Washington State Data 2021 “Spacy NLP” notes

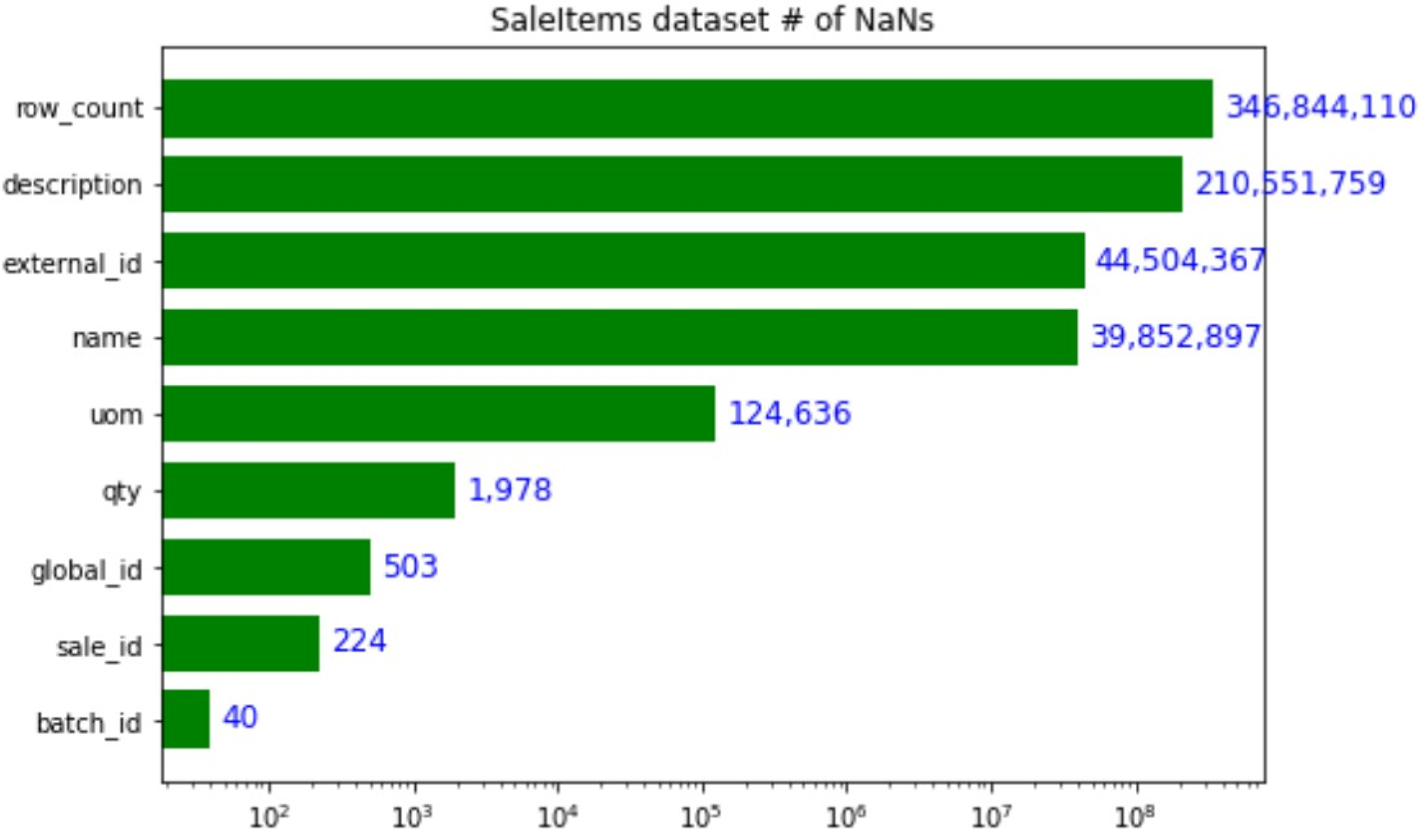
- Notes regarding wrapper python script that runs Charles Rice UFO Software's Spacy NLP code `quantity_from_description.ipynb` which pulls quantity data from description using “Spacy NLP” on augmented SaleItems[0-3].csv dataframes (chunk_df) then writes new dataframes to new csv files. Count NaNs before and after “Spacy NLP” is applied. Dataframe observations.

- issue: error “when df.name==NaN was copied over to df.description”
    - fixed via replacing df.name NaN with “Name”
- spacy NLP code worked fabulously
- df.description has id type fields xxx-xxxx-xxxx-xxx-xxx that prevented name from being copied over and used for “quantity”
- files increased in size as expected

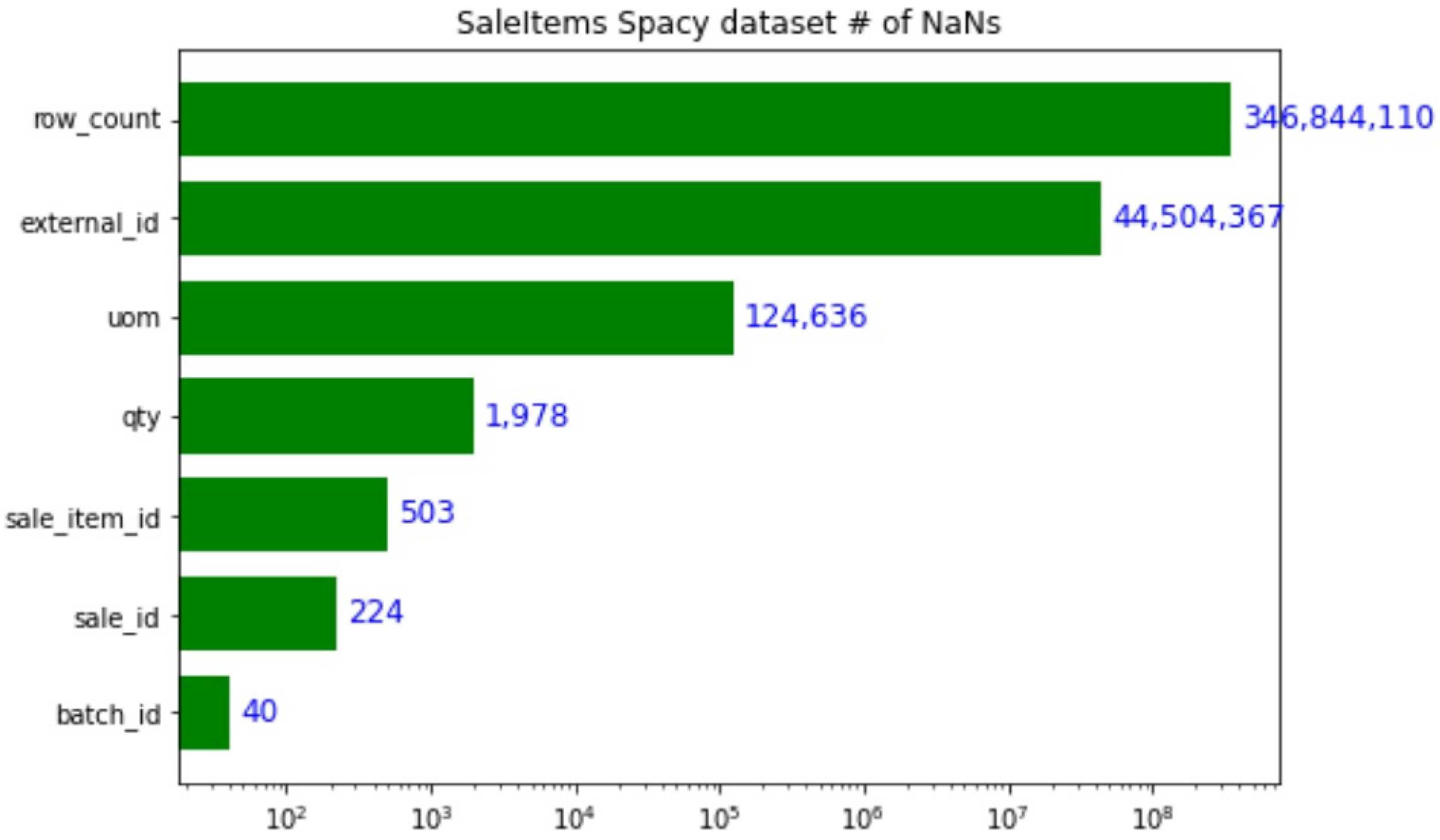
Next: clean up wrapper script, add slide with percentages, save in parquet file format - upload to Google Drive

Candace O’Sullivan-Sutherland
Contributors: Charles Rice, Keegan Skeate, Gerry Pallor
May 5, 2022 created with LibreOffice Impress

Count NaNs on original dataset SaleItems[0-3].csv

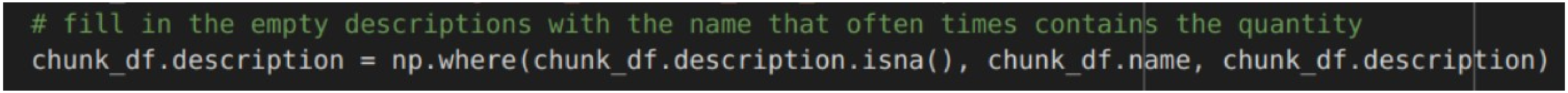


Count NaNs after Python Spacy NLP is applied

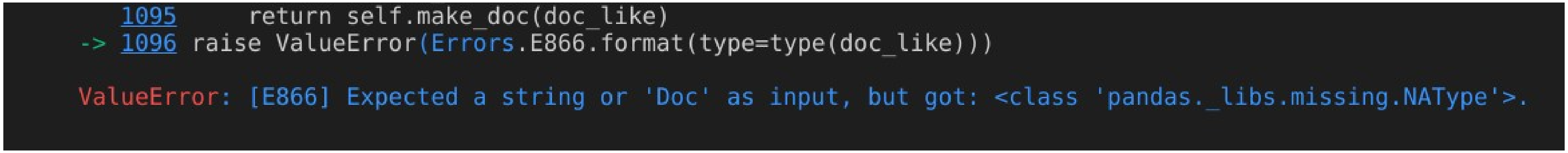


- “global_id” renamed to “sale_item_id”
- “name”, “description” NaNs replaced with “Name”
- “quantity” filled with “Spacy NLP derived” {quantity} or 0

Error when name is NaN



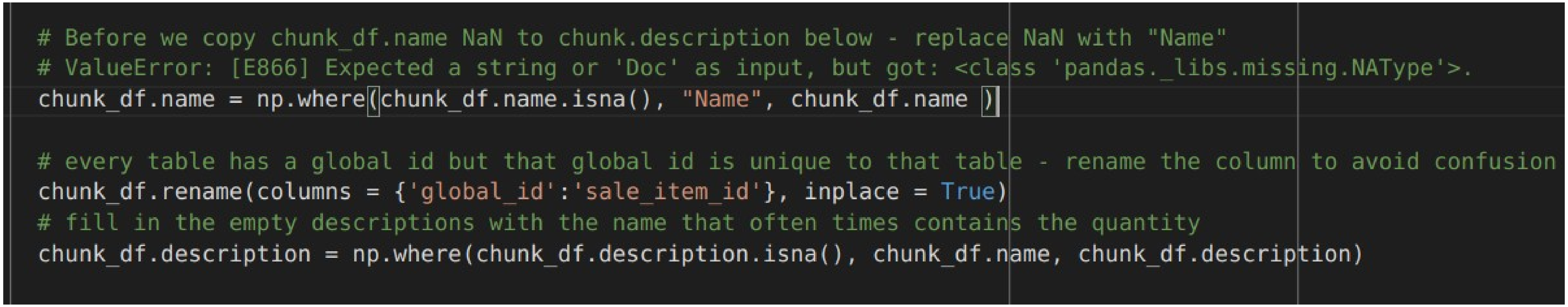
```
# fill in the empty descriptions with the name that often times contains the quantity
chunk_df.description = np.where(chunk_df.description.isna(), chunk_df.name, chunk_df.description)
```



```
    1095     return self.make_doc(doc_like)
 -> 1096 raise ValueError(Errors.E866.format(type=type(doc_like)))

ValueError: [E866] Expected a string or 'Doc' as input, but got: <class 'pandas._libs.missing.NAType'>.
```

# Before we copy chunk_df.name NaN to chunk.description below - replace NaN with "Name"
chunk_df.name = np.where(chunk_df.name.isna(), "nd", chunk_df.name) or test for name.isna()
i.e. if chunk_df.name != 'nan'?

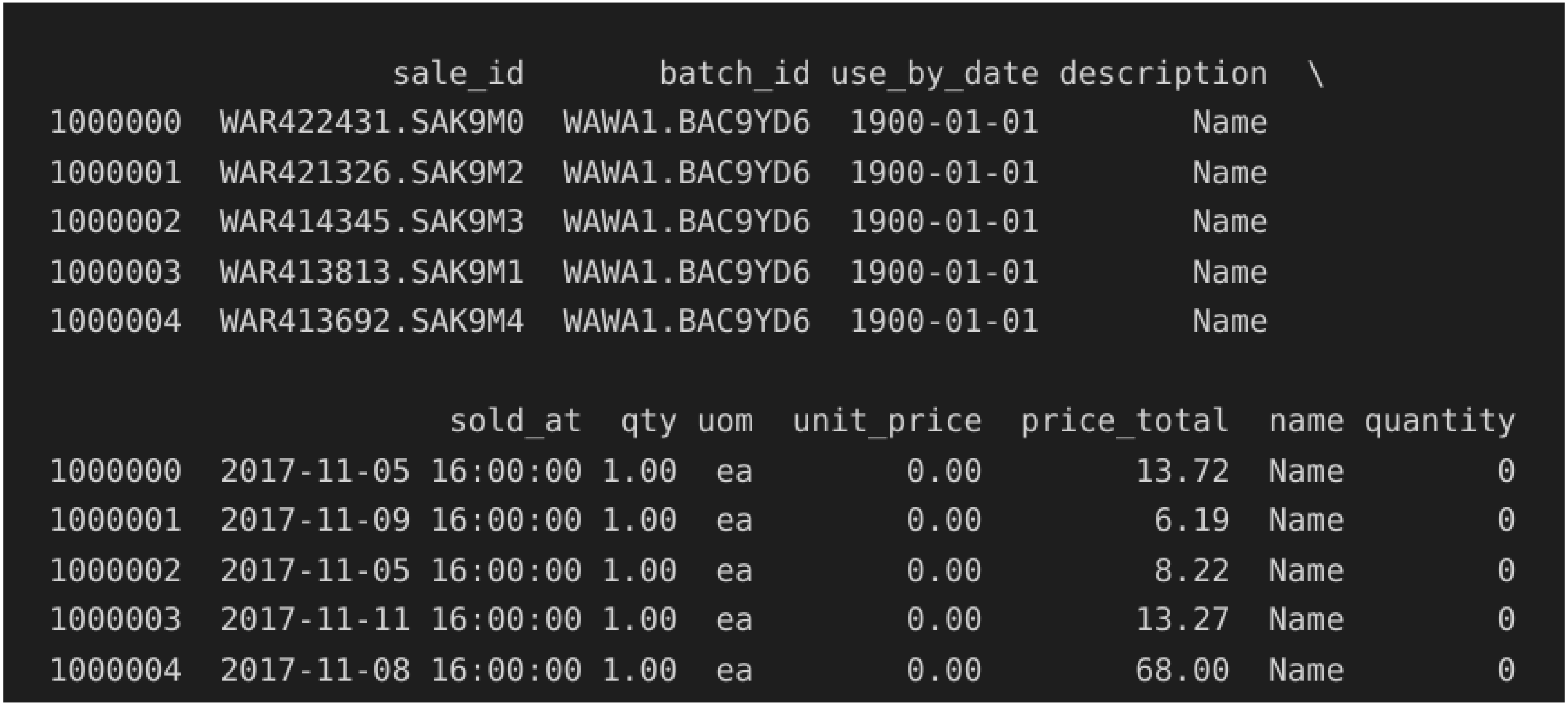
df.name.isna() fill with “Name”eliminates error



```
# Before we copy chunk_df.name NaN to chunk.description below - replace NaN with "Name"
# ValueError: [E866] Expected a string or 'Doc' as input, but got: <class 'pandas._libs.missing.NAType'>.
chunk_df.name = np.where(chunk_df.name.isna(), "Name", chunk_df.name )

# every table has a global id but that global id is unique to that table - rename the column to avoid confusion
chunk_df.rename(columns = {'global_id':'sale_item_id'}, inplace = True)
# fill in the empty descriptions with the name that often times contains the quantity
chunk_df.description = np.where(chunk_df.description.isna(), chunk_df.name, chunk_df.description)
```

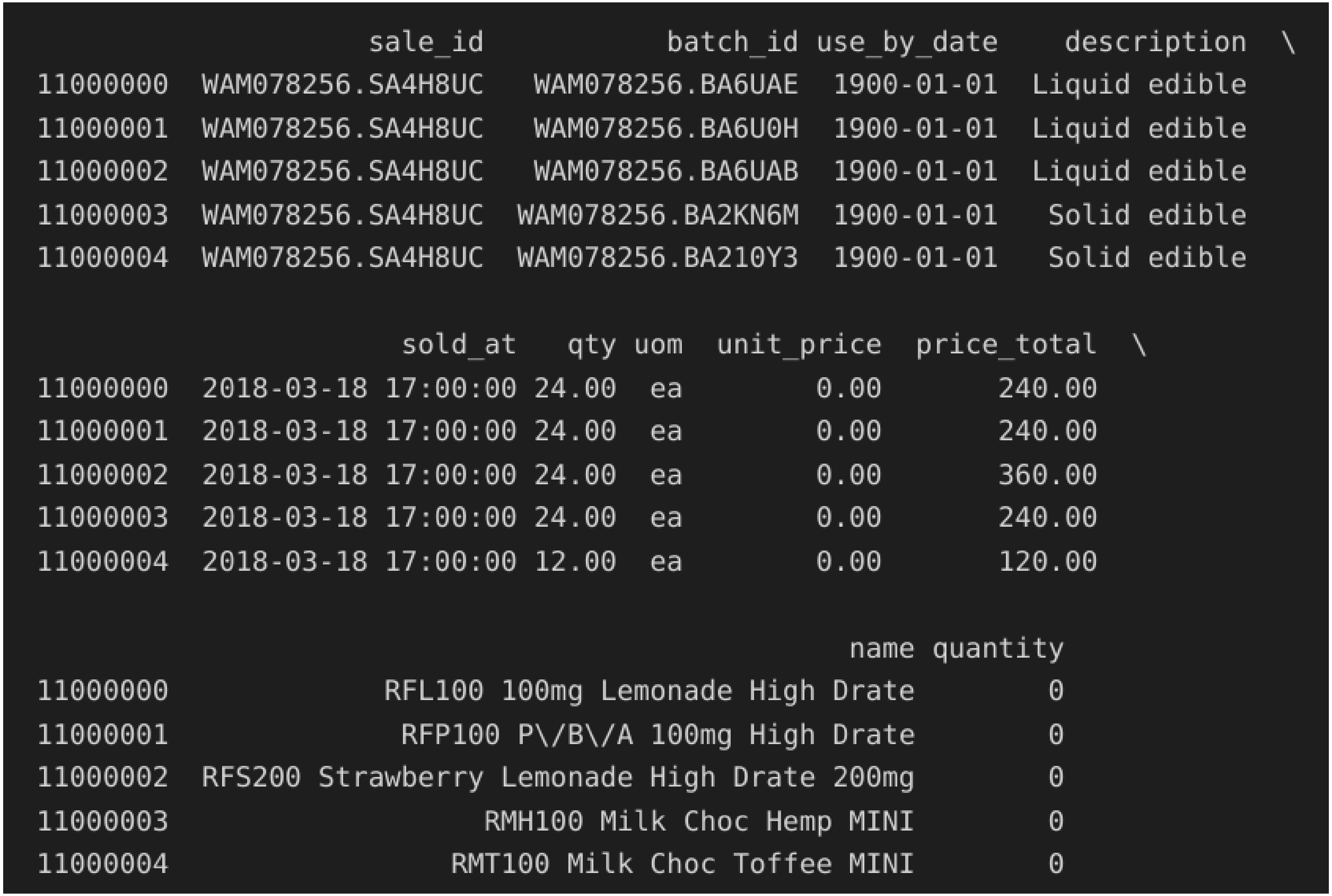
100% reproducible remove the first line get the error

