**Week 5 – Homework Exercise** **1**

1. **Analyzing a Path Analysis Data Source**
   1. Open the browser and sign into SAS Visual Analytics using your student credentials.
   2. Open the **VA2-Practice5.1** report in the **SAS Content** ⇨ **Courses** ⇨ **YVA285** ⇨ **Advanced** ⇨ **Practices** folder. Ensure you are in the editing report view mode.
   3. View data item properties and answer the following question:

How many unique courses did students attend?

**Answer: There is 391 unique courses\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

* 1. Assign the following data items to the specified roles for the path analysis object:

|  |  |
| --- | --- |
| **Event** | **Course Code** |
| **Sequence order** | **Event Delivered (End) Date** |
| **Transaction identifier** | **Student ID** |

* 1. Answer the following question:

What information is displayed when you place your mouse pointer on  in the lower right corner of the object? What do you think this means?

**Answer: When I click ! in lower right corner, I could see that artificial sequence order was generated for 5 path that contains simultaneous events. I think it means that there are some people are taking multiple courses at a day / one day**

* 1. Modify options for the path analysis object to color each link by event and to show only the top 20 paths. Also change the display to a vertical layout.
  2. Answer the following questions:

With which course do a majority of students begin?

**Answer: begin with PG1**

Of those students who start with PG1, how many have not taken another class?

**Answer: 9595 people are not taking another classes**

Of those students who start with PG1, what is the next course that most are likely to take? How many students take this course?

**Answer: most student are taking PG2 after PG1 and 5045 students are taking.**

Are there any students who take PG1 twice? How many students do this?

**Answer: 56 students are re-taking PG1.**

* 1. Save the report. Sign out of Visual Analytics.

End of Exercise

**Week 5 – Homework Exercise 2**

In Exercise 1, you work with a data set named **PVA**. It contains data that represents charitable donations made to an American veterans’ association. The data represent the results of a mail campaign to solicit donations. The data set contains the following information:

* a flag to indicate respondents to the appeal and the dollar amount of their donations (**Target Gift Flag** and **Target Gift Amount**)
* respondents’ PVA promotion and giving history
* demographic data of the respondents

#### PVA Metadata Table

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** |  | **Measurement Level** | **Description** |
| **DemAge** |  | Interval | Age |
| **DemCluster** |  | Nominal | Demographic Cluster |
| **DemGender** |  | Nominal | Gender |
| **DemHomeOwner** |  | Binary | Home Owner |
| **DemMedHomeValue** |  | Interval | Median Home Value Region |
| **DemMedIncome** |  | Interval | Median Income Region |
| **DemPctVeterans** |  | Interval | Percent Veterans Region |
| **GiftAvg36** |  | Interval | Gift Amount Average 36 Months |
| **GiftAvgAll** |  | Interval | Gift Amount Average All Months |
| **GiftAvgCard36** |  | Interval | Gift Amount Average Card 36 Months |
| **GiftAvgLast** |  | Interval | Gift Amount Last |
| **GiftCnt36** |  | Interval | Gift Count 36 Months |
| **GiftCntAll** |  | Interval | Gift Count All Months |
| **GiftCntCard36** |  | Interval | Gift Count Card 36 Months |
| **GiftCntCardAll** |  | Interval | Gift Count Card All Months |
| **GiftTimeFirst** |  | Interval | Time Since First Gift |
| **GiftTimeLast** |  | Interval | Time Since Last Gift |
| **ID** |  | Nominal | Control Number |
| **PromCnt12** |  | Interval | Promotion Count 12 Months |
| **PromCnt36** |  | Interval | Promotion Count 36 Months |
| **PromCntAll** |  | Interval | Promotion Count All Months |
| **PromCntCard12** |  | Interval | Promotion Count Card 12 Months |
| **PromCntCard36** |  | Interval | Promotion Count Card 36 Months |
| **PromCntCardAll** |  | Interval | Promotion Count Card All Months |
| **StatusCat96NK** |  | Nominal | Status Category 96NK |
| **StatusCatStarAll** |  | Binary | Status Category Star All Months |
| **TARGET\_B** |  | Binary | Target Gift Flag |
| **TARGET\_D** |  | Interval | Target Gift Amount |

In the first exercise, you use SAS Visual Analytics to familiarize yourself with the data.

