In [1]:
```
%matplotlib inline
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

In [3]:
```
# Some instances have missing features
# There are three types of plants: Iris-setosa, Iris-virginica, Iris-versicolor
# In this case, we can find mean value of an attribute for each type of plant
# and use it to substitute the missing values
df = pd.read_csv(r'C:\Users\309962\Desktop\IrisMissingData.csv')
```

In [4]:
```
df
```

Out[4]:

	sepal_length	sepal_width	petal_length	petal_width	class
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
5	5.4	3.9	1.7	0.4	Iris-setosa
6	4.6	NaN	1.4	0.3	Iris-setosa
7	5.0	3.4	NaN	0.2	Iris-setosa
8	4.4	2.9	1.4	0.2	Iris-setosa
9	4.9	3.1	1.5	0.1	Iris-setosa
10	5.4	3.7	1.5	0.2	Iris-setosa
11	4.8	3.4	1.6	0.2	Iris-setosa
12	4.8	3.0	1.4	NaN	Iris-setosa
13	4.3	3.0	1.1	0.1	Iris-setosa
14	5.8	4.0	1.2	0.2	Iris-setosa
15	5.7	4.4	1.5	0.4	Iris-setosa
16	5.4	3.9	1.3	0.4	Iris-setosa
17	5.1	3.5	1.4	0.3	Iris-setosa
18	5.7	3.8	1.7	0.3	Iris-setosa
19	5.1	3.8	1.5	0.3	Iris-setosa
20	5.4	3.4	1.7	0.2	Iris-setosa
21	5.1	3.7	1.5	0.4	Iris-setosa
22	4.6	3.6	1.0	0.2	Iris-setosa
23	5.1	3.3	1.7	0.5	Iris-setosa
24	4.8	3.4	1.9	0.2	Iris-setosa
25	5.0	3.0	1.6	0.2	Iris-setosa
26	5.0	3.4	1.6	0.4	Iris-setosa
27	5.2	3.5	1.5	0.2	Iris-setosa
28	5.2	3.4	1.4	0.2	Iris-setosa
29	4.7	3.2	1.6	0.2	Iris-setosa
...	...	...	...	...	...
120	6.9	3.2	5.7	2.3	Iris-virginica
121	5.6	2.8	4.9	2.0	Iris-virginica
122	7.7	2.8	6.7	2.0	Iris-virginica

	sepal_length	sepal_width	petal_length	petal_width	class
123	6.3	2.7	4.9	1.8	Iris-virginica
124	6.7	3.3	5.7	2.1	Iris-virginica
125	7.2	3.2	6.0	1.8	Iris-virginica
126	6.2	2.8	4.8	1.8	Iris-virginica
127	6.1	NaN	4.9	1.8	Iris-virginica
128	6.4	2.8	NaN	2.1	Iris-virginica
129	7.2	3.0	5.8	1.6	Iris-virginica
130	7.4	2.8	6.1	1.9	Iris-virginica
131	7.9	3.8	6.4	2.0	Iris-virginica
132	6.4	2.8	5.6	2.2	Iris-virginica
133	6.3	2.8	5.1	1.5	Iris-virginica
134	6.1	2.6	5.6	1.4	Iris-virginica
135	7.7	3.0	6.1	2.3	Iris-virginica
136	6.3	3.4	5.6	2.4	Iris-virginica
137	6.4	3.1	5.5	1.8	Iris-virginica
138	6.0	3.0	4.8	1.8	Iris-virginica
139	6.9	3.1	5.4	2.1	Iris-virginica
140	6.7	3.1	NaN	2.4	Iris-virginica
141	6.9	3.1	5.1	2.3	Iris-virginica
142	5.8	2.7	5.1	1.9	Iris-virginica
143	6.8	3.2	5.9	2.3	Iris-virginica
144	6.7	3.3	5.7	2.5	Iris-virginica
145	NaN	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 × 5 columns

In [5]:
```
# Look for any columns that have NA
df.isna().any(axis=0)
```

Out[5]:
```
sepal_length     True
sepal_width      True
petal_length     True
petal_width      True
class           False
dtype: bool
```

In [6]:
```
# Look for any rows that have NA
rows_missing_values = df.isna().any(axis=1)
```

In [7]:
```
df[rows_missing_values]
```

Out[7]:

	sepal_length	sepal_width	petal_length	petal_width	class
6	4.6	NaN	1.4	0.3	Iris-setosa
7	5.0	3.4	NaN	0.2	Iris-setosa
12	4.8	3.0	1.4	NaN	Iris-setosa
62	NaN	2.2	4.0	1.0	Iris-versicolor
64	5.6	2.9	3.6	NaN	Iris-versicolor
80	5.5	NaN	NaN	1.1	Iris-versicolor
127	6.1	NaN	4.9	1.8	Iris-virginica
128	6.4	2.8	NaN	2.1	Iris-virginica
140	6.7	3.1	NaN	2.4	Iris-virginica
145	NaN	3.0	5.2	2.3	Iris-virginica

In [8]:
```
# Find Summary Statistics for Each Class
# Impute values based on class
# https://stackoverflow.com/questions/19966018/pandas-filling-missing-values-by-m
group_class = df.groupby('class')
```

In [9]:
```
# First few rows of each group
group_class.head(2)
```

Out[9]:

	sepal_length	sepal_width	petal_length	petal_width	class
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
50	7.0	3.2	4.7	1.4	Iris-versicolor
51	6.4	3.2	4.5	1.5	Iris-versicolor
100	6.3	3.3	6.0	2.5	Iris-virginica
101	5.8	2.7	5.1	1.9	Iris-virginica

In [10]:
```
# Attribute Mean value is different for each group
group_class.mean()
```

Out[10]:

class	sepal_length	sepal_width	petal_length	petal_width
Iris-setosa	5.006000	3.418367	1.463265	0.246939
Iris-versicolor	5.934694	2.777551	4.269388	1.326531
Iris-virginica	6.585714	2.973469	5.550000	2.026000

In [11]:
```
# Compared to mean value for entire dataset
df.mean()
```

Out[11]:
```
sepal_length    5.836486
sepal_width     3.056463
petal_length    3.748630
petal_width     1.205405
dtype: float64
```

In [12]:
```
# For each group, use group level averages to fill missing values
df['sepal_length'] = group_class['sepal_length'].transform(lambda x: x.fillna(x.m
df['sepal_width'] = group_class['sepal_width'].transform(lambda x: x.fillna(x.mea
df['petal_length'] = group_class['petal_length'].transform(lambda x: x.fillna(x.m
df['petal_width'] = group_class['petal_width'].transform(lambda x: x.fillna(x.mea
```

In [13]:
```
# Let's now check the rows that had missing values
df[rows_missing_values]
```

Out[13]:

	sepal_length	sepal_width	petal_length	petal_width	class
6	4.600000	3.418367	1.400000	0.300000	Iris-setosa
7	5.000000	3.400000	1.463265	0.200000	Iris-setosa
12	4.800000	3.000000	1.400000	0.246939	Iris-setosa
62	5.934694	2.200000	4.000000	1.000000	Iris-versicolor
64	5.600000	2.900000	3.600000	1.326531	Iris-versicolor
80	5.500000	2.777551	4.269388	1.100000	Iris-versicolor
127	6.100000	2.973469	4.900000	1.800000	Iris-virginica
128	6.400000	2.800000	5.550000	2.100000	Iris-virginica
140	6.700000	3.100000	5.550000	2.400000	Iris-virginica
145	6.585714	3.000000	5.200000	2.300000	Iris-virginica

In [ ]:
```
```