bazalewski_capstone_recommender

April 28, 2022

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

#from scipy.spatial import distance
#from sklearn.metrics.pairwise import linear_kernel

from sklearn.neighbors import NearestNeighbors
import numpy as np

from statsmodels.stats.outliers_influence import variance_inflation_factor
```

[2]:
```
census_df = pd.read_csv('cleaned_census.csv')
census_df = census_df.drop('Unnamed: 0', axis=1)
census_df = census_df.replace('-',np.NAN)
census_df = census_df.replace('+','')
census_df = census_df.dropna().reset_index(drop=True)
```

[3]:
```
census_df.columns
```

[3]:
```
Index(['ZCTA', 'Total Households', 'Percent Married Couple Family',
       'Percent Married Couple Family with Children',
       'Percent Male Householder', 'Percent Female Householder',
       'Average Household Size', 'Average Family Size',
       'Percent Males Never Married', 'Percent Males Married',
       'Percent Males Divorced', 'Percent Females Never Married',
       'Percent Females Married', 'Percent Females Divorced',
       'Percent High School Grad', 'Percent Assoc Deg',
       'Percent Bachelors Deg', 'Percent Graduate Deg', 'Percent Disabled',
       'Total Pop 16 and Up', 'Percent in Labor Force', 'Unemployment Rate',
       'Percent Private Sector', 'Percent Govt Workers',
       'Percent Self Employed', 'Median Income', 'Mean Income',
       'Per Capita Income', 'Percent 2 Bedroom Homes',
       'Percent 3 Bedroom Homes', 'Percent 4 Bedroom Homes',
       'Median House Value', 'Median Mortgage',
       'Tot Housing Units with Mortgage',
```

```
       'Mortgage Less than 20 Percent of Income',
       'Mortgage Between 20 and 25 Percent of Income',
       'Mortgage Between 25 and 30 Percent of Income',
       'Mortgage Between 30 and 35 Percent of Income',
       'Mortgage More than 35 Percent of Income', 'Total Units Paying Rent',
       'Rent Less than 15 Percent of Income',
       'Rent Between 15 and 20 Percent of Income',
       'Rent Between 20 and 25 Percent of Income',
       'Rent Between 20 and 25 Percent of Income.1',
       'Rent Between 25 and 30 Percent of Income',
       'Rent Between 30 and 35 Percent of Income',
       'Rent More than 35 Percent of Income', 'Total Pop', 'Percent Male',
       'Percent Female', 'Median Age', 'Percent Under 18',
       'Percent 62 and Over', 'Percent 65 and Over', 'Percent White',
       'Percent Black', 'Percent Asian', 'Percent Hispanic'],
      dtype='object')
```

[4]:
```
state_zip_df = pd.read_csv('state_zip.csv')
census_df = census_df.
 ↪merge(state_zip_df,how='left',left_on='ZCTA',right_on='Zipcode')
```

[5]:
```
census_df_labels = census_df[['ZCTA','City','State']]

census_df_labels
```

[5]:

	ZCTA	City	State
0	1001	AGAWAM	MA
1	1002	AMHERST	MA
2	1005	BARRE	MA
3	1007	BELCHERTOWN	MA
4	1010	BRIMFIELD	MA
…	…	…	…
26111	99919	THORNE BAY	AK
26112	99921	CRAIG	AK
26113	99925	KLAWOCK	AK
26114	99926	METLAKATLA	AK
26115	99929	WRANGELL	AK

[26116 rows x 3 columns]

[6]:
```
census_df = census_df.loc[:,(census_df.columns!='ZCTA')&
                                (census_df.columns!='City') &
                                (census_df.columns!='State')]
census_df = census_df.astype(float)
```

[7]:
```
#check for and fix multicollinearity
```

