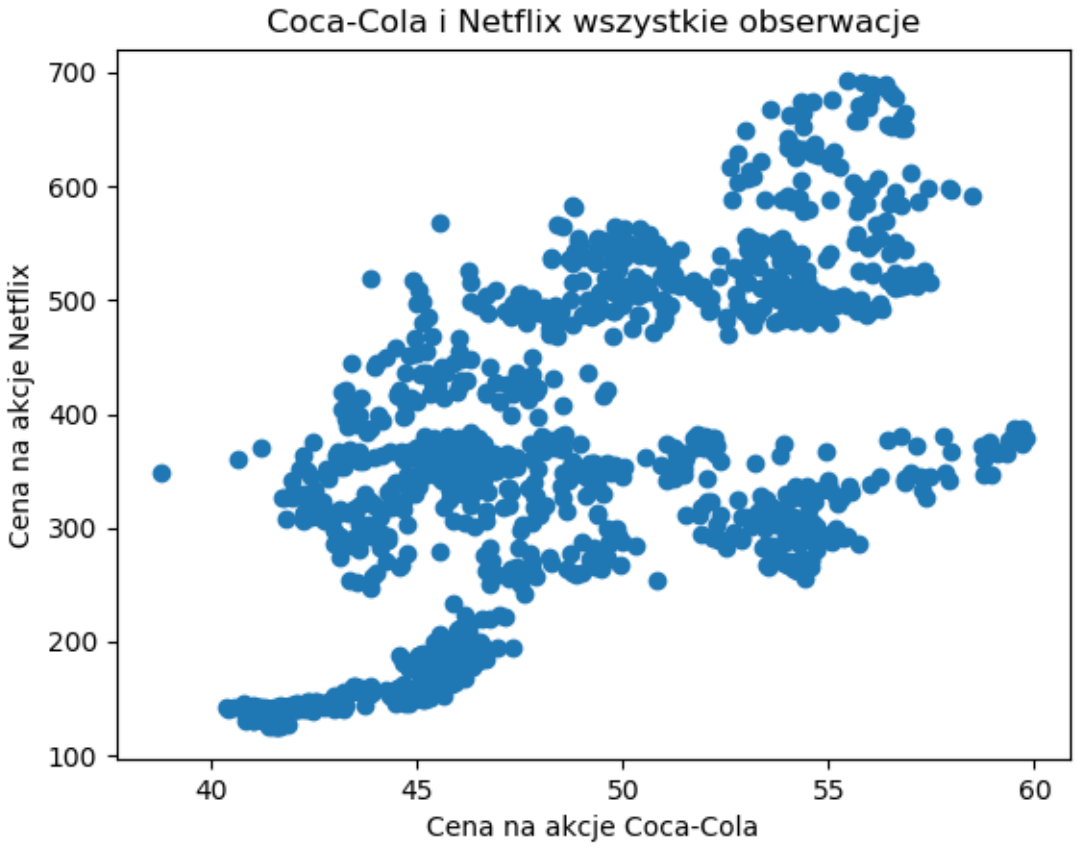
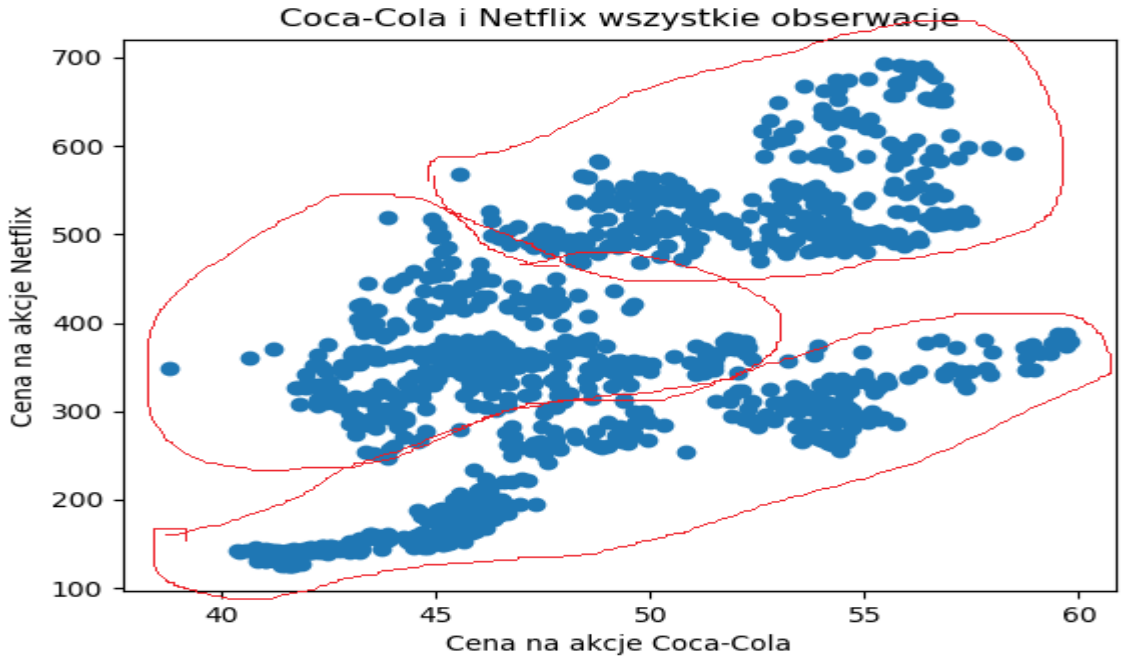
Raport 1

Autor: Kostiantyn Skopych

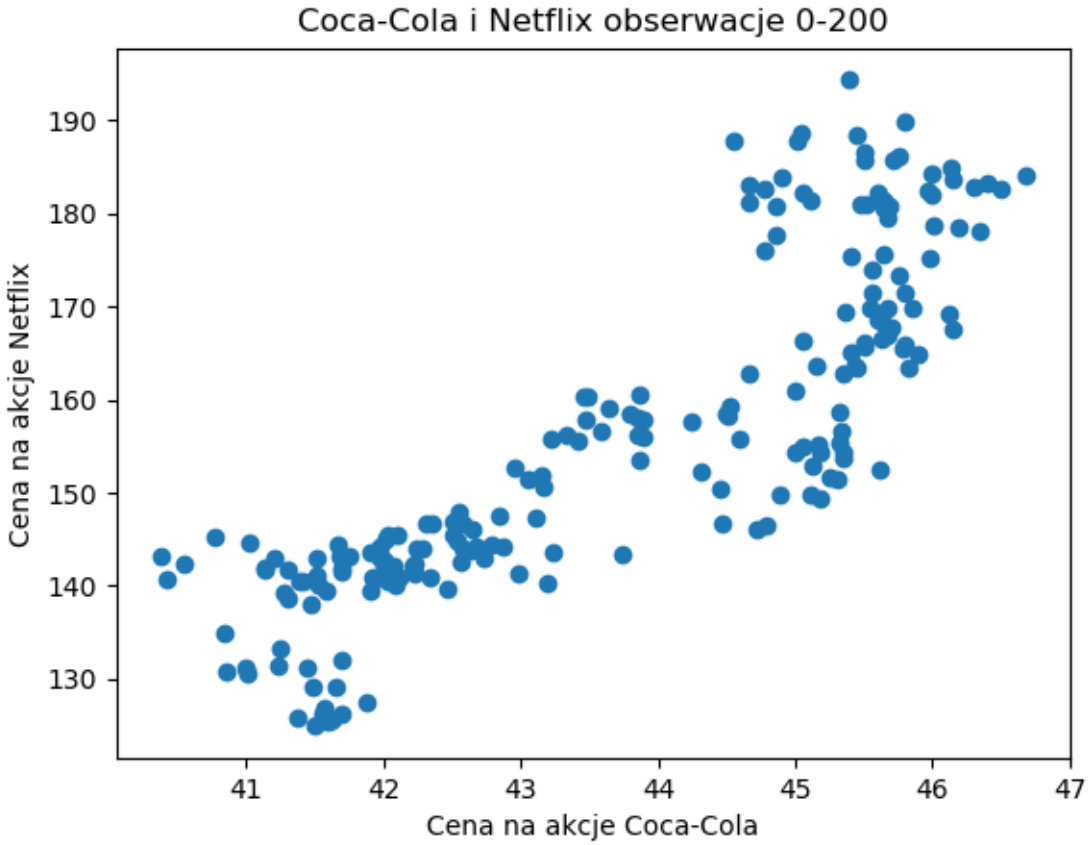
2. Źródło danych: strona https://finance.yahoo.com/. Ceny na akcje firmy Coca-Cola oraz Netflix w ciągu ostatnich 5 lat. Ogólna liczba obserwacji: 1258. Na wykresie poniżej przedstawiony jest wykres tych danych.

