test

December 9, 2024

1 STOR 320: Introduction to Data Science

1.1 EDA Group PLACE_GROUP_NUMBER_HERE (Ex: EDA Group 12)

1.2 Part 1: Data cleaning, merging, and visualization (6 points)

[2]:
```
# hide
from IPython.display import HTML

HTML('''
<script>
$(document).ready(function() {
    // Hide input code cells
    $('div.input').hide();

    // Hide output prompts ("Out []")
    $('.prompt').hide();
});
</script>
''')
```

[2]:
```
<IPython.core.display.HTML object>
```

[3]:
```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import roc_auc_score, precision_recall_curve,␣
 ↪confusion_matrix, classification_report, roc_curve
from sklearn.model_selection import train_test_split
```

[4]:
```
eviction = pd.read_csv('eviction_2020_2024.csv')
eviction.head()
```

[4]:
```
              city          type        GEOID racial_majority    month  \
0  Albuquerque, NM  Census Tract  35001000107           White  01/2020
```

```
1  Albuquerque, NM  Census Tract  35001000107           White  02/2020
2  Albuquerque, NM  Census Tract  35001000107           White  03/2020
3  Albuquerque, NM  Census Tract  35001000107           White  04/2020
4  Albuquerque, NM  Census Tract  35001000107           White  05/2020

   filings_2020  filings_avg last_updated  xcourtcofips
0             8     5.333333   2024-08-10           NaN
1            14     9.000000   2024-08-10           NaN
2            10     5.666667   2024-08-10           NaN
3             5     8.333333   2024-08-10           NaN
4             0     6.666667   2024-08-10           NaN
```

[5]:
```
PIT_2023 = pd.read_excel('PITCOC.xlsb', engine='pyxlsb', sheet_name="2023")
PIT_2023.head()
```

[5]:
```
  CoC Number                                        CoC Name  \
0     AK-500                                  Anchorage CoC
1     AK-501                       Alaska Balance of State CoC
2     AL-500  Birmingham/Jefferson, St. Clair, Shelby Counti…
3     AL-501            Mobile City & County/Baldwin County CoC
4     AL-502                     Florence/Northwest Alabama CoC

              CoC Category                      Count Types  Overall Homeless  \
0  Other Largely Urban CoC  Sheltered and Unsheltered Count            1760.0
1        Largely Rural CoC  Sheltered and Unsheltered Count             854.0
2    Largely Suburban CoC  Sheltered and Unsheltered Count             847.0
3  Other Largely Urban CoC  Sheltered and Unsheltered Count             670.0
4        Largely Rural CoC  Sheltered and Unsheltered Count             195.0

   Overall Homeless - Under 18  Overall Homeless - Age 18 to 24  \
0                        185.0                            161.0
1                        176.0                             66.0
2                         67.0                             42.0
3                        110.0                             19.0
4                         63.0                              9.0

   Overall Homeless - Age 25 to 34  Overall Homeless - Age 35 to 44  \
0                            377.0                            419.0
1                            124.0                            190.0
2                            127.0                            182.0
3                             78.0                            156.0
4                             42.0                             36.0

   Overall Homeless - Age 45 to 54  …  \
0                            315.0  …
1                            144.0  …
2                            180.0  …
```

```
3                            120.0  …
4                             23.0  …

   Overall Homeless Parenting Youth Age 18-24  \
0                                        10.0
1                                         8.0
2                                         2.0
3                                         2.0
4                                         4.0

   Sheltered ES Homeless Parenting Youth Age 18-24  \
0                                              5.0
1                                              8.0
2                                              1.0
3                                              2.0
4                                              2.0

   Sheltered TH Homeless Parenting Youth Age 18-24  \
0                                              5.0
1                                              0.0
2                                              1.0
3                                              0.0
4                                              0.0

   Sheltered Total Homeless Parenting Youth Age 18-24  \
0                                               10.0
1                                                8.0
2                                                2.0
3                                                2.0
4                                                2.0

   Unsheltered Homeless Parenting Youth Age 18-24  \
0                                             0.0
1                                             0.0
2                                             0.0
3                                             0.0
4                                             2.0

   Overall Homeless Children of Parenting Youth  \
0                                       10.0
1                                        7.0
2                                        2.0
3                                        2.0
4                                        4.0

   Sheltered ES Homeless Children of Parenting Youth  \
0                                               5.0
```

```
1                                                 7.0
2                                                 1.0
3                                                 2.0
4                                                 2.0

   Sheltered TH Homeless Children of Parenting Youth  \
0                                                 5.0
1                                                 0.0
2                                                 1.0
3                                                 0.0
4                                                 0.0

   Sheltered Total Homeless Children of Parenting Youth  \
0                                                10.0
1                                                 7.0
2                                                 2.0
3                                                 2.0
4                                                 2.0

   Unsheltered Homeless Children of Parenting Youth
0                                                0.0
1                                                0.0
2                                                0.0
3                                                0.0
4                                                2.0

[5 rows x 645 columns]
```

[6]:
```
eviction.head()
```

[6]:
```
              city          type        GEOID racial_majority    month  \
0  Albuquerque, NM  Census Tract  35001000107           White  01/2020
1  Albuquerque, NM  Census Tract  35001000107           White  02/2020
2  Albuquerque, NM  Census Tract  35001000107           White  03/2020
3  Albuquerque, NM  Census Tract  35001000107           White  04/2020
4  Albuquerque, NM  Census Tract  35001000107           White  05/2020

   filings_2020  filings_avg last_updated  xcourtcofips
0             8     5.333333   2024-08-10           NaN
1            14     9.000000   2024-08-10           NaN
2            10     5.666667   2024-08-10           NaN
3             5     8.333333   2024-08-10           NaN
4             0     6.666667   2024-08-10           NaN
```

[7]:
```
## column names check (1 point)
```

[8]:
```
eviction.columns
```

[8]:
```
Index(['city', 'type', 'GEOID', 'racial_majority', 'month', 'filings_2020',
       'filings_avg', 'last_updated', 'xcourtcofips'],
      dtype='object')
```

[9]:
```
# remove GEOID, last_updated, and xcourtcofips because they're not useful for merging or relevant
eviction = eviction.drop(["GEOID", "last_updated", "xcourtcofips"], axis=1)
eviction.columns
```

[9]:
```
Index(['city', 'type', 'racial_majority', 'month', 'filings_2020',
       'filings_avg'],
      dtype='object')
```

[10]:
```
len(PIT_2023.columns) # len because there's a lot
```

[10]:
```
645
```

[11]:
```
new_cols = PIT_2023.columns[0:25] # columns after 25 talk about sheltered and unsheltered
PIT_sliced = PIT_2023[new_cols][0:386] # truncating rows
PIT_sliced = PIT_sliced.drop("Count Types", axis=1)
PIT_sliced.head()
```

[11]:

	CoC Number	CoC Name	CoC Category	Overall Homeless	Overall Homeless - Under 18	Overall Homeless - Age 18 to 24	Overall Homeless - Age 25 to 34	Overall Homeless - Age 35 to 44	Overall Homeless - Age 45 to 54
0	AK-500	Anchorage CoC	Other Largely Urban CoC	1760.0	185.0	161.0	377.0	419.0	315.0
1	AK-501	Alaska Balance of State CoC	Largely Rural CoC	854.0	176.0	66.0	124.0	190.0	144.0
2	AL-500	Birmingham/Jefferson, St. Clair, Shelby Counti…	Largely Suburban CoC	847.0	67.0	42.0	127.0		
3	AL-501	Mobile City & County/Baldwin County CoC	Other Largely Urban CoC	670.0	110.0	19.0	78.0		
4	AL-502	Florence/Northwest Alabama CoC	Largely Rural CoC	195.0	63.0	9.0	42.0		

```
2                            182.0                              180.0
3                            156.0                              120.0
4                             36.0                               23.0

   Overall Homeless - Age 55 to 64  …  \
0                            223.0  …
1                            123.0  …
2                            187.0  …
3                            140.0  …
4                             16.0  …

   Overall Homeless - Gender that is not Singularly Female or Male  \
0                                                 3.0
1                                                 2.0
2                                                 1.0
3                                                 2.0
4                                                 0.0

   Overall Homeless - Gender Questioning  \
0                                    2.0
1                                    0.0
2                                    0.0
3                                    0.0
4                                    0.0

   Overall Homeless - Non-Hispanic/Non-Latin(o)(a)(x)  \
0                                           1638.0
1                                            717.0
2                                            822.0
3                                            655.0
4                                            186.0

   Overall Homeless - Hispanic/Latin(o)(a)(x)  Overall Homeless - White  \
0                                       122.0                     480.0
1                                       137.0                     316.0
2                                        25.0                     295.0
3                                        15.0                     281.0
4                                         9.0                     130.0

   Overall Homeless - Black, African American, or African  \
0                                           156.0
1                                            25.0
2                                           518.0
3                                           351.0
4                                            48.0

   Overall Homeless - Asian or Asian American  \
```

```
0                                        28.0
1                                        10.0
2                                         1.0
3                                         2.0
4                                         0.0

   Overall Homeless - American Indian, Alaska Native, or Indigenous  \
0                                              755.0
1                                              396.0
2                                               14.0
3                                               11.0
4                                                2.0

   Overall Homeless - Native Hawaiian or Other Pacific Islander  \
0                                               83.0
1                                               16.0
2                                                2.0
3                                                6.0
4                                                0.0

   Overall Homeless - Multiple Races
0                              258.0
1                               91.0
2                               17.0
3                               19.0
4                               15.0

[5 rows x 24 columns]
```

[12]:
```
## missing data check (1 point)
```

[13]:
```
eviction_cols = eviction.columns
for col in eviction_cols:
    slice = eviction[str(col)]
    if slice.hasnans:
        print(f"{str(col)} " + f"({len(slice.unique())})" + ": " + str(slice.
 ↪unique()))
```

```
racial_majority (5): ['White' 'Other' 'Latinx' nan 'Black']
filings_avg (2775): [ 5.33333333  9.          5.66666667 … 79.25       59.75
 69.25      ]
```

[14]:
```
# we can drop filings_avg since we are calculating the average for the year␣
 ↪later
# we can drop racial_majority since we can supplement it with census data if␣
 ↪need be
eviction = eviction.drop(["filings_avg", "racial_majority"], axis=1)
```

```
eviction
```

[14]:
```
                    city          type    month  filings_2020
0       Albuquerque, NM  Census Tract  01/2020             8
1       Albuquerque, NM  Census Tract  02/2020            14
2       Albuquerque, NM  Census Tract  03/2020            10
3       Albuquerque, NM  Census Tract  04/2020             5
4       Albuquerque, NM  Census Tract  05/2020             0
…                   …           …       …              …
598451   Wilmington, DE  Census Tract  04/2024             1
598452   Wilmington, DE  Census Tract  05/2024             9
598453   Wilmington, DE  Census Tract  06/2024             3
598454   Wilmington, DE  Census Tract  07/2024             5
598455   Wilmington, DE  Census Tract  08/2024             1

[598456 rows x 4 columns]
```

[15]:
```
# replacing NaN with zero because values are not categorical
PIT_sliced = PIT_sliced.fillna(0)
PIT_sliced.head()
```

[15]:
```
  CoC Number                                        CoC Name  \
0     AK-500                                  Anchorage CoC
1     AK-501                       Alaska Balance of State CoC
2     AL-500  Birmingham/Jefferson, St. Clair, Shelby Counti…
3     AL-501            Mobile City & County/Baldwin County CoC
4     AL-502                     Florence/Northwest Alabama CoC

              CoC Category  Overall Homeless  Overall Homeless - Under 18  \
0  Other Largely Urban CoC            1760.0                         185.0
1        Largely Rural CoC             854.0                         176.0
2     Largely Suburban CoC             847.0                          67.0
3  Other Largely Urban CoC             670.0                         110.0
4        Largely Rural CoC             195.0                          63.0

   Overall Homeless - Age 18 to 24  Overall Homeless - Age 25 to 34  \
0                            161.0                            377.0
1                             66.0                            124.0
2                             42.0                            127.0
3                             19.0                             78.0
4                              9.0                             42.0

   Overall Homeless - Age 35 to 44  Overall Homeless - Age 45 to 54  \
0                            419.0                            315.0
1                            190.0                            144.0
2                            182.0                            180.0
3                            156.0                            120.0
```

```
4                            36.0                                23.0

   Overall Homeless - Age 55 to 64  …  \
0                           223.0  …
1                           123.0  …
2                           187.0  …
3                           140.0  …
4                            16.0  …

   Overall Homeless - Gender that is not Singularly Female or Male  \
0                                                 3.0
1                                                 2.0
2                                                 1.0
3                                                 2.0
4                                                 0.0

   Overall Homeless - Gender Questioning  \
0                                    2.0
1                                    0.0
2                                    0.0
3                                    0.0
4                                    0.0

   Overall Homeless - Non-Hispanic/Non-Latin(o)(a)(x)  \
0                                             1638.0
1                                              717.0
2                                              822.0
3                                              655.0
4                                              186.0

   Overall Homeless - Hispanic/Latin(o)(a)(x)  Overall Homeless - White  \
0                                       122.0                     480.0
1                                       137.0                     316.0
2                                        25.0                     295.0
3                                        15.0                     281.0
4                                         9.0                     130.0

   Overall Homeless - Black, African American, or African  \
0                                              156.0
1                                               25.0
2                                              518.0
3                                              351.0
4                                               48.0

   Overall Homeless - Asian or Asian American  \
0                                        28.0
1                                        10.0
```

```
2                                        1.0
3                                        2.0
4                                        0.0

   Overall Homeless - American Indian, Alaska Native, or Indigenous  \
0                                              755.0
1                                              396.0
2                                               14.0
3                                               11.0
4                                                2.0

   Overall Homeless - Native Hawaiian or Other Pacific Islander  \
0                                               83.0
1                                               16.0
2                                                2.0
3                                                6.0
4                                                0.0

   Overall Homeless - Multiple Races
0                              258.0
1                               91.0
2                               17.0
3                               19.0
4                               15.0

[5 rows x 24 columns]
```

