Contents

[Input Files & Submissions 1](#_Toc74932936)

[Problem 1. Cluster Analysis 1](#_Toc74932937)

[Problem 2: Market Basket Analysis 4](#_Toc74932938)

[Problem 3. Data Visualization in Tableau Desktop 5](#_Toc74932939)

[Problem 3.1 Bubble Chart 6](#_Toc74932940)

[Problem 3.2 Sales & Shipping Cost 6](#_Toc74932941)

[Problem 3.3 Pick Your Own Question & Design 7](#_Toc74932942)

[Problem 3.4 7](#_Toc74932943)

[Problem 3.4.1 8](#_Toc74932944)

[Problem 3.4.2 8](#_Toc74932945)

[Problem 3.4.3 8](#_Toc74932946)

# Input Files & Submissions

|  |  |  |  |
| --- | --- | --- | --- |
| **Problems** | **Software** | **Input Files** | **Submissions** |
| 1.Cluster Analysis | SAS Enterprise Miner | Bank Direct Marketing | **Results - Output windows of 3 clusters** |
| 2.Market Basket Analysis | SAS Enterprise Miner | Transactions | **Results - Output windows** |
| Problems 1-2 |  |  | **This word document with your answers.** |
| 3.Data Visualization in Tableau Desktop | Tableau Desktop | Problems 3.1-3.3  Global Superstores | **Tableau Public URL #1**  Enter the URL in the assignment textbox. |
|  | Tableau Desktop | Problems 3.4 | **Tableau Public URL #2**  Enter the URL in the assignment textbox. |

# Problem 1. Cluster Analysis

**Instructions:**

* Download the data set **bank\_direct\_marketing** from Canvas, and save the input file in your data folder
* In SAS Enterprise Miner, build a cluster analysis model and answer the questions below where you see *[your answer]*.
* Saving the output for submission - in the Cluster node’s results, maximize the Output window, select File -> Save as

**Bank\_direct\_marketing dataset**

Demographic:

education

customer\_id

age

marital

job

Other information:

Default – default? (yes/no)

Balance – checking account balance

housing – having a mortgage account (yes/no)

loan – having a loan account (yes/no)

Promotion:

Contact – the channel salesperson contacted the customer

Day – contact date

Month – contact month

Duration – number of days since the last contact

Campaign – number of campaigns

pdays – see the detail below

previous – number of previous campaigns

poutcome – the previous outcome

y – currently having an active promotion of some sort (yes/no)

|  |  |
| --- | --- |
| pdays | number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted).” This is a combination of a continuous variable and a categorical variable. He consulted with the domain expert on the team and she suggested that the variable be binned as follows: |

**Bank Direct Marketing Cluster Analysis**

*Question: Do subgroups of customers exist in the bank direct marketing data set? If so, what are the best ways to describe them? What business recommendations can you make based on the results?*

1. Create a new diagram in your project. You can open an existing project or create a new one. Name the diagram **Bank Clustering**.
2. Add a new library if you created a new project. If the library of your existing project point to the data folder that contain the data set **bank\_direct\_marketing**, you don’t need to add a new library.
3. Use the **bank\_direct\_marketing** data as a data source for this clustering and profiling exercise.
4. Determine whether the model roles and measurement levels assigned to the variables are appropriate.

* Select the ellipsis next to **Variables** in the Properties panel.
* Sort the variables by level by clicking the **Level** column.
* Select all the continuous variables, except **customer\_id**.
* Click the **Explore** button and inspect the distributions.
* Click the plot containing the distribution for **balance** to select it.
* Right-click in the same plot and select **Graph** **Properties** ⇨ **Number of X Bins** and change the value to **100**. Click **OK**.
* Repeat for all the other histograms, steps e & f. (**Age** is displayed as a bar chart.). Note: The three most heavily skewed distributions are for **balance**, **campaign**, and **previous**. Although not optimal, we could reduce the skewness of the distributions by taking the log of the variable.

1. Drag a **Transform Variables** (in Modify tab) node onto the diagram and connect it to the **Input Data** source. Apply a log transformation to the following variables:

* balance
* previous
* campaign

*Here is the how-to:*

1. *Select the ellipsis next to* ***Variables****.*
2. *Change Method to Log for balance, previous, and campaign.*
3. *Run the node and do not view the results.*
4. Connect a **Cluster** node to the **Transform Variables** node.
5. Change **Maximum Number of Clusters** to **6**.
6. Change **Use** for all the variables to **No**, except for these:

* **balance**
* **previous**
* **duration**
* **Age**
* **campaign**

1. Run the Cluster node. Drag a **Segment Profile** node onto the diagram workspace and connect it to the **Cluster** node.
2. Run the **Segment Profile** node.
3. **Based on the three most important variables**. Describe the largest cluster (relevant variables, size, and profile description). *[your answer].*
4. Change **Maximum Number of Clusters** to **3** and rerun the **Cluster** node.
5. Inspect the profile results.

**Based on the three most important variables.**

1. If you were to present the results to your internal client, how would you label Clusters 1 and 2? *[your answer]*
2. In your opinion, what should you do about Cluster 3 containing only one observation? *[your answer]*
3. What business recommendations can you make based on the results? *[your answer]*

# Problem 2: Market Basket Analysis

**Problem Description:**

A store is interested in determining the associations between items purchased from the Health and Beauty Aids Department and the Stationery Department. The store chose to conduct a market basket analysis of specific items purchased from these two departments. The **TRANSACTIONS** data set contains information about more than 400,000 transactions made over the past three months. The following products are represented in the data set:

1. bar soap
2. bows
3. candy bars
4. deodorant
5. greeting cards
6. magazines
7. markers
8. pain relievers
9. pencils
10. pens
11. perfume
12. photo processing
13. prescription medications
14. shampoo
15. toothbrushes
16. toothpaste
17. wrapping paper

There are four variables in the data set:

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Model Role** | **Measurement Level** | **Description** |
| **STORE** | Rejected | Nominal | Identification number of the store |
| **TRANSACTION** | ID | Nominal | Transaction identification number |
| **PRODUCT** | Target | Nominal | Product purchased |
| **QUANTITY** | Rejected | Interval | Quantity of this product purchased |

* 1. Create a new diagram. Name the diagram **Transactions**.
  2. Create a new data source. Use the data set **TRANSACTIONS**.
  3. Assign the model role **Rejected** to the variables **STORE** and **QUANTITY**. These variables are not used in this analysis. Assign the **ID** model role to the variable **TRANSACTION** and the **Target** model role to the variable **PRODUCT**.
  4. Add the node for the **TRANSACTIONS** data set and an **Association** node to the diagram.
  5. Change the setting for **Export Rule by ID** to **Yes**.
  6. Do not change the remaining default settings for the Association node and run the analysis.
  7. Examine the results of the association analysis.

What is the highest lift value for the resulting rules? *[your answer]*

Which rule has this value? *[your answer]*

* 1. Examine the top five association rules that have highest life values and suggest at least one recommendation to the store manager. *[your answer]*

# Problem 3. Data Visualization in Tableau Desktop

**Connecting to the Global Superstore data source**

Start Tableau and add **Orders** to the canvas on the right side. Change Connection type to Extract.

## Problem 3.1 Bubble Chart

Create a bubble chart showing average profit and sum of quantity for various product sub-categories. Filter the data by Region, the filter should be shown on the worksheet as a single-value dropdown list. The sum of Quantity should be shown as size of bubble and average of Profit should be shown as color. Each bubble should display the Sub-Category label.

Present a summary of what you find from the bubble chart in the Caption window. Your summary should be between 2-5 lines.

Below are the steps

1. Add a new worksheet
2. Drag and drop Quantity on Size mark
3. Drag and drop Profit on Color mark. Change the function from Sum to Average.
4. Drag and drop Sub-Category on Label mark
5. Click Show Me button (top right) and select bubble chart
6. Drag and drop Region on Filters area. Click the Region filter and select Show Filter. On the right-hand side of the window, click the arrow next to Region filter and select Single Value (dropdown) type.
7. From the menu bar, click Worksheet and select Show Caption. Double-click on the Caption window. In the Edit Caption window, remove the default text and present a summary of what you find interesting from the bubble chart.

You can talk about any region and any sub-category you find interesting.

1. Change the chart title to “Quantity & Profit Bubbles”
2. Rename the worksheet as “**Problem 1**”

## Problem 3.2 Sales & Shipping Cost

Create an appropriate chart to show the sum of Sales and average of Shipping Cost for various Sub-Categories. Filter the data by Country and show filter on the right side.

Present what you find in the Caption window, your summary should be between 2-5 lines.

Rename the worksheet as “**Problem 2**”

## Problem 3.3 Pick Your Own Question & Design

Come up with a business problem or question and create a visualization of your choice to answer the question.

In the caption box, enter your:

1. Business problem or question
2. Audience
3. What you find from the visualization
4. Your recommendation(s) based on what you find

Change the worksheet title to your name and “[Your Name]’s Design #1”, e.g., **Will Smith’s Design #1**.

Name the worksheet **Problem 3**.

***Uploading Your Workbook***

To upload the workbook to Tableau Public, Select Server tab -> Tableau Public -> Save to Tableau Public, and create a new account. Follow the instructions to complete the upload. Make sure your workbook works properly.

## Problem 3.4

Open a new workbook and use the following file as a data source. Select Extract option for Connection.

FILE to use: Facebook Insights Post Level – Tableau

This Excel file contains actual data for Tableau’s Facebook posts across several months in 2014-15. The file is in the exact format that you would see when downloading data from Facebook Insights. (You might need to Google some field names in order to understand all the columns.)

Create Tableau visualizations for items 4.1 – 4.2 below and answer the questions in the caption box.

Pick your own key word(s) and metric(s) for item 4.3.

Rename the worksheets as 4.1, 4.2, and 4.3, respectively.

When finished, upload the workbook to the Tableau Public, copy the URL and submit it to Canvas.

*Hint: See how to create a calculated field in the Twitter example.*

### Problem 3.4.1

Looking at each month in the data, **does including an exclamation point (“!”) in the text of our Facebook post seem to increase the post’s average lifetime “engaged” users** (engaged users are users who clicked anywhere in the post)**?**

### Problem 3.4.2

**Are there certain posting times that seem to rarely receive negative feedback?** Pick an appropriate plot and answer the question in the caption box.

### Problem 3.4.3

Explore the relationship between the inclusion of other key words in the text (data, science, etc.) and other metrics of interest (negative feedback, etc.).