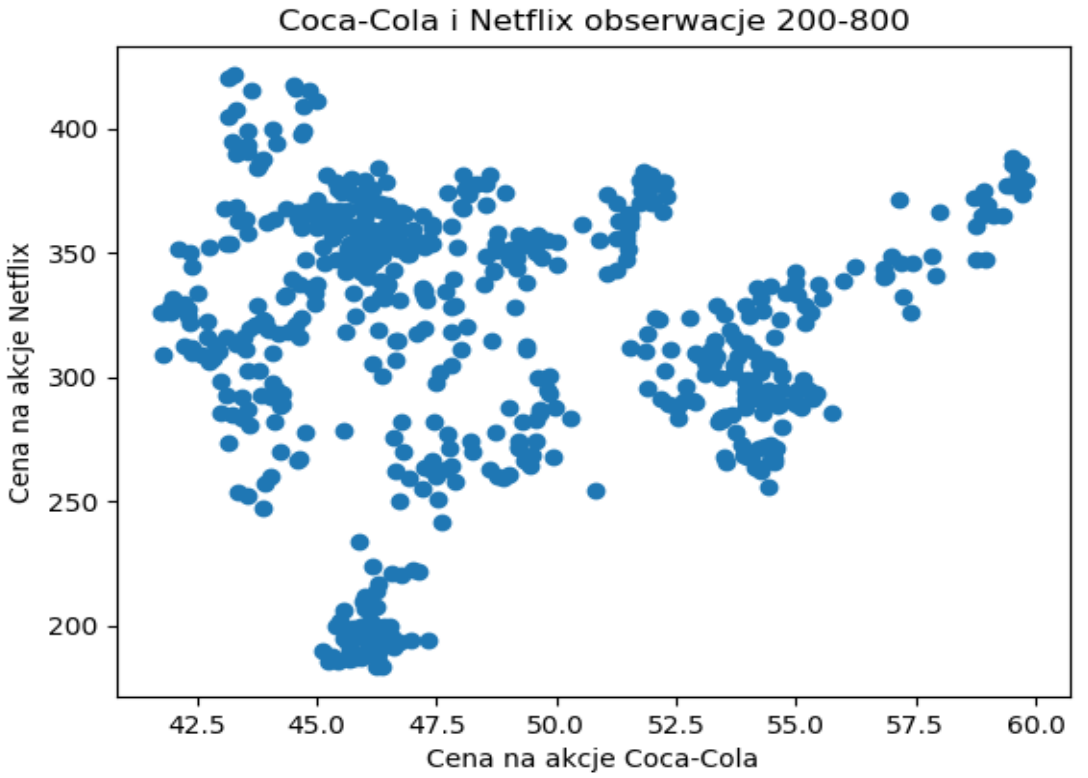


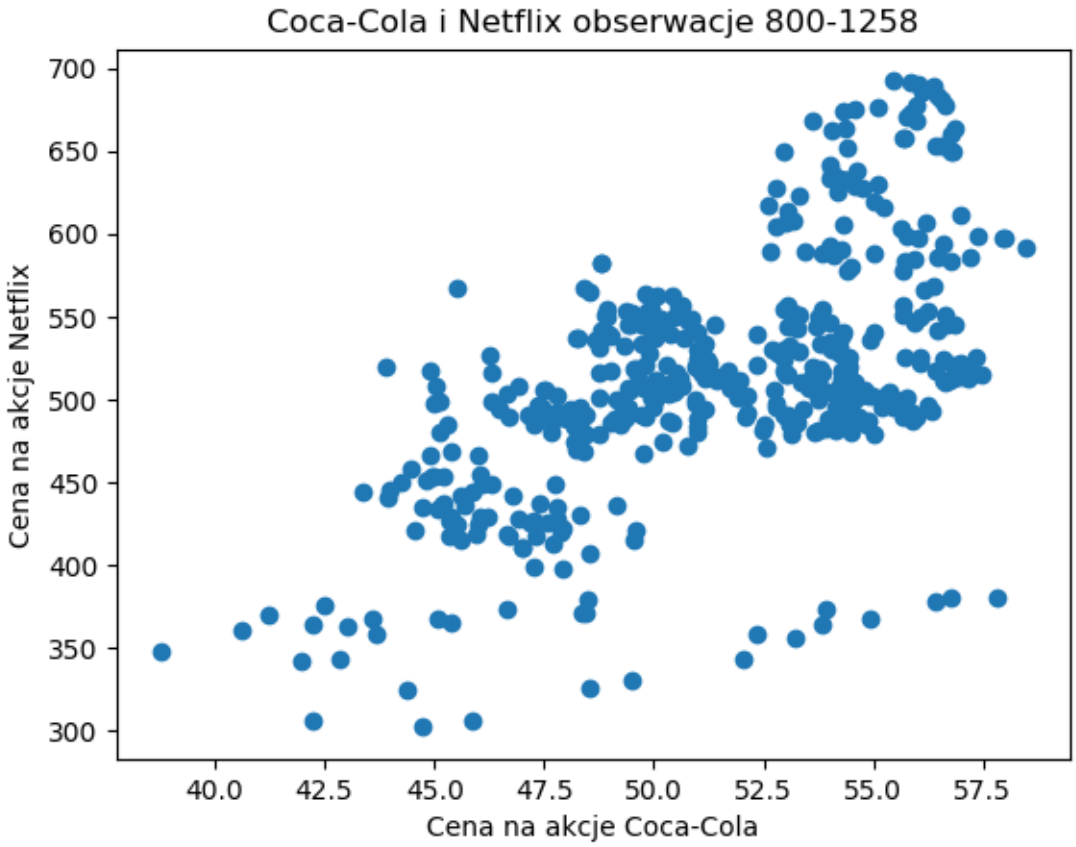

Patrząc na dany wykres widzimy, że są jakieś korelacje między tymi danymi, jednak są one niby pogrupowane. Dlatego chcielibyśmy pogrupować te dane tak jak jest pokazane na poniższym rysunku, żeby nasze badanie było bardziej precyzyjne.