df.name.isna()==True name=“Name” before copied to Description



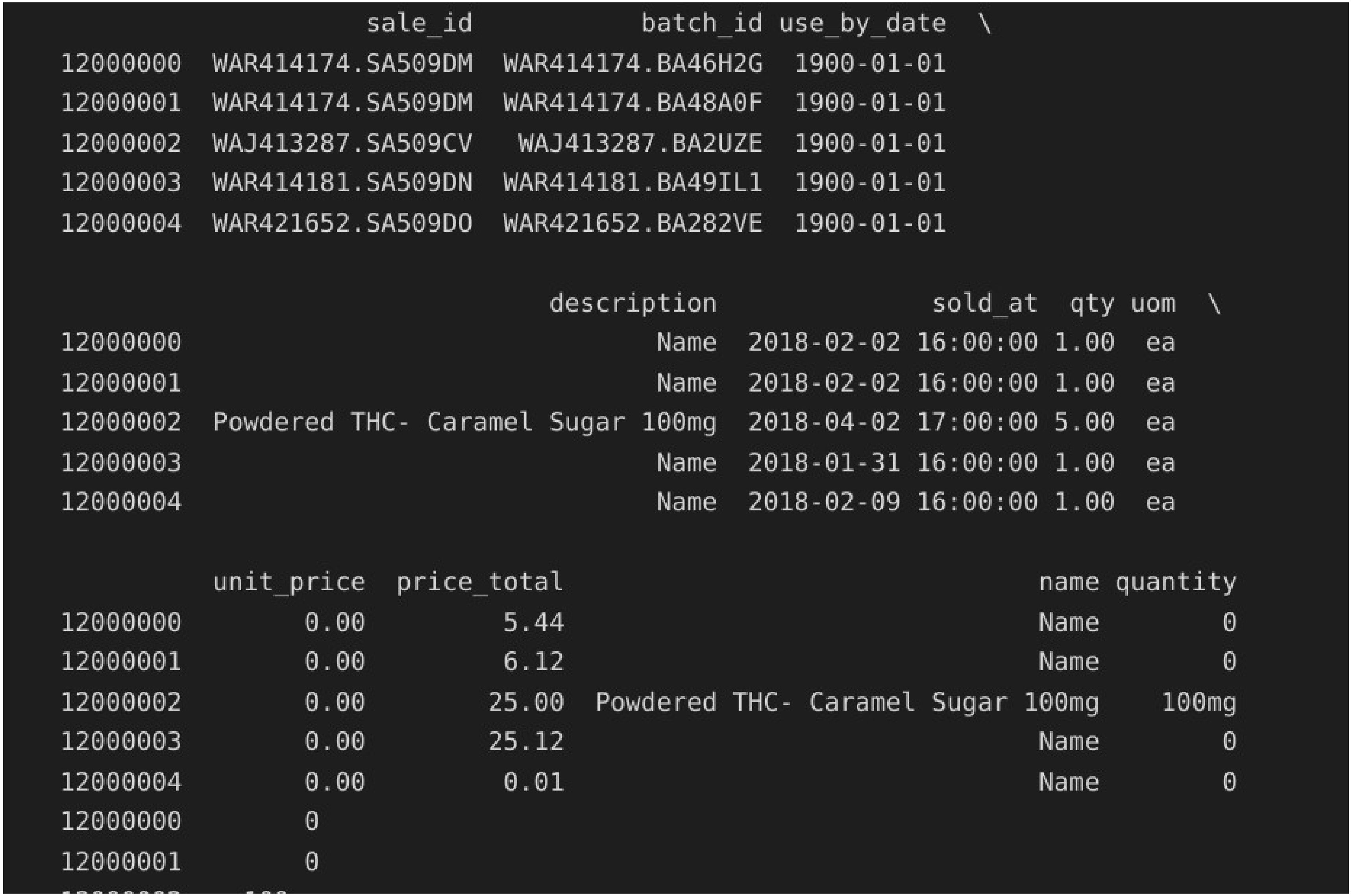
```
                 sale_id       batch_id use_by_date description  \
1000000  WAR422431.SAK9M0  WAWA1.BAC9YD6  1900-01-01        Name   
1000001  WAR421326.SAK9M2  WAWA1.BAC9YD6  1900-01-01        Name   
1000002  WAR414345.SAK9M3  WAWA1.BAC9YD6  1900-01-01        Name   
1000003  WAR413813.SAK9M1  WAWA1.BAC9YD6  1900-01-01        Name   
1000004  WAR413692.SAK9M4  WAWA1.BAC9YD6  1900-01-01        Name   

                    sold_at  qty uom  unit_price  price_total  name quantity  
1000000  2017-11-05 16:00:00 1.00  ea        0.00        13.72  Name        0  
1000001  2017-11-09 16:00:00 1.00  ea        0.00         6.19  Name        0  
1000002  2017-11-05 16:00:00 1.00  ea        0.00         8.22  Name        0  
1000003  2017-11-11 16:00:00 1.00  ea        0.00        13.27  Name        0  
1000004  2017-11-08 16:00:00 1.00  ea        0.00        68.00  Name        0  
```

Works: quantity data was in name



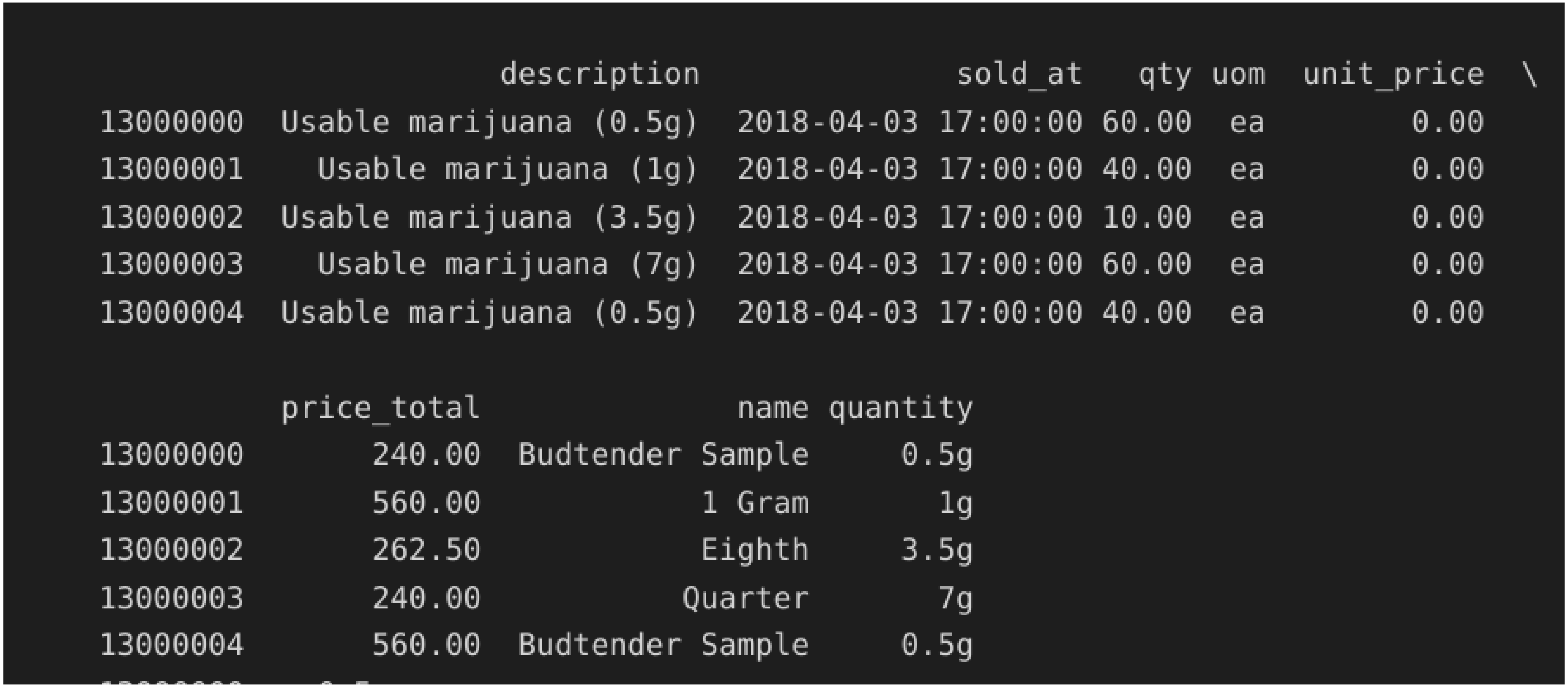
```
                    sale_id           batch_id use_by_date    description  \
11000000  WAM078256.SA4H8UC   WAM078256.BA6UAE  1900-01-01  Liquid edible
11000001  WAM078256.SA4H8UC   WAM078256.BA6U0H  1900-01-01  Liquid edible
11000002  WAM078256.SA4H8UC   WAM078256.BA6UAB  1900-01-01  Liquid edible
11000003  WAM078256.SA4H8UC  WAM078256.BA2KN6M  1900-01-01   Solid edible
11000004  WAM078256.SA4H8UC  WAM078256.BA210Y3  1900-01-01   Solid edible

                      sold_at   qty uom  unit_price  price_total  \
11000000  2018-03-18 17:00:00 24.00  ea        0.00       240.00
11000001  2018-03-18 17:00:00 24.00  ea        0.00       240.00
11000002  2018-03-18 17:00:00 24.00  ea        0.00       360.00
11000003  2018-03-18 17:00:00 24.00  ea        0.00       240.00
11000004  2018-03-18 17:00:00 12.00  ea        0.00       120.00

                                                 name quantity
11000000             RFL100 100mg Lemonade High Drate        0
11000001              RFP100 P\/B\/A 100mg High Drate        0
11000002  RFS200 Strawberry Lemonade High Drate 200mg        0
11000003                   RMH100 Milk Choc Hemp MINI        0
11000004                 RMT100 Milk Choc Toffee MINI        0
```

Works fabulously...



