Uniwersytet Przyrodniczy we Wrocławiu

Wydział Biologii i Hodowli Zwierząt

Kierunek: Bioinformatyka

Specjalność: Biostatystyka i programowanie bioinformatyczne

Przedmiot: Zaawansowane elementy stosowania pakietów statystycznych

---

Analiza oraz modelowanie łącznej kwoty odrzuconych pożyczek na podstawie raportu Lending Club w latach 2007-2012

Autor
Artur Wójtowicz

Prowadzący
dr inż. Anna Mucha

6 lutego 2019

Spis treści

1 Wprowadzenie

1.1 Zbiór danych

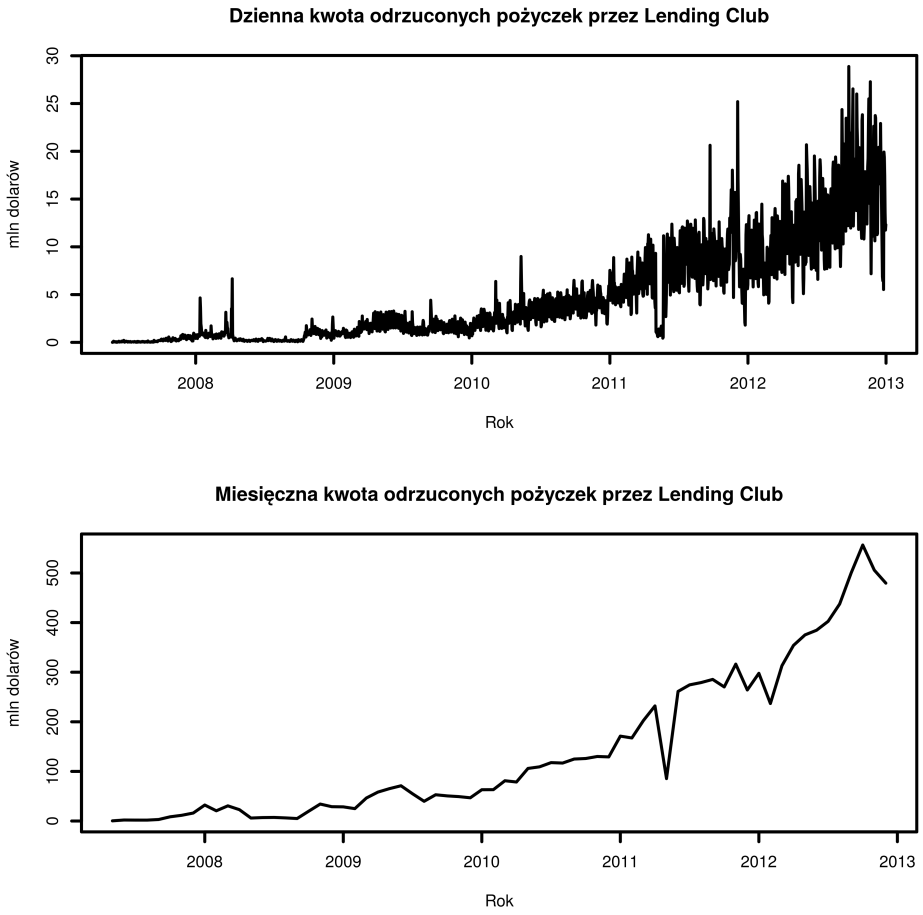
Dane, które zostały użyte w tym sprawozdaniu, pochodzą z głównej strony firmy *Lending Club* - firmy będącej pośrednikiem między pożyczkodawcą a pożyczkobiorcą na terenie USA, zaś dane, są ogólnodostępne na oficjalnej stronie, pod adresem:

`https://www.lendingclub.com/info/download-data.action`

Zbiór danych zawiera ponad 755 tysięcy rekordów odrzuconych pożyczek w latach od 26 maja 2007 do 31 grudnia 2012 oraz dziewięć kolumn, na które składają się:

- *Amount Requested* - Kwota całkowita wnioskowana przez pożyczkobiorcę;
- *Application Date* - Data aplikacji o pożyczkę przez pożyczkobiorcę;
- *Loan Title* - Tytuł pożyczki podaną przez pożyczkobiorcę;
- *Risk Score* - Do 5 listopada 2013 używano tutaj *FICO score* stworzonego przez *Fair Isaac Corporation* do oceny ryzyka udzielenia pożyczki. Po 5 listopada 2013 roku, do oceny ryzyka korzystano z Vantage score;
- *Debt-To-Income Ratio* - Obliczona proporcja względem całkowitych miesięcznych spłat zadłużeń pożyczkodawcy, odliczając od tego zakwaterowanie i wnioskowaną pożyczkę. Wartość ta, jest dzielona przez miesięczne przychody pożyczkobiorcy, które zgłosił;
- *Zip Code* - Pierwsze trzy numery kodu pocztowego pożyczkobiorcy;
- *State* - Stan (w USA), w którym pożyczkodawca składa aplikację o pożyczkę;
- *Employment Length* - Okres w latach zatrudnienia. Możliwe wartości znajdują się pomiędzy 0 a 10, gdzie 0 oznacza mniej niż rok zaś 10 więcej niż dziesięc lat.
- *Policy Code* - W przypadku naszych danych, wartość każdego rekordu tutaj, jest równa 0.

Do analiz szeregów czasowych, potrzebować będziemy jedynie dwóch pierwszych kolumn - *Amount Requested* oraz *Application Date*. W celu ich wyciągnięcia oraz dopasowania, napisany został skrypt w języku *Python*, który tworzy dwa pliki csv. Pierwszy, zawierający dzienną sumę pożyczek wnioskowanych przez pożyczkobiorcę, oraz drugi, zawierający miesięczną sumę pożyczek wnioskowanych przez pożyczkobiorcę.



Rysunek 1: Ogólne przedstawienie danych dziennych oraz miesięcznych w czasie.

