Project 2.1: Data Cleanup

## Step 1: Business and Data Understanding

### Key Decisions:

*Answer these questions*

1. What decisions needs to be made?

Pawdacity is planning to open a new store. Need to decide which city to start up a new business.

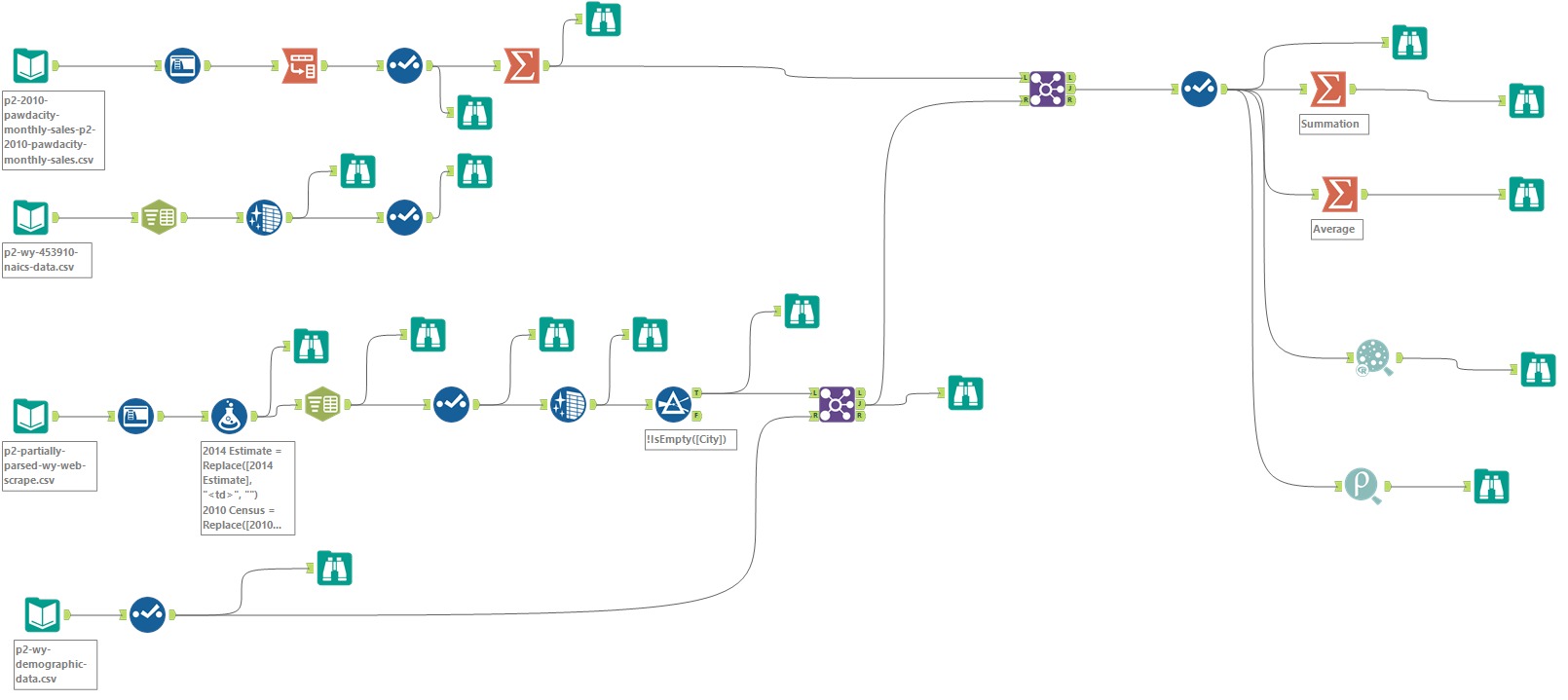
1. What data is needed to inform those decisions?

We have the following files that may be useful for us to make those decisions.

* p2-2010-pawdacity-monthly-sales.csv
* p2-partially-parsed-wy-web-scrape.csv
* p2-wy-453910-naics-data.csv

They contain sales, population and demographics data.

