Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

1. What decisions need 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. An analysis in required in order 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 from different datasets needs to be formatted and blended together and outliers need to be dealt with.

The following information has been provided to inform the decisions:

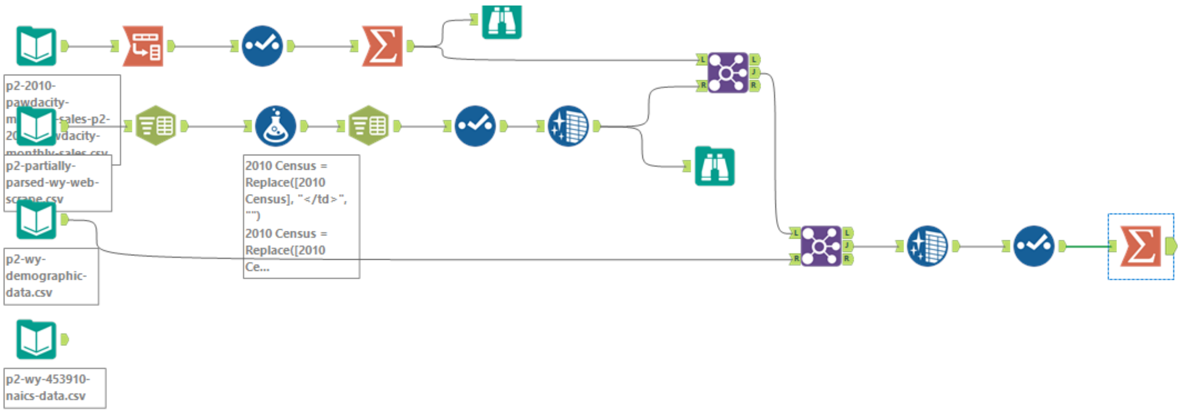
- The monthly sales data for all of the Pawdacity stores for the year 2010.
- NAICS data on the most current sales of all competitor stores where total sales is equal to 12 months of sales.
- A partially parsed data file that can be used for population numbers.
- Demographic data for each city and county in the state of Wyoming.

To build the model and select the appropriate predictor variables, I will create a dataset with the following columns.

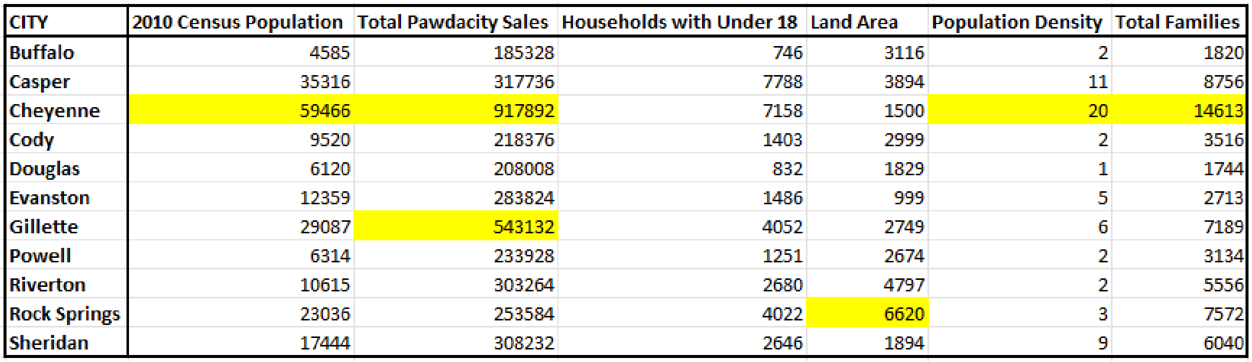
- City
- 2010 Census Population
- Total Pawdacity Sales
- Households with Under 18
- Land Area
- Population Density
- Total Families

Step 2: Building the Training Set

The below workflow was used to build the dataset. I have included my Alteryx workflow to assist with the review.

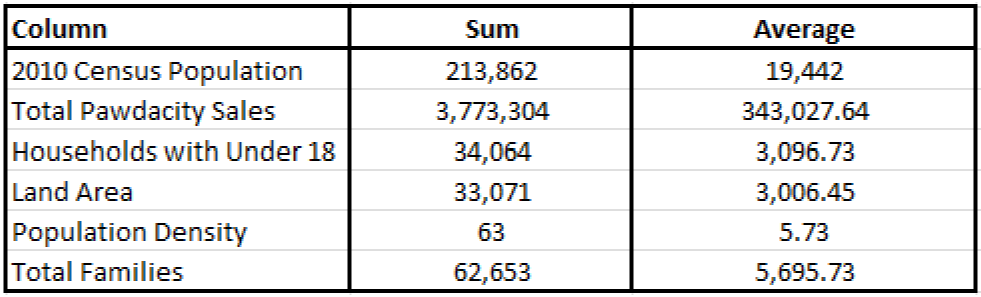

p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales.csv
p2-partially-parsed-wy-web-scrape.csv
2010 Census = Replace([2010 Census], "</td>", "")
2010 Census = Replace([2010 Ce...
p2-wy-demographic-data.csv
p2-wy-453910-naics-data.csv

The below table shows the training set built using the data and model displayed above.