[16]:
```
## Outlier check (1 point)
```

[17]:
```
PIT_sliced.describe()
```

[17]:
```
       Overall Homeless  Overall Homeless - Under 18  \
count        386.000000                   386.000000
mean        3383.958549                   578.341969
std        33685.071349                  5823.125602
min           55.000000                     1.000000
25%          337.000000                    53.000000
50%          656.500000                   109.500000
75%         1549.250000                   226.750000
max       653104.000000                111620.000000

       Overall Homeless - Age 18 to 24  Overall Homeless - Age 25 to 34  \
count                       386.000000                       386.000000
mean                        245.782383                       575.139896
std                        2462.595873                      5768.223300
min                           2.000000                         5.000000
25%                          21.000000                        52.250000
```

```
50%                          44.000000                      107.000000
75%                         102.000000                      244.500000
max                       47436.000000                   111002.000000

       Overall Homeless - Age 35 to 44  Overall Homeless - Age 45 to 54  \
count                        386.000000                      386.000000
mean                         618.119171                      504.088083
std                         6160.189569                     5011.779805
min                            0.000000                        0.000000
25%                           60.000000                       48.000000
50%                          120.500000                       99.500000
75%                          285.500000                      233.000000
max                       119297.000000                    97289.000000

       Overall Homeless - Age 55 to 64  Overall Homeless - Over 64  \
count                        386.000000                  386.000000
mean                         469.958549                  191.735751
std                         4672.930049                 1905.210525
min                            0.000000                    0.000000
25%                           41.250000                   16.000000
50%                           89.500000                   36.000000
75%                          216.750000                   88.500000
max                        90702.000000                37005.000000

       Overall Homeless - Female  Overall Homeless - Male  …  \
count                 386.000000               386.000000  …
mean                 1295.383420              2047.461140  …
std                 12896.424039             20384.594635  …
min                    20.000000                23.000000  …
25%                   134.250000               190.500000  …
50%                   255.000000               389.000000  …
75%                   569.500000               890.500000  …
max                250009.000000            395160.000000  …

       Overall Homeless - Gender that is not Singularly Female or Male  \
count                                          386.000000
mean                                            16.005181
std                                            160.203799
min                                              0.000000
25%                                              0.000000
50%                                              2.000000
75%                                              6.000000
max                                           3089.000000

       Overall Homeless - Gender Questioning  \
count                             386.000000
mean                                3.932642
```

std	39.295757
min	0.000000
25%	0.000000
50%	0.000000
75%	1.000000
max	759.000000

	Overall Homeless - Non-Hispanic/Non-Latin(o)(a)(x)
count	386.000000
mean	2454.756477
std	24261.167747
min	17.000000
25%	292.000000
50%	565.500000
75%	1197.000000
max	473768.000000

	Overall Homeless - Hispanic/Latin(o)(a)(x)	Overall Homeless - White
count	386.000000	386.000000
mean	929.202073	1683.181347
std	9548.451413	16621.001323
min	0.000000	9.000000
25%	23.250000	156.750000
50%	68.000000	344.000000
75%	221.750000	821.000000
max	179336.000000	324854.000000

	Overall Homeless - Black, African American, or African
count	386.000000
mean	1262.300518
std	12835.350265
min	0.000000
25%	70.000000
50%	177.500000
75%	436.000000
max	243624.000000

	Overall Homeless - Asian or Asian American
count	386.000000
mean	59.968912
std	622.043861
min	0.000000
25%	1.000000
50%	4.000000
75%	11.000000
max	11574.000000

```
       Overall Homeless - American Indian, Alaska Native, or Indigenous  \
count                                         386.000000
mean                                          119.772021
std                                          1188.048176
min                                             0.000000
25%                                             2.000000
50%                                            10.000000
75%                                            35.750000
max                                         23116.000000

       Overall Homeless - Native Hawaiian or Other Pacific Islander  \
count                                         386.000000
mean                                           55.502591
std                                           561.061296
min                                             0.000000
25%                                             0.000000
50%                                             2.000000
75%                                             9.000000
max                                         10712.000000

       Overall Homeless - Multiple Races
count                         386.000000
mean                          203.233161
std                          2022.758577
min                             0.000000
25%                            13.000000
50%                            28.000000
75%                            75.750000
max                         39224.000000

[8 rows x 21 columns]
```

[18]:
```
# the max seems quite high for overall homeless
max = PIT_sliced["Overall Homeless"].idxmax()
PIT_sliced.loc[[max]]
```

[18]:
```
    CoC Number CoC Name CoC Category  Overall Homeless  \
385          0    Total            0          653104.0

     Overall Homeless - Under 18  Overall Homeless - Age 18 to 24  \
385                     111620.0                          47436.0

     Overall Homeless - Age 25 to 34  Overall Homeless - Age 35 to 44  \
385                         111002.0                          119297.0

     Overall Homeless - Age 45 to 54  Overall Homeless - Age 55 to 64  …  \
385                          97289.0                           90702.0  …
```

```
     Overall Homeless - Gender that is not Singularly Female or Male  \
385                                                3089.0

     Overall Homeless - Gender Questioning  \
385                                  759.0

     Overall Homeless - Non-Hispanic/Non-Latin(o)(a)(x)  \
385                                           473768.0

     Overall Homeless - Hispanic/Latin(o)(a)(x)  Overall Homeless - White  \
385                                    179336.0                  324854.0

     Overall Homeless - Black, African American, or African  \
385                                           243624.0

     Overall Homeless - Asian or Asian American  \
385                                    11574.0

     Overall Homeless - American Indian, Alaska Native, or Indigenous  \
385                                           23116.0

     Overall Homeless - Native Hawaiian or Other Pacific Islander  \
385                                           10712.0

     Overall Homeless - Multiple Races
385                            39224.0

[1 rows x 24 columns]
```

[19]:
```
# these are totals, so it's okay to remove
PIT_sliced = PIT_sliced.drop(max)
PIT_sliced.describe()
```

[19]:
```
       Overall Homeless  Overall Homeless - Under 18  \
count        385.000000                   385.000000
mean        1696.374026                   289.922078
std         5955.731326                  1343.006959
min           55.000000                     1.000000
25%          337.000000                    53.000000
50%          653.000000                   109.000000
75%         1532.000000                   226.000000
max        88025.000000                 25200.000000

       Overall Homeless - Age 18 to 24  Overall Homeless - Age 25 to 34  \
count                        385.00000                       385.000000
mean                         123.21039                       288.316883
```

std	515.60255	1233.336451
min	2.00000	5.000000
25%	21.00000	52.000000
50%	44.00000	107.000000
75%	102.00000	240.000000
max	9130.00000	19577.000000

	Overall Homeless - Age 35 to 44	Overall Homeless - Age 45 to 54	\
count	385.000000	385.000000	
mean	309.862338	252.698701	
std	1128.114723	852.013239	
min	0.000000	0.000000	
25%	60.000000	48.000000	
50%	120.000000	99.000000	
75%	284.000000	230.000000	
max	16597.000000	13475.000000	

	Overall Homeless - Age 55 to 64	Overall Homeless - Over 64	\
count	385.000000	385.000000	
mean	235.589610	96.116883	
std	797.125492	317.653276	
min	0.000000	0.000000	
25%	41.000000	16.000000	
50%	89.000000	36.000000	
75%	216.000000	87.000000	
max	12841.000000	4721.000000	

	Overall Homeless - Female	Overall Homeless - Male	…	\
count	385.000000	385.000000	…	
mean	649.374026	1026.389610	…	
std	2289.696403	3623.145956	…	
min	20.000000	23.000000	…	
25%	134.000000	190.000000	…	
50%	254.000000	389.000000	…	
75%	568.000000	880.000000	…	
max	37788.000000	49650.000000	…	

	Overall Homeless - Gender that is not Singularly Female or Male	\
count	385.000000	
mean	8.023377	
std	32.806995	
min	0.000000	
25%	0.000000	
50%	2.000000	
75%	6.000000	
max	570.000000	

	Overall Homeless - Gender Questioning
count	385.000000
mean	1.971429
std	7.721164
min	0.000000
25%	0.000000
50%	0.000000
75%	1.000000
max	110.000000

	Overall Homeless - Non-Hispanic/Non-Latin(o)(a)(x)
count	385.000000
mean	1230.566234
std	3186.678681
min	17.000000
25%	292.000000
50%	565.000000
75%	1194.000000
max	41029.000000

	Overall Homeless - Hispanic/Latin(o)(a)(x)	Overall Homeless - White
count	385.000000	385.000000
mean	465.807792	843.776623
std	2882.200067	2071.985499
min	0.000000	9.000000
25%	23.000000	156.000000
50%	68.000000	342.000000
75%	221.000000	809.000000
max	46996.000000	28624.000000

	Overall Homeless - Black, African American, or African
count	385.000000
mean	632.789610
std	3436.753939
min	0.000000
25%	70.000000
50%	174.000000
75%	436.000000
max	59292.000000

	Overall Homeless - Asian or Asian American
count	385.000000
mean	30.062338
std	204.468891
min	0.000000
25%	1.000000
50%	4.000000

```
75%                                    11.000000
max                                  3773.000000

       Overall Homeless - American Indian, Alaska Native, or Indigenous  \
count                                          385.000000
mean                                            60.041558
std                                            185.485240
min                                              0.000000
25%                                              2.000000
50%                                             10.000000
75%                                             35.000000
max                                           2700.000000

       Overall Homeless - Native Hawaiian or Other Pacific Islander  \
count                                          385.000000
mean                                            27.823377
std                                            138.238208
min                                              0.000000
25%                                              0.000000
50%                                              2.000000
75%                                              9.000000
max                                           1683.000000

       Overall Homeless - Multiple Races
count                         385.000000
mean                          101.880519
std                           356.021559
min                             0.000000
25%                            13.000000
50%                            28.000000
75%                            75.000000
max                          4888.000000

[8 rows x 21 columns]
```

[20]:
```
eviction.describe()
```

[20]:
```
        filings_2020
count  598456.000000
mean        4.744833
std        86.956711
min         0.000000
25%         0.000000
50%         1.000000
75%         3.000000
max     14556.000000
```

[21]:
```
max = eviction["filings_2020"].idxmax()
eviction.loc[[max]] # this makes sense, no need to remove
```

[21]:

	city	type	month	filings_2020
460830	New York, NY	Zip Code	01/2020	14556

[22]:
```
## data merging (1 point)
```

[23]:
```
# remove data from eviction that is not from 2023
eviction_2023 = eviction[eviction['month'].str.contains('2023')]
eviction_2023["month"].unique()
```

[23]:
```
array(['01/2023', '02/2023', '03/2023', '04/2023', '05/2023', '06/2023',
       '07/2023', '08/2023', '09/2023', '10/2023', '11/2023', '12/2023'],
      dtype=object)
```

[24]:
```
eviction_city = eviction_2023.groupby("city").mean(numeric_only=True)
eviction_city = eviction_city.reset_index()
eviction_city['state'] = eviction_city['city'].str[-2:]
eviction_city["state"][10] = "FL" # ft. lauderdale was mislabeled
eviction_city["state"][20] = "FL" # miami was mislabeled
eviction_city["state"][26] = "FL" # palm beach was mislabeled
eviction_city
```

```
/var/folders/nw/5zcrqdxs7c57b12ptv8284p80000gn/T/ipykernel_1500/372597069.py:4:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  eviction_city["state"][10] = "FL" # ft. lauderdale was mislabeled
/var/folders/nw/5zcrqdxs7c57b12ptv8284p80000gn/T/ipykernel_1500/372597069.py:5:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  eviction_city["state"][20] = "FL" # miami was mislabeled
/var/folders/nw/5zcrqdxs7c57b12ptv8284p80000gn/T/ipykernel_1500/372597069.py:6:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  eviction_city["state"][26] = "FL" # palm beach was mislabeled
```

[24]:

	city	filings_2020	state
0	Albuquerque, NM	4.560734	NM

1	Austin, TX	8.560688	TX
2	Boston, MA	2.673019	MA
3	Bridgeport, CT	1.642178	CT
4	Charleston, SC	8.527500	SC
5	Cincinnati, OH	4.695301	OH
6	Cleveland, OH	3.372671	OH
7	Columbus, OH	6.018744	OH
8	Dallas, TX	4.892802	TX
9	Fort Worth, TX	5.956454	TX
10	Ft. Lauderdale	3.701754	FL
11	Gainesville, FL	3.346045	FL
12	Greenville, SC	10.136425	SC
13	Hartford, CT	2.340749	CT
14	Houston, TX	5.739814	TX
15	Indianapolis, IN	8.988517	IN
16	Jacksonville, FL	5.681818	FL
17	Kansas City, MO	3.308114	MO
18	Las Vegas, NV	8.877954	NV
19	Memphis, TN	10.232667	TN
20	Miami	2.313795	FL
21	Milwaukee, WI	3.814631	WI
22	Minneapolis-Saint Paul, MN	2.767266	MN
23	Nashville, TN	19.700893	TN
24	New Orleans, LA	2.615489	LA
25	New York, NY	36.728682	NY
26	Palm Beach	2.003788	FL
27	Philadelphia, PA	2.681133	PA
28	Phoenix, AZ	35.612973	AZ
29	Pittsburgh, PA	7.492009	PA
30	Providence, RI	3.785959	RI
31	Richmond, VA	42.004167	VA
32	South Bend, IN	2.735944	IN
33	St Louis, MO	3.816227	MO
34	Tampa, FL	2.764184	FL
35	Wilmington, DE	4.833904	DE

[25]:
```
PIT_sliced_state = PIT_sliced[PIT_sliced['CoC Name'].str.contains('State')]
PIT_sliced_state['state'] = PIT_sliced_state['CoC Number'].str[0:2]
PIT_sliced_state.head()
```

```
/var/folders/nw/5zcrqdxs7c57b12ptv8284p80000gn/T/ipykernel_1500/2207987685.py:2:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  PIT_sliced_state['state'] = PIT_sliced_state['CoC Number'].str[0:2]
```

