Ćwiczenie 3 & 4 & 5

Wstęp

Celem zestawu ćwiczeń jest praktyczna analiza danych pochodzących z kampanii marketingowej oraz wykorzystanie metod statystycznych i modeli predykcyjnych, by odpowiedzieć na pytanie: czy klient zdecyduje się na subskrypcję? Wnioski płynące z tej analizy mogą mieć realne zastosowanie w optymalizacji działań marketingowych, zwiększaniu skuteczności kampanii sprzedażowych oraz lepszym dopasowaniu ofert do oczekiwań klientów.

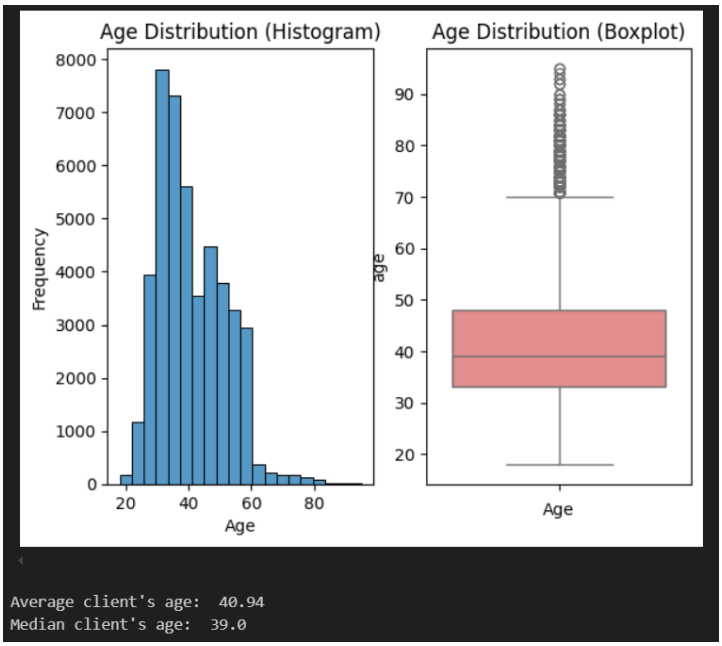
Wykonanie

Na początku procesu zaznajamiania się danymi sprawdzono, jakie cechy każdego klienta znajdują się w badanym datasecie oraz czy występują brakujące wartości:

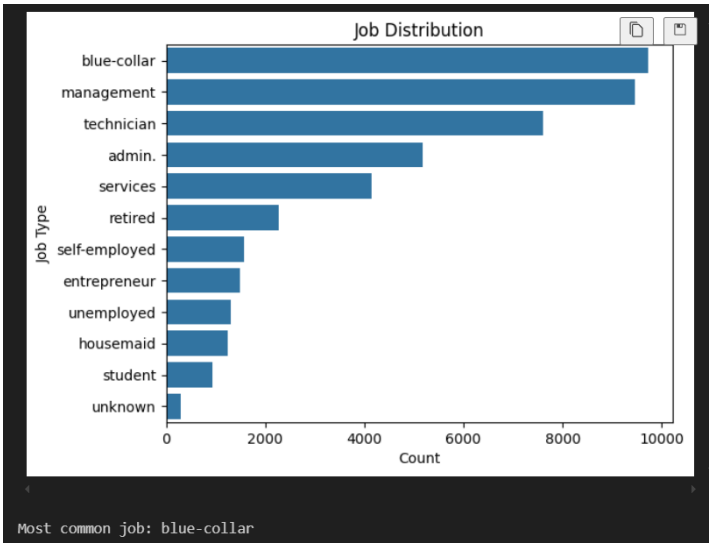
```
age          0
job          0
marital      0
education    0
default      0
balance      0
housing      0
loan         0
contact      0
day          0
month        0
duration     0
campaign     0
pdays        0
previous     0
poutcome     0
y            0
```

Z racji niewystępowania wartości NULL w danych, sprawdzono każdą cechę; jej rozkład, statystyki opisowe i wiele innych parametrów:

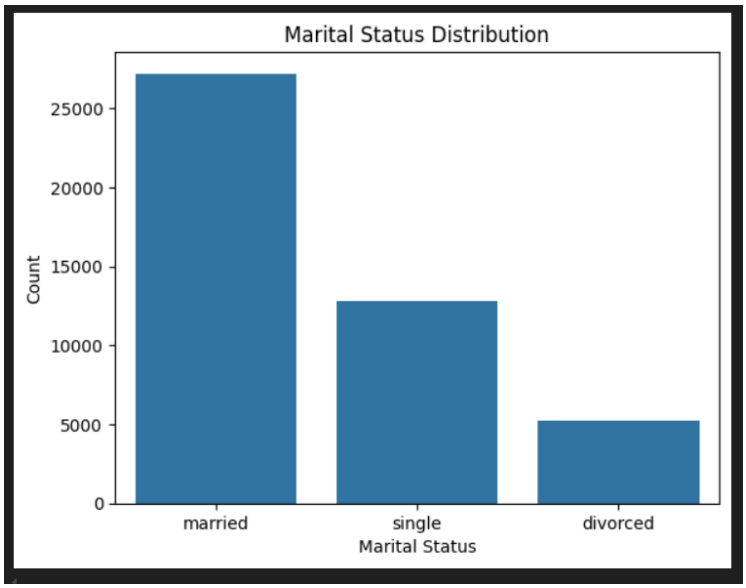
1. Zmienna age



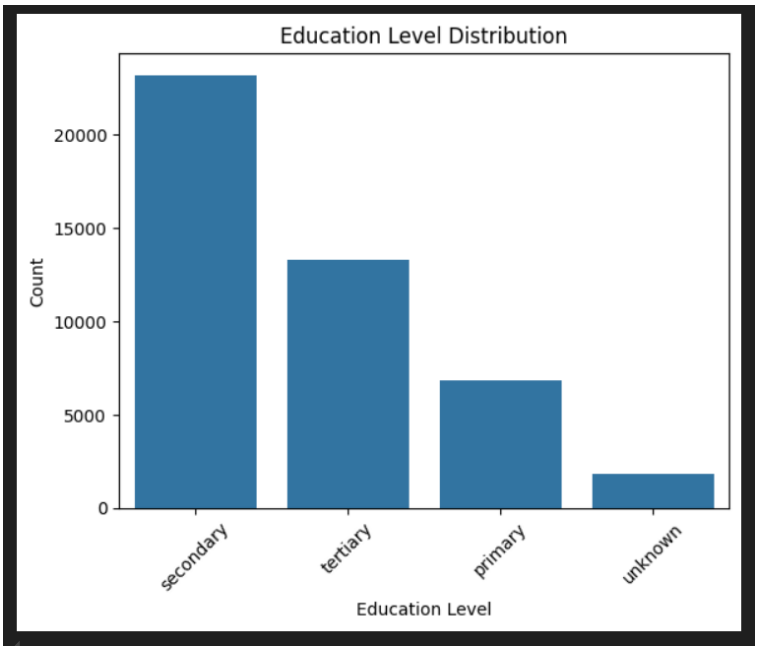
2. Zmienna job



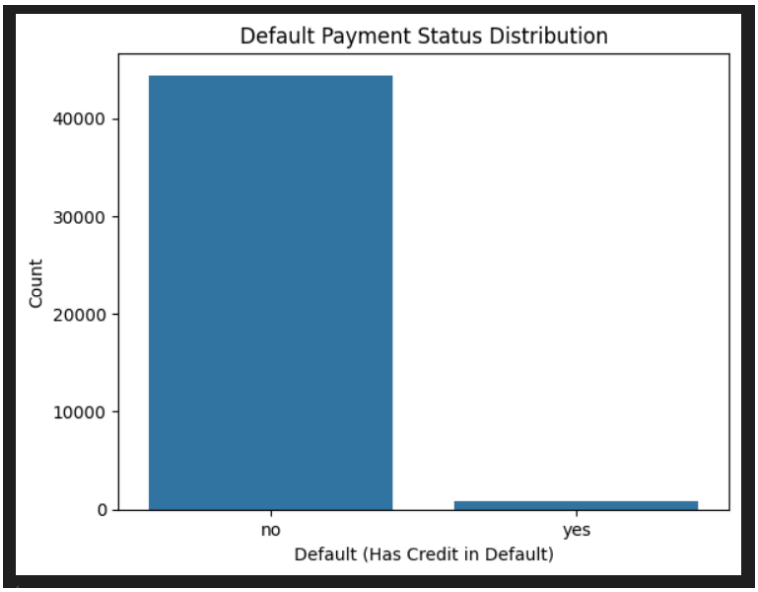
3. Zmienna marital



4. Zmienna education

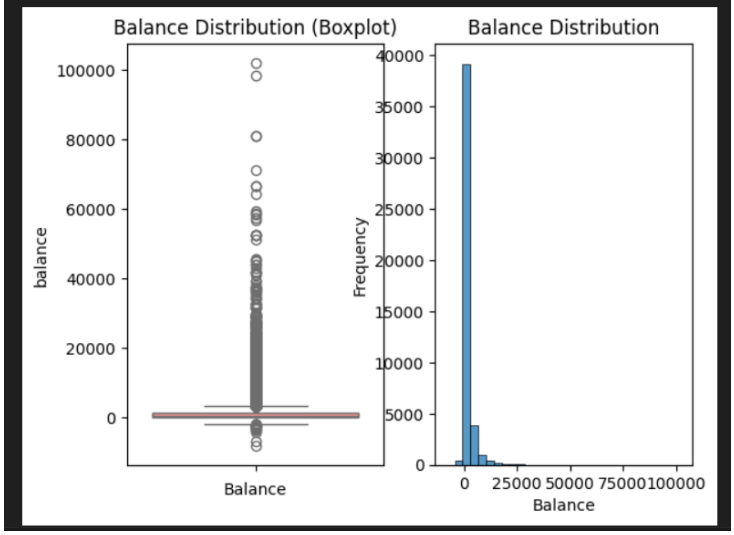


5. Zmienna default



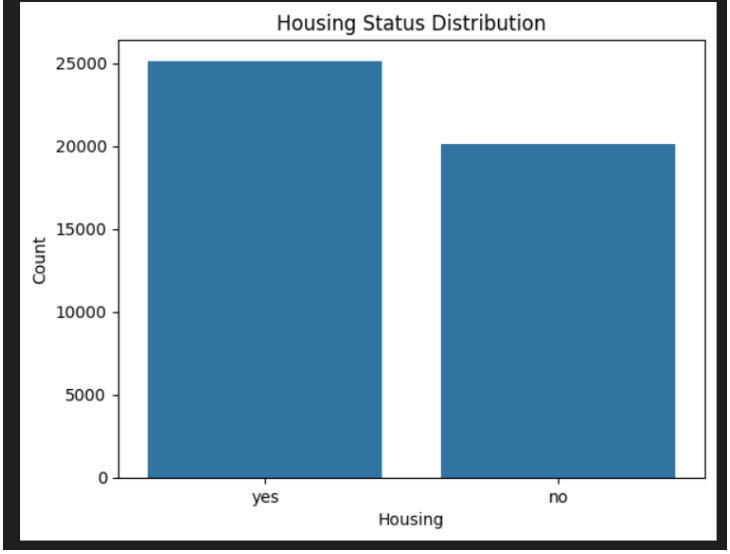
6. Zmienna balance

```
count     45211.000000
mean       1362.272058
std        3044.765829
min       -8019.000000
