Data Quality Plan

After the initial understanding of the data, the second step is to identify any data quality issues. A data quality issue is defined as anything unusual about the data. The most common data quality issues are missing values, irregular cardinality problems, and outliers.

General considerations

The handling strategies have been calibrated for each feature and the final decision follows these guidelines and commonly used techniques:

Missing and invalid values:

- Drop features: remove features that have many missing values
- Drop rows: perform complete case analysis if it affects only a few rows
- Imputation: replace missing value with reasonable estimate (e.g. mean or median)
- New features: Derive a missing indicator feature from features with missing values
- Regression Substitution: use multiple-regression analysis to estimate and predict a missing value based on other values.

Outliers:

clamp transformation: clamps all values above an upper threshold and below a lower threshold to these threshold values, thus removing the outliers. The upper and lower thresholds can be set manually based on domain knowledge or can be calculated from data. We calculated the thresholds from data using mainly the following method: lower threshold to the 1st quartile value minus 1.5 times the inter-quartile range and the upper threshold to the 3rd quartile plus 1.5 times the inter-quartile range. Only in one case we adopted a different approach to establish the boundaries: setting the upper and lower thresholds using the mean value of a feature plus or minus 2 times the standard deviation. It's important to take into consideration that clamp transformation is quite invasive and the impact of this technique should be evaluated by comparing the performance of different models trained on datasets where the transformation has been applied and where it has not.

Categorical Features

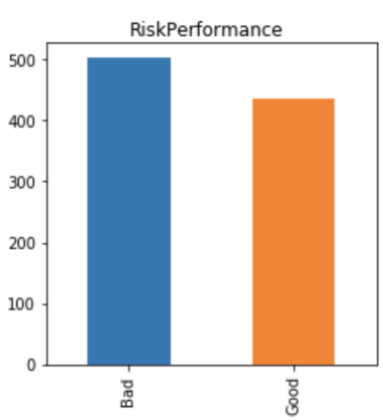
The table below represent a summary of the data quality plan for the categorical features.

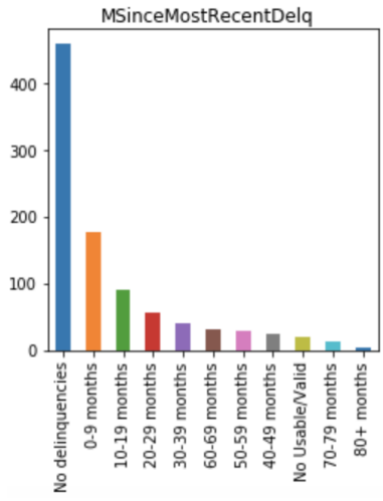
Categorical Feature	Data Quality Issue	Handling Strategy
RiskPerformance	None	Keep as is
MSinceMostRecentDelq	Special Value -8 (2%)	Imputation
MSinceMostRecentInqexcl7days	Special Value -8 (4%)	Imputation
MaxDelqEver	None	Keep as is
MaxDelq/PublicRecLast12M	None	Keep as is

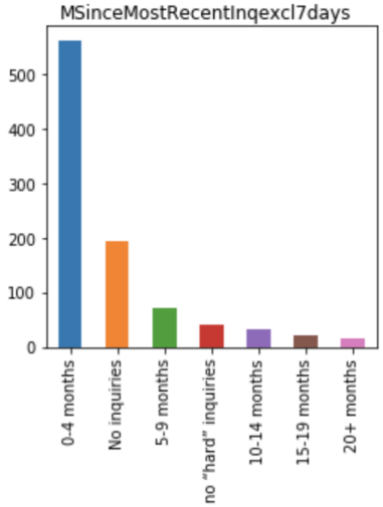
The bar charts graphs prepared for the data quality report helped understand that there are essentially no data quality issues for the categorical features ***RiskPerformance***, ***MaxDelqEver*** and ***MaxDelq*/*PublicRecLast12M***. It's indeed possible to observe that there are no negative values i.e. -9, corresponding to missing data. The cardinality, which shows the number of distinct values present is regular, i.e. it's always more than 1. The feature ***RiskPerformance*** has a low cardinality of 2 as it represents a flag for the target.

