A Project of Applying Different Multivariate Methods on Bank Telemarketing Data

Submitted to

Dr. Megan Orr

Submitted By

Sharmin Hossain

Student ID: 1337949

Index

1.Introduction

The data I have worked on came from direct marketing campaigns of a Portuguese banking institution. Their research goal was to predict if clients will subscribe after their telemarketing.

The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

The business purpose of the research was finding out the characteristics that are helping Bank to make customers successfully subscribe for deposits, which helps in increasing campaign efficiently and selecting high value customers.

Data set is taken from UCI Machine Learning repository. There is total 41188 Rows in the dataset.

Attribute Information:

There is total 21 columns in the main dataset.

For my project purpose, I have worked with 10 of them. Here is the brief description of those 10 variables.

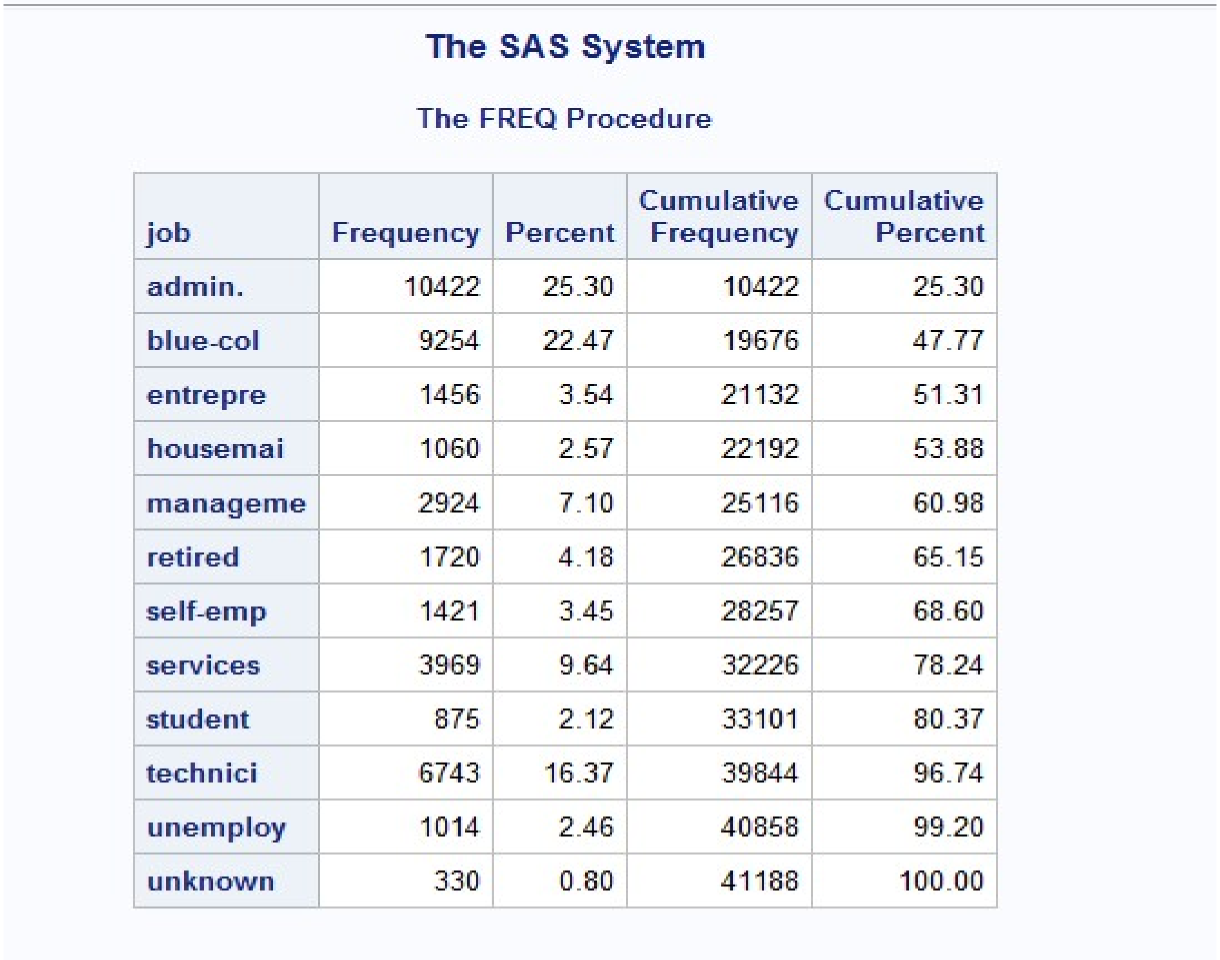
Serial	**Variable Name**	**Variable Type**	**Variable Description**
1	Age	Numeric	
2	Job	Categorical	Type of Job: admin, blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed, unknown
3	Marital	Categorical	divorced, married, single, unknown
4	Education	Categorical	basic.4y, basic.6y, basic.9y, high. School, illiterate, professional. Course, university. Degree, unknown
5	Contact	Categorical	cellular, telephone
6	Duration	Numeric	Last Contact Duration
7	Campaign	Numeric	Number of contacts performed during this campaign and for this client
8	ConsumerPriceIndex	Numeric	Consumer Price Index (Monthly Indicator)
9	ConsumerConfidenceIndex	Numeric	Consumer Confidence Index (Monthly Indicator)
10	Y	Binary	Has the client subscribed a term deposit? (Yes/No)

From this dataset, the answers I am trying to get are:

1. Are all variables equally important for this research? Or can we eliminate some of them using ‘Stepwise’ procedure.
2. What kind of classification can be used in the dataset if we consider “Job” as a grouping variable? Does error rate change if we use both resubstitution and cross validation method?
3. Using different MANOVA methods, can we decide on the equality of population mean vector of numeric variables?
4. How many factors can be used to explain the total variances of those variables? Does rotating can be used to interpret the factors properly?

2.Data Description and Visualization

For Categorical variable, from Figure 1, we can see that admin, blue-collar and technician consists of almost 65% of total frequencies of overall population.



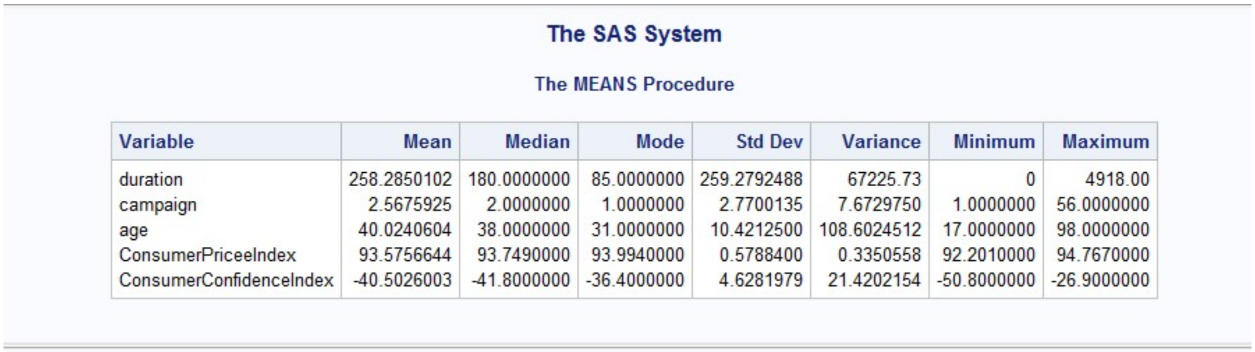
The SAS System

The FREQ Procedure

job	Frequency	Percent	Cumulative Frequency	Cumulative Percent
admin.	10422	25.30	10422	25.30
blue-col	9254	22.47	19676	47.77
entrepre	1456	3.54	21132	51.31
housemai	1060	2.57	22192	53.88
manageme	2924	7.10	25116	60.98
retired	1720	4.18	26836	65.15
self-emp	1421	3.45	28257	68.60
services	3969	9.64	32226	78.24
student	875	2.12	33101	80.37
technici	6743	16.37	39844	96.74
unemploy	1014	2.46	40858	99.20
unknown	330	0.80	41188	100.00

Figure 1: Descriptive Statistics of Categorical Variable

For numerical variables, as we can see from Figure 2 and 3, Age and ConsumerPriceIndex seem to be normally distributed whereas other variables are not. The variance for variables ranges from 0.33 to 67225.73 which seem to be a huge difference.



