**Charotar University of Science and Technology [CHARUSAT]**

**Chandubhai S. Patel Institute of Technology [CSPIT]**

**U & P U. Patel Department of Computer Engineering**

**Lab Manual**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Subject code | : | CE-350 | Semester | : | 6 | Academic Year | : | 2021 |
| Subject name | : | Data Warehouse and Data Mining | | | | | | |

**Practical – 3**

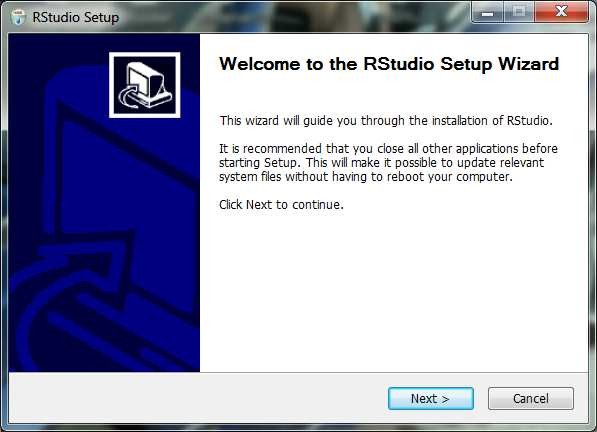
**Aim: Generate Association Rules: [R Programming, Python]**

* **Prerequisite**

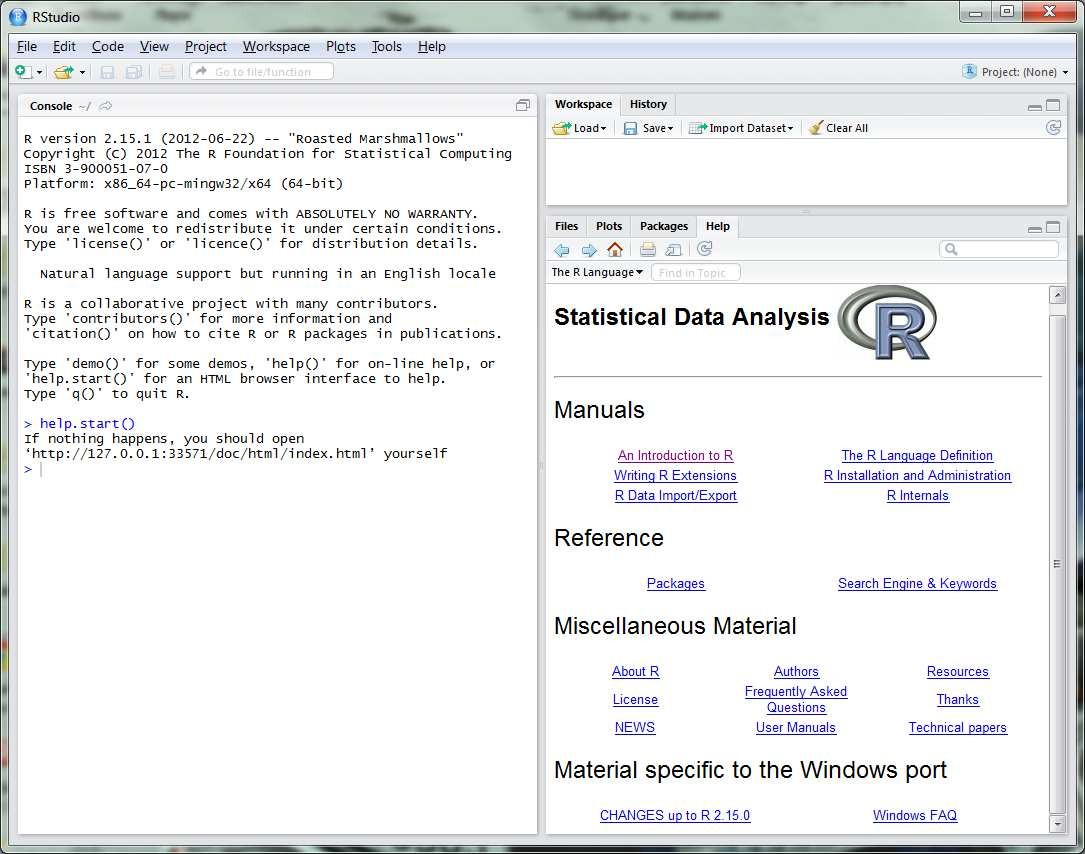
S/W: - R studio, Jupyter Notebook

Little bit knowledge about the python programming, R programming and some theory about that association rules.

* **Step to install R studio:**
* Download the latest version of RStudio IDE for your Windows platform from <http://rstudio.org/download/desktop>*(At the time of writing the latest available version of RStudio is v0.96)*
* Start the installation and follow the steps required by the Setup Wizard



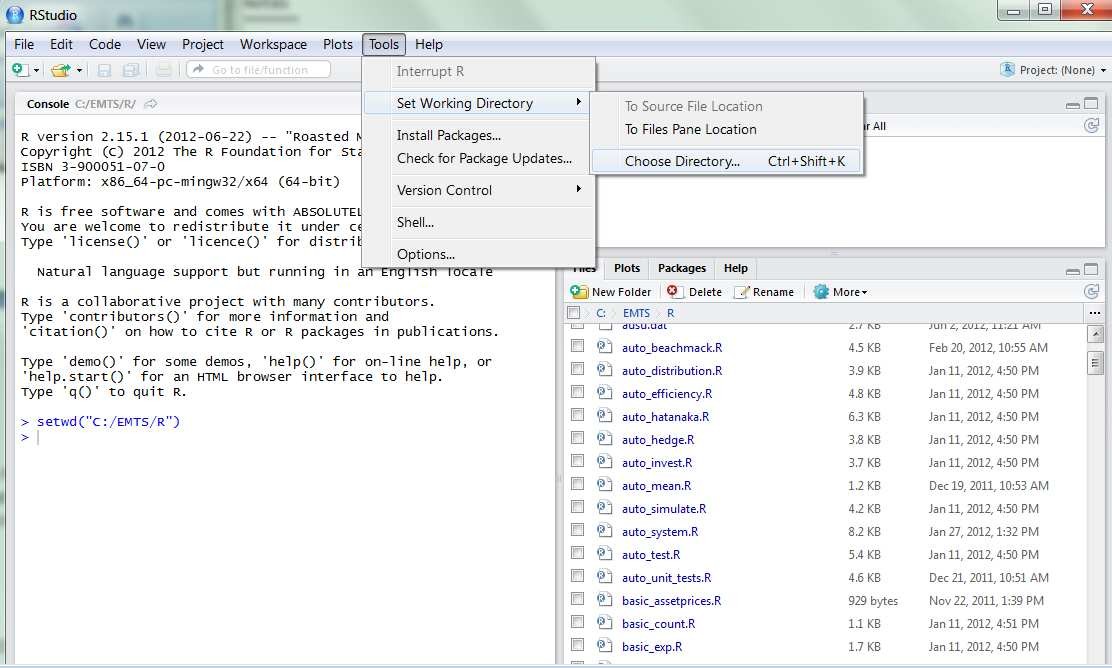
* Once completed, launch RStudio IDE from Start -> All Programs-> RStudio -> RStudio.exe or from your custom installation directory. The default installation directory for RStudio IDE is "C:\Program Files\RStudio\bin\rstudio.exe"
* Type help.start() at the RStudio prompt and press Enter. If you can see the following screen then you have successfully installed and configured RStudio IDE to run with R.



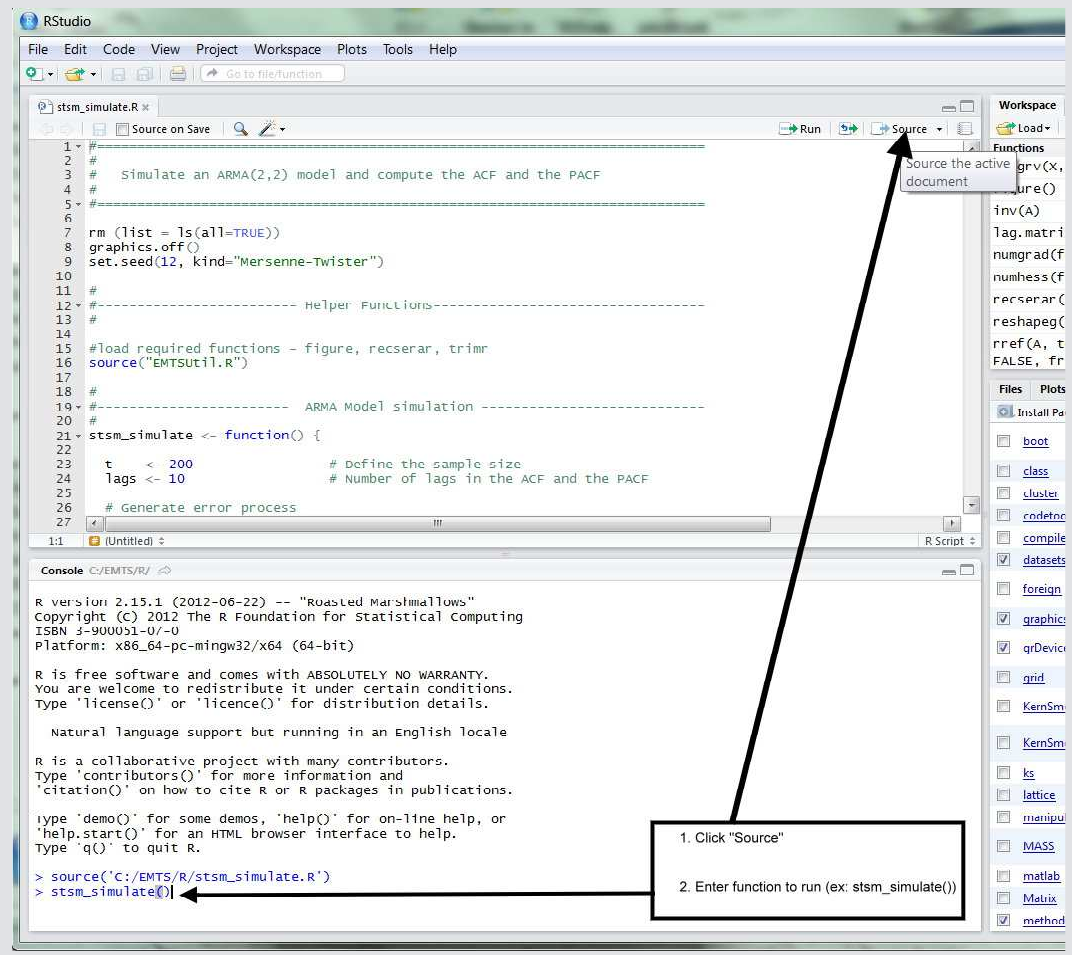
* Extract/Unzip all R-Scripts to a folder location you want.
* All R-Scripts make use of the required utility functions implemented by EMTSUtil.R. Make sure that this file exists in your extracted directory and make sure you have installed the required packages/library for the text book exercises. These can be found in section below (*see Installing additional packages with RStudio IDE)*.
* Set your working directory to your R-scripts using the command,

