Exercise 3 - Introduction to Data Science

Exercise 3.1

(Association Rules) The data in for this assignment was generated using the method by Agrawal and Srikant (random.patterns) to simulate transactions (random.transactions) which contains correlated items.

10,000 transactions occurred with 100 items to choose from. The average length of the transactions is 10 items.

Note: You will need to load the arules and arulesViz R packages to complete this assignment.

- Import the AssociationRules.csv transaction data file.
- Create a frequent item plot, and a frequent item table.
    a. Determine the most frequent item bought in the store.
    b. How many items were bought in the largest transaction?
- Mine the Association rules with a minimum Support of 1% and a minimum Confidence of 0%.
    c. How many rules appear in the data?
    d. How many rules are observed when the minimum confidence is 50%.
    e. Explain how the specified confidence impacts the number of rules.
- Create a scatter plot comparing the parameters support and confidence on the axis, and lift with shading.
    f. Identify the positioning of the interesting rules.
- Compare support and lift.
    g. Create a scatter plot measuring support vs. lift; record your observations.
    h. Where are the rules located that would be considered interesting and useful?
    i. One downside to the Apriori algorithm, is that extraneous rules can be generated that are not particularly useful. Identify where these rules are located on the graph. Explain the relationship between the expected observation of these itemsets and the actual observation of the itemsets.
    j. Using the interaction tool for a scatter plot, identify 3 rules that appear in at least 10% of the transactions by coincidence.
- Identify the most interesting rules by extracting the rules in which the Confidence is >0.8. Observe the output of the data table for the most interesting rules.
    k. Sort the rules stating the highest lift first. Provide the 10 rules with the lowest lift. Do they appear to be coincidental? Why or why not?
- Create a Matrix-based visualization of two measures with colored squares. The two measures should compare confidence and lift (have recorded = FALSE). Note that 4 interesting rules stand out on the graph.
    l. Identify these rules and explain their appearance.
    m. What can you infer about rules represented by a dark blue color?
- Extract the three rules with the highest lift.
    n. Record the Rules. Explain why these rules vary from the rules in Step 3.
    o. Create a Graph-based visualization with items and rules as vertices.

p. Based on your observations, explain how you would expect association rules to relate to order (i.e. the number of items contained in the rule).

- Create a training set from the first 8,000 transactions. Create a testing set from the last 2,000 transactions. Run the algorithm on each dataset. Compare the results.
    q. Justify that the relationships we see are not just an artifact of the data we started with.
    r. Can we conclude that the association rules we found are actually true in the population we are studying?

Exercise 3.2

(Linear Regression) In this assignment, you will analyze linear regression models on different categories of data about average households in the United States.

Gather and Prepare Data: You will be using the same dataset (zcta) in exercise 2.

- Make sure to remove all meanhouseholdincome duplicate rows of data (only females' records should be in the dataset).
- Remove the columns zcta and sex.
- Remove outliers by creating subsets of the original data so that:

$$8 < \text{meaneducation} < 18$$

$$10{,}000 < \text{meanhouseholdincome} < 200{,}000$$

$$0 < \text{meanemployment} < 3$$

$$20 < \text{meanage} < 60$$

- Create a variable called log_income = log10(meanhouseholdincome).
- Rename the columns meanage, meaneducation, and meanemployment as age, education, and employment, respectively.

Linear Regression Analysis: You will be analyzing this data with income as the dependent variable and the other columns as independent variables.

a. Create a scatter plot showing the effect age has on log_income and paste it here. Do you see any linear relationship between the two variables?
b. Create a linear regression model between log_income and age. What is the interpretation of the t-value? What kind of t-value would indicate a significant coefficient?
c. What is the interpretation of the R-squared value? What kind of R-squared value would indicate a good fit?
d. What is the interpretation of the F-statistic? What kind of F-statistic indicates a strong linear regression model?
e. View a detailed summary of the previous model. What is the R-squared value? Does this suggest that the model is a good fit? Why?
f. Create a scatter plot showing the effect education has on log_income. Do you see any linear relationship between the two variables?

g. Analyze a detailed summary of a linear regression model between log_income and education. What is the R-squared value? Is the model a good fit? Is it better than the previous model?
h. Analyze a detailed summary of a linear regression model between the dependent variable log_income, and the independent variables age, education, and employment. Is this model a good fit? Why? What conclusions can be made about the different independent variables?
i. Based on the coefficients of the multiple regression model, by what percentage would income increase/decrease for every unit of education completed, while all other independent variables remained constant?
j. Create a graph that contains a y = x line and uses the multiple regression model to plot the predicted data points against the actual data points of the training set.
k. How well does the model predict across the various income ranges?