Data – The data set chosen contains 8,993 responses from a questionnaire given to patrons of a San Francisco shopping center to collect income and demographic data. The data set was sourced from the arules R library and each row in the data set contains answers to 14 variables that each represent a question on the questionnaire, they are as follows:

Variable	Number of Options on Questionnaire	Type
Income	9	Ordinal
Sex	2	Categorical
Marital status	5	Categorical
Age	7	Ordinal
Education	6	Ordinal
Occupation	9	Categorical
Years in the bay area	5	Ordinal
Dual incomes	3	Categorical
Number in household	9	Ordinal
Number of children	9	Ordinal
Householder status	3	Categorical
Type of home	5	Categorical
Ethnic classification	8	Categorical
Language in home	3	Categorical

Within this data set, which can be accessed using the R command *data(Income)*, it is not uncommon for some variables to have a high degree of correlation. As an example, if a respondent states that their type of home is an apartment, their response to householder status will often be rent (confidence of 0.91). With this being true simply running association rules on the data may not be interesting without a defined problem.

Problem – The information contained in this data set can be used in several different ways to draw insights from the population. Assuming the sample of data is large and diverse enough, it may be acceptable to extrapolate the results and apply them to the entire population of San Francisco. Therefore, using association rules on the income data set may be used to generalize a resident’s income based on other demographic attributes. Association rules will be used to determine the strongest relationships between the different responses to questions in the San Francisco questionnaire that can predict with reasonable accuracy whether a respondent has an income greater than $40,000.

Techniques – Since the responses to all the questions in the survey are either categorical or ordinal, it is possible to carry out an association rule analysis using the Apriori algorithm provided in the arules R package. The Apriori algorithm can only be used on transactional databases. The algorithm was run with a minimum support of 10% and a minimum confidence of 80% to capture a set of rules that contain strong associations between the variables. To target association rules that will predict income patterns it is necessary to set the right-hand side constraint equal to a realistic categorical income figure. From the data this was selected to be $40,000+.

Results – After running the Aprior algorithm on the income data set and setting the right hand constraint to an income of greater than $40,000, here are the strongest association rules ranked by standardized lift.

LHS	Support	Confidence	Lift	£
education=college graduate, dual incomes=yes	0.08566027	0.8249300	2.184984	0.29971989
occupation=professional/managerial, dual incomes=yes	0.11227458	0.8160677	2.161510	0.26427061
dual incomes=yes, householder status=own	0.12609075	0.8156162	2.160315	0.26246472

Conclusions – After running the algorithm and returning the top three association rules with a right hand constraint representative of having an income greater than $40,000 the strongest association rule states that respondents who were college graduates with a dual income had incomes of over $40,000 82.5% of the time. This means that of all the respondents who said they were college graduates with dual incomes, 82.5% of those respondents had incomes over $40,000. The application of the Apriori algorithm to this income data set proves, at least at a preliminary level, that such demographic variables as listed above, can accurately predict a resident’s income category and that collecting more observations as well as more specific data would be appropriate for many different private or public sector applications.

Sources –

Hastie, Trevor, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer, 2017.