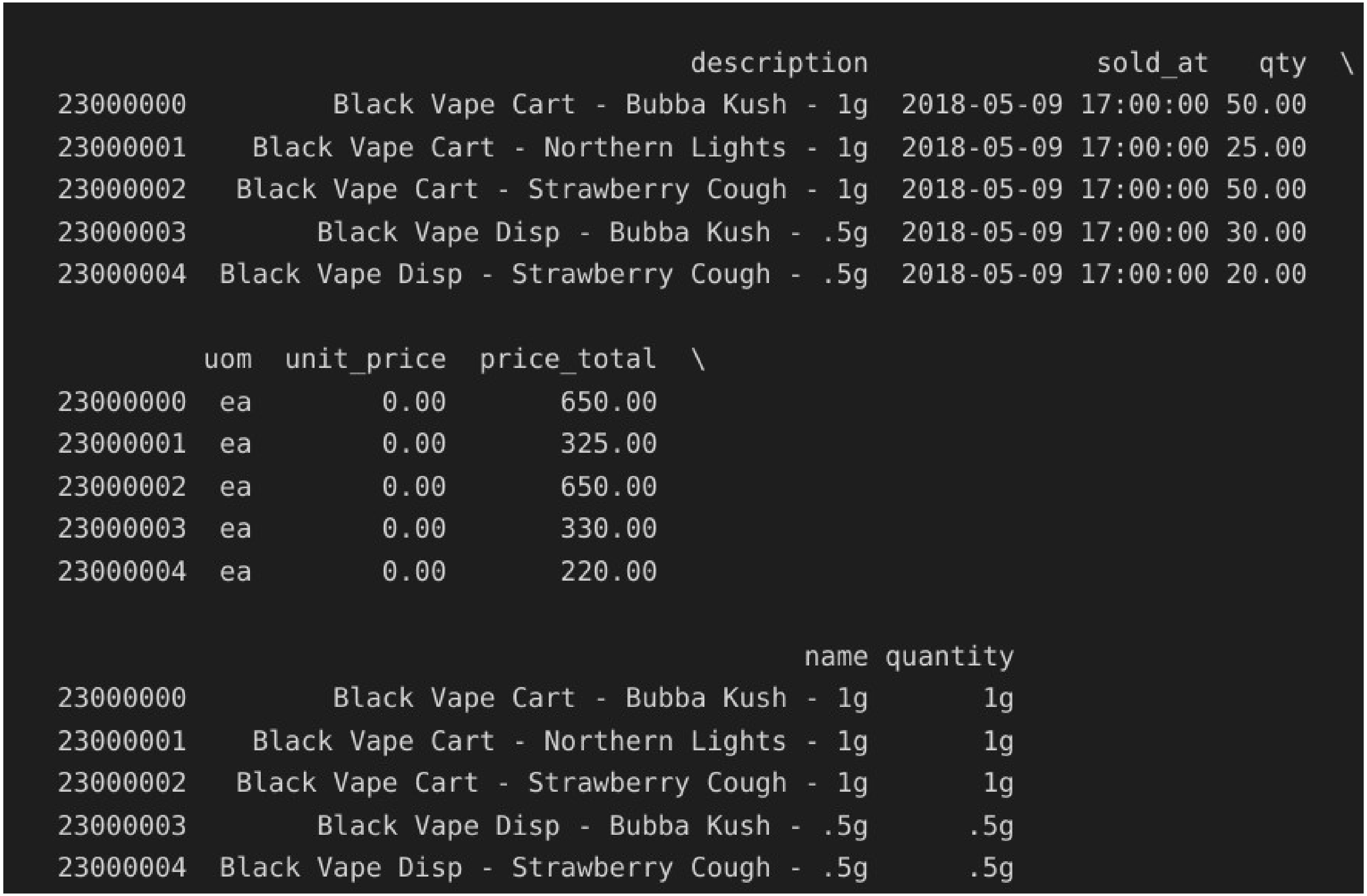
```
                     sale_id           batch_id use_by_date  \
12000000  WAR414174.SA509DM  WAR414174.BA46H2G  1900-01-01
12000001  WAR414174.SA509DM  WAR414174.BA48A0F  1900-01-01
12000002  WAJ413287.SA509CV   WAJ413287.BA2UZE  1900-01-01
12000003  WAR414181.SA509DN  WAR414181.BA49IL1  1900-01-01
12000004  WAR421652.SA509DO  WAR421652.BA282VE  1900-01-01

                                description              sold_at  qty uom  \
12000000                               Name  2018-02-02 16:00:00 1.00  ea
12000001                               Name  2018-02-02 16:00:00 1.00  ea
12000002  Powdered THC- Caramel Sugar 100mg  2018-04-02 17:00:00 5.00  ea
12000003                               Name  2018-01-31 16:00:00 1.00  ea
12000004                               Name  2018-02-09 16:00:00 1.00  ea

          unit_price  price_total                              name quantity
12000000        0.00         5.44                              Name        0
12000001        0.00         6.12                              Name        0
12000002        0.00        25.00  Powdered THC- Caramel Sugar 100mg    100mg
12000003        0.00        25.12                              Name        0
12000004        0.00         0.01                              Name        0
12000000        0
12000001        0
12000002    100
```

Works:



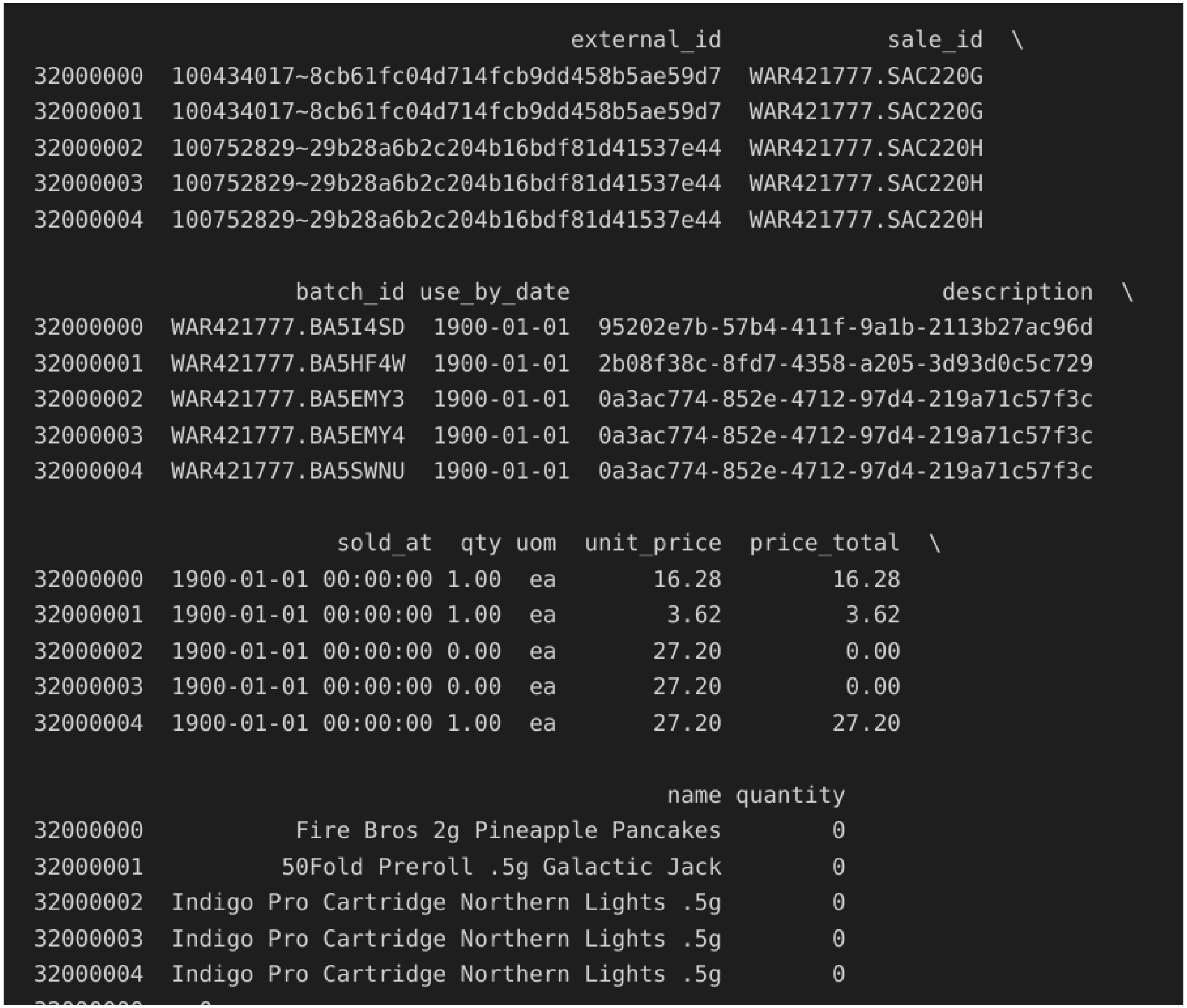
```
                        description             sold_at   qty uom  unit_price  \
13000000  Usable marijuana (0.5g)  2018-04-03 17:00:00 60.00  ea        0.00
13000001    Usable marijuana (1g)  2018-04-03 17:00:00 40.00  ea        0.00
13000002  Usable marijuana (3.5g)  2018-04-03 17:00:00 10.00  ea        0.00
13000003    Usable marijuana (7g)  2018-04-03 17:00:00 60.00  ea        0.00
13000004  Usable marijuana (0.5g)  2018-04-03 17:00:00 40.00  ea        0.00

          price_total              name quantity
13000000       240.00  Budtender Sample     0.5g
13000001       560.00            1 Gram       1g
13000002       262.50            Eighth     3.5g
13000003       240.00           Quarter       7g
13000004       560.00  Budtender Sample     0.5g
```