## Step 2: Building the Training Set

**

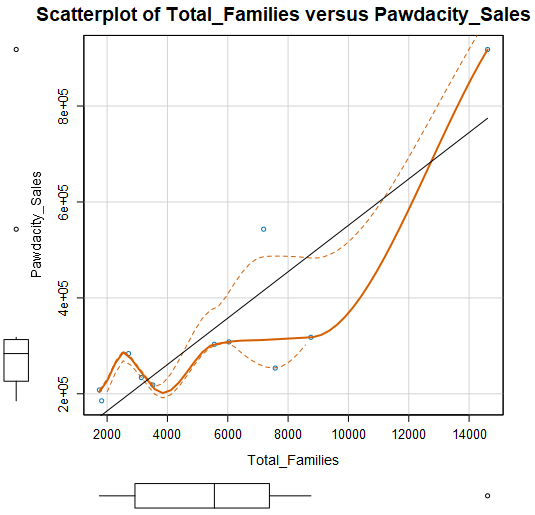
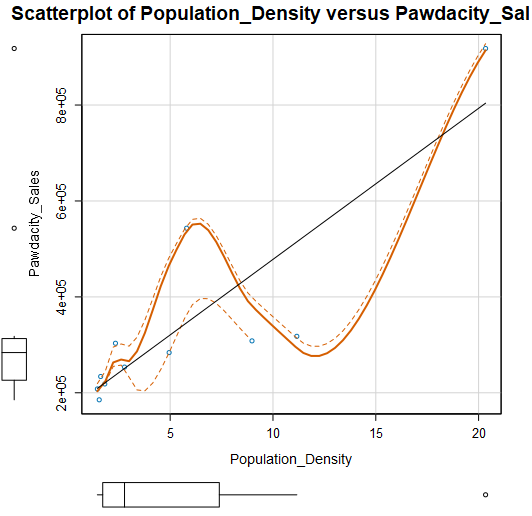
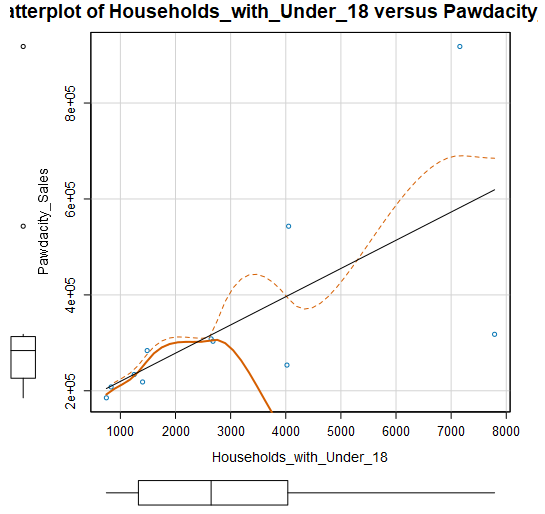
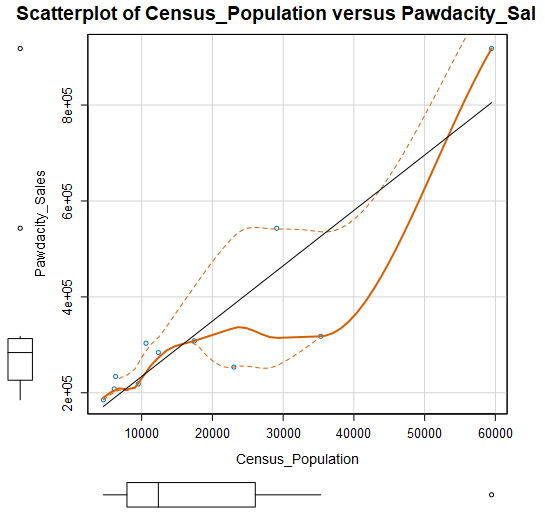
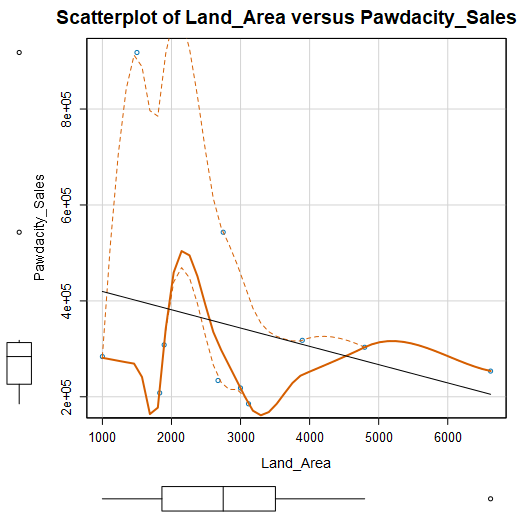
|  |  |  |
| --- | --- | --- |
| **Column** | **Sum** | **Average** |
| *Census Population* | *213,862* | *19442* |
| *Total Pawdacity Sales* | *3,773,304* | *343027.64* |
| *Households with Under 18* | *34,064* | *3096.73* |
| *Land Area* | *33,071* | *3006.49* |
| *Population Density* | *63* | *5.71* |
| *Total Families* | *62,653* | *5695.71* |

## 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.

See figures below.



From the box and scatter plots of Pawdacity Sales shown above, there are 2 points can be regarded as outliers, Cheyenne and Gillette. If we further check the data, it is clear that census population, population density and total families of Cheyenne exceeds 1.5 \* IQR. In my opinion, we can interpret Cheyenne as a high population city, and thus it has high population density and total families. I will prefer regarding it as “abnormal but correct” data point. On the contrary, Gillette city is hard to explain, can be removed.