Data Quality Report - Initial Findings

1 Descriptive statistics for continuous features

	count	mean	std	min	25%	50%	75%	max
ExternalRiskEstimate	942.0	72.150743	10.039486	37.0	65.0	72.0	80.0	93.0
MSinceOldestTradeOpen	942.0	196.628450	101.707191	-8.0	134.0	183.5	252.0	528.0
MSinceMostRecentTradeOpen	942.0	9.325902	11.255838	0.0	3.0	6.0	12.0	145.0
AverageMInFile	942.0	78.936306	32.789643	5.0	58.0	75.5	98.0	256.0
NumSatisfactoryTrades	942.0	21.399151	11.398674	1.0	13.0	20.0	28.0	70.0
NumTrades60Ever2DerogPubRec	942.0	0.587049	1.202354	0.0	0.0	0.0	1.0	12.0
NumTrades90Ever2DerogPubRec	942.0	0.384289	0.905109	0.0	0.0	0.0	0.0	9.0
PercentTradesNeverDelq	942.0	91.898089	12.294483	0.0	88.0	97.0	100.0	100.0
MSinceMostRecentDelq	942.0	8.498938	20.495340	-8.0	-7.0	1.0	18.0	81.0
NumTotalTrades	942.0	23.289809	13.362273	0.0	14.0	21.0	31.0	85.0
NumTradesOpeninLast12M	942.0	1.817410	1.731313	0.0	0.0	1.0	3.0	11.0
PercentInstallTrades	942.0	35.298301	17.510804	0.0	22.0	33.0	46.0	100.0
MSinceMostRecentInqexcl7days	942.0	0.004246	5.912754	-8.0	-7.0	0.0	1.0	24.0
NumInqLast6M	942.0	1.450106	1.871299	0.0	0.0	1.0	2.0	16.0
NumInqLast6Mexcl7days	942.0	1.397028	1.834697	0.0	0.0	1.0	2.0	16.0
NetFractionRevolvingBurden	942.0	33.200637	29.168736	-8.0	7.0	27.0	55.0	112.0
NetFractionInstallBurden	942.0	41.003185	41.344217	-8.0	-8.0	49.5	79.0	142.0
NumRevolvingTradesWBalance	942.0	3.832272	3.339058	-8.0	2.0	3.0	5.0	20.0
NumInstallTradesWBalance	942.0	1.553079	3.510550	-8.0	1.0	2.0	3.0	14.0
NumBank2NatlTradesWHighUtilization	942.0	0.639066	2.535881	-8.0	0.0	0.5	2.0	18.0
PercentTradesWBalance	942.0	66.690021	22.626372	0.0	50.0	67.0	83.0	100.0

Figure 1: continuous features descriptive statistics. Imported from Jupyter Notebook

We have 942 observations for all the continuous features, after having dropped 58 rows containing duplicates with “no bureau record” values.

- **ExternalRiskEstimate** feature represents a “consolidated version of risk markers”, as taken from a consumer credit bureau report. The median appears to be very close to the mean (~72), so that the feature might likely be distributed symmetrically. The standard deviation is around 10, the minimum value is 37 and the maximum 93. The feature has a “monotonically decreasing” constraint, meaning that it is negatively correlated with the probability of a bad target as observed in the data.

- **MSinceOldestTradeOpen** refers to the number of months since the oldest trade (that we assume as each single agreement between the credit institution and the consumer/borrower) was opened. The feature minimum is a special value (-8) referring to no usable or valid trade, and should be addressed. The maximum (528) is likely to be an outlier, as the interquartile range spans from 134 to 252, with an average of approximately 197 months.
- **MSinceMostRecentTradeOpen** refers to the number of months since the most recent trade was opened. The maximum (145) is very far from the rest of the data and might be an outlier. It is interesting to note that the majority of the data (75%) shows a value of less than a year.
- **AverageMInFile** refers to the number of average months in the credit bureau file. Since half of the feature values are concentrated between 58 and 98, both the minimum and the maximum (respectively 5 and 256) should be treated as outliers.
- **NumSatisfactoryTrades** refers to the number of trades considered satisfactory. It appears to be symmetrically as the mean (~21) and the median (20) are close. Minimum and maximum might be outliers and should be investigated.
- **NumTrades60Ever2DerogPubRec** refers to the number of trade lines that record a 60 days past-due payment. This feature presents zero-values for a good portion of observations, as the 75% of records shows less than 1 past-due trade line in the last two months of watch period.
- **NumTrades90Ever2DerogPubRec** refers to the number of trade lines that record a 90 days past-due payment. Similarly to the previous feature, zero-values are by far the most frequent, even further in this case. The observations related to this feature and the previous one might mean that, since the “good” and “bad” scores are quite balanced (more on that later), the vast majority of past-due payments have occurred before the last three months of the watch period.
- **PercentTradesNeverDelq** refers to the percentage of trade lines that have never been delinquent. The vast portion of observations show a high percentage, as the 75% has a percentage higher than 88% of non-delinquent trade lines – having a maximum value of 100%. This should be investigated, as the target feature *RiskPerformance* is quite evenly distributed between “good” and “bad” borrowers.
- **MSinceMostRecentDelq** refers to the number of months passed since the most recent credit delinquency occurred. The minimum (-8) represents no usable/valid trades, so it should be addressed, as well as the maximum value (81) that is an outlier, as the 75% of observations show values smaller than 18 (1.5 years). A good portion of observations show a value of -7, indicating that there have been no delinquencies.
- **NumTotalTrades** refers to the total amount of trades associated with an individual. The interquartile range spans from 14 to 31, and the maximum value might be an outlier (85). The presence of a minimum value of 0 indicates the presence in the dataset of individuals with no associated credit trades.
- **NumTradesOpeninLast12M** refers to the number of trades opened in the last year. Half of the observations show 1 trade or less, while 75% show less than 3 trades opened in the last year – that may tell us that the dataset refers mostly to credit applications with a remarkable amount of credit history. The maximum (11) is an outlier and should be addressed.
- **PercentInstallTrades** refers to the percentage of trades devoted to installments. The 75% of the observations show a value less than 50.
- **MSinceMostRecentInqexcl7days** refers to the number of months passed since the most recent inquiry has occurred (excluding the last 7 days). The feature shows very low levels, as the 75% of observations have less than 1 month since the most recent

enquiry. Nevertheless, many special values (-7 and -8) indicate that there have been no enquiries at all.

- **NumInqLast6M** and **NumInqLast6Mexcl7days** refer to the amount of inquires carried out respectively in the last 6 months and in the 6 months excluding the last 7 days. These two features show almost identical summary statistics, with a quarter of observations with no inquiries carried out in the last six months – coherently with the previous examined feature (*MSinceMostRecentInqexcl7days*).
- **NetFractionRevolvingBurden** refers to the revolving balance divided by credit limit. The minimum of -8 means that for some observation there is no valid trade to be considered, and the maximum of 112 might be an outlier and should be addressed.
- **NetFractionInstallBurden** refers to the installment balance divided by original loan amount. The 25% of observations show a -8 special value, meaning that there are no valid trades to be considered, whereas the maximum of 142 might be an outlier.
- **NumRevolvingTradesWBalance** and **NumInstallTradesWBalance** refer to the number of revolving trades with balance and the number of installment trades with balance, respectively. They show some special -8 values as well, and their summary statistics are quite similar (even the values are slightly higher for the number of revolving trades).
- **NumBank2NatlTradesWHighUtilization** refers to the number of bank trades characterized by high utilization. The 75% of observations show a reasonable amount of less than 2 high utilization trades, while the maximum value of 18 is certainly an outlier. There are still observations with no valid trades to be considered (minimum level of -8).
- **PercentTradesWBalance** refers to the percentage of trades with balance. The good majority (more than 75%) of the observations show a level higher than 50, with an average level of ~67.

