SAS Project 1&2

(Anastasia, Jiyoung)

TASKS:

1: "Analysis of missing data, their shares in time (i.e. stability in time) and comparison of shares between train and valid datasets (i.e. dataset stability) in the form of a detailed tabular report."

In the first part of our report we are going to analyze the missing data in train and valid datasets, its shares in time (monthly, annually) as well as present visualization reports that prove us with insights about our data, which can be useful for our future analysis.

To start with let us take a glimpse of our missing values as they are presented in original data by running a simple SAS code. Below are screenshots (full reports are attached):

Train

The MEANS Procedure

Variable	N	N Miss
act_age	52841	0
app_income	52841	0
app_number_of_children	52841	0
app_spendings	52841	0
act_cus_seniority	52841	0
act_cus_n_loans_hist	52841	0
act_cus_n_statC	52841	0
act_cus_n_statB	52841	0
act_cus_n_loans_act	52841	0
act_cus_pins	52841	0
act_cus_utl	52841	0
act_cus_dueutl	52841	0
act_cus_cc	52841	0
act_state_1_CMax_Days	45552	7289
act_state_2_CMax_Days	44287	8554
act_state_3_CMax_Days	42908	9933
act_state_4_CMax_Days	40968	11873
act_state_5_CMax_Days	39016	13825
act_state_6_CMax_Days	36965	15876
act_state_7_CMax_Days	34843	17998
act_state_8_CMax_Days	32611	20230
act_state_9_CMax_Days	30412	22429
act_state_10_CMax_Days	28377	24464
act_state_11_CMax_Days	26363	26478
act_state_12_CMax_Days	24462	28379
act_state_13_CMax_Days	22413	30428
act_state_14_CMax_Days	20598	32243
act_state_15_CMax_Days	18858	33983
act_state_16_CMax_Days	17196	35645
act_state_17_CMax_Days	15479	37362
act_state_18_CMax_Days	13938	38903
act_state_19_CMax_Days	12405	40436
act_state_20_CMax_Days	10948	41893
act_state_21_CMax_Days	9561	43280
act_state_22_CMax_Days	8196	44645
act_state_23_CMax_Days	6961	45880
act_state_24_CMax_Days	5724	47117
act_state_25_CMax_Days	4449	48392
act_state_26_CMax_Days	4016	48825
act_state_27_CMax_Days	3952	48889
act_state_28_CMax_Days	3989	48852
act_state_29_CMax_Days	4136	48705
act_state_30_CMax_Days	4231	48610
act_state_31_CMax_Days	4270	48571
act_state_32_CMax_Days	4394	48447
act_state_33_CMax_Days	4457	48384
act_state_34_CMax_Days	4540	48301

Valid

The MEANS Procedure

Variable	N	N Miss
act_age	53070	0
app_income	53070	0
app_number_of_children	53070	0
app_spendings	53070	0
act_cus_seniority	53070	0
act_cus_n_loans_hist	53070	0
act_cus_n_statC	53070	0
act_cus_n_statB	53070	0
act_cus_n_loans_act	53070	0
act_cus_pins	53070	0
act_cus_utl	53070	0
act_cus_dueutl	53070	0
act_cus_cc	53070	0
act_state_1_CMax_Days	45827	7243
act_state_2_CMax_Days	44291	8779
act_state_3_CMax_Days	43075	9995
act_state_4_CMax_Days	41134	11936
act_state_5_CMax_Days	39102	13968
act_state_6_CMax_Days	37135	15935
act_state_7_CMax_Days	34978	18092
act_state_8_CMax_Days	32836	20234
act_state_9_CMax_Days	30928	22142
act_state_10_CMax_Days	28723	24347
act_state_11_CMax_Days	26715	26355
act_state_12_CMax_Days	24746	28324
act_state_13_CMax_Days	22687	30383
act_state_14_CMax_Days	20909	32161
act_state_15_CMax_Days	19182	33888
act_state_16_CMax_Days	17466	35604
act_state_17_CMax_Days	15874	37196
act_state_18_CMax_Days	14243	38827
act_state_19_CMax_Days	12658	40412
act_state_20_CMax_Days	11208	41862
act_state_21_CMax_Days	9770	43300
act_state_22_CMax_Days	8439	44631
act_state_23_CMax_Days	7120	45950
act_state_24_CMax_Days	5814	47256
act_state_25_CMax_Days	4564	48506
act_state_26_CMax_Days	4175	48895
act_state_27_CMax_Days	4125	48945
act_state_28_CMax_Days	4110	48960
act_state_29_CMax_Days	4224	48846
act_state_30_CMax_Days	4298	48772
act_state_31_CMax_Days	4389	48681
act_state_32_CMax_Days	4455	48615
act_state_33_CMax_Days	4540	48530

Now we want to look closely at the train dataset and see the percent share of missing values. We are placing screenshots of pdf reports. Full versions are in the attachments.

Missing Data Summary
(for all numeric variables in Train dataset)

Variable	NMiss	N	train_pct_missing
agr12_Iqr_Cncr	52841	0	100%
agr12_Kurtosis_Cncr	52841	0	100%
agr12_Max_Cncr	52841	0	100%
agr12_Mean_Cncr	52841	0	100%
agr12_Median_Cncr	52841	0	100%
agr12_Min_Cncr	52841	0	100%
agr12_N_Cncr	52841	0	100%
agr12_Nmiss_Cncr	52841	0	100%
agr12_Pctl25_Cncr	52841	0	100%
agr12_Pctl5_Cncr	52841	0	100%
agr12_Pctl75_Cncr	52841	0	100%
agr12_Pctl95_Cncr	52841	0	100%
agr12_Range_Cncr	52841	0	100%
agr12_Skewness_Cncr	52841	0	100%
agr12_Std_Cncr	52841	0	100%
agr12_Sum_Cncr	52841	0	100%
agr15_Iqr_Cncr	52841	0	100%
agr15_Kurtosis_Cncr	52841	0	100%
agr15_Max_Cncr	52841	0	100%
agr15_Mean_Cncr	52841	0	100%
agr15_Median_Cncr	52841	0	100%
agr15_Min_Cncr	52841	0	100%
agr15_N_Cncr	52841	0	100%
agr15_Nmiss_Cncr	52841	0	100%
agr15_Pctl25_Cncr	52841	0	100%
agr15_Pctl5_Cncr	52841	0	100%
agr15_Pctl75_Cncr	52841	0	100%
agr15_Pctl95_Cncr	52841	0	100%
agr15_Range_Cncr	52841	0	100%
agr15_Skewness_Cncr	52841	0	100%
agr15_Std_Cncr	52841	0	100%
agr15_Sum_Cncr	52841	0	100%
agr18_Iqr_Cncr	52841	0	100%
agr18_Kurtosis_Cncr	52841	0	100%
agr18_Max_Cncr	52841	0	100%
agr18_Mean_Cncr	52841	0	100%
agr18_Median_Cncr	52841	0	100%

Missing Data Summary
(for all numeric variables in Train dataset)

Variable	NMiss	N	train_pct_missing
agr15_Kurtosis_CMax_Days	47217	5624	89.4%
agr15_Kurtosis_CMin_Days	47217	5624	89.4%
agr15_Max_CMax_Days	47217	5624	89.4%
agr15_Max_CMin_Days	47217	5624	89.4%
agr15_Mean_CMax_Days	47217	5624	89.4%
agr15_Mean_CMin_Days	47217	5624	89.4%
agr15_Median_CMax_Days	47217	5624	89.4%
agr15_Median_CMin_Days	47217	5624	89.4%
agr15_Min_CMax_Days	47217	5624	89.4%
agr15_Min_CMin_Days	47217	5624	89.4%
agr15_N_CMax_Days	47217	5624	89.4%
agr15_N_CMin_Days	47217	5624	89.4%
agr15_Nmiss_CMax_Days	47217	5624	89.4%
agr15_Nmiss_CMin_Days	47217	5624	89.4%
agr15_Pctl25_CMax_Days	47217	5624	89.4%
agr15_Pctl25_CMin_Days	47217	5624	89.4%
agr15_Pctl5_CMax_Days	47217	5624	89.4%
agr15_Pctl5_CMin_Days	47217	5624	89.4%
agr15_Pctl75_CMax_Days	47217	5624	89.4%
agr15_Pctl75_CMin_Days	47217	5624	89.4%
agr15_Pctl95_CMax_Days	47217	5624	89.4%
agr15_Pctl95_CMin_Days	47217	5624	89.4%
agr15_Range_CMax_Days	47217	5624	89.4%
agr15_Range_CMin_Days	47217	5624	89.4%
agr15_Skewness_CMax_Days	47217	5624	89.4%
agr15_Skewness_CMin_Days	47217	5624	89.4%
agr15_Std_CMax_Days	47217	5624	89.4%
agr15_Std_CMin_Days	47217	5624	89.4%
agr15_Sum_CMax_Days	47217	5624	89.4%
agr15_Sum_CMin_Days	47217	5624	89.4%
act_state_35_CMax_Due	47167	5674	89.3%
act_state_35_CMin_Due	47167	5674	89.3%
act_state_24_CMax_Days	47117	5724	89.2%
act_state_24_CMin_Days	47117	5724	89.2%
act_state_36_CMax_Due	47111	5730	89.2%
act_state_36_CMin_Due	47111	5730	89.2%
act_state_24_CMax_Due	46235	6606	87.5%

Missing Data Summary
(for all numeric variables in Valid dataset)

Variable	NMiss	N	valid_pct_missing
agr12_Iqr_Cncr	53070	0	100%
agr12_Kurtosis_Cncr	53070	0	100%
agr12_Max_Cncr	53070	0	100%
agr12_Mean_Cncr	53070	0	100%
agr12_Median_Cncr	53070	0	100%
agr12_Min_Cncr	53070	0	100%
agr12_N_Cncr	53070	0	100%
agr12_Nmiss_Cncr	53070	0	100%
agr12_Pctl25_Cncr	53070	0	100%
agr12_Pctl5_Cncr	53070	0	100%
agr12_Pctl75_Cncr	53070	0	100%
agr12_Pctl95_Cncr	53070	0	100%
agr12_Range_Cncr	53070	0	100%
agr12_Skewness_Cncr	53070	0	100%
agr12_Std_Cncr	53070	0	100%
agr12_Sum_Cncr	53070	0	100%
agr15_Iqr_Cncr	53070	0	100%
agr15_Kurtosis_Cncr	53070	0	100%
agr15_Max_Cncr	53070	0	100%
agr15_Mean_Cncr	53070	0	100%
agr15_Median_Cncr	53070	0	100%
agr15_Min_Cncr	53070	0	100%
agr15_N_Cncr	53070	0	100%
agr15_Nmiss_Cncr	53070	0	100%
agr15_Pctl25_Cncr	53070	0	100%
agr15_Pctl5_Cncr	53070	0	100%
agr15_Pctl75_Cncr	53070	0	100%
agr15_Pctl95_Cncr	53070	0	100%
agr15_Range_Cncr	53070	0	100%
agr15_Skewness_Cncr	53070	0	100%
agr15_Std_Cncr	53070	0	100%
agr15_Sum_Cncr	53070	0	100%
agr18_Iqr_Cncr	53070	0	100%
agr18_Kurtosis_Cncr	53070	0	100%
agr18_Max_Cncr	53070	0	100%
agr18_Mean_Cncr	53070	0	100%
agr18_Median_Cncr	53070	0	100%

Missing Data Summary
(for all numeric variables in Valid dataset)

Variable	NMiss	N	valid_pct_missing
ags3_Mean_CMin_Due	0	53070	0.0%
ags3_Median_CMax_Due	0	53070	0.0%
ags3_Median_CMin_Due	0	53070	0.0%
ags3_Min_CMax_Due	0	53070	0.0%
ags3_Min_CMin_Due	0	53070	0.0%
ags3_N_CMax_Days	0	53070	0.0%
ags3_N_CMax_Due	0	53070	0.0%
ags3_N_CMin_Days	0	53070	0.0%
ags3_N_CMin_Due	0	53070	0.0%
ags3_N_Cncr	0	53070	0.0%
ags3_Nmiss_CMax_Days	0	53070	0.0%
ags3_Nmiss_CMax_Due	0	53070	0.0%
ags3_Nmiss_CMin_Days	0	53070	0.0%
ags3_Nmiss_CMin_Due	0	53070	0.0%
ags3_Nmiss_Cncr	0	53070	0.0%
ags3_Pctl25_CMax_Due	0	53070	0.0%
ags3_Pctl25_CMin_Due	0	53070	0.0%
ags3_Pctl5_CMax_Due	0	53070	0.0%
ags3_Pctl5_CMin_Due	0	53070	0.0%
ags3_Pctl75_CMax_Due	0	53070	0.0%
ags3_Pctl75_CMin_Due	0	53070	0.0%
ags3_Pctl95_CMax_Due	0	53070	0.0%
ags3_Pctl95_CMin_Due	0	53070	0.0%
ags3_Range_CMax_Due	0	53070	0.0%
ags3_Range_CMin_Due	0	53070	0.0%
ags3_Sum_CMax_Due	0	53070	0.0%
ags3_Sum_CMin_Due	0	53070	0.0%
ags3_n_cus_arrears	0	53070	0.0%
ags3_n_cus_arrears_days	0	53070	0.0%
ags3_n_cus_good_days	0	53070	0.0%
ags6_Csev_all	0	53070	0.0%
ags6_Csev_family	0	53070	0.0%
ags6_Csev_health	0	53070	0.0%
ags6_Csev_home	0	53070	0.0%
ags6_Csev_work	0	53070	0.0%
ags6_Iqr_CMax_Days	0	53070	0.0%
ags6_Iqr_CMax_Due	0	53070	0.0%

From our first reports we can see that many variables have huge shares of missing data, so for better results we are dropping variables which have missing values containing more than 97.5% of the variable values.

