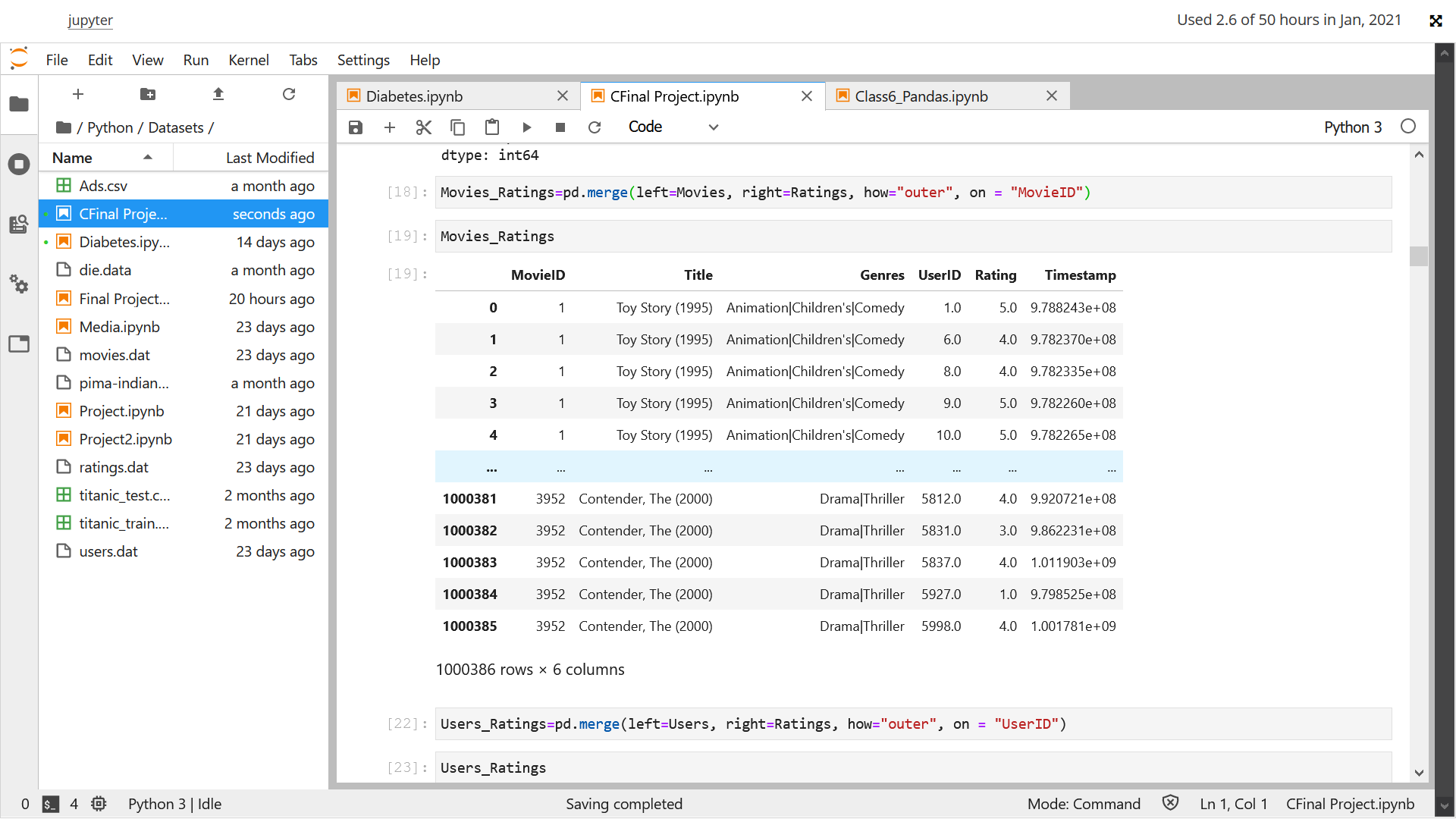
Import three datasets

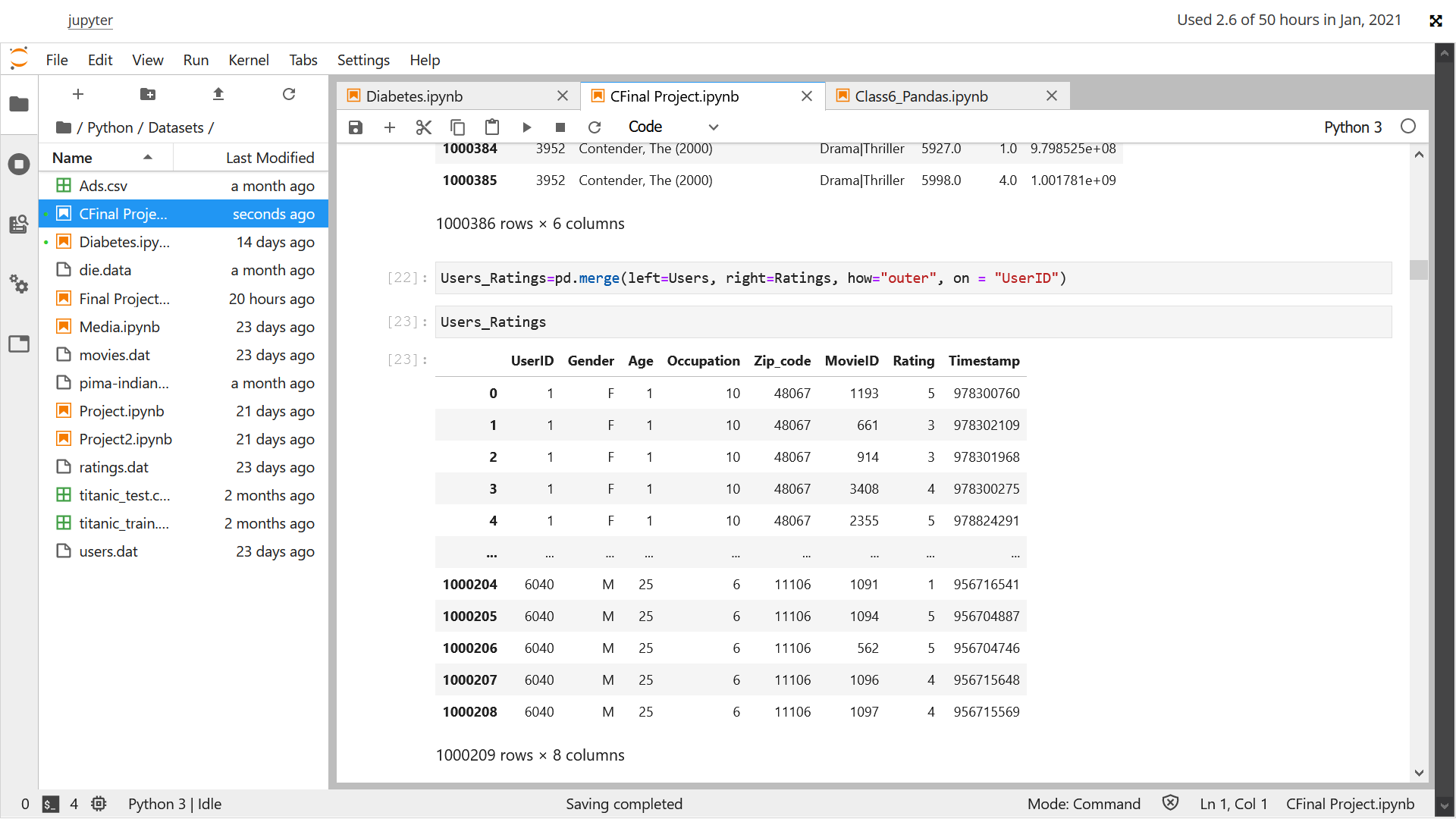
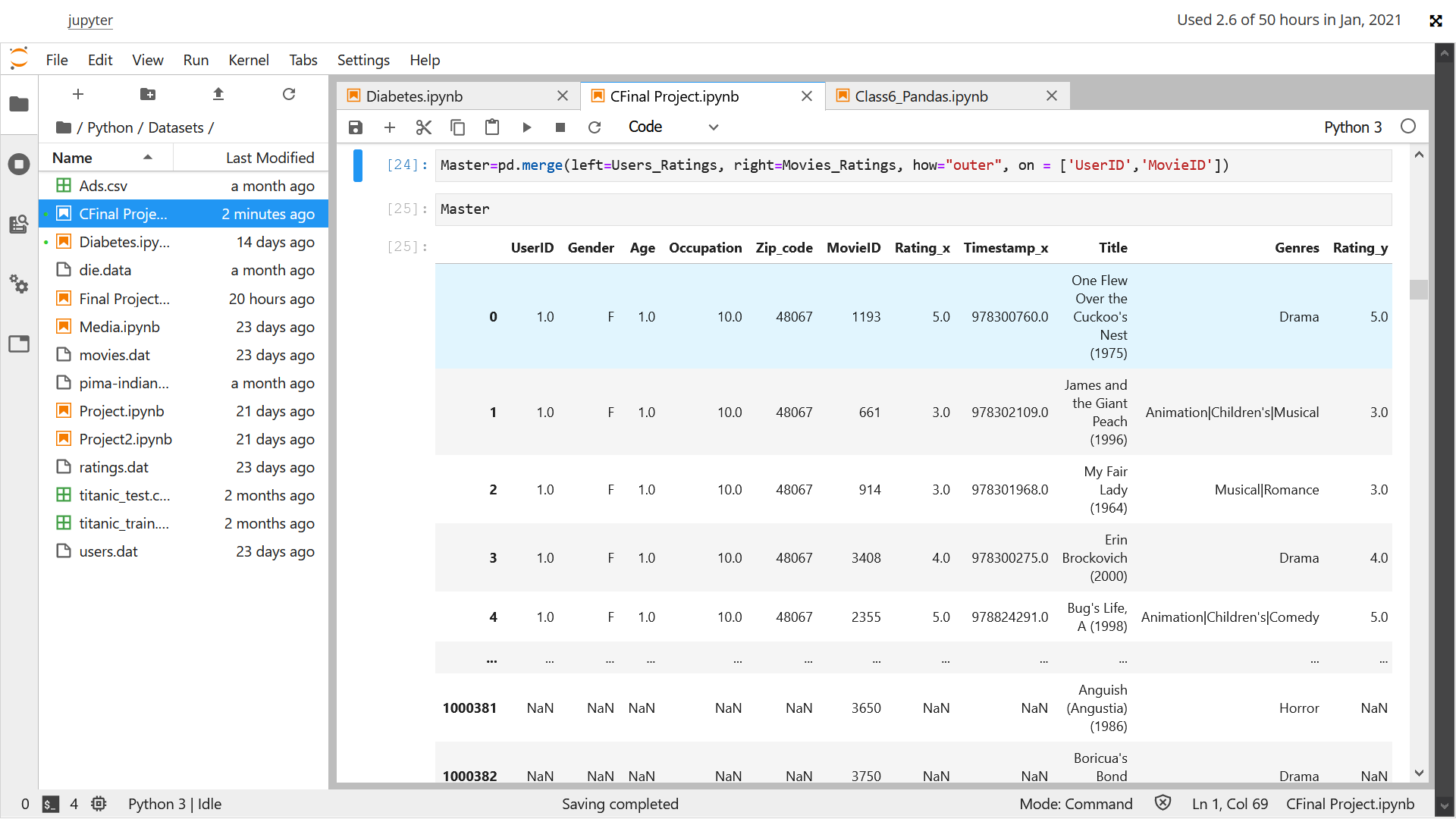