Czyste dane zostały przedstawione na Rysunek 1, gdzie już na starcie możemy zauważyć trend wzrostowy. Problematyczny może tutaj się okazać miesiąc maj w roku 2011, gdzie widzimy intensywny spadek ilości odrzuconych pożyczek. Istnieje wiele przesłanek tak intensywnego spadku, jednakże najpewniejszymi są:

- Usunięcie lub awaria bazy danych;
- Działanie na rzecz wykluczenia danych ogólnodostępnych;

Firma *Lending Club* wystartowała na Facebooku w 2006 roku, a na giełdę wskoczyła dopiero 12 grudnia 2014, więc nie jesteśmy w stanie stwierdzić, czy występowało wtedy załamanie giełdowe dla tej firmy, stąd możliwe jedynie spekulacje dotyczące załamania tendencji wzrostowej.

Jednocześnie możemy zauważyć brak wzorców sezonowych w danych.

1.2 Odstępstwa w danych

Analiza, którą przeprowadzam będzie na niezmodyfikowanych danych udostępnionych przez *Lending Club*, jednakże dobrze jest wiedzieć o odstępstwach w danych. Analizujemy dane jedynie odrzuconych pożyczek dla osób fizycznych.

Lending Club oferuje pożyczki do 40 tysięcy dolarów dla osób prywatnych oraz do 300 tysięcy dolarów dla małych firm. Analizując bezpośrednio ogólnodostepne dane, możemy znaleźć dwa rekordy na wysokość pożyczki 0. Zauważyłem również, że istnieje kilkadziesiąt rekordów odstających wzwyż od zakładanych norm, dokładniej omówię sześć największych kwotowo aplikacji o pożyczkę:

- 10.09.2010 - 1.4 miliona dolarów na zakup samochodu;
- 16.10.2012 - 1.2 miliona dolarów na konsolidację zadłużenia;
- 16.10.2012 - milion dolarów na konsolidację zadłużenia
- 09.06.2011 - 500 tysięcy dolarów na określony cel jako inne;
- 15.06.2011 - 500 tysięcy dolarów na określony cel jako inne;
- 22.08.2012 - 300 tysięcy dolarów na zakup samochód.

Jak możemy zauważyć, dwa z trzech zapytań o pożyczkę powyżej miliona dolarów miała miejsce tego samego dnia, na podobną kwotę. Możliwe jest, że była to ta sama osoba, która sprawdzała swoją zdolność kredytową po

tym, jak zadłużyła się w innych firmach/bankach. Można to traktować jako dane nadmiarowe.

Podobną sytuację możemy zauważyć przy pożyczkach na 500 tysięcy dolarów, różnica dni pomiędzy nimi pozwala nam zakładać, że te zapytania zostały podane przez jedną osobę.

Pożyczki na 300 tysięcy lub 1.4 miliona dolarów na zakup samochodu, mogą wydawać się abstrakcyjne, jednakże jak powyżej, osoba wysyłająca zapytanie mogła sprawdzać jedynie swoją zdolność kredytową, lub pomylić się przy wypełnianiu aplikacji.

Analiza pozostałych czynników takich jak Risk Score, State, Debt-To-Income, przy bardzo podobnych pożyczkach jak wyżej wymieniłem może wskazywać na inną osobę, tak tym czynnikom nie należy wierzyć, ponieważ są one zależne od deklaracji pożyczkobiorcy. W zależności od Stanu, obowiązują inne przepisy. W zależności od deklaracji zarobkowych, mogą różnić się parametry Risk Score, Debt-To-Income.

Łącznie cały zbiór zawiera 218 rekordów, w których kwota przekraczała 40 tysięcy dolarów. Chciałbym nadmienić, że 157 z tych rekordów, czyli 72% miała przeznaczenie - zakup samochodu. Po przeanalizowaniu odstępstw nasuwają mi się dwa wnioski:

- *Lending Club* jest platformą internetową, każda osoba na wpół anonimowo może wypełnić wniosek w celu sprawdzenia swojej zdolności kredytowej. Zapytanie takie może być powielane przez jedną osobę dla wielu Stanów, deklaracji płacowych itd.;
- Mógł wystąpić czynnik ludzki przy wypełnianiu wniosku o pożyczkę lub jego magazynowaniu;

1.3 Kod źródłowy

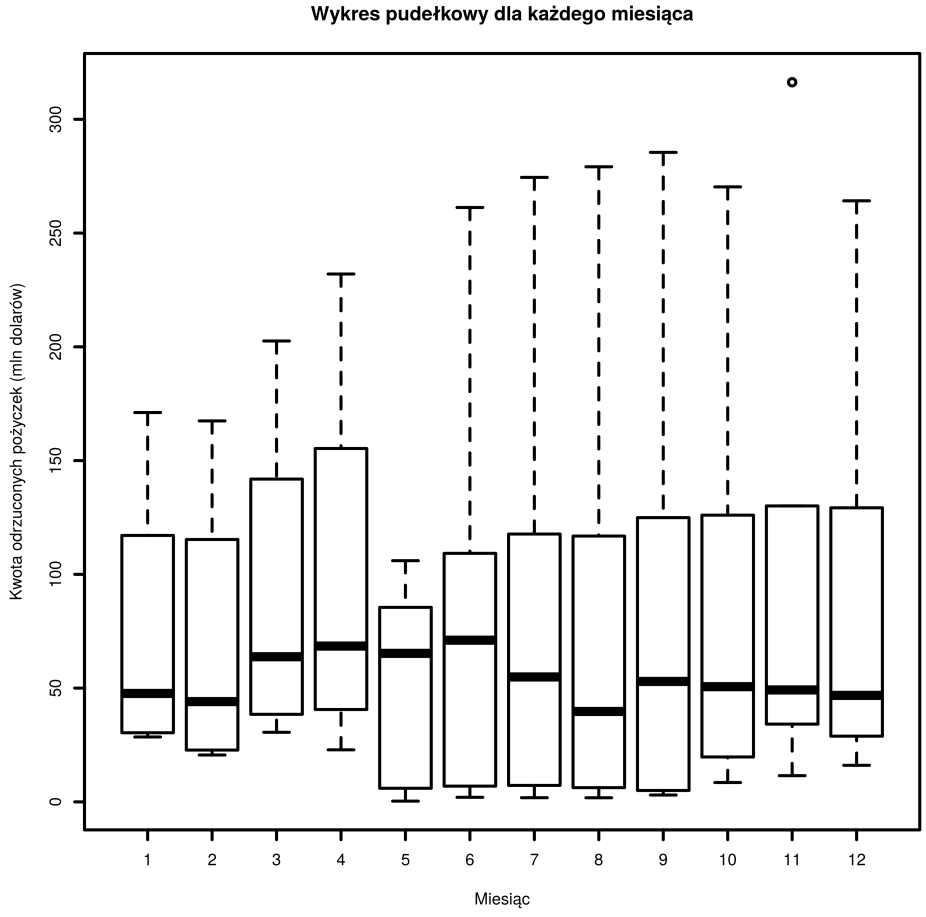
Kod źródłowy wszystkich omawianych skryptów napisanych w R oraz Python a także wszystkie rysunki dostępne są do pobrania na githubie. Link poniżej:

`https://github.com/arturwojtowicz/analiza_odrzuconych_pozyczek`

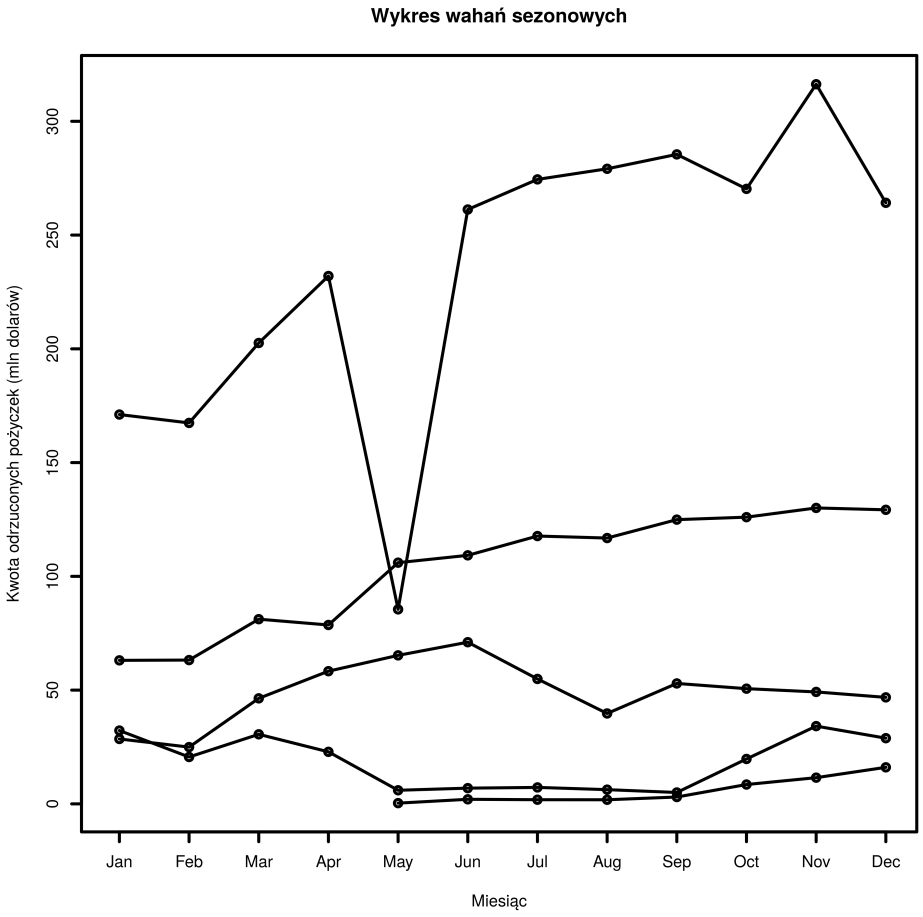
2 Wizualizacja i przekształcenia

2.1 Wykresy sezonowe

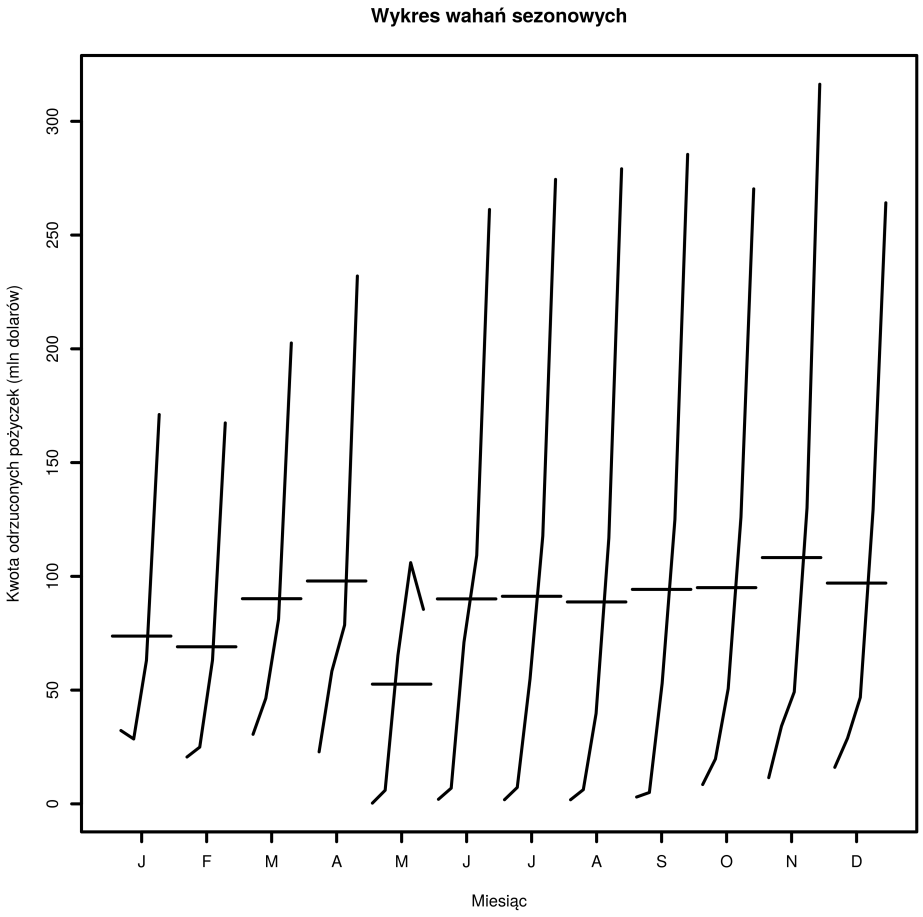
Na wykresie pudełkowym widocznym na Rysunek 2 możemy zauważyć spore różnice w średnich wartościach dla poszczególnych szeregów czasowych. Widzimy również załamanie w piątym miesiącu o czym wspominałem przy omawianiu Rysunek 1. Takiej samej informacji dostarczają nam pozostałe dwa wykresy - Rysunek 3, 4. Zauważalny jest trend wzrostowy. Wahania sezonowe nie są widoczne.



