```
6092.596170    54
6713.133849    23
Name: Age, dtype: int64
```

Statistical Summaries

Descriptive statistics is an essential component of the data science pipeline. By investigating the properties of the dataset, we can gain a better understanding of the data and the relationship between the variables. This information is useful in making decisions about the type of data transformations to carry out or the types of learning algorithms to spot check. Let's see some examples of simple statistical functions in Pandas.

First, we'll create a Pandas dataframe.

```
my_DF = pd.DataFrame(np.random.randint(10,80,[7,4]),\
            columns=['First','Second','Third', 'Fourth'])
'Output':
   First  Second  Third  Fourth
0     47      32     66      52
1     37      66     16      22
2     24      16     63      36
3     70      47     62      12
4     74      61     44      18
5     65      73     21      37
6     44      47     23      13
```

Use the **describe** function to obtain summary statistics of a dataset. Eight statistical measures are displayed. They are count, mean, standard deviation, minimum value, 25th percentile, 50th percentile or median, 75th percentile, and the maximum value.

```
my_DF.describe()
'Output':
           First     Second      Third     Fourth
count   7.000000   7.000000   7.000000   7.000000
mean   51.571429  48.857143  42.142857  27.142857
std    18.590832  19.978560  21.980511  14.904458
min    24.000000  16.000000  16.000000  12.000000
```

```
25%    40.500000  39.500000  22.000000  15.500000
50%    47.000000  47.000000  44.000000  22.000000
75%    67.500000  63.500000  62.500000  36.500000
max    74.000000  73.000000  66.000000  52.000000
```

Correlation

Correlation shows how much relationship exists between two variables. Parametric machine learning methods such as logistic and linear regression can take a performance hit when variables are highly correlated. The correlation values range from –1 to 1, with 0 indicating no correlation at all. –1 signifies that the variables are strongly negatively correlated, while 1 shows that the variables are strongly positively correlated. In practice, it is safe to eliminate variables that have a correlation value greater than –0.7 or 0.7. A common correlation estimate in use is the Pearson's correlation coefficient.

```
my_DF.corr(method='pearson')
'Output':
           First    Second     Third    Fourth
First    1.000000  0.587645 -0.014100 -0.317333
Second   0.587645  1.000000 -0.768495 -0.345265
Third   -0.014100 -0.768495  1.000000  0.334169
Fourth  -0.317333 -0.345265  0.334169  1.000000
```

Skewness

Another important statistical metric is the skewness of the dataset. Skewness is when a bell-shaped or normal distribution is shifted toward the right or the left. Pandas offers a convenient function called **skew()** to check the skewness of each variable. Values close to 0 are more normally distributed with less skew.

```
my_DF.skew()
'Output':
First    -0.167782
Second   -0.566914
Third    -0.084490
Fourth    0.691332
dtype: float64
```