CM3

January 29, 2021

1 Notable Outliers in dataset

1.1 Required Libraries

[1]:
```
import pandas as pd
import seaborn as sns
```

1.2 Heart Dataset

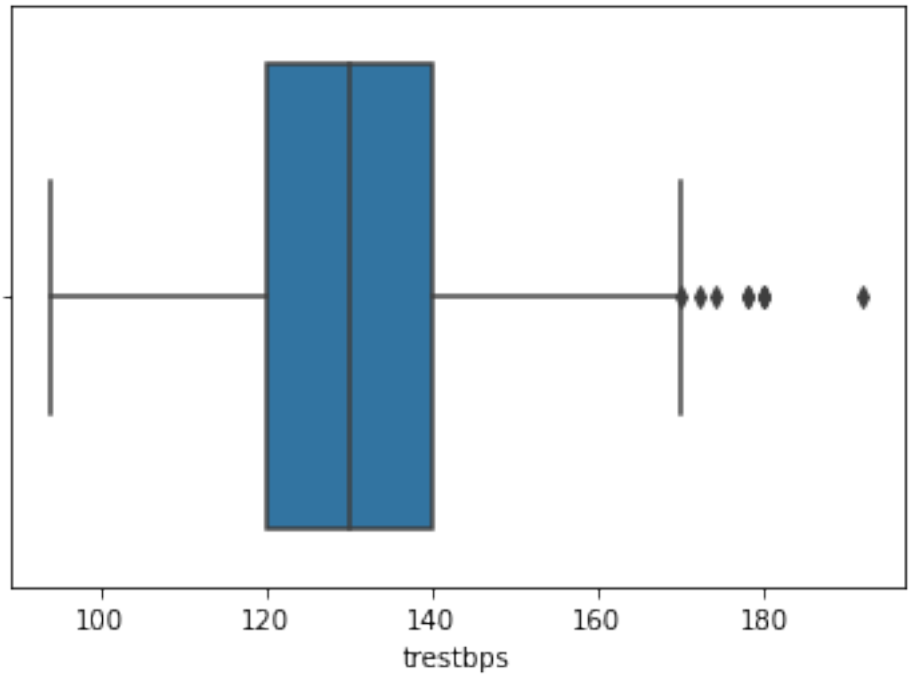
[2]:
```
df_heart = pd.read_csv("heart_disease_missing.csv")
```

1.2.1 Plotting boxplots

Only plotted features, which are having outliers.

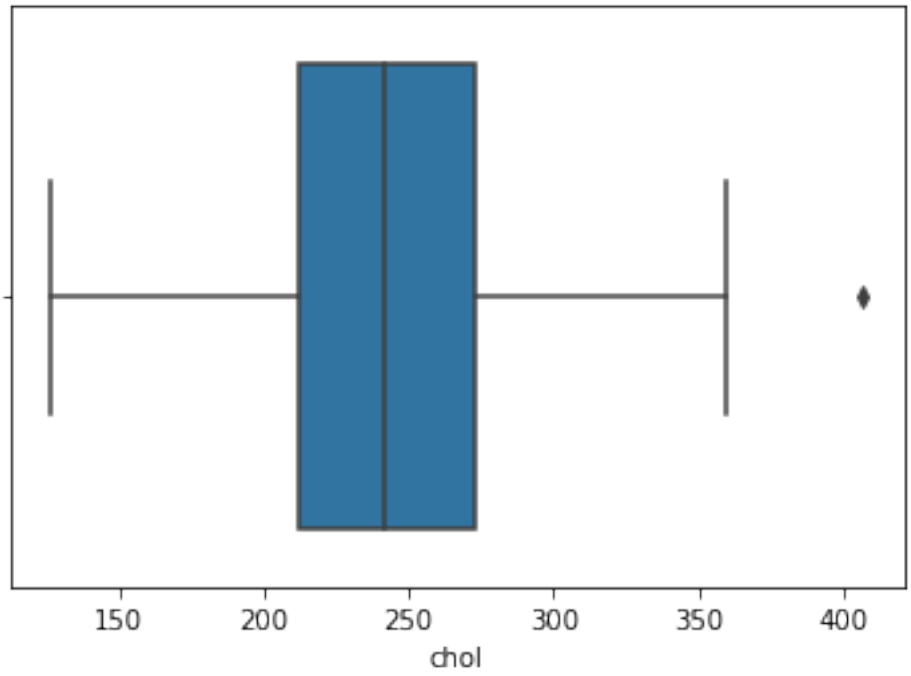
[3]:
```
sns.boxplot(x = df_heart['trestbps'])
```