setwd(dir) For example, setwd("C:\EMTS\R")

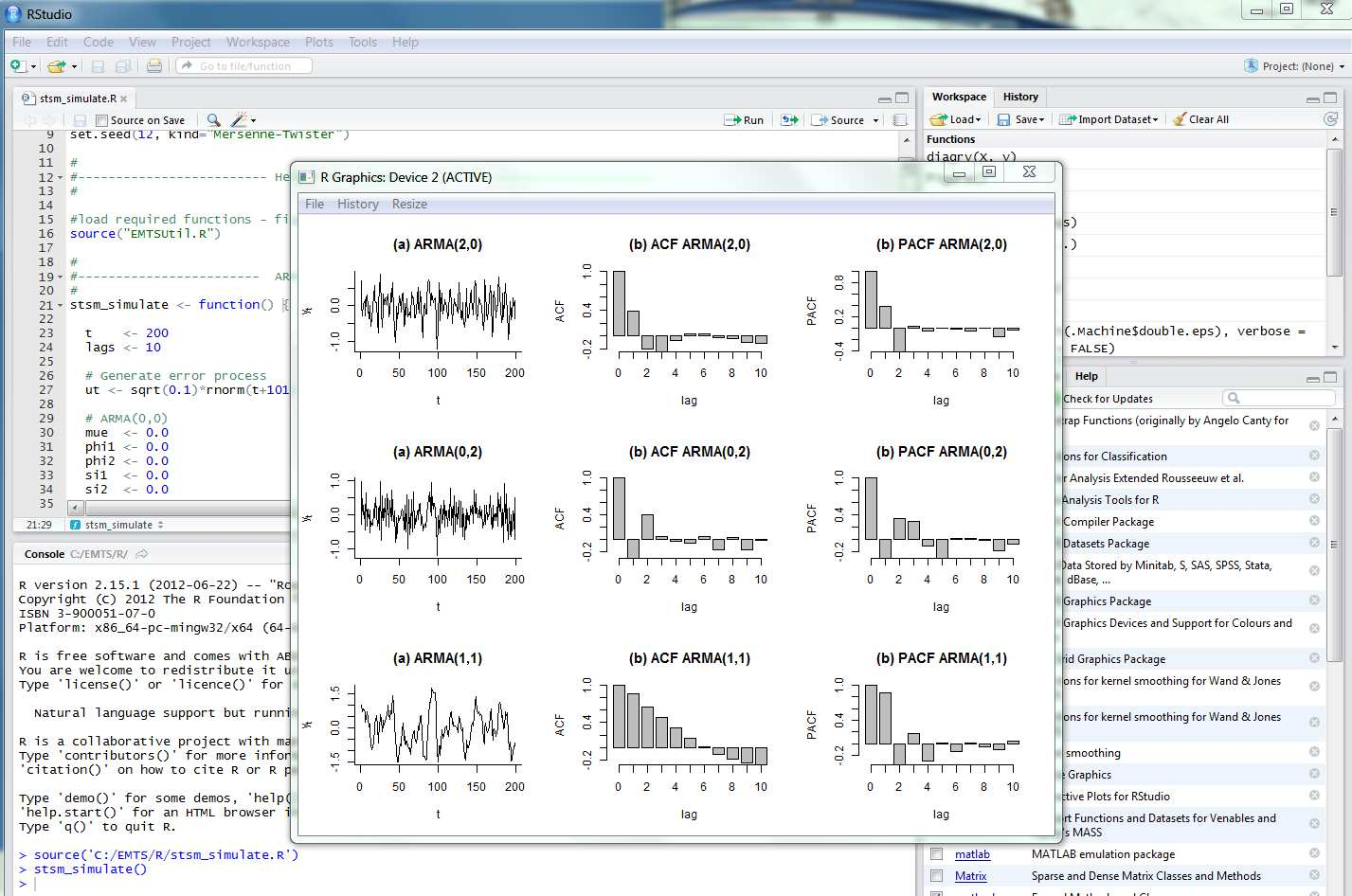
Alternatively you can use RStudio's Tools Menu: Tools -> Set Working Directory -> Choose Directory...



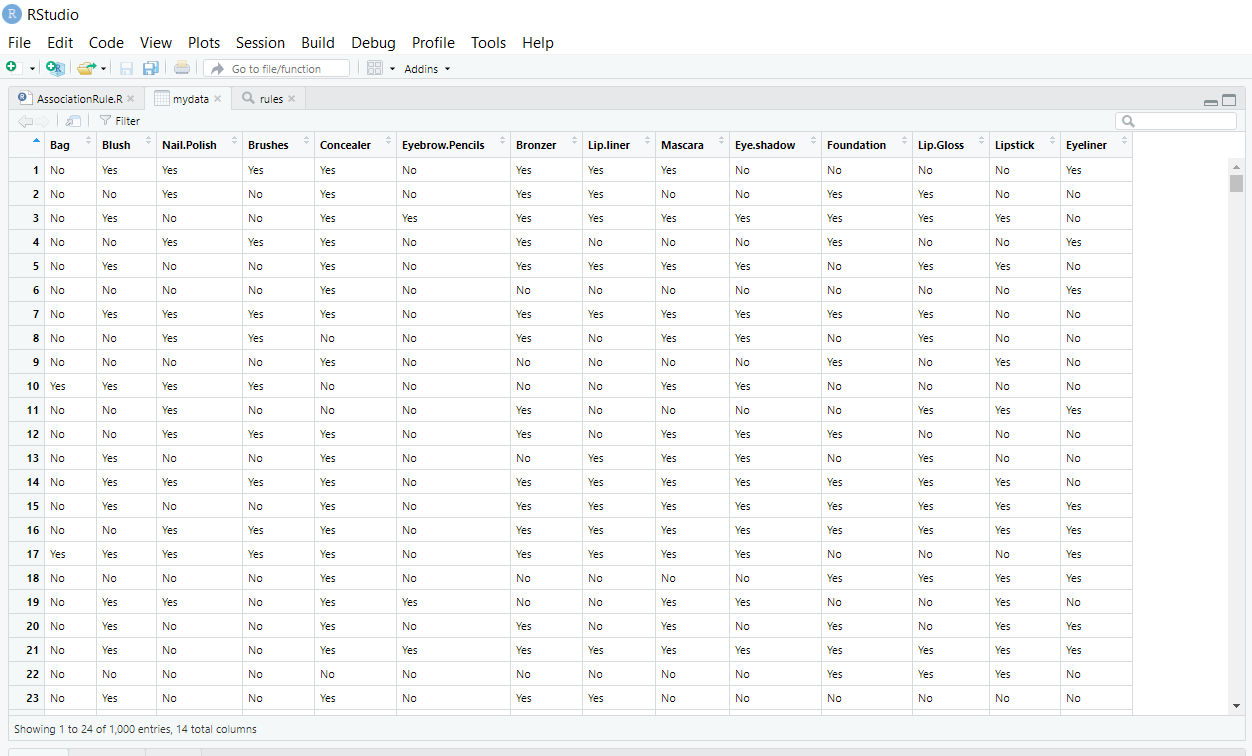
* To run a script, open the script in RStudio's script editor and choose "Source" from the menu or the following command at the Console prompt,

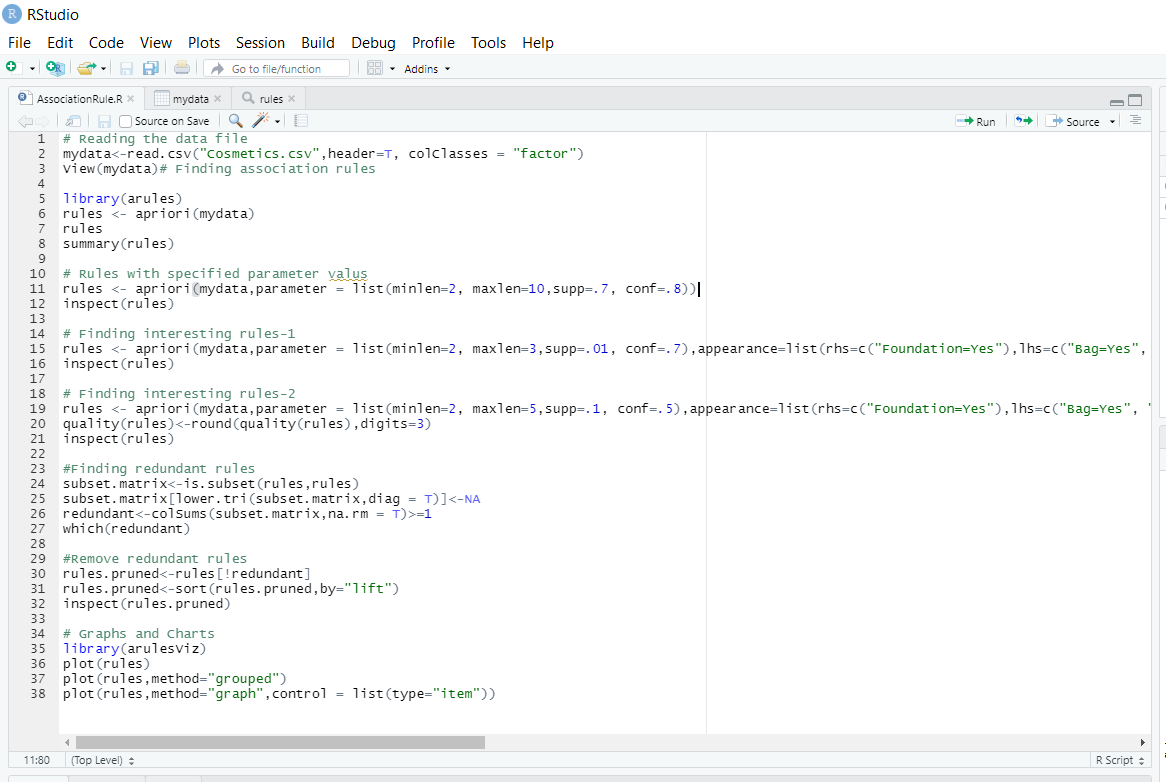


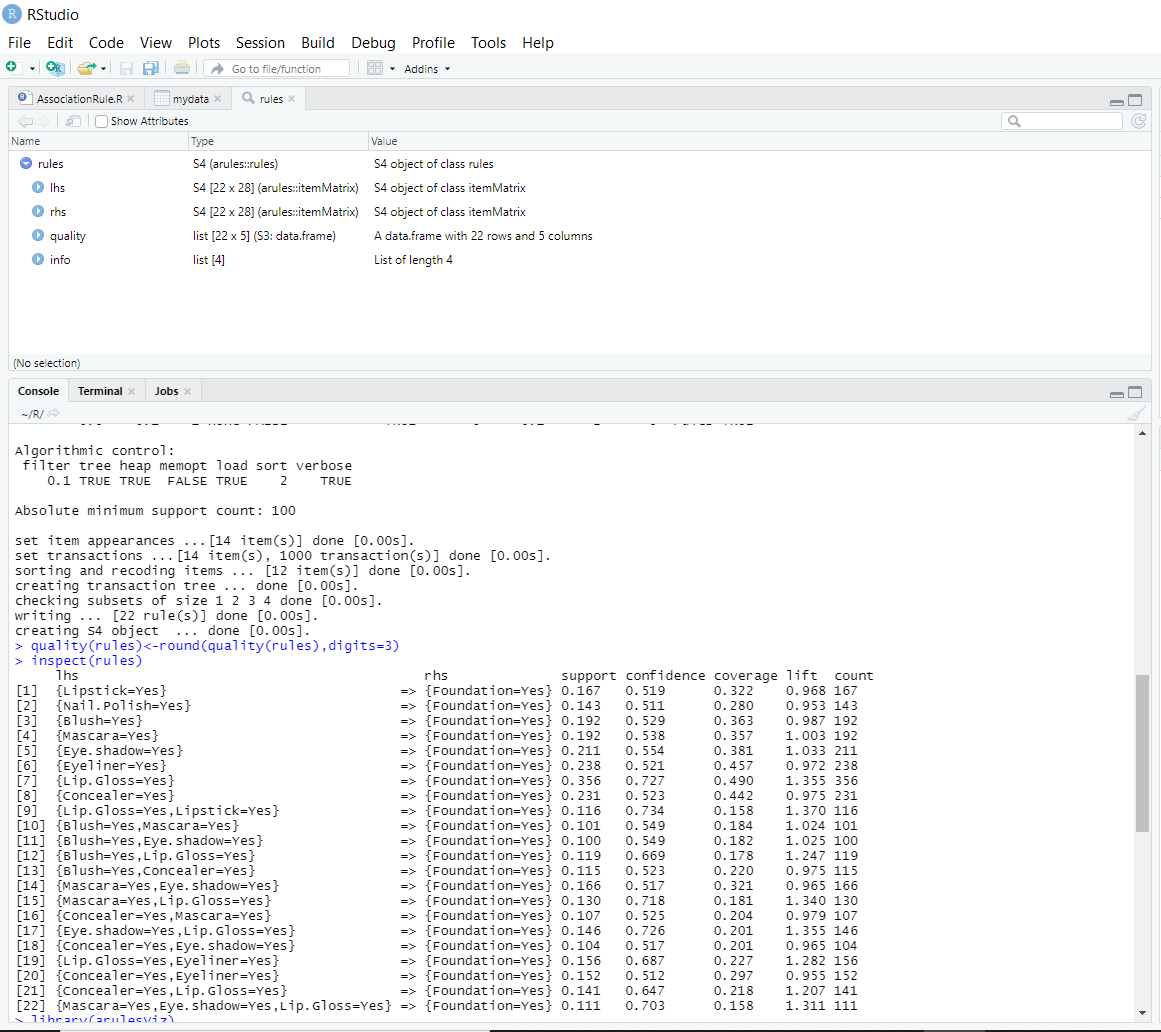
If you can see the results below then you have successfully configured RStudio IDE with R and the required packages for the exercises of the book. If not, please go through steps 4-9 again.

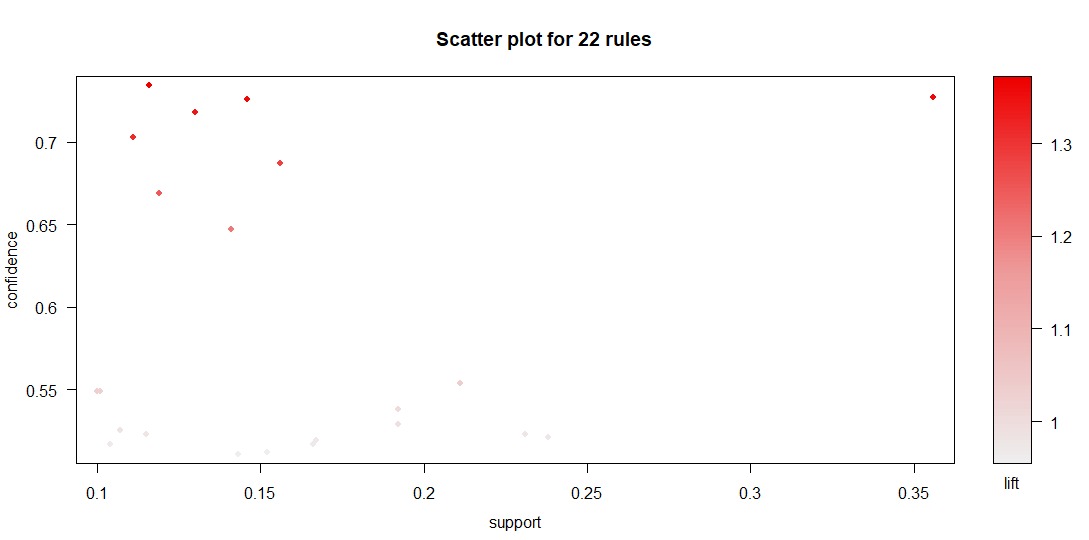


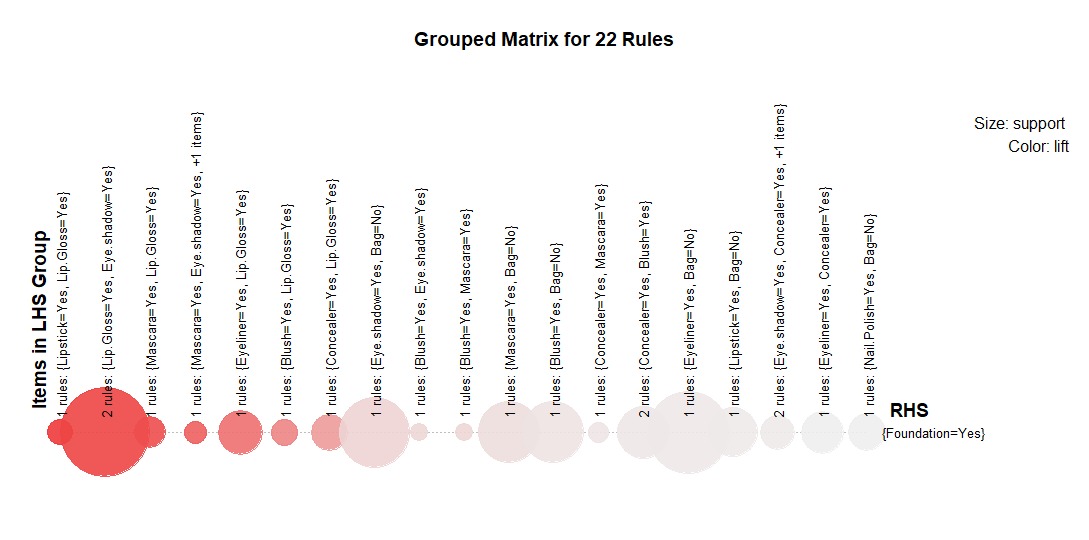
* **Screenshot of Cosmetics Data And Association Rule**

