computational_narrative

May 12, 2024

1 Different Taxes and their development in Germany

Introduction The german federal statistics office (DEStatis) offers several different datasets about the federal tax income of germany. Using the monthly and quarterly tax income data (uncorrected and before allocation), we want to look at the following research questions:

- How has the overall tax income developed over the past 20 years?
- Can we see seasonal variations in the data, specifically in the consumer goods taxes?
- What role did the air traffic tax play in the overall tax income since it's introduction in 2011?
- How do the wage tax and the three assets taxes (wealth, solidarity, inheritance) compare to each other? Has this ratio changed over the past 20 years?

Preprocessing The Data Firstly, we need to process the source data and load it into the notebook. The `process_source_data` function cleans the raw data that can be downloaded from the DEStatis website and transforms it into numerical values, with proper NaN-types where values are missing.

Sometimes, we will also want to look at the tax income for whole years. Therefore, we create another yearly dataset by summing up 4 consecutive values of the quarterly dataset for each tax type.

[ ]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from helper_functions import process_source_data

BASE_DIR = '/home/nils/Documents/Repositories/semester-10/rse/
 ↪rse_24_individual_project/'
```

[ ]:
```
process_source_data('data/71211-0005_$F.csv', 'results/monthly_tax_data.csv',␣
 ↪False)
process_source_data('data/71211-0003_$F.csv', 'results/quarterly_tax_data.csv',␣
 ↪True)

df_m = pd.read_csv(BASE_DIR + 'results/monthly_tax_data.csv', index_col=0).
 ↪astype('Int64')
df_q = pd.read_csv(BASE_DIR + 'results/quarterly_tax_data.csv', index_col=0).
 ↪astype('Int64')
```

```
yearly_data = []
for i in range(0, len(df_q.columns), 4):
    yearly_data.append(df_q.iloc[:, i:i+4].sum(axis=1).values)
df_y = pd.DataFrame(np.transpose(yearly_data), index=df_q.index,␣
 ↪columns=range(1999, 2024))
```

We can confirm the number of rows and columns of all 3 datasets and also look at the number of NaN values (Use pandas' .info() method to find out about the exact column names):

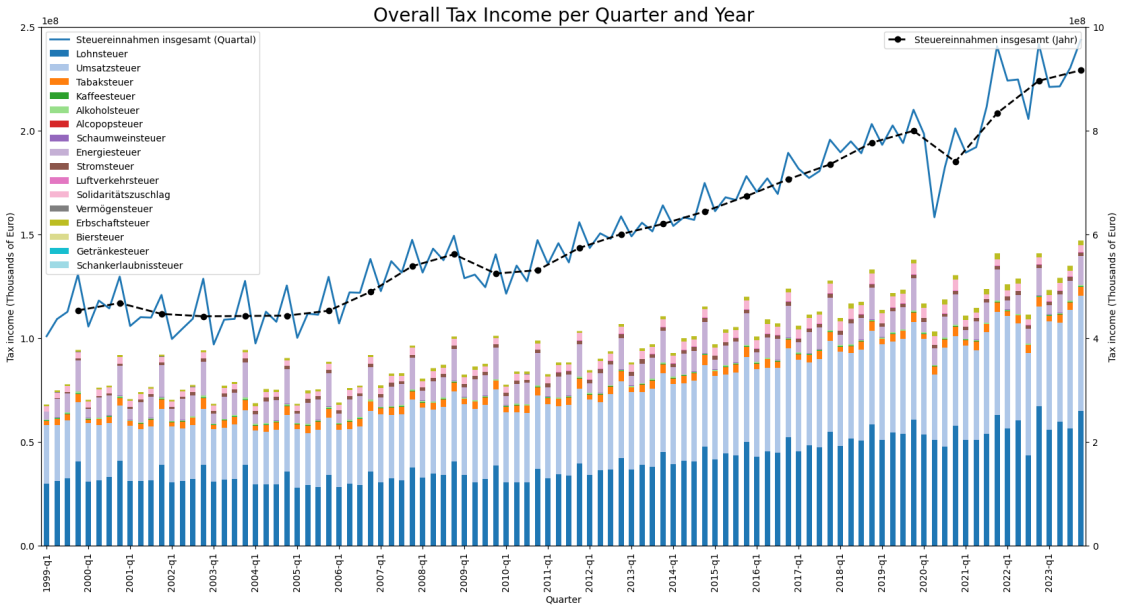
[ ]:
```
print(f'Tax types:\n{df_m.index.values}\n')
print(f'Monthly data: {df_m.shape[0]} rows, {df_m.shape[1]} columns, {df_m.
 ↪isna().sum().sum()/(df_m.shape[0]*df_m.shape[1]):0.2f}% NaN values')
print(f'Quarterly data: {df_q.shape[0]} rows, {df_q.shape[1]} columns, {df_q.
 ↪isna().sum().sum()/(df_q.shape[0]*df_q.shape[1]):0.2f}% NaN values')
print(f'Yearly data: {df_y.shape[0]} rows, {df_y.shape[1]} columns, {df_y.
 ↪isna().sum().sum()/(df_y.shape[0]*df_y.shape[1]):0.2f}% NaN values')
```

```
Tax types:
['Steuereinnahmen insgesamt' 'Lohnsteuer' 'Umsatzsteuer' 'Tabaksteuer'
 'Kaffeesteuer' 'Alkoholsteuer' 'Alcopopsteuer' 'Schaumweinsteuer'
 'Energiesteuer' 'Stromsteuer' 'Luftverkehrsteuer' 'Solidaritätszuschlag'
 'Vermögensteuer' 'Erbschaftsteuer' 'Biersteuer' 'Getränkesteuer'
 'Schankerlaubnissteuer']

Monthly data: 17 rows, 312 columns, 0.24% NaN values
Quarterly data: 17 rows, 100 columns, 0.15% NaN values
Yearly data: 17 rows, 25 columns, 0.00% NaN values
```

Question 1: How has the overall tax income developed over the past 20 years? The total tax income is only available quaterly, so we will use this dataset to answer the first question. The plot below shows a lot of information combined: - The 16 different tax types (excluding the total tax income) are shown as bar plots on the bottom. We can instantly see that income and value added tax make up the biggest chunk of the chosen categories. We can also see a very prevalent pattern: For every year in the dataset, the tax income rises from one quarter to the next, with a significant increase in the fourth quarter - The blue line shows the quarterly total tax income, as taken directly from the dataset. Here, we also see the same pattern of rising tax income from the first to last quarter for each year. The gap between the bars and the blue line also tells us, how much tax income comes from the tax types not included in our selection: This amount seems to be getting bigger over the years, though we would have to look at that theory in more detail for confirmation. - The dotted black line shows the total tax income per year, summed up from the quarterly data. This line therefore has less data points (only one dot for each year) and a different value range (right y axis).

Both the blue and the black line clearly show us that the tax income has increased compared to 20 years ago. We do, however see 1) a period of stagnation (1999-2005) 2) a sharp increase (2006-2008) 3) followed by a short decline (2009-2010), possibly due to the financial crisis in 2008 4) a steady increase (2011-2023) 5) that is only disrupted heavily by the Covid crisis in 2020

[ ]:
```
ax = plt.figure(figsize=(20, 10)).gca()
ax2 = ax.twinx()

# use the pandas bar plot function for convenience and assign it to our␣
 ↪existing axis
df_q[1:].T.plot(ax=ax, kind='bar', stacked=True, colormap='tab20')
# plot the total tax income per quarter (using left y-axis)
ax.plot(df_q.iloc[0, :], linewidth=2, label='Steuereinnahmen insgesamt␣
 ↪(Quartal)')
# and per year on the 4th quarter (using right y-axis)
ax2.plot(range(3, len(df_q.columns), 4), df_y.iloc[0, :], 'o--', color='black',␣
 ↪linewidth=2, label='Steuereinnahmen insgesamt (Jahr)')

ax.legend()
ax2.legend()
# set the x-axis ticks to the quarter numbers
ax.set_xticks(range(0, len(df_q.columns), 4))
# set y-axis limits
ax.set_ylim(0, 2.5e8)
ax2.set_ylim(0, 10e8)
# format the right y-axis to use the same scientific notation unit as the left␣
 ↪y-axis
ax2.ticklabel_format(axis='y', style='sci', scilimits=(8, 8))

# add title and axis labels
ax.set_title('Overall Tax Income per Quarter and Year', fontsize=20)
ax.set_xlabel('Quarter')
ax.set_ylabel('Tax income (Thousands of Euro)')
ax2.set_ylabel('Tax income (Thousands of Euro)')
plt.savefig(BASE_DIR + 'results/question_1.png')
plt.show()
```



Question 2: Can we see seasonal variations in the data, specifically in the consumer goods taxes? To answer the second question, we will look more closely at 6 different taxes and use the monthly dataset. We will also use a different visualization: By showing each year as an individual line plot from January to December, we can see how the consumer goods tax income changes both within a year and over multiple years. The figures each show the years 1999-2019 in a light to darker blue, and the (post-)covid years in red, yellow, and green for distinction. The mean value over all years is shown in a black dotted line.

Looking at the data, we have to keep in mind that tax income values generally increase through the year with a huge spike in December, as we saw earlier. Still, the six figures tell us multiple interesting insights:

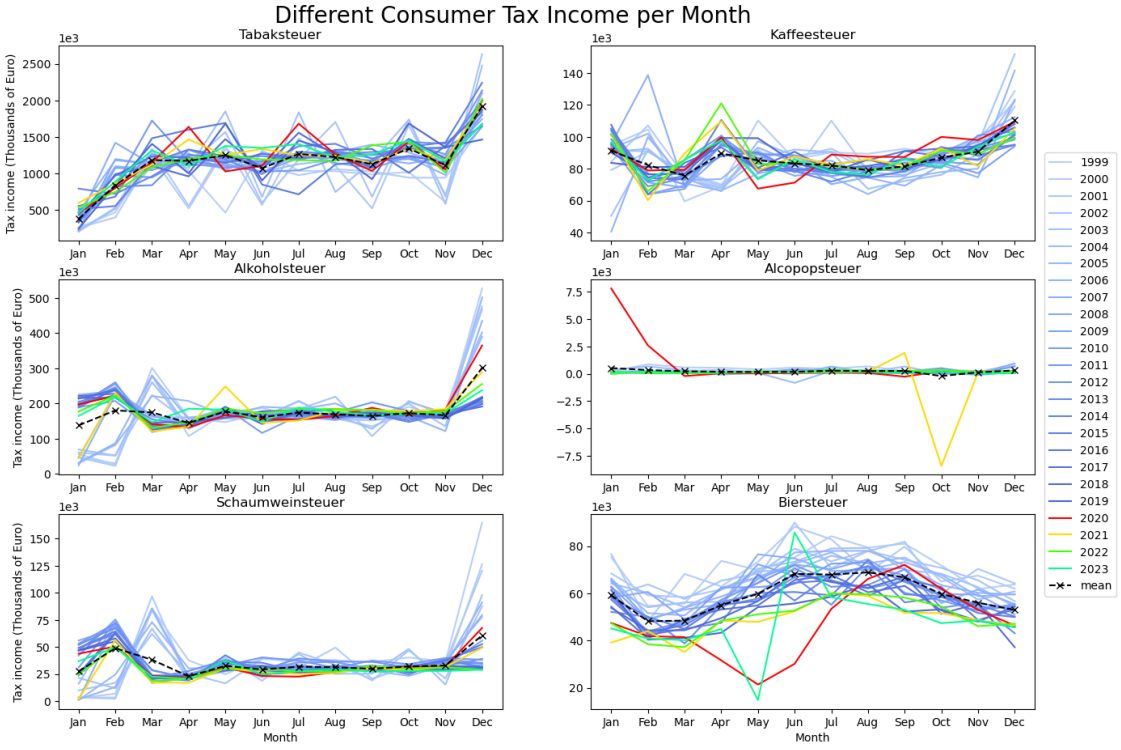
- The tabacco tax income (upper left) shows a clear dip in January and Februar, probably relating to many people's new year's resolutions. We can also see that many people presumably drop their resolutions as early as March, when tax income recoveres to a steady level for the rest of the year.
- Coffe tax income (upper right) shows a distinctive spike in April, which could be related to Easter Holidays, which is a typically celebrated with a brunch and coffee by many families.

- Sparkling wine tax income (lower left) shows clear spikes in March (formerly) and February (more recently). These spikes could be related to carnival celebrations (though the shift from March to February remains unexplained). The December spike in the data is also stronger than for the other tax types, likely due to preparations for new year's celebrations. The sparkling wine tax income pattern also dominates the total alcohol tax income pattern (middle left).
- The beer tax income (lower right) shows a clear seasonal curve, with high values in summer months and lower values during spring and winter (except for January, interestingly). Here we can also see the strongest impact of the Covid pandemic during the early summer months

of 2020. Additionally, we can see a general downwards trend in the beer tax income over the years, with the lighter blue lines consistently lining up above the darker blue lines.
- The alcopop tax income (middle right) is hard to interpret, with small values and little variation and two outliners visible.

[ ]:
```
fig, ax = plt.subplots(figsize=(15, 10), nrows=3, ncols=2)
# indices for the tax types we want to analyze
consumer_tax_indeces = [3, 4, 5, 6, 7, 14]