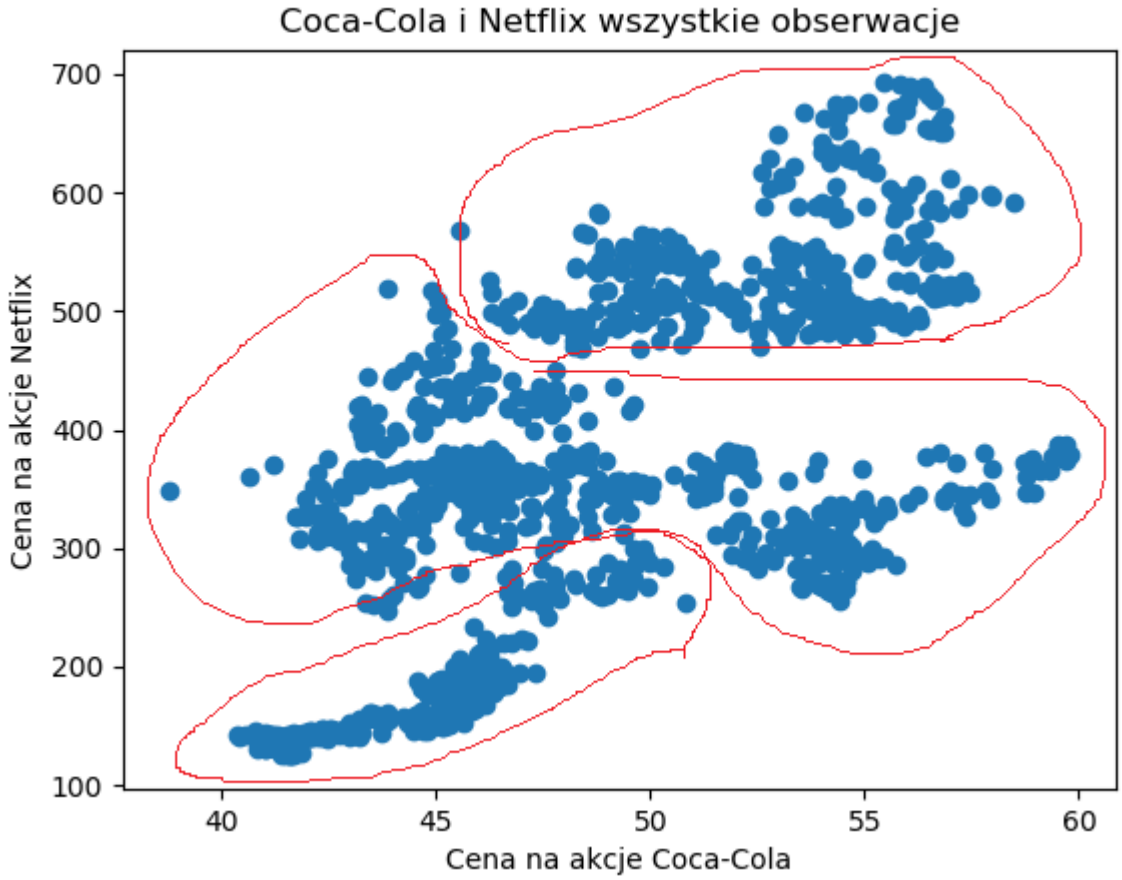
Próbując różne podejścia okazuje się że dane łatwo się grupują w wygodny sposób na podstawie okresu. Dlatego od tego momentu będziemy badali jednocześnie ogólny zbiór danych oraz trzy grupy danych (oznaczane 1,2,3), gdzie grupa 1 obejmuje bserwacje 0-200, grupa 2 – obserwacje 200-800, grupa 3 – obserwacje 800-1258. Liczba obserwacji grupy 1: 200. Liczba obserwacji grupy 2: 600. Liczba obserwacji grupy 3: 458. Na poniższych wykresach możemy zobaczyć jak wyglądają te pogrupowane dane,





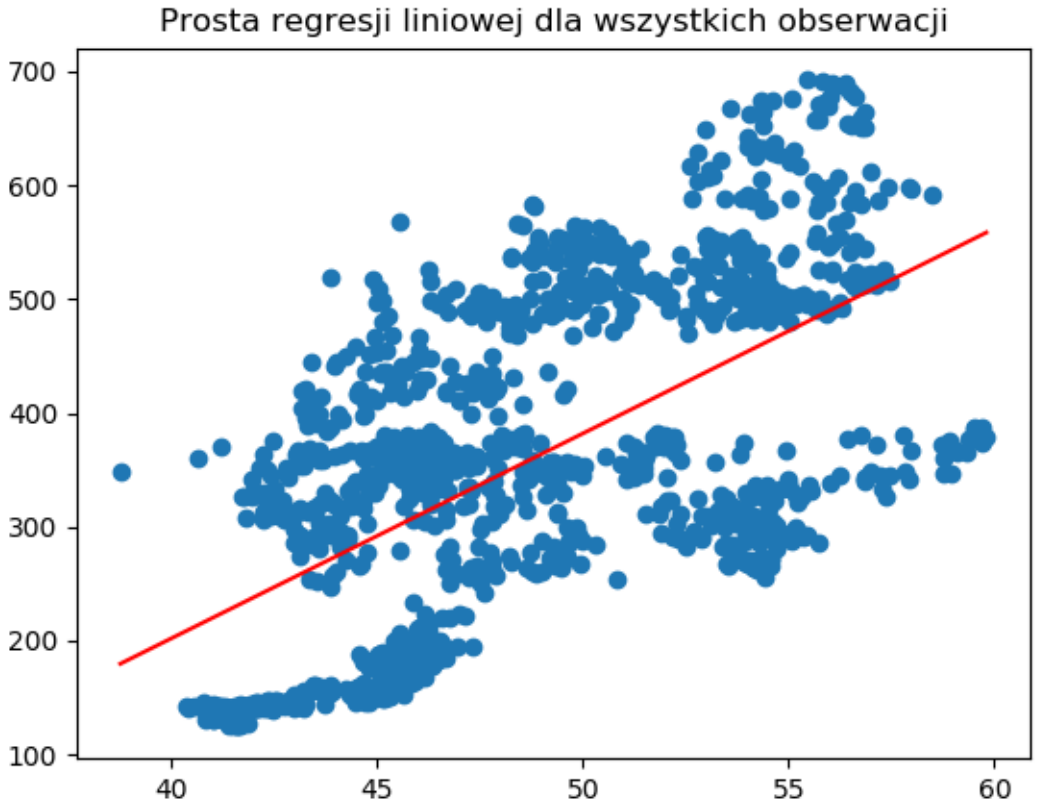

Widzimy, że dane się pogrupowały trochę inaczej niż planowaliśmy, jednak i tak bardziej się nadają do analizy w takiej postaci. Na poniższym rysunku Jest pokazane jak są teraz pogrupowane dane.

