***DQR Report***

1. **High-level description:**

The name of the dataset is ‘NY property data’, it comes from ‘NYC Open Data’. The dataset describes the information about Real Estate Assessment Property data. It covers the record in Nov 2010. There are 32 fields and 1070994 records in the dataset.

1. **Two field summary tables:**

*Numerical Field Table:*

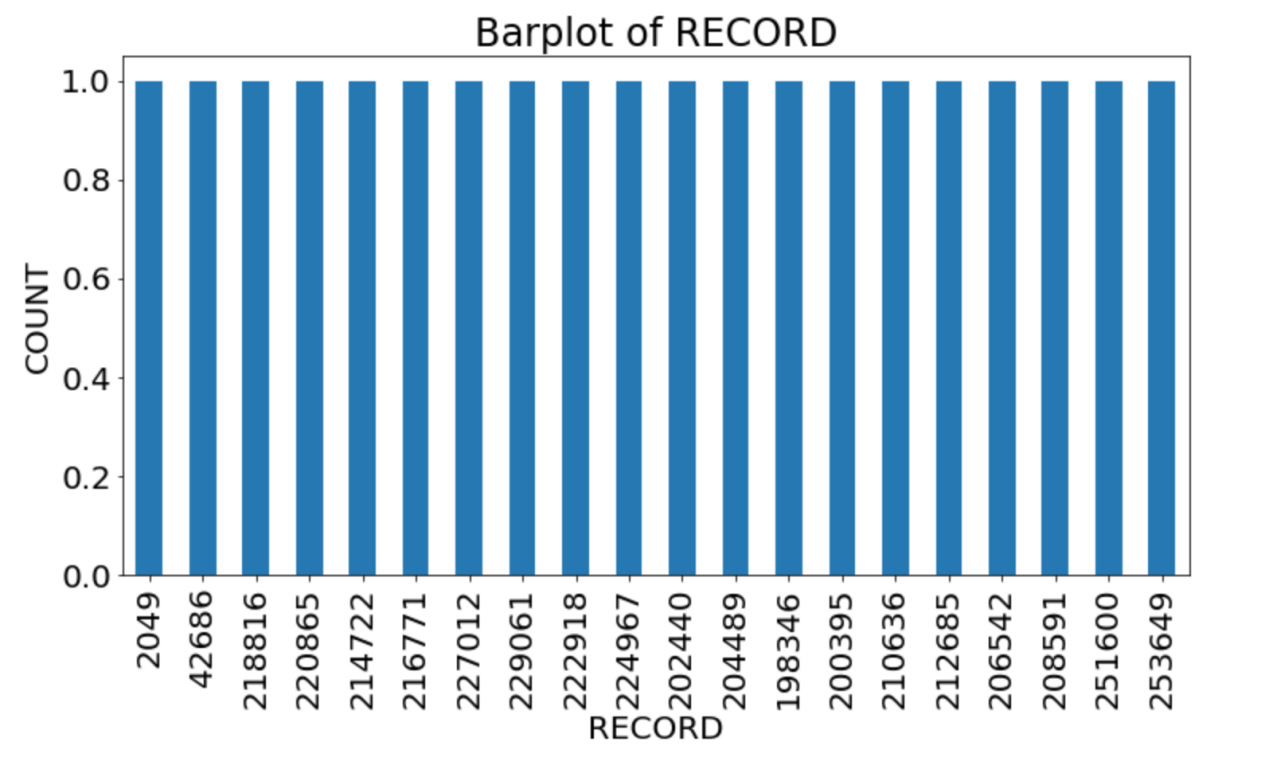
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Field Name** | **%**  **Populated** | **Min** | **Max** | **Mean** | **Stdev** | **%Zero** |
| LTFRONT | 100 | 0 | 9999 | 36.64 | 74.03 | 15.79 |
| LTDEPTH | 100 | 0 | 9999 | 88.86 | 76.4 | 15.89 |
| STORIES | 94.75 | 1 | 119 | 5.01 | 8.37 | 0 |
| FULLVAL | 100 | 0 | 6,150,000,000 | 874,264.51 | 11,582,430.99 | 1.21 |
| AVLAND | 100 | 0 | 2,668,500,000 | 85,067.92 | 4,057,260.06 | 1.21 |
| AVTOT | 100 | 0 | 4,668,309,000 | 227,238.17 | 6,877,529.31 | 1.21 |
| EXLAND | 100 | 0 | 2,668,500,000 | 36,423.89 | 3,981,575.79 | 45.91 |
| EXTOT | 100 | 0 | 4,668,309,000 | 91,186.98 | 6,508,402.82 | 40.39 |
| BLDFRONT | 100 | 0 | 7575 | 23.04 | 35.58 | 21.36 |
| BLDDEPTH | 100 | 0 | 9393 | 39.92 | 42.71 | 21.37 |
| AVLAND2 | 26.4 | 3 | 2,371,005,000 | 246,235.72 | 6,178,962.56 | 0 |
| AVTOT2 | 26.4 | 3 | 4,501,180,000 | 713,911.44 | 11,652,528.95 | 0 |
| EXLAND2 | 8.17 | 1 | 2,371,005,000 | 351,235.68 | 10,802,212.67 | 0 |
| EXTOT2 | 12.22 | 7 | 4,501,180,000 | 656,768.28 | 16,072,510.17 | 0 |
| YEAR | 100 | 2010/11 | 2010/11 | N/A | N/A | 0 |

*Categorical Field Table:*

|  |  |  |  |
| --- | --- | --- | --- |
| **Field Name** | **%**  **Populated** | **#Unique Values** | ***Most Common Value*** |
| RECORD | 100 | 1070994 | N/A |
| BBLE | 100 | 1070994 | N/A |
| BORO | 100 | 5 | 4(Queens) |
| BLOCK | 100 | 13984 | 3944 |
| LOT | 100 | 6366 | 1 |
| EASEMENT | 0.43 | 12 | E(Land Easement) |
| OWNER | 97.04 | 863347 | PARKCHESTER PRESERVAT |
| BLDGCL | 100 | 200 | R4 |
| TAXCLASS | 100 | 11 | 1(1-3 Unit Residence) |
| EXT | 33.08 | 3 | G |
| EXCD1 | 59.62 | 129 | 1017 |
| STADDR | 99.94 | 839280 | 501 SURF AVENUE |
| ZIP | 97.21 | 196 | 10314 |
| EXMPTCL | 1.45 | 14 | X1 |
| EXCD2 | 8.68 | 60 | 1017 |
| PERIOD | 100 | 1 | FINAL |
| VALTYPE | 100 | 1 | AC-TR |

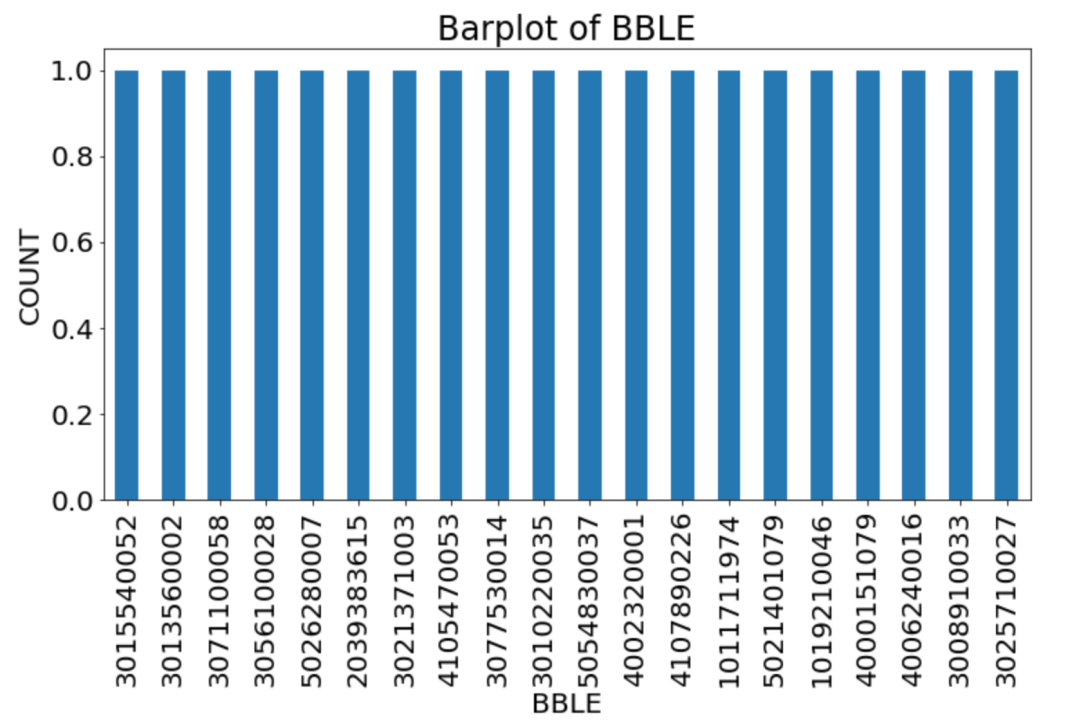
1. **Distribution for each field**
2. **RECORD**

The graph below is the first 20 most repeated observations in the ‘RECORD’ field. However, we can observe that they all have a count of 1. It is because field ‘RECORD’ has unique values for each rows (represent the record of each observations).



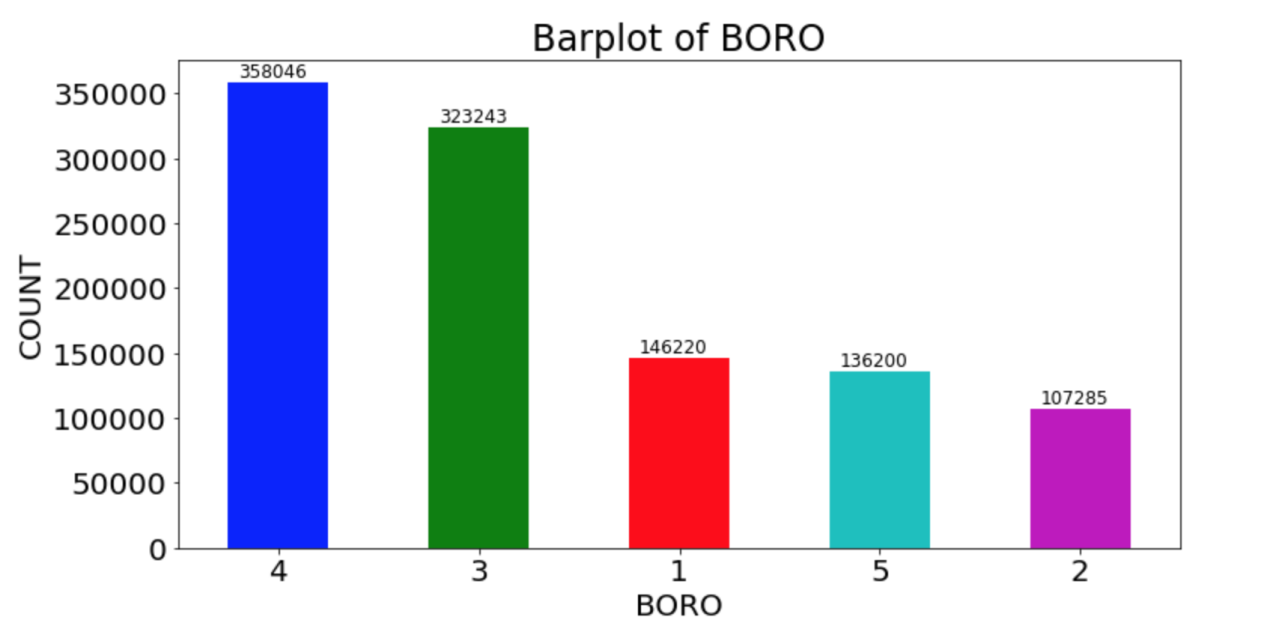
1. **BBLE**

The graph below is the first 20 most repeated observations in the ‘BBLE’ field. However, we can observe that they all have a count of 1 on the y-axis. It is because field ‘BBLE’ has unique values for each rows (represent the File Key).



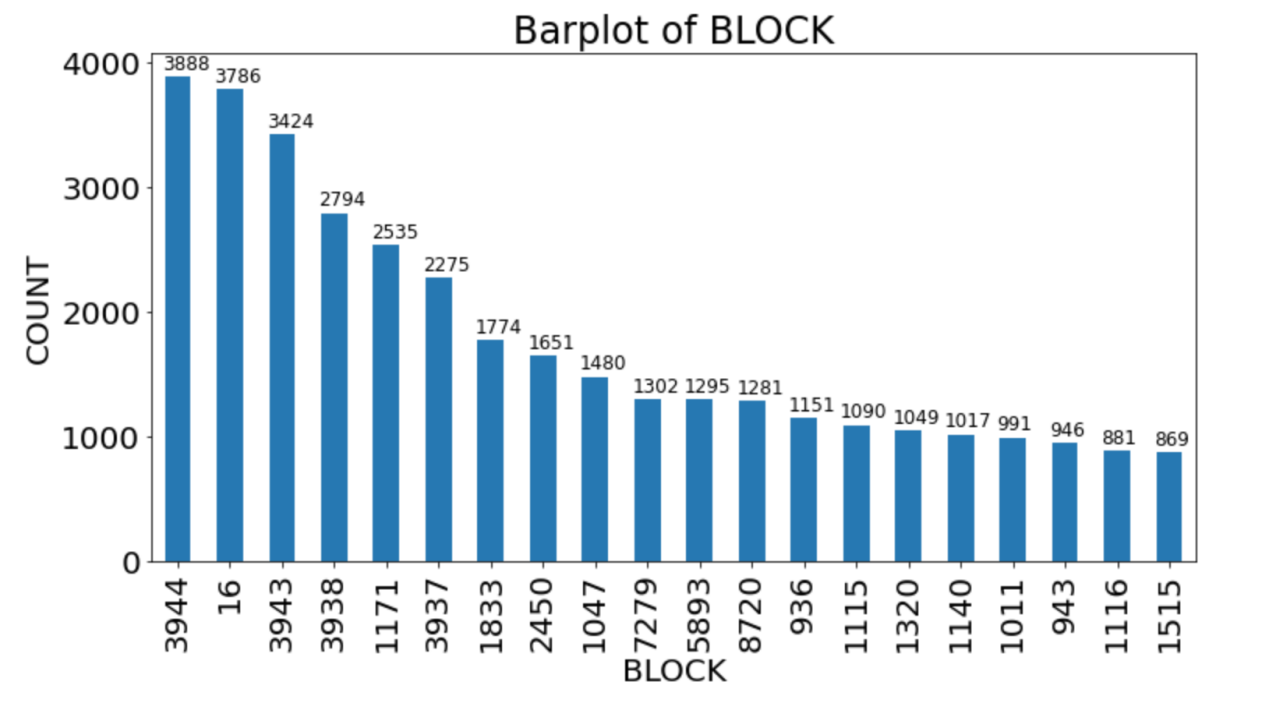
1. **BORO**

The graph below is the top 5 most repeated observations in the ‘RECORD’ field. We can observed that 4 which represent Queens has the highest count.



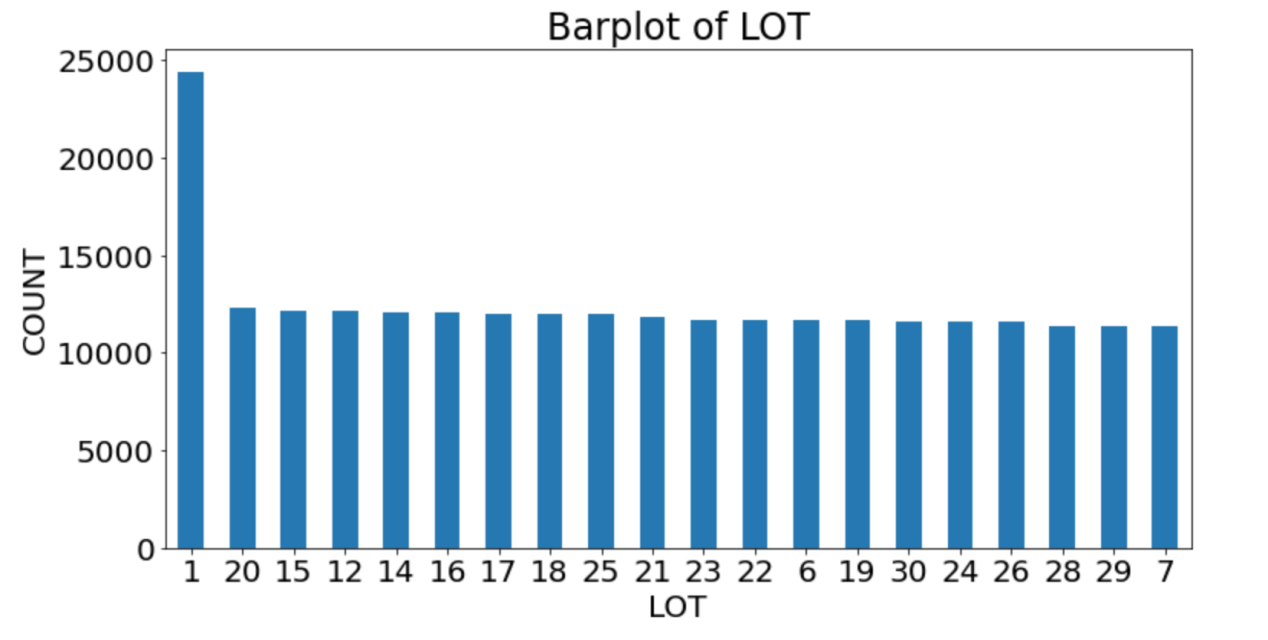
**4. BLOCK**

The graph below is the top 20 most repeated observations in the ‘BLOCK’ field.



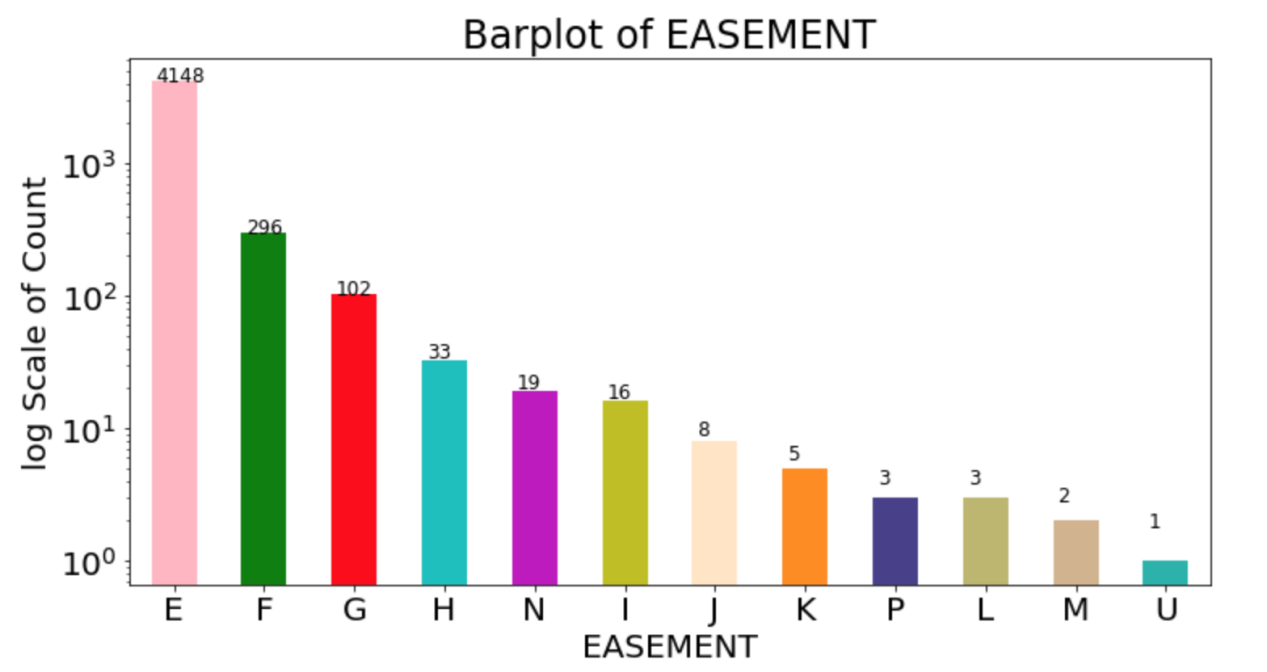
1. **LOT**

The graph below is the top 20 most repeated observations in the ‘LOT’ field.



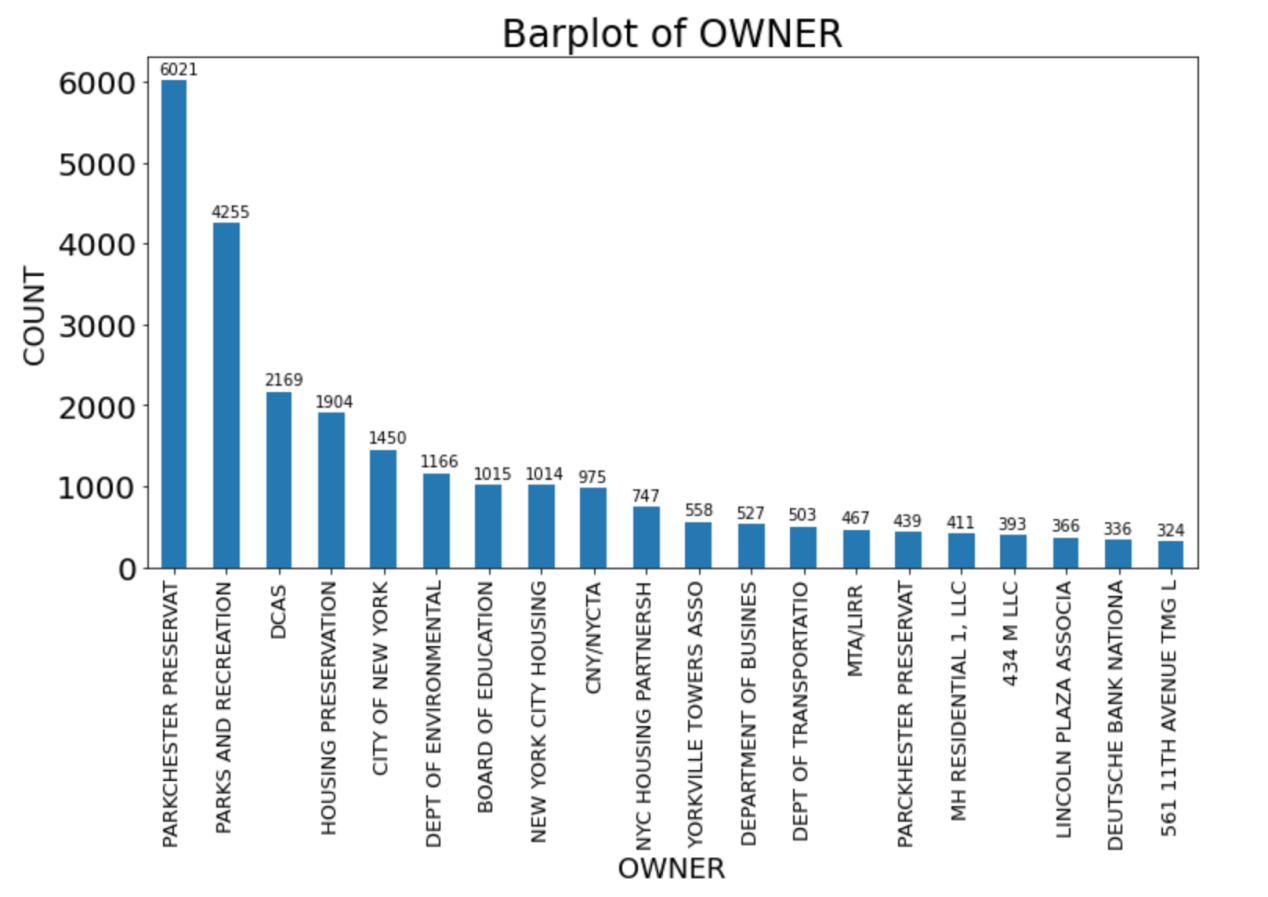
1. **EASEMENT**

The graph below is the top 12 most repeated observations in the ‘EASEMENT’ field with log scaled y-axis for better visualizations.



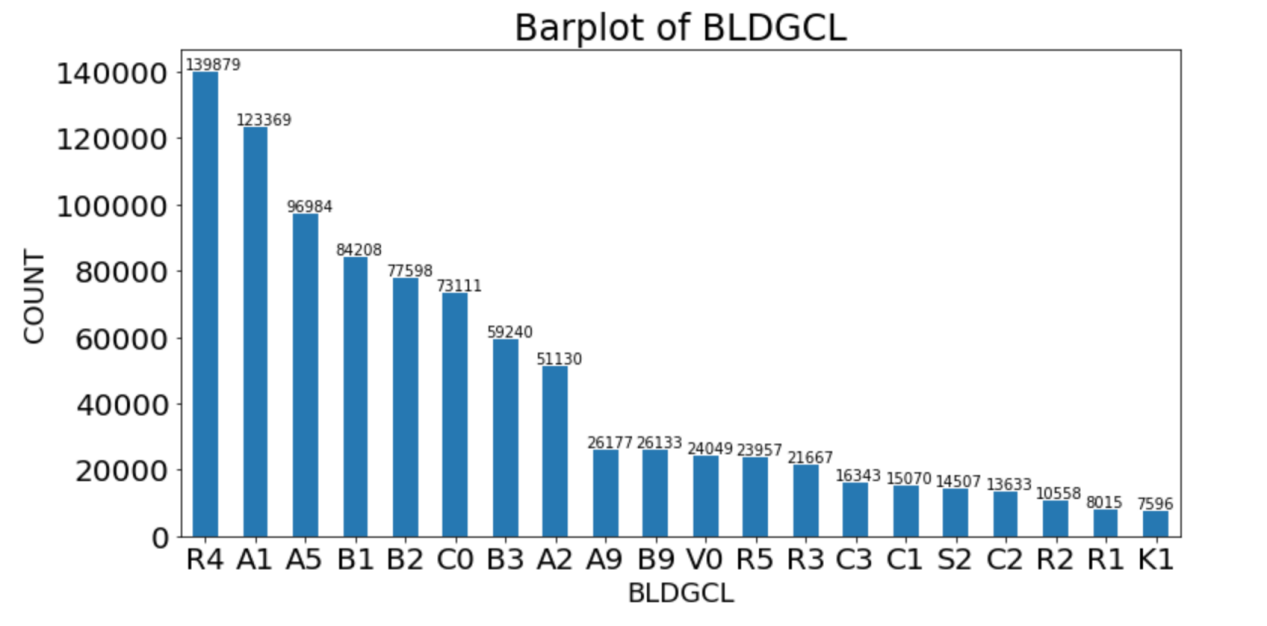
1. **OWNER**

The graph below is the top 20 most repeated observations in the ‘OWNER’ field.



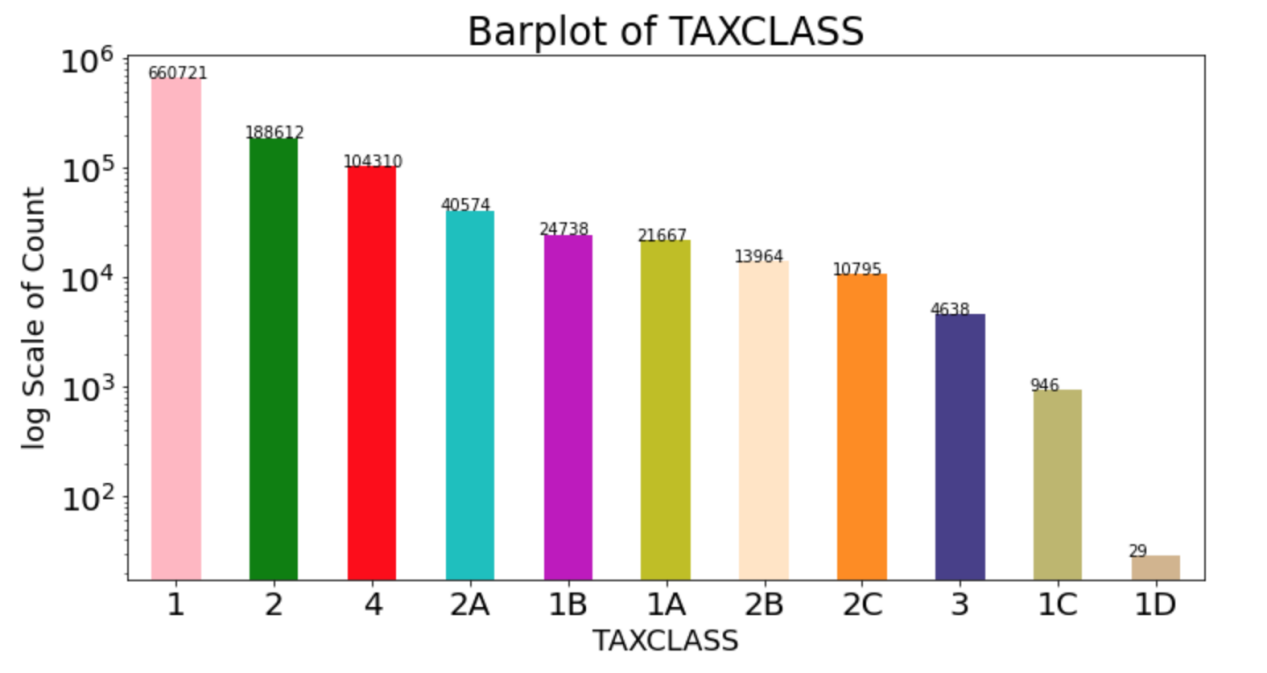
1. **BLDGCL**

The graph below is the top 20 most repeated observations in the ‘BLDGCL’ field.



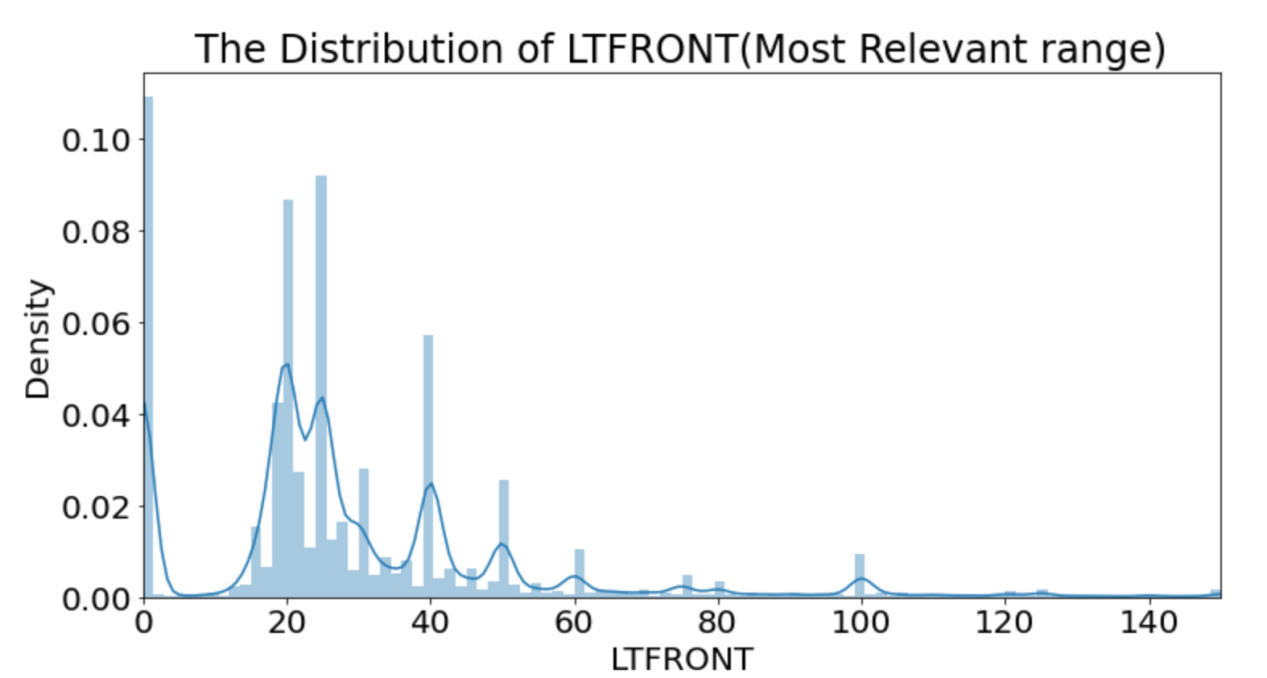
1. **TAXCLASS**

The graph below is the top 11 most repeated observations in the ‘TAXCLASS’ field.



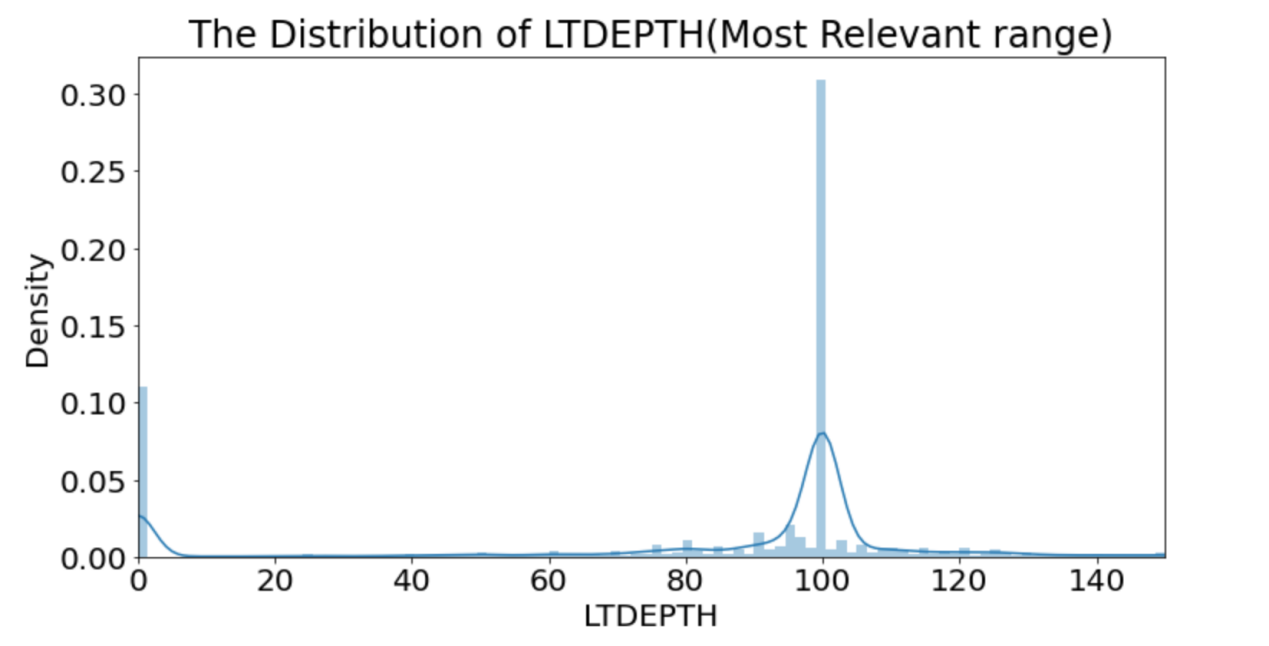
1. **LTFRONT**

After the observation of the most relevant range for LTFRONT, I choose the 0-150 for the x range. The graph below is the distribution of the most relevant LTFRONT observations.