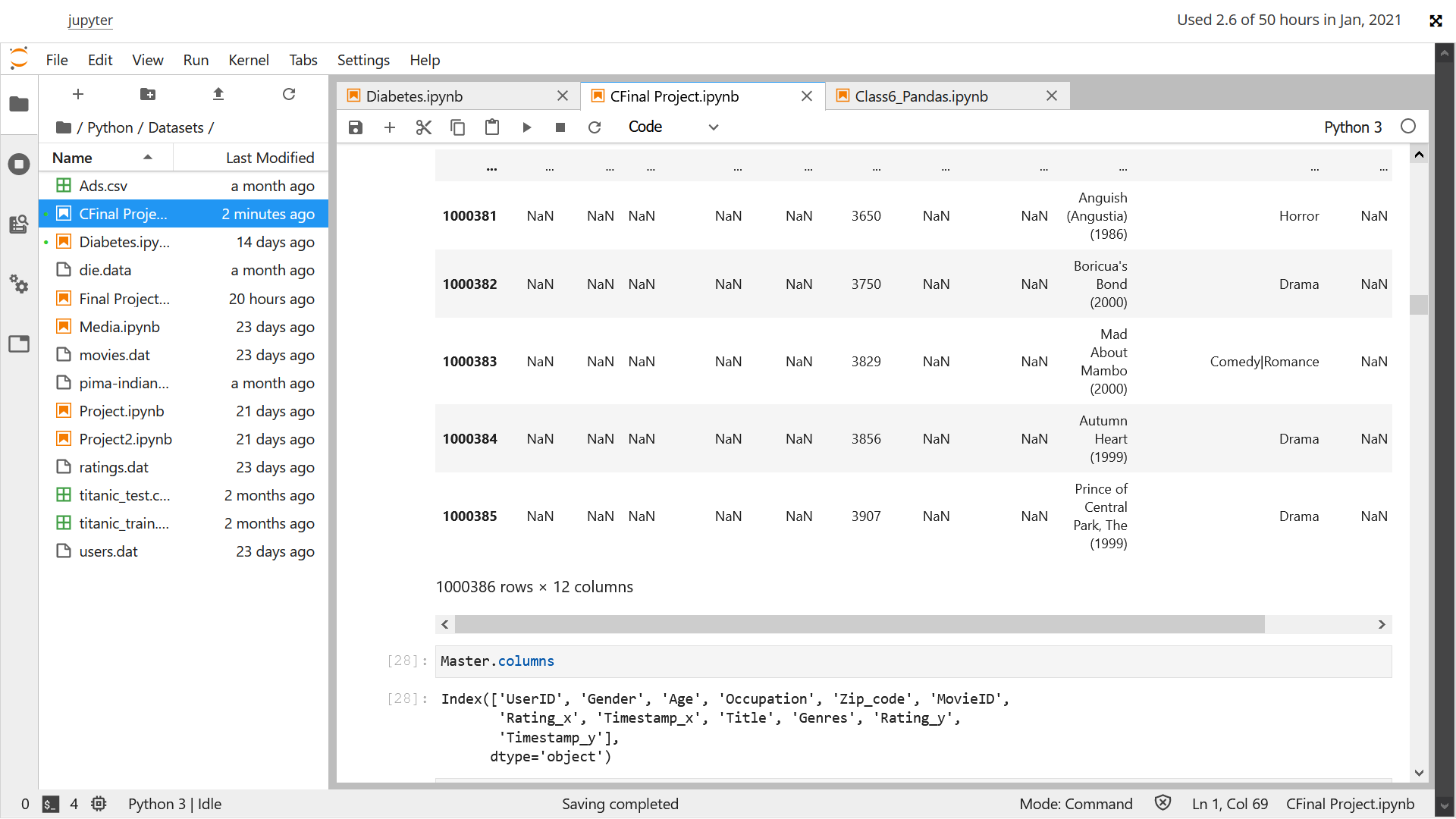


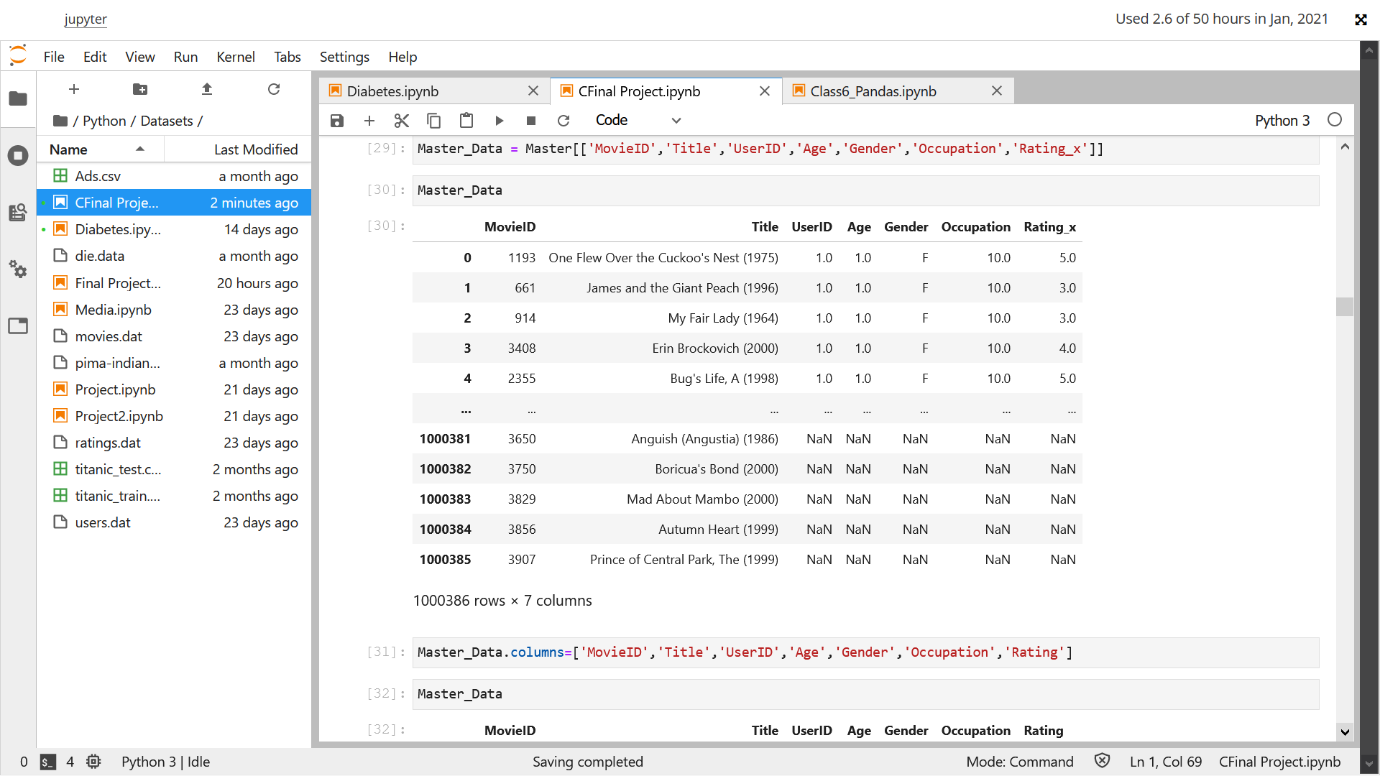


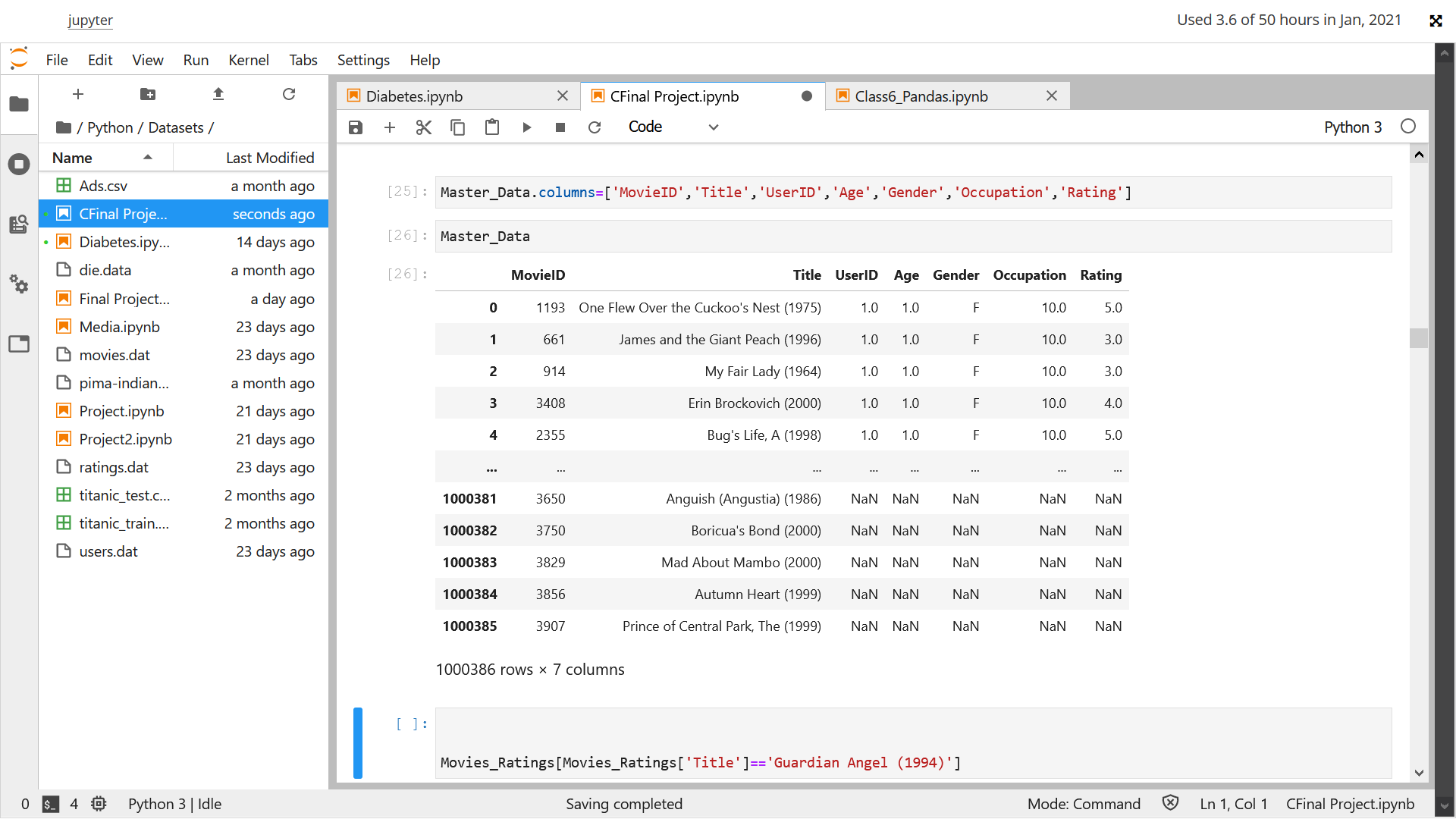
Create a new dataset [Master\_Data] with the following columns MovieID Title UserID Age Gender Occupation Rating. (Hint: (i) Merge two tables at a time. (ii) Merge the tables using two primary keys MovieID & UserId)



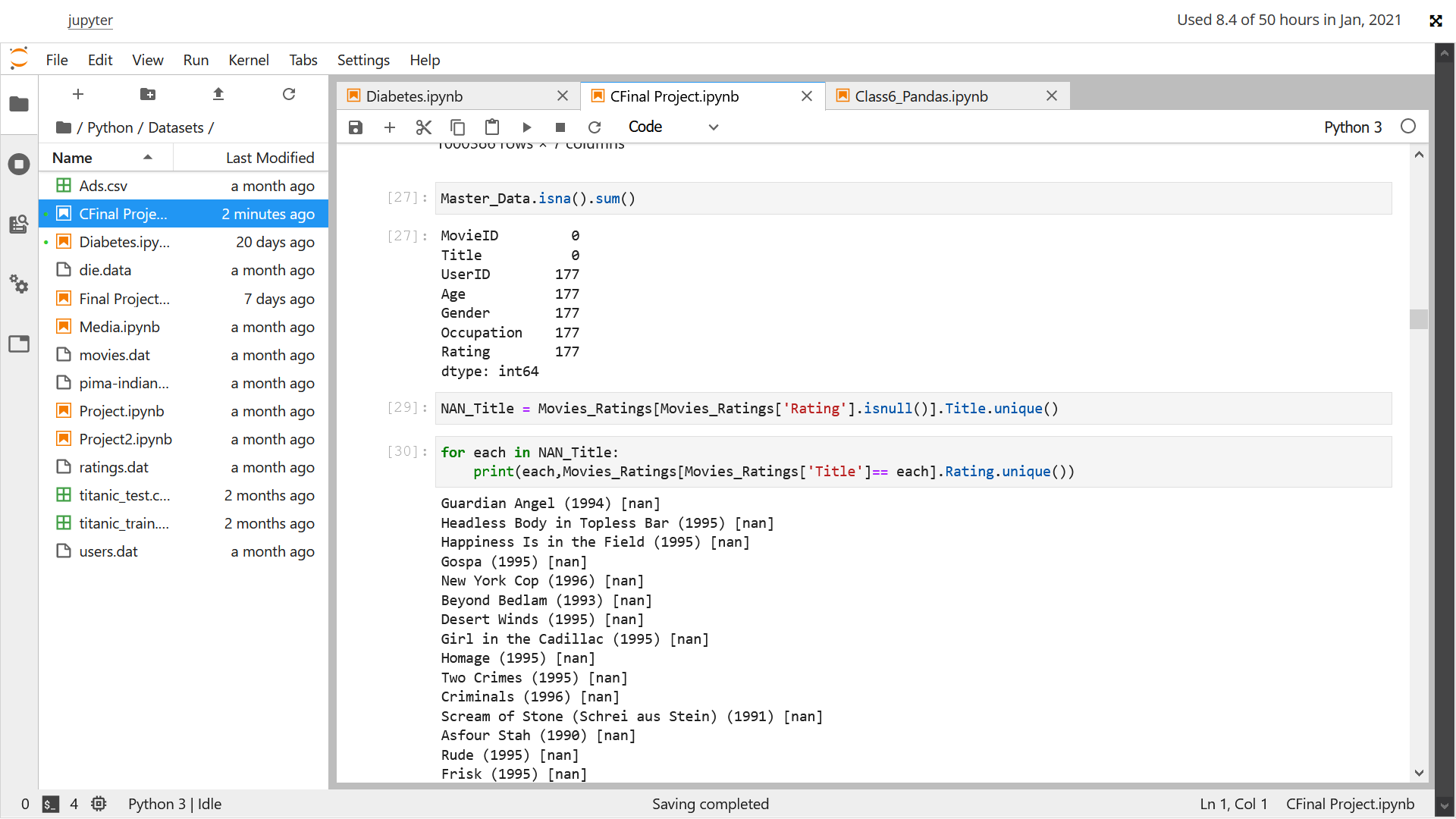




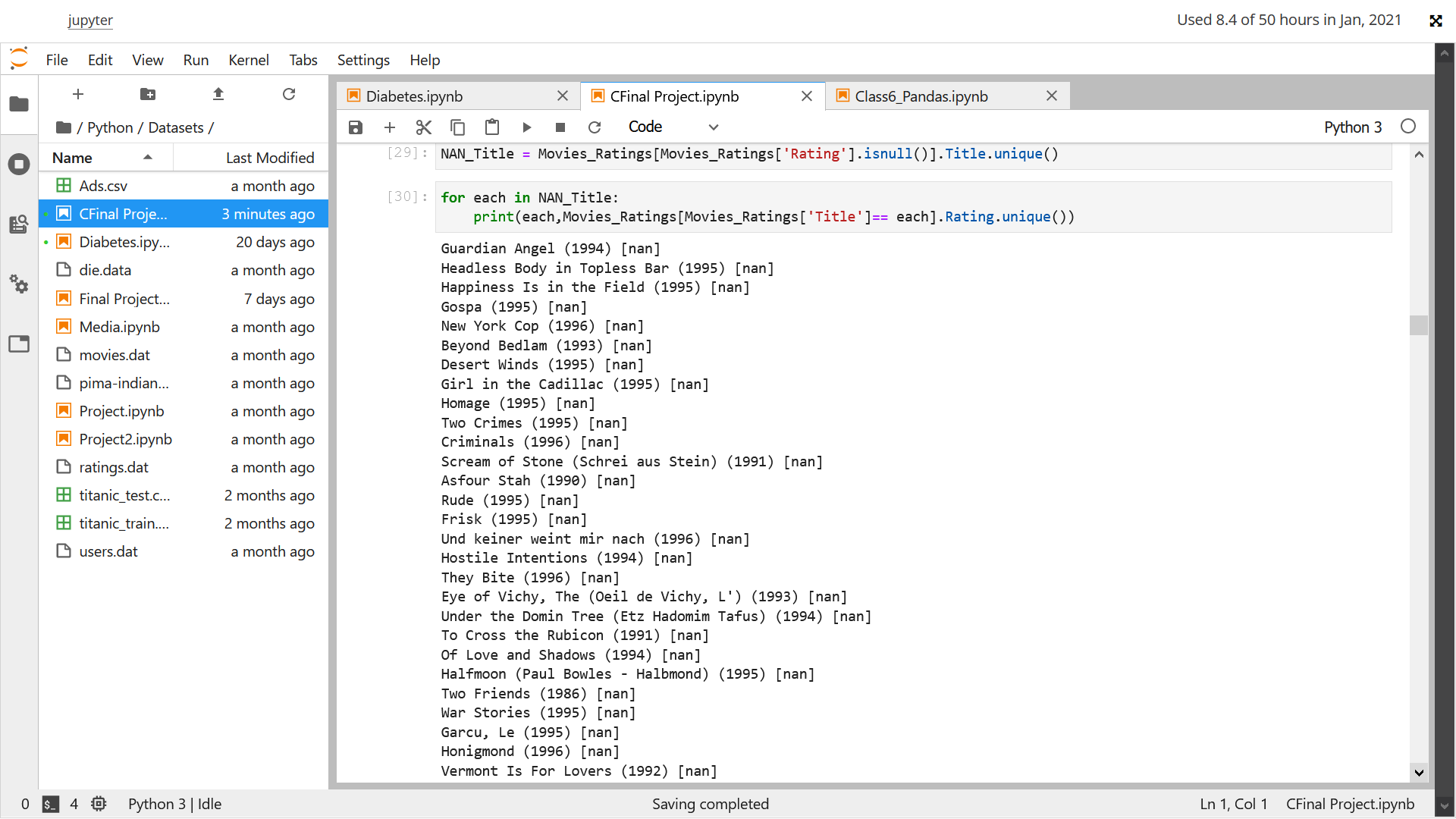


Below is a required Master\_Data

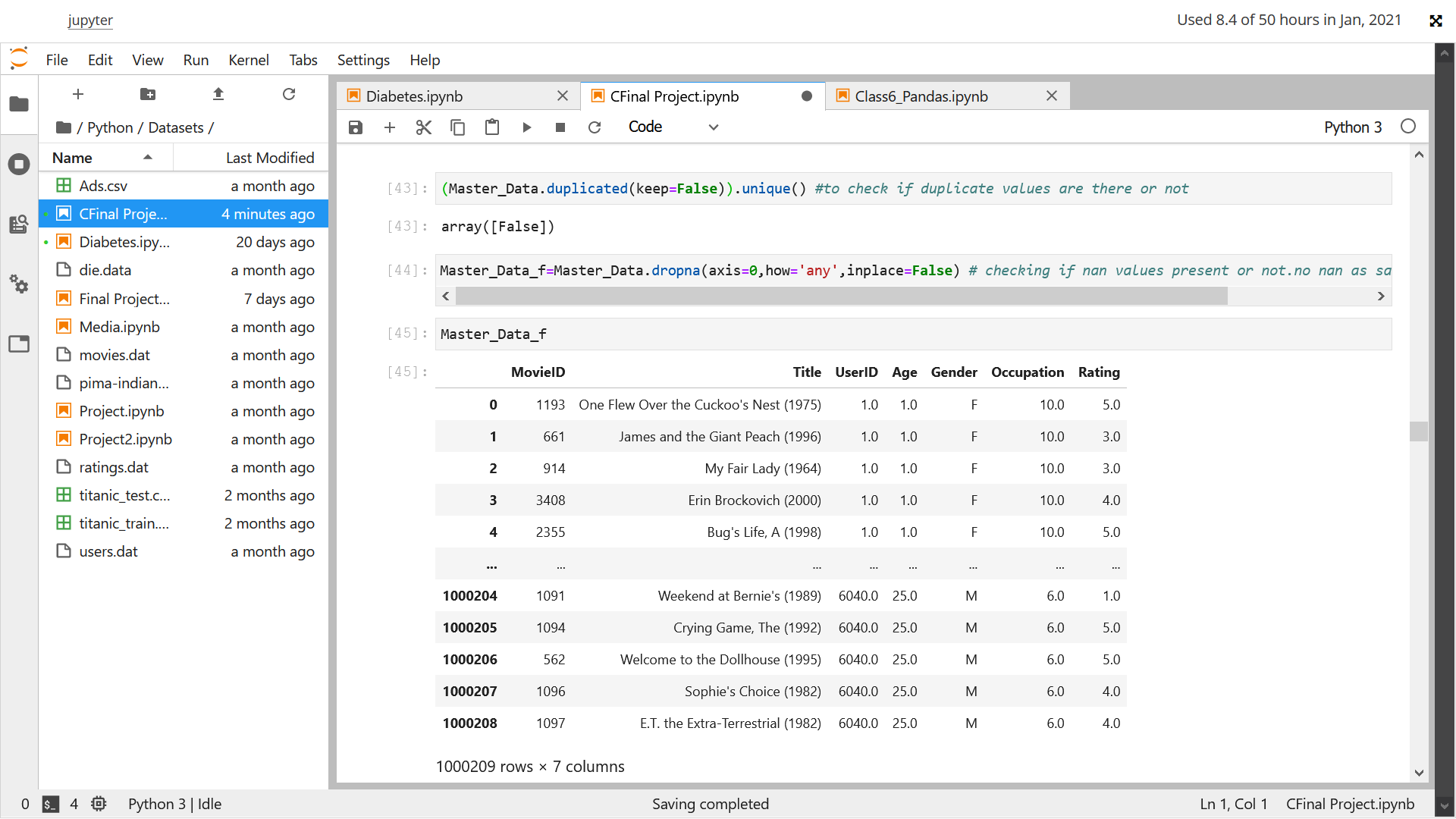
Checking for Nan in Master data :



Titles with no user rating .

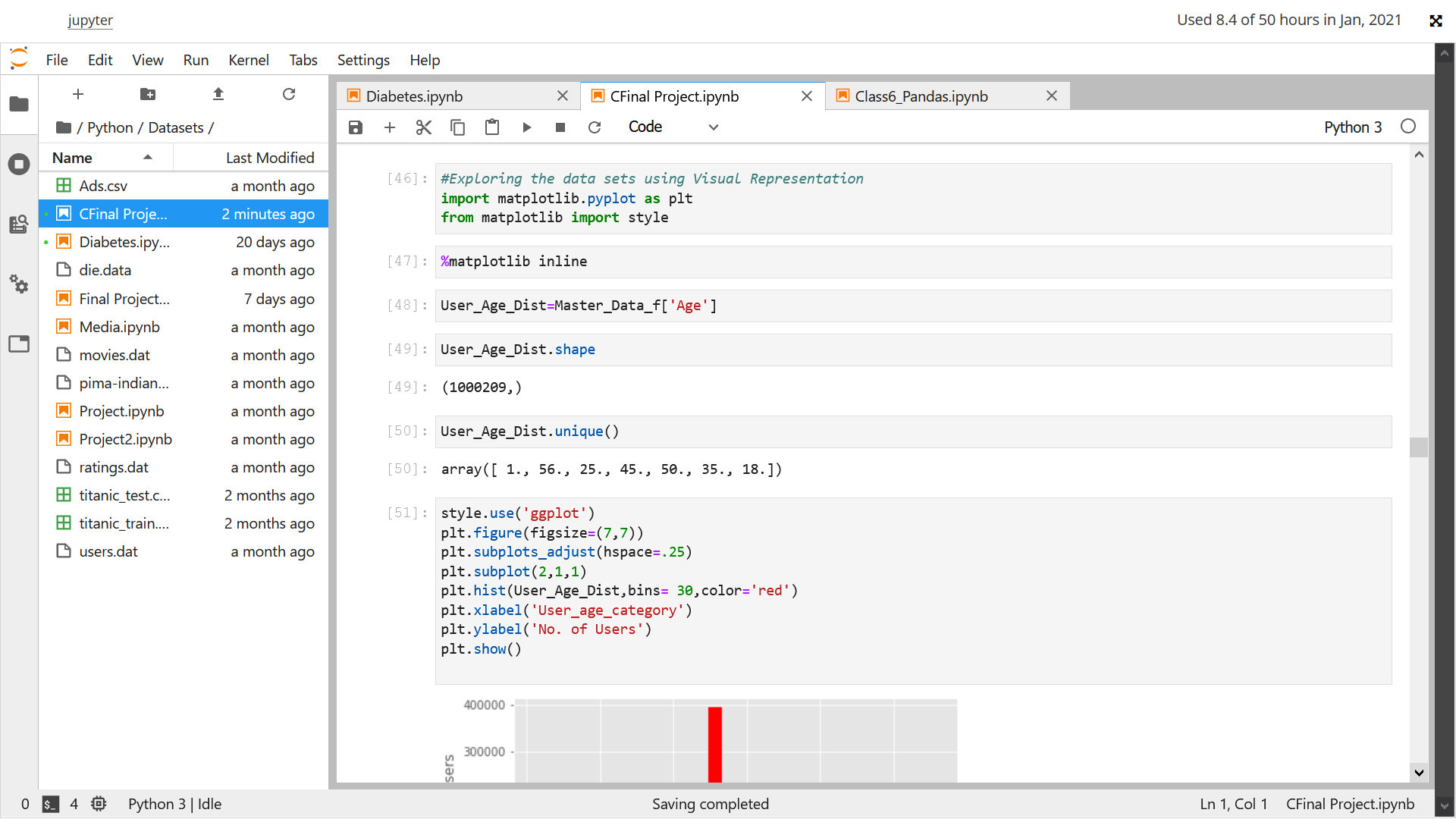


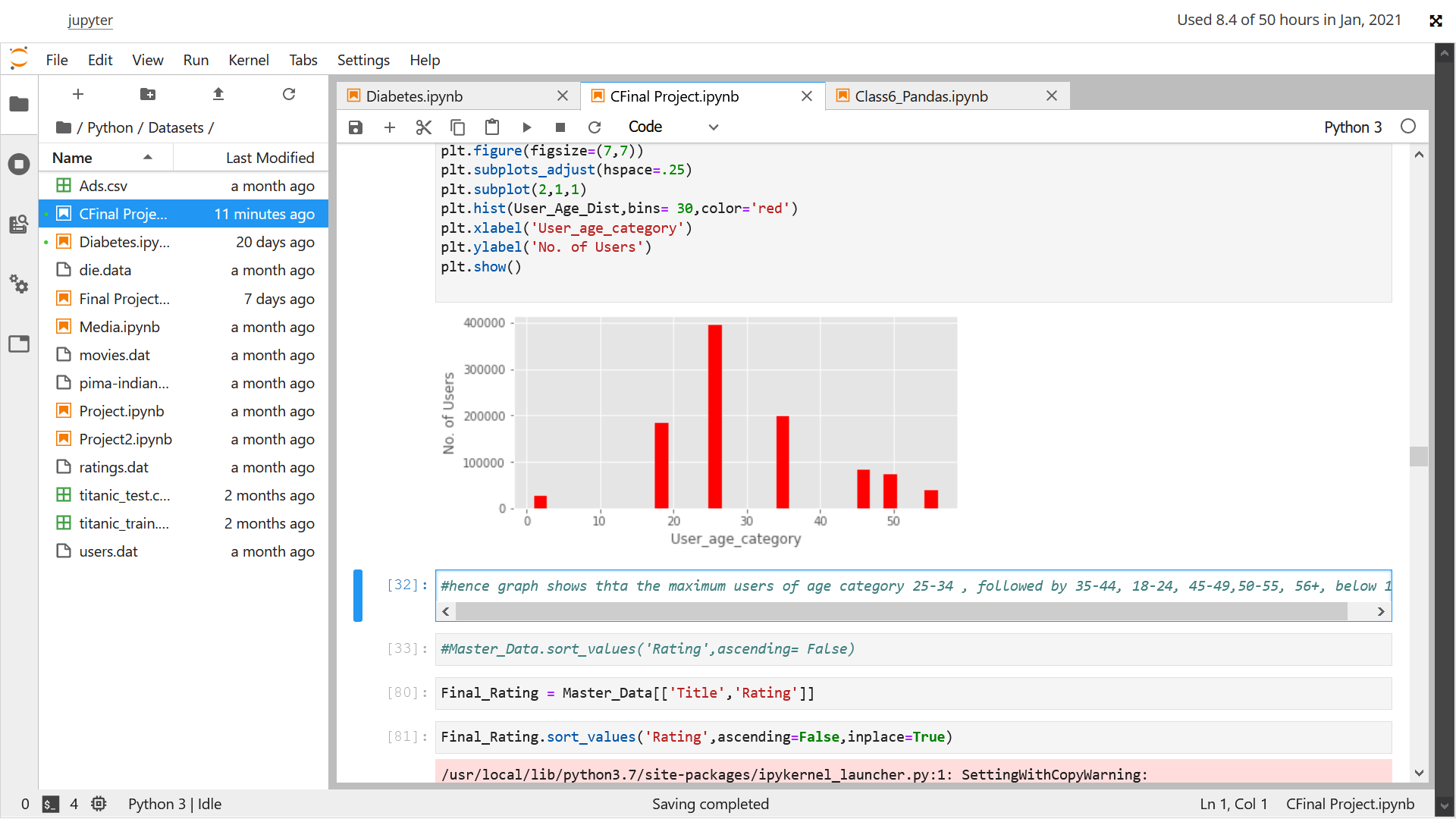
Checking for duplicate values and dropping the Nans as they will not be helpful in the model making



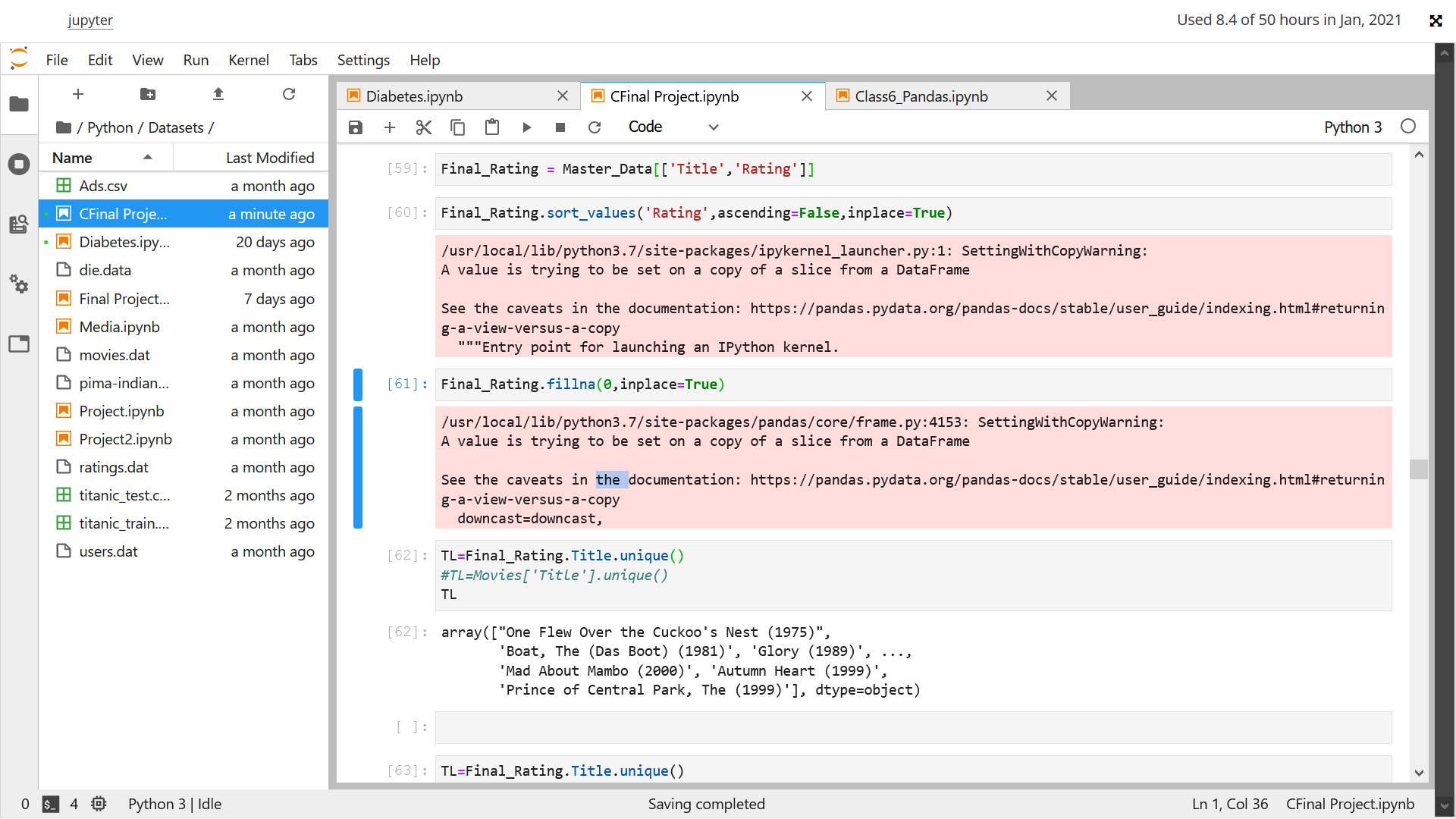
* Explore the datasets using visual representations (graphs or tables), also include your comments on the following:

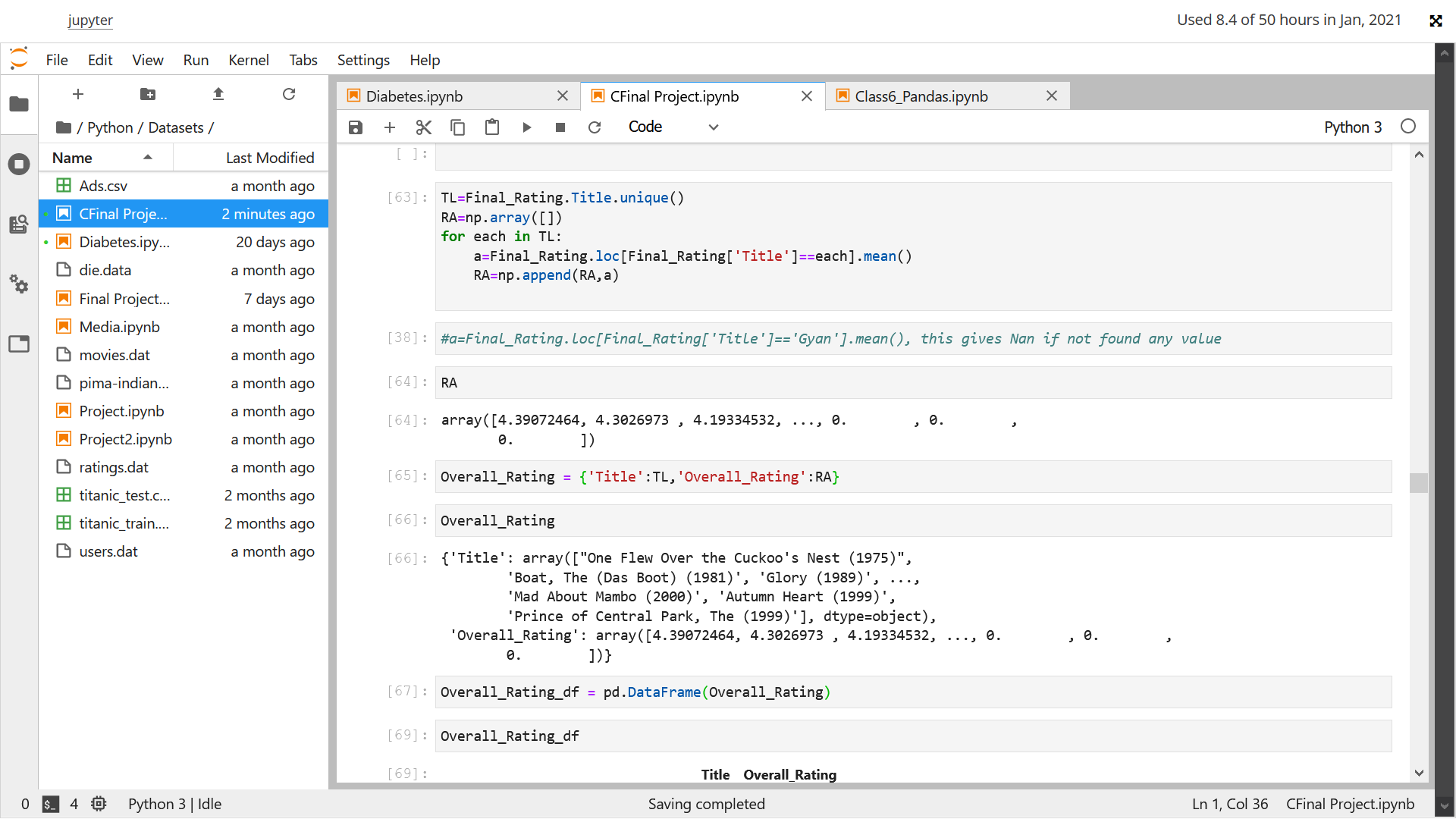
1. User Age Distribution
2. User rating of the movie “Toy Story”
3. Top 25 movies by viewership rating
4. Find the ratings for all the movies reviewed by for a particular user of user id = 2696

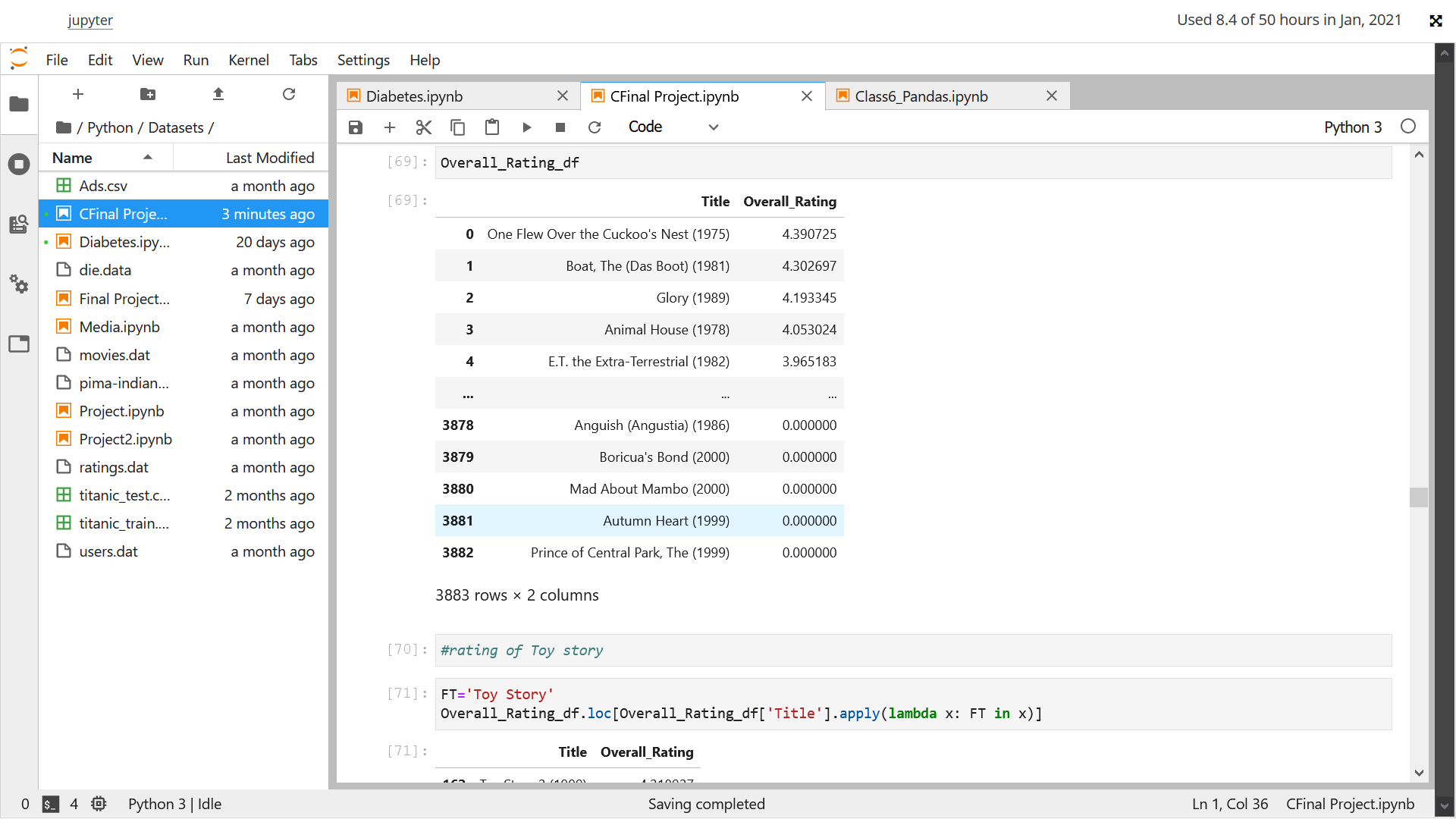


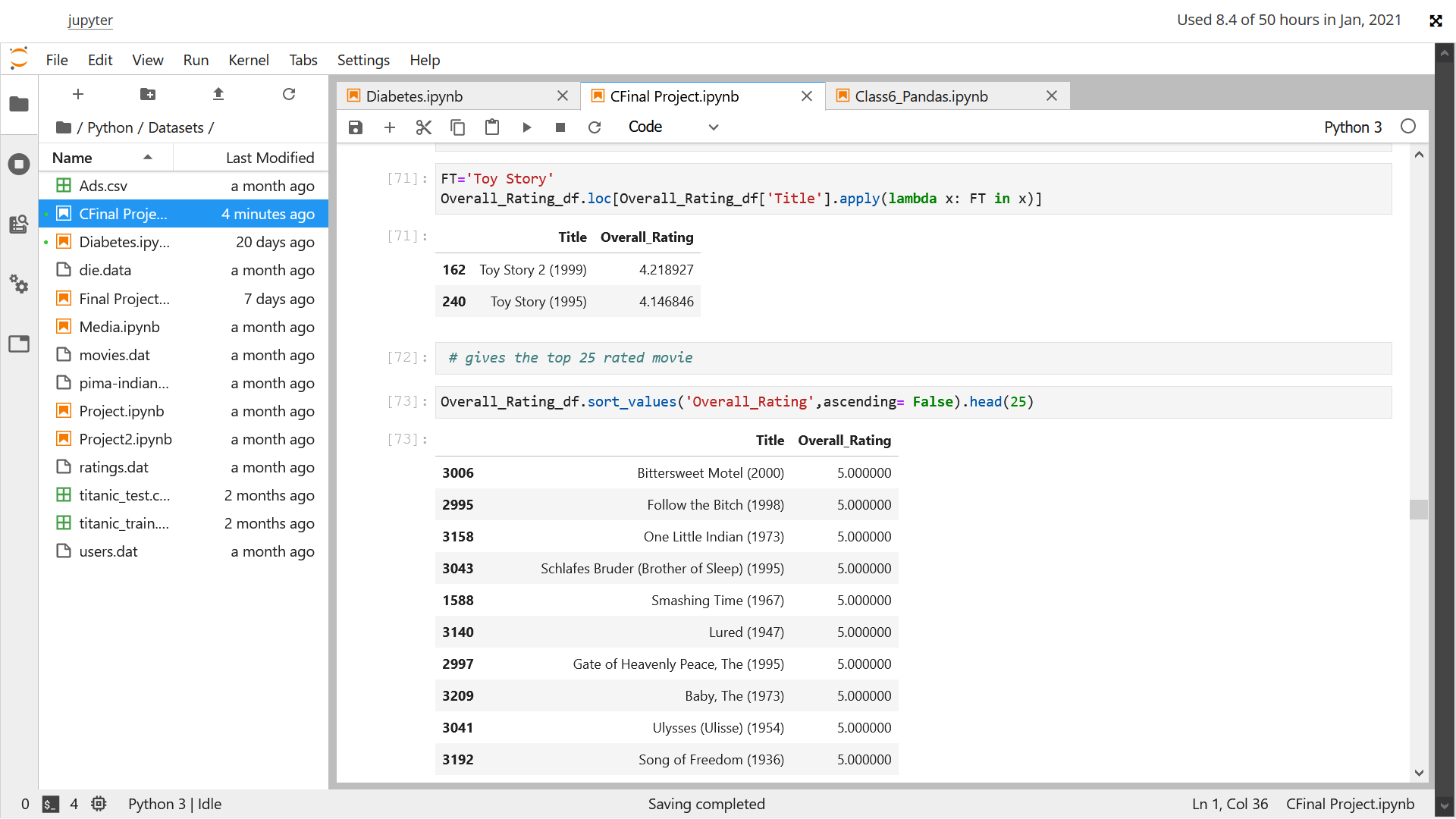
z

hence graph shows thta the maximum users of age category 25-34 , followed by 35-44, 18-24, 45-49,50-55, 56+, below 18

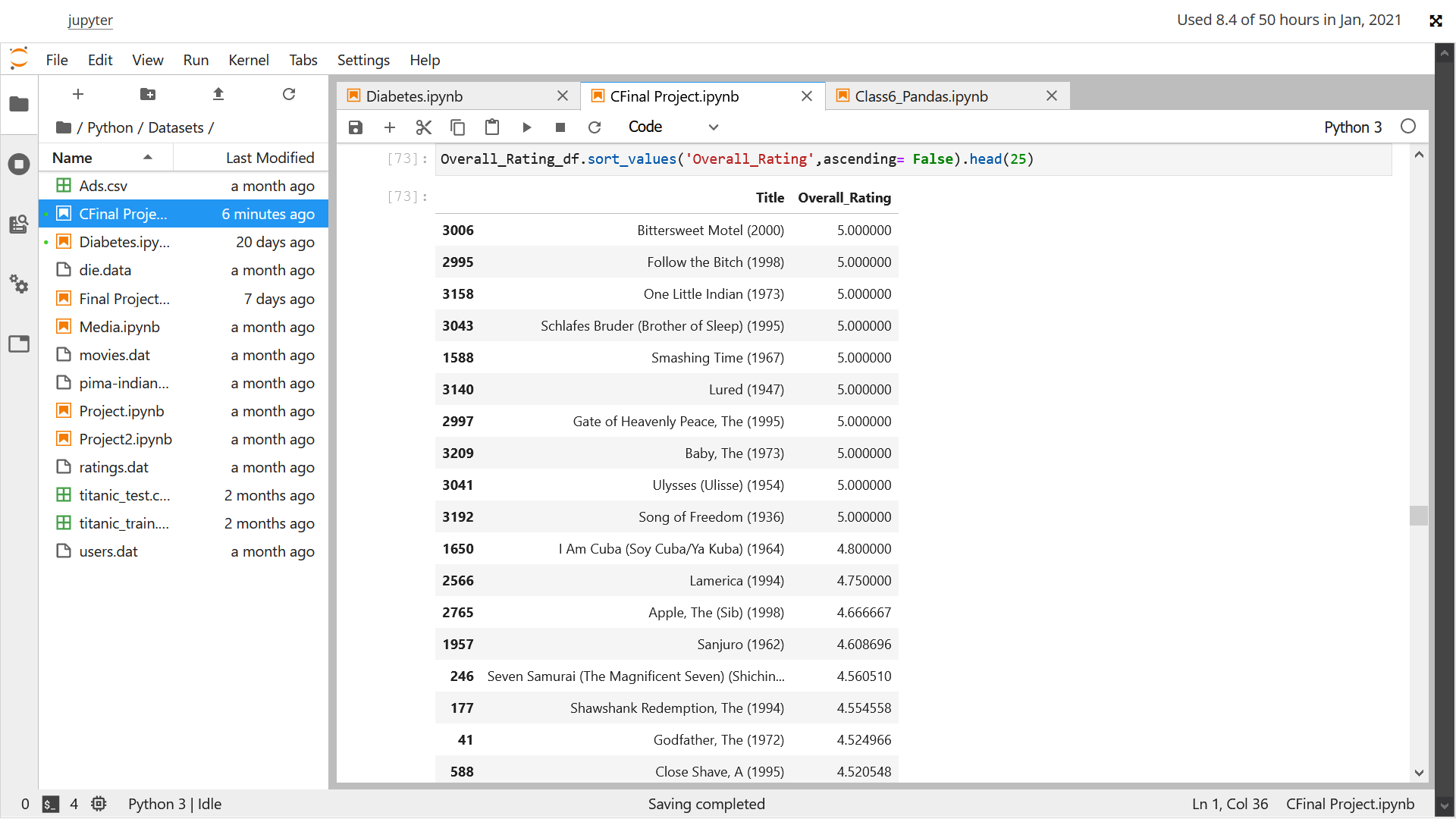




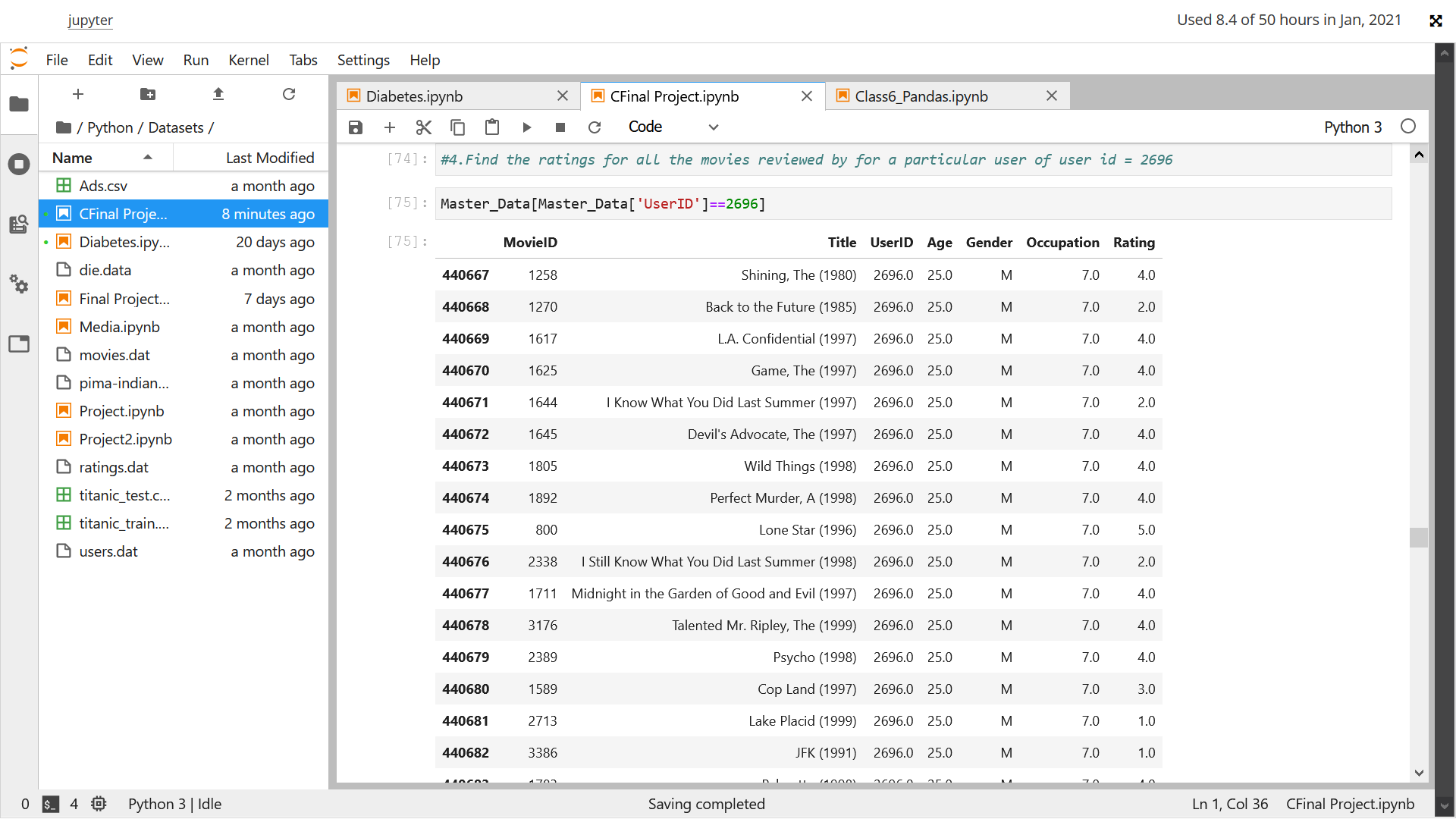




Above gives the User rating for the toy story movie



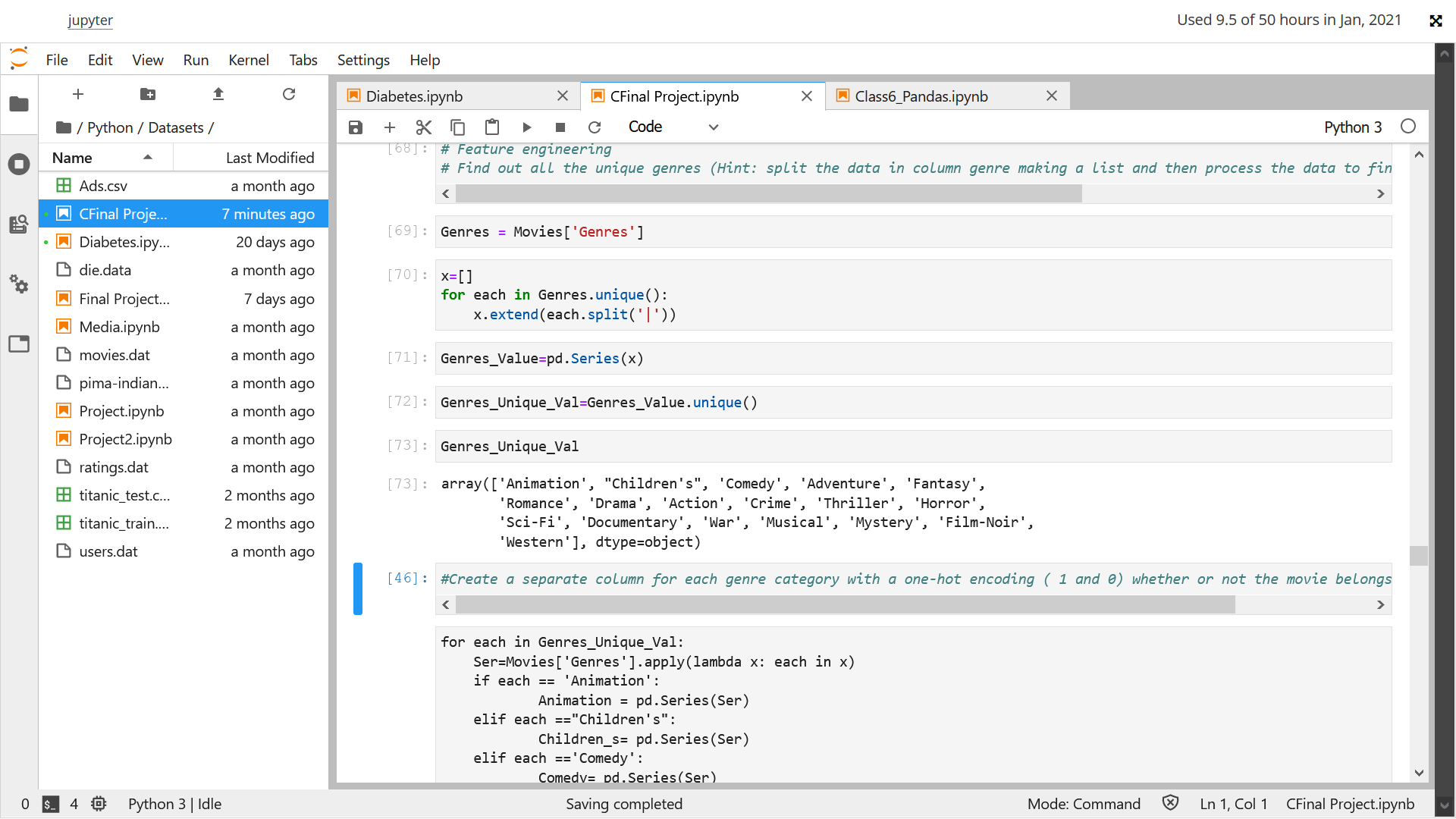
Above represents the rating of top 25 movies as per rating



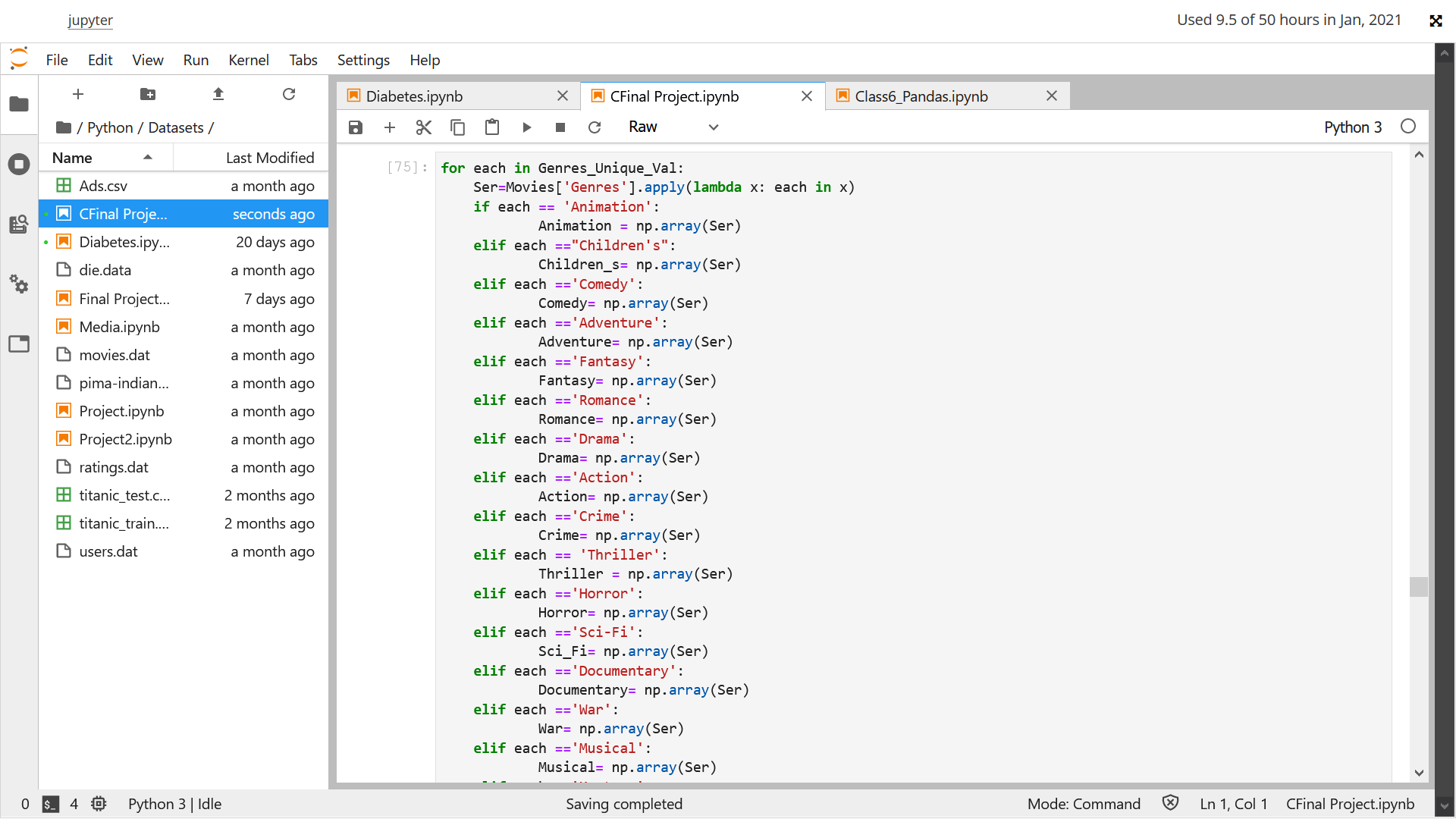
Above represents the ratings given by userid 2696

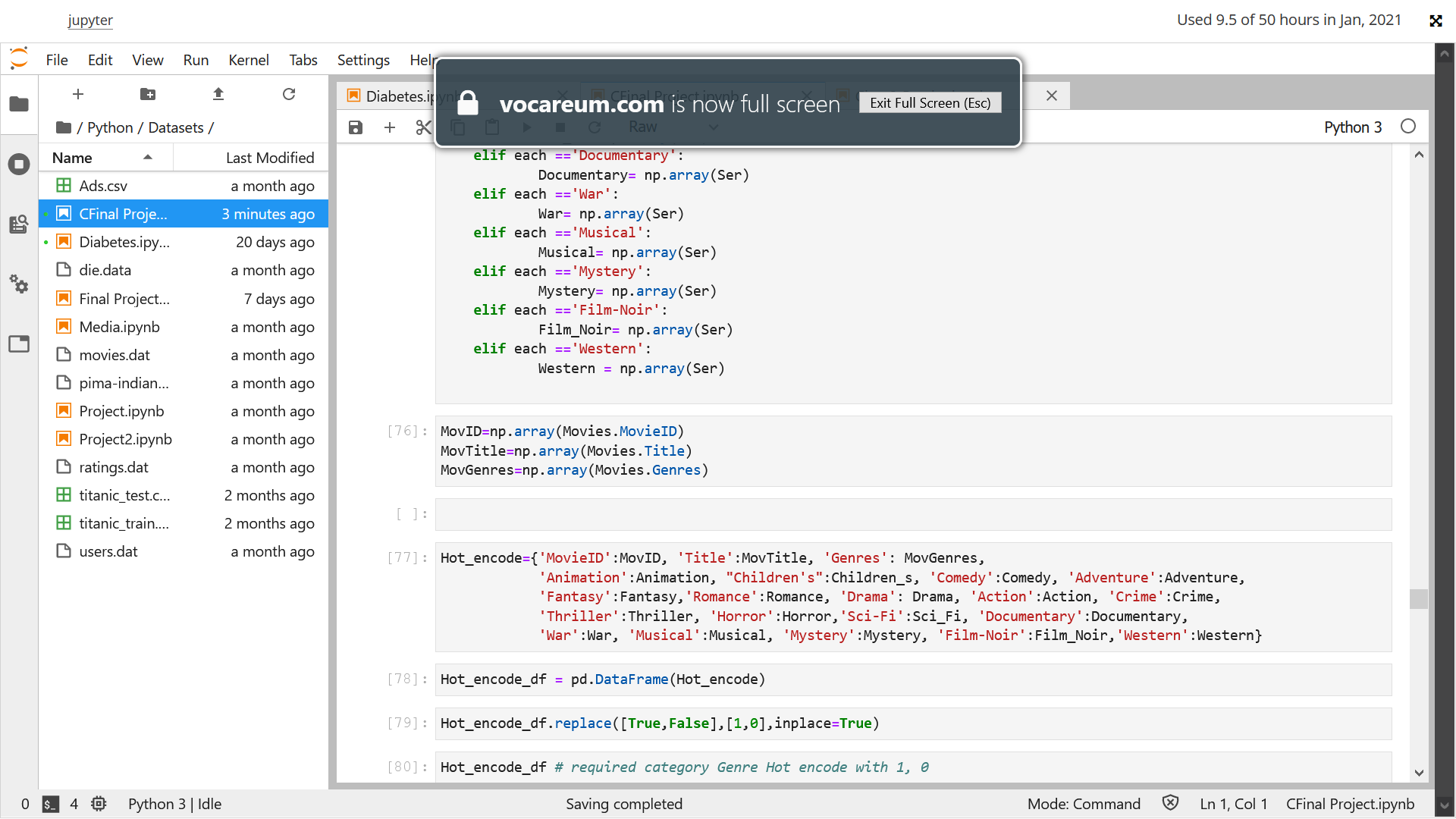
Feature Engineering

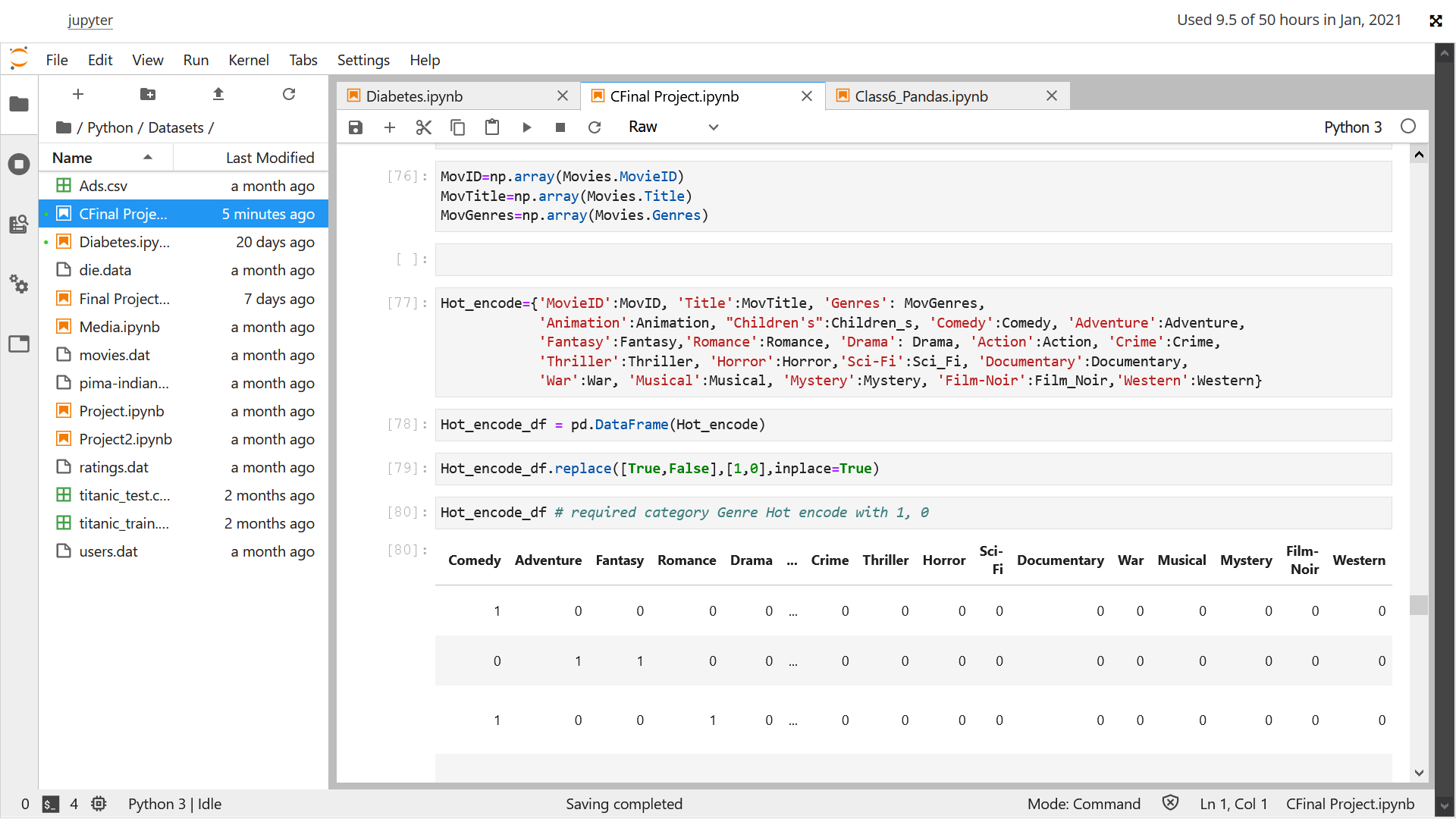
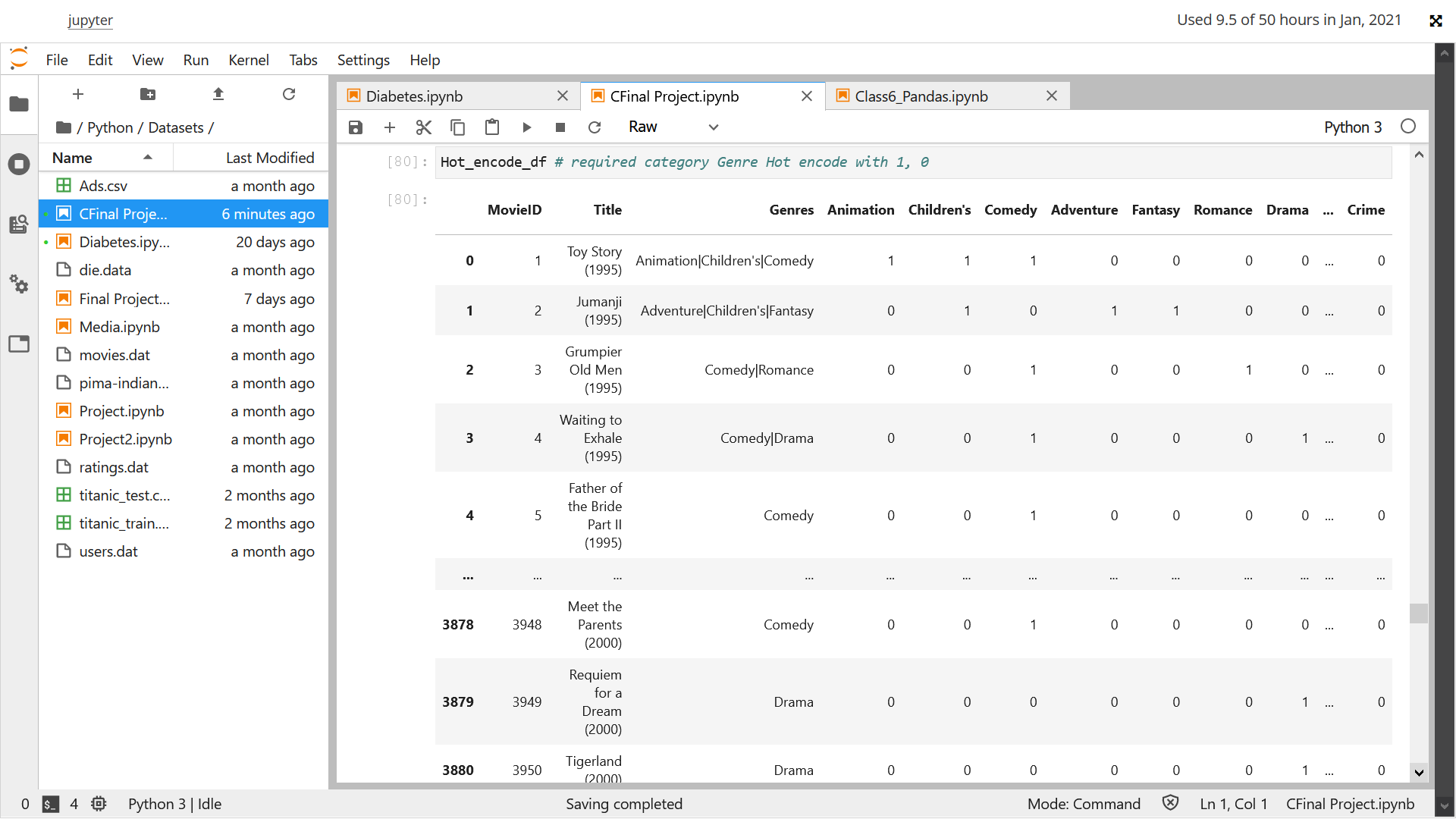
1. Find out all the unique genres (Hint: split the data in column genre making a list and then process the data to find out only the unique categories of genres)
2. Create a separate column for each genre category with a one-hot encoding ( 1 and 0) whether or not the movie belongs to that genre.
3. Determine the features affecting the ratings of any particular movie.
4. Develop an appropriate model to predict the movie ratings



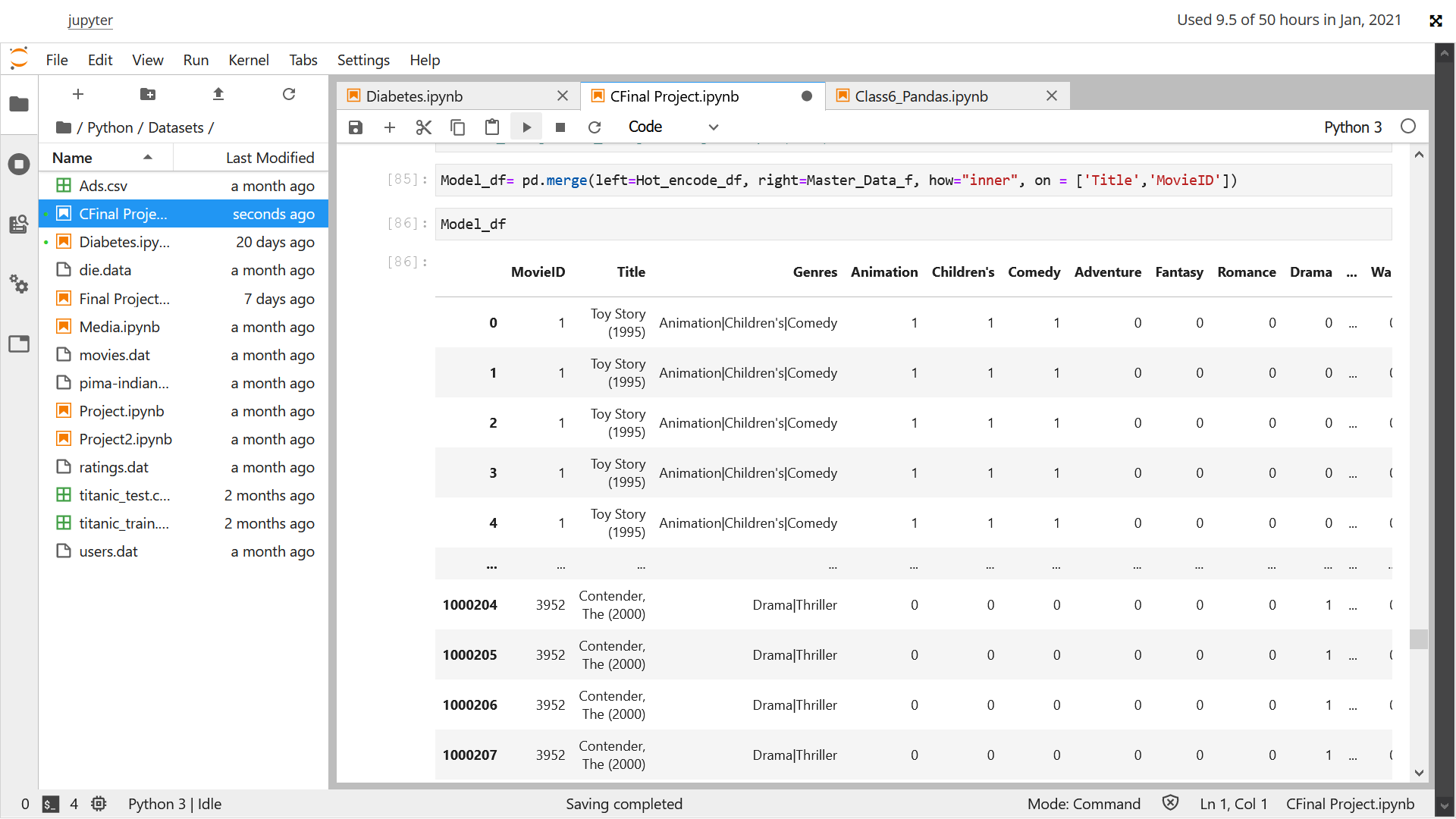
Above represent the unique values of the Genres

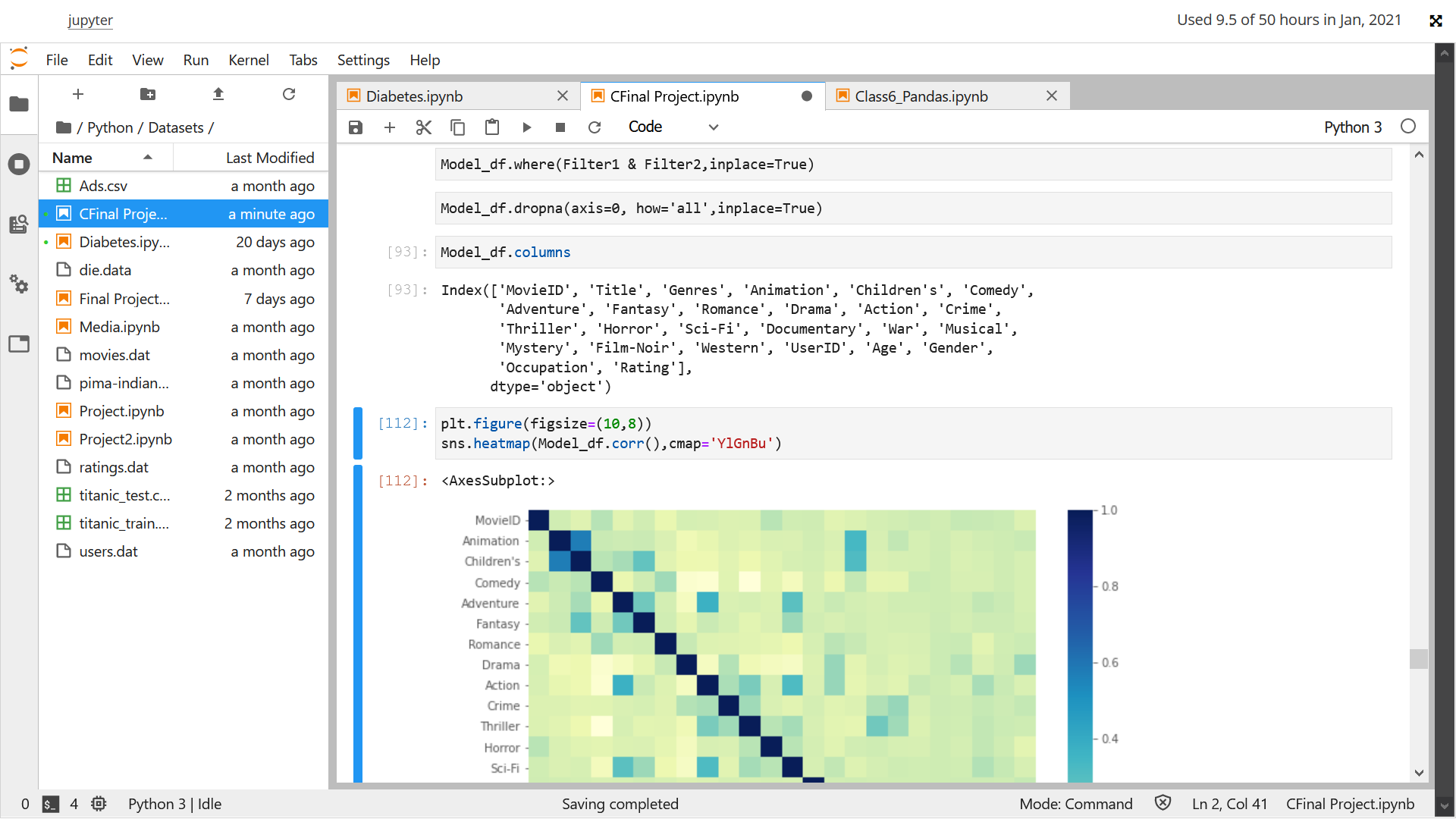




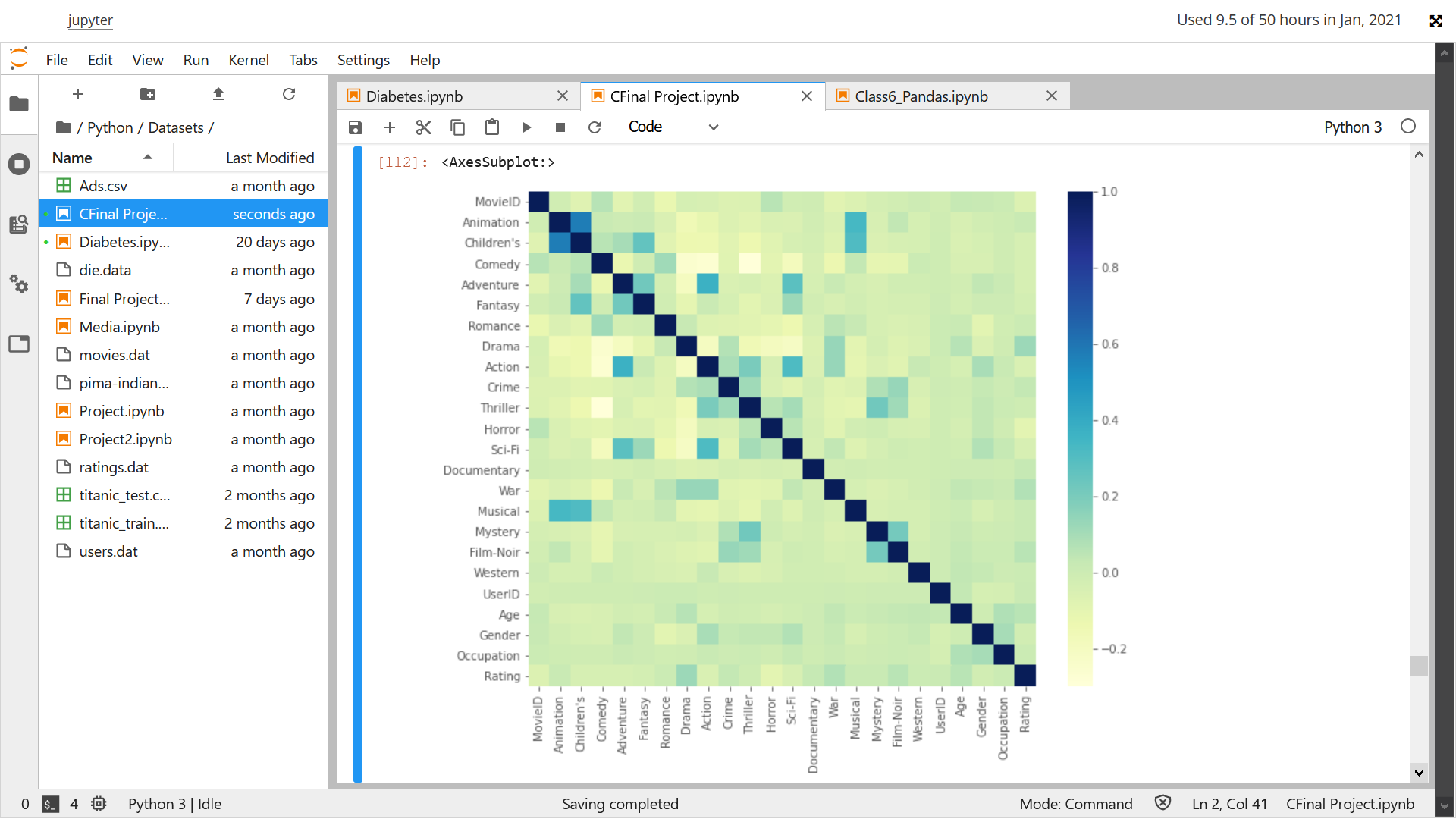
Above is a required Hot encoding

Creating a model and selecting features affecting target

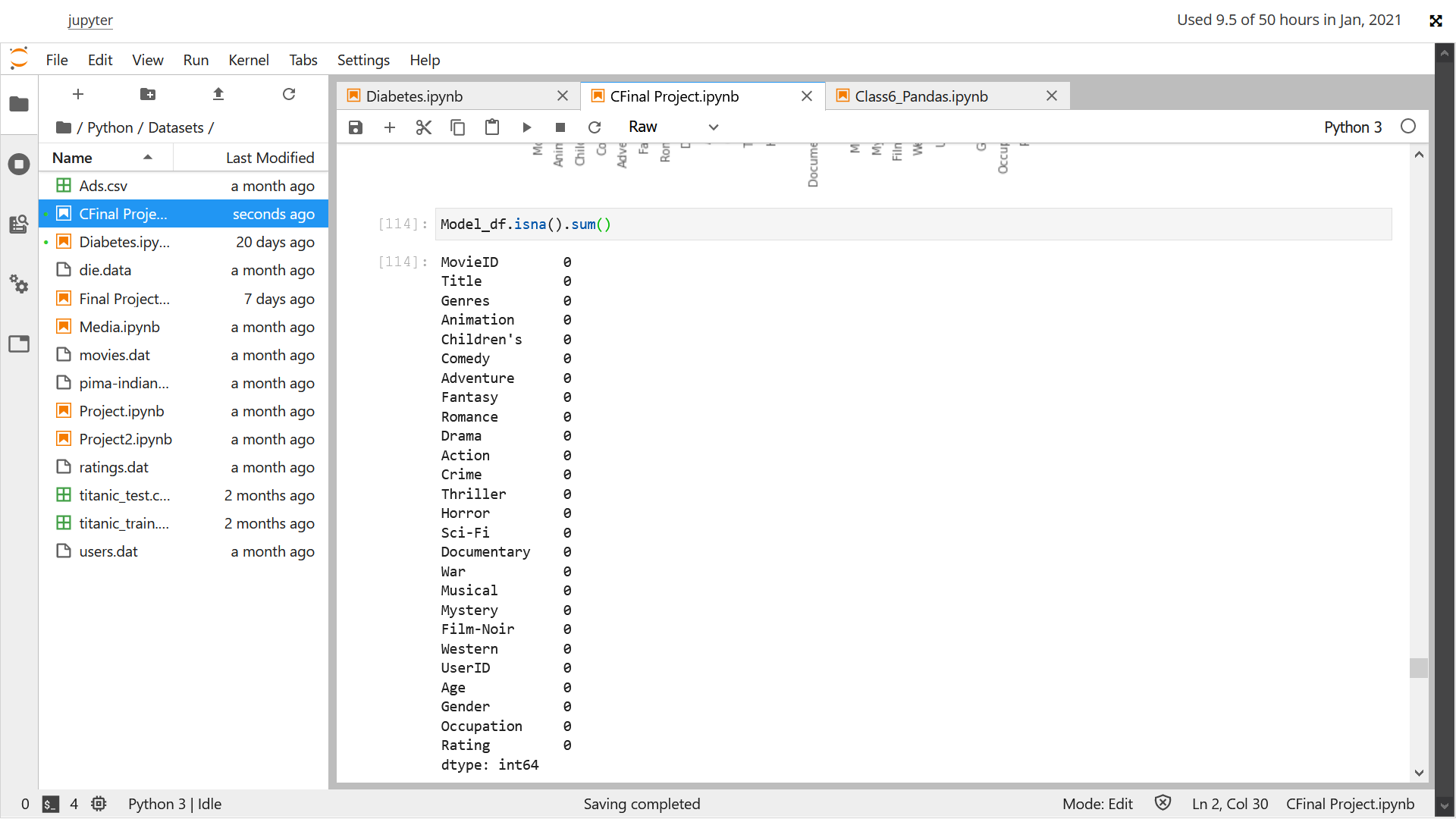




Checking correlation above.