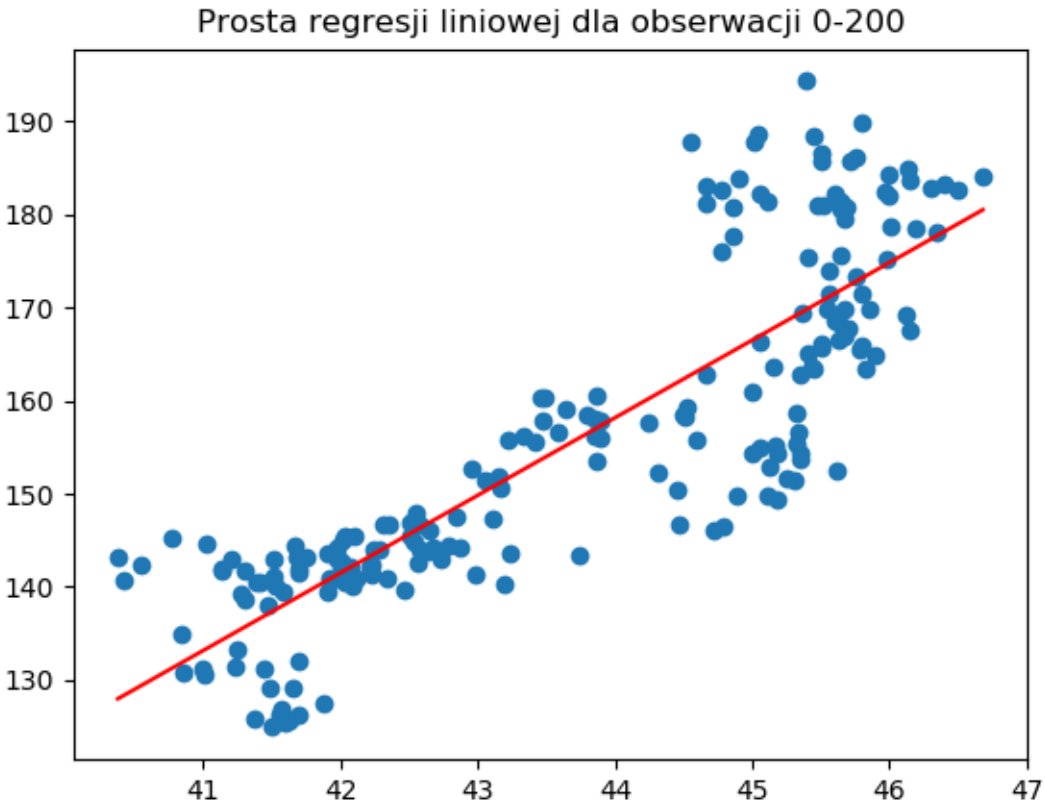


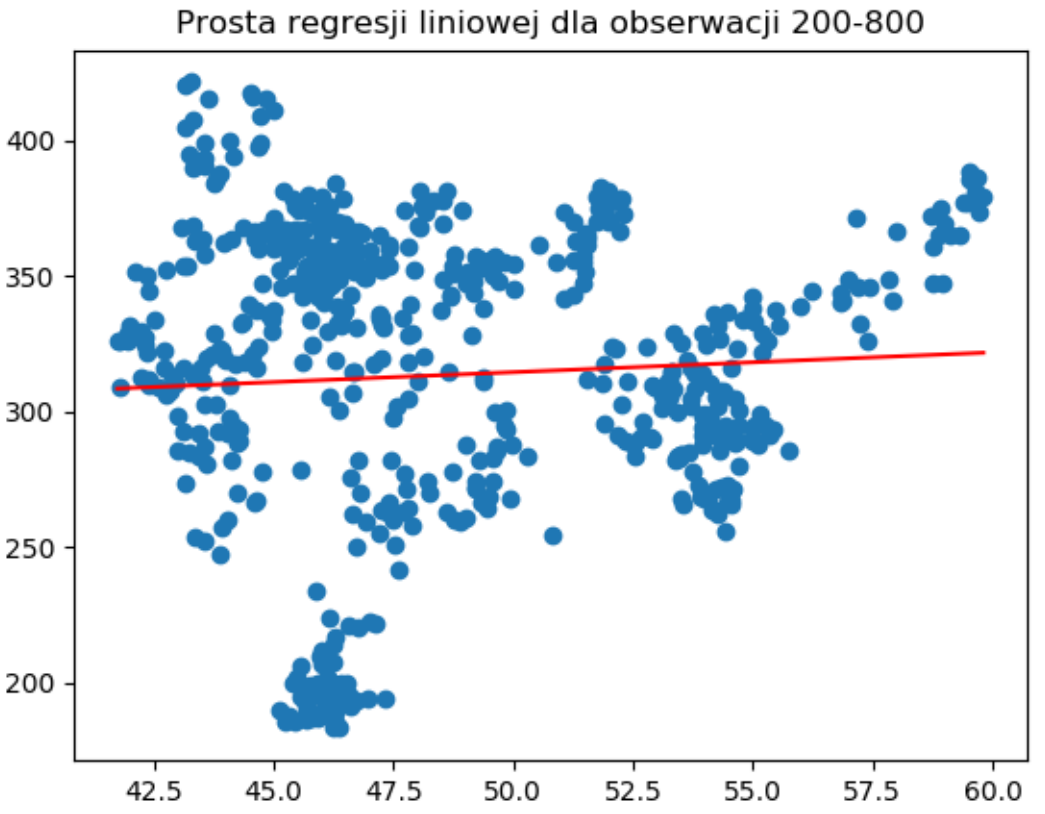
3. Statystyki opisowe.

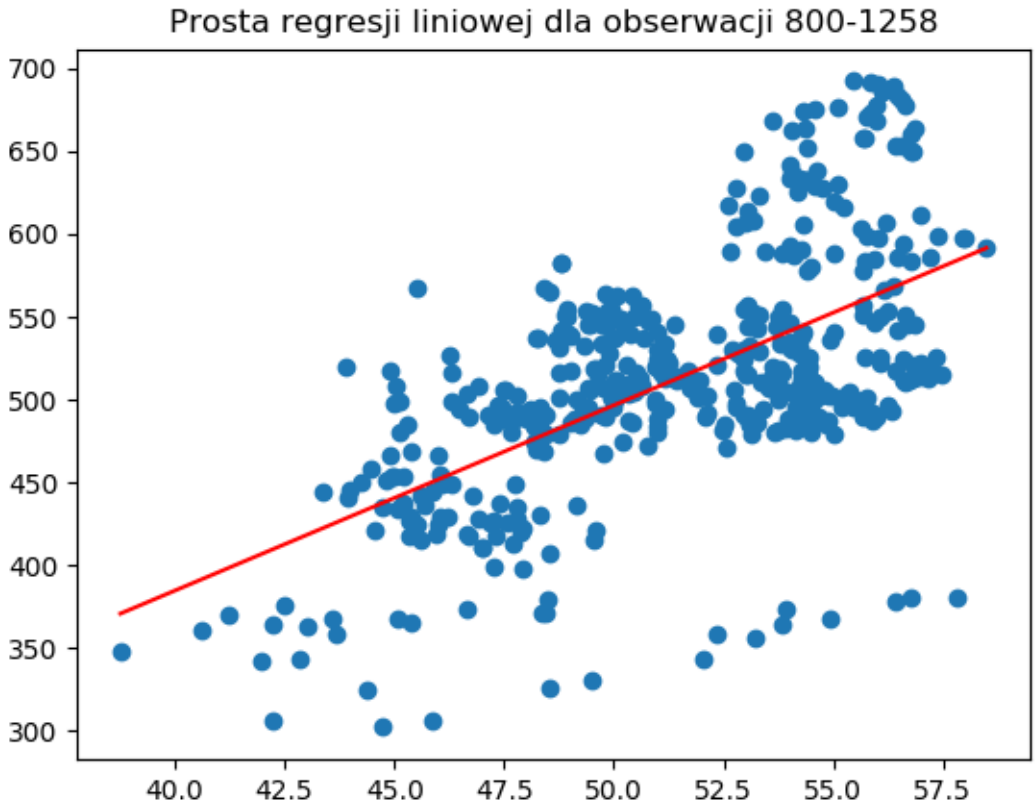
| | Ogólne Coca-Cola | Ogólne Netflix | Grupa 1 Coca-Cola | Grupa 1 Netflix | Grupa 2 Coca-Cola | Grupa 2 Netflix | Grupa 3 Coca-Cola | Grupa 3 Netflix |
|---|---|---|---|---|---|---|---|---|
| Średnia | 48.8 | 361 | 43.7 | 155.9 | 48.5 | 313 | 51.5 | 512 |
| Mediana | 47.8 | 352.2 | 43.9 | 152.8 | 46.8 | 324 | 52.2 | 510 |
| Wariancja | 21.8 | 19694 | 3.1 | 297 | 18.8 | 3163 | 15.5 | 5399 |

4. Na poniższych wykresach są dobrane proste regresji liniowej do naszych danych.




Prosta regresji liniowej dla obserwacji 0-200
190
180
170
160
150
140
130
41
42
43
44
45
46
47


Prosta regresji liniowej dla obserwacji 200-800
400
350
300
250
200
42.5
45.0
47.5
50.0
52.5
55.0
57.5
60.0

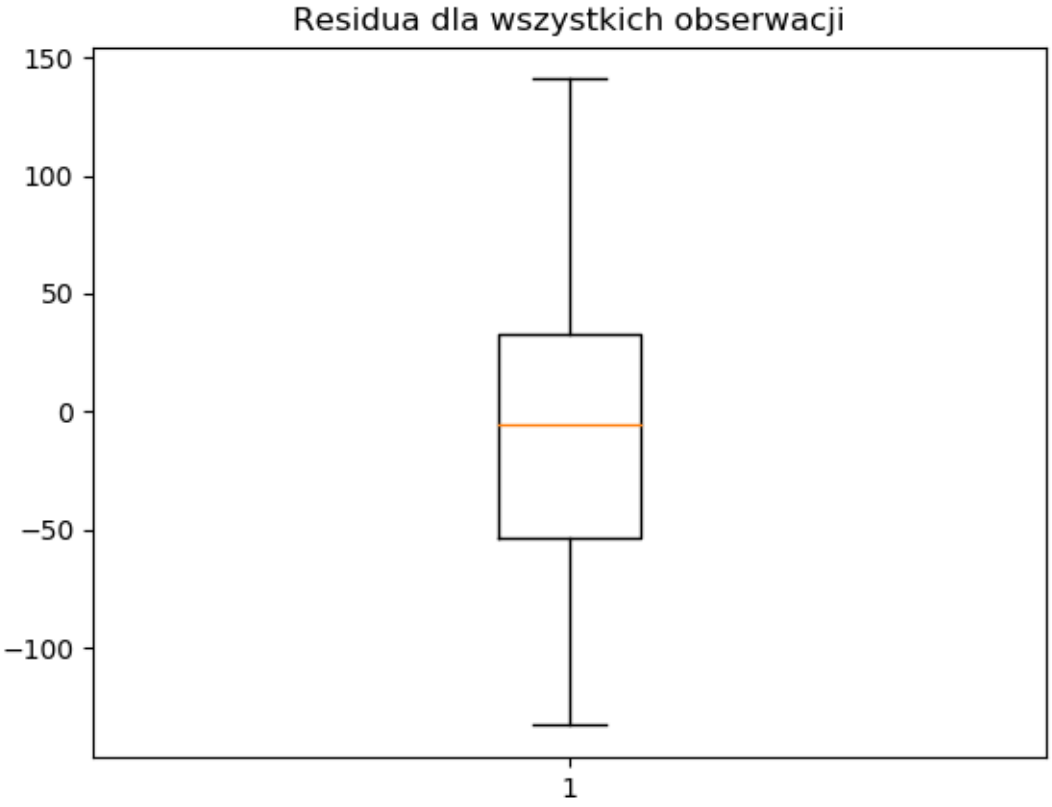


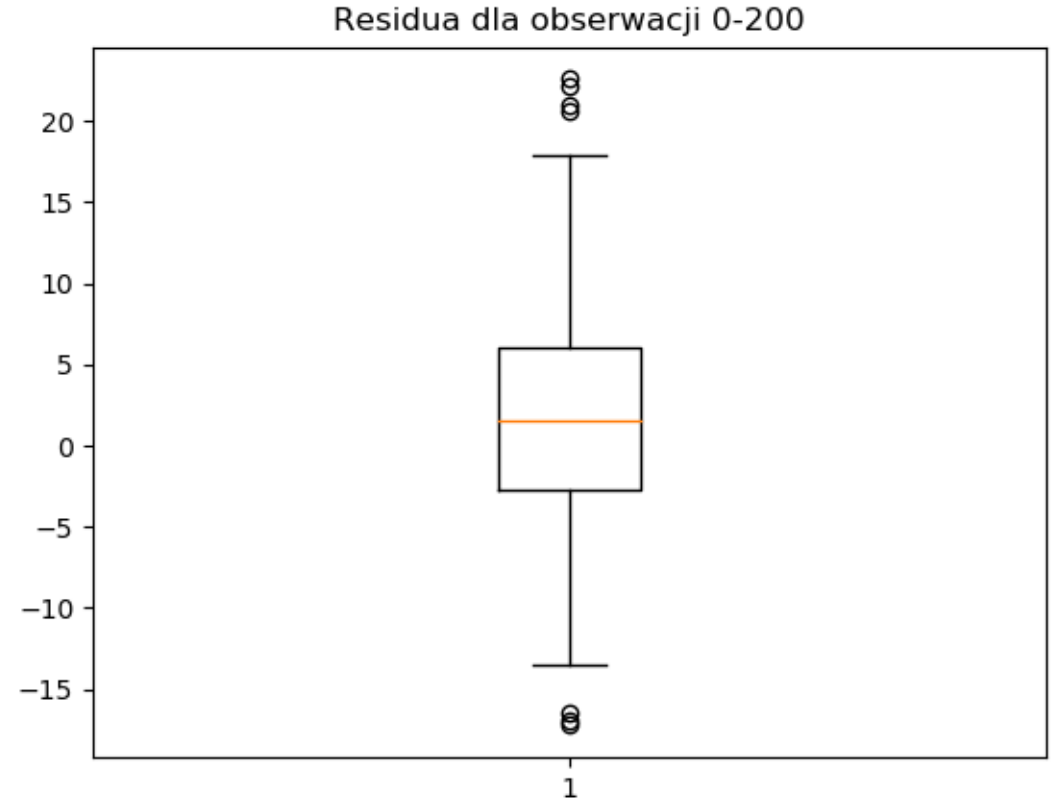

Interesują nas też wartości współczynników determinancji. Dla ogólnego zbioru danych $R^2$= 0.61. Dla grupy 1 $R^2$=0.79. Dla grupy 2 $R^2$=0.005. Dla grupy 3 $R^2$=0.94.

