Data Preparation

1. Loaded file using *pandas.read_csv*. Separator was ";". I used quoting=3 because there were some Unicode errors with quotes in the file which was causing certain elements to be in the wrong column. This way when separating elements *pandas* does not care about quotes. Later used *converters* to remove the quotes around the strings.
2. Data had appropriate types, no steps necessary
3. Found typos using *dataframe.value_counts().* Fixed them using *dataframe.mask().*
4. Removed whitespaces using *converter* during the loading file step
5. Casted data to lowercase using *converter* during the loading file step
6. For sanity checks given the description of the data, I assumed that age < 120, if duration=0 then y=no, and if pdays=999 then poutcome!=success. 2 rows were removed for the age check.
7. Filled missing values using *dataframe.fillna()*

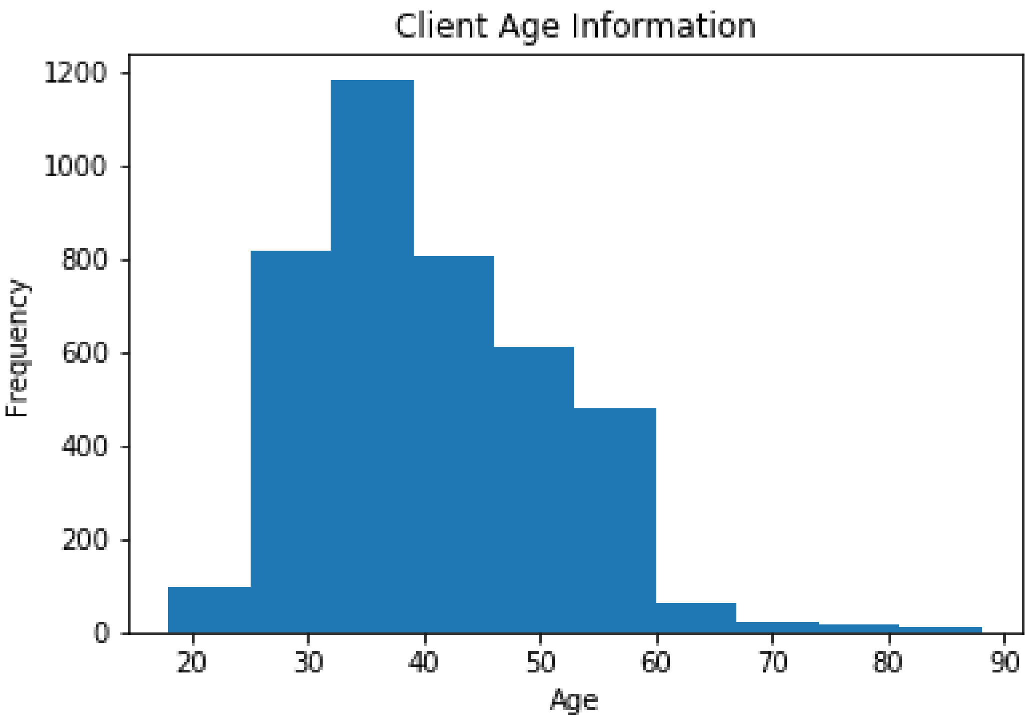
Data Exploration

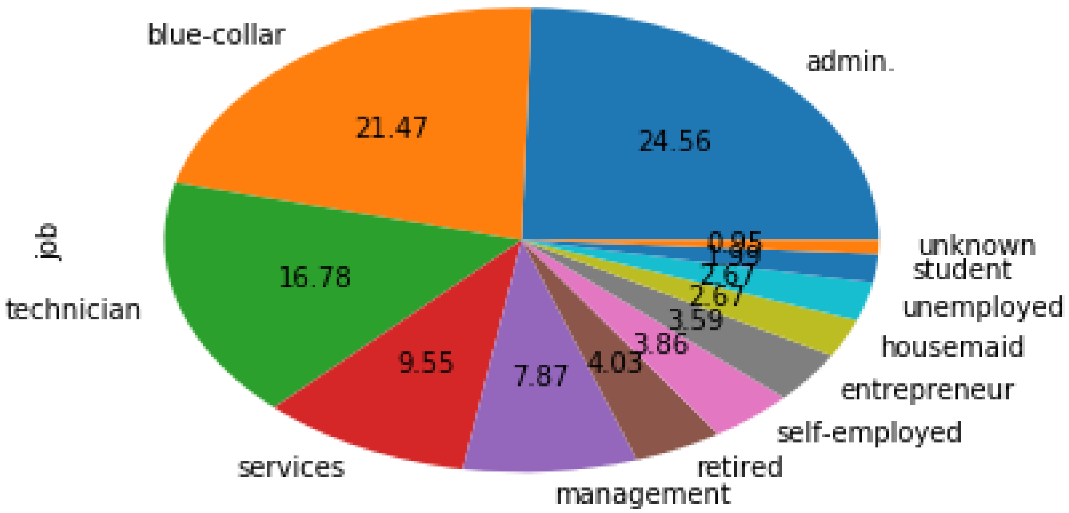
Subsection 1

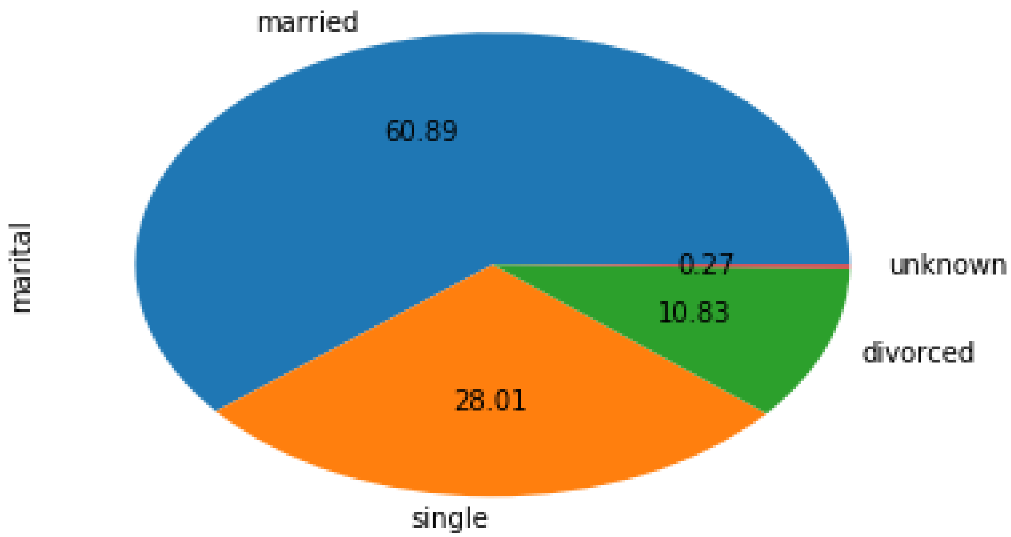
Why I chose a certain type of graph:

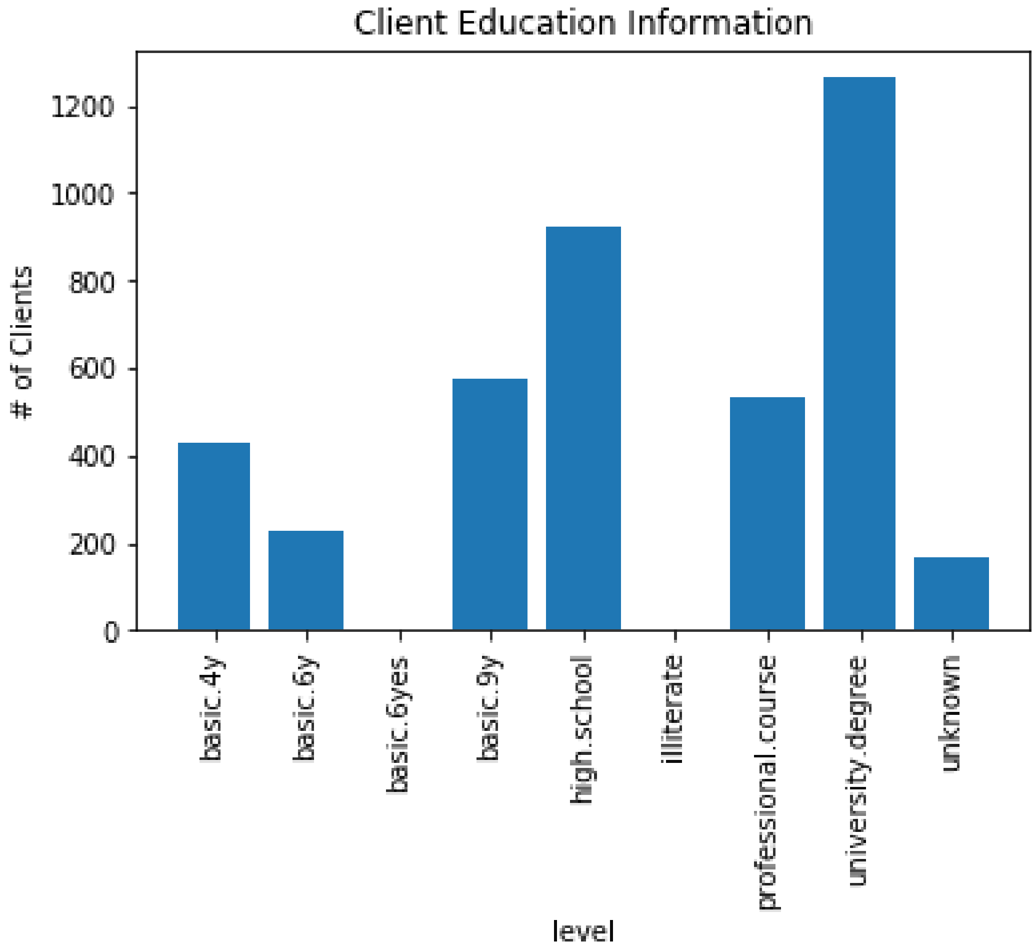
- Pie Graph: used to show proportional relationship of values (Eg: proportion of people who are single vs married). Used for categorical and binary data. Except for *education* because there were couple values where the proportion was too small to see.
- Box plot: used when data had many outliers, in this case, *duration and campaign* columns. Also helps to show the data was left skewed
- Histograms: to show distribution and when data was easily divisible into bins
- Bar graph: used when pie graphs would have been unreadable

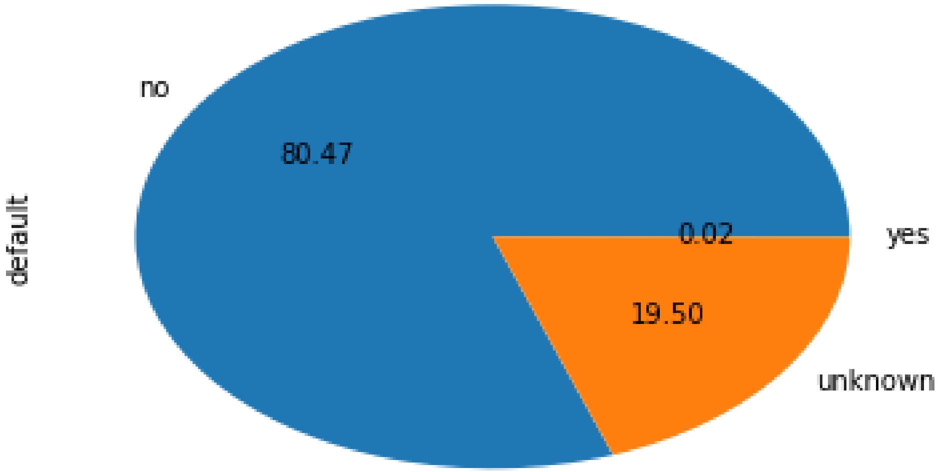
Graphs:

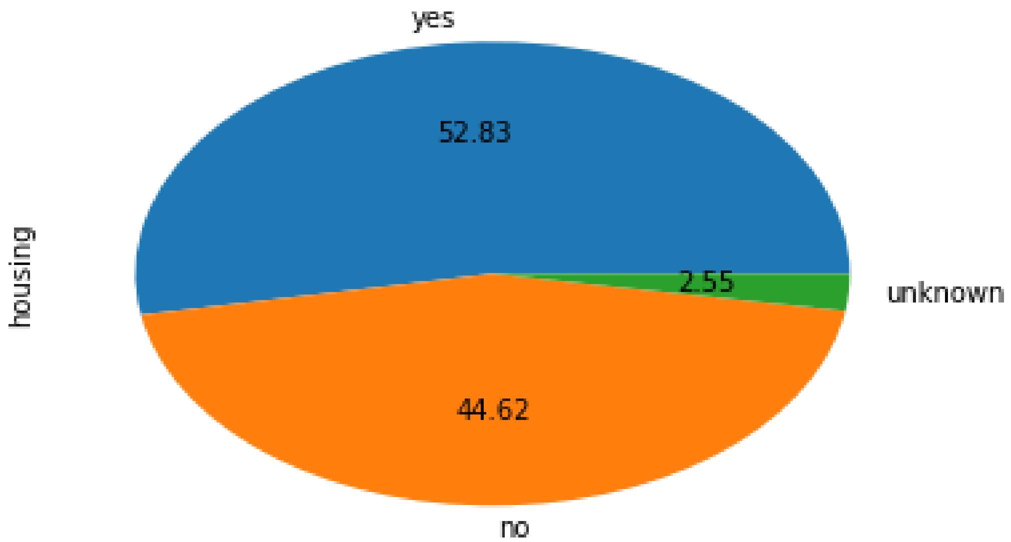

Client Age Information
Frequency
1200
1000
800
600
400
200
0
20
30
40
50
60
70
80
90
Age

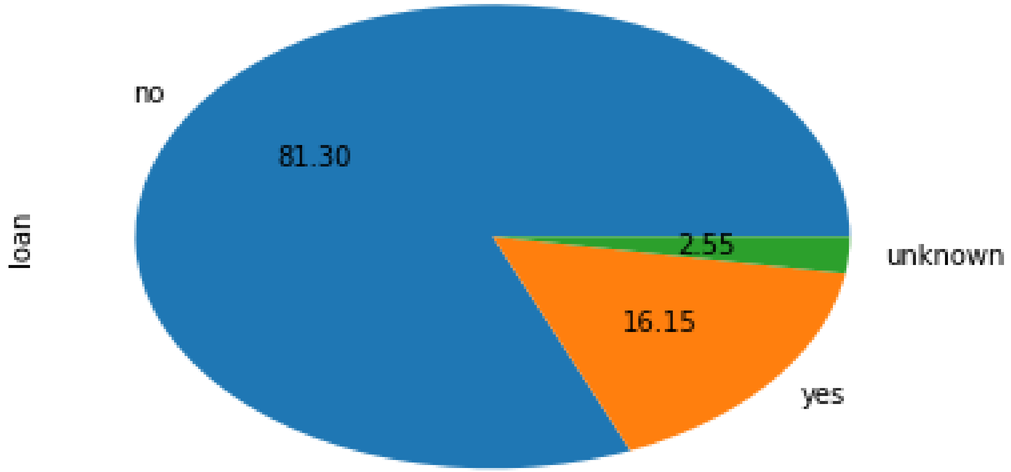

job
blue-collar
admin.
21.47
24.56
0.85
1.99
2.67
2.67
3.59
3.86
4.03
7.87
9.55
16.78
unknown
student
unemployed
housemaid
entrepreneur
self-employed
retired
management
services
technician

