Introduction

You work for an environmental think tank called Repair Our Air (ROA). ROA is formulating policy recommendations to improve the air quality in America, using the Environmental Protection Agency's Air Quality Index (AQI) to guide their decision making. An AQI value close to 0 signals "little to no" public health concern, while higher values are associated with increased risk to public health.

They've tasked you with leveraging AQI data to help them prioritize their strategy for improving air quality in America.

ROA is considering the following decisions. For each, construct a hypothesis test and an accompanying visualization, using your results of that test to make a recommendation:

ROA is considering a metropolitan-focused approach. Within California, they want to know if the mean AQI in Los Angeles County is statistically different from the rest of California. With limited resources, ROA has to choose between New York and Ohio for their next regional office. Does New York have a lower AQI than Ohio? A new policy will affect those states with a mean AQI of 10 or greater. Can you rule out Michigan from being affected by this new policy?

In [1]:

```
# Import relevant packages

import pandas as pd
import numpy as np
from scipy import stats
```

In [2]:

```
# Use read_csv() to import your data

df = pd.read_csv('startdata1.csv')
```

In [3]:

```
df.head()
```

Out[3]:

	Unnamed: 0	date_local	state_name	county_name	city_name	local_site_name	parameter_nan
0	0	2018-01-01	Arizona	Maricopa	Buckeye	BUCKEYE	Carbon monoxi
1	1	2018-01-01	Ohio	Belmont	Shadyside	Shadyside	Carbon monoxi
2	2	2018-01-01	Wyoming	Teton	Not in a city	Yellowstone National Park - Old Faithful Snow ...	Carbon monoxi
3	3	2018-01-01	Pennsylvania	Philadelphia	Philadelphia	North East Waste (NEW)	Carbon monoxi
4	4	2018-01-01	Iowa	Polk	Des Moines	CARPENTER	Carbon monoxi

In [4]:

```
print("Use describe() to summarize AQI")
print(df.describe(include='all'))
```

```
Use describe() to summarize AQI
        Unnamed: 0  date_local  state_name  county_name      city_name  \
count   260.000000         260         260          260            260   
unique         NaN           1          52          149            190   
top            NaN  2018-01-01  California  Los Angeles  Not in a city   
freq           NaN         260          66           14             21   
mean    129.500000         NaN         NaN          NaN            NaN   
std      75.199734         NaN         NaN          NaN            NaN   
min       0.000000         NaN         NaN          NaN            NaN   
25%      64.750000         NaN         NaN          NaN            NaN   
50%     129.500000         NaN         NaN          NaN            NaN   
75%     194.250000         NaN         NaN          NaN            NaN   
max     259.000000         NaN         NaN          NaN            NaN   

       local_site_name   parameter_name   units_of_measure  arithmetic_mean  
\
count              257              260                260       260.000000   
unique             253                1                  1              NaN   
top            Kapolei  Carbon monoxide  Parts per million              NaN   
freq                 2              260                260              NaN   
mean               NaN              NaN                NaN         0.403169   
std                NaN              NaN                NaN         0.317902   
min                NaN              NaN                NaN         0.000000   
25%                NaN              NaN                NaN         0.200000   
50%                NaN              NaN                NaN         0.276315   
75%                NaN              NaN                NaN         0.516009   
max                NaN              NaN                NaN         1.921053   

               aqi  
count   260.000000  
unique         NaN  
top            NaN  
freq           NaN  
mean      6.757692  
std       7.061707  
min       0.000000  
25%       2.000000  
50%       5.000000  
75%       9.000000  
max      50.000000  
```

In [6]:

```
print("For a more thorough examination of observations by state use values
print(df['state_name'].value_counts())
```

```
For a more thorough examination of observations by state use values_counts()
California              66
Arizona                 14
Ohio                    12
Florida                 12
Texas                   10
New York                10
Pennsylvania            10
Michigan                 9
Colorado                 9
Minnesota                7
New Jersey               6
Indiana                  5
North Carolina           4
Massachusetts            4
Maryland                 4
Oklahoma                 4
Virginia                 4
Nevada                   4
Connecticut              4
Kentucky                 3
Missouri                 3
Wyoming                  3
Iowa                     3
Hawaii                   3
Utah                     3
Vermont                  3
Illinois                 3
New Hampshire            2
District Of Columbia     2
New Mexico               2
Montana                  2
Oregon                   2
Alaska                   2
Georgia                  2
Washington               2
Idaho                    2
Nebraska                 2
Rhode Island             2
Tennessee                2
Maine                    2
South Carolina           1
Puerto Rico              1
Arkansas                 1
Kansas                   1
Mississippi              1
Alabama                  1
Louisiana                1
Delaware                 1
South Dakota             1
West Virginia            1
North Dakota             1
Wisconsin                1
Name: state_name, dtype: int64
```

Statistical test

Recall the following steps for conducting hypothesis testing:

Formulate the null hypothesis and the alternative hypothesis.

Set the significance level.

Determine the appropriate test procedure.

Compute the p-value.

Draw your conclusion.

In [8]:
```
# Create dataframes for each sample being compared in your test


ca_la = df[df['county_name']=='Los Angeles']
ca_other = df[(df['state_name']=='California') & (df['county_name']!='Los
```

In [9]:
```
ca_la.head()
```

Out[9]:

	Unnamed: 0	date_local	state_name	county_name	city_name	local_site_name	parameter_nar
33	33	2018-01-01	California	Los Angeles	Lancaster	Lancaster-Division Street	Carbon monoxi
42	42	2018-01-01	California	Los Angeles	Santa Clarita	Santa Clarita	Carbon monoxi
61	61	2018-01-01	California	Los Angeles	Pasadena	Pasadena	Carbon monoxi
76	76	2018-01-01	California	Los Angeles	Los Angeles	LAX Hastings	Carbon monoxi
109	109	2018-01-01	California	Los Angeles	Los Angeles	Los Angeles-North Main Street	Carbon monoxi

In [11]:
```
#    validate
ca_la.state_name.unique()
```

Out[11]:
```
array(['California'], dtype=object)
```

In [12]:

```
ca_other.head()
```

Out[12]:

	Unnamed: 0	date_local	state_name	county_name	city_name	local_site_name	parameter_nam
16	16	2018-01-01	California	San Bernardino	Ontario	Ontario Near Road (Etiwanda)	Carbon monoxid
18	18	2018-01-01	California	Sacramento	Arden-Arcade	Sacramento-Del Paso Manor	Carbon monoxid
26	26	2018-01-01	California	Orange	La Habra	La Habra	Carbon monoxid
27	27	2018-01-01	California	Alameda	Not in a city	Berkeley-Aquatic Park	Carbon monoxid
34	34	2018-01-01	California	Fresno	Fresno	Fresno - Garland	Carbon monoxid

In [16]:

```
ca_other.county_name.nunique()
```

Out[16]:

```
25
```

In [17]:

```
df.county_name.nunique()
```

Out[17]:

```
149
```

Formulate your hypothesis:

Formulate your null and alternative hypotheses:

$H0$: There is no difference in the mean AQI between Los Angeles County and the rest of California.

HA: There is a difference in the mean AQI between Los Angeles County and the rest of California.

In [18]:

```
# For this analysis, the significance level is 5%
significance_level = 0.05
significance_level
```

Out[18]:

```
0.05
```

Here, you are comparing the sample means between two independent samples. Therefore, you will utilize a two-sample t-tes

In [19]:

```
# Compute your p-value here
stats.ttest_ind(a=ca_la['aqi'], b=ca_other['aqi'], equal_var=False)
```

Out[19]:

```
Ttest_indResult(statistic=2.1107010796372014, pvalue=0.049839056842410995)
```

With a p-value (0.049) being less than 0.05 (as your significance level is 5%), reject the null hypothesis in favor of the alternative hypothesis.

Therefore, a metropolitan strategy may make sense in this case.

Hypothesis 2:

With limited resources, ROA has to choose between New York and Ohio for their next regional office. Does New York have a lower AQI than Ohio?

In [21]:

```
# Create dataframes for each sample being compared in your test
ny = df[df['state_name']=='New York']
ohio = df[df['state_name']=='Ohio']
```

Formulate your hypothesis: Formulate your null and alternative hypotheses:

$H0$: The mean AQI of New York is greater than or equal to that of Ohio.

HA: The mean AQI of New York is below that of Ohio.

In [22]:

```
# Here, you are comparing the sample means between two independent samples
tstat, pvalue = stats.ttest_ind(a=ny['aqi'], b=ohio['aqi'], alternative='l
print(tstat)
print(pvalue)
```

```
-1.891850434703295
0.03654034300840755
```

With a p-value (0.030) being less than 0.05 (as your significance level is 5%) and a t-statistic < 0 (-2.02), reject the null hypothesis in favor of the alternative hypothesis.

Therefore, you can conclude at the 5% significance level that New York has a lower mean AQI than Ohio.

Hypothesis 3:

A new policy will affect those states with a mean AQI of 10 or greater. Can you rule out Michigan from being affected by this new policy?

In [24]:

```
# Create dataframes for each sample being compared in your test
michigan = df[df['state_name']=='Michigan']
```

Formulate your hypothesis: Formulate your null and alternative hypotheses here:

$H0$: The mean AQI of Michigan is less than or equal to 10.

HA: The mean AQI of Michigan is greater than 10.

Here, you are comparing one sample mean relative to a particular value in one direction. Therefore, you will utilize a one-sample t-test.

In [25]:

```
# Compute your p-value here
tstat, pvalue = stats.ttest_1samp(michigan['aqi'], 10, alternative='greate
print(tstat)
print(pvalue)
```

```
-1.7395913343286131
0.939940519314011
```

With a p-value (0.060) being greater than 0.05 (as your significance level is 5%) and a t-statistic < 0 (-1.73), fail to reject the null hypothesis.

Therefore, you cannot conclude at the 5% significance level that Michigan's mean AQI is greater than 10. This implies that Michigan would not be affected by the new policy.

Result and evaluation

1. The results indicated that the AQI in Los Angeles County was in fact different from the rest of California.
2. Using a 5% significance level, you can conclude that New York has a lower AQI than Ohio based on the results.
3. Based on the tests, you would fail to reject the null hypothesis, meaning you can't conclude that the mean AQI is greater than 10. Thus, it is unlikely that Michigan would be affected by the new policy.

In [ ]:

```

```