VIRTUAL INTERNSHIP

DATA SCIENCE

LISUM01

Data Cleansing and Transformation

Name: Samuel Alejandro Cueva Lozano
Email: scueval07@gmail.com
Country: Perú

30/07/2021

Sumario

Business Background

Client

- ABC bank: Portuguese banking institution

Problem Description

- ABC Bank wants to sell it's term deposit product to customers and before launching the product they want to know whether a particular customer will buy their product or not (based on customer's past interaction with bank or other Financial Institution).

Missing values

Attributes	Number of records	Number of missing values	% of missing values
age	41188	0	0.0
job	41188	0	0.0
marital	41188	0	0.0
education	41188	0	0.0
default	41188	0	0.0
housing	41188	0	0.0
loan	41188	0	0.0
contact	41188	0	0.0
month	41188	0	0.0
day_of_week	41188	0	0.0
duration	41188	0	0.0
campaign	41188	0	0.0
pdays	41188	0	0.0
previous	41188	0	0.0
poutcome	41188	0	0.0
emp.var.rate	41188	0	0.0
cons.price.idx	41188	0	0.0
cons.conf.idx	41188	0	0.0
euribor3m	41188	0	0.0
nr.employed	41188	0	0.0
y	41188	0	0.0

No missing values found in any attribute

Duplicate values

Using Pandas, the duplicate rows are as follows:

	age	job	marital	education	default	housing
1266	39	blue-collar	married	basic.6y	no	no
12261	36	retired	married	unknown	no	no
14234	27	technician	single	professional.course	no	no
16956	47	technician	divorced	high.school	no	yes
18465	32	technician	single	professional.course	no	yes
20216	55	services	married	high.school	unknown	no
20534	41	technician	married	professional.course	no	yes
25217	39	admin.	married	university.degree	no	no
28477	24	services	single	high.school	no	yes
32516	35	admin.	married	university.degree	no	yes
36951	45	admin.	married	university.degree	no	no
38281	71	retired	single	university.degree	no	no

Note: only some attributes are displayed due to the table size.

This rows will be removed

Outliers

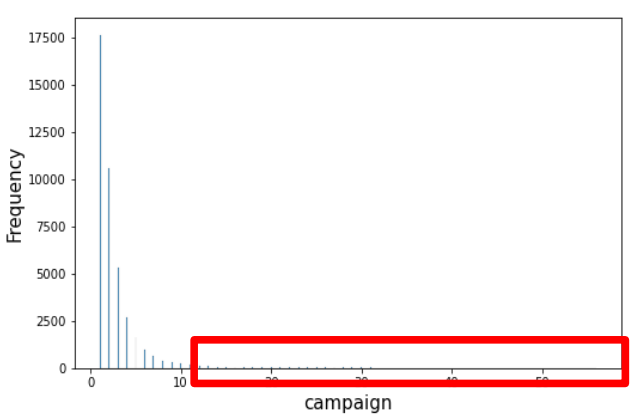
Approaches to dealing with outliers

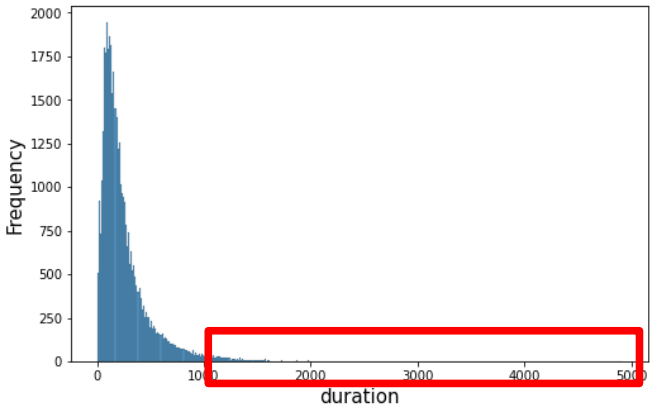
Visualizations and descriptive statistics to detect potential outliers

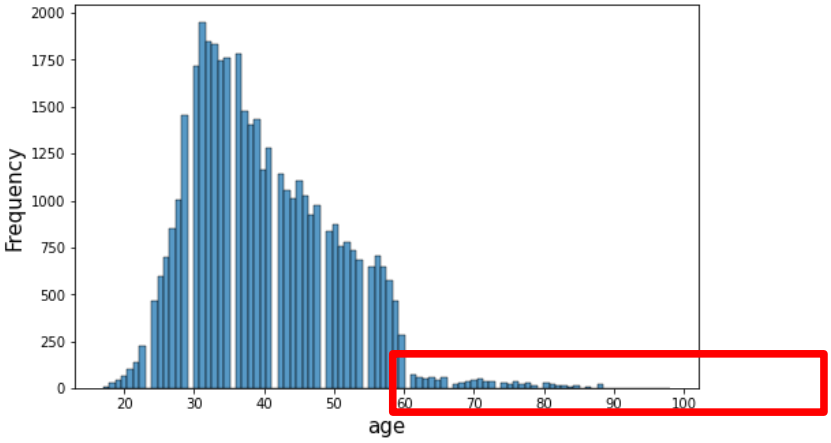
	age	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
count	41176.00000	41176.000000	41176.000000	41176.000000	41176.000000	41176.000000	41176.000000	41176.000000	41176.000000	41176.000000
mean	40.02380	258.315815	2.567879	962.464810	0.173013	0.081922	93.575720	-40.502863	3.621293	5167.034870
std	10.42068	259.305321	2.770318	186.937102	0.494964	1.570883	0.578839	4.627860	1.734437	72.251364
min	17.00000	0.000000	1.000000	0.000000	0.000000	-3.400000	92.201000	-50.800000	0.634000	4963.600000
25%	32.00000	102.000000	1.000000	999.000000	0.000000	-1.800000	93.075000	-42.700000	1.344000	5099.100000
50%	38.00000	180.000000	2.000000	999.000000	0.000000	1.100000	93.749000	-41.800000	4.857000	5191.000000
75%	47.00000	319.000000	3.000000	999.000000	0.000000	1.400000	93.994000	-36.400000	4.961000	5228.100000
max	98.00000	4918.000000	56.000000	999.000000	7.000000	1.400000	94.767000	-26.900000	5.045000	5228.100000

Potential outliers in Age, *Duration* and *Campaign*

Histograms for Age, Campaign and Duration







Filtering by fixed threshold

- There is not any attribute in which to apply this approach is reasonable.
- The age attribute has a maximum value of 98 and this value is correct.

Clipping the attribute at a computed percentile

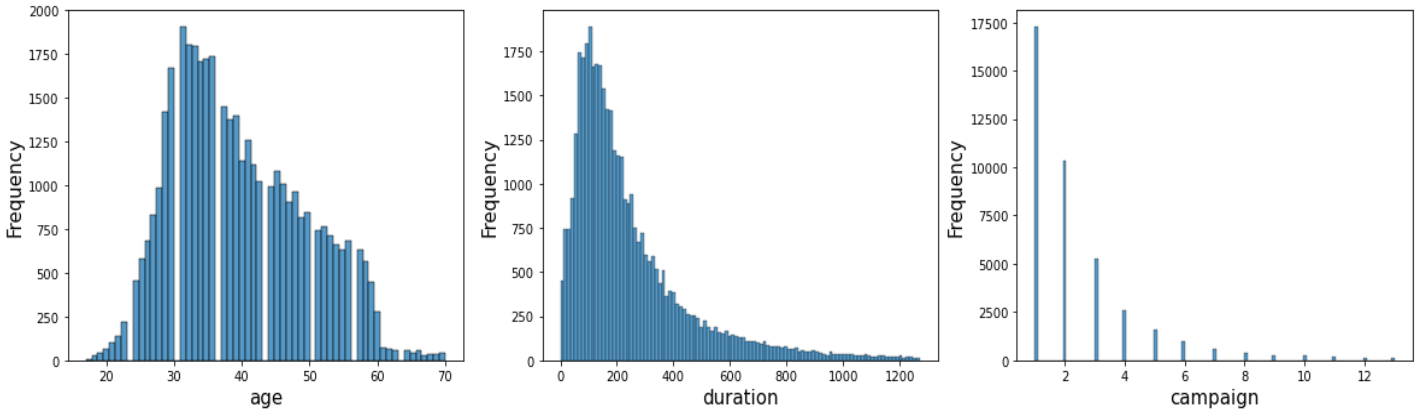
The data will be clipped in 0.99 percentile for *age, duration*, and *campaign* attributes

`

Percentile 99%		
age	duration	campaign
71	1271.25	14

Total values less than percetile 99%		
age	duration	campaign
40755	40764	40701

Histograms after Clipping



Better, but still some large outlier values in *duration* and *campaign*

Log of every value

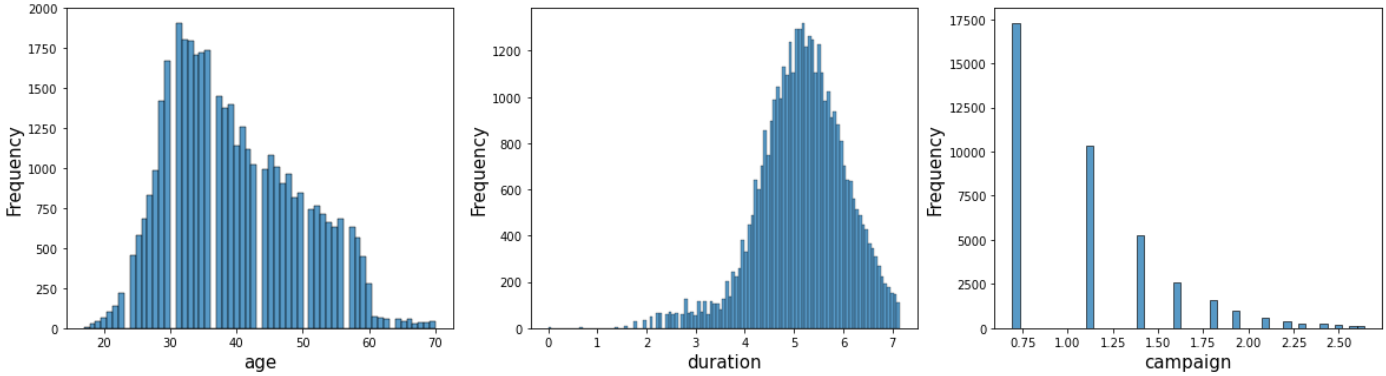
Duration after log scaling

count	mean	Std	Min	25%	50%	75%	max
39872.00000	5.161927	0.881968	0.00	4.644391	5.192957	5.752573	7.148346

Campaign after log scaling

count	mean	Std	Min	25%	50%	75%	max
39872.00000	1.097673	0.449664	0.693147	0.693147	1.098612	1.386294	2.639057

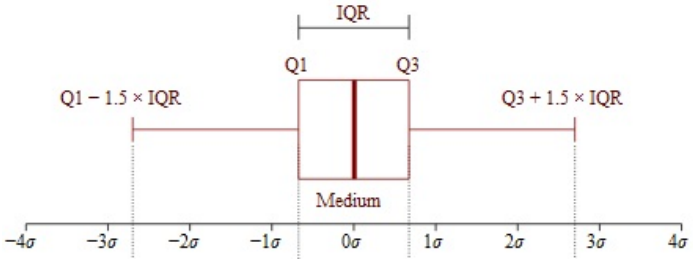
Histograms

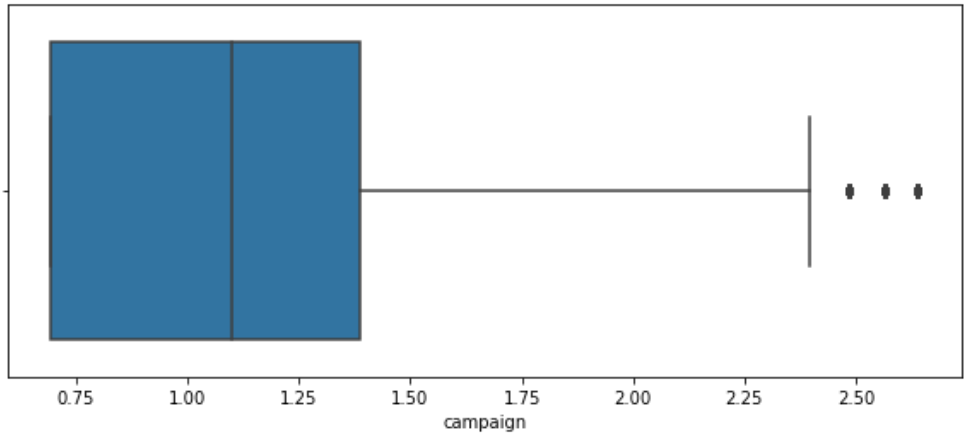


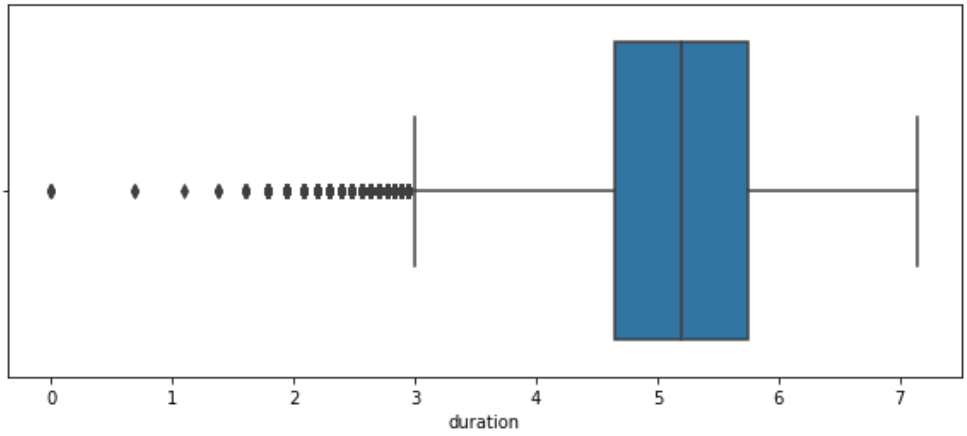
Now, *duration* looks more normal and *campaign* has less tail

IQR Score

The IQR can be used to identify outliers by defining limits on the sample values that are a factor k of the IQR below the 25th percentile or above the 75th percentile. The common value for the factor k is the value 1.5. A factor k of 3 or more can be used to identify values that are extreme outliers or “far outs” when described in the context of box and whisker plots.

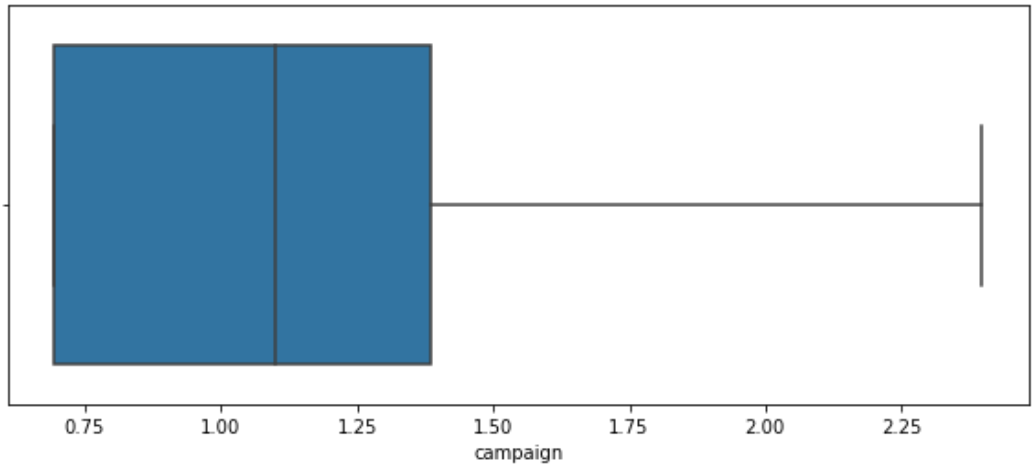


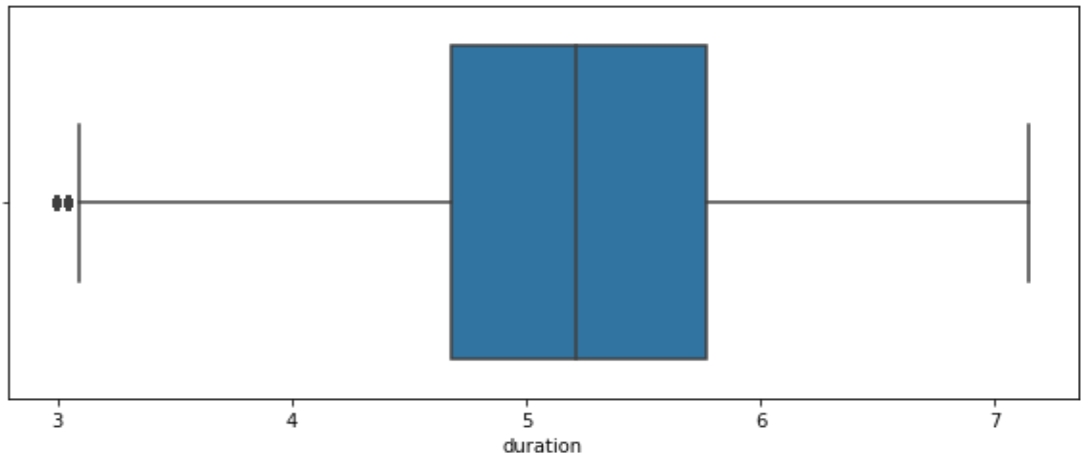


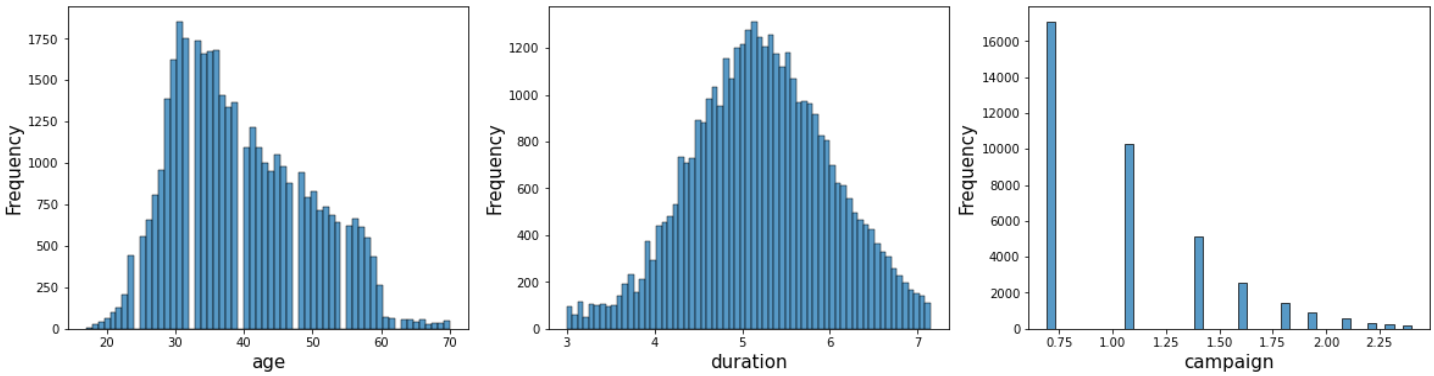


As you can see from the box-and-whisker plots, there are still outliers in the attributes

Box-and-whisker plots after filtering outliers using IQR-Score







Now, our data is more useful than the original data