Data Quality Report - Initial Findings

Descriptive Statistics for Continuous Features

	count	mean	std	min	25%	50%	75%	max
ExternalRiskEstimate	956.0	71.496862	10.508884	-9.0	64.0	71.0	80.0	93.0
MSinceOldestTradeOpen	956.0	189.944561	95.062922	-9.0	134.0	181.0	247.0	538.0
MSinceMostRecentTradeOpen	956.0	9.195607	12.239598	-9.0	3.0	6.0	11.0	184.0
AverageMInFile	956.0	77.534519	33.221419	-9.0	57.0	76.0	94.0	241.0
NumSatisfactoryTrades	956.0	21.043933	11.316955	-9.0	13.0	20.0	28.0	60.0
NumTrades60Ever2DerogPubRec	956.0	0.583682	1.413586	-9.0	0.0	0.0	1.0	17.0
NumTrades90Ever2DerogPubRec	956.0	0.379707	1.160956	-9.0	0.0	0.0	0.0	16.0
PercentTradesNeverDelq	956.0	92.240586	12.635140	-9.0	89.0	97.0	100.0	100.0
MSinceMostRecentDelq	956.0	7.570084	21.037396	-9.0	-7.0	0.0	14.0	81.0
NumTotalTrades	956.0	22.555439	12.783909	-9.0	13.0	21.0	30.0	77.0
NumTradesOpeninLast12M	956.0	1.928870	1.948328	-9.0	1.0	2.0	3.0	14.0
PercentInstallTrades	956.0	34.088912	17.608109	-9.0	21.0	33.0	44.0	100.0
MSinceMostRecentInqexcl7days	956.0	0.176778	5.853457	-9.0	0.0	0.0	1.0	23.0
NumInqLast6M	956.0	1.452929	2.278256	-9.0	0.0	1.0	2.0	29.0
NumInqLast6Mexcl7days	956.0	1.387029	2.220985	-9.0	0.0	1.0	2.0	29.0
NetFractionRevolvingBurden	956.0	34.598326	28.687260	-9.0	8.0	30.0	56.0	131.0
NetFractionInstallBurden	956.0	45.523013	41.410099	-9.0	-8.0	57.5	83.0	165.0
NumRevolvingTradesWBalance	956.0	3.986402	3.355206	-9.0	2.0	4.0	6.0	18.0
NumInstallTradesWBalance	956.0	1.589958	3.296716	-9.0	1.0	2.0	3.0	14.0
NumBank2NatlTradesWHighUtilization	956.0	0.652720	2.522922	-9.0	0.0	1.0	2.0	13.0
PercentTradesWBalance	956.0	66.243724	22.313995	-9.0	50.0	67.0	83.0	100.0

We have a full count of all the continuous features: there are 956 rows, and 21 features. The minimum value for each feature is -9, which has a special meaning of -9 No Bureau Record or No Investigation. This value is not useful for predicting the target feature, and therefore the rows should be removed if over 50% of the features are -9, or imputation used if the percentage is 30 or less. If the percentage is between 30% and 50%, then a decision based on the usefulness of the row will be made.

NetFractionInstallBurden has a -8 No Usable/Valid Trades or Inquiries value, which will be treated like -9 values above.

MSinceMostRecentDelq has a -7 Condition not Met (e.g. No Inquiries, No Delinquencies) value, which does have meaning and shouldn't undergo imputation. A binary feature will be

created to keep a record of the -7 appearance in the row, and the -7 value will be set to NaN so as not to skew our data at a later stage.

The standard deviation for MSinceOldestTradeOpen is rather large, but as the age of the credit institution isn't know, the range of months seen here doesn't seem worrisome. However, all outliers should be investigated further.

NumTrades60Ever2DerogPubRec and NumTrades90Ever2DerogPubRec seems to have very similar values. If there is a correlation between the two of around 90%+, then one of the features should be dropped, as there is redundancy.

NumInqLast6M and NumInqLast6Mexcl7days also need to be investigated for redundancy, and one feature dropped if deemed necessary.

The number of satisfactory trades is quite high which may be worth investigating as an indicator for the target value of RiskPerformance.

Descriptive Statistics for Categorical Features

	count	unique	top	freq
RiskPerformance	956	2	Bad	508
MaxDelq2PublicRecLast12M	956	9	7	412
MaxDelqEver	956	8	8	441

There are three categorical features. What is interesting is that the RiskPerformance target feature is mostly 'Bad', while the majority of trades were satisfactory, as we saw in the last section. It seems as though past satisfactory trades are not weighted as heavily as some other features in determining the target feature, but of course this needs to be investigated further.

MaxDelq2PublicRecLast12M has two values for 'unknown delinquency': 5 and 6, and also has two values for 'all other': 8 and 9. It can be seen that the top value in the dataframe is 7 however, and thus we don't need to worry about 8 and 9. 5 and 6 can be combined as the meaning is the same. Thus, 5 can be replaced with 6.

The top values for MaxDelq2PublicRecLast12M and MaxDelqEver are 7 and 8 respectively, which means 'current and never delinquent'. This shows that the credit company's customers are more often good at repaying their loans than not.

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

AverageMInFile, MSinceOldestTradeOpen, NumSatisfactoryTrades, NumTradesOpeninLast12M, and PercentInstallTrades all appear to be normally distributed.

ExternalRiskEstimate, PercentTradesNeverDelq, and PercentTradesWBalance all seem to be unimodal skewed left.

MSinceMostRecentDelq, MSinceMostRecentInqexcl7days, MSinceMostRecentTradeOpen, NumBank2NatlTradesWHighUtilization, NumInqLast6M,NumInqLast6Mexcl7days, NumInstallTradesWBalance, NumTrades60Ever2DerogPublic, and NumTrades90Ever2DerogPublic seem to be exponential once the special minus values are ignored.

NetFractionInstallBurden appears to be multimodal.

NetFractionRevolvingBurden, NumTotalTrades, and NumRevolvingTradesWBalance appears to be unimodal skewed right.

Box Plots for Continuous Features [plots attached at end of file]

Most of the plots have outliers present. Many of the outliers are multiple standard deviations from the mean. The following features are affected: MSinceOldestTradeOpen, MSinceMostRecentTradeOpen, AverageMInFile, NumSatisfactoryTrades, NumTrades90Ever2DerogPubRec, PercentTradesNeverDelq, MSinceMostRecentDelq, NumTotalTrades, NumTradesOpeninLast12M, PercentInstallTrades, MSinceMostRecentInqexcl7days, NumInqLast6Mexcl7days, NetFractionRevolvingBurden, NumRevolvingTradesWBalance, NumInstallTradesWBalance, NumBank2NatlTradesWHighUtilization, PercentTradesWBalance.

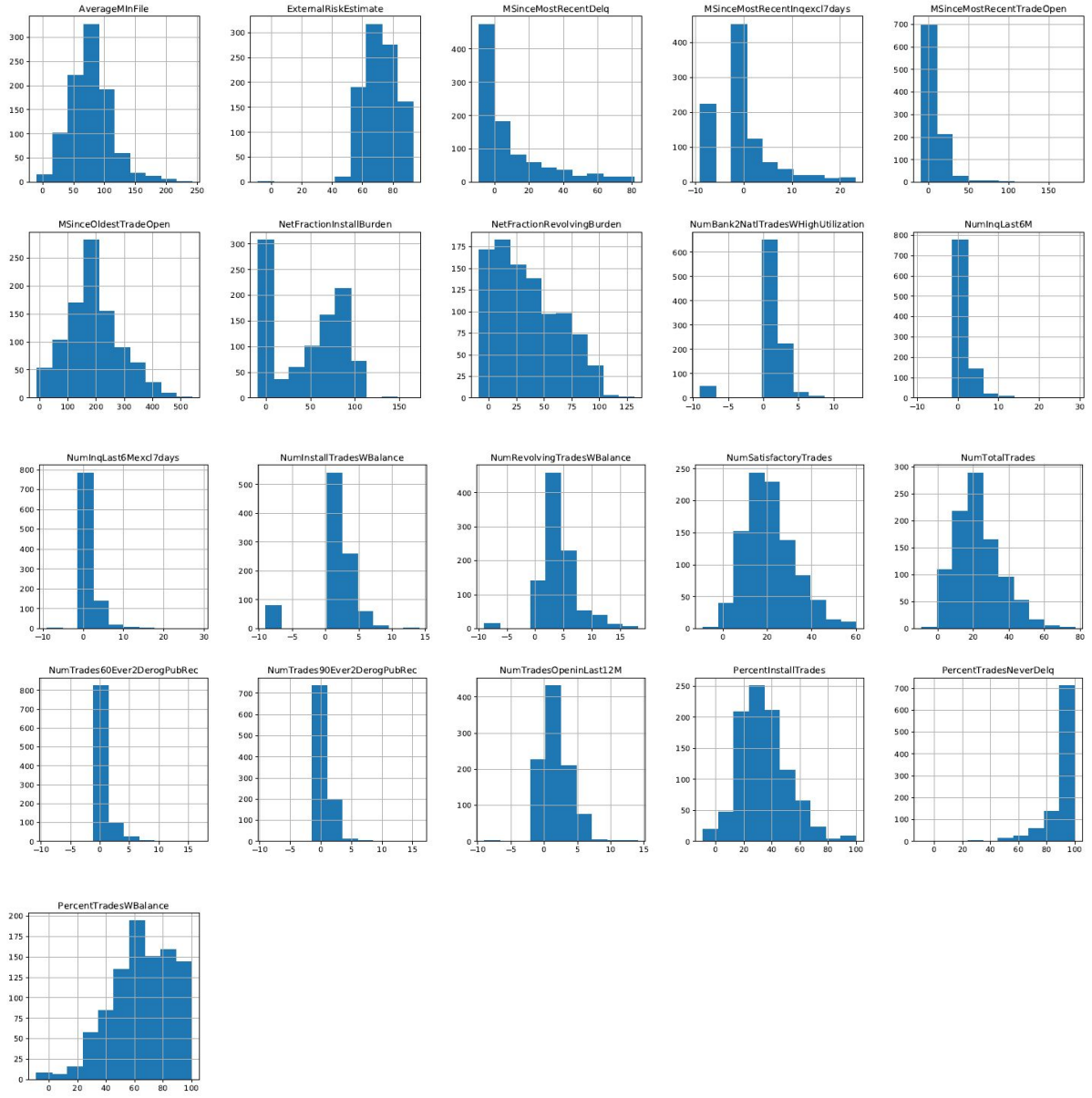
These outliers need investigation, but upon initial inspection, nothing in the data seems unexplainable or impossible. Some people may have opened an account many years before the credit company became popular for example, and thus some people have trades open much longer than the majority of customers.

