pid_to_info_all dataset

Basic Information:

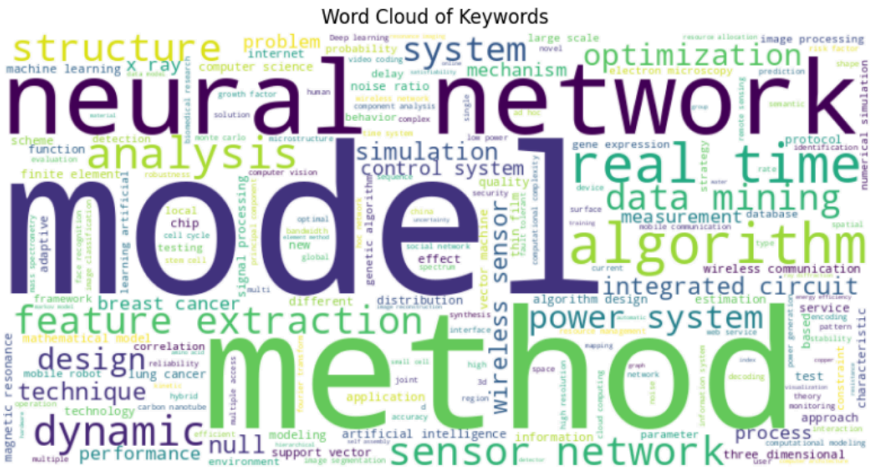
```
DataFrame Info:
<class 'pandas.core.frame.DataFrame'>
Index: 317302 entries, 6IsfnuWU to ebYfGt6j
Data columns (total 7 columns):
 #   Column    Non-Null Count   Dtype
---  ------    --------------   -----
 0   id        317302 non-null  object
 1   title     317302 non-null  object
 2   authors   317302 non-null  object
 3   abstract  317302 non-null  object
 4   keywords  317302 non-null  object
 5   venue     317301 non-null  object
 6   year      317302 non-null  object
dtypes: object(7)
```

```
Missing Values
id          0
title       0
authors     0
abstract    0
keywords    0
venue       1
year        0
dtype: int64
```

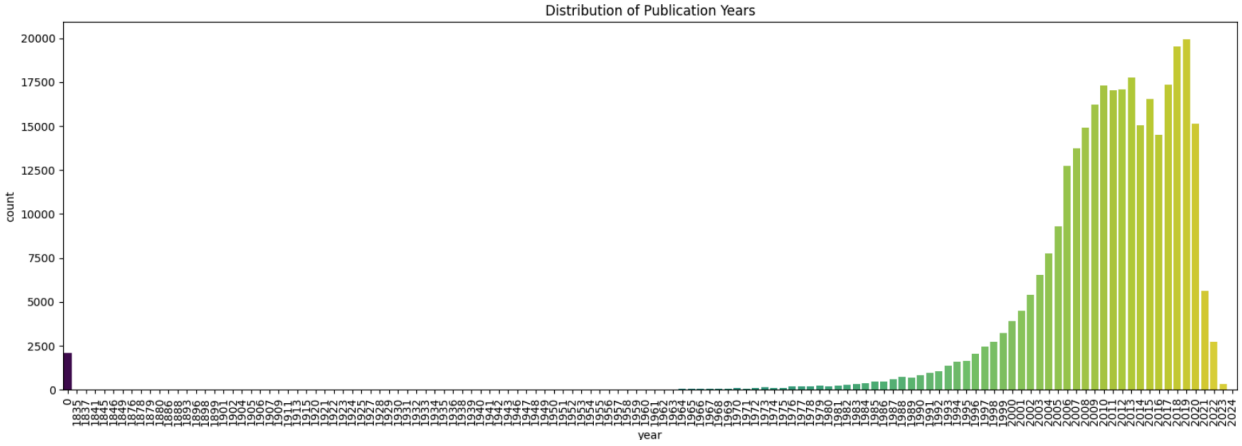
- venue has a null value

```
DataFrame Head:
                id                                              title  \
6IsfnuWU  6IsfnuWU  Probabilistic Skyline Operator over Sliding Wi...
8B8GhlnI  8B8GhlnI  Editorial: Knowledge-Driven Activity Recogniti...
4dZKGwVR  4dZKGwVR  Subscriber Assignment For Wide-Area Content-Ba...
V1JgT3OM  V1JgT3OM  Tree-Based Mining for Discovering Patterns of ...
HMvrPr2W  HMvrPr2W  Protein Function Prediction using Multi-label ...

                                                  authors  \
6IsfnuWU  [{'name': 'Wenjie Zhang', 'org': 'UNSW Sydney'...
8B8GhlnI  [{'name': 'Liming Chen', 'org': ''}, {'name': ...
4dZKGwVR  [{'name': 'Albert Yu', 'org': 'Duke Univ, Dept...
V1JgT3OM  [{'name': 'Zhiwen Yu', 'org': 'Northwestern Po...
HMvrPr2W  [{'name': 'Guoxian Yu', 'org': 'Southwest Univ...

                                                  abstract  \
6IsfnuWU  Skyline computation has many applications incl...
8B8GhlnI
4dZKGwVR  We study the problem of assigning subscribers ...
V1JgT3OM  AbstractDiscovering semantic knowledge is sign...
HMvrPr2W  AbstractHigh-throughput experimental technique...

                                                  keywords  \
6IsfnuWU  [continuous skyline query, probabilistic skyli...
8B8GhlnI                              [activity recognition]
4dZKGwVR  [Monte Carlo approximation algorithm, future e...
V1JgT3OM  [discovering patterns, interaction flow patter...
HMvrPr2W  [heterogeneous proteomic data sets, multilabel...

                                                     venue  year
6IsfnuWU  ICDE '09 Proceedings of the 2009 IEEE Internat...  2009
8B8GhlnI                                       Periodicals  2011
4dZKGwVR  ICDE '11 Proceedings of the 2011 IEEE 27th Int...  2011
V1JgT3OM                                       Periodicals  2012
HMvrPr2W  IEEE/ACM Transactions on Computational Biology...  2013
```