25%          72.000000
50%         448.000000
75%        1428.000000
max      102127.000000
Name: balance, dtype: float64
```



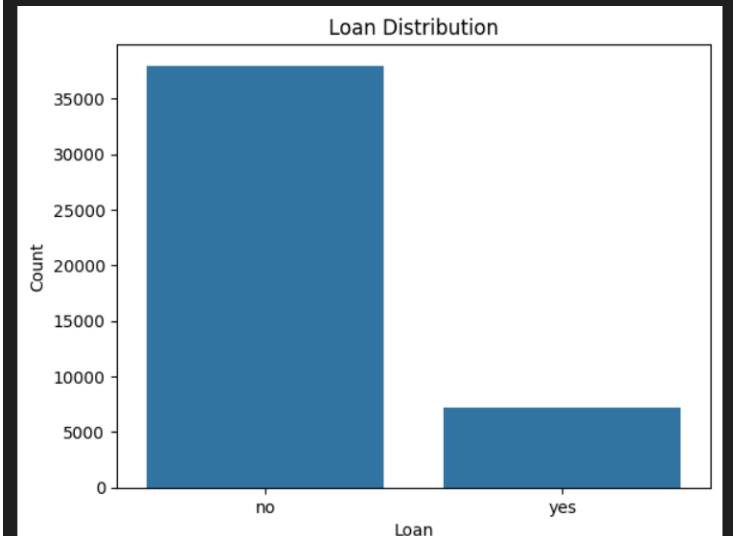
7. Zmienna housing

```
housing
yes    25130
no     20081
Name: count, dtype: int64

Text(0, 0.5, 'Count')
```



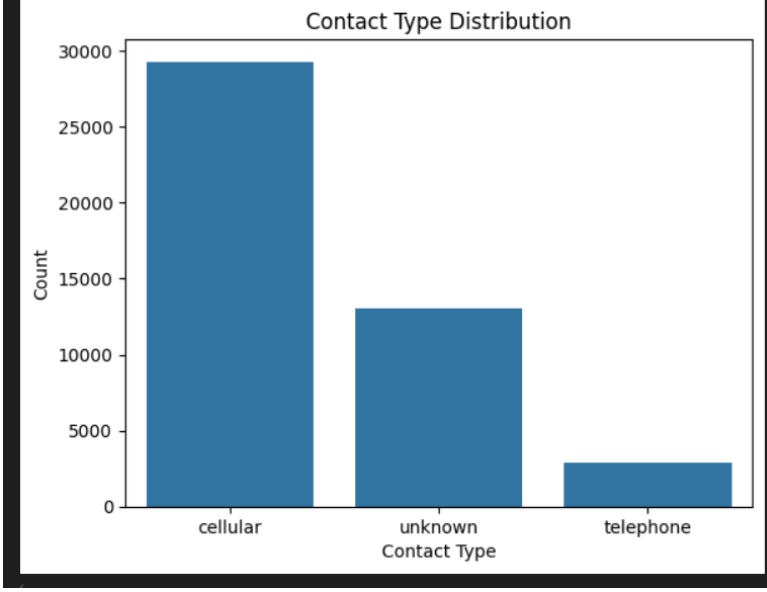
8. Zmienna loan

```
loan
no     37967
yes     7244
Name: count, dtype: int64

Text(0, 0.5, 'Count')
```

