U.S. Population Trends - 2020 to 2022 Analysis

In this project, I analyzed U.S. state population estimates from 2020 to 2022. The analysis uncovers population trends, identifies states with significant changes, and presents the findings in an actionable format.

The workflow demonstrates key data analysis skills, including data cleaning, transformation, exploratory analysis, and insights generation. The data is from the US Census.

0.1 Data Preparation

Convert FIP codes to State Abbreviation

[3]:
```
import os
import pandas as pd

data_dir = "/Users/misbaharshad/Downloads/176818886/"
filename = "NST-EST2022-ALLDATA.csv"
abs_path = os.path.join(data_dir, filename)

df = pd.read_csv(abs_path)
```

[4]:
```
import us

def fip_to_abb(fips_code):
    fips_str = str(fips_code).zfill(2)
    if fips_code == 0:
        return "US"
    elif fips_code == 72:
        return "PR"
    elif fips_code in fips_to_abbrev:
        return fips_to_abbrev[fips_code]
    else:
        state = us.states.lookup(fips_str)
        if state is not None:
            return state.abbr

fips_to_abbrev = {
    0: "US", 72: "PR", 1: "AL", 2: "AK", 4: "AZ", 5: "AR", 6: "CA", 8: "CO",
    9: "CT", 10: "DE", 11: "DC", 12: "FL", 13: "GA", 15: "HI", 16: "ID",
    17: "IL", 18: "IN", 19: "IA", 20: "KS", 21: "KY", 22: "LA", 23: "ME",
```

```
    24: "MD", 25: "MA", 26: "MI", 27: "MN", 28: "MS", 29: "MO", 30: "MT",
    31: "NE", 32: "NV", 33: "NH", 34: "NJ", 35: "NM", 36: "NY", 37: "NC",
    38: "ND", 39: "OH", 40: "OK", 41: "OR", 42: "PA", 44: "RI", 45: "SC",
    46: "SD", 47: "TN", 48: "TX", 49: "UT", 50: "VT", 51: "VA", 53: "WA",
    54: "WV", 55: "WI", 56: "WY"
}

df['STATE'] = df['STATE'].map(fip_to_abb)
```

0.2 High Level Exploratory Data Analysis (EDA)

There are **66 rows and 44 columns**. The rows represent states, regions, or divisions in the U.S. and columns include features like population estimates, natural changes, migration rates, etc.

Categorical Columns: Region, Division, and State. Numerical Columns: Most of the columns are numeric and related to demographic statistics.

Total U.S. Population: 333M in 2022 (increased from 2020 to 2022)

Key columns:

- POPESTIMATE2020, POPESTIMATE2021, POPESTIMATE2022: Population estimates over years.
- NPOPCHG_2020, NPOPCHG_2021: Absolute population changes.
- RNETMIG2021, RNETMIG2022: Net migration rates.
- RDEATH2021, RNATURALCHG2021: Death rates and natural changes (births - deaths).

print(df.head) print(df.shape) print(df.describe) print(df.dtypes)

How Did State Populations Change from 2020 - 2022?

[5]:
```
df = df[[c for c in df.columns if c.endswith('2020') or ('2021') or ('2022')]]
df = df.loc[:, ['STATE'] + [c for c in df.columns if c.
  ↪startswith('POPESTIMATE')]]
```

Filter Top 10 States

[6]:
```
df_top_states = df[['STATE', 'POPESTIMATE2021']].
  ↪sort_values(by='POPESTIMATE2021', ascending=False).head(10)
print(df_top_states)
```

```
   STATE  POPESTIMATE2021
0     US        332031554
7     US        127346029
11    US         78589763
4     US         68836505
8     US         66666348
1     US         57259257
13    US         53321373
5     US         47181948
```

```
3     US          42137512
10    US          41205309
```

41 states saw an increase in the population from 2020 to 2022 and 25 states saw a decrease.

