# Project 2.1: Data Cleanup

## Step 1: Business and Data Understanding

### Key Decisions:

1. What decisions needs to be made?
   The leading pet store Pawdacity, need to decide in which city they should open their new store based on predicted yearly sales. To decide this, data analysis needs to be done first.

2. What data is needed to inform those decisions?
   Data needed to help in making the decision would be:
   - Yearly sales (for 2010) of the pet store across the cities.
   - Demographic data, such as Land area, Population density, Total families and census data on the population of the various cities in Wyoming where the pet store already has branches and also for the cities where the Pet store is considering for expansion
   - Sales details of all competitor stores and their locations

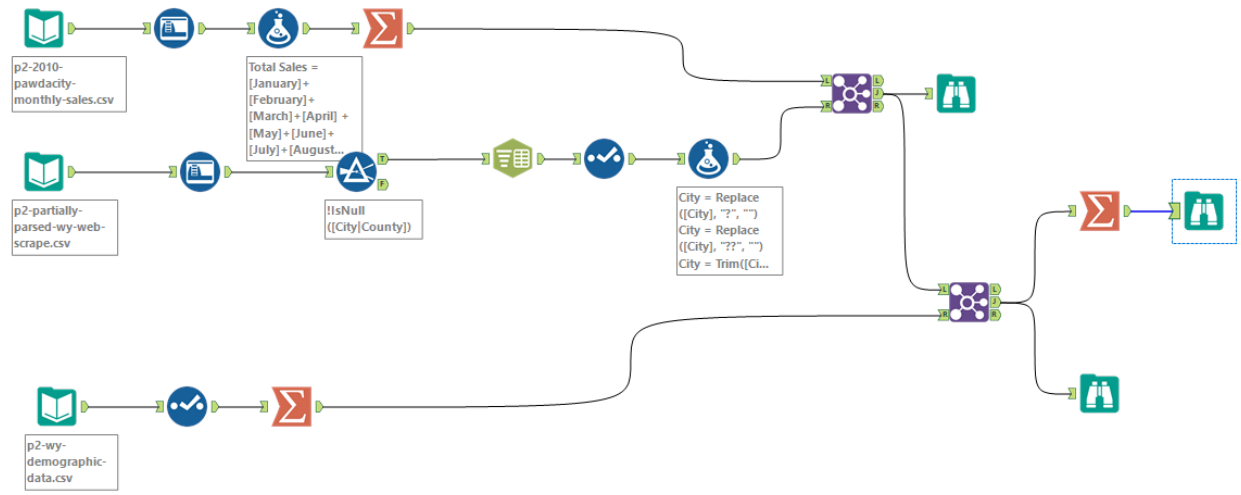
## Step 2: Building the Training Set

| Column | Sum | Average |
|---|---|---|
| *Census Population* | *213,862* | *19442* |
| *Total Pawdacity Sales* | *3,773,304* | *343027.64* |
| *Households with Under 18* | *34,064* | *3096.73* |
| *Land Area* | *33,071* | *3006.45* |
| *Population Density* | *63* | *5.73* |
| *Total Families* | *62,653* | *5695.73* |

*The result from Alteryx:*

| Record # | NAME | Sum_2010 Census | Sum_Sum_Total Sales | Sum_Households with Under 18 | Sum_Total Land Area | Sum_Population Density | Sum_Total Families |
|---|---|---|---|---|---|---|---|
| 1 | Pawdacity | 213862 | 3773304 | 34064 | 33071 | 63 | 62653 |

