Project 2: Data Wrangling for new Shop location

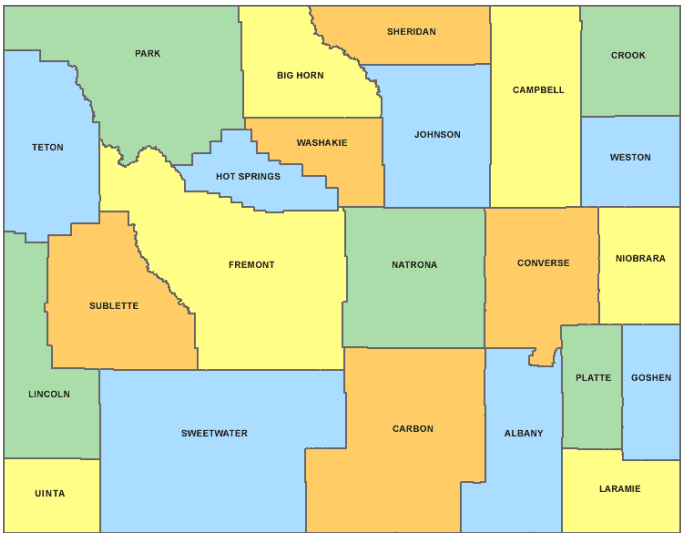
Pawdacity is a leading pet store chain in Wyoming with 13 stores throughout the state. This year, Pawdacity would like to expand and open a 14th store. Your manager has asked you to perform an analysis to recommend the city for Pawdacity's newest store, based on predicted yearly sales.

Your first step in predicting yearly sales is to first format and blend together data from different datasets and deal with outliers.

Your manager has given you the following information to work with:

1. The monthly sales data for all of the Pawdacity stores for the year 2010.
2. NAICS data on the most current sales of all competitor stores where total sales is equal to 12 months of sales.
3. A partially parsed data file that can be used for population numbers.
4. Demographic data (Households with individuals under 18, Land Area, Population Density, and Total Families) for each city and county in the state of Wyoming. For people who are unfamiliar with the US city system, a state contains counties and counties contains one or more cities.

Map of Wyoming Counties



Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?

A proposal for the location of the newest store of Pawdacity, based on the yearly sales statistics.

2. What data is needed to inform those decisions?

- Monthly sales of Pawdacity for year 2010
- Demographic data for the cities in Wyoming.

Step 2: Building the Training Set

The training set was built the following way:
P2-wy-demographic data was the basis for calculationg Population for each city, based on Land Area and Population density. Then the Partially parsed web scraped data was cleaned, removing not needed characters with a series of regexp cleanup, null checks and data enrichment for missing county information. Then, the pawdacity yearly sales numbers were calculated based on the monthly sales statistics. All these 3 inputs were Joined together, thus I could calculate the sums and averages for the fields above.

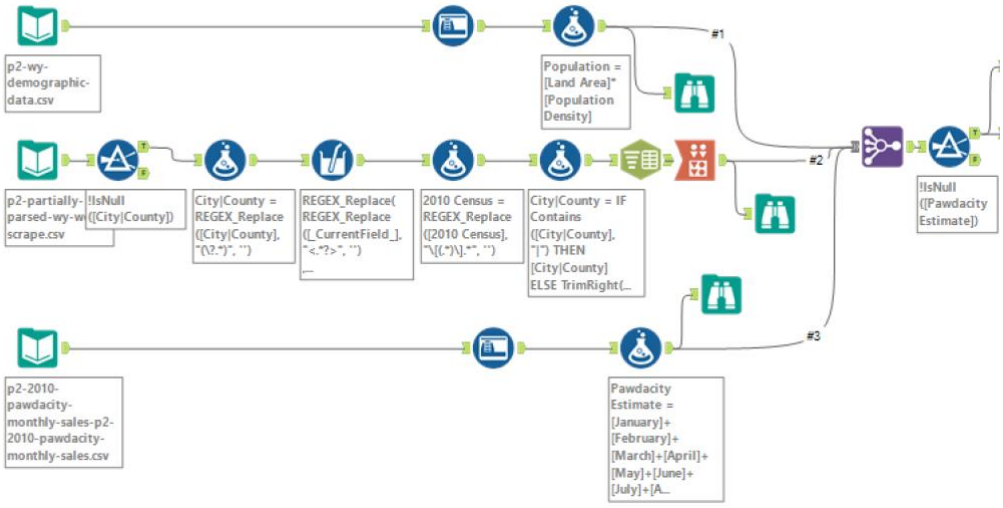


Figure 1 – The training set

The calculated averages were checked based on the output in the red circle on Figure 3 and the results are in Figure 2.

Record	Name	Average
1	2010 Census	19442
2	Pawdacity Estimate	343027.636364
3	Households with Under 18	3096.727273
4	Land Area	3006.489126
5	Population Density	5.709091
6	Total Families	5695.708182

Figure 2 - The calculated averages for the predictor variables

Step 3: Dealing with Outliers

For calculation IQR Interquartile range the Basic data profile information was extracted to get Q1 and Q3 quartile values for the dataset, and then IQR = Q3 - Q1, Upper Fence = Q3 + 1.5 IQR and Lower Fence = Q1 - 1.5 IQR calculation have been applied, see Figure 3 blue circle and Figure 4 numerical results

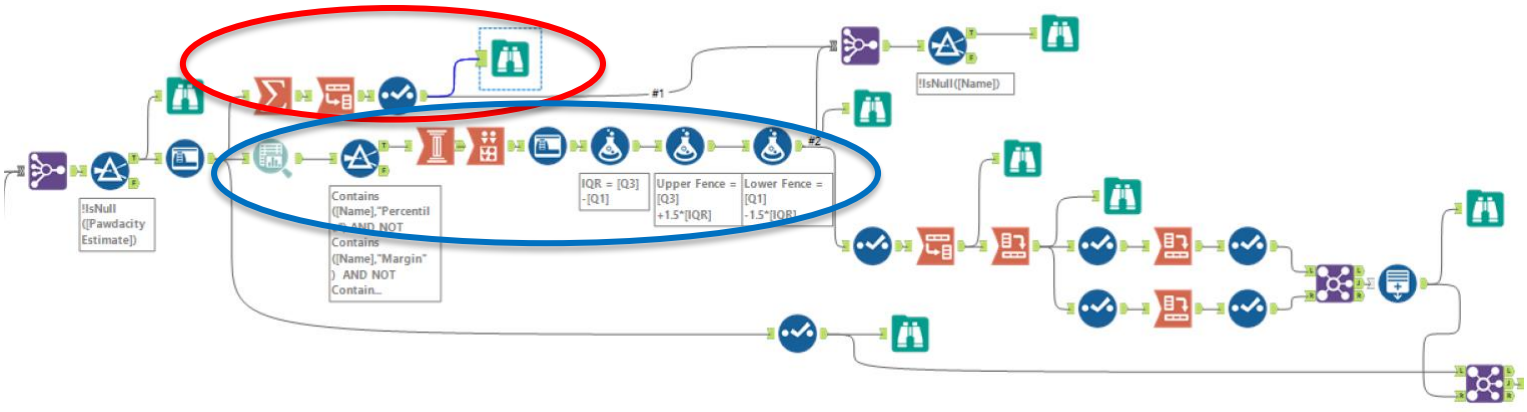


Figure 3 – Extracting the basic statistics of predictor values and calculating IQR values

Record	Name	Sum	FieldName	Q1	Q3	IQR	Upper Fence	Lower Fence
1	2010 Census	213862	2010 Census	7917	26061.5	18144.5	53278.25	-19299.75
2	Households with Under 18	34064	Households with Under 18	1327	4037	2710	8102	-2738
3	Land Area	33071.380389	Land Area	1861.721074	3504.9083	1643.187226	5969.689139	-603.059765
4	Pawdacity Estimate	3773304	Pawdacity Estimate	226152	312984	86832	443232	95904
5	Population Density	62.8	Population Density	1.72	7.39	5.67	15.895	-6.785
6	Total Families	62652.79	Total Families	2923.41	7380.805	4457.395	14066.8975	-3762.6825

Figure 4 – The IQR numerical results

Based on the actual values of the predictor variables for the cities (Figure 5) and the IQR Upper and Lower Fence limits (Figure 4), one can implement Alteryx checks for these values (Figure 6).

