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?

Answer: A recommendation must be made regarding the location of a new store using historical sales data.

2. What data is needed to inform those decisions?

Answer: To perform this analysis I found it useful to put the focus on the following columns (2010 census Population, Households with under 18, Land Area, Population Density and Total Families). These variables act like predictor variables.

This data can be find in the files “p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales”, “p2-partially-parsed-wy-web-scrape” and “p2-wy-demographic-data”

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*	19,442.00
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*	*5.71*
Total Families	*62,653*	5,695.72

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.

Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.

Answer: I indicated two outlier, but one stands out for me, which is the city "Cheyenne" for many reasons. If you follow the IQR Steps and compare it with the City "Cheyenne" you can clearly see that in four categories the "Cheyenne" is way over the upper fence.

City	County	Land Area	Households with Under 18	Population Density	Total Families	Total Sale	2010 census Population
Cheyenne	Laramie	1.500	7158	20,34	14612,64	917.892	59466

	Land Area	Households with Under 18	Population Density	Total Families	Total Sale	2010 census Population
Q1	1.862	1.327	2	2.923	226.152	7.917
Q3	3.505	4.037	7	7.381	312.984	26.062
IQR	1.643	2.710	6	4.457	86.832	18.145
Upper fence	5.970	8.102	16	14.067	443.232	53.278
Lower fence	-603	-2.738	-7	-3.763	95.904	-19.300
Description	Fine	Fine	Over the upper fence	Over the upper fence	Over the upper fence	Over the upper fence

Besides “Cheyenne” there is also the City “Gillette”, which has conspicuity regarding the category total sale.

City	County	Land Area	Households with Under 18	Population Density	Total Families	Total Sale	2010 census Population
Gillette	Campbell	2.749	4052	5,80	7189,43	543.132	29087

Q1	1.862	1.327	2	2.923	226.152	7.917
Q3	3.505	4.037	7	7.381	312.984	26.062
IQR	1.643	2.710	6	4.457	86.832	18.145
Upper fence	5.970	8.102	16	14.067	443.232	53.278
Lower fence	-603	-2.738	-7	-3.763	95.904	-19.300
Description	Fine	Fine	Fine	Fine	Over the upper fence	Fine

My suggestion would be to remove the city “Cheyenne” since there are too many outliers connected to this city. In contrast, I would not remove "Gillette" because the city has only one outlier, which, for one, is not as large as in the case of "Cheyenne" and, in addition, I would not remove "Gillette" because the total number, originally 11 cities, is relatively small and any reduction that is not too justifiable has a large impact on the overall result.