

بسم الله الرحمن الرحيم

**مشروع Data Mining & Warehousing**

مادة: data mining and warehousing

الشعبة:IB8B

|  |  |
| --- | --- |
| الرقم الجامعي | الاسم |
| 3901009 | عبد الرحمن محمود العيد |
| 3901498 | علي عدنان محبت |
| 3901555 | معاذ العوفي |

**First of all general information about data: -**

**1. Title: Haberman's Survival Data**

**2. Sources:**

(a) Donor: Tjen-Sien Lim (limt@stat.wisc.edu)

(b) Date: March 4, 1999

**3**. **Past** **Usage**:

1. Haberman, S. J. (1976). Generalized Residuals for Log-Linear

Models, Proceedings of the 9th International Biometrics

Conference, Boston, pp. 104-122.

2. Landwehr, J. M., Pregibon, D., and Shoemaker, A. C. (1984),

Graphical Models for Assessing Logistic Regression Models (with

discussion), Journal of the American Statistical Association 79: 61-83.

3. Lo, W.-D. (1993). Logistic Regression Trees, PhD thesis,

Department of Statistics, University of Wisconsin, Madison, WI.

**4. Relevant Information:**

The dataset contains cases from a study that was conducted between

1958 and 1970 at the University of Chicago's Billings Hospital on

the survival of patients who had undergone surgery for breast cancer.

**5. Number of Instances: 306**

**6. Number of Attributes: 4 (including the class attribute)**

**7. Attribute Information:**

1. Age of patient at time of operation (numerical)

2. Patient's year of operation (year - 1900, numerical)

3. Number of positive axillary nodes detected (numerical)

4. Survival status (class attribute)

1 = the patient survived 5 years or longer

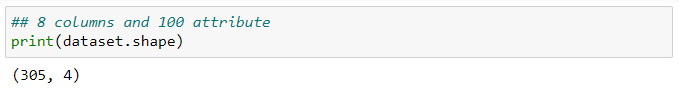
2 = the patient died within 5 year

**8. Missing Attribute Values: None**

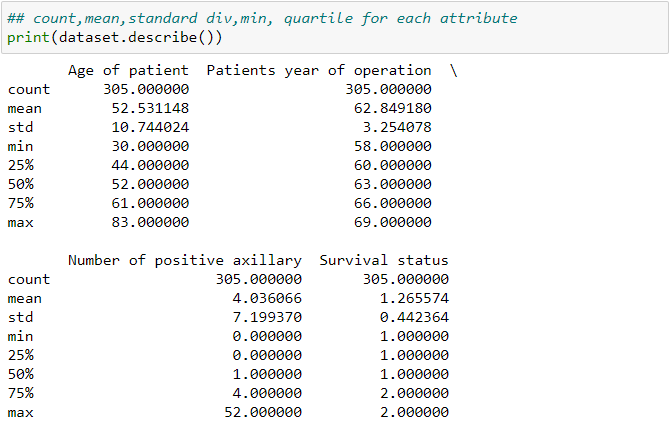
The following cell shows the importing of the dataset chosen for this project which is haberman and printing of the first 9 rows in the dataset using head function.



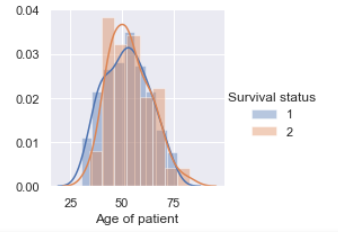
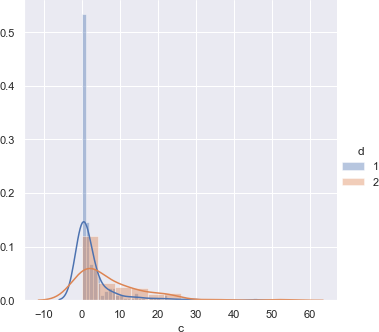
To check the dimensions of the dataset, it is done as follows.

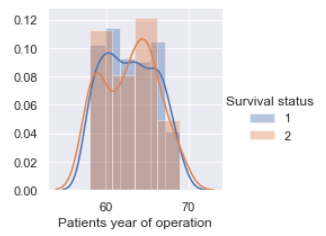


The description of the dataset gives an overview about the mean, std, min, 25%, 50%, 70% and max of each column as follows.

