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 decision that needs to be made here is whether Pawdacity, a pet store chain in Wyoming that has 13 stores in the state should open a 14th score based on predicted annual sales for this store.

2. What data is needed to inform those decisions?

   The data needed to formulate the decision is the monthly sales from all existing Pawdacity stores starting 2010, the NAICS data on the current sales from competitor stores where the total sales is equivalent to 12 months' worth of sales, information regarding the population and demographics of the area served by 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

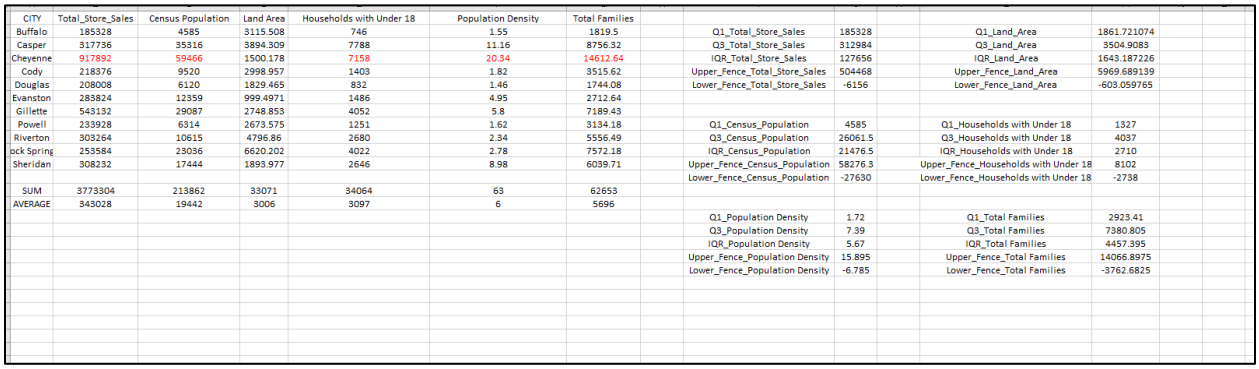
Column	Sum	Average
Census Population	*213,862*	*19442*
Total Pawdacity Sales	*3,773,304*	*343028*
Households with Under 18	*34,064*	*3097*
Land Area	*33,071*	*3006*
Population Density	*63*	*6*
Total Families	*62,653*	*5696*

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 is an outlier that I have chosen to remove. This outlier includes records pertaining to the city of Cheyenne. I came to this conclusion by performing the IQR upper fence and lower fence analysis as given in the project details. The image below displays a screenshot of my excel analysis. The only column value related to Cheyenne that is not an outlier is the land area.



CITY	Total_Store_Sales	Census Population	Land Area	Households with Under 18	Population Density	Total Families
Buffalo	185328	4585	3115.508	746	1.55	1819.5
Casper	317736	35316	3894.309	7788	11.16	8756.32
Cheyenne	917892	59466	1500.178	7158	20.34	14612.64
Cody	218376	9520	2998.957	1403	1.82	3515.62
Douglas	208008	6120	1829.465	832	1.46	1744.08
Evanston	283824	12359	999.4971	1486	4.95	2712.64
Gillette	543132	29087	2748.853	4052	5.8	7189.43
Powell	233928	6314	2673.575	1251	1.62	3134.18
Riverton	303264	10615	4796.86	2680	2.34	5556.49
ock Spring	253584	23036	6620.202	4022	2.78	7572.18
Sheridan	308232	17444	1893.977	2646	8.98	6039.71
SUM	3773304	213862	33071	34064	63	62653
AVERAGE	343028	19442	3006	3097	6	5696

Q1_Total_Store_Sales	185328	Q1_Land_Area	1861.721074
Q3_Total_Store_Sales	312984	Q3_Land_Area	3504.9083
IQR_Total_Store_Sales	127656	IQR_Land_Area	1643.187226
Upper_Fence_Total_Store_Sales	504468	Upper_Fence_Land_Area	5969.689139
Lower_Fence_Total_Store_Sales	-6156	Lower_Fence_Land_Area	-603.059765
Q1_Census_Population	4585	Q1_Households with Under 18	1327
Q3_Census_Population	26061.5	Q3_Households with Under 18	4037
IQR_Census_Population	21476.5	IQR_Households with Under 18	2710
Upper_Fence_Census_Population	58276.3	Upper_Fence_Households with Under 18	8102
Lower_Fence_Census_Population	-27630	Lower_Fence_Households with Under 18	-2738
Q1_Population Density	1.72	Q1_Total Families	2923.41
Q3_Population Density	7.39	Q3_Total Families	7380.805
IQR_Population Density	5.67	IQR_Total Families	4457.395
Upper_Fence_Population Density	15.895	Upper_Fence_Total Families	14066.8975
Lower_Fence_Population Density	-6.785	Lower_Fence_Total Families	-3762.6825

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.