Project 2.1: Data Cleanup

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?

   ✓ Pawdacity, a pet shop, is thinking of opening the 14th shop. The decision to be made is choosing the best location.

2. What data is needed to inform those decisions?

   ✓ The decision depends on the predicted yearly sales per city. To achieve that the following data is needed:

   - Actual yearly sales
   - Population metrics such as density, size, total families and land area
   - Data on the competition

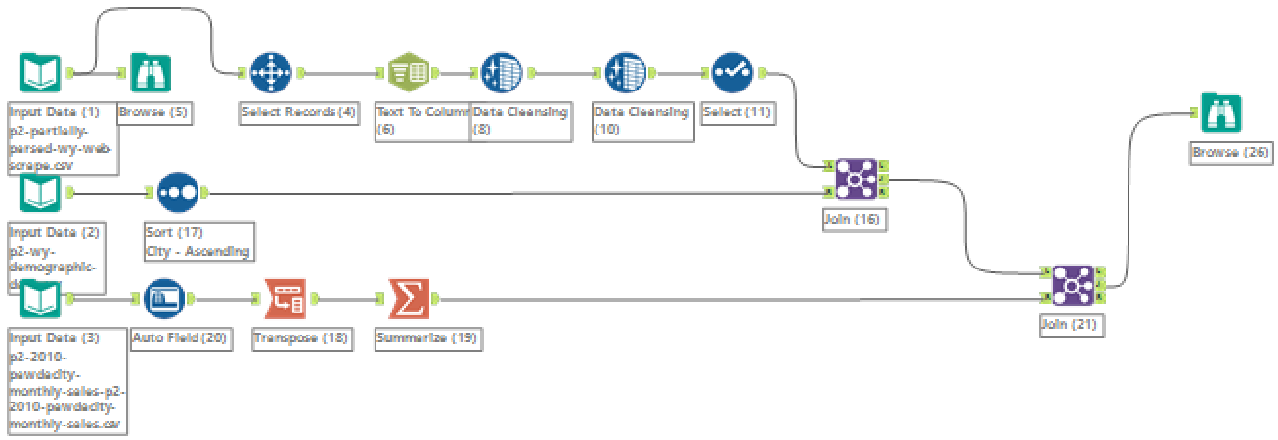
Step 2: Building the Training Set

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	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	5695.71

✓ Below you can find the Alteryx workflow I used for data cleaning and preparation:



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.

- ✓ To identify outliers I used box and whisker graph with Tableau. Observing the graphs (check the following page) I can say that:

2010 Census Population

Only one city that stands out from the all other as an outlier: Cheyenne

Households with under 18 years

There does not appear to be outliers.

Land Area

Only the Rock Springs shows as an outlier.

Population Density

Only one city that stands out from the all other as an outlier: Cheyenne

Total Families

Only in Cheyenne the Total Families data stands as an outlier.

Total Pawdacity Sales

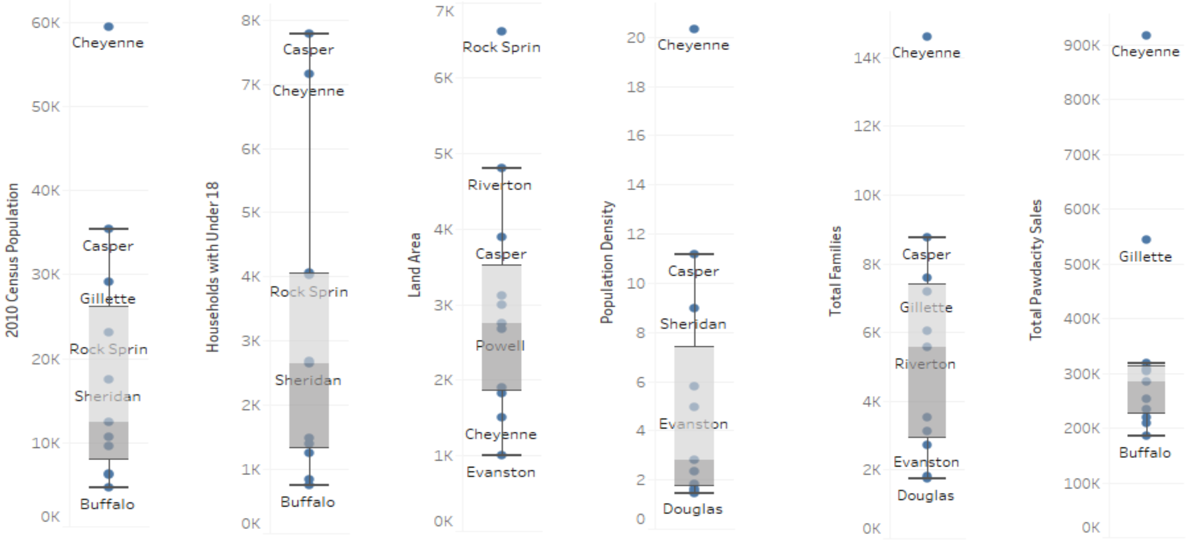
Two outliers: Gillette city and Cheyenne city.

Summary

- ✓ Cheyenne looks like an outlier but analyzing its data it seems that the values considered as outliers come from the fact that Cheyenne is a big city with population density much higher than average. Having said that, its values seem quite consistent so I'd not remove it from the dataset.

- ✓ On the other hand, the sales values for Gillette seem quite high considering its low population density and the fact that it is not as large of a city. Even compared with other cities of the same population density, the sales values are pretty high.

- ✓ Hence the city of Gillete is the outlier I'd remove since it would have a disproportionate effect on statistical analysis which can result in misleading interpretations.


2010 Census Population
60K
50K
40K
30K
20K
10K
0K
Cheyenne
Casper
Gillette
Rock Sprin
Sheridan
Buffalo
Households with Under 18
8K
7K
6K
5K
4K
3K
2K
1K
0K
Casper
Cheyenne
Rock Sprin
Sheridan
Buffalo
Land Area
7K
6K
5K
4K
3K
2K
1K
0K
Rock Sprin
Riverton
Casper
Powell
Cheyenne
Evanston
Population Density
Cheyenne
Casper
Sheridan
Evanston
Douglas
Total Families
14K
12K
10K
8K
6K
4K
2K
0K
Cheyenne
Casper
Gillette
Riverton
Evanston
Douglas
Total Pawdacity Sales
900K
800K
700K
600K
500K
400K
300K
200K
100K
0K
Cheyenne
Gillette
Buffalo