Users_Quality

May 29, 2024

1 Import Libraries

[1]:
```
import pandas as pd
import numpy as np
import json
```

2 Convert JSON into DataFrame Object

[2]:
```
users = pd.read_json('users.json', lines=True)
```

3 Get Overview of Data

[3]:
```
users.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 495 entries, 0 to 494
Data columns (total 7 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   _id           495 non-null    object
 1   active        495 non-null    bool
 2   createdDate   495 non-null    object
 3   lastLogin     433 non-null    object
 4   role          495 non-null    object
 5   signUpSource  447 non-null    object
 6   state         439 non-null    object
dtypes: bool(1), object(6)
memory usage: 23.8+ KB
```

[4]:
```
users.head()
```

[4]:
```
                                   _id  active                  createdDate  \
0  {'$oid': '5ff1e194b6a9d73a3a9f1052'}    True  {'$date': 1609687444800}
1  {'$oid': '5ff1e194b6a9d73a3a9f1052'}    True  {'$date': 1609687444800}
2  {'$oid': '5ff1e194b6a9d73a3a9f1052'}    True  {'$date': 1609687444800}
3  {'$oid': '5ff1e1eacfcf6c399c274ae6'}    True  {'$date': 1609687530554}
```

```
4  {'$oid': '5ff1e194b6a9d73a3a9f1052'}    True  {'$date': 1609687444800}

                   lastLogin      role signUpSource state
0  {'$date': 1609687537858}  consumer        Email    WI
1  {'$date': 1609687537858}  consumer        Email    WI
2  {'$date': 1609687537858}  consumer        Email    WI
3  {'$date': 1609687530597}  consumer        Email    WI
4  {'$date': 1609687537858}  consumer        Email    WI
```

4 Evaluation

4.1 Null/NaN/Missing Values

```
[5]: users.isnull().sum()
```

```
[5]: _id              0
     active           0
     createdDate      0
     lastLogin       62
     role             0
     signUpSource    48
     state           56
     dtype: int64
```

Very few user rows contain null values, good.

4.1.1 Last Login Null records

```
[6]: nulls = users.loc[users['lastLogin'].isnull()]
     nulls
```

```
[6]:                                       _id  active               createdDate  \
     97   {'$oid': '5ff616a68f142f11dd189163'}    True  {'$date': 1609963174996}
     143  {'$oid': '5ffe115404929101d0aaebb2'}    True  {'$date': 1610486100208}
     148  {'$oid': '5ffe115404929101d0aaebb2'}    True  {'$date': 1610486100208}
     170  {'$oid': '5e27526d0bdb6a138c32b556'}    True  {'$date': 1579635309795}
     180  {'$oid': '6002475cfb296c121a81b98d'}    True  {'$date': 1610762076571}
     ..                                      …     …                        …
     381  {'$oid': '60186237c8b50e11d8454d5f'}    True  {'$date': 1612210743551}
     382  {'$oid': '60186237c8b50e11d8454d5f'}    True  {'$date': 1612210743551}
     389  {'$oid': '60217fa799409b11fcf899fe'}    True  {'$date': 1612808103714}
     420  {'$oid': '5fb0a078be5fc9775c1f3945'}    True  {'$date': 1605410936818}
     429  {'$oid': '5fb0a078be5fc9775c1f3945'}    True  {'$date': 1605410936818}

         lastLogin      role signUpSource state
     97        NaN  consumer        Email    KY
     143       NaN  consumer        Email    AL
```

```
148        NaN  consumer        Email    AL
170        NaN  consumer       Google    WI
180        NaN  consumer        Email    WI
..         …        …        …   …
381        NaN  consumer        Email   NaN
382        NaN  consumer        Email   NaN
389        NaN  consumer        Email    WI
420        NaN  consumer       Google    AL
429        NaN  consumer       Google    AL

[62 rows x 7 columns]
```

