

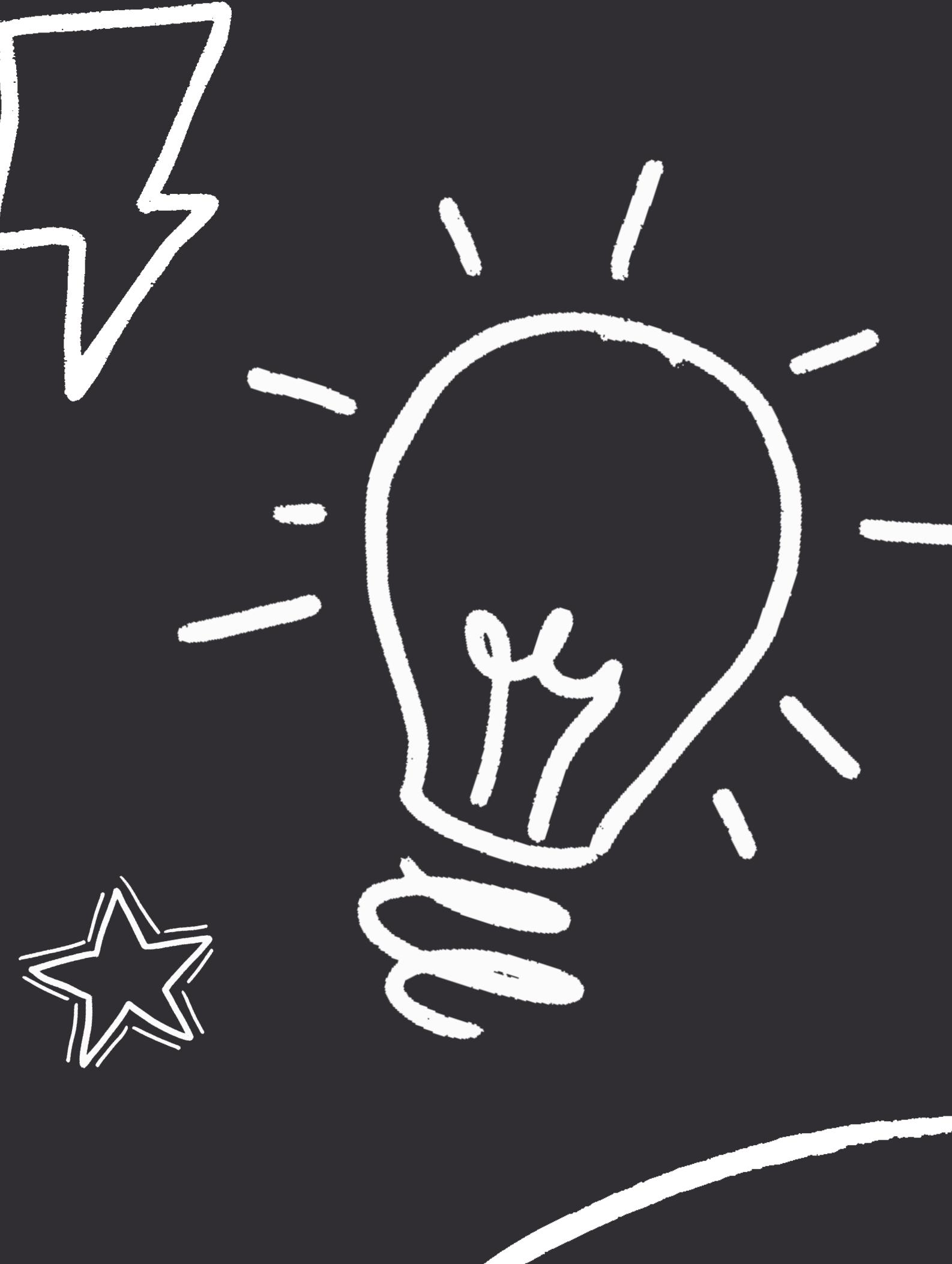
Cyber Security

Presenters:

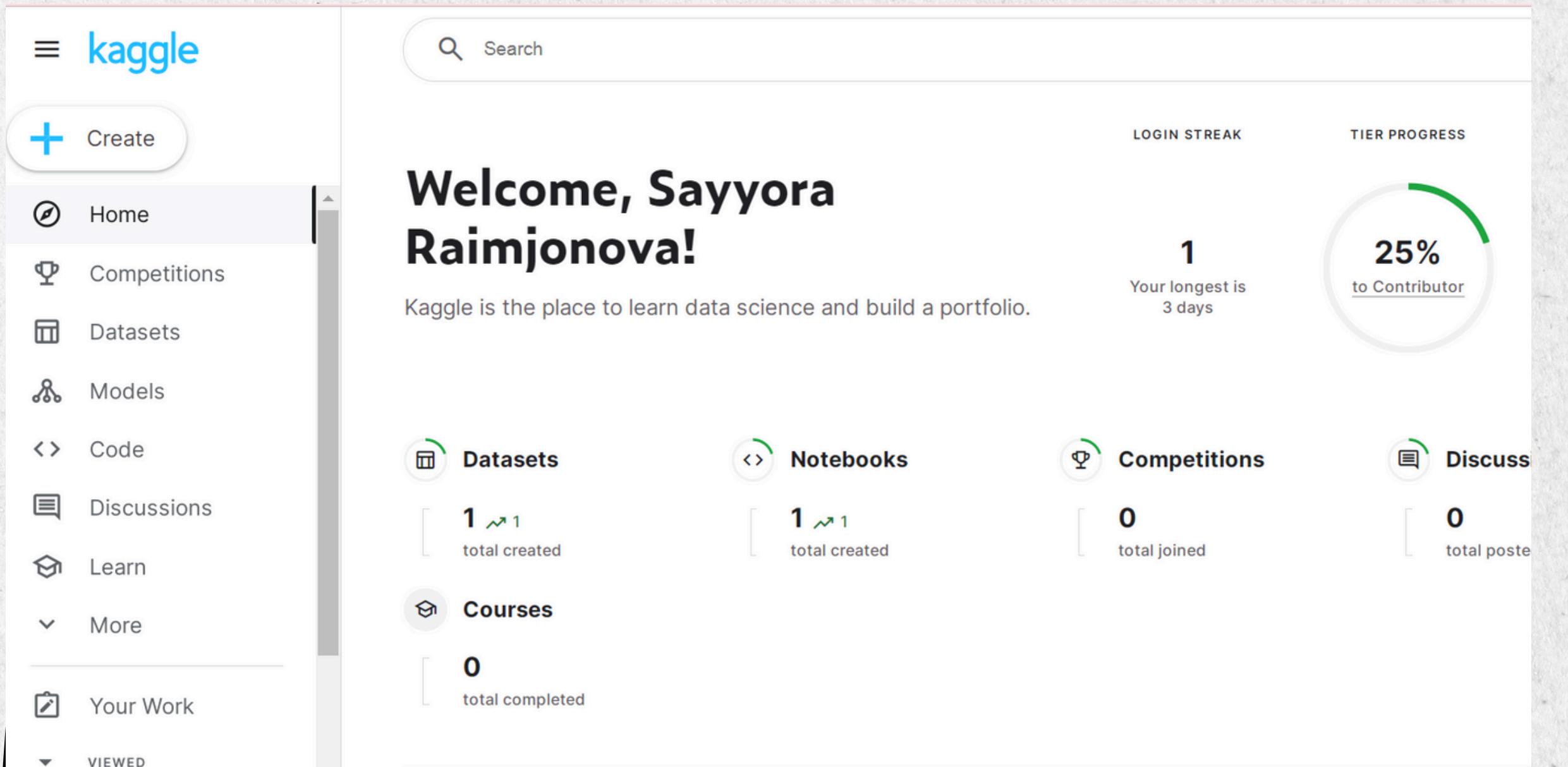
Aibiike, Sayyora, Nadim, Zeesham, Mohammed, Anas,
Yesenia

Date:

19 JULY 2024



KAGGLE



The image shows a screenshot of the Kaggle website. At the top left is a navigation bar with a user icon and the word "kaggle". Below it is a sidebar with links: "Create", "Home", "Competitions", "Datasets", "Models", "Code", "Discussions", "Learn", "More", and "Your Work". A "VIEWED" section is at the bottom of the sidebar. On the right, there's a search bar and a "Login Streak" section showing "1" day and "Your longest is 3 days". A "Tier Progress" circle indicates "25% to Contributor". Below these are four main categories: "Datasets" (1 total created), "Notebooks" (1 total created), "Competitions" (0 total joined), and "Discuss" (0 total posts). There are also sections for "Courses" (0 total completed) and "Your Work".

Kaggle is a website where you can find lots of datasets for different subjects

DATASETS

INCROBO AND 1 COLLABORATOR · UPDATED 9 HOURS AGO

▲ 164 New Notebook Download (5 MB)

Cyber Security Attacks

Consists of 25 varied metrics and 40,000 records

Data Card Code (16) Discussion (2) Suggestions (0)

About Dataset

Welcome to Incrobo's synthetic cyber dataset! Crafted with precision, this dataset offers a realistic representation of travel history, making it an ideal playground for various analytical tasks.

Use the cybersecurity attacks dataset to help you assess the heatmaps, attack signatures, types, and more!

