Matplotlib Tutorial

Part 6

`Creating histogram`

Notes and codes

In [1]:
```
import pandas as pd
from matplotlib import pyplot as plt
```

Loading the data

- When loading a .txt data, you can use pandas read_csv method/function even if it's not a comma separated values. It will be rendered as a csv file.

In [4]:
```
data = pd.read_csv('data.txt')
data
```

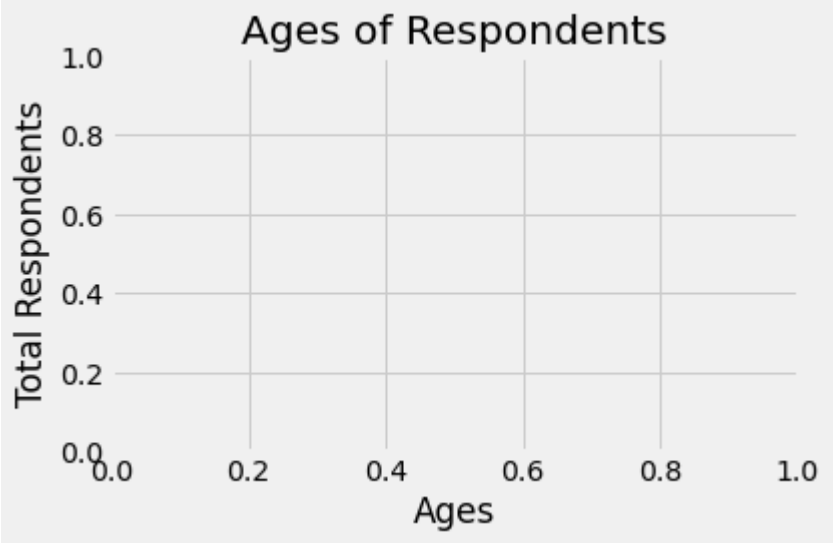
Out[4]:

	Responder_id	Age
0	1	14
1	2	19
2	3	28
3	4	22
4	5	30
...	...	...
79205	87352	59
79206	87386	21
79207	87739	25
79208	88212	40
79209	88863	18

79210 rows × 2 columns

In [6]:

```
import pandas as pd
from matplotlib import pyplot as plt

plt.style.use('fivethirtyeight')

ages = [18, 19, 21, 25, 26, 26, 30, 32, 38, 45, 55]

data = pd.read_csv('data.txt')
ids = data['Responder_id']
ages = data['Age']


# median_age = 29
# color = '#fc4f30'

# plt.legend()

plt.title('Ages of Respondents')
plt.xlabel('Ages')
plt.ylabel('Total Respondents')

plt.tight_layout()

plt.show()
```

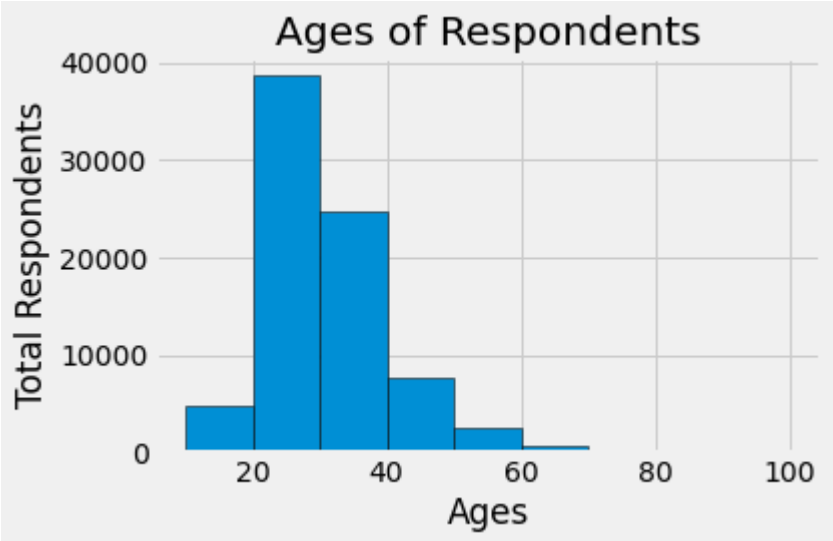


Setting up the *bins*

- **Bins** - It is a type of bar graph. To construct a `histogram`, the first step is to "bin" the *range of values* — that is, divide the entire range of values into a series of intervals — and then count how many values fall into each interval. *In this way, you don't need to plot each values separately. It will just be placed in the bins where they belong.

In [20]:

```
import pandas as pd
from matplotlib import pyplot as plt

plt.style.use('fivethirtyeight')

#data = pd.read_csv('data.txt')
#ids = data['Responder_id']
#ages = data['Age']

bins = [10, 20, 30, 40, 50, 60, 70, 80, 90, 100]

plt.hist(ages, bins=bins, edgecolor='black')

#median_age = 29
#color = '#fc4f30'

#plt.axvline(median_age, color=color, label='Age Median', linewidth=2)

#plt.legend()

plt.title('Ages of Respondents')
plt.xlabel('Ages')
plt.ylabel('Total Respondents')

plt.tight_layout()

plt.show()
```

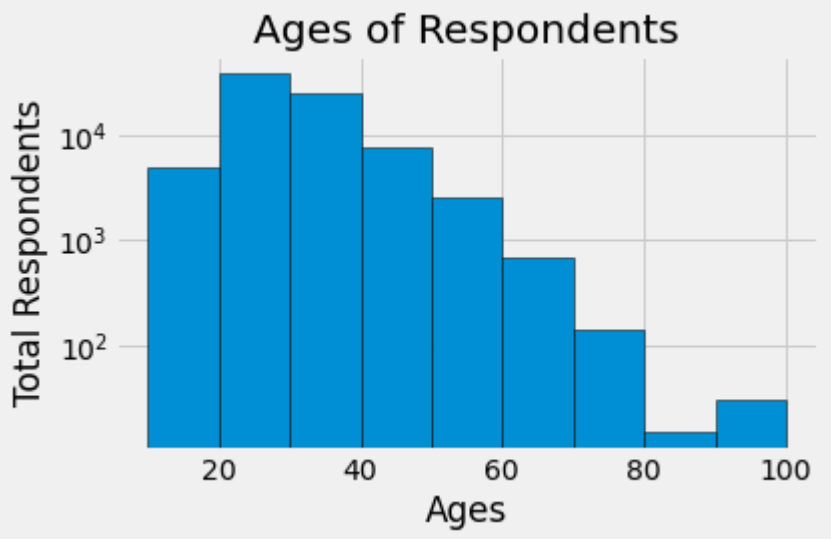


Adding or converting to logarithmic function

- You will see that there are values in the extreme right of the Table which are not shown.
- To solve this, add a `log = True` command.

In [21]:

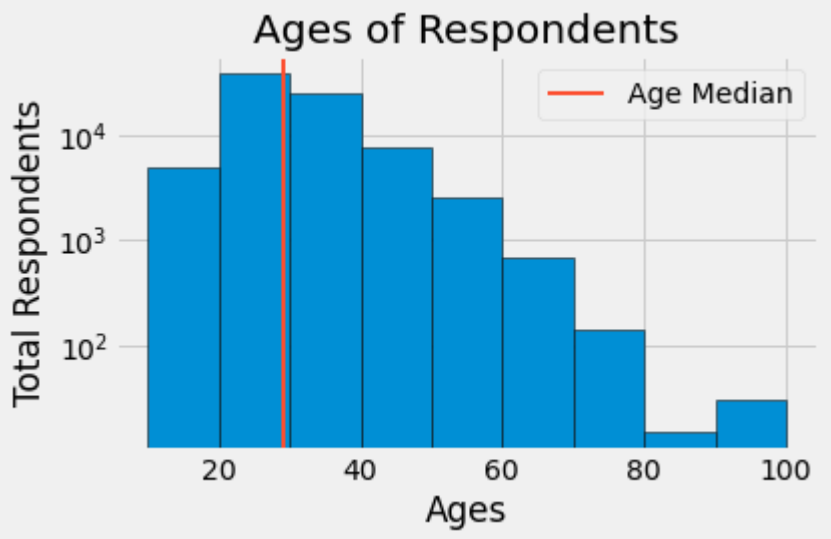
```
import pandas as pd
from matplotlib import pyplot as plt

plt.style.use('fivethirtyeight')

#data = pd.read_csv('data.txt')
#ids = data['Responder_id']
#ages = data['Age']

bins = [10, 20, 30, 40, 50, 60, 70, 80, 90, 100]

plt.hist(ages, bins=bins, edgecolor='black', log=True)

#median_age = 29
#color = '#fc4f30'

#plt.axvline(median_age, color=color, label='Age Median', linewidth=2)

#plt.legend()

plt.title('Ages of Respondents')
plt.xlabel('Ages')
plt.ylabel('Total Respondents')

plt.tight_layout()

plt.show()
```



Adding vertical line along the median

In [24]:

```
import pandas as pd
from matplotlib import pyplot as plt

plt.style.use('fivethirtyeight')

#data = pd.read_csv('data.txt')
#ids = data['Responder_id']
#ages = data['Age']

bins = [10, 20, 30, 40, 50, 60, 70, 80, 90, 100]

plt.hist(ages, bins=bins, edgecolor='black', log=True)

median_age = 29
color = '#fc4f30' #this is color red

plt.axvline(median_age, color=color, label='Age Median', linewidth=2)

plt.legend()

plt.title('Ages of Respondents')
plt.xlabel('Ages')
plt.ylabel('Total Respondents')

plt.tight_layout()

plt.show()
```



Final Codes

In [16]:

```
import pandas as pd
from matplotlib import pyplot as plt

plt.style.use('fivethirtyeight')

data = pd.read_csv('data.txt')
ids = data['Responder_id']
ages = data['Age']

bins = [10, 20, 30, 40, 50, 60, 70, 80, 90, 100]

plt.hist(ages, bins=bins, edgecolor='black', log=True)

median_age = 29
color = '#fc4f30'

plt.axvline(median_age, color=color, label='Age Median', linewidth=2)

plt.legend()

plt.title('Ages of Respondents')
plt.xlabel('Ages')
plt.ylabel('Total Respondents')

plt.tight_layout()

plt.show()
```