

Group Project:

This project is based on a data set called ERIMdata.xlsx that includes about 3,000 households in two midwestern cities in the United States. The data contain demographic information such as household incomes, number of household members, education levels as well as information on the purchases of retail products such as frozen dinners and yogurt. The data were collected by a marketing research firm, AC Nielsen.

Your assignment is to conduct business analysis on the data set using **1) data visualizations** (e.g., charts and tables) and **2) either prediction** (e.g., linear regression) **or classification** (e.g., logistic regression, naïve Bayes, or kNN) **techniques**. The project deliverables from each student will include a **one-page** written report that highlights the key findings and an appendix with additional information (e.g., relevant R code and output). Think of the following CRISP-DM phases and guidelines to plan for your analysis.

Business understanding: Describe the business opportunities that the data present and formulate relevant business questions.

Data understanding: Explore the data set with descriptive analytics tools and use the information to help determine whether or not data wrangling steps are needed for each variable. Examine the possibility of different analysis techniques. Keep in mind the business opportunities and questions formulated in the first phase.

Data preparation: Determine and perform the necessary data wrangling and preparation tasks based on the decision made during the business and data understanding phases.

Modeling: Implement the appropriate techniques and variables, explain the rationale for your selection(s), and present relevant analysis results and interpretation. Experiment with multiple models and identify the best model(s).

Evaluation: Refocus on the business objectives of the project. Evaluate whether the model(s) have properly achieved the business objectives outlined during the business understanding phase. Formulate actionable recommendations based on the findings.

Deployment: Communicate the findings and relevant business insights with a written report incorporate appropriate statistical information and visuals. The main focus of the written report should be placed on providing actionable business recommendations for a managerial and non-technical audience.

DATA DICTIONARY: Description of variables in ERIMData

Variable	Description
HH_ID	The household's identification number
ResType	Types of residence: 1 for Apartment, 2 for Condo, 3 for Single Family, 4 for Multiple Family, 5 for Mobile, and 6 for Other.
ResStatus	Residence status: 1 for owned home, 2 for rented, and 3 for other.
HHInc	The average annual income of a household; there are 14 categories for this variable (1 being the lowest income category).
HHNbr	The number of members in the household.
MWrkHrs	The average hours worked each week by the male adult/household member.
MEdu	Education level of the male adult: values less than 9 imply education levels prior to a college degree, 9 for graduated from college, 10 for attended graduate school, and 11 for post-graduate degree.
FWrkHrs	The average hours worked each week by the female adult/household member.
FEdu	Education level of the female adult. See MEdu for detail.
FBirth	The birth year of the female adult/household member.
MBirth	The birth year of the male adult/household member.
Cable	Whether or not the household has cable TV; 1 if yes, 0 otherwise.
Cats	The number of cats in the household.
Dogs	The number of dogs in the household.
YogExp	A household's yogurt expenditures (in \$)
DinExp	A household's frozen dinner expenditures (in \$)