As we want to compare missing values of both datasets, we are joining results of missing share taking a closer look to share of missing values and the differences of these shares in both dataset:

Missing Data Summary

Friday, January 29, 2021 05:01:35 PM

(for all numeric variables in Valid and Train datasets without variables with share of missing values more than 97.5%)

Variable	trainNMiss	validNMiss	train_percent_miss	valid_percent_miss
act_state_9_Cncr	50288	50508	95.2%	95.2%
act_state_8_Cncr	50208	50485	95.0%	95.1%
act_state_7_Cncr	50149	50426	94.9%	95.0%
act_state_5_Cncr	50079	50351	94.8%	94.9%
act_state_4_Cncr	50076	50348	94.8%	94.9%
act_state_6_Cncr	50053	50339	94.7%	94.9%
act_state_2_Cncr	50033	50306	94.7%	94.8%
act_state_3_Cncr	50057	50282	94.7%	94.7%
act_Cncr	50061	50160	94.7%	94.5%
act_state_1_Cncr	50061	50160	94.7%	94.5%
ags27_Std_Cncr	49892	50153	94.4%	94.5%
ags30_Std_Cncr	49458	49676	93.6%	93.6%
agr18_Iqr_CMax_Days	49500	49574	93.7%	93.4%
agr18_Iqr_CMin_Days	49500	49574	93.7%	93.4%
agr18_Kurtosis_CMax_Days	49500	49574	93.7%	93.4%
agr18_Kurtosis_CMin_Days	49500	49574	93.7%	93.4%
agr18_Max_CMax_Days	49500	49574	93.7%	93.4%
agr18_Max_CMin_Days	49500	49574	93.7%	93.4%
agr18_Mean_CMax_Days	49500	49574	93.7%	93.4%
agr18_Mean_CMin_Days	49500	49574	93.7%	93.4%
agr18_Median_CMax_Days	49500	49574	93.7%	93.4%
agr18_Median_CMin_Days	49500	49574	93.7%	93.4%
agr18_Min_CMax_Days	49500	49574	93.7%	93.4%
agr18_Min_CMin_Days	49500	49574	93.7%	93.4%
agr18_N_CMax_Days	49500	49574	93.7%	93.4%
agr18_N_CMin_Days	49500	49574	93.7%	93.4%
agr18_Nmiss_CMax_Days	49500	49574	93.7%	93.4%
agr18_Nmiss_CMin_Days	49500	49574	93.7%	93.4%
agr18_Pctl25_CMax_Days	49500	49574	93.7%	93.4%
agr18_Pctl25_CMin_Days	49500	49574	93.7%	93.4%
agr18_Pctl5_CMax_Days	49500	49574	93.7%	93.4%
agr18_Pctl5_CMin_Days	49500	49574	93.7%	93.4%
agr18_Pctl75_CMax_Days	49500	49574	93.7%	93.4%
agr18_Pctl75_CMin_Days	49500	49574	93.7%	93.4%
agr18_Pctl95_CMax_Days	49500	49574	93.7%	93.4%
agr18_Pctl95_CMin_Days	49500	49574	93.7%	93.4%

As there are too many variables it will be easier to divide them by groups and check if there are similarities depending on the name of the variable. We check those who start with act, ags, agr.

Friday, January 29, 2021 0

Missing Data Summary for Variables start with act_ (describes state at a given point) (for all numeric variables in Valid and Train datasets)

Variable	trainNMiss	validNMiss	train_percent_miss	valid_percent_miss
act_state_36_CMin_Due	47111	47287	89.2%	89.1%
act_state_24_CMax_Days	47117	47256	89.2%	89.0%
act_state_24_CMin_Days	47117	47256	89.2%	89.0%
act_state_24_CMax_Due	46235	46307	87.5%	87.3%
act_state_24_CMin_Due	46235	46307	87.5%	87.3%
act_state_23_CMax_Days	45880	45950	86.8%	86.6%
act_state_23_CMin_Days	45880	45950	86.8%	86.6%
act_state_23_CMax_Due	44897	44968	85.0%	84.7%
act_state_23_CMin_Due	44897	44968	85.0%	84.7%
act_state_22_CMax_Days	44645	44631	84.5%	84.1%
act_state_22_CMin_Days	44645	44631	84.5%	84.1%
act_state_22_CMax_Due	43586	43600	82.5%	82.2%
act_state_22_CMin_Due	43586	43600	82.5%	82.2%
act_state_21_CMax_Days	43280	43300	81.9%	81.6%
act_state_21_CMin_Days	43280	43300	81.9%	81.6%
act_state_21_CMax_Due	42169	42203	79.8%	79.5%
act_state_21_CMin_Due	42169	42203	79.8%	79.5%
act_state_20_CMax_Days	41893	41862	79.3%	78.9%
act_state_20_CMin_Days	41893	41862	79.3%	78.9%
act_state_20_CMax_Due	40726	40686	77.1%	76.7%
act_state_20_CMin_Due	40726	40686	77.1%	76.7%
act_state_19_CMax_Days	40436	40412	76.5%	76.1%
act_state_19_CMin_Days	40436	40412	76.5%	76.1%
act_state_19_CMax_Due	39173	39189	74.1%	73.8%
act_state_19_CMin_Due	39173	39189	74.1%	73.8%
act_state_18_CMax_Days	38903	38827	73.6%	73.2%
act_state_18_CMin_Days	38903	38827	73.6%	73.2%
act_state_18_CMax_Due	37560	37510	71.1%	70.7%
act_state_18_CMin_Due	37560	37510	71.1%	70.7%
act_state_17_CMax_Days	37362	37196	70.7%	70.1%
act_state_17_CMin_Days	37362	37196	70.7%	70.1%
act_state_17_CMax_Due	35919	35748	68.0%	67.4%
act_state_17_CMin_Due	35919	35748	68.0%	67.4%
act_state_16_CMax_Days	35645	35604	67.5%	67.1%
act_state_16_CMin_Days	35645	35604	67.5%	67.1%
act_state_16_CMax_Due	34133	34028	64.6%	64.1%

Missing Data Summary for Variables start with agr (aggregated information during previous months) (for all numeric variables in Valid and Train datasets)

Variable	trainNMiss	validNMiss	train_percent_miss	valid_percent_miss
ags18_Std_Cncr	51255	51510	97.0%	97.1%
ags21_Std_Cncr	50794	51079	96.1%	96.2%
ags24_Std_Cncr	50358	50590	95.3%	95.3%
ags27_Std_Cncr	49892	50153	94.4%	94.5%
ags30_Std_Cncr	49458	49676	93.6%	93.6%
ags33_Std_Cncr	49006	49169	92.7%	92.6%
ags36_Std_Cncr	48521	48688	91.8%	91.7%
ags3_Iqr_Cncr	44518	44654	84.2%	84.1%
ags3_Max_Cncr	44518	44654	84.2%	84.1%
ags3_Mean_Cncr	44518	44654	84.2%	84.1%
ags3_Median_Cncr	44518	44654	84.2%	84.1%
ags3_Min_Cncr	44518	44654	84.2%	84.1%
ags3_Pctl25_Cncr	44518	44654	84.2%	84.1%
ags3_Pctl5_Cncr	44518	44654	84.2%	84.1%
ags3_Pctl75_Cncr	44518	44654	84.2%	84.1%
ags3_Pctl95_Cncr	44518	44654	84.2%	84.1%
ags3_Range_Cncr	44518	44654	84.2%	84.1%
ags3_Sum_Cncr	44518	44654	84.2%	84.1%
ags3_Skewness_CMin_Due	37953	38144	71.8%	71.9%
ags3_Skewness_CMax_Due	37167	37346	70.3%	70.4%
ags6_Iqr_Cncr	36400	36652	68.9%	69.1%
ags6_Max_Cncr	36400	36652	68.9%	69.1%
ags6_Mean_Cncr	36400	36652	68.9%	69.1%
ags6_Median_Cncr	36400	36652	68.9%	69.1%
ags6_Min_Cncr	36400	36652	68.9%	69.1%
ags6_Pctl25_Cncr	36400	36652	68.9%	69.1%
ags6_Pctl5_Cncr	36400	36652	68.9%	69.1%
ags6_Pctl75_Cncr	36400	36652	68.9%	69.1%
ags6_Pctl95_Cncr	36400	36652	68.9%	69.1%
ags6_Range_Cncr	36400	36652	68.9%	69.1%
ags6_Sum_Cncr	36400	36652	68.9%	69.1%
ags6_Kurtosis_CMin_Due	31237	31361	59.1%	59.1%
ags6_Skewness_CMin_Due	30796	30909	58.3%	58.2%
ags6_Kurtosis_CMax_Due	30279	30455	57.3%	57.4%
ags6_Skewness_CMax_Due	29832	30001	56.5%	56.5%
ags9_Iqr_Cncr	28780	29125	54.5%	54.9%

Missing Data Summary for Variables start with ags (aggregated information during previous months) (for all numeric variables in Valid and Train datasets)

Variable	trainNMiss	validNMiss	train_percent_miss	valid_percent_miss
ags9_Max_Cncr	28780	29125	54.5%	54.9%
ags9_Mean_Cncr	28780	29125	54.5%	54.9%
ags9_Median_Cncr	28780	29125	54.5%	54.9%
ags9_Min_Cncr	28780	29125	54.5%	54.9%
ags9_Pctl25_Cncr	28780	29125	54.5%	54.9%
ags9_Pctl5_Cncr	28780	29125	54.5%	54.9%
ags9_Pctl75_Cncr	28780	29125	54.5%	54.9%
ags9_Pctl95_Cncr	28780	29125	54.5%	54.9%
ags9_Range_Cncr	28780	29125	54.5%	54.9%
ags9_Sum_Cncr	28780	29125	54.5%	54.9%
ags9_Kurtosis_CMin_Due	28199	28193	53.4%	53.1%
ags9_Skewness_CMin_Due	27776	27750	52.6%	52.3%
ags9_Kurtosis_CMax_Due	27248	27285	51.6%	51.4%
ags9_Skewness_CMax_Due	26820	26841	50.8%	50.6%
ags12_Kurtosis_CMin_Due	26503	26547	50.2%	50.0%
ags12_Skewness_CMin_Due	26080	26112	49.4%	49.2%
ags12_Kurtosis_CMax_Due	25602	25690	48.5%	48.4%
ags15_Kurtosis_CMin_Due	25485	25563	48.2%	48.2%
ags12_Skewness_CMax_Due	25174	25254	47.6%	47.6%
ags15_Skewness_CMin_Due	25068	25133	47.4%	47.4%
ags18_Kurtosis_CMin_Due	24874	24889	47.1%	46.9%
ags15_Kurtosis_CMax_Due	24665	24752	46.7%	46.6%
ags18_Skewness_CMin_Due	24461	24462	46.3%	46.1%
ags21_Kurtosis_CMin_Due	24447	24457	46.3%	46.1%
ags15_Skewness_CMax_Due	24243	24321	45.9%	45.8%
ags18_Kurtosis_CMax_Due	24120	24152	45.6%	45.5%
ags24_Kurtosis_CMin_Due	24125	24090	45.7%	45.4%
ags21_Skewness_CMin_Due	24048	24031	45.5%	45.3%
ags21_Kurtosis_CMax_Due	23748	23774	44.9%	44.8%
ags27_Kurtosis_CMin_Due	23827	23753	45.1%	44.8%
ags18_Skewness_CMax_Due	23702	23724	44.9%	44.7%
ags24_Skewness_CMin_Due	23723	23668	44.9%	44.6%
ags30_Kurtosis_CMin_Due	23546	23473	44.6%	44.2%
ags24_Kurtosis_CMax_Due	23470	23448	44.4%	44.2%
ags21_Skewness_CMax_Due	23344	23347	44.2%	44.0%
ags27_Skewness_CMin_Due	23436	23338	44.4%	44.0%

Now let's compare stability in time. Firstly, we would like to check missing values by year. This is just a small part of the report. We can observe how most of the share of missing values is stable in time.

Missing Data Summary by Year
(for all numeric variables in Train dataset)

Friday, January 29, 2021 05:16:36 PM

Variable	pct_missing2004	pct_missing2005	pct_missing2006	pct_missing2007	pct_missing2008
act_state_26_CMin_Days	81.2%	86.5%	89.4%	91.8%	92.6%
act_state_26_CMin_Due	77.4%	82.9%	87.1%	89.8%	90.9%
act_state_26_Cncr	99.0%	99.0%	99.1%	99.2%	99.2%
act_state_27_CMax_Days	80.8%	85.9%	89.2%	91.9%	92.9%
act_state_27_CMax_Due	77.2%	82.3%	87.1%	90.0%	91.1%
act_state_27_CMin_Days	80.8%	85.9%	89.2%	91.9%	92.9%
act_state_27_CMin_Due	77.2%	82.3%	87.1%	90.0%	91.1%
act_state_27_Cncr	99.0%	99.0%	99.1%	99.4%	99.4%
act_state_28_CMax_Days	80.2%	85.4%	89.5%	92.1%	93.0%
act_state_28_CMax_Due	76.6%	81.9%	87.0%	89.9%	90.9%
act_state_28_CMin_Days	80.2%	85.4%	89.5%	92.1%	93.0%

pct_missing2009	pct_missing2010	pct_missing2011	pct_missing2012	pct_missing2013	pct_missing2014
92.7%	93.4%	93.2%	94.5%	94.8%	94.8%
90.9%	91.9%	91.9%	93.0%	93.3%	93.4%
99.3%	99.3%	99.4%	99.5%	99.5%	99.4%
93.1%	93.8%	93.6%	94.8%	94.5%	94.9%
91.1%	92.1%	92.1%	93.2%	93.3%	93.7%
93.1%	93.8%	93.6%	94.8%	94.5%	94.9%
91.1%	92.1%	92.1%	93.2%	93.3%	93.7%
99.6%	99.4%	99.4%	99.7%	99.7%	99.7%
92.9%	94.1%	93.7%	94.4%	94.4%	95.0%
90.9%	92.2%	92.3%	93.1%	93.3%	93.6%
92.9%	94.1%	93.7%	94.4%	94.4%	95.0%

pct_missing2015	pct_missing2016	pct_missing2017	pct_missing2018
95.1%	94.7%	95.2%	95.5%
93.7%	93.6%	94.5%	94.5%
99.2%	99.5%	99.3%	99.6%
95.3%	95.1%	95.8%	95.4%
94.1%	93.7%	95.0%	94.5%
95.3%	95.1%	95.8%	95.4%
94.1%	93.7%	95.0%	94.5%
99.8%	99.7%	99.8%	99.6%
95.1%	94.8%	95.7%	95.8%
93.9%	93.5%	94.9%	94.6%
95.1%	94.8%	95.7%	95.8%

Missing Data Summary by Month
(for all numeric variables in Train dataset)

Variable	pct_missing1	pct_missing2	pct_missing3	pct_missing4	pct_missing5	pct_missing6	pct_missing7	pct_missing8	pct_missing9	pct_missing10	pct_missing11	pct_missing12
act24_n_cind_arrears	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
act27_Ciev_all	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
act27_Ciev_family	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
act27_Ciev_health	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
act27_Ciev_home	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
act27_Ciev_work	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
act27_Cncr_taken	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
act27_n_cind_arrears	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
act30_Ciev_all	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
act30_Ciev_family	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
act30_Ciev_health	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
act30_Ciev_home	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
act30_Ciev_work	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
act30_Cncr_taken	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
act30_n_cind_arrears	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
act33_Ciev_all	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
act33_Ciev_family	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

What about character variables? We have checked the missing values for them as well. The conclusion is that there are no missing values for Variables of character type.