2 Descriptive statistics for categorical features

	count	unique	top	freq
RiskPerformance	942	2	Bad	490
MaxDelq2PublicRecLast12M	942	8	7	388
MaxDelqEver	942	7	8	416

Figure 2: categorical features descriptive statistics. Imported from Jupyter Notebook

We have all the 942 observations for the categorical features as well, after having dropped 58 rows containing duplicates with “no bureau record” values.

- Our target feature **RiskPerformance** is binary since it has two unique values (“good” and “bad”). Since the dataset has been obtained from a stratified random sample of records (each record representing a loan applicant), an almost equal number of accounts with good and bad payment records have been selected – considering a two-year observation window. In fact, on 942 observations there are 490 “bad” applicants (relative frequency of ~52%).

The next two categorical features contain information about the number of days in which an applicant has been considered as delinquent on a loan. Since they are categorized in "levels" - corresponding each to an interval of 30 days – we considered correct to treat them as categorical features, as they can take a fixed number of possible values.

- **MaxDelq2PublicRecLast12M** refers to the maximum delinquency records over the last 12 months. It comes in 8 unique levels spanning from 0 to 7. The majority of observations falls in level 7, consisting of no delinquency recorded over the last 12 months (relative frequency of 41%).
- **MaxDelqEver** refers to the maximum delinquency ever recorded. It comes in 7 unique levels and the most frequent level is 8 – meaning that no delinquency has ever been recorded (relative frequency of 44%).

3 Histograms for continuous features

The plots are attached at the end of the file.

A good portion of plots shows an exponential distribution, greatly affected by negative and zero values, so that for a relevant number of observations there are no records available or any signaled delinquency. Their behavior appears thus to be exponentially decreasing. Examples of features affected by such characteristics are:

- **MSinceMostRecentDelq**
- **MSinceMostRecentTradeOpen**: exceptional few levels above 60, might be outliers.
- **NumInqLast6M**
- **NumInqLast6Mexcl7days**
- **NumTrades60Ever2DerogPubRec** and **NumTrades90Ever2DerogPubRec**: the features show a very similar distribution. They are likely related, as consecutive measures of "defaulting" trades over time levels. The latter contains the former (90 days > 60 days).
- **NumTradesOpeninLast12M**
- **NumBank2NatlTradesWHighUtilization** and **NumInstallTradesWBalance**: moderate frequency of 0 and negative special values. The rest of the values appear distributed in an exponentially decreasing way.
- **NumRevolvingTradesWBalance**: moderate presence of special values.

Remarkably enough, among exponential-shaped distributions, only one feature shows exponentially increasing behavior:

- **PercentTradesNeverDelq**: the vast majority of observations show a percentage of trades never delinquent close to 100. This could represent a fact that needs to be furtherly investigated, as only a rough half of the rows in the dataset show a value of "bad" in the target feature. From this plot it could seem that our dataset is populated by good borrowers, on average.

Some features show a bimodal distribution, as a great portion of their observation is affected by zero or special values:

- **MsinceMostRecentInqexcl7days**: very few levels above 10, long right-tail. Moderate portion of negative special values, the rest of observation has exponentially decreasing distribution.
- **NetFractionInstallBurden**: the distribution appears to be left-skewed even though around one third of observations shows negative special values (-8), meaning that no trades can be considered. There are few outliers in the upper levels as well.

Some features show a right-skewed distribution:

- **NetFractionRevolvingBurden**: mode centered around zero values.
- **NumSatisfactoryTrades** and **NumTotalTrades** are distributed in a very similar way. A possible reason could be that the former is a subset of the latter.
- **MSinceOldestTradeOpen**: slight skewness towards the right tail (towards very high "trade-seniority").

One feature shows a left-skewed distribution:

- **PercentTradesWBalance**: few levels fall under the 50 value, so the vast majority of applicants in the dataset appear to had credit agreements (i.e. trades) with some sort of balance.

Some features show a quasi-normal distribution:

- **AverageMInFile**: slightly right-skewed.
- **ExternalRiskEstimate**: some outliers in the lower tail, characterized by low credit scores.
- **PercentInstallTrades**: with a slightly longer right-tail.

4 Box-plots for continuous features

The plots are attached at the end of the file.

Many box-plots show a relevant amount of outliers. In fact, some features are affected by extreme outliers, distant from the mean with a factor of multiple standard deviation measures.

Examples of such features are: **MSinceMostRecentTradeOpen** (almost flattened box, outliers great in number), **AverageMInFile**, **NumTrades60Ever2DerogPubRec**, **NumTrades90Ever2DerogPubRec** (flattened box, some extreme outliers), **PercentTradesNeverDelq**, **NumInqLast6M**, **NumInqLast6Mexcl7days**.

NumRevolvingTradesWBalance, **NumInstallTradesWBalance** and **NumBank2NatlTradesWHighUtilization** are affected by negative special values.

The great amount of outliers in the majority of the dataset features are indicators of some heterogeneity in the data that we should keep in consideration in the modeling phase.

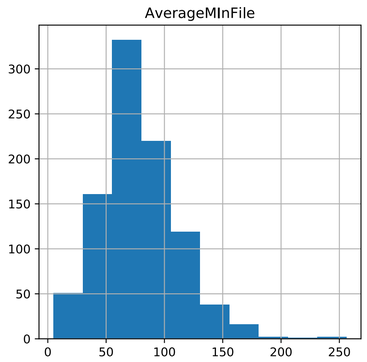
5 Bar charts for categorical features

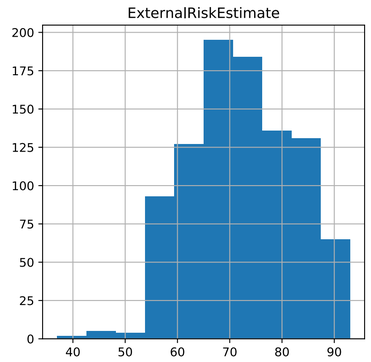
The plots are attached at the end of the file.

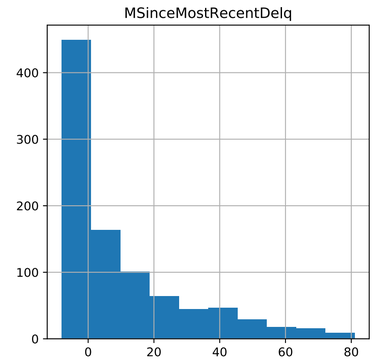
RiskPerformance binary feature shows an evenly distributed bar chart; this is an effect of the aforementioned technique of stratified random sampling through which the dataset was obtained. The frequency of good credit applicants is thus comparable to the frequency of bad credit applicants.