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

The table below shows the sum value of each column and the averages.

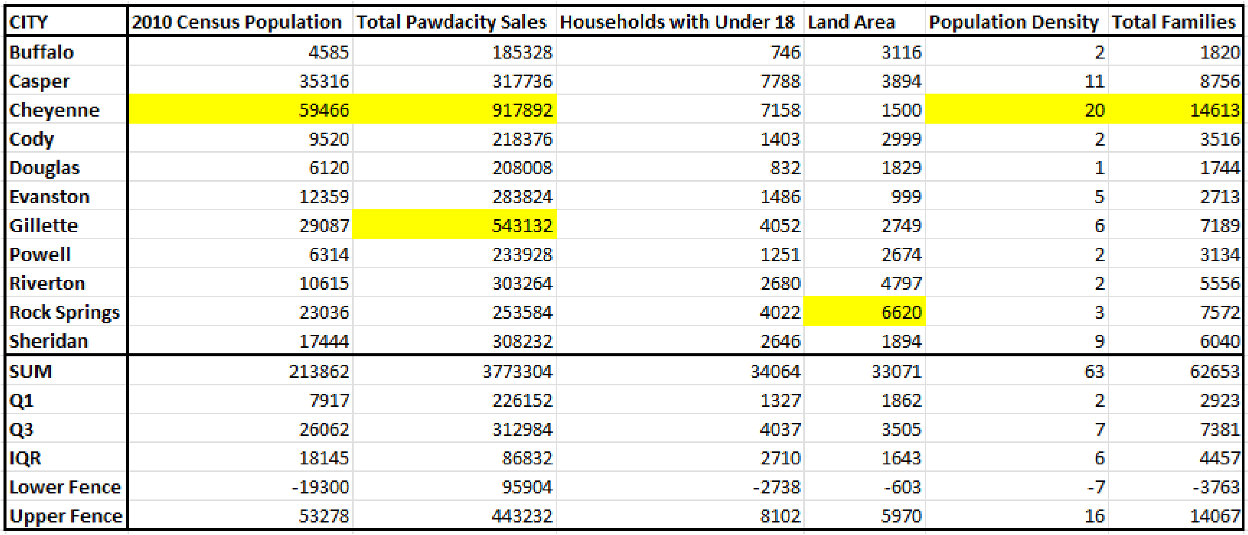


Column	Sum	Average
2010 Census Population	213,862	19,442
Total Pawdacity Sales	3,773,304	343,027.64
Households with Under 18	34,064	3,096.73
Land Area	33,071	3,006.45
Population Density	63	5.73
Total Families	62,653	5,695.73

Step 3: Dealing with Outliers

To deal with the outliers I calculated the upper fence and the lower fence using the following steps:

1 . Calculate 1st quartile Q1 and 3rd quartile Q3 of the dataset. I used the QUARTILE.INC Excel function.
2 . To calculate the Interquartile Range: IQR = Q3 - Q1
3 . Add 1.5 IQR to Q3 to get the upper fence: Upper Fence = Q3 + 1.5 IQR
4 . Subtract 1.5 IQR to Q1 to get the lower fence: Lower Fence = Q1 - 1.5 IQR
5 . Values above the Upper Fence and values below the Lower Fence are outliers.



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
SUM	213862	3773304	34064	33071	63	62653
Q1	7917	226152	1327	1862	2	2923
Q3	26062	312984	4037	3505	7	7381
IQR	18145	86832	2710	1643	6	4457
Lower Fence	-19300	95904	-2738	-603	-7	-3763
Upper Fence	53278	443232	8102	5970	16	14067

As can be seen from the dataset above there are three cities with outliers, Cheyenne, Gillette and Rock Springs.

I chose to remove Cheyenne from the dataset as the city displays outliers in four of the six columns. Compared to the other two cities with outliers, (Gillette and Rock Springs) Cheyenne stands out as the city which should be removed from the dataset. As the dataset created is small with only 11 cities then it is recommended that only one city should be removed.