# Project 2.1: Data Cleanup

## Step 1: Business and Data Understanding

### Key Decisions:

*Answer these questions*

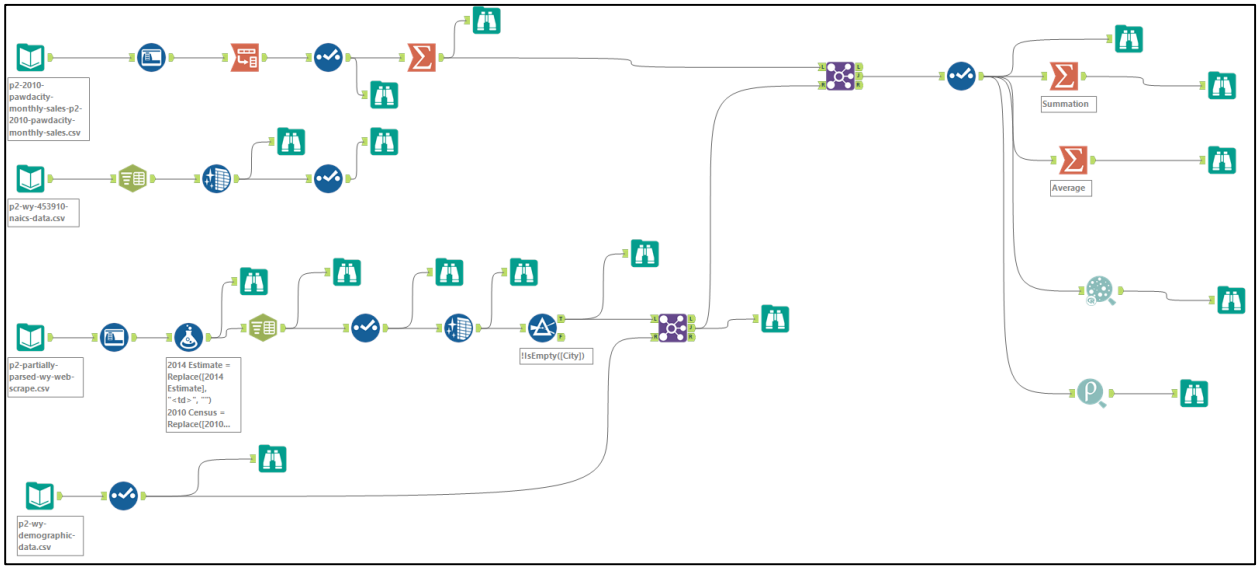
1. What decisions needs to be made?
   Pawdacity is planning to open a new store. Need to decide which city to start up a new business.