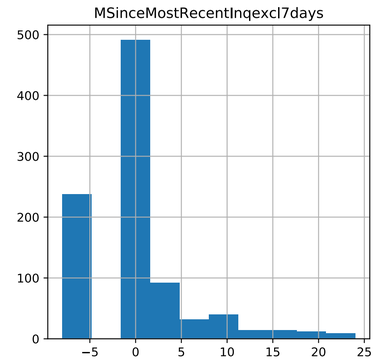
MaxDelq2PublicRecLast12M and **MaxDelqEver** are related features as they differ only in the time measurement interval. They are dominated by “current and never delinquent” observations, accounting for more than one third. It is interesting to note that for the first feature, the second most frequent level corresponds to “unknown delinquency” while for **MaxDelqEver** it corresponds to “30-days delinquent”. This could mean two opposite things:

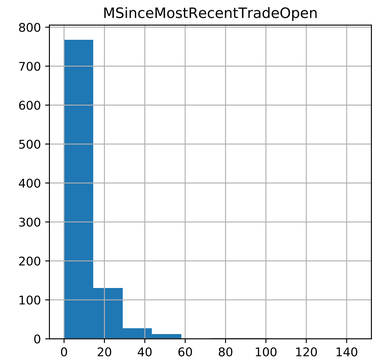
- many delinquency episodes are older than 1 year
- the **MaxDelq2PublicRecLast12M** feature is flawed by a large amount of “unknown delinquency” levels that might well represent delinquencies (false negatives).

