

DV01 - Datasets. Basic Chart Types

'1. Make some data'

--> Downloaded the data as csv.

'2. Reproducible Research'

- Are older students taller?

--> There does not seem to be a direct correlation between these two properties.

In [129]:

```
import numpy as np
import matplotlib as mpl
import pandas as pd
from matplotlib import pyplot as plt

# we need the following line to indicate that the plots should be shown inline with
%matplotlib inline

# Survey class data
df = pd.read_csv('survey.csv', sep=',', low_memory=False, encoding = 'ISO-8859-1')

df.head()
```

Out[129]:

	Timestamp	How much do you like Fontys so far?	What is your main study profile?	What aspect of Data Visualization is most interesting to you?	How much effort do you intend to invest into this subject?	Your age (in years)	Your height (in cm)	Your starsign	conti
0	9/1/2017 12:19:47	NaN	Software Engineering	(web) programming	10	21	420.69	Leo	Eu
1	9/1/2017 12:22:32	NaN	Software Engineering	exploring/analysing	8	19	178.00	Cancer	Eu
2	9/1/2017 12:23:13	NaN	Technology	designing and storytelling	7	20	182.00	Scorpio	Eu
3	9/1/2017 12:23:21	NaN	Software Engineering	(web) programming	8	21	197.00	Cancer	Eu
4	9/1/2017 12:26:08	NaN	Mathematics	designing and storytelling	8	23	183.00	Aquarius	Eu

In [96]:

```
df_filtered = df[df['Your age (in years)'] > 15]
df_filtered = df_filtered[df_filtered['Your age (in years)'] < 50]

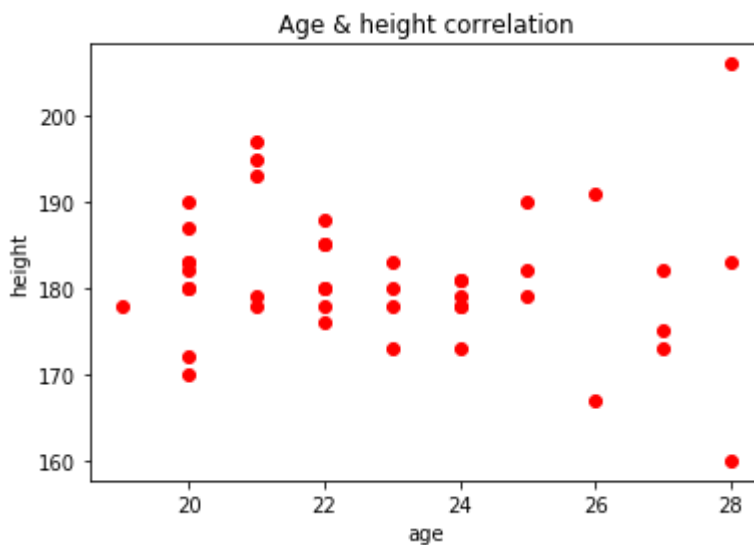
df_filtered = df_filtered[df_filtered['Your height (in cm)'] > 100]
df_filtered = df_filtered[df_filtered['Your height (in cm)'] < 300]

x_age = df_filtered['Your age (in years)']
y_height = df_filtered['Your height (in cm)']

plt.plot(x_age, y_height, 'ro')
plt.title('Age & height correlation')
plt.ylabel('height')
plt.xlabel('age')
```

Out[96]:

<matplotlib.text.Text at 0x7f75dddd9cc0>



- Make a histogram for every quantitative attribute. What are these charts saying?

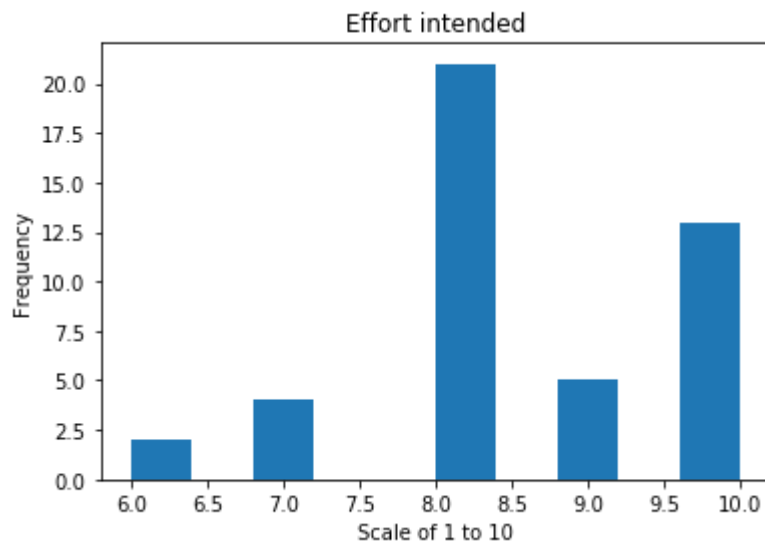
--> .

In [108]:

```
effort = df['How much effort do you intend to invest into this subject? ']  
  
plt.hist(effort)  
plt.title("Effort intended")  
plt.xlabel("Scale of 1 to 10")  
plt.ylabel("Frequency")
```

Out[108]:

<matplotlib.text.Text at 0x7f75dd3520b8>

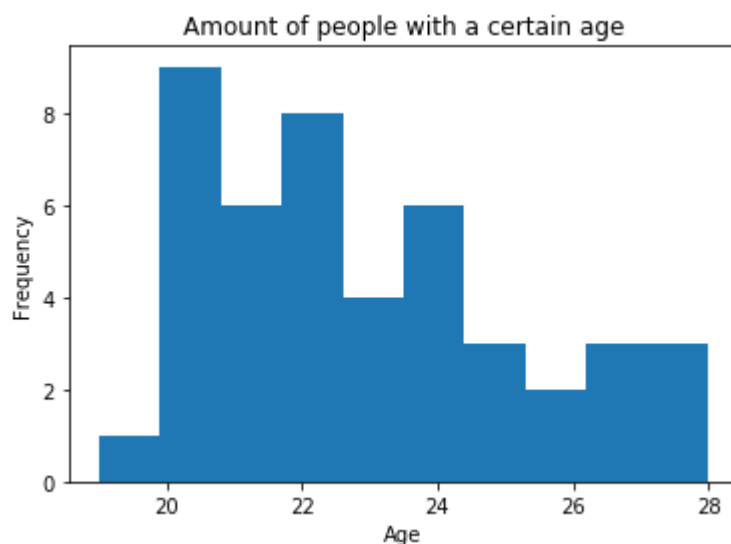


In [114]:

```
age = df['Your age (in years)']  
  
plt.hist(age)  
plt.title("Amount of people with a certain age")  
plt.xlabel("Age")  
plt.ylabel("Frequency")
```

Out[114]:

<matplotlib.text.Text at 0x7f75dd8c3940>

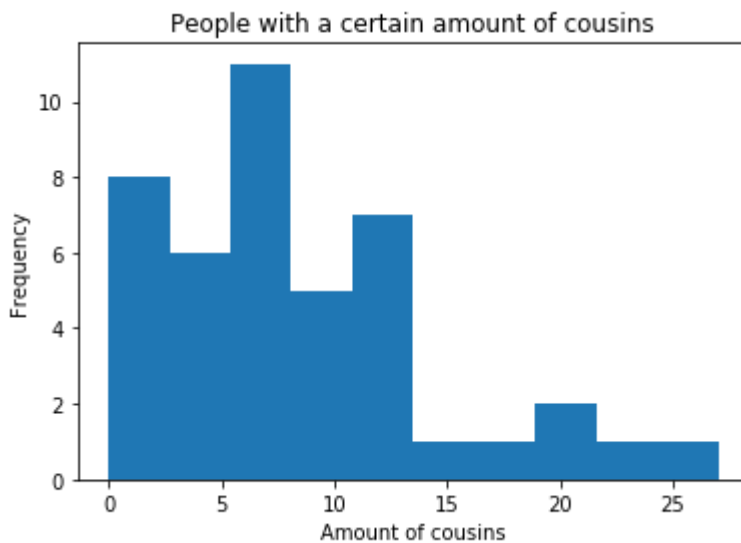


In [149]:

```
df_filtered = df[df['How many cousins do you have?'].apply(lambda x: x.isnumeric())  
df_filtered = df[df['How many cousins do you have?'].apply(lambda x: len(x) < 5)]  
df_filtered = df_filtered[df_filtered['How many cousins do you have?'].astype(int)]  
  
cousins = df_filtered['How many cousins do you have?'].astype(int)  
  
plt.hist(cousins)  
plt.title("People with a certain amount of cousins")  
plt.xlabel("Amount of cousins")  
plt.ylabel("Frequency")
```

Out[149]:

<matplotlib.text.Text at 0x7f75dca49978>



'3. Attribute types'

Examine each attribute

In [151]:

```
df.dtypes
```

Out[151]:

```
Timestamp                                obje
ct
How much do you like Fontys so far?      float
64
What is your main study profile?         obje
ct
What aspect of Data Visualization is most interesting to you?  obje
ct
How much effort do you intend to invest into this subject?     int
64
Your age (in years)                                int
64
Your height (in cm)                                float
64
Your starsign                                     obje
ct
Your continent                                   obje
ct
How many cousins do you have?                 obje
ct
Unnamed: 10                                     obje
ct
Do you like spicy food?                       obje
ct
dtype: object
```

4, 5, 6, 7, 8 - Done.

Find your dataset!

Assignment: Choose a dataset to visualize

Good dataset to visualize?

- As clean as possible
- Well-explained attributes, to avoid confusion and inaccuracy
- Many attributes of various types
- Generic attributes --> easy linking to other datasets
- Interesting topic to me, so I can ask relevant questions

In order to choose a suitable dataset for myself to work with, I started brainstorming about my interests. Generally, this came down to these subjects:

- Software
- Soccer
- Technology
- Guitar
- Blockchain
- Games (League of Legends)

The one thing that most of these have in common, is Modern Technological Development. I started playing guitar by watching youtube videos, and I can watch my favorite soccer games on TV, so in a sense for me, even those have something to do with modern technological development.

Then I started to think about a dataset to accompany this interest. The reason we're all able to benefit from these developments, has everything to do with the availability of the internet, raising my first question:

What's going on with global internet usage across the years?

After a really long search, I was unable to acquire a suitable dataset to find an answer in for this question. Some of the websites I looked at:

- <https://ourworldindata.org/internet/> (<https://ourworldindata.org/internet/>)
- <http://www.kdnuggets.com/datasets/index.html> (<http://www.kdnuggets.com/datasets/index.html>)
- https://catalog.data.gov/dataset?res_format=CSV&groups=consumer9350&groups_limit=0&page=1 (https://catalog.data.gov/dataset?res_format=CSV&groups=consumer9350&groups_limit=0&page=1)
- The most promising one: <https://catalog.data.gov/dataset/current-population-survey-internet-and-computer-use-supplement/resource/5e311174-e05a-4901-b57f-402310477268> (<https://catalog.data.gov/dataset/current-population-survey-internet-and-computer-use-supplement/resource/5e311174-e05a-4901-b57f-402310477268>) --> however, I was unable to retrieve the set due to technical issues with the website.

In the end, I did find an equally as interesting dataset which covers every Reddit post on the subreddit 'worldnews', from 25-01-2008 to 2016-11-22. I've decided to use this dataset instead.

The accompanying topic with this dataset is, obviously,

World News

Source: <https://www.kaggle.com/rootuser/worldnews-on-reddit>
(<https://www.kaggle.com/rootuser/worldnews-on-reddit>)