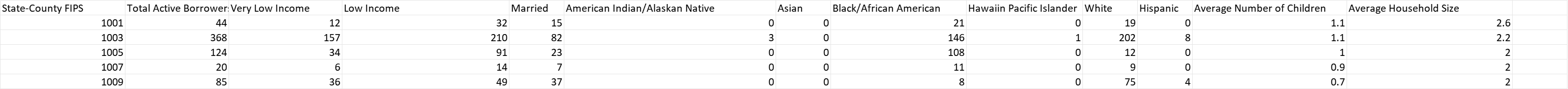
1. **Select a supervised and unsupervised learning of features of data of week2.**

To choose the features of data of week2 as supervised and unsupervised learning, we need to understand their definitions. Both supervised and unsupervised learning aim for feature extraction but in different approaches.

**Supervised Learning:**

Supervised learning is a machine learning technique that is uses previously labeled data to train algorithms for classification. Labeled data indicates that each input (feature vector) has a corresponding output (label or value), which the algorithm should be trained to anticipate.

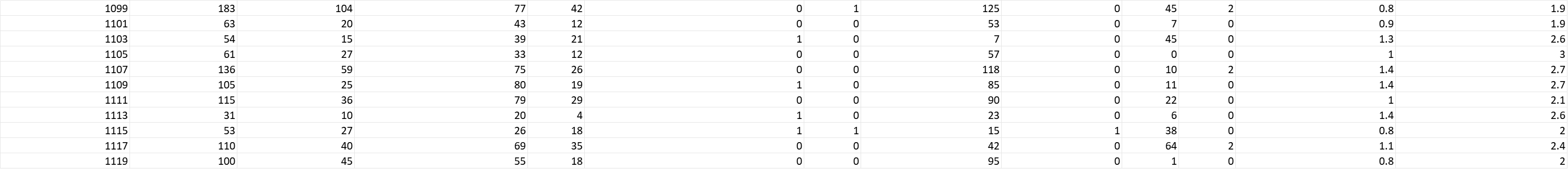
**For the given Data-Week2** State-County FIPS, Total Active Borrowers, Very Low Income, Low Income, Married, American Indian/Alaskan Native, Asian, Black/African American, Hawaiin Pacific Islander, White, Hispanic, Average Number of Children, Average Household size are the labels of the data week2 used for supervised learning algorithms. As this is labeled data this features are used to train the model using supervised learning approaches like Random forest, Naive Bayes, Support vector machines, Logistic regression etc.



**Unsupervised Learning:**

Unsupervised learning is a machine learning technique that is uses neither classified nor labeled data to train algorithm to act on the data without supervision.

**For the given Data-Week2** the below given unlabeled data that can be considered for unsupervised learning.



1. **How to split data to train and test data in Python?**

To split data to train and test data in python we generally use train\_test\_split from sklearn.model\_selection library. Below is illustration for splitting data to train and test.

I considered Data-week2 as feature and Low Income as target variable.

import pandas as pd

from sklearn.model\_selection import train\_test\_split

data=pd.read\_csv("Data-Week2.csv")

labels=data["Low Income"]

#split ratio defines the ratio of data that is considered for testing(Lets consider 70% for training, 30% for testing)