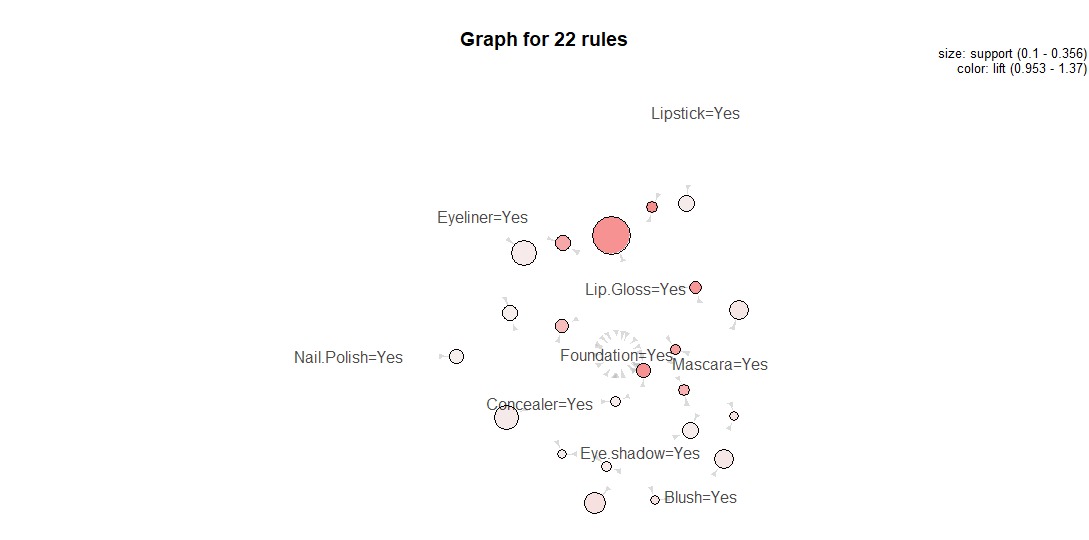






****

****

****

* **Reference link of web material**

<file:///C:/Users/Admin/Downloads/RStudio_setup_guide_draft1.pdf>

<https://www.youtube.com/watch?v=_V8eKsto3Ug>

* **5 Question/Answers**

1. **Explain what is R?**

R is data analysis software which is used by analysts, quants, statisticians, data scientists and others.

1. **List out some of the function that R provides?**

* Mean
* Median
* Distribution
* Covariance
* Regression
* Non-linear
* Mixed Effects

**3) Mention what does not ‘R’ language do?**

* Though R programming can easily connects to DBMS is not a database
* R does not consist of any graphical user interface
* Though it connects to Excel/Microsoft Office easily, R language does not provide any spreadsheet view of data

**4) What are some advantages of R?**

* It’s important to be familiar with the advantages and disadvantages of certain languages and ecosystems. R is no exception. So what are the advantages of R?
* Its open-source nature. This qualifies as both an advantage and disadvantage for various reasons, but being open source means it’s widely accessible, free to use, and extensible.
* Its package ecosystem. The built-in functionality available via R packages means you don’t have to spend a ton of time reinventing the wheel as a data scientist.
* Its graphical and statistical aptitude. By many people’s accounts, R’s graphing capabilities are unmatched

### 5) What are the disadvantages of R?

Just as you should know what R does well, you should understand its failings.

Memory and performance. In comparison to [Python](https://www.springboard.com/blog/data-science-with-python/), R is often said to be the lesser language in terms of memory and performance. This is disputable, and many think it’s no longer relevant as 64-bit systems dominate the marketplace.

Open source. Being open source has its disadvantages as well as its advantages. For one, there’s no governing body managing R, so there’s no single source for support or quality control. This also means that sometimes the packages developed for R are not the highest quality.

Security. R was not built with security in mind, so it must rely on external resources to mind these gaps.

* **Conclusion**

From this practical we have learnt how to R programming works and analyse the association rule in R studio software and also learnt the different functions of R studio.

|  |  |  |  |
| --- | --- | --- | --- |
| **Prepared By:** | Rajiv Kumar Gupta (18CE137) | **Date:** | 03-07-2021 |

**Aprioro Algorithm:**

Apriori algorithm was the first algorithm that was proposed for frequent itemset mining. It was later improved by R Agarwal and R Srikant and came to be known as Apriori. This algorithm uses two steps “join” and “prune” to reduce the search space. It is an iterative approach to discover the most frequent itemsets.

**Apriori says:**

The probability that item I is not frequent is if:

* P(I) < minimum support threshold, then I is not frequent.
* P (I+A) < minimum support threshold, then I+A is not frequent, where A also belongs to itemset.
* If an itemset set has value less than minimum support then all of its supersets will also fall below min support, and thus can be ignored. This property is called the Antimonotone property.