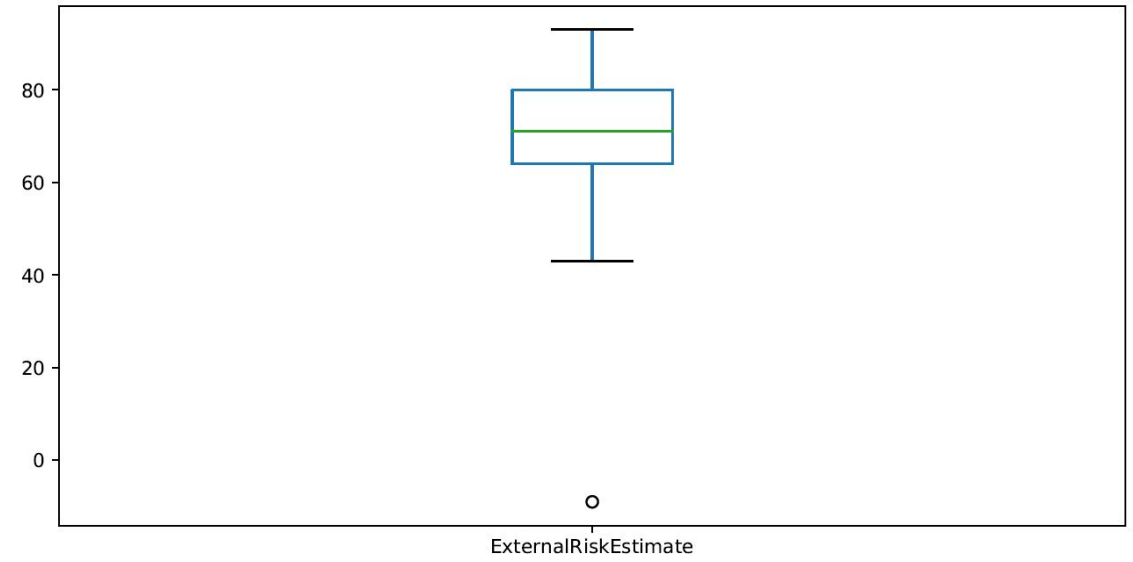
Bar Plots for Categorical Features [plots attached at end of file]

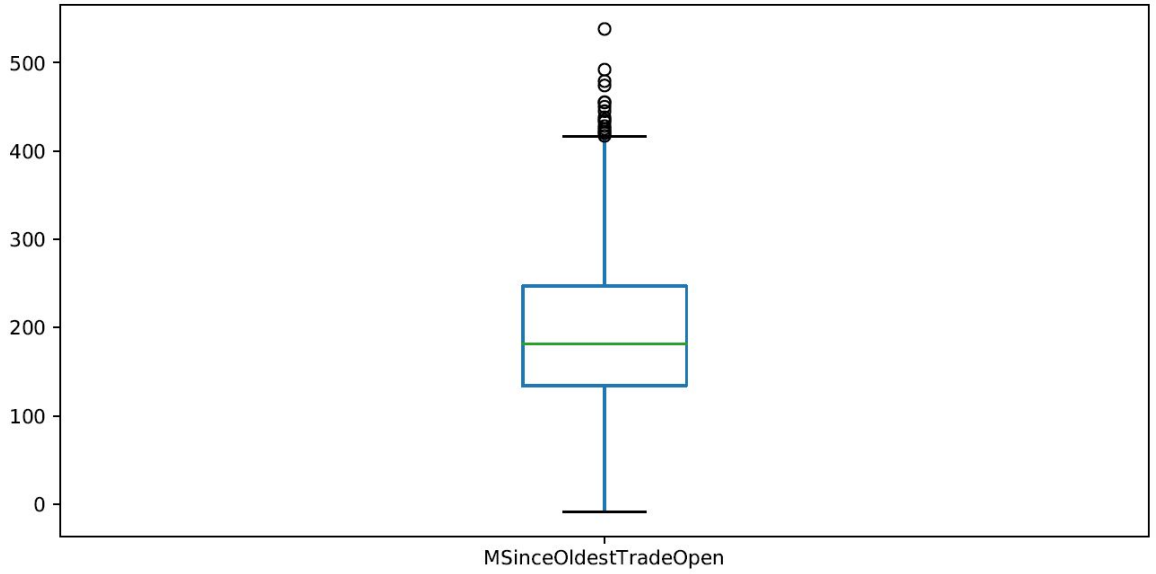
It can be seen that the outcome of the target feature is roughly 50/50, yet it must be noted that slightly more of RiskPerformance's values are bad. This is interesting as the number of satisfactory trades is very high. Investigation will take place to determine with greater accuracy what affects the target feature.

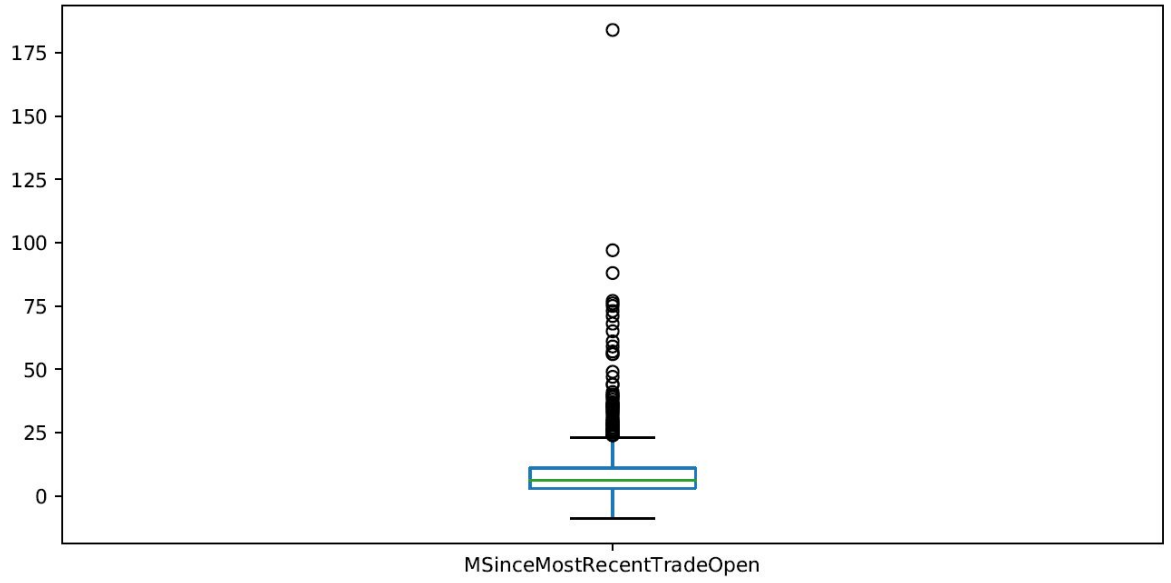
Looking at MaxDelq2PublicRecLast12M, we find that current and never delinquent is the most common value, and then somewhat surprisingly, unknown delinquency is the second most common value. It seems odd that something as important as delinquency figures are unknown, but we have to deal with the data we are dealt.

