In [1]:
```
from pathlib import Path
import plotly.express as px
import pandas as pd
import numpy as np
from matplotlib.pyplot as plt
import seaborn as sns
from statsmodels.tsa.stattools import grangercausalitytests, adfuller
from scipy import signal
```

In [2]:
```
DATA_PREPARED_DIR = "./data/prepared/"
MERGED_DATA = "life_expectancy.csv"
STANDARDIZED_MERGED_DATA = "full_dataset_standardized.csv"
```

In [3]:
```
df = pd.read_csv(Path(DATA_PREPARED_DIR, STANDARDIZED_MERGED_DATA))
df.head()
```

Out[3]:

	year	sex	cancer [deaths]	country	cardiovascular disease [deaths]	diabetes mellitus [deaths]	chronic respiratory diseases [deaths]	diseases of digestive system [deaths]	population	life expectancy [age]	non-communicable chronic disease [deaths]
0	1996	1.0	27929.9760	Netherlands	328211472	4974282	36.279796	28.725629	15795930.0	74.783	688.442258
1	1996	2.0	28973154	Netherlands	332467051	5.548057	27.588842	37.587548	7815000.0	80.378	613.921791
2	1996	3.0	27893748	Netherlands	330.362634	5.769293	41.775860	33.205628	15335058.0	77.642	650.771063
3	1997	1.0	28456284	Netherlands	314.948694	4.392102	53.007993	27.573481	7778400.0	74.963	664.541874
4	1997	2.0	27176604	Netherlands	322.449946	7.108284	28.618868	36.529738	7855200.0	80.435	606.486884

Data description

Describe the data.

Here I compare The Netherlands and Japan. I do not include Canada, as they did not report all data to the World Health Organization. Thus, a means of comparing Canada with the others, is not there.

Only the data containing both male and females is used.

The standardized dataset contains **deaths per 100,000 people**.

In [4]:
```
df = df[df['sex'] == 3]

n1_df = df[df['country'] == 'Netherlands']
jp_df = df[df['country'] == 'Japan']
```

In [5]:
```
cols = ['cancer [deaths]', 'cardiovascular disease [deaths]',
        'diabetes mellitus [deaths]', 'chronic respiratory diseases [deaths]',
        'diseases of digestive system [deaths]',
        'non-communicable chronic disease [deaths]','life expectancy [age]']
```

In [6]:
```
n1_df[cols].describe()
```

Out[6]:

	cancer [deaths]	cardiovascular disease [deaths]	diabetes mellitus [deaths]	chronic respiratory diseases [deaths]	diseases of digestive system [deaths]	non-communicable chronic disease [deaths]	life expectancy [age]
count	23.000000	23.000000	23.000000	23.000000	23.000000	23.000000	23.000000
mean	247.214576	262.635648	5.658420	39.750268	31.242719	586.520070	79.911729
std	10.031153	37.814654	0.671944	2.295987	3.011367	33.150573	1.539977
min	234.959360	218.197000	4.744663	34.769355	25.931557	545.265789	77.642000
25%	238.189652	228.934523	5.171462	38.571394	29.018845	561.207495	78.481500
50%	242.862522	249.343367	5.677068	39.736618	32.133619	570.087493	80.001000
75%	255.887315	297.022982	5.986461	41.286157	33.209077	612.721905	81.302000
max	263.872960	330.362634	7.590592	44.400010	36.091776	650.771063	82.140000

In [7]:
```
jp_df[cols].describe()
```

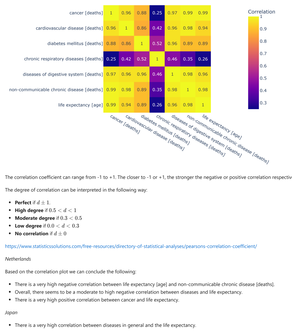
Out[7]:

	cancer [deaths]	cardiovascular disease [deaths]	diabetes mellitus [deaths]	chronic respiratory diseases [deaths]	diseases of digestive system [deaths]	non-communicable chronic disease [deaths]	life expectancy [age]
count	23.000000	23.000000	23.000000	23.000000	23.000000	23.000000	23.000000
mean	263.934819	259.060949	6.013263	15.468307	34.495504	573.580363	82.453000
std	27.463117	16.106512	0.502569	0.782565	3.845211	47.598389	1.264192
min	217.458260	227.499997	5.160246	14.600241	29.592542	501.195583	80.162000
25%	240.248552	243.446086	6.197753	14.973091	30.804036	526.745082	81.541500
50%	264.858072	261.047705	6.409248	15.245062	34.156423	576.130809	82.575000
75%	286.402945	272.129577	6.901244	15.700816	37.659980	619.313954	83.373000
max	300.547994	283.780282	7.861240	17.573716	42.009918	645.879952	84.470000