Top Keywords:

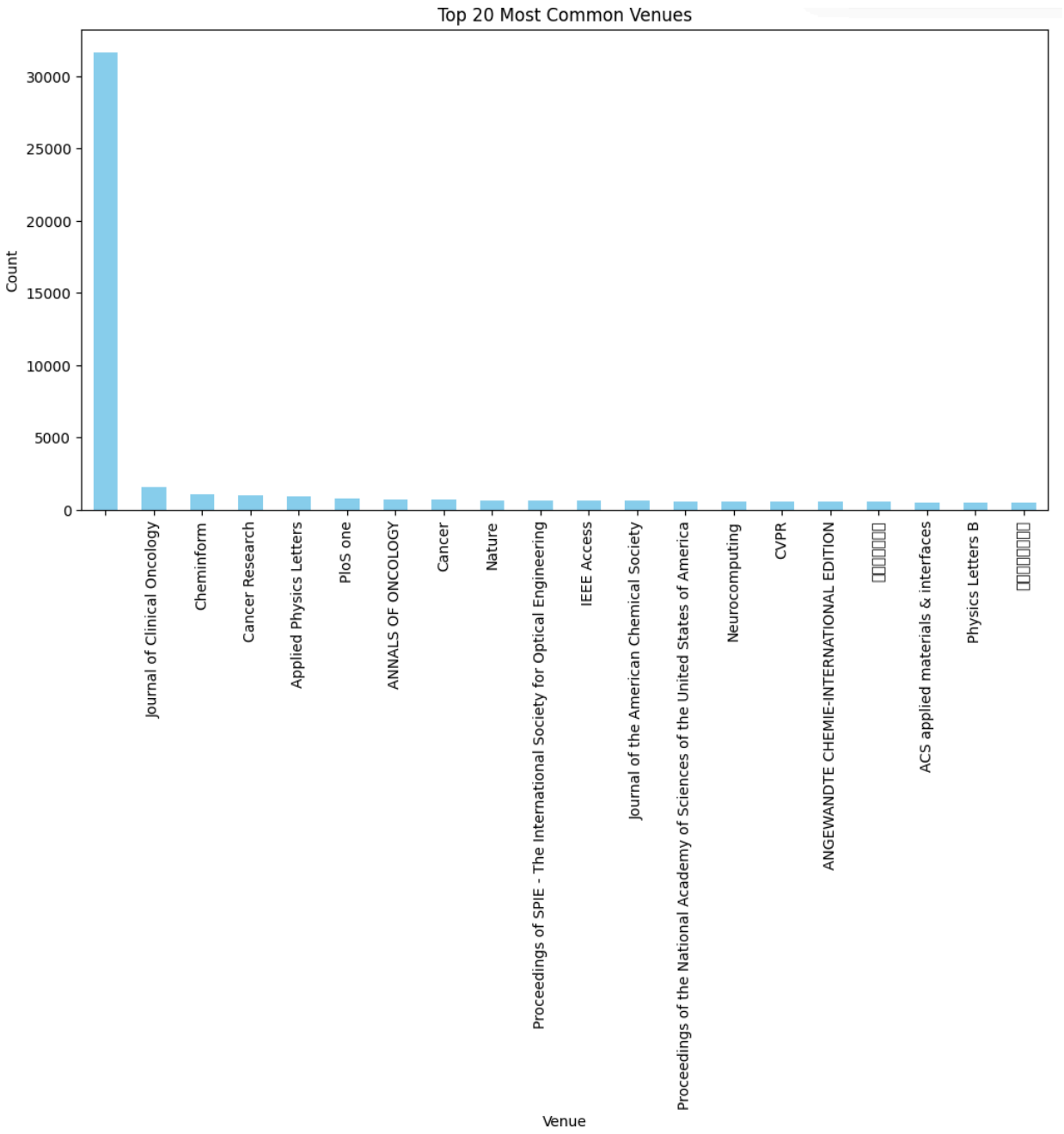


Distribution over the years:



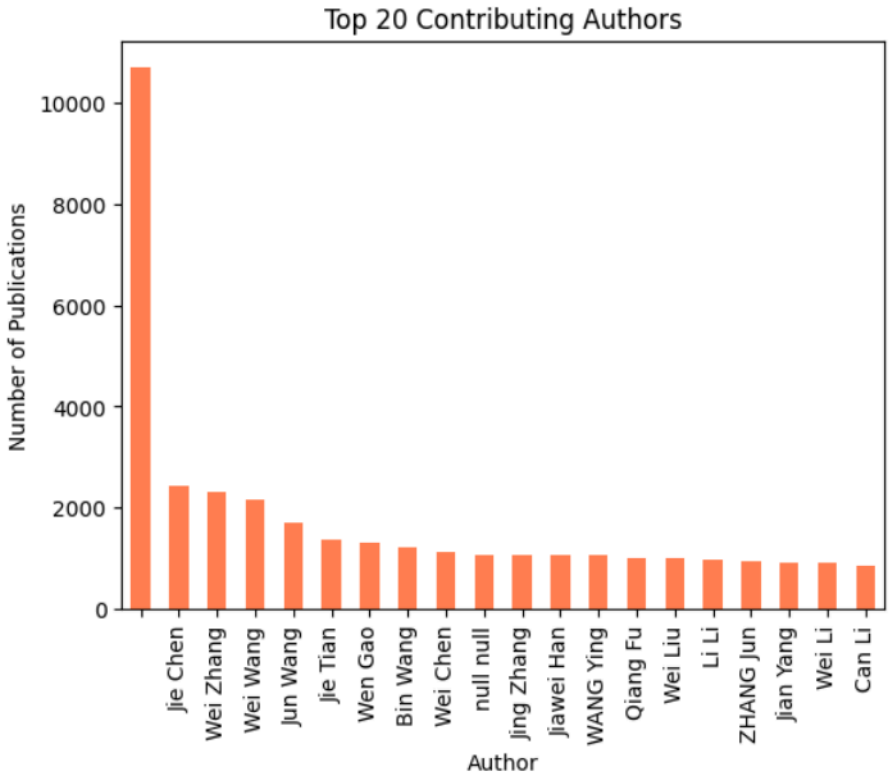
- Some values are around 0 => some data points don't have the year mentioned
- Majority are after 1990s
- Use case: Can use the years to determine wrongly assigned papers based comparing with the topic that author is writing about around that period?

