Project 2.1: Data Cleanup

## Step 1: Business and Data Understanding

## Pawdacity is a leading pet store chain in Wyoming with 13 stores throughout the state. This year, Pawdacity would like to expand and open a 14th store. The manager has asked you to perform an analysis to recommend the city for Pawdacity’s newest store, based on predicted yearly sales.

The manager has given you the following information to work with:

1. The monthly sales data for all of the Pawdacity stores for the year 2010.
2. NAICS data on the most current sales of all competitor stores where total sales is equal to 12 months of sales.
3. A partially parsed data file that can be used for population numbers.
4. Demographic data (Households with individuals under 18, Land Area, Population Density, and Total Families) for each city and county in the state of Wyoming.

The data required for the analysis are:

* **City-** The name of the city.
* **2010 Census Population**- The value of the census population.
* **Total Pawdacity Sales-** The total sales of the pet stores.
* **Household with Under 18-** Household with individual under the age 18.
* **Land Area-** The area of the land of the city.
* **Population Density-** The population density of the city.
* **Total Families-** Number of families.

## Step 2: Building the Training Set

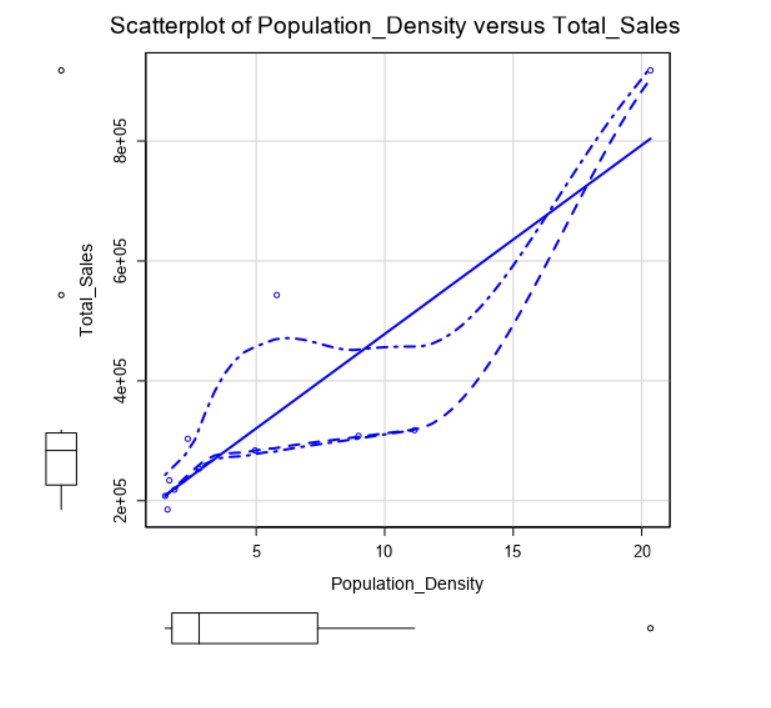
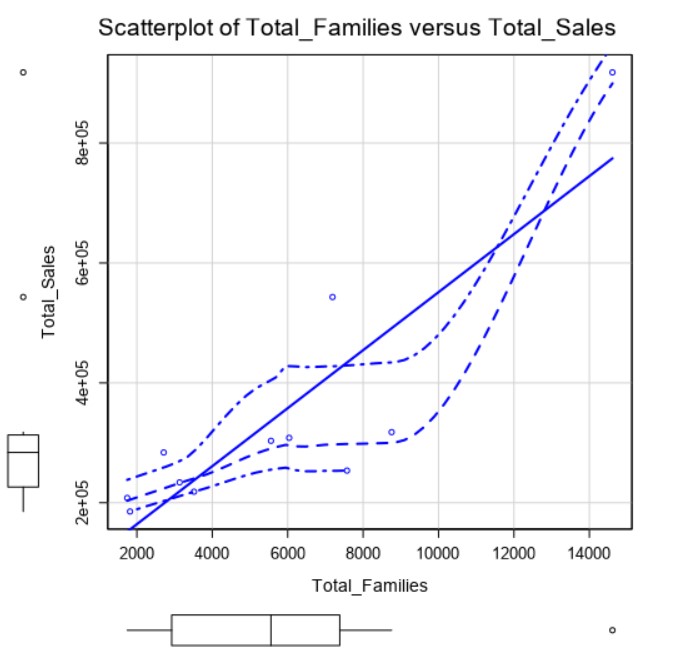
After getting the raw data from 4 different sources, the data needs to be first cleaned in order to remove the dirty data- incorrect, duplicate, missing data, etc. Now, the clean data needs to be formatted to get it in a presentable format. This formatted data from different sources are then joined together to form a single data source.

