Part 3 – Work to be completed with your group (70 points total)

Work with your group to complete the following problems from the 1st edition of Data Mining for Business Analytics Concepts, Techniques and Applications in R.

Chapter 2 (10 Points)

Problems 2.1c, 2.1d, 2.6, 2.7, 2.8

Data Exploration (20 points)

1. Import the “Universities2019.csv” file to R Studio and write R code to produce summary statistics of all numeric variables. For each numeric variable, include the following.

Variable name

Minimum

Maximum

Median

Mean

Standard Deviation

Number of Missing Values

Paste a screenshot of your summary statistics below.

1. Explore the dataset and identify three issues with values found in the dataset. An issue is something that is not logically possible based on your knowledge of the domain and your understanding of the definition of each variable.
2. Create a new column called “AdmitRate” and add it to the data set as the last column. Admit rate should be calculated by dividing “Admits” by “Applicants” to get the percent of students admitted. Paste a screen shot of the first 10 rows of the dataset with the new column below.
3. Create a histogram of the Applicants variable. Paste a screen shot below. Interpret the result
4. Create a histogram of AdmitRate. Paste a screenshot below. Interpret the result.
5. Create a side by side Box Plot for Enrolled using Type as the by variable. Interpret the result.
6. Create a scatter plot of AdmitRate and GraduationRate. Interpret the result.
7. Create a color-coded Tableau visualization with the count of nthe umber of schools by state. Include a filter to filter results by Type of University. The results should be displayed on a map of the US and each state should be shaded to indicate the number of schools in the state.
8. Provide 3 insights on the results of a – g. An insight should be an observation made based on your understanding of the value of individual variables or the association between variables.
9. Paste a copy of your R code below.

Association Rules (15 points)

1. Import the Cosmetics.csv file into R Studio and browse the dataset to understand it. What does each row represent?
2. Perform an association rules analysis with a minimum support of 3%, a minimum threshold confidence of 25%, and a maximum rule length of 2 items. Paste a screenshot of the top 10 rules in descending order by lift below.
3. Explain the first rule in b.
4. Perform an association rules analysis with a minimum support of 3%, a minimum threshold confidence of 25%, and a maximum rule length of 3 items. Paste a screenshot of the top 10 rules in descending order by lift below.
5. Explain the first rule in d.
6. Provide the owner of the store selling cosmetics with a recommendation on how the rules could be used to improve his or her business. Be specific and actionable with your recommendation. In other words, your recommendation should be something the store owner can immediately implement to improve his or her business.
7. Paste a copy of the R code below.

Recommender System (15 Points)

Identifying Course Combinations. (modified version of problems 14.2 14.3). Consider the data in the file CourseTopics.csv. These data are for purchases of online statistics courses in Statistics.com. Each row represents the courses attended by a single customer. The firm wishes to assess alternative sequencings and bundling of courses.

a. Use association rules to analyze the data, and interpret the top 2 rules.

b. We want to provide a course recommendation to a student who purchased the Regression course. Apply item-based collaborative filtering to the data. What are the results?

c. We want to provide a course recommendation to user 33 (row 33). Apply user-based collaborative filtering to the data. Who are the top 3 nearest users? What product would you recommend?

d. Calculate a user-based correlation matrix for all users in the dataset. Using the user-based correlation matrix, who are the top 3 correlated users with user 33 (row 33)? Does this differ you’re your answer in part c? Why or why not?

Each group should submit a single Word document containing answers to the questions listed above.

Presentation to Class (10 points)

The day the assignment is due, be prepared to present any one of the answers above to the class. You do ***not*** need to prepare a PowerPoint with your answers and should instead be prepared to use your Microsoft Word solution file, R and/or Tableau to help answer the question. I will ask questions to each group that presents and you will be graded on both the answer to the homework question and your answers to my questions.

You should also be prepared to present your answers to any of the questions contained in the homework.

Required Datasets

Universities2022.csv

Cosmetics.csv

Coursetopics.csv