Project 2.1: Data Cleanup
Author : Neavil Porus A

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?

   Pawdacity, a leading pet store chain in **Wyoming**, needs a recommendation on where to open its 14th store. So we need to develop a analytical dataset.

2. What data is needed to inform those decisions?

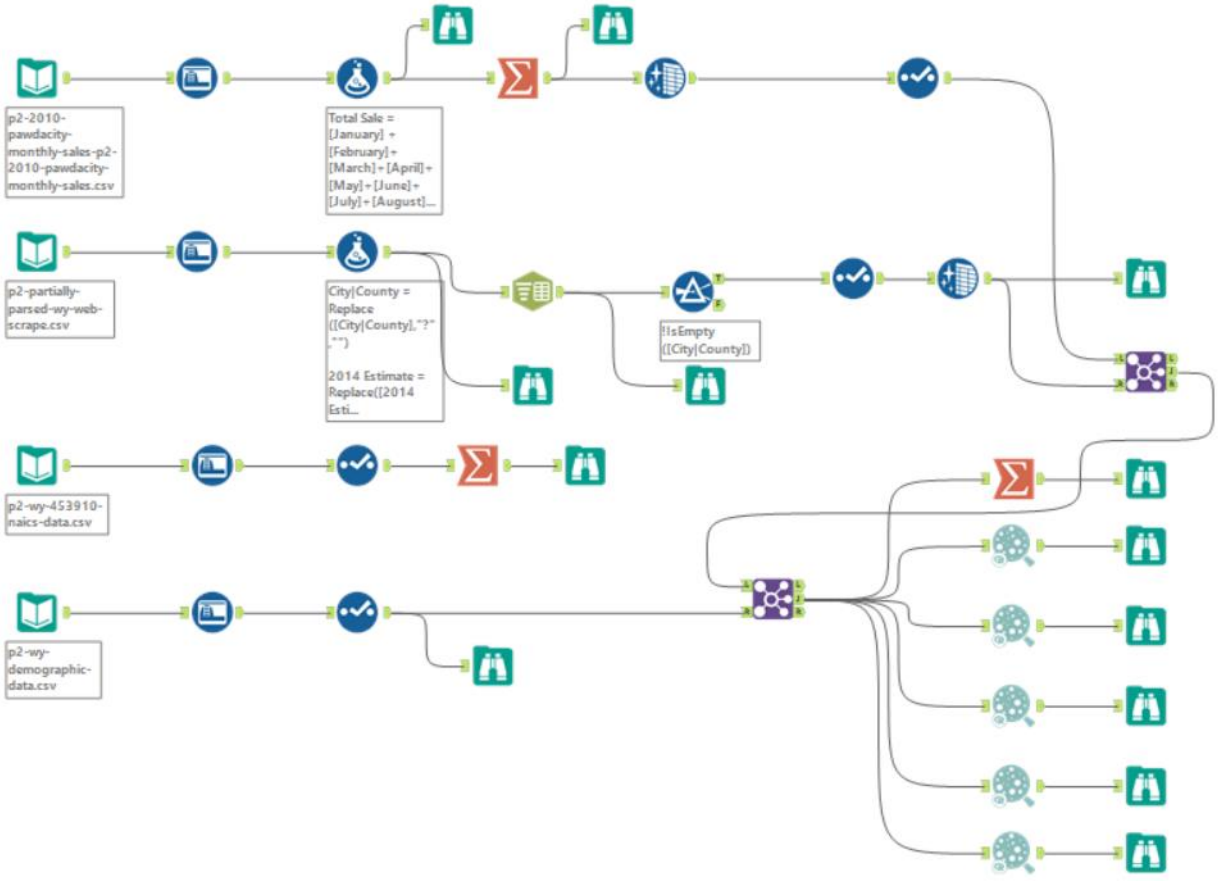
   Data required in order to make an informed decision are given to us in the form of

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

Step 2: Building the Training Set

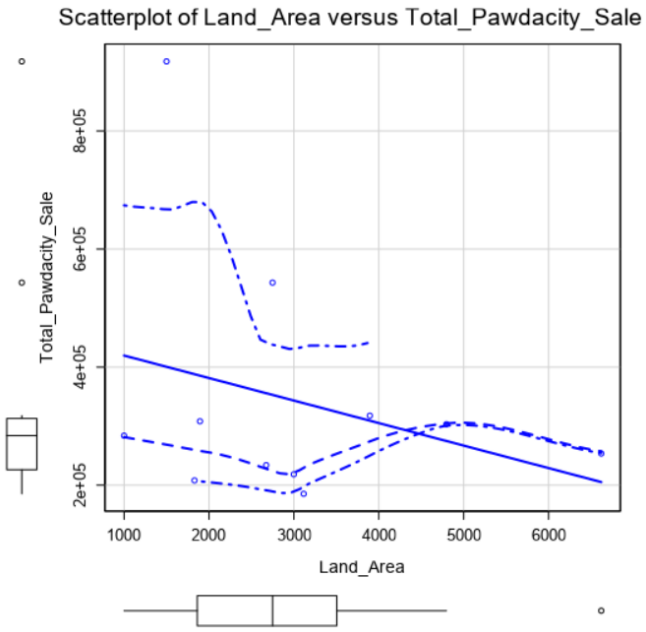
After Completing the building of the training set, the values derived are as follows:

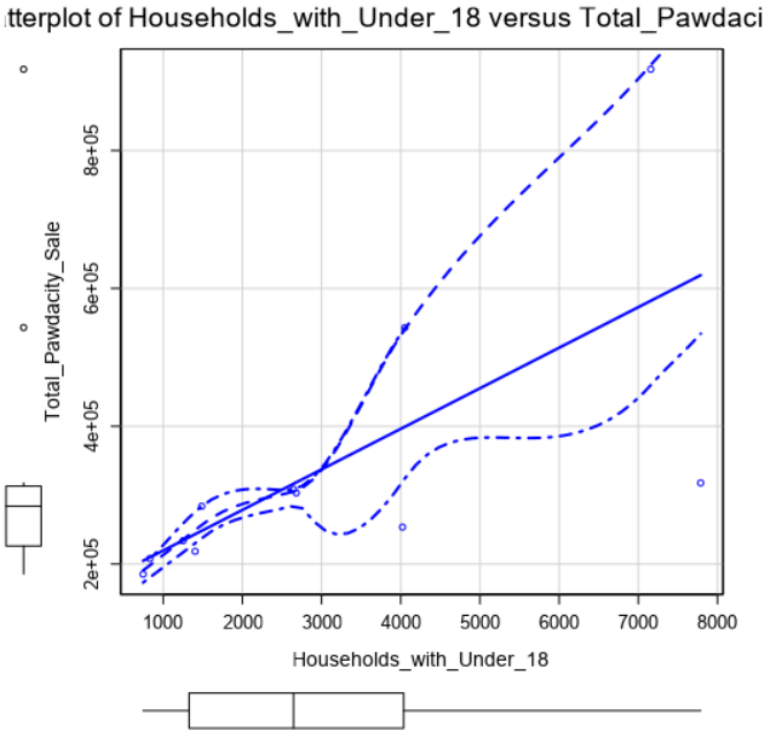
Column	Sum	Average
Census Population	213,862	19442
Total Pawdacity Sales	3,773,304	343,027.63
Households with Under 18	34,064	3,096.72
Land Area	33,071	3,006.48
Population Density	63	5.70
Total Families	62,653	5695.70

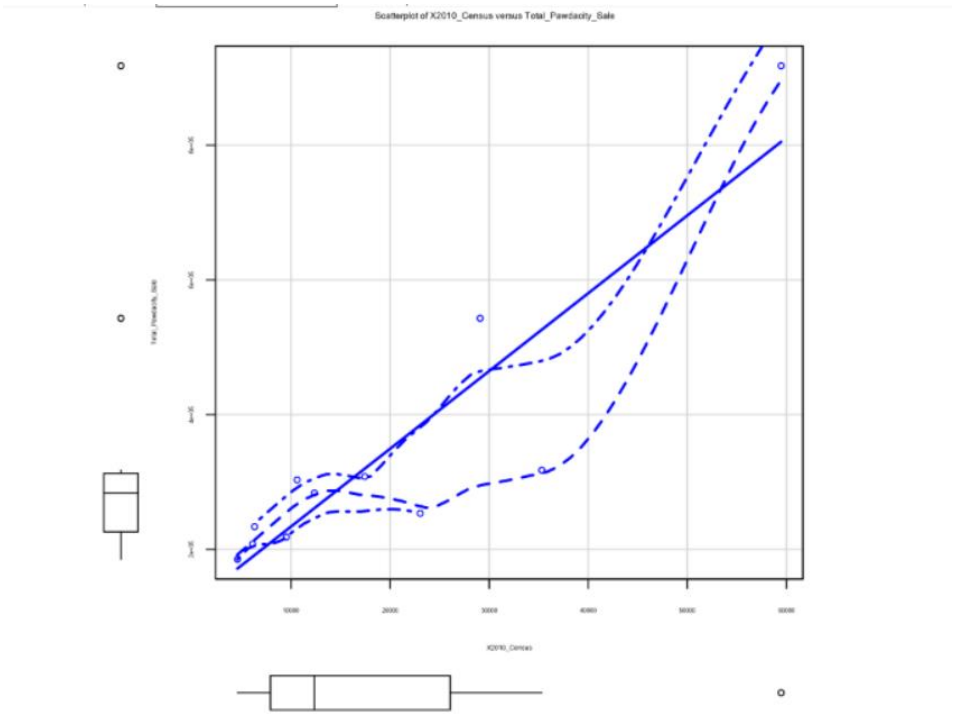


Step 3: Dealing with Outliers

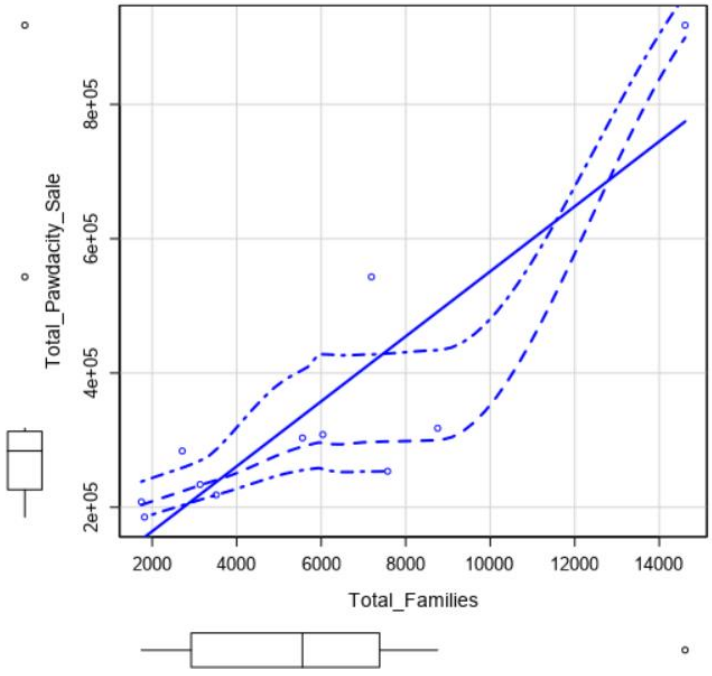
My Dealing with the outliers are furnished with pictorial representation as follows:

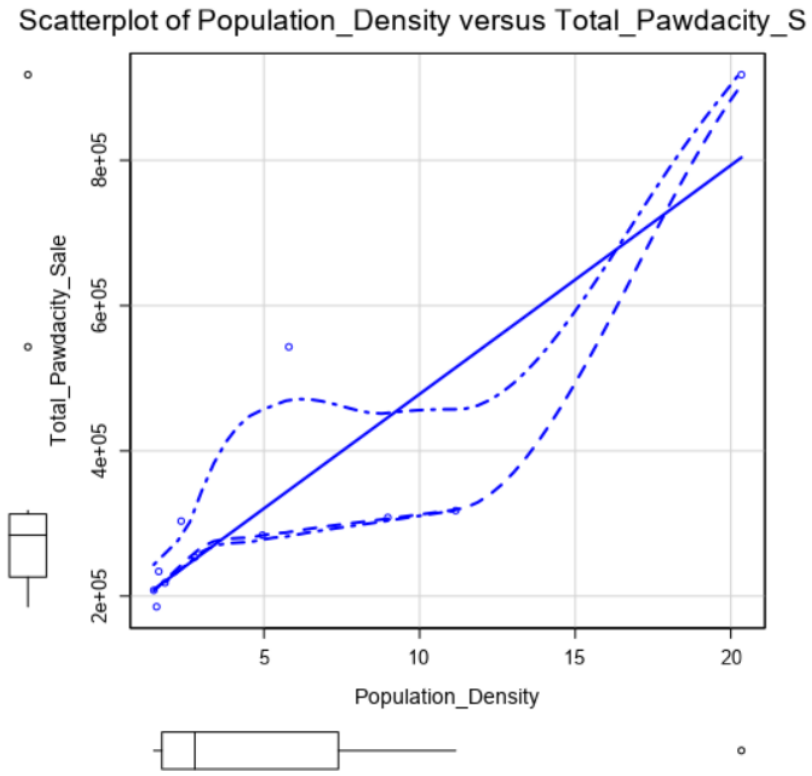






Scatterplot of Total_Families versus Total_Pawdacity_Sale





From the data acquired from scatterplots above, and the data extracted,

There are 3 cities that are outlier in the training set which are : **Cheyenne, Rock Springs and Gillette**

Cheyenne can be **retained** because it can be a big city (the city Cheyenne outlies in total sales , total population , population density and total families).

Gillette can be **removed** which outlies in total sales but have all other things in interquartile range and it also seems abnormal to have high sales with all other variables in proper range .

Rock Springs can be **retained** due to its larger land area **[Big City]** (it outlies in the land area)

The city of **Gillette, Rock springs and Cheyenne** seems to be the **possible outliers** as their sales data are higher than the other cities.

Few points to keep in mind is that these cities also have a higher number of stores and a larger population as well.
(This could be a logical reason as to why these cities have higher sales compared to the rest cities of Wyoming.)

Other factors such as **median income** for the **population of the cities** are **not available**. The dataset given to us is **limited and small** and from the **plot data available**, I believe **Gillette city can be removed.**