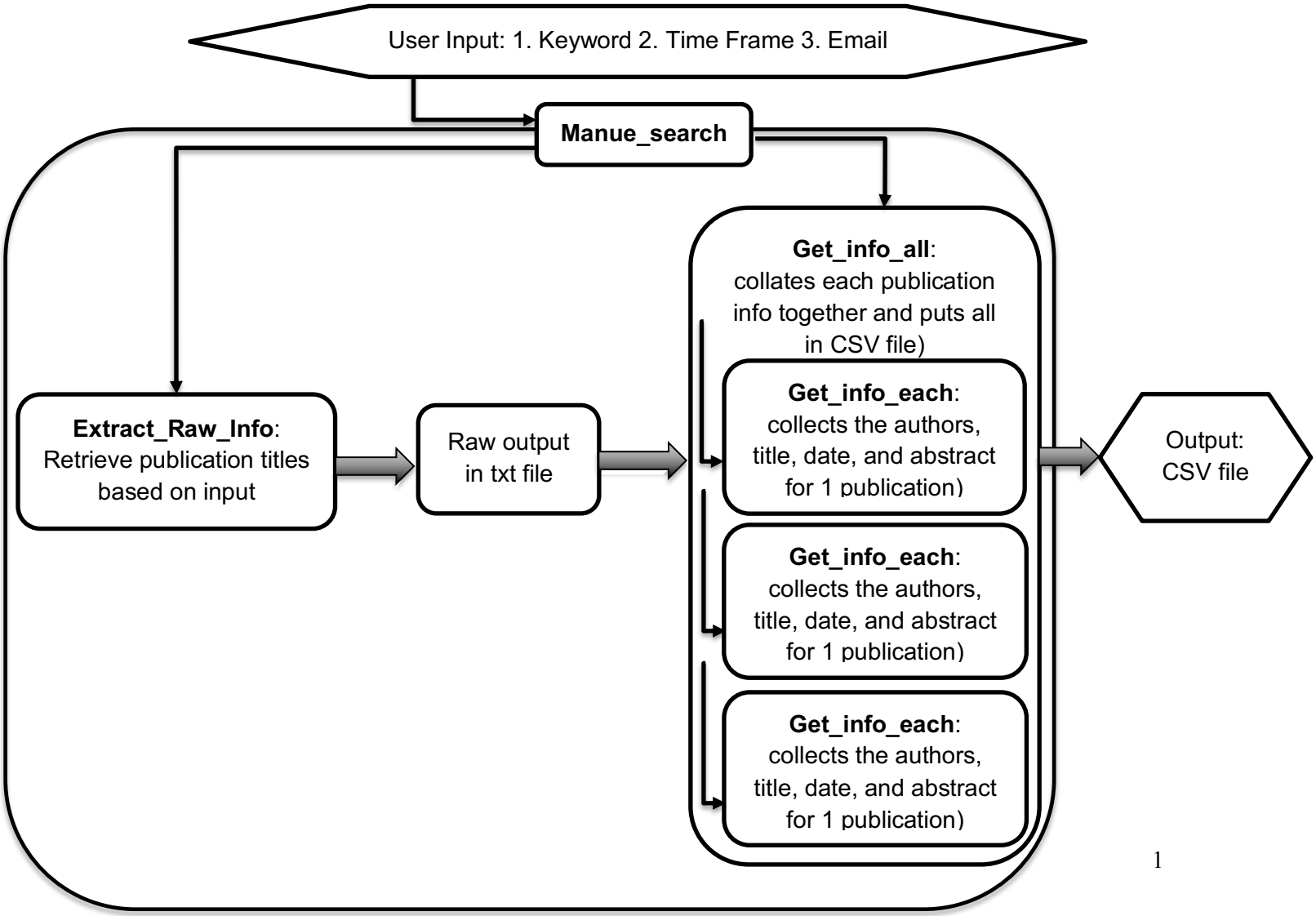
Emille Benavides, Perrin Hines, Theresa Nguyen, Carlos Vera Recio, Jinhao Zou
Introduction to Data Science
December 16, 2020