Top 20 Venues



- Some weird values here (e.g. the rightmost venue), not sure if it's because we're reading a json file or the way it's formatted?
- Most of the values have no venue, could be due to the way the file is formatted so need to look into how to read the json file so that this issue won't occur
- Useful case: Compare the venue with topic/abstract/keywords and remove those papers that don't match

Top 20 Authors

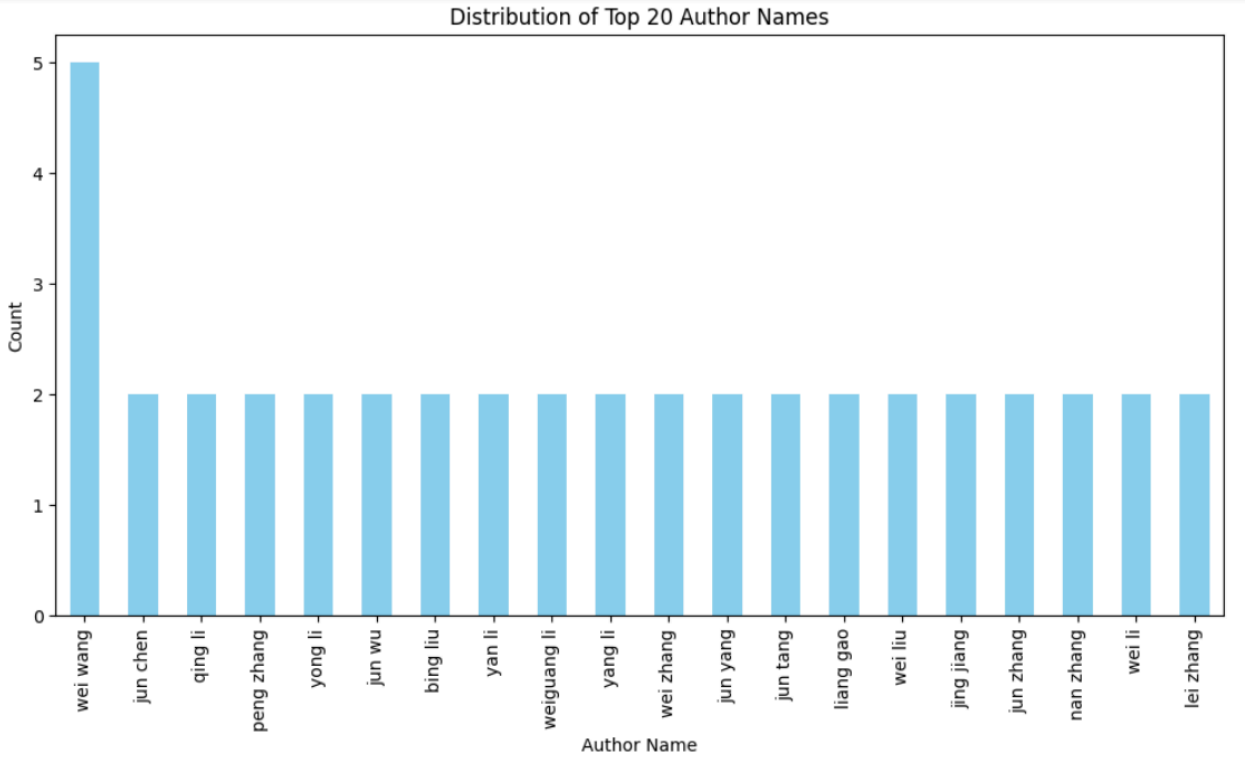


- There is a null null author again could be an issue with the way the json is being read because this feature we know has no null values

