Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Key Decisions:

Answer these questions

1. What decisions needs to be made?

Business decisions need to be made around opening a 14th store for Pawdacity. We currently have 4 sets of data that contains information related to 13 Pawdacity stores. The data contains information on location, sales, population and demographics. The initial phase is to extract relevant fields of information clean and merge the data to create a new data set that can be used to determine the viability of a 14th store based on the location.

2. What data is needed to inform those decisions?

The following sets of data needs to be extracted for 11 stores across different geographic locations.

Column
Census Population
Total Pawdacity Sales
Households with Under 18
Land Area
Population Density
Total Families

Once we have extracted the following fields in the desired format we will have to utilize IQR to determine the outliers using the upper and lower fence values.

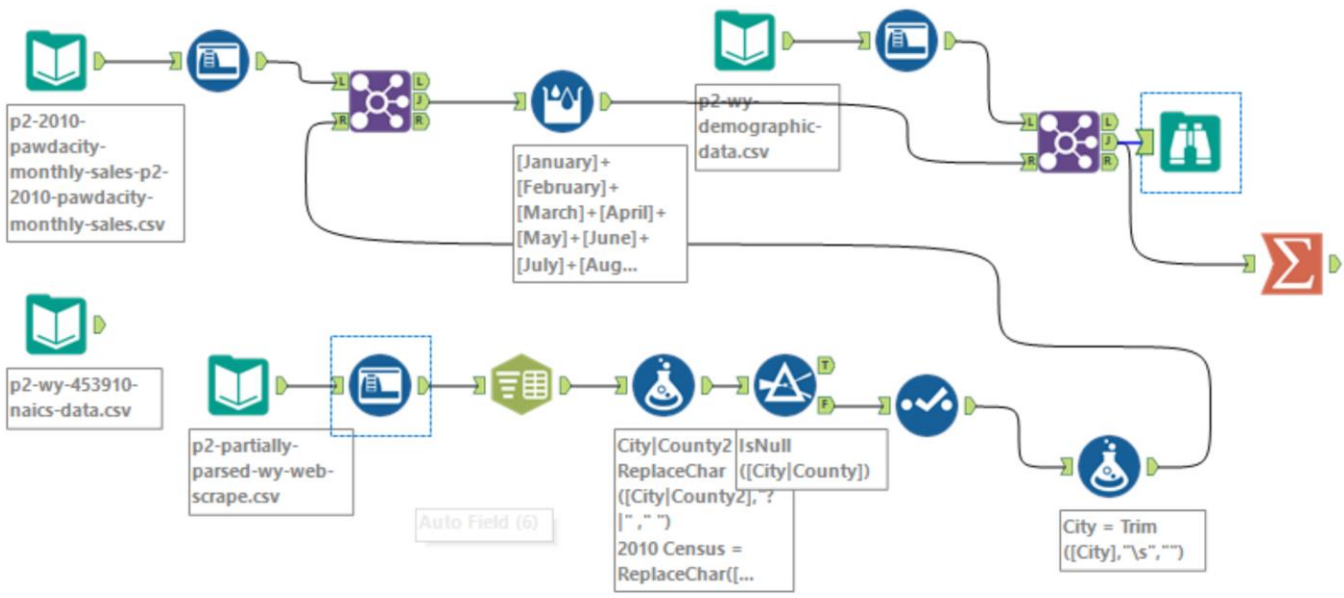
Step 2: Building the Training Set

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*	
Total Pawdacity Sales	*3,773,304*	
Households with Under 18	*34,064*	
Land Area	*33,071*	
Population Density	*63*	
Total Families	*62,653*	

Alteryx WorkSpace Project 2



Results - Browse (29) - Input

7 of 7 Fields | Cell Viewer | 11 records displayed, 3580 bytes | Data | Metadata

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

Output Data Project 2

1st quartile Q1
3rd quartile Q3 function used QUARTILE.INC
Interquartile Range: IQR = Q3 - Q1
Upper Fence = Q3 + 1.5 IQR
Lower Fence = Q1 - 1.5 IQR
Values above Upper Fence and values below the Lower Fence are outliers

City	2010 Census Population	Total Pawdacity Sales	Land Area	Households with Under 18	Population Density	Total Families	
Buffalo	4585.00	185328.00	3115.51	746.00	1.55	1819.50	**Legend : Red Data Fills are Outliers**
Casper	35316.00	317736.00	3894.31	7788.00	11.16	8756.32	
Cheyenne	59466.00	917892.00	1500.18	7158.00	20.34	14612.64	
Cody	9520.00	218376.00	2998.96	1403.00	1.82	3515.62	
Douglas	6120.00	208008.00	1829.47	832.00	1.46	1744.08	
Evanston	12359.00	283824.00	999.50	1486.00	4.95	2712.64	
Gillette	29087.00	543132.00	2748.85	4052.00	5.80	7189.43	
Powell	6314.00	233928.00	2673.57	1251.00	1.62	3134.18	
Riverton	10615.00	303264.00	4796.86	2680.00	2.34	5556.49	
Rock Springs	23036.00	253584.00	6620.20	4022.00	2.78	7572.18	
Sheridan	17444.00	308232.00	1893.98	2646.00	8.98	6039.71	
SUM	213862.00	3773304.00	33071.38	34064.00	62.80	62652.79	
AVERAGE	19442.00	343027.64	3006.49	3096.73	5.71	5695.71	
Q1	6314.00	218376.00	1829.47	1251.00	1.62	2712.64	
Q3	29087.00	317736.00	3894.31	4052.00	8.98	7572.18	
IQR	22773.00	99360.00	2064.84	2801.00	7.36	4859.54	
Upper Fence	63246.50	466776.00	6991.58	8253.50	20.02	14861.49	
Lower Fence	-27845.50	69336.00	-1267.80	-2950.50	-9.42	-4576.67	

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.

Utilizing the Upper and Lower Fences from the IQR data. We find the following outliers:

City	2010 Census Population	Total Pawdacity Sales	Land Area	Households with Under 18	Population Density	Total Families
Cheyenne	59466.00	917892.00	1500.18	7158.00	20.34	14612.64
Gillette	29087.00	543132.00	2748.85	4052.00	5.80	7189.43

Cheyenne: The Total sales volume does correlate to the overall population/population density. This can be included in the data set and does not need to be either removed or imputed.

Gillette: The sales volume in comparison to the population/population density has little correlation. The data can be removed with out an issue. Imputing the data could cause an issue with final analysis.