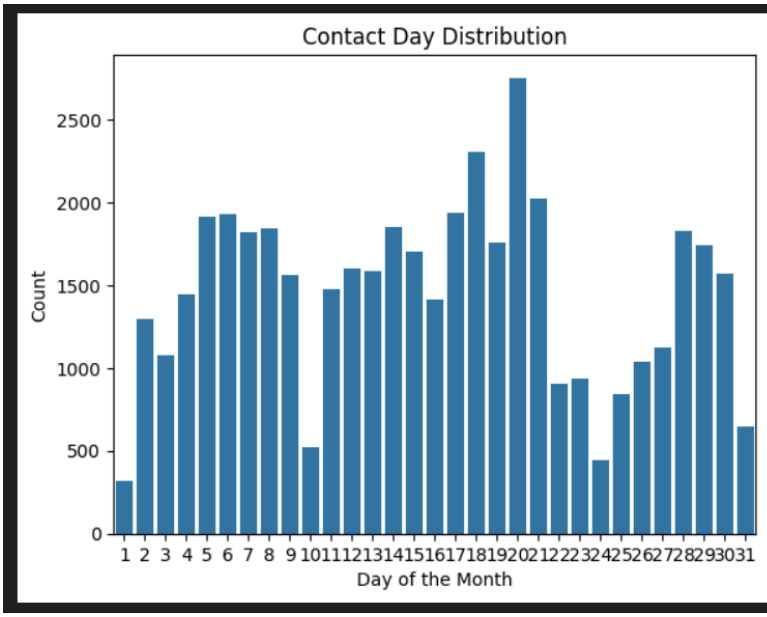


9. Zmienna contact

```
contact
cellular     29285
unknown      13020
telephone     2906
Name: count, dtype: int64
```

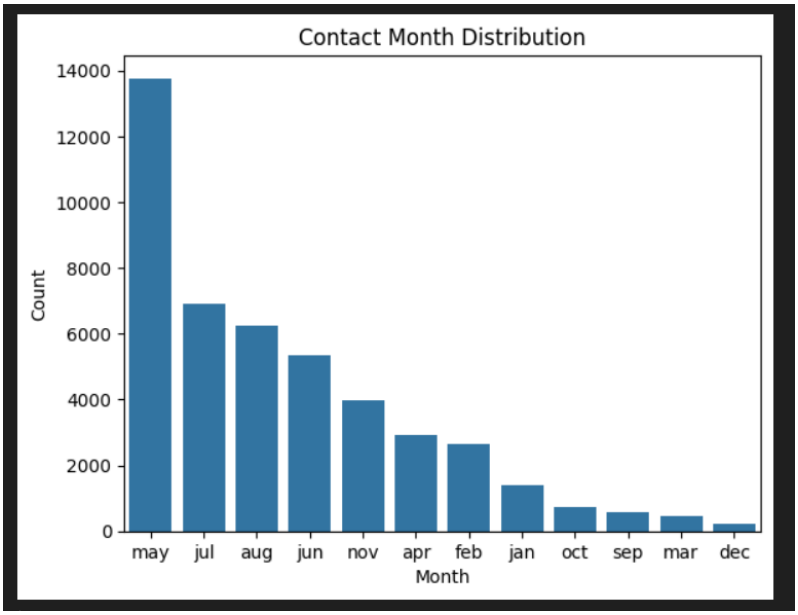


10. Zmienna day



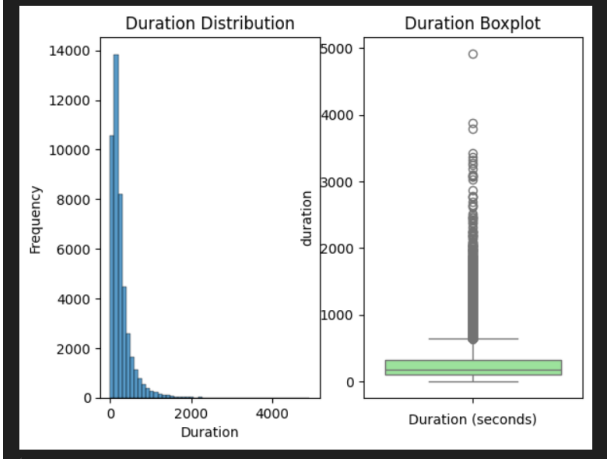
11. Zmienna month

```
month
may    13766
jul     6895
aug     6247
jun     5341
nov     3970
apr     2932
feb     2649
jan     1403
oct      738
sep      579
mar      477
dec      214
Name: count, dtype: int64
```



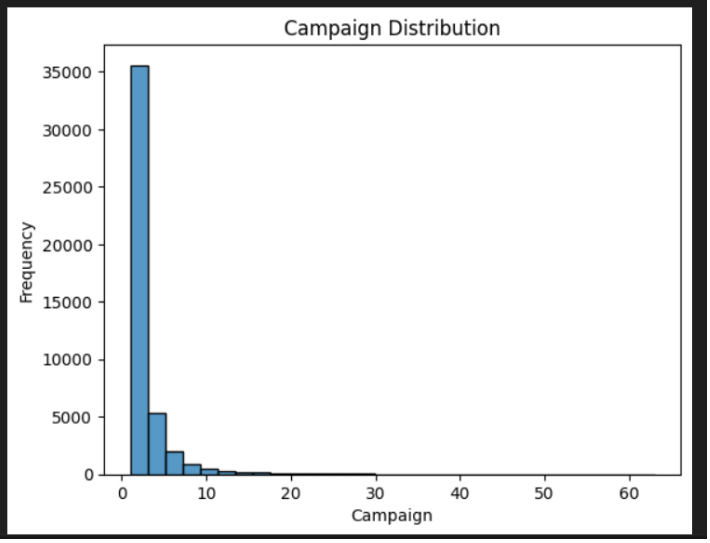
12. Zmienna duration

```
count    45211.000000
mean       258.163080
std        257.527812
min          0.000000
25%        103.000000
50%        180.000000
75%        319.000000
max       4918.000000
Name: duration, dtype: float64
```



