ASSIGNMENT-1

Topic: CLUSTERING

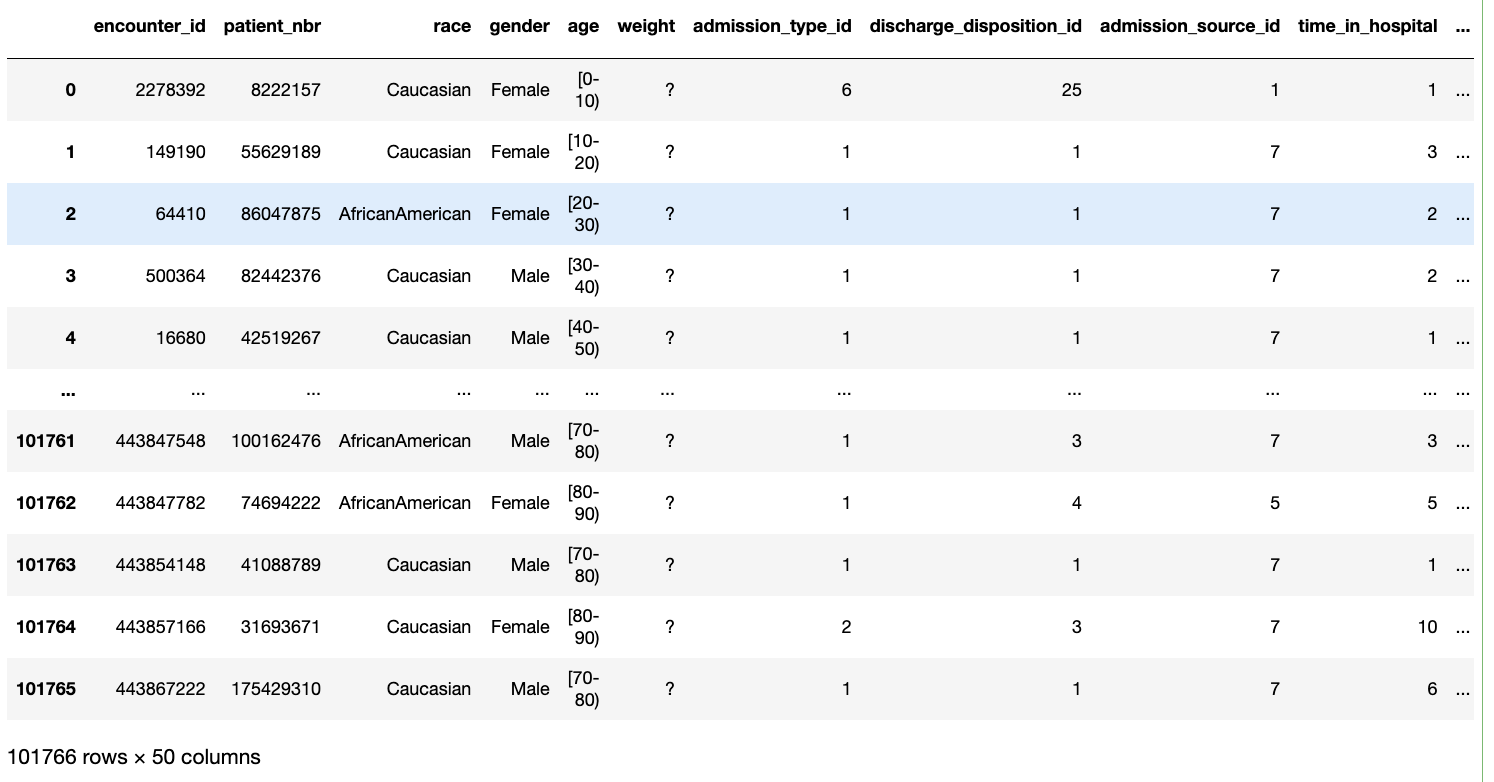
Name: Dixit Mj

USN No: 21BSR18041

Course: 4th Sem DSA

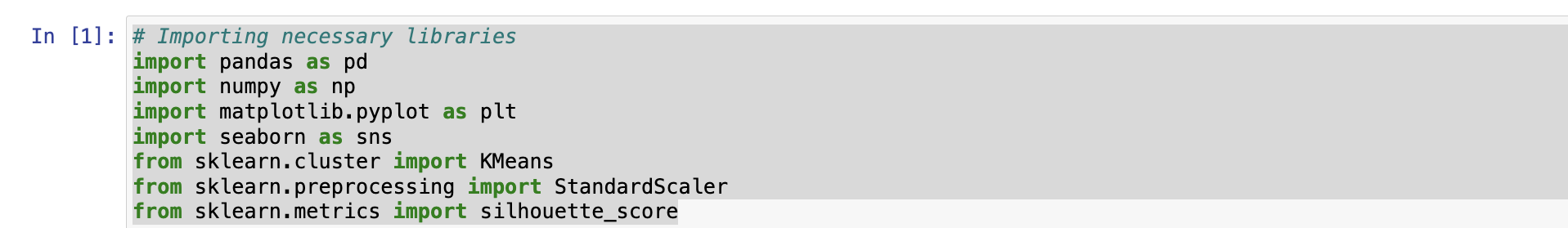
Jain university Jc Road

SAMPLE DATA: diabetic\_data.csv

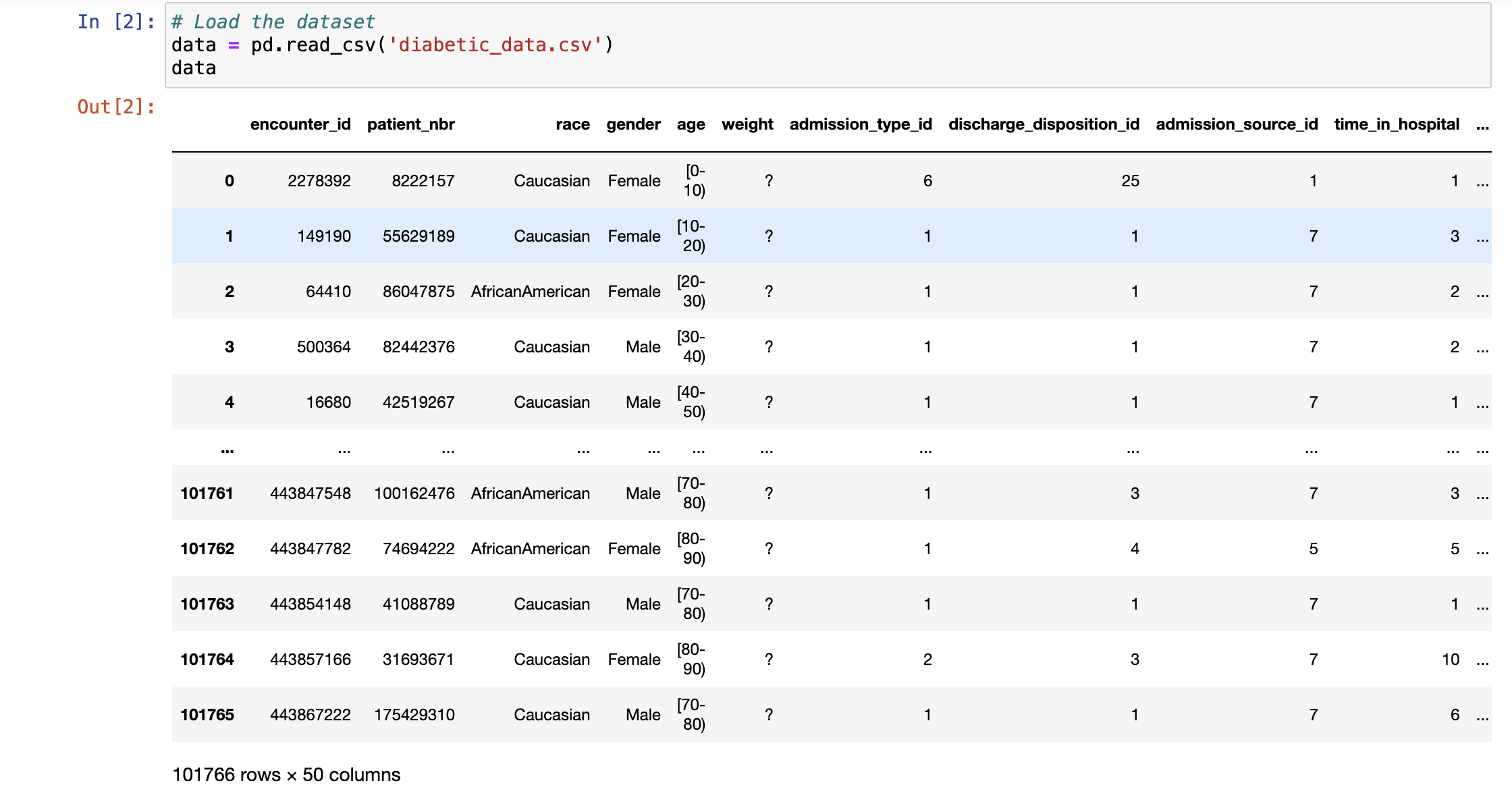


TO EXPLAIN CODE AND ITS FUNCTIONALITY

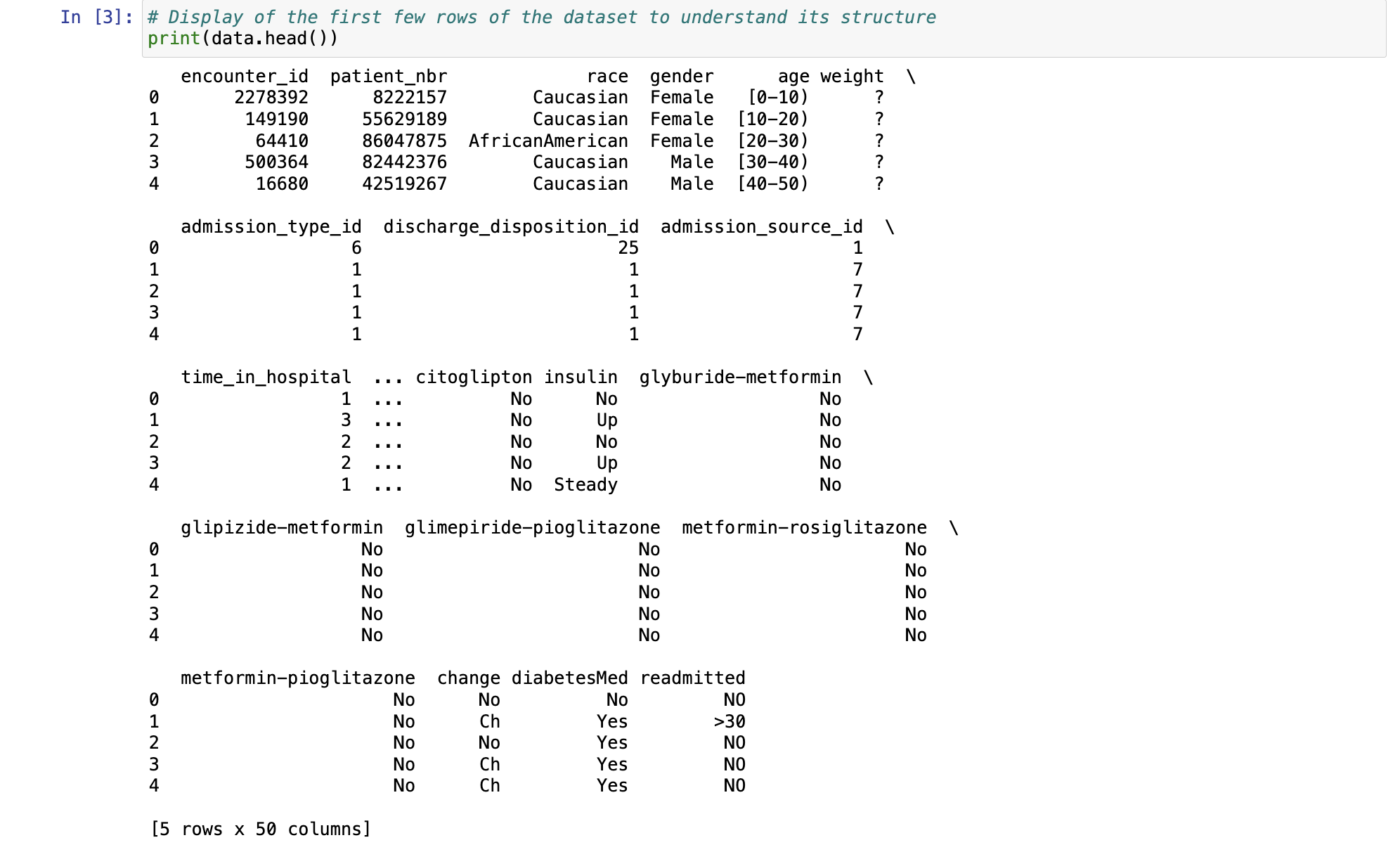
