Data Quality Report - Initial Findings

Descriptive Statistics for Continuous Features

	count	mean	std	min	25%	50%	75%	max
ExternalRiskEstimate	943.0	71.817603	10.473129	-9.0	64.0	72.0	80.0	92.0
MSinceOldestTradeOpen	943.0	196.037116	101.039153	-8.0	129.0	181.0	252.5	590.0
MSinceMostRecentTradeOpen	943.0	9.667020	13.337026	0.0	3.0	6.0	12.0	184.0
AverageMInFile	943.0	79.463415	36.269141	7.0	57.0	75.0	96.0	258.0
NumSatisfactoryTrades	943.0	20.923648	11.026435	1.0	13.0	20.0	27.0	74.0
NumTrades60Ever2DerogPubRec	943.0	0.617179	1.306850	0.0	0.0	0.0	1.0	12.0
NumTrades90Ever2DerogPubRec	943.0	0.379639	1.009556	0.0	0.0	0.0	0.0	12.0
PercentTradesNeverDelq	943.0	92.217391	11.887317	20.0	88.5	97.0	100.0	100.0
MSinceMostRecentDelq	943.0	7.202545	20.321783	-8.0	-7.0	-7.0	14.0	81.0
NumTotalTrades	943.0	22.686108	12.637905	0.0	14.0	21.0	29.0	78.0
NumTradesOpeninLast12M	943.0	1.889714	1.891014	0.0	0.0	1.0	3.0	17.0
PercentInstallTrades	943.0	34.397667	17.641026	0.0	21.0	33.0	45.5	100.0
MSinceMostRecentInqexcl7days	943.0	0.511135	5.954501	-8.0	0.0	0.0	2.0	24.0
NumInqLast6M	943.0	1.650053	3.009630	0.0	0.0	1.0	2.0	66.0
NetFractionRevolvingBurden	943.0	34.415695	29.199131	-8.0	7.0	30.0	57.5	104.0
NetFractionInstallBurden	943.0	42.091198	41.348169	-8.0	-8.0	53.0	79.0	138.0
NumRevolvingTradesWBalance	943.0	3.830329	3.106881	-8.0	2.0	3.0	5.0	29.0
NumInstallTradesWBalance	943.0	1.652174	3.337033	-8.0	1.0	2.0	3.0	23.0
NumBank2NatlTradesWHighUtilization	943.0	0.600212	2.342780	-8.0	0.0	1.0	2.0	13.0
PercentTradesWBalance	943.0	65.994698	22.382056	-8.0	50.0	67.0	83.0	100.0

While count is full for all features, this does not mean that the dataset contains a full set of continuous values across all features for each accounts, with negative values of -8 and -9 being used to represent where there was no useable data associated with an account for a feature. -7 is also a special, non-continuous value used to represent where a condition has not been met (e.g. no delinquencies). For the same reason, the minimum values displayed in the table for external risk estimate, months since oldest trade, months since most recent delinquency, months since last inquiry excluding 7 days and net fraction install burden (-7, -8 or -9) are not actually representative of the minimum range of these continuous feature.

A significant proportion of the values across these features are 0. Number of trades 60+ ever and number of trades 90+ ever contain at least 50% and 75% 0 values respectively, which may simply indicate that the majority of accounts never make trades of these sizes. Interestingly, at

least 50% of accounts have had 1 inquiry in the last 6 months, which may perhaps suggest that most accounts are inquired into as a procedure regularly over certain time intervals. MSinceMostRecentInqexcl7days offers more insight into this aspect of the data, revealing that the majority of accounts have had an inquiry within the current month, and that at least 75% have been the subject of an inquiry in the last 2 months.

Months since most recent delinquency contains at least 50% -7 values, meaning that the majority of accounts have never had a delinquent trade.

Net fraction install burden contains 25% -8 values, which is the special value used in the continuous columns to represent no useable or valid trades/inquiries, meaning that at least 25% of the values for this feature are effectively missing. Investigating this could be worthwhile as this is a significant amount of missing data.

There is a significant gap between the 75% quartile and max for many of these features, indicating that there are some extreme outliers among the values in the dataset. For example, at least 75% of values in months since most recent trade open are 12 months or less while the max is greater than 15 years. Another example of an extremely outlying max value can be seen in number of inquiries in the last 6 months, with 75% of accounts having had 2 or less inquiries while one account has had 66 inquiries in the same period, which may be an error.

Standard deviation is not extreme in any of the features except for months since oldest trade open, where it is ~101. However, this high standard deviation is not unreasonable, considering that accounts will have a broad range of account ages.

Percentage trades never delinquent is 100% for at least 75% of accounts, which on it’s own would indicate that 75% accounts have a perfect record in respect to delinquency. However, only months since most recent delinquency is only negative (i.e. no delinquency ever) up to the 50% quartile. This apparent contradiction in the dataset warrants further investigation, especially as this may lead to the discovery of other problems in the dataset which may not have been identified in this report.

Descriptive Statistics for Categorical Features

	count	unique	top	freq
RiskPerformance	943	2	Bad	486
MaxDelq2PublicRecLast12M	943	8	7	408
MaxDelqEver	943	7	8	436

We have a full count of all categorical features.

Looking at our target feature risk performance, which has two unique values ‘good’ and ‘bad’, we can see 486 out of the 943 accounts are rated ‘Bad’. This equates to a nearly 50/50 split between bad and good values (roughly 51.5/48.5).

The top value in the maximum delinquency/public record last 12 months is 7, accounting for 408 out of the 943 accounts in the dataset. Unfortunately, alone this tells us essentially nothing, as this value is used to denote both accounts who have never had a delinquent trade and accounts who currently have a delinquent trade.

The top value in maximum delinquency ever is 8, accounting for 436 of the 943 accounts in the dataset. Again, this value has been used to represent both accounts with a current delinquency trade and accounts with no delinquent trades in their history, so it is also unhelpful to us without additional information to determine which of these two possible meanings is represented by this value in each row it appears in.

Histograms for Continuous Features [plots attached at end of file]

Several of the continuous features in the dataset are quite normally distributed, including average months in file (centred between 50 and 100), external risk estimate (centred between 60 and 80), months since oldest trade open (centred just below 200, although the positioning of the lowest bin for this feature is not ideal as will be discussed soon in this report), number of satisfactory trades and number of total trades (both centred around 20) and percent installment trades (centred just below 40).

Months since most recent delinquency is exponentially decreasing, peaking somewhere below zero and diving sharply above 0. However, this doesn’t tell us anything about the data, as this largest bin unfortunately combines 0 values, which represent an immediately recent/ongoing delinquency, with negative values, which represent the absence of a delinquency. This unfortunate binning of continuous values with numbers representing unrelated categorical

values also affects, months since oldest trade open, net fraction install burden, net fraction revolving burden, number of installment trades with balance, number of revolving trades with balance and percent trades with balance.

Number of bank/national trades with high utilisation and months since most recent inquiry excluding last 7 days also seem to be (very steeply) normally distributed (both centering around 0), ignoring the fact that both display a total void of values between 0 and somewhere between -5 and -10. Given that the minimum value for these features is -8 and the value -7 and -8 for continuous features represents a condition not being met or no usable trades or inquiries, we can gather from this that roughly a quarter of accounts have never been the subject of an inquiry, while a significantly smaller but non-negligible proportion of accounts have no bank/national trades with high utilisation in their account history.

Months since most recent trade open, number of inquiries in the last 6 months, number of trades 60+ and number of trades 90+ ever are nearly entirely concentrated in the lowest bin, indicating that the overwhelming majority of accounts trade frequently, have had few to no inquiries in the last 6 months and made little or no 60+ (and less 90+) trades with their accounts.

Percentage trades never delinquent is exponentially increasing (bar a complete absence of values between the late 20s to late 30s), spiking sharply going from the second highest to highest bin. This indicates that most trades in the account history of the overwhelming majority of accounts have not been delinquent, with only an extremely small subset of accounts having a history of mostly delinquent trades. Number of trades open in the last 12 months is also exponential decreasing (though significantly more gradually), with nearly all accounts falling in the first three bins representing 0 to around 5 trades.

Boxplots for Continuous Features [plots attached at end of file]

Most continuous features in this dataset contain a significant amount of outliers, with the only exceptions being external risk estimate, with a single outlier multiple standard deviations below the low cut-off points, and net fraction revolving burden and net fraction install burden which both have no outliers. In all of the other features, some of the outliers are far outside the cut-off points and many standard deviations away from the mean, with the most extreme examples of features with a high concentration of far outliers being months since most recent delinquency, followed by average months in file and months since oldest trade open. Features with particularly high numbers of outliers include months since oldest trade open, months since most recent trade open, average months in file, percentage trades never delinquent, months since most recent delinquency, number of total trades and months since most recent inquiry excluding last 7 days.

The lack of multiple outliers in external risk estimate shows us that valid values for this feature (the single outlier is a -9 value) do not have a huge range, seemingly no less than 40 and no

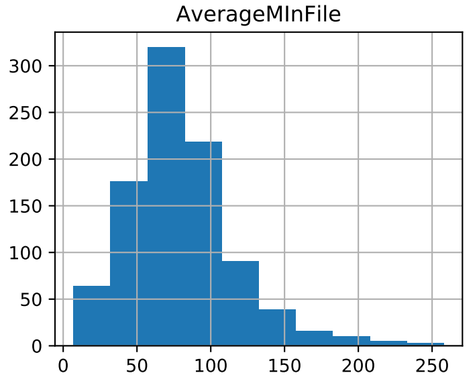
more than 100. This seems reasonable, as it is likely that these external assessments are done on scale of 0 to 100, or standardised to such a scale if provided through external organisations, with 100 representing a theoretical 0% risk and 0 representing the reverse. The interquartile range for months since most recent delinquency (-7 to roughly 16) is not representative of any real trend as it groups in both -7 values representing no delinquencies with low positive values representing a very recent delinquency. The large number of far outlying positive values in months since most recent trade open may be capturing some accounts which have been abandoned by their owners but have not been closed for whatever reason. The interquartile range (0 to roughly 2.3) for months since most recent inquiry is interesting, as it suggests that the majority of accounts are currently or have recently been the subject of an inquiry, while the outliers suggest that some accounts have somehow avoided this for up to 24 months.

Bar Plots for Categorical Features [plots attached at end of file]

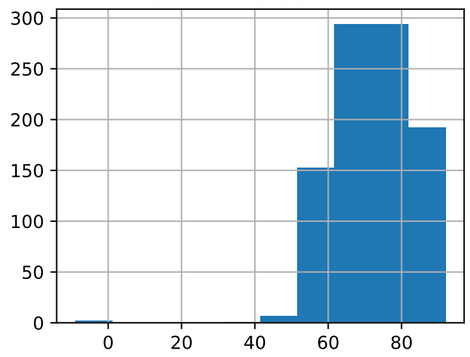
The two categories, bad and good, in the risk performance plot are roughly equal in size, meaning that there is a nearly even split between these values in this features as was previously concluded in the descriptive statistics for categorical features section.

In the plots for max delinquency/public record last 12 months and max delinquency ever, the value representing “current and never delinquent” (7 and 8 respectively) has the majority. This is both unsurprising and unhelpful, as both of these values represent two different categories merged together. Therefore, we are unfortunately unable to learn the frequencies of these two functionally different categories of accounts from our plot. If these categories are ignored, both plots are dominated by the value 6, which has a different meaning for both features, representing “unknown delinquency” in max delinquency/public record last 12 months and “30 days delinquent” in max delinquency ever. The meaning of “unknown delinquency” is unclear but it seems likely that it has been used to represent missing data. This brings the usefulness of the plot of this feature in to question, as while the rest it’s values have clear meanings and represent binning of different ranges of days (bar the 0 value, which represents a “derogatory comment”), these amount to less than half of the data for this feature. The plot for max delinquency ever gives us more to work with (after the bars for 8/current and never delinquent and 7/unknown delinquency are ignored), giving us a clear visual representation of the maximum number of days delinquent for accounts (6 = 30 days, 5 = 60 days, 4 = 90 days, 3 = 120+ days, 2 = “derogatory comment”) who are not current delinquents but have some delinquent trade in their history.

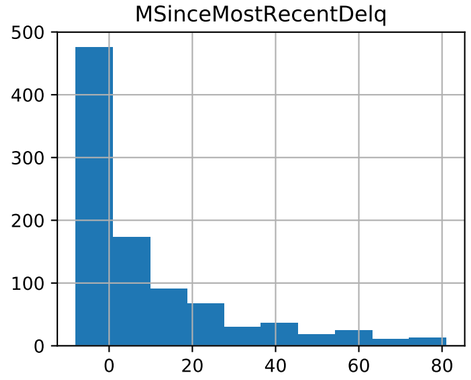
AverageMInFile



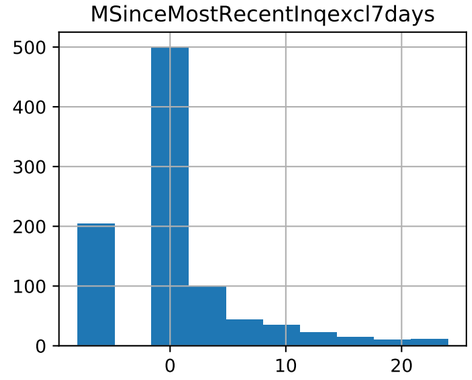
ExternalRiskEstimate



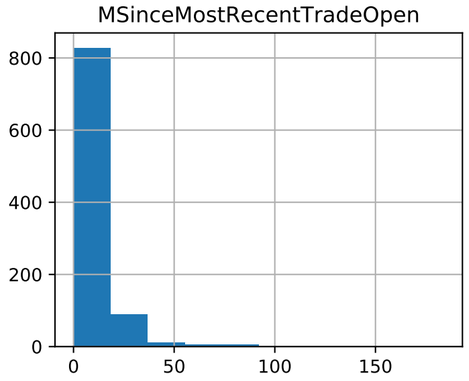
MSinceMostRecentDelq



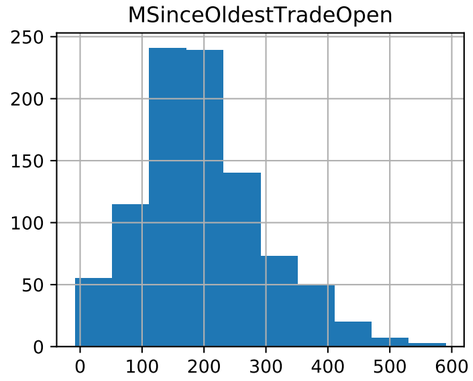
MSinceMostRecentInqexcl7days



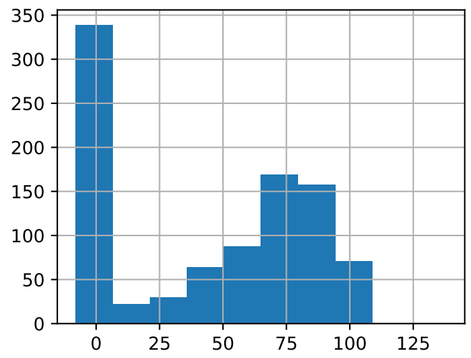
MSinceMostRecentTradeOpen



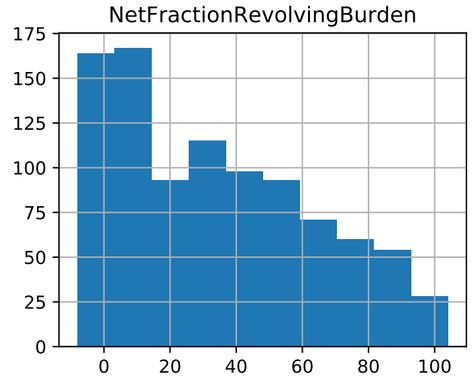
MSinceOldestTradeOpen



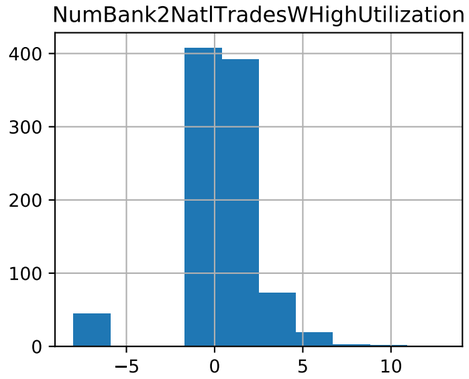
NetFractionInstallBurden



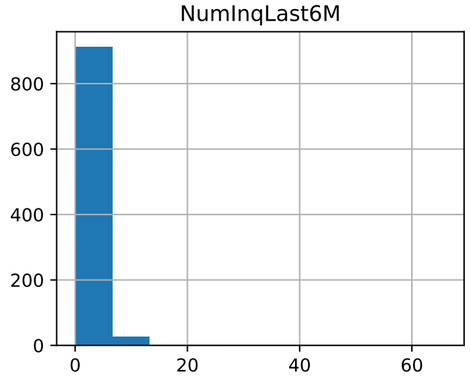
NetFractionRevolvingBurden



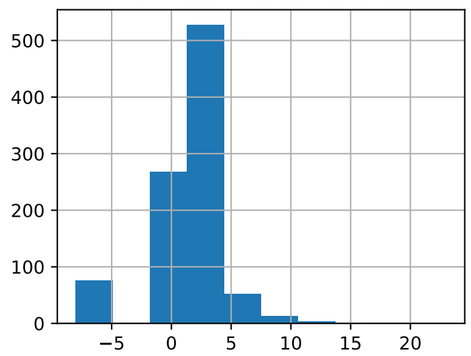
NumBank2NatlTradesWHighUtilization



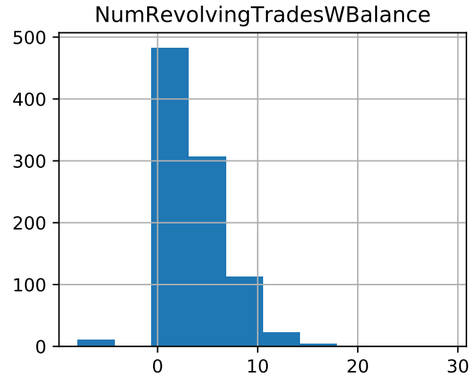
NumInqLast6M



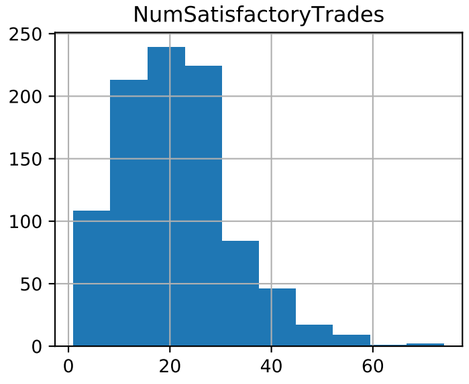
NumInstallTradesWBalance



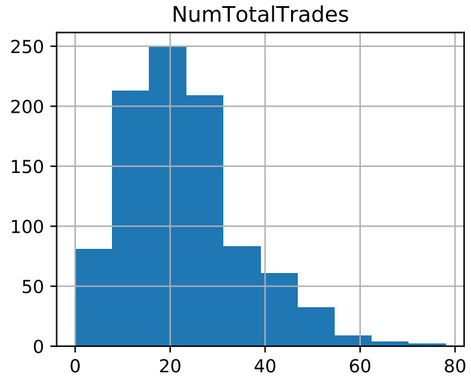
NumRevolvingTradesWBalance



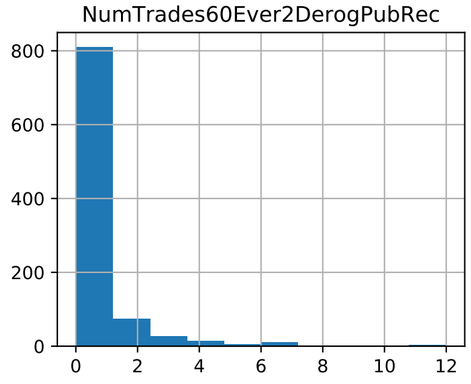
NumSatisfactoryTrades



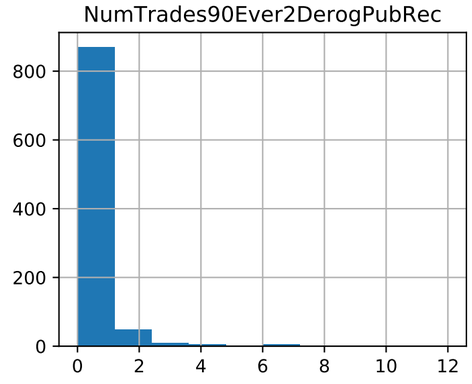
NumTotalTrades



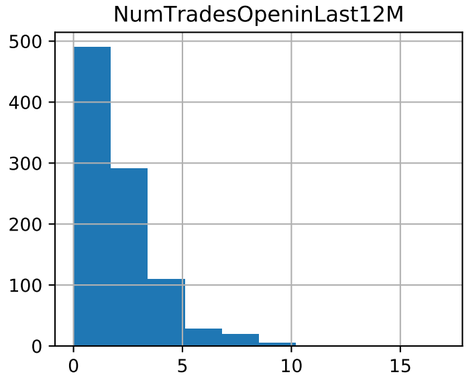
NumTrades60Ever2DerogPubRec



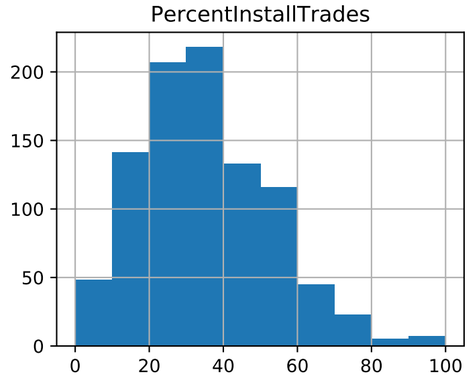
NumTrades90Ever2DerogPubRec



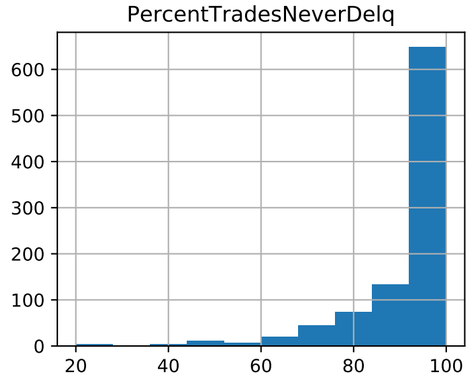
NumTradesOpeninLast12M



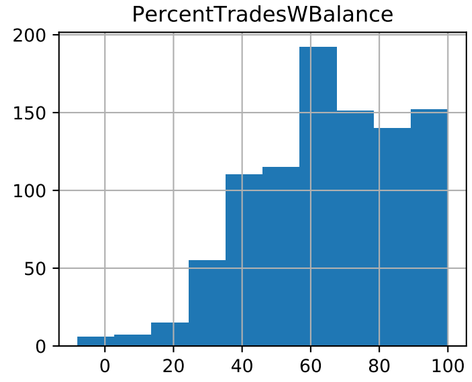
PercentInstallTrades

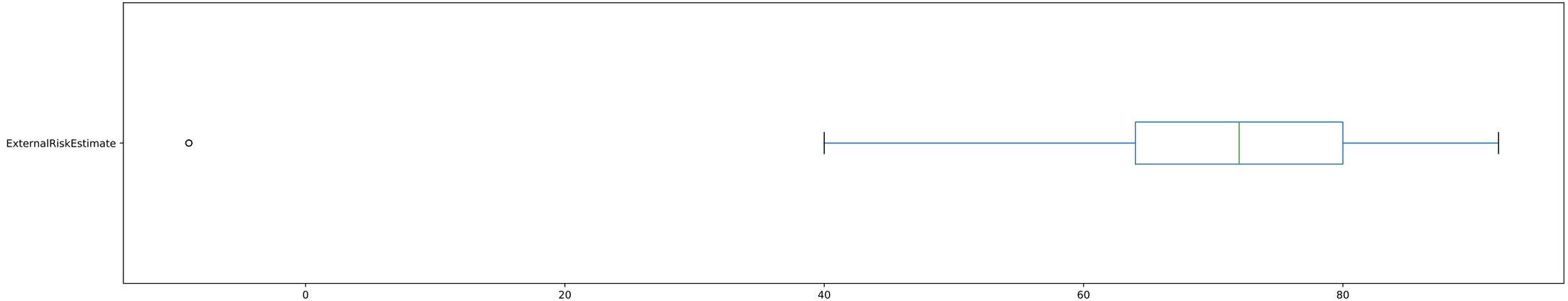


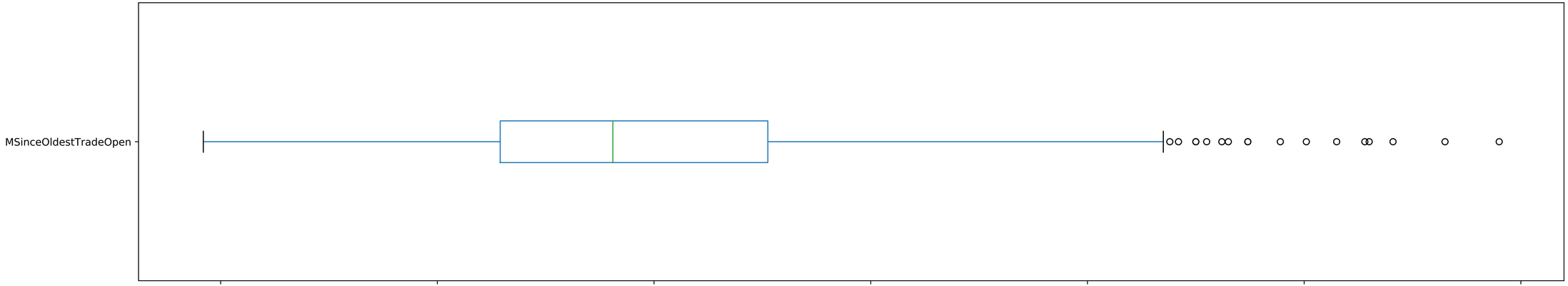
PercentTradesNeverDelq

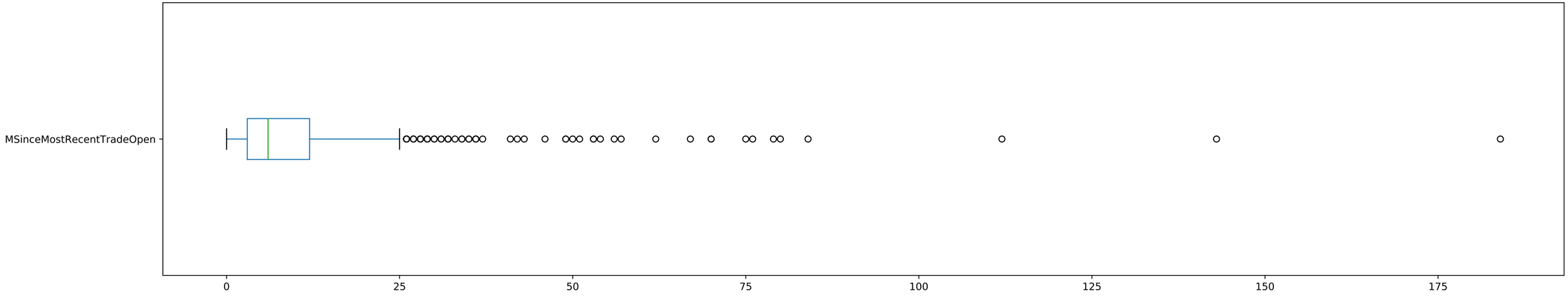


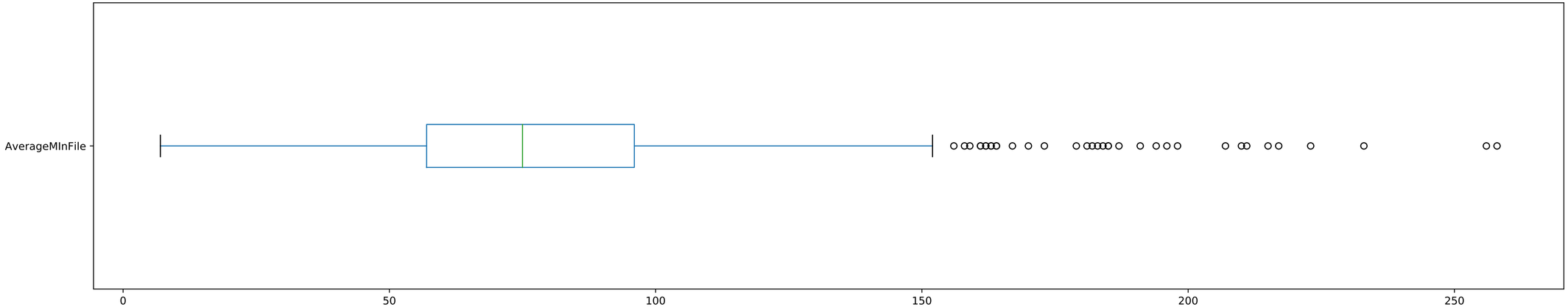
PercentTradesWBalance

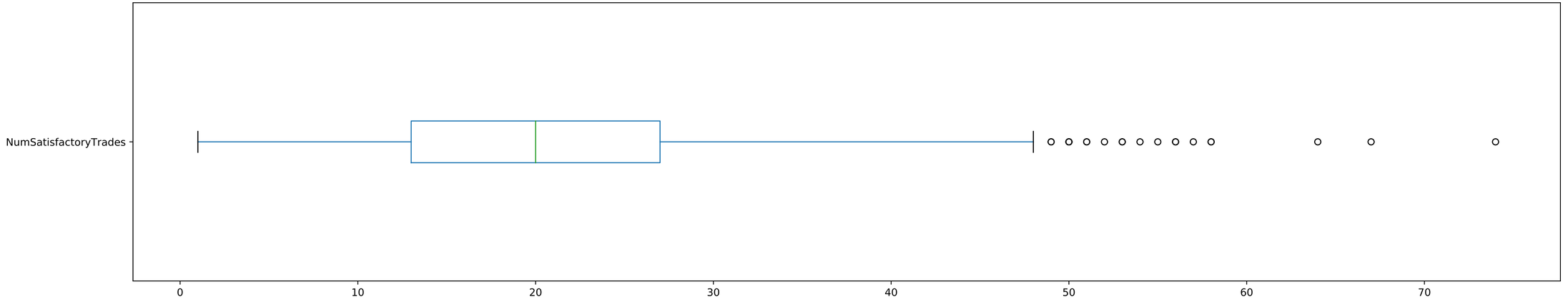


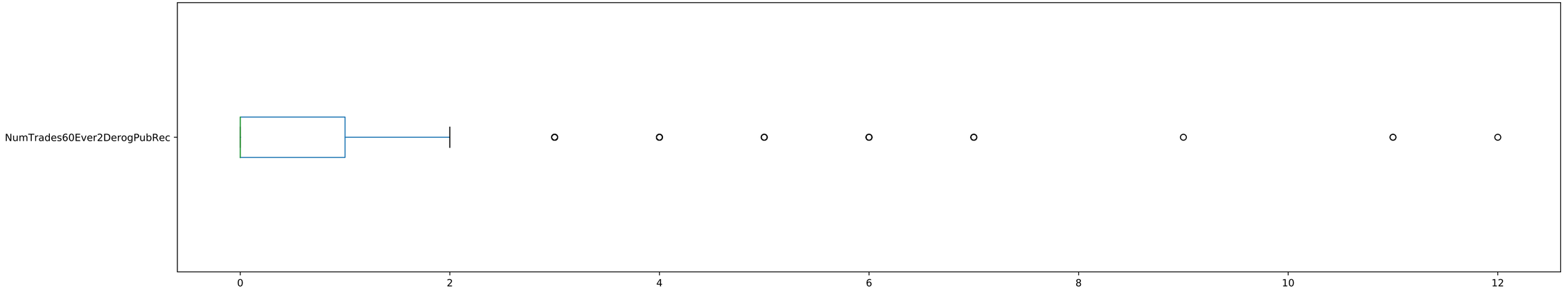

ExternalRiskEstimate
0
20
40
60
80

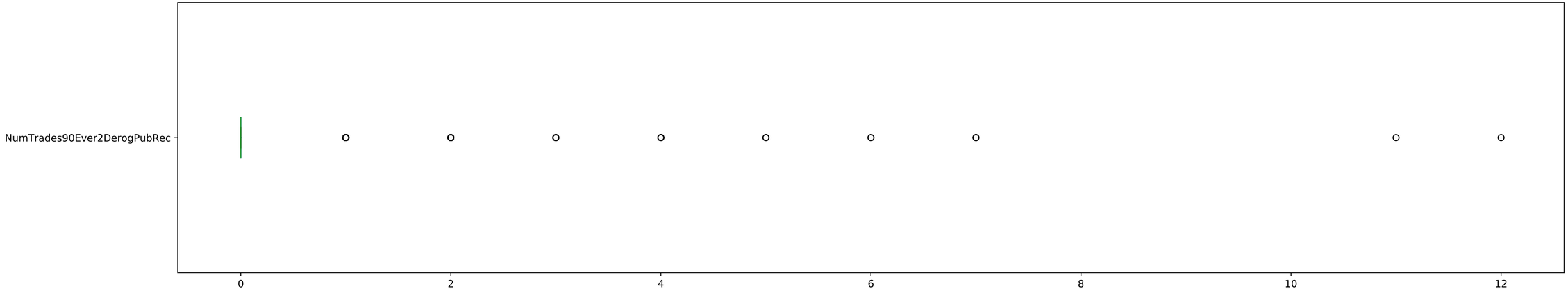

MSinceOldestTradeOpen
0
100
200
300
400
500
600

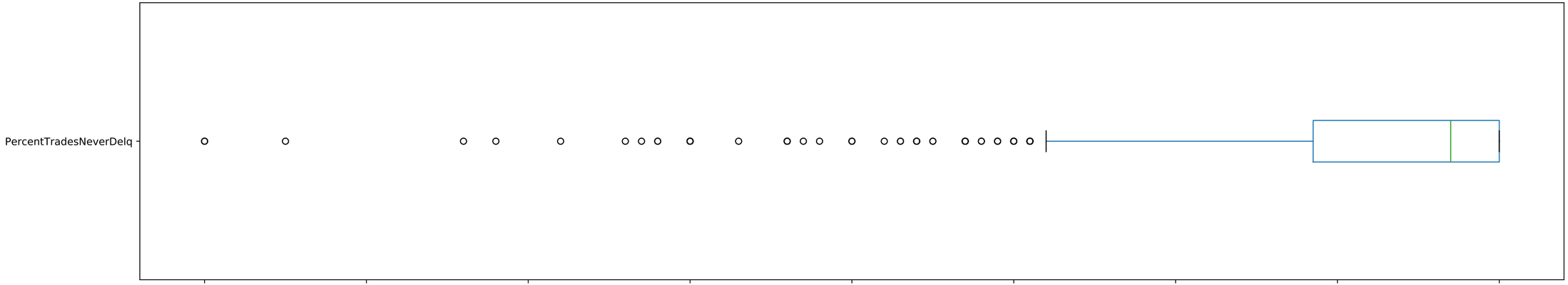

MSinceMostRecentTradeOpen
0
25
50
75
100
125
150
175

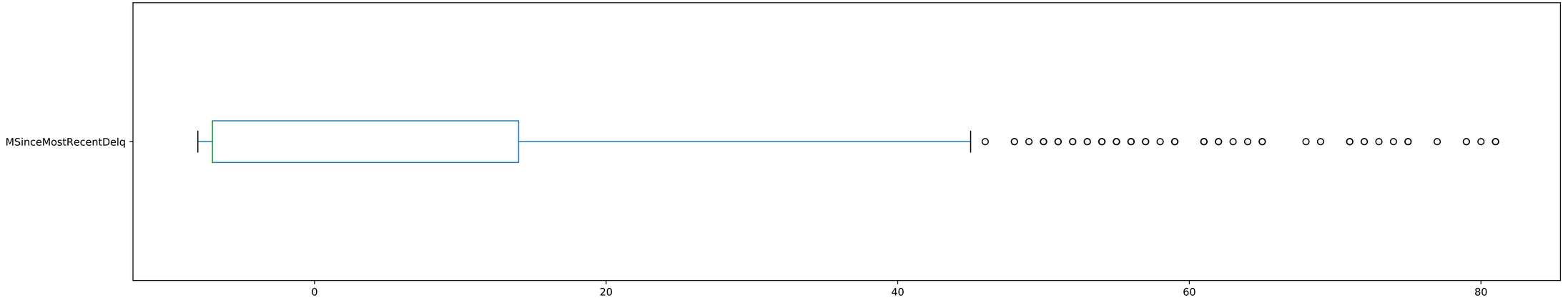

AverageMInFile
0
50
100
150
200
250

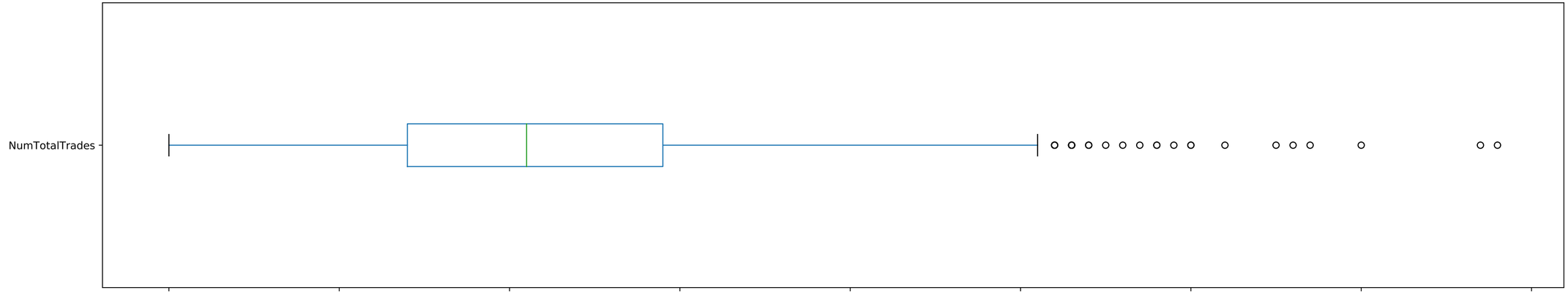

NumSatisfactoryTrades
0
10
20
30
40
50
60
70

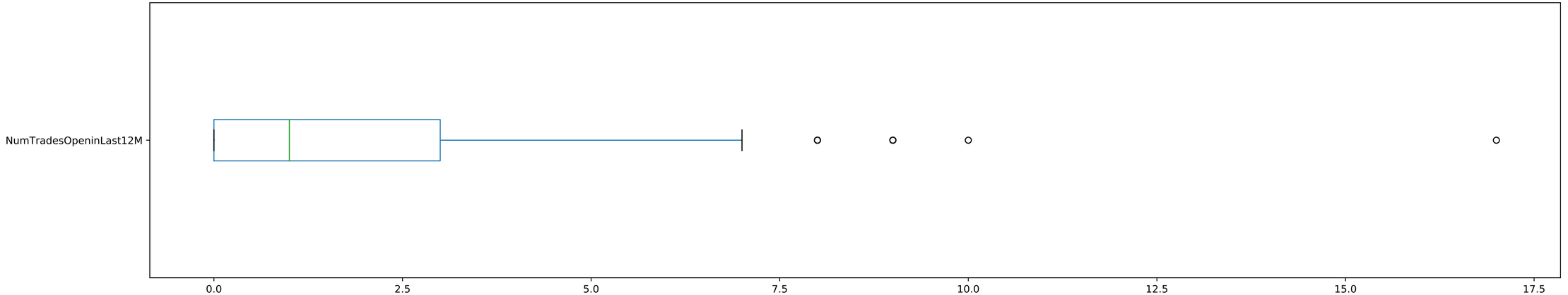

NumTrades60Ever2DerogPubRec
0
2
4
6
8
10
12

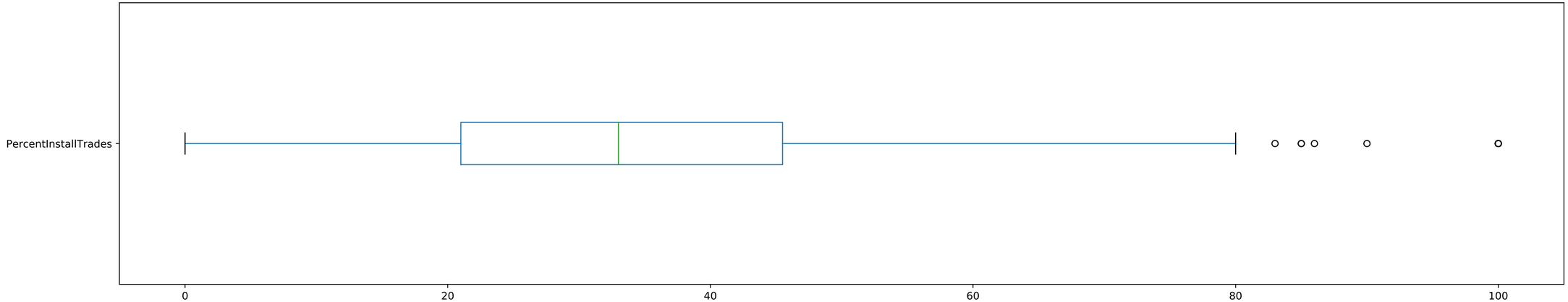

NumTrades90Ever2DerogPubRec
0
2
4
6
8
10
12

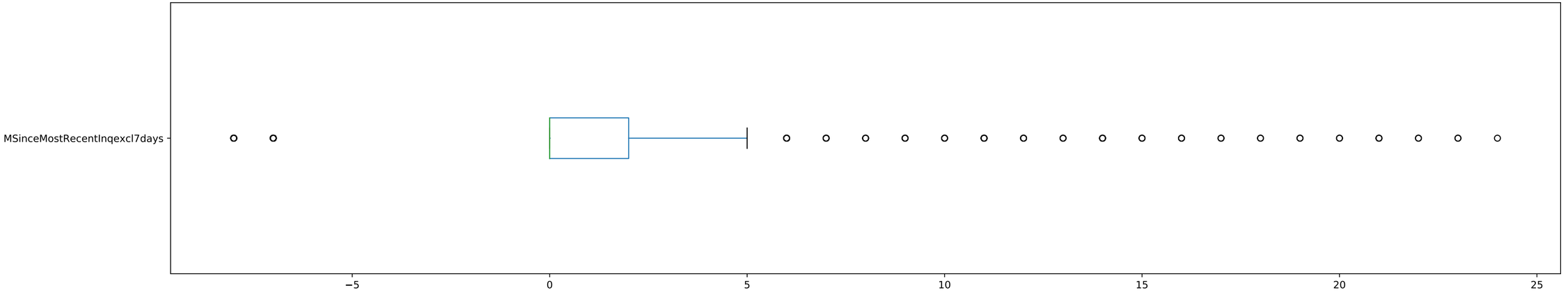

PercentTradesNeverDelq
20
30
40
50
60
70
80
90
100

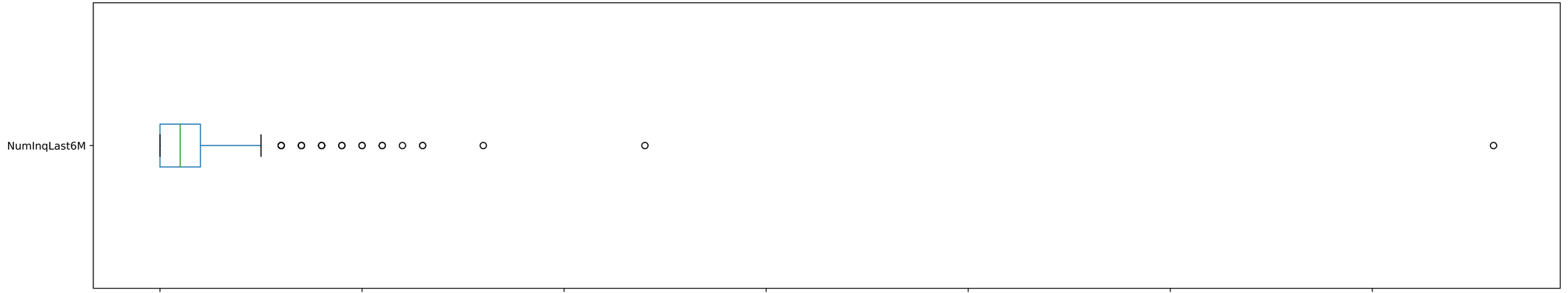

