Center for
Data Science

NYU

DS-UA 112
Introduction to Data Science

Lecture 6

Cleaning Data – Missing or Inconsistent Values

Agenda

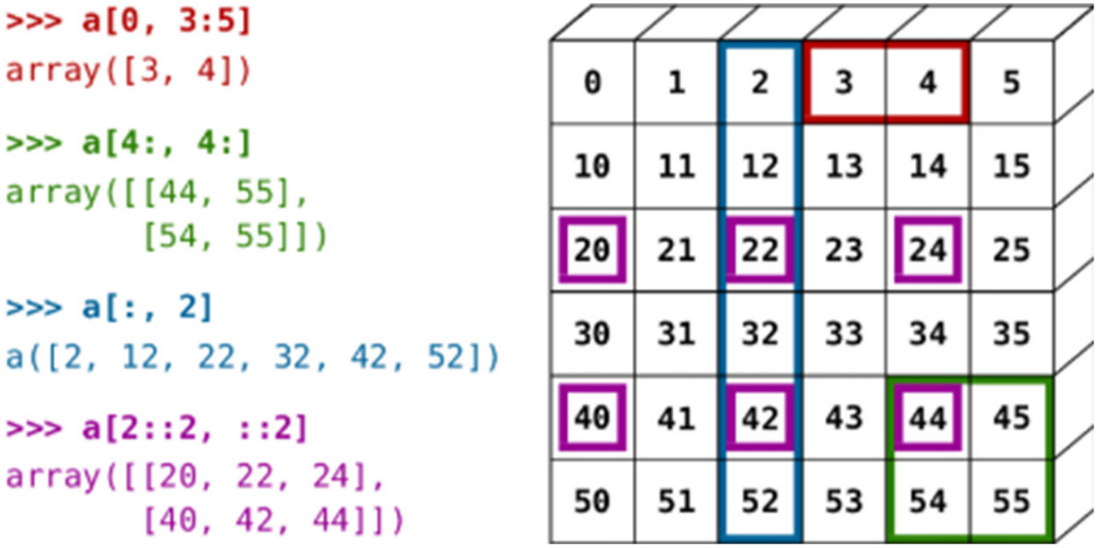
- Review
- Lesson
- Demo

Reminders

- Survey 2
- Homework
  - Homework 2
  - Forum
  - Grader Contact Information
- Final Exam

Review

- numpy

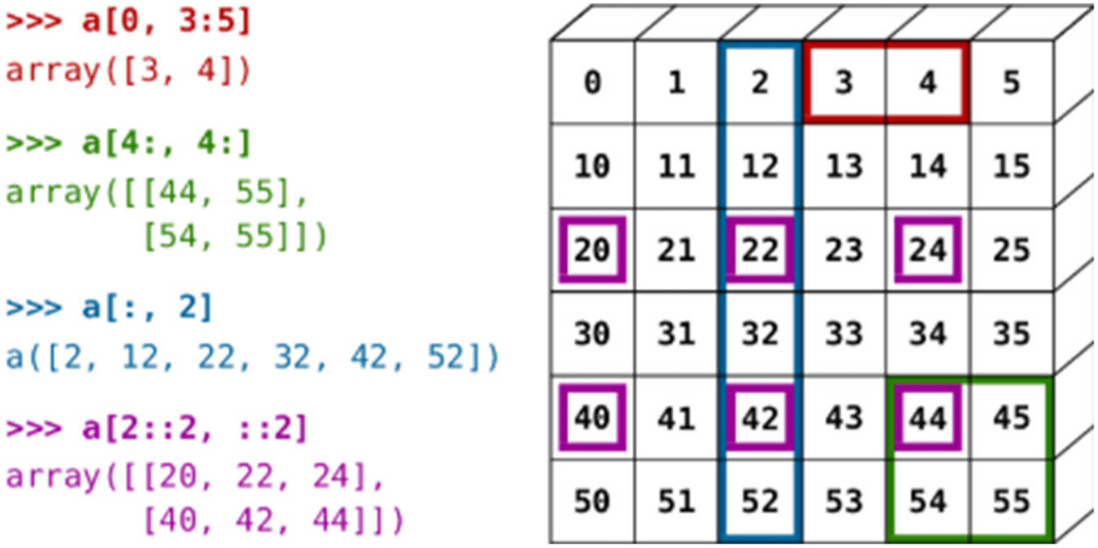


```
>>> a[0, 3:5]
array([3, 4])

>>> a[4:, 4:]
array([[44, 55],
       [54, 55]])

>>> a[:, 2]
a([2, 12, 22, 32, 42, 52])

>>> a[2::2, ::2]
array([[20, 22, 24],
       [40, 42, 44]])
```

0	1	2	3	4	5
10	11	12	13	14	15
20	21	22	23	24	25
30	31	32	33	34	35
40	41	42	43	44	45
50	51	52	53	54	55

Review

- numpy
- Fixed
  - Size
  - Data Type

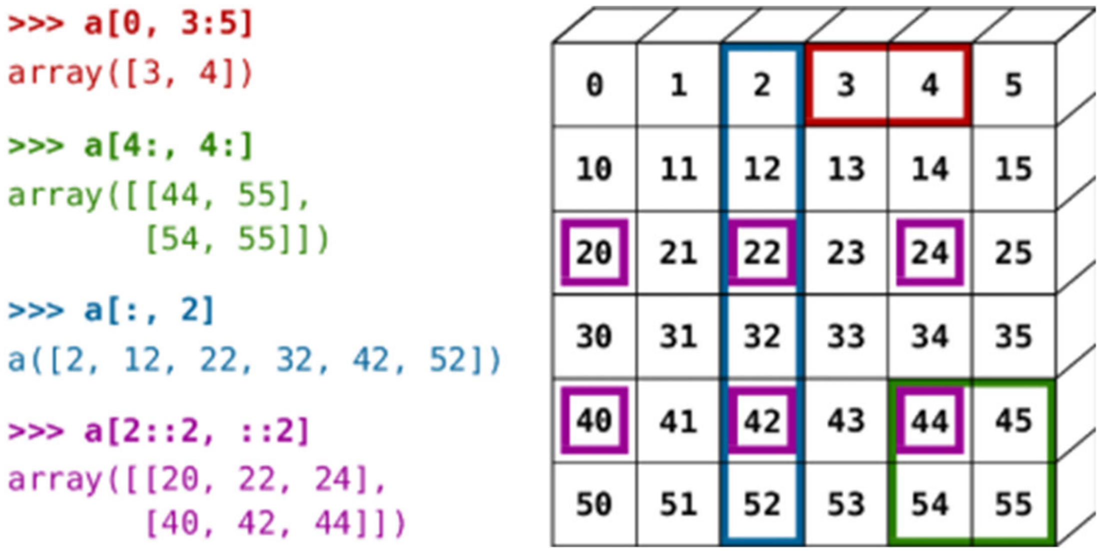


```
>>> a[0, 3:5]
array([3, 4])

>>> a[4:, 4:]
array([[44, 55],
       [54, 55]])

>>> a[:, 2]
a([2, 12, 22, 32, 42, 52])

>>> a[2::2, ::2]
array([[20, 22, 24],
       [40, 42, 44]])
```

0	1	2	3	4	5
10	11	12	13	14	15
20	21	22	23	24	25
30	31	32	33	34	35
40	41	42	43	44	45
50	51	52	53	54	55

Review

- numpy
- Fixed
    - Size
    - Data Type
- Element by Element
    - Broadcasting
    - Vectorization

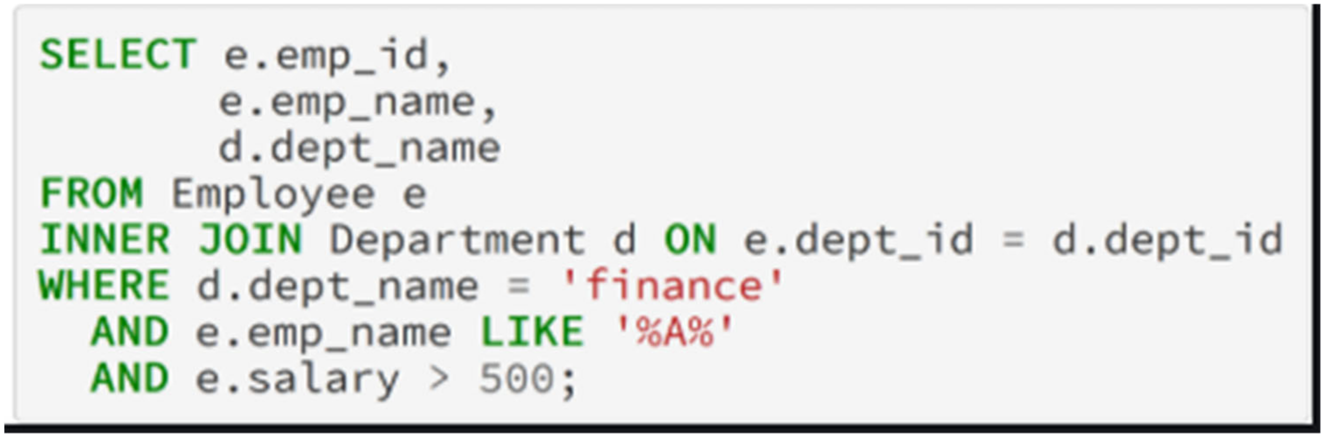


```
>>> a[0, 3:5]
array([3, 4])

>>> a[4:, 4:]
array([[44, 55],
       [54, 55]])

>>> a[:, 2]
a([2, 12, 22, 32, 42, 52])

>>> a[2::2, ::2]
array([[20, 22, 24],
       [40, 42, 44]])
```

0	1	2	3	4	5
10	11	12	13	14	15
20	21	22	23	24	25
30	31	32	33	34	35
40	41	42	43	44	45
50	51	52	53	54	55

Review

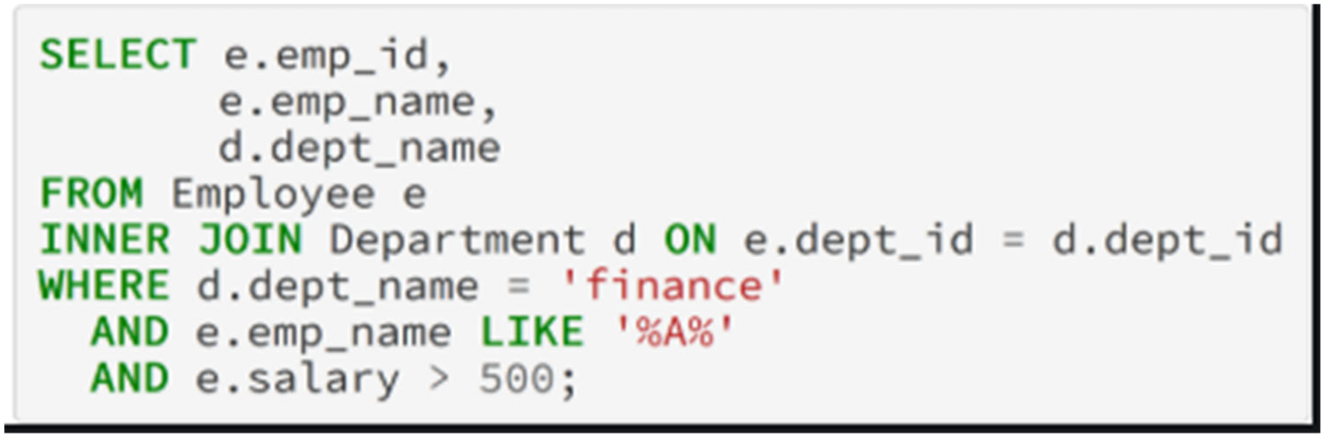
- SQL



```sql
SELECT e.emp_id,
       e.emp_name,
       d.dept_name
FROM Employee e
INNER JOIN Department d ON e.dept_id = d.dept_id
WHERE d.dept_name = 'finance'
  AND e.emp_name LIKE '%A%'
  AND e.salary > 500;
```

Review

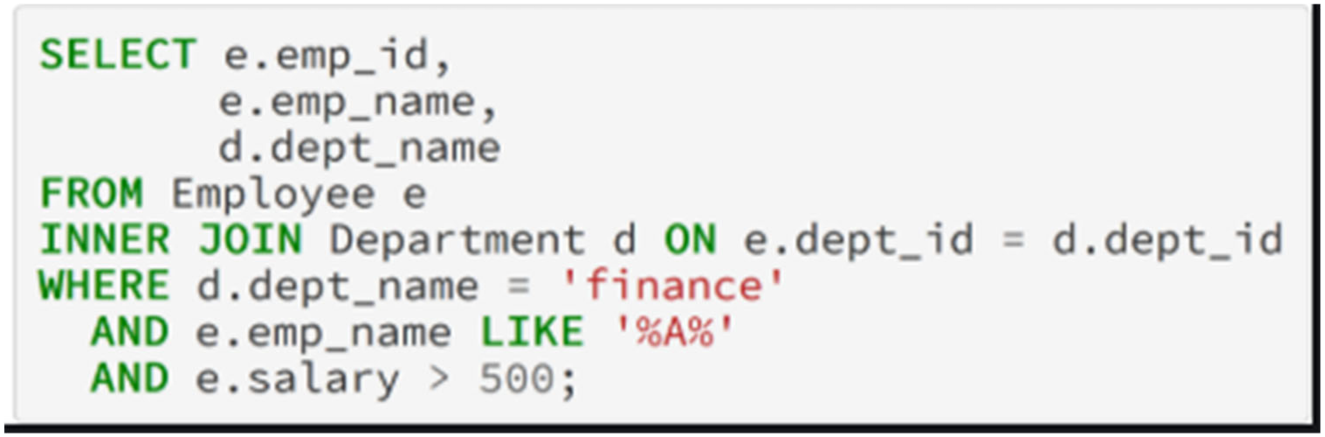
- SQL
  - Tables
    - Rows
    - Columns



```
SELECT e.emp_id,
       e.emp_name,
       d.dept_name
FROM Employee e
INNER JOIN Department d ON e.dept_id = d.dept_id
WHERE d.dept_name = 'finance'
  AND e.emp_name LIKE '%A%'
  AND e.salary > 500;
```

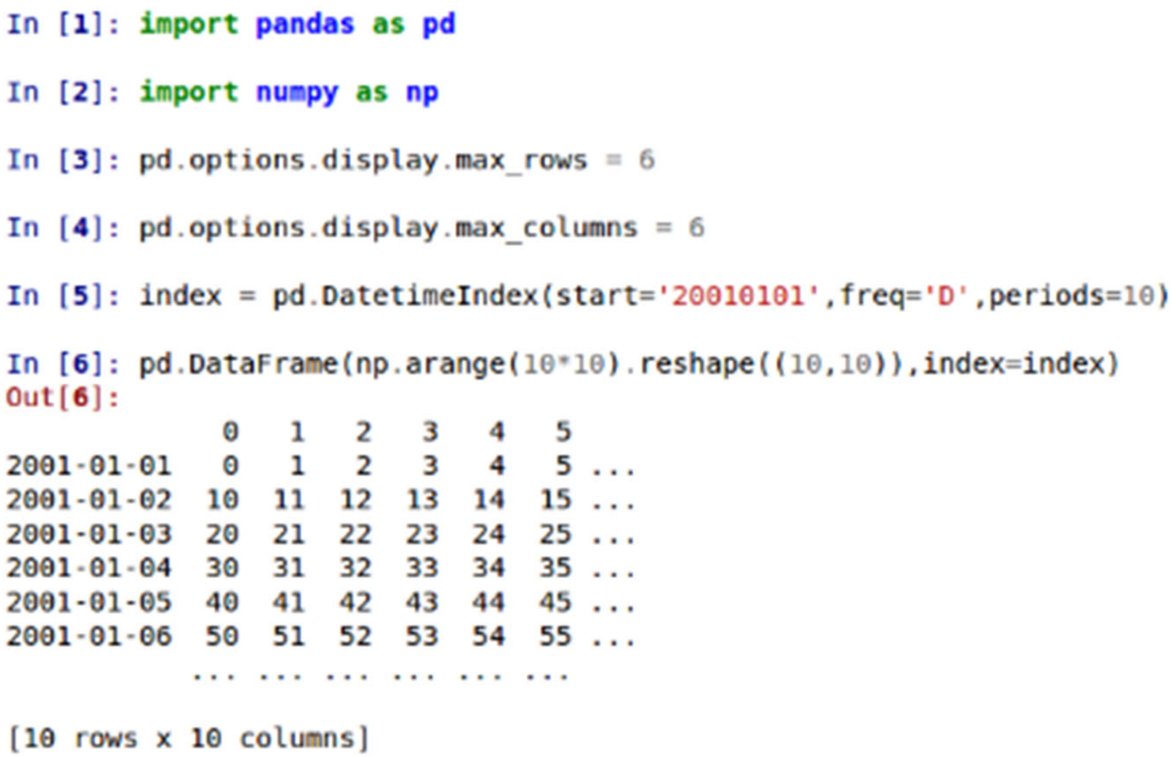
Review

- SQL
  - Tables
    - Rows
    - Columns
  - Keys
    - Primary
    - Foreign



```sql
SELECT e.emp_id,
       e.emp_name,
       d.dept_name
FROM Employee e
INNER JOIN Department d ON e.dept_id = d.dept_id
WHERE d.dept_name = 'finance'
  AND e.emp_name LIKE '%A%'
  AND e.salary > 500;
```

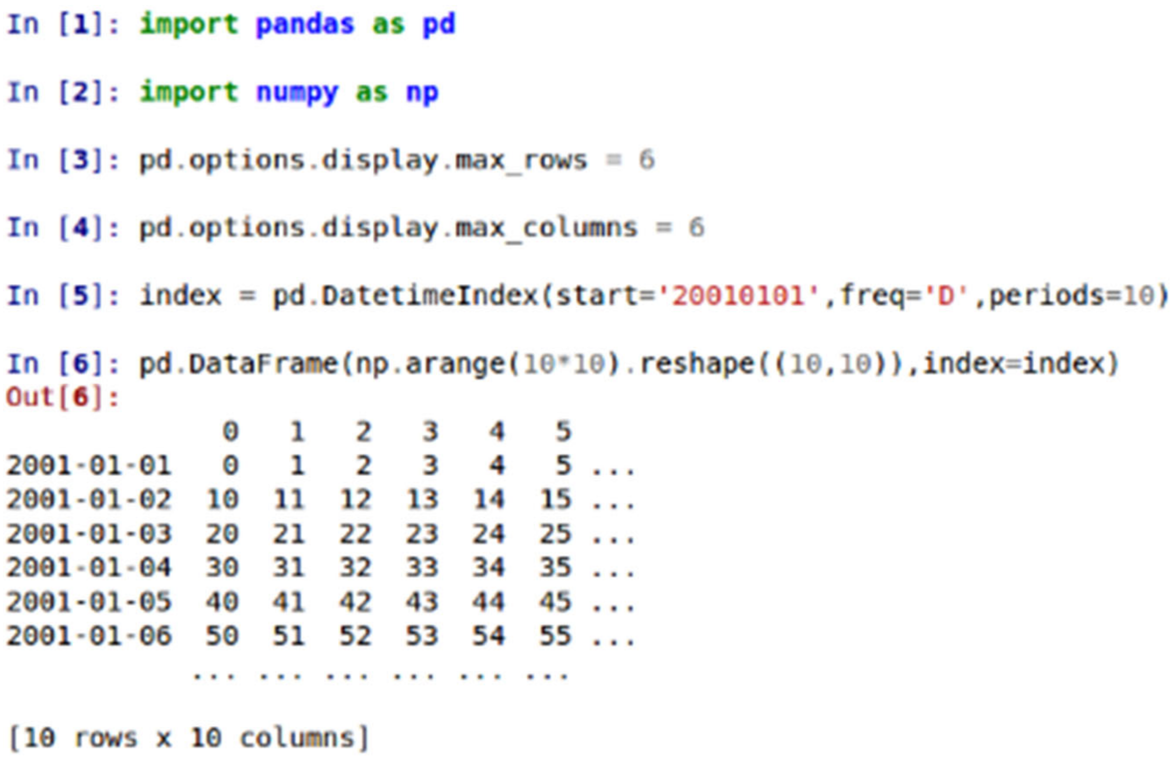
Review



```
In [1]: import pandas as pd

In [2]: import numpy as np

In [3]: pd.options.display.max_rows = 6

In [4]: pd.options.display.max_columns = 6

In [5]: index = pd.DatetimeIndex(start='20010101',freq='D',periods=10)

In [6]: pd.DataFrame(np.arange(10*10).reshape((10,10)),index=index)
Out[6]:
             0   1   2   3   4   5
2001-01-01   0   1   2   3   4   5 ...
2001-01-02  10  11  12  13  14  15 ...
2001-01-03  20  21  22  23  24  25 ...
2001-01-04  30  31  32  33  34  35 ...
2001-01-05  40  41  42  43  44  45 ...
2001-01-06  50  51  52  53  54  55 ...
           ... ... ... ... ... ...

[10 rows x 10 columns]
```

Review

- ▶ Similarities
- ▶ Differences



```
In [1]: import pandas as pd

In [2]: import numpy as np

In [3]: pd.options.display.max_rows = 6

In [4]: pd.options.display.max_columns = 6

In [5]: index = pd.DatetimeIndex(start='20010101',freq='D',periods=10)

In [6]: pd.DataFrame(np.arange(10*10).reshape((10,10)),index=index)
Out[6]:
             0   1   2   3   4   5
2001-01-01   0   1   2   3   4   5 ...
2001-01-02  10  11  12  13  14  15 ...
2001-01-03  20  21  22  23  24  25 ...
2001-01-04  30  31  32  33  34  35 ...
2001-01-05  40  41  42  43  44  45 ...
2001-01-06  50  51  52  53  54  55 ...
           ... ... ... ... ... ...

[10 rows x 10 columns]
```

Demo

- Remember
  - Solve problems with data…using programming!

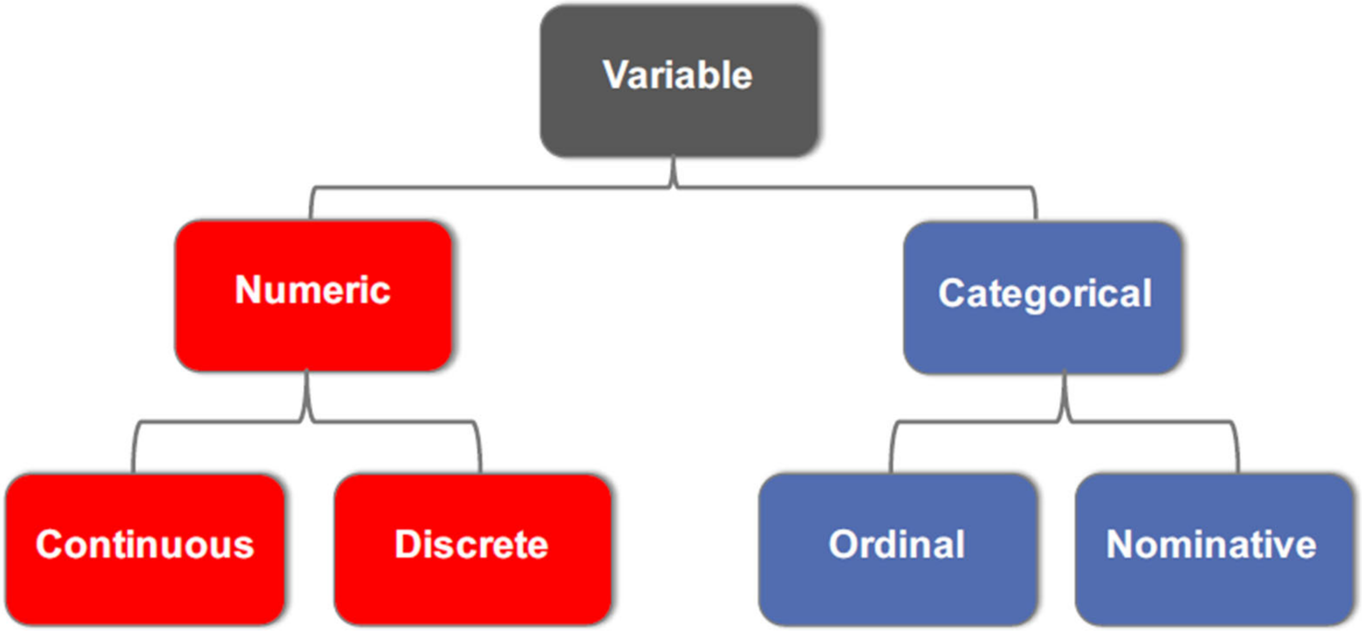
Demo

- Remember
  - Solve problems with data…using programming!
  - Focus on common practices between languages!

Demo

- Remember
  - Solve problems with data…using programming!
  - Focus on common practices between languages!
  - Many tools for data science…in particular, many uses for Python!

Lesson



Variable

Numeric

Categorical

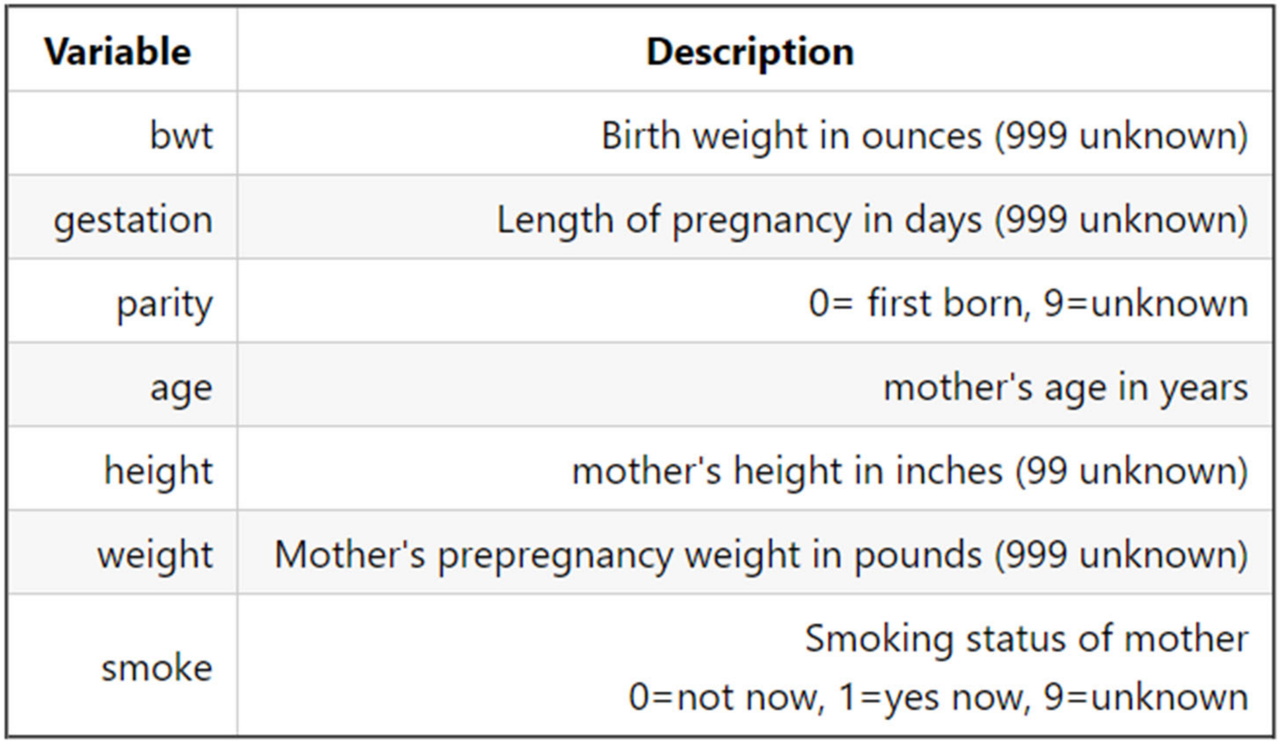
Continuous

Discrete

Ordinal

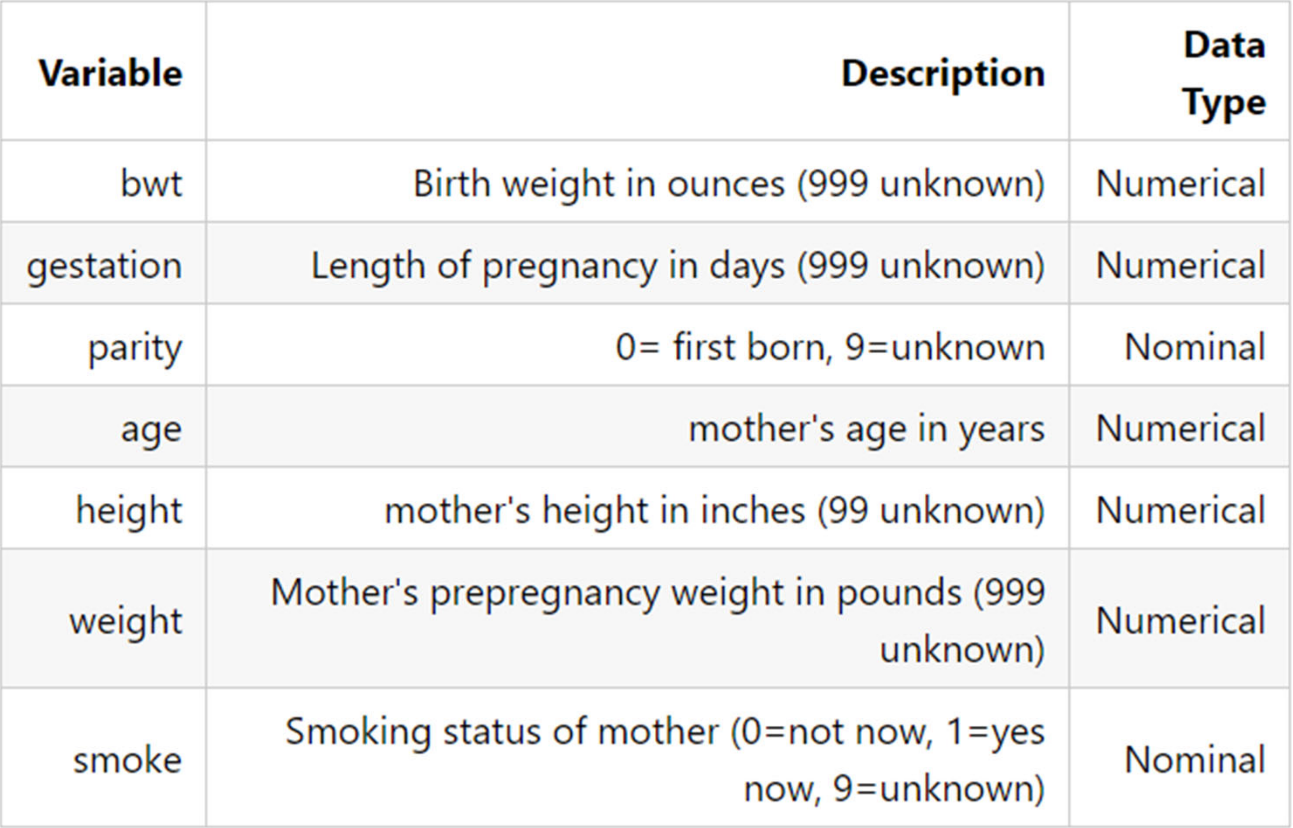
Nominative

Lesson



Variable	Description
bwt	Birth weight in ounces (999 unknown)
gestation	Length of pregnancy in days (999 unknown)
parity	0= first born, 9=unknown
age	mother's age in years
height	mother's height in inches (99 unknown)
weight	Mother's prepregnancy weight in pounds (999 unknown)
smoke	Smoking status of mother 0=not now, 1=yes now, 9=unknown

Lesson



Variable	Description	Data Type
bwt	Birth weight in ounces (999 unknown)	Numerical
gestation	Length of pregnancy in days (999 unknown)	Numerical
parity	0= first born, 9=unknown	Nominal
age	mother's age in years	Numerical
height	mother's height in inches (99 unknown)	Numerical
weight	Mother's prepregnancy weight in pounds (999 unknown)	Numerical
smoke	Smoking status of mother (0=not now, 1=yes now, 9=unknown)	Nominal

Demo

- Missing Values

Take-Aways

- What to do about bias?
  - Avoid it

Take-Aways

- What to do about bias?
  - Avoid it
  - Adjust it

Take-Aways

- What to do about bias?
  - Avoid it
  - Adjust it
  - Expect it

Take-Aways

- What to do about bias?
  - Avoid it
  - Adjust it
  - Expect it
- What to do about missing or inconsistent data?
  - Avoid it

Take-Aways

- What to do about bias?
  - Avoid it
  - Adjust it
  - Expect it
- What to do about missing or inconsistent data?
  - Avoid it
  - Adjust it

Take-Aways

- What to do about bias?
  - Avoid it
  - Adjust it
  - Expect it
- What to do about missing or inconsistent data?
  - Avoid it
  - Adjust it
  - Expect it