Milestone4_DSC540

September 22, 2024

Said Moussadeq

24-May-2024

DSC540, Milestone 4

#

Drug Overdose Death Rates by Demographic Factors

Step 1: Import Libraries and Define API Callg

[1]:
```
import requests
import pandas as pd

# APi Key
api_key = '4NfqOzr4JAEhwg7yYbF9T8bitAthLsciBCk9xz0q'

# API URL
url = 'https://data.cdc.gov/api/views/95ax-ymtc/rows.json?accessType=DOWNLOAD'

# Make the API call
headers = {'X-Api-Key': api_key}
response = requests.get(url, headers=headers)
```

0.0.1 Step 2: Parse JSON Response and Convert to DataFrame

{-}

[2]:
```
# Check if the request was successful
if response.status_code == 200:
    # Parse the JSON response
    data = response.json()

    # Extract the column names and data
    columns = data['meta']['view']['columns']
    column_names = [col['name'] for col in columns]
    records = data['data']

    # Convert to DataFrame
```

```
    df = pd.DataFrame(records, columns=column_names)
else:
    # Handle errors
    print(f"Error: {response.status_code}")
    print(response.json())

display(df.head())
```

```
                sid                                    id  position  \
0  row-w7x6~i8kd~fdci  00000000-0000-0000-091E-F8A8604226D7         0
1  row-6mkc~muk7-9fmw  00000000-0000-0000-F699-11F36B7A415C         0
2  row-wbfi.8jvw_cqn5  00000000-0000-0000-DEE4-DCEF5E21F087         0
3  row-frax~qhwm.qinq  00000000-0000-0000-5C52-A0D010FE654B         0
4  row-gjau_mmtj.87pn  00000000-0000-0000-44D0-F81B01114109         0

   created_at created_meta  updated_at updated_meta meta  \
0  1651187677         None  1651187677         None  { }
1  1651187677         None  1651187677         None  { }
2  1651187677         None  1651187677         None  { }
3  1651187677         None  1651187677         None  { }
4  1651187677         None  1651187677         None  { }

                  INDICATOR                      PANEL  … STUB_NAME  \
0  Drug overdose death rates  All drug overdose deaths  …      Total
1  Drug overdose death rates  All drug overdose deaths  …      Total
2  Drug overdose death rates  All drug overdose deaths  …      Total
3  Drug overdose death rates  All drug overdose deaths  …      Total
4  Drug overdose death rates  All drug overdose deaths  …      Total

  STUB_NAME_NUM   STUB_LABEL STUB_LABEL_NUM  YEAR YEAR_NUM       AGE AGE_NUM  \
0             0  All persons            0.1  1999        1  All ages     1.1
1             0  All persons            0.1  2000        2  All ages     1.1
2             0  All persons            0.1  2001        3  All ages     1.1
3             0  All persons            0.1  2002        4  All ages     1.1
4             0  All persons            0.1  2003        5  All ages     1.1

  ESTIMATE  FLAG
0      6.1  None
1      6.2  None
2      6.8  None
3      8.2  None
4      8.9  None

[5 rows x 23 columns]
```

Step 3: Select Relevant Columns and Make a Copy

[3]:
```
# Select only the specified columns and make a copy to avoid␣
 ↪SettingWithCopyWarning
columns_to_keep = ['INDICATOR', 'PANEL', 'UNIT', 'STUB_NAME', 'STUB_LABEL',␣
 ↪'YEAR', 'AGE', 'ESTIMATE']
df = df[columns_to_keep].copy()

# Display the formatted DataFrame
display(df.head())
```

```
                  INDICATOR                       PANEL  \
0  Drug overdose death rates  All drug overdose deaths
1  Drug overdose death rates  All drug overdose deaths
2  Drug overdose death rates  All drug overdose deaths
3  Drug overdose death rates  All drug overdose deaths
4  Drug overdose death rates  All drug overdose deaths

                                              UNIT STUB_NAME   STUB_LABEL  \
0  Deaths per 100,000 resident population, age-ad…     Total  All persons
1  Deaths per 100,000 resident population, age-ad…     Total  All persons
2  Deaths per 100,000 resident population, age-ad…     Total  All persons
3  Deaths per 100,000 resident population, age-ad…     Total  All persons
4  Deaths per 100,000 resident population, age-ad…     Total  All persons

   YEAR       AGE ESTIMATE
0  1999  All ages      6.1
1  2000  All ages      6.2
2  2001  All ages      6.8
3  2002  All ages      8.2
4  2003  All ages      8.9
```

Step 4: Rename Columns

[4]:
```
# Rename columns
df.rename(columns={
    'INDICATOR': 'Indicator',
    'PANEL': 'Type_Of_Substance',
    'UNIT': 'Unit',
    'STUB_NAME': 'Group',
    'STUB_LABEL': 'Subgroup',
    'ESTIMATE': 'Death_Rate'
}, inplace=True)


# Display the renamed Columns
display(df.head())
```

```
                   Indicator         Type_Of_Substance  \
0  Drug overdose death rates  All drug overdose deaths
1  Drug overdose death rates  All drug overdose deaths
2  Drug overdose death rates  All drug overdose deaths
3  Drug overdose death rates  All drug overdose deaths
4  Drug overdose death rates  All drug overdose deaths

                                                 Unit  Group     Subgroup  \
0  Deaths per 100,000 resident population, age-ad…  Total  All persons
1  Deaths per 100,000 resident population, age-ad…  Total  All persons
2  Deaths per 100,000 resident population, age-ad…  Total  All persons
3  Deaths per 100,000 resident population, age-ad…  Total  All persons
4  Deaths per 100,000 resident population, age-ad…  Total  All persons

   YEAR       AGE Death_Rate
0  1999  All ages        6.1
1  2000  All ages        6.2
2  2001  All ages        6.8
3  2002  All ages        8.2
4  2003  All ages        8.9
```

