# 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?

Make recommendation for a city to open a new store based on predicted sales from 2010.

2. What data is needed to inform those decisions?

To predict sales based on geospatial locations, need to know for the cities currently with Pawdacity stores, the population demographics including population of city, land area, population density, as well as total number of families and households with members under 18 as there may be more pet owners with families and children. Additionally, total sales of these Pawdacity stores for 2010 are needed.

## 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* | *3006.49* |
| *Population Density* | *63* | *5.71* |
| *Total Families* | *62,653* | *5,695.71* |

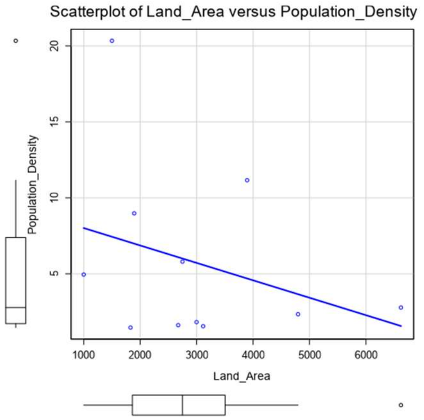
## Step 3: Dealing with Outliers

*Answer these questions*

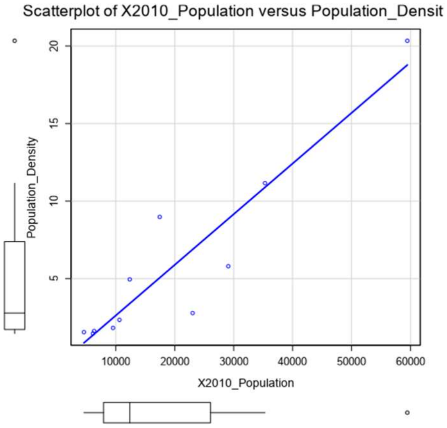
Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.

Outliers identified.

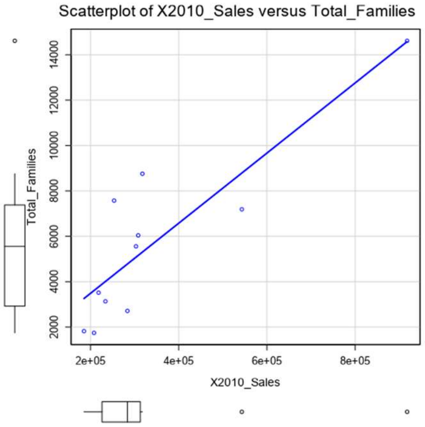
| City | Variable | Exceeds |
|---|---|---|
| **Rock Springs** | Land Area | Upper Fence |
| **Cheyenne** | Population Density | Upper Fence |
| **Cheyenne** | Total Families | Upper Fence |
| **Cheyenne** | Population | Upper Fence |
| **Cheyenne** | 2010 Sales | Upper Fence |
| **Gillette** | 2010 Sales | Upper Fence |



From the scatterplot of Land Area vs Population Density, the outlier for Land Area is Rock Springs, while for Population Density is Cheyenne. While Rock Springs is identified to be outlier for Land Area, it is not that much larger than the Upper Fence (6620 > 5969) and is still reasonable, so it should not be removed.

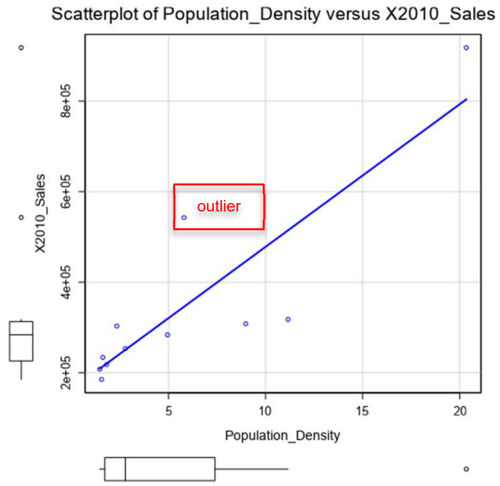


From the scatterplot of 2010 Population vs Population Density, the outlier for both variables is Cheyenne. The correlation between the 2 variables for Cheyenne mostly fits on the regression for the other cities and this is a case where Cheyenne is just a large 2010 Population with a large Population Density. It should not be removed.



From the scatterplot of 2010 Sales vs Total Families, the outlier for Total Families is Cheyenne and for 2010 Sales are Cheyenne and Gillette. The correlation between the 2 variables for Cheyenne mostly fits on the regression for the other cities and this is a case where Cheyenne just had a larger Total Families and hence larger 2010 Sales. It should not be removed.

The same can said for Gillette too.



From the scatterplot of Population Density vs 2010 Sales, the outlier for Population Density is Cheyenne and for 2010 Sales are Cheyenne and Gillette. The correlation between the 2 variables for Cheyenne mostly fits on the regression for the other cities and this is a case where Cheyenne just had a larger Population Density and hence larger 2010 Sales. It should not be removed.

However, for Gillette, the 2010 Sales appears abnormally higher than typical for its Population Density. For other cities with similar Population Density, their 2010 Sales are almost 2 times less than Gillette's. Given there are only 11 cities, removing Gillette will severely compromise the statistical power of the model. Instead 2010 Sales for Gillette should be imputed using a regression based on 2010 Sales and the Population Density, predicting Gillette's 2010 Sales from its Population Density.