Data Profile Dataframe - Notebook v15

data/credit\_train.csv

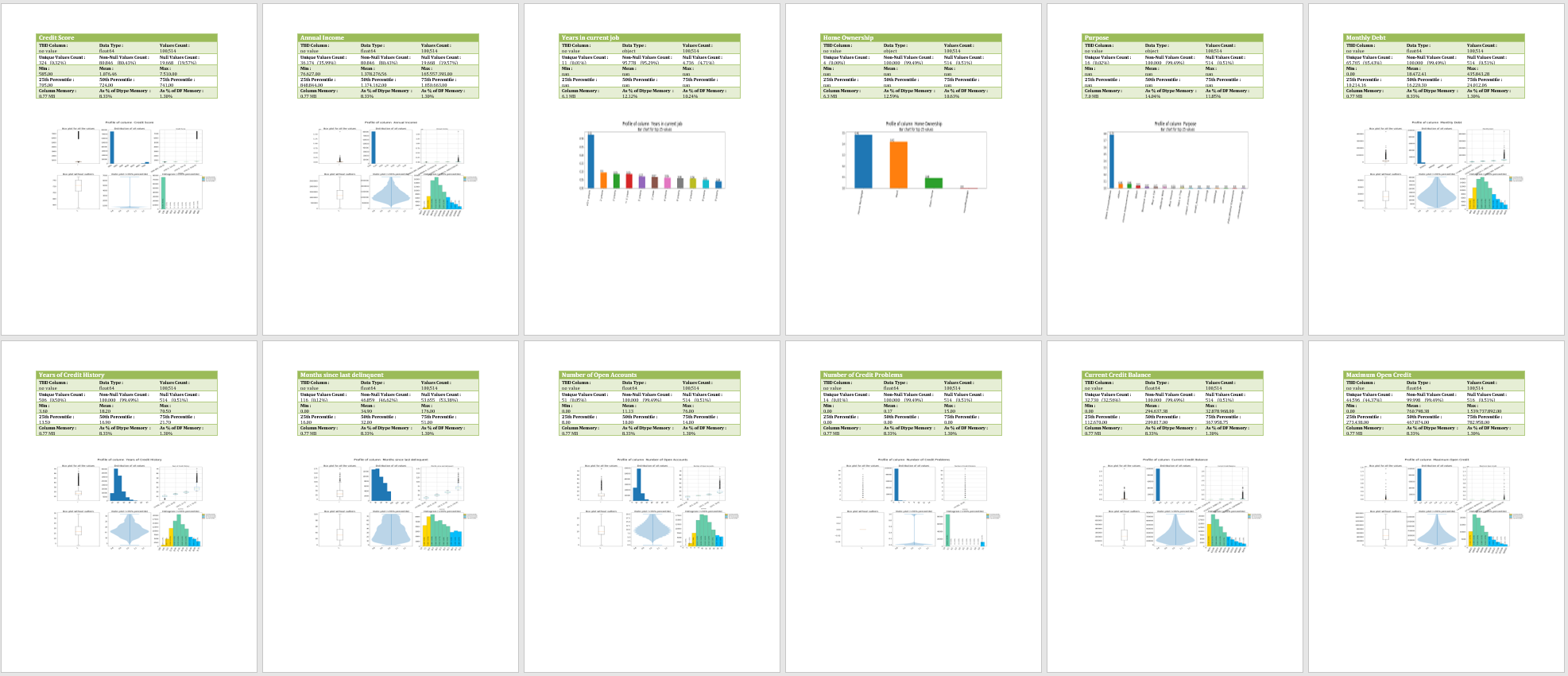
The main objective of this notebook is **only** to understand raw data profile. i.e. data type, min & max values, ranges, unique values, etc.

In consequent notebooks we will explore further on how to make decisions to make the data tidy and perform the data transformations based on the understanding of the data profile.

*The code is largely kept generic so that it could be used with any shape of data.*

# The Game Changer - Data Profile Dataframe (DPD)

The game changer for exploratory data analysis is the final **Data Profile Dataframe** that is generated which combines **all** the information required to inform data cleaning, tidy data and optimisations (memory and processing) decisions. Instead of using various Pandas commands at different instances and going back and forth to cross refer information, Data Profile Dataframe brings all information into a single dataframe. This will be very useful when reviewing the data profile with the business subject matter or other team members as all information related to data profile is in a single easy to understand format.



Understanding the data is **the critical step** in preparing the data to be used for analytics. As many experts will point out the data preparation and transforming the data into a tidy format takes about 80% of the effort in any data analytics or data analysis project.

**Understanding the data requires good understanding of the domain and/or access to a subjectmatter expert (SME) to help make decisions about data quality and data usage:**

* What are the columns and what do they mean?
* How to interpret each columns and possible values of a column?
* Should the columns be renamed (and cleaned e.g. trim)?
* Are there columns that may have similar information that could be dropped in favour of one master column?
* Can columns with no values (or all empty) be dropped?
* Can columns which have more than certain threshold of blank values be dropped?
* Can rows that have missing values for certain columns or combination of columns be dropped?

i.e. the row is meaningless wihtout those values.

* Can the numeric data type columns be converted / down casted to optimise memory usage based on the data values?
* or will there be outliers possibly in future data sets that we cannot do this?
* Can the min and max values be used to determine the lowest possible data type?
* Can some string/object columns be converted to Category types?
* based on count of unique values
* Can any columns be discarded that may not be required for analytics?

Columns Data Profile Summary

# Dataset shape

|  |  |
| --- | --- |
| No.of rows | No.of columns |
| 100,514 | 19 |

# Dataframe columns summary

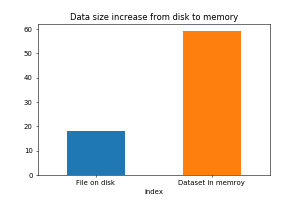
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| column\_name | col\_data\_type | non\_null\_values | null\_values | count |
| Loan ID | object | 100,000 | 514 | 100,514 |
| Customer ID | object | 100,000 | 514 | 100,514 |
| Loan Status | object | 100,000 | 514 | 100,514 |
| Current Loan Amount | float64 | 100,000 | 514 | 100,514 |
| Term | object | 100,000 | 514 | 100,514 |
| Credit Score | float64 | 80,846 | 19,668 | 100,514 |
| Annual Income | float64 | 80,846 | 19,668 | 100,514 |
| Years in current job | object | 95,778 | 4,736 | 100,514 |
| Home Ownership | object | 100,000 | 514 | 100,514 |
| Purpose | object | 100,000 | 514 | 100,514 |
| Monthly Debt | float64 | 100,000 | 514 | 100,514 |
| Years of Credit History | float64 | 100,000 | 514 | 100,514 |
| Months since last delinquent | float64 | 46,859 | 53,655 | 100,514 |
| Number of Open Accounts | float64 | 100,000 | 514 | 100,514 |
| Number of Credit Problems | float64 | 100,000 | 514 | 100,514 |
| Current Credit Balance | float64 | 100,000 | 514 | 100,514 |
| Maximum Open Credit | float64 | 99,998 | 516 | 100,514 |
| Bankruptcies | float64 | 99,796 | 718 | 100,514 |
| Tax Liens | float64 | 99,990 | 524 | 100,514 |

Memory Usage Profile

# Data file size on disk vs. dataset size in memory

|  |  |
| --- | --- |
| Description | Size in MB |
| Data file size on disk | 17.94 |
| Dataset size in memory | 59.12 |

Dataset increase in memory : **229.51%**



# Dataframe column types and size in memory

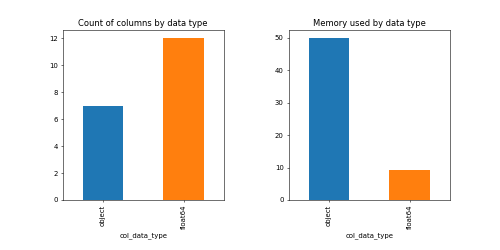
|  |  |  |  |
| --- | --- | --- | --- |
| col\_data\_type | dtype\_count | dtype\_total | dtype\_%\_total\_mem |
| object | 7.00 | 49.92 | 84.44 |
| float64 | 12.00 | 9.20 | 15.56 |

"col\_data\_type" : Column data type

"dtype\_count" : Number of oclumns in the dataset of the given data type

"dtype\_total" : Total memory in MB for the given data type

"dtype\_%\_total\_mem" : Percentage of the memory used by the given data type out of the total memory used by the dataset



In a memory heavy datasets the above information can shed light into which data type you need to focus if you need to optimise the memory usage.

e.g. may be convert "object" datatype to "category" type if the cardinality is low or may be down cast "float64" to float16 or smaller.

