

Creating Analytical Dataset

Step 1: Business and Data Understanding

What decisions need to be made?

This year, Pawdacity would like to expand and open a 14th store. We have to recommend the best city for sales of Pawdacity's new store. For it, we are supposed to clean the data and format it. So that, we are able to predict the yearly sales.

What data is needed to inform those decisions?

We are working on these three datasets:

p2-2010-pawdacity-monthly-sales.csv - This file contains all of the monthly sales for all Pawdacity stores for 2010.

p2-partially-parsed-wy-web-scrape.csv - This is a partially parsed data file that can be used for population numbers.

p2-wy-demographic-data.csv - This file contains demographic data for each city and county in Wyoming.

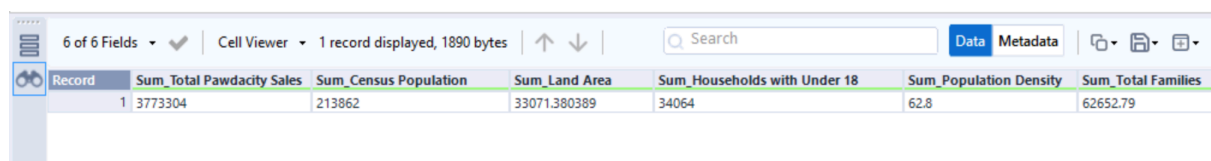
We will need seven columns to perform our final analysis:

- Census Population
- Total Pawdacity Sales
- Households with Under 18
- Land Area
- Population Density
- Total Families

Step 2: Building the Training Set

Summarising the dataset:

Sum of the predictor variables after data wrangling and joins:



The screenshot shows a data viewer interface with a toolbar at the top containing icons for list, search, and other functions. Below the toolbar, a table displays one record of summarized data. The table has seven columns: Record, Sum_Total Pawdacity Sales, Sum_Census Population, Sum_Land Area, Sum_Households with Under 18, Sum_Population Density, and Sum_Total Families. The first row shows the values for record 1.

Record	Sum_Total Pawdacity Sales	Sum_Census Population	Sum_Land Area	Sum_Households with Under 18	Sum_Population Density	Sum_Total Families
1	3773304	213862	33071.380389	34064	62.8	62652.79

Summary of the dataset:

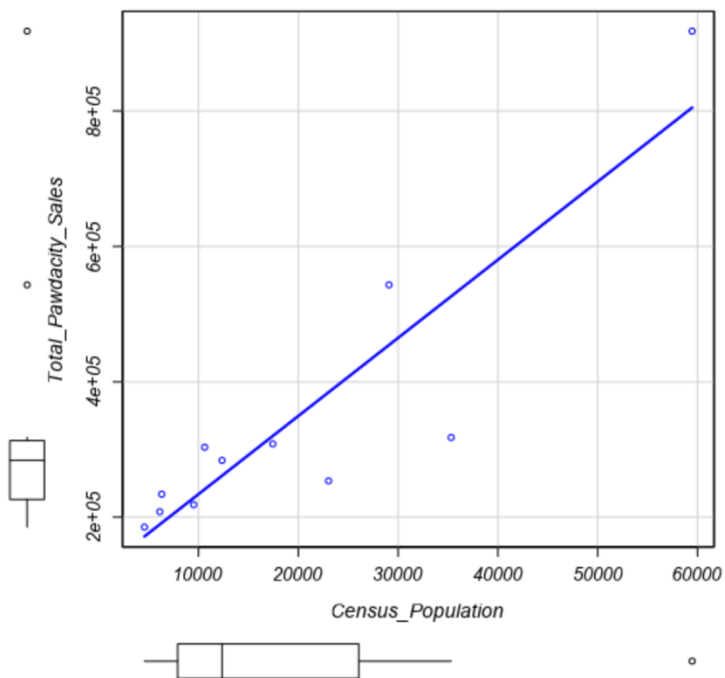
Record	Name	Min	Max	Median	Std. Dev.	Mean
1	Census Population	4585	59466	12359	16616.02	19442
2	Households with Under 18	746	7788	2646	2453.003	3096.727
3	Land Area	999.4971	6620.201916	2748.8529	1617.46	3006.489
4	Population Density	1.46	20.34	2.78	5.849685	5.709091
5	Total Families	1744.08	14612.64	5556.49	3816.05	5695.708
6	Total Pawdacity Sales	185328	917892	283824	213538.7	343027.6

Step 3: Dealing with Outliers

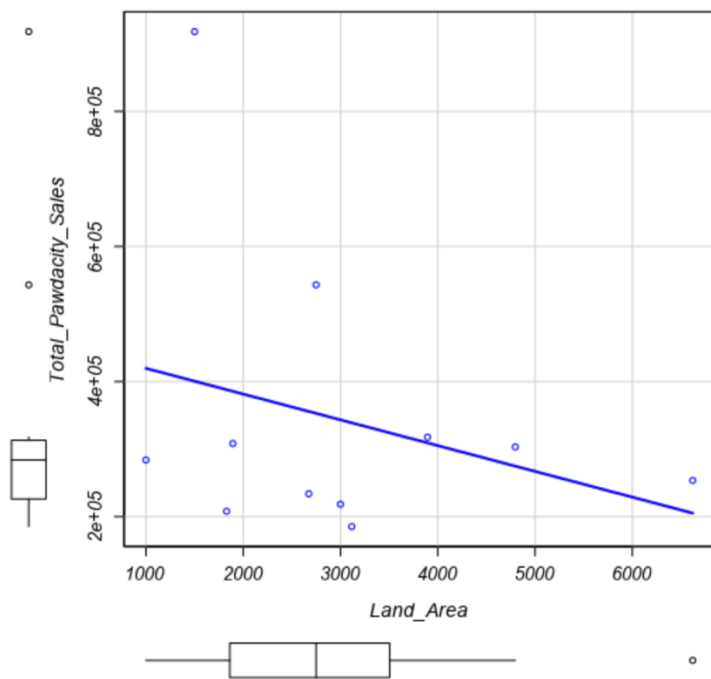
Outliers

Scatterplots of possible predictor values vs target variable (Total_Pawdacity_Sales) tells the story of outliers and which predictor variables can be used in further analysis.

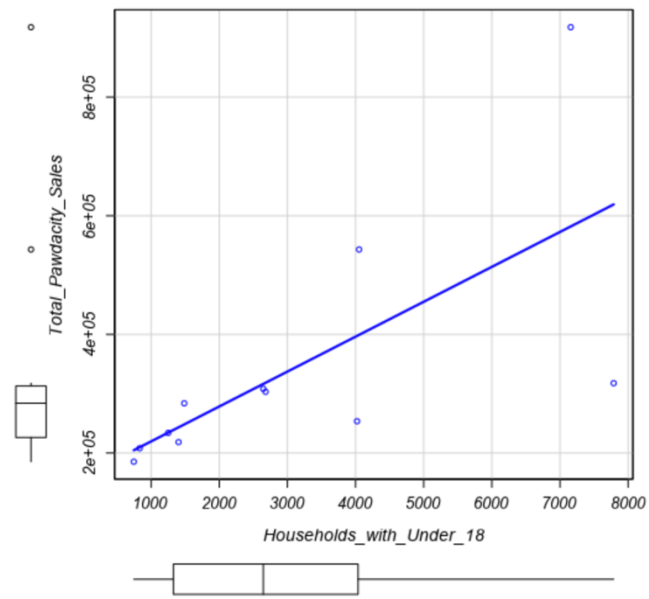
Scatterplot of Census_Population versus Total_Pawdacity_Sales



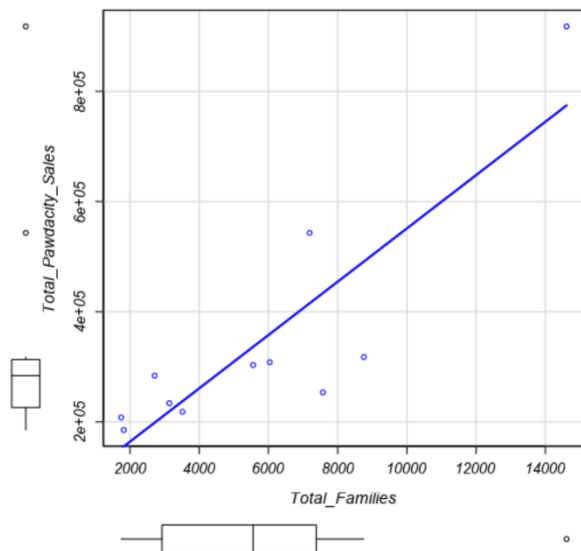
Scatterplot of Land_Area versus Total_Pawdacity_Sales



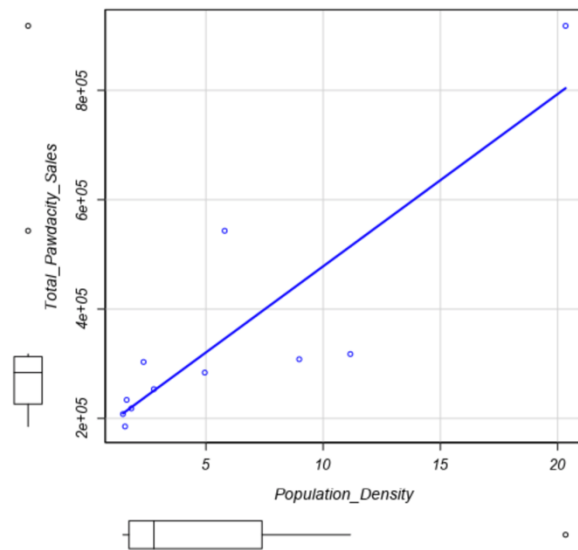
Scatterplot of Households_with_Under_18 versus Total_Pawdacity_Sales



Scatterplot of Total_Families versus Total_Pawdacity_Sale



Scatterplot of Population_Density versus Total_Pawdacity_Sales



3

Calculated Upper Fence for all the variables:

Pawdacity_Sales_Upper_Fence	Census_Population_Upper_Fence	Land_Area_Upper_Fence	Total_Families_Upper_Fence	Households_with_under_18_Upper_Fence	Population_Density_Upper_Fence
443232	53278.25	5969.689139	14066.8975	8102	15.895

Pawdacity_Sales	Census_Population	Land_Area	Total_Families	Households_with_Under_18	Pop_Density
443232	53278.25	5969.68	14066.89	8102	15.89

Cities like Cheyenne, Rock Springs and Gillette did have some outliers. Gillette is having "Pawdacity_Sales" outlier while Cheyenne shows outliers in "Pawdacity_Sales", "Census_Population", "Population_Density", and "Total_Families".

Cheyenne is highly populated and it is understandable why it has sky-rocketing sales. On the other hand, Gillette has only its sales data outside the upper fence while other datapoints are under the outlier range.

So, Gillette is to be outcasted from the dataset.