WOJSKOWA AKADEMIA TECHNICZNA

im. Jarosława Dąbrowskiego

WYDZIAŁ CYBERNETYKI

Metody Eksploracji Danych

Laboratorium 1

Osoba realizująca: Piotr Gdula	**Prowadzący: Romuald Hoffmann**
Grupa: WCY21IJ1S1	**Data ćwiczenia: 09.11.2023**

ZADANIE 1

Pytania badawcze, założenie:

1. Przeanalizować model zależności przychodów firmy w stosunku do zatrudnienia wraz z wyznaczeniem parametrów liniowych.
2. Przeanalizować model zależności liczby użytkowników w danych latach wraz z wyznaczeniem parametrów liniowych oraz predykcją na następne lata.
3. Przeanalizować model zależności przychodów w danych latach, przekształcić model tej funkcji nieliniowej w postać liniową wraz z wyznaczeniem parametrów oraz predykcją na następne lata.

Przygotowanie danych:

1. Wczytanie oraz inspekcja danych:

	rok	kwartal	liczba_uzytkownikow_w_mln
0	2008	1	NaN
1	2008	2	NaN
2	2008	3	100.0
3	2008	4	NaN
4	2009	1	197.0

	rok	przychod_w_mln	zysk_w_mln	zatrudnienie
0	2007	153	-138	450
1	2008	272	-56	850
2	2009	777	229	1218
3	2010	1974	606	2127
4	2011	3711	1000	3200

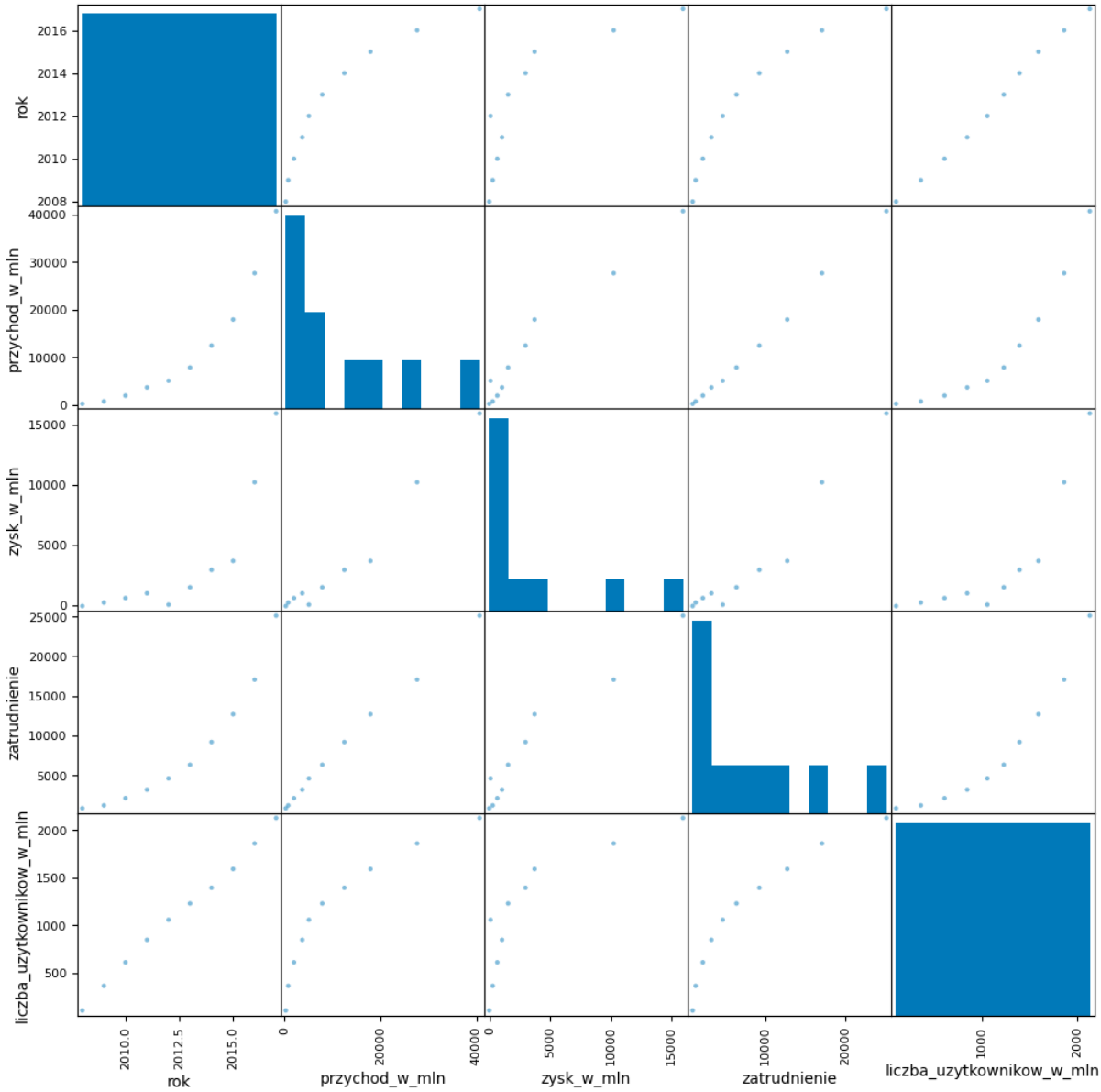
2. Przygotowanie dataframe'ów do połączenia oraz uzupełnienie danych:

	rok	przychod_w_mln	zysk_w_mln	zatrudnienie	liczba_uzytkownikow_w_mln
0	2008	272	-56	850	100.0
1	2009	777	229	1218	360.0
2	2010	1974	606	2127	608.0
3	2011	3711	1000	3200	845.0
4	2012	5089	53	4619	1056.0
5	2013	7872	1500	6337	1228.0
6	2014	12466	2940	9199	1393.0
7	2015	17928	3688	12691	1591.0
8	2016	27638	10217	17048	1860.0
9	2017	40653	15934	25105	2129.0

