**Creating-Customer-Segments**

Machine Learning Nanodegree Project Udacity

**Introduction**

This repo contains all my work for Project 2 of Udacity's Machine Learning Basic Nanodegree Program. In this project I applied 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. I first explored the data by selecting a small subset to sample and determine if any product categories highly correlate with one another. 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 compared the segmentation found with an additional labelling and consider ways this information could assist the wholesale distributor with future service changes.

**Disclaimer:**

As a CS minor student of IIT Kharagpur and a long-time self-taught learner, I have completed many CS related MOOCs on Coursera, Udacity, Udemy, and Edx. I do understand the hard time you spend on understanding new concepts and debugging your program. Here I released these solutions, which are **only for your reference purpose**. It may help you to save some time. And I hope you don't copy any part of the code (the programming assignments are fairly easy if you read the instructions carefully), see the solutions before you start your own adventure. This Project is almost one of the simplest Machine Learning Project I have ever taken, but the simplicity is based on the fabulous course content and structure. It's a treasure given by Udacity team.

**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 pre-process 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 labelling and consider ways this information could assist the wholesale distributor with future service changes.

**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:**

1. How to apply pre-processing techniques such as feature scaling and outlier detection.
2. How to interpret data points that have been scaled, transformed, or reduced from PCA.
3. How to analyse PCA dimensions and construct a new feature space.
4. How to optimally cluster a set of data to find hidden patterns in a dataset.
5. How to assess information given by cluster data and use it in a meaningful way.

**Problem Statement:**

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 cancelling 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.