



**IE School of Human Sciences & Technology**  
***Master's in Big Data & Business Analytics***

***Course Name:***

**Machine Learning I (Unsupervised Assignment)**

***Supervised by:***

**Prof. Álvaro José Méndez López**

***Submitted by:***

**MANUEL MARINA HERRERA**

## TABLE OF CONTENT

<b>EXECUTIVE SUMMARY .....</b>	<b>1</b>
<b>DATA PREPARATION .....</b>	<b>2</b>
<b>MODELING .....</b>	<b>2</b>
Initial Modeling Phase .....	2
The Final Model .....	3
<b>TARGET CLUSTERS .....</b>	<b>4</b>
<b>RECOMMENDATIONS &amp; CONCLUSIONS .....</b>	<b>4</b>
<b>BIBLIOGRAPHY .....</b>	<b>5</b>
<b>APPENDIX .....</b>	<b>5</b>

## EXECUTIVE SUMMARY

This assignment focuses on finding similar groups of customers in a telecommunication firm. The Marketing Manager provided us a dataset of customers who, at some point, purchased a mobile phone. The dataset is made up of 4431 rows of different customers and their information. The information contained in the dataset is provided by 15 features. Concluding, using this information we segmented the customers into similar groups to find shared characteristics among customers more likely to churn.

The main findings of our research are 1) customers using international calls are unaware of the international discount and 2) churn rates for women are more than three times higher compared to men. After all, people show a mismatch between their phone usage and the contract design. This is the reason why marketing recommendations are made, such as offering more customized packages, changing features of existing packages and increasing the awareness of the international discount.

## DATA PREPARATION

Before starting the clustering, the team decided to perform data manipulation to a) combine information from multiple columns into one and to b) find more variables that might be relevant for finding clusters or to use them as profiling variables later.

The following variables were created:

- **national calls = local + long\_distance:** Local calls and long distance calls are done in the same country, thus are different from international calls.
- **family size: marital + children + 1:** Family size consists of the spouse of the customer, the amount of children and the customer.
- **childrenyesno: 1 = children, 0 = no children:** Change the amount of children to having children or not.
- **income: low (< 25,000), middle (25,000 – 75,000) and high (> 75,000):** Divide the different incomes into three groups (categorization of the variable).
- **6 family types: single, couple, monoparental with one child, monoparental with two children, married with one child and married with two children:** Create six different family types (maximal number of children in the dataset is 2, Formula in Appendix).

After data manipulation, a correlation analysis was performed for all variables in the dataset and the newly created variables to avoid including redundant information. No high correlation between variables was found (Appendix, Figure 1).

## MODELING

### INITIAL MODELING PHASE

We first tried to find variables relevant for finding clusters with high churn rates by running the stepwise-method using churn as the dependent variable, to obtain only relevant variables for linear regression (variables with p-value < 0.05). The table shows 9 variables which are significant: gender, internat, longdist, national\_min, local, est\_inco, age, pay\_Bank, middle income (Appendix, Figure 2).

Running a segmentation with a sample representing 80% of the records and the 4 variables with the lowest p-value (gender, internat, national\_min, est\_inco) returns a model with a good silhouette (0.4)

and clusters with high churn rates (Cluster 0-2) that in total accounted for 85% of churn in the sample (Appendix, Figure 3).

Unfortunately, we were unable to find both useful labels for the clusters and useful insights for the marketing department to run promotions targeted to clusters with high churn rates. This is the reason why the model was dropped.

#### THE FINAL MODEL

The second and final approach was to split the dataset into 80% training and 20% testing data. A trial and error method was followed. For each iteration the most relevant variables were kept and for each new iteration different variables were introduced. After several trials, the team found a model from which valuable insights can be extracted as all clusters show clear differences between churn rates (either very low, <20% or very high, > 70%). In addition, these clusters can be labeled easily.

#### The design of the model was as follows:

General settings: Initially, the model was trained with 80% of the dataset. The “Random (approx. ratio)” sampling method was selected. After, the remaining 20% of the dataset was used for cross-validation.

Main algorithm:

- K-means clustering: The number of clusters were set to 5, 6 and 7 as after several iterations, the team found the best silhouettes and results were obtained with the above mentioned number of clusters.
- Hierarchical clustering: The number of clusters was set to 5, 6 and 7 for the same reasons as mentioned above.
- Interactive clustering: The cluster number was set to 6.

Features handled:

- Gender (0 = male, 1 = female): The variable is relevant for the model as it is the most relevant variable in terms of p-value. Moreover, the churn rate for women is much higher (66.58%) than for men (20.16%).
- Internat: minutes of international calls: Internat is the variable with the second lowest p-value obtained by the stepwise-method.
- Int\_disc: (1 = international discount used, 0 = international discount not used)
- Marital: (1 = married, 0 = not married).
- Features rescaling: AVG-STD rescaling (there is no need to manage missing values as there are none).

The outlier threshold was set to 2%.

The results for the sample simulation were promising as it had a silhouette of 0.77 and several clusters with a high churn rate. Therefore, the team decided to apply the model to the whole dataset. Choosing this model with the previously mentioned variables resulted in 7 + 1 clusters (outlier cluster) with a silhouette of 0.77, well-balanced, and with a high/low churn rate differentiation in all of them (Appendix, Figure 5). The team found 4 clusters (Appendix, Figure 6 - 9) where the number of customers churning is higher than the number of customers staying. These clusters were called “Target Clusters” as these are the clusters to target for future marketing campaigns to reduce churn rates. The target clusters account for 81.71% of total churn. The table can be found in the appendix.

All the variables had an importance above 15% (int\_disc 36%, gender 27%, marital 24%, Internat 16%) and did not indicate high correlation. (Appendix, figure 4)

After checking the usability of the model for our goal, the last step was finding the correct labels for each Target Cluster. The team examined the characteristics of each segmentation feature per cluster (Appendix, Figure 11 - 20). Profiling variables were introduced to obtain more information and insights into each cluster. The profiling variables used are age, children, familytype, est\_inco, pay\_CreditCard, billtype and local.

## TARGET CLUSTERS

### *Cluster 1: Family Women Overpaying.*

This cluster is formed strictly by women as gender mean is 1. Most customers in this cluster have a family: 70% have children, 61% are married and 25% are married with one child (family type appendix). The biggest age group is 70-75 years old (11%), with an international discount plan, without hardly making any international calls. Therefore, it can be assumed, this is a cluster of clients with a family that are overpaying for services they do not use.

### *Cluster 2: High-Medium Class Married Women.*

Cluster 2 mostly contains married middle-aged women, with an age-peak (10%) between 55 and 60 years, with medium to high income. The customers in this cluster have children in 68% of cases (one-child family 32%, two-child family 35%). The majority does not use international discounts and pays with credit card frequently (61%).

### *Cluster 5: Single Women with Descendants.*

The underlying cluster contains unmarried women with descendants in 63% of the cases. Therefore it is assumed they form a mono-parental family. The biggest age group in this cluster is 75-90 years old so they might have been divorced or widowed, which could serve as an explanation for the high share of mono-parental families in this cluster. Moreover, no customer uses the international discount and the cluster's preferred payment method is credit card. As they can be located in the medium-high income group, they are interesting from a marketing perspective.

### *Cluster 6: Uninformed International Female Callers.*

The last target cluster presents a higher percentage of women (64%) with a medium-high income. 59% of the cluster members are married, 62% have children (with a varied family type distribution). The most important and interesting characteristic of this cluster is that they are heavy users of international calls but without having an international discount.

### *Outliers Cluster: Business Men.*

The Outliers Cluster consists of men, who use international calls and the international discount. Moreover, they call more (more average minutes per day) in comparison with other clusters. In addition, most of the Outliers use the budget package.

## RECOMMENDATIONS & CONCLUSIONS

Our segmentation analysis enabled the company to explain 1581 out of 1935 total records (81%) of churning for the telecom company. Taking a step back, we can say that this company's dataset has a churning rate of 43.6%, estimated to be double of the telecom industry's churn rate annually (Statista, 2020), assuming this company operates in North America or Europe. Lastly, our recommendations to reduce churning are made without insight on profitability, which should guide decision-making.

Our first recommendation is to collect more data that we assume has a high likelihood to explain churning and was not taken into account for our analysis. In order to give better recommendations, we would like to obtain data that could explain possible customer dissatisfaction such as, the frequency of interactions with customer service, amount of open tickets per user, average time spent to deal with issues, time spent at the company and more should be collected. Additionally, information on the specifics of subscription packages such as the amount of GB of data included, possible unlimited calls and price structures may be relevant to increase the depth of our analysis.

