# 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?
   3 data sheets available:
   *p2-2010-pawdacity-monthly-sales.csv*
   *p2-partially-parsed-wy-web-scrape.csv*
   *p2-wy-453910-naics-data.csv*
   We need to figure out what data from the above is needed for us to predict the next store site

2. What data is needed to inform those decisions?
   We need the following data from the data provided:

- 2010 Census Population:
- Total Sales:
- Households with Under 18:
- Land Area:
- Population Density:
- Total Families:

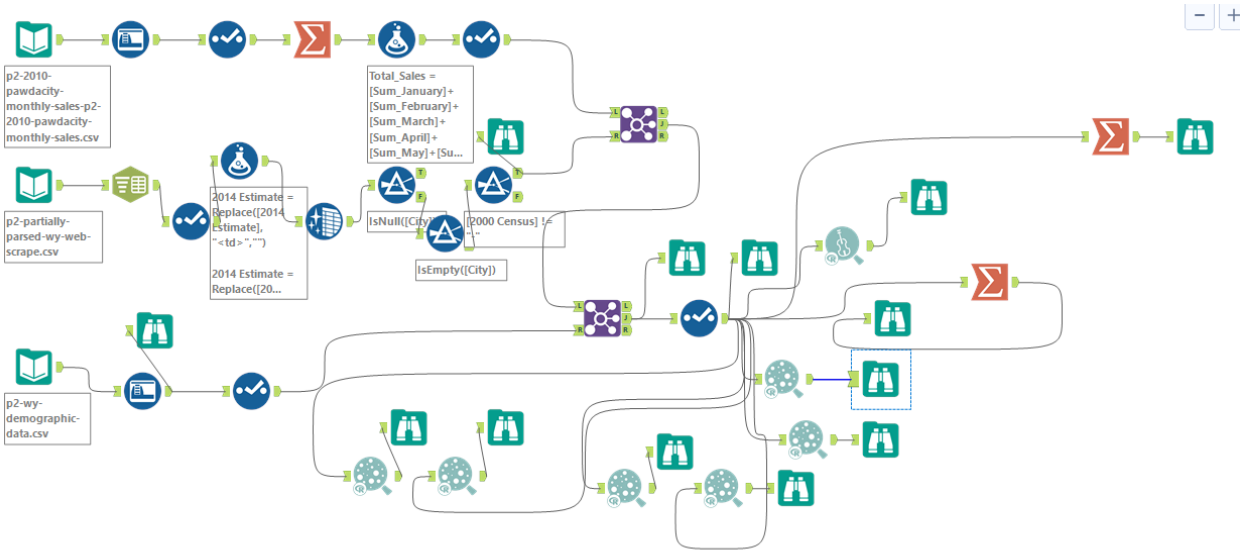
## Step 2: Building the Training Set

*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* | *19,442.00* |
| *Total Pawdacity Sales* | *3,773,304* | *343,027.64* |
| *Households with Under 18* | *34,064* | *3,096.73* |
| *Land Area* | *33,071* | *3,006.45* |