[25]:

	CoC Number	CoC Name	CoC Category	\
1	AK-501	Alaska Balance of State CoC	Largely Rural CoC	
9	AL-507	Alabama Balance of State CoC	Largely Rural CoC	
12	AR-503	Arkansas Balance of State CoC	Largely Rural CoC	
15	AZ-500	Arizona Balance of State CoC	Largely Rural CoC	
62	CO-500	Colorado Balance of State CoC	Largely Rural CoC	

	Overall Homeless	Overall Homeless - Under 18	\
1	854.0	176.0	
9	283.0	77.0	
12	871.0	91.0	
15	2386.0	295.0	
62	2201.0	203.0	

	Overall Homeless - Age 18 to 24	Overall Homeless - Age 25 to 34	\
1	66.0	124.0	
9	25.0	53.0	
12	66.0	161.0	
15	117.0	317.0	
62	120.0	394.0	

	Overall Homeless - Age 35 to 44	Overall Homeless - Age 45 to 54	\
1	190.0	144.0	
9	60.0	38.0	
12	196.0	190.0	
15	447.0	473.0	
62	495.0	460.0	

	Overall Homeless - Age 55 to 64	…	\
1	123.0	…	
9	13.0	…	
12	133.0	…	
15	457.0	…	
62	380.0	…	

	Overall Homeless - Gender Questioning	\
1	0.0	
9	4.0	
12	3.0	
15	5.0	
62	2.0	

	Overall Homeless - Non-Hispanic/Non-Latin(o)(a)(x)	\
1	717.0	
9	258.0	
12	832.0	
15	1798.0	

```
62                                          1577.0

    Overall Homeless - Hispanic/Latin(o)(a)(x)  Overall Homeless - White  \
1                                        137.0                     316.0
9                                         25.0                      91.0
12                                        39.0                     619.0
15                                       588.0                    1662.0
62                                       624.0                    1660.0

    Overall Homeless - Black, African American, or African  \
1                                                   25.0
9                                                  170.0
12                                                 233.0
15                                                 142.0
62                                                  90.0

    Overall Homeless - Asian or Asian American  \
1                                         10.0
9                                          0.0
12                                         2.0
15                                        14.0
62                                         8.0

    Overall Homeless - American Indian, Alaska Native, or Indigenous  \
1                                                  396.0
9                                                    1.0
12                                                   3.0
15                                                 266.0
62                                                 217.0

    Overall Homeless - Native Hawaiian or Other Pacific Islander  \
1                                                   16.0
9                                                    0.0
12                                                   5.0
15                                                  16.0
62                                                  10.0

    Overall Homeless - Multiple Races  state
1                                91.0     AK
9                                21.0     AL
12                                9.0     AR
15                              286.0     AZ
62                              216.0     CO

[5 rows x 25 columns]
```

[26]:
```
merge = pd.merge(PIT_sliced_state, eviction_city, on = "state")
merge.head()
```

[26]:
```
  CoC Number                          CoC Name          CoC Category  \
0     AZ-500      Arizona Balance of State CoC     Largely Rural CoC
1     CT-505  Connecticut Balance of State CoC  Largely Suburban CoC
2     CT-505  Connecticut Balance of State CoC  Largely Suburban CoC
3     DE-500            Delaware Statewide CoC  Largely Suburban CoC
4     IN-502      Indiana Balance of State CoC     Largely Rural CoC

   Overall Homeless  Overall Homeless - Under 18  \
0            2386.0                         295.0
1            2418.0                         454.0
2            2418.0                         454.0
3            1245.0                         335.0
4            4398.0                         923.0

   Overall Homeless - Age 18 to 24  Overall Homeless - Age 25 to 34  \
0                            117.0                            317.0
1                            190.0                            378.0
2                            190.0                            378.0
3                             66.0                            195.0
4                            238.0                            709.0

   Overall Homeless - Age 35 to 44  Overall Homeless - Age 45 to 54  \
0                            447.0                            473.0
1                            430.0                            393.0
2                            430.0                            393.0
3                            189.0                            193.0
4                            848.0                            789.0

   Overall Homeless - Age 55 to 64  …  \
0                            457.0  …
1                            431.0  …
2                            431.0  …
3                            211.0  …
4                            704.0  …

   Overall Homeless - Hispanic/Latin(o)(a)(x)  Overall Homeless - White  \
0                                       588.0                    1662.0
1                                       730.0                    1381.0
2                                       730.0                    1381.0
3                                       107.0                     385.0
4                                       221.0                    3018.0

   Overall Homeless - Black, African American, or African  \
0                                              142.0
```

```
1                                                832.0
2                                                832.0
3                                                773.0
4                                               1113.0

   Overall Homeless - Asian or Asian American  \
0                                        14.0
1                                        12.0
2                                        12.0
3                                         4.0
4                                        19.0

   Overall Homeless - American Indian, Alaska Native, or Indigenous  \
0                                                266.0
1                                                 33.0
2                                                 33.0
3                                                  3.0
4                                                 20.0

   Overall Homeless - Native Hawaiian or Other Pacific Islander  \
0                                                 16.0
1                                                  6.0
2                                                  6.0
3                                                  2.0
4                                                 30.0

   Overall Homeless - Multiple Races  state             city  filings_2020
0                              286.0     AZ       Phoenix, AZ     35.612973
1                              154.0     CT    Bridgeport, CT      1.642178
2                              154.0     CT      Hartford, CT      2.340749
3                               78.0     DE    Wilmington, DE      4.833904
4                              198.0     IN  Indianapolis, IN      8.988517

[5 rows x 27 columns]
```

[27]:
```
PIT_sliced
```

[27]:
```
    CoC Number                                          CoC Name  \
0       AK-500                                    Anchorage CoC
1       AK-501                     Alaska Balance of State CoC
2       AL-500  Birmingham/Jefferson, St. Clair, Shelby Counti…
3       AL-501            Mobile City & County/Baldwin County CoC
4       AL-502                  Florence/Northwest Alabama CoC
..         …                                               …
380     WV-500                      Wheeling, Weirton Area CoC
381     WV-501            Huntington/Cabell, Wayne Counties CoC
382     WV-503  Charleston/Kanawha, Putnam, Boone, Clay Counti…
```

```
383     WV-508                West Virginia Balance of State CoC
384     WY-500                             Wyoming Statewide CoC

                CoC Category  Overall Homeless  Overall Homeless - Under 18  \
0    Other Largely Urban CoC            1760.0                         185.0
1          Largely Rural CoC             854.0                         176.0
2       Largely Suburban CoC             847.0                          67.0
3    Other Largely Urban CoC             670.0                         110.0
4          Largely Rural CoC             195.0                          63.0
..                       …                 …                             …
380        Largely Rural CoC             113.0                          37.0
381        Largely Rural CoC             244.0                          18.0
382     Largely Suburban CoC             293.0                          20.0
383        Largely Rural CoC             766.0                          50.0
384        Largely Rural CoC             532.0                          37.0

     Overall Homeless - Age 18 to 24  Overall Homeless - Age 25 to 34  \
0                              161.0                             377.0
1                               66.0                             124.0
2                               42.0                             127.0
3                               19.0                              78.0
4                                9.0                              42.0
..                               …                                …
380                              3.0                              23.0
381                             19.0                              51.0
382                             40.0                              53.0
383                             74.0                             168.0
384                             62.0                              73.0

     Overall Homeless - Age 35 to 44  Overall Homeless - Age 45 to 54  \
0                              419.0                             315.0
1                              190.0                             144.0
2                              182.0                             180.0
3                              156.0                             120.0
4                               36.0                              23.0
..                               …                                …
380                             24.0                              16.0
381                             72.0                              53.0
382                             66.0                              72.0
383                            209.0                             148.0
384                            116.0                             126.0

     Overall Homeless - Age 55 to 64  …  \
0                              223.0  …
1                              123.0  …
2                              187.0  …
3                              140.0  …
```

```
4                                16.0  …
..                                …  …
380                               9.0  …
381                              25.0  …
382                              28.0  …
383                              78.0  …
384                              75.0  …

     Overall Homeless - Gender that is not Singularly Female or Male  \
0                                                    3.0
1                                                    2.0
2                                                    1.0
3                                                    2.0
4                                                    0.0
..                                                   …
380                                                  0.0
381                                                  0.0
382                                                  0.0
383                                                  5.0
384                                                  1.0

     Overall Homeless - Gender Questioning  \
0                                      2.0
1                                      0.0
2                                      0.0
3                                      0.0
4                                      0.0
..                                     …
380                                    0.0
381                                    0.0
382                                    0.0
383                                    4.0
384                                    1.0

     Overall Homeless - Non-Hispanic/Non-Latin(o)(a)(x)  \
0                                                1638.0
1                                                 717.0
2                                                 822.0
3                                                 655.0
4                                                 186.0
..                                                  …
380                                               113.0
381                                               238.0
382                                               292.0
383                                               750.0
384                                               462.0
```

	Overall Homeless - Hispanic/Latin(o)(a)(x)	Overall Homeless - White	\
0	122.0	480.0	
1	137.0	316.0	
2	25.0	295.0	
3	15.0	281.0	
4	9.0	130.0	
..	…	…	
380	0.0	83.0	
381	6.0	206.0	
382	1.0	230.0	
383	16.0	691.0	
384	70.0	419.0	

	Overall Homeless - Black, African American, or African	\
0	156.0	
1	25.0	
2	518.0	
3	351.0	
4	48.0	
..	…	
380	17.0	
381	19.0	
382	42.0	
383	59.0	
384	47.0	

	Overall Homeless - Asian or Asian American	\
0	28.0	
1	10.0	
2	1.0	
3	2.0	
4	0.0	
..	…	
380	0.0	
381	2.0	
382	0.0	
383	0.0	
384	14.0	

	Overall Homeless - American Indian, Alaska Native, or Indigenous	\
0	755.0	
1	396.0	
2	14.0	
3	11.0	
4	2.0	
..	…	
380	4.0	

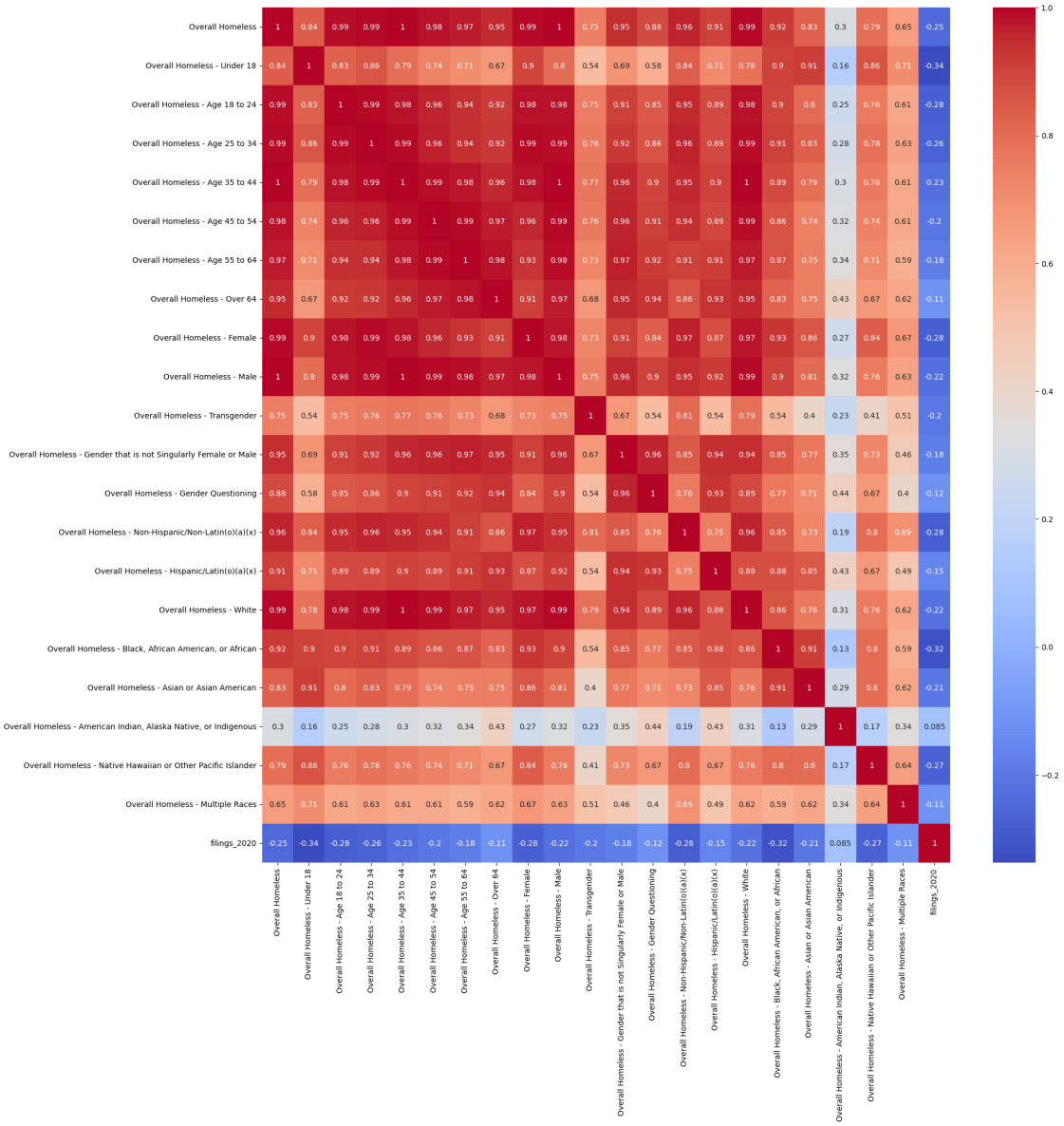
```
381                                                 3.0
382                                                 0.0
383                                                 4.0
384                                                22.0

     Overall Homeless - Native Hawaiian or Other Pacific Islander  \
0                                                   83.0
1                                                   16.0
2                                                    2.0
3                                                    6.0
4                                                    0.0
..                                                   …
380                                                  0.0
381                                                  1.0
382                                                  0.0
383                                                  1.0
384                                                  2.0

     Overall Homeless - Multiple Races
0                                258.0
1                                 91.0
2                                 17.0
3                                 19.0
4                                 15.0
..                                 …
380                                9.0
381                               13.0
382                               21.0
383                               11.0
384                               28.0

[385 rows x 24 columns]
```