**The steps followed in the Apriori Algorithm of data mining are:**

1. **Join Step**: This step generates (K+1) itemset from K-itemsets by joining each item with itself.
2. **Prune Step**: This step scans the count of each item in the database. If the candidate item does not meet minimum support, then it is regarded as infrequent and thus it is removed. This step is performed to reduce the size of the candidate itemsets.

#### **Steps In Apriori**

Apriori algorithm is a sequence of steps to be followed to find the most frequent itemset in the given database. This data mining technique follows the join and the prune steps iteratively until the most frequent itemset is achieved. A minimum support threshold is given in the problem or it is assumed by the user.

**#1)** In the first iteration of the algorithm, each item is taken as a 1-itemsets candidate. The algorithm will count the occurrences of each item.

**#2)** Let there be some minimum support, min\_sup ( eg 2). The set of 1 – itemsets whose occurrence is satisfying the min sup are determined. Only those candidates which count more than or equal to min\_sup, are taken ahead for the next iteration and the others are pruned.

**#3)** Next, 2-itemset frequent items with min\_sup are discovered. For this in the join step, the 2-itemset is generated by forming a group of 2 by combining items with itself.

**#4)** The 2-itemset candidates are pruned using min-sup threshold value. Now the table will have 2 –itemsets with min-sup only.

**#5)** The next iteration will form 3 –itemsets using join and prune step. This iteration will follow antimonotone property where the subsets of 3-itemsets, that is the 2 –itemset subsets of each group fall in min\_sup. If all 2-itemset subsets are frequent then the superset will be frequent otherwise it is pruned.

**#6)** Next step will follow making 4-itemset by joining 3-itemset with itself and pruning if its subset does not meet the min\_sup criteria. The algorithm is stopped when the most frequent itemset is achieved.

**Some fields where Apriori is used:**

1. **In Education Field:** Extracting association rules in data mining of admitted students through characteristics and specialties.
2. **In the Medical field:** For example Analysis of the patient’s database.
3. **In Forestry:** Analysis of probability and intensity of forest fire with the forest fire data.
4. Apriori is used by many companies like Amazon in the **Recommender System** and by Google for the auto-complete feature.

### Frequent Pattern Growth Algorithm

This algorithm is an improvement to the Apriori method. A frequent pattern is generated without the need for candidate generation. FP growth algorithm represents the database in the form of a tree called a frequent pattern tree or FP tree.

This tree structure will maintain the association between the itemsets. The database is fragmented using one frequent item. This fragmented part is called “pattern fragment”. The itemsets of these fragmented patterns are analyzed. Thus with this method, the search for frequent itemsets is reduced comparatively.

### FP Tree

Frequent Pattern Tree is a tree-like structure that is made with the initial itemsets of the database. The purpose of the FP tree is to mine the most frequent pattern. Each node of the FP tree represents an item of the itemset.

The root node represents null while the lower nodes represent the itemsets. The association of the nodes with the lower nodes that is the itemsets with the other itemsets are maintained while forming the tree.

### Frequent Pattern Algorithm Steps

The frequent pattern growth method lets us find the frequent pattern without candidate generation.

**Let us see the steps followed to mine the frequent pattern using frequent pattern growth algorithm:**

**#1)** The first step is to scan the database to find the occurrences of the itemsets in the database. This step is the same as the first step of Apriori. The count of 1-itemsets in the database is called support count or frequency of 1-itemset.

**#2)** The second step is to construct the FP tree. For this, create the root of the tree. The root is represented by null.

**#3)**The next step is to scan the database again and examine the transactions. Examine the first transaction and find out the itemset in it. The itemset with the max count is taken at the top, the next itemset with lower count and so on. It means that the branch of the tree is constructed with transaction itemsets in descending order of count.

**#4)** The next transaction in the database is examined. The itemsets are ordered in descending order of count. If any itemset of this transaction is already present in another branch (for example in the 1st transaction), then this transaction branch would share a common prefix to the root.

This means that the common itemset is linked to the new node of another itemset in this transaction.

**#5)** Also, the count of the itemset is incremented as it occurs in the transactions. Both the common node and new node count is increased by 1 as they are created and linked according to transactions.

**#6)** The next step is to mine the created FP Tree. For this, the lowest node is examined first along with the links of the lowest nodes. The lowest node represents the frequency pattern length 1. From this, traverse the path in the FP Tree. This path or paths are called a conditional pattern base.

Conditional pattern base is a sub-database consisting of prefix paths in the FP tree occurring with the lowest node (suffix).

**#7)** Construct a Conditional FP Tree, which is formed by a count of itemsets in the path. The itemsets meeting the threshold support are considered in the Conditional FP Tree.

**#8)** Frequent Patterns are generated from the Conditional FP Tree.