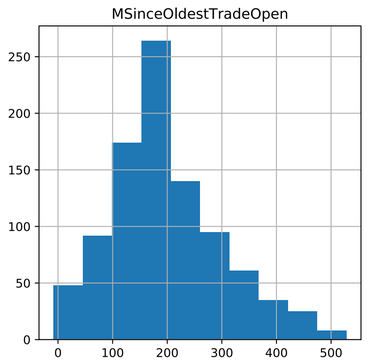

AverageMInFile

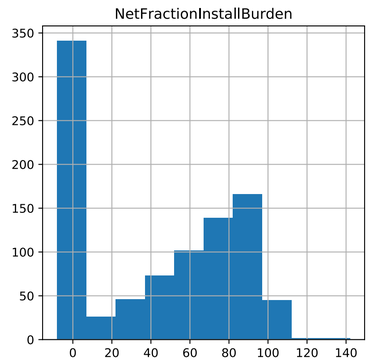

ExternalRiskEstimate

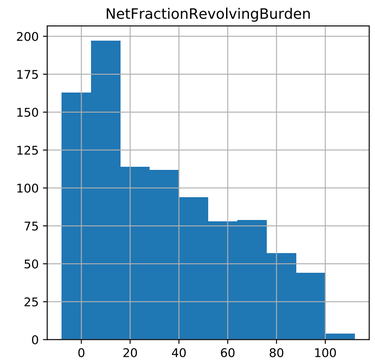

MSinceMostRecentDelq

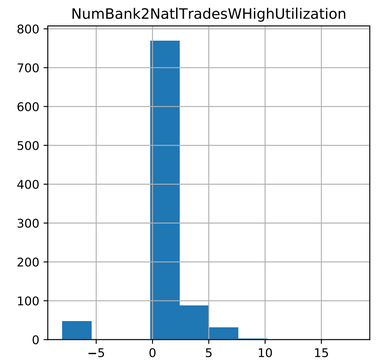

MSinceMostRecentInqexcl7days

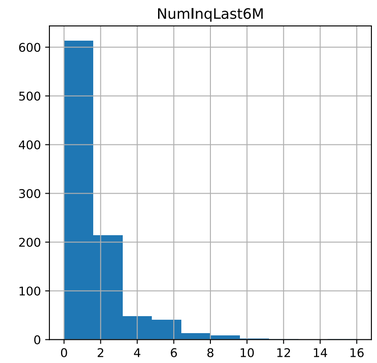

MSinceMostRecentTradeOpen

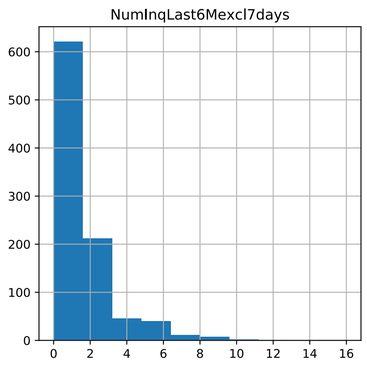

MSinceOldestTradeOpen

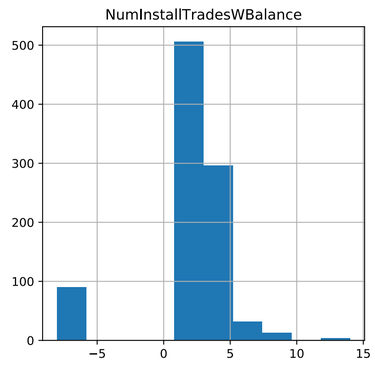

NetFractionInstallBurden

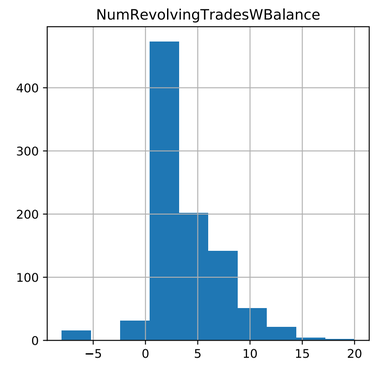

NetFractionRevolvingBurden

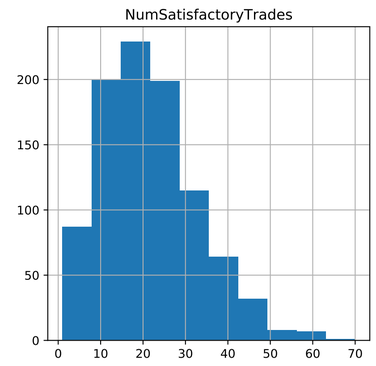

NumBank2NatlTradesWHighUtilization

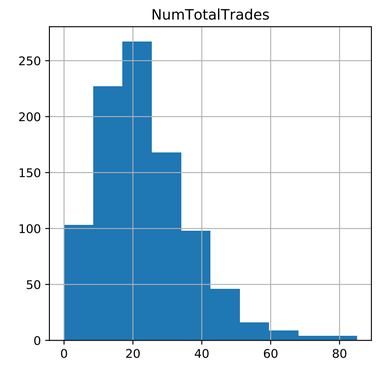