Works:



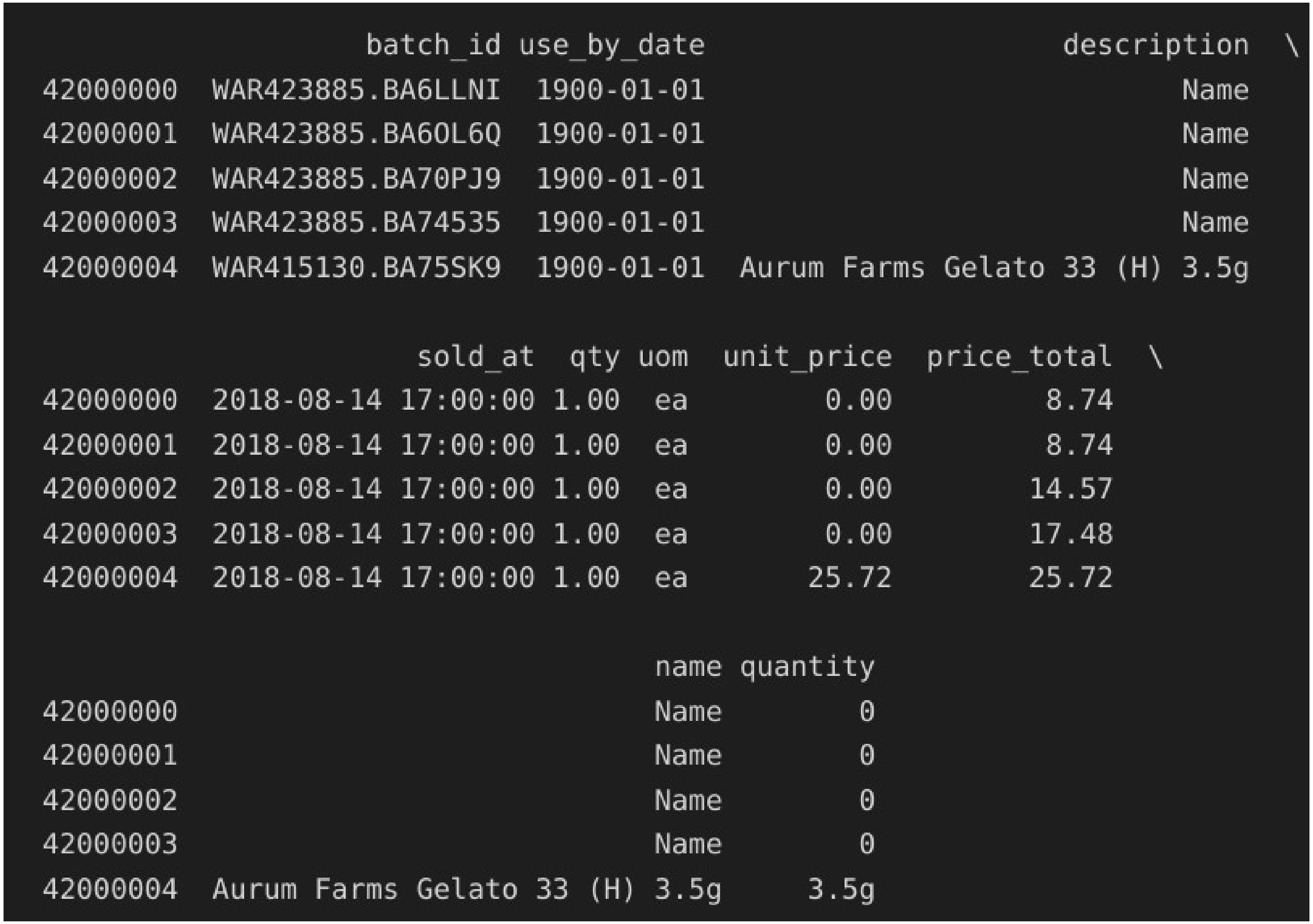
```
                                      description             sold_at   qty  \
23000000         Black Vape Cart - Bubba Kush - 1g  2018-05-09 17:00:00 50.00
23000001    Black Vape Cart - Northern Lights - 1g  2018-05-09 17:00:00 25.00
23000002   Black Vape Cart - Strawberry Cough - 1g  2018-05-09 17:00:00 50.00
23000003        Black Vape Disp - Bubba Kush - .5g  2018-05-09 17:00:00 30.00
23000004  Black Vape Disp - Strawberry Cough - .5g  2018-05-09 17:00:00 20.00

         uom  unit_price  price_total  \
23000000  ea        0.00       650.00
23000001  ea        0.00       325.00
23000002  ea        0.00       650.00
23000003  ea        0.00       330.00
23000004  ea        0.00       220.00


                                             name quantity
23000000         Black Vape Cart - Bubba Kush - 1g       1g
23000001    Black Vape Cart - Northern Lights - 1g       1g
23000002   Black Vape Cart - Strawberry Cough - 1g       1g
23000003        Black Vape Disp - Bubba Kush - .5g      .5g
23000004  Black Vape Disp - Strawberry Cough - .5g      .5g
```

Works: Description had garbage, name has the data



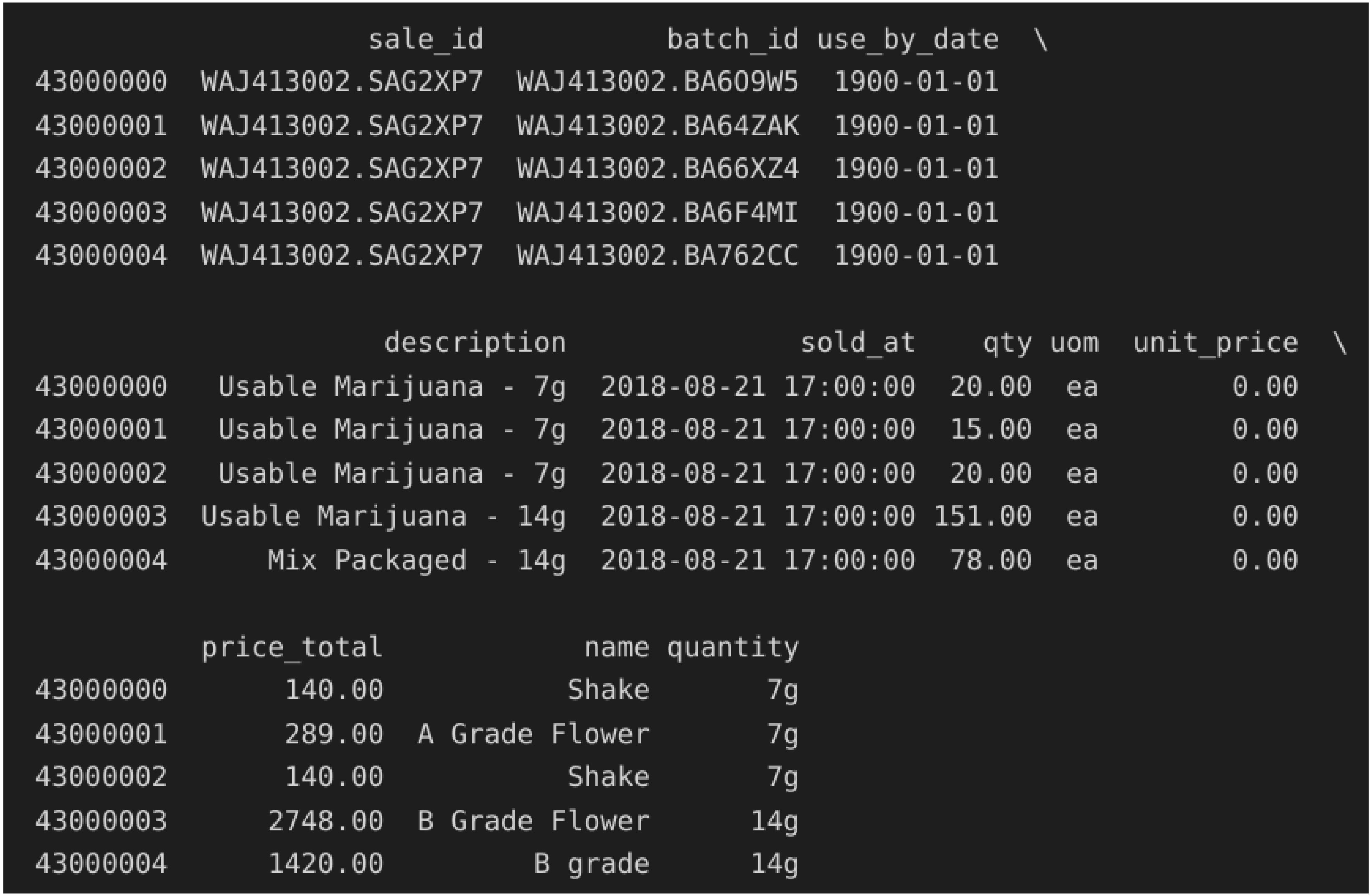
```
                                       external_id             sale_id  \
32000000  100434017~8cb61fc04d714fcb9dd458b5ae59d7  WAR421777.SAC220G
32000001  100434017~8cb61fc04d714fcb9dd458b5ae59d7  WAR421777.SAC220G
32000002  100752829~29b28a6b2c204b16bdf81d41537e44  WAR421777.SAC220H
32000003  100752829~29b28a6b2c204b16bdf81d41537e44  WAR421777.SAC220H
32000004  100752829~29b28a6b2c204b16bdf81d41537e44  WAR421777.SAC220H

                   batch_id use_by_date                           description  \
32000000  WAR421777.BA5I4SD  1900-01-01  95202e7b-57b4-411f-9a1b-2113b27ac96d
32000001  WAR421777.BA5HF4W  1900-01-01  2b08f38c-8fd7-4358-a205-3d93d0c5c729
32000002  WAR421777.BA5EMY3  1900-01-01  0a3ac774-852e-4712-97d4-219a71c57f3c
32000003  WAR421777.BA5EMY4  1900-01-01  0a3ac774-852e-4712-97d4-219a71c57f3c
32000004  WAR421777.BA5SWNU  1900-01-01  0a3ac774-852e-4712-97d4-219a71c57f3c

                      sold_at  qty uom  unit_price  price_total  \
32000000  1900-01-01 00:00:00 1.00  ea       16.28        16.28
32000001  1900-01-01 00:00:00 1.00  ea        3.62         3.62
32000002  1900-01-01 00:00:00 0.00  ea       27.20         0.00
32000003  1900-01-01 00:00:00 0.00  ea       27.20         0.00
32000004  1900-01-01 00:00:00 1.00  ea       27.20        27.20

                                              name quantity
32000000           Fire Bros 2g Pineapple Pancakes        0
32000001          50Fold Preroll .5g Galactic Jack        0
32000002  Indigo Pro Cartridge Northern Lights .5g        0
32000003  Indigo Pro Cartridge Northern Lights .5g        0
32000004  Indigo Pro Cartridge Northern Lights .5g        0
```

Works



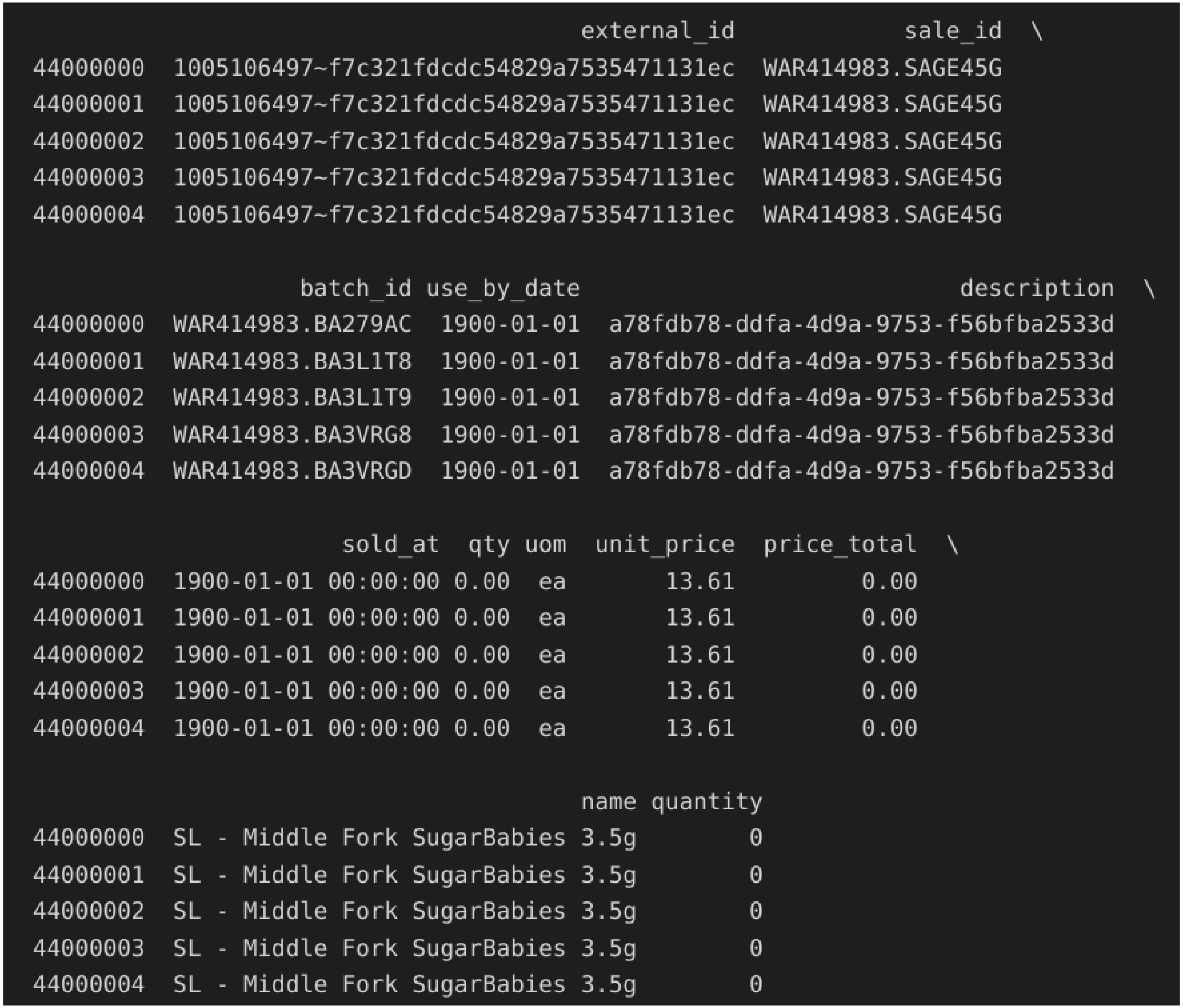
```
                   batch_id use_by_date                     description  \
42000000  WAR423885.BA6LLNI  1900-01-01                            Name   
42000001  WAR423885.BA6OL6Q  1900-01-01                            Name   
42000002  WAR423885.BA70PJ9  1900-01-01                            Name   
42000003  WAR423885.BA74535  1900-01-01                            Name   
42000004  WAR415130.BA75SK9  1900-01-01  Aurum Farms Gelato 33 (H) 3.5g   

                      sold_at  qty uom  unit_price  price_total  \
42000000  2018-08-14 17:00:00 1.00  ea        0.00         8.74   
42000001  2018-08-14 17:00:00 1.00  ea        0.00         8.74   
42000002  2018-08-14 17:00:00 1.00  ea        0.00        14.57   
42000003  2018-08-14 17:00:00 1.00  ea        0.00        17.48   
42000004  2018-08-14 17:00:00 1.00  ea       25.72        25.72   

                                    name quantity  
42000000                            Name        0  
42000001                            Name        0  
42000002                            Name        0  
42000003                            Name        0  
42000004  Aurum Farms Gelato 33 (H) 3.5g     3.5g  
```

Works



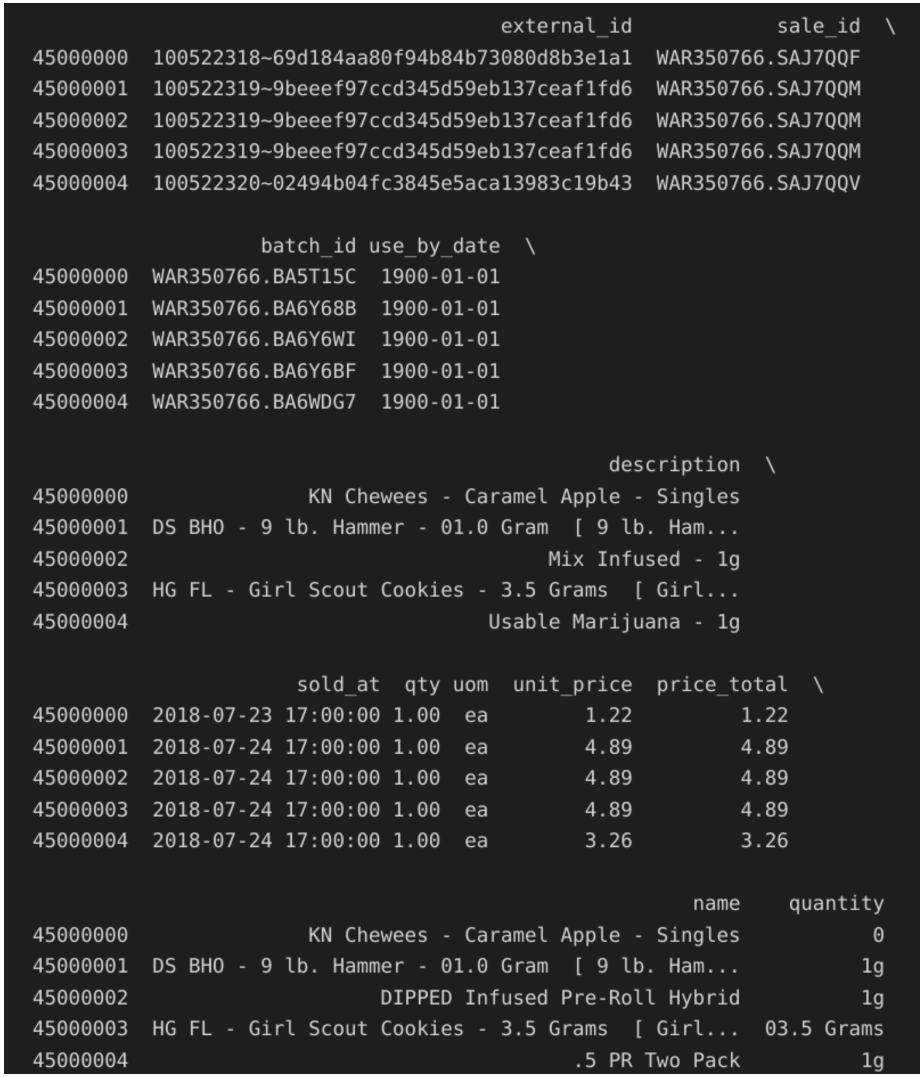
```
                    sale_id           batch_id use_by_date  \
43000000  WAJ413002.SAG2XP7  WAJ413002.BA6O9W5  1900-01-01
43000001  WAJ413002.SAG2XP7  WAJ413002.BA64ZAK  1900-01-01
43000002  WAJ413002.SAG2XP7  WAJ413002.BA66XZ4  1900-01-01
43000003  WAJ413002.SAG2XP7  WAJ413002.BA6F4MI  1900-01-01
43000004  WAJ413002.SAG2XP7  WAJ413002.BA762CC  1900-01-01

                     description              sold_at    qty uom  unit_price  \
43000000   Usable Marijuana - 7g  2018-08-21 17:00:00  20.00  ea        0.00
43000001   Usable Marijuana - 7g  2018-08-21 17:00:00  15.00  ea        0.00
43000002   Usable Marijuana - 7g  2018-08-21 17:00:00  20.00  ea        0.00
43000003  Usable Marijuana - 14g  2018-08-21 17:00:00 151.00  ea        0.00
43000004      Mix Packaged - 14g  2018-08-21 17:00:00  78.00  ea        0.00

          price_total            name quantity
43000000       140.00           Shake       7g
43000001       289.00  A Grade Flower       7g
43000002       140.00           Shake       7g
43000003      2748.00  B Grade Flower      14g
43000004      1420.00         B grade      14g
```

