Stockholm School of Economics
Center for Data Analytics

Joel Smedberg
Rickard Sandberg

Assignment 1 – Analyzing Market Survey Data

Formal Requirements

Please read the formal requirements *carefully*. Points may be docked if your hand-in does not conform to these specifications. Submit your answers on Canvas in **three (3)** parts:

1. A PDF with your written answers to the assignment.
    a. The PDF should be named <Your Student Number>_<Partner_number> _assignment2.pdf (e.g. 23456_98765_assignment2.pdf)
    b. The PDF should consist of a maximum of **3 pages** of written text.
    c. Use a reasonable font, standard margins, and line spacing - size 12. Be kind to our eyes.
    d. **All figures/tables/graphs/plots you need to support your answers should be called figures and numbered sequentially in an Appendix** (e.g. Figure 1, Figure 2, etc.). The appendix can be of any length.
    e. If you infer things in your answer from a figure in your appendix, you should explicitly refer to it in the text, e.g. *"The quick brown fox jumps over the lazy dog (see Figure 3)"*.
    f. Only add the figures you need to support your answers. Less is more.
3. A Python (.py) or Jupyter Notebook (.ipynb) file containing your code for generating the JSON file. Name it <Your Student Number>_<partners number>_assignment2.<Your Chosen File Format> (e.g. 23456_98765_assignment2.py, 23456_98765_assignment2.ipynb)

To the extent possible, segment your code file into sections that correspond to: Setup, Question 1, Question 2, and Question 3. This will make it easier to inspect your code. You may use any external library as long as it's installable by conda or pip. You may work in teams of two, but both members should understand all aspects of what is handed in. Groups are graded as one unit but can, on very rare occasions, be graded individually at the course director's discretion.

You may discuss methodology with other teams, but you should create the content of the submission from start to finish.

The evaluation of your assignment will consider, among other things:

- Written answers: Justification of and understanding of the chosen features, model, and evaluation. Do visual aids explain your findings and/or your case?

- Python code file:
    - Code quality and readability as evident from variable/function names, control flow, and comments (Please only comment in English)

Stockholm School of Economics
Center for Data Analytics

Joel Smedberg
Rickard Sandberg

Introduction

Marketing analysts often investigate differences between groups of people by asking various questions. Do men or women subscribe to our service at a higher rate? Which demographic segment can best afford our product? Does the product appeal more to homeowners or renters? The answers help us understand the market, target customers effectively, and evaluate outcomes of marketing activities such as promotions.

Your task in this assignment is to conduct basic clustering and principal component analysis on Marketing Survey data for Company XYZ based on 300 respondents.

Stockholm School of Economics
Center for Data Analytics

Joel Smedberg
Rickard Sandberg

Data Description

Download the *SegmentData.xlsx* from the course webpage. This data set contains the six features:

1. Age
2. Gender
3. Income
4. Number of children
5. Own their homes: “OwnNo” = No, “OwnYes” = Yes
6. Subscribing to offered service: “SubNo” = No, “SubYes” = Yes

The sample size is $n = 300$, and the data are further divided into the four segments:

S1. Suburb mix, segment size $n_1 = 100$
S2. Urban hip, segment size $n_2 = 50$
S3. Travelers, segment size $n_3 = 80$
S4. Moving up, segment size $n_4 = 70$

Question 1: K-means Clustering

i. Exploratory Data Analysis (EDA): Report summary statistics for all variables. Plot a histogram for each variable. Identify any outliers. Provide a correlation matrix for relevant variables. Briefly summarize your EDA.

ii. Apply the K-means clustering algorithm to these data with K=4. Report the sizes of each cluster and the mean values for all features in each cluster. Comment briefly on any differences in cluster means for the features.

iii. Visualizing the cluster solution: Plot the cluster solution with respect to the variables Age and Income. Include the cluster centroids in the plot. Briefly comment on the results.

iv. Continuing visualization of the cluster solution: Provide box plots for Income across the different clusters. Also, provide box plots for Age across the different clusters. Comment briefly on these results.

Joel Smedberg
Rickard Sandberg

v. Test the hypothesis that the mean Income is the same for all clusters. Also, test the hypothesis that the mean Age is the same for all clusters. Hint: use ANOVA testing. What are your conclusions?

vi. Determine the "optimal" number of clusters. Provide a brief commentary on your answer.

vii. With respect to your cluster solution, how many of the respondents are correctly classified regarding their segmentation information?

Question 2: PCA

i. Use the same data as in Question 1. Extract two principal components using standardized data. State the extracted principal components and comment briefly on their loadings.

ii. Calculate the proportion of variance extracted by these two principal components.

iii. Plot the principal component scores with the cluster membership as determined in Question 1.

iv. In this plot, identify the individual with the highest score for the first principal component, and compare the features of this individual with the individual having the highest score on the second principal component. Briefly comment on any difference.