colect_data_aeat

December 8, 2024

[18]:
```
import pandas as pd
import matplotlib.pyplot as plt
```

[19]:
```
# url_dict = {}
# for year in range(2008,2023):
#     url_dict[year] = {'Total': None,
#    '0-0.5': None,
#    '0.5-1': None,
#    '1-1.5': None,
#    '1.5-2': None,
#    '2-2.5': None,
#    '2.5-3': None,
#    '3-3.5': None,
#    '3.5-4': None,
#    '4-4.5': None,
#    '4.5-5': None,
#    '5-7.5': None,
#    '7.5-10': None,
#    '>10': None,}
# url_dict
```

[20]:
```
base_url = "https://sede.agenciatributaria.gob.es/AEAT/Contenidos_Comunes/
 ↪La_Agencia_Tributaria/Estadisticas/Publicaciones/sites/mercado/{year}/{id}.
 ↪html"
```

[21]:
```
ids = {2008: {'Total': "jrubik7125",
  '0-0.5': "jrubik2817",
  '0.5-1': "jrubikeae7",
  '1-1.5': "jrubikec1a",
  '1.5-2': "jrubikc3cf",
  '2-2.5': "jrubik298d",
  '2.5-3': "jrubikd50c",
  '3-3.5': "jrubik8cce",
  '3.5-4': "jrubike09c",
  '4-4.5': "jrubikde50",
  '4.5-5': "jrubik3627",
  '5-7.5': "jrubik128d",
  '7.5-10': "jrubik8727",
```

```
 '>10': "jrubik9bb9"},
2009: {'Total': "jrubikf409",
 '0-0.5': "jrubika3ae",
 '0.5-1': "jrubik9d7d",
 '1-1.5': "jrubik29cb",
 '1.5-2': "jrubikcd1d",
 '2-2.5': "jrubik30dc",
 '2.5-3': "jrubikd8f8",
 '3-3.5': "jrubikfc68",
 '3.5-4': "jrubika62b",
 '4-4.5': "jrubikc952",
 '4.5-5': "jrubik33d4",
 '5-7.5': "jrubik4eb0",
 '7.5-10': "jrubikd4e1",
 '>10': "jrubikfae1"},
2010: {'Total': "jrubik7be4",
 '0-0.5': "jrubikf843",
 '0.5-1': "jrubik1b6f",
 '1-1.5': "jrubik811a",
 '1.5-2': "jrubikdb42",
 '2-2.5': "jrubikade6",
 '2.5-3': "jrubikc670",
 '3-3.5': "jrubikae56",
 '3.5-4': "jrubikebdf",
 '4-4.5': "jrubika251",
 '4.5-5': "jrubik650a",
 '5-7.5': "jrubik42d3",
 '7.5-10': "jrubik84b0",
 '>10': "jrubik8443"},
2011: {'Total': "jrubikf30da5f83d15fb9603527d60e94641fae8825e497",
 '0-0.5': "jrubikf537326ad6a699215a89b2d4bbccc816a43822d3f",
 '0.5-1': "jrubikf445dcbea10c984b641d96aaaef0554436f6eddfa",
 '1-1.5': "jrubikf484c592473528e0435e771ad33ab67f480be17a7",
 '1.5-2': "jrubik160aa2a427ab5513fca5088448bd173d5afa1182",
 '2-2.5': "jrubikf3f1e201eb5cffe47022ed3774b6a04816f04d071",
 '2.5-3': "jrubikf1081c591058370d5f5f8695f30838888398f122a",
 '3-3.5': "jrubik20697493f52d176237411ba171d8035245c94f13",
 '3.5-4': "jrubik73208c0fa991a81d46bb70a2a9ef72d8dee320a2",
 '4-4.5': "jrubik7956f32ccda8c388aa2ba52bc69e4ccf5bb5eab5",
 '4.5-5': "jrubikf5bae79fa7da04ac694e6d17de72d49689b3ffb28",
 '5-7.5': "jrubik2f78defde6daadb30da02fcb2c34627ece2510be",
 '7.5-10': "jrubikf7ca27fa5d7ff0b9ebb6c52130d832e833fe5a62a",
 '>10': "jrubikf4863f2dfc2f810804adc3c435efdaffa98517cf5"},
2012: {'Total': "jrubik3c8b4fca9b22cbe4a1825a98283f957b044fb82b",
 '0-0.5': "jrubikf1b957c66594f27aaf5a216ea029769a59420344d",
 '0.5-1': "jrubikf49afba57c2e192fc261b4fcf8c26a9114b47e477",
 '1-1.5': "jrubik2641df70420048a88eb98ace07b8d1afa965e7b0",
```

```
 '1.5-2': "jrubik7ad66ada29e5eb9bbc9b15cc2062937d0f47b3ce",
 '2-2.5': "jrubik7d8eadc6c3802d42ec240ee7cecce9074e507bda",
 '2.5-3': "jrubikf4ab46480512e12fe18cd12af556fd39c752cdb18",
 '3-3.5': "jrubik7a7d4b684c1201375ce7d357b9cc02960f453aea",
 '3.5-4': "jrubik1eb713093a8483e810479068b0aebc10f1e28215",
 '4-4.5': "jrubikf543fdebcf969f938a0e1533c9b06117bd6cb9943",
 '4.5-5': "jrubik29214bf3deffd6d61e3def744f3a11a9d263237e",
 '5-7.5': "jrubikf7742db7062158f15f35f368df4dc30350da20ade",
 '7.5-10': "jrubikf7542ae1c9c1551404499beb3eaf8d8c7af229e6d",
 '>10': "jrubikfb619b6910a79484ed676993db08b74d26373be"},
2013: {'Total': "jrubikf5590486e5658081a33152a847e7ca9eadf67bd00",
 '0-0.5': "jrubikf2ab98d9cb1ab23eb019611cc5aba8be2f13cfd4e",
 '0.5-1': "jrubikf1a6c7401b35b9bac7710850f1a223ceca47ae53e",
 '1-1.5': "jrubik23d16ae24abac04715a6b3246a2714231fd5cc2a",
 '1.5-2': "jrubikf4b877f46f3ec57dd3a33aaab616bbaafe71472e5",
 '2-2.5': "jrubik44a2137afbbc26c5bb39558c605f5aad38fa936b",
 '2.5-3': "jrubikf34a5bd86b4baa36e16dd60970e38bd2b213ac539",
 '3-3.5': "jrubikfe36ef3fd9b62b3dc725f2cb827686dcf9ea6ff",
 '3.5-4': "jrubik5c8351b308df59a74e4a8f5bf013139d6eaa28fb",
