02_eda_pubmed

March 16, 2025

[23]:

```
import pandas as pd
pubmed = pd.read_csv('data/all_pubmed.csv') # make sure pubmed has a column␣
  ↪named 'category'
print(pubmed.columns)
pubmed.head()
```

```
Index(['pubmed_id', 'title', 'keywords', 'journal', 'abstract', 'methods',
       'results', 'conclusions', 'publication_date', 'category'],
      dtype='object')
```

[23]:

```
   pubmed_id                                              title  \
0   40073121  Targeting the NPY/NPY1R signaling axis in muta…
1   40069621  The value of preoperative RDW for post-pancrea…
2   40069616  Protocol of the IMPACT study: randomized, mult…
3   40066089  Association between human leukocyte antigen E …
4   40065459  Oncological and Survival Endpoints in Cancer C…

                                            keywords  \
0                                                NaN
1  Pancreatic ductal adenocarcinoma, Post-pancrea…
2  Atezolizumab, Bevacizumab, Conversion, Hepatoc…
3  HLA-E, cancer, human leukocyte antigen, immuno…
4  adverse events, cachexia, cancer, clinical tri…

                                      journal  \
0                            Science advances
1                                  BMC cancer
2                                  BMC cancer
3                       Frontiers in oncology
4  Journal of cachexia, sarcopenia and muscle

                                            abstract methods  \
0  Pancreatic cancer (PC) is a highly metastatic …     NaN
1  Pancreatic ductal adenocarcinoma (PDAC) is a h…     NaN
2  Atezolizumab plus bevacizumab is recommended a…     NaN
3  Immunotherapy has gained momentum with the dis…     NaN
4  In patients receiving anti-cancer treatment, c…     NaN
```

```
                                             results  \
0                                                NaN
1  A total of 2268 patients were analyzed. We fou…
2                                                NaN
3  After screening 657 articles, 11 studies were …
4  Fifty-seven trials were eligible, totalling 97…

                                         conclusions publication_date  \
0                                                NaN       2025-03-12
1  The preoperative RDW may be a useful marker fo…       2025-03-12
2                                                NaN       2025-03-12
3  This systematic review highlights that HLA-E e…       2025-03-11
4  In CC trials, oncological endpoints were mostl…       2025-03-11

                       category
0  Pubmed_Pancreatic_Cancer.csv
1  Pubmed_Pancreatic_Cancer.csv
2  Pubmed_Pancreatic_Cancer.csv
3  Pubmed_Pancreatic_Cancer.csv
4  Pubmed_Pancreatic_Cancer.csv
```

[24]:

```python
pubmed['category'] = (
    pubmed['category']
    .str.replace("Pubmed_", "", regex=False)
    .str.replace(".csv", "", regex=False)
    .str.replace("_", " ", regex=False)
    .str.replace("-", " ")  # Optional: Replace hyphens with spaces if needed
    .str.title()  # Capitalize each word
)
pubmed['category'].value_counts()
```

[24]:

```
category
Pancreatic Cancer              9831
Influenza                      8905
Hepatitis                      7087
Malaria                        6855
Endometriosis                  2839
Duchenne Muscular Dystrophy    1423
Drug Resistant Tuberculosis    1274
Chagas Disease                  680
Breast Cancer                   171
Alzheimer                        60
Name: count, dtype: int64
```

for rare diseases

[17]:

```
pubmed = pubmed[pubmed['category'].isin(['Pancreatic Cancer', 'Endometriosis',␣
  ↪'Chagas Disease','Drug Resistant Tuberculosis', 'Duchenne Muscular␣
  ↪Dystrophy'])]
```

[18]:

```
import pandas as pd


# Convert publication_date to datetime
pubmed['publication_date'] = pd.to_datetime(pubmed['publication_date'],␣
  ↪errors="coerce")

### 1) Number of samples per category
category_counts = pubmed['category'].value_counts()
print("Number of samples per category:")
print(category_counts)
print("\n")

### 2) Top 10 journal venues per category
top_journals = pubmed.groupby('category')['journal'].value_counts().
  ↪groupby(level=0).head(10)
print("Top 10 journal venues per category:")
print(top_journals)
print("\n")

### 3) Earliest, latest publication date, and elapsed days per category
date_range = pubmed.groupby('category')['publication_date'].agg(['min', 'max'])
date_range['elapsed_days'] = (date_range['max'] - date_range['min']).dt.days  #␣
  ↪Compute elapsed days
print("Earliest, latest publication date, and elapsed days per category:")
print(date_range)
```

```
Number of samples per category:
category
Pancreatic Cancer              9831
Endometriosis                  2839
Duchenne Muscular Dystrophy    1423
Drug Resistant Tuberculosis    1274
Chagas Disease                  680
Name: count, dtype: int64


Top 10 journal venues per category:
category                     journal
Chagas Disease               PLoS neglected tropical diseases
30
                             PloS one
23
```

```
                             Arquivos brasileiros de cardiologia
18
                             Antimicrobial agents and chemotherapy
15
                             Memorias do Instituto Oswaldo Cruz
13
                             Clinical infectious diseases : an official
publication of the Infectious Diseases Society of America
12
                             The American journal of tropical medicine and
hygiene
12
                             Revista da Sociedade Brasileira de Medicina
Tropical
11
                             Journal of acquired immune deficiency syndromes
(1999)
10
                             The Lancet. Infectious diseases
10
Drug Resistant Tuberculosis  The international journal of tuberculosis and lung
disease : the official journal of the International Union against Tuberculosis
and Lung Disease     90
                             Antimicrobial agents and chemotherapy
79
                             PloS one
48
                             Clinical infectious diseases : an official
publication of the Infectious Diseases Society of America
44
                             The European respiratory journal
35
                             Trials
26
                             The Lancet. Infectious diseases
25
                             The New England journal of medicine
21
                             American journal of respiratory and critical care
medicine
20
                             BMC infectious diseases
17
Duchenne Muscular Dystrophy  Neuromuscular disorders : NMD
113
                             Muscle & nerve
71
                             Journal of neuromuscular diseases
```