These decision need further information on column cardinality and max/min values which are covered in the next few sections.

# Memory used by "object" data type

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| column\_name | col\_memory | %\_of\_dtype\_mem | %\_of\_total\_memory | unique\_values\_count |
| Loan Status | 6.43 | 12.87 | 10.87 | 2.00 |
| Term | 6.38 | 12.78 | 10.79 | 2.00 |
| Home Ownership | 6.29 | 12.59 | 10.63 | 4.00 |
| Years in current job | 6.05 | 12.12 | 10.24 | 11.00 |
| Purpose | 7.01 | 14.04 | 11.85 | 16.00 |
| Loan ID | 8.88 | 17.80 | 15.03 | 81,999.00 |
| Customer ID | 8.88 | 17.80 | 15.03 | 81,999.00 |

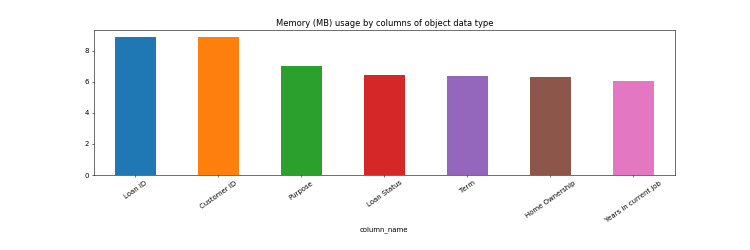
"column\_name" : Name of the column in the dataframe

"col\_memory" : Memory used by the given column

"%\_of\_dtype\_mem" : Percentage of memory used by the given column out of memory used by the column data type

"%\_of\_total\_memory" : Percentage of the memory used by the given column out of the total memory used by the dataset

"unique\_values\_count" : Count of the unique values for the given column



Analysing how many unique values an 'object' column has will be useful to detrminewhich columns are good candidates for \*Categorical\* data type. In combination with the total memory used by 'object'data type and each 'object' data type column, decisions can be made on converting them Category type.Object or string data type columns with low cardinality is suitable for Category type.**The threshold of 'low cardinality' depends on the domain of the data and data usage patterns.**

# Memory used by "Non-Object" data type

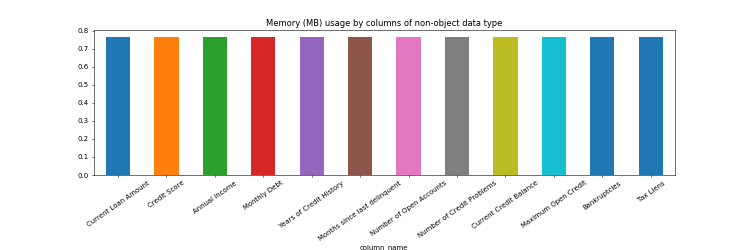
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| column\_name | col\_memory | %\_of\_dtype\_mem | %\_of\_total\_memory | unique\_values\_count |
| Bankruptcies | 0.77 | 8.33 | 1.30 | 8.00 |
| Tax Liens | 0.77 | 8.33 | 1.30 | 12.00 |
| Number of Credit Problems | 0.77 | 8.33 | 1.30 | 14.00 |
| Number of Open Accounts | 0.77 | 8.33 | 1.30 | 51.00 |
| Months since last delinquent | 0.77 | 8.33 | 1.30 | 116.00 |
| Credit Score | 0.77 | 8.33 | 1.30 | 324.00 |
| Years of Credit History | 0.77 | 8.33 | 1.30 | 506.00 |
| Current Loan Amount | 0.77 | 8.33 | 1.30 | 22,004.00 |
| Current Credit Balance | 0.77 | 8.33 | 1.30 | 32,730.00 |
| Annual Income | 0.77 | 8.33 | 1.30 | 36,174.00 |
| Maximum Open Credit | 0.77 | 8.33 | 1.30 | 44,596.00 |
| Monthly Debt | 0.77 | 8.33 | 1.30 | 65,765.00 |

"column\_name" : Name of the column in the dataframe

"col\_memory" : Memory used by the given column

"%\_of\_dtype\_mem" : Percentage of memory used by the given column out of memory used by the column data type

"%\_of\_total\_memory" : Percentage of the memory used by the given column out of the total memory used by the dataset



By analysing the min and max values of the numeric columns decions can be made to downcast the data type to more memory efficient storage types.

# Columns with non-null values less than 75.00%

The columns should contain at least 75,386 (75.00%) non-empty rows out of 100,514 rows to be considered useful.

The non-empty values threshold can be set using the threshold\_perc variable in the code.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| column\_name | col\_data\_type | col\_memory | non\_null\_values | %\_of\_non\_nulls | null\_values | %\_of\_nulls |
| Months since last delinquent | float64 | 0.77 | 46,859.00 | 46.62 | 53,655.00 | 53.38 |

"column\_name" : Name of the column in the dataframe

"col\_data\_type" : Data type of the given column

"col\_memory" : Memory used by the given column

"non\_null\_values" : Count of non-null values in the given column

"%\_of\_non\_nulls" : Percentage of the non-null values out of total values for the given column

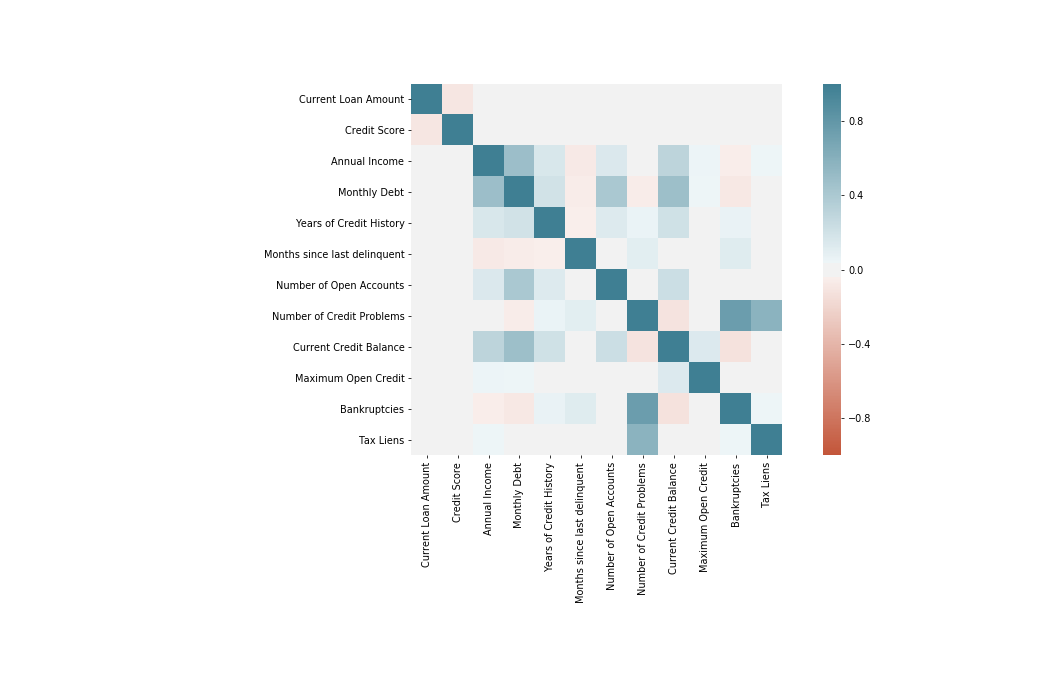
"null\_values" : Count of null values in the given column

"%\_of\_nulls" : Percentage of the null values out of total values for the given column

Generally columns with large percentage of empty values can be \*dropped\* from the dataset as they will not add any value to the analysis.

