Missing value handling in Python

MISSING VALUE HANDLING IN PYTHON

```
import pandas as pd
import os
```

Missing value handling in Python

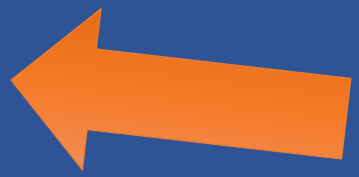
In this session

1. Replace missing values with the mean
2. Replace missing values with the median
3. Replace missing values with an interpolated estimate
4. Replace missing values with a constant
5. Replace missing values using imputation
6. Replace missing values with a missing rank
7. Replace missing values with a dummy
8. Replace missing values with 0
9. Create an indicator variable for "missing."
10. Replace missing values with a string
11. Add an indicator variable showing which strings are considered "missing."
12. Delete columns that are missing too many values to be useful
13. Delete rows that are missing critical values

Missing value handling in Python

We need data that is:

- Relevant
- Connected
- Accurate
- Enough to work with



Missing value handling in Python

Example of missing values dataset CSV file

missing_values.csv

	A	B	C	D	E	F	G	H	I
1		age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact
2	Tony	48	27		1	5	shrimp		Pepper
3	Donald	67	25	86	10	2	beef		Jane
4	Henry	69	21	95	6	1	chicken	62	Janet
5	Janet	62	21	110	3	1	beef		Henry
6	Nick		17		4				
7	Bruce	37	14	63		1	veggie		NA
8	Steve	83		77	7	1	chicken		n/a
9	Clint	27	9	118	9		shrimp	3	None
10	Wanda	19	7	52	2	2	shrimp		empty
11	Natasha	26	4	162	5	3			_
12	Carol		3	127	11	1	veggie	1	""
13	Mandy	44	2	68	8	1	chicken		null

Missing value handling in Python

General commands

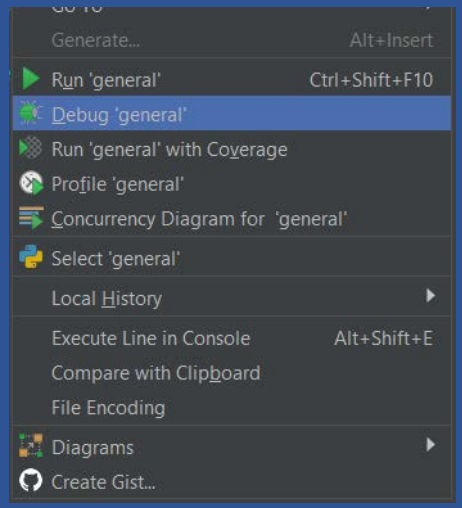
```python
import pandas as pd
import os

# General useful commands
os.chdir("d:\\temp")    # change current directory
# Read CSV to pandas Data frame
df = pd.read_csv('missing_values.csv')
print(type(df))     # show data type of df
print(list(df))     # show column list
print(df)           # show all rows & columns
# column projection limit
print(df[['name', 'age', 'income']])
# row limit
print(df[['name', 'age', 'income']].head(n=5))
```