*Alteryx workflow developed to build the data set:*




# Step 3: Dealing with Outliers

*Answer these questions*

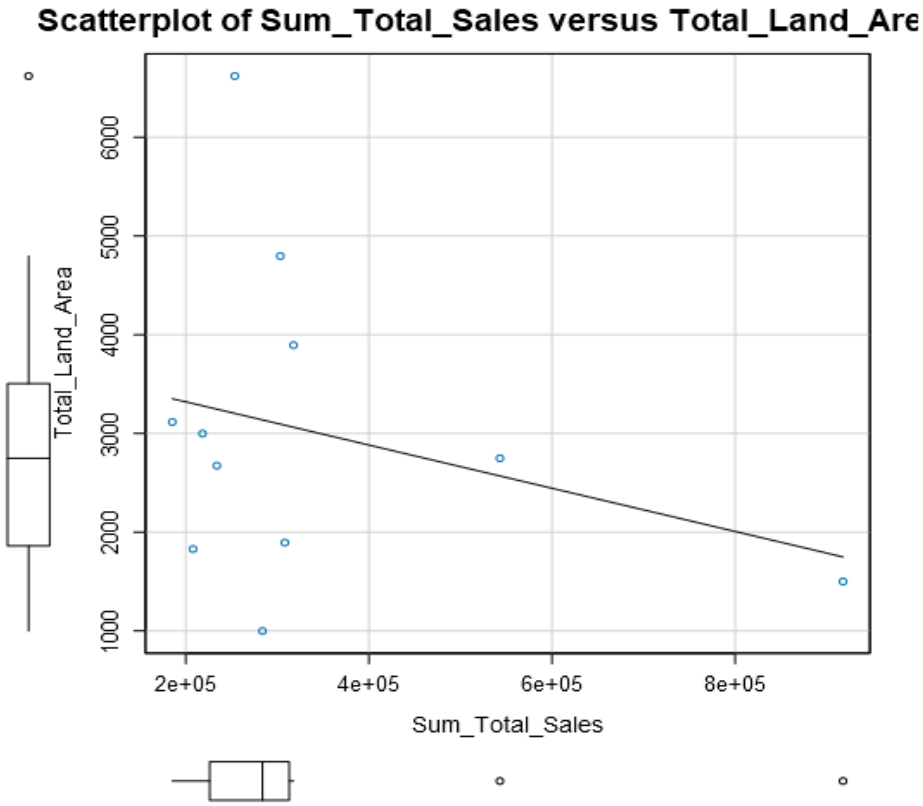
Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.

| NAME | CITY | 2010 Census | Sum_Total Sales | Households with Under 18 | Total Land Area | Population Density | Total Families |
|---|---|---|---|---|---|---|---|
| Pawdacity | Buffalo | 4585 | 185328 | 746 | 3116 | 2 | 1820 |
| Pawdacity | Casper | 35316 | 317736 | 7788 | 3894 | 11 | 8756 |
| Pawdacity | Cheyenne | 59466 | 917892 | 7158 | 1500 | 20 | 14613 |
| Pawdacity | Cody | 9520 | 218376 | 1403 | 2999 | 2 | 3516 |
| Pawdacity | Douglas | 6120 | 208008 | 832 | 1829 | 1 | 1744 |
| Pawdacity | Evanston | 12359 | 283824 | 1486 | 999 | 5 | 2713 |
| Pawdacity | Gillette | 29087 | 543132 | 4052 | 2749 | 6 | 7189 |
| Pawdacity | Powell | 6314 | 233928 | 1251 | 2674 | 2 | 3134 |
| Pawdacity | Riverton | 10615 | 303264 | 2680 | 4797 | 2 | 5556 |
| Pawdacity | Rock Springs | 23036 | 253584 | 4022 | 6620 | 3 | 7572 |
| Pawdacity | Sheridan | 17444 | 308232 | 2646 | 1894 | 9 | 6040 |

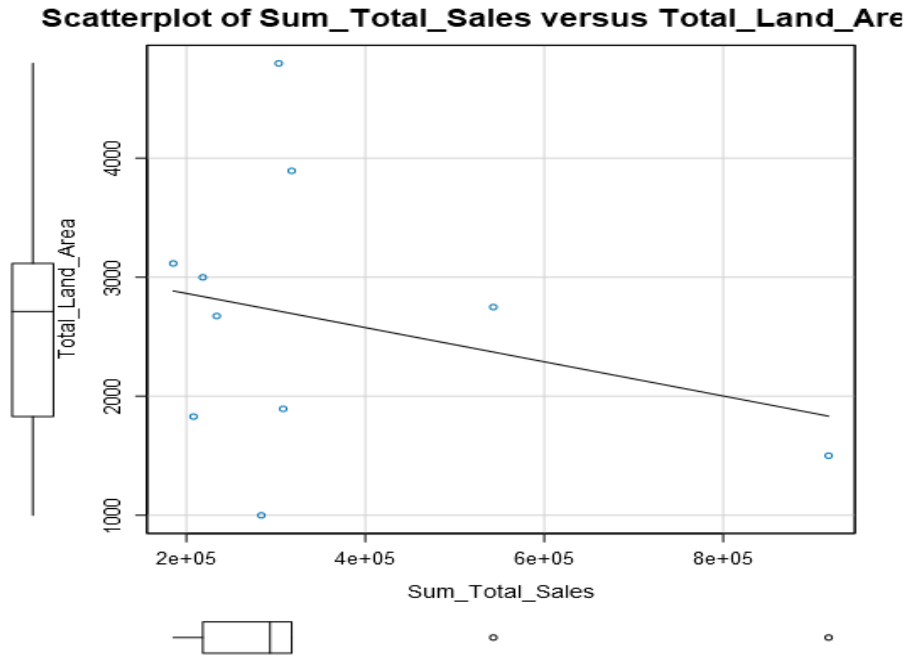
| | | | | | | |
|---|---|---|---|---|---|---|
| Average | 19442 | 343027.6364 | 3096.727273 | 3006.454545 | 5.727272727 | 5695.727273 |
| Quartile 1 | 7917 | 226152 | 1327 | 1861.5 | 2 | 2923.5 |
| Quartile 3 | 26061.5 | 312984 | 4037 | 3505 | 7.5 | 7380.5 |
| IQR | 18144.5 | 86832 | 2710 | 1643.5 | 5.5 | 4457 |
| Upper Fence | 53278.25 | 443232 | 8102 | 5970.25 | 15.75 | 14066 |
| Lower Fence | -19299.75 | 95904 | -2738 | -603.75 | -6.25 | -3762 |

Yes, there are outliers in the data set, which are highlighted in Yellow in the above screen shot.

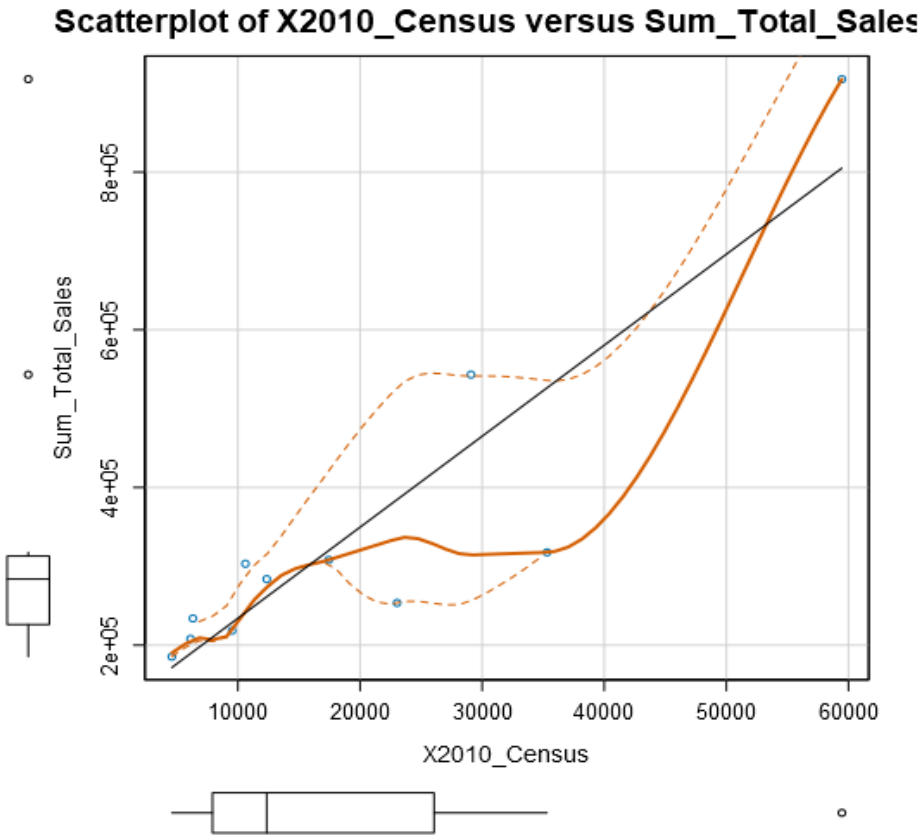
Scatterplot of Total Sales Vs. Land Area of All the cities



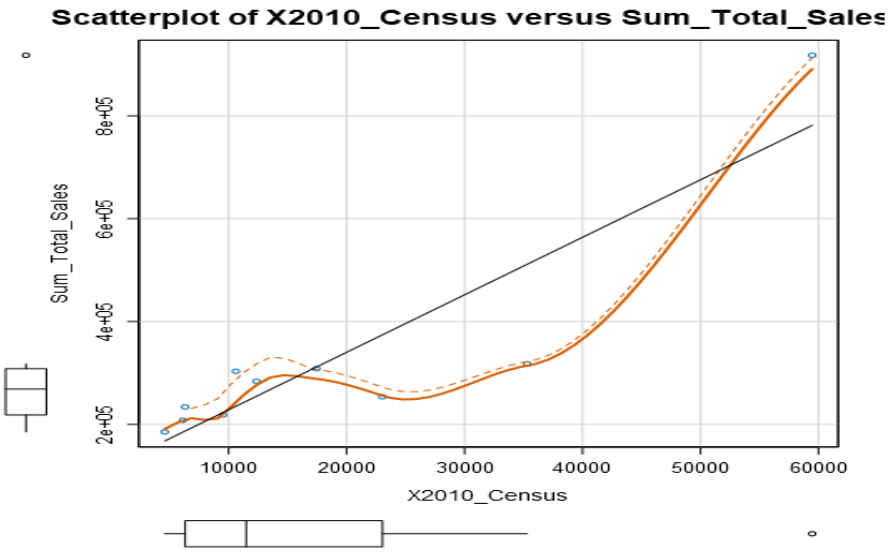

Scatterplot of Total Sales Vs. Land Area of All the cities except Rock Springs



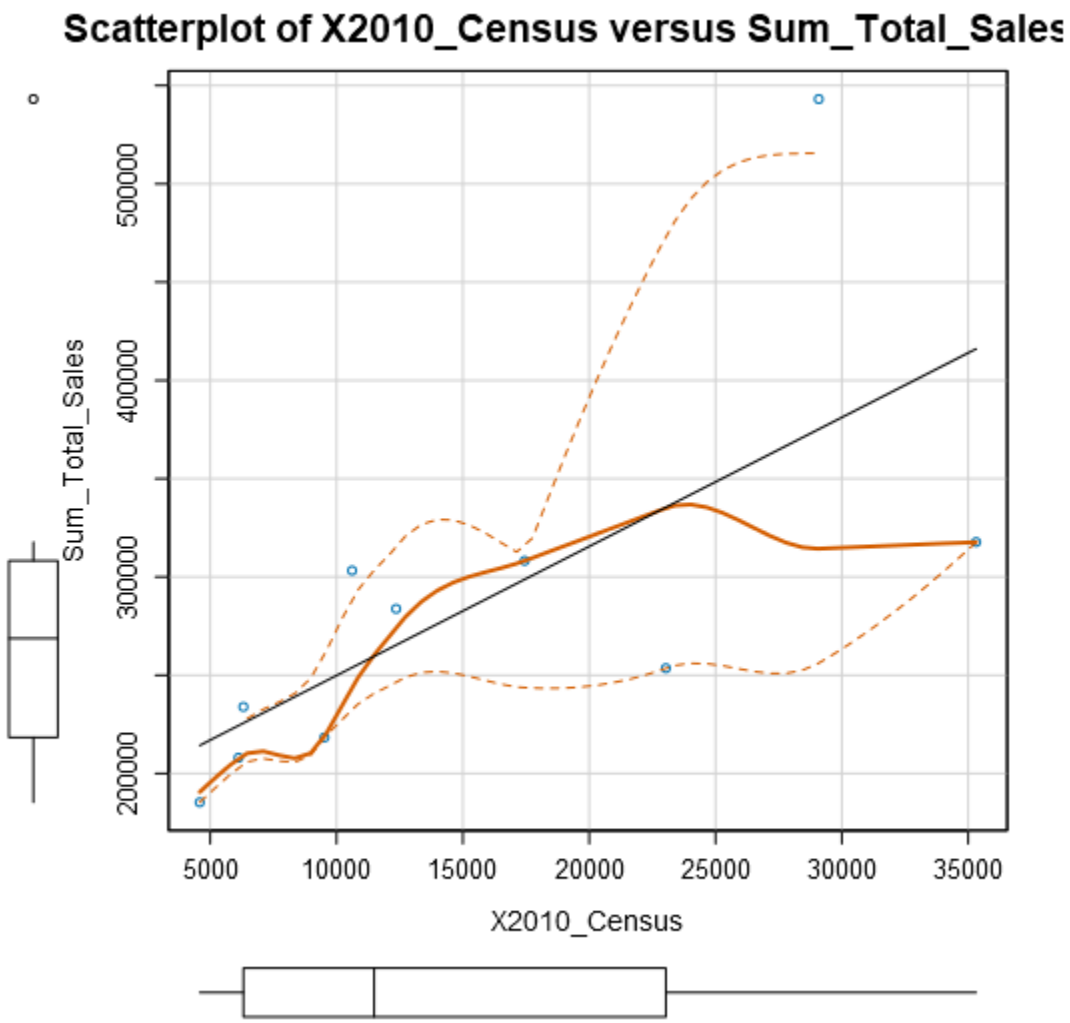
Scatterplot of Total Sales Vs. Census for all the cities:




Scatterplot of Total Sales Vs. Census for all the cities excluding City “Gillette”



Scatterplot of Total Sales Vs. Census for all the cities excluding City “Cheyenne”




Three cities Cheyenne, Gillette and Rock Springs have outliers. Of these, Total land area is least significant to the Total sales. Also, looking at the scatterplot of Total Sales with the Land Area, it can be noted that Rock Springs is in-line with the trend of other cities. So, even though Rock Spring city data is an outlier it need not be imputed.

This leaves us with Gillette and Cheyenne. For Cheyenne, almost all the values are in outlier range except for Households with under 18 and Total land area, but those two variable values are least significant. Also, given the population size, the sales value may even be normal/actual and not really an outlier. In addition, if Cheyenne city’s values are removed/imputed there is a high probability that the model will be skewed.

For Gillette city, the sales value looks highly skewed compared to other cities with similar population and so the total sales may not be significantly related to the Population metrics. Hence, Gillette city’s data can be removed.