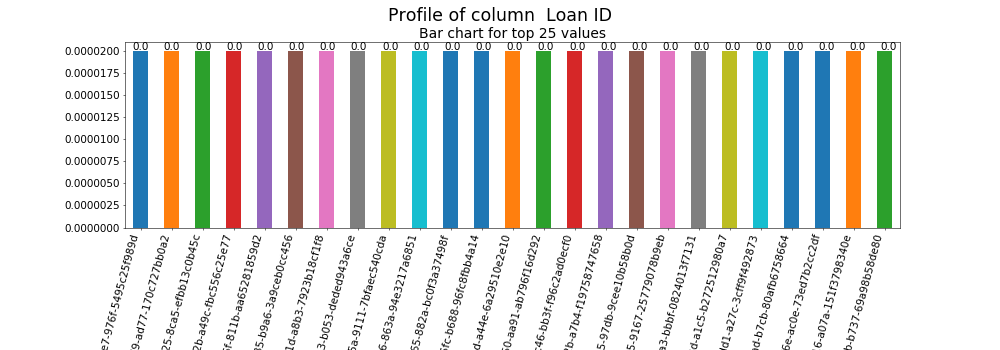
**But this depends on the domian of the dataset and usage pattern of the columns/data.**

Data correlation plot

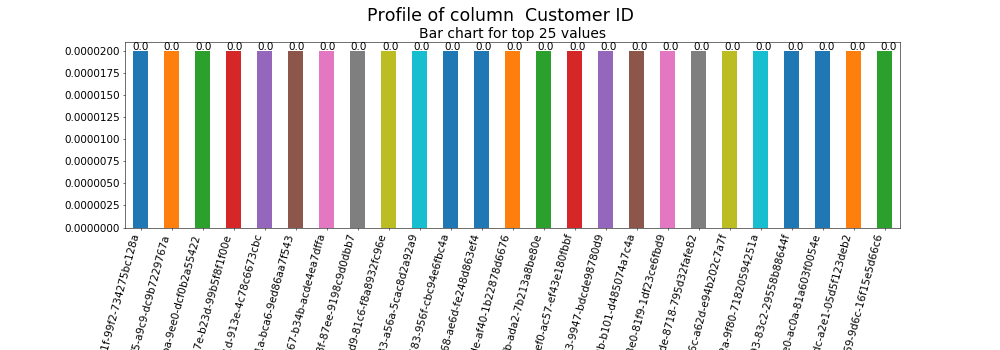


Column Data Profile Details

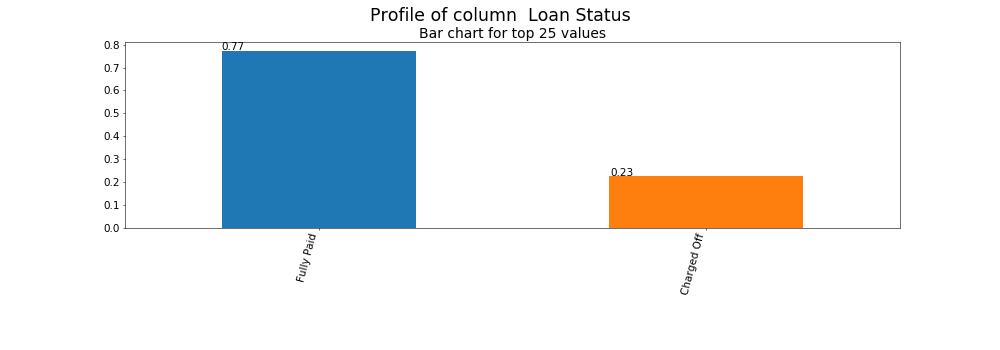
|  |  |  |
| --- | --- | --- |
| Loan ID | | |
| TBD Column : no value | **Data Type :** object | **Values Count :** 100,514 |
| Unique Values Count :  81,999 (81.58%) | **Non-Null Values Count :** 100,000 (99.49%) | **Null Values Count :** 514 (0.51%) |
| Min :  nan | **Mean :** nan | **Max :** nan |
| 25th Percentile :  nan | **50th Percentile :** nan | **75th Percentile :** nan |
| Column Memory :  8.9 MB | **As % of Dtype Memory :** 17.80% | **As % of DF Memory :** 15.03% |



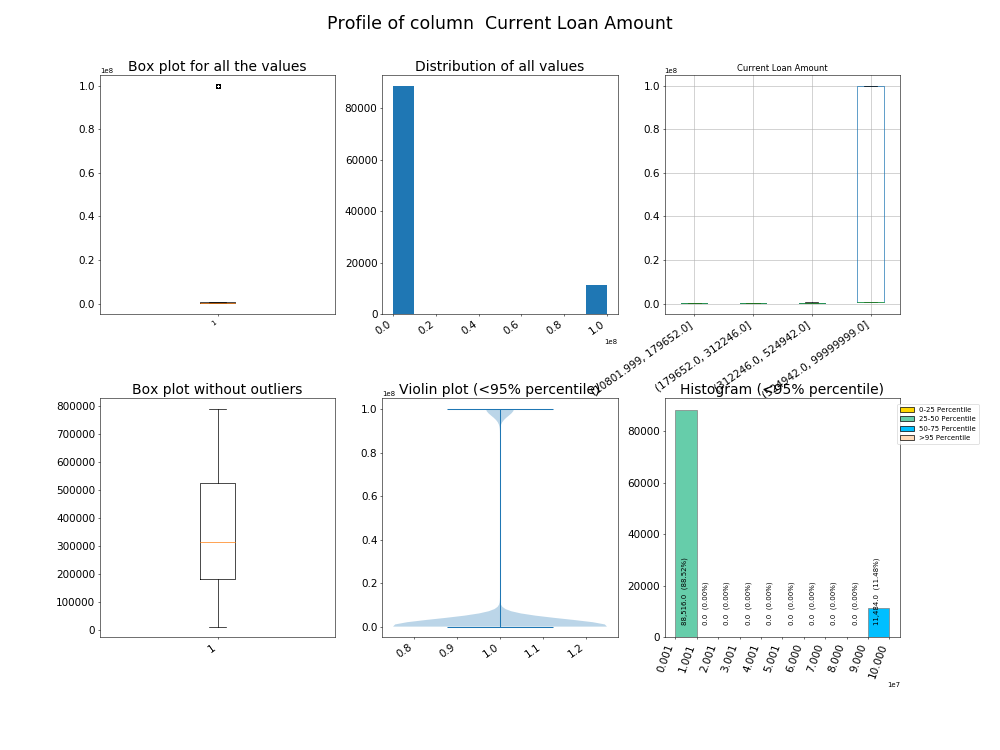
|  |  |  |
| --- | --- | --- |
| Customer ID | | |
| TBD Column : no value | **Data Type :** object | **Values Count :** 100,514 |
| Unique Values Count :  81,999 (81.58%) | **Non-Null Values Count :** 100,000 (99.49%) | **Null Values Count :** 514 (0.51%) |
| Min :  nan | **Mean :** nan | **Max :** nan |
| 25th Percentile :  nan | **50th Percentile :** nan | **75th Percentile :** nan |
| Column Memory :  8.9 MB | **As % of Dtype Memory :** 17.80% | **As % of DF Memory :** 15.03% |



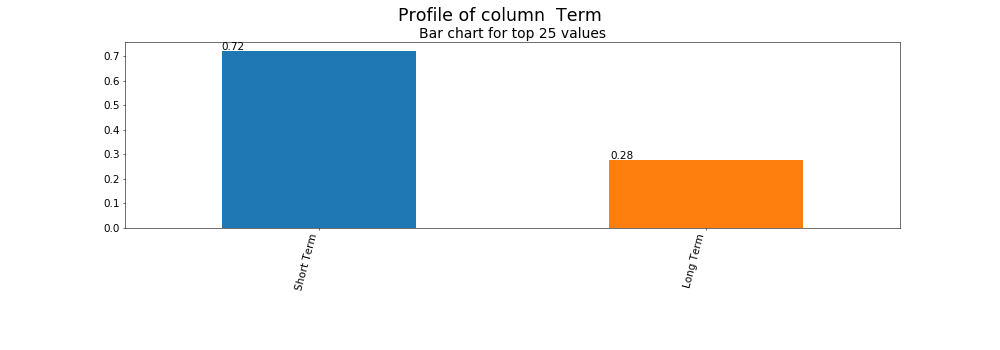
|  |  |  |
| --- | --- | --- |
| Loan Status | | |
| TBD Column : no value | **Data Type :** object | **Values Count :** 100,514 |
| Unique Values Count :  2 (0.00%) | **Non-Null Values Count :** 100,000 (99.49%) | **Null Values Count :** 514 (0.51%) |
| Min :  nan | **Mean :** nan | **Max :** nan |
| 25th Percentile :  nan | **50th Percentile :** nan | **75th Percentile :** nan |
| Column Memory :  6.4 MB | **As % of Dtype Memory :** 12.87% | **As % of DF Memory :** 10.87% |