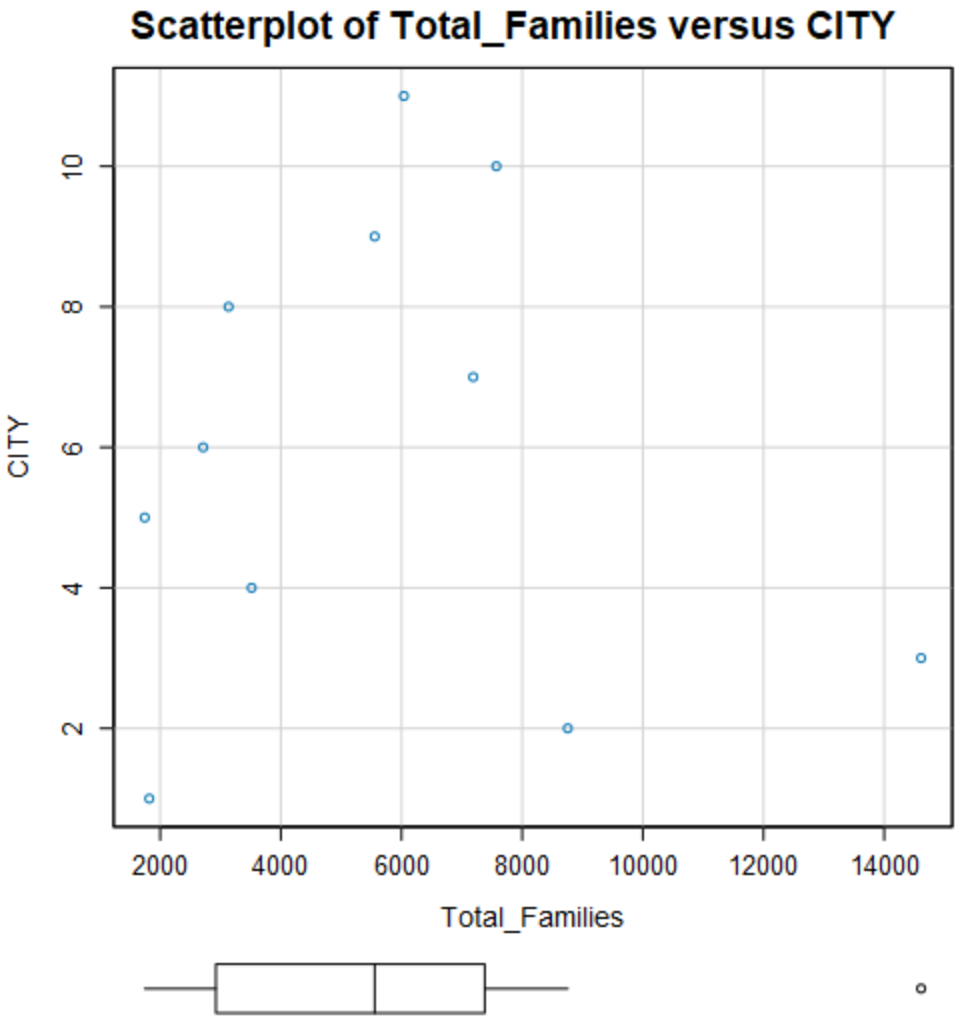
| | | |
|---|---|---|
| *Population Density* | *63* | *5.73* |
| *Total Families* | *62,653* | *5695.73* |

