Assignment Case	
COMP6140 | COMP6140001 | COMP6140049 Data Mining	BINUS UNIVERSITY Software Laboratory Center
Computer Science	**Example Case**
Valid on -	**Revision 00**

Soal
Case

The Convenient Store

The Convenient Store is a conveniently placed store where 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 the data that is 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:

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

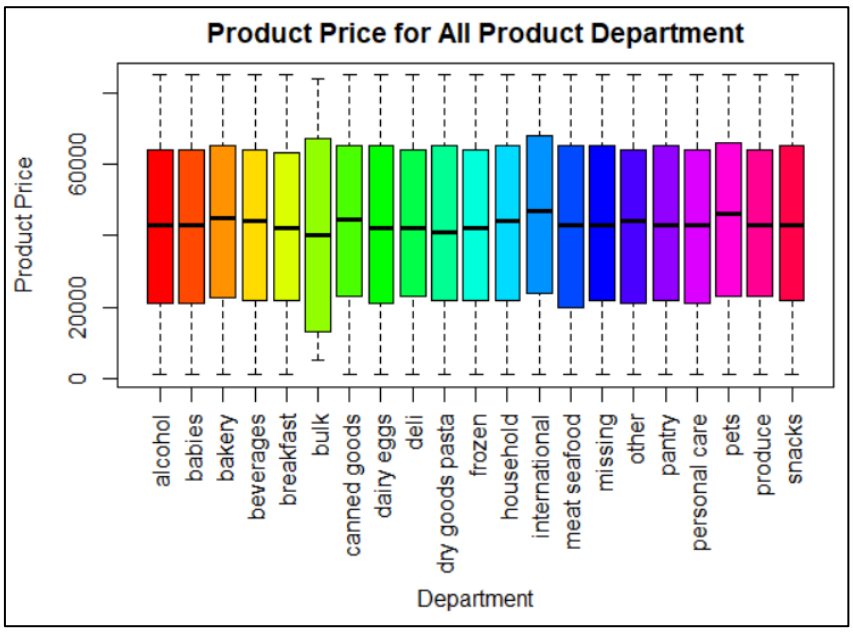

Figure 1. Product Price for all Product Department Chart

b. Show **top 5 department** based on its **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**.

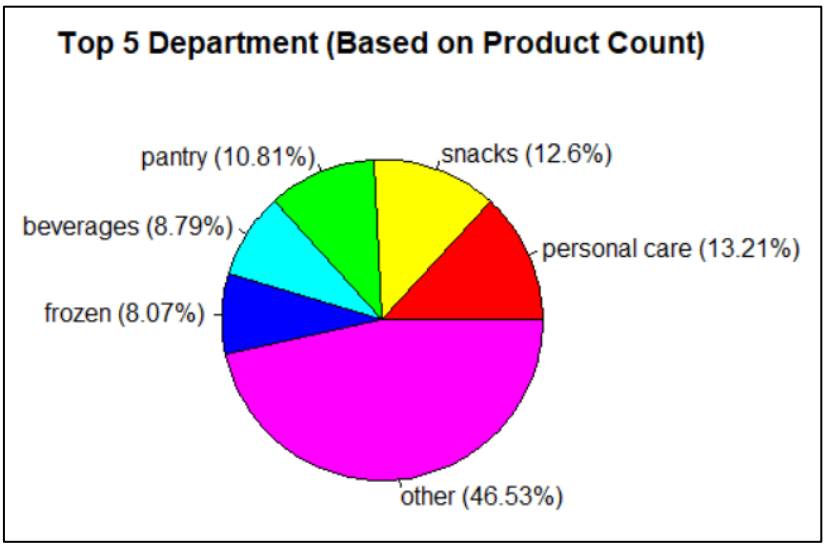

Figure 2. Top 5 Department based on Product Count Chart

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

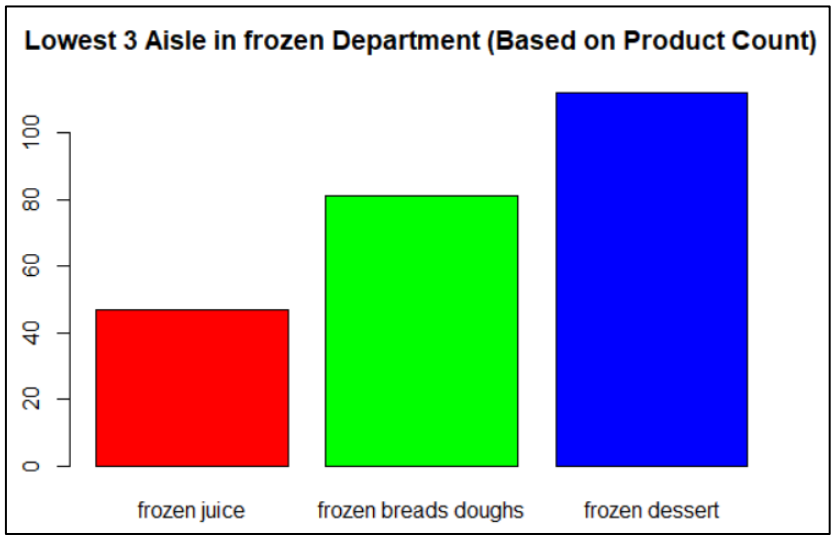

Figure 3. Lowest 3 Aisle based on Product Count Chart

2. 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:

a. **Data Preprocessing**

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

- Remove all **product** which department **is not alcohol**
- Remove all **product** which aisle is **Specialty wines champagnes**
- Remove all **duplicated** data for the analysis

b. **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**.

c. **Data Mining**

- Show **frequent product** using **Apriori** algorithm with **minimum support**: **0.04** based on the data that have already pre-processed

	items	support	transIdenticalToItemsets	count
[1]	{India Pale Ale}	0.04444444	0.03333333	4
[2]	{Chardonnay}	0.08888889	0.05555556	8
[3]	{Cabernet Sauvignon}	0.05555556	0.03333333	5
[4]	{Pinot Noir}	0.08888889	0.06666667	8
[5]	{Pinot Grigio}	0.05555556	0.00000000	5
[6]	{Vodka}	0.10000000	0.05555556	9
[7]	{Amber Ale}	0.05555556	0.00000000	5
[8]	{Beer}	0.12222222	0.04444444	11
[9]	{Sauvignon Blanc}	0.16666667	0.12222222	15
[10]	{Pinot Grigio,Vodka}	0.04444444	0.02222222	4
[11]	{Amber Ale,Beer}	0.05555556	0.03333333	5

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.

	lhs		rhs	support	confidence	lift
[1]	{Pinot Grigio}	=>	{Vodka}	0.04444444	0.8	8.000000
[2]	{Amber Ale}	=>	{Beer}	0.05555556	1.0	8.181818

Figure 5. Association Rules Result

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

If you do not understand, please ask your assistant!