Rysunek 2: Wykres pudełkowy każdego miesiąca dla analizowanego szeregu czasowego



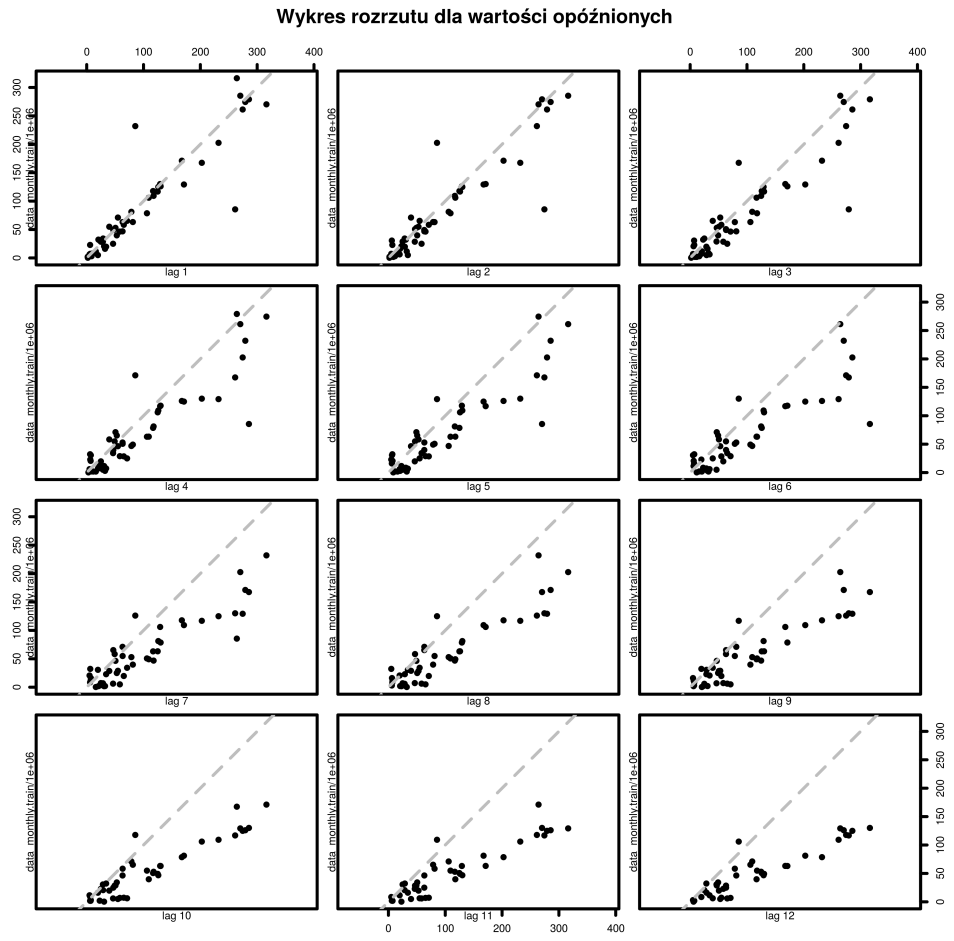
Rysunek 3: Wykres wahań sezonowych każdego miesiąca dla analizowanego szeregu czasowego.



Rysunek 4: Wykres wahań miesięcznych dla analizwanego szeregu czasowego.