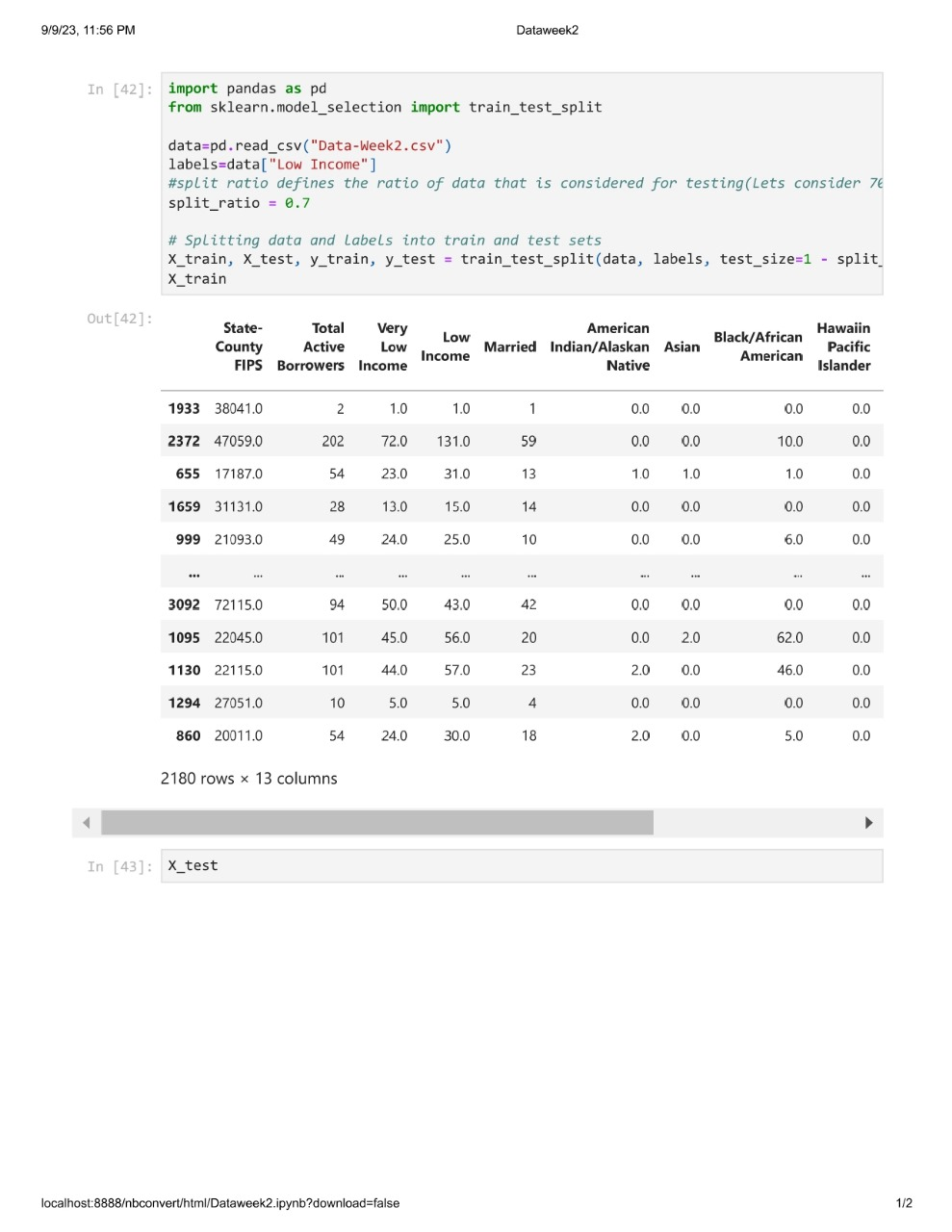
split\_ratio = 0.7

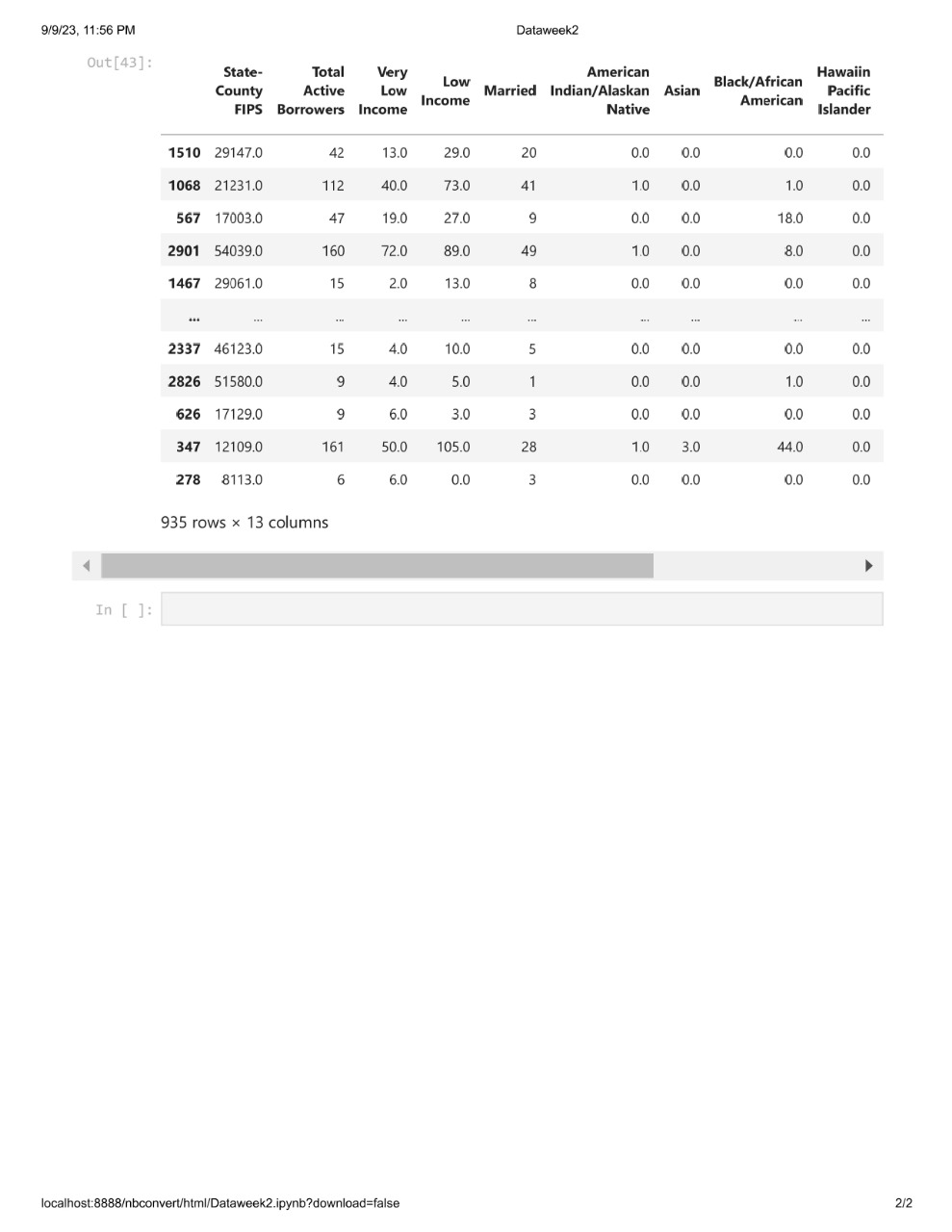
# Splitting data and labels into train and test sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(data, labels, test\_size=1 - split\_ratio, random\_state=42)

X\_train

X\_test





In the above python code X\_train and X\_test represent the training data which is 70% of the total data and remaining 30% of the data for testing.

In the above case we can notice 2180 rows (which is 70%)for training and 935 rows for testing(which is 30%).