**WOJSKOWA AKADEMIA TECHNICZNA**

Metody eksploracji danych

Labolatoria 1.1

Sprawozdanie

Temat: Facebook

**Prowadzący: dr inż. Romuald Hoffmann**

Grupa: WCY21IJ1N1

Wykonali: Paweł Karczewski, Adam Tarkowski

Data przeprowadzenia laboratorium: 19.11.2023

Data oddania sprawozdania: 13.12.2023

Treść zadania

Mamy zgromadzone dane dot. dobrze znanego przedsiębiorstwa internetowego Meta w tym odnoszące się do portalu „Facebook”, które zostały zawarte w poniższych tabelach. Dane te dotyczą liczby użytkowników (klientów) w rozliczeniu na kwartały w poszczególnych latach. Pozostałe dane to przychody i zysk liczone w milionach dolarów amerykańskich. Ostatnia wielkość to zatrudnienie.

W analizie proszę:

1) Proszę przeanalizować dane i zastanowić się nad tym „Co chcemy zbadać i dlaczego?”, tzn. na jakie pytania chcemy sobie odpowiedzieć.

2) W powyższym kontekście, na gruncie regresji, należy zaproponować model lub modele badające wybrane zależności i wyliczyć m. in. ich parametry strukturalne, odchylenia standardowe, miary dopasowania, czy też przebadać hipotezy, itp.

3) Analizę proszę na początek przeprowadzić na podstawie danych, które zostały przedstawione w powyższych tabelach.

4) Proponowane modele należy uzasadnić oraz wskazać ewentualny sposób ich praktycznego użycia w tym np. do predykcji.

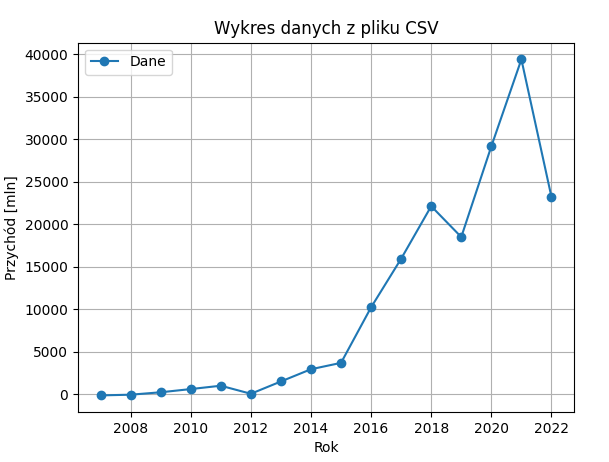
5) Na podstawie opracowanych modeli i przeprowadzonych obliczeń sformułować własne wnioski.

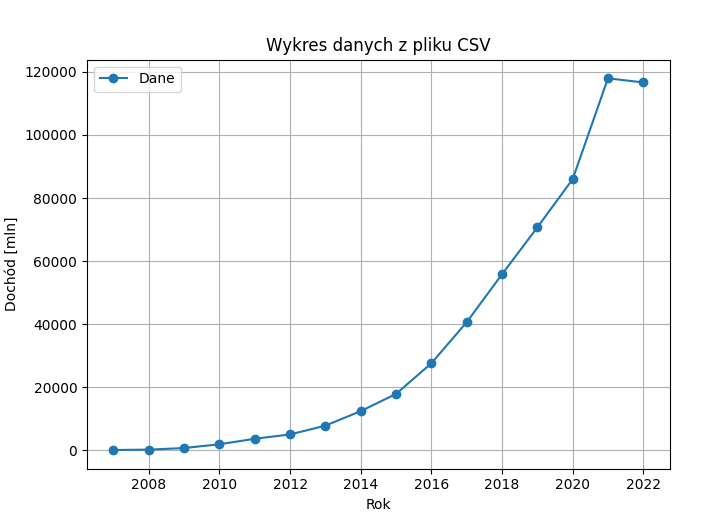
6) Następnie proszę odnaleźć w ogólnie dostępnych źródłach dane: a) za lata 2018-2020 i sprawdzić zachowanie się (wykorzystanie) modeli np. procesie predykcji za lata 2018-2020. b) za lata 2021-2022 i sprawdzić zachowanie się (wykorzystanie) modeli np. procesie predykcji za lata 2021-2022.

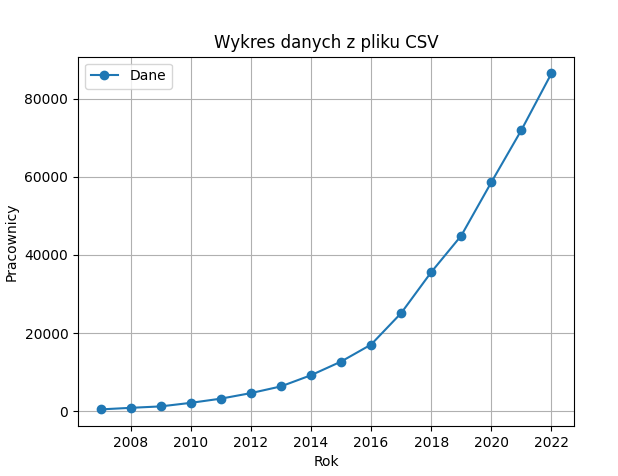
7) Wyniki swoich analiz zawierających: postawione pytania „badawcze”, hipotezy, wzory, udokumentowane wyniki obliczeń, wnioski, itp. proszę zawrzeć w postaci sprawozdania. Obliczenia można przeprowadzić w dowolnie wybranym narzędziu, np. które do tej pory wykorzystywaliśmy na zajęciach. Do sprawozdania proszę dodać jako załączniki wszystkie pliki z obliczeniami.

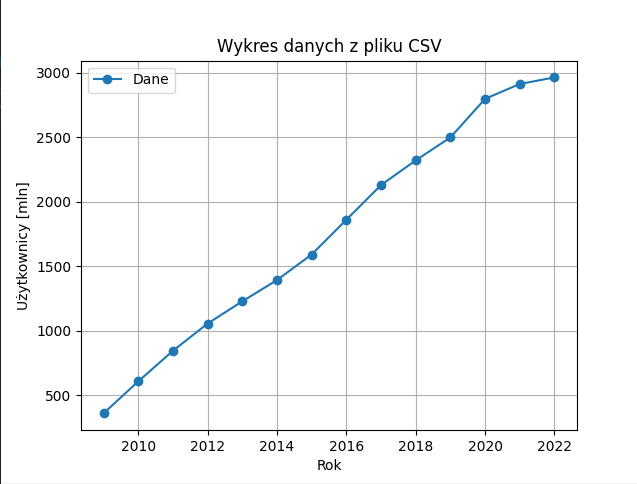
# Facebook – pytania

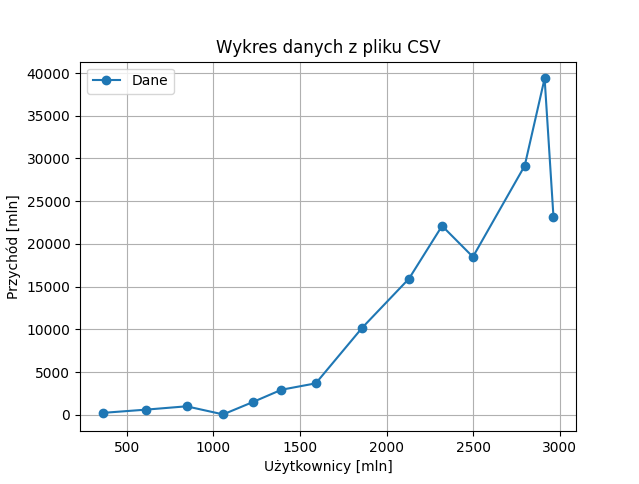
Proces analizy warto rozpocząć od przejrzenia danych w postaci graficznej. Pozwoliło nam to zaobserwować zależność między zmiennymi.

