[Customer Segmentation Clustering Project]

By [Bingtao Zou]

with support from

[Professor Mahesh Nagarajan]

[August 23, 2023]

EXECUTIVE SUMMARY

In today's data-driven business landscape, effective customer segmentation is pivotal. This project focuses on implementing customer segmentation within the banking sector through advanced data analytics. By leveraging machine learning techniques, I aimed to enhance marketing campaign precision, understand customer behavior deeply, and drive targeted initiatives for a leading bank.

The project was motivated by a prior internship at a Chinese bank, where I observed significant customer engagement variations among different age groups. Younger clients (18-35) had distinct preferences and behaviors, yet the bank lacked tailored marketing, resulting in below-average engagement. This highlighted the critical need for customer segmentation.

Utilizing the Bank Marketing dataset, I applied the k-prototypes clustering algorithm to factor in both categorical and numeric variables. I identified seven distinct customer clusters, offering opportunities for targeted marketing initiatives. These clusters ranged from elderly, retired clients with a high affinity for term deposits to younger, educated professionals with varied preferences.

This project provides actionable insights for the bank. Tailoring marketing campaigns to each cluster's specific needs and preferences can lead to improved outcomes, enhanced customer engagement, and satisfaction. However, customer segmentation is dynamic and requires ongoing refinement in response to market and customer behavior changes.

In conclusion, this project represents a meaningful attempt at revamping marketing strategies in banking. The granular customer segments identified can deepen relationships and optimize marketing efforts, reshaping the institution's approach to customer engagement and contributing to long-term success.

TABLE OF CONTENTS

INTRODUCTION

In today's data-driven world, organizations across various industries are increasingly recognizing the significance of customer segmentation in understanding their clientele and tailoring services and marketing campaigns to meet their specific needs. As a data enthusiast with a passion for unlocking insights from complex datasets, I embarked on a personal research project aimed at leveraging the power of data analytics to conduct customer segmentation within the banking industry.

The reason I chose the banking industry is related to my previous professional experience, where I completed an internship at a bank in China. During my time there, I conducted research on a particular client group within the retail banking clients, those who are 18 to 35 years old. This client group was selected because their level of engagement with the bank was significantly lower compared to clients from other age groups. One of the most important insights I generated from my research was that these younger generation clients have very different preferences with regard to types of services and channels of engagement. However, the bank had very few targeted marketing campaigns. This resulted in a level of engagement that is below the industry average. These findings led me to understand the importance of customer segmentation and how it allows businesses to cater to the different preferences of different customers.

The main purpose of this project is to segment the customers of this banking institution and gain a comprehensive understanding of these customer groups. I seek to reveal the distinct characteristics of each customer group to facilitate more customized marketing. Ultimately, I aim to improve the overall effectiveness of the marketing campaign of this bank.

DESCRIPTION OF DATA

The data set used is the Bank Marketing data from the UCI Machine Learning Repository. This data set is "related with direct marketing campaigns (phone calls) of a Portuguese banking institution", and the original goal of the data is for a classification algorithm to "predict if the client will subscribe a term deposit" (Moro et al., 2014). Since my goal is to segment the clients instead of predicting the outcome of the marketing campaign, I made some necessary changes to this data set.

A data dictionary with a full list of all the variables is available in Appendix A. There are in total 20 variables related to each customer as well as one target variable, whether the client subscribed to a term deposit. This data set also includes a few socio-economic variables such as consumer price index and Euribor rate (variables 16-20), these variables were not used because they are not directly associated with each customer and therefore were not used.

The database contains four data sets, some with smaller sizes or fewer variables. The one used in this case is the full data set with 41188 rows and 21 columns. While there are no missing values in the data, there are certain variables that contain the value "unknown" in some cases. Since the data set is big enough, the rows containing "unknown" values were removed. After removing these rows, there are 30488 rows left.

EXPLORATORY DATA ANALYSIS

To start the exploratory data analysis, I plotted the distribution for all the variables in the dataset as well as their relationship with each other, including the ones that will not be included in the model, to gain a high-level understanding of all these variables and their roles in the original study. This is an essential step in any data analytics project since it serves as a foundational bedrock where the entire analytical process will be built.

1. Age: The distribution of the age variable is right-skewed, with a peak at the age of early 30s. Most of the clients are between the age of 25 and 60, as the count of clients outside of this range dwindles quickly. As one of the few numerical variables in the data set, this will be an important variable to include in the model.
2. Job: There are 11 categories of jobs in total, including the category “unemployed”. The top 3 common ones are administration, blue-collar, and technician. Together they account for almost 20000 people, or 65% of all the clients.
3. Marital: There are only 3 categories for the marital status variable, “married”, “single”, and “divorced”. Over half of the clients are married.
4. Education: There are 7 categories in total for education. The most common ones are “university degree”, “high school”, and “professional course”. Around 30% of all clients have a university degree, around 25% graduated high school, whereas only 11 people are illiterate.
5. Default: This is a binary variable with only “yes” or “no” values, and only 3 out of the 30488 clients have defaulted. Since over 99% of all clients have the same value for this variable, it will not be included in the model.
6. Housing: This variable is also binary, signifying whether the client has housing loan with the bank. It has close to equal distribution between yes and no, with slightly more clients who have housing loan compared to those who don’t.
7. Loan: Also a binary variable, showing whether a client has a personal loan with the bank. Over 80% of clients don’t have personal loans, and the rest do.

The other variables are either related to the previous marketing campaign or not specific to each customer level, and thus won’t be included in our customer segmentation model. Finally, I decided to use the variables “age”, “job”, “marital”, “education”, “housing”, and “loan” to construct the model.

APPROACH & METHODS

In selecting the appropriate method for customer segmentation in this project, I took several considerations into account to ensure the best-fit approach for the nature of this dataset. The dataset predominantly consists of categorical variables about customer characteristics, with a single numeric variable, age, also included. This unique data composition presented an opportunity to explore clustering methods beyond the well-known k-means and k-nearest neighbors (KNN) algorithms.

While k-means and KNN are popular and widely-used clustering techniques, they are predominantly designed for numeric data, which may not work well when handling datasets with a mix of categorical and numeric variables. To address this, the k-prototypes algorithm emerged as a suitable choice. K-prototypes is an extension of k-means that accommodates mixed data types, effectively combining the strengths of k-means and k-modes for categorical variables. This aligns well with our dataset's composition and allows for a more accurate division of customer segments.

The selection of an appropriate gamma value in the k-prototypes algorithm played a crucial role in determining the weight assigned to the numeric variable. A γ=0 means that the model will perform clustering solely on the numeric variables, and the weight of the numeric variables decreases as the value of γ goes up. Initial experimentation with different gamma values revealed that a gamma value of 5 disproportionately emphasized the numeric variable (age) in the clustering process, as shown in the resulting clusters that almost every cluster has the same values for the categorical variables and only different ages. While age is certainly an important factor in the model, it shouldn't be the only deciding factor. A gamma value of 10 was chosen as it balanced the influence of categorical and numeric variables more effectively, contributing to a more balanced and meaningful customer segmentation outcome.

To determine the optimal number of clusters, an elbow curve was generated by plotting the within-cluster sum of squares against different k values (figure 10). Based on the elbow curve, it seems like a k value of 3 or 4 should be chosen. However, given the context of customer segmentation, relying solely on the elbow curve to decide the number of clusters might oversimplify the complexity of customer behavior. Considering the size of the dataset (with over 30,000 clients), we might need a larger number of clusters than the optimal value for the

distance/cost tradeoff. Consequently, a more nuanced approach was employed. A Python for loop was designed to create multiple clustering algorithms, with k values ranging from 3 to 8. A subsequent for-loop was created to visualize the distribution across clusters as well as print the center of each cluster in each case. This approach enabled a comprehensive evaluation of different cluster configurations, considering both the data-driven results and the business context of customer segments. A final decision on the number of clusters would then be made after consideration of the results.

RESULTS & DISCUSSION

It's a quite challenging task to examine the results from a k-prototypes clustering. After some research, I decided to first plot a bar chart to visualize the distribution of the clients in each cluster. On top of that, for each cluster in every scenario of cluster number k, I created a table that prints the mean and median of the numeric variables, and the mode as well as the percentage of instances that match the mode for all categorical variables. The results are in Figures 12-17.

- In the context of the 3-cluster scenario, it becomes evident that the clusters are predominantly formed by the age variable. This phenomenon is illustrated in Figure 12. The cluster centers distinctly emerge around clients in their 30s, 40s, and 50s, respectively. What's worth noticing is that these customer segments had negligible discrepancies with regard to categorical variables such as housing loans and education levels. Intriguingly, approximately half of all clients fall into a single cluster, presenting an undesirable outcome for precise segmentation. To address this limitation, it requires additional clusters to be formed to enable the formulation of more refined and targeted marketing campaigns.
- Similarly, in the 4-cluster scenario, the discernible distinction among the groups remains primarily the age variable, recreating the patterns seen in the 3-cluster configuration. Figure 13 visualizes this setup, where the key divergence stems from the segmentation of clients into four age groups. Across these clusters, the categorical variables—such as education, housing, and loan attributes—persistently exhibit consistent values at the cluster centers. For instance, all cluster centers were comprised of administrative roles for the job variable, except for the group with an average age of 59, as this specific group predominantly consists of retired clients.
- The results from the 5 and 6-cluster scenarios show significant resemblance. A very distinct group of 70+ years old customers has started to emerge. Despite being a relatively small cluster, comprising only a few hundred individuals, its separation from the rest of the clients is clear-cut. A typical client in this group is over 70 years old, married and retired, and possess basic 4-year education. Another distinctive characteristic of this group is that they have a remarkably high response rate to the marketing campaign, which was conducted by phone. Around 46% of these clients ended up subscribing to a term deposit, while the average for all the other groups is around 10%. However, other than that, the rest of the groups are still homogeneous to some degree. While subtle variations in attributes like housing loans, education, and

employment begin to emerge across the groups, the percentage of instances that match the mode for most of the categorical variables is still not high enough, showing the overall diversity in these variables remains limited.

- The 7-cluster configuration seems more promising compared to the previous scenarios. As can be seen in Figure 16, the distribution of clients across clusters is reasonably equal, except for the very distinct group of older clients. All of these clusters have their distinctive characteristics, and I will discuss each of them separately.
    - Cluster 0: Notably, this cluster exhibits an elevated average age of approximately 47.10 years, signifying an older demographic compared to the general clientele, which has an average age of 39 years. Despite this higher age, the average outcome, representing the subscription percentage for term deposits, stands at around 9.74%, which is lower than the overall mean of 12.66%. This implies that customers within this cluster are relatively less responsive to term deposit campaigns. An interesting observation pertains to the prevalent attributes associated with this cluster. A typical customer within this segment is more likely to be married (69.50%), have an administrative job (31.07%), have a level of education that is comparable to the overall average, and refrain from holding housing (54.65%), and personal loans (84.91%). While mirroring broader trends, this cluster represents a distinct subset of customers characterized by their advanced age, marital status, and conservative financial behaviors that contribute to the lower response rate to the term deposit campaign.
    - Cluster 1: The average age of approximately 55.43 years surpasses the general mean of 39 years, marking this cluster as older in age. This cluster displays an average outcome, reflecting a term deposit subscription rate of around 14.98%, which is slightly higher than the overall mean of 12.66%. This suggests a stronger receptiveness to term deposit campaigns within this segment. The distribution of job categories in this cluster is rather even, with the most common one making up only 21%. Additionally, marital status remains predominantly "married" with a mode percentage of approximately 75.02%. Similarly, an average percentage of clients (32.98%) hold a university degree. Moreover, more people have housing loans than not (58.22%), while the personal loan variable aligns with the overall trend, primarily indicating "no" responses (84.81%). Overall, this cluster embodies an older demographic with a heightened interest in term deposits, predominantly married, and have a housing loan.
    - Cluster 2: This customer cluster exhibits distinct attributes that define a relatively youthful segment within the larger clientele. With an average age of approximately 36.25 years, the cluster is younger than the overall mean of 39

years. Although term deposit engagement slightly trails the average with an average subscription rate of 10.37% (below the 12.66% mean), the dominant "technician" job category (35.71%) and prevalence of university degrees (38.71%) characterize this group. Marital status leans towards "married" (61.29%), and a vast majority do not hold housing loans (73.82%) or personal loans (85.49%). Overall, this cluster's distinct attributes, including technical professions and higher education, offer valuable insights for targeted marketing strategies.

- Cluster 3: This particular customer cluster represents the youngest group of all, with an average age of approximately 27.56 years and a median age of 28.0 years. In terms of term deposit subscription, this cluster is just above average, with an average subscription rate of 15.53%, exceeding the overall mean of 12.66%. The distribution of job categories is still quite even, with administration still being the most prevalent. Additionally, a significant proportion of customers within this cluster are single (74.41%) and hold university degrees (35.18%). Housing and loan attributes align with the broader trends, with a preference for "no" responses in both categories (housing: 55.89%, loan: 84.47%). In summary, this cluster represents a youthful segment with an inclination towards term deposit subscriptions, probably driven by the prevalence of higher education and single marital status. These attributes offer valuable insights for targeted marketing strategies catering to this specific demographic.
- Cluster 4: This customer cluster presents a middle-aged demographic, with an average age of approximately 40.83 years. This figure almost coincides with the overall average of 39 years. However, the cluster demonstrates the lowest level of engagement with term deposits, with an average subscription rate of 9.27% compared to the overall mean of 12.66%. The most prevalent job category within this segment is "blue-collar" (29.35%), indicating a substantial representation of manual laborers. Marital status predominantly leans towards married as well (67.39%), while education levels are most commonly "high school" (34.10%). This group also has a higher-than-average proportion of clients that have housing loans (69.42%). In summary, this cluster represents a middle-aged group with characteristics typical of blue-collar workers, who graduated high school, and with an inclination towards homeownership while avoiding personal loans. Their lower-than-average term deposit engagement suggests potential opportunities for tailored marketing strategies to increase subscription rates within this segment.

- Cluster 5: This particular customer cluster is another relatively young demographic, with an average age of approximately 32.63. In terms of term deposit subscription, this cluster displays a slightly lower-than-average subscription rate of 11.71%. The dominant job category is still administrative (40.65%). Marital status leans towards "married" (61.65%), and a significant proportion holds university degrees (42.91%), reflecting higher educational attainment within this segment. Additionally, housing loan preferences trend heavily towards "yes" responses (77.22%), and loan attributes again align with the broader trend of "no" responses (82.77%). In summary, this cluster represents a relatively young demographic with a somewhat low engagement with term deposits, combined with a prevalence of administrative jobs, higher education, and homeownership. These attributes present opportunities for targeted marketing strategies tailored to this specific segment.
- Cluster 6: This customer cluster is distinctive due to its advanced age, with an average age of approximately 73.66 years, almost doubling the overall mean age of 39.03 years. Term deposit engagement within this cluster is extremely high, with an average subscription rate of 46.54%, well above the overall mean of 12.66%. The overwhelming majority of individuals in this cluster are "retired" (88.22%), aligning with the higher age demographic. Marital status leans towards "married" (69.91%), reflecting a relatively stable relationship status among this group. Education levels are predominantly "basic.4y" (51.03%), indicating a lower level of formal education. Housing preferences trend towards "yes" responses (55.51%), while loan attributes conform to the broader trend of "no" responses (84.86%). In summary, this cluster represents an elderly demographic characterized by retirement, a propensity for term deposits, and a preference for homeownership, despite a lower level of formal education. These attributes offer valuable insights for tailored marketing strategies aimed at engaging this segment.

- In summary, the 7-cluster configuration quite reasonably separates the bank’s clientele. Most of the clusters have distinctive characteristics, instead of being divided by age alone. This gives us enough insights to curate targeted marketing strategies towards these customer segments and ideally, be able to improve the success rate.
- Compared to the 7-cluster configuration, the 8-cluster configuration seems less desirable. While there’s one additional cluster, the distance between the clusters didn’t improve significantly. For many of the clusters, the distribution of important variables

like job category and education is rather sparse. Overall, the clusters did not exhibit clear-cut distinctions.

CONCLUSION

In conclusion, this project has provided valuable insights into customer segmentation within the banking industry, underscoring the importance of data analytics in tailoring marketing campaigns and services to meet the specific needs of diverse customer groups. By leveraging machine learning algorithms, we aimed to enhance the effectiveness of marketing strategies for a banking institution.

Our analysis of the Bank Marketing dataset revealed nuanced customer segments, each characterized by distinct attributes and behaviors. Through extensive exploratory data analysis and the application of the k-prototypes clustering algorithm, we identified seven distinct customer clusters, each offering unique opportunities for targeted marketing initiatives.

These clusters ranged from older, retired clients with a high propensity for term deposits to younger, well-educated professionals with varied preferences. Notably, our approach went beyond age-based segmentation, taking into account job categories, marital status, education levels, and housing and personal loan status to create more nuanced customer profiles.

The results of this project demonstrate the potential for data-driven customer segmentation to enhance marketing strategies within the banking industry. By tailoring campaigns to the specific needs and preferences of each cluster, the banking institution can expect to improve the success rate of its marketing initiatives and better serve its diverse clientele.

However, it's essential to emphasize that customer segmentation is an ongoing process that should evolve with changing customer behaviors and market dynamics. Regularly revisiting and refining these segments will ensure that marketing efforts remain effective and responsive to customer needs.

In summary, this project represents a significant step towards optimizing marketing strategies in the banking industry through data-driven customer segmentation. The insights gained from this analysis provide a solid foundation for future marketing campaigns, highlighting the potential for improved customer engagement and satisfaction.

REFERENCES

1. S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, In press, http://dx.doi.org/10.1016/j.dss.2014.03.001

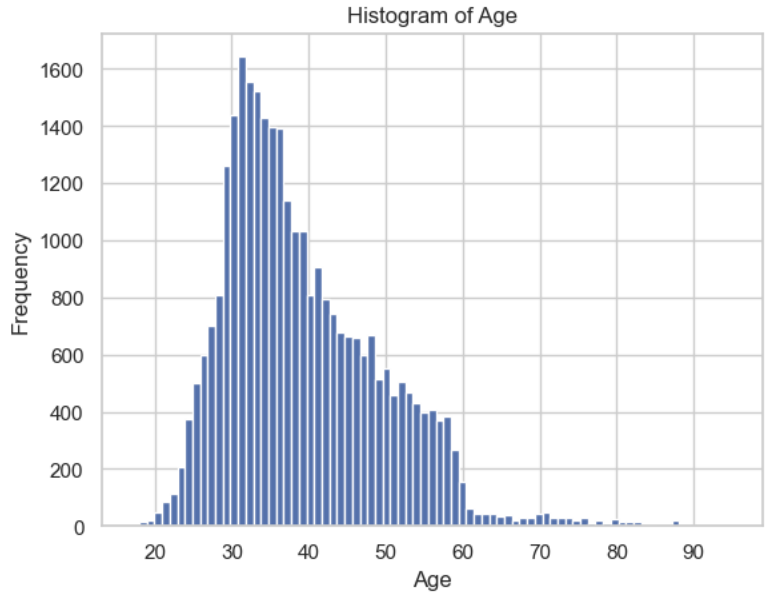
APPENDIX A. [DATA DICTIONARY]

1 - age (numeric)
2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
5 - default: has credit in default? (categorical: 'no','yes','unknown')
6 - housing: has housing loan? (categorical: 'no','yes','unknown')
7 - loan: has personal loan? (categorical: 'no','yes','unknown')
related with the last contact of the current campaign:
8 - contact: contact communication type (categorical: 'cellular','telephone')
9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
10 - day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.
other attributes:
12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
14 - previous: number of contacts performed before this campaign and for this client (numeric)
15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')
social and economic context attributes
16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)
17 - cons.price.idx: consumer price index - monthly indicator (numeric)
18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)
19 - euribor3m: euribor 3 month rate - daily indicator (numeric)
20 - nr.employed: number of employees - quarterly indicator (numeric)

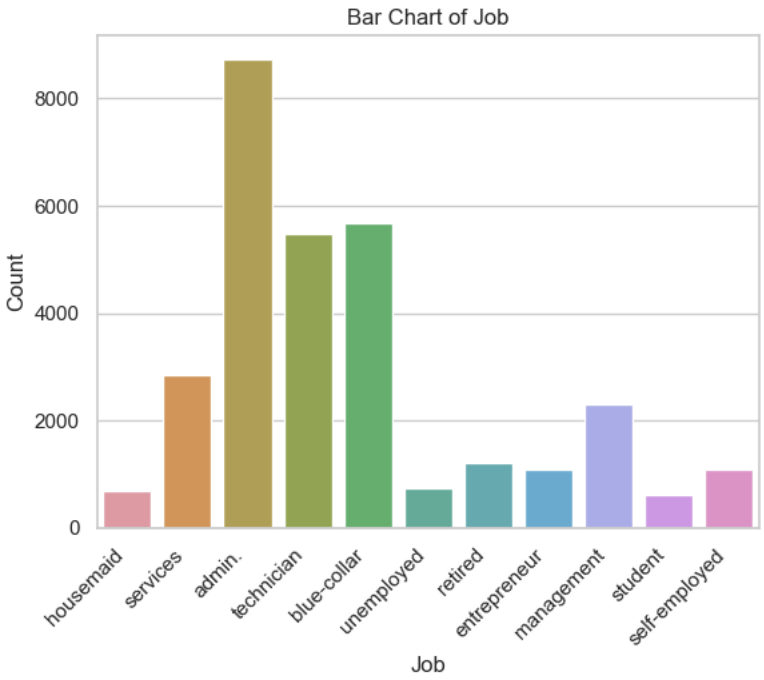
Output variable (desired target):
21 - y - has the client subscribed a term deposit? (binary: 'yes','no')