for i, tax_index in enumerate(consumer_tax_indeces):
    # create a new dataframe with years as row and months as columns
    # only for the selected tax type for easier visualization
    data = pd.DataFrame(index=range(1999, 2024), columns=range(1, 13))
    for year in range(0, 25):
        data.iloc[year, :] = df_m.iloc[tax_index, year*12:year*12+12]

    # two different color maps for pre and post covid data
    colors = iter([plt.cm.coolwarm(a) for a in reversed(range(0, 100, 100//
 ↪21))])
    covid_colors = iter([plt.cm.hsv(a) for a in range(0, 150, 150//4)])

    # index for the subplot location
    plot_loc = (i // 2, i % 2)

    for year in range(0, 25):
        # skip the first 6 years of the 'Alcopops' tax type
        # as it was only introduced in 2005
        if tax_index == 6 and year < 7:
            continue

        # use different colors for pre and post covid data
        if year > 20:
            ax[plot_loc].plot(range(0, 12), data.iloc[year, :], label=str(1999␣
 ↪+ year), color=next(covid_colors), linewidth=1.5)
        else:
            ax[plot_loc].plot(range(0, 12), data.iloc[year, :], label=str(1999␣
 ↪+ year), color=next(colors))

    # plot the mean of all years for each month
    monthly_mean = [data.iloc[:, month].mean() for month in range(0, 12)]
    ax[plot_loc].plot(range(0, 12), monthly_mean, 'x--', label='mean',␣
 ↪color='black')

