

Projekt z przedmiotu Eksploracja Danych
Pierwszy etap: Zrozumienie problemu + Zrozumienie danych
Breast Cancer Wisconsin (Diagnostic) Data Set
Michał Sieczczyński

Ogólny opis zbioru

Zbiór danych zawiera charakterystykę pacjentów ze zdiagnozowanym nowotworem piersi. Pojedynczy wiersz odpowiada jednemu pacjentowi o unikalnym identyfikatorze. Dla każdego pacjenta pobrano komórki guza za pomocą biopsji, a następnie ich zdjęcia spod mikroskopu zostały przeanalizowane komputerowo. Poszczególne kolumny zawierają informacje czy zdiagnozowany nowotwór jest złośliwy czy łagodny, a także wartości numeryczne opisujące wygląd jądra komórkowego na zdjęciu spod mikroskopu. Kolumny podzielone są na trzy kategorie: “_mean” (z ang. średnia) - wartość uśredniona względem wszystkich jąder komórkowych, “_se” (z ang. standard error)- błąd standardowy wszystkich jąder komórkowych (błąd standardowy to odchylenie standardowe podzielone przez pierwiastek z liczby próbek) oraz “_worst” (z ang. najgorsze) - średnia trzech największych wartości.

Określenie celu eksploracji i kryteriów sukcesu

Celem eksploracji jest predykcja, który pacjent ma nowotwór złośliwy a który łagodny. Docelowo rozwiązywanym problemem będzie klasyfikacja binarna. Dodatkowym celem jest określenie które atrybuty mają największy wpływ na predykcję.

W przypadku danego problemu istotną metryką będzie czułość, która informuje o tym jak dobrze model klasyfikuje, czy nowotwór jest złośliwy. Nie chcielibyśmy, aby pacjent z nowotworem złośliwy otrzymał diagnozę, że nowotwór jest łagodny. Czułość jest metryką o poniższym wzorze:

$$CZUŁOŚĆ = \frac{TP}{TP + FN},$$

gdzie:

TP (z ang. true positive)- ilość przypadków prawdziwie pozytywnych

FN (z ang. false negative) - ilość przypadków fałszywie negatywnych

Natomiast wykorzystanie samej czułości może okazać się niewystarczające, ponieważ tracimy informację o tym, jak są klasyfikowane próbki negatywne - w tym przypadku pacjenci z nowotworem łagodnym. Dlatego należy zastosować dodatkową metrykę swoistości, która informuje jak dobrze klasyfikowane są próbki negatywne. Poniżej wzór danej metryki:

$$SWOISTOŚĆ = \frac{TN}{TN + FP},$$

gdzie:

TN (z ang. true negative) - ilość przypadków prawdziwie negatywnych

FP (z ang. false positive) - ilość przypadków fałszywie pozytywnych

Ewentualny błąd fałszywie pozytywny, czyli zaklasyfikowanie osoby z nowotworem łagodnym jako złośliwy potencjalnie wiązałoby się z dokonaniem dalszych badań, dlatego

ważniejszą metryką w przypadku tej analiza byłaby czułość. Sukces zostanie osiągnięty, jeżeli model uzyska czułość na poziomie 80% a swoistość na poziomie 60%.

Charakterystyka zbioru danych

Pochodzenie:

<https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data>

Format:

.csv

Liczba przykładów:

569

Ilość zbiorów danych:

1

Opis atrybutów

W ogólności w tym zbiorze danych każdy atrybut numeryczny (oprócz identyfikatora pacjenta) podzielony jest na 3 kategorie:

- “_mean”: średnia wartość względem wszystkich jąder komórkowych,
- “_se”: błąd standardowy wszystkich jąder komórkowych (odchylenie standardowe podzielone przez pierwiastek z liczby próbek),
- “_worst”: średnia z 3 największych wartości

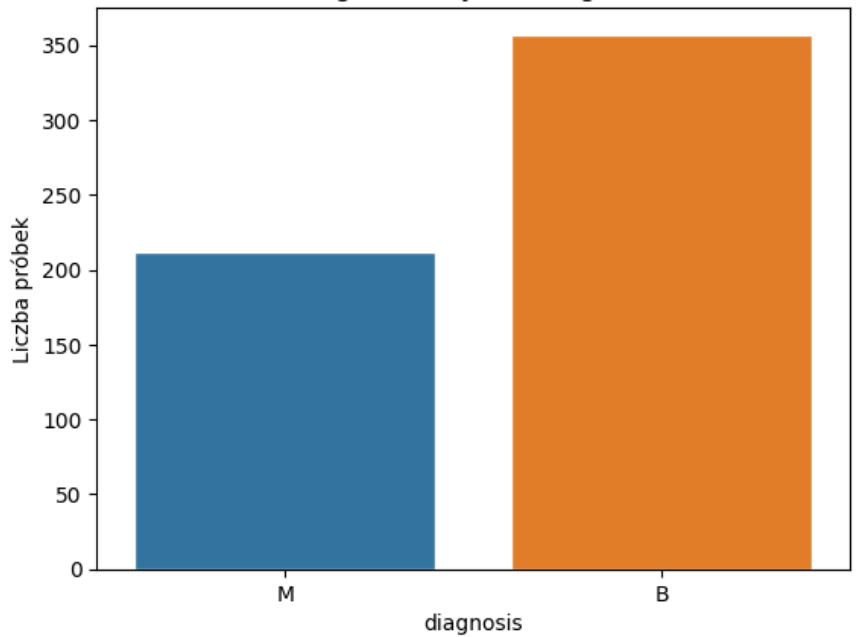
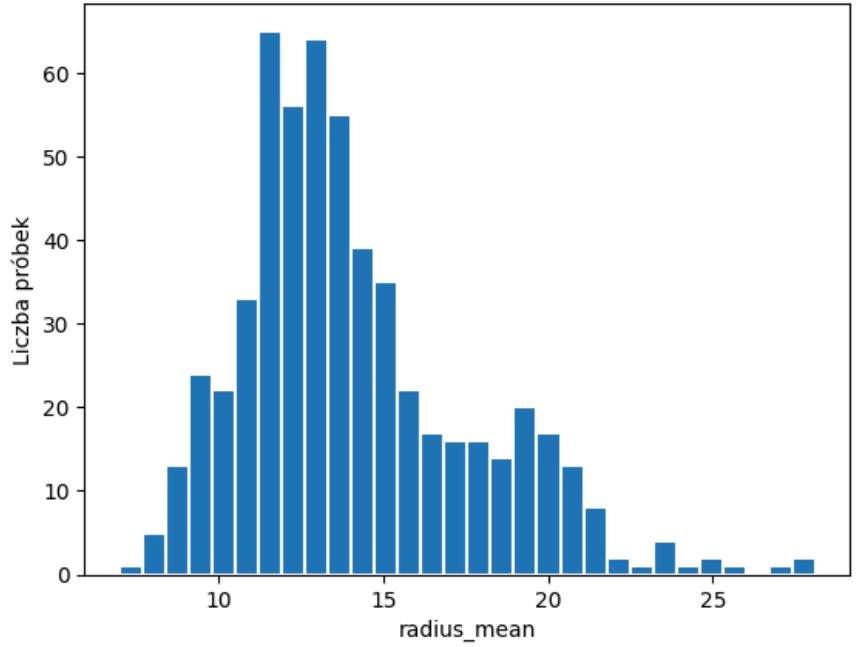
Poniżej zamieszczono opis poszczególnych atrybutów zarówno numerycznych jak i nominalnych. Atrybuty numeryczne opisane zostały w odniesieniu do pojedynczego jądra komórkowego - docelowe atrybuty odpowiadają średniej, błędowi standardowemu oraz średniej z 3 największych wartości. Jednostki danych atrybutów niestety nie zostały podane przez twórców zbioru danych - pochodzą one z analizy obrazów, więc możliwe, że nie odpowiadają fizycznym jednostkom i należy je uznać za względne.

Nazwa	Typ	Znaczenie
id	Numeryczny	Numer identyfikujący danego pacjenta
diagnosis	Nominalny	Atrybut informuje, czy nowotwór danego pacjenta jest złośliwy, czy łagodny. Nowotwór złośliwy często stanowi zagrożenie życia, ma zdolność do agresywnego wzrostu i może powodować przerzuty do innych narządów. Nowotwór łagodny zazwyczaj nie stanowi bezpośredniego zagrożenie życia, rozprzestrzeniają się powoli i nie powodują przerzutów do innych narządów. M (z ang. Malignant) oznacza pacjentów z nowotworem złośliwym. B (z ang. Benign) oznacza pacjentów z nowotworem łagodnym

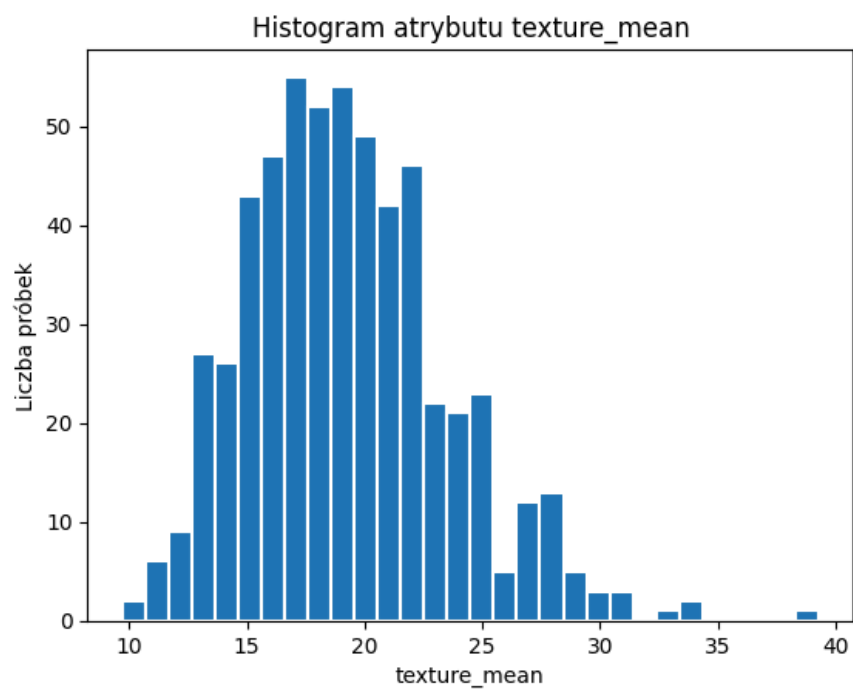
radius_mean/ radius_se/ radius_worst	Numeryczny	Promień jądra komórki - średni dystans od środka do punktów na brzegu danego jądra komórkowego.
texture_mean/ texture_se/ texture_worst	Numeryczny	Odchylenie standardowe wartości pikseli w odcieniach szarości jądra komórkowego
perimeter_mean/ perimeter_se/ perimeter_worst/	Numeryczny	Obwód jądra komórkowego
area_mean/ area_se/ area_worst	Numeryczny	Powierzchnia jądra komórkowego
smoothness_mean/ smoothness_se/ smoothness_worst	Numeryczny	Miara gładkości krawędzi jądra określana jako lokalna zmiana w długości promienia. Niska wartość oznacza, że kontur jest gładki, wysoka, że kontur jest poszarpany
compactness_mean/ compactness_se/ compactness_worst	Numeryczny	Miara zwartości jądra określana jako: $\frac{perimeter^2}{area} - 1$
concavity_mean/ concavity_se/ concavity_worst	Numeryczny	Miara wklęsłości krawędzi informująca o tym jak bardzo kształt jądra odchyła się do wewnątrz
concave points_mean/ concave points_se/ concave points_worst	Numeryczny	Liczba punktów wklęsłych na krawędzi jądra
symmetry_mean/ symmetry_se/ symmetry_worst	Numeryczny	Miara symetrii jądra - jak bardzo kształt różni się po przecięciu wzdłuż osi
fractal dimension_mean/ fractal dimension_se/ fractal dimension_worst	Numeryczny	Wymiar fraktalny kształtu jądra komórkowego. Jeżeli jądra jest gładkie wartość ta może być niska, natomiast jeśli kształt jest poszarpany i nieregularny - wartość może być wysoka.

Wyniki eksploracyjnej analizy danych

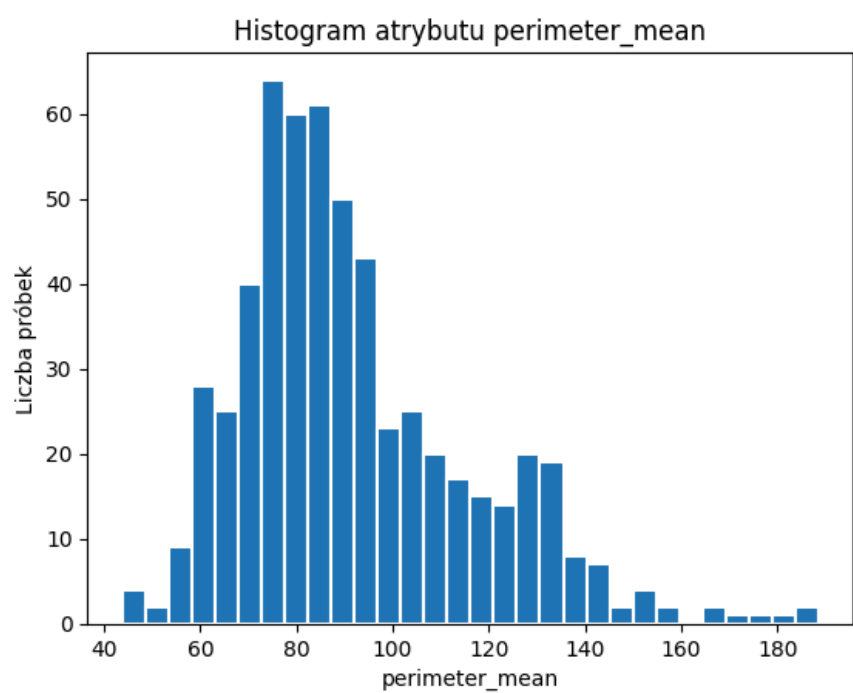
Rozkłady wartości atrybutów

Atrybut	Histogram																																																																																						
diagnosis	<p>Histogram atrybutu diagnosis</p>  <p>Liczba próbek</p> <p>diagnosis</p> <table border="1"><thead><tr><th>diagnosis</th><th>Liczba próbek</th></tr></thead><tbody><tr><td>M</td><td>210</td></tr><tr><td>B</td><td>355</td></tr></tbody></table>	diagnosis	Liczba próbek	M	210	B	355																																																																																
diagnosis	Liczba próbek																																																																																						
M	210																																																																																						
B	355																																																																																						
radius_mean	<p>Histogram atrybutu radius_mean</p>  <p>Liczba próbek</p> <p>radius_mean</p> <table border="1"><thead><tr><th>radius_mean</th><th>Liczba próbek</th></tr></thead><tbody><tr><td>7.5</td><td>1</td></tr><tr><td>8.0</td><td>5</td></tr><tr><td>8.5</td><td>13</td></tr><tr><td>9.0</td><td>24</td></tr><tr><td>9.5</td><td>22</td></tr><tr><td>10.0</td><td>33</td></tr><tr><td>10.5</td><td>65</td></tr><tr><td>11.0</td><td>56</td></tr><tr><td>11.5</td><td>64</td></tr><tr><td>12.0</td><td>55</td></tr><tr><td>12.5</td><td>39</td></tr><tr><td>13.0</td><td>35</td></tr><tr><td>13.5</td><td>22</td></tr><tr><td>14.0</td><td>17</td></tr><tr><td>14.5</td><td>16</td></tr><tr><td>15.0</td><td>16</td></tr><tr><td>15.5</td><td>14</td></tr><tr><td>16.0</td><td>20</td></tr><tr><td>16.5</td><td>17</td></tr><tr><td>17.0</td><td>13</td></tr><tr><td>17.5</td><td>8</td></tr><tr><td>18.0</td><td>2</td></tr><tr><td>18.5</td><td>1</td></tr><tr><td>19.0</td><td>4</td></tr><tr><td>19.5</td><td>1</td></tr><tr><td>20.0</td><td>2</td></tr><tr><td>20.5</td><td>1</td></tr><tr><td>21.0</td><td>1</td></tr><tr><td>21.5</td><td>1</td></tr><tr><td>22.0</td><td>2</td></tr><tr><td>22.5</td><td>1</td></tr><tr><td>23.0</td><td>1</td></tr><tr><td>23.5</td><td>1</td></tr><tr><td>24.0</td><td>1</td></tr><tr><td>24.5</td><td>1</td></tr><tr><td>25.0</td><td>1</td></tr><tr><td>25.5</td><td>1</td></tr><tr><td>26.0</td><td>1</td></tr><tr><td>26.5</td><td>1</td></tr><tr><td>27.0</td><td>1</td></tr><tr><td>27.5</td><td>1</td></tr><tr><td>28.0</td><td>1</td></tr></tbody></table>	radius_mean	Liczba próbek	7.5	1	8.0	5	8.5	13	9.0	24	9.5	22	10.0	33	10.5	65	11.0	56	11.5	64	12.0	55	12.5	39	13.0	35	13.5	22	14.0	17	14.5	16	15.0	16	15.5	14	16.0	20	16.5	17	17.0	13	17.5	8	18.0	2	18.5	1	19.0	4	19.5	1	20.0	2	20.5	1	21.0	1	21.5	1	22.0	2	22.5	1	23.0	1	23.5	1	24.0	1	24.5	1	25.0	1	25.5	1	26.0	1	26.5	1	27.0	1	27.5	1	28.0	1
radius_mean	Liczba próbek																																																																																						
7.5	1																																																																																						
8.0	5																																																																																						
8.5	13																																																																																						
9.0	24																																																																																						
9.5	22																																																																																						
10.0	33																																																																																						
10.5	65																																																																																						
11.0	56																																																																																						
11.5	64																																																																																						
12.0	55																																																																																						
12.5	39																																																																																						
13.0	35																																																																																						
13.5	22																																																																																						
14.0	17																																																																																						
14.5	16																																																																																						
15.0	16																																																																																						
15.5	14																																																																																						
16.0	20																																																																																						
16.5	17																																																																																						
17.0	13																																																																																						
17.5	8																																																																																						
18.0	2																																																																																						
18.5	1																																																																																						
19.0	4																																																																																						
19.5	1																																																																																						
20.0	2																																																																																						
20.5	1																																																																																						
21.0	1																																																																																						
21.5	1																																																																																						
22.0	2																																																																																						
22.5	1																																																																																						
23.0	1																																																																																						
23.5	1																																																																																						
24.0	1																																																																																						
24.5	1																																																																																						
25.0	1																																																																																						
25.5	1																																																																																						
26.0	1																																																																																						
26.5	1																																																																																						
27.0	1																																																																																						
27.5	1																																																																																						
28.0	1																																																																																						

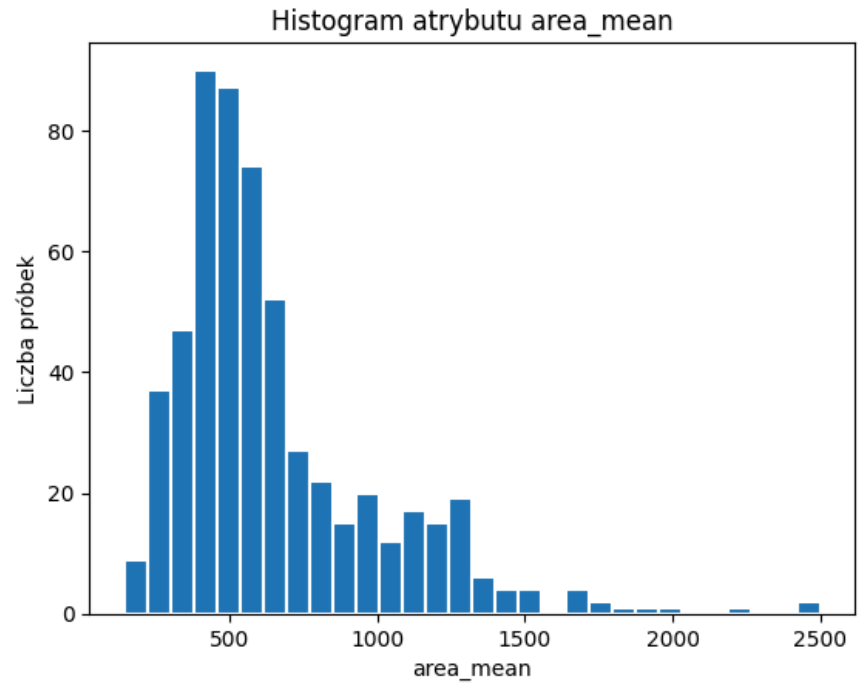
texture_mean



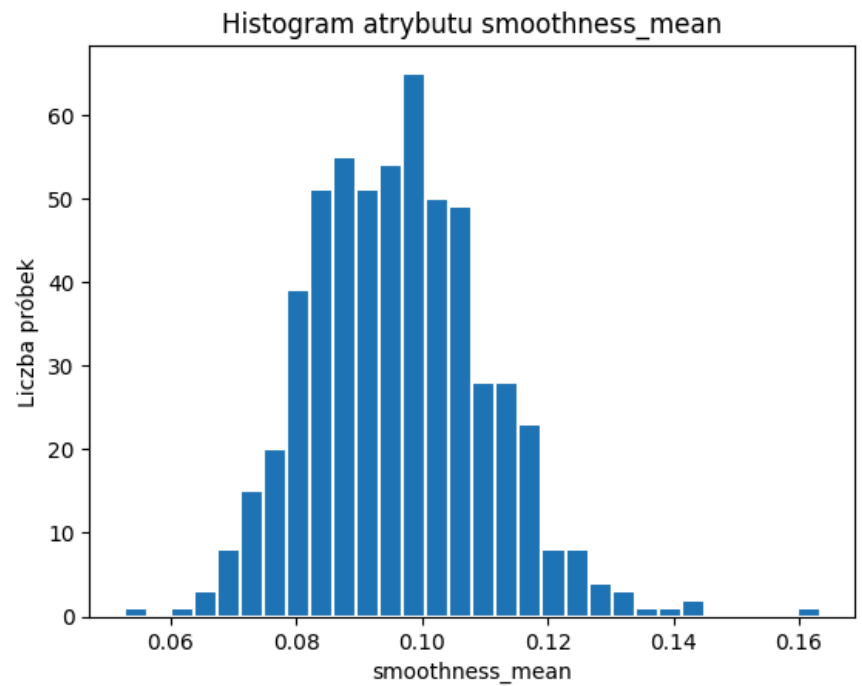
perimeter_mean



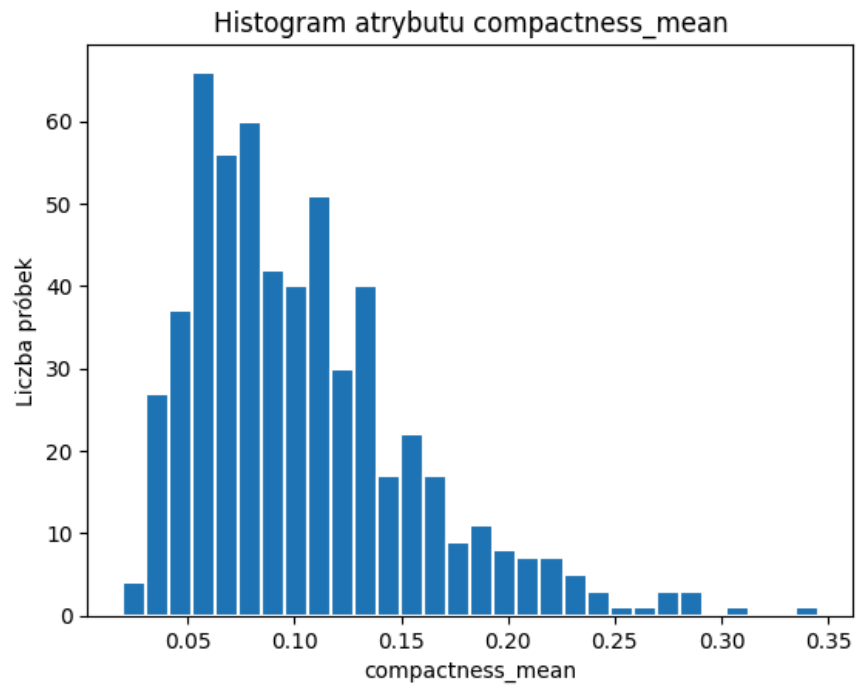
area_mean



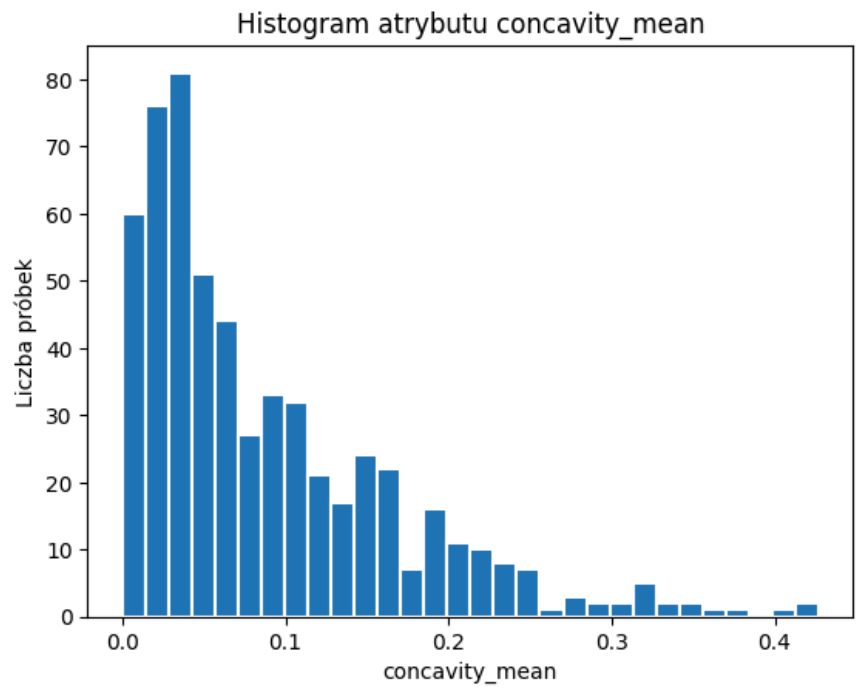
smoothness_mean



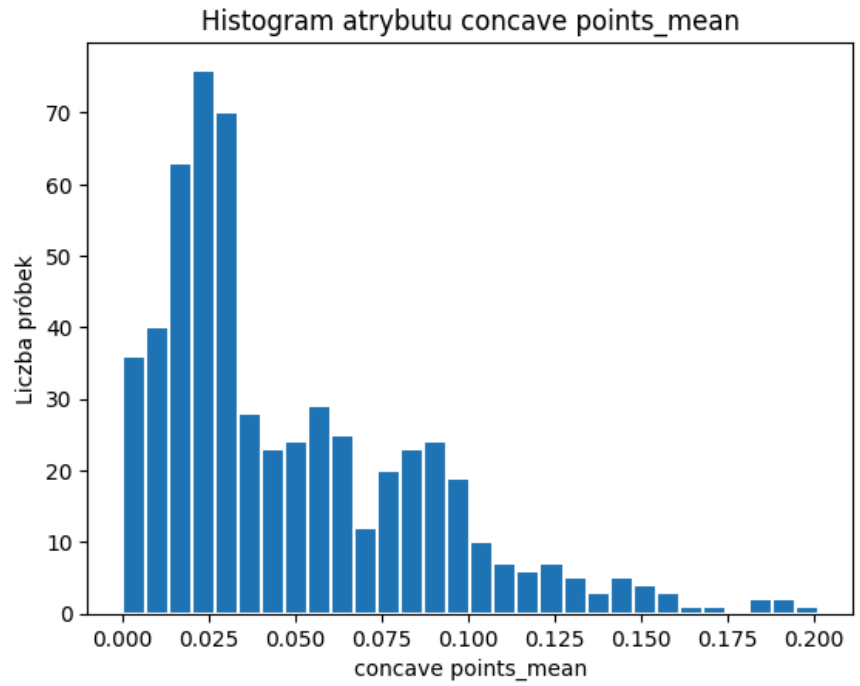
compactness_mean



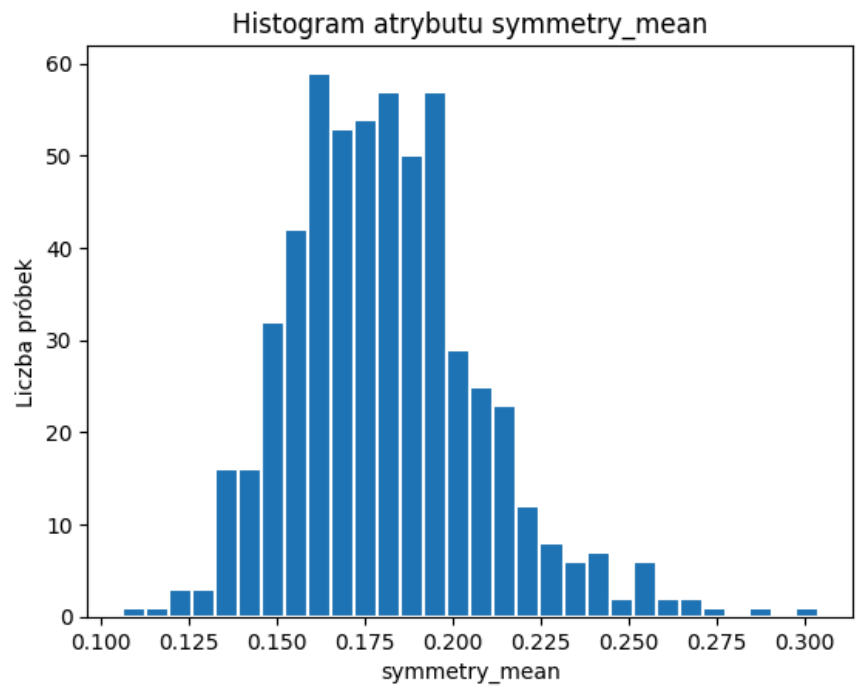
concavity_mean



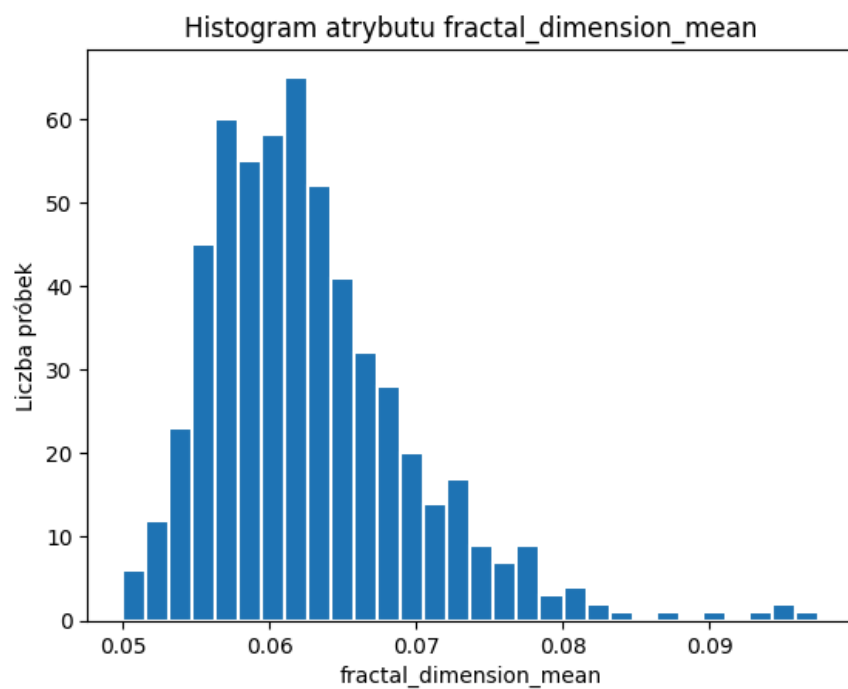
concave points



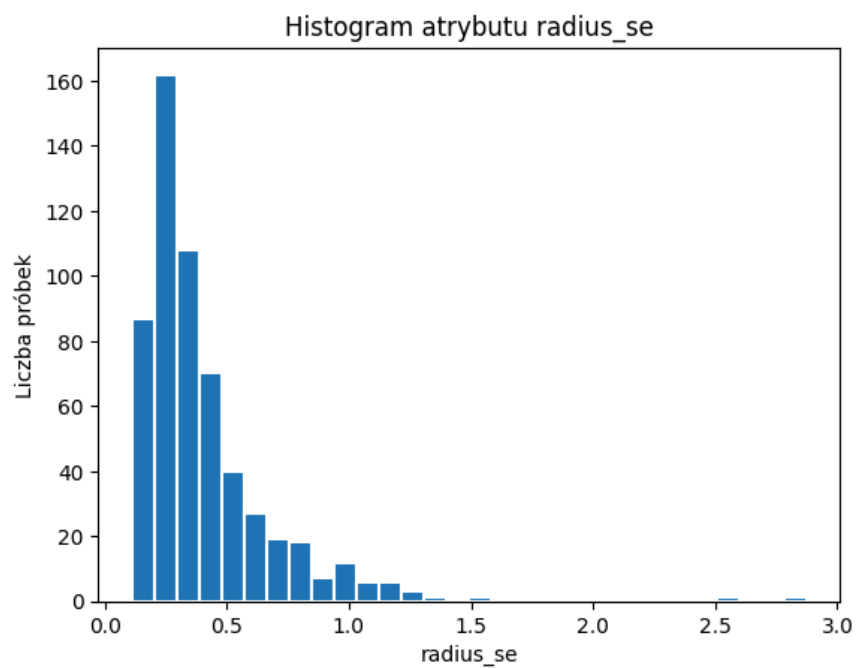
symmetry_mean



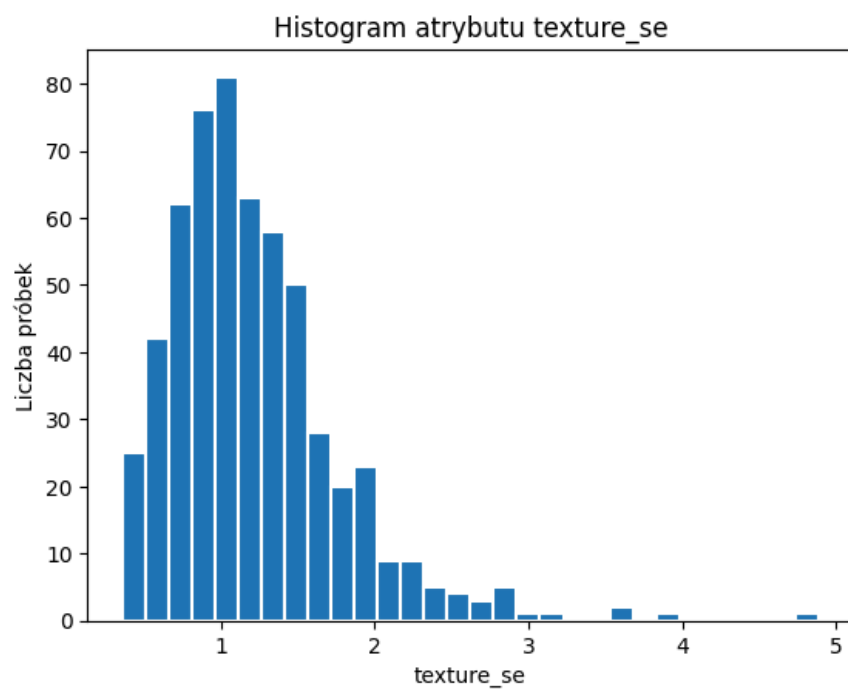
fractal_dimension_mean



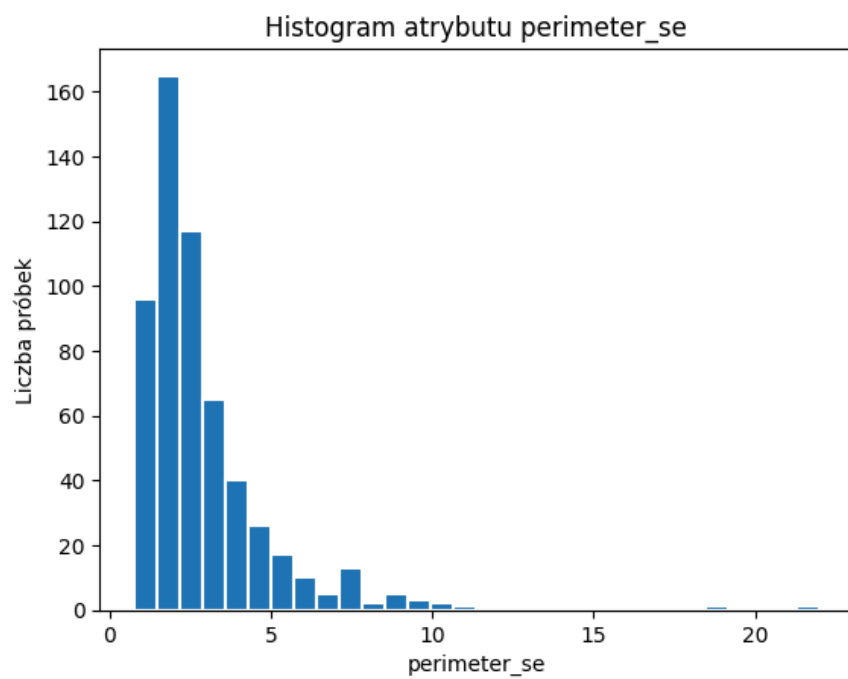
radius_se



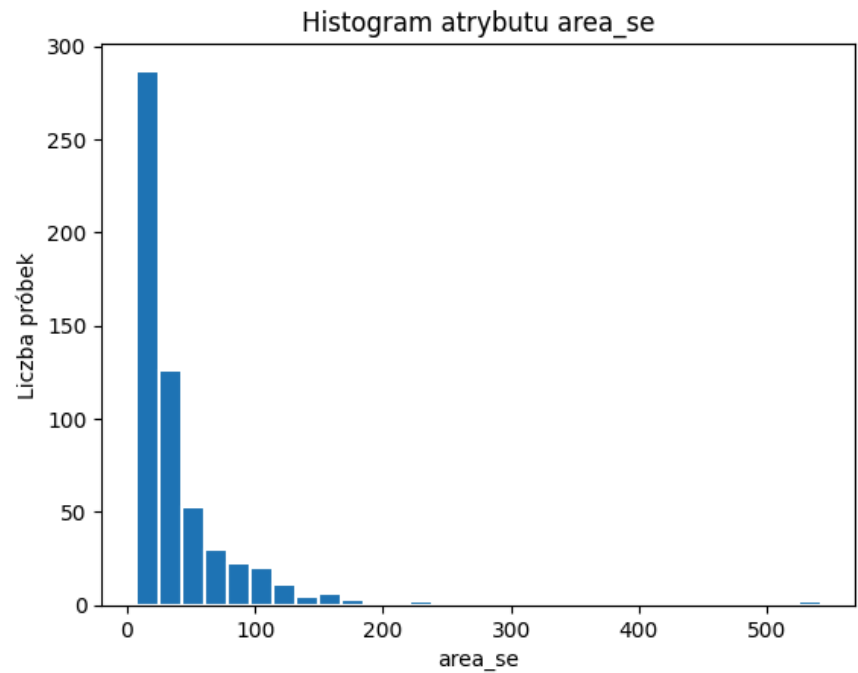
texture_se



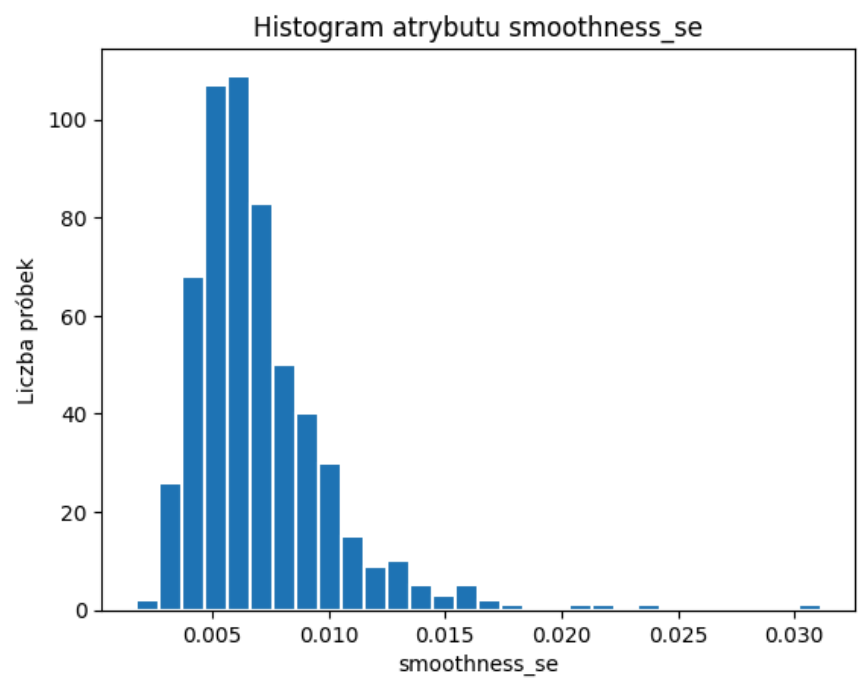
peremiter_se



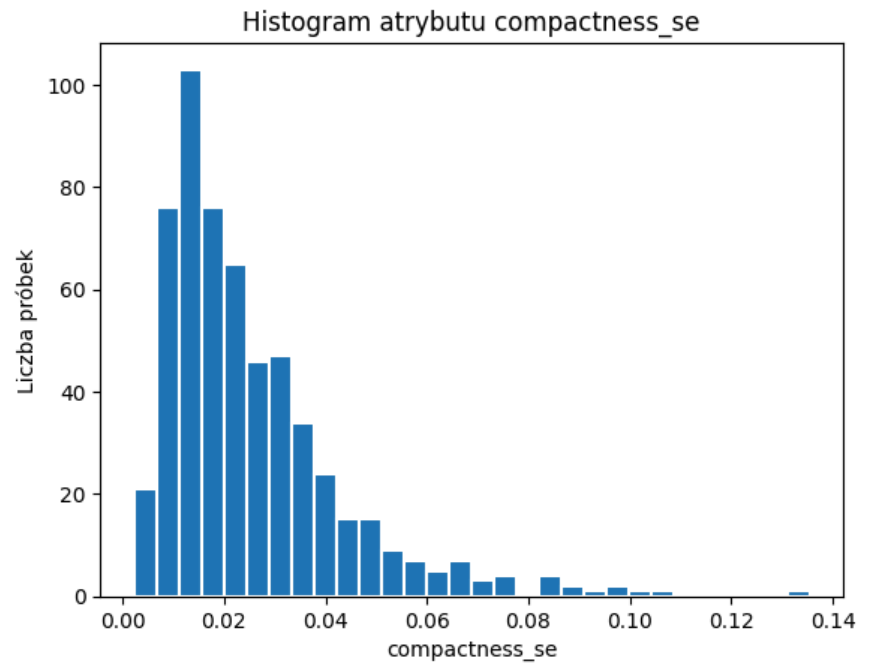
area_se



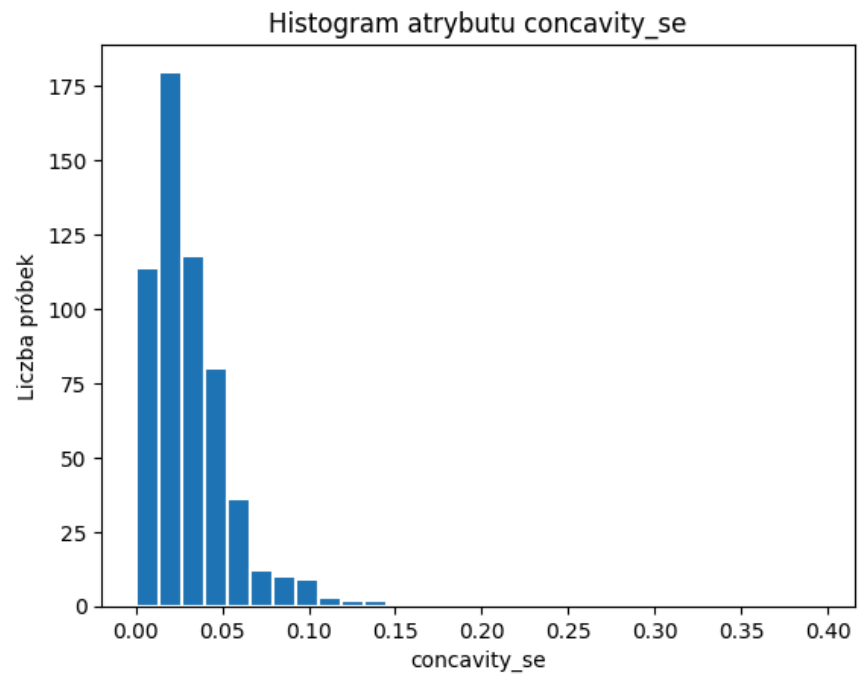
smoothness_se



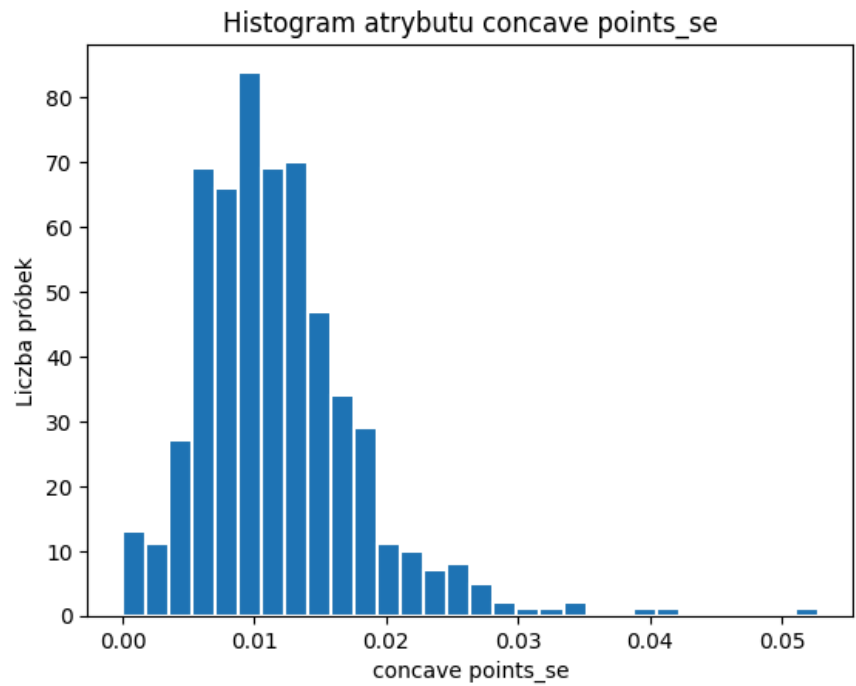
compactness_se



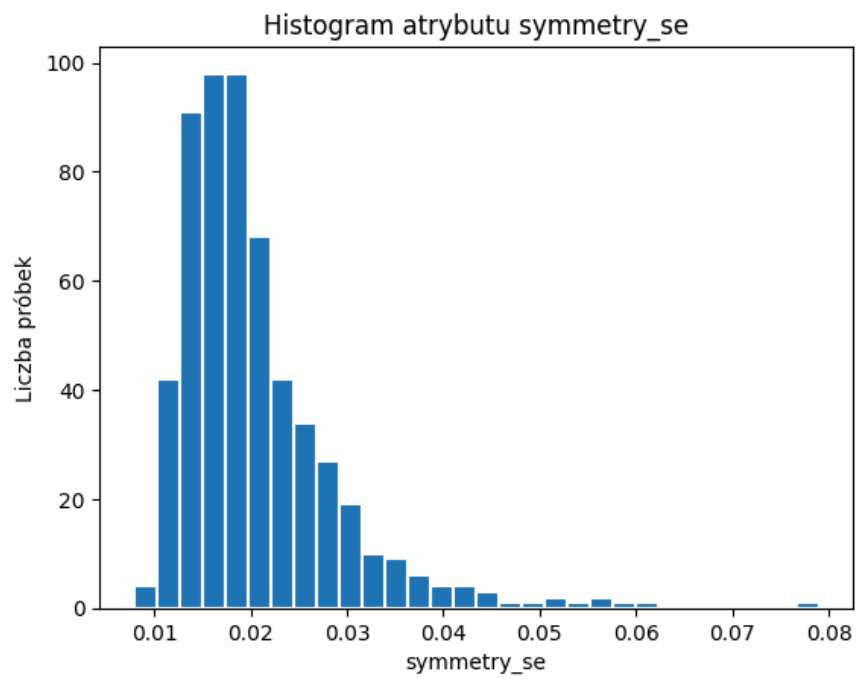
concavity_se



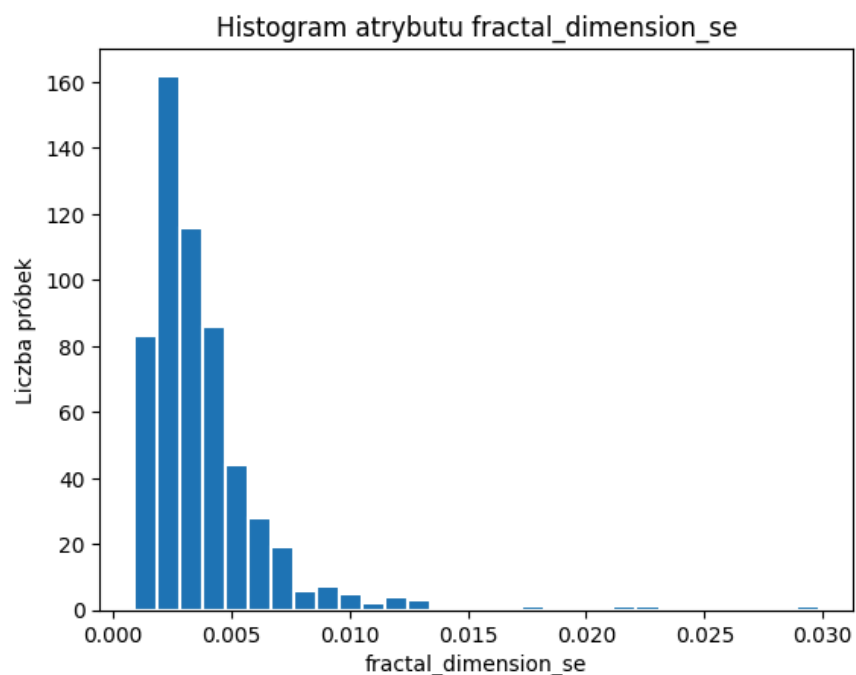
concave points_se



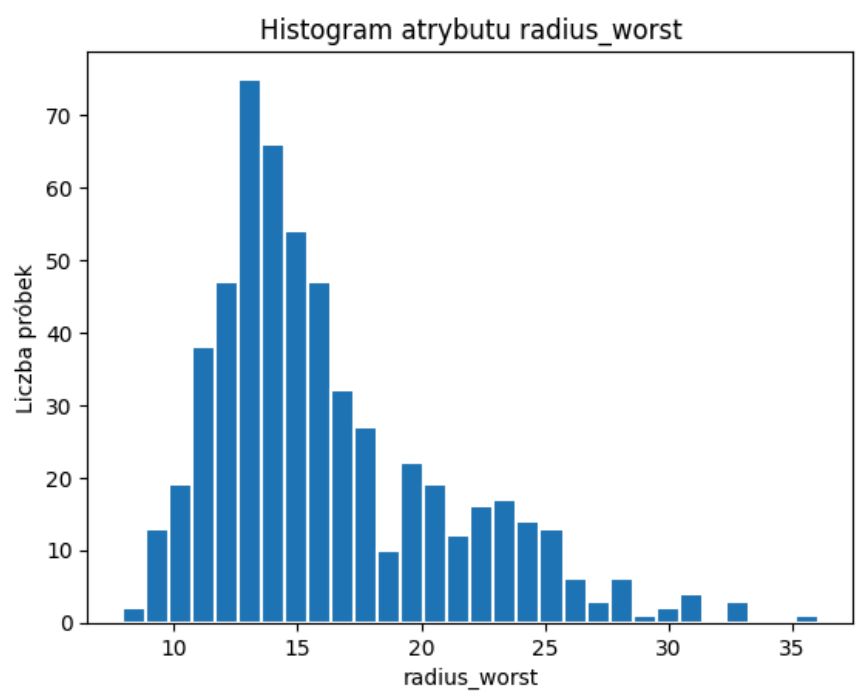
symmetry_se



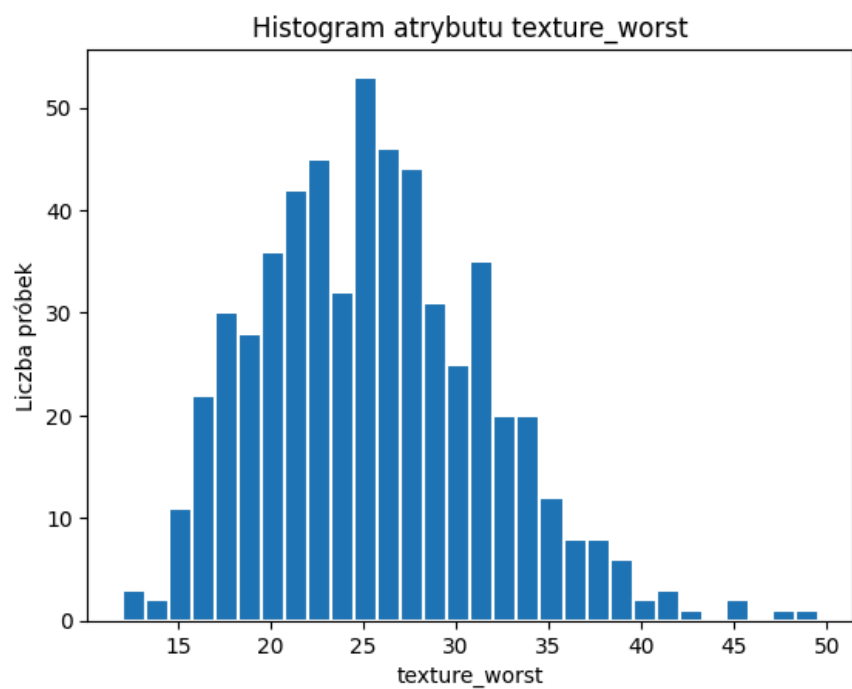
fractal_dimension_se



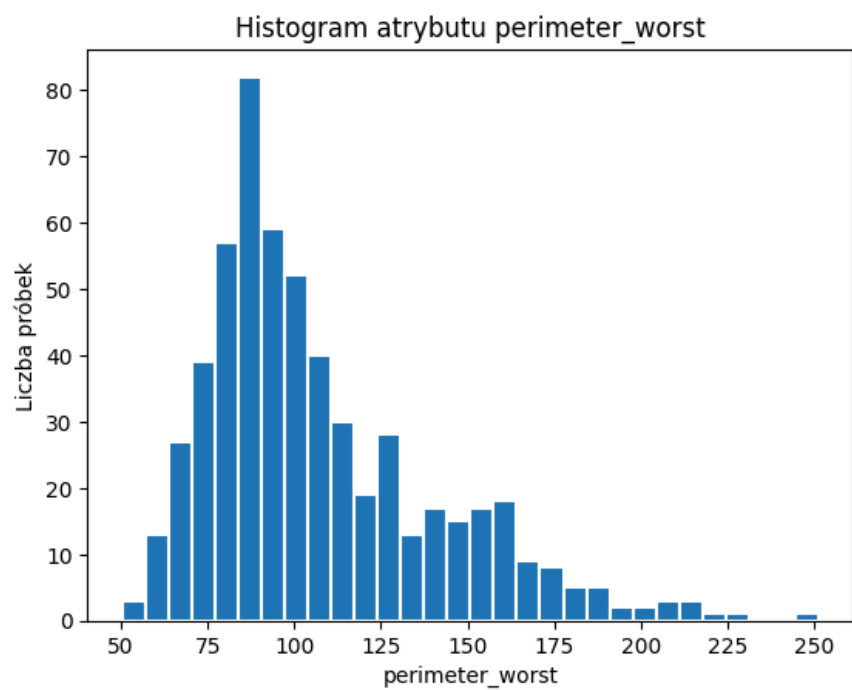
radius_worst



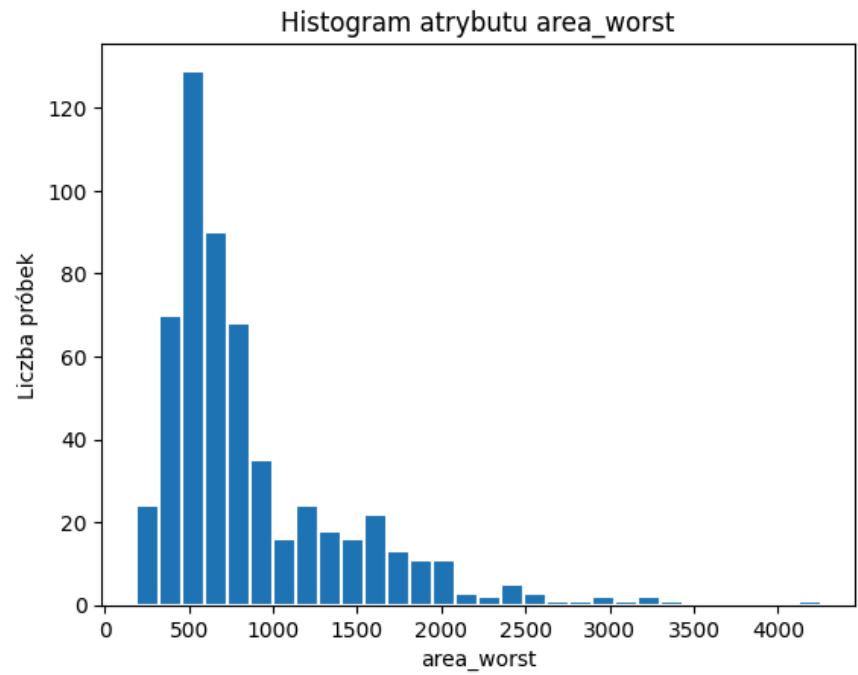
texture_worst



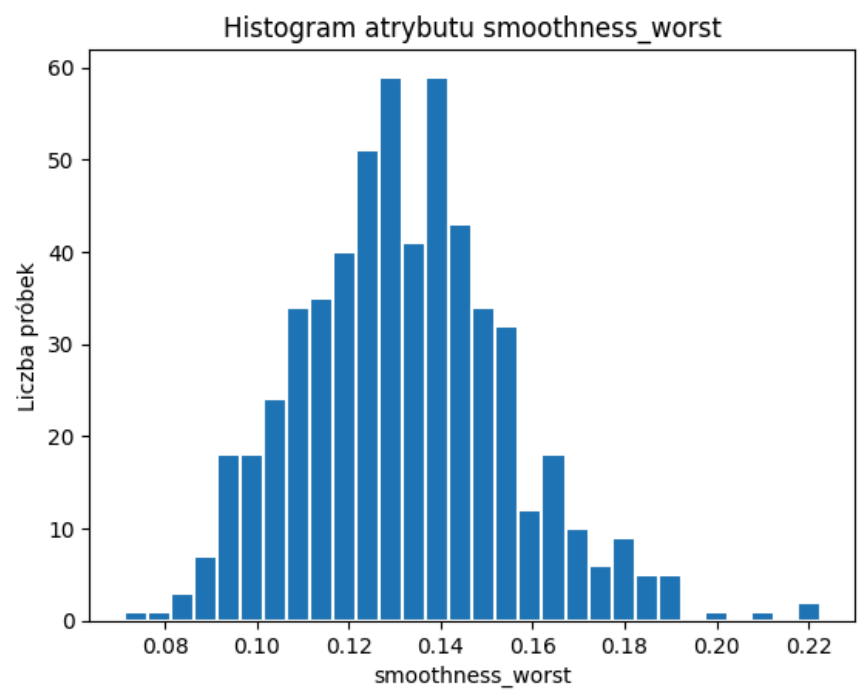
peremiter_worst



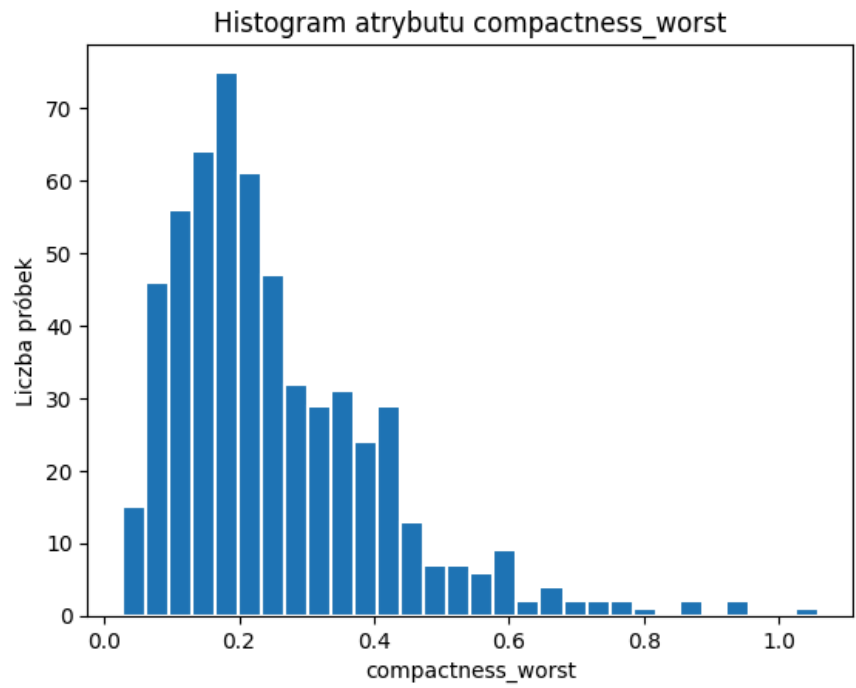
area_worst



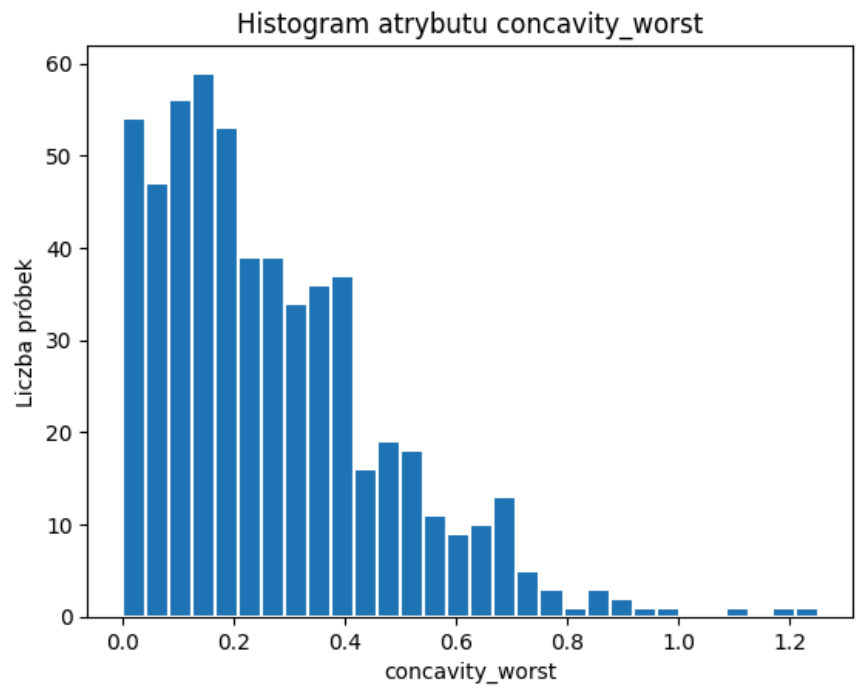
smoothness_worst



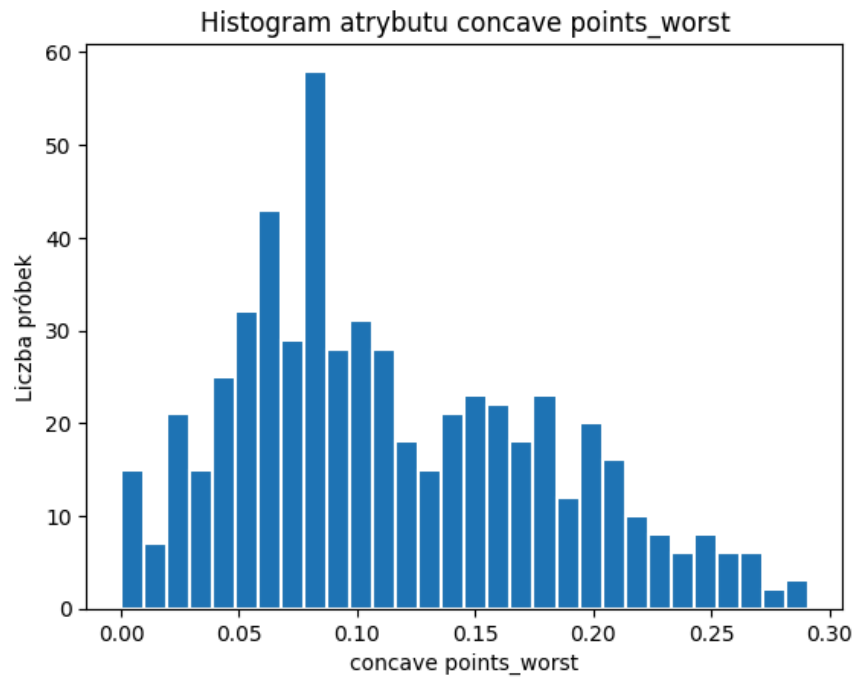
compactness_worst



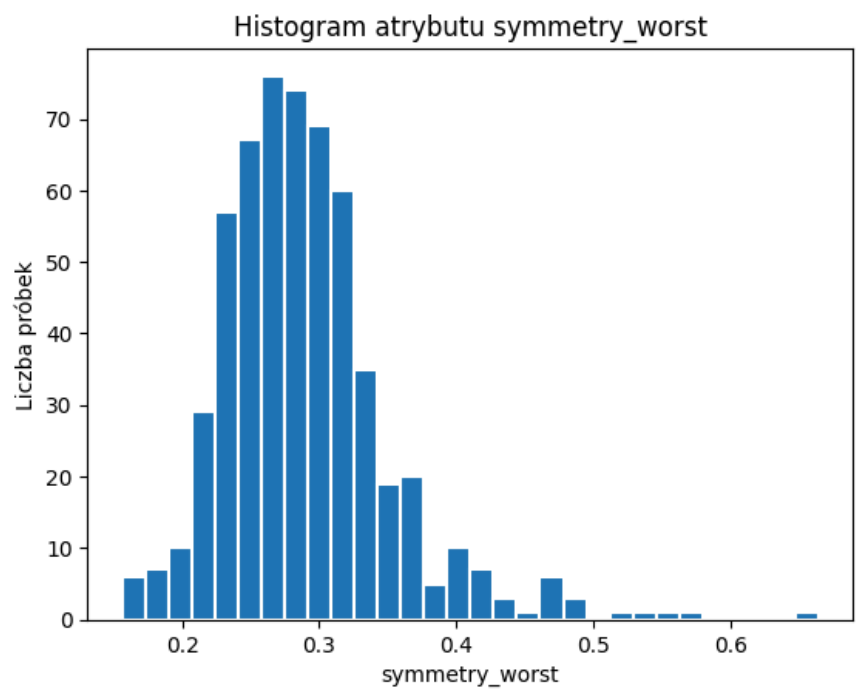
concavity_worst

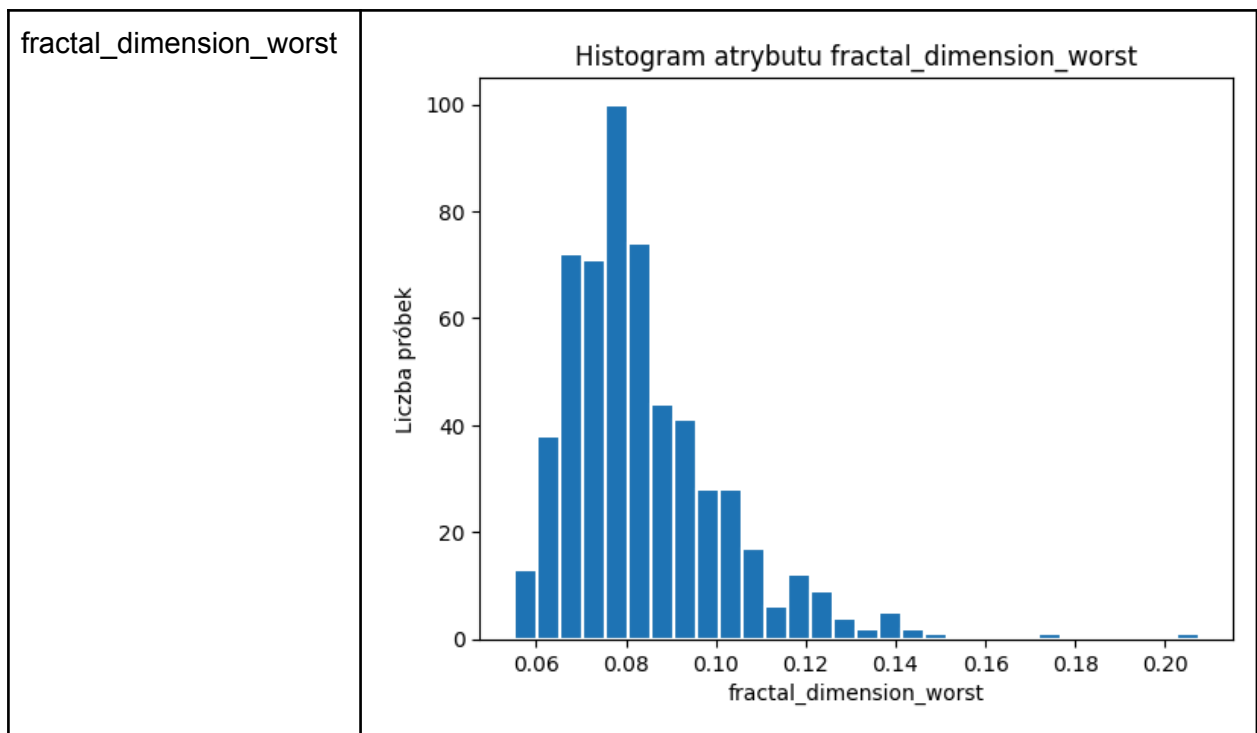


concave points_worst



symmetry_worst





Wnioski:

- W przypadku atrybutu diagnosis widać, że liczba pacjentów z łagodnym nowotworem jest większa od liczby pacjentów z nowotworem złośliwym. Klasy nie są zbalansowane, co może mieć wpływ na dalsze analizy.
- Rozkłady wszystkich analizowanych atrybutów numerycznych wydają się nie być normalne. Najczęściej mamy do czynienia z rozkładem prawoskośnym - rozkład jest przesunięty w lewo a dłuższy ogon znajduje się po prawej stronie wykresu. Atrybuty smoothness_mean oraz smoothness_worst mają rozkłady najbardziej zbliżone do normalnego.
- Komórki nowotworu złośliwego są najczęściej większe i bardziej nieregularne od zdrowych komórek, ponieważ mają uszkodzone mechanizmy związane z kontrolą wzrostu oraz podziałem komórki. W związku z czym wysokie, nieliczne wartości (ogon rozkładu) atrybutów _mean oraz _worst może wskazywać na pacjentów z nowotworem złośliwym.
- Wysokie, nieliczne wartości (ogon rozkładu) dla atrybutów _se może świadczyć o większej różnorodności populacji komórek, co również może być oznaką nowotworu złośliwego, ponieważ komórki dla tego typu nowotworu są najczęściej bardziej zróżnicowane - mają różne rozmiary, kształty czy stopień uszkodzenia DNA.

Sprawdzenie testem statystycznym - *dodatkowo*

Określenie, czy rozkład atrybutu jest normalny czy nie po prostu na niego patrząc najczęściej spełnia swoje zadanie. Jednakże w razie wątpliwości i w celu bardziej rzetelnego określenia normalności rozkładu można zastosować test statystyczny Shapiro-Wilka. Hipotezą zerową tego testu brzmi: dana próba pochodzi z populacji o rozkładzie normalnym. Po obliczeniu wartości p należy ją porównać z przyjmowanym progiem (najczęściej 0,05). Jeżeli wartość p jest większa od 0,05 oznacza to, że dane pasują do hipotezy zerowej - możemy uznać rozkład danego atrybutu za normalny. Natomiast, jeżeli wartość p jest

