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.

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 with 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.

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.

Software and Libraries

This project uses the following software and Python libraries:

- **Python**
- **NumPy**
- **pandas**
- **scikit-learn** (v0.17)
- **matplotlib**
  You will also need to have software installed to run and execute a **Jupyter Notebook**.
  If you do not have Python installed yet, it is highly recommended that you install the **Anaconda** distribution of Python, which already has the above packages and more included.

Starting the Project

For this assignment, you can find the `customer_segments` folder containing the necessary project files on the **Machine Learning projects GitHub**, under the `projects` folder. You may download all of the files for projects we'll use in this Nanodegree program directly from this repo. Please make sure that you use the most recent version of project files when completing a project!
This project contains three files:

- `customer_segments.ipynb`: This is the main file where you will be performing your work on the project.
- `customers.csv`: The project dataset. You'll load this data in the notebook.
- `visuals.py`: This Python script provides supplementary visualizations for the project. Do not modify.

In the Terminal or Command Prompt, navigate to the folder containing the project files, and then use the command `jupyter notebook customer_segments.ipynb` to open up a browser window or tab to work with your notebook. Alternatively, you can use the command `jupyter notebook` or `ipython notebook` and navigate to the notebook file in the browser window that opens. Follow the instructions in the notebook and answer each question presented to successfully complete the project.
A **README** file has also been provided with the project files which may contain additional necessary information or instruction for the project.

Submitting the Project

Evaluation

Your project will be reviewed by a Udacity reviewer against the **Creating Customer Segments project rubric**. Be sure to review this rubric thoroughly and self-evaluate your project before submission. All criteria found in the rubric must be *meeting specifications* for you to pass.

Submission Files

Following files would be needed for evaluation:

- The `customer_segments.ipynb` notebook file with all questions answered and all code cells executed and displaying output.
- An **HTML** export of the project notebook with the name **report.html**. This file *must* be present for your project to be evaluated.

When you are ready to submit your project, There are three ways in which your project can be submitted for evaluation.

1. If you ran the notebook from your **local machine** collect the above files and compress them into a single archive for upload.
2. You could supply the above files on your **GitHub Repo** in a folder named `customer_segments` for ease of access. This would build a good Github profile in parallel.
3. If you worked using the **workspace inside the classroom** you can submit your project directly for review using the submit button at the end of project, just make sure you download the HTML report to local machine and upload it back into workspace BEFORE submitting your report.

PROJECT SPECIFICATION

Creating Customer Segments

Data Exploration

CRITERIA	MEETS SPECIFICATIONS
Question 1 Selecting Samples	Three separate samples of the data are chosen and their establishment representations are proposed based on the statistical description of the dataset.

CRITERIA	MEETS SPECIFICATIONS
Question 2 Feature Relevance	A prediction score for the removed feature is accurately reported. Justification is made for whether the removed feature is relevant.
Question 3 Feature Distributions	Student identifies features that are correlated and compares these features to the predicted feature. Student further discusses the data distribution for those features.

Data Preprocessing

CRITERIA	MEETS SPECIFICATIONS
Feature Scaling	Feature scaling for both the data and the sample data has been properly implemented in code.
Question 4 Outlier Detection	Student identifies extreme outliers and discusses whether the outliers should be removed. Justification is made for any data points removed.

Feature Transformation

CRITERIA	MEETS SPECIFICATIONS
Question 5 Principal Component Analysis	The total variance explained for two and four dimensions of the data from PCA is accurately reported. The first four dimensions are interpreted as a representation of customer spending with justification.
Dimensionality Reduction	PCA has been properly implemented and applied to both the scaled data and scaled sample data for the two-dimensional case in code.

Clustering

CRITERIA	MEETS SPECIFICATIONS
Question 6 Clustering Algorithm	The Gaussian Mixture Model and K-Means algorithms have been compared in detail. Student's choice of algorithm is justified based on the characteristics of the algorithm and data.
Question 7 Creating Clusters	Several silhouette scores are accurately reported, and the optimal number of clusters is chosen based on the best reported score. The cluster visualization provided produces the optimal number of clusters based on the clustering algorithm chosen.
Question 8 Data Recovery	The establishments represented by each customer segment are proposed based on the statistical description of the dataset. The inverse transformation and inverse scaling has been properly implemented and applied to the cluster centers in code.
Question 9 Sample Predictions	Sample points are correctly identified by customer segment, and the predicted cluster for each sample point is discussed.

Conclusion

CRITERIA	MEETS SPECIFICATIONS
Question 10 A/B Test	Student correctly identifies how an A/B test can be performed on customers after a change in the wholesale distributor's service.
Question 11 Predicting Additional Data	Student discusses with justification how the clustering data can be used in a supervised learner for new predictions.
Question 12 Comparing Customer Data	Comparison is made between customer segments and customer 'Channel' data. Discussion of customer segments being identified by 'Channel' data is provided, including whether this representation is consistent with previous results.