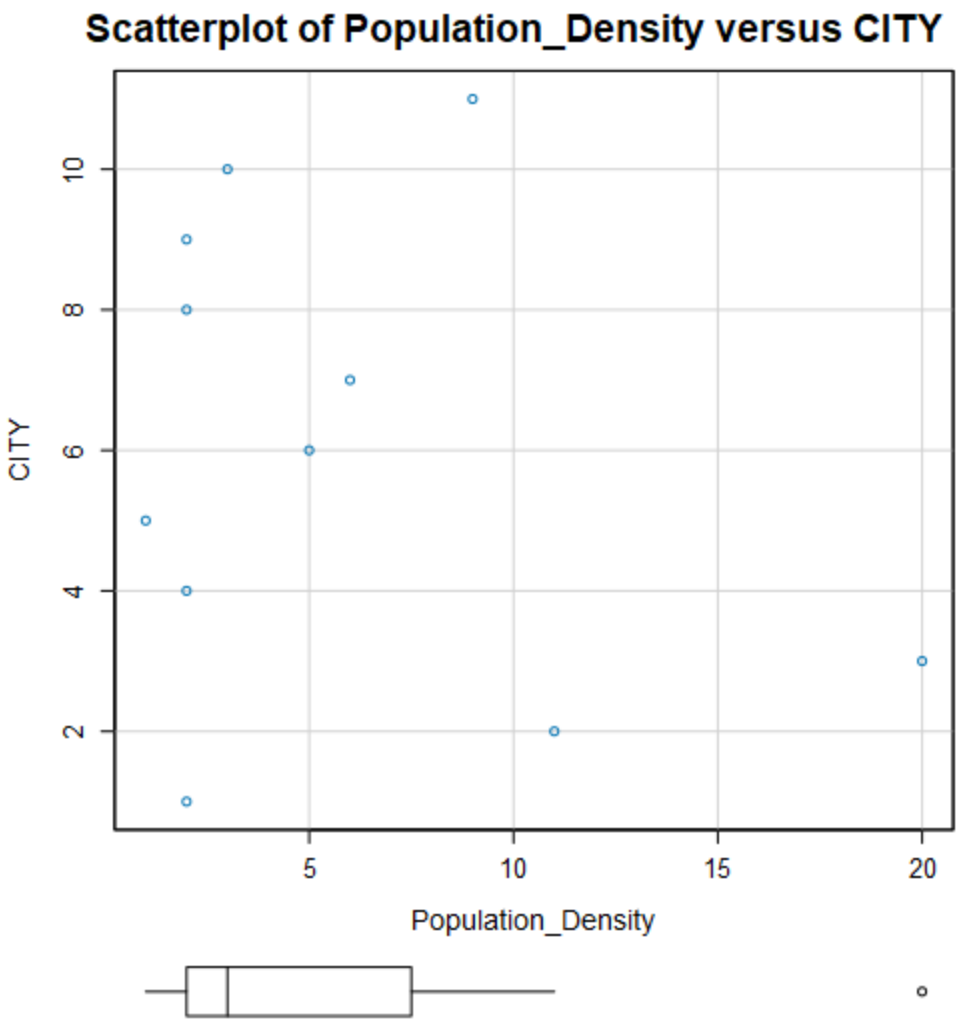


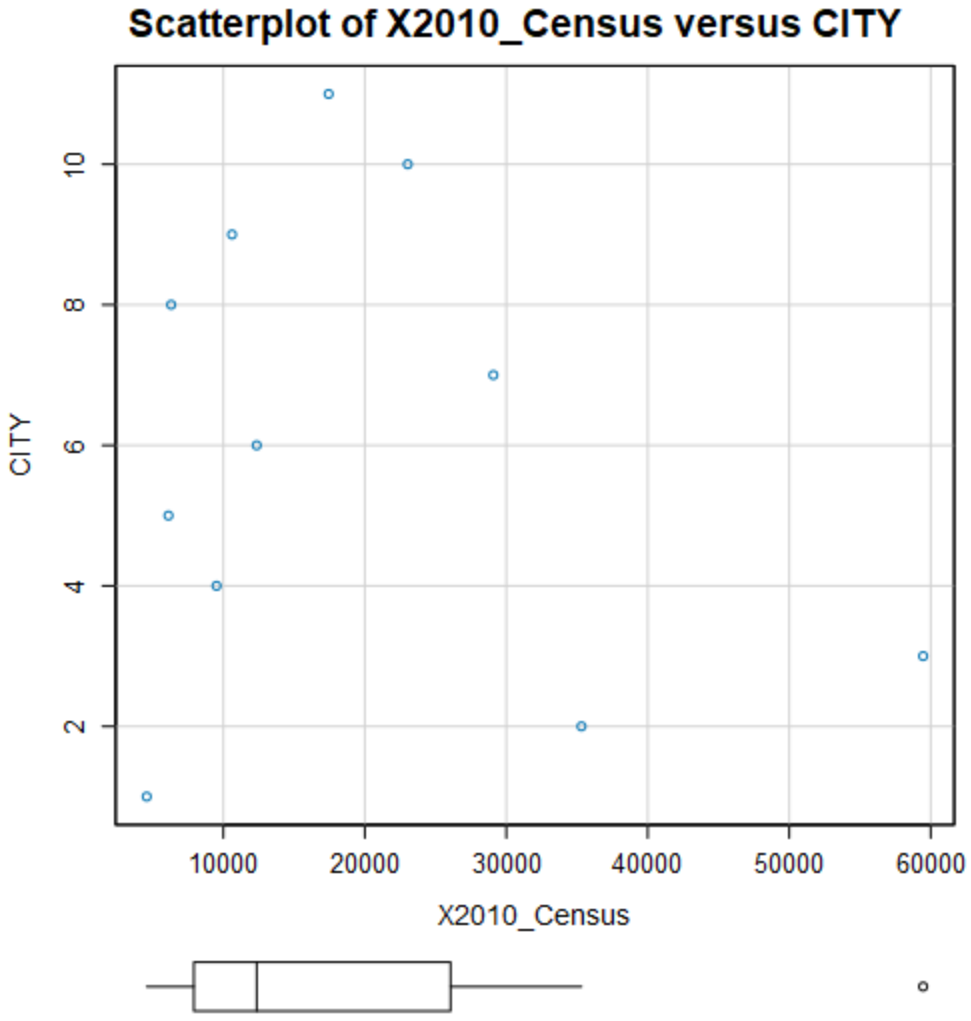

## Step 3: Dealing with Outliers

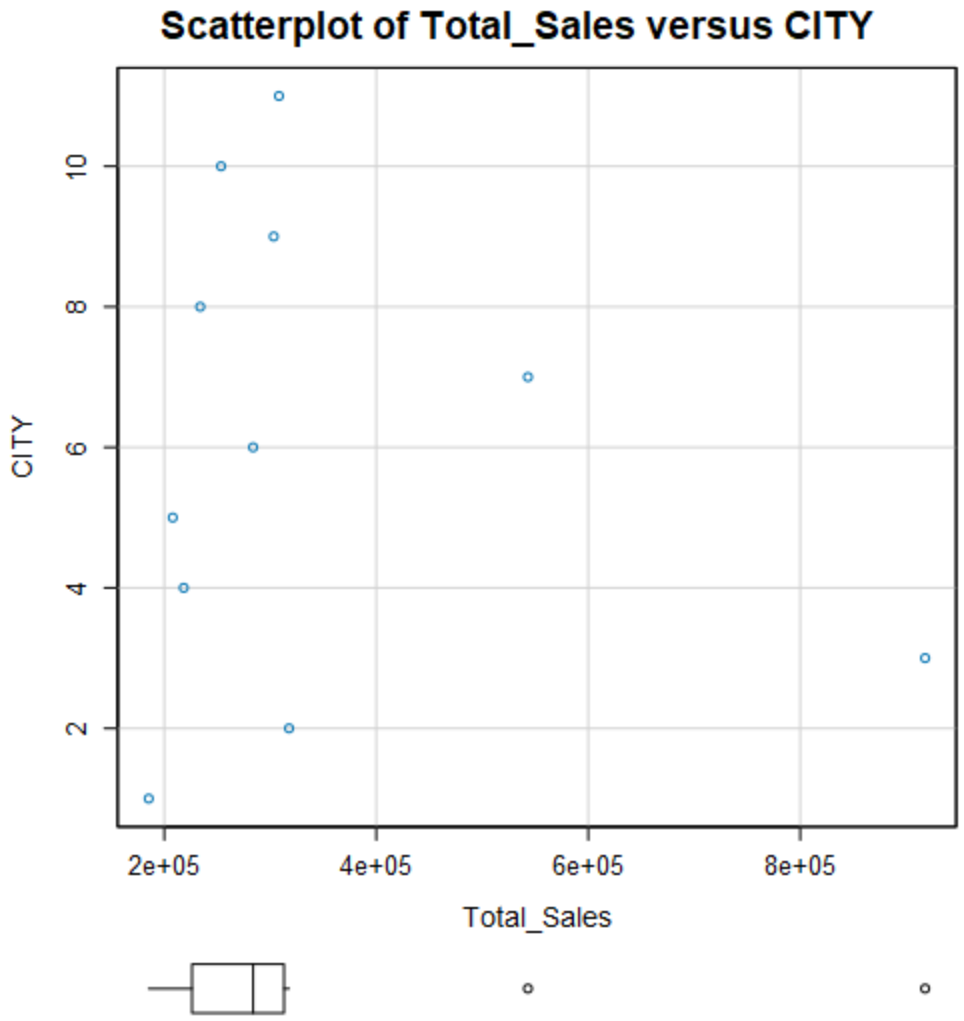
*Answer these questions*

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.
My one outlier to remove is: City #3 Cheyenne
Cheyenne has outliers 4 out of the 6 categories: Total Families, Population Density, 2010 Census & Total Sales.


Scatterplot of Total_Families versus CITY
CITY
2
4
6
8
10
2000
4000
6000
8000
10000
12000
14000
Total_Families


Scatterplot of Population_Density versus CITY
CITY
2
4
6
8
10
5
10
15
20
Population_Density


Scatterplot of X2010_Census versus CITY
CITY
2
4
6
8
10
10000
20000
30000
40000
50000
60000
X2010_Census




## Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.