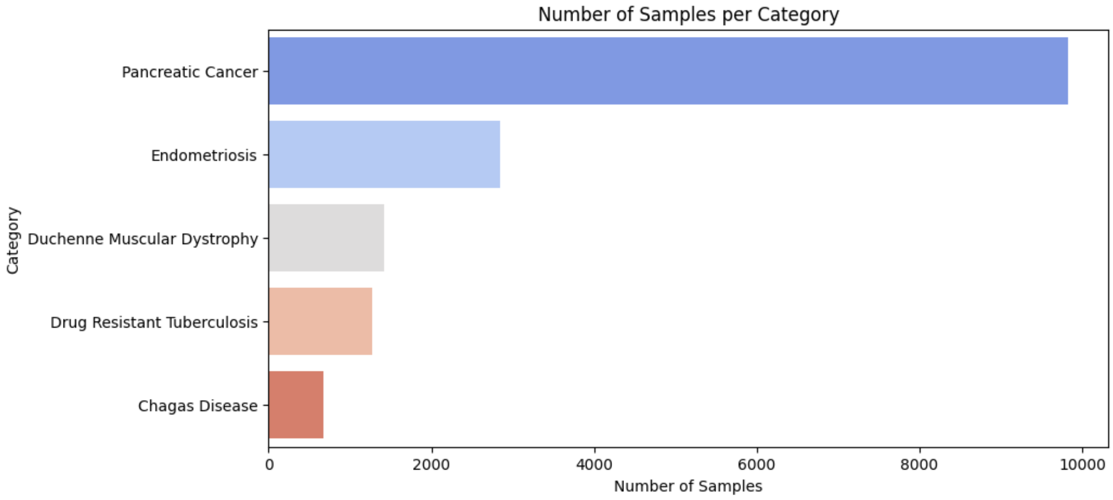
		54
	PloS one	47
	Neurology	46
	Human gene therapy	23
	Methods in molecular biology (Clifton, N.J.)	22
	Molecular therapy : the journal of the American Society of Gene Therapy	20
	Orphanet journal of rare diseases	20
	International journal of molecular sciences	15
Endometriosis	Fertility and sterility	311
	Human reproduction (Oxford, England)	232
	The Cochrane database of systematic reviews	83
	Journal of minimally invasive gynecology	78
	European journal of obstetrics, gynecology, and reproductive biology	62
	American journal of obstetrics and gynecology	57
	Archives of gynecology and obstetrics	57
	Obstetrics and gynecology	45
	Acta obstetricia et gynecologica Scandinavica	41
	International journal of gynaecology and obstetrics: the official organ of the International Federation of Gynaecology and Obstetrics	40
Pancreatic Cancer	Journal of clinical oncology : official journal of the American Society of Clinical Oncology	231
	Clinical cancer research : an official journal of the American Association for Cancer Research	180
	British journal of cancer	169
	Cancer chemotherapy and pharmacology	168

```
                             Annals of oncology : official journal of the
European Society for Medical Oncology
156
                             BMC cancer
146
                             Investigational new drugs
142
                             Cancer
137
                             European journal of cancer (Oxford, England : 1990)
136
                             Annals of surgery
130
Name: count, dtype: int64


Earliest, latest publication date, and elapsed days per category:
                                   min        max  elapsed_days
category
Chagas Disease              1969-01-01 2025-02-28         20512
Drug Resistant Tuberculosis 1965-05-01 2025-03-11         21864
Duchenne Muscular Dystrophy 1966-01-01 2025-03-11         21619
Endometriosis               1964-05-27 2025-03-12         22204
Pancreatic Cancer           1966-07-01 2025-03-12         21439
```

