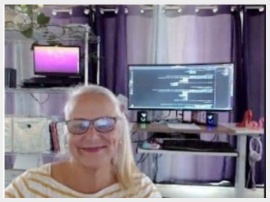
Washington State Cannabis Datasets Nov 2021

- Dirty Data Investigation:
  - Count and Plot the number of NaNs

    “CountAndPlot#NaNs.py”

Data Detective:  Candace O’Sullivan(Sutherland)

https://github.com/candy-o/Cannabis_Data_Science_Research

WashingtonStateCannabisDatasetsNov2021#NaNs.pdf

Links to the Datasets

Data sources:

- WA State Traceability Data January 2018 - November 2021 Link(s)

  https://lcb.app.box.com/s/e89t59s0yb558tjoncjsid710oirqbgd?page=1

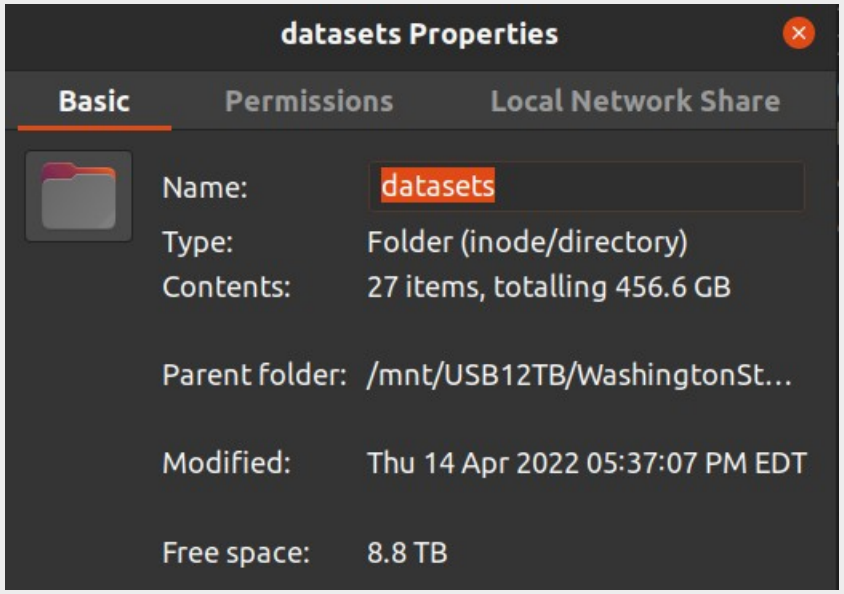
  https://lcb.app.box.com/s/e89t59s0yb558tjoncjsid710oirqbgd?page=2

Data Guide:

- Washington State Leaf Data Systems Guide Link

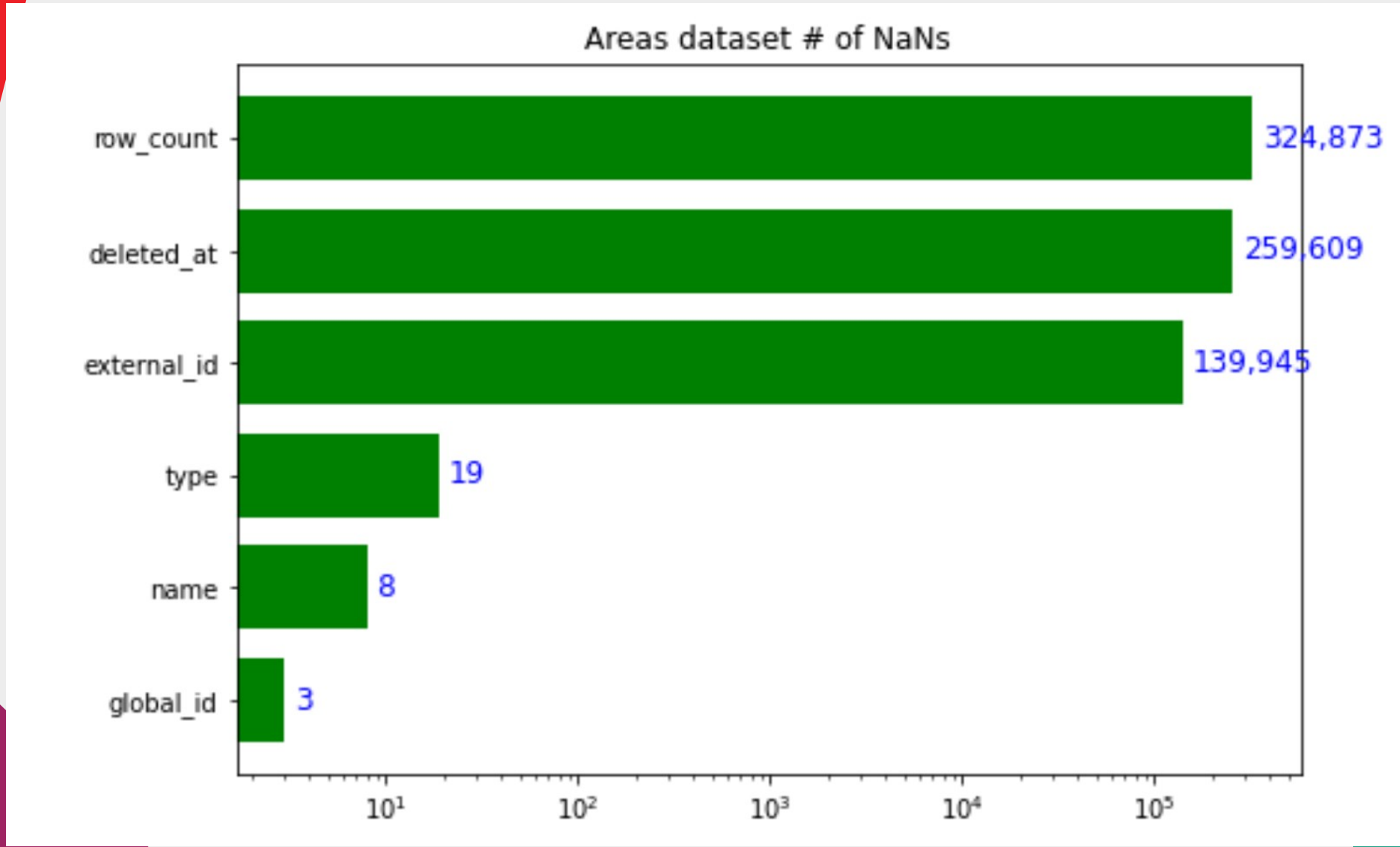
  https://lcb.wa.gov/sites/default/files/publications/Marijuana/traceability/WALeafDataSystems_UserManual_v1.37.5_AddendumC_LicenseeUser.pdf

