Project 2.1: Data Cleanup

Make a copy of this document. Complete each section. When you are ready, save your file as a PDF document and submit it here:
https://classroom.udacity.com/nanodegrees/nd008/parts/8d60a887-d4c1-4b0e-8873-b2f36435eb39/project

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (250 word limit)

Key Decisions:

Answer these questions

1. What decisions needs to be made?

We want to help Pawdacity's in deciding which city the new store should be opened, based on predicted yearly sales.

How Do I Complete this Project?

2. What data is needed to inform those decisions?
   After blending and cleaning the separated data files, I prepared the following fields that is needed for analysis purpose:
   - Census Population
   - Total Pawdacity Sales
   - Households with Under 18
   - Land Area
   - Population Density
   - Total Families

Step 2: Building the Training Set

Build your training set given the data provided to you. Your column sums of your dataset should match the sums in the table below.

In addition provide the averages on your data set here to help reviewers check your work. You should round up to two decimal places, ex: 1.24

Column	Sum	Average
Census Population	*213,862*	*19442*
Total Pawdacity Sales	*3,773,304*	*343027.64*
Households with Under 18	*34,064*	*3096.73*
Land Area	*33,071*	*3006.49*
Population Density	*63*	*5.71*
Total Families	*62,653*	*5695.71*

Step 3: Dealing with Outliers

Answer these questions

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.

Yes there are tow cities outlier, Cheyenne and Gillette .
I used IQR for identifying outliers, as shown below :

G24

	A	B	C	D	E	F	G	H
1	**CITY**	**2010 Census**	**Total_Pawdacity_Sales**	**Households with Under 18**	**Land Area**	**Population Density**	**Total Families**	
2	Buffalo	4585	185328	746	3115.5075	1.55	1819.5	
3	Casper	35316	317736	7788	3894.3091	11.16	8756.32	
4	Cheyenne	59466	917892	7158	1500.1784	20.34	14612.64	
5	Cody	9520	218376	1403	2998.95696	1.82	3515.62	
6	Douglas	6120	208008	832	1829.4651	1.46	1744.08	
7	Evanston	12359	283824	1486	999.4971	4.95	2712.64	
8	Gillette	29087	543132	4052	2748.8529	5.8	7189.43	
9	Powell	6314	233928	1251	2673.57455	1.62	3134.18	
10	Riverton	10615	303264	2680	4796.859815	2.34	5556.49	
11	Rock Springs	23036	253584	4022	6620.201916	2.78	7572.18	
12	Sheridan	17444	308232	2646	1893.977048	8.98	6039.71	
13								
14								
15	**Q3**	29087	317736	4052	3894.3091	8.98	7572.18	
16	**Q1**	6314	218376	1251	1829.4651	1.62	2712.64	
17	**IQR**	22773	99360	2801	2064.844	7.36	4859.54	
18	**upper fence**	63246.5	466776	8253.5	6991.5751	20.02	14861.49	
19	**lower fence**	-27845.5	69336	-2950.5	-1267.8009	-9.42	-4576.67	
20								
21								
22								

For Cheyenne city I keep this outlier since the other variables (Population Density, And Total Families) are high, so it makes sense to lead for high sales.
For Gillette city, I removed this city because the other variables in average.