[7]:
```
def pop_changes22(x):
    return x['POPESTIMATE2022'] - x['POPESTIMATE2020']

df['POPCHANGE'] = df.apply(pop_changes22, axis=1)

gained = (df['POPCHANGE'] > 0).sum()
lost = (df['POPCHANGE'] < 0).sum()

print('number of states that gained population from 2020 to 2022:', gained)
print('number of states that lost population from 2020 to 2022:', lost)
```

```
number of states that gained population from 2020 to 2022: 41
number of states that lost population from 2020 to 2022: 25
```

Four states saw a change of less than 1,000 people in their population from 2020 - 2022: Arkansas, Washington D.C., Kansas, and North Dakota.

Arkansas and D.C. saw an increase, whereas Kansas and North Dakota saw a decrease in population.

[8]:
```
df[df['POPCHANGE'].abs() < 1000]
```

[8]:

	STATE	POPESTIMATE2020	POPESTIMATE2021	POPESTIMATE2022	POPCHANGE
15	AK	732923	734182	733583	660
22	DC	670868	668791	671803	935
30	KS	2937919	2937922	2937150	-769
48	ND	779518	777934	779261	-257

[9]:
```
def popchange_zscore(df):
    popchange_mean = df['POPCHANGE'].mean()
    popchange_std = df['POPCHANGE'].std()
    zscore = (df['POPCHANGE'] - popchange_mean) / popchange_std
    return df[(zscore.abs() > 1)][['STATE', 'POPCHANGE']].
 ↪sort_values(by='POPCHANGE', ascending=False)

popchange_zscore(df)
```

[9]:

	STATE	POPCHANGE
7	US	2265579
0	US	1776045
8	US	1288139
10	US	822005
57	TX	797098
23	FL	655221
1	US	-408492
13	US	-425401

```
46    NY    -431145
3     US    -463567
18    CA    -472311
```

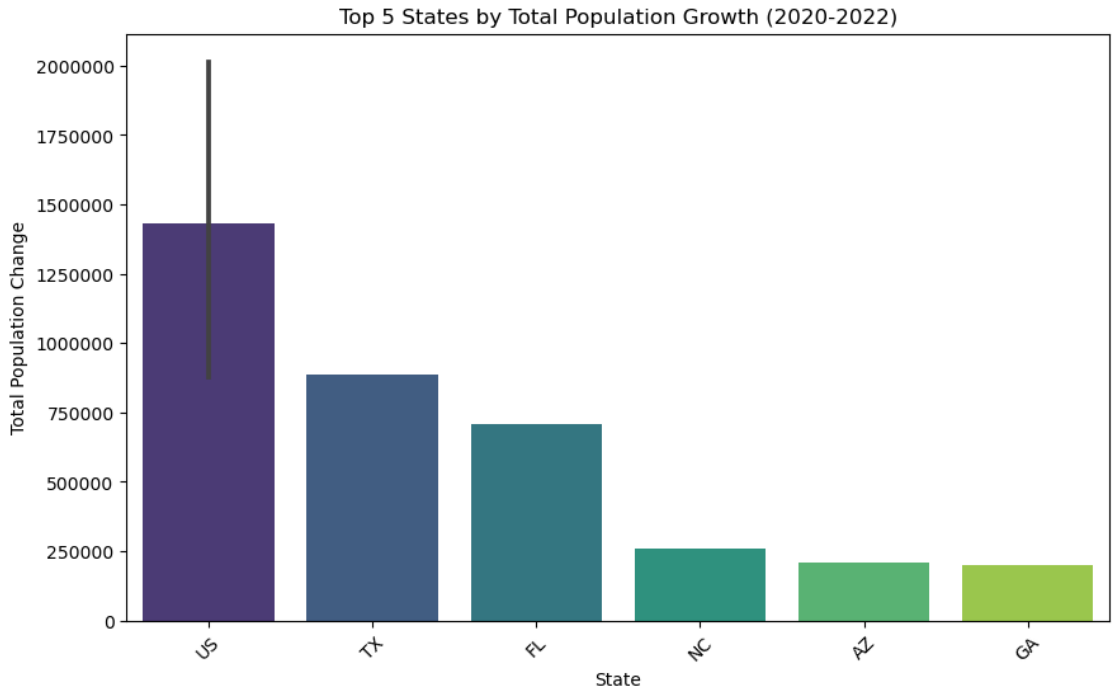
0.3 Visualizing the Change

Which States Saw a Population Increase? The population of Texas grew more from 2020 to 2022 than any other state in the county, by almost 1 million people.

[10]:
```
import matplotlib.pyplot as plt
import seaborn as sns
import matplotlib as mpl
```

[11]:
```
df = pd.read_csv(abs_path)
df['STATE'] = df['STATE'].map(fip_to_abb)
```

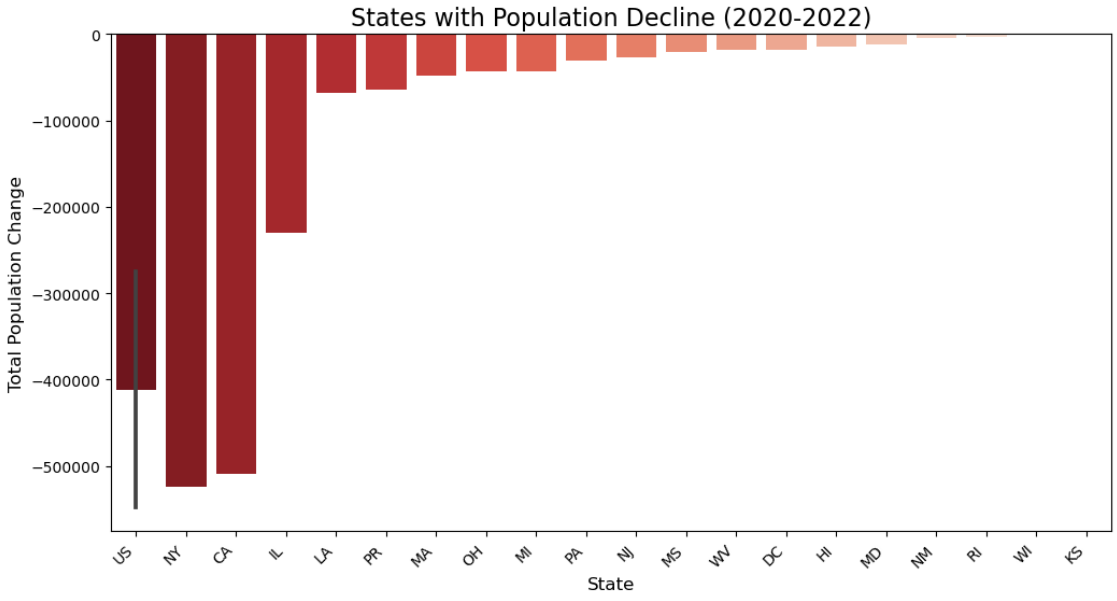
[12]:
```
df['Total_Pop_Change'] = df[['NPOPCHG_2020', 'NPOPCHG_2021', 'NPOPCHG_2022']].
 ↪sum(axis=1)

top_states = df.sort_values('Total_Pop_Change', ascending=False).head(10)

plt.figure(figsize=(10, 6))
sns.barplot(data=top_states, x='STATE', y='Total_Pop_Change', palette='viridis')
plt.title('Top 5 States by Total Population Growth (2020-2022)')
plt.ylabel('Total Population Change')
plt.xlabel('State')
plt.xticks(rotation=45)

plt.ticklabel_format(style='plain', axis='y')
plt.show()
```



Which States Saw a Population Decline? The top states that saw a major decline in population were New York and California, followed by Illinois by less than half the decline.

