Exploring data one variable at time

a) Nie wszystkie dane się zgadzają. Występują braki danych w kolumnach, np. REASON. Różnice są widoczne w nazwach oraz ilościach kolumn. Spowodowane jest to np. tym, że nie widać sensu nadawania wartościom nominalnym takich kategorii jak minimum czy maksimum.

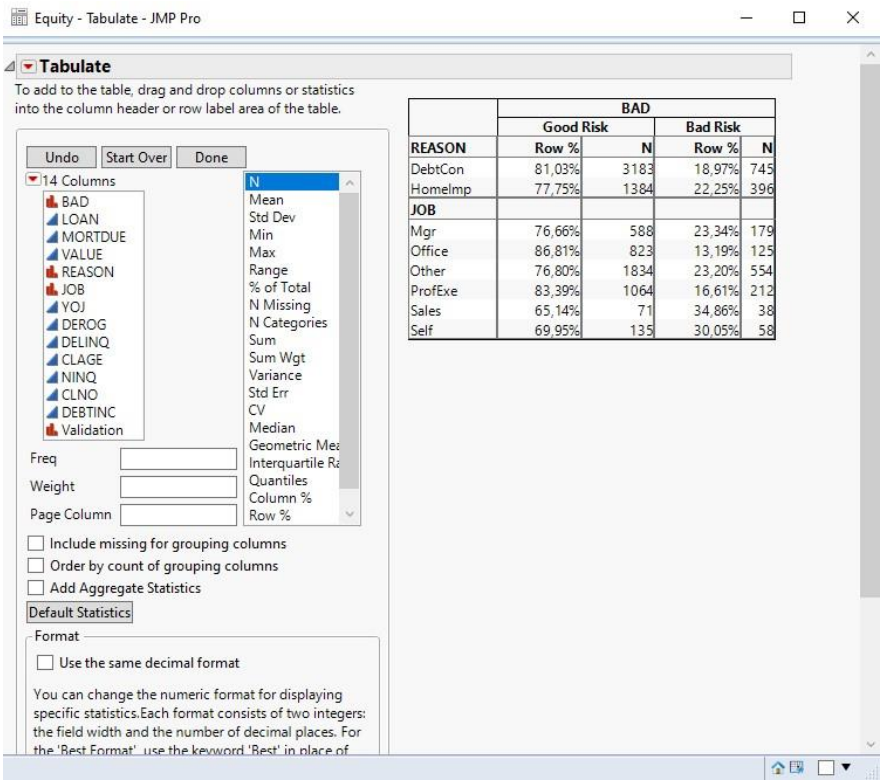
Columns	N	N Missing	N Categories	Min	Max	Mean	Std Dev
BAD	5960	0	2	.	.	.	.
LOAN	5960	0	.	1100	89900	18607,969799	11207,480417
MORTDUE	5442	518	.	2063	399550	73760,8172	44457,609458
VALUE	5848	112	.	8000	855909	101776,04874	57385,775334
REASON	5708	252	2	.	.	.	.
JOB	5681	279	6	.	.	.	.
YOJ	5445	515	.	0	41	8,9222681359	7,5739822489
DEROG	5252	708	.	0	10	0,2545696877	0,8460467771
DELINQ	5380	580	.	0	15	0,4494423792	1,1272659176
CLAGE	5652	308	.	0	1168,2335609	179,76627519	85,810091764
NINQ	5450	510	.	0	17	1,1860550459	1,7286749712
CLNO	5738	222	.	0	71	21,296096201	10,138933192
DEBTINC	4693	1267	.	0,52	203,31	33,780	8,6017461863
Validation	5960	0	3	.	.	.	.

b) Niektóre zmienne posiadają wartości równe 0, np. ktoś, kto nigdy nie pracował nie będzie posiadał dodatniej liczby lat pracy. Duża ilość takich osób wpłynąć może na to, że pierwszy kwartyl również będzie wynosił 0.

Columns	N	N Missing	N Categories	Min	Max	Mean	Std Dev	Median	Lower Quartile	Upper Quartile	Interquartile Range
BAD	5960	0	2	.	.	.	.	.	.	.	.
LOAN	5960	0	.	1100	89900	18607,969799	11207,480417	16300	11100	23300	12200
MORTDUE	5442	518	.	2063	399550	73760,8172	44457,609458	65019	46267,5	91493,75	45226,25
VALUE	5848	112	.	8000	855909	101776,04874	57385,775334	89235,5	66062,5	119838,75	53776,25
REASON	5708	252	2	.	.	.	.	.	.	.	.
JOB	5681	279	6	.	.	.	.	.	.	.	.
YOJ	5445	515	.	0	41	8,9222681359	7,5739822489	7	3	13	10
DEROG	5252	708	.	0	10	0,2545696877	0,8460467771	0	0	0	0
DELINQ	5380	580	.	0	15	0,4494423792	1,1272659176	0	0	0	0
CLAGE	5652	308	.	0	1168,2335609	179,76627519	85,810091764	173,46666667	115,08969143	231,58738906	116,49769763
NINQ	5450	510	.	0	17	1,1860550459	1,7286749712	1	0	2	2
CLNO	5738	222	.	0	71	21,296096201	10,138933192	20	14,75	26	11,25
DEBTINC	4693	1267	.	0,52	203,31	33,780	8,6017461863	34,818	29,137	39,005	9,869
Validation	5960	0	3	.	.	.	.	.	.	.	.

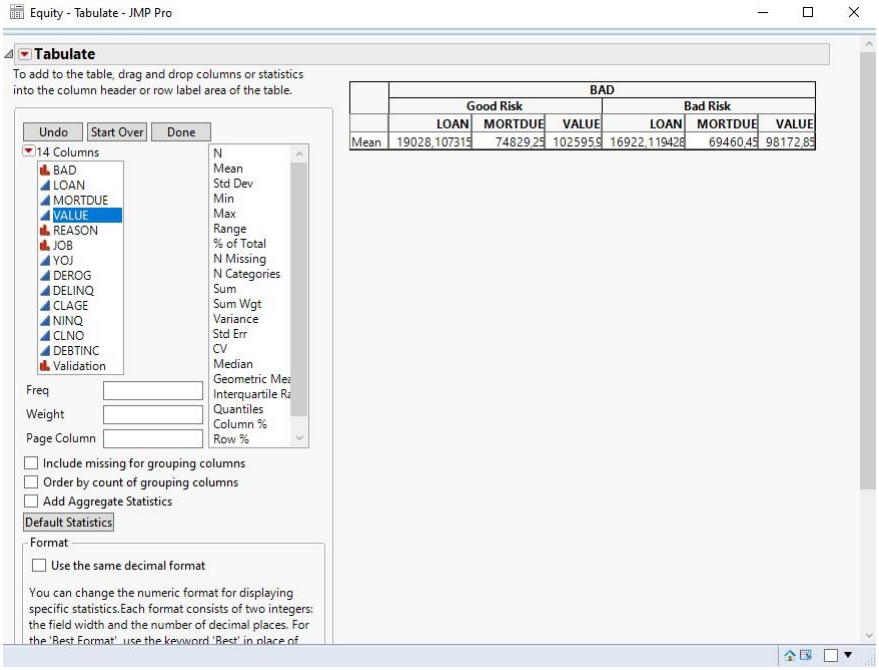
c) Najwiekszą wadą modalnej jest jej niestabilnośd ze względu na dane. Natomiast mediana jest odporna na obserwacje odstające, bo nie zależy od wartości krańcowych.

Exploring data two variables at a time - tabulate a)



	BAD			
	Good Risk		Bad Risk	
REASON	Row %	N	Row %	N
DebtCon	81,03%	3183	18,97%	745
HomeImp	77,75%	1384	22,25%	396
JOB				
Mgr	76,66%	588	23,34%	179
Office	86,81%	823	13,19%	125
Other	76,80%	1834	23,20%	554
ProfExe	83,39%	1064	16,61%	212
Sales	65,14%	71	34,86%	38
Self	69,95%	135	30,05%	58