Friday, Ja

Frequency Count of Character variables

The FREQ Procedure

cid	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0000019242	3	0.01	18283	34.60
0000019243	2	0.00	18285	34.60
0000019245	2	0.00	18287	34.61
0000019247	1	0.00	18288	34.61
0000019249	1	0.00	18289	34.61
0000019250	2	0.00	18291	34.62
0000019255	1	0.00	18292	34.62
0000019259	3	0.01	18295	34.62
0000019261	1	0.00	18296	34.62
0000019262	1	0.00	18297	34.63
0000019265	1	0.00	18298	34.63
0000019268	2	0.00	18300	34.63
0000019269	1	0.00	18301	34.63
0000019271	2	0.00	18303	34.64
0000019273	1	0.00	18304	34.64
0000019274	2	0.00	18306	34.64
0000019275	1	0.00	18307	34.65
0000019276	1	0.00	18308	34.65
0000019278	1	0.00	18309	34.65
0000019279	1	0.00	18310	34.65
0000019280	2	0.00	18312	34.65
0000019282	3	0.01	18315	34.66
0000019283	1	0.00	18316	34.66
0000019288	1	0.00	18317	34.66
0000019292	3	0.01	18320	34.67
0000019294	1	0.00	18321	34.67
0000019295	3	0.01	18324	34.68
0000019298	2	0.00	18326	34.68
0000019299	1	0.00	18327	34.68
0000019302	5	0.01	18332	34.69
0000019304	1	0.00	18333	34.69
0000019306	2	0.00	18335	34.70
0000019307	1	0.00	18336	34.70
0000019308	1	0.00	18337	34.70
0000019309	1	0.00	18338	34.70

Frequency Count of Character variables

The FREQ Procedure

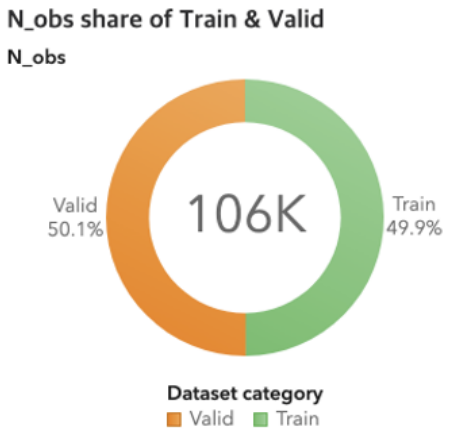
cid	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0000000001	1	0.00	1	0.00
0000000002	1	0.00	2	0.00
0000000003	2	0.00	4	0.01
0000000004	1	0.00	5	0.01
0000000005	1	0.00	6	0.01
0000000007	5	0.01	11	0.02
0000000009	1	0.00	12	0.02
0000000010	2	0.00	14	0.03
0000000011	1	0.00	15	0.03
0000000014	2	0.00	17	0.03
0000000016	1	0.00	18	0.03
0000000017	3	0.01	21	0.04
0000000018	2	0.00	23	0.04
0000000019	4	0.01	27	0.05
0000000020	1	0.00	28	0.05
0000000021	2	0.00	30	0.06
0000000022	3	0.01	33	0.06
0000000023	2	0.00	35	0.07
0000000025	1	0.00	36	0.07
0000000026	3	0.01	39	0.07
0000000027	1	0.00	40	0.08
0000000028	2	0.00	42	0.08
0000000029	6	0.01	48	0.09
0000000030	1	0.00	49	0.09
0000000032	1	0.00	50	0.09
0000000033	1	0.00	51	0.10
0000000035	1	0.00	52	0.10
0000000036	2	0.00	54	0.10
0000000038	1	0.00	55	0.10
0000000040	1	0.00	56	0.11
0000000041	4	0.01	60	0.11
0000000042	2	0.00	62	0.12
0000000043	3	0.01	65	0.12
0000000045	2	0.00	67	0.13
0000000046	3	0.01	70	0.13
0000000047	4	0.01	74	0.14

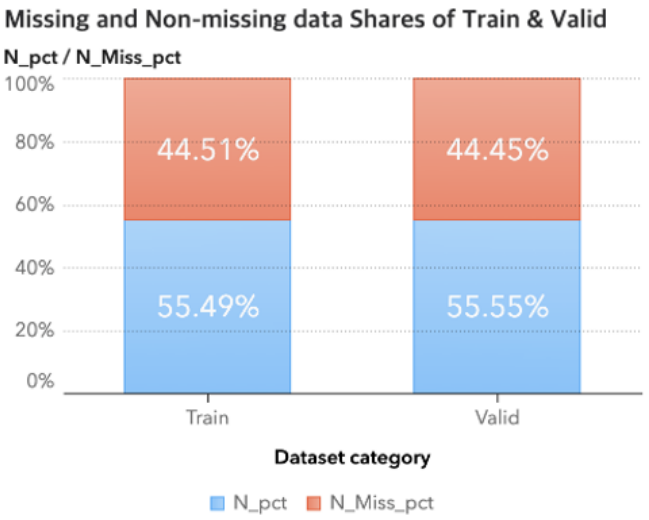
2: "Missing data analysis in the form of visualized graphical reports to quickly identify variables with more and fewer missing data and their different stability."

Dataset Stability (Comparison of Train & Valid datasets)

For the Task 2 'visualization of missing data analysis', we used SAS® Visual Analytics. First of all, we started from checking overall distribution of the dataset. Our dataset consists of 52,841 train obs

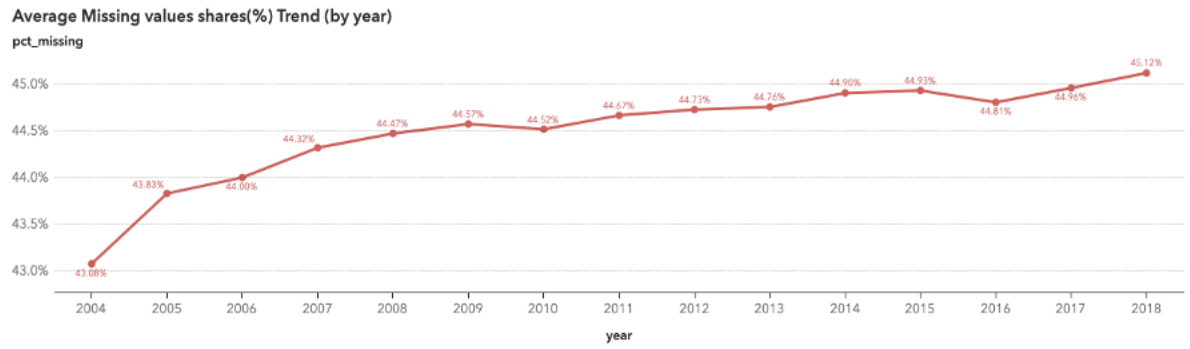
and 53,070 valid obs with almost 50%:50% proportion out of total 106K obs. The next plot shows the comparison of shares between train and valid dataset. Both Train and Valid dataset have a similar shape of data distribution with 44% of missing and 55% of non-missing values.

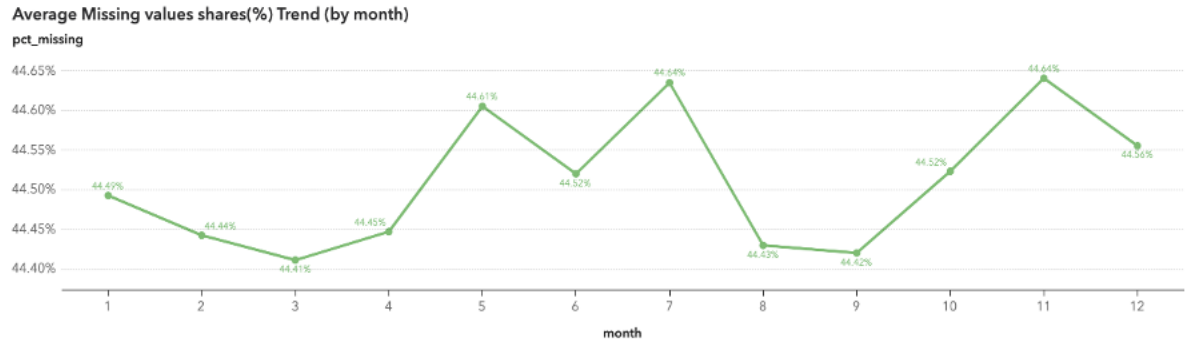




Stability in time (Trend of shares of missing values in time)

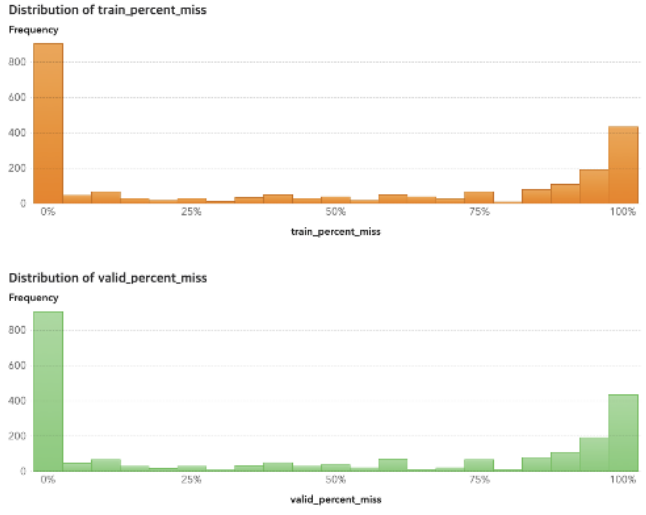
For the next step, we drew the plot showing how the shares of missing values are changing by time in 2 aspects based on the variable ‘period’: 1) by Year; 2) by Month. According to the plot, the average missing values shares has been steadily increased by year but mostly in the same level of share except for 2004. For monthly basis, we can see there is no specific pattern of steady increase or decrease. We could find some months like May, July and November have more missing values shares than others in average.

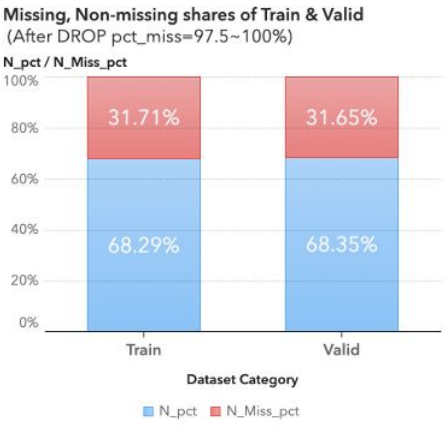




Could there be variables that are almost empty and not worth analysing?

In Task 1, we created the report table which shows missing data summary. Based on that, we drew histogram showing the distribution of variables based on missing percentage. In both train and valid, there are quite a lot of variables with range of missing percentage with 97.5%-100%, which actually means that they are totally empty and cannot even impute with other values so that cannot derive any useful insights with analysis. Therefore, we decided to drop all these 435 variables with 97.5%-100% of missing percentage from both datasets. After dropping them, we could see the missing value shares of both dataset have decreased a lot comparing to the original dataset.

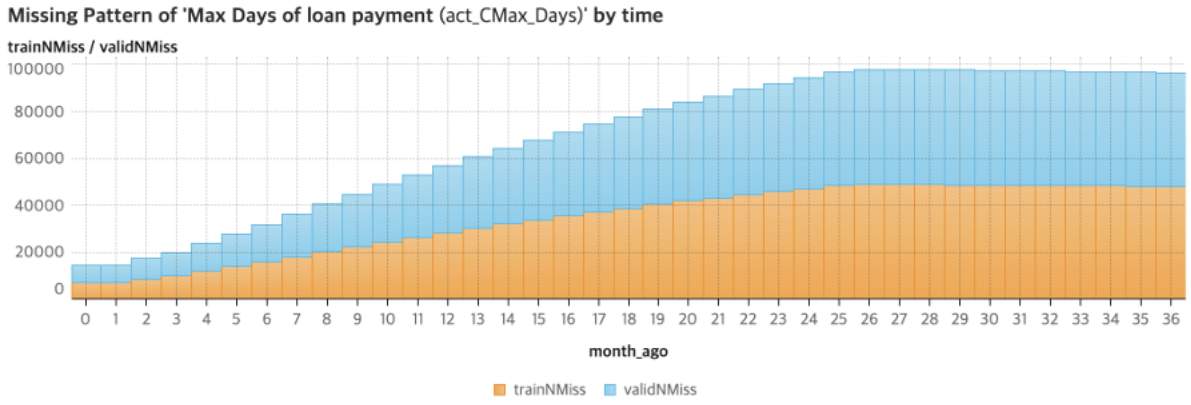


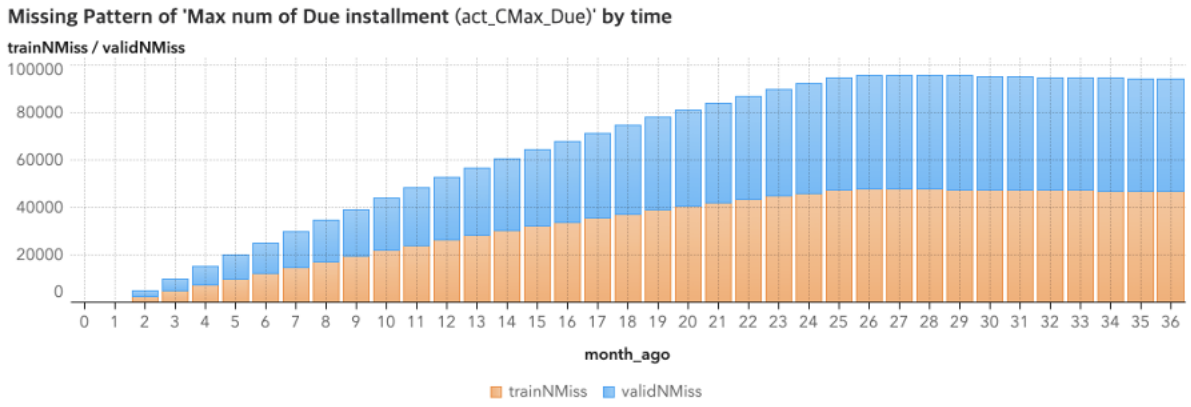


Is there any pattern of missing data?

