|  |  |
| --- | --- |
| **Assignment Case** |  |
| COMP6140  Data Mining |
| **Computer Science** | **O193-COMP6140-BO01-01** |
| ***Valid on*** *Odd Semester Year 2018/2019* | **Revision 00** |

## Soal

*Case*

**Restaurant Cuisines**

**Bluejack Data Center** is a specialist data center in Jakarta. **Bluejack Data Center** is a data center that can **analyze** and **visualize** data for resulting a new data that can help client to improve their business.

Now, **Bluejack Data Center** is in restaurant cuisines data research to get the **frequent cuisines** data that available in a restaurant. **Bluejack Data Center** are provided by some data in csv format as following:

* **chefmozcuisine**.**csv**

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Data Type** | **Description** |
| placeId | Integer | The unique identifier of the corresponding restaurant. |
| Rcuisine | Character | Cuisines that available in the restaurant. |

* **geoplaces2**.**csv**

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Data Type** | **Description** |
| placeId | Integer | The unique identifier of the corresponding restaurant. |
| latitude | Numeric | Latitude point of the corresponding restaurant. |
| longitude | Numeric | Longitude point of the corresponding restaurant. |
| the\_geom\_meter | Character | The geometry of the corresponding restaurant. |
| name | Character | The name of the corresponding restaurant. |
| address | Character | The address of the corresponding restaurant. |
| city | Character | The city of the corresponding restaurant. |
| state | Character | The state of the corresponding restaurant. |
| country | Character | The country of the corresponding restaurant. |
| fax | Character | The fax number of the corresponding restaurant. |
| zip | Character | The zip code of the corresponding restaurant. |
| alcohol | Character | Define whether the restaurant serve alcohol or not. |
| smoking area | Character | Define whether the restaurant has smoking area or not. |
| dress\_code | Character | Define what dress code the customer must use. |
| accessibility | Character | Define the restaurant accessibility. |
| price | Character | Define the price level of the corresponding restaurant. |
| url | Character | The website’s url of the restaurant. |
| Rambience | Character | Define the restaurant ambience. |
| franchise | Character | Status that define the restaurant is a franchise or not.  (“f”: not a franchise, “t”: a franchise) |
| area | Character | Define whether the restaurant is an open or closed area. |
| other\_services | Character | Define whether the restaurant has other services or not. |

* **rating\_final**.**csv**

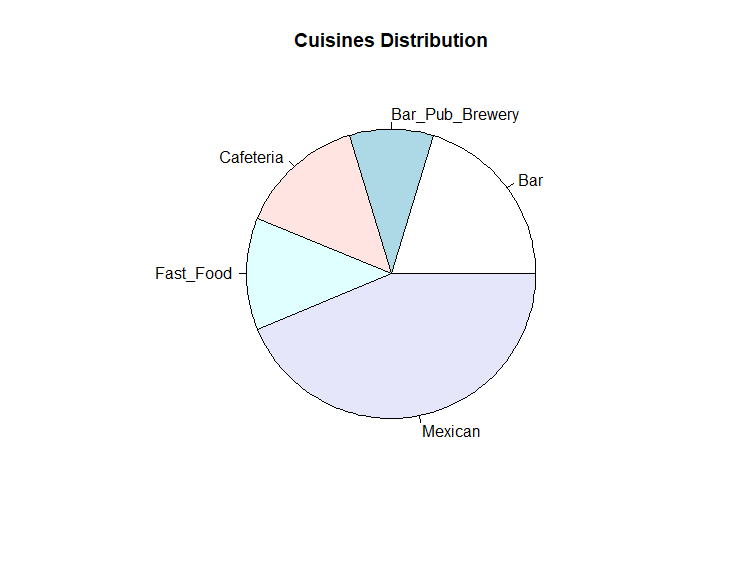
|  |  |  |
| --- | --- | --- |
| **Attribute** | **Data Type** | **Description** |
| userId | Character | The unique identifier of user that give rating value. |
| placeID | Integer | The unique identifier of the corresponding restaurant. |
| rating | Integer | The general rating of the corresponding restaurant. |
| food\_rating | Integer | The food rating of the corresponding restaurant. |
| service\_rating | Integer | The service rating of the corresponding restaurant. |

As the data scientist in **Bluejack Data Center**, you are asked to do some task below:

1. **Data Visualization**

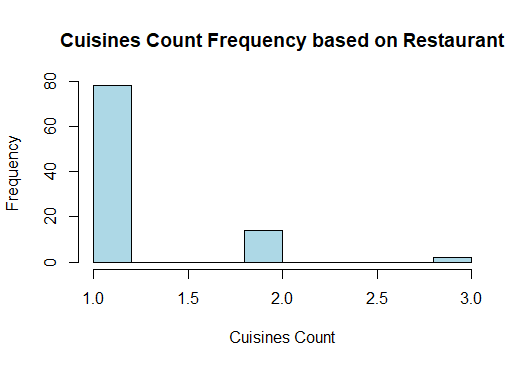
To help the client understand the data easier, you are asked to **visualize** the data in **graph** **form**. Some data that needed to be visualized are:

* 1. Show **cuisines distribution** where the **cuisines** are provided by **more** **than** **5** **restaurants**.



**Figure 1**.Cuisines Distribution Chart

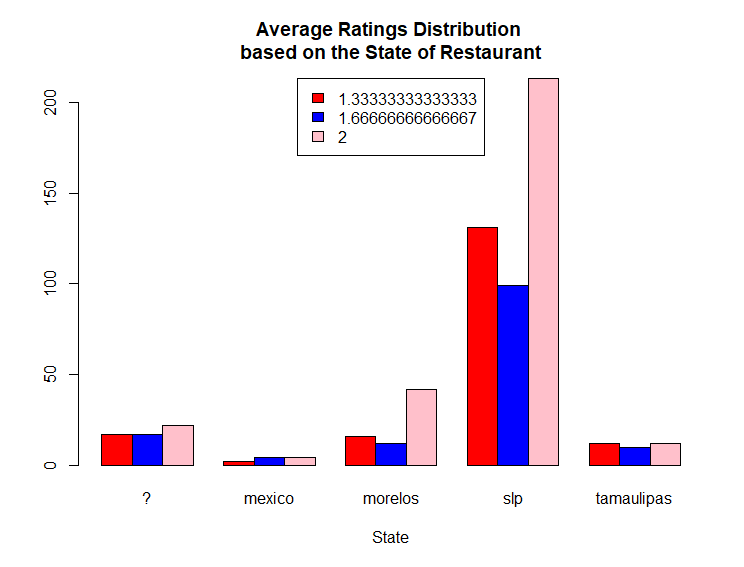
* 1. Show the **frequency** of **restaurant** **cuisines** **count**.



**Figure 2**. Cuisine Count Frequency based on Restaurant Chart

* 1. Show **restaurant state** distribution that the **average** **rating** is **above** **1**.**2**.

**Average Rating** = **AVG** (**rating**, **food**\_**rating**, **service**\_**rating**)



**Figure 3**. Restaurant Average Ratings Distribution

based on State Chart

1. **Frequent Pattern Analysis**

To help clients know what cuisines are popular nowadays in the food business, you are asked to do frequent pattern analysis to search for **frequent cuisines** that is available in a Restaurant. To get the **frequent cuisines** data, use the “**chefmozcuisine**.**csv**” file and do some step below:

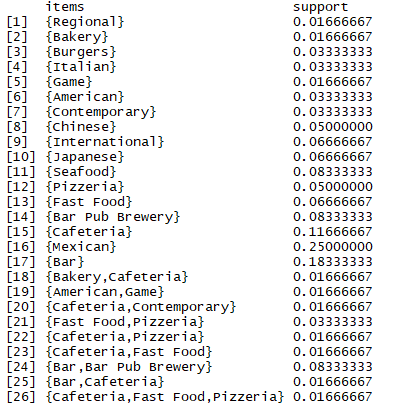
* 1. **Data Preprocessing**

In the **Data Preprocessing** phase, there are some data that can’t be used for the further analysis. Do the following task to **cleanse** the data:

* **Remove** all **restaurants** thatis a **franchise restaurant**.
* **Remove** all **restaurants** that provide**s other services** in the restaurant.
* **Remove** all **restaurants** thatthe **country** is **not defined**.
* **Replace underscore (‘\_’)** inthe **cuisine name** to **space** (‘ ’).
  1. **Data Transformation**

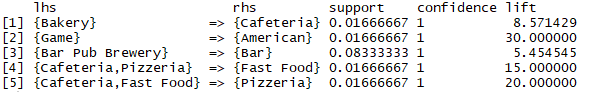
In this phase, you need to **change** **the** **data**, so it is suitable to be used in the **Apriori** analysis. Prepared the **cuisines** data in term of the **cuisine** **name**.

* 1. **Data Mining**
* Show **frequent** **cuisines** using **Apriori** algorithm with **minimum support**: **0**.**008** based on the data that have already pre-processed.



**Figure 4**. Frequent Cuisines using Apriori

* Show the **association** **rules** using **minimum confidence**: **0**.**8** based on the frequent cuisines that resulted from step above.



**Figure 5**. Association Rules

**References**:

* https://archive.ics.uci.edu/ml/datasets/Restaurant+%26+consumer+data

**Files to be collected**:

* R Studio Project Folder consist of:
  + - .RData
    - .Rhistory
    - .Rproj
    - .r
    - .Rproj