[3]:
```
<AxesSubplot:xlabel='trestbps'>
```



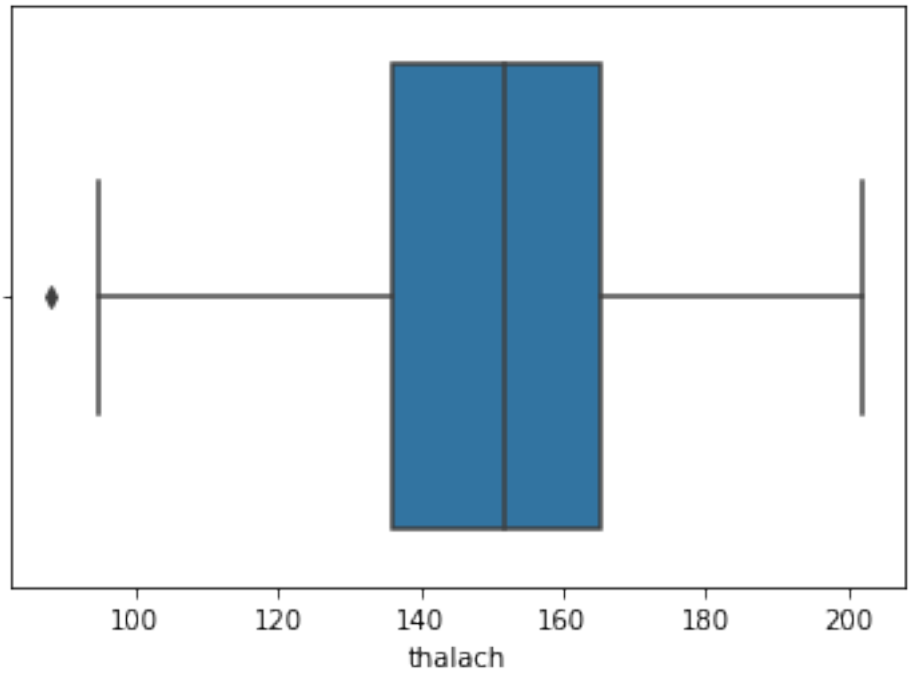
[4]:
```
sns.boxplot(x = df_heart['chol'])
```

[4]: <AxesSubplot:xlabel='chol'>



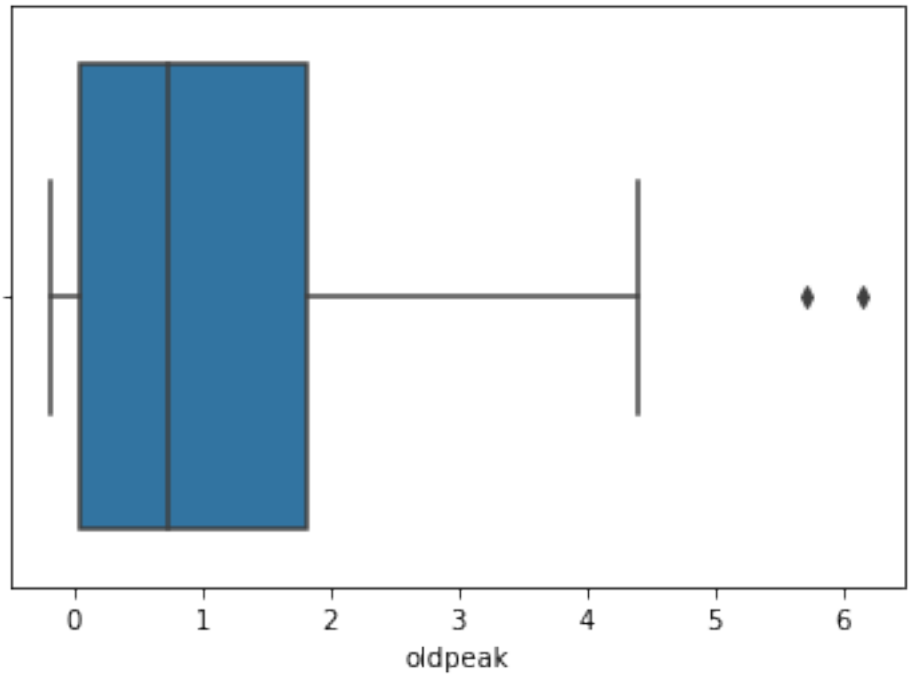
[5]:
```
sns.boxplot(x = df_heart['thalach'])
```

[5]: <AxesSubplot:xlabel='thalach'>



[6]:
```
sns.boxplot(x = df_heart['oldpeak'])
```

[6]: <AxesSubplot:xlabel='oldpeak'>



As we know, outliers increase variability in data, which is not good for statistical significance. So, it is always advisable to replace outliers, either with one of the smoothning method or nearest data point. For Heart dataset, as We can see, “trestbps, chol, thalach, and oldpeak” are having outliers. From which ‘chol’ and ‘oldpeak’ seems to be having notable affect on data spread. So those should be replaced.

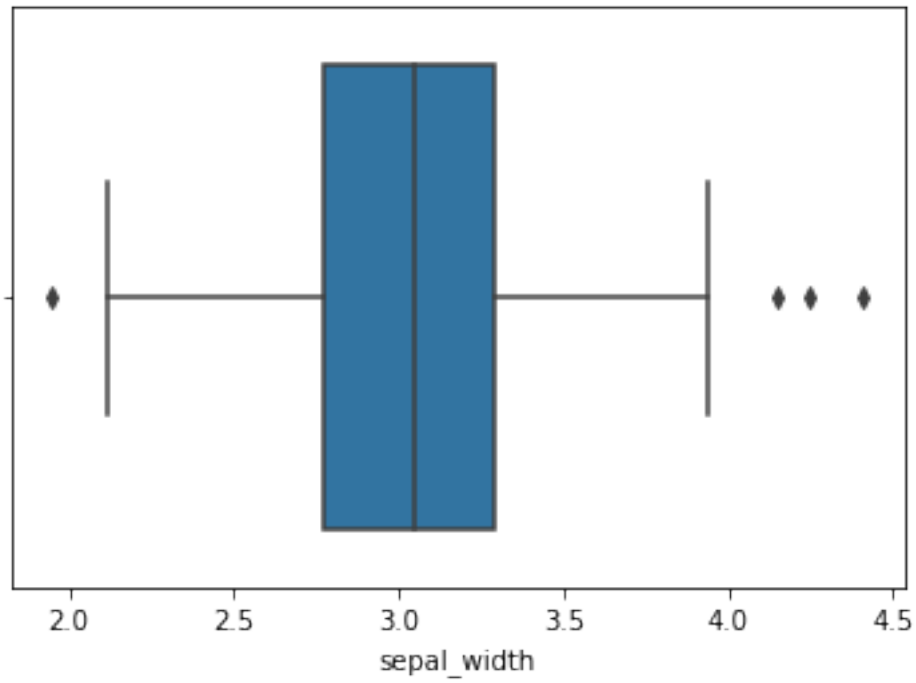
1.3 Iris Dataset

[7]:
```
df_iris = pd.read_csv("iris_dataset_missing.csv")
```

1.3.1 Plotting boxplot

[8]:
```
sns.boxplot(x = df_iris['sepal_width'])
```

[8]:
```
<AxesSubplot:xlabel='sepal_width'>
```



For Iris dataset, Only sepal_width is having outliers, which are not very far, but outside of range. So, replacing these, will make our dataset free of outliers.

1.4 References

https://seaborn.pydata.org/generated/seaborn.boxplot.html
https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.read_csv.html