The SAS System

The MEANS Procedure

Variable	Mean	Median	Mode	Std Dev	Variance	Minimum	Maximum
duration	258.2850102	180.0000000	85.0000000	259.2792488	67225.73	0	4918.00
campaign	2.5675925	2.0000000	1.0000000	2.7700135	7.6729750	1.0000000	56.0000000
age	40.0240604	38.0000000	31.0000000	10.4212500	108.6024512	17.0000000	98.0000000
ConsumerPriceeIndex	93.5756644	93.7490000	93.9940000	0.5788400	0.3350558	92.2010000	94.7670000
ConsumerConfidenceIndex	-40.5026003	-41.8000000	-36.4000000	4.6281979	21.4202154	-50.8000000	-26.9000000

Figure 2: Descriptive Statistics of Numerical Variables

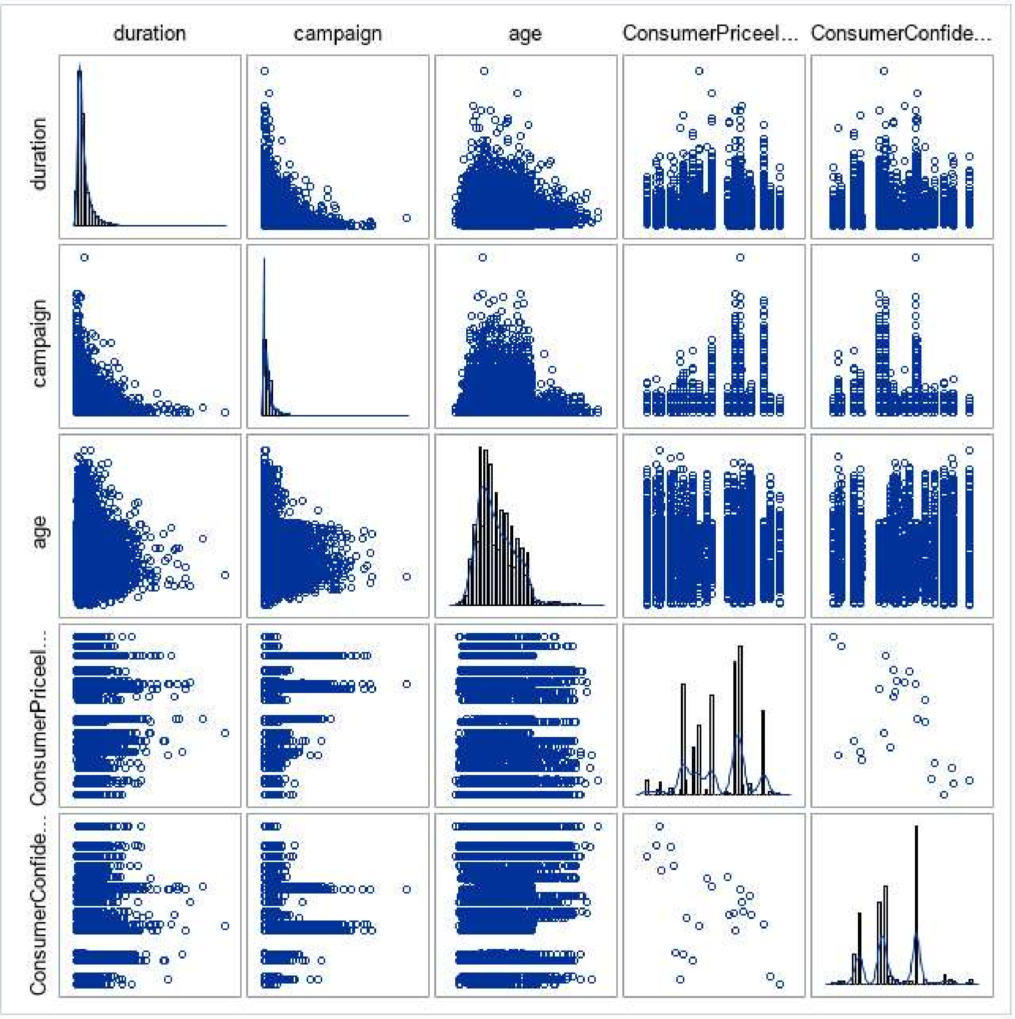
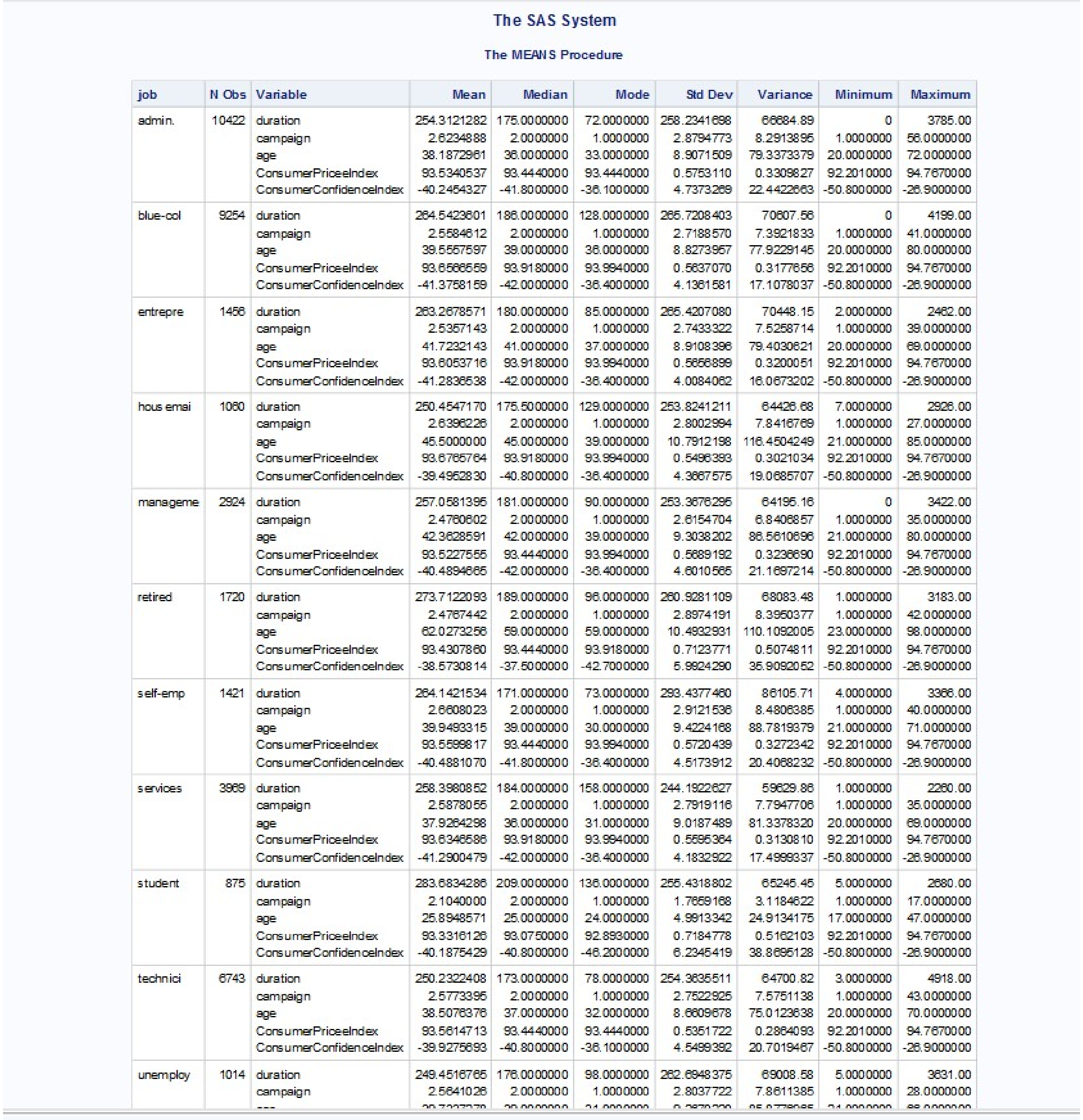


Figure 3: Graphical representation of Numerical Variables

If we make a descriptive summary of group wise numerical variables, we can see from Figure 4 & 5 that there are many outliers for duration and campaign variable. For age, there are very few numbers of outliers and consumer price index & consumer confidence index seem more stable in terms of mean, median and percentiles for every job sector.



The SAS System

The MEANS Procedure

job	N Obs	Variable	Mean	Median	Mode	Std Dev	Variance	Minimum	Maximum
admin.	10422	duration	254.3121282	175.0000000	72.0000000	258.2341698	66684.89	0	3785.00
		campaign	2.6234888	2.0000000	1.0000000	2.8794773	8.2913895	1.0000000	56.0000000
		age	38.1872961	36.0000000	33.0000000	8.9071509	79.3373379	20.0000000	72.0000000
		ConsumerPriceeIndex	93.5340537	93.4440000	93.4440000	0.5753110	0.3309827	92.2010000	94.7670000
		ConsumerConfidenceIndex	-40.2454327	-41.8000000	-36.1000000	4.7373269	22.4422663	-50.8000000	-26.9000000
blue-col	9254	duration	264.5423601	186.0000000	128.0000000	265.7208403	70607.56	0	4199.00
		campaign	2.5584612	2.0000000	1.0000000	2.7188570	7.3921833	1.0000000	41.0000000
		age	39.5557597	39.0000000	36.0000000	8.8273957	77.9229145	20.0000000	80.0000000
		ConsumerPriceeIndex	93.6566559	93.9180000	93.9940000	0.5637070	0.3177656	92.2010000	94.7670000
		ConsumerConfidenceIndex	-41.3758159	-42.0000000	-36.4000000	4.1361581	17.1078037	-50.8000000	-26.9000000
entrepre	1456	duration	263.2678571	180.0000000	85.0000000	265.4207080	70448.15	2.0000000	2462.00
		campaign	2.5357143	2.0000000	1.0000000	2.7433322	7.5258714	1.0000000	39.0000000
		age	41.7232143	41.0000000	37.0000000	8.9108396	79.4030621	20.0000000	69.0000000
		ConsumerPriceeIndex	93.6053716	93.9180000	93.9940000	0.5656899	0.3200051	92.2010000	94.7670000
		ConsumerConfidenceIndex	-41.2836538	-42.0000000	-36.4000000	4.0084062	16.0673202	-50.8000000	-26.9000000
hous emai	1060	duration	250.4547170	175.5000000	129.0000000	253.8241211	64426.68	7.0000000	2926.00
		campaign	2.6396226	2.0000000	1.0000000	2.8002994	7.8416769	1.0000000	27.0000000
		age	45.5000000	45.0000000	39.0000000	10.7912198	116.4504249	21.0000000	85.0000000
		ConsumerPriceeIndex	93.6765764	93.9180000	93.9940000	0.5496393	0.3021034	92.2010000	94.7670000
		ConsumerConfidenceIndex	-39.4952830	-40.8000000	-36.4000000	4.3667575	19.0685707	-50.8000000	-26.9000000
manageme	2924	duration	257.0581395	181.0000000	90.0000000	253.3676295	64195.16	0	3422.00
		campaign	2.4760602	2.0000000	1.0000000	2.6154704	6.8406857	1.0000000	35.0000000
		age	42.3628591	42.0000000	39.0000000	9.3038202	86.5610696	21.0000000	80.0000000
		ConsumerPriceeIndex	93.5227555	93.4440000	93.9940000	0.5689192	0.3236690	92.2010000	94.7670000
		ConsumerConfidenceIndex	-40.4894665	-42.0000000	-36.4000000	4.6010565	21.1697214	-50.8000000	-26.9000000
retired	1720	duration	273.7122093	189.0000000	96.0000000	260.9281109	68083.48	1.0000000	3183.00
		campaign	2.4767442	2.0000000	1.0000000	2.8974191	8.3950377	1.0000000	42.0000000
		age	62.0273256	59.0000000	59.0000000	10.4932931	110.1092005	23.0000000	98.0000000
		ConsumerPriceeIndex	93.4307860	93.4440000	93.9180000	0.7123771	0.5074811	92.2010000	94.7670000
		ConsumerConfidenceIndex	-38.5730814	-37.5000000	-42.7000000	5.9924290	35.9092052	-50.8000000	-26.9000000
self-emp	1421	duration	264.1421534	171.0000000	73.0000000	293.4377460	86105.71	4.0000000	3366.00
		campaign	2.6608023	2.0000000	1.0000000	2.9121536	8.4806385	1.0000000	40.0000000
		age	39.9493315	39.0000000	30.0000000	9.4224168	88.7819379	21.0000000	71.0000000
		ConsumerPriceeIndex	93.5598817	93.4440000	93.9940000	0.5720439	0.3272342	92.2010000	94.7670000
		ConsumerConfidenceIndex	-40.4881070	-41.8000000	-36.4000000	4.5173912	20.4068232	-50.8000000	-26.9000000
services	3969	duration	258.3980852	184.0000000	158.0000000	244.1922627	59629.86	1.0000000	2260.00
		campaign	2.5878055	2.0000000	1.0000000	2.7919116	7.7947706	1.0000000	35.0000000
		age	37.9264298	36.0000000	31.0000000	9.0187489	81.3378320	20.0000000	69.0000000
		ConsumerPriceeIndex	93.6346586	93.9180000	93.9940000	0.5595364	0.3130810	92.2010000	94.7670000
		ConsumerConfidenceIndex	-41.2900479	-42.0000000	-36.4000000	4.1832922	17.4999337	-50.8000000	-26.9000000
student	875	duration	283.6834286	209.0000000	136.0000000	255.4318802	65245.45	5.0000000	2680.00
		campaign	2.1040000	2.0000000	1.0000000	1.7659168	3.1184622	1.0000000	17.0000000
		age	25.8948571	25.0000000	24.0000000	4.9913342	24.9134175	17.0000000	47.0000000
		ConsumerPriceeIndex	93.3316126	93.0750000	92.8930000	0.7184778	0.5162103	92.2010000	94.7670000
		ConsumerConfidenceIndex	-40.1875429	-40.8000000	-46.2000000	6.2345419	38.8695128	-50.8000000	-26.9000000
technici	6743	duration	250.2322408	173.0000000	78.0000000	254.3635511	64700.82	3.0000000	4918.00
		campaign	2.5773395	2.0000000	1.0000000	2.7522925	7.5751138	1.0000000	43.0000000
		age	38.5076376	37.0000000	32.0000000	8.6609678	75.0123638	20.0000000	70.0000000
		ConsumerPriceeIndex	93.5614713	93.4440000	93.4440000	0.5351722	0.2864093	92.2010000	94.7670000
		ConsumerConfidenceIndex	-39.9275693	-40.8000000	-36.1000000	4.5499392	20.7019467	-50.8000000	-26.9000000
unemploy	1014	duration	249.4516765	176.0000000	98.0000000	262.6948375	69008.58	5.0000000	3631.00
		campaign	2.5641026	2.0000000	1.0000000	2.8037722	7.8611385	1.0000000	28.0000000

Figure 4: Descriptive Statistics of Numerical Variables using Job as a grouping variable