[28]:
```
## data transformation, normalization, and cleaning (1 point)
```

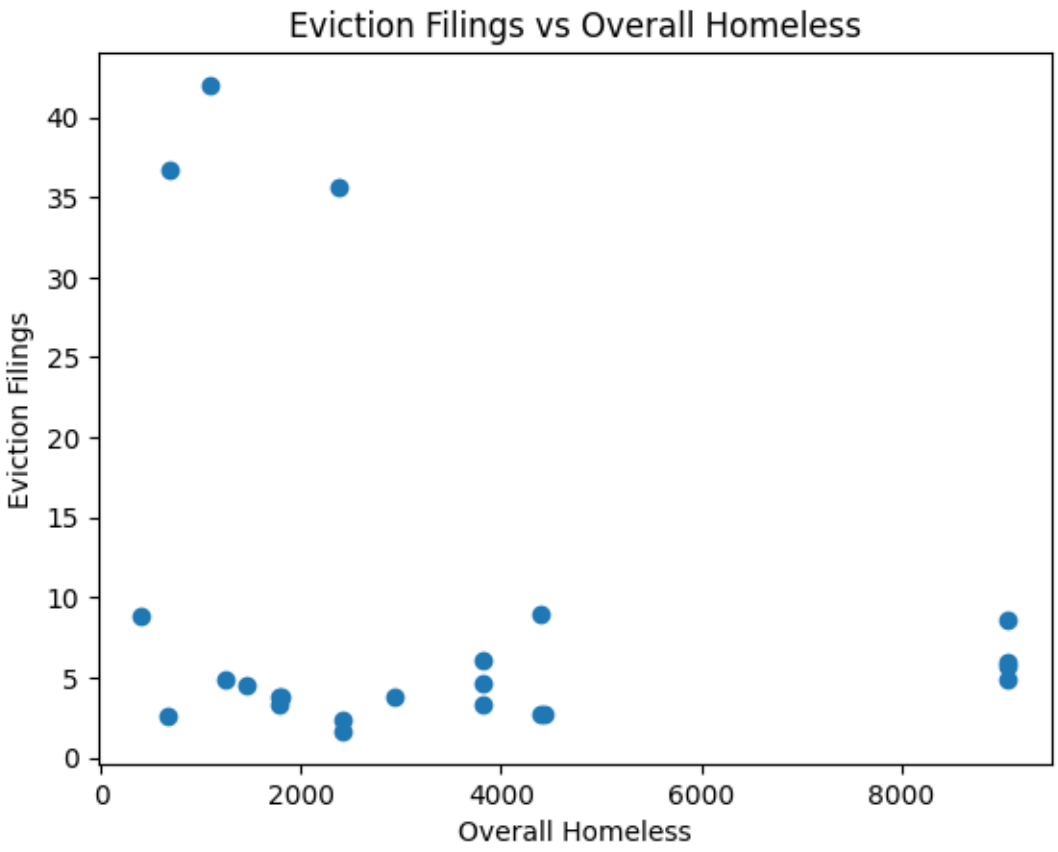
[29]:
```
# sets up data for corrplots and purely quantitative analysis
merge_numeric = merge.select_dtypes(include=[np.number])
```

[30]:
```
## Exploratory data visualization (1 point)
```

[31]:
```
plt.figure(figsize=(20,20))
sns.heatmap(data=merge_numeric.corr(), annot=True, cmap='coolwarm')
plt.show()
```



[32]:
```
import matplotlib.pyplot as plt
plt.scatter(x=merge["Overall Homeless"], y=merge["filings_2020"])
plt.ylabel("Eviction Filings")
plt.xlabel("Overall Homeless")
plt.title("Eviction Filings vs Overall Homeless")
plt.show()
```



High multicollinearity & low n, unfit for regression or advanced models.

1.3 Part 2: Answer questions from the proposals (8 points)

Each plot should be followed by a paragraph of explanation and observation.

1.3.1 Creator: Ivy Nangalia

Question: What demographics of people are more likely to be homeless today?

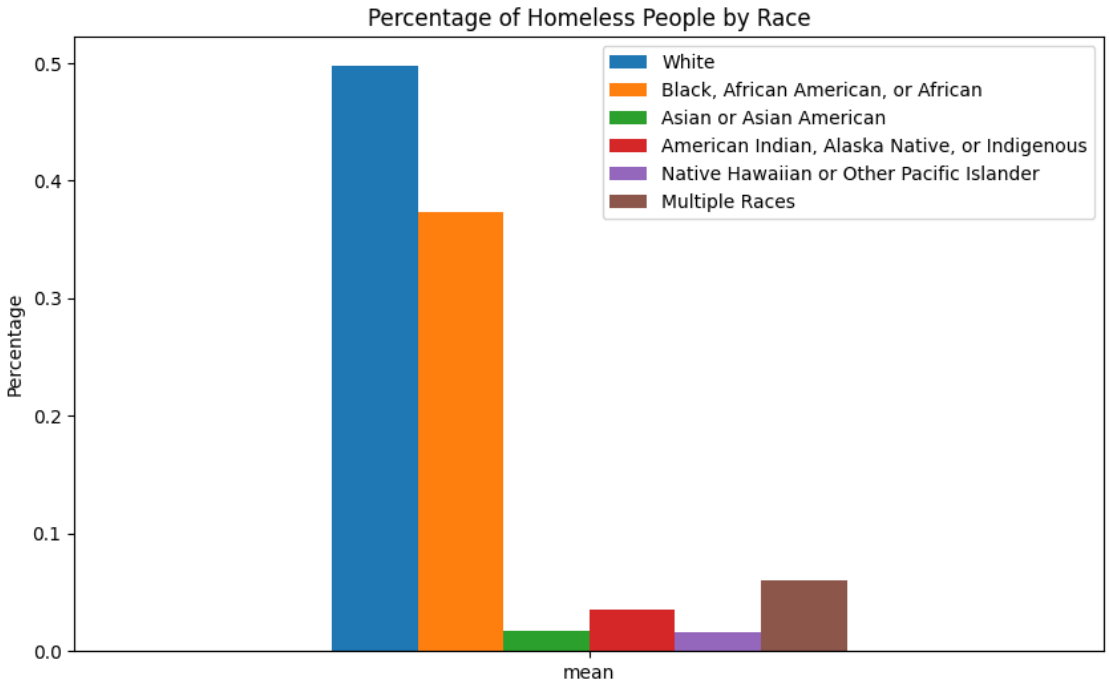
[33]:
```
means = PIT_sliced.describe()[1:2]
race_cols = means.columns[15:22]
total_sum = means[race_cols].sum(axis=1)
race_percent = means[race_cols].div(total_sum, axis=0)
race_percent.reset_index()
```

[33]:
```
  index  Overall Homeless - White  \
0  mean                    0.4974

   Overall Homeless - Black, African American, or African  \
0                                           0.373025
```

```
   Overall Homeless - Asian or Asian American  \
0                                    0.017722

   Overall Homeless - American Indian, Alaska Native, or Indigenous  \
0                                           0.035394

   Overall Homeless - Native Hawaiian or Other Pacific Islander  \
0                                           0.016402

   Overall Homeless - Multiple Races
0                           0.060058
```

[34]:
```
# renaming columns
race_percent = race_percent.rename(columns={'Overall Homeless - White': 'White',
                                            'Overall Homeless - Black, African␣
 ↪American, or African': 'Black, African American, or African',
                                            'Overall Homeless - Asian or Asian␣
 ↪American': 'Asian or Asian American',
                                            'Overall Homeless - American␣
 ↪Indian, Alaska Native, or Indigenous': 'American Indian, Alaska Native, or␣
 ↪Indigenous',
                                            'Overall Homeless - Native Hawaiian␣
 ↪or Other Pacific Islander': 'Native Hawaiian or Other Pacific Islander',
                                            'Overall Homeless - Multiple Races':
 ↪ 'Multiple Races'})
```

[35]:
```
race_percent.plot(kind='bar', figsize=(10, 6))
plt.ylabel('Percentage')
plt.title('Percentage of Homeless People by Race')
plt.xticks(rotation=0)
plt.show()
```



Answer: It seems that White people have the highest percentage of the homeless population (49.7%) followed closely by Black people (37.3%). White people making up the largest share of the homeless population makes sense since the majority of the American population is White. However, only 13.7% of the population is Black, which implies some disproportionate factors affecting Black people and their housing security. This makes sense considering the long history of racism and racist policies enacted in the United States. Moreover, I'd argue that White people are less likely to be homeless considering that they make up 75.3% of the population but only 49.7% of the homeless population, implying some systemic factors that improve the housing security of White people as compared to others.

Note: the population data is from the US Census.

1.3.2 Interpreter 2: Ximing Sun

Question: What trends do we see in racial segregation?

Answer:

[36]:
```
racial = merge.iloc[:, 16:26]
racial
```

[36]:
```
    Overall Homeless - Non-Hispanic/Non-Latin(o)(a)(x)  \
0                                              1798.0
1                                              1688.0
2                                              1688.0
3                                              1138.0
```

4	4177.0
5	4177.0
6	662.0
7	2906.0
8	1725.0
9	1725.0
10	620.0
11	349.0
12	604.0
13	3655.0
14	3655.0
15	3655.0
16	1404.0
17	5676.0
18	5676.0
19	5676.0
20	5676.0
21	1029.0
22	2674.0

	Overall Homeless - Hispanic/Latin(o)(a)(x)	Overall Homeless - White	\
0	588.0	1662.0	
1	730.0	1381.0	
2	730.0	1381.0	
3	107.0	385.0	
4	221.0	3018.0	
5	221.0	3018.0	
6	16.0	273.0	
7	1526.0	2153.0	
8	67.0	1347.0	
9	67.0	1347.0	
10	828.0	900.0	
11	61.0	361.0	
12	83.0	553.0	
13	168.0	3027.0	
14	168.0	3027.0	
15	168.0	3027.0	
16	406.0	1140.0	
17	3389.0	6533.0	
18	3389.0	6533.0	
19	3389.0	6533.0	
20	3389.0	6533.0	
21	59.0	653.0	
22	266.0	1933.0	

	Overall Homeless - Black, African American, or African	\
0	142.0	

1	832.0
2	832.0
3	773.0
4	1113.0
5	1113.0
6	385.0
7	1917.0
8	328.0
9	328.0
10	136.0
11	16.0
12	87.0
13	580.0
14	580.0
15	580.0
16	436.0
17	2095.0
18	2095.0
19	2095.0
20	2095.0
21	366.0
22	603.0

	Overall Homeless - Asian or Asian American	\
0	14.0	
1	12.0	
2	12.0	
3	4.0	
4	19.0	
5	19.0	
6	1.0	
7	78.0	
8	5.0	
9	5.0	
10	4.0	
11	3.0	
12	1.0	
13	7.0	
14	7.0	
15	7.0	
16	12.0	
17	59.0	
18	59.0	
19	59.0	
20	59.0	
21	6.0	
22	35.0	

	Overall Homeless - American Indian, Alaska Native, or Indigenous
0	266.0
1	33.0
2	33.0
3	3.0
4	20.0
5	20.0
6	4.0
7	19.0
8	14.0
9	14.0
10	339.0
11	17.0
12	3.0
13	34.0
14	34.0
15	34.0
16	33.0
17	142.0
18	142.0
19	142.0
20	142.0
21	13.0
22	186.0

	Overall Homeless - Native Hawaiian or Other Pacific Islander
0	16.0
1	6.0
2	6.0
3	2.0
4	30.0
5	30.0
6	1.0
7	40.0
8	23.0
9	23.0
10	6.0
11	0.0
12	1.0
13	11.0
14	11.0
15	11.0
16	4.0
17	33.0
18	33.0
19	33.0

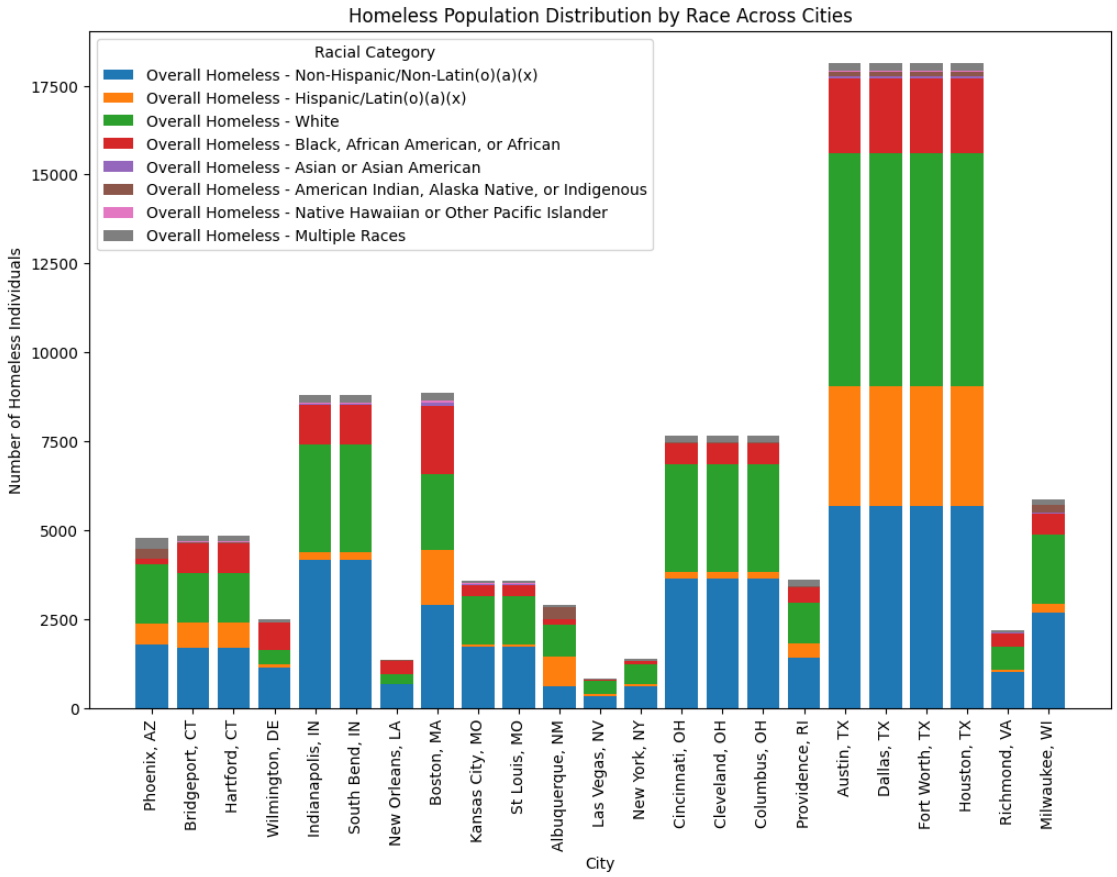
```
20                                                33.0
21                                                 0.0
22                                                10.0
```

	Overall Homeless - Multiple Races	state	city
0	286.0	AZ	Phoenix, AZ
1	154.0	CT	Bridgeport, CT
2	154.0	CT	Hartford, CT
3	78.0	DE	Wilmington, DE
4	198.0	IN	Indianapolis, IN
5	198.0	IN	South Bend, IN
6	14.0	LA	New Orleans, LA
7	225.0	MA	Boston, MA
8	75.0	MO	Kansas City, MO
9	75.0	MO	St Louis, MO
10	63.0	NM	Albuquerque, NM
11	13.0	NV	Las Vegas, NV
12	42.0	NY	New York, NY
13	164.0	OH	Cincinnati, OH
14	164.0	OH	Cleveland, OH
15	164.0	OH	Columbus, OH
16	185.0	RI	Providence, RI
17	203.0	TX	Austin, TX
18	203.0	TX	Dallas, TX
19	203.0	TX	Fort Worth, TX
20	203.0	TX	Houston, TX
21	50.0	VA	Richmond, VA
22	173.0	WI	Milwaukee, WI

[37]:
```
## Add more cells if your group has more than two interpreters
```

[38]:
```
racial_categories = [
    'Overall Homeless - Non-Hispanic/Non-Latin(o)(a)(x)',
    'Overall Homeless - Hispanic/Latin(o)(a)(x)',
    'Overall Homeless - White',
    'Overall Homeless - Black, African American, or African',
    'Overall Homeless - Asian or Asian American',
    'Overall Homeless - American Indian, Alaska Native, or Indigenous',
    'Overall Homeless - Native Hawaiian or Other Pacific Islander',
    'Overall Homeless - Multiple Races'
]

fig, ax = plt.subplots(figsize=(12, 8))
for category in racial_categories:
    ax.bar(racial['city'], racial[category], label=category,␣
 ↪bottom=racial[racial_categories].cumsum(axis=1)[category] - racial[category])
```

```
ax.set_xlabel("City")
ax.set_ylabel("Number of Homeless Individuals")
ax.set_title("Homeless Population Distribution by Race Across Cities")
ax.legend(title="Racial Category")
plt.xticks(rotation=90)

plt.show()
```



[39]:
```
PIT_sliced["state"] = PIT_sliced["CoC Number"].apply(lambda x: x[:2])

racial_categories = [
    'Overall Homeless - Non-Hispanic/Non-Latin(o)(a)(x)',
    'Overall Homeless - Hispanic/Latin(o)(a)(x)',
    'Overall Homeless - White',
    'Overall Homeless - Black, African American, or African',
    'Overall Homeless - Asian or Asian American',
    'Overall Homeless - American Indian, Alaska Native, or Indigenous',
    'Overall Homeless - Native Hawaiian or Other Pacific Islander',
    'Overall Homeless - Multiple Races',
```

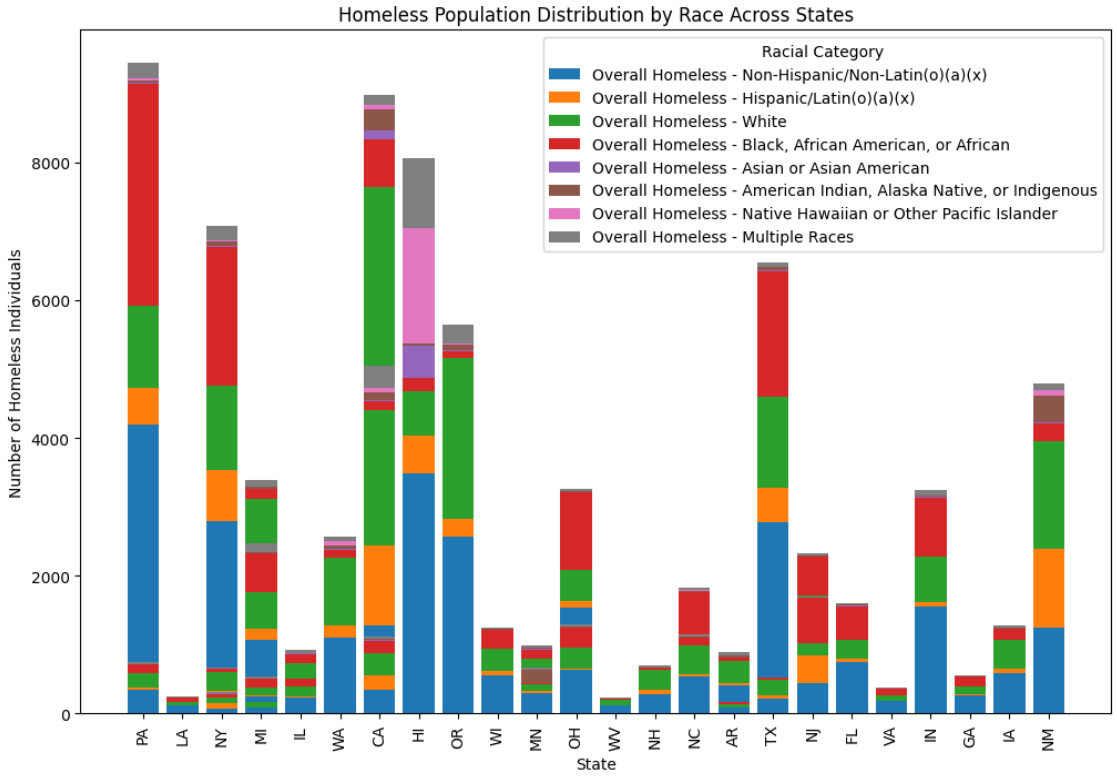
```
    "state"
]


race_country = PIT_sliced.iloc[:, 16:26]
race_country_state = race_country.groupby("state").mean()

#state validation:
states = ["AK", "AL", "AR", "AZ", "CA", "CO", "CT", "DC",
          "DE", "FL", "GA", "HI", "IA", "ID", "IL", "IN", "KS",
          "KY", "LA", "MA", "MD", "ME", "MI", "MN", "MO", "MS", "MT", "NC",
          "ND", "NE", "NH", "NJ", "NM", "NV", "NY", "OH", "OK", "OR", "PA",␣
 ↪"RI",
          "SC", "SD", "TN", "TX", "UT", "VA", "VT", "WA", "WI", "WV", "WY"]
for state in race_country_state.index:
    if state not in states:
        race_country_state.drop(state, inplace=True)

len(race_country_state)
race_country_state["total"] = race_country_state.sum(axis=1)

race_country_state_pct = race_country_state.drop("total", axis=1).
 ↪div(race_country_state["total"], axis=0) * 100

hisp_pct = pd.DataFrame(race_country_state_pct.iloc[:, 1])

import plotly.express as px

state_names = {
    "AK": "Alaska", "AL": "Alabama", "AR": "Arkansas", "AZ": "Arizona",
    "CA": "California", "CO": "Colorado", "CT": "Connecticut", "DC": "District␣
 ↪of Columbia",
    "DE": "Delaware", "FL": "Florida", "GA": "Georgia", "HI": "Hawaii",
    "IA": "Iowa", "ID": "Idaho", "IL": "Illinois", "IN": "Indiana",
    "KS": "Kansas", "KY": "Kentucky", "LA": "Louisiana", "MA": "Massachusetts",
    "MD": "Maryland", "ME": "Maine", "MI": "Michigan", "MN": "Minnesota",
    "MO": "Missouri", "MS": "Mississippi", "MT": "Montana", "NC": "North␣
 ↪Carolina",
    "ND": "North Dakota", "NE": "Nebraska", "NH": "New Hampshire", "NJ": "New␣
 ↪Jersey",
    "NM": "New Mexico", "NV": "Nevada", "NY": "New York", "OH": "Ohio",
    "OK": "Oklahoma", "OR": "Oregon", "PA": "Pennsylvania", "RI": "Rhode␣
 ↪Island",
    "SC": "South Carolina", "SD": "South Dakota", "TN": "Tennessee", "TX":␣
 ↪"Texas",
    "UT": "Utah", "VA": "Virginia", "VT": "Vermont", "WA": "Washington",
```

```
    "WI": "Wisconsin", "WV": "West Virginia", "WY": "Wyoming"
}

hisp_data = hisp_pct.reset_index()
hisp_data["state_name"] = hisp_data["state"].map(state_names)

fig = px.choropleth(
    hisp_data,
    locations="state",
    locationmode="USA-states",
    color="Overall Homeless - Hispanic/Latin(o)(a)(x)",
    color_continuous_scale="Viridis",
    range_color=[0, 30],
    title="Percentage of Hispanic/Latin(o)(a)(x) Homeless Population by State",
    labels={"Overall Homeless - Hispanic/Latin(o)(a)(x)": "%"}
    )

fig.update_layout(
    title_x=0.5,
    geo_scope="usa",
    width=1200,
    height=800
    )

fig.show()
```

[142]:

[40]:
```
# get a random sample of 40 CoC numbers and their corresponding homeless racial␣
 ↪demographics, and plot them

np.random.seed(89)
racial = PIT_sliced.copy().sample(40)

racial_categories = [
    'Overall Homeless - Non-Hispanic/Non-Latin(o)(a)(x)',
    'Overall Homeless - Hispanic/Latin(o)(a)(x)',
    'Overall Homeless - White',
    'Overall Homeless - Black, African American, or African',
    'Overall Homeless - Asian or Asian American',
    'Overall Homeless - American Indian, Alaska Native, or Indigenous',
    'Overall Homeless - Native Hawaiian or Other Pacific Islander',
    'Overall Homeless - Multiple Races'
]

fig, ax = plt.subplots(figsize=(12, 8))
for category in racial_categories:
```

```
    ax.bar(racial["state"], racial[category], label=category,␣
 ↪bottom=racial[racial_categories].cumsum(axis=1)[category] - racial[category])

ax.set_xlabel("State")
ax.set_ylabel("Number of Homeless Individuals")
ax.set_title("Homeless Population Distribution by Race Across States")
ax.legend(title="Racial Category")
plt.xticks(rotation=90)

plt.show()
```



[41]:
```
racial.head(15)
```

[41]:

	CoC Number	CoC Name
303	PA-500	Philadelphia CoC
147	LA-507	Alexandria/Central Louisiana CoC
269	NY-523	Glens Falls, Saratoga Springs/Saratoga, Washin…
189	MI-518	Livingston County CoC
129	IL-517	Aurora, Elgin/Kane County CoC
374	WA-504	Everett/Snohomish County CoC
58	CA-611	Oxnard, San Buenaventura/Ventura County CoC
108	HI-501	Honolulu City and County CoC

295	OR-500	Eugene, Springfield/Lane County CoC
124	IL-512	Bloomington/Central Illinois CoC
379	WI-503	Madison/Dane County CoC
194	MN-502	Rochester/Southeast Minnesota CoC
280	OH-502	Cleveland/Cuyahoga County CoC
380	WV-500	Wheeling, Weirton Area CoC
232	NH-502	Nashua/Hillsborough County CoC

	CoC Category	Overall Homeless	Overall Homeless - Under 18
303	Major City CoC	4725.0	825.0
147	Largely Rural CoC	122.0	15.0
269	Largely Suburban CoC	332.0	30.0
189	Largely Suburban CoC	88.0	36.0
129	Largely Suburban CoC	461.0	81.0
374	Largely Suburban CoC	1285.0	202.0
58	Other Largely Urban CoC	2441.0	141.0
108	Largely Suburban CoC	4028.0	583.0
295	Other Largely Urban CoC	2824.0	230.0
124	Largely Rural CoC	256.0	67.0
379	Other Largely Urban CoC	624.0	148.0
194	Largely Rural CoC	496.0	144.0
280	Largely Suburban CoC	1629.0	281.0
380	Largely Rural CoC	113.0	37.0
232	Largely Suburban CoC	348.0	99.0

	Overall Homeless - Age 18 to 24	Overall Homeless - Age 25 to 34
303	409.0	778.0
147	3.0	15.0
269	27.0	79.0
189	12.0	7.0
129	32.0	70.0
374	70.0	235.0
58	100.0	390.0
108	328.0	492.0
295	199.0	442.0
124	32.0	38.0
379	33.0	105.0
194	39.0	76.0
280	95.0	278.0
380	3.0	23.0
232	15.0	51.0

	Overall Homeless - Age 35 to 44	Overall Homeless - Age 45 to 54
303	968.0	753.0
147	39.0	24.0
269	81.0	48.0
189	20.0	8.0

```
129                              79.0                                81.0
374                             297.0                               202.0
58                              513.0                               509.0
108                             681.0                               727.0
295                             564.0                               589.0
124                              32.0                                30.0
379                             120.0                                82.0
194                              96.0                                63.0
280                             309.0                               248.0
380                              24.0                                16.0
232                              69.0                                40.0

     Overall Homeless - Age 55 to 64  …  \
303                             768.0  …
147                              22.0  …
269                              44.0  …
189                               4.0  …
129                              77.0  …
374                             211.0  …
58                              498.0  …
108                             895.0  …
295                             552.0  …
124                              24.0  …
379                              83.0  …
194                              51.0  …
280                             310.0  …
380                               9.0  …
232                              43.0  …

     Overall Homeless - Gender Questioning  \
303                                    8.0
147                                    0.0
269                                    0.0
189                                    1.0
129                                    0.0
374                                    2.0
58                                     0.0
108                                    2.0
295                                   17.0
124                                    0.0
379                                    1.0
194                                    0.0
280                                    0.0
380                                    0.0
232                                    0.0

     Overall Homeless - Non-Hispanic/Non-Latin(o)(a)(x)  \
```

303		4199.0
147		119.0
269		307.0
189		88.0
129		269.0
374		1106.0
58		1275.0
108		3479.0
295		2563.0
124		228.0
379		561.0
194		446.0
280		1532.0
380		113.0
232		280.0

	Overall Homeless - Hispanic/Latin(o)(a)(x)	Overall Homeless - White \
303	526.0	1197.0
147	3.0	45.0
269	25.0	267.0
189	0.0	88.0
129	192.0	278.0
374	179.0	981.0
58	1166.0	1969.0
108	549.0	656.0
295	261.0	2338.0
124	28.0	133.0
379	63.0	314.0
194	50.0	293.0
280	97.0	453.0
380	0.0	83.0
232	68.0	285.0

	Overall Homeless - Black, African American, or African \
303	3214.0
147	72.0
269	51.0
189	0.0
129	129.0
374	112.0
58	188.0
108	181.0
295	96.0
124	116.0
379	275.0
194	141.0
280	1133.0

380		17.0
232		34.0

	Overall Homeless - Asian or Asian American	\
303	26.0	
147	1.0	
269	3.0	
189	0.0	
129	5.0	
374	18.0	
58	9.0	
108	472.0	
295	9.0	
124	0.0	
379	8.0	
194	4.0	
280	5.0	
380	0.0	
232	2.0	

	Overall Homeless - American Indian, Alaska Native, or Indigenous	\
303	32.0	
147	0.0	
269	2.0	
189	0.0	
129	10.0	
374	38.0	
58	57.0	
108	31.0	
295	85.0	
124	0.0	
379	9.0	
194	21.0	
280	2.0	
380	4.0	
232	1.0	

	Overall Homeless - Native Hawaiian or Other Pacific Islander	\
303	20.0	
147	0.0	
269	0.0	
189	0.0	
129	0.0	
374	65.0	
58	57.0	
108	1683.0	
295	14.0	

```
124                                                 1.0
379                                                 1.0
194                                                 1.0
280                                                 4.0
380                                                 0.0
232                                                 4.0

     Overall Homeless - Multiple Races  state
303                              236.0     PA
147                                4.0     LA
269                                9.0     NY
189                                0.0     MI
129                               39.0     IL
374                               71.0     WA
58                               161.0     CA
108                             1005.0     HI
295                              282.0     OR
124                                6.0     IL
379                               17.0     WI
194                               36.0     MN
280                               32.0     OH
380                                9.0     WV
232                               22.0     NH

[15 rows x 25 columns]
```

[42]:
```
totals = race_country_state["total"]
```

[43]:
```
len(race_country_state)
race_country_state["total"] = race_country_state.sum(axis=1)

race_country_state_pct = race_country_state.drop("total", axis=1).
 ↪div(race_country_state["total"], axis=0) * 100

hisp_pct = pd.DataFrame(race_country_state_pct.iloc[:, 1])
hisp_pct
#race_country_state_pct
```

[43]:
```
       Overall Homeless - Hispanic/Latin(o)(a)(x)
state
AK                                       2.477047
AL                                       0.877724
AR                                       1.293599
AZ                                       6.894009
CA                                       9.227035
CO                                       6.523998
CT                                       7.728027
```

DC	2.453271
DE	2.148594
FL	4.238197
GA	1.386855
HI	3.338422
IA	2.318130
ID	5.265448
IL	7.499791
IN	1.200765
KS	3.060594
KY	1.138271
LA	0.978227
MA	8.752155
MD	1.867008
ME	1.285815
MI	1.803379
MN	3.252711
MO	1.745447
MS	0.712831
MT	2.330119
NC	1.358417
ND	2.136480
NE	3.330626
NH	2.201966
NJ	7.526306
NM	12.844872
NV	4.208978
NY	12.029312
OH	1.174688
OK	2.570998
OR	3.204746
PA	3.456515
RI	5.607735
SC	1.116457
SD	1.716069
TN	0.784048
TX	7.787559
UT	5.838080
VA	2.558793
VT	1.168437
WA	3.781745
WI	2.288624
WV	0.406073
WY	3.289474

[44]:
```
import plotly.express as px
```

```
state_names = {
    "AK": "Alaska", "AL": "Alabama", "AR": "Arkansas", "AZ": "Arizona",
    "CA": "California", "CO": "Colorado", "CT": "Connecticut", "DC": "District␣
 ↪of Columbia",
    "DE": "Delaware", "FL": "Florida", "GA": "Georgia", "HI": "Hawaii",
    "IA": "Iowa", "ID": "Idaho", "IL": "Illinois", "IN": "Indiana",
    "KS": "Kansas", "KY": "Kentucky", "LA": "Louisiana", "MA": "Massachusetts",
    "MD": "Maryland", "ME": "Maine", "MI": "Michigan", "MN": "Minnesota",
    "MO": "Missouri", "MS": "Mississippi", "MT": "Montana", "NC": "North␣
 ↪Carolina",
    "ND": "North Dakota", "NE": "Nebraska", "NH": "New Hampshire", "NJ": "New␣
 ↪Jersey",
    "NM": "New Mexico", "NV": "Nevada", "NY": "New York", "OH": "Ohio",
    "OK": "Oklahoma", "OR": "Oregon", "PA": "Pennsylvania", "RI": "Rhode␣
 ↪Island",
    "SC": "South Carolina", "SD": "South Dakota", "TN": "Tennessee", "TX":␣
 ↪"Texas",
    "UT": "Utah", "VA": "Virginia", "VT": "Vermont", "WA": "Washington",
    "WI": "Wisconsin", "WV": "West Virginia", "WY": "Wyoming"
}

hisp_data = hisp_pct.reset_index()
hisp_data["state_name"] = hisp_data["state"].map(state_names)

fig = px.choropleth(
    hisp_data,
    locations="state",
    locationmode="USA-states",
    color="Overall Homeless - Hispanic/Latin(o)(a)(x)",
    color_continuous_scale="Viridis",
    range_color=[0, 30],
    title="Percentage of Hispanic/Latin(o)(a)(x) Homeless Population by State",
    labels={"Overall Homeless - Hispanic/Latin(o)(a)(x)": "Percentage"}
    )

fig.update_layout(
    title_x=0.5,
    geo_scope="usa",
    width=1200,
    height=800
    )

fig.show()
```

[45]:
```
totals
```

[45]:
```
state
AK    2614.000000
AL     826.000000
AR    1043.600000
AZ    9491.333333
CA    8245.409091
CO    7219.500000
CT    3015.000000
DC    9844.000000
DE    2490.000000
FL    2278.222222
GA    2732.000000
HI    6223.000000
IA    1768.666667
ID    2298.000000
IL    1257.578947
IN    6017.000000
KS    1213.000000
KY    3177.333333
LA     905.428571
MA    3190.166667
MD    1173.000000
ME    8516.000000
MI     899.700000
MN    1678.600000
MO    1729.500000
MS     654.666667
MT    4356.000000
NC    1625.666667
ND    1568.000000
NE    1641.333333
NH    1627.333333
NJ    1283.000000
NM    3842.000000
NV    5777.333333
NY    8600.000000
OH    2530.222222
OK    1162.000000
OR    5035.500000
PA    1569.500000
RI    3620.000000
SC    2026.500000
SD    2564.000000
TN    1843.000000
TX    4977.636364
UT    2458.000000
VA     845.125000
```

```
VT    3295.000000
WA    9345.333333
WI    2430.500000
WV     708.000000
WY    1064.000000
Name: total, dtype: float64
```

[46]:
```
totals["state"] = totals.index
```

[47]:
```
race_country_state
```

[47]:

state	Overall Homeless - Non-Hispanic/Non-Latin(o)(a)(x)
AK	1177.500000
AL	398.500000
AR	494.800000
AZ	3437.000000
CA	2601.090909
CO	2667.750000
CT	1041.500000
DC	4439.000000
DE	1138.000000
FL	946.000000
GA	1290.222222
HI	2696.000000
IA	802.333333
ID	907.000000
IL	440.157895
IN	2864.000000
KS	532.250000
KY	1516.333333
LA	435.000000
MA	1036.666667
MD	542.700000
ME	4039.000000
MI	417.400000
MN	730.100000
MO	804.375000
MS	318.000000
MT	1975.000000
NC	768.666667
ND	717.000000
NE	711.333333
NH	742.000000
NJ	448.375000
NM	934.000000
NV	2402.333333

NY	2230.958333
OH	1205.666667
OK	521.250000
OR	2195.000000
PA	676.250000
RI	1404.000000
SC	968.000000
SD	1194.000000
TN	892.600000
TX	1713.545455
UT	942.000000
VA	379.312500
VT	1570.500000
WA	3965.833333
WI	1104.000000
WV	348.250000
WY	462.000000

state	Overall Homeless - Hispanic/Latin(o)(a)(x)	Overall Homeless - White
AK	129.500000	398.000000
AL	14.500000	168.625000
AR	27.000000	339.200000
AZ	1308.666667	3003.666667
CA	1521.613636	2190.568182
CO	942.000000	2413.750000
CT	466.000000	826.500000
DC	483.000000	587.000000
DE	107.000000	385.000000
FL	193.111111	614.037037
GA	75.777778	511.444444
HI	415.500000	647.500000
IA	82.000000	562.333333
ID	242.000000	951.000000
IL	188.631579	291.315789
IN	144.500000	1842.000000
KS	74.250000	419.500000
KY	72.333333	1128.666667
LA	17.714286	168.428571
MA	558.416667	768.583333
MD	43.800000	194.500000
ME	219.000000	2044.000000
MI	32.450000	213.900000
MN	109.200000	299.900000
MO	60.375000	491.000000
MS	9.333333	143.000000
MT	203.000000	1480.000000

NC	44.166667	341.166667
ND	67.000000	352.000000
NE	109.333333	528.333333
NH	71.666667	725.000000
NJ	193.125000	244.750000
NM	987.000000	1228.000000
NV	486.333333	1661.333333
NY	2069.041667	1231.333333
OH	59.444444	657.444444
OK	59.750000	330.250000
OR	322.750000	1945.375000
PA	108.500000	368.562500
RI	406.000000	1140.000000
SC	45.250000	495.000000
SD	88.000000	372.000000
TN	28.900000	571.800000
TX	775.272727	1479.727273
UT	287.000000	935.000000
VA	43.250000	167.562500
VT	77.000000	1417.500000
WA	706.833333	2841.000000
WI	111.250000	684.750000
WV	5.750000	302.500000
WY	70.000000	419.000000

state	Overall Homeless - Black, African American, or African \
AK	90.500000
AL	223.375000
AR	146.600000
AZ	1032.333333
CA	1212.931818
CO	607.750000
CT	558.500000
DC	4091.000000
DE	773.000000
FL	462.777778
GA	796.111111
HI	123.000000
IA	221.333333
ID	25.000000
IL	293.842105
IN	981.000000
KS	116.500000
KY	377.000000
LA	269.285714
MA	721.583333

MD	352.100000
ME	2013.000000
MI	199.050000
MN	315.200000
MO	309.125000
MS	168.333333
MT	63.000000
NC	420.416667
ND	102.000000
NE	176.000000
NH	41.333333
NJ	363.375000
NM	200.000000
NV	930.000000
NY	2741.291667
OH	529.444444
OK	123.875000
OR	194.875000
PA	361.250000
RI	436.000000
SC	465.750000
SD	63.000000
TN	310.100000
TX	878.909091
UT	109.666667
VA	214.625000
VT	133.000000
WA	782.333333
WI	398.000000
WV	34.250000
WY	47.000000

state	Overall Homeless - Asian or Asian American \
AK	19.000000
AL	1.500000
AR	1.400000
AZ	33.666667
CA	159.363636
CO	33.750000
CT	7.000000
DC	40.000000
DE	4.000000
FL	6.888889
GA	4.777778
HI	301.500000
IA	9.333333

ID	4.000000
IL	10.368421
IN	13.500000
KS	3.000000
KY	5.333333
LA	2.142857
MA	15.083333
MD	6.400000
ME	17.000000
MI	2.550000
MN	19.300000
MO	5.000000
MS	2.333333
MT	10.000000
NC	5.000000
ND	9.000000
NE	12.333333
NH	4.666667
NJ	5.125000
NM	12.000000
NV	57.333333
NY	36.375000
OH	4.444444
OK	3.000000
OR	19.125000
PA	4.062500
RI	12.000000
SC	3.000000
SD	7.000000
TN	3.000000
TX	22.363636
UT	13.000000
VA	8.312500
VT	18.000000
WA	57.166667
WI	12.750000
WV	0.500000
WY	14.000000

state	Overall Homeless - American Indian, Alaska Native, or Indigenous
AK	575.500000
AL	4.625000
AR	10.000000
AZ	325.000000
CA	195.204545
CO	207.250000

	Overall Homeless - Native Hawaiian or Other Pacific Islander \
CT	19.000000
DC	73.000000
DE	3.000000
FL	12.222222
GA	6.444444
HI	25.000000
IA	28.666667
ID	80.000000
IL	7.526316
IN	19.500000
KS	18.500000
KY	12.000000
LA	4.000000
MA	10.500000
MD	9.600000
ME	34.000000
MI	5.050000
MN	96.100000
MO	11.750000
MS	2.666667
MT	461.000000
NC	12.916667
ND	276.000000
NE	44.666667
NH	9.666667
NJ	9.312500
NM	359.500000
NV	79.000000
NY	39.416667
OH	10.111111
OK	77.625000
OR	124.500000
PA	5.812500
RI	33.000000
SC	9.500000
SD	779.000000
TN	8.500000
TX	37.818182
UT	72.666667
VA	3.750000
VT	41.500000
WA	344.166667
WI	52.750000
WV	2.750000
WY	22.000000

state	
AK	49.500000
AL	1.375000
AR	6.000000
AZ	35.333333
CA	77.568182
CO	100.500000
CT	3.500000
DC	52.000000
DE	2.000000
FL	4.333333
GA	2.222222
HI	1168.000000
IA	6.000000
ID	10.500000
IL	3.263158
IN	18.000000
KS	3.250000
KY	3.666667
LA	1.857143
MA	11.166667
MD	2.900000
ME	5.000000
MI	1.000000
MN	4.500000
MO	8.500000
MS	1.666667
MT	18.000000
NC	3.000000
ND	0.000000
NE	3.666667
NH	3.333333
NJ	2.562500
NM	40.500000
NV	52.333333
NY	12.250000
OH	4.222222
OK	5.625000
OR	46.875000
PA	2.812500
RI	4.000000
SC	2.000000
SD	9.000000
TN	1.000000
TX	8.727273
UT	31.666667
VA	1.125000

VT	3.000000
WA	203.000000
WI	3.000000
WV	0.500000
WY	2.000000

state	Overall Homeless - Multiple Races	total
AK	174.500000	5228.000000
AL	13.500000	1652.000000
AR	18.600000	2087.200000
AZ	315.666667	18982.666667
CA	287.068182	16490.818182
CO	246.750000	14439.000000
CT	93.000000	6030.000000
DC	79.000000	19688.000000
DE	78.000000	4980.000000
FL	38.851852	4556.444444
GA	45.000000	5464.000000
HI	846.500000	12446.000000
IA	56.666667	3537.333333
ID	78.500000	4596.000000
IL	22.473684	2515.157895
IN	134.500000	12034.000000
KS	45.750000	2426.000000
KY	62.000000	6354.666667
LA	7.000000	1810.857143
MA	68.166667	6380.333333
MD	21.000000	2346.000000
ME	145.000000	17032.000000
MI	28.300000	1799.400000
MN	104.300000	3357.200000
MO	39.375000	3459.000000
MS	9.333333	1309.333333
MT	146.000000	8712.000000
NC	30.333333	3251.333333
ND	45.000000	3136.000000
NE	55.666667	3282.666667
NH	29.666667	3254.666667
NJ	16.375000	2566.000000
NM	81.000000	7684.000000
NV	108.666667	11554.666667
NY	239.333333	17200.000000
OH	59.444444	5060.444444
OK	40.625000	2324.000000
OR	187.000000	10071.000000
PA	42.250000	3139.000000

```
RI                                185.000000   7240.000000
SC                                 38.000000   4053.000000
SD                                 52.000000   5128.000000
TN                                 27.100000   3686.000000
TX                                 61.272727   9955.272727
UT                                 67.000000   4916.000000
VA                                 27.187500   1690.250000
VT                                 34.500000   6590.000000
WA                                445.000000  18690.666667
WI                                 64.000000   4861.000000
WV                                 13.500000   1416.000000
WY                                 28.000000   2128.000000
```

[48]:
```
race_country_state['Overall Homeless - Total'] = race_country_state[
    [
        'Overall Homeless - Non-Hispanic/Non-Latin(o)(a)(x)',
    'Overall Homeless - Hispanic/Latin(o)(a)(x)',
    'Overall Homeless - White',
    'Overall Homeless - Black, African American, or African',
    'Overall Homeless - Asian or Asian American',
    'Overall Homeless - American Indian, Alaska Native, or Indigenous',
    'Overall Homeless - Native Hawaiian or Other Pacific Islander',
    'Overall Homeless - Multiple Races'
    ]
].sum(axis=1)

state_homeless_totals = race_country_state.groupby('state').
 ↪sum(numeric_only=True)['Overall Homeless - Total']
state_homeless_totals = state_homeless_totals.reset_index()

fig = px.choropleth(
    state_homeless_totals,
    locations='state',
    locationmode="USA-states",
    color='Overall Homeless - Total',
    color_continuous_scale="YlOrRd",

    scope="usa",
    labels={'Overall Homeless - Total': 'Total Homeless Population'},
    title="Total Homeless Population by State in the USA"
)

fig.update_layout(
    title_x=0.5,
    geo_scope="usa",
    width=1200,
    height=800
```

```
    )

fig.show()
```

1.3.3 Deliverer: FIRSTNAME LASTNAME

Question:

[ ]:

Answer:

[ ]:

[ ]:

1.4 Part 3: Follow-up Questions (4 points)

1.4.1 New Questions Based Off Initial Investigation

- Q1: WRITE_QUESTION_HERE
- Q2: WRITE_QUESTION_HERE

[ ]:

1.5 Summary (2 points)

GIVE A 2 PARAGRAPH SUMMARY.

PARAGRAPH 1 SHOULD DESCRIBE WHAT YOU LEARNED ABOUT YOUR DATA FROM INVESTIGATING THE INITIAL QUESTIONS. DID YOU FIND ANYTHING UNUSUAL IN YOUR DATA? DID ANYTHING SURPRISE YOU? WHICH OF THE INITIAL QUESTIONS WERE HELPFUL IN LEADING YOU TO MORE QUESTIONS?

PARAGRAPH 2 SHOULD SUMMARIZE WHAT YOU LEARNED FROM INVESTIGATING THE FOLLOW-UP QUESTIONS. WHY ARE THESE FOLLOW-UP QUESTIONS INTERESTING FOR INVESTIGATION? DESCRIBE THE TABLES/FIGURES YOU USED TO EXPLORE ANSWERS TO THESE FOLLOW-UP QUESTIONS? WHAT DID YOU LEARN FROM THE TABLES/FIGURES REGARDING THE FOLLOW-UP QUESTIONS YOU PROPOSED?

[ ]:

[49]:
```
# get data from new york
new_york_slice = PIT_sliced_state.loc[269:270]
new_york_slice
```

[49]:
```
     CoC Number                                      CoC Name  \
270     NY-525  New York Balance of State Continuum of Care

           CoC Category  Overall Homeless  Overall Homeless - Under 18  \
270  Largely Rural CoC             687.0                         138.0
```

```
     Overall Homeless - Age 18 to 24  Overall Homeless - Age 25 to 34  \
270                             69.0                             151.0

     Overall Homeless - Age 35 to 44  Overall Homeless - Age 45 to 54  \
270                            164.0                              94.0

     Overall Homeless - Age 55 to 64  …  \
270                             53.0  …

     Overall Homeless - Gender Questioning  \
270                                    0.0

     Overall Homeless - Non-Hispanic/Non-Latin(o)(a)(x)  \
270                                              604.0

     Overall Homeless - Hispanic/Latin(o)(a)(x)  Overall Homeless - White  \
270                                        83.0                      553.0

     Overall Homeless - Black, African American, or African  \
270                                              87.0

     Overall Homeless - Asian or Asian American  \
270                                         1.0

     Overall Homeless - American Indian, Alaska Native, or Indigenous  \
270                                              3.0

     Overall Homeless - Native Hawaiian or Other Pacific Islander  \
270                                              1.0

     Overall Homeless - Multiple Races  state
270                              42.0     NY

[1 rows x 25 columns]
```

[50]:
```
nyc_slice = PIT_sliced[PIT_sliced["CoC Number"]=="NY-600"]
total_homeless = nyc_slice["Overall Homeless"].values[0]

# Calculate demographic percentages
race_cols = ["Overall Homeless - White",
             "Overall Homeless - Black, African American, or African",
             "Overall Homeless - Asian or Asian American",
             "Overall Homeless - American Indian, Alaska Native, or Indigenous",
             "Overall Homeless - Native Hawaiian or Other Pacific Islander",
             "Overall Homeless - Multiple Races"]

race_probs = []
```

```
for col in race_cols:
    race_probs.append(nyc_slice[col].values[0] / total_homeless)

# Create simulated population
np.random.seed(42)
n_simulated = 88025

# Simulate race based on observed proportions
simulated_race = np.random.choice(len(race_cols), size=n_simulated,␣
 ↪p=race_probs)

# Create dataframe with simulated data
simulated_df = pd.DataFrame({
    "race": [race_cols[i].replace("Overall Homeless - ", "") for i in␣
 ↪simulated_race]
})

# Add age based on NYC proportions
age_cols = ["Overall Homeless - Under 18",
           "Overall Homeless - Age 18 to 24",
           "Overall Homeless - Age 25 to 34",
           "Overall Homeless - Age 35 to 44",
           "Overall Homeless - Age 45 to 54",
           "Overall Homeless - Age 55 to 64",
           "Overall Homeless - Over 64"]

age_probs = []
for col in age_cols:
    age_probs.append(nyc_slice[col].values[0] / total_homeless)

simulated_df["age"] = np.random.choice(
    [col.replace("Overall Homeless - ", "").replace("Age ", "") for col in␣
 ↪age_cols],
    size=n_simulated,
    p=age_probs
)

# Add gender based on NYC proportions
gender_cols = ["Overall Homeless - Female",
              "Overall Homeless - Male",
              "Overall Homeless - Transgender",
              "Overall Homeless - Gender that is not Singularly Female or Male",
              "Overall Homeless - Gender Questioning"]

gender_probs = []
for col in gender_cols:
    gender_probs.append(nyc_slice[col].values[0] / total_homeless)
```

```
simulated_df["gender"] = np.random.choice(
    [col.replace("Overall Homeless - ", "") for col in gender_cols],
    size=n_simulated,
    p=gender_probs
)


simulated_df["homeless"] = 1
```

[51]:
```
race_probs
```

[51]:
```
[0.2504742970746947,
 0.6735813689292814,
 0.008952002272081795,
 0.008838398182334565,
 0.0026242544731610337,
 0.05552967906844646]
```

[52]:
```
simulated_df
```

[52]:

	race	age	gender	homeless
0	Black, African American, or African	35 to 44	Male	1
1	Multiple Races	Under 18	Female	1
2	Black, African American, or African	25 to 34	Male	1
3	Black, African American, or African	35 to 44	Female	1
4	White	Under 18	Male	1
…	…	…	…	…
88020	Black, African American, or African	25 to 34	Female	1
88021	White	Under 18	Male	1
88022	Asian or Asian American	18 to 24	Male	1
88023	Black, African American, or African	35 to 44	Male	1
88024	Black, African American, or African	Under 18	Female	1

[88025 rows x 4 columns]

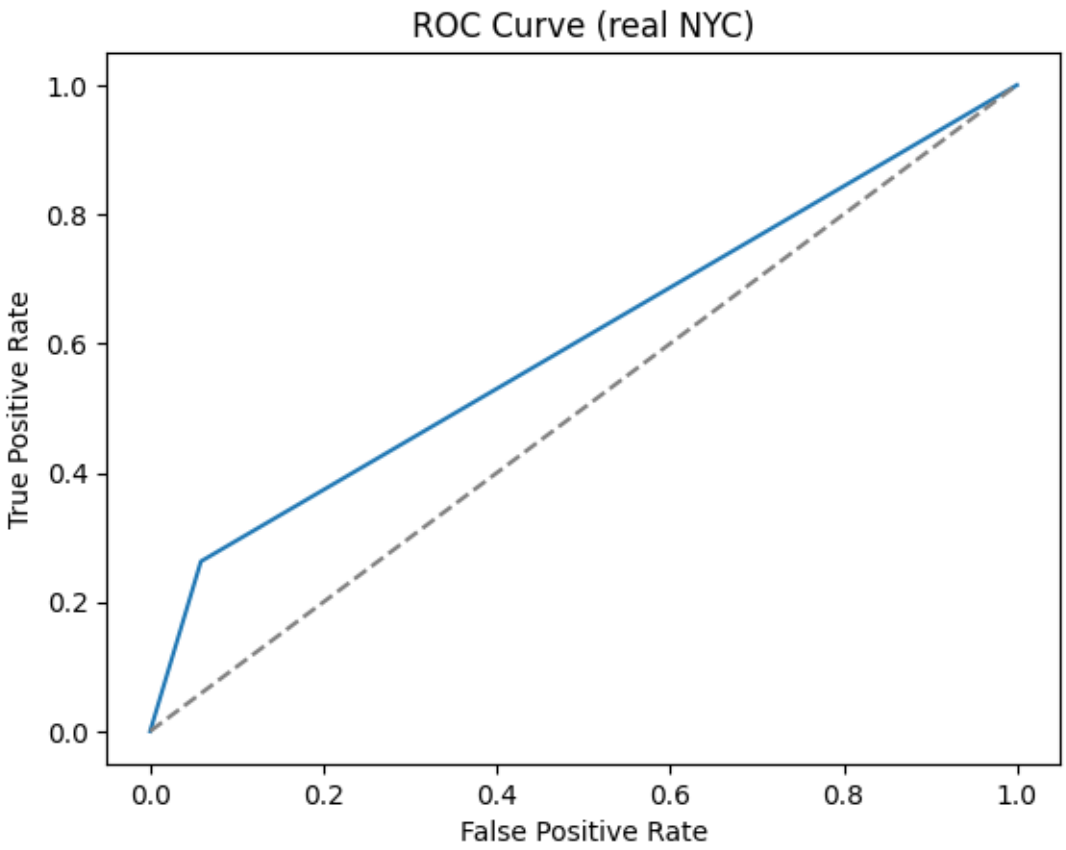
[53]:
```
n_non_homeless = 8_800_000 # Approximate NYC population


non_homeless_df = pd.DataFrame()

race_probs = {
    "White": 0.375,
    "Black, African American, or African": 0.231,
    "Asian or Asian American": 0.145,
    "Multiple Races": 0.089,
    "American Indian, Alaska Native, or Indigenous": 0.006,
```

```
    "Native Hawaiian or Other Pacific Islander": 0.001,
    "Hispanic or Latino": 0.29 - 0.137
}

non_homeless_df["race"] = np.random.choice(
    list(race_probs.keys()),
    size=n_non_homeless,
    p=list(race_probs.values())
)

# Age distribution based on NYC census
age_dist = {
    "Under 18": 0.21,
    "18 to 24": 0.10,
    "25 to 34": 0.17,
    "35 to 44": 0.14,
    "45 to 54": 0.13,
    "55 to 64": 0.13,
    "Over 64": 0.12
}

non_homeless_df["age"] = np.random.choice(
    list(age_dist.keys()),
    size=n_non_homeless,
    p=list(age_dist.values())
)

# Gender distribution based on NYC census
gender_dist = {
    "Female": 0.52 - (0.0012 +  0.0007 + 0.0001)/2,
    "Male": 0.48 - (0.0012 +  0.0007 + 0.0001)/2,
    "Other": 0.0012 +  0.0007 + 0.0001
}

non_homeless_df["gender"] = np.random.choice(
    list(gender_dist.keys()),
    size=n_non_homeless,
    p=list(gender_dist.values())
)

non_homeless_df["homeless"] = 0

# Combine homeless and non-homeless populations
simulated_df = pd.concat([simulated_df, non_homeless_df], ignore_index=True)
#simulated_df.drop("ethnicity", axis=1, inplace=True)
```