Is the lack of data (a missing value) a random loss or maybe an information?

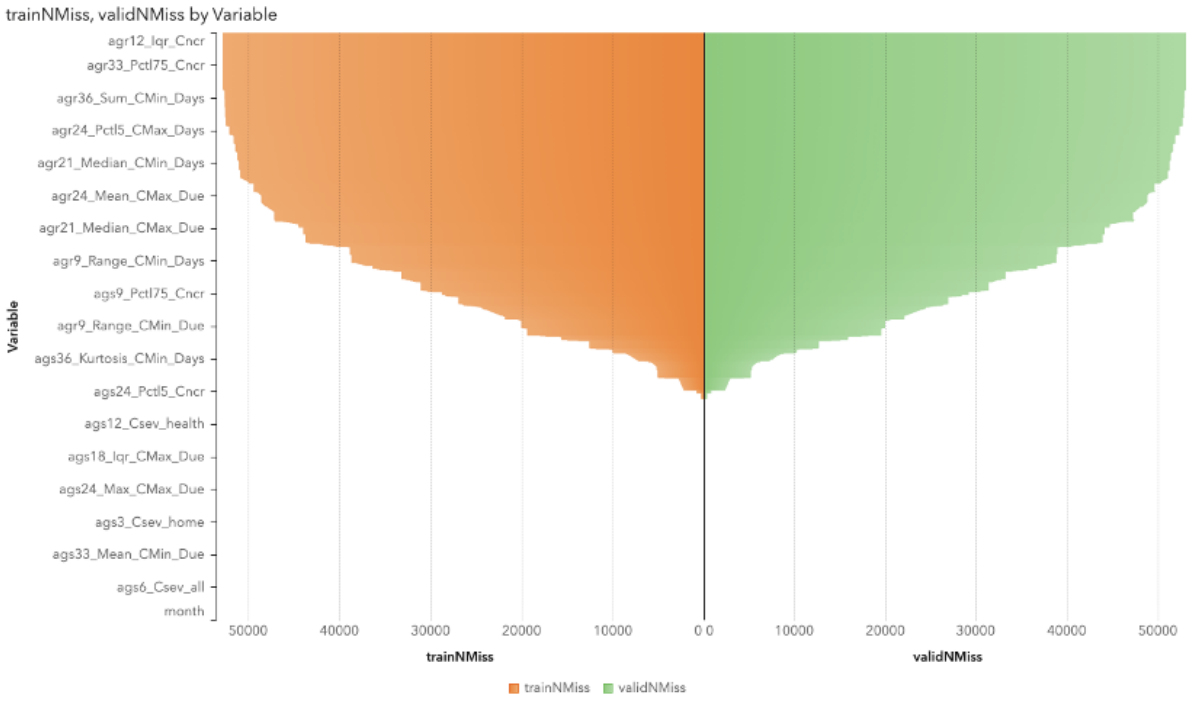
Some variables had the pattern of missing data and analyzing this pattern helped us to decide what variables to take or drop furthermore. For example, one of our variable ‘act_CMax_Days’(we regarded as act_state_0_CMax_Days in here) and ‘act_state_1~36_CMax_Days’ show that the share of missing values has been increased steadily by time(month). In here, x axis ‘month_ago’ means minus month from observed ‘period’ because this act_ variables arehistorical status data according to our labels.xlsx file. In detail, as the data is from more past, more months ago, there are more missing values. This can be interpreted that majority of credits taken were done in less than 36 instalments. So that’s why there are more missing values as time goes by. We could observe same pattern in ‘act_CMax_Due’ variable too. Since these act_ group variables can give us this kind of good information about credit loan payment pattern, we decided to keep all of these act_ group variables even though they have quite a lot of missing values.

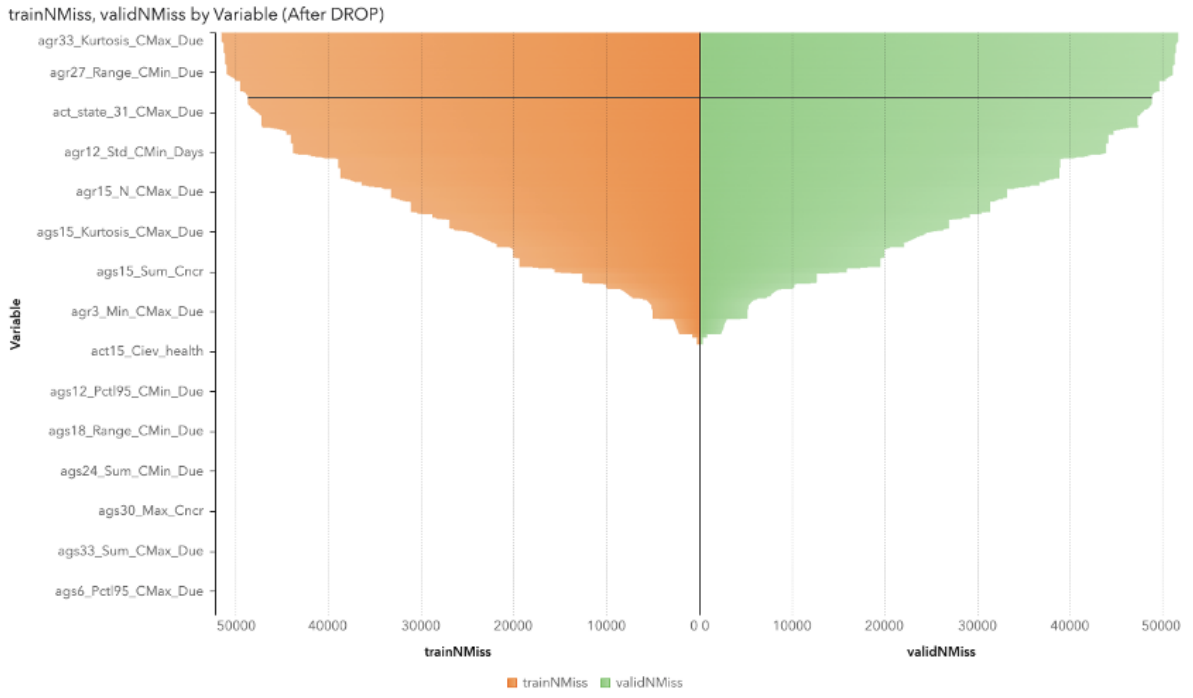




Ranking of variables by the number of missing data
Are there variables that are similar to each other and perhaps it is enough to choose a representative one only?

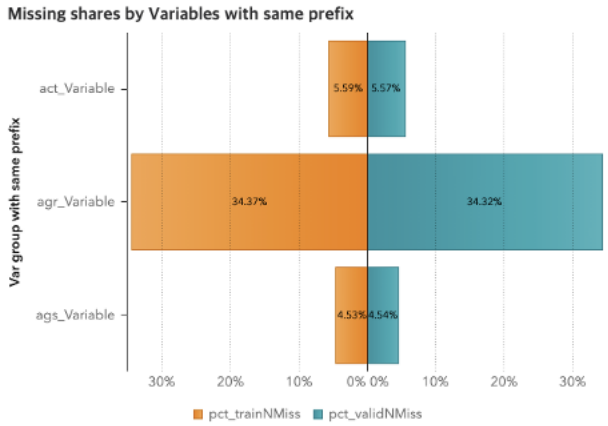
After all those variable dropping and keeping processes, we drew plots showing ‘which variables havethe largest and smallest shares of missing values’. One is the rank from original data and another one is after DROP. Since we still have too many variables, it doesn’t show all variables who have 0 missing values but we can clearly see that some of agr_ variables have largest shares of missing value.

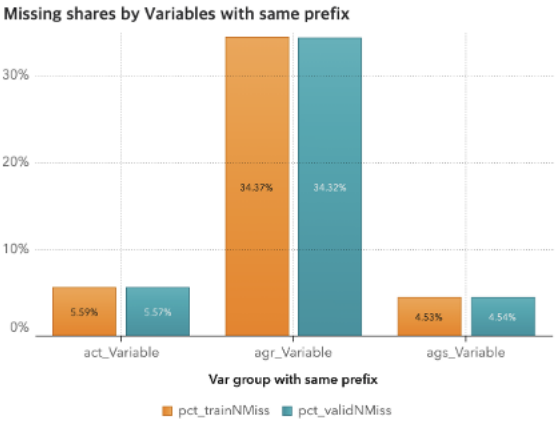




Based on these results, we decided to compare the missing shares grouped by Variables with same prefix of their names. Therefore, we categorized our variables into 3 groups: 1) agr_ ;2) ags_ ;3) act_. As you can see in the plot below, for the original dataset, Variables with prefix 'agr_' has the largest shares of missing data in our dataset. However, for the new dataset that we dropped all unnecessary variables based on our logic, we can see variables with prefix 'ags_' has the largest shares of missing data. I think it is because we already dropped all the variables with 97.5%-100% missing percentage and that may contained large amount of agr_ variables. However, we decide not to drop all these ags_ group variables even though it contains similar information with agr_ group variables. It is because of the following reasons: 1) Deleting all these ags_ variables may affect analysis result a lot; 2) We also thought of imputation, but without good background knowledge in credit loan

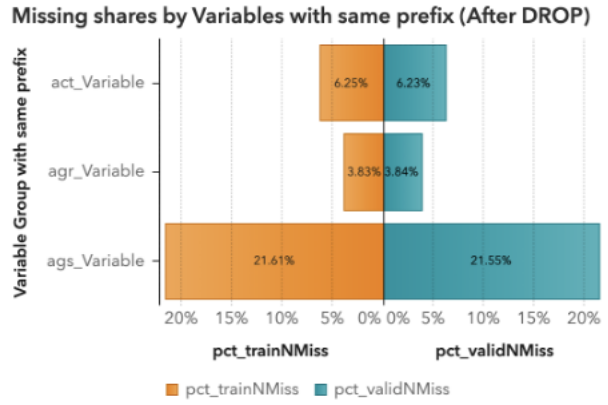
field, we have limitation to proceed imputation. Because it may cause a distortion of analysis results to impute all missing values with 'mean(average)'. For these realistic reasons, we decided to just keep it in our dataset.

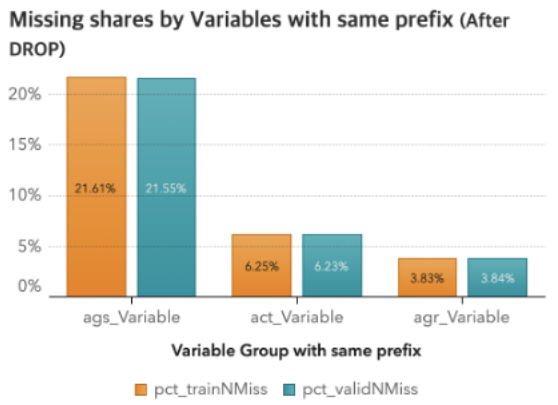




The data-driven content object enables you to incorporate your own content, like a calendar object, into a SAS Visual Analytics report.

	Var group with same prefix	Count of Variable	pct_trainNMiss	pct_validNMiss	Sum of trainNMiss	Sum of validNMiss
☐	act_Variable	280	0.05587941556263153	0.0557359726006219	6808982	6820936
☐	agr_Variable	960	0.34367187868404836	0.3431911509304424	41876881	41999534
☐	ags_Variable	1056	0.04534733658173953	0.04541663949706576 5	5525634	5558062





The data-driven content object enables you to incorporate your own content, like a calendar object, into a SAS Visual Analytics report.

	Variable Group with same prefix	Count of Variable	pct_trainNMiss	pct_validNMiss	Sum of trainNMiss	Sum of validNMiss
☐	act_Variable	268	0.06249313777379859	0.06231653342091166	6178416	6187656
☐	agr_Variable	1023	0.03828223777652856	0.03836638821068389	3784793	3809551
☐	ags_Variable	570	0.21605359426099563	0.21551908942708203	21360249	21399746

SAS Project 3&4

(Konsta, Michał)

TASKS:

3: "Analysis of untypical values, outliers or, in general, irregularities in distributions, their shares in time (i.e. stability in time) and comparison of shares between train and valid datasets (i.e. dataset stability) in the form of a detailed tabular report."

4: "Analysis of untypical data in the form of visualized graphical reports in order to quickly identify variables with greater and lesser and less untypical and irregular data and with their different stability."

The dataset we are using in this project is very large (More than 2300 variables!). It contains 52 841 observations in training data and 53 070 observations in validation data. Therefore it is very likely that the data contains some irregularities (untypical values, outliers…). In this chapter we are going to take a closer look on those values and perform an analysis to get a better understanding of our data.

To get a good overview of this dataset and the properties of its variables, it is good to create a table, which shows the amount of observations, mean, minimum, median, first quartile, interquartile range, third quartile and maximum.

Variable	N	Mean	Minimum	Median	Lower Quartile	Quartile Range	Upper Quartile	Maximum
act_age	52841	58.371908	18.000000	58.000000	51.000000	15.000000	66.000000	93.000000
app_income	52841	2092.796219	300.000000	1644.000000	873.000000	1774.000000	2647.000000	16373
app_number_of_children	52841	1.072690	0	1.000000	0	2.000000	2.000000	3.000000
app_spendings	52841	564.415132	0	380.000000	200.000000	500.000000	700.000000	7720.000000
act_cus_seniority	52841	47.047274	1.000000	17.000000	8.000000	73.000000	81.000000	223.000000

One common way to detect outliers is by using interquartile range (IQR). This is calculated by subtracting Q1 (first quartile) from Q3 (third quartile). To find mild outliers, IQR is multiplied by 1,5. This multiplication is subtracted from Q1 and values below that are mild outliers. The same is done on the other end by adding that value to Q3 and classifying values after that the same way. To find extreme outliers, the exact same operation is done, but this time IQR is multiplied by 3.

The following graphs show the amount/share of variables that contain at least one mild/extreme outlier. We extracted binary variables from this analysis because they can't really contain outliers.