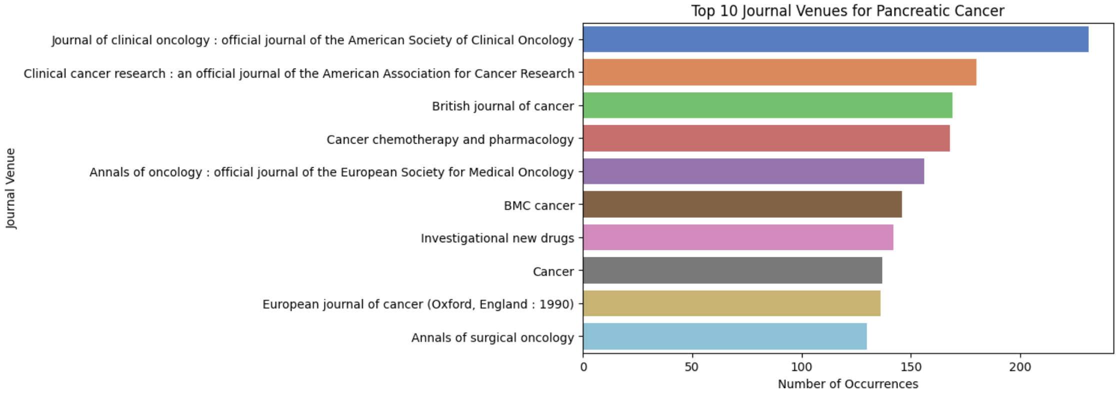
[ ]:

```python
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd


# Convert publication_date to datetime
pubmed['publication_date'] = pd.to_datetime(pubmed['publication_date'])

# 1) Number of Samples per Category
plt.figure(figsize=(10, 5))
sns.countplot(y=pubmed['category'], palette="coolwarm",␣
  ↪order=pubmed['category'].value_counts().index)
plt.xlabel("Number of Samples")
plt.ylabel("Category")
plt.title("Number of Samples per Category")
plt.show()

categories = pubmed['category'].unique()

for category in categories:
    plt.figure(figsize=(8, 5))
    category_data = pubmed[pubmed['category'] == category]['journal'].
  ↪value_counts().head(10)

    sns.barplot(y=category_data.index, x=category_data.values, palette="muted")
    plt.xlabel("Number of Occurrences")
    plt.ylabel("Journal Venue")
    plt.title(f"Top 10 Journal Venues for {category}")
    plt.show()

# 3) Earliest, Latest Publication Date, and Elapsed Days per Category
date_range = pubmed.groupby('category')['publication_date'].agg(['min', 'max']).
  ↪reset_index()
date_range['elapsed_days'] = (date_range['max'] - date_range['min']).dt.days

plt.figure(figsize=(12, 6))
sns.barplot(data=date_range, y="category", x="elapsed_days", palette="viridis")
plt.xlabel("Elapsed Days (Difference Between Earliest & Latest Publication)")
plt.ylabel("Category")
plt.title("Publication Date Range & Elapsed Days per Category")
plt.show()
```

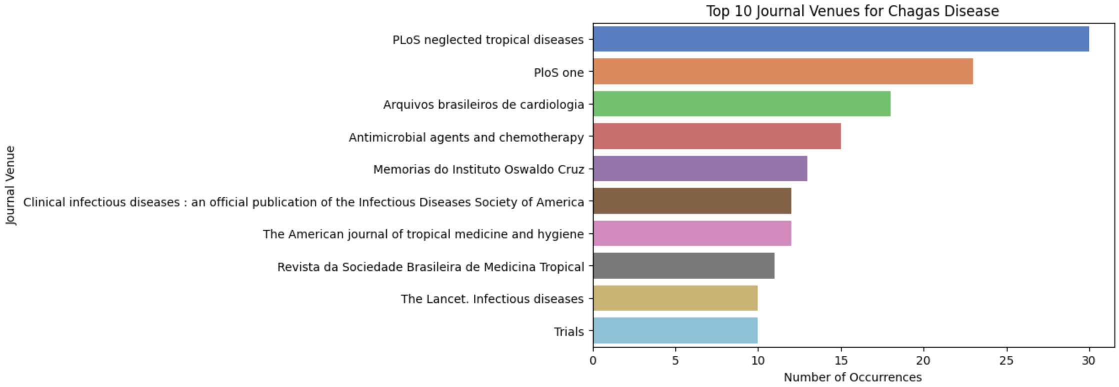
```
/tmp/ipykernel_8893/4216411280.py:11: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  sns.countplot(y=pubmed['category'], palette="coolwarm",
order=pubmed['category'].value_counts().index)
```



```
/tmp/ipykernel_8893/4216411280.py:34: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  sns.barplot(y=category_data.index, x=category_data.values, palette="muted")
```

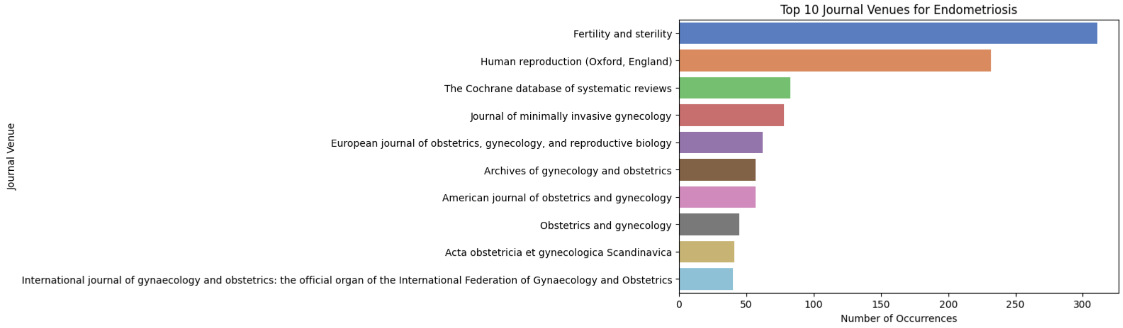


```
/tmp/ipykernel_8893/4216411280.py:34: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  sns.barplot(y=category_data.index, x=category_data.values, palette="muted")
```

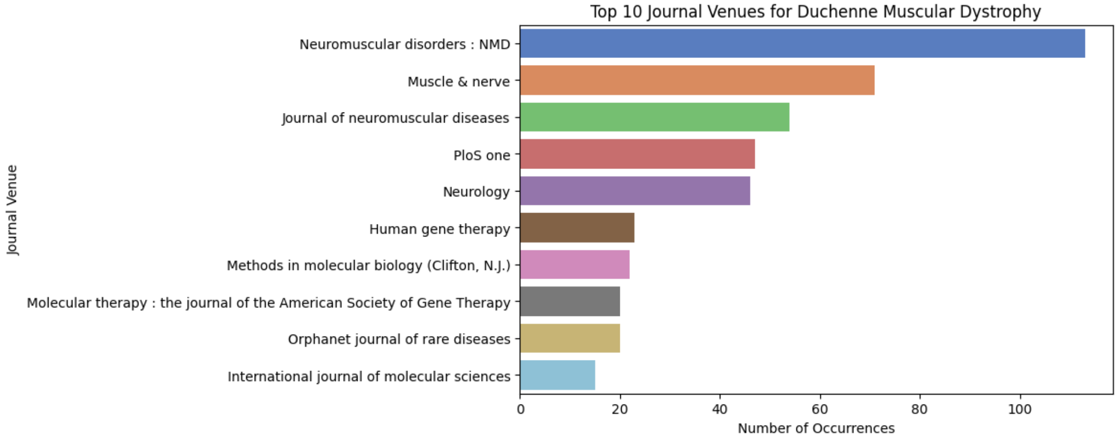


```
/tmp/ipykernel_8893/4216411280.py:34: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
```

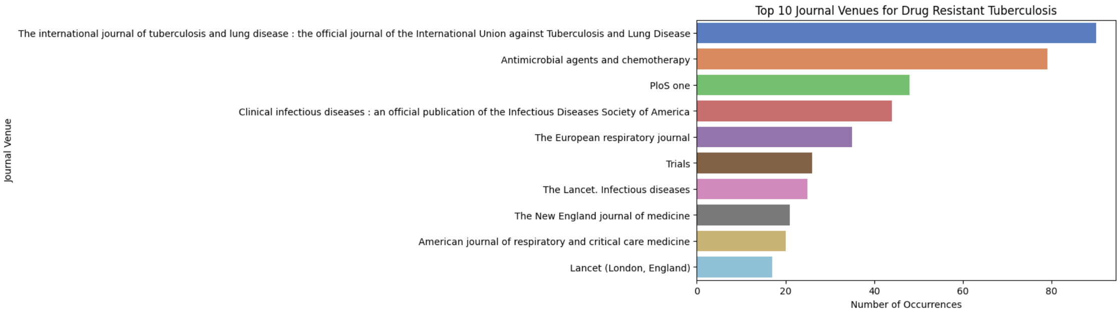
```
effect.

  sns.barplot(y=category_data.index, x=category_data.values, palette="muted")
```



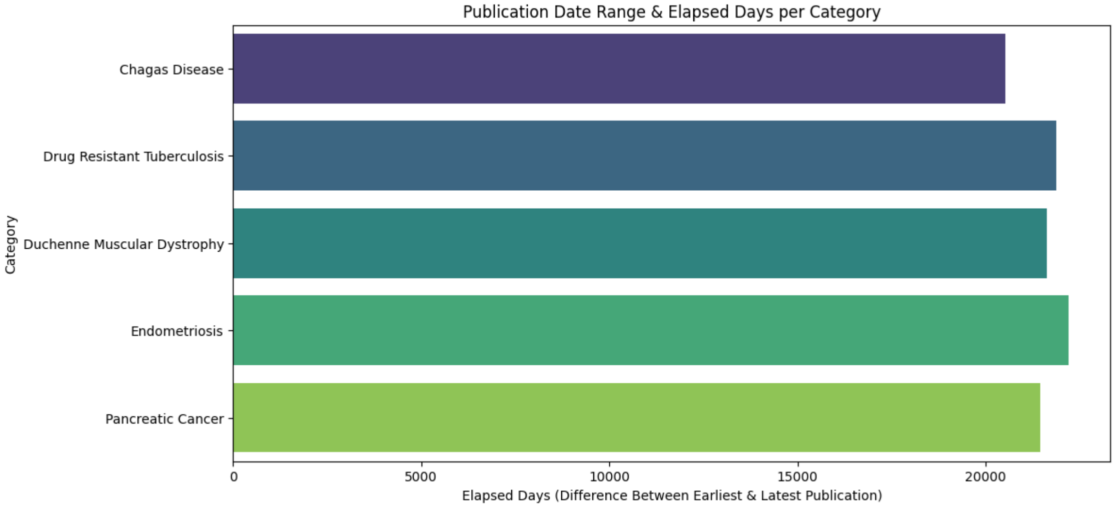
```
/tmp/ipykernel_8893/4216411280.py:34: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  sns.barplot(y=category_data.index, x=category_data.values, palette="muted")
```



```
/tmp/ipykernel_8893/4216411280.py:34: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  sns.barplot(y=category_data.index, x=category_data.values, palette="muted")
```



```
/tmp/ipykernel_8893/4216411280.py:45: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  sns.barplot(data=date_range, y="category", x="elapsed_days",
palette="viridis")
```



1 For Common Disease

[25]:
```
pubmed = pubmed[~pubmed['category'].isin(['Pancreatic Cancer', 'Endometriosis',␣
  ↪'Chagas Disease', 'Drug Resistant Tuberculosis', 'Duchenne Muscular␣
  ↪Dystrophy'])]
```

[26]:
```
import pandas as pd
```

```python
# Convert publication_date to datetime
pubmed['publication_date'] = pd.to_datetime(pubmed['publication_date'], errors="coerce")

### 1) Number of samples per category
category_counts = pubmed['category'].value_counts()
print("Number of samples per category:")
print(category_counts)
print("\n")

### 2) Top 10 journal venues per category
top_journals = pubmed.groupby('category')['journal'].value_counts().groupby(level=0).head(10)
print("Top 10 journal venues per category:")
print(top_journals)
print("\n")

### 3) Earliest, latest publication date, and elapsed days per category
date_range = pubmed.groupby('category')['publication_date'].agg(['min', 'max'])
date_range['elapsed_days'] = (date_range['max'] - date_range['min']).dt.days  # Compute elapsed days
print("Earliest, latest publication date, and elapsed days per category:")
print(date_range)
```

```
Number of samples per category:
category
Influenza        8905
Hepatitis        7087
Malaria          6855
Breast Cancer     171
Alzheimer          60
Name: count, dtype: int64


Top 10 journal venues per category:
category       journal
Alzheimer      Alzheimer's & dementia : the journal of the Alzheimer's
Association                                                      13
               The journal of prevention of Alzheimer's disease
8
               Alzheimer's & dementia (New York, N. Y.)
5
               International psychogeriatrics
5
               The American journal of geriatric psychiatry : official journal
of the American Association for Geriatric Psychiatry      3
```

```
               Alzheimer's research & therapy
2
               JAMA neurology
2
               Aging clinical and experimental research
1
               Alzheimer's & dementia (Amsterdam, Netherlands)
1
               American journal of Alzheimer's disease and other dementias
1
Breast Cancer  Journal of clinical oncology : official journal of the American
Society of Clinical Oncology                              9
               Breast (Edinburgh, Scotland)
6
               The oncologist
5
               Clinical breast cancer
4
               The Lancet. Oncology
4
               Annals of surgical oncology
3
               Breast cancer research and treatment
3
               Cancers
3
               Clinical cancer research : an official journal of the American
Association for Cancer Research                            3
               Expert review of pharmacoeconomics & outcomes research
3
Hepatitis      Vaccine
402
               Journal of hepatology
245
               Hepatology (Baltimore, Md.)
221
               Journal of viral hepatitis
171
               Zhonghua gan zang bing za zhi = Zhonghua ganzangbing zazhi =
Chinese journal of hepatology                              142
               World journal of gastroenterology
125
               PloS one
109
               The Pediatric infectious disease journal
104
               Lancet (London, England)
98
```

