**Project 2.1: Data Cleanup**

## **Step 1: Business and Data Understanding**

### Key Decisions:

1. **What decisions needs to be made?**

The Decision which we need to made is **“In which city to open a new Pawdacity Store**”.

1. **What data is needed to inform those decisions?**

* The data needed to inform those decision is the current data from the 11 existing stores.
* The Project details says to use Census Population, Total Pawdacity Sales, Households with under 18, Land Area, Population Density, Total Families.

## **Step 2: Building the Training Set**

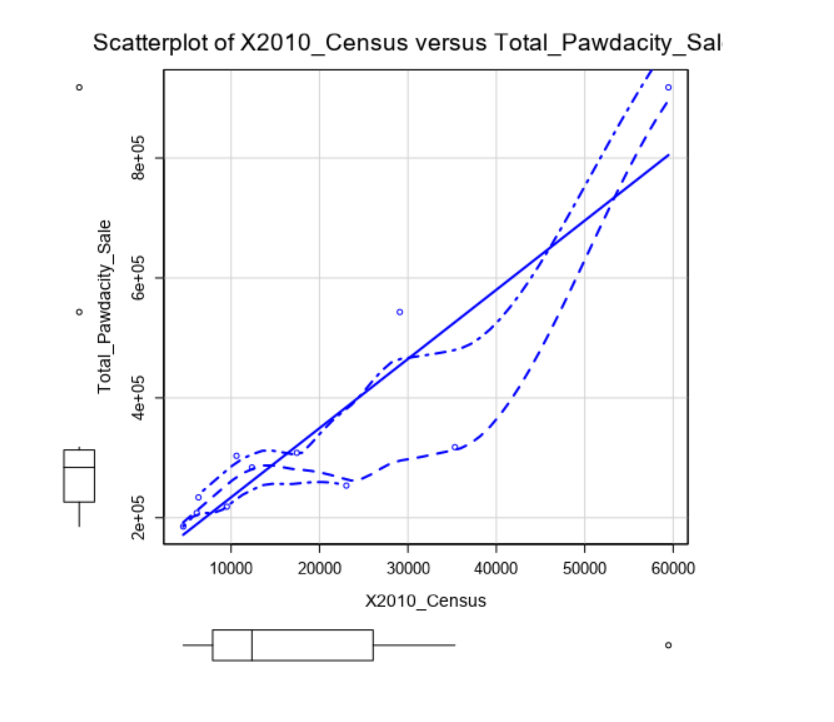
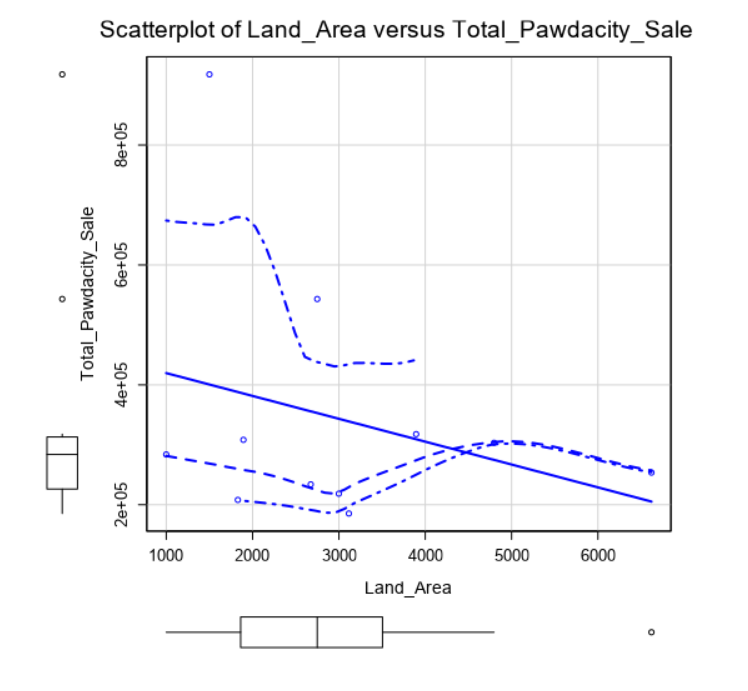
After performing the data cleansing with Alteryx on the given four datasets, the averages for the variables are mentioned below:

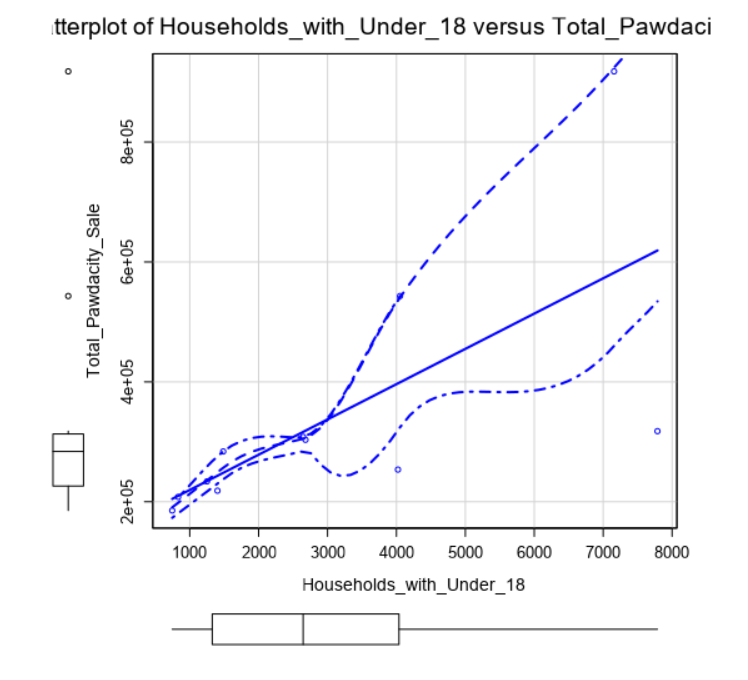
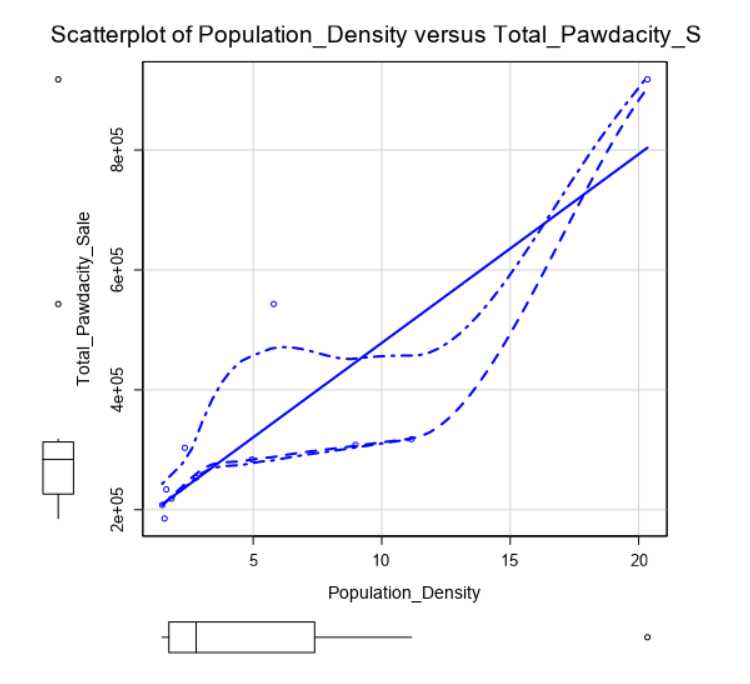
|  |  |  |
| --- | --- | --- |
| **Column** | **Sum** | **Average** |
| **Census Population** | **213,862** | **19,442** |
| **Total Pawdacity Sales** | **3,773,304** | **343,027.64** |
| **Households with Under 18** | **34,064** | **3,096.73** |
| **Land Area** | **33,071** | **3,006.49** |
| **Population Density** | **63** | **5.71** |
| **Total Families** | **62,653** | **5,695.71** |