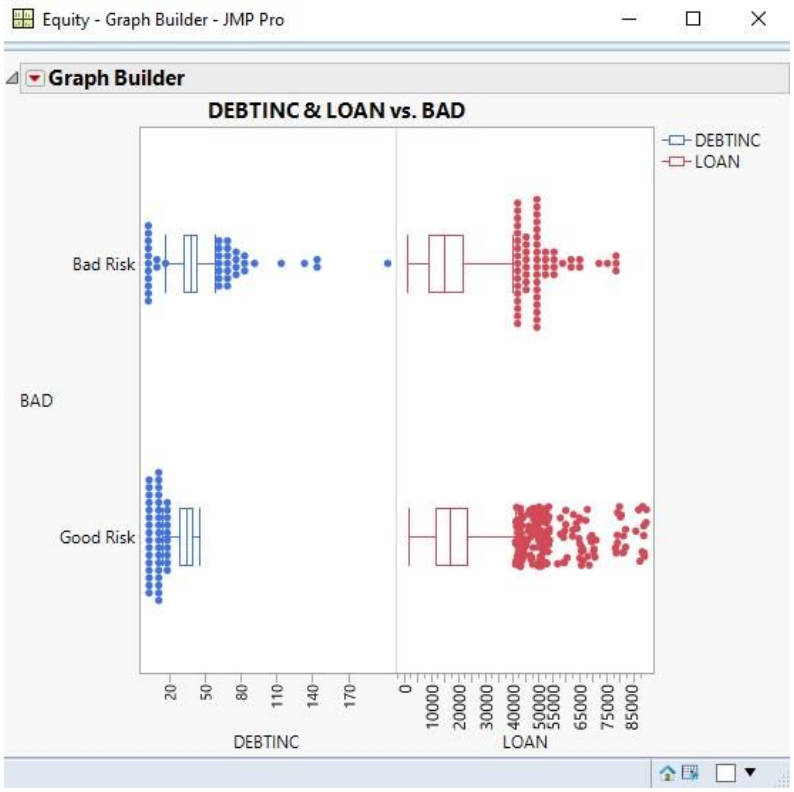
b)



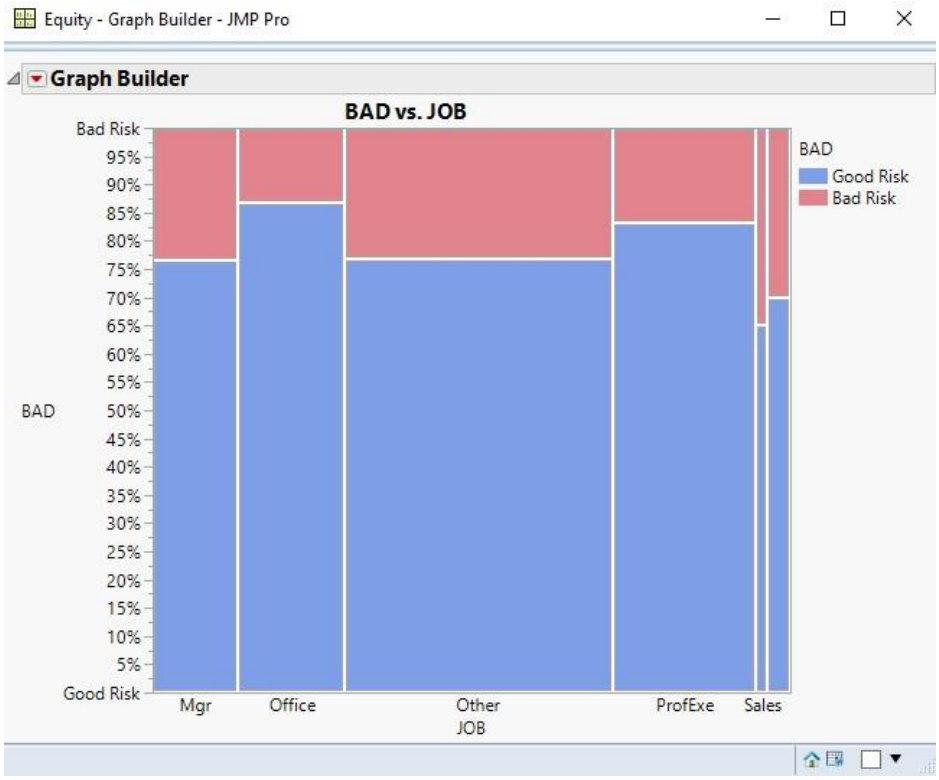
	BAD					
	Good Risk			Bad Risk		
	LOAN	MORTDUE	VALUE	LOAN	MORTDUE	VALUE
Mean	19028,107315	74829,25	102595,9	16922,119428	69460,45	98172,85

Exploring data two variables at a time–graph

a)

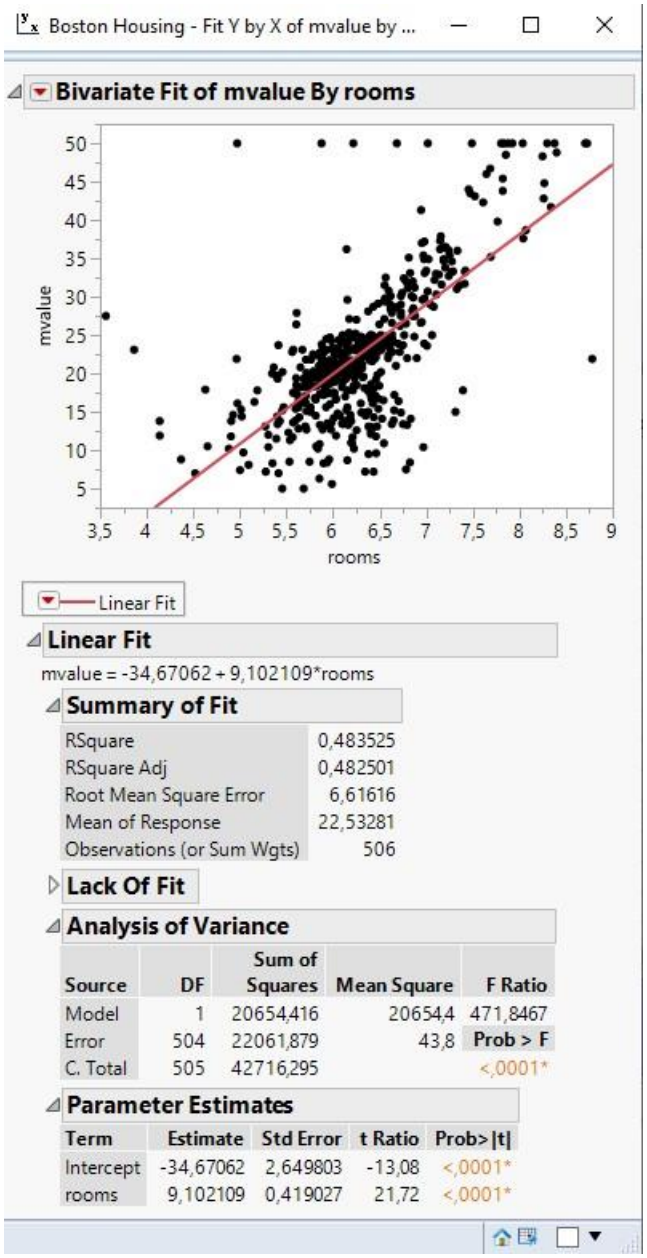


b)

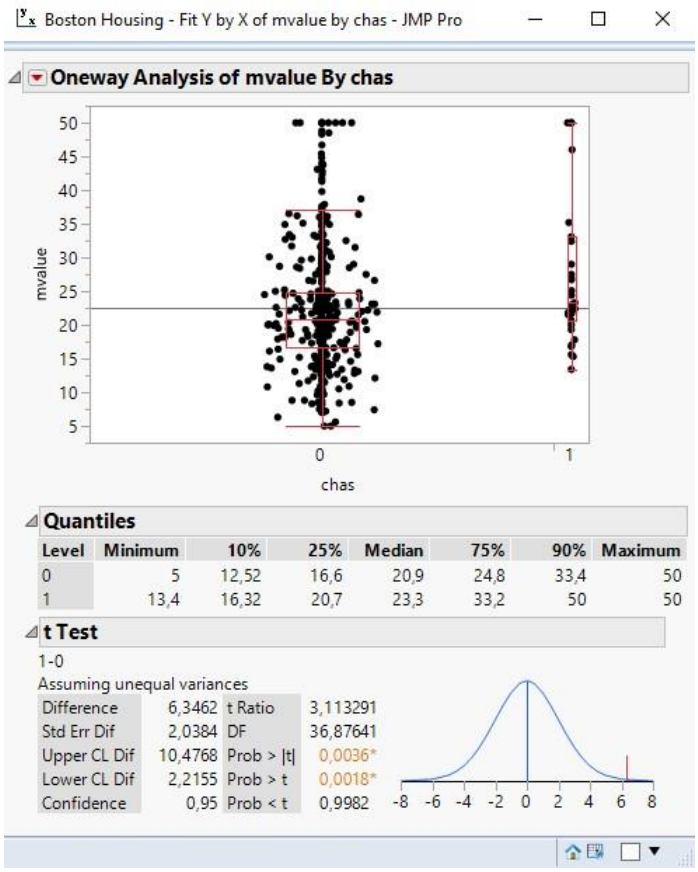


Exercise 2

a)