1. **Using SAS Visual Analytics**
   1. Sign in to SAS Viya for Learners.
   2. Select **Explore and Visualize Data** to begin exploring the data.
   3. Select the **PVA** data source.

How many columns of data does **PVA** contain? 29 columns

* 1. Select the **Data** pane on the left of the canvas (if it is not open).
     1. How many category variables does **PVA** contain? 5 category variables
     2. How many measure variables does **PVA** contain (excluding Frequency)? 24 without Frequency
     3. How many distinct control numbers does the data contain? 107K (106,546)
     4. How many distinct demographic clusters are represented in the data? 54
  2. Select the measure variables that begin with **Gift.** Add the measures into the analysis workspace. The result will be a correlation matrix. As a good analyst, you realize you should inspect your measures before proceeding. Click  (**Actions**) and select **View measure details** in the Data pane to view the properties of all the measures in the data.
     1. Which measure variables that begin with **Gift** have missing values? Gift Amount Average Card 36 months
     2. How many missing values are there for **Target Gift Amount**? 53273
     3. What is the average value for **Promotion Count All Months**? 48.48
  3. Close the measure details view, and start a new Page. Create an auto chart with **Status Category 96NK**. How many observations represent those with **Status Category 96NK = F**? 7260
  4. Change **Target Gift Flag** from a *measure* to a *category*. It is a binary indicator that represents a response to a mailing, where *1* indicates that customers did respond.

Create a bar chart. How many females responded to the campaign? 28699

(Hint: Create a bar chart on a new page. Add **Gender** as a Group or Lattice columns or Lattice rows role.)

* 1. Create a histogram of **Target Gift Amount** on a new page.
     1. Is the variable skewed? YES
     2. If so, in which direction? Skewed to right
  2. Save the report. Click  (**Menu**) ⇨ **Save As**. Save the report in My Folder. Click **Save**.

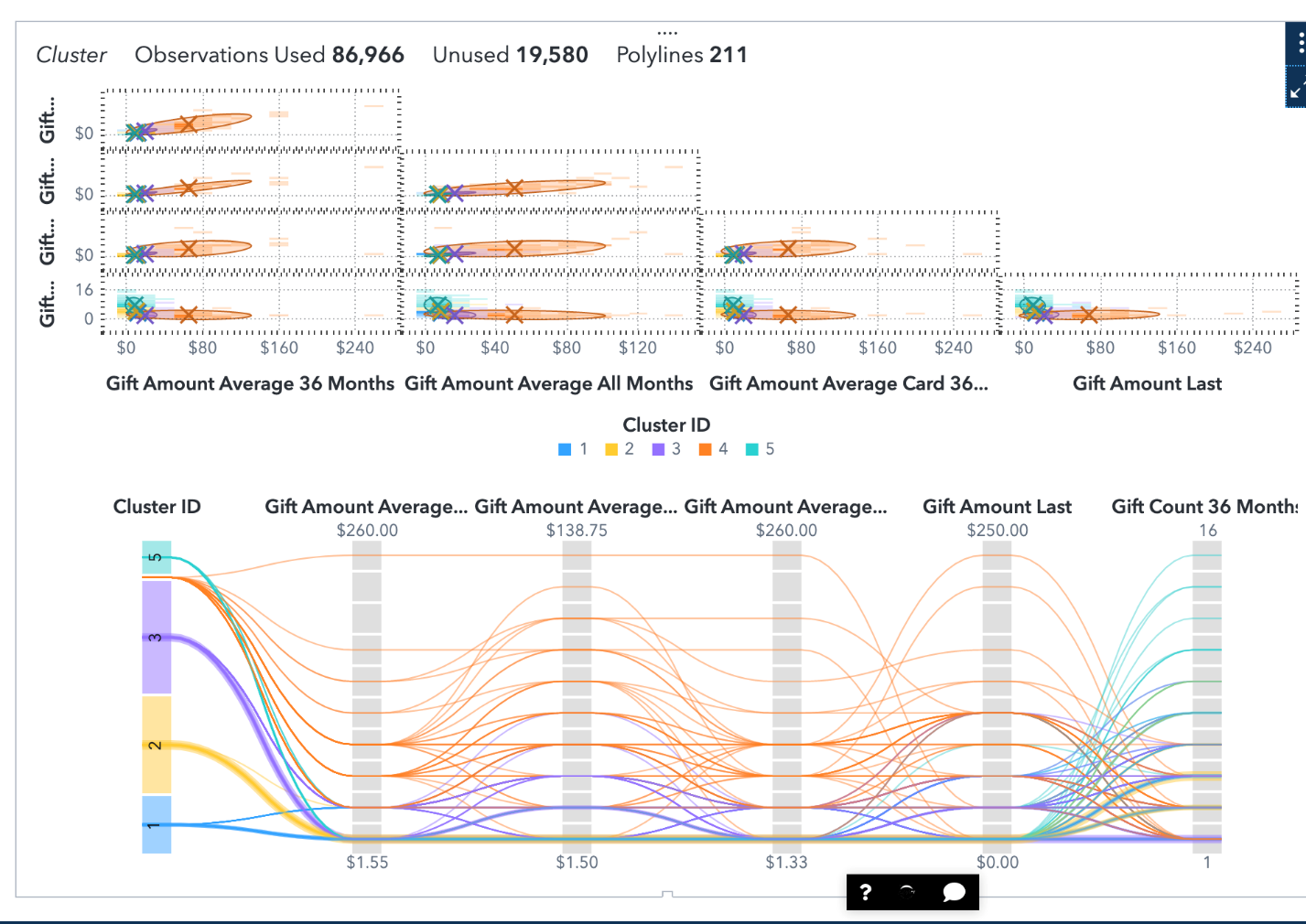
**Week 5 – Homework Exercise 3**

1. **Performing a Cluster Analysis**

This practice uses the **PVA** data set that you explored previously.

* 1. Perform a cluster analysis for all the variables that begin with **Gift**. Experiment with three-, four-, and five-cluster solutions.
  2. Examine the details table for the five-cluster solution. Which cluster ID had the largest variability? Please include a screenshot of the table to support your answer.

Cluster 3 is the largest.



* 1. Which two of the cluster centroids are closest?

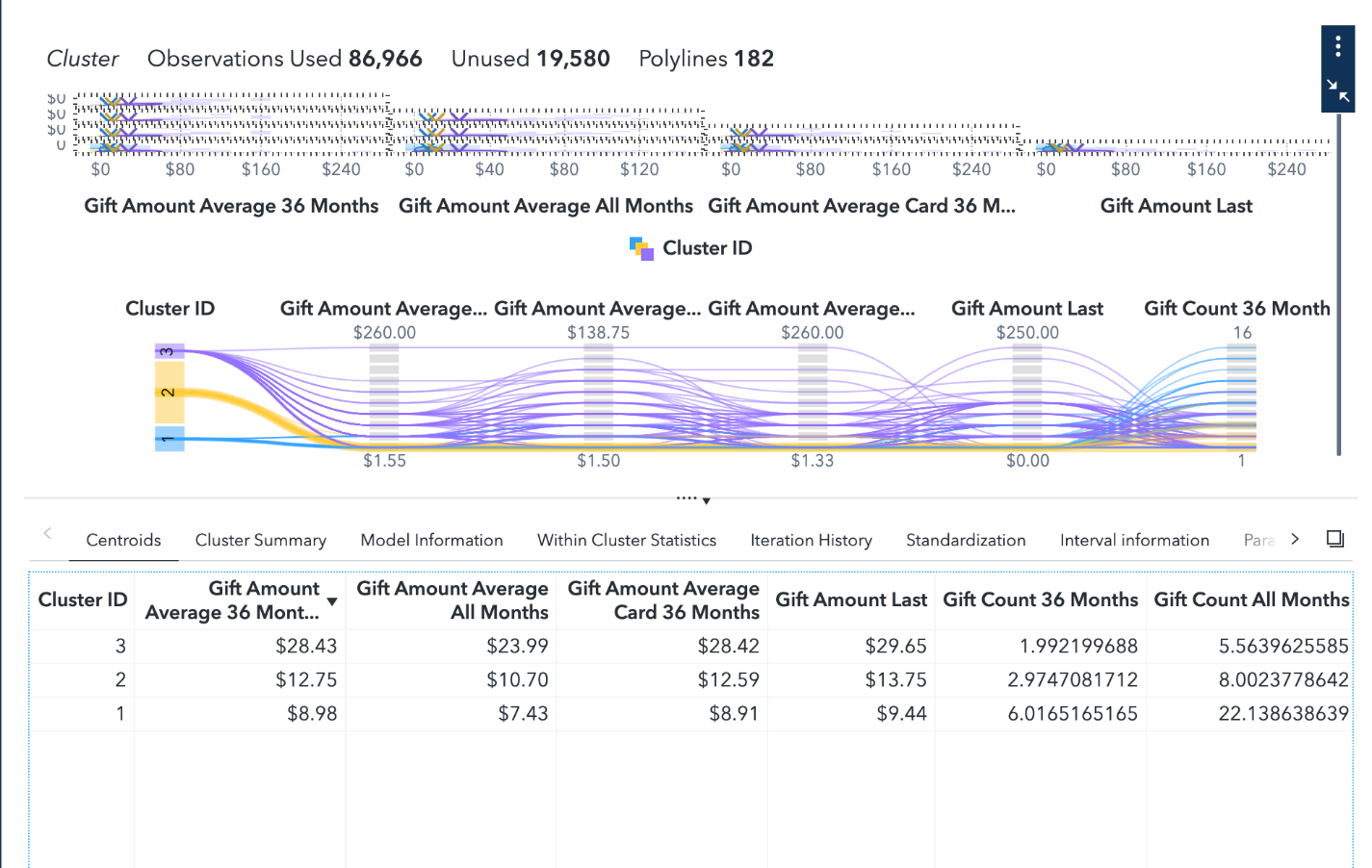
Cluster 1 and 2 are the closest

* 1. For the four-cluster solution, derive a Cluster ID variable. How many distinct levels are   
     in the new variable?

There are five different distinct level

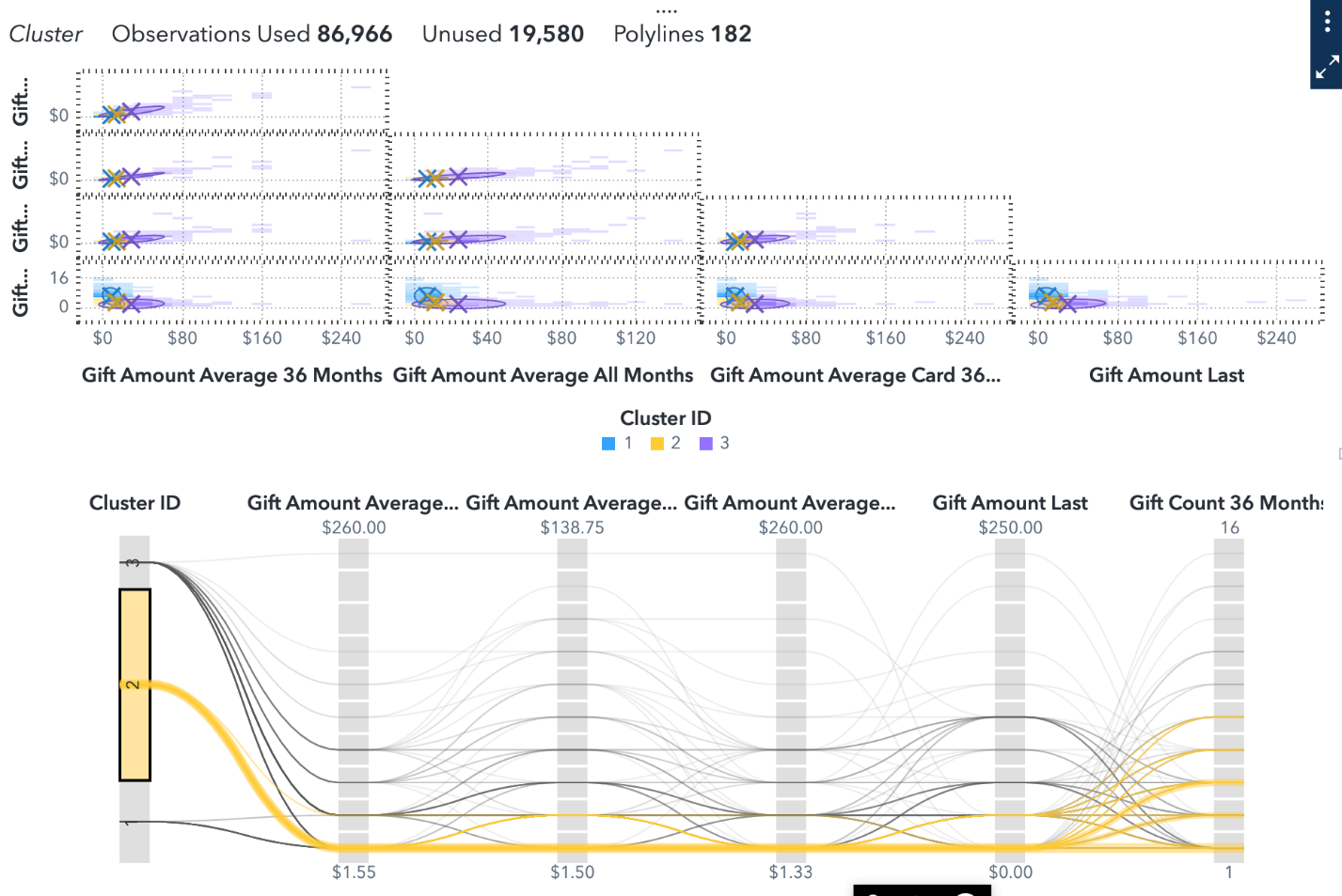
* 1. Create an auto chart of the new cluster variable. What does the -1 category represent?  
     It shows the missing values
  2. Examine the parallel coordinates plot for the three-cluster solution. Which cluster gave   
     the most money per donation? Please share a screenshot to support your answer.

Cluster 3 had the most money per donation.



