Week 8: Deliverables

Individual Project:

I have done this project individually

Name: Nkululeko Freedom Mqadi

Email:mqadinf@gmail.com

Country:South Africa

College/Company:Deviare

Specialization:Data Science

• Problem description

Before we do any Exploratory Data Analysis (EDA), model building and model selection, it is very crucial to understand the data. Having some data understanding assists us to understand the data types available in a dataset, dimensionality of the data, to detect outliers, finding null/missing values and many more.

Using the "**bank-additional-full.csv"** dataset, I will provide some description of the available features available in a dataset, outline the type of data available for analysis and most importantly reveal some problems available in the data and how to tackle these problems.

• Data understanding

As mentioned above, the dataset consists of direct marketing campaigns data of a banking institution. The dataset was picked from UCI Machine Learning Repository which is an amazing source for publicly available datasets. There were four variants of the datasets out of which we chose **" bank-additional-full.csv"** which consists of 41188 data points with 20 independent variables out of which 10 are numeric features and 10 are categorical features. The list of features available to us are given below:

- **Data set description**

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')

Type of data available for analysis:

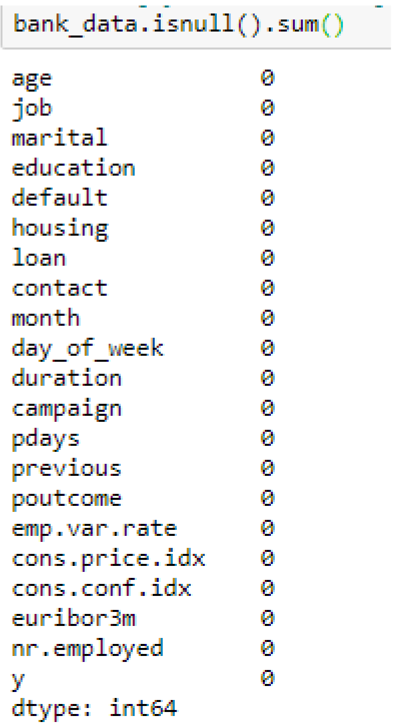
The type of data available for analysis is the **Exploratory Data Analysis(EDA).**The goal of using **Exploratory Data Analysis(EDA**) is:

- To explore data and find relationships between variables which were previously unknown.
- It’s also helps you to discover relationships between measures in your data, which are evidence for the existence of the correlation.

Problems available in the data

Null/missing values

There are none **null/missing values** in a dataset, and we can confirm this as shown below:



```
bank_data.isnull().sum()
```

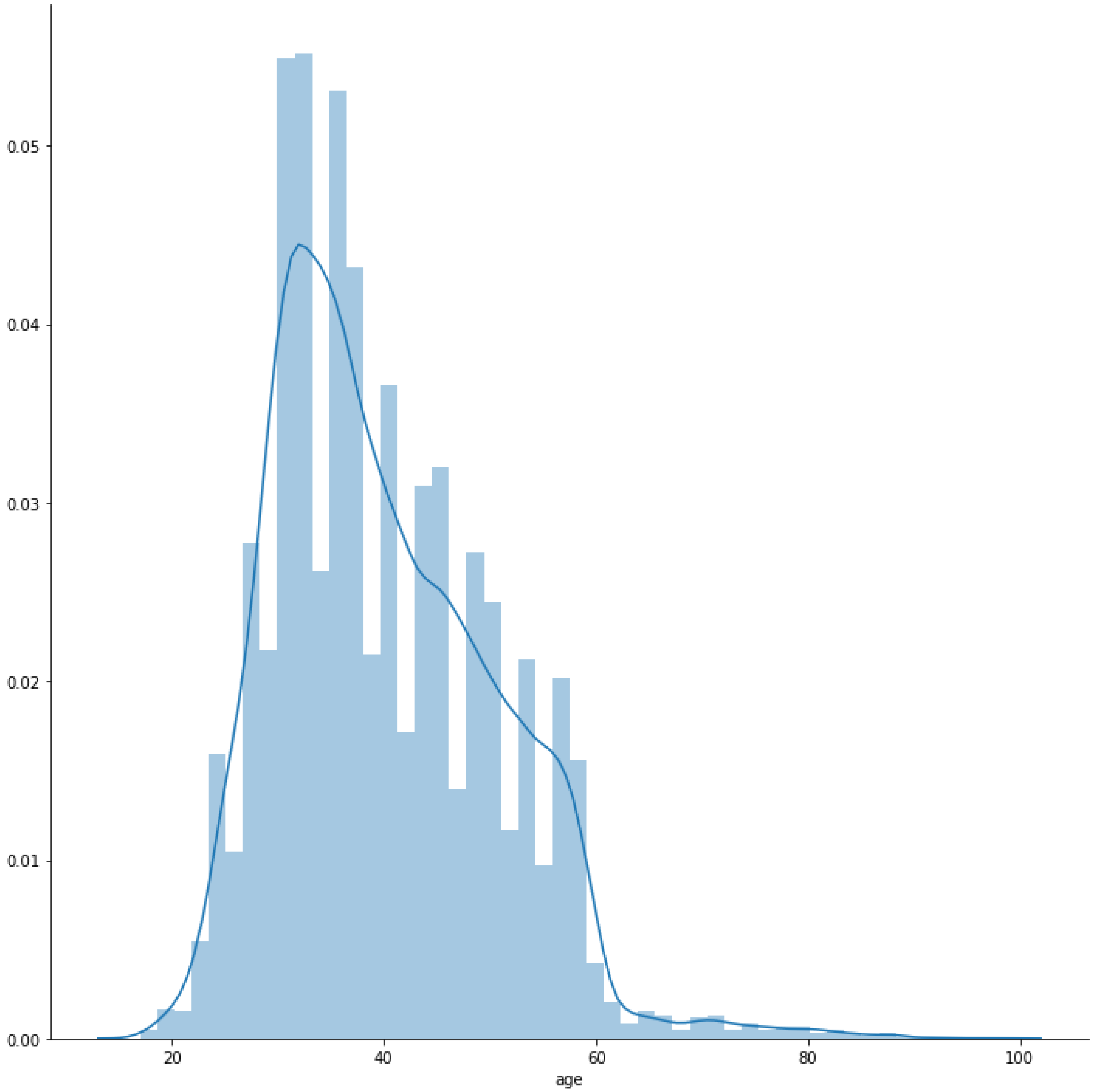
```
age               0
job               0
marital           0
education         0
default           0
housing           0
loan              0
contact           0
month             0
day_of_week       0
duration          0
campaign          0
pdays             0
previous          0
poutcome          0
emp.var.rate      0
cons.price.idx    0
cons.conf.idx     0
euribor3m         0
nr.employed       0
y                 0
dtype: int64
```

➢ Outliers

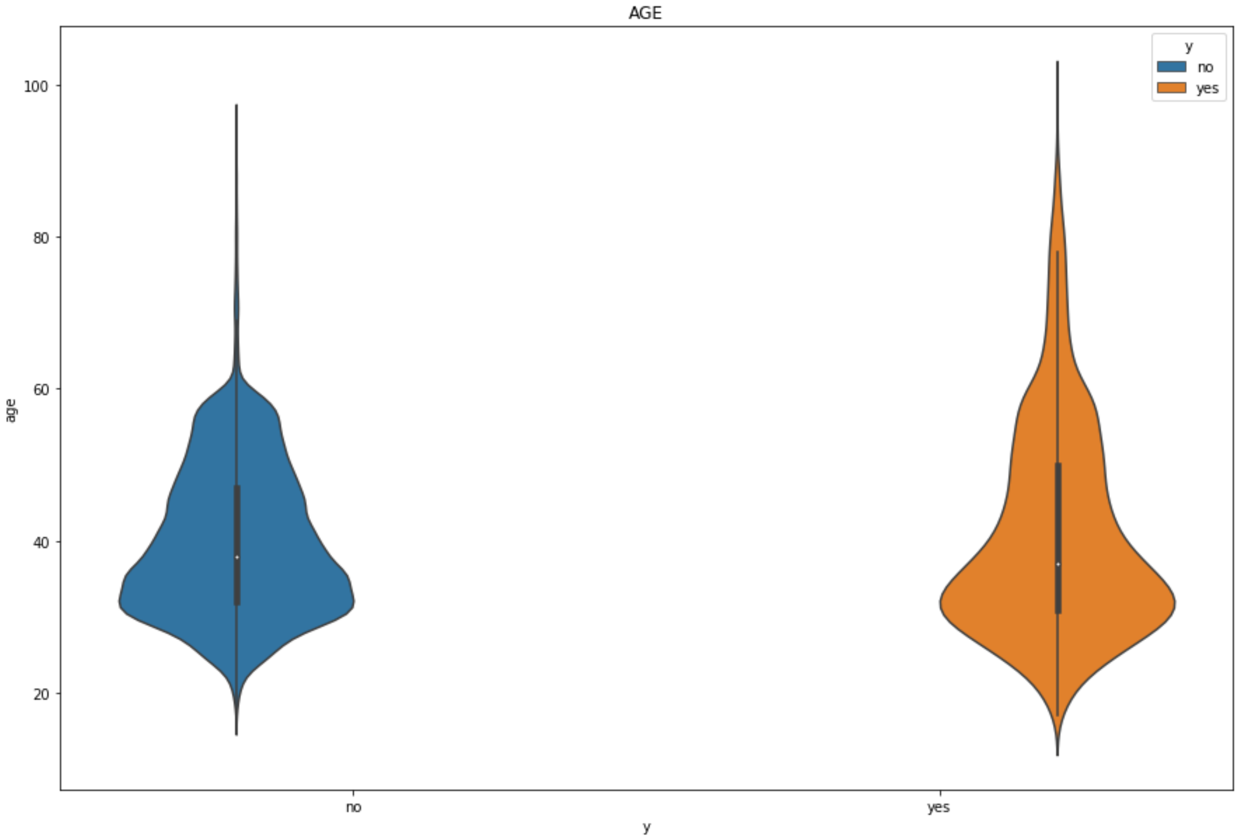
There are a variety of ways to check for outliers in a dataset. We can check for outliers using Boxplot, Kdeplot and Violin plot. To check for the outliers requires that we look for numerical values only, and however we did find some outliers on certain numerical values. There are 10 numerical columns having 5 columns of integer datatype and 5 columns on float datatype.

