notebook_5_medicare_opioid_and_nonopioid_prescriber_cleanup_and

October 20, 2019

1 Medicare opioid and non-opioid prescriber datasets cleanup and town join

1.0.1 Goals:

- Clean up the Medicare Part D datasets (both the opioid prescriber and general prescription datasets)
- Merge these datasets with the town names in the opioid overdose deaths dataset

1.0.2 Outputs:

Inidividual prescribers by zip code, associated with town from opioid overdose death count dataset:

Local paths:

- 2013 dataset: `data/tidy_data/medicare_partD_opioid_prescriber_2013_w_zip_MAtown_v1.csv`
- 2014 dataset: `data/tidy_data/medicare_partD_opioid_prescriber_2014_w_zip_MAtown_v1.csv`
- 2015 dataset: `data/tidy_data/medicare_partD_opioid_prescriber_2015_w_zip_MAtown_v1.csv`
- 2016 dataset: `data/tidy_data/medicare_partD_opioid_prescriber_2016_w_zip_MAtown_v1.csv`
- 2017 dataset: `data/tidy_data/medicare_partD_opioid_prescriber_2017_w_zip_MAtown_v1.csv`

Benzodiazepine prescription data years 2013-2017, with each prescriber associated with MA opioid overdose death town

- Local path: `data/tidy_data/med_partD_benzo_indiv_pres_w_town_merge_13_to_17.csv`

Summarized benzodiazepine prescription data years 2013-2017, grouped by town (from opioid overdose death dataset), year, and drug (out of the 3 benzo drugs in the dataset, by geneeric name)

- Local path: `data/tidy_data/med_partD_benzo_sum_w_town_merge_13_to_17.csv`
- pdf report (in case notebook doesn't run): products/notebook_5_medicare_opioid_and_nonopioid_prescrib

[1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

```
import seaborn as sns
import geopandas as gpd
sns.set_style('darkgrid')
sns.set(font_scale=1.5)
```

[2]:
```
# cms opioid prescriber data
opi_pres_13_raw = pd.read_csv("../../data/raw_data/
 ↪medicare_prescription_opioids/
 ↪Medicare_Part_D_Opioid_Prescriber_Summary_File_2013.csv")
opi_pres_14_raw = pd.read_csv("../../data/raw_data/
 ↪medicare_prescription_opioids/
 ↪Medicare_Part_D_Opioid_Prescriber_Summary_File_2014.csv")
opi_pres_15_raw = pd.read_csv("../../data/raw_data/
 ↪medicare_prescription_opioids/
 ↪Medicare_Part_D_Opioid_Prescriber_Summary_File_2015.csv")
opi_pres_16_raw = pd.read_csv("../../data/raw_data/
 ↪medicare_prescription_opioids/
 ↪Medicare_Part_D_Opioid_Prescriber_Summary_File_2016.csv")
opi_pres_17_raw = pd.read_csv("../../data/raw_data/
 ↪medicare_prescription_opioids/
 ↪Medicare_Part_D_Opioid_Prescriber_Summary_File_2017.csv")
```

[3]:
```
# cms non-opioid prescriber data (big file - takes a while)
all_pres_17_raw = pd.read_csv("../../data/raw_data/
 ↪medicare_prescription_all_drugs/PartD_Prescriber_PUF_NPI_Drug_17.txt",␣
 ↪sep='\t')
```

[4]:
```
# zipcode - town association attempts
## zipcode - town lists (copy paste from websites)
zip_town_raw = pd.read_csv('../../data/raw_data/
 ↪shapefiles_and_geography_related/ma_town_zipcode_list.txt', sep="(",␣
 ↪header=None)
zip_town_raw_alt = pd.read_csv('../../data/raw_data/
 ↪shapefiles_and_geography_related/ma_town_zipcode_list_alt.txt', sep='\t')
## shapfile of ma postal zipcodes
ma_postzip_map = gpd.read_file('../../data/raw_data/
 ↪shapefiles_and_geography_related/zipcodes_nt/ZIPCODES_NT_POLY.shp')
# compare zipcodes and town names to the overdose deaths dataset (351 towns)
ma_over_death = pd.read_csv("../../data/tidy_data/
 ↪overdose_death_count_acs_merge.csv")
```

[5]:
```
# easier column names for opi prescriber dfs
lwr_column_names =[x.lower().replace(' ', '_') for x in opi_pres_13_raw.columns]
raw_opi_dflist = [opi_pres_13_raw, opi_pres_14_raw, opi_pres_15_raw,␣
 ↪opi_pres_16_raw, opi_pres_17_raw]
for x in raw_opi_dflist:
    x.columns = lwr_column_names
```

[6]:
```
# number of rows in each file:
for x in raw_opi_dflist:
    print(x.shape)
```

```
(1049299, 11)
(1072978, 11)
(1102253, 11)
(1131550, 11)
(1162898, 11)
```

[7]:
```
# what the columns look like:
opi_pres_17_raw.head()
```

[7]:

	npi	nppes_provider_last_name	nppes_provider_first_name	nppes_provider_zip_code	nppes_provider_state	specialty_description	total_claim_count	opioid_claim_count	opioid_prescribing_rate	long-acting_opioid_claim_count	long-acting_opioid_prescribing_rate
0	1003000126	ENKESHAFI	ARDALAN	21502.0	MD	Internal Medicine	677	25.0	3.69	NaN	NaN
1	1003000142	KHALIL	RASHID	43623.0	OH	Anesthesiology	1946	1040.0	53.44	200.0	19.23
2	1003000167	ESCOBAR	JULIO	89403.0	NV	Dentist	55	14.0	25.45	0.0	0.00
3	1003000175	REYES-VASQUEZ	BELINDA	91744.0	CA	Dentist	18	0.0	0.00	0.0	NaN
4	1003000282	BLAKEMORE	ROSIE	37243.0	TN	Nurse Practitioner	90	0.0	0.00	0.0	NaN

[8]:
```
all_pres_17_raw.columns
```

[8]:
```
Index(['npi', 'nppes_provider_last_org_name', 'nppes_provider_first_name',
       'nppes_provider_city', 'nppes_provider_state', 'specialty_description',
       'description_flag', 'drug_name', 'generic_name', 'bene_count',
       'total_claim_count', 'total_30_day_fill_count', 'total_day_supply',
       'total_drug_cost', 'bene_count_ge65', 'bene_count_ge65_suppress_flag',
       'total_claim_count_ge65', 'ge65_suppress_flag',
```

```
       'total_30_day_fill_count_ge65', 'total_day_supply_ge65',
       'total_drug_cost_ge65'],
      dtype='object')
```

The opioid prescriber dataset comes with a zip code, but the dataset for all drug comes only with a town name.

Questions to address:

- do the NPI's (unique provider IDs - I think) match between the 2 prescriber datasets?
- do the town names in the all-drugs prescriber dataset match the town names in the opioid overdose data?

Potential merge options: * merge 1: opioid overdose death counts + all-drug prescribers on town -> merge 2: merge to opioid prescribers by npi * merge 1: opioid prescribers to town (using postal map) to associate zipcodes with town -> merge 2: merge to opioid overdose deaths -> use merge 1 to associate all-drug prescribers with opioid overdose death town based on npi

[9]:
```
opi_pres_17_MA = opi_pres_17_raw[opi_pres_17_raw['nppes_provider_state'] ==␣
 ↪'MA'].copy()
all_pres_17_MA = all_pres_17_raw[all_pres_17_raw['nppes_provider_state'] ==␣
 ↪'MA'].copy()
# all-drug prescriber dataset is very large, focus on the prescribers/providers␣
 ↪only for now
all_pres_17_MA_prov = all_pres_17_MA.iloc[:, 0:6].copy().drop_duplicates()
```

[10]:
```
display(opi_pres_17_MA.head())
display(all_pres_17_MA_prov.head())
```

```
           npi nppes_provider_last_name nppes_provider_first_name  \
36   1003002312                 HOPKINS                  PATRICIA
118  1003007477                   ABDOW                  KIMBERLY
132  1003008244              RAJBHANDARI                      RUMA
189  1003011610                     RAY                     ALAKA
224  1003012766                    KANO                   ZACHARY

     nppes_provider_zip_code nppes_provider_state specialty_description  \
36                    2169.0                   MA          Rheumatology
118                   1609.0                   MA    Nurse Practitioner
132                   2115.0                   MA      Gastroenterology
189                   2114.0                   MA     Internal Medicine
224                   2445.0                   MA               Dentist

     total_claim_count  opioid_claim_count  opioid_prescribing_rate  \
36                4487               513.0                    11.43
118               5314                 0.0                     0.00
132                 56                 0.0                     0.00
189               1993                62.0                     3.11
224                 19                 0.0                     0.00
```

```
     long-acting_opioid_claim_count  long-acting_opioid_prescribing_rate
36                             84.0                                16.37
118                             0.0                                  NaN
132                             0.0                                  NaN
189                             NaN                                  NaN
224                             0.0                                  NaN


            npi nppes_provider_last_org_name nppes_provider_first_name  \
757   1003002312                     HOPKINS                   PATRICIA
2478  1003007477                       ABDOW                   KIMBERLY
2997  1003008244                  RAJBHANDARI                       RUMA
4569  1003011610                         RAY                     ALAKA
5158  1003012766                        KANO                    ZACHARY

     nppes_provider_city nppes_provider_state specialty_description
757              QUINCY                   MA          Rheumatology
2478          WORCESTER                   MA    Nurse Practitioner
2997             BOSTON                   MA      Gastroenterology
4569             BOSTON                   MA     Internal Medicine
5158          BROOKLINE                   MA               Dentist
```

[11]:
```
# NPI sanity check
print(len(set(opi_pres_17_MA['npi']) - set(all_pres_17_MA['npi'])))
print(len(set(all_pres_17_MA['npi']) - set(opi_pres_17_MA['npi'])))
print(len(set(opi_pres_17_MA['npi']) - set(all_pres_17_raw['npi'])))
print(len(set(all_pres_17_MA['npi'])))
print(len(set(opi_pres_17_MA['npi'])))
```

```
8430
0
8430
27300
35730
```

Interesting - There are 8,430 more unique prescribers in the opioid prescription dataset than in the all-drugs prescriber dataset

Not sure why that would be the case.

Maybe a difference in specialties?

[12]:
```
print(opi_pres_17_MA.shape)
print(all_pres_17_MA_prov.shape)
opi_pres_MA_miss = opi_pres_17_MA[~ opi_pres_17_MA['npi'].
 ↪isin(set(all_pres_17_MA['npi']))].copy()
display(opi_pres_MA_miss['specialty_description'].value_counts())
display(all_pres_17_MA['specialty_description'].value_counts())
```

```
(35730, 11)
(27300, 6)
```

```
Internal Medicine                                                  1355
Nurse Practitioner                                                 1011
Dentist                                                             890
Student in an Organized Health Care Education/Training Program      890
Physician Assistant                                                 550
                                                                   ...
Sleep Medicine                                                        1
Assistant, Podiatric                                                  1
Physical Therapist in Private Practice                                1
Community/Behavioral Health                                           1
Social Worker                                                         1
Name: specialty_description, Length: 97, dtype: int64
```

```
Internal Medicine                                   242242
Nurse Practitioner                                   85862
Family Practice                                      83128
Psychiatry                                           26665
Physician Assistant                                  26472
                                                     ...
Hospital                                                 1
Assistant, Podiatric                                     1
Registered Dietitian or Nutrition Professional           1
Thoracic Surgery (Cardiothoracic Vascular Surgery)       1
Midwife                                                  1
Name: specialty_description, Length: 96, dtype: int64
```

Top specialties for opioid prescribers missing from the all-drugs prescriber dataset are Internal Medicine and Nurse Practitioner, but those are also present in the all-drugs prescriber dataset. Maybe specialty is not the reason for difference?

The difference may be in how Medicare created these datasets - maybe different benefit plans? I tried to read the methodology on the website, but it wasn't clear to me.

Moving on

Q: are the towns in the all-drugs prescriber database a match for the names in the opioid overdose death count dataset?

[13]:
```
# number of towns in the opioid overdose death data missing in the all-drug␣
 ↪prescriber dataset:
print(len(set(ma_over_death['city_death']) -␣
 ↪set(all_pres_17_MA_prov['nppes_provider_city'].str.lower())))
# number of towns in the all-drugs prescriber dataset missing from the opioid␣
 ↪overdose death dataset
print(len(set(all_pres_17_MA_prov['nppes_provider_city'].str.lower()) -␣
 ↪set(ma_over_death['city_death'])))
```

```
85
194
```

Expected that this would be an issue - the towns don't match well - likely different definitions of town/cities

Have to go with this strategy:

merge 1: opioid prescribers to town (using postal map) to associate zipcodes with town -> merge 2: merge to opioid overdose deaths -> use merge 1 to associate all-drug prescribers with opioid overdose death town based on npi

1.0.3 Step 1: try to associate zip codes in the opioid prescriber dataset with town in the opioid overdose death count dataset

But first, zip codes were imported as floats and missing leading zero that is typicaly seen in a zip code - fix this

[14]:
```
opi_pres_17_MA['nppes_provider_zip_code'] = [x.zfill(5) for x in␣
 ↪list(opi_pres_17_MA['nppes_provider_zip_code'].astype(int).astype(str))]
opi_pres_17_MA.head()
```

[14]:

	npi	nppes_provider_last_name	nppes_provider_first_name	nppes_provider_zip_code	nppes_provider_state	specialty_description	total_claim_count	opioid_claim_count	opioid_prescribing_rate	long-acting_opioid_claim_count	long-acting_opioid_prescribing_rate
36	1003002312	HOPKINS	PATRICIA	02169	MA	Rheumatology	4487	513.0	11.43	84.0	16.37
118	1003007477	ABDOW	KIMBERLY	01609	MA	Nurse Practitioner	5314	0.0	0.00	0.0	NaN
132	1003008244	RAJBHANDARI	RUMA	02115	MA	Gastroenterology	56	0.0	0.00	0.0	NaN
189	1003011610	RAY	ALAKA	02114	MA	Internal Medicine	1993	62.0	3.11	NaN	NaN
224	1003012766	KANO	ZACHARY	02445	MA	Dentist	19	0.0	0.00	0.0	NaN

2 zip code - town association lists copied from websites - will either work?

[15]:
```
print(zip_town_raw.shape)
print(zip_town_raw_alt.shape)
print(zip_town_raw.columns)
print(zip_town_raw_alt.columns)
```

```
(1093, 2)
(682, 4)
Int64Index([0, 1], dtype='int64')
Index(['ZIP Code ', 'City ', 'County ', 'Type'], dtype='object')
```

[16]:
```
display(zip_town_raw.head())
display(zip_town_raw_alt.head())
```

	0	1
0	02351 Abington	781)
1	02018 Accord	781)
2	01718 Acton	978)
3	01719 Acton	978)
4	01720 Acton	978)

	ZIP Code	City	County	Type
0	ZIP Code 01001	Agawam	Hampden	Standard
1	ZIP Code 01002	Amherst	Hampshire	Standard
2	ZIP Code 01003	Amherst	Hampshire	Standard
3	ZIP Code 01004	Amherst	Hampshire	P.O. Box
4	ZIP Code 01005	Barre	Worcester	Standard

Both potentially need cleaning, if they're useful

[17]:
```
# pull out zip code to separate column
zip_town = zip_town_raw.iloc[:, [0]].copy()
zip_town.columns = ['col1']
zip_town['zip'] = [x[:5] for x in zip_town['col1']]
zip_town.head()
```

[17]:

	col1	zip
0	02351 Abington	02351
1	02018 Accord	02018
2	01718 Acton	01718
3	01719 Acton	01719
4	01720 Acton	01720

[18]:
```
# now pull out town name
zip_town['town'] = [x[6:] for x in zip_town['col1']]
# towns have white spaces at the end (carryover from formatting)
zip_town['town'] = zip_town['town'].str.strip().str.lower()
zip_town.drop('col1', axis=1, inplace=True)
```

```
zip_town.head()
```

[18]:
```
     zip      town
0  02351  abington
1  02018    accord
2  01718     acton
3  01719     acton
4  01720     acton
```

[19]:
```
ma_over_death.head()
```

[19]:
```
  city_death  2014  2015  2016  2017  2018    tot_pop_13    tot_pop_14  \
0   abington     0     6     1     3     5  16109.285714  16150.714286
1      acton     1     2     3     0     1  22580.142857  22798.857143
2   acushnet     0     4     2     4     0  10363.000000  10383.000000
3      adams     2     3     1     0     4   8367.571429   8328.428571
4     agawam     1     2     0     4     8  27684.428571  27705.571429

     tot_pop_15    tot_pop_16  tot_pop_17  over_65_count  over_65_prop  \
0  16192.142857  16233.571429       16275          2469      0.151705
1  23017.571429  23236.285714       23455          4001      0.170582
2  10403.000000  10423.000000       10443          2431      0.232788
3   8289.285714   8250.142857        8211          1764      0.214834
4  27726.714286  27747.857143       27769          6195      0.223090

   med_house_inc  mean_house_inc  less_than_hs_ed  at_or_below_pov_prop  \
0   87156.000000    98809.035505         5.405643              0.035754
1  139890.466667   156680.203867         2.456531              0.038315
2   69624.714286    80333.175842        18.297315              0.040828
3   48445.400000    60968.594660        11.862182              0.110854
4   65490.125000    79464.234446         7.748863              0.094819

   pop_struggling_prop urb_v_rur town_status
0             0.100408     rural       grown
1             0.041747     rural       grown
2             0.178406     rural       grown
3             0.144597     rural      shrunk
4             0.142656     rural       grown
```

[20]:
```
print(len(set(zip_town['town'])))
print(len(set(ma_over_death['city_death'])))
```

```
806
347
```

[21]:
```
np.unique(zip_town['town'])[:20]
```

[21]:
```
array(['abington', 'accord', 'acton', 'acushnet', 'adams',
       'aetna life & casualty co', 'agawam', 'alford', 'allendale',
       'allmerica', 'allston', 'amesbury', 'amherst', 'andover',
```

```
       'aquinnah', 'arlington', 'arlington heights', 'ashburnham',
       'ashby', 'ashfield'], dtype=object)
```

Town names are too granular, try the other one

[22]:
```
zip_town_raw_alt.columns = [x.lower().strip() for x in list(zip_town_raw_alt.
 ↪columns)]
```

[23]:
```
len(np.unique(zip_town_raw_alt['city']))
# still too many towns
```

[23]:
```
512
```

[24]:
```
# what are the zip code types?
zip_town_raw_alt['type'].value_counts()
```

[24]:
```
Standard    492
P.O. Box    156
Unique       34
Name: type, dtype: int64
```

[25]:
```
# try filtering down - maybe that will help?
zip_stand = zip_town_raw_alt[zip_town_raw_alt['type'] == 'Standard'].copy()
print(len(set(zip_stand['city'])))
# closer count - try matching what's missing
zip_stand['city'] = zip_stand['city'].str.lower()
zip_stand.head()
```

```
417
```

[25]:

	zip code	city	county	type
0	ZIP Code 01001	agawam	Hampden	Standard
1	ZIP Code 01002	amherst	Hampshire	Standard
2	ZIP Code 01003	amherst	Hampshire	Standard
4	ZIP Code 01005	barre	Worcester	Standard
5	ZIP Code 01007	belchertown	Hampshire	Standard

[26]:
```
# mismatches
print(len(set(ma_over_death['city_death']) - set(zip_stand['city'])))
print(len(set(zip_stand['city']) - set(ma_over_death['city_death'])))
# still a lot of mismatches - maybe there's a better source?
```

```
39
109
```

[27]:
```
ma_postzip_map.head()
```

[27]:

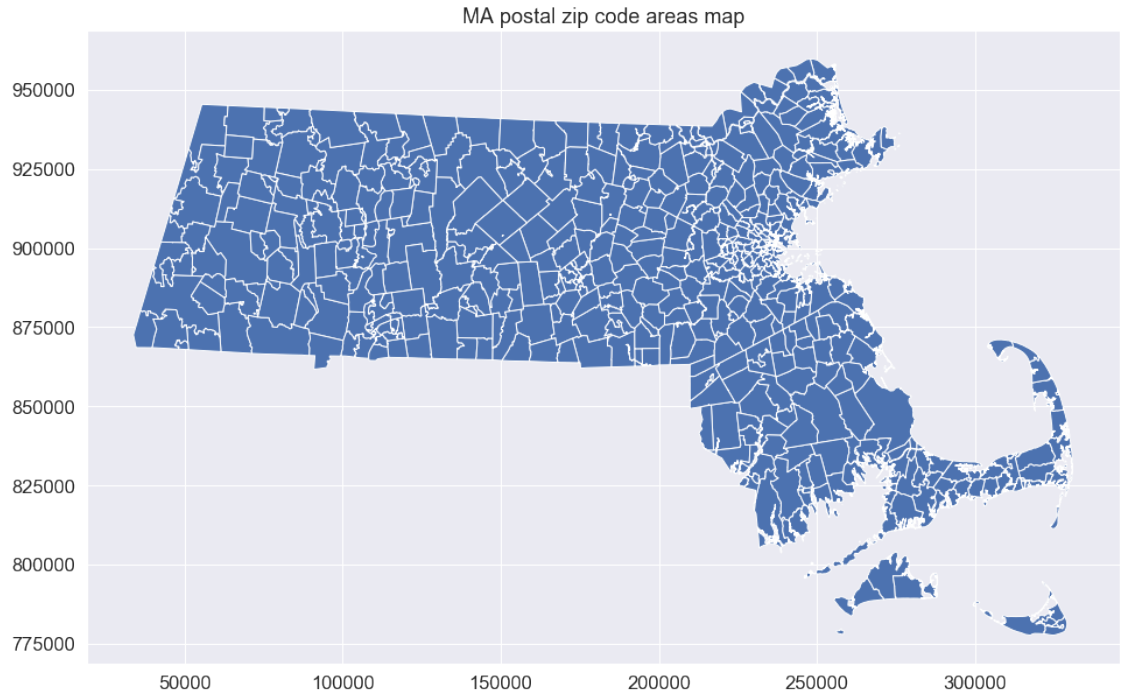
	POSTCODE	PC_NAME	PC_TYPE	PA_NAME	PA_FIPS	\
0	01331	ATHOL	NON UNIQUE	ATHOL	02515	
1	01085	WESTFIELD	NON UNIQUE	WESTFIELD	76030	
2	01370	SHELBURNE FALLS	NON UNIQUE	SHELBURNE FALLS	61205	
3	01235	HINSDALE	NON UNIQUE	HINSDALE	30280	

```
4    02747  NORTH DARTMOUTH  NON UNIQUE  NORTH DARTMOUTH   47450

            CITY_TOWN     COUNTY  AREA_SQMI    SHAPE_AREA     SHAPE_LEN  \
0      ATHOL, TOWN OF  WORCESTER  58.084870  1.504391e+08  66459.581259
1           WESTFIELD    HAMPDEN  55.938928  1.448812e+08  61329.577100
2  SHELBURNE, TOWN OF   FRANKLIN  48.804057  1.264019e+08  70885.011312
3   HINSDALE, TOWN OF  BERKSHIRE  47.757052  1.236902e+08  49286.404925
4  DARTMOUTH, TOWN OF    BRISTOL  47.495414  1.230126e+08  66614.835420

                                              geometry
0  POLYGON ((147583.6014999971 930766.1334999986,...
1  POLYGON ((101952.2437999994 882113.238499999, ...
2  POLYGON ((107090.9670000002 933358.561999999, ...
3  POLYGON ((77728.64249999821 914027.9114000015,...
4  POLYGON ((242894.990199998 830521.8995000012, ...
```

[28]:
```
ma_postzip_map.plot(figsize=(16,10))
plt.title('MA postal zip code areas map')
plt.show()
```



[29]:
```
print(len(set(ma_postzip_map['CITY_TOWN'])))
print(len(set(ma_postzip_map['PC_NAME'])))
```

```
335
458
```

[30]:
```
# fix the city_town column - this looks to be what I might be looking for
ma_postzip_map['CITY_TOWN'] = ma_postzip_map['CITY_TOWN'].str.lower().str.
 ↪replace(', town of', '')
```

[31]:
```
# mismatches - what's in the postal map that's not in the opioid overdose␣
 ↪dataset?:
set(ma_postzip_map['CITY_TOWN']) - set(ma_over_death['city_death'])
```

[31]:
```
{'charlemont',
 'manchester by the sea',
 'monroe',
 'mt washington',
 'worthington'}
```

[32]:
```
# what's in the opioid overdose dataset that's not in the postal map?
set(ma_over_death['city_death']) - set(ma_postzip_map['CITY_TOWN'])
```

[32]:
```
{'alford',
 'aquinnah',
 'clarksburg',
 'granville',
 'hawley',
 'leyden',
 'manchester',
 'montgomery',
 'mount washington',
 'new ashford',
 'pelham',
 'peru',
 'phillipston',
 'tisbury',
 'washington',
 'west tisbury',
 'westhampton'}
```

Much more manageable, few fixes

- manchester and manchester by the sea seem to be the same town (wikipedia)
- mt washington and mount washington - different spelling
- charlemont was one of the towns without a census block association (notebook 3)

[33]:
```
ma_over_death['city_death_match'] = ma_over_death['city_death'].str.
 ↪replace('manchester', 'manchester by the sea').str.replace('mount␣
 ↪washington', 'mt washington')
```

[34]:
```
# death counts for towns/cities I'm missing
ma_over_death[ma_over_death['city_death_match'].
 ↪isin(set(ma_over_death['city_death_match']) -␣
 ↪set(ma_postzip_map['CITY_TOWN'].str.lower().str.replace(', town of', '')))]
```

[34]:

	city_death	2014	2015	2016	2017	2018	tot_pop_13	tot_pop_14	\
5	alford	0	0	0	0	0	458.428571	446.571429	
9	aquinnah	0	0	0	0	0	452.000000	499.000000	
62	clarksburg	0	0	0	1	0	1710.571429	1713.428571	
109	granville	0	0	0	0	0	1606.285714	1619.714286	
126	hawley	0	0	0	0	0	930.857143	942.142857	
153	leyden	0	0	0	0	0	1809.571429	1805.428571	
190	montgomery	0	0	0	0	0	822.571429	817.428571	
196	new ashford	0	0	0	0	0	273.428571	288.571429	
226	pelham	0	0	0	0	0	1302.142857	1295.857143	
229	peru	0	0	0	1	0	831.571429	826.428571	
231	phillipston	0	0	0	0	0	1664.000000	1658.000000	
292	tisbury	0	2	1	1	2	4013.714286	4035.285714	
309	washington	0	0	0	0	0	521.285714	515.714286	
323	west tisbury	1	0	1	0	0	2601.571429	2555.428571	
327	westhampton	0	0	0	0	0	1697.857143	1728.142857	

	tot_pop_15	tot_pop_16	...	over_65_count	over_65_prop	\
5	434.714286	422.857143	...	219	0.532847	
9	546.000000	593.000000	...	66	0.103125	
62	1716.285714	1719.142857	...	494	0.286876	
109	1633.142857	1646.571429	...	302	0.181928	
126	953.428571	964.714286	...	267	0.273566	
153	1801.285714	1797.142857	...	443	0.247072	
190	812.285714	807.142857	...	148	0.184539	
196	303.714286	318.857143	...	81	0.242515	
226	1289.571429	1283.285714	...	343	0.268598	
229	821.285714	816.142857	...	115	0.141800	
231	1652.000000	1646.000000	...	239	0.145732	
292	4056.857143	4078.428571	...	654	0.159512	
309	510.142857	504.571429	...	158	0.316633	
323	2509.285714	2463.142857	...	746	0.308647	
327	1758.428571	1788.714286	...	336	0.184717	

	med_house_inc	mean_house_inc	less_than_hs_ed	at_or_below_pov_prop	\
5	105625.00	148519.270833	1.453488	0.048662	
9	80250.00	103304.316547	9.667674	0.217188	
62	60558.00	72516.151094	7.096774	0.095404	
109	86000.00	93351.480263	4.056291	0.083942	
126	59167.00	71880.150754	9.433962	0.118191	
153	65588.00	81706.330749	5.201465	0.114607	
190	74000.00	91171.739130	3.398058	0.037406	
196	84583.00	92367.213115	7.053942	0.078788	
226	92250.00	116789.122137	2.277433	0.040031	
229	68636.00	77355.714286	8.307210	0.067818	
231	75893.00	80461.406518	9.430894	0.058716	
292	48170.75	95017.493314	5.420318	0.068771	

```
309       86389.00    91221.120690         3.960396              0.028169
323       92237.00   116438.371218         1.171303              0.037332
327       86591.00   100600.450450         2.215432              0.037486

     pop_struggling_prop  urb_v_rur town_status city_death_match
5               0.024331      rural      shrunk           alford
9               0.050000      rural       grown         aquinnah
62              0.132635      rural   uncertain       clarksburg
109             0.065693      rural       grown        granville
126             0.157246      rural       grown           hawley
153             0.112921      rural   uncertain           leyden
190             0.087282      rural   uncertain       montgomery
196             0.042424      rural       grown      new ashford
226             0.032967      rural   uncertain           pelham
229             0.108508      rural   uncertain             peru
231             0.148012      rural   uncertain      phillipston
292             0.226138      rural       grown          tisbury
309             0.072435      rural   uncertain       washington
323             0.138423      rural      shrunk     west tisbury
327             0.068357      rural       grown      westhampton

[15 rows x 21 columns]
```

These are pretty small towns - for some, it looks like they're considered to be a part of other towns?

Have to keep moving on, afraid will have to lose them

[35]:
```
# how many zipcodes in prescriber dataset not in the postal map?
print(len(set(opi_pres_17_MA['nppes_provider_zip_code']) -␣
 ↪set(ma_postzip_map['POSTCODE'])))
# other way around?
print(len(set(ma_postzip_map['POSTCODE']) -␣
 ↪set(opi_pres_17_MA['nppes_provider_zip_code'])))
```

```
74
105
```

[36]:
```
# try to fix mismatches - how many prescribers are associated with zip codes␣
 ↪not in the postal map shapefile?
zip_miss = set(opi_pres_17_MA['nppes_provider_zip_code']) -␣
 ↪set(ma_postzip_map['POSTCODE'])
# total number of unique npi in 2017 MA opioid prescriber dataset:
print(len(set(opi_pres_17_MA['npi'])))
# number missing from the postal zip codes map
print(len(set(opi_pres_17_MA[opi_pres_17_MA['nppes_provider_zip_code'].
 ↪isin(zip_miss)]['npi'])))
```

```
35730
2316
```

[37]:
```
opi_pres_17_MA[opi_pres_17_MA['nppes_provider_zip_code'].isin(zip_miss)]
```

[37]:

```
               npi nppes_provider_last_name nppes_provider_first_name  \
303      1003015686                   BLUME                    DEBORAH
769      1003041971                   IRONS                    HILLARY
1516     1003080151                  NICORA                  AGNIESZKA
1662     1003090143                   DUBIN                     JOSEPH
1995     1003108259                KORAPATI                     SOWMYA
...             ...                     ...                        ...
1159959  1992854707                 ABRAMOV                 KONSTANTIN
1160834  1992891352                DEPETERS                   FRANKLIN
1161153  1992902977                 CROWLEY                    JILLIAN
1162263  1992964183               MACLACHLAN                       LARA
1162578  1992979645                THANAWALA                      RUCHI

        nppes_provider_zip_code nppes_provider_state  \
303                       01199                   MA
769                       01655                   MA
1516                      01199                   MA
1662                      01805                   MA
1995                      01655                   MA
...                         ...                  ...
1159959                   01655                   MA
1160834                   01655                   MA
1161153                   01199                   MA
1162263                   01805                   MA
1162578                   01199                   MA

                                  specialty_description  total_claim_count  \
303                                 Physician Assistant               1562
769                                  Emergency Medicine                 53
1516                                 Emergency Medicine                216
1662                                  Internal Medicine                139
1995                                Hematology-Oncology                 15
...                                                 ...                ...
1159959                                     Nephrology               1519
1160834                           Diagnostic Radiology                 68
1161153                             Physician Assistant                107
1162263                                      Urology               1487
1162578  Student in an Organized Health Care Education/...                 20

         opioid_claim_count  opioid_prescribing_rate  \
303                    28.0                     1.79
769                     NaN                      NaN
1516                    NaN                      NaN
1662                    NaN                      NaN
1995                    NaN                      NaN
```

```
...                     ...                     ...
1159959                27.0                    1.78
1160834                 0.0                    0.00
1161153                18.0                   16.82
1162263                 NaN                     NaN
1162578                16.0                   80.00

         long-acting_opioid_claim_count  long-acting_opioid_prescribing_rate
303                                 0.0                                 0.0
769                                 0.0                                 NaN
1516                                0.0                                 NaN
1662                                0.0                                 NaN
1995                                NaN                                 NaN
...                                 ...                                 ...
1159959                             NaN                                 NaN
1160834                             0.0                                 NaN
1161153                             0.0                                 0.0
1162263                             0.0                                 NaN
1162578                             0.0                                 0.0

[2316 rows x 11 columns]
```

[38]:
```
# subset map to try and match up zip codes and towns
postzip_town_match = ma_postzip_map.iloc[:, 0:7]
display(postzip_town_match.head())
```

```
  POSTCODE          PC_NAME     PC_TYPE          PA_NAME PA_FIPS  CITY_TOWN  \
0    01331            ATHOL  NON UNIQUE            ATHOL   02515      athol
1    01085        WESTFIELD  NON UNIQUE        WESTFIELD   76030  westfield
2    01370  SHELBURNE FALLS  NON UNIQUE  SHELBURNE FALLS   61205  shelburne
3    01235         HINSDALE  NON UNIQUE         HINSDALE   30280   hinsdale
4    02747  NORTH DARTMOUTH  NON UNIQUE  NORTH DARTMOUTH   47450  dartmouth

      COUNTY
0  WORCESTER
1    HAMPDEN
2   FRANKLIN
3  BERKSHIRE
4    BRISTOL
```

[39]:
```
# PC and PA name columns look identical - are they?
print(sum(postzip_town_match['PC_NAME'] != postzip_town_match['PA_NAME']))
display(postzip_town_match[postzip_town_match['PC_NAME'] !=␣
 ↪postzip_town_match['PA_NAME']])
postzip_town_match = postzip_town_match.apply(lambda x: x.str.lower())
postzip_town_match.head()
```

```
3
```

```
    POSTCODE            PC_NAME              PC_TYPE       PA_NAME PA_FIPS  \
497    01441               TYCO  UNIQUE ORGANIZATION   WESTMINSTER   76975
520    02357  STONEHILL COLLEGE  UNIQUE ORGANIZATION  NORTH EASTON   47625
530    01063      SMITH COLLEGE  UNIQUE ORGANIZATION   NORTHAMPTON   46330

       CITY_TOWN     COUNTY
497  westminster  WORCESTER
520       easton    BRISTOL
530  northampton  HAMPSHIRE
```

[39]:
```
   POSTCODE          PC_NAME     PC_TYPE          PA_NAME PA_FIPS  CITY_TOWN  \
0    01331            athol  non unique            athol   02515      athol
1    01085        westfield  non unique        westfield   76030  westfield
2    01370  shelburne falls  non unique  shelburne falls   61205  shelburne
3    01235         hinsdale  non unique         hinsdale   30280   hinsdale
4    02747  north dartmouth  non unique  north dartmouth   47450  dartmouth

      COUNTY
0  worcester
1    hampden
2   franklin
3  berkshire
4    bristol
```

[40]:
```
print(set(ma_over_death['city_death_match']) -␣
 ↪set(postzip_town_match['CITY_TOWN']))
print(set(postzip_town_match['CITY_TOWN']) -␣
 ↪set(ma_over_death['city_death_match']))
```

```
{'clarksburg', 'aquinnah', 'new ashford', 'tisbury', 'peru', 'phillipston',
'alford', 'west tisbury', 'westhampton', 'granville', 'pelham', 'hawley',
'washington', 'montgomery', 'leyden'}
{'charlemont', 'worthington', 'monroe'}
```

[41]:
```
len(zip_miss)
```

[41]:
```
74
```

Maybe the zip code - town association files will be useful for filling in these missing town - zip code associations that do match the opioid overdose death dataset?

[42]:
```
print(len(set(ma_over_death['city_death_match']) - set(zip_town['town'])))
print(len(set(ma_over_death['city_death_match']) - set(zip_town_raw_alt['city'].
 ↪str.lower())))
#zip_town town names are more similar to the opioid overdose death count towns␣
 ↪- try that one
```

```
4
28
```

[43]:
```
print(len(zip_miss))
print(len(zip_town[zip_town['zip'].isin(zip_miss)]))
```

```
74
36
```

[44]:
```
zip_fill = zip_town[zip_town['zip'].isin(zip_miss) & zip_town['town'].
 ↪isin(set(ma_over_death['city_death']))]
zip_fill
```

[44]:

	zip	town
88	02117	boston
111	02199	boston
118	02211	boston
124	02241	boston
143	02185	braintree
148	02325	bridgewater
167	01805	burlington
177	02238	cambridge
298	02334	easton
309	02722	fall river
313	02541	falmouth
457	01842	lawrence
482	01853	lowell
491	01910	lynn
687	01061	northampton
733	01202	pittsfield
873	01102	springfield
889	01144	springfield
892	01199	springfield
923	02575	tisbury
1015	02575	west tisbury
1082	01614	worcester
1086	01655	worcester

[45]:
```
zip_town_fin = postzip_town_match[['POSTCODE', 'CITY_TOWN']]
zip_town_fin.columns = list(zip_fill.columns)
zip_town_fin = pd.concat([zip_town_fin, zip_fill]).reindex()
print(zip_town_fin.shape)
print(zip_town_fin.drop_duplicates().shape)
zip_town_fin.head()
```

```
(572, 2)
(572, 2)
```

[45]:

	zip	town
0	01331	athol
1	01085	westfield
2	01370	shelburne
3	01235	hinsdale
4	02747	dartmouth

[46]:
```
print(len(set(zip_town_fin['town']) - set(ma_over_death['city_death_match'])))
print(len(set(ma_over_death['city_death_match']) - set(zip_town_fin['town'])))
print(set(zip_town_fin['town']) - set(ma_over_death['city_death_match']))
print(set(ma_over_death['city_death_match']) - set(zip_town_fin['town']))
# getting closer
```

```
3
13
{'charlemont', 'worthington', 'monroe'}
{'clarksburg', 'aquinnah', 'new ashford', 'peru', 'phillipston', 'alford',
'westhampton', 'granville', 'pelham', 'hawley', 'washington', 'montgomery',
'leyden'}
```

[47]:
```
# zip codes that have a town association - but that town name is not in the␣
 ↪opioid overdose death data
# do these zip codes have another association that matches the opioid overdose␣
 ↪death data?
zip_miss_town = zip_town[zip_town['zip'].isin(zip_miss) & ~zip_town['town'].
 ↪isin(set(ma_over_death['city_death']))].copy()
display(zip_miss_town)
# used manual database look up to look these up
zip_miss_town['alt_town'] = ['springfield', 'bridgewater', 'orleans',
                    'boston', 'cambridge', 'danvers',
                    'burlington', 'plymouth', 'marshfield',
                    'eastham', 'northampton', 'winchendon', 'winchendon']
zip_miss_town
```

	zip	town
54	01199	baystate medical
149	02325	bridgewater state college
282	02643	east orleans
323	02241	fleet bank boston
383	02238	harvard square
391	01937	hathorne
445	01805	lahey clinic medical center
498	02345	manomet
509	02051	marshfield hills
651	02651	north eastham
662	01061	north hampton
1055	01477	winchdon springs

```
1057  01477           winchendon springs
```

[47]:

	zip	town	alt_town
54	01199	baystate medical	springfield
149	02325	bridgewater state college	bridgewater
282	02643	east orleans	orleans
323	02241	fleet bank boston	boston
383	02238	harvard square	cambridge
391	01937	hathorne	danvers
445	01805	lahey clinic medical center	burlington
498	02345	manomet	plymouth
509	02051	marshfield hills	marshfield
651	02651	north eastham	eastham
662	01061	north hampton	northampton
1055	01477	winchdon springs	winchendon
1057	01477	winchendon springs	winchendon

[48]:
```
# are all of these new alternative town names in the opioid overdose death dataset?
set(zip_miss_town['alt_town']) - set(ma_over_death['city_death_match'])
# yes
```

[48]:
```
set()
```

[49]:
```
display(zip_town_fin)
# modify to combine these two dfs
zip_miss_town = zip_miss_town.drop('town', axis=1)
zip_miss_town.columns = zip_town_fin.columns
```

	zip	town
0	01331	athol
1	01085	westfield
2	01370	shelburne
3	01235	hinsdale
4	02747	dartmouth
...	...	...
892	01199	springfield
923	02575	tisbury
1015	02575	west tisbury
1082	01614	worcester
1086	01655	worcester

[572 rows x 2 columns]

[50]:
```
zip_town_fin = pd.concat([zip_town_fin, zip_miss_town]).drop_duplicates()
print(zip_town_fin.shape)
zip_town_fin.head()
```

```
(578, 2)
```

[50]:

	zip	town
0	01331	athol
1	01085	westfield
2	01370	shelburne
3	01235	hinsdale
4	02747	dartmouth

[51]:
```
print(len(set(zip_town_fin['town']) - set(ma_over_death['city_death_match'])))
print(len(set(ma_over_death['city_death_match']) - set(zip_town_fin['town'])))
print(set(zip_town_fin['town']) - set(ma_over_death['city_death_match']))
print(set(ma_over_death['city_death_match']) - set(zip_town_fin['town']))
```

```
3
13
{'charlemont', 'worthington', 'monroe'}
{'clarksburg', 'aquinnah', 'new ashford', 'peru', 'phillipston', 'alford',
'westhampton', 'granville', 'pelham', 'hawley', 'washington', 'montgomery',
'leyden'}
```

[52]:
```
# how many zip codes associated with more than 1 town?
zip_town_counts = zip_town_fin['zip'].value_counts().reset_index()
zip_town_counts.columns = ['zip', 'count']
print(zip_town_counts[zip_town_counts['count'] > 1].shape)
zip_town_counts[zip_town_counts['count'] > 1]
```

```
(23, 2)
```

[52]:

	zip	count
0	02467	3
1	01082	3
2	02148	2
3	01010	2
4	02135	2
5	01096	2
6	01070	2
7	02151	2
8	01434	2
9	01235	2
10	01008	2
11	02575	2
12	01223	2
13	01050	2
14	02718	2
15	01247	2
16	01002	2

```
17  02136      2
18  02532      2
19  01011      2
20  01026      2
21  01432      2
22  01039      2
```

[53]:
```
# number of opi prescribers now missing with this better zip code - town␣
 ↪association list:
print(opi_pres_17_MA[opi_pres_17_MA['nppes_provider_zip_code'].
 ↪isin(set(opi_pres_17_MA['nppes_provider_zip_code']) -␣
 ↪set(zip_town_fin['zip']))].shape)
# down from 2k + missing to only 52 - great!
```

```
(52, 11)
```

1.0.4 Step 2: Associate all opioid prescribers with opioid overdose death town name (join by zip code) and 2017 all-drug prescribers with zip code (based on npi from opioid prescriber dataset) and then with town (based on zip code)

- how many all-drug prescribers don't have an associated town?
- check for duplicate town assignments to prescribers

[54]:
```
# give the all drug prescribers zipcodes
all_pres_17_MA_prov = all_pres_17_MA_prov.merge(opi_pres_17_MA[['npi',␣
 ↪'nppes_provider_zip_code']], on='npi', how='left')
display(all_pres_17_MA_prov.info())
display(all_pres_17_MA_prov.head())
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 27300 entries, 0 to 27299
Data columns (total 7 columns):
npi                           27300 non-null int64
nppes_provider_last_org_name  27299 non-null object
nppes_provider_first_name     27300 non-null object
nppes_provider_city           27300 non-null object
nppes_provider_state          27300 non-null object
specialty_description         27300 non-null object
nppes_provider_zip_code       27300 non-null object
dtypes: int64(1), object(6)
memory usage: 1.7+ MB

None
```

```
          npi nppes_provider_last_org_name nppes_provider_first_name  \
0  1003002312                      HOPKINS                  PATRICIA
1  1003007477                        ABDOW                  KIMBERLY
```

```
2  1003008244                  RAJBHANDARI                       RUMA
3  1003011610                          RAY                      ALAKA
4  1003012766                         KANO                    ZACHARY

  nppes_provider_city nppes_provider_state specialty_description  \
0              QUINCY                    MA          Rheumatology
1           WORCESTER                    MA    Nurse Practitioner
2              BOSTON                    MA      Gastroenterology
3              BOSTON                    MA     Internal Medicine
4           BROOKLINE                    MA               Dentist

  nppes_provider_zip_code
0                   02169
1                   01609
2                   02115
3                   02114
4                   02445
```

[55]:
```
# number of all drug prescribers missing/being dropped because of zip code:
print(len(set(all_pres_17_MA_prov['npi'])))
print(all_pres_17_MA_prov[all_pres_17_MA_prov['nppes_provider_zip_code'].
 ↪isin(set(all_pres_17_MA_prov['nppes_provider_zip_code']) -␣
 ↪set(zip_town_fin['zip']))].shape)
# only 41 - great
```

```
27300
(41, 7)
```

[56]:
```
opi_pres_17_MA['nppes_provider_last_name'] =␣
 ↪opi_pres_17_MA['nppes_provider_last_name'].str.lower()
opi_pres_17_MA['nppes_provider_first_name'] =␣
 ↪opi_pres_17_MA['nppes_provider_first_name'].str.lower()
opi_pres_17_MA['specialty_description'] =␣
 ↪opi_pres_17_MA['specialty_description'].str.lower()
print(opi_pres_17_MA.shape)
opi_pres_17_town = opi_pres_17_MA.merge(zip_town_fin, how='inner',␣
 ↪left_on='nppes_provider_zip_code', right_on='zip')
print(opi_pres_17_town.shape)
opi_pres_17_town.head()
```

```
(35730, 11)
(37069, 13)
```

[56]:
```
          npi nppes_provider_last_name nppes_provider_first_name  \
0  1003002312                  hopkins                   patricia
1  1003083270                   kabadi                     mitesh
```

```
2  1003291121                    blair                    meghan
3  1003834433                     nair                      anil
4  1003992397                  carolan                  patricia

  nppes_provider_zip_code nppes_provider_state specialty_description  \
0                   02169                   MA          rheumatology
1                   02169                   MA            cardiology
2                   02169                   MA    nurse practitioner
3                   02169                   MA             neurology
4                   02169                   MA               dentist

   total_claim_count  opioid_claim_count  opioid_prescribing_rate  \
0               4487               513.0                    11.43
1               1363                 0.0                     0.00
2                905                 0.0                     0.00
3               1133                 NaN                      NaN
4                 54                 NaN                      NaN

   long-acting_opioid_claim_count  long-acting_opioid_prescribing_rate    zip  \
0                            84.0                                16.37  02169
1                             0.0                                  NaN  02169
2                             0.0                                  NaN  02169
3                             NaN                                  NaN  02169
4                             0.0                                  NaN  02169

     town
0  quincy
1  quincy
2  quincy
3  quincy
4  quincy
```

[57]:
```
# number of opi prescribers w/more than 1 town association:
print(opi_pres_17_MA[opi_pres_17_MA['nppes_provider_zip_code'].
 ↪isin(set(zip_town_counts[zip_town_counts['count'] > 1]['zip']))].shape)
# number of all drug prescribers w/more than 1 town association:
print(all_pres_17_MA_prov[all_pres_17_MA_prov['nppes_provider_zip_code'].
 ↪isin(set(zip_town_counts[zip_town_counts['count'] > 1]['zip']))].shape)
```

```
(1156, 11)
(876, 7)
```

[58]:
```
opi_pres_17_cols = list(opi_pres_17_town.columns)
# reoder columns and drop extra zipcode col that appeared because of merge
opi_pres_17_town = opi_pres_17_town[opi_pres_17_cols[0:4] + opi_pres_17_cols[-1:
 ↪] + opi_pres_17_cols[4:10]]
opi_pres_17_town.head()
```

[58]:
```
          npi nppes_provider_last_name nppes_provider_first_name  \
0  1003002312                  hopkins                  patricia
1  1003083270                   kabadi                    mitesh
2  1003291121                    blair                    meghan
3  1003834433                     nair                      anil
4  1003992397                  carolan                  patricia

  nppes_provider_zip_code    town nppes_provider_state specialty_description  \
0                   02169  quincy                   MA          rheumatology
1                   02169  quincy                   MA            cardiology
2                   02169  quincy                   MA    nurse practitioner
3                   02169  quincy                   MA             neurology
4                   02169  quincy                   MA               dentist

   total_claim_count  opioid_claim_count  opioid_prescribing_rate  \
0               4487               513.0                    11.43
1               1363                 0.0                     0.00
2                905                 0.0                     0.00
3               1133                 NaN                      NaN
4                 54                 NaN                      NaN

   long-acting_opioid_claim_count
0                            84.0
1                             0.0
2                             0.0
3                             NaN
4                             0.0
```

[59]:
```
npi_town_count = opi_pres_17_town['npi'].value_counts().reset_index()
npi_town_count.columns = ['npi', 'count']
display(npi_town_count.head())
npi_town_count[npi_town_count['count'] > 1].shape
```

```
          npi  count
0  1962471714      3
1  1689687097      3
2  1679530083      3
3  1720064272      3
4  1306882410      3
```

[59]:
```
(1156, 2)
```

Will need to deal with duplicates at some point - not sure why this is happening (dealt with notebook 6 actually)

Clean up other opioid prescriber datasets (filter for MA only, lower case string, etc - associate with town):

[60]:
```
# other opioid dataset processing:
opi_pres_13_MA = opi_pres_13_raw[opi_pres_13_raw['nppes_provider_state'] ==␣
 ↪'MA'].copy()
opi_pres_14_MA = opi_pres_14_raw[opi_pres_14_raw['nppes_provider_state'] ==␣
 ↪'MA'].copy()
opi_pres_15_MA = opi_pres_15_raw[opi_pres_15_raw['nppes_provider_state'] ==␣
 ↪'MA'].copy()
opi_pres_16_MA = opi_pres_16_raw[opi_pres_16_raw['nppes_provider_state'] ==␣
 ↪'MA'].copy()
```

[61]:
```
# zip code fix for all files
MA_opi_dflist = [opi_pres_13_MA, opi_pres_14_MA, opi_pres_15_MA, opi_pres_16_MA]
for x in MA_opi_dflist:
    x['nppes_provider_zip_code'] = [x.zfill(5) for x in␣
 ↪list(x['nppes_provider_zip_code'].astype(int).astype(str))]
```

[62]:
```
# lowercase names
for x in MA_opi_dflist:
    x['nppes_provider_last_name'] = x['nppes_provider_last_name'].str.lower()
for x in MA_opi_dflist:
    x['nppes_provider_first_name'] = x['nppes_provider_first_name'].str.lower()
for x in MA_opi_dflist:
    x['specialty_description'] = x['specialty_description'].str.lower()
```

[63]:
```
# how many providers missing from the zip code -town association df for each of␣
 ↪these?
print(len(set(opi_pres_13_MA['nppes_provider_zip_code']) -␣
 ↪set(zip_town_fin['zip'])))
print(len(set(opi_pres_14_MA['nppes_provider_zip_code']) -␣
 ↪set(zip_town_fin['zip'])))
print(len(set(opi_pres_15_MA['nppes_provider_zip_code']) -␣
 ↪set(zip_town_fin['zip'])))
print(len(set(opi_pres_16_MA['nppes_provider_zip_code']) -␣
 ↪set(zip_town_fin['zip'])))
```

```
53
46
49
46
```

Looks like the vast majority of the opioid prescribers are in the zip code - town association dataset - good news.

Perform zip code - town merges, check if each went as expected:

[64]:
```
print(opi_pres_13_MA.shape)
opi_pres_13_town = opi_pres_13_MA.merge(zip_town_fin, how='inner',␣
 ↪left_on='nppes_provider_zip_code', right_on='zip')
print(opi_pres_13_town.shape)
opi_pres_13_town.head()
```

```
# also duplicates here
```

```
(32791, 11)
(34086, 13)
```

[64]:
```
          npi nppes_provider_last_name nppes_provider_first_name  \
0  1003002312                  hopkins                  patricia
1  1003083270                   kabadi                    mitesh
2  1003834433                     nair                      anil
3  1003895269                 angelini                   domenic
4  1003992397                   carolan                  patricia

  nppes_provider_zip_code nppes_provider_state specialty_description  \
0                   02169                   MA     internal medicine
1                   02169                   MA            cardiology
2                   02169                   MA             neurology
3                   02169                   MA               dentist
4                   02169                   MA               dentist

   total_claim_count  opioid_claim_count  opioid_prescribing_rate  \
0               4139               522.0                    12.61
1                 40                 0.0                     0.00
2               1217                 NaN                      NaN
3                 14                 0.0                     0.00
4                 37                 NaN                      NaN

   long-acting_opioid_claim_count  long-acting_opioid_prescribing_rate    zip  \
0                           104.0                                19.92  02169
1                             0.0                                  NaN  02169
2                             NaN                                  NaN  02169
3                             0.0                                  NaN  02169
4                             0.0                                  NaN  02169

     town
0  quincy
1  quincy
2  quincy
3  quincy
4  quincy
```

[65]:
```
print(opi_pres_14_MA.shape)
opi_pres_14_town = opi_pres_14_MA.merge(zip_town_fin, how='inner',␣
 ↪left_on='nppes_provider_zip_code', right_on='zip')
print(opi_pres_14_town.shape)
opi_pres_14_town.head()
```

```
(33380, 11)
(34734, 13)
```

[65]:
```
          npi nppes_provider_last_name nppes_provider_first_name  \
0  1003001660                   newton                    robert
1  1003841198                   kaplan                      mark
2  1003842543                reiner,jr                  marshall
3  1013101740                      lee                     peter
4  1013125731                lipetsker                   nickolay

  nppes_provider_zip_code nppes_provider_state  specialty_description  \
0                   02446                   MA                urology
1                   02446                   MA  obstetrics/gynecology
2                   02446                   MA     pediatric medicine
3                   02446                   MA                dentist
4                   02446                   MA                dentist

   total_claim_count  opioid_claim_count  opioid_prescribing_rate  \
0                 12                 0.0                      0.0
1                 76                 0.0                      0.0
2                 45                 NaN                      NaN
3                 30                 0.0                      0.0
4                134                 0.0                      0.0

   long-acting_opioid_claim_count  long-acting_opioid_prescribing_rate    zip  \
0                             0.0                                  NaN  02446
1                             0.0                                  NaN  02446
2                             0.0                                  NaN  02446
3                             0.0                                  NaN  02446
4                             0.0                                  NaN  02446

        town
0  brookline
1  brookline
2  brookline
3  brookline
4  brookline
```

[66]:
```
print(opi_pres_15_MA.shape)
opi_pres_15_town = opi_pres_15_MA.merge(zip_town_fin, how='inner',␣
 ↪left_on='nppes_provider_zip_code', right_on='zip')
print(opi_pres_15_town.shape)
opi_pres_15_town.head()
```

```
(34081, 11)
(35416, 13)
```

[66]:
```
          npi nppes_provider_last_name nppes_provider_first_name  \
0  1003002312                  hopkins                  patricia
1  1003083270                   kabadi                    mitesh
2  1003291121                    blair                    meghan
```

```
3  1003834433                    nair                      anil
4  1003895269                 angelini                   domenic

  nppes_provider_zip_code nppes_provider_state specialty_description  \
0                   02169                   MA     internal medicine
1                   02169                   MA            cardiology
2                   02169                   MA    nurse practitioner
3                   02169                   MA             neurology
4                   02169                   MA               dentist

   total_claim_count  opioid_claim_count  opioid_prescribing_rate  \
0               4183               495.0                    11.83
1                906                 0.0                     0.00
2                 45                 0.0                     0.00
3               1079                16.0                     1.48
4                 17                 NaN                      NaN

   long-acting_opioid_claim_count  long-acting_opioid_prescribing_rate    zip  \
0                            99.0                                 20.0  02169
1                             0.0                                  NaN  02169
2                             0.0                                  NaN  02169
3                             0.0                                  0.0  02169
4                             0.0                                  NaN  02169

     town
0  quincy
1  quincy
2  quincy
3  quincy
4  quincy
```

[67]:
```
print(opi_pres_16_MA.shape)
opi_pres_16_town = opi_pres_16_MA.merge(zip_town_fin, how='inner',␣
 ↪left_on='nppes_provider_zip_code', right_on='zip')
print(opi_pres_16_town.shape)
opi_pres_16_town.head()
```

```
(35029, 11)
(36357, 13)
```

[67]:
```
          npi nppes_provider_last_name nppes_provider_first_name  \
0  1003002312                 hopkins                  patricia
1  1003083270                  kabadi                    mitesh
2  1003291121                   blair                    meghan
3  1003834433                    nair                      anil
4  1003895269                 angelini                   domenic

  nppes_provider_zip_code nppes_provider_state  \
```

```
0                    02169                    MA
1                    02169                    MA
2                    02169                    MA
3                    02169                    MA
4                    02169                    MA

                specialty_description  total_claim_count  opioid_claim_count  \
0                         rheumatology               4634               593.0
1  cardiovascular disease (cardiology)               1147                 0.0
2                  nurse practitioner                888                 0.0
3                           neurology               1096                14.0
4                             dentist                 26                 0.0

   opioid_prescribing_rate  long-acting_opioid_claim_count  \
0                    12.80                           106.0
1                     0.00                             0.0
2                     0.00                             0.0
3                     1.28                             0.0
4                     0.00                             0.0

   long-acting_opioid_prescribing_rate    zip    town
0                                17.88  02169  quincy
1                                  NaN  02169  quincy
2                                  NaN  02169  quincy
3                                 0.00  02169  quincy
4                                  NaN  02169  quincy
```

[68]:
```
print(list(opi_pres_13_town.columns) == opi_pres_17_cols)
#opi_town_dflist = [opi_pres_13_town, opi_pres_14_town, opi_pres_15_town,␣
 ↪opi_pres_16_town]
#for x in opi_town_dflist:
#    x = x[opi_pres_17_cols[0:4] + opi_pres_17_cols[-1:] + opi_pres_17_cols[4:
 ↪10]]
```

```
True
```

[69]:
```
opi_pres_13_town = opi_pres_13_town[opi_pres_17_cols[0:4] + opi_pres_17_cols[-1:
 ↪] + opi_pres_17_cols[4:10]]
opi_pres_14_town = opi_pres_14_town[opi_pres_17_cols[0:4] + opi_pres_17_cols[-1:
 ↪] + opi_pres_17_cols[4:10]]
opi_pres_15_town = opi_pres_15_town[opi_pres_17_cols[0:4] + opi_pres_17_cols[-1:
 ↪] + opi_pres_17_cols[4:10]]
opi_pres_16_town = opi_pres_16_town[opi_pres_17_cols[0:4] + opi_pres_17_cols[-1:
 ↪] + opi_pres_17_cols[4:10]]
```

[70]:
```
opi_pres_13_town.head()
```

[70]:
```
          npi nppes_provider_last_name nppes_provider_first_name  \
0  1003002312                  hopkins                  patricia
1  1003083270                   kabadi                    mitesh
2  1003834433                     nair                      anil
3  1003895269                 angelini                   domenic
4  1003992397                  carolan                  patricia

  nppes_provider_zip_code    town nppes_provider_state specialty_description  \
0                   02169  quincy                   MA     internal medicine
1                   02169  quincy                   MA            cardiology
2                   02169  quincy                   MA             neurology
3                   02169  quincy                   MA               dentist
4                   02169  quincy                   MA               dentist

   total_claim_count  opioid_claim_count  opioid_prescribing_rate  \
0               4139               522.0                    12.61
1                 40                 0.0                     0.00
2               1217                 NaN                      NaN
3                 14                 0.0                     0.00
4                 37                 NaN                      NaN

   long-acting_opioid_claim_count
0                           104.0
1                             0.0
2                             NaN
3                             0.0
4                             0.0
```

[71]:
```
# does dropping duplicates do anything at all?
print(opi_pres_13_town.shape)
print(opi_pres_13_town.drop_duplicates().shape)
print(opi_pres_14_town.shape)
print(opi_pres_14_town.drop_duplicates().shape)
print(opi_pres_15_town.shape)
print(opi_pres_15_town.drop_duplicates().shape)
print(opi_pres_16_town.shape)
print(opi_pres_16_town.drop_duplicates().shape)
print(opi_pres_17_town.shape)
print(opi_pres_17_town.drop_duplicates().shape)
# nope
```

```
(34086, 11)
(34086, 11)
(34734, 11)
(34734, 11)
(35416, 11)
(35416, 11)
(36357, 11)
```

```
(36357, 11)
(37069, 11)
(37069, 11)
```

[72]:
```
# create year column for each df
opi_pres_13_town['year'] = ['2013'] * len(opi_pres_13_town.index)
opi_pres_14_town['year'] = ['2014'] * len(opi_pres_14_town.index)
opi_pres_15_town['year'] = ['2015'] * len(opi_pres_15_town.index)
opi_pres_16_town['year'] = ['2016'] * len(opi_pres_16_town.index)
opi_pres_17_town['year'] = ['2017'] * len(opi_pres_17_town.index)
```

[73]:
```
opi_pres_17_town.sample(20)
```

[73]:

	npi	nppes_provider_last_name	nppes_provider_first_name	\
33270	1184174187	ambler	christopher	
20476	1265484190	saeed	asad	
27236	1205064417	korinow	doron	
18039	1932501798	black	brenda	
24279	1245342583	cohen	jeffrey	
24072	1366686628	gonzalez casals	abel	
30154	1093755522	hession	brian	
18178	1649533761	choi	jinyoung	
14914	1366970618	ning	ying	
27701	1932224995	gorenberg	david	
9813	1104117431	vandaam	percival	
18191	1831157577	zakaria	sarah	
34202	1366880882	garabedian	laurie	
497	1043397813	weintraub	joanne	
33818	1790890705	bonavita	andrew	
17332	1841381027	radisic-basovic	marina	
13844	1023116043	chou	rosemary	
16501	1699864413	desai	neelam	
6372	1689718009	javer	robert	
225	1659372787	goldin	dennis	

	nppes_provider_zip_code	town	nppes_provider_state	\
33270	01005	barre	MA	
20476	02301	brockton	MA	
27236	01950	newburyport	MA	
18039	01824	chelmsford	MA	
24279	01453	leominster	MA	
24072	01104	springfield	MA	
30154	01040	holyoke	MA	
18178	01610	worcester	MA	
14914	01608	worcester	MA	
27701	02540	falmouth	MA	
9813	02149	everett	MA	
18191	01610	worcester	MA	

```
34202                  02035   foxborough                  MA
497                    02115       boston                  MA
33818                  01106   longmeadow                  MA
17332                  02740  new bedford                  MA
13844                  01805   burlington                  MA
16501                  02215       boston                  MA
6372                   02481    wellesley                  MA
225                    02169       quincy                  MA

        specialty_description  total_claim_count  opioid_claim_count  \
33270      nurse practitioner                253                 NaN
20476       internal medicine               1697                 0.0
27236      emergency medicine                109                18.0
18039      nurse practitioner                325                 NaN
24279                  dentist                 15                 NaN
24072               psychiatry               6995                 0.0
30154      emergency medicine                487                76.0
18178                  dentist                 11                 0.0
14914       internal medicine                 17                 NaN
27701  psychiatry & neurology                420                 NaN
9813      physician assistant                 58                 NaN
18191         family practice                130                 0.0
34202         family practice               1122                32.0
497        nurse practitioner                490                 0.0
33818                  dentist                157                 NaN
17332               psychiatry               2282                 0.0
13844       internal medicine                 88                 0.0
16501     hematology-oncology                171                 NaN
6372                   dentist                142                37.0
225         internal medicine                 16                 0.0

       opioid_prescribing_rate  long-acting_opioid_claim_count  year
33270                      NaN                             0.0  2017
20476                     0.00                             0.0  2017
27236                    16.51                             0.0  2017
18039                      NaN                             NaN  2017
24279                      NaN                             0.0  2017
24072                     0.00                             0.0  2017
30154                    15.61                             0.0  2017
18178                     0.00                             0.0  2017
14914                      NaN                             0.0  2017
27701                      NaN                             0.0  2017
9813                       NaN                             0.0  2017
18191                     0.00                             0.0  2017
34202                     2.85                             0.0  2017
497                       0.00                             0.0  2017
33818                      NaN                             0.0  2017
```

```
17332                    0.00                          0.0  2017
13844                    0.00                          0.0  2017
16501                     NaN                          0.0  2017
6372                    26.06                          0.0  2017
225                      0.00                          0.0  2017
```

Write result to csv - but will have to edit these to remove multiple prescriber - town associations

[74]:
```
#opi_pres_13_town.to_csv("../../data/tidy_data/
 ↪medicare_partD_opioid_prescriber_2013_w_zip_MAtown_v1.csv", index=False)
#opi_pres_14_town.to_csv("../../data/tidy_data/
 ↪medicare_partD_opioid_prescriber_2014_w_zip_MAtown_v1.csv", index=False)
#opi_pres_15_town.to_csv("../../data/tidy_data/
 ↪medicare_partD_opioid_prescriber_2015_w_zip_MAtown_v1.csv", index=False)
#opi_pres_16_town.to_csv("../../data/tidy_data/
 ↪medicare_partD_opioid_prescriber_2016_w_zip_MAtown_v1.csv", index=False)
#opi_pres_17_town.to_csv("../../data/tidy_data/
 ↪medicare_partD_opioid_prescriber_2017_w_zip_MAtown_v1.csv", index=False)
```

1.0.5 Step 3: All drug MA prescribers

This section was modified to rely on the zip code - town associations derived from notebook 6, where the issue of opioid prescribers being linked to more than 1 town was dealt with.

[75]:
```
all_pres_17_MA_prov.head()
```

[75]:
```
          npi nppes_provider_last_org_name nppes_provider_first_name  \
0  1003002312                      HOPKINS                   PATRICIA
1  1003007477                        ABDOW                   KIMBERLY
2  1003008244                   RAJBHANDARI                       RUMA
3  1003011610                          RAY                     ALAKA
4  1003012766                         KANO                    ZACHARY

  nppes_provider_city nppes_provider_state specialty_description  \
0              QUINCY                    MA          Rheumatology
1           WORCESTER                    MA    Nurse Practitioner
2              BOSTON                    MA      Gastroenterology
3              BOSTON                    MA     Internal Medicine
4           BROOKLINE                    MA               Dentist

  nppes_provider_zip_code
0                   02169
1                   01609
2                   02115
3                   02114
4                   02445
```

[76]:
```

```

```
med_opi_pres_no_town_dup = pd.read_csv("../../data/tidy_data/
 ↪medicare_partD_opioid_prescriber_all_years_no_ziptown_duplicates.csv")
med_opi_pres_no_town_dup.head()
```

[76]:

	npi	nppes_provider_last_name	nppes_provider_first_name	nppes_provider_zip_code	town	specialty_description	total_claim_count	opioid_claim_count	year	calc_opioid_rate
0	1003001660	newton	robert	2446	brookline	urology	12	0.0	2014	0.000000
1	1003002312	hopkins	patricia	2169	quincy	internal medicine	4139	522.0	2013	12.611742
2	1003002312	hopkins	patricia	2169	quincy	internal medicine	4467	542.0	2014	12.133423
3	1003002312	hopkins	patricia	2169	quincy	internal medicine	4183	495.0	2015	11.833612
4	1003002312	hopkins	patricia	2169	quincy	rheumatology	4634	593.0	2016	12.796720

[77]:
```
print(len(set(all_pres_17_MA['npi']) - set(med_opi_pres_no_town_dup['npi'])))
print(len(set(all_pres_17_MA['npi'])))
```

```
39
27300
```

[78]:
```
print(len(set(all_pres_17_MA['npi']) -␣
 ↪set(med_opi_pres_no_town_dup[med_opi_pres_no_town_dup['year'] ==␣
 ↪2017]['npi'])))
```

```
41
```

[79]:
```
# drop previously assigned zip codes just in case
all_pres_17_MA_prov = all_pres_17_MA_prov.drop('nppes_provider_zip_code',␣
 ↪axis=1).apply(lambda x: x.astype(str).str.lower())
```

[80]:
```
all_pres_17_MA_prov['npi'] = all_pres_17_MA_prov['npi'].astype(int)
```

[82]:
```
len(set(all_pres_17_MA_prov['npi']) - set(med_opi_pres_no_town_dup['npi']))
```

[82]: 39

[83]:
```
# focus on providers only
all_pres_17_prov = all_pres_17_MA_prov.drop(['nppes_provider_city',␣
 ↪'nppes_provider_state', 'specialty_description'], axis=1)
```

```
print(all_pres_17_prov.shape)
print(all_pres_17_prov.drop_duplicates().shape)
all_pres_17_prov.head()
```

```
(27300, 3)
(27300, 3)
```

[83]:

	npi	nppes_provider_last_org_name	nppes_provider_first_name
0	1003002312	hopkins	patricia
1	1003007477	abdow	kimberly
2	1003008244	rajbhandari	ruma
3	1003011610	ray	alaka
4	1003012766	kano	zachary

[86]:
```
#Are the NPI unique? Are the combinations of provider npi/last name/ first name␣
↪unique?
print(len(set(all_pres_17_MA_prov['npi'])))
print(len(all_pres_17_MA_prov[['npi', 'nppes_provider_last_org_name',␣
↪'nppes_provider_first_name']].index))
```

```
27300
27300
```

[87]:
```
med_opi_pres_no_town_dup.head()
```

[87]:

	npi	nppes_provider_last_name	nppes_provider_first_name	nppes_provider_zip_code	town	specialty_description	total_claim_count	opioid_claim_count	year	calc_opioid_rate
0	1003001660	newton	robert	2446	brookline	urology	12	0.0	2014	0.000000
1	1003002312	hopkins	patricia	2169	quincy	internal medicine	4139	522.0	2013	12.611742
2	1003002312	hopkins	patricia	2169	quincy	internal medicine	4467	542.0	2014	12.133423
3	1003002312	hopkins	patricia	2169	quincy	internal medicine	4183	495.0	2015	11.833612
4	1003002312	hopkins	patricia	2169	quincy	rheumatology	4634	593.0	2016	12.796720

[88]:
```
# narrow down file to only needed columns:
```

```
opi_pres_17_npi_town_match =␣
↪med_opi_pres_no_town_dup[med_opi_pres_no_town_dup['year'] == 2017][['npi',␣
↪'nppes_provider_last_name', 'nppes_provider_first_name', 'town']].copy()
opi_pres_17_npi_town_match.head()
```

[88]:

	npi	nppes_provider_last_name	nppes_provider_first_name	town
5	1003002312	hopkins	patricia	quincy
9	1003007477	abdow	kimberly	worcester
14	1003008244	rajbhandari	ruma	boston
19	1003011610	ray	alaka	boston
25	1003012766	kano	zachary	brookline

[89]:
```
all_pres_17_MA.head()
```

[89]:

	npi	nppes_provider_last_org_name	nppes_provider_first_name	nppes_provider_city	nppes_provider_state	specialty_description	description_flag	drug_name	generic_name	bene_count	...	total_30_day_fill_count	total_day_supply	total_drug_cost	bene_count_ge65	bene_count_ge65_suppress_flag	total_claim_count_ge65
757	1003002312	HOPKINS	PATRICIA	QUINCY	MA	Rheumatology	S	ADVAIR DISKUS	FLUTICASONE/SALMETEROL	NaN	...	34.0	1020	11877.07	NaN	*	14.0
758	1003002312	HOPKINS	PATRICIA	QUINCY	MA	Rheumatology	S	ALLOPURINOL	ALLOPURINOL	21.0	...	136.9	4106	841.33	NaN	#	NaN
759	1003002312	HOPKINS	PATRICIA	QUINCY	MA	Rheumatology	S	ALPRAZOLAM	ALPRAZOLAM	NaN	...	25.0	735	213.58	NaN	#	NaN
760	1003002312	HOPKINS	PATRICIA	QUINCY	MA	Rheumatology	S	AMLODIPINE BESYLATE	AMLODIPINE BESYLATE	19.0	...	125.0	3750	272.86	NaN	#	NaN
761	1003002312	HOPKINS	PATRICIA	QUINCY	MA	Rheumatology	S	ARMOUR THYROID	THYROID,PORK	NaN	...	28.1	772	321.92	NaN	#	NaN

```
     ge65_suppress_flag total_30_day_fill_count_ge65  total_day_supply_ge65  \
757                 NaN                         34.0                 1020.0
758                   #                          NaN                    NaN
759                   #                          NaN                    NaN
760                   #                          NaN                    NaN
761                   #                          NaN                    NaN

     total_drug_cost_ge65
757              11877.07
758                   NaN
759                   NaN
760                   NaN
761                   NaN

[5 rows x 21 columns]
```

Drug ideas source: https://mhc.cpnp.org/doi/full/10.9740/mhc.2016.05.120
Focus on alprazolam (Xanax), diazepam (Valium), and lorazepam (Ativan)

[90]:
```
# alprazolam 2017 prescribers:
alprazolam_17_pres = all_pres_17_MA[all_pres_17_MA['generic_name'].str.lower().
 ↪str.find('alprazolam') >= 0]
print(alprazolam_17_pres.shape)
alprazolam_17_pres.head()
```

```
(3912, 21)
```

[90]:
```
              npi nppes_provider_last_org_name nppes_provider_first_name  \
759    1003002312                      HOPKINS                   PATRICIA
2478   1003007477                        ABDOW                   KIMBERLY
8824   1003023284                      MCELROY                    ALLEGRA
15603  1003047473                       HASSEY                    SHERINE
19690  1003062647                     BEAUZILE                     RONALD

      nppes_provider_city nppes_provider_state specialty_description  \
759             QUINCY                    MA          Rheumatology
2478         WORCESTER                    MA    Nurse Practitioner
8824           MASHPEE                    MA    Nurse Practitioner
15603         WESTFORD                    MA    Nurse Practitioner
19690     BELCHERTOWN                     MA     Internal Medicine

      description_flag   drug_name generic_name  bene_count  ...  \
759                  S  ALPRAZOLAM   ALPRAZOLAM         NaN  ...
2478                 S  ALPRAZOLAM   ALPRAZOLAM        15.0  ...
8824                 S  ALPRAZOLAM   ALPRAZOLAM         NaN  ...
15603                S  ALPRAZOLAM   ALPRAZOLAM         NaN  ...
19690                S  ALPRAZOLAM   ALPRAZOLAM         NaN  ...
```

```
       total_30_day_fill_count  total_day_supply  total_drug_cost  \
759                       25.0               735           213.58
2478                      75.0              2160           324.14
8824                      12.0               350            57.18
15603                     17.0               492            89.98
19690                     22.0               660            94.03

       bene_count_ge65  bene_count_ge65_suppress_flag total_claim_count_ge65  \
759                NaN                              #                    NaN
2478               NaN                              *                    NaN
8824               NaN                              *                    NaN
15603              NaN                              #                    NaN
19690              NaN                              *                   14.0

       ge65_suppress_flag total_30_day_fill_count_ge65  total_day_supply_ge65  \
759                     #                          NaN                    NaN
2478                    *                          NaN                    NaN
8824                    *                          NaN                    NaN
15603                   #                          NaN                    NaN
19690                 NaN                         22.0                  660.0

       total_drug_cost_ge65
759                     NaN
2478                    NaN
8824                    NaN
15603                   NaN
19690                 94.03

[5 rows x 21 columns]
```

Note: the `drug_name` column refers to the brand name and the generic_name usually refers to the active ingredient(s). Where the drug_name and generic_name are the same, those are usually considered generic versions of the drug (as opposed to brand name).

How many of the drugs are generic vs brand name prescriptions?

[92]:
```
alprazolam_17_pres['drug_name'].value_counts()
```

[92]:
```
ALPRAZOLAM        3850
ALPRAZOLAM ER       31
XANAX               22
ALPRAZOLAM XR        4
ALPRAZOLAM ODT       3
XANAX XR             2
Name: drug_name, dtype: int64
```

Most are generic prescriptions, that's not surprising because this is medicare - generic drugs are far more affordable.

Repeat for other 2 benzo drugs:

[93]:
```
diazepam_17_pres = all_pres_17_MA[all_pres_17_MA['generic_name'].str.lower().
 ↪str.find('diazepam') >= 0]
print(diazepam_17_pres.shape)
diazepam_17_pres.head()
```

```
(2612, 21)
```

[93]:

	npi	nppes_provider_last_org_name	nppes_provider_first_name	\
777	1003002312	HOPKINS	PATRICIA	
2495	1003007477	ABDOW	KIMBERLY	
8843	1003023284	MCELROY	ALLEGRA	
14526	1003044272	BENDER	ELISE	
19713	1003062647	BEAUZILE	RONALD	

	nppes_provider_city	nppes_provider_state	specialty_description	\
777	QUINCY	MA	Rheumatology	
2495	WORCESTER	MA	Nurse Practitioner	
8843	MASHPEE	MA	Nurse Practitioner	
14526	BRAINTREE	MA	Family Practice	
19713	BELCHERTOWN	MA	Internal Medicine	

	description_flag	drug_name	generic_name	bene_count	...	\
777	S	DIAZEPAM	DIAZEPAM	NaN	...	
2495	S	DIAZEPAM	DIAZEPAM	17.0	...	
8843	S	DIAZEPAM	DIAZEPAM	NaN	...	
14526	S	DIAZEPAM	DIAZEPAM	NaN	...	
19713	S	DIAZEPAM	DIAZEPAM	NaN	...	

	total_30_day_fill_count	total_day_supply	total_drug_cost	\
777	14.0	360	56.56	
2495	97.6	2920	781.21	
8843	24.0	490	61.77	
14526	16.0	453	25.39	
19713	17.0	490	54.91	

	bene_count_ge65	bene_count_ge65_suppress_flag	total_claim_count_ge65	\
777	NaN	*	14.0	
2495	NaN	*	NaN	
8843	NaN	#	NaN	
14526	NaN	*	16.0	
19713	NaN	#	NaN	

	ge65_suppress_flag	total_30_day_fill_count_ge65	total_day_supply_ge65	\
777	NaN	14.0	360.0	
2495	*	NaN	NaN	
8843	#	NaN	NaN	
14526	NaN	16.0	453.0	

```
19713                   #                              NaN                       NaN

       total_drug_cost_ge65
777                   56.56
2495                    NaN
8843                    NaN
14526                 25.39
19713                   NaN

[5 rows x 21 columns]
```

[94]:
```
diazepam_17_pres['drug_name'].value_counts()
```

[94]:
```
DIAZEPAM           2597
VALIUM               14
DIASTAT ACUDIAL       1
Name: drug_name, dtype: int64
```

[95]:
```
lorazepam_17_pres = all_pres_17_MA[all_pres_17_MA['generic_name'].str.lower().
 ↪str.find('lorazepam') >= 0]
print(lorazepam_17_pres.shape)
lorazepam_17_pres.head()
```

```
(6637, 21)
```

[95]:
```
             npi nppes_provider_last_org_name nppes_provider_first_name  \
803   1003002312                      HOPKINS                   PATRICIA
2511  1003007477                        ABDOW                   KIMBERLY
4590  1003011610                          RAY                      ALAKA
6946  1003015686                        BLUME                    DEBORAH
8878  1003023284                      MCELROY                    ALLEGRA

     nppes_provider_city nppes_provider_state specialty_description  \
803                QUINCY                    MA          Rheumatology
2511            WORCESTER                    MA    Nurse Practitioner
4590               BOSTON                    MA     Internal Medicine
6946           SPRINGFIELD                   MA   Physician Assistant
8878              MASHPEE                    MA    Nurse Practitioner

     description_flag  drug_name generic_name  bene_count  ...  \
803                 S  LORAZEPAM    LORAZEPAM        26.0  ...
2511                S  LORAZEPAM    LORAZEPAM        59.0  ...
4590                S  LORAZEPAM    LORAZEPAM         NaN  ...
6946                S  LORAZEPAM    LORAZEPAM         NaN  ...
8878                S  LORAZEPAM    LORAZEPAM        13.0  ...

      total_30_day_fill_count  total_day_supply  total_drug_cost  \
803                     139.0              4028           907.84
2511                    272.0              7863          1645.60
```

```
4590                    20.0               516            49.13
6946                    22.5               600            57.82
8878                    41.0               837           108.39

      bene_count_ge65  bene_count_ge65_suppress_flag total_claim_count_ge65  \
803               NaN                              #                     NaN
2511             17.0                            NaN                    58.0
4590              NaN                              *                    18.0
6946              NaN                              #                     NaN
8878              NaN                              *                    14.0

      ge65_suppress_flag total_30_day_fill_count_ge65  total_day_supply_ge65  \
803                    #                          NaN                    NaN
2511                 NaN                         80.0                 2355.0
4590                 NaN                         20.0                  516.0
6946                   #                          NaN                    NaN
8878                 NaN                         14.0                  251.0

      total_drug_cost_ge65
803                    NaN
2511                494.91
4590                 49.13
6946                   NaN
8878                 19.62

[5 rows x 21 columns]
```

[96]:
```
lorazepam_17_pres['drug_name'].value_counts()
```

[96]:
```
LORAZEPAM             6614
LORAZEPAM INTENSOL      13
ATIVAN                  10
Name: drug_name, dtype: int64
```

[97]:
```
# bind rows together into 1 2017 dataset
benzo_pres_17 = pd.concat([alprazolam_17_pres, diazepam_17_pres,␣
 ↪lorazepam_17_pres]).drop(['nppes_provider_city', 'nppes_provider_state'],␣
 ↪axis=1)
benzo_pres_17.head()
```

[97]:
```
              npi nppes_provider_last_org_name nppes_provider_first_name  \
759    1003002312                      HOPKINS                   PATRICIA
2478   1003007477                        ABDOW                   KIMBERLY
8824   1003023284                      MCELROY                    ALLEGRA
15603  1003047473                       HASSEY                    SHERINE
19690  1003062647                     BEAUZILE                     RONALD

      specialty_description description_flag   drug_name generic_name  \
759             Rheumatology                S  ALPRAZOLAM   ALPRAZOLAM
```

```
2478     Nurse Practitioner                S  ALPRAZOLAM   ALPRAZOLAM
8824     Nurse Practitioner                S  ALPRAZOLAM   ALPRAZOLAM
15603    Nurse Practitioner                S  ALPRAZOLAM   ALPRAZOLAM
19690     Internal Medicine                S  ALPRAZOLAM   ALPRAZOLAM

       bene_count  total_claim_count  total_30_day_fill_count  \
759           NaN                 21                     25.0
2478         15.0                 63                     75.0
8824          NaN                 12                     12.0
15603         NaN                 17                     17.0
19690         NaN                 14                     22.0

       total_day_supply  total_drug_cost  bene_count_ge65  \
759                 735           213.58              NaN
2478               2160           324.14              NaN
8824                350            57.18              NaN
15603               492            89.98              NaN
19690               660            94.03              NaN

      bene_count_ge65_suppress_flag  total_claim_count_ge65  \
759                               #                     NaN
2478                              *                     NaN
8824                              *                     NaN
15603                             #                     NaN
19690                             *                    14.0

      ge65_suppress_flag  total_30_day_fill_count_ge65  total_day_supply_ge65  \
759                    #                           NaN                    NaN
2478                   *                           NaN                    NaN
8824                   *                           NaN                    NaN
15603                  #                           NaN                    NaN
19690                NaN                          22.0                  660.0

       total_drug_cost_ge65
759                     NaN
2478                    NaN
8824                    NaN
15603                   NaN
19690                 94.03
```

[98]:
```
# what's in here and what's missing?
benzo_pres_17.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 13161 entries, 759 to 25205770
Data columns (total 19 columns):
npi                              13161 non-null int64
nppes_provider_last_org_name     13161 non-null object
```

```
nppes_provider_first_name        13161 non-null object
specialty_description            13161 non-null object
description_flag                 13161 non-null object
drug_name                        13161 non-null object
generic_name                     13161 non-null object
bene_count                       5640 non-null float64
total_claim_count                13161 non-null int64
total_30_day_fill_count          13161 non-null float64
total_day_supply                 13161 non-null int64
total_drug_cost                  13161 non-null float64
bene_count_ge65                  1828 non-null float64
bene_count_ge65_suppress_flag    11333 non-null object
total_claim_count_ge65           7397 non-null float64
ge65_suppress_flag               5764 non-null object
total_30_day_fill_count_ge65     7397 non-null float64
total_day_supply_ge65            7397 non-null float64
total_drug_cost_ge65             7397 non-null float64
dtypes: float64(8), int64(3), object(8)
memory usage: 2.0+ MB
```

[100]:
```
# drop cols with a lot of NA - don't have a lot of time to dig into it
# total claim counts is probably what I'm after anyway, or maybe drug costs for␣
 ↪medicare utilization
benzo_pres_17_sub = benzo_pres_17.dropna(axis=1).drop('description_flag',␣
 ↪axis=1)
print(benzo_pres_17_sub.shape)
benzo_pres_17_sub.head()
```

```
(13161, 10)
```

[100]:

	npi	nppes_provider_last_org_name	nppes_provider_first_name	specialty_description	drug_name	generic_name	total_claim_count	total_30_day_fill_count	total_day_supply	total_drug_cost
759	1003002312	HOPKINS	PATRICIA	Rheumatology	ALPRAZOLAM	ALPRAZOLAM	21	25.0	735	213.58
2478	1003007477	ABDOW	KIMBERLY	Nurse Practitioner	ALPRAZOLAM	ALPRAZOLAM	63	75.0	2160	324.14
8824	1003023284	MCELROY	ALLEGRA	Nurse Practitioner	ALPRAZOLAM	ALPRAZOLAM	12			
15603	1003047473	HASSEY	SHERINE	Nurse Practitioner	ALPRAZOLAM	ALPRAZOLAM	17			
19690	1003062647	BEAUZILE	RONALD	Internal Medicine	ALPRAZOLAM	ALPRAZOLAM	14			

```
8824                      12.0              350           57.18
15603                     17.0              492           89.98
19690                     22.0              660           94.03
```

[101]:
```
for x in list(benzo_pres_17_sub.columns)[1:6]:
    benzo_pres_17_sub[x] = benzo_pres_17_sub[x].str.lower()
```

[102]:
```
benzo_pres_17_sub.head()
```

[102]:
```
             npi nppes_provider_last_org_name nppes_provider_first_name  \
759    1003002312                      hopkins                  patricia
2478   1003007477                        abdow                  kimberly
8824   1003023284                      mcelroy                   allegra
15603  1003047473                       hassey                   sherine
19690  1003062647                     beauzile                    ronald

      specialty_description   drug_name generic_name  total_claim_count  \
759             rheumatology  alprazolam   alprazolam                 21
2478     nurse practitioner  alprazolam   alprazolam                 63
8824     nurse practitioner  alprazolam   alprazolam                 12
15603    nurse practitioner  alprazolam   alprazolam                 17
19690     internal medicine  alprazolam   alprazolam                 14

       total_30_day_fill_count  total_day_supply  total_drug_cost
759                       25.0               735           213.58
2478                      75.0              2160           324.14
8824                      12.0               350            57.18
15603                     17.0               492            89.98
19690                     22.0               660            94.03
```

[103]:
```
opi_pres_17_npi_town_match.head()
```

[103]:
```
           npi nppes_provider_last_name nppes_provider_first_name       town
5   1003002312                   hopkins                  patricia     quincy
9   1003007477                     abdow                  kimberly  worcester
14  1003008244               rajbhandari                      ruma     boston
19  1003011610                       ray                     alaka     boston
25  1003012766                      kano                   zachary  brookline
```

[104]:
```
# assign 2017 benzodiazepine prescribers a town
print(benzo_pres_17_sub.shape)
print(opi_pres_17_npi_town_match.shape)
benzo_town_merge = benzo_pres_17_sub.merge(opi_pres_17_npi_town_match,␣
 ↪how="inner", on="npi", suffixes=["_benz", "_opi"])
print(benzo_town_merge.shape)
print(benzo_town_merge.columns)
benzo_town_merge.head()
```

```
(13161, 10)
(35678, 4)
```

```
(13138, 13)
Index(['npi', 'nppes_provider_last_org_name', 'nppes_provider_first_name_benz',
       'specialty_description', 'drug_name', 'generic_name',
       'total_claim_count', 'total_30_day_fill_count', 'total_day_supply',
       'total_drug_cost', 'nppes_provider_last_name',
       'nppes_provider_first_name_opi', 'town'],
      dtype='object')
```

[104]:

	npi	nppes_provider_last_org_name	nppes_provider_first_name_benz	specialty_description	drug_name	generic_name	total_claim_count	total_30_day_fill_count	total_day_supply	total_drug_cost	nppes_provider_last_name	nppes_provider_first_name_opi	town
0	1003002312	hopkins	patricia	rheumatology	alprazolam	alprazolam	21	25.0	735	213.58	hopkins	patricia	quincy
1	1003002312	hopkins	patricia	rheumatology	diazepam	diazepam	14	14.0	360	56.56	hopkins	patricia	quincy
2	1003002312	hopkins	patricia	rheumatology	lorazepam	lorazepam	129	139.0	4028	907.84	hopkins	patricia	quincy
3	1003007477	abdow	kimberly	nurse practitioner	alprazolam	alprazolam	63	75.0	2160	324.14	abdow	kimberly	worcester
4	1003007477	abdow	kimberly	nurse practitioner	diazepam	diazepam	97	97.6	2920	781.21	abdow	kimberly	worcester

[105]:
```
# merge checks - do the provider names match?
print(sum(benzo_town_merge['nppes_provider_last_org_name'] !=␣
 ↪benzo_town_merge['nppes_provider_last_name']))
print(sum(benzo_town_merge['nppes_provider_first_name_benz'] !=␣
 ↪benzo_town_merge['nppes_provider_first_name_opi']))
benzo_town_merge.drop(['nppes_provider_last_name',␣
 ↪'nppes_provider_first_name_opi'], axis=1, inplace=True)
```

```
0
0
```

Merge worked well - names match

[106]:
```
benzo_town_merge_17 = benzo_town_merge.copy()
```

Some basic exploration:

[108]:
```
benzo_town_merge.drop(['nppes_provider_last_org_name',␣
 ↪'nppes_provider_first_name_benz'], axis=1, inplace=True)
```

[109]:
```
benzo_town_merge.head()
```

[109]:

	npi	specialty_description	drug_name	generic_name	total_claim_count	total_30_day_fill_count	total_day_supply	total_drug_cost	town
0	1003002312	rheumatology	alprazolam	alprazolam	21	25.0	735	213.58	quincy
1	1003002312	rheumatology	diazepam	diazepam	14	14.0	360	56.56	quincy
2	1003002312	rheumatology	lorazepam	lorazepam	129	139.0	4028	907.84	quincy
3	1003007477	nurse practitioner	alprazolam	alprazolam	63	75.0	2160	324.14	worcester
4	1003007477	nurse practitioner	diazepam	diazepam	97	97.6	2920	781.21	worcester

[110]:
```
# add up claim info by town and drug:
benzo_town_sum = benzo_town_merge.groupby(['town', 'generic_name']).sum().
 ↪reset_index().drop('npi', axis=1)
print(benzo_town_sum.shape)
benzo_town_sum.head()
```

```
(729, 6)
```

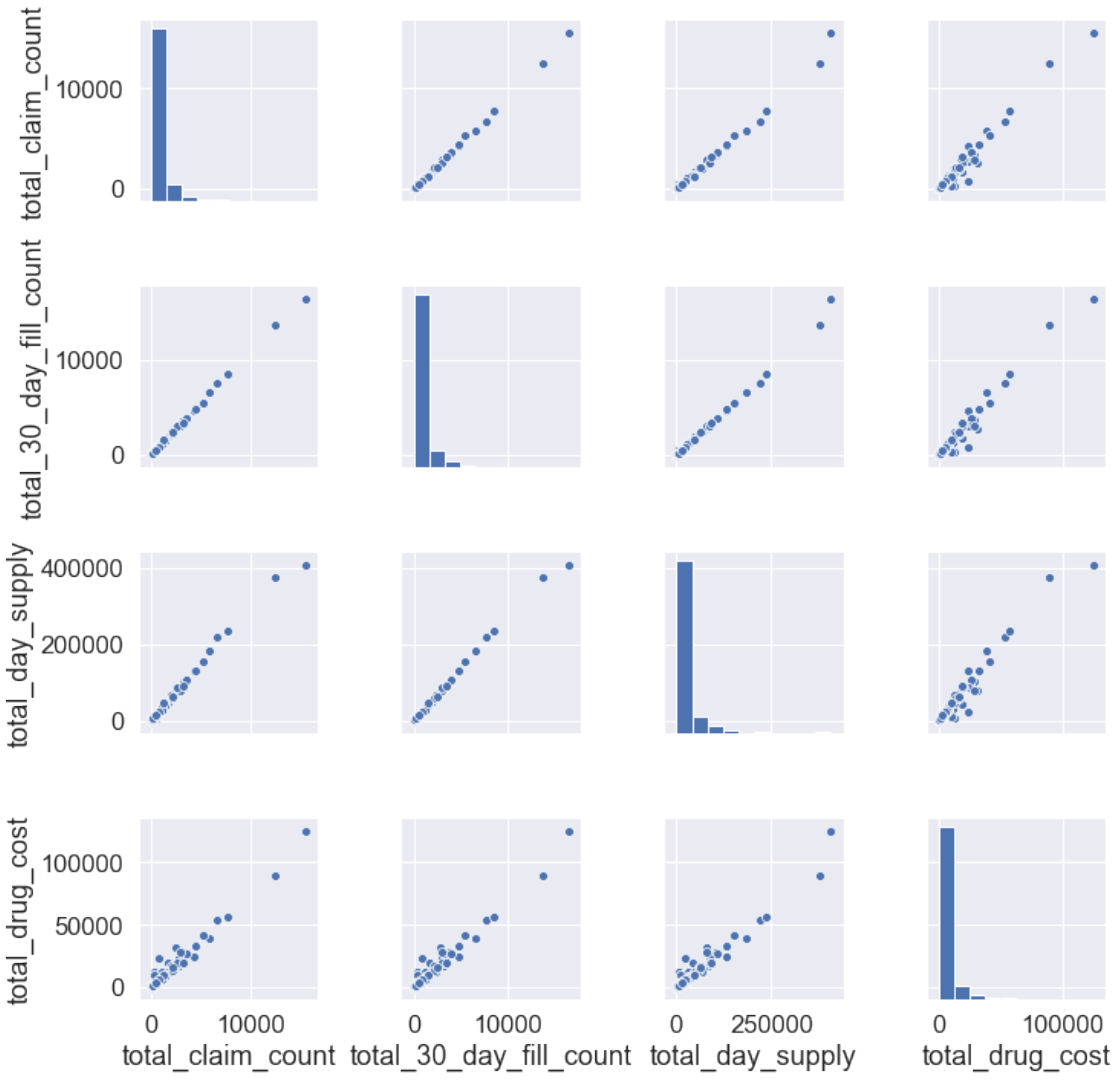
[110]:

	town	generic_name	total_claim_count	total_30_day_fill_count	total_day_supply	total_drug_cost
0	abington	alprazolam	275	309.5	8384	1627.81
1	abington	diazepam	82	84.0	1908	6791.29
2	abington	lorazepam	398	441.0	11206	2986.91
3	acton	alprazolam	235	269.0	6430	825.62

```
4              5038           881.19
```

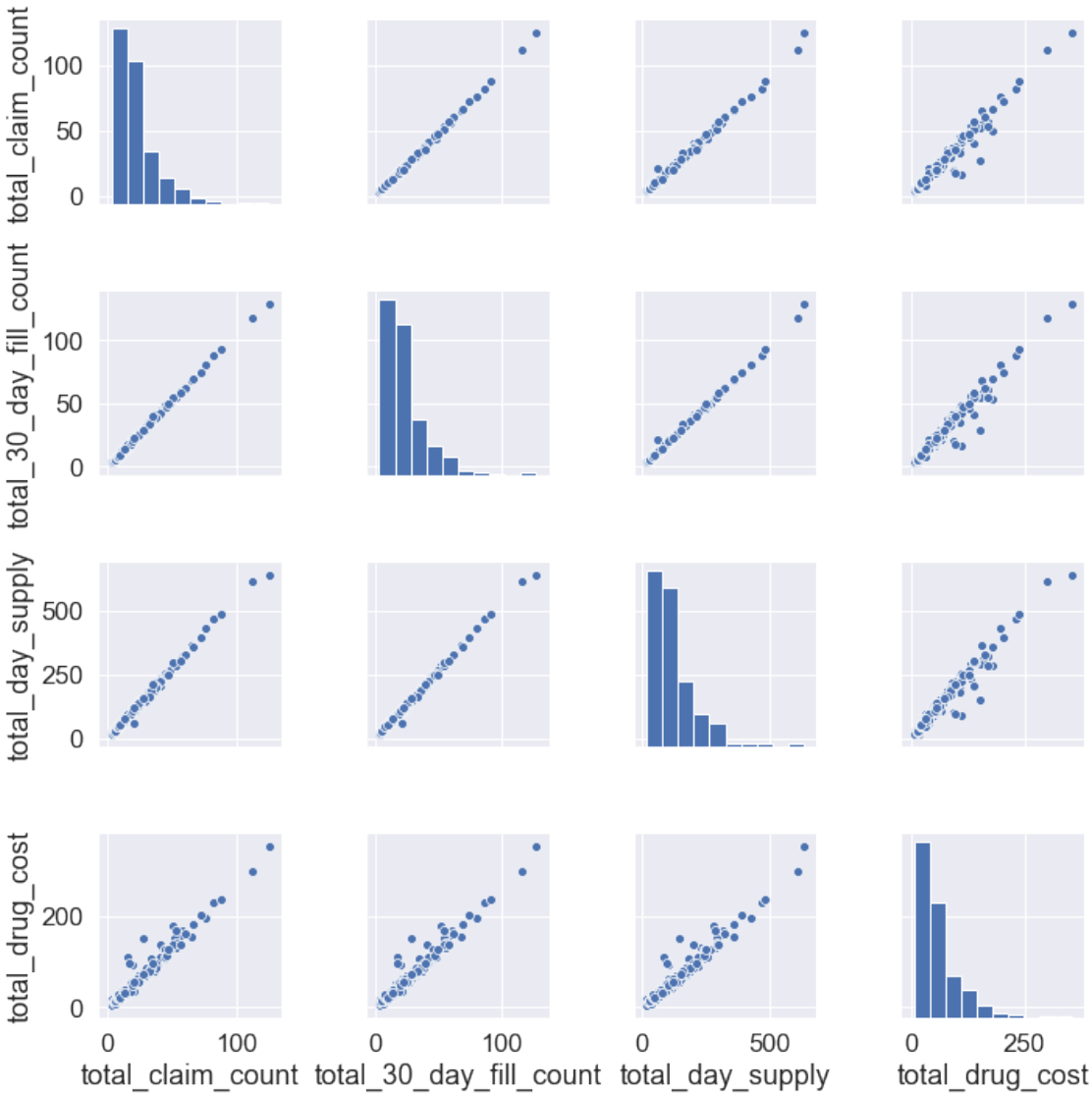
[111]:
```
sns.pairplot(benzo_town_sum[benzo_town_sum['generic_name'] == 'alprazolam'])
```

[111]:
```
<seaborn.axisgrid.PairGrid at 0x162d6b11668>
```



[112]:
```
# skewed distribution - take square root
sns.pairplot(benzo_town_sum[benzo_town_sum['generic_name'] == 'alprazolam'].
 ↪iloc[:, 2:].apply(np.sqrt))
```

[112]:
```
<seaborn.axisgrid.PairGrid at 0x162d7e550b8>
```



[113]:
```
benzo_town_sum_norm = benzo_town_sum.copy()
for x in list(benzo_town_sum_norm.columns)[2:]:
    benzo_town_sum_norm[x] = np.sqrt(benzo_town_sum_norm[x])
benzo_town_sum_norm.head()
```

[113]:

	town	generic_name	total_claim_count	total_30_day_fill_count	total_day_supply	total_drug_cost
0	abington	alprazolam	16.583124	17.592612	91.564185	40.346127
1	abington	diazepam	9.055385	9.165151		
2	abington	lorazepam	19.949937	21.000000		
3	acton	alprazolam	15.329710	16.401219		
4	acton	diazepam	13.453624	14.053469		

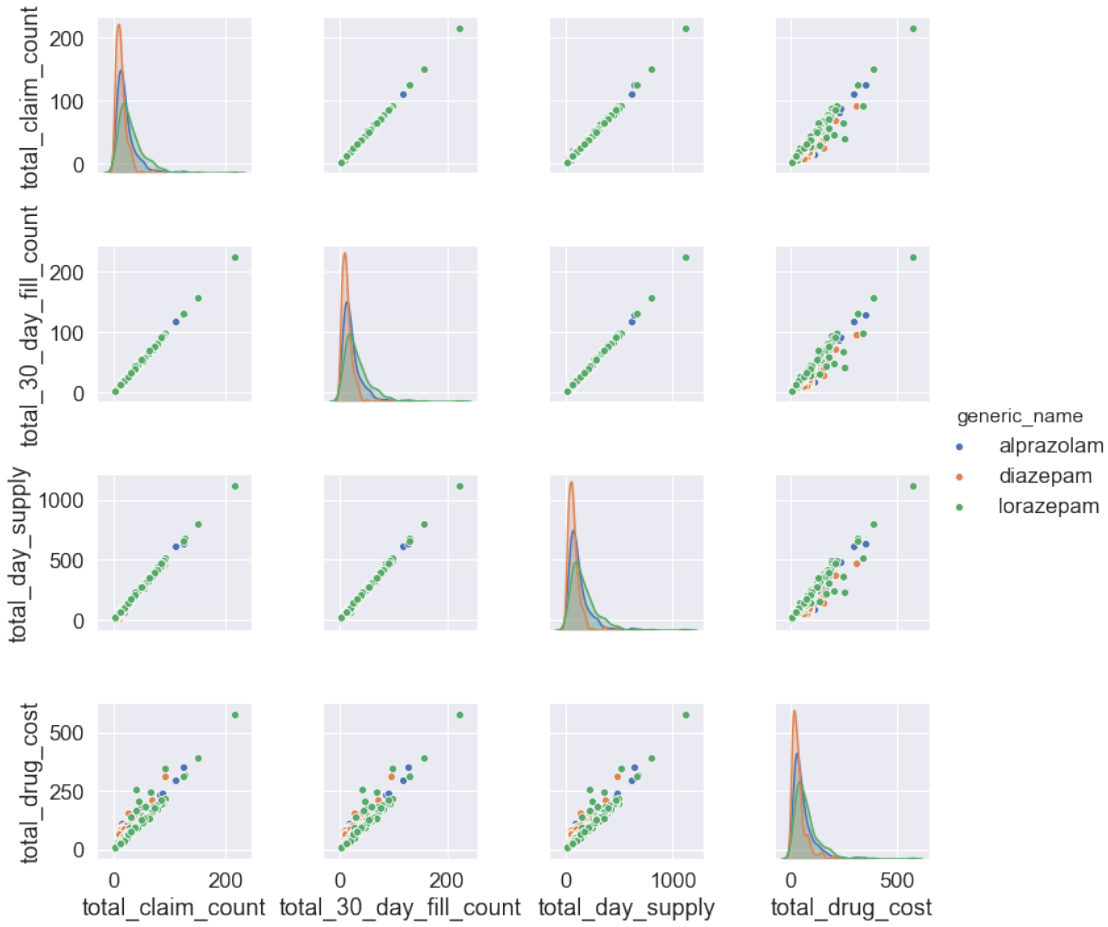
1	43.680659	82.409283
2	105.858396	54.652630
3	80.187281	28.733604
4	70.978870	29.684845

All of the variable columns (claim count, total 30 day fill count, total supply, and total drug cost) seem to be highly correlated with each other.

Is this true for all 3 drugs?

[114]:
```
sns.pairplot(benzo_town_sum_norm, hue="generic_name")
```

[114]:
```
<seaborn.axisgrid.PairGrid at 0x162de374748>
```



Yes, all of the prescription data is highly skewed and all 4 features are very correlated with each other, for all 3 drugs.

1.1 Clean older all-drug prescriber data

This is not the best way to do it, this was just a fast way to get throug these big datasets (2.7-3GB each). Basic strategy was to import each file one by one, pull out and clean the benzo prescriber data, and then remove the raw file from memory.

[115]:
```
del all_pres_17_raw
```

[116]:
```
all_pres_16_raw =  pd.read_csv("../../data/raw_data/
↪medicare_prescription_all_drugs/PartD_Prescriber_PUF_NPI_Drug_16.txt",␣
↪sep='\t')
```

[117]:
```
all_pres_16_raw.columns
```

[117]:
```
Index(['npi', 'nppes_provider_last_org_name', 'nppes_provider_first_name',
       'nppes_provider_city', 'nppes_provider_state', 'specialty_description',
       'description_flag', 'drug_name', 'generic_name', 'bene_count',
       'total_claim_count', 'total_30_day_fill_count', 'total_day_supply',
       'total_drug_cost', 'bene_count_ge65', 'bene_count_ge65_suppress_flag',
       'total_claim_count_ge65', 'ge65_suppress_flag',
       'total_30_day_fill_count_ge65', 'total_day_supply_ge65',
       'total_drug_cost_ge65'],
      dtype='object')
```

[118]:
```
all_pres_16_MA = all_pres_16_raw[all_pres_16_raw['nppes_provider_state'] ==␣
↪'MA']
all_pres_16_MA.head()
```

[118]:

	npi	nppes_provider_last_org_name	nppes_provider_first_name	nppes_provider_city	nppes_provider_state	specialty_description	description_flag	drug_name	generic_name	bene_count	...	total_30_day_fill_count	total_day_supply
713	1003002312	HOPKINS	PATRICIA	QUINCY	MA	Rheumatology	S	ACETAMINOPHEN-CODEINE	ACETAMINOPHEN WITH CODEINE	NaN	...	16.0	480
714	1003002312	HOPKINS	PATRICIA	QUINCY	MA	Rheumatology	S	ADVAIR DISKUS	FLUTICASONE/SALMETEROL	NaN	...	39.0	1170
715	1003002312	HOPKINS	PATRICIA	QUINCY	MA	Rheumatology	S	ALLOPURINOL	ALLOPURINOL	14.0	...	90.8	2673
716	1003002312	HOPKINS	PATRICIA	QUINCY	MA	Rheumatology	S	ALPRAZOLAM	ALPRAZOLAM	NaN	...	25.0	750
717	1003002312	HOPKINS	PATRICIA	QUINCY	MA	Rheumatology	S	AMLODIPINE BESYLATE	AMLODIPINE BESYLATE	14.0	...	125.0	3750

```
     total_drug_cost  bene_count_ge65  bene_count_ge65_suppress_flag  \
713           536.19              NaN                              *
714         13618.31              NaN                              *
715           469.08              NaN                              #
716           278.12              NaN                              #
717           421.39              NaN                              #

    total_claim_count_ge65  ge65_suppress_flag  total_30_day_fill_count_ge65  \
713                   16.0                 NaN                          16.0
714                   13.0                 NaN                          39.0
715                    NaN                   #                           NaN
716                    NaN                   #                           NaN
717                   66.0                 NaN                         102.0

     total_day_supply_ge65  total_drug_cost_ge65
713                  480.0               536.19
714                 1170.0             13618.31
715                    NaN                  NaN
716                    NaN                  NaN
717                 3060.0               384.86

[5 rows x 21 columns]
```

[119]:
```
alprazolam_16_pres = all_pres_16_MA[all_pres_16_MA['generic_name'].str.lower().
 ↪str.find('alprazolam') >= 0]
print(alprazolam_16_pres.shape)
display(alprazolam_16_pres['drug_name'].value_counts())
diazepam_16_pres = all_pres_16_MA[all_pres_16_MA['generic_name'].str.lower().
 ↪str.find('diazepam') >= 0]
print(diazepam_16_pres.shape)
display(diazepam_16_pres['drug_name'].value_counts())
lorazepam_16_pres = all_pres_16_MA[all_pres_16_MA['generic_name'].str.lower().
 ↪str.find('lorazepam') >= 0]
print(lorazepam_16_pres.shape)
display(lorazepam_16_pres['drug_name'].value_counts())
```

```
(3921, 21)

ALPRAZOLAM             3866
XANAX                    26
ALPRAZOLAM ER            22
ALPRAZOLAM ODT            3
ALPRAZOLAM XR             2
ALPRAZOLAM INTENSOL       1
XANAX XR                  1
Name: drug_name, dtype: int64
```

```
(2583, 21)

DIAZEPAM    2567
VALIUM        16
Name: drug_name, dtype: int64


(6654, 21)

LORAZEPAM             6633
LORAZEPAM INTENSOL      11
ATIVAN                  10
Name: drug_name, dtype: int64
```

[120]:
```
benzo_pres_16 = pd.concat([alprazolam_16_pres, diazepam_16_pres,␣
↪lorazepam_16_pres]).drop(['nppes_provider_city', 'nppes_provider_state'],␣
↪axis=1)
display(benzo_pres_16.head())
display(benzo_pres_16.info())
```

```
             npi nppes_provider_last_org_name nppes_provider_first_name  \
716    1003002312                     HOPKINS                  PATRICIA
2446   1003007477                       ABDOW                  KIMBERLY
4518   1003011610                         RAY                     ALAKA
8597   1003023284                     MCELROY                   ALLEGRA
15671  1003047473                      HASSEY                   SHERINE

      specialty_description description_flag   drug_name generic_name  \
716             Rheumatology                S  ALPRAZOLAM   ALPRAZOLAM
2446     Nurse Practitioner                S  ALPRAZOLAM   ALPRAZOLAM
4518      Internal Medicine                S  ALPRAZOLAM   ALPRAZOLAM
8597     Nurse Practitioner                S  ALPRAZOLAM   ALPRAZOLAM
15671    Nurse Practitioner                S  ALPRAZOLAM   ALPRAZOLAM

       bene_count  total_claim_count  total_30_day_fill_count  \
716           NaN                 21                     25.0
2446          NaN                 26                     30.0
4518          NaN                 14                     25.0
8597          NaN                 13                     13.0
15671        11.0                 16                     22.0

       total_day_supply  total_drug_cost  bene_count_ge65  \
716                 750           278.12              NaN
2446                876           253.86              NaN
4518                633            34.17              NaN
8597                370            21.10              NaN
15671               656            76.62              NaN
```

```
      bene_count_ge65_suppress_flag  total_claim_count_ge65  \
716                               #                     NaN
2446                              *                     NaN
4518                              *                    14.0
8597                              *                     NaN
15671                             #                     NaN

      ge65_suppress_flag  total_30_day_fill_count_ge65  total_day_supply_ge65  \
716                    #                           NaN                    NaN
2446                   *                           NaN                    NaN
4518                 NaN                          25.0                  633.0
8597                   *                           NaN                    NaN
15671                  #                           NaN                    NaN

       total_drug_cost_ge65
716                     NaN
2446                    NaN
4518                  34.17
8597                    NaN
15671                   NaN


<class 'pandas.core.frame.DataFrame'>
Int64Index: 13158 entries, 716 to 24960927
Data columns (total 19 columns):
npi                              13158 non-null int64
nppes_provider_last_org_name     13158 non-null object
nppes_provider_first_name        13158 non-null object
specialty_description            13158 non-null object
description_flag                 13158 non-null object
drug_name                        13158 non-null object
generic_name                     13158 non-null object
bene_count                       5613 non-null float64
total_claim_count                13158 non-null int64
total_30_day_fill_count          13158 non-null float64
total_day_supply                 13158 non-null int64
total_drug_cost                  13158 non-null float64
bene_count_ge65                  1826 non-null float64
bene_count_ge65_suppress_flag    11332 non-null object
total_claim_count_ge65           7275 non-null float64
ge65_suppress_flag               5883 non-null object
total_30_day_fill_count_ge65     7275 non-null float64
total_day_supply_ge65            7275 non-null float64
total_drug_cost_ge65             7275 non-null float64
dtypes: float64(8), int64(3), object(8)
memory usage: 2.0+ MB

None
```

[121]:
```
benzo_pres_16_sub = benzo_pres_16.dropna(axis=1).drop('description_flag',␣
↪axis=1)
print(benzo_pres_16_sub.shape)
benzo_pres_16_sub.head()
```

```
(13158, 10)
```

[121]:

	npi	nppes_provider_last_org_name	nppes_provider_first_name	specialty_description	drug_name	generic_name	total_claim_count	total_30_day_fill_count	total_day_supply	total_drug_cost
716	1003002312	HOPKINS	PATRICIA	Rheumatology	ALPRAZOLAM	ALPRAZOLAM	21	25.0	750	278.12
2446	1003007477	ABDOW	KIMBERLY	Nurse Practitioner	ALPRAZOLAM	ALPRAZOLAM	26	30.0	876	253.86
4518	1003011610	RAY	ALAKA	Internal Medicine	ALPRAZOLAM	ALPRAZOLAM	14	25.0	633	34.17
8597	1003023284	MCELROY	ALLEGRA	Nurse Practitioner	ALPRAZOLAM	ALPRAZOLAM	13	13.0	370	21.10
15671	1003047473	HASSEY	SHERINE	Nurse Practitioner	ALPRAZOLAM	ALPRAZOLAM	16	22.0	656	76.62

[122]:
```
for x in list(benzo_pres_16_sub.columns)[1:6]:
    benzo_pres_16_sub[x] = benzo_pres_16_sub[x].str.lower()
```

[123]:
```
opi_pres_16_npi_town_match =␣
↪med_opi_pres_no_town_dup[med_opi_pres_no_town_dup['year'] == 2016][['npi',␣
↪'nppes_provider_last_name', 'nppes_provider_first_name', 'town']].copy()
print(benzo_pres_16_sub.shape)
print(opi_pres_16_npi_town_match.shape)
benzo_town_merge_16 = benzo_pres_16_sub.merge(opi_pres_16_npi_town_match,␣
↪how="inner", on="npi", suffixes=["_benz", "_opi"])
print(benzo_town_merge_16.shape)
print(benzo_town_merge_16.columns)
benzo_town_merge_16.head()
```

```
(13158, 10)
(34977, 4)
(13132, 13)
Index(['npi', 'nppes_provider_last_org_name', 'nppes_provider_first_name_benz',
```

```
       'specialty_description', 'drug_name', 'generic_name',
       'total_claim_count', 'total_30_day_fill_count', 'total_day_supply',
       'total_drug_cost', 'nppes_provider_last_name',
       'nppes_provider_first_name_opi', 'town'],
      dtype='object')
```

[123]:

	npi	nppes_provider_last_org_name	nppes_provider_first_name_benz	specialty_description	drug_name	generic_name	total_claim_count	total_30_day_fill_count	total_day_supply	total_drug_cost	nppes_provider_last_name	nppes_provider_first_name_opi	town
0	1003002312	hopkins	patricia	rheumatology	alprazolam	alprazolam	21	25.0	750	278.12	hopkins	patricia	quincy
1	1003002312	hopkins	patricia	rheumatology	diazepam	diazepam	19	19.0	507	89.89	hopkins	patricia	quincy
2	1003002312	hopkins	patricia	rheumatology	lorazepam	lorazepam	142	150.0	4380	782.41	hopkins	patricia	quincy
3	1003007477	abdow	kimberly	nurse practitioner	alprazolam	alprazolam	26	30.0	876	253.86	abdow	kimberly	worcester
4	1003007477	abdow	kimberly	nurse practitioner	diazepam	diazepam	25	25.0	750	121.84	abdow	kimberly	worcester

[124]:
```
print(sum(benzo_town_merge_16['nppes_provider_last_org_name'] !=␣
 ↪benzo_town_merge_16['nppes_provider_last_name']))
print(sum(benzo_town_merge_16['nppes_provider_first_name_benz'] !=␣
 ↪benzo_town_merge_16['nppes_provider_first_name_opi']))
benzo_town_merge_16.drop(['nppes_provider_last_name',␣
 ↪'nppes_provider_first_name_opi'], axis=1, inplace=True)
```

```
0
0
```

[125]:
```
print(benzo_town_merge_17.shape)
print(benzo_town_merge_16.shape)
display(benzo_town_merge_17.head())
display(benzo_town_merge_16.head())
```

```
(13138, 11)
(13132, 11)
```

```
         npi nppes_provider_last_org_name nppes_provider_first_name_benz  \
0  1003002312                      hopkins                       patricia
1  1003002312                      hopkins                       patricia
2  1003002312                      hopkins                       patricia
3  1003007477                        abdow                       kimberly
4  1003007477                        abdow                       kimberly

  specialty_description   drug_name generic_name  total_claim_count  \
0          rheumatology  alprazolam   alprazolam                 21
1          rheumatology   diazepam     diazepam                 14
2          rheumatology   lorazepam    lorazepam                129
3    nurse practitioner  alprazolam   alprazolam                 63
4    nurse practitioner   diazepam     diazepam                 97

   total_30_day_fill_count  total_day_supply  total_drug_cost       town
0                     25.0               735           213.58     quincy
1                     14.0               360            56.56     quincy
2                    139.0              4028           907.84     quincy
3                     75.0              2160           324.14  worcester
4                     97.6              2920           781.21  worcester
```

```
         npi nppes_provider_last_org_name nppes_provider_first_name_benz  \
0  1003002312                      hopkins                       patricia
1  1003002312                      hopkins                       patricia
2  1003002312                      hopkins                       patricia
3  1003007477                        abdow                       kimberly
4  1003007477                        abdow                       kimberly

  specialty_description   drug_name generic_name  total_claim_count  \
0          rheumatology  alprazolam   alprazolam                 21
1          rheumatology   diazepam     diazepam                 19
2          rheumatology   lorazepam    lorazepam                142
3    nurse practitioner  alprazolam   alprazolam                 26
4    nurse practitioner   diazepam     diazepam                 25

   total_30_day_fill_count  total_day_supply  total_drug_cost       town
0                     25.0               750           278.12     quincy
1                     19.0               507            89.89     quincy
2                    150.0              4380           782.41     quincy
3                     30.0               876           253.86  worcester
4                     25.0               750           121.84  worcester
```

2016 done - 2015 next

[126]:
```
del all_pres_16_raw
```

[127]:
```
all_pres_15_raw =  pd.read_csv("../../data/raw_data/
↪medicare_prescription_all_drugs/PartD_Prescriber_PUF_NPI_Drug_15.txt",␣
↪sep='\t')
```

[128]:
```
all_pres_15_MA = all_pres_15_raw[all_pres_15_raw['nppes_provider_state'] ==␣
↪'MA']
alprazolam_15_pres = all_pres_15_MA[all_pres_15_MA['generic_name'].str.lower().
↪str.find('alprazolam') >= 0]
print(alprazolam_15_pres.shape)
display(alprazolam_15_pres['drug_name'].value_counts())
diazepam_15_pres = all_pres_15_MA[all_pres_15_MA['generic_name'].str.lower().
↪str.find('diazepam') >= 0]
print(diazepam_15_pres.shape)
display(diazepam_15_pres['drug_name'].value_counts())
lorazepam_15_pres = all_pres_15_MA[all_pres_15_MA['generic_name'].str.lower().
↪str.find('lorazepam') >= 0]
print(lorazepam_15_pres.shape)
display(lorazepam_15_pres['drug_name'].value_counts())
```

```
(3793, 21)

ALPRAZOLAM        3741
ALPRAZOLAM ER       24
XANAX               21
ALPRAZOLAM ODT       3
ALPRAZOLAM XR        2
XANAX XR             2
Name: drug_name, dtype: int64


(2588, 21)

DIAZEPAM    2575
VALIUM        13
Name: drug_name, dtype: int64


(6463, 21)

LORAZEPAM             6419
LORAZEPAM INTENSOL      31
ATIVAN                  13
Name: drug_name, dtype: int64
```

[129]:
```

```

```
benzo_pres_15 = pd.concat([alprazolam_15_pres, diazepam_15_pres,␣
 ↪lorazepam_15_pres]).drop(['nppes_provider_city', 'nppes_provider_state'],␣
 ↪axis=1)
display(benzo_pres_15.info())
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 12844 entries, 652 to 24521569
Data columns (total 19 columns):
npi                                12844 non-null int64
nppes_provider_last_org_name       12844 non-null object
nppes_provider_first_name          12844 non-null object
specialty_description              12844 non-null object
description_flag                   12844 non-null object
drug_name                          12844 non-null object
generic_name                       12844 non-null object
bene_count                         5267 non-null float64
total_claim_count                  12844 non-null int64
total_30_day_fill_count            12844 non-null float64
total_day_supply                   12844 non-null int64
total_drug_cost                    12844 non-null float64
bene_count_ge65                    1809 non-null float64
bene_count_ge65_suppress_flag      11035 non-null object
total_claim_count_ge65             7110 non-null float64
ge65_suppress_flag                 5734 non-null object
total_30_day_fill_count_ge65       7110 non-null float64
total_day_supply_ge65              7110 non-null float64
total_drug_cost_ge65               7110 non-null float64
dtypes: float64(8), int64(3), object(8)
memory usage: 2.0+ MB

None
```

[130]:
```
benzo_pres_15_sub = benzo_pres_15.dropna(axis=1).drop('description_flag',␣
 ↪axis=1)
print(benzo_pres_15_sub.shape)
benzo_pres_15_sub.head()
```

```
(12844, 10)
```

[130]:

	npi	nppes_provider_last_org_name	nppes_provider_first_name	\
652	1003002312	HOPKINS	PATRICIA	
14460	1003044272	BENDER	ELISE	
19502	1003062647	BEAUZILE	RONALD	
43176	1003164310	SIMPSON	SOMATRA	
50083	1003218041	LYONS	PATRICK	

```
      specialty_description   drug_name generic_name  total_claim_count  \
652       Internal Medicine  ALPRAZOLAM   ALPRAZOLAM                 26
14460       Family Practice  ALPRAZOLAM   ALPRAZOLAM                 15
19502     Internal Medicine  ALPRAZOLAM   ALPRAZOLAM                 16
43176    Nurse Practitioner  ALPRAZOLAM   ALPRAZOLAM                184
50083    Nurse Practitioner  ALPRAZOLAM   ALPRAZOLAM                 16

       total_30_day_fill_count  total_day_supply  total_drug_cost
652                       32.0               960           263.50
14460                     19.0               439            51.41
19502                     16.0               452           241.53
43176                    184.0              5450          1092.11
50083                     16.0               418            60.67
```

[131]:
```
for x in list(benzo_pres_15_sub.columns)[1:6]:
    benzo_pres_15_sub[x] = benzo_pres_15_sub[x].str.lower()
```

[132]:
```
opi_pres_15_npi_town_match =␣
 ↪med_opi_pres_no_town_dup[med_opi_pres_no_town_dup['year'] == 2015][['npi',␣
 ↪'nppes_provider_last_name', 'nppes_provider_first_name', 'town']].copy()
print(benzo_pres_15_sub.shape)
print(opi_pres_15_npi_town_match.shape)
benzo_town_merge_15 = benzo_pres_15_sub.merge(opi_pres_15_npi_town_match,␣
 ↪how="inner", on="npi", suffixes=["_benz", "_opi"])
print(benzo_town_merge_15.shape)
print(benzo_town_merge_15.columns)
benzo_town_merge_15.head()
```

```
(12844, 10)
(34026, 4)
(12820, 13)
Index(['npi', 'nppes_provider_last_org_name', 'nppes_provider_first_name_benz',
       'specialty_description', 'drug_name', 'generic_name',
       'total_claim_count', 'total_30_day_fill_count', 'total_day_supply',
       'total_drug_cost', 'nppes_provider_last_name',
       'nppes_provider_first_name_opi', 'town'],
      dtype='object')
```

[132]:
```
          npi nppes_provider_last_org_name nppes_provider_first_name_benz  \
0  1003002312                      hopkins                       patricia
1  1003002312                      hopkins                       patricia
2  1003002312                      hopkins                       patricia
3  1003044272                       bender                          elise
4  1003044272                       bender                          elise

  specialty_description   drug_name generic_name  total_claim_count  \
0     internal medicine  alprazolam   alprazolam                 26
1     internal medicine    diazepam     diazepam                 17
```

```
2     internal medicine   lorazepam    lorazepam                122
3       family practice  alprazolam   alprazolam                 15
4       family practice   lorazepam    lorazepam                 17

   total_30_day_fill_count  total_day_supply  total_drug_cost  \
0                     32.0               960           263.50
1                     17.0               501            65.54
2                    132.0              3773           707.43
3                     19.0               439            51.41
4                     21.0               538            43.52

  nppes_provider_last_name nppes_provider_first_name_opi       town
0                  hopkins                      patricia     quincy
1                  hopkins                      patricia     quincy
2                  hopkins                      patricia     quincy
3                   bender                         elise  braintree
4                   bender                         elise  braintree
```

[133]:
```
print(sum(benzo_town_merge_15['nppes_provider_last_org_name'] !=␣
 ↪benzo_town_merge_15['nppes_provider_last_name']))
print(sum(benzo_town_merge_15['nppes_provider_first_name_benz'] !=␣
 ↪benzo_town_merge_15['nppes_provider_first_name_opi']))
benzo_town_merge_15.drop(['nppes_provider_last_name',␣
 ↪'nppes_provider_first_name_opi'], axis=1, inplace=True)
```

```
0
0
```

[134]:
```
print(benzo_town_merge_17.shape)
print(benzo_town_merge_16.shape)
print(benzo_town_merge_15.shape)
```

```
(13138, 11)
(13132, 11)
(12820, 11)
```

[135]:
```
del all_pres_15_raw
```

[136]:
```
all_pres_14_raw =  pd.read_csv("../../data/raw_data/
 ↪medicare_prescription_all_drugs/PartD_Prescriber_PUF_NPI_Drug_14.txt",␣
 ↪sep='\t')
```

[137]:
```
all_pres_14_MA = all_pres_14_raw[all_pres_14_raw['nppes_provider_state'] ==␣
 ↪'MA']
alprazolam_14_pres = all_pres_14_MA[all_pres_14_MA['generic_name'].str.lower().
 ↪str.find('alprazolam') >= 0]
print(alprazolam_14_pres.shape)
display(alprazolam_14_pres['drug_name'].value_counts())
```

```
diazepam_14_pres = all_pres_14_MA[all_pres_14_MA['generic_name'].str.lower().
 ↪str.find('diazepam') >= 0]
print(diazepam_14_pres.shape)
display(diazepam_14_pres['drug_name'].value_counts())
lorazepam_14_pres = all_pres_14_MA[all_pres_14_MA['generic_name'].str.lower().
 ↪str.find('lorazepam') >= 0]
print(lorazepam_14_pres.shape)
display(lorazepam_14_pres['drug_name'].value_counts())
```

```
(3760, 21)

ALPRAZOLAM             3708
ALPRAZOLAM ER            24
XANAX                    19
ALPRAZOLAM XR             4
ALPRAZOLAM ODT            3
ALPRAZOLAM INTENSOL       1
XANAX XR                  1
Name: drug_name, dtype: int64


(2591, 21)

DIAZEPAM           2580
VALIUM               10
DIASTAT ACUDIAL       1
Name: drug_name, dtype: int64


(6293, 21)

LORAZEPAM             6274
ATIVAN                  10
LORAZEPAM INTENSOL       9
Name: drug_name, dtype: int64
```

[138]:
```
benzo_pres_14 = pd.concat([alprazolam_14_pres, diazepam_14_pres,␣
 ↪lorazepam_14_pres]).drop(['nppes_provider_city', 'nppes_provider_state'],␣
 ↪axis=1)
display(benzo_pres_14.info())
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 12644 entries, 610 to 24117229
Data columns (total 19 columns):
npi                              12644 non-null int64
nppes_provider_last_org_name     12644 non-null object
nppes_provider_first_name        12644 non-null object
```

```
specialty_description            12644 non-null object
description_flag                 12644 non-null object
drug_name                        12644 non-null object
generic_name                     12644 non-null object
bene_count                       5009 non-null float64
total_claim_count                12644 non-null int64
total_30_day_fill_count          12644 non-null float64
total_day_supply                 12644 non-null int64
total_drug_cost                  12644 non-null float64
bene_count_ge65                  1746 non-null float64
bene_count_ge65_suppress_flag    10898 non-null object
total_claim_count_ge65           6850 non-null float64
ge65_suppress_flag               5794 non-null object
total_30_day_fill_count_ge65     6850 non-null float64
total_day_supply_ge65            6850 non-null float64
total_drug_cost_ge65             6850 non-null float64
dtypes: float64(8), int64(3), object(8)
memory usage: 1.9+ MB

None
```

[139]:
```
benzo_pres_14_sub = benzo_pres_14.dropna(axis=1).drop('description_flag',␣
 ↪axis=1)
print(benzo_pres_14_sub.shape)
benzo_pres_14_sub.head()
```

```
(12644, 10)
```

[139]:

	npi	nppes_provider_last_org_name	nppes_provider_first_name	\
610	1003002312	HOPKINS	PATRICIA	
8082	1003023284	MCELROY	ALLEGRA	
18800	1003062647	BEAUZILE	RONALD	
25590	1003086976	KHERA	VANDANA	
39132	1003164310	SIMPSON	SOMATRA	

	specialty_description	drug_name	generic_name	total_claim_count	\
610	Internal Medicine	ALPRAZOLAM	ALPRAZOLAM	22	
8082	Nurse Practitioner	ALPRAZOLAM	ALPRAZOLAM	11	
18800	Internal Medicine	ALPRAZOLAM	ALPRAZOLAM	16	
25590	Internal Medicine	ALPRAZOLAM	ALPRAZOLAM	22	
39132	Nurse Practitioner	ALPRAZOLAM	ALPRAZOLAM	82	

	total_30_day_fill_count	total_day_supply	total_drug_cost
610	28.0	840	191.83
8082	11.0	320	58.73
18800	16.0	448	116.43
25590	22.0	514	82.94

```
39132                    83.5              2348           462.21
```

[140]:
```
for x in list(benzo_pres_14_sub.columns)[1:6]:
    benzo_pres_14_sub[x] = benzo_pres_14_sub[x].str.lower()
```

[141]:
```
opi_pres_14_npi_town_match =␣
 ↪med_opi_pres_no_town_dup[med_opi_pres_no_town_dup['year'] == 2014][['npi',␣
 ↪'nppes_provider_last_name', 'nppes_provider_first_name', 'town']].copy()
print(benzo_pres_14_sub.shape)
print(opi_pres_14_npi_town_match.shape)
benzo_town_merge_14 = benzo_pres_14_sub.merge(opi_pres_14_npi_town_match,␣
 ↪how="inner", on="npi", suffixes=["_benz", "_opi"])
print(benzo_town_merge_14.shape)
print(benzo_town_merge_14.columns)
benzo_town_merge_14.head()
```

```
(12644, 10)
(33329, 4)
(12621, 13)
Index(['npi', 'nppes_provider_last_org_name', 'nppes_provider_first_name_benz',
       'specialty_description', 'drug_name', 'generic_name',
       'total_claim_count', 'total_30_day_fill_count', 'total_day_supply',
       'total_drug_cost', 'nppes_provider_last_name',
       'nppes_provider_first_name_opi', 'town'],
      dtype='object')
```

[141]:

	npi	nppes_provider_last_org_name	nppes_provider_first_name_benz	specialty_description	drug_name	generic_name	total_claim_count	total_30_day_fill_count	total_day_supply	total_drug_cost	nppes_provider_last_name	nppes_provider_first_name_opi	town
0	1003002312	hopkins	patricia	internal medicine	alprazolam	alprazolam	22	28.0	840	191.83			
1	1003002312	hopkins	patricia	internal medicine	diazepam	diazepam	26	26.0	755	166.64			
2	1003002312	hopkins	patricia	internal medicine	lorazepam	lorazepam	107	114.0	3393	676.93			
3	1003023284	mcelroy	allegra	nurse practitioner	alprazolam	alprazolam	11	11.0	320	58.73			
4	1003023284	mcelroy	allegra	nurse practitioner	diazepam	diazepam	17	17.0	481	55.05			

```
0                  hopkins                       patricia   quincy
1                  hopkins                       patricia   quincy
2                  hopkins                       patricia   quincy
3                  mcelroy                        allegra  mashpee
4                  mcelroy                        allegra  mashpee
```

[142]:
```
print(sum(benzo_town_merge_14['nppes_provider_last_org_name'] !=␣
 ↪benzo_town_merge_14['nppes_provider_last_name']))
print(sum(benzo_town_merge_14['nppes_provider_first_name_benz'] !=␣
 ↪benzo_town_merge_14['nppes_provider_first_name_opi']))
benzo_town_merge_14.drop(['nppes_provider_last_name',␣
 ↪'nppes_provider_first_name_opi'], axis=1, inplace=True)
```

```
0
0
```

[143]:
```
print(benzo_town_merge_17.shape)
print(benzo_town_merge_16.shape)
print(benzo_town_merge_15.shape)
print(benzo_town_merge_14.shape)
```

```
(13138, 11)
(13132, 11)
(12820, 11)
(12621, 11)
```

2015 done - 2014 now

[144]:
```
del all_pres_14_raw
```

[145]:
```
all_pres_13_raw =  pd.read_csv("../../data/raw_data/
 ↪medicare_prescription_all_drugs/PartD_Prescriber_PUF_NPI_Drug_13.txt",␣
 ↪sep='\t')
```

[146]:
```
all_pres_13_MA = all_pres_13_raw[all_pres_13_raw['nppes_provider_state'] ==␣
 ↪'MA']
alprazolam_13_pres = all_pres_13_MA[all_pres_13_MA['generic_name'].str.lower().
 ↪str.find('alprazolam') >= 0]
print(alprazolam_13_pres.shape)
display(alprazolam_13_pres['drug_name'].value_counts())
diazepam_13_pres = all_pres_13_MA[all_pres_13_MA['generic_name'].str.lower().
 ↪str.find('diazepam') >= 0]
print(diazepam_13_pres.shape)
display(diazepam_13_pres['drug_name'].value_counts())
lorazepam_13_pres = all_pres_13_MA[all_pres_13_MA['generic_name'].str.lower().
 ↪str.find('lorazepam') >= 0]
print(lorazepam_13_pres.shape)
display(lorazepam_13_pres['drug_name'].value_counts())
```

```
(3558, 21)

ALPRAZOLAM             3509
ALPRAZOLAM ER            26
XANAX                    14
ALPRAZOLAM XR             4
ALPRAZOLAM ODT            2
ALPRAZOLAM INTENSOL       2
XANAX XR                  1
Name: drug_name, dtype: int64


(2537, 21)

DIAZEPAM    2529
VALIUM         8
Name: drug_name, dtype: int64


(6088, 21)

LORAZEPAM             6069
LORAZEPAM INTENSOL      12
ATIVAN                   7
Name: drug_name, dtype: int64
```

[147]:
```
benzo_pres_13 = pd.concat([alprazolam_13_pres, diazepam_13_pres,␣
 ↪lorazepam_13_pres]).drop(['nppes_provider_city', 'nppes_provider_state'],␣
 ↪axis=1)
display(benzo_pres_13.info())
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 12183 entries, 551 to 23642799
Data columns (total 19 columns):
npi                             12183 non-null int64
nppes_provider_last_org_name    12183 non-null object
nppes_provider_first_name       12183 non-null object
specialty_description           12183 non-null object
description_flag                12183 non-null object
drug_name                       12183 non-null object
generic_name                    12183 non-null object
bene_count                      4763 non-null float64
total_claim_count               12183 non-null int64
total_30_day_fill_count         12183 non-null float64
total_day_supply                12183 non-null int64
total_drug_cost                 12183 non-null float64
bene_count_ge65                 1696 non-null float64
```

```
bene_count_ge65_suppress_flag    10487 non-null object
total_claim_count_ge65           6628 non-null float64
ge65_suppress_flag               5555 non-null object
total_30_day_fill_count_ge65     6628 non-null float64
total_day_supply_ge65            6628 non-null float64
total_drug_cost_ge65             6628 non-null float64
dtypes: float64(8), int64(3), object(8)
memory usage: 1.9+ MB

None
```

[148]:
```
benzo_pres_13_sub = benzo_pres_13.dropna(axis=1).drop('description_flag',␣
 ↪axis=1)
print(benzo_pres_13_sub.shape)
benzo_pres_13_sub.head()
```

```
(12183, 10)
```

[148]:

	npi	nppes_provider_last_org_name	nppes_provider_first_name	\
551	1003002312	HOPKINS	PATRICIA	
9410	1003030586	KSHEERSAGAR	PANKAJ	
12142	1003042441	MANZO	MARK	
17125	1003062647	BEAUZILE	RONALD	
23373	1003086976	KHERA	VANDANA	

	specialty_description	drug_name	generic_name	total_claim_count	\
551	Internal Medicine	ALPRAZOLAM	ALPRAZOLAM	29	
9410	Family Practice	ALPRAZOLAM	ALPRAZOLAM	15	
12142	Physician Assistant	ALPRAZOLAM	ALPRAZOLAM	17	
17125	Internal Medicine	ALPRAZOLAM	ALPRAZOLAM	18	
23373	Internal Medicine	ALPRAZOLAM	ALPRAZOLAM	17	

	total_30_day_fill_count	total_day_supply	total_drug_cost
551	35.0	947	183.70
9410	15.0	450	59.09
12142	17.0	500	77.42
17125	18.0	481	113.03
23373	17.0	495	72.85

[149]:
```
for x in list(benzo_pres_13_sub.columns)[1:6]:
    benzo_pres_13_sub[x] = benzo_pres_13_sub[x].str.lower()
```

[150]:
```
opi_pres_13_npi_town_match =␣
 ↪med_opi_pres_no_town_dup[med_opi_pres_no_town_dup['year'] == 2013][['npi',␣
 ↪'nppes_provider_last_name', 'nppes_provider_first_name', 'town']].copy()
print(benzo_pres_13_sub.shape)
print(opi_pres_13_npi_town_match.shape)
```

```
benzo_town_merge_13 = benzo_pres_13_sub.merge(opi_pres_13_npi_town_match,␣
 ↪how="inner", on="npi", suffixes=["_benz", "_opi"])
print(benzo_town_merge_13.shape)
print(benzo_town_merge_13.columns)
benzo_town_merge_13.head()
```

```
(12183, 10)
(32734, 4)
(12162, 13)
Index(['npi', 'nppes_provider_last_org_name', 'nppes_provider_first_name_benz',
       'specialty_description', 'drug_name', 'generic_name',
       'total_claim_count', 'total_30_day_fill_count', 'total_day_supply',
       'total_drug_cost', 'nppes_provider_last_name',
       'nppes_provider_first_name_opi', 'town'],
      dtype='object')
```

[150]:

	npi	nppes_provider_last_org_name	nppes_provider_first_name_benz	specialty_description	drug_name	generic_name	total_claim_count	total_30_day_fill_count	total_day_supply	total_drug_cost	nppes_provider_last_name	nppes_provider_first_name_opi	town
0	1003002312	hopkins	patricia	internal medicine	alprazolam	alprazolam	29	35.0	947	183.70	hopkins	patricia	quincy
1	1003002312	hopkins	patricia	internal medicine	diazepam	diazepam	16	16.0	460	86.53	hopkins	patricia	quincy
2	1003002312	hopkins	patricia	internal medicine	lorazepam	lorazepam	72	87.0	2496	510.97	hopkins	patricia	quincy
3	1003030586	ksheersagar	pankaj	family practice	alprazolam	alprazolam	15	15.0	450	59.09	ksheersagar	pankaj	worcester
4	1003042441	manzo	mark	physician assistant	alprazolam	alprazolam	17	17.0	500	77.42	manzo	mark	grafton

[151]:
```
print(sum(benzo_town_merge_13['nppes_provider_last_org_name'] !=␣
 ↪benzo_town_merge_13['nppes_provider_last_name']))
print(sum(benzo_town_merge_13['nppes_provider_first_name_benz'] !=␣
 ↪benzo_town_merge_13['nppes_provider_first_name_opi']))
```

```
benzo_town_merge_13.drop(['nppes_provider_last_name',␣
 ↪'nppes_provider_first_name_opi'], axis=1, inplace=True)
```

```
0
0
```

[152]:
```
print(benzo_town_merge_17.shape)
print(benzo_town_merge_16.shape)
print(benzo_town_merge_15.shape)
print(benzo_town_merge_14.shape)
print(benzo_town_merge_13.shape)
```

```
(13138, 11)
(13132, 11)
(12820, 11)
(12621, 11)
(12162, 11)
```

[153]:
```
benzo_town_merge_17['year'] = 2017
benzo_town_merge_16['year'] = 2016
benzo_town_merge_15['year'] = 2015
benzo_town_merge_14['year'] = 2014
benzo_town_merge_13['year'] = 2013
```

[154]:
```
benzo_pres_town_all = pd.concat([benzo_town_merge_17, benzo_town_merge_16,␣
 ↪benzo_town_merge_15, benzo_town_merge_14, benzo_town_merge_13])
print(benzo_pres_town_all.shape)
display(benzo_pres_town_all.head())
sum(benzo_pres_town_all[['npi', 'year']].drop_duplicates()['npi'].
 ↪value_counts() > 5)
```

```
(63873, 12)
```

	npi	nppes_provider_last_org_name	nppes_provider_first_name_benz	\
0	1003002312	hopkins	patricia	
1	1003002312	hopkins	patricia	
2	1003002312	hopkins	patricia	
3	1003007477	abdow	kimberly	
4	1003007477	abdow	kimberly	

	specialty_description	drug_name	generic_name	total_claim_count	\
0	rheumatology	alprazolam	alprazolam	21	
1	rheumatology	diazepam	diazepam	14	
2	rheumatology	lorazepam	lorazepam	129	
3	nurse practitioner	alprazolam	alprazolam	63	
4	nurse practitioner	diazepam	diazepam	97	

	total_30_day_fill_count	total_day_supply	total_drug_cost	town	year
0	25.0	735	213.58	quincy	2017
1	14.0	360	56.56	quincy	2017
2	139.0	4028	907.84	quincy	2017
3	75.0	2160	324.14	worcester	2017
4	97.6	2920	781.21	worcester	2017

[154]:
```
0
```

[155]:
```
#benzo_pres_town_all.to_csv("../../data/tidy_data/
↪med_partD_benzo_indiv_pres_w_town_merge_13_to_17.csv", index=False)
```

[156]:
```
benzo_town_year_sum = benzo_pres_town_all.groupby(['town', 'generic_name',␣
↪'year']).sum().reset_index().drop('npi', axis=1)
print(benzo_town_year_sum.shape)
benzo_town_year_sum.head()
```

```
(3653, 7)
```

[156]:

	town	generic_name	year	total_claim_count	total_30_day_fill_count	\
0	abington	alprazolam	2013	457	495.0	
1	abington	alprazolam	2014	413	442.8	
2	abington	alprazolam	2015	385	410.0	
3	abington	alprazolam	2016	284	316.5	
4	abington	alprazolam	2017	275	309.5	

	total_day_supply	total_drug_cost
0	13567	3185.74
1	12428	2427.40
2	11241	2234.07
3	8777	1534.62
4	8384	1627.81

[157]:
```
#benzo_town_year_sum.to_csv("../../data/tidy_data/
↪med_partD_benzo_sum_w_town_merge_13_to_17.csv", index=False)
```