Dla roku 2007 nie było podanych wartości co do liczny użytkowników, dlatego postanowiłem usunąć ten rok z dalszych badań. Na koniec roku 2008 nie było podanych danych, ile użytkowników wtedy było, dlatego została tam wpisana wartość z końca 3 kwartału tego samego roku

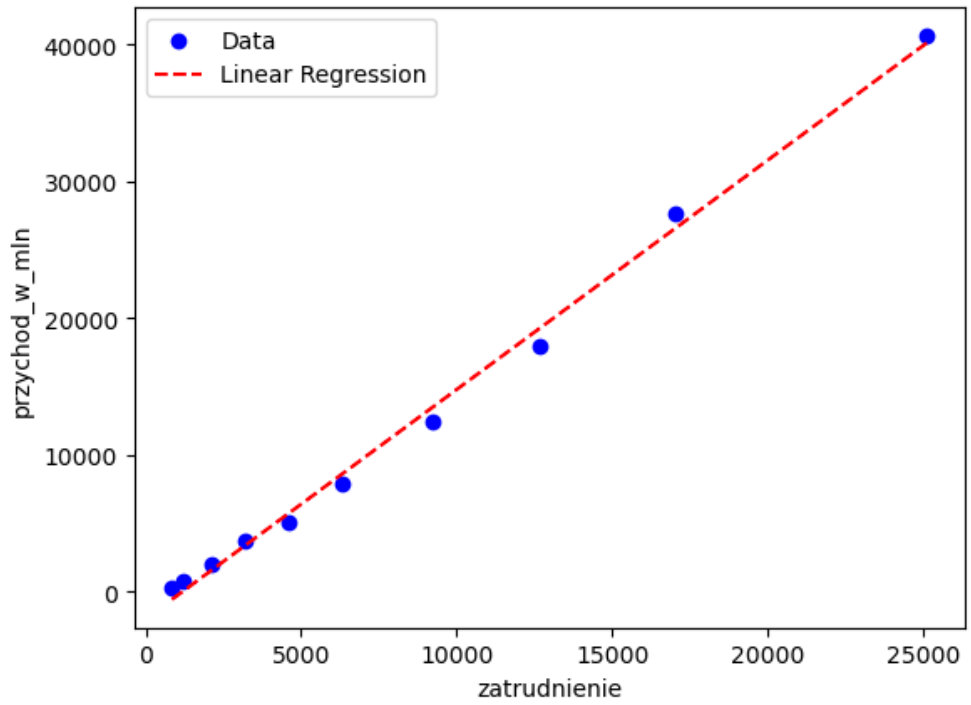
Badanie danych:

1. Wykres krzyżowania się danych:



Z wykresu na pierwszy rzut oka widać, że zależność **liczby użytkowników w danych latach** od **zatrudnienia od przychodów** wyglądają na zależność liniową. Ponad to zależność **przychodów w danych latach** wygląda jak funkcja wykładnicza, którą postaram się przekształcić w równanie liniowe.

Wykres 1: Zależność przychodów od zatrudnienia w latach 2008-2017



Własności liniowe:

slope: 1.6765939768483806

intercept: -1976.128412844546

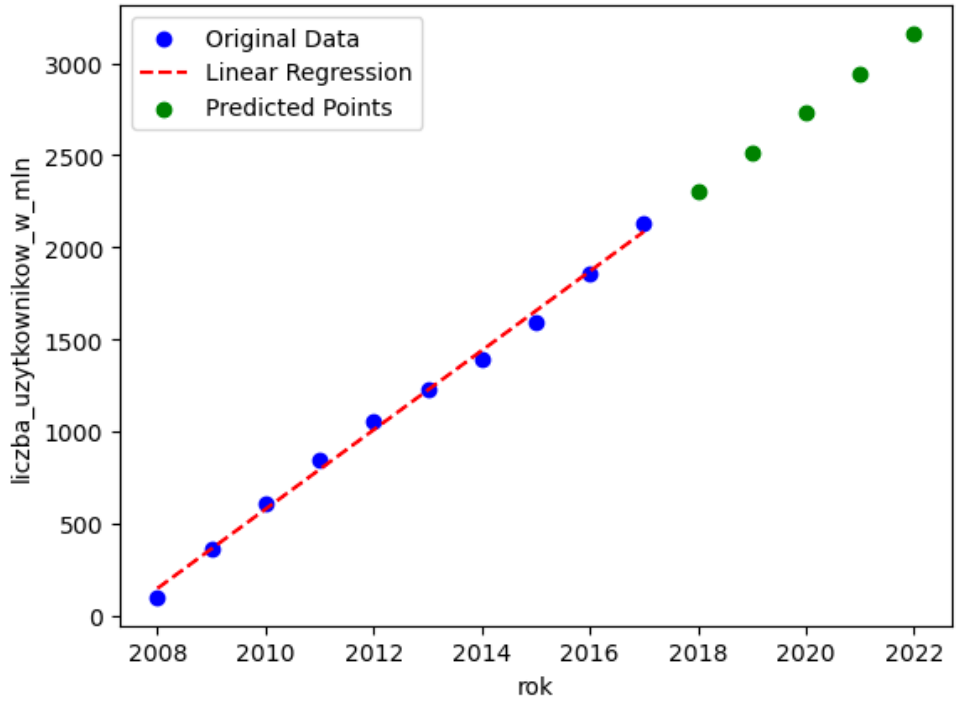
r^2: 0.9958073517828487

p_value: 8.46333266119875e-11

Wnioski:

Z wyznaczonych własności liniowych widać, że punkty mają korelacje liniową. Można również wstępnie zauważyć sinusoidalność punktów względem wyznaczonej prostej.

