## PROJECT 2: CREATING CUSTOMER\_SEGMENTS

**PROJECT DESCRIPTION**

This is the 2nd main project for the Machine Learning Engineer Nano-degree. In this project I apply unsupervised learning techniques on product spending data collected for customers of a wholesale distributor to identify customer segments hidden in the data. Supervised learning techniques are also mentioned in this project.

Here, I first explore the data to determine if any product categories highly correlate with one another by observing a small subset of the data and also by plotting a scatter matrix.  Afterwards, I pre-processed the data by scaling each product category and then identifying (and removing) unwanted outliers. With the good, clean customer spending data, I applied PCA transformations to the data and implement clustering algorithms to segment the transformed customer data. Finally, I compare the segmentation found with an additional labelling and consider ways this information could assist the wholesale distributor with future service changes.

### INSTALL

This project requires **Python 2.7** and the following Python libraries installed:

* [NumPy](http://www.numpy.org/)
* [Pandas](http://pandas.pydata.org/)
* [matplotlib](http://matplotlib.org/)
* [scikit-learn](http://scikit-learn.org/stable/)

You will also need to have Anaconda software installed to run and execute a [jupyter notebook](http://ipython.org/notebook.html) file.

**CODE**

The main code for this project is located in the ‘customer\_segments.ipynb ‘ notebook file. Additional supporting code for visualizing the necessary graphs can be found in ‘visuals.py’. Additionally, the ‘report.html’ file contains a snapshot of the main code in the jupyter notebook with all code cells executed. ’customer.csv’ dataset file is used to complete the work.

**RUN**

In a terminal or command window, navigate to the top-level project directory customer\_segments and run one of the following command:

“ jupyter notebook customer\_segments.ipynb ”

This will open the Jupyter Notebook software and project file in your browser.

**DATA**

The customer segments data is included as a selection of 440 data points collected on data found from clients of a wholesale distributor in Lisbon, Portugal. More information can be found on the  [Machine Learning Repository](https://archive.ics.uci.edu/ml/datasets/Wholesale+customers) of udacity.

Note (m.u.) is shorthand formonetary units*.*

**Features:**

1. Fresh: annual spending (m.u.) on fresh products (Continuous);
2. Milk: annual spending (m.u.) on milk products (Continuous);
3. Grocery: annual spending (m.u.) on grocery products (Continuous);
4. Frozen: annual spending (m.u.) on frozen products (Continuous);
5. Detergents\_Paper: annual spending (m.u.) on detergents and paper products (Continuous);
6. Delicatessen: annual spending (m.u.) on and delicatessen products (Continuous);