**Project Overview**

In this project you will apply unsupervised learning techniques on product spending data collected for customers of a wholesale distributor in Lisbon, Portugal to identify customer segments hidden in the data. You will first explore the data by selecting a small subset to sample and determine if any product categories highly correlate with one another. Afterwards, you will preprocess the data by scaling each product category and then identifying (and removing) unwanted outliers. With the good, clean customer spending data, you will apply PCA transformations to the data and implement clustering algorithms to segment the transformed customer data. Finally, you will compare the segmentation found with an additional labeling and consider ways this information could assist the wholesale distributor with future service changes.

Prepare for this project with our supplementary course on [**Unsupervised Learning**](https://classroom.udacity.com/courses/ud727-nd).

**Project Highlights**

This project is designed to give you a hands-on experience with unsupervised learning and work towards developing conclusions for a potential client on a real-world dataset. Many companies today collect vast amounts of data on customers and clientele, and have a strong desire to understand the meaningful relationships hidden in their customer base. Being equipped with this information can assist a company engineer future products and services that best satisfy the demands or needs of their customers.

Things you will learn by completing this project:

* How to apply preprocessing techniques such as feature scaling and outlier detection.
* How to interpret data points that have been scaled, transformed, or reduced from PCA.
* How to analyze PCA dimensions and construct a new feature space.
* How to optimally cluster a set of data to find hidden patterns in a dataset.
* How to assess information given by cluster data and use it in a meaningful way.

**Project Description**

A wholesale distributor recently tested a change to their delivery method for some customers, by moving from a morning delivery service five days a week to a cheaper evening delivery service three days a week. Initial testing did not discover any significant unsatisfactory results, so they implemented the cheaper option for all customers. Almost immediately, the distributor began getting complaints about the delivery service change and customers were canceling deliveries — losing the distributor more money than what was being saved. You’ve been hired by the wholesale distributor to find what types of customers they have to help them make better, more informed business decisions in the future. Your task is to use unsupervised learning techniques to see if any similarities exist between customers, and how to best segment customers into distinct categories.

For this assignment, you can find the project files as a downloadable in the Resources section ascustomer\_segments.zip. You may also visit our [**Projects GitHub**](https://github.com/udacity/machine-learning/tree/master/projects) to have access to all of the projects available for this Nanodegree. While some code has already been implemented to get you started, you will need to implement additional functionality to successfully answer all of the questions included in the notebook. You can find the included questions for reference on the following slide.

**Software and Libraries**

For this project, you will need to have the following software installed:

* [**Python 2.7**](https://www.python.org/download/releases/2.7/)
* [**NumPy**](http://www.numpy.org/)
* [**pandas**](http://pandas.pydata.org/)
* [**matplotlib**](http://matplotlib.org/)
* [**scikit-learn**](http://scikit-learn.org/stable/)
* [**iPython Notebook**](http://ipython.org/notebook.html)

**Deliverables**

The following files should be included as your submission, and can be packaged as a single zip file for convenience:

* The customer\_segments.ipynb file with fully implemented, functional code, with all code blocks executed and showing output.
* An HTML or PDF report of the project (you may either export the notebook as HTML with your answers included, or submit a separate PDF with only the questions and your answers).

## Questions and Report Structure

### Data Exploration

* What kind of establishment (customer) could each of the three samples chosen from the customer data represent?
* Which feature did you attempt to predict? What was the reported prediction score? Is this feature relevant for identifying a specific customer?
* Are there any pairs of features which exhibit some degree of correlation? Does this confirm or deny your suspicions about the relevance of the feature you attempted to predict? How is the data for those features distributed?

### Data Preprocessing

* Are there any data points considered outliers for more than one feature? Should these data points be removed from the dataset? If any data points were added to the outliers list to be removed, explain why.
* How much variance in the data is explained in total by the first and second principal component? What about the first four principal components? Using the visualization provided in the notebook, discuss what the first four dimensions best represent in terms of customer spending.

### Clustering

* What are the advantages to using a K-Means clustering algorithm? What are the advantages to using a Gaussian Mixture Model clustering algorithm? Given your observations about the wholesale customer data so far, which of the two algorithms will you use and why?
* Report the silhouette score for several cluster numbers you tried. Of these, which number of clusters has the best silhouette score?
* Consider the total purchase cost of each product category for the representative data points found from the cluster centers, and reference the statistical description of the dataset at the beginning of the notebook. What set of establishments could each of the customer segments represent?
* For each sample point, which customer segment from the question above best represents it? Are the predictions for each sample point consistent with this?

### Conclusion

* Companies often run [**A/B tests**](https://en.wikipedia.org/wiki/A/B_testing) when making small changes to their products or services. If the wholesale distributor wanted to change its delivery service from 5 days a week to 3 days a week, how would you use the structure of the data to help them decide on a group of customers to test?
* Assume the wholesale distributor wanted to predict some other feature for each customer based on the purchasing information available. How could the wholesale distributor use the structure of the data to assist a supervised learning analysis?
* How well does the clustering algorithm and number of clusters you've chosen compare to the underlying distribution of Hotel/Restaurant/Cafe customers to Retailer customers? Are there customer segments that would be classified as purely 'Retailers' or 'Hotels/Restaurants/Cafes' by this distribution? Would you consider these classifications as consistent with your previous definition of the customer segments?