Project 2.1: Data Cleanup

Provide an explanation of the key decisions that need to be made. (250 word limit)

Key Decisions:

Answer these questions

1. What decisions need to be made?
   Management needs to decide on which city in Wyoming to open Pawdacity's newest store based on predicted yearly sales.

2. What data is needed to inform those decisions?
   The management requires the decision to be based on predicted yearly sales. Therefore to predict sales we shall require past sales data of all Pawdacity's stores, demographic data of cities in Wyoming, data on Wyoming population in various cities, and sales data of competitor stores.

Step 2: Building the Training Set

City	Sales	2010 Census	Land Area	Households with Under 18	Population Density	Total Families
Buffalo	185328	4585	3115.51	746	1.55	1819.5
Casper	317736	35316	3894.31	7788	11.16	8756.32
Cheyenne	917892	59466	1500.18	7158	20.34	14612.64
Cody	218376	9520	2998.96	1403	1.82	3515.62
Douglas	208008	6120	1829.47	832	1.46	1744.08
Evanston	283824	12359	999.50	1486	4.95	2712.64
Gillette	543132	29087	2748.85	4052	5.8	7189.43
Powell	233928	6314	2673.57	1251	1.62	3134.18
Riverton	303264	10615	4796.86	2680	2.34	5556.49
Rock Springs	253584	23036	6620.20	4022	2.78	7572.18
Sheridan	308232	17444	1893.98	2646	8.98	6039.71
Totals	3773304	213862	33071.38	34064	62.8	62652.79

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

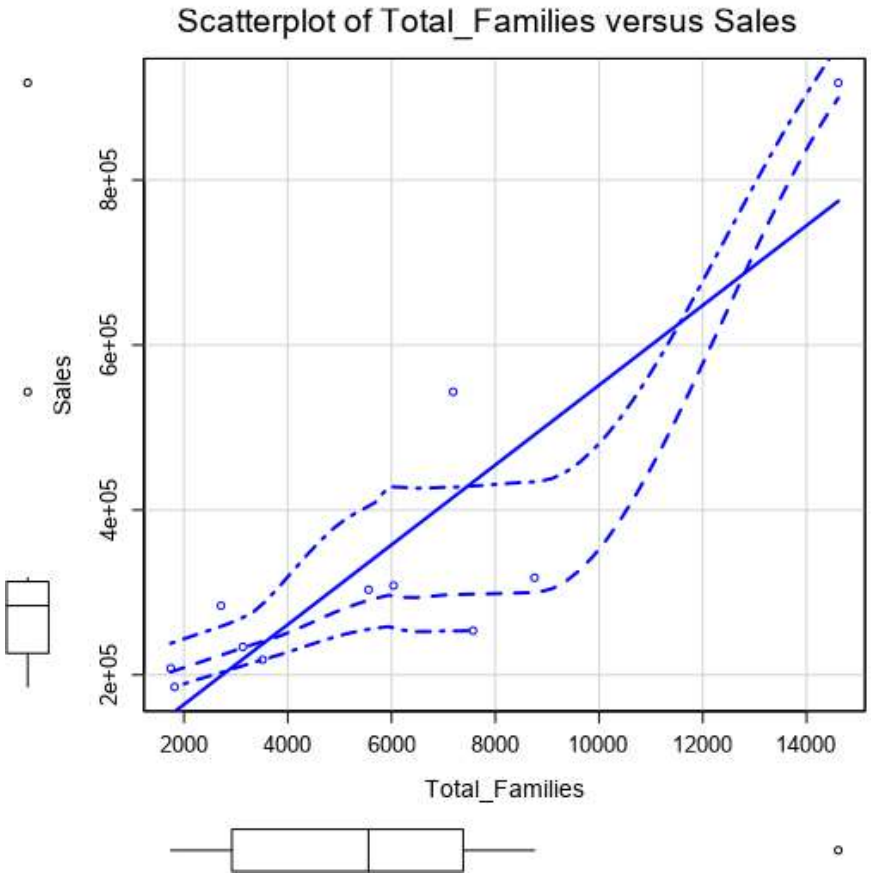
Column	Sum	Average
Census Population	*213,862*	*19442*
Total Pawdacity Sales	*3,773,304*	*343027.64*
Households with Under 18	*34,064*	*3096.73*
Land Area	*33,071*	*3006.49*
Population Density	*63*	*5.71*
Total Families	*62,653*	*5695.71*

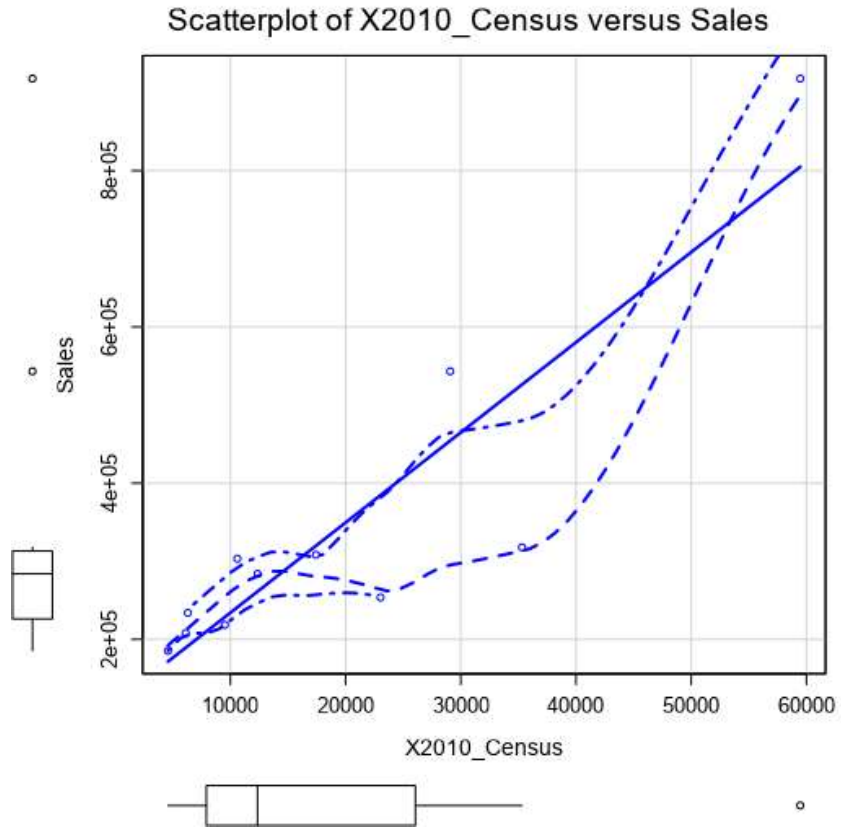
Step 3: Dealing with Outliers

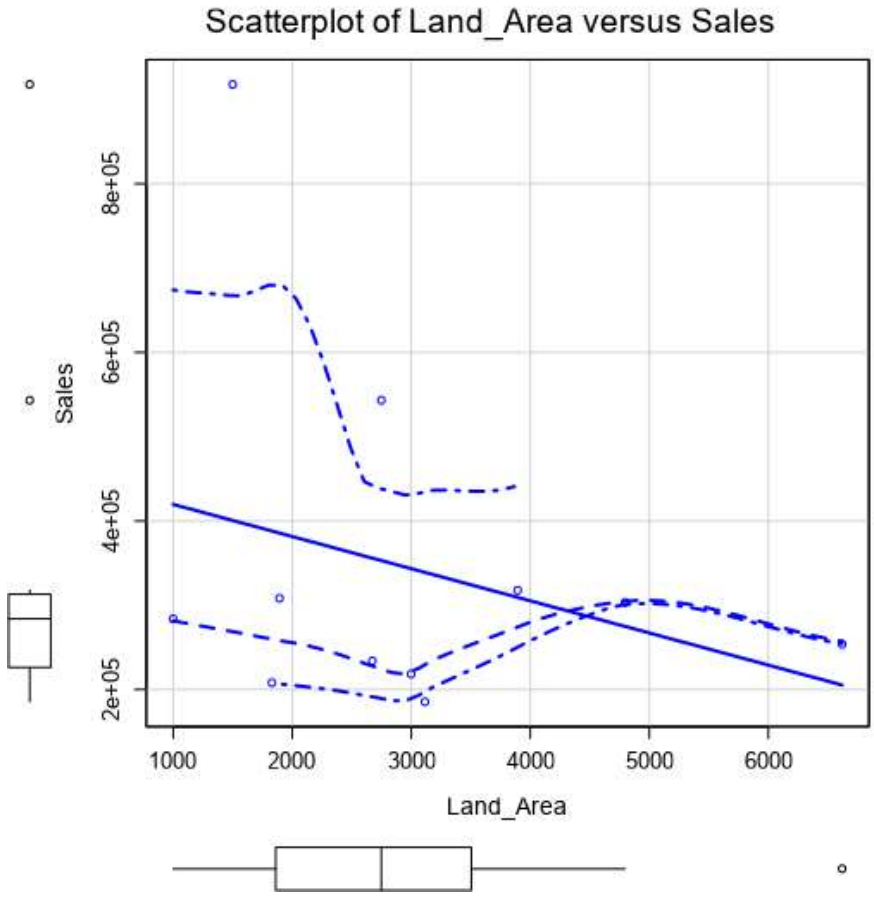
Answer these questions

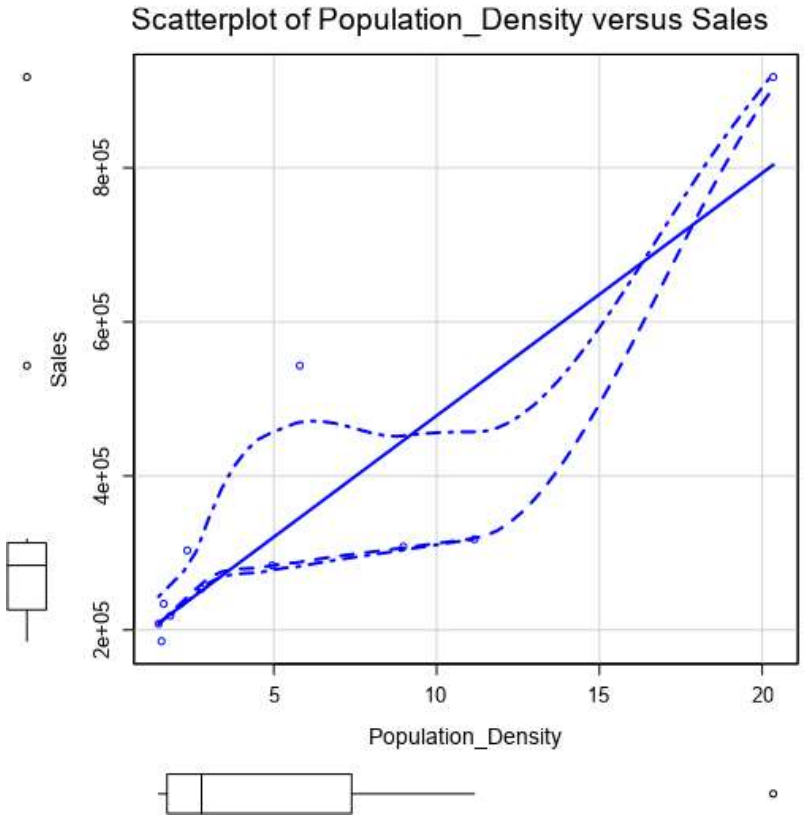
Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.

I utilized the following scatter plots to identify outliers:




Scatterplot of X2010_Census versus Sales
Sales
2e+05
4e+05
6e+05
8e+05
10000
20000
30000
40000
50000
60000
X2010_Census


Scatterplot of Land_Area versus Sales
Sales
2e+05
4e+05
6e+05
8e+05
1000
2000
3000
4000
5000
6000
Land_Area



The city of **Cheyenne** has outliers values across the following fields:

Field	**Outlier Value**
2010_Census	59,466
Population_density	20.34
Total_families	14,612.64
Sales	91,7892

Because the city of **Cheyenne** has relatively massive outlier values across multiple fields it will lessen the model's ability to make predictions as keeping it in the model will skew all other predictions.
Therefore it would be best to filter out and remove the record of the city of **Cheyenne.**

The record of the city of **Rock springs** has an outlier value in the field of **land_area** with a value of 6620.20. Based on the fitted line on the scatter plot, the outlier is in line with the relationship, so I'd leave it in.

The record of the city of **Gillette** has an outlier value in field of **sales** with a value of 54,3132. Rather than removing the record of this city I will build a model with and without this record to compare the effect of the outlier.