# 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

*A leading pet (retail) store chain named "Pawdacity" in Wyoming, USA has 13 stores across the state and would like to expand and open a 14th store in same state. Objective is to analyze historic data, make yearly sales prediction and recommend a new location (city) for chain's newest store.*

*Listed below is criteria for choosing the right city:*

- *The new store should be located in a new city; that means there should be no existing Pawdacity store in that city.*
- *The total sales for the entire competition in this new city should be less than $500,000.*
- *The new city where we want to recommend this new store must have a population over 4,000 people (based upon the 2014 US Census estimate).*
- *The predicted yearly sales for this city must be over $200,000.*
- *The recommended city must have highest predicted sales from all options available.*

### Key Decisions:

*Answer these questions*

1. What decisions needs to be made?

*Pawdacity, a leading pet store chain in Wyoming, needs recommendation on where to open its 14th store*

2. What data is needed to inform those decisions?

*Some of the data required in order to inform this decision are,*

1. *City,*
2. *2010 census population,*
3. *Pawdacity sales in other stores,*
4. *Competitor sales,*
5. *Household with under 18,*
6. *Land area,*
7. *Population density and*
8. *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.*

*After the data cleaning activities the following training dataset was found. The R script attached was used to clean the data.*

| | city | county | land_area | Households.with.Under.18 | Population.Density | Total.Families | total_sales | X2010.Census |
|---|---|---|---|---|---|---|---|---|
| **1** | Gillette | Campbell | 2748.853 | 4052 | 5.8 | 7189.43 | 543132 | 29,087 |
| **2** | Douglas | Converse | 1829.465 | 832 | 1.46 | 1744.08 | 208008 | 6,120 |
| **3** | Riverton | Fremont | 4796.86 | 2680 | 2.34 | 5556.49 | 303264 | 10,615 |
| **4** | Buffalo | Johnson | 3115.508 | 746 | 1.55 | 1819.5 | 185328 | 4,585 |
| **5** | Cheyenne | Laramie | 1500.178 | 7158 | 20.34 | 14612.64 | 917892 | 59,466 |
| **6** | Casper | Natrona | 3894.309 | 7788 | 11.16 | 8756.32 | 317736 | 35,316 |
| **7** | Cody | Park | 2998.957 | 1403 | 1.82 | 3515.62 | 218376 | 9,520 |
| **8** | Powell | Park | 2673.575 | 1251 | 1.62 | 3134.18 | 233928 | 6,314 |
| **9** | Sheridan | Sheridan | 1893.977 | 2646 | 8.98 | 6039.71 | 308232 | 17,444 |
| **10** | Rock Springs | Sweetwater | 6620.202 | 4022 | 2.78 | 7572.18 | 253584 | 23,036 |
| **11** | Evanston | Uinta | 999.4971 | 1486 | 4.95 | 2712.64 | 283824 | 12,359 |

*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*

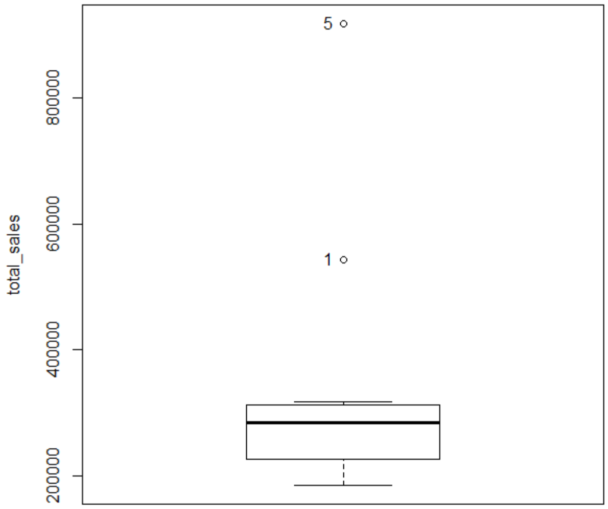
| **Column** | **Sum** | **Average** |
|---|---|---|
| *Census Population* | ***213,862*** | ***19,442*** |
| *Total Pawdacity Sales* | ***3,773,304*** | ***343,027.63*** |
| *Households with Under 18* | ***34,064*** | ***3096.72*** |
| *Land Area* | ***33,071*** | ***3006.48*** |
| *Population Density* | ***63*** | ***5.70*** |
| *Total Families* | ***62,653*** | ***5695.70*** |

## Step 3: Dealing with Outliers

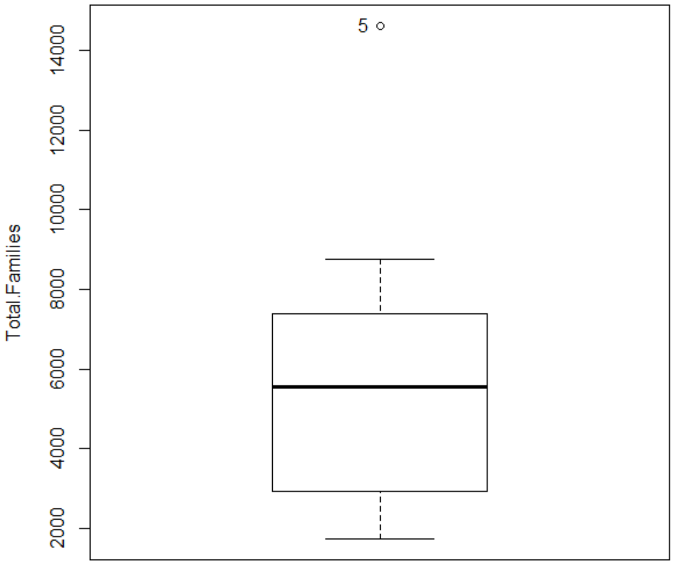
*Answer these questions*

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.

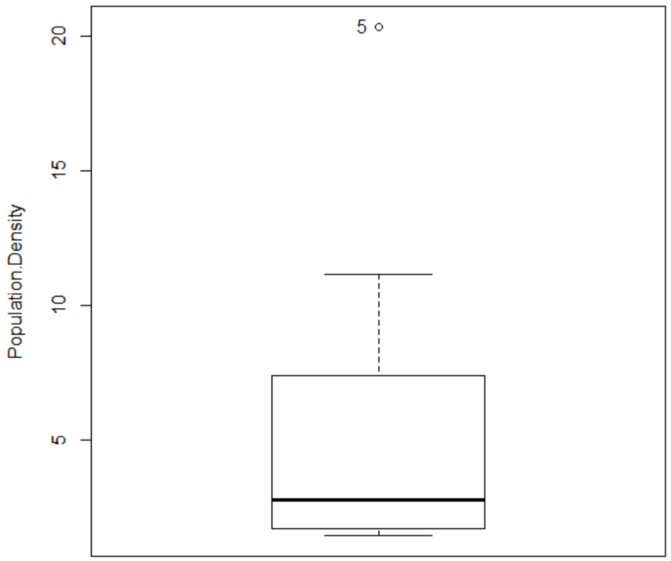
Investigation of outliers are carried out by box plots as follows,



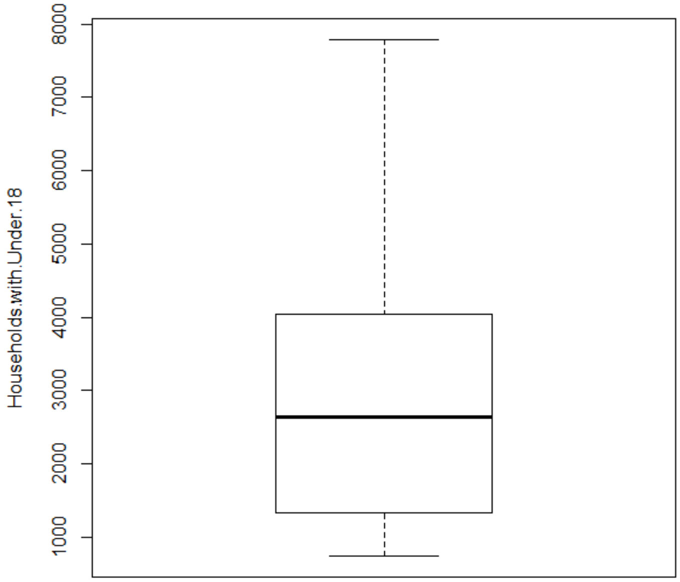
Total sales two outliers, data set 1 and 5 corresponding to cities, Gillette & Cheyenne.



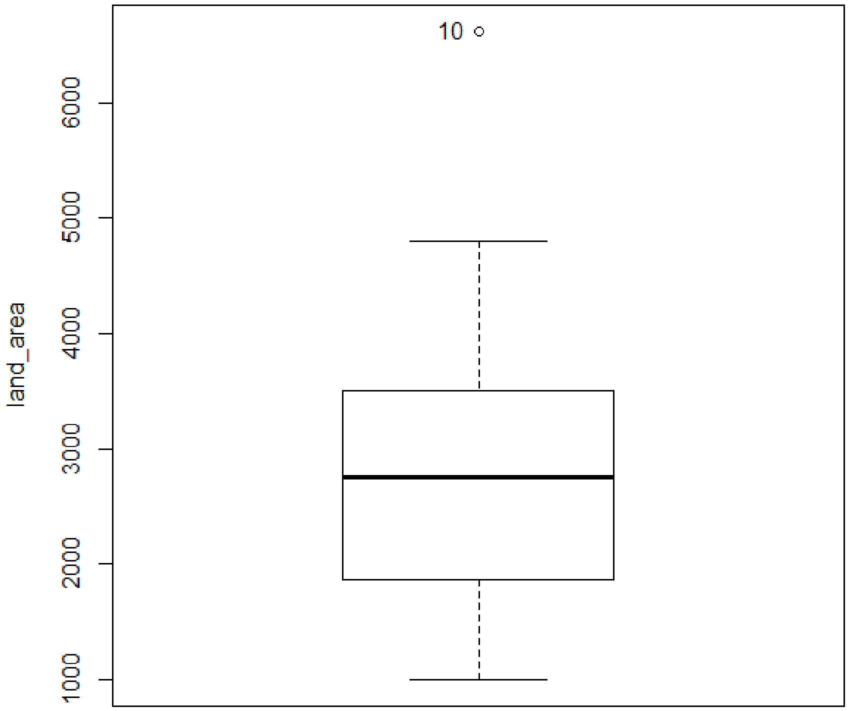
Total Families, one outlier, data set 5 corresponding to city Cheyenne.



Population Density, one outlier, data set 5 corresponding to city Cheyenne.



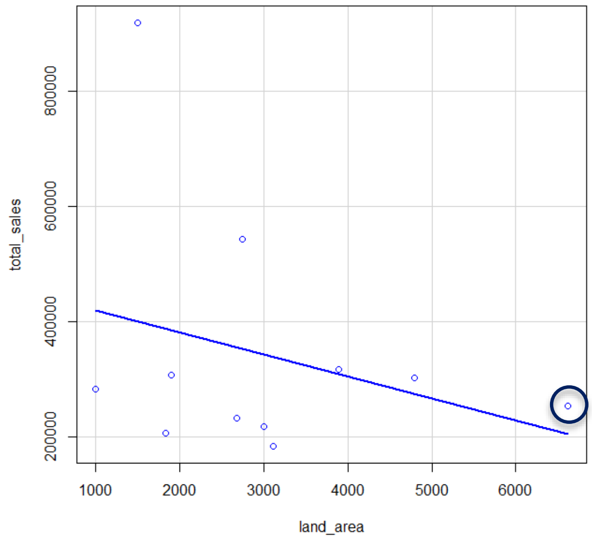
No outliers for Household under 18.

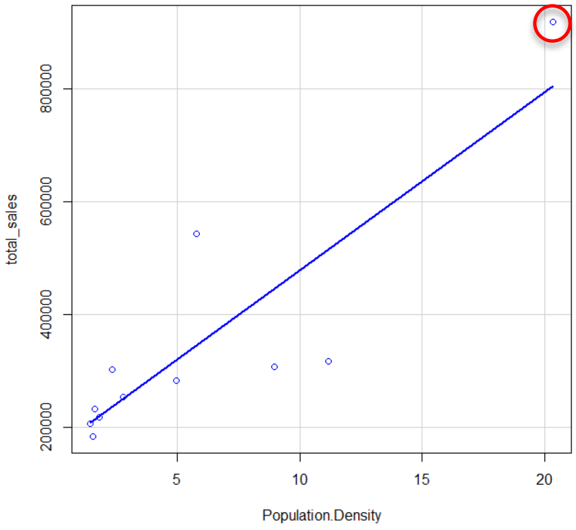


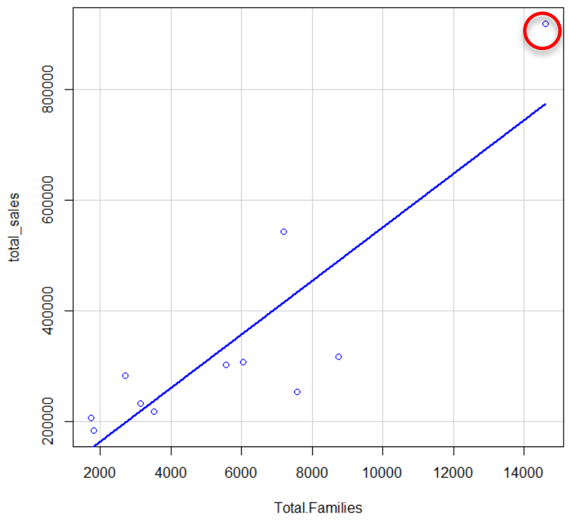
Land Area, one outlier, data set 10, corresponding to city Rock Springs.

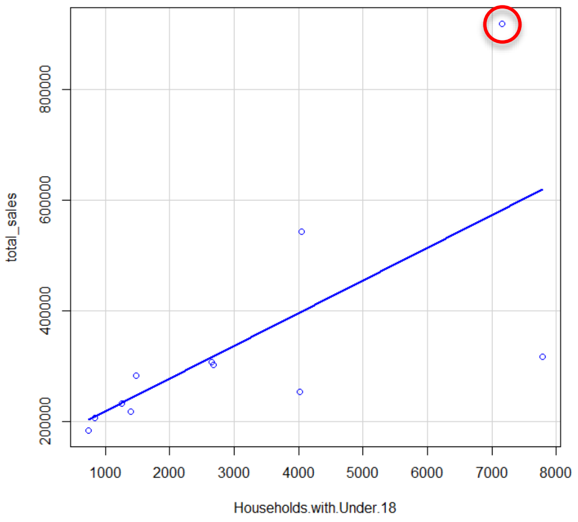
**How to deal with the outliers (only remove or impute one outlier) ?**

Now let us take a close look on the data of the three cities mentioned above, **Gillette, Cheyenne and Rock Springs**. How the total sales correlate with the other variables?
Let us see how a **predictor variable**, **total sales** correlate with other **suspected dependent variables, land area, population density and total families & household under 18**.
We can see from the plots below; the data related to city of **Cheyenne (**marked in red).
**Cheyenne** seems to be **big city** from the data set, will be helpful to predict similar big cities.
**Hence kept in the dataset**.
Similarly, the data related to city of **Rock Springs** (marked in **black**) is very close to the fitted model. **Hence kept in the dataset.**
**The outlier related to city of Gillette, being the predictor variable itself with no logical explanations. There is a possibility that it may affect the future model. hence decided to remove from dataset.**









## 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.