*Cluster 1: Women Overpaying.*

These women seem to be paying for a service they do not use. Therefore, either removing the international discount or including more benefits in the package could enable them to change their feelings and perception towards the company. Moreover, since this cluster seems to correspond to a family group, these clients could be targeted with a potentially personalized family package.

*Cluster 2: Medium-High Middle Class Married Women*

As the data does not allow us to infer why the group may churn, we would recommend the telecom company to gather additional data. Our hypothesis is that this cluster may be dissatisfied with pricing or/and customer service. It could be interesting to consider a promotion/package for retired people and people near retirement with special customer service. More data is needed.

*Cluster 5: Single Women with Descendants.*

One noticeable element about this cluster is the high proportion of people above 50, particularly between 70 and 90. Our hypothesis is that cluster 5 may churn because their descendants are recommending them to switch companies, frustrated about our customer service or product offering. Similarly to cluster 2, it could be interesting to consider a promotion/package for retired people and people near retirement with special customer service. More data is needed.

*Cluster 6: Uninformed International Female Callers.*

These users make a lot of international calls but don't have any discounts. Therefore it can be assumed that they don't know about the international discount. Improving communication about the international discount may help decrease the churn rate.

*Outliers Cluster: Business Men.*

Higher time spent on calls than all other clusters, using international discount and performing international calls, yet churning. Our hypothesis is that churning may come from bad customer service, or that our service/billtype doesn't fit their needs. We recommend gathering more data and to explore promotions or updates regarding the billtype called Budget, used by about  $\frac{2}{3}$  of them.

Our segmentation analysis shows that clusters can be created out of socioeconomic variables, usage of discounts and usage of service. Nevertheless, to have more meaningful insights and recommendations we require additional data, in particular about customer service and package contents.

## BIBLIOGRAPHY

<https://puresoftware.com/application-testing-for-a-global-telecom-company/>, accessed 27/11/2021.

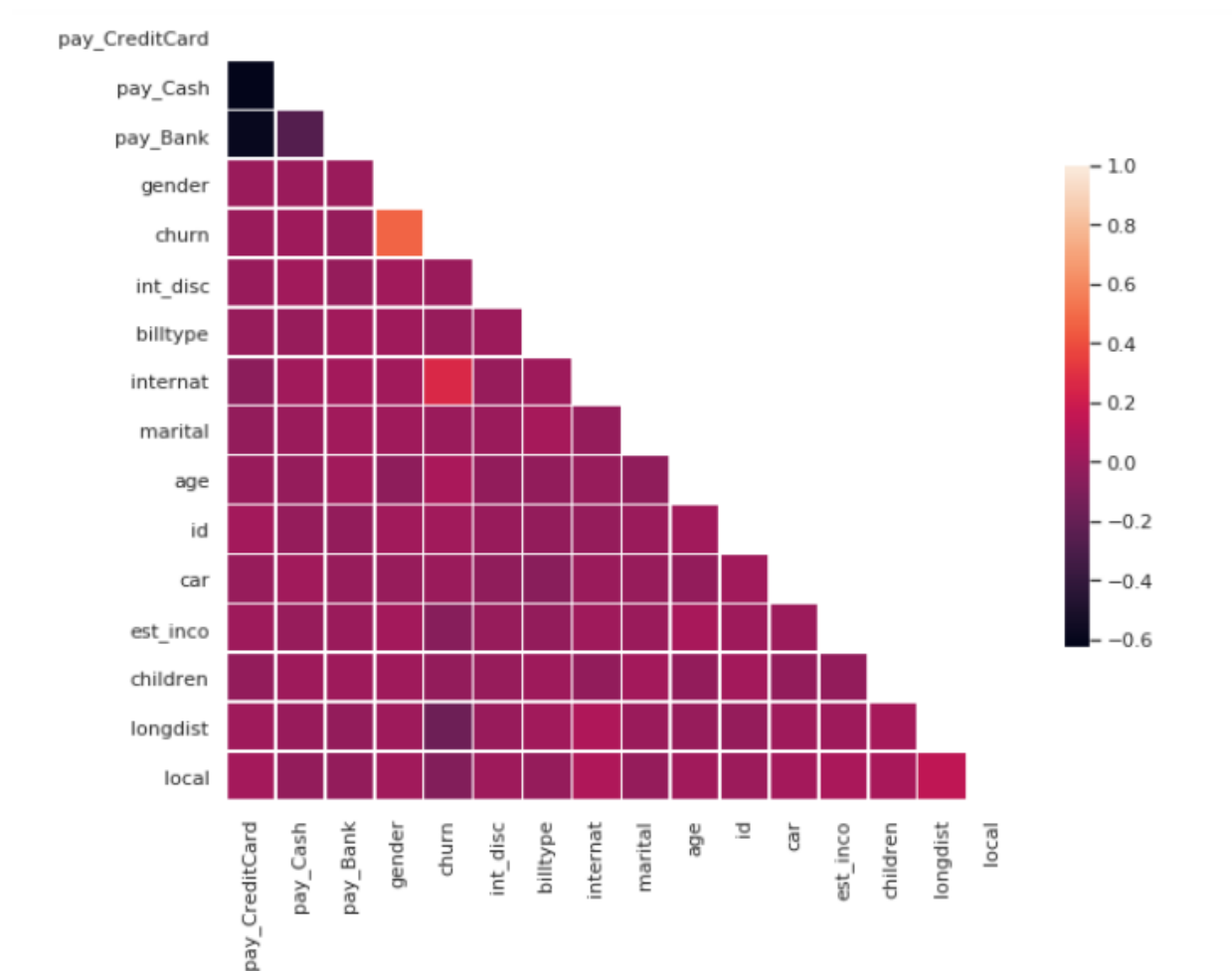
Statista. "Customer churn rate in the US in 2020 by Industry." Statista.com, <https://www.statista.com/statistics/816735/customer-churn-rate-by-industry-us/>, accessed 28/11/2021.