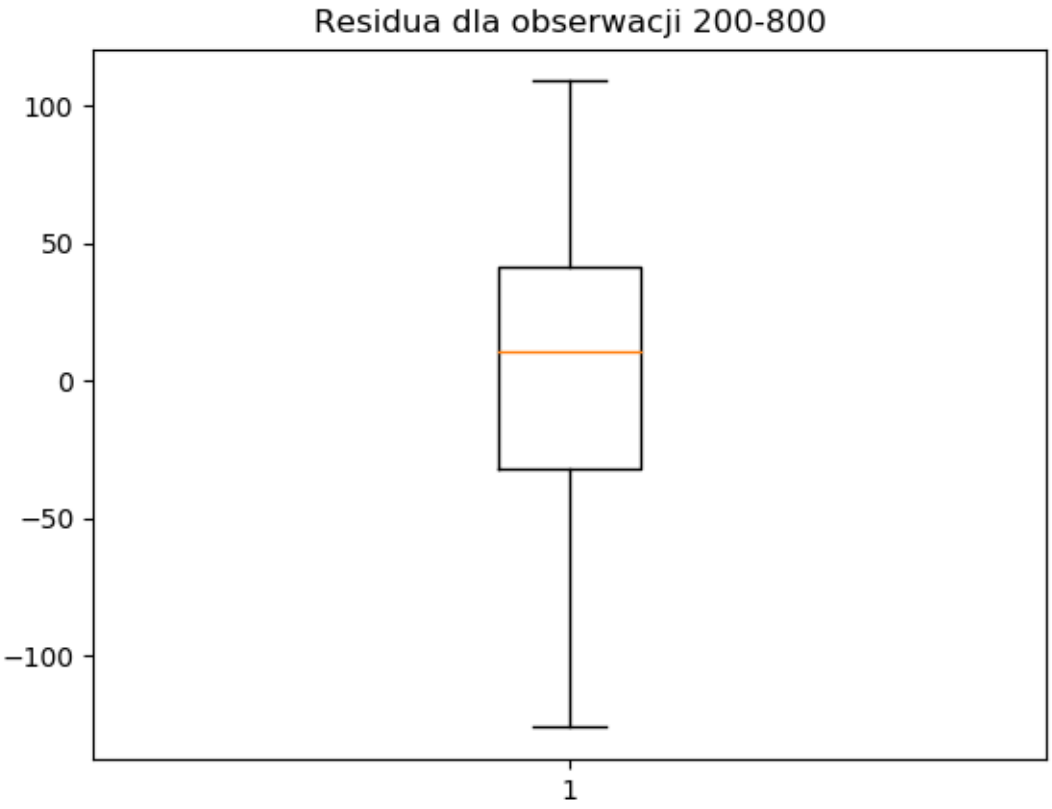
5. W tabeli poniżej są obliczone przedziały ufności dla parametrów $\beta_0$, $\beta_1$.

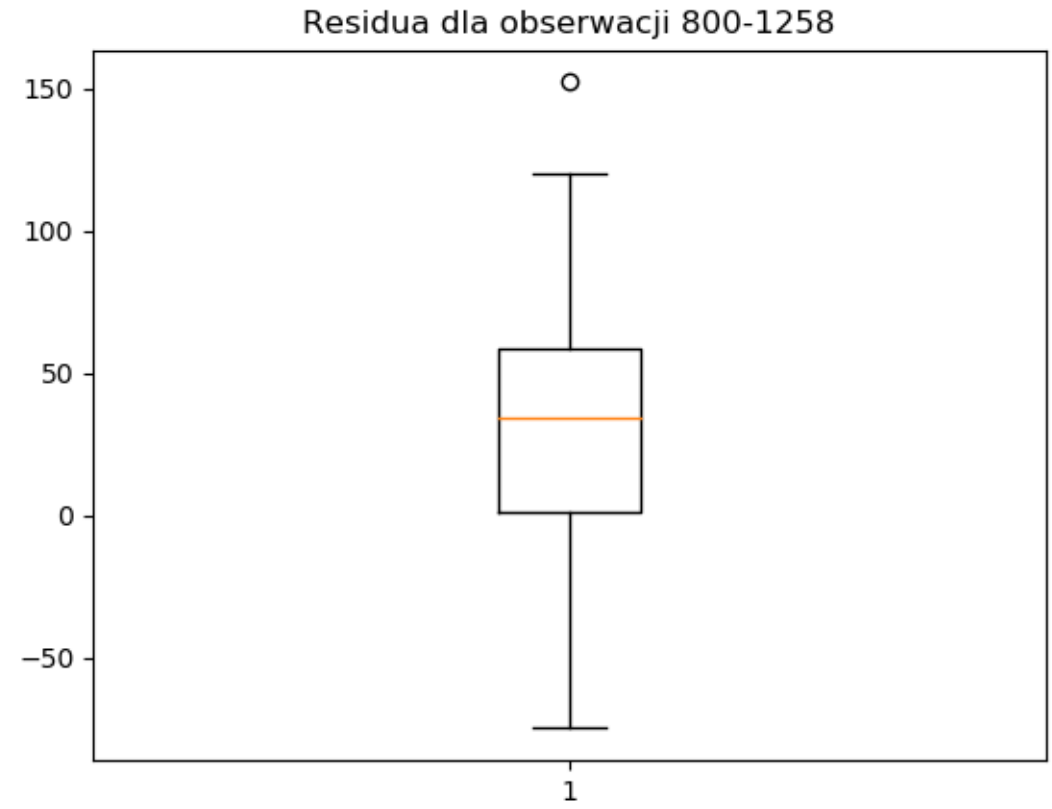
| | Ogólny zbiór | Grupa 1 | Grupa 2 | Grupa 3 |
|---|---|---|---|---|
| Dolne ograniczenie $\beta_1$ | 17.3 | 7.8 | -0.3 | 10 |
| Górne ograniczenie $\beta_1$ | 18.7 | 8.9 | 1.7 | 12.4 |
| Dolne ograniczenie $\beta_0$ | -550 | -233 | 228 | -123 |
| Górne ograniczenie $\beta_0$ | -485 | -183 | 327 | -1.5 |