Detecting outliers for the variable age

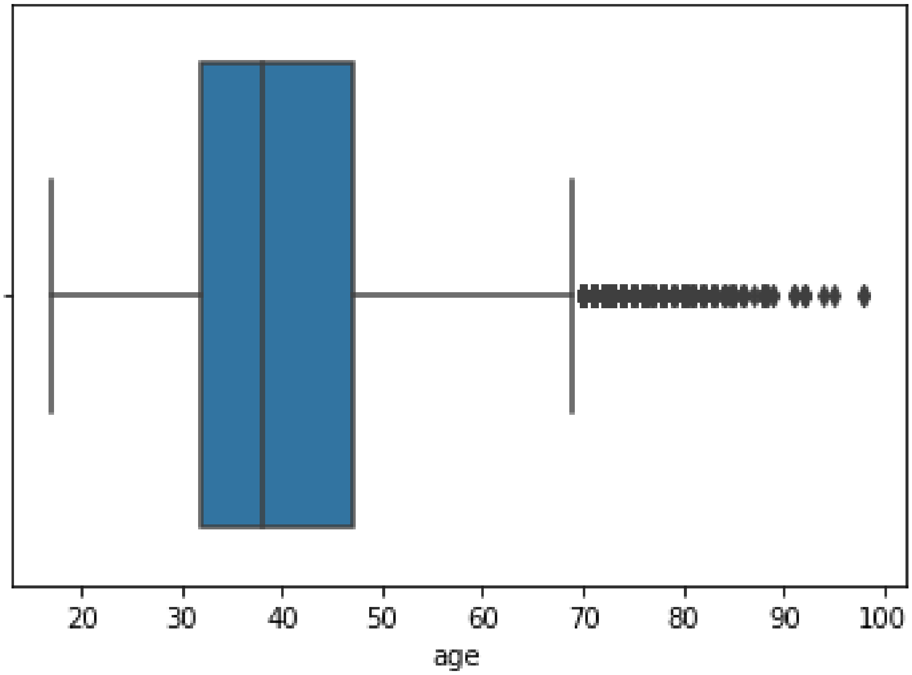
- Using the Kde plot there is an evidence of outliers after the age of 60 as shown below.



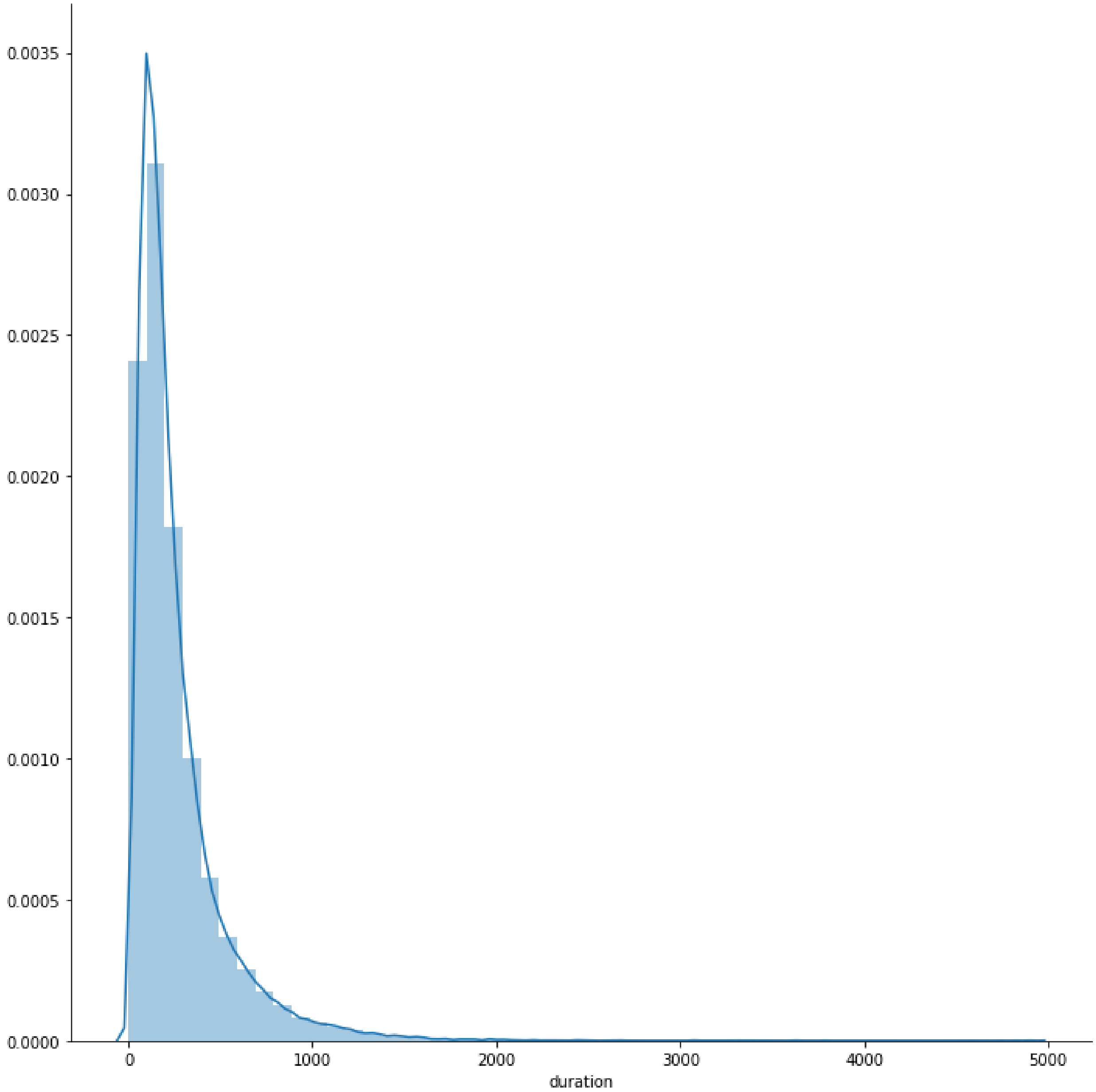
- Using the Violin plot as shown below, it is clearly visible that there are outliers present for both the class. In No class, outliers are present above age 70 and for Yes class, outliers are present above age 75.



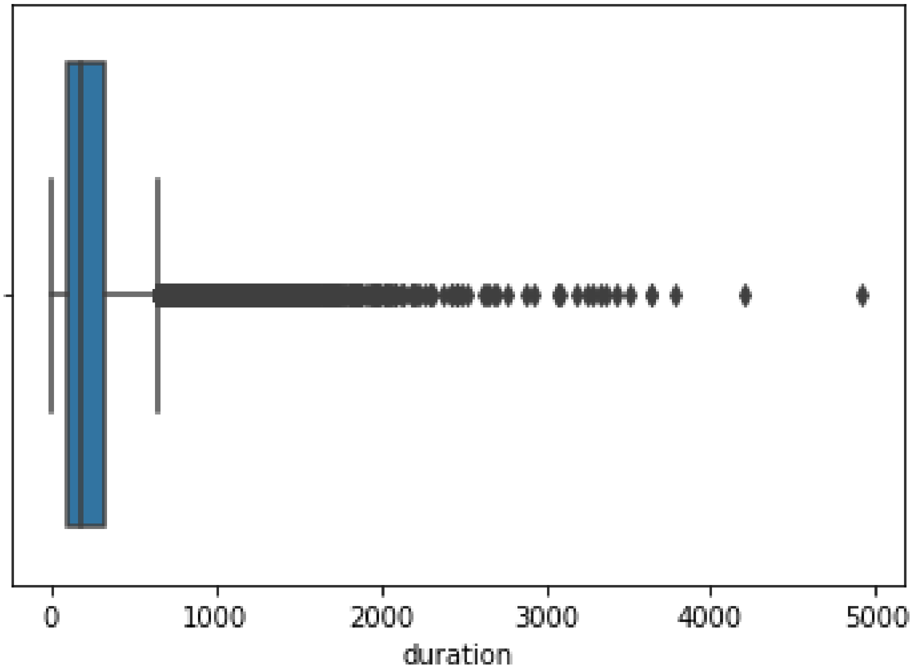
- Using the Box plot we can say that outliers are present after the age of 70 years.

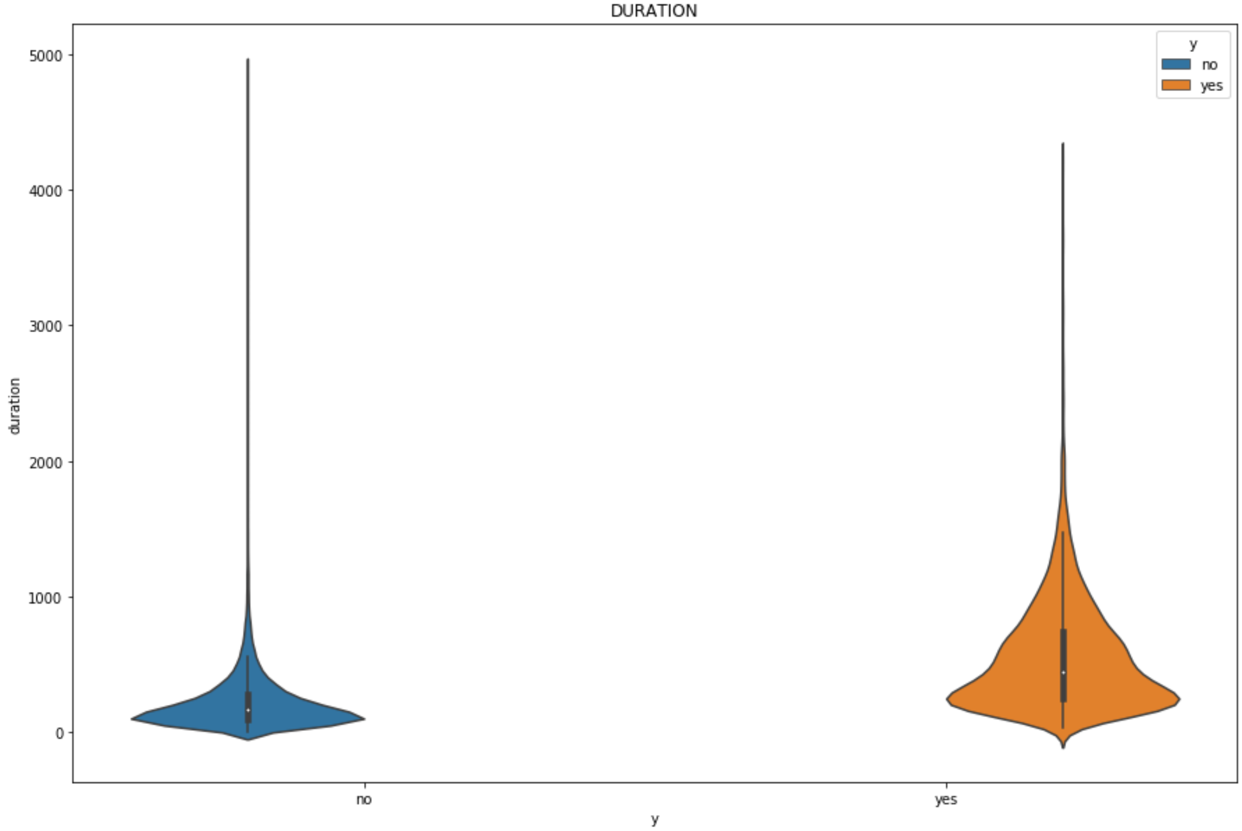


- **Detecting outliers for the variable duration**
  Using the Kde plot, it is evidence that the outliers are present where the duration of calls are more than 1000.



