Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?

We are trying to determine the appropriate location for Pawdacity to open a new store. In order to determine the location we are trying to project sales in different locations so we can finalize on the location for new store.

2. What data is needed to inform those decisions?

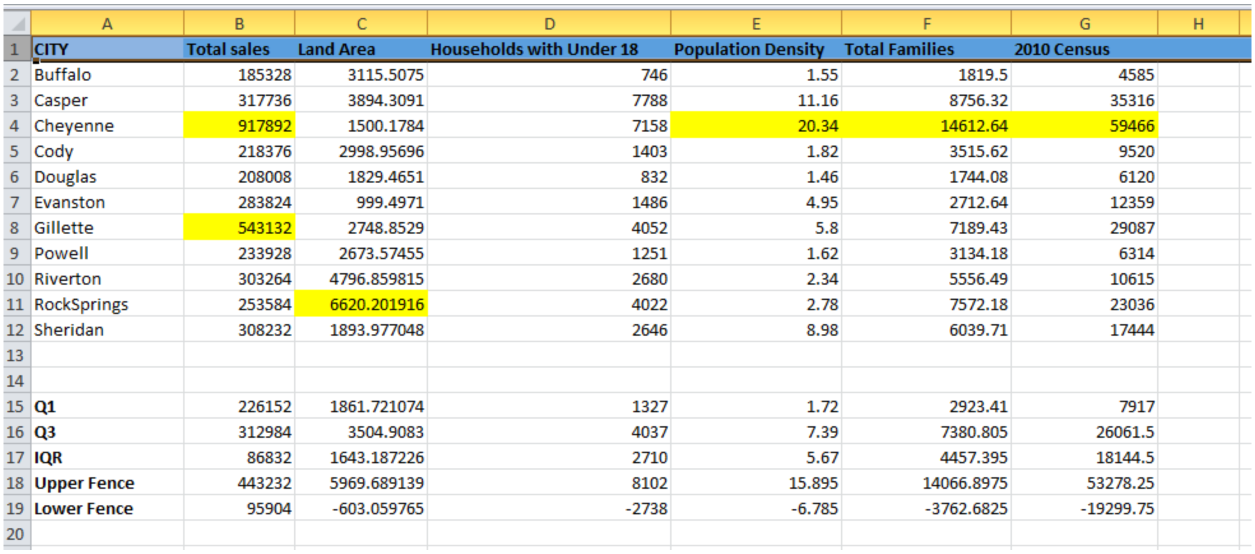
In order to project/forecast sales volume, we need past sales data, data about different cities and its population, demographics etc. In addition we will also need information of other competitor stores in different cities so we understand the market better. For this we will total sales of all Pawdicity existing stores, Population, Population density, Demographics of customers like households with under 18 years, total number of families with pets, Land area etc. In addition it would be nice to know the competitor data and sales volumes for each competitor in different cities.

Step 2: Building the Training Set

Column	**Sum**	**Average**
Census Population	*213,862*	*19442*
Total Pawdacity Sales	*3,773,304*	*343027.64*
Households with Under 18	*34,064*	*3006.49*
Land Area	*33,071*	*3069.73*
Population Density	*63*	*5.71*
Total Families	*62,653*	*5695.71*

Step 3: Dealing with Outliers

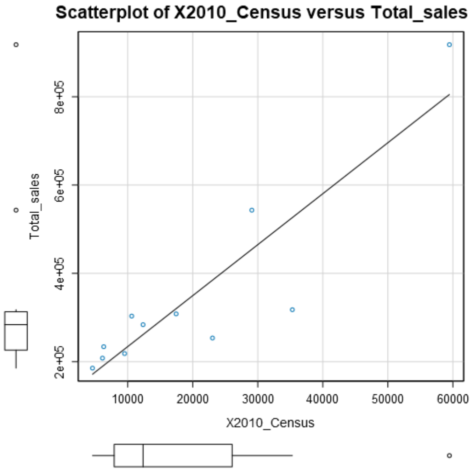
As suggested I have used the IQR method to calculate upper and lower fence values for each field and was able to determine that ‘Cheyenne’ is the only one outlier city present in this dataset.