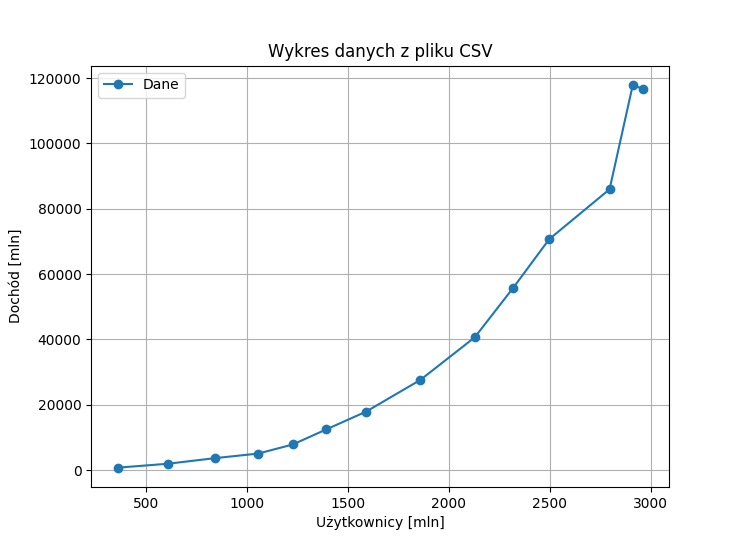


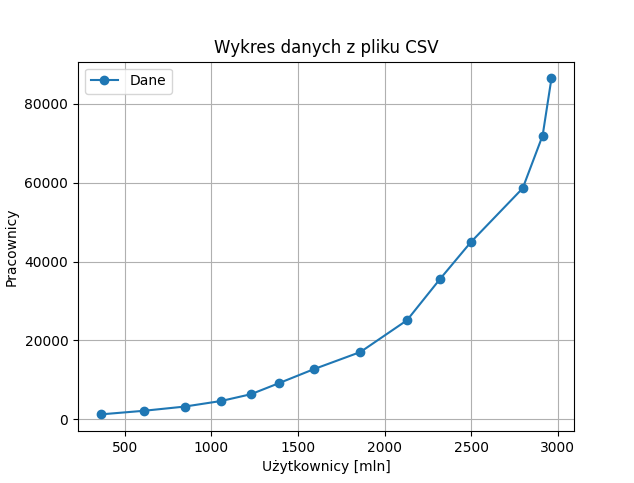


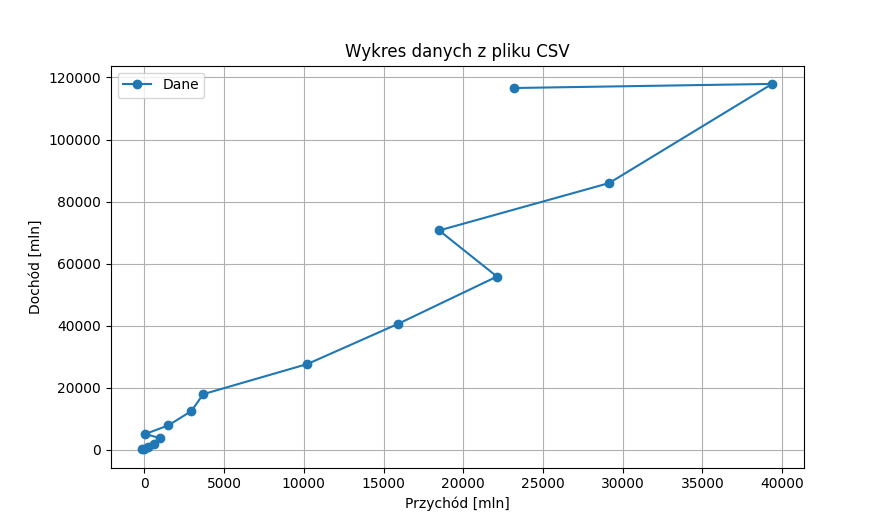


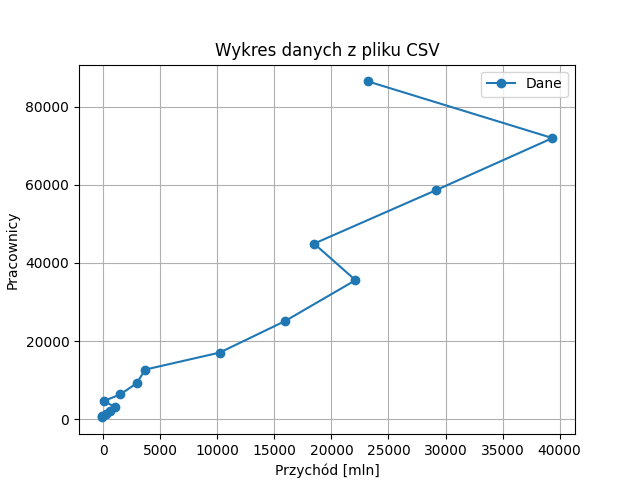


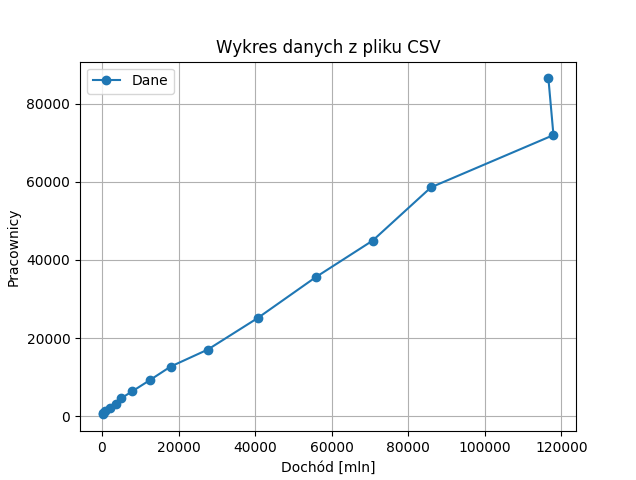












Na podstawie tak zwizualizowanych danych możemy postawić pewne hipotezy:

1. Ilość użytkowników rośnie w sposób liniowy względem czasu.
2. W sposób wykładniczy względem ilości użytkowników rosną takie wartości jak ilość pracowników, przychód oraz dochód.
3. Ilość pracowników, przychód oraz dochód są względem siebie zależne.

Potwierdzenie danych hipotez dałoby nam podstawę pod wywnioskowanie danych w przyszłości. Wartościowe pytania które można sobie zadać to:

1. Ile będziemy mieli użytkowników w kolejnych latach? Pomogłoby nam to w tworzeniu planów biznesowych oraz stwarzaniu infrastruktury, która byłaby w stanie udźwignąć ruch na naszym systemie.
2. Jaki będzie przychód zakładając konkretną liczbę użytkowników? Dałoby nam to podstawy do oszacowania ile będzie warta nasza firma w przypadku konkretnej ilości użytkowników.
3. Ilu pracowników będziemy mieli przy danej ilości użytkowników? Pomogłoby nam to zoptymalizować procesy rekrutacyjne.