1. Importing the Necessary Libraries (code)

Explanation: Here we are importing the necessary libraries.

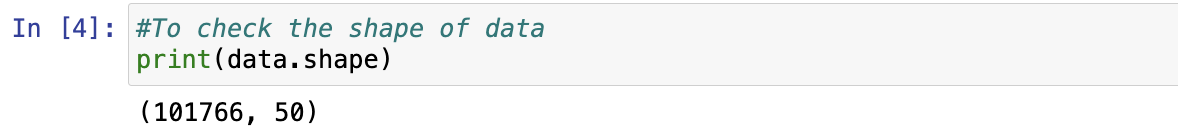
1. Loading the Data set (code)

Explanation: Here we are loading the dataset from the CSV file.

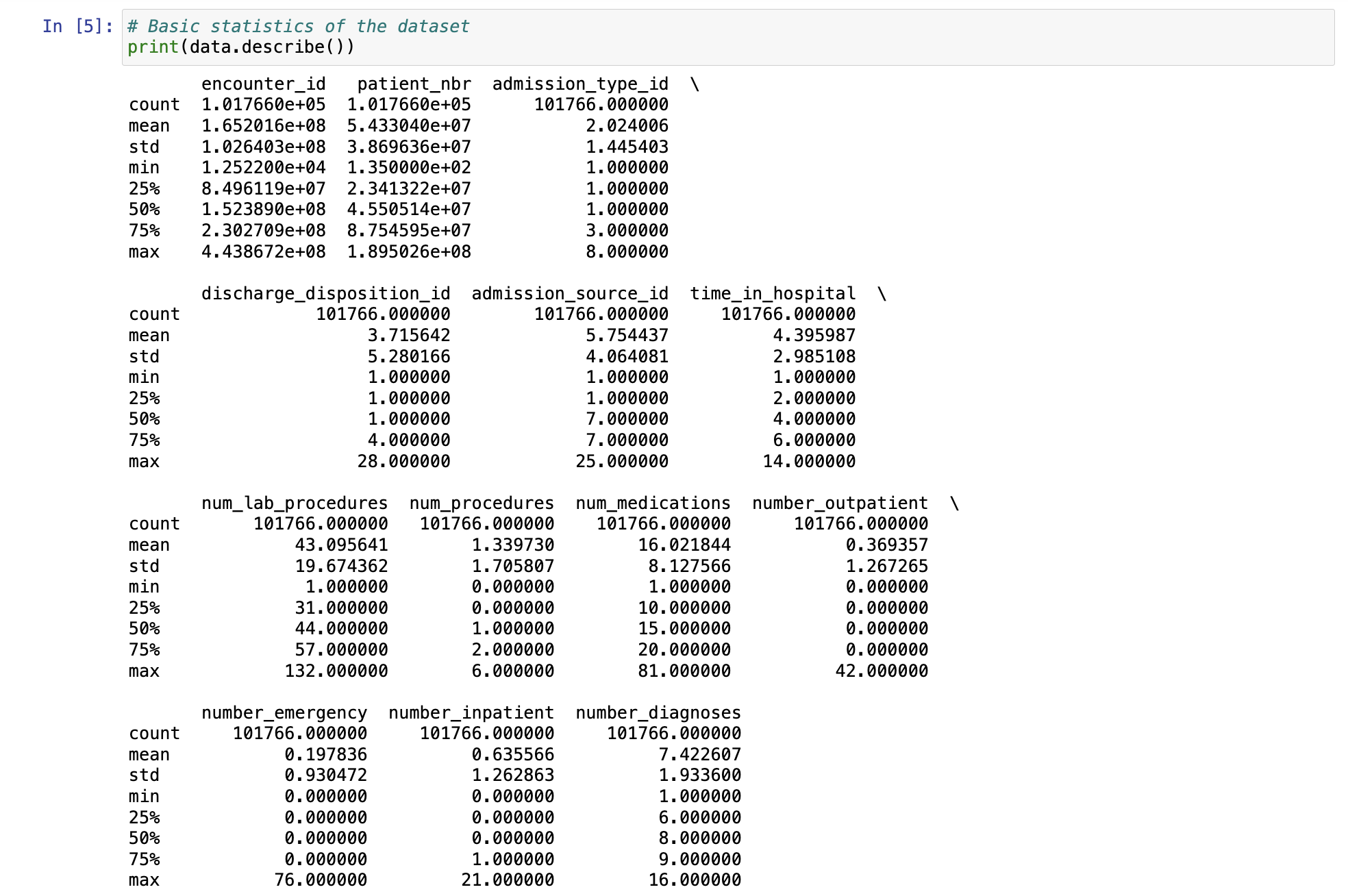
1. Displaying of the first few rows of the dataset (code)

