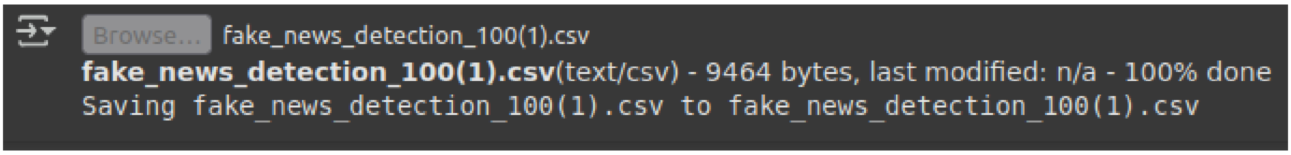
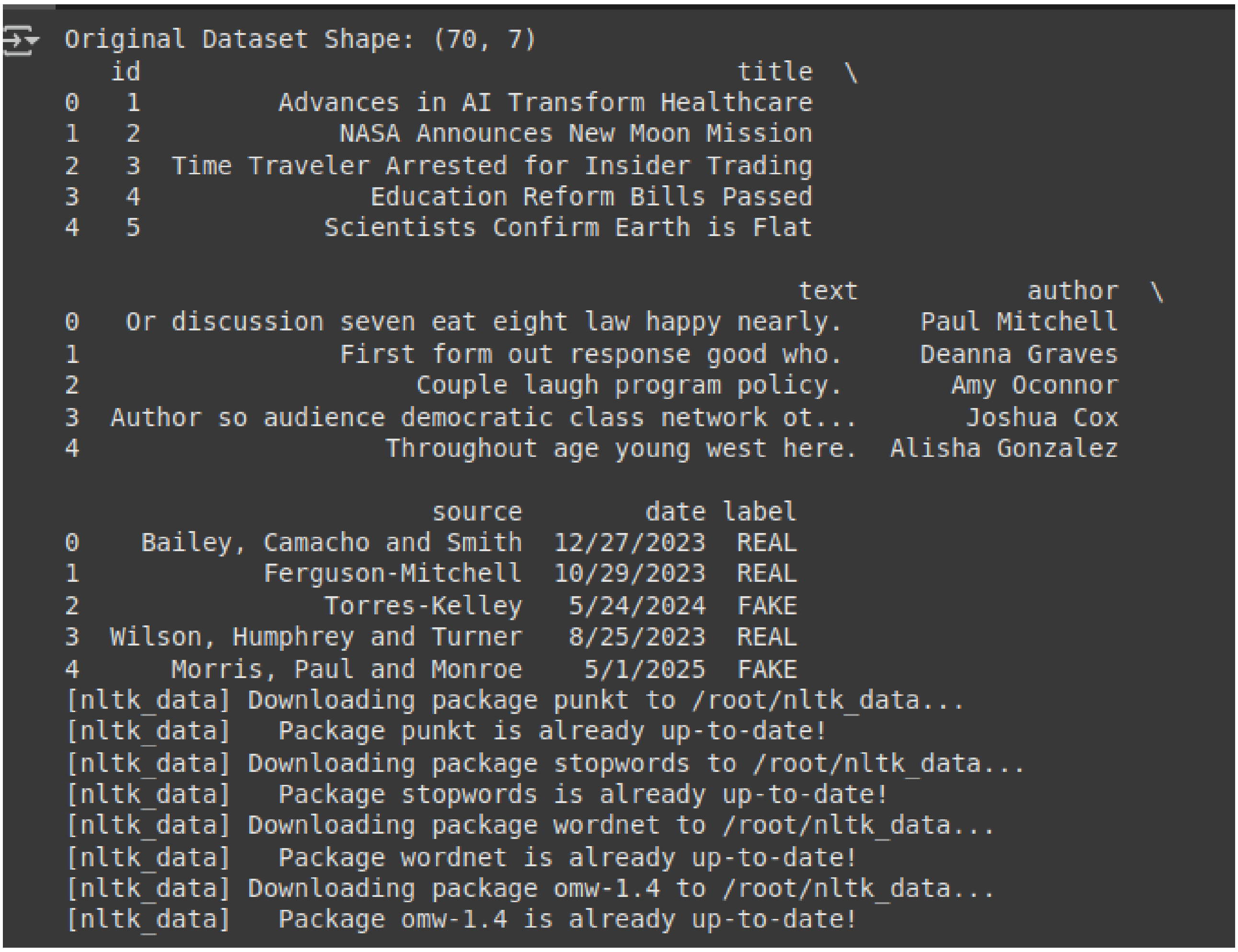
```
from google.colab import files
uploaded = files.upload()
```