# Facebook – Modele startowe

Na początku możemy stworzyć pierwsze modele na podstawie ograniczonej ilości danych. Mamy więc dane z lat 2008-2017 dla ilości użytkowników oraz 2007-2017 dla ilości zatrudnionych pracowników, przychodu oraz dochodu. Warto jest sprawdzić czy wygenerowane początkowo modele dadzą nam podstawę pod wnioskowanie przyszłości. Dane chcielibyśmy podzielić na 2 grupy: uczące I weryfikujące. Te pierwsze będą potrzebne do wygenerowania modelu jaki mamy nadzieje będzie oddawał rzeczywistość w tym tą z przyszłości. Do weryfikacji czy modele mają wartość posłużą nam dane z drugiej grupy.

# Facebook – Przyrost liczby użytkowników

Przyrost liczby użytkowników wydaje się być liniowy względem czasu. Przy użyciu biblioteki scikit-learn w środowisku python wyliczona została regresja liniowa przy użyciu metody najmniejszych kwadratów. Zmienną objaśniającą jest u nas czas, który został zamieniona do wartości numerycznej jako ilość milisekund od 1 stycznia 1970 r.   
  
Przy danych uczących do 2015 roku współczynnik determinacji wyszedł bardzo wysoki bo: 0.99. Wskazuje to, że model był bardzo dobrze dobrany, a przy testowaniu i ocenie predykcji można stwierdzić, że w miarę dobrze przewiduje przyszłość bo średni absolutny błąd procentowy nie przekracza 5%.

Co ciekawe model uczony na szerszych zakresach nie był mocno lepszy. Tzn przy uczeniu go do 2017 miał co prawda lepszą predykcję o ok 1 punkt procentowy dla lat 2018-2020, ale już uczenie do 2020 dało gorsze wyniki dla lat 2021-2022.

Wynik działania programu:

Trening modelu 2008-2015

Współczynnik determinacji: 0.9939773540699818

Testowanie modelu 2016-2017:

Date Year Quarter Users\_in\_mln GeneratedByModel Error Error\_squared

29 2016-01-01 2016 1 1654 1685.861154 -31.861154 1015.133105

30 2016-04-01 2016 2 1712 1737.380917 -25.380917 644.190953

31 2016-07-01 2016 3 1788 1788.900681 -0.900681 0.811226

32 2016-10-01 2016 4 1860 1840.986595 19.013405 361.509551

33 2017-01-01 2017 1 1936 1893.072510 42.927490 1842.769371

34 2017-04-01 2017 2 2006 1944.026123 61.973877 3840.761476

35 2017-07-01 2017 3 2072 1995.545886 76.454114 5845.231517

36 2017-10-01 2017 4 2129 2047.631801 81.368199 6620.783805

Błąd średni: 27.94929162863582

Błąd średni bezwzględny: 42.4849794568222

Błąd średniokwadratowy: 2521.3988753729413

Pierwiastek błędu średniokwadratowego: 50.21353279120024

Odchylenie standardowe błedu: 41.71613564114501

Średni absolutny błąd procentowy: 2.16249519659545 %

Predykcja 2018-2020:

Date Year Quarter Users\_in\_mln GeneratedByModel Error Error\_squared

37 2018-01-01 2018 1 2196 2099.717716 96.282284 9270.278243

38 2018-04-01 2018 2 2234 2150.671328 83.328672 6943.667551

39 2018-07-01 2018 3 2271 2202.191092 68.808908 4734.665857

40 2018-10-01 2018 4 2320 2254.277007 65.722993 4319.511868

41 2019-01-01 2019 1 2375 2306.362921 68.637079 4711.048563

42 2019-04-01 2019 2 2414 2357.316534 56.683466 3213.015353

43 2019-07-01 2019 3 2449 2408.836297 40.163703 1613.123018

44 2019-10-01 2019 4 2498 2460.922212 37.077788 1374.762357

45 2020-01-01 2020 1 2603 2513.008127 89.991873 8098.537225

46 2020-04-01 2020 2 2701 2564.527890 136.472110 18624.636681

47 2020-07-01 2020 3 2740 2616.047654 123.952346 15364.184070

48 2020-10-01 2020 4 2797 2668.133569 128.866431 16606.557077

Błąd średni: 82.99897109018684

Błąd średni bezwzględny: 82.99897109018684

Błąd średniokwadratowy: 7906.165655288348

Pierwiastek błędu średniokwadratowego: 88.91662192913284

Odchylenie standardowe błedu: 31.89571214534452

Średni absolutny błąd procentowy: 3.331720581758577 %

Predykcja 2021-2022:

Date Year Quarter Users\_in\_mln GeneratedByModel Error Error\_squared

49 2021-01-01 2021 1 2853 2720.219484 132.780516 17630.665517

50 2021-04-01 2021 2 2895 2771.173096 123.826904 15333.102156

51 2021-07-01 2021 3 2910 2822.692860 87.307140 7622.536771

52 2021-10-01 2021 4 2912 2874.778774 37.221226 1385.419637

53 2022-01-01 2022 1 2936 2926.864689 9.135311 83.453903

54 2022-04-01 2022 2 2934 2977.818302 -43.818302 1920.043548

55 2022-07-01 2022 3 2958 3029.338065 -71.338065 5089.119531

56 2022-10-01 2022 4 2963 3081.423980 -118.423980 14024.239018

Błąd średni: 19.586343833729643

Błąd średni bezwzględny: 77.98143046426003

Błąd średniokwadratowy: 7886.072510062893

Pierwiastek błędu średniokwadratowego: 88.80356135911946

Odchylenie standardowe błedu: 86.61667071233929

Średni absolutny błąd procentowy: 2.677856783429592 %

Trening modelu 2008-2017

Współczynnik determinacji: 0.9959129089016896

Predykcja 2018-2020:

Date Year Quarter Users\_in\_mln GeneratedByModel Error Error\_squared

37 2018-01-01 2018 1 2196 2123.681120 72.318880 5230.020463

38 2018-04-01 2018 2 2234 2175.562553 58.437447 3414.935255

39 2018-07-01 2018 3 2271 2228.020446 42.979554 1847.242060

40 2018-10-01 2018 4 2320 2281.054800 38.945200 1516.728619

41 2019-01-01 2019 1 2375 2334.089154 40.910846 1673.697357

42 2019-04-01 2019 2 2414 2385.970587 28.029413 785.648016

43 2019-07-01 2019 3 2449 2438.428480 10.571520 111.757035

44 2019-10-01 2019 4 2498 2491.462834 6.537166 42.734543

45 2020-01-01 2020 1 2603 2544.497188 58.502812 3422.579068

46 2020-04-01 2020 2 2701 2596.955081 104.044919 10825.345187

47 2020-07-01 2020 3 2740 2649.412974 90.587026 8206.009223

48 2020-10-01 2020 4 2797 2702.447328 94.552672 8940.207767

Błąd średni: 53.86812143652438

Błąd średni bezwzględny: 53.86812143652438

Błąd średniokwadratowy: 3834.742049406741

Pierwiastek błędu średniokwadratowego: 61.92529410028459

Odchylenie standardowe błedu: 30.544517385393455

Średni absolutny błąd procentowy: 2.153624028008359 %

Trening modelu 2008-2020

Współczynnik determinacji: 0.9975542416507277

Predykcja 2021-2022:

Date Year Quarter Users\_in\_mln GeneratedByModel Error Error\_squared

49 2021-01-01 2021 1 2853 2800.298549 52.701451 2777.442887

50 2021-04-01 2021 2 2895 2853.425137 41.574863 1728.469203

51 2021-07-01 2021 3 2910 2907.142021 2.857979 8.168046

52 2021-10-01 2021 4 2912 2961.449199 -49.449199 2445.223322

53 2022-01-01 2022 1 2936 3015.756378 -79.756378 6361.079854

54 2022-04-01 2022 2 2934 3068.882966 -134.882966 18193.414525

55 2022-07-01 2022 3 2958 3122.599849 -164.599849 27093.110403

56 2022-10-01 2022 4 2963 3176.907028 -213.907028 45756.216659

Błąd średni: -68.18264106395259

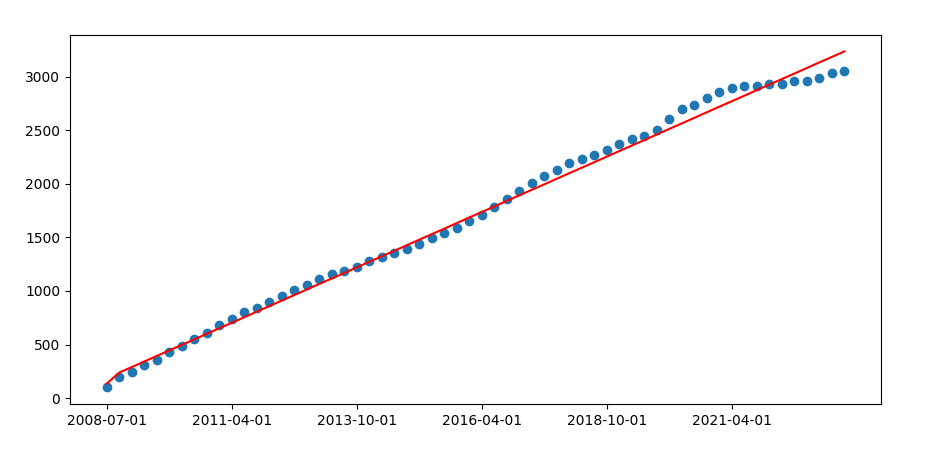
Błąd średni bezwzględny: 92.46621418430618

Błąd średniokwadratowy: 13045.390612447212

Pierwiastek błędu średniokwadratowego: 114.21642006492416

Odchylenie standardowe błedu: 91.63251644471748

Średni absolutny błąd procentowy: 3.1471531502428136 %



# Facebook – Przychód zależny od liczby użytkowników

Dla predykcji przychodu w zależności od liczby użytkowników sprawdzone zostało kilka modeli:

* Potęgowy
* Wykładniczy
* Wielomianowy
* Kwadratowy
* Liniowy

Wszystkie z tych modeli dało się dopasować do różnych przedziałów dany z współczynnikiem determinacji przekraczającym 0.9 oprócz liniowego, który był w okolicach 0.87. Można byłoby wywnioskować, że tak samo dobrze pójdzie z predykcją.

Jak się okazało wielomianowy, a w szczególności kwadratowy miały najlepszą predykcję. Nie była to jakaś idealna ale przy sprawdzaniu średniego absolutnego błędu procentowego utrzymywało się to na poziomie 10-12%.  
  
Co ciekawe mimo, że np taki model jak liniowy miał gorsze dopasowanie (0.89 vs 0.97 przy mierzeniu tego współczynnikiem R^2) to średni absolutny błąd procentowy był mniejszy dla modelu liniowego (35 % vs 51 %).

Wynik działania programu:

Trenowanie modelu dla lat 2009-2015 ze zmienną objaśniającą: 'Users\_in\_mln' na zminną objaśnianą: 'Revenue\_in\_mln'...

Stopnie swobody: 1.0, 5.0

Wartość krytyczna dla alpha 0.05: 6.607890973703367

Model wykładniczy y = a^x \* b zlinearyzowny do: ln(y) = ln(a)x + ln(b)

Współczynnik determinacji dla modelu wykładniczego R^2: 0.9866091493432528

Współczynnik korelacji wielorakiej dla modelu wykładniczego R: 0.993282008969886

Wartość F dla modelu wykładniczego: 368.38927362920595

Wartość F dla modelu wykładniczego z biblioteki statsmodel: 368.38927362920305

Wartość p-value dla modelu wykładniczego z biblioteki statsmodel: 7.079255748105451e-06

Model potęgowy y = x^a \* b zlinearyzowny do: ln(y) = a\*ln(x) + ln(b)

Współczynnik determinacji dla modelu potęgowego: 0.9820218478779258

Współczynnik korelacji wielorakiej dla modelu potęgowego: 0.9909701548875858

Wartość F dla modelu potęgowego: 273.11534611840466

Wartość F dla modelu potęgowego z biblioteki statsmodel: 273.11534611840506

Wartość p-value dla modelu potęgowego z biblioteki statsmodel: 1.4809844060040637e-05

Model wielomianowy y = ax^4 + bx^3 + cx^2 + dx + e

Współczynnik determinacji dla modelu wielomianowego: 0.9975923447194053

Współczynnik korleacji wielorakiej dla modelu wielomianowego: 0.9987954468856

Współczynniki dla modelu wielomianowego: [[ 4.78968365e+01 -7.92438574e-02 5.66699375e-05 -1.20388114e-08]]

Wartość F dla modelu wielomianowego: 2071.7092533134287

Wartość F dla modelu wielomianowego z biblioteki statsmodel: 207.17092045948576

Wartość p-value dla modelu wielomianowego z biblioteki statsmodel: 0.004809513869932095

Model kwadratowy y = ax^2 + bx + c

Współczynnik determinacji dla modelu kwadratowego: 0.9903232810850041

Współczynnik korleacji wielorakiej dla modelu kwadratowego: 0.9951498787042101

Współczynniki dla modelu kwadratowego: [[-12.30787977 0.01311398]]

Wartość F dla modelu kwadratowego: 511.7040650784601

Wartość F dla modelu kwadratowego z biblioteki statsmodel: 204.68162603138438

Wartość p-value dla modelu kwadratowego z biblioteki statsmodel: 9.363888895983885e-05

Model liniowy y = ax + b

Współczynnik determinacji dla modelu liniowego: 0.8776161608980293

Współczynnik korleacji wielorakiej dla modelu liniowego: 0.936811699808467

Współczynniki dla modelu liniowego: [[13.23054179]]

Wartość F dla modelu liniowego: 35.85506743936983

Wartość F dla modelu liniowego z biblioteki statsmodel: 35.855067439369826

Wartość p-value dla modelu liniowego z biblioteki statsmodel: 0.0018628743458618327

Testowanie modelu (wyniki błędu dla różnych modeli) dla lat 2016 - 2018 ...

Średni błąd kwadratowy dla modelu wykładniczego ręcznego: 1724488634.4543622

Średni absolutny błąd procentowy: 77.57677868630694 %

Średni błąd kwadratowy dla modelu wykładniczego: 1724488634.4543622

Średni absolutny błąd procentowy: 77.57677868630694 %

Średni błąd kwadratowy dla modelu potęgowego: 268654129.9738634

Średni absolutny błąd procentowy: 34.23799325165022 %

Średni błąd kwadratowy dla modelu wielomianowego: 164361092.33793393

Średni absolutny błąd procentowy: 19.313449691860534 %

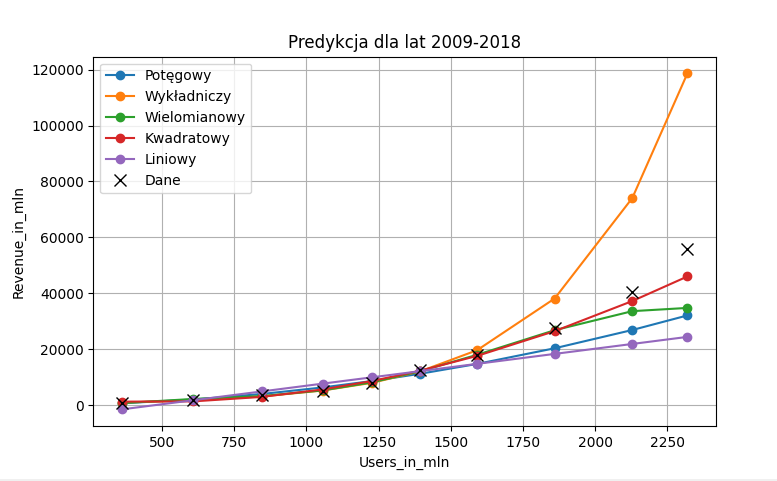
Średni błąd kwadratowy dla modelu kwadratowego: 36378913.48222748

Średni absolutny błąd procentowy: 10.050103338909533 %

Średni błąd kwadratowy dla modelu liniowego: 474883778.8206248

Średni absolutny błąd procentowy: 45.338209554946815 %

Najlepszy model: kwadratowy, z błędem średnio kwadratowym: 36378913.48222748



Trenowanie modelu dla lat 2009-2017 ze zmienną objaśniającą: 'Users\_in\_mln' na zminną objaśnianą: 'Revenue\_in\_mln'...

Stopnie swobody: 1.0, 7.0

Wartość krytyczna dla alpha 0.05: 5.591447851220736

Model wykładniczy y = a^x \* b zlinearyzowny do: ln(y) = ln(a)x + ln(b)

Współczynnik determinacji dla modelu wykładniczego R^2: 0.9803167089871461

Współczynnik korelacji wielorakiej dla modelu wykładniczego R: 0.9901094429340355

Wartość F dla modelu wykładniczego: 348.63158596947716

Wartość F dla modelu wykładniczego z biblioteki statsmodel: 348.63158596947784

Wartość p-value dla modelu wykładniczego z biblioteki statsmodel: 3.1378065359426406e-07

Model potęgowy y = x^a \* b zlinearyzowny do: ln(y) = a\*ln(x) + ln(b)

Współczynnik determinacji dla modelu potęgowego: 0.9803544640970017

Współczynnik korelacji wielorakiej dla modelu potęgowego: 0.990128508880035

Wartość F dla modelu potęgowego: 349.3150445252882

Wartość F dla modelu potęgowego z biblioteki statsmodel: 349.31504452528884

Wartość p-value dla modelu potęgowego z biblioteki statsmodel: 3.11674494343813e-07

Model wielomianowy y = ax^4 + bx^3 + cx^2 + dx + e

Współczynnik determinacji dla modelu wielomianowego: 0.9993642120737372

Współczynnik korleacji wielorakiej dla modelu wielomianowego: 0.9996820554925137

Współczynniki dla modelu wielomianowego: [[ 1.91584626e+01 -2.94217572e-02 2.20319164e-05 -3.71074406e-09]]

Wartość F dla modelu wielomianowego: 11002.960571516278

Wartość F dla modelu wielomianowego z biblioteki statsmodel: 1571.851506032622

Wartość p-value dla modelu wielomianowego z biblioteki statsmodel: 1.212164863606747e-06

Model kwadratowy y = ax^2 + bx + c

Współczynnik determinacji dla modelu kwadratowego: 0.9975942780402791

Współczynnik korleacji wielorakiej dla modelu kwadratowego: 0.998796414711366

Współczynniki dla modelu kwadratowego: [[-1.65626617e+01 1.53958573e-02]]

Wartość F dla modelu kwadratowego: 2902.7294355712443

Wartość F dla modelu kwadratowego z biblioteki statsmodel: 1244.0269009591336

Wartość p-value dla modelu kwadratowego z biblioteki statsmodel: 1.3923111385243545e-08

Model liniowy y = ax + b

Współczynnik determinacji dla modelu liniowego: 0.8714151219051808

Współczynnik korleacji wielorakiej dla modelu liniowego: 0.9334961820517429

Współczynniki dla modelu liniowego: [[21.69265874]]

Wartość F dla modelu liniowego: 47.43874974814815

Wartość F dla modelu liniowego z biblioteki statsmodel: 47.438749748148155

Wartość p-value dla modelu liniowego z biblioteki statsmodel: 0.00023392706084919034

Testowanie modelu (wyniki błędu dla różnych modeli) dla lat 2018 - 2020 ...

Średni błąd kwadratowy dla modelu wykładniczego ręcznego: 7361879272.102246

Średni absolutny błąd procentowy: 89.79481175846935 %

Średni błąd kwadratowy dla modelu wykładniczego: 7361879272.102302

Średni absolutny błąd procentowy: 89.79481175846979 %

Średni błąd kwadratowy dla modelu potęgowego: 502938596.3469685

Średni absolutny błąd procentowy: 31.04201857312911 %

Średni błąd kwadratowy dla modelu wielomianowego: 86967637.75845495

Średni absolutny błąd procentowy: 12.460090537150858 %

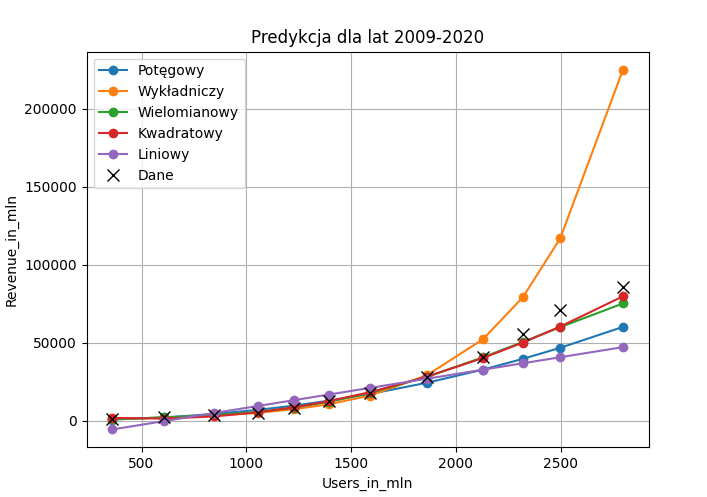
Średni błąd kwadratowy dla modelu kwadratowego: 60095950.42134187

Średni absolutny błąd procentowy: 10.770049459673585 %

Średni błąd kwadratowy dla modelu liniowego: 925668080.7351221

Średni absolutny błąd procentowy: 40.624875340425184 %

Najlepszy model: kwadratowy, z błędem średnio kwadratowym: 60095950.42134187



Trenowanie modelu dla lat 2009-2020 ze zmienną objaśniającą: 'Users\_in\_mln' na zminną objaśnianą: 'Revenue\_in\_mln'...

