Aberrant Data

Lesson 3

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Aberrant Data

Overview

Data Cleaning

- Removal
- Imputation

Data Types

Missing Values

- Null Values
- Removal

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Data Cleaning

Removal

Erroneous Inputs – an example

Python Numpy library – for working with numbers

```
import numpy as np
```

Examine an array of age of preschooler in daycare

```
x = np.array([2, 1, 1, 99, 1, 5, 3, 1, 4, 3])
```

–99 is an outlier

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Outliers

2+ standard deviations from the mean

–Gaussian distribution has 95% of values within 2 stds

Use numpy for mean & standard deviation

```
LimitHi = np.mean(x) + 2*np.std(x)
LimitLo = np.mean(x) - 2*np.std(x)
```


High and low limits

Name	Type	Size	Value
LimitHi	float64	1	70.062035789317619
LimitLo	float64	1	-46.062035789317619
x	int32	(10,)	array([2, 1, 1, 99, 1, 5, 3, 1, 4, 3])

What values are the Outliers?

Flag: A Boolean array

```
FlagGood = (x >= LimitLo) & (x <= LimitHi)
```

Element-wise comparison

–Checks each value in the array

Element-wise logical AND

–Operator &

```
array([ True,  True,  True, False,  True,
True,  True,  True,  True,  True],
dtype=bool)
```

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Indexing an array with a Boolean array

```
x[FlagGood]
```

Show the value in x if the FlagGood is True:

```
array([2, 1, 1, 1, 5, 3, 1, 4, 3])
```

Recall original array:

x	int32	(10,)	array([2, 1, 1, 99, 1, 5, 3, 1, 4, 3])

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Remove outlier from data set

```
x = x[FlagGood]
```

How big is the new data set?

Name	Type	Size	Value
FlagGood	bool	(10,)	ndarray object of numpy module
LimitHi	float64	1	70.062035789317619
LimitLo	float64	1	-46.062035789317619
x	int32	(9,)	array([2, 1, 1, 1, 5, 3, 1, 4, 3])

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Data Cleaning

Replacement

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Imputation

Replacing an outlier value with a guess

Mean imputation

–Arithmetic mean replaces every outlier

Replacing outliers – an example

```
import numpy as np
x = np.array([2, 1, 1, 99., 1, 5, 3, 1, 4, 3])
```

–99 is the outlier

```
LimitHi = np.mean(x) + 2*np.std(x)
LimitLo = np.mean(x) - 2*np.std(x)
```

–Limits within 2 standard deviations from the mean

Flag the Outlier Values

```
FlagBad = (x < LimitLo) | (x > LimitHi)
```

- –True for every outlier in the array
- –False for values within the limits

Element-wise operator OR

- –Operator |

```
array([False, False, False,  True, False,
False, False, False, False, False], dtype=bool)
```

- –4th value is the outlier

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Replace outliers with the mean

Use the Flag as the index on the array:

```
x[FlagBad] = np.mean(x)
```

```
array([  2.,   1.,   1.,  12.,   1.,   5.,
3.,   1.,   4.,   3.])
```

Mean of the dataset = 12

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Imputation with outlier influence

12 is too large for age of preschooler.

Need mean of values that are not outliers

`FlagGood = ~FlagBad`

–Complement operator ~

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Imputation without outlier influence

Mean without outliers

`np.mean(x[FlagGood])`

```
np.mean(x[FlagGood])

2.3333333333333335
```

Replacement:

`x[FlagBad] = np.mean(x[FlagGood])`

```
array([ 2.        ,  1.        ,  1.        ,  2.33333333,  1.        ,
        5.        ,  3.        ,  1.        ,  4.        ,  3.        ])
```

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Median Imputation

Median is less sensitive to outliers

Before

```
array([  2.,   1.,   1.,  99.,   1.,   5.,   3.,   1.,   4.,   3.])
```

```
x[FlagBad] = np.median(x)
```

After

```
array([ 2. ,  1. ,  1. ,  2.5,  1. ,  5. ,  3. ,  1. ,  4. ,  3. ])
```

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Summary

>Use Boolean flags to create new arrays
- –FlagBad are values outside of the limits
- –FlagGood = ~FlagBad , values inside the limits

>Remove Outliers
- –Keep the FlagGood values

>Replace Outliers (Imputation)
- –Use the mean without the outlier
- –Use the median even with the outlier