MSinceMostRecentDelq
0
20
40
60
80

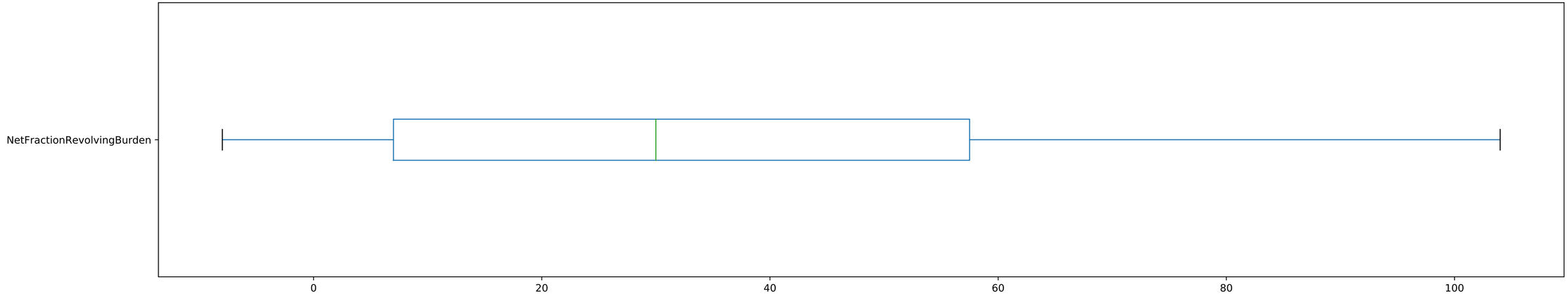

NumTotalTrades
0
10
20
30
40
50
60
70
80

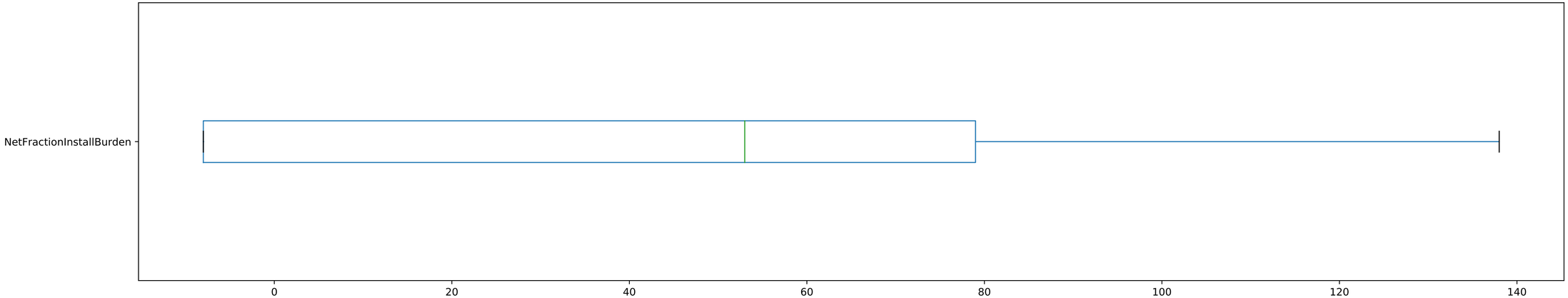

NumTradesOpeninLast12M
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5

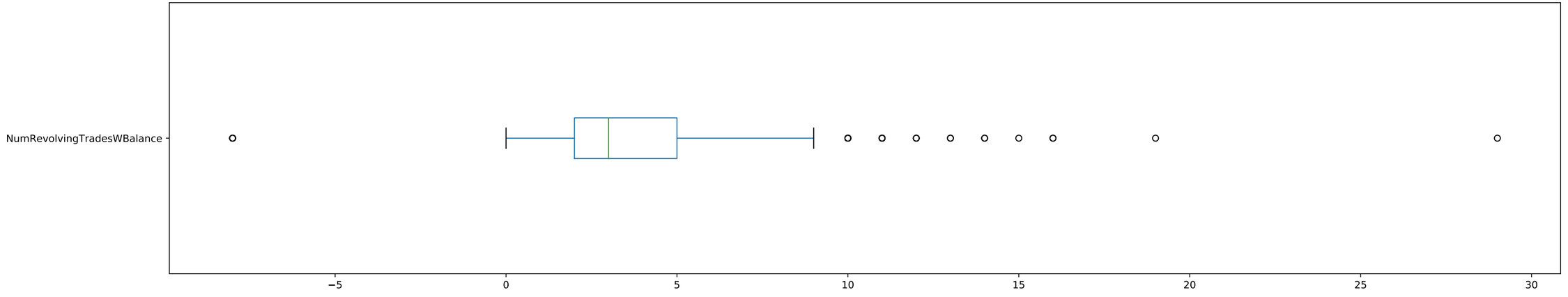

PercentInstallTrades
0
20
40
60
80
100

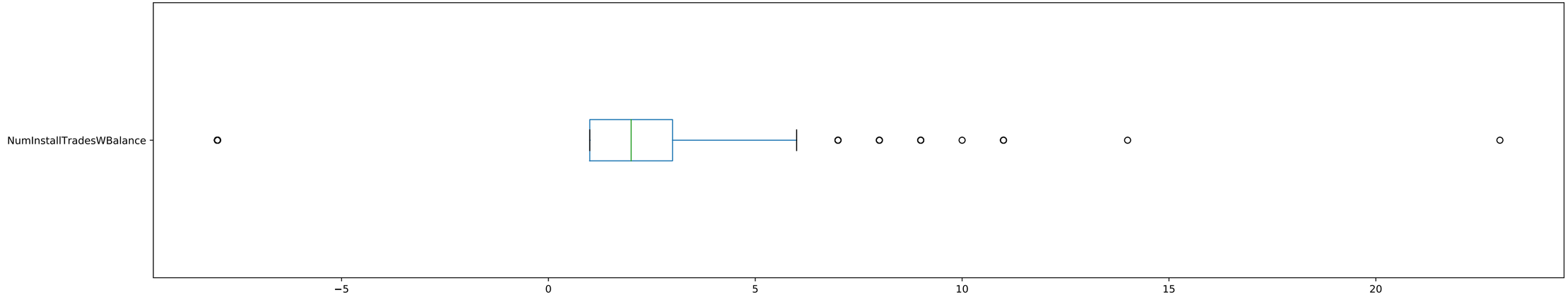

MSinceMostRecentInqexcl7days
−5
0
5
10
15
20
25

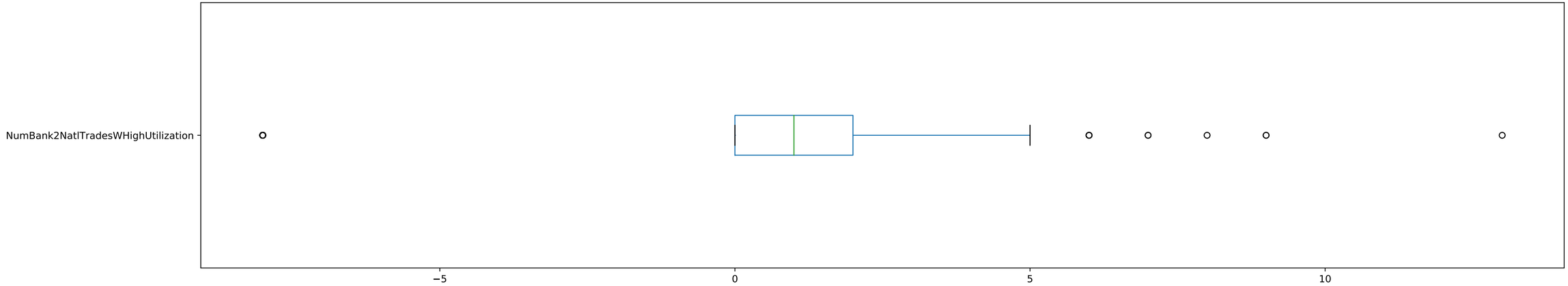

NumInqLast6M
0
10
20
30
40
50
60

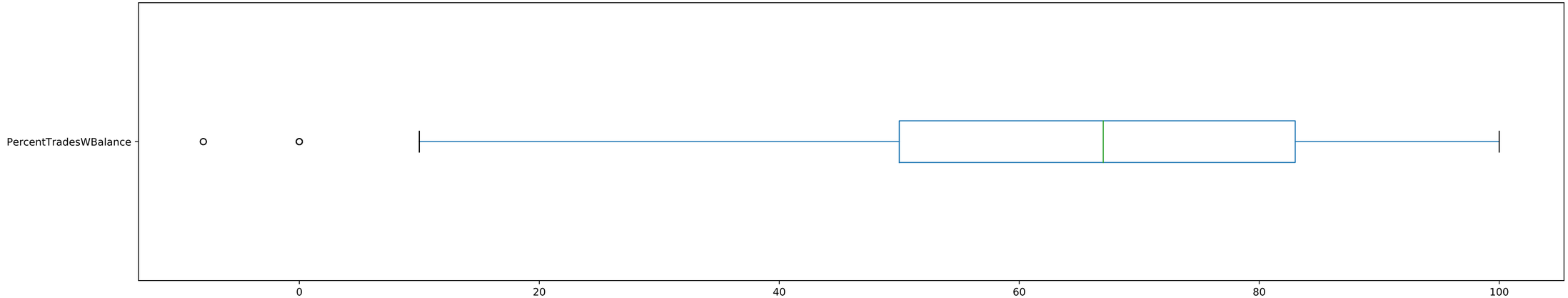

NetFractionRevolvingBurden
0
20
40
60
80
100

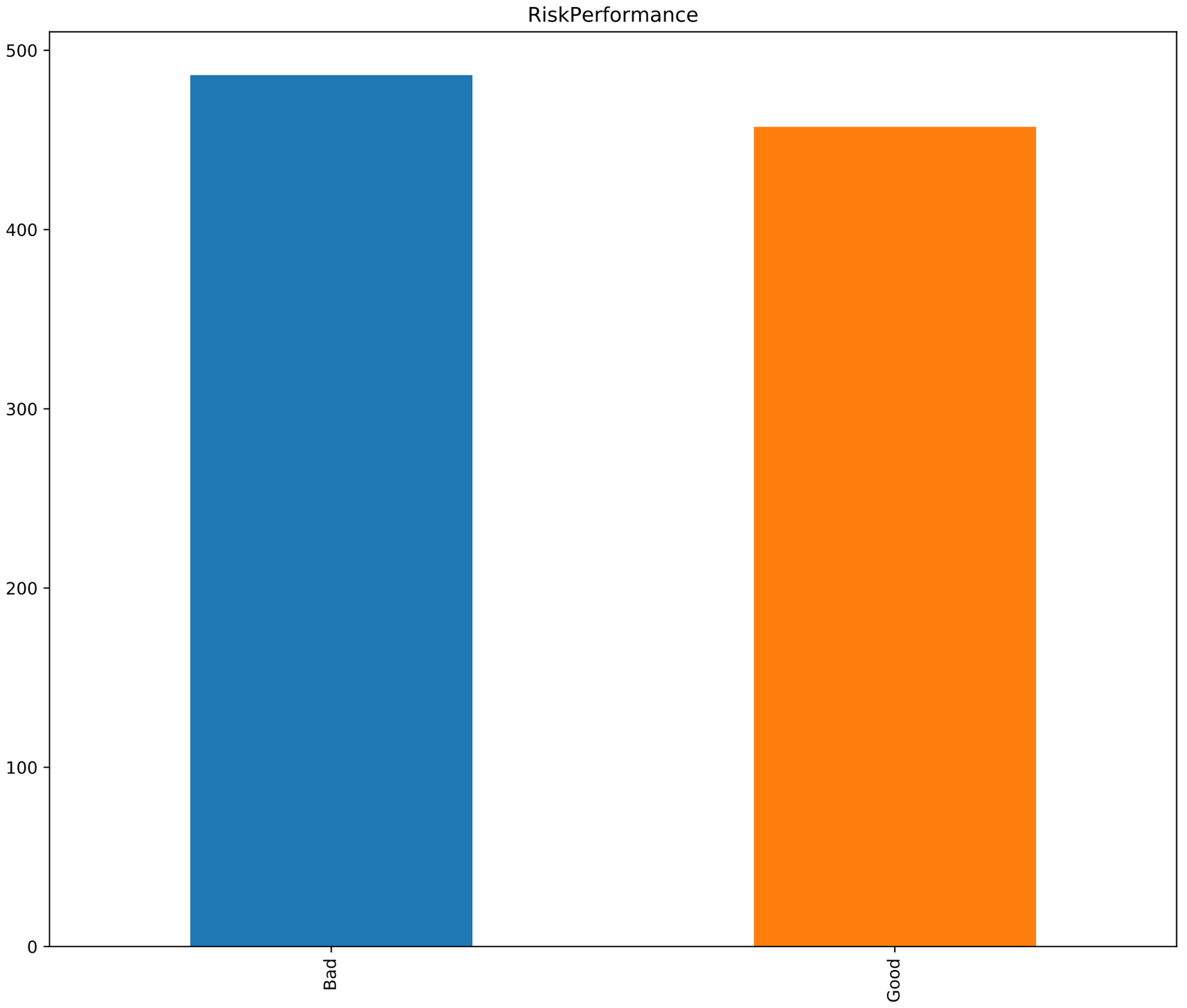

NetFractionInstallBurden
0
20
40
60
80
100
120
140

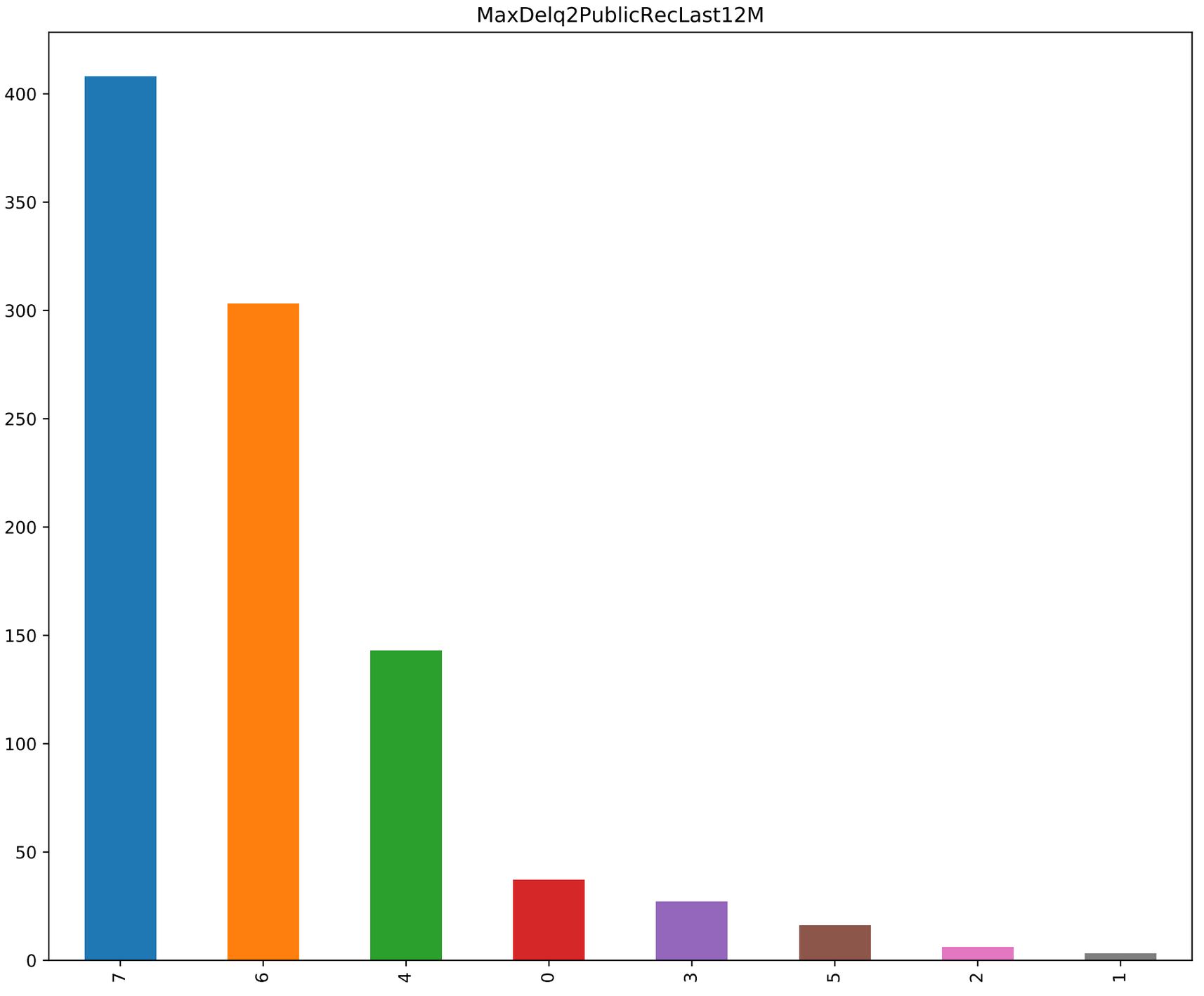

NumRevolvingTradesWBalance
−5
0
5
10
15
20
25
30

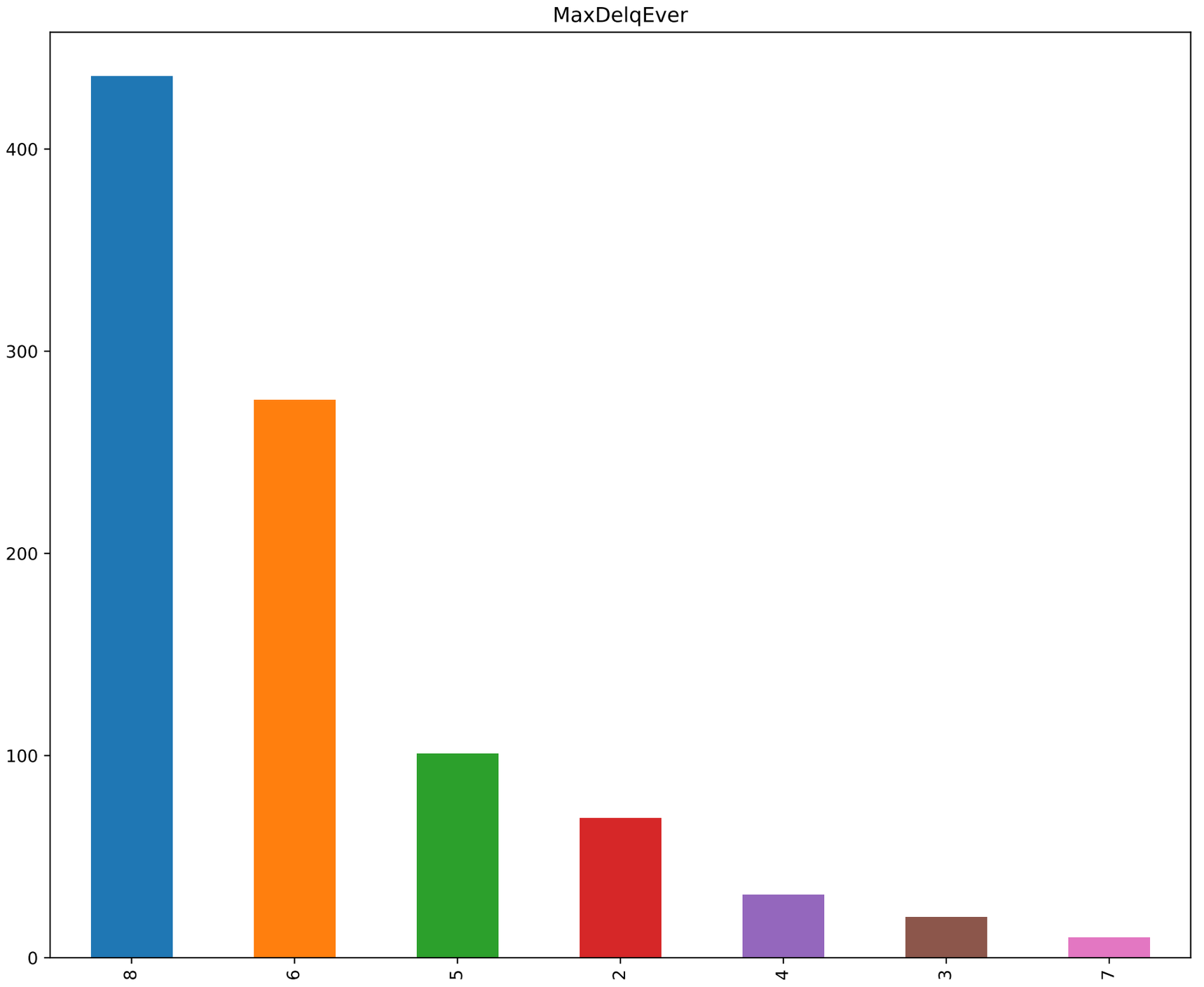

NumInstallTradesWBalance
−5
0
5
10
15
20


NumBank2NatlTradesWHighUtilization
−5
0
5
10


PercentTradesWBalance
0
20
40
60
80
100


RiskPerformance
500
400
300
200
100
0
Bad
Good


MaxDelq2PublicRecLast12M
400
350
300
250
200
150
100
50
0
7
6
4
0
3
5
2
1


MaxDelqEver
400
300
200
100
0
8
6
5
2
4
3
7