Stopnie swobody: 1.0, 10.0

Wartość krytyczna dla alpha 0.05: 4.9646027437307145

Model wykładniczy y = a^x \* b zlinearyzowny do: ln(y) = ln(a)x + ln(b)

Współczynnik determinacji dla modelu wykładniczego R^2: 0.9706993883831571

Współczynnik korelacji wielorakiej dla modelu wykładniczego R: 0.9852407768576964

Wartość F dla modelu wykładniczego: 331.2898041435994

Wartość F dla modelu wykładniczego z biblioteki statsmodel: 331.2898041435988

Wartość p-value dla modelu wykładniczego z biblioteki statsmodel: 5.380899038568536e-09

Model potęgowy y = x^a \* b zlinearyzowny do: ln(y) = a\*ln(x) + ln(b)

Współczynnik determinacji dla modelu potęgowego: 0.9816670180252234

Współczynnik korelacji wielorakiej dla modelu potęgowego: 0.9907911071589326

Wartość F dla modelu potęgowego: 535.4649992979032

Wartość F dla modelu potęgowego z biblioteki statsmodel: 535.4649992979055

Wartość p-value dla modelu potęgowego z biblioteki statsmodel: 5.135826271447456e-10

Model wielomianowy y = ax^4 + bx^3 + cx^2 + dx + e

Współczynnik determinacji dla modelu wielomianowego: 0.997590083532853

Współczynnik korleacji wielorakiej dla modelu wielomianowego: 0.9987943149281803

Współczynniki dla modelu wielomianowego: [[ 5.16394266e+01 -6.98397088e-02 4.05026450e-05 -6.31200487e-09]]

Wartość F dla modelu wielomianowego: 4139.521419652506

Wartość F dla modelu wielomianowego z biblioteki statsmodel: 724.4162480798933

Wartość p-value dla modelu wielomianowego z biblioteki statsmodel: 3.0860622379578967e-09

Model kwadratowy y = ax^2 + bx + c

Współczynnik determinacji dla modelu kwadratowego: 0.9952057535479212

Współczynnik korleacji wielorakiej dla modelu kwadratowego: 0.9975999967661995

Współczynniki dla modelu kwadratowego: [[-2.21769482e+01 1.82158901e-02]]

Wartość F dla modelu kwadratowego: 2075.8335298269712

Wartość F dla modelu kwadratowego z biblioteki statsmodel: 934.1250884221388

Wartość p-value dla modelu kwadratowego z biblioteki statsmodel: 3.657982301363278e-11

Model liniowy y = ax + b

Współczynnik determinacji dla modelu liniowego: 0.8858066304059365

Współczynnik korleacji wielorakiej dla modelu liniowego: 0.9411730076909008

Współczynniki dla modelu liniowego: [[35.37447882]]

Wartość F dla modelu liniowego: 77.5707585786125

Wartość F dla modelu liniowego z biblioteki statsmodel: 77.57075857861243

Wartość p-value dla modelu liniowego z biblioteki statsmodel: 5.024235788762229e-06

Testowanie modelu (wyniki błędu dla różnych modeli) dla lat 2021 - 2022 ...

Średni błąd kwadratowy dla modelu wykładniczego ręcznego: 3587130286.201193

Średni absolutny błąd procentowy: 50.51250087289563 %

Średni błąd kwadratowy dla modelu wykładniczego: 3587130286.2011185

Średni absolutny błąd procentowy: 50.512500872895096 %

Średni błąd kwadratowy dla modelu potęgowego: 1114983366.963056

Średni absolutny błąd procentowy: 28.389278059953998 %

Średni błąd kwadratowy dla modelu wielomianowego: 516040101.21051025

Średni absolutny błąd procentowy: 19.292569254445723 %

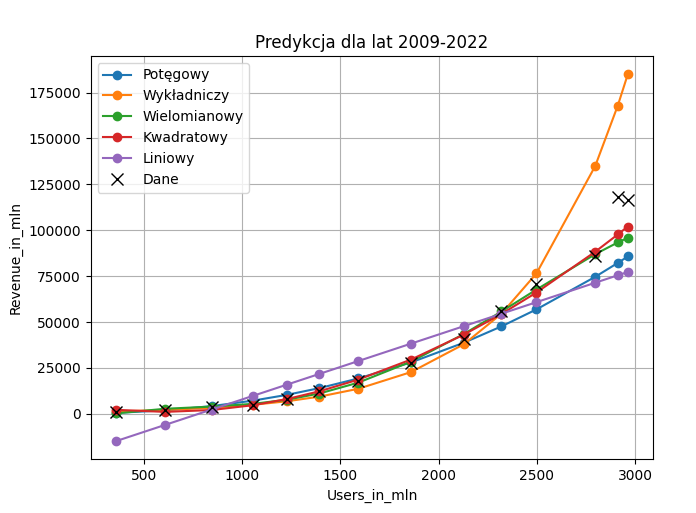
Średni błąd kwadratowy dla modelu kwadratowego: 310690119.04535395

Średni absolutny błąd procentowy: 14.823575919887638 %

Średni błąd kwadratowy dla modelu liniowego: 1674171491.213306

Średni absolutny błąd procentowy: 34.85944002038941 %

Najlepszy model: kwadratowy, z błędem średnio kwadratowym: 310690119.04535395



# Facebook - Ilość potrzebnych pracowników przy danej liczbie użytkowników

Podobnie jak w przypadku przychodu sprawdzone zostały te same modele:

* Potęgowy
* Wykładniczy
* Wielomianowy
* Kwadratowy
* Liniowy

Tu również dopasowanie było wysokie I tak samo jak ostatnio dopasowanie liniowego modelu była troszkę słabsza od reszty.

Tu jednak do najlepszych modeli wielomianowego I kwadratowego dołączył wykładniczy. Jeżeli chodzi o średni błąd kwadratowy to był nawet najlepszy. Tu jednak zależy od badanego okresu:

* Przy uczeniu do 2015 roku I testowaniu na 2016-2018 najlepszy był kwadratowy przy średnim absolutnym błędzie procentowym 7,23%
* Przy uczeniu do 2017 I testowaniu 2018-2020 najlepszym był wielomianowy z błędem 10,77%
* Przy uczeniu do 2020 I testowaniu 2021-2022 najlepsze były 3 modele wielomianowy, kwadratowy oraz wykładniczy z błędami między 14 a 17 %.

Wynik działania programu:

Trenowanie modelu dla lat 2009-2015 ze zmienną objaśniającą: 'Users\_in\_mln' na zminną objaśnianą: 'Employees'...

Stopnie swobody: 1.0, 5.0

Wartość krytyczna dla alpha 0.05: 6.607890973703367

Model wykładniczy y = a^x \* b zlinearyzowny do: ln(y) = ln(a)x + ln(b)

Współczynnik determinacji dla modelu wykładniczego R^2: 0.9983315148786045

Współczynnik korelacji wielorakiej dla modelu wykładniczego R: 0.9991654091683742

Wartość F dla modelu wykładniczego: 2991.730348915582

Wartość F dla modelu wykładniczego z biblioteki statsmodel: 2991.730348915625

Wartość p-value dla modelu wykładniczego z biblioteki statsmodel: 3.863168596570764e-08

Model potęgowy y = x^a \* b zlinearyzowny do: ln(y) = a\*ln(x) + ln(b)

Współczynnik determinacji dla modelu potęgowego: 0.9612886710587323

Współczynnik korelacji wielorakiej dla modelu potęgowego: 0.9804532987647766

Wartość F dla modelu potęgowego: 124.16115609427746

Wartość F dla modelu potęgowego z biblioteki statsmodel: 124.16115609427717

Wartość p-value dla modelu potęgowego z biblioteki statsmodel: 0.00010152537520167999

Model wielomianowy y = ax^4 + bx^3 + cx^2 + dx + e

Współczynnik determinacji dla modelu wielomianowego: 0.99918392283369

Współczynnik korleacji wielorakiej dla modelu wielomianowego: 0.9995918781351166

Współczynniki dla modelu wielomianowego: [[ 2.42360029e+01 -3.95087283e-02 2.93311810e-05 -6.28739383e-09]]

Wartość F dla modelu wielomianowego: 6121.871583244837

Wartość F dla modelu wielomianowego z biblioteki statsmodel: 612.187134095857

Wartość p-value dla modelu wielomianowego z biblioteki statsmodel: 0.001631488415169364

Model kwadratowy y = ax^2 + bx + c

Współczynnik determinacji dla modelu kwadratowego: 0.9951440605186055

Współczynnik korleacji wielorakiej dla modelu kwadratowego: 0.9975690755624923

Współczynniki dla modelu kwadratowego: [[-6.52078155 0.00794623]]

Wartość F dla modelu kwadratowego: 1024.6668686167643

Wartość F dla modelu kwadratowego z biblioteki statsmodel: 409.8667474467018

Wartość p-value dla modelu kwadratowego z biblioteki statsmodel: 2.358014824696643e-05

Model liniowy y = ax + b

Współczynnik determinacji dla modelu liniowego: 0.902255137786032

Współczynnik korleacji wielorakiej dla modelu liniowego: 0.9498711164079219

Współczynniki dla modelu liniowego: [[8.95385828]]

Wartość F dla modelu liniowego: 46.15358379711833

Wartość F dla modelu liniowego z biblioteki statsmodel: 46.153583797118394

Wartość p-value dla modelu liniowego z biblioteki statsmodel: 0.0010517443589205385

Testowanie modelu (wyniki błędu dla różnych modeli) dla lat 2016 - 2018 ...

Średni błąd kwadratowy dla modelu wykładniczego ręcznego: 122069823.92304885

Średni absolutny błąd procentowy: 36.88536189916826 %

Średni błąd kwadratowy dla modelu wykładniczego: 122069823.92304635

Średni absolutny błąd procentowy: 36.88536189916786 %

Średni błąd kwadratowy dla modelu potęgowego: 132126098.21822216

Średni absolutny błąd procentowy: 36.09242730354275 %

Średni błąd kwadratowy dla modelu wielomianowego: 46887702.18482941

Średni absolutny błąd procentowy: 16.27481656789269 %

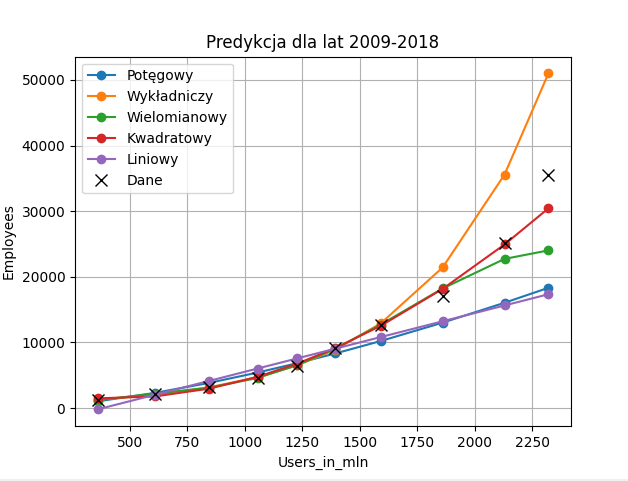
Średni błąd kwadratowy dla modelu kwadratowego: 9270149.76506053

Średni absolutny błąd procentowy: 7.226525598966381 %

Średni błąd kwadratowy dla modelu liniowego: 145733880.4608532

Średni absolutny błąd procentowy: 37.142856722959586 %

Najlepszy model: kwadratowy, z błędem średnio kwadratowym: 9270149.76506053



Trenowanie modelu dla lat 2009-2017 ze zmienną objaśniającą: 'Users\_in\_mln' na zminną objaśnianą: 'Employees'...

Stopnie swobody: 1.0, 7.0

Wartość krytyczna dla alpha 0.05: 5.591447851220736

Model wykładniczy y = a^x \* b zlinearyzowny do: ln(y) = ln(a)x + ln(b)

Współczynnik determinacji dla modelu wykładniczego R^2: 0.991765144446521

Współczynnik korelacji wielorakiej dla modelu wykładniczego R: 0.9958740605350261

Wartość F dla modelu wykładniczego: 843.045268498082

Wartość F dla modelu wykładniczego z biblioteki statsmodel: 843.0452684980748

Wartość p-value dla modelu wykładniczego z biblioteki statsmodel: 1.479529534546654e-08

Model potęgowy y = x^a \* b zlinearyzowny do: ln(y) = a\*ln(x) + ln(b)

Współczynnik determinacji dla modelu potęgowego: 0.9627081702094471

Współczynnik korelacji wielorakiej dla modelu potęgowego: 0.9811769311441475

Wartość F dla modelu potęgowego: 180.70867611793355

Wartość F dla modelu potęgowego z biblioteki statsmodel: 180.70867611793207

Wartość p-value dla modelu potęgowego z biblioteki statsmodel: 2.95786906384857e-06

Model wielomianowy y = ax^4 + bx^3 + cx^2 + dx + e

Współczynnik determinacji dla modelu wielomianowego: 0.9977637698129549

Współczynnik korleacji wielorakiej dla modelu wielomianowego: 0.9988812591158945

Współczynniki dla modelu wielomianowego: [[-5.73682820e+00 1.00624566e-02 -3.09132256e-06 9.33486084e-10]]

Wartość F dla modelu wielomianowego: 3123.268091608944

Wartość F dla modelu wielomianowego z biblioteki statsmodel: 446.1811554853008

Wartość p-value dla modelu wielomianowego z biblioteki statsmodel: 1.4979810832658389e-05

Model kwadratowy y = ax^2 + bx + c

Współczynnik determinacji dla modelu kwadratowego: 0.9969400262684088

Współczynnik korleacji wielorakiej dla modelu kwadratowego: 0.9984688409101251

Współczynniki dla modelu kwadratowego: [[-6.70455172 0.00793967]]

Wartość F dla modelu kwadratowego: 2280.601337139596

Wartość F dla modelu kwadratowego z biblioteki statsmodel: 977.4005730598263

Wartość p-value dla modelu kwadratowego z biblioteki statsmodel: 2.865187810571574e-08

Model liniowy y = ax + b

Współczynnik determinacji dla modelu liniowego: 0.9007129496926258

Współczynnik korleacji wielorakiej dla modelu liniowego: 0.9490589811453374

Współczynniki dla modelu liniowego: [[13.02378627]]

Wartość F dla modelu liniowego: 63.50264841516902

Wartość F dla modelu liniowego z biblioteki statsmodel: 63.50264841516903

Wartość p-value dla modelu liniowego z biblioteki statsmodel: 9.344719324816804e-05

Testowanie modelu (wyniki błędu dla różnych modeli) dla lat 2018 - 2020 ...

Średni błąd kwadratowy dla modelu wykładniczego ręcznego: 361918810.41488504

Średni absolutny błąd procentowy: 28.410585330298698 %

Średni błąd kwadratowy dla modelu wykładniczego: 361918810.4148931

Średni absolutny błąd procentowy: 28.41058533029908 %

Średni błąd kwadratowy dla modelu potęgowego: 429235425.528491

Średni absolutny błąd procentowy: 41.98108216067332 %

Średni błąd kwadratowy dla modelu wielomianowego: 23989670.929040197

Średni absolutny błąd procentowy: 10.768155779918072 %

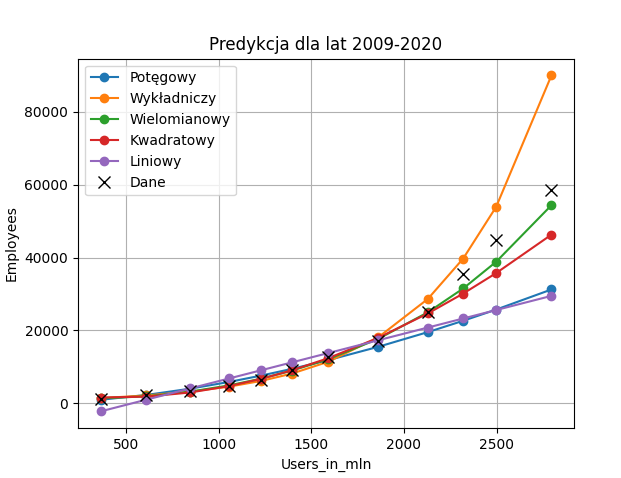
Średni błąd kwadratowy dla modelu kwadratowego: 88901962.56595369

Średni absolutny błąd procentowy: 18.976002656980114 %

Średni błąd kwadratowy dla modelu liniowego: 458668352.38589746

Średni absolutny błąd procentowy: 42.4863221837998 %

Najlepszy model: wielomianowy, z błędem średnio kwadratowym: 23989670.929040197



Trenowanie modelu dla lat 2009-2020 ze zmienną objaśniającą: 'Users\_in\_mln' na zminną objaśnianą: 'Employees'...

Stopnie swobody: 1.0, 10.0

Wartość krytyczna dla alpha 0.05: 4.9646027437307145

Model wykładniczy y = a^x \* b zlinearyzowny do: ln(y) = ln(a)x + ln(b)

Współczynnik determinacji dla modelu wykładniczego R^2: 0.9907380973477166

Współczynnik korelacji wielorakiej dla modelu wykładniczego R: 0.99535827587242

Wartość F dla modelu wykładniczego: 1069.6917626352497

Wartość F dla modelu wykładniczego z biblioteki statsmodel: 1069.691762635255

Wartość p-value dla modelu wykładniczego z biblioteki statsmodel: 1.6837764223385264e-11

Model potęgowy y = x^a \* b zlinearyzowny do: ln(y) = a\*ln(x) + ln(b)

Współczynnik determinacji dla modelu potęgowego: 0.9579058798626506

Współczynnik korelacji wielorakiej dla modelu potęgowego: 0.9787266624868511

Wartość F dla modelu potęgowego: 227.5628702386673

Wartość F dla modelu potęgowego z biblioteki statsmodel: 227.56287023866716

Wartość p-value dla modelu potęgowego z biblioteki statsmodel: 3.311066257900626e-08

Model wielomianowy y = ax^4 + bx^3 + cx^2 + dx + e

Współczynnik determinacji dla modelu wielomianowego: 0.9971310334374348

Współczynnik korleacji wielorakiej dla modelu wielomianowego: 0.9985644863690251

Współczynniki dla modelu wielomianowego: [[ 2.21145139e+01 -2.60396269e-02 1.44981024e-05 -1.82699948e-09]]

Wartość F dla modelu wielomianowego: 3475.5756530877234

Wartość F dla modelu wielomianowego z biblioteki statsmodel: 608.225739175659

Wartość p-value dla modelu wielomianowego z biblioteki statsmodel: 5.679126489622861e-09

Model kwadratowy y = ax^2 + bx + c

Współczynnik determinacji dla modelu kwadratowego: 0.9936266716919372

Współczynnik korleacji wielorakiej dla modelu kwadratowego: 0.9968082421870001

Współczynniki dla modelu kwadratowego: [[-1.67923065e+01 1.24709083e-02]]

Wartość F dla modelu kwadratowego: 1559.0388940656235

Wartość F dla modelu kwadratowego z biblioteki statsmodel: 701.5675023295305

Wartość p-value dla modelu kwadratowego z biblioteki statsmodel: 1.3171896470069028e-10

Model liniowy y = ax + b

Współczynnik determinacji dla modelu liniowego: 0.8702941572743408

Współczynnik korleacji wielorakiej dla modelu liniowego: 0.932895576832874

Współczynniki dla modelu liniowego: [[22.6083795]]

Wartość F dla modelu liniowego: 67.09752922349841

Wartość F dla modelu liniowego z biblioteki statsmodel: 67.09752922349841

Wartość p-value dla modelu liniowego z biblioteki statsmodel: 9.567629837642307e-06

Testowanie modelu (wyniki błędu dla różnych modeli) dla lat 2021 - 2022 ...

Średni błąd kwadratowy dla modelu wykładniczego ręcznego: 167963468.660008

Średni absolutny błąd procentowy: 16.286000165651025 %

Średni błąd kwadratowy dla modelu wykładniczego: 167963468.66000015

Średni absolutny błąd procentowy: 16.28600016565062 %

Średni błąd kwadratowy dla modelu potęgowego: 1052874756.3271794

Średni absolutny błąd procentowy: 39.7289307840846 %

Średni błąd kwadratowy dla modelu wielomianowego: 172744917.29900718

Średni absolutny błąd procentowy: 14.376188763954012 %

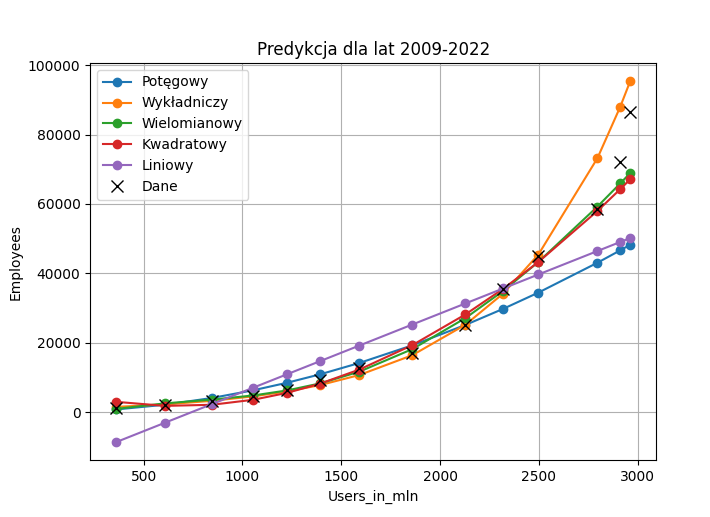
Średni błąd kwadratowy dla modelu kwadratowego: 217345217.69739613

Średni absolutny błąd procentowy: 16.565778642631678 %

Średni błąd kwadratowy dla modelu liniowego: 922390435.8193268

Średni absolutny błąd procentowy: 36.93353111497719 %

Najlepszy model: wykładniczy, z błędem średnio kwadratowym: 167963468.66000015



# Wnioski

Badając takie dane można się wiele dowiedzieć:

1. Ciężko jest oszacować przyszłość ale przy pozwoleniu sobie na pewien zakres błędu są modele które się lepiej sprawdzają a inne gorzej.
2. To, że model do pewnego czasu jest dobry nie oznacza że przy nowych danych inny nie jest lepszy.
3. Mimo wielu dobrze dopasowanych modeli nie możemy stwierdzić, że w przyszłości wskażą dobre dane.

# Wzory

**Model Potęgowy :**

**Model Wykładniczy:**

**Model Wielomianowy (4 stopnia):**

**Model Kwadratowy:**

**Model Liniowy:**

gdzie:

* y to zmienna objaśniana (np. liczba pracowników),
* x to zmienna objaśniająca (np. liczba użytkowników),
* a,b,c,d,e to współczynniki modelu, które są dostosowywane podczas procesu dopasowywania modelu do danych,