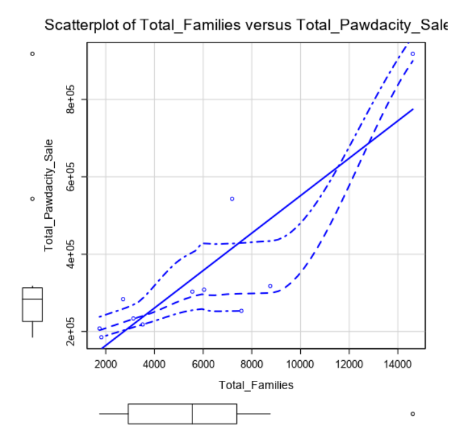
## **Step 3: Dealing with Outliers**

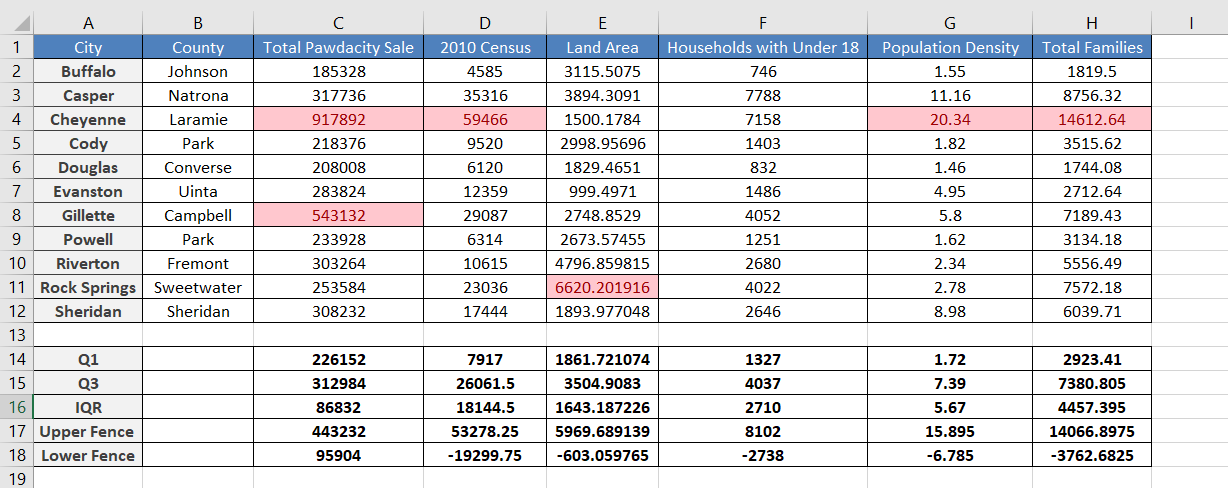
Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.

For better Understanding of Outliers, I have used Scatter Plot of Total Pawdacity sales Vs the other given variables:





From the Scatter Plot above and the Data Extracted, some of the observations are:

* Even though Cheyenne is flagged out as outlier but it is big city compared to the rest of the cities. Cheyenne’s values for the different fields are larger in comparison with other cities even though it has a smaller number of stores.
* Comparing Gillette’s with rest of the cities, its total sales are not in the proportion with other demographic fields such as population. If a city has a large sale, we would expect it to have a large population to drive those sales which isn’t the case.
* For Land Area, Rock Spring was an outlier.

So, after considering all of the above point I will Suggest to remove Gillette from the dataset in order to get unbiased model.

**My Alteryx Work Flow:**