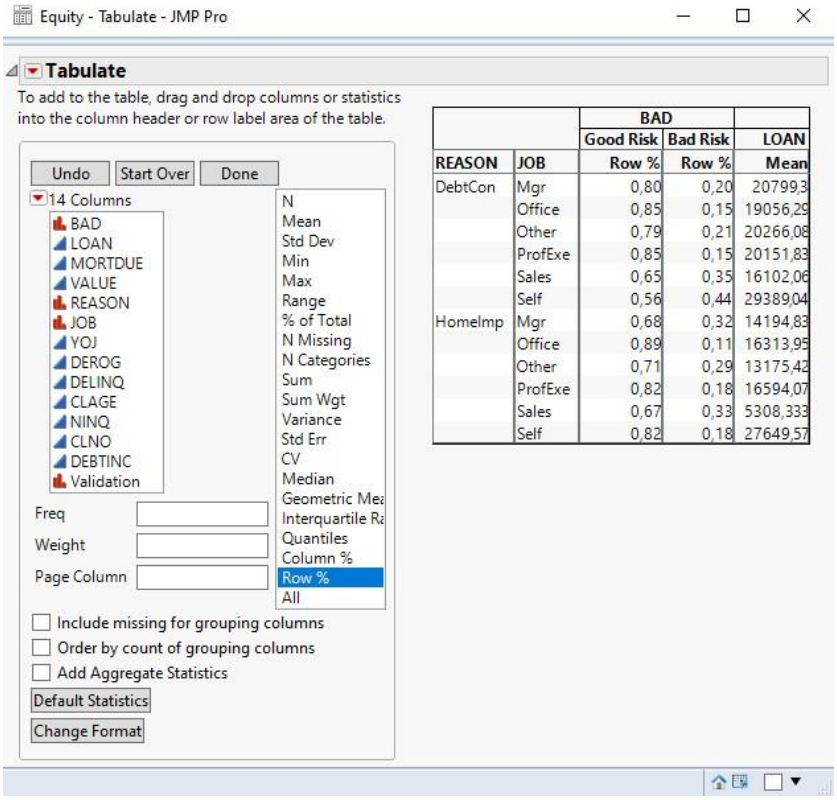

Boston Housing - Fit Y by X of mvalue by ...
Bivariate Fit of mvalue By rooms
mvalue
rooms
Linear Fit
Linear Fit
mvalue = -34,67062 + 9,102109*rooms
Summary of Fit
RSquare 0,483525
RSquare Adj 0,482501
Root Mean Square Error 6,61616
Mean of Response 22,53281
Observations (or Sum Wgts) 506
Lack Of Fit
Analysis of Variance
Source DF Sum of Squares Mean Square F Ratio
Model 1 20654,416 20654,4 471,8467
Error 504 22061,879 43,8 Prob > F
C. Total 505 42716,295 <,0001*
Parameter Estimates
Term Estimate Std Error t Ratio Prob>|t|
Intercept -34,67062 2,649803 -13,08 <,0001*
rooms 9,102109 0,419027 21,72 <,0001*

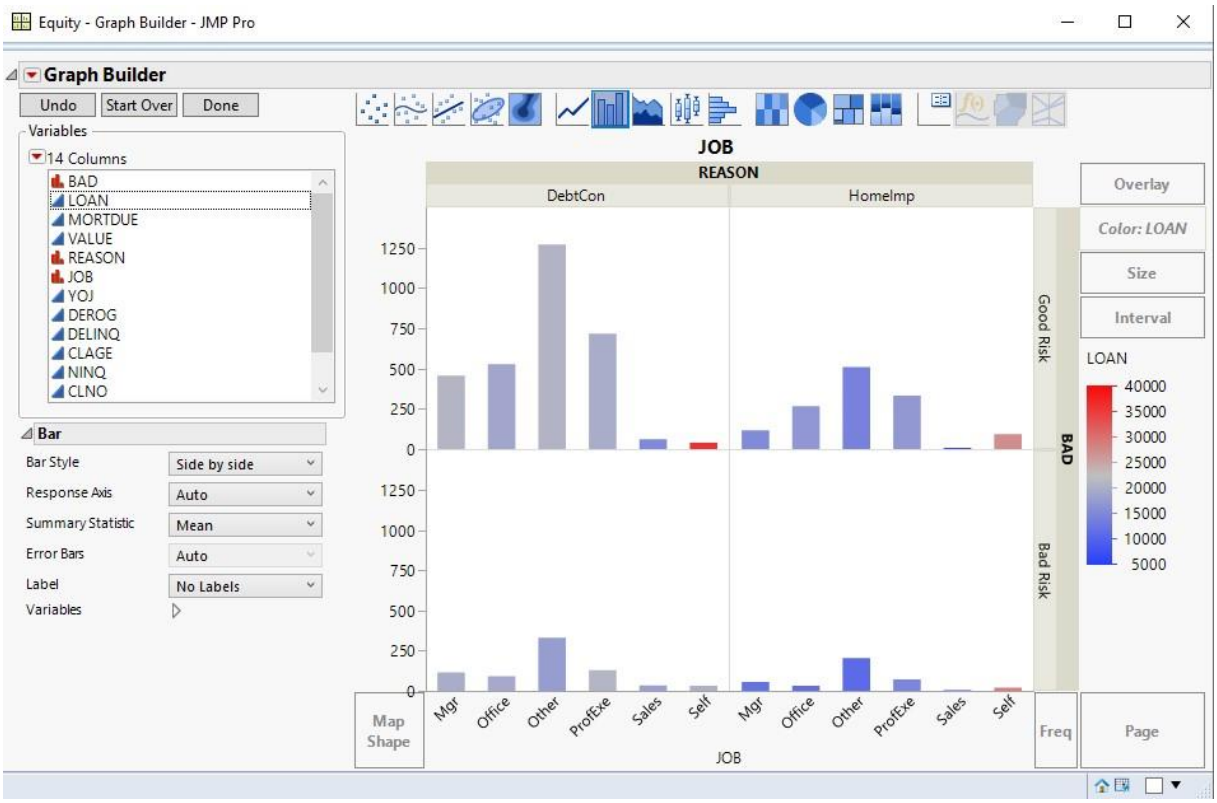
b)



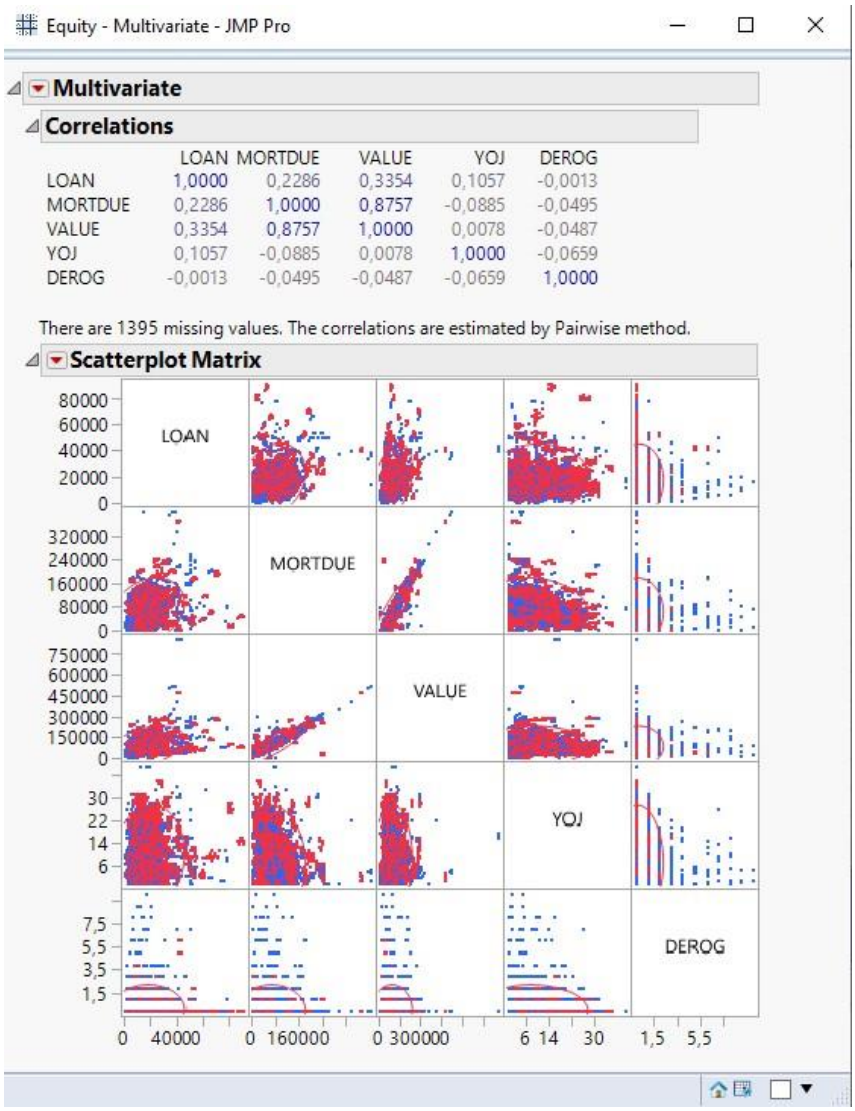
Exercise 3 – Exploring data three or more variables at a time

a)