Datasets unzipped: 456.6GB

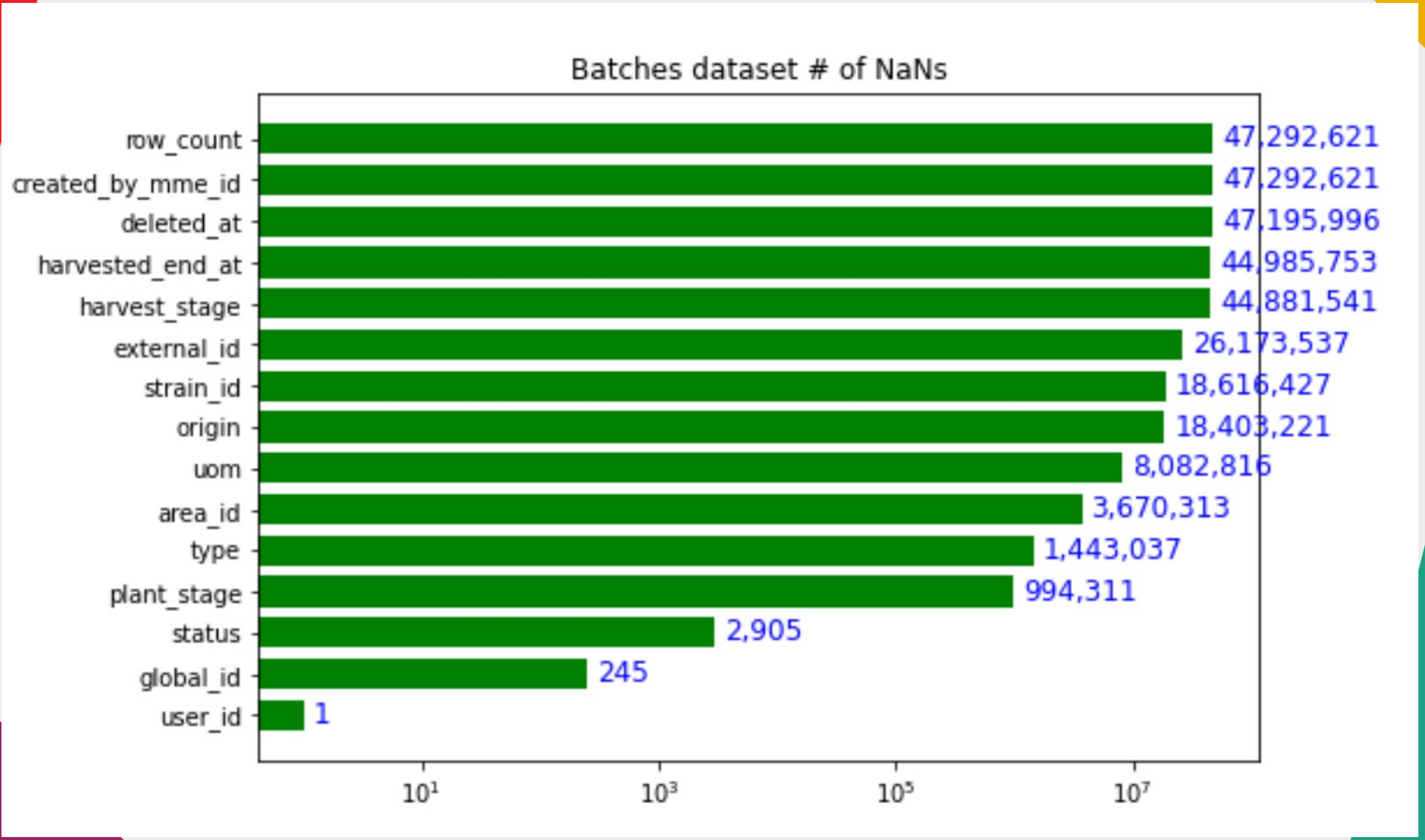


Areas_0.csv	83.4 MB	15 Nov 2021
Batches_0.csv	27.7 GB	15 Nov 2021
Disposals_0.csv	10.2 GB	15 Nov 2021
Inventories_0.csv	32.5 GB	15 Nov 2021
InventoryAdjustments_0.csv	24.9 GB	15 Nov 2021
InventoryAdjustments_1.csv	24.6 GB	15 Nov 2021
InventoryAdjustments_2.csv	22.8 GB	15 Nov 2021
InventoryTransferItems_0.csv	15.6 GB	15 Nov 2021
InventoryTransfers_0.csv	718.4 MB	15 Nov 2021
InventoryTypes_0.csv	15.2 GB	15 Nov 2021
LabResults_0.csv	1.2 GB	19 Nov 2021
LabResults_1.csv	1.1 GB	19 Nov 2021
LabResults_2.csv	138.0 MB	19 Nov 2021
LabRetests_0.csv	288 bytes	15 Nov 2021
Licensees_0.csv	1.3 MB	15 Nov 2021
MmeUser_0.csv	5.8 MB	15 Nov 2021
Plants_0.csv	13.2 GB	15 Nov 2021
SaleItems_0.csv	43.7 GB	15 Nov 2021
SaleItems_1.csv	49.3 GB	15 Nov 2021
SaleItems_2.csv	49.6 GB	15 Nov 2021
SaleItems_3.csv	42.9 GB	16 Nov 2021
Sales_0.csv	31.9 GB	16 Nov 2021
Sales_1.csv	38.0 GB	16 Nov 2021
Sales_2.csv	11.0 GB	16 Nov 2021
Strains_0.csv	270.1 MB	16 Nov 2021
Taxes_0.csv	130 bytes	16 Nov 2021
Users_0.csv	7.1 MB	16 Nov 2021

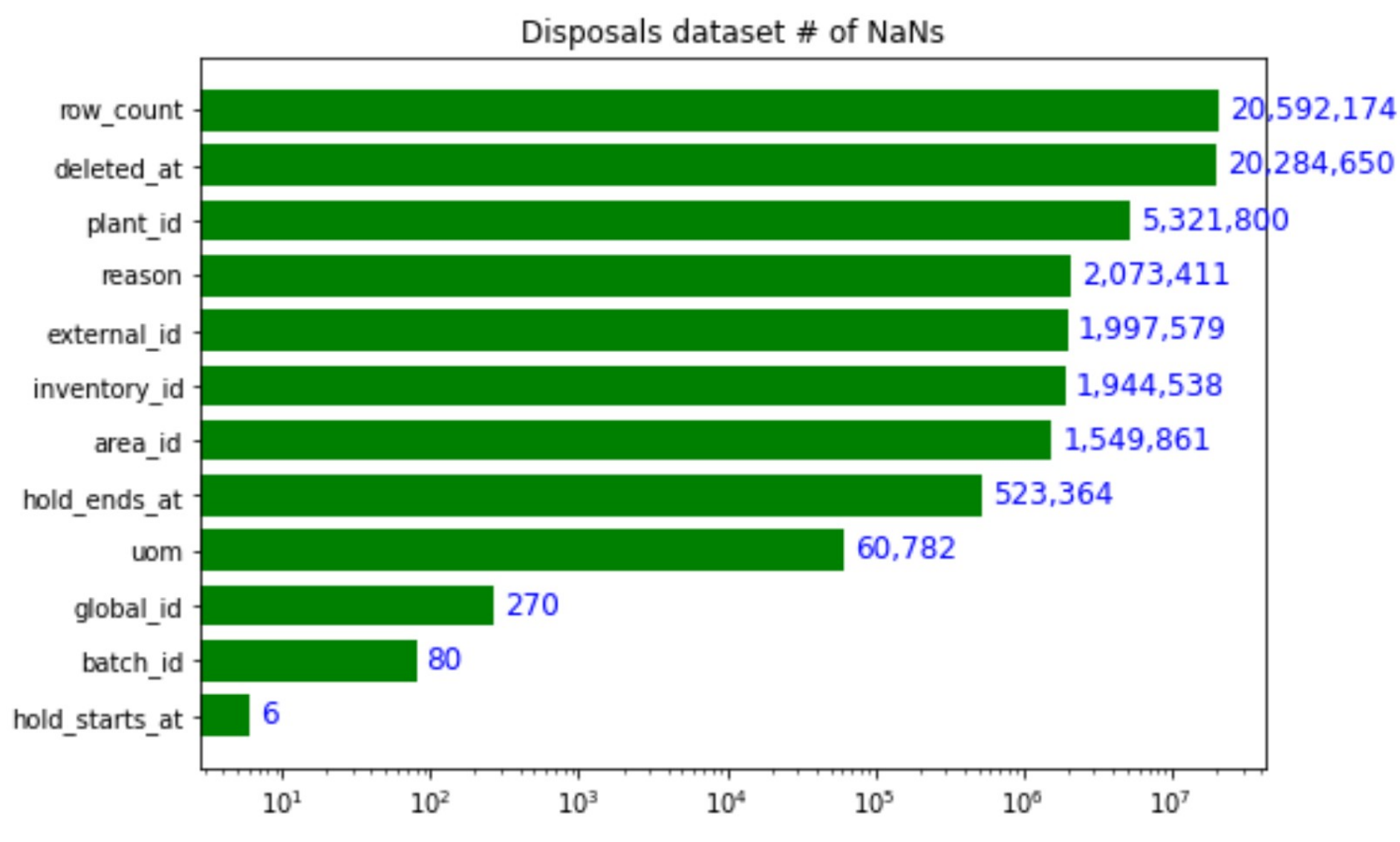
Areas_0.csv


Areas dataset # of NaNs
row_count
deleted_at
external_id
type
name
global_id
324,873
259,609
139,945
19
8
3
10^1
10^2
10^3
10^4
10^5

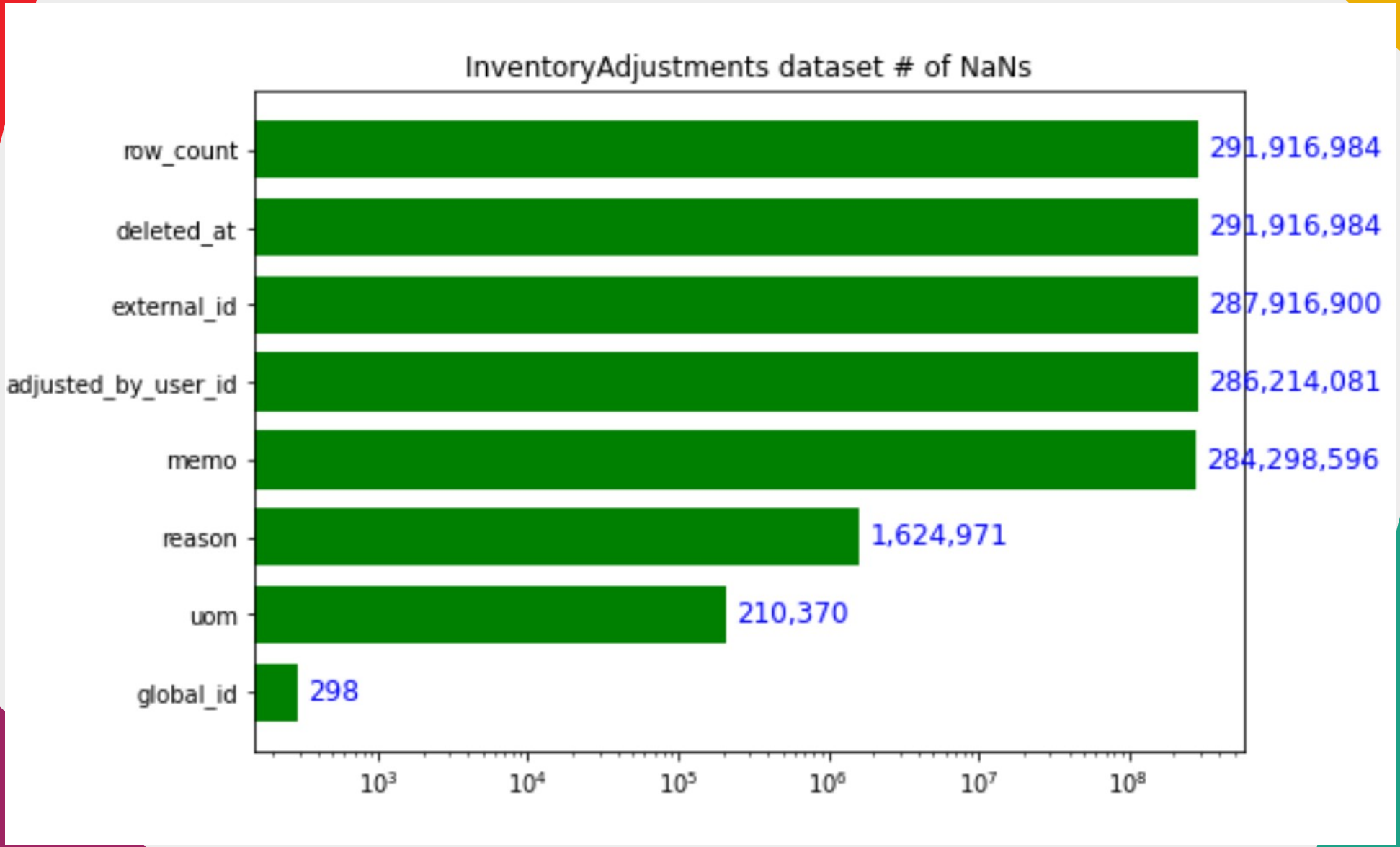
Batches_0.csv



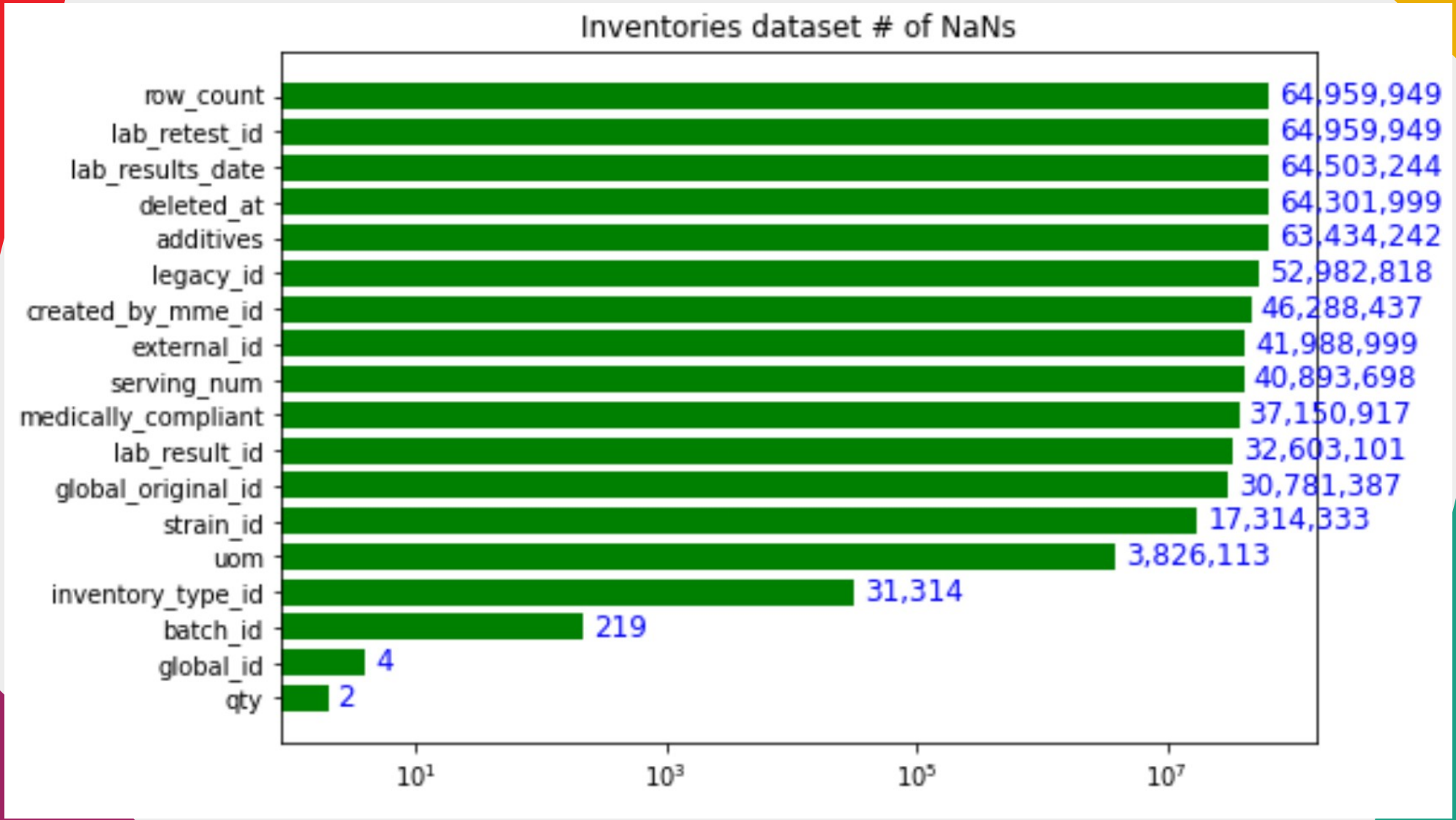
Disposals_0.csv



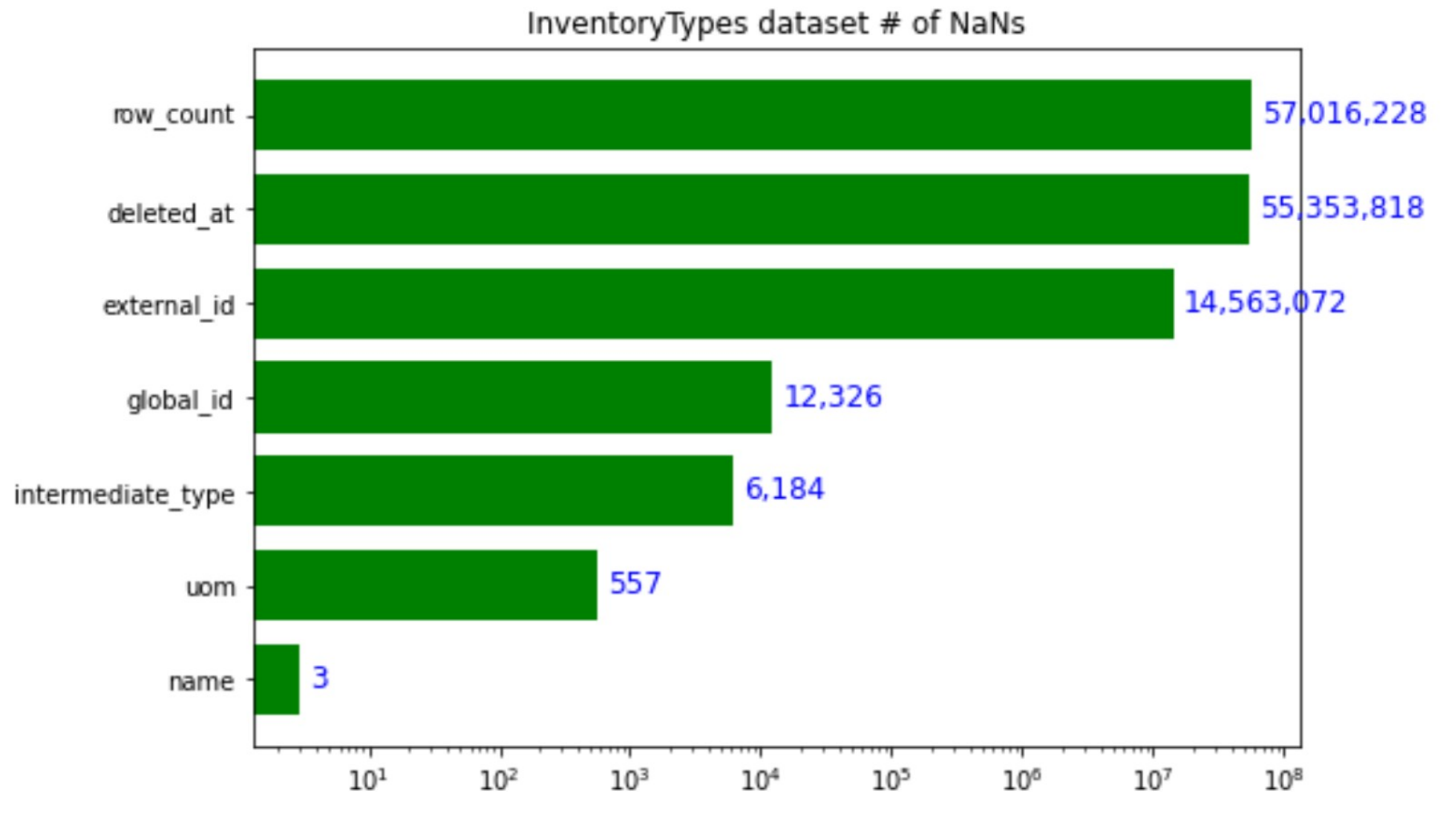
InventoryAdjustments_[0-2].csv



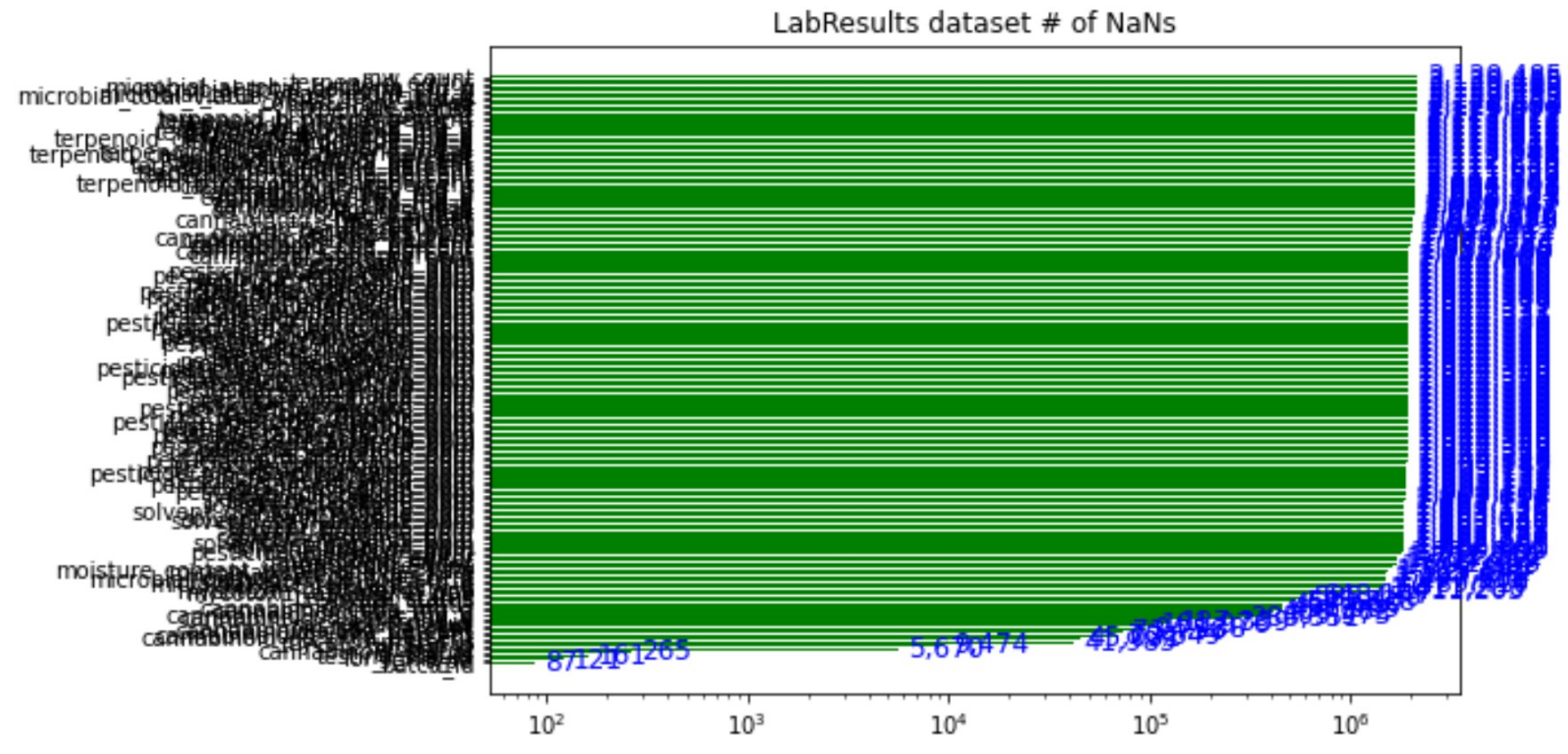
