Data Wrangling I

May 1, 2022

[1]:
```
import pandas as pd
import numpy as np
```

[2]:
```
iris = pd.read_csv("IRIS.csv")
```

1 Data Pre-processing

[3]:
```
iris.head()
```

[3]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

[4]:
```
iris.tail(5)
```

[4]:

	sepal_length	sepal_width	petal_length	petal_width	species
145	6.7	3.0	5.2	2.3	Iris-virginica
146	6.3	2.5	5.0	1.9	Iris-virginica
147	6.5	3.0	5.2	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica
149	5.9	3.0	5.1	1.8	Iris-virginica

[5]:
```
iris.index
```

[5]:
```
RangeIndex(start=0, stop=150, step=1)
```

[6]:
```
iris.columns
```

[6]:
```
Index(['sepal_length', 'sepal_width', 'petal_length', 'petal_width',
       'species'],
      dtype='object')
```

[7]:
```
iris.shape
```

[7]:
```
(150, 5)
```

[8]:
```
iris.dtypes
```

[8]:
```
sepal_length    float64
sepal_width     float64
petal_length    float64
petal_width     float64
species          object
dtype: object
```

[9]:
```
iris.columns.values
```

[9]:
```
array(['sepal_length', 'sepal_width', 'petal_length', 'petal_width',
       'species'], dtype=object)
```

[10]:
```
iris.describe(include='all')
```

[10]:

	sepal_length	sepal_width	petal_length	petal_width	species
count	150.000000	150.000000	150.000000	150.000000	150
unique	NaN	NaN	NaN	NaN	3
top	NaN	NaN	NaN	NaN	Iris-setosa
freq	NaN	NaN	NaN	NaN	50
mean	5.843333	3.054000	3.758667	1.198667	NaN
std	0.828066	0.433594	1.764420	0.763161	NaN
min	4.300000	2.000000	1.000000	0.100000	NaN
25%	5.100000	2.800000	1.600000	0.300000	NaN
50%	5.800000	3.000000	4.350000	1.300000	NaN
75%	6.400000	3.300000	5.100000	1.800000	NaN
max	7.900000	4.400000	6.900000	2.500000	NaN

[11]:
```
iris['sepal_length'].values
```

[11]:
```
array([5.1, 4.9, 4.7, 4.6, 5. , 5.4, 4.6, 5. , 4.4, 4.9, 5.4, 4.8, 4.8,
       4.3, 5.8, 5.7, 5.4, 5.1, 5.7, 5.1, 5.4, 5.1, 4.6, 5.1, 4.8, 5. ,
       5. , 5.2, 5.2, 4.7, 4.8, 5.4, 5.2, 5.5, 4.9, 5. , 5.5, 4.9, 4.4,
       5.1, 5. , 4.5, 4.4, 5. , 5.1, 4.8, 5.1, 4.6, 5.3, 5. , 7. , 6.4,
       6.9, 5.5, 6.5, 5.7, 6.3, 4.9, 6.6, 5.2, 5. , 5.9, 6. , 6.1, 5.6,
       6.7, 5.6, 5.8, 6.2, 5.6, 5.9, 6.1, 6.3, 6.1, 6.4, 6.6, 6.8, 6.7,
       6. , 5.7, 5.5, 5.5, 5.8, 6. , 5.4, 6. , 6.7, 6.3, 5.6, 5.5, 5.5,
       6.1, 5.8, 5. , 5.6, 5.7, 5.7, 6.2, 5.1, 5.7, 6.3, 5.8, 7.1, 6.3,
       6.5, 7.6, 4.9, 7.3, 6.7, 7.2, 6.5, 6.4, 6.8, 5.7, 5.8, 6.4, 6.5,
       7.7, 7.7, 6. , 6.9, 5.6, 7.7, 6.3, 6.7, 7.2, 6.2, 6.1, 6.4, 7.2,
       7.4, 7.9, 6.4, 6.3, 6.1, 7.7, 6.3, 6.4, 6. , 6.9, 6.7, 6.9, 5.8,
       6.8, 6.7, 6.7, 6.3, 6.5, 6.2, 5.9])
```

[12]:
```
iris.sort_values(by='sepal_length')
```

[12]:

	sepal_length	sepal_width	petal_length	petal_width	species
13	4.3	3.0	1.1	0.1	Iris-setosa
42	4.4	3.2	1.3	0.2	Iris-setosa
38	4.4	3.0	1.3	0.2	Iris-setosa
8	4.4	2.9	1.4	0.2	Iris-setosa
41	4.5	2.3	1.3	0.3	Iris-setosa
..	…	…	…	…	…
122	7.7	2.8	6.7	2.0	Iris-virginica
118	7.7	2.6	6.9	2.3	Iris-virginica
117	7.7	3.8	6.7	2.2	Iris-virginica
135	7.7	3.0	6.1	2.3	Iris-virginica
131	7.9	3.8	6.4	2.0	Iris-virginica

[150 rows x 5 columns]