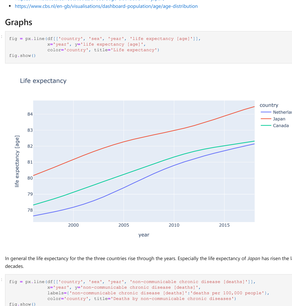
Plot Heatmap

In [8]:
```
n1_corr = n1_df[cols].corr()
jp_corr = jp_df[cols].corr()
```

In [9]:
```
px.imshow(n1_corr.values.round(decimals=2),
          labels=dict(color="Correlation"),
          x = n1_corr.index,
          y = n1_corr.index,
          text_auto=True)
```



In [10]:
```
px.imshow(jp_corr.values.round(decimals=2),
          labels=dict(color="Correlation"),
          x = jp_corr.index,
          y = jp_corr.index,
          text_auto=True)
```



The correlation coefficient can range from -1 to +1. The closer to -1 or +1, the stronger the negative or positive correlation respectively.

The degree of correlation can be interpreted in the following way:

- **Perfect** if $d \pm 1$.
- **High degree** if $0.5 < d < 1$
- **Moderate degree** if $0.3 < d < 0.5$
- **Low degree** if $0.0 < d < 0.3$
- **No correlation** if $d \pm 0$

https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/pearsons-correlation-coefficient/

Netherlands

Based on the correlation plot we can conclude the following:

- There is a very high negative correlation between the expectancy [age] and non-communicable chronic disease [deaths].
- Overall, there seems to be a moderate to high negative correlation between diseases and life expectancy.
- There is a very high positive correlation between cancer and life expectancy.

Japan

- There is a very high correlation between diseases in general and the life expectancy.

We see The Netherlands show a very high negative correlation, while Japan shows a very high positive correlation between diseases and life expectancy. This can be due to demographic factors. In japan, 25-30% of people are 65+ years old. This is in contrast with the netherlands, where only around 20% are more than 65 years old.

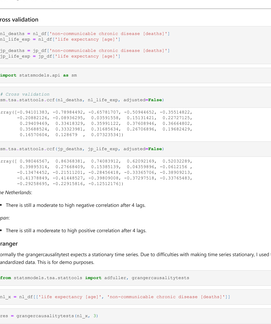
Conclusion: There seems to be a high (positive/negative) correlation between non-communicable chronic diseases and life expectancy.

Sources:

- https://www.statista.com/statistics/270087/age-distribution-in-japan/
- https://www.cbs.nl/en-gb/visualisaties/dashboard-population/age-and-gender-distribution

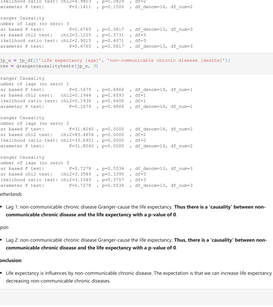
Graphs

In [11]:
```
fig = px.line(df[('country', 'sex', 'year', 'life expectancy [age]')],
              x='year', y='life expectancy [age]',
              color='country', title='life expectancy')

fig.show()
```



In general the life expectancy for the the three countries rise through the years. Especially the life expectancy of Japan has risen the last decades.

In [12]:
```
fig = px.line(df[('country', 'sex', 'year', 'non-communicable chronic disease [deaths]')],
              x='year', y='non-communicable chronic disease [deaths]',
              labels={'non-communicable chronic disease [deaths]':'deaths per 100,000 people'},
              color='country', title='Deaths by non-communicable chronic diseases')

fig.show()
```



Canada did not provide data before 2000 and after 2005. In general a decrease in diseases can be seen. However, Japan has an increase. This may be due to demographic factors as stated earlier.

The decrease in The Netherlands also seems to slow down. This indicates a change in demographic composition or the difficulties of preventing non-communicable diseases.