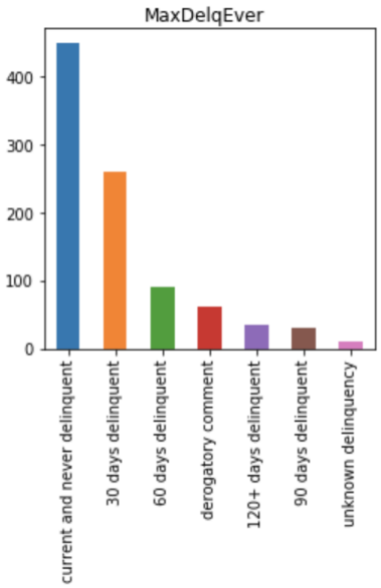
We have observed that the features ***MSinceMostRecentDelq*** and ***MSinceMostRecentInqexcl7days*** have the same issue related to a small percentage of special values -8.
The value -8 means no Usable/Valid Accounts Trades or Inquiries and can indicate inactive, or very old account.
Since the invalid data in quite low (2% and 4%), we use imputation to assign this value to its explicit meaning.
We decide to use imputation over the other possible approaches because, even if the account is very old, some delinquency could have happened in the past and dropping every row associated with this value would cause an excessive data loss.

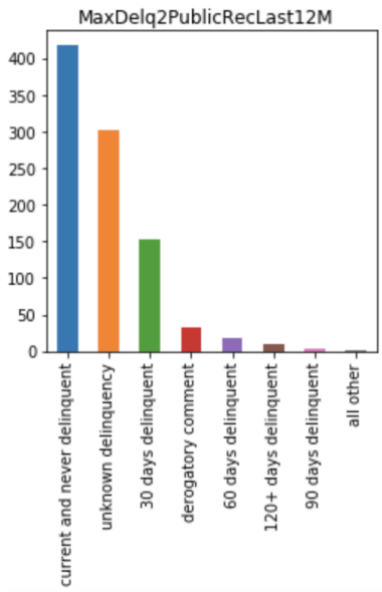
The following graphs represent the updated categorical features.











Continuous Features

The table below represent a summary of the data quality plan for the continuous features.

Numeric Feature	Data Quality Issue	Handling Strategy
ExternalRiskEstimate	Special Value -9 (0.2%)	Imputation (mean)
MSinceOldestTradeOpen	Special Value -8 (3%) & Outliers (High)	Imputation (max) & Clamp
MSinceMostRecentTradeOpen	Outliers (High)	Clamp transformation
AverageMInFile	Outliers (High)	Clamp transformation
NumSatisfactoryTrades	Outliers (High)	Clamp transformation
NumTrades60Ever2DerogPubRec	Outliers (High)	Clamp transformation
NumTrades90Ever2DerogPubRec	Outliers (High)	Clamp transformation
PercentTradesNeverDelq	Outliers (Low)	Clamp transformation
NumTotalTrades	Outliers (High)	Clamp transformation
NumTradesOpeninLast12M	Outliers (High)	Clamp transformation
PercentInstallTrades	Outliers (High)	Clamp transformation
NumInqLast6M	Outliers (High)	Clamp transformation
NumInqLast6Mexcl7days	Outliers (High)	Clamp transformation
NetFractionRevolvingBurden	Special Value -8 (2%) & outlier (high)	Imputation (median)
NetFractionInstallBurden	Special Value -8 (34%)	Drop Feature
NumRevolvingTradesWBalance	Special Value -8 (2%) & Outliers (High)	Imputation (mean) & Clamp
NumInstallTradesWBalance	Special Value -8 (9%) & Outliers (High)	Imputation (mean) & Clamp
NumBank2NatlTradesWHighUtilization	Special Value -8 (7%) & Outliers (High)	Imputation (mean) & Clamp
PercentTradesWBalance	Special Value -8 (0.2%) & Outliers (Low)	Imputation (mean)

We can observe that every continuous feature shows one or more data quality issues. In general, no irregular cardinality is observed and the issues are due to outliers and presence of special values (-9, and -8) which indicates missing data, invalid accounts trades or inquiries.

We'll now discuss in more details the data issues and the handling strategy adopted for every continuous feature.

Note: for every feature on which imputation was applied, since the values to be replaced were negative (-8 or -9), these special values were first replaced by a null value. This first step was taken in order to avoid that a feature statistic (e.g. mean or median), that is used in the imputation process, is affected by these negative values.

ExternalRiskEstimate **Special Value -9 (0.2%)**

Missing data is very low therefore the missing values were replaced by imputation using the mean.