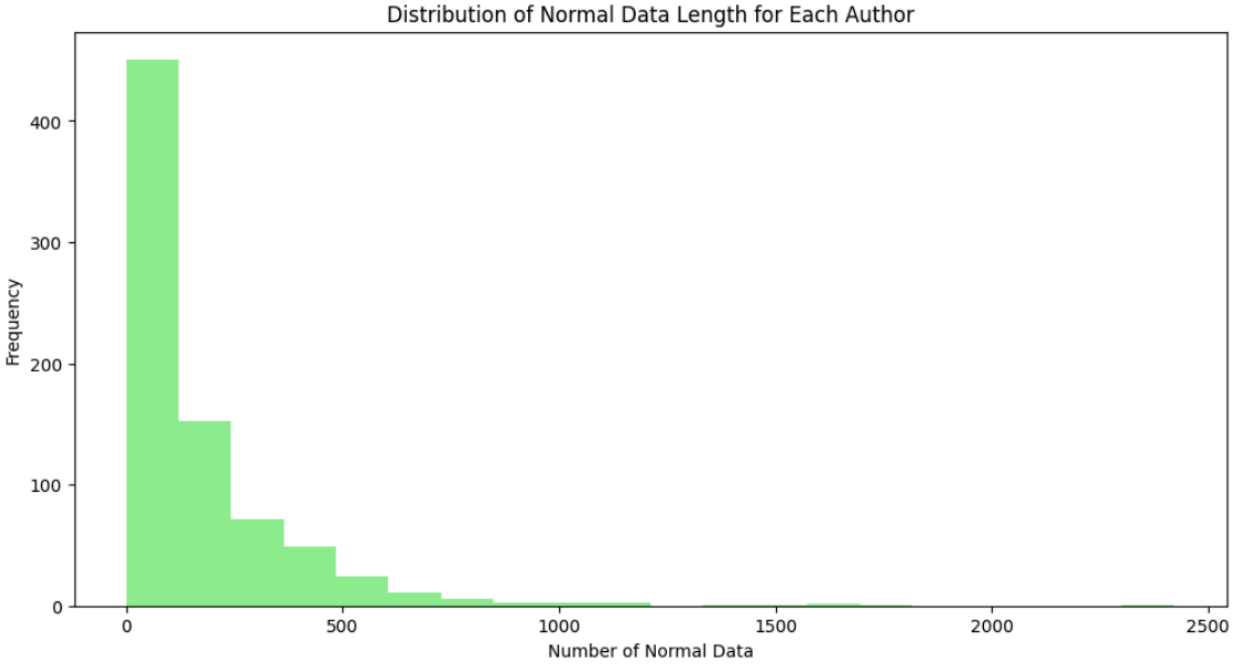
Train Author Dataset

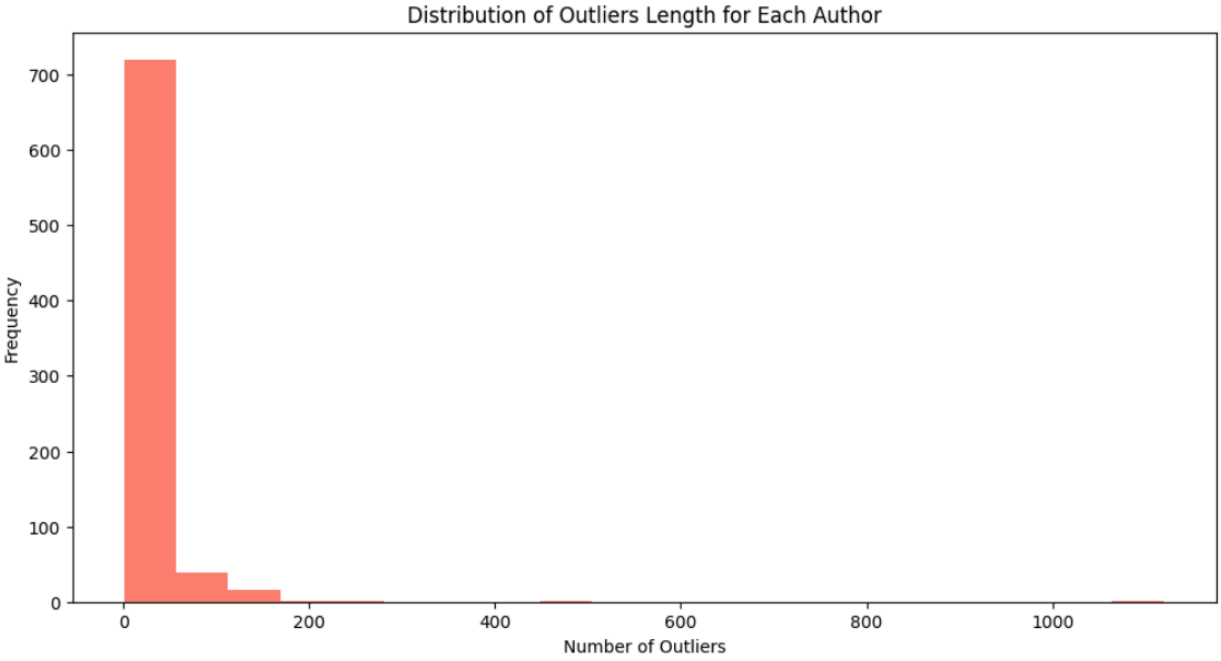
```
DataFrame Info:
<class 'pandas.core.frame.DataFrame'>
Index: 779 entries, Iki037dt to UeEnjWiz
Data columns (total 3 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   name         779 non-null    object
 1   normal_data  779 non-null    object
 2   outliers     779 non-null    object
dtypes: object(3)
memory usage: 24.3+ KB
None

DataFrame Head:
                    name                                        normal_data  \
Iki037dt  atsushi ochiai  [YzOCpPTO, AblgcGjH, B5aouLse, u1G7wBEv, W7w6P...
ZihzMro7     mingwu yang  [C58t0yYu, sWIRnfR3, HJW8h2mo, 0Ptx4O5n, fU4vB...
WXMYBk3c  jianzhao huang  [lJAIOXO4, fYJcce0K, ZaeOFAcI, kg9xDSXm, T37S3...
WrCODHhe     xuebiao yao  [3fYoJb1W, wjt8Y8ho, pPx6o7KZ, xgRarLPn, 9w9yz...
k3uSCGEE    shunlin tang  [gTeQer76, mVk2vmmN, TLKSll8D, Eg5NcmZ2, kM5Ip...

                                                 outliers
Iki037dt  [XL3wd3CP, BTKTiJp2, JxSjl5xc, 0jyMLgRt, uHWx8...
ZihzMro7  [qK8llKzD, I0eTdaAG, nFebDDiR, 903CyaNQ, Q45WM...
WXMYBk3c  [HwaUxOes, nvELwvhl, 6Z6SRTQh, R1yeZqOY, qnwco...
WrCODHhe  [OtmIuFFb, wnP8OmXf, IZ1qVc9S, YccNQrlZ, sLO7c...
k3uSCGEE  [xPmu4CGB, buwfccml, fBPzgpof, HgjM9QKW, rPk5S...
Number of Unique Authors: 779
```

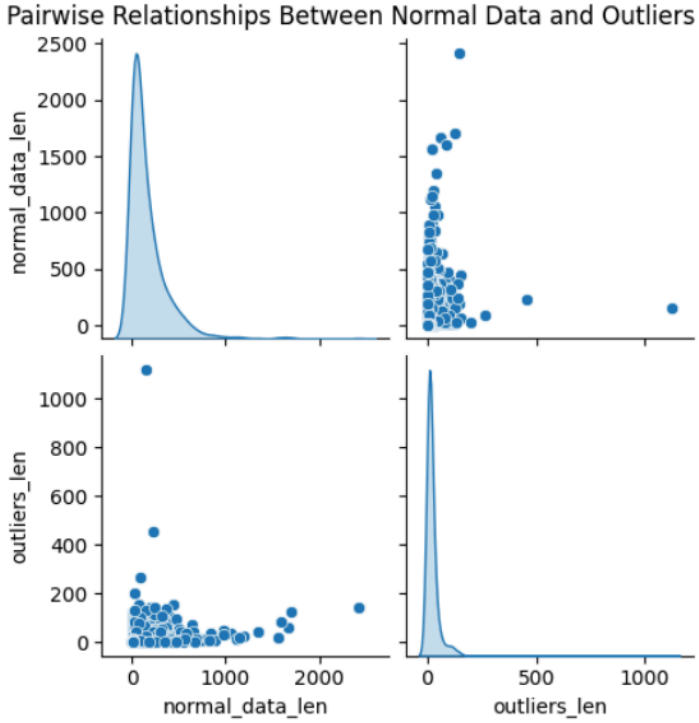
Top 20 Authors



Other visualizations




Distribution of Outliers Length for Each Author
Frequency
Number of Outliers
0
100
200
300
400
500
600
700
0
200
400
600
800
1000


Pairwise Relationships Between Normal Data and Outliers
normal_data_len
outliers_len
0
500
1000
1500
2000
2500
0
200
400
600
800
1000
0
1000
2000
0
500
1000