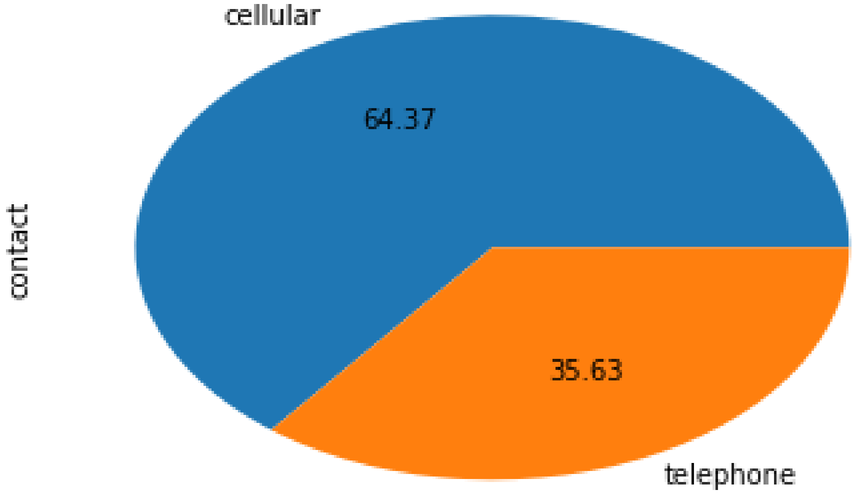

married
60.89
marital
0.27
unknown
10.83
divorced
28.01
single

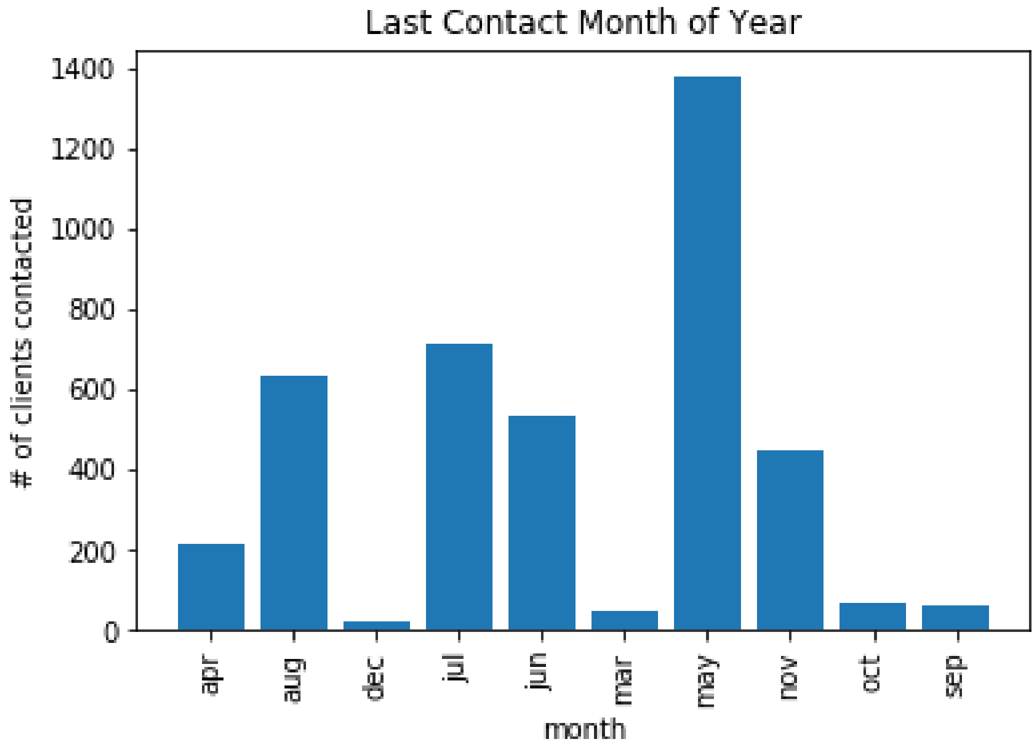

Client Education Information
# of Clients
0
200
400
600
800
1000
1200
basic.4y
basic.6y
basic.6yes
basic.9y
high.school
illiterate
professional.course
university.degree
unknown
level