- Using the box plot below, we also see the presence of outliers where the duration of calls is more than 1000.

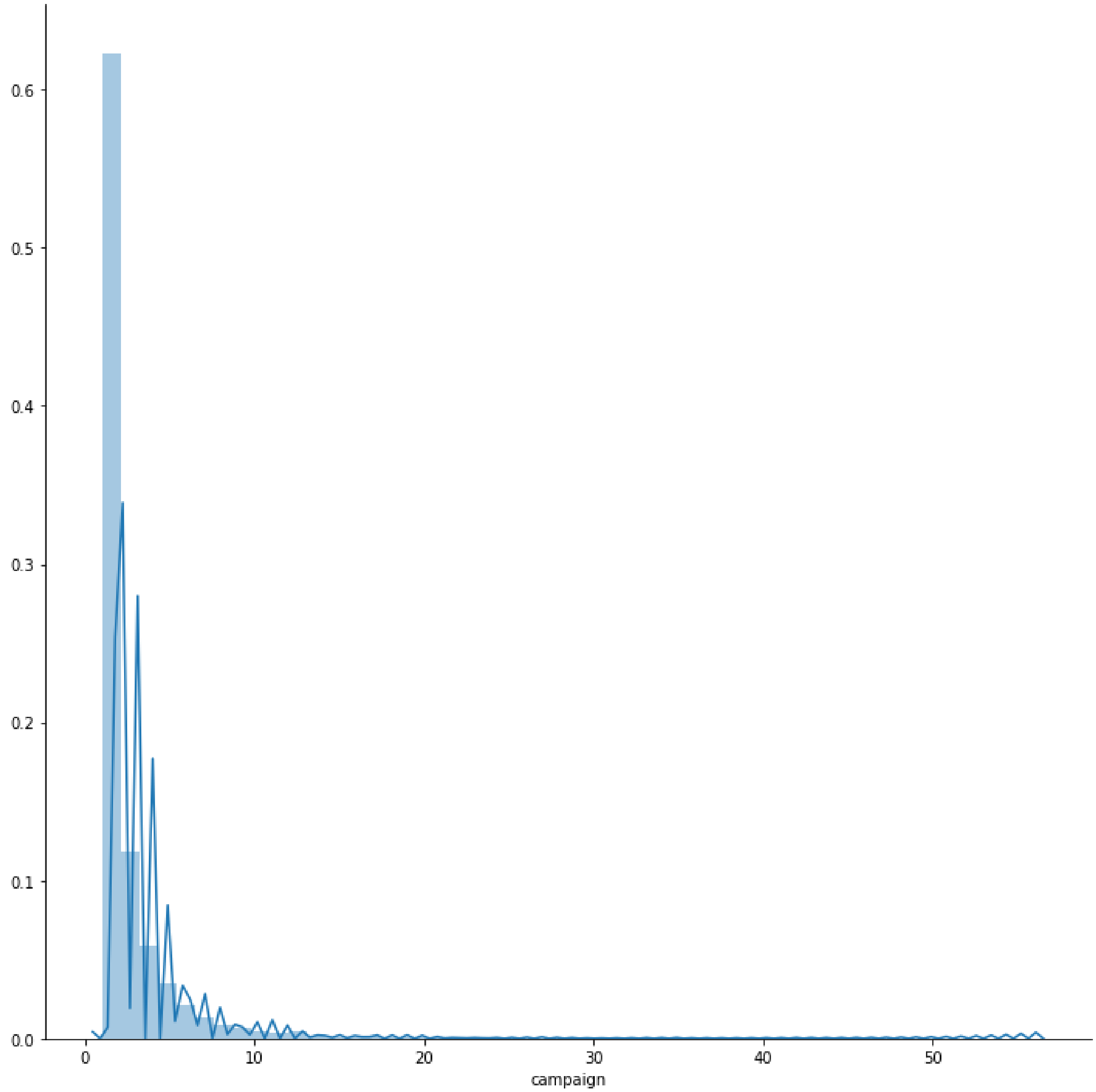

0
1000
2000
3000
4000
5000
duration


DURATION
y
no
yes
5000
4000
3000
2000
1000
0
duration
no
yes
y

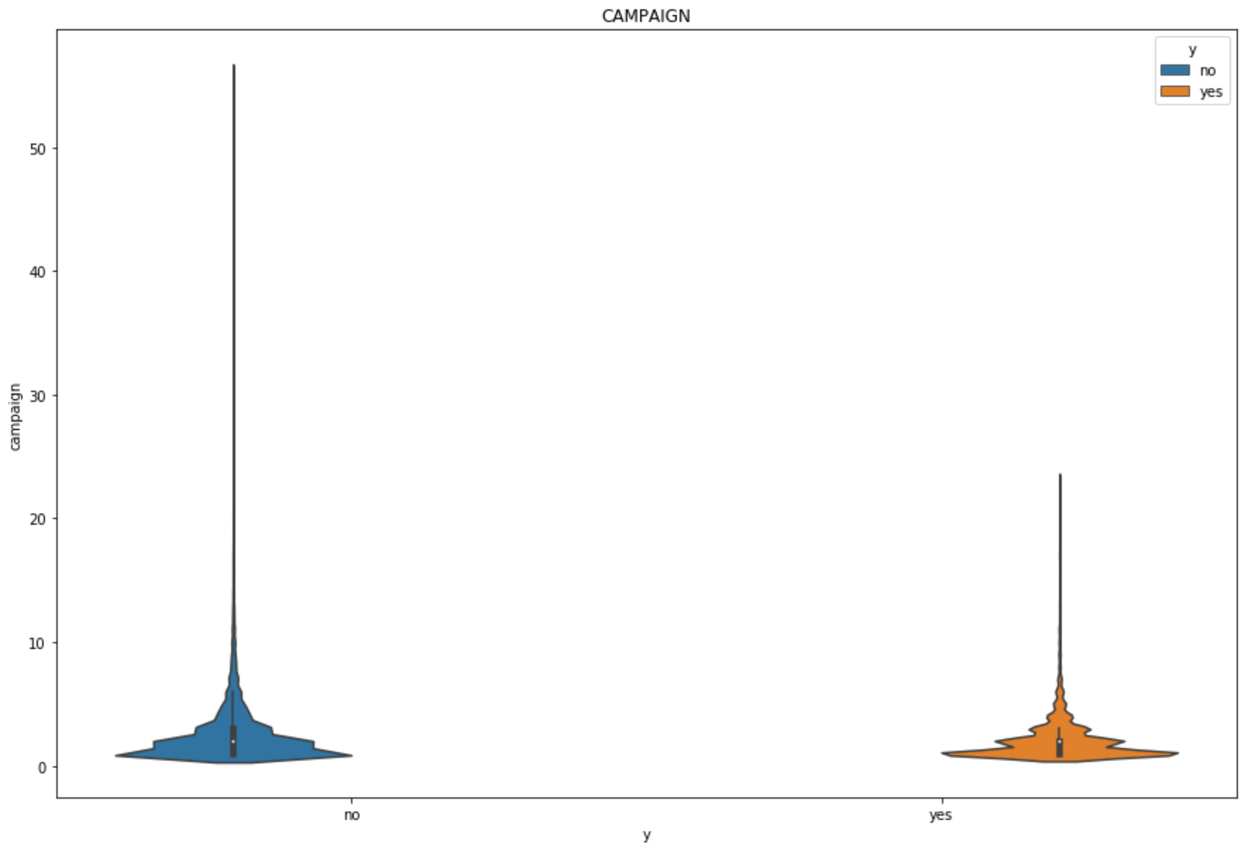
- As shown above for the Violin plot, any duration of call with class labels as no, more than 1000 are considered as outliers, while with class labels as yes, more than 1500 would be considered as outliers.

Detecting outliers for the variable campaign

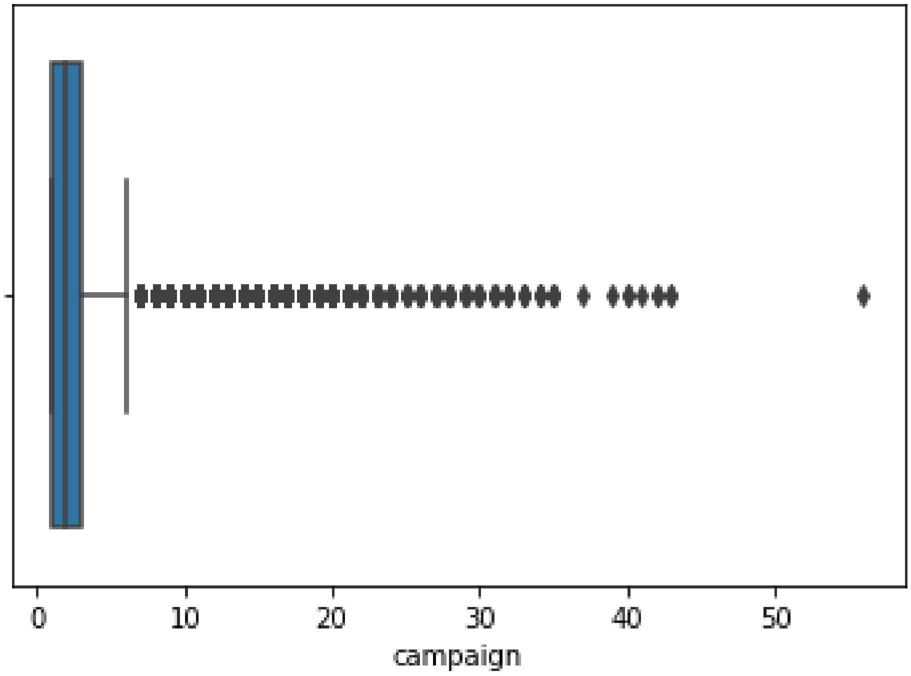
- Using the Kde plot as shown below, it is not very clear that outliers are present or not.



- Using the Violin plot as shown below, for any class labels, campaigns more than 10 are considered as outliers. From the above plot there are so many outliers are present for No class than yes class.

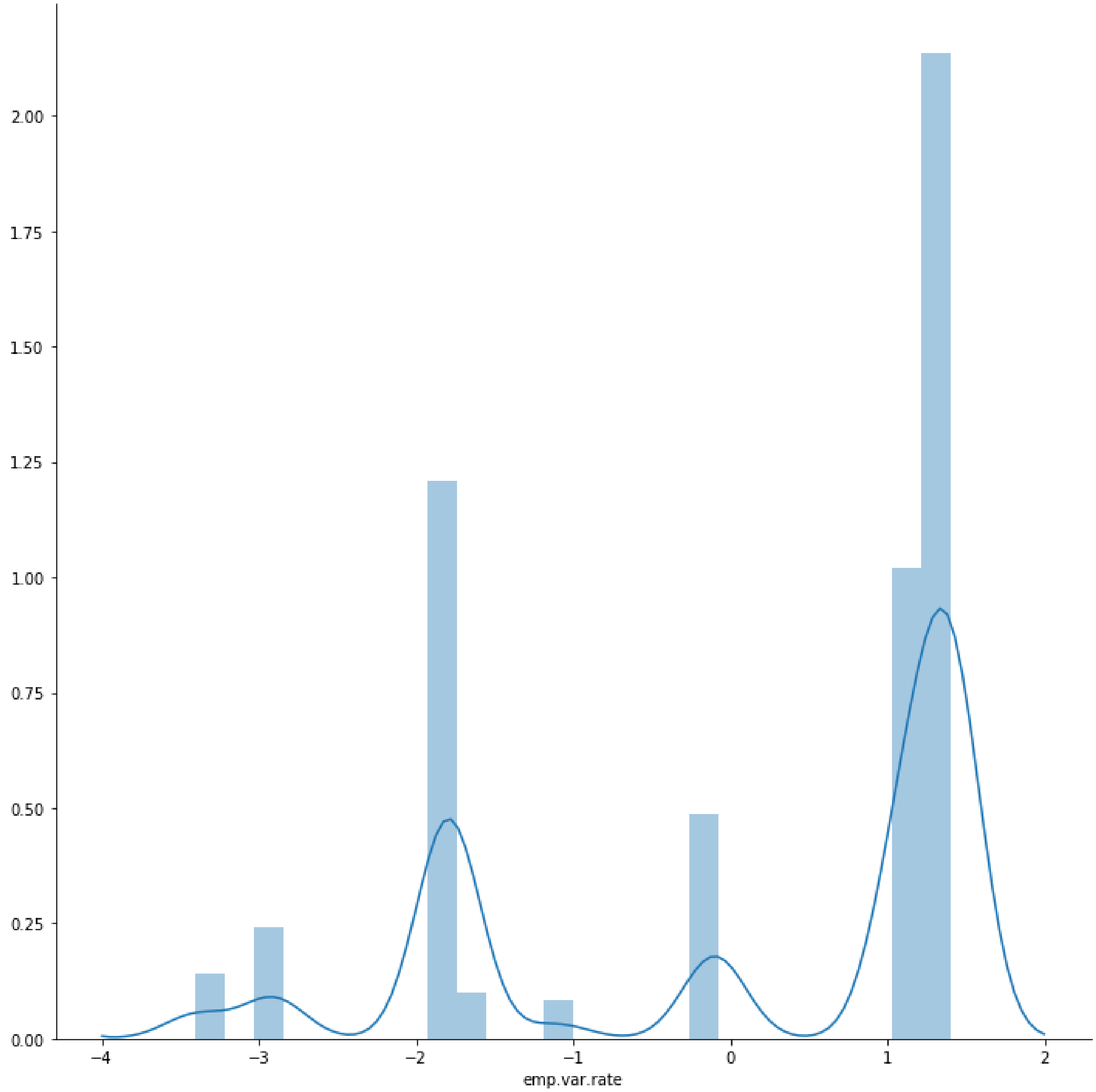


- Using the box plot as shown below, campaigns more than 10 are considered as outliers.

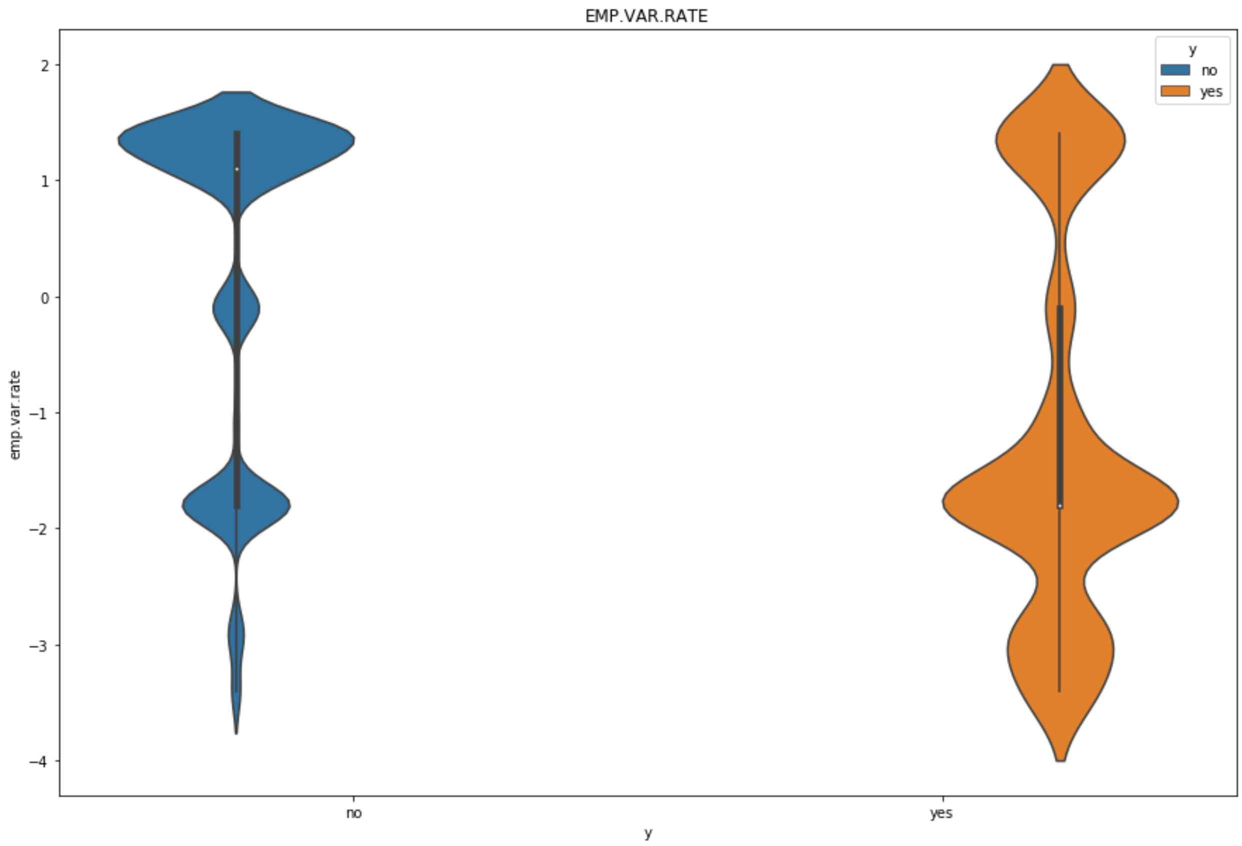


Detecting outliers for the variable emp.var.rate

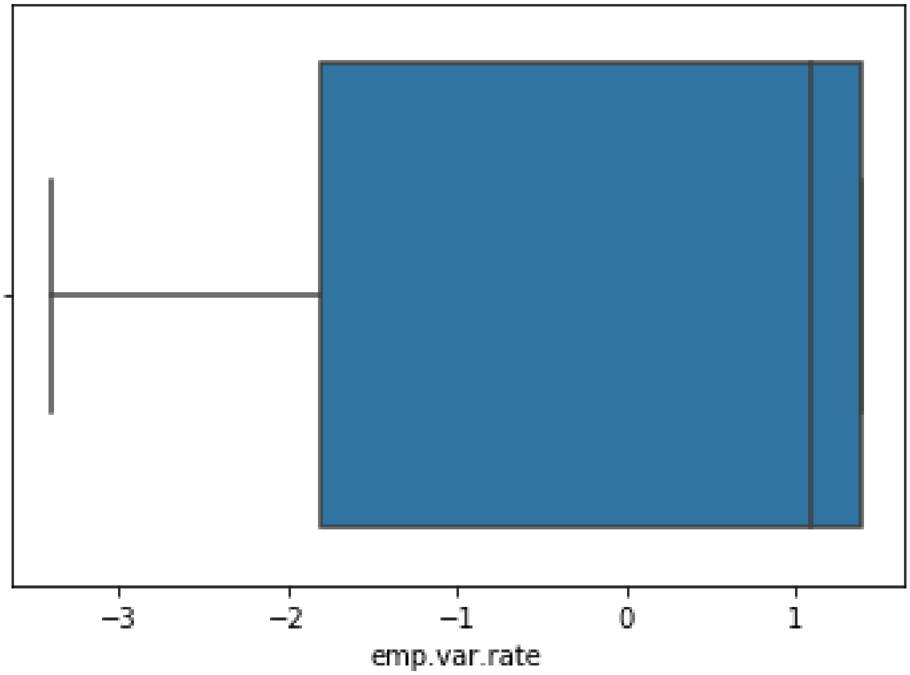
- Using the Kde plot, we are not able to determine whether outliers are present.



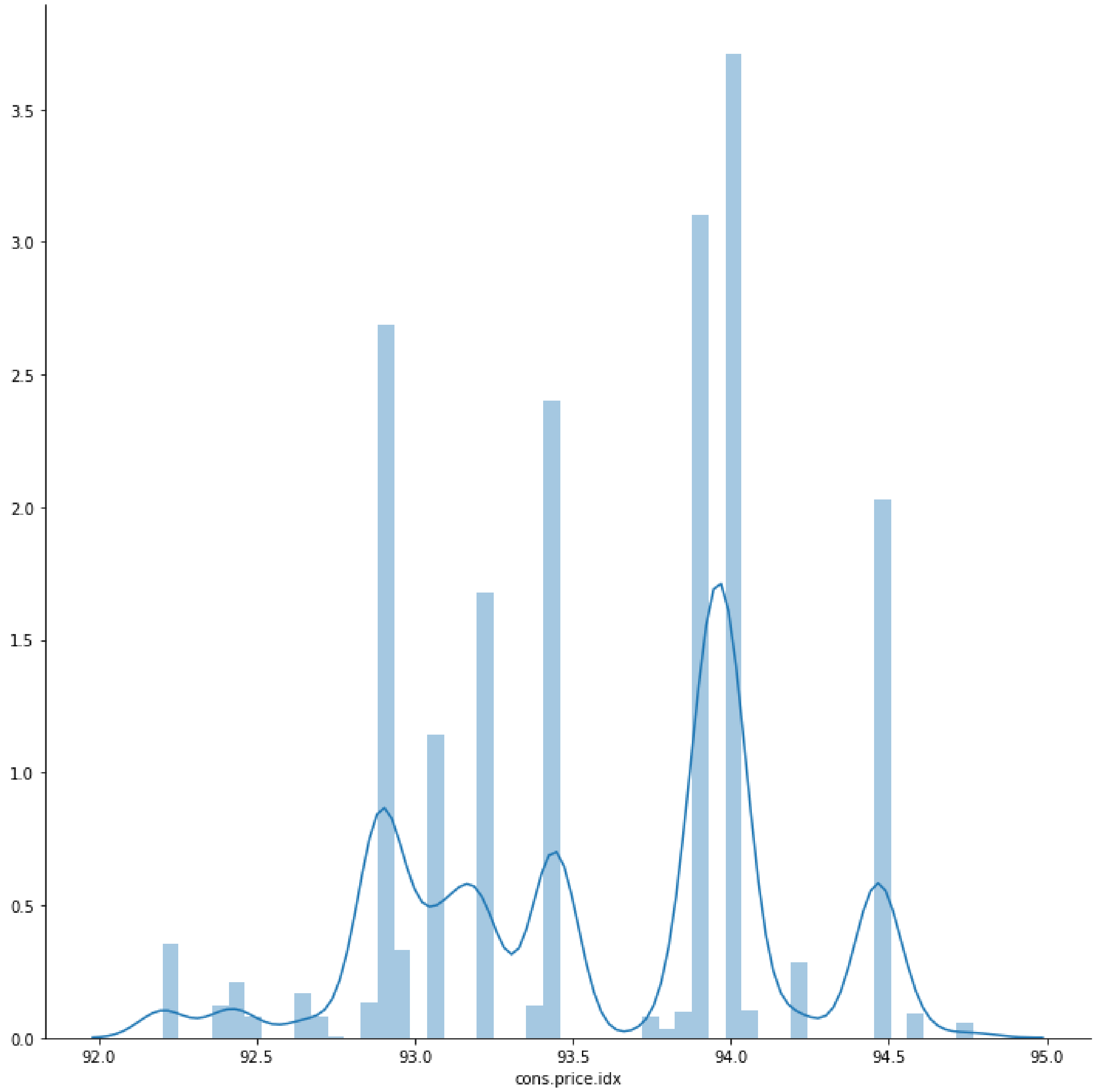
- Using the Violin plot as shown below, there are no outliers present.