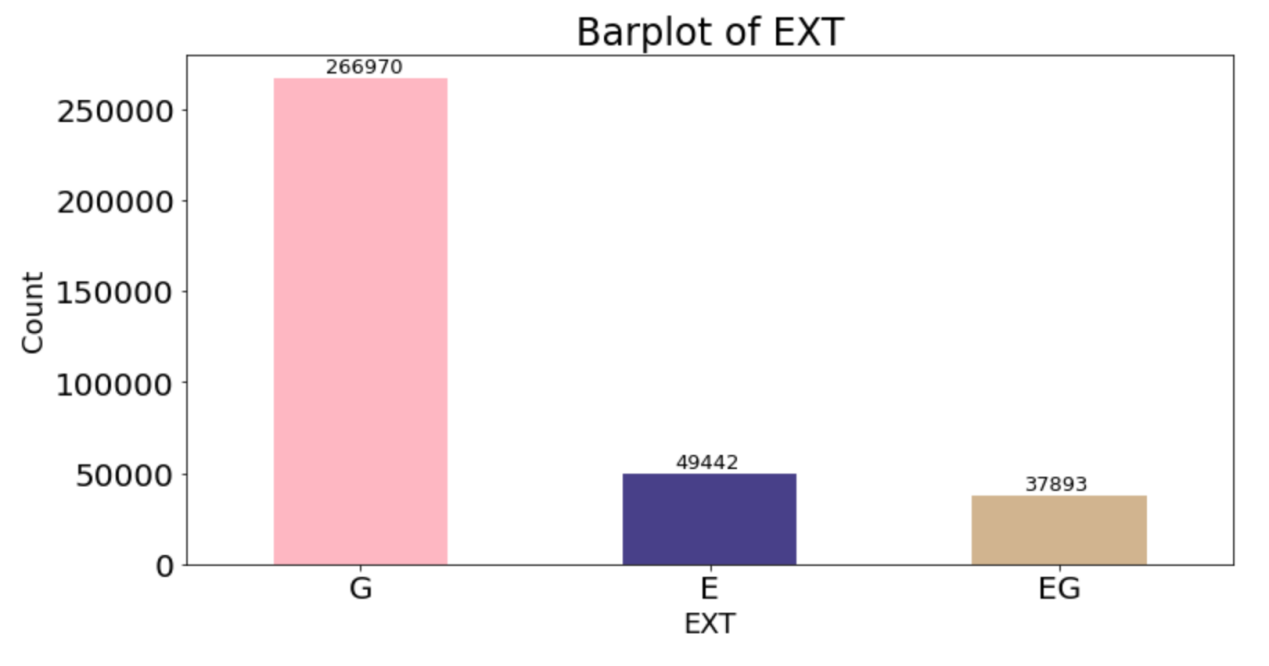
1. **LTDEPTH**

After the observation of the most relevant range for LTDEPTH, I choose the 0-150 for the x range. The graph below is the distribution of the most relevant LTDEPTH observations.



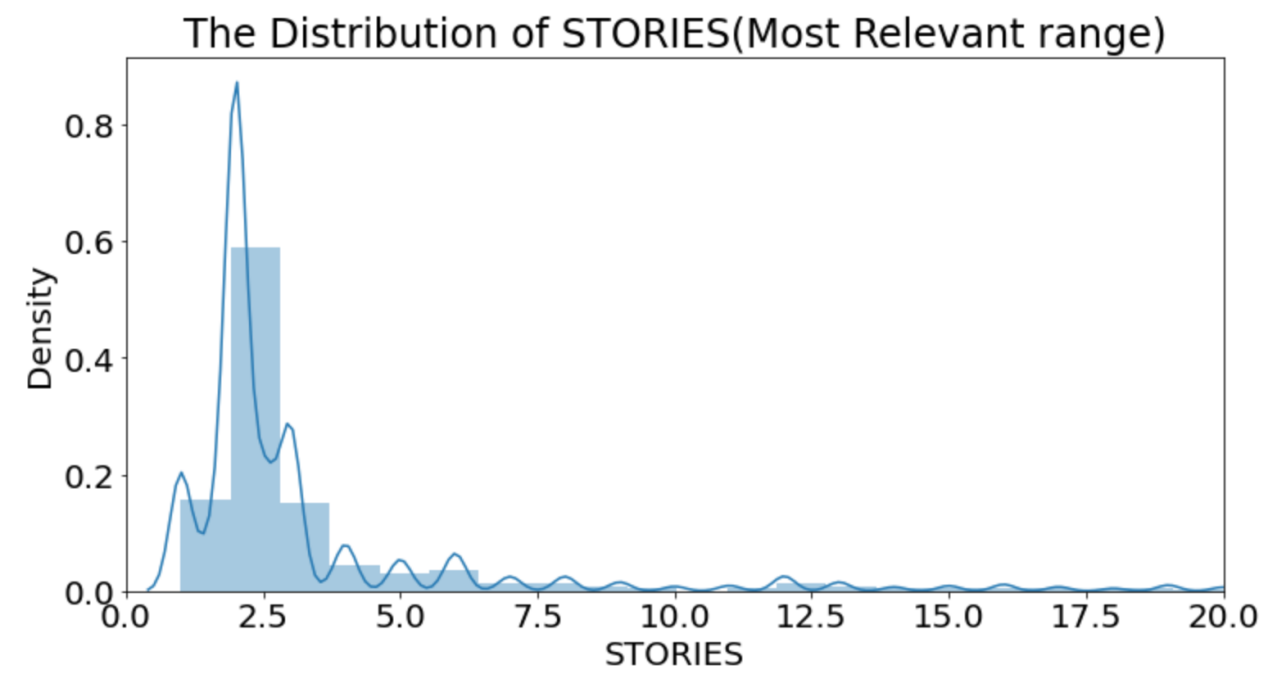
1. **EXT**

The graph below is the top 3 most repeated observations in the ‘EXT’ field.



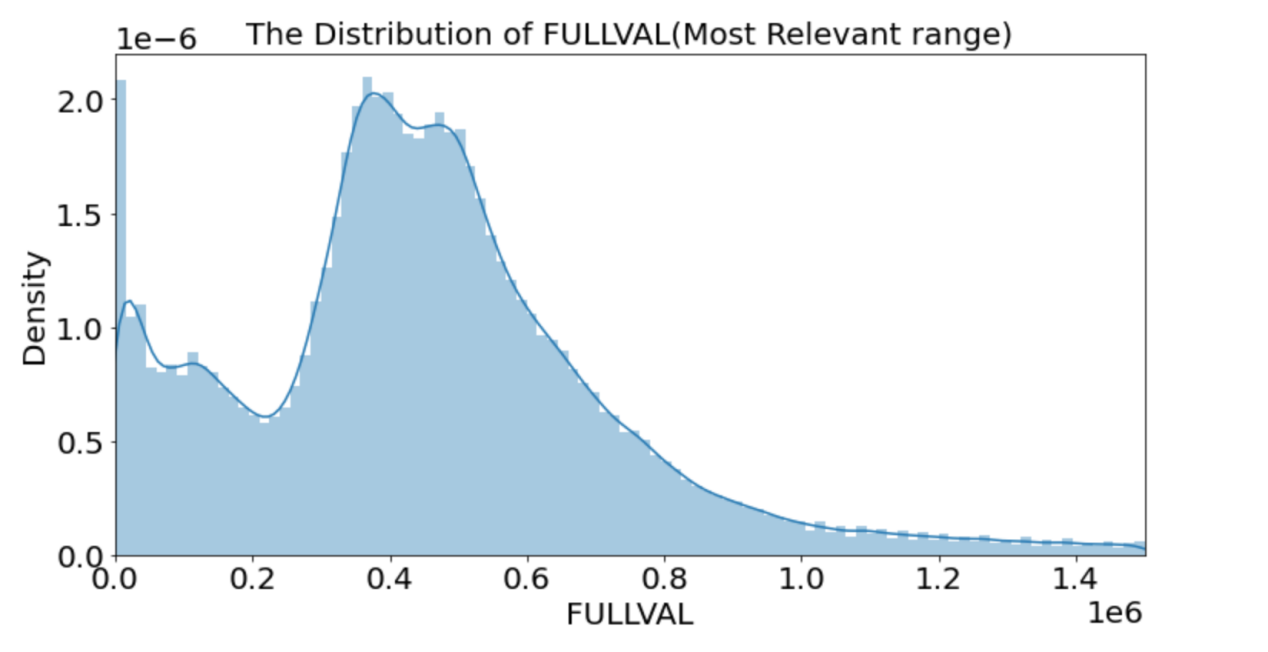
1. **STORIES**

After the observation of the most relevant range for STORIES, I choose 0-20 for the x range. The graph below is the distribution of the most relevant LTFRONT observations.



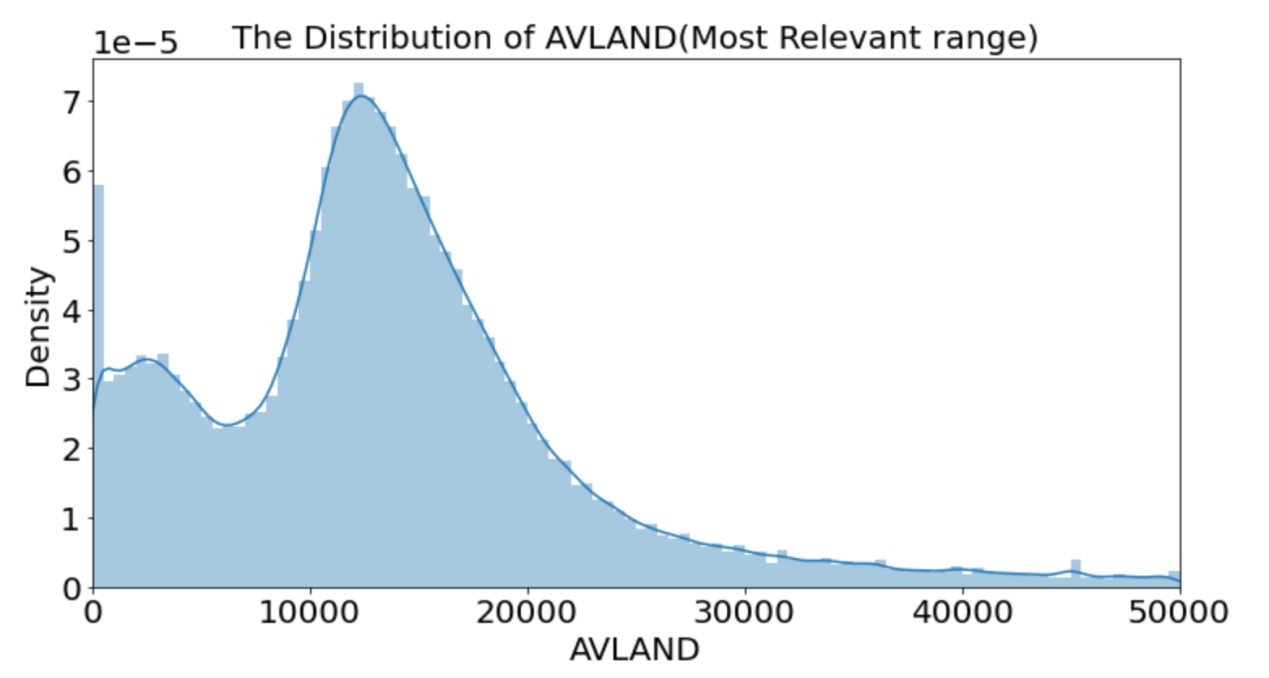
1. **FULLVAL**

After the observation of the most relevant range for FULLVAL and the consideration of 0 values, I choose 0-1500000 for the x range. The graph below is the distribution of the most relevant FULLVAL observations.