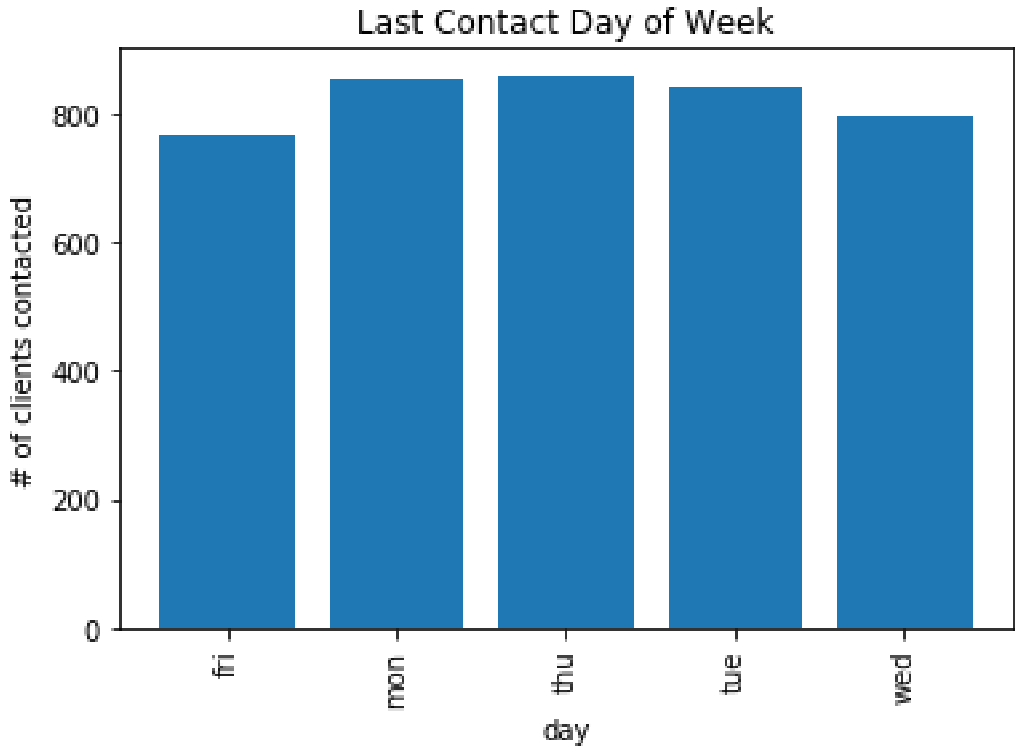

default
no
80.47
0.02
yes
19.50
unknown

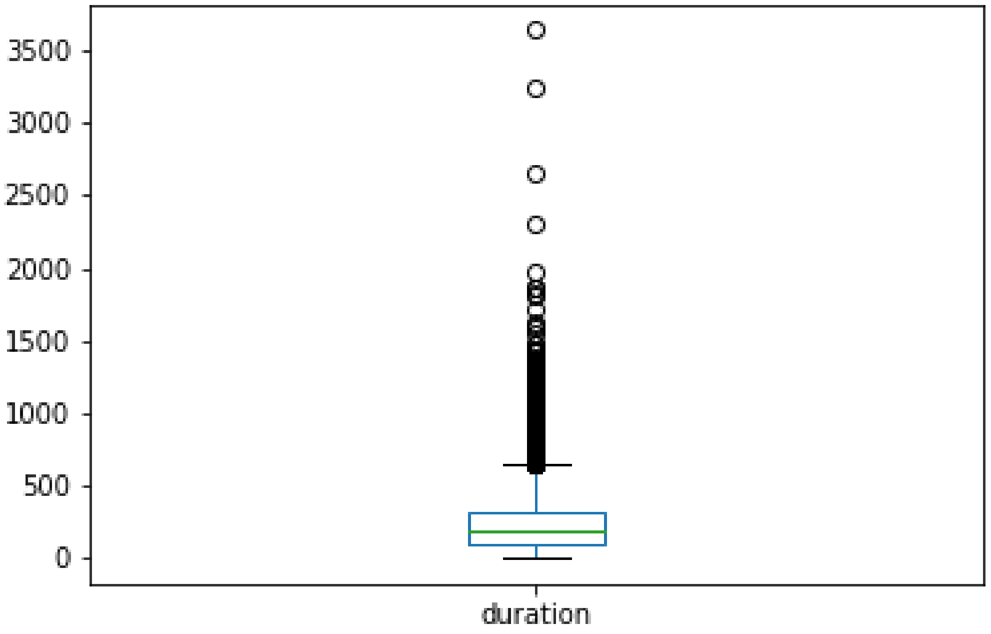

yes
52.83
housing
2.55
unknown
44.62
no

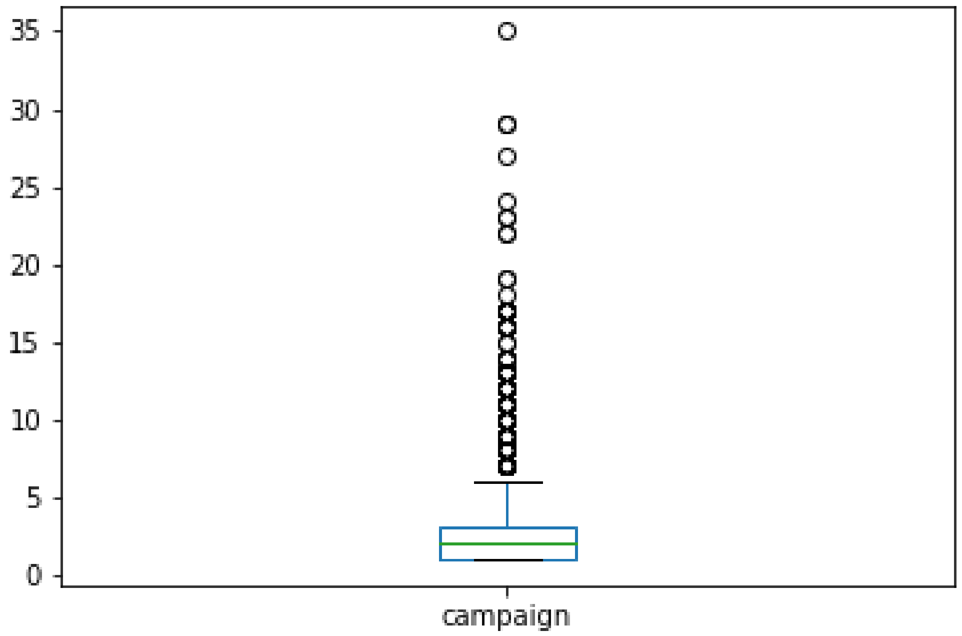

loan
no
81.30
2.55
unknown
16.15
yes

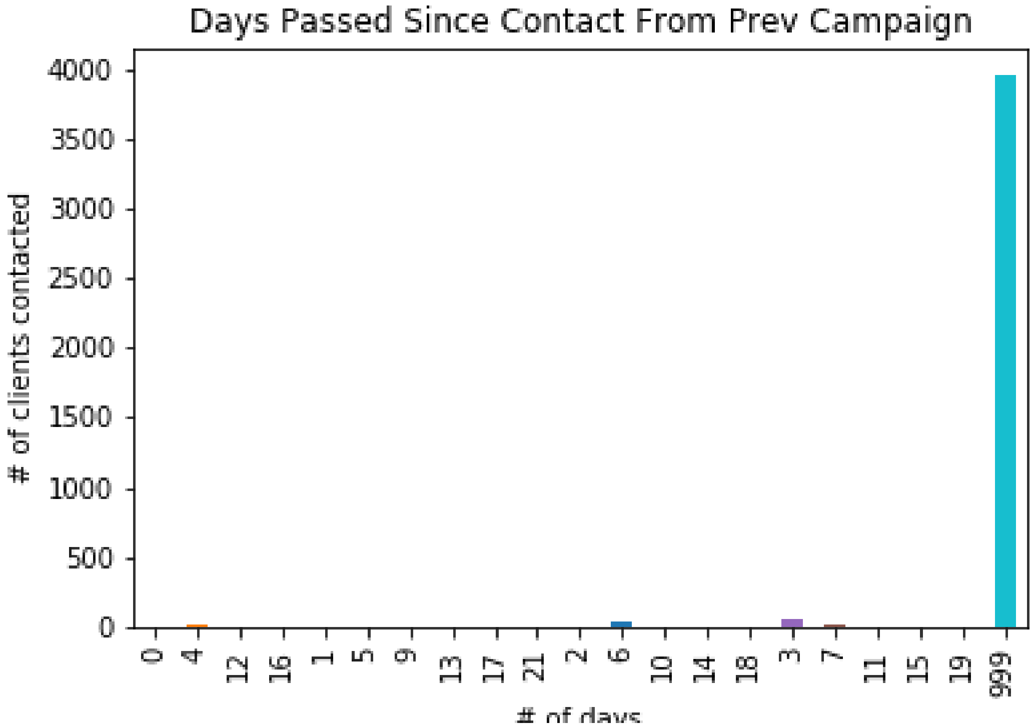

contact
cellular
64.37
35.63
telephone


Last Contact Month of Year
# of clients contacted
1400
1200
1000
800
600
400
200
0
apr
aug
dec
jul
jun
mar
may
nov
oct
sep
month