mniejsza od 0,05 to należy odrzucić hipotezę zerową - uznajemy, że rozkład atrybutu nie jest normalny. Wartość p można rozumieć jako prawdopodobieństwo uzyskania takiego wyniku, przy założeniu, że hipoteza zerowa jest prawdziwa. Na przykład: jeśli p jest małe (np. mniejsze niż 0,05), to istnieje małe prawdopodobieństwo, że takie dane zostałyby zaobserwowane, jeśli rozkład byłby normalny – co może sugerować, że rozkład nie jest normalny.

Atrybut	Wartość p dla testu Shapiro-Wilka
radius_mean	3,11e-14
texture_mean	7,28e-08
perimeter_mean	7,01e-15
area_mean	3,20e-22
smoothness_mean	8,60e-05
compactness_mean	3,97e-17
concavity_mean	1,34e-21
concave points_mean	1,40e-19
symmetry_mean	7,88e-09
fractal_dimension_mean	1,96e-16
radius_se	1,22e-28
texture_se	3,56e-19
perimeter_se	7,59e-30
area_se	2,65e-35
smoothness_se	1,36e-23
compactness_se	1,08e-23
concavity_se	1,10e-31
concave points_se	7,83e-17
symmetry_se	3,13e-24
fractal_dimension_se	8,55e-31
radius_worst	1,70e-17
texture_worst	2,57e-06
perimeter_worst	1,37e-17
area_worst	5,60e-25
smoothness_worst	2,10e-04
compactness_worst	1,25e-19
concavity_worst	4,54e-17
concave points_worst	1,99e-10
symmetry_worst	3,23e-17

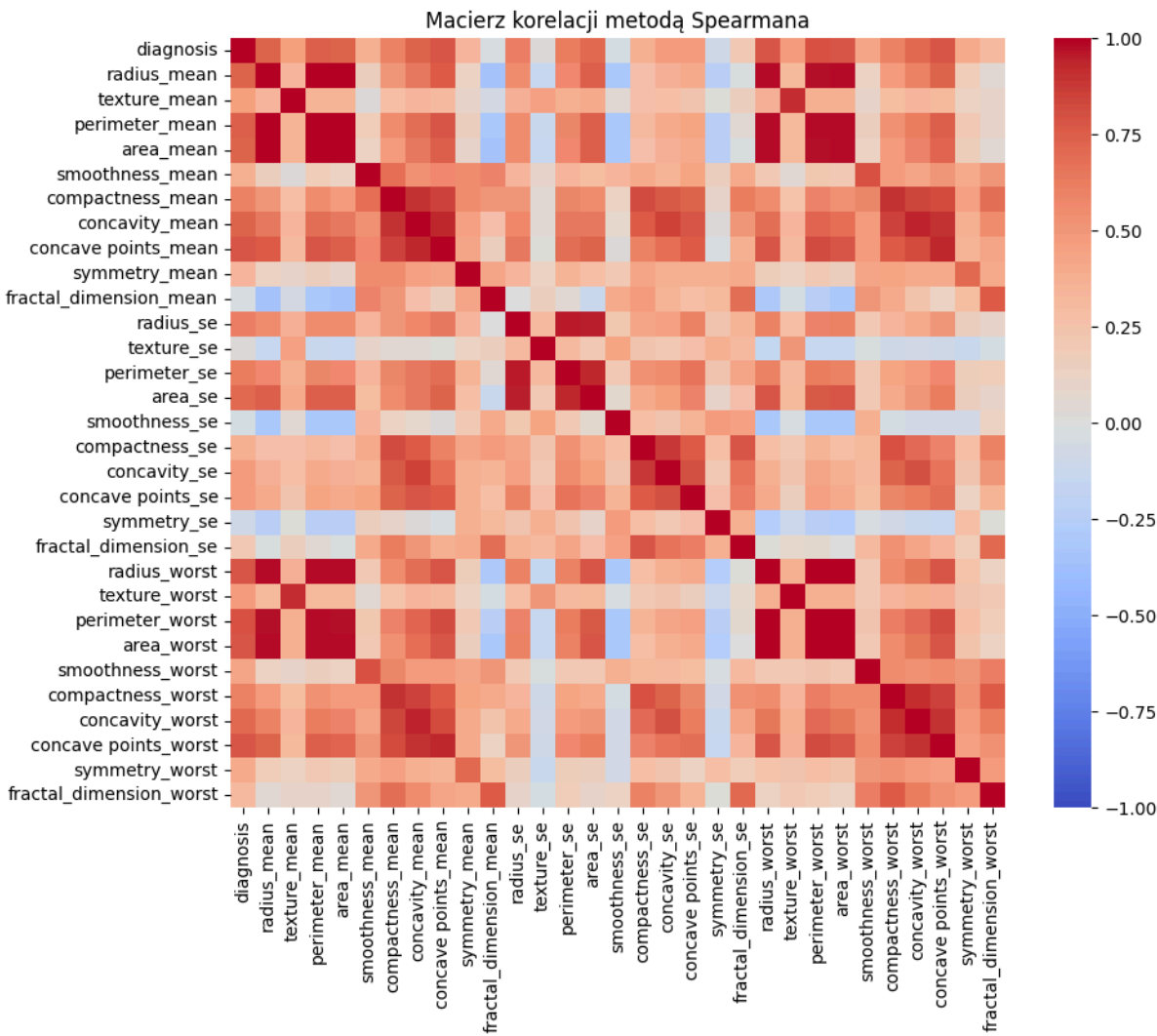
fractal_dimension_worst	9,20e-20
-------------------------	----------

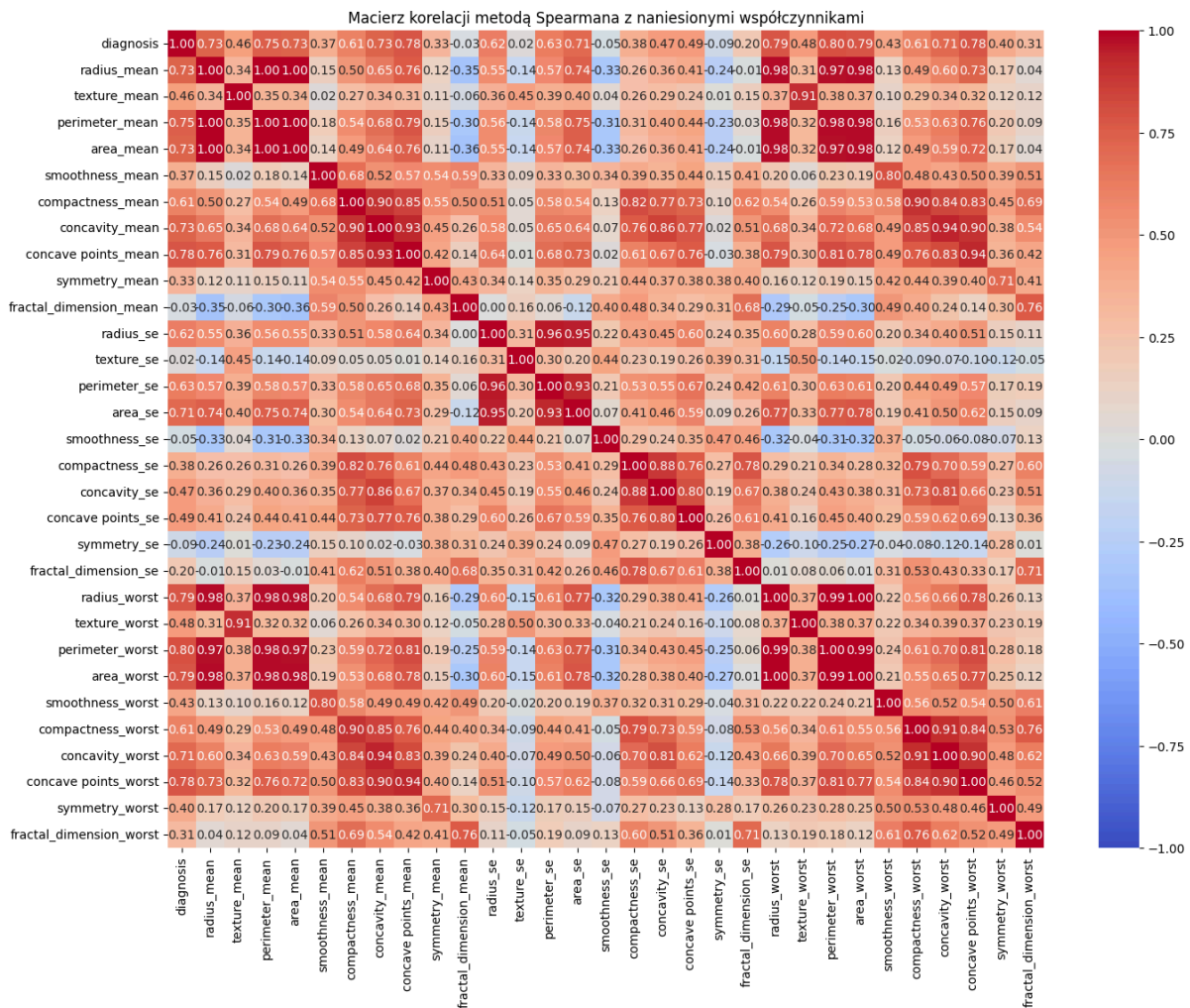
Wnioski

Na podstawie powyższej tabeli możemy powiedzieć z dużą pewnością, że wszystkie atrybuty numeryczne nie mają rozkładu normalnego. Dodatkowo widać, że atrybuty smoothness_mean oraz smoothness_worst mają jedne z najwyższych wartości p, lecz wciąż mniejsze od 0,05.