## APPENDIX

### *Formula 1: Data manipulation*

```
if(and(marital==0,children==1),"one-child-  
monoparental",if(and(marital==1,children==0),"couple",if(and(marital==1,children==1),"one-  
child-family",if(and(marital==1,children==2),"two_child-  
family",if(and(marital==0,children==2),"two-child-  
monoparental",if(and(marital==0,children==0),"single",0))))))
```

*Figure 1: Correlation analysis*



**Figure 2: Linear regression output**

```
In [40]: * # Fills X (removing variables with missing values)
selected_fields=df.drop(labels=["churn"],axis=1)
# Sets y
y = df['churn']

result = stepwise_selection(selected_fields, y)

Add gender with p-value 6.6162e-240
Add internat with p-value 8.68782e-81
Add longdist with p-value 7.85959e-55
Add national_min with p-value 7.36047e-14
Add local with p-value 3.56309e-28
Add est_inco with p-value 8.83769e-12
Add age with p-value 1.22133e-11
Add pay_Bank with p-value 0.00359815
Add mediumincomefamily with p-value 0.00682506

In [41]: print('resulting features:')
print(result)

resulting features:
['gender', 'internat', 'longdist', 'national_min', 'local', 'est_inco', 'age', 'pay_Bank', 'mediumincomefamily']
```

**Figure 3: First segmentation clusters**

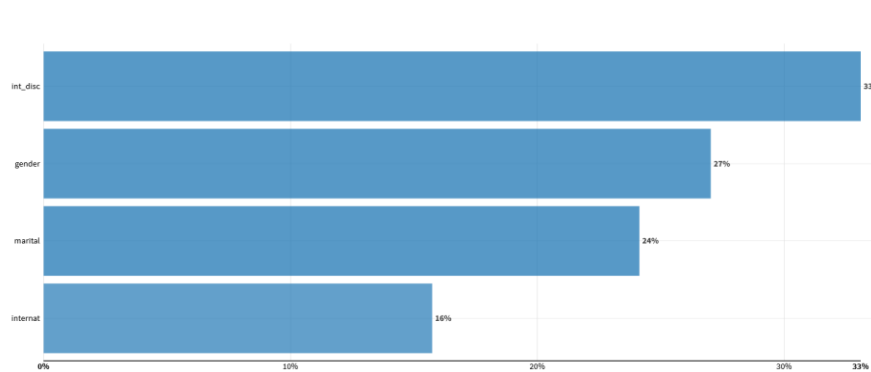
Run: In DSS

Count by cluster\_labels and churn

3587 records

	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4	cluster_5	cluster_outli...
0-0.5	293	260	87	510	398	468	37
0.5-1.0	563	425	300	139	18	59	30
	856	685	387	649	416	527	67

**Figure 4: General Feature information**



**Figure 5: Second segmentation clusters**

Run: In DSS

Count by cluster\_labels and churn

	cluster_3	cluster_2	cluster_1	cluster_0	cluster_4	cluster_5	cluster_6	cluster_outli...
0-0.5	675	294	243	507	495	180	75	27
0.5-1.0	141	507	432	84	84	354	288	45
	816	801	675	591	579	534	363	72

**Figure 6: Cluster 1**



### Family women overpaying

Family women with childrens, int\_discount and with an age-pick between 70-75

#### Observations

- **int\_disc** is in average **222% greater** : mean of 1 against 0.31 globally
- **gender** is in average **97.46% greater** : mean of 1 against 0.51 globally
- **internat** is in average **70.28% smaller** : mean of 0.25 against 0.84 globally

**Figure 7: Cluster 2**

### High-medium class married women

Women married with childrens between 55-60 years old medium high class

#### Observations

- **gender** is in average **97.46% greater** : mean of 1 against 0.51 globally
- **marital** is in average **68.61% greater** : mean of 1 against 0.59 globally
- The mean of int\_disc is 0 against 0.31 across all clusters

**Figure 8: Cluster 5**

### Single women with descendants

Single women with children (62%), with a slight age-pick between 75 and 90 years old, and with medium-high income

#### Observations

- The mean of marital is 0 against 0.59 across all clusters
- **gender** is in average **97.46% greater** : mean of 1 against 0.51 globally
- The mean of int\_disc is 0 against 0.31 across all clusters

**Figure 9: Cluster 6**

### Uninformed international female callers

Mainly women international callers without discount

#### Observations

- **internat** is in average **762% greater** : mean of 7.21 against 0.84 globally
- **int\_disc** is in average **54.79% smaller** : mean of 0.14 against 0.31 globally
- **gender** is in average **27.29% greater** : mean of 0.64 against 0.51 globally

**Figure 10: Cluster Outliers**

### Informed International business men callers


Men international callers with discount

#### Observations

- **internat** is in average **750% greater** : mean of 7.11 against 0.84 globally
- **int\_disc** is in average **222% greater** : mean of 1 against 0.31 globally
- The mean of gender is 0 against 0.51 across all clusters

**Figure 11: Marital vs Custer\_labels**


Run: In DSS

Count by cluster\_labels and local 

	cluster_3	cluster_2	cluster_1	cluster_0	cluster_4	cluster_5	cluster_6	cluster_outli...
0-0.5	0%	0%	39%	100%	39%	100%	41%	54%
0.5-1.0	100%	100%	61%	0%	61%	0%	59%	46%

**Figure 12: Children vs Cluster\_labels**

Run: In DSS

Count by cluster\_labels and childrens 

	cluster_3	cluster_2	cluster_1	cluster_0	cluster_4	cluster_5	cluster_6	cluster_outli...
0-0.7	35%	32%	30%	34%	39%	37%	38%	38%
0.7-1.3	28%	33%	38%	30%	30%	32%	33%	29%
1.3-2.0	36%	35%	32%	36%	31%	31%	29%	33%

**Figure 13: Age vs Cluster\_labels**

Run: In DSS

Count by cluster\_labels and Age 

	cluster_3	cluster_2	cluster_1	cluster_0	cluster_4	cluster_5	cluster_6	cluster_outli...
15-20	3%	4%	3%	2%	3%	1%	4%	0%
20-25	5%	8%	6%	5%	7%	5%	4%	8%
25-30	6%	4%	6%	8%	6%	8%	7%	4%
30-35	5%	7%	9%	5%	6%	4%	8%	4%
35-40	6%	6%	5%	9%	3%	6%	5%	4%
40-45	6%	6%	4%	4%	6%	8%	8%	8%
45-50	7%	7%	8%	6%	8%	3%	7%	4%
50-55	7%	6%	7%	6%	4%	6%	4%	8%
55-60	7%	10%	7%	7%	8%	5%	7%	13%
60-65	7%	4%	8%	8%	11%	8%	7%	4%
65-70	8%	6%	8%	6%	3%	4%	6%	17%
70-75	4%	2%	11%	6%	6%	7%	8%	8%
75-80	7%	9%	4%	6%	6%	9%	7%	4%
80-85	6%	6%	6%	6%	7%	8%	6%	4%
85-90	4%	6%	3%	8%	7%	9%	2%	0%
90-95	7%	6%	5%	8%	5%	5%	6%	4%
95-100	6%	2%	0%	3%	5%	3%	5%	4%
100-105	0%	0%	0%	0%	0%	0%	0%	0%


**Figure 14: Family type vs Cluster\_labels**

Run: In DSS


Count by cluster\_labels and FamilyType 

	cluster_3	cluster_2	cluster_1	cluster_0	cluster_4	cluster_5	cluster_6	cluster_outli...
two_child-family	36%	35%	19%	0%	21%	0%	20%	13%
couple	35%	32%	17%	0%	23%	0%	21%	17%
one-child-family	28%	33%	25%	0%	17%	0%	17%	17%
single	0%	0%	13%	34%	16%	37%	17%	21%
one-child-mon...	0%	0%	13%	30%	13%	32%	16%	13%
two-child-mon...	0%	0%	13%	36%	10%	31%	9%	21%


**Figure 15: Est\_inco vs Cluster\_labels**

Run: In DSS		Count by cluster_labels and est_inco 						
	cluster_3	cluster_2	cluster_1	cluster_0	cluster_4	cluster_5	cluster_6	cluster_outli...
110-25k	22%	25%	23%	28%	29%	21%	24%	17%
25k-50k	25%	21%	28%	25%	23%	23%	19%	21%
50k-75k	28%	27%	27%	23%	22%	29%	29%	42%
75k-100k	25%	27%	22%	23%	26%	26%	28%	21%


**Figure 16: Pay\_CreditCard vs Cluster\_labels**

Run: In DSS		Count by cluster_labels and pay_CreditCard 						
	cluster_3	cluster_2	cluster_1	cluster_0	cluster_4	cluster_5	cluster_6	cluster_outli...
0-0.5	44%	39%	45%	39%	40%	38%	54%	33%
0.5-1.0	56%	61%	55%	61%	60%	62%	46%	67%

**Figure 17: Local vs Cluster\_labels (% and total records)**

Run: In DSS		Count by cluster_labels and local 						
	cluster_3	cluster_2	cluster_1	cluster_0	cluster_4	cluster_5	cluster_6	cluster_outli...
0-23	35%	40%	41%	38%	40%	37%	26%	17%
23-45	26%	21%	16%	23%	22%	24%	17%	29%
45-68	15%	15%	15%	13%	12%	10%	23%	17%
68-90	9%	10%	11%	11%	7%	12%	15%	13%
90-113	5%	5%	6%	6%	7%	4%	6%	8%
113-135	2%	3%	3%	5%	4%	3%	5%	13%
135-158	3%	2%	3%	2%	2%	4%	2%	4%
158-180	1%	1%	1%	1%	3%	2%	1%	0%
180-203	1%	1%	0%	1%	1%	1%	2%	0%
203-225	1%	0%	0%	1%	0%	1%	2%	0%
225-248	0%	0%	1%	0%	1%	0%	0%	0%
248-270	0%	0%	0%	1%	0%	1%	0%	0%
270-293	0%	0%	1%	0%	0%	0%	1%	0%
293-315	0%	0%	0%	0%	1%	1%	1%	0%
315-338	0%	0%	0%	0%	0%	0%	0%	0%
338-361	0%	0%	0%	0%	0%	0%	0%	0%
361-383	0%	0%	0%	0%	0%	0%	0%	0%
383-406	0%	0%	0%	0%	0%	1%	0%	0%
406-428	0%	0%	0%	0%	0%	0%	0%	0%
428-451	0%	0%	0%	0%	0%	1%	0%	0%


Run: In DSS

Count by cluster\_labels and local 

	cluster_3	cluster_2	cluster_1	cluster_0	cluster_4	cluster_5	cluster_6	cluster_outli...	
0-23	288	318	276	225	234	195	93	12	1641
23-45	216	168	111	135	126	126	60	21	963
45-68	120	117	102	78	69	54	84	12	636
68-90	75	81	75	63	39	63	54	9	459
90-113	42	39	39	36	42	21	21	6	246
113-135	18	24	21	27	24	18	18	9	159
135-158	21	15	18	9	12	24	6	3	108
158-180	12	12	9	6	18	12	3	0	72
180-203	6	9	3	6	6	3	9	0	42
203-225	6	3	0	3	0	3	9	0	24
225-248	0	3	9	0	6	0	0	0	18
248-270	3	3	3	3	0	6	0	0	18
270-293	3	3	6	0	0	0	3	0	15
293-315	3	3	0	0	3	3	3	0	15
315-338	3	3	0	0	0	0	0	0	6
338-361	0	0	0	0	0	0	0	0	0
361-383	0	0	0	0	0	0	0	0	0
383-406	0	0	0	0	0	3	0	0	3


**Figure 18: Internat vs Cluster\_labels**

Run: In DSS

Count by cluster\_labels and internat 


	cluster_3	cluster_2	cluster_1	cluster_0	cluster_4	cluster_5	cluster_6	cluster_outli...
0-0.7	94%	93%	91%	95%	96%	96%	0%	0%
0.7-1.3	1%	2%	2%	2%	2%	2%	0%	0%
1.3-2.0	1%	2%	1%	1%	1%	1%	0%	0%
2.0-2.7	0%	0%	2%	1%	1%	1%	0%	0%
2.7-3.3	2%	1%	1%	1%	1%	2%	0%	0%
3.3-4.0	1%	1%	1%	1%	0%	0%	0%	0%
4.0-4.6	1%	0%	2%	1%	0%	0%	5%	8%
4.6-5.3	0%	0%	0%	0%	0%	0%	9%	8%
5.3-6.0	0%	0%	0%	0%	0%	0%	7%	13%
6.0-6.6	0%	0%	0%	0%	0%	0%	19%	17%
6.6-7.3	0%	0%	0%	0%	0%	0%	11%	8%
7.3-8.0	0%	0%	0%	0%	0%	0%	15%	13%
8.0-8.6	0%	0%	0%	0%	0%	0%	12%	0%
8.6-9.3	0%	0%	0%	0%	0%	0%	12%	17%
9.3-10.0	0%	0%	0%	0%	0%	0%	11%	17%

**Figure 19: National\_min vs Cluster\_labels**

Run: In DSS Count by cluster\_labels and national\_min 

	cluster_3	cluster_2	cluster_1	cluster_0	cluster_4	cluster_5	cluster_6	cluster_outli...
0-24	22%	23%	21%	20%	20%	21%	7%	0%
24-48	27%	28%	26%	29%	33%	29%	23%	33%
48-72	21%	17%	18%	19%	17%	13%	21%	13%
72-96	13%	13%	13%	13%	8%	15%	22%	21%
96-120	6%	8%	10%	9%	6%	7%	12%	13%
120-144	4%	3%	5%	4%	7%	5%	6%	13%
144-168	3%	3%	1%	3%	2%	4%	2%	8%
168-192	1%	1%	3%	2%	4%	2%	2%	0%
192-216	0%	1%	0%	1%	2%	1%	2%	0%
216-240	1%	0%	0%	1%	1%	1%	2%	0%
240-264	1%	1%	1%	0%	0%	0%	1%	0%
264-288	0%	0%	0%	1%	1%	1%	0%	0%
288-312	1%	1%	0%	0%	0%	0%	2%	0%
312-336	0%	0%	0%	0%	0%	1%	0%	0%
336-360	0%	0%	0%	0%	1%	0%	0%	0%
360-384	0%	0%	0%	0%	0%	0%	0%	0%
384-408	0%	0%	0%	0%	0%	0%	0%	0%
408-432	0%	0%	0%	0%	0%	1%	0%	0%
432-456	0%	0%	0%	0%	0%	0%	0%	0%
456-480	0%	0%	0%	0%	0%	1%	0%	0%

**Figure 20: BillType vs Cluster\_labels**

Run: In DSS		Count by cluster_labels and billtype 						
	cluster_3	cluster_2	cluster_1	cluster_0	cluster_4	cluster_5	cluster_6	cluster_outli...
0-0.5	50%	46%	48%	51%	52%	53%	47%	33%
0.5-1.0	50%	54%	52%	49%	48%	47%	53%	67%