# Prep Exercise (PE09) Associated Rules Mining

### General Instructions

1. For this exercise you will answer all of the questions in this document and turn it in to Blackboard.
2. Before you get started make sure to read Chapter 17 of *An Introduction to Data Science* and execute the code throughout the chapter to gain familiarity.
3. Getting Started:
   1. Up until now we have done relatively simple data exploration and modelling, here is where we enter the territory of machine learning. This week we will be examining associated rules mining, a form of machine learning that discovers relationships between variables within large or small datasets. A practical example of associated rules mining is seen in the way grocery stores operate, maximizing their profits by organizing their stores in a manner that maximizes the probability that a person who purchases “x” will also purchase “y”. However, in this PE and homework we will be examining associations within a dataset containing titanic data.
   2. As usual we will use this Prep Ex to set you up for the homework exercises and test your knowledge of materials within the chapter reading. Let’s begin…

# IST 687, Standard Homework Heading

#

# Student name: Thadhani Hitesh Chandrakumar

# Homework number: PE09

# Date due: 31st Oct 2019 11:59 PM

#

# Attribution statement: (choose the statements that are true)

# 1. I did this work by myself, with help from the book and the professor

# 2. I did this work with help from the book and the professor and these Internet sources: www.en.wikipedia.org, www.machinelearningmastery.com

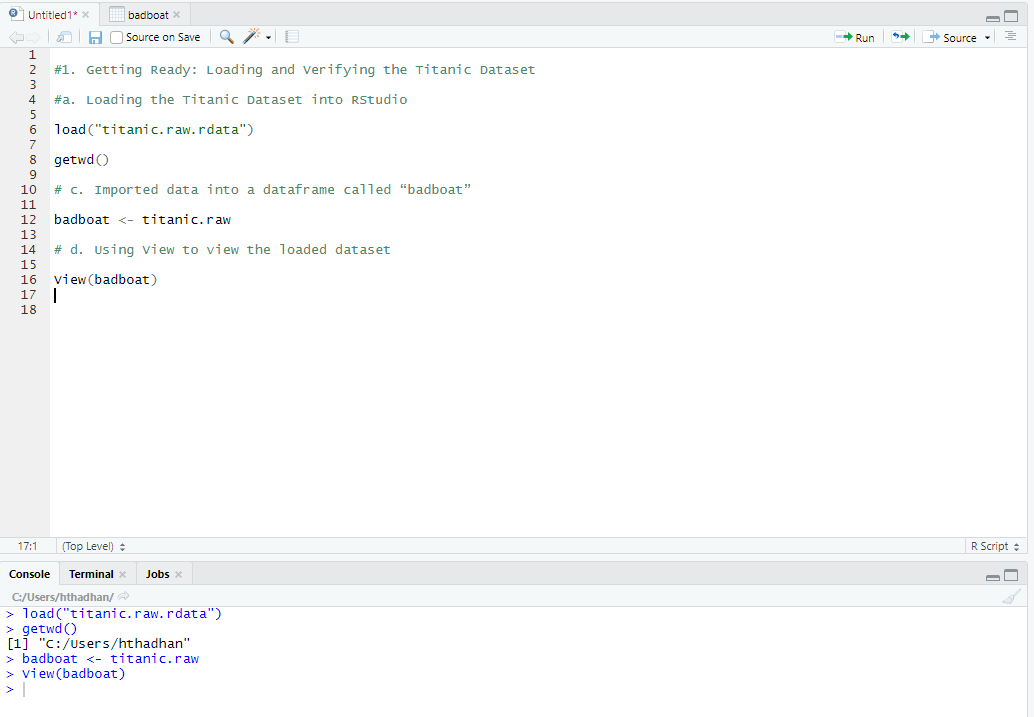
# 3. I did this work with coaching from <Name of another student> but did not cut and paste any code

### Prep Exercise

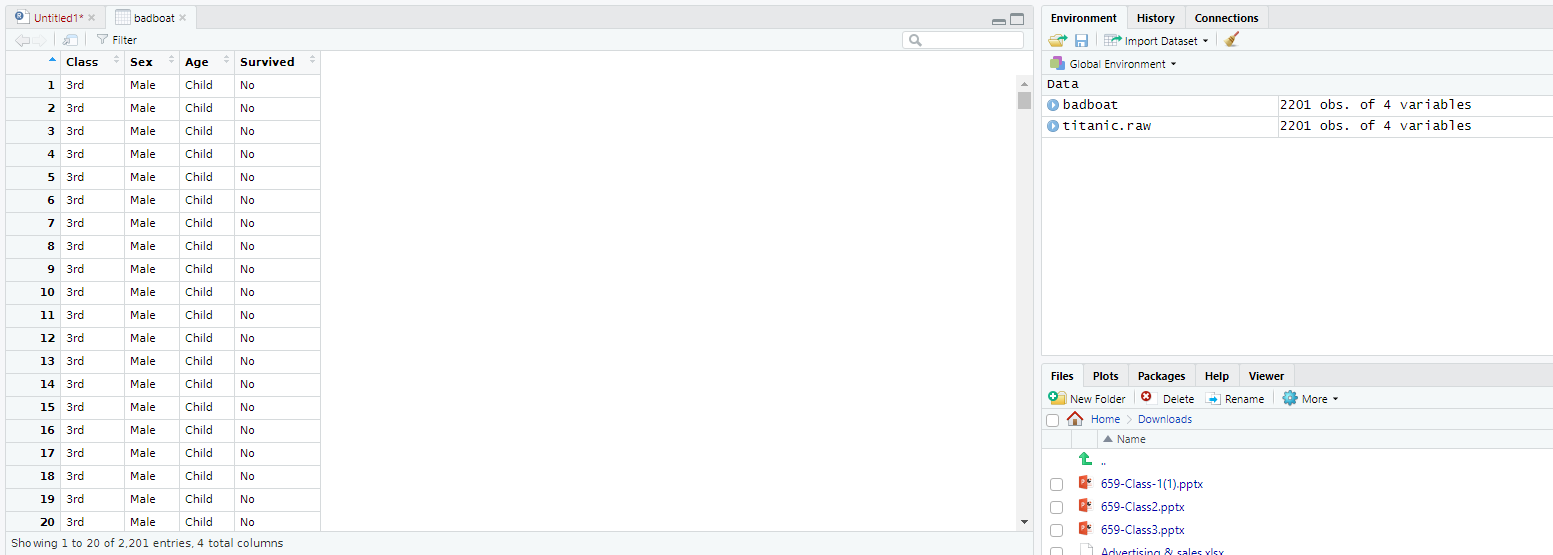
1. **Getting Ready: Loading and Verifying the Titanic Dataset**
   1. An R dataset containing the titanic data is available on the Blackboard site. Download it to your computer and use the *load()* command to bring it into your RStudio environment.
   2. If you are having trouble using the *load()* command, make sure you downloaded the titanic data to your working directory. You can check what is your working directory in R using the code below:

**getwd()**

* 1. Load the imported data into a dataframe called “badboat”.



* 1. Run the *View()* command to verify that the dataframe has been loaded correctly and place a screenshot of the dataset below (do not worry about screenshotting the entire dataframe, a portion will do).



1. **Understanding Terminology that will be used in this PE and HW.**
   1. In a paragraph or two explain the concept of a sparse matrix and how to identify whether or not data is a sparse matrix.

Sparse Matrix is a matrix where most of the elements are zero. Sparsity which is defined as number of zero entry elements divided by total elements of the matrix is a score used to quantify a matrix as sparse or not. In a matrix when the sparsity is more than 0.5 i.e. 50% of the elements are zero then it can be considered as a sparse matrix. In practice most of the large matrices are sparse matrices having large number of zero elements. They are mainly used where there are too large matrices difficult to handle due to majority of zero entries and using Sparse matrix can be used to compress them easily.

* 1. In a few sentences explain what a dense matrix is and its defining feature.

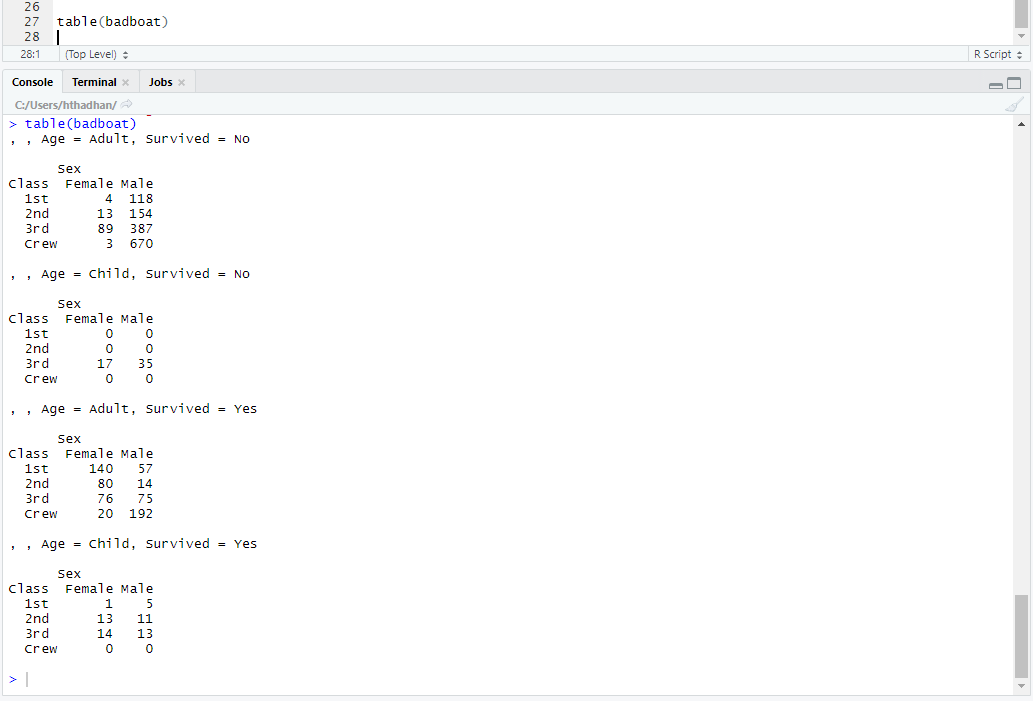
Dense Matrix is a matrix in which most of the elements are not zero as compared to Sparse matrix where most of the elements are zero. Dense matrices have density which is defined as

Density = 1 – Sparsity of the matrix

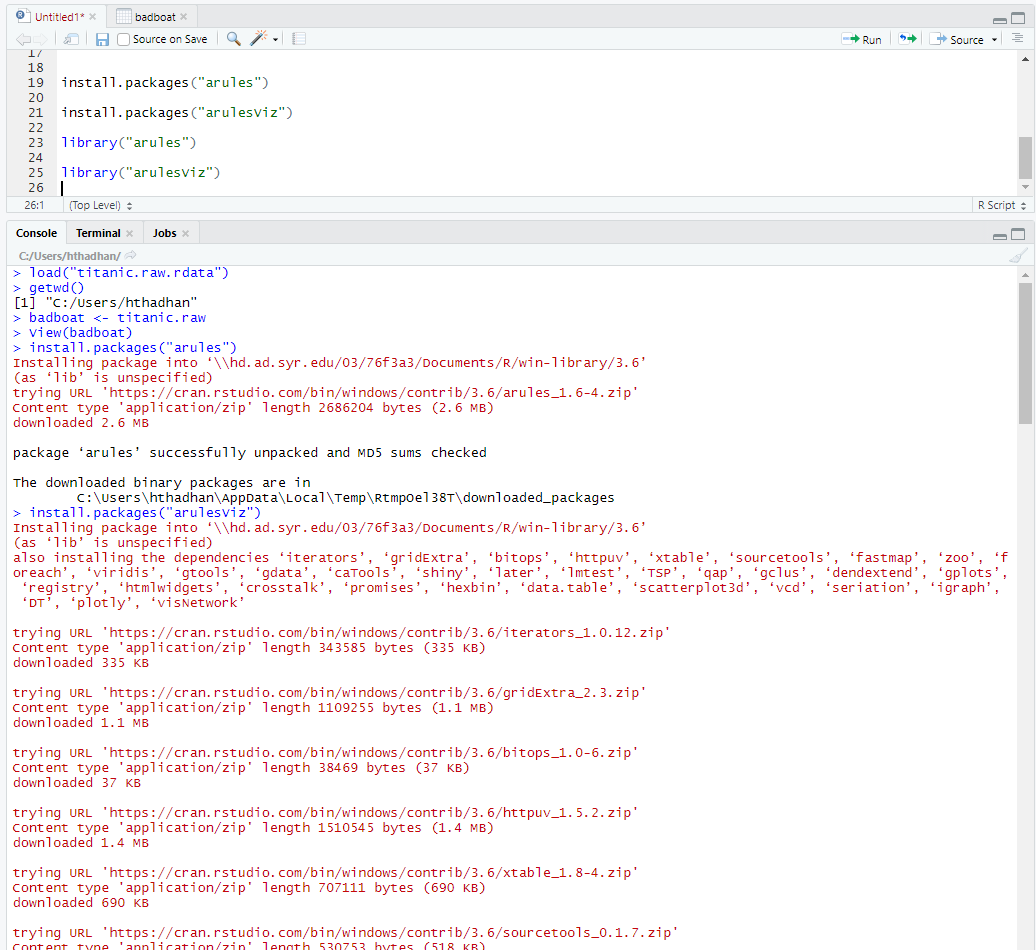
This means that there only a few elements which are zero and the matrix is dense here because the density is very high majority of the elements are non-zero.

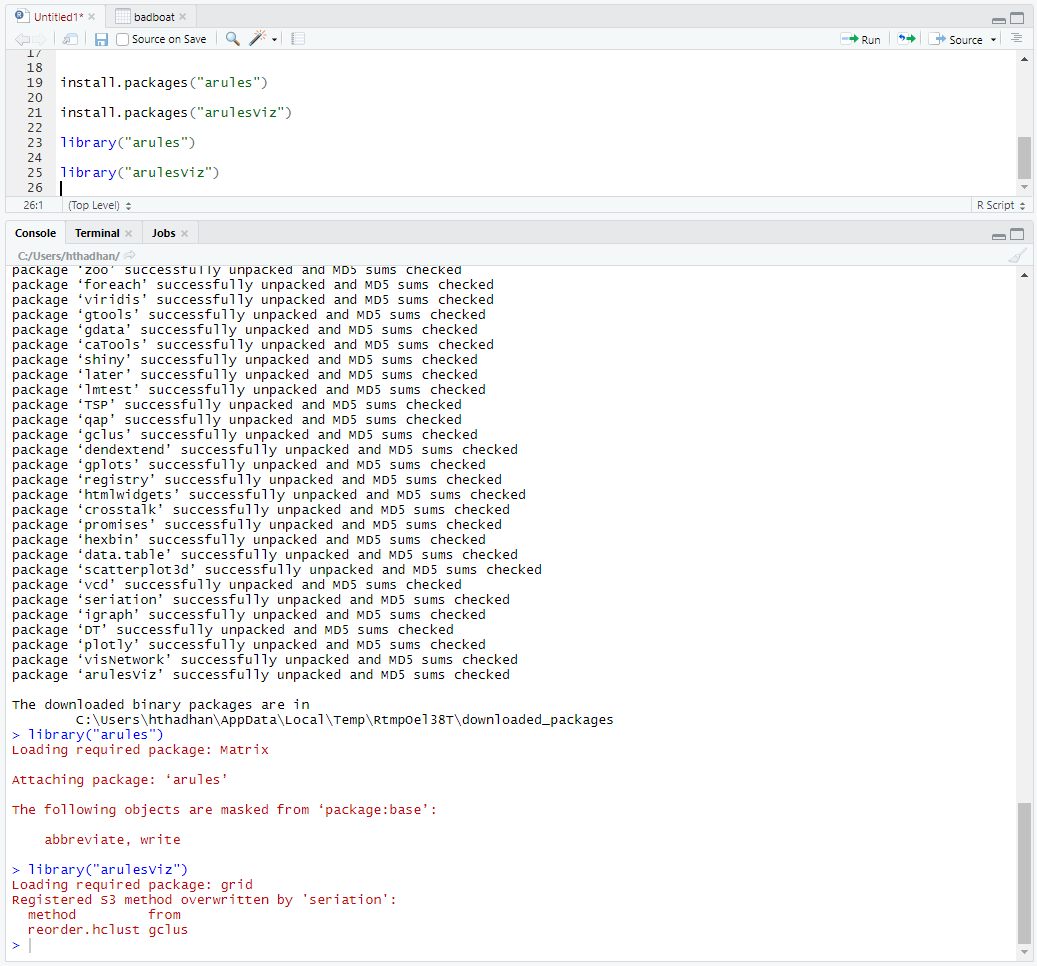
* 1. Explain contingency tables and how you would go about creating one in RStudio (if needed, you can google “contingency table in r”, there are many resources on the web). How would you go about reading /understanding a contingency table?

Contingency tables are a way to redraw a large table into a format which allows the users to get overall summary of the whole data in a better manner. We can use table() function to create a contingency table in R using vector, data frame or matrix. We can also create a contingency table from specific rows, columns of a data frame or a matrix. Here in the screenshot I have created a contingency table from badboat data frame which has titanic data in it. This gives a very concise information about the survival of the passenger and how it is related to all other categorical variables in short it summarizes the relationship between passenger’s gender, survival, age group and the class of the travel. This tells the exact number/ frequency of the variables and we can easily read it to understand what the data is briefly trying to convey without having to perform any operation on the data.



1. **Loading Necessary Packages.**
   1. In the homework portion of this week’s lesson you will need two packages; *arules* and *arulesViz*.
   2. Install both packages and check/update them using the *library()* command to ensure the packages exist. Paste a screenshot of the code and respective output below.





1. **List any additional resources that you used here.**
2. **Be sure to save your R file as this will become the starting code for your homework.**

***You must submit all Prep Exercises to blackboard prior to the deadline specified for each assignment.*** PE assignments are due on the evening prior to the lecture class. Late PE assignments will not be accepted for credit.

*Materials provided by the instructors of this course are copyrighted 2018 by Jeffrey Saltz and Jeffrey Stanton, or by the respective instructor that produced them. These materials are provided to enrolled students for their exclusive use during the semester of instruction. Uploading instructor-produced materials to Internet sites (e.g., Coursehero) without permission is considered a breach of professional ethics and will be reported as a possible academic integrity violation.*