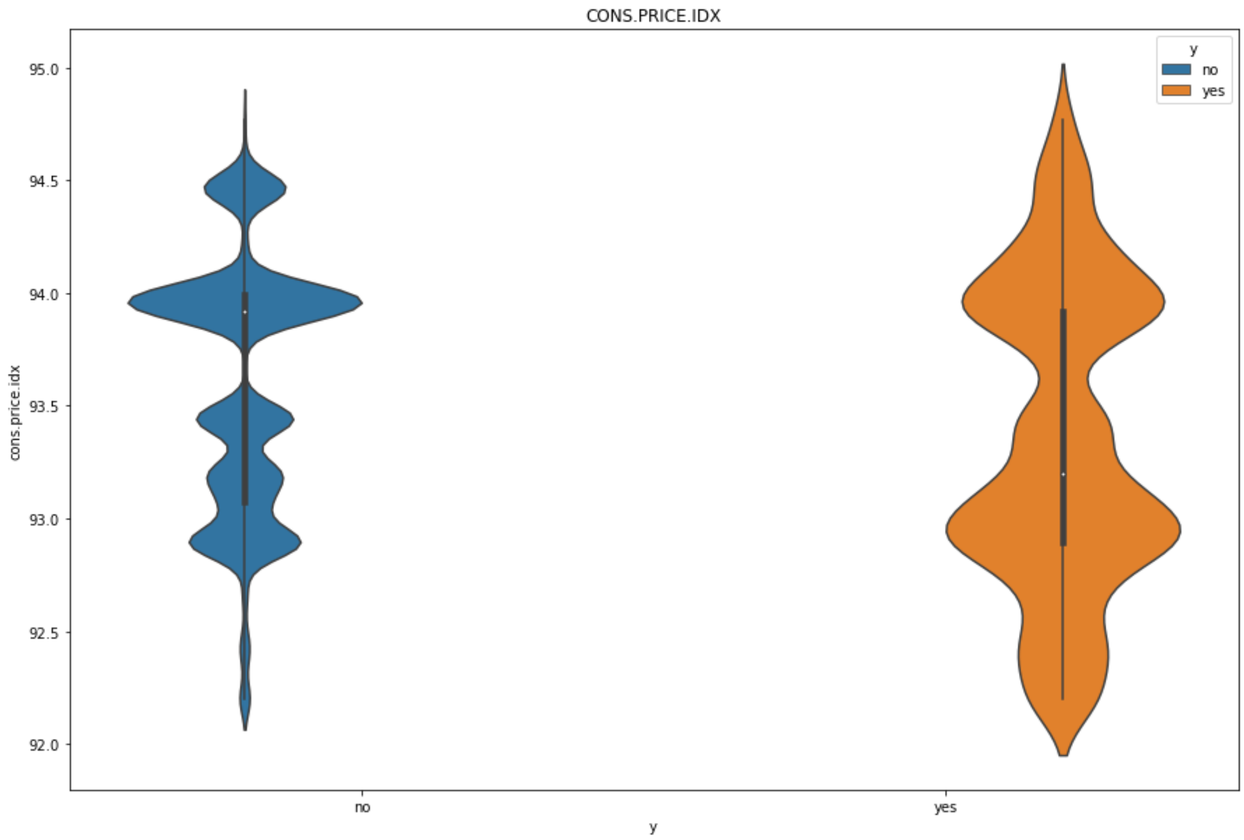


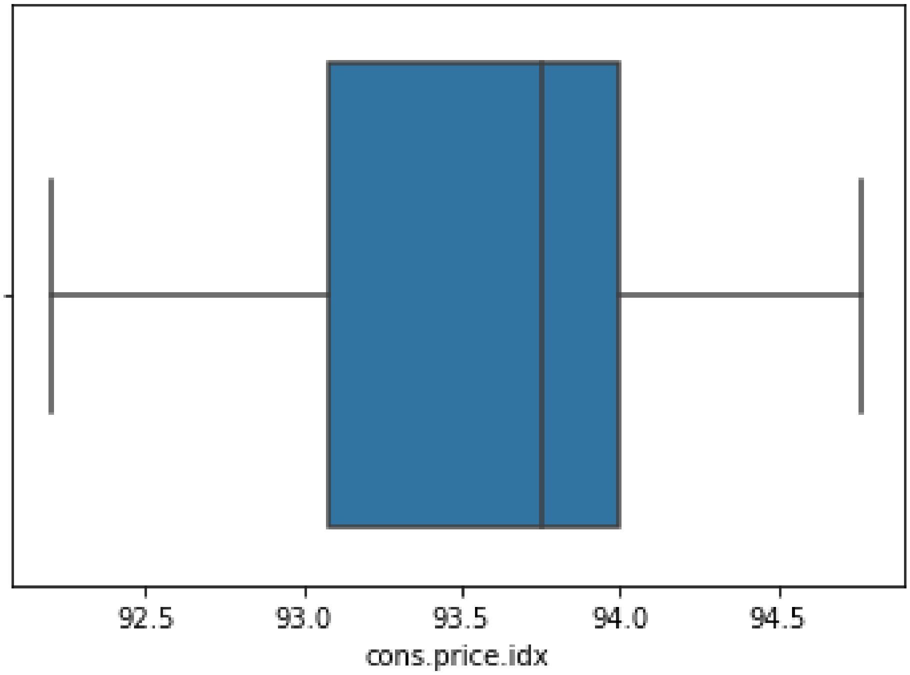
- Using the box plot as shown below, we also see no outliers present.