    # set the same scientific notation unit and month names for all subplots
    ax[plot_loc].ticklabel_format(axis='y', style='sci', scilimits=(3, 3))
    ax[plot_loc].set_xticks(range(0, 12, 1), labels=['Jan', 'Feb', 'Mar',␣
 ↪'Apr', 'May', 'Jun', 'Jul', 'Aug', 'Sep', 'Oct', 'Nov', 'Dec'])
```

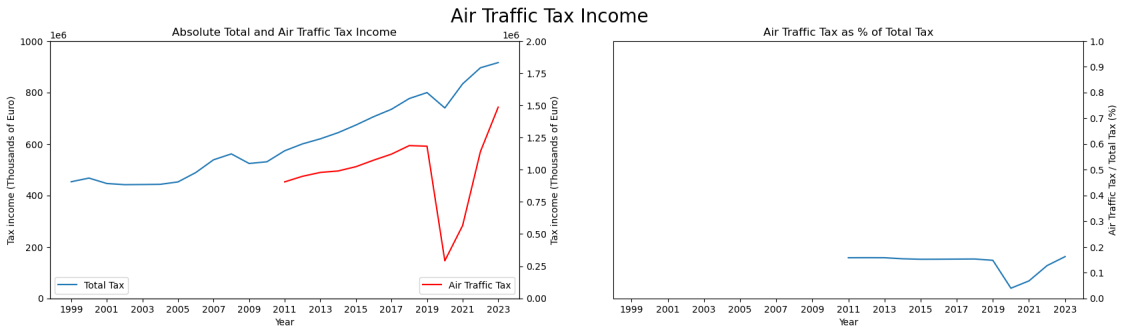
```
    ax[plot_loc].set_title(df_m.index[tax_index])
    if i == 0 or i == 2 or i == 4:
        ax[plot_loc].set_ylabel('Tax income (Thousands of Euro)')
    if i == 5 or i == 4:
        ax[plot_loc].set_xlabel('Month')

# add single legend for all subplots
handles, labels = ax[0, 0].get_legend_handles_labels()
plt.figlegend(handles, labels, loc='right')

fig.suptitle('Different Consumer Tax Income per Month', fontsize=20)
fig.subplots_adjust(right=0.93, top=0.93)
plt.savefig(BASE_DIR + 'results/question_2.png')
plt.show()
```



Question 3: What role did the air traffic tax play in the overall tax income since it's introduction in 2011? For the third question, we only look at the yearly tax income. To achieve this, we can simply sum up 4 subsequent entries from the quarterly dataset and store them in a new dataframe. Plotting the absolute values (left), we can see that the air traffic tax income developed similar to the total tax income. The break in the income due to Covid, is, however much sharper for the air traffic tax income compared to the total tax income. This makes sense, since in 2020 almost no air traffic was possible for significant parts of the year, especially in the tourism sector.

The percentual values (right) show us that air traffic actually makes up only a very small portion of total tax income: Except for the dip due to Covid, air traffic is responsible for a steady 0.16 percent of the total tax income.

[ ]:
```
fig, ax = plt.subplots(figsize=(20, 5), nrows=1, ncols=2)

# plot total tax income
ax[0].plot(df_y.iloc[0, :], label='Total Tax')
# and air traffic tax income in the same plot
ax2 = ax[0].twinx()
ax2.plot(df_y.iloc[10, 12:], 'r', label='Air Traffic Tax')
# plot air traffic tax as percentage of total tax income in second plot
ax[1].plot(df_y.iloc[10, 12:] / df_y.iloc[0, :] * 100)

# figure adjustments
ax[0].legend(loc='lower left')
ax[0].ticklabel_format(axis='y', style='sci', scilimits=(6, 6))
ax[0].set_ylim(0, 1000e6)
ax[0].set_xticks(range(1999, 2024, 2))
ax[0].set_ylabel('Tax income (Thousands of Euro)')
ax[0].set_xlabel('Year')
ax[0].set_title('Absolute Total and Air Traffic Tax Income')

ax2.legend(loc='lower right')
ax2.set_ylim(0, 2e6)
ax2.set_ylabel('Tax income (Thousands of Euro)')

ax[1].set_ylim(0, 1)
ax[1].yaxis.set_ticks(np.arange(0, 1.1, 0.1))
ax[1].yaxis.set_label_position('right')
ax[1].yaxis.tick_right()
ax[1].set_xlim(1998, 2024)
ax[1].set_xticks(range(1999, 2024, 2))
ax[1].set_ylabel('Air Traffic Tax / Total Tax (%)')
ax[1].set_xlabel('Year')
ax[1].set_title('Air Traffic Tax as % of Total Tax')

