1.Analiza braków danych, ich udziałów w czasie (czyli stabilności w czasie) i porównywania udziałów pomiędzy zbiorami train i valid (czyli stabilności na zbiorach) w postaci szczegółowego raportu tabelarycznego.

Analiza braków - zmienne posiadające ok. 100% braków (próba 50 zmiennych numerycznych)

Obs.	nazwa	N_braki_train	N_braki_valid	proc_braki_train	proc_braki_valid	roznica_braki_train	roznica_braki_valid
1	act_state_28_Cncr	52640	52855	100%	100%	0%	0%
2	act_state_29_Cncr	52621	52873	100%	100%	0%	0%
3	act_state_30_Cncr	52639	52830	100%	100%	0%	0%
4	act_state_31_Cncr	52625	52835	100%	100%	0%	0%
5	act_state_32_Cncr	52596	52805	100%	100%	0%	0%
6	act_state_33_Cncr	52609	52790	100%	99%	0%	0%
7	act_state_34_Cncr	52584	52824	100%	100%	0%	0%
8	agr12_Iqr_Cncr	52841	53070	100%	100%	0%	0%
9	agr12_Kurtosis_Cncr	52841	53070	100%	100%	0%	0%
10	agr12_Max_Cncr	52841	53070	100%	100%	0%	0%
11	agr12_Mean_Cncr	52841	53070	100%	100%	0%	0%
12	agr12_Median_Cncr	52841	53070	100%	100%	0%	0%
13	agr12_Min_Cncr	52841	53070	100%	100%	0%	0%
14	agr12_N_Cncr	52841	53070	100%	100%	0%	0%
15	agr12_Nmiss_Cncr	52841	53070	100%	100%	0%	0%
16	agr12_Pctl25_Cncr	52841	53070	100%	100%	0%	0%
17	agr12_Pctl5_Cncr	52841	53070	100%	100%	0%	0%
18	agr12_Pctl75_Cncr	52841	53070	100%	100%	0%	0%
19	agr12_Pctl	52841	53070	100%	100%	0%	0%

	95_Cncr						
20	agr12_Range_Cncr	52841	53070	100%	100%	0%	0%
21	agr12_Skewness_Cncr	52841	53070	100%	100%	0%	0%
22	agr12_Std_Cncr	52841	53070	100%	100%	0%	0%
23	agr12_Sum_Cncr	52841	53070	100%	100%	0%	0%
24	agr15_Iqr_Cncr	52841	53070	100%	100%	0%	0%
25	agr15_Kurtosis_Cncr	52841	53070	100%	100%	0%	0%
26	agr15_Max_Cncr	52841	53070	100%	100%	0%	0%
27	agr15_Mean_Cncr	52841	53070	100%	100%	0%	0%
28	agr15_Median_Cncr	52841	53070	100%	100%	0%	0%
29	agr15_Min_Cncr	52841	53070	100%	100%	0%	0%
30	agr15_N_Cncr	52841	53070	100%	100%	0%	0%
31	agr15_Nmiss_Cncr	52841	53070	100%	100%	0%	0%
32	agr15_Pctl25_Cncr	52841	53070	100%	100%	0%	0%
33	agr15_Pctl5_Cncr	52841	53070	100%	100%	0%	0%
34	agr15_Pctl75_Cncr	52841	53070	100%	100%	0%	0%
35	agr15_Pctl95_Cncr	52841	53070	100%	100%	0%	0%
36	agr15_Range_Cncr	52841	53070	100%	100%	0%	0%
37	agr15_Skewness_Cncr	52841	53070	100%	100%	0%	0%
38	agr15_Std_Cncr	52841	53070	100%	100%	0%	0%
39	agr15_Sum_Cncr	52841	53070	100%	100%	0%	0%
40	agr18_Iqr_Cncr	52841	53070	100%	100%	0%	0%
41	agr18_Kurtosis_Cncr	52841	53070	100%	100%	0%	0%
42	agr18_Max_Cncr	52841	53070	100%	100%	0%	0%

43	agr18_Mean_Cncr	52841	53070	100%	100%	0%	0%
44	agr18_Median_Cncr	52841	53070	100%	100%	0%	0%
45	agr18_Min_Cncr	52841	53070	100%	100%	0%	0%
46	agr18_N_Cncr	52841	53070	100%	100%	0%	0%
47	agr18_Nmiss_Cncr	52841	53070	100%	100%	0%	0%
48	agr18_Pctl25_Cncr	52841	53070	100%	100%	0%	0%
49	agr18_Pctl5_Cncr	52841	53070	100%	100%	0%	0%
50	agr18_Pctl75_Cncr	52841	53070	100%	100%	0%	0%

Analiza braków - zmienne posiadające 0% braków (próba 50 zmiennych numerycznych)

Obs.	nazwa	N_braki_train	N_braki_valid	proc_braki_train	proc_braki_valid	roznica_braki_train	roznica_braki_valid
1	act12_Ciev_all	0	0	0%	0%	0%	0%
2	act12_Ciev_family	0	0	0%	0%	0%	0%
3	act12_Ciev_health	0	0	0%	0%	0%	0%
4	act12_Ciev_home	0	0	0%	0%	0%	0%
5	act12_Ciev_work	0	0	0%	0%	0%	0%
6	act12_Cncr_taken	0	0	0%	0%	0%	0%
7	act12_n_cind_arrears	0	0	0%	0%	0%	0%
8	act15_Ciev_all	0	0	0%	0%	0%	0%

9	act15_Ciev_family	0	0	0%	0%	0%	0%
10	act15_Ciev_health	0	0	0%	0%	0%	0%
11	act15_Ciev_home	0	0	0%	0%	0%	0%
12	act15_Ciev_work	0	0	0%	0%	0%	0%
13	act15_Cncr_taken	0	0	0%	0%	0%	0%
14	act15_n_cind_arrears	0	0	0%	0%	0%	0%
15	act18_Ciev_all	0	0	0%	0%	0%	0%
16	act18_Ciev_family	0	0	0%	0%	0%	0%
17	act18_Ciev_health	0	0	0%	0%	0%	0%
18	act18_Ciev_home	0	0	0%	0%	0%	0%
19	act18_Ciev_work	0	0	0%	0%	0%	0%
20	act18_Cncr_taken	0	0	0%	0%	0%	0%
21	act18_n_cind_arrears	0	0	0%	0%	0%	0%
22	act21_Ciev_all	0	0	0%	0%	0%	0%
23	act21_Ciev_family	0	0	0%	0%	0%	0%
24	act21_Ciev_health	0	0	0%	0%	0%	0%
25	act21_Ciev_home	0	0	0%	0%	0%	0%

26	act21_Ciev_work	0	0	0%	0%	0%	0%
27	act21_Cncr_taken	0	0	0%	0%	0%	0%
28	act21_n_cind_arrears	0	0	0%	0%	0%	0%
29	act24_Ciev_all	0	0	0%	0%	0%	0%
30	act24_Ciev_family	0	0	0%	0%	0%	0%
31	act24_Ciev_health	0	0	0%	0%	0%	0%
32	act24_Ciev_home	0	0	0%	0%	0%	0%
33	act24_Ciev_work	0	0	0%	0%	0%	0%
34	act24_Cncr_taken	0	0	0%	0%	0%	0%
35	act24_n_cind_arrears	0	0	0%	0%	0%	0%
36	act27_Ciev_all	0	0	0%	0%	0%	0%
37	act27_Ciev_family	0	0	0%	0%	0%	0%
38	act27_Ciev_health	0	0	0%	0%	0%	0%
39	act27_Ciev_home	0	0	0%	0%	0%	0%
40	act27_Ciev_work	0	0	0%	0%	0%	0%
41	act27_Cncr_taken	0	0	0%	0%	0%	0%
42	act27_n_cind_arrears	0	0	0%	0%	0%	0%

43	act30_ Ciev_all	0	0	0%	0%	0%	0%
44	act30_ Ciev_fa mily	0	0	0%	0%	0%	0%
45	act30_ Ciev_he alth	0	0	0%	0%	0%	0%
46	act30_ Ciev_ho me	0	0	0%	0%	0%	0%
47	act30_ Ciev_w ork	0	0	0%	0%	0%	0%
48	act30_ Cncr_ta ken	0	0	0%	0%	0%	0%
49	act30_n _cind_a rrears	0	0	0%	0%	0%	0%
50	act33_ Ciev_all	0	0	0%	0%	0%	0%

Opis zmiennych:

N_braki_train – liczba braków w zbiorze train
N_braki_valid – liczba braków w zbiorze valid

proc_braki_train – procentowa liczba braków w zbiorze train
proc_braki_valid – procentowa liczba braków w zbiorze valid

roznica_braki_train – wskaźnik jak zmieniały się braki w czasie dla zbioru train (Jeśli różnica była mała np. 2 % oznacza to, że braki w poszczególnych miesiącach nie odbiegały od średniej braków w czasie)
oznica_braki_valid– wskaźnik jak zmieniały się braki w czasie dla zbioru valid (Jeśli różnica była mała np. 2 % oznacza to, że braki w poszczególnych miesiącach nie odbiegały od średniej braków w czasie)

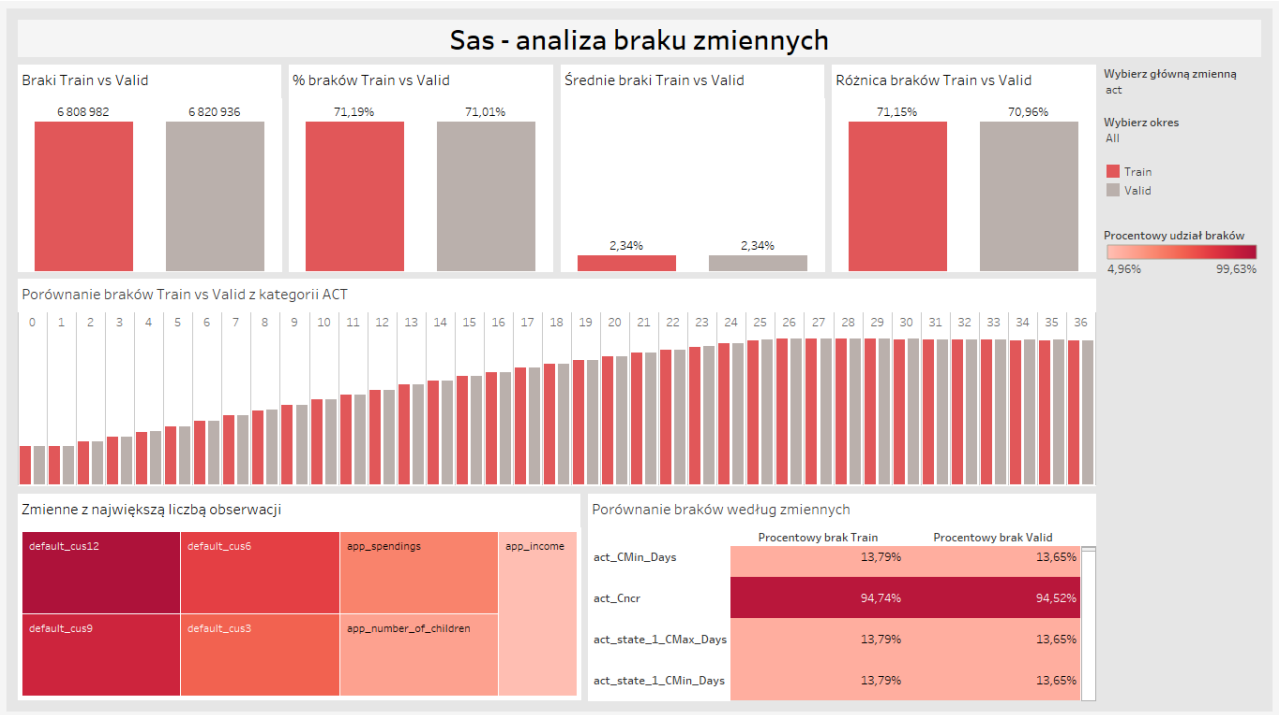
2. Analiza braków danych w postaci zwizualizowanych graficznych raportów w celu szybkiego identyfikowania zmiennych z większą i mniejszą liczbą braków danych oraz z ich różną stabilnością.

Analiza braków - część interaktywna.

Poniższa wizualizacja opiera się o kod SAS, który został następnie przekształcony w plik CSV i wykorzystany w Tableau. Jest to aplikacja, która pozwala na tworzenie interaktywnych i przyjaznych użytkownikowi raportów. Rozwiązania Business Inteligence pozwalają na szybkie podejmowanie decyzji i łatwe dostrzeganie różnic, poprzez zaznaczanie odpowiednich miar z wykorzystaniem koloru lub kształtu.

Uruchom poniższy link. W przypadku problemów, prosilibyśmy o kontakt z grupą.

https://public.tableau.com/app/profile/tomaszk/viz/SASProjekt/Dashboard1



Interpretacja zestawu wykresów” Sas - analiza braku zmiennych”.