NumInqLast6M

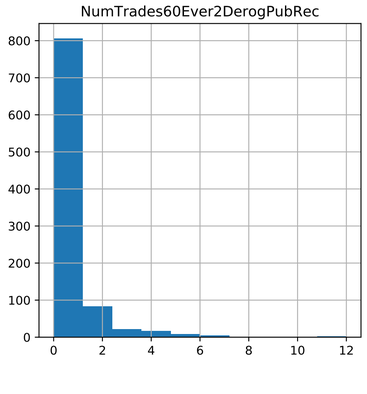

NumInqLast6Mexcl7days

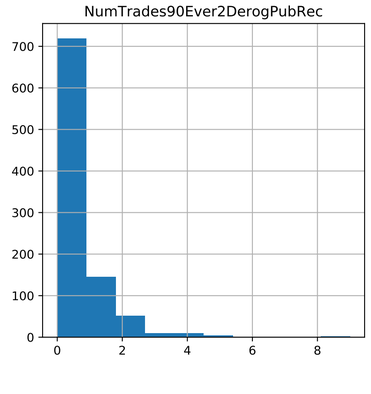

NumInstallTradesWBalance

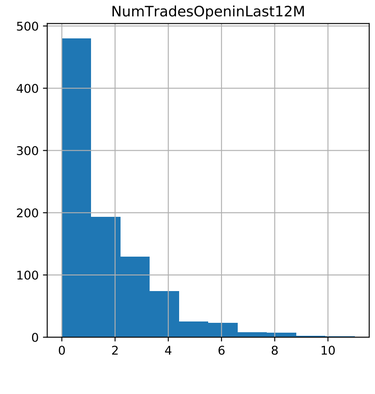

NumRevolvingTradesWBalance

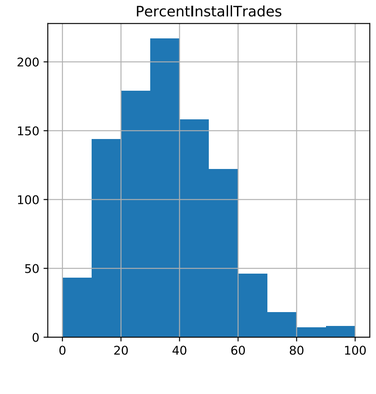

NumSatisfactoryTrades

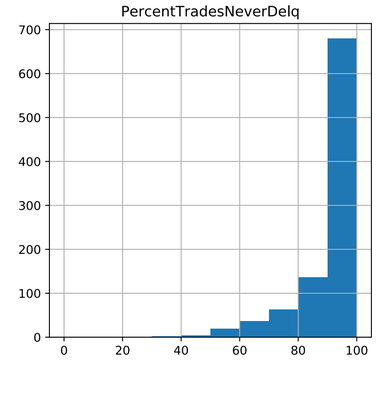

NumTotalTrades

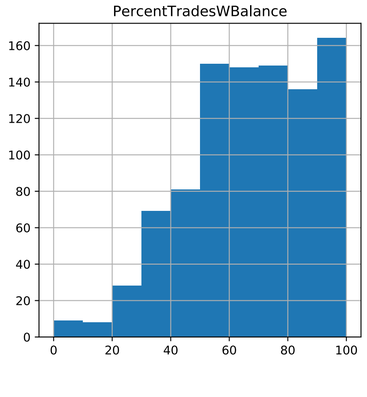

NumTrades60Ever2DerogPubRec

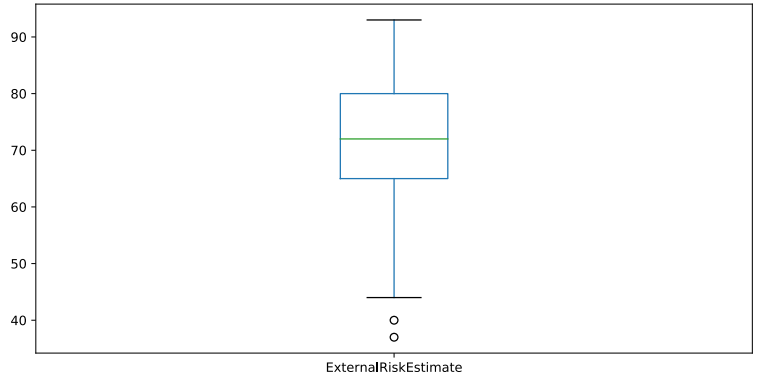

NumTrades90Ever2DerogPubRec

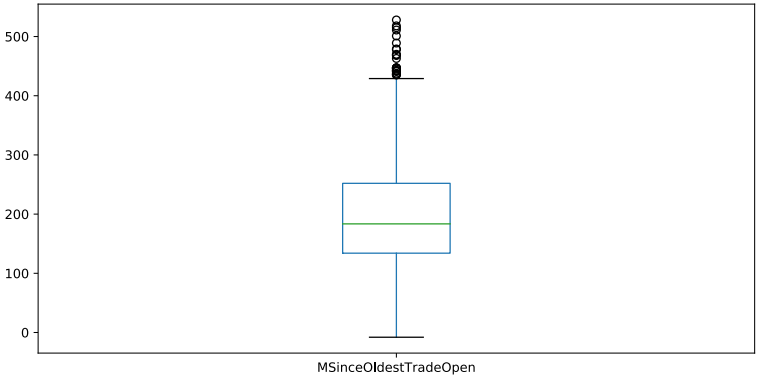

NumTradesOpeninLast12M

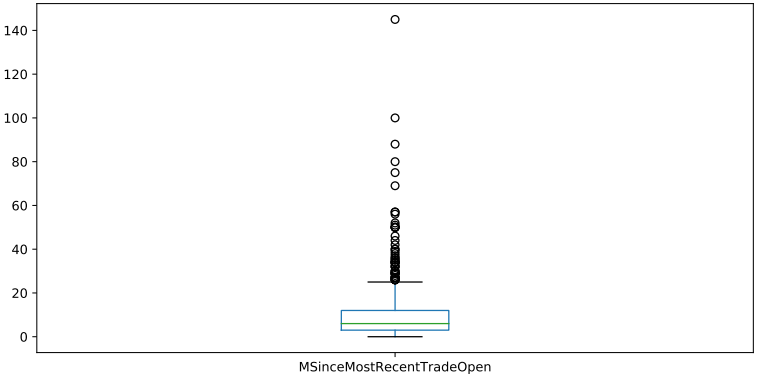

PercentInstallTrades

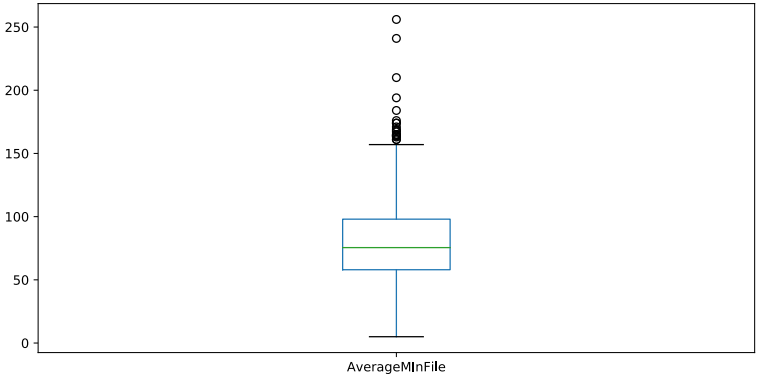

PercentTradesNeverDelq

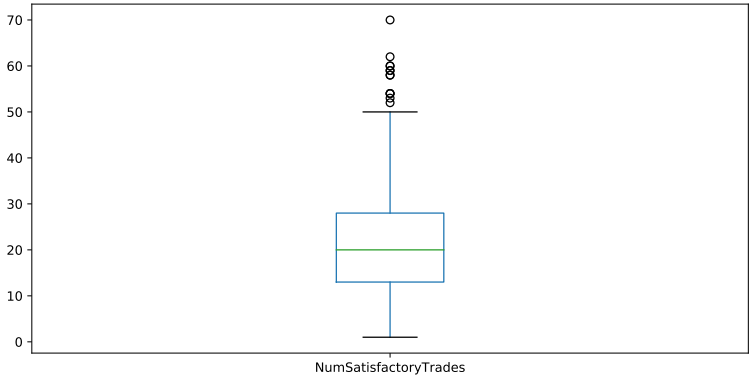

PercentTradesWBalance

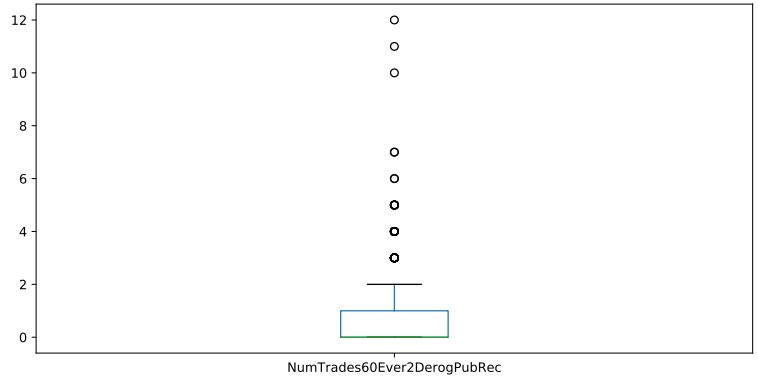

90
80
70
60
50
40
ExternalRiskEstimate

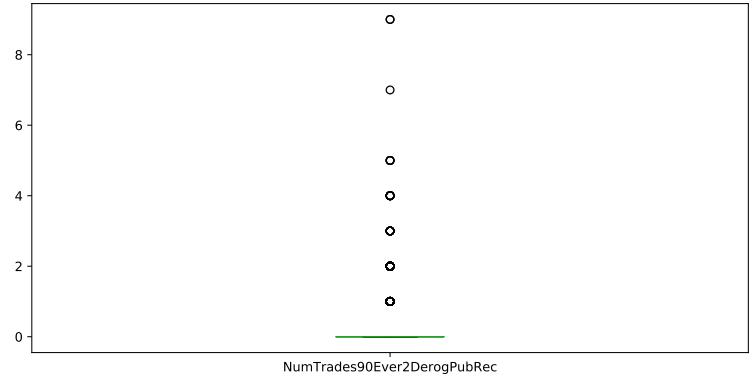

500
400
300
200
100
0
MSinceOldestTradeOpen

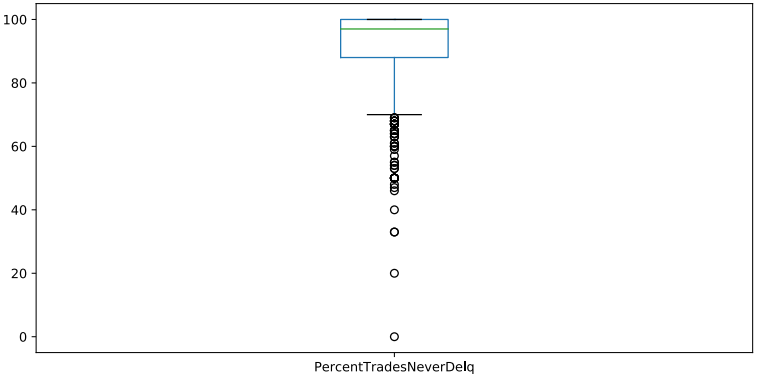

140
120
100
80
60
40
20
0
MSinceMostRecentTradeOpen

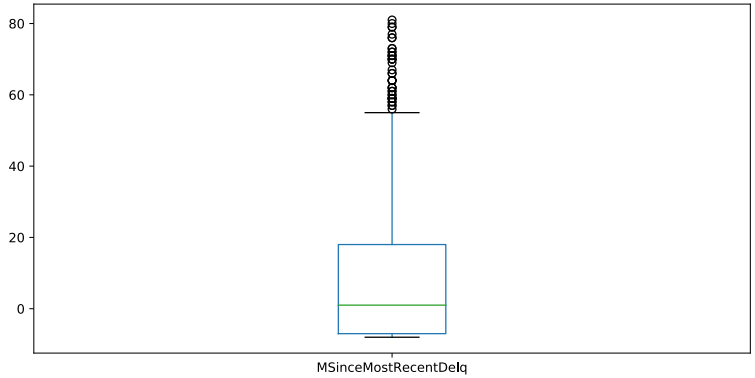

250
200
150
100
50
0
AverageMInFile

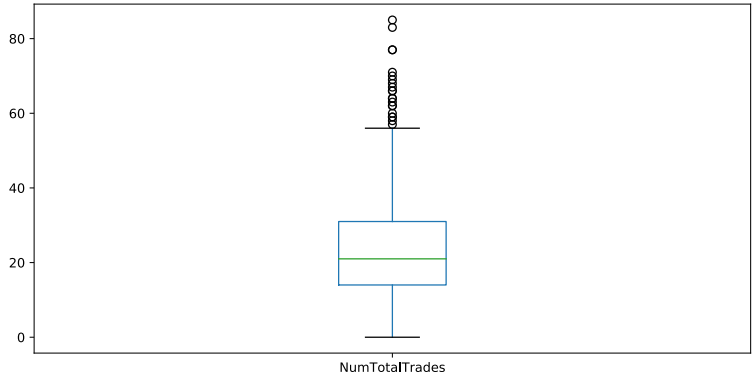

70
60
50
40
30
20
10
0
NumSatisfactoryTrades

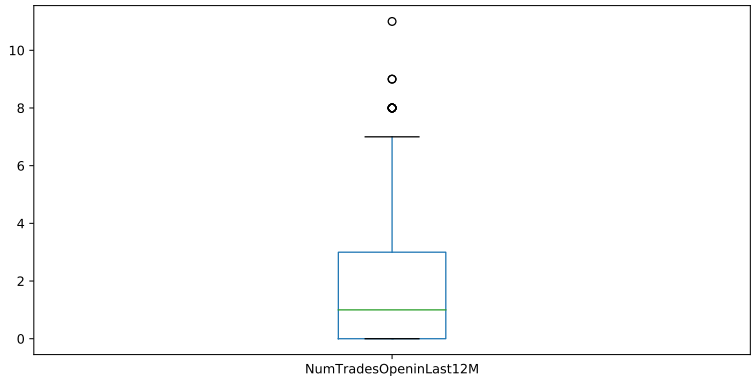

12
10
8
6
4
2
0
NumTrades60Ever2DerogPubRec

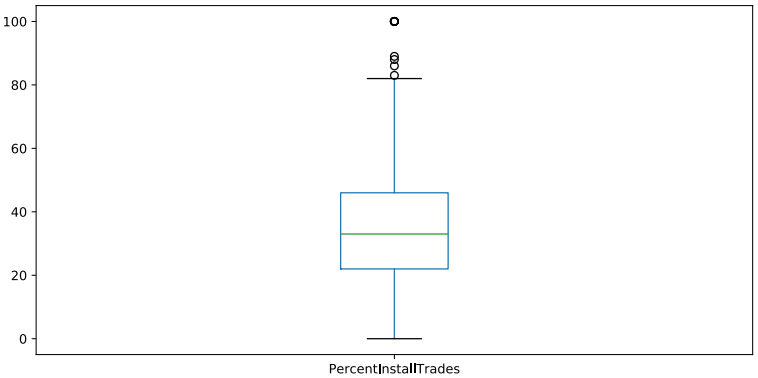

8
6
4
2
0
NumTrades90Ever2DerogPubRec

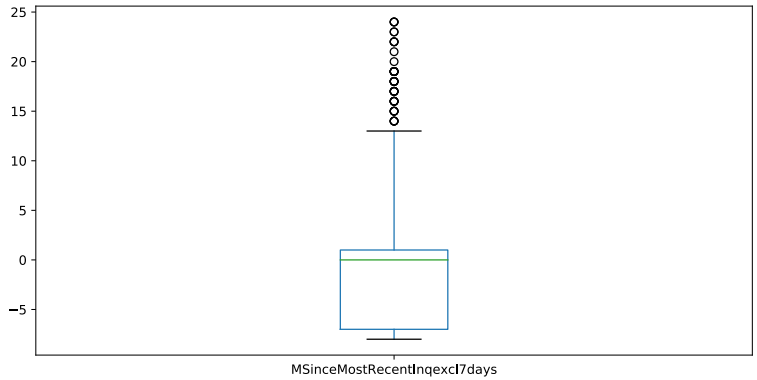

100
80
60
40
20
0
PercentTradesNeverDelq

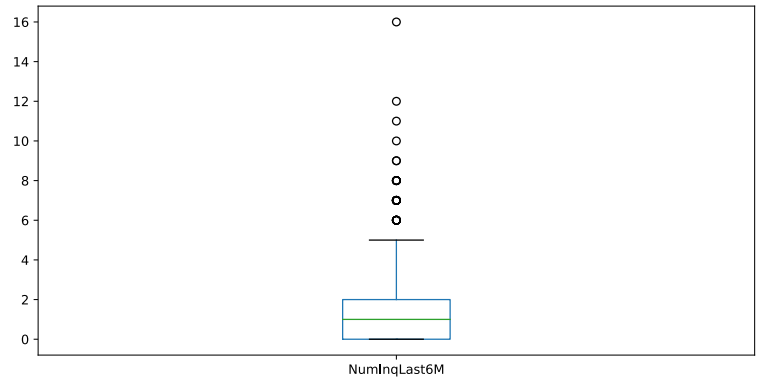

80
60
40
20
0
MSinceMostRecentDelq

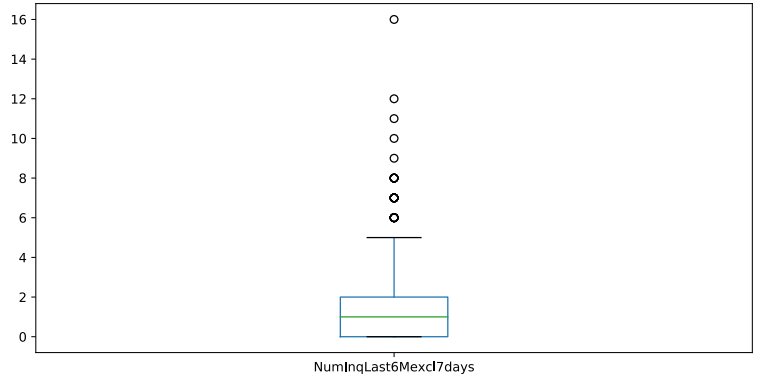

80
60
40
20
0
NumTotalTrades

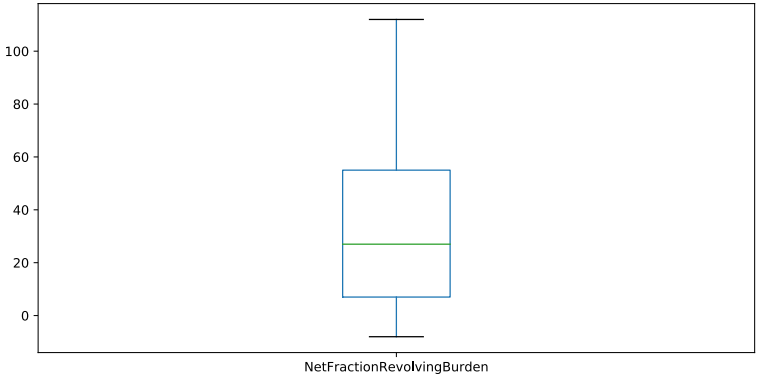

10
8
6
4
2
0
NumTradesOpeninLast12M

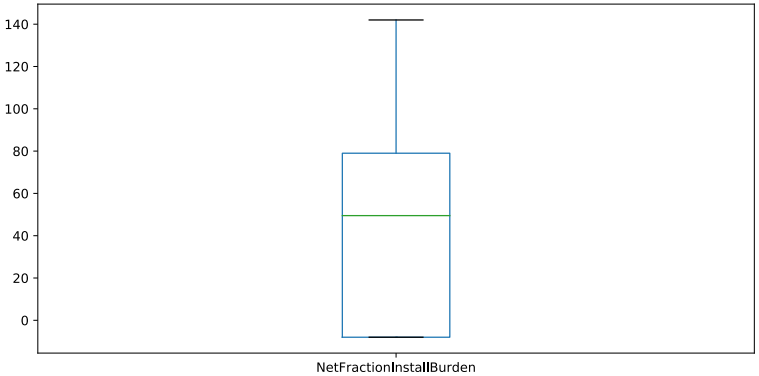

100
80
60
40
20
0
PercentInstallTrades

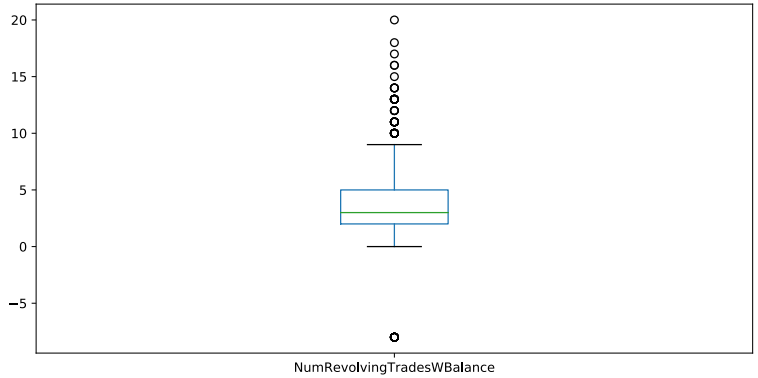

25
20
15
10
5
0
−5
MSinceMostRecentInqexcl7days

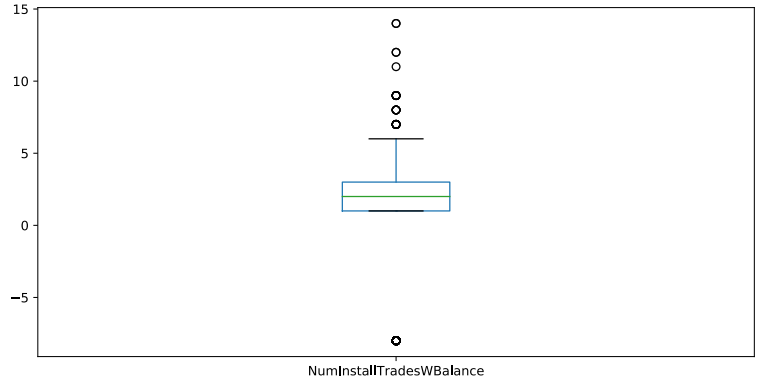

16
14
12
10
8
6
4
2
0
NumInqLast6M

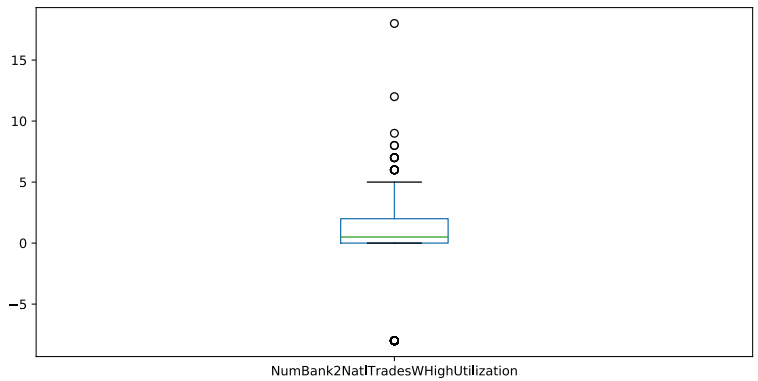

