

## Project 2.1: Data Cleanup

### Step 1: Business and Data Understanding

The decision that needs to be made is to determine which city is the most suitable for new 14<sup>th</sup> Pawdacity pet store based on predicted yearly sales.

#### Key Decisions:

1. What decisions needs to be made?

Which city is the most suitable for new 14<sup>th</sup> Pawdacity pet store based on predicted yearly sales.

2. What data is needed to inform those decisions?

Predicted sales for the new store. The data provided to predict sales:

- 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

*Build your training set given the data provided to you. Your column sums of your dataset should match the sums in the table below.*

*In addition provide the averages on your data set here to help reviewers check your work. You should round up to two decimal places, ex: 1.24*

Column	Sum	Average
Census Population	213,862	19442
Total Pawdacity Sales	3,773,304	343027.64
Households with Under 18	34,064	3096.73
Land Area	33,071	3006.45
Population Density	63	5.73
Total Families	62,653	5695.73

### 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.

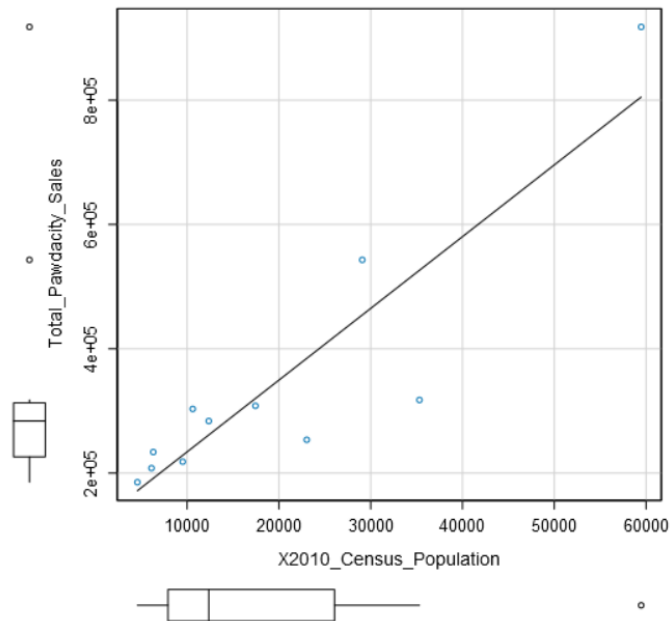
I would remove the data for the city **Gillette** as it is the outlier for Total Sales. Even though Cheyenne is the outlier for Census Population, Total Sales, Population Density, and Total

Families predictor variables. It seems like Cheyenne is a bigger and denser city with more people living compare to other cities. Rock Springs is the outlier for Land Area, but I will keep it as it is in line with the trend. Please see below for scatterplots and analysis using IQR steps.

City		2010 Census Population	Total Pawdacity Sales	Households with Under 18	Land Area	Population Density	Total Families
	Buffalo	4585	185328	746	3116	2	1820
	Casper	35316	317736	7788	3894	11	8756
	Cheyenne	59466	917892	7158	1500	20	14613
	Cody	9520	218376	1403	2999	2	3516
	Douglas	6120	208008	832	1829	1	1744
	Evanston	12359	283824	1486	999	5	2713
	Gillette	29087	543132	4052	2749	6	7189
	Powell	6314	233928	1251	2674	2	3134
	Riverton	10615	303264	2680	4797	2	5556
	Rock Springs	23036	253584	4022	6620	3	7572
	Sheridan	17444	308232	2646	1894	9	6040
Q <sub>1</sub>	25th Percentile	7917	226152	1327	1861.5	2	2923.5
Q <sub>3</sub>	75th Percentile	26061.5	312984	4037	3505	7.5	7380.5
IQR	Q <sub>3</sub> - Q <sub>1</sub>	18144.5	86832	2710	1643.5	5.5	4457
Lower	Q <sub>1</sub> - 1.5×IQR	-19299.75	95904	-2738	-603.75	-6.25	-3762
Upper	Q <sub>3</sub> + 1.5×IQR	53278.25	443232	8102	5970.25	15.75	14066

