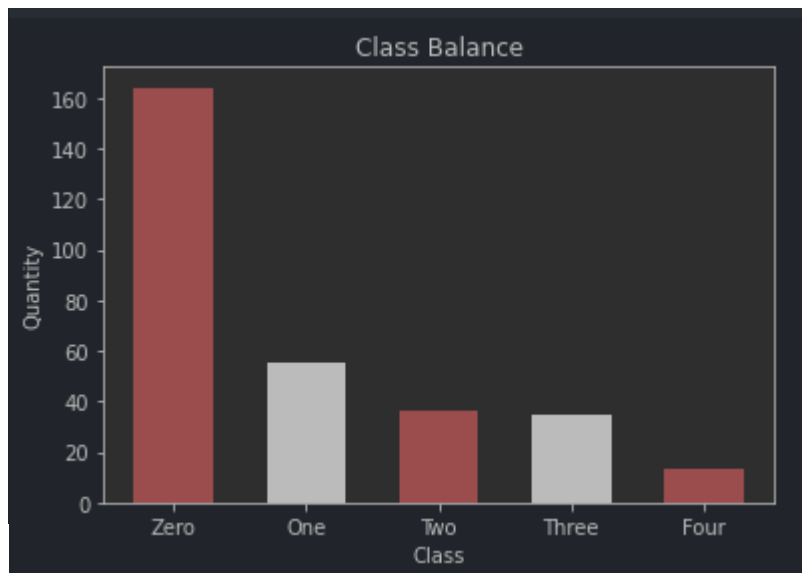


Raport 1

Krzysztof Maciejewski 260449

Analiza eksploracyjna

Czy zbiór jest zbalansowany pod względem liczby próbek na klasy?



1. Wykres zależności liczby próbek od klasy

Po przeprowadzeniu analizy liczby próbek należących do danej klasy, zauważyłem że najwięcej próbek przynależy do klasy 0, która mówi o jej braku. Pozostałe cztery klasy mają podobnie zbalansowane wartości.

Jakie są średnie i odchylenia cech liczbowych?

age	54.438944
trestbps	131.689769
chol	246.693069
thalach	149.607261
oldpeak	1.039604
ca	0.672241

Tabela przedstawia średnie wartości cech liczbowych

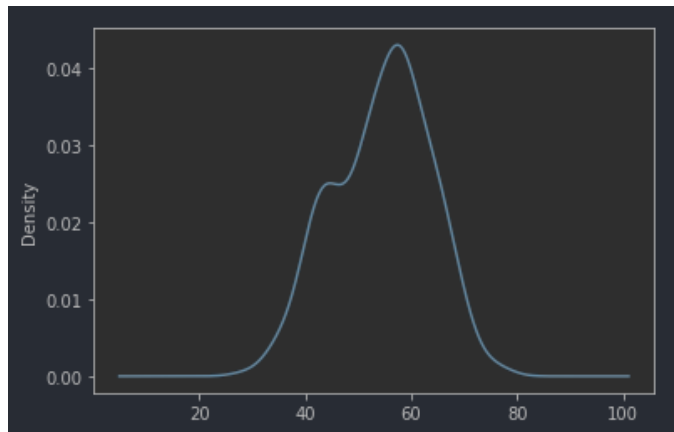
Średnie wartości cech liczbowych znacznie się od siebie różnią co pozwala stwierdzić, że w przyszłości trzeba będzie je znormalizować aby umożliwić wzajemne porównywanie i dalszą analizę.

age	9.038662
trestbps	17.599748
chol	51.776918
thalach	22.875003
oldpeak	1.161075
ca	0.937438