Group 10: Capstone Project Report

Section 1: Program Design

This program begins with a crawler module that collects information on papers published to PubMed, an archive of biomedical and life sciences journal literature. The crawler module prompts the user for a keyword and time frame for publication date on PubMed, searches for publications within the criteria specified, and collects the publication titles, list of authors, real-time publication dates, and abstracts. The module then creates a CSV file containing the retrieved data. To perform these tasks, the crawler module uses four functions: Manue_search, Get_info_each, Extract_Raw_Info, and Get_info_all. The first function, Manue_search, is the main function of the crawler module. It collects the keyword and dates that will be used as criteria for the PubMed search. It calls two functions, Extract_Raw_Info and Get_info_all, to generate the CSV file. The first function, Extract_Raw_Info, is used to extract the name of papers which contain the input keyword and are published on PubMed during the specified time window, and puts the details in a txt file. The next function, Get_info_all, uses the Get_info_each function to collect the required information for all papers from the txt file and generate a CSV file. The function Get_info_each takes one paper name

and extracts the required information for that paper, including the real publication date (rather than the PubMed publication date). This can be seen in the below workflow diagram:



After the crawler module is complete, the next phase of the program is a database module that creates a database in SQLite using the CSV file obtained from the crawler module. This module also prompts the user to search for publications in the SQLite database based on the author's name.

The final piece of the program offers visualizations for all the data extracted in the crawler module, using the CSV file that was generated. The visualization module uses the Matplotlib and Seaborn packages to create a histogram that aids in recognizing trends over time, such as the number of publications per year (based on input keywords and time frame from the crawler module). Finally, the visualization model also gives basic summary statistics on number of publications per year including the mean, median, standard deviation, and first to third quartile range.

Section 2: Implementation Details

The following is a list of packages, tools, libraries, and modules that are used in this program and their description:

- **Biopython**: A set of libraries that contain tools for biological computation. Biopython contains a module named Entrez. This package needs to be installed prior to running the program.
- **Entrez:** A Biopython package that provides a code to access NCBI through the web. This makes it possible to look up and extract data from journals and articles published to PubMed.
- **Urllib**: Urllib is the URL handling module for python. Urllib is a package that collects several modules for working with URLs. This program uses the error module to print a "Please retry" message when the program is unable to connect to PUBMED.
- **NumPy**: A package used for scientific computing with the use of arrays. It offers tools for quantum computing, statistical computing, bioinformatics, mathematical analysis, and much more. It is most commonly used for data analysis.
- **Pandas**: Pandas is an open-source library built on top of NumPy. It provides easy-to-use data structures and data analysis tools for the Python programming language. It allows for fast analysis and data cleaning and preparation.
- **CSV:** This function allows reading and writing of CSV files in Python. A CSV (Comma Separated Values) file is a plain type of text file that arranges data using commas to separate the data.
- **SQLite3**: A module that allows Python to work with SQL databases. It can be used to create databases, tables, and data insertion.
- **Pathlib**: A package that helps deal with file paths. This module offers classes representing filesystem paths with semantics appropriate for different operating systems.
- **SciPy:** A Python library used for scientific and technical computing built on top of NumPy. It offers utility functions for optimization, statistics, and signal processing.
- **Matplotlib**: A plotting library for the Python programming language and its numerical mathematics extension NumPy. It can be used to create static, animated, and interactive visualizations.
- **Seaborn**: A data visualization library based on matplotlib. It is used for visualization of statistical data and can be used to create graphs, scatter plots, histograms, and much more.
- **Ipyfilechooser**: This module is a Python file chooser widget for use in Jupyter/IPython in conjunction with ipywidgets. It displays a widget that allows you to choose what file to upload. This package needs to be installed prior to running the program.

Section 3: Results

Part One (Crawler Module)

Figure 1. Example of output from a PubMed search using the crawler module based on the keyword, “HIV,” and specified time frame, “2020/08/29 – 2020/08/30.” Please see Jupyter Notebook for full details.

```
Manue_search()
```

```
Please type in a keyword you are interested in: eg. HIV: HIV
Please type in a time window for searching in format yyyy/mm/dd - yyyy/mm/dd: eg. 2020/08/29 - 2020/08/30: 2020/08/
29 - 2020/08/30
Please type your email address:x
Found 76 results
Going to download record 1 to 10
Going to download record 11 to 20
Going to download record 21 to 30
Going to download record 31 to 40
Going to download record 41 to 50
Going to download record 51 to 60
Going to download record 61 to 70
Going to download record 71 to 76
The number of papers extracted is: 76
```

Part Two (Database and Query Module)

Table 1. Example of the SQLite Database (titled “database 2”) and table titled “HIV_Papers” created in the database module. The Title, Author, and Abstract columns are all shortened due to the length of entries.

