In [1]:
```
import pandas
```

Introductie

In dit notebook wordt data uit twee csv files samengevoegd en dit resulteert in de file: gemeentedata_merged.csv
Het gaat om:

- gegevens_gemeenten_2021.csv
  Bevat gegevens m.b.t. o.a. woningprijs, bevolkingssamenstelling, inkomen, werkloosheid, criminaliteit per gemeente in Nederland
- Uitslag_alle_gemeenten_TK20210317.csv
  Bevat gegevens m.b.t. de verkiezingsuitslag van de Tweede Kamer verkiezingen van 2021 per gemeente in Nederland

Importeren data

In [2]:
```
data_diverse = pandas.read_csv('./data/gegevens_gemeenten_2021.csv', sep=
election_data = pandas.read_csv('./data/Uitslag_alle_gemeenten_TK20210317
```

In [3]:
```
data_diverse = data_diverse.sort_values('Gemeenten')
election_data = election_data.sort_values('RegioNaam')
```

In [4]:
```
len(data_diverse)
```

Out[4]:
```
344
```

In [5]:
```
len(election_data)
```

Out[5]:
```
355
```

In [6]:
```
data_diverse.head(5)
```

Out[6]:

	Gemeenten	Vraagprijs aangeboden woningen|2021	Bevolking totaal|2021	Leeftijd bevolking|2021	Bevolkingsdichtheid|2021	Mig
252	's-Hertogenbosch	488020	155490	42	1414	
0	Aa en Hunze	469107	25399	47,3	92	
1	Aalsmeer	658553	31991	42,2	1590	
2	Aalten	347721	27120	44,5	281	
3	Achtkarspelen	357519	27900	42,3	273	

In [7]:
```
election_data.head(5)
```

Out[7]:

	RegioNaam	RegioCode	AmsterdamseCode	OuderRegioNaam	OuderRegioCode	Kies
258	's-Gravenhage	G0518	11434	's-Gravenhage	K12	
259	's-Hertogenbosch	G0796	10054	's-Hertogenbosch	K18	
0	Aa en Hunze	G1680	10787	Assen	K3	
1	Aalsmeer	G0358	11264	Haarlem	K10	
2	Aalten	G0197	11046	Arnhem	K7	

5 rows × 47 columns

Vergelijking gemeentenamen tussen de twee data sheets

In [8]:
```
result_inner_join = pandas.merge(data_diverse.iloc[:, [0]], election_data
                                 how='inner',
                                 left_on='Gemeenten', right_on='RegioNaam
```

In [9]:
```
len(result_inner_join)
```

Out[9]: 338

We zien dat 338 namen hetzelfde zijn.

In [10]:
```
result_left_join = pandas.merge(data_diverse.iloc[:, [0]], election_data.
                                how='left',
                                left_on='Gemeenten', right_on='RegioNaam'
```

In [11]:
```
result_left_join[(result_left_join.RegioNaam != result_left_join.Gemeente
```

Out[11]:

	Gemeenten	RegioNaam
30	Bergen (L.)	NaN
31	Bergen (NH.)	NaN
67	Den Haag	NaN
72	Dijk en Waard	NaN
157	Land van Cuijk	NaN
177	Maashorst	NaN

- We zien dat 6 rijen in data_diverse een gemeentenaam hebben die niet gematched kan worden op een gemeentenaam in election_data

In [12]:
```
result_right_join = pandas.merge(data_diverse.iloc[:, [0]], election_data
                    how='right',
                    left_on='Gemeenten', right_on='RegioNaam')
```

In [13]:
```
result_right_join[(result_right_join.Gemeenten != result_right_join.Regio
```

Out[13]:

	Gemeenten	RegioNaam
0	NaN	's-Gravenhage
28	NaN	Beemster
32	NaN	Bergen
33	NaN	Bergen
45	NaN	Bonaire
49	NaN	Boxmeer
63	NaN	Cuijk
112	NaN	Grave
128	NaN	Heerhugowaard
162	NaN	Landerd
165	NaN	Langedijk
193	NaN	Mill en Sint Hubert
254	NaN	Saba
261	NaN	Sint Anthonis
262	NaN	Sint Eustatius
290	NaN	Uden
322	NaN	Weesp

- We zien dat 17 rijen in election_data een gemeentenaam hebben die niet gematched kan worden op een gemeentenaam in data_diverse

Verklaring van de waargenomen verschillen

- Sommige namen konden (net) niet gematched worden, wegens een (klein) verschil in naamgeving

Bergen (L.)
Bergen

Bergen (NH.)
Bergen

Den Haag
's-Gravenhage

- Bijzondere gemeenten

Bonaire
Saba
Sint Eustatius

- Opgeheven gemeenten die werden ondergebracht bij een andere gemeente

Beemster (gefuseerd met Purmerend)
Weesp (gefuseerd met Amsterdam)

- Samengevoegde gemeenten

Dijk en Waard
Heerhugowaard
Langedijk

Land van Cuijk
Boxmeer
Cuijk
Grave
Mill en Sint Hubert
Sint Anthonis

Maashorst
Landerd
Uden

- Wat betreft de gemeentelijke herindeling:

  De gegevens m.b.t. 2021 werden gedownload in 2022, en voor de gegevens die gedownload werden van de website www.waarstaatjegemeente.nl/jive geldt dat standaard uitgegaan wordt van de actuele gemeentelijke indeling, ook voor gegevens uit voorgaande jaren.
  De gemeentelijke herindeling na de verkiezingen van 2021 verklaart de bovengenoemde verschillen.

  Op: https://www.cbs.nl/nl-nl/onze-diensten/methoden/classificaties/overig/gemeentelijke-indelingen-per-jaar lezen we: 'Per 1 januari 2022 is het aantal gemeenten door gemeentelijke herindelingen afgenomen tot 345. En per 24 maart 2022 tot 344 gemeenten.'

  Om e.e.a. te corrigeren dienen we de verkiezingsdata te converteren, zodat ook hier wordt uitgegaan van de gemeentelijke indeling, zoals die gold begin oktober 2022.

Opschonen van de data

Deel I

Eerst schonen we de kolomnamen op:

- In beide sheets de kolom met gemeentenamen de kolomnaam 'Gemeente' geven

In [14]:
```
data_diverse.rename(columns={'Gemeenten': 'Gemeente'}, inplace=True)
election_data.rename(columns={'RegioNaam': 'Gemeente'}, inplace=True)
```

- in de sheet data_diverse diverse kolomnamen aanpassen

In [15]:
```
data_diverse.rename(columns={'Vraagprijs aangeboden woningen|2021': 'Gemi
                             'Bevolking totaal|2021': 'Bevolking totaal',
                             'Leeftijd bevolking|2021': 'Gemiddelde leeft
                             'Bevolkingsdichtheid|2021': 'Bevolkingsdicht
                             'Totaal misdrijven|2021': 'Misdrijven totaal
                             'Misdrijven - Diefstal/inbraak woning|2021':
                             'Werkloosheidspercentage|2021': 'Werklooshei
                             'Migratieachtergrond - Nederlandse achtergro
                             'Migratieachtergrond - Westers|2021': 'Alloc
                             'Migratieachtergrond - Niet-westers|2021': '
                             'Migratieachtergrond - Totaal|2021': 'Alloch
                             }, inplace=True)
```

- In de sheet election_data elke partijnaam vervangen door een bondigere variant

In [16]:
```
election_data.rename(columns={'PVV (Partij voor de Vrijheid)': 'PVV',
                              'SP (Socialistische Partij)': 'SP',
                              'Partij van de Arbeid (P.v.d.A.)': 'PvdA',
                              'Forum voor Democratie': 'FvD',
                              'Partij voor de Dieren': 'PvdD',
                              'Staatkundig Gereformeerde Partij (SGP)': '
                              'Trots op Nederland (TROTS)': 'TROTS',
                              'Blanco (Zeven, A.J.L.B.)': 'Blanco',
                              'LP (Libertaire Partij)': 'LP',
                              'DE FEESTPARTIJ (DFP)': 'DFP',
                              'Partij van de Eenheid': 'PvdE',
                              'Partij voor de Republiek': 'PvdR',
                             }, inplace=True)
```

- Enkele correcties m.b.t. het format van waarden: een komma vervangen door een punt

In [17]:
```
for colname in ['Gemiddelde leeftijd bevolking',
                'Werkloosheidspercentage',
                ]:

    data_diverse[colname] = data_diverse[colname].apply(lambda value: str
```

- In geval van 'Besteedbaar inkomen per huishouden|2019' ligt het iets complexer, omdat we hier ook met missing data te maken hebben.
  - Sommige cellen hebben een waarde '?'
  - We converteren deze eerst naar '0', daarna converteren we alle waarden in de gehele kolom naar float
  - Vervolgens berekenen we het gemiddelde en zetten we alle cellen, met waarde 0 op dit gemiddelde.

In [18]:
```
filter_qm = data_diverse['Besteedbaar inkomen per huishouden|2019'].isin(
```

In [19]:
```
data_diverse[filter_qm]
```

Out[19]:

	Gemeente	Gemiddelde vraagprijs aangeboden woningen	Bevolking totaal	Gemiddelde leeftijd bevolking	Bevolkingsdichtheid	Autochtonen totaal
12	Ameland	554807	3746	44.2	63	3439
234	Renswoude	518110	5556	37.8	302	5092
245	Rozendaal	925931	1726	45.1	62	1458
250	Schiermonnikoog	520548	931	49.9	23	830
271	Terschelling	493591	4870	45.1	57	4500
298	Vlieland	385608	1194	43.8	30	997

In [20]:
```
data_diverse['Besteedbaar inkomen per huishouden|2019'] = data_diverse['B
data_diverse['Besteedbaar inkomen per huishouden|2019'] = data_diverse['B
data_diverse['Besteedbaar inkomen per huishouden|2019'] = data_diverse['B
```

In [21]:
```
mean = round(data_diverse['Besteedbaar inkomen per huishouden|2019'].mean
```

In [22]:
```
data_diverse['Besteedbaar inkomen per huishouden|2019'] = data_diverse['B
```

- Nacontrole

In [23]:
```
filter_mean = data_diverse['Besteedbaar inkomen per huishouden|2019'].isi
```

In [24]:
```
data_diverse[filter_mean]
```

Out[24]:

	Gemeente	Gemiddelde vraagprijs aangeboden woningen	Bevolking totaal	Gemiddelde leeftijd bevolking	Bevolkingsdichtheid	Autochtonen totaal
12	Ameland	554807	3746	44.2	63	3439
45	Boxtel	515983	32973	44.0	477	27444
126	Hellendoorn	354764	35932	43.7	261	33366
231	Raalte	351904	37911	44.2	222	35287
234	Renswoude	518110	5556	37.8	302	5092
245	Rozendaal	925931	1726	45.1	62	1458
250	Schiermonnikoog	520548	931	49.9	23	830
271	Terschelling	493591	4870	45.1	57	4500
298	Vlieland	385608	1194	43.8	30	997

- Alle overige ontbrekende waarden geven we een waarde 0

In [25]:
```
election_data = election_data.fillna(0)
data_diverse = data_diverse.fillna(0)
```

- Het betreft de inbraak gegevens van Schiermonnikoog
- Verder alle cellen in de verkiezingsdata, waar verzuimd werd een 0 in te voeren (0 stemmen)

Deel II

- Een aantal gemeenten moeten qua naam gelijk getrokken worden

In [26]:
```
for name, regiocode, new_name in [
                ['\'s-Gravenhage', 'G0518', 'Den Haag'],
                ['Bergen', 'G0373', 'Bergen (NH.)'],
                ['Bergen', 'G0893', 'Bergen (L.)'],
                      ]:

    filter1 = election_data['Gemeente'].isin([name])
    filter2 = election_data['RegioCode'].isin([regiocode])
    index = election_data[filter1][filter2].index
    election_data.loc[index, 'Gemeente'] = new_name
```

```
<ipython-input-26-5a47c70f2024>:9: UserWarning: Boolean Series key will b
e reindexed to match DataFrame index.
  index = election_data[filter1][filter2].index
```

- De bijzondere gemeenten komen te vervallen in election_data

In [27]:
```
election_data = election_data.drop(election_data[(election_data.Gemeente.
```

- Nu het probleem oplossen met de samengevoegde gemeenten:
  In election_data dienen
  - de nieuwe gemeentenamen toegevoegd te worden
  - de uitgefaseerde gemeentenamen verwijderd te worden
  - de aantallen van de uitgefaseerde gemeenten toegevoegd te worden aan de gemeente waarin ze werden opgenomen

In [28]:
```
conversion_table = {
    'Dijk en Waard': ['Heerhugowaard', 'Langedijk'],
    'Land van Cuijk': ['Boxmeer', 'Cuijk', 'Grave', 'Mill en Sint Hubert'
    'Maashorst': ['Landerd', 'Uden'],
    'Purmerend': ['Purmerend', 'Beemster'],
    'Amsterdam': ['Amsterdam', 'Weesp'],
}
```

In [29]:
```
def find_new_municipalities(conversion_table):
    return [municipality for municipality in conversion_table if municipa


def find_updated_municipalities(conversion_table):
    return [municipality for municipality in conversion_table if municipa


def find_phased_out_municipalities(conversion_table, updated_municipaliti

    phased_out_municipalities = []

    for municipality_ in conversion_table:
        for municipality in conversion_table[municipality_]:
            if municipality not in updated_municipalities:
                phased_out_municipalities.append(municipality)

    return phased_out_municipalities


def add_new_municipalities(new_municipalities, conversion_table, election

    for new_municipality in new_municipalities:
        phased_out_municipalities = conversion_table[new_municipality]

        municipality_to_add = election_data[(election_data.Gemeente.isin(
        municipality_to_add.Gemeente = new_municipality
        election_data = election_data.append(municipality_to_add, ignore_

        indices_rows2drop = election_data[(election_data.Gemeente.isin(ph
        election_data = election_data.drop(indices_rows2drop)

    return election_data


def update_municipalities(updated_municipalities, conversion_table, elect

    for updated_municipality in updated_municipalities:

        municipalities2merge = conversion_table[updated_municipality]

        municipality_to_add = election_data[(election_data.Gemeente.isin(
        election_data = election_data.drop(election_data[(election_data.G
        municipality_to_add.Gemeente = updated_municipality
        election_data = election_data.append(municipality_to_add, ignore_

    return election_data


new_municipalities = find_new_municipalities(conversion_table)
updated_municipalities = find_updated_municipalities(conversion_table)
phased_out_municipalities = find_phased_out_municipalities(conversion_tab

election_data = add_new_municipalities(new_municipalities, conversion_tab
election_data = update_municipalities(updated_municipalities, conversion_
```

Nacontrole

In [30]:
```
len(data_diverse)
```

Out[30]: 344

In [31]:
```
len(election_data)
```

Out[31]: 344

In [32]:
```
result_inner_join = pandas.merge(data_diverse.iloc[:, [0]], election_data
                                 how='inner',
                                 left_on='Gemeente', right_on='Gemeente')
```

In [33]:
```
len(result_inner_join)
```

Out[33]: 344

Samenvoegen data

In [34]:
```
gemeentedata = pandas.merge(data_diverse, election_data, on='Gemeente')
```

In [35]:
```
gemeentedata.head(4)
```

Out[35]:

	Gemeente	Gemiddelde vraagprijs aangeboden woningen	Bevolking totaal	Gemiddelde leeftijd bevolking	Bevolkingsdichtheid	Autochtonen totaal	Al
0	's-Hertogenbosch	488020	155490	42.0	1414	122514	
1	Aa en Hunze	469107	25399	47.3	92	23719	
2	Aalsmeer	658553	31991	42.2	1590	25117	
3	Aalten	347721	27120	44.5	281	23997	

4 rows × 59 columns

In [36]:
```
gemeentedata.columns
```

Out[36]:
```
Index(['Gemeente', 'Gemiddelde vraagprijs aangeboden woningen',
       'Bevolking totaal', 'Gemiddelde leeftijd bevolking',
       'Bevolkingsdichtheid', 'Autochtonen totaal', 'Allochtonen Westers
',
       'Allochtonen Niet Westers', 'Allochtonen totaal', 'Misdrijven tota
al',
       'Diefstal/inbraak', 'Werkloosheidspercentage',
       'Besteedbaar inkomen per huishouden|2019', 'RegioCode',
       'AmsterdamseCode', 'OuderRegioNaam', 'OuderRegioCode',
       'Kiesgerechtigden', 'Opkomst', 'OngeldigeStemmen', 'BlancoStemmen
',
       'GeldigeStemmen', 'VVD', 'D66', 'PVV', 'CDA', 'SP', 'PvdA',
       'GROENLINKS', 'FvD', 'PvdD', 'ChristenUnie', 'Volt', 'JA21', 'SGP
',
       'DENK', '50PLUS', 'BBB', 'BIJ1', 'CODE ORANJE', 'NIDA', 'Splinter
',
       'Piratenpartij', 'JONG', 'TROTS', 'Lijst Henk Krol', 'NLBeter',
       'Blanco', 'LP', 'OPRECHT', 'JEZUS LEEFT', 'DFP',
       'U-Buntu Connected Front', 'Vrij en Sociaal Nederland', 'PvdE',
       'Wij zijn Nederland', 'PvdR', 'Modern Nederland', 'De Groenen'],
      dtype='object')
```

Export to csv

In [37]:
```
gemeentedata.to_csv('./data/gemeentedata_merged.csv', index=False, sep=';
```