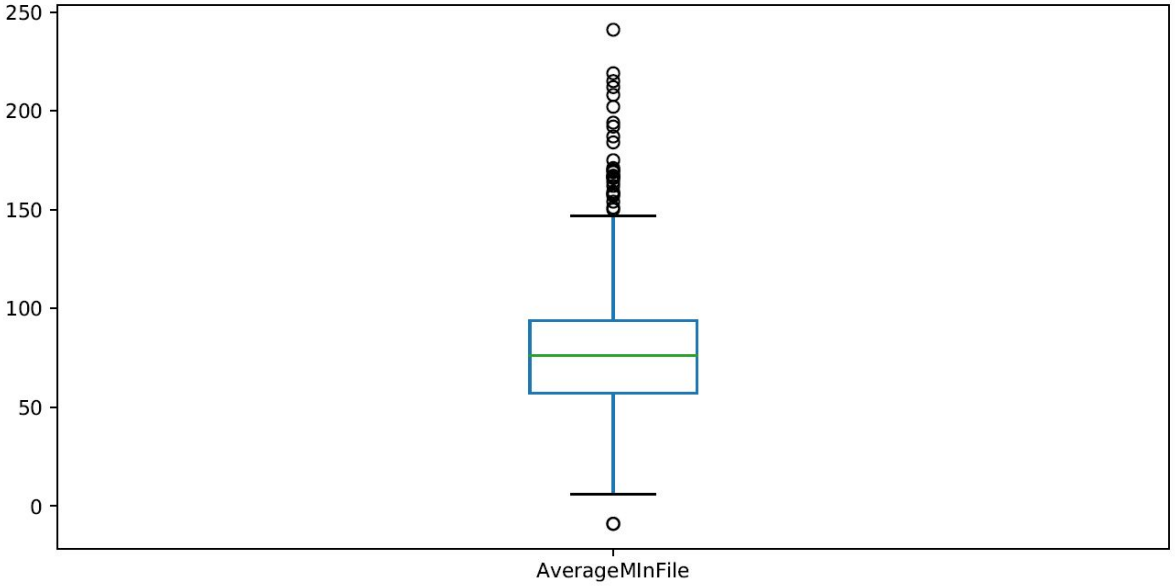
MaxDelqEver is more in line with expectations, as current and never delinquent is the most common value, followed by 30 days delinquent as the second highest value. Unknown delinquencies are very low. This may be because the credit company had more time to process customer data over many years rather than in just the last 12 months.