6. Analiza residuów. Na wykresach poniżej są przedstawione boxploty residuów dla odpowiednich grup danych.


Residua dla wszystkich obserwacji
150
100
50
0
−50
−100
1


Residua dla obserwacji 0-200
20
15
10
5
0
−5
−10
−15
1

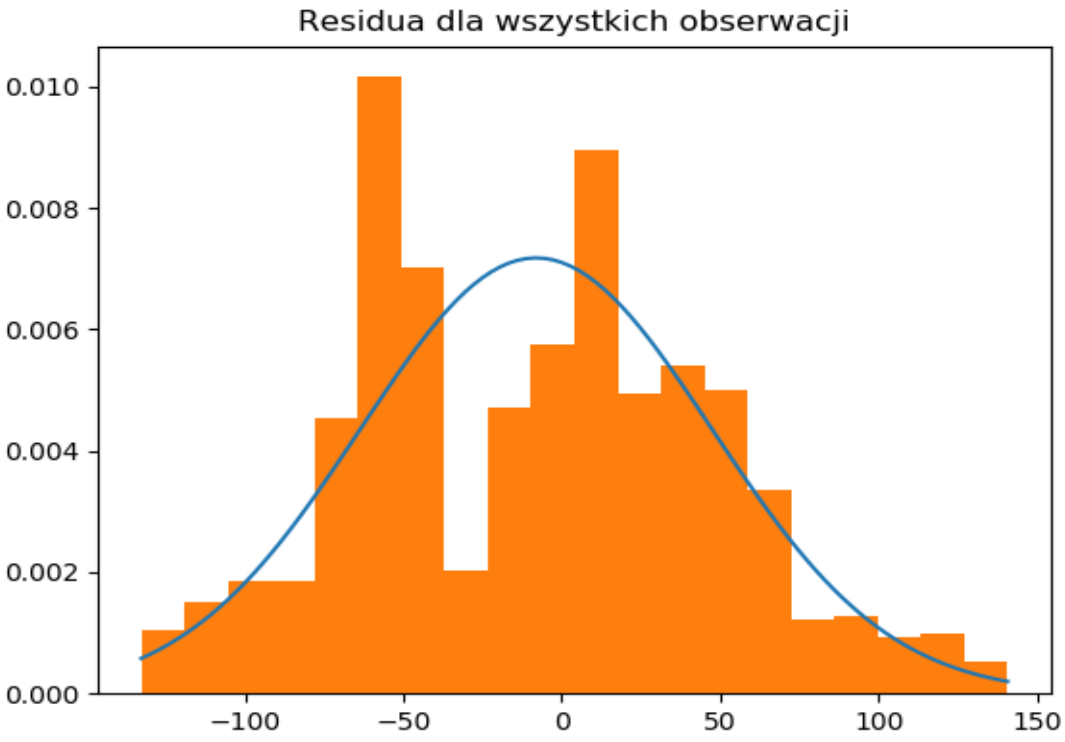


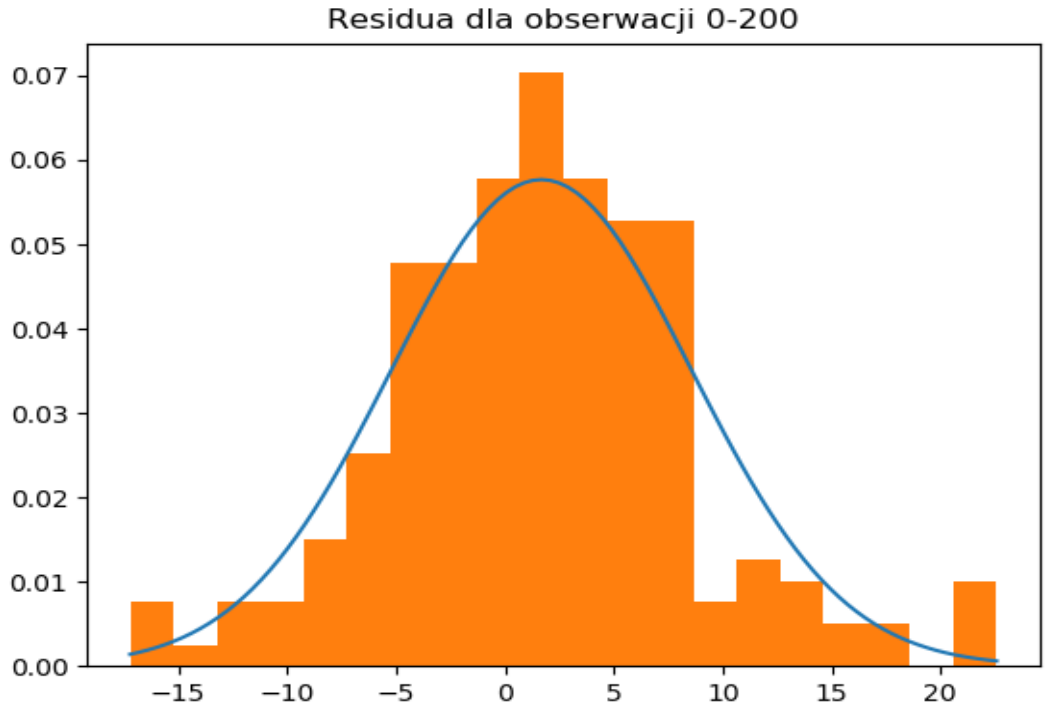



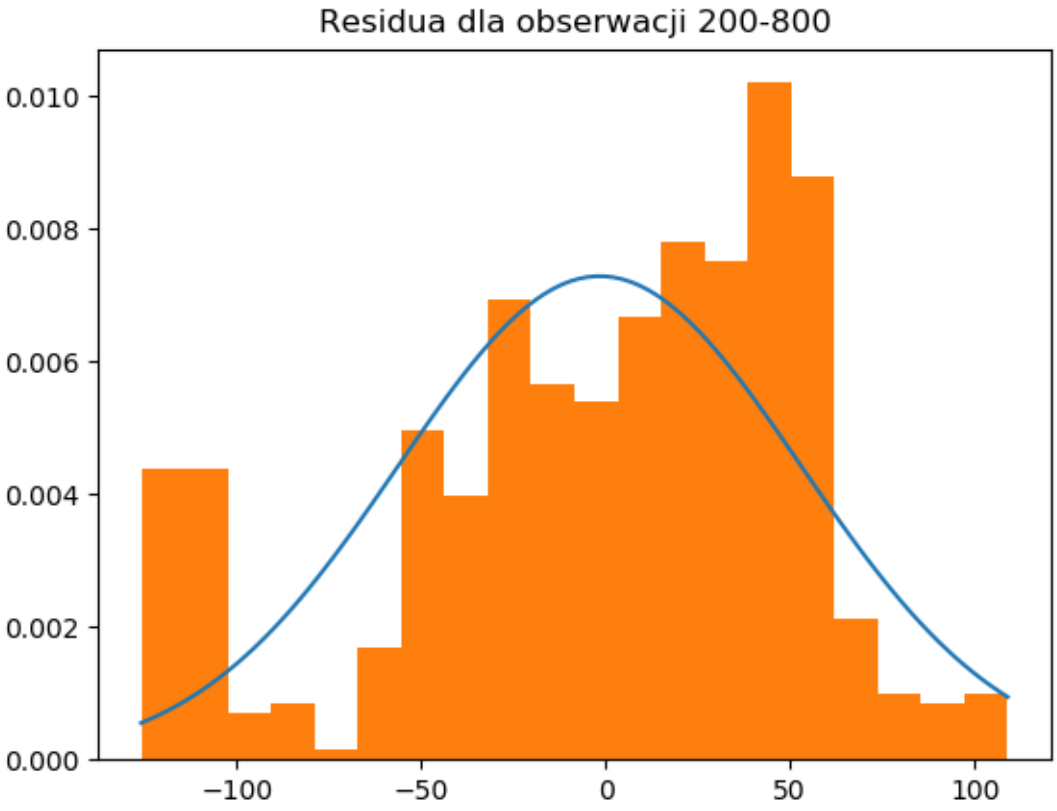

W tabeli poniżej są policzone wartości średnie, wariancje residuów oraz p-wartości dla testu Jarque-Bera. Tego testu używamy dla sprawdzenia, czy residua mają rozkład normalny.