Inventories_0.csv



InventoryTypes_0.csv

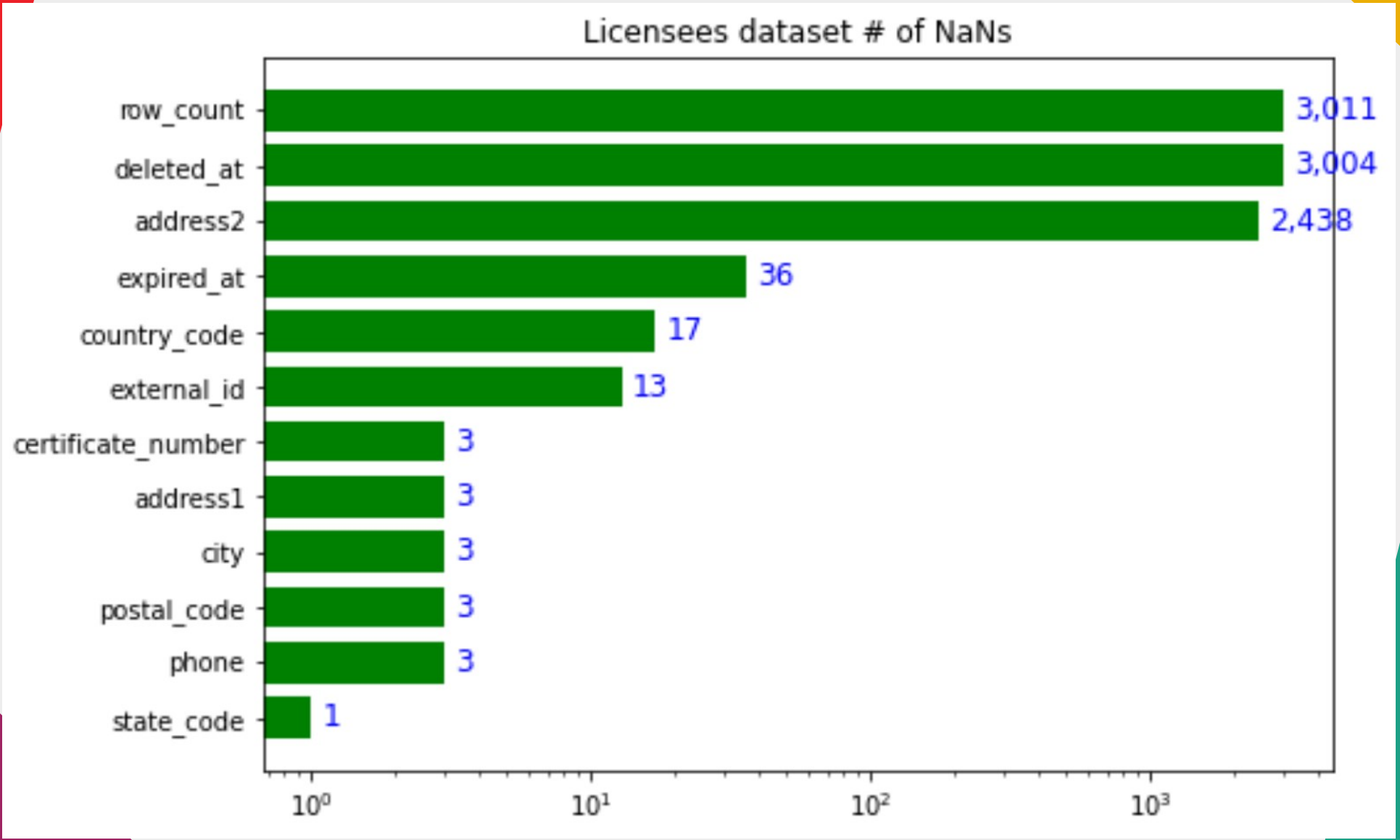


LabResults_[0-2].csv

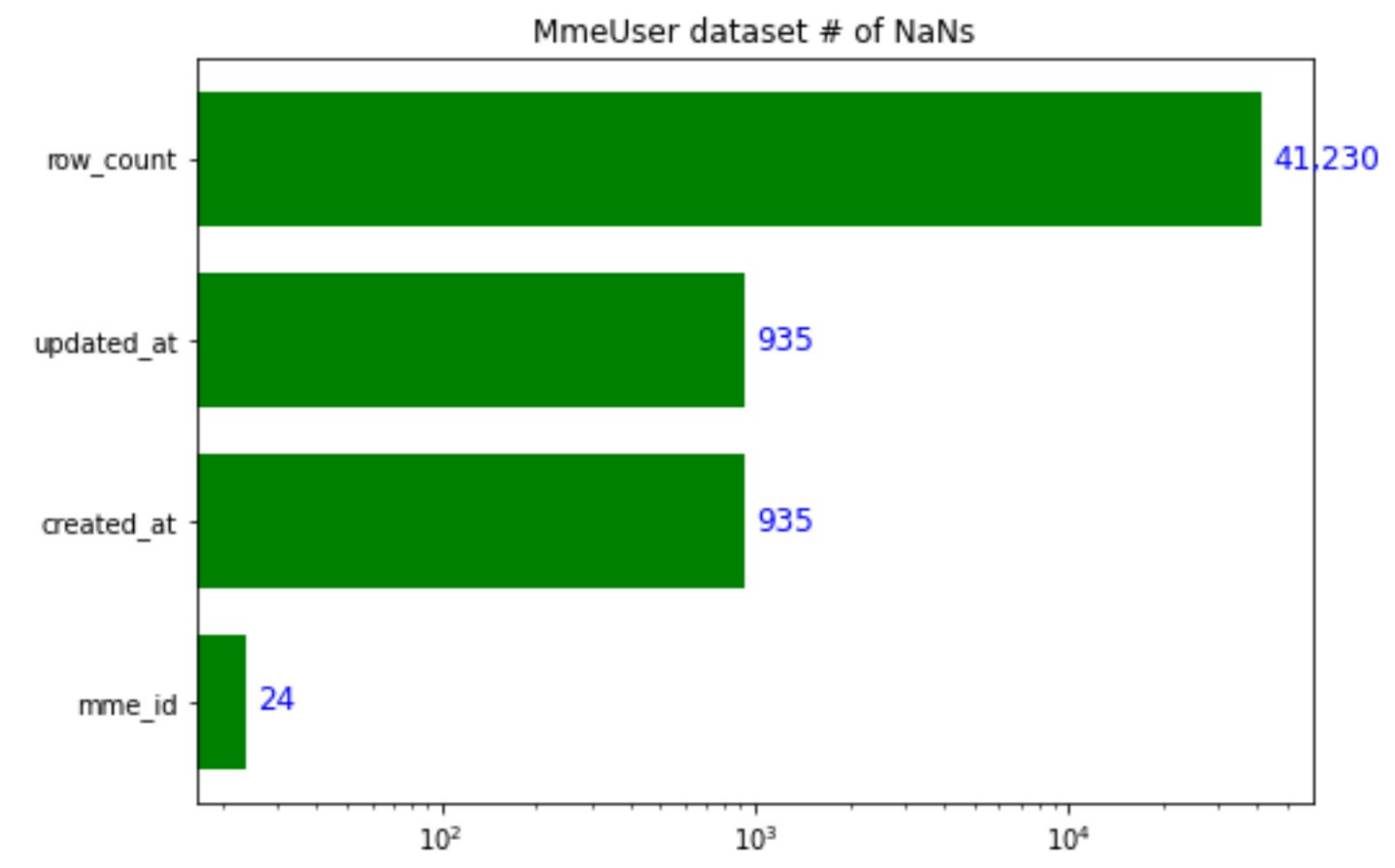


HaHa – Yes - This slide needs work, but clearly the dataset suffers from NaNs