2.2 Wykresy autokorelacji

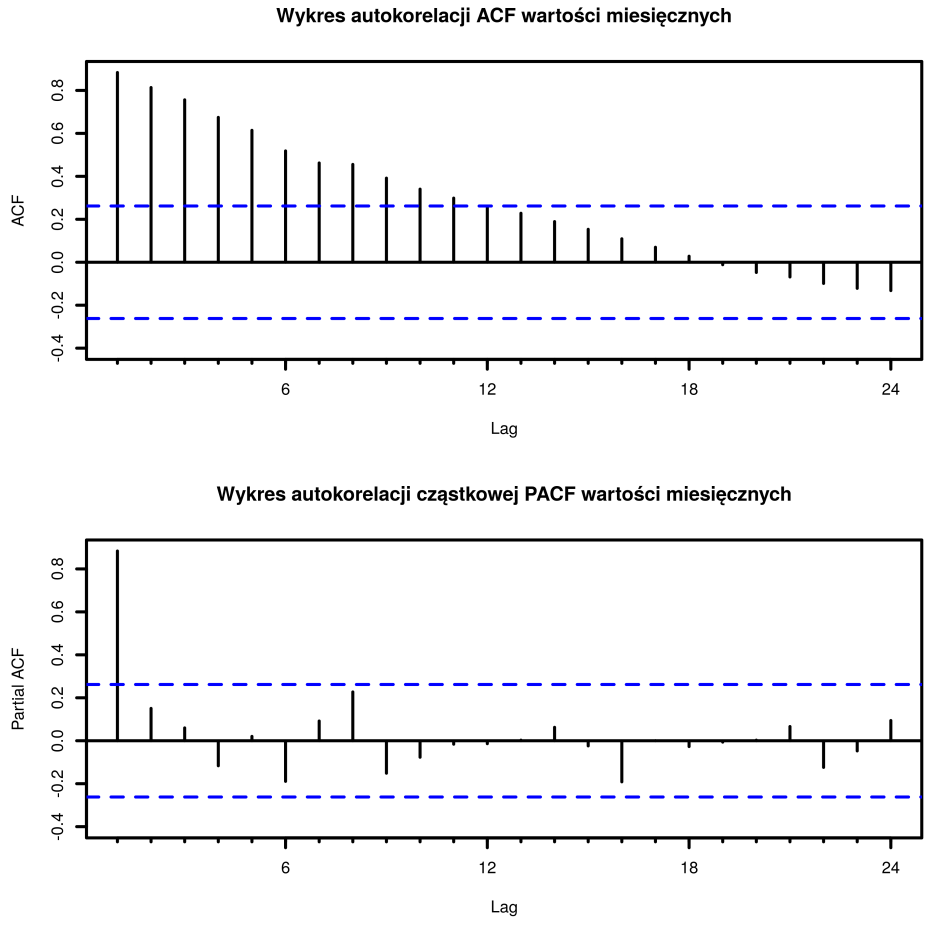
Poniżej widzimy wykres rozrzutu dla wartości opóźnionych danych miesięcznych analizowanego szeregu czasowego. Jak możemy zauważyć dla opóźnienia 1 mamy silną korelację czasową, co oznacza, że trend rzeczywiście istnieje. Analizując opóźnienie 12, widzimy, że liniowość nie jest zachowana, przez co mamy kolejną przesłankę do wykluczenia sezonowości w naszych danych.



Rysunek 5: Wykres rozrzutu dla wartości opóźnionych dla danych miesięcznych analizowanego szeregu czasowego.

Wykresy autokorelacji ACF oraz autokorelacji cząstkowej PACF dają nam również informację o tym, że w naszych danych występuje trend wzrostowy poprzez:

- ACF - Wysoka, dodatnia wartość początkowa, istotna statystycznie zanikająca;
- PACF - Wartości nie wychodzące poza przedziały ufności.

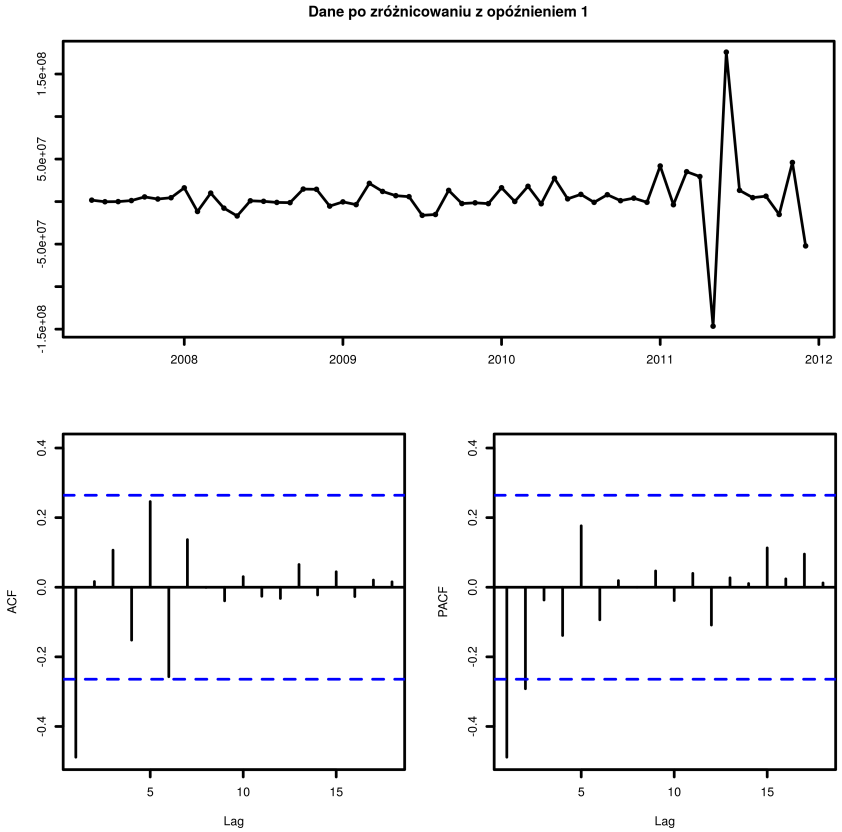


Rysunek 6: Wykres autokorelacji (ACF) oraz autokorelacji cząstkowej (PACF) dla danych miesięcznych analizowanego szeregu czasowego.

2.3 Różnicowanie szeregu czasowego

Za pomocą przekształceń typu: różnicowanie, dekompozycja, transformacja BoxaCoxa. Jesteśmy w stanie pozbyć się trendów oraz sezonowości z naszych danych. Przekształcone dane, w których pozbyliśmy się regularnych tendencji, są znacznie lepsze do analizy oraz interpretacji.

Na przedstawionym poniżej wykresie (Rysunek 7) widoczny jest szereg czasowy, po zastosowaniu różnicowania z opóźnieniem jeden. Po różnicowaniu pozbyliśmy się trendu wzrostowego, co widoczne jest na wykresach autokorelacji ACF oraz autokorelacji cząstkowej PACF. Dzięki nim, możemy wytypować modele. Są to kolejno **AR(2)**, **MA(1)**.



Rysunek 7: Wykres różnicowania z opóźnieniem jeden dla danych miesięcznych analizowanego szeregu czasowego.

