Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

1. What decisions needs to be made?

To select the best city choice to open a new Pawdacity store.

2. What data is needed to inform those decisions?
    - The monthly sales data for all of the Pawdacity stores
    - partially parsed data file for population numbers.
    - sales of all competitor stores where total sales is equal to 12 months of sales
    - Demographic data.

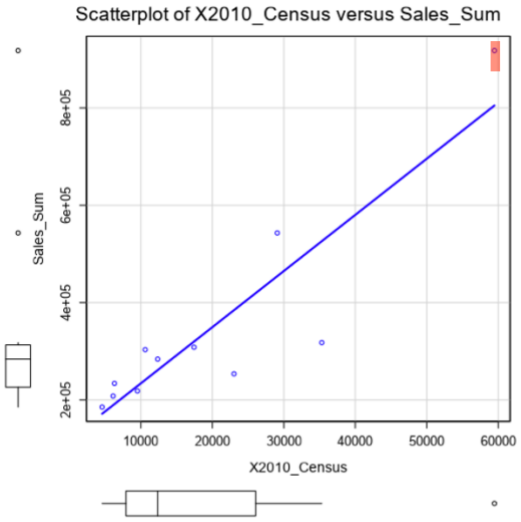
Step 2: Building the Training Set

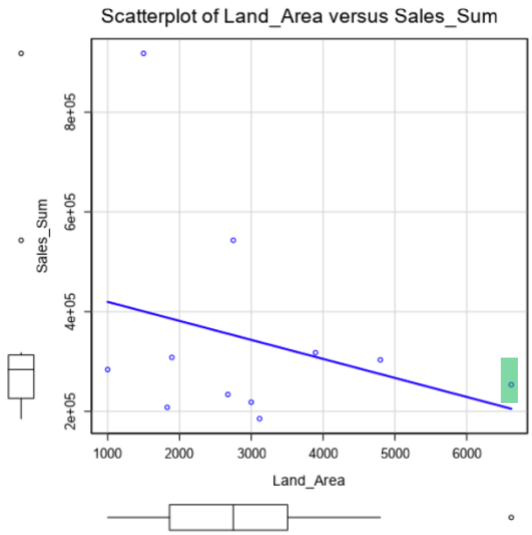
Column	Sum	Average
Census Population	*213,862*	*19442*
Total Pawdacity Sales	*3,773,304*	*343027.63*
Households with Under 18	*34,064*	*3096.72*
Land Area	*33,071*	*3006.48*
Population Density	*63*	*5.723*
Total Families	*62,653*	*5695.72*

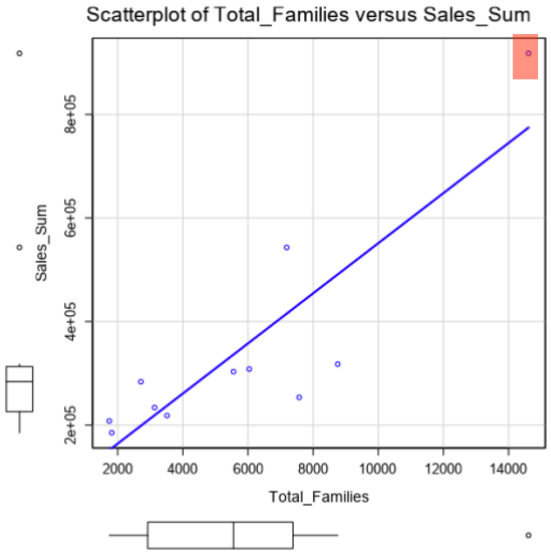
Record	Average_Total_Sales	Average_2010 Census	Average_Land Area	Average_Households with Under 18	Average_Population Density	Average_Total Families
1	343027.636364	19442	3006.489126	3096.727273	5.727273	5695.727273

