Project 2.1: Data Cleanup

Chandan Mishra

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?

   Pawdacity is a leading pet store in Wyoming and has 13 Stores throughout the state. The chain would like to open a 14th store. Based on the yearly sales data, the city for new store needs to be predicted.

   The data needs to be properly formatted and blended together from various datasets:

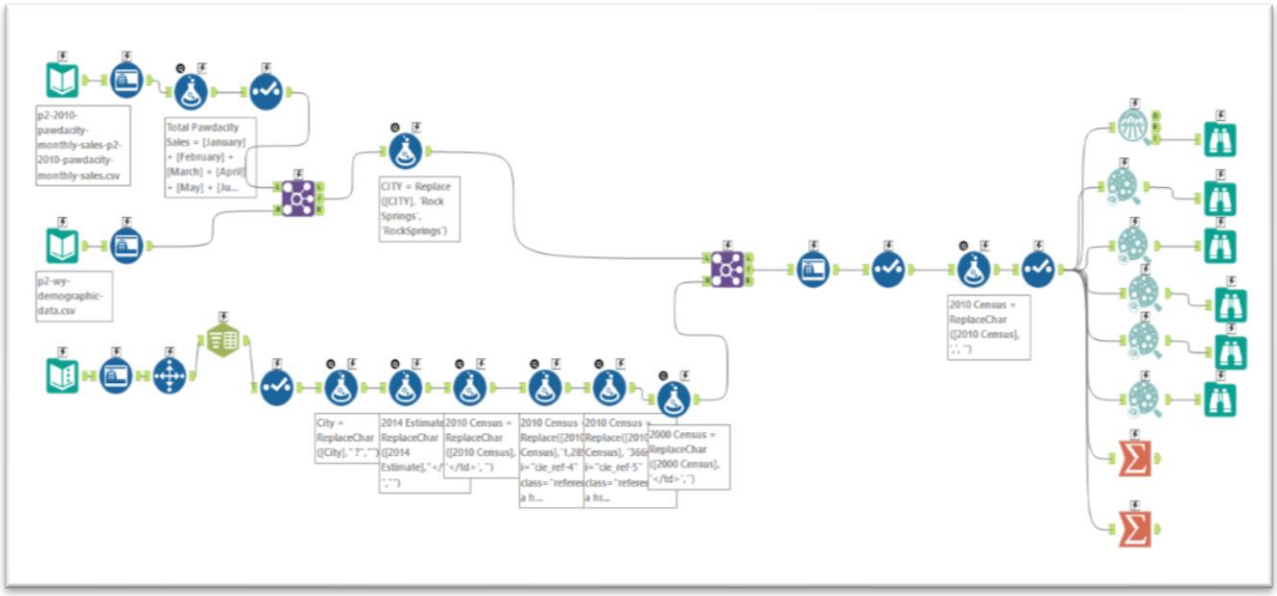
   a) Monthly Sales data for all the Pawdacity stores in Wyoming for the year 2010. Total pawdacity sales needs to be calculated by adding all the monthly sales.

   b) Partially parsed population data. The data needs to be cleaned by removing null values, removing random string characters, and removing spaces.

2. What data is needed to inform those decisions?

   Since the city for new store needs to be predicted, the monthly sales data needs to be converted into annual sales on city level.

   The population data and demographic data needs to be blended with the city where Pawdacity already has the operations.



Step 2: Building the Training Set

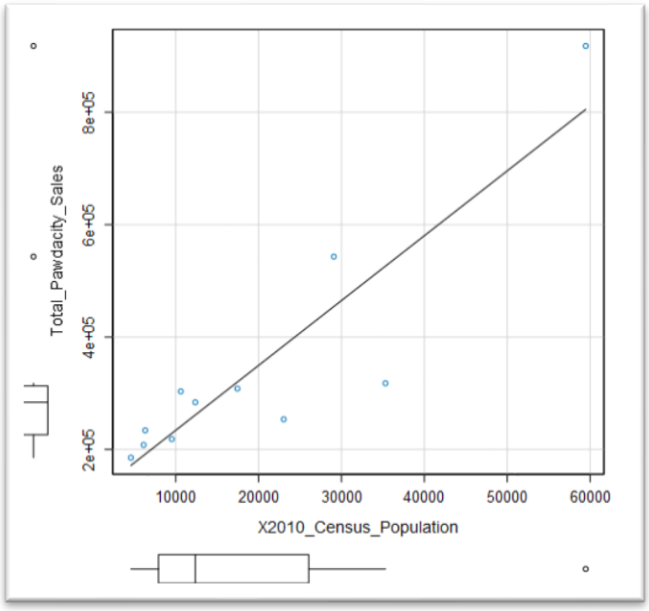
Column	Sum	Average
Census Population	*213,862*	*19442*
Total Pawdacity Sales	*3,773,304*	*343027.64*
Households with Under 18	*34,064*	*3096.72*
Land Area	*33,071*	*3006.49*
Population Density	*63*	*5.71*
Total Families	*62,653*	*5695.71*

Step 3: Dealing with Outliers

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.

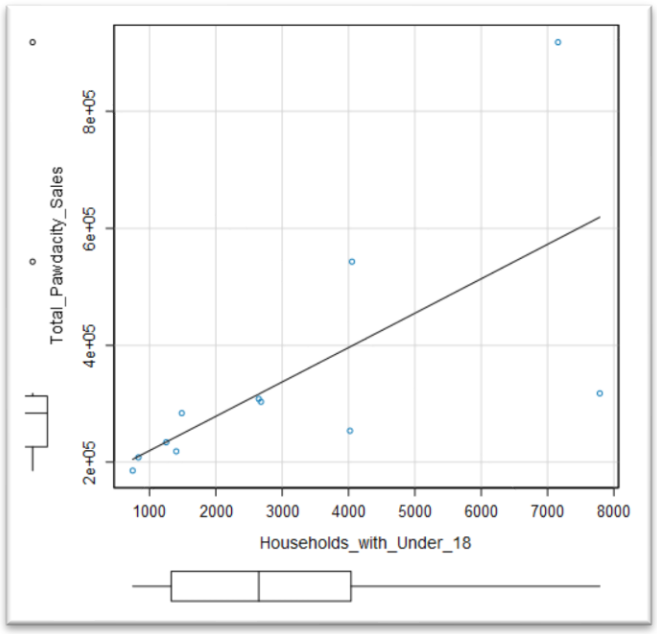
There are outliers present in the dataset.

Census Population vs Total Pawdacity Sales:



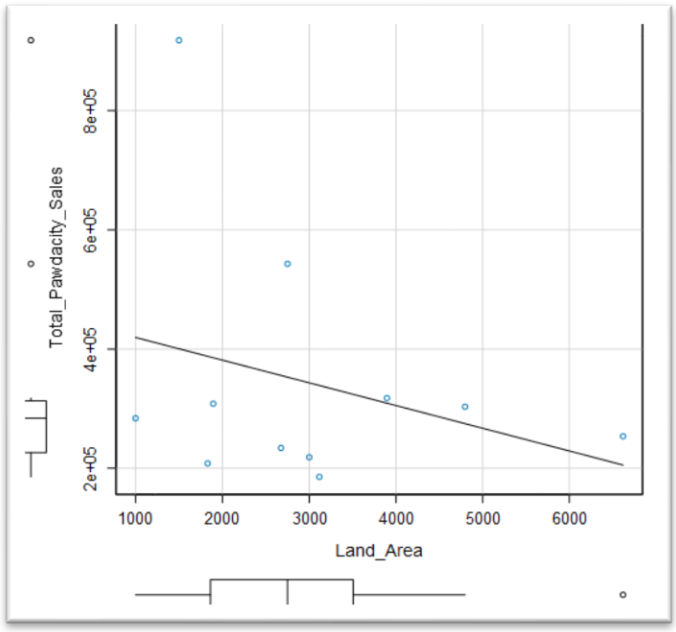
2010 census population has value 59466 for city Cheyenne which is an outlier based on the scatter plot.

Households with under 18 vs Total Pawdacity Sales:



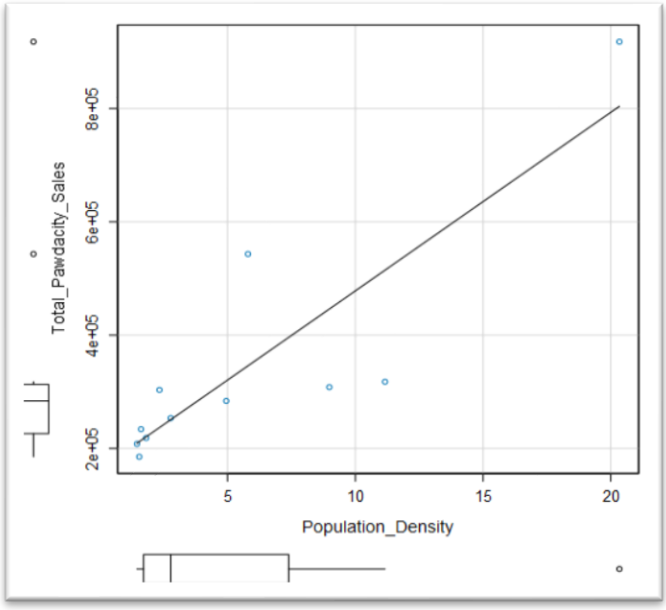
There is no outlier present in this scatter plot.

Land Area vs Total Pawdacity Sales:



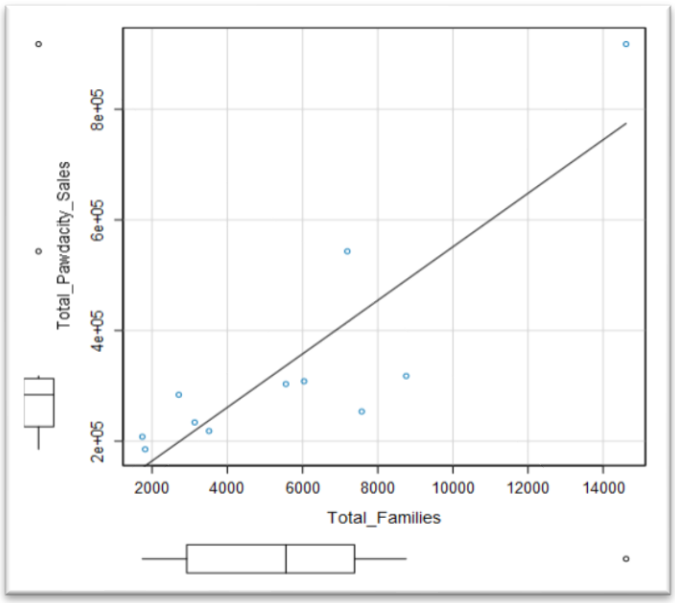
The land area for Rock Springs of 6620 is an outlier here.

Population density vs Total Pawdacity Sales:



Population density having value 20 for city Cheyenne is being shown as an outlier.

Total Families vs Total Pawdacity Sales:



Total families having value 14613 for city Cheyenne is an outlier in total families.

Total Pawdacity Sales vs City: Chenyenne has total sales of $917,892 while Gillette has total sales of $543,132. Gillette's population is 29087. Both are outliers. In Gillette, the total pawdacity sales is only

outlier. While in Cheyenne, four of the five predictor variables are an outlier which might mean it's a big city and no need to remove it from the prediction.

Since the city can be big population concentration wise and can accommodate large number of people in a small area, we can't drop Cheyenne. Since Gillette has only one outlier in the sales, we would need to impute total pawdacity sales for this city.