Explanation: Here we are displaying the first few rows of the dataset.

1. Checking the Shape of data (code)

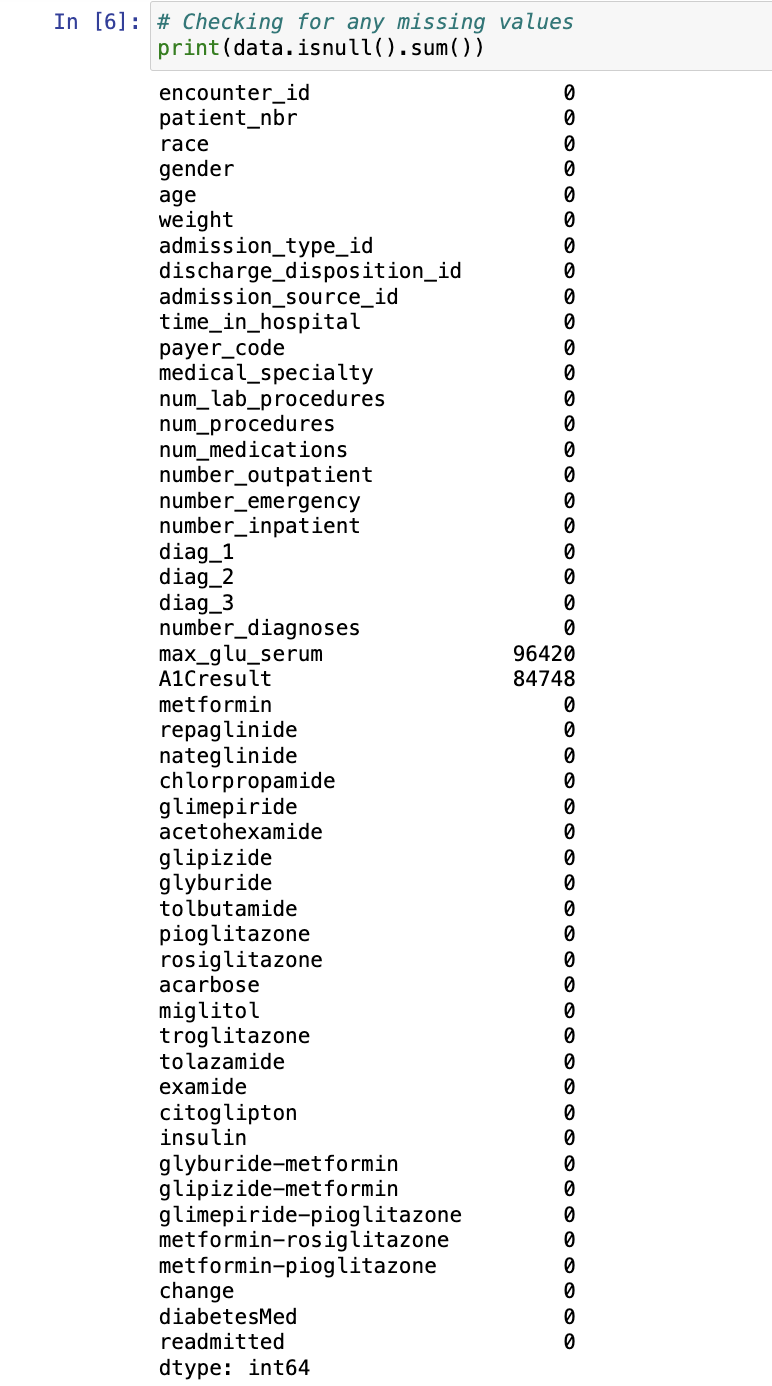
Explanation: Here we can see the number of rows and number of columns in the data

1. Describing the basic statistics (code)



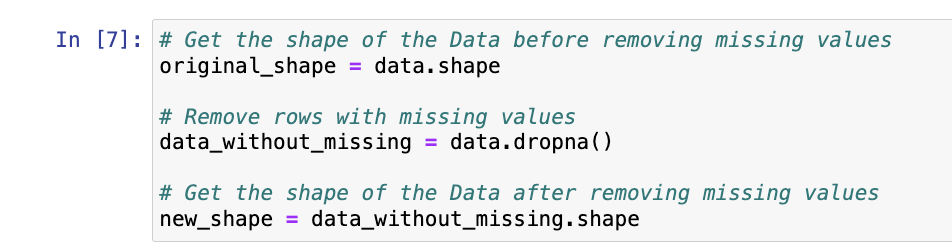
Explanation: Here the basic statistics gives us a quick overview of the distribution and range values of each numerical column in our dataset. By this we can understand the central tendency, variability and distribution of the data.

1. Checking of missing values (code)



Explanation: Here we checking the missing values in the dataset

1. Dealing with the Missing values (code)

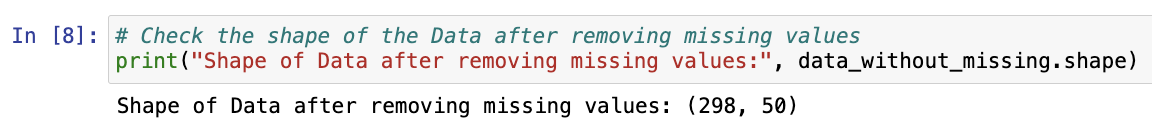


Explanation: Here firstly we are taking the shape of data before removing the missing values and naming it has original\_shape, after that we are removing the

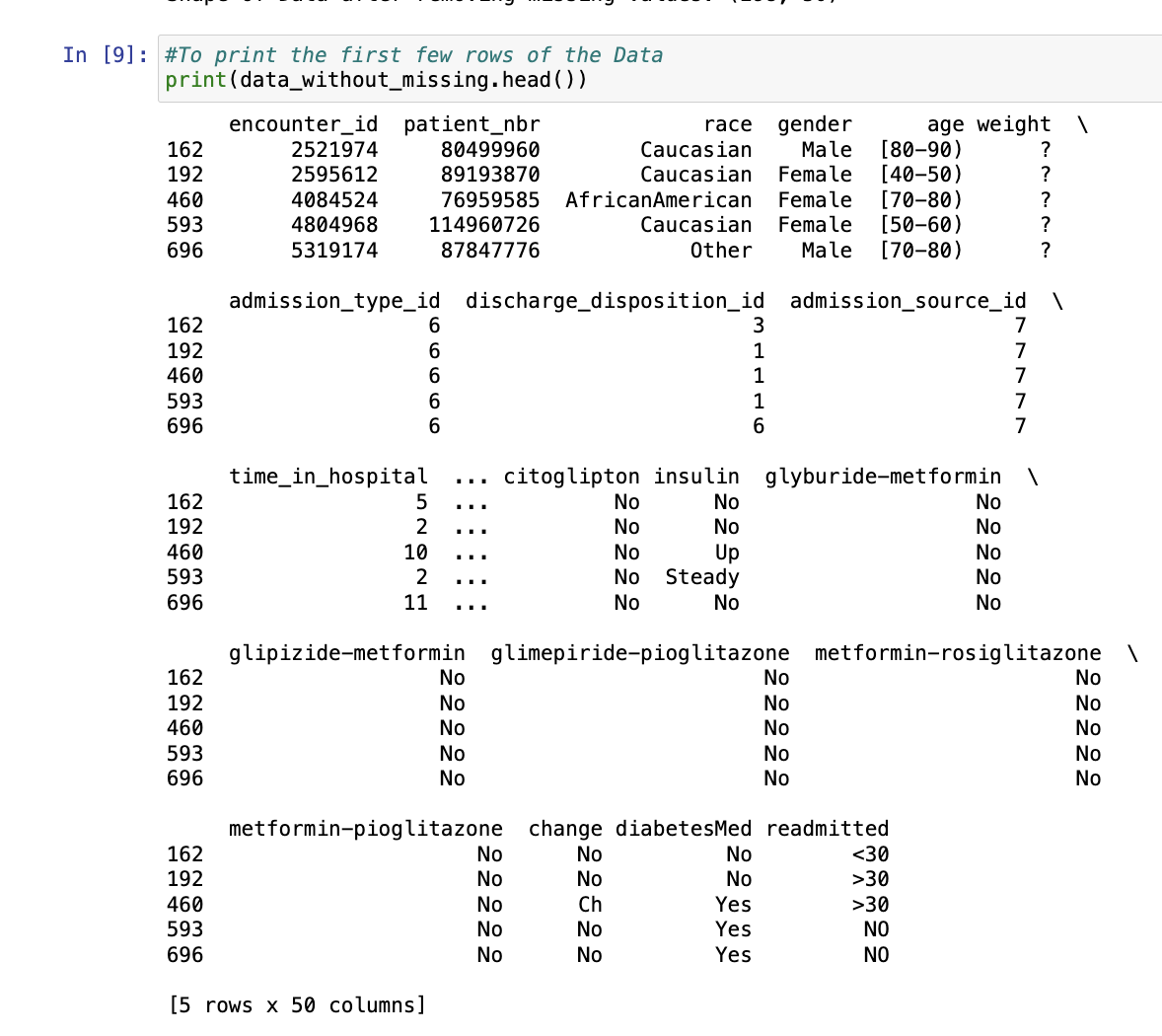
missing values and naming the data set as data\_without\_missing.

**Here we removed the missing values because the data set had a greater number of missing values as if we were able to understand that patients which were coming to the hospital we not diagnosed with diabetes.**

1. Checking the shape of the data after removing the missing values (code)

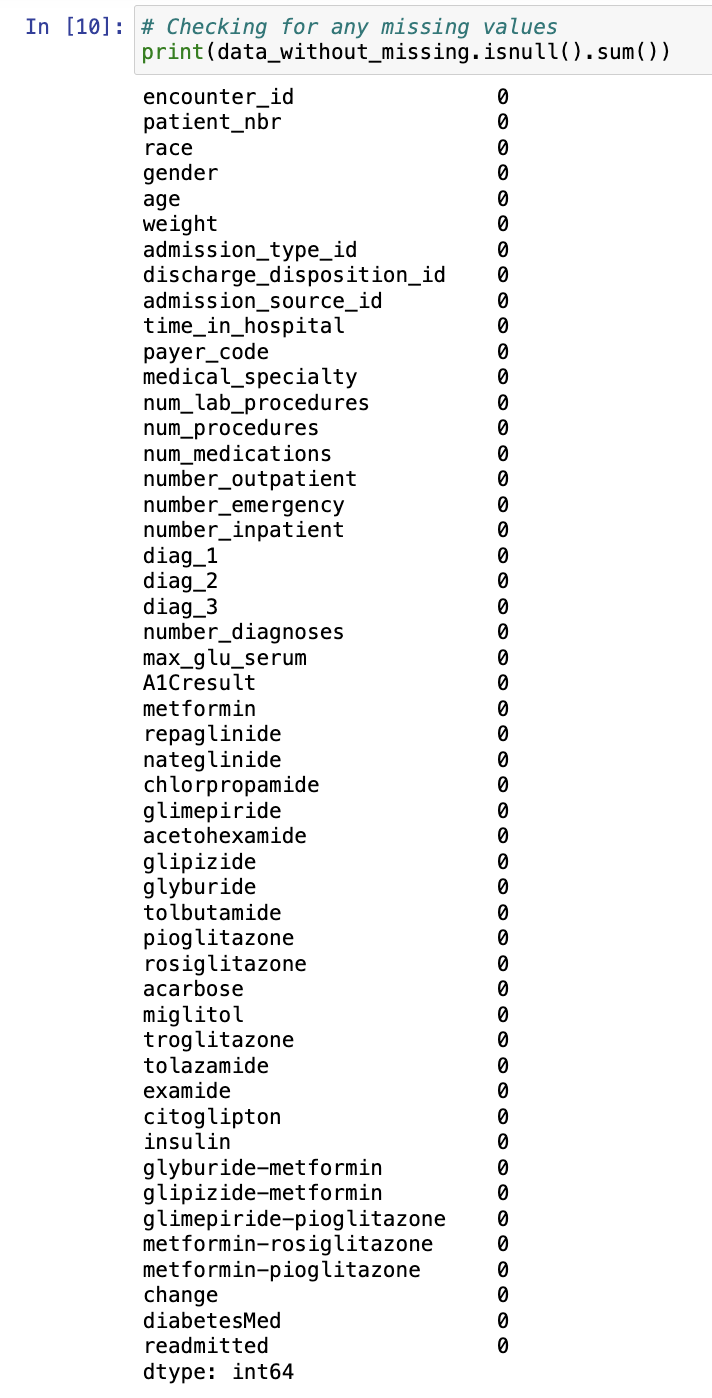
Explanation: Here we can see that the missing values were removed, and the number of rows is 298 and number of columns are 50.

1. Checking the first few rows of the data\_without\_missing (code)



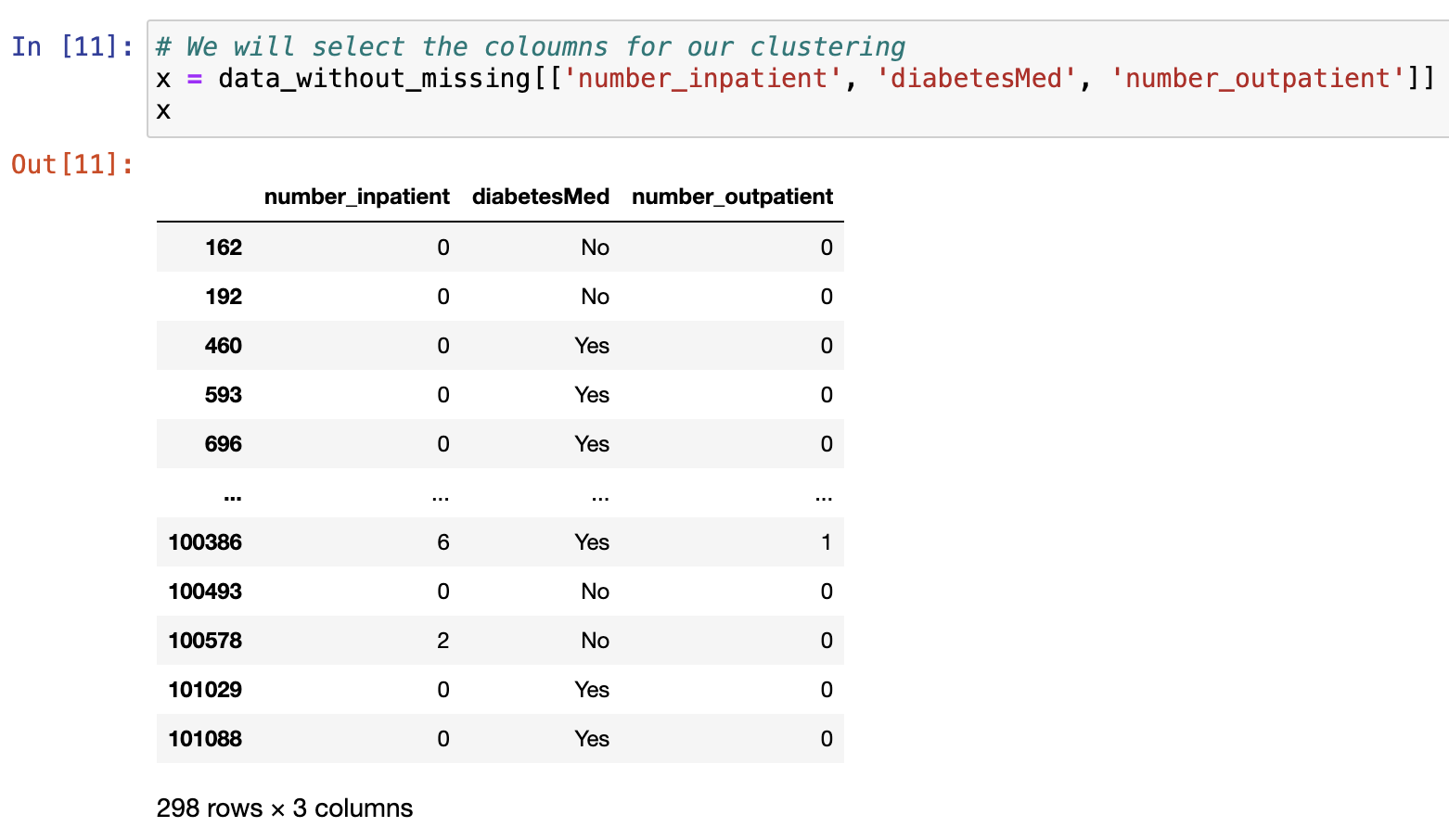
Explanation: Here we can see the first few rows of the data without the missing values.