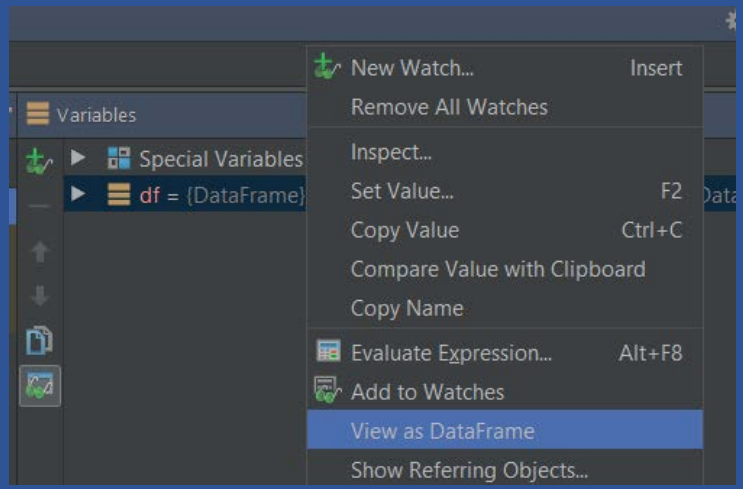
Missing value handling in Python

View variable as Data Frame

	name	age	years_senio...	income	parking_sp...	attending_...	entree	pets	emergency...
0	Tony	48.00000	27.00000	nan	1.00000	5.00000	shrimp	nan	Pepper
1	Donald	67.00000	25.00000	86.00000	10.00000	2.00000	beef	nan	Jane
2	Henry	69.00000	21.00000	95.00000	6.00000	1.00000	chicken	62.00000	Janet
3	Janet	62.00000	21.00000	110.00000	3.00000	1.00000	beef	nan	Henry
4	Nick	nan	17.00000	nan	4.00000	nan	nan	nan	nan
5	Bruce	37.00000	14.00000	63.00000	nan	1.00000	veggie	nan	nan
6	Steve	83.00000	nan	77.00000	7.00000	1.00000	chicken	nan	n/a
7	Clint	27.00000	9.00000	118.00000	9.00000	nan	shrimp	3.00000	None
8	Wanda	19.00000	7.00000	52.00000	2.00000	2.00000	shrimp	nan	empty
9	Natasha	26.00000	4.00000	162.00000	5.00000	3.00000	nan	nan	_
10	Carol	nan	3.00000	127.00000	11.00000	1.00000	veggie	1.00000	
11	Mandy	44.00000	2.00000	68.00000	8.00000	1.00000	chicken	nan	null

```
1   + import ...
3
4     # General useful commands
5     os.chdir("d:\\temp")     # change
6     # Read CSV to pandas Data frame
7     df = pd.read_csv('missing_values
8     print(type(df))          # show data
9     print(list(df))          # show colum
10    print(df)                # show all r
```





Missing value handling in Python

Replace missing values with the mean

```
1   import pandas as pd
2   import os
3
4   # Replace missing values with the mean (age)
5   # column = age
6   os.chdir("d:\\temp")
7   df = pd.read_csv('missing_values.csv')
8   print(type(df))
9   print(list(df))
10  print(df[['name', 'age', 'income']])
11  # replace missing value with mean
12  # and create a new column 'age1'
13  df['age1'] = df[['age']].fillna(df.mean()['age':'age'])
14  print(df[['age', 'age1']])
15  print("end")
16  print("***")
```

```
   age  age1
48.0  48.0
67.0  67.0
69.0  69.0
62.0  62.0
 NaN  48.2
37.0  37.0
83.0  83.0
27.0  27.0
19.0  19.0
26.0  26.0
 NaN  48.2
44.0  44.0
```

Missing value handling in Python

Replace missing values with the median

```
1   import pandas as pd
2   import os
3
4   # Replace missing values with the median
5   # column = age
6   os.chdir("d:\\temp")
7   df = pd.read_csv('missing_values.csv')
8   print(type(df))
9   print(list(df))
10  print(df[['name', 'age', 'income']])
11  df['age1'] = df[['age']].fillna(df.median()['age':'age'])
12  print(df[['age', 'age1']])
13  print("end")
14  print("***")
```

```
   age  age1
0  48.0  48.0
1  67.0  67.0
2  69.0  69.0
3  62.0  62.0
4   NaN  46.0
5  37.0  37.0
6  83.0  83.0
7  27.0  27.0
8  19.0  19.0
9  26.0  26.0
10  NaN  46.0
11 44.0  44.0
```

