|  |  |
| --- | --- |
| **Assignment Case** |  |
| COMP6140  Data Mining |
| **Computer Science** | **O213-COMP6140-JE04-01** |
| ***Valid on*** *Odd Semester Year 2020/2021* | **Revision 00** |

1. Seluruh mahasiswa tidak diperkenankan untuk:

*All students are not allowed to:*

* + - Melihat sebagian atau seluruh jawaban mahasiswa lain,

*Seeing a part or the whole answer from other student*

* + - Menyadur sebagian maupun seluruh jawaban dari buku,

*Adapted a part or the whole answer from the book*

* + - Mendownload sebagian maupun seluruh jawaban dari internet,

*Downloading a part or the whole answer from the internet,*

* + - Mengerjakan soal yang tidak sesuai dengan tema yang ada di soal,

*Working with another theme which is not in accordance with the existing theme in the matter of the case,*

* + - Melakukan tindakan kecurangan lainnya,

*Committing other dishonest actions,*

* + - Secara sengaja maupun tidak sengaja melakukan segala tindakan kelalaian yang menyebabkan hasil karyanya berhasil dicontek oleh orang lain / kelompok lain.

*Accidentally or intentionally conduct any failure action that cause the results of the project was copied by someone else / other groups.*

1. Jika mahasiswa terbukti melakukan tindakan seperti yang dijelaskan butir 1 di atas, maka **nilai mahasiswa** yang melakukan kecurangan (menyontek maupun dicontek) akan di – **NOL** – kan.

*If the student is proved to the actions described in point 1 above, the score of the student which committed dishonest acts (cheating or being cheated) will be “Zero”*

1. Perhatikan jadwal pengumpulan jawaban, segala jenis pengumpulan jawaban di luar jadwal tidak dilayani.

*Pay attention to the submission schedule, all kinds of submission outside the schedule will not be accepted*

1. Persentase penilaiaan untuk matakuliah ini adalah sebagai berikut:

*Marking percentage for this subject is described as follows:*

|  |  |  |
| --- | --- | --- |
| **Tugas Mandiri**  *Assignment* | **Proyek**  *Project* | **UAP**  *Final Exam* |
| 40% | - | 60% |

1. Software yang digunakan pada matakuliah ini adalah sebagai berikut:

*Software will be used in this subject are described as follows:*

|  |
| --- |
| **Software**  *Software* |
| R 4.0.2  RStudio 1.3.9.959  RapidMiner Studio 9.7 |

## Ekstensi file yang harus disertakan dalam pengumpulan tugas mandiri dan proyek untuk matakuliah ini adalah sebagai berikut:

*File extensions should be included in assignment and project collection for this subject are described as follows:*

|  |  |
| --- | --- |
| **Tugas Mandiri**  *Assignment* | **UAP**  *Final Exam* |
| Folder Project (RDATA, RHISTORY, RPROJ, R, RPROJ.USER) | RMP, PROPERTIES |

## Soal

*Case*

**The Convenient Store**

**The Convenient Store** is a conveniently placed store were many shoppers go to buy their daily goods**. The Convenient Store** has chosen you to help them **analyze** and **visualize** all their transaction data so that the shop manager knows what needs improvements.

**The Convenient Store** requested you to research from the data that are available. The data is in csv format as the following:

* **orders**.**csv**

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Data Type** | **Description** |
| order\_id | Integer | The id of the transaction |
| product\_id | Integer | The id of the product |

* **products**.**csv**

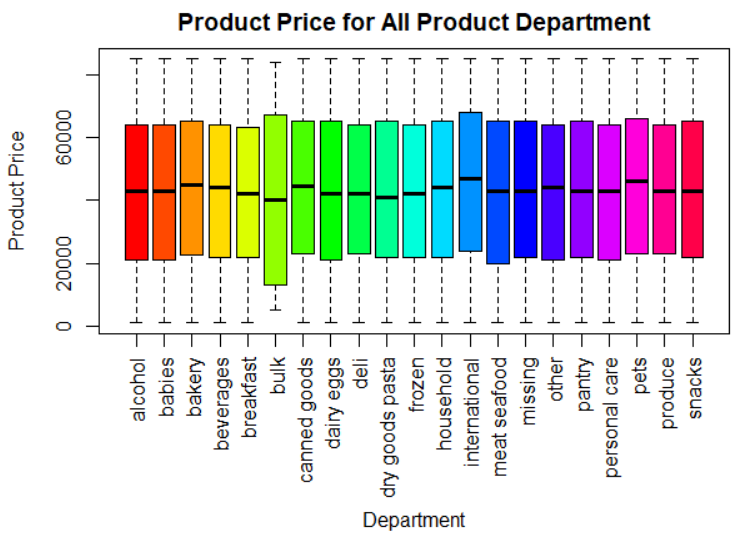
|  |  |  |
| --- | --- | --- |
| **Attribute** | **Data Type** | **Description** |
| product\_id | Integer | The id of the product |
| product\_name | Character | The name of the product |
| aisle | Character | The aisle where the product is located |
| department | Character | The department where the product is located |
| product\_price | Integer | The price of the product |

You are asked to help him analyze and visualize the data based on specification below:

1. **Data Visualization**

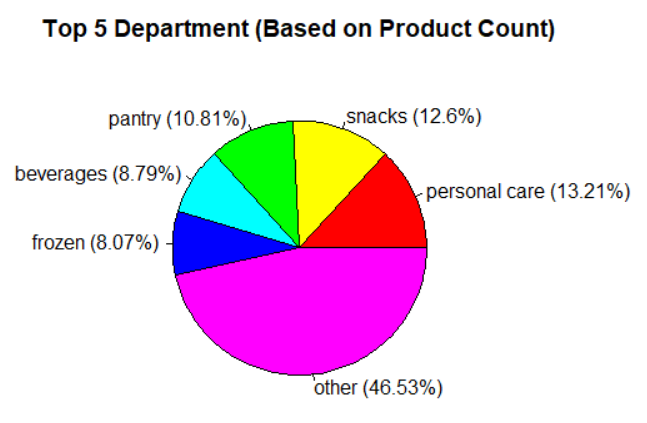
To describe the data better, you are asked to visualize the data in graph form. Some data that needed to be visualized are:

* 1. Show the **Product Price** for **all Product Department.**



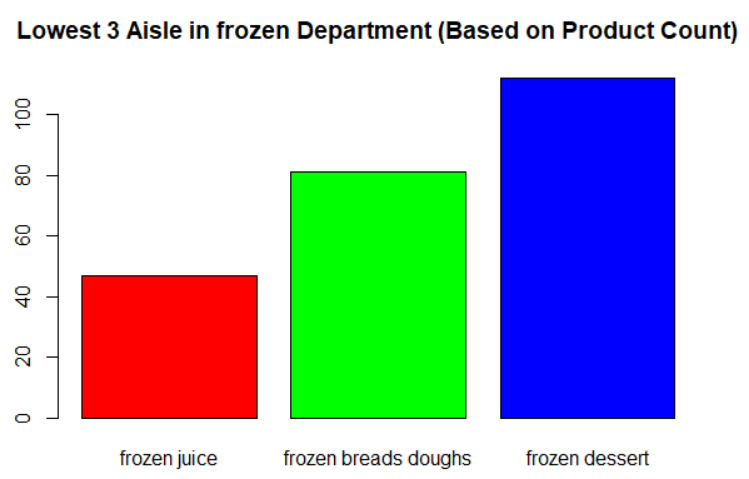
**Figure 1. Product Price for all Product Department Chart**

* 1. Show **top 5** **department** basedonits **product count**. Product which is **not** in the **top 5 department** will go into “**other**” categories. For each area, put **label** to show **percentage of the department** mentioned along with **department name**.



***Figure 2. Top 5 Department based on Product Count Chart***

* 1. Show the **lowest 3 aisle** based on its **product count**. Take only data in which the **department** is **frozen**.



**Figure 3. Lowest 3 Aisle based on Product Count Chart**

1. **Frequent Pattern Analysis**

You are asked to do frequent pattern analysis to know the **frequent product** that the people bought. To get the frequent product, use “**orders.csv**” and “**products.csv**” and follow all steps below:

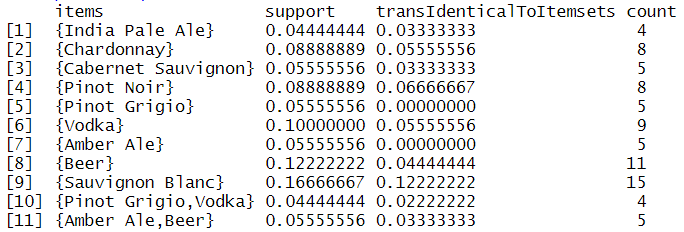
* 1. **Data Preprocessing**

In thisphase, some data can't be used for further analysis. Do the following task to **cleanse** the data:

* Removeall **product** which department **is** **not** **alcohol**
* Removeall **product** which aisle is **Specialty wines champagnes**
* Removeall **duplicated** data for the analysis
  1. **Data Transformation**

In this phase, you need to change the data, so it is suitable to be used in the **Apriori** analysis. Prepare the product data in terms of the **product’s name**.

* 1. **Data Mining**
* Show **frequent product** using **Apriori** algorithm with **minimum support**: **0**.**04** based on the data that have already pre-processed



**Figure 4. Frequent Product Result using Apriori**

* Show the **association** **rules** using **minimum confidence**: **0**.**5** based on the **frequent product** that resulted from step above.



**Figure 5. Association Rules Result**

**References**:

https://www.kaggle.com/c/instacart-market-basket-analysis/data

**Files to be collected**:

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

**If you do not understand, please ask your assistant!**