In [1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
```

The variables for the 'penguins' dataset include species, island, bill_length_mm, bill_depth_mm, flipper_length_mm, body_mass_g, and sex.

In [3]:
```
penguins = sns.load_dataset('penguins')
penguins.info()
penguins.head()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 344 entries, 0 to 343
Data columns (total 7 columns):
species              344 non-null object
island               344 non-null object
bill_length_mm       342 non-null float64
bill_depth_mm        342 non-null float64
flipper_length_mm    342 non-null float64
body_mass_g          342 non-null float64
sex                  333 non-null object
dtypes: float64(4), object(3)
memory usage: 18.9+ KB
```

Out[3]:

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	MALE
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	FEMALE
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	FEMALE
3	Adelie	Torgersen	NaN	NaN	NaN	NaN	NaN
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0	FEMALE

There are 344 rows and 7 columns, .shape returns the number of rows and columns while len() returns only the number of rows.

In [8]:
```
penguins.shape
```

Out[8]: (344, 7)

In [9]:
```
len(penguins)
```

Out[9]: 344

According to .value_counts(), the highest number of observations for an island is 168 observations for Biscoe. Dream is close behind but Torgersen only has 52. This may be because Torgersen is smaller or less habitable for the species. This can be explored further.

In [11]:
```
penguins['island'].value_counts()
```

Out[11]:
```
Biscoe       168
Dream        124
Torgersen     52
Name: island, dtype: int64
```

*For the species Gentoo, there is a mean bill length of about 47.50 mm and a standard deviation of about 3.08, while there is a mean bill depth of about 14.98 mm and a standard deviation of about 0.98. Since the standard deviation of bill length is close to 3 and the standard deviation of bill depth is almost 1, there is less variation of bill depth across the sample. Bill depths are very similar while there are more differences in bill length.

In [17]:
```
penguins.filter(['bill_length_mm', 'bill_depth_mm', 'species']) \
.query('species == "Gentoo"') \
.agg(['mean', 'std'])
```

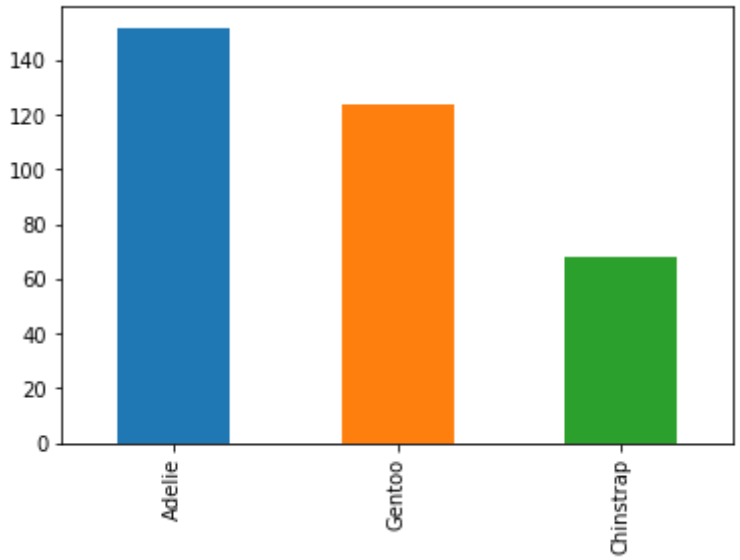
Out[17]:

	bill_length_mm	bill_depth_mm
mean	47.504878	14.982114
std	3.081857	0.981220

The barplot below compares the counts of species with each other. It indicates that Adelie has the highest number of population in the sample.

In [24]:
```
# Create a barplot by displaying counts
penguins['species'].value_counts().plot(kind = "bar")
```

Out[24]: <matplotlib.axes._subplots.AxesSubplot at 0x7f568cfd2710>



The histogram below indicated that the highest frequency of flipper lengths among all species was around the 190-195 mm range. There is a dip in the middle around 200-210 mm causing the plot to appear bimodular. This may be because flipper lengths are arranged in groups depending on the species and there aren't any species whose means land around this area.

In [23]:
```
# Create a histogram
penguins['flipper_length_mm'].plot(kind = "hist")
```

Out[23]: <matplotlib.axes._subplots.AxesSubplot at 0x7f568d0f8518>