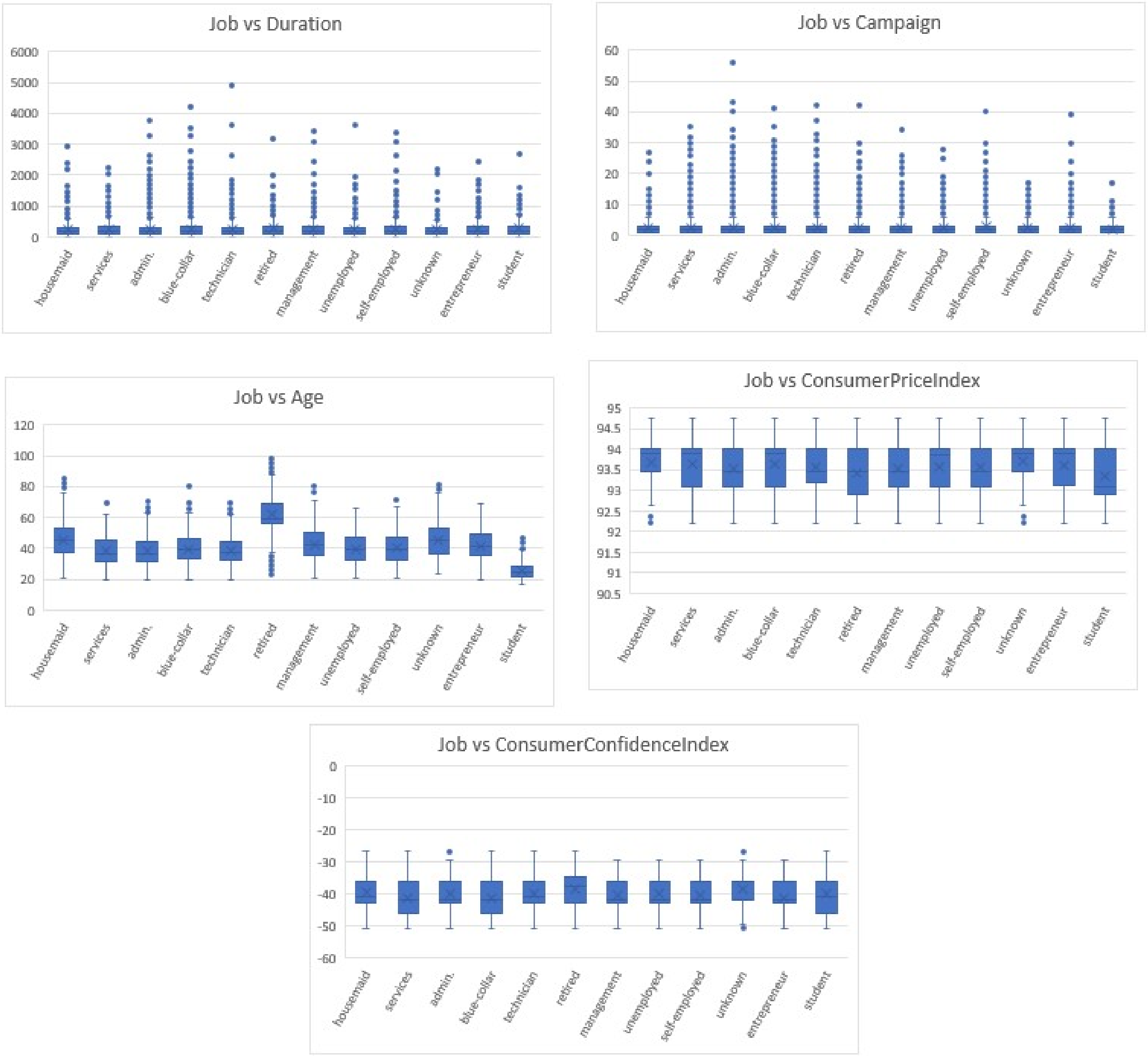


Figure 5: Graphical representation of Numerical Variables using Job as a grouping variable

3.Methods and Results:

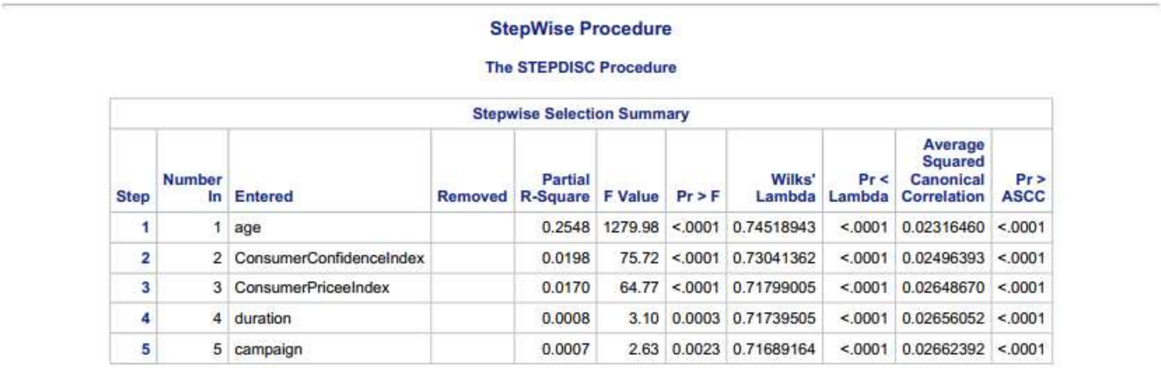
3.1. Research Question 1:

Are all variables equally important for this research? Or can we eliminate some of them using some procedure?

Method Used: Stepwise Procedure.

Method Description: In many applications, there are many dependent variables available for analysis and researchers are interested in discarding redundant variables for separating the groups. Stepwise Procedure is a combination of forward selection and backward elimination. Variables are added one at a time. At each step, the variables are reexamined to see if any previously selected variable has become redundant in the presence of the recently added variables. The procedure stops when the largest partial F-statistic among the remaining variables available for entry fails to exceed the preset significance threshold.

SAS Interpretation: Applying Stepwise procedure in SAS, we can see that there were 6 steps in total before getting into conclusion for this dataset. And no variables were removed from the procedure. So, age, consumer confidence index, consumer price index, duration, campaign- these 5 variables will be used for the further procedures.



StepWise Procedure

The STEPDISC Procedure

Stepwise Selection Summary

Step	Number In	Entered	Removed	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
1	1	age		0.2548	1279.98	<.0001	0.74518943	<.0001	0.02316460	<.0001
2	2	ConsumerConfidenceIndex		0.0198	75.72	<.0001	0.73041362	<.0001	0.02496393	<.0001
3	3	ConsumerPriceeIndex		0.0170	64.77	<.0001	0.71799005	<.0001	0.02648670	<.0001
4	4	duration		0.0008	3.10	0.0003	0.71739505	<.0001	0.02656052	<.0001
5	5	campaign		0.0007	2.63	0.0023	0.71689164	<.0001	0.02662392	<.0001

Figure 6: Stepwise Selection Summary

3.2. Research Question 2:

What kind of classification can be used in the dataset if we consider “Job” as a grouping variable? Does error rate change if we use both resubstitution and cross validation summary?

Method Used: Discrim Procedure

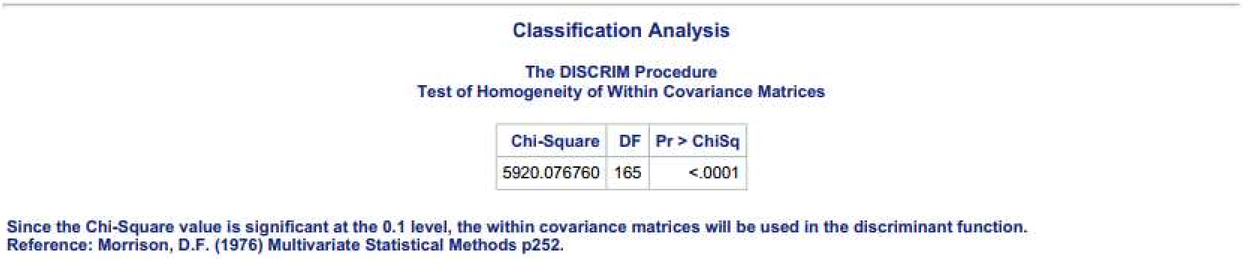
Method Description: Given a set of observations that contains one or more quantitative variables and a classification variable which indexes groups of observations, the DISCRIM procedure develops a discriminant criterion to classify each observation into one of the groups. If we are not sure of the equality of population covariance matrices, we can use a Chi-square test to identify if we should use Linear discriminant function or quadratic function to classify the observations.

To judge the ability of classification procedures to predict group membership, the probability of misclassification, or error rate, is usually used. We have used two types of error rate here – one is called resubstitution and another one is cross validation method.

Using resubstitution method, the classification rule is applied to each observation vector and this observation is assigned to a group. Then, the number of misclassifications is counted. The proportion of misclassifications resulting from resubstitution is call the apparent error rate.

In the cross-validation method, all but one observation vector is used to determine the classification rule and then the omitted observation is classified into one of the groups using the classification rule based on the N-1 observations. This procedure is repeated until every observation has been classified in this manner.

SAS Interpretation: As the Discrim procedure shows (Figure 7) the p value of chi-square test is less than 0.0001, we can safely say that quadratic classification was applied to this dataset.