 '4-4.5': "jrubik6a1a89d9536abdc93d7099ab35cd10933c89de2a",
 '4.5-5': "jrubik2be0320ab9f3fd12275524b4d027f8499ea6b109",
 '5-7.5': "jrubik7b5fa59e7941ad81086471d6a6fb62a44a235b9a",
 '7.5-10': "jrubikf57820074c607557183a07496f4c3a89f38f3560b",
 '>10': "jrubikf672dea112a71a7bdd98b51fb687def23bf808a3f"},
2014: {'Total': "jrubik2722bf59c799ead0bffc44ea5c6b253bfb1d8a40",
 '0-0.5': "jrubik1770a8239e1127215a62947b4f966df8bbc8e229",
 '0.5-1': "jrubikf3c87d861ff46ca9289a300105a0f43e6d7204c34",
 '1-1.5': "jrubikf7a1d97d86c9df5c11295305582966284 98af3e89",
 '1.5-2': "jrubikf50095bf505bc0bc5440c85f454079356cd0055e0",
 '2-2.5': "jrubik1182a0bcbe4985dc094be4ebf5f27d9631acab95",
 '2.5-3': "jrubik2171ba2048bd64b18372291a113ee1c9d079f54a",
 '3-3.5': "jrubik2cc85694ab97cbffc106b1b96a70469705fd19db",
 '3.5-4': "jrubik335ef607bc8c3c27e0f2b77ba7a5d0f3300139d7",
 '4-4.5': "jrubikf591e56a3a74a2501cc6359ba69d59925e8b0235b",
 '4.5-5': "jrubikfd0c811aa4dcf1d59660f3e98a7f9ef196f8a43f",
 '5-7.5': "jrubik7662664e538ff9d0c2154485cdc5fcaf27e8a09d",
 '7.5-10': "jrubik6096e603513a2416cf64b2812a4f09f73d42d488",
 '>10': "jrubikf593b92a7ae63f4724f42d1880cb2753879c8db38"},
2015: {'Total': "jrubik55627506bfba609ffe04c5a96a9ead39ac5e9d26",
 '0-0.5': "jrubikf10f245073b453e2a7f9456faae1abe39248a524e",
 '0.5-1': "jrubik71111e090dcd46994a94baf6672a208c783e6920",
 '1-1.5': "jrubikf6e712d257901a10cf18b95d65b1316b9df66daf3",
 '1.5-2': "jrubikf427f03ede600934bdbdab86cd8a9ff54c50b21ba",
 '2-2.5': "jrubikf1b9bda005d1525c220f0dfc16ef6bbd55d14b6d",
 '2.5-3': "jrubikf5cb3007238be3718d72298d76e3b07d0dcd0400a",
 '3-3.5': "jrubik2d83b362e719936f790aa24645b4ad1e07840eb2",
 '3.5-4': "jrubikd51d85d85814c08b3d2055d4bf1345ac004e51",
```

```
 '4-4.5': "jrubikf45db518359749c0ef77317f1d0faf7a80508f49f",
 '4.5-5': "jrubikf46433aeb27a7103bb9bb8bbe60a0318deac951ad",
 '5-7.5': "jrubikf5415f39251df95c52754436d8819280622979d8d",
 '7.5-10': "jrubikf4e1899587d4c9adb151a3d090eb83aeff41d82d8",
 '>10': "jrubik3750e2645b7e8314942b749feba5dc934afba236"},
2016: {'Total': "jrubik10c01fd8c4e99c1fc3916ab34f8f6250666781b3",
 '0-0.5': "jrubik297b44e169160f65104e6e60befa97d0b3a597e6",
 '0.5-1': "jrubikf378f7c568effb312a394660f67df3cb99375d8fc",
 '1-1.5': "jrubikf5c968bc7c081a48bf33c0301360c6d99e30869",
 '1.5-2': "jrubikf1d4340548498f5ad8a8c684fc20ac2b6912dcc15",
 '2-2.5': "jrubik78a91723a1e73ed96603ad187356aa3073a8d1e0",
 '2.5-3': "jrubik63263f307c1a76f0e914e9ea4c15193fcd21ef75",
 '3-3.5': "jrubikf319be59de9c13ef7a685bb80363413e8b323f300",
 '3.5-4': "jrubik3d1497d464c690646c813f9d5b59339f3fa20370",
 '4-4.5': "jrubikf7e508ec18e27cb715f3abb16b1651979d6aae916",
 '4.5-5': "jrubikf45a1a9e9e2c9cd39bda94296bde59da9786dd63f",
 '5-7.5': "jrubik6c57a083f1266a872c6d570169abc782ddd49225",
 '7.5-10': "jrubik6cf6b1c07e7817586034 95c57dc22a259fb014d",
 '>10': "jrubik56823f859c8521ce66c9484d53de365b4345678b"},
2017: {'Total': "jrubik6b321d30366f388b9379cc09b8b592ab62db52cb",
 '0-0.5': "jrubik67dc0ce9c75e4c5ed0988eb1c93fc87f0856367f",
 '0.5-1': "jrubikf55b4e6caba0d32706e0036571e10e06a9e7a0dec",
 '1-1.5': "jrubikf2666b0a27700486f22fec980006efb8d90e53e7b",
 '1.5-2': "jrubik1173a1f407a53e028df5def5ea4726bebec4d853",
 '2-2.5': "jrubikf615438bb3b9f35d4431559fcaf20919f80c02d4b",
 '2.5-3': "jrubikf4a953aef1cf1f3182af6cd544cdc5630d9cb36d0",
 '3-3.5': "jrubikf2895401504bcc936bc5f5a1eb82599ab60844fb2",
 '3.5-4': "jrubikf68e06f3125175bd845668aa975f2a4a0d027fbfb",
 '4-4.5': "jrubikfcd25e4047dfd9ff2d415d7ed114c73e91d7117d",
 '4.5-5': "jrubik4b38cc19303af45f2b0f9b72cc1f090f1e8b9748",
 '5-7.5': "jrubikf72cb63e6b940cb21f80902c8146b1868b34b7cf8",
 '7.5-10': "jrubik2929595de537b250d247e211baae82d84f0617da",
 '>10': "jrubikf60d68cce808c5382be7dc69f89eae6e4a926a67a"},
2018: {'Total': "jrubikf13275bf1ef5d96e7febc25aa100156774ff5c8ad",
 '0-0.5': "jrubik1adc1d27f48ee29d14587ecb8b94e27476b3a174",
 '0.5-1': "jrubikf5a03dd6555efb3b7f9d5f84c23f74ab36de45c31",
 '1-1.5': "jrubikf1c26a7b4a4643c2c5a9d31103d15813de144a23b",
 '1.5-2': "jrubik5c065ac90866a050d362ffe345d971509d429202",