	Title	Author	Publish_date	Abstract
1	High microbial translocation...	Kantamala, Doungnapa; Praparattanapan, Jutarat; Ta...	2020 Dec	BACKGROUND: Individuals residing in areas with high prevalence of foodborne infection
2	High sleep-related breathin...	Chen, Chang-Chun; Lin, Cheng-Yu; Chen, Yen-Chin;...	2020 Nov	BACKGROUND: Sleep-related breathing disorders (SRBD) not only adversely impact card
3	Design, synthesis and SAR ...	Patel, Manoj; Cianci, Christopher; Allard, Christopher...	2020 Nov 1	The design, synthesis and structure-activity relationships associated with a series of C2-
4	Building resilient and respo...	Veepanattu, P; Singh, S; Mendelson, M; Nampoothiri...	2020 Nov	Research, collaboration, and knowledge exchange are critical to global efforts to tackle an
5	Association of maternal and...	Sevenoaks, Tatum; Wedderburn, Catherine J; Donald...	2020 Aug 26	HIV-exposed uninfected (HEU) children may have altered immune regulation and poorer
6	Head-to-head comparison ...	Tiwari, Aseem Kumar; Upadhyay, Anand Prakash; Ar...	2020 Nov	Safe blood transfusion being the cornerstone of any Blood Transfusion Services requires
7	Risk factors for COVID-19 d...	Boulle, Andrew; Davies, Mary-Ann; Hussey, Hannah; ...	2020 Aug 29	BACKGROUND: Risk factors for COVID-19 death in sub-Saharan Africa and the effects of
8	Development of a composit...	Acharya, Shrikala; Setia, Maninder Singh; Palkar, Am...	2020 Aug 29	A key recommendation of the National AIDS Control Programme-IV of India was to develo
9	HIV-1 Integrase Inhibitors: A...	Scarsi, Kimberly K; Havens, Joshua P; Podany, Antho...	2020 Nov	The newest class of antiretrovirals for all persons living with HIV are the integrase strand t
10	Comorbidities and HCV coin...	Garagiola, Elisabetta; Foglia, Emanuela; Ferrario, Luc...	2020 Aug 29	BACKGROUND: Since HIV+ treatment has become more effective, the average age of pe
11	Phlorotannins as HIV Vpu in...	Langarizadeh, Mohammad Amin; Abiri, Ardavan; Gha...	2020 Aug 29	The importance of new effective treatment methodologies for human immunodeficiency v
12	Assessing changes in vitami...	Crutchley, Rustin D; Jacobs, David M; Gathe, Joseph...	2020 Aug 27	BACKGROUND: Vitamin D deficiency is common in HIV population and has been associat
13	Prevalence of Asymptomati...	Bhati, Rajendra; Sirohi, Pramendra; Sejoo, Bharat; Ku...	2020 Aug 27	OBJECTIVE: Cryptococcal meningitis is an important cause of morbidity and mortality in l
14	Recently acquired infection ...	Quiner, Claire; Bruhn, Roberta; Grebe, Eduard; Di Ger...	2020 Aug 28	BACKGROUND: Monitoring of transfusion-transmissible infections in the blood supply is e
15	Exploring Relative Preferenc...	Eshun-Wilson, I; Kim, H-Y; Schwartz, S; Conte, M; Gl...	2020 Oct	PURPOSE OF REVIEW: Aligning HIV treatment services with patient preferences can prom
16	Mental Health, Social Influe...	Wood, Sarah M; Morales, Knashawn H; Metzger, Davi...	2020 Aug 28	The effects of mental health comorbidities and social support on the HIV pre-exposure pr
17	Coexistent HIV infection is n...	de Carvalho, Fabricio Rodrigues Torres; Ho, Yeh-Li; J...	2020 Nov	NULL
18	Retrospective Hospital-bas...	Onkarappa, Saroja A; Panpalia, Nikhil G; Naik, Karkal R	2020 Jul-Aug	Background: Patients with retroviral disease are prone to opportunistic infections (OIs) of
19	Barriers in accessing HIV ca...	Djiadeu, Pascal; Yusuf, Abban; Ongolo-Zogo, Cleme...	2020 Aug 27	INTRODUCTION: In 2001, 50%–55% of French-speaking minority communities did not ha
20	Self-reported disability in re...	Prynn, Josephine E; Dube, Albert; Mkandawire, Jose...	2020 Aug 27	OBJECTIVES: We investigated whether self-reported disability was associated with mort
21	[Sternal tuberculosis: an un...	Koffi, M-O; Djaha, K J-M; Kone, A; Djegbeton, A; Kpa...	2020 Oct	INTRODUCTION: Osteo-articular tuberculosis mainly affects the spine. Sternal localizatio
22	Interpersonal reactivity inde...	Sack, Daniel E; Frisby, Michael B; Diemer, Matthew A;...	2020 Aug 28	BACKGROUND: The ability to understand another's emotions and act appropriately, empa
23	Social determinants of ment...	Bhalla, Ish P; Stefanovics, Elina A; Rosenheck, Robert...	2020 Aug 28	BACKGROUND: Since deinstitutionalization in the 1950s-1970s, public mental health care
24	Understanding long-term HI...	Freeman, Robert; Gwadz, Marya; Wilton, Leo; Collins...	2020 Aug 28	BACKGROUND: Persons living with HIV (PLWH) are living longer, although racial/ethnic an
25	Efficacy of HIV intervention...	Chen, Dahui; Luo, Ganfeng; Meng, Xiaojun; Wang, Zi...	2020 Aug 28	BACKGROUND: Factory workers in low vulnerable to HIV transmission. Interventions are n
26	Water Extract of Agastache ...	Jang, Seon-A; Hwang, Youn-Hwan; Kim, Taesoo; Yan...	2020 Aug 26	Estrogen deficiency in postmenopausal women causes homeostatic imbalance of bone, r

Figure 2. Example of a query of the SQLite database table, “HIV_Papers,” for publication titles by author name.

```
while True:
    input_author = input('Enter an author name, e.g. Nguyen: ')
    result = c.execute(f"SELECT Author, Title FROM HIV_Papers WHERE Author LIKE '%{input_author}%'")
    res = result.fetchall()
    if len(res)!=0:
        for row in res:
            print(row)
        break
    else:
        print("Sorry, the author name you entered does not match any authors.")
```

```
Enter an author name, e.g. Nguyen: Farooque
Sorry, the author name you entered does not match any authors.
Enter an author name, e.g. Nguyen: Nguyen
('Nguyen, Huy; Gazy, Nicky; Venketaraman, Vishwanath', 'A Role of Intracellular Toll-Like Receptors (3, 7, and 9) i
n Response to Mycobacterium tuberculosis and Co-Infection with HIV.')
('Trang, Nguyen Thu; Jauffret-Roustide, Marie; Giang, Le Minh; Visier, Laurent', 'How to be self-reliant in a stigm
atising context? Challenges facing people who inject drugs in Vietnam.')
```

Part Three (Visualization Module)

Figure 3. This is an example of choosing which CSV file to read using the File Chooser widget.

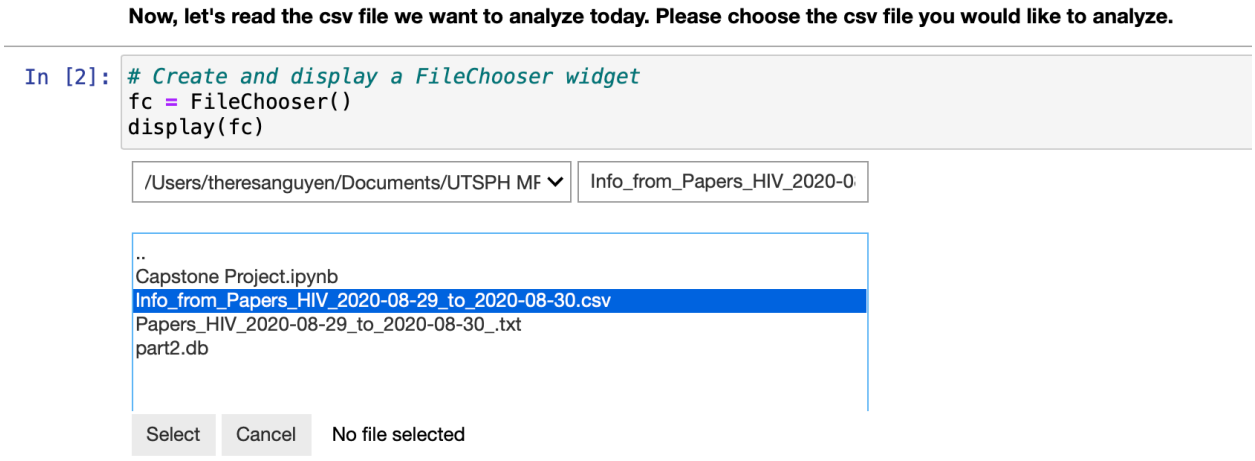
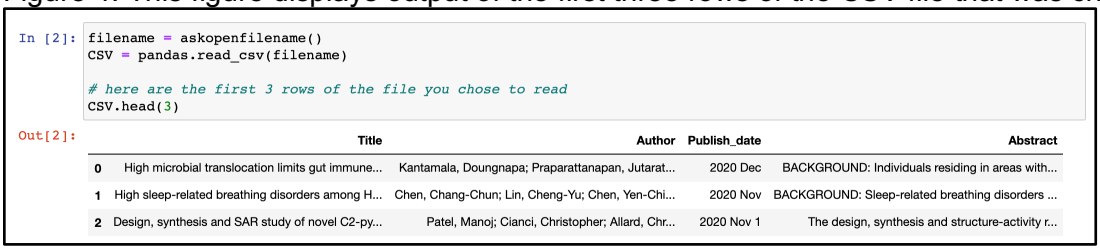


Figure 4. This figure displays output of the first three rows of the CSV file that was chosen.



```
In [2]: filename = askopenfilename()
        CSV = pandas.read_csv(filename)

        # here are the first 3 rows of the file you chose to read
        CSV.head(3)
```

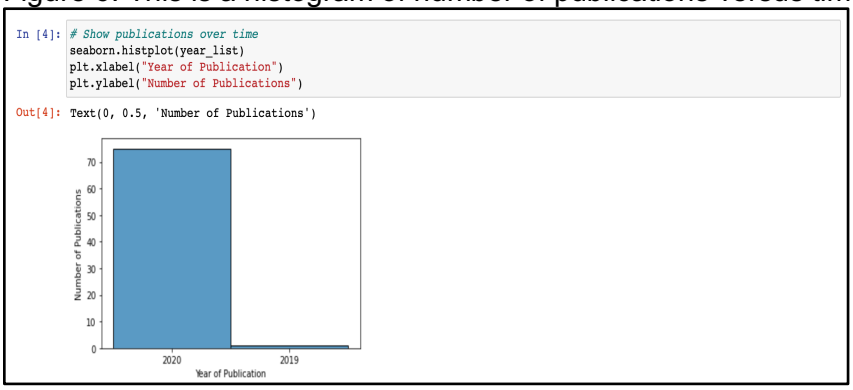
Out[2]:

	Title	Author	Publish_date	Abstract
0	High microbial translocation limits gut immune...	Kantamala, Doungnapa; Praparattanapan, Jutarat...	2020 Dec	BACKGROUND: Individuals residing in areas with...
1	High sleep-related breathing disorders among H...	Chen, Chang-Chun; Lin, Cheng-Yu; Chen, Yen-Chi...	2020 Nov	BACKGROUND: Sleep-related breathing disorders ...
2	Design, synthesis and SAR study of novel C2-py...	Patel, Manoj; Cianci, Christopher; Allard, Chr...	2020 Nov 1	The design, synthesis and structure-activity r...

Figure 5. This figure displays the output for number of publications in each year generated in the CSV file. There is 1 publication from 2019 and 75 publications from 2020. Please see the Jupyter Notebook for full details.

	Year	Publications
0	2019	1
1	2020	75

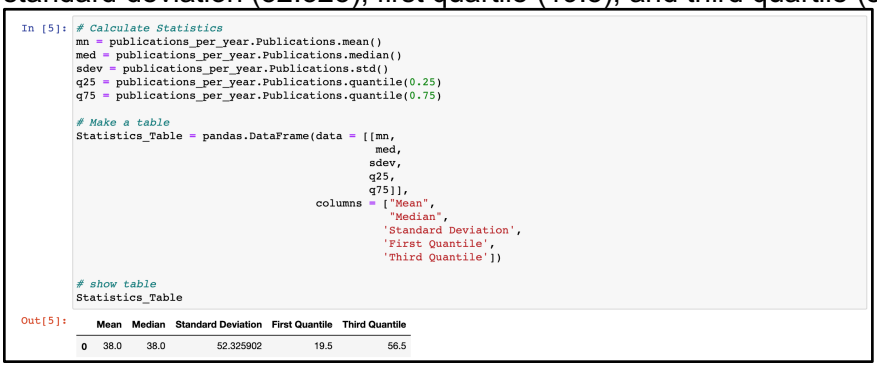
Figure 6. This is a histogram of number of publications versus time in years.



```
In [4]: # Show publications over time
        seaborn.histplot(year_list)
        plt.xlabel("Year of Publication")
        plt.ylabel("Number of Publications")
```

Out[4]: Text(0, 0.5, 'Number of Publications')

Figure 7. This figure displays an output of the summary statistics for the number of publications per year. The output denotes the mean (38.0 publications per year), median (38.0 publications), standard deviation (52.325), first quartile (19.5), and third quartile (56.5) of the user's query.



In [5]:
```
# Calculate Statistics
mn = publications_per_year.Publications.mean()
med = publications_per_year.Publications.median()
sdev = publications_per_year.Publications.std()
q25 = publications_per_year.Publications.quantile(0.25)
q75 = publications_per_year.Publications.quantile(0.75)

# Make a table
Statistics_Table = pandas.DataFrame(data = [[mn,
                                             med,
                                             sdev,
                                             q25,
                                             q75]],
                                    columns = ["Mean",
                                               "Median",
                                               'Standard Deviation',
                                               'First Quantile',
                                               'Third Quantile'])

# show table
Statistics_Table
```

Out[5]:

	Mean	Median	Standard Deviation	First Quantile	Third Quantile
0	38.0	38.0	52.325902	19.5	56.5

Section 4: User Manual/Guide

This program is a tool that collects academic journal information from PubMed to assist your research efforts and interests. By entering a keyword and date range into the tool, you will receive a listed table of results (publication titles, authors, publication date, and abstract) tailored to your request. The tool then allows you to search the table for an author's name and return a list of that searched author's publication titles. Lastly, the program allows you to visualize the results from the PubMed search by returning summary statistics and creating histograms. To use the program and begin your search, you will need to install biopython and ipyfilechooser.

Keyword Search Entry

You will enter a keyword to start your publication search. If there are results that include the searched keyword, then a CSV file will be generated. You should enter a keyword into the ***"Please type in a keyword you are interested in: eg. HIV"*** search box. Any entry is acceptable. The entry is **not** case-sensitive. It is recommended that singular "words" or acronyms are used (eg. medicine, HIV) to yield desired search results. Phrases (eg. emergency department wait), while acceptable, may yield little to no results.

Date Range Entry

Next, you will enter the date range of publication on PubMed to search. The date range can be between 2020/01/01 - 2020/12/01. You will enter your desired date range into the **"Please type in a time window for searching in format yyyy/mm/dd - yyyy/mm/dd: eg. 2020/08/29 - 2020/08/30"** search box. It is imperative that your date is formatted as seen in the example. The entered date range must be a range and not a singular date. If these guidelines are not followed, results will not generate.

Email Entry

Next, you will enter your email address into the "***Please type your email address"*** search box. This field is optional. Your email address is used by biopython to note your usage of their package and to send a query to NCBI (National Center for Biotechnology Information) for data. This information will be protected, and you will not receive spam emails or promotions.

CSV File Download

After entering your information into the above fields, the program will find results based on your keyword and date range entries. If there are publications that include your keyword and are within your desired date range, these publications and the details of these publications will be generated into a CSV file. The CSV file will automatically download into your working folder for immediate view and use. The CSV file is user-friendly as it can be opened in most spreadsheet oriented programs (Microsoft Excel, Google Sheets, etc.). The publication details in the CSV file will include publication title, author(s), real publication date, and the abstract of the publication. The real publication date differs from the searched date range as the searched date range is for the publication date on PubMed. If there are no publication results based on your entries, then no file will be downloaded. If a file is not downloaded, you can repeat the process by entering a new keyword.

Publication Search by Author Name

After the CSV file is generated, it can be loaded into a database for query. The program will prompt you to enter an author's name in the ***"Enter an author name, e.g. Nguyen"*** search box. Any publications that have an author with the entered name will appear along with co-authors and the title of the author's publication. If you enter a name that does not match any of the authors, then you will receive the message: ***"Sorry, the author name you entered does not match any authors".*** You will then be prompted to enter another last name to yield a matched result.

Visualization Dashboard and Summary Statistics

You will use the Visualization Dashboard to display graphical tools and summary statistics for the publications generated by your search. This dashboard features tools which allow you to see outputs of your first three rows of your generated CSV file, a breakdown of publication date by year, publication date trends, and the summary statistics of publication numbers by year (mean, median, standard deviation, first quartile, and third quartile) of your search results.