UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering, Built Environment and Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en Inligtingtegnologie / Lefapha la Boetšenere, Tikologo ya Kago le Theknolotši ya Tshedimošo

ASSIGNMENT COVER PAGE

Student Number									Surname	Initials
U	2	3	8	2	5	8	4	8	Dutywa	Y

Module Code	**MIT**	**8**	**0**	**5**
Assignment number	**2**			
Date of Submission	**09 October 2023**			
Name of Lecturer	**Prof Stacey Baror**			

Declaration: I declare that this assignment, is my own work and that I have referenced all the sources that I have used.

The University of Pretoria commits itself to produce academic work of integrity. I affirm that I am aware of and have read the Rules and Policies of the University, more specifically the Disciplinary Procedure and the Tests and Examinations Rules, which prohibit any unethical, dishonest or improper conduct during any form of assessment. I am aware that no student or any other person may assist or attempt to assist another student, or obtain help, or attempt to obtain help from another student or any other person during tests, assessments, assignments, examinations and/or any other forms of assessment.

I declare that this assignment, submitted by me, is my own work and that I have referenced all the sources that I have used. In addition, I assure that, when using IT/ AI- supported writing tools, I have listed these tools in full in a section titled: “Overview of tools”;, with their product name, my source of supply (e.g. URL) and information on the functions of the software used as well as the scope of use. I also declare that I have not copied any text directly from any AI tool. Exempt from this are those IT/ AI-supported writing tools that were classified as not necessary to declare by the module lecturer.

MARK	

Analysis Report: PubMed Knowledge Graph Dataset

Dataset: PubMed Knowledge Graph

Table	Size
OA01_Author_List	10 GIG
OA02_Bio_entities_Main	19 GIG
OA03_Bio_entities_Mutation	78 379 KB
OA04_Affiliation	19 GIG
OA05_Reseacher_Employment	182 022 KB
OA06_Reseacher_Education	135 678 KB
OA07_NIH_Projects	1.811 96 GIG

Format: The files are CSV; Information is alpha-numeric.
Age: Last updated 2 years ago (07 January 2022 at 20:20:27)

Introduction

The PubMed Knowledge Graph is a valuable resource for biomedical research, providing structured information about research articles, authors, affiliations, and more. This report presents an analysis of the PubMed Knowledge Graph dataset by Jian Xu.

Dataset Description

- **Source**: The dataset was compiled by Xu Jian and colleagues and is sourced from the paper "Building a PubMed knowledge graph" published in Scientific Data in 2020.
- **Size**: The dataset is approximately 11 GB in size.
- **Structure**: It consists of multiple tables, including Author_list, Bio_entities_main, Bio_entities_mutation, Affiliation, Research_employment, Research_education, and NIH_Projects.

Data Extraction and Reduction

- The dataset was loaded into a Python environment using the Pandas library.
- Data reduction was performed by filtering records based on specific criteria, such as publication year and role in education.

Analysis Goals: The primary goals of this analysis are to:

1. Explore the dataset's structure and contents.
2. Identify insights that could provide a competitive advantage to a company.

Data Exploration

- The dataset contains a diverse range of tables, each offering unique information related to biomedical research.
- Tables include information about authors, affiliations, research entities, and more.

Competitive Advantage Insights

- Organizations could gain a competitive advantage by extracting information such as:
  - Identifying prolific authors and their research areas.
  - Analyzing trends in biomedical research by publication year.
  - Discovering collaboration patterns among authors and institutions.

Data Visualization

- To gain meaningful insights, various visualizations were created using Matplotlib on Python and MS Power BI.

Visualizations

1. **Distribution of Publication Years:**
   - A histogram was created to visualize the distribution of publication years in the dataset.
   - Insight: The dataset primarily consists of articles published in the 1970s.