Classification Analysis

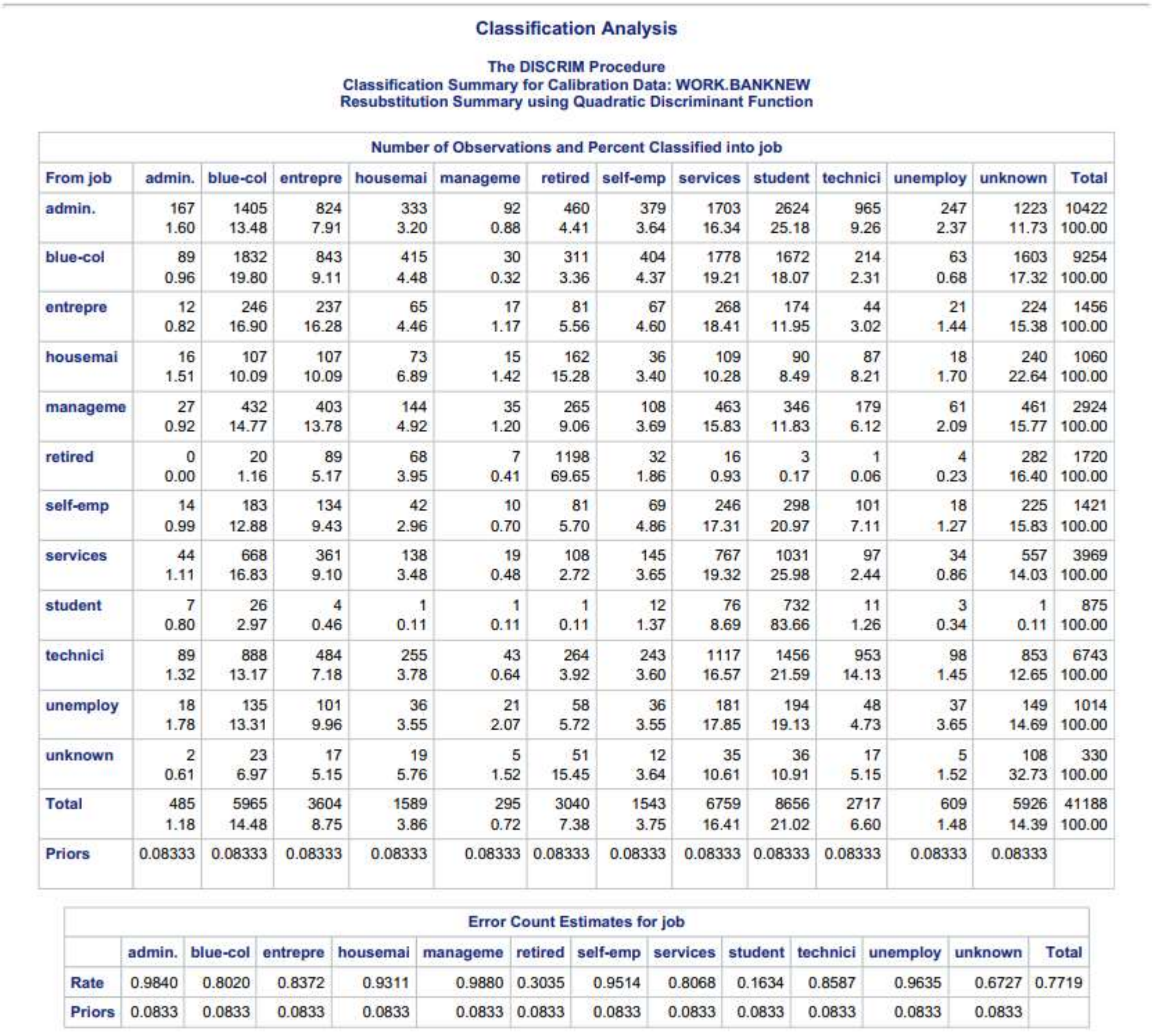
The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
5920.076760	165	<.0001

Since the Chi-Square value is significant at the 0.1 level, the within covariance matrices will be used in the discriminant function.
Reference: Morrison, D.F. (1976) Multivariate Statistical Methods p252.

Figure 7: Test of Homogeneity of within covariance Matrices

From Figure 8 & 9, we can see that there is not much difference between resubstitution & cross validation method in terms of overall error rate. All the variables also showed similar error rates individually.



Classification Analysis

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.BANKNEW
Resubstitution Summary using Quadratic Discriminant Function

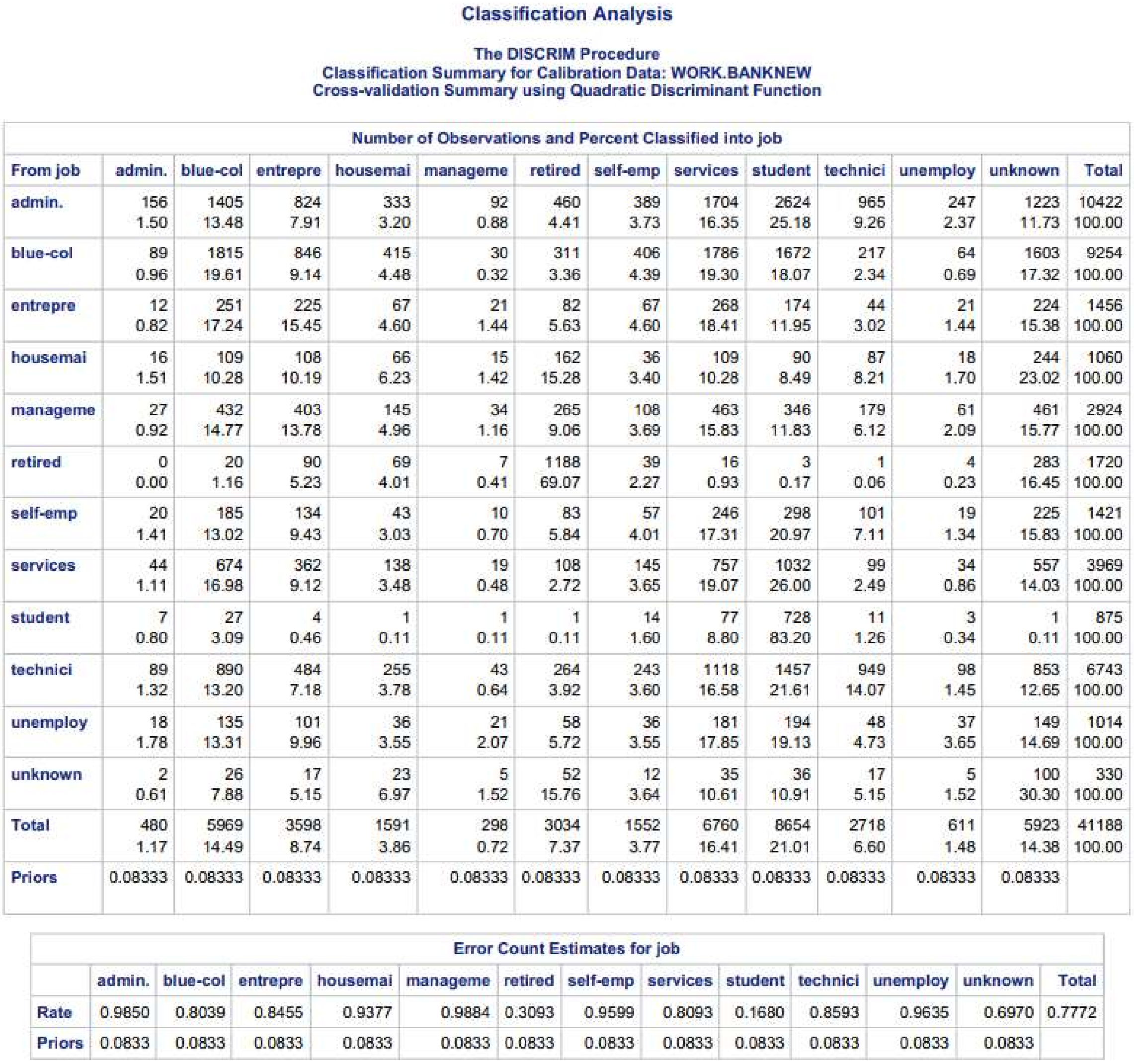
Number of Observations and Percent Classified into job