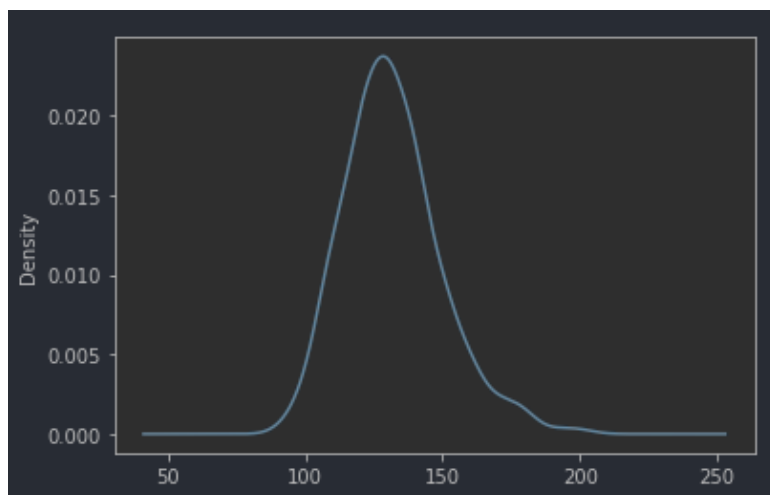
Tabela przedstawia odchylenia cech liczbowych

Wartości pozwalają stwierdzić, że takie dane jak np. wskaźnik cholesterolu mają bardzo zróżnicowane wartości i są mocno rozproszone.

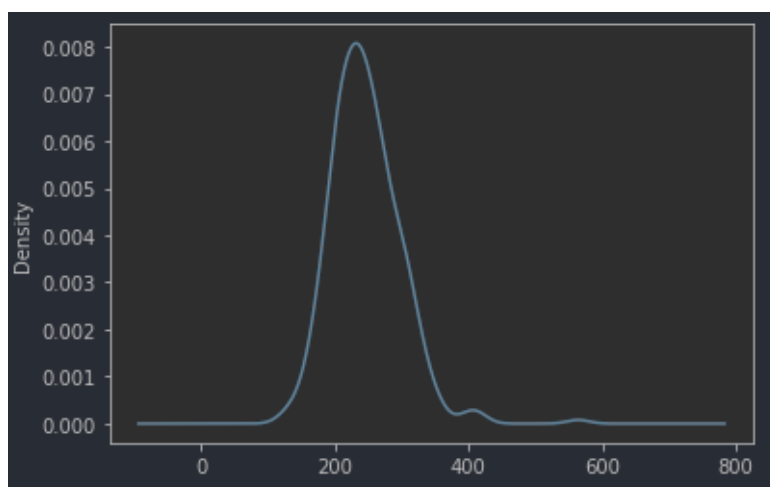
Dla cech liczbowych: czy ich rozkład jest w przybliżeniu normalny?



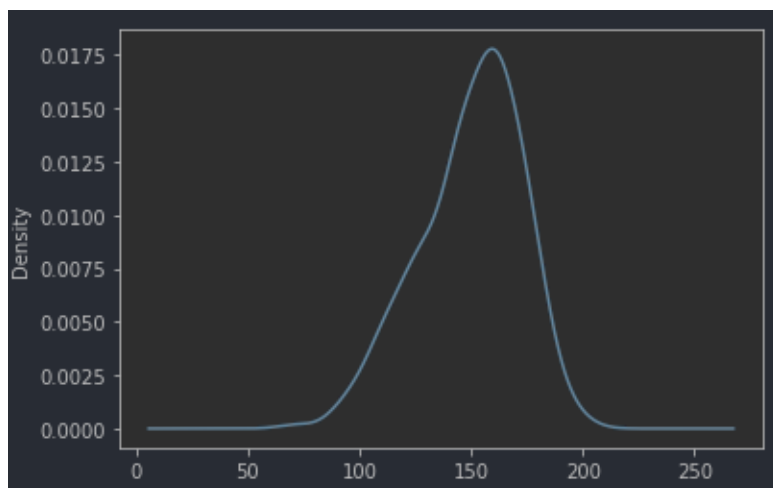
Rozkład wieku



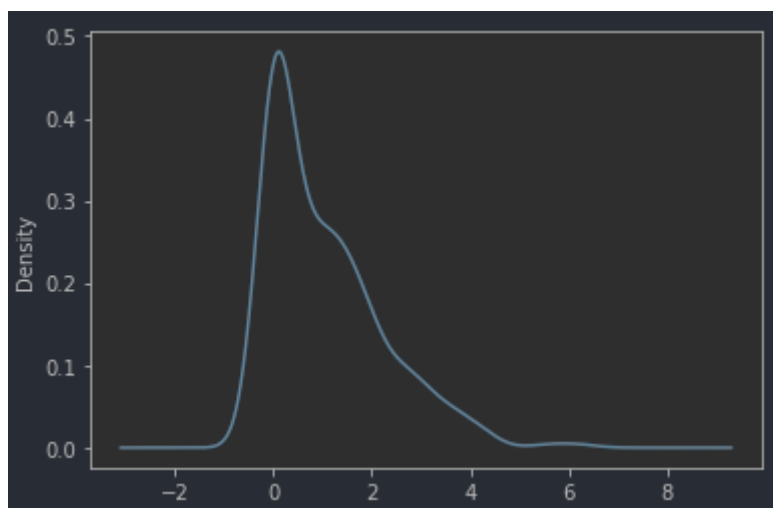
Rozkład resting blood pressure



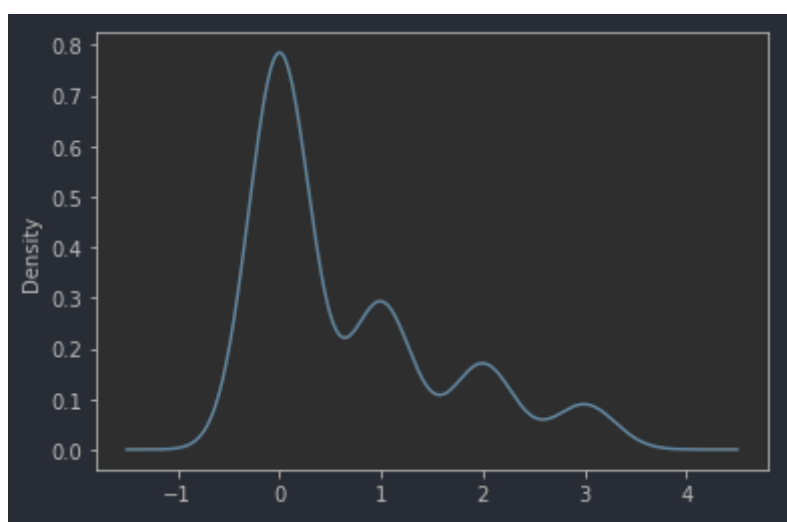
Rozkład cholesterolu



Rozkład maximum heart rate



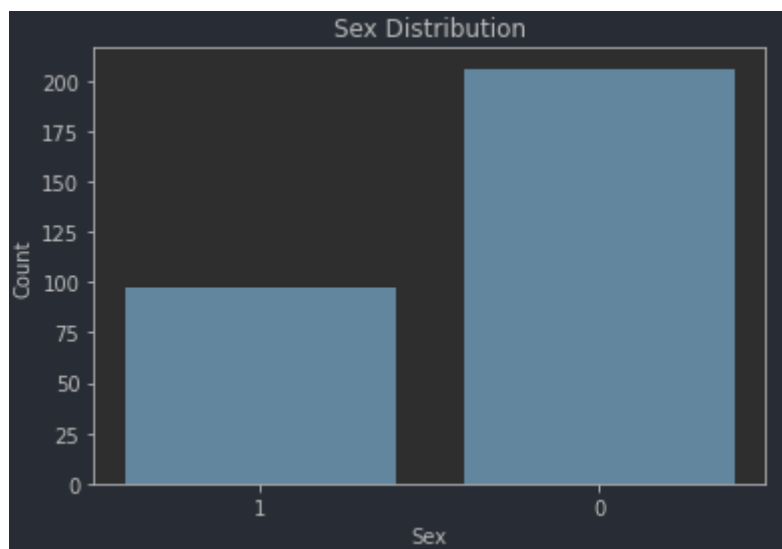
Rozkład depression induced by exercise



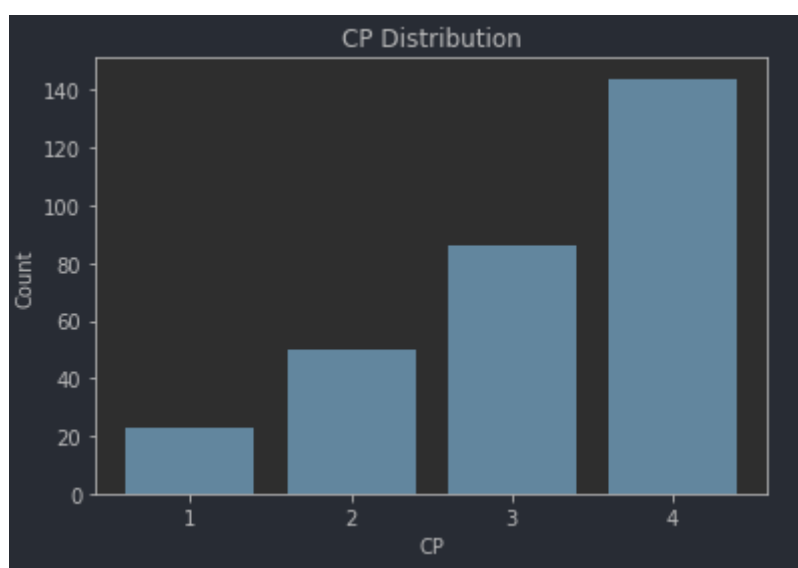
Rozkład cechy ca

Większość rozkładów dla cech liczbowych jest w przybliżeniu normalnych. Najbardziej od kształtu „dzwona” odbiegają dwa ostatnie wykresy. Wartości koncentrują się w nich wokół 0.

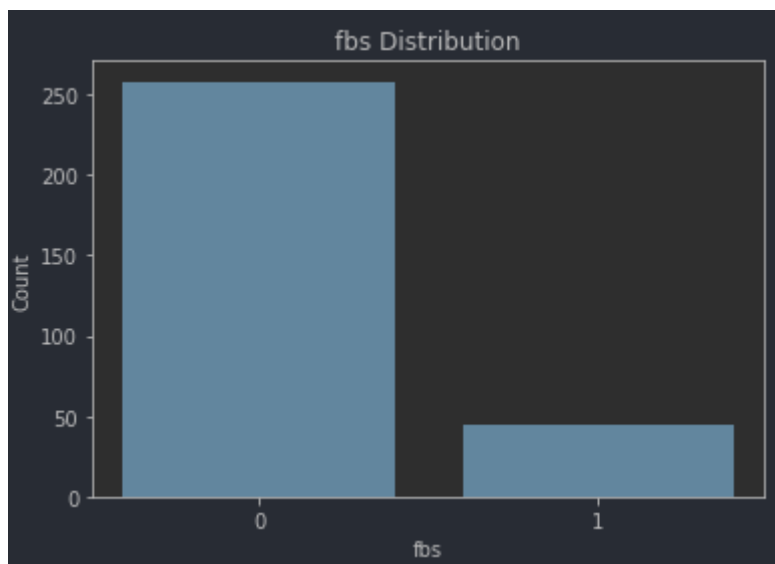
Dla cech kategorycznych: czy rozkład jest w przybliżeniu równomierny?



