UDACITY'S Predictive Analytics for Business NANODEGREE
Project 2.1: Data Cleanup

Daiane Ucceli Kreitlow

09 September, 2023

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions need to be made?

   Recommend the city for Pawdacity's newest store, based on predicted yearly sales.

2. What data is needed to inform those decisions?

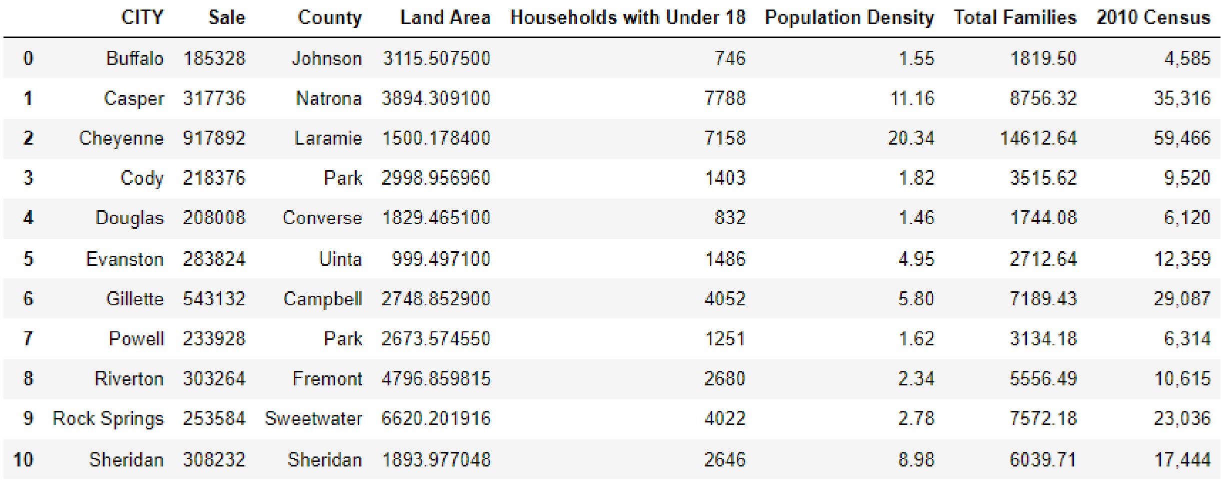
   First step in predicting yearly sales is to format and blend together data from different datasets and deal with outliers.

   It's necessary to build a training dataset with columns:

   - City
   - 2010 Census Population
   - Total Pawdacity Sales
   - Households Under 18
   - Land Area
   - Population Density
   - Total Families

   The data provided has information at store level, but we only need data at city level.

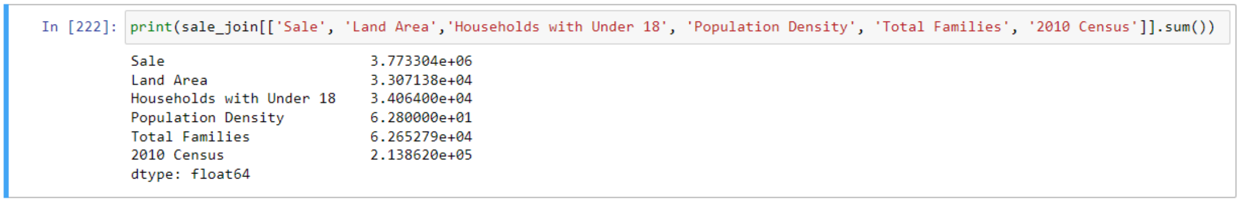
After clean data, my dataset is:



	CITY	Sale	County	Land Area	Households with Under 18	Population Density	Total Families	2010 Census
0	Buffalo	185328	Johnson	3115.507500	746	1.55	1819.50	4,585
1	Casper	317736	Natrona	3894.309100	7788	11.16	8756.32	35,316
2	Cheyenne	917892	Laramie	1500.178400	7158	20.34	14612.64	59,466
3	Cody	218376	Park	2998.956960	1403	1.82	3515.62	9,520
4	Douglas	208008	Converse	1829.465100	832	1.46	1744.08	6,120
5	Evanston	283824	Uinta	999.497100	1486	4.95	2712.64	12,359
6	Gillette	543132	Campbell	2748.852900	4052	5.80	7189.43	29,087
7	Powell	233928	Park	2673.574550	1251	1.62	3134.18	6,314
8	Riverton	303264	Fremont	4796.859815	2680	2.34	5556.49	10,615
9	Rock Springs	253584	Sweetwater	6620.201916	4022	2.78	7572.18	23,036
10	Sheridan	308232	Sheridan	1893.977048	2646	8.98	6039.71	17,444

where Sale is **Total Pawdacity Sales.**

Performing the sum of numerical variables, I had the answers:



```
In [222]: print(sale_join[['Sale', 'Land Area','Households with Under 18', 'Population Density', 'Total Families', '2010 Census']].sum())

          Sale                        3.773304e+06
          Land Area                   3.307138e+04
          Households with Under 18    3.406400e+04
          Population Density          6.280000e+01
          Total Families              6.265279e+04
          2010 Census                 2.138620e+05
          dtype: float64
```

Step 2: Building the Training Set

Column	**Sum**	**Average**
Census Population	*213,862*	*19,442.0*
Total Pawdacity Sales	*3,773,304*	*343,027.6*
Households Under 18	*34,064*	*3,096.7*
Land Area	*33,071*	*3,006.5*
Population Density	*63*	*5.7*
Total Families	*62,653*	*5,695.7*

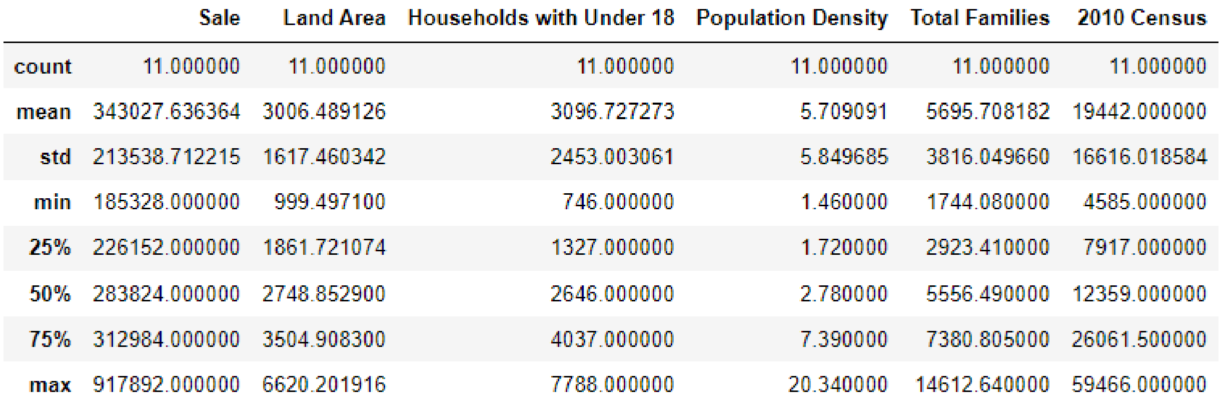
Step 3: Dealing with Outliers

Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.

Using the IQR methods to determine if there are outlier cities for each of the variables.

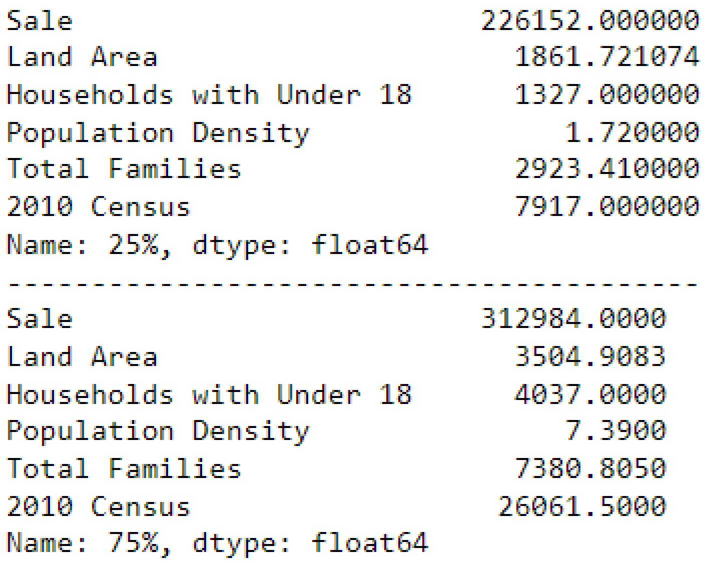
So:

Describing values:



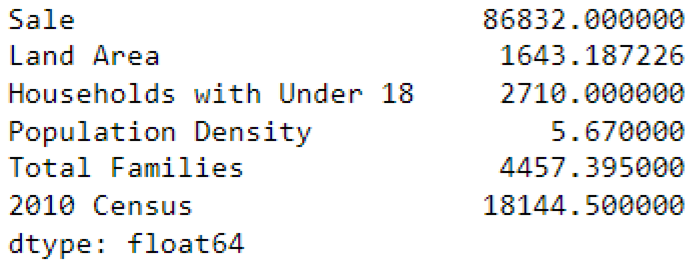
	Sale	Land Area	Households with Under 18	Population Density	Total Families	2010 Census
count	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000
mean	343027.636364	3006.489126	3096.727273	5.709091	5695.708182	19442.000000
std	213538.712215	1617.460342	2453.003061	5.849685	3816.049660	16616.018584
min	185328.000000	999.497100	746.000000	1.460000	1744.080000	4585.000000
25%	226152.000000	1861.721074	1327.000000	1.720000	2923.410000	7917.000000
50%	283824.000000	2748.852900	2646.000000	2.780000	5556.490000	12359.000000
75%	312984.000000	3504.908300	4037.000000	7.390000	7380.805000	26061.500000
max	917892.000000	6620.201916	7788.000000	20.340000	14612.640000	59466.000000

1. **Q1 e Q3:**



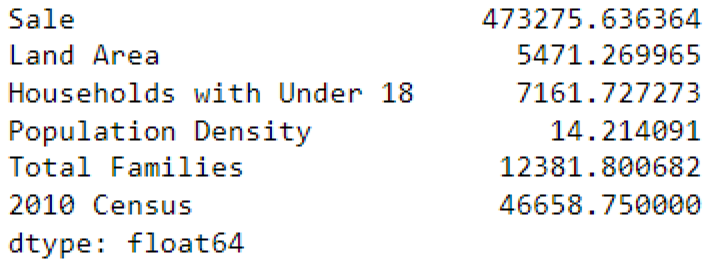
```
Sale                          226152.000000
Land Area                       1861.721074
Households with Under 18        1327.000000
Population Density                 1.720000
Total Families                  2923.410000
2010 Census                     7917.000000
Name: 25%, dtype: float64
-----------------------------------------
Sale                          312984.0000
Land Area                       3504.9083
Households with Under 18        4037.0000
Population Density                 7.3900
Total Families                  7380.8050
2010 Census                    26061.5000
Name: 75%, dtype: float64
```

2. **IQR = Q3 - Q1:**



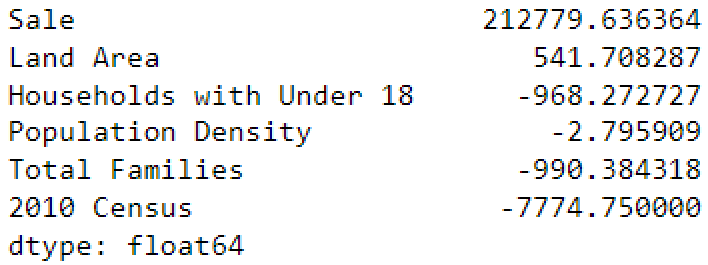
```
Sale                          86832.000000
Land Area                      1643.187226
Households with Under 18       2710.000000
Population Density                5.670000
Total Families                 4457.395000
2010 Census                   18144.500000
dtype: float64
```

3. **Upper Fence = Q3 + 1.5 * IQR:**



```
Sale                         473275.636364
Land Area                      5471.269965
Households with Under 18       7161.727273
Population Density               14.214091
Total Families                12381.800682
2010 Census                   46658.750000
dtype: float64
```

4. **Lower Fence = Q1 - 1.5 * IQR:**

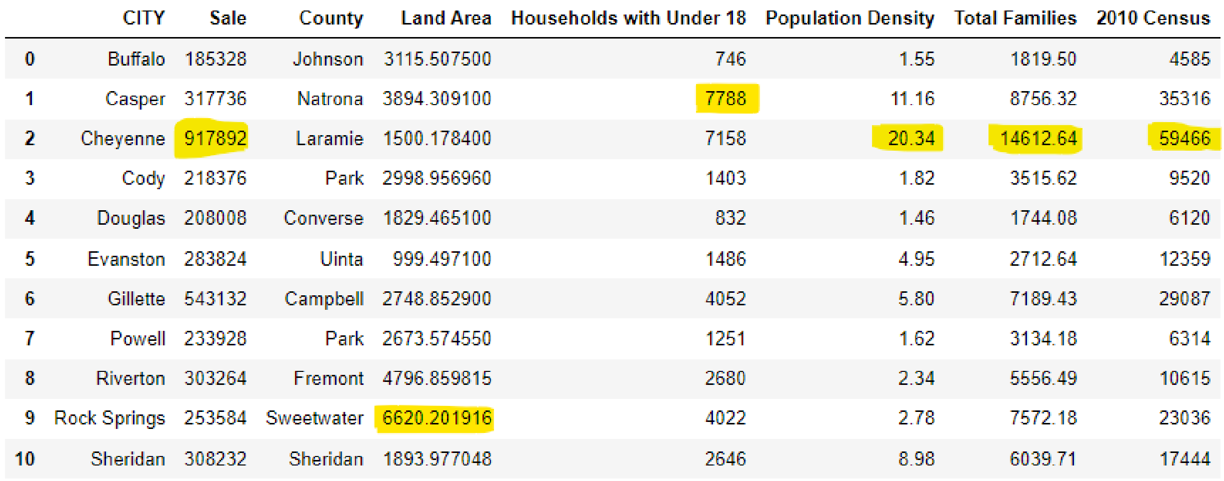


```
Sale                         212779.636364
Land Area                       541.708287
Households with Under 18       -968.272727
Population Density               -2.795909
Total Families                 -990.384318
2010 Census                   -7774.750000
dtype: float64
```

Values above the Upper Fence and values below the Lower Fence are outliers:

Variables	**Too High**	**Too Low**
Census Population	*46,658.75*	*-7,774.75*
Total Pawdacity Sales	*473,275.63*	*212779.63*
Households Under 18	*7,161.72*	*-968.27*
Land Area	*5,471.27*	*541.71*
Population Density	*14.21*	*-2.79*
Total Families	*12,381.80*	*-990.38*

Observing the High Values of the Upper Fence for each variable, Cheyenne is too high in 4 variables.



	CITY	Sale	County	Land Area	Households with Under 18	Population Density	Total Families	2010 Census
0	Buffalo	185328	Johnson	3115.507500	746	1.55	1819.50	4585
1	Casper	317736	Natrona	3894.309100	7788	11.16	8756.32	35316
2	Cheyenne	917892	Laramie	1500.178400	7158	20.34	14612.64	59466
3	Cody	218376	Park	2998.956960	1403	1.82	3515.62	9520
4	Douglas	208008	Converse	1829.465100	832	1.46	1744.08	6120
5	Evanston	283824	Uinta	999.497100	1486	4.95	2712.64	12359
6	Gillette	543132	Campbell	2748.852900	4052	5.80	7189.43	29087
7	Powell	233928	Park	2673.574550	1251	1.62	3134.18	6314
8	Riverton	303264	Fremont	4796.859815	2680	2.34	5556.49	10615
9	Rock Springs	253584	Sweetwater	6620.201916	4022	2.78	7572.18	23036
10	Sheridan	308232	Sheridan	1893.977048	2646	8.98	6039.71	17444

At least data from Cheyenne’s City should be removed from the dataset for the reasons above.