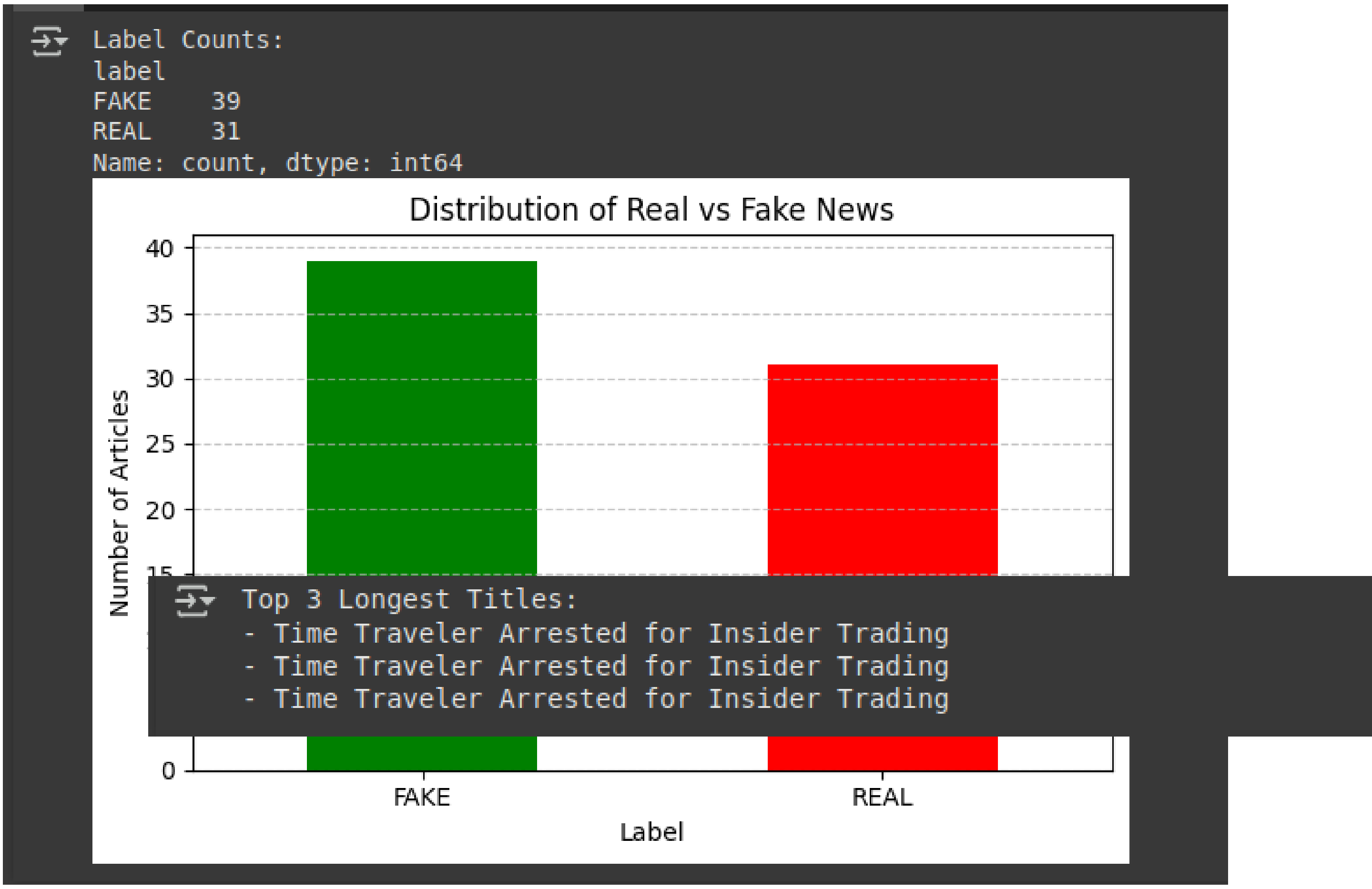
```
fake_news_detection_100(1).csv(text/csv) - 9464 bytes, last modified: n/a - 100% done
Saving fake_news_detection_100(1).csv to fake_news_detection_100(1).csv
```

```
import pandas as pd
import re
import string
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer

# Download NLTK resources (only the first time)
nltk.download('punkt')
nltk.download('stopwords')
nltk.download('wordnet')
nltk.download('omw-1.4')

# Load the dataset
file_path = "fake_news_detection_100(1).csv"
df = pd.read_csv(file_path)

print("Original Dataset Shape:", df.shape)

# Preview the data
print(df.head())

# Initialize lemmatizer and stopwords
lemmatizer = WordNetLemmatizer()
stop_words = set(stopwords.words('english'))

# Function to clean and preprocess text
def clean_text(text):
# Lowercase
text = text.lower()
# Remove punctuation and numbers
text = re.sub(r'[^a-z\s]', '', text)
```



```
Original Dataset Shape: (70, 7)
   id                                  title  \
0   1         Advances in AI Transform Healthcare
1   2             NASA Announces New Moon Mission
2   3   Time Traveler Arrested for Insider Trading
3   4               Education Reform Bills Passed
4   5            Scientists Confirm Earth is Flat

                                              text          author  \
0   Or discussion seven eat eight law happy nearly.     Paul Mitchell
1                 First form out response good who.     Deanna Graves
2                      Couple laugh program policy.       Amy Oconnor
3  Author so audience democratic class network ot...        Joshua Cox
4                    Throughout age young west here.  Alisha Gonzalez

                       source        date label
0    Bailey, Camacho and Smith  12/27/2023  REAL
1            Ferguson-Mitchell  10/29/2023  REAL
2                Torres-Kelley   5/24/2024  FAKE
3  Wilson, Humphrey and Turner   8/25/2023  REAL
4      Morris, Paul and Monroe    5/1/2025  FAKE
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package omw-1.4 to /root/nltk_data...
[nltk_data]   Package omw-1.4 is already up-to-date!
```