4.1.2 signUpSource Null records

[7]:

```
nulls = users.loc[users['signUpSource'].isnull()]
nulls
```

[7]:

```
                                      _id  active              createdDate  \
388  {'$oid': '55308179e4b0eabd8f99caa2'}    True  {'$date': 1429242233186}
395  {'$oid': '59c124bae4b0299e55b0f330'}    True  {'$date': 1505830074302}
396  {'$oid': '59c124bae4b0299e55b0f330'}    True  {'$date': 1505830074302}
397  {'$oid': '59c124bae4b0299e55b0f330'}    True  {'$date': 1505830074302}
398  {'$oid': '59c124bae4b0299e55b0f330'}    True  {'$date': 1505830074302}
399  {'$oid': '59c124bae4b0299e55b0f330'}    True  {'$date': 1505830074302}
400  {'$oid': '59c124bae4b0299e55b0f330'}    True  {'$date': 1505830074302}
401  {'$oid': '59c124bae4b0299e55b0f330'}    True  {'$date': 1505830074302}
402  {'$oid': '59c124bae4b0299e55b0f330'}    True  {'$date': 1505830074302}
403  {'$oid': '59c124bae4b0299e55b0f330'}    True  {'$date': 1505830074302}
405  {'$oid': '59c124bae4b0299e55b0f330'}    True  {'$date': 1505830074302}
406  {'$oid': '59c124bae4b0299e55b0f330'}    True  {'$date': 1505830074302}
407  {'$oid': '59c124bae4b0299e55b0f330'}    True  {'$date': 1505830074302}
409  {'$oid': '59c124bae4b0299e55b0f330'}    True  {'$date': 1505830074302}
410  {'$oid': '59c124bae4b0299e55b0f330'}    True  {'$date': 1505830074302}
411  {'$oid': '59c124bae4b0299e55b0f330'}    True  {'$date': 1505830074302}
412  {'$oid': '59c124bae4b0299e55b0f330'}    True  {'$date': 1505830074302}
413  {'$oid': '59c124bae4b0299e55b0f330'}    True  {'$date': 1505830074302}
414  {'$oid': '59c124bae4b0299e55b0f330'}    True  {'$date': 1505830074302}
422  {'$oid': '5a43c08fe4b014fd6b6a0612'}    True  {'$date': 1514389647059}
423  {'$oid': '5a43c08fe4b014fd6b6a0612'}    True  {'$date': 1514389647059}
424  {'$oid': '5a43c08fe4b014fd6b6a0612'}    True  {'$date': 1514389647059}
425  {'$oid': '5a43c08fe4b014fd6b6a0612'}    True  {'$date': 1514389647059}
426  {'$oid': '5a43c08fe4b014fd6b6a0612'}    True  {'$date': 1514389647059}
428  {'$oid': '5a43c08fe4b014fd6b6a0612'}    True  {'$date': 1514389647059}
430  {'$oid': '5a43c08fe4b014fd6b6a0612'}    True  {'$date': 1514389647059}
431  {'$oid': '5a43c08fe4b014fd6b6a0612'}    True  {'$date': 1514389647059}
462  {'$oid': '5964eb07e4b03efd0c0f267b'}    True  {'$date': 1499785991771}
475  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
```

```
476  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
477  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
478  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
479  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
480  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
481  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
482  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
483  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
484  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
485  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
486  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
487  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
488  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
489  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
490  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
491  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
492  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
493  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
494  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}

                    lastLogin         role signUpSource state
388  {'$date': 1525713820003}     consumer          NaN    WI
395  {'$date': 1612802578117}  fetch-staff          NaN    WI
396  {'$date': 1612802578117}  fetch-staff          NaN    WI
397  {'$date': 1612802578117}  fetch-staff          NaN    WI
398  {'$date': 1612802578117}  fetch-staff          NaN    WI
399  {'$date': 1612802578117}  fetch-staff          NaN    WI
400  {'$date': 1612802578117}  fetch-staff          NaN    WI
401  {'$date': 1612802578117}  fetch-staff          NaN    WI
402  {'$date': 1612802578117}  fetch-staff          NaN    WI
403  {'$date': 1612802578117}  fetch-staff          NaN    WI
405  {'$date': 1612802578117}  fetch-staff          NaN    WI
406  {'$date': 1612802578117}  fetch-staff          NaN    WI
407  {'$date': 1612802578117}  fetch-staff          NaN    WI
409  {'$date': 1612802578117}  fetch-staff          NaN    WI
410  {'$date': 1612802578117}  fetch-staff          NaN    WI
411  {'$date': 1612802578117}  fetch-staff          NaN    WI
412  {'$date': 1612802578117}  fetch-staff          NaN    WI
413  {'$date': 1612802578117}  fetch-staff          NaN    WI
414  {'$date': 1612802578117}  fetch-staff          NaN    WI
422  {'$date': 1613146957155}     consumer          NaN   NaN
423  {'$date': 1613146957155}     consumer          NaN   NaN
424  {'$date': 1613146957155}     consumer          NaN   NaN
425  {'$date': 1613146957155}     consumer          NaN   NaN
426  {'$date': 1613146957155}     consumer          NaN   NaN
428  {'$date': 1613146957155}     consumer          NaN   NaN
430  {'$date': 1613146957155}     consumer          NaN   NaN
```

```
431  {'$date': 1613146957155}     consumer          NaN   NaN
462  {'$date': 1614884869770}  fetch-staff          NaN    IL
475  {'$date': 1614963143204}  fetch-staff          NaN   NaN
476  {'$date': 1614963143204}  fetch-staff          NaN   NaN
477  {'$date': 1614963143204}  fetch-staff          NaN   NaN
478  {'$date': 1614963143204}  fetch-staff          NaN   NaN
479  {'$date': 1614963143204}  fetch-staff          NaN   NaN
480  {'$date': 1614963143204}  fetch-staff          NaN   NaN
481  {'$date': 1614963143204}  fetch-staff          NaN   NaN
482  {'$date': 1614963143204}  fetch-staff          NaN   NaN
483  {'$date': 1614963143204}  fetch-staff          NaN   NaN
484  {'$date': 1614963143204}  fetch-staff          NaN   NaN
485  {'$date': 1614963143204}  fetch-staff          NaN   NaN
486  {'$date': 1614963143204}  fetch-staff          NaN   NaN
487  {'$date': 1614963143204}  fetch-staff          NaN   NaN
488  {'$date': 1614963143204}  fetch-staff          NaN   NaN
489  {'$date': 1614963143204}  fetch-staff          NaN   NaN
490  {'$date': 1614963143204}  fetch-staff          NaN   NaN
491  {'$date': 1614963143204}  fetch-staff          NaN   NaN
492  {'$date': 1614963143204}  fetch-staff          NaN   NaN
493  {'$date': 1614963143204}  fetch-staff          NaN   NaN
494  {'$date': 1614963143204}  fetch-staff          NaN   NaN
```