2010 Census Population

Scatterplot of X2010\_Census\_Population versus Total\_Pawdacity\_Sales

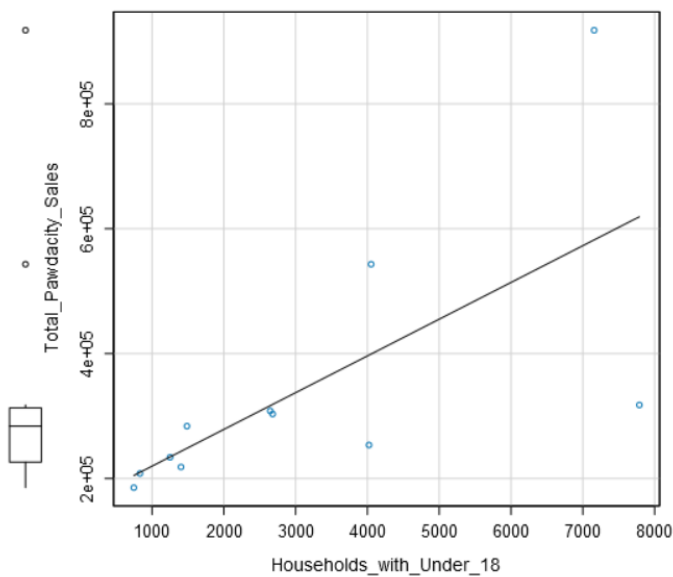


Lower  $Q_1 - 1.5 \times IQR$  -19299.8  
 Upper  $Q_3 + 1.5 \times IQR$  53278.25

The city Cheyenne has 2010 Census Population of 59466 which is outside of upper fence. The data is probably right, and Cheyenne is probably bigger city with more people. This is also outlier for Total Sales. This makes sense since I'd expect the relationship to behave this way.

### Households with Under 18

Scatterplot of Households\_with\_Under\_18 versus Total\_Pawdacity\_Sales



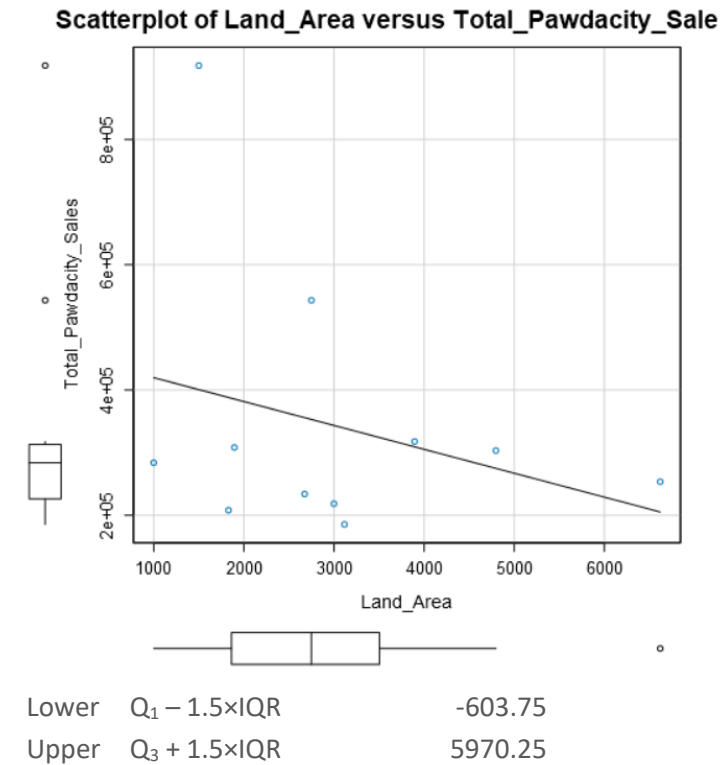
Lower  $Q_1 - 1.5 \times IQR$  -2738

Upper  $Q_3 + 1.5 \times IQR$

8102

Everything is between lower fence and upper fence. There doesn't seem to be any outliers for Households with Under 18 predictor variables.

