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 business decision here is recommending a city in Wyoming State for opening a new store for Pawdacity. This would be based on predicted yearly sales.

To predict yearly sales, we would need to build a regression model basis data of existing cities. We would also need demographic data on the target cities where we want to predict sales and hence arrive at a decision. The data would include variables like:

1. City Name
2. Annual Sales
3. 2010 Census Population
4. Land Area
5. Households with under 18
6. Population Density
7. Total Families

2. **What data is needed to inform those decisions?**

We would require present sales and demographic data of the cities where Pawdacity has operations.

We would also need the demographic data of the cities which are under consideration for expansion.

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*	*19442*
Total Pawdacity Sales	*3,773,304*	*343027.64*
Households with Under 18	*34,064*	*3096.73*
Land Area	*33,071*	*3006.49*
Population Density	*63*	*5.70*
Total Families	*62,653*	*5695.71*

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.

I exported the output after cleaning and joining the relevant data into an excel file. The cells highlighted in Red are above the upper fence and are hence outliers.

Cheyenne seems to be a very high population density city and Rock Springs seems to be a very low population density area. In Cheyenne, multiple parameters are outliers. In Rock Springs, only the Land Area is an outlier.

In view of the above, I suggest we remove Cheyenne as an outlier city.

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

The Upper and lower fence calculations are as below:

	Census Population	Total Pawdacity Sales	Households with Under 18	Land Area	Population Density	Total Families
1st Quartile	7,917	226,152	1,327	1,862	2	2,923
3rd Quartile	26,062	312,984	4,037	3,505	7	7,381
IOQ	18,145	86,832	2,710	1,643	6	4,457
Upper Fence	53,278	443,232	8,102	5,970	16	14,067
Lower Fence	(19,300)	95,904	(2,738)	(603)	(7)	(3,763)

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.