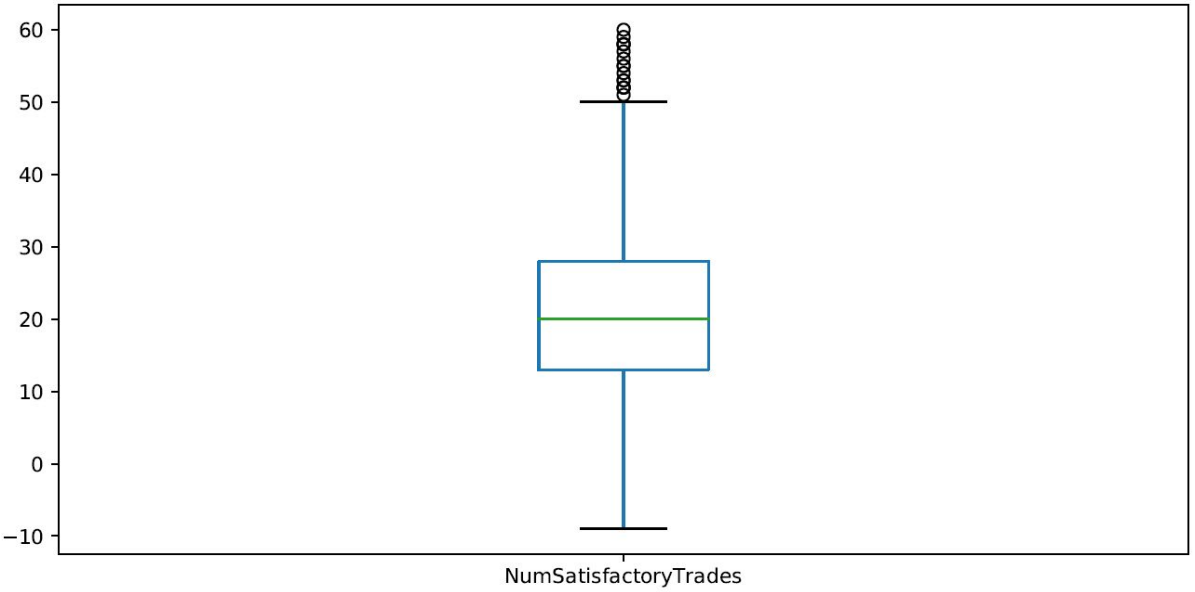


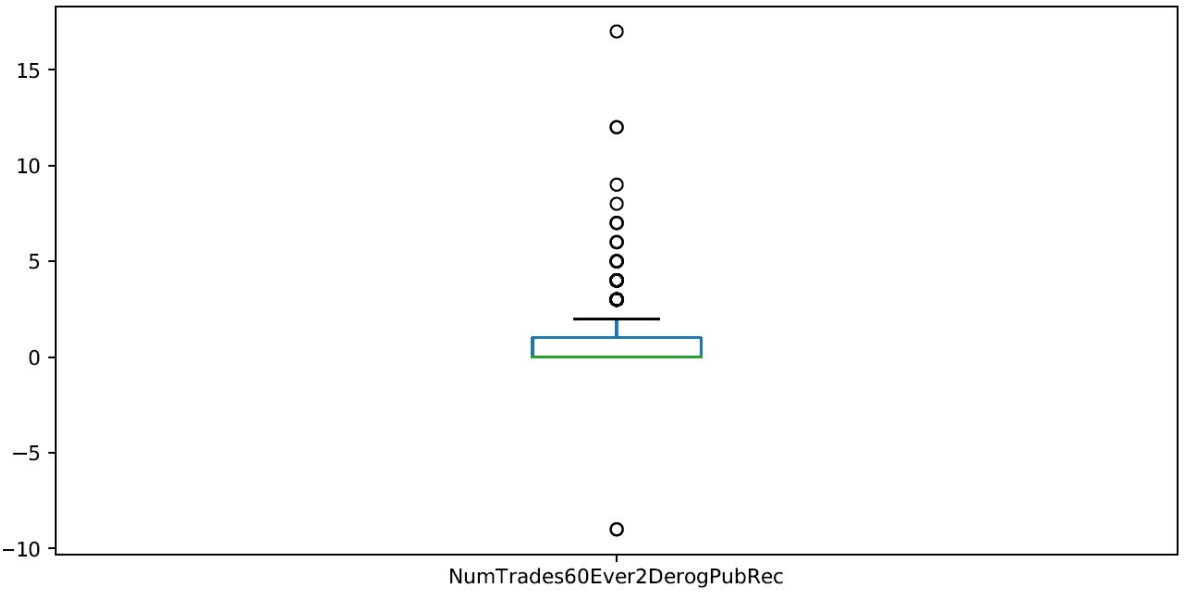

80
60
40
20
0
ExternalRiskEstimate

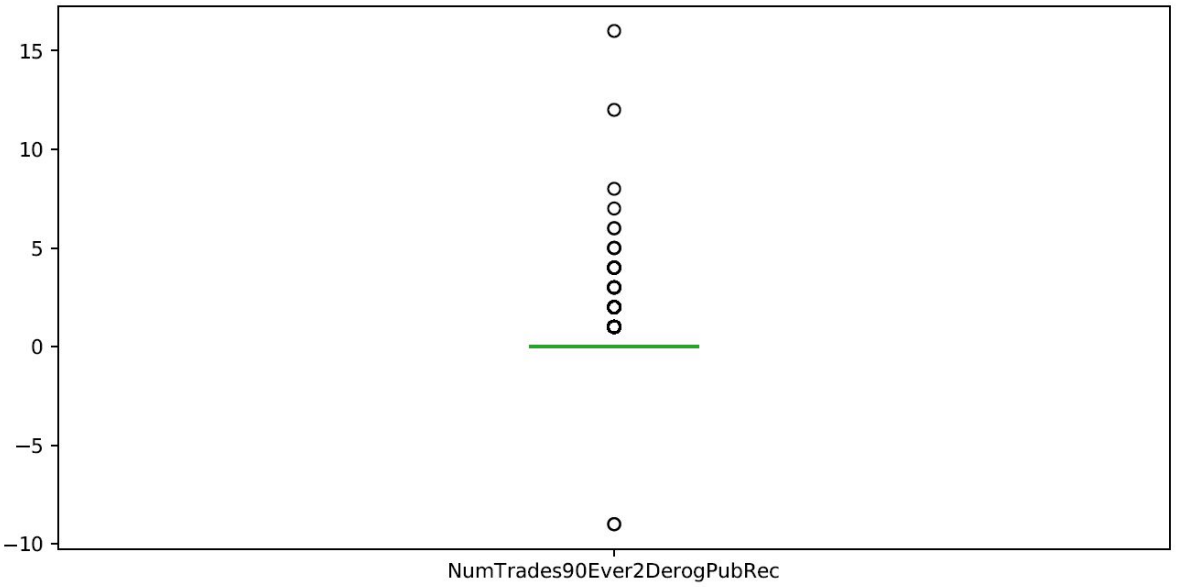

500
400
300
200
100
0
MSinceOldestTradeOpen

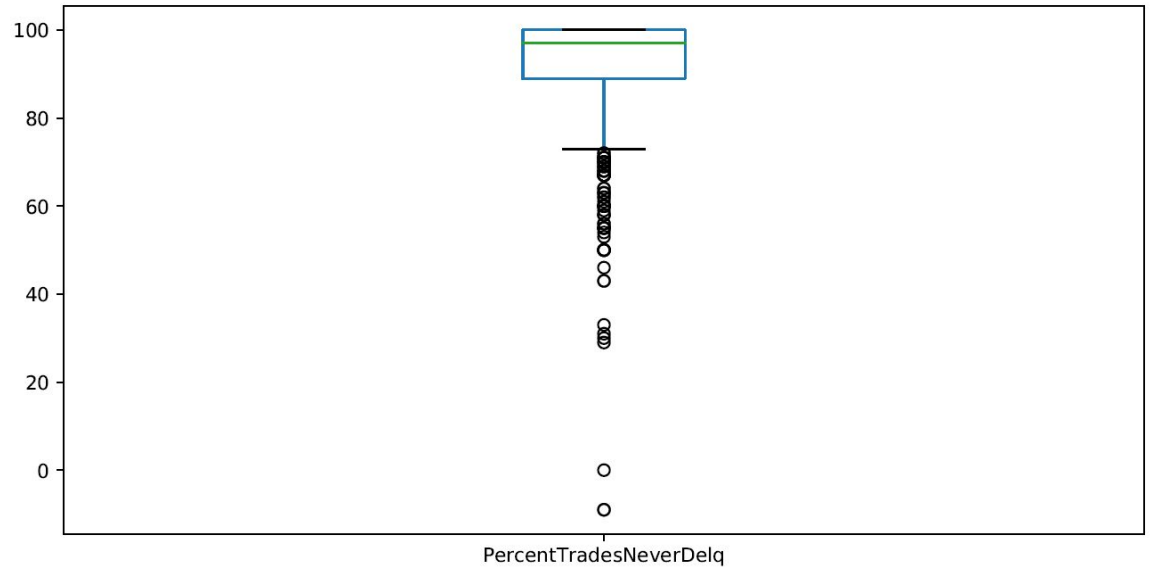

175
150
125
100
75
50
25
0
MSinceMostRecentTradeOpen

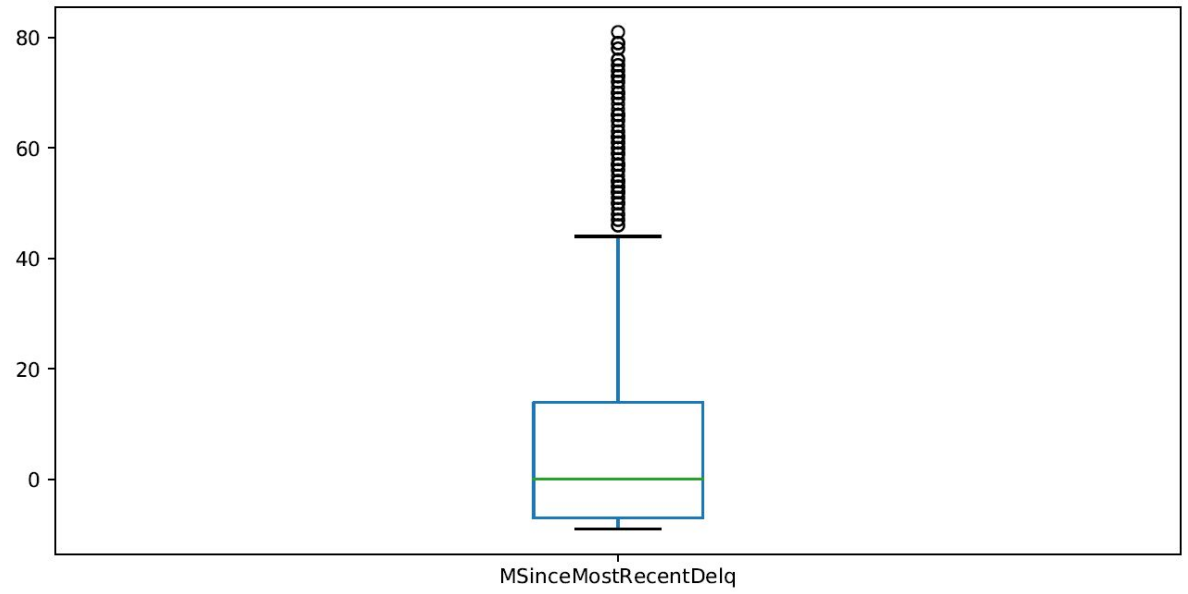

250
200
150