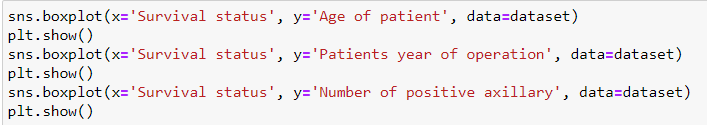


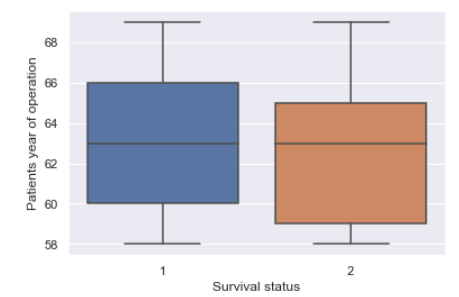
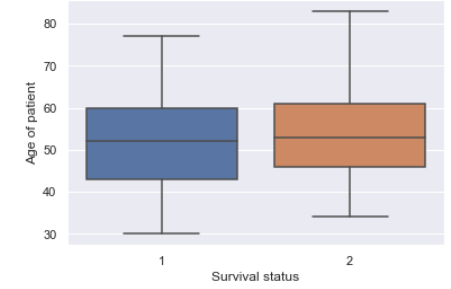
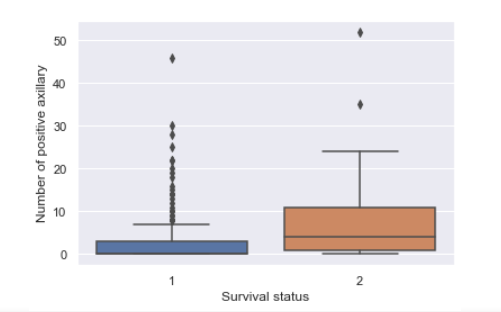
For visualizing the distribution of the data, the following is used. The distribution shows that the data is normally distributed.

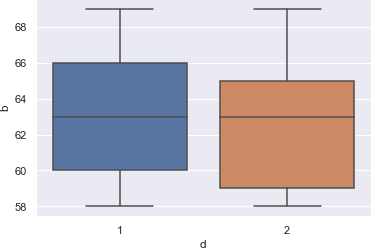


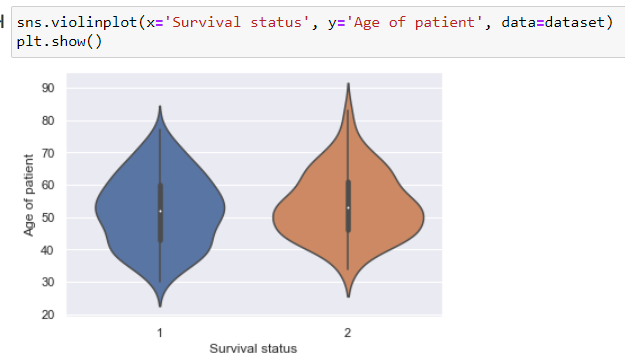


Also, a box plot is used in order to determine the quartiles in the data for the attributes