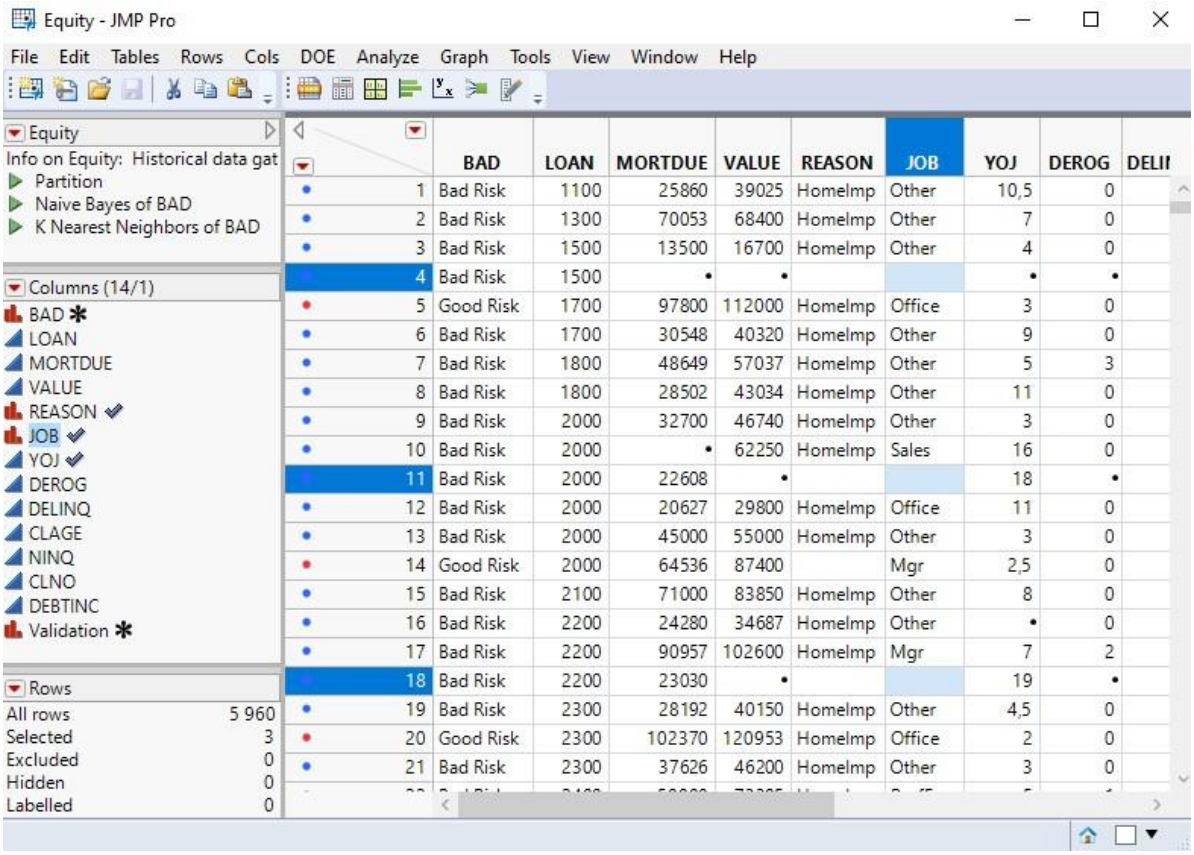


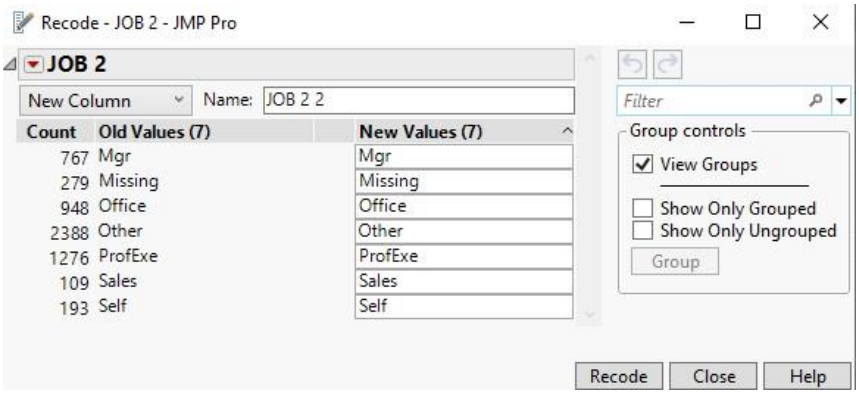

Equity - Graph Builder - JMP Pro
Graph Builder
Undo
Start Over
Done
Variables
14 Columns
BAD
LOAN
MORTDUE
VALUE
REASON
JOB
YOJ
DEROG
DELINQ
CLAGE
NINQ
CLNO
Bar
Bar Style
Side by side
Response Axis
Auto
Summary Statistic
Mean
Error Bars
Auto
Label
No Labels
Variables
JOB
REASON
DebtCon
HomeImp
Good Risk
Bad Risk
BAD
Overlay
Color: LOAN
Size
Interval
LOAN
40000
35000
30000
25000
20000
15000
10000
5000
1250
1000
750
500
250
0
Mgr
Office
Other
ProfExe
Sales
Self
Map Shape
Freq
Page
JOB

b)

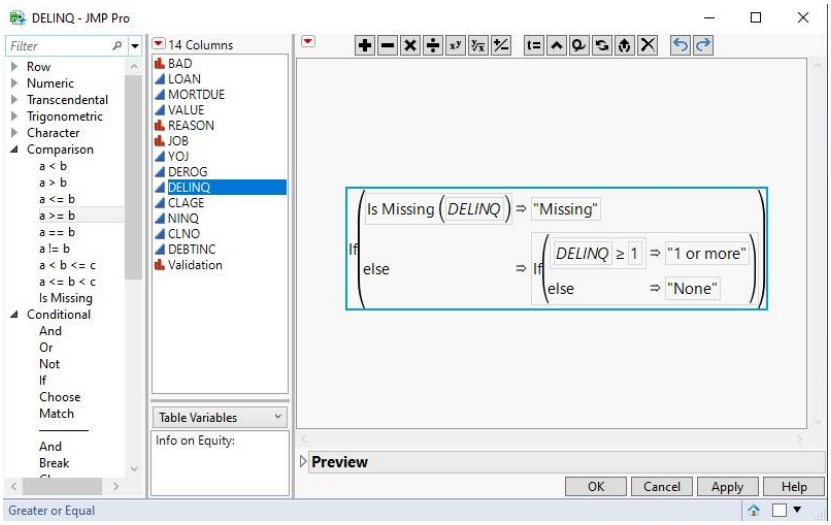

Equity - Multivariate - JMP Pro
Multivariate
Correlations
LOAN MORTDUE VALUE YOJ DEROG
LOAN 1,0000 0,2286 0,3354 0,1057 -0,0013
MORTDUE 0,2286 1,0000 0,8757 -0,0885 -0,0495
VALUE 0,3354 0,8757 1,0000 0,0078 -0,0487
YOJ 0,1057 -0,0885 0,0078 1,0000 -0,0659
DEROG -0,0013 -0,0495 -0,0487 -0,0659 1,0000
There are 1395 missing values. The correlations are estimated by Pairwise method.
Scatterplot Matrix
LOAN
MORTDUE
VALUE
YOJ
DEROG
80000
60000
40000
20000
0
320000
240000
160000
80000
0
750000
600000
450000
300000
150000
0
30
22
14
6
7,5
5,5
3,5
1,5
0 40000
0 160000
0 300000
6 14 30
1,5 5,5

Exercise 4 – Preparing data for modeling



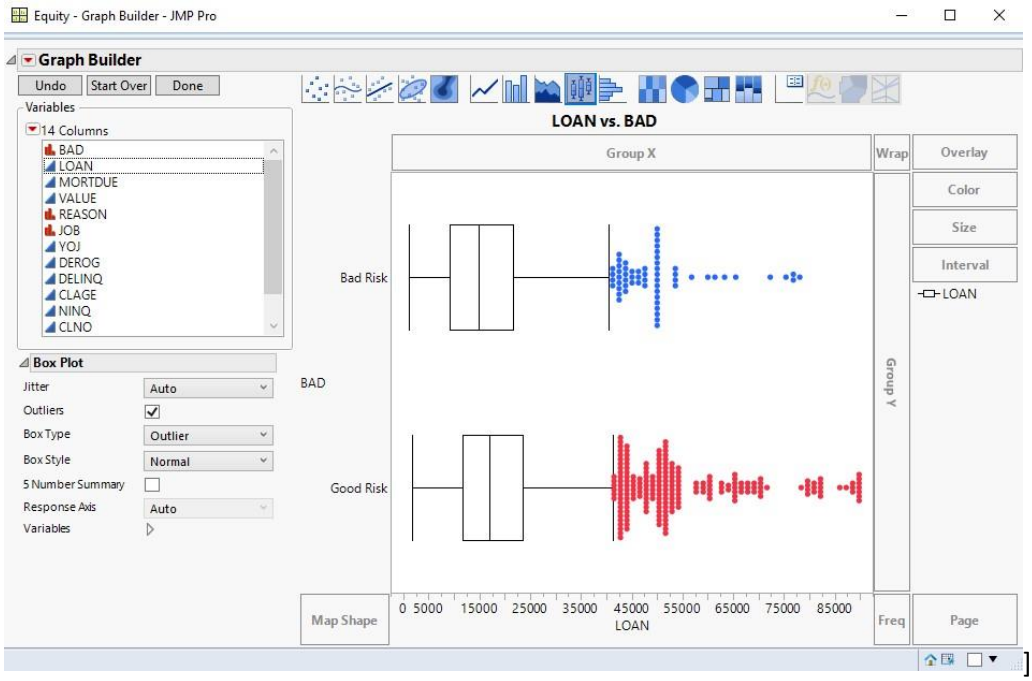


c)



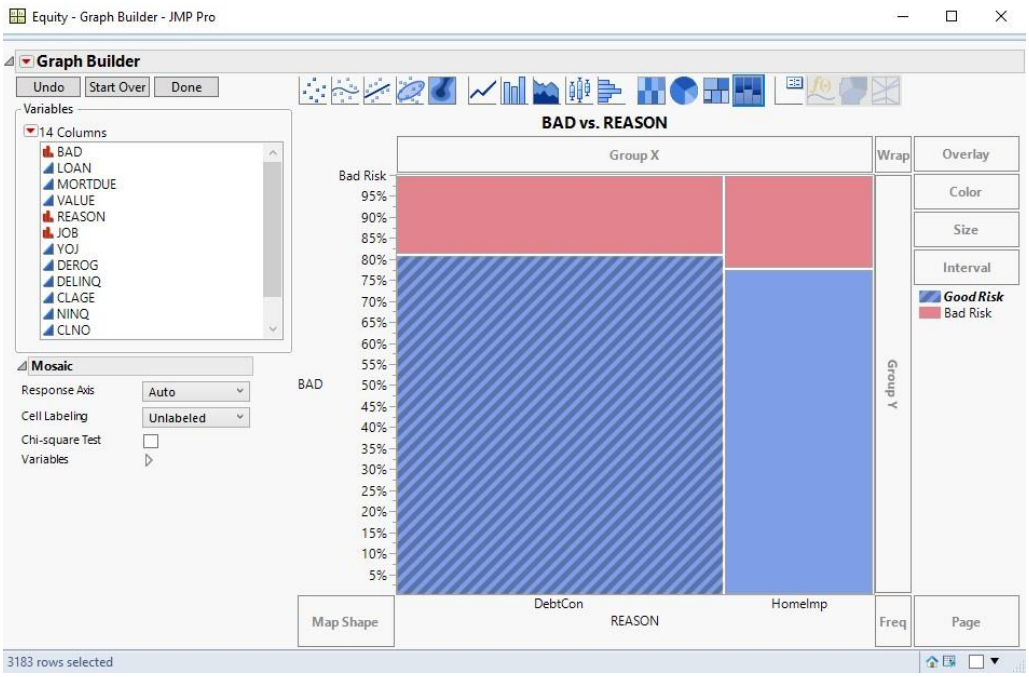
Exercise 5

a)



Osobny podział dla dwóch grup - brak powiązania.

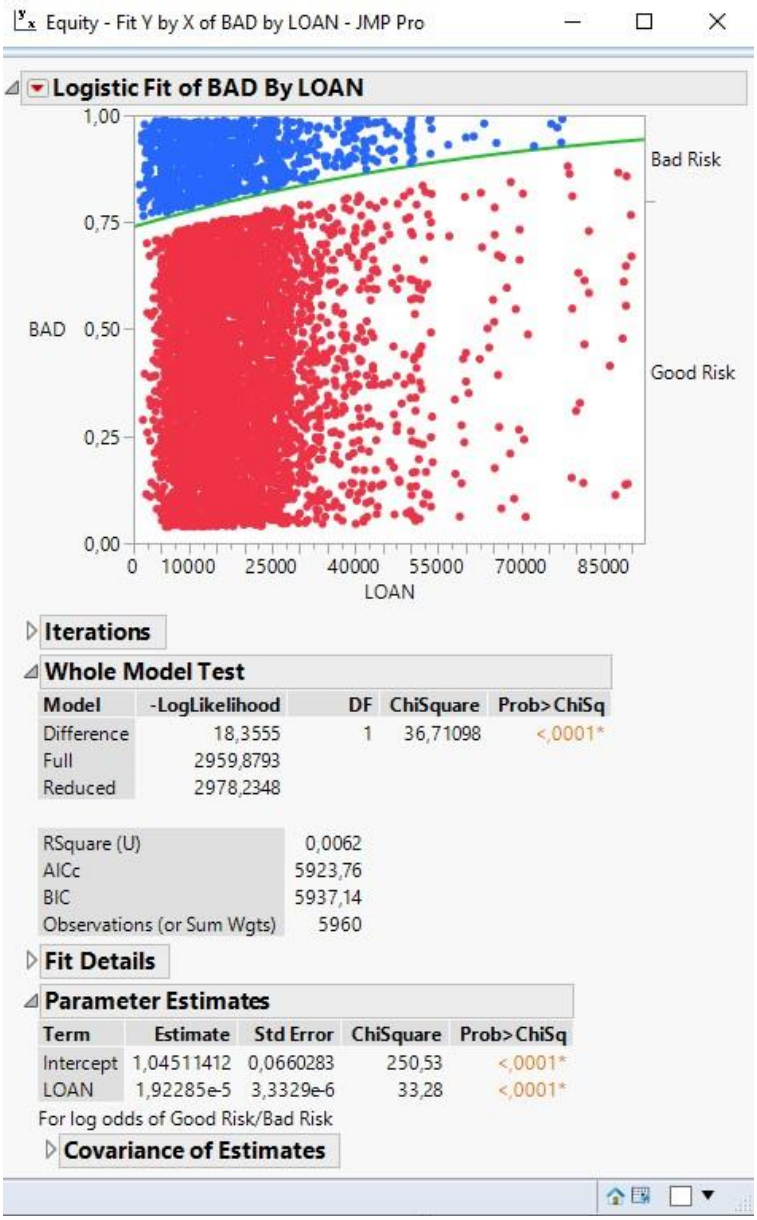
b)



81% of DebtCon to good risk, 19% DebtCon of bad risk. 77.8% home improvement good risk and 22.2% bad risk.

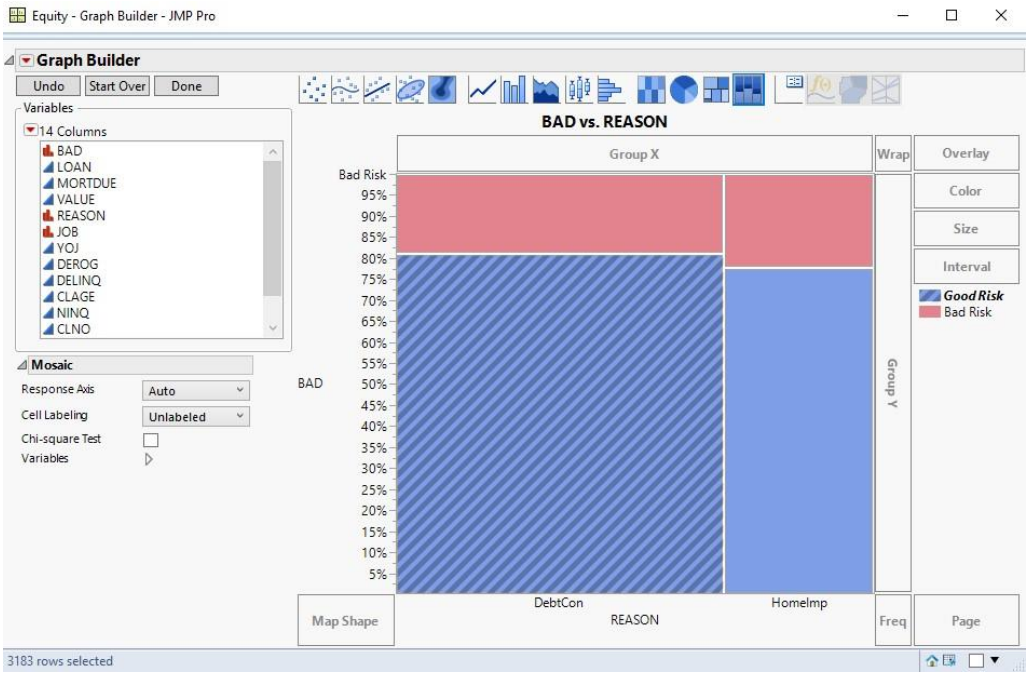
Powiązanie pomiędzy BAD i REASON jest takie, że w zależności od REASON wyszczególnione są osoby które są w grupie Good RIsk i Bad Risk.

