Data Exploration 1-19

January 20, 2021

0.1 Exploring Zip Code Data for My Final Project

This Jupyter file contains some emploratory analysis of water debt data by zip code merged with Amercian Community Survey data. In this file, I will trim some of the data (as it is a large set) to Zip Code, the number of houses per zip that have 1000 or more water debt, median household income, and % of rened properties per zip. The data is divided into quartiles and the file concludes with a scatterplot.

Import pandas module

[68]:
```
import pandas as pd
```

Upload Zip Code Data

[69]:
```
zipdata = pd.read_excel('Data/ZipCode & Debt Data 1-19.xlsx')
```

How many rows and columns are in the data?

[70]:
```
zipdata.shape
```

[70]:
```
(1073, 52)
```

There are too many data, so we will filter it only to variables of interest: Zip Codes, number of households with more than $1000 in debt, median household income, and percent of the population that rents their home.

[71]:
```
relevant_columns = ['Zip Codes', 'Sum of More than $1000', 'mhhi', '% Renter␣
↪Pop']
```

Make a Copy

[67]:
```
zip_trim = zipdata[relevant_columns].copy()
zip_trim
```

[67]:

	Zip Codes	Sum of More than $1000	mhhi	% Renter Pop
0	90001	709.0	38521.0	0.647257
1	90002	1779.0	35410.0	0.621999
2	90003	3437.0	37226.0	0.692044
3	90004	1116.0	48754.0	0.803900
4	90005	426.0	35149.0	0.912764

```
…              …                        …       …           …
1068     90033-2053                     0.0      NaN         NaN
1069   92780, 92705                    18.0      NaN         NaN
1070  95608 & 95628                     6.0      NaN         NaN
1071        (blank)                     5.0      NaN         NaN
1072    Grand Total                155118.8      NaN         NaN

[1073 rows x 4 columns]
```

Determine the types of data in the new set.

[18]:
```
zip_trim.info
```

[18]:
```
<bound method DataFrame.info of           Zip Codes  Sum of More than $1000
mhhi  % Renter Pop
0              90001                   709.0  38521.0      0.647257
1              90002                  1779.0  35410.0      0.621999
2              90003                  3437.0  37226.0      0.692044
3              90004                  1116.0  48754.0      0.803900
4              90005                   426.0  35149.0      0.912764
…              …                       …       …           …
1068     90033-2053                     0.0      NaN         NaN
1069   92780, 92705                    18.0      NaN         NaN
1070  95608 & 95628                     6.0      NaN         NaN
1071        (blank)                     5.0      NaN         NaN
1072    Grand Total                155118.8      NaN         NaN

[1073 rows x 4 columns]>
```

Show first 5 lines of the new dataframe

[37]:
```
zip_trim.head()
```

[37]:
```
  Zip Codes  Sum of More than $1000     mhhi  % Renter Pop
0     90001                   709.0  38521.0      0.647257
1     90002                  1779.0  35410.0      0.621999
2     90003                  3437.0  37226.0      0.692044
3     90004                  1116.0  48754.0      0.803900
4     90005                   426.0  35149.0      0.912764
```

I wanted to see the quantiles for the data to get a better sense of the spread. The spread of ”Sum of More than $1,000 is extremely high, showing great variation in that variable. Median household income seems normaly scaled, and generally most people seem to own as opposed to rent in the respective zip codes.

[50]:
```
zip_trim.quantile([0.25,0.5,0.75])
```

[50]:

	Sum of More than $1000	mhhi	% Renter Pop
0.25	3.0	50060.5	0.281793
0.50	17.0	69049.0	0.390885
0.75	73.0	94093.0	0.528756

Run a value count, breaking up the Median Household Income into 12 bins. This shows that most folks in the sample were middle class households, making 42,000 - 62,500.

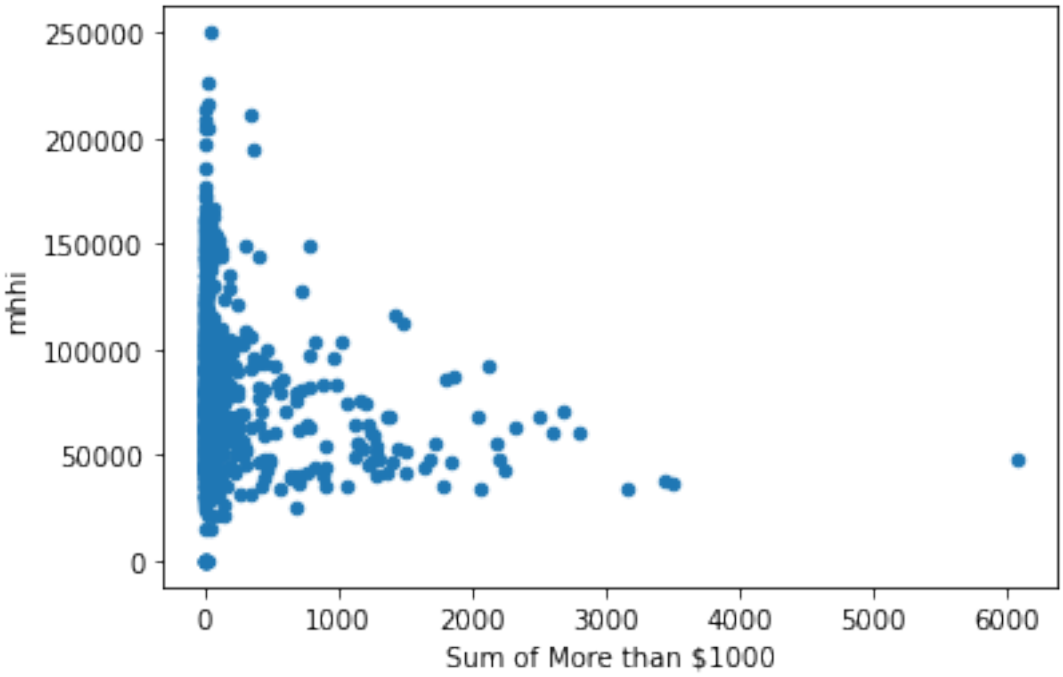
[51]:
```
zip_trim['mhhi'].value_counts(bins=12)
```

[51]:
```
(41666.833, 62500.25]       298
(62500.25, 83333.667]       244
(83333.667, 104167.083]     179
(20833.417, 41666.833]      112
(104167.083, 125000.5]       82
(125000.5, 145833.917]       43
(145833.917, 166667.333]     40
(-250.002, 20833.417]        18
(208334.167, 229167.583]      5
(187500.75, 208334.167]       4
(166667.333, 187500.75]       4
(229167.583, 250001.0]        1
Name: mhhi, dtype: int64
```

Finally, we'll make a scatterplot of Sum of More than 1000 (which is the frequency of households per zip code that have more than 1 thousand in water will debt) by Median Houssehold Income. The data might show that as income decreases, the sum of areas with more than 1000 in debt increases, but more research must be done to determine this.

[62]:
```
zip_trim.plot.scatter(x='Sum of More than $1000', y='mhhi')
```

[62]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fb6bae08a60>
```


250000
200000
150000
100000
50000
0
mhhi
0
1000
2000
3000
4000
5000
6000
Sum of More than $1000