Project 2.1: Data Cleanup

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?
   The decision needs to be made is, “In which city of the Wyoming state the Pawdacity’s newest store should be opened?”

2. What data is needed to inform those decisions?
   To inform those decisions, we need to calculate the average of the total Pawdacity Sales of each city and make a prediction on it in order to know about the city to recommands.

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

Based on the 11 cities given in the **The monthly sales data for all of the Pawdacity stores for the year 2010,** in order to build the new dataset, here are the result found.

Column	Sum	Average
Census Population	*213,862*	19,442
Total Pawdacity Sales	*3,773,304*	*343,027.64*
Households with Under 18	*34,064*	3,097
Land Area	*33,071*	3,006
Population Density	*63*	6
Total Families	*62,653*	5,696

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.

City	2010 Census Population	Total Pawdacity Sales	Land Area	Households with Under 18	Population Density	Total Families		**1st quartile Q1**	**3rd quartile Q3**	**Interquartile Range**	**upper fence**	**lower fence**
Buffalo	4,585	185328	3,115.51	746.00	1.55	1819.5	195,596	**1014.375**	**49770.75**	**48756.38**	**122,905.31**	**-72120.19**
Casper	35,316	317736	3,894.31	7,788.00	11.16	8756.32	373,502	**4867.731825**	**105921**	**101053.3**	**257,500.90**	**-146712.17**
Cheyenne	59,466	917892	1,500.18	7,158.00	20.34	14612.64	1,000,649	**2914.6338**	**274072.5**	**271157.9**	**680,809.30**	**-403822.17**
Cody	9,520	218376	2,998.96	1,403.00	1.82	3515.62	235,815	**1801.98924**	**61734**	**59932.01**	**151,632.02**	**-88096.03**
Douglas	6,120	208008	1,829.47	832.00	1.46	1744.08	218,535	**1060.02**	**56592**	**55531.98**	**139,889.97**	**-82237.95**
Evanston	12,359	283824	999.50	1,486.00	4.95	2712.64	301,386	**1121.122825**	**80225.25**	**79104.13**	**198,881.44**	**-117535.07**
Gillette	29,087	543132	2,748.85	4,052.00	5.8	7189.43	586,215	**3074.639675**	**157598.25**	**154523.6**	**389,383.67**	**-228710.78**
Powell	6,314	233928	2,673.57	1,251.00	1.62	3134.18	247,302	**1606.643638**	**63217.5**	**61610.86**	**155,633.78**	**-90809.64**
Riverton	10,615	303264	4,796.86	2,680.00	2.34	5556.49	326,915	**3209.214954**	**83777.25**	**80568.04**	**204,629.30**	**-117642.84**
Rock Springs	23,036	253584	6,620.20	4,022.00	2.78	7572.18	294,837	**4671.550479**	**80673**	**76001.45**	**194,675.17**	**-109330.62**
Sheridan	17,444	308232	1,893.98	2,646.00	8.98	6039.71	336,265	**2081.982786**	**90141**	**88059.02**	**222,229.53**	**-130006.54**
SUM	**213,862.00**	**3,773,304.00**	**33,071.38**	**34,064.00**	**62.80**	**62,652.79**						
AVERAGE	**19,442**	**343,028**	**3,006**	**3,097**	**6**	**5,696**						

According to the IQR method values above the Upper Fence and values below the Lower Fence are outliers. Thus, in our dataset except the Buffalo city, all others cities appeared outliers. However, as our data is small, we are going to remove only the city which appeared most to be outlier and it is **Cheyenne.**