Seems like the missing signup source rows are more often fetch staff, I assume test data.

What's the percentage of fetch-staff roles for all null signup source?

[8]:
```python
# 48 is the number of missing signUpSource rows
print((nulls['role'].value_counts() / 48) * 100.0)
```

```
role
fetch-staff    81.25
consumer       18.75
Name: count, dtype: float64
```

Far more fetch-staff roles for the missing data, let's check the missing states too

4.1.3 State missing records

[9]:
```python
nulls = users.loc[users['state'].isnull()]
nulls
```

[9]:
```
                                       _id  active               createdDate  \
344  {'$oid': '60145ff384231211ce796d51'}    True  {'$date': 1611948019722}
350  {'$oid': '60145ff384231211ce796d51'}    True  {'$date': 1611948019722}
375  {'$oid': '60186237c8b50e11d8454d5f'}    True  {'$date': 1612210743551}
376  {'$oid': '60186237c8b50e11d8454d5f'}    True  {'$date': 1612210743551}
378  {'$oid': '60186237c8b50e11d8454d5f'}    True  {'$date': 1612210743551}
381  {'$oid': '60186237c8b50e11d8454d5f'}    True  {'$date': 1612210743551}
```

```
382  {'$oid': '60186237c8b50e11d8454d5f'}    True  {'$date': 1612210743551}
422  {'$oid': '5a43c08fe4b014fd6b6a0612'}    True  {'$date': 1514389647059}
423  {'$oid': '5a43c08fe4b014fd6b6a0612'}    True  {'$date': 1514389647059}
424  {'$oid': '5a43c08fe4b014fd6b6a0612'}    True  {'$date': 1514389647059}
425  {'$oid': '5a43c08fe4b014fd6b6a0612'}    True  {'$date': 1514389647059}
426  {'$oid': '5a43c08fe4b014fd6b6a0612'}    True  {'$date': 1514389647059}
428  {'$oid': '5a43c08fe4b014fd6b6a0612'}    True  {'$date': 1514389647059}
430  {'$oid': '5a43c08fe4b014fd6b6a0612'}    True  {'$date': 1514389647059}
431  {'$oid': '5a43c08fe4b014fd6b6a0612'}    True  {'$date': 1514389647059}
432  {'$oid': '5fbc35711d967d1222cbfefc'}    True  {'$date': 1606169969509}
433  {'$oid': '5fbc35711d967d1222cbfefc'}    True  {'$date': 1606169969509}
434  {'$oid': '5fbc35711d967d1222cbfefc'}    True  {'$date': 1606169969509}
455  {'$oid': '5fa41775898c7a11a6bcef3e'}    True  {'$date': 1604589429396}
457  {'$oid': '5fa41775898c7a11a6bcef3e'}    True  {'$date': 1604589429396}
458  {'$oid': '5fa41775898c7a11a6bcef3e'}    True  {'$date': 1604589429396}
459  {'$oid': '5fa41775898c7a11a6bcef3e'}    True  {'$date': 1604589429396}
460  {'$oid': '5fa41775898c7a11a6bcef3e'}    True  {'$date': 1604589429396}
461  {'$oid': '5fa41775898c7a11a6bcef3e'}    True  {'$date': 1604589429396}
463  {'$oid': '5fa41775898c7a11a6bcef3e'}    True  {'$date': 1604589429396}
464  {'$oid': '5fa41775898c7a11a6bcef3e'}    True  {'$date': 1604589429396}
465  {'$oid': '5fa41775898c7a11a6bcef3e'}    True  {'$date': 1604589429396}
466  {'$oid': '5fa41775898c7a11a6bcef3e'}    True  {'$date': 1604589429396}
467  {'$oid': '5fa41775898c7a11a6bcef3e'}    True  {'$date': 1604589429396}
468  {'$oid': '5fa41775898c7a11a6bcef3e'}    True  {'$date': 1604589429396}
469  {'$oid': '5fa41775898c7a11a6bcef3e'}    True  {'$date': 1604589429396}
470  {'$oid': '5fa41775898c7a11a6bcef3e'}    True  {'$date': 1604589429396}
471  {'$oid': '5fa41775898c7a11a6bcef3e'}    True  {'$date': 1604589429396}
472  {'$oid': '5fa41775898c7a11a6bcef3e'}    True  {'$date': 1604589429396}
473  {'$oid': '5fa41775898c7a11a6bcef3e'}    True  {'$date': 1604589429396}
474  {'$oid': '5fa41775898c7a11a6bcef3e'}    True  {'$date': 1604589429396}
475  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
476  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
477  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
478  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
479  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
480  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
481  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
482  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
483  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
484  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
485  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
486  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
487  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
488  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
489  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
490  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
491  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
```

```
492  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
493  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
494  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}

                    lastLogin         role signUpSource state
344                       NaN     consumer        Email   NaN
350                       NaN     consumer        Email   NaN
375                       NaN     consumer        Email   NaN
376                       NaN     consumer        Email   NaN
378                       NaN     consumer        Email   NaN
381                       NaN     consumer        Email   NaN
382                       NaN     consumer        Email   NaN
422  {'$date': 1613146957155}     consumer          NaN   NaN
423  {'$date': 1613146957155}     consumer          NaN   NaN
424  {'$date': 1613146957155}     consumer          NaN   NaN
425  {'$date': 1613146957155}     consumer          NaN   NaN
426  {'$date': 1613146957155}     consumer          NaN   NaN
428  {'$date': 1613146957155}     consumer          NaN   NaN
430  {'$date': 1613146957155}     consumer          NaN   NaN
431  {'$date': 1613146957155}     consumer          NaN   NaN
432  {'$date': 1614313551057}  fetch-staff        Email   NaN
433  {'$date': 1614313551057}  fetch-staff        Email   NaN
434  {'$date': 1614313551057}  fetch-staff        Email   NaN
455  {'$date': 1614873722026}  fetch-staff        Email   NaN
457  {'$date': 1614873722026}  fetch-staff        Email   NaN
458  {'$date': 1614873722026}  fetch-staff        Email   NaN
459  {'$date': 1614873722026}  fetch-staff        Email   NaN
460  {'$date': 1614873722026}  fetch-staff        Email   NaN
461  {'$date': 1614873722026}  fetch-staff        Email   NaN
463  {'$date': 1614873722026}  fetch-staff        Email   NaN
464  {'$date': 1614873722026}  fetch-staff        Email   NaN
465  {'$date': 1614873722026}  fetch-staff        Email   NaN
466  {'$date': 1614873722026}  fetch-staff        Email   NaN
467  {'$date': 1614873722026}  fetch-staff        Email   NaN
468  {'$date': 1614873722026}  fetch-staff        Email   NaN
469  {'$date': 1614873722026}  fetch-staff        Email   NaN
470  {'$date': 1614873722026}  fetch-staff        Email   NaN
471  {'$date': 1614873722026}  fetch-staff        Email   NaN
472  {'$date': 1614873722026}  fetch-staff        Email   NaN
473  {'$date': 1614873722026}  fetch-staff        Email   NaN
474  {'$date': 1614873722026}  fetch-staff        Email   NaN
475  {'$date': 1614963143204}  fetch-staff          NaN   NaN
476  {'$date': 1614963143204}  fetch-staff          NaN   NaN
477  {'$date': 1614963143204}  fetch-staff          NaN   NaN
478  {'$date': 1614963143204}  fetch-staff          NaN   NaN
479  {'$date': 1614963143204}  fetch-staff          NaN   NaN
480  {'$date': 1614963143204}  fetch-staff          NaN   NaN
```

```
481  {'$date': 1614963143204}  fetch-staff         NaN   NaN
482  {'$date': 1614963143204}  fetch-staff         NaN   NaN
483  {'$date': 1614963143204}  fetch-staff         NaN   NaN
484  {'$date': 1614963143204}  fetch-staff         NaN   NaN
485  {'$date': 1614963143204}  fetch-staff         NaN   NaN
486  {'$date': 1614963143204}  fetch-staff         NaN   NaN
487  {'$date': 1614963143204}  fetch-staff         NaN   NaN
488  {'$date': 1614963143204}  fetch-staff         NaN   NaN
489  {'$date': 1614963143204}  fetch-staff         NaN   NaN
490  {'$date': 1614963143204}  fetch-staff         NaN   NaN
491  {'$date': 1614963143204}  fetch-staff         NaN   NaN
492  {'$date': 1614963143204}  fetch-staff         NaN   NaN
493  {'$date': 1614963143204}  fetch-staff         NaN   NaN
494  {'$date': 1614963143204}  fetch-staff         NaN   NaN
```