Record	City	Land Area	Households with Under 18	Population Density	Total Families	Population	2010 Census	Pawdacity Estimate
1	Buffalo	3115.5075	746	1.55	1819.5	4829.036625	4585	185328
2	Casper	3894.3091	7788	11.16	8756.32	43460.489556	35316	317736
3	Cheyenne	1500.1784	7158	20.34	14612.64	30513.628656	59466	917892
4	Cody	2998.95696	1403	1.82	3515.62	5458.101667	9520	218376
5	Douglas	1829.4651	832	1.46	1744.08	2671.019046	6120	208008
6	Evanston	999.4971	1486	4.95	2712.64	4947.510645	12359	283824
7	Gillette	2748.8529	4052	5.8	7189.43	15943.34682	29087	543132
8	Powell	2673.57455	1251	1.62	3134.18	4331.190771	6314	233928
9	Riverton	4796.859815	2680	2.34	5556.49	11224.651967	10615	303264
10	Rock Springs	6620.201916	4022	2.78	7572.18	18404.161326	23036	253584
11	Sheridan	1893.977048	2646	8.98	6039.71	17007.913891	17444	308232

Figure 5 – The actual values of predictor variables

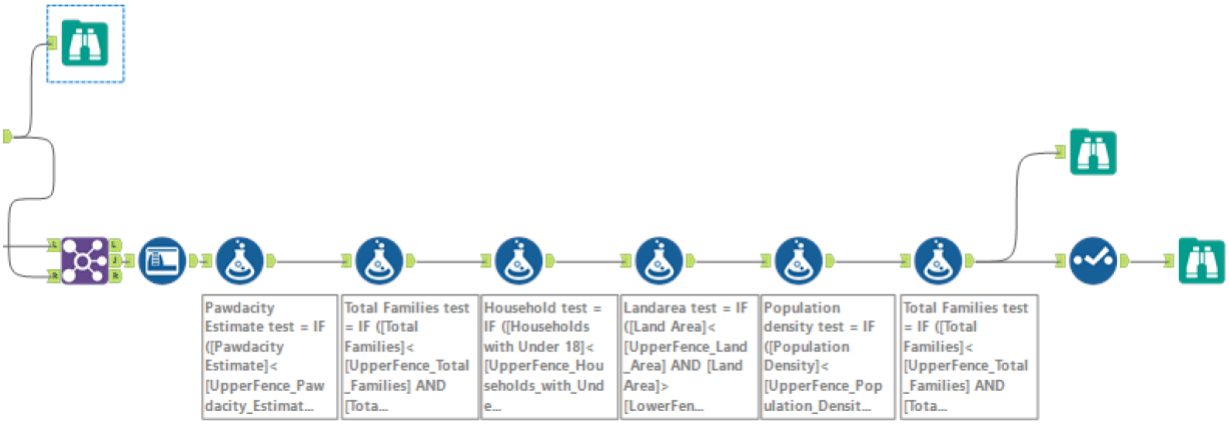


Figure 6 – Alteryx checks for the IQR fences

When checking the output of the predictor value checks against the IQR fences (see Figure 7), one can see three cities from outlier results, Cheyenne has 3 outlier values, while Gillette and Rock Spring have one. Since the dataset is quite small and Cheyenne has 3 outliers:

- 917892 Pawdacity Estimate (see Figure 5) exceeding the upper IQR limit of 443232 (see Figure 4)
- 14613 total family number (Figure 5) exceeding the upper IQR limit of 14067 (see Figure 4)
- 20.34 Population density (see Figure 5) exceeding the upper IQR limit of 20.9 (see Figure 4)

therefore I would recommend to remove **Cheyenne** from the dataset, as it is most likely that there has been errors in collecting the data.

Record	City	Pawdacity Estimate test	Total Families test	Household test	Landarea test	Population density test
1	Buffalo	OK	OK	OK	OK	OK
2	Casper	OK	OK	OK	OK	OK
3	Cheyenne	NOT OK	NOT OK	OK	OK	NOT OK
4	Cody	OK	OK	OK	OK	OK
5	Douglas	OK	OK	OK	OK	OK
6	Evanston	OK	OK	OK	OK	OK
7	Gillette	NOT OK	OK	OK	OK	OK
8	Powell	OK	OK	OK	OK	OK
9	Riverton	OK	OK	OK	OK	OK
10	Rock Springs	OK	OK	OK	NOT OK	OK
11	Sheridan	OK	OK	OK	OK	OK

Figure 7 – The results of the predictor values against the IQR Upper and Lower Limits