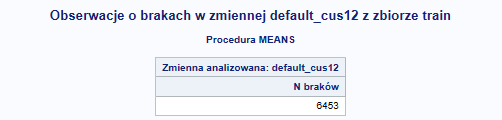
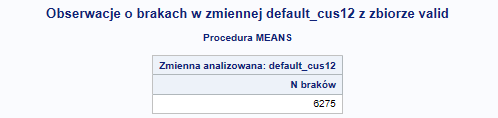
**Projekt z przedmiotu Podstawowe**   
**i zaawansowane programowanie**   
**oraz statystyka w SAS**

# **Preselekcja danych**

Niżej wymienione działania zastosowano w przypadku obu zbiorów, *train* oraz *valid*.

Na początku usunięto ze zbioru obserwacje, które miały brakujące wartości zmiennej celu *default\_cus12*. Dla zbioru *train* znaleziono 6453 obserwacji z brakami w zmiennej *default\_cus12*, natomiast dla zbioru *valid* 6275 braków. W ten sposób liczba obserwacji w zbiorze *train* zmalała z 52841 do 46388 obserwacji, a w zbiorze *valid* z 53070do 46795 obserwacji.





Kolejnym krokiem było usunięcie zmiennych, które w procesie analizy zbiorów danych okazały się niepotrzebne. Były to następujące zmienne:

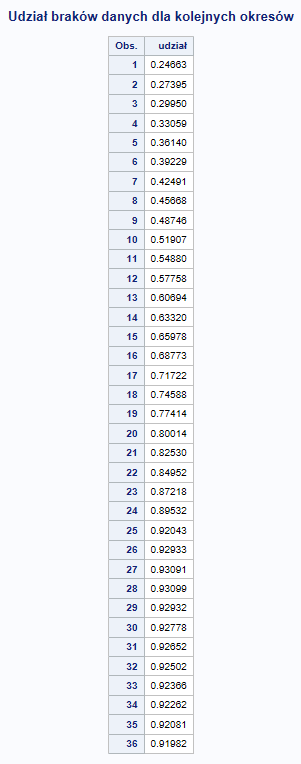
* default\_cus3, default\_cus6, default\_cus9 – zmienne celu nieanalizowane w tym projekcie,
* zmienne z przedrostkiem agr – w przypadku tych zmiennych, aby uniknąć brakujących wartości wszystkie poprzedzające dany okres wartości stanów również muszą być niebrakujące. Dysponując jednak zmiennymi z przedrostkiem ags, dla których wystarczy, że jedna wartość dla stanów poprzedzających nie jest brakująca, a zmienna ags też będzie miała wartość. Dodatkowo, zmienne z przedrostkiem agr, ze względu na ich charakterystykę cechują bardzo wysokie wartości braków zmiennych, co i tak uniemożliwiłoby skorzystanie z nich w dalszym procesie,
* act\_Cmax\_Days – powielenie wartości zmiennej act\_state\_1\_Cmax\_Days,
* act\_Cmax\_Due – powielenie wartości zmiennej act\_state\_1\_Cmax\_Due,
* act\_CMin\_Days – powielenie wartości zmiennej act\_state\_1\_Cmin\_Days,
* act\_CMin\_Due – powielenie wartości zmiennej act\_state\_1\_Cmin\_Due,
* zmienne z prefixem agsX\_Sum – posiadając pojedyncze wartości w zmiennych z prefixem act zsumowanie może się odbyć za ich pomocą,
* zmienne z prefixem agsX\_Range – posiadając zmienne agsX\_Max oraz agsX\_Min można policzyć zakres,
* zmienne z prefixem agsX\_Iqr – posiadając zmienne agsX\_Pctl25 oraz agsX\_Pctl75 można policzyć za ich pomocą rozstęp międzyćwiartkowy,
* zmienne agsX\_Csev\_all – zsumowanie może się odbyć za pomocą pojedynczych zmiennych *family*, *work*, *health*, *home,*
* zmienne actX\_Csev\_all – zsumowanie może się odbyć za pomocą pojedynczych zmiennych *family*, *work*, *health*, *home*,
* act\_cus\_loan\_number – zmienna według informacji w pliku labels.xlsx jest niepoprawnie policzona.

Po odrzuceniu wyżej wymienionych zmiennych w obu zbiorach pozostało 1138 zmiennych.

# **Analiza braków danych**

## Stabilność w czasie

Dla zmiennych o przedrostku act\_state\_X, gdzie X jest liczbą oznaczająca miesiąc, dla którego wyznaczano daną zmienną, występujących w interwale jednomiesięcznym, policzono procentowy udział braków danych według miesiąca, dla których wyliczano dane statystyki. Na poniższej tabeli widać, że wraz z upływem czasu udział braków danych ulega powiększeniu.

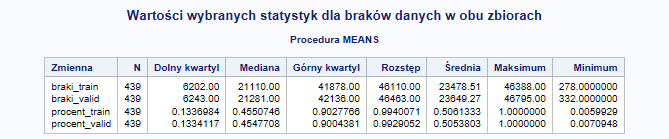


Następnie policzono udział braków danych dla zmiennych o przedrostku actX i agsX, gdzie X jest liczbą oznaczająca miesiąc, dla którego wyznaczano daną zmienną, występujących w interwale trzymiesięcznym. Na poniższej tabeli widać, że wraz z upływem czasu udział braków danych ulega zmniejszeniu. Wynika to z faktu, że pomimo rosnącego wraz z upływającym czasem udziału braków danych w zmiennych act\_state\_X, to i tak przybywało ich sumarycznie, co pozwalało wyznaczać statystyki takie jak percentyle, skośność czy kurtoza.

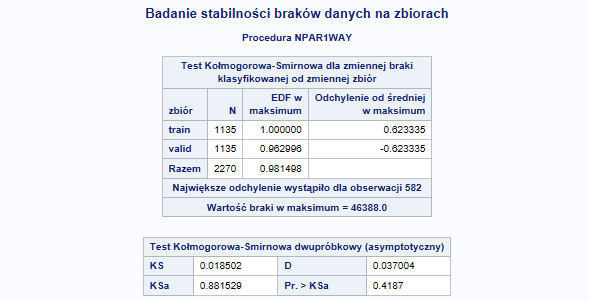


## Stabilność na zbiorach

W przypadku obu zbiorów zanotowano 439 zmiennych z brakami danych, tych samych dla obu zbiorów. Z powodu znaczącej objętości raportu analizującego liczby braków zmiennych w obu zbiorach jest on załączony oddzielnie do tego raportu jako *Załącznik2.pdf*. Poniżej zostały przedstawione wartości wybranych statystyk dla braków danych w obu zbiorach, które są bardzo zbliżone do siebie.