Remember, this is just a sample! If you're intrigued and want access to the complete dataset or have specific

C	D
tre Destination IP A	Source Port
2 84.9.164.252	31225
98 66.191.137.154	17245
198.219.82.17	16811
) 101.228.192.255	20018

Usability 9.41

License Apache 2.0

Update frequency Monthly

A dataset is a collection of data, usually organized in a table or a structured format

ANACONDA



Products

Solutions

Resources

Partn

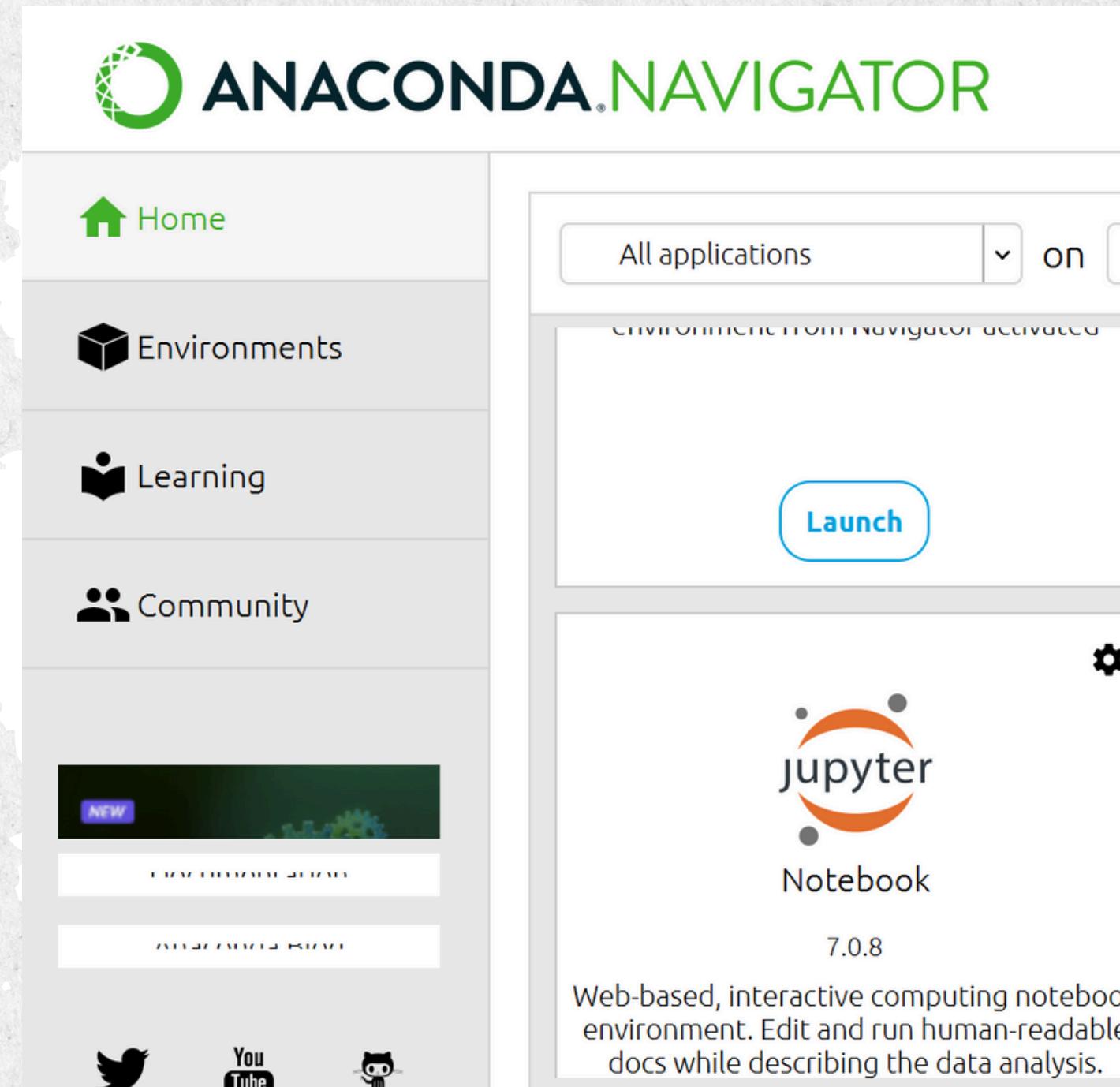
Distribution

Free Download*

Register to get everything you need to get started on your workstation including Cloud Notebooks, Navigator, AI Assistant, Learning and more.

Anaconda is a distribution that makes it easy to use the Python and R programming languages for scientific computing.

JUPYTER NOTEBOOK



Jupyter Notebook: An interactive web-based interface where you can write and execute code, visualize data, and document your analysis in one place.