```
vif_df = pd.DataFrame()

vif_df["feature"] = census_df.columns

vif_df["VIF"] = [variance_inflation_factor(census_df.values, i)
                          for i in range(len(census_df.columns))]

print(vif_df)
print('\n')
```

```
                                      feature           VIF
0                               Total Households  2.492275e+02
1                   Percent Married Couple Family  7.595637e+02
2    Percent Married Couple Family with Children  3.981199e+01
3                         Percent Male Householder  6.289623e+01
4                       Percent Female Householder  1.105114e+02
5                           Average Household Size  7.348391e+02
6                              Average Family Size  5.989733e+02
7                      Percent Males Never Married  1.990573e+02
8                            Percent Males Married  7.146141e+02
9                           Percent Males Divorced  2.846997e+01
10                  Percent Females Never Married  7.852323e+01
11                        Percent Females Married  4.838344e+02
12                       Percent Females Divorced  2.005092e+01
13                       Percent High School Grad  4.855859e+01
14                              Percent Assoc Deg  1.086419e+01
15                         Percent Bachelors Deg  2.434007e+01
16                          Percent Graduate Deg  1.380715e+01
17                              Percent Disabled  2.026857e+01
18                           Total Pop 16 and Up  7.836541e+02
19                         Percent in Labor Force  1.512342e+02
20                              Unemployment Rate  4.611772e+00
21                         Percent Private Sector  1.106870e+04
22                           Percent Govt Workers  5.283610e+02
23                          Percent Self Employed  1.546322e+02
24                                  Median Income  1.250984e+02
25                                    Mean Income  3.095268e+02
26                              Per Capita Income  2.187790e+02
27                         Percent 2 Bedroom Homes  3.004610e+01
28                         Percent 3 Bedroom Homes  5.022813e+01
29                         Percent 4 Bedroom Homes  1.882163e+01
30                              Median House Value  1.490568e+01
31                                 Median Mortgage  7.137184e+01
32                 Tot Housing Units with Mortgage  5.900618e+01
33       Mortgage Less than 20 Percent of Income  6.091782e+05
34  Mortgage Between 20 and 25 Percent of Income  6.688573e+04
35  Mortgage Between 25 and 30 Percent of Income  3.244927e+04
```

```
36  Mortgage Between 30 and 35 Percent of Income  1.665414e+04
37       Mortgage More than 35 Percent of Income  1.301769e+05
38                         Total Units Paying Rent  3.226654e+01
39             Rent Less than 15 Percent of Income  3.758597e+01
40       Rent Between 15 and 20 Percent of Income  1.163395e+05
41       Rent Between 20 and 25 Percent of Income  6.835094e+04
42    Rent Between 20 and 25 Percent of Income.1  5.620552e+04
43       Rent Between 25 and 30 Percent of Income  4.251237e+04
44       Rent Between 30 and 35 Percent of Income  2.760454e+04
45            Rent More than 35 Percent of Income  3.182756e+05
46                                      Total Pop  5.592547e+02
47                                   Percent Male  1.155990e+06
48                                 Percent Female  1.171290e+06
49                                     Median Age  2.795888e+02
50                                Percent Under 18  9.047179e+01
51                             Percent 62 and Over  2.489791e+02
52                             Percent 65 and Over  1.771691e+02
53                                   Percent White  1.299885e+02
54                                   Percent Black  6.580474e+00
55                                   Percent Asian  3.148621e+00
56                                Percent Hispanic  3.426226e+00
57                                         Zipcode  7.209993e+00
```

[8]:
```
census_df_subset = census_df[[
       'Percent Married Couple Family with Children',
       'Percent Males Divorced',
       'Percent Bachelors Deg',
       'Percent Disabled',
       'Unemployment Rate',
       'Percent Govt Workers',
       'Percent Self Employed',
       'Percent 4 Bedroom Homes',
       'Median House Value',
       'Total Pop',
       'Percent Black', 'Percent Asian', 'Percent Hispanic',
       'Mortgage Between 20 and 25 Percent of Income',
       'Mortgage Between 30 and 35 Percent of Income',
       'Mortgage More than 35 Percent of Income',
       'Rent Between 15 and 20 Percent of Income',
       'Rent Between 20 and 25 Percent of Income',
       'Rent Between 25 and 30 Percent of Income',
       'Rent Between 30 and 35 Percent of Income']]

vif_df = pd.DataFrame()
vif_df["feature"] = census_df_subset.columns
```

```
vif_df["VIF"] = [variance_inflation_factor(census_df_subset.values, i)
                          for i in range(len(census_df_subset.columns))]

print(vif_df)
```

```
                                        feature        VIF
0    Percent Married Couple Family with Children  10.183023
1                          Percent Males Divorced   7.270229
2                           Percent Bachelors Deg   9.199497
3                                Percent Disabled  10.461332
4                              Unemployment Rate   3.878860
5                            Percent Govt Workers   5.254223
6                           Percent Self Employed   3.798134
7                         Percent 4 Bedroom Homes   7.984256
8                              Median House Value   5.145396
9                                       Total Pop   2.808340
10                                Percent Black   1.704563
11                                Percent Asian   1.878822
12                             Percent Hispanic   1.998047
13  Mortgage Between 20 and 25 Percent of Income   5.520180
14  Mortgage Between 30 and 35 Percent of Income   2.813584
15       Mortgage More than 35 Percent of Income   6.729897
16      Rent Between 15 and 20 Percent of Income   3.164907
17      Rent Between 20 and 25 Percent of Income   2.859721
18      Rent Between 25 and 30 Percent of Income   2.691519
19      Rent Between 30 and 35 Percent of Income   2.278236
```

[9]:
```
def find_KNN(df,df_y,knn,lookup,state='All'):

    if state != 'All':
        df = pd.concat([df.loc[df_y['State']==state],df.
↪loc[df_y['ZCTA']==lookup]])
        df_y = pd.
↪concat([df_y[df_y['State']==state],df_y[df_y['ZCTA']==lookup]]).
↪reset_index(drop=True)
        df = df.reset_index(drop=True)

    X = df.to_numpy()
    nbrs = NearestNeighbors(n_neighbors=knn, algorithm='ball_tree').fit(X)
    distances, indices = nbrs.kneighbors(X)

    df_y[df_y['ZCTA']==lookup]
    i = df_y[df_y['ZCTA']==lookup].index.values[0]
    zips = df_y.iloc[indices[i][1:knn+1]]
    print(zips)
    return zips
```

0.1 Tests

[10]:
```
#full model test, Brooklyn zip code
zips = find_KNN(census_df,census_df_labels,5,11201)
```

	ZCTA	City	State
2091	10019	NEW YORK	NY
5372	22102	MC LEAN	VA
368	2445	BROOKLINE	MA
381	2467	CHESTNUT HILL	MA

[11]:
```
#full model test, Brooklyn zip code, Virgina results
zips = find_KNN(census_df,census_df_labels,5,11201,'VA')
```

	ZCTA	City	State
58	22102	MC LEAN	VA
79	22207	ARLINGTON	VA
69	22182	VIENNA	VA
77	22205	ARLINGTON	VA

[12]:
```
#variable subset test, Brooklyn zip code
zips = find_KNN(census_df_subset,census_df_labels,5,11201)
```

	ZCTA	City	State
23638	90019	LOS ANGELES	CA
23940	92024	ENCINITAS	CA
25149	96816	HONOLULU	HI
2306	11221	BROOKLYN	NY

[13]:
```
#variable subset test, Brooklyn zip code, Virgina results
zips = find_KNN(census_df_subset,census_df_labels,5,11201,'VA')
```

	ZCTA	City	State
79	22207	ARLINGTON	VA
58	22102	MC LEAN	VA
57	22101	MC LEAN	VA
439	24011	ROANOKE	VA

0.2 Suggested Areas

[14]:
```
zip_codes = [15317,15227]

states = ['NY','NJ','OH','WV','MD','VA','NC','SC','GA','FL']

zips = pd.DataFrame()
```

[15]:
```
#variable subset, All States
for i in zip_codes:
```

```
    find_KNN(census_df_subset,census_df_labels,5,i)
```

```
        ZCTA           City State
24728  95301        ATWATER    CA
22100  80014         AURORA    CO
8318   33060  POMPANO BEACH    FL
24045  92307   APPLE VALLEY    CA
        ZCTA          City State
2521   12010     AMSTERDAM    NY
19390  70607  LAKE CHARLES    LA
10904  43512      DEFIANCE    OH
17084  62226    BELLEVILLE    IL
```

[16]:
```
#variable subset, selected states
for i in states:
    for j in zip_codes:
        zips = pd.
 ↪concat([zips,find_KNN(census_df_subset,census_df_labels,5,j,i)])
```

```
       ZCTA          City State
1252  14534     PITTSFORD    NY
655   12603  POUGHKEEPSIE    NY
632   12553   NEW WINDSOR    NY
437   11967       SHIRLEY    NY
       ZCTA         City State
446   12010    AMSTERDAM    NY
803   13045     CORTLAND    NY
1293  14623    ROCHESTER    NY
555   12304  SCHENECTADY    NY
     ZCTA          City State
337  8080        SEWELL    NJ
317  8054  MOUNT LAUREL    NJ
91   7201     ELIZABETH    NJ
249  7860        NEWTON    NJ
     ZCTA                 City State
472  8759  MANCHESTER TOWNSHIP    NJ
301  8030      GLOUCESTER CITY    NJ
376  8232        PLEASANTVILLE    NJ
362  8110           PENNSAUKEN    NJ
      ZCTA        City State
426  44145    WESTLAKE    OH
92   43201    COLUMBUS    OH
99   43209    COLUMBUS    OH
763  45244  CINCINNATI    OH
      ZCTA        City State
188  43512    DEFIANCE    OH
759  45240  CINCINNATI    OH
521  44460       SALEM    OH
```

```
225  43606      TOLEDO    OH
      ZCTA           City State
266  26508     MORGANTOWN    WV
95   25414   CHARLES TOWN    WV
90   25403    MARTINSBURG    WV
102  25430  KEARNEYSVILLE    WV
      ZCTA        City State
146  25701  HUNTINGTON    WV
150  25705  HUNTINGTON    WV
151  25801     BECKLEY    WV
4    24740   PRINCETON    WV
      ZCTA          City State
78   20785   HYATTSVILLE    MD
61   20748  TEMPLE HILLS    MD
141  21060   GLEN BURNIE    MD
77   20784   HYATTSVILLE    MD
      ZCTA        City State
192  21217   BALTIMORE    MD
216  21502  CUMBERLAND    MD
221  21532   FROSTBURG    MD
204  21229   BALTIMORE    MD
      ZCTA          City State
303  23321    CHESAPEAKE    VA
168  22801  HARRISONBURG    VA
305  23323    CHESAPEAKE    VA
131  22602    WINCHESTER    VA
      ZCTA          City State
586  24501     LYNCHBURG    VA
356  23607  NEWPORT NEWS    VA
440  24012       ROANOKE    VA
394  23847       EMPORIA    VA
      ZCTA        City State
577  28607       BOONE    NC
350  28105    MATTHEWS    NC
478  28411  WILMINGTON    NC
124  27511        CARY    NC
      ZCTA            City State
609  28658          NEWTON    NC
373  28150          SHELBY    NC
340  28086  KINGS MOUNTAIN    NC
374  28152          SHELBY    NC
      ZCTA        City State
135  29414  CHARLESTON    SC
290  29715   FORT MILL    SC
86   29205    COLUMBIA    SC
7    29016  BLYTHEWOOD    SC
      ZCTA        City State
304  29801       AIKEN    SC
```

```
244  29640      EASLEY    SC
146  29440  GEORGETOWN    SC
70   29154      SUMTER    SC
      ZCTA     City State
174  30316  ATLANTA    GA
356  30809    EVANS    GA
11   30019   DACULA    GA
50   30082   SMYRNA    GA
      ZCTA            City State
6    30012         CONYERS    GA
54   30088  STONE MOUNTAIN    GA
391  31021          DUBLIN    GA
472  31404        SAVANNAH    GA
      ZCTA           City State
377  33060  POMPANO BEACH    FL
440  33183          MIAMI    FL
347  33014        HIALEAH    FL
116  32309    TALLAHASSEE    FL
      ZCTA              City State
605  33709  SAINT PETERSBURG    FL
315  32905          PALM BAY    FL
5    32025         LAKE CITY    FL
798  34472             OCALA    FL
```

[17]:
```
zips
```

[17]:
```
       ZCTA              City State
1252  14534         PITTSFORD    NY
655   12603      POUGHKEEPSIE    NY
632   12553       NEW WINDSOR    NY
437   11967           SHIRLEY    NY
446   12010         AMSTERDAM    NY
…     …                 …   …
116   32309       TALLAHASSEE    FL
605   33709  SAINT PETERSBURG    FL
315   32905          PALM BAY    FL
5     32025         LAKE CITY    FL
798   34472             OCALA    FL

[80 rows x 3 columns]
```

[18]:
```
zips.to_csv('recommended_zips.csv')
```

[ ]:
```

```