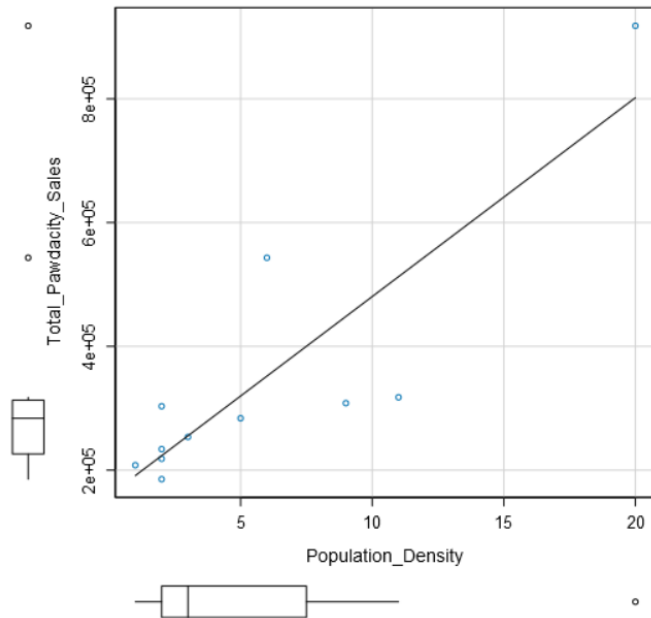
### Land Area



Rock Springs has a value 6620 which is above upper fence. But it seems to be in line with the trend, and not dramatically different.

## Population Density

Scatterplot of Population\_Density versus Total\_Pawdacity\_Sales



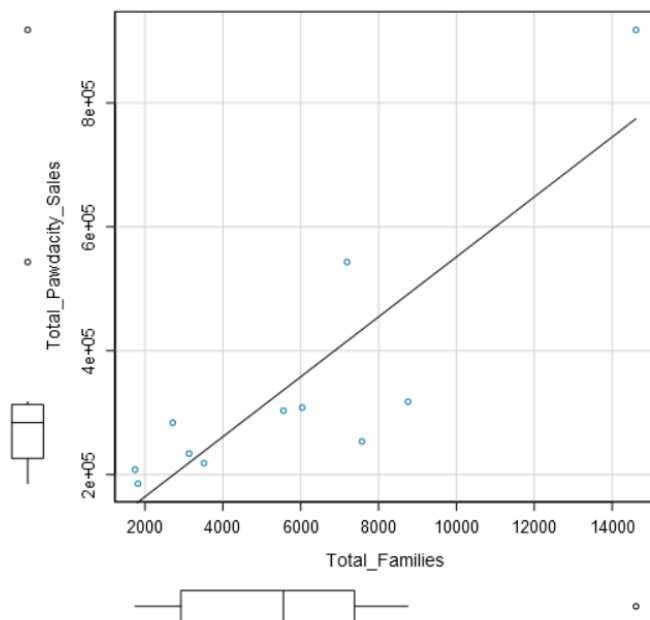
Lower  $Q_1 - 1.5 \times IQR$  -6.25

Upper  $Q_3 + 1.5 \times IQR$  15.75

The city Cheyenne has population density of 20 which is outside of upper fence.

## Total Families

Scatterplot of Total\_Families versus Total\_Pawdacity\_Sales



Lower  $Q_1 - 1.5 \times IQR$  -3762

Upper  $Q_3 + 1.5 \times IQR$  14066

Again, the city Cheyenne has total families of 14613 which is above upper fence.