Każda z brakujących została przypisana do jednej z 4 kategori: act, agr, ags, oraz Zmiennej Default. Dodatkowo każda z kategorii posiada okresy czasowe dla danej zmiennej. Chcąc poznać poziom wybrakowania zmiennych można dostosowywać parametry wybierając konkretne zmienne lub zaznaczając opcję All która ukaże całość danych.

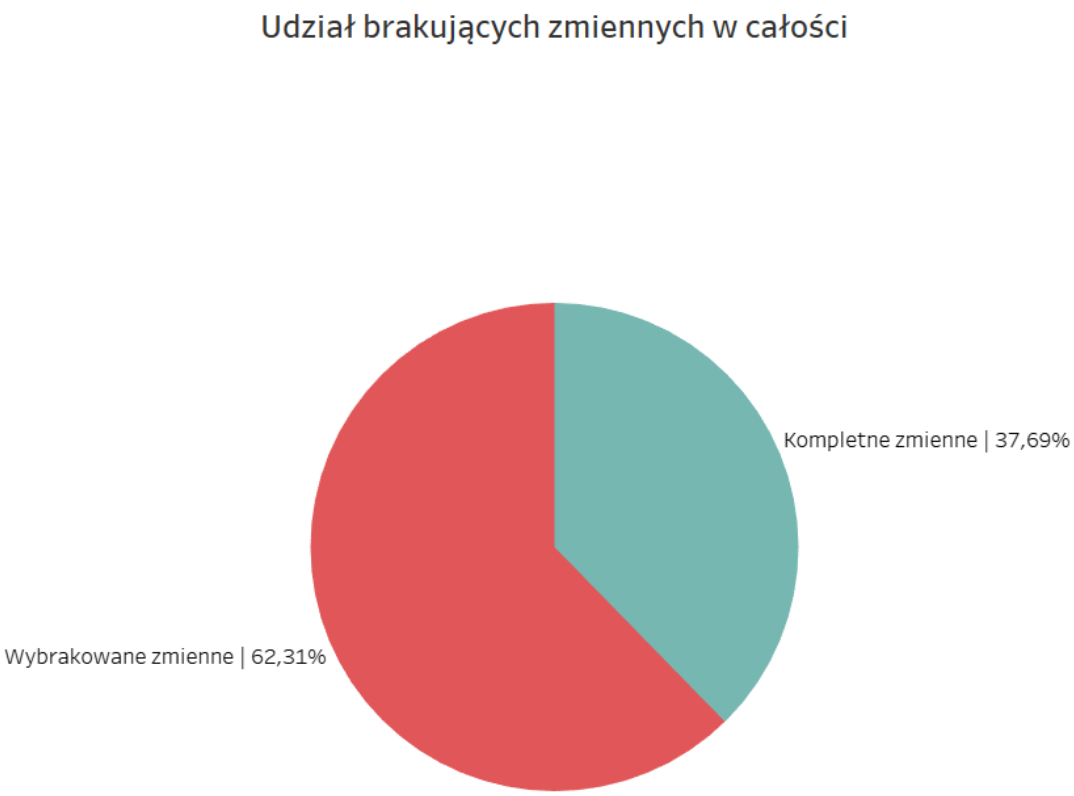
Podstawowe miary to:

- Numeryczna liczba braków w zbiorze.
- Procentowa liczba braków w zbiorze.
- Średnia liczba braków dla danej zmiennej w czasie w wartości procentowej.
- Opis zmian braków w czasie w wyrażeniu procentowym.

Analiza braków - część statystyczna.

W wskazanych danych znajduje się 2310 zmiennych, na które składa się 52841 obserwacji zbioru Train, oraz 53070 obserwacji zbioru Valid. Pierwszym krokiem było wyfiltrowanie danych które były niepełne. Tym samym liczba zmiennych spadła z 2310 do 1435. Efektem tego są dwa wnioski:

- 63.31 % wszystkich zmiennych jest niekompletna (1435)
- 37.69 % wszystkich zmiennych posiada komplet danych (868)

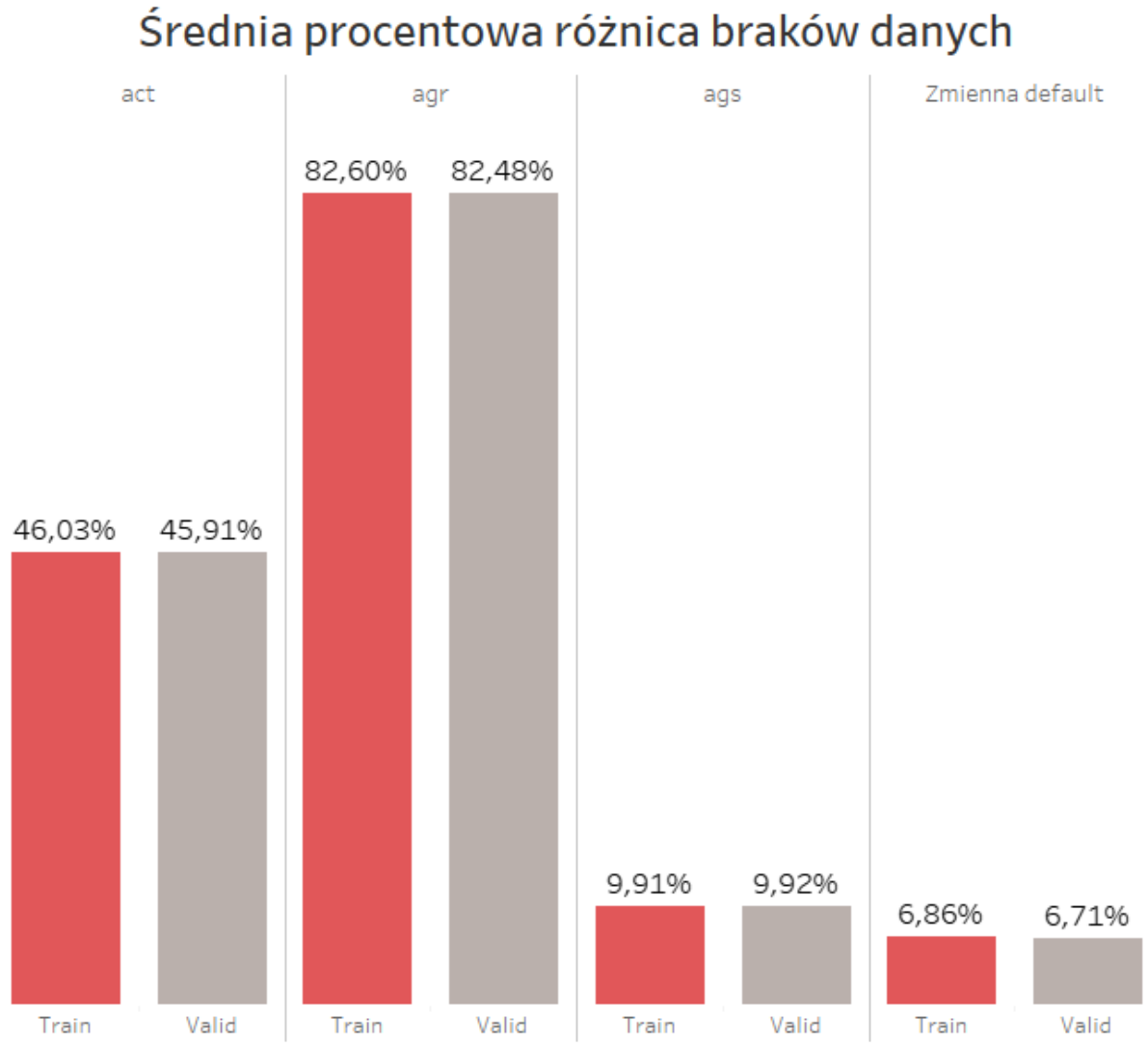


Każdą z brakującch zmiennych można skategoryzować do jednej z czterech grup:

- Act
- Agr
- Ags
- Default

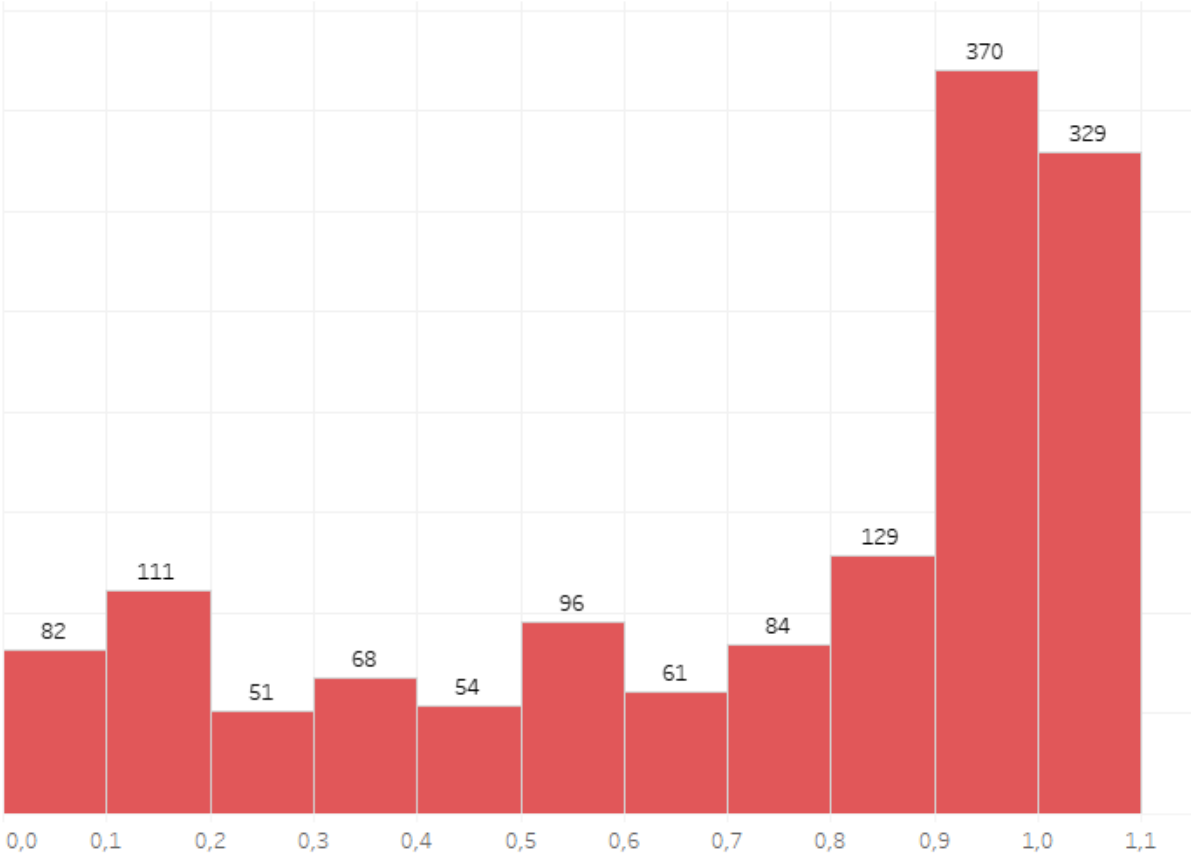
Dodatkowo każda z wyżej wymienionych pozycji posiada wartość Time, jest to miesiąc, w którym dana obserwacja została zbadana. Zależnie od zmiennej, mogą one przyjmować wartości od 1 do 36.

Poniższy wykres przedstawia średnią procentową różnicę braków danych według kategorii. Można zauważyć, iż dane ze zbioru Train są większe w przypadku act, agr oraz zmiennej default. Jedynie w zmiennej ags procentowy udział braków danych jest większy od zbioru Train.