TRAIN:

Number of Variables with Mild Outliers	Number of Variables without Mild Outliers	Number of All Non-binary Variables
1577	311	1888
83,53 %	16,47 %	100,00 %

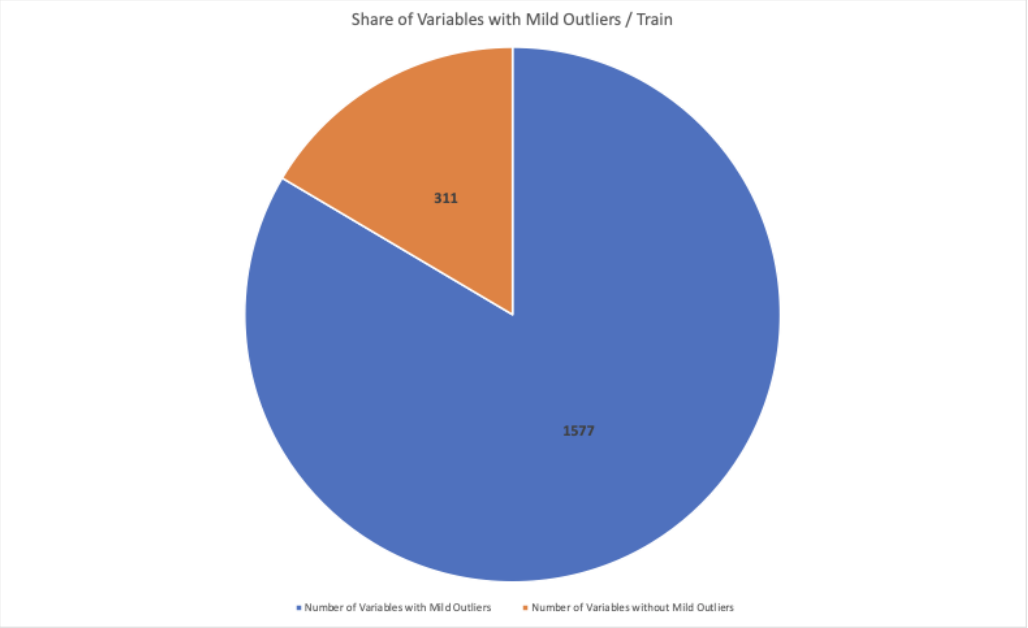
Number of Variables with Extreme Outlier	Number of Variables without Extreme Outliers	Number of All Non-binary Variables
1230	658	1888
65,15 %	34,85 %	100,00 %

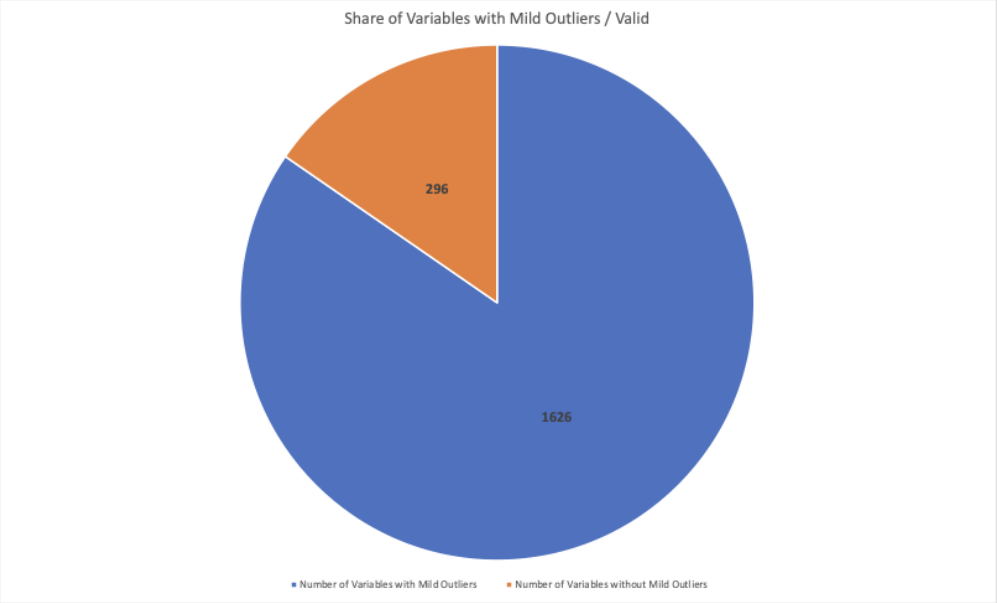
VALIDATION:

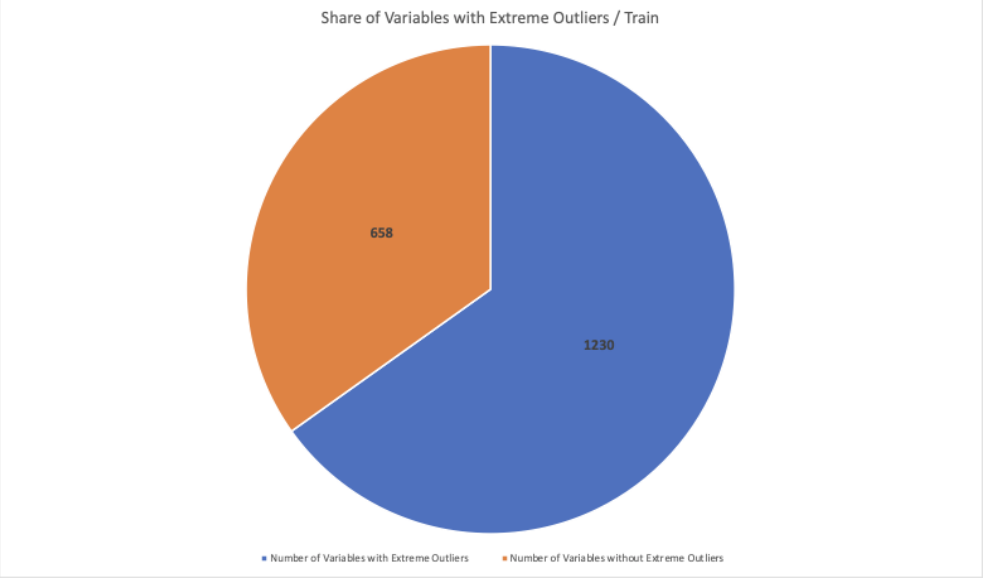
Number of Variables with Mild Outliers	Number of Variables without Mild Outliers	Number of All Non-binary Variables
1626	296	1922
84,60 %	15,40 %	100,00 %

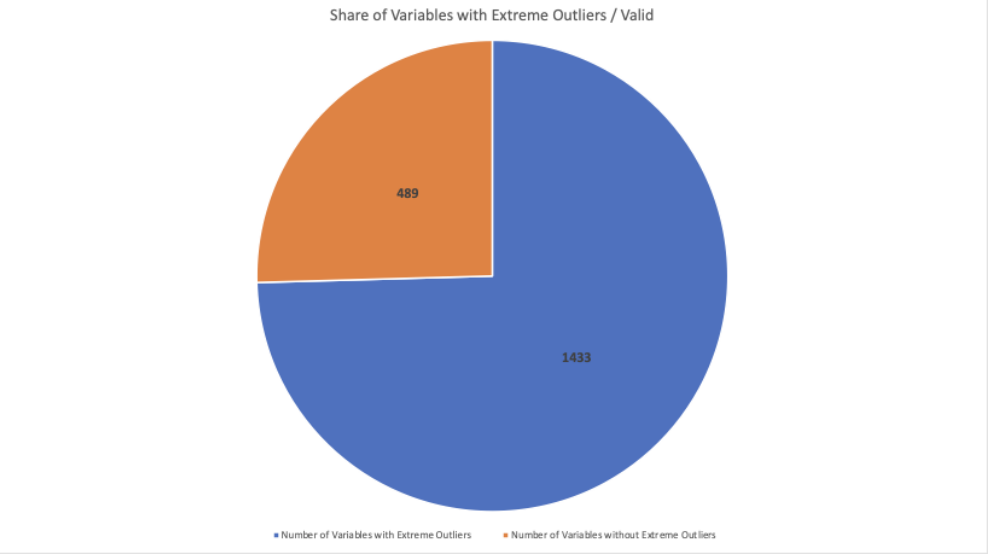
Number of Variables with Extreme Outliers	Number of Variables without Extreme Outliers	Number of All Non-binary Variables
1433	489	1922
74,56 %	25,44 %	100,00 %

We've also decided to show the same statistics in the pie chart so that the high percentage of outliers was even better visualised.




Share of Variables with Mild Outliers / Valid
296
1626
Number of Variables with Mild Outliers
Number of Variables without Mild Outliers


Share of Variables with Extreme Outliers / Train
658
1230
Number of Variables with Extreme Outliers
Number of Variables without Extreme Outliers


Share of Variables with Extreme Outliers / Valid
489
1433
Number of Variables with Extreme Outliers
Number of Variables without Extreme Outliers

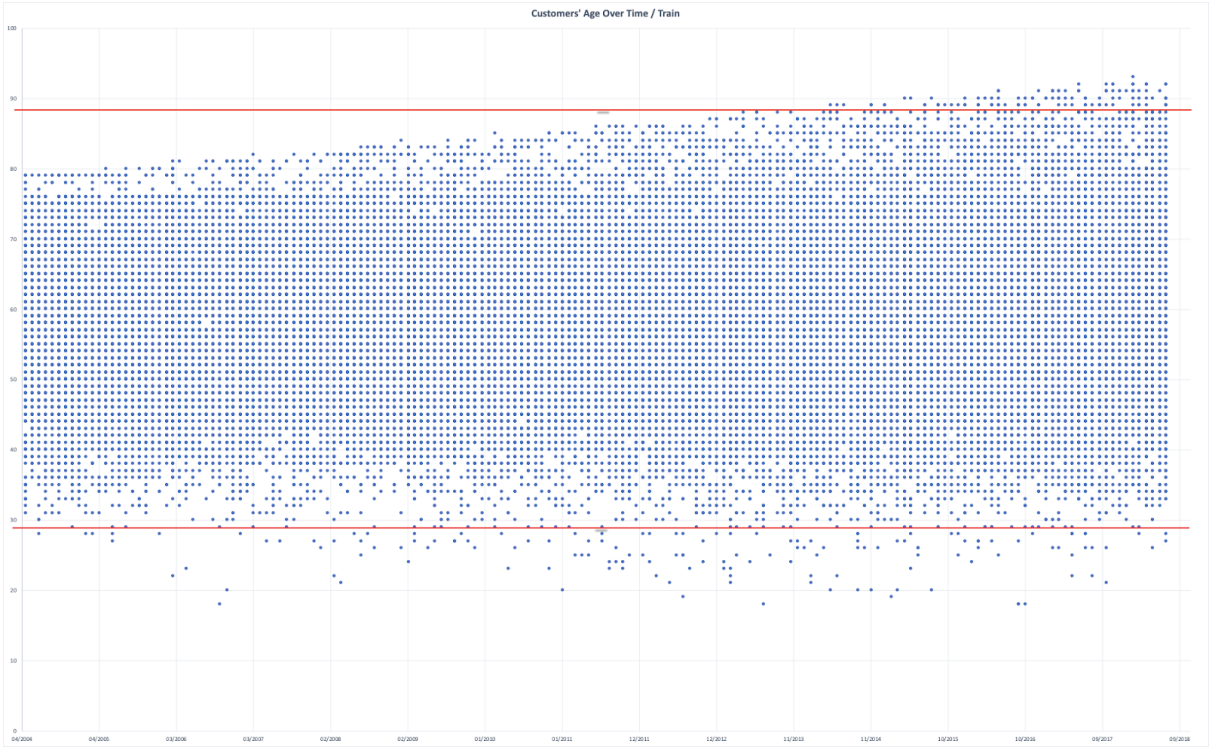
As we can see from these tables, validation data has more outliers. The difference of extreme outliers (74,56 % vs. 65,15%) between train and validation data sets is actually quite big.

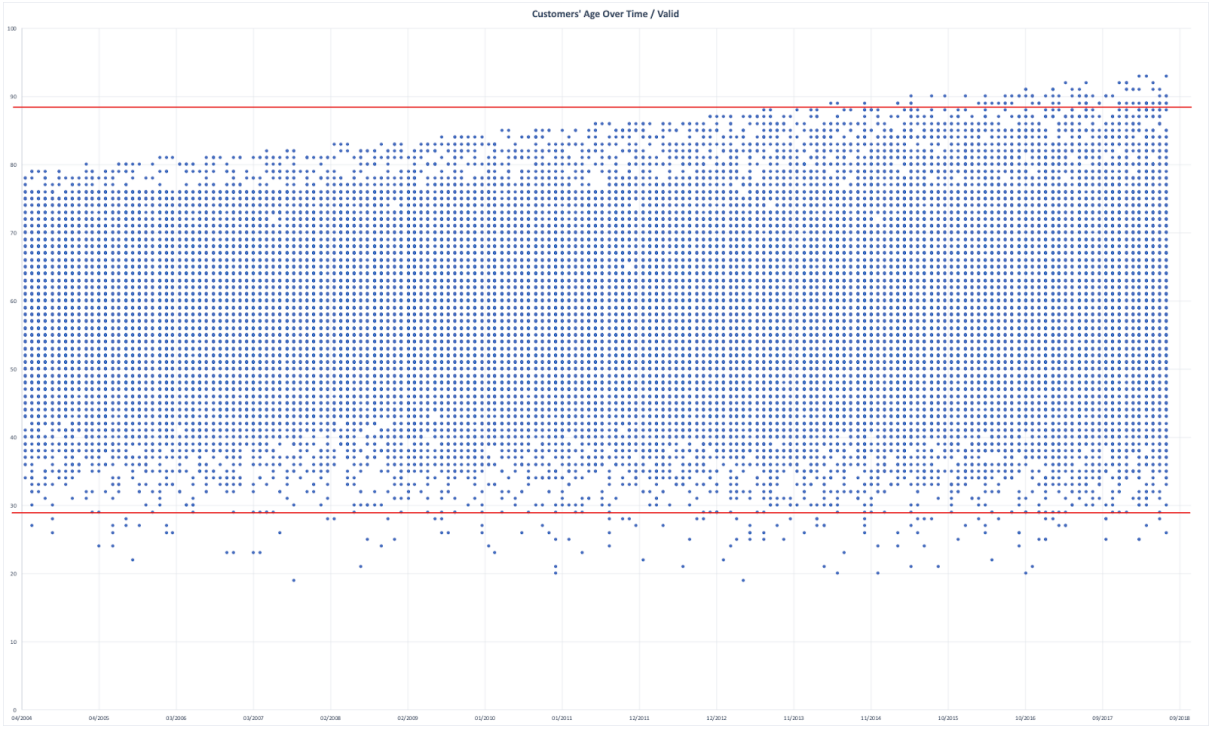
We chose a few important variables to investigate them closer. They are **age**, **income** and **spendings** *(52841 observations train, 53071 observations valid)*. The following summarisations show how many of their observations could be classified as outliers. The thresholds are the same as we calculated before using IQR.

TRAIN_DATA	Age	Income	Spendings
Mild Outlier Threshold (low)	28,5	-1788,0	-550,0
Mild Outlier Threshold (high)	88,5	5308,0	1450,0
Mild Outlier Count	**267**	**3058**	**3708**
Extreme Outlier Threshold(low)	6,0	-4449,0	-1300,0
Extreme Outlier Threshold(high)	111,0	7969,0	2200,0
Extreme Outlier Count	**0**	**911**	**1537**

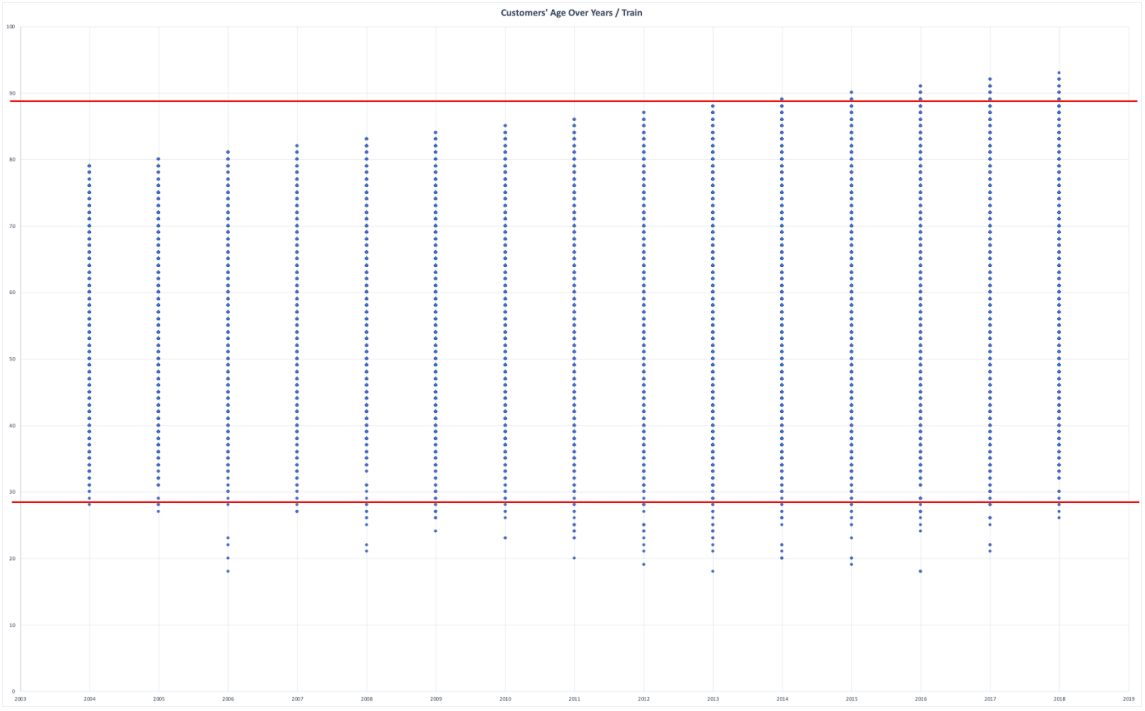
VALID_DATA	Age	Income	Spendings
Mild Outlier Threshold (low)	28,5	-1804,0	-550,0
Mild Outlier Threshold (high)	88,5	5316,0	1450,0
Mild Outlier Count	**213**	**3077**	**3734**
Extreme Outlier Threshold(low)	6,0	-4474,0	-1300,0
Extreme Outlier Threshold(high)	111,0	7986,0	2200,0
Extreme Outlier Count	**0**	**930**	**1486**

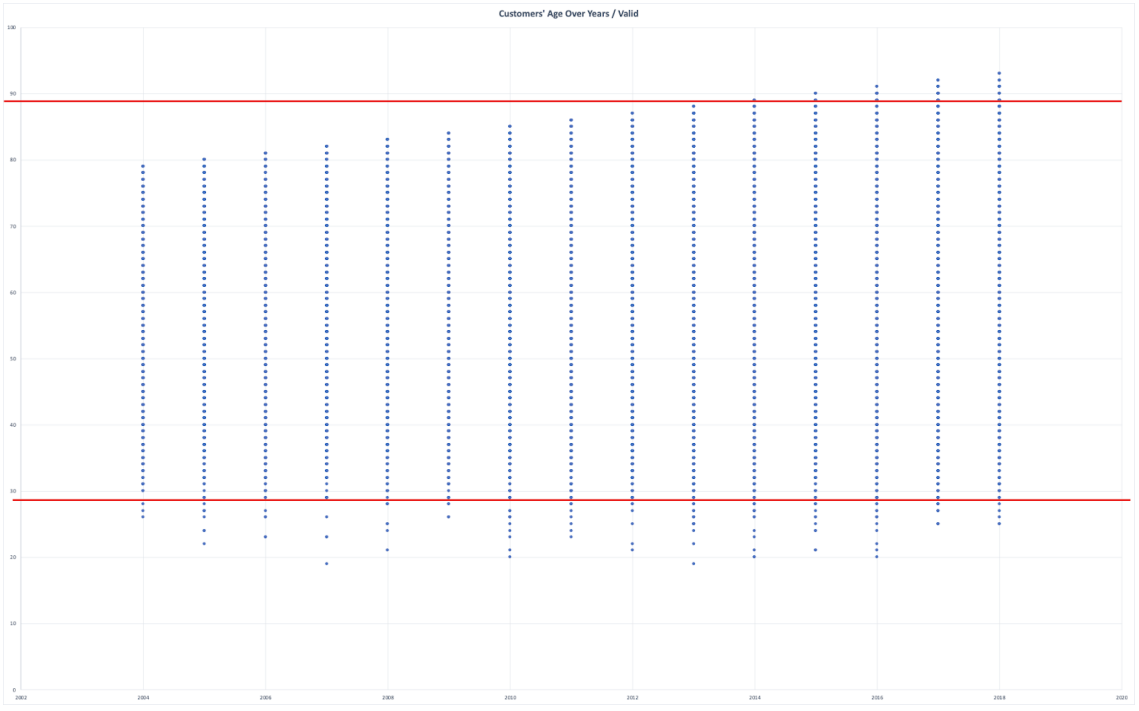
That’s how the scatter plot representing the age of the consumers looks like. The red lines show the mild outliers boundaries.





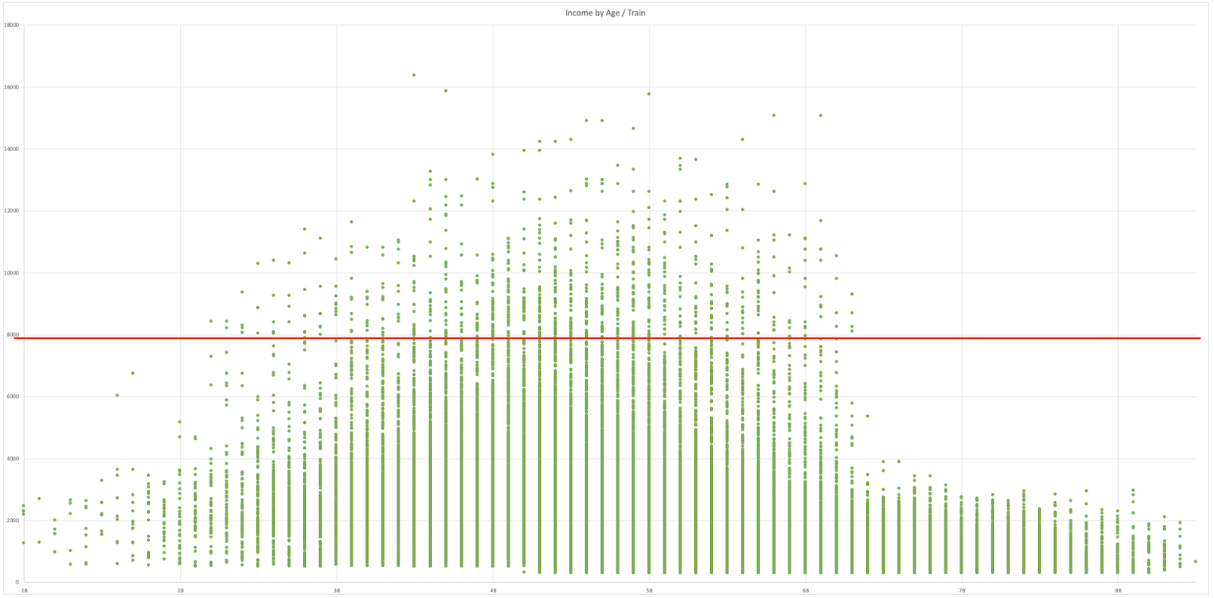
We also grouped them in the time series to see how age is changing over years and what is its time stability.

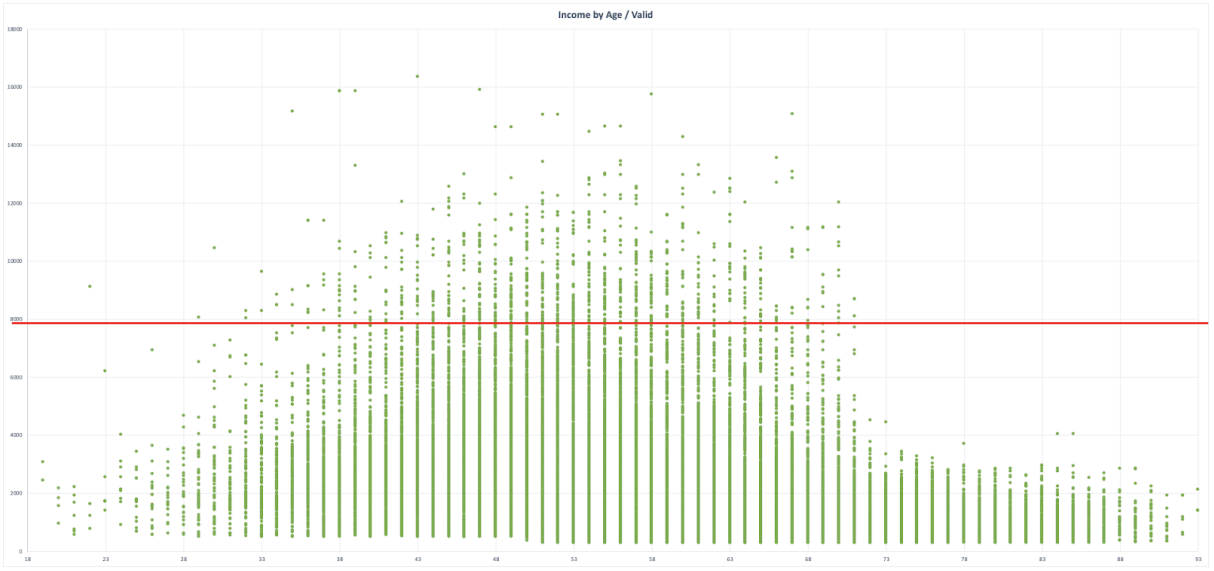




The conclusions were the following: Society is ageing and more old people are willing to take the credits. What's more, they are also very good consumers because they have certain incomes (pensions) and people (children) who will take care of their debts if needed so they should not be excluded from the model. But those that are younger than 28.5 were both unstable in time and had more difficulties with due payments so we recommend to exclude from the model but it is not necessary.

We also checked how the upper extreme outlier in income is influencing the model regarding the groups of age.

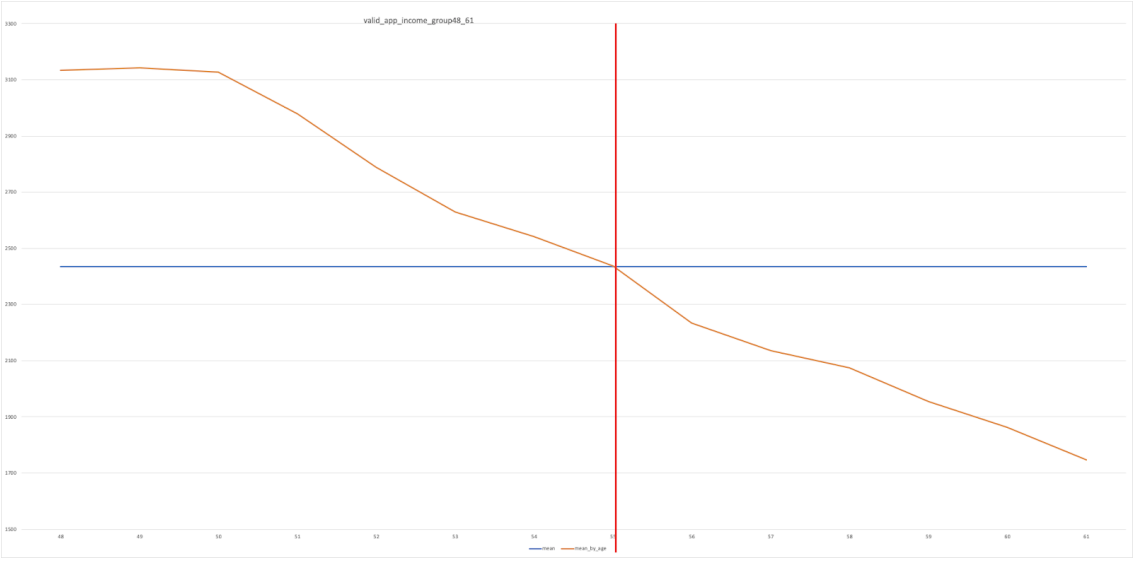


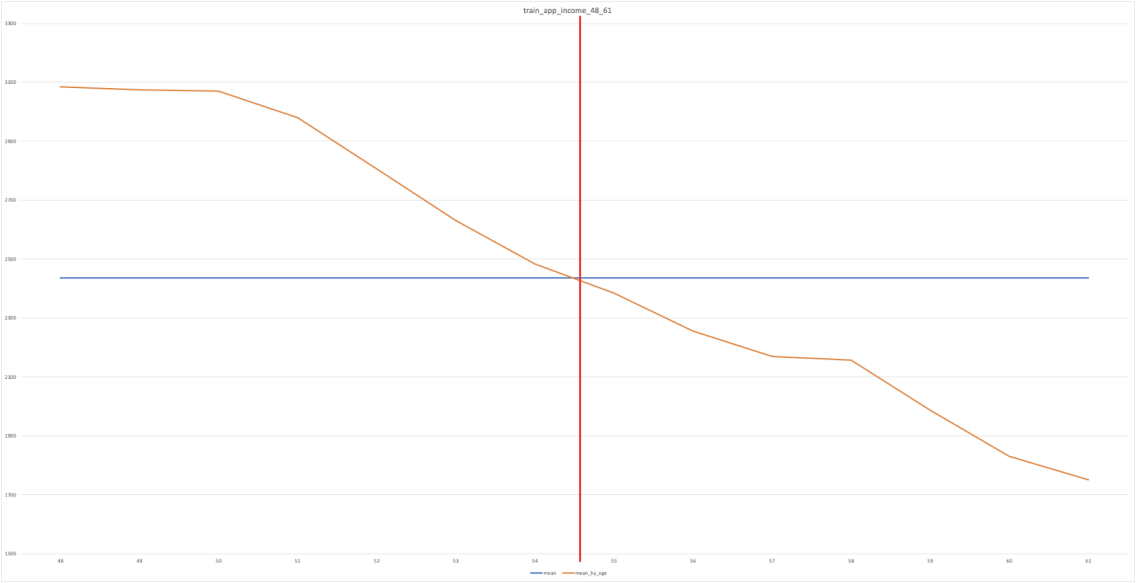


As we see that the mean for the middle-aged people is much higher than for the rest, we decided to group them. We saw two important characteristics for the groups 34 - 47 and 48 - 61:


valid_app_income_group34_47
3500
3400
3300
3200
3100
3000
2900
2800
2700
2600
2500
34
35
36
37
38
39
40
41
42
43
44
45
46
47
mean
mean_by_age


train_app_income_group34_47
3200
3100
3000
2900
2800
2700
2600
2500
34
35
36
37
38
39
40
41
42
43
44
45
46
47
mean
mean_by_age