1. Checking of any missing values incase left out (code)

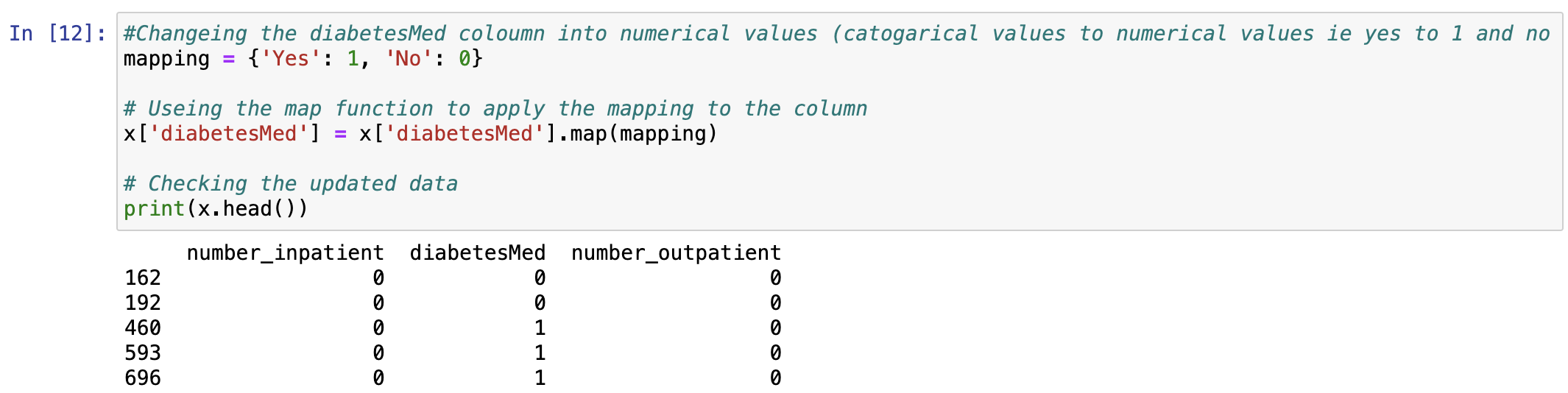


Explanation: Here we are checking the missing values in the new data in case any are left out.

1. Selecting of columns for our clustering (code)

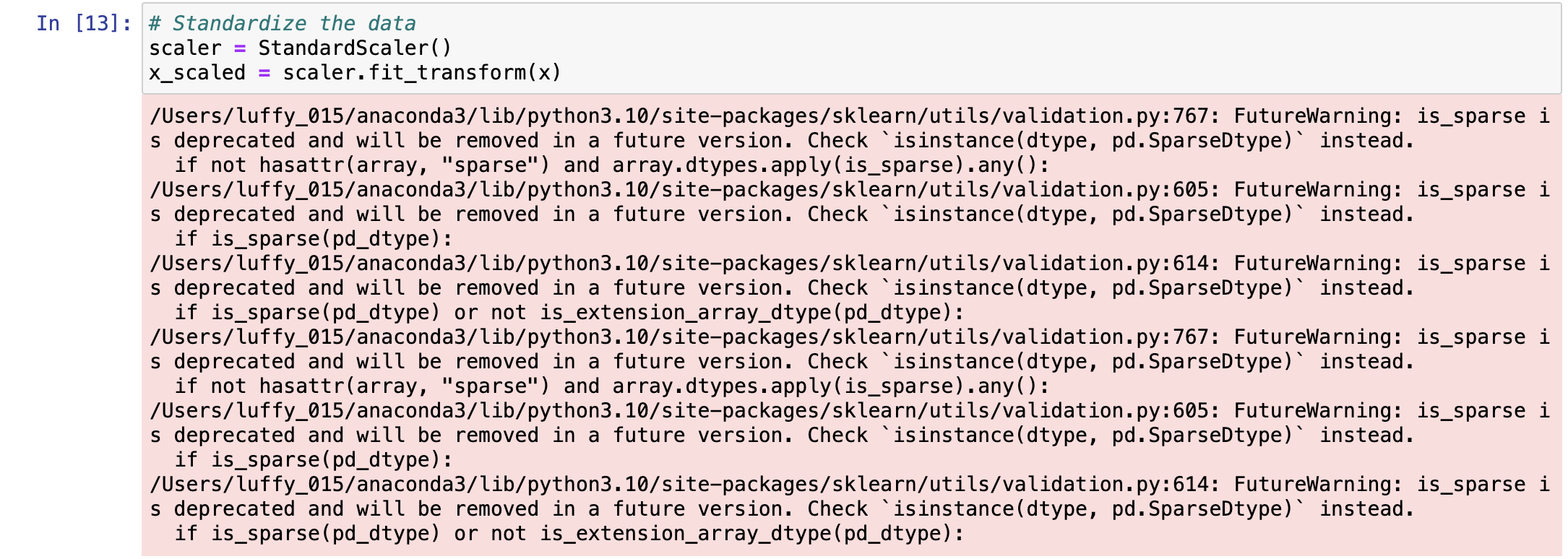
 Explanation: Here we are selecting the columns for our clustering that is number\_inpatient, number\_outpatient and diabetesMed.