[4]:

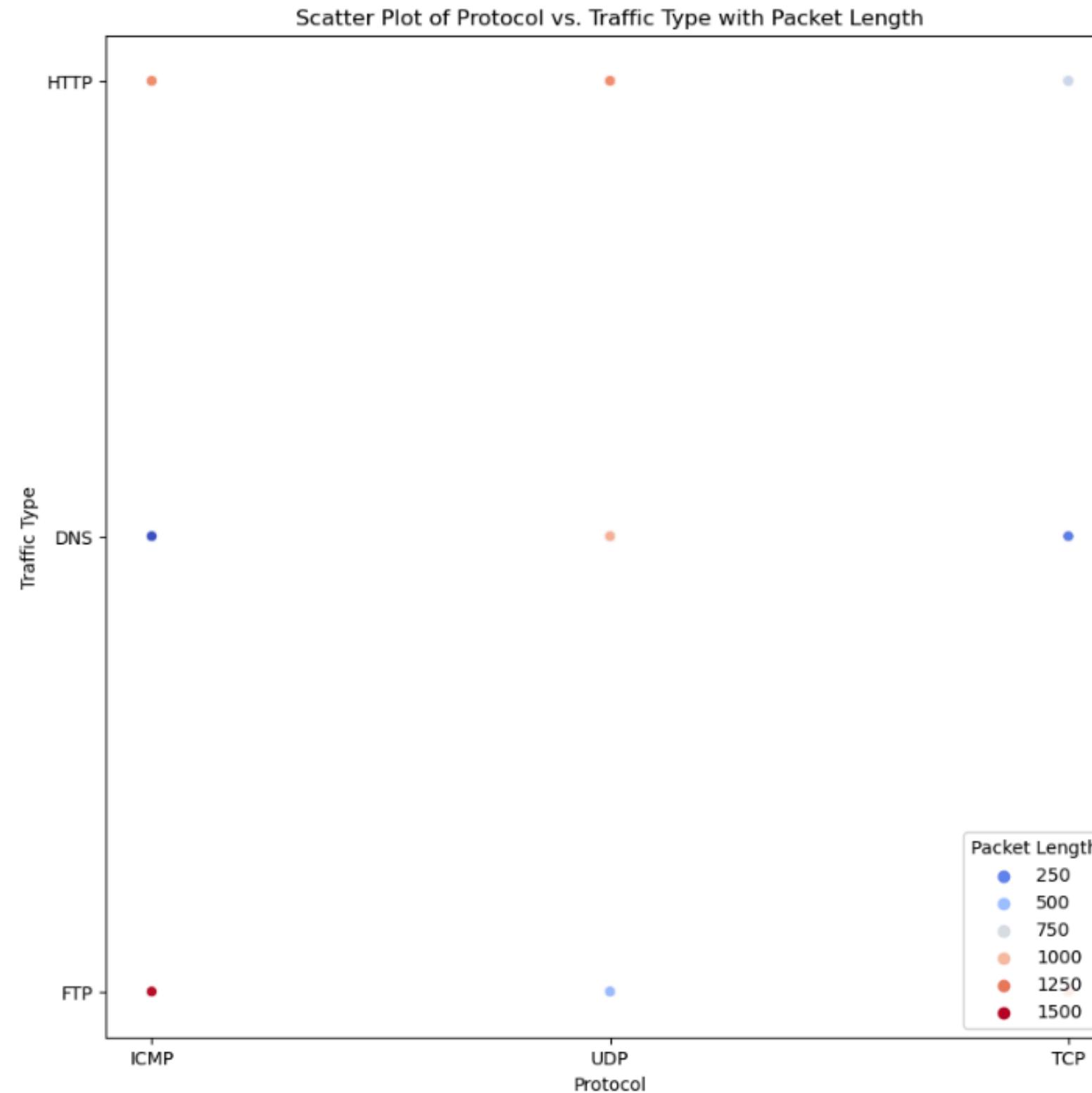
```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40000 entries, 0 to 39999
Data columns (total 25 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   Timestamp        40000 non-null   object  
 1   Source IP Address 40000 non-null   object  
 2   Destination IP Address 40000 non-null   object  
 3   Source Port       40000 non-null   int64  
 4   Destination Port 40000 non-null   int64  
 5   Protocol          40000 non-null   object  
 6   Packet Length     40000 non-null   int64  
 7   Packet Type       40000 non-null   object  
 8   Traffic Type      40000 non-null   object  
 9   Payload Data      40000 non-null   object  
 10  Malware Indicators 20000 non-null   object  
 11  Anomaly Scores    40000 non-null   float64 
 12  Alerts/Warnings   19933 non-null   object  
 13  Attack Type       40000 non-null   object  
 14  Attack Signature   40000 non-null   object  
 15  Action Taken      40000 non-null   object  
 16  Severity Level    40000 non-null   object  
 17  User Information   40000 non-null   object  
 18  Device Information 40000 non-null   object  
 19  Network Segment    40000 non-null   object  
 20  Geo-location Data 40000 non-null   object  
 21  Proxy Information  20149 non-null   object  
 22  Firewall Logs     20039 non-null   object  
 23  IDS/IPS Alerts     19950 non-null   object  
 24  Log Source         40000 non-null   object  
dtypes: float64(1), int64(3), object(21)
memory usage: 7.6+ MB
```

In this project, we analyze 40,000 records of cybersecurity attack data across 25 varied metrics to gain insights and improve our security measures.

```
[5]: plt.figure(figsize=(10, 10))
sns.scatterplot(data=df, x='Protocol', y='Traffic Type', hue='Packet Length', palette='coolwarm')
plt.title('Scatter Plot of Protocol vs. Traffic Type with Packet Length')
plt.xlabel('Protocol')
plt.ylabel('Traffic Type')

# plt.legend(title='Packet Length')
plt.show()
```

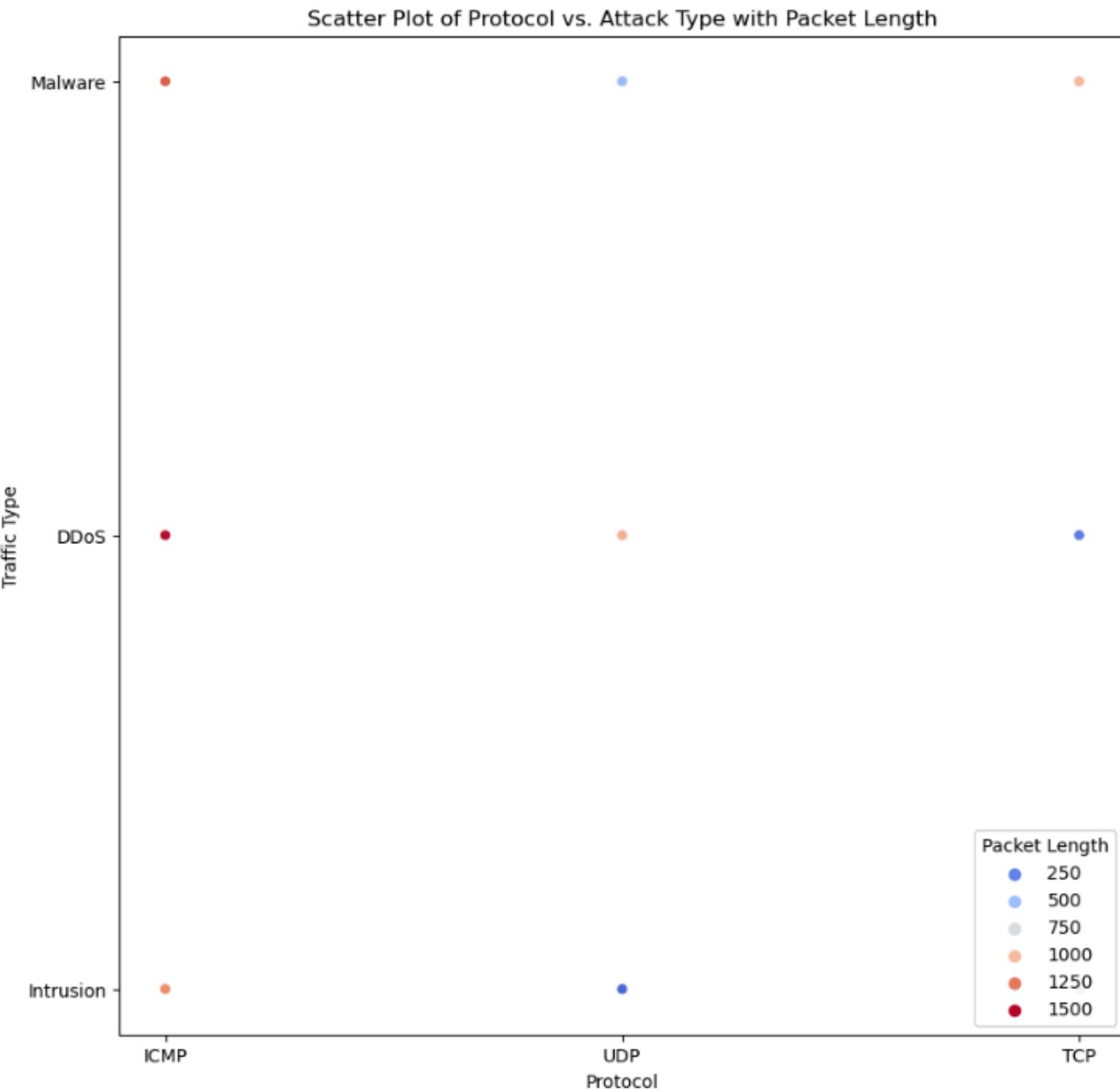


Protocol Analysis and ICMP Dominance:

- Visualization: Scatter Plot of Protocol vs. Traffic Type with Packet Length.
- Insight: The scatter plot revealed that the majority of attacks were conducted using the ICMP (Internet Control Message Protocol) protocol. This protocol is commonly used for network diagnostics.

```
[14]: plt.figure(figsize=(10, 10))
sns.scatterplot(data=df, x='Protocol', y='Attack Type', hue='Packet Length', palette='coolwarm')
plt.title('Scatter Plot of Protocol vs. Attack Type with Packet Length')
plt.xlabel('Protocol')
plt.ylabel('Traffic Type')

# plt.legend(title='Packet Length')
plt.show()
```

