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?
   The decisions to be made are:
   - whether or not to open a new store (14th), and
   - where to open the new store if the recommendation is to open

2. What data is needed to inform those decisions?
   The data required to inform those decisions include data on:
   - Locations of current stores
   - Populations of the cities
   - Population of Pawdacity's target demographic
   - Previous sales from all the stores

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

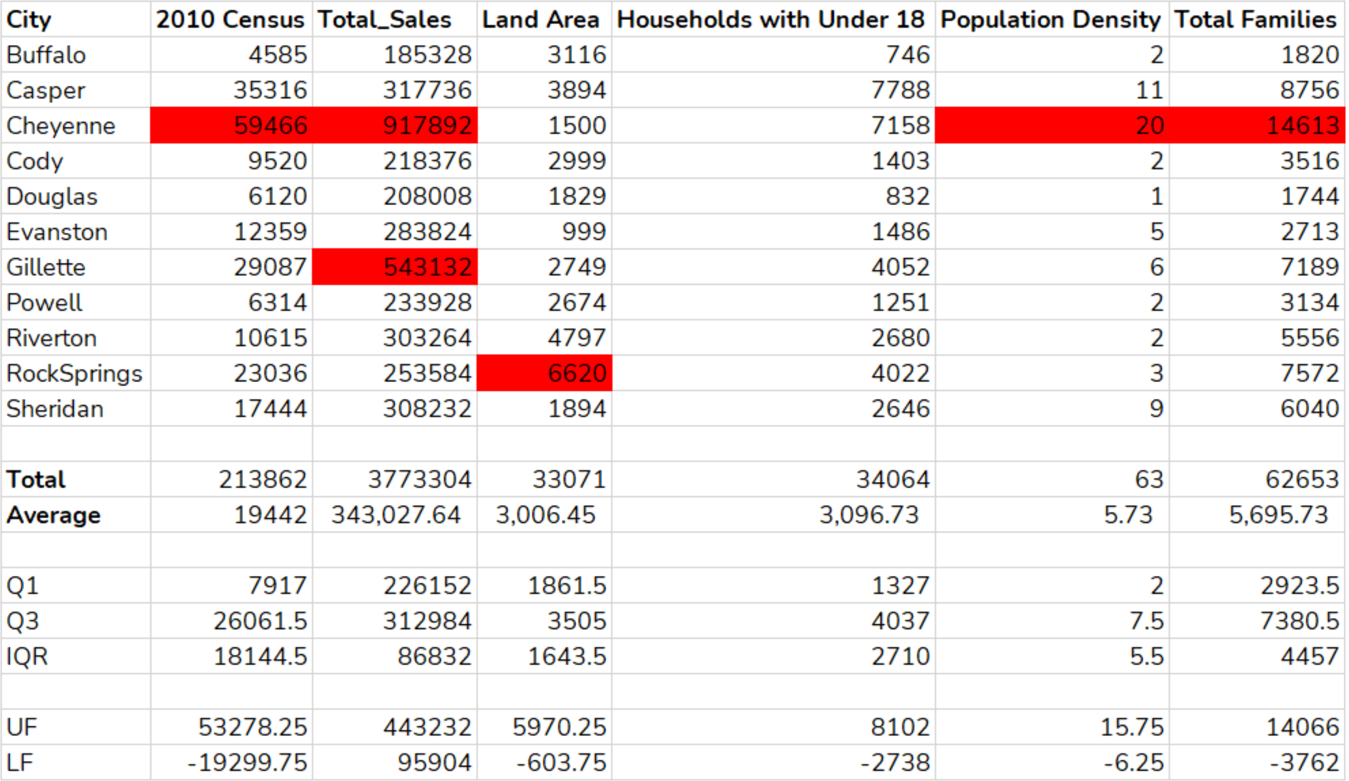
Answer these questions

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 outliers in the training dataset. Cheyenne has outliers for all its variables except land area and households with under 18. Total sales has a second outlier, from Gillette (543,132). RockSprings is an outlier under land area (6,620).

Of all three, due to the frequency and magnitude of its outliers, I would recommend removal of Cheyenne as it is likely to distort any analysis and models built on the data that includes it. Removing any more values will affect the analysis as the remaining data may not be adequate to support decisions. An example is total sales, where the removal of both outliers will leave us with little data for further analysis and/or prediction.

The above are summarized in the image below, with the red values representing outliers:



City	2010 Census	Total_Sales	Land Area	Households with Under 18	Population Density	Total Families
Buffalo	4585	185328	3116	746	2	1820
Casper	35316	317736	3894	7788	11	8756
Cheyenne	59466	917892	1500	7158	20	14613
Cody	9520	218376	2999	1403	2	3516
Douglas	6120	208008	1829	832	1	1744
Evanston	12359	283824	999	1486	5	2713
Gillette	29087	543132	2749	4052	6	7189
Powell	6314	233928	2674	1251	2	3134
Riverton	10615	303264	4797	2680	2	5556
RockSprings	23036	253584	6620	4022	3	7572
Sheridan	17444	308232	1894	2646	9	6040
Total	213862	3773304	33071	34064	63	62653
Average	19442	343,027.64	3,006.45	3,096.73	5.73	5,695.73
Q1	7917	226152	1861.5	1327	2	2923.5
Q3	26061.5	312984	3505	4037	7.5	7380.5
IQR	18144.5	86832	1643.5	2710	5.5	4457
UF	53278.25	443232	5970.25	8102	15.75	14066
LF	-19299.75	95904	-603.75	-2738	-6.25	-3762

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.

Kwame Odoi Otchere