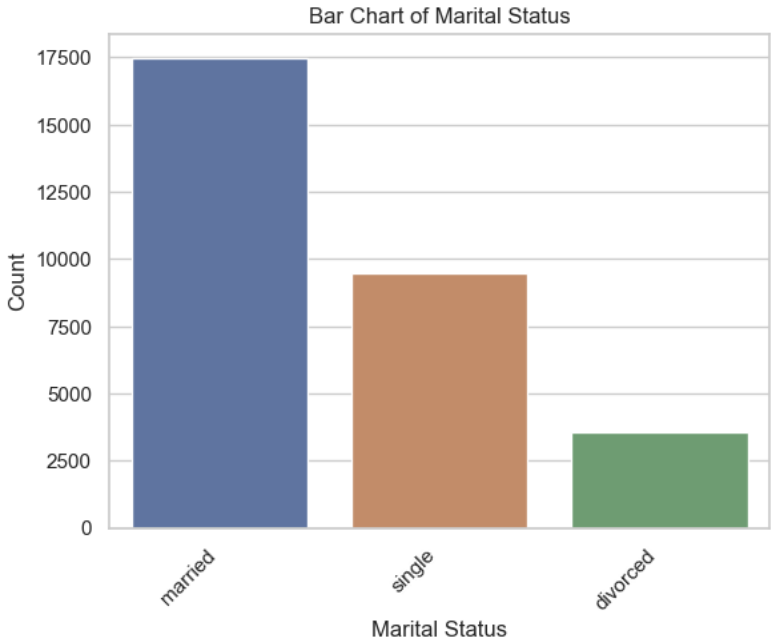
APPENDIX B. [VISUALIZATIONS]

1.

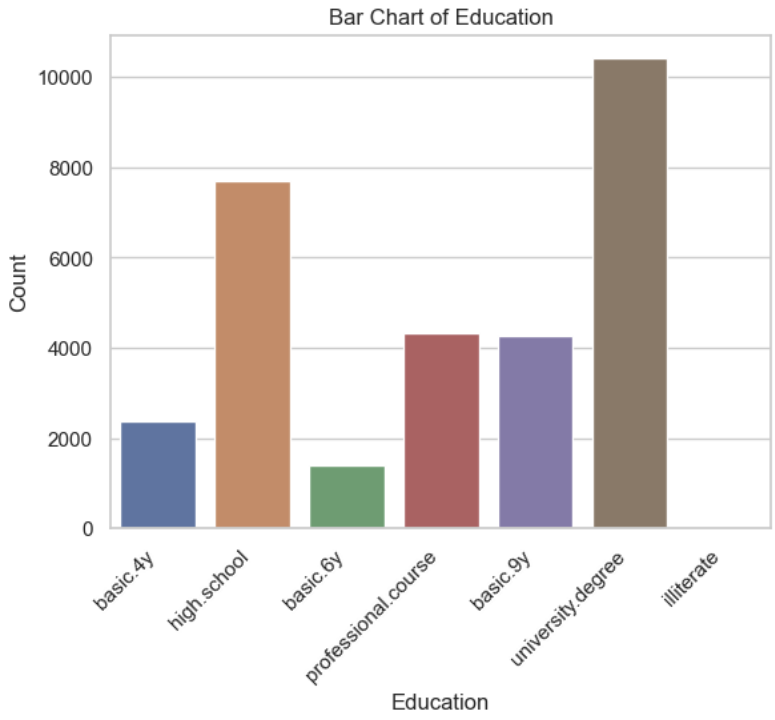


2.



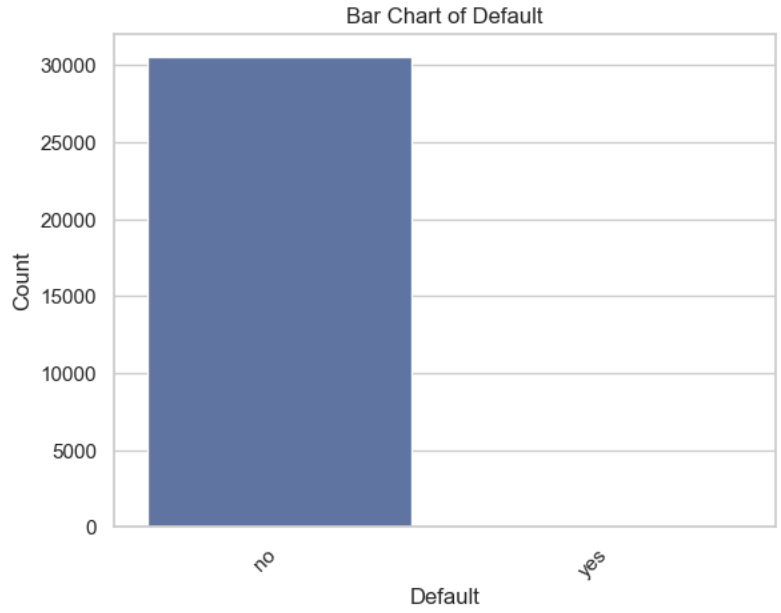

Bar Chart of Marital Status
17500
15000
12500
10000
7500
5000
2500
0
Count
married
single
divorced
Marital Status

3.

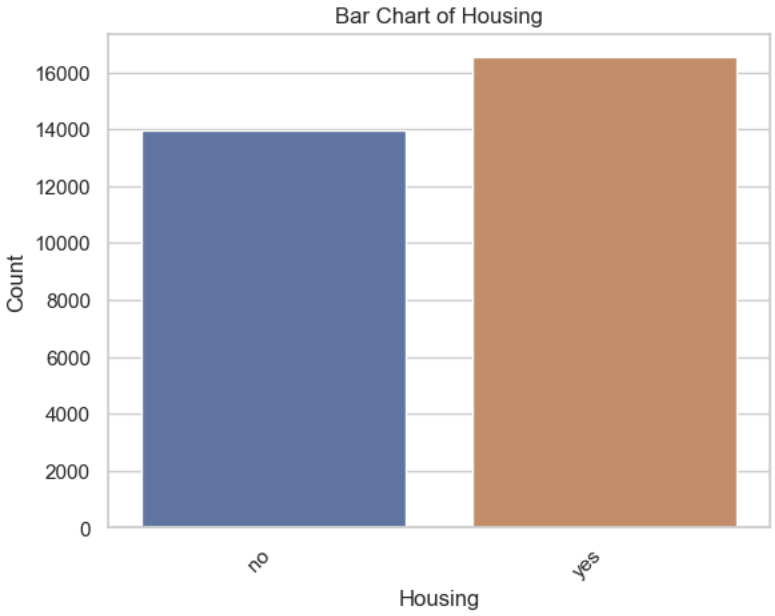

Bar Chart of Education
10000
8000
6000
4000
2000
0
Count
basic.4y
high.school
basic.6y
professional.course
basic.9y
university.degree
illiterate
Education

4.

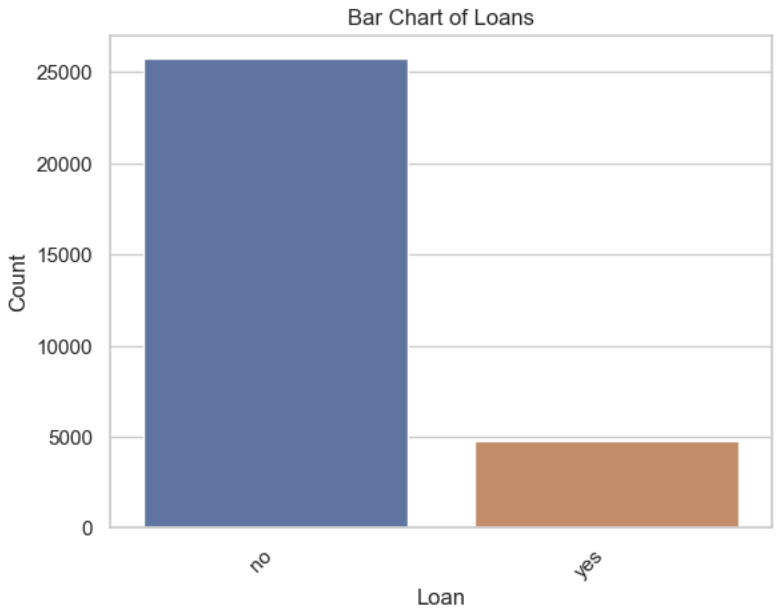
5.