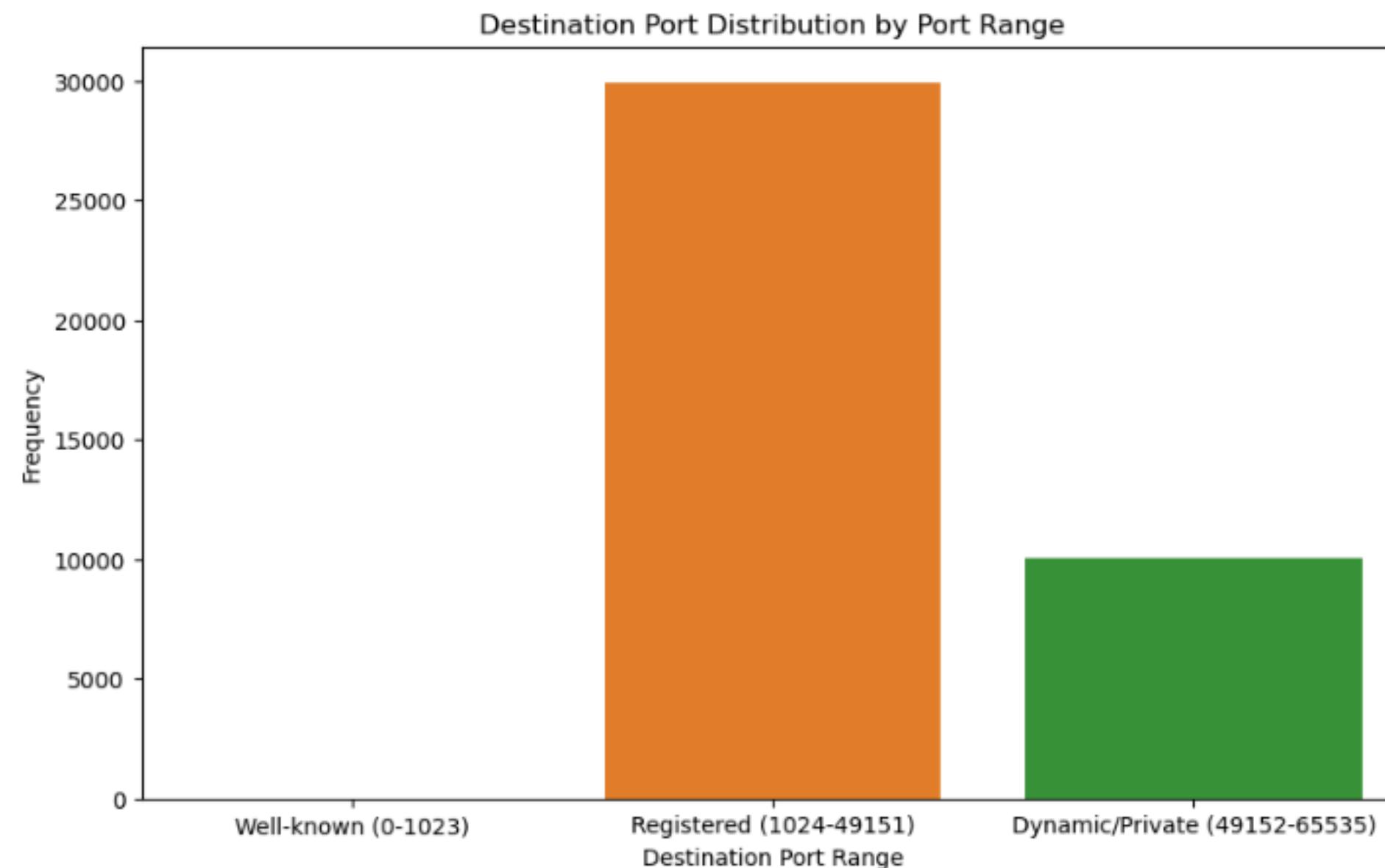


Attack Types and Packet Length:

- Visualization: Scatter Plot of Protocol vs. Attack Type with Packet Length.
- Insight: This highlights that DDoS (Distributed Denial of Service) attacks tend to carry the largest packet lengths and are primarily conducted over the ICMP protocol. This insight underscores the need for robust defenses against DDoS attacks, particularly focusing on ICMP traffic.

```
[13]: df['Destination Port Range'] = df['Destination Port'].apply(categorize_port)

plt.figure(figsize=(10, 6))
sns.countplot(data=df, x='Destination Port Range', order=['Well-known (0-1023)', 'Registered (1024-49151)', 'Dynamic/Private (49152-65535)'])
plt.title('Destination Port Distribution by Port Range')
plt.xlabel('Destination Port Range')
plt.ylabel('Frequency')
plt.show()
```

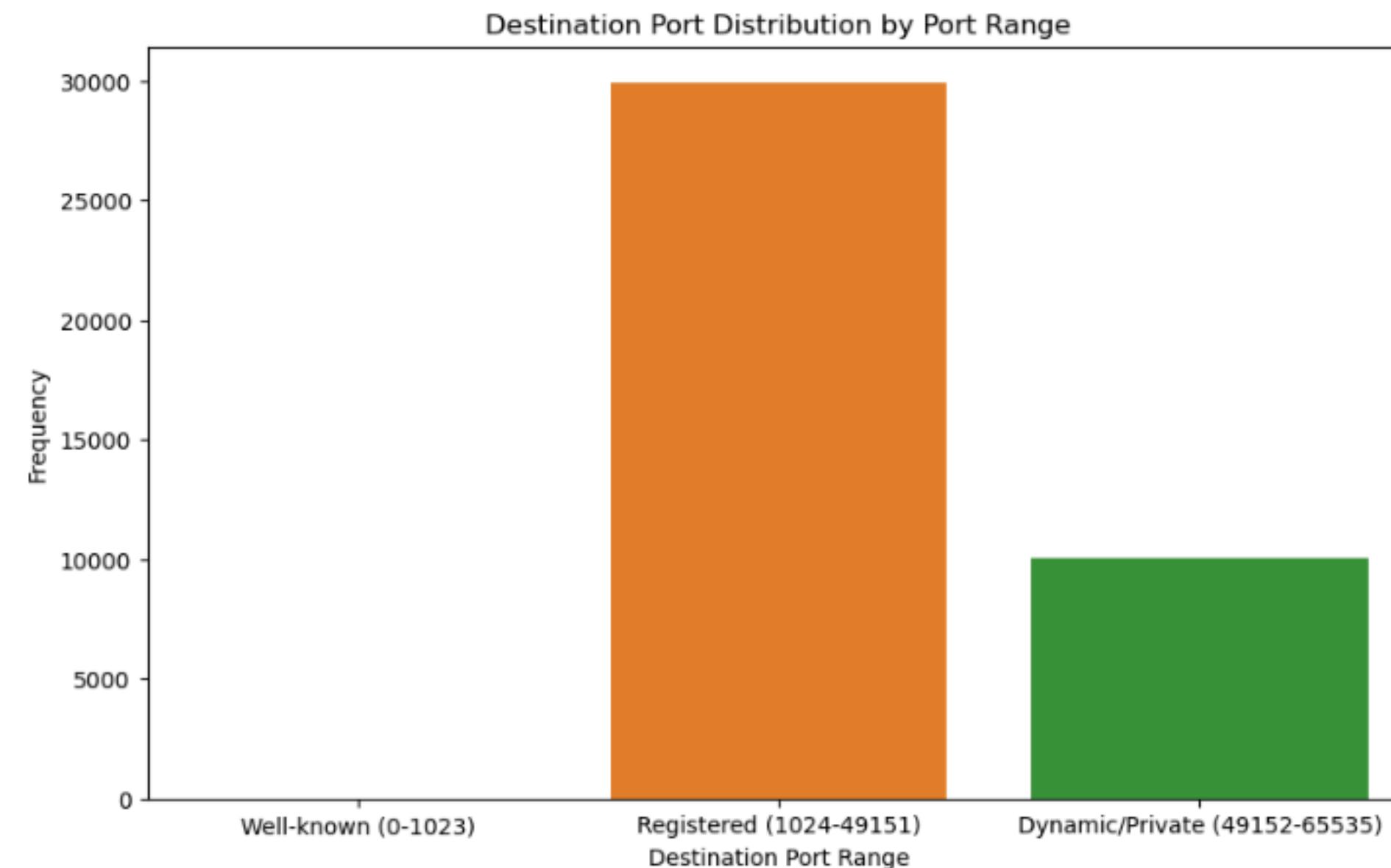


Port Range Analysis:

- Visualization: Histogram of Destination Port Distribution by Port Range.
- Insight: The histogram showed that most attacks target registered ports (1024-49151) and dynamic/private ports (49152-65535). This finding suggests that security measures should be particularly stringent for these port ranges, as they are frequent targets for cyber attacks.

