Data Quality Report - Initial Findings

Descriptive Statistics for continuous Feature

	count	mean	std	min	25%	50%	75%	max
AverageMInFile	939.0	78.570820	34.428568	5.0	56.0	75.0	96.0	245.0
PercentTradesNeverDelq	939.0	92.716720	11.654435	27.0	90.5	98.0	100.0	100.0
PercentInstallTrades	939.0	33.903088	17.344379	0.0	21.0	33.0	45.0	100.0
NetFractionRevolvingBurden	939.0	34.592119	29.089355	-8.0	9.0	30.0	56.0	135.0
NetFractionInstallBurden	939.0	42.795527	40.946332	-8.0	-8.0	51.0	80.0	190.0
PercentTradesWBalance	939.0	66.370607	21.439383	-8.0	50.0	67.0	82.5	100.0
ExternalRiskEstimate	939.0	72.023429	10.205327	-9.0	65.0	71.0	80.0	93.0
MSinceOldestTradeOpen	939.0	195.873269	106.183049	-8.0	128.5	183.0	259.0	565.0
MSinceMostRecentTradeOpen	939.0	9.367412	11.716637	0.0	3.0	6.0	12.0	152.0
NumSatisfactoryTrades	939.0	21.684771	11.766323	1.0	13.0	20.0	29.0	78.0
NumTrades60Ever2DerogPubRec	939.0	0.588924	1.224886	0.0	0.0	0.0	1.0	11.0
NumTrades90Ever2DerogPubRec	939.0	0.391906	0.954050	0.0	0.0	0.0	0.0	7.0
MSinceMostRecentDelq	939.0	8.002130	21.769043	-8.0	-7.0	-7.0	15.0	83.0
NumTotalTrades	939.0	22.992545	13.272563	0.0	14.0	21.0	31.0	100.0
NumTradesOpeninLast12M	939.0	1.903088	1.938330	0.0	0.0	1.0	3.0	19.0
MSinceMostRecentInqexcl7days	939.0	0.106496	5.919804	-8.0	0.0	0.0	1.0	23.0
NumInqLast6M	939.0	1.566560	2.072407	0.0	0.0	1.0	2.0	20.0
NumInqLast6Mexcl7days	939.0	1.510117	2.046541	0.0	0.0	1.0	2.0	20.0
NumRevolvingTradesWBalance	939.0	4.063898	3.310216	-8.0	2.0	3.0	5.5	21.0
NumInstallTradesWBalance	939.0	1.686901	3.301474	-8.0	1.0	2.0	3.0	13.0
NumBank2NatlTradesWHighUtilization	939.0	0.710330	2.522227	-8.0	0.0	1.0	2.0	12.0

Showing the full count of the continuous features.

```
RiskPerformance                  catego
ExternalRiskEstimate                int
MSinceOldestTradeOpen               int
MSinceMostRecentTradeOpen           int
AverageMInFile                    float
NumSatisfactoryTrades               int
NumTrades60Ever2DerogPubRec         int
NumTrades90Ever2DerogPubRec         int
PercentTradesNeverDelq            float
MSinceMostRecentDelq                int
MaxDelq2PublicRecLast12M         catego
MaxDelqEver                      catego
NumTotalTrades                      int
NumTradesOpeninLast12M              int
PercentInstallTrades              float
MSinceMostRecentInqexcl7days        int
NumInqLast6M                        int
NumInqLast6Mexcl7days               int
NetFractionRevolvingBurden        float
NetFractionInstallBurden          float
NumRevolvingTradesWBalance          int
NumInstallTradesWBalance            int
NumBank2NatlTradesWHighUtilization  int
PercentTradesWBalance             float
```

As a starting point all the columns have been converted into the type that would describe them better. Only three Columns are categorical to represent the dependent variables (ex. as either 0 or 1).
The numerical features have been further divided into 'Integer' and 'floating point' type for ratios or percentages.
There is a large amount of negative values on every column that will have to be considered for the data-cleaning step.

Each negative value meaning:

-9 No Bureau Record or No Investigation
-8 No Usable/Valid Trades or Inquiries
-7 Condition not Met (e.g. No Inquiries, No Delinquencies)

Out of the total 1000 rows, we have now 939 observations for the continuous features. Many features have substantial amounts of zeros, while some have outliers that will be considered in the report.

For a good half of all cases considered, a user has been kept not less then 75 months on average on file.

NumBank2NatlTradesWHighUtilization, ***NumTrades90Ever2DerogPubRec***, ***NumTrades60Ever2DerogPubR***, ***NumTradesOpeninLast12M*** containing many zero values, as high as 50% of data for most of them. All of these features indicate the frequency of trades, possibly credit applications, averaging from 1 or 2 trades, with some outliers as high as 23.

MSinceMostRecentDelq this feature contains 50% of '-7' values, meaning that no delinquency has occurred. Although, some outliers are present in the dataset, that needs to be addressed later on.
This number most likely corresponds to the users rated 'good'.

PercentTradesNeverDelq shows a percentage with three quarters of customers over the 90% already.

There are some unusable negative values that will be advisable to reduce or re-address using appropriate methods.

Descriptive statistics for categorical values

	count	unique	top	freq
RiskPerformance	939	2	Bad	475
MaxDelq2PublicRecLast12M	939	9	7	404
MaxDelqEver	939	7	8	436

Risk performances where a user paid as negotiated (within a period of 24 months) has been represented by the two dependent values of 'Bad' and 'Good', and are almost perfectly divided between the two, to represent all the users.

MaxDelq2PublicRecLast12M and **MaxDelqEver** are divided into levels from 1 to 7 for the first and 1 to 8 for the latter, to represent the level at which the user is placed in respect to having any delinquency, over the past 12 months for the former and ever for the latter.

Looking at the dataset, there is a 52% of 'Bad' in terms of risk performance.
Over the past 12 months the majority of observations show no delinquency, while for the whole period taken into consideration there is a majority of user that has no delinquency.

Histograms for all the continuous features

The average time for a user to be kept on file seems to be normally distributed with a light right skewedness (**AverageMInFile**).

Right skewed distribution, hence less then the median.

- **NumSatisfactoryTrades**, **NumTotalTrades**, **PercentInstallTrades, MSinceOldestTradeOpen**

Left skewed, hence more then the median.

- **PercentTradesWBalance**, **ExternalRiskEstimate**,

Decreasing with exponential behaviours.

- **MSinceMostRecentDelq, MSinceMostRecentTradeOpen, NumInqLast6M**, **NumInqLast6Mexcl7days**, **NumTrades90Ever2DerogPubRec, NumTrades60Ever2DerogPubRec, NumTradesOpeninLast12M**

PercentTradesNeverDelq shows an exponential increase behaviour that will be interesting to address later on.

NetFractionRevolvingBurden is bimodal.

Exponentially decreasing.

- **MSinceMostRecentInqexcl7days**, **NetFractionInstallBurden**, **NumInstallTradesWBalance** **NumBank2NatlTradesWHighUtilization**, **NumRevolvingTradesWBalance**.
  In these cases there is some noise that contaminates the data, and needs to be adjusted in future steps.

Box plots for all the continuous features

When displaying the dataset using a box plot method, it becomes apparent that there are many outliers on each of the columns provided. Most of these outliers have a great distance from the mean, and could be a good representation of variegation, and thus make sense within the context.

Bar chart for all the categorical features

Looking at the categorical features on the bar chart, the 'good' and 'bad' frequency is comparable, if not even for the number of occurrences, making for a good ratio when wanting to train the system on a later stage.

MaxDelq2PublicRecLast12M shows an overabundant number of '7s' for current and never delinquent, accounting for almost half of the observations on the past 12 months.

MaxDelqEver differ only for the greater interval, and is divided on half by non-delinquent (8), while the closes measure (6) states for the 30 days delinquent. Possibly indicating the half of the users has never been delinquent, while the portion represented by the 6s has been delinquent during the past 30 days.

MaxDelq2PublicRecLast12M

value	meaning
0	derogatory comment
1	120+ days delinquent
2	90 days delinquent
3	60 days delinquent
4	30 days delinquent
5, 6	unknown delinquency
7	current and never delinquent
8, 9	all other

MaxDelqEver

value	meaning
1	No such value
2	derogatory comment
3	120+ days delinquent
4	90 days delinquent
5	60 days delinquent
6	30 days delinquent
7	unknown delinquency
8	current and never delinquent
9	all other