**Project – Identify Customer Segments**

**Project Introduction**

Welcome to the project for this part of the Nanodegree program. Now that you’ve learned about unsupervised learning techniques that allow you to summarize the dimensions of your data and cluster unlabelled data into groups with similar properties, it’s time to apply your knowledge to some real-life data.

In this project, our Bertelsmann partners AZ Direct and Arvato Financial Solutions have provided two datasets one with demographic information about the people of Germany, and one with that same information for customers of a mail-order sales company. You’ll look at relationships between demographics features, organize the population into clusters, and see how prevalent customers are in each of the segments obtained.

Before you begin, Timo will provide some background about the dataset, how it was prepared, and the tasks performed by a data scientist using this data.

We hope you’re excited to test your skills on some real-life data. Don’t forget that you’ll need to make use of more than what was covered in the preceding course. You’ll need to use wrangling techniques to organize the data into a form you can work with, and you’ll need to interpret the output of your operations.

The remainder of this lesson will provide you instructions for completing the project, and a rubric detailing how the project will be evaluated. Rise up to the challenge, and let’s see what patterns you can find in the data!

**Project Overview**

In this project, you will work with real-life data provided to us by our Bertelsmann partners AZ Direct and Arvato Finance Solution. The data here concerns a company that performs mail-order sales in Germany. Their main question of interest is to identify facets of the population that are most likely to be purchasers of their products for a mailout campaign. Your job as a data scientist will be to use unsupervised learning techniques to organize the general population into clusters, then use those clusters to see which of them comprise the main user base for the company. Prior to applying the machine learning methods, you will also need to assess and clean the data in order to convert the data into a usable form.

**What will I need to install?**

It is highly recommended that you use the project workspace to complete and submit this project. Should you decide to download the data, you will need to remove it from your computer via the agreement with Arvato Bertlesmann. If you choose to use your machine to complete this project, below is a list of the requirements.

This project uses Python 3 and is designed to be completed through the Jupyter Notebooks IDE. It is highly recommended that you use the Anaconda distribution to install Python, since the distribution includes all necessary Python libraries as well as Jupyter Notebooks. The following libraries are expected to be used in this project:

* **NumPy**
* **pandas**
* **Sklearn** / scikit-learn
* **Matplotlib** (for data visualization)
* **Seaborn** (for data visualization)

**Why this project?**

The unsupervised learning branch of machine learning is key in the organization of large and complex datasets. While unsupervised learning lies in contrast to supervised learning in the fact that unsupervised learning lacks objective output classes or values, it can still be important in converting the data into a form that can be used in a supervised learning task. Dimensionality reduction techniques can help surface the main signals and associations in your data, providing supervised learning techniques a more focused set of features upon which to apply their work. Clustering techniques are useful for understanding how the data points themselves are organized. These clusters might themselves be a useful feature in a directed supervised learning task. This project will give you hands-on experience with a real-life task that makes use of these techniques, focusing on the unsupervised work that goes into understanding a dataset.

In addition, the dataset presented in this project requires a number of assessment and cleaning steps before you can apply your machine learning methods. In workplace contexts, you will frequently need to work with data that is untidy or needs pre-processing before standard algorithms and models can be applied. The ability to perform data wrangling and the ability to make decisions on data that you work with are both valuable skills that you will practice in this project.

**Project Details**

Below, you will find an outline of the steps that you will take in this project. It is recommended that you read through all of these instructions before starting on the project so you can build a high-level understanding of the tasks ahead!

**Step 0: The Data**

There are a number of files used for this project available in the workspace on the next page. **It is strongly recommended that you complete and turn in this project from the workspace.** As you will see from the agreement on the next page, if you choose to download a .zip file you will need to **REMOVE** the files from your computer via the agreement with Arvato Bartlesmann after completing the project, as the data are proprietary.

The files available for the project are:

* Udacity\_AZDIAS\_Subset.csv: Demographic data for the general population of Germany; 891211 persons (rows) x 85 features (columns).
* Udacity\_CUSTOMERS\_Subset.csv: Demographic data for customers of a mail-order company; 191652 persons (rows) x 85 features (columns).
* Data\_Dictionary.md: Information file about the features in the provided datasets.
* AZDIAS\_Feature\_Summary.csv: Summary of feature attributes for demographic data.
* Identify\_Customer\_Segments.ipynb: Jupyter Notebook divided into sections and guidelines for completing the project. The notebook provides more details and tips than the outline given here.
* After downloading the data and unzipping it to a working folder of your choice, use the Terminal or Command Prompt to navigate to your selected folder and execute the command jupyter notebook Identify\_Customer\_Segments.ipynb to open up a browser window or tab to work on the project notebook.

**Step 1: Preprocessing**

When you start an analysis, you must first explore and understand the data that you are working with. In this (and the next) step of the project, you’ll be working with the general demographics data. As part of your investigation of dataset properties, you must attend to a few key points:

* How are missing or unknown values encoded in the data? Are there certain features (columns) that should be removed from the analysis because of missing data? Are there certain data points (rows) that should be treated separately from the rest?
* Consider the level of measurement for each feature in the dataset (e.g. categorical, ordinal, numeric). What assumptions must be made in order to use each feature in the final analysis? Are there features that need to be re-encoded before they can be used? Are there additional features that can be dropped at this stage?

You will create a cleaning procedure that you will apply first to the general demographic data, then later to the customers data.

**Step 2: Feature Transformation**

Now that your data is clean, you will use dimensionality reduction techniques to identify relationships between variables in the dataset, resulting in the creation of a new set of variables that account for those correlations. In this stage of the project, you will attend to the following points:

* The first technique that you should perform on your data is feature scaling. What might happen if we don’t perform feature scaling before applying later techniques you’ll be using?
* Once you’ve scaled your features, you can then apply principal component analysis (PCA) to find the vectors of maximal variability. How much variability in the data does each principal component capture? Can you interpret associations between original features in your dataset based on the weights given on the strongest components? How many components will you keep as part of the dimensionality reduction process?

You will use the **sklearn** library to create objects that implement your feature scaling and PCA dimensionality reduction decisions.

**Step 3: Clustering**

Finally, on your transformed data, you will apply clustering techniques to identify groups in the general demographic data. You will then apply the same clustering model to the customers dataset to see how market segments differ between the general population and the mail-order sales company. You will tackle the following points in this stage:

* Use the k-means method to cluster the demographic data into groups. How should you make a decision on how many clusters to use?
* Apply the techniques and models that you fit on the demographic data to the customers data: data cleaning, feature scaling, PCA, and k-means clustering. Compare the distribution of people by cluster for the customer data to that of the general population. Can you say anything about which types of people are likely consumers for the mail-order sales company?

**sklearn** will continue to be used in this part of the project, to perform your k-means clustering. In the end, you will export the completed notebook with your work as an HTML file, which will serve as a report documenting your approach and findings.

**Step 4: Review and Submit the Project**

Before you submit your project for review, you should first review the [**project rubric**](http://review.udacity.com/#!/projects/f5062c28-1399-403e-af18-0e5b0e4587fd/rubric). Your project will be evaluated by an Udacity reviewer using this criterion, so you should make a self-evaluation as well. In order for your project to pass review, it must meet all specifications.

Once you are ready for project submission, export an HTML version of your notebook and add to your workspace submission, then submit your entire project from the workspace using in the button in the bottom right.

**Project Submission**

The data and design for this project were provided by Arvato Financial Services. You will apply unsupervised learning techniques on demographic and spending data for a sample of German households. You will preprocess the data, apply dimensionality reduction techniques, and implement clustering algorithms to segment customers with the goal of optimizing customer outreach for a mail order company.

**Evaluation**

Use the instructions in the project lesson to complete all steps in the project. Once you're done, review your work against the [project rubric](http://review.udacity.com/#!/projects/f5062c28-1399-403e-af18-0e5b0e4587fd/rubric). Check that all of your work and its documentation meets specifications on all rubric points – if you see room for improvement in any category for which you do not meet specifications, go back and make some revisions. Remember that your project will be evaluated by a Udacity reviewer, and it must meet specifications on all rubric points in order for the submission to pass.

**Submission**

Once you are ready to submit your project, collect the following files in a .zip archive:

1. A copy of the Identify\_Customer\_Segments.ipynb notebook with completed code cells and filled out Discussion sections. If you worked on the project from the workspace in the lesson, you can download a copy of the notebook by going to the workspace and navigating to the File -> Download as -> Notebook (.ipynb) option.
2. An HTML export of the above project notebook. This can be obtained by navigating to File -> Download as -> HTML (.html) in the notebook editor.

**PROJECT RUBRIC**

**Pre-processing**

|  |  |
| --- | --- |
| **CRITERIA** | **MEETS SPECIFICATIONS** |
| Have data values representing missing values been encoded consistently? | All missing values have been re-encoded in a consistent way as NaNs. |
| Has the quality of the data been assessed by looking at missing values by columns? | Columns with a large amount of missing values have been removed from the analysis. Patterns in missing values have been identified between other columns. |
| Has the quality of the data been assessed by looking at missing values by rows? | The data has been split into two parts based on how much data is missing from each row. The subsets have been compared to see if they are qualitatively different from one another. |
| Have categorical features been properly processed? | Categorical features have been explored and handled based on if they are binary or multi-level. |
| Have mixed-type features been properly processed? | Mixed-type features have been explored, resulting in re-engineered features. |
| Has the dataset been cleaned to include only relevant columns? | Dataset includes all original features with appropriate data types and re-engineered features. Features that are not formatted for further analysis have been excluded. |
| Has a function performing cleaning operations been created? | A function applying pre-processing operations has been created, so that the same steps can be applied to the general and customer demographics alike. |

**Feature Transformation**

|  |  |
| --- | --- |
| **CRITERIA** | **MEETS SPECIFICATIONS** |
| Has feature scaling been applied to the data? | Feature scaling has been properly applied to the demographics data. Imputation has been performed to remove remaining missing values. |
| Have an appropriate number of features been selected following dimensionality reduction? | Principal component analysis has been applied to the data to create transformed features. A variability analysis has been performed to justify a decision on the number of features to retain. |
| Have relationships between original features been investigated for the data? | Weights on at least three principal components are used to make inferences on correlations between original features of the data. General meanings are ascribed to principal components where applicable. |

**Clustering**

|  |  |
| --- | --- |
| **CRITERIA** | **MEETS SPECIFICATIONS** |
| Have an appropriate number of clusters been set? | Multiple cluster counts have been tested on the general demographics data, and the average point-centroid distances have been reported. A decision on the number of clusters to use is made and justified. |
| Have the customer demographics data been handled in a fashion consistent with the general demographics data? | Cleaning, feature transformation, dimensionality reduction, and clustering models are applied properly to the customer demographics data. |
| What customer segments are popular or unpopular with the mail-order sales company? | A comparison is made between the general population and customers to identify segments of the population that are central to the sales company's base as well as those that are not. |

**Suggestions to Make Your Project Stand Out!**

* Use the data dictionary to understand the meaning of variables in the dataset.
* Use plotting packages like matplotlib and seaborn as necessary to visualize trends in the data.
* Add additional cells to your notebook to organize your code in manageable pieces and make notes on your observations.