From job	admin.	blue-col	entrepre	housemai	manageme	retired	self-emp	services	student	technici	unemploy	unknown	Total
admin.	167 1.60	1405 13.48	824 7.91	333 3.20	92 0.88	460 4.41	379 3.64	1703 16.34	2624 25.18	965 9.26	247 2.37	1223 11.73	10422 100.00
blue-col	89 0.96	1832 19.80	843 9.11	415 4.48	30 0.32	311 3.36	404 4.37	1778 19.21	1672 18.07	214 2.31	63 0.68	1603 17.32	9254 100.00
entrepre	12 0.82	246 16.90	237 16.28	65 4.46	17 1.17	81 5.56	67 4.60	268 18.41	174 11.95	44 3.02	21 1.44	224 15.38	1456 100.00
housemai	16 1.51	107 10.09	107 10.09	73 6.89	15 1.42	162 15.28	36 3.40	109 10.28	90 8.49	87 8.21	18 1.70	240 22.64	1060 100.00
manageme	27 0.92	432 14.77	403 13.78	144 4.92	35 1.20	265 9.06	108 3.69	463 15.83	346 11.83	179 6.12	61 2.09	461 15.77	2924 100.00
retired	0 0.00	20 1.16	89 5.17	68 3.95	7 0.41	1198 69.65	32 1.86	16 0.93	3 0.17	1 0.06	4 0.23	282 16.40	1720 100.00
self-emp	14 0.99	183 12.88	134 9.43	42 2.96	10 0.70	81 5.70	69 4.86	246 17.31	298 20.97	101 7.11	18 1.27	225 15.83	1421 100.00
services	44 1.11	668 16.83	361 9.10	138 3.48	19 0.48	108 2.72	145 3.65	767 19.32	1031 25.98	97 2.44	34 0.86	557 14.03	3969 100.00
student	7 0.80	26 2.97	4 0.46	1 0.11	1 0.11	1 0.11	12 1.37	76 8.69	732 83.66	11 1.26	3 0.34	1 0.11	875 100.00
technici	89 1.32	888 13.17	484 7.18	255 3.78	43 0.64	264 3.92	243 3.60	1117 16.57	1456 21.59	953 14.13	98 1.45	853 12.65	6743 100.00
unemploy	18 1.78	135 13.31	101 9.96	36 3.55	21 2.07	58 5.72	36 3.55	181 17.85	194 19.13	48 4.73	37 3.65	149 14.69	1014 100.00
unknown	2 0.61	23 6.97	17 5.15	19 5.76	5 1.52	51 15.45	12 3.64	35 10.61	36 10.91	17 5.15	5 1.52	108 32.73	330 100.00
Total	485 1.18	5965 14.48	3604 8.75	1589 3.86	295 0.72	3040 7.38	1543 3.75	6759 16.41	8656 21.02	2717 6.60	609 1.48	5926 14.39	41188 100.00
Priors	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	

Error Count Estimates for job

	admin.	blue-col	entrepre	housemai	manageme	retired	self-emp	services	student	technici	unemploy	unknown	Total
Rate	0.9840	0.8020	0.8372	0.9311	0.9880	0.3035	0.9514	0.8068	0.1634	0.8587	0.9635	0.6727	0.7719
Priors	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	

Figure 8: Resubstitution Summary



Classification Analysis

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.BANKNEW
Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into job

From job	admin.	blue-col	entrepre	housemai	manageme	retired	self-emp	services	student	technici	unemploy	unknown	Total
admin.	156 1.50	1405 13.48	824 7.91	333 3.20	92 0.88	460 4.41	389 3.73	1704 16.35	2624 25.18	965 9.26	247 2.37	1223 11.73	10422 100.00
blue-col	89 0.96	1815 19.61	846 9.14	415 4.48	30 0.32	311 3.36	406 4.39	1786 19.30	1672 18.07	217 2.34	64 0.69	1603 17.32	9254 100.00
entrepre	12 0.82	251 17.24	225 15.45	67 4.60	21 1.44	82 5.63	67 4.60	268 18.41	174 11.95	44 3.02	21 1.44	224 15.38	1456 100.00
housemai	16 1.51	109 10.28	108 10.19	66 6.23	15 1.42	162 15.28	36 3.40	109 10.28	90 8.49	87 8.21	18 1.70	244 23.02	1060 100.00
manageme	27 0.92	432 14.77	403 13.78	145 4.96	34 1.16	265 9.06	108 3.69	463 15.83	346 11.83	179 6.12	61 2.09	461 15.77	2924 100.00
retired	0 0.00	20 1.16	90 5.23	69 4.01	7 0.41	1188 69.07	39 2.27	16 0.93	3 0.17	1 0.06	4 0.23	283 16.45	1720 100.00
self-emp	20 1.41	185 13.02	134 9.43	43 3.03	10 0.70	83 5.84	57 4.01	246 17.31	298 20.97	101 7.11	19 1.34	225 15.83	1421 100.00
services	44 1.11	674 16.98	362 9.12	138 3.48	19 0.48	108 2.72	145 3.65	757 19.07	1032 26.00	99 2.49	34 0.86	557 14.03	3969 100.00
student	7 0.80	27 3.09	4 0.46	1 0.11	1 0.11	1 0.11	14 1.60	77 8.80	728 83.20	11 1.26	3 0.34	1 0.11	875 100.00
technici	89 1.32	890 13.20	484 7.18	255 3.78	43 0.64	264 3.92	243 3.60	1118 16.58	1457 21.61	949 14.07	98 1.45	853 12.65	6743 100.00
unemploy	18 1.78	135 13.31	101 9.96	36 3.55	21 2.07	58 5.72	36 3.55	181 17.85	194 19.13	48 4.73	37 3.65	149 14.69	1014 100.00
unknown	2 0.61	26 7.88	17 5.15	23 6.97	5 1.52	52 15.76	12 3.64	35 10.61	36 10.91	17 5.15	5 1.52	100 30.30	330 100.00
Total	480 1.17	5969 14.49	3598 8.74	1591 3.86	298 0.72	3034 7.37	1552 3.77	6760 16.41	8654 21.01	2718 6.60	611 1.48	5923 14.38	41188 100.00
Priors	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	

Error Count Estimates for job

	admin.	blue-col	entrepre	housemai	manageme	retired	self-emp	services	student	technici	unemploy	unknown	Total
Rate	0.9850	0.8039	0.8455	0.9377	0.9884	0.3093	0.9599	0.8093	0.1680	0.8593	0.9635	0.6970	0.7772
Priors	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	

Figure 9: Cross-validation Summary

3.3. Research Question 3:

Using different MANOVA methods, can we decide on the equality of population mean vector of numeric variables?

Method Used: GLM Procedure.

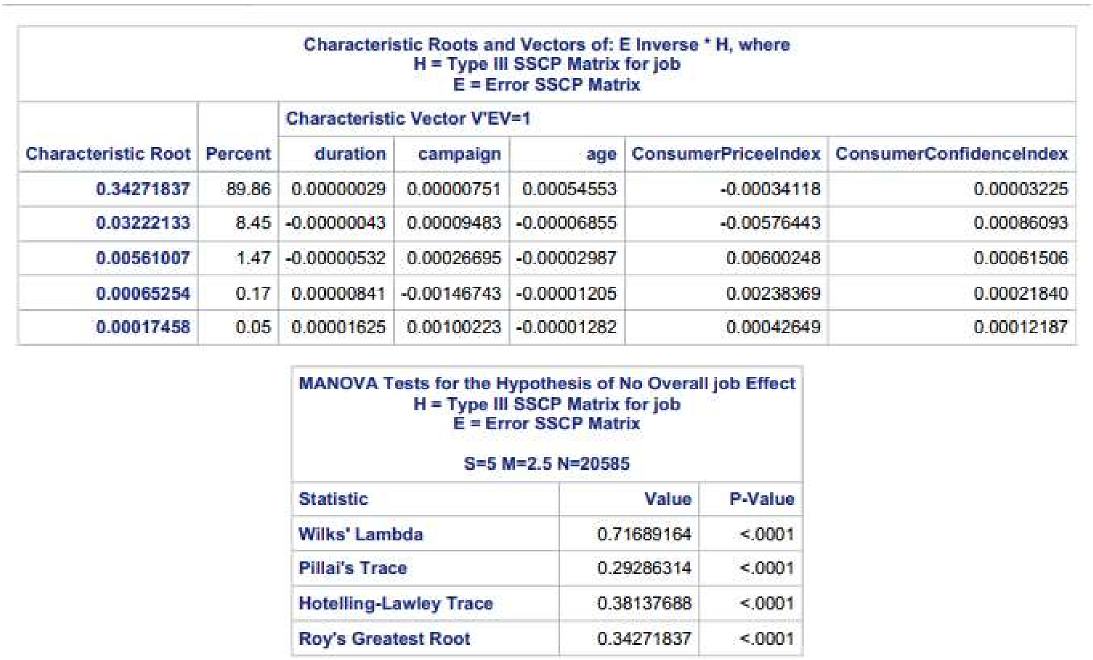
Method Description: I used this process to test for significant difference between at least two mean vectors. The hypothesis I want to test is