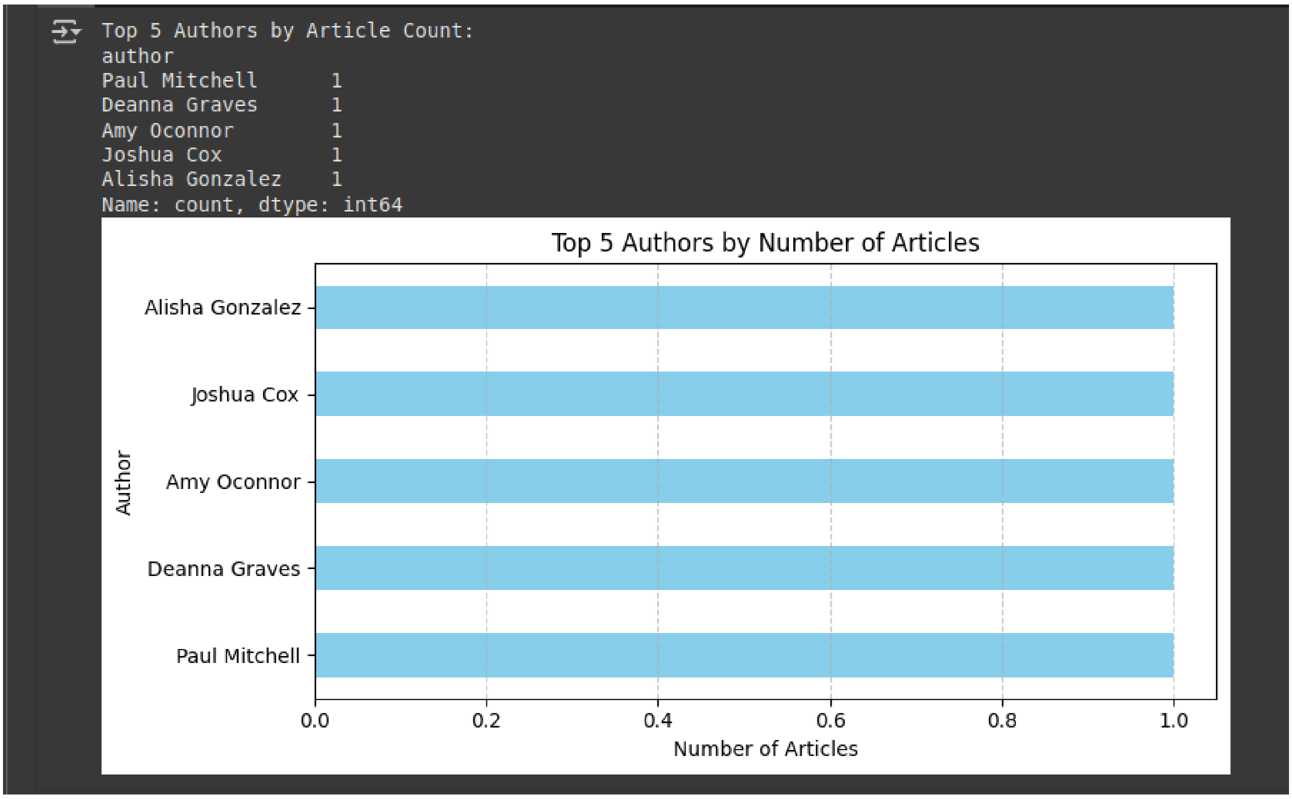
```
import pandas as pd
import matplotlib.pyplot as plt

# Load the dataset
file_path = "fake_news_detection_100(1).csv"
df = pd.read_csv(file_path)

# Count label values
label_counts = df['label'].value_counts()
print("Label Counts:")
print(label_counts)

# Create a bar chart
plt.figure(figsize=(6, 4))
label_counts.plot(kind='bar', color=['green', 'red'])
plt.title('Distribution of Real vs Fake News')
plt.xlabel('Label')
plt.ylabel('Number of Articles')
plt.xticks(rotation=0)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()

# Show the plot
plt.show()
```



```
Label Counts:
label
FAKE    39
REAL    31
Name: count, dtype: int64
```

