

# Activity\_Explore descriptive statistics

July 4, 2024

## 1 Activity: Explore descriptive statistics

### 1.1 Introduction

Data professionals often use descriptive statistics to understand the data they are working with and provide collaborators with a summary of the relative location of values in the data, as well as information about its spread.

For this activity, you are a member of an analytics team for the United States Environmental Protection Agency (EPA). You are assigned to analyze data on air quality with respect to carbon monoxide, a major air pollutant. The data includes information from more than 200 sites, identified by state, county, city, and local site names. You will use Python functions to gather statistics about air quality, then share insights with stakeholders.

### 1.2 Step 1: Imports

Import the relevant Python libraries `pandas` and `numpy`.

```
[1]: # Import relevant Python libraries.  
  
    ### YOUR CODE HERE ###  
    import pandas as pd  
    import numpy as np
```

The dataset provided is in the form of a `.csv` file named `c4_epa_air_quality.csv`. It contains a subset of data from the U.S. EPA. As shown in this cell, the dataset has been automatically loaded in for you. You do not need to download the `.csv` file, or provide more code, in order to access the dataset and proceed with this lab. Please continue with this activity by completing the following instructions.

```
[2]: # RUN THIS CELL TO IMPORT YOUR DATA.  
  
    ### YOUR CODE HERE  
    epa_data = pd.read_csv("c4_epa_air_quality.csv", index_col = 0)
```

Hint 1

Refer to the video about loading data in Python.

Hint 2

There is a function in the `pandas` library that allows you to read in data from a `.csv` file and load it into a `DataFrame`.

Hint 3

Use the `read_csv` function from the `pandas` library. The `index_col` parameter can be set to 0 to read in the first column as an index (and to avoid "Unnamed: 0" appearing as a column in the resulting `DataFrame`).

### 1.3 Step 2: Data exploration

To understand how the dataset is structured, display the first 10 rows of the data.

```
[3]: # Display first 10 rows of the data.
```

```
### YOUR CODE HERE
epa_data.head(10)
```

```
[3]:   date_local   state_name   county_name   city_name \
0  2018-01-01      Arizona      Maricopa      Buckeye
1  2018-01-01        Ohio      Belmont      Shadyside
2  2018-01-01      Wyoming        Teton  Not in a city
3  2018-01-01  Pennsylvania  Philadelphia  Philadelphia
4  2018-01-01        Iowa        Polk      Des Moines
5  2018-01-01      Hawaii      Honolulu  Not in a city
6  2018-01-01      Hawaii      Honolulu  Not in a city
7  2018-01-01  Pennsylvania        Erie        Erie
8  2018-01-01      Hawaii      Honolulu      Honolulu
9  2018-01-01      Colorado      Larimer  Fort Collins

                                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
5                                Kapolei   Carbon monoxide
6                                Kapolei   Carbon monoxide
7                                NaN   Carbon monoxide
8                                Honolulu   Carbon monoxide
9      Fort Collins - CSU - S. Mason   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
```

3	Parts per million	0.300000	3
4	Parts per million	0.215789	3
5	Parts per million	0.994737	14
6	Parts per million	0.200000	2
7	Parts per million	0.200000	2
8	Parts per million	0.400000	5
9	Parts per million	0.300000	6

Hint 1

Refer to the video about exploratory data analysis in Python.

Hint 2

There is a function in the **pandas** library that allows you to get a specific number of rows from the top of a DataFrame.

Hint 3

Use the **head()** function from the **pandas** library.

**Question:** What does the **aqi** column represent?

[aqi column represent Air Quality Index (AQI)]

Now, get a table that contains some descriptive statistics about the data.

```
[4]: # Get descriptive stats.

### YOUR CODE HERE
epa_data.describe()
```

```
[4]:
```

	arithmetic_mean	aqi
count	260.000000	260.000000
mean	0.403169	6.757692
std	0.317902	7.061707
min	0.000000	0.000000
25%	0.200000	2.000000
50%	0.276315	5.000000
75%	0.516009	9.000000
max	1.921053	50.000000

Hint 1

Refer to the video about descriptive statistics in Python.

Hint 2

There is a function in the **pandas** library that allows you to generate a table of basic descriptive statistics about the numeric columns in a DataFrame.

Hint 3

Use the **describe()** function from the **pandas** library.

**Question:** Based on the table of descriptive statistics, what do you notice about the count value for the `aqi` column?

[aqi column have 260 count values. that means there are 260 aqi measurement represented in this dataset]

**Question:** What do you notice about the 25th percentile for the `aqi` column?

This is an important measure for understanding where the aqi values lie.

[25th percentile for the aqi colum is 2 that means 25% of the aqi value are bellow 2]

**Question:** What do you notice about the 75th percentile for the `aqi` column?

This is another important measure for understanding where the aqi values lie.

[75th percentile for the aqi colum is 9 that means 75% of the aqi value are bellow 9]

## 1.4 Step 3: Statistical tests

Next, get some descriptive statistics about the states in the data.

```
[14]: # Get descriptive stats about the states in the data.  
  
### YOUR CODE HERE  
descriptive_stats= epa_data['state_name'].describe()  
descriptive_stats
```

```
[14]: count          260  
       unique         52  
       top    California  
       freq          66  
       Name: state_name, dtype: object
```

Hint 1

Refer to the video about descriptive statistics in Python.

Hint 2

There is a function in the `pandas` library that allows you to generate basic descriptive statistics about a `DataFrame` or a column you are interested in.

Hint 3

Use the `describe()` function from the `pandas` library. Note that this function can be used: - “on a `DataFrame` (to find descriptive statistics about the numeric columns)” - “directly on a column containing categorical data (to find pertinent descriptive statistics)”

**Question:** What do you notice while reviewing the descriptive statistics about the states in the data?

Note: Sometimes you have to individually calculate statistics. To review to that approach, use the `numpy` library to calculate each of the main statistics in the preceding table for the `aqi` column.

[There are 260 state values, and 52 of them are unique. California is the most commonly occurring state in the data, with a frequency of 66]

## 1.5 Step 4. Results and evaluation

Now, compute the mean value from the `aqi` column.

```
[16]: # Compute the mean value from the aqi column.
```

```
### YOUR CODE HERE
mean_value= epa_data['aqi'].mean()
mean_value
```

```
[16]: 6.757692307692308
```

Hint 1

Refer to the video about descriptive statistics in Python.

Hint 2

There is a function in the `numpy` library that allows you to get the mean value from an array or a Series of values.

Hint 3

Use the `mean()` function from the `numpy` library.

**Question:** What do you notice about the mean value from the `aqi` column?

This is an important measure, as it tells you what the average air quality is based on the data.

[Write your response here. Double-click (or enter) to edit.]

Next, compute the median value from the `aqi` column.

```
[17]: # Compute the median value from the aqi column.
```

```
### YOUR CODE HERE
mean_value= epa_data['aqi'].median()
mean_value
```

```
[17]: 5.0
```

Hint 1

Refer to the video about descriptive statistics in Python.

Hint 2

There is a function in the `numpy` library that allows you to get the median value from an array or a series of values.

Hint 3

Use the `median()` function from the `numpy` library.

**Question:** What do you notice about the median value from the `aqi` column?

This is an important measure for understanding the central location of the data.

[Write your response here. Double-click (or enter) to edit.]

Next, identify the minimum value from the `aqi` column.

```
[21]: # Identify the minimum value from the aqi column.  
  
### YOUR CODE HERE  
np.min(epa_data['aqi'])
```

[21]: 0

Hint 1

Refer to the video about descriptive statistics in Python.

Hint 2

There is a function in the `numpy` library that allows you to get the minimum value from an array or a Series of values.

Hint 3

Use the `min()` function from the `numpy` library.

**Question:** What do you notice about the minimum value from the `aqi` column?

This is an important measure, as it tell you the best air quality observed in the data.

[Write your response here. Double-click (or enter) to edit.]

Now, identify the maximum value from the `aqi` column.

```
[22]: # Identify the maximum value from the aqi column.  
  
### YOUR CODE HERE  
np.max(epa_data['aqi'])
```

[22]: 50

Hint 1

Refer to the video about descriptive statistics in Python.

Hint 2

There is a function in the `numpy` library that allows you to get the maximum value from an array or a Series of values.

Hint 3

Use the `max()` function from the `numpy` library.

**Question:** What do you notice about the maximum value from the `aqi` column?

This is an important measure, as it tells you which value in the data corresponds to the worst air quality observed in the data.

[Write your response here. Double-click (or enter) to edit.]

Now, compute the standard deviation for the `aqi` column.

By default, the `numpy` library uses 0 as the Delta Degrees of Freedom, while `pandas` library uses 1. To get the same value for standard deviation using either library, specify the `ddof` parameter to 1 when calculating standard deviation.

```
[23]: # Compute the standard deviation for the aqi column.  
  
### YOUR CODE HERE  
np.std(epa_data['aqi'])
```

```
[23]: 7.048113390607722
```

Hint 1

Refer to the video section about descriptive statistics in Python.

Hint 2

There is a function in the `numpy` library that allows you to get the standard deviation from an array or a series of values.

Hint 3

Use the `std()` function from the `numpy` library. Make sure to specify the `ddof` parameter as 1. To read more about this function, refer to its documentation in the references section of this lab.

**Question:** What do you notice about the standard deviation for the `aqi` column?

This is an important measure of how spread out the `aqi` values are.

[Write your response here. Double-click (or enter) to edit.]

## 1.6 Considerations

**What are some key takeaways that you learned during this lab?**

[Write your response here. Double-click (or enter) to edit.]

**How would you present your findings from this lab to others? Consider the following relevant points noted by AirNow.gov as you respond:** - “AQI values at or below 100 are generally thought of as satisfactory. When AQI values are above 100, air quality is considered to be unhealthy—at first for certain sensitive groups of people, then for everyone as AQI values increase.” - “An AQI of 100 for carbon monoxide corresponds to a level of 9.4 parts per million.”

[Write your response here. Double-click (or enter) to edit.]

**What summary would you provide to stakeholders? Use the same information provided previously from AirNow.gov as you respond.**

[Write your response here. Double-click (or enter) to edit.]

## References

[Air Quality Index - A Guide to Air Quality and Your Health](#). (2014,February)

[Numpy.Std — NumPy v1.23 Manual](#)

US EPA, OAR. (2014, 8 July).*Air Data: Air Quality Data Collected at Outdoor Monitors Across the US*.

**Congratulations!** You’ve completed this lab. However, you may not notice a green check mark next to this item on Coursera’s platform. Please continue your progress regardless of the check mark. Just click on the “save” icon at the top of this notebook to ensure your work has been logged.