1. Changing the diabetesMed, categorical values into numerical values (code)

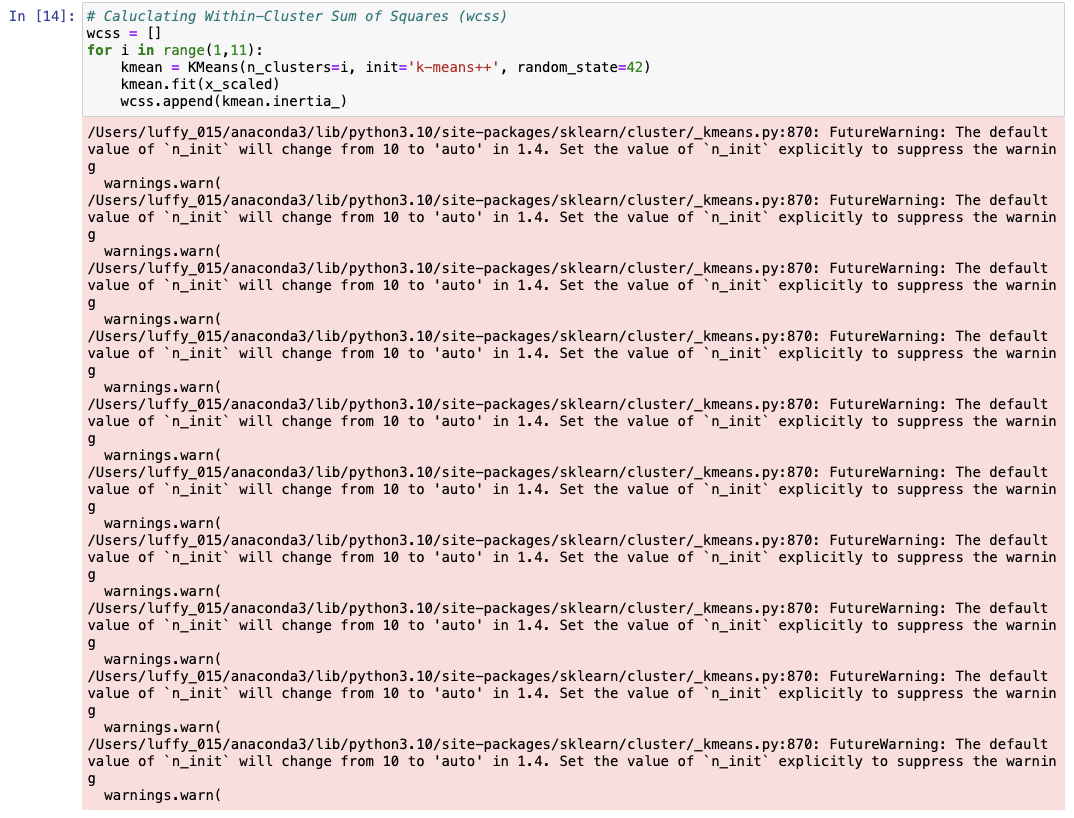
Explanation: Here we are converting the values in the 'diabetesMed' column from categorical ('Yes' and 'No') to numerical (1.0 and 0.0). Where we have assigned value 1.0 for Yes and 0.0 for No. After that we are viewing the head of the data,

**Here the conversion is necessary because the 'diabetesMed' column is likely to represent whether a patient was prescribed diabetes medication ('Yes') or not ('No'). By converting 'Yes' to 1.0 and 'No' to 0.0, we are transforming the categorical variable into a binary numerical variable, which can then be standardized along with other numerical features in the dataset.**

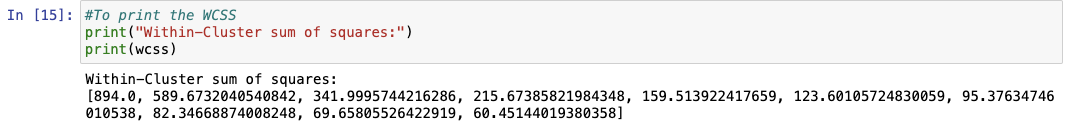
1. **Standardizing the data**

Explanation: Here we are Preprocessing the data to regularize the data and use it effectively.

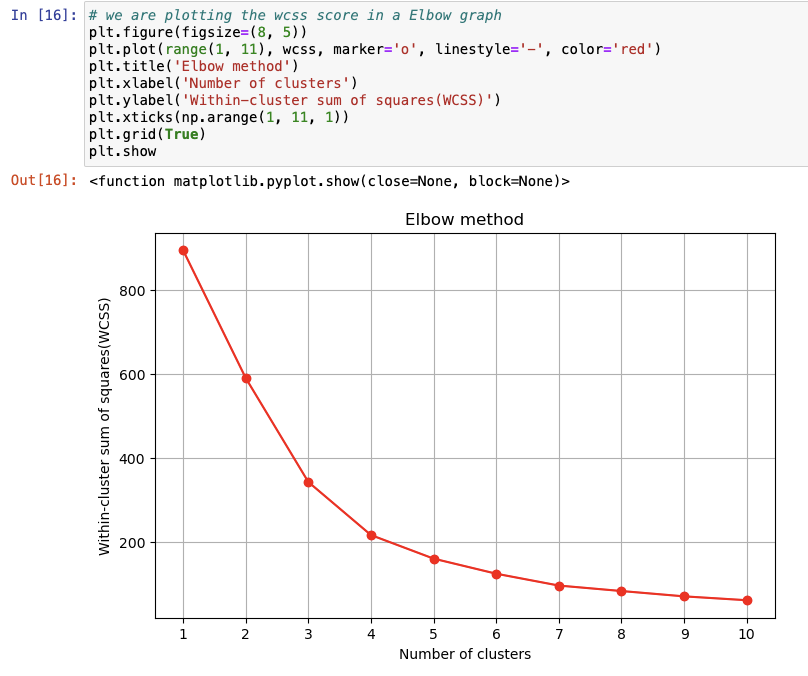
1. Calculating within cluster sum of square {wcss} (code)

Explanation: Here we are calculating the within cluster sum of square to determine the optimal number of clusters in the data set. **With that here we are using k-means++ algorithm instead of k-means because this leads to a better convergence and more accurate clustering results, and it gives us more stable, accurate, and faster convergence in the K-means clustering process.**

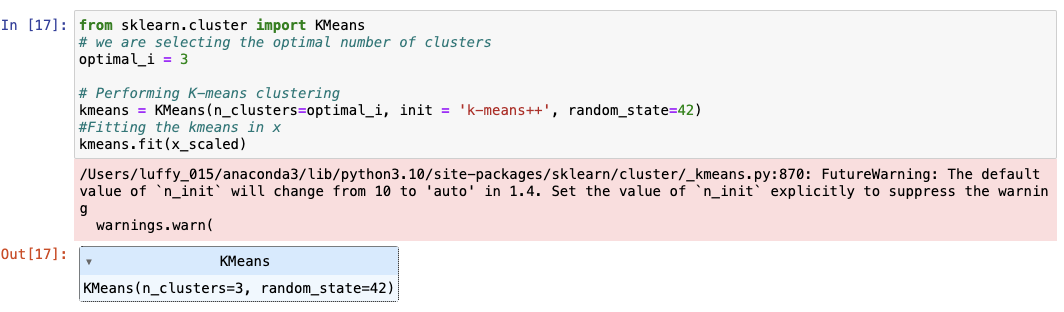
1. Printing the within cluster sum of squares (code)

 Explanation: Here we can see the result of the wcss.

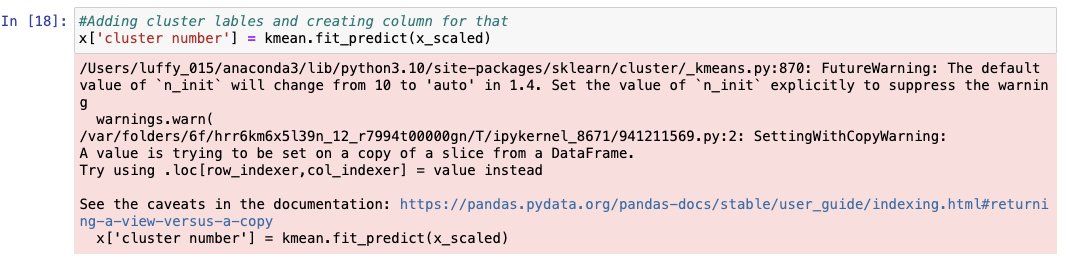
1. Plotting the wcss score in Lbow graph (code)

Explanation: Here we are plotting the Within-Cluster Sum of Squares (WCSS) scores on the "Elbow Method" graph. Here this method helps us to determine the optimal number of clusters by looking for the elbow point.

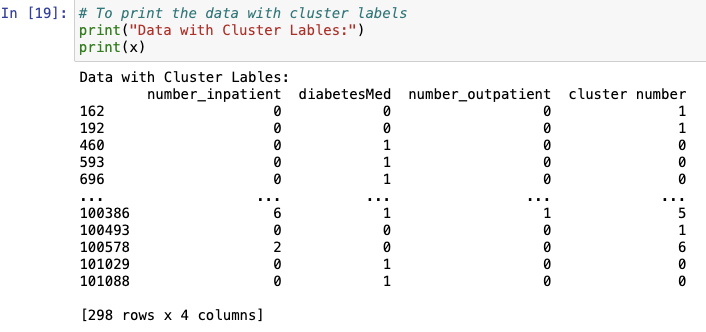
1. Performing K-means clustering (code)

Explanation: Here we are performing k means clustering by considering 4 clusters by initiating k-means++ algorithm and fitting it in x

1. Adding the cluster labels and creating column (code)

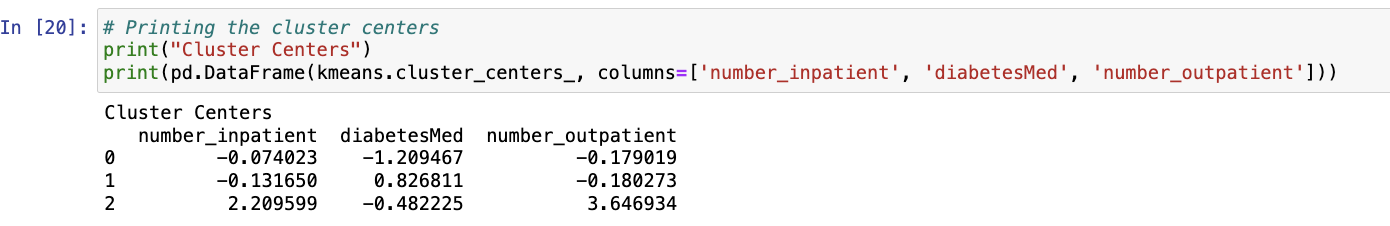
Explanation: Here we are adding the cluster labels for each data point and assigning them to a new column called 'cluster number' in the DataFrame ’x’.

1. Printing the number of clusters (code)

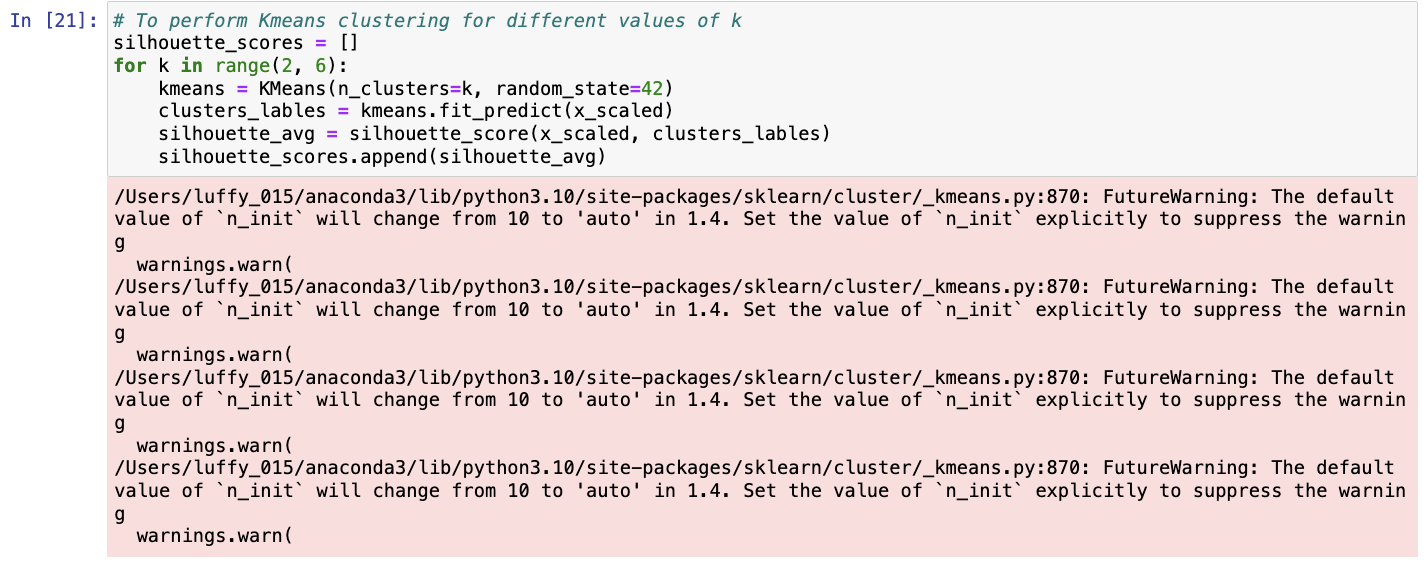


Explanation: For the above code we are getting the output by predicting the cluster labels for each data point and assigning them in cluster number.

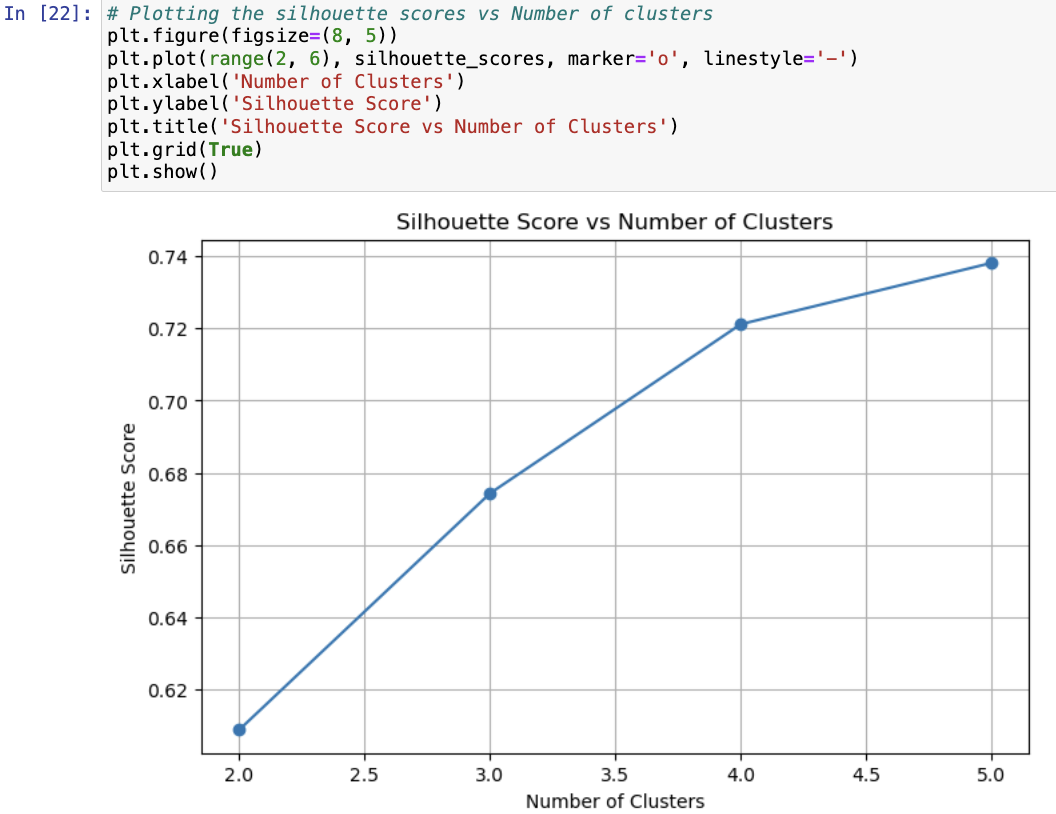
1. Printing the cluster centers

Explanation: Here we are finding out the mean of each cluster to gain insights into the characteristics of each cluster and understand the typical values of each feature within each cluster. This will help us to interpret and visualize the clustering results and identify meaningful patterns and differences between clusters.

1. To perform kmeans clustering with different values of k (code)

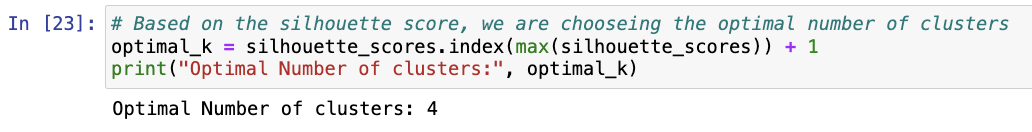
Explanation: Here we are using silhouette score to determine the optimal number of clusters. **we are using silhouette score because it helps us to measure the similarity of the clusters with its own clusters compared to other clusters. It ranges from -1 to 1, where higher score indicates better clustering.**

1. Plotting the graph (code)

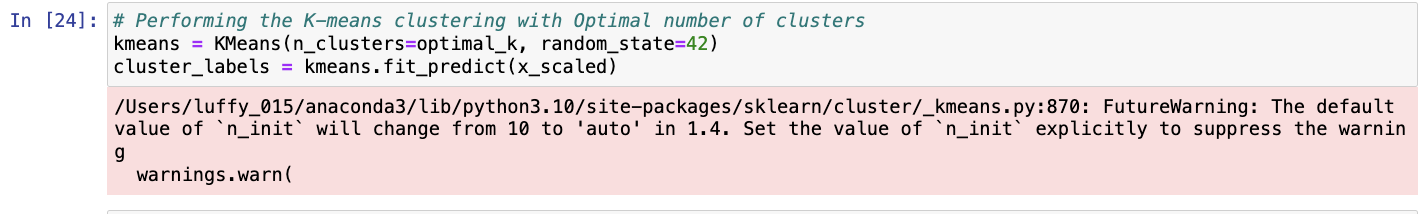


Explanation: Here we are plotting the graph based on number of clusters and silhouette score, where we can visually identify the optimal number of clusters.

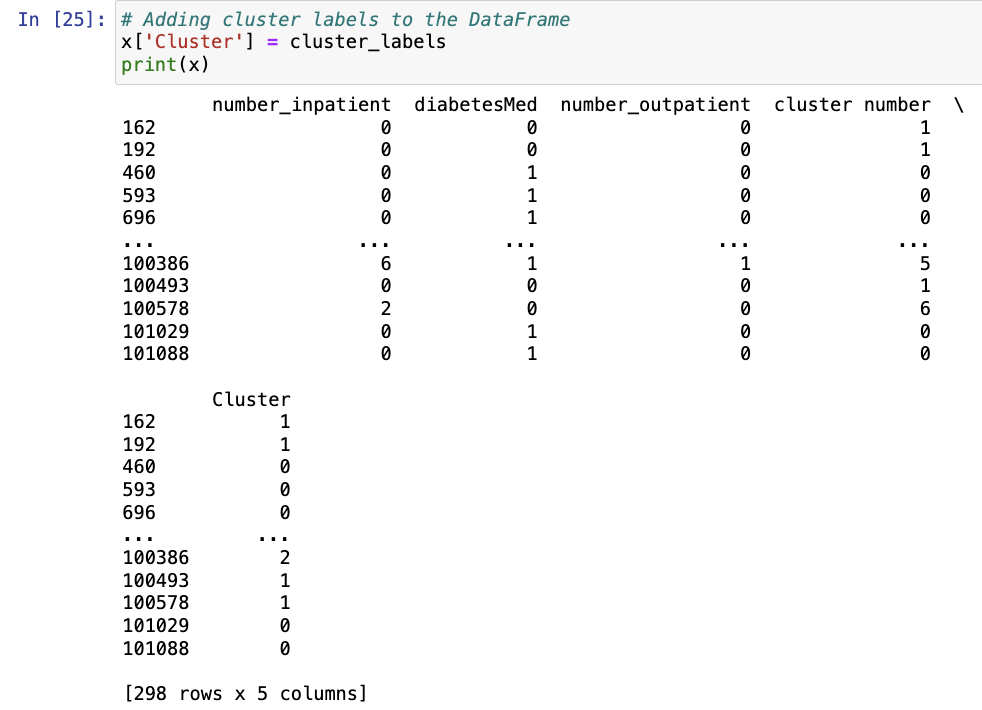
1. Choosing of number of clusters (code)

Explanation: Here we are choosing the optimal number of clusters based on silhouette score.

1. Performing K-means clustering

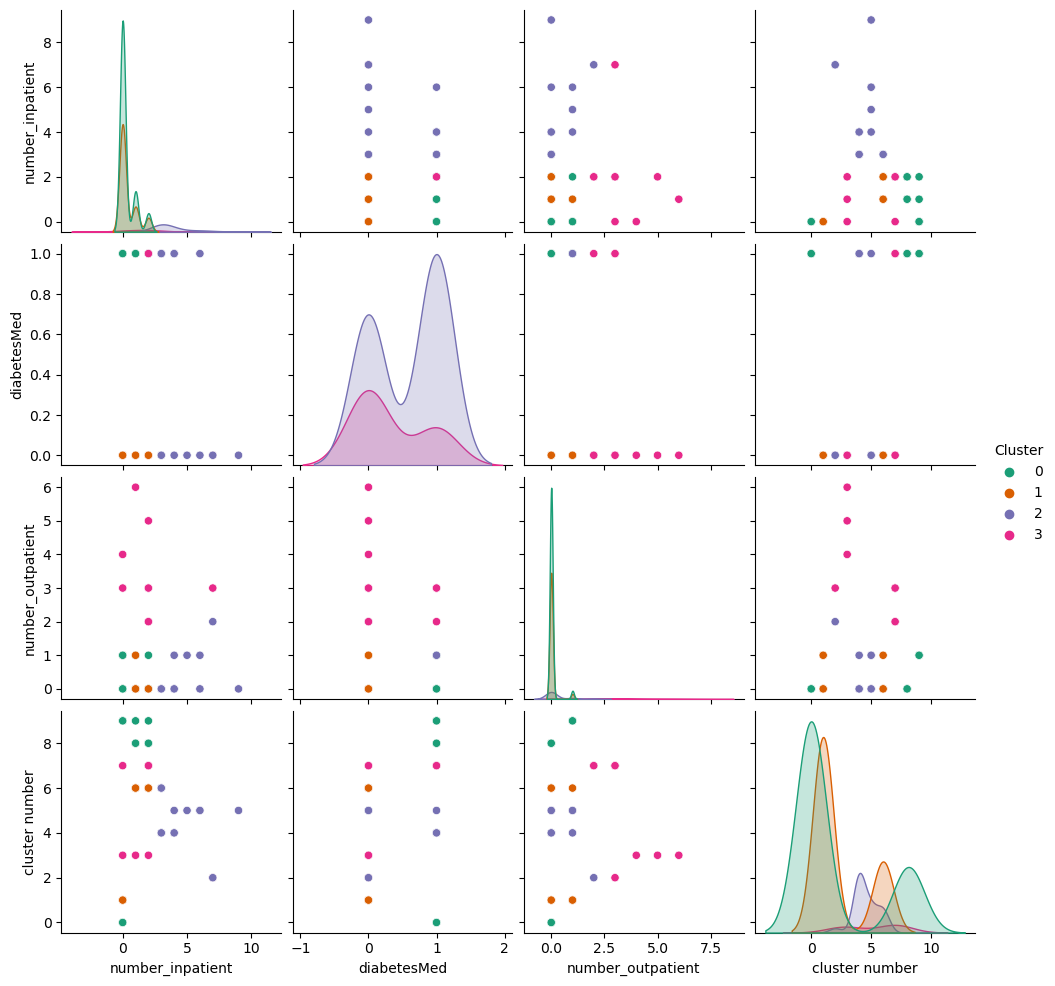
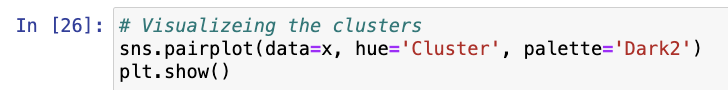
Explanation: Here we are performing the K-means clustering.

1. Adding the cluster labels (code)



Explanation: Here based on the clustering results obtained from the K-means algorithm we are creating additional column 'Cluster' that indicates the cluster assignment for each data point based on the K-means clustering. This will help us to analyze the clustering results.

1. Visualizing the clusters



Explanation: Here we are creating pair plot to visualize the clusters and gain insights into the relationship between different variables.

Top 10 Insights Derived After Applying Clustering

Techniques To The Dataset

1. Here the data is clustered into groups based on the number of inpatient visits, outpatient visits, and diabetes medication.
2. We can observe that there are 4 distinct clusters observer in data indicating different patient profiles.
3. Patients with higher numbers of inpatient visits tend to belong to one cluster, while those with lower numbers belong to another.
4. Patients who are on diabetes medication tend to be evenly distributed among the clusters.
5. The majority of patients have zero outpatient visits, which is evident in both clusters.
6. Here we can see a mixture of patients with and without diabetes medication in each cluster. By this we can say that medication alone may not be a determining factor for clustering.
7. The patients with higher inpatient visits may have more complex health conditions, leading to clustering in one group.
8. This clustering may help us in identifying high-risk patients who frequently require inpatient care.
9. Each cluster represents a different segment of patient.
10. Insights from clustering can guide us to personalized healthcare.