**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**](https://review.udacity.com/#!/rubrics/2538/view). Your project will be evaluated by an Udacity reviewer using this criteria, 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.