Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Pawdacity, a pet store chain in Wyoming, would like to expand and open a 14th store. We have to perform an analysis to recommend the city for Pawdacity's newest store based on predicted yearly sales and data from other different datasets.

Key Decisions:

1. What decisions needs to be made?

The decision that needs to be made is in which city, in Wyoming, the newest Pawdacity's pet store should be opened, based on predicted yearly sales and data from other different datasets.

2. What data is needed to inform those decisions?

In order to recommend the city for Pawdacity's newest store, we require the following data for each city and county in the state of Wyoming:

- All of the Pawdacity stores sales data;
- Competitor stores sales data;
- Demographic data, such as:
    - Population numbers
    - Households with individuals under 18;
    - Land Area;
    - Population Density;
    - Total Families.

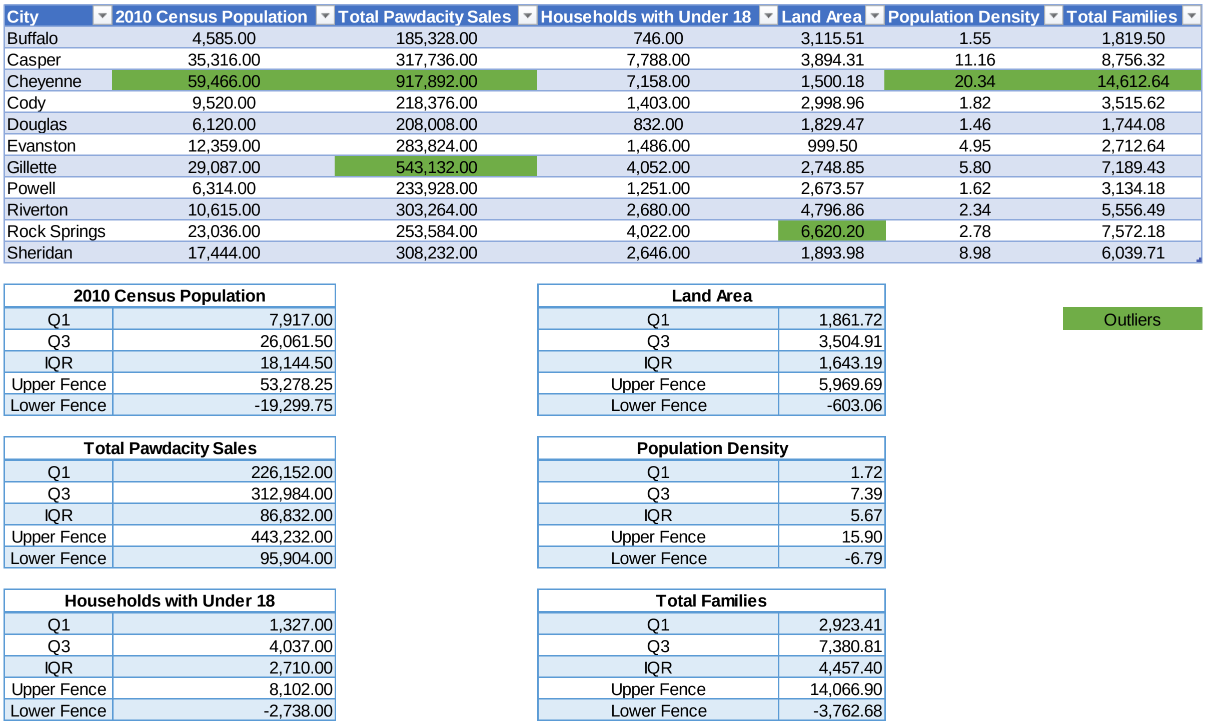
Step 2: Building the Training Set

Column	Sum	Average
Census Population	213,862	19,442
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

Step 3: Dealing with Outliers

Once I have created the dataset, I used the IQR method to determine if there are outlier cities. According to my results there are 3 cities that contain outliers, which are: Cheyenne, Gillette and Rock Springs.

An overview of the results is available in Pic 3.1.



City	2010 Census Population	Total Pawdacity Sales	Households with Under 18	Land Area	Population Density	Total Families
Buffalo	4,585.00	185,328.00	746.00	3,115.51	1.55	1,819.50
Casper	35,316.00	317,736.00	7,788.00	3,894.31	11.16	8,756.32
Cheyenne	59,466.00	917,892.00	7,158.00	1,500.18	20.34	14,612.64
Cody	9,520.00	218,376.00	1,403.00	2,998.96	1.82	3,515.62
Douglas	6,120.00	208,008.00	832.00	1,829.47	1.46	1,744.08
Evanston	12,359.00	283,824.00	1,486.00	999.50	4.95	2,712.64
Gillette	29,087.00	543,132.00	4,052.00	2,748.85	5.80	7,189.43
Powell	6,314.00	233,928.00	1,251.00	2,673.57	1.62	3,134.18
Riverton	10,615.00	303,264.00	2,680.00	4,796.86	2.34	5,556.49
Rock Springs	23,036.00	253,584.00	4,022.00	6,620.20	2.78	7,572.18
Sheridan	17,444.00	308,232.00	2,646.00	1,893.98	8.98	6,039.71

2010 Census Population	
Q1	7,917.00
Q3	26,061.50
IQR	18,144.50
Upper Fence	53,278.25
Lower Fence	-19,299.75

Land Area	
Q1	1,861.72
Q3	3,504.91
IQR	1,643.19
Upper Fence	5,969.69
Lower Fence	-603.06

🟩 Outliers

Total Pawdacity Sales	
Q1	226,152.00
Q3	312,984.00
IQR	86,832.00
Upper Fence	443,232.00
Lower Fence	95,904.00

Population Density	
Q1	1.72
Q3	7.39
IQR	5.67
Upper Fence	15.90
Lower Fence	-6.79

Households with Under 18	
Q1	1,327.00
Q3	4,037.00
IQR	2,710.00
Upper Fence	8,102.00
Lower Fence	-2,738.00

Total Families	
Q1	2,923.41
Q3	7,380.81
IQR	4,457.40
Upper Fence	14,066.90
Lower Fence	-3,762.68

Fig. 3.1. IQR results.

The first outlier city, Cheyenne, is a big city with a high population density, therefore, it is expected that there will be a higher total sales for this city.

The second outlier city, Gillette, appears to be a small city with high sales, which can influence our model.

The third outlier city, Rock Springs, only has a higher land area, but the population density is lower.

In the end, taking all the above into consideration I have chosen to remove the outlier city Gillette because it can impact our model.