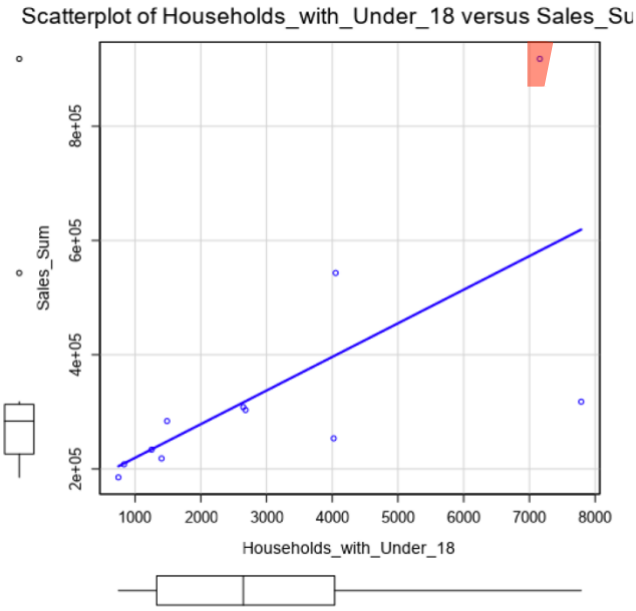
Record	Sum_Total_Sales	Sum_2010 Census	Sum_Land Area	Sum_Households with Under 18	Sum_Population Density	Sum_Total Families
1	3773304	213862	33071.380389	34064	63	62653

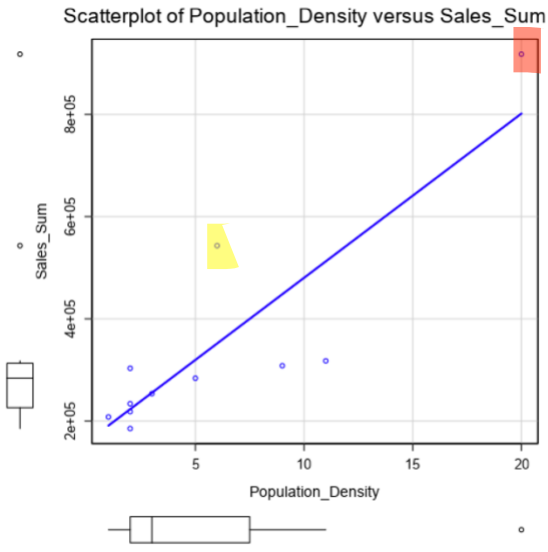
Step 3: Dealing with Outliers











- Red: Cheyenne
- Green: Rock Spring
- Yellow :Gillite

Record	CITY	Total_Sales	2010 Census	Land Area	Households with Under 18	Population Density	Total Families
1	Buffalo	185328	4585	3115.5075	746	2	1820
2	Casper	317736	35316	3894.3091	7788	11	8756
3	Cheyenne	917892	59466	1500.1784	7158	20	14613
4	Cody	218376	9520	2998.95696	1403	2	3516
5	Douglas	208008	6120	1829.4651	832	1	1744
6	Evanston	283824	12359	999.4971	1486	5	2713
7	Gillette	543132	29087	2748.8529	4052	6	7189
8	Powell	233928	6314	2673.57455	1251	2	3134
9	Riverton	303264	10615	4796.859815	2680	2	5556
10	Rock Springs	253584	23036	6620.201916	4022	3	7572
11	Sheridan	308232	17444	1893.977048	2646	9	6040

Activate Windows

ilies	sales_sum	2010Censu	LandArea	Household	population	TotalFamilies
Q1	226152	7917	1861.721	1327	2	2923.5
Q3	312984	26061.5	3504.908	4037	7.5	7380.5
IQR	86832	18144.5	1643.187	2710	5.5	4457
upper fenc	443232	53278.25	5969.689	8102	15.75	14066
lower fence	95904	-19299.8	-603.06	-2738	-8.25	-3762

Are there any cities that are outliers in the training set? Yes Rock Spring, Cheyenne, gillette
Which outlier have you chosen to remove or impute? I can see that Cheyenne is an outlier in the training set as shown on the table and plots, its is an outlier depending on many features 2010 Census total_sales, population density and total families unlike the other cities who have outlier in 1 or two features. Therefore, Cheyenne records should be imputed.