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 [3]:

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

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

In [5]:

```python
# Explore your dataframe `df` here:



print("Use head() to show a sample of data")
print(df.head())

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

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

print('for a more')
```

```
Use head() to show a sample of data
   Unnamed: 0  date_local    state_name   county_name      city_name  \
0           0  2018-01-01       Arizona      Maricopa        Buckeye   
1           1  2018-01-01          Ohio       Belmont      Shadyside   
2           2  2018-01-01       Wyoming         Teton  Not in a city   
3           3  2018-01-01  Pennsylvania  Philadelphia   Philadelphia   
4           4  2018-01-01          Iowa          Polk     Des Moines   

                                     local_site_name   parameter_name  \
0                                            BUCKEYE  Carbon monoxide   
1                                          Shadyside  Carbon monoxide   
2  Yellowstone National Park - Old Faithful Snow ...  Carbon monoxide   
3                             North East Waste (NEW)  Carbon monoxide   
4                                          CARPENTER  Carbon monoxide   

    units_of_measure  arithmetic_mean  aqi  
0  Parts per million         0.473684    7  
1  Parts per million         0.263158    5  
2  Parts per million         0.111111    2  
```

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 [6]:

```python
# 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') & (aqi['county_name']!='Los
```

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 [8]:

```python
# For this analysis, the significance level is 5%

significance_level = 0.05
significance_level
```

Out[8]:

```
0.05
```

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

In [9]:

```python
# Compute your p-value here


stats.ttest_ind(a=ca_la['aqi'], b=ca_other['aqi'], equal_var=False)
```

Out[9]:

```
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 [10]:

```python
# Create dataframes for each sample being compared in your test


ny = aqi[aqi['state_name']=='New York']
ohio = aqi[aqi['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 [12]:

```
# 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 [13]:

```
# Create dataframes for each sample being compared in your test



michigan = aqi[aqi['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 [14]:

```
# 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.

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.