# Project 2: Create an Analytical Dataset

By: Ivy Huong Nguyen

## Step 1: Business and Data Understanding

### Key Decisions:

1. What decisions needs to be made?
   The pet store Pawdacity wants to expand its chain by opening another store within the Wyoming area. In order to decide where would be the next most profitable location (obtain highest total sales), the store manager wants me to comb through previous data, clean it and use it to make a recommendation for the new store's location. In order to accomplish this task, I will need to go through some data-cleaning steps before fitting the data into a model for prediction. The cleaning-steps include:
   - Reformat the datasets if necessary
   - Delete or impute all the null values present
   - Merge and join datasets
   - Eliminate outliers before fitting the data into a model
2. What data is needed to inform those decisions?
   - The total sales at each present location: this fact can help us decide which location is in need of a secondary store more. For example, if a particular location has a tremendously high volume of sales going on, there might be it is a good idea to open a second store to slow down the traffic at that store. Doing this can help maintaining and improving customer service for Pawdacity.
   - The total population (census data and total families) of a city
   - The number of competitors in a city
   - The number of households with members who are under 18: This fact is important since family who has kids tends to have pets in their house
   - Population density

## Step 2: Building the Training Set

**Table 1.** Training Set before removing outlier

| CITY | Total_Sales | Households with Under 18 | Land Area | Population Density | Total Families | 2010 Census |
|---|---|---|---|---|---|---|
| Buffalo | 185328 | 746 | 3116 | 2 | 1820 | 4585 |
| Casper | 317736 | 7788 | 3894 | 11 | 8756 | 35316 |
| Cheyenne | 917892 | 7158 | 1500 | 20 | 14613 | 59466 |
| Cody | 218376 | 1403 | 2999 | 2 | 3516 | 9520 |
| Douglas | 208008 | 832 | 1829 | 1 | 1744 | 6120 |
| Evanston | 283824 | 1486 | 999 | 5 | 2713 | 12359 |
| Gillette | 543132 | 4052 | 2749 | 6 | 7189 | 29087 |
| Powell | 233928 | 1251 | 2674 | 2 | 3134 | 6314 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Riverton | 303264 | 2680 | 4797 | 2 | 5556 | 10615 |
| Rock Springs | 253584 | 4022 | 6620 | 3 | 7572 | 23036 |
| Sheridan | 308232 | 2646 | 1894 | 9 | 6040 | 17444 |

**Table 2.** Training Set after removing outlier using 'Total Sales' as a criteria

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

**Table 3.** Statistical Summary for all predictor variables

| Column | Sum | Average |
|---|---|---|
| Census Population | 213,862 | 19442 |
| Total Pawdacity Sales | 3,773,304 | 343,027.64 |
| 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 |

## Step 3: Dealing with Outliers

After combing through the dataset, I decided to eliminate the city Gillette from the training dataset. To decide which city to remove, I first calculated the first quartile (Q1), third quartile (Q3), and the Interquartile Range (IQR) of each predictor variable. The predictor variables are: census population, households with under 18, land area, population density, and total families. I then used the IQR to calculate the upper fence (Q3+1.5IQR), and the lower fence (Q1-1.5IQR) for each predictor variable (Table 4).

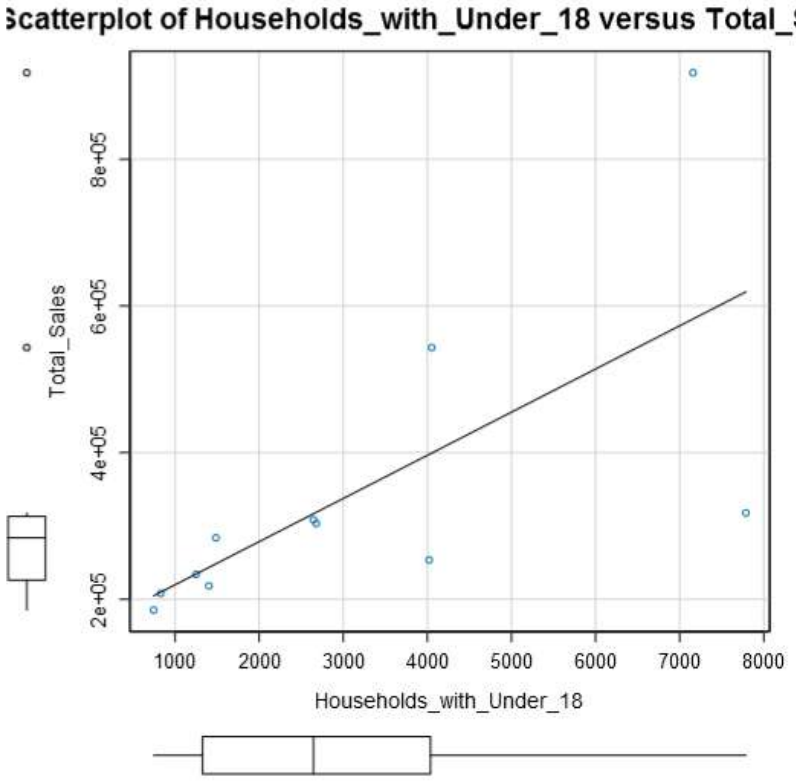
**Table 4.** Median, Q1, Q3, IQR, lower and upper fence values of all predictor variables.

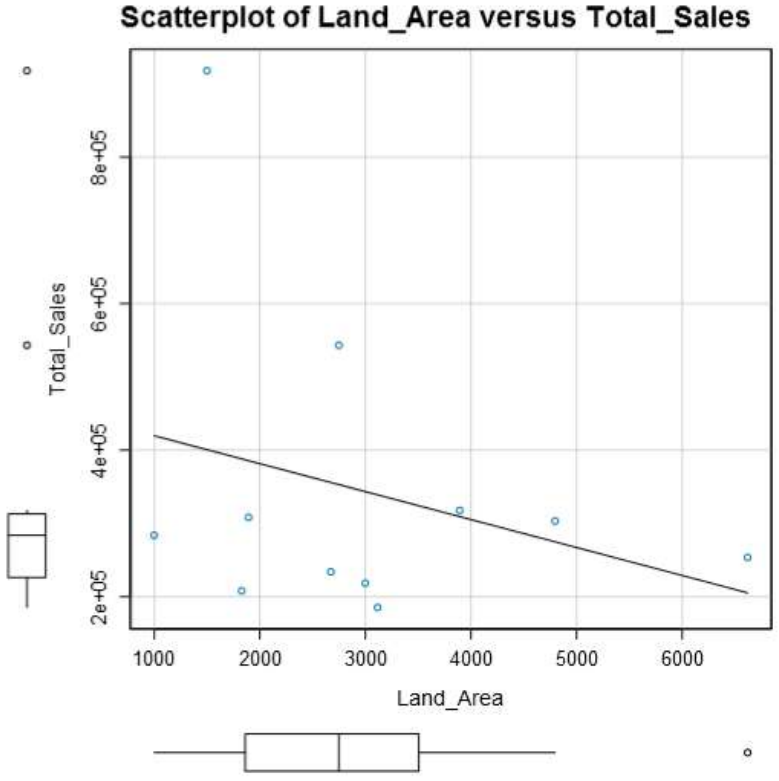
| | Total_Sales | Households with Under 18 | Land Area | Population Density | Total Families | 2010 Census |
|---|---|---|---|---|---|---|
| Median | 283824 | 2646 | 2749 | 3 | 5556 | 12359 |
| Q1 | 185328 | 746 | 999 | 1 | 1744 | 4585 |
| Q3 | 317736 | 4052 | 3894 | 9 | 7572 | 29087 |

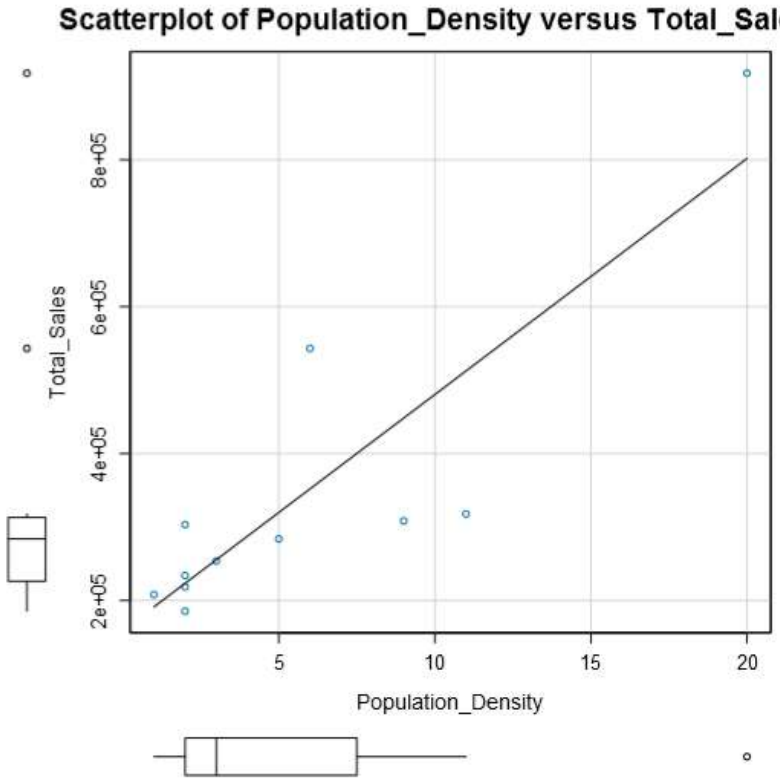
| | | | | | | |
|---|---|---|---|---|---|---|
| IQR | 132408 | 3306 | 2895 | 8 | 5828 | 24502 |
| Upper Fence | 516348 | 9011 | 8236.5 | 21 | 16314 | 65840 |
| Lower Fence | -145584 | 2187 | 1396.5 | 6.5 | 3212 | 17624.5 |

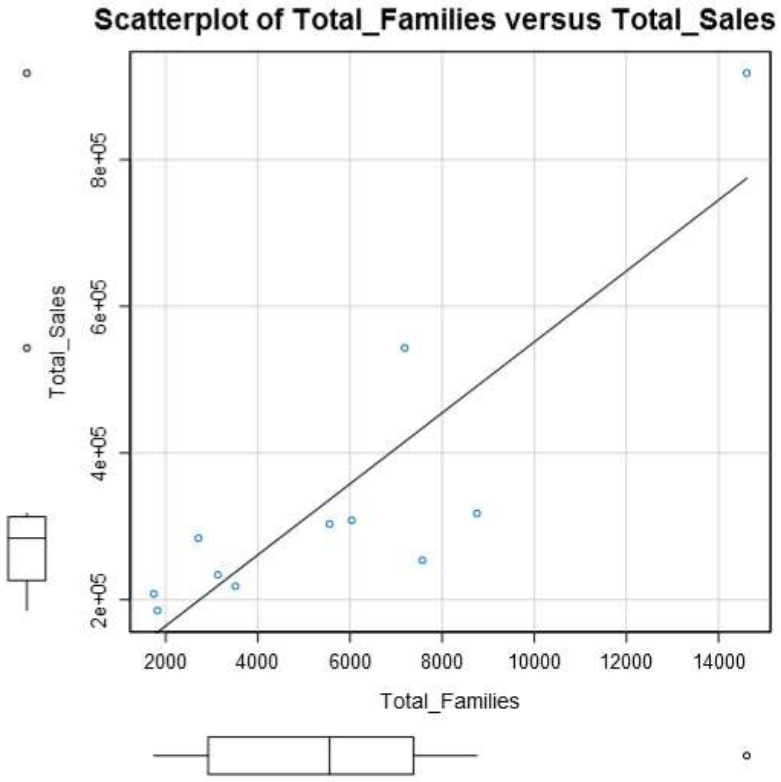
According the 1.5 IQR rule, records that have values less than the lower fence and greater than the upper fence are outliers. However, if I strictly followed this rule then more than one record would be eliminated, which is not applicable for such a small dataset like this case. Thus, to successfully decide which variable I should use to focus on removing an outlier, I first plotted each predictor variable against the target variable (total sales) to get a sense how the data is distributed. (Figures below) I noticed land area has a negative slope while the other predictor variables have positive slope in their linearity with the target variable. However, in all the predictor variables, land area should not have a large impact on the total sales of a store. Therefore, using land area as a criteria for removing an outlier is not really appropriate. The overall point for the project is to predict location that could potentially give the best total sales; thus, total sales would make a great variable to use in choosing an outlier. I then applied the 1.5IQR rule to the 'total sale' column. The result showed that Cheyenne and Gillette are the two cities with an abnormal total sales. However, the data point of Cheyenne city makes perfect sense since there is a high number of census for this city and therefore it is expected that the total sales should be higher. One exception that I noticed is the city Casper, which has a really high household with under 18 but a low total sales. This could be some interesting fact about this city, such as a lot of competitors. By using human logic, scatterplot figures, and the 1.5IQR rule, I decided to filter records based on the 'total sales' upper and lower fence. The city that I removed is Gillette. The final training set is:

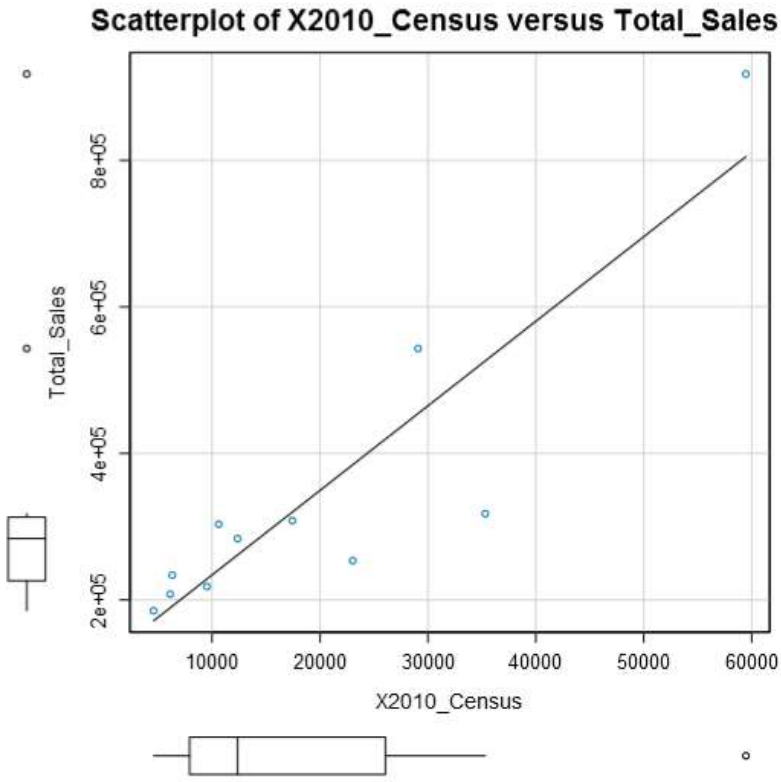
| **CITY** | **Total_Sales** | **Households with Under 18** | **Land Area** | **Population Density** | **Total Families** | **2010 Census** |
|---|---|---|---|---|---|---|
| Buffalo | 185328 | 746 | 3116 | 2 | 1820 | 4585 |
| Casper | 317736 | 7788 | 3894 | 11 | 8756 | 35316 |
| Cheyenne | 917892 | 7158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


Scatterplot of Households_with_Under_18 versus Total_S
Total_Sales
2e+05
4e+05
6e+05
8e+05
1000
2000
3000
4000
5000
6000
7000
8000
Households_with_Under_18


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


Scatterplot of Population_Density versus Total_Sales
Total_Sales
2e+05
4e+05
6e+05
8e+05
5
10
15
20
Population_Density


Scatterplot of Total_Families versus Total_Sales
Total_Sales
2e+05
4e+05
6e+05
8e+05
2000
4000
6000
8000
10000
12000
14000
Total_Families


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