c) 1)



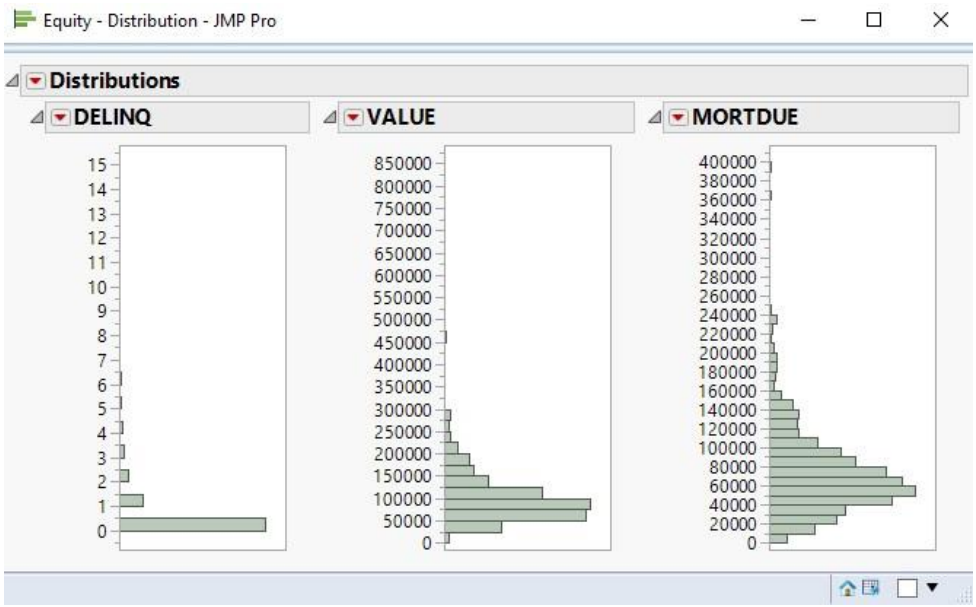
W przypadku mniejszych kwot pożyczek (10000 – 25000) jest mniejsze ryzyko wystąpienia Bad Risk, co oznacza, że pożyczanie kwot w tym przedziale jest dużo bezpieczniejsze. Im większa kwota pożyczki, tym mniej chętnych na tę kwotę.

c) 2)



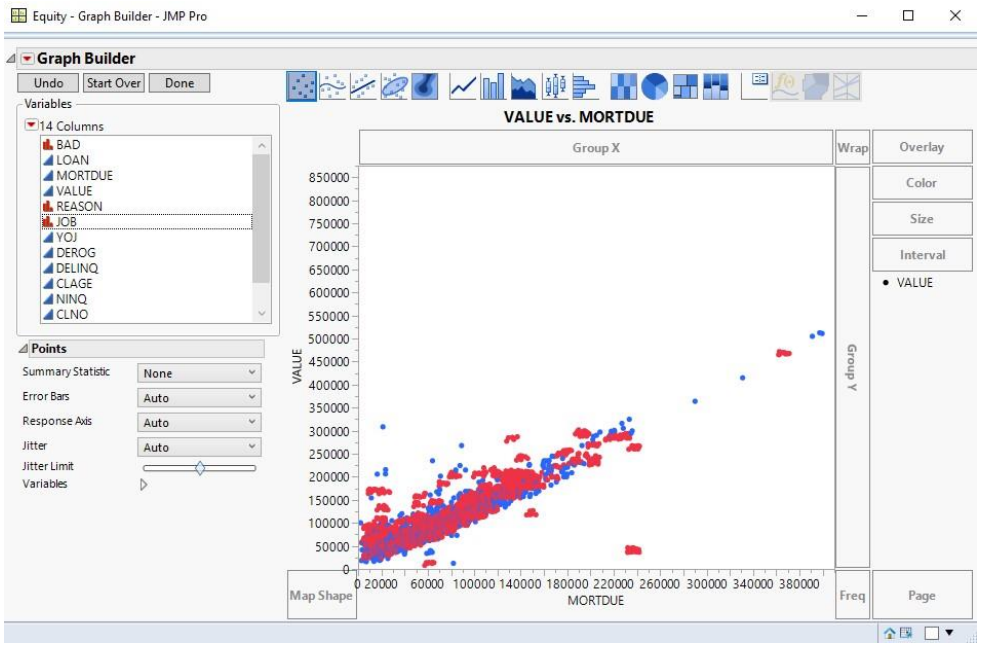
Powiązanie pomiędzy BAD i REASON jest takie, że w zależności od REASON wyszczególnione sa osoby które są w grupie Good Risk i Bad Risk.

d)



Delinq ukazuje wykres słupkowy, natomiast value i mortdue to histogramy, ponieważ słupki są połączone – jest to asymetria lewostronna.

f) 1)



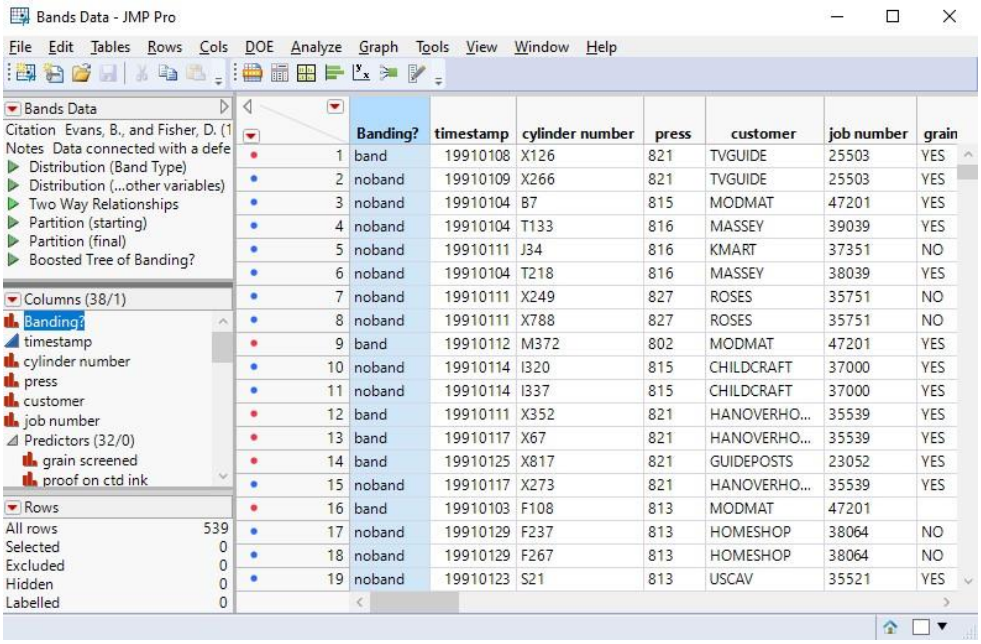
Wykres punktowy przedstawia dane w najdokładniejszy sposób.

f) 2)

Ma sens, ponieważ dane są wtedy najbardziej czytelne na wykresie.

Excercise 6

a)



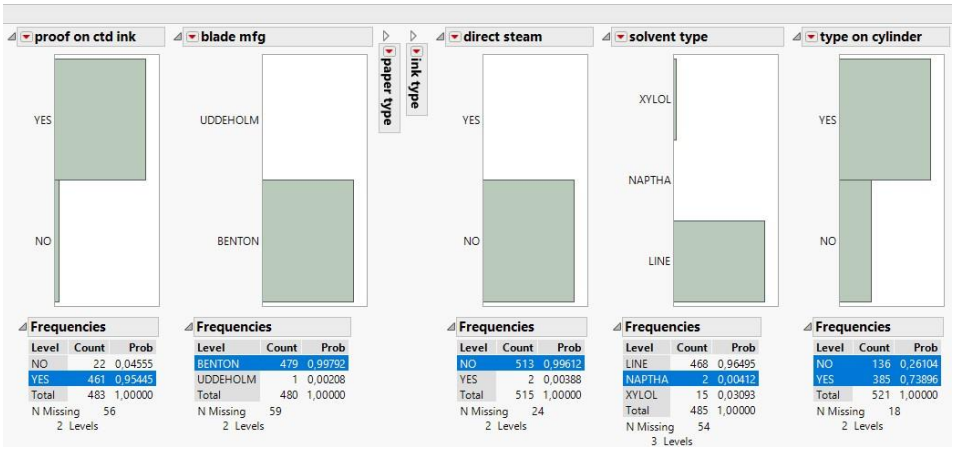
	Banding?	timestamp	cylinder number	press	customer	job number	grain
1	band	19910108	X126	821	TVGUIDE	25503	YES
2	noband	19910109	X266	821	TVGUIDE	25503	YES
3	noband	19910104	B7	815	MODMAT	47201	YES
4	noband	19910104	T133	816	MASSEY	39039	YES
5	noband	19910111	J34	816	KMART	37351	NO
6	noband	19910104	T218	816	MASSEY	38039	YES
7	noband	19910111	X249	827	ROSES	35751	NO
8	noband	19910111	X788	827	ROSES	35751	NO
9	band	19910112	M372	802	MODMAT	47201	YES
10	noband	19910114	I320	815	CHILDCRAFT	37000	YES
11	noband	19910114	I337	815	CHILDCRAFT	37000	YES
12	band	19910111	X352	821	HANOVERHO...	35539	YES
13	band	19910117	X67	821	HANOVERHO...	35539	YES
14	band	19910125	X817	821	GUIDEPOSTS	23052	YES
15	noband	19910117	X273	821	HANOVERHO...	35539	YES
16	band	19910103	F108	813	MODMAT	47201	
17	noband	19910129	F237	813	HOMESHOP	38064	NO
18	noband	19910129	F267	813	HOMESHOP	38064	NO
19	noband	19910123	S21	813	USCAV	35521	YES

37 zmiennych kategorycznych oraz 20 ciągłych.

b) 1)

Między 2-10 kategorii - press, paper type, ink type, press type, solvent type, unit number, cylinder size, paper mill location.

b) 2)

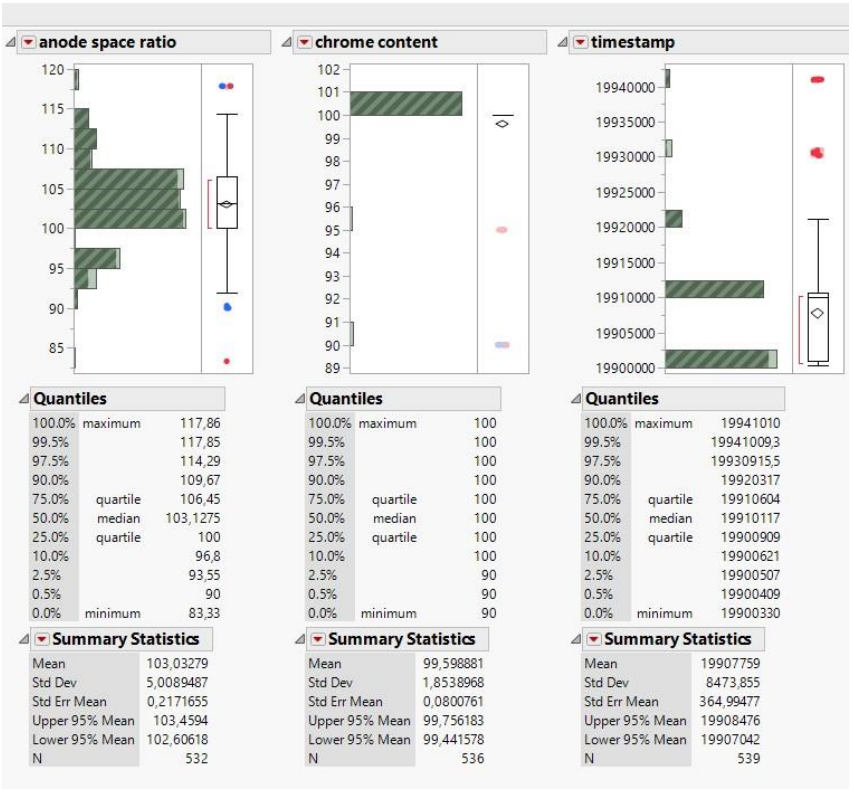


Proof on ctd ink, blade mfg, direct steam, solvent type, type on cylinder

c) 1)

Esa Voltage – asymetira lewostronna, Wax – asymetria prawostronna, Amperage – wykres słupkowy

c) 2)



Dla chrome content – 100, liczba elementów wynosi 510

Dla 95 – liczba elementów wynosi 9

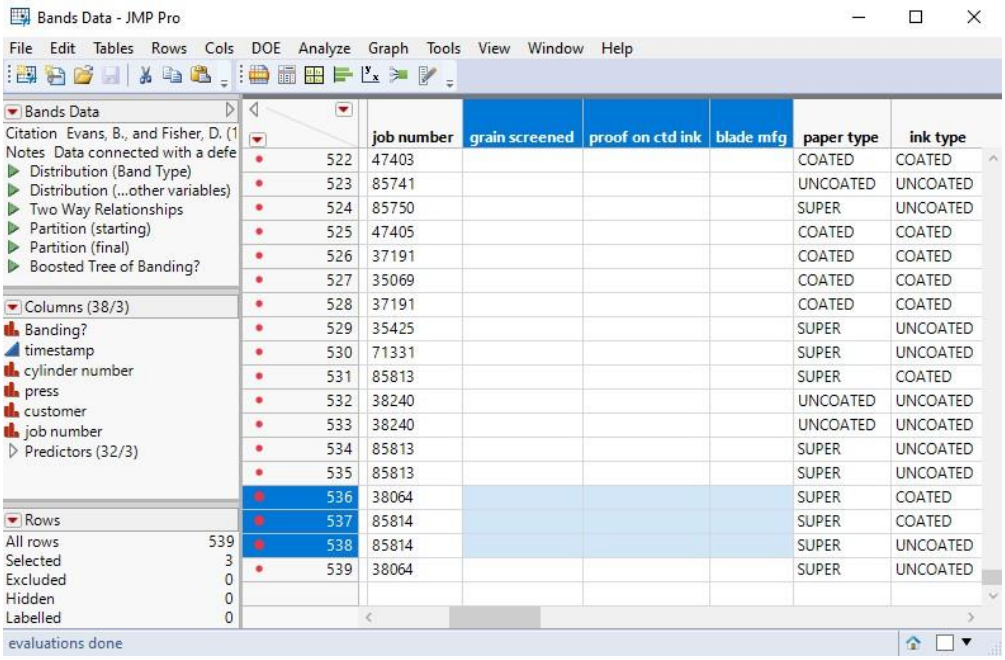
Dla 90 – liczba elementów wynosi 17

Można wybrać pojedyńcze elementy z kategorii co powoduje ukazanie z którymi elementami z (innych kategorii) jest powiązanie.

d)

W miejscu gdzie jest ciągła, jest to relacja 1 do wielu, a tam gdzie jest kategoryczny jest 1:1

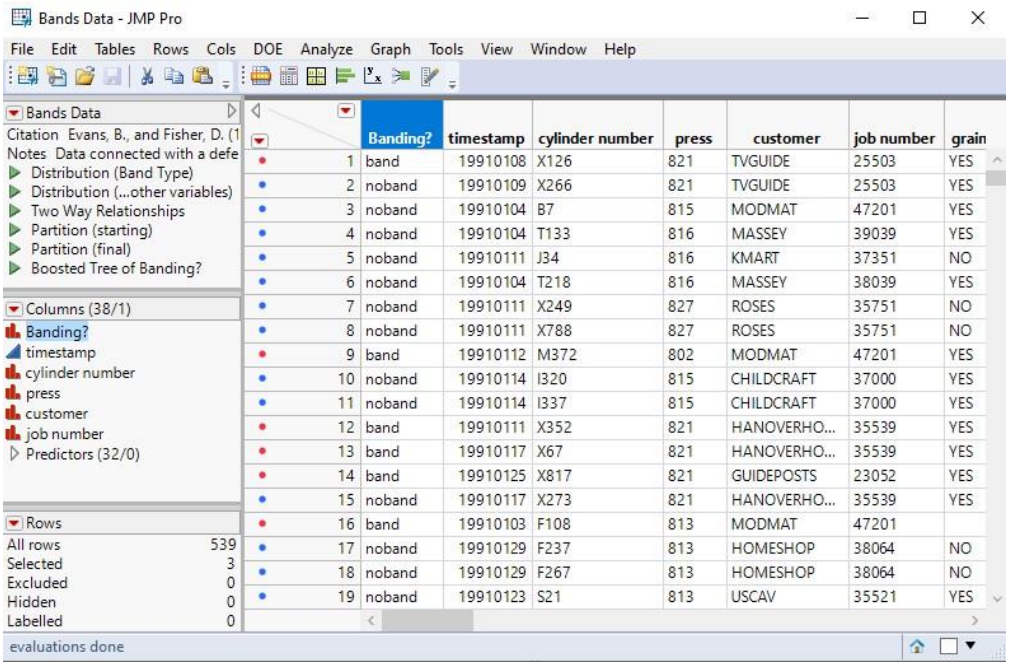
e)



	job number	grain screened	proof on ctd ink	blade mfg	paper type	ink type
522	47403				COATED	COATED
523	85741				UNCOATED	UNCOATED
524	85750				SUPER	UNCOATED
525	47405				COATED	COATED
526	37191				COATED	COATED
527	35069				COATED	COATED
528	37191				COATED	COATED
529	35425				SUPER	UNCOATED
530	71331				SUPER	UNCOATED
531	85813				SUPER	COATED
532	38240				UNCOATED	UNCOATED
533	38240				UNCOATED	UNCOATED
534	85813				SUPER	UNCOATED
535	85813				SUPER	UNCOATED
536	38064				SUPER	COATED
537	85814				SUPER	COATED
538	85814				SUPER	UNCOATED
539	38064				SUPER	UNCOATED