MSinceOldestTradeOpen **Special Value -8 (3%) & Outliers (High)**

-8 for accounts/trades means inactive, or very old. Since -8 stands for a very old value, we decide to replace, using imputation, the special value with the max value of the feature as it represents a higher value in terms of months. We take this approach as invalid data is quite low and dropping the feature or the rows will cause too much data loss.
Another data issue is given by outliers and this is resolved by clamping the values.

MSinceMostRecentTradeOpen **Outliers (High) Clamp transformation**

No missing data only outliers, clamping is applied using inter-quartile range.

AverageMInFile **Outliers (High)**

No missing data only outliers, clamping is applied using inter-quartile range.

NumSatisfactoryTrades **Outliers (High)**

No missing data only outliers, clamping is applied using inter-quartile range.

NumTrades60Ever/DerogPubRec **Outliers (High)**

No missing data only outliers, clamping is applied using inter-quartile range.

NumTrades90Ever/DerogPubRec **Outliers (High)**

No missing data only outliers, clamping is applied. In this case however a different technique has been used to determine the thresholds. This because, since 70% of the values is 0, using the quartiles and the IQR, both limits would be set at 0. We therefore the another technique previously described which adopts mean and standard deviation to determine the upper and lower bounds.

PercentTradesNeverDelq **Outliers (Low)**

No missing data only outliers, clamping is applied using inter-quartile range.

NumTotalTrades **Outliers (High)**

No missing data only outliers, clamping is applied using inter-quartile range.

NumTradesOpeninLast12M **Outliers (High)**

No missing data only outliers, clamping is applied using inter-quartile range.

PercentInstallTrades **Outliers (High)**

No missing data only outliers, clamping is applied using inter-quartile range.

NumInqLast6M **Outliers (High)**

No missing data only outliers, clamping is applied using inter-quartile range.

NumInqLast6Mexcl7days **Outliers (High)**

No missing data only outliers, clamping is applied using inter-quartile range.

NetFractionRevolvingBurden **Special Value -8 (2%) & outlier (high)**

Since the percentage of special values is quite low, these are replaced with the median value of the feature because, even if the value means no valid inquires, we consider a better approach to assign the median value rather than dropping the rows. Another data issue is given by outliers and this is resolved by clamping the values.

NetFractionInstallBurden **Special Value -8 (34%)**

Net Fraction Installment Burden. This is installment balance divided by original loan amount.
Invalid data is high and above 30%. There are various approaches that should be taken into consideration:

- in this scenario the use of imputation is not recommended as it would not be accurate and could change the central tendency of the feature too much;
- complete case analysis, which deletes all the rows but also in this case the data loss would be too high;
- derive a missing indicator feature, a binary feature that flags whether the value was present or missing in the original feature but for a feature that represent a fraction this would not probably be relevant;
- apply Regression Substitution which uses multiple-regression analysis to estimate a missing value. Regression substitution predicts the missing value from the other values.

After these considerations and taken into account that the feature represents the fraction between installment balance divided by original loan amount, and does not essentially give any direct information about previous legal/illegal behaviour (e.g delinquency, late payments, etc…), the feature will be dropped.

NumRevolvingTradesWBalance **Special Value -8 (2%) & Outliers (High)**

Invalid data is quite low, dropping the feature or the rows will cause too much data loss, we use imputation replacing the missing values with the mean. Another data issue is given by outliers and this is resolved by clamping the values.

NumInstallTradesWBalance **Special Value -8 (9%) & Outliers (High)**

Invalid data is quite low, dropping the feature or the rows will cause too much data loss, we use imputation replacing the missing values with the mean. Another data issue is given by outliers and this is resolved by clamping the values.