[ ]:
```
df = simulated_df.copy()


X = pd.get_dummies(df[["race", "age", "gender"]], drop_first=True)
y = df["homeless"]


X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,␣
 ↪random_state=42, stratify=y)


scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)


model = LogisticRegression(
    class_weight="balanced",
    max_iter=1000,
    random_state=42,
    C=0.1
)
model.fit(X_train, y_train)

y_pred_proba = model.predict_proba(X_test_scaled)[:, 1]


precisions, recalls, thresholds = precision_recall_curve(y_test, y_pred_proba)
f1_scores = 2 * (precisions * recalls) / (precisions + recalls)
optimal_threshold = thresholds[np.argmax(f1_scores[:-1])]


y_pred = (y_pred_proba >= optimal_threshold).astype(int)


print(f"Optimal threshold: {optimal_threshold:.3f}")
print("\nClassification Report with Optimal Threshold:")
print(classification_report(y_test, y_pred, zero_division=0))


roc_auc = roc_auc_score(y_test, y_pred_proba)
print(f"\nROC AUC Score: {roc_auc:.3f}")


feature_importance = pd.DataFrame({
    "feature": X.columns,
    "importance": abs(model.coef_[0])
```

```
})
feature_importance = feature_importance.sort_values("importance",␣
 ↪ascending=False)
print("\nTop 10 Most Important Features:")
print(feature_importance.head(10))


plt.figure(figsize=(8, 6))
sns.heatmap(confusion_matrix(y_test, y_pred),
            annot=True,
            fmt="d",
            cmap="Blues",
            xticklabels=["Not Homeless", "Homeless"],
            yticklabels=["Not Homeless", "Homeless"])
plt.title("Confusion Matrix")
plt.ylabel("True Label")
plt.xlabel("Predicted Label")
plt.show()
```

[272]:
```
fpr, tpr, thresholds = roc_curve(y_test, y_pred)

# use the model to make an roc curve

x = np.linspace(0, 1, 100)
plt.plot(fpr, tpr)
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Curve (real NYC)')

# random chance line
plt.plot(x,x, color="grey", linestyle="--")

plt.show()
```



[273]:
```
# hypothetical scenario with same number of homeless and non homeless people

n_non_homeless = 88025

nyc_slice = PIT_sliced[PIT_sliced["CoC Number"]=="NY-600"]
total_homeless = nyc_slice["Overall Homeless"].values[0]

# Calculate demographic percentages
race_cols = ["Overall Homeless - White",
             "Overall Homeless - Black, African American, or African",
             "Overall Homeless - Asian or Asian American",
             "Overall Homeless - American Indian, Alaska Native, or Indigenous",
             "Overall Homeless - Native Hawaiian or Other Pacific Islander",
             "Overall Homeless - Multiple Races"]

race_probs = []
for col in race_cols:
    race_probs.append(nyc_slice[col].values[0] / total_homeless)
```

```
# Create simulated population
np.random.seed(42)
n_simulated = 88025

# Simulate race based on observed proportions
simulated_race = np.random.choice(len(race_cols), size=n_simulated,␣
 ↪p=race_probs)

# Create dataframe with simulated data
simulated_df = pd.DataFrame({
    "race": [race_cols[i].replace("Overall Homeless - ", "") for i in␣
 ↪simulated_race]
})

# Add age based on NYC proportions
age_cols = ["Overall Homeless - Under 18",
            "Overall Homeless - Age 18 to 24",
            "Overall Homeless - Age 25 to 34",
            "Overall Homeless - Age 35 to 44",
            "Overall Homeless - Age 45 to 54",
            "Overall Homeless - Age 55 to 64",
            "Overall Homeless - Over 64"]

age_probs = []
for col in age_cols:
    age_probs.append(nyc_slice[col].values[0] / total_homeless)

simulated_df["age"] = np.random.choice(
    [col.replace("Overall Homeless - ", "").replace("Age ", "") for col in␣
 ↪age_cols],
    size=n_simulated,
    p=age_probs
)

# Add gender based on NYC proportions
gender_cols = ["Overall Homeless - Female",
               "Overall Homeless - Male",
               "Overall Homeless - Transgender",
               "Overall Homeless - Gender that is not Singularly Female or Male",
               "Overall Homeless - Gender Questioning"]

gender_probs = []
for col in gender_cols:
    gender_probs.append(nyc_slice[col].values[0] / total_homeless)

simulated_df["gender"] = np.random.choice(
    [col.replace("Overall Homeless - ", "") for col in gender_cols],
```

```
    size=n_simulated,
    p=gender_probs
)


simulated_df["homeless"] = 1


non_homeless_df = pd.DataFrame()

race_probs = {
    "White": 0.375,
    "Black, African American, or African": 0.231,
    "Asian or Asian American": 0.145,
    "Multiple Races": 0.089,
    "American Indian, Alaska Native, or Indigenous": 0.006,
    "Native Hawaiian or Other Pacific Islander": 0.001,
    "Hispanic or Latino": 0.29 - 0.137
}

non_homeless_df["race"] = np.random.choice(
    list(race_probs.keys()),
    size=n_non_homeless,
    p=list(race_probs.values())
)

# Age distribution based on NYC census
age_dist = {
    "Under 18": 0.21,
    "18 to 24": 0.10,
    "25 to 34": 0.17,
    "35 to 44": 0.14,
    "45 to 54": 0.13,
    "55 to 64": 0.13,
    "Over 64": 0.12
}

non_homeless_df["age"] = np.random.choice(
    list(age_dist.keys()),
    size=n_non_homeless,
    p=list(age_dist.values())
)

# Gender distribution based on NYC census
gender_dist = {
    "Female": 0.52 - (0.0012 +  0.0007 + 0.0001)/2,
    "Male": 0.48 - (0.0012 +  0.0007 + 0.0001)/2,
```

```
    "Other": 0.0012 +  0.0007 + 0.0001
}

non_homeless_df["gender"] = np.random.choice(
    list(gender_dist.keys()),
    size=n_non_homeless,
    p=list(gender_dist.values())
)

non_homeless_df["homeless"] = 0

# Combine homeless and non-homeless populations
simulated_df = pd.concat([simulated_df, non_homeless_df], ignore_index=True)
#simulated_df.drop("ethnicity", axis=1, inplace=True)
```

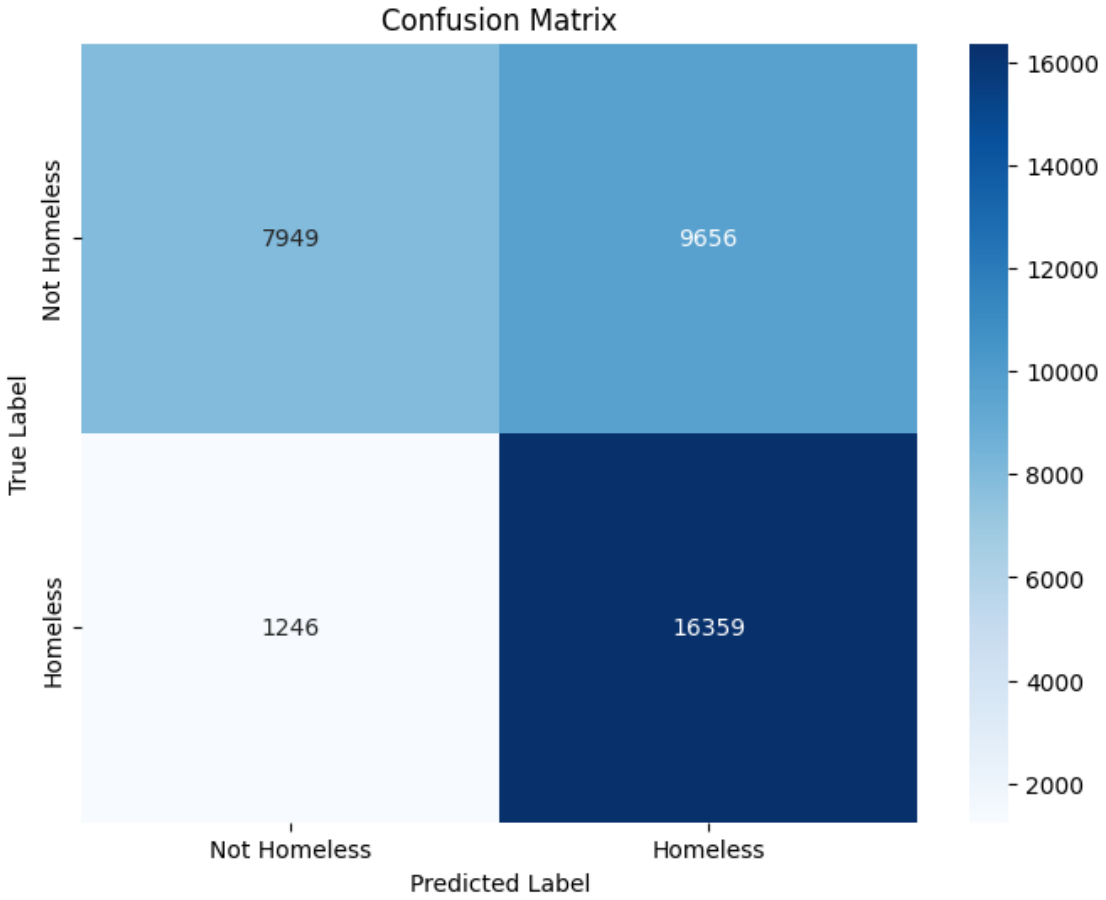
[274]:
```
df = simulated_df.copy()


X = pd.get_dummies(df[["race", "age", "gender"]], drop_first=True)
y = df["homeless"]


X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,␣
 ↪random_state=42, stratify=y)


scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)


model = LogisticRegression(
    class_weight="balanced",
    max_iter=1000,
    random_state=42,
    C=0.1
)
model.fit(X_train, y_train)

y_pred_proba = model.predict_proba(X_test_scaled)[:, 1]


precisions, recalls, thresholds = precision_recall_curve(y_test, y_pred_proba)
f1_scores = 2 * (precisions * recalls) / (precisions + recalls)
optimal_threshold = thresholds[np.argmax(f1_scores[:-1])]
```

```
y_pred = (y_pred_proba >= optimal_threshold).astype(int)


print(f"Optimal threshold: {optimal_threshold:.3f}")
print("\nClassification Report with Optimal Threshold:")
print(classification_report(y_test, y_pred, zero_division=0))


roc_auc = roc_auc_score(y_test, y_pred_proba)
print(f"\nROC AUC Score: {roc_auc:.3f}")


feature_importance = pd.DataFrame({
    "feature": X.columns,
    "importance": abs(model.coef_[0])
})
feature_importance = feature_importance.sort_values("importance",␣
 ↪ascending=False)
print("\nTop 10 Most Important Features:")
print(feature_importance.head(10))


plt.figure(figsize=(8, 6))
sns.heatmap(confusion_matrix(y_test, y_pred),
            annot=True,
            fmt="d",
            cmap="Blues",
            xticklabels=["Not Homeless", "Homeless"],
            yticklabels=["Not Homeless", "Homeless"])
plt.title("Confusion Matrix")
plt.ylabel("True Label")
plt.xlabel("Predicted Label")
plt.show()
```

```
Optimal threshold: 0.473

Classification Report with Optimal Threshold:
              precision    recall  f1-score   support

           0       0.86      0.45      0.59     17605
           1       0.63      0.93      0.75     17605

    accuracy                           0.69     35210
   macro avg       0.75      0.69      0.67     35210
weighted avg       0.75      0.69      0.67     35210


ROC AUC Score: 0.803
```

```
Top 10 Most Important Features:
                                              feature  importance
2                             race_Hispanic or Latino    5.321921
0                        race_Asian or Asian American    2.855276
16                                 gender_Transgender    2.612394
15                                       gender_Other    1.689474
10                                        age_Over 64    1.105533
1             race_Black, African American, or African    0.964928
13  gender_Gender that is not Singularly Female or…    0.816621
4       race_Native Hawaiian or Other Pacific Islander    0.703543
3                                 race_Multiple Races    0.624683
5                                         race_White    0.525241
```

```
/Library/Frameworks/Python.framework/Versions/3.12/lib/python3.12/site-
packages/sklearn/base.py:493: UserWarning:

X does not have valid feature names, but LogisticRegression was fitted with
feature names
```



[ ]:

[275]:
```
fpr, tpr, thresholds = roc_curve(y_test, y_pred)

# use the model to make an roc curve

x = np.linspace(0, 1, 100)
plt.plot(fpr, tpr)
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Curve (fake NYC)')

# random chance line
plt.plot(x,x, color="grey", linestyle="--")

plt.show()
```