[13]:
```
df['Total_Pop_Change'] = df[['NPOPCHG_2020', 'NPOPCHG_2021', 'NPOPCHG_2022']].
 ↪sum(axis=1)

decline_data = df[df['Total_Pop_Change'] < 0]
```

[14]:
```
decline_data = decline_data.sort_values('Total_Pop_Change')

plt.figure(figsize=(12, 6))
sns.barplot(data=decline_data, x='STATE', y='Total_Pop_Change',␣
 ↪palette='Reds_r')

plt.title('States with Population Decline (2020-2022)', fontsize=16)
plt.ylabel('Total Population Change', fontsize=12)
plt.xlabel('State', fontsize=12)
plt.xticks(rotation=45, ha='right')

plt.ticklabel_format(style='plain', axis='y')
plt.show()
```



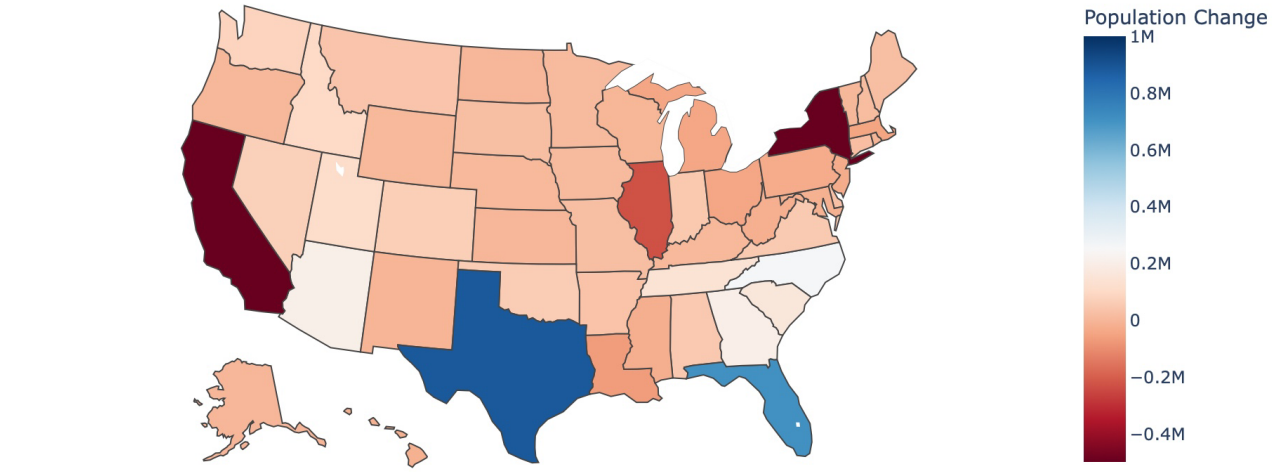
0.4 Analysis of Population Change

An interesting trend emerging from the data reveals that states with high population density and living costs, particularly in urban centers (i.e. New York and California), experienced the steepest population declines. Conversely, states with more affordable living costs saw significant population increases.

This pattern aligns closely with real-world shifts triggered by the COVID-19 pandemic. The pandemic spurred a rise in remote work and a reduction in public interaction, prompting many to relocate to less densely populated and more affordable areas. Further research could explore how these COVID-era migration patterns have reshaped population distributions.

[16]:
```
import plotly.express as px

state_pop_changes = df[['STATE', 'Total_Pop_Change']]
# Create the map with adjusted scale
fig = px.choropleth(
    state_pop_changes,
    locations='STATE',
    locationmode="USA-states",
    color='Total_Pop_Change',
    color_continuous_scale="RdBu",
    range_color=[-500000, 1000000],  # Adjusting the range from -500k to 1M
    title="Net Population Change by State (2020-2022)",
    scope="usa",
    labels={'Total_Pop_Change': 'Population Change'}
)
```

```
fig.show()
```


Net Population Change by State (2020-2022)
Population Change
1M
0.8M
0.6M
0.4M
0.2M
0
−0.2M
−0.4M