Cross validation

In [13]:
```
n1_death = n1_df['non-communicable chronic disease [deaths]']
n1_life_exp = n1_df['life expectancy [age]']

jp_death = jp_df['non-communicable chronic disease [deaths]']
jp_life_exp = jp_df['life expectancy [age]']
```

In [14]:
```
import statsmodels.api as sm
```

In [15]:
```
# Cross validation
sm.tsa.statsmodels.ccf(n1_death, n1_life_exp, adjusted=False)
```

Out[15]:
```
array([-0.04312803, -0.78984492, -0.65767075, -0.50364652, -0.35514822,
       -0.22197149, -0.12154154, -0.05177304, 0.00000000, 0.05177304,
        0.12154154, 0.22197149, 0.35514822, 0.50364652, 0.65767075,
        0.78984492, 0.04312803, 0.33323961, 0.31485634, 0.26708896, 0.19682426,
        0.15570056, 0.12829171, 0.07332161])
```

In [16]:
```
sm.tsa.statsmodels.ccf(jp_death, jp_life_exp, adjusted=False)
```

Out[16]:
```
array([ 0.96606567, 0.84368381, 0.74639312, 0.62931809, 0.52032289,
       0.39865016, 0.27648609, 0.15385196, 0.04319896, -0.06421156,
      -0.13476422, -0.21511201, -0.28456416, -0.33365706, -0.38995213,
      -0.41379845, -0.41488527, -0.40280606, -0.37297514, -0.33767483,
      -0.29256495, -0.22915614, -0.12121276])
```

The Netherlands:

- There is still a moderate to high negative correlation after 4 lags.

Japan:

- There is still a moderate to high positive correlation after 4 lags.

Granger

Normally the grangercausalitytest expects a stationary time series. Due to difficulties with making time series stationary, I used the standardized data. This is for demo purposes.

In [18]:
```
from statsmodels.tsa.stattools import adfuller, grangercausalitytests
```

In [20]:
```
n1_x = n1_df['life expectancy [age]', 'non-communicable chronic disease [deaths]']
```

In [21]:
```
res = grangercausalitytests(n1_x, 3)
```

```
Granger Causality
number of lags (no zero) 1
ssr based F test:         F=0.8457 , p=0.0000  , df_denom=19, df_num=1
ssr based chi2 test:   chi2=0.7687 , p=0.0000  , df=1
likelihood ratio test: chi2=26.3174 , p=0.0000  , df=1
parameter F test:         F=0.8457 , p=0.0000  , df_denom=19, df_num=1

Granger Causality
number of lags (no zero) 2
ssr based F test:         F=2.1411 , p=0.1500  , df_denom=16, df_num=2
ssr based chi2 test:   chi2=4.9933 , p=0.0824  , df=2
likelihood ratio test: chi2=8.3831 , p=0.0151  , df=2
parameter F test:         F=2.1411 , p=0.1500  , df_denom=16, df_num=2

Granger Causality
number of lags (no zero) 3
ssr based F test:         F=0.4765 , p=0.5817  , df_denom=13, df_num=3
ssr based chi2 test:   chi2=2.1225 , p=0.3731  , df=3
likelihood ratio test: chi2=2.9021 , p=0.4071  , df=3
parameter F test:         F=0.4765 , p=0.5817  , df_denom=13, df_num=3
```

In [23]:
```
jp_x = jp_df['life expectancy [age]', 'non-communicable chronic disease [deaths]']
res = grangercausalitytests(jp_x, 3)
```

```
Granger Causality
number of lags (no zero) 1
ssr based F test:         F=0.1679 , p=0.6866  , df_denom=19, df_num=1
ssr based chi2 test:   chi2=0.1844 , p=0.6703  , df=1
likelihood ratio test: chi2=0.1936 , p=0.6600  , df=1
parameter F test:         F=0.1679 , p=0.6866  , df_denom=19, df_num=1

Granger Causality
number of lags (no zero) 2
ssr based F test:         F=31.8340 , p=0.0000  , df_denom=16, df_num=2
ssr based chi2 test:   chi2=74.2793 , p=0.0000  , df=2
likelihood ratio test: chi2=39.6000 , p=0.0000  , df=2
parameter F test:         F=31.8340 , p=0.0000  , df_denom=16, df_num=2

Granger Causality
number of lags (no zero) 3
ssr based F test:         F=2.7278 , p=0.5534  , df_denom=13, df_num=3
ssr based chi2 test:   chi2=9.3589 , p=0.0249  , df=3
likelihood ratio test: chi2=9.1240 , p=0.0277  , df=3
parameter F test:         F=2.7278 , p=0.5534  , df_denom=13, df_num=3
```

Netherlands

- Lag 1: non-communicable chronic disease Granger-cause the life expectancy. **Thus there is a 'causality' between non-communicable chronic disease and the life expectancy with a p-value of 0**.

Japan

- Lag 2: non-communicable chronic disease Granger-cause the life expectancy. **Thus, there is a 'causality' between non-communicable chronic disease and the life expectancy with a p-value of 0**.

Conclusion:

- Life expectancy is influences by non-communicable chronic disease. The expectation is that we can increase life expectancy by decreasing non-communicable chronic diseases.

In [24]:
```
```