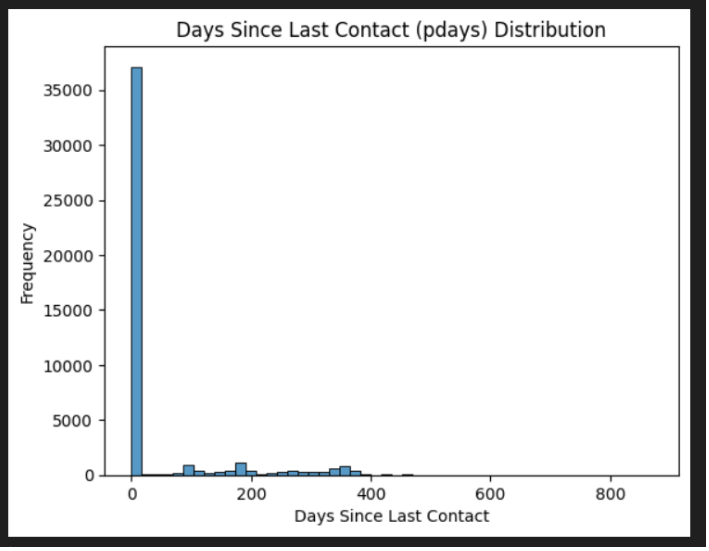
13. Zmienna campaign

```
count    45211.000000
mean         2.763841
std          3.098021
min          1.000000
25%          1.000000
50%          2.000000
75%          3.000000
max         63.000000
Name: campaign, dtype: float64
```



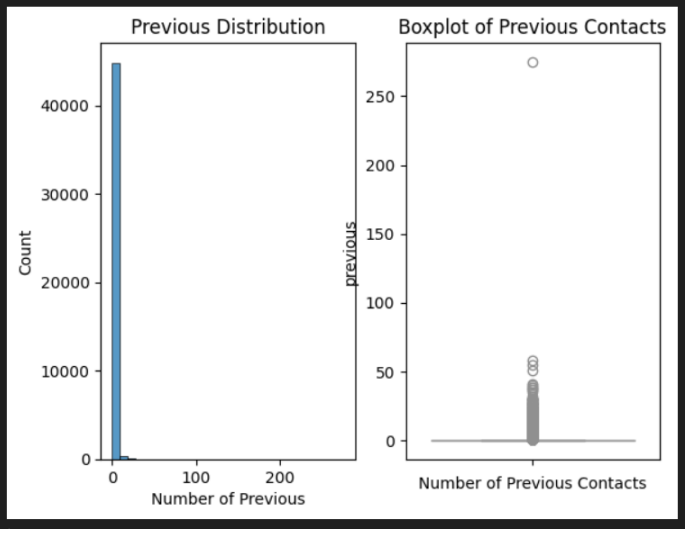
14. Zmienna pdays

```
count    45211.000000
mean        40.197828
std        100.128746
min         -1.000000
25%         -1.000000
50%         -1.000000
75%         -1.000000
max        871.000000
Name: pdays, dtype: float64
```



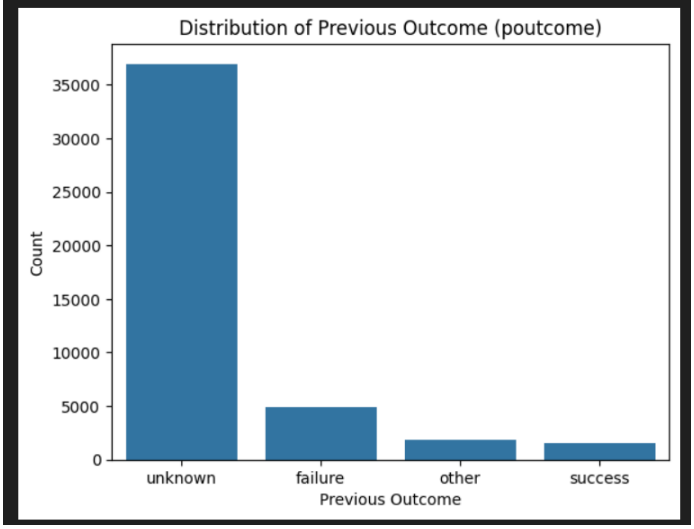
15. Zmienna previous

```
count    45211.000000
mean         0.580323
std          2.303441
min          0.000000
25%          0.000000
50%          0.000000
75%          0.000000
max        275.000000
Name: previous, dtype: float64
```



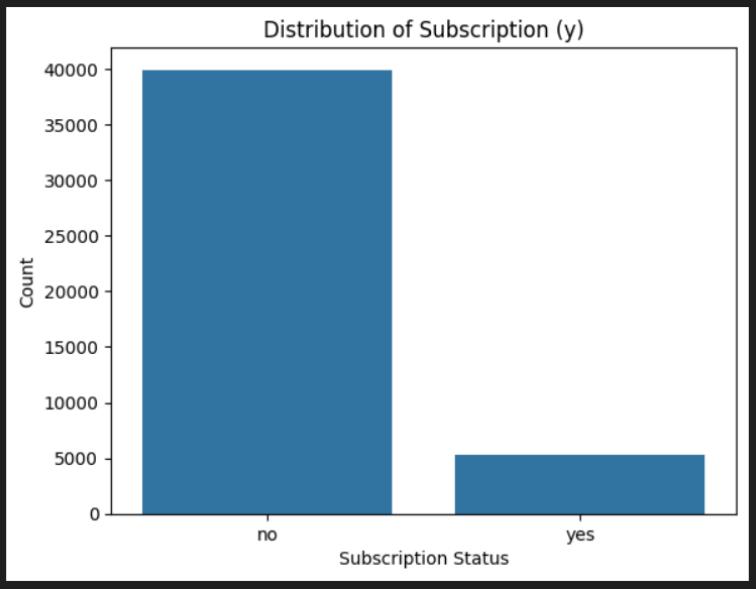
16. Zmienna outcome

```
poutcome
unknown    36959
failure     4901
other       1840
success     1511
Name: count, dtype: int64
```



17. Zmienna y (zmienna zależna)

```
y
no      39922
yes      5289
Name: count, dtype: int64
```




```
Percentage of clients who subscribed: 11.70%
```

Podczas procesu zaznajamiania się z danymi stwierdzono występowanie wielu wartości odstających (outlier-ów). Ich występowanie mogłoby negatywnie wpłynąć na modelowanie, stąd podjęto decyzję o pozbyciu się ich ze zbioru danych. W tym celu wykorzystano poniższe wzory do wyznaczenia dolnej i górnej granicy każdej zmiennej niezależnej:

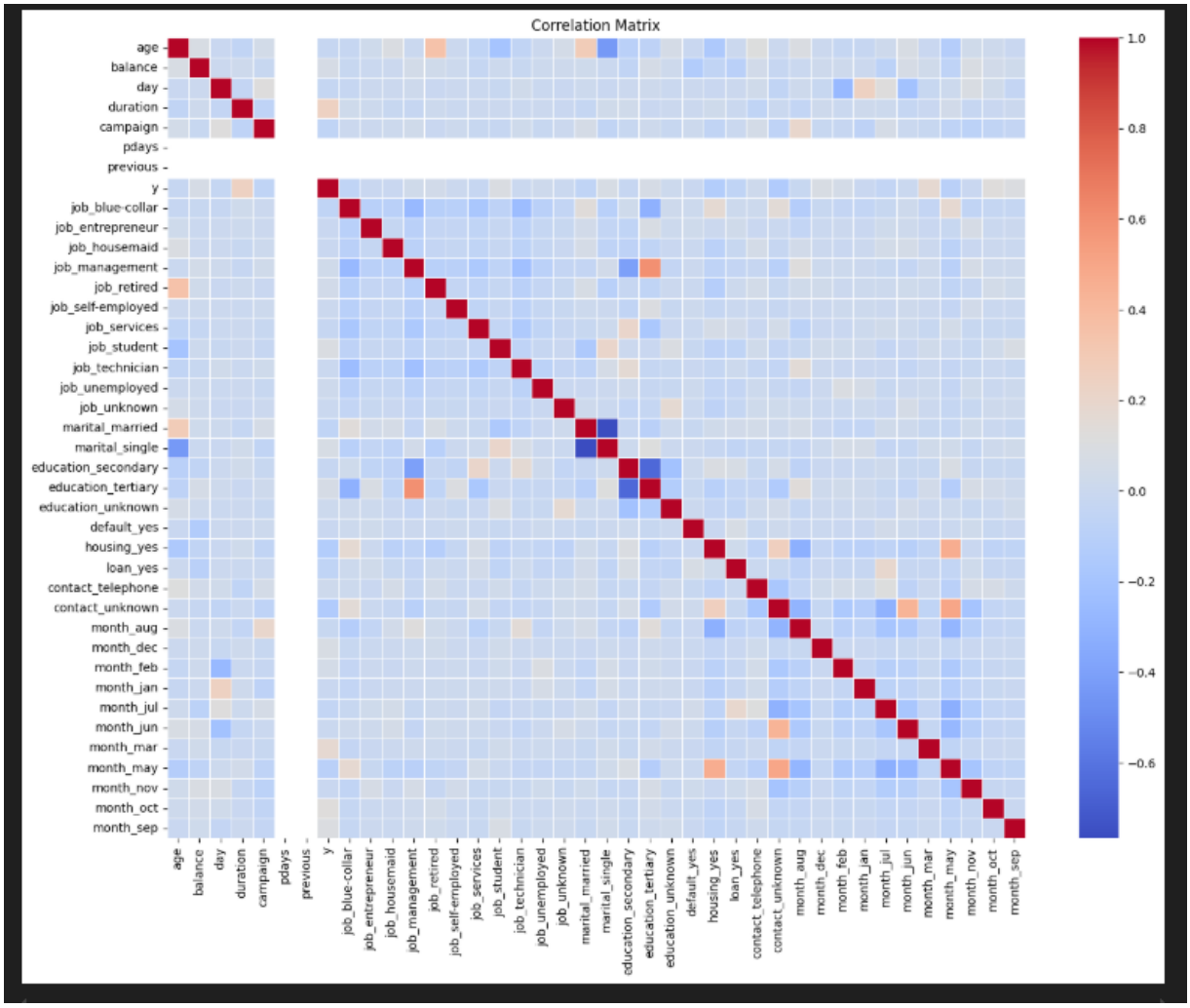
Dolna granica = Kwantyl(0.25) – 1.5*IQR

Górna granica = Kwantyl(0.75) + 1.5*IQR

Przed wykonaniem usunięcia outlierów zbiór danych liczył 45.211 rekordów. Po tym procesie ich liczbę zredukowano do 28.069 rekordów.

W następnym korku wykonano one-hot encoding, czyli proces konwersji danych kategorycznych na dane numeryczne. Procesowi temu uległa również zmienna zależna y. Klienci którzy wykupili subskrypcję oznaczono jako 1 (zamiast yes), a tych którzy nie zdecydowali się na zakup oznaczono 0 (zamiast no).

W dalszym etapie przeanalizowano korelację między zmiennymi. Wykonano w tym celu macierz korelacji:



W tym miejscu zakończono proces przygotowywania danych do modelowania.

W kolejnych krokach skupiono się na nauce modelów uczenia maszynowego, mających na celu przewidzenie czy klient wykupi subskrypcję, czy nie. Wykorzystano wiele różnych rodzajów modeli, a następnie przyjrzano się statystykom je opisującym, w celu określenia, który z nich najlepiej nadaje się do badanego problemu

Pierwszym utworzonym i przebadanym modelem był model lasu losowego. Poniżej zaprezentowano wyniki (dokładność i macierz korelacji):

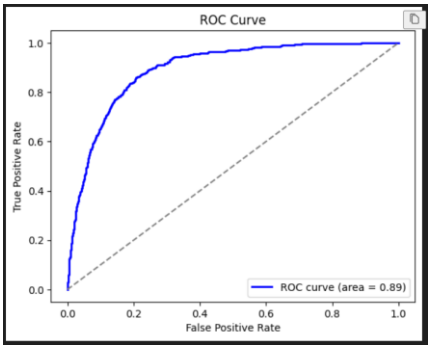
```
Accuracy: 0.95
Confusion matrix:
[[5273   24]
 [ 267   50]]
```

Drugim wykorzystanym modelem do przewidywania subskrypcji była regresja logistyczna (z wykorzystaniem wszystkich dostępnych cech):

```
Accuracy: 0.925
Confusion matrix:
[[5946  120]
 [ 373  113]]

Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.98      0.96      6066
           1       0.48      0.23      0.31       486

    accuracy                           0.92      6552
   macro avg       0.71      0.61      0.64      6552
weighted avg       0.91      0.92      0.91      6552
```

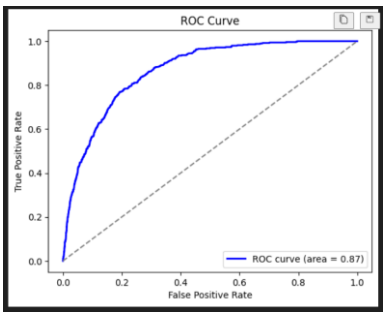
Trzecim wykorzystanym modelem do przewidywania subskrypcji była regresja logistyczna (z wykorzystaniem wszystkich dostępnych cech) przy wykorzystaniu funkcji skalującej dane:

```
Accuracy: 0.927
Confusion matrix:
[[5953  113]
 [ 368  118]]

Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.98      0.96      6066
           1       0.51      0.24      0.33       486

    accuracy                           0.93      6552
   macro avg       0.73      0.61      0.65      6552
weighted avg       0.91      0.93      0.91      6552
```

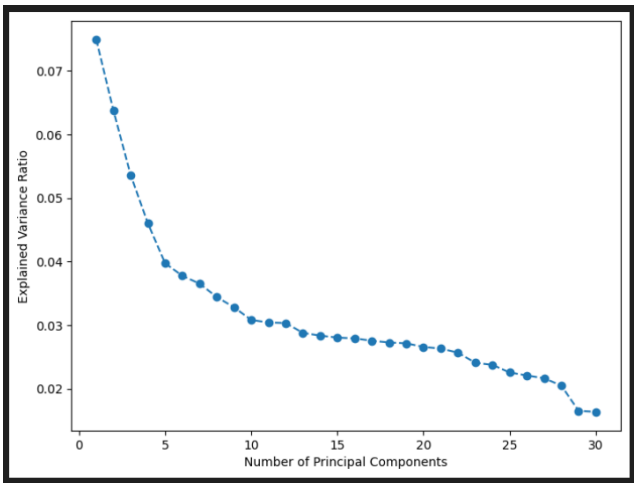


Czwartym wykorzystanym modelem do przewidywania subskrypcji była regresja logistyczna (z wykorzystaniem dostępnych cech dobrze korelujących się ze zmienną y) przy wykorzystaniu funkcji skalującej dane:

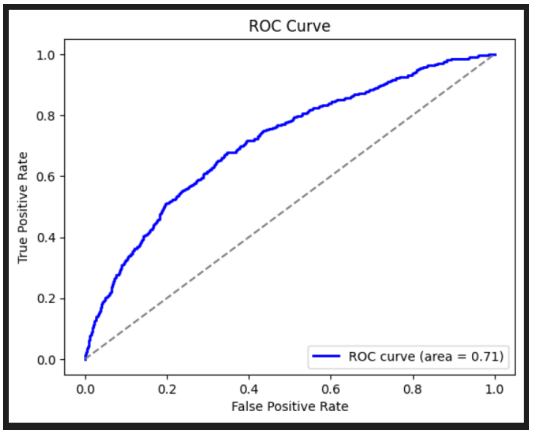
```
Wybrane cechy: ['duration', 'housing_yes', 'contact_unknown', 'month_mar', 'month_oct']
Accuracy: 0.925
Confusion matrix:
[[5959  107]
 [ 387   99]]

Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.98      0.96      6066
           1       0.48      0.20      0.29       486

    accuracy                           0.92      6552
   macro avg       0.71      0.59      0.62      6552
weighted avg       0.91      0.92      0.91      6552
```



Piątym wykorzystanym modelem do przewidywania subskrypcji była regresja logistyczna (z wykorzystaniem 5 komponentów utworzonych z pomocą algorytmu PCA) przy wykorzystaniu funkcji skalującej dane:

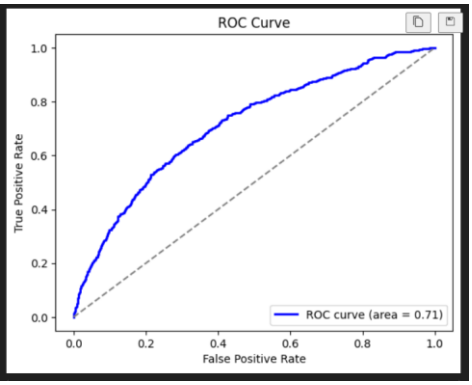



```
Number of components PCA: 5
Accuracy: 0.93

Confusion Matrix:
[[6054   12]
 [ 479    7]]

Classification Report:
              precision    recall  f1-score   support

           0       0.93      1.00      0.96      6066
           1       0.37      0.01      0.03       486

    accuracy                           0.93      6552
   macro avg       0.65      0.51      0.49      6552
weighted avg       0.89      0.93      0.89      6552
```



Szóstym wykorzystanym modelem do przewidywania subskrypcji była regresja logistyczna (z wykorzystaniem 5 komponentów utworzonych z pomocą algorytmu PCA) przy wykorzystaniu funkcji skalującej dane oraz algorytmu SMOTE do wyrównania liczebności obu kategorii zmiennej y:

```
Number of components PCA: 5
Accuracy: 0.86

Confusion Matrix:
[[5493  573]
 [ 335  151]]

Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.91      0.92      6066
           1       0.21      0.31      0.25       486

    accuracy                           0.86      6552
   macro avg       0.58      0.61      0.59      6552
weighted avg       0.89      0.86      0.87      6552
```



Siódmym wykorzystanym modelem do przewidywania subskrypcji była sztuczna sieć neuronowa; klasyfikator MLP złożony z 10 neuronów w jednej ukrytej warstwie (wykorzystując algorytm służący do standaryzacji danych):

✓ **Dokładność (accuracy):** 92.38%

✓ **Macierz klasyfikacji:**

	precision	recall	f1-score	support
0	0.95	0.97	0.96	6066
1	0.48	0.29	0.36	486
accuracy			0.92	6552
macro avg	0.71	0.63	0.66	6552
weighted avg	0.91	0.92	0.92	6552

Ósmym wykorzystanym modelem do przewidywania subskrypcji była sztuczna sieć neuronowa; klasyfikator MLP złożony z 10 neuronów w pięciu ukrytych warstwach (wykorzystując algorytm służący do standaryzacji danych):

✓ **Dokładność (accuracy):** 92.43%

✓ **Macierz klasyfikacji:**

	precision	recall	f1-score	support
0	0.95	0.97	0.96	6066
1	0.49	0.34	0.40	486
accuracy			0.92	6552
macro avg	0.72	0.66	0.68	6552
weighted avg	0.91	0.92	0.92	6552

Dziewiątym wykorzystanym modelem do przewidywania subskrypcji była sztuczna sieć neuronowa; prosty Perceptron (wykorzystując algorytm służący do standaryzacji danych):

Wyniki Modelu Perceptron

✓ **Dokładność (accuracy):** 91.41%

✓ **Macierz klasyfikacji:**

	precision	recall	f1-score	support
0	0.93	0.98	0.95	6066
1	0.32	0.14	0.19	486
accuracy			0.91	6552
macro avg	0.63	0.56	0.57	6552
weighted avg	0.89	0.91	0.90	6552

WNIOSKI I PODSUMOWANIE

Podsumowując, przeprowadzono złożony proces analizy i procesowania danych w celu przygotowania ich do procesu modelowania. Skupiono się przede wszystkich na korelacji między zmiennymi, prostych miarach statystycznych, wykresach boxplot i histogramach. Z danych usunięto wartości odstające.

Na podstawie tak przygotowanych danych wykonano dziewięć różnych modelów uczenia maszynowego dedykowanych do problemu klasyfikacji. Przebadano, który model najlepiej przewiduje wartość zmiennej mówiącej o decyzji klienta (zakup subskrypcji bądź brak zakupu).

Moim zdaniem najlepszym modelem okazała się sieć neuronowa z 5 warstwami ukrytymi. Wartość dokładności okazała się wybitnie wysoka (93% rekordów zostało dobrze przypisanych do kategorii). Model ten kapitalnie radzi sobie z przewidywaniem kategorii 1 (wykup subskrypcji). Nieco gorzej przewiduje kategorię 0 (brak wykupu), choć statystyki wskazują, iż w tej kwestii jest on lepszy od reszty przebadanych modeli. Proponowałbym użycie tego właśnie modelu. Niewiele słabiej wypadły modele regresji logistycznej i lasu losowego (choć statystyki je opisujące również wskazują na wysoką dokładność, to modele te nie radzą sobie z przewidywaniem wartości dla kategorii 0).