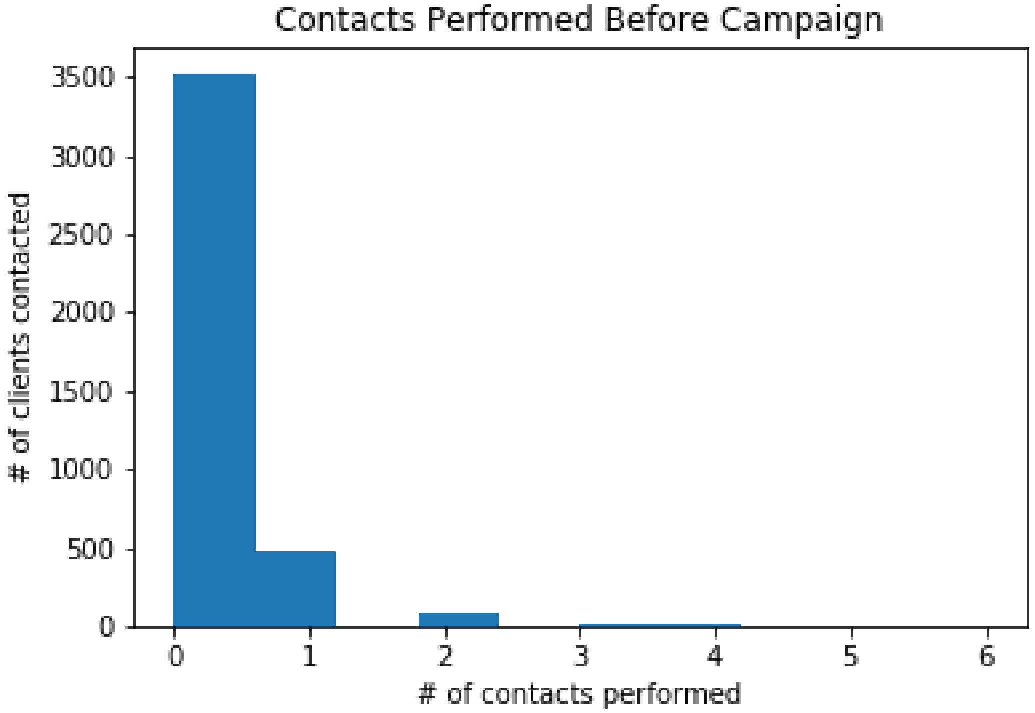

Last Contact Day of Week
# of clients contacted
800
600
400
200
0
fri
mon
thu
tue
wed
day


3500
3000
2500
2000
1500
1000
500
0
duration

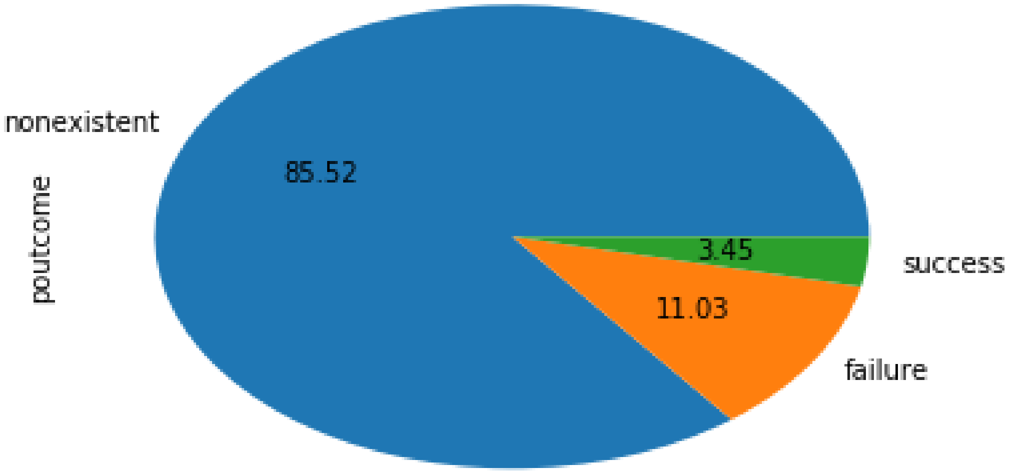

35
30
25
20
15
10
5
0
campaign

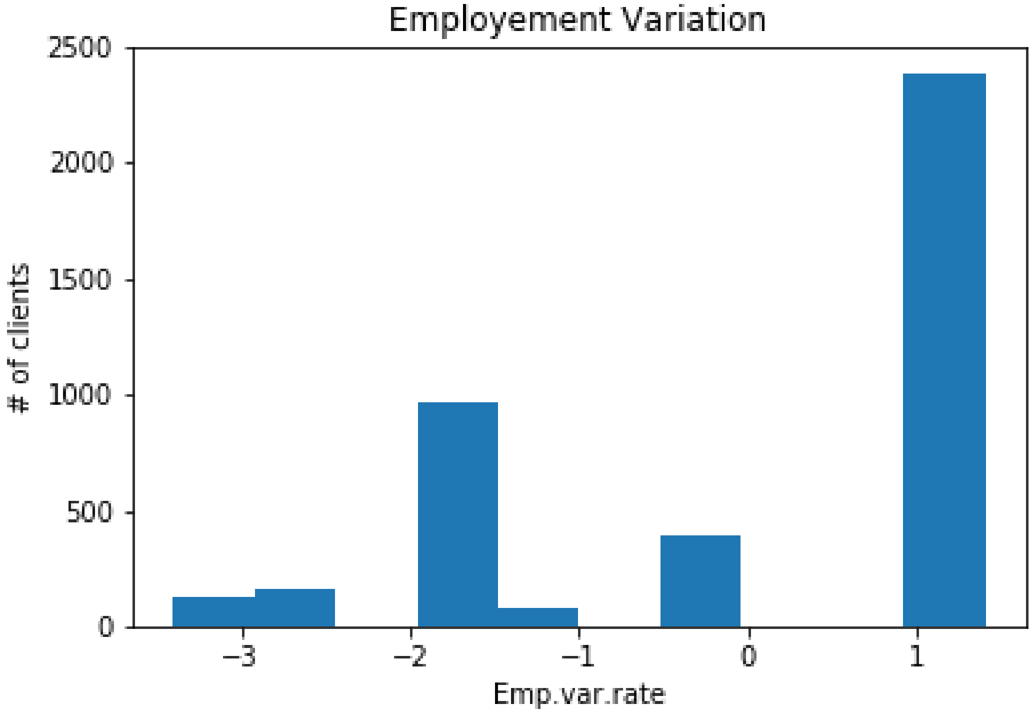


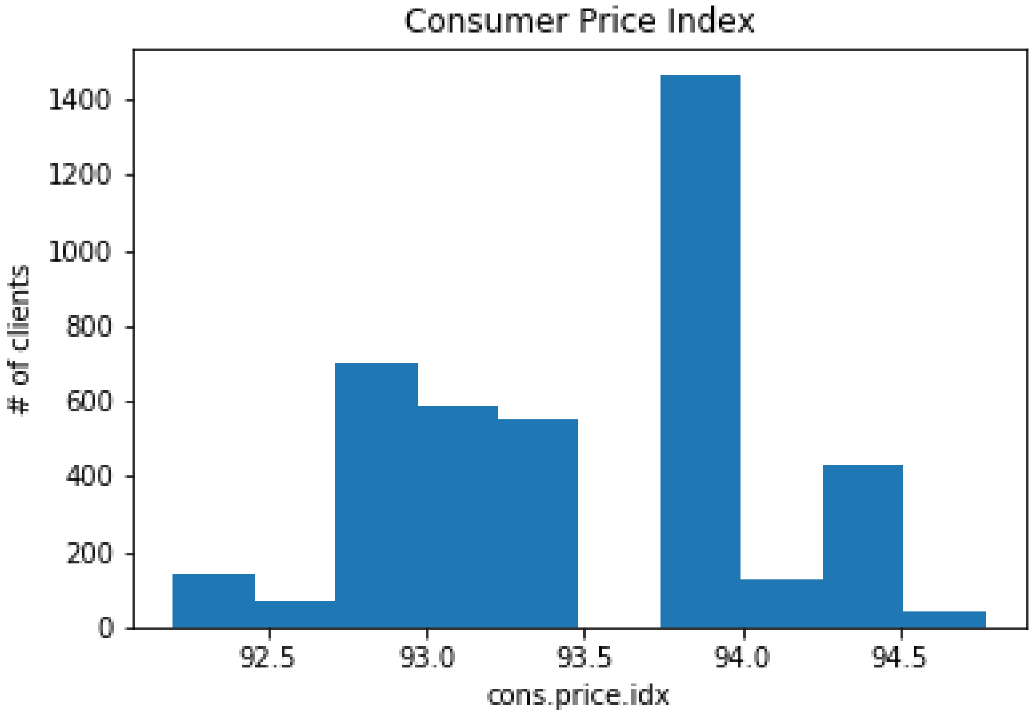
Days Passed Since Contact From Prev Campaign