Korelacje pomiędzy wartościami atrybutów

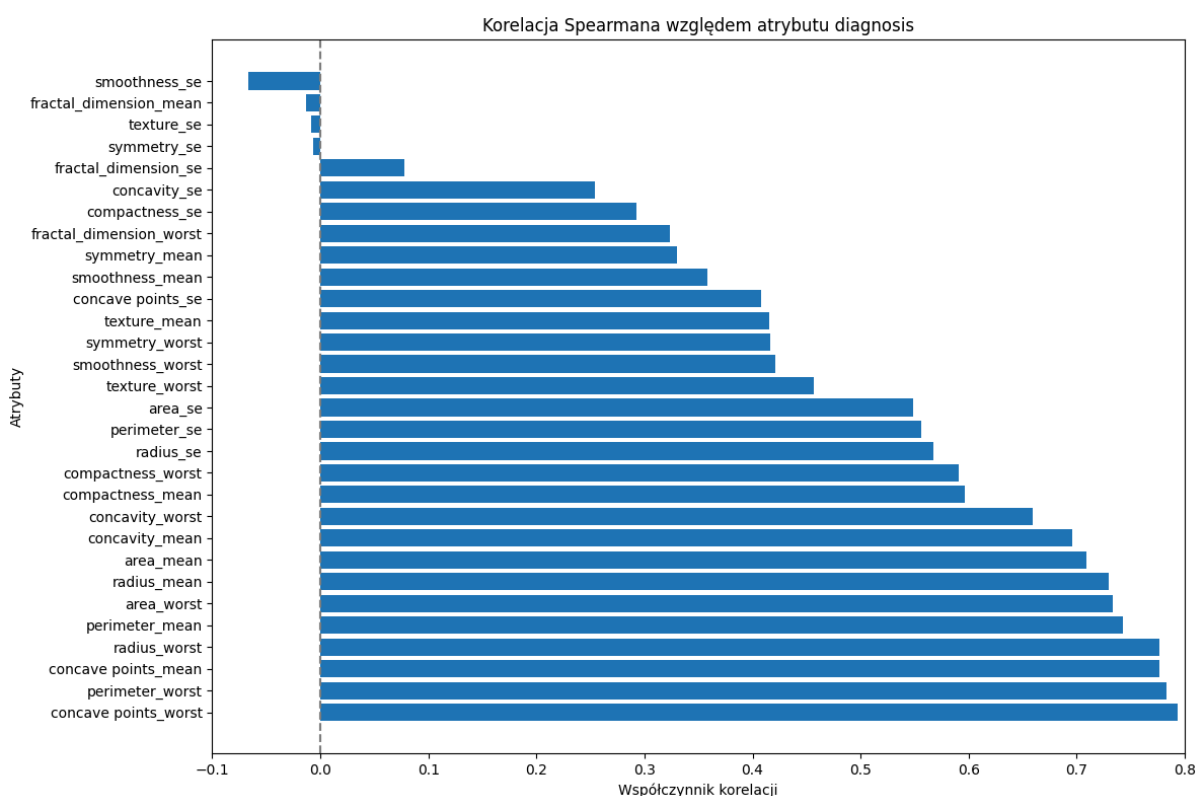
W celu obliczenia korelacji z atrybutem diagnosis zamieniłem klasę M (nowotwór złośliwy) na 1 i klasę B (nowotwór łagodny) na 0. W związku z faktem, że żaden atrybut nie miał rozkładu normalnego stosuję korelację Spearmana.





sens. Ciekawe jest również to, że kategorie _se tych atrybutów są skorelowane z kategoriami _mean oraz _worsť, z czego wynika, że wraz ze zwiększaniem się nieregularności jąder komórek w bardzo dużym stopniu zwiększa się ich różnorodność nieregularności. W tym przypadku również korelacja dodatnia również może być wytłumaczona poprzez pojawianie się większej ilości ekstremalnych przypadków.

Korelacja z atrybutem diagnosis



Wnioski:

W odniesieniu do korelacji poszczególnych atrybutów do atrybutu celu - diagnosis, można zauważyć, że cechy związane z występowaniem wklęsłości (concave points_worst, concave points_mean, concavity_worst, concavity_mean), a także związane z rozmiarem jądra komórki (perimeter_worst, perimeter_mean, radius_worst, radius_mean, area_worst, area_mean) cechują się silną korelacją. Dodatkowo atrybuty compactness_mean a także compactness_worst jak również area_se, perimeter_se oraz radius_se również mają silną korelację z atrybutem diagnosis. Oznacza to, że wysokie wartości tych atrybutów są częściej obserwowane dla nowotworów złośliwych niż dla nowotworów łagodnych. Pozostałe atrybuty związane z teksturą czy regularnością jąder komórkowych a także reszta atrybutów w kategorii _se mają mniejszy związek z rodzajem nowotworu.

Uwagi na temat jakości danych

Dane brakujące:
brak

Dane niespójne:

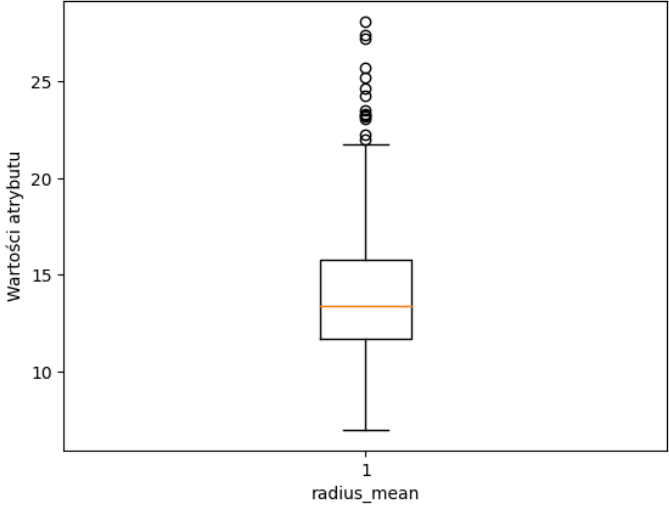
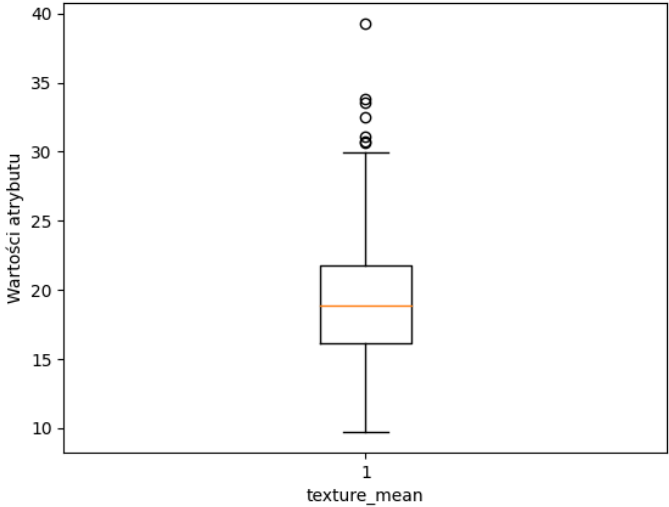
Ze względu na to, że dane są uśrednione lub policzony jest ich błąd standardowy, nie jest możliwe sprawdzenie spójności danych na poziomie pojedynczych komórek. Jednakże ze względu na wysokie korelacje pomiędzy atrybutami dotyczącymi wielkości (np. radius_mean, perimeter_mean, area_mean; radius_worst, perimeter_worst, area_worst; radius_se, perimeter_se, area_se), a także wklęsłości (np. concavity_mean, concave points_mean; concavity_se, concave points_se; concavity_worst, concave points_worst) możemy założyć, że dane są spójne. Dana korelacja nie daje 100% pewności braku błędów w danych, jednakże dane powiązania wydają się w oczywisty sposób logiczne.

Dane niezrozumiałe:

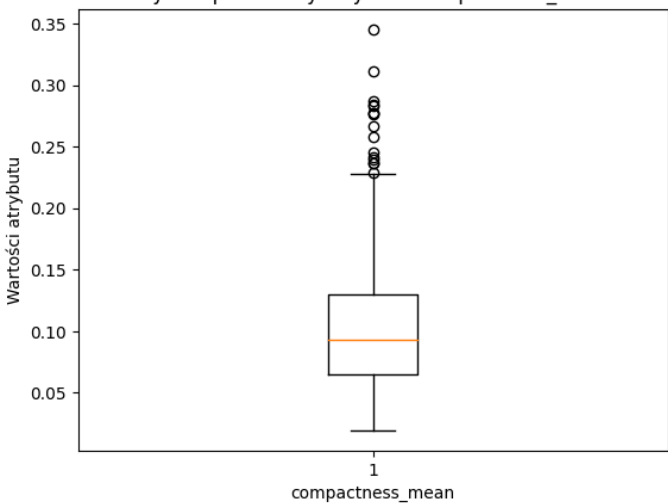
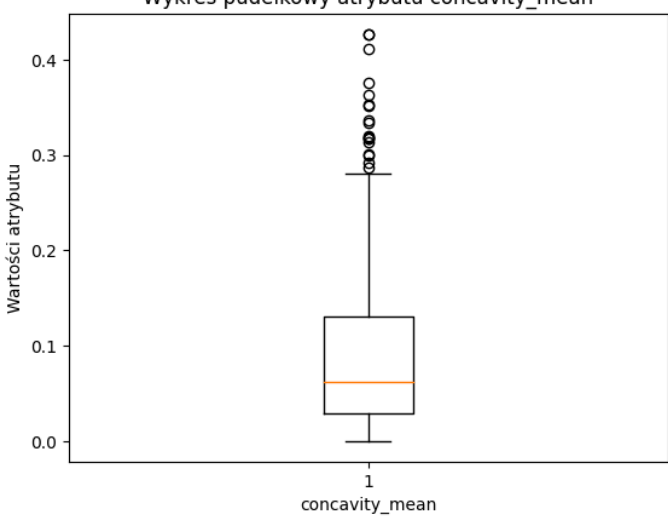
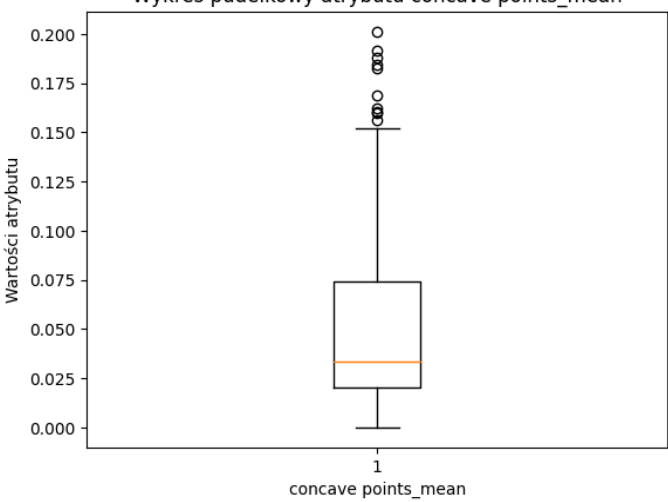
W ogólności dane są zrozumiałe na dość intuicyjnym poziomie. Dla wielu atrybutów rodzą się pytania o ich dokładny sposób liczenia. Nie są podane dokładne wzory, przykładowo dla atrybutów fractal_dimension, symmetry, concavity czy smoothness. Nie wiadomo jak do końca liczony jest obwód jądra komórkowego, jego powierzchnia czy centrum w przypadku atrybutu radius. Nie ma również podanych jednostek poszczególnych atrybutów. Jednakże na potrzeby tego projektu takie intuicyjne zrozumienie atrybutów wydaje się być wystarczające i można dojść do interesujących wniosków.

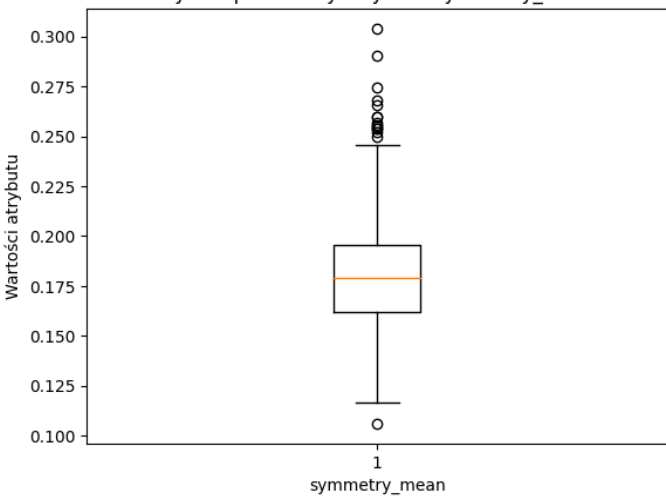
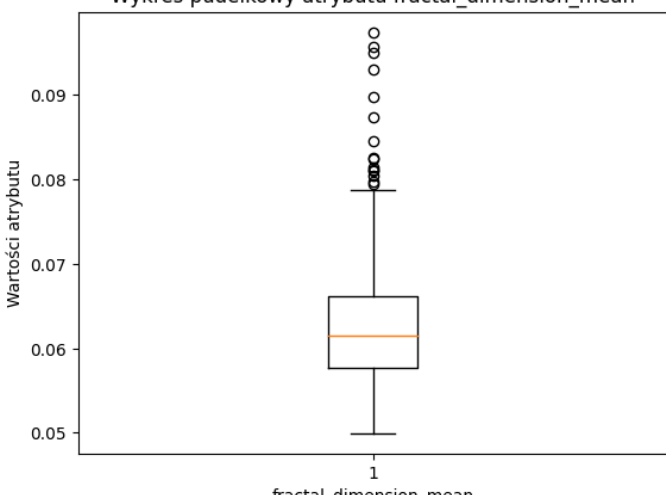
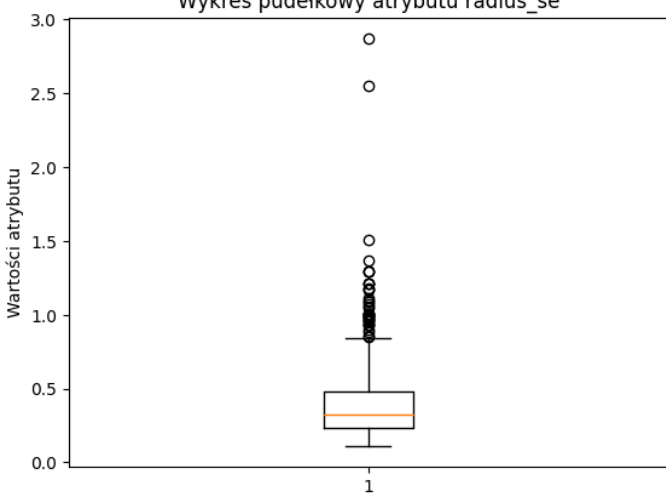
Punkty oddalone:

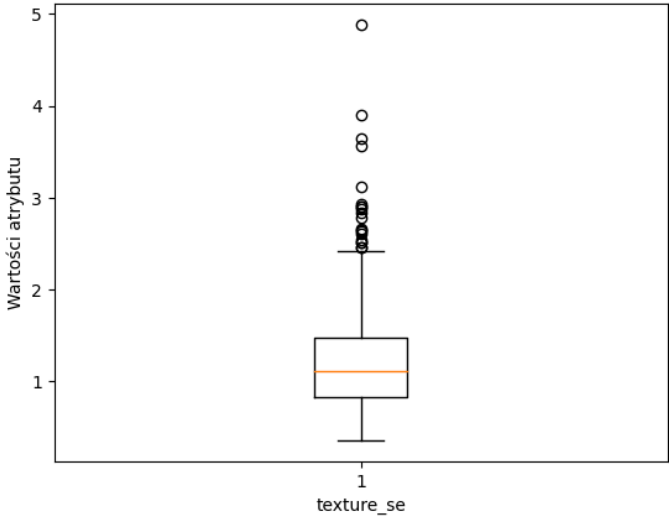
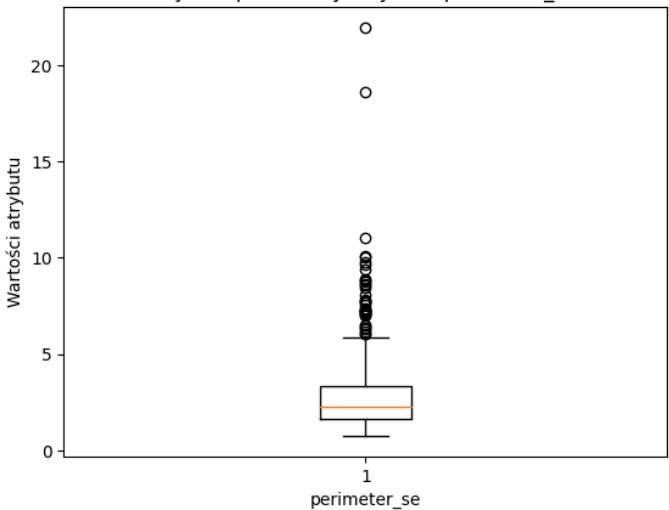
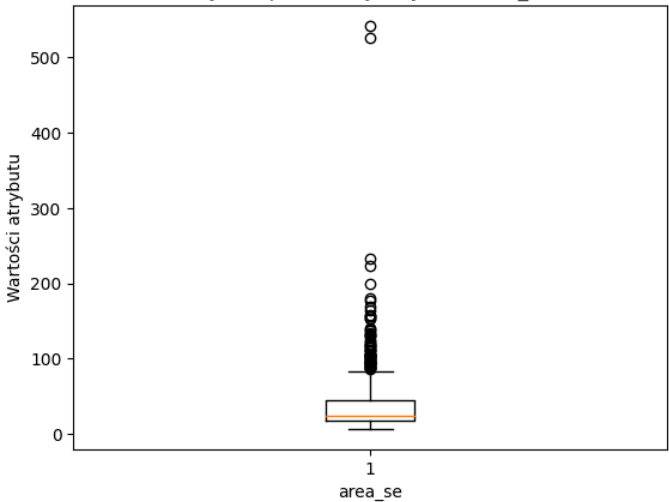
Analizując punkty oddalone można dojść do wniosku, że dane są stosunkowo dobrej jakości. Średnia liczba punktów oddalonych względem atrybutów numerycznych wynosi ok. 20, co stanowi ok. 4% wszystkich danych. Atrybutami o największej liczbie punktów oddalonych są atrybuty area_se (65 punktów oddalonych: ok 11% danych) oraz radius_se i perimeter_se (oba atrybuty mają po 38 punktów oddalonych: ok. 7% wszystkich danych).

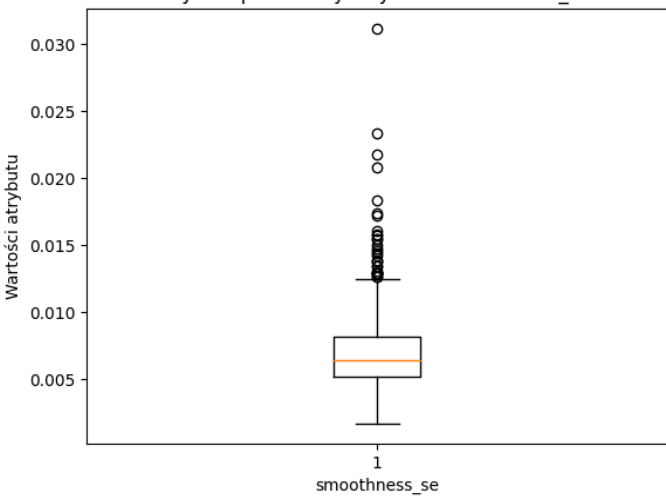
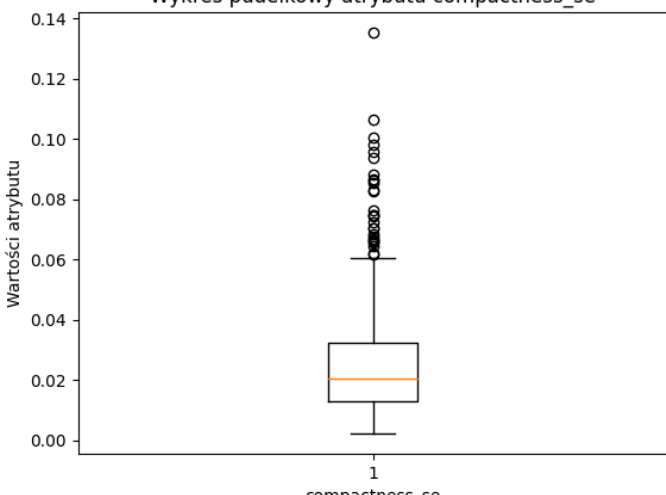
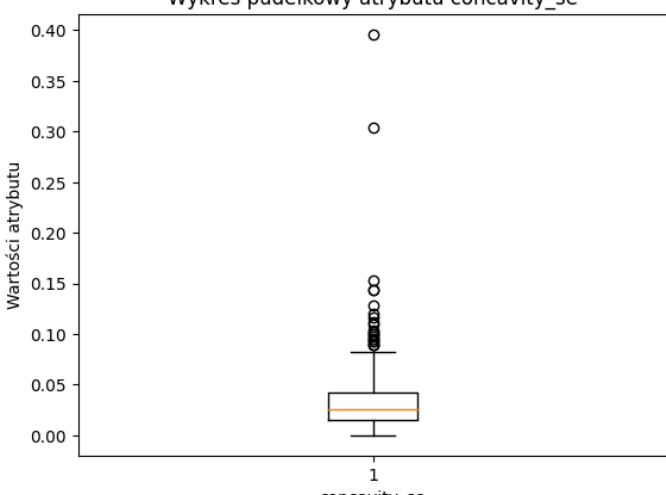
Atrybut	Wykres pudełkowy	Wartości
radius_mean	<p>Wykres pudełkowy atrybutu radius_mean</p> 	<p>Mediana: 13,37</p> <p>Przedział wartości występujących najczęściej: [6,98;21,90]</p> <p>Liczba punktów oddalonych: 14</p>
texture_mean	<p>Wykres pudełkowy atrybutu texture_mean</p> 	<p>Mediana: 18,84</p> <p>Przedział wartości występujących najczęściej: [9,71;30,25]</p> <p>Liczba punktów oddalonych: 7</p>

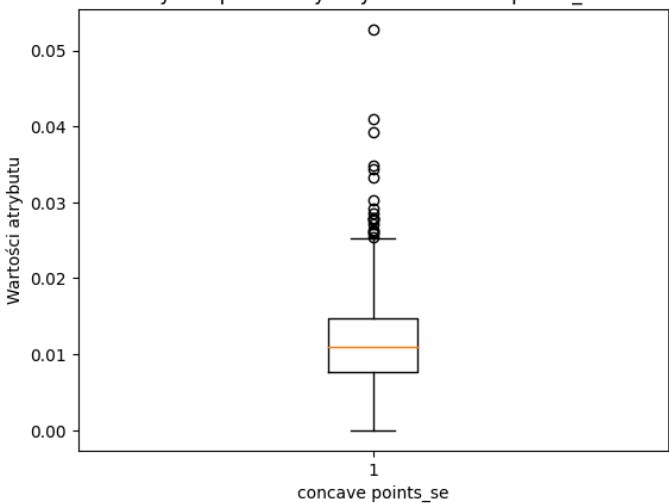
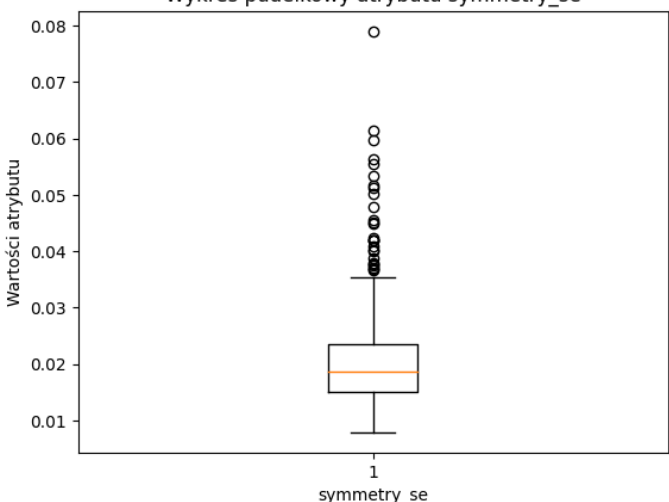
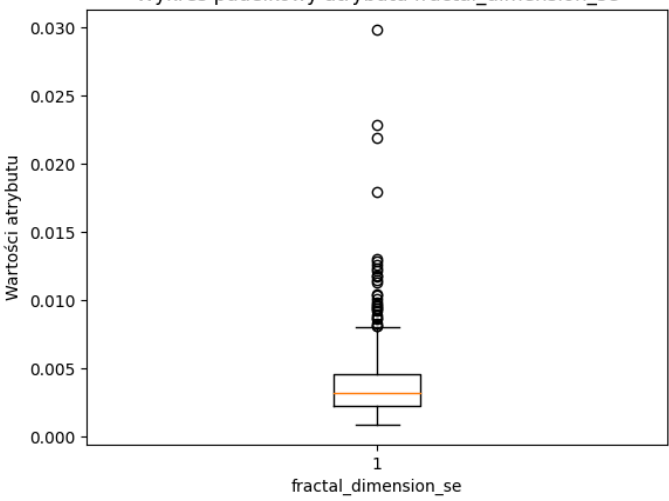
perimeter_mean	<p>Wykres pudełkowy atrybutu perimeter_mean</p>	<p>Mediana: 86,24</p> <p>Przedział wartości występujących najczęściej: [43,79;147,50]</p> <p>Liczba punktów oddalonych: 13</p>
area_mean	<p>Wykres pudełkowy atrybutu area_mean</p>	<p>Mediana: 555,10</p> <p>Przedział wartości występujących najczęściej: [143,50;1326,30]</p> <p>Liczba punktów oddalonych: 25</p>
smoothness_mean	<p>Wykres pudełkowy atrybutu smoothness_mean</p>	<p>Mediana: 0,10</p> <p>Przedział wartości występujących najczęściej: [0,06;0,13]</p> <p>Liczba punktów oddalonych: 6</p>

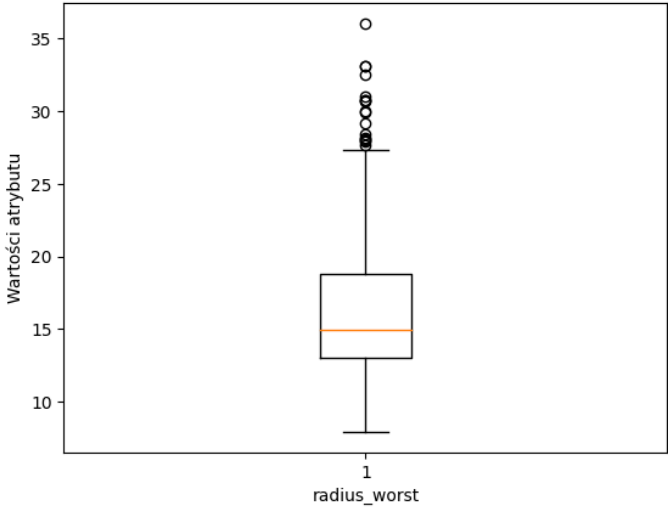
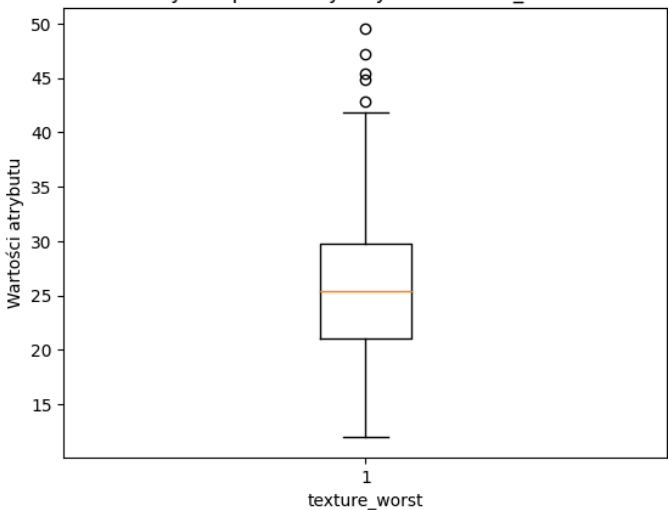
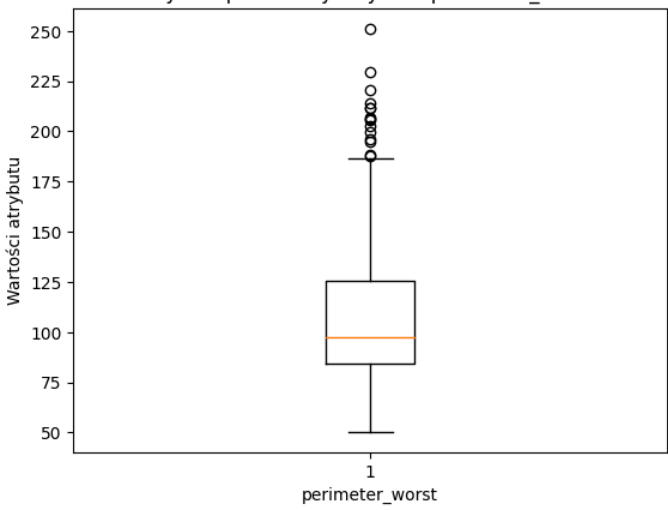
compactness_mean	<p>Wykres pudełkowy atrybutu compactness_mean</p> 	<p>Mediana: 0,09</p> <p>Przedział wartości występujących najczęściej: [0,02;0,23]</p> <p>Liczba punktów oddalonych: 16</p>
concavity_mean	<p>Wykres pudełkowy atrybutu concavity_mean</p> 	<p>Mediana: 0,09</p> <p>Przedział wartości występujących najczęściej: [0,00;0,28]</p> <p>Liczba punktów oddalonych: 18</p>
concave points_mean	<p>Wykres pudełkowy atrybutu concave points_mean</p> 	<p>Mediana: 0,03</p> <p>Przedział wartości występujących najczęściej: [0,00;0,15]</p> <p>Liczba punktów oddalonych: 10</p>