Let's compare the fetch-staff to consumers again for state missing data

[10]:
```python
# 56 is the number of missing signUpSource rows
print((nulls['role'].value_counts() / 56 * 100.0))
```

```
role
fetch-staff    73.214286
consumer       26.785714
Name: count, dtype: float64
```

Far more roles are fetch-staff for the missing data again. This encourages the test-data hypothesis

4.2 Unique Values

[11]:
```python
users_json = users.copy()
```

[12]:
```python
# Convert dictionary columns to JSON strings to allow nunique() to run
users_json._id = users_json._id.apply(json.dumps)
users_json.createdDate = users_json.createdDate.apply(json.dumps)
users_json.lastLogin = users_json.lastLogin.apply(json.dumps)
```

[13]:
```python
users_json.nunique()
```

[13]:
```
_id            212
active           2
createdDate    212
lastLogin      173
role             2
signUpSource     2
state            8
dtype: int64
```

Big issue, there's no unique identifier in this dataset. Which _ids are the most often occuring?

[14]:
```
users['_id'].value_counts()
```

[14]:
```
_id
{'$oid': '54943462e4b07e684157a532'}    20
{'$oid': '5fc961c3b8cfca11a077dd33'}    20
{'$oid': '5ff5d15aeb7c7d12096d91a2'}    18
{'$oid': '5fa41775898c7a11a6bcef3e'}    18
{'$oid': '59c124bae4b0299e55b0f330'}    18
                                        ..
{'$oid': '6004a965e257124ec6b9a39f'}     1
{'$oid': '600746fd6e64691717e8cfb5'}     1
{'$oid': '60074b996e64691717e8f11a'}     1
{'$oid': '60074246325c8a12289e22a0'}     1
{'$oid': '60088e5d633aab121bb8e5cf'}     1
Name: count, Length: 212, dtype: int64
```

[15]:
```
# Do the most-repeated _ids correlate to the null data?
users.loc[(users['_id'] == {'$oid': '54943462e4b07e684157a532'} )]
```

[15]:
```
                                      _id  active                createdDate  \
475  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
476  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
477  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
478  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
479  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
480  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
481  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
482  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
483  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
484  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
485  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
486  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
487  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
488  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
489  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
490  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
491  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
492  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
493  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}
494  {'$oid': '54943462e4b07e684157a532'}    True  {'$date': 1418998882381}

                    lastLogin         role signUpSource state
475  {'$date': 1614963143204}  fetch-staff          NaN   NaN
476  {'$date': 1614963143204}  fetch-staff          NaN   NaN
477  {'$date': 1614963143204}  fetch-staff          NaN   NaN
478  {'$date': 1614963143204}  fetch-staff          NaN   NaN
479  {'$date': 1614963143204}  fetch-staff          NaN   NaN
```

```
480  {'$date': 1614963143204}  fetch-staff          NaN   NaN
481  {'$date': 1614963143204}  fetch-staff          NaN   NaN
482  {'$date': 1614963143204}  fetch-staff          NaN   NaN
483  {'$date': 1614963143204}  fetch-staff          NaN   NaN
484  {'$date': 1614963143204}  fetch-staff          NaN   NaN
485  {'$date': 1614963143204}  fetch-staff          NaN   NaN
486  {'$date': 1614963143204}  fetch-staff          NaN   NaN
487  {'$date': 1614963143204}  fetch-staff          NaN   NaN
488  {'$date': 1614963143204}  fetch-staff          NaN   NaN
489  {'$date': 1614963143204}  fetch-staff          NaN   NaN
490  {'$date': 1614963143204}  fetch-staff          NaN   NaN
491  {'$date': 1614963143204}  fetch-staff          NaN   NaN
492  {'$date': 1614963143204}  fetch-staff          NaN   NaN
493  {'$date': 1614963143204}  fetch-staff          NaN   NaN
494  {'$date': 1614963143204}  fetch-staff          NaN   NaN
```

They do!

5 Conclusions

Dataset Users is full of test data, not production data. The clues that give this away are:

1. Multiple records contain the same _id, which has been used as a uuid, a universally-unique identifier, in the other datasets.
2. The rows with the repeated _id has a higher correlation to null/missing data in the other rows, namely signUpSource & state.
3. Rows where role == ‘fetch-staff’ have a much higher correlation to null/missing signUpSources & states.