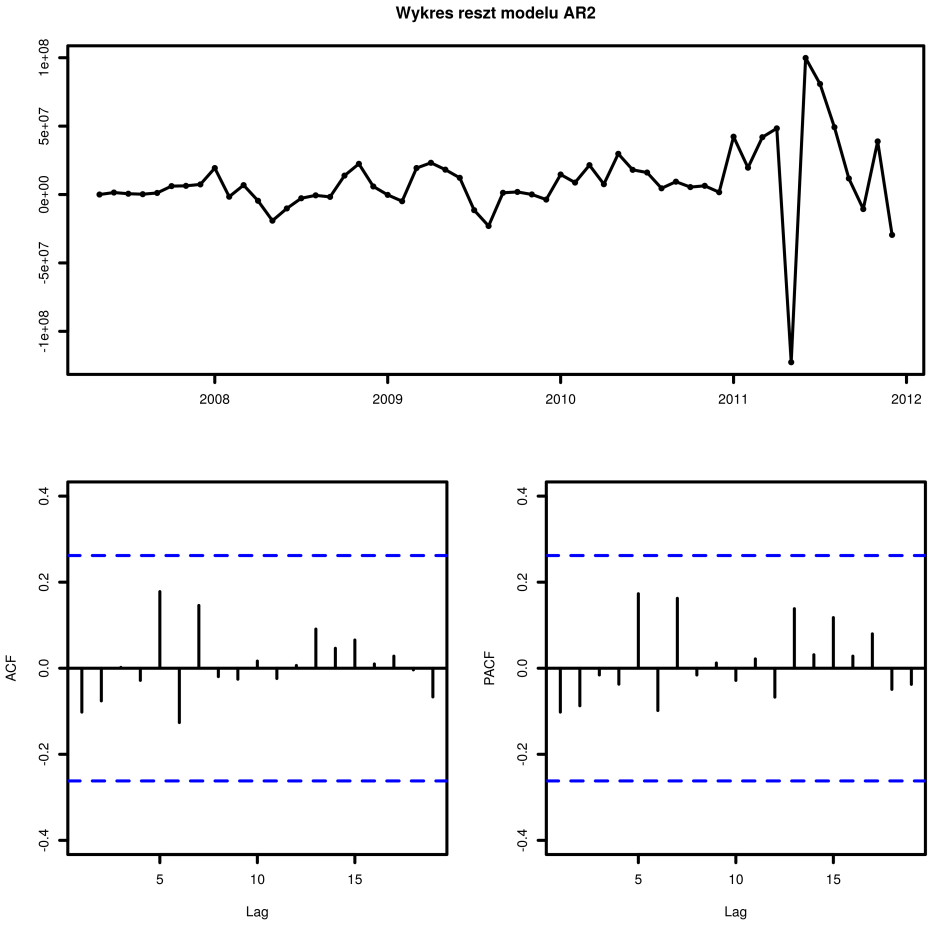
3 Diagnostyka wybranych modeli

Mając na uwadze wybór jak najlepszego modelu, postanowiłem użyć w pakiecie R funkcji *auto.arima*, która wybiera optymalne różnicowanie oraz najlepszy model dla naszych danych. Funkcja ta zwróciła w efekcie model **MA(2)** z różnicowaniem jeden, czyli taki sam, jaki został przed chwilą wybrany przeze mnie do dalszych analiz.

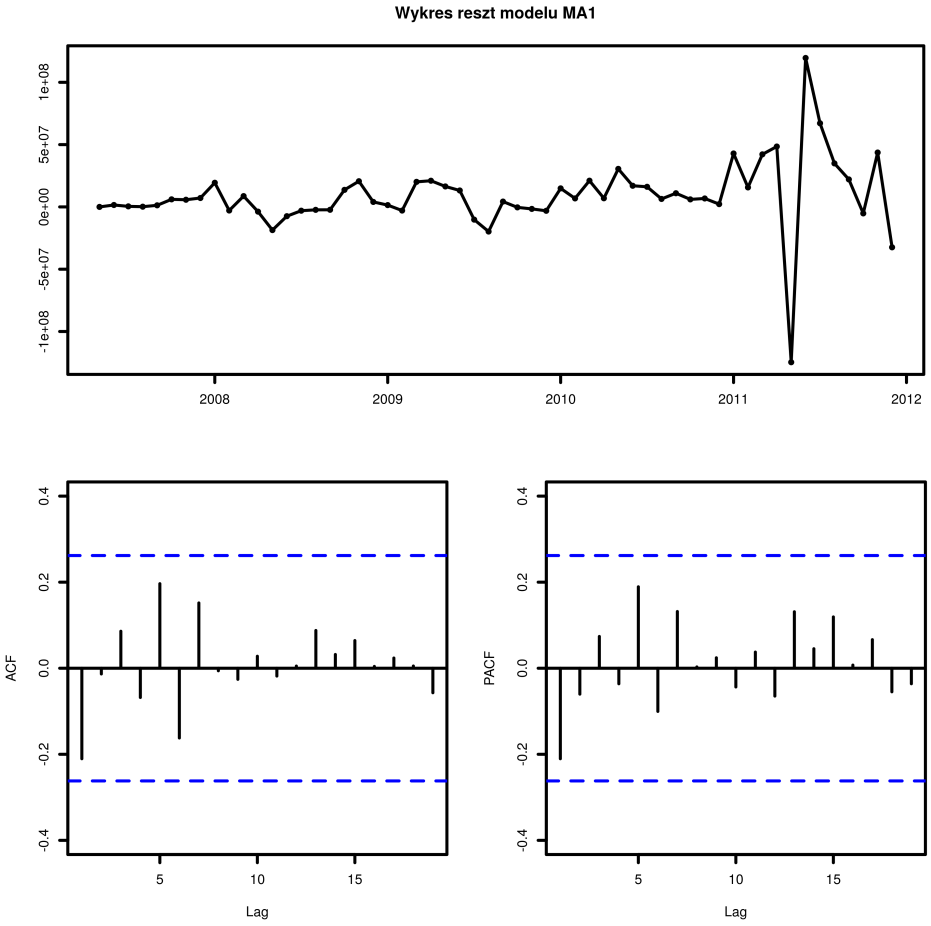
3.1 Diagnostyka reszt modeli

Dla obu testowanych modeli, wykonana została diagnostyka reszt. Na poniższych wykresach (Rysunki 8 i 9) nie widać nieregularnych wzorców ani niejednorodności wariancji.

Z ACF i PACF wynika, że nie ma istotnych korelacji



Rysunek 8: Wykres reszt modelu AR(2).



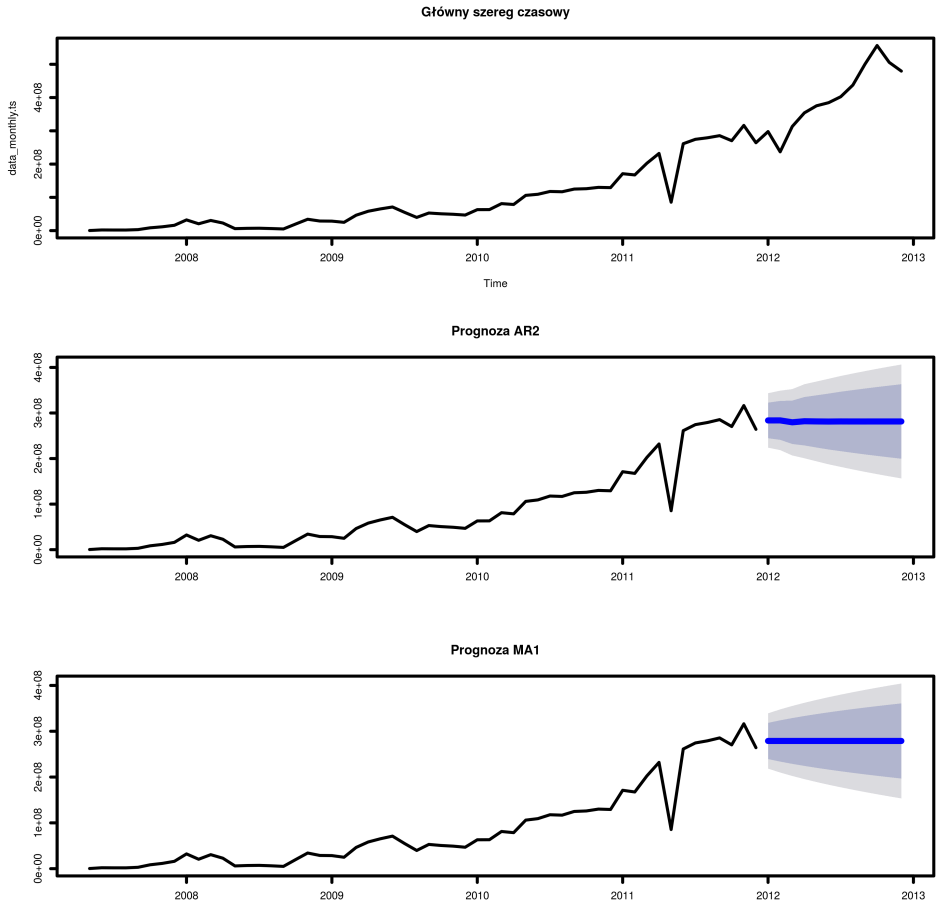
Rysunek 9: Wykres reszt modelu MA(1).

3.2 Prognozowanie

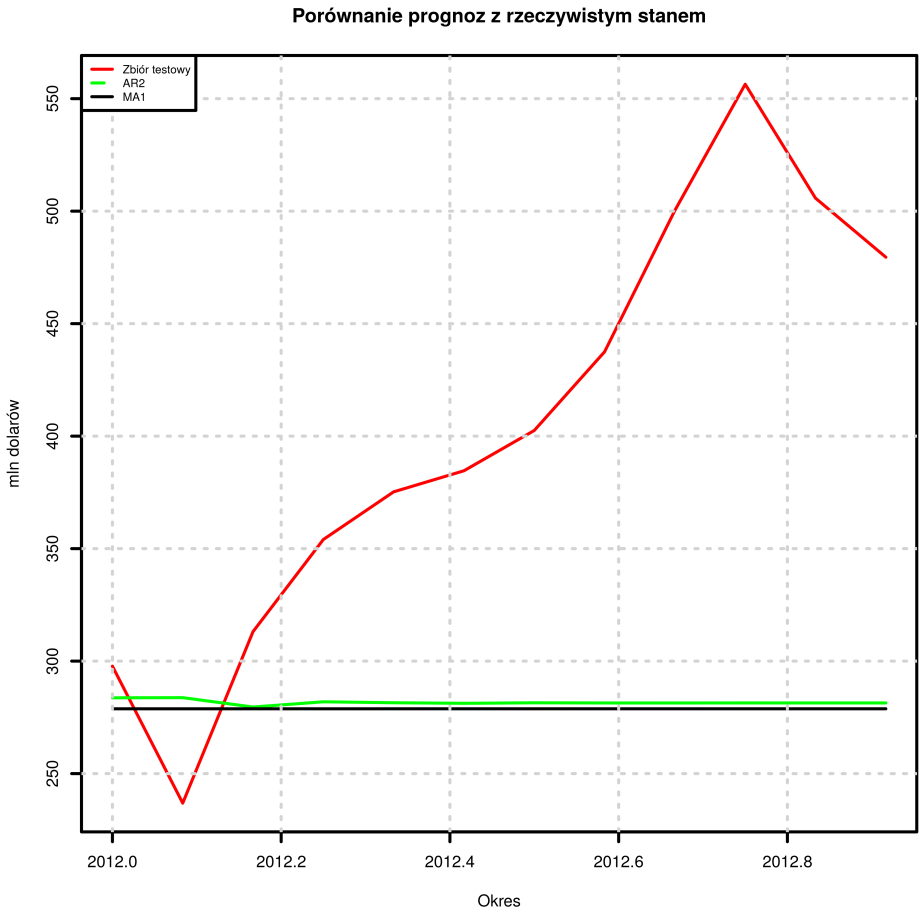
3.2.1 ARIMA

Prognozowanie zostało wykonane dla dwóch modeli ARIMA(2,1,0) oraz ARIMA(0,1,1), które możemy zauważyć na przedstawionych poniżej wykresach Rysunek 10 i 11. Niestety za pomocą tych wykresów nie jesteśmy w stanie stwierdzić, który model jest lepiej dopasowany.

W tym celu na stronie 17 wyznaczyłem oceny wykorzystując do tego kryteria błędów predykcji.



Rysunek 10: Analizowany szereg czasowy przedstawiony osobno z wyznaczonymi prognozami dla modeli.



Rysunek 11: Przyrównanie analizowanego szeregu czasowego z wyznaczonymi prognozami dla modeli.

Wykorzystałem tutaj cztery rodzaje błędów predykcji:

- **MAE** (ang. *Mean Absolute Error*) - Średni błąd całkowity;
- **RMSE** (ang. *Root Mean Squared Error*) - Błąd średniokwadratowy;
- **MAPE** (ang. *Mean Absolute Percentage Error*) - Procent średniego błędu całkowitego;
- **MASE** (ang. *Mean Absolute Scaled Error*) - Wyskalowany średni błąd całkowity.

W poniższej tabeli widać, że najniższą wartością większości błędów oprócz średniokwadratowego reprezentował sobą model **MA(1)**.

	MAE	**RMSE**	**MAPE**	**MASE**
AR(2)	17691188.13	29636744.22	33.49	0.29
MA(1)	17645678.03	30238184.06	32.97	0.29

Tablica 1: Błędy predykcji modeli ARIM.A

3.2.2 Metoda Holta

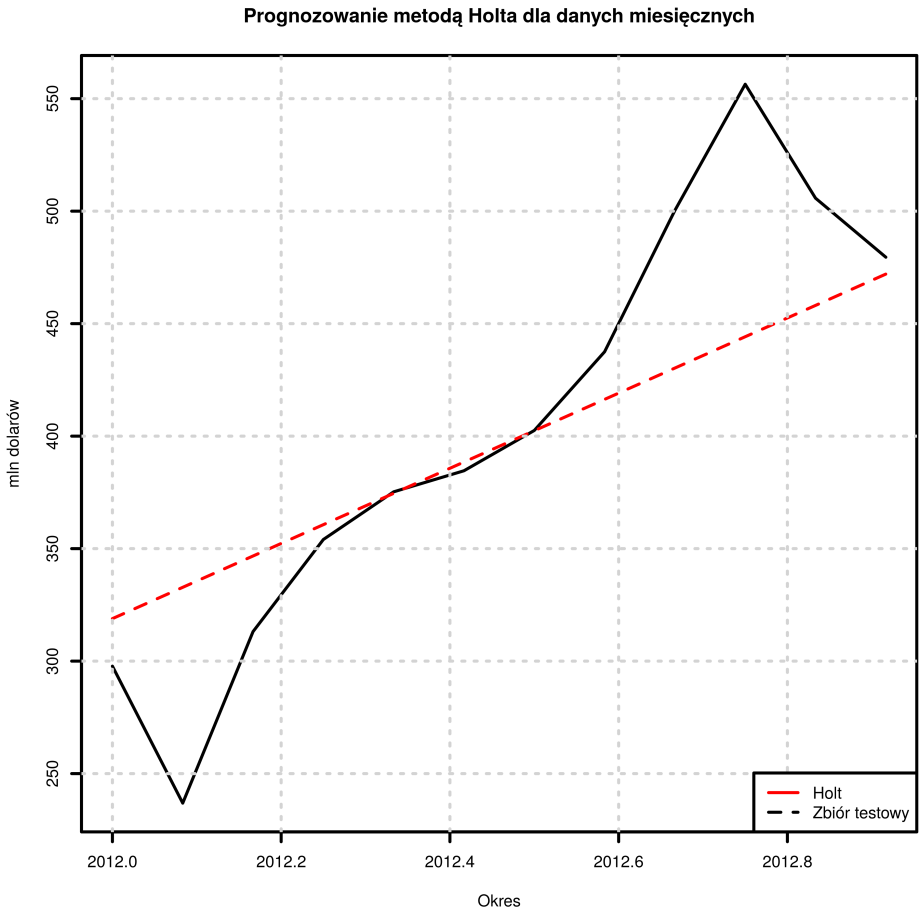
Oceniając jakość modeli, postanowiłem skorzystać z prognozowania metodą Holta, która polega na wygładzaniu wykładniczym. Metoda ta została przeze mnie wybrana, ponieważ w danych występował trend, jednakże nie posiadaliśmy sezonowości.

Prognoza ma charakter liniowy i jest przedstawiona na poniższym wykresie - Rysunek 12.

	MAE	**RMSE**	**MAPE**	**MASE**
AR(2)	18451960.86	27605036.61	76.86	0.30

Tablica 2: Błędy predykcji modelu Holta.

W powyższej tableli widzimy wcześniej omówione błędy predykcji dla metody Holta. Jak można łatwo zauważyć, metoda ta uzyskała jedynie niższy błąd średniokwadratowy. W ogólnej ocenie modelu wcale nie był on lepszy od wcześniej analizowanych modeli ARIMA.



Rysunek 12: Prognoza analizowanego szeregu metodą Holta z jego przyrównaniem do analizowanego szeregu czasowego.

4 Wnioski

Wykonane analizy pozwoliły nam wybrać optymalny model do oszacowania prognozy. Przetestowane zostały 3 modele, z czego 2 były oparte na modelu ARIMA oraz jeden na modelu Holta.

Model Holta chociaż z pierwszego punktu widzenia, patrząc na Rysunek 12, może wydawać się lepiej dopasowany od modeli ARIMA, tak błędy predykcji (w tym dwa najważniejsze MAPE oraz MASE) wykazały, że jest on gorzej dopasowany.

Z modeli ARIMA najlepiej dopasowanym jest model ARIMA(0,1,1), ponieważ ma niższy procent średniego błędu całkowitego.

Dopasowany przeze mnie model nie jest idealny, możliwe, że jest to spowodowane moim błędem w ocenie lub za małą grupą uczącą się, dzięki której moglibyśmy wyznaczyć dokładniejszy model.

Spis rysunków

Spis tablic