100
50
0
AverageMInFile

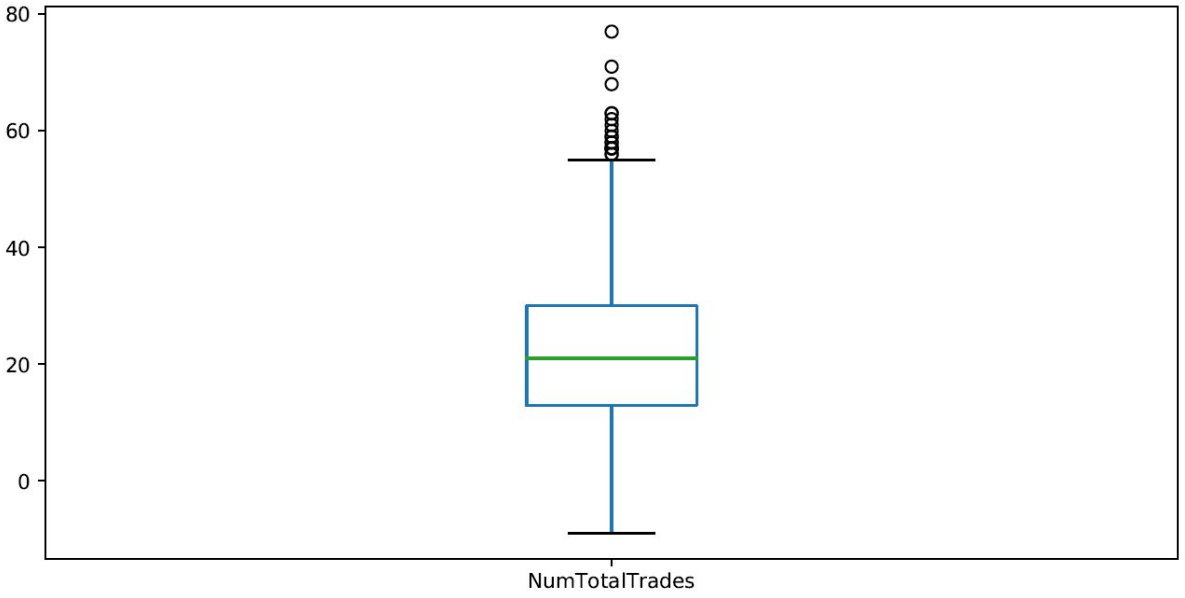

60
50
40
30
20
10
0
−10
NumSatisfactoryTrades

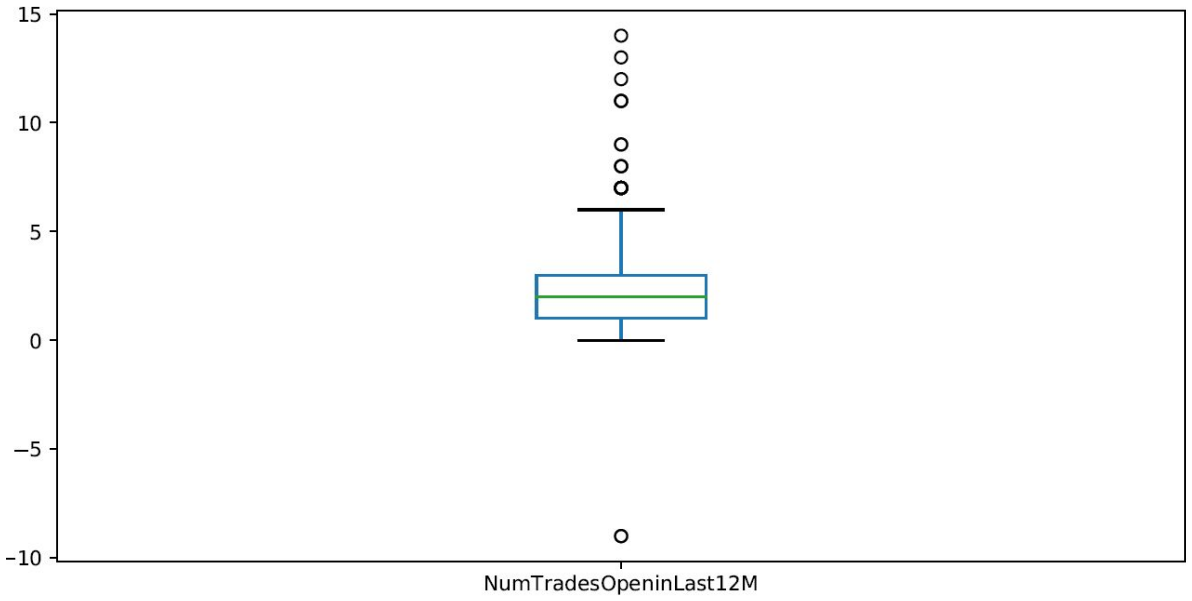

15
10
5
0
−5
−10
NumTrades60Ever2DerogPubRec

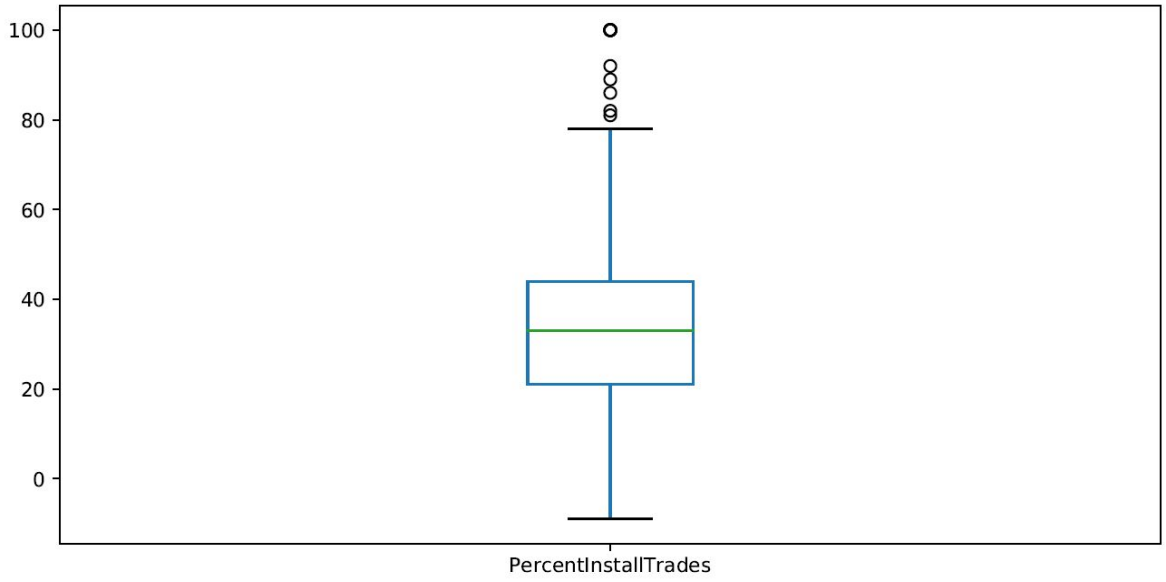

15
10
5
0
−5
−10
NumTrades90Ever2DerogPubRec

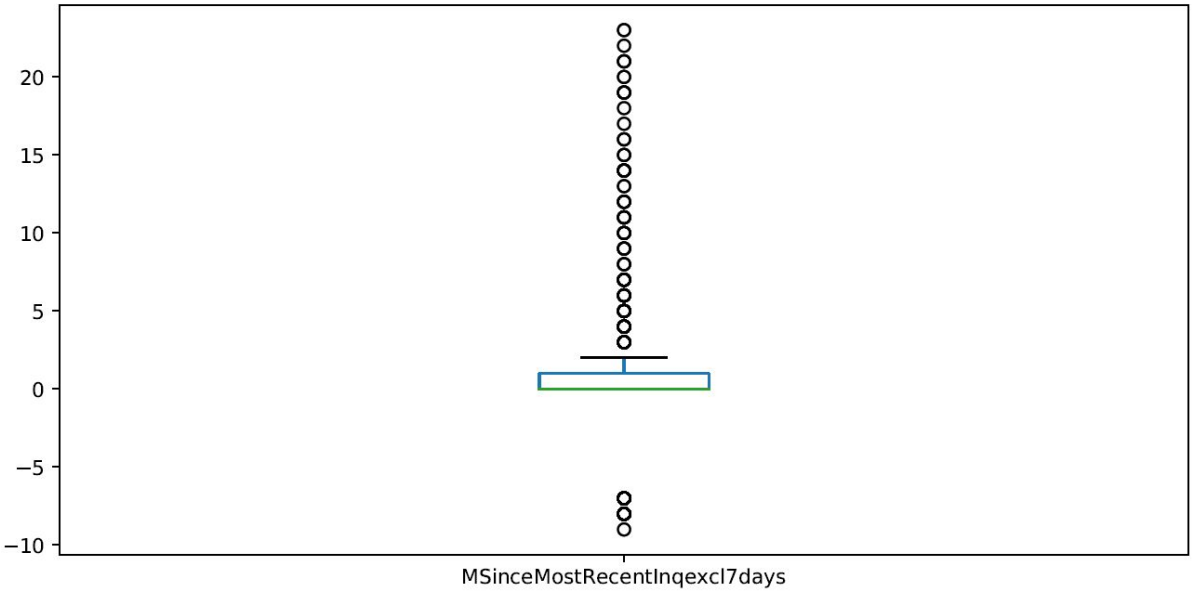

100
80
60
40
20
0
