

PROJECT

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (250 word limit)

Key Decisions:

Answer these questions

1. What decisions needs to be made?

Business Objective :- 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. Perform 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?

Data Understanding :- The data is distributed in four csv files.

1. The monthly sales data for all 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.

Files 1,2,3 will be used for analysis a data preparation.

Data Preparation :- The files are first input into Alteryx. We then solve data issues and format the data.

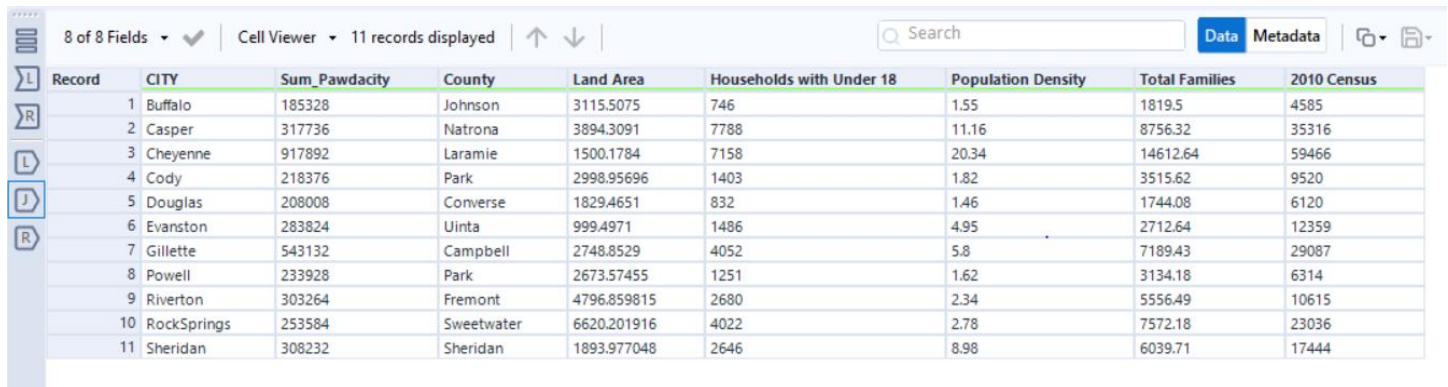
1. Check the data types of each dataset and convert the ones required.
2. Format and clean data using Alteryx *Data Preparation* Tools.
3. Blend the dataset using join
4. Check for outliers using *outlier test* in excel and *scatter plots* in alteryx.
5. Select Predictor variables for training the model
6. Train the Model
7. Output Results

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.

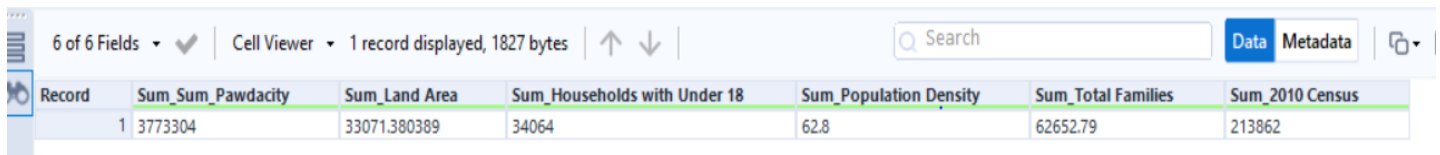
Building the Training Set

Averages(Values for all the store location based on city)



Record	CITY	Sum_Pawdacity	County	Land Area	Households with Under 18	Population Density	Total Families	2010 Census
1	Buffalo	185328	Johnson	3115.5075	746	1.55	1819.5	4585
2	Casper	317736	Natrona	3894.3091	7788	11.16	8756.32	35316
3	Cheyenne	917892	Laramie	1500.1784	7158	20.34	14612.64	59466
4	Cody	218376	Park	2998.95696	1403	1.82	3515.62	9520
5	Douglas	208008	Converse	1829.4651	832	1.46	1744.08	6120
6	Evanston	283824	Uinta	999.4971	1486	4.95	2712.64	12359
7	Gillette	543132	Campbell	2748.8529	4052	5.8	7189.43	29087
8	Powell	233928	Park	2673.57455	1251	1.62	3134.18	6314
9	Riverton	303264	Fremont	4796.859815	2680	2.34	5556.49	10615
10	RockSprings	253584	Sweetwater	6620.201916	4022	2.78	7572.18	23036
11	Sheridan	308232	Sheridan	1893.977048	2646	8.98	6039.71	17444

The above figure represents the joined dataset for 11 city locations after cleaning and blending(using inner join) the three original datasets.