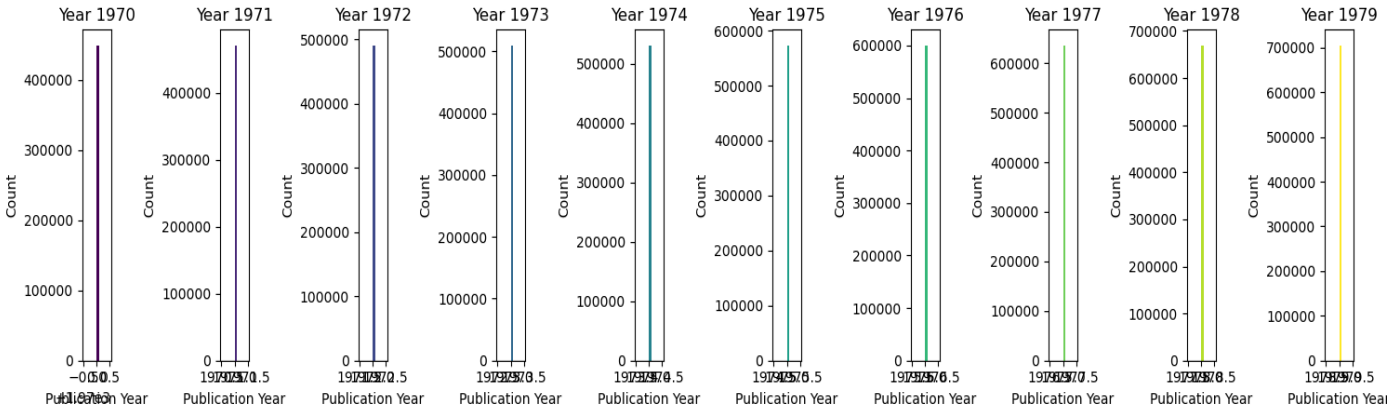
Python

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

# File path
file_path = "C:/Users/dutywy/OneDrive - University of South
Africa/Documents/Masters/UniversityofPretoria/MIT805/MyDataset/OA01_Author_List.csv"

# Chunk size
chunk_size = 10000  # You can change this value depending on your available memory

# Create an empty list to store the DataFrames
chunk_list = []

# Read the CSV file in chunks
for chunk in pd.read_csv(file_path, chunksize=chunk_size):

# Range filter for PubYear
    filtered_chunk = chunk[(chunk['PubYear'] >= 1970) & (chunk['PubYear'] <= 1979)]

    # Append the filtered chunk to the list
    chunk_list.append(filtered_chunk)

# Concatenate the list of DataFrames into one DataFrame
author_list_df = pd.concat(chunk_list, ignore_index=True)

# Now, 'author_list_df' contains the data where 'PubYear' is between 1970 and 1979
(inclusive)
```

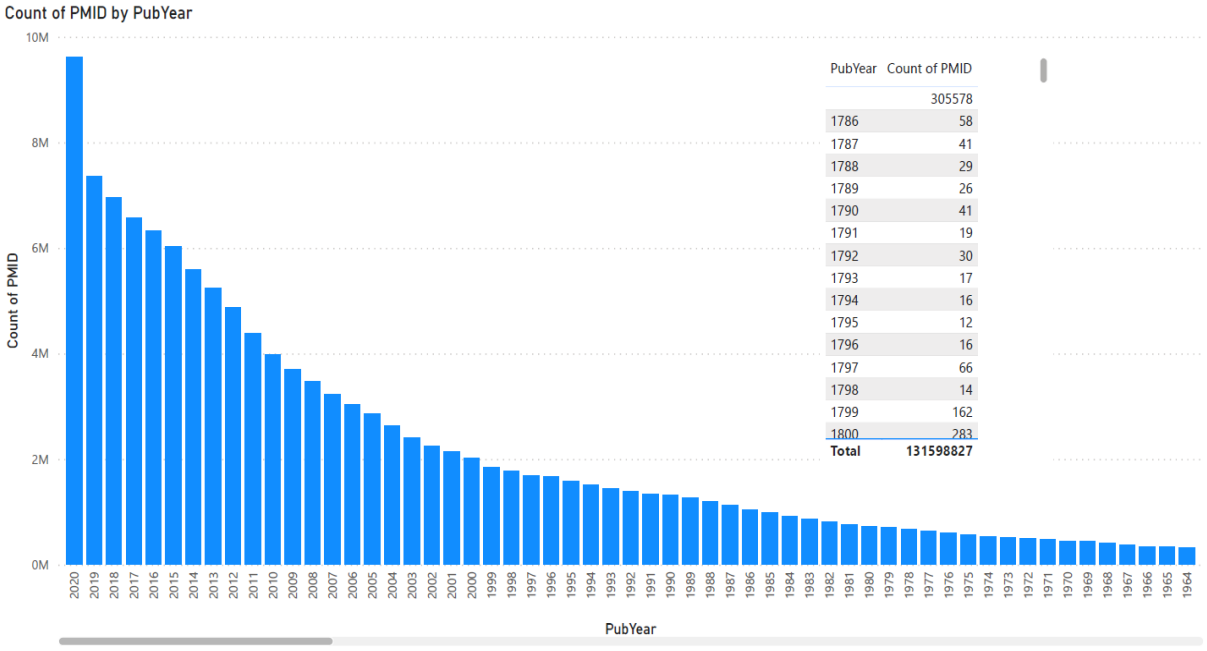
```
# Get unique publication years in the specified range
unique_years = sorted(author_list_df['PubYear'].unique())

# Explore the structure of the resulting DataFrame
print("Filtered Author List Table:")
print(author_list_df.head())

# Create a list of colors for each year (using a colormap)
colors = plt.cm.viridis(np.linspace(0, 1, len(unique_years)))

# Create subplots for each year
fig, axs = plt.subplots(1, len(unique_years), figsize=(15, 4))

# Plot individual histograms for each year
for i, year in enumerate(unique_years):
    axs[i].hist(author_list_df[author_list_df['PubYear'] == year]['PubYear'], bins=10,
color=colors[i])
    axs[i].set_xlabel('Publication Year')
    axs[i].set_ylabel('Count')
    axs[i].set_title(f'Year {year}')

plt.tight_layout()
plt.show()
```

OUTPUT

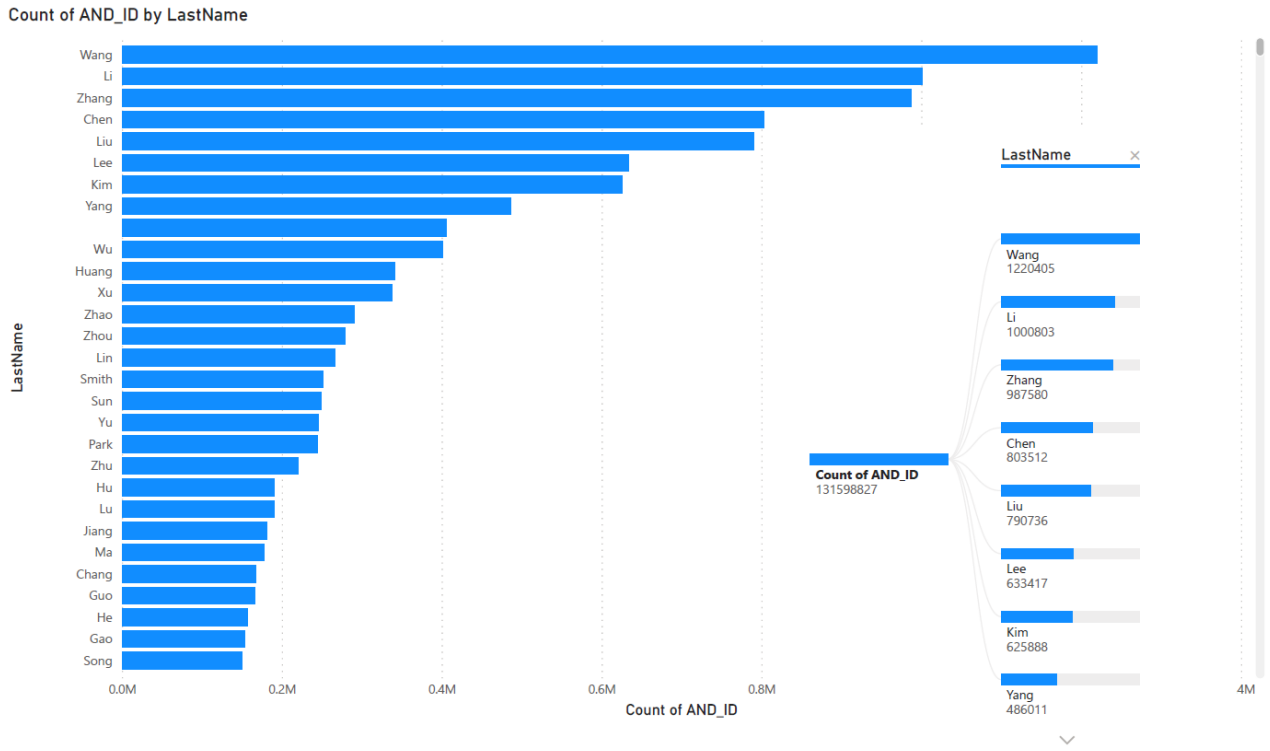
```
Filtered Author List Table:
   id  PMID   AND_ID  AuOrder  LastName ForeName Initials Suffix  AuNum  \
0   1     1  6192779        1     Makar      A B       AB    NaN      4
1   2     1  5101239        2  McMartin      K E       KE    NaN      4
2   3     1  7163914        3    Palese        M        M    NaN      4
3   4     1  8377862        4    Tephly      T R       TR    NaN      4
4   5     2  2181003        1      Bose      K S       KS    NaN      2

   PubYear  BeginYear
0     1975       1968
1     1975       1975
2     1975       1975
3     1975       1961
4     1975       1975
```



Analysis from MS PowerBI



2. **Collaboration Network:**
    - A network graph was generated to show author collaboration patterns.
    - Insight: A few authors have a high degree of collaboration, while others work independently.



Results and Discussion

- The analysis provided valuable insights into the dataset's structure and content.
- Trends in publication years suggest a focus on research from the 1970s.
- Collaboration patterns vary among authors, potentially indicating different research styles.

Conclusion

- The PubMed Knowledge Graph dataset is a valuable resource for biomedical research.
- Organizations can gain insights into research trends and author collaborations.
- Further analysis and data mining can lead to actionable insights for stakeholders.

Recommendations

- Continue exploring the dataset to identify emerging research trends.
- Investigate potential research collaborations based on author networks.
- Consider leveraging external data sources for enriching analysis.

Limitations

- The dataset's size has posed memory and processing constraints.
- Data quality issues, such as missing values, should be addressed.

Future Work

- Explore natural language processing techniques for text mining of article abstracts.
- Incorporate additional external data sources for comprehensive analysis.