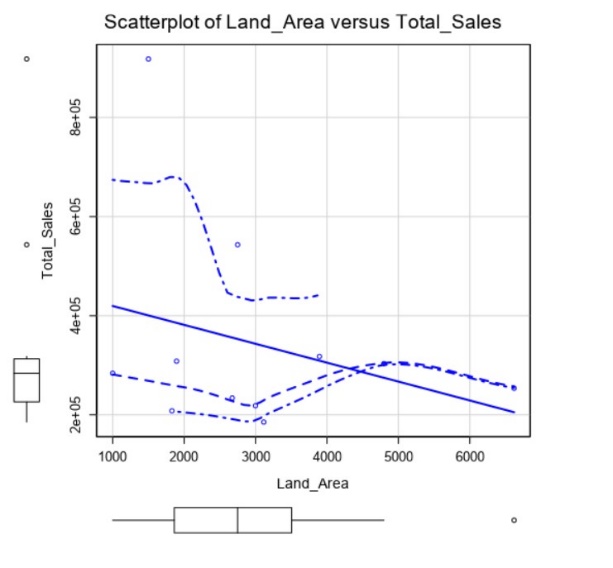
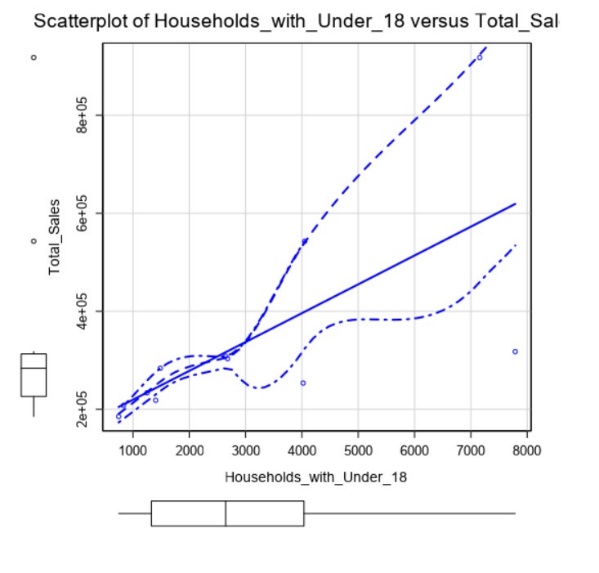
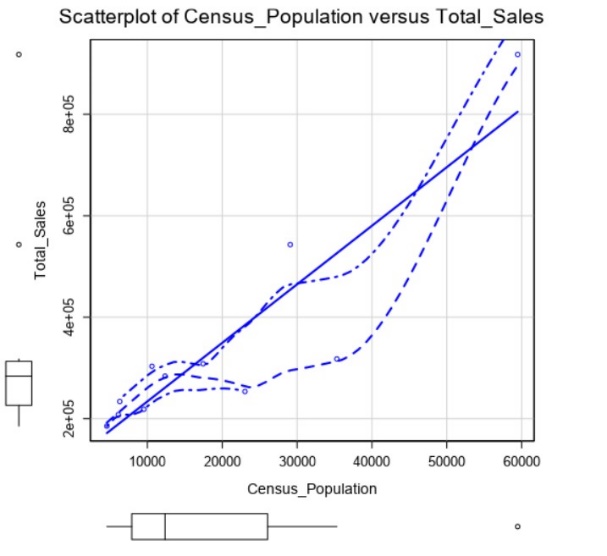
After working with the training set, we find the sum and average of the data given in the table below:

|  |  |  |
| --- | --- | --- |
| **Column** | **Sum** | **Average** |
| *Census Population* | *213,862* | *19,442* |
| *Total Pawdacity Sales* | *3,773,304* | *343,027.63* |
| *Households with Under 18* | *34,064* | *3096.72* |
| *Land Area* | *33,071* | *3006.48* |
| *Population Density* | *63* | *5.709* |
| *Total Families* | *62,653* | *5695.70* |

## Step 3: Dealing with Outliers

The scatterplots and the boxplots are found between the target variable and the predictor variables.



After creating the dataset, we need to identify the outliers. The IQR method is used here to find the outliers. In this method we need to find the upper fence and the lower fence,

Upper fence= Q3(3rd Quartile) + 1.5 x Interquartile Range

Lower fence =Q1(1st Quartile) – 1.5 x Interquartile Range

The values that are above the upper fence or below the lower fence are considered as the outliers.

Our dataset looks like this:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| City | Total  Sales | Census  Population | Land  Area | Households with  Under 18 | Population  Density | Total  Families |
| Buffalo | 185328 | 4585 | 3115.5075 | 746 | 1.55 | 1819.5 |
| Casper | 317736 | 35316 | 3894.3091 | 7788 | 11.16 | 8756.32 |
| Cheyenne | 917892 | 59466 | 1500.1784 | 7158 | 20.34 | 14612.64 |
| Cody | 218376 | 9520 | 2998.957 | 1403 | 1.82 | 3515.62 |
| Douglas | 208008 | 6120 | 1829.4651 | 832 | 1.46 | 1744.08 |
| Evanston | 283824 | 12359 | 999.4971 | 1486 | 4.95 | 2712.64 |
| Gillette | 543132 | 29087 | 2748.8529 | 4052 | 5.8 | 7189.43 |
| Powell | 233928 | 6314 | 2673.5746 | 1251 | 1.62 | 3134.18 |
| Riverton | 303264 | 10615 | 4796.8598 | 2680 | 2.34 | 5556.49 |
| Rock Springs | 253584 | 23036 | 6620.2019 | 4022 | 2.78 | 7572.18 |
| Sheridan | 308232 | 17444 | 1893.977 | 2646 | 8.98 | 6039.71 |

The upper and lower fence for each data field are:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Data Field | Total Sales | Census Population | Land Area | Household with Under 18 | Population Density | Total Families |
| First Quartile | 226152 | 7917 | 1861.721074 | 1327 | 1.72 | 2923.41 |
| Third Quartile | 312984 | 26061.5 | 3504.9083 | 4037 | 7.39 | 7380.805 |
| Inter Quartile Range | 86832 | 18144.5 | 1643.187226 | 2710 | 5.67 | 4457.395 |
| Upper Fence | 443232 | 53278.25 | 5969.689139 | 8102 | 15.895 | 14066.898 |
| Lower Fence | 95904 | -19299.8 | -603.059765 | -2738 | -6.785 | -3762.683 |
|  |  |  |  |  |  |  |

If we analyze each data field:

1. For *Total Sales*:

The outliers are Cheyenne and Gillette.

2. For *Census Population*:

Cheyenne is the outlier.

3. For *Land Area*:

Rock Springs is the outlier.

4.For *Population Density*:

Again Cheyenne is the outlier.

5.For *Total\_Families*:

Cheyenne is the outlier, but the value is almost nearer to the upper fence.

6.There is no outlier for the *Household\_Under 18*.

Although Cheyenne looks like the outlier, but if we analyse the data carefully, we find that the population density is very high for Cheyenne. Hence that might be the reason for the high sales. With the increase in the population density, the sales will also increase.

Gillette is to be considered as the outlier here. It has a low population density, but a high sale value. When compared to the other cities of the same population densities , the high sales value looks abnormal.

Hence the outlier here is the city of Gillette.

The Alteryx workflow is mentioned below:

