Project 2.1: Data Cleanup

Make a copy of this document. Complete each section. When you are ready, save your file as a PDF document and submit it here: <https://classroom.udacity.com/nanodegrees/nd008/parts/8d60a887-d4c1-4b0e-8873-b2f36435eb39/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?

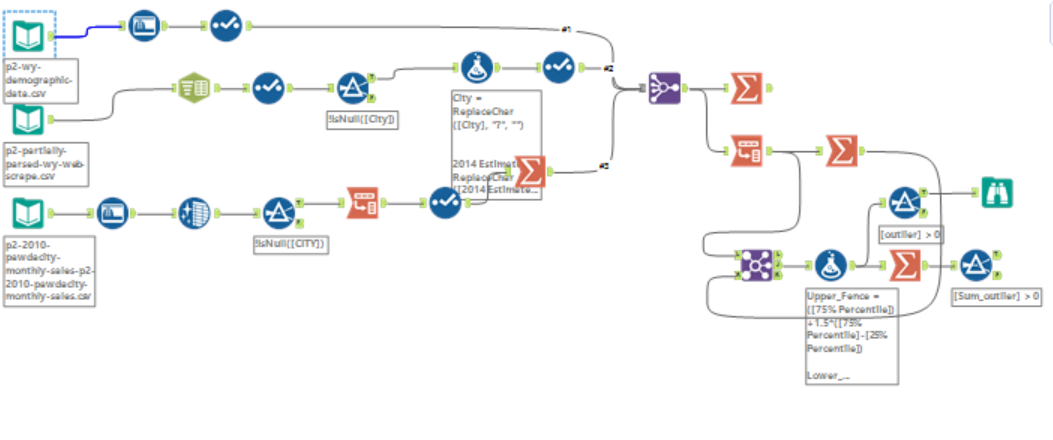
Pawdacity is a Wyoming pet store chain, the manger wants to open a new store. So based on historic data of current stores, a decision will be made to locate 14th store.

1. What data is needed to inform those decisions?

To choose the location, we need to have census population, total sale for each of store, household under 18, land area, number total families, also we need competitor sales information.

## 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*

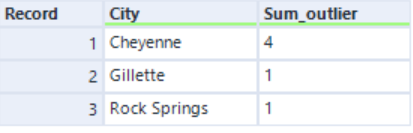
|  |  |  |
| --- | --- | --- |
| **Column** | **Sum** | **Average** |
| *Census Population* | *213,862* |  |
| *Total Pawdacity Sales* | *3,773,304* |  |
| *Households with Under 18* | *34,064* |  |
| *Land Area* | *33,071* |  |
| *Population Density* | *63* |  |
| *Total Families* | *62,653* |  |

## 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.

This table is the outlier cities.



In total there are 6 outliers in 3 cities.

The Cheyenne contribute most amount. I think because this is large cities, so the population size and number of total families are much more than other cities. So, I don't think this is outlier.

The Gillette has one outlier. This city has less population than Casper, less population density and a smaller number of families. But the sale amount is much more than Casper, which is abnormal, So, I will consider Gillette as outlier.

Before you Submit

Please check your answers against the requirements of the project dictated by the [rubric](https://review.udacity.com/#!/rubrics/382/view) here. Reviewers will use this rubric to grade your project.