Lung_Cancer_EDA_orosco

February 3, 2023

[1]:
```
# Using an API to retrieve data
# Christine  Orosco
```

[2]:
```
import pandas as pd
import numpy as np
import json
import requests as res
from sodapy import Socrata  # API libararies for access to endpoint
```

0.1 Read Data using API

[3]:
```
# Access the data set on the Socrata DB endpoint
# https://chronicdata.cdc.gov/resource/u9ek-bct3.json?yearstart=2010
# grab data from years 2010 - 2018
```

[4]:
```
# read in variables from file
# read json file to load API Token value
with open('~/Socrata_API.json') as f:
    data = json.load(f)
    MyAppToken = data['APIToken']
    MyPassword = data['password']
    year = 2010
```

[6]:
```
# Use  Requests with anonymous request versus the soapy wrapper

url = 'https://chronicdata.cdc.gov/resource/u9ek-bct3.json?
 ↪&$limit=5000&yearstart=2010&question=Cancer of the lung and bronchus,␣
 ↪mortality'
resp = res.get(url)
data = json.loads(resp.text)
```

[7]:
```
# Save data to output file
out = json.dumps(resp.text)
```

[8]:
```
# Write out file
with open('~/cdc_data.json', 'w') as f:
    f.write(out)
```

[9]:
```
# Convert json file to Dataframe
df = pd.DataFrame.from_records(data)
df
```

[9]:

```
                stratification1                    datavaluetype  \
0                        Female             Average Annual Number
1                       Overall         Average Annual Crude Rate
2                       Overall         Average Annual Crude Rate
3                          Male             Average Annual Number
4                      Hispanic             Average Annual Number
…                             …                                …
1243                    Overall             Average Annual Number
1244                    Overall  Average Annual Age-adjusted Rate
1245          Black, non-Hispanic             Average Annual Number
1246  Asian or Pacific Islander         Average Annual Crude Rate
1247                    Overall  Average Annual Age-adjusted Rate

                                              geolocation  \
0     {'latitude': '47.52227862900048', 'human_addre…
1     {'latitude': '45.254228894000505', 'human_addr…
2     {'latitude': '41.56266102000046', 'human_addre…
3     {'latitude': '44.6613195430005', 'human_addres…
4     {'latitude': '35.68094058000048', 'human_addre…
…                                                   …
1243  {'latitude': '47.52227862900048', 'human_addre…
1244  {'latitude': '44.56744942400047', 'human_addre…
1245  {'latitude': '39.29058096400047', 'human_addre…
1246  {'latitude': '40.79373015200048', 'human_addre…
1247  {'latitude': '40.06021014100048', 'human_addre…

     stratificationcategory1 yearend  \
0                     Gender    2014
1                    Overall    2014
2                    Overall    2014
3                     Gender    2014
4              Race/Ethnicity    2014
…                         …       …
1243                 Overall    2014
1244                 Overall    2014
1245           Race/Ethnicity    2014
1246           Race/Ethnicity    2014
1247                 Overall    2014

                                        question         datasource  \
0     Cancer of the lung and bronchus, mortality  Death Certificate
1     Cancer of the lung and bronchus, mortality  Death Certificate
2     Cancer of the lung and bronchus, mortality  Death Certificate
```

```
3     Cancer of the lung and bronchus, mortality  Death Certificate
4     Cancer of the lung and bronchus, mortality  Death Certificate
…                                              …                 …
1243  Cancer of the lung and bronchus, mortality  Death Certificate
1244  Cancer of the lung and bronchus, mortality  Death Certificate
1245  Cancer of the lung and bronchus, mortality  Death Certificate
1246  Cancer of the lung and bronchus, mortality  Death Certificate
1247  Cancer of the lung and bronchus, mortality  Death Certificate

     stratificationcategoryid1 locationid questionid  …   topic  \
0                       GENDER         53     CAN8_2  …  Cancer
1                      OVERALL         23     CAN8_2  …  Cancer
2                      OVERALL         09     CAN8_2  …  Cancer
3                       GENDER         26     CAN8_2  …  Cancer
4                         RACE         47     CAN8_2  …  Cancer
…                            …        …         …  …     …
1243                   OVERALL         53     CAN8_2  …  Cancer
1244                   OVERALL         41     CAN8_2  …  Cancer
1245                      RACE         24     CAN8_2  …  Cancer
1246                      RACE         42     CAN8_2  …  Cancer
1247                   OVERALL         39     CAN8_2  …  Cancer

     stratificationid1  locationdesc datavalue   datavaluetypeid  \
0                 GENF    Washington      1456        AVGANNNMBR
1                  OVR         Maine      71.9     AVGANNCRDRATE
2                  OVR   Connecticut      47.6     AVGANNCRDRATE
3                 GENM      Michigan      3171        AVGANNNMBR
4                  HIS     Tennessee        10        AVGANNNMBR
…                  …            …       …               …
1243               OVR    Washington      3107        AVGANNNMBR
1244               OVR        Oregon      44.2  AVGANNAGEADJRATE
1245               BLK      Maryland       683        AVGANNNMBR
1246              APIO  Pennsylvania      13.8     AVGANNCRDRATE
1247               OVR          Ohio      52.8  AVGANNAGEADJRATE

     lowconfidencelimit highconfidencelimit datavalueunit datavaluefootnote  \
0                   NaN                 NaN           NaN               NaN
1                  69.9                  74   per 100,000               NaN
2                  46.6                48.7   per 100,000               NaN
3                   NaN                 NaN           NaN               NaN
4                   NaN                 NaN           NaN               NaN
…                   …                   …             …                 …
1243                NaN                 NaN           NaN               NaN
1244               43.4                45.1   per 100,000               NaN
1245                NaN                 NaN           NaN               NaN
1246               12.2                15.5   per 100,000               NaN
1247               52.3                53.4   per 100,000               NaN
```

```
     datavaluefootnotesymbol
0                        NaN
1                        NaN
2                        NaN
3                        NaN
4                        NaN
…                        …
1243                     NaN
1244                     NaN
1245                     NaN
1246                     NaN
1247                     NaN

[1248 rows x 24 columns]
```

[10]:
```
df.columns
```

[10]:
```
Index(['stratification1', 'datavaluetype', 'geolocation',
       'stratificationcategory1', 'yearend', 'question', 'datasource',
       'stratificationcategoryid1', 'locationid', 'questionid', 'locationabbr',
       'yearstart', 'datavaluealt', 'topicid', 'topic', 'stratificationid1',
       'locationdesc', 'datavalue', 'datavaluetypeid', 'lowconfidencelimit',
       'highconfidencelimit', 'datavalueunit', 'datavaluefootnote',
       'datavaluefootnotesymbol'],
      dtype='object')
```

[11]:
```
# Create subset to work with
col_list = ['locationabbr', 'yearstart', 'yearend',\
            'stratification1', \
            'datavaluetype', 'datavalue', 'datavaluealt']
df_subset = df[col_list].copy()
```

[12]:
```
# sort DF
df_subset.sort_values('locationabbr', inplace=True)
```

0.2 Data cleaning

[13]:
```
# Rename columns
df_subset.rename(columns={'locationabbr':'State_Code', \
                          'stratification1': 'Strat',\
                          'datavaluetype': 'Value_Type',\
                          'datavalue': 'Value', \
                          'datavaluealt': 'Alt_Value'}, inplace=True)
```

[14]:
```
# Change column values to merge two columns into one (Unit and Value Type)
df_subset.replace({'Value_Type': 'Average Annual Age-adjusted Rate'},\
                  {'Value_Type':'AA_Age_ARate_100k'}, regex=True, inplace=True)
df_subset.replace({'Value_Type': 'Average Annual Crude Rate'},\
                  {'Value_Type':'AA_CRate_100k'}, regex=True, inplace=True)
df_subset.replace({'Value_Type': 'Average Annual Number'},\
                  {'Value_Type':'AA_Nbr'}, regex=True, inplace=True)
```

[15]:
```
# Missing Values
# Check Sample
df_subset[df_subset['State_Code'] == 'AK']
```

[15]:

	State_Code	yearstart	yearend	Strat	\
1116	AK	2010	2014	Male	
719	AK	2010	2014	American Indian or Alaska Native	
932	AK	2010	2014	White, non-Hispanic	
783	AK	2010	2014	Overall	
550	AK	2010	2014	Overall	
567	AK	2010	2014	Male	
351	AK	2010	2014	White, non-Hispanic	
521	AK	2010	2014	Black, non-Hispanic	
1148	AK	2010	2014	Overall	
1162	AK	2010	2014	Asian or Pacific Islander	
1095	AK	2010	2014	American Indian or Alaska Native	
1018	AK	2010	2014	Black, non-Hispanic	
468	AK	2010	2014	Female	
726	AK	2010	2014	Hispanic	
917	AK	2010	2014	Hispanic	
404	AK	2010	2014	White, non-Hispanic	
407	AK	2010	2014	Black, non-Hispanic	
978	AK	2010	2014	Male	
694	AK	2010	2014	American Indian or Alaska Native	
416	AK	2010	2014	Asian or Pacific Islander	
1221	AK	2010	2014	Female	
49	AK	2010	2014	Female	
286	AK	2010	2014	Hispanic	
579	AK	2010	2014	Asian or Pacific Islander	

	Value_Type	Value	Alt_Value
1116	AA_CRate_100k	36.8	36.8
719	AA_Nbr	49	49
932	AA_Age_ARate_100k	46	46
783	AA_Age_ARate_100k	47.6	47.6
550	AA_CRate_100k	35	35
567	AA_Age_ARate_100k	54.6	54.6
351	AA_CRate_100k	38.8	38.8
521	AA_Nbr	6	6

```
1148             AA_Nbr   255       255
1162  AA_Age_ARate_100k  32.3      32.3
1095      AA_CRate_100k  39.3      39.3
1018      AA_CRate_100k  18.5      18.5
468              AA_Nbr   115       115
726              AA_Nbr   NaN       NaN
917   AA_Age_ARate_100k   NaN       NaN
404              AA_Nbr   187       187
407   AA_Age_ARate_100k  49.2      49.2
978              AA_Nbr   140       140
694   AA_Age_ARate_100k    67        67
416              AA_Nbr    11        11
1221      AA_CRate_100k  33.1      33.1
49    AA_Age_ARate_100k  41.8      41.8
286       AA_CRate_100k   NaN       NaN
579       AA_CRate_100k  20.2      20.2
```

[16]:
```
# Change Value and Alt_Value NaN to 0
df_subset['Value'] = df_subset['Value'].fillna(0)
df_subset['Alt_Value'] = df_subset['Alt_Value'].fillna(0)
```

[17]:
```
# Change datatypes
col_list = ['yearstart', 'yearend','Value', 'Alt_Value' ]
df_subset[col_list] = df_subset[col_list].apply(pd.to_numeric)
```

[19]:
```
df_subset[df_subset['State_Code'] == 'AK']
```

[19]:
```
     State_Code  yearstart  yearend                            Strat  \
1116         AK       2010     2014                             Male
719          AK       2010     2014  American Indian or Alaska Native
932          AK       2010     2014             White, non-Hispanic
783          AK       2010     2014                          Overall
550          AK       2010     2014                          Overall
567          AK       2010     2014                             Male
351          AK       2010     2014             White, non-Hispanic
521          AK       2010     2014             Black, non-Hispanic
1148         AK       2010     2014                          Overall
1162         AK       2010     2014      Asian or Pacific Islander
1095         AK       2010     2014  American Indian or Alaska Native
1018         AK       2010     2014             Black, non-Hispanic
468          AK       2010     2014                           Female
726          AK       2010     2014                         Hispanic
917          AK       2010     2014                         Hispanic
404          AK       2010     2014             White, non-Hispanic
407          AK       2010     2014             Black, non-Hispanic
978          AK       2010     2014                             Male
694          AK       2010     2014  American Indian or Alaska Native
```

```
416          AK       2010     2014         Asian or Pacific Islander
1221         AK       2010     2014                           Female
49           AK       2010     2014                           Female
286          AK       2010     2014                         Hispanic
579          AK       2010     2014         Asian or Pacific Islander

             Value_Type  Value  Alt_Value
1116      AA_CRate_100k   36.8       36.8
719              AA_Nbr   49.0       49.0
932   AA_Age_ARate_100k   46.0       46.0
783   AA_Age_ARate_100k   47.6       47.6
550       AA_CRate_100k   35.0       35.0
567   AA_Age_ARate_100k   54.6       54.6
351       AA_CRate_100k   38.8       38.8
521              AA_Nbr    6.0        6.0
1148             AA_Nbr  255.0      255.0
1162  AA_Age_ARate_100k   32.3       32.3
1095      AA_CRate_100k   39.3       39.3
1018      AA_CRate_100k   18.5       18.5
468              AA_Nbr  115.0      115.0
726              AA_Nbr    0.0        0.0
917   AA_Age_ARate_100k    0.0        0.0
404              AA_Nbr  187.0      187.0
407   AA_Age_ARate_100k   49.2       49.2
978              AA_Nbr  140.0      140.0
694   AA_Age_ARate_100k   67.0       67.0
416              AA_Nbr   11.0       11.0
1221      AA_CRate_100k   33.1       33.1
49    AA_Age_ARate_100k   41.8       41.8
286       AA_CRate_100k    0.0        0.0
579       AA_CRate_100k   20.2       20.2
```

0.2.1 Reshape dataframe using pivot_table

[20]:
```
# Create a pivot_table to get aggregates for unqiue rows (long format)
index_col = ['State_Code', 'Strat', 'Value_Type']
data_list = ['Value', 'Alt_Value', 'yearstart', 'yearend']
df_pivot = df_subset.pivot_table(index=index_col, values=data_list, aggfunc=sum)
```

[21]:
```
# Sort the Dataframe
df_pivot.sort_index(inplace=True)
```

[22]:
```
df_pivot.head(24)
```

[22]:
```
                                                          Alt_Value  \
State_Code Strat                            Value_Type
AK         American Indian or Alaska Native AA_Age_ARate_100k       67.0
```

		AA_CRate_100k	39.3	
		AA_Nbr	49.0	
	Asian or Pacific Islander	AA_Age_ARate_100k	32.3	
		AA_CRate_100k	20.2	
		AA_Nbr	11.0	
	Black, non-Hispanic	AA_Age_ARate_100k	49.2	
		AA_CRate_100k	18.5	
		AA_Nbr	6.0	
	Female	AA_Age_ARate_100k	41.8	
		AA_CRate_100k	33.1	
		AA_Nbr	115.0	
	Hispanic	AA_Age_ARate_100k	0.0	
		AA_CRate_100k	0.0	
		AA_Nbr	0.0	
	Male	AA_Age_ARate_100k	54.6	
		AA_CRate_100k	36.8	
		AA_Nbr	140.0	
	Overall	AA_Age_ARate_100k	47.6	
		AA_CRate_100k	35.0	
		AA_Nbr	255.0	
	White, non-Hispanic	AA_Age_ARate_100k	46.0	
		AA_CRate_100k	38.8	
		AA_Nbr	187.0	

			Value	yearend	\
State_Code	Strat	Value_Type			
AK	American Indian or Alaska Native	AA_Age_ARate_100k	67.0	2014	
		AA_CRate_100k	39.3	2014	
		AA_Nbr	49.0	2014	
	Asian or Pacific Islander	AA_Age_ARate_100k	32.3	2014	
		AA_CRate_100k	20.2	2014	
		AA_Nbr	11.0	2014	
	Black, non-Hispanic	AA_Age_ARate_100k	49.2	2014	
		AA_CRate_100k	18.5	2014	
		AA_Nbr	6.0	2014	
	Female	AA_Age_ARate_100k	41.8	2014	
		AA_CRate_100k	33.1	2014	
		AA_Nbr	115.0	2014	
	Hispanic	AA_Age_ARate_100k	0.0	2014	
		AA_CRate_100k	0.0	2014	
		AA_Nbr	0.0	2014	
	Male	AA_Age_ARate_100k	54.6	2014	
		AA_CRate_100k	36.8	2014	
		AA_Nbr	140.0	2014	
	Overall	AA_Age_ARate_100k	47.6	2014	
		AA_CRate_100k	35.0	2014	
		AA_Nbr	255.0	2014	

```
            White, non-Hispanic              AA_Age_ARate_100k   46.0     2014
                                             AA_CRate_100k       38.8     2014
                                             AA_Nbr             187.0     2014
```

```
                                                                yearstart
State_Code Strat                               Value_Type
AK         American Indian or Alaska Native    AA_Age_ARate_100k       2010
                                               AA_CRate_100k           2010
                                               AA_Nbr                  2010
           Asian or Pacific Islander           AA_Age_ARate_100k       2010
                                               AA_CRate_100k           2010
                                               AA_Nbr                  2010
           Black, non-Hispanic                 AA_Age_ARate_100k       2010
                                               AA_CRate_100k           2010
                                               AA_Nbr                  2010
           Female                              AA_Age_ARate_100k       2010
                                               AA_CRate_100k           2010
                                               AA_Nbr                  2010
           Hispanic                            AA_Age_ARate_100k       2010
                                               AA_CRate_100k           2010
                                               AA_Nbr                  2010
           Male                                AA_Age_ARate_100k       2010
                                               AA_CRate_100k           2010
                                               AA_Nbr                  2010
           Overall                             AA_Age_ARate_100k       2010
                                               AA_CRate_100k           2010
                                               AA_Nbr                  2010
           White, non-Hispanic                 AA_Age_ARate_100k       2010
                                               AA_CRate_100k           2010
                                               AA_Nbr                  2010
```

