**Santander Bank Product Recommendation**

1. Introduction

We are provided with 1.5 years of customer data from Santander bank to predict which products their existing customers will use in the next month. To support needs for a range of financial decisions, Santander Bank offers a lending hand to their customers through personalized product recommendations based on their past behavior and that of similar customers. With this, Santander can better meet the individual needs of all customers and ensure their satisfaction no matter where they are in life.



1. Data Retrieval

***Dataset Link:*** [*https://www.kaggle.com/c/santander-product-recommendation/data*](https://www.kaggle.com/c/santander-product-recommendation/data)

We are provided with 1.5 years of customer data from Santander bank to predict what new products customers will purchase. We have divided our data set in two part having one column as “Customer Code” same in both the data sets: -

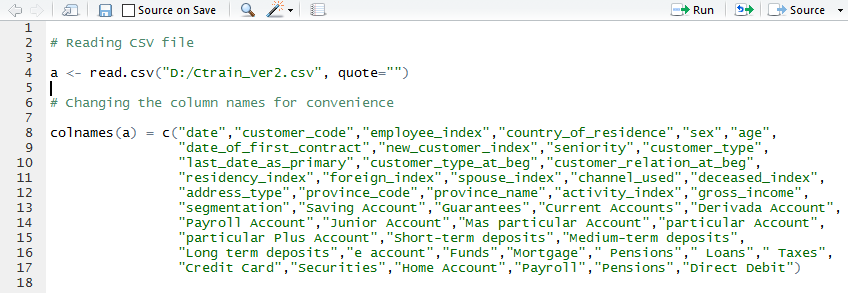
* Customer Demographics
* Products (Different Accounts Opened by Customers - Transactional Data)

1. Data Pre-Process
   1. *Rename the Data Fields and Reduce the transactions*

We have the 30 data fields whose names are written in Spanish language which were hard to understand for us. So, we changed their labels into English language using R Console. We used R Console because our dataset is too large to open in .xlsx or .csv format having more than 600K transactions.

Also, we have reduced the number of transactions from 400K to 30K customers details so that it is easy to handle in .csv format.

**R Code: -**



*Customer Demographics Data Fields*

|  |  |  |  |
| --- | --- | --- | --- |
| S.No. | Column Name (Spanish) | Column Name (English) | Description |
| 1. | fecha\_dato | Transactional Date | The table is partitioned for this column |
| 2. | Ncodpers | customer\_code | Customer code |
| 3. | ind\_empleado | employee\_index | Employee index: A active, B ex employed, F filial, N not employee, P pasive |
| 4. | pais\_residencia | country\_of\_residence | Customer's Country residence |
| 5. | Sexo | Sex | Customer's sex |
| 6. | Age | Age | Age |
| 7. | fecha\_alta | date\_of\_first\_contract | The date in which the customer became as the first holder of a contract in the bank |
| 8. | ind\_nuevo | new\_customer\_index | New customer Index. 1 if the customer registered in the last 6 months. |
| 9. | Antiguedad | Seniority | Customer seniority (in months) |
| 10. | Indrel | customer\_type | 1 (First/Primary), 99 (Primary customer during the month but not at the end of the month) |
| 11. | ult\_fec\_cli\_1t | last\_date\_as\_primary | Last date as primary customer (if he isn't at the end of the month) |
| 12. | indrel\_1mes | customer\_type\_at\_beg | Customer type at the beginning of the month ,1 (First/Primary customer), 2 (co-owner), P (Potential),3 (former primary), 4(former co-owner) |
| 13. | tiprel\_1mes | customer\_relation\_at\_beg | Customer relation type at the beginning of the month, A (active), I (inactive), P (former customer), R (Potential) |
| 13. | Indresi | residency\_index | Residence index (S (Yes) or N (No) if the residence country is the same than the bank country) |
| 14. | Indext | foreign\_index | Foreigner index (S (Yes) or N (No) if the customer's birth country is different than the bank country) |
| 15. | Conyuemp | spouse\_index | Spouse index. 1 if the customer is spouse of an employee |
| 16. | canal\_entrada | channel\_used | channel used by the customer to join |
| 17. | Indfall | deceased\_index | Deceased index. N/S |
| 18. | Tipodom | address\_type | Address type. 1, primary address |
| 19. | cod\_prov | province\_code | Province code (customer's address) |
| 20. | Nomprov | province\_name | Province name |
| 21. | ind\_actividad\_cliente | activity\_index | Activity index (1, active customer; 0, inactive customer) |
| 22. | Renta | gross\_income | Gross income of the household |
| 23. | Segment | segmentation | segmentation: 01 - VIP, 02 - Individuals 03 - college graduated |

*Product Purchased Data Fields*

|  |  |  |  |
| --- | --- | --- | --- |
| S.No. | Column Name (Spanish) | Column Name (English) | Description |
| 1. | Ncodpers | customer\_code | Customer code |
| 2. | ind\_ahor\_fin\_ult1 | Saving Account | Saving Account |
| 3. | ind\_aval\_fin\_ult1 | Guarantees | Guarantees |
| 4. | ind\_cco\_fin\_ult1 | Current Accounts | Current Accounts |
| 5. | ind\_cder\_fin\_ult1 | Derivada Account | Derivada Account |
| 6. | ind\_cno\_fin\_ult1 | Payroll Account | Payroll Account |
| 7. | ind\_ctju\_fin\_ult1 | Junior Account | Junior Account |
| 8. | ind\_ctma\_fin\_ult1 | Más particular Account | Más particular Account |
| 9. | ind\_ctop\_fin\_ult1 | particular Account | particular Account |
| 10. | ind\_ctpp\_fin\_ult1 | particular Plus Account | particular Plus Account |
| 11. | ind\_deco\_fin\_ult1 | Short-term deposits | Short-term deposits |
| 12. | ind\_deme\_fin\_ult1 | Medium-term deposits | Medium-term deposits |
| 13. | ind\_dela\_fin\_ult1 | Long-term deposits | Long-term deposits |
| 14. | ind\_ecue\_fin\_ult1 | e-account | e-account |
| 15. | ind\_fond\_fin\_ult1 | Funds | Funds |
| 16. | ind\_hip\_fin\_ult1 | Mortgage | Mortgage |
| 17. | ind\_plan\_fin\_ult1 | Pensions | Pensions |
| 18. | ind\_pres\_fin\_ult1 | Loans | Loans |
| 19. | ind\_reca\_fin\_ult1 | Taxes | Taxes |
| 20. | ind\_tjcr\_fin\_ult1 | Credit Card | Credit Card |
| 21. | ind\_valo\_fin\_ult1 | Securities | Securities |
| 22. | ind\_viv\_fin\_ult1 | Home Account | Home Account |
| 23. | ind\_nomina\_ult1 | Payroll | Payroll |
| 24. | ind\_nom\_pens\_ult1 | Pensions | Pensions |
| 25 | ind\_recibo\_ult1 | Direct Debit | Direct Debit |

* 1. *Categorization and Encoding of Customer Demographics Data Fields*

We have categorized and encoded 10 important fields of customer demographics so that we can use them as input variables in our clustering process. After this we have saved them in a separate CSV file. Following are the fields:-

* **Sex**: Values(V/H)

|  |  |
| --- | --- |
| SEX | ENCODE |
| V | 0 |
| H | 1 |

* **Age**: Multivalues (18-99)

|  |  |
| --- | --- |
| AGE RANGE | ENCODE |
| 0-20 | 0 |
| 21-40 | 1 |
| 41-60 | 2 |
| 61-80 | 3 |
| 81-100 | 4 |

* **New\_customer\_index/new:**

|  |  |
| --- | --- |
| NEW CUSTOMER INDEX | ENCODE |
| YES | 0 |
| NO | 1 |

* **Seniority**: Multivalues

|  |  |
| --- | --- |
| SENIORITY | ENCODE |
| 0-49 | 0 |
| 50-99 | 1 |
| 100-149 | 2 |
| 150-199 | 3 |
| 200-260 | 4 |

* **Customer\_relation\_at\_beg/relation:**

|  |  |
| --- | --- |
| RELATION | ENCODE |
| A | 0 |
| I | 1 |
| P | 2 |

* **Residency\_index/Residency:**

|  |  |
| --- | --- |
| RESIDENCY | ENCODE |
| S | 0 |
| N | 1 |

* **Foreign\_index/Foreign:**

|  |  |
| --- | --- |
| RESIDENCY | ENCODE |
| S | 0 |
| N | 1 |

* **Activity\_index:** Multivalues (1-7)
* **Gross\_income:** Encoded in terms of USA Gov division for low-class, middle-class, top-class

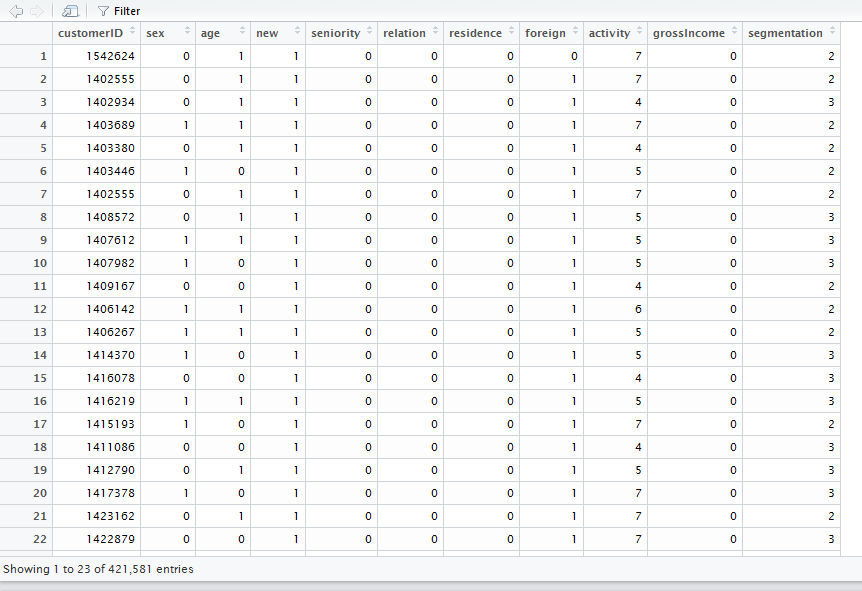
|  |  |
| --- | --- |
| INCOME RANGE | ENCODE |
| 0 - 19,999 | 0 |
| 20,000 - 34,999 | 1 |
| 35,000 - 59,999 | 2 |
| 60,000 - 99,999 | 3 |
| 100,000 – 149,999 | 4 |
| 150,000+ | 5 |

* **Segmentation**:

|  |  |
| --- | --- |
| SEGMENT | ENCODE |
| 01-VIP | 1 |
| 02-INDIVIDUALS | 2 |
| 03-COLLEGE GRADUATE | 3 |

So, our new sheet is now viewed as: -

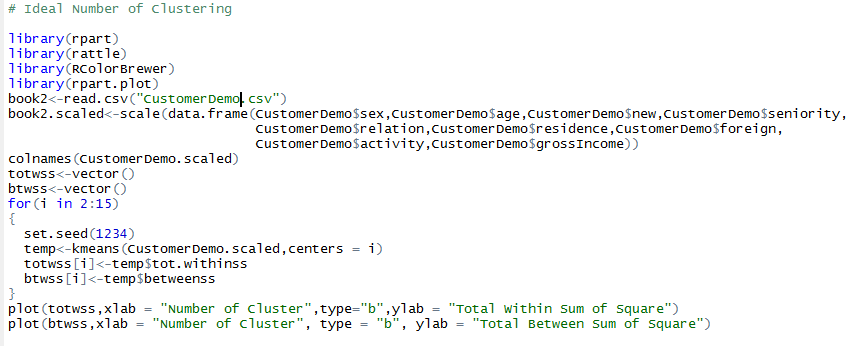




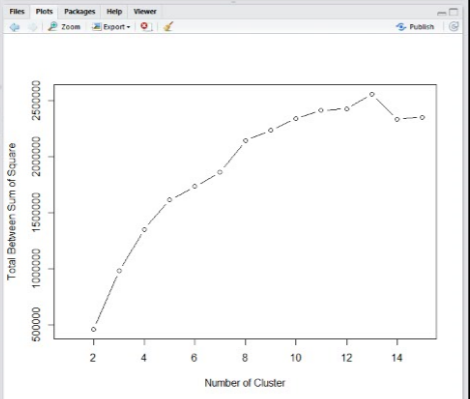
1. K-Means clustering

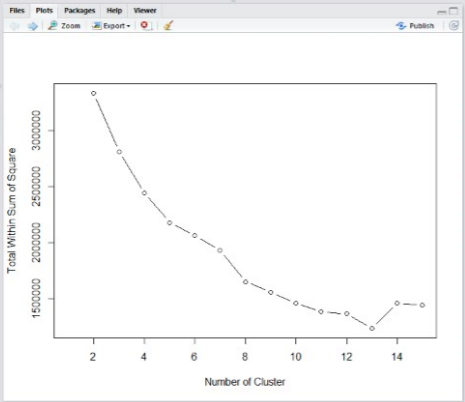
*Optimum Number of Clusters:* For cluster analysis, we have to select the optimum number of clusters.

**R Code: -**



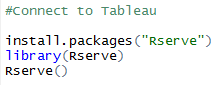
We got following 2 graphs to choose optimum number of clusters:



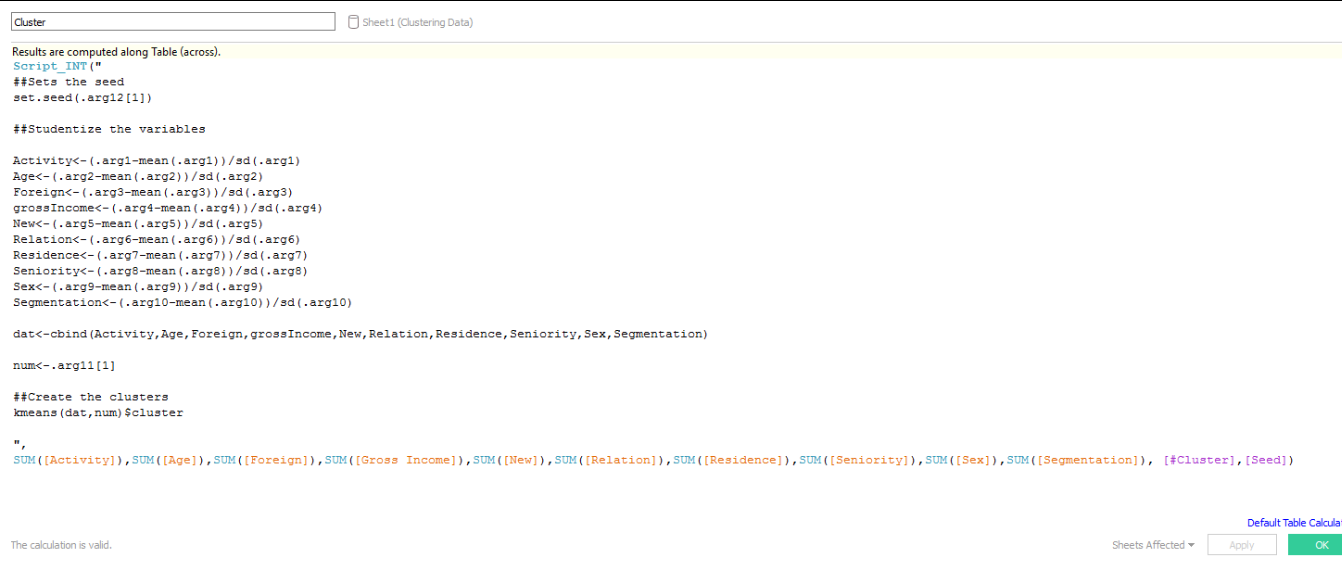


From here we have chosen **8 as our Optimum Number of Clusters.**

1. Tableau – R Integration



We have made the clusters in tableau using following code:



**Cluster visualization on Tableau:**

**screenshot**

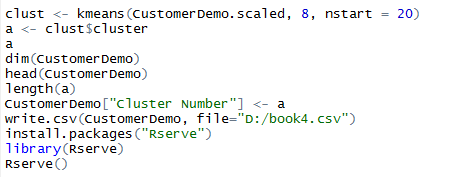
**Observation:** As soon as we visualize the clusters in tableau, we found Cluster 1-6 are better to analyze and we can ignore cluster 7 & cluster 8 because customers coming in these clusters doesn’t open any account.

1. Market Basket Analysis (Association Analysis)

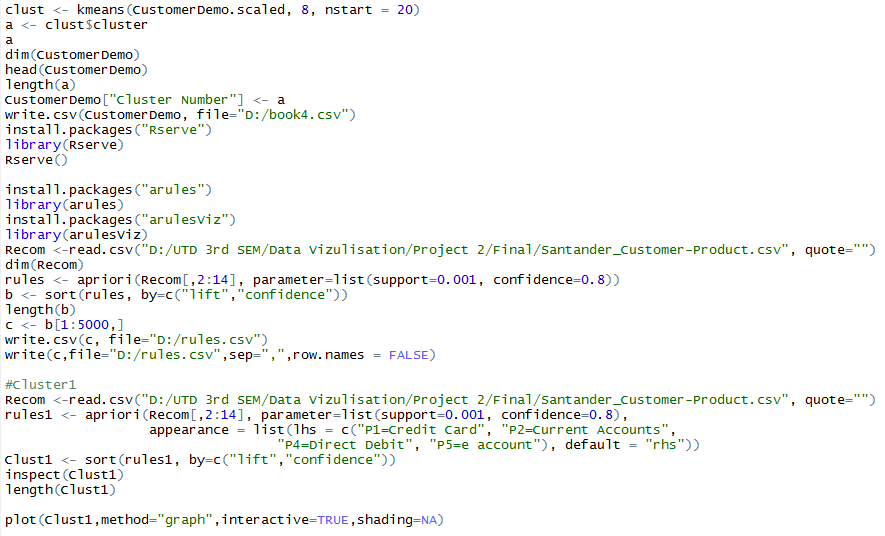
Now when we have done our clustering and come-up with few insights, we are ready to find out the **association rules** between the products based on customer profiles or our 6 clusters.

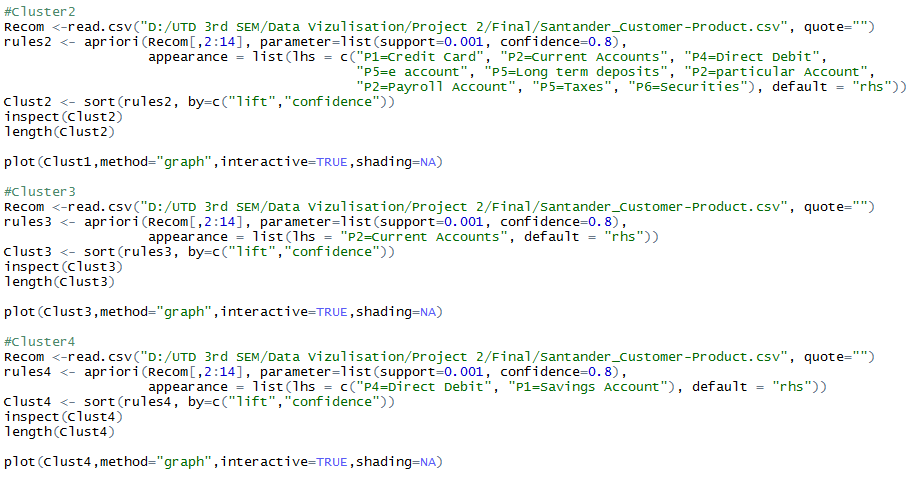
Doing this, we can do market-basket analysis of our product (Type of Account) and then recommend the future products to our customers based on their past purchase behavior.

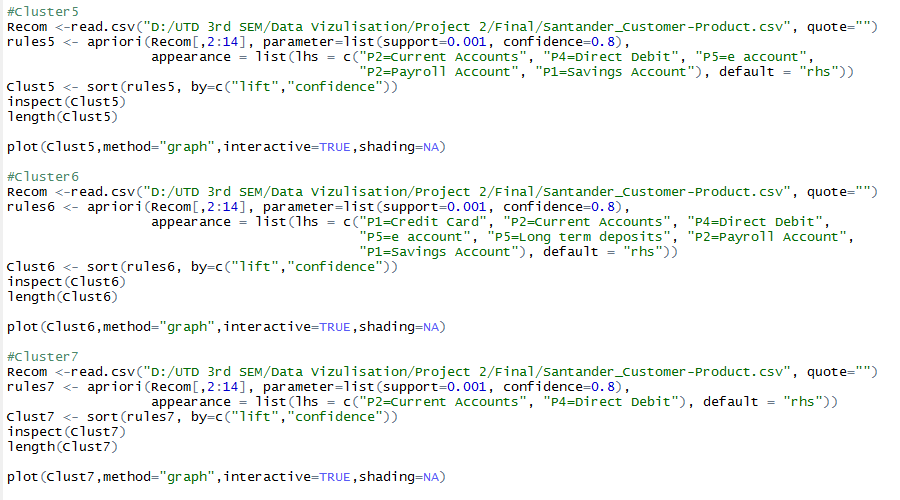
**Step1:** Add number of clusters in our data set.



**Step 2:** We made the association rules separately for each cluster based on its behavior whose **R Code** is written below.

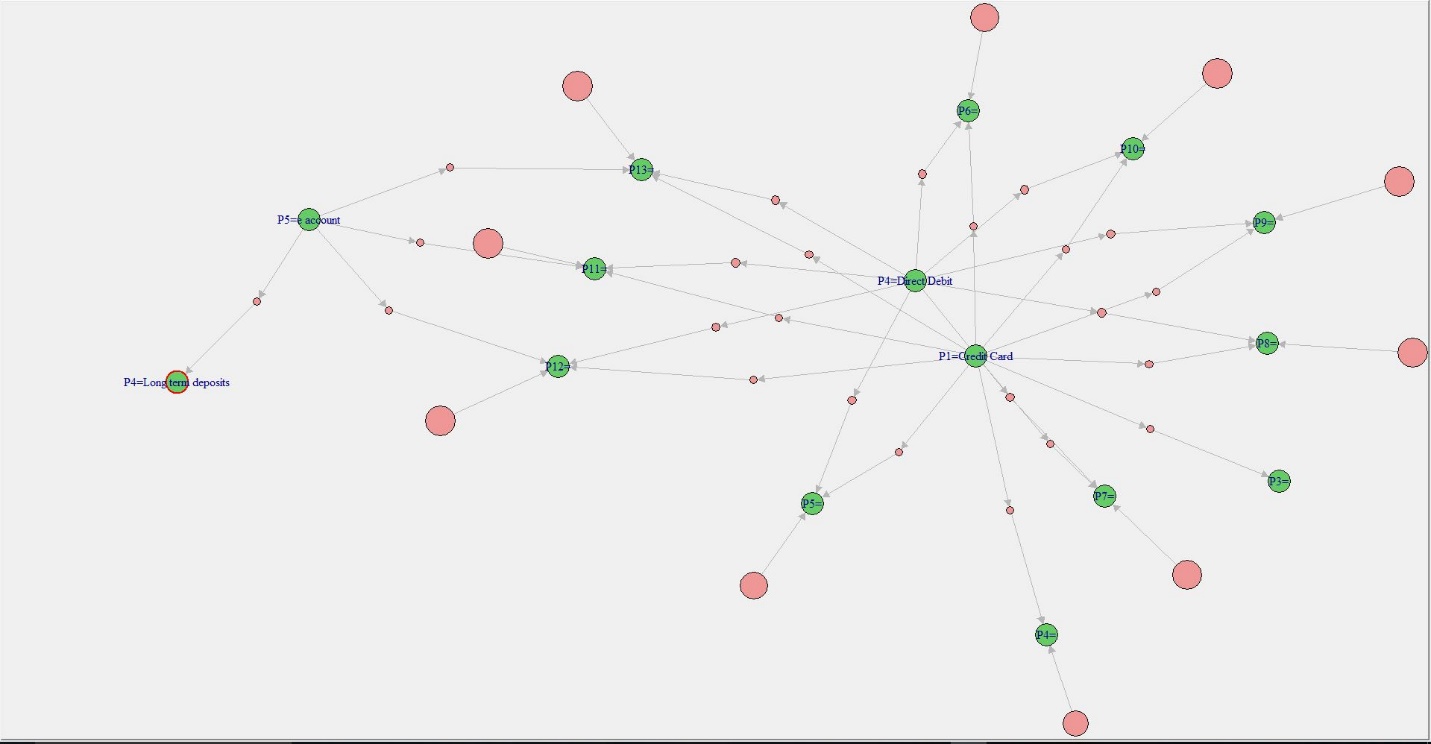




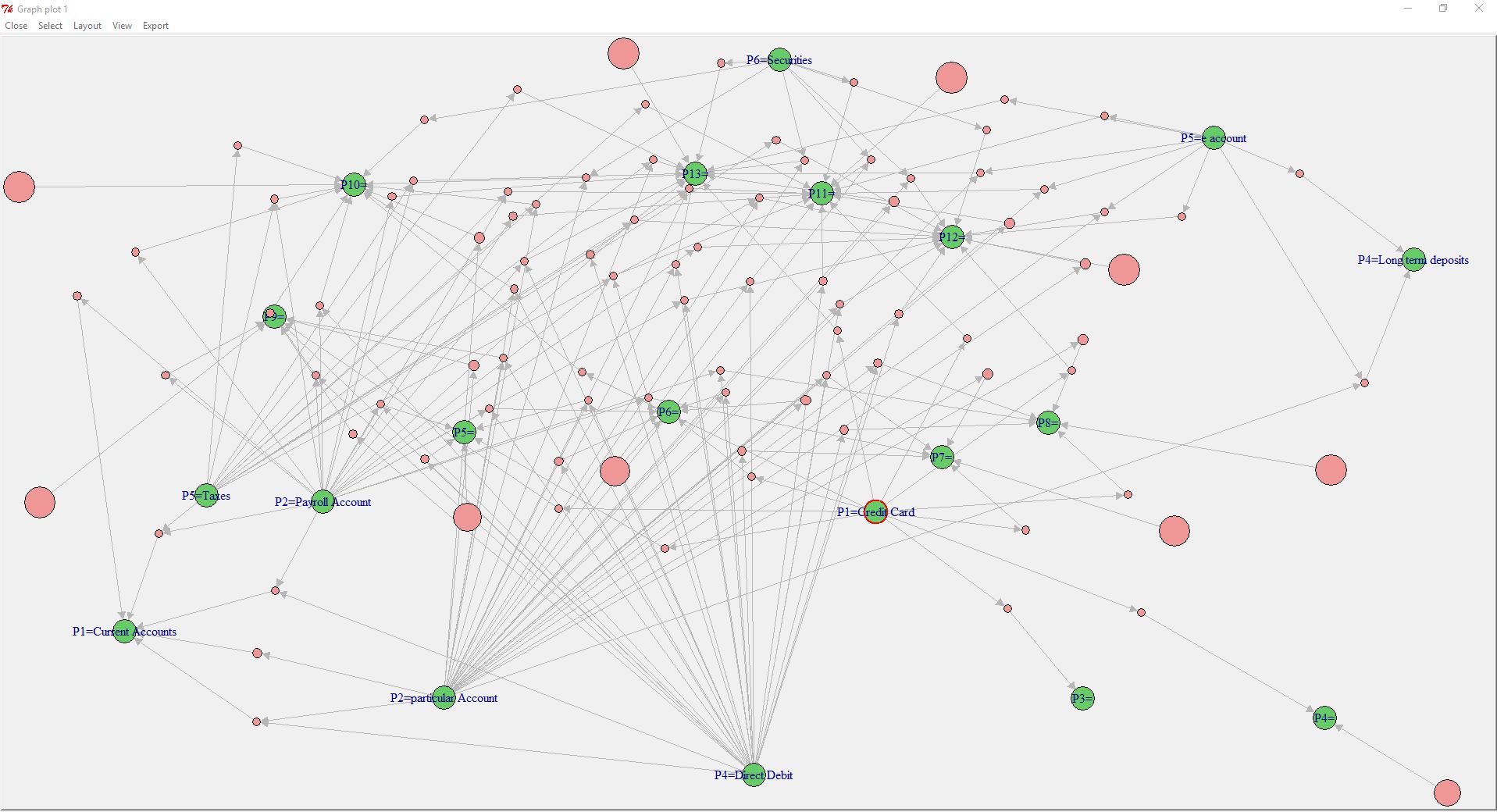


**Association Rules Plots:**

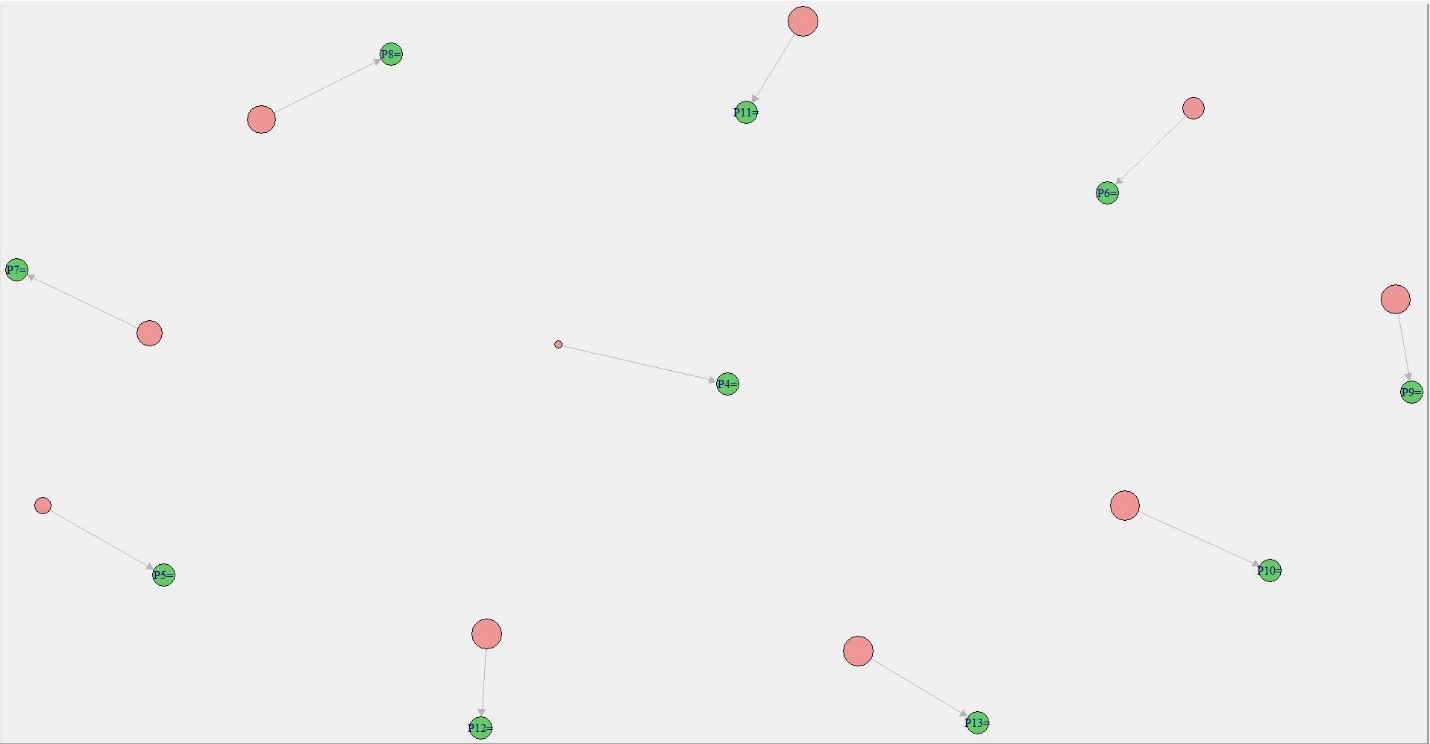
**Cluster 1**



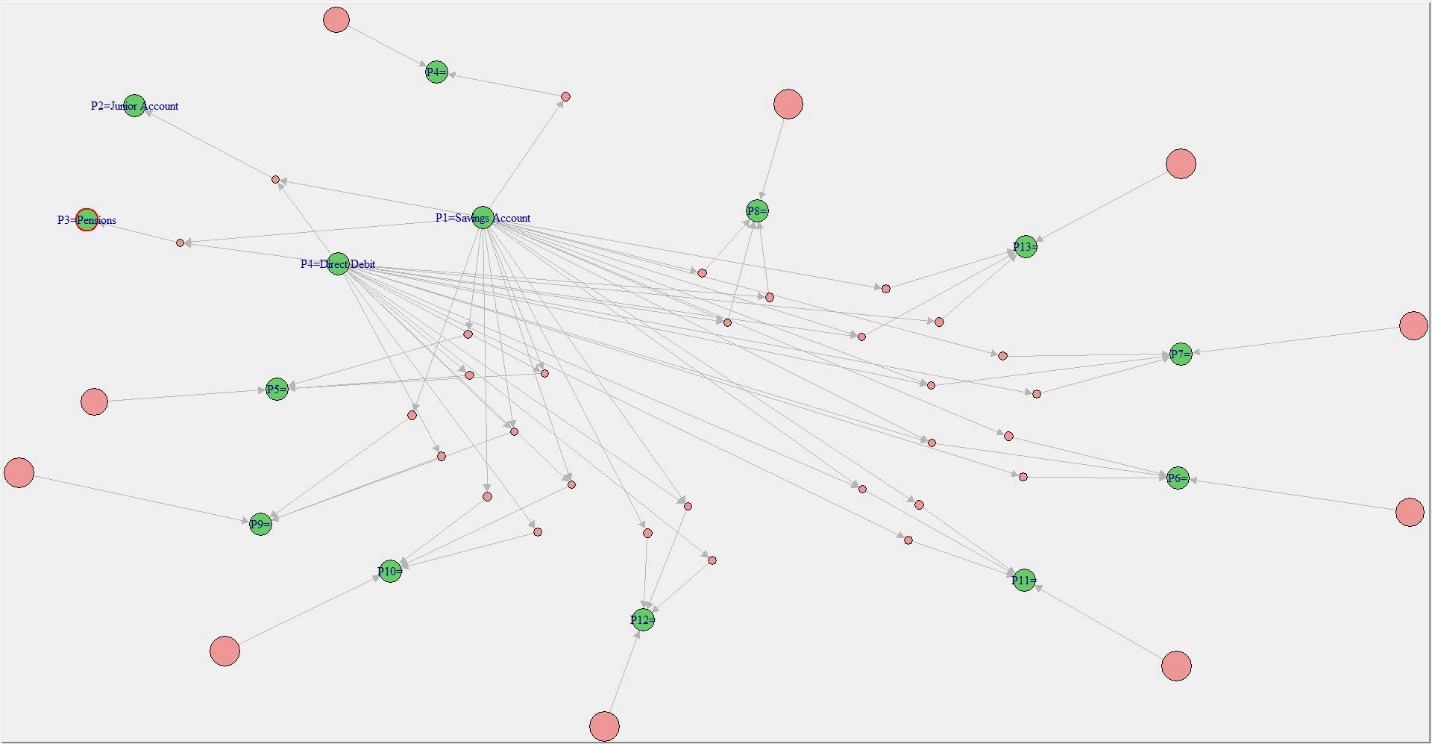
**Cluster 2**

****

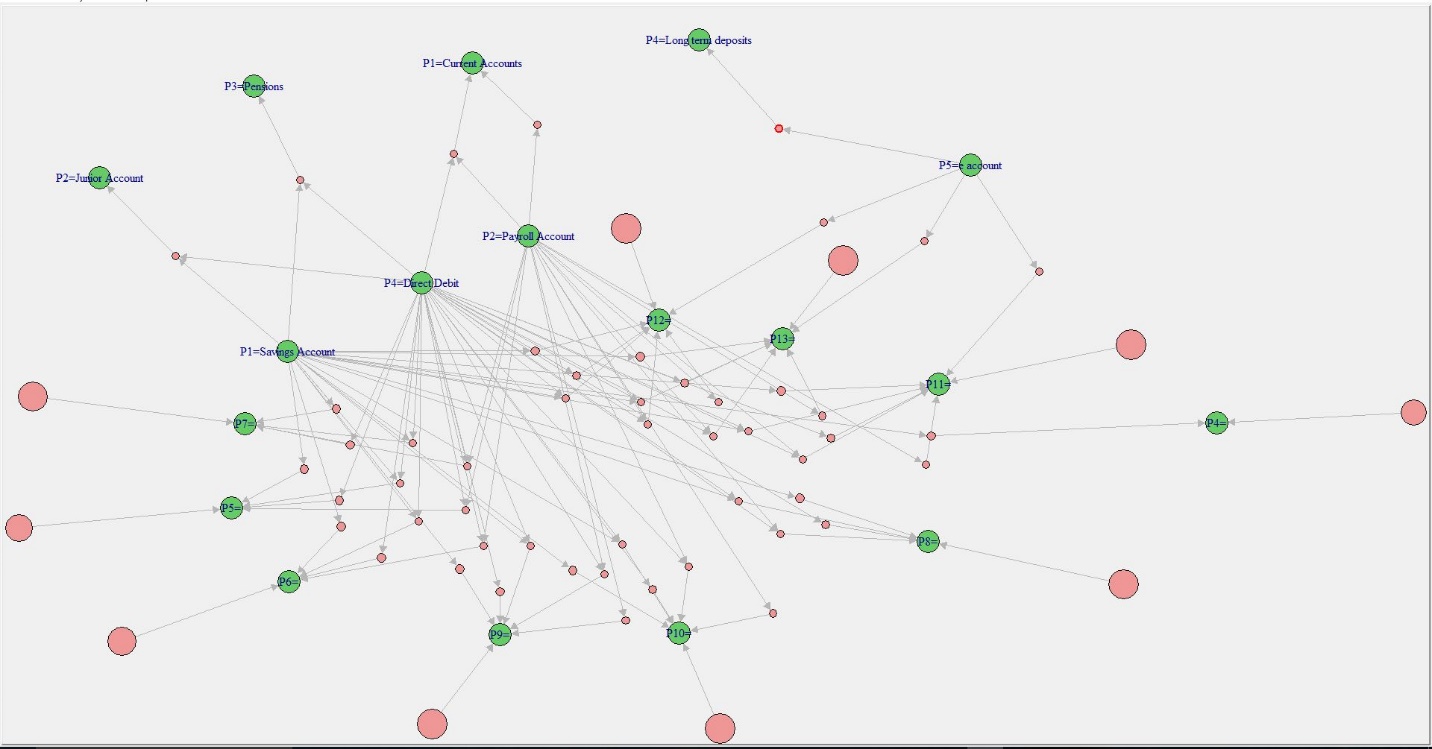
**Cluster 3**

****

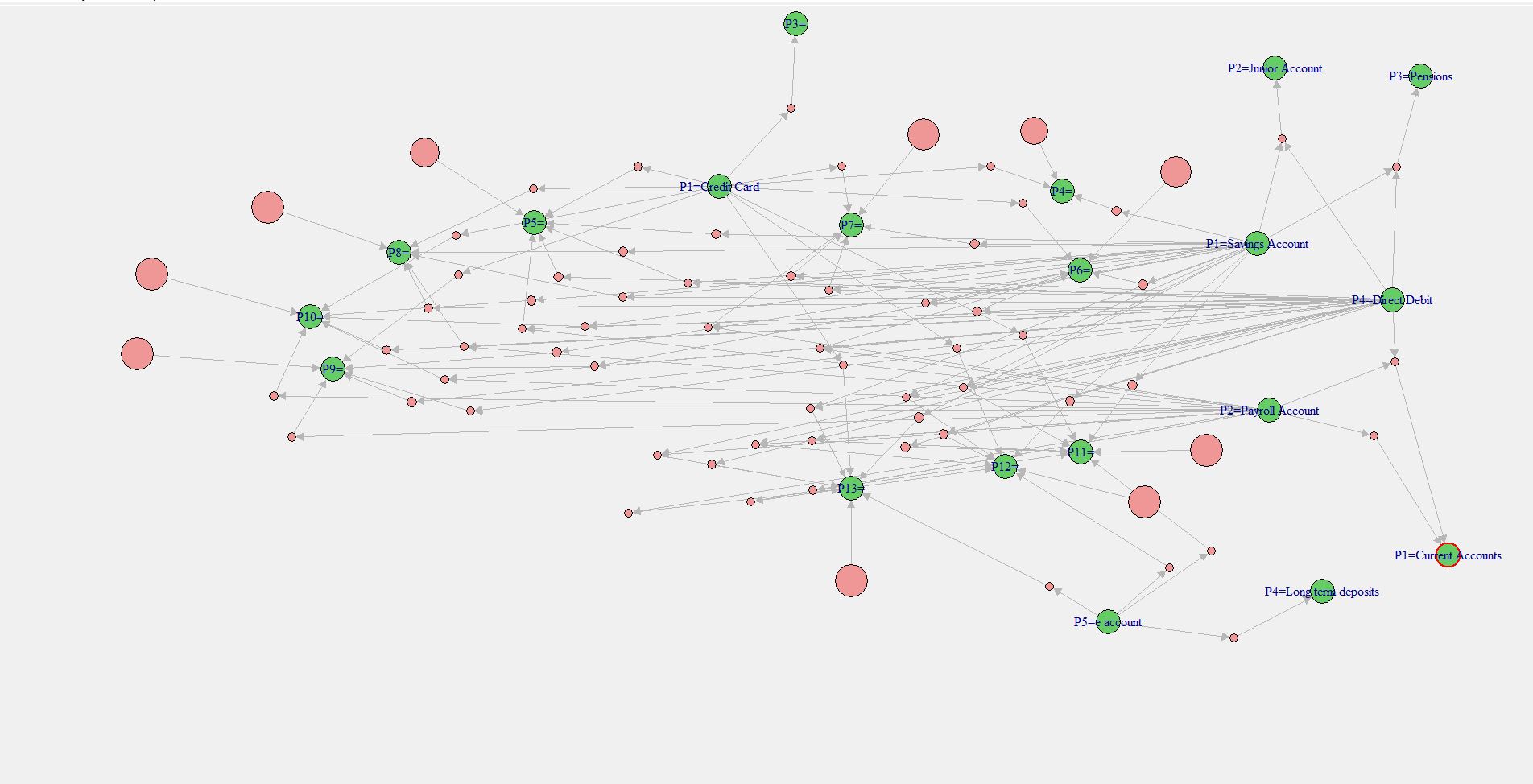
**Cluster 4**

****

**Cluster 5**

****

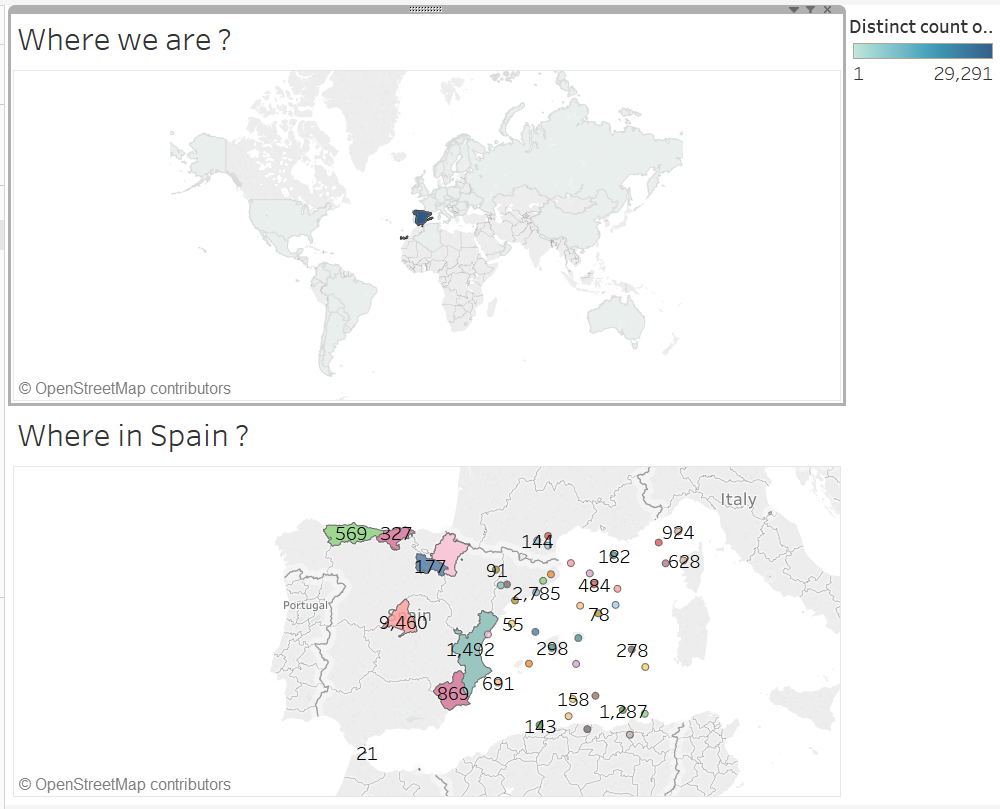
**Cluster 6**

****

**Observation:** From these plots we observed that Cluster 4, 5 and 6 are the key clusters having following profiles of customers.

**Tableau Story Points:**

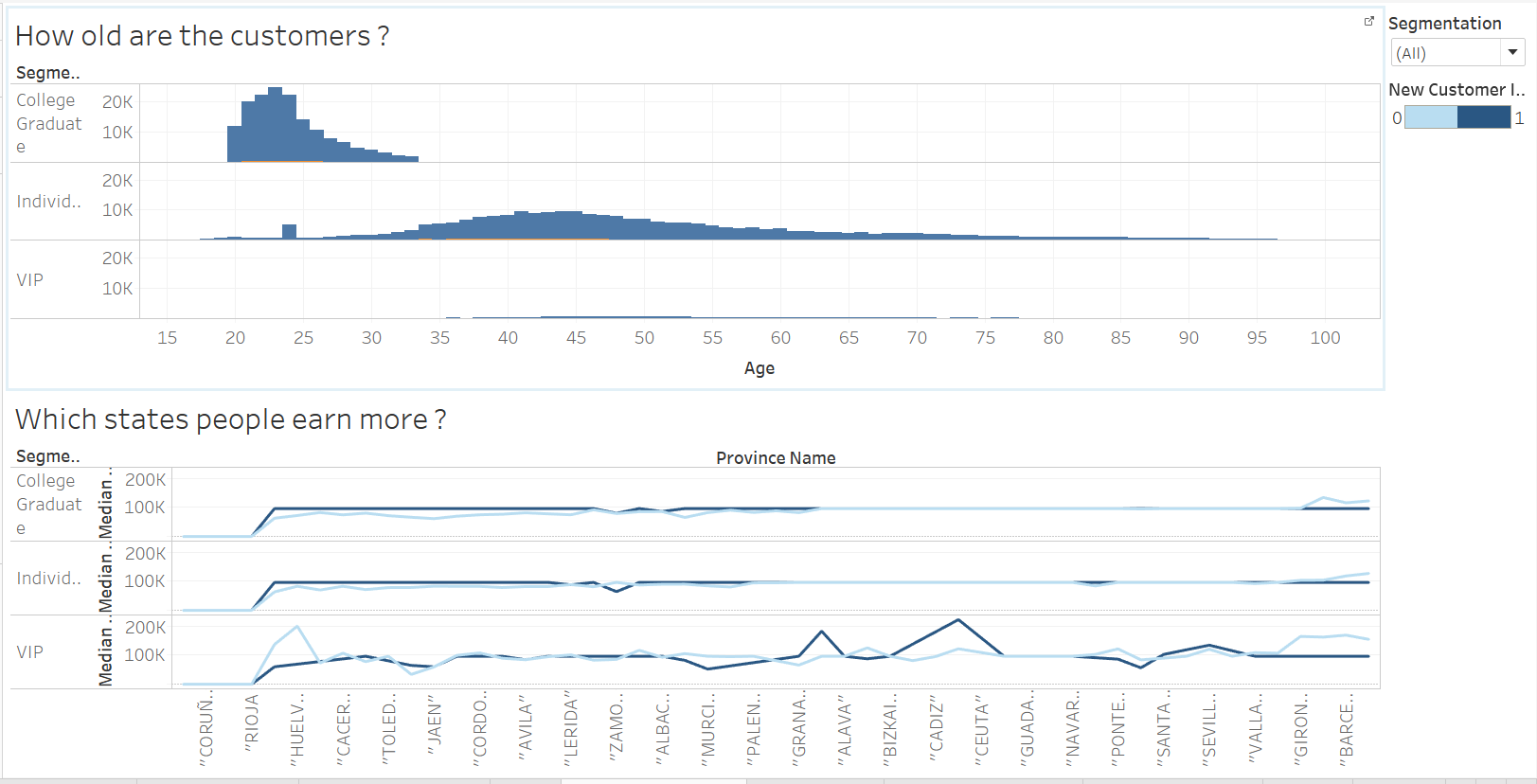
1. **Map**



**Insights –**

**Where we are in the world: We are focusing on Market of Spain and their provinces for Santander Bank.**

1. **Age Distribution**



**Insights –**

**How old are the customers?**

College graduate – the age ranges from 20 – 35 with most of the customers mainly of age 23.

Individual – The age ranges from 18 – 97 with bimodal distribution with 35 – 55. One more observation is that customers of age 24 are more in number and after digging into the data we found that 60 percent of them belongs to capital of Spain.

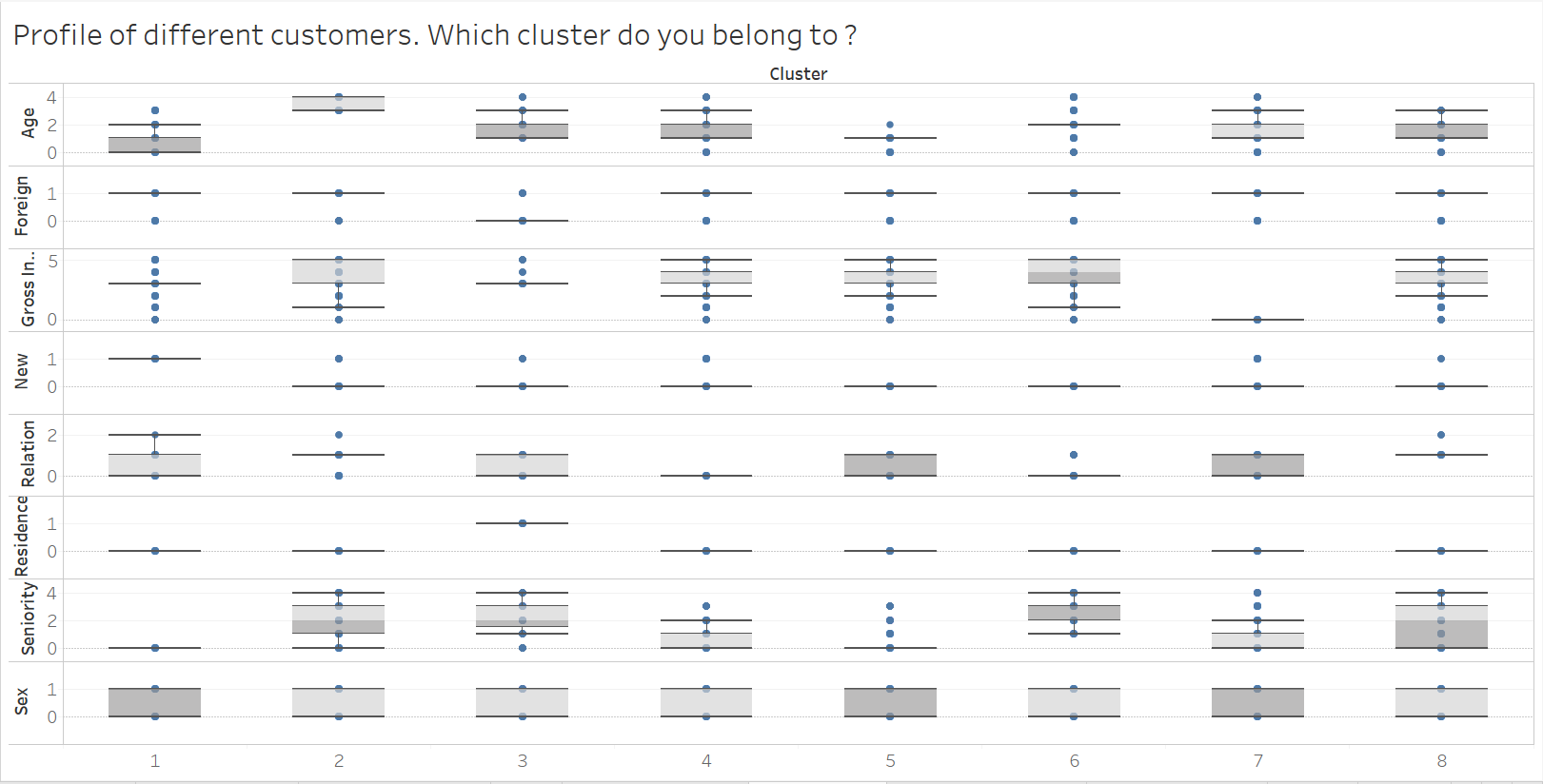
VIP – The age ranges 35 – 77 and the number customers are very less.

**Which state people earn more?**

There are customers who belong from Cantabria and Huelva whose Gross income is more than the trend and they are from VIP segment.

College graduate and Individual earns less than VIP group.

1. **Clustering**

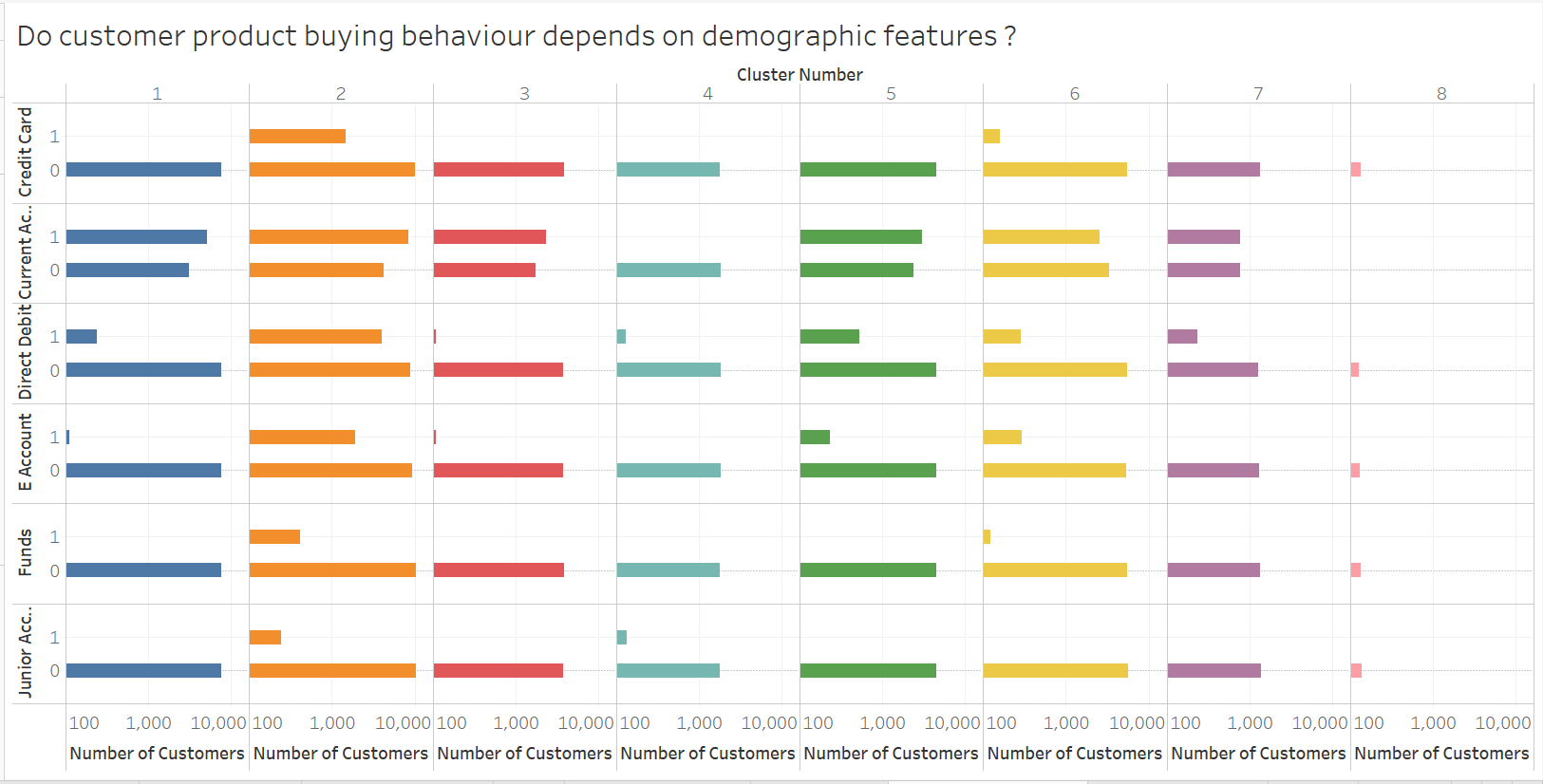


|  |  |  |  |
| --- | --- | --- | --- |
| Variables | Cluster 2 | Cluster 5 | Cluster 6 |
| Age (years) | 61-90 | 21-40 | 41-60 |
| Gross\_income | 60,000 & above | 35,000 – 149,999 | 100,000 – 149,999 |
| New\_Customer\_index | Yes | No | No |
| Customer\_Relation | Inactive | Inactive | Inactive |
| Residency\_index | Yes | Yes | Yes |
| Seniority (months) | 150-199 | 50-99 | 150-199 |
| Sex | Both | Both | Both |

1. Insights
2. In Cluster 2, most of the customers are of age 61-90 whose gross income is more than 60k and they are new customers\* of Santander Bank and their seniority is 150-199 month old. All the customers are from Spain consisting of both male and female.
3. In Cluster 5, most of the customers are of age 21- 40 and their gross income range is 35k to 150k. They are old customers of Santander Bank and their seniority is 50-99 month old. All the customers are from Spain consisting of both male and female.
4. In Cluster 6, most of the customers are of age 41- 60 and their gross income range is 100k to 150k. They are old customers of Santander Bank and their seniority is 150-199 month old. All the customers are from Spain consisting of both male and female.

\*New\_customer\_index – It means they have recently purchased a new product being an old customer.

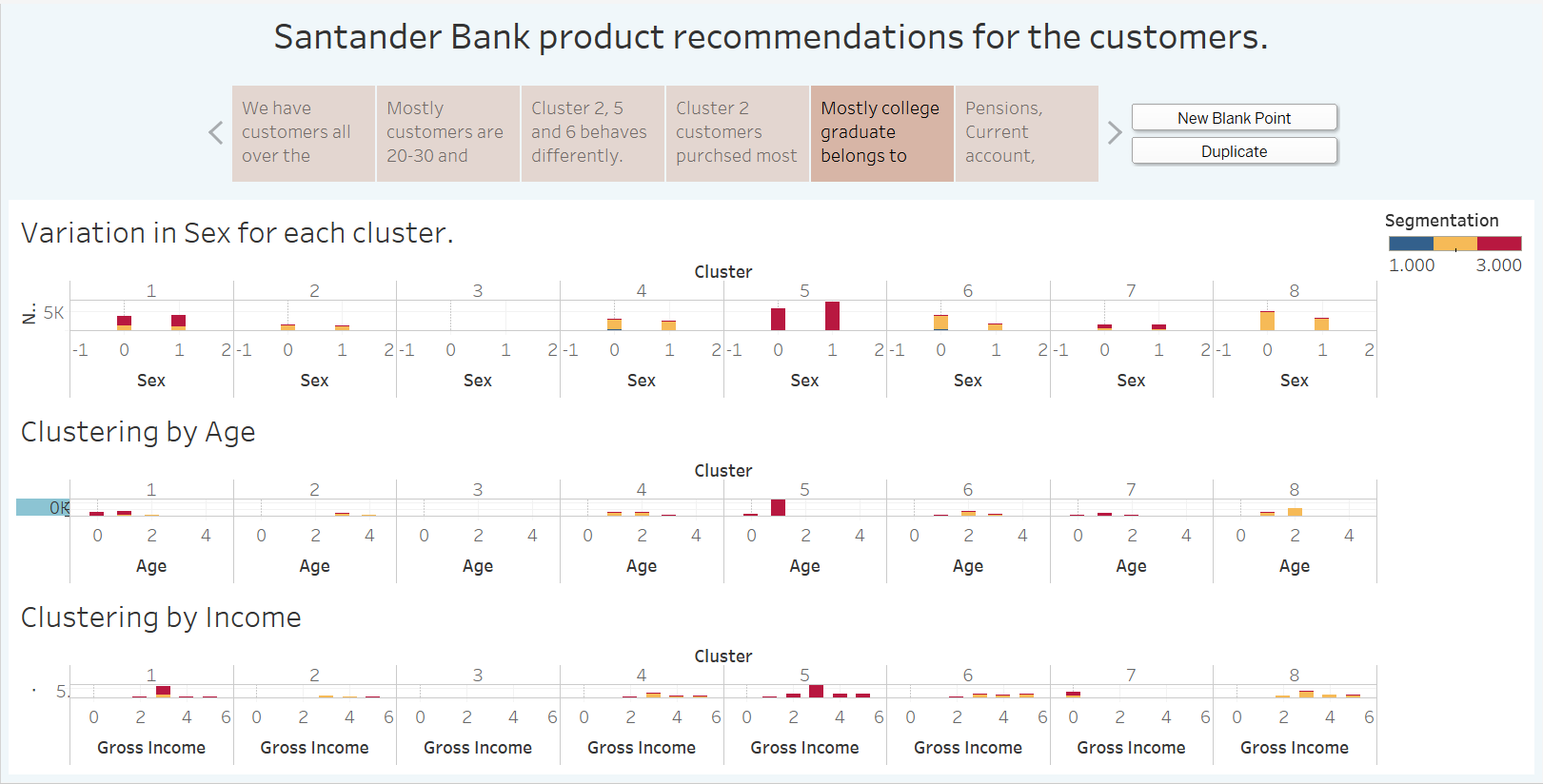
1. **Product Purchase**



**Insight:**

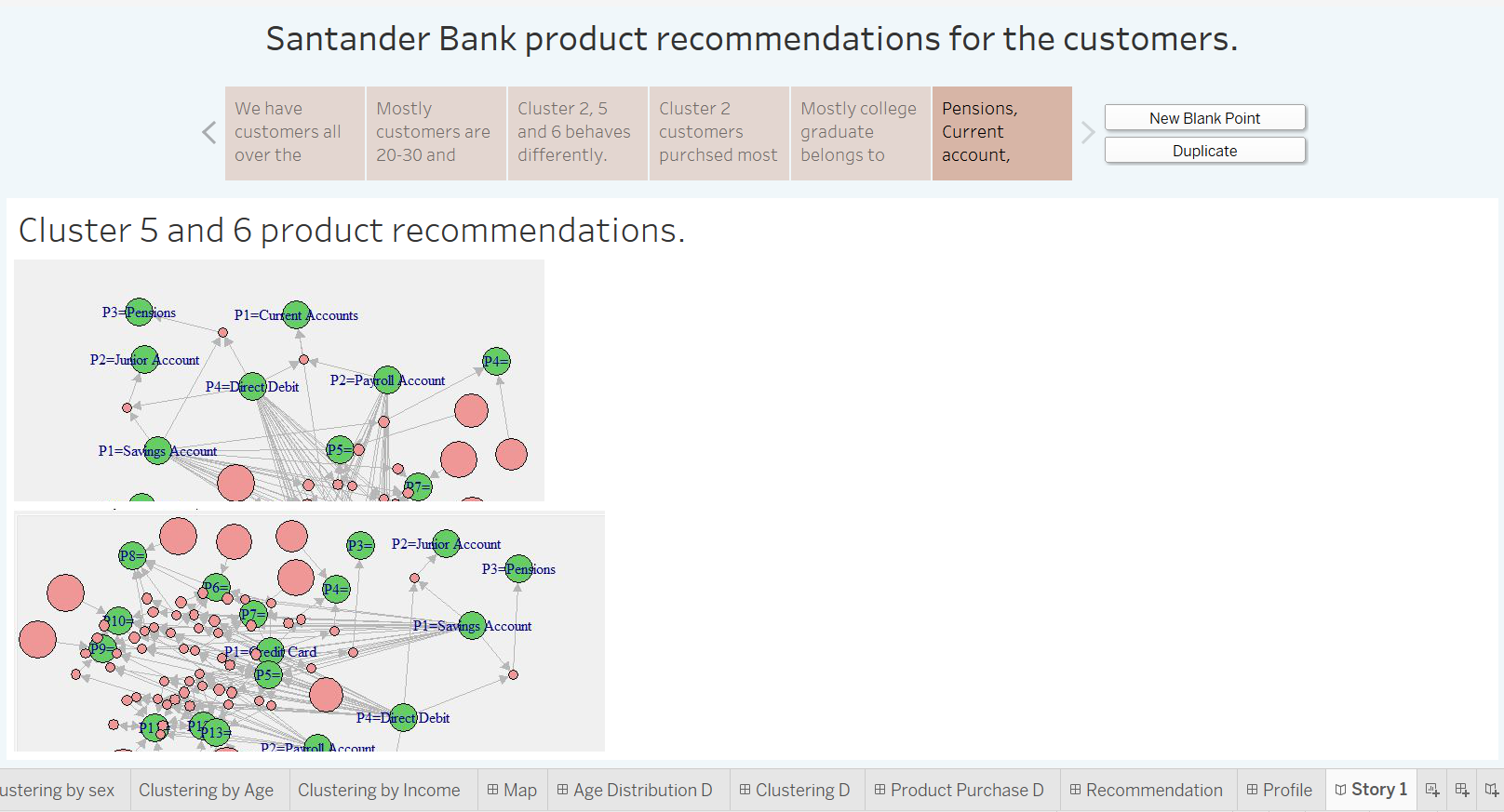
**Cluster 2 customers purchased most of the products followed by Cluster 5 and 6. But, cluster 7 and 8 have not purchased much products.**

1. **Variation of Sex Age Income for different clusters**



**Mostly college graduates belong to cluster 5. They are mostly female and have opened account at young age.**

1. **Cluster 5 and 6 product recommendations**



**We have done Market Basket Analysis which can be seen above.**