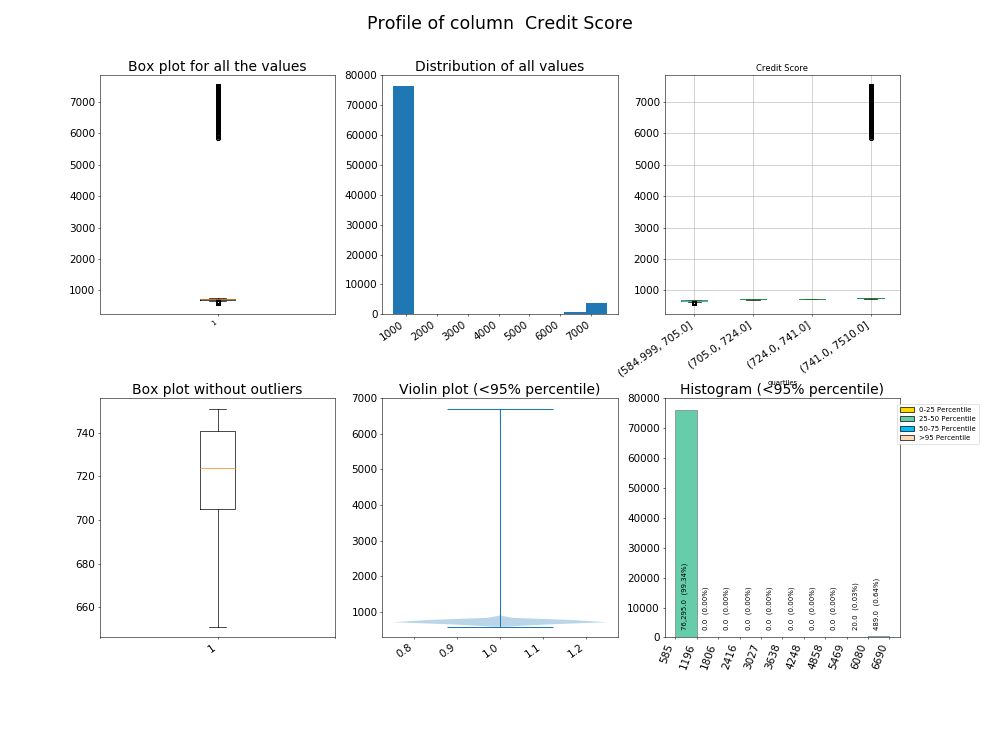
|  |  |  |
| --- | --- | --- |
| Current Loan Amount | | |
| TBD Column : no value | **Data Type :** float64 | **Values Count :** 100,514 |
| Unique Values Count :  22,004 (21.89%) | **Non-Null Values Count :** 100,000 (99.49%) | **Null Values Count :** 514 (0.51%) |
| Min :  10,802.00 | **Mean :** 11,760,447.39 | **Max :** 99,999,999.00 |
| 25th Percentile :  179,652.00 | **50th Percentile :** 312,246.00 | **75th Percentile :** 524,942.00 |
| Column Memory :  0.77 MB | **As % of Dtype Memory :** 8.33% | **As % of DF Memory :** 1.30% |



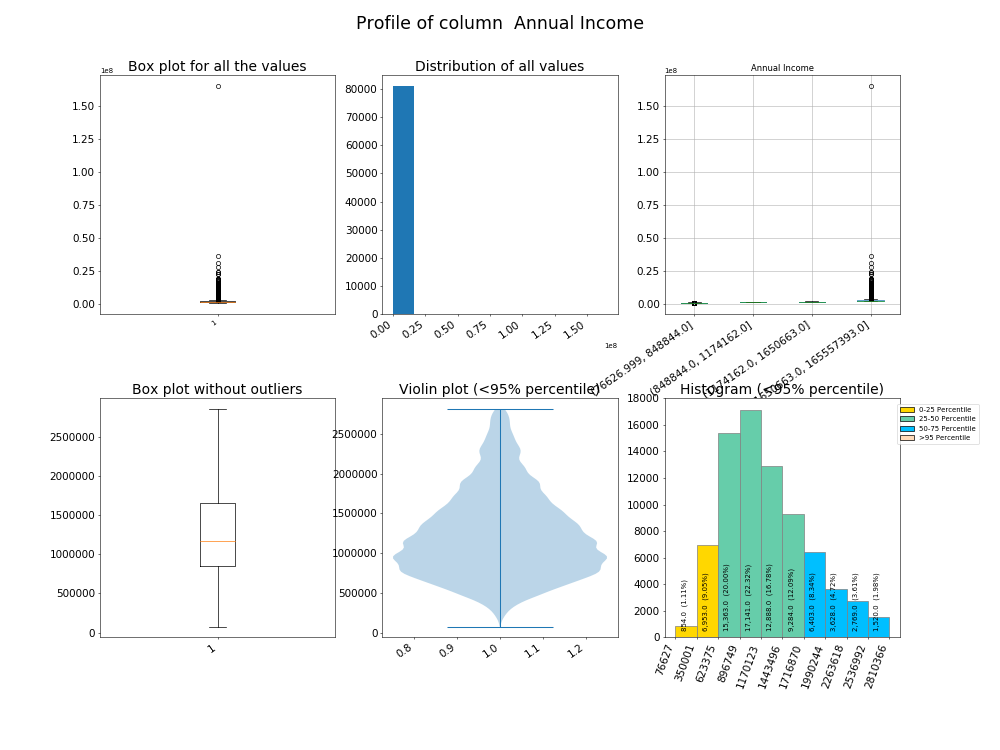
|  |  |  |
| --- | --- | --- |
| Term | | |
| TBD Column : no value | **Data Type :** object | **Values Count :** 100,514 |
| Unique Values Count :  2 (0.00%) | **Non-Null Values Count :** 100,000 (99.49%) | **Null Values Count :** 514 (0.51%) |
| Min :  nan | **Mean :** nan | **Max :** nan |
| 25th Percentile :  nan | **50th Percentile :** nan | **75th Percentile :** nan |
| Column Memory :  6.4 MB | **As % of Dtype Memory :** 12.78% | **As % of DF Memory :** 10.79% |



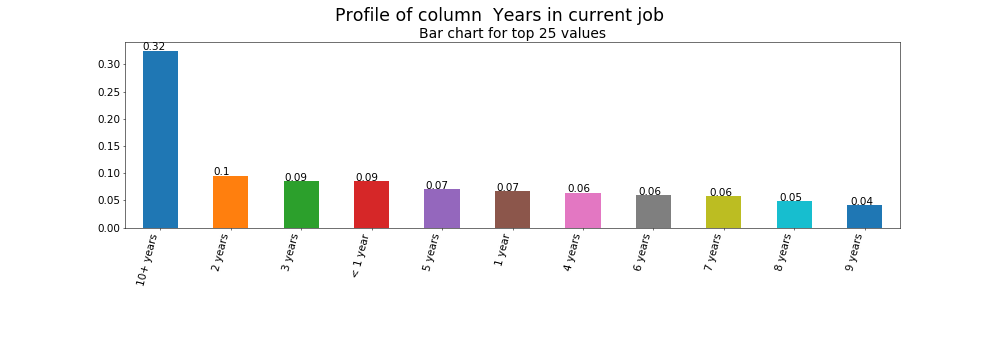
|  |  |  |
| --- | --- | --- |
| Credit Score | | |
| TBD Column : no value | **Data Type :** float64 | **Values Count :** 100,514 |
| Unique Values Count :  324 (0.32%) | **Non-Null Values Count :** 80,846 (80.43%) | **Null Values Count :** 19,668 (19.57%) |
| Min :  585.00 | **Mean :** 1,076.46 | **Max :** 7,510.00 |
| 25th Percentile :  705.00 | **50th Percentile :** 724.00 | **75th Percentile :** 741.00 |
| Column Memory :  0.77 MB | **As % of Dtype Memory :** 8.33% | **As % of DF Memory :** 1.30% |



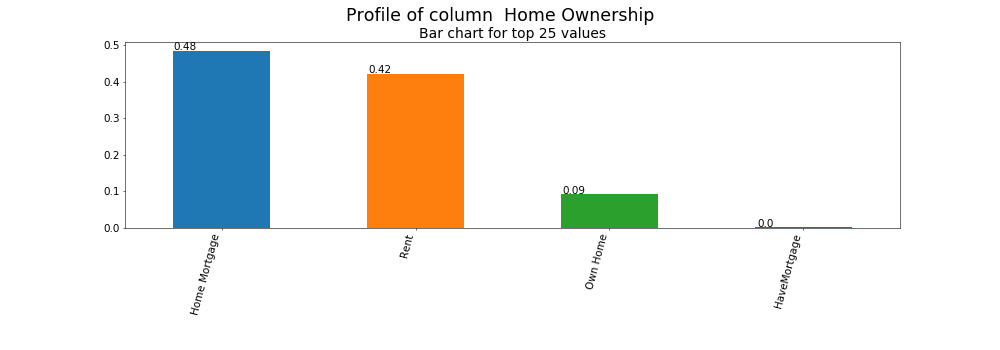
|  |  |  |
| --- | --- | --- |
| Annual Income | | |
| TBD Column : no value | **Data Type :** float64 | **Values Count :** 100,514 |
| Unique Values Count :  36,174 (35.99%) | **Non-Null Values Count :** 80,846 (80.43%) | **Null Values Count :** 19,668 (19.57%) |
| Min :  76,627.00 | **Mean :** 1,378,276.56 | **Max :** 165,557,393.00 |
| 25th Percentile :  848,844.00 | **50th Percentile :** 1,174,162.00 | **75th Percentile :** 1,650,663.00 |
| Column Memory :  0.77 MB | **As % of Dtype Memory :** 8.33% | **As % of DF Memory :** 1.30% |



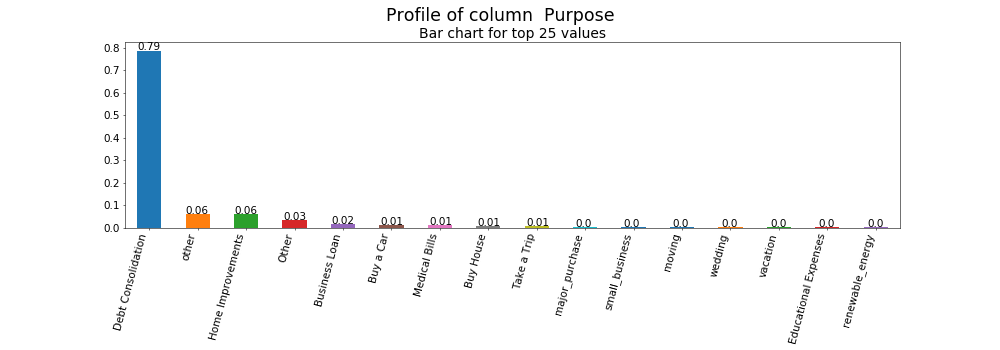
|  |  |  |
| --- | --- | --- |
| Years in current job | | |
| TBD Column : no value | **Data Type :** object | **Values Count :** 100,514 |
| Unique Values Count :  11 (0.01%) | **Non-Null Values Count :** 95,778 (95.29%) | **Null Values Count :** 4,736 (4.71%) |
| Min :  nan | **Mean :** nan | **Max :** nan |
| 25th Percentile :  nan | **50th Percentile :** nan | **75th Percentile :** nan |
| Column Memory :  6.1 MB | **As % of Dtype Memory :** 12.12% | **As % of DF Memory :** 10.24% |



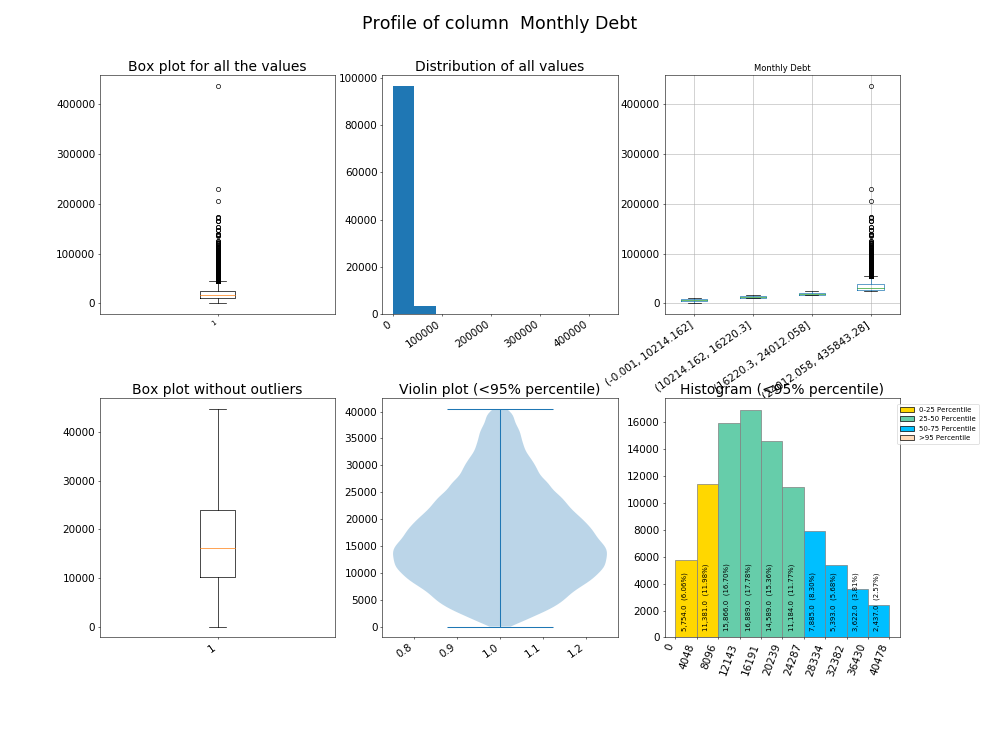
|  |  |  |
| --- | --- | --- |
| Home Ownership | | |
| TBD Column : no value | **Data Type :** object | **Values Count :** 100,514 |
| Unique Values Count :  4 (0.00%) | **Non-Null Values Count :** 100,000 (99.49%) | **Null Values Count :** 514 (0.51%) |
| Min :  nan | **Mean :** nan | **Max :** nan |
| 25th Percentile :  nan | **50th Percentile :** nan | **75th Percentile :** nan |
| Column Memory :  6.3 MB | **As % of Dtype Memory :** 12.59% | **As % of DF Memory :** 10.63% |



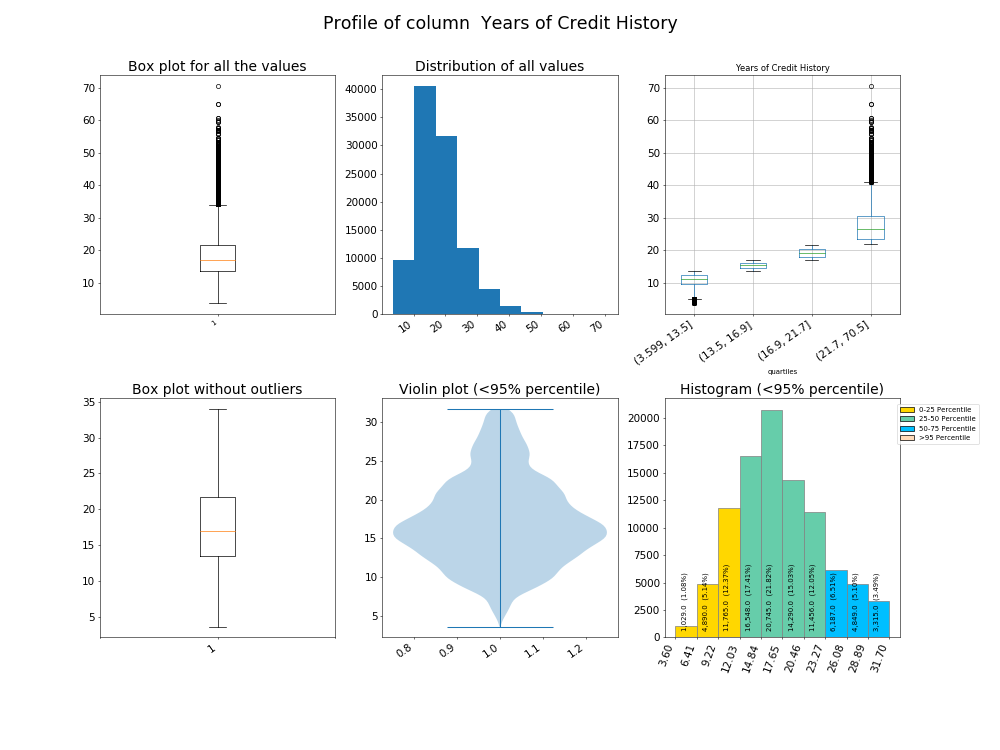
|  |  |  |
| --- | --- | --- |
| Purpose | | |
| TBD Column : no value | **Data Type :** object | **Values Count :** 100,514 |
| Unique Values Count :  16 (0.02%) | **Non-Null Values Count :** 100,000 (99.49%) | **Null Values Count :** 514 (0.51%) |
| Min :  nan | **Mean :** nan | **Max :** nan |
| 25th Percentile :  nan | **50th Percentile :** nan | **75th Percentile :** nan |
| Column Memory :  7.0 MB | **As % of Dtype Memory :** 14.04% | **As % of DF Memory :** 11.85% |



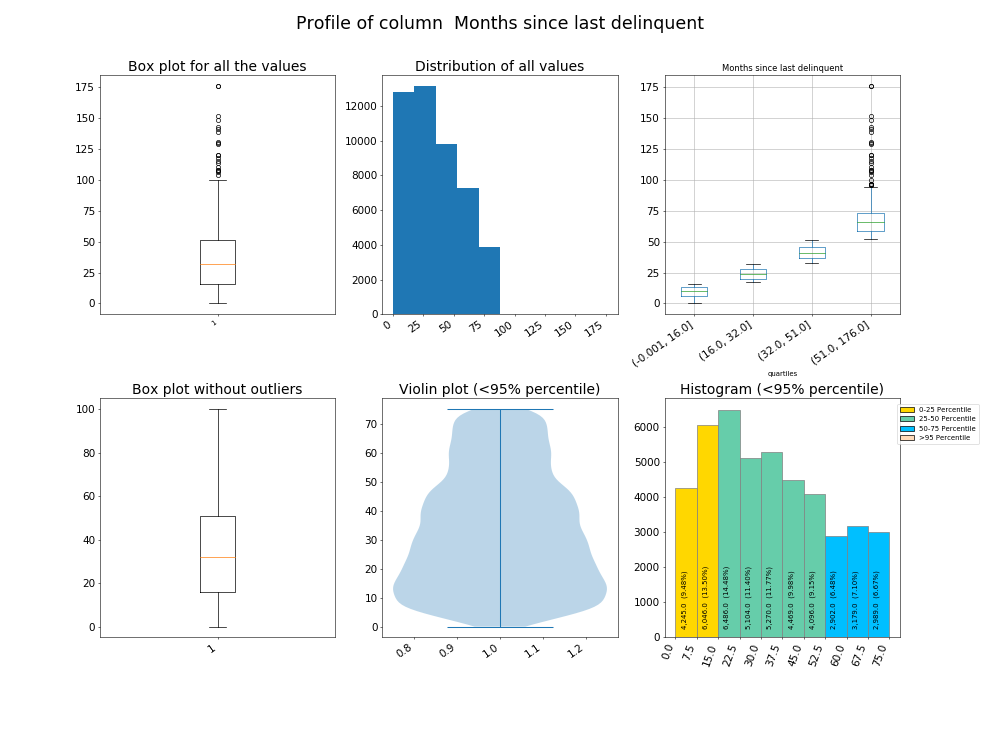
|  |  |  |
| --- | --- | --- |
| Monthly Debt | | |
| TBD Column : no value | **Data Type :** float64 | **Values Count :** 100,514 |
| Unique Values Count :  65,765 (65.43%) | **Non-Null Values Count :** 100,000 (99.49%) | **Null Values Count :** 514 (0.51%) |
| Min :  0.00 | **Mean :** 18,472.41 | **Max :** 435,843.28 |
| 25th Percentile :  10,214.16 | **50th Percentile :** 16,220.30 | **75th Percentile :** 24,012.06 |
| Column Memory :  0.77 MB | **As % of Dtype Memory :** 8.33% | **As % of DF Memory :** 1.30% |



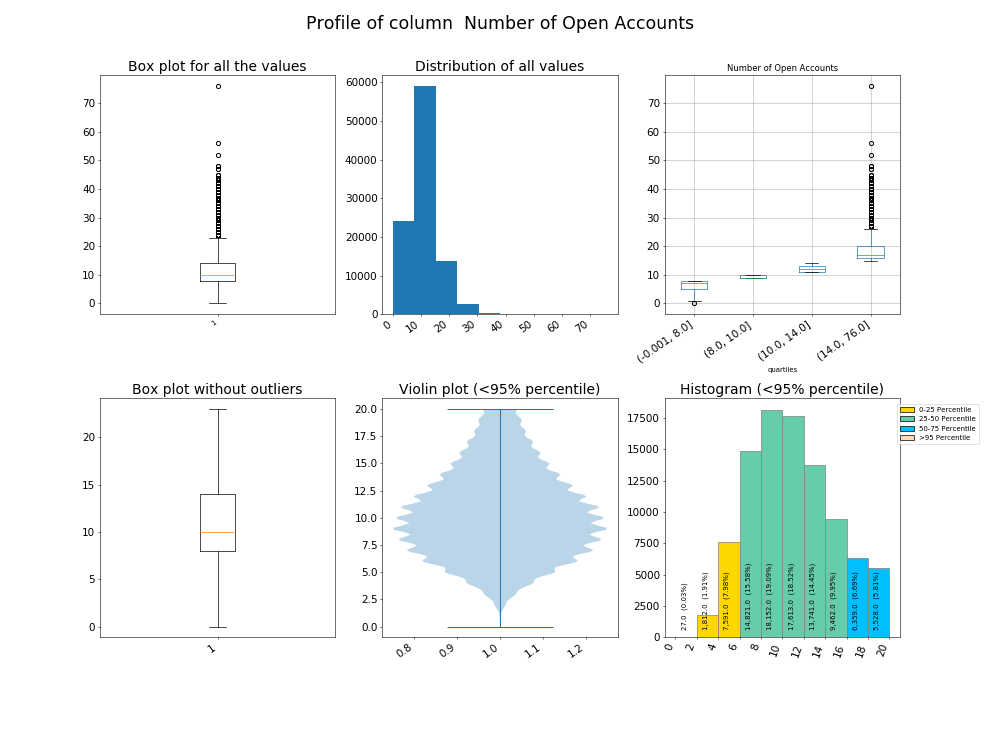
|  |  |  |
| --- | --- | --- |
| Years of Credit History | | |
| TBD Column : no value | **Data Type :** float64 | **Values Count :** 100,514 |
| Unique Values Count :  506 (0.50%) | **Non-Null Values Count :** 100,000 (99.49%) | **Null Values Count :** 514 (0.51%) |
| Min :  3.60 | **Mean :** 18.20 | **Max :** 70.50 |
| 25th Percentile :  13.50 | **50th Percentile :** 16.90 | **75th Percentile :** 21.70 |
| Column Memory :  0.77 MB | **As % of Dtype Memory :** 8.33% | **As % of DF Memory :** 1.30% |



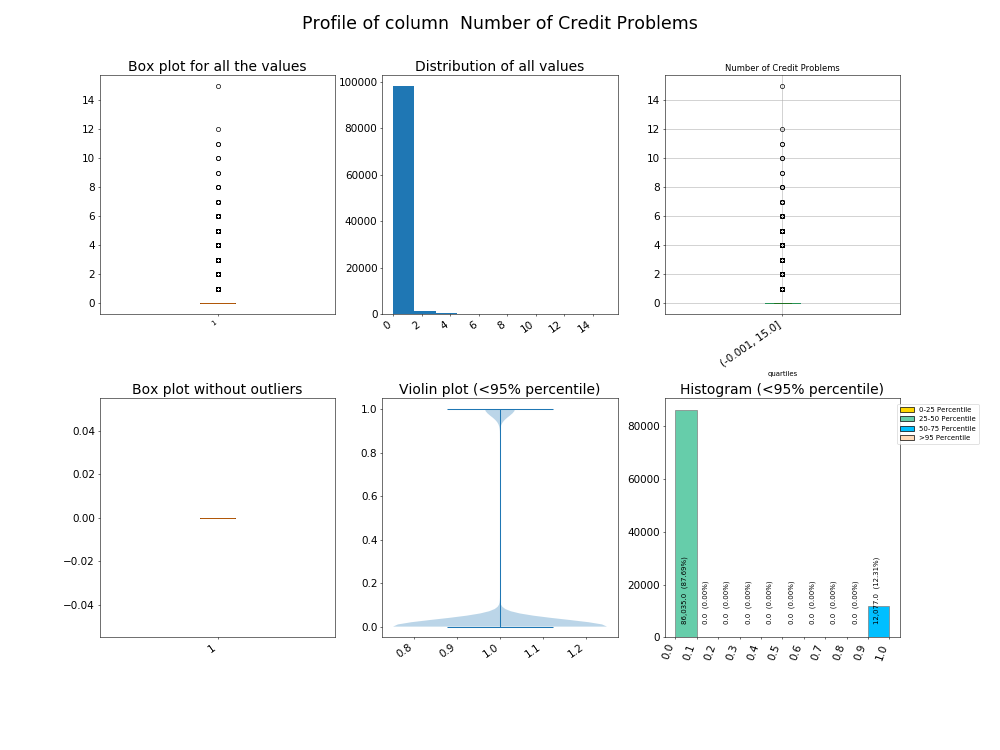
|  |  |  |
| --- | --- | --- |
| Months since last delinquent | | |
| TBD Column : no value | **Data Type :** float64 | **Values Count :** 100,514 |
| Unique Values Count :  116 (0.12%) | **Non-Null Values Count :** 46,859 (46.62%) | **Null Values Count :** 53,655 (53.38%) |
| Min :  0.00 | **Mean :** 34.90 | **Max :** 176.00 |
| 25th Percentile :  16.00 | **50th Percentile :** 32.00 | **75th Percentile :** 51.00 |
| Column Memory :  0.77 MB | **As % of Dtype Memory :** 8.33% | **As % of DF Memory :** 1.30% |



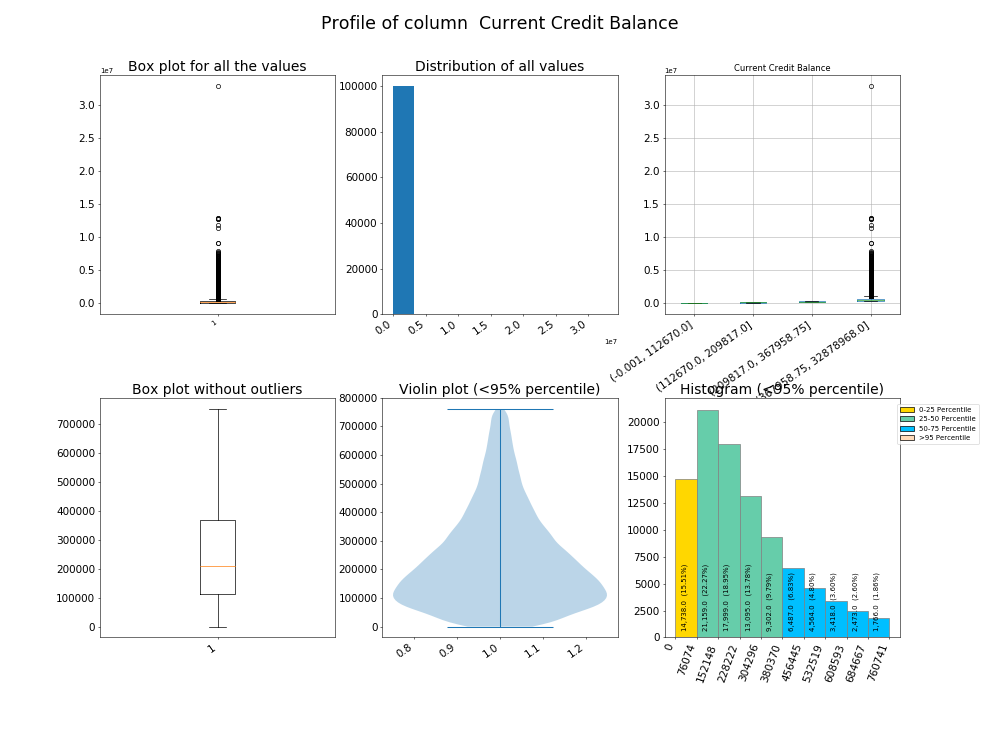
|  |  |  |
| --- | --- | --- |
| Number of Open Accounts | | |
| TBD Column : no value | **Data Type :** float64 | **Values Count :** 100,514 |
| Unique Values Count :  51 (0.05%) | **Non-Null Values Count :** 100,000 (99.49%) | **Null Values Count :** 514 (0.51%) |
| Min :  0.00 | **Mean :** 11.13 | **Max :** 76.00 |
| 25th Percentile :  8.00 | **50th Percentile :** 10.00 | **75th Percentile :** 14.00 |
| Column Memory :  0.77 MB | **As % of Dtype Memory :** 8.33% | **As % of DF Memory :** 1.30% |



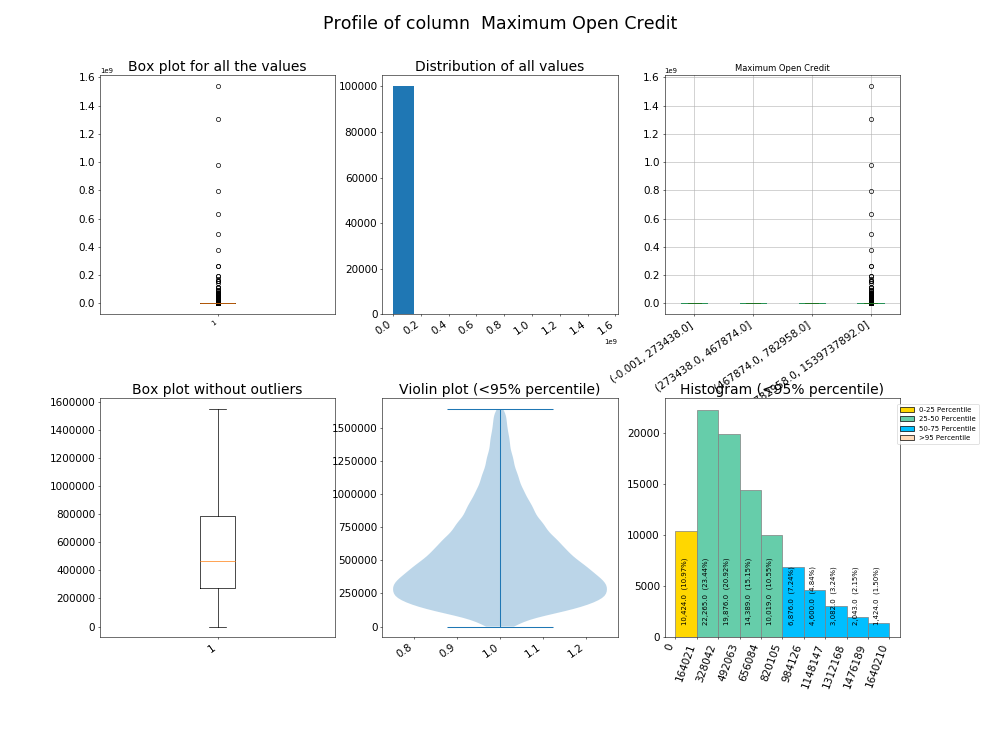
|  |  |  |
| --- | --- | --- |
| Number of Credit Problems | | |
| TBD Column : no value | **Data Type :** float64 | **Values Count :** 100,514 |
| Unique Values Count :  14 (0.01%) | **Non-Null Values Count :** 100,000 (99.49%) | **Null Values Count :** 514 (0.51%) |
| Min :  0.00 | **Mean :** 0.17 | **Max :** 15.00 |
| 25th Percentile :  0.00 | **50th Percentile :** 0.00 | **75th Percentile :** 0.00 |
| Column Memory :  0.77 MB | **As % of Dtype Memory :** 8.33% | **As % of DF Memory :** 1.30% |



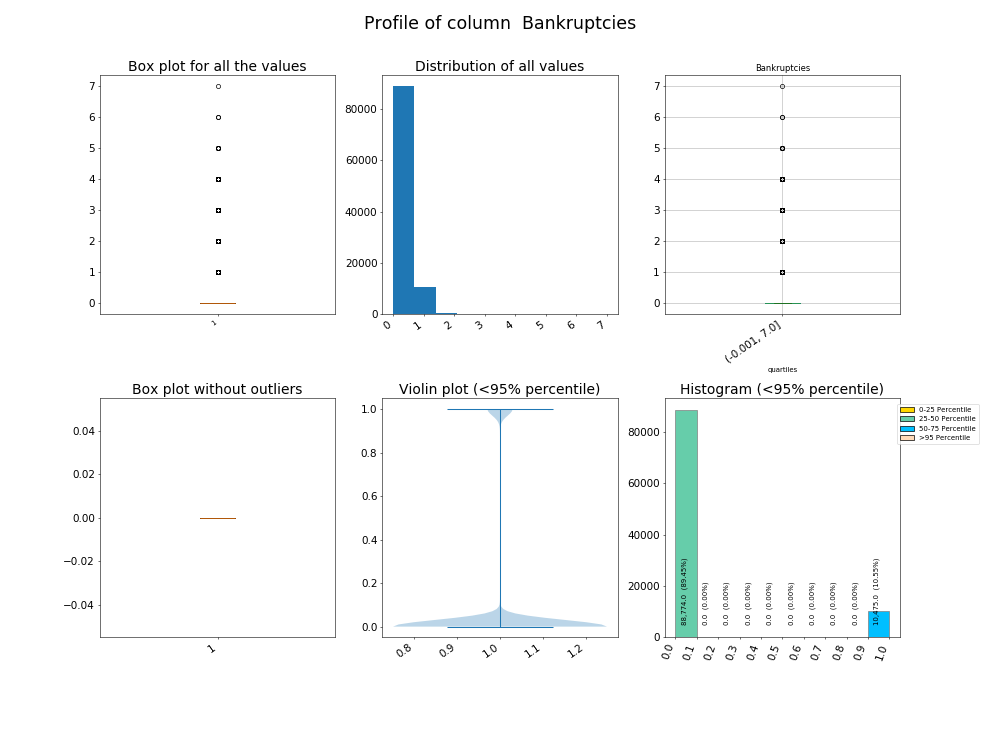
|  |  |  |
| --- | --- | --- |
| Current Credit Balance | | |
| TBD Column : no value | **Data Type :** float64 | **Values Count :** 100,514 |
| Unique Values Count :  32,730 (32.56%) | **Non-Null Values Count :** 100,000 (99.49%) | **Null Values Count :** 514 (0.51%) |
| Min :  0.00 | **Mean :** 294,637.38 | **Max :** 32,878,968.00 |
| 25th Percentile :  112,670.00 | **50th Percentile :** 209,817.00 | **75th Percentile :** 367,958.75 |
| Column Memory :  0.77 MB | **As % of Dtype Memory :** 8.33% | **As % of DF Memory :** 1.30% |



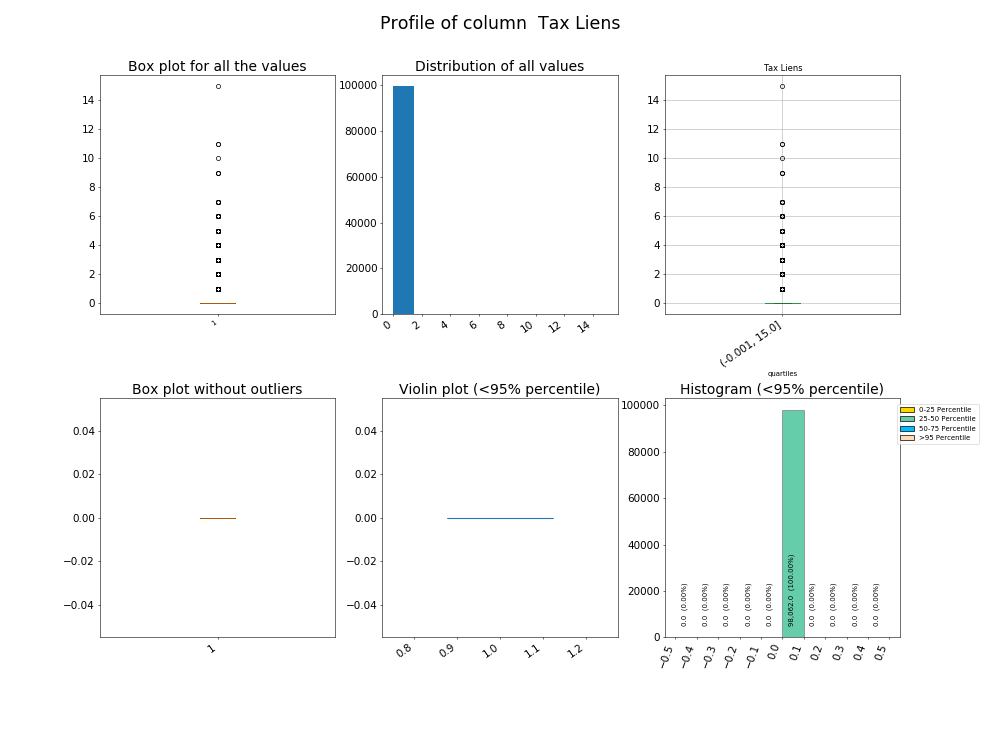
|  |  |  |
| --- | --- | --- |
| Maximum Open Credit | | |
| TBD Column : no value | **Data Type :** float64 | **Values Count :** 100,514 |
| Unique Values Count :  44,596 (44.37%) | **Non-Null Values Count :** 99,998 (99.49%) | **Null Values Count :** 516 (0.51%) |
| Min :  0.00 | **Mean :** 760,798.38 | **Max :** 1,539,737,892.00 |
| 25th Percentile :  273,438.00 | **50th Percentile :** 467,874.00 | **75th Percentile :** 782,958.00 |
| Column Memory :  0.77 MB | **As % of Dtype Memory :** 8.33% | **As % of DF Memory :** 1.30% |



|  |  |  |
| --- | --- | --- |
| Bankruptcies | | |
| TBD Column : no value | **Data Type :** float64 | **Values Count :** 100,514 |
| Unique Values Count :  8 (0.01%) | **Non-Null Values Count :** 99,796 (99.29%) | **Null Values Count :** 718 (0.71%) |
| Min :  0.00 | **Mean :** 0.12 | **Max :** 7.00 |
| 25th Percentile :  0.00 | **50th Percentile :** 0.00 | **75th Percentile :** 0.00 |
| Column Memory :  0.77 MB | **As % of Dtype Memory :** 8.33% | **As % of DF Memory :** 1.30% |



|  |  |  |
| --- | --- | --- |
| Tax Liens | | |
| TBD Column : no value | **Data Type :** float64 | **Values Count :** 100,514 |
| Unique Values Count :  12 (0.01%) | **Non-Null Values Count :** 99,990 (99.48%) | **Null Values Count :** 524 (0.52%) |
| Min :  0.00 | **Mean :** 0.03 | **Max :** 15.00 |
| 25th Percentile :  0.00 | **50th Percentile :** 0.00 | **75th Percentile :** 0.00 |
| Column Memory :  0.77 MB | **As % of Dtype Memory :** 8.33% | **As % of DF Memory :** 1.30% |



Segun Olivera (2005) los problemas de calidad de datos en bases de datos relacionales se puede tratar cuatro niveles de gradualidad: varias fuentes de datos, varias relaciones (tablas), una simple relacion, y nivel de atributo/tupla.

1. Problemas anivel de atributo / tuplas

Este tipo de problema se pueden dar en los siguientes elementos i) un atributo de una tupla, ii) un atributo en muchas tuplas (a nivel de columna) iii) varios atributos a nivel de una tupla (a nivel de fila)

* 1. Un atributo en una tupla
     1. Valores perdidos, por ejemplo, cuando falta un valor en un atributo en cuya definicion se establecio como no nulo. Este tipo de problema tiene que ver con la dimension de la completitud
     2. Violacion de sintaxis, cuando por ejemplo se establece una expresion regular para el formato de un atributo de correo electronico de la forma [usuario@subdominio.dominio](mailto:usuario@subdominio.dominio) y se almacena “ismael@uclm”. Este tipo de problema tiene que ver con la dimension de la precision sintactica
     3. Valores incorrectos, por ejemplo, cuando la base de datos se tiene un atributo nombre autor, y se sabe que es “Mario Piattini” aparece “Mariano Piattini”. Este tipo de problema tiene que ver con la dimension de la presencia semantica.

Imagen que contiene captura de pantalla, texto

Descripción generada automáticamente

Imagen que contiene captura de pantalla, texto

Descripción generada automáticamente

Imagen que contiene captura de pantalla

Descripción generada automáticamente

Imagen que contiene captura de pantalla

Descripción generada automáticamente

Imagen que contiene texto, periódico, captura de pantalla

Descripción generada automáticamente

Imagen que contiene texto, captura de pantalla, interior

Descripción generada automáticamente

Imagen que contiene captura de pantalla, texto, periódico

Descripción generada automáticamente

Imagen que contiene texto, periódico

Descripción generada automáticamente

Imagen que contiene texto

Descripción generada automáticamente