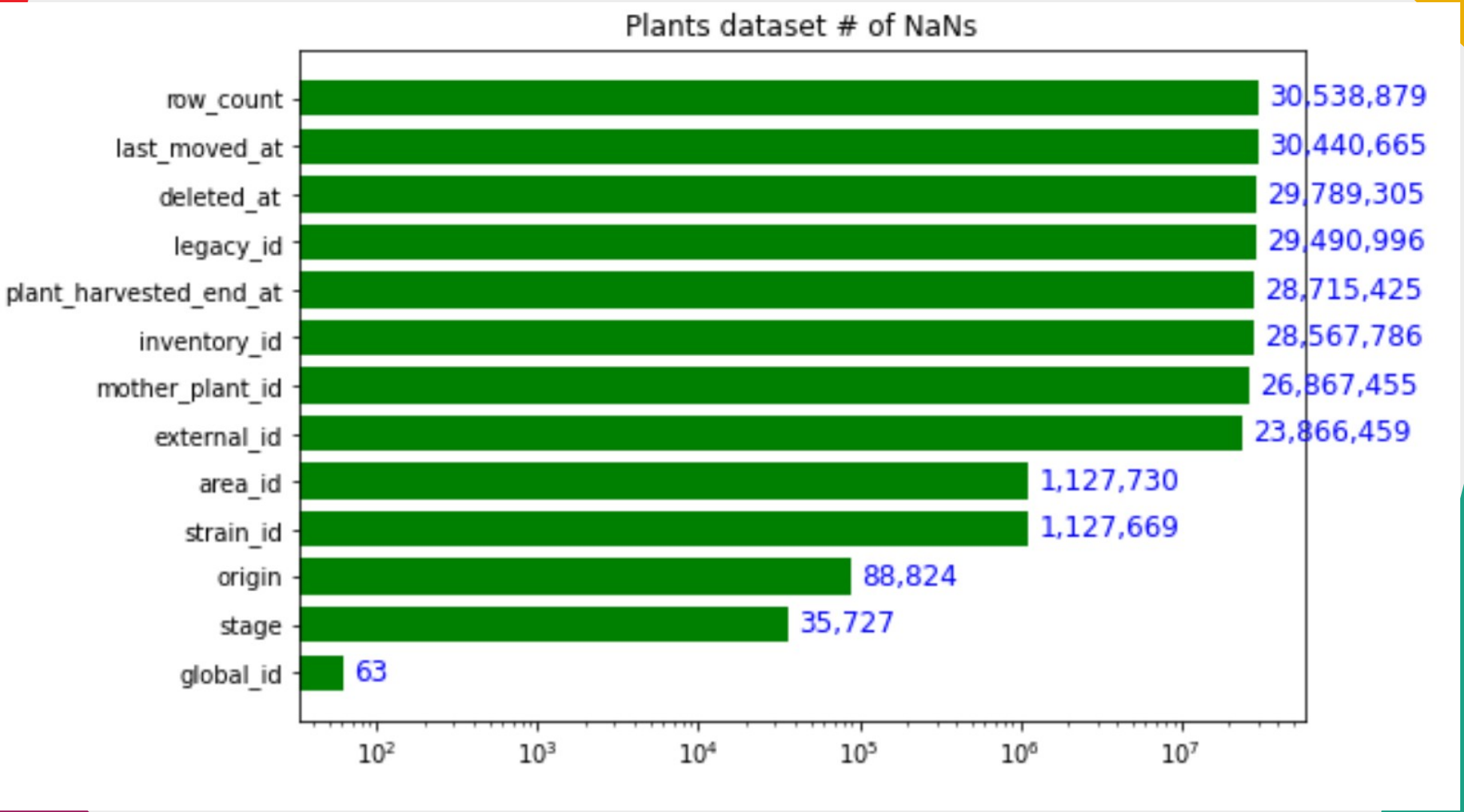
Licensees_0.csv



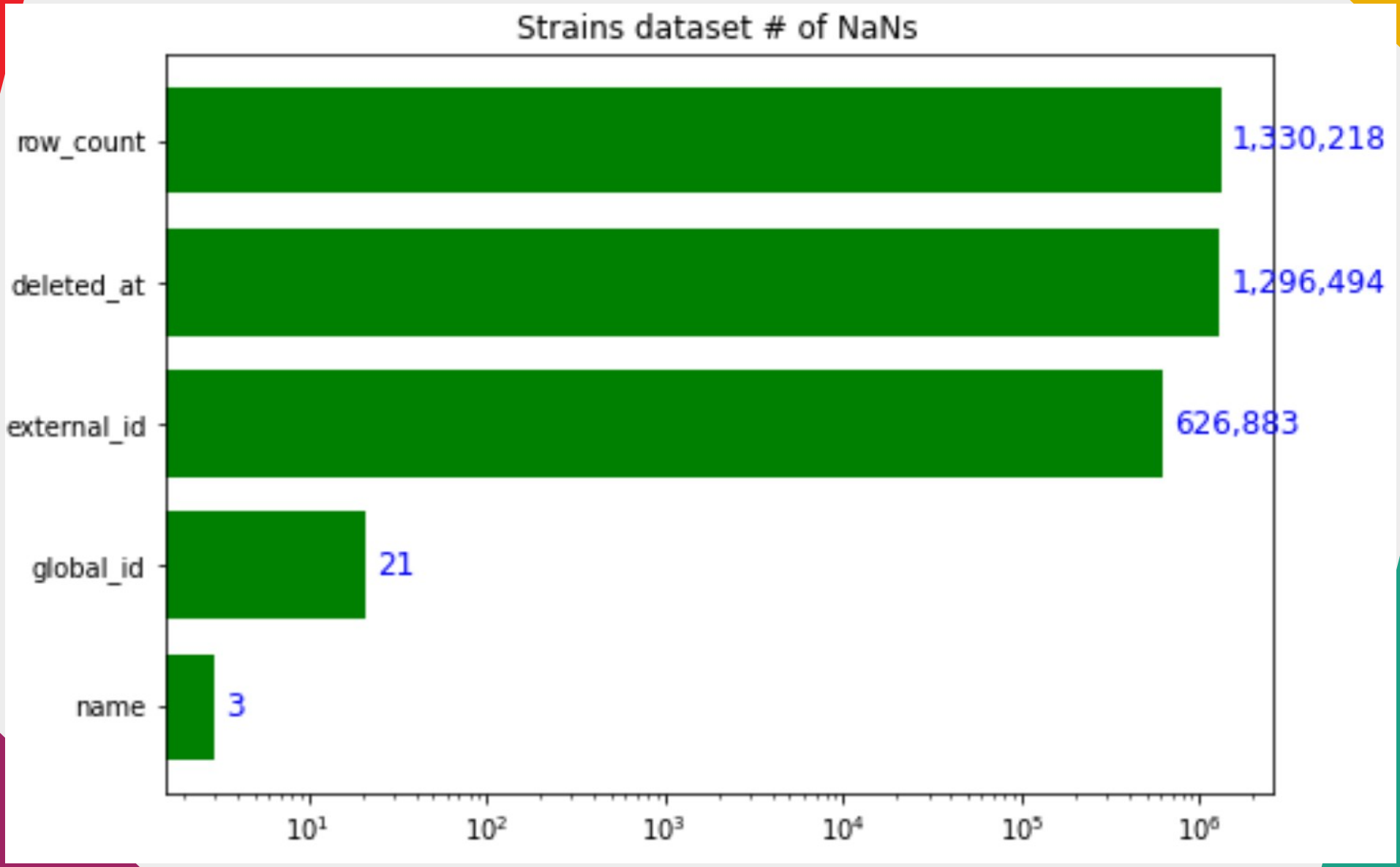
MmeUser_0.csv



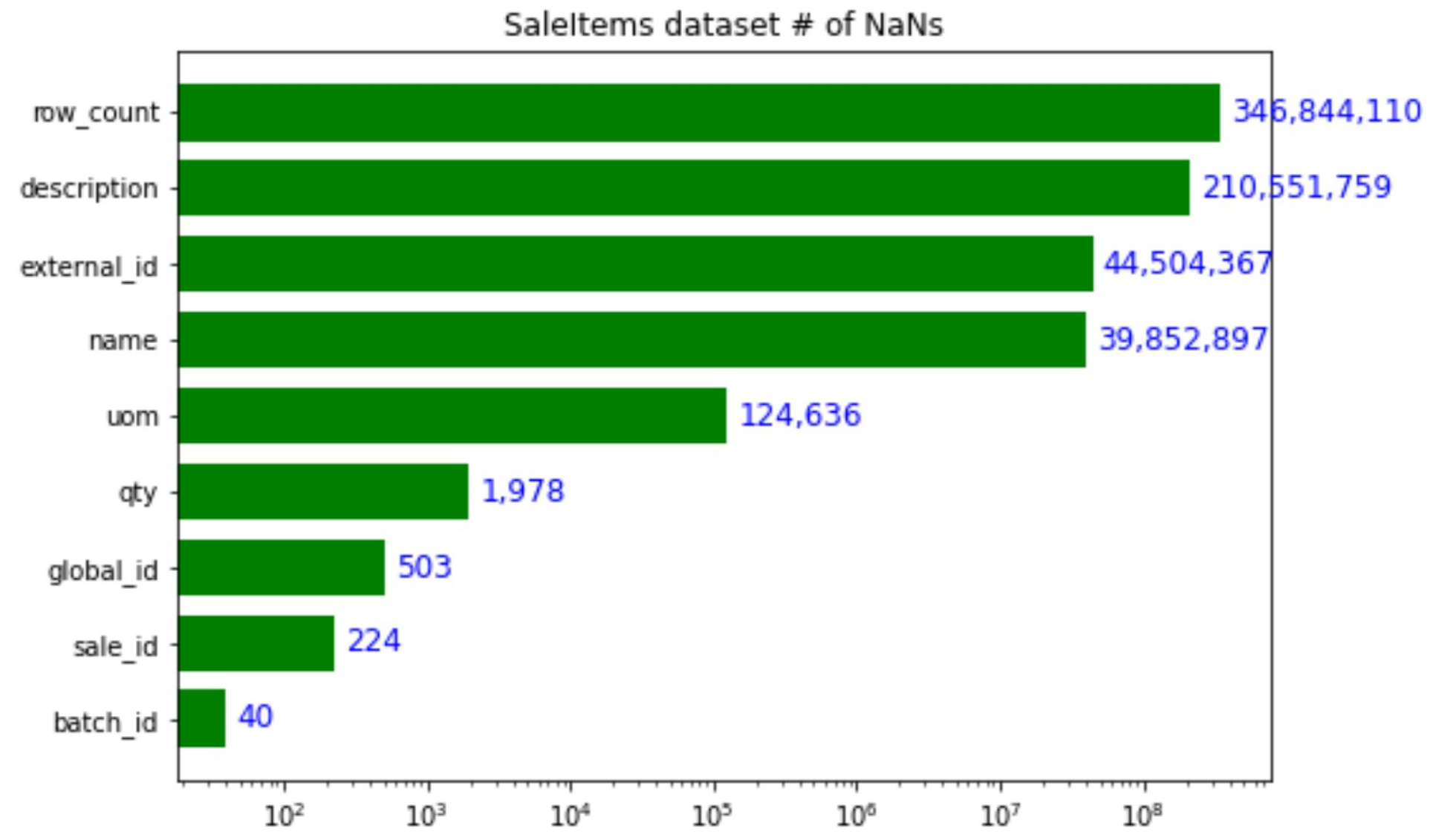
Plants_0.csv


Plants dataset # of NaNs
row_count
last_moved_at
deleted_at
legacy_id
plant_harvested_end_at
inventory_id
mother_plant_id
external_id
area_id
strain_id
origin
stage
global_id
30,538,879
30,440,665
29,789,305
29,490,996
28,715,425
28,567,786
26,867,455
23,866,459
1,127,730
1,127,669
88,824
35,727
63
10^2
10^3
10^4
10^5
10^6
10^7