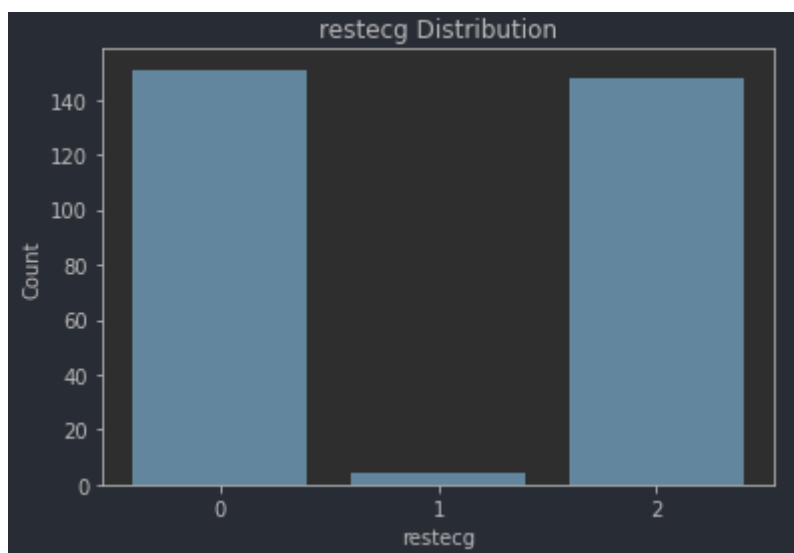
Rozkład cechy kategorycznej płeć



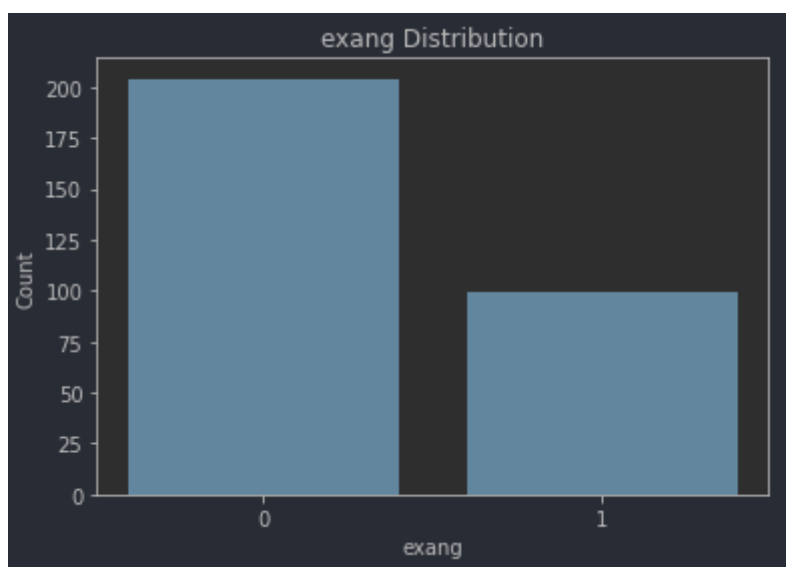
Rozkład cechy kategorycznej CP



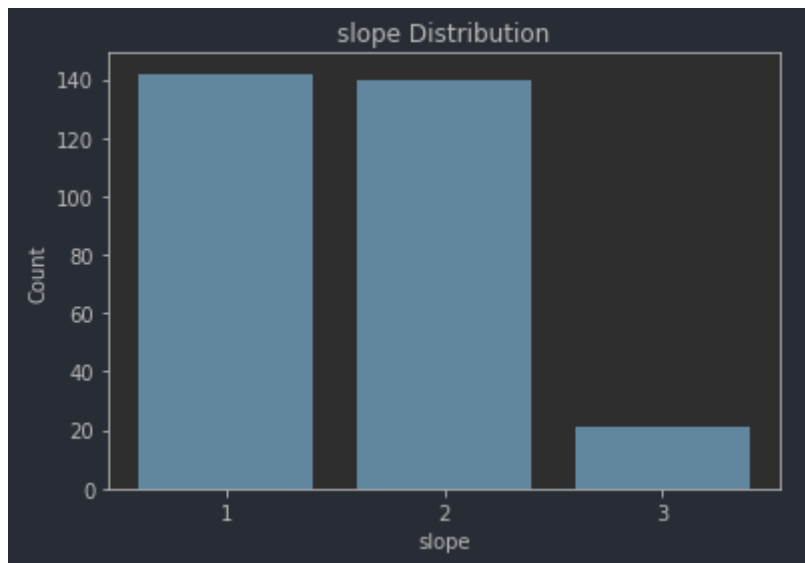
Rozkład cechy katerycznej fbs



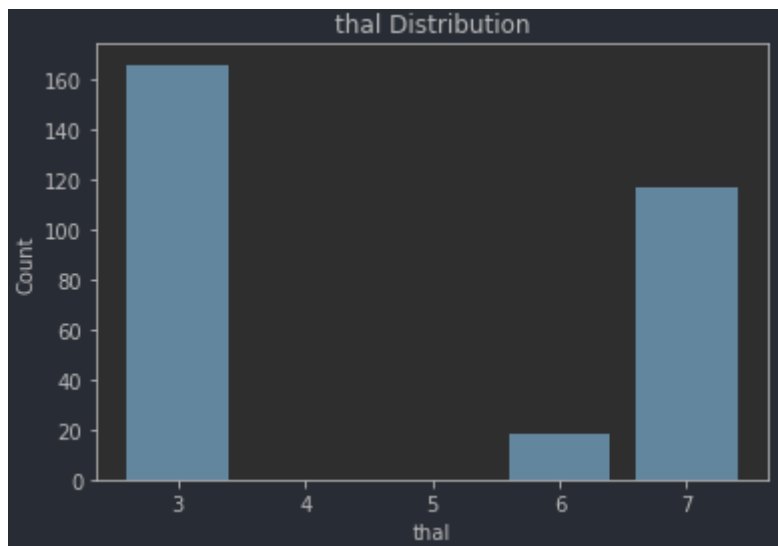
Rozkład cechy katerycznej restecg



Rozkład cechy katerycznej exang



Rozkład cechy slope



Rozkład cechy thal

Jak pokazują przedstawione rozkłady cech katerycznych, żaden z nich nie jest rozkładem jednostajnym. Wartości tych cech występują nierównomiernie.

Czy występują cechy brakujące i jaką strategię możemy zastosować żeby je zastąpić?

W danych następujące cechy mają brakujące wartości: ca (cecha liczbowa) oraz thal (cecha kateryczna). Cecha ca to liczba głównych naczyń zabarwionych za pomocą fluoroskopii. Po sprawdzeniu ile wartości jest brakujących, okazało się że dla ca są to 4 wartości, a dla thal 2 wartości. Stosunkowo jest to ilość, która nie jest duża więc nie warto byłoby w tych przypadkach odrzucać całą kolumnę. Ponieważ liczba wierszy z brakującymi wartościami jest mała, można by usunąć te wiersze.

Dla cechy liczbowej w przypadku rozkładu normalnego skuteczną strategią pozwalającą na nieodrzucać wierszy byłoby wstawienie w miejsce brakującej wartości średniej z danej kolumny. Jednakże jak wynika z wykresu rozkładu cechy ca jej rozkład nie przypomina rozkładu normalnego,

dlatego w tym przypadku lepsze okazałoby się wstawienie mediany która jest mniej wrażliwa na wartości odstające.

Inną strategią było by wypełnienie brakujących kolumn w wierszach losowo wybraną wartością spośród wierszy posiadających wszystkie wartości. Technika ta jest odpowiednia zarówno dla cech liczbowych jak i kategoriycznych.

Ostatnią strategią jest technika zwana Multiple Imputation, która pozwala na wyznaczenie brakujących wartości biorąc pod uwagę pozostałe wartości kolumn. Na początku zastępujemy brakujące wartości prostą strategią jak np. średnią danej kolumny. Brakujące wartości są wyznaczone na podstawie modelu regresji, w którym brakująca zmienna jest zmienną zależną, a pozostałe zmienne są zmiennymi niezależnymi. Następnie kolejna brakująca wartość jest używana jako zmienna zależna, a pozostałe jako niezależne. Proces trwa dopóki wszystkie brakujące wartości nie zostaną uwzględnione jako zmienne zależne. Po wyznaczeniu ich wartości początkowe tymczasowe wartości wyznaczone za pomocą prostej strategii są zastępowane przewidywaniami z modelu regresji. Proces zastępowania jest wykonywany kilka razy i wartości są aktualizowane po każdym z nich, aż do momentu gdy najlepiej odzwierciedlają relacje zidentyfikowane w danych.