1.- Question or problem definition.

We will try to identify using the data presented in the exercise database whether there is a direct relationship between the work performed by each of the bank's customers and their account balance in order to identify whether there is a group that is more attractive to the bank because of the banking security they offer.

2.- Acquire training and testing data.

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls.

3.- Wrangle, prepare, cleanse the data.

In order to be able to answer this question we will extract from the database the work column where the type of work performed by the client appears and the balance column which indicates what has happened in the last year in the client's bank account, either positive or negative.

We will then divide this information into thirteen columns, one for each type of work available, and with the values of all the available balances.

4.- Analyze, identify patterns, and explore the data.

the first step in analyzing the data was to find out some statistical data on the different jobs to see if there were any differences.

admin.	
Media	1226.736
Error típico	108.4067
Mediana	430
Moda	0
Desviación estándar	2370.119
Varianza de la muestra	5617465
Curtosis	31.25918
Coeficiente de asimetría	4.748747
Rango	23138
Mínimo	-967
Máximo	22171
Suma	586380
Cuenta	478

blue-collar	
Media	1085.162
Error típico	66.33322
Mediana	408.5
Moda	0
Desviación estándar	2040.218
Varianza de la muestra	4162490
Curtosis	18.09018
Coeficiente de asimetría	3.750573
Rango	17753
Mínimo	-1400
Máximo	16353
Suma	1026563
Cuenta	946

entrepreneur	
Media	1645.125
Error típico	342.654
Mediana	365.5
Moda	0
Desviación estándar	4441.304
Varianza de la muestra	19725178
Curtosis	45.66067
Coeficiente de asimetría	6.049316
Rango	44127
Mínimo	-2082
Máximo	42045
Suma	276381
Cuenta	168

housemaid	
Media	2083.804
Error típico	435.0217
Mediana	296.5
Moda	0
Desviación estándar	4603.837
Varianza de la muestra	21195312
Curtosis	13.79546
Coeficiente de asimetría	3.538975
Rango	27724
Mínimo	-759
Máximo	26965
Suma	233386
Cuenta	112

management	
Media	1766.929
Error típico	104.9747
Mediana	577
Moda	0
Desviación estándar	3267.733
Varianza de la muestra	10678079
Curtosis	17.04516
Coeficiente de asimetría	3.642617
Rango	29105
Mínimo	-1746
Máximo	27359
Suma	1712154
Cuenta	969

retired	
Media	2319.191
Error típico	385.4989
Mediana	672.5
Moda	0
Desviación estándar	5846.38
Varianza de la muestra	34180158
Curtosis	86.43204
Coeficiente de asimetría	8.071526
Rango	72394
Mínimo	-1206
Máximo	71188
Suma	533414
Cuenta	230

self-employed	
Media	1392.41
Error típico	183.3003
Mediana	483
Moda	0
Desviación estándar	2479.641
Varianza de la muestra	6148619
Curtosis	12.65537
Coeficiente de asimetría	3.183122
Rango	19743
Mínimo	-3313
Máximo	16430
Suma	254811
Cuenta	183

services	
Media	1103.957
Error típico	119.7439
Mediana	288
Moda	0
Desviación estándar	2445.24
Varianza de la muestra	5979199
Curtosis	42.60244
Coeficiente de asimetría	5.402962
Rango	27596
Mínimo	-1202
Máximo	26394
Suma	460350
Cuenta	417

student	
Media	1543.821
Error típico	281.4887
Mediana	422.5
Moda	0
Desviación estándar	2579.887
Varianza de la muestra	6655815
Curtosis	5.134617
Coeficiente de asimetría	2.368667
Rango	11785
Mínimo	-230
Máximo	11555
Suma	129681
Cuenta	84

technician	
Media	1330.996
Error típico	94.91109
Mediana	434.5
Moda	0
Desviación estándar	2630.253
Varianza de la muestra	6918233
Curtosis	30.61018
Coeficiente de asimetría	4.494781
Rango	29413
Mínimo	-1680
Máximo	27733
Suma	1022205
Cuenta	768

unemployed	
Media	1089.422
Error típico	149.5767
Mediana	473.5
Moda	0
Desviación estándar	1692.268
Varianza de la muestra	2863770
Curtosis	6.674153
Coeficiente de asimetría	2.350104
Rango	9891
Mínimo	-872
Máximo	9019
Suma	139446
Cuenta	128

Analyzing these first statistical data, we realize that there are notable differences in the balances if we group the sample into jobs.

We know that, on average, retirees and household employees are the people with the highest balance, but they are also the jobs in which there is the highest variance.

5.- Model

To verify that the differences between the different papers are significant and that there is a considerable difference depending on the paper, we performed an ANOVA analysis.

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
admin.	477	564209	1182.828	4705775
blue-collar	945	1010210	1069.005	3919703
entrepreneur	167	234336	1403.21	9952900
housemaid	111	206421	1859.649	15709349
management	968	1684795	1740.491	10011117
retired	229	462226	2018.454	13436982
self-employed	182	238381	1309.786	4926393
services	416	433956	1043.163	4448730
student	83	118126	1423.205	5500018
technician	767	994472	1296.574	6016070
unemployed	127	130427	1026.984	2383535

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	4.22E+08	10	42215888	6.243729	1.47E-09	1.83282
Within Groups	3.02E+10	4461	6761327			
Total	3.06E+10	4471				

With the ANOVA analysis done, we can prove that with a 95% certainty there are notable differences in the means of the variables or jobs we are evaluating.