```
#Importing necessary libraries
import pandas as pd
import numpy as np
import json


file_name = '/content/users.json'


# Load the JSON file
with open(file_name, 'r') as file:
    data = [json.loads(line) for line in file]


# Convert to DataFrame
df = pd.DataFrame(data)


df.head()
```

	_id	active	createdDate	lastLogin	role	signUpSou
0	{'$oid': '5ff1e194b6a9d73a3a9f1052'}	True	{'$date': 1609687444800}	{'$date': 1609687537858}	consumer	Er
1	{'$oid': '5ff1e194b6a9d73a3a9f1052'}	True	{'$date': 1609687444800}	{'$date': 1609687537858}	consumer	Er
2	{'$oid': '5ff1e194b6a9d73a3a9f1052'}	True	{'$date': 1609687444800}	{'$date': 1609687537858}	consumer	Er

Next steps: Generate code with df View recommended plots

```
# Check for missing values
missing_values = df.isnull().sum()
print("Missing Values:\n", missing_values)
```

```
Missing Values:
 _id              0
active            0
createdDate       0
lastLogin        62
role              0
signUpSource     48
state            56
dtype: int64
```

Double-click (or enter) to edit

```
# Count the total number of rows
total_rows = len(df)

# Count the number of duplicate rows based on '_id'
duplicate_rows = df.duplicated(subset=['_id']).sum()

# Calculate the percentage of duplicate records
percentage_duplicates = (duplicate_rows / total_rows) * 100

print(f"Percentage of Duplicate Records: {percentage_duplicates:.2f}%")
```

```
Percentage of Duplicate Records: 57.17%
```

```
# Check data types
data_types = df.dtypes
print("Data Types:\n", data_types)
```

```
Data Types:
 _id              object
active             bool
createdDate      object
lastLogin        object
role             object
signUpSource     object
```

```
state           object
dtype: object
```