Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?
   A decision needs to be made about which city in Wyoming should Pawdacity should expand and open a 14th store base on predicted yearly sales.
2. What data is needed to inform those decisions?
   We need data regarding yearly sales and demographics (i.e. census population, number of households with under 18, land area, population density, and total number of family) for each city in order to predict yearly sales for each city.

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.

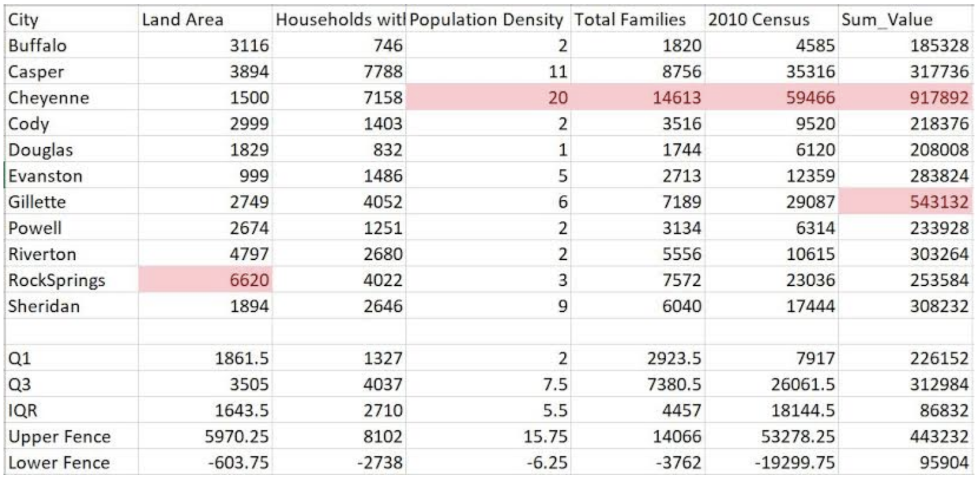
In addition provide the averages on your data set here to help reviewers check your work. You should round up to two decimal places, ex: 1.24

Column	**Sum**	**Average**
Census Population	*213,862*	*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

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 method for each attribute (Census Population, Total Pawdacity Sales, Households with Under 18, Land Area, Population Density, Total Families) of the cities, I have calculated the upper fence and lower fence and determined that the values above the Upper Fence and values below the Lower Fence are outliers.



City	Land Area	Households witl	Population Density	Total Families	2010 Census	Sum_Value
Buffalo	3116	746	2	1820	4585	185328
Casper	3894	7788	11	8756	35316	317736
Cheyenne	1500	7158	20	14613	59466	917892
Cody	2999	1403	2	3516	9520	218376
Douglas	1829	832	1	1744	6120	208008
Evanston	999	1486	5	2713	12359	283824
Gillette	2749	4052	6	7189	29087	543132
Powell	2674	1251	2	3134	6314	233928
Riverton	4797	2680	2	5556	10615	303264
RockSprings	6620	4022	3	7572	23036	253584
Sheridan	1894	2646	9	6040	17444	308232
Q1	1861.5	1327	2	2923.5	7917	226152
Q3	3505	4037	7.5	7380.5	26061.5	312984
IQR	1643.5	2710	5.5	4457	18144.5	86832
Upper Fence	5970.25	8102	15.75	14066	53278.25	443232
Lower Fence	-603.75	-2738	-6.25	-3762	-19299.75	95904

In this case, three cities Cheyenne, Gillette, and Rock Springs have outliers in the training set.

From an examination of the fence points and the data, we could see that Cheyenne stands out as the most number of outliers. Predictor variables, Population Density, Total families, 2010 Census and Total Pawdacity Sales exceeds the upper fence. Even though Cheyenne compares to have a small Land Area among other cities, the high explanatory variables explains the high Total Pawdaicty Sales. Since Cheyenne still holds the linear relationship, we do not consider Cheyenne as a outlier city.

On the other hand, for Gillette, all of the predictor variables, except for Total Pawdacity Sales, are within the upper and lower fence. Even though the demographic data is all within the range, the Total Pawdacity Sales exceeds the upper fence which stands out as an outlier value that should be removed.

Rock Springs also has a variable, Land Area, that exceeds the upper fence yet, all of the other predictor variables, including Population Density, are within the upper and lower fence. Since Population Density is heavily relies on the data of 2010 Census and Land Area and is within the range, we do not consider Rock Springs as an outlier city.

All in all, I have chosen Gillette as the outlier city in the training set to remove.