H_0: $\mu_{1=}$ $\mu_{2=\ldots\ldots\ldots.}$ $= \mu_k$ vs H_1=at least two population mean vector differ

From GLM procedure, we can test this hypothesis by 4 different methods. (Wilks' lambda, Pillai's trace, Hotelling-Lawley Trace & Roy's greatest root)

SAS Interpretation: We can see from the hypothesis test that for all methods mentioned here (Wilks' lambda, Pillai's trace, Hotelling-Lawley Trace & Roy's greatest root), the P-values are significant. So, we can reject our null hypothesis and conclude that there is significant difference among the mean vectors of population variables. We can also rank the variables based on their coefficient of 1st discriminate function as:

Age>ConsumerPriceIndex>ConsumerConfidenceIndex>Campaign>Duration



Characteristic Roots and Vectors of: E Inverse * H, where
H = Type III SSCP Matrix for job
E = Error SSCP Matrix

		Characteristic Vector V'EV=1				
Characteristic Root	Percent	duration	campaign	age	ConsumerPriceIndex	ConsumerConfidenceIndex
0.34271837	89.86	0.00000029	0.00000751	0.00054553	-0.00034118	0.00003225
0.03222133	8.45	-0.00000043	0.00009483	-0.00006855	-0.00576443	0.00086093
0.00561007	1.47	-0.00000532	0.00026695	-0.00002987	0.00600248	0.00061506
0.00065254	0.17	0.00000841	-0.00146743	-0.00001205	0.00238369	0.00021840
0.00017458	0.05	0.00001625	0.00100223	-0.00001282	0.00042649	0.00012187

MANOVA Tests for the Hypothesis of No Overall job Effect
H = Type III SSCP Matrix for job
E = Error SSCP Matrix

S=5 M=2.5 N=20585

Statistic	Value	P-Value
Wilks' Lambda	0.71689164	<.0001
Pillai's Trace	0.29286314	<.0001
Hotelling-Lawley Trace	0.38137688	<.0001
Roy's Greatest Root	0.34271837	<.0001

Figure 10: Applying MANOVA Method

3.4. Research Question 4:

How many factors can be used to explain the total variances of those variables? Does rotating can be used to interpret the factors for better?

Method Used: Principal Component Method in the Factor Procedure & Orthogonal Rotation using Varimax.

Method Description: Principal component method is a factor extraction method used to form uncorrelated linear combinations of the observed variables. The first component has maximum variance. Successive components explain progressively smaller portions of the variance and are all uncorrelated with each other. Principal components analysis is used to obtain the initial factor solution.

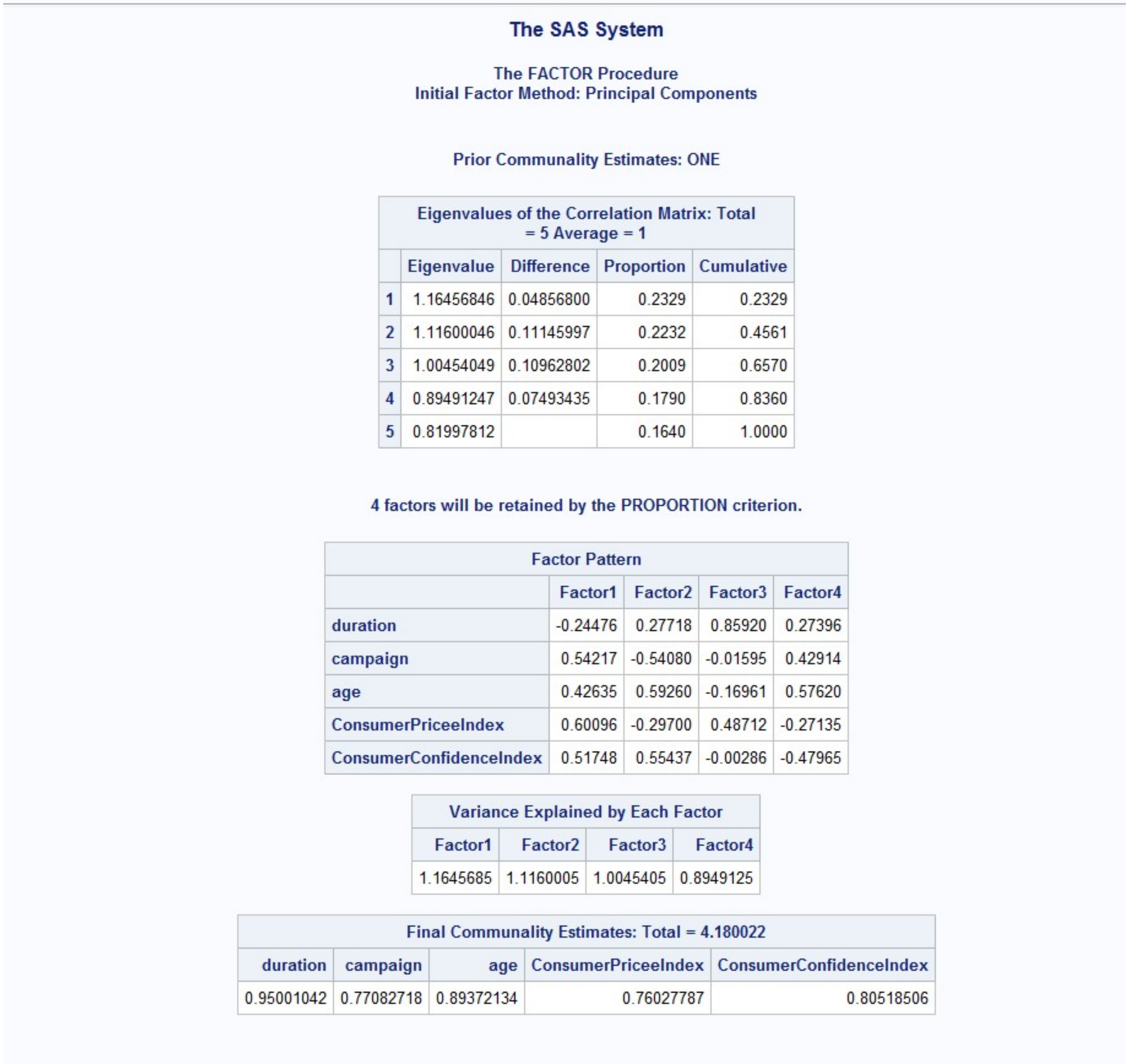
Varimax Method is an orthogonal rotation method that minimizes the number of variables that have high loadings on each factor. This method simplifies the interpretation of the factors.

SAS Interpretation: 4 factors will be retained for this dataset.

Before rotating, we can see that duration can be explained by Factor 3.

Campaign, Age & Consumer Confidence Index is associated with Factor1, Factor2 & Factor 4 almost equally.

Consumer Price Index can be explained by both Factor 1 & Factor 3.



The SAS System

The FACTOR Procedure
Initial Factor Method: Principal Components

Prior Communality Estimates: ONE

Eigenvalues of the Correlation Matrix: Total = 5 Average = 1				
	Eigenvalue	Difference	Proportion	Cumulative
1	1.16456846	0.04856800	0.2329	0.2329
2	1.11600046	0.11145997	0.2232	0.4561
3	1.00454049	0.10962802	0.2009	0.6570
4	0.89491247	0.07493435	0.1790	0.8360
5	0.81997812		0.1640	1.0000

4 factors will be retained by the PROPORTION criterion.

