Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

1. What decisions needs to be made?

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. So, they need to be performing an analysis to recommend the city for Pawdacity's newest store, based on predicted yearly sales.

2. What data is needed to inform those decisions?

The data needed in this business problem are:

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

Step 2: Building the Training Set

After cleaning up and blending all data together, below is the result of training dataset:

7 of 7 Fields | Cell Viewer | 11 records displayed | Data | Metadata

Record #	City	2010 Census_Population	Padacity_Sales	Household_with_Under_18	Land_Area	Population_Density	Total_Families
1	Buffalo	4585	185328	746	3115.5075	1.55	1819.5
2	Casper	35316	317736	7788	3894.3091	11.16	8756.32
3	Cheyenne	59466	917892	7158	1500.1784	20.34	14612.64
4	Cody	9520	218376	1403	2998.95696	1.82	3515.62
5	Douglas	6120	208008	832	1829.4651	1.46	1744.08
6	Evanston	12359	283824	1486	999.4971	4.95	2712.64
7	Gillette	29087	543132	4052	2748.8529	5.8	7189.43
8	Powell	6314	233928	1251	2673.57455	1.62	3134.18
9	Riverton	10615	303264	2680	4796.859815	2.34	5556.49
10	Rock Springs	23036	253584	4022	6620.201916	2.78	7572.18
11	Sheridan	17444	308232	2646	1893.977048	8.98	6039.71

Figure 1: The Training Dataset

The sum of all records matches the sum provided:

Recor...	Sum_2010 Cens...	Sum_Padacity_Sales	Sum_Househo...	Sum_Land_Area	Sum_Pop...	Sum_Total_Families
1	213862	3773304	34064	33071	63	62653

Figure 2: Fields Totals

Dataset SUMs and AVERAGEs:

Column	Sum	Average
Census Population	*213,862*	**19,442.00**
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**

Table 1:Fields Average Values

Step 3: Dealing with Outliers

To identify outliers, we should calculate the upper fence and the lower fence values before. The table below contain these values for each field:

	Total_Sales	Census_ Population	Household_with_Under_18	Land_Area	Population_Density	Total_Families
upper fence	443232	53278.25	8102	5969.689	15.895	14066.8975
lower fence	95904	-19299.75	-2738	-603.06	-6.785	-3762.6825

Table 2: Upper Fence and Lower Fence Values

After the outliers' values were determined and compared to the records, we found six outliers values. The outliers found are: **Pawdacity Sales for Gillette and Cheyenne, Land Area for Rock Springs, Census Population, Population Density, and Total Families for Cheyenne.**

City	Total_Sales	Census_ Population	Household_with _Under_18	Land_Area	Population_Density	Total_Families
Cheyenne	917892	59466	7158	1500.178	20.34	14612.64
Gillette	543132	29087	4052	2748.853	5.8	7189.43
Rock Springs	253584	23036	4022	6620.202	2.78	7572.18

Table 3: Dataset Outliers

As tables shown above we found three cities with an outliers data, the cities are: **Cheyenne, Gillette, and Rock Springs.** To illustrate the effect of these outliers data on the dataset, we have included these scatter plots depending on the **Pawdacity total sales** as target variable.

Cheyenne appears always as the highest point at all pivots. **Cheyenne** has high effects on **dataset** because of its high values, we think removing a valuable record like Cheyenne will effects our analysis. **Rock Springs** had only one **outlier** (6620.2), which is not appears abnormal comparing with the upper fence (5969.68) of land_area, they are close. We prefer to keep Cheyenne and Rock Springs. And removing **Gillette** record because it appears at the periphery of the pivot. So, there wouldn't be a significant impact on predictive model of the **dataset** if it has removed.

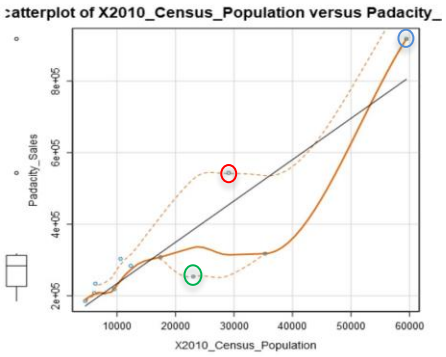

Figure 3: Census_Population VS Padacit_Sales

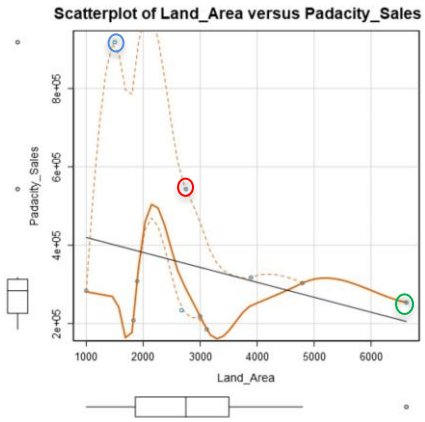

Figure 4: Land_Area VS Padacit_Sales

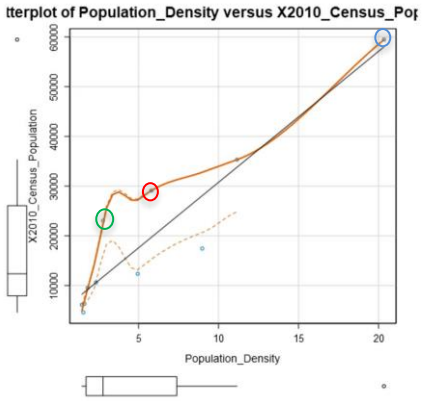

Figure 6:Population_density VS Padacit_Sales

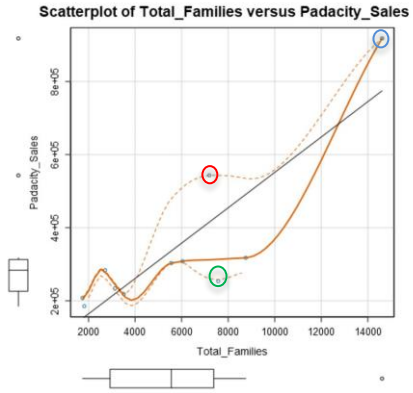

Figure 5:Total_Families VS Padacit_Sales

- **Scatter Plots represent the impact below:**

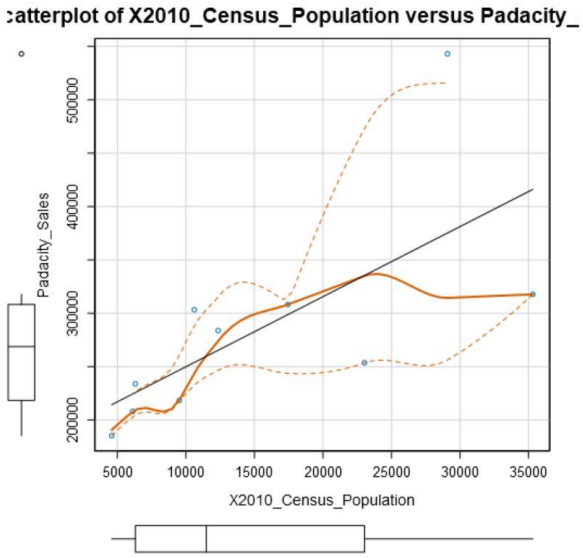

Figure 7: Census_Population VS Padacity_Sales after removing Cheyenne

once we removed the value of Cheyenne, there was a significant negative setback in the pivot, leading to the impact of the predictive results needed by Padacity.

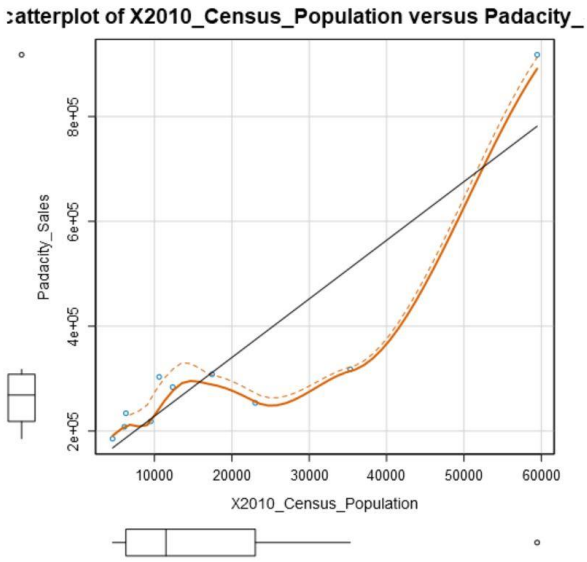

Figure 8: Census_Population VS Padacit_Sales after removing Gillette

There wouldn't be a significant change or impact. So, we think it is the correct choice to be removed as an outlier.