PercentTradesNeverDelq

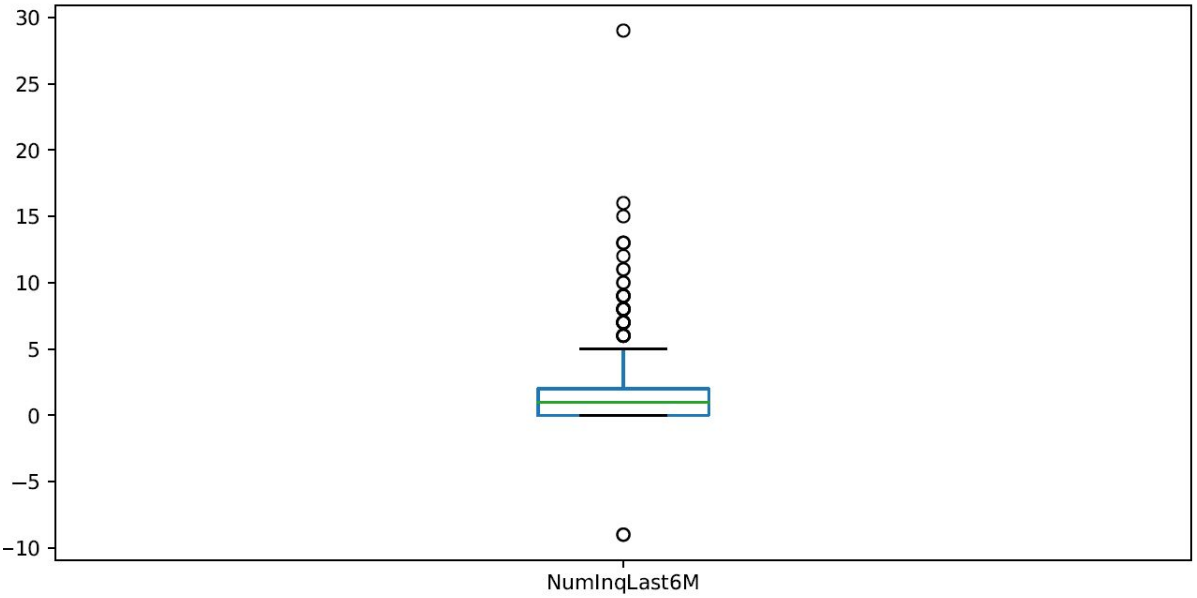

80
60
40
20
0
MSinceMostRecentDelq

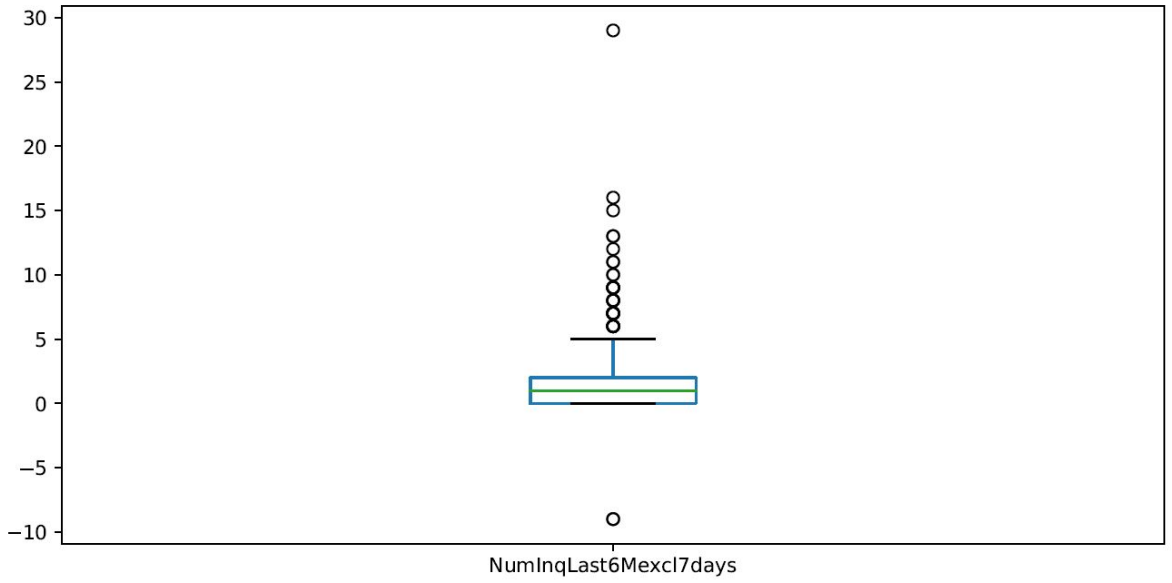

80
60
40
20
0
NumTotalTrades

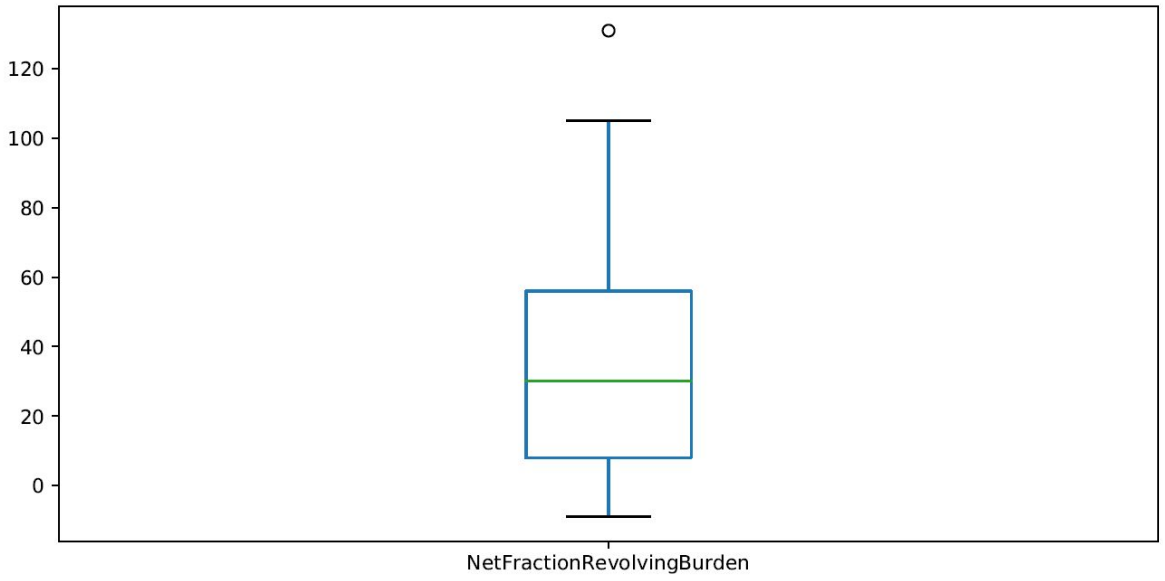

15
10
5
0
−5
−10
NumTradesOpeninLast12M

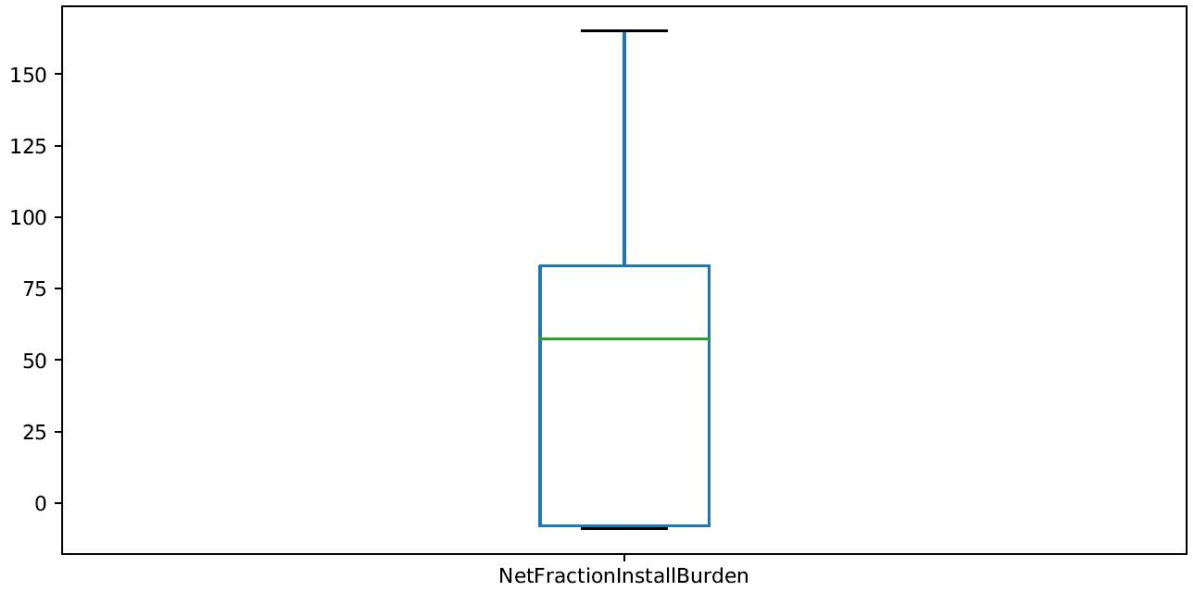

100
80
60
40
20
0
PercentInstallTrades

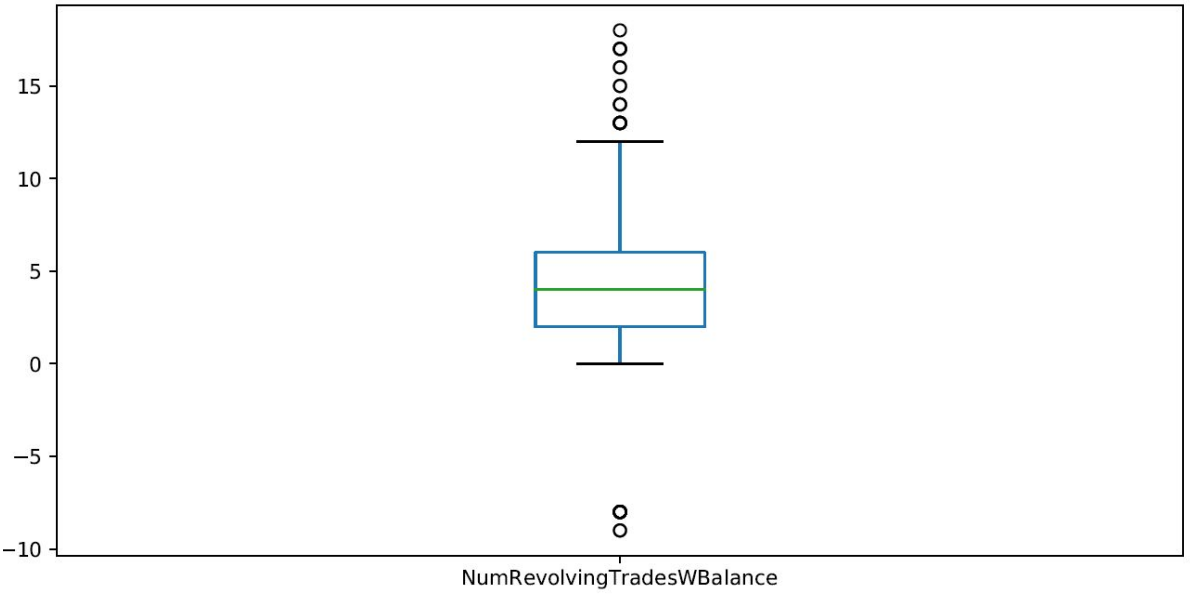

