

**Project Report**

**Customer Classification based on the**

**historical purchase data**

**Submitted to: Submitted by:**

Mr. Sanjay Kumar Akash Gupta

**UID:** 15745 **Reg. No.** 11702783

**Roll no.** 36

**Section:** KM030

**Dataset Information:**

**Name:** Wholesale customers Data Set

**URL to download:** <http://archive.ics.uci.edu/ml/datasets/Wholesale+customers>

Alternate url:<https://www.kaggle.com/binovi/wholesale-customers-data-set>

**Source:** Margarida G. M. S. Cardoso, margarida.cardoso '@' iscte.pt, ISCTE-IUL,

Lisbon, Portugal

No. of records(rows) : 440

**Attributes:** There are total 6 attributes in the dataset namely:

* **FRESH:** annual spending (m.u.) on fresh products (Continuous)
* **MILK:** annual spending (m.u.) on milk products (Continuous)
* **GROCERY:** annual spending (m.u.) on grocery products (Continuous)
* **FROZEN:** annual spending (m.u.)on frozen products (Continuous)
* **DETERGENTS\_PAPER:** annual spending (m.u.) on detergents and paper products (Continuous)
* **DELICATESSEN:** annual spending (m.u.) on and delicatessen products (Continuous)

**Libraries Used:**

1. **NumPy:** NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays.  
   Numpy is used in this project to find logarithms, percentiles and other statistical values of data for better observation.
2. **Pandas:**  Pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool,built on top of the Python programming language.   
   In this project pandas is used to read a csv file which contains the data set, to create dataframes and for plotting scatter plot matrices.
3. **Matplotlib:** Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+.  
   In this project matplotlib is used to plot correlation matrix between six attributes depicting how each is dependent on other.
4. **PCA:** Principal component analysis is a technique for feature extraction — so it combines our input variables in a specific way, then we can drop the “least important” variables while still retaining the most valuable parts of all of the variables! As an added benefit, each of the “new” variables after PCA are all independent of one another. This is a benefit because the assumptions of a linear model require our independent variables to be independent of one another. If we decide to fit a linear regression model with these “new” variables (see “principal component regression” below), this assumption will necessarily be satisfied.
5. **Scikit Learn:** Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python.

It is licensed under a permissive simplified BSD license and is distributed under many Linux distributions, encouraging academic and commercial use.

The library is built upon the SciPy (Scientific Python) that must be installed before you can use scikit-learn. This stack that includes:

* 1. NumPy: Base n-dimensional array package
  2. SciPy: Fundamental library for scientific computing
  3. Matplotlib: Comprehensive 2D/3D plotting
  4. IPython: Enhanced interactive console
  5. Sympy: Symbolic mathematics
  6. Pandas: Data structures and analysis

Extensions or modules for SciPy care conventionally named SciKits. As such, the module provides learning algorithms and is named scikit-learn.

The vision for the library is a level of robustness and support required for use in production systems. This means a deep focus on concerns such as easy of use, code quality, collaboration, documentation and performance.

Although the interface is Python, c-libraries are leverage for performance such as numpy for arrays and matrix operations, LAPACK, LibSVM and the careful use of cython.  
Some popular groups of models provided by scikit-learn include:

* Clustering: for grouping unlabeled data such as KMeans.
* Cross Validation: for estimating the performance of supervised models on unseen data.
* Datasets: for test datasets and for generating datasets with specific properties for investigating model behavior.
* Dimensionality Reduction: for reducing the number of attributes in data for summarization, visualization and feature selection such as Principal component analysis.
* Ensemble methods: for combining the predictions of multiple supervised models.
* Feature extraction: for defining attributes in image and text data.
* Feature selection: for identifying meaningful attributes from which to create supervised models.
* Parameter Tuning: for getting the most out of supervised models.
* Manifold Learning: For summarizing and depicting complex multi-dimensional data.
* Supervised Models: a vast array not limited to generalized linear models, discriminate analysis, naive bayes, lazy methods, neural networks, support vector machines and decision trees.

**Github Link:** https://github.com/akash-404/Customer-Segmentation-ML-

All the graphs and observations are mentioned in the attached file.

**File Name:** UCI wholesale customer.ipynb

**References:**

1. <https://www.kaggle.com/binovi/wholesale-customers-data-set>
2. <http://archive.ics.uci.edu/ml/datasets/Wholesale+customers>
3. <https://subhayo.wordpress.com/2019/09/10/customer-segmentation-using-k-means-clustering/>
4. <https://sajalsharma.com/portfolio/customer_segments>
5. <https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781789807943/5/ch05lvl1sec46/understanding-the-wholesale-customer-dataset-and-the-segmentation-problem>
6. <https://moosend.com/blog/customer-segmentation-ecommerce/>
7. <https://scikit-learn.org/stable/>
8. <https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c>
9. <https://setosa.io/ev/principal-component-analysis/>
10. <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html>
11. <https://www.geeksforgeeks.org/confusion-matrix-machine-learning/>
12. <https://www.displayr.com/what-is-a-correlation-matrix/>