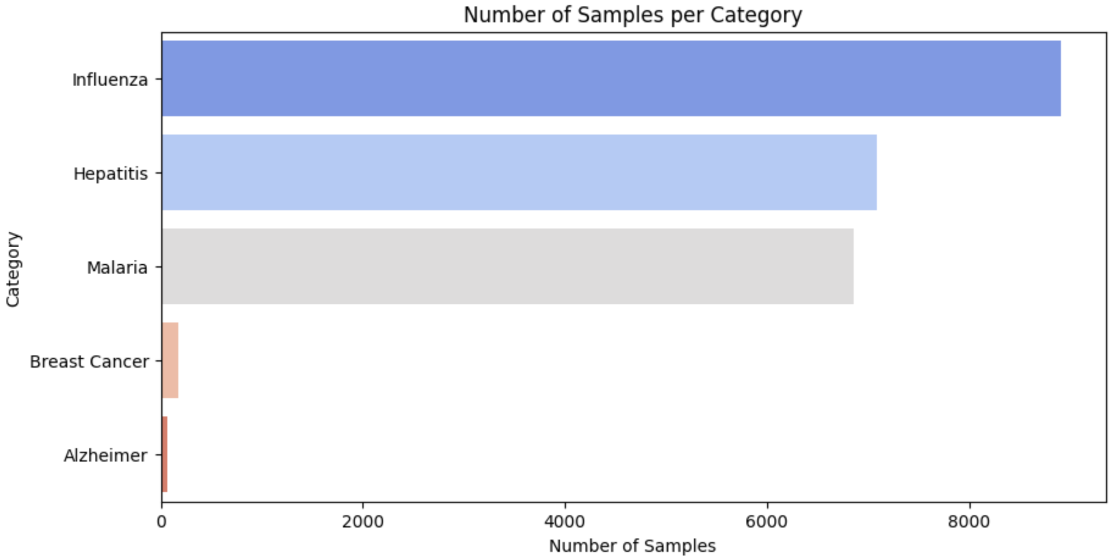
```
               Antiviral therapy
90
Influenza      Vaccine
889
               The Journal of infectious diseases
326
               The Pediatric infectious disease journal
283
               Human vaccines & immunotherapeutics
218
               PloS one
200
               Clinical infectious diseases : an official publication of the
Infectious Diseases Society of America                    161
               Antimicrobial agents and chemotherapy
121
               Lancet (London, England)
116
               The Cochrane database of systematic reviews
112
               The Japanese journal of antibiotics
107
Malaria        Malaria journal
693
               The American journal of tropical medicine and hygiene
471
               PloS one
331
               Transactions of the Royal Society of Tropical Medicine and
Hygiene                                                      255
               The Journal of infectious diseases
217
               Antimicrobial agents and chemotherapy
200
               Lancet (London, England)
190
               Vaccine
176
               Clinical infectious diseases : an official publication of the
Infectious Diseases Society of America                    166
               Tropical medicine & international health : TM & IH
161
Name: count, dtype: int64


Earliest, latest publication date, and elapsed days per category:
                     min        max  elapsed_days
category
```

```
Alzheimer     1996-01-01 2025-03-15         10666
Breast Cancer 1977-12-01 2025-02-17         17245
Hepatitis     1966-05-09 2025-03-09         21489
Influenza     1953-06-20 2025-03-09         26195
Malaria       1945-12-29 2025-03-13         28929
```

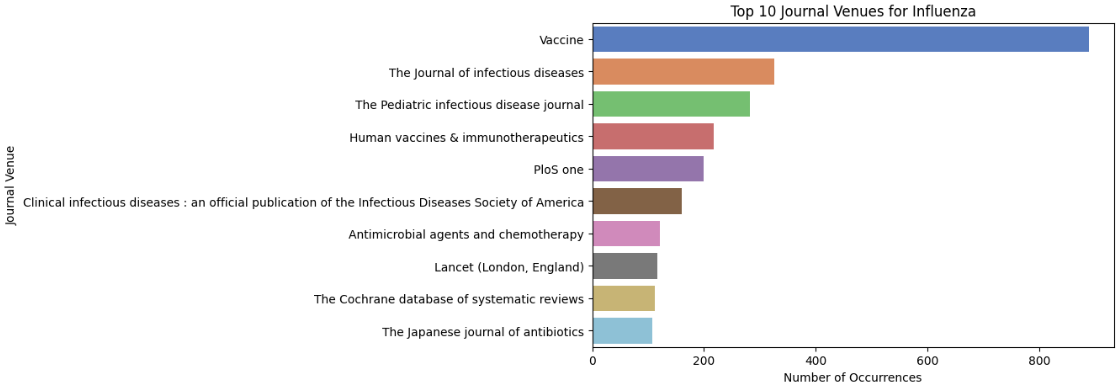
[27]:

```
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd


# Convert publication_date to datetime
pubmed['publication_date'] = pd.to_datetime(pubmed['publication_date'])

# 1) Number of Samples per Category
plt.figure(figsize=(10, 5))
sns.countplot(y=pubmed['category'], palette="coolwarm",␣
 ↪order=pubmed['category'].value_counts().index)
plt.xlabel("Number of Samples")
plt.ylabel("Category")
plt.title("Number of Samples per Category")
plt.show()

categories = pubmed['category'].unique()

for category in categories:
    plt.figure(figsize=(8, 5))
    category_data = pubmed[pubmed['category'] == category]['journal'].
 ↪value_counts().head(10)

    sns.barplot(y=category_data.index, x=category_data.values, palette="muted")
    plt.xlabel("Number of Occurrences")
    plt.ylabel("Journal Venue")
    plt.title(f"Top 10 Journal Venues for {category}")
    plt.show()

# 3) Earliest, Latest Publication Date, and Elapsed Days per Category
date_range = pubmed.groupby('category')['publication_date'].agg(['min', 'max']).
 ↪reset_index()
date_range['elapsed_days'] = (date_range['max'] - date_range['min']).dt.days

plt.figure(figsize=(12, 6))
sns.barplot(data=date_range, y="category", x="elapsed_days", palette="viridis")
plt.xlabel("Elapsed Days (Difference Between Earliest & Latest Publication)")
plt.ylabel("Category")
plt.title("Publication Date Range & Elapsed Days per Category")
plt.show()
```