20
15
10
5
0
−5
−10
MSinceMostRecentInqexcl7days

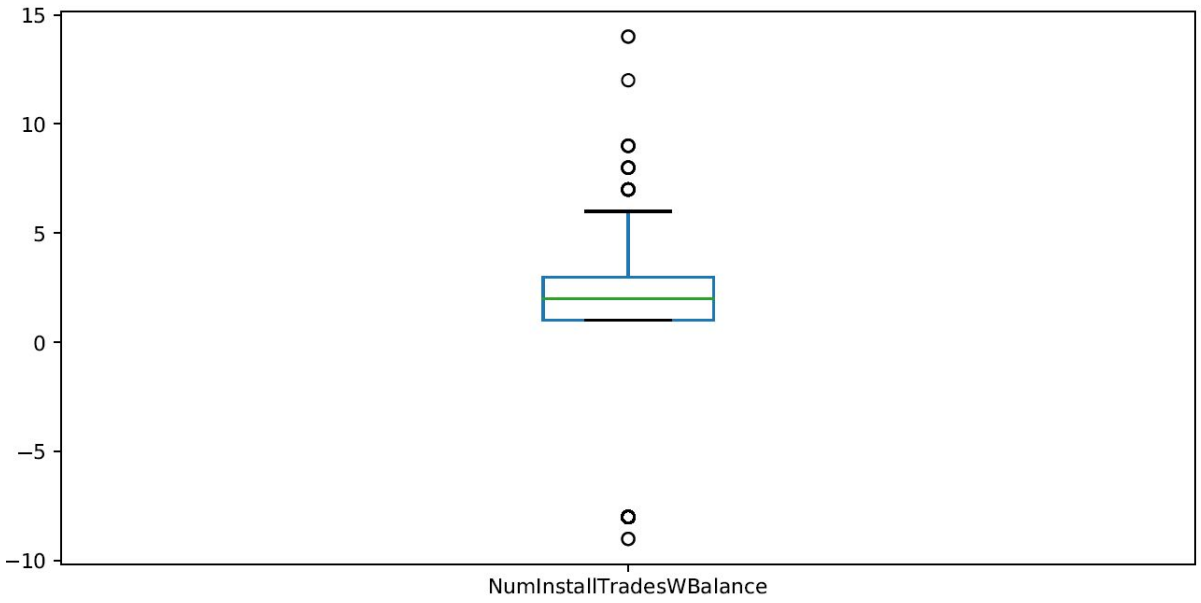

30
25
20
15
10
5
0
−5
−10
NumInqLast6M

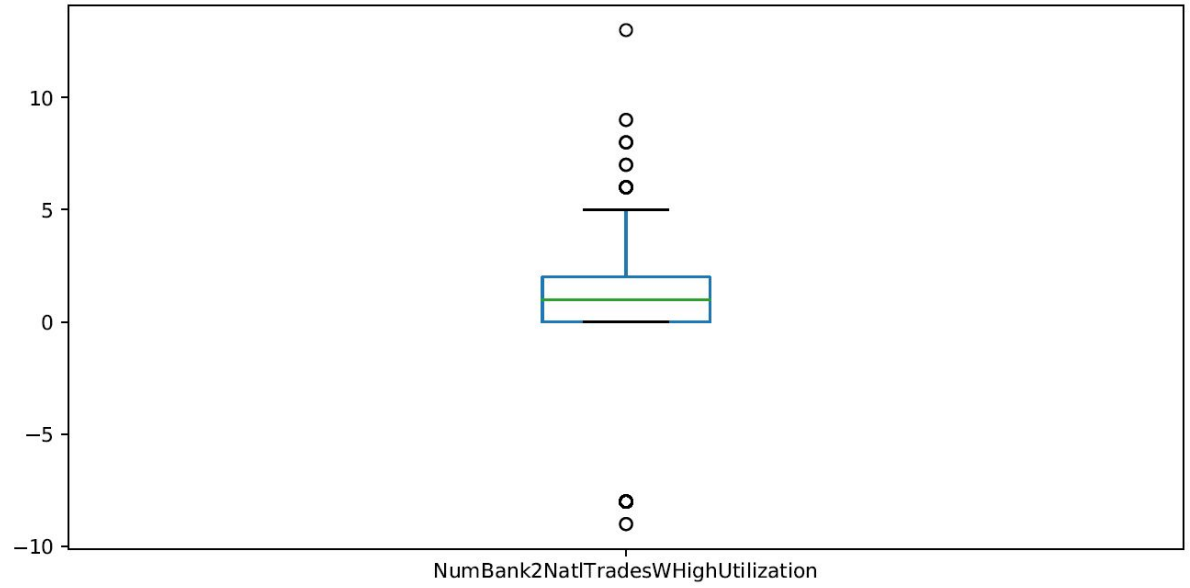

30
25
20
15
10
5
0
−5
−10
NumInqLast6Mexcl7days

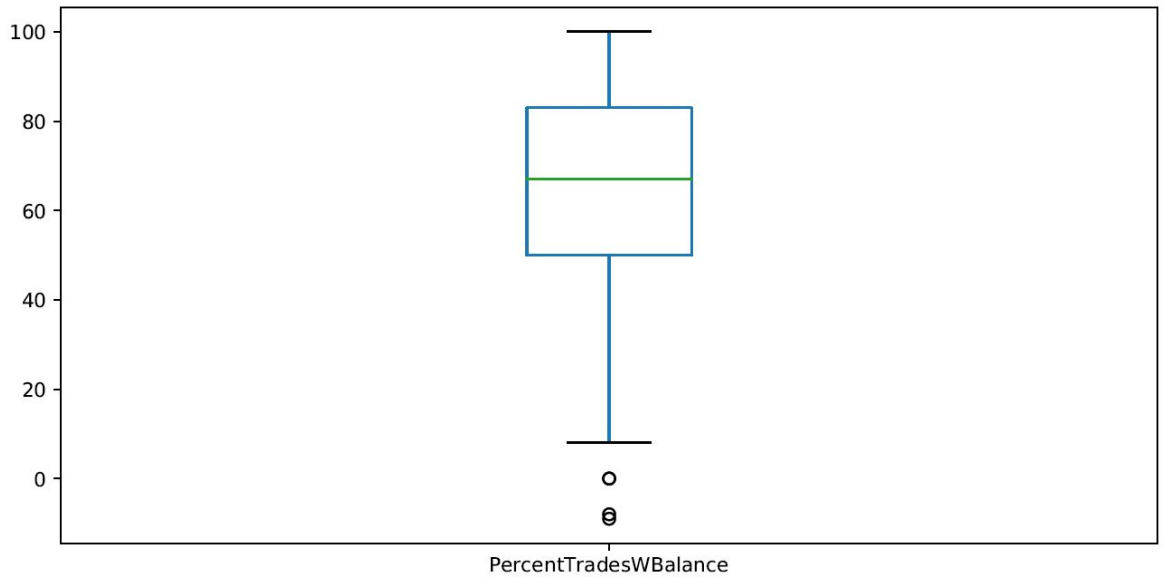

120
100
80
60
40
20
0
NetFractionRevolvingBurden

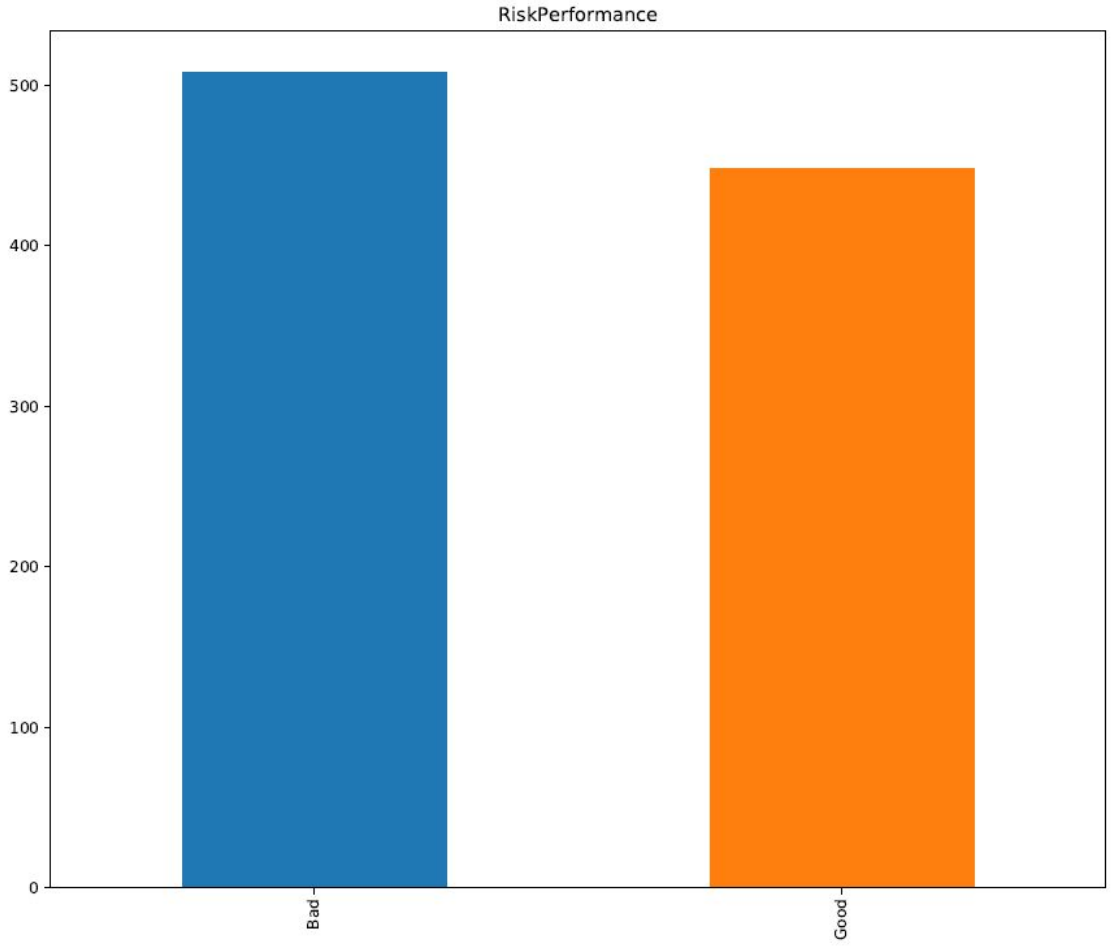