NumBank/NatlTradesWHighUtilization **Special Value -8 (7%) & Outliers (High)**

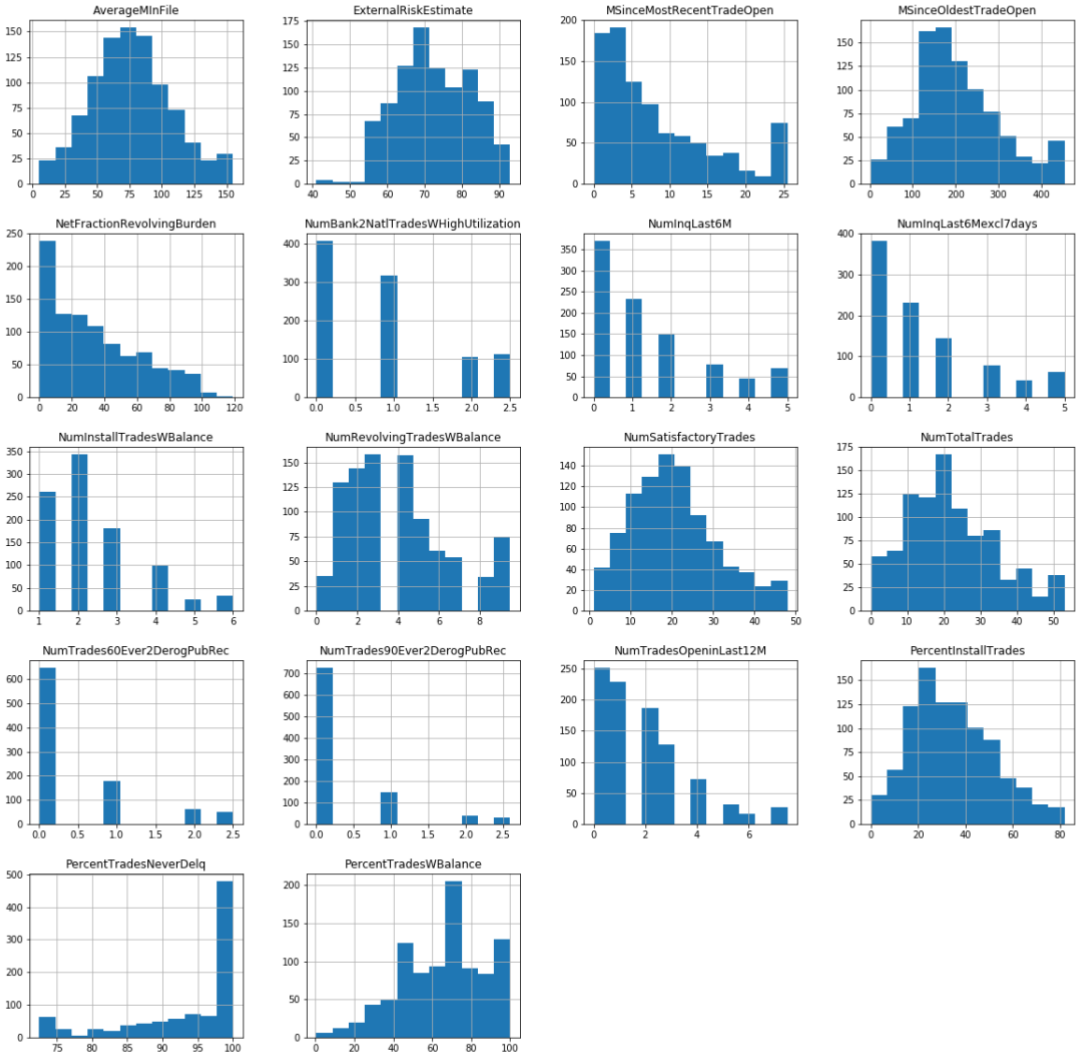
Missing data is quite low, dropping the feature or the rows will cause too much data loss, we use imputation replacing the missing values with the mean. Another data issue is given by outliers and this is resolved by clamping the values.

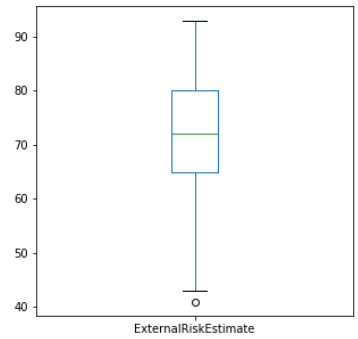
PercentTradesWBalance **Special Value -8 (0.2%) & Outliers (Low)**

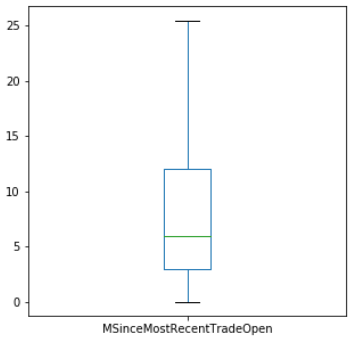
Missing data is very low, dropping the feature or the rows will cause too much data loss, we use imputation replacing the missing values with the mean. Another data issue is given by outliers and this is resolved by clamping the values.