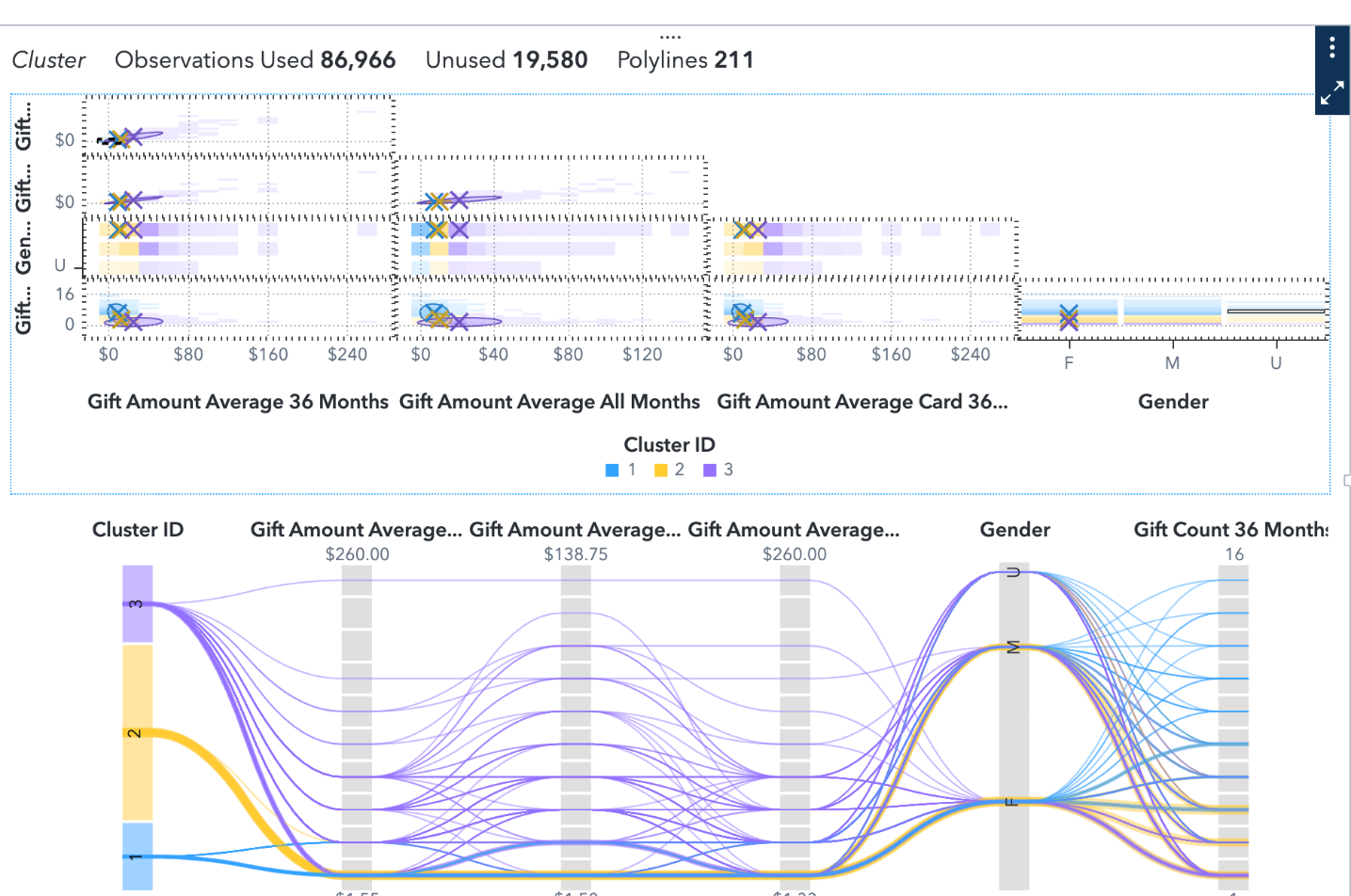
* 1. Which cluster donated most frequently? Please share a screenshot to support your answer. (Hint: Expand the number of visible roles for the plot.)

Cluster 2 had donated most frequently.



* 1. For the three-cluster solution, replace the variable **Gift Amount Last** with the variable **Gender**. In the new model, are the three clusters identified *Male*, *Female*, or *Unknown*? Please share a screenshot to support your answer.

Yes, three cluster are identified as male, female, and unknown.



* 1. Save the report. Click  (**Menu**) ⇨ **Save As**. Save the report in My Folder. Click **Save**.

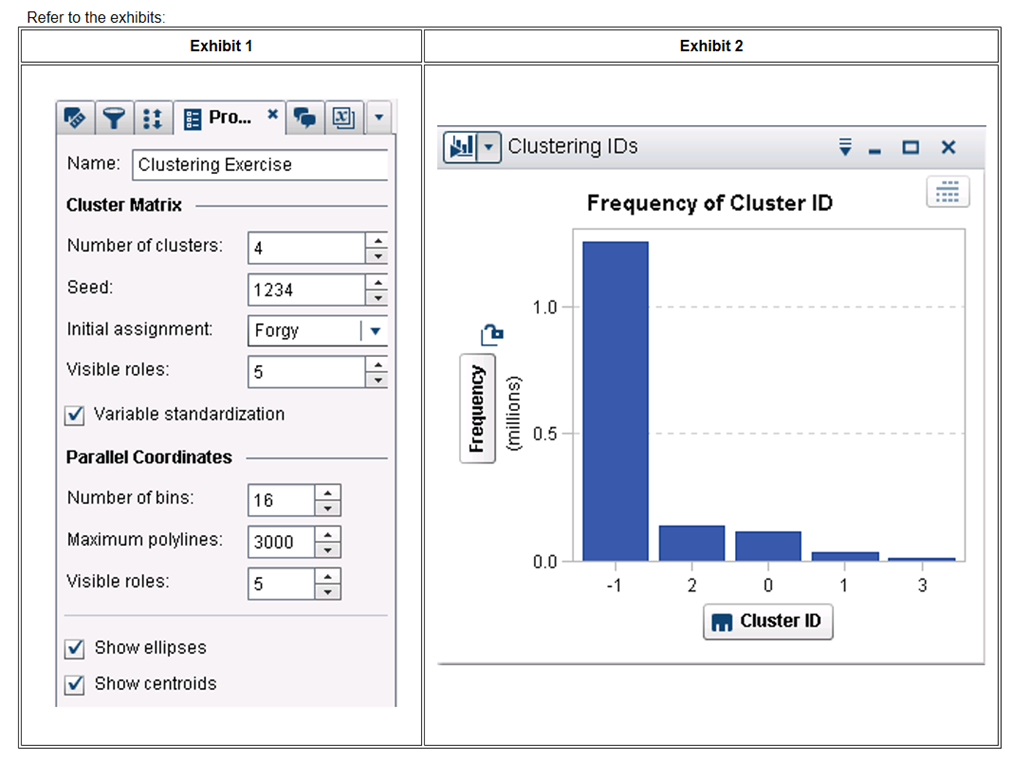
**Week 5 – Homework Exercise 4**

1. When specifying Cluster Roles for Clustering Analysis in Visual Statistics, which condition is correct?
   1. Interaction terms are allowed
   2. A target variable is required
   3. At least two measures are required
   4. Category items are allowed

**Answer:** At least two measures are required

*(Simply copy & paste your answer from the options provided)*

1. An online retailer wants to identify the following customers: bargain hunters, impulse shoppers, big spenders, and loyal repeaters. The retailer has big data on 6 demographic variables and 9 RFM behavioral variables. To begin a cluster analysis in SAS Visual Statistics, how many clusters should be selected?
   1. 4
   2. 6
   3. 9
   4. 15
   5. **Answer:** 4
2. Refer to the exhibits:



An analyst has created a cluster model based on the settings in Exhibit 1 and chose to derive a Cluster ID variable. Why are there 5 Cluster IDs in Exhibit 2 based on these settings?

* 1. Because the Cluster Matrix Visible Roles setting is 5
  2. Because there are observations with missing values
  3. Because Variable Standardization has been chosen
  4. Because the Parallel Coordinates Visible Roles setting is 5
  5. **Answer: B** Because there are observations with missing values

1. In the Cluster Model Options tab, what is the purpose of variable standardization?
   1. To filter out selected category variables
   2. To filter out observations with excessive values
   3. To transform category variables
   4. To transform measures of different scales
   5. **Answer:** To transform measures of different scales
2. Which statement is true regarding Clustering analysis?
   1. Sorting data improves cluster assignments
   2. Each observation is assigned to only one cluster
   3. Character variables create easier to understand clusters
   4. Significant domain knowledge is required to understand cluster output
   5. **Answer:** Each observation is assigned to only one cluster
3. Which statement is true regarding the number of polylines created in a Parallel Coordinates plot during Cluster Analysis?
   1. It is calculated using the number of variables visible in the plot
   2. It is calculated using the variable with the greatest number of bins
   3. It cannot exceed 3,000
   4. It is calculated using the number of variables leveraged in the cluster analysis
   5. **Answer:** It is calculated using the number of variables visible in the plot
4. Cluster Analysis in SAS Visual Statistics is categorized as which type of machine learning?
   1. Supervised learning
   2. Semi-supervised learning
   3. Unsupervised learning
   4. Reinforcement learning
   5. **Answer:** Unsupervised learning