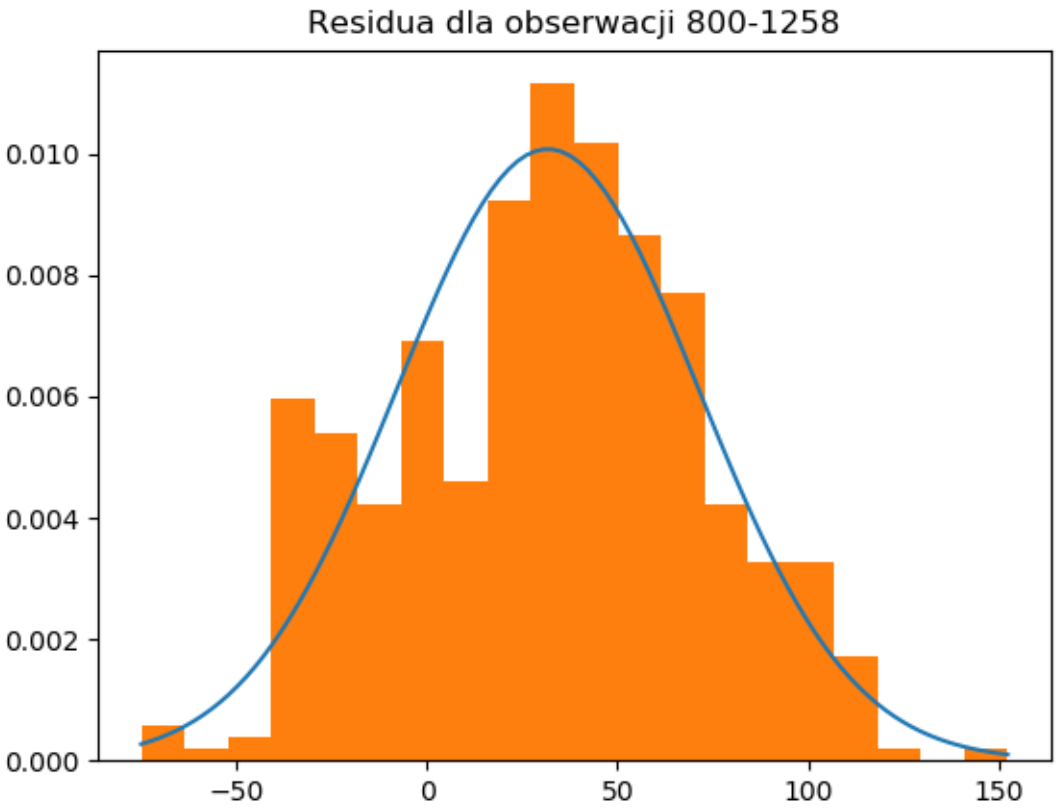
| | Ogólna grupa | Grupa 1 | Grupa 2 | Grupa 3 |
|---|---|---|---|---|
| Średnia | -8 | 1.7 | -1.6 | 31.7 |
| Wariancja | 3086 | 47.8 | 2993 | 1566 |
| p-wartość | 0.00003 | 0.02 | 0 | 0.09 |

Na wykresach poniżej widać porównanie gęstości rozkładu residuów z gęstością rozkładu normalnego.




Residua dla obserwacji 200-800
0.010
0.008
0.006
0.004
0.002
0.000
−100
−50
0
50
100


Residua dla obserwacji 800-1258
0.010
0.008
0.006
0.004
0.002
0.000
−50
0
50
100
150

Wnioski

Odrazu możemy powiedzieć, że rozdzielenie na grupy danych miało pozytywne skutki. Możemy na podstawie tego zobaczyć korelację między ceną na akcje Coca-Coli a ceną na akcje Netflixa, której nie było tak dobrze widać na wszystkich obserwacjach. Ze statystyk opisowych widzimy, że wariancje znacznie się zmniejszyły po podzieleniu danych na grupy. Oznacza to, że dane w poszczególnych grupach są bardziej skupione wokół pewnych wartości, co jest całkiem oczekiwane przy ciałości cen na akcje. Przy tym wartości średnie słabo się zmieniły. Przechodząc do analizy regresji liniowej, możemy zobaczyć, że dopasowanie prostej jest okropne na pełym okresie danych, jednak sytuacja nieco zmienia się przy podziale na grupy. Dobrze widac, że w pierwszym okresie czasowym dane dość dobrze dopasowują się do prostej regresji liniowej. Gorzej jest w grupie 2, sytuacja w tej grupie wygląda prawie tak samo jak i w ogólnej grupie. W grupie 3 dopasowanie nie jest idealne, jednak widać że te dane jednak trochę układają się w prostą. Współczynnik determinancji pokazuje zależność między danymi. Im bliżej on jest 1, tym o mocniejszej zależności możemy mówić. Ogólnie uważa się, że możemy rozmawiać w ogóle o jakiejś zależności, kiedy ten współczynnik jest większy od 0.5. Na ogólnym zbiorze danych współczynnik jest akurat trochę powyżej 0.5. W grupie 1 on znajduje się gdzieś na środku między 1 a 0.5. W grupie 2 współczynnik determinancji wnosi prawie 0, a w grupie 3 wynosi prawie 1. Stąd możemy wywnioskować, że na ogólnym zbiorze danych występuje słaba zależność między cenami na akcje, w 1 okresie zależność jest średnia, w 2 okresie tej zależności nie ma, w 3 okresie zależność jest dość mocna. Przedziały ufności są dość krótkie tylko w grupie 1, natomiast w pozostałych grupach te przydziały są bardzo szyrokie, co mówi o tym, że tylko w 1 okresie czasowym zbiór danych dobrze dopasowuje się do prostej regresji liniowej. To samo możemy wywnioskować z analizy residuów. Z boxplotów oraz wartości wariancji widzimy, że wariancja jest mała tylko dla grupy 1, dla pozostałych jest bardzo duża, a dla grupy 3 nawet nie układa się wokół zera. Dla pozostałych jednak residua układają się wokół zera. Residua w grupie 3 mają rozkład normalny. Mówią o tym zarówno test Jarque-Bera jak i porównanie gęstości. W grupie 1 gęstość wygląda podobno do rozkładu normalnego, jednak p-wartość testu Jarque-Bera sugeruje odrzucenie hipotezy o tym, że residua mają rozkład normalny. W ogólnej grupie oraz grupie 2 residua nie mają rozkładu normalnego. Przechodząc do wniosków ogólnych, chcę powiedzieć, że widać pewne zależności między ceną na akcje Coca-Coli oraz Netflixa. Jednak te zależności występują nie wszędzie. Patrząc na wszystkie dane za ostatnie 5 lat dość trudno je odnaleźć. Jednak jeśli podzielimy ten okres na trzy krótsze, będzie widać, że w pierwszym oraz 3 okresie są pewne zależności między tymi cenami. Jednak na drugim okresie nie widzimy prawie żadnych zależności. Warto też zwrócić uwagę na ten fakt, że warto zawsze stosować kilka różnych metod analizy danych, ponieważ one mogą doprowadzać do różnych wniosków. Warto też zadać sobie pytanie, dlaczego dane są skorelowane tylko w pierwszym i trzecim okresie czasowym, jednak są niezależne w drugim? Można to wyjaśnić tym, że ogólnie giełda papierów wartościowych zwykle zachowuje się w takim sam sposób dla wszystkich dużych firm, a dokładnie ceny na ich akcje zwykle rośną z czasem. Jednak w pewnym okresach z pewnych powodów poszczególne firmy mogą znajdować się w kryzysie, wtedy prawdopodobnie nie zobaczymy żadnych zależności między różnymi firmami. Jednak warto przeprowadzić dodatkowe badania (niekoniecznie statystyczne) dla udzielenia lepszej odpowiedzi na to pytanie.

## Użyte wzory i definicje

Średnia z danych jest zwykłą średnią arytmetyczną.

Mediana - wartość cechy w szeregu uporządkowanym, powyżej i poniżej której znajduje się jednakowa liczba obserwacji. Ona jest też drugim kwartylem.

Wariancja jest średnią arytmetyczną kwadratów odchyleń (różnic) poszczególnych wartości cechy od wartości oczekiwanej.

Regresja liniowa - w modelowaniu statystycznym, metody oparte o liniowe kombinacje zmiennych i parametrów dopasowujących model do danych. Dopasowana linia lub krzywa regresji reprezentuje oszacowaną wartość oczekiwaną zmiennej. Przy konkretnych wartościach innej zmiennej lub zmiennych .

**Współczynnik determinacji $R^2$** – jedna z miar jakości dopasowania modelu do danych uczących. Obecnie, współczynnik determinacji wykorzystuje się głównie w celach pomocniczych.

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

Gdzie $\hat{Y}_i$ jest dopasowaniem regresji liniowej, $\bar{Y}$ jest średnią z Y.

Współczynniki prostej regresji liniowej:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

Residuum – błąd dopasowania prsotej regresji liniowej

$$e_i = Y_i - \hat{Y}_i$$