[23]:
```
# Create a subset of the df_subset with only the overall values to add a␣
 ↪percent value column from the avg annual crude rate
subset_all = df_subset[(df_subset['Strat'] == 'Overall') &␣
 ↪(df_subset['Value_Type'] == 'AA_CRate_100k')].copy()
subset_all.head(6)
```

[23]:

	State_Code	yearstart	yearend	Strat	Value_Type	Value	Alt_Value
550	AK	2010	2014	Overall	AA_CRate_100k	35.0	35.0
1101	AL	2010	2014	Overall	AA_CRate_100k	65.4	65.4
985	AR	2010	2014	Overall	AA_CRate_100k	71.9	71.9
1043	AZ	2010	2014	Overall	AA_CRate_100k	41.8	41.8
878	CA	2010	2014	Overall	AA_CRate_100k	33.1	33.1
621	CO	2010	2014	Overall	AA_CRate_100k	30.7	30.7

[ ]:
```
### Write out dataframe to a flat file
```

[30]:
```
# Write subset to csv file
subset_csv = df_subset[(df_subset['Strat'] == 'Overall') &␣
 ↪(df_subset['Value_Type'] == 'AA_Nbr')].copy()
```

[31]:
```
subset_csv.drop(columns=['Alt_Value'], axis=1, inplace=True)
```

[ ]:
```

```

[32]:
```
subset_csv.to_csv('~/api_deaths.csv', index = False)
```

[ ]:
```
### Add new column to store the rate and reshape data from long to wide
```

[25]:
```
# Add New Column
rate = 1000
subset_all['Percent_Deaths'] = subset_all.apply(lambda row: row['Value'] /␣
 ↪rate, axis=1)
subset_all.reindex
subset_all.head(5)
```

[25]:

	State_Code	yearstart	yearend	Strat	Value_Type	Value	Alt_Value	\
550	AK	2010	2014	Overall	AA_CRate_100k	35.0	35.0	
1101	AL	2010	2014	Overall	AA_CRate_100k	65.4	65.4	
985	AR	2010	2014	Overall	AA_CRate_100k	71.9	71.9	
1043	AZ	2010	2014	Overall	AA_CRate_100k	41.8	41.8	
878	CA	2010	2014	Overall	AA_CRate_100k	33.1	33.1	

	Percent_Deaths
550	0.0350
1101	0.0654
985	0.0719
1043	0.0418
878	0.0331

[26]:
```
# Reshape data into wide format
index_col = ['State_Code', 'Strat']
data_list = ['Value', 'Alt_Value']
df_pivot2 = df_subset.pivot_table(index=index_col, values=data_list,␣
 ↪columns='Value_Type')
```

[27]:
```
df_pivot2
```

[27]:

		Alt_Value	
Value_Type		AA_Age_ARate_100k	AA_CRate_100k
State_Code	Strat		
AK	American Indian or Alaska Native	67.0	39.3
	Asian or Pacific Islander	32.3	20.2
	Black, non-Hispanic	49.2	18.5

```
           Female                                        41.8          33.1
           Hispanic                                       0.0           0.0
…                                                         …             …
WY         Female                                        32.5          37.5
           Hispanic                                      26.6          11.9
           Male                                          41.9          41.4
           Overall                                       36.7          39.5
           White, non-Hispanic                           37.2          43.5

                                                                      Value  \
Value_Type                                  AA_Nbr AA_Age_ARate_100k
State_Code Strat
AK         American Indian or Alaska Native   49.0              67.0
           Asian or Pacific Islander          11.0              32.3
           Black, non-Hispanic                 6.0              49.2
           Female                            115.0              41.8
           Hispanic                            0.0               0.0
…                                              …                 …
WY         Female                            106.0              32.5
           Hispanic                            6.0              26.6
           Male                              121.0              41.9
           Overall                           227.0              36.7
           White, non-Hispanic               215.0              37.2


Value_Type                                  AA_CRate_100k AA_Nbr
State_Code Strat
AK         American Indian or Alaska Native          39.3   49.0
           Asian or Pacific Islander                 20.2   11.0
           Black, non-Hispanic                       18.5    6.0
           Female                                    33.1  115.0
           Hispanic                                   0.0    0.0
…                                                     …      …
WY         Female                                    37.5  106.0
           Hispanic                                  11.9    6.0
           Male                                      41.4  121.0
           Overall                                   39.5  227.0
           White, non-Hispanic                       43.5  215.0

[416 rows x 6 columns]
```

[ ]: