In [172…
```
# Import Library
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import statsmodels.api as sm
from scipy import stats
```

In [174…
```
aqi = pd.read_csv(r'C:\Users\HP\Desktop\Advance Data Analyst\4. The Power of Stats\3. Module 3\3. Work with sampling distribution\Files\c4_epa_air_quality.csv')
aqi.head(10)
```

Out[174…

	Unnamed: 0	date_local	state_name	county_name	city_name	local_site_name	parameter_name	units_of_measure	arithmetic_mean	aqi
0	0	2018-01-01	Arizona	Maricopa	Buckeye	BUCKEYE	Carbon monoxide	Parts per million	0.473684	7
1	1	2018-01-01	Ohio	Belmont	Shadyside	Shadyside	Carbon monoxide	Parts per million	0.263158	5
2	2	2018-01-01	Wyoming	Teton	Not in a city	Yellowstone National Park - Old Faithful Snow ...	Carbon monoxide	Parts per million	0.111111	2
3	3	2018-01-01	Pennsylvania	Philadelphia	Philadelphia	North East Waste (NEW)	Carbon monoxide	Parts per million	0.300000	3
4	4	2018-01-01	Iowa	Polk	Des Moines	CARPENTER	Carbon monoxide	Parts per million	0.215789	3
5	5	2018-01-01	Hawaii	Honolulu	Not in a city	Kapolei	Carbon monoxide	Parts per million	0.994737	14
6	6	2018-01-01	Hawaii	Honolulu	Not in a city	Kapolei	Carbon monoxide	Parts per million	0.200000	2
7	7	2018-01-01	Pennsylvania	Erie	Erie	NaN	Carbon monoxide	Parts per million	0.200000	2
8	8	2018-01-01	Hawaii	Honolulu	Honolulu	Honolulu	Carbon monoxide	Parts per million	0.400000	5
9	9	2018-01-01	Colorado	Larimer	Fort Collins	Fort Collins - CSU - S. Mason	Carbon monoxide	Parts per million	0.300000	6

In [176…
```
# Explore the `aqi` DataFrame.

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

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

```
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
For a more thorough examination of observations by state use values_counts()
state_name
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: count, dtype: int64
```

In [178…
```
# Summarize the mean AQI for RRE states.

# Create a list of RRE states.

rre_states = ['California','Florida','Michigan','Ohio','Pennsylvania','Texas']

# Subset `aqi` to only consider these states.

aqi_rre = aqi[aqi['state_name'].isin(rre_states)]

# Find the mean aqi for each of the RRE states.

aqi_rre.groupby(['state_name']).agg({"aqi":"mean","state_name":"count"}) #alias as aqi_rre
```

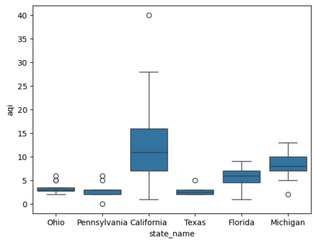
Out[178…

state_name	aqi	state_name
California	12.121212	66
Florida	5.500000	12
Michigan	8.111111	9
Ohio	3.333333	12
Pennsylvania	2.900000	10
Texas	2.700000	10

In [180…
```
# Create an in-line visualization showing the distribution of aqi by state_name

sns.boxplot(x=aqi_rre["state_name"],y=aqi_rre["aqi"])
```

Out[180…
```
<Axes: xlabel='state_name', ylabel='aqi'>
```



In [181…
```
# Find the mean aqi for your state.

aqi_ca = aqi[aqi['state_name']=='California']

sample_mean = aqi_ca['aqi'].mean()
sample_mean
```

Out[181…
```
12.121212121212121
```

In [184…
```
# Input your confidence level.

confidence_level = 0.95
confidence_level
```

Out[184…
```
0.95
```

In [186…
```
# Calculate your margin of error.

# Begin by identifying the z associated with your chosen confidence level.

z_value = 1.96

# Next, calculate your standard error.

standard_error = aqi_ca['aqi'].std() / np.sqrt(aqi_ca.shape[0])
print("standard error:")
print(standard_error)

# Lastly, use the preceding result to calculate your margin of error.

margin_of_error = standard_error * z_value
print("margin of error:")
print(margin_of_error)
```

```
standard error:
0.8987209641127412
margin of error:
1.7614930896609726
```

In [188…
```
# Calculate your confidence interval (upper and lower limits).

upper_ci_limit = sample_mean + margin_of_error
lower_ci_limit = sample_mean - margin_of_error
(lower_ci_limit, upper_ci_limit)
```

Out[188…
```
(10.359719031551148, 13.882705210873095)
```

In [ ]:
```
# key takeaways

# Based on the mean AQI for RRE states, California and Michigan were most likely to have experienced a mean AQI above 10.
# With California experiencing the highest sample mean AQI in the data, it appears to be the state most likely to be affected by the policy change.
# Constructing a confidence interval allowed you to estimate the sample mean AQI with a certain degree of confidence.

# What findings would you share with others?

# Present this notebook to convey the analytical process and describe the methodology behind constructing the confidence interval.
# Convey that a confidence interval at the 95% level of confidence from this sample data yielded [10.36 , 13.88], which provides the interpretation "given the observed sample AQI measurements, there is a 95% confidence that the popula
# Share how varying the confidence level changes the interval. For example, if you varied the confidence level to 99%, the confidence interval would become [9.80 , 14.43].

# What would you convey to external stakeholders?
```

```
# Explain statistical significance at a high level.
# Describe California's observed mean AQI and suggest focusing on that state.
# Share the result of the 95% confidence interval, describing what this means relative to the threshold of 10.
# Convey any potential shortcomings of this analysis, such as the short time period being referenced.
```