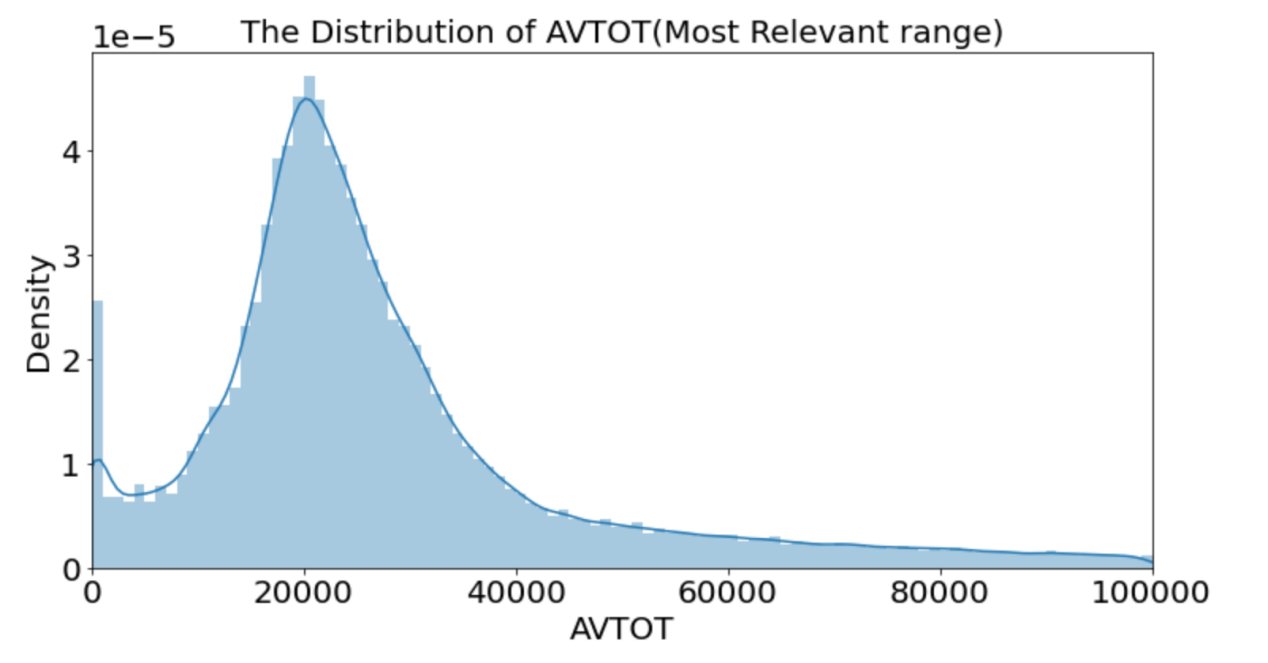
1. **AVLAND**

After the observation of the most relevant range for AVLAND and the consideration of 0 values, I choose 0-50000 for the x range. The graph below is the distribution of the most relevant AVLAND observations.



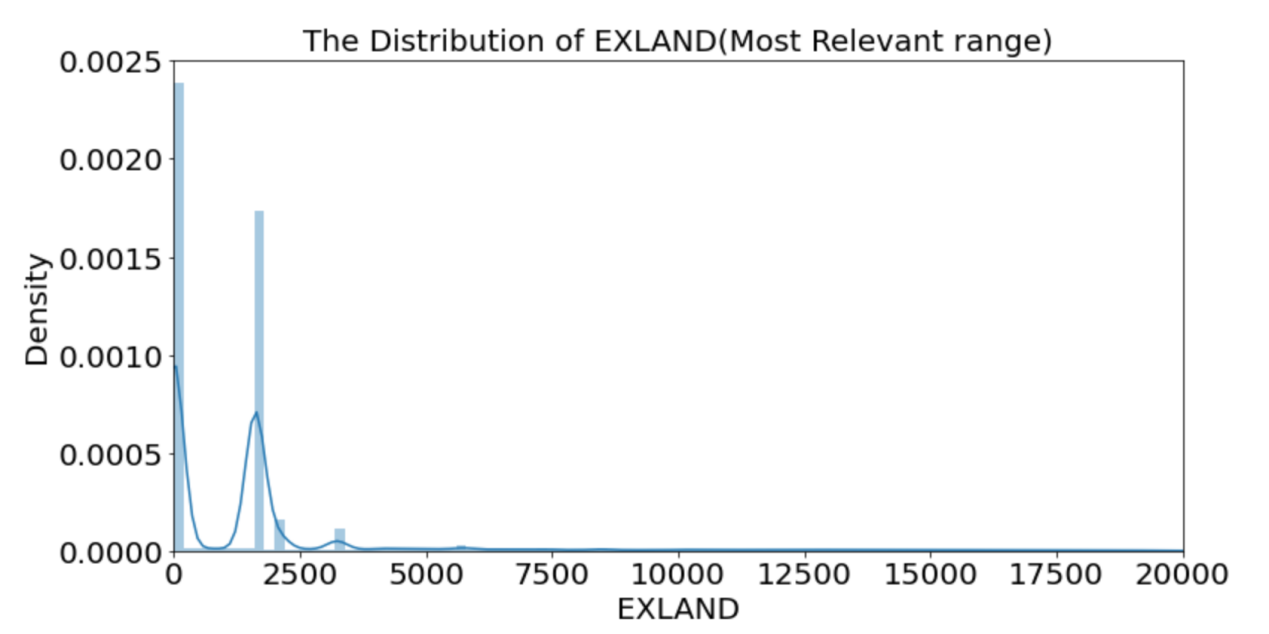
1. **AVTOT**

After the observation of the most relevant range for AVTOT and the consideration of 0 values, I choose 0-100000 for the x range. The graph below is the distribution of the most relevant AVTOT observations.



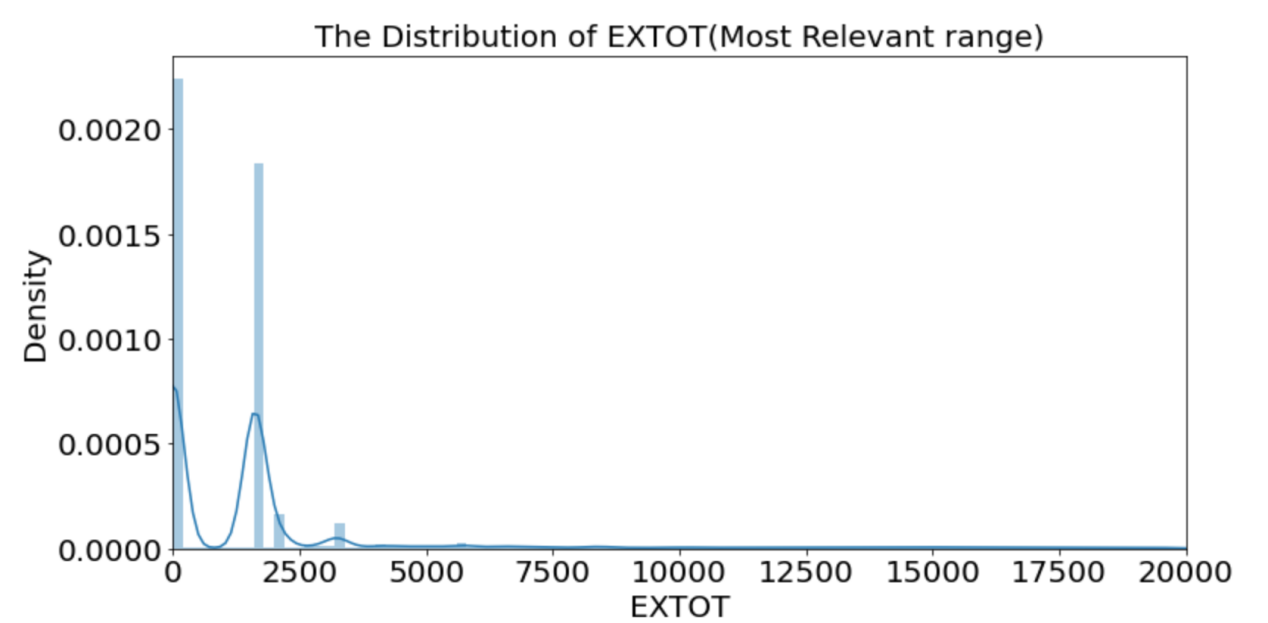
1. **EXLAND**

After the observation of the most relevant range for EXLAND and the consideration of 0 values, I choose 0-20000 for the x range. The graph below is the distribution of the most relevant EXLAND observations.



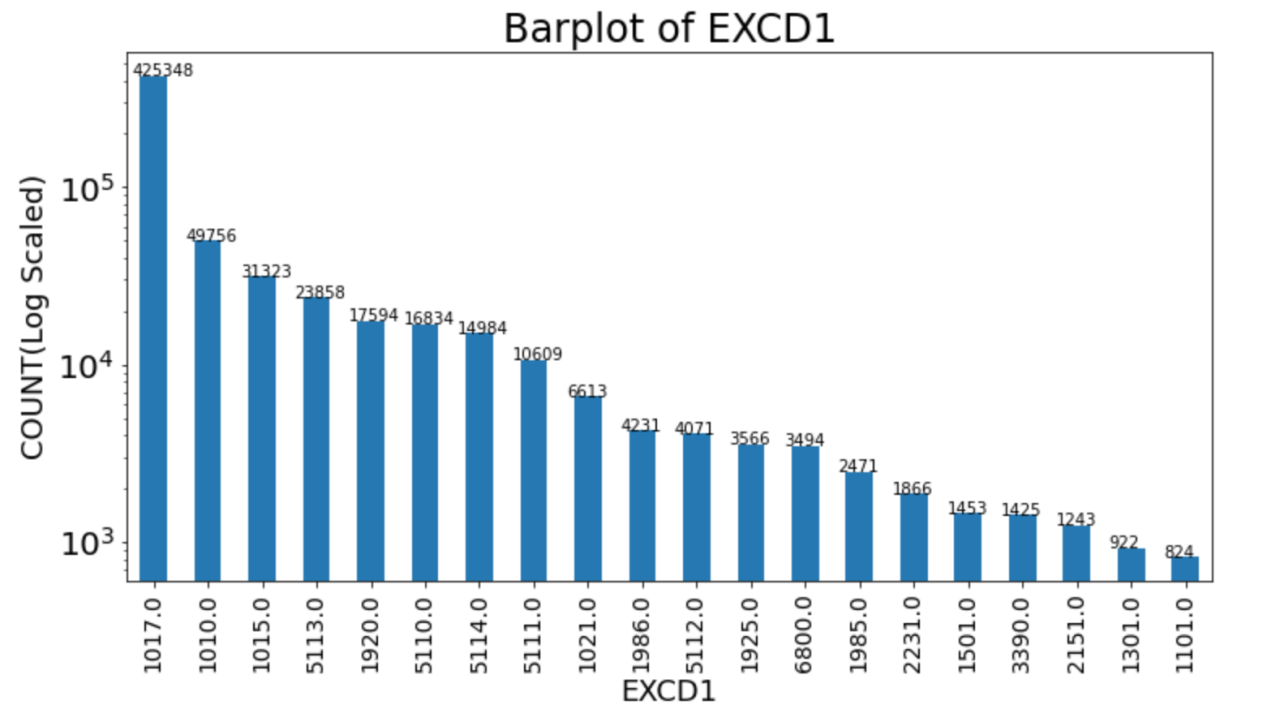
1. **EXTOT**

After the observation of the most relevant range for EXTOT and the consideration of 0 values, I choose 0-20000 for the x range. The graph below is the distribution of the most relevant EXLAND observations.



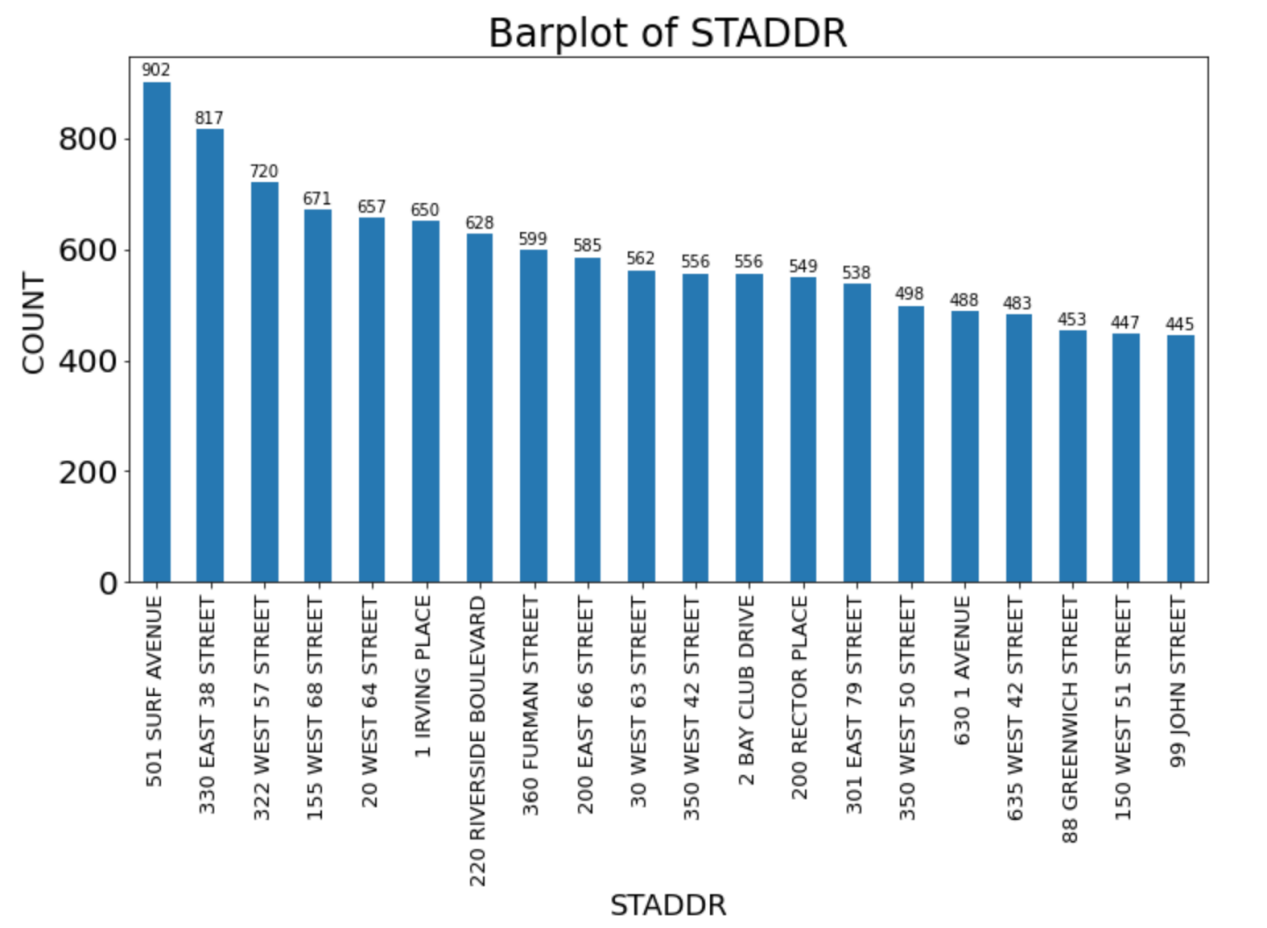
1. **EXCD1**

The graph below is the top 20 most repeated observations in the ‘EXCD1’ field, with y-axis log scaled.



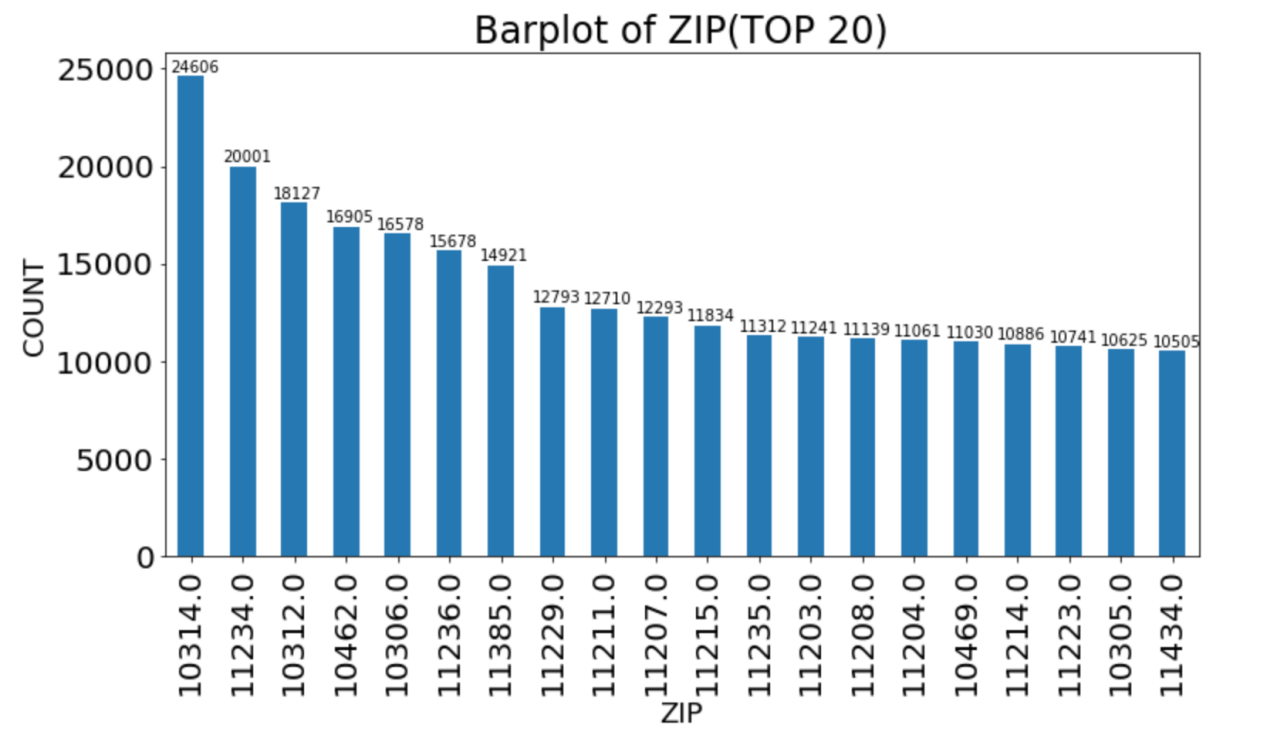
1. **STADDR**

The graph below is the top 20 most repeated observations in the ‘STADDR’ field.



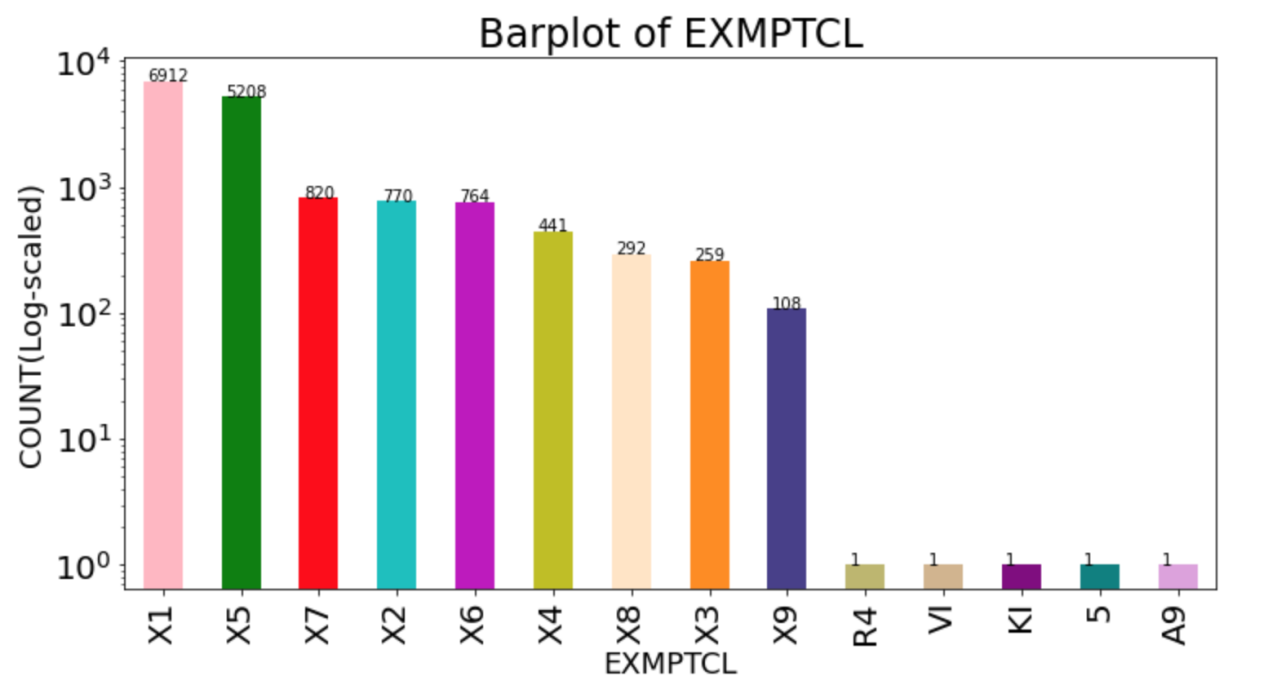
1. **ZIP**

The graph below is the top 20 most repeated observations in the ‘ZIP’ field.



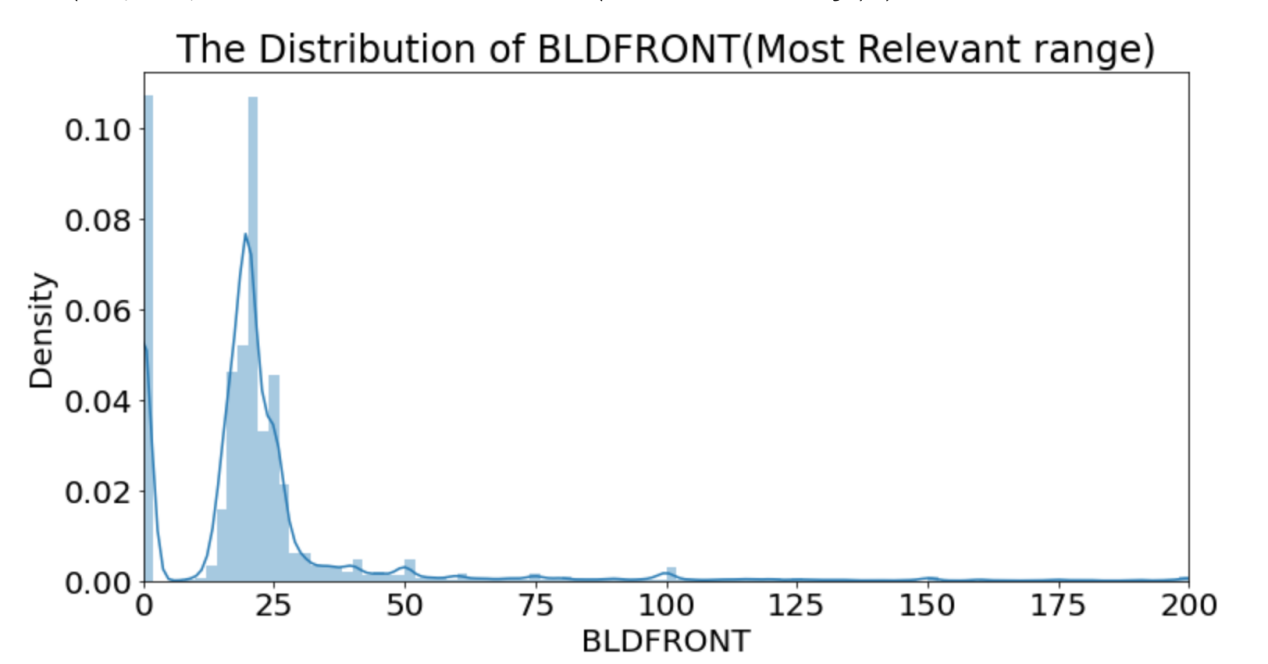
1. **EXMPTCL**

The graph below is the top 14 most repeated observations in the ‘EXMPTCL’ field with log scaled y-axis for better visualizations.



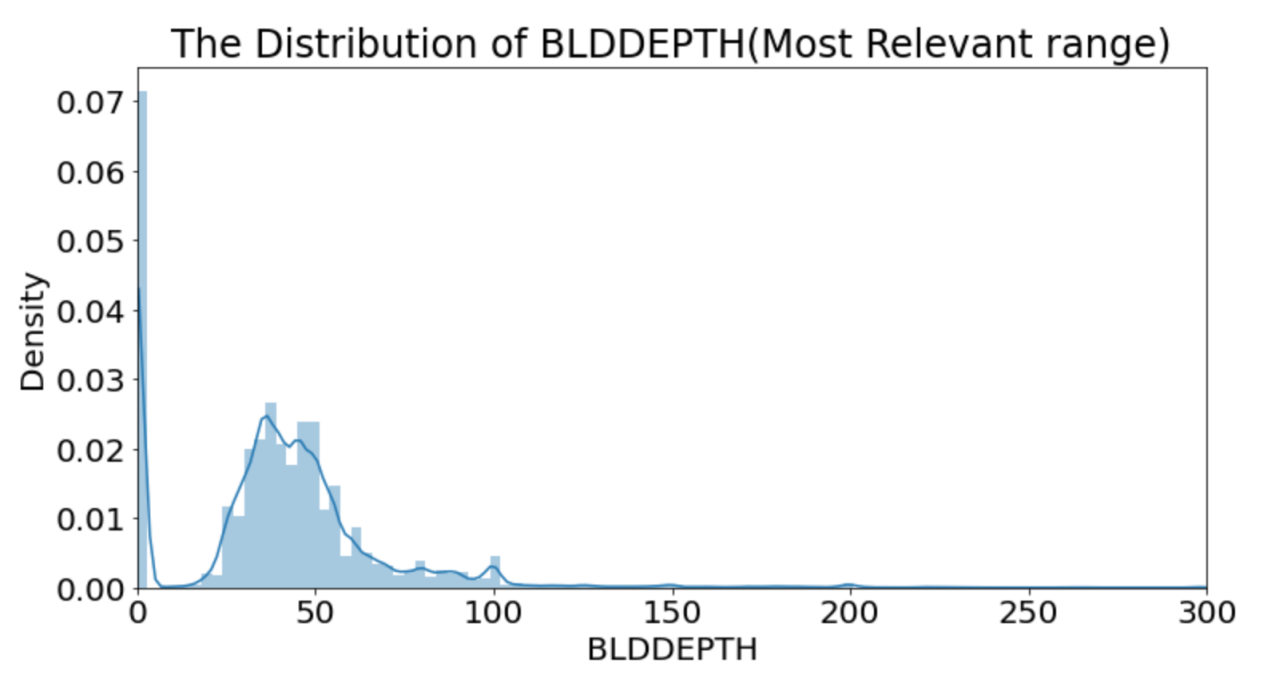
1. **BLDFRONT**

After the observation of the most relevant range for BLDFRONT, I choose 0-200 for the x range. The graph below is the distribution of the most relevant BLDFRONT observations.



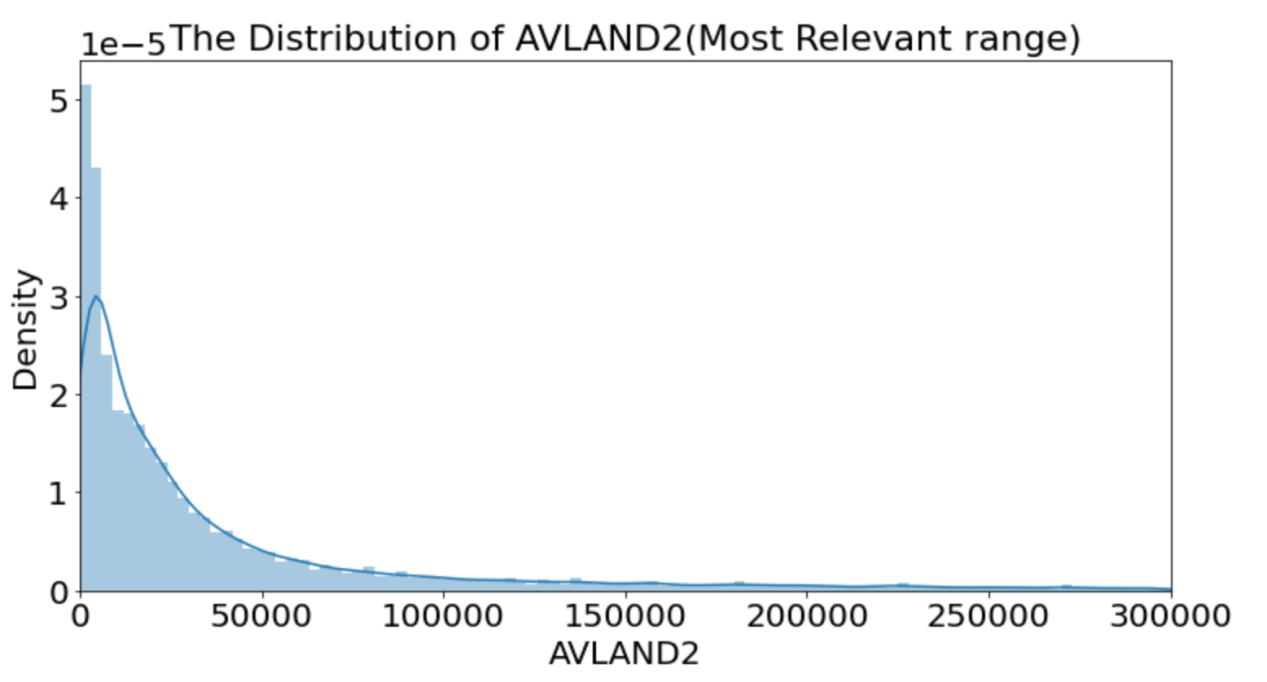
1. **BLDDEPTH**

After the observation of the most relevant range for BLDDEPTH, I choose 0-300 for the x range. The graph below is the distribution of the most relevant BLDDEPTH observations.



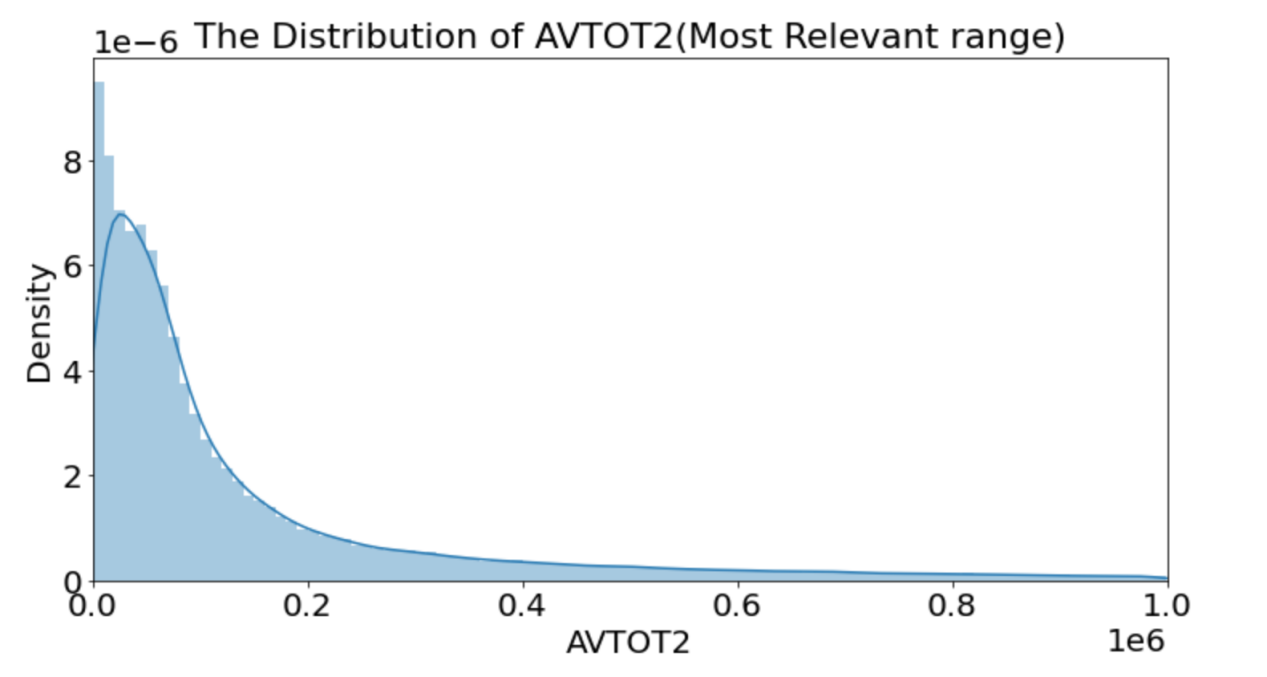
1. **AVLAND2**

After the observation of the most relevant range for AVLAND2, I choose 0-300000 for the x range. The graph below is the distribution of the most relevant AVLAND2 observations.



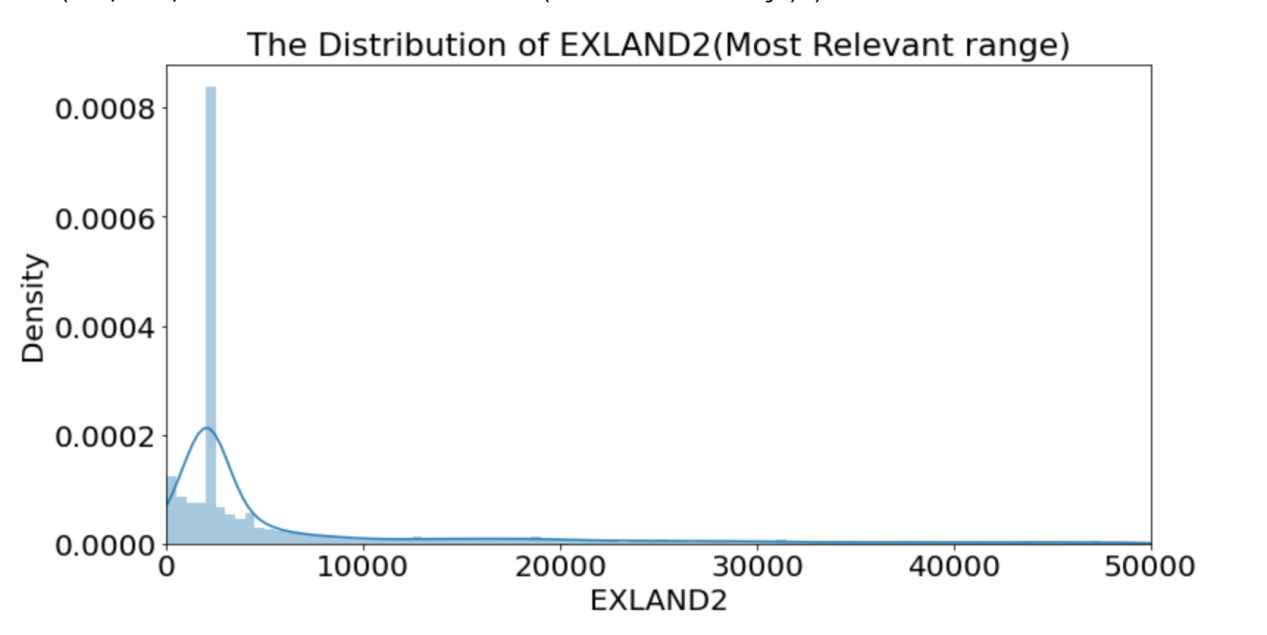
1. **AVTOT2**

After the observation of the most relevant range for AVTOT2, I choose 0-1000000 for the x range. The graph below is the distribution of the most relevant AVTOT2 observations.



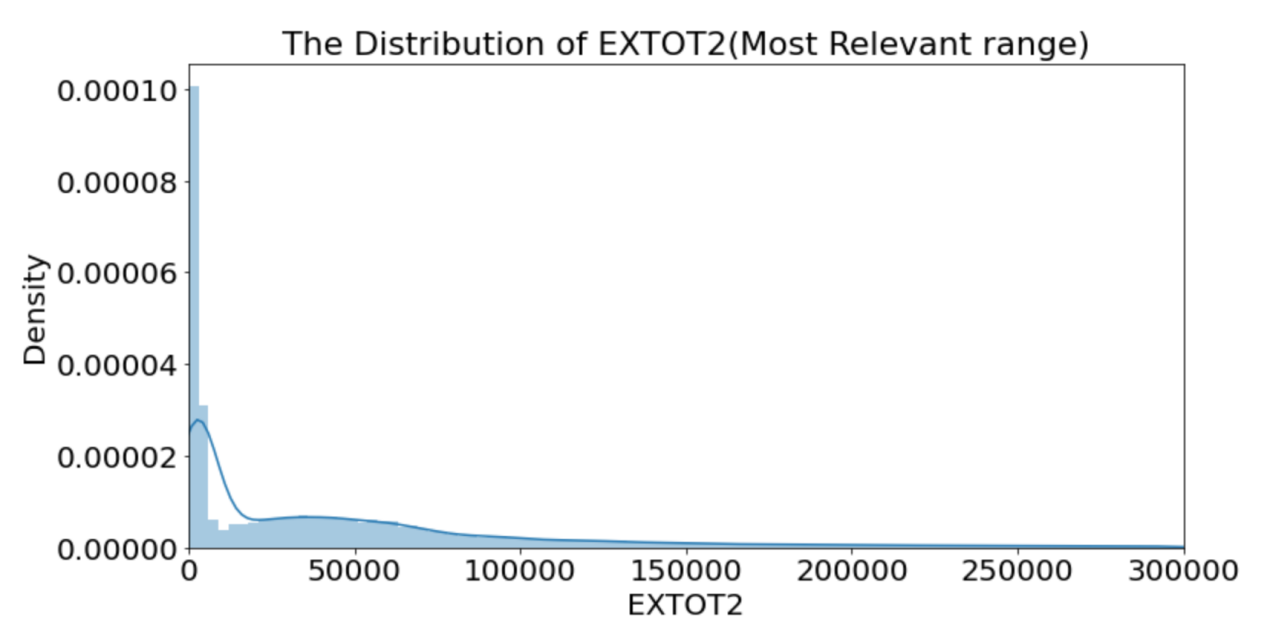
1. **EXLAND2**

After the observation of the most relevant range for EXLAND2, I choose 0-50000 for the x range. The graph below is the distribution of the most relevant EXLAND2 observations.



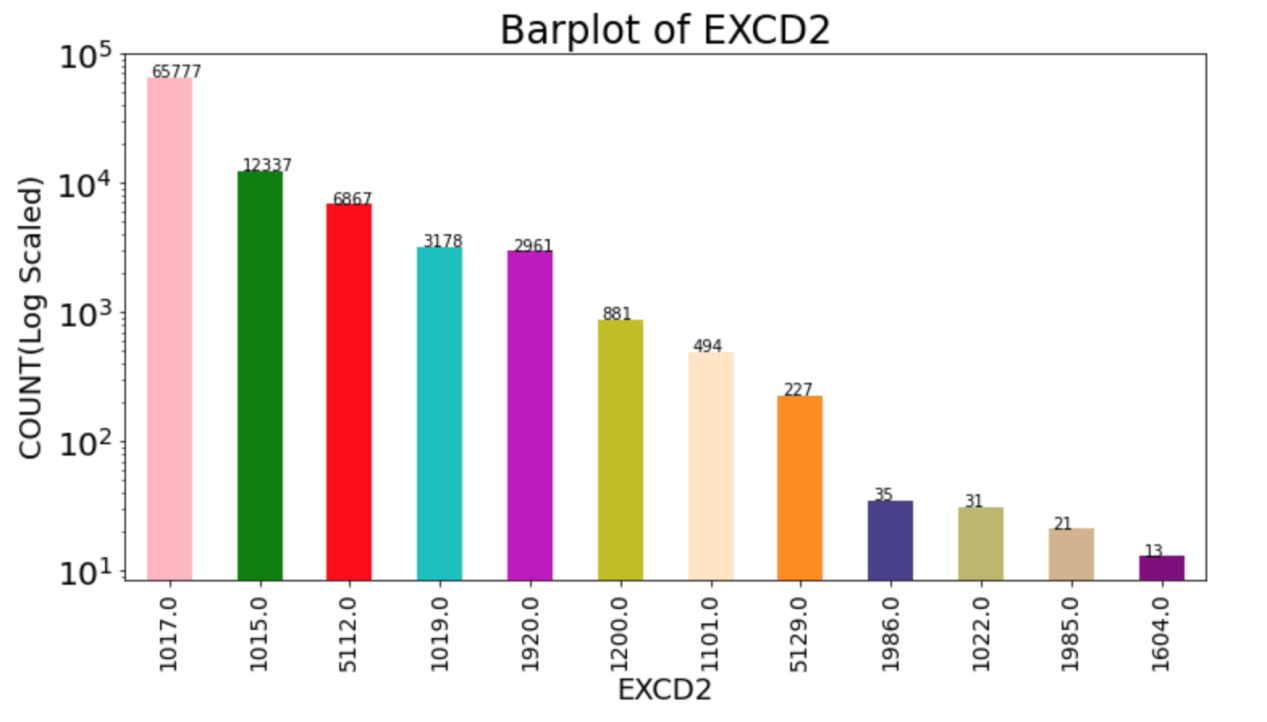
1. **EXTOT2**

After the observation of the most relevant range for EXTOT2, I choose 0-30000 for the x range. The graph below is the distribution of the most relevant EXTOT2 observations.



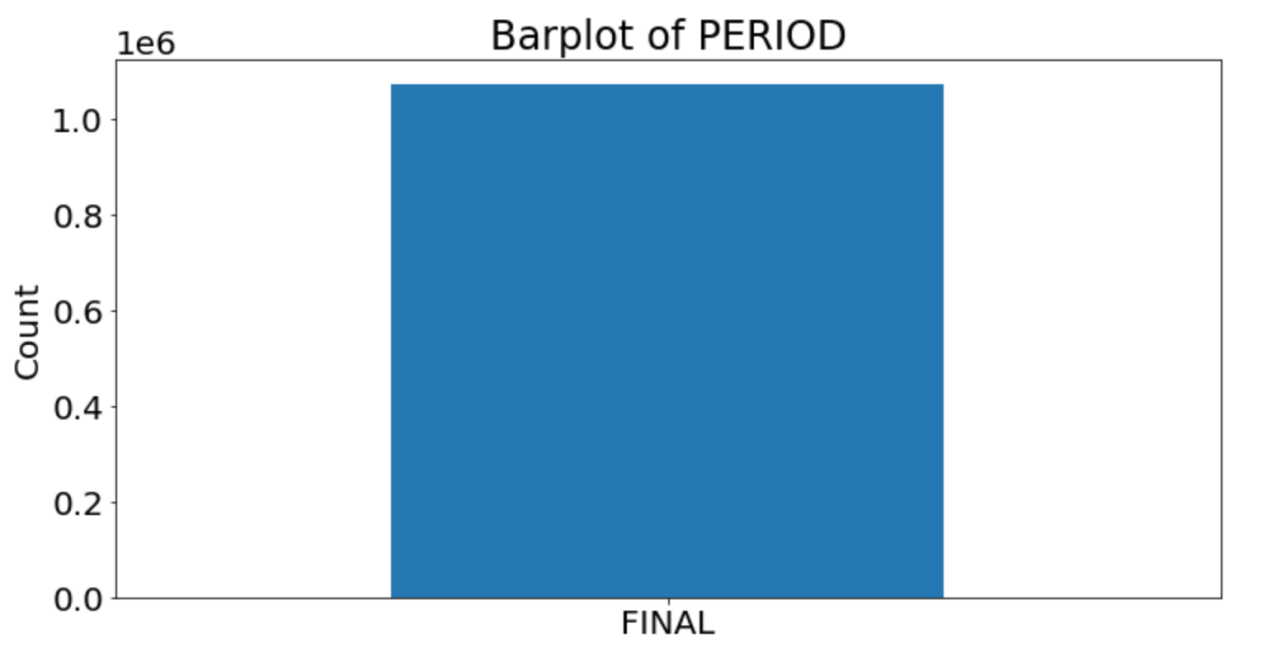
1. **EXCD2**

The graph below is the top 12 most repeated observations in the ‘EXCD1’ field, with y-axis log scaled.



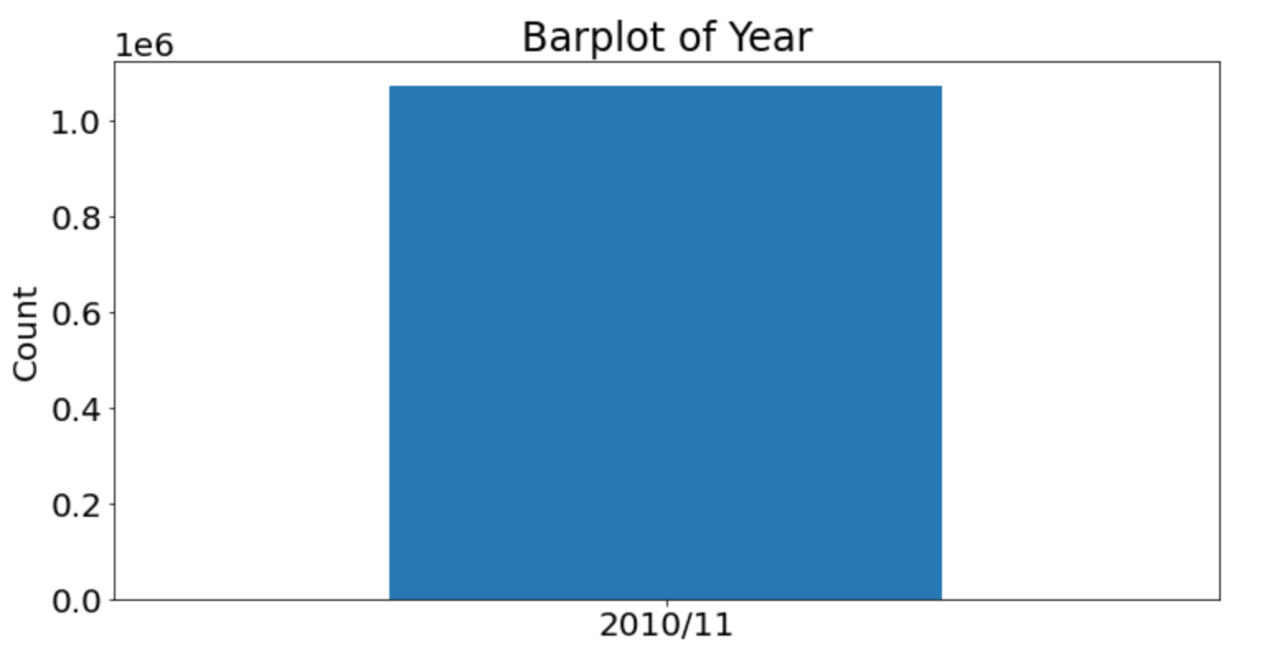
1. **PERIOD**

The graph below is the bar plot of PERIOD. There is only one bar in the graph below since PERIOD has only one value “FINAL”.



1. **YEAR**

The graph below is the bar plot of YEAR. There is only one bar in the graph below since YEAR has only one value “2010/11”.



1. **VALTYPE**

The graph below is the bar plot of VALTYPE. There is only one bar in the graph below since VALTYPE has only one value “AC-TR”.