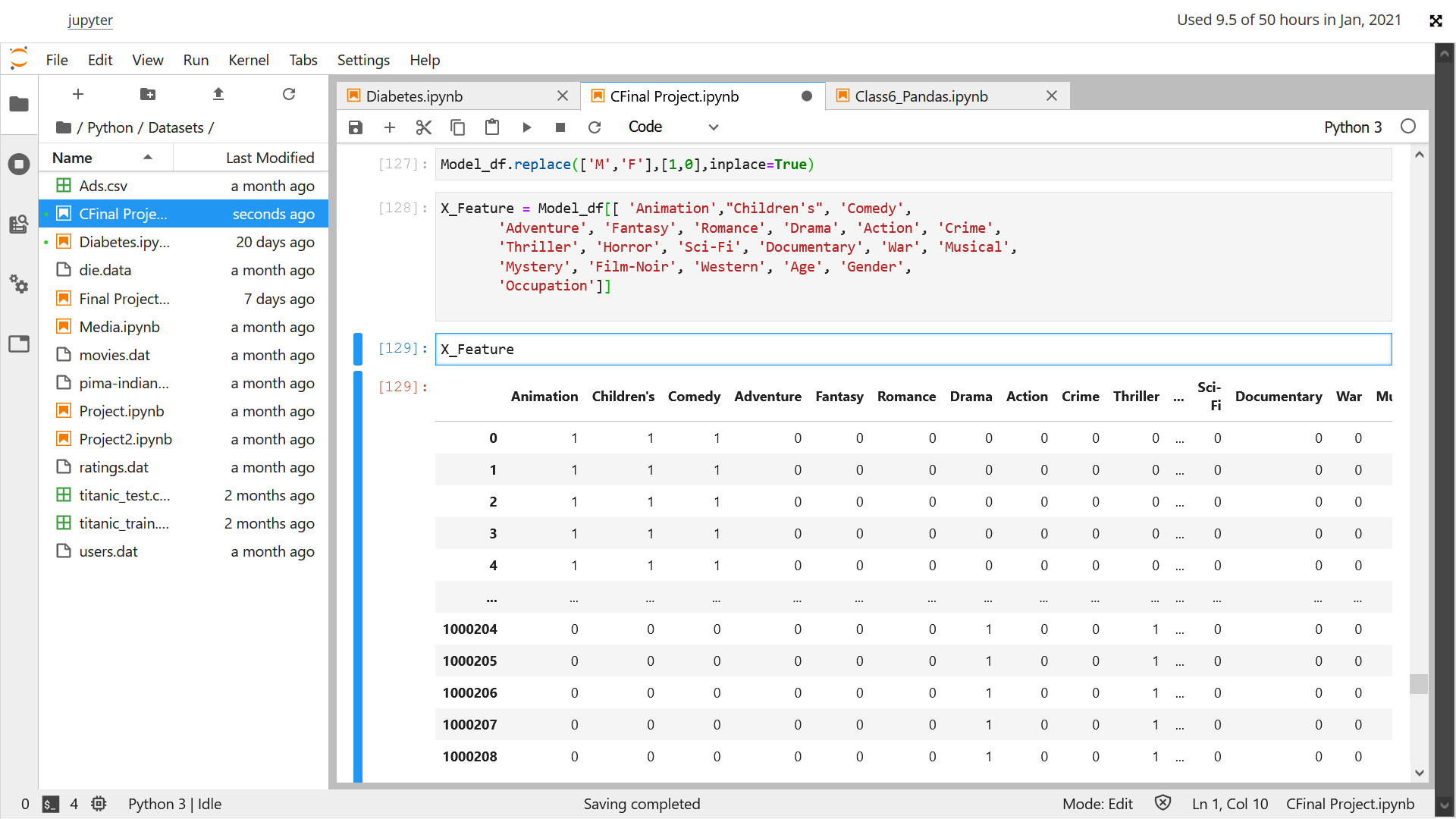


We can see in the heatmap that all the Genres and user attributes like age, Gender and occupation are slightly related with rating , so we can say that they are affecting the rating

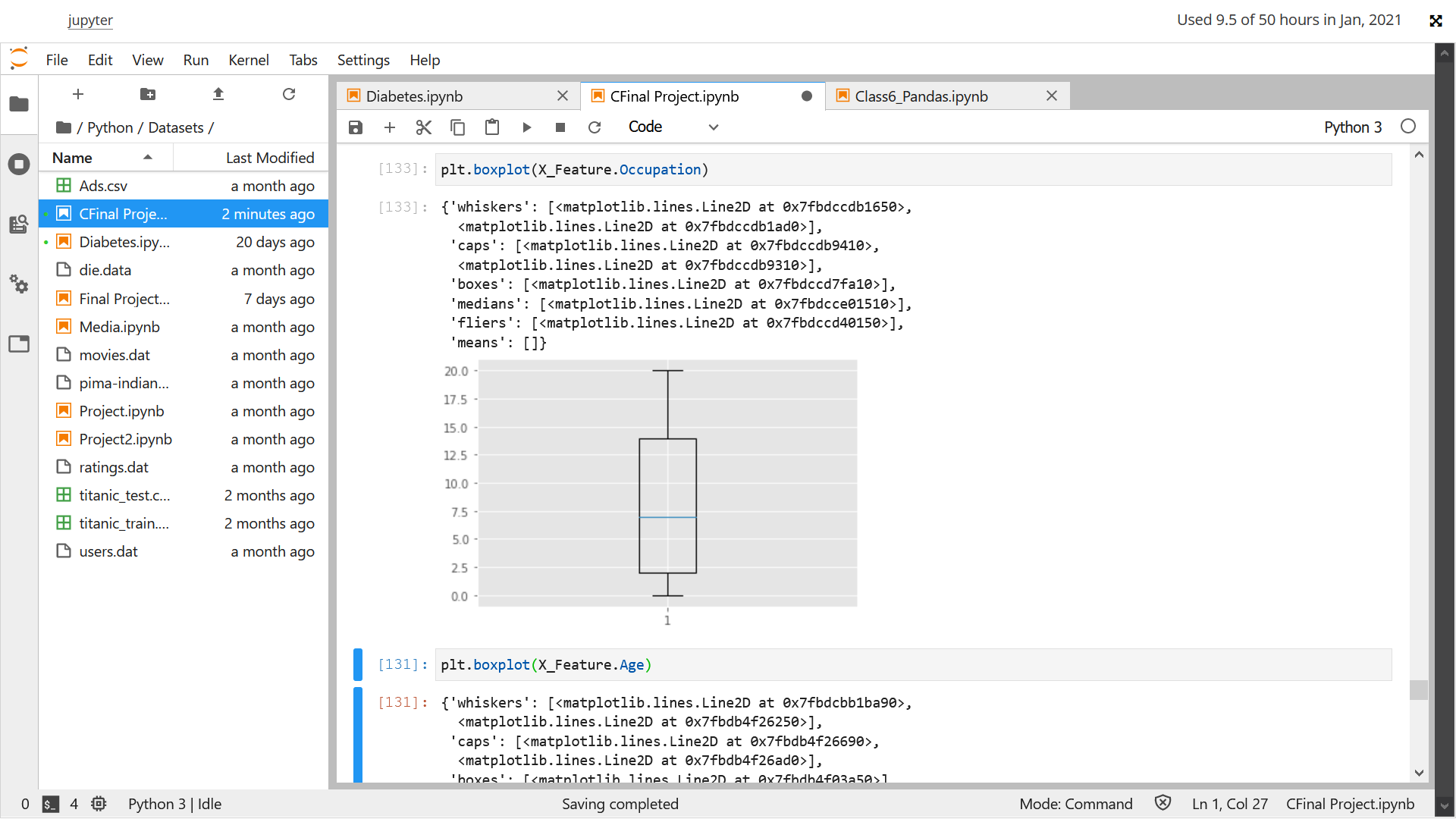


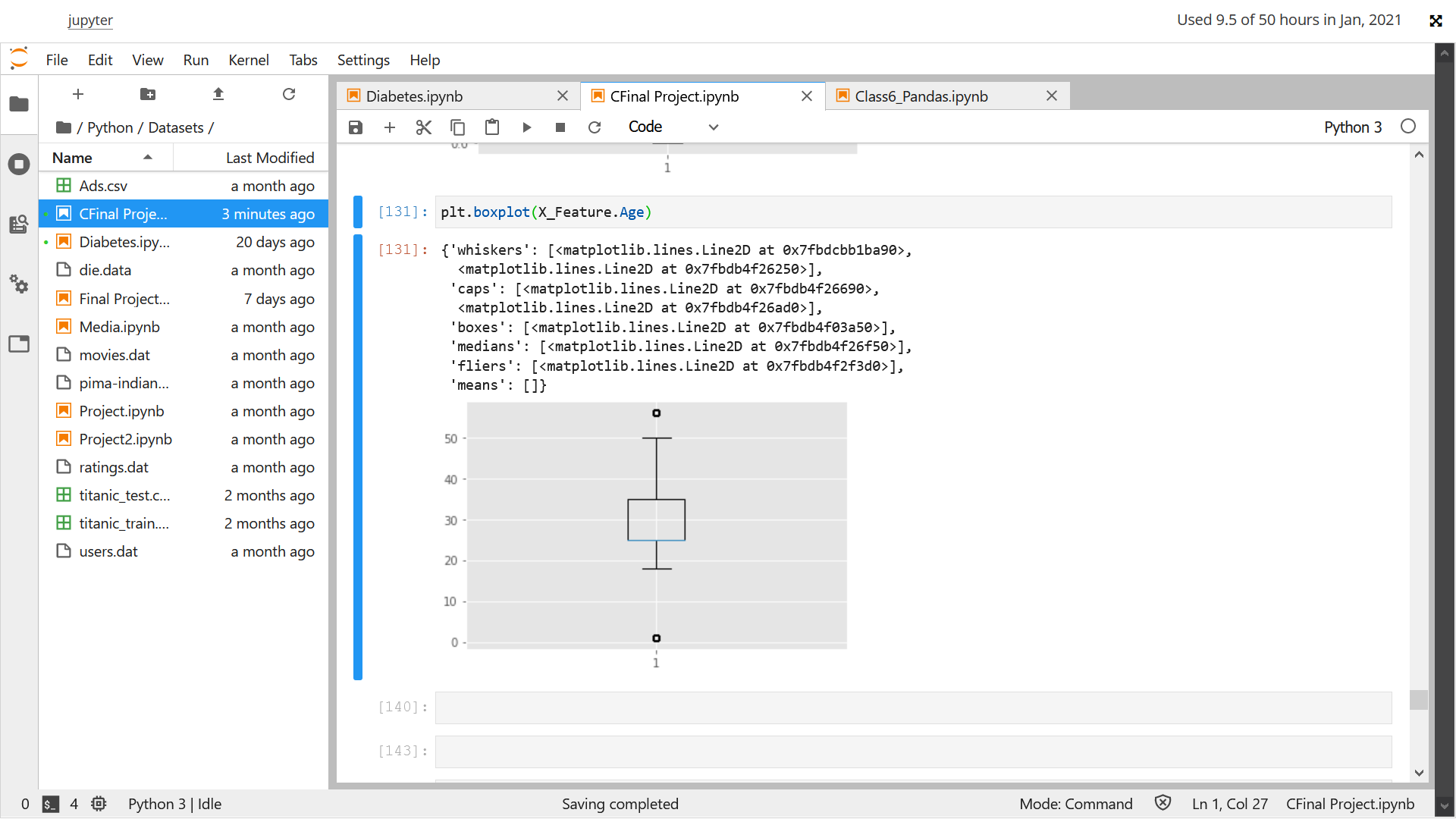
Checking nan values, none present

replacing Categorical value Gender with numerical value and starting to make model