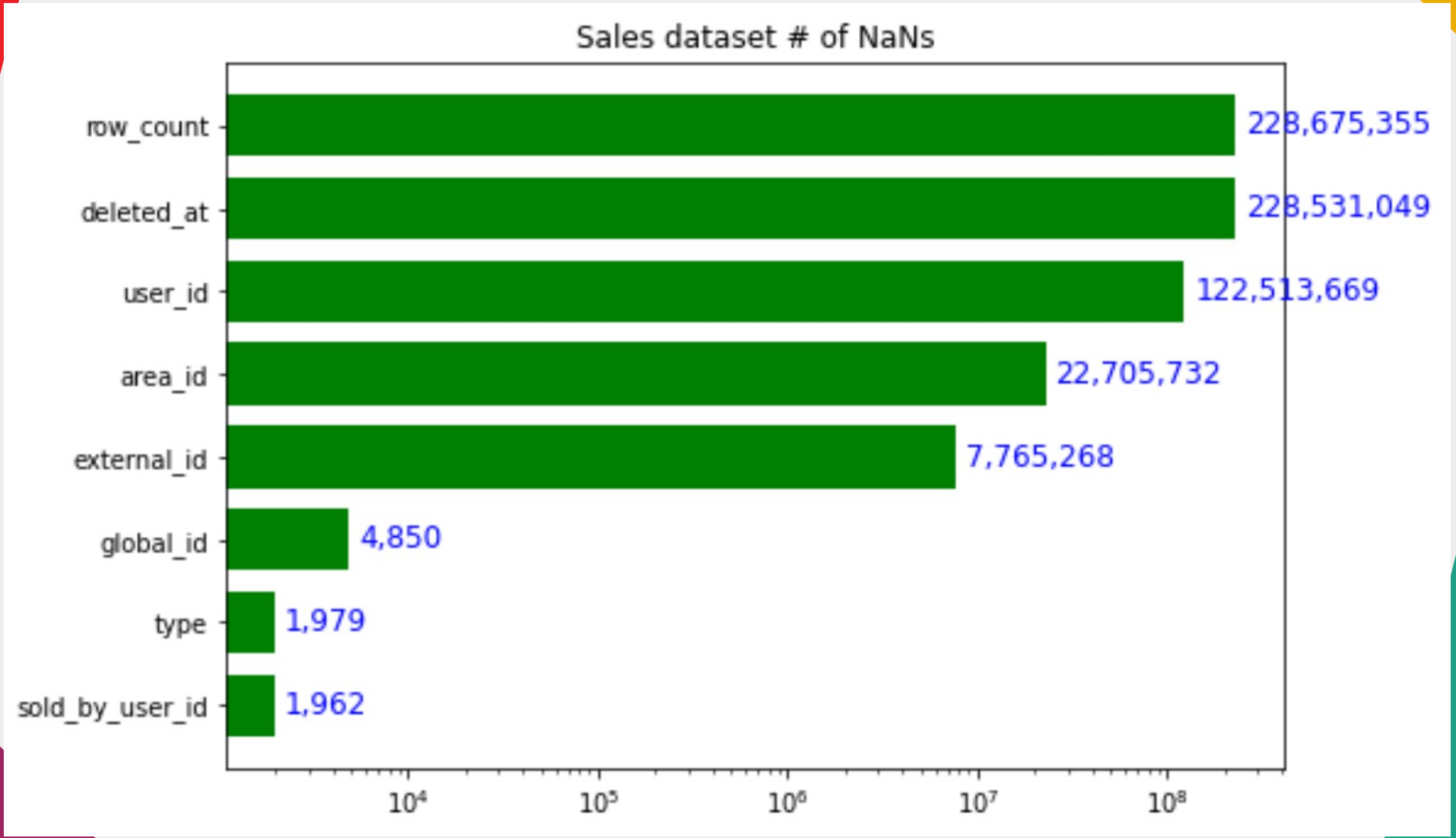
Strains_0.csv



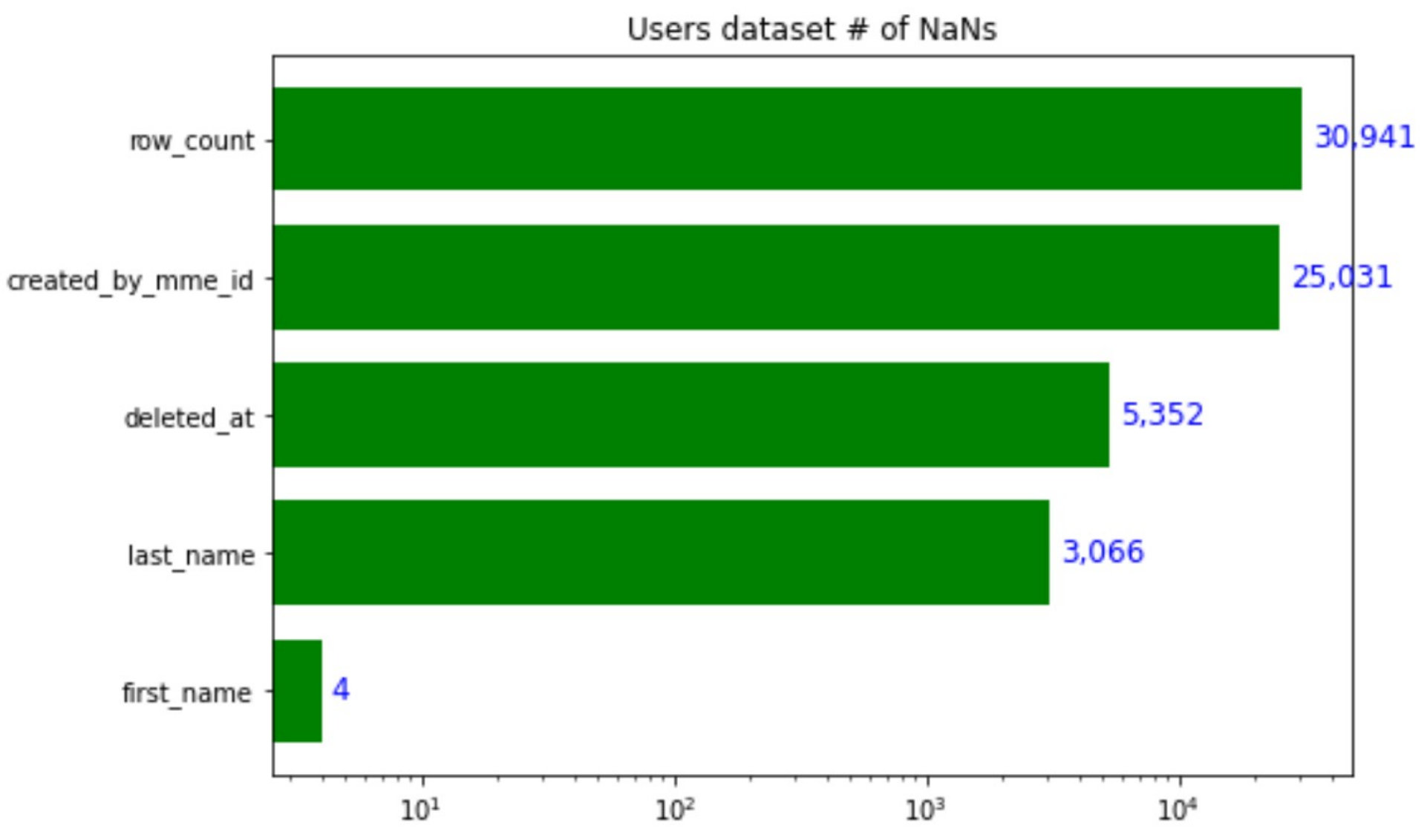
SaleItems_[0-3].csv



Sales_[0-2].csv


Sales dataset # of NaNs
row_count
deleted_at
user_id
area_id
external_id
global_id
type
sold_by_user_id
228,675,355
228,531,049
122,513,669
22,705,732
7,765,268
4,850
1,979
1,962
10^4
10^5
10^6
10^7
10^8

Users_0.csv


Users dataset # of NaNs
row_count
created_by_mme_id
deleted_at
last_name
first_name
30,941
25,031
5,352
3,066
4
10^1
10^2
10^3
10^4