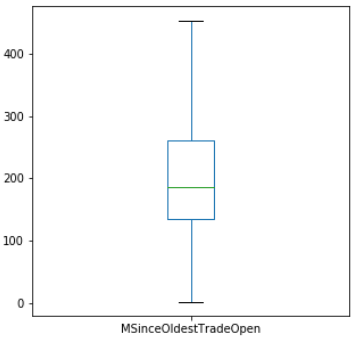
The update/cleaned dataframe was saved as:
'CreditRisk_clean_round2_5Mar2019_DataQualityPlan'

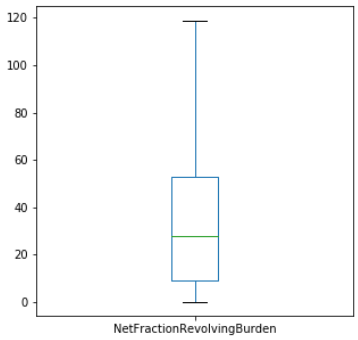
The following pages contain the graph representations, histograms and box plots, of the updated continuous features. We can observe that the issues related to missing values and/or special values have been resolved for every feature. Furthermore, also the issues related to outliers have been all resolved, with the only exception of the feature ***NumTrades90Ever/DerogPubRec*** which has most of its values equal to zero and therefore the box plot appears as a line at value 0 and a few outliers are still present.

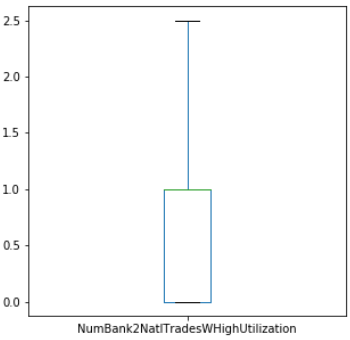

AverageMInFile
ExternalRiskEstimate
MSinceMostRecentTradeOpen
MSinceOldestTradeOpen
NetFractionRevolvingBurden
NumBank2NatlTradesWHighUtilization
NumInqLast6M
NumInqLast6Mexcl7days
NumInstallTradesWBalance
NumRevolvingTradesWBalance
NumSatisfactoryTrades
NumTotalTrades
NumTrades60Ever2DerogPubRec
NumTrades90Ever2DerogPubRec
NumTradesOpeninLast12M
PercentInstallTrades
PercentTradesNeverDelq
PercentTradesWBalance

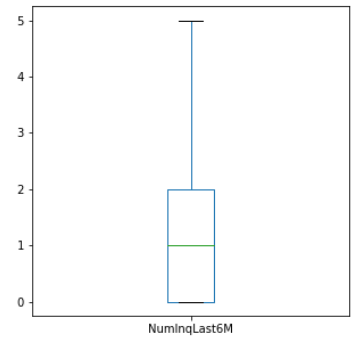

90
80
70
60
50
40
ExternalRiskEstimate

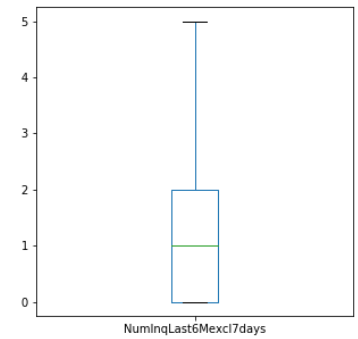

25
20
15
10
5
0
MSinceMostRecentTradeOpen

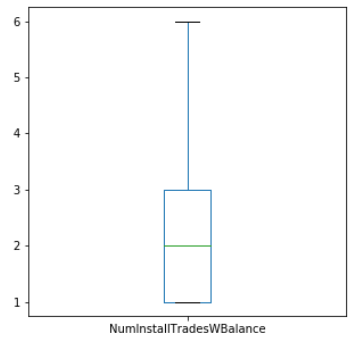

400
300
200
100
0
MSinceOldestTradeOpen

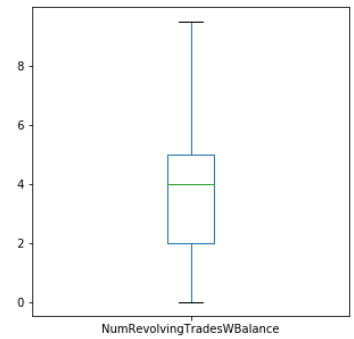

120
100
80
60
40
20
0
NetFractionRevolvingBurden

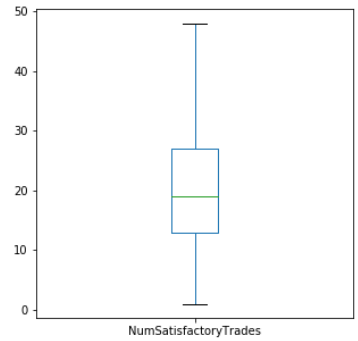

2.5
2.0
1.5
1.0
0.5
0.0
NumBank2NatlTradesWHighUtilization

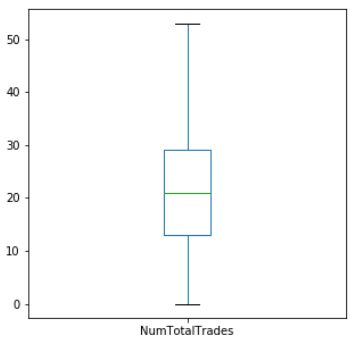

5
4
3
2
1
0
NumInqLast6M

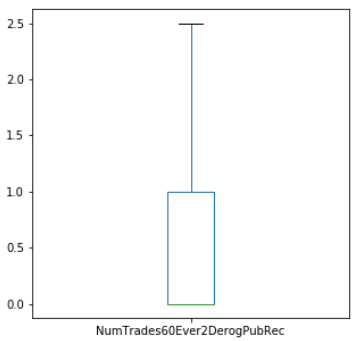

5
4
3
2
1
0
NumInqLast6Mexcl7days

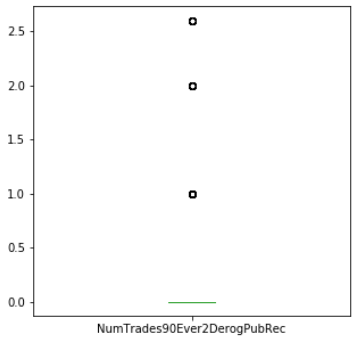

6
5
4
3
2
1
NumInstallTradesWBalance

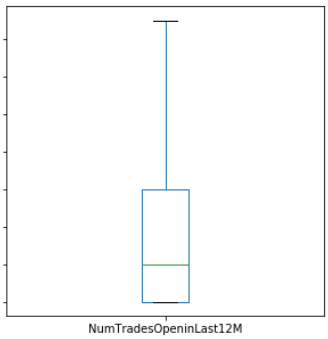

8
6
4
2
0
NumRevolvingTradesWBalance

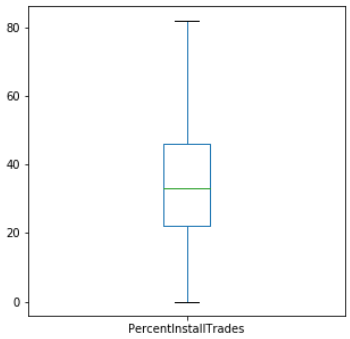

50
40
30
20
10
0
NumSatisfactoryTrades

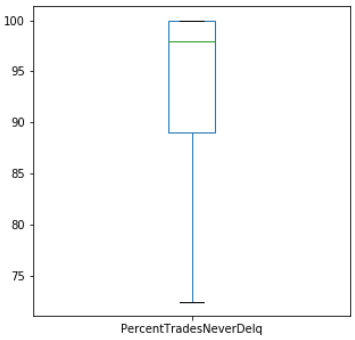

50
40
30
20
10
0
NumTotalTrades

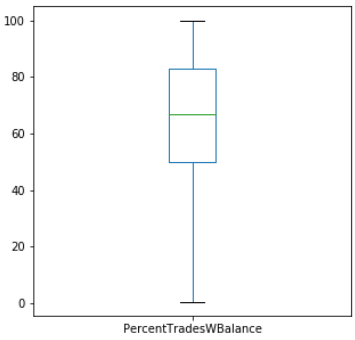

2.5
2.0
1.5
1.0
0.5
0.0
NumTrades60Ever2DerogPubRec


2.5
2.0
1.5
1.0
0.5
0.0
NumTrades90Ever2DerogPubRec


7
6
5
4
3
2
1
0
NumTradesOpeninLast12M


80
60
40
20
0
PercentInstallTrades


100
95
90
85
80
75
PercentTradesNeverDelq


100
80
60
40
20
0
PercentTradesWBalance