Bar Chart of Default
30000
25000
20000
15000
10000
5000
0
Count
no
yes
Default

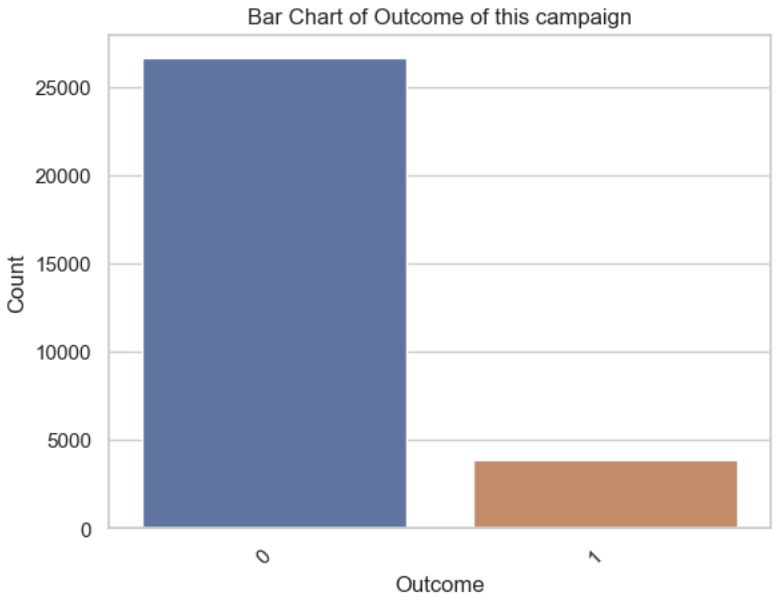
6.

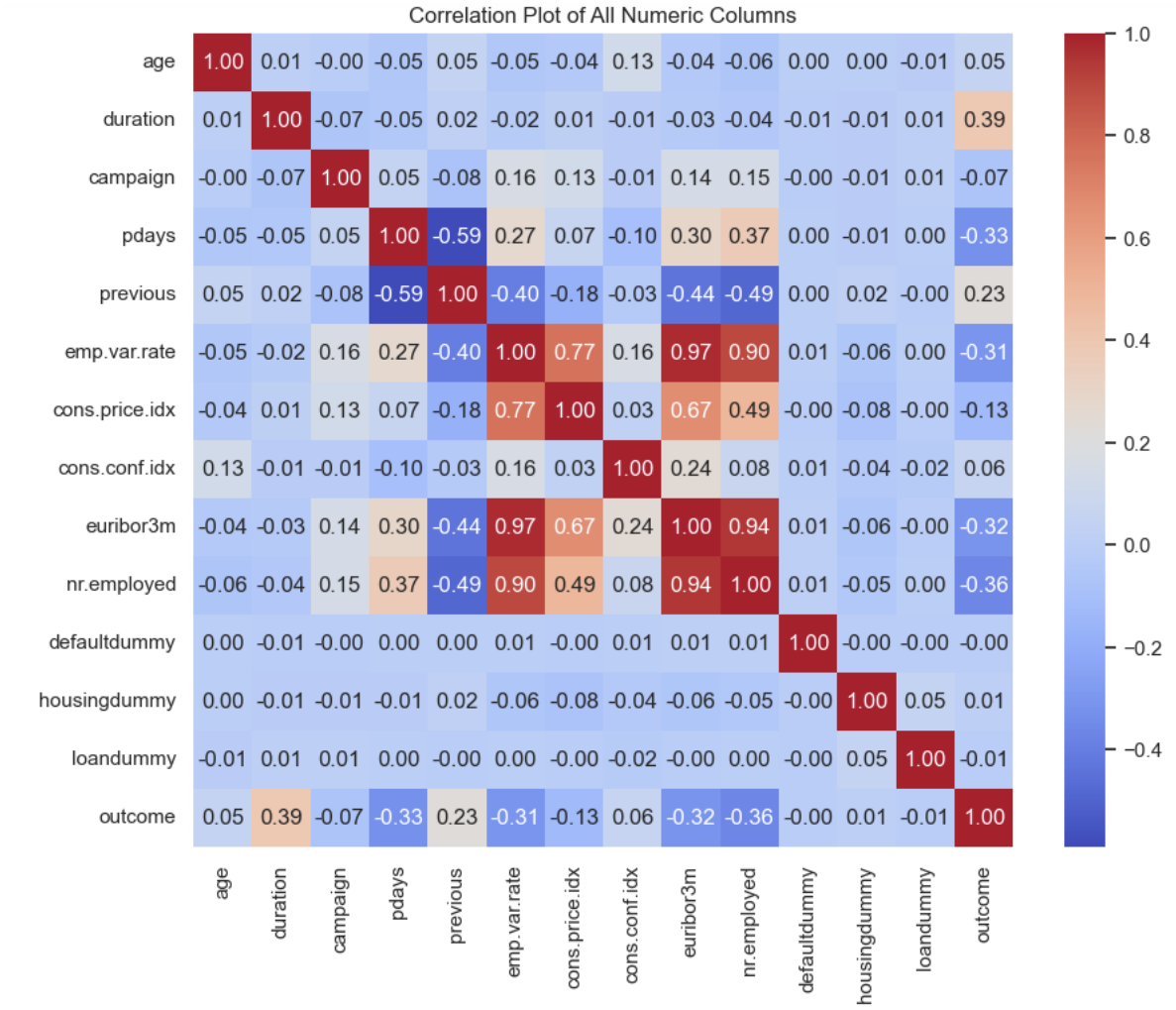

Bar Chart of Housing
16000
14000
12000
10000
8000
6000
4000
2000
0
Count
no
yes
Housing

7.


Bar Chart of Loans
25000
20000
15000
10000
5000
0
Count
no
yes
Loan

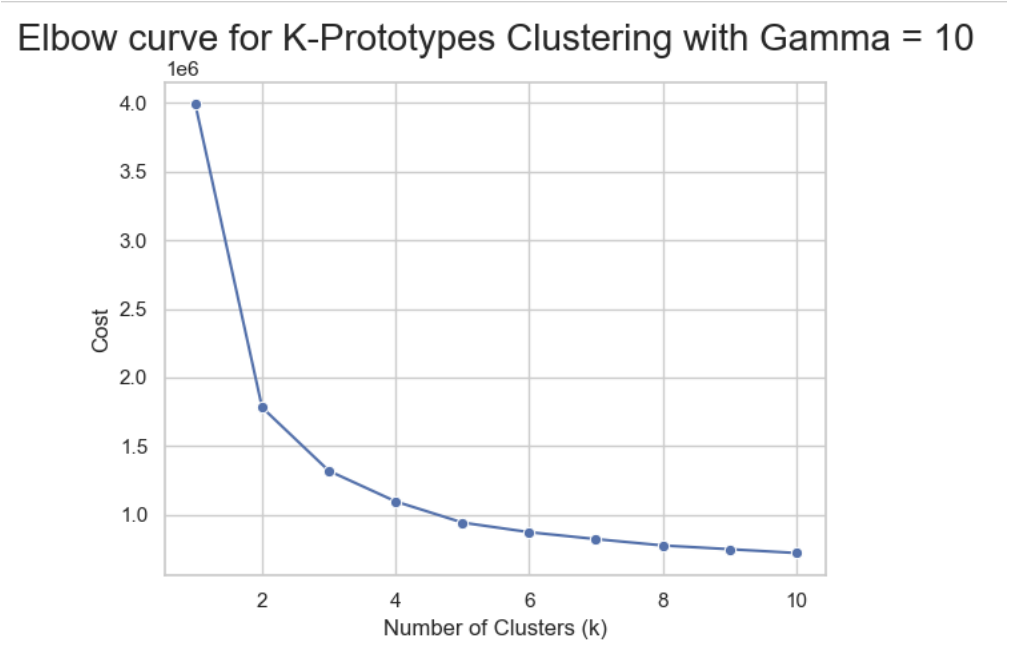
8.