Detecting outliers for the variable cons.price.idx

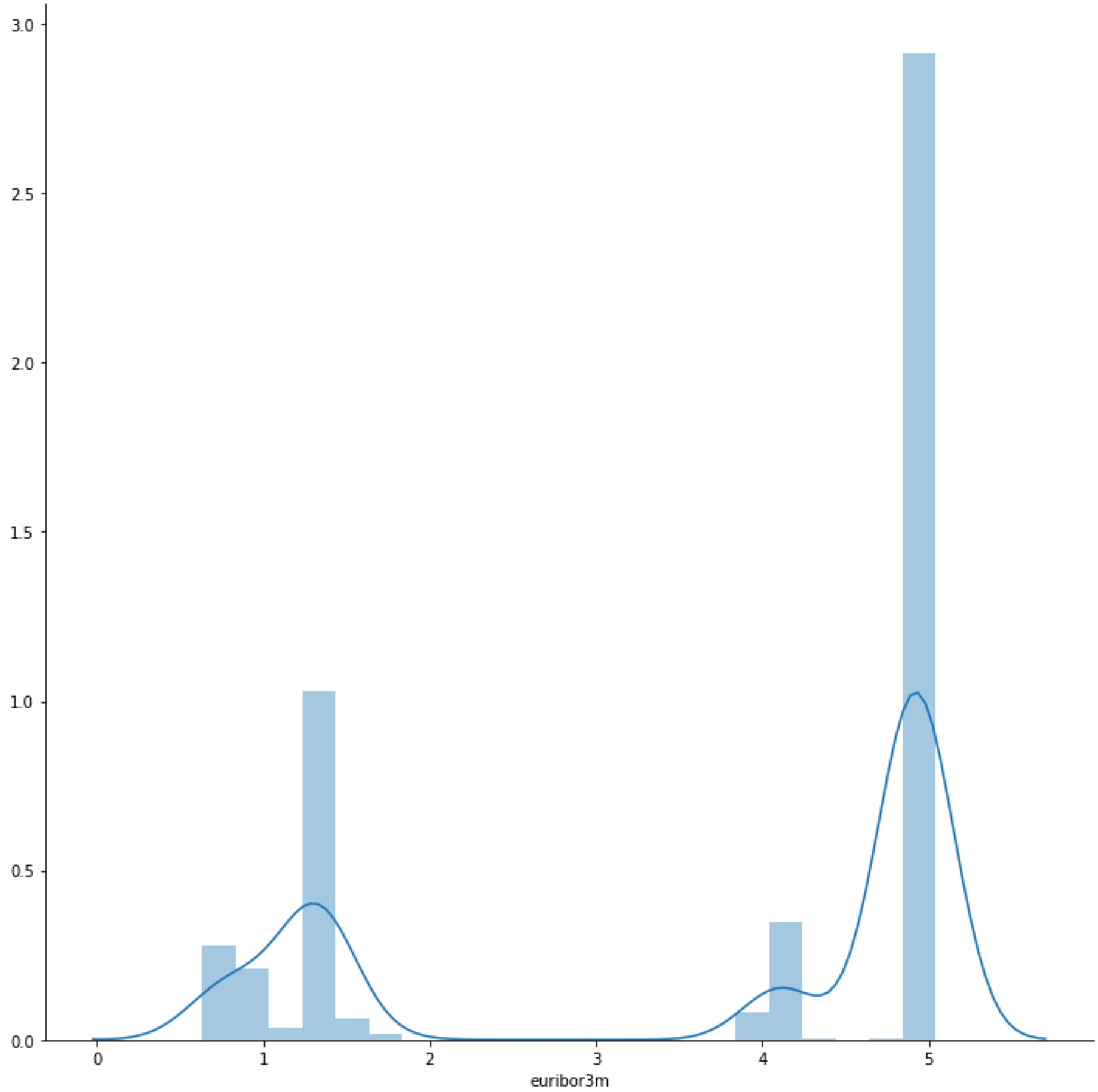


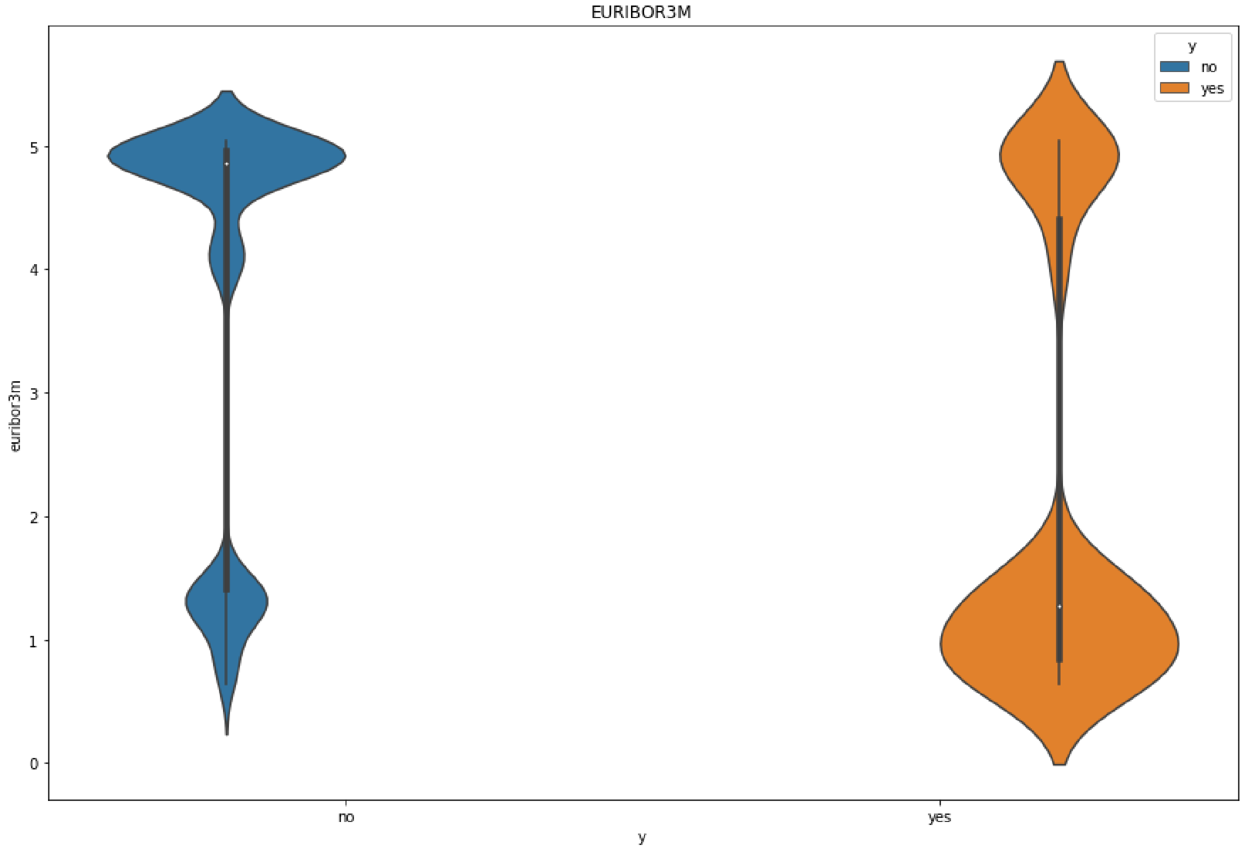


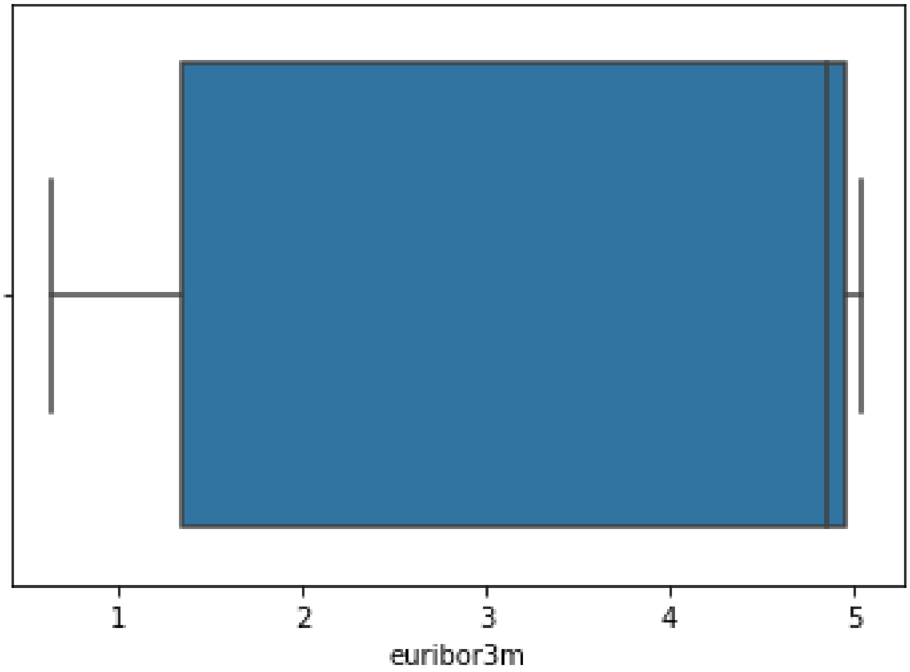


- From the above 3 plots, **cons.price.idx** does not contain any outliers.