5. Handle Missing Values

[5]:
```
# Handle missing values by dropping rows with missing 'Death_Rate'
df.dropna(subset=['Death_Rate'], inplace=True)

# Display the DataFrame after handling missing values
display(df.head())
```

```
                   Indicator         Type_Of_Substance  \
0  Drug overdose death rates  All drug overdose deaths
1  Drug overdose death rates  All drug overdose deaths
2  Drug overdose death rates  All drug overdose deaths
3  Drug overdose death rates  All drug overdose deaths
4  Drug overdose death rates  All drug overdose deaths

                                                 Unit  Group     Subgroup  \
0  Deaths per 100,000 resident population, age-ad…  Total  All persons
1  Deaths per 100,000 resident population, age-ad…  Total  All persons
2  Deaths per 100,000 resident population, age-ad…  Total  All persons
3  Deaths per 100,000 resident population, age-ad…  Total  All persons
4  Deaths per 100,000 resident population, age-ad…  Total  All persons

   YEAR       AGE Death_Rate
0  1999  All ages        6.1
1  2000  All ages        6.2
2  2001  All ages        6.8
3  2002  All ages        8.2
```

```
4  2003  All ages        8.9
```

Step 6: Convert Data Types

[6]:
```
# Convert data types
df['Year'] = df['YEAR'].astype(int)
df['Death_Rate'] = df['Death_Rate'].astype(float)
```

0.0.2 Step 7: Drop Redundant Columns

{-}

[7]:
```
# Drop the original 'YEAR' column after conversion
df.drop(columns=['YEAR'], inplace=True)
```

Step 8: Display Cleaned DataFrame

[8]:
```
# Display the cleaned DataFrame
print(df.head())
```

```
                  Indicator         Type_Of_Substance  \
0  Drug overdose death rates  All drug overdose deaths
1  Drug overdose death rates  All drug overdose deaths
2  Drug overdose death rates  All drug overdose deaths
3  Drug overdose death rates  All drug overdose deaths
4  Drug overdose death rates  All drug overdose deaths

                                                Unit  Group     Subgroup  \
0  Deaths per 100,000 resident population, age-ad…  Total  All persons
1  Deaths per 100,000 resident population, age-ad…  Total  All persons
2  Deaths per 100,000 resident population, age-ad…  Total  All persons
3  Deaths per 100,000 resident population, age-ad…  Total  All persons
4  Deaths per 100,000 resident population, age-ad…  Total  All persons

        AGE  Death_Rate  Year
0  All ages         6.1  1999
1  All ages         6.2  2000
2  All ages         6.8  2001
3  All ages         8.2  2002
4  All ages         8.9  2003
```

1 Summary

For the seek of data extraction and data cleaning, we fetched data from the CDC API using an API key and converted the JSON response into a DataFrame. We selected relevant columns (`INDICATOR`, `PANEL`, `UNIT`, `STUB_NAME`, `STUB_LABEL`, `YEAR`, `AGE`, `ESTIMATE`) and made a copy to avoid `SettingWithCopyWarning`. We then renamed these columns to more descriptive names for clarity (`Indicator`, `Drug_Type`, `Unit`, `Group`, `Subgroup`, `Death_Rate`). To ensure data quality,

we handled missing values by dropping rows with missing `Death_Rate` and converted the `Year` and `Death_Rate` columns to appropriate data types (integer and float, respectively). Finally, we removed the redundant `YEAR` column after conversion. These steps were taken to clean and standardize the data, making it ready for analysis by ensuring that it is consistent, complete, and properly formatted. The data was originally outsourced from the CDC, and compliance with legal and regulatory guidelines like HIPAA is a must to ensure privacy and proper data usage. Data transformations, such as dropping rows with missing Death_Rate values, carry the risk of biasing results, and assumptions were made about the validity of year values and the impact of missing data. Ethical acquisition are involved such as adhering to terms of use and ensuring no personal data misuse. Mitigating ethical implications includes maintaining transparency, responsible usage for public health, compliance with legal guidelines, and engagement to align with ethical standards.

[9]:
```
import sqlite3
import pandas as pd

# Assuming the DataFrame 'df' is already created and cleaned as described␣
 ↪earlier.

# Create a SQLite database connection
conn = sqlite3.connect('drug_overdose_deaths.db')
cursor = conn.cursor()

# Convert the DataFrame to SQL
df.to_sql('overdose_deaths', conn, if_exists='replace', index=False)

# Verify by querying the first few rows from the table
cursor.execute("SELECT * FROM overdose_deaths LIMIT 5")
rows = cursor.fetchall()

# Print the rows
for row in rows:
    print(row)

# Close the connection
conn.close()
```

```
('Drug overdose death rates', 'All drug overdose deaths', 'Deaths per 100,000
resident population, age-adjusted', 'Total', 'All persons', 'All ages', 6.1,
1999)
('Drug overdose death rates', 'All drug overdose deaths', 'Deaths per 100,000
resident population, age-adjusted', 'Total', 'All persons', 'All ages', 6.2,
2000)
('Drug overdose death rates', 'All drug overdose deaths', 'Deaths per 100,000
resident population, age-adjusted', 'Total', 'All persons', 'All ages', 6.8,
2001)
('Drug overdose death rates', 'All drug overdose deaths', 'Deaths per 100,000
resident population, age-adjusted', 'Total', 'All persons', 'All ages', 8.2,
2002)
```

```
('Drug overdose death rates', 'All drug overdose deaths', 'Deaths per 100,000
resident population, age-adjusted', 'Total', 'All persons', 'All ages', 8.9,
2003)
```

[ ]: