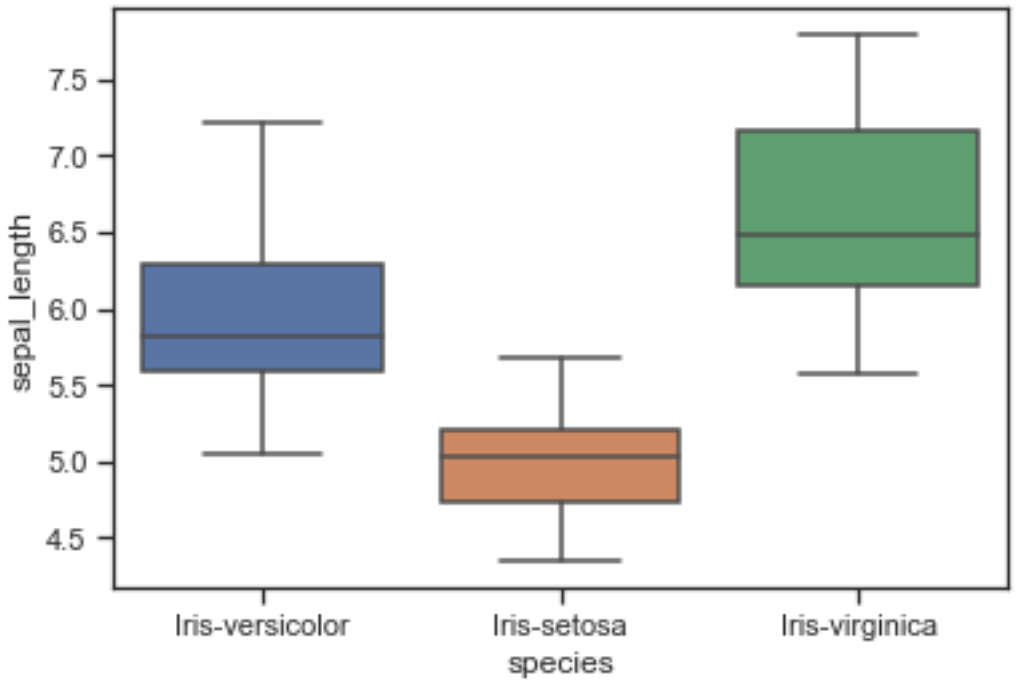
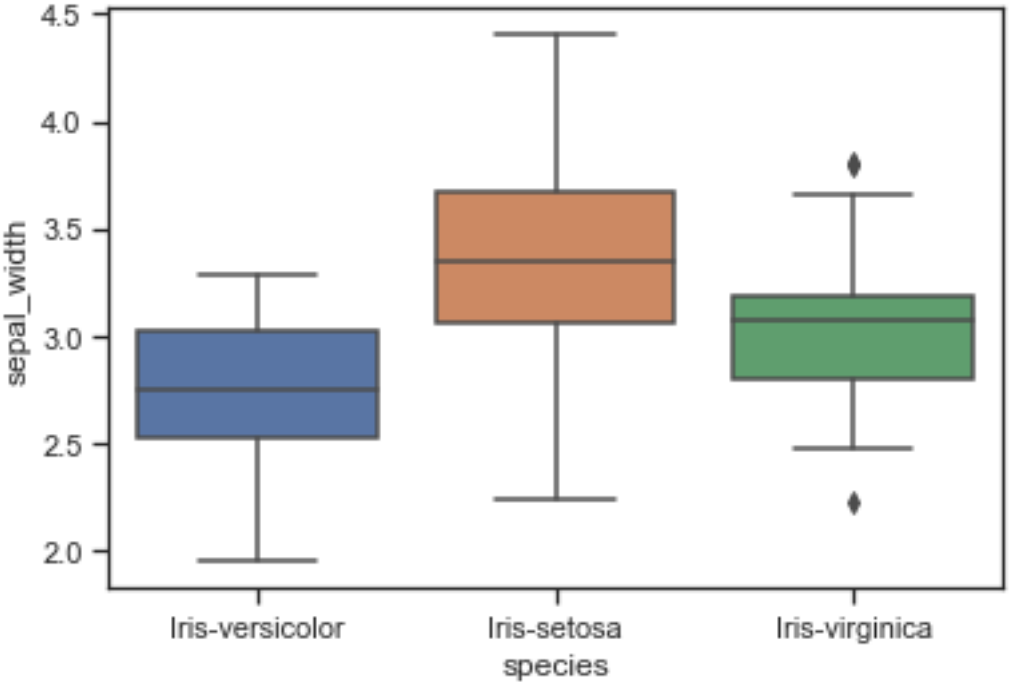
[CM3]