Nie są informacyjne ponieważ dla niektórych „bandingów” nie pokazują żadnych informacji

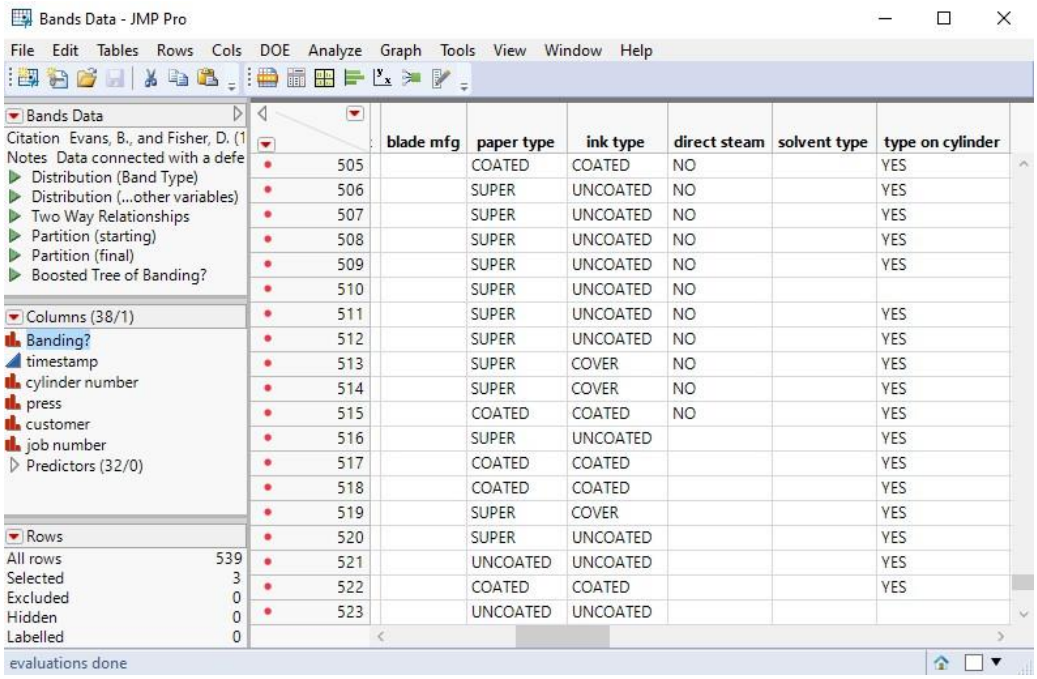
f)



	Banding?	timestamp	cylinder number	press	customer	job number	grain
1	band	19910108	X126	821	TVGUIDE	25503	YES
2	noband	19910109	X266	821	TVGUIDE	25503	YES
3	noband	19910104	B7	815	MODMAT	47201	YES
4	noband	19910104	T133	816	MASSEY	39039	YES
5	noband	19910111	J34	816	KMART	37351	NO
6	noband	19910104	T218	816	MASSEY	38039	YES
7	noband	19910111	X249	827	ROSES	35751	NO
8	noband	19910111	X788	827	ROSES	35751	NO
9	band	19910112	M372	802	MODMAT	47201	YES
10	noband	19910114	I320	815	CHILDCRAFT	37000	YES
11	noband	19910114	I337	815	CHILDCRAFT	37000	YES
12	band	19910111	X352	821	HANOVERHO...	35539	YES
13	band	19910117	X67	821	HANOVERHO...	35539	YES
14	band	19910125	X817	821	GUIDEPOSTS	23052	YES
15	noband	19910117	X273	821	HANOVERHO...	35539	YES
16	band	19910103	F108	813	MODMAT	47201	
17	noband	19910129	F237	813	HOMESHOP	38064	NO
18	noband	19910129	F267	813	HOMESHOP	38064	NO
19	noband	19910123	S21	813	USCAV	35521	YES

Są powiązane

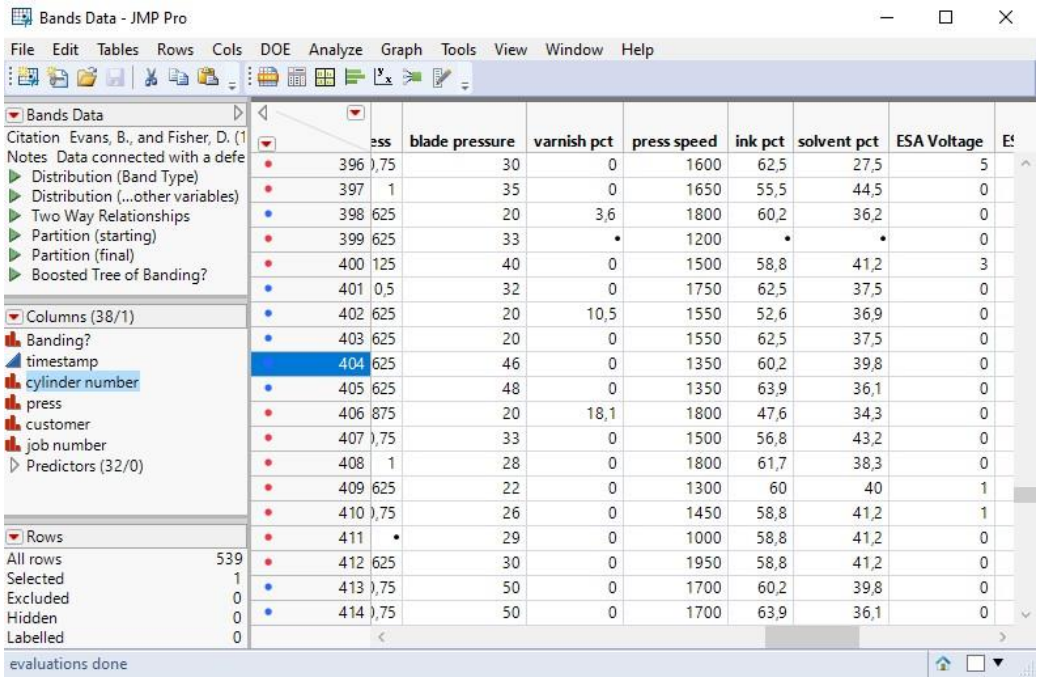
g)



	blade mfg	paper type	ink type	direct steam	solvent type	type on cylinder
505		COATED	COATED	NO		YES
506		SUPER	UNCOATED	NO		YES
507		SUPER	UNCOATED	NO		YES
508		SUPER	UNCOATED	NO		YES
509		SUPER	UNCOATED	NO		YES
510		SUPER	UNCOATED	NO		
511		SUPER	UNCOATED	NO		YES
512		SUPER	UNCOATED	NO		YES
513		SUPER	COVER	NO		YES
514		SUPER	COVER	NO		YES
515		COATED	COATED	NO		YES
516		SUPER	UNCOATED			YES
517		COATED	COATED			YES
518		COATED	COATED			YES
519		SUPER	COVER			YES
520		SUPER	UNCOATED			YES
521		UNCOATED	UNCOATED			YES
522		COATED	COATED			YES
523		UNCOATED	UNCOATED			

Problem istnieje – występują brakujace wartości które są widoczne w data table. Data table ukazuje, że dla niektorych kategorii wartości mają puste pola. Problematyczne pola są puste przez co nie możemy odczytać niektórych wartości. Powinna byd jakakolwiek informacja, czy puste pole jest uzasadnione, np. wartość null, czy pole rzeczywiście jest puste i nie ma żadnej wartości.

h)



	ess	blade pressure	varnish pct	press speed	ink pct	solvent pct	ESA Voltage	ES
396	),75	30	0	1600	62,5	27,5	5	
397	1	35	0	1650	55,5	44,5	0	
398	625	20	3,6	1800	60,2	36,2	0	
399	625	33	•	1200	•	•	0	
400	125	40	0	1500	58,8	41,2	3	
401	0,5	32	0	1750	62,5	37,5	0	
402	625	20	10,5	1550	52,6	36,9	0	
403	625	20	0	1550	62,5	37,5	0	
404	625	46	0	1350	60,2	39,8	0	
405	625	48	0	1350	63,9	36,1	0	
406	875	20	18,1	1800	47,6	34,3	0	
407	),75	33	0	1500	56,8	43,2	0	
408	1	28	0	1800	61,7	38,3	0	
409	625	22	0	1300	60	40	1	
410	),75	26	0	1450	58,8	41,2	1	
411	•	29	0	1000	58,8	41,2	0	
412	625	30	0	1950	58,8	41,2	0	
413	),75	50	0	1700	60,2	39,8	0	
414	),75	50	0	1700	63,9	36,1	0	

Kropki powinny przedstawiać jakąkolwiek wartość, aby łatwiej było je zidentyfikować.