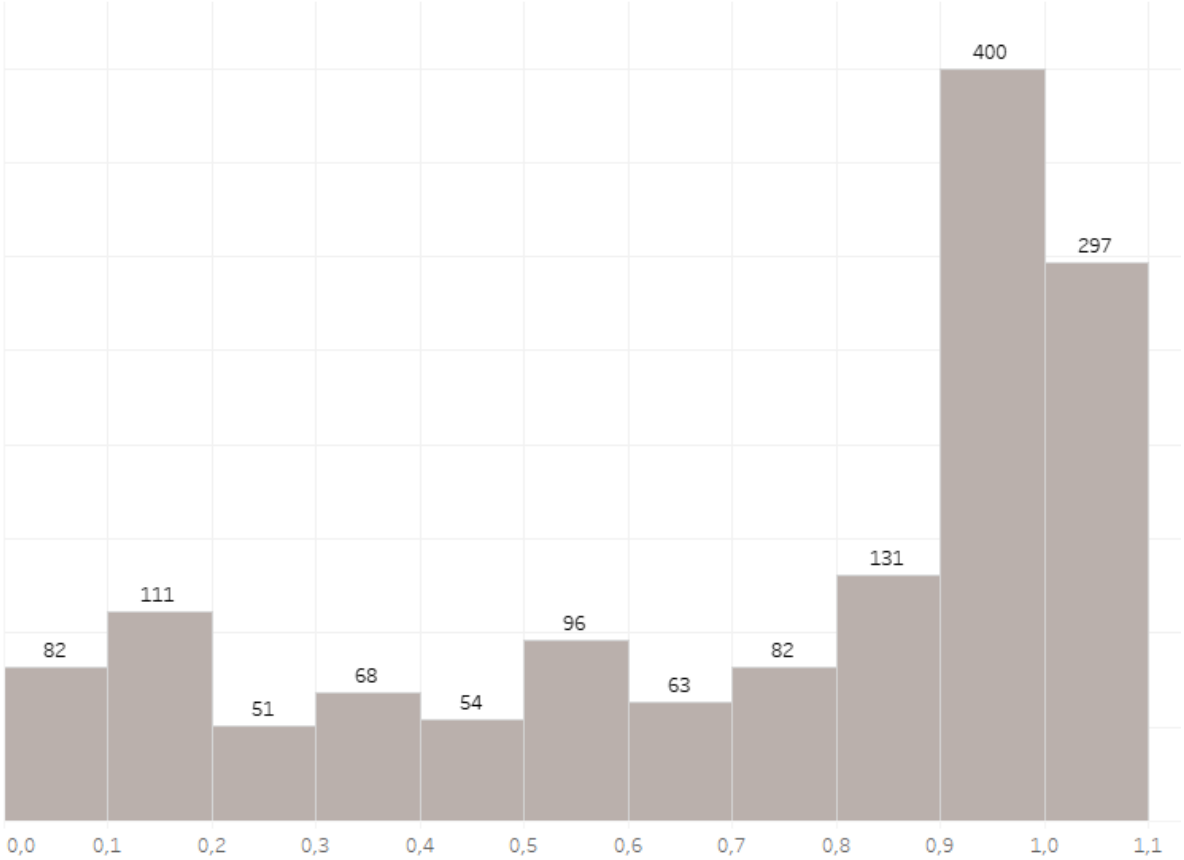


Poniższe dwa wykresy prezentują brak danych w zmiennych train oraz valid. Można tutaj zauważyć, iż w zbiorze Train nlisko 49,2 % danych jest wybrakowanych w przedziale 90-100%. Dla porównania, w zbiorze valid ten wskaźnik wynosi ok 48,5 %.

Ile zmiennych zbioru TRAIN odpowiada za procentowy udział w braku danych?



Ile zmiennych zbioru VALID odpowiada za procentowy udział w braku danych?



3. Analiza wartości nietypowych, odstających lub ogólnie nieregularności rozkładów, ich udziałów w czasie (czyli stabilności w czasie) i porównywania udziałów pomiędzy zbiorami train i valid (czyli stabilności na zbiorach) w postaci szczegółowego raportu tabelarycznego.

Zmienne charakteryzowały się dużą stabilnością na zbiorach. Na początku zostały porównane wybrane zmienne tekstowe w celu porównania zbiorów train i valid.

Sprawdzenie zmiennych tesktowych

Podgląd zmiennych tekstowych w zbiorze train

Procedura FREQ

app_char_cars	Liczebność	Procent
No	10169	19.24
Owner	42672	80.76

app_char_city	Liczebność	Procent
Big	18009	34.08
Large	8818	16.69
Medium	17433	32.99
Small	8581	16.24

app_char_home_status	Liczebność	Procent
Owner	36197	68.50
Rental	6841	12.95
With parents	9803	18.55

app_char_job_code	Liczebność	Procent
Contract	4033	7.63
Owner company	5760	10.90
Permanent	16983	32.14
Retired	26065	49.33

app_char_marital_status	Liczebność	Procent
Maried	37135	70.28
Single	15706	29.72

Podgląd zmiennych tekstowych w zbiorze valid

Procedura FREQ

app_char_cars	Liczebność	Procent
No	10238	19.29
Owner	42832	80.71

app_char_city	Liczebność	Procent
Big	18029	33.97
Large	8980	16.92
Medium	17475	32.93
Small	8586	16.18

app_char_home_status	Liczebność	Procent
Owner	36565	68.90
Rental	6734	12.69
With parents	9771	18.41

app_char_job_code	Liczebność	Procent
Contract	3926	7.40
Owner company	5880	11.08
Permanent	16864	31.78
Retired	26400	49.75

app_char_marital_status	Liczebność	Procent
Maried	37113	69.93
Single	15957	30.07

Udział procentowy jest inny, ale są w podobnych proporcjach.

Podgląd zmiennych tekstowych w zbiorze valid

Procedura FREQ

act_cus_loan_number	Liczebność	Procent
1	52746	99.82
2	95	0.18

Podgląd zmiennych tekstowych w zbiorze train

Procedura FREQ

act_cus_loan_number	Liczebność	Procent
1	52746	99.82
2	95	0.18

Podgląd zmiennych tekstowych w zbiorze valid

Procedura MEANS

Zmienna	N	N braków	Średnia	Odch. std.
act_age	53070	0	58.4961937	11.1306192
app_income	53070	0	2090.91	1766.47
app_number_of_children	53070	0	1.0684379	0.9832247
app_spendings	53070	0	563.4038063	617.1927008

Podgląd zmiennych tekstowych w zbiorze train

Procedura MEANS

Zmienna	N	N braków	Średnia	Odch. std.
act_age	52841	0	58.3719082	11.1397574
app_income	52841	0	2092.80	1754.77
app_number_of_children	52841	0	1.0726898	0.9827602
app_spendings	52841	0	564.4151322	620.8676737

Na załączonych tabelach ze zbioru train i valid nie ma braków danych. Dane te też mają bardzo podobne wartości.

Jak widać wyżej zbiory są te do siebie bardzo podobne, są stabilne, nie zaobserwowano nieregularności rozkładów.

Nie ma zmiennych odbiegających od spodziewanego wieku.

Za pomocą funkcji distinct zostały zbadane unikalne wartości w zbiorze train i w zbiorze valid

default_cus3	default_cus6	default_cus9	default_cus12
I	I	I	I
I	I	I	0
I	I	I	1
I	I	0	0
I	I	1	0
I	I	1	1
I	0	0	0
I	1	0	0
I	1	1	0
I	1	1	1
0	I	I	I
0	I	I	0
0	I	I	1
0	I	0	0
0	I	1	0
0	I	1	1
0	0	I	I
0	0	I	0
0	0	I	1
0	0	0	I
0	0	0	0
0	0	0	1
0	0	1	0
0	0	1	1
0	1	0	0
0	1	1	0
0	1	1	1
1	I	I	I
1	I	I	0
1	I	I	1
1	I	0	0
1	I	1	0

W zbiorze zostały zauważone występowanie I, które zostało oznakowane jako braków danych za pomocą funkcji strip,

```
/*znakowanie "I" jako braków danych*/
data tr3;
set tr2;
if strip(default_cus3) = "I" then default_cus3=.;
if strip(default_cus6) = "I" then default_cus6=.;
if strip(default_cus9) = "I" then default_cus9=.;
if strip(default_cus12) = "I" then default_cus12=.;
run;
```

Za pomocą funkcji select count(distinct period) zostało zbadane ilość unikalnych czasów, gdyż wiele razy występowały te same czasy w obserwacjach.

Mamy 174 taki samych wartości godzin kiedy zostały wygenerowane statystki w obydwóch zbiorach.

Sprawdzamy czy zmienne odstające w zbiorze train są tak samo rozłożone jak w zbiorze valid

Zostały wyrzucone zmienne agr, ags i default za pomocą kodu

```
data abc;
 file remove;
 if 0 then set train;
 length varname $32;
do until (varname='varname');
  call vnext(varname);
  if index(varname,'Ciev') then put varname;  /*zapis nazw zm
  if index(varname,'Cncr') then put varname;
  if index(varname,'arrears') then put varname;
end;
run;
```

Zostały te zmienne wypisane po powyższym kodzie.

Rozrzut zmienności na nich byłby bardzo duży. Jeżeli zmienna ma tylko 0 i 1, jest odbierane jako duży przedział.

Zostały zaprezentowane te wartości dla tych zmiennych na poniższej tabeli. Widać, że są w nich jedynie wartości brakujące. Nie ma żadnych wartości odstających.

Wspolczynnik zmiennosci w zbiorze valid

Procedura FREQ

act3_n_cind_arrears	Liczebność	Procent
0	38971	73.75
1	13870	26.25

act6_n_cind_arrears	Liczebność	Procent
0	43207	81.77
1	9634	18.23

act9_n_cind_arrears	Liczebność	Procent
0	45650	86.39
1	7191	13.61

act12_n_cind_arrears	Liczebność	Procent
0	47404	89.71
1	5437	10.29

act15_n_cind_arrears	Liczebność	Procent
0	48699	92.16
1	4142	7.84

act18_n_cind_arrears	Liczebność	Procent
0	49683	94.02
1	3158	5.98

act21_n_cind_arrears	Liczebność	Procent
0	50323	95.23
1	2518	4.77

Została zaprezentowana część tabeli ze względu na jej rozmiar. Nie ma wartości odstających w tych zmiennych. Są tylko missingi 0,1. Dlatego były one usunięte ze zbioru. Zostało wtedy 166 zmiennych.

Następnie analizowany jest okrojony zbiór train2.

oznaczenie zmiennych o bardzo duzej zmiennosci

Where | Query Builder | Tasks

CV2

	variable	CV	a
1	nobs	52841	1
2	ags21_Min_CMin_Due	6374.7699671	.
3	ags36_Min_CMax_Due	6142.8239059	.
4	ags33_Pctl5_CMin_Due	5886.7946986	.
5	ags36_Pctl5_CMin_Due	5567.246583	.
6	ags30_Pctl5_CMin_Due	5397.4836921	.
7	ags33_Min_CMax_Due	5272.7202437	.
8	ags30_Min_CMax_Due	5015.2624022	.
9	ags24_Pctl5_CMin_Due	4930.2490357	.
10	ags27_Pctl5_CMin_Due	4913.4672569	.
11	ags27_Min_CMax_Due	4507.088414	.
12	ags24_Min_CMax_Due	3829.9259979	.
13	ags21_Pctl5_CMin_Due	3694.4559065	.
14	ags36_Pctl5_CMax_Due	3446.6299559	.
15	agr24_Pctl5_CMin_Due	3238.440707	.
16	ags33_Pctl5_CMax_Due	3173.6552044	.
17	ags18_Pctl5_CMin_Due	3016.734902	.
18	ags18_Min_CMin_Due	3016.734902	.
19	ags30_Pctl5_CMax_Due	2885.6896789	.
20	agr30_Pctl5_CMin_Due	2818.6870797	.
21	ags21_Min_CMax_Due	2726.2675996	.
22	ags27_Pctl5_CMax_Due	2660.806452	.
23	agr21_Min_CMin_Due	2602.3655864	.
24	agr27_Pctl5_CMin_Due	2519.2584558	.
25	agr6_Kurtosis_CMin_Days	2517.7367443	.
26	ags6_Kurtosis_CMin_Days	2487.7835553	.

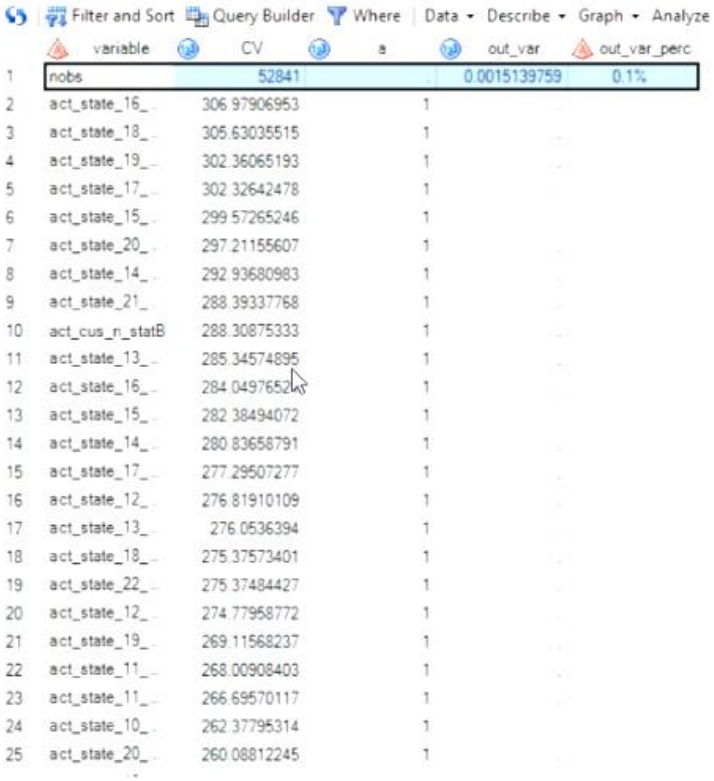
Została następnie obliczona liczba zmiennych o współczynniku zmienności większym niż 100

Współczynnik zmienności w zbiorze train

80

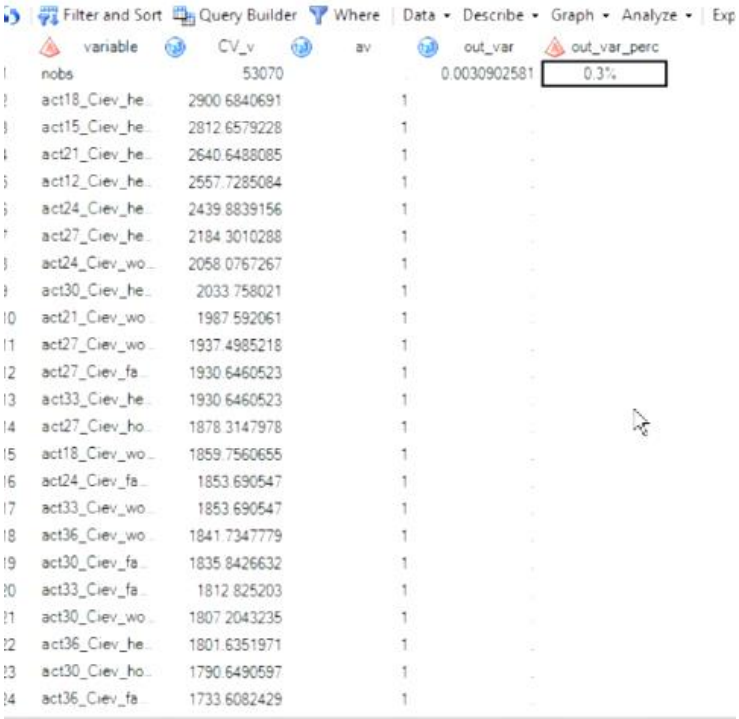
Page Break

Następnie zostało obliczone jaką część czesc zbioru stanowia zmiennej o bardzo duzym zroznnicowaniu



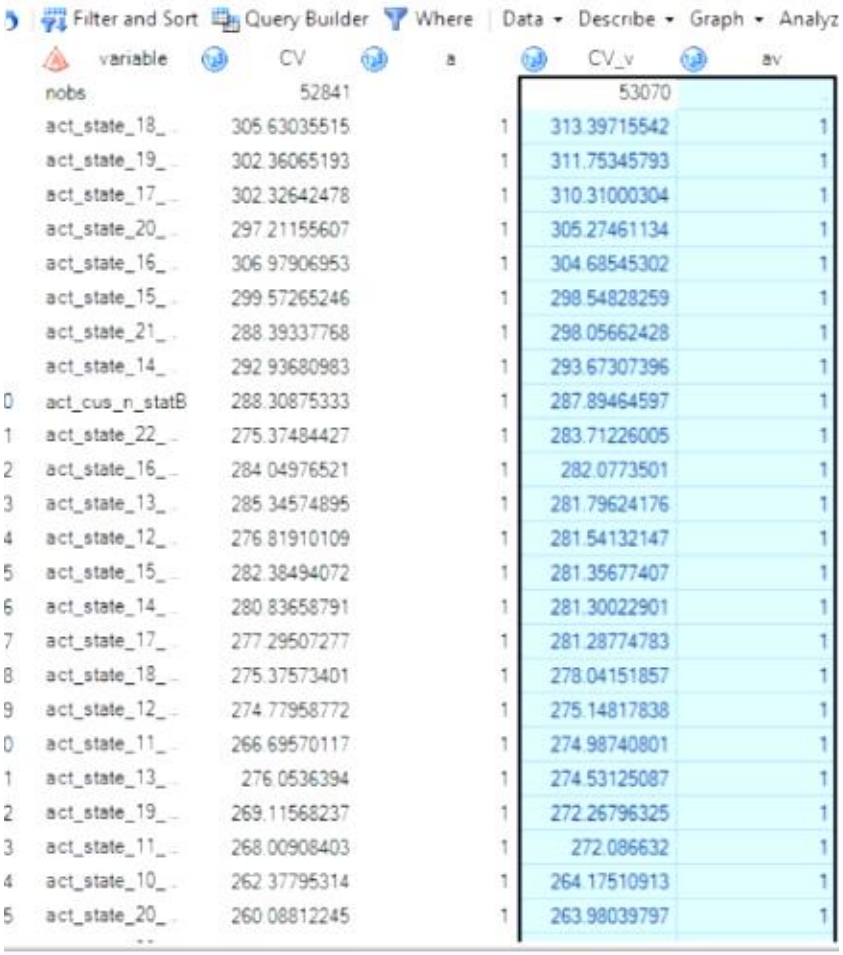
	variable	CV	a	out_var	out_var_perc
1	nobs	52841		0.0015139759	0.1%
2	act_state_16_..	306.97906953	1		
3	act_state_18_..	305.63035515	1		
4	act_state_19_..	302.36065193	1		
5	act_state_17_..	302.32642478	1		
6	act_state_15_..	299.57265246	1		
7	act_state_20_..	297.21155607	1		
8	act_state_14_..	292.93680983	1		
9	act_state_21_..	288.39337768	1		
10	act_cus_n_statB	288.30875333	1		
11	act_state_13_..	285.34574895	1		
12	act_state_16_..	284.0497652	1		
13	act_state_15_..	282.38494072	1		
14	act_state_14_..	280.83658791	1		
15	act_state_17_..	277.29507277	1		
16	act_state_12_..	276.81910109	1		
17	act_state_13_..	276.0536394	1		
18	act_state_18_..	275.37573401	1		
19	act_state_22_..	275.37484427	1		
20	act_state_12_..	274.77958772	1		
21	act_state_19_..	269.11568237	1		
22	act_state_11_..	268.00908403	1		
23	act_state_11_..	266.69570117	1		
24	act_state_10_..	262.37795314	1		
25	act_state_20_..	260.08812245	1		

Zostały wykonane podobne działania dla zbioru valid



	variable	CV_v	av	out_var	out_var_perc
1	nobs	53070		0.0030902581	0.3%
2	act18_Ciev_he..	2900.6840691	1		
3	act15_Ciev_he..	2812.6579228	1		
4	act21_Ciev_he..	2640.6488085	1		
5	act12_Ciev_he..	2557.7285084	1		
6	act24_Ciev_he..	2439.8839156	1		
7	act27_Ciev_he..	2184.3010288	1		
8	act24_Ciev_wo..	2058.0767267	1		
9	act30_Ciev_he..	2033.758021	1		
10	act21_Ciev_wo..	1987.592061	1		
11	act27_Ciev_wo..	1937.4985218	1		
12	act27_Ciev_fa..	1930.6460523	1		
13	act33_Ciev_he..	1930.6460523	1		
14	act27_Ciev_ho..	1878.3147978	1		
15	act18_Ciev_wo..	1859.7560655	1		
16	act24_Ciev_fa..	1853.690547	1		
17	act33_Ciev_wo..	1853.690547	1		
18	act36_Ciev_wo..	1841.7347779	1		
19	act30_Ciev_fa..	1835.8426632	1		
20	act33_Ciev_fa..	1812.825203	1		
21	act30_Ciev_wo..	1807.2043235	1		
22	act36_Ciev_he..	1801.6351971	1		
23	act30_Ciev_ho..	1790.6490597	1		
24	act36_Ciev_fa..	1733.6082429	1		

Porównanie train i valid po złączeniu dówch zbiorów. Zmienne av została oznaczona dla zbioru valid



	variable	CV	a	CV_v	av
	nobs	52841		53070	
	act_state_18_	305.63035515	1	313.39715542	1
	act_state_19_	302.36065193	1	311.75345793	1
	act_state_17_	302.32642478	1	310.31000304	1
	act_state_20_	297.21155607	1	305.27461134	1
	act_state_16_	306.97906953	1	304.68545302	1
	act_state_15_	299.57265246	1	298.54828259	1
	act_state_21_	288.39337768	1	298.05662428	1
	act_state_14_	292.93680983	1	293.67307396	1
0	act_cus_n_statB	288.30875333	1	287.89464597	1
1	act_state_22_	275.37484427	1	283.71226005	1
2	act_state_16_	284.04976521	1	282.0773501	1
3	act_state_13_	285.34574895	1	281.79624176	1
4	act_state_12_	276.81910109	1	281.54132147	1
5	act_state_15_	282.38494072	1	281.35677407	1
6	act_state_14_	280.83658791	1	281.30022901	1
7	act_state_17_	277.29507277	1	281.28774783	1
8	act_state_18_	275.37573401	1	278.04151857	1
9	act_state_12_	274.77958772	1	275.14817838	1
0	act_state_11_	266.69570117	1	274.98740801	1
1	act_state_13_	276.0536394	1	274.53125087	1
2	act_state_19_	269.11568237	1	272.26796325	1
3	act_state_11_	268.00908403	1	272.086632	1
4	act_state_10_	262.37795314	1	264.17510913	1
5	act_state_20_	260.08812245	1	263.98039797	1

1 są przy zmiennych, który współczynnik zmienności wynosi 100, co wkazuje na zbioru o dużym zróżnicowaniu.

Sprawdzam czy a równa się av, czyli czy zróżnicowanie na zbiorze valid i train jest takie samo.

Zbiory train i valid są podobnie zróżnicowane.

Współczynnik zmienności w zbiorze valid

NAME OF FORMER VARIABLE	CV	CV_v
act_state_16_CMin_Due	306.9791	304.6855
act_state_18_CMin_Due	305.6304	313.3972
act_state_19_CMin_Due	302.3607	311.7535
act_state_17_CMin_Due	302.3264	310.31
act_state_15_CMin_Due	299.5727	298.5483
act_state_20_CMin_Due	297.2116	305.2746
act_state_14_CMin_Due	292.9368	293.6731
act_state_21_CMin_Due	288.3934	298.0566
act_cus_n_statB	288.3088	287.8946
act_state_13_CMin_Due	285.3457	281.7962
act_state_16_CMax_Due	284.0498	282.0774
act_state_15_CMax_Due	282.3849	281.3568
act_state_14_CMax_Due	280.8366	281.3002
act_state_17_CMax_Due	277.2951	281.2877
act_state_12_CMin_Due	276.8191	281.5413
act_state_13_CMax_Due	276.0536	274.5313
act_state_18_CMax_Due	275.3757	278.0415
act_state_22_CMin_Due	275.3748	283.7123
act_state_12_CMax_Due	274.7796	275.1482
act_state_19_CMax_Due	269.1157	272.268
act_state_11_CMax_Due	268.0091	272.0866
act_state_11_CMin_Due	266.6957	274.9874
act_state_10_CMax_Due	262.378	264.1751
act_state_20_CMax_Due	260.0881	263.9804
act_state_10_CMin_Due	259.3299	262.0878
act_state_9_CMax_Due	258.4902	258.4257
act_state_23_CMin_Due	258.2736	262.9644

Zmienne o bardzo dużym zróżnicowaniu posegregowane od najwyższego współczynnika zmienności, wsród tych które zróżnicowanie jest duże

Zmienne o największym zróżnicowaniu na obu zbiorach, gdzie współczynnik zmienności jest powyżej 100. Jest ich 80. Wśród nich jest zmienna spendigs.

	app_spendings	act_cus_seniority	act_cus_n_loans_hist	act_cus_n_statC	act_cus_n_statB	act_cus_dueutl	act_state_1_CMax_Due	act_state_2_CMax_Due	act_state_3_CMax_Due
1	740	50	22	1	20	0	0	0	0
2	160	50	8	1	6	0.0416666667	1	0	0
3	160	50	10	1	5	0.0520833333	3	2	1
4	1180	50	6	1	4	0	0	0	0
5	200	50	7	4	1	0.0625	3	2	1
6	520	49	3	2	0	0	0	0	0
7	440	49	6	2	3	0.0416666667	1	1	0
8	260	49	5	0	4	0.125	3	2	1
9	100	48	8	3	2	0	0	1	1
10	360	48	8	2	3	0.0138888889	1	1	1
11	460	48	5	0	3	0.0833333333	3	2	2
12	1420	48	6	0	5	0.0416666667	1	0	7
13	140	47	4	1	0	0	0	0	0
14	400	47	6	3	1	0.0555555556	2	1	1
15	320	47	4	2	1	0	0	0	0
16	540	47	4	1	1	0	0	0	0
17	280	47	4	2	1	0	0	0	1
18	80	47	4	2	1	0	0	0	0
19	180	47	4	0	2	0.0833333333	3	2	1
20	180	46	5	1	3	0	0	0	0
21	900	46	4	1	2	0.0833333333	2	3	2
22	380	45	4	1	2	0	0	0	0
23	180	45	4	0	3	0.0416666667	1	1	0
24	160	45	2	0	1	0.0416666667	1	1	1

Obserwacje odstające

Filter and Sort | Query Builder | Where | Data | Describe | Graph | Analyze | Export | Send To