[13]:
```
iris[50:55]
```

[13]:

	sepal_length	sepal_width	petal_length	petal_width	species
50	7.0	3.2	4.7	1.4	Iris-versicolor
51	6.4	3.2	4.5	1.5	Iris-versicolor
52	6.9	3.1	4.9	1.5	Iris-versicolor
53	5.5	2.3	4.0	1.3	Iris-versicolor
54	6.5	2.8	4.6	1.5	Iris-versicolor

[14]:
```
iris.sort_index(axis=1)
```

[14]:

	petal_length	petal_width	sepal_length	sepal_width	species
0	1.4	0.2	5.1	3.5	Iris-setosa
1	1.4	0.2	4.9	3.0	Iris-setosa
2	1.3	0.2	4.7	3.2	Iris-setosa
3	1.5	0.2	4.6	3.1	Iris-setosa
4	1.4	0.2	5.0	3.6	Iris-setosa
..	…	…	…	…	…
145	5.2	2.3	6.7	3.0	Iris-virginica
146	5.0	1.9	6.3	2.5	Iris-virginica
147	5.2	2.0	6.5	3.0	Iris-virginica
148	5.4	2.3	6.2	3.4	Iris-virginica
149	5.1	1.8	5.9	3.0	Iris-virginica

[150 rows x 5 columns]

[15]:
```
iris
```

[15]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa

```
3             4.6          3.1           1.5          0.2     Iris-setosa
4             5.0          3.6           1.4          0.2     Iris-setosa
..            …            …             …            …            …
145           6.7          3.0           5.2          2.3  Iris-virginica
146           6.3          2.5           5.0          1.9  Iris-virginica
147           6.5          3.0           5.2          2.0  Iris-virginica
148           6.2          3.4           5.4          2.3  Iris-virginica
149           5.9          3.0           5.1          1.8  Iris-virginica

[150 rows x 5 columns]
```

[16]:
```
iris.iloc[1:2]
```

[16]:
```
   sepal_length  sepal_width  petal_length  petal_width      species
1           4.9          3.0           1.4          0.2  Iris-setosa
```

[17]:
```
iris.iloc[15:5:-2]
```

[17]:
```
    sepal_length  sepal_width  petal_length  petal_width      species
15           5.7          4.4           1.5          0.4  Iris-setosa
13           4.3          3.0           1.1          0.1  Iris-setosa
11           4.8          3.4           1.6          0.2  Iris-setosa
9            4.9          3.1           1.5          0.1  Iris-setosa
7            5.0          3.4           1.5          0.2  Iris-setosa
```

[18]:
```
iris.loc[0:5,['sepal_length','species']]
```

[18]:
```
   sepal_length      species
0           5.1  Iris-setosa
1           4.9  Iris-setosa
2           4.7  Iris-setosa
3           4.6  Iris-setosa
4           5.0  Iris-setosa
5           5.4  Iris-setosa
```

[19]:
```
iris.iloc[3:5,0:2]
```

[19]:
```
   sepal_length  sepal_width
3           4.6          3.1
4           5.0          3.6
```

[20]:
```
iris.iloc[[1,2,5,78,105],[0,1,4]]
```

[20]:
```
    sepal_length  sepal_width          species
1            4.9          3.0      Iris-setosa
2            4.7          3.2      Iris-setosa
5            5.4          3.9      Iris-setosa
```

```
78            6.0          2.9  Iris-versicolor
105           7.6          3.0   Iris-virginica
```

[21]:
```
iris.iloc[2]
```

[21]:
```
sepal_length            4.7
sepal_width             3.2
petal_length            1.3
petal_width             0.2
species         Iris-setosa
Name: 2, dtype: object
```

[22]:
```
iris.isnull().sum()
```

[22]:
```
sepal_length    0
sepal_width     0
petal_length    0
petal_width     0
species         0
dtype: int64
```

[23]:
```
iris.isna().sum()
```

[23]:
```
sepal_length    0
sepal_width     0
petal_length    0
petal_width     0
species         0
dtype: int64
```

[24]:
```
iris.sepal_length.isnull()
```

[24]:
```
0      False
1      False
2      False
3      False
4      False
       …
145    False
146    False
147    False
148    False
149    False
Name: sepal_length, Length: 150, dtype: bool
```

2 Data Formating

[25]:
```
iris.dtypes
```

[25]:
```
sepal_length    float64
sepal_width     float64
petal_length    float64
petal_width     float64
species          object
dtype: object
```

[26]:
```
iris["sepal_length"] = iris["sepal_length"].astype("int")
```

[27]:
```
iris.dtypes
```

[27]:
```
sepal_length      int32
sepal_width     float64
petal_length    float64
petal_width     float64
species          object
dtype: object
```

[28]:
```
iris.head(5)
```

[28]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5	3.5	1.4	0.2	Iris-setosa
1	4	3.0	1.4	0.2	Iris-setosa
2	4	3.2	1.3	0.2	Iris-setosa
3	4	3.1	1.5	0.2	Iris-setosa
4	5	3.6	1.4	0.2	Iris-setosa

