Project 2.1: Data Cleanup

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?

Ans: Pawdacity is a leading pet store chain in Wyoming with 13 stores throughout the state. Based on the predicted yearly sales, whether opening of 14th store will be profitable or not, has to be predicted.

2. What data is needed to inform those decisions?

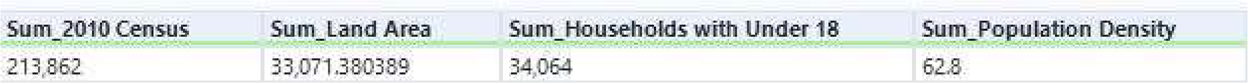
Ans: We need the data of the monthly sales of all the Pawdacity stores, and the current sales of all competitor stores.
Also, we need the population records of each city. Plus the Demographic data like Land Area, Population Density, Households with individuals under 18 and Total Families for each city in the state of Wyoming.

Step 2: Building the Training Set

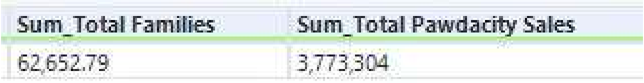
Build your training set given the data provided to you. Your column sums of your dataset should match the sums in the table below.

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

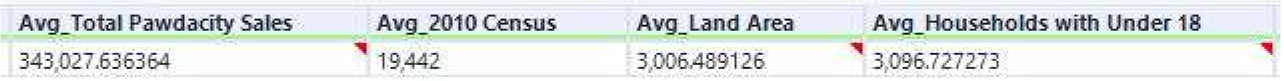
Column	**Sum**	**Average**
Census Population	*213,862*	*19,442*
Total Pawdacity Sales	*3,773,304*	*343,027.64*
Households with Under 18	*34,064*	*3,096.73*
Land Area	*33,071*	*3,006.49*
Population Density	*63*	*5.71*
Total Families	*62,653*	*5,695.71*



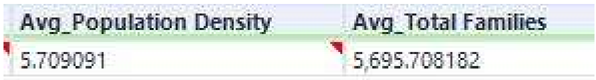
Sum_2010 Census	Sum_Land Area	Sum_Households with Under 18	Sum_Population Density
213,862	33,071.380389	34,064	62.8



Sum_Total Families	Sum_Total Pawdacity Sales
62,652.79	3,773,304



Avg_Total Pawdacity Sales	Avg_2010 Census	Avg_Land Area	Avg_Households with Under 18
343,027.636364	19,442	3,006.489126	3,096.727273



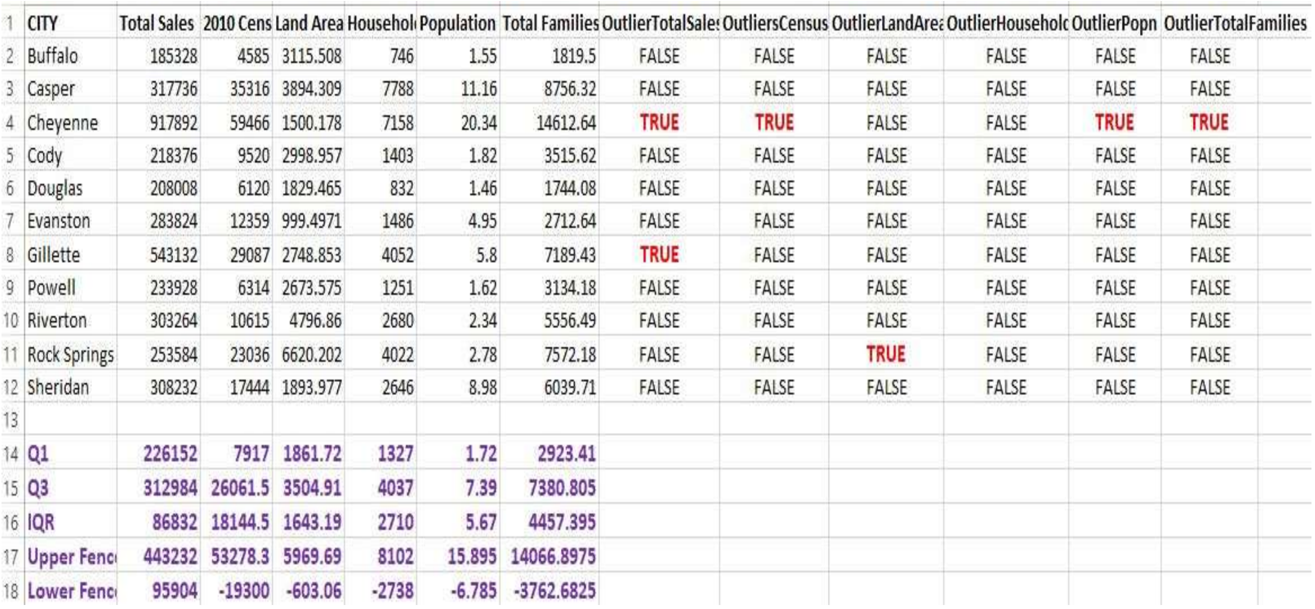
Avg_Population Density	Avg_Total Families
5.709091	5,695.708182

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.

Ans: I found 3 outlier cities in the training set that are: **Cheyenne, Gillette & RockSprings.**
I used Excel for finding these outliers.
First, I ignored the city Cheyenne because it outlies in 2010 Census, Total Pawdacity Sales, Population Density and Total Families, which indicates it's a big city.
Second, I ignored RockSprings city because it outlies only in Land Area not in Total Sales.
Hence I decided to **remove the city Gillette** because it outlies in Total Pawdacity Sales only, and have all other values within the quartile ranges which is abnormal.



1	CITY	Total Sales	2010 Cens	Land Area	Househol	Population	Total Families	OutlierTotalSale	OutliersCensus	OutlierLandAre	OutlierHousehol	OutlierPopn	OutlierTotalFamilies
2	Buffalo	185328	4585	3115.508	746	1.55	1819.5	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
3	Casper	317736	35316	3894.309	7788	11.16	8756.32	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
4	Cheyenne	917892	59466	1500.178	7158	20.34	14612.64	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE
5	Cody	218376	9520	2998.957	1403	1.82	3515.62	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
6	Douglas	208008	6120	1829.465	832	1.46	1744.08	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
7	Evanston	283824	12359	999.4971	1486	4.95	2712.64	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
8	Gillette	543132	29087	2748.853	4052	5.8	7189.43	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
9	Powell	233928	6314	2673.575	1251	1.62	3134.18	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
10	Riverton	303264	10615	4796.86	2680	2.34	5556.49	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
11	Rock Springs	253584	23036	6620.202	4022	2.78	7572.18	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
12	Sheridan	308232	17444	1893.977	2646	8.98	6039.71	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
13													
14	Q1	226152	7917	1861.72	1327	1.72	2923.41						
15	Q3	312984	26061.5	3504.91	4037	7.39	7380.805						
16	IQR	86832	18144.5	1643.19	2710	5.67	4457.395						
17	Upper Fenc	443232	53278.3	5969.69	8102	15.895	14066.8975						
18	Lower Fenc	95904	-19300	-603.06	-2738	-6.785	-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.