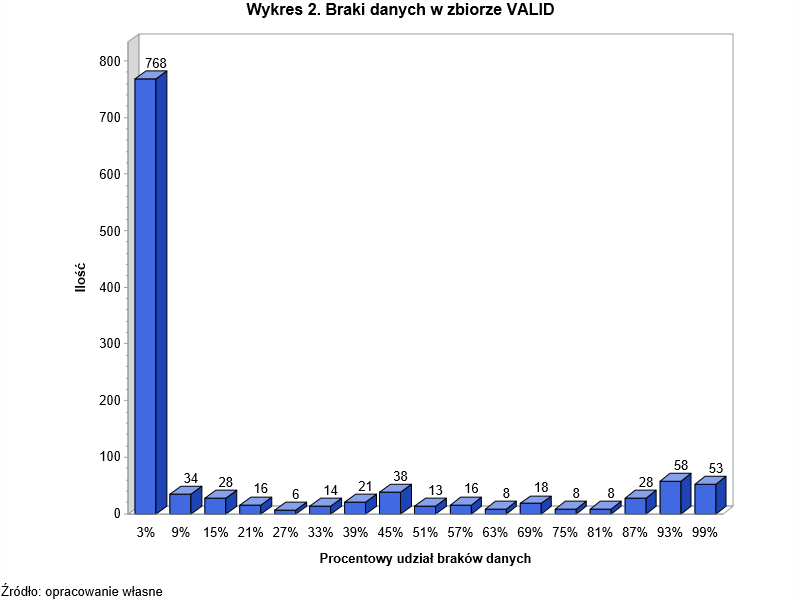
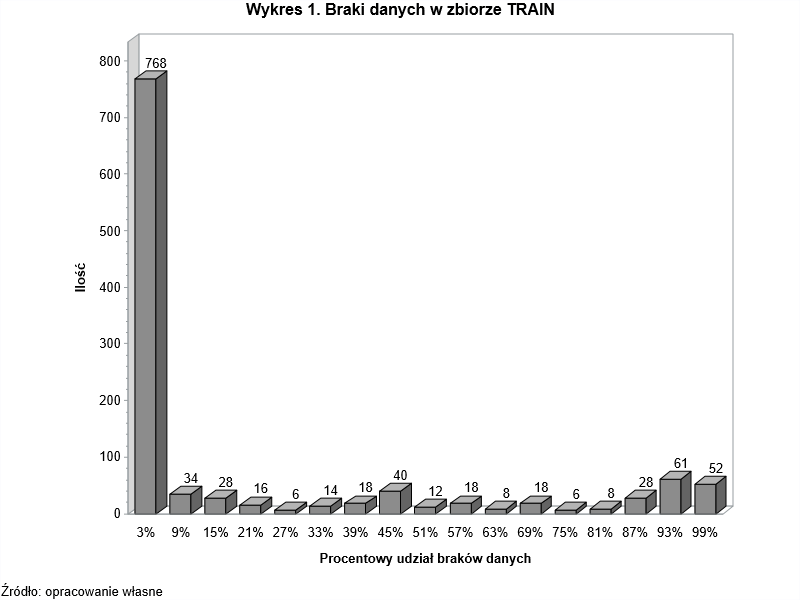
W celu zbadania zgodności rozkładów posłużono się testem Kołmogorowa-Smirnowa za pomocą proc npar1way.

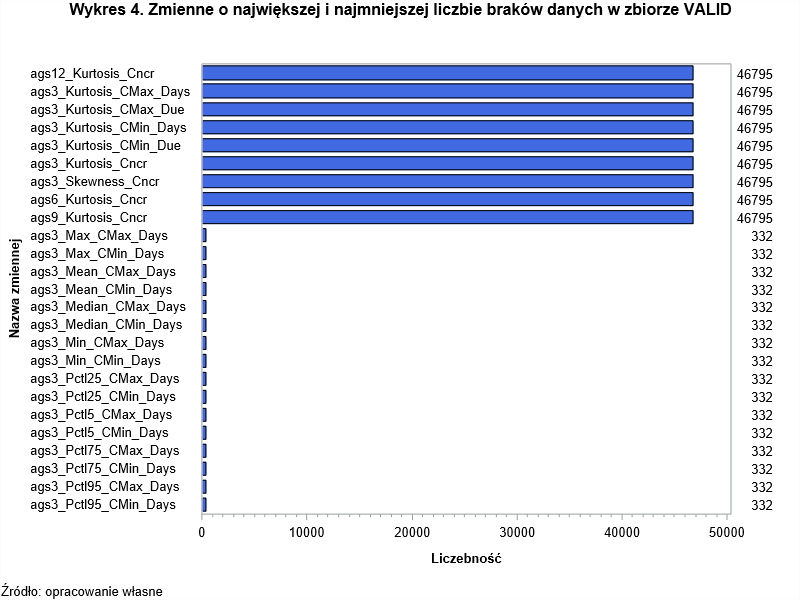
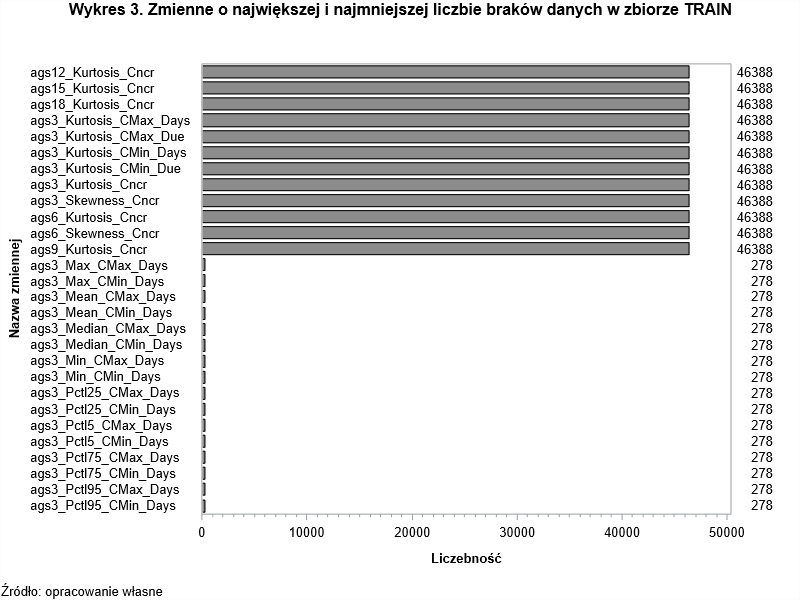