150
125
100
75
50
25
0
NetFractionInstallBurden

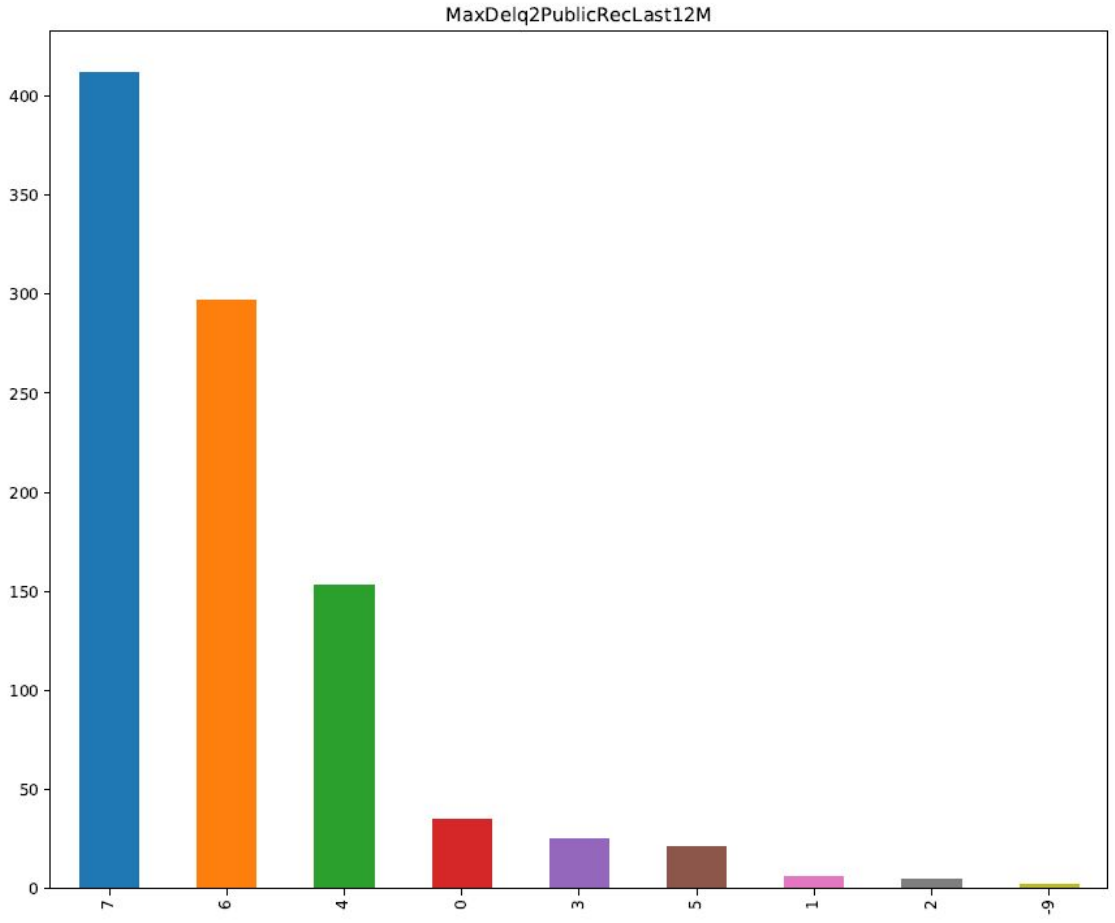

15
10
5
0
−5
−10
NumRevolvingTradesWBalance

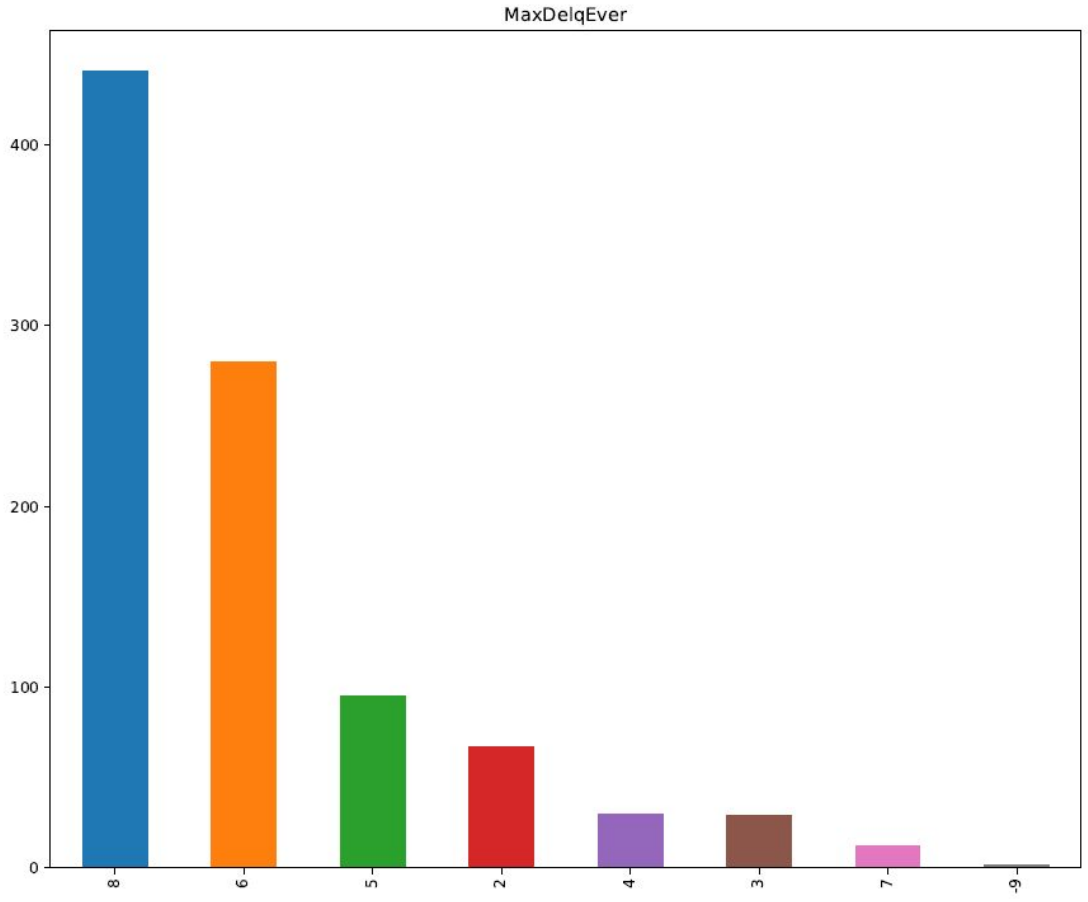

15
10
5
0
−5
−10
NumInstallTradesWBalance


10
5
0
−5
−10
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
3
5
1
2
-9


MaxDelqEver
400
300
200
100
0
8
6
5
2
4
3
7
-9