Appendix 3

In [ ]:
```
import pandas as pd
from bs4 import BeautifulSoup
import os
import json
from datetime import datetime
from datetime import timedelta
from datetime import date
import locale
```

2 methods

1. convert read out the information out of a htm or html file and convert it to a csv.
2. itterate through the folder structure, call first method to compute csv with the according information and create master data frame for analysis

In [ ]:
```
folder = 'folder'

def htm_to_csv(html_file, csv_file, id):
    with open(html_file, encoding='utf8') as infile:
        soup = BeautifulSoup(infile, "html.parser")

        csv_file = pd.read_csv(csv_file)

        df = pd.DataFrame()

        offers_text = soup.find(class_='css-1it8m2y').get_text()
        n_offers = int(offers_text.split()[0])


        df['insurance_company'] = [x.get_text() for x in soup.find_all(class_='css

        df['insurance_model'] = [x.get_text() for x in soup.find_all(class_='css-1

        df['user_rating'] = [x.get_text() for x in soup.find_all(class_='css-1r2y0

        df['price'] = [x.get_text() for x in soup.find_all(class_='css-vibb24')][:

        df['date_of_birth'] = csv_file['date_of_birth'][0]
        df['date_of_drivers_license'] = csv_file['date_of_drivers_license'][0]
        df['gender'] = csv_file['gender'][0]
        df['nationality'] = csv_file['nationality'][0]
        df['id'] = id

        df.set_index('id', inplace=True)

        return(df)

def get_all(rootdir):

    counter = 0
    master_df = pd.DataFrame()

    for subdir, dirs, files in os.walk(rootdir):
        html = ''
        csv = ''
        for file in files:
```

```
            file_path = os.path.join(subdir, file)
            if file_path.endswith('.htm') or file_path.endswith('.html'):
                html = file_path
            elif file_path.endswith('.csv'):
                csv = file_path

        if html != '' and csv != '':
            # print(csv)
            # print(html)
            counter += 1
            try:
                master_df = pd.concat([master_df, htm_to_csv(html, csv, counter)],
            except ValueError:
                print("Value Error at" + html)
            except:
                print("Different Error has occured")

            html = ''
            csv = ''
    print(counter)
    return(master_df)

master_df = get_all(folder)

today = date.today()

# calculate age from date of birth

age_list = []
for dob in master_df['date_of_birth']:
    dob = datetime.strptime(dob, '%Y-%m-%d').date()
    age = today.year - dob.year - ((today.month, today.day) < (dob.month, dob.day)
    age_list.append(age)

master_df['age'] = age_list

# calculate time since drivers license from date of diverslicense

dsd_list = []
for dod in master_df['date_of_drivers_license']:
    dod = datetime.strptime(dod, '%Y-%m-%d').date()
    age = today.year - dod.year - ((today.month, today.day) < (dod.month, dod.day)
    dsd_list.append(age)

master_df['time_since_dl'] = dsd_list

price_floats = []

for raw_price in master_df['price']:
    locale.setlocale(locale.LC_ALL, 'de_CH.UTF8')
    conv = locale.localeconv()
    raw_numbers = raw_price.strip(conv['currency_symbol'])
    raw_numbers = raw_numbers.replace("'", "")
    amount = locale.atof(raw_numbers)
    price_floats.append(amount)

master_df['price'] = price_floats

#htm_to_csv('html_files/test_html.htm')
master_df.to_csv('master_df.csv')
```