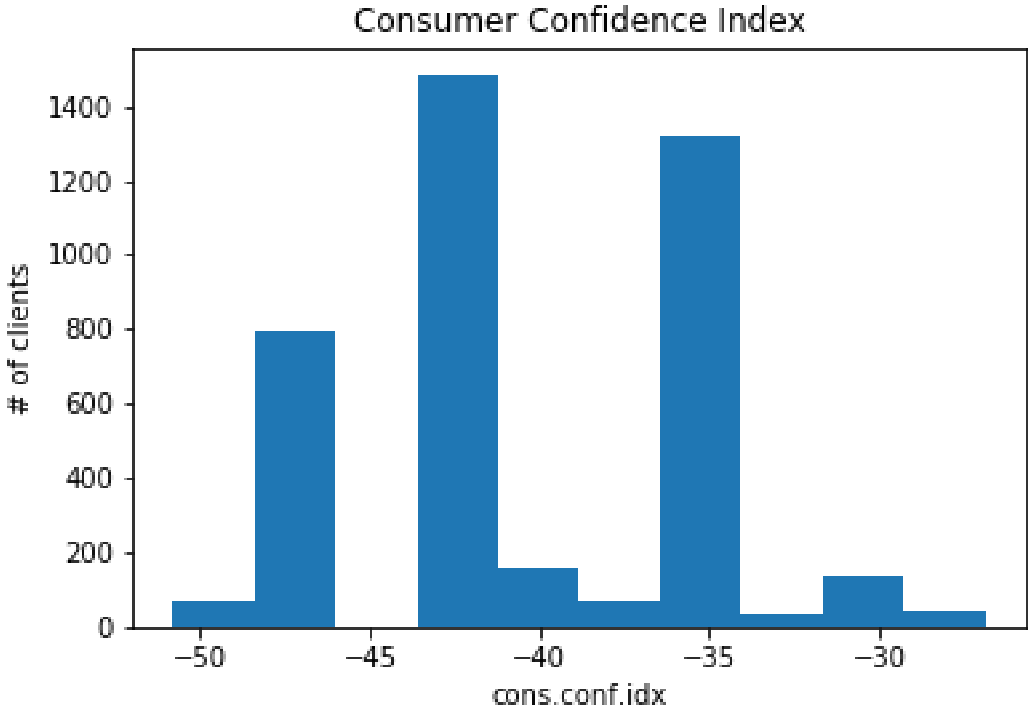


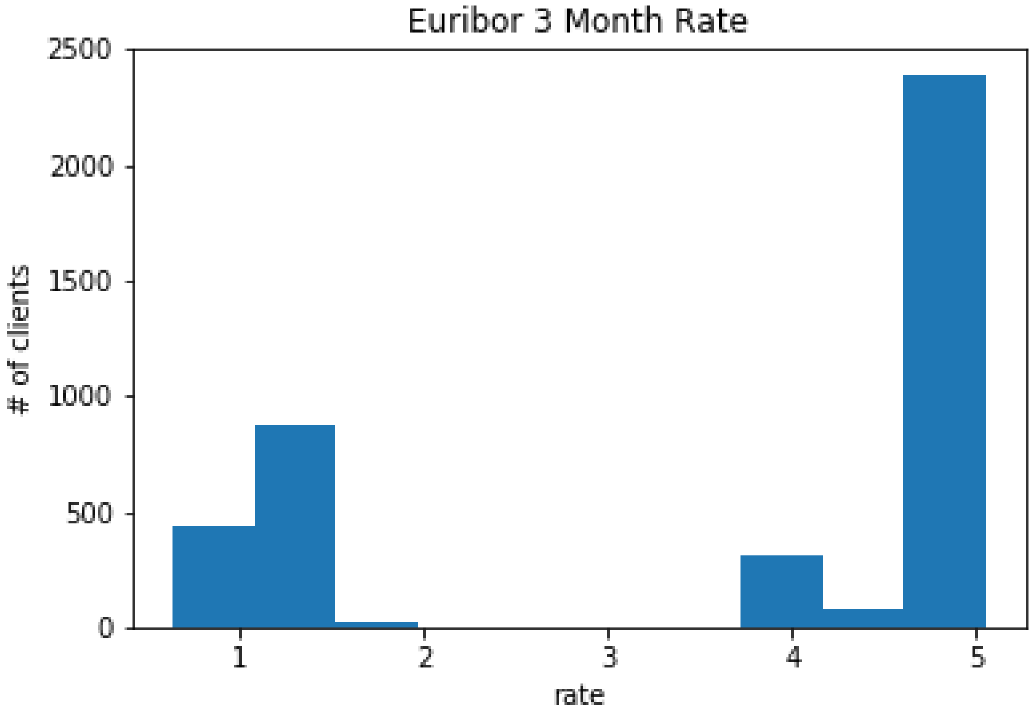
Contacts Performed Before Campaign


nonexistent
85.52
poutcome
3.45
success
11.03
failure

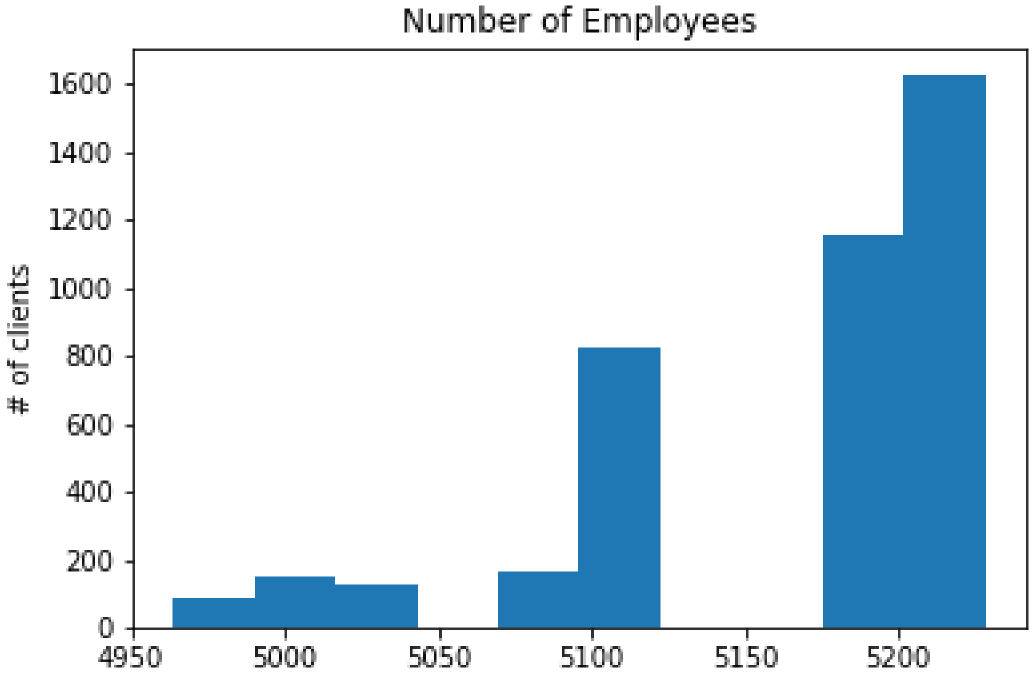

Employement Variation
2500
2000
1500
1000
500
0
# of clients
−3
−2
−1
0
1
Emp.var.rate


Consumer Price Index
# of clients
0
200
400
600
800
1000
1200
1400
92.5
93.0
93.5
94.0
94.5
cons.price.idx

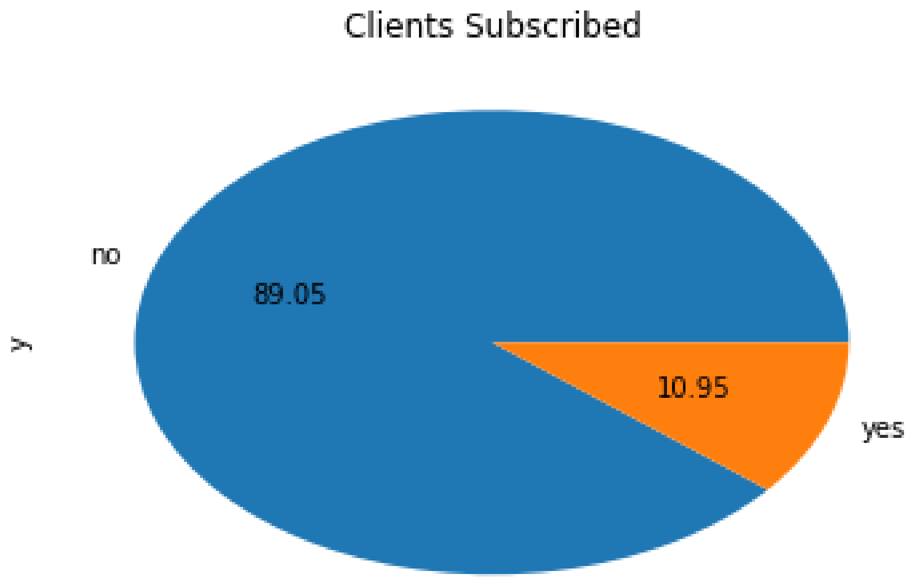

Consumer Confidence Index
# of clients
0
200
400
600
800
1000
1200
1400
−50
−45
−40
−35
−30
cons.conf.idx



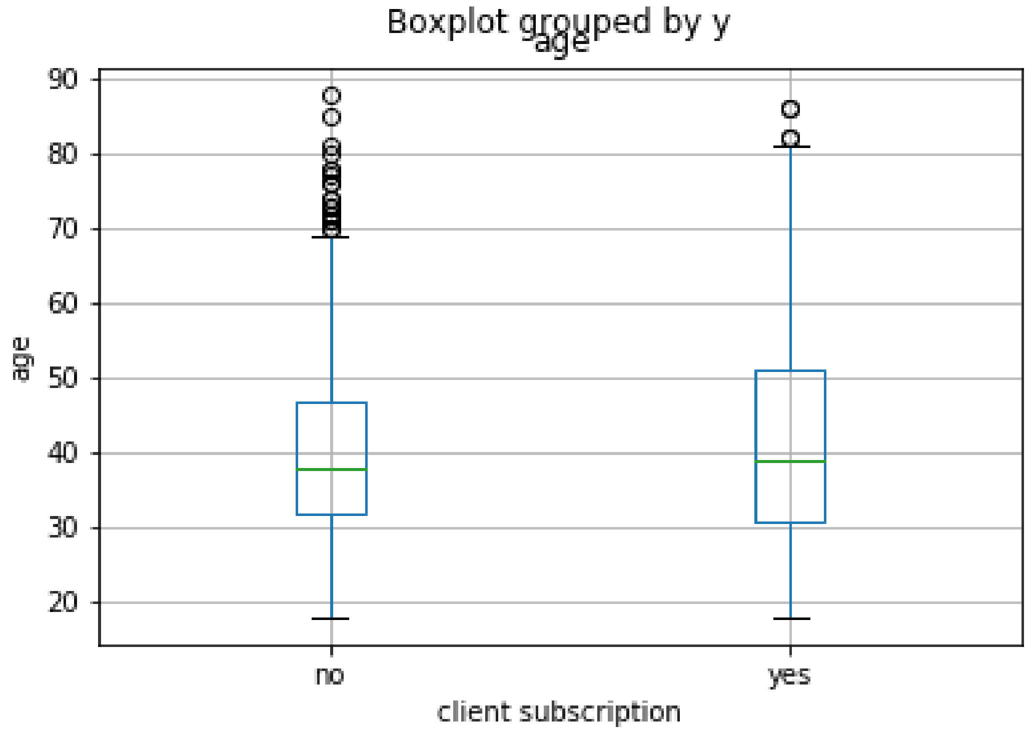
Euribor 3 Month Rate



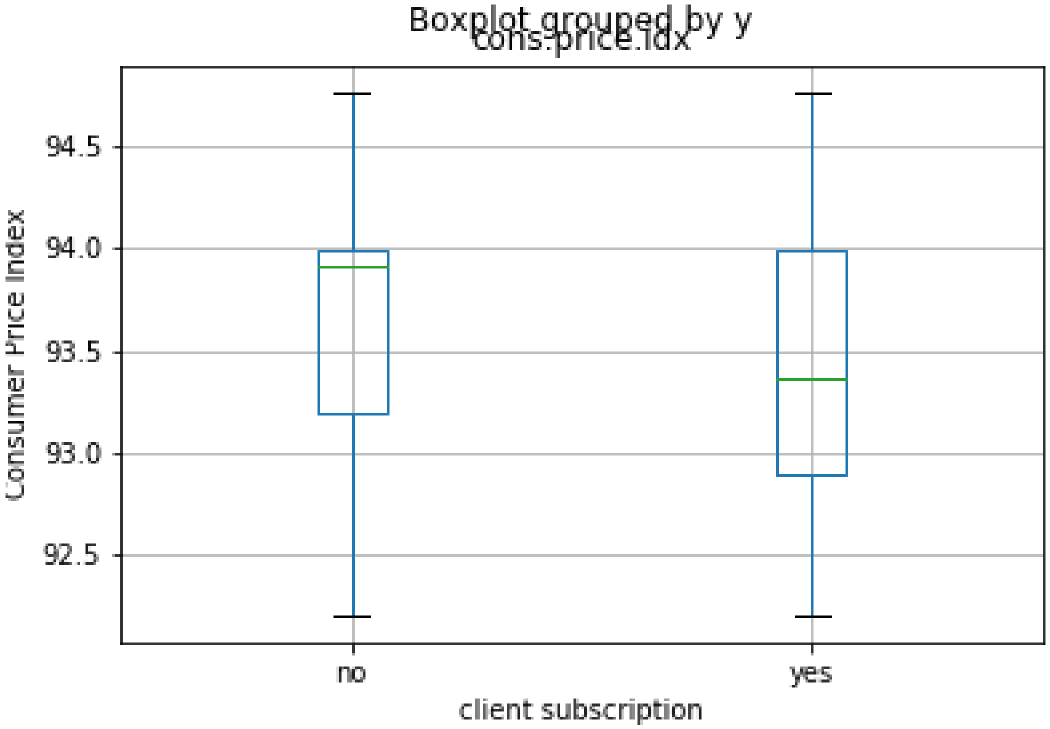
Number of Employees



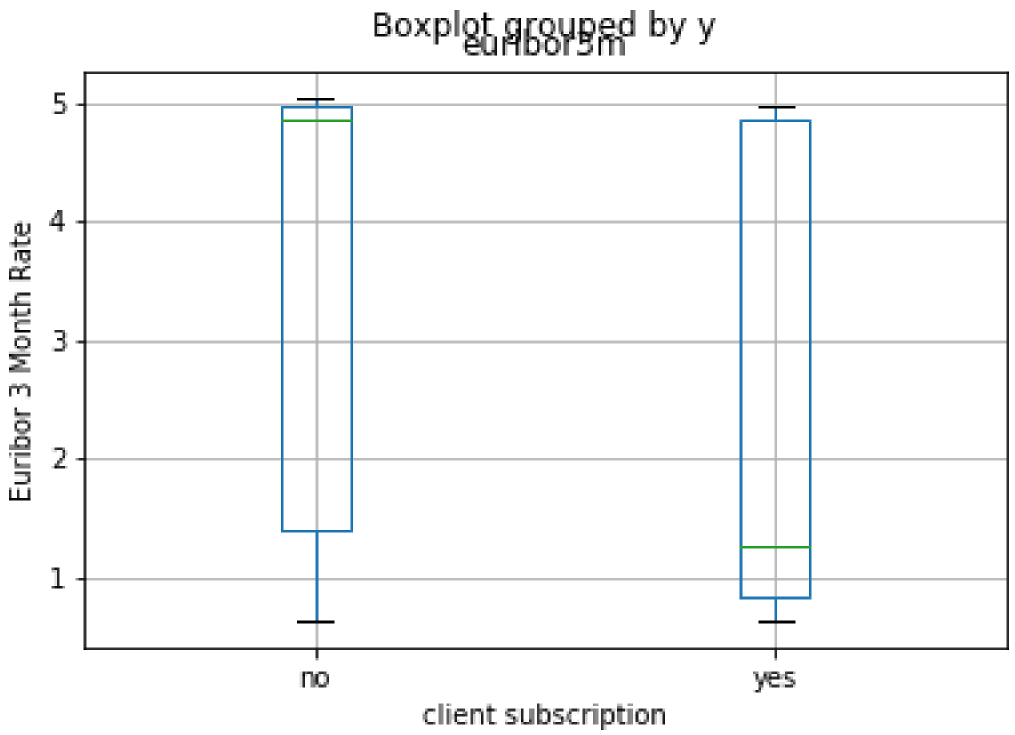
Subsection 2



Hypothesis: Older clients tend to subscribe rather than say no.
Why?: The boxplot for clients who did not subscribe has the upper whisker at age 70. This means that 75% of people who said no are under the age of 70. Also notice the upper whisker on boxplot for yes. It is at age 80. It seems that there is not relationship between middle age people and subscription shown in this graph because the median is around the same place. Also the IQR is roughly the same.

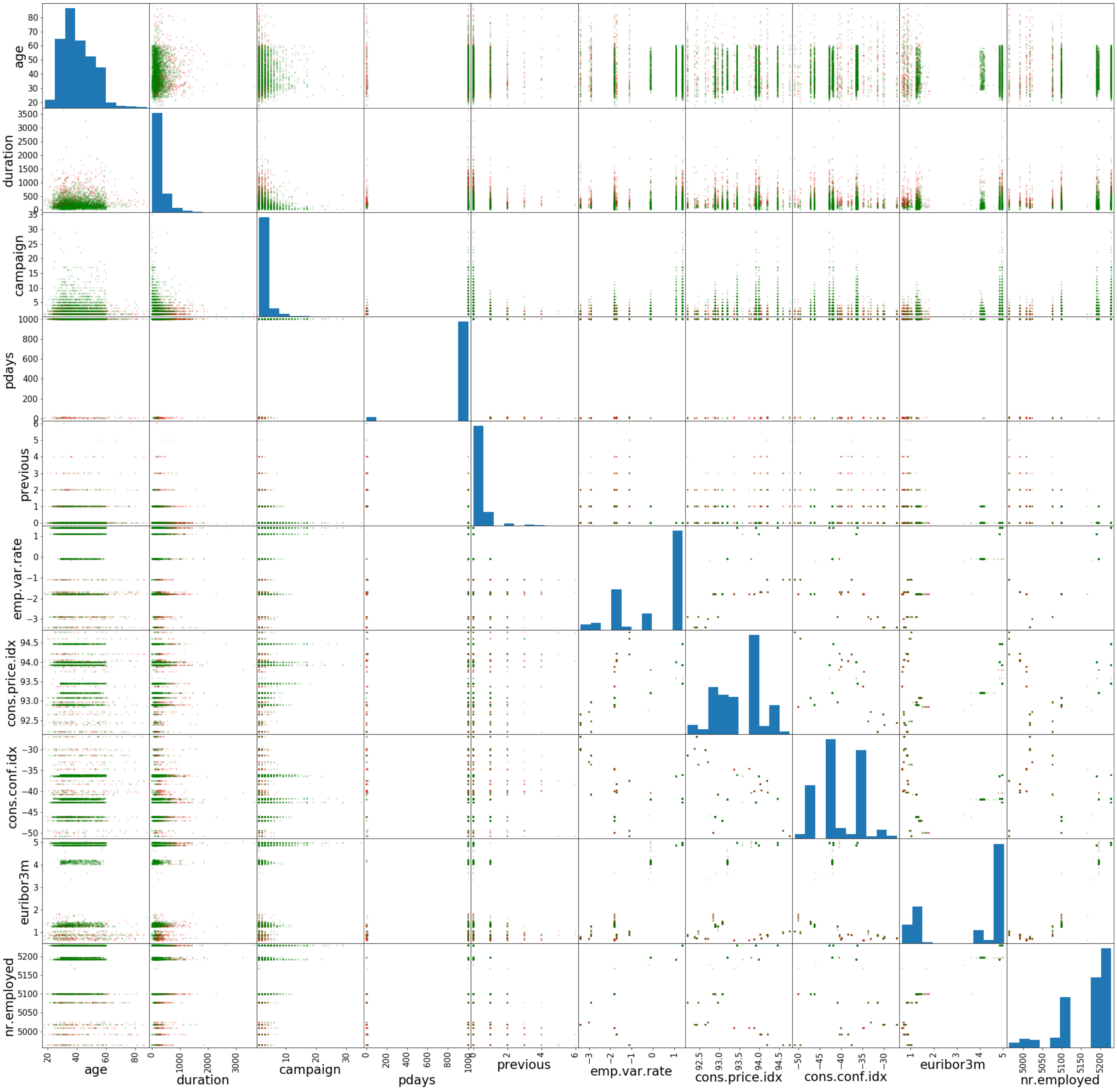


Hypothesis: clients with high consumer price index will not subscribe.
Why?: Notice the median for clients who said yes is much lower than other boxplot. Everything else is pretty much the same. We can see both graphs have similar distributions but the clients who said no are more densely populated around the values 93.2 to 94.0. Therefore it is possible that high consumer price index leads to clients saying no.



Hypothesis: clients with high Euribor 3 month rate will not subscribe.
Why?: Notice the median for clients who said yes is much lower than other boxplot. Everything else is pretty much the same. We can see that clients who said no have a median of 4.8 which is near the max value. Thus it is possible that clients with a high rate tend to decline subscription.

Subsection 3
Scatter Matrix


age
duration
campaign
pdays
previous
emp.var.rate
cons.price.idx
cons.conf.idx
euribor3m
nr.employed