# Project 2.1: Data Cleanup

Valencia Becerra Gerardo

## Step 1: Business and Data Understanding

1. What decisions needs to be made?

   The *Pawdacity* executives and employees want to know where (in Wyoming) a pet store would get more sales, to open a new *Pawdacity* store there.

2. What data is needed to inform those decisions?

   We want to make a decision based on predicted yearly sales, and we are provided with the following information to work with:

   - Monthly sales data for all the Pawdacity stores for 2010. We need a new column called "*Total Pawdacity Sales"* which is the sum of all the monthly sales per city (*p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales.csv*).
   - A partially parsed data file that can be used for population numbers. We must clean the data first in order to get the value of the *2010 Census* column (*p2-partially-parsed-wy-web-scrape.csv*).
   - Demographic data (Households with individuals under 18, Land Area, Population Density, and Total Families) for each city and county in the state of Wyoming. (*p2-wy-demographic-data.csv*)
   - NAICS data on the most current sales of all competitor stores where total sales is equal to 12 months of sales.
     (p2-wy-453910-naics-data.csv)

   We'll analyze these data for each city in Wyoming, no other cities are needed for now. For the Data Cleanup part, this last dataset (NAICS data) was not needed.

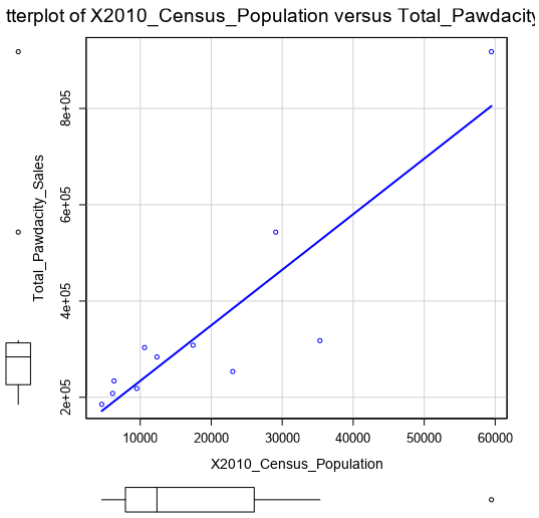
## Step 2: Building the Training Set

| 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* | 3,006.49 |
| *Population Density* | *63* | 5.71 |
| *Total Families* | *62,653* | 5,695.71 |

# Step 3: Dealing with Outliers

At first glance, Cheyenne and Gillette seem to be outliers due to the Total Pawdacity Sale values contained in the box and whisker plot range from 185,328 to 317,736, and these two cities have 917,892 and 543,132 respectively. However, the scatterplots of *2010 Census Population, Land Area, Population Density* and *Total Families,* all of these vs Total Pawdacity Sale which is the value we'll try to predict, show only one outlier:

## 2010 Census Population

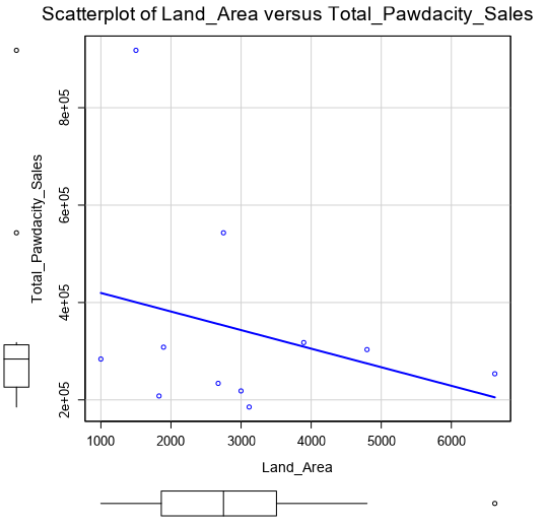



The only outlier here is the Cheyenne value. However, its value of 59,466 fits with the regression line. Also, the *2000 Census* column in the original dataset has a value of 53,011 and the *2014 Estimate* value is 62,845, so our *2010 Census* value seems to be right and is unlikely to be a typo. So, let's use the IQR method to determine if it's an outlier or not. We can see that the Upper Fence is 53,278.25, so the value of 59,466 surpasses this limit and we can conclude that the Cheyenne city is an outlier in this case.

| Values | Notes |
|---|---|
| 4585 | Min Value |
| 7917 | 25th percentile |
| 12359 | 50th percentile |
| 26061.5 | 75th percentile |
| 59466 | Max Value |

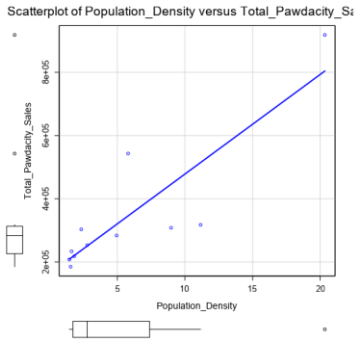
| | |
|---|---|
| IQR | 18144.5 |
| Upper Fence | 53278.25 |
| Lower Fence | -19299.75 |

## Land Area



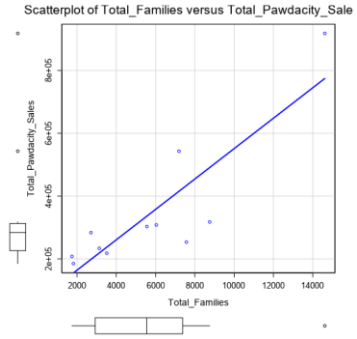

In this scatterplot, we can see that neither Cheyenne nor Gillette are the outliers here (with respect to the land area value). The one value at the right corresponds to Rock Springs, with a value of 6,620.20. We could also use the IQR method to determine if Rock Springs is an outlier or not, but it's also important to note that the Cheyenne value doesn't fit well in the trend shown by the regression line.

## Population Density



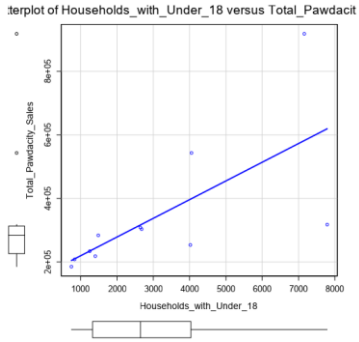

Another scatterplot in which the only outlier is the Cheyenne city with a population density of 20.34, almost twice the value of the highest one in the box-and-whisker plot, Casper, with a value of 11.16.

## Total Families




Once more, the outlier here is the city of Cheyenne with a value of 14,612.64 total families. This time it seems to follow the trend of the regression line.

## Households with Under 18




The *Households with Under 18* box-and-whisker plot doesn't show any obvious outlier, we can see it has a very long right whisker with the values of Cheyenne and Casper being the highest ones, with 7,158 and 7,788 households respectively.

Analyzing the Cheyenne city outlier: 1. It was above the Upper Fence of the IQR method in the 2010 Census Population plot, 2. It didn't follow the trend in the Land Area plot 3. Even though it seems to follow the trend in the Population Density and Total Families plots, this value may likely be affecting the regression line slope. So, out of the other possible outliers, we can remove the Cheyenne outlier to make the predictive model next.