2. 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



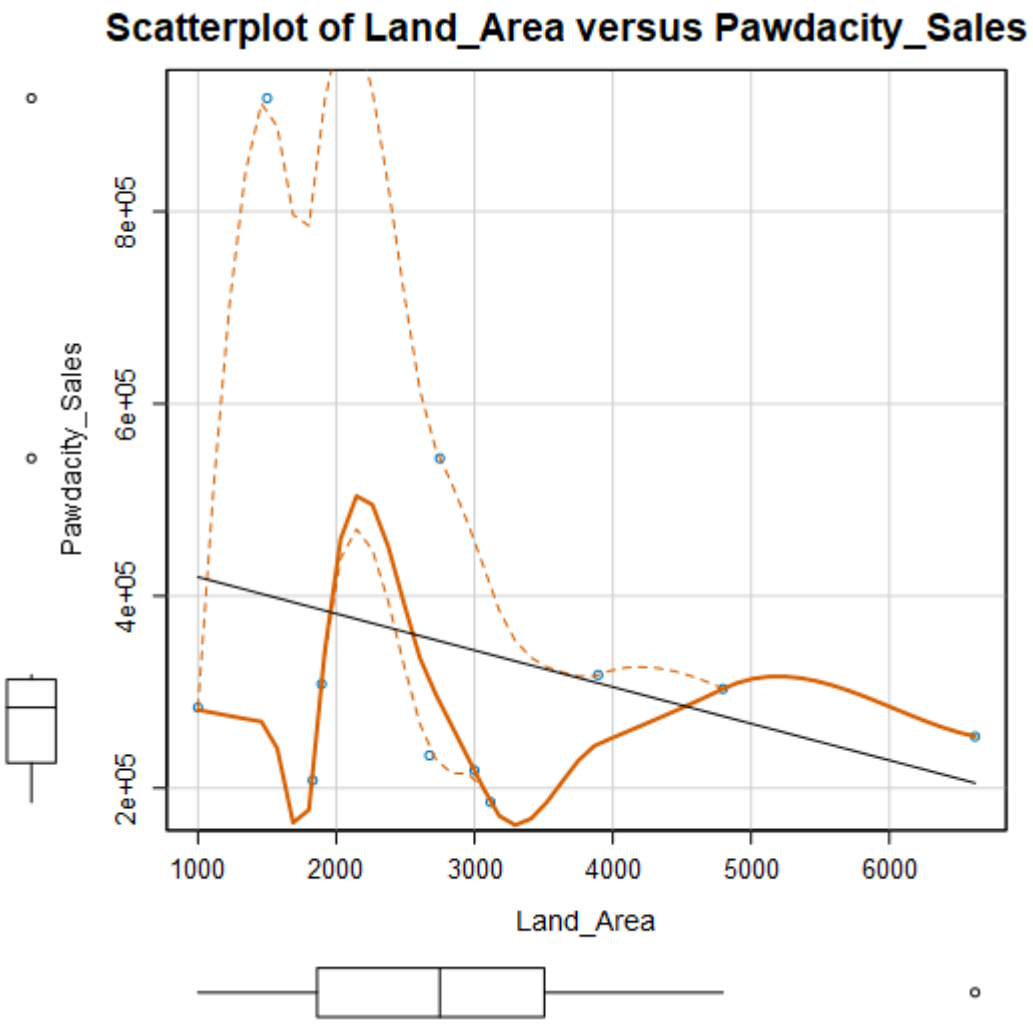

| Column | Sum | Average |
|---|---|---|
| *Census Population* | *213,862* | *19442* |
| *Total Pawdacity Sales* | *3,773,304* | *343027.64* |
| *Households with Under 18* | *34,064* | *3096.73* |
| *Land Area* | *33,071* | *3006.49* |
| *Population Density* | *63* | *5.71* |
| *Total Families* | *62,653*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

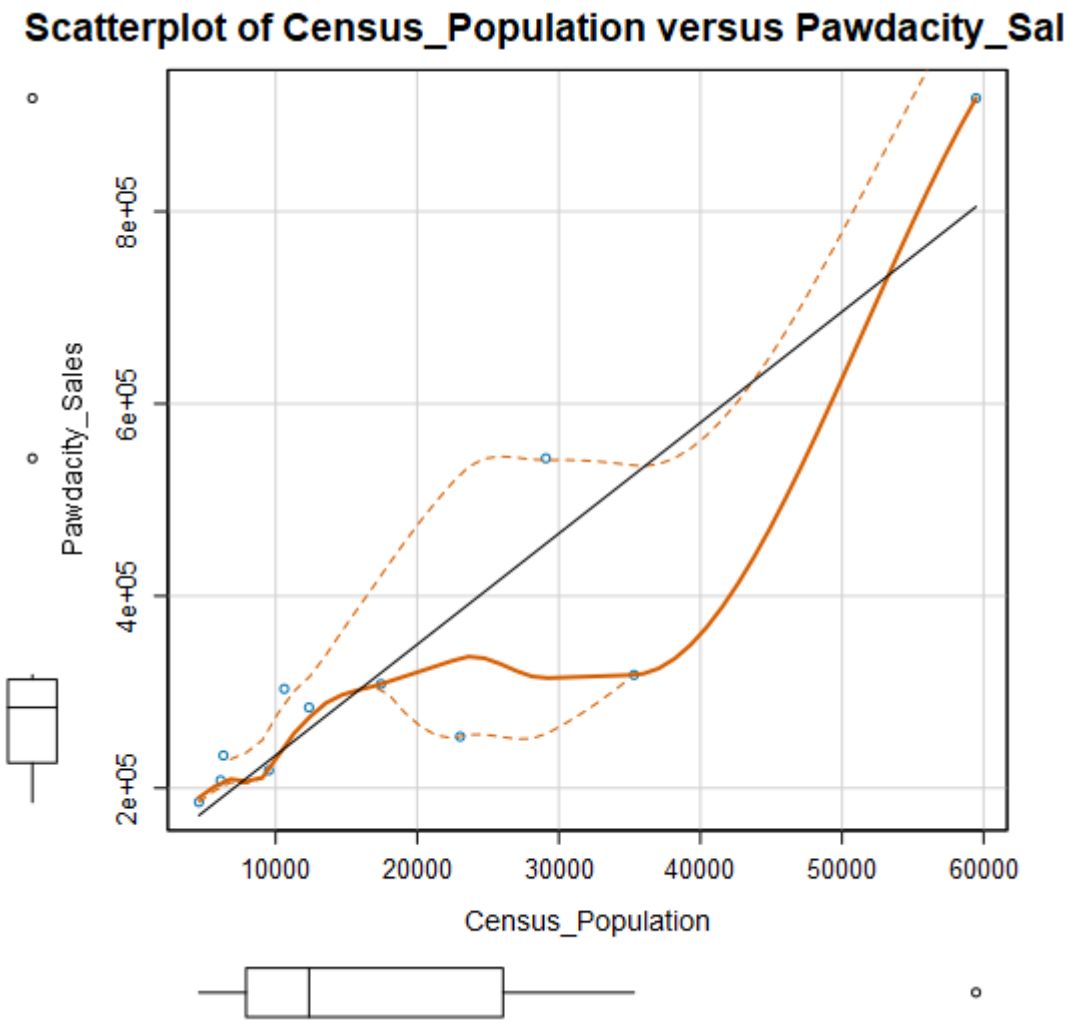
## Step 3: Dealing with Outliers

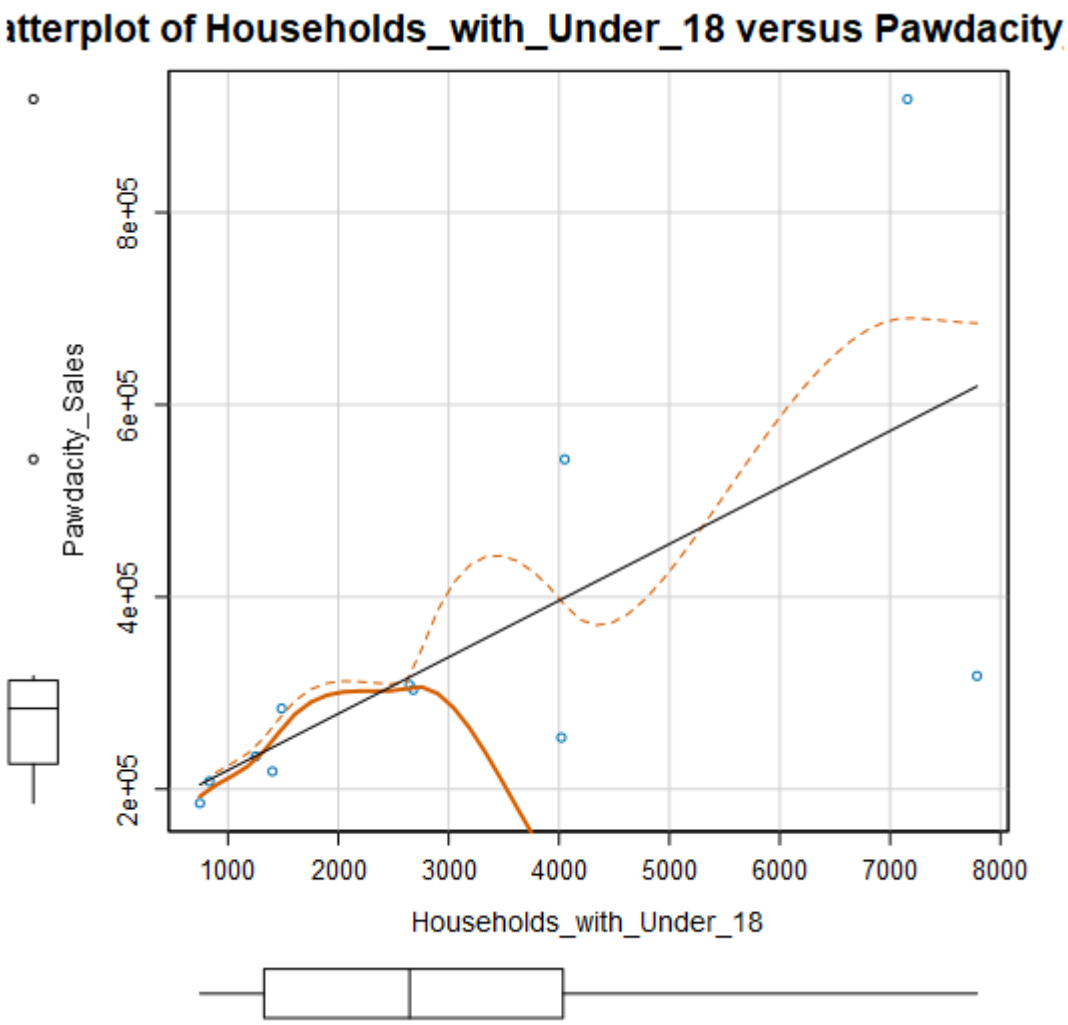
*Answer these questions*

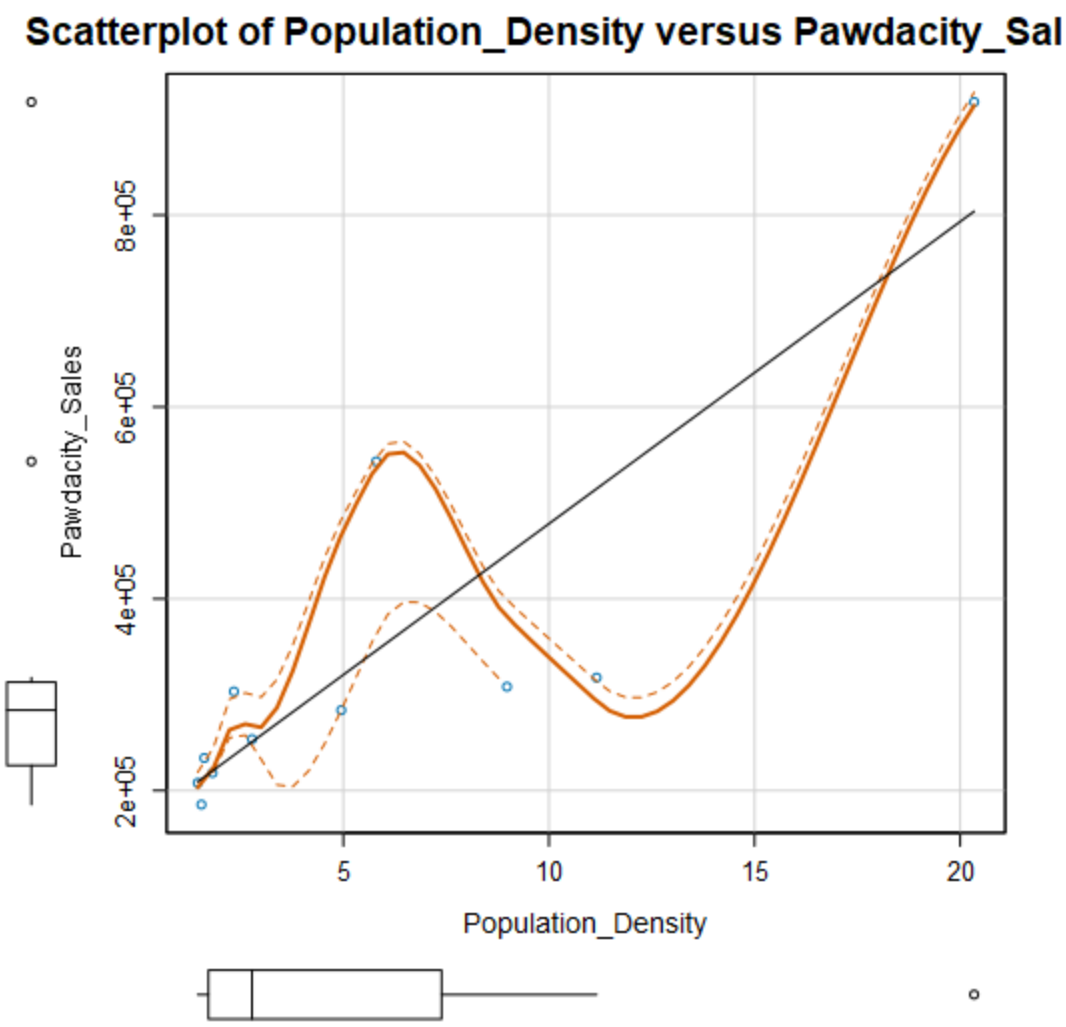
Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.