Wykres 2: Liczba użytkowników w latach 2008-2017 wraz z predykcjami na następne lata



Wlasnosci liniowe:

slope: 215.10303030303032

intercept: -431777.8484848485

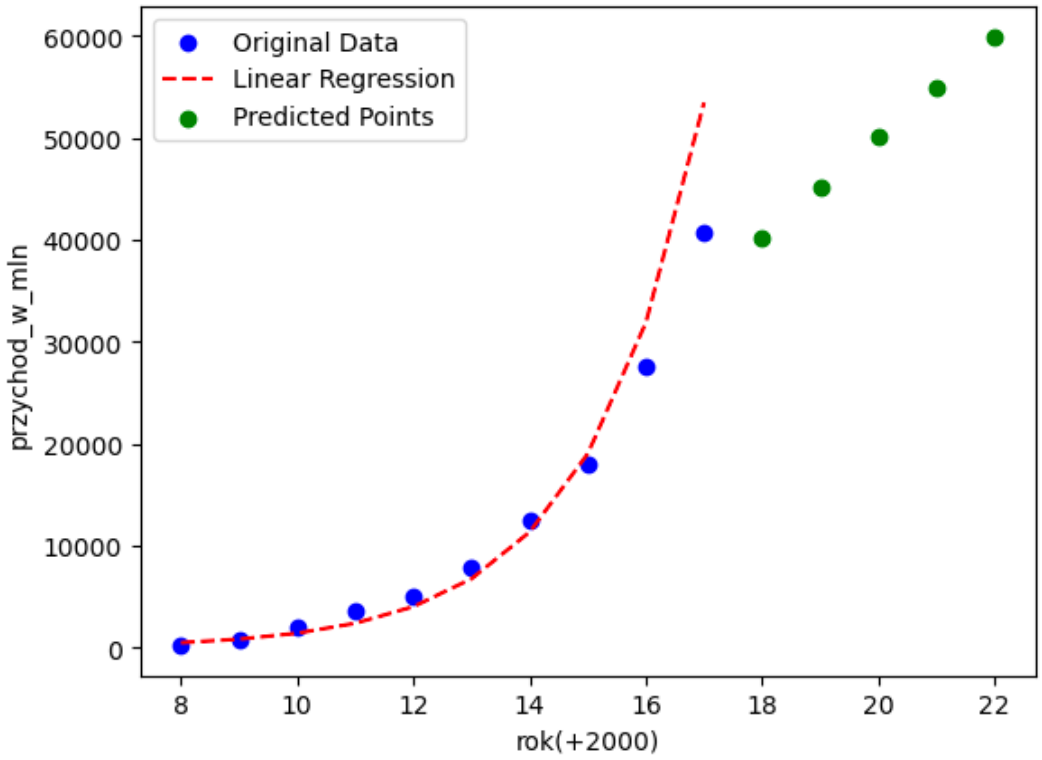
r^2: 0.9957554019333756

p_value: 8.890845390981214e-11

Wnioski:

Z wyznaczonych własności liniowych widać, że punkty mają korelacje liniową, chociaż współczynnik r^2 jest mniejszy niż przy poprzednim wykresie. Tutaj też można również wstępnie zauważyć sinusoidalność punktów względem wyznaczonej prostej. Zastosowana została najprostszy sposób predykcji następnych punktów, dlatego przewidywane punkty leżą na przedłużeniu wyznaczonej prostej.

Wykres 3: Przychody w latach 2008-2017 wraz z próbą predykcji na następne lata



Wlasnosci liniowe:

slope: 4917.171588114933

intercept: -48282.74752935466

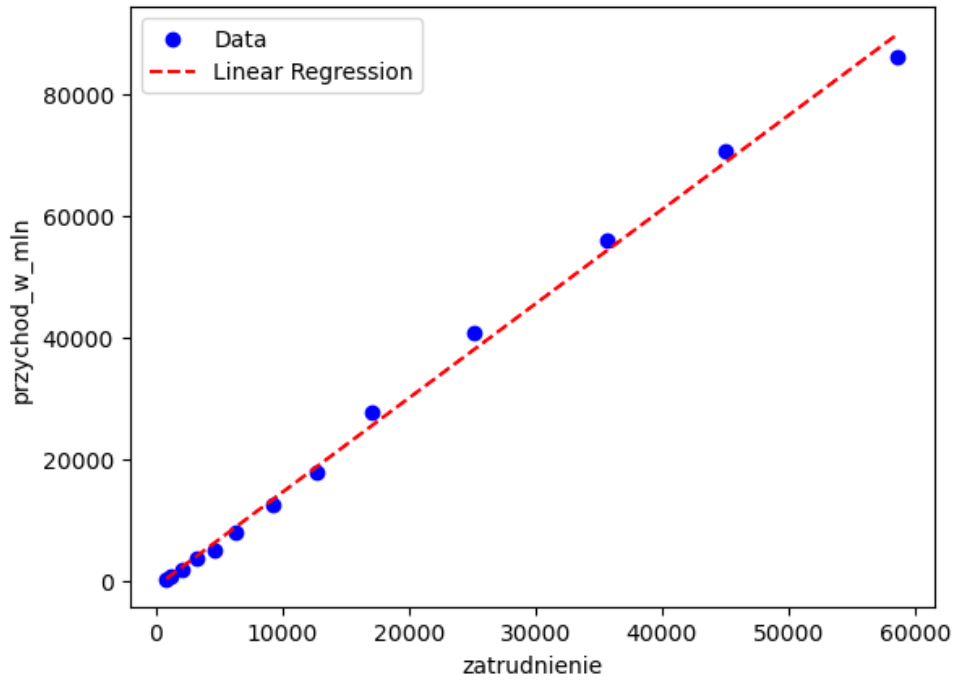
r^2: 0.7386640765938253

p_value: 0.0014356217146781344

Wnioski:

Po kilku próbach przekształcenia funkcji nieliniowej do postaci liniowej najlepszym rozwiązaniem okazało się to ze strony: https://www.geeksforgeeks.org/how-to-do-exponential-and-logarithmic-curve-fitting-in-python/. Rok został pomniejszony o 2000, gdyż przed pomniejszeniem rok powodował błędy przy obliczeniach, ale nie ma to wpływu na wyznaczoną liniowość funkcji. Predykcja punktów została zaimplementowana tak samo jak na wykresie 2, przez co przy funkcji pierwotnie nieliniowej, tylko przekształconej do takowej, punkty nie wydają się dobrze odzwierciedlać następnych punktów.

Wykres 4: Zależność przychodów od zatrudnienia w latach 2008-2020



Wlasnosci liniowe:

slope: 1.5482910287612728

intercept: -931.4051329537288

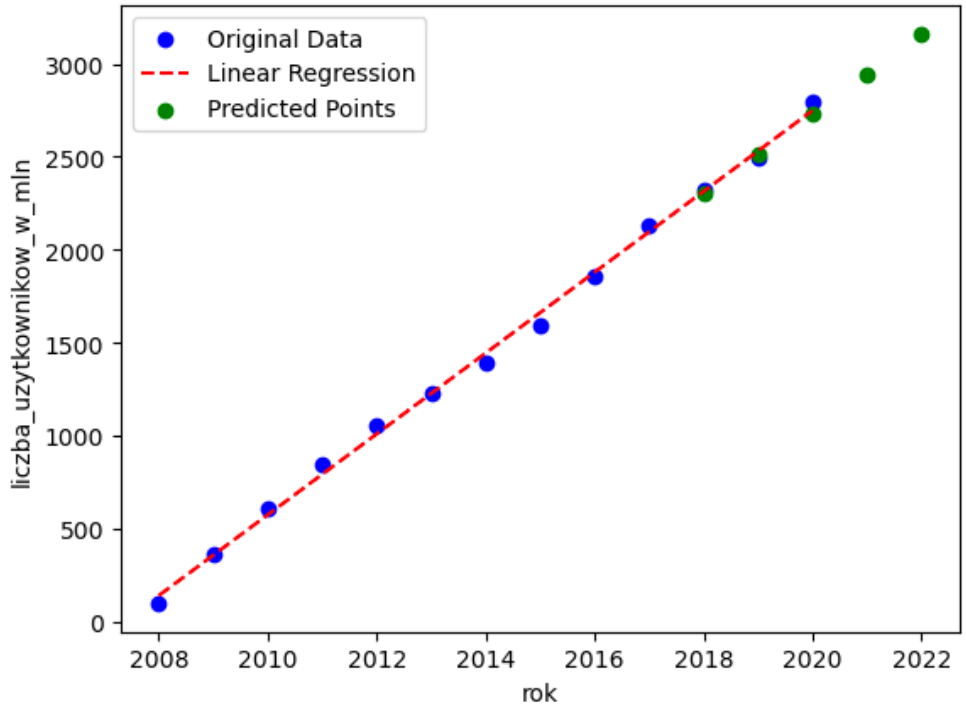
r^2: 0.9962576124082747

p_value: 1.0577762594876189e-14

Wnioski:

Po dodaniu wartości z lat 2018-2020 można zauważyć, że współczynnik liniowości wzrósł względem modelu z wykresu 1, oraz sinusoidalność względem wyznaczonej prostej, którą można było zauważyć, nadal występuje.

Wykres 5: Liczba użytkowników w latach 2008-2020 wraz z predykcjami



Wlasnosci liniowe:

slope: 217.26923076923077

intercept: -436135.23076923075

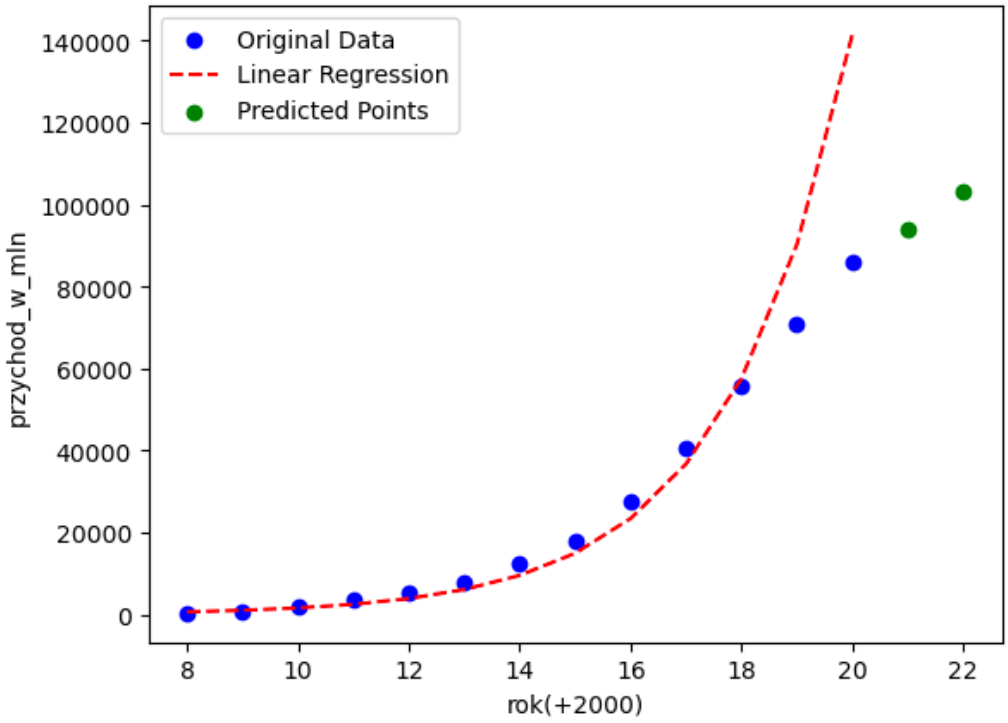
r^2: 0.9976554126689156

p_value: 8.075759339503507e-16

Wnioski:

Po dodaniu wartości z lat 2018-2020 można zauważyć, że współczynnik liniowości wzrósł względem modelu z wykresu 2. Ponad to wcześniejsze predykcje w są bardzo bliskie faktycznych danych oraz wcześniej zauważona sinusoidalność względem wyznaczonej prostej nadal wystepuje.

Wykres 6: Przychody w latach 2008-2020 wraz z predykcjami



Wlasnosci liniowe:

slope: 9149.1420325169

intercept: -98137.66277497167

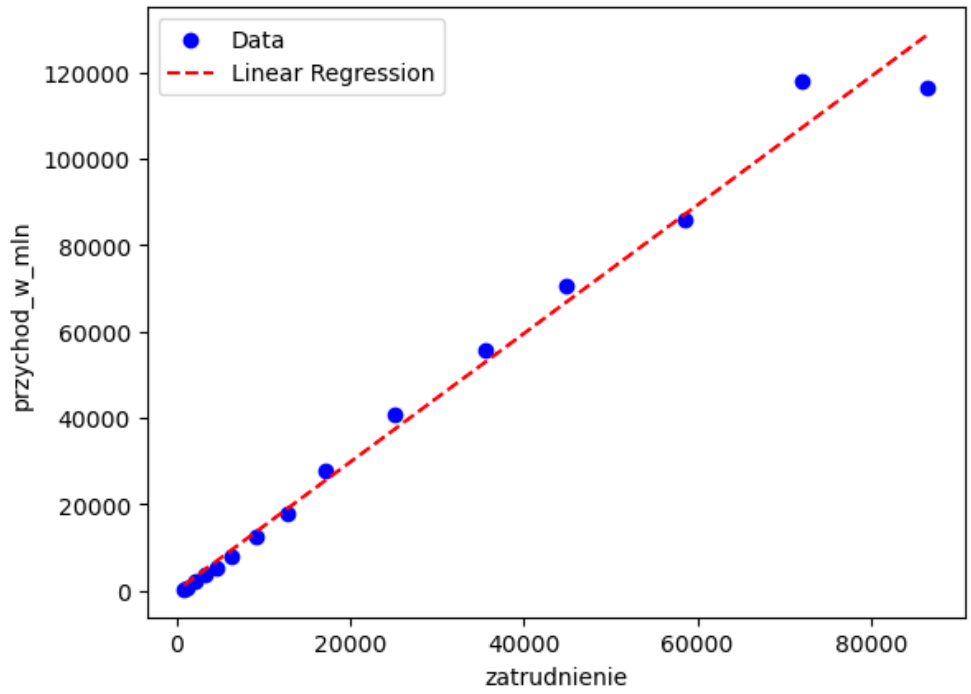
r^2: 0.690686534716997

p_value: 0.00043146329466444593

Wnioski:

Po dodaniu wartości z lat 2018-2020 zmalał współczynnik liniowości. Predykcja wartości z modelu na wykresie 3 nie miała większego sensu do sprawdzania jej z rzeczywistymi wartościami, dlatego jeszcze raz, teraz na obecnych danych, została przeprowadzona predykcja na następne lata.

Wykres 7: Zależność przychodów od zatrudnienia w latach 2008-2022



Wlasnosci liniowe:

slope: 1.4888458581139832

intercept: -20.81068801954825

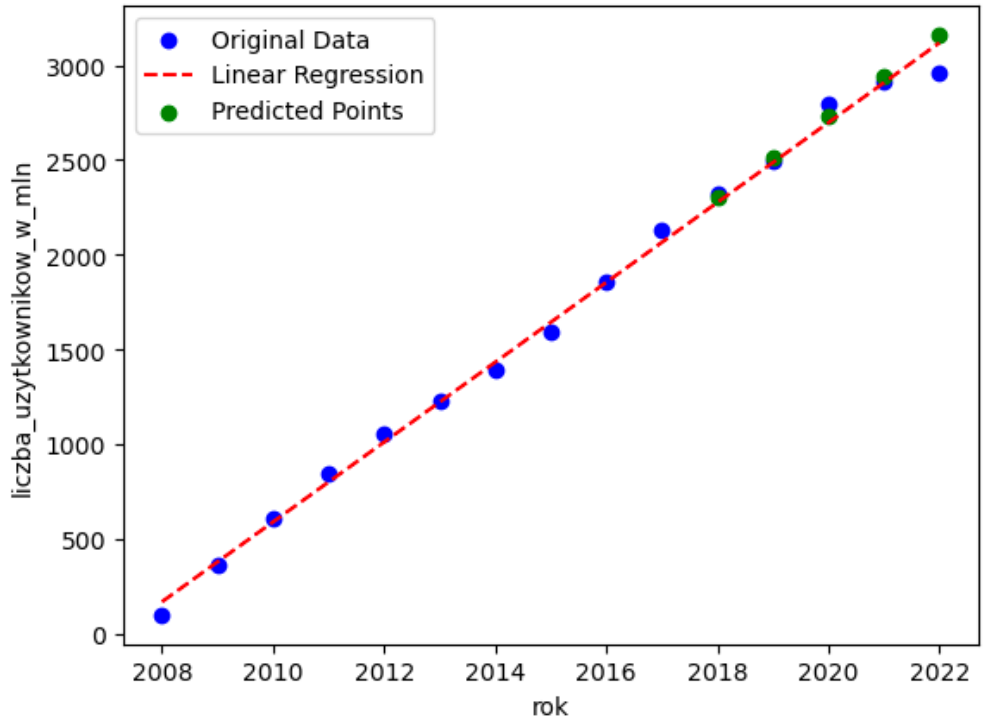
r^2: 0.9871899635830972

p_value: 1.0917589086765404e-13

Wnioski:

Po dodaniu wartości z lat 2021-2022 można zauważyć, że liniowość wykresu zmniejszyła się. Nadal można stwierdzić, że stworzony model jest dobry do predykcji przychodów od zatrudnienia na następne lata. Wcześniej zauważona sinusoidalność względem wyznaczonej prostej jest już tak zauważalna jak na model z mniejszą ilością danych.

Wykres 8: Liczba użytkowników w latach 2008-2022 wraz z predykcjami



Wlasnosci liniowe:

slope: 210.61071428571427

intercept: -422736.58928571426

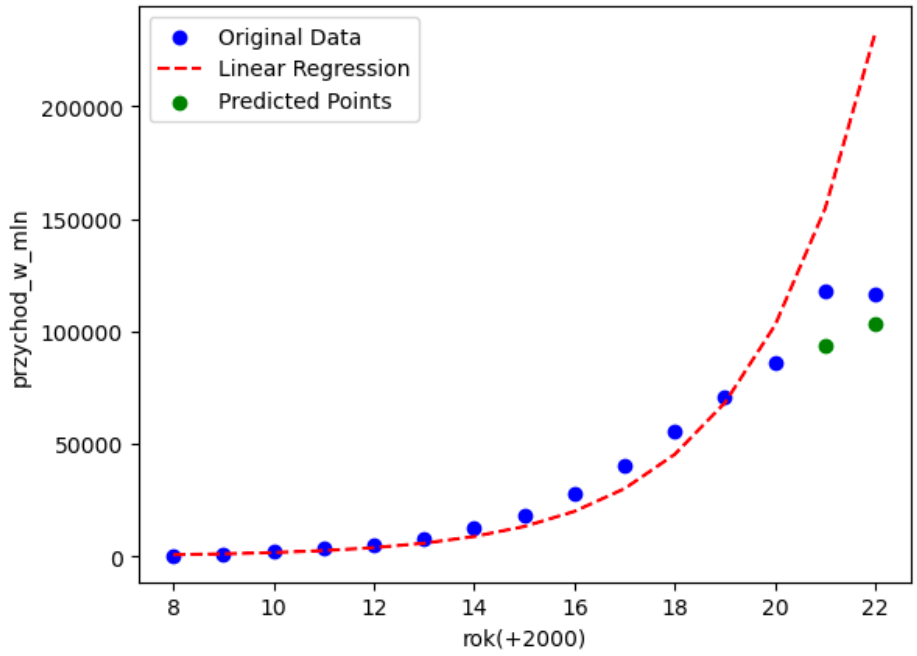
r^2: 0.9956596696011435

p_value: 9.579465849842614e-17

Wnioski:

Po dodaniu wartości z lat 2021-2022 można zauważyć, że liniowość wykresu zmniejszyła się w małym stopniu. Nadal można stwierdzić, że stworzony model jest dobry do predykcji przychodów od zatrudnienia na następne lata, gdyż predykcja, która była tworzona na podstawie danych z lat 2008-2017, dobrze odzwierciedla faktyczne dane.

Wykres 9: Przychody w latach 2008-2020 wraz z predykcją



Wlasnosci liniowe:

slope: 12493.918801730857

intercept: -141252.91181165312

r^2: 0.6712416658176005

p_value: 0.0001860337934117293

Wnioski:

Po dodaniu wartości z lat 2021-2022 można zauważyć, że liniowość wykresu zmniejszyła sie. Predykcja na podstawie danych z modelu na wykresie 6, tak samo jak predykcja z modelu na wykresie 3, nie jest dobrym odzwierciedleniem faktycznych danych.

Wnioski końcowe:

- Dla modeli z wykresów 1,4,7 mogę stwierdzić, że udało mi się wyznaczyć odpowiednio model regresji liniowej dla podanych danych. Dane układały się w sposób sinusoidalny do wyznaczonej prostej, ale przy zwiększeniu próby badawczej zależność zanika. Własności liniowe przy tych 3 wykresach były na bardzo wysokim poziomie.

- Dla modeli z wykresów 2,5,8 mogę również stwierdzić, że udało mi się wyznaczyć odpowiednio model regresji liniowej dla podanych danych. Predykcja punktów z wykresu 2 dobrze odzwierciedlała faktyczne punkty. Dane układały się w sposób sinusoidalny do wyznaczonej prostej, ale przy zwiększeniu próby badawczej zależność zanika. Własności liniowe przy tych 3 wykresach były na bardzo wysokim poziomie.

- Dla modeli z wykresów 3,6,9 mogę stwierdzić, że udało mi się przekształcić funkcję nieliniową do takiej postaci, aby móc liczyć z niej własności regresji, oraz odpowiednio wyznaczyć jej własności. Predykcja zastosowana przy wyznaczaniu następnych wartości nie pokrywała się z faktycznymi danymi i następnym razem zastosował inny sposób na predykcję. Po dodaniu danych z lat 2021-2022 widać, że funkcja nie wygląda już tak wykładniczo jak na początku.

ZADANIE 2

Pytania badawcze, założenie:

1. Zbudować model pozwalający przewidzieć %bodyfat na podstawie zmiennych objaśniających.
2. Zaproponować model pozwalający przewidzieć %bodyfat dla innych zmiennych objaśniających.
3. Zaproponować model pozwalający przewidzieć inną zmienną objaśnianą na podstawie zmiennych objaśniających.

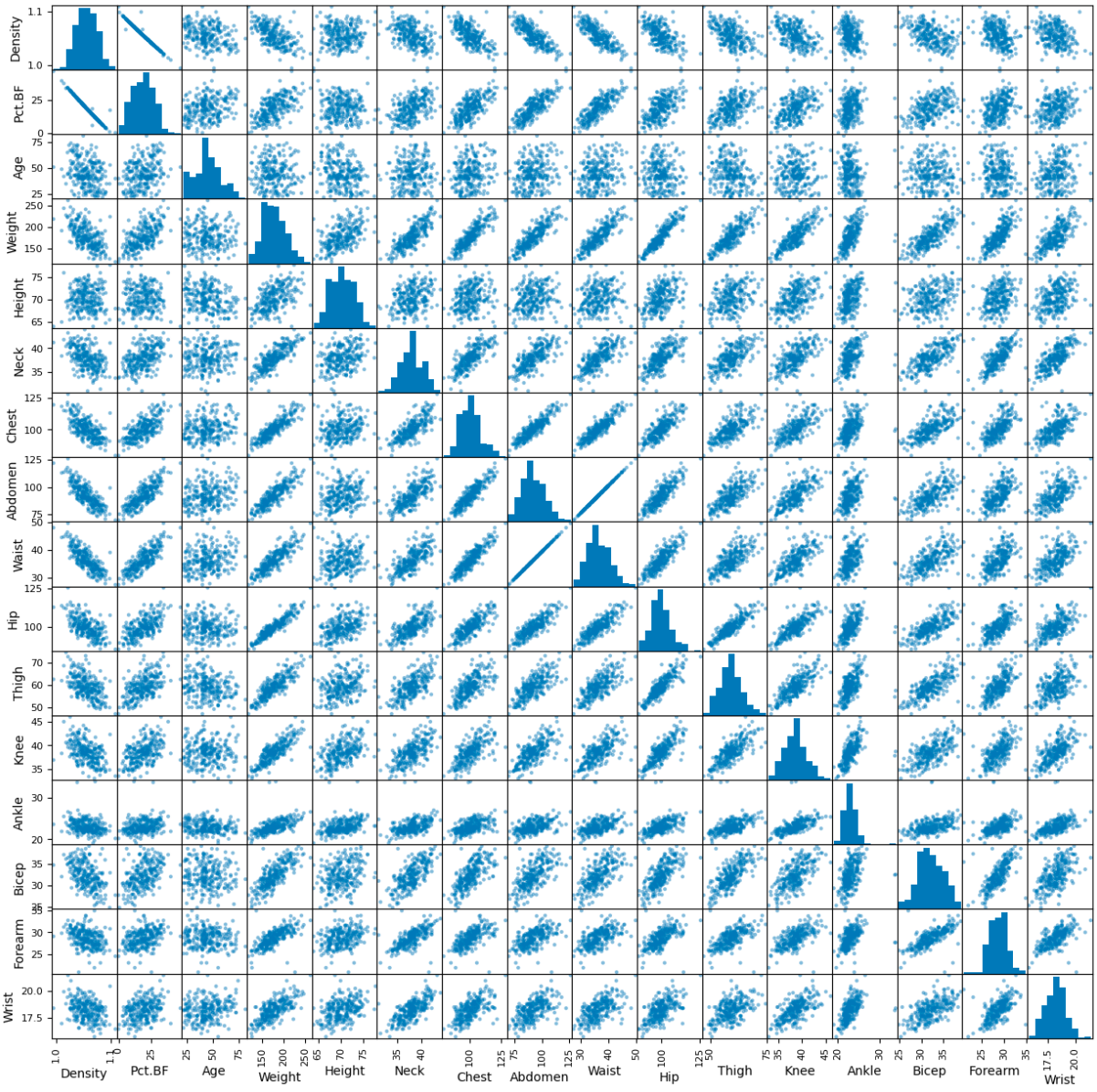
Przygotowanie danych:

	Density	Pct.BF	Age	Weight	Height	Neck	Chest	Abdomen	Waist	Hip	Thigh	Knee	Ankle	Bicep	Forearm	Wrist
0	1.0708	12.3	23	154.25	67.75	36.2	93.1	85.2	33.543307	94.5	59.0	37.3	21.9	32.0	27.4	17.1
1	1.0853	6.1	22	173.25	72.25	38.5	93.6	83.0	32.677165	98.7	58.7	37.3	23.4	30.5	28.9	18.2
2	1.0414	25.3	22	154.00	66.25	34.0	95.8	87.9	34.606299	99.2	59.6	38.9	24.0	28.8	25.2	16.6
3	1.0751	10.4	26	184.75	72.25	37.4	101.8	86.4	34.015748	101.2	60.1	37.3	22.8	32.4	29.4	18.2
4	1.0340	28.7	24	184.25	71.25	34.4	97.3	100.0	39.370079	101.9	63.2	42.2	24.0	32.2	27.7	17.7
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
245	1.0736	11.0	70	134.25	67.00	34.9	89.2	83.6	32.913386	88.8	49.6	34.8	21.5	25.6	25.7	18.5
246	1.0236	33.6	72	201.00	69.75	40.9	108.5	105.0	41.338583	104.5	59.6	40.8	23.2	35.2	28.6	20.1
247	1.0328	29.3	72	186.75	66.00	38.9	111.1	111.5	43.897638	101.7	60.3	37.3	21.5	31.3	27.2	18.0
248	1.0399	26.0	72	190.75	70.50	38.9	108.3	101.3	39.881890	97.8	56.0	41.6	22.7	30.5	29.4	19.8
249	1.0271	31.9	74	207.50	70.00	40.8	112.4	108.5	42.716535	107.1	59.3	42.2	24.6	33.7	30.0	20.9

250 rows × 16 columns

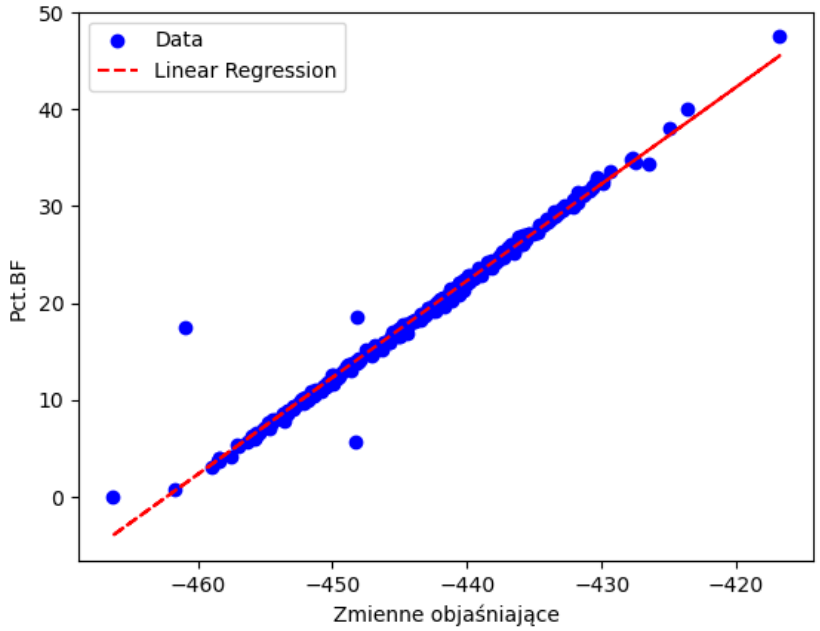
Badanie danych:

1. Wykres krzyżowania się danych:



Z wykresu na pierwszy rzut oka widać, że zależność **Density** od **Pct.BF** wygląda na zależność liniową. Ponad to **Weight** oraz **Height** są podstawowymi danymi do obliczania **Pct.BF**, dlatego też na podstawie tych trzech zmiennych objaśniających postaram się stworzyć model pozwalający przewidzieć **Pct.BF**. **Waist**, **Neck** oraz **Age** również są danymi, które bardzo często są wykorzystywane do obliczania **Pct.BF**, a dlatego, że ich zależność w stosunku do **Pct.BF** wydaje się względnie liniowa, na ich podstawie postaram się zbudować drugi model do przewidywania **Pct.BF**. W trzecim modelu postaram się przewidywać **Abdomen** na podstawie **Waist**, **Hip** oraz **Chest**.

Wykres 1: Zależność Pct.BF od Density, Weight oraz Height



Wlasnosci liniowe:

slope: 0.9999999999999786

intercept: 462.2960054253641

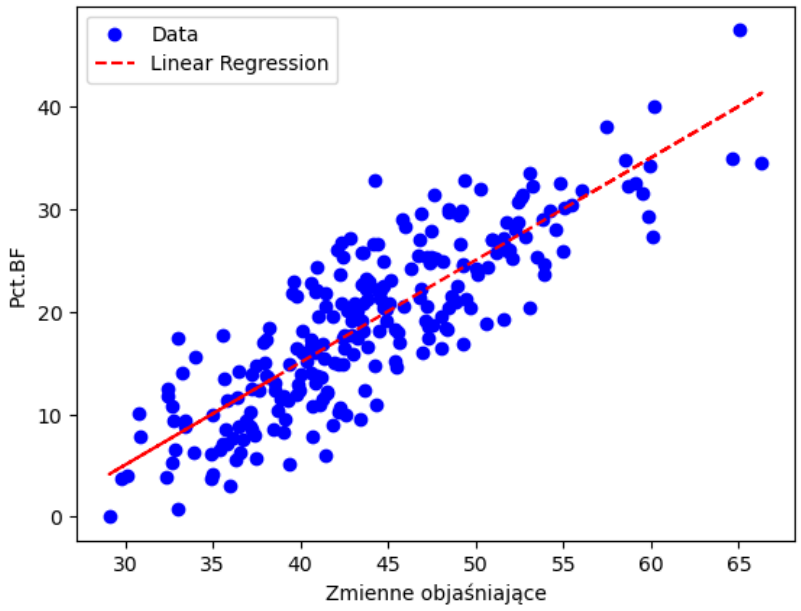
r^2: 0.976342813565221

p_value: 1.2046382164799312e-203

Wnioski:

Dobrane zmienne objaśniające są podstawowe przy wyliczaniu Pct.BF, dlatego też zostały użyte. Na wykresie widać, że oprócz pojedynczych punktów całą reszta układa się w sposób liniowy. Na podstawie wybranych danych objaśniających można śmiało przewidywać Pct.BF, gdyż korelacja liniowa wynosi ponad 0.97.

Wykres 2: Zależność Pct.BF od Waist, Neck oraz Age



Wlasnosci liniowe:

slope: 1.0000000000000004

intercept: -24.93279166161728

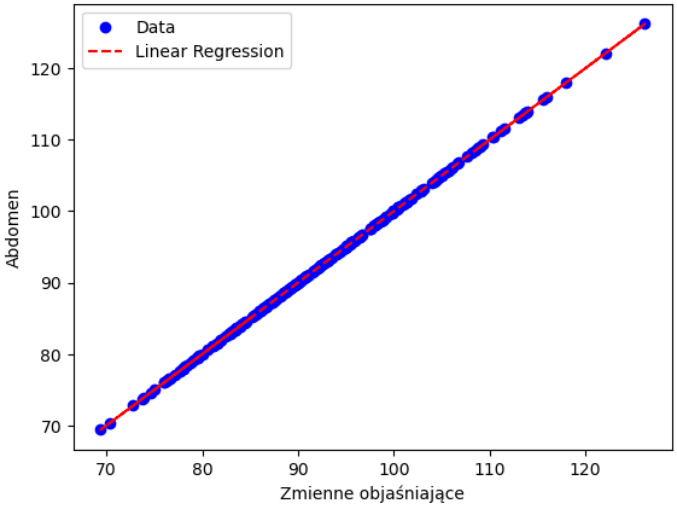
r^2: 0.7120281221662406

p_value: 5.454284442817139e-69

Wnioski:

Dobrane inne zmienne objaśniające są również bardzo często używane przy wyliczaniu Pct.BF, dlatego też zostały użyte. Na wykresie widać, że punkty oscylują wokół wyznaczonej prostej, bez punktów które widocznie odstawałyby od reszty. Na podstawie wybranych danych objaśniających można próbować przewidywać Pct.BF, gdyż korelacja liniowa wynosi ponad 0.71.

Wykres 3: Zależność Abdomen od Waist, Hip oraz Chest



Wlasnosci liniowe:

slope: 0.9999999999999992

intercept: -5.824783215757634e-07

r^2: 0.9999999999999944

p_value: 0.0

Wnioski:

Dobrane zmienne objaśniające do wyznaczenia Abdomen na pierwszy rzut oka na wykresie, gdzie krzyżowałem ze sobą wszystkie dane, układały się w sposób liniowy ze zmienną objaśnianą. Na wykresie widać, że punkty układają się w prostą, bez punktów które widocznie odstawałyby od reszty. Na podstawie wybranych danych objaśniających można śmiało przewidywać Abdomen, gdyż korelacja liniowa wynosi ponad 0.99.

Wnioski końcowe:

Dla danych pytań badawczych udało mi się:

- Zbudować model pozwalający przewidzieć %bodyfat na podstawie zmiennych objaśniających.
- Zaproponować model pozwalający przewidzieć %bodyfat dla innych zmiennych objaśniających.
- Zaproponować model pozwalający przewidzieć inną zmienną objaśnianą na podstawie zmiennych objaśniających.

Dla pierwszego i trzeciego punktu udało mi się wyznaczyć bardzo dobre modele do predykcji zmiennych objaśnianych na podstawie dobranych zmiennych objaśniających. Dla drugiego punktu udało mi się wyznaczyć odpowiedni model, aby w miarę dobrze przewidywać zmienną objaśnianą na podstawie zmiennych objaśniających.