Detecting outliers for the variable euribor3m

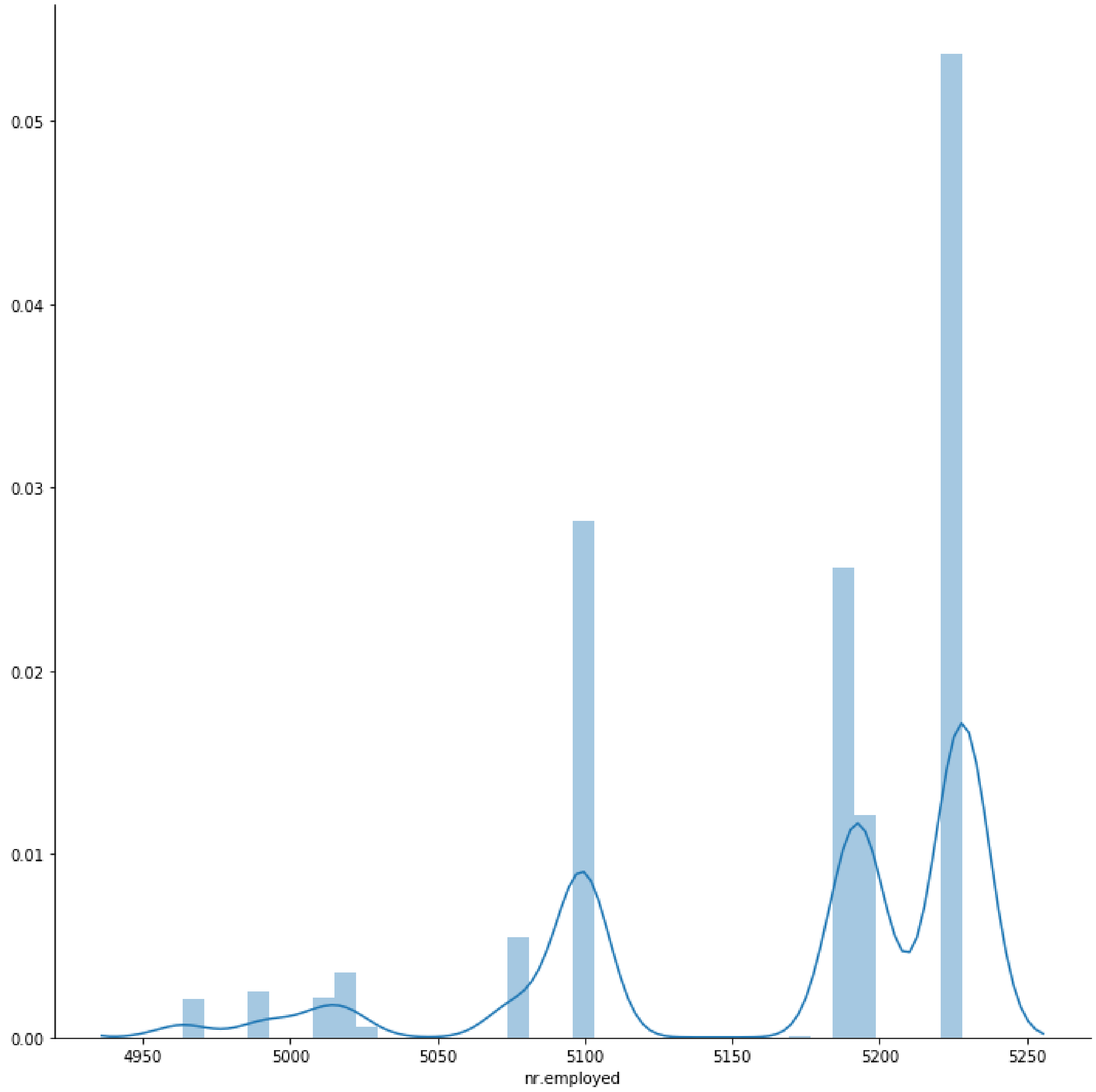


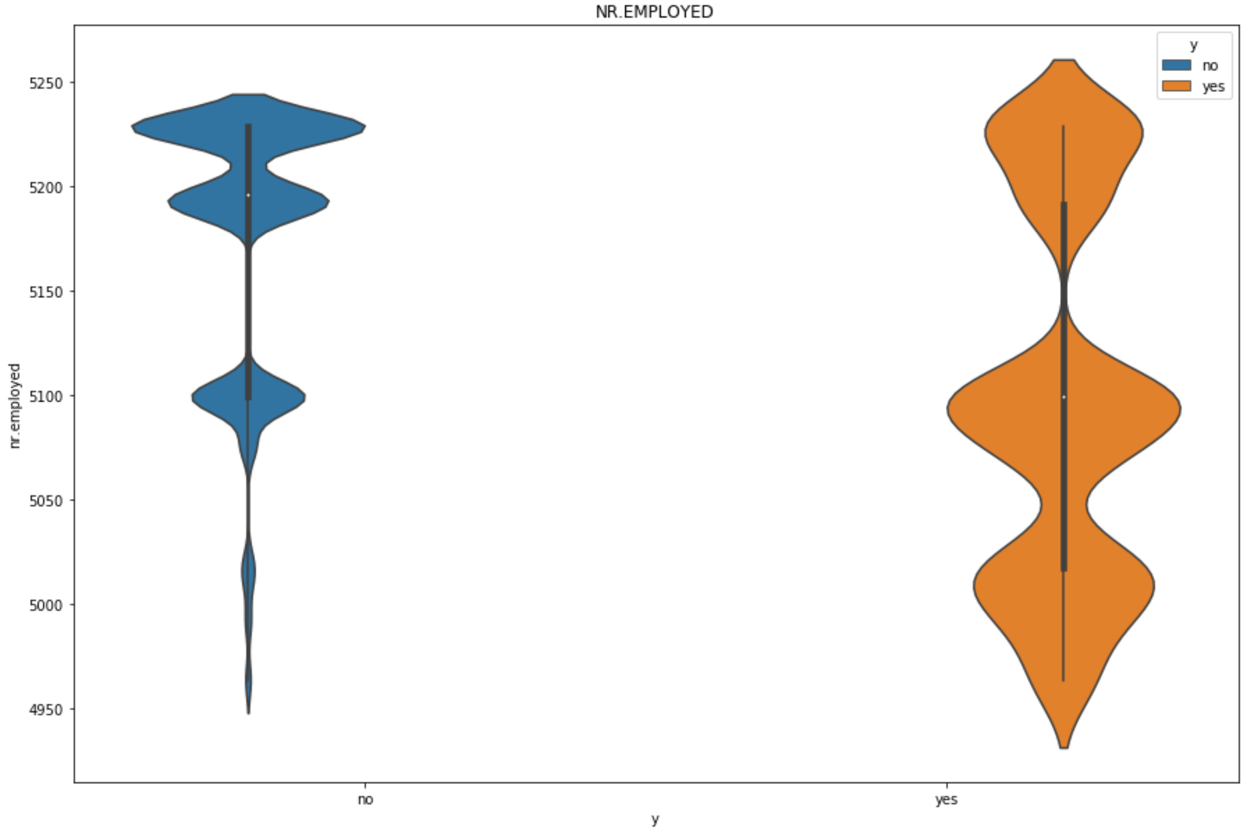


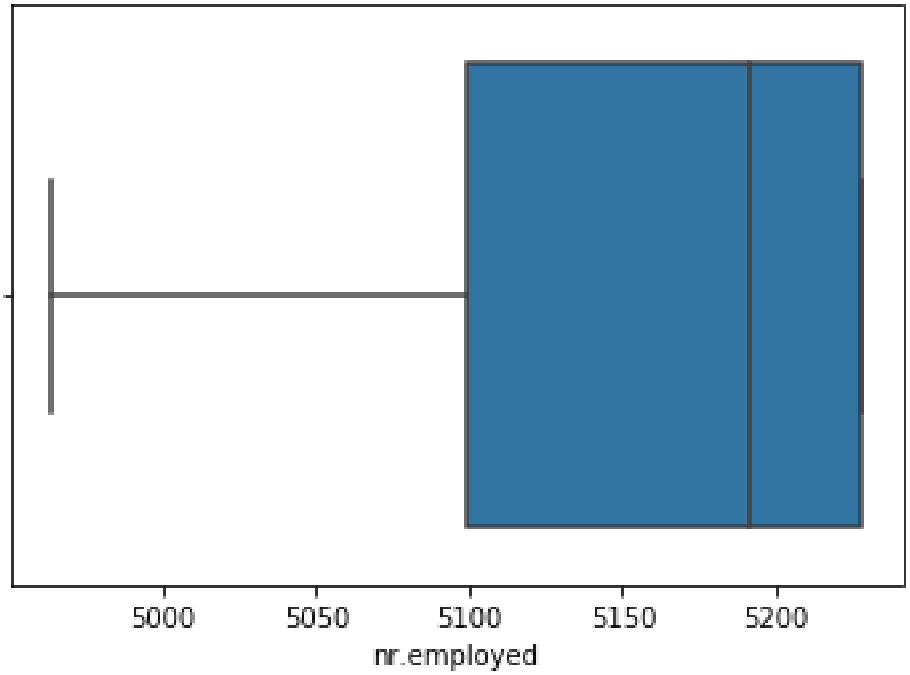


- From the above 3 plots, we can see that Euribor3m does not contain outliers.