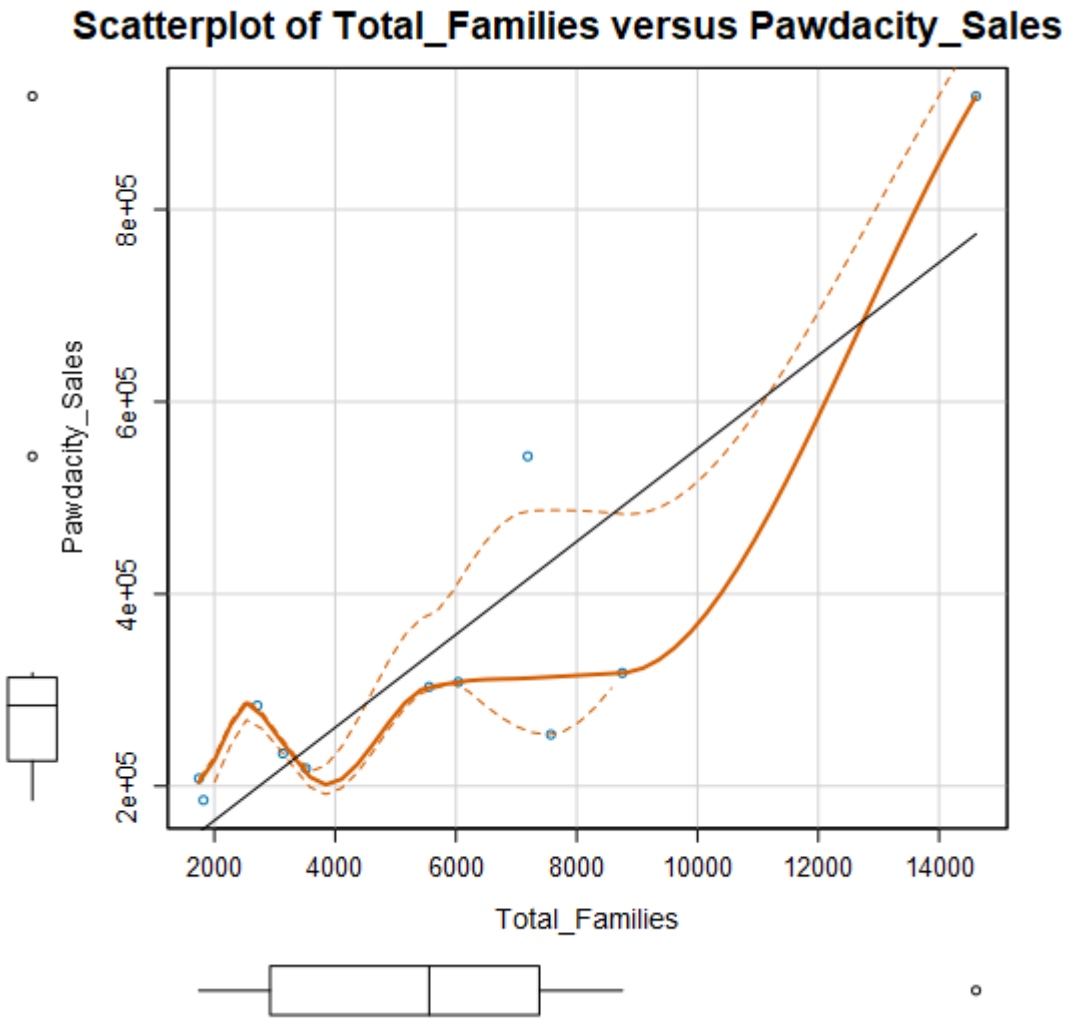
See figures below.


Scatterplot of Land_Area versus Pawdacity_Sales
Pawdacity_Sales
2e+05
4e+05
6e+05
8e+05
1000
2000
3000
4000
5000
6000
Land_Area


Scatterplot of Census_Population versus Pawdacity_Sal
Pawdacity_Sales
2e+05
4e+05
6e+05
8e+05
10000
20000
30000
40000
50000
60000
Census_Population


atterplot of Households_with_Under_18 versus Pawdacity
Pawdacity_Sales
2e+05
4e+05
6e+05
8e+05
1000
2000
3000
4000
5000
6000
7000
8000
Households_with_Under_18


Scatterplot of Population_Density versus Pawdacity_Sal
Pawdacity_Sales
2e+05
4e+05
6e+05
8e+05
5
10
15
20
Population_Density




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.