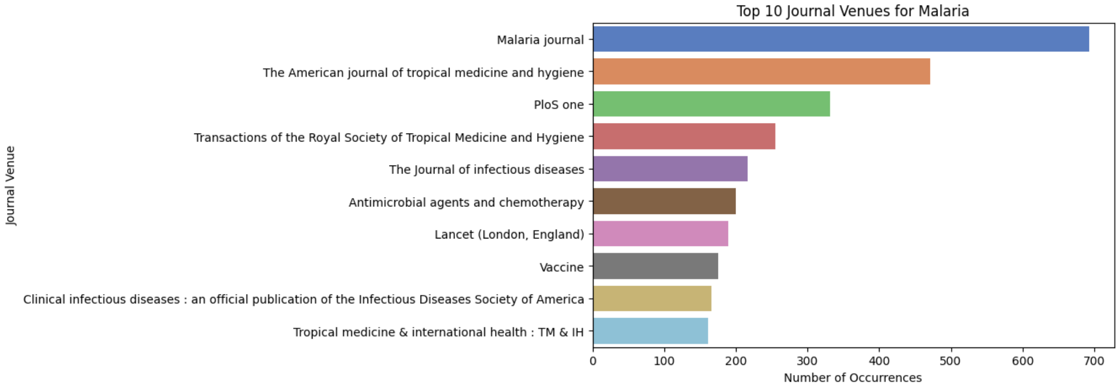
```
/tmp/ipykernel_8893/2142337901.py:11: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  sns.countplot(y=pubmed['category'], palette="coolwarm",
order=pubmed['category'].value_counts().index)
```



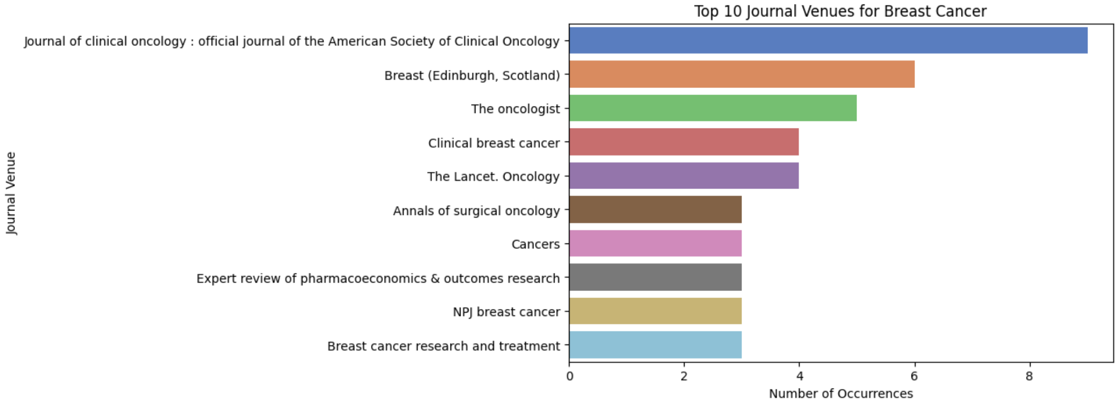
```
/tmp/ipykernel_8893/2142337901.py:23: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  sns.barplot(y=category_data.index, x=category_data.values, palette="muted")
```



```
/tmp/ipykernel_8893/2142337901.py:23: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  sns.barplot(y=category_data.index, x=category_data.values, palette="muted")
```

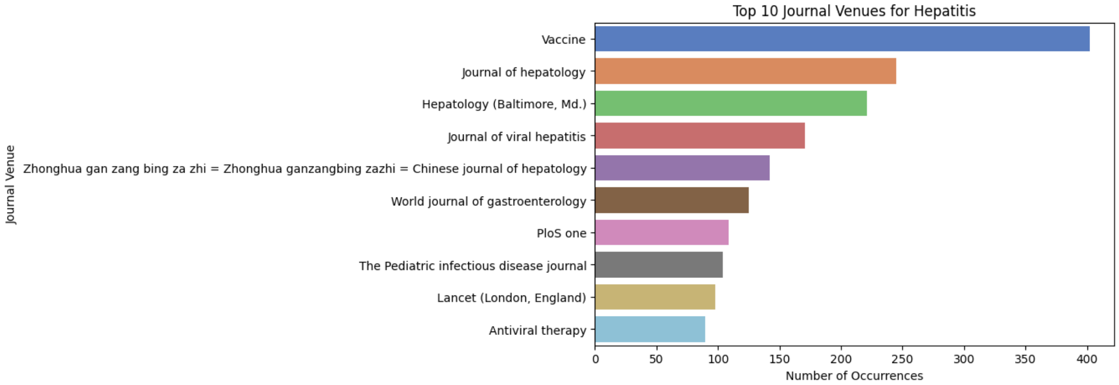


```
/tmp/ipykernel_8893/2142337901.py:23: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  sns.barplot(y=category_data.index, x=category_data.values, palette="muted")
```

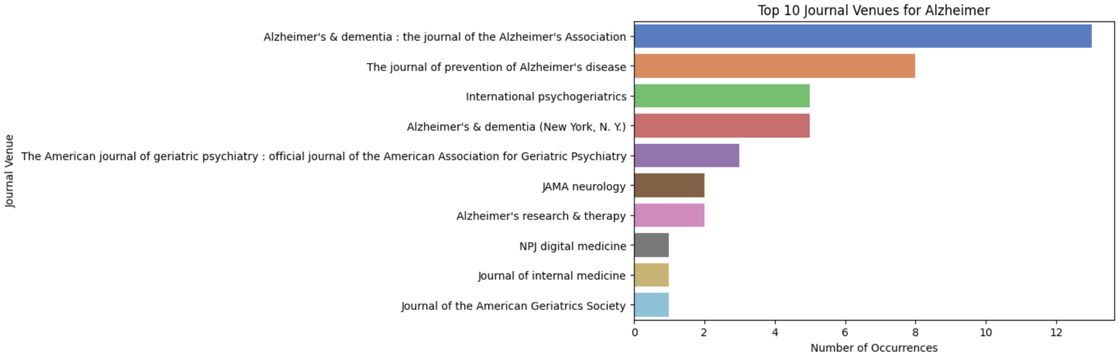


```
/tmp/ipykernel_8893/2142337901.py:23: FutureWarning:
```

```
Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  sns.barplot(y=category_data.index, x=category_data.values, palette="muted")
```

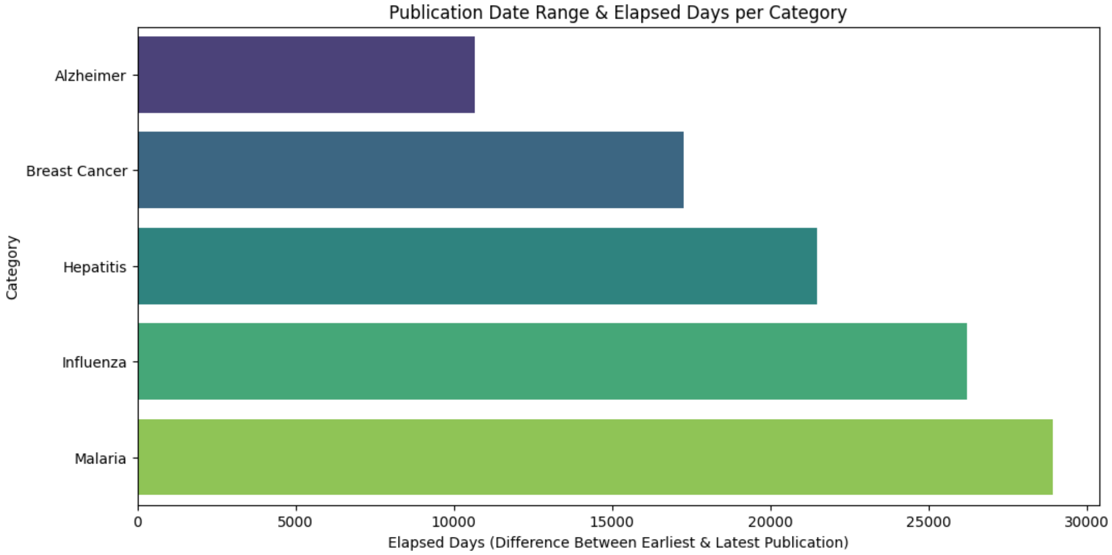


```
/tmp/ipykernel_8893/2142337901.py:23: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  sns.barplot(y=category_data.index, x=category_data.values, palette="muted")
```



```
/tmp/ipykernel_8893/2142337901.py:34: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.
```

```
  sns.barplot(data=date_range, y="category", x="elapsed_days",
palette="viridis")
```


Publication Date Range & Elapsed Days per Category
Category
Alzheimer
Breast Cancer
Hepatitis
Influenza
Malaria
Elapsed Days (Difference Between Earliest & Latest Publication)