[29]:
```
iris["species"] = iris["species"].astype("string")
```

[30]:
```
iris.dtypes
```

[30]:
```
sepal_length      int32
sepal_width     float64
petal_length    float64
petal_width     float64
species          string
dtype: object
```

[31]:
```
iris.head(5)
```

[31]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5	3.5	1.4	0.2	Iris-setosa
1	4	3.0	1.4	0.2	Iris-setosa

```
2             4          3.2           1.3          0.2  Iris-setosa
3             4          3.1           1.5          0.2  Iris-setosa
4             5          3.6           1.4          0.2  Iris-setosa
```

3 Data Normalization

[32]:
```
from sklearn import preprocessing
```

[33]:
```
iris = pd.read_csv("IRIS.csv")
```

[34]:
```
iris.head()
```

[34]:
```
   sepal_length  sepal_width  petal_length  petal_width      species
0           5.1          3.5           1.4          0.2  Iris-setosa
1           4.9          3.0           1.4          0.2  Iris-setosa
2           4.7          3.2           1.3          0.2  Iris-setosa
3           4.6          3.1           1.5          0.2  Iris-setosa
4           5.0          3.6           1.4          0.2  Iris-setosa
```

[35]:
```
min_max_scaler = preprocessing.MinMaxScaler()
max_abs_scaler = preprocessing.MaxAbsScaler()
```

[36]:
```
x = iris.iloc[0:,:4]
y = iris.iloc[0:,:4]
```

[37]:
```
x_scaler = min_max_scaler.fit_transform(x)
y_scaler = max_abs_scaler.fit_transform(y)
```

[38]:
```
df_normalized = pd.DataFrame(x_scaler)
df_normalize = pd.DataFrame(y_scaler)
```

[39]:
```
df_normalized
```

[39]:
```
            0         1         2         3
0    0.222222  0.625000  0.067797  0.041667
1    0.166667  0.416667  0.067797  0.041667
2    0.111111  0.500000  0.050847  0.041667
3    0.083333  0.458333  0.084746  0.041667
4    0.194444  0.666667  0.067797  0.041667
..        …         …         …         …
145  0.666667  0.416667  0.711864  0.916667
146  0.555556  0.208333  0.677966  0.750000
147  0.611111  0.416667  0.711864  0.791667
148  0.527778  0.583333  0.745763  0.916667
149  0.444444  0.416667  0.694915  0.708333
```

[150 rows x 4 columns]

[40]:
```
df_normalize
```

[40]:

	0	1	2	3
0	0.645570	0.795455	0.202899	0.08
1	0.620253	0.681818	0.202899	0.08
2	0.594937	0.727273	0.188406	0.08
3	0.582278	0.704545	0.217391	0.08
4	0.632911	0.818182	0.202899	0.08
..	…	…	…	…
145	0.848101	0.681818	0.753623	0.92
146	0.797468	0.568182	0.724638	0.76
147	0.822785	0.681818	0.753623	0.80
148	0.784810	0.772727	0.782609	0.92
149	0.746835	0.681818	0.739130	0.72

[150 rows x 4 columns]

4 Handing categorical variable

[41]:
```
iris['species'].unique()
```

[41]:
```
array(['Iris-setosa', 'Iris-versicolor', 'Iris-virginica'], dtype=object)
```

[42]:
```
label_encoder = preprocessing.LabelEncoder()
```

[43]:
```
iris["species"] = label_encoder.fit_transform(iris["species"])
```

[44]:
```
iris
```

[44]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0
2	4.7	3.2	1.3	0.2	0
3	4.6	3.1	1.5	0.2	0
4	5.0	3.6	1.4	0.2	0
..	…	…	…	…	…
145	6.7	3.0	5.2	2.3	2
146	6.3	2.5	5.0	1.9	2
147	6.5	3.0	5.2	2.0	2
148	6.2	3.4	5.4	2.3	2
149	5.9	3.0	5.1	1.8	2

[150 rows x 5 columns]

[45]:
```
one_hot_iris = pd.get_dummies(iris, prefix="species", columns=["species"],␣
↪drop_first=True)
```

[46]:
```
one_hot_iris
```

[46]:

	sepal_length	sepal_width	petal_length	petal_width	species_1	species_2
0	5.1	3.5	1.4	0.2	0	0
1	4.9	3.0	1.4	0.2	0	0
2	4.7	3.2	1.3	0.2	0	0
3	4.6	3.1	1.5	0.2	0	0
4	5.0	3.6	1.4	0.2	0	0
..	…	…	…	…	…	…
145	6.7	3.0	5.2	2.3	0	1
146	6.3	2.5	5.0	1.9	0	1
147	6.5	3.0	5.2	2.0	0	1
148	6.2	3.4	5.4	2.3	0	1
149	5.9	3.0	5.1	1.8	0	1

[150 rows x 6 columns]