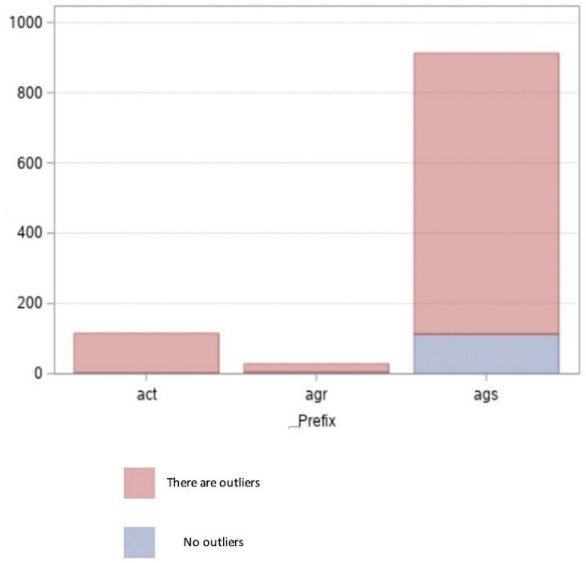

valid_app_income_group48_61
3300
3100
2900
2700
2500
2300
2100
1900
1700
1500
48
49
50
51
52
53
54
55
56
57
58
59
60
61
mean
mean_by_age

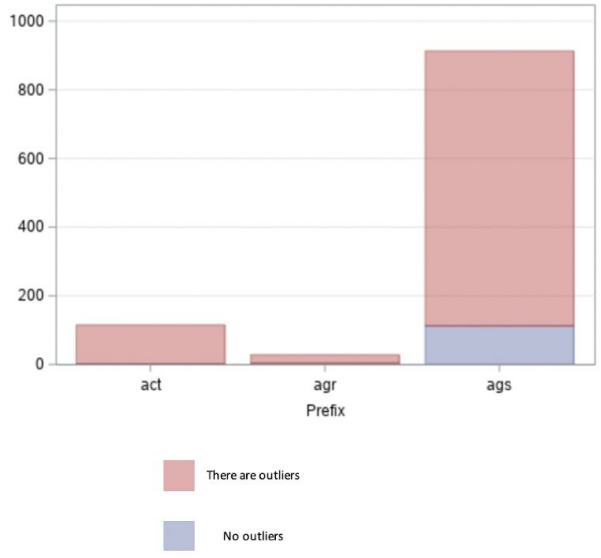


We could see that for these grouped variables, the mean for the people aged 41 - 55 was higher than for the rest. So we recommend not to exclude all the extreme variables of the people of this age, but of those that are between 18 - 40 and 56 - 100.

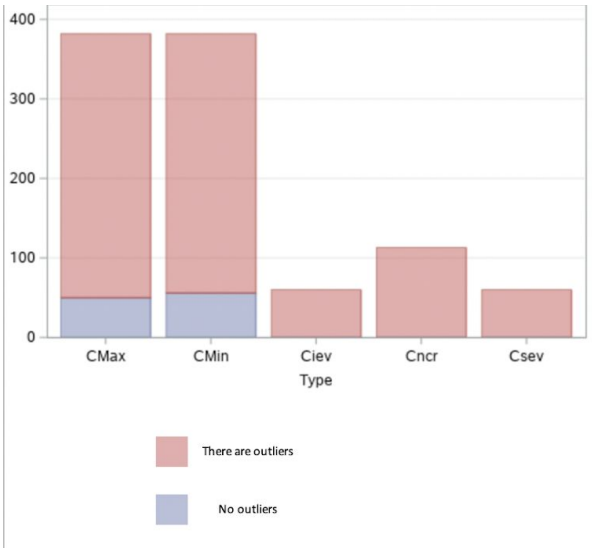
The same was done for the “spending” variable and the same conclusions were made for its stability in time and over different groups of age. Here the mean was increased for the people between 44 and 55, so the lower boundary was moved.

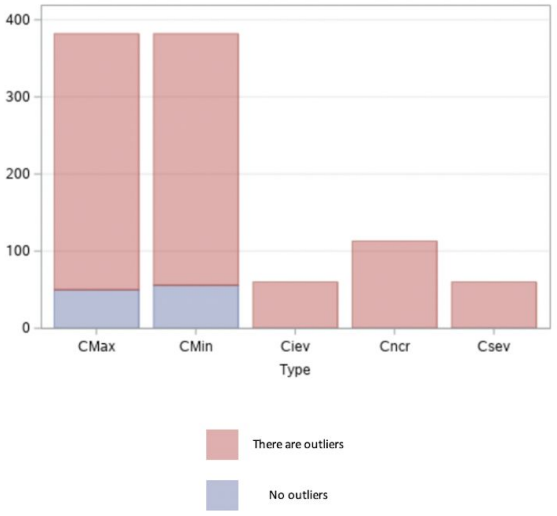
Finally, we also grouped the ‘act.’, ‘agr.’ and ‘ags.’ outliers by their prefix to check how many of these possess mild outliers.





The same was done for the types such as ‘CMax’, ‘CMin’, ‘Ciev’, ‘Cncr’ and ‘Csev’:





The graphical visualisation let us quickly conclude that the number of variables with mild outliers in these datasets is dominating. However, as we also checked that the ratio of missing data in these variables is enormous, we decided to recommend leaving them in their first stage.

SAS Project 5

(Yuqing&Likala)

TASKS:

Analysis of the relationship between the defaut_cus12 target function and the customer's characteristics. Identification of the variables influencing the objective function, their interdependencies, correlations, dependencies, etc. In this case, the final report may consist of both tabular and graphical reports visualizing the goodness of the predictors.

My solution

>Step0

Considering there is too much missing data, what we need to do is to clean the dataset. Hence we only choose some variables whose missing value is lower than 50% since we think of variables with higher than 50% missing value is meaningless.

Virables	N	N Missing	Count	Missing Ratio
act_state_12_CMax_Due	26427	26414	52841	49.99%
act_state_12_CMin_Due	26427	26414	52841	49.99%
ags12_Skewness_CMin_Due	26761	26080	52841	49.36%
ags12_Kurtosis_CMax_Due	27239	25602	52841	48.45%
ags15_Kurtosis_CMin_Due	27356	25485	52841	48.23%
ags12_Skewness_CMax_Due	27667	25174	52841	47.64%
ags15_Skewness_CMin_Due	27773	25068	52841	47.44%
ags18_Kurtosis_CMin_Due	27967	24874	52841	47.07%
ags15_Kurtosis_CMax_Due	28176	24665	52841	46.68%
act_state_10_CMax_Days	28377	24464	52841	46.30%
act_state_10_CMin_Days	28377	24464	52841	46.30%
ags18_Skewness_CMin_Due	28380	24461	52841	46.29%
ags21_Kurtosis_CMin_Due	28394	24447	52841	46.27%
act_state_11_CMax_Due	28551	24290	52841	45.97%
act_state_11_CMin_Due	28551	24290	52841	45.97%
ags15_Skewness_CMax_Due	28598	24243	52841	45.88%
ags24_Kurtosis_CMin_Due	28716	24125	52841	45.66%
ags18_Kurtosis_CMax_Due	28721	24120	52841	45.65%
ags21_Skewness_CMin_Due	28793	24048	52841	45.51%
ags27_Kurtosis_CMin_Due	29014	23827	52841	45.09%
ags21_Kurtosis_CMax_Due	29093	23748	52841	44.94%
ags24_Skewness_CMin_Due	29118	23723	52841	44.90%
ags18_Skewness_CMax_Due	29139	23702	52841	44.86%
ags30_Kurtosis_CMin_Due	29295	23546	52841	44.56%

Now, we get a new dataset.

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
default_cus12	46388	0.33151	0.47076	15378	0	1.00000
act_age	52841	58.37191	11.13976	3084430	18.00000	93.00000
app_income	52841	2093	1755	110585445	300.00000	16373
app_number_of_children	52841	1.07269	0.98276	56682	0	3.00000
app_spendings	52841	564.41513	620.86767	29824260	0	7720
act_cus_seniority	52841	47.04727	57.30013	2486025	1.00000	223.00000
act_cus_n_loans_hist	52841	1.92854	1.96576	101906	1.00000	49.00000
act_cus_n_statC	52841	0.40363	0.90792	21328	0	11.00000
act_cus_n_statB	52841	0.48108	1.38701	25421	0	46.00000
act_cus_n_loans_act	52841	1.04563	0.23358	55252	1.00000	5.00000
act_cus_pins	52841	10.41718	7.21061	550454	0	61.00000
act_cus_utl	52841	0.41610	0.27882	21987	0	0.97917
act_cus_dueutl	52841	0.01873	0.03127	989.90625	0	0.12500
act_cus_cc	52841	0.52385	0.33333	27681	0.01432	5.03000
act_state_1_CMax_Days	45552	12.81004	2.35230	583523	-1.00000	26.00000
act_state_2_CMax_Days	44287	12.80285	2.34550	567000	-1.00000	26.00000
act_state_3_CMax_Days	42908	12.78920	2.32373	548759	-1.00000	25.00000
act_state_4_CMax_Days	40968	12.77641	2.31327	523424	-1.00000	25.00000

(The picture shows part of data…)

>Step1

Firstly, all the customer's details variables (numerical) and defaut_cus12 variables for correlation analysis, take out Pearson correlation coefficient absolute value greater than 0.3 variables. The specific variables can be found in the table below:

Variables	Count
act_	8
agr_	52
ags_	54
Total	114

Considering that there are still many variables, so we need to narrow the range of correlation coefficients. Therefore, we choose variables whose absolute value is greater than 0.4.

The specific variables can be found in table below:

Variables	Pearson	r	N
act_cus_dueutl	0.46059	<.0001	46388
act_state_1_CMax_Due	0.47192	<.0001	46388
act_CMax_Due	0.47192	<.0001	46388
act_state_1_CMin_Due	0.43899	<.0001	46388
act_CMin_Due	0.43899	<.0001	46388
agr3_Pctl75_CMax_Due	0.41645	<.0001	42229
ags3_Pctl75_CMax_Due	0.40744	<.0001	46388
agr3_Pctl95_CMax_Due	0.41645	<.0001	42229
ags3_Pctl95_CMax_Due	0.40744	<.0001	46388
agr3_Mean_CMax_Due	0.43035	<.0001	42229
ags3_Mean_CMax_Due	0.4203	<.0001	46388
agr3_Max_CMax_Due	0.41645	<.0001	42229
ags3_Max_CMax_Due	0.40744	<.0001	46388
agr3_Sum_CMax_Due	0.43035	<.0001	42229
ags3_Sum_CMax_Due	0.4198	<.0001	46388
agr3_Pctl75_CMin_Due	0.41188	<.0001	42229
ags3_Pctl75_CMin_Due	0.40307	<.0001	46388
agr3_Pctl95_CMin_Due	0.41188	<.0001	42229
ags3_Pctl95_CMin_Due	0.40307	<.0001	46388
agr3_Mean_CMin_Due	0.41242	<.0001	42229
ags3_Mean_CMin_Due	0.40346	<.0001	46388
agr3_Max_CMin_Due	0.41188	<.0001	42229
ags3_Max_CMin_Due	0.40307	<.0001	46388
agr3_Sum_CMin_Due	0.41242	<.0001	42229
ags3_Sum_CMin_Due	0.40298	<.0001	46388
ags3_n_cus_arrears	0.40878	<.0001	46388

*Further consideration of the correlation between the same type of variables, if the correlation is large, you can further reduce the variables:

>Step2

We choose the most typical variable according to Pearson in three groups.

In act_ group, the variable act_cMax_Due has the strongest correlation.

act_	r	n	n
act_cus_dueutl	0.46059	<.0001	46388
act_state_1_CMax_Due	0.47192	<.0001	46388
act_CMax_Due	0.47192	<.0001	46388
act_state_1_CMin_Due	0.43899	<.0001	46388
act_CMin_Due	0.43899	<.0001	46388

The SAS System

The CORR Procedure

5 Variables: act_CMax_Due act_state_1_CMax_Due act_cus_dueutl act_CMin_Due act_state_1_CMin_Due

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
act_CMax_Due	52841	0.46534	0.77106	24589	0	3.00000
act_state_1_CMax_Due	52841	0.46534	0.77106	24589	0	3.00000
act_cus_dueutl	52841	0.01873	0.03127	989.90625	0	0.12500
act_CMin_Due	52841	0.43444	0.74771	22956	0	3.00000
act_state_1_CMin_Due	52841	0.43444	0.74771	22956	0	3.00000

In agr_ group, the variable agr3_Mean_CMax_Due has the strongest correlation.

agr_	r	n	n
agr3_Pctl75_CMax_Due	0.41645	<.0001	42229
agr3_Pctl95_CMax_Due	0.41645	<.0001	42229
agr3_Mean_CMax_Due	0.43035	<.0001	42229
agr3_Max_CMax_Due	0.41645	<.0001	42229
agr3_Sum_CMax_Due	0.43035	<.0001	42229
agr3_Pctl75_CMin_Due	0.41188	<.0001	42229
agr3_Pctl95_CMin_Due	0.41188	<.0001	42229
agr3_Mean_CMin_Due	0.41242	<.0001	42229
agr3_Max_CMin_Due	0.41188	<.0001	42229
agr3_Sum_CMin_Due	0.41242	<.0001	42229

The SAS System

The CORR Procedure

10 Variables:	agr3_Pctl75_CMax_Due agr3_Pctl95_CMax_Due agr3_Mean_CMax_Due agr3_Max_CMax_Due agr3_Sum_CMax_Due agr3_Pctl75_CMin_Due agr3_Pctl95_CMin_Due agr3_Mean_CMin_Due agr3_Max_CMin_Due agr3_Sum_CMin_Due

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
agr3_Pctl75_CMax_Due	47768	0.65318	0.96062	31201	0	7.00000
agr3_Pctl95_CMax_Due	47768	0.65318	0.96062	31201	0	7.00000
agr3_Mean_CMax_Due	47768	0.43789	0.67180	20917	0	5.66667
agr3_Max_CMax_Due	47768	0.65318	0.96062	31201	0	7.00000
agr3_Sum_CMax_Due	47768	1.31368	2.01541	62752	0	17.00000
agr3_Pctl75_CMin_Due	47768	0.59102	0.84611	28232	0	7.00000
agr3_Pctl95_CMin_Due	47768	0.59102	0.84611	28232	0	7.00000
agr3_Mean_CMin_Due	47768	0.39644	0.61015	18937	0	4.33333
agr3_Max_CMin_Due	47768	0.59102	0.84611	28232	0	7.00000
agr3_Sum_CMin_Due	47768	1.18931	1.83046	56811	0	13.00000

In ags_ group, the variable agr3_Mean_CMax_Due has the strongest correlation.

ags_	r	n	n
ags3_Pctl75_CMax_Due	0.40744	<.0001	46388
ags3_Pctl95_CMax_Due	0.40744	<.0001	46388
ags3_Mean_CMax_Due	0.4203	<.0001	46388
ags3_Max_CMax_Due	0.40744	<.0001	46388
ags3_Sum_CMax_Due	0.4198	<.0001	46388
ags3_Pctl75_CMin_Due	0.40307	<.0001	46388
ags3_Pctl95_CMin_Due	0.40307	<.0001	46388
ags3_Mean_CMin_Due	0.40346	<.0001	46388
ags3_Max_CMin_Due	0.40307	<.0001	46388
ags3_Sum_CMin_Due	0.40298	<.0001	46388
ags3_n_cus_arrears	0.40878	<.0001	46388

The SAS System

The CORR Procedure

11 Variables:	ags3_Pctl75_CMax_Due ags3_Pctl95_CMax_Due ags3_Mean_CMax_Due ags3_Max_CMax_Due ags3_Sum_CMax_Due ags3_Pctl75_CMin_Due ags3_Pctl95_CMin_Due ags3_Mean_CMin_Due ags3_Max_CMin_Due ags3_Sum_CMin_Due ags3_n_cus_arrears

Simple Statistics

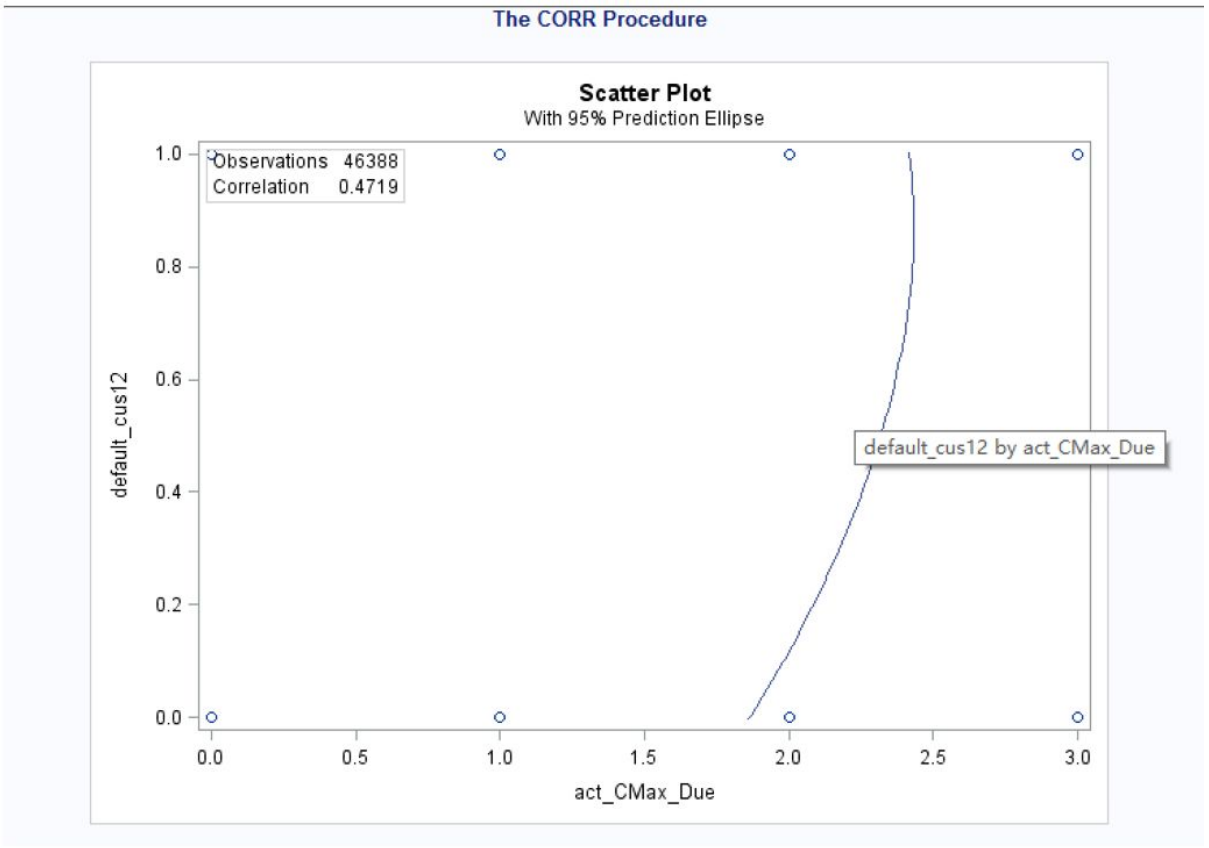
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
ags3_Pctl75_CMax_Due	52841	0.59628	0.93537	31508	0	7.00000
ags3_Pctl95_CMax_Due	52841	0.59628	0.93537	31508	0	7.00000
ags3_Mean_CMax_Due	52841	0.39876	0.65190	21071	0	5.66667
ags3_Max_CMax_Due	52841	0.59628	0.93537	31508	0	7.00000
ags3_Sum_CMax_Due	52841	1.19337	1.95408	63059	0	17.00000
ags3_Pctl75_CMin_Due	52841	0.54004	0.82573	28536	0	7.00000
ags3_Pctl95_CMin_Due	52841	0.54004	0.82573	28536	0	7.00000
ags3_Mean_CMin_Due	52841	0.36125	0.59224	19089	0	4.33333
ags3_Max_CMin_Due	52841	0.54004	0.82573	28536	0	7.00000
ags3_Sum_CMin_Due	52841	1.08088	1.77476	57115	0	13.00000
ags3_n_cus_arrears	52841	0.88831	1.22018	46939	0	3.00000

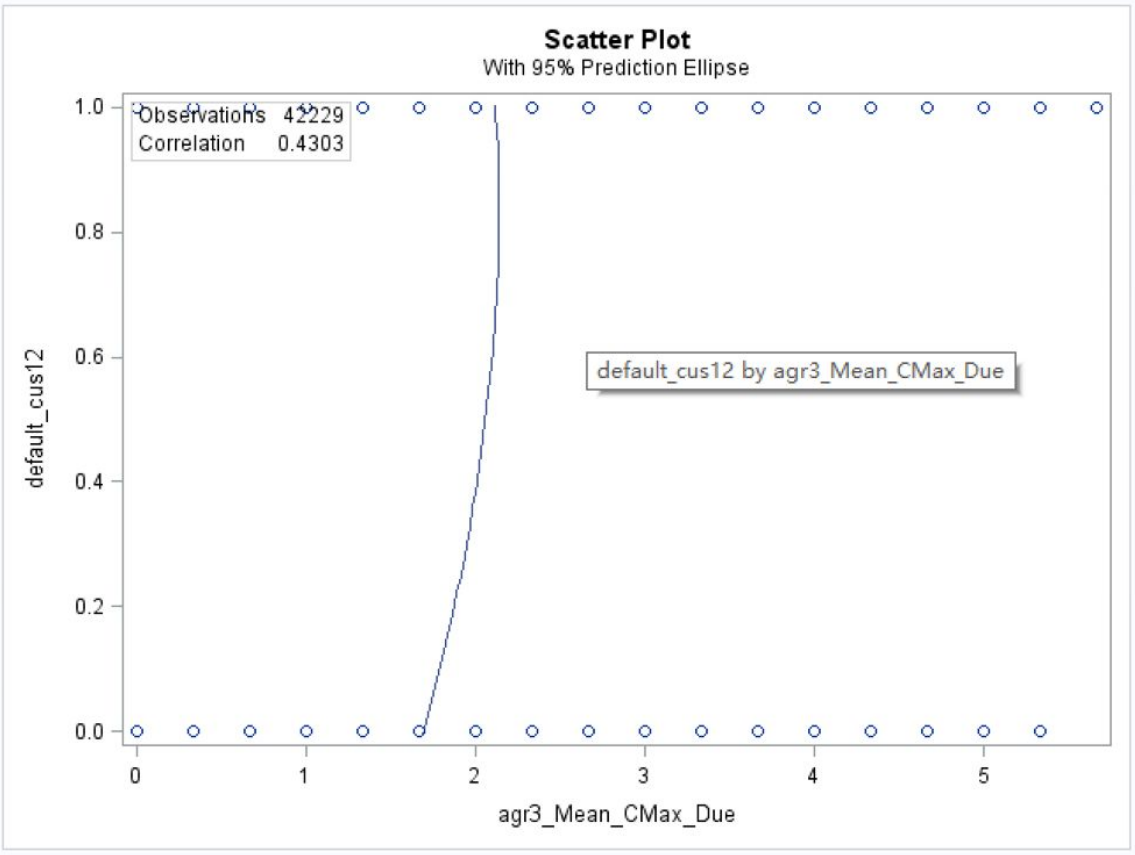
To sum up :

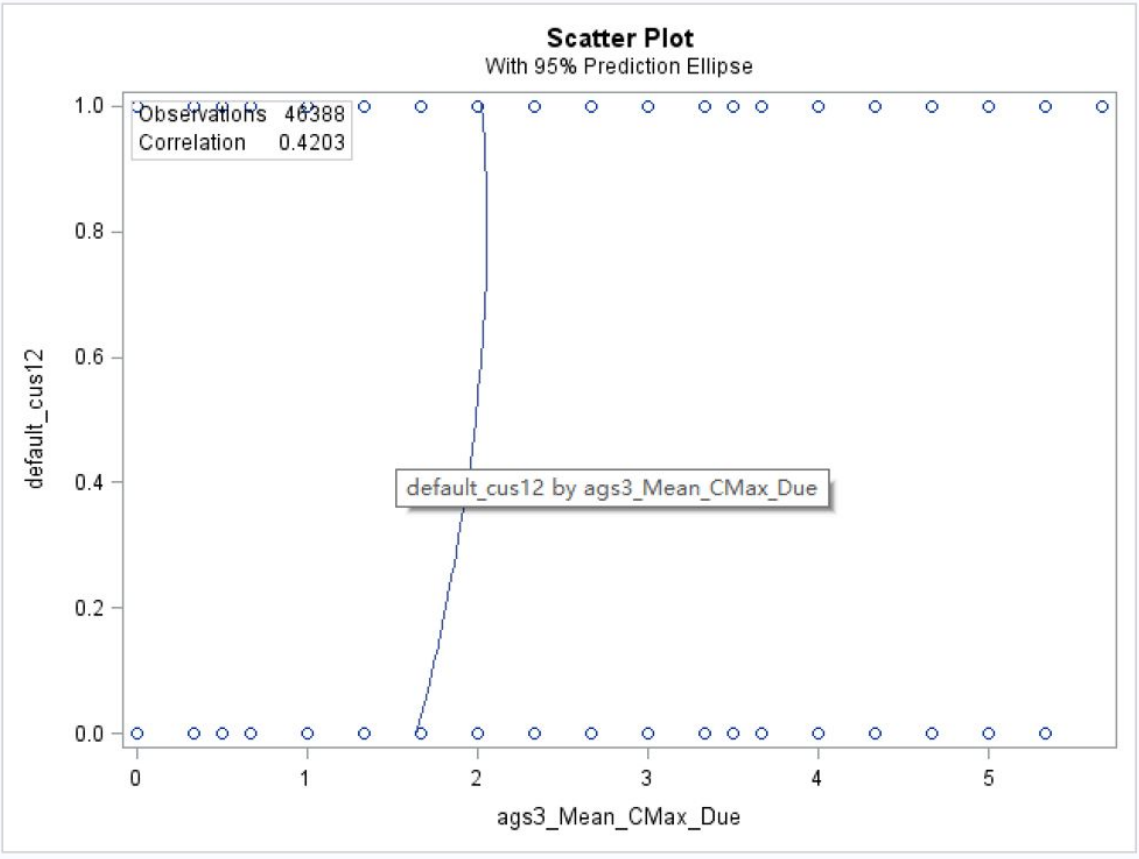
Variables	r	n	n
act_CMax_Due	0.47192	<.0001	46388
agr3_Mean_CMax_Due	0.43035	<.0001	42229
ags3_Mean_CMax_Due	0.4203	<.0001	46388

>Step3

Plot the graphs below:




Scatter Plot
With 95% Prediction Ellipse
Observations 42229
Correlation 0.4303
default_cus12 by agr3_Mean_CMax_Due
default_cus12
agr3_Mean_CMax_Due
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5


Scatter Plot
With 95% Prediction Ellipse
Observations 46388
Correlation 0.4203
default_cus12 by ags3_Mean_CMax_Due
default_cus12
ags3_Mean_CMax_Due
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5