data analysis week 7 assignment

May 12, 2025

[6]:
```
from sklearn.datasets import load_iris
import pandas as pd

iris = load_iris()
df = pd.DataFrame(iris.data, columns=iris.feature_names)
df['species'] = iris.target_names[iris.target]
print("first 5 rows")
display(df.head())

print("\nData types:")
display(df.dtypes)

print("\nMissing values:")
display(df.isnull().sum())
```

```
   sepal length (cm)  sepal width (cm)  petal length (cm)  petal width (cm)  \
0                5.1               3.5                1.4               0.2
1                4.9               3.0                1.4               0.2
2                4.7               3.2                1.3               0.2
3                4.6               3.1                1.5               0.2
4                5.0               3.6                1.4               0.2

  species
0  setosa
1  setosa
2  setosa
3  setosa
4  setosa


Data types:

sepal length (cm)    float64
sepal width (cm)     float64
petal length (cm)    float64
petal width (cm)     float64
species               object
dtype: object
```

```
Missing values:

sepal length (cm)    0
sepal width (cm)     0
petal length (cm)    0
petal width (cm)     0
species              0
dtype: int64
```

[14]:
```
# 1. Compute basic statistics for numerical columns
print("Basic Statistics for Numerical Columns:")
display(df.describe())

# 2. Group by species and calculate mean values
print("\nMean Values by Species:")
species_means = df.groupby('species').mean()
display(species_means)

# 3.group by species and calculate median values
print("\nMedian Values by Species:")
species_median = df.groupby('species').median()
display(species_median)

# 4. Group by species and calculate std values
print("\nstd Values by Species:")
species_std = df.groupby('species').std()
display(species_std)


#5. Compare petal width across species
print("\nPetal Width Comparison:")
petal_stats = df.groupby('species')['petal width (cm)'].agg(['mean', 'median',␣
 ↪'std'])
display(petal_stats)

# 6. Interesting findings
print("\nKey Findings:")
print("- Setosa has the smallest petals (mean width: {:.2f} cm)".format(
    species_means.loc['setosa', 'petal width (cm)']))
print("- Virginica has the largest sepals (mean length: {:.2f} cm)".format(
    species_means.loc['virginica', 'sepal length (cm)']))
print("- Versicolor's petal dimensions are intermediate between setosa and␣
 ↪virginica")
```

```
Basic Statistics for Numerical Columns:

       sepal length (cm)  sepal width (cm)  petal length (cm)  \
count         150.000000        150.000000         150.000000
```

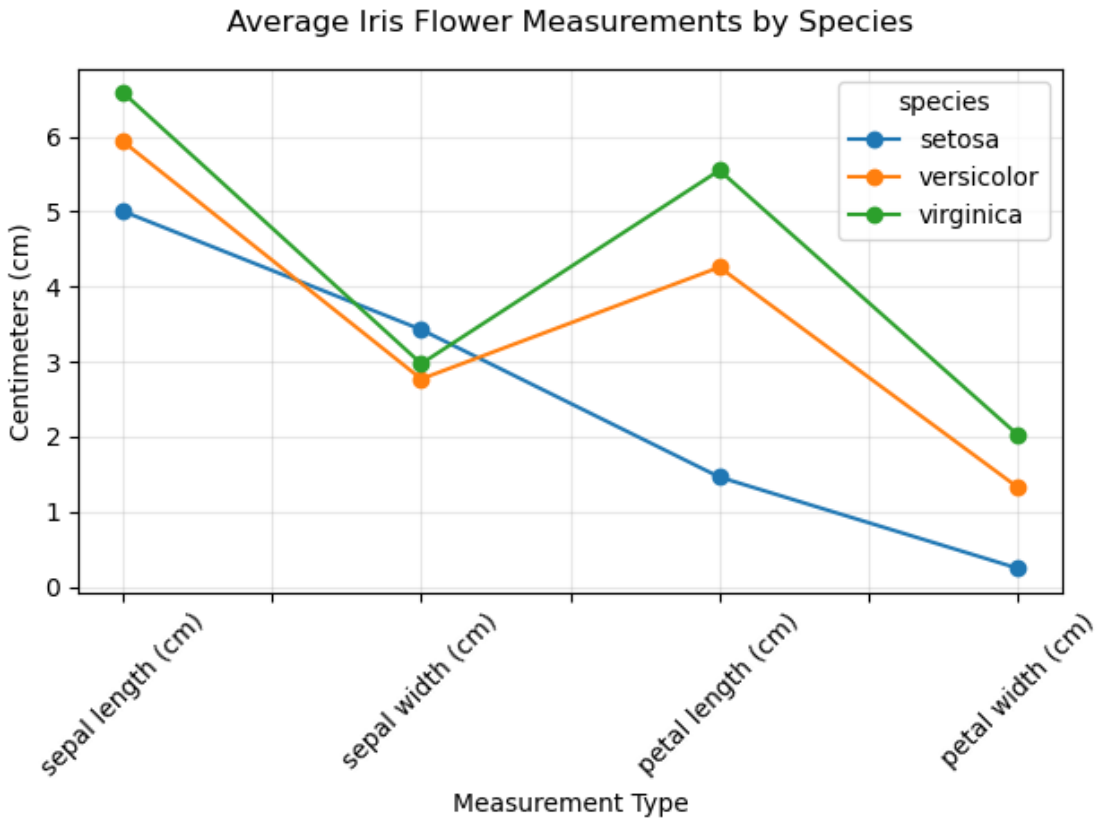
```
mean            5.843333          3.057333           3.758000
std             0.828066          0.435866           1.765298
min             4.300000          2.000000           1.000000
25%             5.100000          2.800000           1.600000
50%             5.800000          3.000000           4.350000
75%             6.400000          3.300000           5.100000
max             7.900000          4.400000           6.900000

       petal width (cm)
count        150.000000
mean           1.199333
std            0.762238
min            0.100000
25%            0.300000
50%            1.300000
75%            1.800000
max            2.500000


Mean Values by Species:

            sepal length (cm)  sepal width (cm)  petal length (cm)  \
species
setosa                  5.006             3.428              1.462
versicolor              5.936             2.770              4.260
virginica               6.588             2.974              5.552

            petal width (cm)
species
setosa                 0.246
versicolor             1.326
virginica              2.026


Median Values by Species:

            sepal length (cm)  sepal width (cm)  petal length (cm)  \
species
setosa                    5.0               3.4               1.50
versicolor                5.9               2.8               4.35
virginica                 6.5               3.0               5.55

            petal width (cm)
species
setosa                   0.2
versicolor               1.3
virginica                2.0


std Values by Species:
```

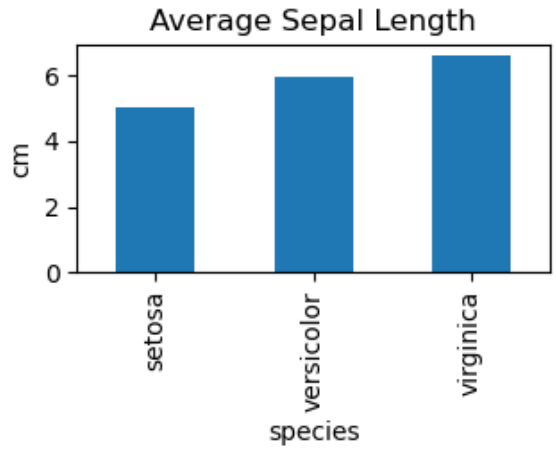
```
            sepal length (cm)  sepal width (cm)  petal length (cm)  \
species
setosa                0.352490          0.379064           0.173664
versicolor            0.516171          0.313798           0.469911
virginica             0.635880          0.322497           0.551895

            petal width (cm)
species
setosa              0.105386
versicolor          0.197753
virginica           0.274650


Petal Width Comparison:

             mean  median       std
species
setosa      0.246     0.2  0.105386
versicolor  1.326     1.3  0.197753
virginica   2.026     2.0  0.274650


Key Findings:
- Setosa has the smallest petals (mean width: 0.25 cm)
- Virginica has the largest sepals (mean length: 6.59 cm)
- Versicolor's petal dimensions are intermediate between setosa and virginica
```

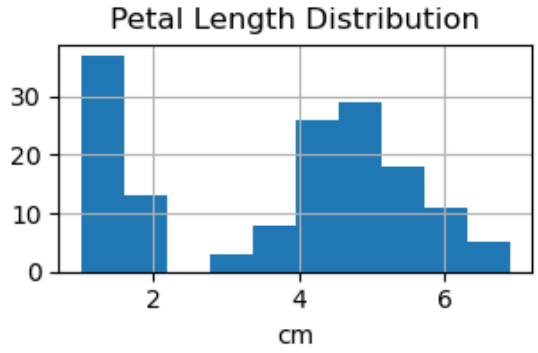
[21]:
```
 #1.line chart
plt.figure(figsize=(8, 5))
species_means.T.plot(marker='o')

plt.title('Average Iris Flower Measurements by Species', pad=15)
plt.xlabel('Measurement Type')
plt.ylabel('Centimeters (cm)')
plt.xticks(rotation=45)
plt.grid(True, alpha=0.3)

plt.tight_layout()
plt.show()
# 2. Bar Chart
plt.subplot(2, 2, 2)
df.groupby('species')['sepal length (cm)'].mean().plot(kind='bar')
plt.title('Average Sepal Length')
plt.ylabel('cm')

# 3. Histogram
plt.subplot(2, 2, 3)
df['petal length (cm)'].hist()
plt.title('Petal Length Distribution')
```

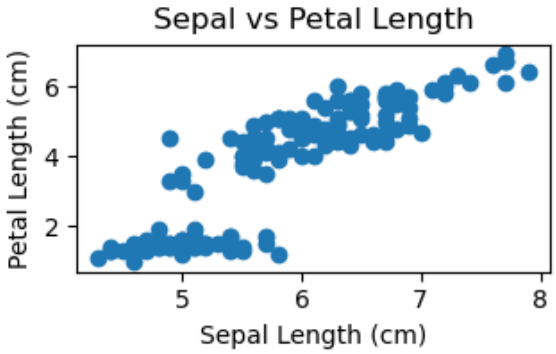
```
plt.xlabel('cm')

# 4. Scatter Plot
plt.subplot(2, 2, 4)
plt.scatter(df['sepal length (cm)'], df['petal length (cm)'])
plt.title('Sepal vs Petal Length')
plt.xlabel('Sepal Length (cm)')
plt.ylabel('Petal Length (cm)')

# Adjust layout and display
plt.tight_layout()
plt.show()
```

```
<Figure size 800x500 with 0 Axes>
```




Average Sepal Length
cm
0
2
4
6
setosa
versicolor
virginica
species


Petal Length Distribution
0
10
20
30
2
4
6
cm


Sepal vs Petal Length
Petal Length (cm)
2
4
6
5
6
7
8
Sepal Length (cm)

[ ]:

[ ]: