Project 2.1: Data Cleanup

## 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?

As per the Pawdacity data provided on monthly sales, census population, competitor’s sales and population demographics, we need to analyse and recommend the city for pawdacity’s newest store based on predicted yearly sales. We need total sales for year 2010, total 2010 Census population, total population density, total land area, total families and total household and individuals under 18 for each city respectively.

1. What data is needed to inform those decisions?

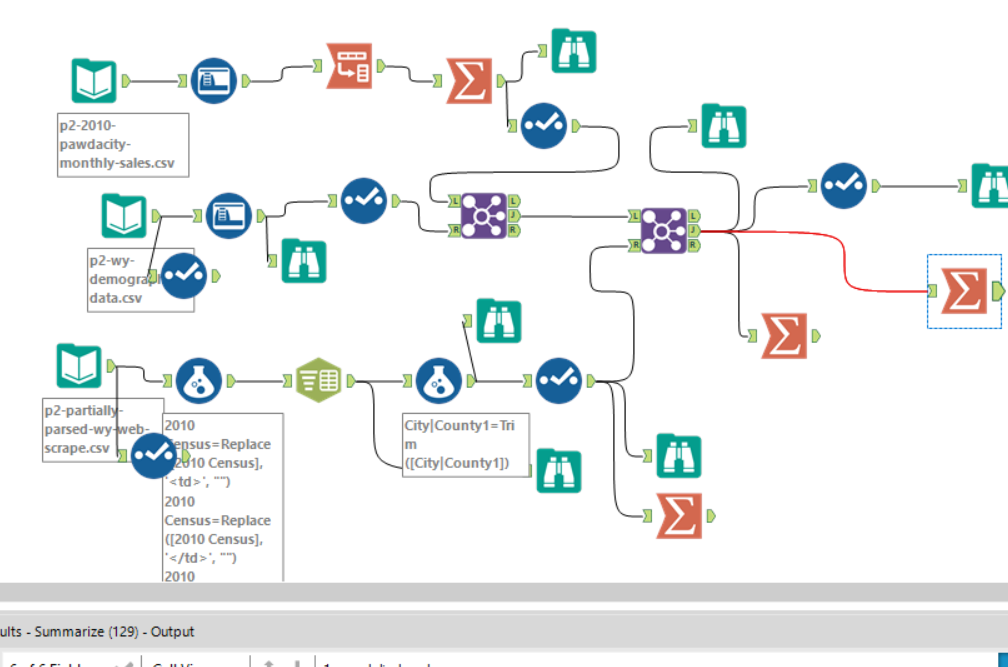
Information needed to inform the decisions are:

* Total Pawdacity sales from monthly sales data provided for year 2010
* Cean data to get Census population
* Join census population, Total sales and population demographic data to get final training data set to build regression model to predict sales
* Get rid of outliers which may affect the model and Pawdacity sale/\.

## 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*



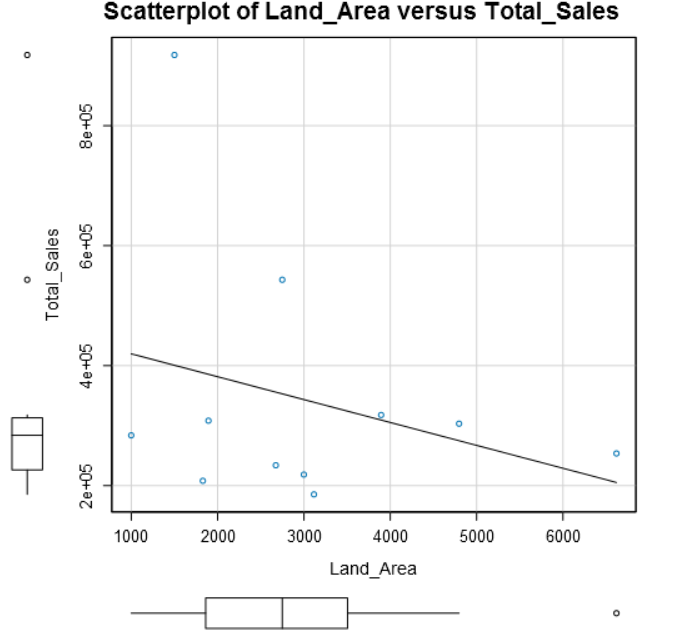
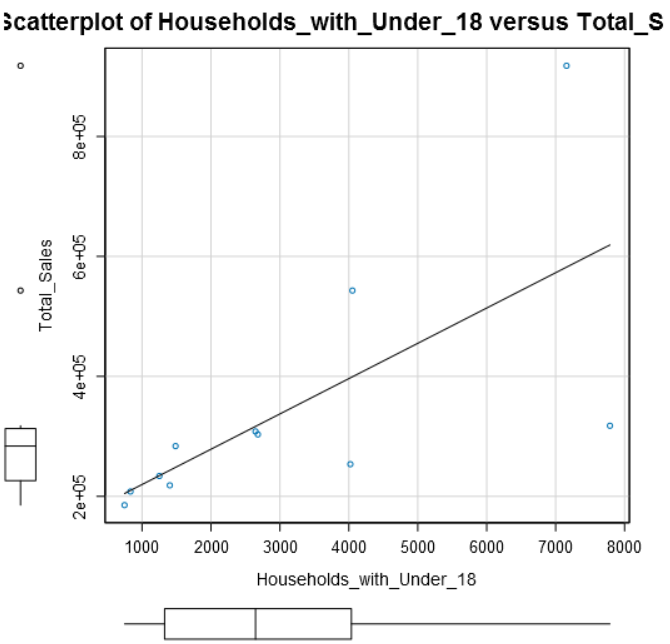
|  |  |  |
| --- | --- | --- |
| **Column** | **Sum** | **Average** |
| *Census Population* | *213,862* | *19442* |
| *Total Pawdacity Sales* | *3,773,304* | *343027.63* |
| *Households with Under 18* | *34,064* | *3096.72* |
| *Land Area* | *33,071* | *3006.45* |
| *Population Density* | *63* | *5.72* |
| *Total Families* | *62,653* | *5695.72* |

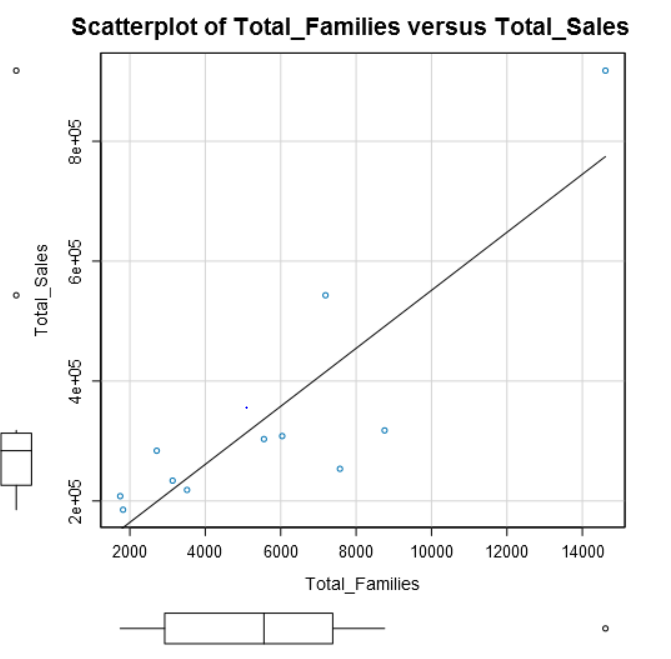
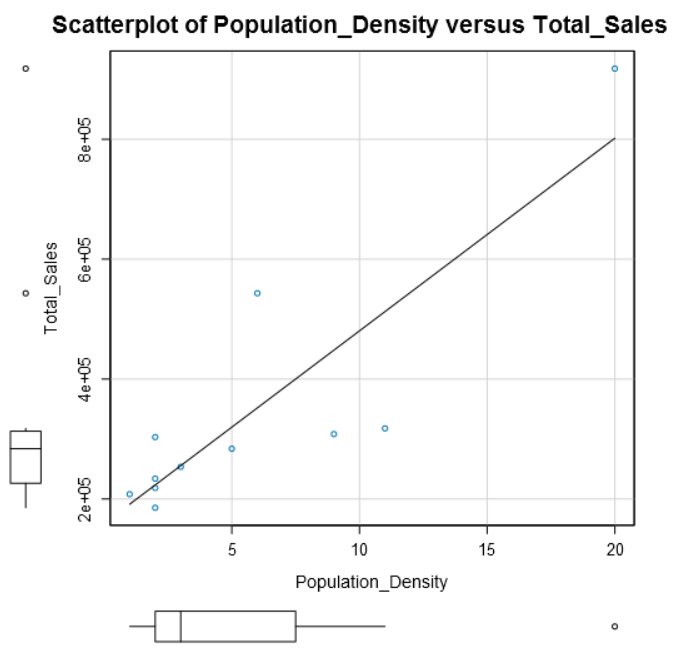
## 

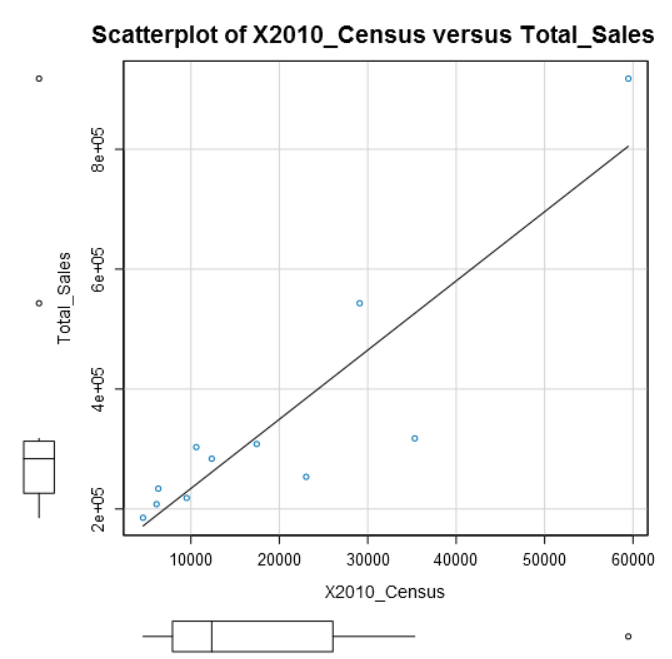
## Step 3: Dealing with Outliers

*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.

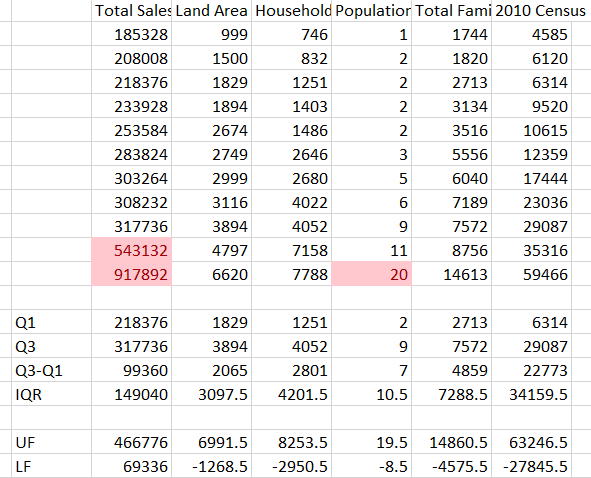
 





According to the scatter plots where target variable is compared with the predictor variable, it seems Cheyenne city is an outlier in almost every variable.

As per the calculation for each variable, by calculating quartiles (Q1,Q3), IQR and upper fence and lower fence, I can point out the outliers.



* It is being analysed that Cheyenne and Gillette city have outliers when seen through Total sales and Population density variables and the rest of the variables do not have any outliers as variable ranges fall with in upper and lower fences.
* Moreover, if we check for city Cheyenne that how much it is contributing in the sales then it is been noticed that maximum sales are coming from that city as it has highest population density, families, census population and households under 18. Therefore, we prefer keeping the Cheyenne city as it has a major effect on the Pawdacity sales and not removing it despite being an outlier.
* If we check Gillette city, it is been identified as an outlier when seen through Total sales and it has medium contribution towards sales so if removed from the list of cities that effect total sales, it will not impact much on the sales.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| CITY | Total Sales | Land Area | Households with Under 18 | Population Density | Total Families | 2010 Census |
| Buffalo | 185328 | 3116 | 746 | 2 | 1820 | 4585 |
| Casper | 317736 | 3894 | 7788 | 11 | 8756 | 35316 |
| Cheyenne | 917892 | 1500 | 7158 | 20 | 14613 | 59466 |
| Cody | 218376 | 2999 | 1403 | 2 | 3516 | 9520 |
| Douglas | 208008 | 1829 | 832 | 1 | 1744 | 6120 |
| Evanston | 283824 | 999 | 1486 | 5 | 2713 | 12359 |
| Gillette | 543132 | 2749 | 4052 | 6 | 7189 | 29087 |
| Powell | 233928 | 2674 | 1251 | 2 | 3134 | 6314 |
| Riverton | 303264 | 4797 | 2680 | 2 | 5556 | 10615 |
| Rock Springs | 253584 | 6620 | 4022 | 3 | 7572 | 23036 |
| Sheridan | 308232 | 1894 | 2646 | 9 | 6040 | 17444 |

* Hence, we will remove Gillette city