16
14
12
10
8
6
4
2
0
NumInqLast6Mexcl7days

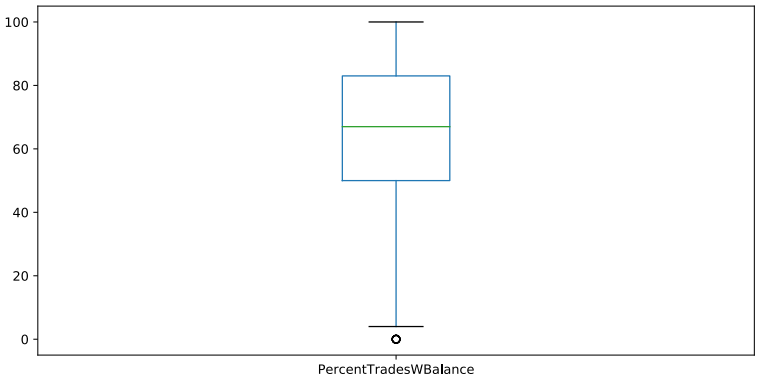

100
80
60
40
20
0
NetFractionRevolvingBurden

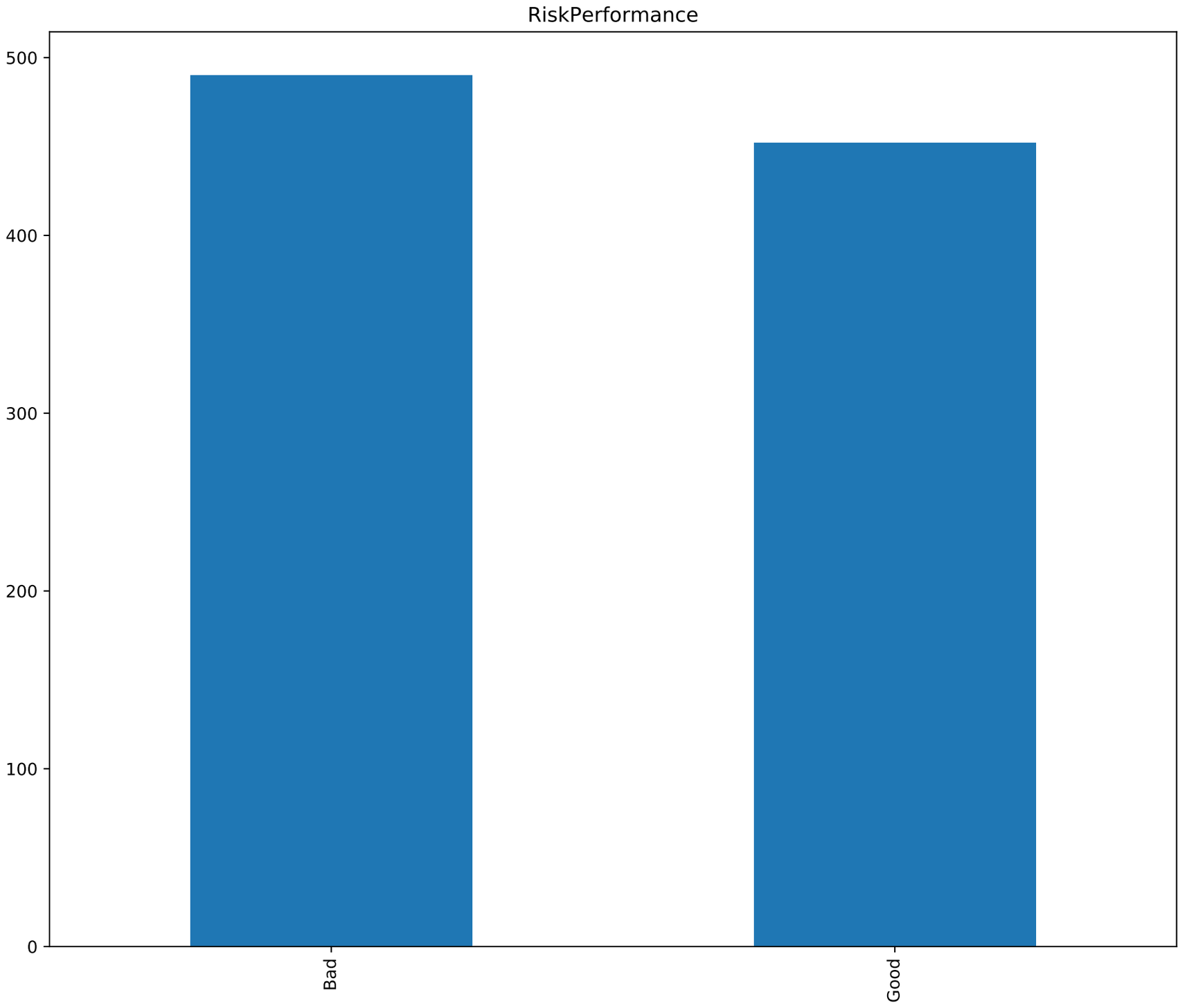

140
120
100
80
60
40
20
0
NetFractionInstallBurden

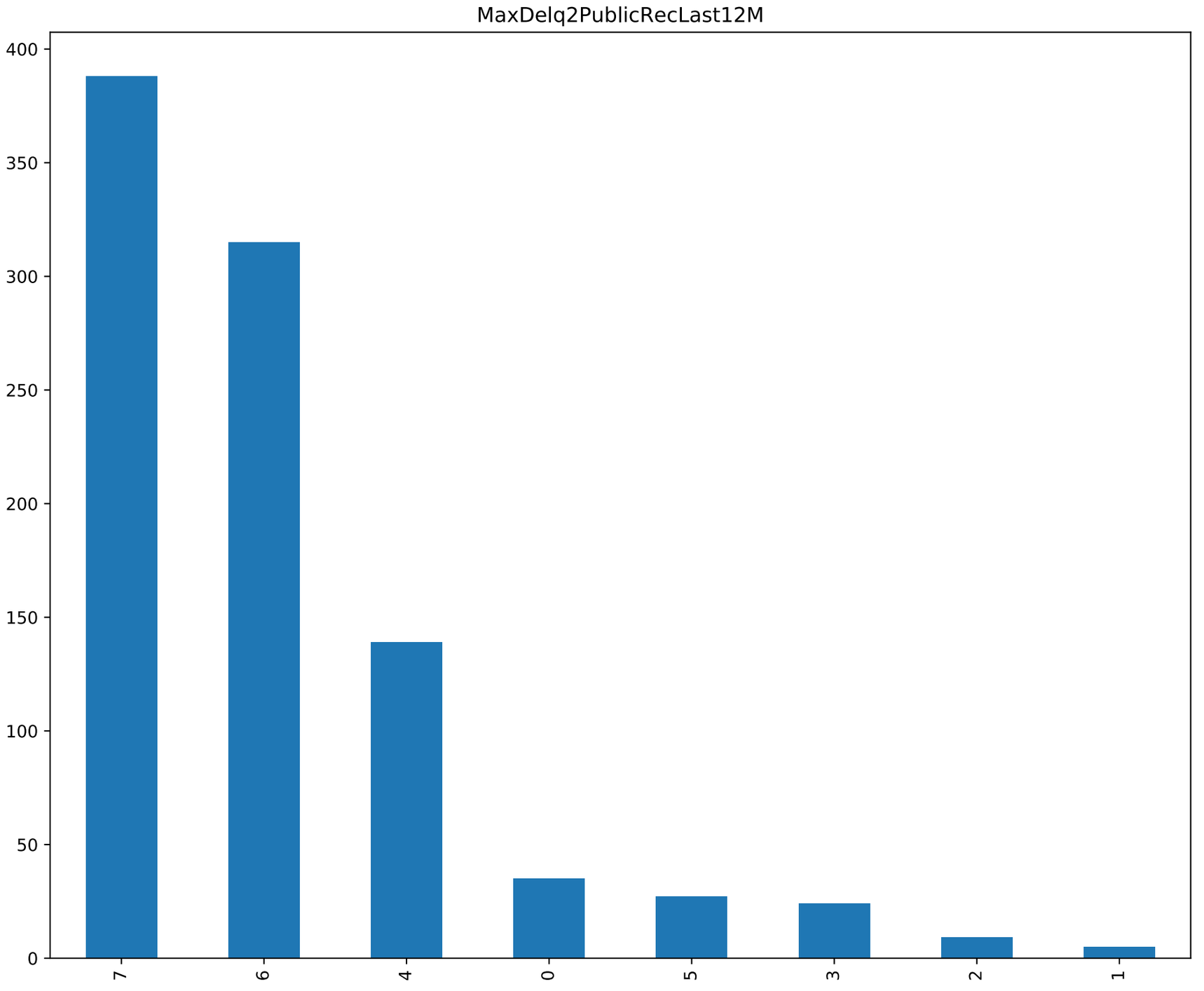

20
15
10
5
0
−5
NumRevolvingTradesWBalance

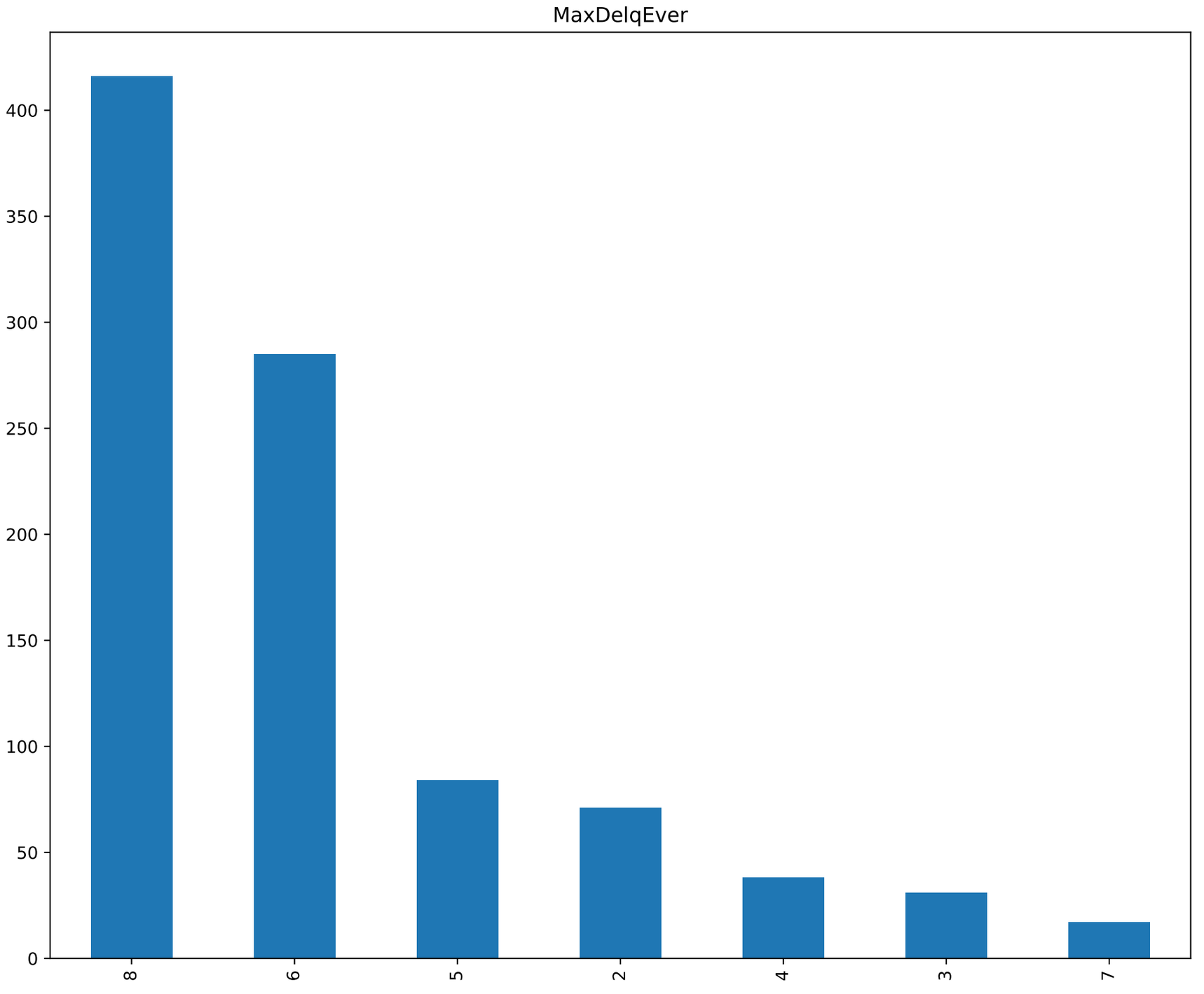

15
10
5
0
−5
NumInstallTradesWBalance


15
10
5
0
−5
NumBank2NatlTradesWHighUtilization


100
80
60
40
20
0
PercentTradesWBalance


RiskPerformance
500
400
300
200
100
0
Bad
Good


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


MaxDelqEver
400
350
300
250
200
150
100
50
0
8
6
5
2
4
3
7