Works but… Description has garbage, name has the data



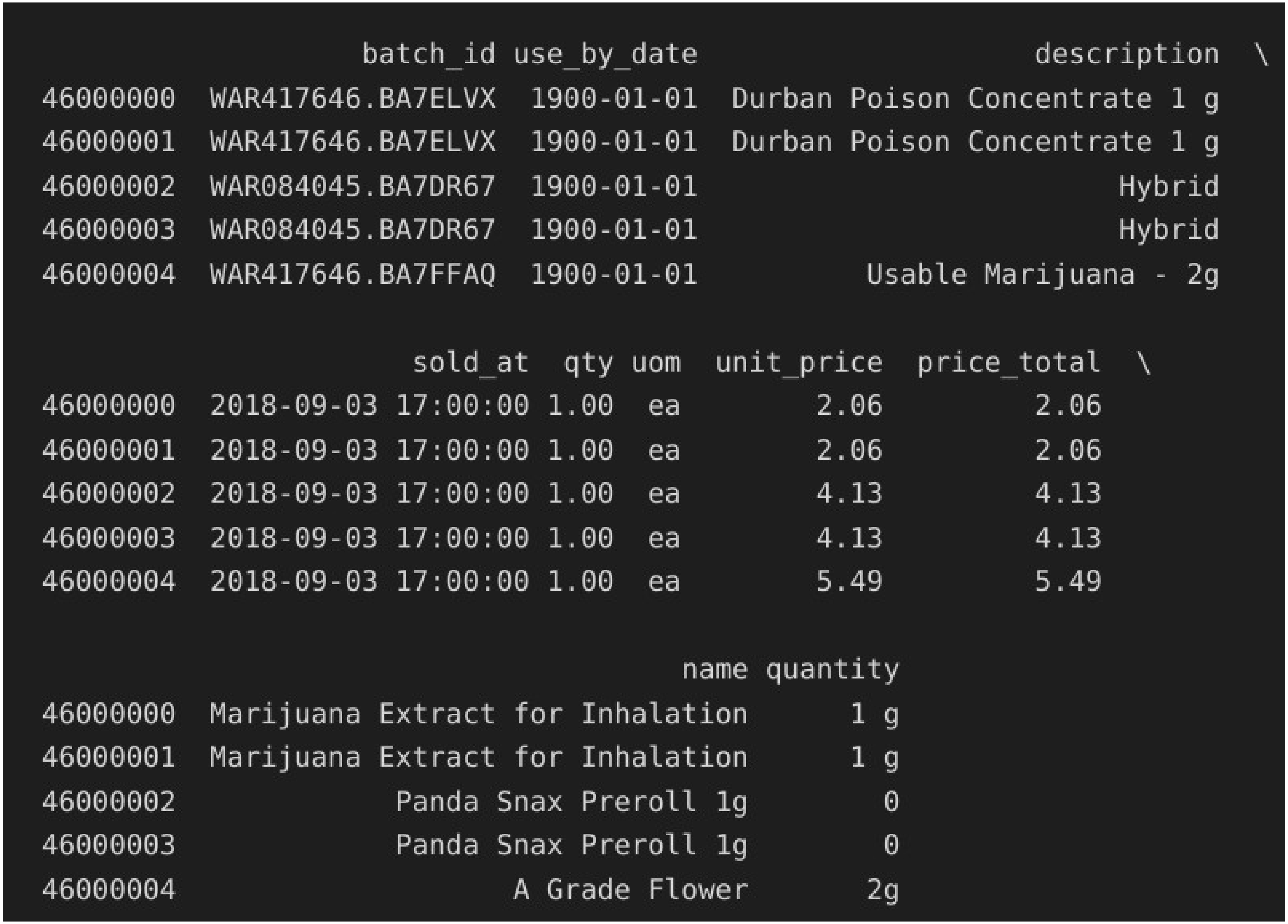
```
                                       external_id            sale_id  \
44000000  1005106497~f7c321fdcdc54829a7535471131ec  WAR414983.SAGE45G   
44000001  1005106497~f7c321fdcdc54829a7535471131ec  WAR414983.SAGE45G   
44000002  1005106497~f7c321fdcdc54829a7535471131ec  WAR414983.SAGE45G   
44000003  1005106497~f7c321fdcdc54829a7535471131ec  WAR414983.SAGE45G   
44000004  1005106497~f7c321fdcdc54829a7535471131ec  WAR414983.SAGE45G   

                   batch_id use_by_date                           description  \
44000000  WAR414983.BA279AC  1900-01-01  a78fdb78-ddfa-4d9a-9753-f56bfba2533d   
44000001  WAR414983.BA3L1T8  1900-01-01  a78fdb78-ddfa-4d9a-9753-f56bfba2533d   
44000002  WAR414983.BA3L1T9  1900-01-01  a78fdb78-ddfa-4d9a-9753-f56bfba2533d   
44000003  WAR414983.BA3VRG8  1900-01-01  a78fdb78-ddfa-4d9a-9753-f56bfba2533d   
44000004  WAR414983.BA3VRGD  1900-01-01  a78fdb78-ddfa-4d9a-9753-f56bfba2533d   

                      sold_at  qty uom  unit_price  price_total  \
44000000  1900-01-01 00:00:00 0.00  ea       13.61         0.00   
44000001  1900-01-01 00:00:00 0.00  ea       13.61         0.00   
44000002  1900-01-01 00:00:00 0.00  ea       13.61         0.00   
44000003  1900-01-01 00:00:00 0.00  ea       13.61         0.00   
44000004  1900-01-01 00:00:00 0.00  ea       13.61         0.00   

                                       name quantity  
44000000  SL - Middle Fork SugarBabies 3.5g        0  
44000001  SL - Middle Fork SugarBabies 3.5g        0  
44000002  SL - Middle Fork SugarBabies 3.5g        0  
44000003  SL - Middle Fork SugarBabies 3.5g        0  
44000004  SL - Middle Fork SugarBabies 3.5g        0  
```

Works



```
                                       external_id           sale_id  \
45000000  100522318~69d184aa80f94b84b73080d8b3e1a1  WAR350766.SAJ7QQF
45000001  100522319~9beeef97ccd345d59eb137ceaf1fd6  WAR350766.SAJ7QQM
45000002  100522319~9beeef97ccd345d59eb137ceaf1fd6  WAR350766.SAJ7QQM
45000003  100522319~9beeef97ccd345d59eb137ceaf1fd6  WAR350766.SAJ7QQM
45000004  100522320~02494b04fc3845e5aca13983c19b43  WAR350766.SAJ7QQV

                   batch_id use_by_date  \
45000000  WAR350766.BA5T15C  1900-01-01
45000001  WAR350766.BA6Y68B  1900-01-01
45000002  WAR350766.BA6Y6WI  1900-01-01
45000003  WAR350766.BA6Y6BF  1900-01-01
45000004  WAR350766.BA6WDG7  1900-01-01

                                                description  \
45000000               KN Chewees - Caramel Apple - Singles
45000001  DS BHO - 9 lb. Hammer - 01.0 Gram  [ 9 lb. Ham...
45000002                                   Mix Infused - 1g
45000003  HG FL - Girl Scout Cookies - 3.5 Grams  [ Girl...
45000004                              Usable Marijuana - 1g

                      sold_at  qty uom  unit_price  price_total  \
45000000  2018-07-23 17:00:00 1.00  ea        1.22         1.22
45000001  2018-07-24 17:00:00 1.00  ea        4.89         4.89
45000002  2018-07-24 17:00:00 1.00  ea        4.89         4.89
45000003  2018-07-24 17:00:00 1.00  ea        4.89         4.89
45000004  2018-07-24 17:00:00 1.00  ea        3.26         3.26

                                                       name    quantity
45000000               KN Chewees - Caramel Apple - Singles           0
45000001  DS BHO - 9 lb. Hammer - 01.0 Gram  [ 9 lb. Ham...          1g
45000002                     DIPPED Infused Pre-Roll Hybrid          1g
45000003  HG FL - Girl Scout Cookies - 3.5 Grams  [ Girl...  03.5 Grams
45000004                                     .5 PR Two Pack          1g
```

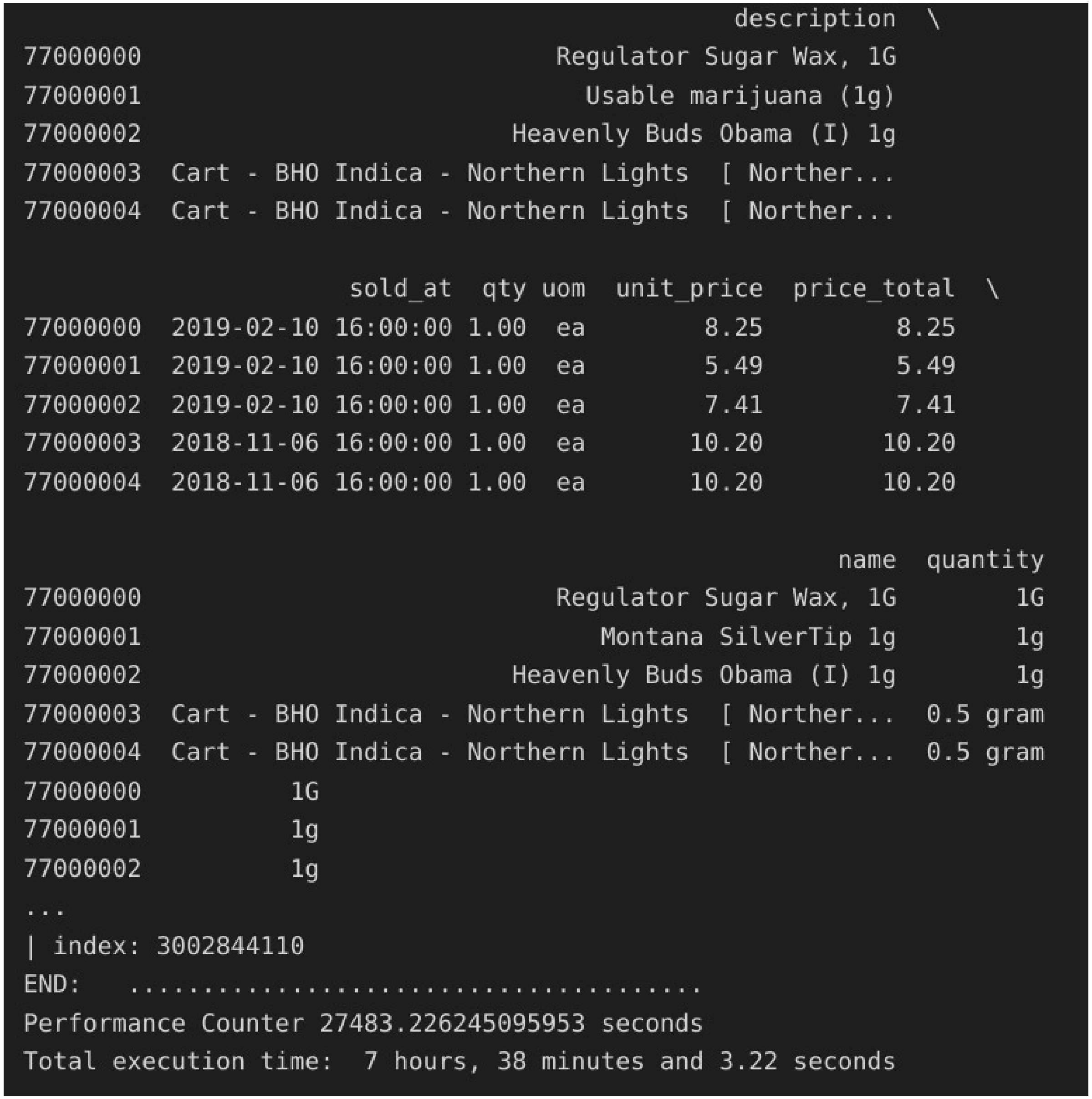
Works: but quantity data was in name on 2



```
                  batch_id use_by_date                     description  \
46000000  WAR417646.BA7ELVX  1900-01-01  Durban Poison Concentrate 1 g
46000001  WAR417646.BA7ELVX  1900-01-01  Durban Poison Concentrate 1 g
46000002  WAR084045.BA7DR67  1900-01-01                          Hybrid
46000003  WAR084045.BA7DR67  1900-01-01                          Hybrid
46000004  WAR417646.BA7FFAQ  1900-01-01          Usable Marijuana - 2g

                     sold_at  qty uom  unit_price  price_total  \
46000000  2018-09-03 17:00:00 1.00  ea        2.06         2.06
46000001  2018-09-03 17:00:00 1.00  ea        2.06         2.06
46000002  2018-09-03 17:00:00 1.00  ea        4.13         4.13
46000003  2018-09-03 17:00:00 1.00  ea        4.13         4.13
46000004  2018-09-03 17:00:00 1.00  ea        5.49         5.49

                                     name quantity
46000000  Marijuana Extract for Inhalation      1 g
46000001  Marijuana Extract for Inhalation      1 g
46000002               Panda Snax Preroll 1g        0
46000003               Panda Snax Preroll 1g        0
46000004                     A Grade Flower       2g
```

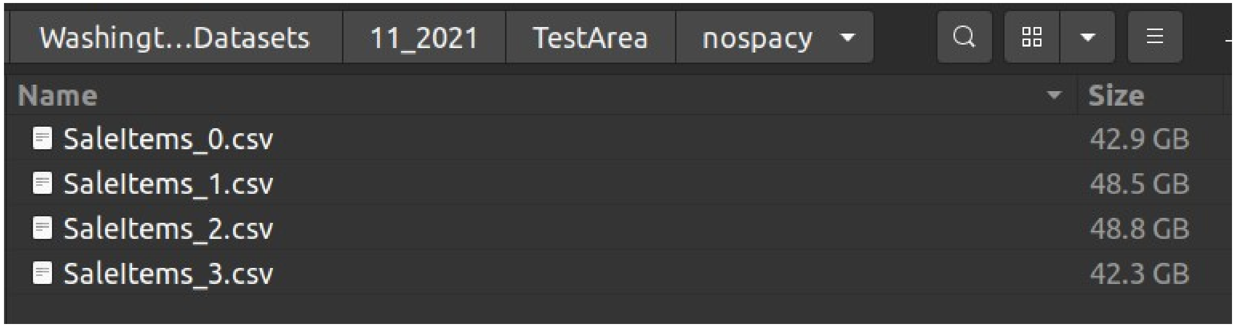
Run time: 7 hours 38 minutes

(**HP Omen 2080s 32GB ram Ubuntu 20-04.1 env Conda Rapids 22-02 Python 3.8.1 11TB USB drive)**

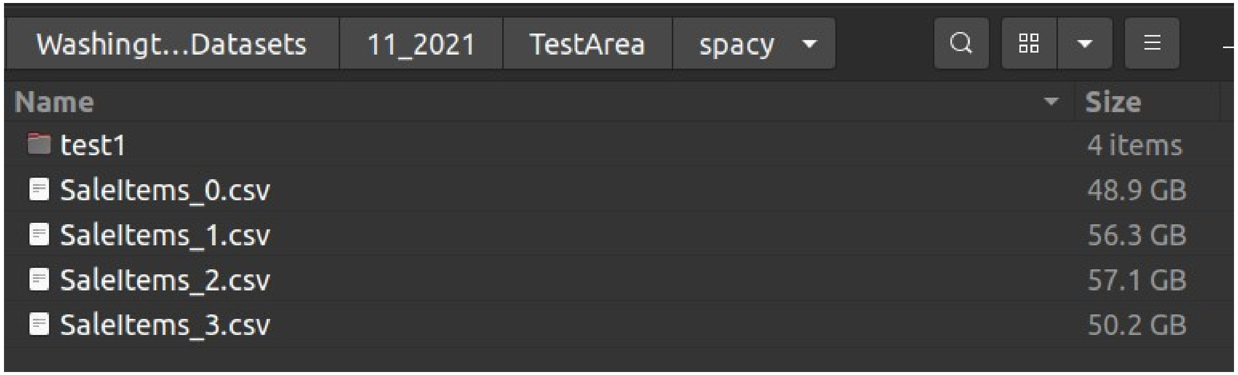


```
                                                description  \
77000000                            Regulator Sugar Wax, 1G
77000001                              Usable marijuana (1g)
77000002                         Heavenly Buds Obama (I) 1g
77000003  Cart - BHO Indica - Northern Lights  [ Norther...
77000004  Cart - BHO Indica - Northern Lights  [ Norther...

                      sold_at  qty uom  unit_price  price_total  \
77000000  2019-02-10 16:00:00 1.00  ea        8.25         8.25
77000001  2019-02-10 16:00:00 1.00  ea        5.49         5.49
77000002  2019-02-10 16:00:00 1.00  ea        7.41         7.41
77000003  2018-11-06 16:00:00 1.00  ea       10.20        10.20
77000004  2018-11-06 16:00:00 1.00  ea       10.20        10.20

                                                        name  quantity
77000000                            Regulator Sugar Wax, 1G        1G
77000001                                Montana SilverTip 1g       1g
77000002                         Heavenly Buds Obama (I) 1g        1g
77000003  Cart - BHO Indica - Northern Lights  [ Norther...  0.5 gram
77000004  Cart - BHO Indica - Northern Lights  [ Norther...  0.5 gram
77000000          1G
77000001          1g
77000002          1g
...
| index: 3002844110
END:   .......................................
Performance Counter 27483.226245095953 seconds
Total execution time:  7 hours, 38 minutes and 3.22 seconds
```

File sizes spacy vs. nospacy



Name	Size
SaleItems_0.csv	42.9 GB
SaleItems_1.csv	48.5 GB
SaleItems_2.csv	48.8 GB
SaleItems_3.csv	42.3 GB



Name	Size
test1	4 items
SaleItems_0.csv	48.9 GB
SaleItems_1.csv	56.3 GB
SaleItems_2.csv	57.1 GB
SaleItems_3.csv	50.2 GB

This is to be expected since we are adding quantity column and filling df.name.isna() and/or df.description.isna() with “Name”

To Do:

Fabulous Suggestions:
Gerry Pallor - add plot that shows NaN percentages
Charles Rice - write to parquet file format