fig.suptitle('Air Traffic Tax Income', fontsize=20)
plt.savefig(BASE_DIR + 'results/question_3.png')
plt.show()
```



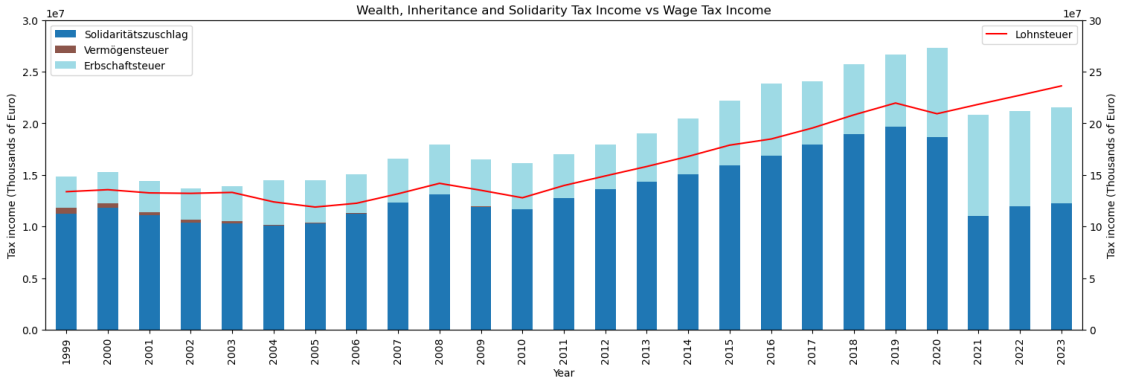
Question 4: How do the wage tax and the three weatlh taxes (asset, solidarity, inheritance) compare to each other? Has this ratio changed over the past 20 years? As before, we will look at the yearly data to answer this question. Firstly we can plot all 4 tax incomes at once (top): The bars show the composition of solidarity, asset, and inheritance tax, while the red line shows the wage tax, which has a scale multiplied by 10. The bars and the line follow a similar pattern overall, with the notable exceptions of 1) the asset tax almost disappearing after 2005 and 2) a decoupling of the two categories after 2020.

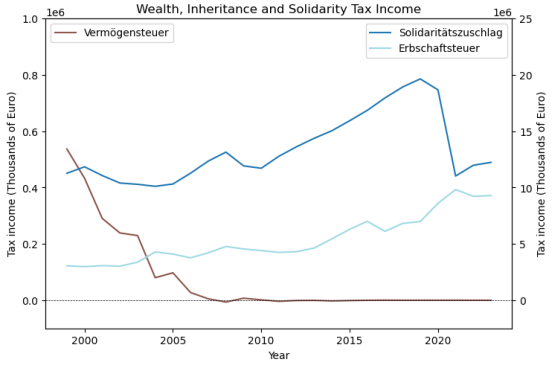
To further investigate this, we can take a closer look at the three wealth taxes (lower left): Here, we can see that the asset tax does indeed almost disappear after 2009, even reaching values below 0 before and after that. The solidarity and inheritance tax, which also have a 25 times higher scale, increase over the years, although with different patterns. The most interesting difference is, that while the inheritance tax income increases in 2021, the solidarity tax drops significantly to its lowest value in the time range.

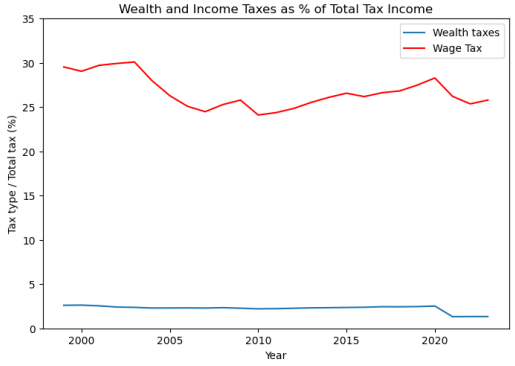
To compare the influence of these taxes on the total tax income, we can again look at their fractions (lower right): Here we can clearly see, that the three wealth taxes make up a steady 2.5% of the total tax income, with a sharp decline in 2021 due to the reduced solidarity tax income. The wage tax makes up a varying fraction of 25-30% of the total tax income, with a sharp decline between 2003 and 2007.

[ ]:
```
fig, ax = plt.subplot_mosaic("AA;BC", constrained_layout=True)

# Figure 1
# use pandas plot function for convenience
df_y.iloc[[11, 12, 13], :].T.plot(ax=ax['A'], kind='bar', stacked=True,␣
 ↪colormap='tab20', figsize=(15, 10))
ax['A'].set_ylim(0, 3e7)
ax['A'].set_xlabel('Year')
ax['A'].set_ylabel('Tax income (Thousands of Euro)')
ax['A'].set_title('Wealth, Inheritance and Solidarity Tax Income vs Wage Tax␣
 ↪Income')
ax['A'].legend()

# right axis for wage tax income
```

```
ax2 = ax['A'].twinx()
ax2.plot(df_y.iloc[1, :].to_list(), 'r', label='Lohnsteuer')
ax2.set_ylim(0, 30e7)
ax2.ticklabel_format(axis='y', style='sci', scilimits=(7, 7))
ax2.set_ylabel('Tax income (Thousands of Euro)')
ax2.legend()

# Figure 2
# only plot the asset tax on left axis
ax['B'].plot(df_y.iloc[12, :], label='Vermögensteuer', color=plt.cm.tab20(10))
ax['B'].axhline(0, linestyle='--', color='black', linewidth=0.5)
ax['B'].set_ylim(-1e5, 1e6)
ax['B'].ticklabel_format(axis='y', style='sci', scilimits=(6, 6))
ax['B'].set_xlabel('Year')
ax['B'].set_ylabel('Tax income (Thousands of Euro)')
ax['B'].set_title('Wealth, Inheritance and Solidarity Tax Income')
ax['B'].legend(loc='upper left')

# right axis for inheritance and solidarity tax income
ax2 = ax['B'].twinx()
ax2.plot(df_y.iloc[11, :], label='Solidaritätszuschlag', color=plt.cm.tab20(0))
ax2.plot(df_y.iloc[13, :], label='Erbschaftsteuer', color=plt.cm.tab20(20))
ax2.set_ylim(-25e5, 25e6)
ax2.ticklabel_format(axis='y', style='sci', scilimits=(6, 6))
ax2.set_ylabel('Tax income (Thousands of Euro)')
ax2.legend(loc = 'upper right')

# Figure 3
# plot wealth and wage tax as percentage of total tax income
ax['C'].plot(df_y.iloc[11:13, :].sum(axis=0) / df_y.iloc[0, :] * 100,␣
 ↪label='Wealth taxes')
ax['C'].plot(df_y.iloc[1, :] / df_y.iloc[0, :] * 100, label='Wage Tax',␣
 ↪color='r')
ax['C'].set_ylim(0, 35)
ax['C'].set_xlabel('Year')
ax['C'].set_ylabel('Tax type / Total tax (%)')
ax['C'].set_title('Wealth and Income Taxes as % of Total Tax Income')
ax['C'].legend()

plt.savefig(BASE_DIR + 'results/question_4.png')
plt.show()
```


Wealth, Inheritance and Solidarity Tax Income vs Wage Tax Income
1e7
1e7
Solidaritätszuschlag
Vermögensteuer
Erbschaftsteuer
Lohnsteuer
Tax income (Thousands of Euro)
Tax income (Thousands of Euro)
Year


Wealth, Inheritance and Solidarity Tax Income
1e6
1e6
Vermögensteuer
Solidaritätszuschlag
Erbschaftsteuer
Tax income (Thousands of Euro)
Tax income (Thousands of Euro)
Year


Wealth and Income Taxes as % of Total Tax Income
Wealth taxes
Wage Tax
Tax type / Total tax (%)
Year