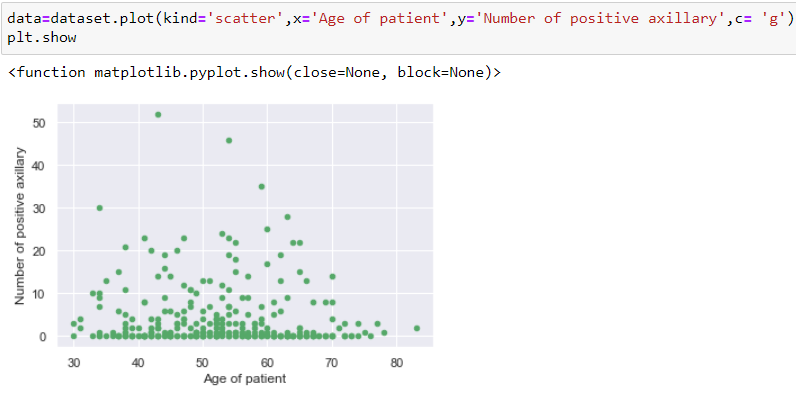


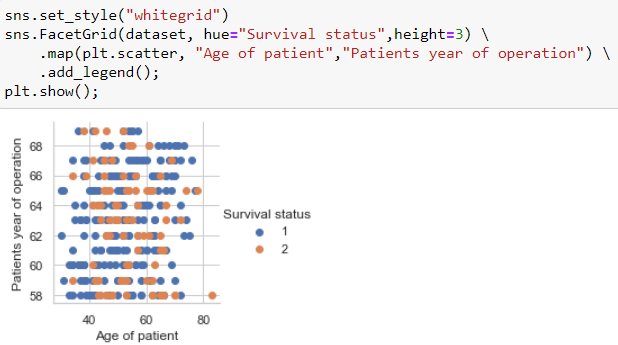




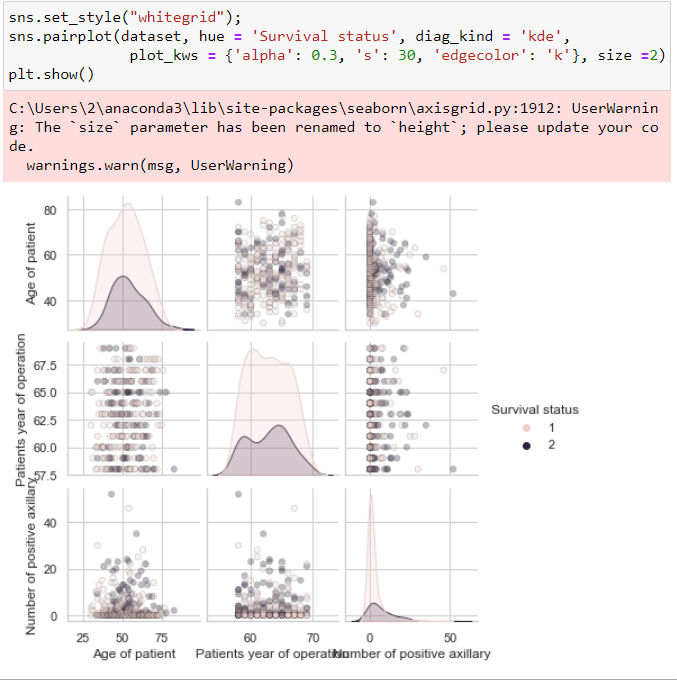


Also, the following scatter plot clearly shows that are outliers in the data as most of the data lies under approximately 25.

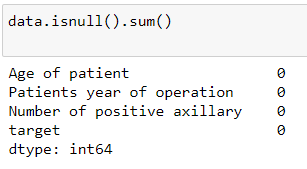




The falling pair plot creates a grid of Axes such that each attribute in data will by shared in the y-axis across a single row and in the x-axis across a single column.

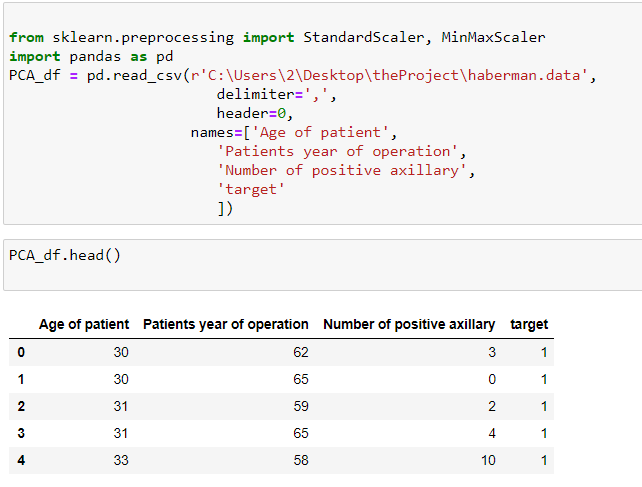


Checking if there are any null values.

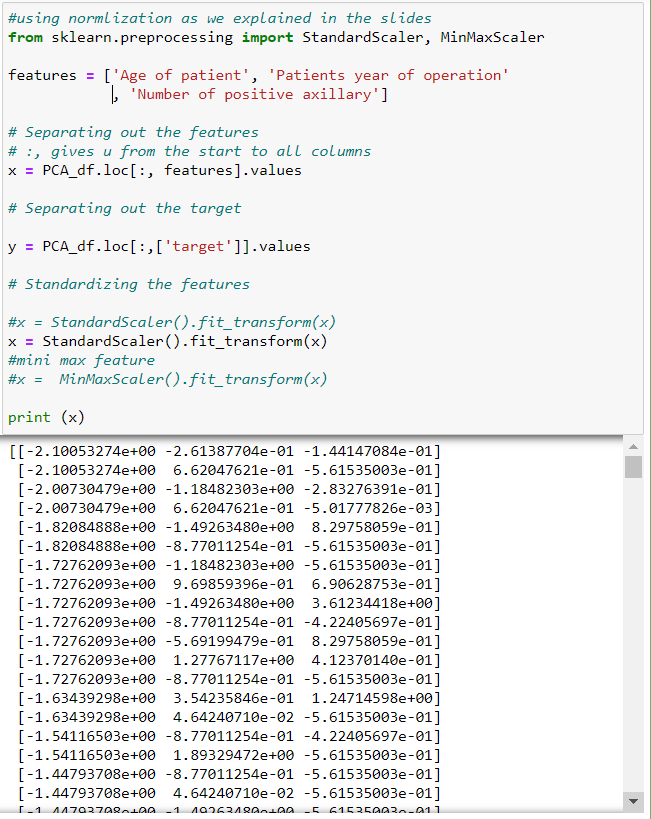


No null values and data type is int64

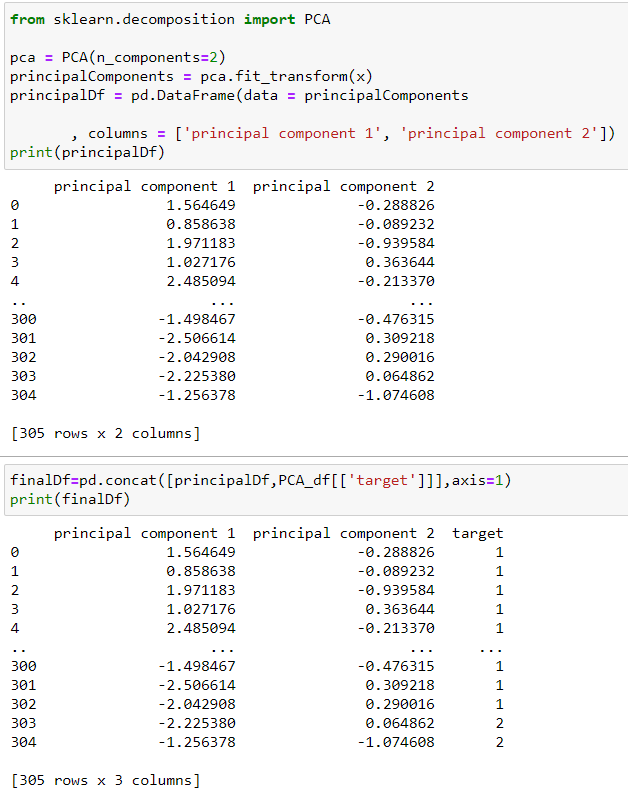
PCA is used to reduce the dimensionality of the dataset, increasing interpretability but at the same time minimizing information loss.

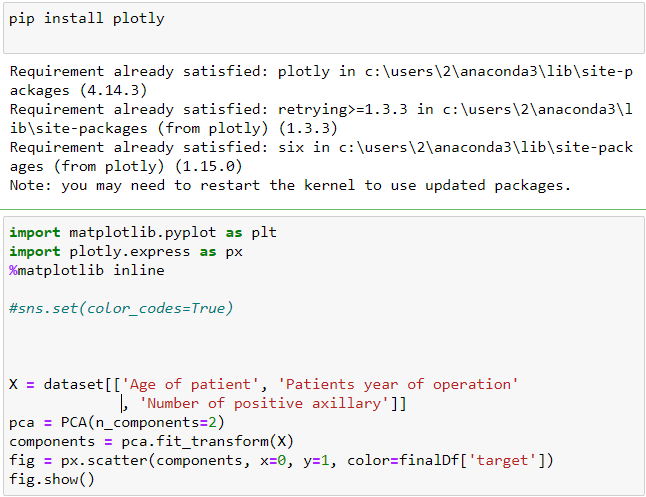


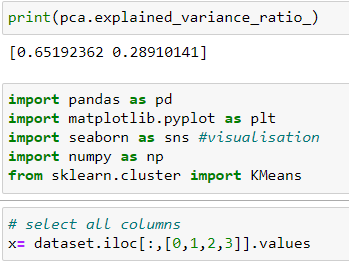
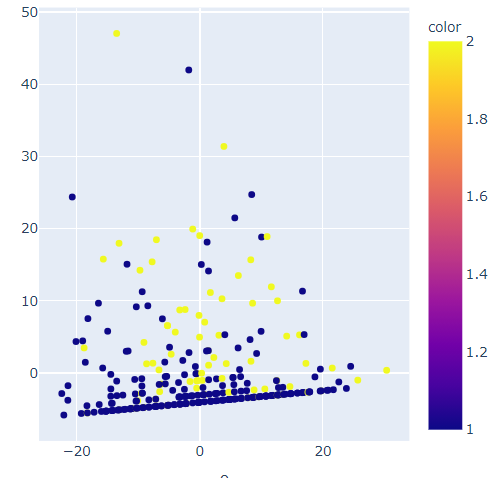
In order to preprocess the data, the StandardScaler and MinMaxScaler are used to normalize the data as follows.



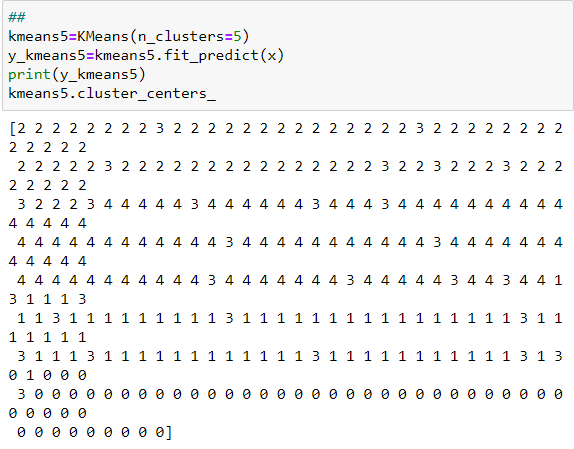
The PCA are constructed as follows.

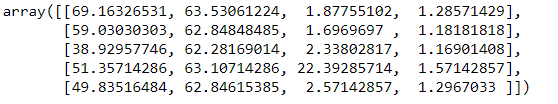


****

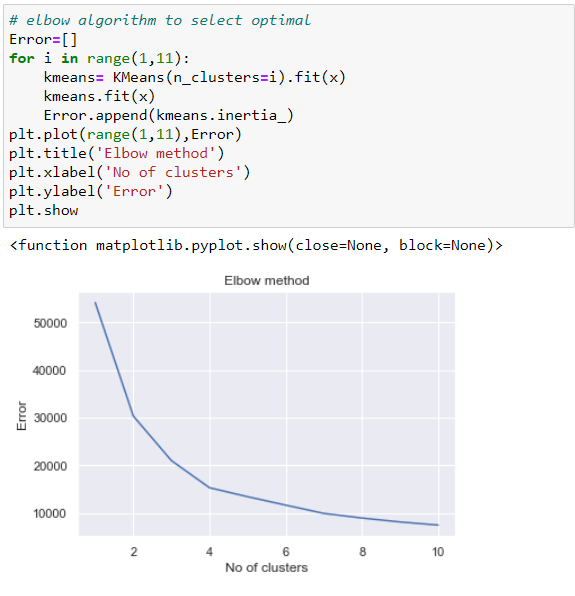
****

KMeans is a method of clustering to determine the clusters which the data fall in. Therefore, 5 clusters have been built where points are added to each cluster as follow.

****



The following Elbow algorithm is used in order to choose the optimal number of clusters for the dataset.

****

Thus, 3 clusters are used to model the data as follows using Kmeans.