Bar Chart of Outcome of this campaign
25000
20000
15000
10000
5000
0
Count
0
1
Outcome

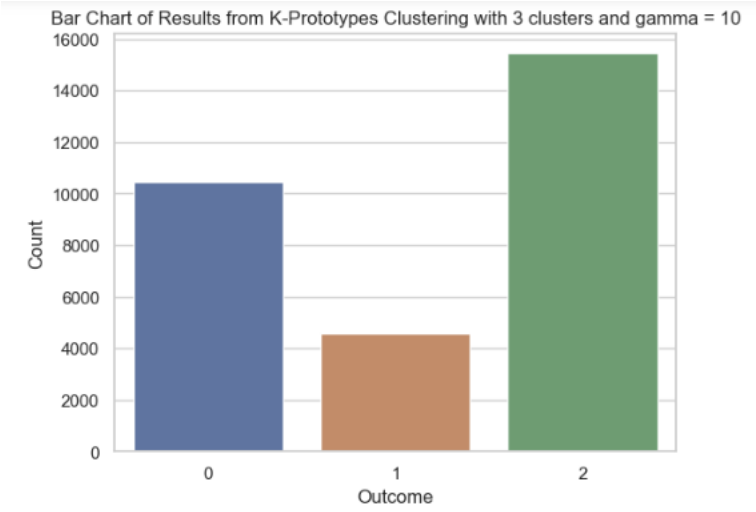

Correlation Plot of All Numeric Columns
age 1.00 0.01 -0.00 -0.05 0.05 -0.05 -0.04 0.13 -0.04 -0.06 0.00 0.00 -0.01 0.05
duration 0.01 1.00 -0.07 -0.05 0.02 -0.02 0.01 -0.01 -0.03 -0.04 -0.01 -0.01 0.01 0.39
campaign -0.00 -0.07 1.00 0.05 -0.08 0.16 0.13 -0.01 0.14 0.15 -0.00 -0.01 0.01 -0.07
pdays -0.05 -0.05 0.05 1.00 -0.59 0.27 0.07 -0.10 0.30 0.37 0.00 -0.01 0.00 -0.33
previous 0.05 0.02 -0.08 -0.59 1.00 -0.40 -0.18 -0.03 -0.44 -0.49 0.00 0.02 -0.00 0.23
emp.var.rate -0.05 -0.02 0.16 0.27 -0.40 1.00 0.77 0.16 0.97 0.90 0.01 -0.06 0.00 -0.31
cons.price.idx -0.04 0.01 0.13 0.07 -0.18 0.77 1.00 0.03 0.67 0.49 -0.00 -0.08 -0.00 -0.13
cons.conf.idx 0.13 -0.01 -0.01 -0.10 -0.03 0.16 0.03 1.00 0.24 0.08 0.01 -0.04 -0.02 0.06
euribor3m -0.04 -0.03 0.14 0.30 -0.44 0.97 0.67 0.24 1.00 0.94 0.01 -0.06 -0.00 -0.32
nr.employed -0.06 -0.04 0.15 0.37 -0.49 0.90 0.49 0.08 0.94 1.00 0.01 -0.05 0.00 -0.36
defaultdummy 0.00 -0.01 -0.00 0.00 0.00 0.01 -0.00 0.01 0.01 0.01 1.00 -0.00 -0.00 -0.00
housingdummy 0.00 -0.01 -0.01 -0.01 0.02 -0.06 -0.08 -0.04 -0.06 -0.05 -0.00 1.00 0.05 0.01
loandummy -0.01 0.01 0.01 0.00 -0.00 0.00 -0.00 -0.02 -0.00 0.00 -0.00 0.05 1.00 -0.01
outcome 0.05 0.39 -0.07 -0.33 0.23 -0.31 -0.13 0.06 -0.32 -0.36 -0.00 0.01 -0.01 1.00
age duration campaign pdays previous emp.var.rate cons.price.idx cons.conf.idx euribor3m nr.employed defaultdummy housingdummy loandummy outcome
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4

9.
10.

11.


Elbow curve for K-Prototypes Clustering with Gamma = 10
1e6
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Cost
2
4
6
8
10
Number of Clusters (k)

12.

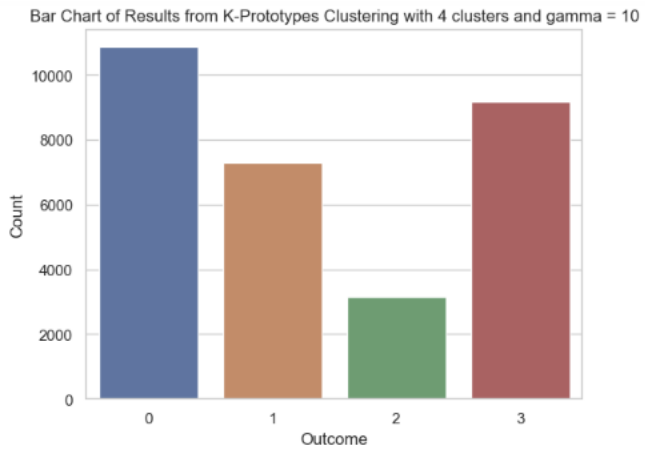


cluster3		0	\
age	mean	42.753275	
	median	42.0	
outcome	mean	0.097829	
	median	0.0	
job	<lambda>	(admin., 27.54135985464282)	
marital	<lambda>	(married, 69.14985177393133)	
education	<lambda>	(university.degree, 30.46762933919862)	
housing	<lambda>	(yes, 53.54308118963373)	
loan	<lambda>	(no, 85.07220043989672)	

cluster3		1	\
age	mean	57.346053	
	median	56.0	
outcome	mean	0.184693	
	median	0.0	
job	<lambda>	(retired, 25.2725686873092)	
marital	<lambda>	(married, 74.13868294810293)	
education	<lambda>	(university.degree, 30.87658089838639)	
housing	<lambda>	(yes, 55.342346271260354)	
loan	<lambda>	(no, 84.53990405582206)	

cluster3		2
age	mean	31.070702
	median	31.0
outcome	mean	0.12878
	median	0.0
job	<lambda>	(admin., 32.09452897377792)
marital	<lambda>	(single, 49.51116866299773)
education	<lambda>	(university.degree, 37.61735189381677)
housing	<lambda>	(yes, 54.28293946260926)
loan	<lambda>	(no, 83.8264810618323)

13.