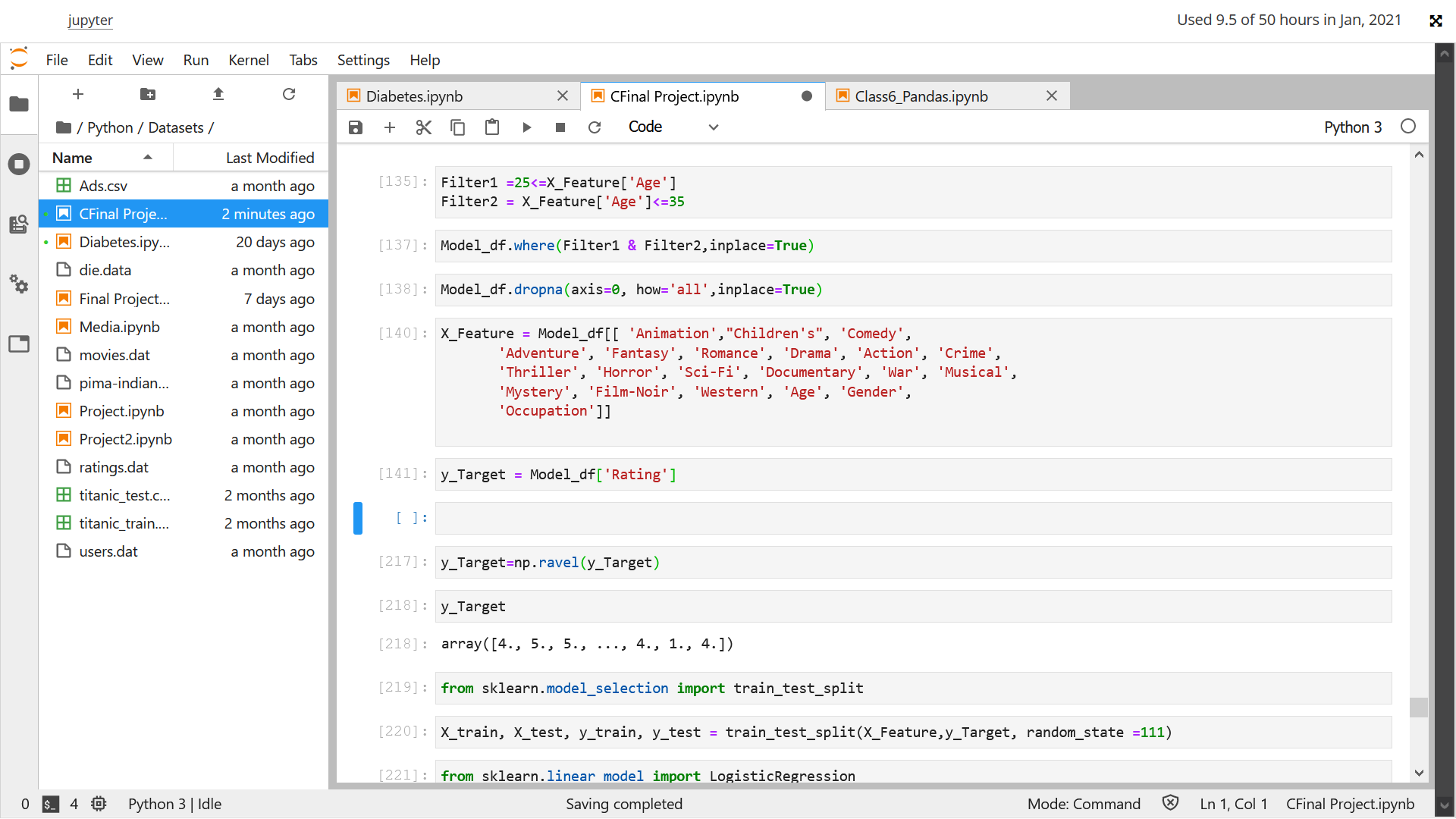


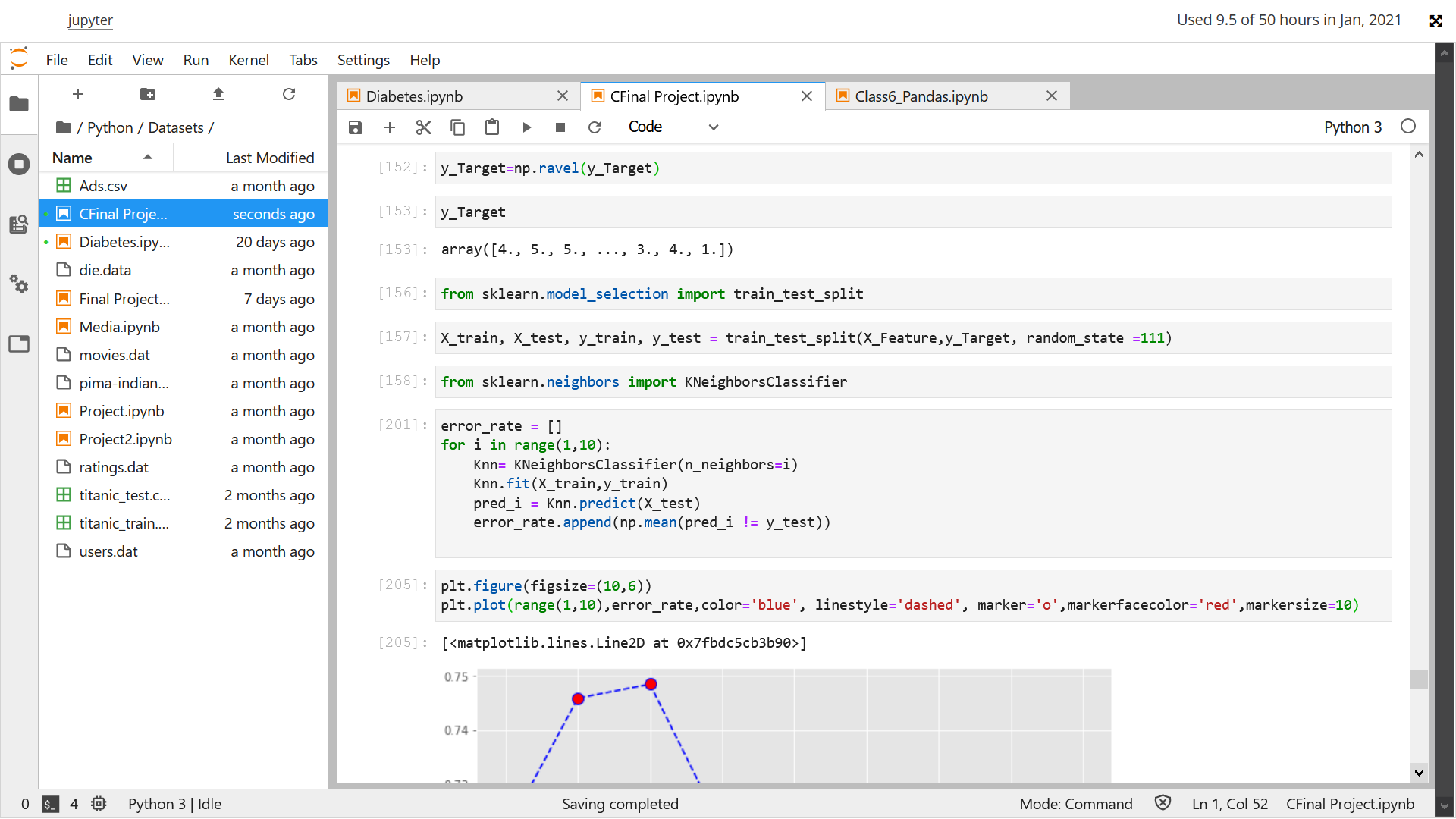
Checking outliers

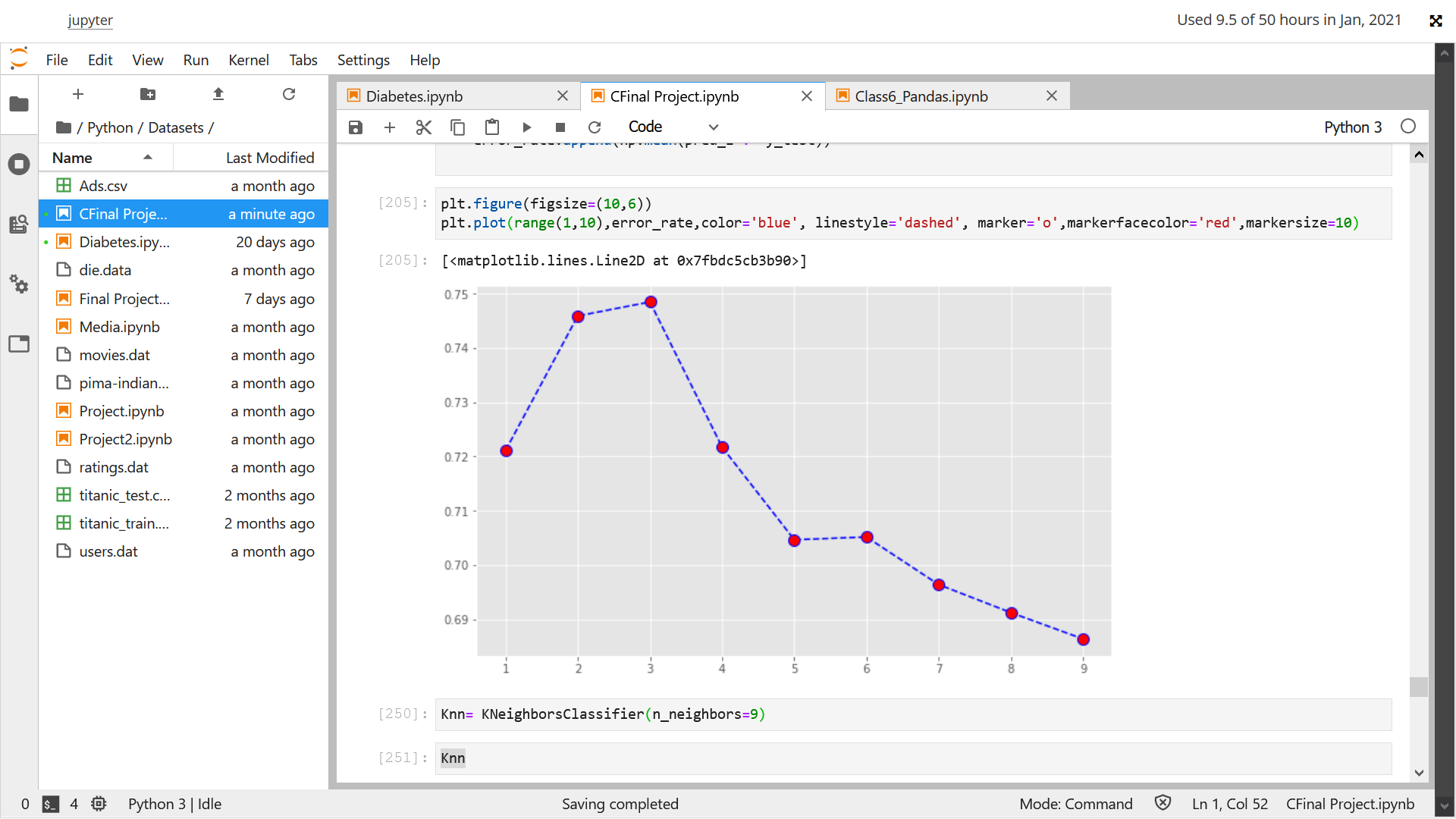




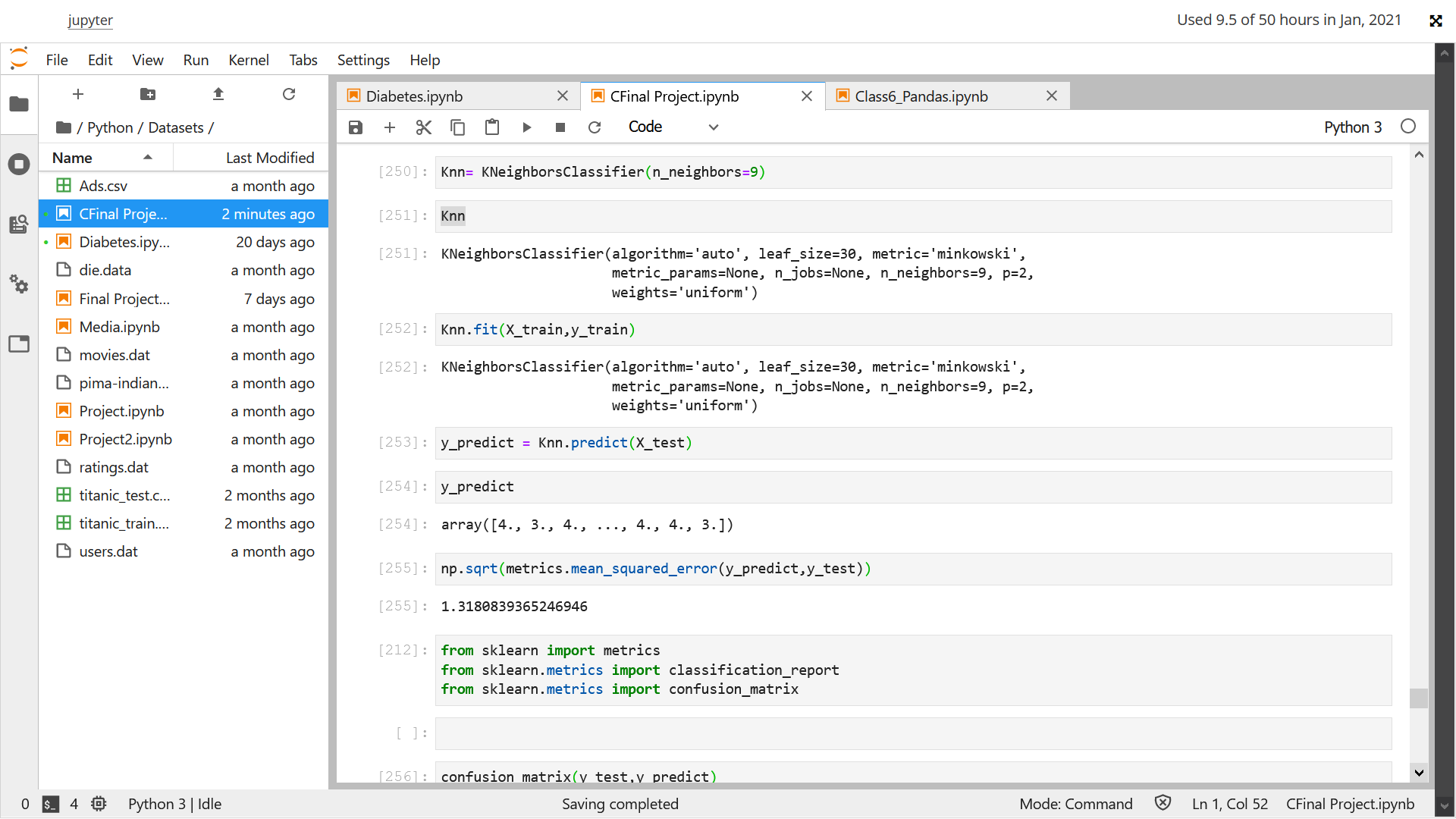
Removing outliers from Model\_Df

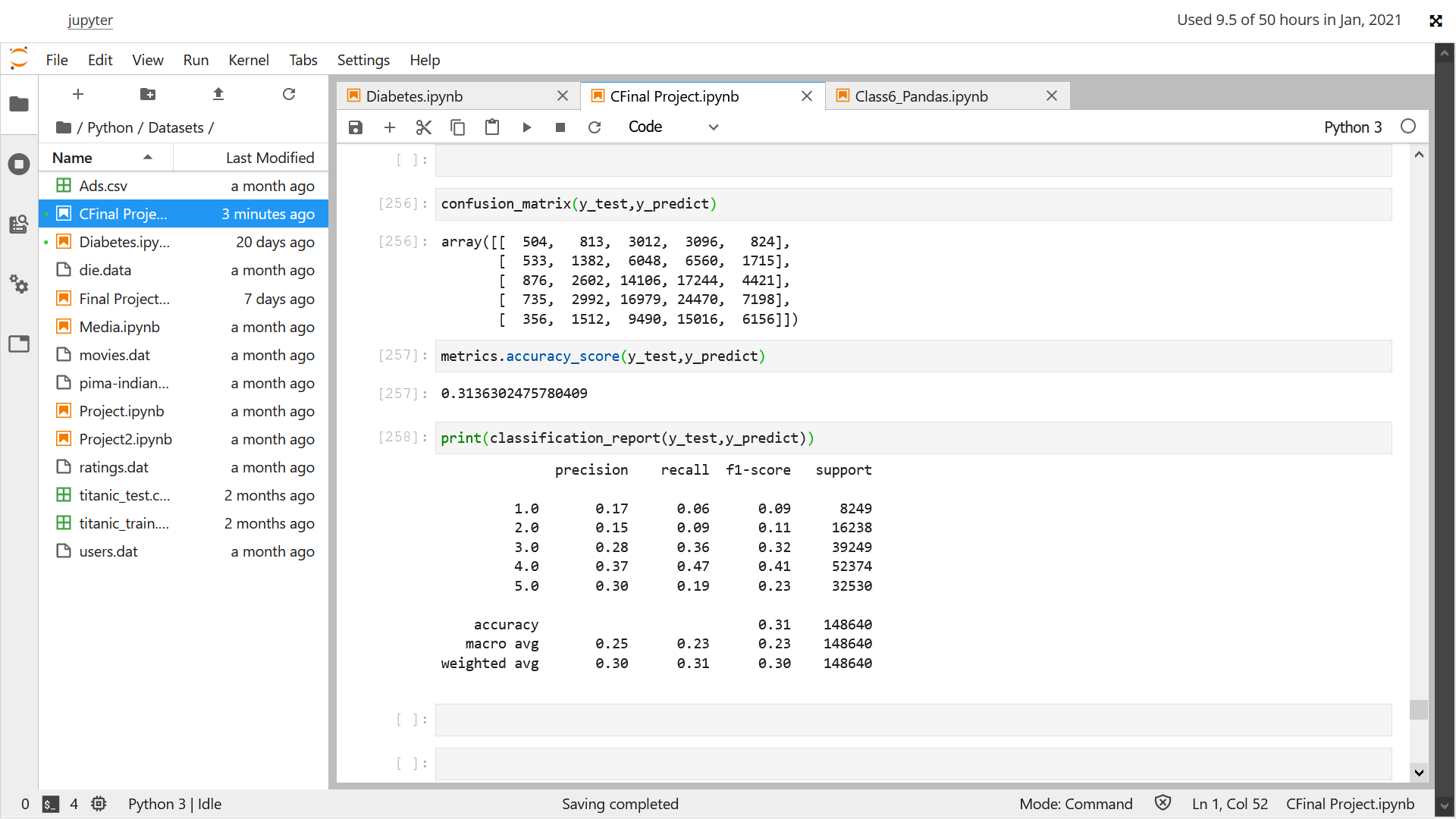






Choosing K-=9 as per the graph above





Result of regression model.