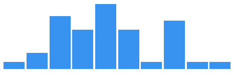
Deepnote Work / Deaths_in_population Published at Nov 29, 2022 Unlisted

```
import numpy as np
import pandas as pd
```

```
deaths_df = pd.read_csv('data/da5e144e-7525-4f18-9def-833f7ced4994.csv')
population_df = pd.read_csv('data/co-est2020.csv')
cig_df = pd.read_csv('Cig_Data2.csv').loc[:57,:'Cig_Rates']
cig_df['County'] = cig_df['County'].str.strip()
cig_df
#population_df
#deaths_df
```

	County object	Cig_Rates float64
	Alameda .... 1.7% Alpine .... 1.7% 56 others 96.6%	0.06 - 0.25 
0	Alameda	0.1
1	Alpine	0.15
2	Amador	0.15
3	Butte	0.17
4	Calaveras	0.15
5	Colusa	0.2
6	Contra Costa	0.12
7	Del Norte	0.2
8	El Dorado	0.15
9	Fresno	0.15

- EACH ROW REPRESENTS A PART OF THE POPULATION WHO DIED IN A COUNTY
- Year: Year of death
- County: Name of County where the people either died or were residents from depending on

Geography type

- Geography_Type:

Residence: California residents even if death happened outside California

Occurrence: happened in California even if people were non-residents

- Strata: General demographic or category. Total Population represents the whole population in County
- Strata_Name: Specific General demographic or category people were categorized as.
- Cause: ALL = All causes, rest are causes of death
- Cause_Desc: All causes (total) if cause is ALL, rest are descriptions of cause of death
- Count: Number of events/deaths in the rows category
- Annotation_Code: Code when Count is NaN, blank=no annotation, 1=cell suppressed for small numbers, 2 = cell suppressed for complementary cell, 3 = no data is available, 4 = statistically unstable value.
- Annotation_Desc: Description of code (already provided above)

Data Cleaning

```
"""
State-level FIPS codes have two digits, county-level FIPS codes have five digits of which
first two are the FIPS code of the state to which the county belongs.
"""
str_state_num=population_df[['STATE']].astype(str)
zero_extend_str_state = ['0'+x if len(x)==1 else x for x in str_state_num['STATE']]
str_county_num=population_df[['COUNTY']].astype(str)
zero_extend_str_county = ['00'+x if len(x)==1 else '0'+x if len(x)==2 else x for x in str_
str_state_num['STATE'] = zero_extend_str_state
str_county_num['COUNTY'] = zero_extend_str_county
FIPS =  str_state_num['STATE'] + str_county_num['COUNTY']
population_df['FIPS'] = FIPS

deaths_df = deaths_df[deaths_df['Year']!=2020]
cali_df = population_df[population_df['STNAME'] == 'California'].loc[:,'CTYNAME':]
cali_df.drop(['CENSUS2010POP','ESTIMATESBASE2010','POPESTIMATE2010','POPESTIMATE2011','POP
import re
pattern = r'([\w, ]+) County'
cali_df['CTYNAME'] = cali_df['CTYNAME'].str.extract(pattern)
cali_df.drop(191, inplace=True)
```

```
cali_df.rename(columns={"POPESTIMATE2014": "2014", "POPESTIMATE2015": "2015", "POPESTIMATE
cali_df
```

	CTYNAME object	2014 int64	2015 int64	2016 int64	2017 int64	2
	Alameda 1.7% Alpine 1.7% 56 others 96.6%	1083 - 10033449 	1080 - 10077263 	1053 - 10094865 	1116 - 10092365 	1
242	Sutter	94721	95224	95769	96161	
243	Tehama	62797	63150	63468	63827	
244	Trinity	13126	13094	12827	12727	
245	Tulare	454858	456794	458991	462072	
246	Tuolumne	53830	53599	53729	53953	
247	Ventura	842113	845599	846921	848264	
248	Yolo	208368	211998	215569	218470	
249	Yuba	73527	74039	74920	76575	

```python
# Forloop to get population per year

population_per_year_county = []
fips_lst = []
cig_lst = []
df = deaths_df[['Year','County']]
for index in df.index:
    year = df['Year'][index]
    county = df['County'][index]
    pop_row = cali_df[cali_df['CTYNAME']==county]
    population = list(pop_row[f'{year}'])
    fips = list(pop_row['FIPS'])

    cig_row = cig_df[cig_df['County']==county]
    cigs = list(cig_row['Cig_Rates'])
    population_per_year_county.append(population[0])
    fips_lst.append(fips[0])
    cig_lst.append(cigs[0])
population_per_year_county = np.array(population_per_year_county)
fips_lst = np.array(fips_lst)
cig_lst = np.array(cig_lst)

deaths_df['Population'] = population_per_year_county
deaths_df['Fips'] = fips_lst
deaths_df['Cig_Rate'] = cig_lst
```

```
deaths_df
```

	Year int64	County object	Geography_Type ...	Strata object	Strata_Name obje...	C
0	2014	Alameda	Occurrence	Total Population	Total Population	A
1	2014	Alameda	Occurrence	Age	Under 1 year	A
2	2014	Alameda	Occurrence	Age	1-4 years	A
3	2014	Alameda	Occurrence	Age	5-14 years	A
4	2014	Alameda	Occurrence	Age	15-24 years	A
5	2014	Alameda	Occurrence	Age	25-34 years	A
6	2014	Alameda	Occurrence	Age	35-44 years	A
7	2014	Alameda	Occurrence	Age	45-54 years	A
8	2014	Alameda	Occurrence	Age	55-64 years	A
9	2014	Alameda	Occurrence	Age	65-74 years	A

```
deaths_df.isna().sum()
```

```
Year                  0
County                0
Geography_Type        0
Strata                0
Strata_Name           0
Cause                 0
Cause_Desc            0
Count             38849
Annotation_Code   87823
Annotation_Desc   87823
Population            0
Fips                  0
Cig_Rate              0
dtype: int64
```

```
deaths_df['Count'].replace(np.NaN,1,inplace=True)
deaths_df.drop(['Annotation_Code','Annotation_Desc'],axis=1,inplace=True)
```

```
deaths_df.to_csv(r'main_df.csv', index = False)
```