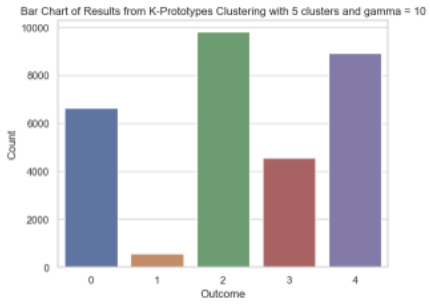

```
cluster4                                                    0  \
age       mean                                      36.354462
          median                                         36.0
outcome   mean                                       0.102484
          median                                          0.0
job       <lambda>            (admin., 29.420423183072675)
marital   <lambda>            (married, 67.42410303587857)
education <lambda>  (university.degree, 34.590616375344986)
housing   <lambda>                  (yes, 53.9834406623735)
loan      <lambda>                   (no, 83.95584176632934)

cluster4                                                    1  \
age       mean                                      46.927121
          median                                         47.0
outcome   mean                                        0.10129
          median                                          0.0
job       <lambda>            (admin., 27.642053252813614)
marital   <lambda>            (married, 70.38155366456216)
education <lambda>  (university.degree, 30.963491627779305)
housing   <lambda>                (yes, 54.172385396651116)
loan      <lambda>                  (no, 84.69667856162503)

cluster4                                                    2  \
age       mean                                      59.76067
          median                                         57.0
outcome   mean                                       0.213405
          median                                          0.0
job       <lambda>          (retired, 35.03003477711034)
marital   <lambda>          (married, 74.73917167246286)
education <lambda>  (university.degree, 29.84508378122036)
housing   <lambda>                (yes, 54.8846032247866)
loan      <lambda>                  (no, 85.1406892190958)

cluster4                                                    3
age       mean                                      28.775221
          median                                         29.0
outcome   mean                                       0.145272
          median                                          0.0
job       <lambda>            (admin., 32.56625586214418)
marital   <lambda>            (single, 66.07045479332534)
education <lambda>  (university.degree, 37.64859853855383)
housing   <lambda>                (yes, 54.20438433853201)
loan      <lambda>                 (no, 84.30581306576508)
```

14.



cluster5		0
age	mean	44.35584
	median	44.0
outcome	mean	0.093896
	median	0.0
job	<lambda>	(admin., 27.65636774679729)
marital	<lambda>	(married, 68.45516201959306)
education	<lambda>	(university.degree, 29.751318764129614)
housing	<lambda>	(yes, 56.63903541823663)
loan	<lambda>	(no, 84.77769404672193)

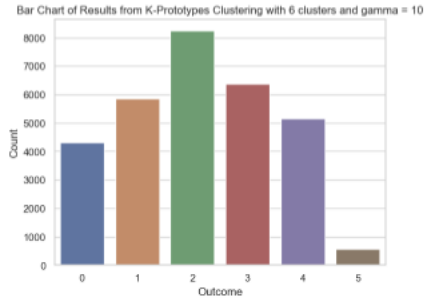
cluster5		1
age	mean	73.056042
	median	72.0
outcome	mean	0.460595
	median	0.0
job	<lambda>	(retired, 86.5148861646234 8)
marital	<lambda>	(married, 71.2784588441331)
education	<lambda>	(basic.4y, 49.036777583187394)
housing	<lambda>	(yes, 55.51663747810858)
loan	<lambda>	(no, 85.11383537653239)

cluster5		2
age	mean	35.652218
	median	36.0
outcome	mean	0.103621
	median	0.0
job	<lambda>	(admin., 30.015298317185106)
marital	<lambda>	(married, 65.46659867414584)
education	<lambda>	(university.degree, 35.604283528811834)
housing	<lambda>	(no, 53.69709331973483)
loan	<lambda>	(no, 84.33452320244773)

cluster5		3
age	mean	54.502852
	median	54.0
outcome	mean	0.143045
	median	0.0
job	<lambda>	(admin., 22.8389644580956 56)
marital	<lambda>	(married, 74.44054409828873)
education	<lambda>	(university.degree, 32.77753400614304)
housing	<lambda>	(yes, 55.33128565160158)
loan	<lambda>	(no, 84.73014480035103)

cluster5		4
age	mean	28.695706
	median	29.0
outcome	mean	0.146317
	median	0.0
job	<lambda>	(admin., 32.57091602197556)
marital	<lambda>	(single, 65.04092387038905)
education	<lambda>	(university.degree, 37.5266285458011)
housing	<lambda>	(yes, 60.36551182868034)
loan	<lambda>	(no, 83.84348021078596)

15.



cluster6		0
age	mean	54.758637
	median	54.0
outcome	mean	0.144911
	median	0.0
job	<lambda>	(admin., 22.37421748203107)
marital	<lambda>	(married, 74.5420820774403)
education	<lambda>	(university.degree, 32.80779040111291)
housing	<lambda>	(yes, 58.47437978205426)
loan	<lambda>	(no, 84.41919777417111)

cluster6		1
age	mean	38.632803
	median	39.0
outcome	mean	0.103814
	median	0.0
job	<lambda>	(admin., 34.70155635368565)
marital	<lambda>	(married, 63.057978450487425)
education	<lambda>	(university.degree, 42.22678296562339)
housing	<lambda>	(yes, 67.6928339319309)
loan	<lambda>	(no, 84.52197708226441)

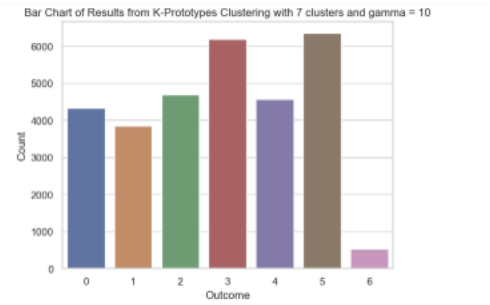
cluster6		2
age	mean	28.570267
	median	29.0
outcome	mean	0.153155
	median	0.0
job	<lambda>	(admin., 37.81553398058252)
marital	<lambda>	(single, 68.81067961165049)
education	<lambda>	(university.degree, 45.0)
housing	<lambda>	(yes, 62.026699029126206)
loan	<lambda>	(no, 83.84708737864077)

cluster6		3
age	mean	33.793407
	median	34.0
outcome	mean	0.096703
	median	0.0
job	<lambda>	(blue-collar, 30.894819466248038)
marital	<lambda>	(married, 68.50863422291994)
education	<lambda>	(high.school, 33.6734693877551)
housing	<lambda>	(no, 65.66718995290424)
loan	<lambda>	(no, 84.03453689167975)

cluster6		4
age	mean	45.752671
	median	46.0
outcome	mean	0.094424
	median	0.0
job	<lambda>	(admin., 28.288323295123373)
marital	<lambda>	(married, 69.72993977074023)
education	<lambda>	(university.degree, 29.764911598989702)
housing	<lambda>	(no, 52.865747037109)
loan	<lambda>	(no, 85.27297454828056)

cluster6		5
age	mean	73.056042
	median	72.0
outcome	mean	0.460595
	median	0.0
job	<lambda>	(retired, 86.51488616462348)
marital	<lambda>	(married, 71.2784588441331)
education	<lambda>	(basic.4y, 49.036777583187394)
housing	<lambda>	(yes, 55.51663747810858)
loan	<lambda>	(no, 85.11383537653239)

16.



cluster7		0
age	mean	47.101342
	median	47.0
outcome	mean	0.097409
	median	0.0
job	<lambda>	(admin., 31.07357704766312)
marital	<lambda>	(married, 69.50485886163813)
education	<lambda>	(university.degree, 35.23831559463211)
housing	<lambda>	(no, 54.65062471078205)
loan	<lambda>	(no, 84.91439148542341)

cluster7		1
age	mean	55.425371
	median	55.0
outcome	mean	0.149779
	median	0.0
job	<lambda>	(admin., 21.62021359729096)
marital	<lambda>	(married, 75.01953633758791)
education	<lambda>	(university.degree, 32.97733784839802)
housing	<lambda>	(yes, 58.218286011982286)
loan	<lambda>	(no, 84.81375358166189)

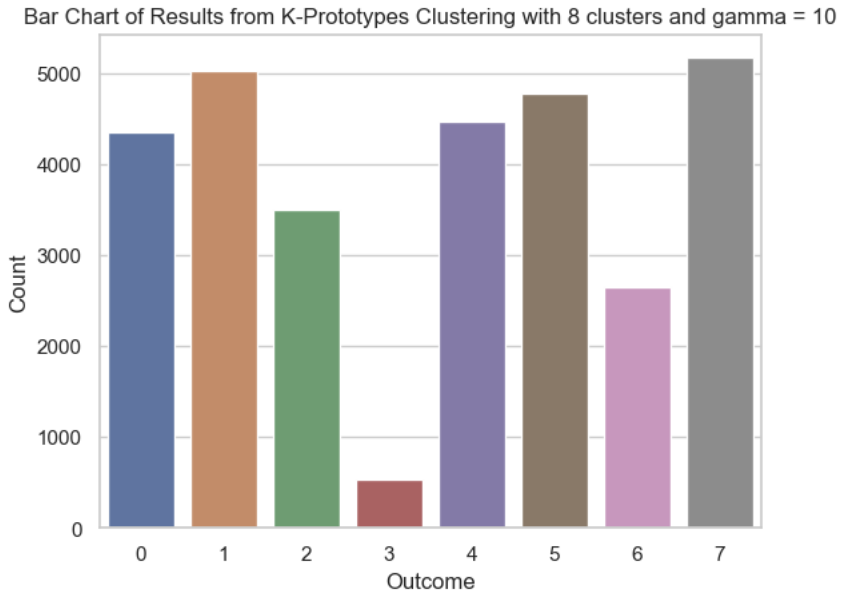
cluster7		2
age	mean	36.247976
	median	36.0
outcome	mean	0.103749
	median	0.0
job	<lambda>	(technician, 35.705155517682144)
marital	<lambda>	(married, 61.29100979974436)
education	<lambda>	(university.degree, 38.70899020025564)
housing	<lambda>	(no, 73.81763953983808)
loan	<lambda>	(no, 85.49211759693225)

cluster7		3
age	mean	27.561307
	median	28.0
outcome	mean	0.15529
	median	0.0
job	<lambda>	(admin., 32.57845357489486)
marital	<lambda>	(single, 74.40957618893562)
education	<lambda>	(university.degree, 35.1827887415076)
housing	<lambda>	(no, 55.88806211582013)
loan	<lambda>	(no, 84.4710449692656)

cluster7		4
age	mean	40.829178
	median	41.0
outcome	mean	0.092738
	median	0.0
job	<lambda>	(blue-collar, 29.352580927384075)
marital	<lambda>	(married, 67.38845144356955)
education	<lambda>	(high.school, 34.09886264216973)
housing	<lambda>	(yes, 69.4225721784777)
loan	<lambda>	(no, 84.2957130358705)

cluster7		5
age	mean	32.627049
	median	33.0
outcome	mean	0.117119
	median	0.0
job	<lambda>	(admin., 40.65258511979823)
marital	<lambda>	(married, 61.64880201765448)
education	<lambda>	(university.degree, 42.90668348045397)
housing	<lambda>	(yes, 77.22257250945775)
loan	<lambda>	(no, 82.77112232030265)

cluster7		6
age	mean	73.659813
	median	72.0
outcome	mean	0.465421
	median	0.0
job	<lambda>	(retired, 88.22429906542057)
marital	<lambda>	(married, 69.90654205607477)
education	<lambda>	(basic.4y, 51.02803738317757)
housing	<lambda>	(yes, 55.51401869158879)
loan	<lambda>	(no, 84.85981308411215)


Bar Chart of Results from K-Prototypes Clustering with 8 clusters and gamma = 10
Count
5000
4000
3000
2000
1000
0
0
1
2
3
4
5
6
7
Outcome

17.

```
cluster8                                                    0  \
age       mean                                      47.64814
          median                                        48.0
outcome   mean                                      0.101056
          median                                         0.0
job       <lambda>       (blue-collar, 25.81534221405604)
marital   <lambda>            (married, 71.08406063389985)
education <lambda>  (university.degree, 31.25861276986679)
housing   <lambda>               (yes, 60.587965089572805)
loan      <lambda>                (no, 84.56591639871382)

cluster8                                                    1  \
age       mean                                     30.112192
          median                                        30.0
outcome   mean                                      0.146545
          median                                         0.0
job       <lambda>           (admin., 41.163621922160445)
marital   <lambda>             (single, 72.85544082605242)
education <lambda>  (university.degree, 49.48371723590151)
housing   <lambda>               (yes, 73.49086576648133)
loan      <lambda>                (no, 83.77680698967434)

cluster8                                                    2  \
age       mean                                     55.855019
          median                                        56.0
outcome   mean                                      0.152988
          median                                         0.0
job       <lambda>            (admin., 21.961681441235346)
marital   <lambda>            (married, 74.69259365170146)
education <lambda>  (university.degree, 34.029167858164136)
housing   <lambda>                (yes, 55.3617386331141)
loan      <lambda>                 (no, 84.67257649413783)

cluster8                                                    3  \
age       mean                                     73.765595
          median                                        72.0
outcome   mean                                      0.459357
          median                                         0.0
job       <lambda>     (retired, 88.468809073724)
marital   <lambda>   (married, 69.75425330812854)
education <lambda>  (basic.4y, 50.85066162570888)
housing   <lambda>       (yes, 55.19848771266541)
loan      <lambda>        (no, 84.87712665406427)

cluster8                                                    4  \
age       mean                                     32.689078
          median                                        33.0
outcome   mean                                      0.109002
          median                                         0.0
job       <lambda>            (admin., 41.61268706723252)
marital   <lambda>           (married, 69.08644181371454)
education <lambda>  (university.degree, 43.82398927853473)
housing   <lambda>                (no, 86.28545901273174)
loan      <lambda>                (no, 84.90060308242127)

cluster8                                                    5  \
age       mean                                    41.152188
          median                                       41.0
outcome   mean                                     0.093992
          median                                        0.0
job       <lambda>       (admin., 34.1427674272556)
marital   <lambda>     (married, 66.35963994138581)
education <lambda>  (high.school, 35.31505128741888)
housing   <lambda>           (no, 62.50785011513502)
loan      <lambda>            (no, 85.6813899937199)

cluster8                                                    6  \
age       mean                                    25.179516
          median                                       25.0
outcome   mean                                     0.167045
          median                                        0.0
job       <lambda>  (blue-collar, 26.303854875283445)
marital   <lambda>       (single, 72.97808012093726)
education <lambda>   (high.school, 46.03174603174603)
housing   <lambda>           (yes, 53.77928949357521)
loan      <lambda>            (no, 83.56009070294785)

cluster8                                                    7
age       mean                                     36.144045
          median                                        36.0
outcome   mean                                     0.101315
          median                                        0.0
job       <lambda>         (blue-collar, 29.25367362722351)
marital   <lambda>              (married, 65.98994586233566)
education <lambda>  (university.degree, 33.507347254447026)
housing   <lambda>               (yes, 79.77571539056459)
loan      <lambda>                (no, 83.21732405259087)
```