 '2-2.5': "jrubikf2143d8eb6a2a2de6dbb441e86916cf97702e73a2",
 '2.5-3': "jrubikf4afccc9716dcdb2c47804d8daab15127b82b5359",
 '3-3.5': "jrubikf8939df248692783906133c64bc0503777ac726f",
 '3.5-4': "jrubikf56c80411e80af54648629e3b87ec07c300d9ee67",
 '4-4.5': "jrubik5c89e9356572639c63feff7d6a0a19c4681374fc",
 '4.5-5': "jrubik2c30f4843fabe8ac4635bb6d4dfe4a633ca20c8e",
 '5-7.5': "jrubik1a4f1ae16413e820e5f5996139d520d5d3dcabd9",
 '7.5-10': "jrubikf5bbae21ffcf400829820d9b972eed2e6dc9e01fb",
 '>10': "jrubik2393a9f7915721dd5374053624b3bf6968728c99"},
```

```
2019: {'Total': "jrubik7d746b5c70fcbedffe15a1f9af327deecaa11188",
 '0-0.5': "jrubikf7a2012a21b05126271da8fda7216208f3ad39553",
 '0.5-1': "jrubik730f04f8988a1bdd96a9afd23a5eba9d4fcddf8a",
 '1-1.5': "jrubik646e93bde1008afa38d7b8cc83fa47ef8c496f90",
 '1.5-2': "jrubikf443a5044d782a72d101a442d5a0041900522e449",
 '2-2.5': "jrubik24af4eefe4b15af6c1aa3bb3a4381224b03ac54e",
 '2.5-3': "jrubik232b132ac1ffa06e9c4297724f4604d7e76f7141",
 '3-3.5': "jrubikf6c522e0823a4a79896496bebb74e8b7dbd6d5618",
 '3.5-4': "jrubikf371e3d135a1cf29542c5e9ed20579f0d41654864",
 '4-4.5': "jrubikf53da22d345259a6cdc2d372afffc6b621c1d84c2",
 '4.5-5': "jrubikf477a1dc4952f99941e8c9accf95d2c8e9eed4d01",
 '5-7.5': "jrubikcc2c60520011a6b6713292920ae1c7a989e48db",
 '7.5-10': "jrubik1018083eb6eb9d0132b14db829b34b8a0b8fb521",
 '>10': "jrubik5c1bacb2f607c48d1380784a4158e3a4de72a6ac"},
2020: {'Total': "jrubik21acaa390f5744160c4e962a8b423875bf6cf180",
 '0-0.5': "jrubik1997c06f719a7f829a095c3a5800e8219f72f66b",
 '0.5-1': "jrubik21ca28f7695a2d5bc227c647ed428899782e4077",
 '1-1.5': "jrubik59f2a99eb26f92daf6a5a439dd8416c1abd6598e",
 '1.5-2': "jrubik15324b736d2e079a152af6bfb7e6cdfe7f6371e6",
 '2-2.5': "jrubikf16a0d3167c7d5de17e8709274b49bf8b927a62e0",
 '2.5-3': "jrubike3ec316d57e6faf6277c958ea07433e7c9aaf40",
 '3-3.5': "jrubik3e8048fd56089d4c4cae835351335b039656515a",
 '3.5-4': "jrubik3c74041a9a97a1a362d6e03d7dd5748f73338d6b",
 '4-4.5': "jrubik611e66d003aadd17cad1c6f055254a3e4ccdc371",
 '4.5-5': "jrubikf4f06d4f22adc6b9b072fd01b11eadf4cfd0908bf",
 '5-7.5': "jrubik32bc5133e5985f27177f36aba26508036b0810d0",
 '7.5-10': "jrubik3eb5cfabbddc127ea38b58318e8e2bee5b990615",
 '>10': "jrubik3c5965de51c7d23be13fea7dd08f4b98937f71ab"},
2021: {'Total': "jrubik4e41fb505966eabffec242df253d8ea082d12628",
 '0-0.5': "jrubikf58d914eb2a814036834a526935a84001cf781660",
 '0.5-1': "jrubikf4c08eb5d719daedfa1ba10ae206a68e9d6b6146a",
 '1-1.5': "jrubik5b8a6936c98db5fcd438a284b8e282f2c0752fc6",
 '1.5-2': "jrubikf6c63cc9da6659f8f7d878e58838ea7a5adc196f1",
 '2-2.5': "jrubik3d7ff622e7eca89c90df7715c049c426390dddcd",
 '2.5-3': "jrubikf55db8121a6dc822a65841755243b68944c56bcb6",
 '3-3.5': "jrubikf6461c93af80473fe76471c8fc86f2d04683eb8a5",
 '3.5-4': "jrubikf2e1ac29b694ce48274bab47f3d6a7110c0df7549",
 '4-4.5': "jrubik7221bf7774256a375dfbb5857ebeae9f5362e949",
 '4.5-5': "jrubikf7b596103e3c49782f185e99dc7263fcefae21106",
 '5-7.5': "jrubikf123f8eca72dddfdc556d91c792a36536cab57aff",
 '7.5-10': "jrubikf2a44318af52697212912fde8d992768391cac700",
 '>10': "jrubik61c0db0345206254f0a611e8d9bb7b694586a263"},
2022: {'Total': "jrubikf40b4df54aee0b0d2d19c4d019958bfeeab3fbc31",
 '0-0.5': "jrubikf0cbd83967a5667ea711b3bd22ac1f41db9c4db",
 '0.5-1': "jrubikf2747f38a914be03287683d9f6ad1594acc3beacd",
 '1-1.5': "jrubik1986fedb861e0bf14ea19e019cd0aaed6ff4e8c7",
 '1.5-2': "jrubikf6c189cf10415f3b44d5804ab1052619c8fdce0b",
```

```
  '2-2.5': "jrubik3fc5bfd14b94f1df31515ad4e78428bf122fb357",
  '2.5-3': "jrubikf5a2ac8b0cf398103170d1936518cac691acad67e",
  '3-3.5': "jrubik5433d13baab5730d7bcb62d88749a55861dbb300",
  '3.5-4': "jrubikf47fea45fafd4aef9d5107edb24d6ab0dfb8951cf",
  '4-4.5': "jrubikf53f743e24630f9fa620df67e55097daf0c887272",
  '4.5-5': "jrubikf6ad1ae215327d2b53e8636ef7e34144ee02b0eb0",
  '5-7.5': "jrubik174cfa163acf328b8d0cf6a36176bf8281247c6b",
  '7.5-10': "jrubikf71213292377967 56e0659d1126157005e532f47e",
  '>10': "jrubik3d3e3e2c8b51ee0655ac784b23312d9f5167a6c9"}}
```

[22]:
```
complete_table = pd.DataFrame()
for year, smis in ids.items():
    for smi, id in smis.items():
        print(year, smi)
        tables = pd.read_html(base_url.format(year = year, id = id),thousands='.
 ↪', decimal=',', encoding='ISO-8859-1')
        df = tables[0]
        df["year"] = year
        df["smi"] = smi
        complete_table = pd.concat([complete_table, df], ignore_index=True)
```

```
2008 Total
2008 0-0.5
2008 0.5-1
2008 1-1.5
2008 1.5-2
2008 2-2.5
2008 2.5-3
2008 3-3.5
2008 3.5-4
2008 4-4.5
2008 4.5-5
2008 5-7.5
2008 7.5-10
2008 >10
2009 Total
2009 0-0.5
2009 0.5-1
2009 1-1.5
2009 1.5-2
2009 2-2.5
2009 2.5-3
2009 3-3.5
2009 3.5-4
2009 4-4.5
2009 4.5-5
2009 5-7.5
2009 7.5-10
```

```
2009 >10
2010 Total

/Users/diegomolina/.virtualenvs/master_uned/lib/python3.11/site-
packages/bs4/__init__.py:228: UserWarning: You provided Unicode markup but also
provided a value for from_encoding. Your from_encoding will be ignored.
  warnings.warn("You provided Unicode markup but also provided a value for
from_encoding. Your from_encoding will be ignored.")

2010 0-0.5

/Users/diegomolina/.virtualenvs/master_uned/lib/python3.11/site-
packages/bs4/__init__.py:228: UserWarning: You provided Unicode markup but also
provided a value for from_encoding. Your from_encoding will be ignored.
  warnings.warn("You provided Unicode markup but also provided a value for
from_encoding. Your from_encoding will be ignored.")

2010 0.5-1

/Users/diegomolina/.virtualenvs/master_uned/lib/python3.11/site-
packages/bs4/__init__.py:228: UserWarning: You provided Unicode markup but also
provided a value for from_encoding. Your from_encoding will be ignored.
  warnings.warn("You provided Unicode markup but also provided a value for
from_encoding. Your from_encoding will be ignored.")

2010 1-1.5

/Users/diegomolina/.virtualenvs/master_uned/lib/python3.11/site-
packages/bs4/__init__.py:228: UserWarning: You provided Unicode markup but also
provided a value for from_encoding. Your from_encoding will be ignored.
  warnings.warn("You provided Unicode markup but also provided a value for
from_encoding. Your from_encoding will be ignored.")

2010 1.5-2

/Users/diegomolina/.virtualenvs/master_uned/lib/python3.11/site-
packages/bs4/__init__.py:228: UserWarning: You provided Unicode markup but also
provided a value for from_encoding. Your from_encoding will be ignored.
  warnings.warn("You provided Unicode markup but also provided a value for
from_encoding. Your from_encoding will be ignored.")

2010 2-2.5

/Users/diegomolina/.virtualenvs/master_uned/lib/python3.11/site-
packages/bs4/__init__.py:228: UserWarning: You provided Unicode markup but also
provided a value for from_encoding. Your from_encoding will be ignored.
  warnings.warn("You provided Unicode markup but also provided a value for
from_encoding. Your from_encoding will be ignored.")

2010 2.5-3

/Users/diegomolina/.virtualenvs/master_uned/lib/python3.11/site-
packages/bs4/__init__.py:228: UserWarning: You provided Unicode markup but also
provided a value for from_encoding. Your from_encoding will be ignored.
```

```
  warnings.warn("You provided Unicode markup but also provided a value for
from_encoding. Your from_encoding will be ignored.")

2010 3-3.5

/Users/diegomolina/.virtualenvs/master_uned/lib/python3.11/site-
packages/bs4/__init__.py:228: UserWarning: You provided Unicode markup but also
provided a value for from_encoding. Your from_encoding will be ignored.
  warnings.warn("You provided Unicode markup but also provided a value for
from_encoding. Your from_encoding will be ignored.")

2010 3.5-4

/Users/diegomolina/.virtualenvs/master_uned/lib/python3.11/site-
packages/bs4/__init__.py:228: UserWarning: You provided Unicode markup but also
provided a value for from_encoding. Your from_encoding will be ignored.
  warnings.warn("You provided Unicode markup but also provided a value for
from_encoding. Your from_encoding will be ignored.")

2010 4-4.5

/Users/diegomolina/.virtualenvs/master_uned/lib/python3.11/site-
packages/bs4/__init__.py:228: UserWarning: You provided Unicode markup but also
provided a value for from_encoding. Your from_encoding will be ignored.
  warnings.warn("You provided Unicode markup but also provided a value for
from_encoding. Your from_encoding will be ignored.")

2010 4.5-5

/Users/diegomolina/.virtualenvs/master_uned/lib/python3.11/site-
packages/bs4/__init__.py:228: UserWarning: You provided Unicode markup but also
provided a value for from_encoding. Your from_encoding will be ignored.
  warnings.warn("You provided Unicode markup but also provided a value for
from_encoding. Your from_encoding will be ignored.")

2010 5-7.5

/Users/diegomolina/.virtualenvs/master_uned/lib/python3.11/site-
packages/bs4/__init__.py:228: UserWarning: You provided Unicode markup but also
provided a value for from_encoding. Your from_encoding will be ignored.
  warnings.warn("You provided Unicode markup but also provided a value for
from_encoding. Your from_encoding will be ignored.")

2010 7.5-10

/Users/diegomolina/.virtualenvs/master_uned/lib/python3.11/site-
packages/bs4/__init__.py:228: UserWarning: You provided Unicode markup but also
provided a value for from_encoding. Your from_encoding will be ignored.
  warnings.warn("You provided Unicode markup but also provided a value for
from_encoding. Your from_encoding will be ignored.")

2010 >10

/Users/diegomolina/.virtualenvs/master_uned/lib/python3.11/site-
packages/bs4/__init__.py:228: UserWarning: You provided Unicode markup but also
```

```
provided a value for from_encoding. Your from_encoding will be ignored.
  warnings.warn("You provided Unicode markup but also provided a value for
from_encoding. Your from_encoding will be ignored.")
```

```
2011 Total
2011 0-0.5
2011 0.5-1
2011 1-1.5
2011 1.5-2
2011 2-2.5
2011 2.5-3
2011 3-3.5
2011 3.5-4
2011 4-4.5
2011 4.5-5
2011 5-7.5
2011 7.5-10
2011 >10
2012 Total
2012 0-0.5
2012 0.5-1
2012 1-1.5
2012 1.5-2
2012 2-2.5
2012 2.5-3
2012 3-3.5
2012 3.5-4
2012 4-4.5
2012 4.5-5
2012 5-7.5
2012 7.5-10
2012 >10
2013 Total
2013 0-0.5
2013 0.5-1
2013 1-1.5
2013 1.5-2
2013 2-2.5
2013 2.5-3
2013 3-3.5
2013 3.5-4
2013 4-4.5
2013 4.5-5
2013 5-7.5
2013 7.5-10
2013 >10
2014 Total
2014 0-0.5
```

2014 0.5-1
2014 1-1.5
2014 1.5-2
2014 2-2.5
2014 2.5-3
2014 3-3.5
2014 3.5-4
2014 4-4.5
2014 4.5-5
2014 5-7.5
2014 7.5-10
2014 >10
2015 Total
2015 0-0.5
2015 0.5-1
2015 1-1.5
2015 1.5-2
2015 2-2.5
2015 2.5-3
2015 3-3.5
2015 3.5-4
2015 4-4.5
2015 4.5-5
2015 5-7.5
2015 7.5-10
2015 >10
2016 Total
2016 0-0.5
2016 0.5-1
2016 1-1.5
2016 1.5-2
2016 2-2.5
2016 2.5-3
2016 3-3.5
2016 3.5-4
2016 4-4.5
2016 4.5-5
2016 5-7.5
2016 7.5-10
2016 >10
2017 Total
2017 0-0.5
2017 0.5-1
2017 1-1.5
2017 1.5-2
2017 2-2.5
2017 2.5-3
2017 3-3.5

2017 3.5-4
2017 4-4.5
2017 4.5-5
2017 5-7.5
2017 7.5-10
2017 >10
2018 Total
2018 0-0.5
2018 0.5-1
2018 1-1.5
2018 1.5-2
2018 2-2.5
2018 2.5-3
2018 3-3.5
2018 3.5-4
2018 4-4.5
2018 4.5-5
2018 5-7.5
2018 7.5-10
2018 >10
2019 Total
2019 0-0.5
2019 0.5-1
2019 1-1.5
2019 1.5-2
2019 2-2.5
2019 2.5-3
2019 3-3.5
2019 3.5-4
2019 4-4.5
2019 4.5-5
2019 5-7.5
2019 7.5-10
2019 >10
2020 Total
2020 0-0.5
2020 0.5-1
2020 1-1.5
2020 1.5-2
2020 2-2.5
2020 2.5-3
2020 3-3.5
2020 3.5-4
2020 4-4.5
2020 4.5-5
2020 5-7.5
2020 7.5-10
2020 >10

```
2021 Total
2021 0-0.5
2021 0.5-1
2021 1-1.5
2021 1.5-2
2021 2-2.5
2021 2.5-3
2021 3-3.5
2021 3.5-4
2021 4-4.5
2021 4.5-5
2021 5-7.5
2021 7.5-10
2021 >10
2022 Total
2022 0-0.5
2022 0.5-1
2022 1-1.5
2022 1.5-2
2022 2-2.5
2022 2.5-3
2022 3-3.5
2022 3.5-4
2022 4-4.5
2022 4.5-5
2022 5-7.5
2022 7.5-10
2022 >10
```

[23]:
```
SMI = pd.read_excel("../data/smi/smi_2008_2024.xlsx", sheet_name="Datos",␣
↪skiprows=4)
```

[24]:
```
SMI.rename(columns={'Unnamed: 1': 'periodo', 'Unnamed: 2': 'smi_14'},␣
↪inplace=True)
SMI.drop(["Unnamed: 0"], axis=1, inplace=True)
```

[25]:
```
rename_dict = {'Unnamed: 0': 'ccaa','Region': 'ccaa', 'Percepciones por␣
↪personas': 'percepciones', 'Asalariados': 'asalariados',
               'Salario Medio Anual': 'sma', 'Salarios':'salarios', 'year':␣
↪'periodo', 'Percepciones por persona': 'percepciones_persona'}
complete_table = complete_table.rename(columns=rename_dict, inplace=False).
↪merge(SMI, how="left", on=["periodo"])
```

[26]:
```
complete_table
```

[26]:

	ccaa	asalariados	percepciones_persona	salarios	\
0	Total	19310627	1.34	366818775121	

```
1                  Andalucía      3459137                  1.41   53933824506
2                     Aragón       624119                  1.35   12388667131
3     Principado de Asturias       439553                  1.32    8663847179
4              Illes Balears       496056                  1.32    8917118969
…                        …            …                     …           …
3719    Madrid, Comunidad de        37496                  1.18   10324752096
3720       Murcia, Región de          664                  1.37     144174288
3721                Rioja, La          174                  1.20      37503780
3722                    Ceuta           26                  2.00       5630484
3723                  Melilla           22                  1.36       3580904

          sma  periodo    smi  smi_14
0       18996     2008  Total   600.0
1       15592     2008  Total   600.0
2       19850     2008  Total   600.0
3       19711     2008  Total   600.0
4       17976     2008  Total   600.0
…         …       …     …       …
3719   275356     2022    >10  1000.0
3720   217130     2022    >10  1000.0
3721   215539     2022    >10  1000.0
3722   216557     2022    >10  1000.0
3723   162768     2022    >10  1000.0

[3724 rows x 8 columns]
```

[27]:
```python
#Rename de ccaa
mapping_ccaa = {
    'Andalucía': 'Andalucía',
    'Aragón': 'Aragón',
    'Principado de Asturias': 'Asturias, Principado de',
    'Illes Balears': 'Balears, Illes',
    'Canarias': 'Canarias',
    'Cantabria': 'Cantabria',
    'Castilla - La Mancha': 'Castilla - La Mancha',
    'Castilla y León': 'Castilla y León',
    'Cataluña': 'Cataluña',
    'Extremadura': 'Extremadura',
    'Galicia': 'Galicia',
    'Comunidad de Madrid': 'Madrid, Comunidad de',
    'Región de Murcia': 'Murcia, Región de',
    'La Rioja': 'Rioja, La',
    'Comunidad Valenciana': 'Comunitat Valenciana',
    'Ciudad de Ceuta': 'Ceuta',
    'Ciudad de Melilla': 'Melilla',
    'Asturias, Principado de': 'Asturias, Principado de',
    'Balears,Illes': 'Balears, Illes',
```

```
    'Madrid, Comunidad de': 'Madrid, Comunidad de',
    'Murcia, Región de': 'Murcia, Región de',
    'Rioja, La': 'Rioja, La',
    'Comunitat Valenciana': 'Comunitat Valenciana',
}
complete_table['ccaa'] = complete_table['ccaa'].replace(mapping_ccaa)
```

[28]:
```
complete_table
```

[28]:
```
                         ccaa  asalariados  percepciones_persona  \
0                       Total     19310627                  1.34
1                   Andalucía      3459137                  1.41
2                      Aragón       624119                  1.35
3     Asturias, Principado de       439553                  1.32
4              Balears, Illes       496056                  1.32
…                         …           …                     …
3719     Madrid, Comunidad de        37496                  1.18
3720        Murcia, Región de          664                  1.37
3721                Rioja, La          174                  1.20
3722                    Ceuta           26                  2.00
3723                  Melilla           22                  1.36

          salarios     sma  periodo    smi  smi_14
0     366818775121   18996     2008  Total   600.0
1      53933824506   15592     2008  Total   600.0
2      12388667131   19850     2008  Total   600.0
3       8663847179   19711     2008  Total   600.0
4       8917118969   17976     2008  Total   600.0
…              …       …        …      …
3719   10324752096  275356     2022    >10  1000.0
3720     144174288  217130     2022    >10  1000.0
3721      37503780  215539     2022    >10  1000.0
3722       5630484  216557     2022    >10  1000.0
3723       3580904  162768     2022    >10  1000.0

[3724 rows x 8 columns]
```

[29]:
```
complete_table.to_csv("../processed_data/salarios/salarios_smis_aeat.csv",␣
 ↪index=False)
```

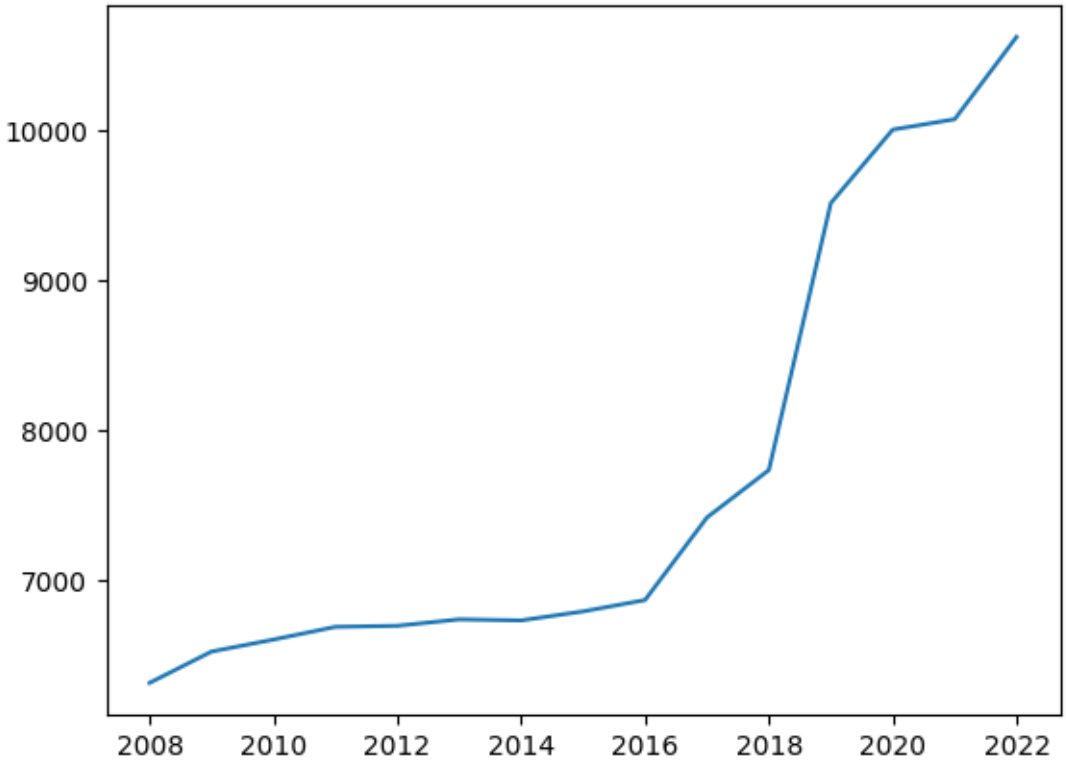
[30]:
```
subset = complete_table[(complete_table["Region"] == "Total") &␣
 ↪(complete_table["smi"] == "0.5-1")]
```

```
---------------------------------------------------------------------------
KeyError                                  Traceback (most recent call last)
File ~/.virtualenvs/master_uned/lib/python3.11/site-packages/pandas/core/indexes/
 ↪base.py:3805, in Index.get_loc(self, key)
```

```
   3804 try:
-> 3805     return self._engine.get_loc(casted_key)
   3806 except KeyError as err:

File index.pyx:167, in pandas._libs.index.IndexEngine.get_loc()

File index.pyx:196, in pandas._libs.index.IndexEngine.get_loc()

File pandas/_libs/hashtable_class_helper.pxi:7081, in pandas._libs.hashtable.
 ↪PyObjectHashTable.get_item()

File pandas/_libs/hashtable_class_helper.pxi:7089, in pandas._libs.hashtable.
 ↪PyObjectHashTable.get_item()

KeyError: 'Region'

The above exception was the direct cause of the following exception:

KeyError                                  Traceback (most recent call last)
Cell In[30], line 1
----> 1 subset = complete_table[(complete_table["Region"] == "Total") &␣
 ↪(complete_table["smi"] == "0.5-1")]

File ~/.virtualenvs/master_uned/lib/python3.11/site-packages/pandas/core/frame.
 ↪py:4102, in DataFrame.__getitem__(self, key)
   4100 if self.columns.nlevels > 1:
   4101     return self._getitem_multilevel(key)
-> 4102 indexer = self.columns.get_loc(key)
   4103 if is_integer(indexer):
   4104     indexer = [indexer]

File ~/.virtualenvs/master_uned/lib/python3.11/site-packages/pandas/core/indexe
 ↪base.py:3812, in Index.get_loc(self, key)
   3807     if isinstance(casted_key, slice) or (
   3808         isinstance(casted_key, abc.Iterable)
   3809         and any(isinstance(x, slice) for x in casted_key)
   3810     ):
   3811         raise InvalidIndexError(key)
-> 3812     raise KeyError(key) from err
   3813 except TypeError:
   3814     # If we have a listlike key, _check_indexing_error will raise
   3815     #  InvalidIndexError. Otherwise we fall through and re-raise
   3816     #  the TypeError.
   3817     self._check_indexing_error(key)

KeyError: 'Region'
```

[ ]:
```
plt.plot(subset["year"], subset["Salario Medio Anual"])
```

[ ]:
```
[<matplotlib.lines.Line2D at 0x12ad61890>]
```



[ ]:

[ ]:
```
SMI
```

[ ]:

	Unnamed: 0	Unnamed: 1	Unnamed: 2
0	NaN	2008	600.0
1	NaN	2009	624.0
2	NaN	2010	633.3
3	NaN	2011	641.4
4	NaN	2012	641.4
5	NaN	2013	645.3
6	NaN	2014	645.3
7	NaN	2015	648.6
8	NaN	2016	655.2
9	NaN	2017	707.7
10	NaN	2018	735.9
11	NaN	2019	900.0
12	NaN	2020	950.0

```
13         NaN        2021       965.0
14         NaN        2022      1000.0
15         NaN        2023      1080.0
16         NaN        2024      1134.0
```

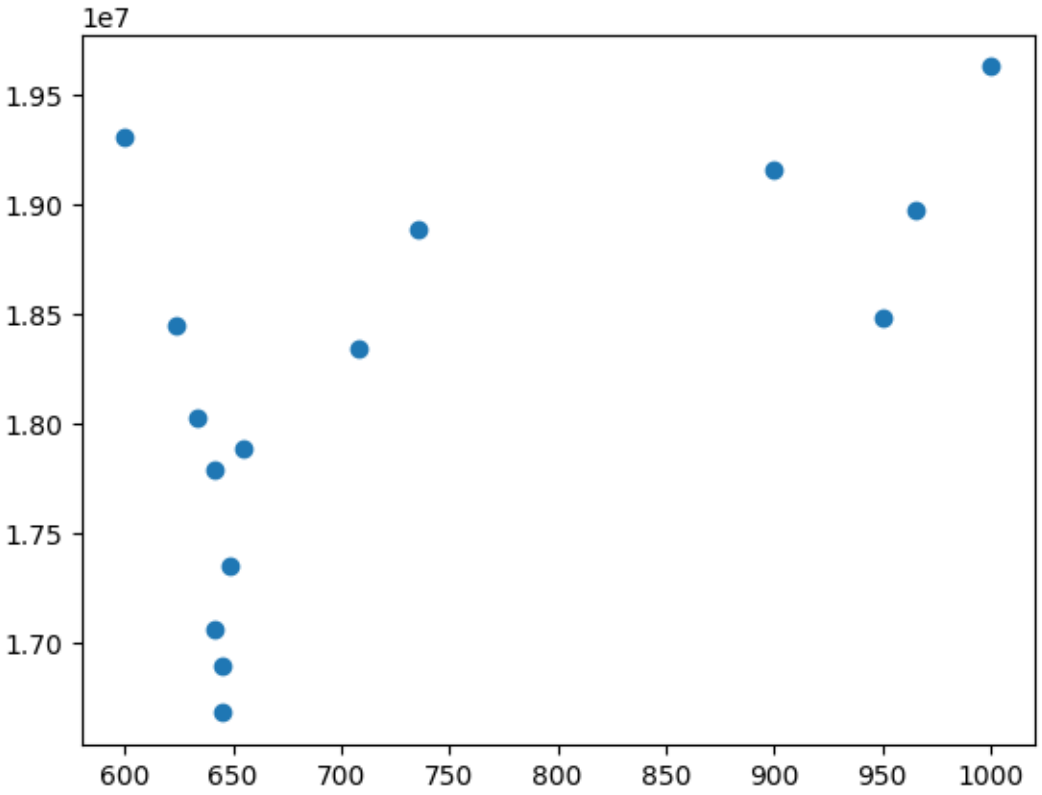
[ ]:
```
merged = pd.merge(subset, SMI, how="left", on=["year"])
```

[ ]:
```
merged
```

[ ]:
```
   Region  Asalariados  Percepciones por persona      Salarios  \
0   Total     19310627                      1.34  366818775121
1   Total     18451827                      1.26  352145124910
2   Total     18024554                      1.25  344505602339
3   Total     17788121                      1.27  339789084815
4   Total     17063142                      1.25  317397482413
5   Total     16682061                      1.26  308695720646
6   Total     16899024                      1.30  311279195843
7   Total     17349558                      1.34  323487828331
8   Total     17888520                      1.36  336938113674
9   Total     18343199                      1.39  351677788541
10  Total     18889515                      1.46  374174646222
11  Total     19157767                      1.41  394004653900
12  Total     18484793                      1.32  378988970687
13  Total     18974247                      1.37  408307759112
14  Total     19628877                      1.40  447160135525

    Salario Medio Anual  year    smi  Unnamed: 0  smi_14
0                 18996  2008  Total         NaN   600.0
1                 19085  2009  Total         NaN   624.0
2                 19113  2010  Total         NaN   633.3
3                 19102  2011  Total         NaN   641.4
4                 18601  2012  Total         NaN   641.4
5                 18505  2013  Total         NaN   645.3
6                 18420  2014  Total         NaN   645.3
7                 18645  2015  Total         NaN   648.6
8                 18835  2016  Total         NaN   655.2
9                 19172  2017  Total         NaN   707.7
10                19809  2018  Total         NaN   735.9
11                20566  2019  Total         NaN   900.0
12                20503  2020  Total         NaN   950.0
13                21519  2021  Total         NaN   965.0
14                22781  2022  Total         NaN  1000.0
```

[ ]:
```
plt.scatter(merged['smi_14'], merged['Asalariados'])
```

[ ]:
```
<matplotlib.collections.PathCollection at 0x12f294690>
```



[ ]:
```
merged[["smi_14", "Asalariados"]].corr()
```

[ ]:
```
               smi_14  Asalariados
smi_14       1.000000     0.605428
Asalariados  0.605428     1.000000
```

[ ]:
```
merged
```

[ ]:
```
    Region  Asalariados  Percepciones por persona      Salarios  \
0   Total     19310627                      1.34  366818775121
1   Total     18451827                      1.26  352145124910
2   Total     18024554                      1.25  344505602339
3   Total     17788121                      1.27  339789084815
4   Total     17063142                      1.25  317397482413
5   Total     16682061                      1.26  308695720646
6   Total     16899024                      1.30  311279195843
7   Total     17349558                      1.34  323487828331
8   Total     17888520                      1.36  336938113674
9   Total     18343199                      1.39  351677788541
10  Total     18889515                      1.46  374174646222
11  Total     19157767                      1.41  394004653900
```

```
12  Total     18484793                    1.32  378988970687
13  Total     18974247                    1.37  408307759112
14  Total     19628877                    1.40  447160135525

    Salario Medio Anual  year    smi  Unnamed: 0  smi_14
0                 18996  2008  Total         NaN   600.0
1                 19085  2009  Total         NaN   624.0
2                 19113  2010  Total         NaN   633.3
3                 19102  2011  Total         NaN   641.4
4                 18601  2012  Total         NaN   641.4
5                 18505  2013  Total         NaN   645.3
6                 18420  2014  Total         NaN   645.3
7                 18645  2015  Total         NaN   648.6
8                 18835  2016  Total         NaN   655.2
9                 19172  2017  Total         NaN   707.7
10                19809  2018  Total         NaN   735.9
11                20566  2019  Total         NaN   900.0
12                20503  2020  Total         NaN   950.0
13                21519  2021  Total         NaN   965.0
14                22781  2022  Total         NaN  1000.0
```

[ ]:
```
rename_dict = {'Region': 'ccaa', 'Percepciones por personas': 'percepciones',␣
 ↪'Asalariados': 'asalariados', 'Salario Medio Anual': 'sma', 'Salarios':
 ↪'salarios', 'year': 'periodo'}
merged.rename(columns=rename_dict, inplace=True)
```

[ ]:
```
merged
```

[ ]:
```
     ccaa  asalariados  Percepciones por persona      salarios    sma  \
0   Total     19310627                      1.34  366818775121  18996
1   Total     18451827                      1.26  352145124910  19085
2   Total     18024554                      1.25  344505602339  19113
3   Total     17788121                      1.27  339789084815  19102
4   Total     17063142                      1.25  317397482413  18601
5   Total     16682061                      1.26  308695720646  18505
6   Total     16899024                      1.30  311279195843  18420
7   Total     17349558                      1.34  323487828331  18645
8   Total     17888520                      1.36  336938113674  18835
9   Total     18343199                      1.39  351677788541  19172
10  Total     18889515                      1.46  374174646222  19809
11  Total     19157767                      1.41  394004653900  20566
12  Total     18484793                      1.32  378988970687  20503
13  Total     18974247                      1.37  408307759112  21519
14  Total     19628877                      1.40  447160135525  22781

    periodo    smi  Unnamed: 0  smi_14
0      2008  Total         NaN   600.0
```

```
1      2009  Total         NaN   624.0
2      2010  Total         NaN   633.3
3      2011  Total         NaN   641.4
4      2012  Total         NaN   641.4
5      2013  Total         NaN   645.3
6      2014  Total         NaN   645.3
7      2015  Total         NaN   648.6
8      2016  Total         NaN   655.2
9      2017  Total         NaN   707.7
10     2018  Total         NaN   735.9
11     2019  Total         NaN   900.0
12     2020  Total         NaN   950.0
13     2021  Total         NaN   965.0
14     2022  Total         NaN  1000.0
```

[ ]:
```
merged.ccaa.unique()
```

[ ]:
```
array(['Total'], dtype=object)
```

[ ]:
```

```