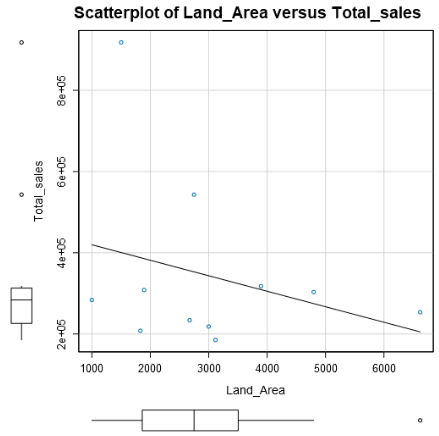
	A	B	C	D	E	F	G	H
1	CITY	Total sales	Land Area	Households with Under 18	Population Density	Total Families	2010 Census	
2	Buffalo	185328	3115.5075	746	1.55	1819.5	4585	
3	Casper	317736	3894.3091	7788	11.16	8756.32	35316	
4	Cheyenne	917892	1500.1784	7158	20.34	14612.64	59466	
5	Cody	218376	2998.95696	1403	1.82	3515.62	9520	
6	Douglas	208008	1829.4651	832	1.46	1744.08	6120	
7	Evanston	283824	999.4971	1486	4.95	2712.64	12359	
8	Gillette	543132	2748.8529	4052	5.8	7189.43	29087	
9	Powell	233928	2673.57455	1251	1.62	3134.18	6314	
10	Riverton	303264	4796.859815	2680	2.34	5556.49	10615	
11	RockSprings	253584	6620.201916	4022	2.78	7572.18	23036	
12	Sheridan	308232	1893.977048	2646	8.98	6039.71	17444	
13								
14								
15	Q1	226152	1861.721074	1327	1.72	2923.41	7917	
16	Q3	312984	3504.9083	4037	7.39	7380.805	26061.5	
17	IQR	86832	1643.187226	2710	5.67	4457.395	18144.5	
18	Upper Fence	443232	5969.689139	8102	15.895	14066.8975	53278.25	
19	Lower Fence	95904	-603.059765	-2738	-6.785	-3762.6825	-19299.75	
20								

Are there any cities that are outliers in the training set?

Total sales volume: Based on the Total sales volume, we see two outliers which are Gillette and Cheyenne, however the value for Gillette is higher than the upper fence but not abnormally high and this can still be a possible number for sales volume. For Gillette, this makes sense since we'd expect the relationship to behave this way. Based on the fitted line, the outliers are in line with the relationship, so we'd leave it in.



Land area: For the land area variable we see that Rock Springs city has higher value than the upper fence again however it is not abnormally high value and just that it is city with bigger area. So we can keep Rock Springs too and not remove this outlier. Again, this makes sense since we'd expect the relationship to behave this way. Based on the fitted line, the outliers are in line with the relationship, so we'd leave it in.



Households with under 18: There are no outliers in this field.

Population density, Total Families and 2010 Census:

We can consider 'Cheyenne' to be the outlier; this has an abnormally high value for Total sales volume. This could be because it has high population and population density values. But if we look at the population (census 2010) we see that it has higher population when compared to other cities for such small land area which really doesn't make sense. If we observe other values in the dataset we can see that the average population lies in 10000 to 30000 ranges and this is definitely a higher value which doesn't make sense. This can be a problem and can skew our results if we consider keeping this value.
So we need to either correct this by going back to source or impute or delete this outlier.

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.

So, Cheyanne is the outlier city where we need to either go back to source to see which data has been reported incorrectly or impute the correct values so the result is not skewed. If we cannot decide to do either then it is better to remove this value from the dataset.