Record	Sum_Sum_Pawdacity	Sum_Land Area	Sum_Households with Under 18	Sum_Population Density	Sum_Total Families	Sum_2010 Census
1	3773304	33071.380389	34064	62.8	62652.79	213862

- Census Population: 213,862
- Total Pawdacity Sales: 3,773,304
- Households with Under 18: 34,064
- Land Area: 33,071
- Population Density: 63
- Total Families: 62,653

The above results are the sum of values of all cities, where Pawdacity has store locations.

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	19442.00
Total Pawdacity Sales	3,773,304	343027.64
Households with Under 18	34,064	3096.73
Land Area	33,071	3006.49
Population Density	63	5.71
Total Families	62,653	5695.71

6 of 6 Fields Cell Viewer 1 record displayed, 1804 bytes Search Data Metadata						
Record	Avg_Sales	Avg_Land Area	Avg_Households with Under 18	Avg_Population Density	Avg_Total Families	Avg_2010 Census
1	343027.636364	3006.489126	3096.727273	5.709091	5695.708182	19442

8 of 8 Fields Cell Viewer 11 records displayed Search Data Metadata Copy Paste								
Record	CITY	Sum_Pawdacity	County	Land Area	Households with Under 18	Population Density	Total Families	2010 Census
1	Buffalo	185328	Johnson	3115.5075	746	1.55	1819.5	4585
2	Casper	317736	Natrona	3894.3091	7788	11.16	8756.32	35316
3	Cheyenne	917892	Laramie	1500.1784	7158	20.34	14612.64	59466
4	Cody	218376	Park	2998.95696	1403	1.82	3515.62	9520
5	Douglas	208008	Converse	1829.4651	832	1.46	1744.08	6120
6	Evanston	283824	Uinta	999.4971	1486	4.95	2712.64	12359
7	Gillette	543132	Campbell	2748.8529	4052	5.8	7189.43	29087
8	Powell	233928	Park	2673.57455	1251	1.62	3134.18	6314
9	Riverton	303264	Fremont	4796.859815	2680	2.34	5556.49	10615
10	RockSprings	253584	Sweetwater	6620.201916	4022	2.78	7572.18	23036
11	Sheridan	308232	Sheridan	1893.977048	2646	8.98	6039.71	17444

The above figure displays the average for each column in the dataset(before removing outliers) and we attain a dataset of 11 rows.

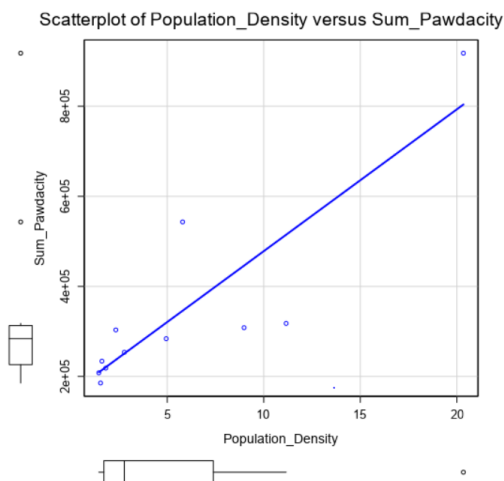
Step 3: Dealing with Outliers

Answer these questions

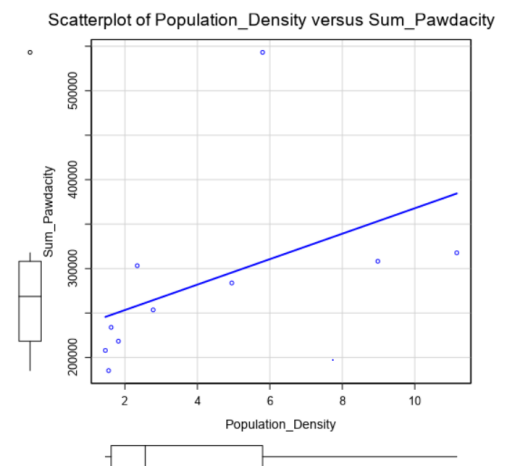
Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.

	A	B	C	D	E	F	G	H
1	CITY	County	Sum_Pawdacity	Land Area	Households with Under 18	Population Density	Total Families	2010 Census
2	Buffalo	Johnson	185328	3115.5075	746	1.55	1819.5	4585
3	Casper	Natrona	317736	3894.3091	7788	11.16	8756.32	35316
4	Cheyenne	Laramie	917892	1500.1784	7158	20.34	14612.64	59466
5	Cody	Park	218376	2998.95696	1403	1.82	3515.62	9520
6	Douglas	Converse	208008	1829.4651	832	1.46	1744.08	6120
7	Evanston	Uinta	283824	999.4971	1486	4.95	2712.64	12359
8	Gillette	Campbell	543132	2748.8529	4052	5.8	7189.43	29087
9	Powell	Park	233928	2673.57455	1251	1.62	3134.18	6314
10	Riverton	Fremont	303264	4796.859815	2680	2.34	5556.49	10615
11	RockSprings	Sweetwater	253584	6620.201916	4022	2.78	7572.18	23036
12	Sheridan	Sheridan	308232	1893.977048	2646	8.98	6039.71	17444
13								
14	Q1		218376	1829.4651	1251	1.62	2712.64	6314
15	Q3		317736	3894.3091	4052	8.98	7572.18	29087
16	IQR		99360	2064.844	2801	7.36	4859.54	22773
17	Upper Fence		466776	6991.5751	8253.5	20.02	14861.49	63246.5
18	Lower Fence		69336	-1267.8009	-2950.5	-9.42	-4576.67	-27845.5

The above spreadsheet outputs the record of the final dataset. The values are categorized according to *CITY*. On the bottom section, we have performed the IQR method to recognize outliers in the dataset. The first quartile, 3rd quartile, IQR, lower fence and the upper fence are mentioned. The values that fall outside of the upper and lower fence are formatted in the color white and their city name is marked with yellow. Here we can see there are two outliers, **Cheyenne** and **Gillette**. Let us compare the direction slope of regressions line, with and without the outlier “**Cheyenne**”.



With Outlier



Without Outlier

FIG.1

Now, let us compare the slope of the regression line with and without the outlier “**Gillette**”

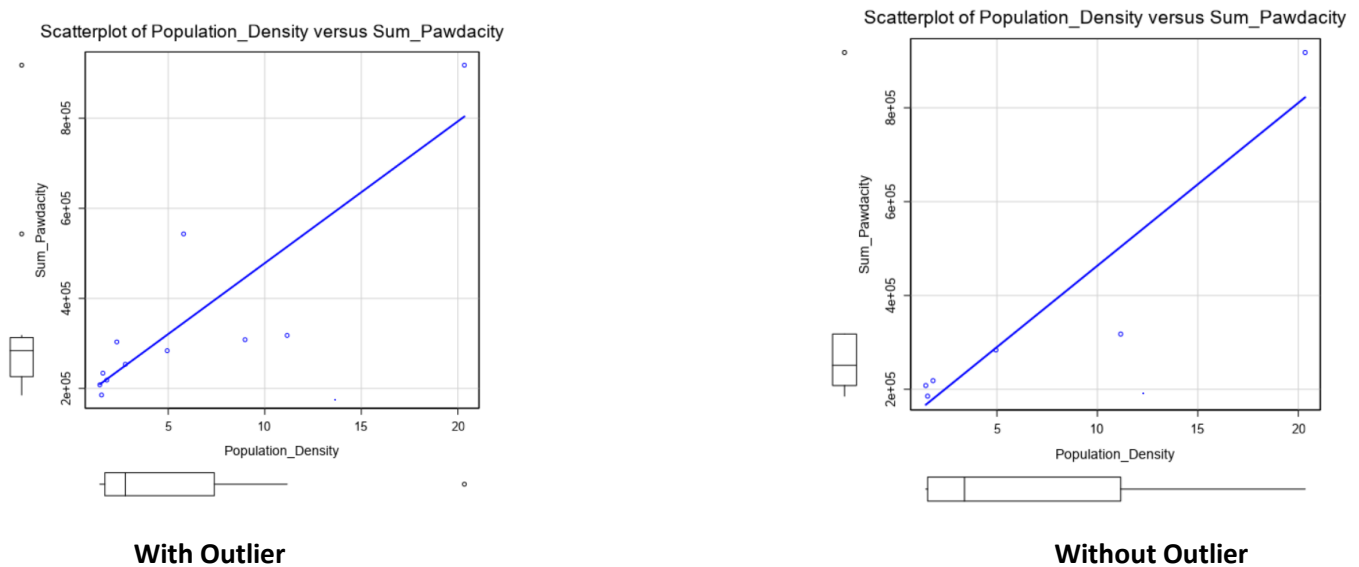


FIG 2.

After performing the IQR method we compare the distribution of values in the x-variable of the above scatter plots with that of the y-variable (population density in the x-variable and sum of sales for all pawdacity locations on the y-variable). From **Figure 1**, we notice that Cheyenne has an abnormal pop. Density and sum of sales. But this doesn't necessarily mean the city(data point) is an outlier. Moreover, this city can lead to insightful analysis. Hence, this outlier is included for the training model. From **Figure 2**, we notice that Gillette's regression line is not affected by the presence or absence of the outlier. Hence, the city “**Gillette**” can be removed from the training model since it is an insignificant outlier for the predictive model.

Alteryx Workflow