Factor Pattern				
	Factor1	Factor2	Factor3	Factor4
duration	-0.24476	0.27718	0.85920	0.27396
campaign	0.54217	-0.54080	-0.01595	0.42914
age	0.42635	0.59260	-0.16961	0.57620
ConsumerPriceeIndex	0.60096	-0.29700	0.48712	-0.27135
ConsumerConfidenceIndex	0.51748	0.55437	-0.00286	-0.47965

Variance Explained by Each Factor			
Factor1	Factor2	Factor3	Factor4
1.1645685	1.1160005	1.0045405	0.8949125

Final Communality Estimates: Total = 4.180022				
duration	campaign	age	ConsumerPriceeIndex	ConsumerConfidenceIndex
0.95001042	0.77082718	0.89372134	0.76027787	0.80518506

Figure 11: Factor Procedure before Rotation

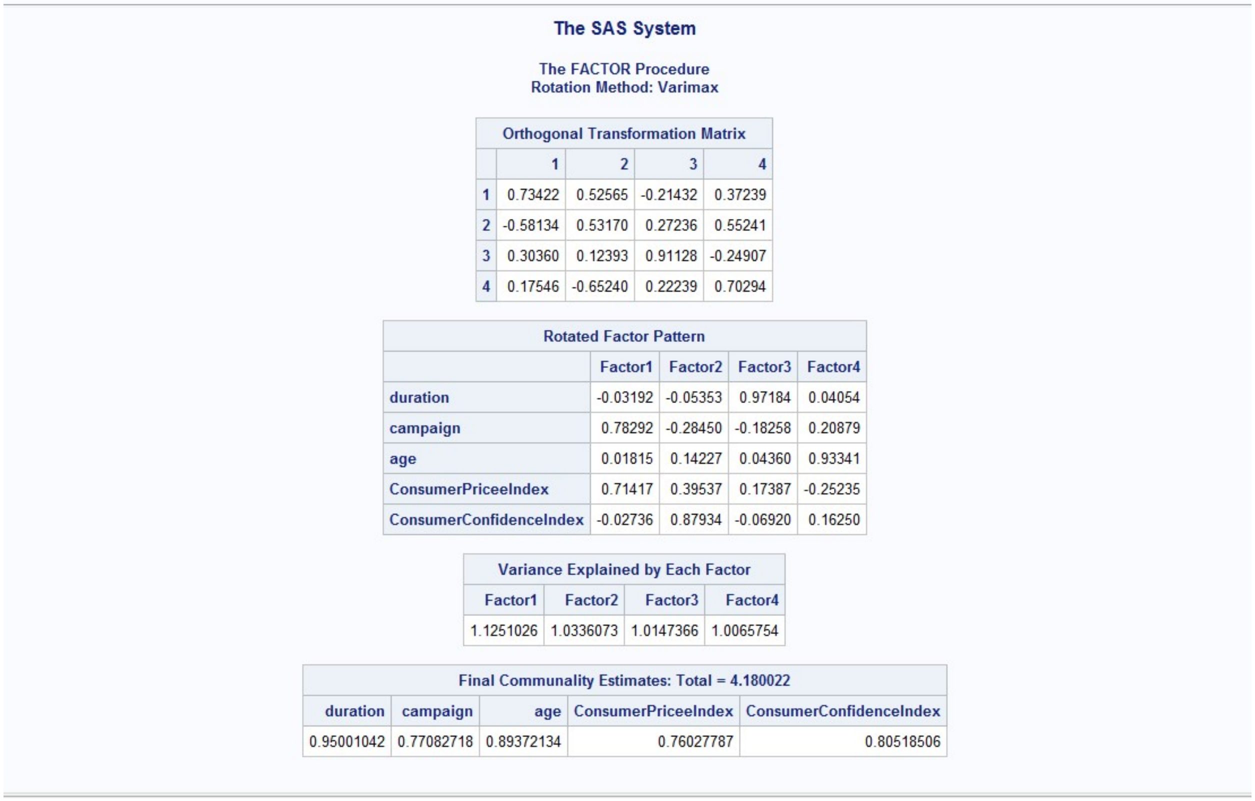
After rotating, we can see that duration is still explained by Factor 3.

Campaign & Consumer Price Index are associated with Factor 1.

Age is explained by Factor 4.

Consumer Confidence Index is associated with Factor 2.

So, except for duration, every variable association with factors are more interpretable after rotation. Also for duration, the proportion gets better after applying rotation.



The SAS System

The FACTOR Procedure
Rotation Method: Varimax

Orthogonal Transformation Matrix

	1	2	3	4
1	0.73422	0.52565	-0.21432	0.37239
2	-0.58134	0.53170	0.27236	0.55241
3	0.30360	0.12393	0.91128	-0.24907
4	0.17546	-0.65240	0.22239	0.70294

Rotated Factor Pattern

	Factor1	Factor2	Factor3	Factor4
duration	-0.03192	-0.05353	0.97184	0.04054
campaign	0.78292	-0.28450	-0.18258	0.20879
age	0.01815	0.14227	0.04360	0.93341
ConsumerPriceeIndex	0.71417	0.39537	0.17387	-0.25235
ConsumerConfidenceIndex	-0.02736	0.87934	-0.06920	0.16250

Variance Explained by Each Factor

Factor1	Factor2	Factor3	Factor4
1.1251026	1.0336073	1.0147366	1.0065754

Final Communality Estimates: Total = 4.180022

duration	campaign	age	ConsumerPriceeIndex	ConsumerConfidenceIndex
0.95001042	0.77082718	0.89372134	0.76027787	0.80518506

Figure 12: Factor Procedure after Rotation

4.Conclusions:

From our multivariate analysis, we can see that all the numeric variables will be used in the process we have done so far. We can use quadratic discriminant function to classify the observations though error rates will be almost similar for resubstitution & cross-validation method. From one-way MANOVA, we can see that at least two population mean vectors differs significantly using four different methods. And four factors will be used to explain the variables. After rotation, the interpretation of factor & variable association will be more straightforward.

5.Reference:

1. UCI Machine Learning Repository: https://data.world/uci/bank-marketing
2. SPSS Statistics Documentation: https://www.ibm.com/docs/en/spss-statistics/24.0.0?topic=option-factor-analysis
3. Performing Exploratory Data Analysis with SAS and Python: https://www.analyticsvidhya.com/blog/2021/06/exploratory-data-analysis-with-sas-and-python/
4. SAS Help Center: https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.3/casref/p1k1no304sy91on1qjvks0ggelb5.htm
5. Data Driven Approach to Predict Success of Bank- Marketing: https://medium.com/mlearning-ai/data-driven-approach-to-predict-success-of-bank-marketing-31791cad8f81

6.SAS Code & Output:

```
/*Reading Text File*/
DATA banknew;
INFILE 'banknew2.txt' dlm=',' firstobs=2 lrecl=32767;
INPUT job$ marital$ education$ contact$ duration campaign age ConsumerPriceeIndex
ConsumerConfidenceIndex y$;


/*Categorical Variable_Descriptive Statistics*/
proc freq data=banknew;
tables job;
run;


/*Numerical Variable _Descriptive Statistics*/
proc means data=banknew mean median mode std var min max;
run;


/*Categorical+ Numerical Variable _Descriptive Statistics*/
proc means data=banknew mean median mode std var min max;
class job;
var duration campaign age ConsumerPriceeIndex ConsumerConfidenceIndex;
run;


/*Scatter Plot by Group*/
proc sgscatter data=banknew;
matrix duration campaign age ConsumerPriceeIndex ConsumerConfidenceIndex
/diagonal=(histogram kernel);
run;


/*Stepwise Procedure*/
PROC STEPDISC STEPWISE;
 CLASS job;
Run;
```

```
/*Classification Analysis'*/
PROC DISCRIM LIST CROSSVALIDATE POOL=TEST;
 CLASS job;
 VAR duration campaign age ConsumerPriceeIndex ConsumerConfidenceIndex;
RUN;


/* MANOVA */
PROC GLM;
 CLASS job;
MODEL duration campaign age ConsumerPriceeIndex ConsumerConfidenceIndex=job;
 MANOVA H=job/PRINTE PRINTH MSTAT=EXACT;
RUN;


/*Principal Component Analysis of Factor Method*/
proc factor method=prin rotate=varimax proportion=0.8 corr;
 var duration campaign age ConsumerPriceeIndex ConsumerConfidenceIndex;
run;
```