Detecting outliers for the variable nr. employed

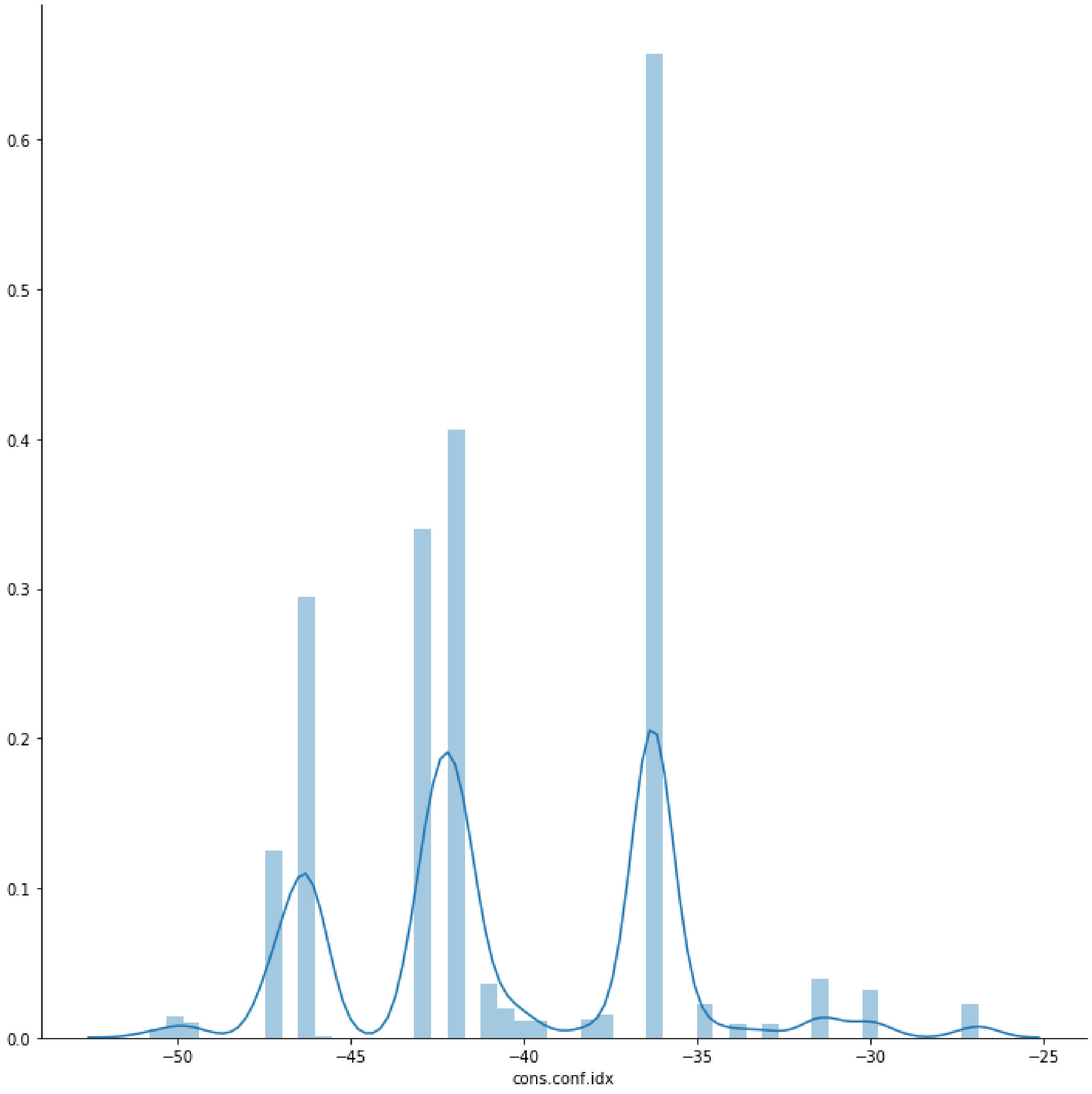


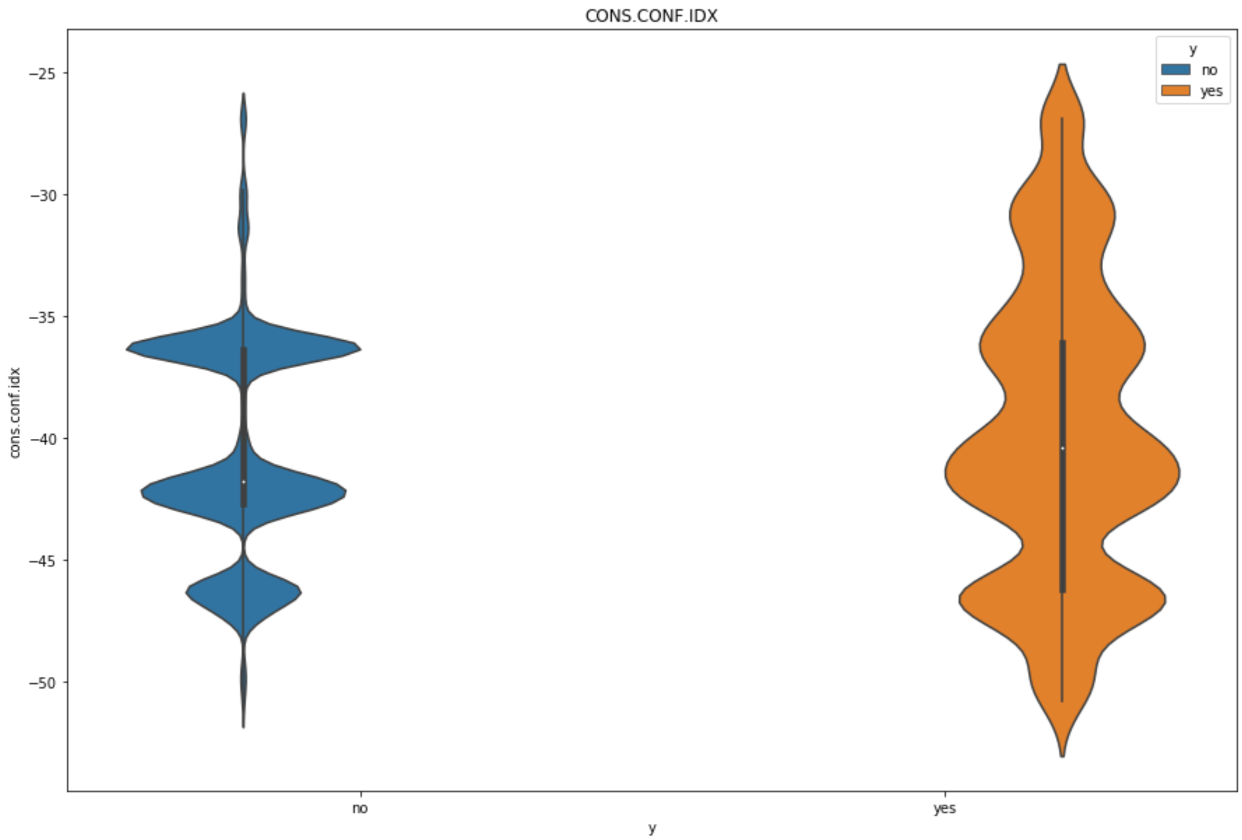


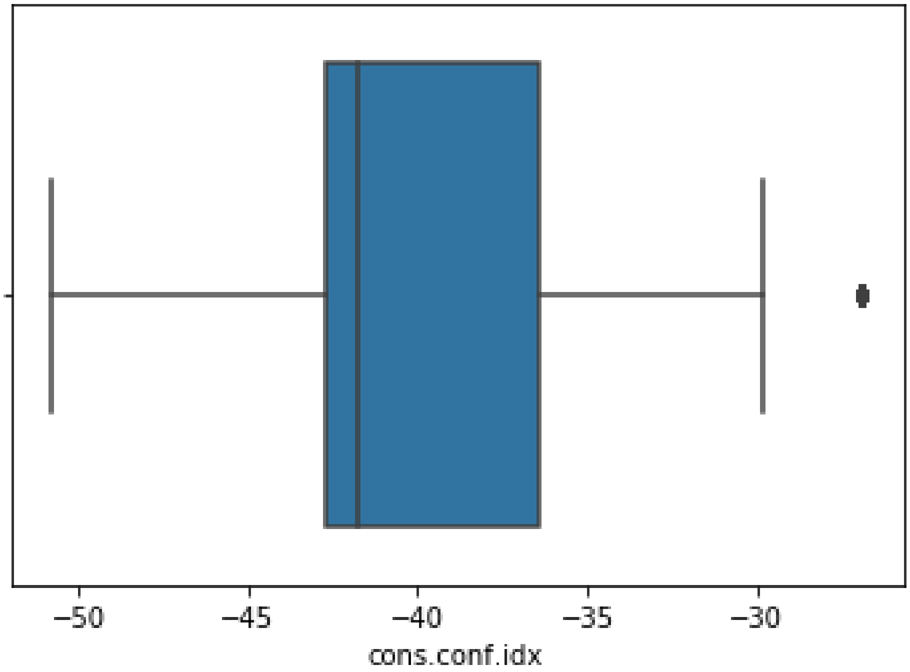


From the above 3 plots as we can see, nr. employed does not contain outliers.

Detecting outliers for the variable cons.conf.idx



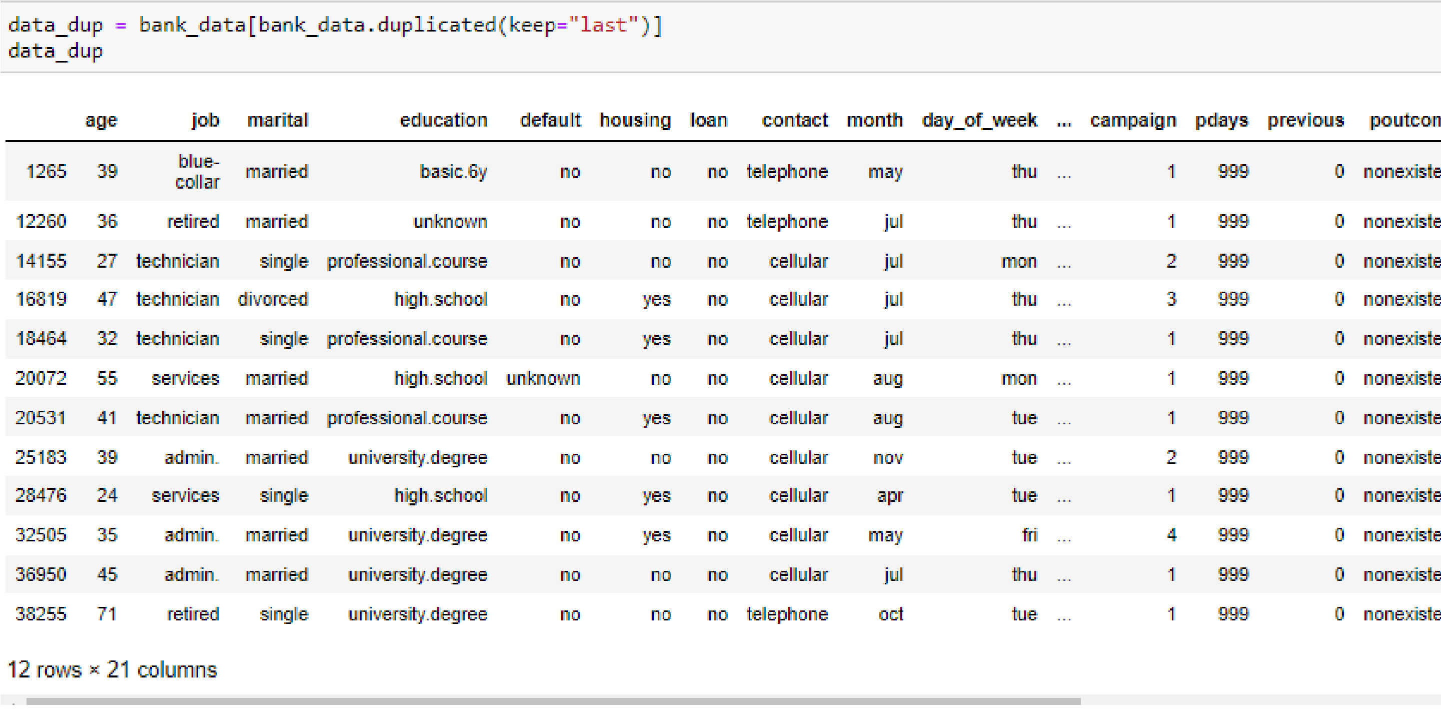




- In cons.conf.idx feature for class labels no, there is an outlier present when value above -30.

➢ Duplicate values on a dataset

- According to the dataset, we have 12 rows which are duplicates. We can attest to this as shown below.



```
data_dup = bank_data[bank_data.duplicated(keep="last")]
data_dup
```

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	...	campaign	pdays	previous	poutcon
1265	39	blue-collar	married	basic.6y	no	no	no	telephone	may	thu	...	1	999	0	nonexiste
12260	36	retired	married	unknown	no	no	no	telephone	jul	thu	...	1	999	0	nonexiste
14155	27	technician	single	professional.course	no	no	no	cellular	jul	mon	...	2	999	0	nonexiste
16819	47	technician	divorced	high.school	no	yes	no	cellular	jul	thu	...	3	999	0	nonexiste
18464	32	technician	single	professional.course	no	yes	no	cellular	jul	thu	...	1	999	0	nonexiste
20072	55	services	married	high.school	unknown	no	no	cellular	aug	mon	...	1	999	0	nonexiste
20531	41	technician	married	professional.course	no	yes	no	cellular	aug	tue	...	1	999	0	nonexiste
25183	39	admin.	married	university.degree	no	no	no	cellular	nov	tue	...	2	999	0	nonexiste
28476	24	services	single	high.school	no	yes	no	cellular	apr	tue	...	1	999	0	nonexiste
32505	35	admin.	married	university.degree	no	yes	no	cellular	may	fri	...	4	999	0	nonexiste
36950	45	admin.	married	university.degree	no	no	no	cellular	jul	thu	...	1	999	0	nonexiste
38255	71	retired	single	university.degree	no	no	no	telephone	oct	tue	...	1	999	0	nonexiste

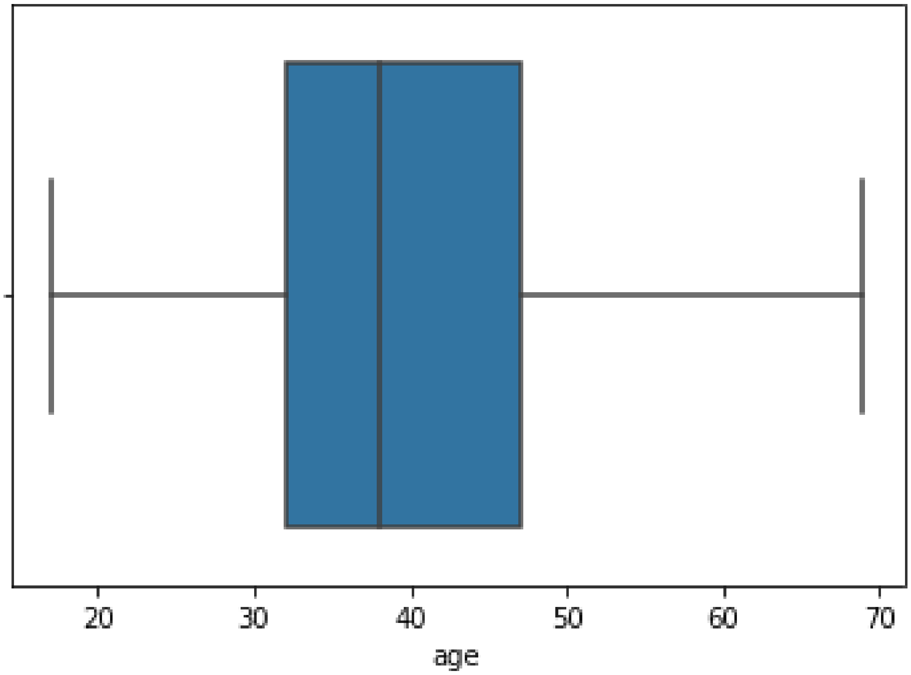
12 rows × 21 columns

- **What approaches can be used to solve the above mentioned problems?**
- Some problems were identified in the data, and which were as follows:
  - ➢ **Outliers in the data**
  - ➢ **Duplicates in the data**

➢ Outliers in the data

- One of the ways to treat the outliers for the numerical values is by creating a filter based on the boxplot obtained and apply the filter to the data. For an example, for the variable **age,** as the outliers are present after the age of 70 years, we can create a filter based by saying age must be less than 70 and age should not exceed 70 years.

An outlier is treated after we've created a filter based and applied the filter to the data.



We can perform some outlier treatment through the rest of the variables where the outliers are present.

- Another alternative approach which can be applied, is by converting the numerical values to categorical. This means numerical values can be binned. For an example for the variable **age,** we can convert **age** to **age_group.**

- Feature scaling is also one of the very helpful approaches for data preprocessing using **feature scailing: standardization.** Standardization is a popular feature scaling method, which gives data the property of a standard normal distribution (also known as Gaussian distribution). All features are standardized on the normal distribution (a mathematical model). **StandardScaler** standardizes the various data features to a uniform scale to allow them to be compared and processed together. This approach is used when splitting between independent and dependent variables

- **Duplicates in the data**

- To treat duplicate values in a row, the best choice would be to drop these duplicate rows.

Github Repo link: https://github.com/Nkululeko353/Week-8-Deliverables