**Predictive Analytics for Business Nanodegree**

**Project: Create an Analytical Dataset**

**By Samuel Otisi**

**Step 1: Business and Data Understanding**

**Key Decisions:**

1. What decisions needs to be made?

The key decision that needs to be made is that Pawdacity store managers are looking for the best city to open a 14thstore coupled with the already existing 13 stores in the state of Wyoming.

1. What data is needed to inform those decisions?

Previous monthly sales data from across various cities in Wyoming will be needed to build a model that will predict yearly sales of Pawdacity stores. The Sales data will be blended, formatted and cleaned with census and demographic data across the cities in Wyoming.

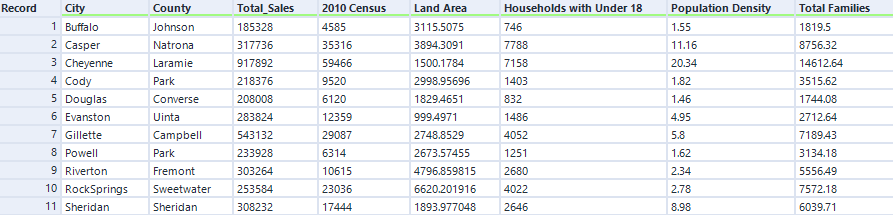
**Step 2: Building the Training Set**

After blending and cleaning the dataset I came up with the following data as a sum and averages of their respective data fields. The result obtained is in line with the expected results from the project reviewers.

|  |  |  |
| --- | --- | --- |
| **Column** | **Sum** | **Average** |
| *Census Population* | *213862* | *19442* |
| *Total Pawdacity Sales* | *3773304* | *343027.64* |
| *Households with Under 18* | *34064* | *3096.73* |
| *Land Area* | *33071.38* | *3006.49* |
| *Population Density* | *62.80* | *5.71* |
| *Total Families* | *62652.79* | *5695.71* |

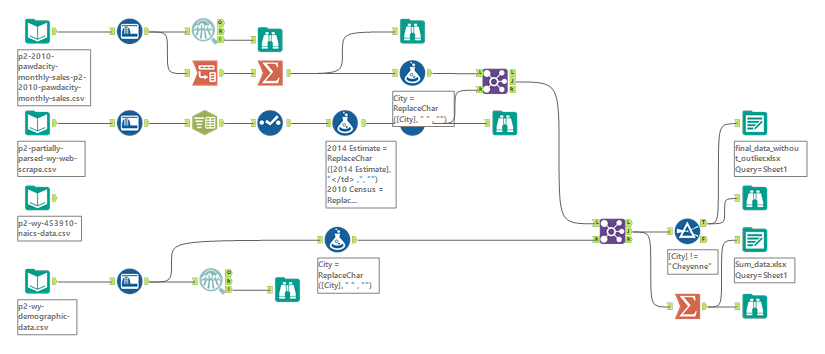
After cleaning up the datasets I came up with 11 rows of data, the table below is the final dataset before removing the outlier.

**Figure 1. final dataset after cleaning and blending.**



Also, the following image is the accompanied alteryx workflow, showing how I arrived at the above table. All files are attached in the zip file.

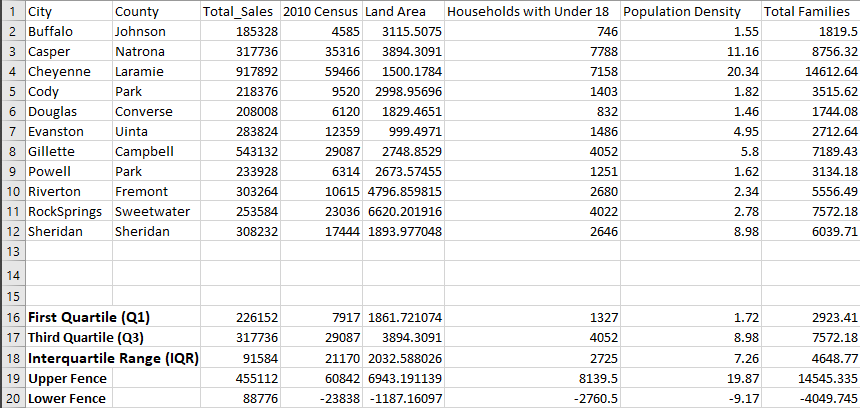
**Figure 2. Alteryx workflow.**



**Step 3: Dealing with Outliers**

I have used the box and whisker method to detect an outlier in final cleaned dataset, according to my workout there is an outlier in the Cheyenne City, the total sales, Population density and Total families fileds are above the threshold for evaluating outliers. There is also another outlier in the Gillette City for the Total Sales field, but I have decided to keep it since I have only small dataset. This is the result from my calculations with excel.

**Figure 3. result.**



from the above image, 917892 and 543132 are greater than 455112 in the Total sales column, which is the upper fence of the box and whisker, therefore they are considered as outliers. Although I will keep one of the Cities, that is Gillette since the dataset is small and will go ahead and remove the Cheyenne row and finally have 10 rows of dataset as shown below.

**Figure 4. final dataset after removing the outlier observation.**

