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


file_name = '/content/brands.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	barcode	category	categoryCode	cpg	name	topBrand	
0	{'$oid': '601ac115be37ce2ead437551'}	511111019862	Baking	BAKING	{'$id': {'$oid': '601ac114be37ce2ead437550'}, ...	test brand @1612366101024	False	
1	{'$oid': '601c5460be37ce2ead43755f'}	511111519928	Beverages	BEVERAGES	{'$id': {'$oid': '5332f5fbe4b03c9a25efd0ba'}, ...	Starbucks	False	S
2	{'$oid': '601ac142be37ce2ead43755d'}	511111819905	Baking	BAKING	{'$id': {'$oid': '601ac142be37ce2ead437559'}, ...	test brand @1612366146176	False	B @161
3	{'$oid': '601ac142be37ce2ead43755a'}	511111519874	Baking	BAKING	{'$id': {'$oid': '601ac142be37ce2ead437559'}, ...	test brand @1612366146051	False	B @161
4	{'$oid': '601ac142be37ce2ead43755e'}	511111319917	Candy & Sweets	CANDY_AND_SWEETS	{'$id': {'$oid': '5332fa12e4b03c9a25efd1e7'}, ...	test brand @1612366146827	False	B @161

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
barcode         0
category      155
categoryCode  650
cpg             0
name            0
topBrand      612
brandCode     234
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: 0.00%
```

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

```
Data Types:
 _id             object
barcode         object
category        object
categoryCode    object
cpg             object
name            object
topBrand        object
brandCode       object
dtype: object
```

```
# Listing categorical columns
categorical_columns = ['category', 'categoryCode', 'topBrand', 'brandCode']

# Dictionary to store distinct value counts
distinct_value_counts = {}

# Iterate through each categorical column
for col in categorical_columns:
    # Count distinct values
    distinct_count = df[col].nunique()
    # Store the count in the dictionary
    distinct_value_counts[col] = distinct_count

# Print the results
print("Number of distinct values in each categorical column:")
for col, count in distinct_value_counts.items():
    print(f"{col}: {count}")
```

```
Number of distinct values in each categorical column:
category: 23
categoryCode: 14
topBrand: 2
brandCode: 897
```