MAX_	MEAN_	STD_	lower	upper	flag	up_outliers	low_outliers
7	0.2723433825	0.8360371815	-2.235768162	2.780454927		4.219545073	2.235768162
7	0.2992605196	0.9146309889	-2.444632447	3.0431534863		3.9568465137	2.4446324471
7	0.3220661399	0.9738012802	-2.599337701	3.2434699805		3.7565300195	2.5993377007
7	0.2840089824	0.8586342025	-2.291893625	2.8599115899		4.1400884101	2.2918936251
7	0.258775351	0.775220183	-2.066885198	2.5844359		4.4155641	2.066885198
7	0.3526207181	1.0480295233	-2.791467852	3.496709288		3.503290712	2.7914678518
7	0.2520955948	0.7384807931	-1.963346785	2.4675379741		4.5324620259	1.9633467845
7	0.3900862069	1.1249827879	-2.984862157	3.7650345706		3.2349654294	2.9848621568
46	0.4810847637	1.3870094847	-3.67994369	4.6421132178		41.357886782	3.6799436904
7	0.2516995659	0.7182140114	-1.902942468	2.4063416001		4.5936583999	1.9029424683
7	0.3760423348	1.0681473692	-2.828399773	3.5804844424		3.4195155576	2.8283997728
7	0.3566692668	1.0071802975	-2.664871626	3.3782101593		3.6217898407	2.6648716257
7	0.3436329588	0.9650470764	-2.55150827	3.238774188		3.761225812	2.5515082704
7	0.3918567545	1.0865994726	-2.867941663	3.6516551723		3.3483448277	2.8679416633
7	0.2389223143	0.6613826028	-1.745225494	2.2230701227		4.7769298773	1.7452254941
7	0.3378655091	0.932690034	-2.460204593	3.1359356111		3.8640643889	2.4602045929
7	0.4182972319	1.1518890726	-3.037369986	3.8739644497		3.1260355503	3.0373699859
7	0.4399783901	1.2115898064	-3.194791029	4.0747478093		2.9252521907	3.1947910291
7	0.3183108185	0.8746531547	-2.305648646	2.9422702826		4.0577297174	2.3056486456
7	0.4622475856	1.2439807442	-3.269694647	4.1941898182		2.8058101818	3.269694647
7	0.3104269553	0.8319724394	-2.185490363	2.8063442735		4.1936557265	2.1854903629
7	0.2370845154	0.6322942108	-1.659798117	2.1339671478		4.8660328522	1.659798117
7	0.3117410221	0.817939713	-2.142078117	2.7655601611		4.2344398389	2.1420781169
7	0.5101114321	1.3267392462	-3.470106307	4.4903291707		2.5096708293	3.4701063065

Wszystko poniżej wartości lower jest odstającą i wszystko powyżej upper

NAME OF FORMER VARIABLE	perc_all
act_cus_n_statB	97.91
act_cus_n_loans_hist	96.13
act_state_11_CMin_Due	93.23
act_state_12_CMin_Due	93.17
act_state_10_CMin_Due	93.1
act_state_9_CMin_Due	93.02
act_state_8_CMin_Due	92.9
act_state_13_CMin_Due	92.81
act_state_14_CMin_Due	92.8
act_state_7_CMin_Due	92.77
act_cus_n_statC	92.66
act_state_15_CMin_Due	92.61
act_state_6_CMin_Due	92.54
act_state_16_CMin_Due	92.22
act_state_5_CMin_Due	92.07
act_state_17_CMin_Due	91.89
act_state_4_CMin_Due	91.51
act_state_18_CMin_Due	91.45
act_state_11_CMax_Due	91.13
act_state_10_CMax_Due	91.09
act_state_9_CMax_Due	91.06
act_state_8_CMax_Due	90.97
act_state_7_CMax_Due	90.91
act_state_12_CMax_Due	90.91
act_state_19_CMin_Due	90.8
act_state_3_CMin_Due	90.74
act_state_6_CMax_Due	90.7

Zmienne odstające są ustawione od najwyższej. Pokazany jest udział procentowy dla każdej zmiennej.

Niżej zostały zaprezentowane zmienne odstające na dwóch zbiorach. Nie ma tu zmiennych odstających tylko w zbiorze train albo tylko w valid. Widać, że procent dla tej samej zmiennej dla obu zbiorów jest podobny.

	variable	perc_all	perc_all_v
1	act_cus_n_statB	97.91	97.92
2	act_cus_n_loa	96.13	96.23
3	act_state_11_	93.23	93.12
4	act_state_12_	93.17	93
5	act_state_10_	93.1	93.21
6	act_state_9_C	93.02	93.23
7	act_state_8_C	92.9	92.99
8	act_state_13_	92.81	92.82
9	act_state_14_	92.8	92.75
10	act_state_7_C	92.77	92.74
11	act_cus_n_statC	92.66	93.16
12	act_state_15_	92.61	92.68
13	act_state_6_C	92.54	92.55
14	act_state_16_	92.22	92.47
15	act_state_5_C	92.07	92.15
16	act_state_17_	91.89	92.31
17	act_state_4_C	91.51	91.56
18	act_state_18_	91.45	91.8
19	act_state_11_	91.13	91.01
20	act_state_10_	91.09	91.17
21	act_state_9_C	91.06	91.26
22	act_state_8_C	90.97	91.07
23	act_state_7_C	90.91	90.85
24	act_state_12_	90.91	90.87
25	act_state_19_	90.8	91.18

Najwięcej zmiennych odstających ma zmienna zaznaczona na tabeli.

Została ta zmienne zaprezentowana jeszcze poniżej

Quantiles (Definition 5)	
Level	**Quantile**
100% Max	46
99%	7
95%	3
90%	1
75% Q3	0
50% Median	0
25% Q1	0
10%	0
5%	0
1%	0
0% Min	0

4. Analiza nietypowych danych w postaci zwizualizowanych graficznych raportów w celu szybkiego identyfikowania zmiennych z większą i mniejszą nietypowością i nieregularnością danych oraz z ich różną stabilnością.

Przed przystąpieniem do wizualizacji dane zostały uporządkowane poprzez usunięcie brakujących wartości funkcji celu oraz danych zagregowanych, które również wykazywały braki.

Następnie użyta została procedura means w celu obliczenia dla poszczególnych zmiennych wartości takich jak:

- liczba obserwacji;

- mediana;

- rozstęp międzykwartylowy;

- centyl 25 oraz 75

- minimum oraz maksimum

Powyższe punkty zostały wykonane zarówno dla zbioru abt_sam_beh_train jak i abt_sam_beh_valid.

Dla obu zbiorów danych wygenerowane zostały również tabele zawierające po 20 zmiennych o najwyższym rozstępie międzykwartylowym.

Poniżej znajdują się wykresy, które pozwalają przeanalizować wartości odstające dla wybranych zmiennych. Można zaobserwować, że w obu zbiorach wartości rozkładają się podobnie.

Zbiór train

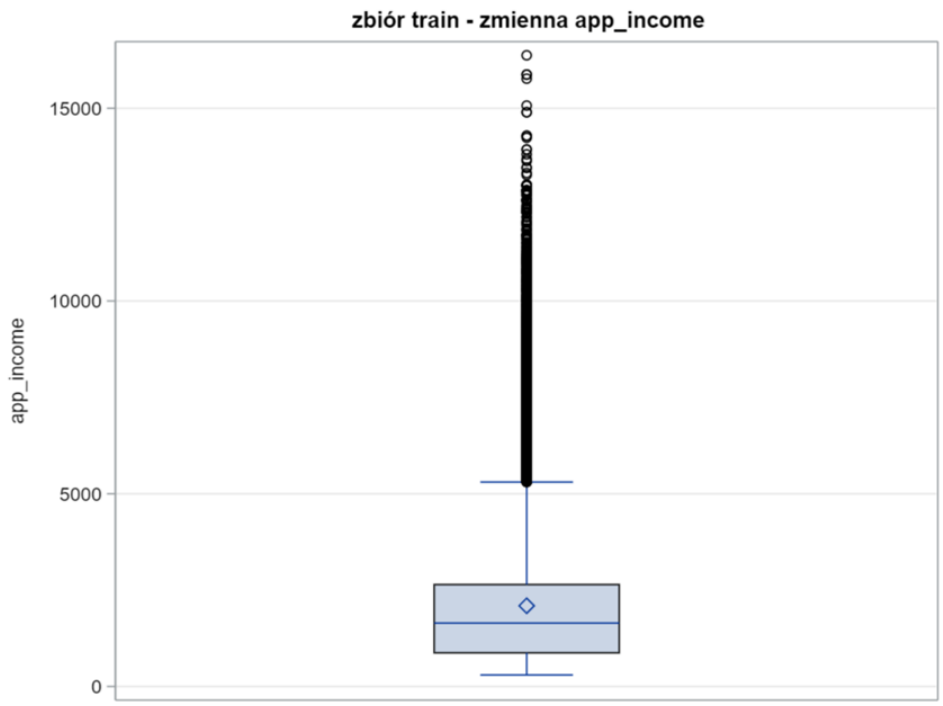
Obs.	Variable	N	Median	QRange	P25	P75	Min	Max	LowerBound	UpperBound
1	app_income	46388	1644.500000	1772.000000	873.000000	2645.000000	300.000000	16373	-1785.0	5303.0
2	app_spendings	46388	380.000000	500.000000	200.000000	700.000000	0	7720.000000	-550.0	1450.0
3	ags36_Sum_CMax_Days	46388	154.000000	151.000000	80.000000	231.000000	15.000000	528.000000	-146.5	457.5
4	ags36_Sum_CMin_Days	46388	154.000000	150.000000	80.000000	230.000000	15.000000	480.000000	-145.0	455.0
5	ags33_Sum_CMax_Days	46388	152.000000	149.000000	79.000000	228.000000	15.000000	482.000000	-144.5	451.5
6	ags30_Sum_CMax_Days	46388	151.000000	148.000000	78.000000	226.000000	15.000000	438.000000	-144.0	448.0
7	ags33_Sum_CMin_Days	46388	152.000000	148.000000	79.000000	227.000000	15.000000	441.000000	-143.0	449.0
8	ags30_Sum_CMin_Days	46388	150.000000	147.000000	78.000000	225.000000	15.000000	401.000000	-142.5	445.5
9	ags27_Sum_CMax_Days	46388	149.000000	145.000000	78.000000	223.000000	15.000000	397.000000	-139.5	440.5
10	ags27_Sum_CMin_Days	46388	148.000000	145.000000	77.000000	222.000000	15.000000	369.000000	-140.5	439.5
11	ags24_Sum_CMax_Days	46388	147.000000	143.000000	77.000000	220.000000	15.000000	354.000000	-137.5	434.5
12	ags24_Sum_CMin_Days	46388	147.000000	142.000000	77.000000	219.000000	15.000000	328.000000	-136.0	432.0
13	ags21_Sum_CMax_Days	46388	145.000000	138.000000	76.000000	214.000000	15.000000	311.000000	-131.0	421.0
14	ags21_Sum_CMin_Days	46388	145.000000	137.000000	76.000000	213.000000	15.000000	289.000000	-129.5	418.5
15	ags18_Sum_CMax_Days	46388	143.000000	122.000000	75.000000	197.000000	15.000000	268.000000	-108.0	380.0
16	ags18_Sum_CMin_Days	46388	143.000000	120.000000	75.000000	195.000000	15.000000	252.000000	-105.0	375.0
17	ags15_Sum_CMax_Days	46388	137.000000	97.000000	74.000000	171.000000	15.000000	227.000000	-71.5	316.5
18	ags15_Sum_CMin_Days	46388	137.000000	95.000000	74.000000	169.000000	15.000000	214.000000	-68.5	311.5
19	act_cus_seniority	46388	18.000000	73.000000	9.000000	82.000000	1.000000	223.000000	-100.5	191.5
20	ags12_Sum_CMax_Days	46388	120.000000	67.000000	73.000000	140.000000	15.000000	184.000000	-27.5	240.5

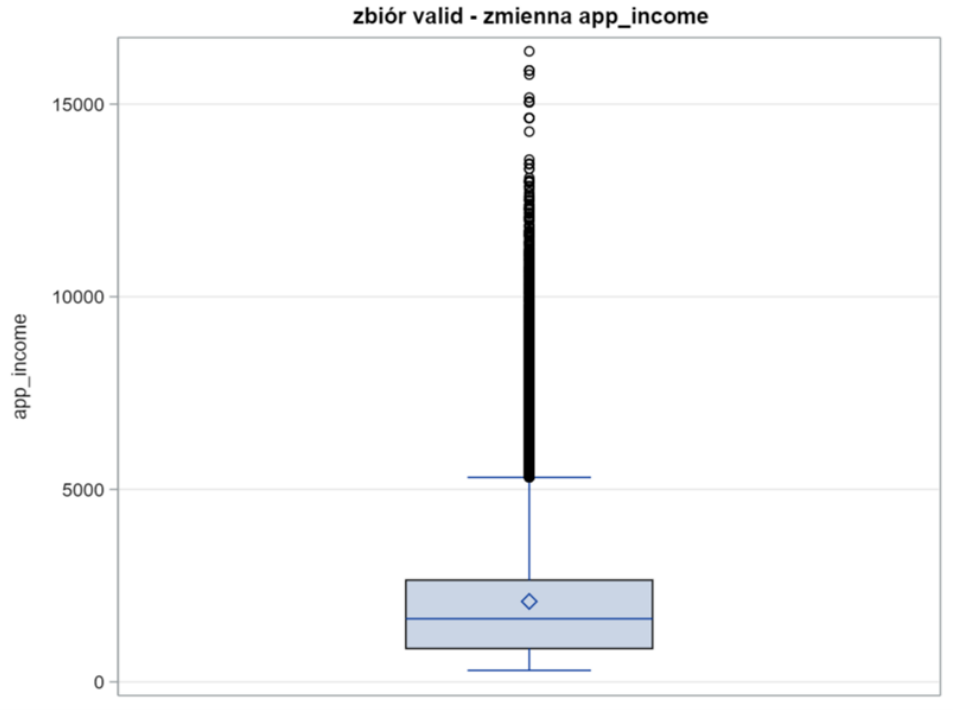
Tabela zawierająca 20 zmiennych ze zbioru abt_sam_beh_train o najwyższym rozstępie międzykwartylowym

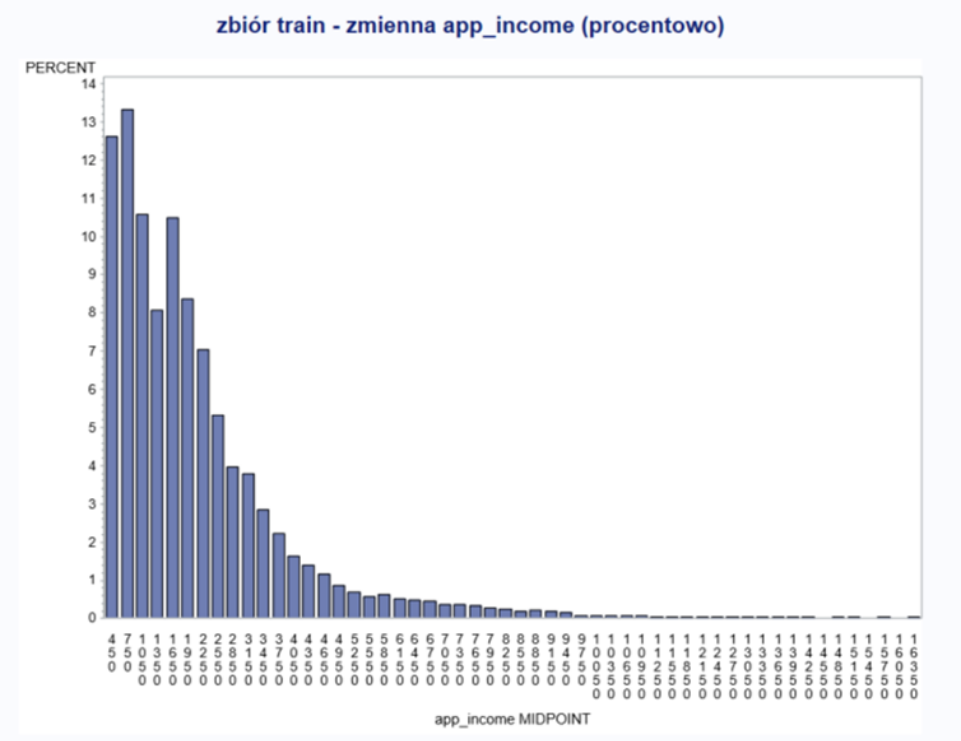
Zbiór valid

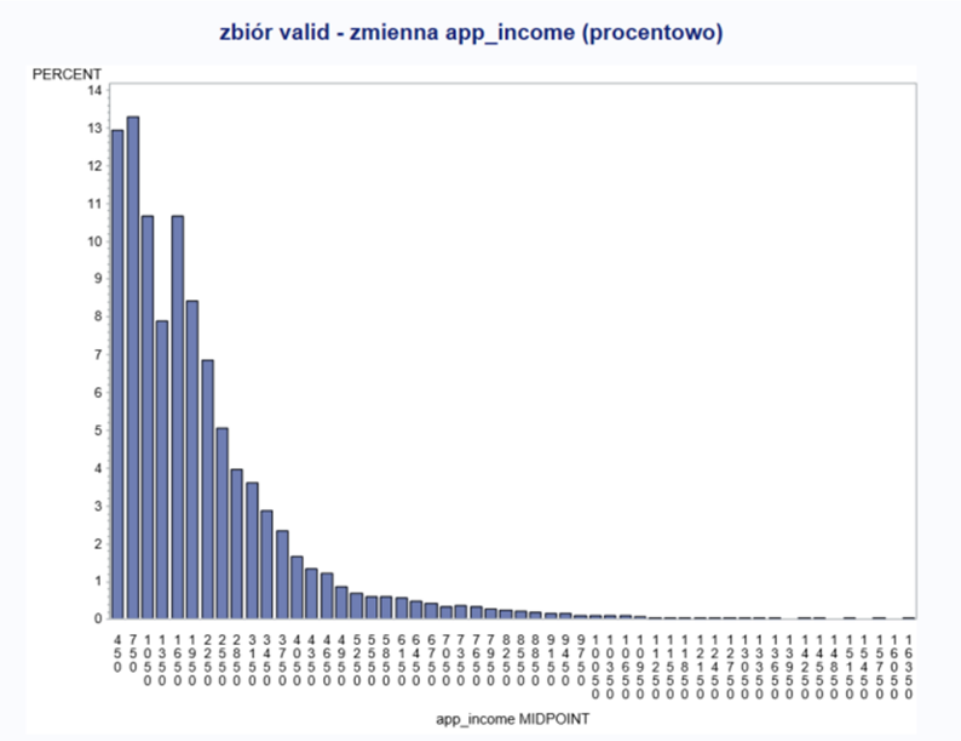
Obs.	Variable	N	Median	QRange	P25	P75	Min	Max	LowerBound	UpperBound
1	app_income	46795	1640.000000	1778.000000	867.000000	2645.000000	300.000000	16373	-1800.0	5312.0
2	app_spendings	46795	380.000000	500.000000	200.000000	700.000000	0	7720.000000	-550.0	1450.0
3	ags36_Sum_CMax_Days	46795	155.000000	152.000000	80.000000	232.000000	15.000000	524.000000	-148.0	460.0
4	ags36_Sum_CMin_Days	46795	155.000000	151.000000	80.000000	231.000000	15.000000	479.000000	-146.5	457.5
5	ags33_Sum_CMax_Days	46795	153.000000	150.000000	79.000000	229.000000	15.000000	482.000000	-146.0	454.0
6	ags33_Sum_CMin_Days	46795	153.000000	149.000000	79.000000	228.000000	15.000000	441.000000	-144.5	451.5
7	ags30_Sum_CMax_Days	46795	151.000000	148.000000	78.000000	226.000000	15.000000	437.000000	-144.0	448.0
8	ags30_Sum_CMin_Days	46795	151.000000	147.000000	78.000000	225.000000	15.000000	402.000000	-142.5	445.5
9	ags27_Sum_CMax_Days	46795	149.000000	146.000000	78.000000	224.000000	15.000000	394.000000	-141.0	443.0
10	ags27_Sum_CMin_Days	46795	149.000000	145.000000	78.000000	223.000000	15.000000	370.000000	-139.5	440.5
11	ags24_Sum_CMax_Days	46795	147.000000	144.000000	77.000000	221.000000	15.000000	348.000000	-139.0	437.0
12	ags24_Sum_CMin_Days	46795	147.000000	143.000000	77.000000	220.000000	15.000000	331.000000	-137.5	434.5
13	ags21_Sum_CMax_Days	46795	146.000000	139.000000	76.000000	215.000000	15.000000	309.000000	-132.5	423.5
14	ags21_Sum_CMin_Days	46795	145.000000	137.000000	76.000000	213.000000	15.000000	292.000000	-129.5	418.5
15	ags18_Sum_CMax_Days	46795	144.000000	122.000000	75.000000	197.000000	15.000000	269.000000	-108.0	380.0
16	ags18_Sum_CMin_Days	46795	143.000000	121.000000	75.000000	196.000000	15.000000	255.000000	-106.5	377.5
17	ags15_Sum_CMax_Days	46795	138.000000	97.000000	74.000000	171.000000	15.000000	224.000000	-71.5	316.5
18	ags15_Sum_CMin_Days	46795	137.000000	96.000000	74.000000	170.000000	15.000000	215.000000	-70.0	314.0
19	act_cus_seniority	46795	18.000000	74.000000	9.000000	83.000000	1.000000	223.000000	-102.0	194.0
20	ags12_Sum_CMax_Days	46795	120.000000	67.000000	73.000000	140.000000	15.000000	185.000000	-27.5	240.5

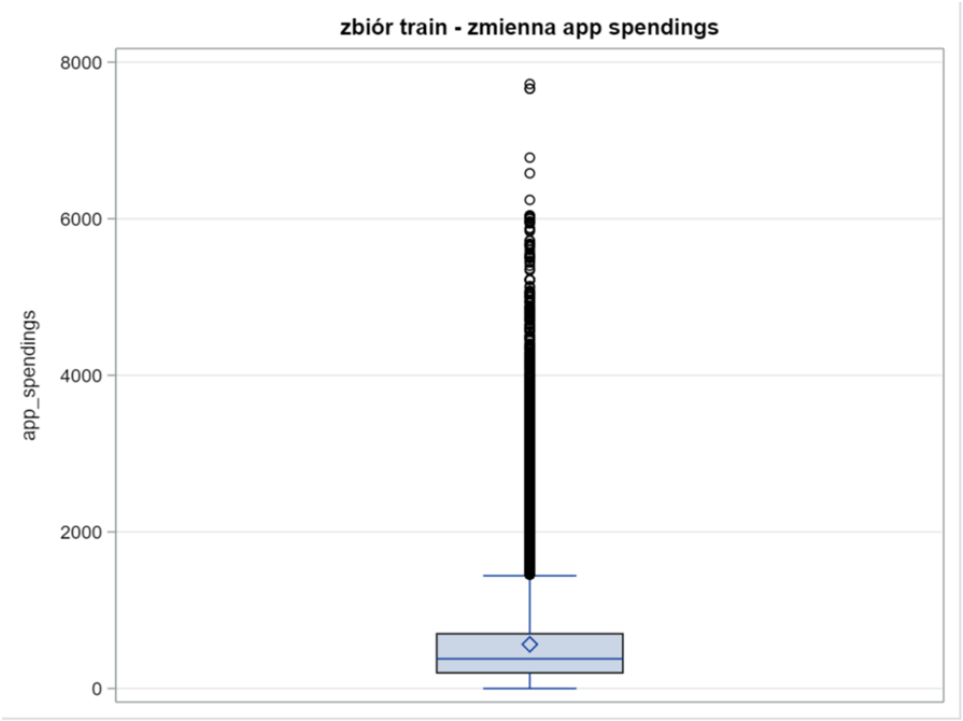
Tabela zawierająca 20 zmiennych ze zbioru abt_sam_beh_valid o najwyższym rozstępie międzykwartylowym

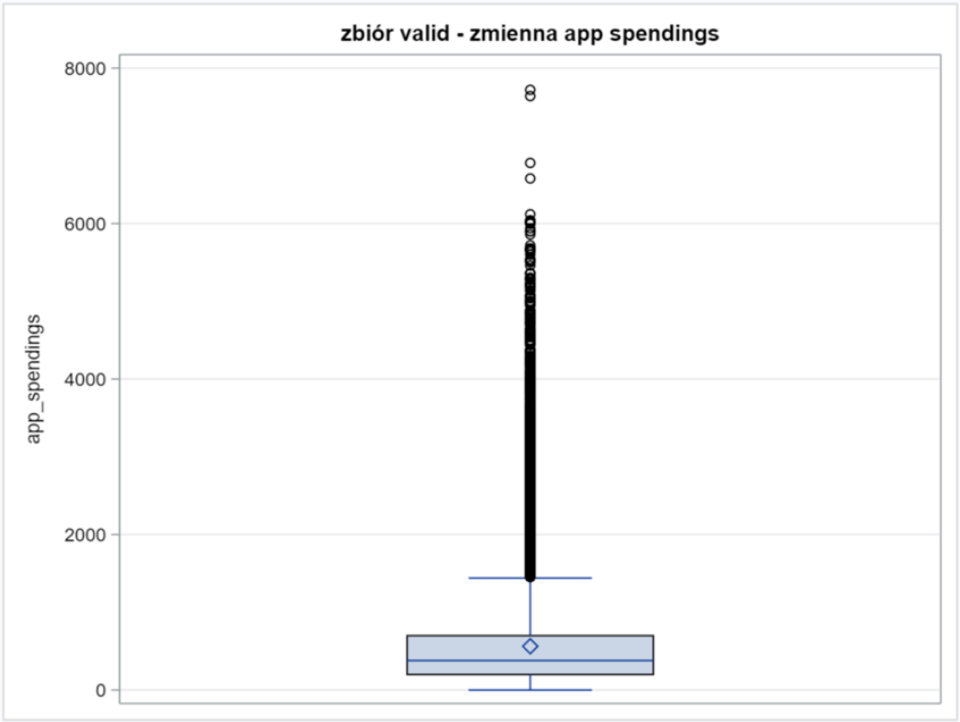

zbiór train - zmienna app_income
app_income
15000
10000
5000
0

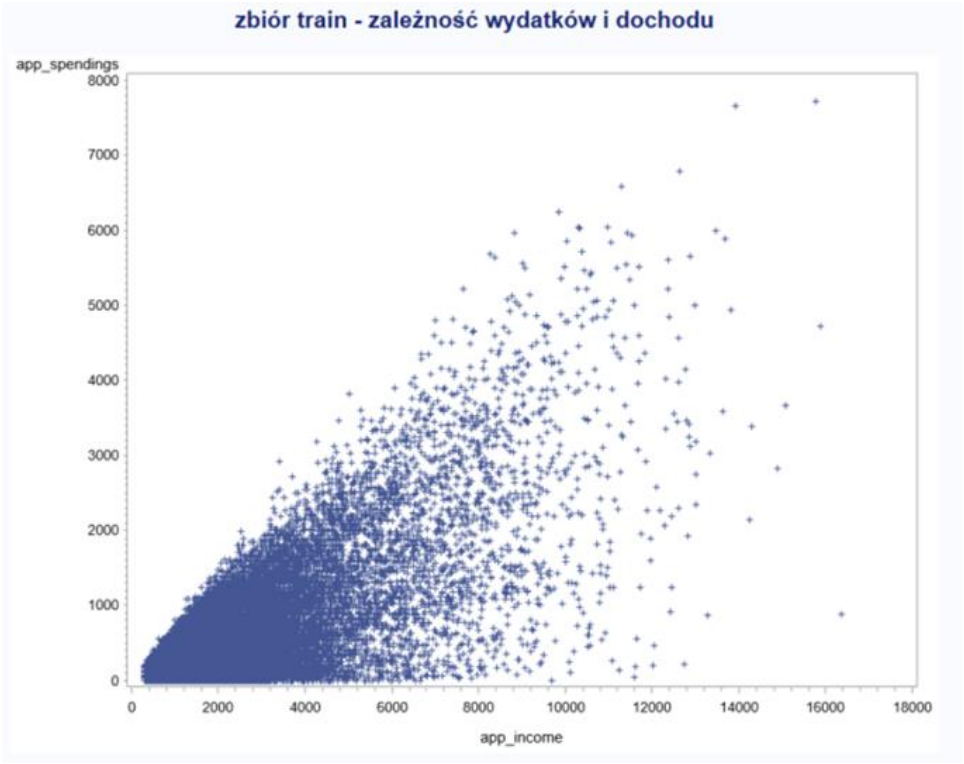

zbiór valid - zmienna app_income
app_income
15000
10000
5000
0

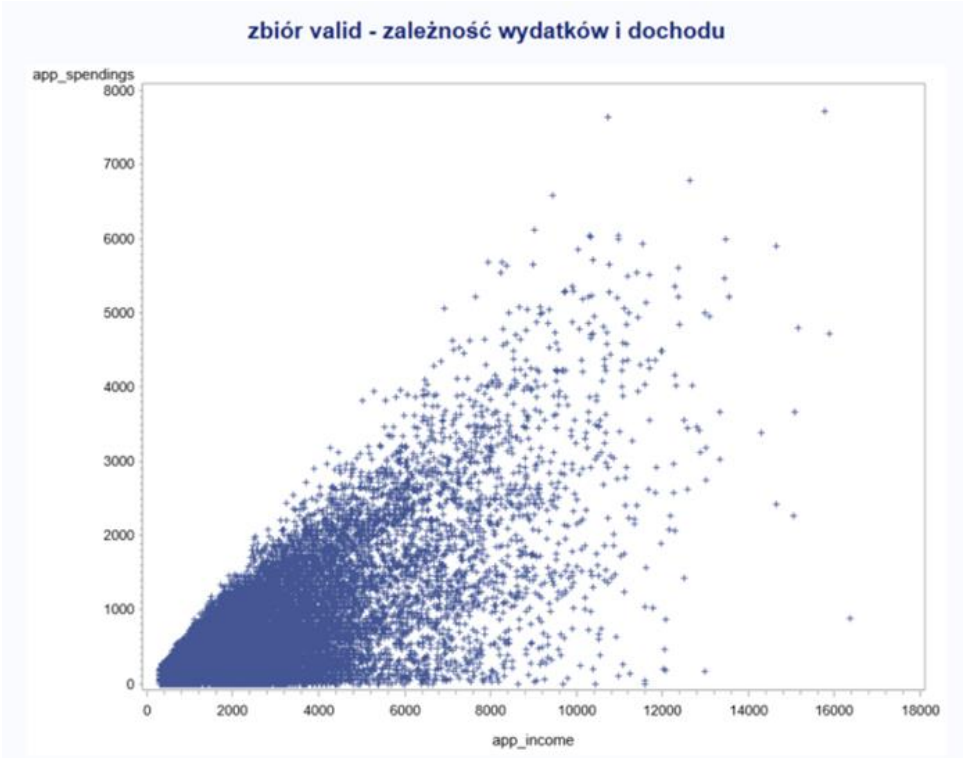

zbiór train - zmienna app_income (procentowo)
PERCENT
app_income MIDPOINT

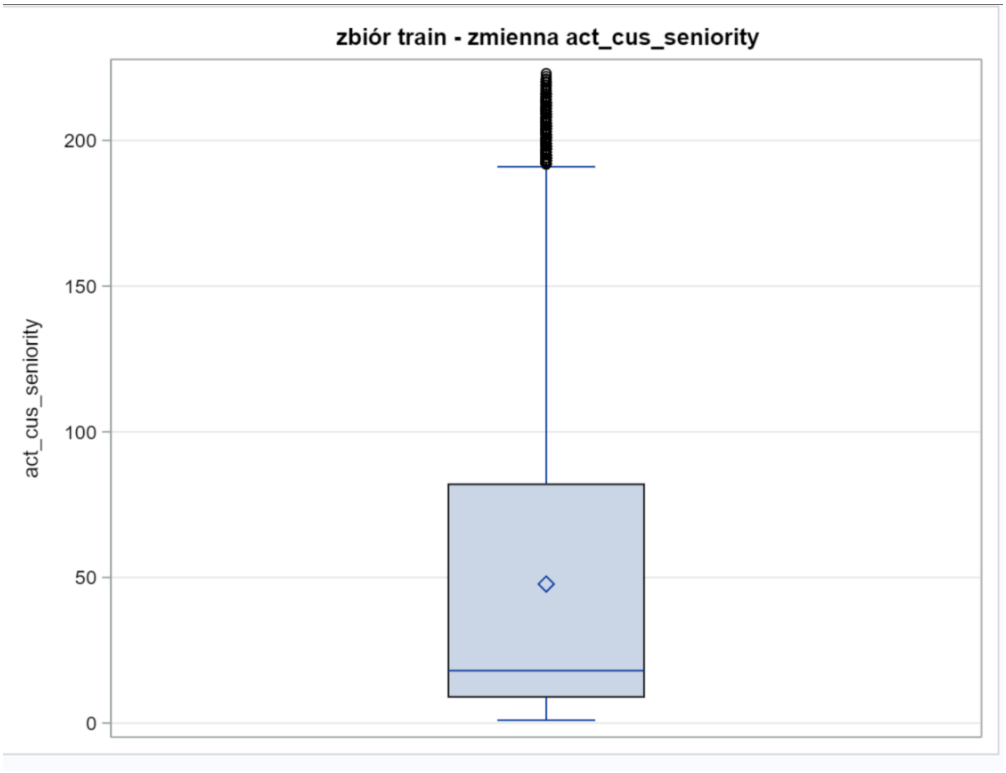

zbiór valid - zmienna app_income (procentowo)
PERCENT
app_income MIDPOINT

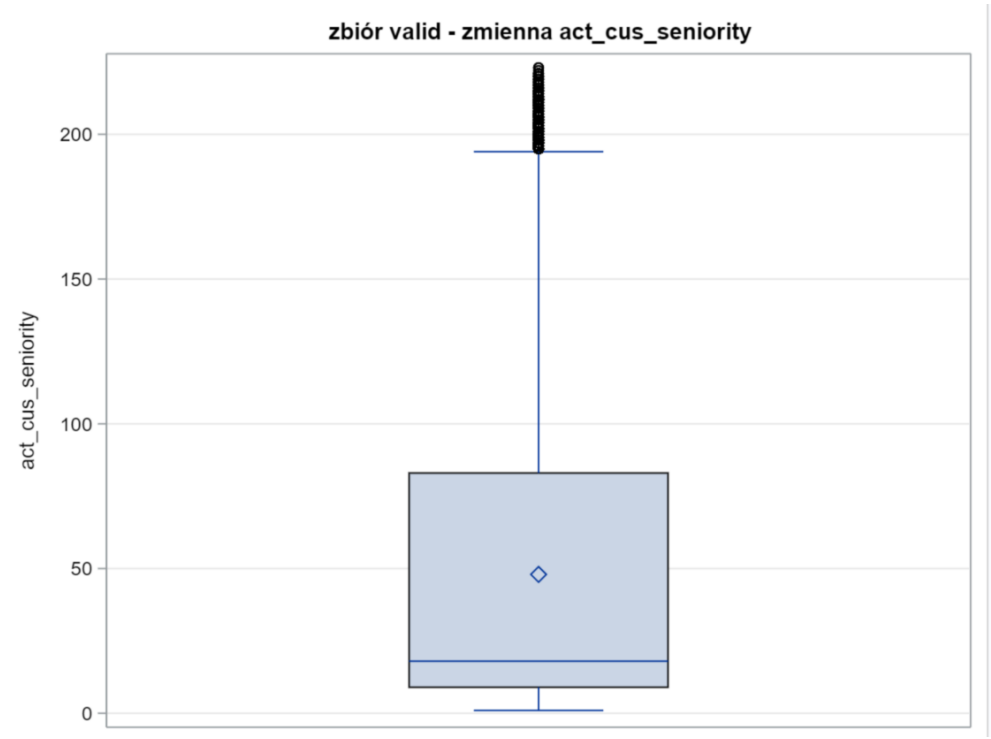

zbiór train - zmienna app spendings
app_spendings
8000
6000
4000
2000
0


zbiór valid - zmienna app spendings
app_spendings
8000
6000
4000
2000
0


zbiór train - zależność wydatków i dochodu
app_spendings
8000
7000
6000
5000
4000
3000
2000
1000
0
0
2000
4000
6000
8000
10000
12000
14000
16000
18000
app_income


zbiór valid - zależność wydatków i dochodu
app_spendings
8000
7000
6000
5000
4000
3000
2000
1000
0
0
2000
4000
6000
8000
10000
12000
14000
16000
18000
app_income


zbiór train - zmienna act_cus_seniority
act_cus_seniority
200
150
100
50
0


zbiór valid - zmienna act_cus_seniority
act_cus_seniority
200
150
100
50
0

5. Analiza zależności pomiędzy funkcją celu defaut_cus12 a charakterystykami klienta. Identyfikacja zmiennych wpływających na funkcję celu, ich współzależności, korelacji, zależności itp. W tym wypadku finalny raport może składać się i z raportów tabelarycznych, i graficznych wizualizujących dobroć predyktorów.

Funkcja celu *default_cus12* – Czy klient od punktu obserwacji w ciągu pierwszych 12 miesięcy wszedł w opóźnienia więcej niż 3 razy. Zmienna przyjmuje poniższe wartości:

- 1 – klient w ciągu ostatnich 12 miesięcy wszedł w opóźnienia ponad 3 razy;
- 0 – klient miał opóźnienie tylko raz w ciągu 12 miesięcy;
- .i – inne przypadki;
- . – brak historii z ostatnich 12 miesięcy.

Charakterystyki klienta:

1. zmienne zaczynające się od *app_* - cechy z aplikacji (głównie zmienne opisowe);
2. zmienne zaczynające się od *act_* - opisujące stan z danego punktu czasowego;
3. zmienne zaczynające się od agr_ - zmienne agregujące informacje z wielu miesięcy;
4. zmienne zaczynające się od *ags_* - zmienne agregujące informacje z wielu miesięcy.

Przygotowanie danych do analizy korelacji:

1. ze zbioru zostały usunięte:

- braki danych funkcji celu;
- inne przypadki „.i" funkcji celu - z uwagi na brak konkretnej informacji co kryje się pod pojęciem „inne przypadki";
- Zmienne objaśniające, których braki wynoszą ponad 60%;
- Inne funkcje celu: default_cus3, default_cus6, default_cus9;
- Wszystkie zmienne agr_.

2. Zmienne opisowe zostały zmienione na zmienne numeryczne:

- App_char_job_code;
- App_char_marital_stat;

- App_ char_city;
- App_char_home_status;
- App_char_cars.

Korelacja – definicja

Korelacja określa współzależność między zmiennymi, celem jej analizy jest stwierdzenie czy między badanymi zmiennymi zachodzą jakieś zależności oraz jaka jest ich siła. Siłę związku między zmiennymi wyznacza m.in. współczynnik korelacji Pearsona. Współczynnik ten należy interpretować następująco:

Znak współczynnika informuje o kierunku zależności:

- dodatni: zależność liniowa dodatnia;

- ujemny: zależność liniowa ujemna.

Wartość współczynnika korelacji informuje o sile zależności:

- mniejsza od 0,20: brak związku liniowego;

- 0,20 – 0,40: zależność linowa niska;

- 0,40 – 0,7: zależność liniowa umiarkowana;

- 0,7 – 0,9: zależność liniowa silna;

- powyżej 0,9: zależność liniowa bardzo silna.

Kształtowanie się korelacji w zbiorze valid

Po usunięciu ze zbioru valid zbędnych zmiennych objaśniających pozostało 1145 zmiennych. Jak widać na poniższym wykresie oraz w tabeli dla ok. 80% zmiennych korelacja z funkcją celu default_cus12 waha się od –0.20 do 0.20 co wskazuje na brak większego związku między nimi. Dla ok. 18,5% zmiennych korelacja jest na poziomie niskim, a jedynie dla 14 zmiennych jest powyżej

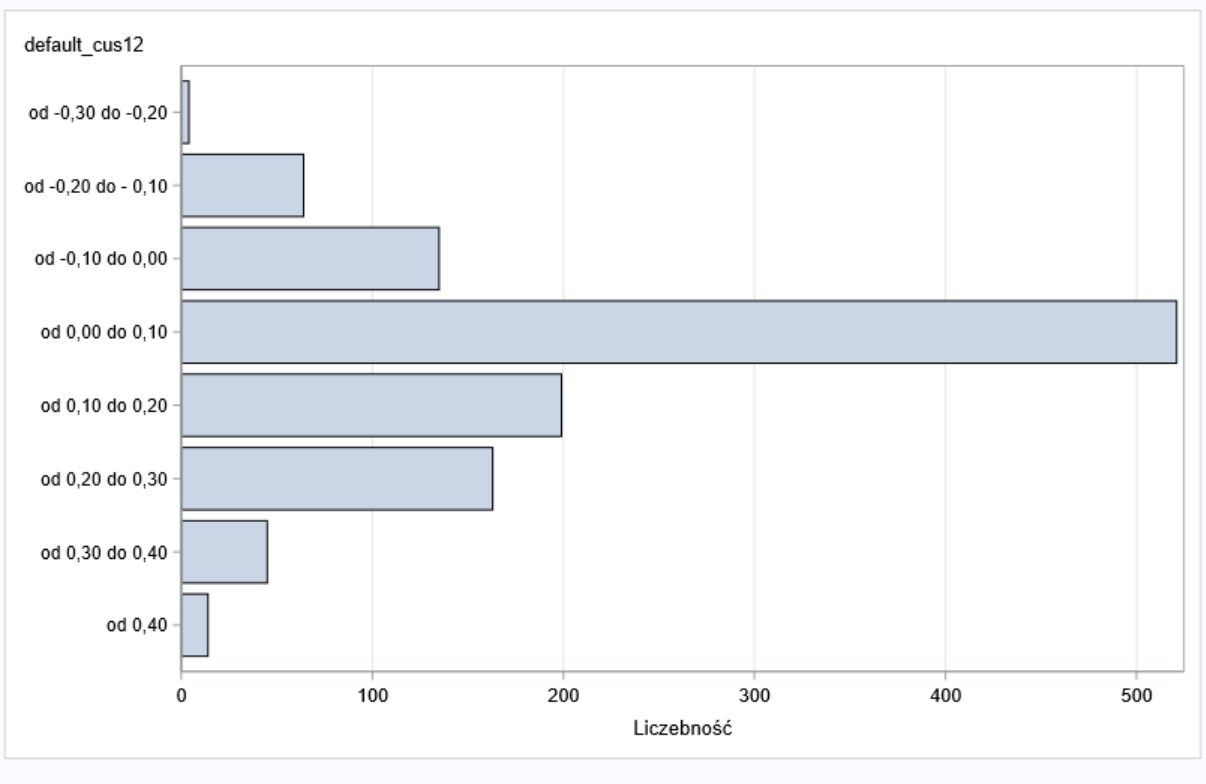
progu 0.40 - umiarkowana korelacja. Z uwagi na brak wyników wskazujących na silną korelację, między zmienną objaśniającą, a funkcją celu do dalszej analizy zostały wzięte zmienne, które posiadają umiarkowany współczynnik korelacji.

Tabela częstości współczynnika korelacji w zbiorze valid:

Procedura FREQ

default_cus12	Liczebność	Procent	Liczebność skumulowana	Procent skumulowany
od -0,30 do -0,20	4	0.35	4	0.35
od -0,20 do - 0,10	64	5.59	68	5.94
od -0,10 do 0,00	135	11.79	203	17.73
od 0,00 do 0,10	521	45.50	724	63.23
od 0,10 do 0,20	199	17.38	923	80.61
od 0,20 do 0,30	163	14.24	1086	94.85
od 0,30 do 0,40	45	3.93	1131	98.78
od 0,40	14	1.22	1145	100.00

Przedstawienie częstości współczynnika korelacji na histogramie:



Zmienne o najwyższym współczynniku korelacji w zbiorze valid:

name	default_cus12
act_cus_dueutl	0.4535571103
act_state_1_CMax_Due	0.4654605824
act_state_1_CMin_Due	0.4311463586
ags3_Pctl75_CMax_Due	0.4032008975
ags3_Pctl95_CMax_Due	0.4032008975
ags3_Mean_CMax_Due	0.4145473767
ags3_Max_CMax_Due	0.4032008975
ags3_Sum_CMax_Due	0.4143789536
ags3_Pctl75_CMin_Due	0.4036577549
ags3_Pctl95_CMin_Due	0.4036577549
ags3_Mean_CMin_Due	0.4004595947
ags3_Max_CMin_Due	0.4036577549
ags3_Sum_CMin_Due	0.4003432239
ags3_n_cus_arrears	0.403764389

Korelacja między zmiennymi objaśniającymi

W następnym kroku sprawdziliśmy, czy istnieje współzależność pomiędzy zmiennymi objaśniającymi, które mają największy wpływ na funkcję celu. Poniższe wyniki wskazują, że w zbiorze valid występuje problem współliniowości pomiędzy niektórymi zmiennymi (współczynnik korelacji jest równy 1). W zbiorze valid problem ten występuje między innymi w parach:

- Ags3_Pctl75_CMax_Due i Ags3_Pctl95_CMax_Due
- Ags3_Pctl75_CMax_Due i ags3_max_Cmax_Due
- Ags3_Pctl95_CMax_Due i ags3_max_Cmax_Due
- Ags3_pctl75_Cmin_Due i ags3_pctl95_cmin_due
- Ags3_max_cmin_due i Ags3_pctl75_Cmin_Due
- Ags3_max_cmin_due i ags3_pctl95_cmin_due

Kształtowanie się korelacji w zbiorze train

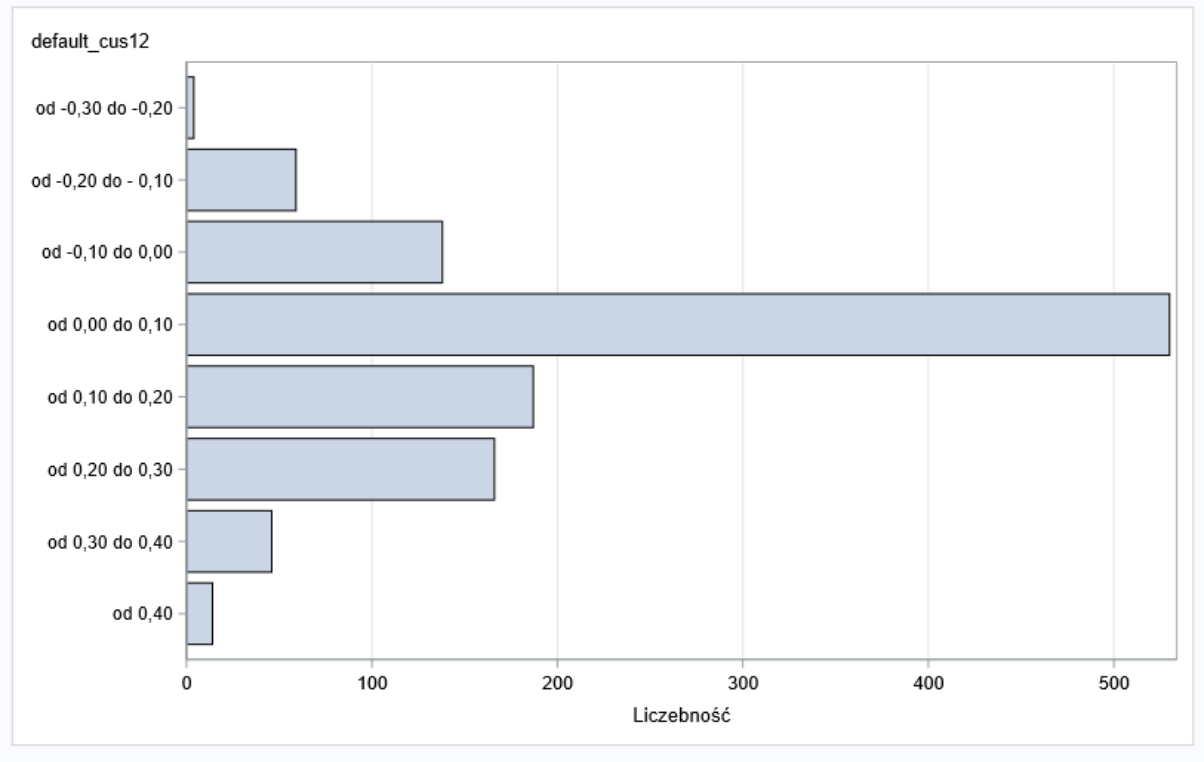
Analogiczna analiza współczynnika korelacji została przeprowadzona na zbiorze train. Po usunięciu zbędnych zmiennych objaśniających w zbierze train pozostało 1144 zmiennych. Współczynnik korelacji pomiędzy funkcją celu *default*_cus12, a zmiennymi objaśniającymi dla ok. 80% przypadków waha się od -0.20 do 0.20. Natomiast dla ponad 18% zmiennych korelacja jest na niskim poziomie, należy do przedziałów od –0.30 do –0.20 lub od 0.20 do 0.40. W zbiorze train zarówno jak w zbiorze valid do dalszej analizy zostały wyciągnięte zmienne, których współczynnik korelacji jest na poziomie umiarkowanym.

Tabela częstości współczynnika korelacji w zbiorze train

Procedura FREQ

default_cus12	Liczebność	Procent	Liczebność skumulowana	Procent skumulowany
od -0,30 do -0,20	4	0.35	4	0.35
od -0,20 do - 0,10	59	5.16	63	5.51
od -0,10 do 0,00	138	12.06	201	17.57
od 0,00 do 0,10	530	46.33	731	63.90
od 0,10 do 0,20	187	16.35	918	80.24
od 0,20 do 0,30	166	14.51	1084	94.76
od 0,30 do 0,40	46	4.02	1130	98.78
od 0,40	14	1.22	1144	100.00

Przedstawienie częstości współczynnika korelacji na histogramie:



Zmienne o najwyższym współczynniku korelacji w zbiorze train:

name	default_cus12
act_cus_dueutl	0.460586491
act_state_1_CMax_Due	0.471924497
act_state_1_CMin_Due	0.438992701
ags3_Pctl75_CMax_Due	0.4074393671
ags3_Pctl95_CMax_Due	0.4074393671
ags3_Mean_CMax_Due	0.4202990705
ags3_Max_CMax_Due	0.4074393671
ags3_Sum_CMax_Due	0.4198011801
ags3_Pctl75_CMin_Due	0.4030736394
ags3_Pctl95_CMin_Due	0.4030736394
ags3_Mean_CMin_Due	0.4034623466
ags3_Max_CMin_Due	0.4030736394

ags3_Sum_CMin_Due	0.4029823468
ags3_n_cus_arrears	0.4087763068

Korelacja między zmiennymi objaśniającymi:

Podobnie jak w zbiorze valid w zbiorze train przy sprawdzaniu korelacji pomiędzy zmiennymi objaśniającymi pojawił się problem współliniowości dla poniższych par zmiennych:

- Ags3_pctl75_Cmax_Due i ags3_pctl95_cmax_due
- Ags3_pctl75_Cmax_Due i ags3_max_cmax_due
- Ags3_pctl95_cmax_due i ags3_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