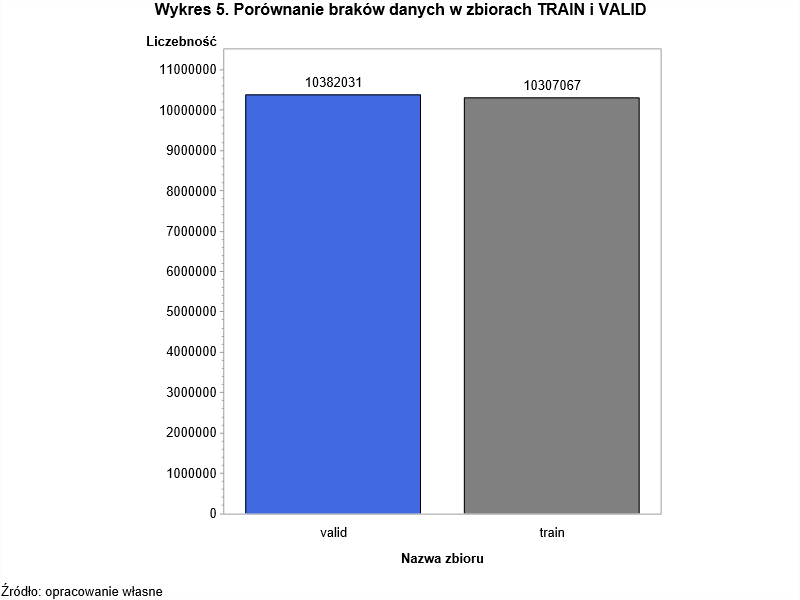
Dla otrzymanego wyniku p-value równego 0,4187 nie mamy podstaw do odrzucenia hipotezy zerowej mówiącej o zgodności rozkładów braków danych w obu zbiorach.[[1]](#footnote-2)

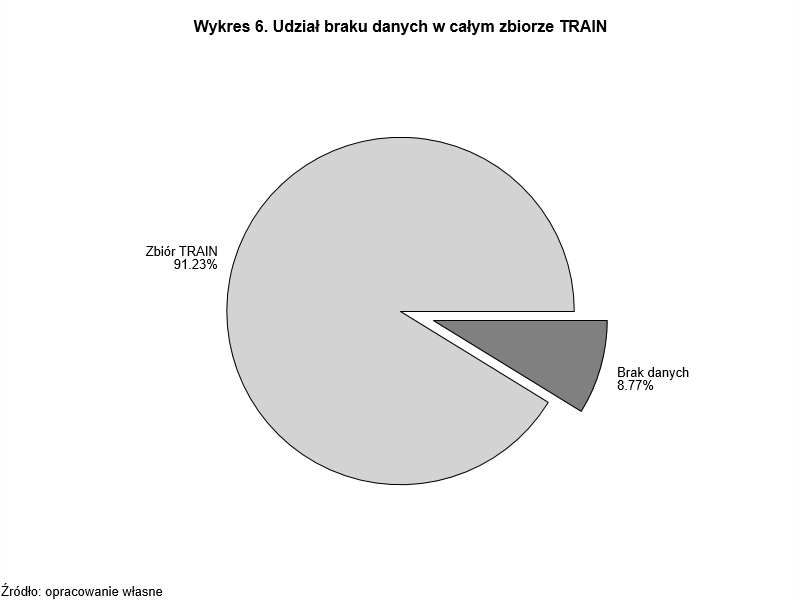
Podjęto decyzję o usunięciu z obu zbiorów tych zmiennych, w przypadku których dla co najmniej 50% obserwacji występował brak danych. W wyniku tej operacji w każdym ze zbiorów zostały 932 zmienne.

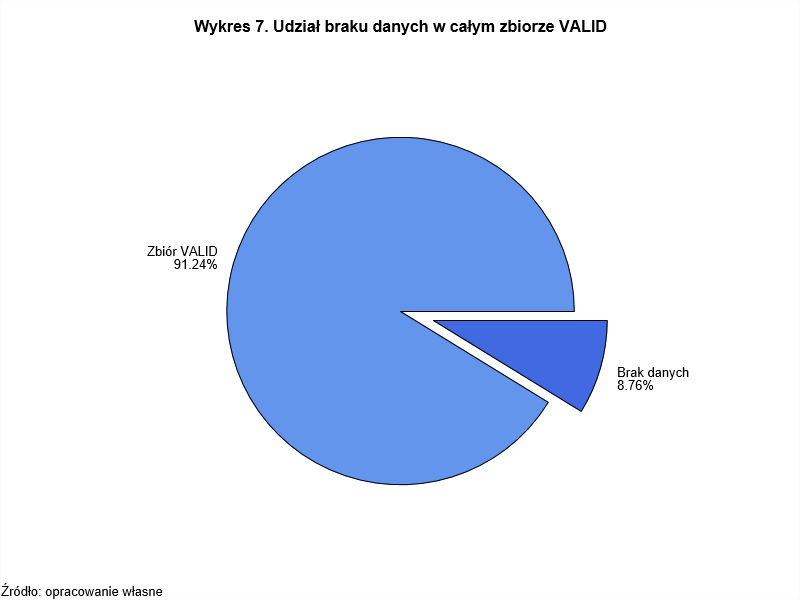
Analiza braków danych zmiennych w postaci zwizualizowanych graficznie raportów









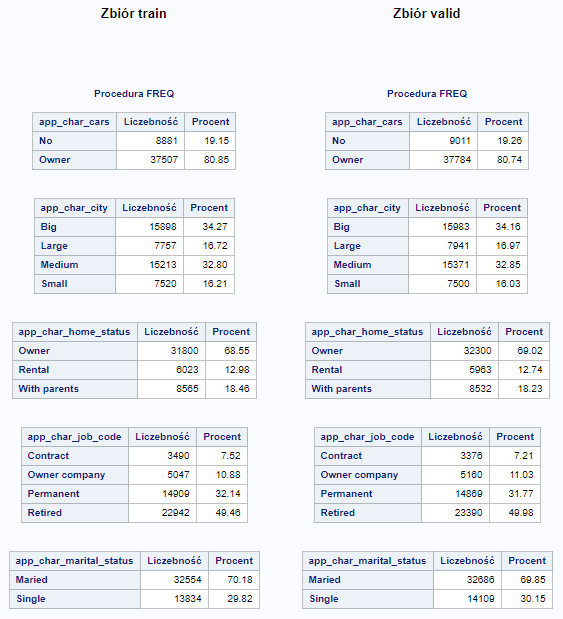


# **Analiza wartości nietypowych**

Dla zmiennych, które nie zostały odrzucone w etapie preselekcji oraz analizy braków danych, przeprowadzona została analiza wartości nietypowych. W pierwszej kolejności z obu zbiorów wyodrębniono zmienne znakowe: app\_char\_cars, app\_char\_city, app\_char\_home\_status, app\_char\_job\_code oraz app\_char\_marital\_status i dokonano ich eksploracji, wykorzystując w tym celu procedurę FREQ. Udziały poszczególnych wartości tych zmiennych były dla obu zbiorów bardzo zbliżone:

* zarówno w zbiorze valid, jak i train, ok. 80% klientów było posiadaczami samochodów;
* w obu zbiorach najwięcej było klientów, którzy mieszkali w bardzo dużych miastach, jednak w obu przypadkach udział osób mieszkających w miastach średnich był tylko niewiele mniejszy;
* prawie 70% klientów stanowili właściciele mieszkań/domów, osób mieszkających w wynajętych nieruchomościach było zaś mniej niż tych mieszkających z rodzicami;
* w obu zbiorach prawie połowę klientów stanowiły osoby na emeryturze, a tylko nieco ponad 7% osoby nie posiadające stałej umowy o pracę;
* zarówno w zbiorze valid, jak i train, ok. 70 procent klientów stanowiły osoby zamężne.

Analizując zmienne znakowe, zwłaszcza takie, które mogą przyjmować zaledwie kilka różnych wartości, trudno mówić o obserwacjach odstających czy nietypowości danych. Zestawienie wyników procedury FREQ dla zbiorów valid i train pozwala jednak wysunąć wniosek, iż zmienne te charakteryzują się dużą stabilnością na zbiorach. Udział poszczególnych kategorii zmiennych znakowych został przedstawiony w tabeli 1.

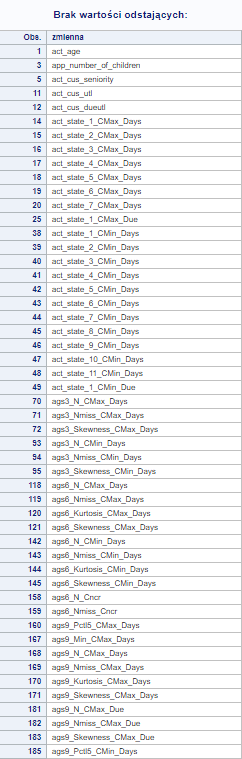


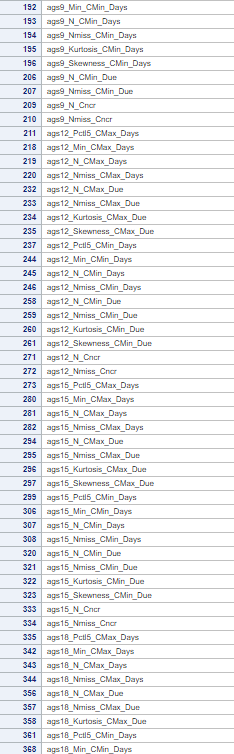
**Tabela 1 Wynik procedury FREQ dla zmiennych znakowych**źródło: opracowanie własne

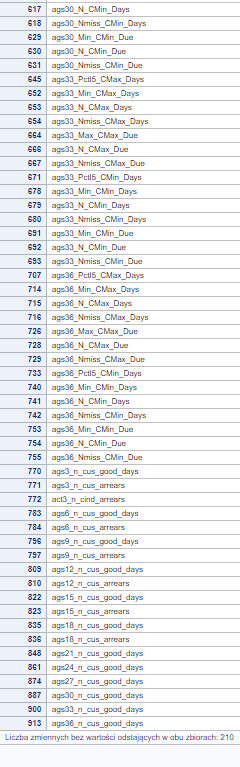
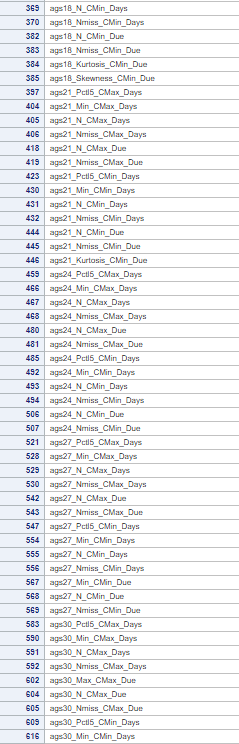
Kolejnym krokiem była analiza wartości nietypowych dla zmiennych numerycznych. Do identyfikacji obserwacji odstających na poszczególnych zmiennych zastosowano następującą metodę: dla każdej zmiennej obliczono pierwszy (Q1) i trzeci kwartyl (Q3) oraz rozstęp międzyćwiartkowy (IQR=Q3-Q1), a następnie dla każdej obserwacji sprawdzono, czy jej wartość dla danej zmiennej mieści się w przedziale od Q1 - 3\*IQR do Q3 + 3\*IQR. Jeśli tak, to oznaczana była ona jako wartość nieodstająca (poprzez przypisanie wartości 0). Jeśli nie, wówczas oznaczana była jako wartość odstająca (poprzez przypisanie wartości 1). Dzięki zastosowaniu tej metody możliwe było ustalenie, poprzez obliczenie sumy, ile razy każda z analizowanych zmiennych przyjmowała wartości nietypowe.

Operację tę przeprowadzono zarówno na zbiorze valid, jak i train. Tak jak w przypadku zmiennych znakowych, pomiędzy zbiorami nie występuje duża różnica w udziale obserwacji nietypowych dla poszczególnych zmiennych numerycznych. Zidentyfikowano 210 zmiennych, dla których w żadnym ze zbiorów nie występowały obserwacje odstające. Z uwagi na najmniejszą nietypowość danych, zmienne te mogą dostarczać cennych informacji o klientach i okazać się przydatne w modelowaniu ryzyka. Zostały one przedstawione w tabeli 2.

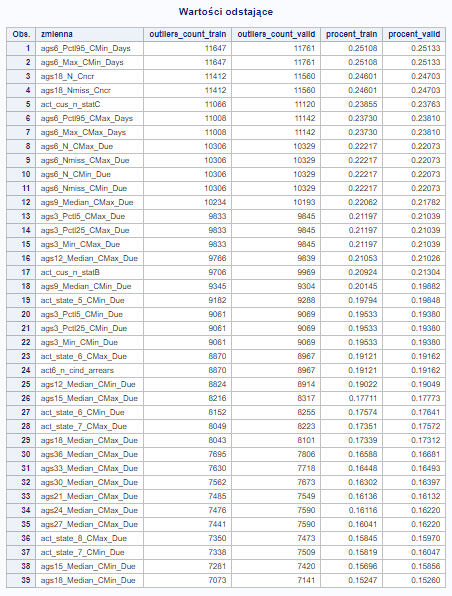
**Tabela 2 Zmienne, dla których nie odnotowano wartości odstających w żadnym ze zbiorów**  
źródło: opracowanie własne



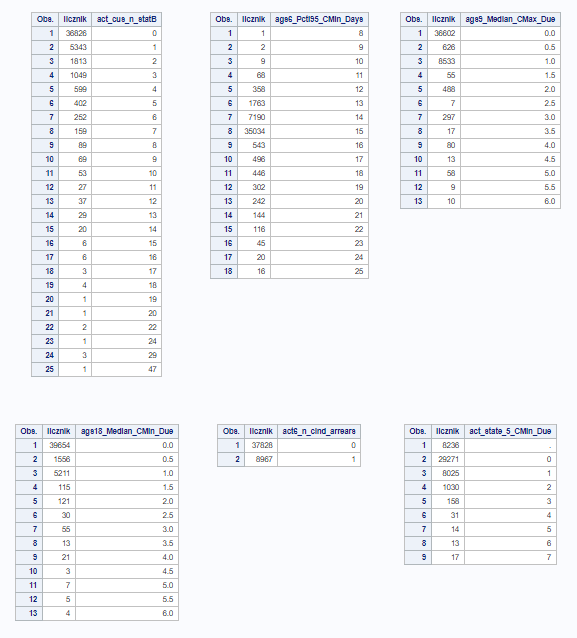




Niektóre z analizowanych zmiennych charakteryzują się natomiast bardzo dużą nietypowością danych. Dla niektórych atrybutów klienta wartości odstające, wyznaczone metodą kwartylową, stanowiły aż do 25% wszystkich obserwacji. Zmienne o udziale obserwacji odstających powyżej 15% zostały przedstawione w tabeli 3.

Spośród zaprezentowanych w tabeli 2 zmiennych o największym udziale wartości odstających wybrano losowo sześć charakterystyk, w celu dokonania bardziej szczegółowej analizy rozkładów. Uzyskane dane zaprezentowano w tabeli 3. Analiza przeprowadzona została na obserwacjach ze zbioru valid, jednak z uwagi na wspomnianą już stabilność na zbiorach, wnioski z niej płynące odnoszą się również do zbioru train.

**Tabela 3 Zmienne o największym udziale wartości odstających**źródło: opracowanie własne



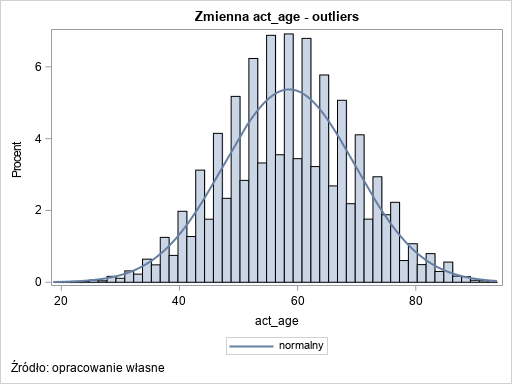
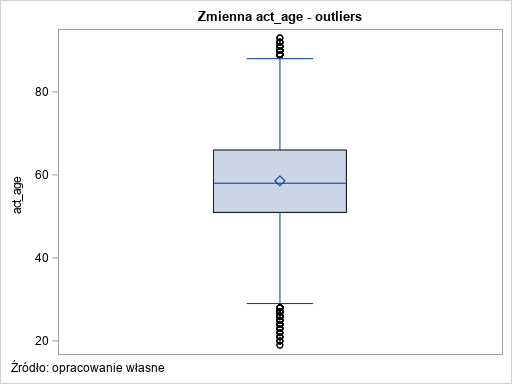
**Tabela 4 Rozkłady przykładowych zmiennych o dużym udziale wartości odstających**źródło: opracowanie własne

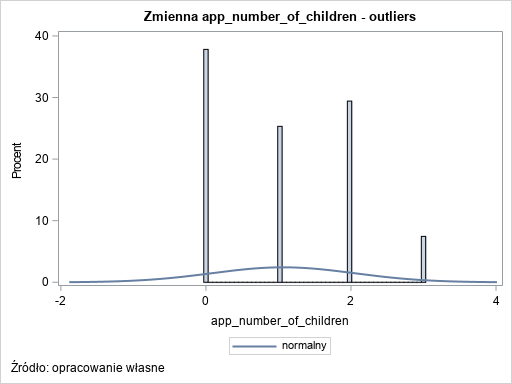
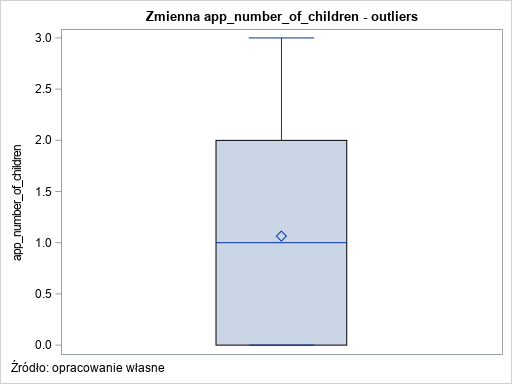
Zauważono, że w przypadku wszystkich przeanalizowanych zmiennych zdecydowanie dominowały pojedyncze ich wartości, co sprawiło, że niemalże każdy inny wynik identyfikowany był jako odstający. Zmienne, które dla tak dużej liczby obserwacji przyjmują taką samą wartość, nie mają zbyt dużej wartości predykcyjnej i raczej nie powinny być wykorzystywane do budowy modeli.

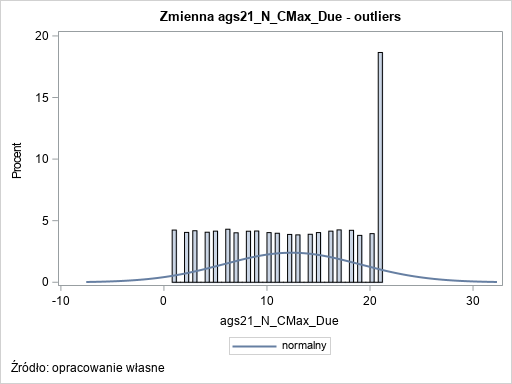
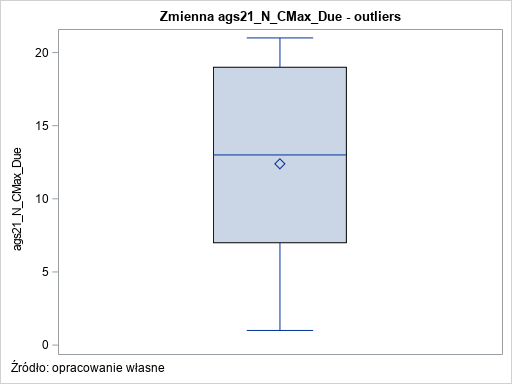
W przypadku zmiennej act6\_n\_cind\_arrears okazało się, że przyjmuje ona tylko 2 wartości – 0,1. Tego typu zmienne powinny zostać wyodrębnione i włączone do ewentualnego modelu jako zmienne klasyfikujące, a nie ciągłe.

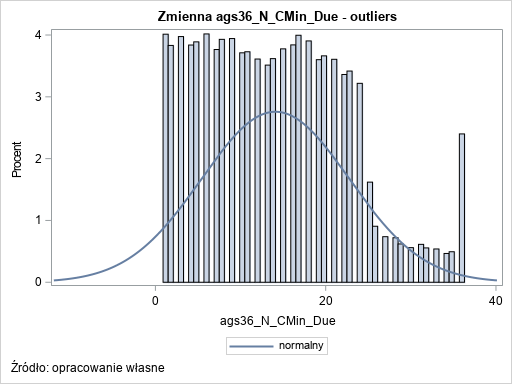
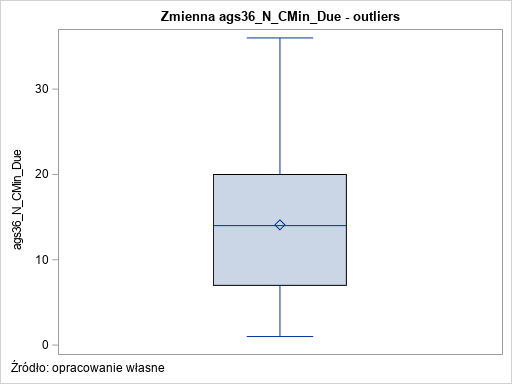
## Analiza nietypowych danych w postaci zwizualizowanych graficznie raportów

### Wykresy wartości odstających dla przykładowych zmiennych numerycznych z niedużym ich odsetkiem

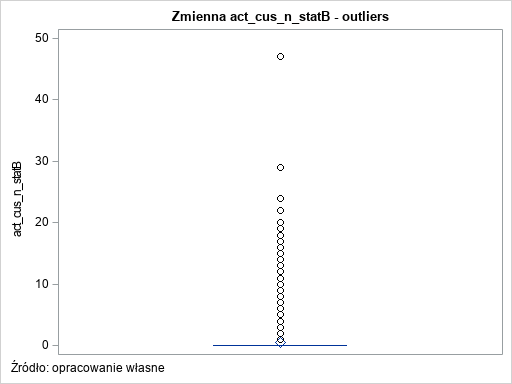
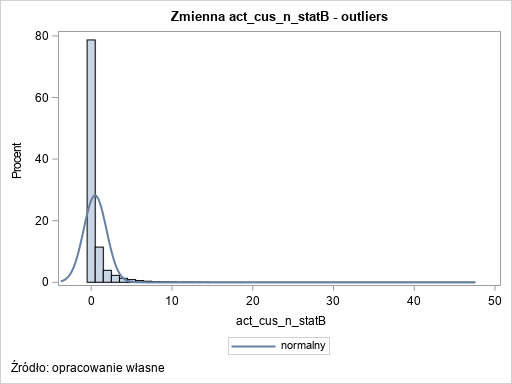


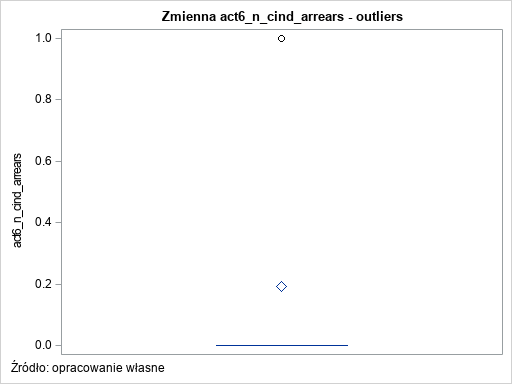
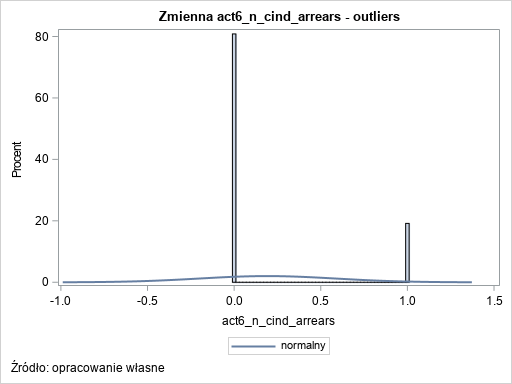


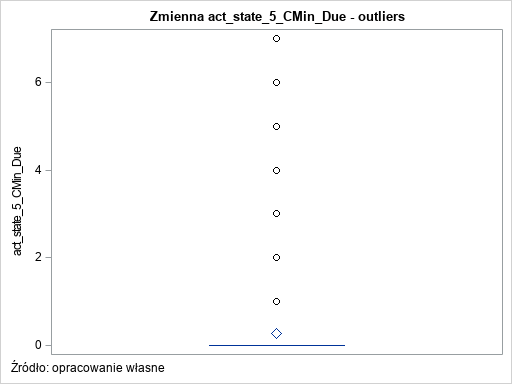
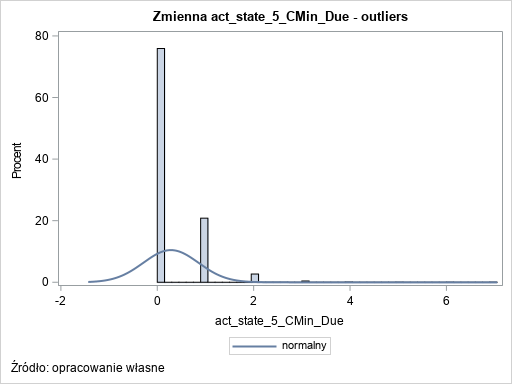


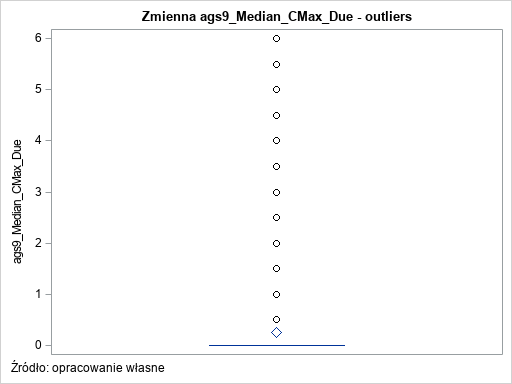
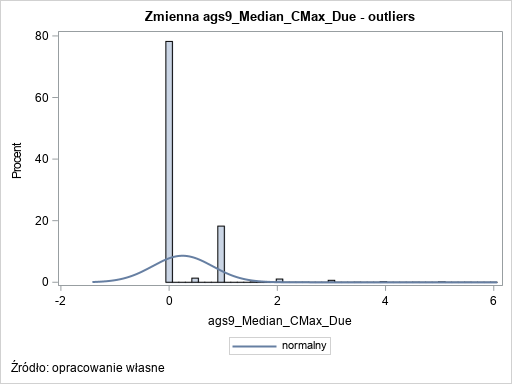


### Wykresy wartości odstających dla przykładowych zmiennych numerycznych z dużym ich odsetkiem

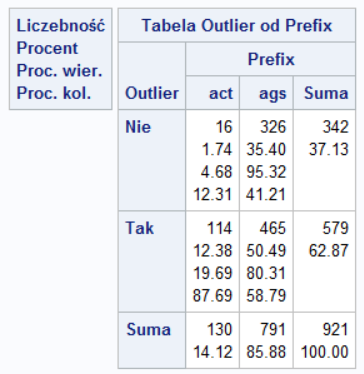
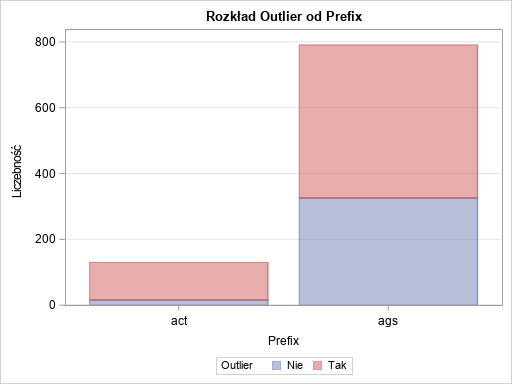
 

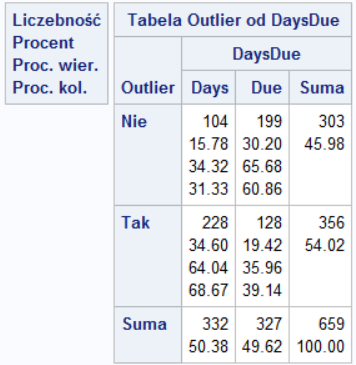
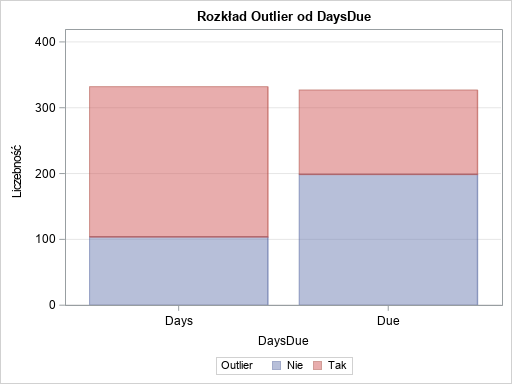
 

### Liczebność wartości odstających w poszczególnych grupach zmiennych

# **Analiza zależności pomiędzy funkcją celu a charakterystykami klienta**

W celu dokonania wyboru predyktorów, które mają największą siłę prognostyczną w zakresie odróżniania kredytów dobrych i złych, zastosowano następujące miary:

* współczynnik Giniego – do oceny siły zależności pomiędzy zmiennymi numerycznymi   
  a zmienną celu *default\_cus12*
* współczynnik V-Cramera – do oceny siły zależności pomiędzy zmiennymi tekstowymi   
  a zmienną celu *default\_cus12*

Współczynnik Giniego umożliwia bezpośrednie porównywanie jakości modeli. Miara ta przyjmuje wartości od 0 do 1. Im wyższe są wartości współczynnika Giniego, tym większą zdolność poprawnego rozróżniania kredytów dobrych i złych ma analizowany predyktor (lub model scoringowy). Idealny model ma indeks Giniego równy 1. Z kolei model, który losowo przypisuje prawdopodobieństwo defaultu dla klienta ma indeks Giniego równy 0. Przyjmuje się, że wartości współczynnika Giniego poniżej 0,35 świadczą, że model zatracił zdolność klasyfikacyjną w zakresie rozróżniania kredytów dobrych i złych. W praktyce model oceny kredytowej z Gini równym 0,4 jest uznawany za dobry.

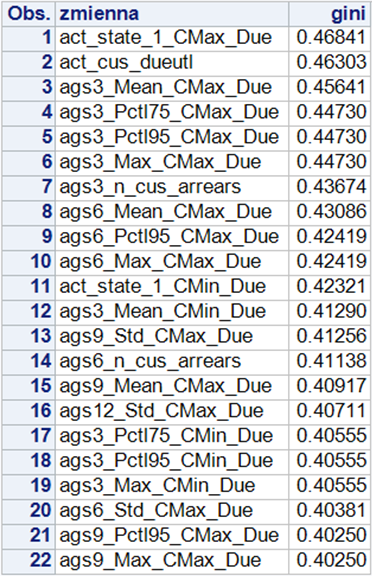
Współczynnik V-Cramera przyjmuje wartości z przedziału od 0 do 1. Im ta wartość jest bliższa 0, tym mniejsza siła związku pomiędzy zmiennymi. Z kolei im wartość tego współczynnika jest bliższa 1, tym siła związku jest większa.

Badanie wpływu zmiennych numerycznych na zmienną celu

Do oceny mocy predykcyjnej zmiennych numerycznych oszacowano modele regresji logistycznej zawierające tylko jedną zmienną objaśniającą. Predyktorem w tych modelach były poszczególne charakterystyki klienta a zmienną objaśnianą była zmienna *default\_cus12.*

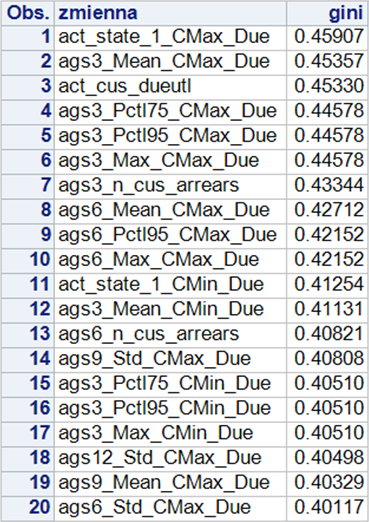
Początkowo dla zbioru train wybrano 22 a dla zbioru valid 20 najlepszych zmiennych względem kryterium mocy predykcyjnej Giniego. Jako punkt odcięcia ustalono 0.4.

**Tabela 5. Predyktory o współczynniku Giniego > 0.4 (zbiór train).**



Źródło: opracowanie własne

**Tabela 6. Predyktory o współczynniku Giniego > 0.4 (zbiór train).**



Źródło: opracowanie własne

Następnie zbadano czy wybrane zmienne nie są ze sobą skorelowane. Wykorzystano do tego celu współczynnik korelacji liniowej Pearsona. Jego wartość mieści się w przedziale <-1;1>. Wartości 1   
i –1 oznaczają zależność liniową pomiędzy dwiema cechami a 0 – całkowity brak zależności. Okazało się, że liniowa zależność występuje między następującymi zmiennymi:

* **zbiór train**
* *ags3\_Pctl75\_CMax\_Due* i *ags3\_Pctl95\_CMax\_Due*,
* *ags3\_Pctl75\_CMax\_Due* i *ags3\_Max\_CMax\_Due*,
* *ags3\_Pctl95\_CMax\_Du*e i *ags3\_Max\_CMax\_Due*,
* *ags6\_Pctl95\_CMax\_Due* i *ags6\_Max\_CMax\_Due*,
* *ags3\_Pctl75\_CMin\_Due* i *ags3\_Pctl95\_CMin\_Due*,
* *ags3\_Pctl75\_CMin\_Due* i *ags3\_Max\_CMin\_Due*,
* *ags3\_Pctl95\_CMin\_Due* i *ags3\_Max\_CMin\_Due*,
* *ags9\_Pctl95\_CMax\_Due* i *ags9\_Max\_CMax\_Due*,
* **zbiór valid**
* *ags3\_Pctl75\_CMax\_Due* i a*gs3\_Pctl95\_CMax\_Due*,
* *ags3\_Pctl75\_CMax\_Due* i *ags3\_Max\_CMax\_Due*,
* *ags3\_Pctl95\_CMax\_Due* i *ags3\_Max\_CMax\_Due*,
* *ags6\_Pctl95\_CMax\_Due* i *ags6\_Max\_CMax\_Due*,
* *ags3\_Pctl75\_CMin\_Due* i *ags3\_Pctl95\_CMin\_Due*,
* *ags3\_Pctl75\_CMin\_Due* i *ags3\_Max\_CMin\_Due*,
* *ags3\_Pctl95\_CMin\_Due* i *ags3\_Max\_CMin\_Due*.

Badanie wpływu zmiennych tekstowych na zmienną celu - zbiór train

**Tabela 7. Statystyki dla zmiennej *app\_char\_cars* od *default\_cus12.***



Źródło: opracowanie własne

**Tabela 8. Statystyki dla zmiennej *app\_char\_city* od *default\_cus12.***



Źródło: opracowanie własne

**Tabela 9. Statystyki dla zmiennej *app\_char\_home\_status* od *default\_cus12.***



Źródło: opracowanie własne

**Tabela 10. Statystyki dla zmiennej *app\_char\_job\_code* od *default\_cus12.***



Źródło: opracowanie własne

**Tabela 11. Statystyki dla zmiennej *app\_char\_marital\_status* od *default\_cus12.***



Źródło: opracowanie własne

Badanie wpływu zmiennych tekstowych na zmienną celu - zbiór valid

**Tabela 12. Statystyki dla zmiennej *app\_char\_cars* od *default\_cus12.***



Źródło: opracowanie własne

**Tabela 13. Statystyki dla zmiennej *app\_char\_city* od *default\_cus12.***



Źródło: opracowanie własne

**Tabela 14. Statystyki dla zmiennej *app\_char\_home\_status* od *default\_cus12.***



Źródło: opracowanie własne

**Tabela 15. Statystyki dla zmiennej *app\_char\_job\_code* od *default\_cus12.***



Źródło: opracowanie własne

**Tabela 16. Statystyki dla zmiennej *app\_char\_marital\_status* od *default\_cus12.***



Źródło: opracowanie własne

1. https://v8doc.sas.com/sashtml/stat/chap47/sect24.htm [dostęp: 18.01.2021 r.] [↑](#footnote-ref-2)