Project 2.1: Data Cleanup

https://classroom.udacity.com/nanodegrees/nd008/parts/8d60a887-d4c1-4b0e-8873-b2f36435eb39/project

Step 1: Business and Data Understanding

Key Decisions:

1. **What decisions needs to be made?**
   Pawdacity needs to decide the best location to open a new store in the state of Wyoming. This decision will be based primarily on the predicted yearly sales of each city.

2. **What data is needed to inform those decisions?**
   To estimate future sales of each city we will use historical performance of the current stores as well as geographic and demographic data of each city (2010 Census Population, Number of Households with Under 18, Land Area, Population Density and Total Families.

Step 2: Building the Training Set

Column	Sum	Average
Census Population	*213,862*	*19,442*
Total Pawdacity Sales	*3,773,304*	*343,027.64*
Households with Under 18	*34,064*	*3,096.73*
Land Area	*33,071*	*3,006.45*
Population Density	*63*	*5.73*
Total Families	*62,653*	*5,695.73*

Step 3: Dealing with Outliers

Three cities presented values that can be considered outliers by the IQR method: Cheyenne, Gillette and Rock Springs. But while Gillette and Rock Springs presented outliers for only one metric each, and still relatively close to the respective upper fences, Cheyenne had four out of the six metrics as outliers, and all of them with a big difference in comparison to the upper fence.
Since it could have a negative effect on our model, we should remove the entry relative from Cheyenne from our input.

Appendix:

CITY	Census Population	Total Pawdacity Sales	Households with Under 18	Land Area	Population Density	Total Families
Buffalo	4.585	185.328	746	3.116	2	1.820
Casper	35.316	317.736	7.788	3.894	11	8.756
Cheyenne	59.466	917.892	7.158	1.500	20	14.613
Cody	9.520	218.376	1.403	2.999	2	3.516
Douglas	6.120	208.008	832	1.829	1	1.744
Evanston	12.359	283.824	1.486	999	5	2.713
Gillette	29.087	543.132	4.052	2.749	6	7.189
Powell	6.314	233.928	1.251	2.674	2	3.134
Riverton	10.615	303.264	2.680	4.797	2	5.556
Rock Springs	23.036	253.584	4.022	6.620	3	7.572
Sheridan	17.444	308.232	2.646	1.894	9	6.040
Sum	213.862	3.773.304	34.064	33.071	63	62.653
Avg	19.442,00	343.027,64	3.096,73	3.006,45	5,73	5.695,73
1st Quartile	7.917	226.152	1.327	1.862	2	2.924
3rd Quartile	26.062	312.984	4.037	3.505	8	7.381
IQR Range	18.145	86.832	2.710	1.644	6	4.457
Lower Fence	-19.300	95.904	-2.738	-604	-6	-3.762
Upper Fence	53.278	443.232	8.102	5.970	16	14.066

Outliers

* many of the Lower Fence calculations resulted in values smaller than zero, which are not possible, so the analysis was made only considering the Upper fence in these cases.