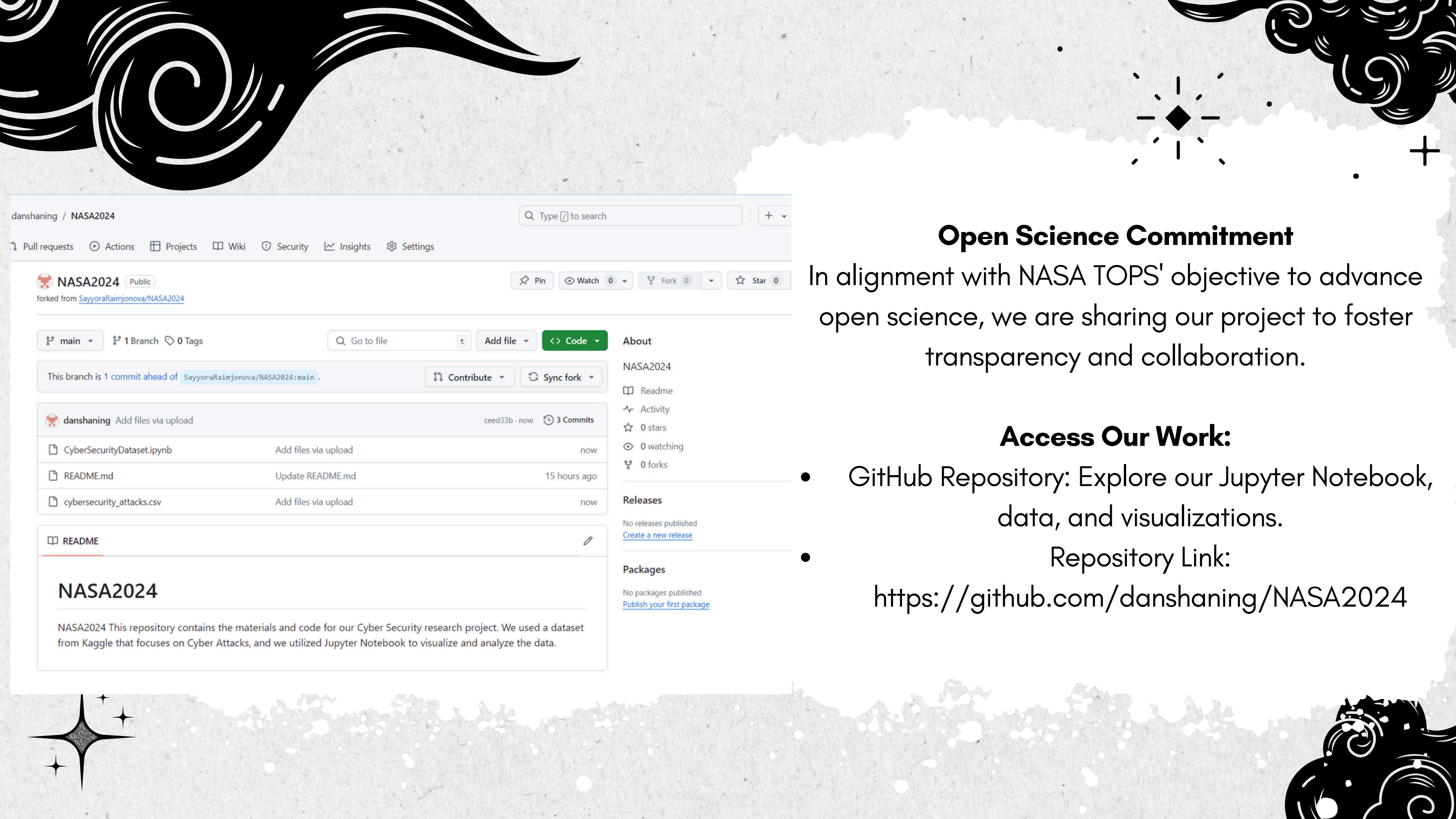
```
[13]: df['Destination Port Range'] = df['Destination Port'].apply(categorize_port)

plt.figure(figsize=(10, 6))
sns.countplot(data=df, x='Destination Port Range', order=['Well-known (0-1023)', 'Registered (1024-49151)', 'Dynamic/Private (49152-65535)'])
plt.title('Destination Port Distribution by Port Range')
plt.xlabel('Destination Port Range')
plt.ylabel('Frequency')
plt.show()
```



Port Range Analysis:

- Visualization: Histogram of Destination Port Distribution by Port Range.
- Insight: The histogram showed that most attacks target registered ports (1024-49151) and dynamic/private ports (49152-65535). This finding suggests that security measures should be particularly stringent for these port ranges, as they are frequent targets for cyber attacks.



Open Science Commitment

In alignment with NASA TOPS' objective to advance open science, we are sharing our project to foster transparency and collaboration.

Access Our Work:

- GitHub Repository: Explore our Jupyter Notebook, data, and visualizations.

Repository Link:

<https://github.com/danshaning/NASA2024>

Project Impact:

- Hands-on Experience: This project provided invaluable hands-on experience with large datasets and the use of tools like Kaggle for data sourcing.
- Skill Development: We enhanced our skills in data analysis and visualization, critical for cybersecurity research and response strategies.
- Collaborative Learning: Working on this project as a team fostered collaboration and improved our ability to analyze complex data sets collectively.