Ironhack Student Portal

preview.my.ironhack.com/lms/courses/course-

Summarizing the Data

In order to look at multiple descriptive statistics at once, we can use the `describe` function. This function will show us the count, mean, standard deviation, minimum, 25th percentile, median (50th percentile), 75th percentile, and the maximum. We will only get this summary for the numeric (integer or float) columns.

```
animals.describe()
          brainwt       bodywt
count    62.000000    62.000000
mean    198.794290   283.135355
std     899.182313   930.278876
min       0.005000     0.140000
25%       0.600000     4.250000
50%       3.342500    17.250000
75%      48.201250   165.998250
max    6654.180000  5711.860000
```

We can also look at each statistic separately but still for all columns.

Maximum:

```
animals.max()
brainwt               6654.18
bodywt                5711.86
animal     Yellow-bellied_marmot
dtype: object
```

Minimum:

```
animals.min()
brainwt             0.005
bodywt               0.14
animal     African_elephant
dtype: object
```

Mean:

```
animals.mean()
brainwt    198.794290
bodywt     283.135355
dtype: float64
```

Standard Deviation:

```
animals.std()
brainwt    899.182313
bodywt     930.278876
dtype: float64
```

Median:

```
animals.median()
brainwt     3.3425
bodywt     17.2500
dtype: float64
```

25th Percentile:

```
animals.quantile(0.25)
brainwt    0.60
bodywt     4.25
Name: 0.25, dtype: float64
```

75th Percentile:

```
animals.quantile(0.75)
brainwt     48.20125
bodywt     165.99825
Name: 0.75, dtype: float64
```

We can look at the summary statistics for each column separately. We do this by subsetting the column. For example:

```
animals['bodywt'].mean()
283.13535483870976
```

What does this summary tell us about the data?

The two main measures of central tendency are the mean and the median (50th percentile). If they are close in value, it means that the data is symmetrically distributed around the mean. If the mean is greater than the median, our data is right skewed. If the median is greater than the mean, the data is left skewed. In our case, both columns have a very large mean compared to the median. This means that there are a few outliers that influence the mean and cause the data to be skewed.

Box Plot

Box plots are a visualization of a distribution of a dataset based on a 5-number summary: minimum, 25th percentile, mean, median, 75th percentile and maximum.

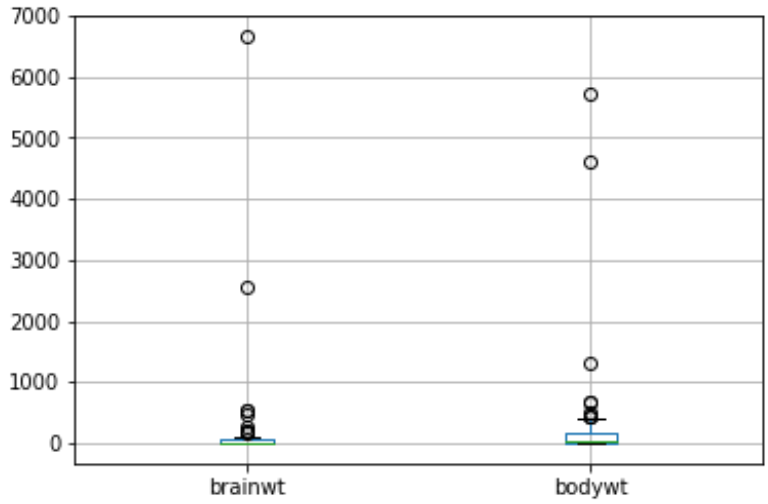
The data between the 25th and 75th percentiles is drawn inside the box. We draw a boundary outside of the box called the whiskers. Box plots also indicate how extreme outliers are by plotting them as individual points.

This plot gives us a visual summary of the data and shows us whether the data is symmetric or skewed.

We can use the matplotlib library to plot visualizations using Pandas. Additionally, if we are using Jupyter notebooks, we must indicate to plot the graph in the notebook using the `%matplotlib inline` command.

```
import matplotlib
%matplotlib inline

animals.boxplot()
```

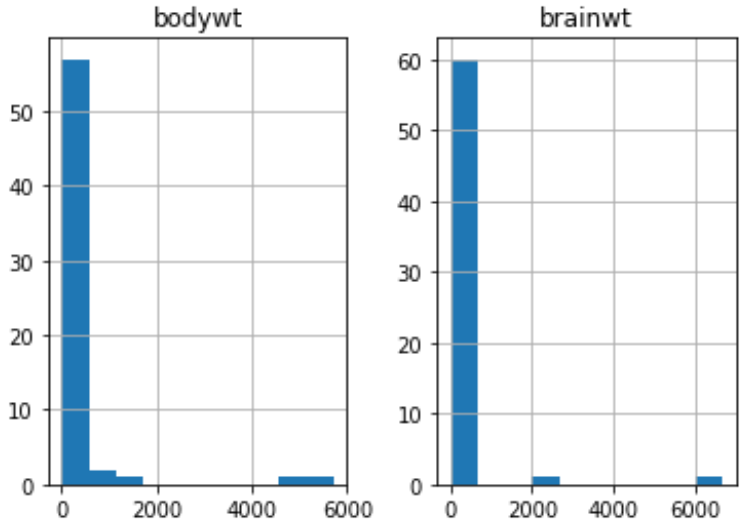


Our box plot confirms our initial suspicion that the data is skewed. We have small boxes for both brain weight and body weight and lots of outliers outside of the whiskers.

Histogram

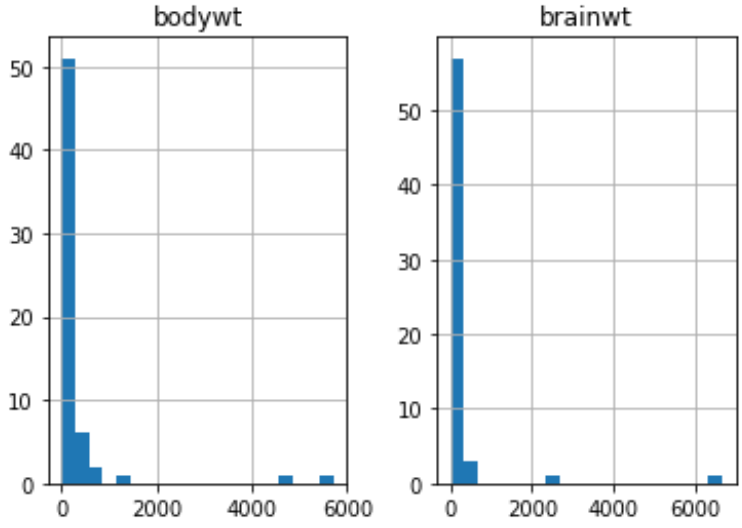
Histograms are a great way to look at the frequency distribution of our dataset. In our initial descriptive statistics lesson, we chose the bin size for our histogram and generated the bins ourselves. The default bin size in Pandas is 10. We can also manually determine the bin size by passing the number of bins to the function.

```
animals.hist()
```



Here is an example of a histogram with 20 bins:

```
animals.hist(bins=20)
```



These two histograms tell us that the bulk of the data is on the left end of the scale and we have a number of large outliers.