# Identify Customer Segments

## Introduction:

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. The goal is to look at relationships between demographics features, organize the population into clusters, and see how prevalent customers are in each of the segments obtained.

## Table of Contents:

This project contains the following file:

- `Identify Customer Segments.ipynb` : This python file has all the code that was used to put through the entire analysis.

In the Terminal or Command Prompt, navigate to the folder containing the project files, and then use the command `jupyter notebook finding\_donors.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.

## Summary of Results of Analysis:

Performed data preprocessing. Assessed missing values in rows and columns. Handled missing values based on the assessment. Rows and columns with lots of missing values were removed. Other missing values were replaced by the median value of the column. Selected and re-encoded features for the analysis. Re-encoded categorical features. Dropped irrelevant features from the analysis. Engineered mixed type features to a usable form.

Applied feature scaling so that principal components are not influenced by the natural differences in scale or features. Used Principal Component Analysis to perform dimensionality reduction. Interpreted principal components by analyzing weights of the features.

Used k-means to generate clusters in order to compare customer data to demographics data.

Clusters of customer data belongs to higher income, with medium to high affinity for socially minded, family minded and religious Personality and low affinity for dominant minded people.

Clusters of the general population belongs to Average to high share of 1-2 and 3-5 family homes. Single or couple smaller family. Not dominant minded.Financial topology is low and have low financial interest, high number of professional academic title holders in family. It belongs to lower income segment.

Looking at the bar plot in the notebook, we can clearly see that there isn't good match between the general population and the

customers population.

## Software:

This project uses the following software and Python libraries:

- [Python 2.7] (https://www.python.org/download/releases/2.7/)

- [NumPy] (http://www.numpy.org/)

- [Pandas] (http://pandas.pydata.org/)

- [scikit-learn] (http://scikit-learn.org/stable/)

- [matplotlib] (http://matplotlib.org/)

You will also need to have software installed to run and execute a [Jupyter Notebook] (http://ipython.org/notebook.html)

If you do not have Python installed yet, it is highly recommended that you install the [Anaconda] (http://continuum.io/downloads) distribution of Python, which already has the above packages and more included. Make sure that you select the Python 2.7 installer and not the Python 3.x installer.