Missing value handling in Python

Replace missing values with an interpolated estimate

```
    import pandas as pd
    import os

    # Replace missing values with an interpolated estimate
    # column = years_seniority
    os.chdir("d:\\temp")
    df = pd.read_csv('missing_values.csv')
    print(type(df))
    print(list(df))
    print(df[['name', 'age', 'years_seniority']])
    df['years_seniority1'] = df[['years_seniority']].fillna(11.5)
    print(df[['years_seniority', 'years_seniority1']])
    print("end")
    print("***")
```

```
   years_seniority  years_seniority1
0             27.0              27.0
1             25.0              25.0
2             21.0              21.0
3             21.0              21.0
4             17.0              17.0
5             14.0              14.0
6              NaN              11.5
7              9.0               9.0
8              7.0               7.0
9              4.0               4.0
10             3.0               3.0
11             2.0               2.0
```

Missing value handling in Python

Replace missing values with a constant

```
1   import pandas as pd
2   import os
3
4   # Replace missing values with a constant
5   # column = income
6   os.chdir("d:\\temp")
7   df = pd.read_csv('missing_values.csv')
8   print(type(df))
9   print(list(df))
10  print(df[['name', 'age', 'income']])
11  df['income1'] = df[['income']].fillna(250)
12  print(df[['income', 'income1']])
13  print("end")
14  print("***")
```

income	income1
NaN	250.0
86.0	86.0
95.0	95.0
110.0	110.0
NaN	250.0
63.0	63.0
77.0	77.0
118.0	118.0
52.0	52.0
162.0	162.0
127.0	127.0
68.0	68.0

Missing value handling in Python

Replace missing values using imputation (MICE)


```
[[  48.          27.          91.01255587]
 [  67.          25.          86.        ]
 [  69.          21.          95.        ]
 [  62.          21.         110.        ]
 [  49.20878949  17.          90.94503833]
 [  37.          14.          63.        ]
 [  83.          26.37257548  77.        ]
 [  27.           9.         118.        ]
 [  19.           7.          52.        ]
 [  26.           4.         162.        ]
 [  25.13935223   3.         127.        ]
 [  44.           2.          68.        ]]
```

```python
import mice
import pandas as pd
import os
import numpy as np

os.chdir("d:\\temp")
df = pd.read_csv('missing_values.csv')
df2 = df[['age', 'years_seniority', 'income']]
a = np.array(df2)
x = mice.MICE().complete(a)
print(df2)
print('----------------')
print(x)
```

Missing value handling in Python

Replace missing values with a missing rank

```
1   import pandas as pd
2   import os
3
4   # Replace missing values with a missing rank
5   # column = parking_space
6   os.chdir("d:\\temp")
7   df = pd.read_csv('missing_values.csv')
8   print(type(df))
9   print(list(df))
10  print(df[['name', 'age', 'parking_space']])
11  # Missing one might be 12
12  df['park1'] = df[['parking_space']].fillna(12)
13  print(df[['parking_space', 'park1']])
14  print("end")
15  print("***")
```

parking_space	park1
1.0	1.0
10.0	10.0
6.0	6.0
3.0	3.0
4.0	4.0
NaN	12.0
7.0	7.0
9.0	9.0
2.0	2.0
5.0	5.0
11.0	11.0
8.0	8.0

Missing value handling in Python

Replace missing values with a dummy

```
1   import pandas as pd
2   import os
3
4   # Replace missing values with a dummy
5   # column = parking_space
6   os.chdir("d:\\temp")
7   df = pd.read_csv('missing_values.csv')
8   print(type(df))
9   print(list(df))
10  print(df[['name', 'age', 'parking_space']])
11  # dummy is -99
12  df['park1'] = df[['parking_space']].fillna(-99)
13  print(df[['parking_space', 'park1']])
14  print("end")
15  print("***")
```

parking_space	park1
1.0	1.0
10.0	10.0
6.0	6.0
3.0	3.0
4.0	4.0
NaN	-99.0
7.0	7.0
9.0	9.0
2.0	2.0
5.0	5.0
11.0	11.0
8.0	8.0

Missing value handling in Python

Replace missing values with 0

```
1   import pandas as pd
2   import os
3
4   # Replace missing values with 0
5   # column = attending_party
6   os.chdir("d:\\temp")
7   df = pd.read_csv('missing_values.csv')
8   print(type(df))
9   print(list(df))
10  print(df[['name', 'age', 'attending_party']])
11  df['party'] = df[['attending_party']].fillna(0)
12  print(df[['attending_party', 'party']])
13  print("end")
```

```
   attending_party  party
0              5.0    5.0
1              2.0    2.0
2              1.0    1.0
3              1.0    1.0
4              NaN    0.0
5              1.0    1.0
6              1.0    1.0
7              NaN    0.0
8              2.0    2.0
9              3.0    3.0
10             1.0    1.0
11             1.0    1.0
```

Missing value handling in Python

Create an indicator variable for "missing"

```
1   import pandas as pd
2   import os
3
4   # Create an indicator variable for "missi
5   # column = pets
6   os.chdir("d:\\temp")
7   df = pd.read_csv('missing_values.csv')
8   df['pets1'] = df[['pets']].fillna(0)
9   df['pets2'] = df[['pets1']].isin([0])
10  print(df[['name', 'pets', 'pets1', 'pets2']])
11  print("end")
```

name	pets	pets1	pets2
Tony	NaN	0.0	True
Donald	NaN	0.0	True
Henry	62.0	62.0	False
Janet	NaN	0.0	True
Nick	NaN	0.0	True
Bruce	NaN	0.0	True
Steve	NaN	0.0	True
Clint	3.0	3.0	False
Wanda	NaN	0.0	True
Natasha	NaN	0.0	True
Carol	1.0	1.0	False
Mandy	NaN	0.0	True

Missing value handling in Python

Replace missing values with a string

```
emergency_contact      e1
           Pepper  Pepper
             Jane    Jane
            Janet   Janet
            Henry   Henry
              NaN      no
              NaN      no
              n/a     n/a
             None    None
            empty   empty
                _       _
               ""      ""
             null    null
```

```
1   import pandas as pd
2   import os
3
4   # Replace missing values with a string
5   # column = emergency_contact
6   os.chdir("d:\\temp")
7   df = pd.read_csv('missing_values.csv')
8   df['e1'] = df[['emergency_contact']].fillna('no')
9   print(df[['emergency_contact', 'e1']])
10  print("end")
```

Missing value handling in Python

Add an indicator variable showing which strings are considered "missing."

```
1   import pandas as pd
2   import os
3
4   # Add an indicator variable showing which
5   # strings are considered "missing."
6   # column = emergency_contact
7   os.chdir("d:\\temp")
8   df = pd.read_csv('missing_values.csv')
9   k = ['NA', 'n/a', 'None', 'empty', '_', '""', 'null']
10  df['e1'] = df[['emergency_contact']].isin(k)
11  print(df[['name', 'emergency_contact', 'e1']])
12  print("end")
```

	name	emergency_contact	e1
0	Tony	Pepper	False
1	Donald	Jane	False
2	Henry	Janet	False
3	Janet	Henry	False
4	Nick	NaN	False
5	Bruce	NaN	False
6	Steve	n/a	True
7	Clint	None	True
8	Wanda	empty	True
9	Natasha	_	True
10	Carol	""	True
11	Mandy	null	True

Missing value handling in Python

Delete columns that are missing too many values to be useful

```
1   import pandas as pd
2   import os
3
4   # Delete columns that are missing too
5   # many values to be useful
6   # column = pets
7   os.chdir("d:\\temp")
8   df = pd.read_csv('missing_values.csv')
9   del df['pets']
10  print(df)
11  print("end")
```

Missing value handling in Python

Delete rows that are missing critical values

```
1   import pandas as pd
2   import os
3
4   # Delete row that are missing too
5   # many values to be useful
6   os.chdir("d:\\temp")
7   df = pd.read_csv('missing_values.csv')
8   df = df.dropna(how='any')   # 'all'
9   print(df)
10  print("end")
```

Missing value handling in Python

More information on Missing value handling in Python

Pandas 0.20.1 documentation: Working with missing data

https://pandas.pydata.org/pandas-docs/stable/missing_data.html

Source code

https://github.com/laploy/ML/blob/master/missing%20value%20python%20msvs.zip