u6511286_JanPoonthong_541_W3

June 30, 2023

[633]:
```
import pandas as pd
import matplotlib.pyplot as plt
```

[634]:
```
taxi_owner = pd.read_pickle('data-sets/taxi_owners.p')
```

[635]:
```
taxi_owner.head()
```

[635]:

	rid	vid	owner	address	zip
0	T6285	6285	AGEAN TAXI LLC	4536 N. ELSTON AVE.	60630
1	T4862	4862	MANGIB CORP.	5717 N. WASHTENAW AVE.	60659
2	T1495	1495	FUNRIDE, INC.	3351 W. ADDISON ST.	60618
3	T4231	4231	ALQUSH CORP.	6611 N. CAMPBELL AVE.	60645
4	T5971	5971	EUNIFFORD INC.	3351 W. ADDISON ST.	60618

[636]:
```
print(taxi_owner)
```

```
        rid   vid             owner                 address    zip
0     T6285  6285    AGEAN TAXI LLC     4536 N. ELSTON AVE.  60630
1     T4862  4862      MANGIB CORP.  5717 N. WASHTENAW AVE.  60659
2     T1495  1495     FUNRIDE, INC.     3351 W. ADDISON ST.  60618
3     T4231  4231      ALQUSH CORP.   6611 N. CAMPBELL AVE.  60645
4     T5971  5971    EUNIFFORD INC.     3351 W. ADDISON ST.  60618
…     …   …               …                     …    …
3514  T4453  4453   IMAGIN CAB CORP     3351 W. ADDISON ST.  60618
3515   T121   121  TRIBECA CAB CORP     4536 N. ELSTON AVE.  60630
3516  T3465  3465  AMIR EXPRESS INC     3351 W. ADDISON ST.  60618
3517  T1962  1962  KARY CAB COMPANY     4707 N. KENTON AVE.  60630
3518  T1031  1031       NECT 42 LLC    6500 N. WESTERN AVE.  60645

[3519 rows x 5 columns]
```

[637]:
```
taxi_owner.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3519 entries, 0 to 3518
Data columns (total 5 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   rid      3519 non-null   object
```

```
 1   vid      3519 non-null   object
 2   owner    3519 non-null   object
 3   address  3519 non-null   object
 4   zip      3519 non-null   object
dtypes: object(5)
memory usage: 137.6+ KB
```

[638]:
```
taxi_owner.describe()
```

[638]:
```
          rid   vid              owner              address    zip
count    3519  3519               3519                 3519   3519
unique   3519  3519               2375                  317     44
top     T6285  6285  CHICAGO SEVEN INC  3351 W. ADDISON ST.  60618
freq        1     1                 21                  639    798
```

[639]:
```
taxi_owner.shape
```

[639]:
```
(3519, 5)
```

[640]:
```
taxi_owner.values
```

[640]:
```
array([['T6285', '6285', 'AGEAN TAXI LLC', '4536 N. ELSTON AVE.',
        '60630'],
       ['T4862', '4862', 'MANGIB CORP.', '5717 N. WASHTENAW AVE.',
        '60659'],
       ['T1495', '1495', 'FUNRIDE, INC.', '3351 W. ADDISON ST.', '60618'],
       …,
       ['T3465', '3465', 'AMIR EXPRESS INC', '3351 W. ADDISON ST.',
        '60618'],
       ['T1962', '1962', 'KARY CAB COMPANY', '4707 N. KENTON AVE.',
        '60630'],
       ['T1031', '1031', 'NECT 42 LLC', '6500 N. WESTERN AVE.', '60645']],
      dtype=object)
```

[641]:
```
taxi_owner.columns
```

[641]:
```
Index(['rid', 'vid', 'owner', 'address', 'zip'], dtype='object')
```

[642]:
```
taxi_owner.index
```

[642]:
```
RangeIndex(start=0, stop=3519, step=1)
```

[643]:
```
homelessness_df = pd.read_csv("data-sets/homelessness.csv", index_col=0)
homelessness_df
```

[643]:
```
                region                state  individuals  family_members  \
0   East South Central              Alabama       2570.0           864.0
1              Pacific               Alaska       1434.0           582.0
```

2	Mountain	Arizona	7259.0	2606.0
3	West South Central	Arkansas	2280.0	432.0
4	Pacific	California	109008.0	20964.0
5	Mountain	Colorado	7607.0	3250.0
6	New England	Connecticut	2280.0	1696.0
7	South Atlantic	Delaware	708.0	374.0
8	South Atlantic	District of Columbia	3770.0	3134.0
9	South Atlantic	Florida	21443.0	9587.0
10	South Atlantic	Georgia	6943.0	2556.0
11	Pacific	Hawaii	4131.0	2399.0
12	Mountain	Idaho	1297.0	715.0
13	East North Central	Illinois	6752.0	3891.0
14	East North Central	Indiana	3776.0	1482.0
15	West North Central	Iowa	1711.0	1038.0
16	West North Central	Kansas	1443.0	773.0
17	East South Central	Kentucky	2735.0	953.0
18	West South Central	Louisiana	2540.0	519.0
19	New England	Maine	1450.0	1066.0
20	South Atlantic	Maryland	4914.0	2230.0
21	New England	Massachusetts	6811.0	13257.0
22	East North Central	Michigan	5209.0	3142.0
23	West North Central	Minnesota	3993.0	3250.0
24	East South Central	Mississippi	1024.0	328.0
25	West North Central	Missouri	3776.0	2107.0
26	Mountain	Montana	983.0	422.0
27	West North Central	Nebraska	1745.0	676.0
28	Mountain	Nevada	7058.0	486.0
29	New England	New Hampshire	835.0	615.0
30	Mid-Atlantic	New Jersey	6048.0	3350.0
31	Mountain	New Mexico	1949.0	602.0
32	Mid-Atlantic	New York	39827.0	52070.0
33	South Atlantic	North Carolina	6451.0	2817.0
34	West North Central	North Dakota	467.0	75.0
35	East North Central	Ohio	6929.0	3320.0
36	West South Central	Oklahoma	2823.0	1048.0
37	Pacific	Oregon	11139.0	3337.0
38	Mid-Atlantic	Pennsylvania	8163.0	5349.0
39	New England	Rhode Island	747.0	354.0
40	South Atlantic	South Carolina	3082.0	851.0
41	West North Central	South Dakota	836.0	323.0
42	East South Central	Tennessee	6139.0	1744.0
43	West South Central	Texas	19199.0	6111.0
44	Mountain	Utah	1904.0	972.0
45	New England	Vermont	780.0	511.0
46	South Atlantic	Virginia	3928.0	2047.0
47	Pacific	Washington	16424.0	5880.0
48	South Atlantic	West Virginia	1021.0	222.0

49	East North Central	Wisconsin	2740.0	2167.0
50	Mountain	Wyoming	434.0	205.0

	state_pop
0	4887681
1	735139
2	7158024
3	3009733
4	39461588
5	5691287
6	3571520
7	965479
8	701547
9	21244317
10	10511131
11	1420593
12	1750536
13	12723071
14	6695497
15	3148618
16	2911359
17	4461153
18	4659690
19	1339057
20	6035802
21	6882635
22	9984072
23	5606249
24	2981020
25	6121623
26	1060665
27	1925614
28	3027341
29	1353465
30	8886025
31	2092741
32	19530351
33	10381615
34	758080
35	11676341
36	3940235
37	4181886
38	12800922
39	1058287
40	5084156
41	878698
42	6771631

```
43   28628666
44    3153550
45     624358
46    8501286
47    7523869
48    1804291
49    5807406
50     577601
```

[644]:
```
homelessness_df.head()
```

[644]:

	region	state	individuals	family_members	state_pop
0	East South Central	Alabama	2570.0	864.0	4887681
1	Pacific	Alaska	1434.0	582.0	735139
2	Mountain	Arizona	7259.0	2606.0	7158024
3	West South Central	Arkansas	2280.0	432.0	3009733
4	Pacific	California	109008.0	20964.0	39461588

[645]:
```
homelessness_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 51 entries, 0 to 50
Data columns (total 5 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   region          51 non-null     object
 1   state           51 non-null     object
 2   individuals     51 non-null     float64
 3   family_members  51 non-null     float64
 4   state_pop       51 non-null     int64
dtypes: float64(2), int64(1), object(2)
memory usage: 2.4+ KB
```

[646]:
```
homelessness_df.describe()
```

[646]:

	individuals	family_members	state_pop
count	51.000000	51.000000	5.100000e+01
mean	7225.784314	3504.882353	6.405637e+06
std	15991.025083	7805.411811	7.327258e+06
min	434.000000	75.000000	5.776010e+05
25%	1446.500000	592.000000	1.777414e+06
50%	3082.000000	1482.000000	4.461153e+06
75%	6781.500000	3196.000000	7.340946e+06
max	109008.000000	52070.000000	3.946159e+07

[647]:
```
homelessness_df.shape
```

[647]:
```
(51, 5)
```

[648]:
```
homelessness_df.values
```

[648]:
```
array([['East South Central', 'Alabama', 2570.0, 864.0, 4887681],
       ['Pacific', 'Alaska', 1434.0, 582.0, 735139],
       ['Mountain', 'Arizona', 7259.0, 2606.0, 7158024],
       ['West South Central', 'Arkansas', 2280.0, 432.0, 3009733],
       ['Pacific', 'California', 109008.0, 20964.0, 39461588],
       ['Mountain', 'Colorado', 7607.0, 3250.0, 5691287],
       ['New England', 'Connecticut', 2280.0, 1696.0, 3571520],
       ['South Atlantic', 'Delaware', 708.0, 374.0, 965479],
       ['South Atlantic', 'District of Columbia', 3770.0, 3134.0, 701547],
       ['South Atlantic', 'Florida', 21443.0, 9587.0, 21244317],
       ['South Atlantic', 'Georgia', 6943.0, 2556.0, 10511131],
       ['Pacific', 'Hawaii', 4131.0, 2399.0, 1420593],
       ['Mountain', 'Idaho', 1297.0, 715.0, 1750536],
       ['East North Central', 'Illinois', 6752.0, 3891.0, 12723071],
       ['East North Central', 'Indiana', 3776.0, 1482.0, 6695497],
       ['West North Central', 'Iowa', 1711.0, 1038.0, 3148618],
       ['West North Central', 'Kansas', 1443.0, 773.0, 2911359],
       ['East South Central', 'Kentucky', 2735.0, 953.0, 4461153],
       ['West South Central', 'Louisiana', 2540.0, 519.0, 4659690],
       ['New England', 'Maine', 1450.0, 1066.0, 1339057],
       ['South Atlantic', 'Maryland', 4914.0, 2230.0, 6035802],
       ['New England', 'Massachusetts', 6811.0, 13257.0, 6882635],
       ['East North Central', 'Michigan', 5209.0, 3142.0, 9984072],
       ['West North Central', 'Minnesota', 3993.0, 3250.0, 5606249],
       ['East South Central', 'Mississippi', 1024.0, 328.0, 2981020],
       ['West North Central', 'Missouri', 3776.0, 2107.0, 6121623],
       ['Mountain', 'Montana', 983.0, 422.0, 1060665],
       ['West North Central', 'Nebraska', 1745.0, 676.0, 1925614],
       ['Mountain', 'Nevada', 7058.0, 486.0, 3027341],
       ['New England', 'New Hampshire', 835.0, 615.0, 1353465],
       ['Mid-Atlantic', 'New Jersey', 6048.0, 3350.0, 8886025],
       ['Mountain', 'New Mexico', 1949.0, 602.0, 2092741],
       ['Mid-Atlantic', 'New York', 39827.0, 52070.0, 19530351],
       ['South Atlantic', 'North Carolina', 6451.0, 2817.0, 10381615],
       ['West North Central', 'North Dakota', 467.0, 75.0, 758080],
       ['East North Central', 'Ohio', 6929.0, 3320.0, 11676341],
       ['West South Central', 'Oklahoma', 2823.0, 1048.0, 3940235],
       ['Pacific', 'Oregon', 11139.0, 3337.0, 4181886],
       ['Mid-Atlantic', 'Pennsylvania', 8163.0, 5349.0, 12800922],
       ['New England', 'Rhode Island', 747.0, 354.0, 1058287],
       ['South Atlantic', 'South Carolina', 3082.0, 851.0, 5084156],
       ['West North Central', 'South Dakota', 836.0, 323.0, 878698],
       ['East South Central', 'Tennessee', 6139.0, 1744.0, 6771631],
       ['West South Central', 'Texas', 19199.0, 6111.0, 28628666],
       ['Mountain', 'Utah', 1904.0, 972.0, 3153550],
```

```
        ['New England', 'Vermont', 780.0, 511.0, 624358],
        ['South Atlantic', 'Virginia', 3928.0, 2047.0, 8501286],
        ['Pacific', 'Washington', 16424.0, 5880.0, 7523869],
        ['South Atlantic', 'West Virginia', 1021.0, 222.0, 1804291],
        ['East North Central', 'Wisconsin', 2740.0, 2167.0, 5807406],
        ['Mountain', 'Wyoming', 434.0, 205.0, 577601]], dtype=object)
```

[649]:
```
homelessness_df.columns
```

[649]:
```
Index(['region', 'state', 'individuals', 'family_members', 'state_pop'],
dtype='object')
```

[650]:
```
homelessness_df.index
```

[650]:
```
Index([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14, 15, 16, 17,
       18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35,
       36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50],
      dtype='int64')
```

1 Ex 1

[651]:
```
homelessness_ind = homelessness_df.sort_values(by="individuals").head(5)
print(homelessness_ind)
```

	region	state	individuals	family_members	state_pop
50	Mountain	Wyoming	434.0	205.0	577601
34	West North Central	North Dakota	467.0	75.0	758080
7	South Atlantic	Delaware	708.0	374.0	965479
39	New England	Rhode Island	747.0	354.0	1058287
45	New England	Vermont	780.0	511.0	624358

2 Ex 2

[652]:
```
homelessness_fam = homelessness_df.sort_values(by="family_members",␣
↪ascending=False).head(5)
print(homelessness_fam)
```

	region	state	individuals	family_members	state_pop
32	Mid-Atlantic	New York	39827.0	52070.0	19530351
4	Pacific	California	109008.0	20964.0	39461588
21	New England	Massachusetts	6811.0	13257.0	6882635
9	South Atlantic	Florida	21443.0	9587.0	21244317
43	West South Central	Texas	19199.0	6111.0	28628666

3 Ex 3

[653]:
```
homelessness_reg_fam = homelessness_df.sort_values(by=["region",␣
↪"family_members"], ascending=[True, False]).head(5)
print(homelessness_reg_fam)
```

```
                region       state  individuals  family_members  state_pop
13  East North Central    Illinois       6752.0          3891.0   12723071
35  East North Central        Ohio       6929.0          3320.0   11676341
22  East North Central    Michigan       5209.0          3142.0    9984072
49  East North Central   Wisconsin       2740.0          2167.0    5807406
14  East North Central     Indiana       3776.0          1482.0    6695497
```

4 Ex 4

[654]:
```
state_fam = homelessness_df[["state", "family_members"]].head()
print(state_fam)
```

```
        state  family_members
0     Alabama           864.0
1      Alaska           582.0
2     Arizona          2606.0
3    Arkansas           432.0
4  California         20964.0
```

5 Ex 5

[655]:
```
ind_gt_10k = homelessness_df[homelessness_df.individuals > 10_000]
print(ind_gt_10k)
```

```
                region       state  individuals  family_members  state_pop
4              Pacific  California     109008.0         20964.0   39461588
9       South Atlantic     Florida      21443.0          9587.0   21244317
32        Mid-Atlantic    New York      39827.0         52070.0   19530351
37             Pacific      Oregon      11139.0          3337.0    4181886
43  West South Central       Texas      19199.0          6111.0   28628666
47             Pacific  Washington      16424.0          5880.0    7523869
```

6 Ex 6

[656]:
```
mountain_reg = homelessness_df[homelessness_df.region == "Mountain"]
print(mountain_reg)
```

```
      region       state  individuals  family_members  state_pop
2   Mountain     Arizona       7259.0          2606.0    7158024
5   Mountain    Colorado       7607.0          3250.0    5691287
12  Mountain       Idaho       1297.0           715.0    1750536
26  Mountain     Montana        983.0           422.0    1060665
```

```
28  Mountain       Nevada        7058.0           486.0    3027341
31  Mountain   New Mexico        1949.0           602.0    2092741
44  Mountain         Utah        1904.0           972.0    3153550
50  Mountain      Wyoming         434.0           205.0     577601
```

7 Ex 7

[657]:
```
fam_it_1k_pac = homelessness_df[(homelessness_df.family_members < 1_000) &␣
 ↪(homelessness_df.region == "Pacific")]
print(fam_it_1k_pac)
```

```
    region   state  individuals  family_members  state_pop
1  Pacific  Alaska       1434.0           582.0     735139
```

8 Ex 8

[658]:
```
south_mid_atlantic = homelessness_df[(homelessness_df.region == "South␣
 ↪Atlantic") | (homelessness_df.region == "Mid-Atlantic")]
south_mid_atlantic
```

[658]:
```
             region                 state  individuals  family_members  \
7   South Atlantic              Delaware        708.0           374.0
8   South Atlantic  District of Columbia       3770.0          3134.0
9   South Atlantic               Florida      21443.0          9587.0
10  South Atlantic               Georgia       6943.0          2556.0
20  South Atlantic              Maryland       4914.0          2230.0
30    Mid-Atlantic            New Jersey       6048.0          3350.0
32    Mid-Atlantic              New York      39827.0         52070.0
33  South Atlantic        North Carolina       6451.0          2817.0
38    Mid-Atlantic          Pennsylvania       8163.0          5349.0
40  South Atlantic        South Carolina       3082.0           851.0
46  South Atlantic              Virginia       3928.0          2047.0
48  South Atlantic         West Virginia       1021.0           222.0

    state_pop
7      965479
8      701547
9    21244317
10   10511131
20    6035802
30    8886025
32   19530351
33   10381615
38   12800922
40    5084156
46    8501286
48    1804291
```

9 Ex 9

[659]:
```
mojave_homelessness = homelessness_df[homelessness_df['state'].
  ↪isin(["Arizona","California","Nevada", "Utah"])]
print(mojave_homelessness)
```

```
      region       state  individuals  family_members  state_pop
2   Mountain     Arizona       7259.0          2606.0    7158024
4    Pacific  California     109008.0         20964.0   39461588
28  Mountain      Nevada       7058.0           486.0    3027341
44  Mountain        Utah       1904.0           972.0    3153550
```

10 Ex 10

[660]:
```
homelessness_df['total'] = homelessness_df['individuals'] +␣
  ↪homelessness_df['family_members']
homelessness_df
```

[660]:

	region	state	individuals	family_members	\
0	East South Central	Alabama	2570.0	864.0	
1	Pacific	Alaska	1434.0	582.0	
2	Mountain	Arizona	7259.0	2606.0	
3	West South Central	Arkansas	2280.0	432.0	
4	Pacific	California	109008.0	20964.0	
5	Mountain	Colorado	7607.0	3250.0	
6	New England	Connecticut	2280.0	1696.0	
7	South Atlantic	Delaware	708.0	374.0	
8	South Atlantic	District of Columbia	3770.0	3134.0	
9	South Atlantic	Florida	21443.0	9587.0	
10	South Atlantic	Georgia	6943.0	2556.0	
11	Pacific	Hawaii	4131.0	2399.0	
12	Mountain	Idaho	1297.0	715.0	
13	East North Central	Illinois	6752.0	3891.0	
14	East North Central	Indiana	3776.0	1482.0	
15	West North Central	Iowa	1711.0	1038.0	
16	West North Central	Kansas	1443.0	773.0	
17	East South Central	Kentucky	2735.0	953.0	
18	West South Central	Louisiana	2540.0	519.0	
19	New England	Maine	1450.0	1066.0	
20	South Atlantic	Maryland	4914.0	2230.0	
21	New England	Massachusetts	6811.0	13257.0	
22	East North Central	Michigan	5209.0	3142.0	
23	West North Central	Minnesota	3993.0	3250.0	
24	East South Central	Mississippi	1024.0	328.0	
25	West North Central	Missouri	3776.0	2107.0	
26	Mountain	Montana	983.0	422.0	
27	West North Central	Nebraska	1745.0	676.0	

28	Mountain	Nevada	7058.0	486.0
29	New England	New Hampshire	835.0	615.0
30	Mid-Atlantic	New Jersey	6048.0	3350.0
31	Mountain	New Mexico	1949.0	602.0
32	Mid-Atlantic	New York	39827.0	52070.0
33	South Atlantic	North Carolina	6451.0	2817.0
34	West North Central	North Dakota	467.0	75.0
35	East North Central	Ohio	6929.0	3320.0
36	West South Central	Oklahoma	2823.0	1048.0
37	Pacific	Oregon	11139.0	3337.0
38	Mid-Atlantic	Pennsylvania	8163.0	5349.0
39	New England	Rhode Island	747.0	354.0
40	South Atlantic	South Carolina	3082.0	851.0
41	West North Central	South Dakota	836.0	323.0
42	East South Central	Tennessee	6139.0	1744.0
43	West South Central	Texas	19199.0	6111.0
44	Mountain	Utah	1904.0	972.0
45	New England	Vermont	780.0	511.0
46	South Atlantic	Virginia	3928.0	2047.0
47	Pacific	Washington	16424.0	5880.0
48	South Atlantic	West Virginia	1021.0	222.0
49	East North Central	Wisconsin	2740.0	2167.0
50	Mountain	Wyoming	434.0	205.0

	state_pop	total
0	4887681	3434.0
1	735139	2016.0
2	7158024	9865.0
3	3009733	2712.0
4	39461588	129972.0
5	5691287	10857.0
6	3571520	3976.0
7	965479	1082.0
8	701547	6904.0
9	21244317	31030.0
10	10511131	9499.0
11	1420593	6530.0
12	1750536	2012.0
13	12723071	10643.0
14	6695497	5258.0
15	3148618	2749.0
16	2911359	2216.0
17	4461153	3688.0
18	4659690	3059.0
19	1339057	2516.0
20	6035802	7144.0
21	6882635	20068.0

```
22    9984072    8351.0
23    5606249    7243.0
24    2981020    1352.0
25    6121623    5883.0
26    1060665    1405.0
27    1925614    2421.0
28    3027341    7544.0
29    1353465    1450.0
30    8886025    9398.0
31    2092741    2551.0
32   19530351   91897.0
33   10381615    9268.0
34     758080     542.0
35   11676341   10249.0
36    3940235    3871.0
37    4181886   14476.0
38   12800922   13512.0
39    1058287    1101.0
40    5084156    3933.0
41     878698    1159.0
42    6771631    7883.0
43   28628666   25310.0
44    3153550    2876.0
45     624358    1291.0
46    8501286    5975.0
47    7523869   22304.0
48    1804291    1243.0
49    5807406    4907.0
50     577601     639.0
```

11 Ex 11

[661]:
```
homelessness_df['p_individuals'] = homelessness_df['individuals'] /␣
↪homelessness_df['total']
homelessness_df.head()
```

[661]:
```
                region       state  individuals  family_members  state_pop  \
0  East South Central     Alabama       2570.0           864.0    4887681
1             Pacific      Alaska       1434.0           582.0     735139
2            Mountain     Arizona       7259.0          2606.0    7158024
3  West South Central    Arkansas       2280.0           432.0    3009733
4             Pacific  California     109008.0         20964.0   39461588

      total  p_individuals
0    3434.0       0.748398
1    2016.0       0.711310
```

```
2    9865.0       0.735834
3    2712.0       0.840708
4  129972.0       0.838704
```

12 Ex 12

[662]:
```
homelessness_df['indiv_per_10k'] = 10_000 * homelessness_df['individuals'] /␣
↪homelessness_df['state_pop']
homelessness_df.head()
```

[662]:
```
               region       state  individuals  family_members  state_pop  \
0  East South Central     Alabama       2570.0           864.0    4887681
1             Pacific      Alaska       1434.0           582.0     735139
2            Mountain     Arizona       7259.0          2606.0    7158024
3  West South Central    Arkansas       2280.0           432.0    3009733
4             Pacific  California     109008.0         20964.0   39461588

      total  p_individuals  indiv_per_10k
0    3434.0       0.748398       5.258117
1    2016.0       0.711310      19.506515
2    9865.0       0.735834      10.141067
3    2712.0       0.840708       7.575423
4  129972.0       0.838704      27.623825
```

[663]:
```
high_homelessness = homelessness_df[homelessness_df['indiv_per_10k'] > 20]
high_homelessness
```

[663]:
```
            region                state  individuals  family_members  \
4          Pacific           California     109008.0         20964.0
8   South Atlantic  District of Columbia       3770.0          3134.0
11         Pacific               Hawaii       4131.0          2399.0
28        Mountain               Nevada       7058.0           486.0
32    Mid-Atlantic             New York      39827.0         52070.0
37         Pacific               Oregon      11139.0          3337.0
47         Pacific           Washington      16424.0          5880.0

    state_pop     total  p_individuals  indiv_per_10k
4    39461588  129972.0       0.838704      27.623825
8      701547    6904.0       0.546060      53.738381
11    1420593    6530.0       0.632619      29.079406
28    3027341    7544.0       0.935578      23.314189
32   19530351   91897.0       0.433387      20.392363
37    4181886   14476.0       0.769481      26.636307
47    7523869   22304.0       0.736370      21.829195
```

[664]:
```
high_homelessness_srt = high_homelessness.sort_values(by="indiv_per_10k",␣
↪ascending=False)
high_homelessness_srt
```

[664]:

	region	state	individuals	family_members	state_pop	total	p_individuals	indiv_per_10k
8	South Atlantic	District of Columbia	3770.0	3134.0	701547	6904.0	0.546060	53.738381
11	Pacific	Hawaii	4131.0	2399.0	1420593	6530.0	0.632619	29.079406
4	Pacific	California	109008.0	20964.0	39461588	129972.0	0.838704	27.623825
37	Pacific	Oregon	11139.0	3337.0	4181886	14476.0	0.769481	26.636307
28	Mountain	Nevada	7058.0	486.0	3027341	7544.0	0.935578	23.314189
47	Pacific	Washington	16424.0	5880.0	7523869	22304.0	0.736370	21.829195
32	Mid-Atlantic	New York	39827.0	52070.0	19530351	91897.0	0.433387	20.392363

[665]:
```
print(high_homelessness_srt[["state", "indiv_per_10k"]])
```

	state	indiv_per_10k
8	District of Columbia	53.738381
11	Hawaii	29.079406
4	California	27.623825
37	Oregon	26.636307
28	Nevada	23.314189
47	Washington	21.829195
32	New York	20.392363

[666]:
```
sales = pd.read_csv("data-sets/sales_subset.csv", index_col=0)
sales
```

[666]:

	store	type	department	date	weekly_sales	is_holiday
0	1	A	1	2010-02-05	24924.50	False
1	1	A	1	2010-03-05	21827.90	False
2	1	A	1	2010-04-02	57258.43	False
3	1	A	1	2010-05-07	17413.94	False
4	1	A	1	2010-06-04	17558.09	False
…	…	…	…	…	…	…
10769	39	A	99	2011-12-09	895.00	False
10770	39	A	99	2012-02-03	350.00	False
10771	39	A	99	2012-06-08	450.00	False
10772	39	A	99	2012-07-13	0.06	False
10773	39	A	99	2012-10-05	915.00	False

```
       temperature_c  fuel_price_usd_per_l  unemployment
0           5.727778              0.679451         8.106
1           8.055556              0.693452         8.106
2          16.816667              0.718284         7.808
3          22.527778              0.748928         7.808
4          27.050000              0.714586         7.808
…                …                     …             …
10769       9.644444              0.834256         7.716
10770      15.938889              0.887619         7.244
10771      27.288889              0.911922         6.989
10772      25.644444              0.860145         6.623
10773      22.250000              0.955511         6.228

[10774 rows x 9 columns]
```

[667]:
```
sales.head()
```

[667]:
```
   store type  department        date  weekly_sales  is_holiday  \
0      1    A           1  2010-02-05      24924.50       False
1      1    A           1  2010-03-05      21827.90       False
2      1    A           1  2010-04-02      57258.43       False
3      1    A           1  2010-05-07      17413.94       False
4      1    A           1  2010-06-04      17558.09       False

   temperature_c  fuel_price_usd_per_l  unemployment
0       5.727778              0.679451         8.106
1       8.055556              0.693452         8.106
2      16.816667              0.718284         7.808
3      22.527778              0.748928         7.808
4      27.050000              0.714586         7.808
```

[668]:
```
sales.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 10774 entries, 0 to 10773
Data columns (total 9 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   store                 10774 non-null  int64
 1   type                  10774 non-null  object
 2   department            10774 non-null  int64
 3   date                  10774 non-null  object
 4   weekly_sales          10774 non-null  float64
 5   is_holiday            10774 non-null  bool
 6   temperature_c         10774 non-null  float64
 7   fuel_price_usd_per_l  10774 non-null  float64
 8   unemployment          10774 non-null  float64
```

```
dtypes: bool(1), float64(4), int64(2), object(2)
memory usage: 768.1+ KB
```

[669]:
```
print(sales['weekly_sales'].mean())
```

```
23843.95014850566
```

[670]:
```
print(sales['weekly_sales'].median())
```

```
12049.064999999999
```

[671]:
```
print(sales['date'].max())
```

```
2012-10-26
```

[672]:
```
print(sales['date'].min())
```

```
2010-02-05
```

[673]:
```
sales_1_1 = sales[(sales['department'] == 1) & (sales['store'] == 1)]
sales_1_1
```

[673]:

	store	type	department	date	weekly_sales	is_holiday	temperature_c	fuel_price_usd_per_l	unemployment
0	1	A	1	2010-02-05	24924.50	False	5.727778	0.679451	8.106
1	1	A	1	2010-03-05	21827.90	False	8.055556	0.693452	8.106
2	1	A	1	2010-04-02	57258.43	False	16.816667	0.718284	7.808
3	1	A	1	2010-05-07	17413.94	False	22.527778	0.748928	7.808
4	1	A	1	2010-06-04	17558.09	False	27.050000	0.714586	7.808
5	1	A	1	2010-07-02	16333.14	False	27.172222	0.705076	7.787
6	1	A	1	2010-08-06	17508.41	False	30.644444	0.693980	7.787
7	1	A	1	2010-09-03	16241.78	False	27.338889	0.680772	7.787
8	1	A	1	2010-10-01	20094.19	False	22.161111	0.687640	7.838
9	1	A	1	2010-11-05	34238.88	False	14.855556	0.710359	7.838
10	1	A	1	2010-12-03	22517.56	False	9.594444	0.715378	7.838
11	1	A	1	2011-01-07	15984.24	False	9.038889	0.786176	7.742

[674]:
```
sales_1_1 = sales_1_1.sort_values('date', ascending = True)
sales_1_1
```

[674]:

	store	type	department	date	weekly_sales	is_holiday	\
0	1	A	1	2010-02-05	24924.50	False	
1	1	A	1	2010-03-05	21827.90	False	
2	1	A	1	2010-04-02	57258.43	False	
3	1	A	1	2010-05-07	17413.94	False	
4	1	A	1	2010-06-04	17558.09	False	
5	1	A	1	2010-07-02	16333.14	False	
6	1	A	1	2010-08-06	17508.41	False	
7	1	A	1	2010-09-03	16241.78	False	
8	1	A	1	2010-10-01	20094.19	False	
9	1	A	1	2010-11-05	34238.88	False	
10	1	A	1	2010-12-03	22517.56	False	
11	1	A	1	2011-01-07	15984.24	False	

	temperature_c	fuel_price_usd_per_l	unemployment
0	5.727778	0.679451	8.106
1	8.055556	0.693452	8.106
2	16.816667	0.718284	7.808
3	22.527778	0.748928	7.808
4	27.050000	0.714586	7.808
5	27.172222	0.705076	7.787
6	30.644444	0.693980	7.787
7	27.338889	0.680772	7.787
8	22.161111	0.687640	7.838
9	14.855556	0.710359	7.838
10	9.594444	0.715378	7.838
11	9.038889	0.786176	7.742

[675]:
```
sales_1_1['cum_weekly_sales'] = sales['weekly_sales'].cumsum()
sales_1_1
```

[675]:

	store	type	department	date	weekly_sales	is_holiday	\
0	1	A	1	2010-02-05	24924.50	False	
1	1	A	1	2010-03-05	21827.90	False	
2	1	A	1	2010-04-02	57258.43	False	
3	1	A	1	2010-05-07	17413.94	False	
4	1	A	1	2010-06-04	17558.09	False	
5	1	A	1	2010-07-02	16333.14	False	
6	1	A	1	2010-08-06	17508.41	False	
7	1	A	1	2010-09-03	16241.78	False	
8	1	A	1	2010-10-01	20094.19	False	
9	1	A	1	2010-11-05	34238.88	False	
10	1	A	1	2010-12-03	22517.56	False	
11	1	A	1	2011-01-07	15984.24	False	

```
    temperature_c  fuel_price_usd_per_l  unemployment  cum_weekly_sales
0        5.727778              0.679451         8.106          24924.50
1        8.055556              0.693452         8.106          46752.40
2       16.816667              0.718284         7.808         104010.83
3       22.527778              0.748928         7.808         121424.77
4       27.050000              0.714586         7.808         138982.86
5       27.172222              0.705076         7.787         155316.00
6       30.644444              0.693980         7.787         172824.41
7       27.338889              0.680772         7.787         189066.19
8       22.161111              0.687640         7.838         209160.38
9       14.855556              0.710359         7.838         243399.26
10       9.594444              0.715378         7.838         265916.82
11       9.038889              0.786176         7.742         281901.06
```

[676]:
```
sales_1_1['cum_max_sales'] = sales['weekly_sales'].cummax()
sales_1_1
```

[676]:
```
    store type  department        date  weekly_sales  is_holiday  \
0       1    A           1  2010-02-05      24924.50       False
1       1    A           1  2010-03-05      21827.90       False
2       1    A           1  2010-04-02      57258.43       False
3       1    A           1  2010-05-07      17413.94       False
4       1    A           1  2010-06-04      17558.09       False
5       1    A           1  2010-07-02      16333.14       False
6       1    A           1  2010-08-06      17508.41       False
7       1    A           1  2010-09-03      16241.78       False
8       1    A           1  2010-10-01      20094.19       False
9       1    A           1  2010-11-05      34238.88       False
10      1    A           1  2010-12-03      22517.56       False
11      1    A           1  2011-01-07      15984.24       False

    temperature_c  fuel_price_usd_per_l  unemployment  cum_weekly_sales  \
0        5.727778              0.679451         8.106          24924.50
1        8.055556              0.693452         8.106          46752.40
2       16.816667              0.718284         7.808         104010.83
3       22.527778              0.748928         7.808         121424.77
4       27.050000              0.714586         7.808         138982.86
5       27.172222              0.705076         7.787         155316.00
6       30.644444              0.693980         7.787         172824.41
7       27.338889              0.680772         7.787         189066.19
8       22.161111              0.687640         7.838         209160.38
9       14.855556              0.710359         7.838         243399.26
10       9.594444              0.715378         7.838         265916.82
11       9.038889              0.786176         7.742         281901.06

    cum_max_sales
```

```
0        24924.50
1        24924.50
2        57258.43
3        57258.43
4        57258.43
5        57258.43
6        57258.43
7        57258.43
8        57258.43
9        57258.43
10       57258.43
11       57258.43
```

[677]:
```
print(sales_1_1[["date", "weekly_sales", "cum_weekly_sales", "cum_max_sales"]])
```

```
          date  weekly_sales  cum_weekly_sales  cum_max_sales
0   2010-02-05      24924.50          24924.50       24924.50
1   2010-03-05      21827.90          46752.40       24924.50
2   2010-04-02      57258.43         104010.83       57258.43
3   2010-05-07      17413.94         121424.77       57258.43
4   2010-06-04      17558.09         138982.86       57258.43
5   2010-07-02      16333.14         155316.00       57258.43
6   2010-08-06      17508.41         172824.41       57258.43
7   2010-09-03      16241.78         189066.19       57258.43
8   2010-10-01      20094.19         209160.38       57258.43
9   2010-11-05      34238.88         243399.26       57258.43
10  2010-12-03      22517.56         265916.82       57258.43
11  2011-01-07      15984.24         281901.06       57258.43
```

13 Ex 13

[678]:
```
store_types = sales.drop_duplicates(subset=["store", "type"])
store_types.head()
```

[678]:
```
       store type  department        date  weekly_sales  is_holiday  \
0          1    A           1  2010-02-05      24924.50       False
901        2    A           1  2010-02-05      35034.06       False
1798       4    A           1  2010-02-05      38724.42       False
2699       6    A           1  2010-02-05      25619.00       False
3593      10    B           1  2010-02-05      40212.84       False

       temperature_c  fuel_price_usd_per_l  unemployment
0           5.727778              0.679451         8.106
901         4.550000              0.679451         8.324
1798        6.533333              0.686319         8.623
2699        4.683333              0.679451         7.259
3593       12.411111              0.782478         9.765
```

[679]:
```
store_depts = sales.drop_duplicates(subset=["store", "department"])
store_depts.head()
```

[679]:
```
    store type  department        date  weekly_sales  is_holiday  \
0       1    A           1  2010-02-05      24924.50       False
12      1    A           2  2010-02-05      50605.27       False
24      1    A           3  2010-02-05      13740.12       False
36      1    A           4  2010-02-05      39954.04       False
48      1    A           5  2010-02-05      32229.38       False

    temperature_c  fuel_price_usd_per_l  unemployment
0        5.727778              0.679451         8.106
12       5.727778              0.679451         8.106
24       5.727778              0.679451         8.106
36       5.727778              0.679451         8.106
48       5.727778              0.679451         8.106
```

[680]:
```
holiday_dates = sales[sales['is_holiday'] == True].drop_duplicates('date')
holiday_dates['date']
```

[680]:
```
498     2010-09-10
691     2011-11-25
2315    2010-02-12
6735    2012-09-07
6810    2010-12-31
6815    2012-02-10
6820    2011-09-09
Name: date, dtype: object
```

14 Ex 14

[681]:
```
print(store_types['type'].value_counts())
```

```
type
A    11
B     1
Name: count, dtype: int64
```

[682]:
```
print(store_types['type'].value_counts(normalize=True))
```

```
type
A    0.916667
B    0.083333
Name: proportion, dtype: float64
```

[683]:
```
print(store_depts['department'].value_counts(sort=True))
```

```
department
1     12
```

```
55    12
72    12
71    12
67    12
      ..
37    10
48     8
50     6
39     4
43     2
Name: count, Length: 80, dtype: int64
```

[684]:
```
print(store_depts['department'].value_counts(normalize=True))
```

```
department
1     0.012917
55    0.012917
72    0.012917
71    0.012917
67    0.012917
        …
37    0.010764
48    0.008611
50    0.006459
39    0.004306
43    0.002153
Name: proportion, Length: 80, dtype: float64
```

[685]:
```
sales_all = sales["weekly_sales"].sum()
sales_all
```

[685]: 256894718.89999998

[686]:
```
sales_A = sales[sales["type"] == "A"]["weekly_sales"].sum()
sales_A
```

[686]: 233716315.01

[687]:
```
sales_B = sales[sales["type"] == "B"]["weekly_sales"].sum()
sales_B
```

[687]: 23178403.89

[688]:
```
sales_C = sales[sales["type"] == "C"]["weekly_sales"].sum()
sales_C
```

[688]: 0.0

[689]:
```
sales_propn_by_type = [sales_A / sales_all, sales_B / sales_all, sales_C /␣
↪sales_all]

print(sales_propn_by_type)
```

```
[0.9097746968515047, 0.09022530314849538, 0.0]
```

15 Ex 15

I saw [0.9097746968515047, 0.09022530314849538, 0.0] as a ouput, by that being the output. We can say that the 'type A sale' and 'type B sale' are most popular on the other hand 'type C sale' didn't make any sales

[690]:
```
import numpy as np
```

[691]:
```
sales_stats = sales.groupby('type')['weekly_sales'].agg([min, max, np.mean, np.
↪median])
print(sales_stats)
```

```
         min        max          mean    median
type
A    -1098.0  293966.05  23674.667242  11943.92
B     -798.0  232558.51  25696.678370  13336.08
```

[692]:
```
unemp_fuel_stats = sales.groupby("type")[['unemployment',␣
↪'fuel_price_usd_per_l']].agg([min, max, np.mean, np.median])
unemp_fuel_stats
```

[692]:
```
     unemployment                          fuel_price_usd_per_l            \
              min    max      mean median                  min       max
type
A          3.879  8.992  7.972611  8.067             0.664129  1.107410
B          7.170  9.765  9.279323  9.199             0.760023  1.107674


          mean    median
type
A     0.744619  0.735455
B     0.805858  0.803348
```

[693]:
```
print(sales.pivot_table(values='weekly_sales', index='department',␣
↪columns='type', fill_value=0))
```

```
type                    A             B
department
1            30961.725379   44050.626667
2            67600.158788  112958.526667
3            17160.002955   30580.655000
4            44285.399091   51219.654167
```

```
5            34821.011364   63236.875000
…                     …            …
95          123933.787121   77082.102500
96           21367.042857    9528.538333
97           28471.266970    5828.873333
98           12875.423182     217.428333
99             379.123659       0.000000

[80 rows x 2 columns]
```

[694]:
```
temperatures = pd.read_csv("data-sets/temperatures.csv", index_col=0)
temperatures
```

[694]:
```
             date     city        country  avg_temp_c
0      2000-01-01  Abidjan  Côte D'Ivoire      27.293
1      2000-02-01  Abidjan  Côte D'Ivoire      27.685
2      2000-03-01  Abidjan  Côte D'Ivoire      29.061
3      2000-04-01  Abidjan  Côte D'Ivoire      28.162
4      2000-05-01  Abidjan  Côte D'Ivoire      27.547
…             …       …              …           …
16495  2013-05-01     Xian          China      18.979
16496  2013-06-01     Xian          China      23.522
16497  2013-07-01     Xian          China      25.251
16498  2013-08-01     Xian          China      24.528
16499  2013-09-01     Xian          China         NaN

[16500 rows x 4 columns]
```

[695]:
```
temperatures_ind = temperatures.set_index('city')
temperatures_ind
```

[695]:
```
               date        country  avg_temp_c
city
Abidjan  2000-01-01  Côte D'Ivoire      27.293
Abidjan  2000-02-01  Côte D'Ivoire      27.685
Abidjan  2000-03-01  Côte D'Ivoire      29.061
Abidjan  2000-04-01  Côte D'Ivoire      28.162
Abidjan  2000-05-01  Côte D'Ivoire      27.547
…               …              …           …
Xian     2013-05-01          China      18.979
Xian     2013-06-01          China      23.522
Xian     2013-07-01          China      25.251
Xian     2013-08-01          China      24.528
Xian     2013-09-01          China         NaN

[16500 rows x 3 columns]
```

[696]:
```
temperatures_ind.reset_index()
```

[696]:
```
           city        date        country  avg_temp_c
0       Abidjan  2000-01-01  Côte D'Ivoire      27.293
1       Abidjan  2000-02-01  Côte D'Ivoire      27.685
2       Abidjan  2000-03-01  Côte D'Ivoire      29.061
3       Abidjan  2000-04-01  Côte D'Ivoire      28.162
4       Abidjan  2000-05-01  Côte D'Ivoire      27.547
…           …           …              …           …
16495     Xian  2013-05-01          China      18.979
16496     Xian  2013-06-01          China      23.522
16497     Xian  2013-07-01          China      25.251
16498     Xian  2013-08-01          China      24.528
16499     Xian  2013-09-01          China         NaN

[16500 rows x 4 columns]
```

[697]:
```
temperatures_ind.reset_index(drop=True)
```

[697]:
```
             date        country  avg_temp_c
0      2000-01-01  Côte D'Ivoire      27.293
1      2000-02-01  Côte D'Ivoire      27.685
2      2000-03-01  Côte D'Ivoire      29.061
3      2000-04-01  Côte D'Ivoire      28.162
4      2000-05-01  Côte D'Ivoire      27.547
…             …              …           …
16495  2013-05-01          China      18.979
16496  2013-06-01          China      23.522
16497  2013-07-01          China      25.251
16498  2013-08-01          China      24.528
16499  2013-09-01          China         NaN

[16500 rows x 3 columns]
```

[698]:
```
cities = ["Moscow", "Saint Petersburg"]
print(temperatures[temperatures['city'].isin(cities)])
```

```
             date              city country  avg_temp_c
10725  2000-01-01            Moscow  Russia      -7.313
10726  2000-02-01            Moscow  Russia      -3.551
10727  2000-03-01            Moscow  Russia      -1.661
10728  2000-04-01            Moscow  Russia      10.096
10729  2000-05-01            Moscow  Russia      10.357
…             …               …       …           …
13360  2013-05-01  Saint Petersburg  Russia      12.355
13361  2013-06-01  Saint Petersburg  Russia      17.185
13362  2013-07-01  Saint Petersburg  Russia      17.234
13363  2013-08-01  Saint Petersburg  Russia      17.153
13364  2013-09-01  Saint Petersburg  Russia         NaN
```

```
[330 rows x 4 columns]
```

[699]:
```
print(temperatures_ind.loc[cities])
```

```
                        date country  avg_temp_c
city
Moscow             2000-01-01  Russia      -7.313
Moscow             2000-02-01  Russia      -3.551
Moscow             2000-03-01  Russia      -1.661
Moscow             2000-04-01  Russia      10.096
Moscow             2000-05-01  Russia      10.357
…                         …       …           …
Saint Petersburg   2013-05-01  Russia      12.355
Saint Petersburg   2013-06-01  Russia      17.185
Saint Petersburg   2013-07-01  Russia      17.234
Saint Petersburg   2013-08-01  Russia      17.153
Saint Petersburg   2013-09-01  Russia         NaN

[330 rows x 3 columns]
```

16 Ex 16

[700]:
```
temperatures_ind = temperatures.set_index(['country', 'city'])
temperatures_ind
```

[700]:
```
                               date  avg_temp_c
country       city
Côte D'Ivoire Abidjan  2000-01-01      27.293
              Abidjan  2000-02-01      27.685
              Abidjan  2000-03-01      29.061
              Abidjan  2000-04-01      28.162
              Abidjan  2000-05-01      27.547
…                               …           …
China         Xian     2013-05-01      18.979
              Xian     2013-06-01      23.522
              Xian     2013-07-01      25.251
              Xian     2013-08-01      24.528
              Xian     2013-09-01         NaN

[16500 rows x 2 columns]
```

[701]:
```
rows_to_keep = [('Brazil', 'Rio De Janeiro'), ('Pakistan', 'Lahore')]
print(temperatures_ind.loc[rows_to_keep])
```

```
                                 date  avg_temp_c
country  city
Brazil   Rio De Janeiro  2000-01-01      25.974
         Rio De Janeiro  2000-02-01      26.699
```

```
         Rio De Janeiro  2000-03-01      26.270
         Rio De Janeiro  2000-04-01      25.750
         Rio De Janeiro  2000-05-01      24.356
…                              …          …
Pakistan Lahore          2013-05-01      33.457
         Lahore          2013-06-01      34.456
         Lahore          2013-07-01      33.279
         Lahore          2013-08-01      31.511
         Lahore          2013-09-01         NaN

[330 rows x 2 columns]
```