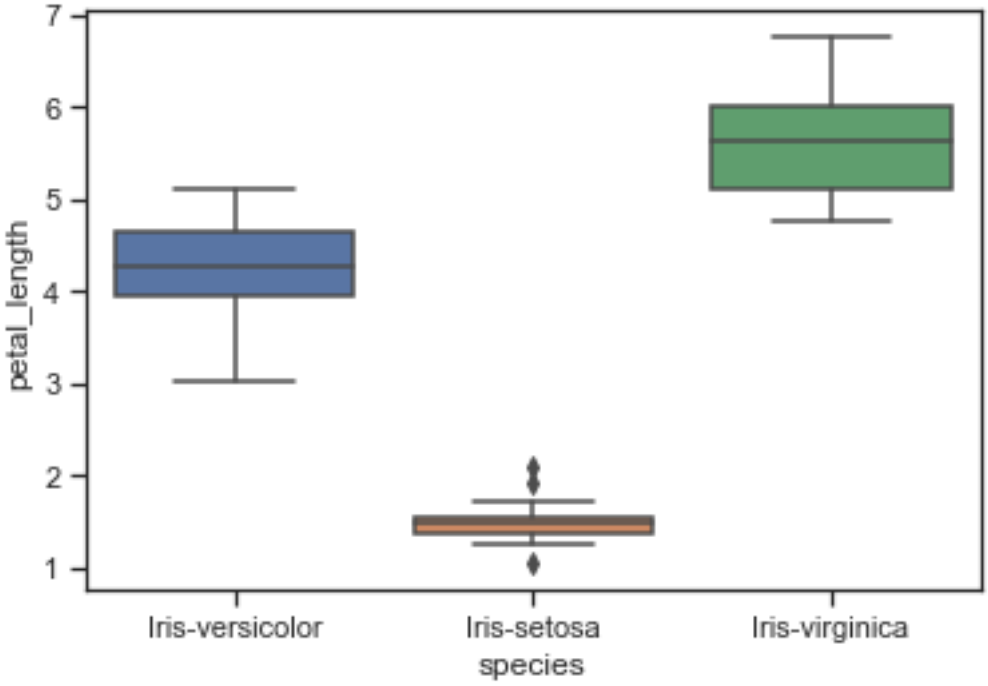
Checking for notable outliers using "Box Plots"

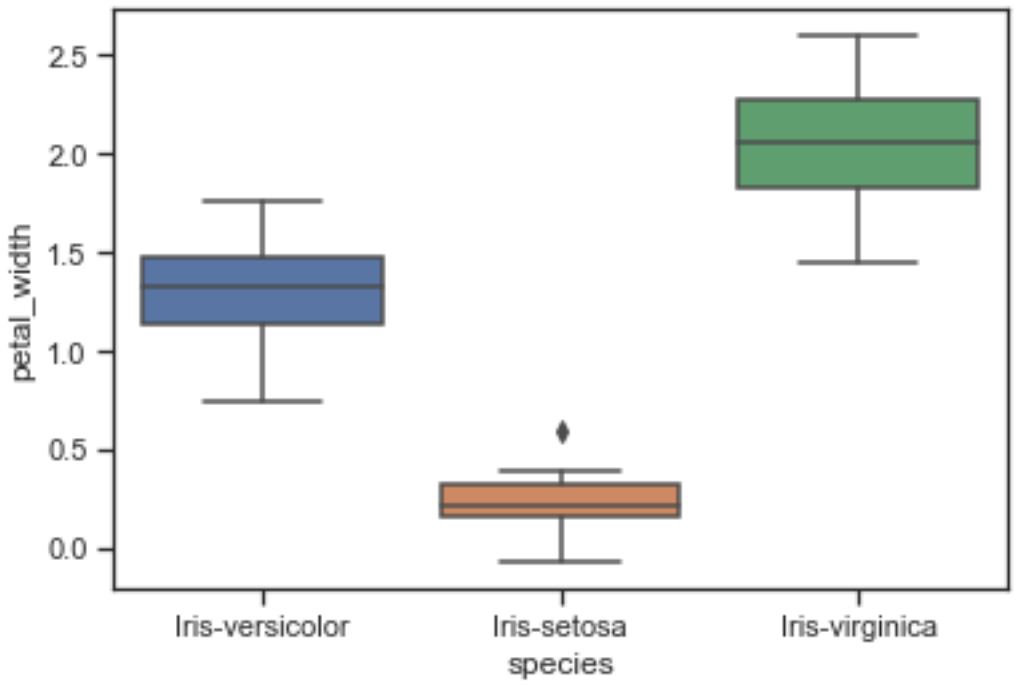
Through box plots, we find the minimum, lower quartile (25th percentile), median (50th percentile), upper quartile (75th percentile), and a maximum of an continuos variable.

[12]:
```
for column in df_iris.columns[:-1]:
    plt.figure()
    ax = sns.boxplot(x='species', y=column, data=df_iris)
    plt.show()
```




4.5
4.0
3.5
3.0
2.5
2.0
sepal_width
Iris-versicolor
Iris-setosa
Iris-virginica
species


7
6
5
4
3
2
1
petal_length
Iris-versicolor
Iris-setosa
Iris-virginica
species



From the "Box Plot" visualization. We observe that:

- there are couple of outliers in sepal width of Iris-virginica.
- there are few outliers in petal length of Iris-setosa.
- there is one outlier in petal width of Iris-setosa.

Checking for outliers using IQR

[13]:
```
# finding outliers using Inter Quartile Range (IQR)
for column in df_iris.columns[0:-1]:
    for specie in df_iris['species'].unique():
        q25 = df_iris[column][df_iris['species'] == specie].quantile(0.25)
        q75 = df_iris[column][df_iris['species'] == specie].quantile(0.75)
        iqr = q75 - q25
        print(specie.upper(), '-', column.upper())
        print('Percentiles: 25th = %.3f, 75th = %.3f, IQR = %.3f' % (q25, q75, iqr))

        # Calculate the outlier cutoff
        cut_off = iqr * 1.5
        lower, upper = q25 - cut_off, q75 + cut_off

        # Identify outliers
        df_iris2 = pd.DataFrame(df_iris[df_iris['species'] == specie][column])

        count = len(df_iris2[df_iris2[column] < lower].index)
        count += len(df_iris2[df_iris2[column] > upper].index)
        print('Identified outliers: ', count)

        # replacing outliers with NaN (Will be later replaced with feature mean)
```

```
        for index in df_iris2[df_iris2[column] < lower].index:
            df_iris.loc[index, column] = np.nan
        for index in df_iris2[df_iris2[column] > upper].index:
            df_iris.loc[index, column] = np.nan
```

```
IRIS-VERSICOLOR - SEPAL_LENGTH
Percentiles: 25th = 5.594, 75th = 6.296, IQR = 0.701
Identified outliers:  0
IRIS-SETOSA - SEPAL_LENGTH
Percentiles: 25th = 4.742, 75th = 5.213, IQR = 0.471
Identified outliers:  0
IRIS-VIRGINICA - SEPAL_LENGTH
Percentiles: 25th = 6.156, 75th = 7.166, IQR = 1.010
Identified outliers:  0
IRIS-VERSICOLOR - SEPAL_WIDTH
Percentiles: 25th = 2.527, 75th = 3.025, IQR = 0.498
Identified outliers:  0
IRIS-SETOSA - SEPAL_WIDTH
Percentiles: 25th = 3.059, 75th = 3.668, IQR = 0.608
Identified outliers:  0
IRIS-VIRGINICA - SEPAL_WIDTH
Percentiles: 25th = 2.803, 75th = 3.182, IQR = 0.379
Identified outliers:  3
IRIS-VERSICOLOR - PETAL_LENGTH
Percentiles: 25th = 3.934, 75th = 4.640, IQR = 0.706
Identified outliers:  0
IRIS-SETOSA - PETAL_LENGTH
Percentiles: 25th = 1.364, 75th = 1.542, IQR = 0.179
Identified outliers:  3
IRIS-VIRGINICA - PETAL_LENGTH
Percentiles: 25th = 5.094, 75th = 6.010, IQR = 0.917
Identified outliers:  0
IRIS-VERSICOLOR - PETAL_WIDTH
Percentiles: 25th = 1.141, 75th = 1.482, IQR = 0.341
Identified outliers:  0
IRIS-SETOSA - PETAL_WIDTH
Percentiles: 25th = 0.161, 75th = 0.324, IQR = 0.163
Identified outliers:  1
IRIS-VIRGINICA - PETAL_WIDTH
Percentiles: 25th = 1.828, 75th = 2.281, IQR = 0.454
Identified outliers:  0
```

As observed in the “Box Plot”, we can see outliers in sepal width, petal length and petal width. The outliers are handled by replacing with feature mean.