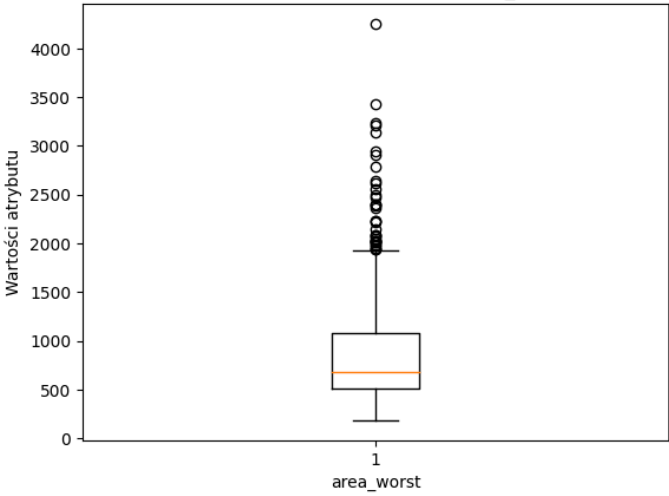
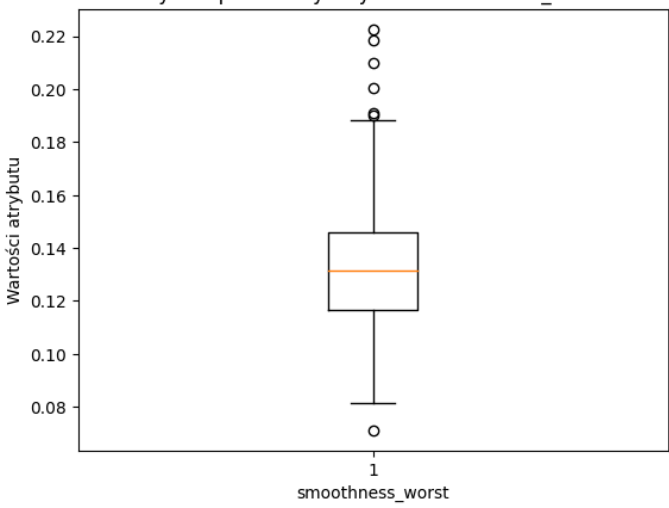
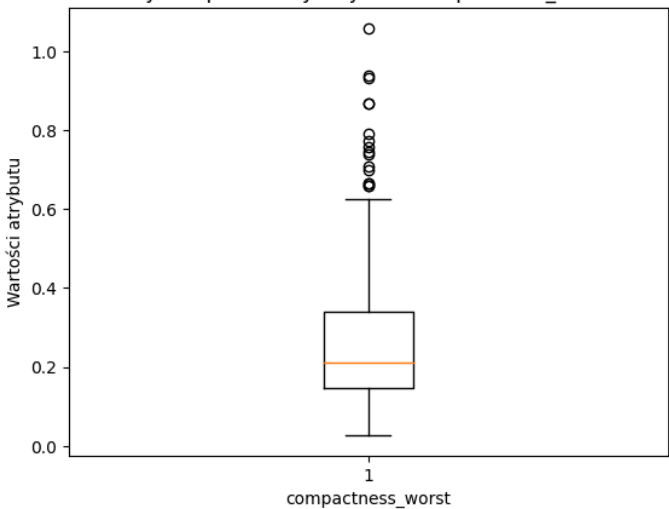
<p>symmetry_mean</p>	<p>Wykres pudełkowy atrybutu symmetry_mean</p> 	<p>Mediana: 0,18</p> <p>Przedział wartości występujących najczęściej: [0,11;0,25]</p> <p>Liczba punktów oddalonych: 15</p>
<p>fractal_dimension_mean</p>	<p>Wykres pudełkowy atrybutu fractal_dimension_mean</p> 	<p>Mediana: 0,06</p> <p>Przedział wartości występujących najczęściej: [0,05;0,08]</p> <p>Liczba punktów oddalonych: 15</p>
<p>radius_se</p>	<p>Wykres pudełkowy atrybutu radius_se</p> 	<p>Mediana: 0,32</p> <p>Przedział wartości występujących najczęściej: [0,11;0,85]</p> <p>Liczba punktów oddalonych: 38</p>

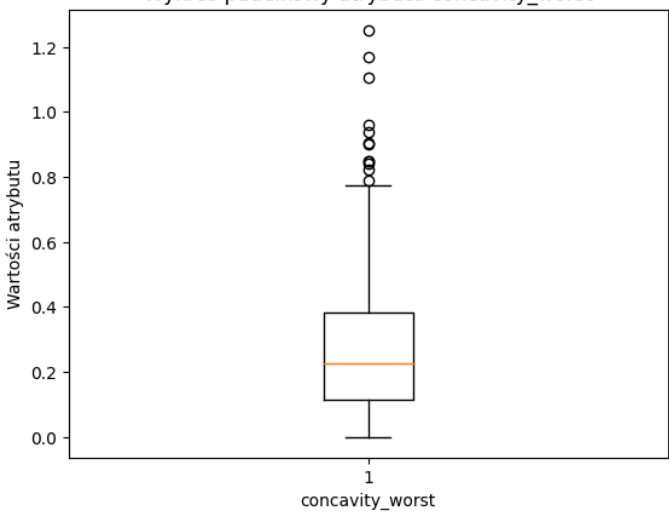
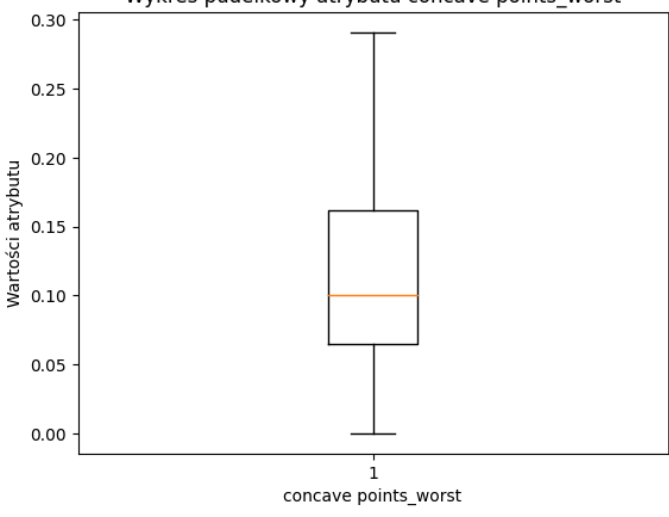
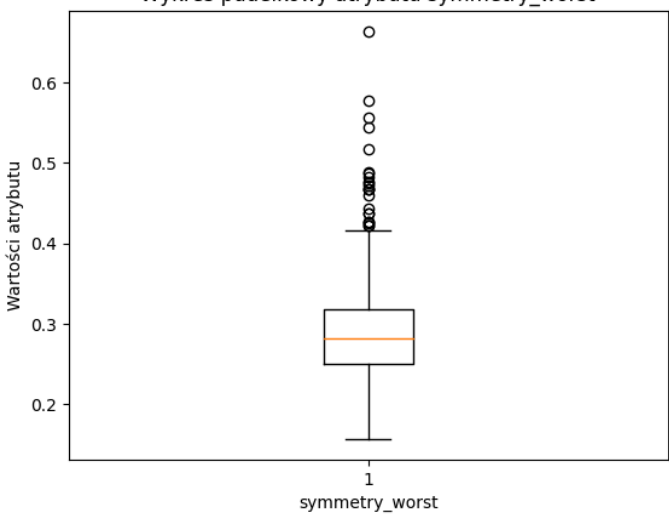
texture_se	<p>Wykres pudełkowy atrybutu texture_se</p> 	<p>Mediana: 1,11</p> <p>Przedział wartości występujących najczęściej: [0,36;2,43]</p> <p>Liczba punktów oddalonych: 20</p>
perimeter_se	<p>Wykres pudełkowy atrybutu perimeter_se</p> 	<p>Mediana: 2,29</p> <p>Przedział wartości występujących najczęściej: [0,76;5,98]</p> <p>Liczba punktów oddalonych: 38</p>
area_se	<p>Wykres pudełkowy atrybutu area_se</p> 	<p>Mediana: 24,53</p> <p>Przedział wartości występujących najczęściej: [6,80;86,20]</p> <p>Liczba punktów oddalonych: 65</p>

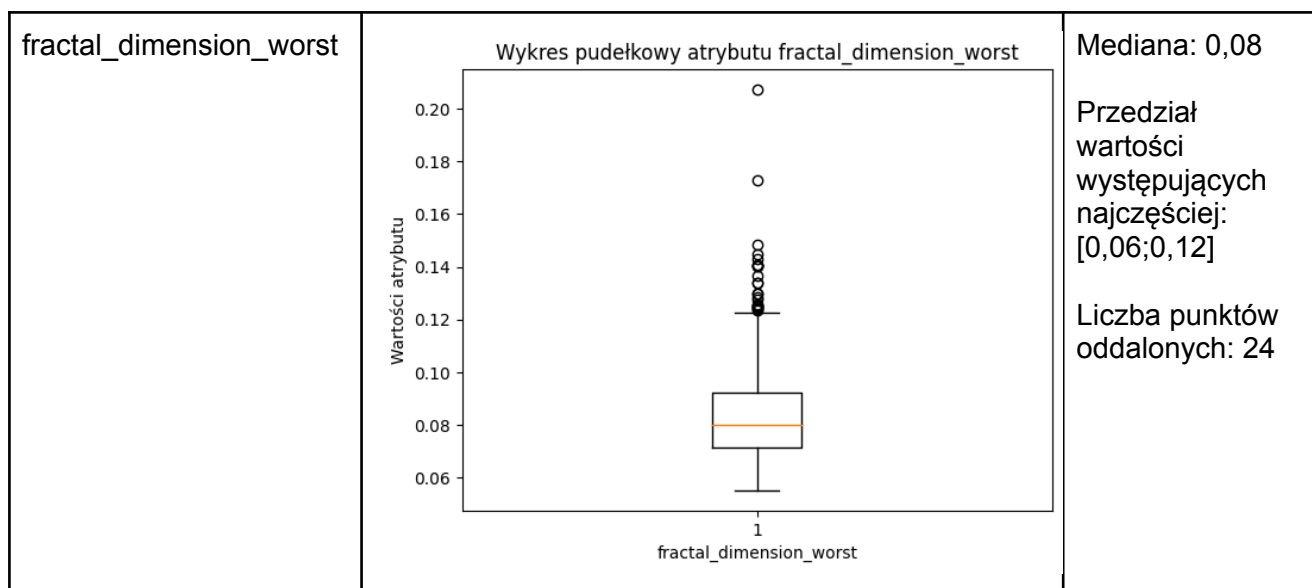
smoothness_se	<p>Wykres pudełkowy atrybutu smoothness_se</p> 	<p>Mediana: 0,006</p> <p>Przedział wartości występujących najczęściej: [0,002;0,013]</p> <p>Liczba punktów oddalonych: 30</p>
compactness_se	<p>Wykres pudełkowy atrybutu compactness_se</p> 	<p>Mediana: 0,020</p> <p>Przedział wartości występujących najczęściej: [0,002;0,062]</p> <p>Liczba punktów oddalonych: 28</p>
concavity_se	<p>Wykres pudełkowy atrybutu concavity_se</p> 	<p>Mediana: 0,03</p> <p>Przedział wartości występujących najczęściej: [0,00;0,08]</p> <p>Liczba punktów oddalonych: 22</p>

concave points_se	<p>Wykres pudełkowy atrybutu concave points_se</p> 	<p>Mediana: 0,01</p> <p>Przedział wartości występujących najczęściej: [0,00;0,03]</p> <p>Liczba punktów oddalonych: 19</p>
symmetry_se	<p>Wykres pudełkowy atrybutu symmetry_se</p> 	<p>Mediana: 0,020</p> <p>Przedział wartości występujących najczęściej: [0,008;0,036]</p> <p>Liczba punktów oddalonych: 27</p>
fractal_dimension_se	<p>Wykres pudełkowy atrybutu fractal_dimension_se</p> 	<p>Mediana: 0,003</p> <p>Przedział wartości występujących najczęściej: [0,001;0,008]</p> <p>Liczba punktów oddalonych: 28</p>

radius_worst	<p>Wykres pudełkowy atrybutu radius_worst</p> 	<p>Mediana: 14,97</p> <p>Przedział wartości występujących najczęściej: [7,93;27,46]</p> <p>Liczba punktów oddalonych: 17</p>
texture_worst	<p>Wykres pudełkowy atrybutu texture_worst</p> 	<p>Mediana: 25,41</p> <p>Przedział wartości występujących najczęściej: [12,02;42,68]</p> <p>Liczba punktów oddalonych: 5</p>
perimeter_worst	<p>Wykres pudełkowy atrybutu perimeter_worst</p> 	<p>Mediana: 97,66</p> <p>Przedział wartości występujących najczęściej: [50,41;187,34]</p> <p>Liczba punktów oddalonych: 15</p>

area_worst	<p>Wykres pudełkowy atrybutu area_worst</p> 	<p>Mediana: 686,50</p> <p>Przedział wartości występujących najczęściej: [185,2;1937,10]</p> <p>Liczba punktów oddalonych: 35</p>
smoothness_worst	<p>Wykres pudełkowy atrybutu smoothness_worst</p> 	<p>Mediana: 0,13</p> <p>Przedział wartości występujących najczęściej: [0,07;0,19]</p> <p>Liczba punktów oddalonych: 7</p>
compactness_worst	<p>Wykres pudełkowy atrybutu compactness_worst</p> 	<p>Mediana: 0,21</p> <p>Przedział wartości występujących najczęściej: [0,03;0,63]</p> <p>Liczba punktów oddalonych: 16</p>

concavity_worst	<p>Wykres pudełkowy atrybutu concavity_worst</p> 	<p>Mediana: 0,23</p> <p>Przedział wartości występujących najczęściej: [0,00;0,79]</p> <p>Liczba punktów oddalonych: 12</p>
concave points_worst	<p>Wykres pudełkowy atrybutu concave points_worst</p> 	<p>Mediana: 0,10</p> <p>Przedział wartości występujących najczęściej: [0,00;0,29]</p> <p>Liczba punktów oddalonych: 0</p>
symmetry_worst	<p>Wykres pudełkowy atrybutu symmetry_worst</p> 	<p>Mediana: 0,28</p> <p>Przedział wartości występujących najczęściej: [0,16;0,42]</p> <p>Liczba punktów oddalonych: 23</p>



Podsumowanie

Na początku przeprowadzono analizę rozkładów, z której wynika, że atrybuty numeryczne nie mają rozkładów normalnych a klasy atrybutu celu nie są zbalansowane. W związku z tym, że rozkłady nie były normalne zastosowano metodę Spearmana do analizy korelacji. Z analizy korelacji wynika, że występują grupy atrybutów silnie ze sobą skorelowane - pierwsza grupa: radius_mean, radius_worst, perimeter_mean, perimeter_worst, area_mean, area_worst, druga grupa: radius_se, area_se, perimeter_se, trzecia grupa: concavity_mean, concavity_worst, concavity_se, concave points_mean, concave points_worst, concave points_se, compactness_worst, compactness_mean, compactness_se. Ze względu na dużą korelację danych atrybutów, przekazują one podobną informację, co może być redundantne dla modelu.

Następnie obliczono korelację względem atrybutu celu - diagnosis. Atrybutami, które silnie korelują ze zmienną diagnosis są w kolejności: concave points_worst, perimeter_worst, concave points_mean, radius_worst, area_mean, concavity_mean, concavity_worst, compactness_mean, compactness_worst, radius_se, perimeter_se oraz area_se. Są to atrybuty należące do grup silnie skorelowanych zmiennych. W związku z czym można usunąć niektóre atrybuty pozostawiając tylko jeden w grupie, przykładowo biorąc pod uwagę wartości korelacji ze zmienną diagnosis - concave points_worst, perimeter_worst i radius_se.

Dodatkowo większość atrybutów ma korelację dodatnią z rodzajem nowotworu, z czego wynika, że ich wyższe wartości częściej występują wśród pacjentów z nowotworem złośliwym (ponieważ do obliczania korelacji M zostało zamieniona na 1 a B na 0). Zgadza się to z biologicznymi właściwościami komórek nowotworu złośliwego, które mają uszkodzone geny odpowiedzialne za kontrolę wzrostu oraz podziałem komórki, a także mogą mieć różny stopień uszkodzenia DNA, przez co ich jądra komórkowe mogą być większe i bardziej nieregularne od zdrowych.

Następnie dokonano oceny jakości danych. Głównym mankamentem danych jest nieprecyzyjne wyjaśnienie kolumn - nie podano wzorów ich dokładnego wyznaczania oraz jednostek. Dlatego ciężko jest ocenić, czy występowanie punktów oddalonych ma sens czy też nie. Jednakże ze względu na dodatnią korelację atrybutów do rodzaju nowotworu, występowanie outlierów może świadczyć o złośliwości nowotworu, a zatem może być

przydatne dla modelu. W związku z czym najlepszą decyzją byłoby pozostawienie wartości odstających.

Biorąc pod uwagę stosunkowo dobrą jakość danych, uważam, że cel eksploracji jest możliwy do spełnienia.