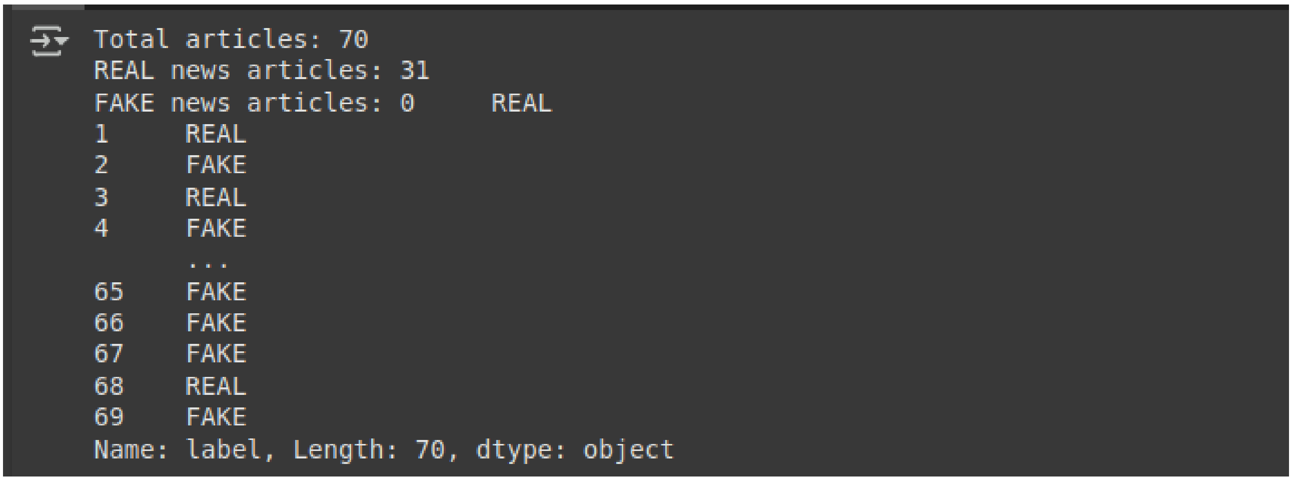
```
Top 3 Longest Titles:
- Time Traveler Arrested for Insider Trading
- Time Traveler Arrested for Insider Trading
- Time Traveler Arrested for Insider Trading
```

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

# Load the dataset
df = pd.read_csv("fake_news_detection_100(1).csv")

# Count articles by author
top_authors = df['author'].value_counts().head(5)
print("Top 5 Authors by Article Count:")
print(top_authors)

# Plot horizontal bar chart
plt.figure(figsize=(8, 4))
top_authors.plot(kind='barh', color='skyblue')
plt.title('Top 5 Authors by Number of Articles')
plt.xlabel('Number of Articles')
plt.ylabel('Author')
plt.grid(axis='x', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()import pandas as pd
df = pd.read_csv("fake_news_detection_100(1).csv")
df['date'] = pd.to_datetime(df['date'])
latest = df.sort_values('date', ascending=False).iloc[0]
print("Most recent article:", latest['title'], "| Date:", latest['date'].date())
```



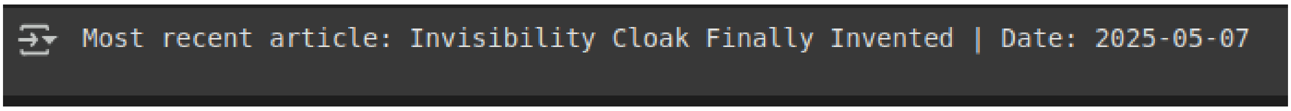
```
Top 5 Authors by Article Count:
author
Paul Mitchell      1
Deanna Graves      1
Amy Oconnor        1
Joshua Cox         1
Alisha Gonzalez    1
Name: count, dtype: int64
```

```
import pandas as pd
df = pd.read_csv("fake_news_detection_100(1).csv")
print("Total articles:", len(df))
print("REAL news articles:", (df['label'] == 'REAL').sum())
print("FAKE news articles:", (df['label']))
```



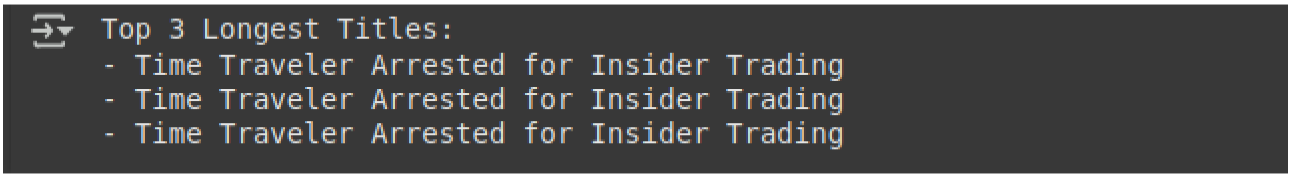
```
Total articles: 70
REAL news articles: 31
FAKE news articles: 0     REAL
1     REAL
2     FAKE
3     REAL
4     FAKE
      ...
65    FAKE
66    FAKE
67    FAKE
68    REAL
69    FAKE
Name: label, Length: 70, dtype: object
```

```
import pandas as pd
df = pd.read_csv("fake_news_detection_100(1).csv")
print("Total articles:", len(df))
print("REAL news articles:", (df['label'] == 'REAL').sum())
print("FAKE news articles:", (df['label']))
```



```
Most recent article: Invisibility Cloak Finally Invented | Date: 2025-05-07
```

```
import pandas as pd
df = pd.read_csv("fake_news_detection_100(1).csv")
df['title_length'] = df['title'].apply(len)
longest_titles = df.sort_values('title_length', ascending=False).head(3)
print("Top 3 Longest Titles:")
for title in longest_titles['title']:
print("-", title)
```



```
Top 3 Longest Titles:
- Time Traveler Arrested for Insider Trading
- Time Traveler Arrested for Insider Trading
- Time Traveler Arrested for Insider Trading
```