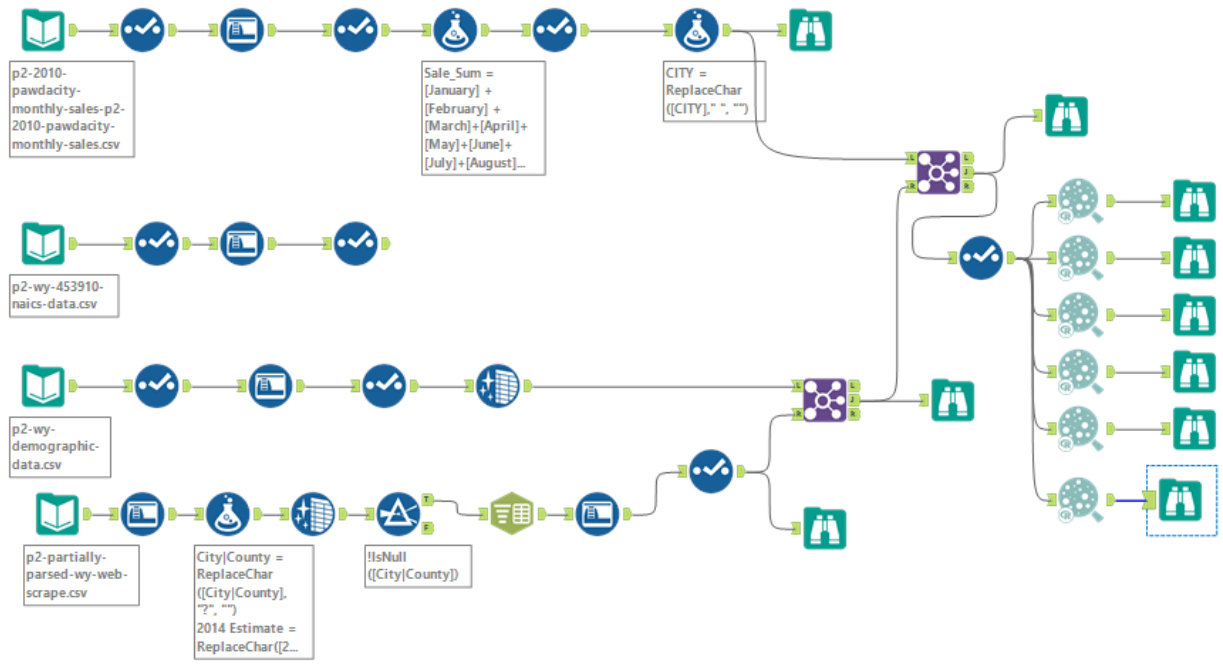
Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Alteryx WorkFlow



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?
   Census Population, Total Pawdacity 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

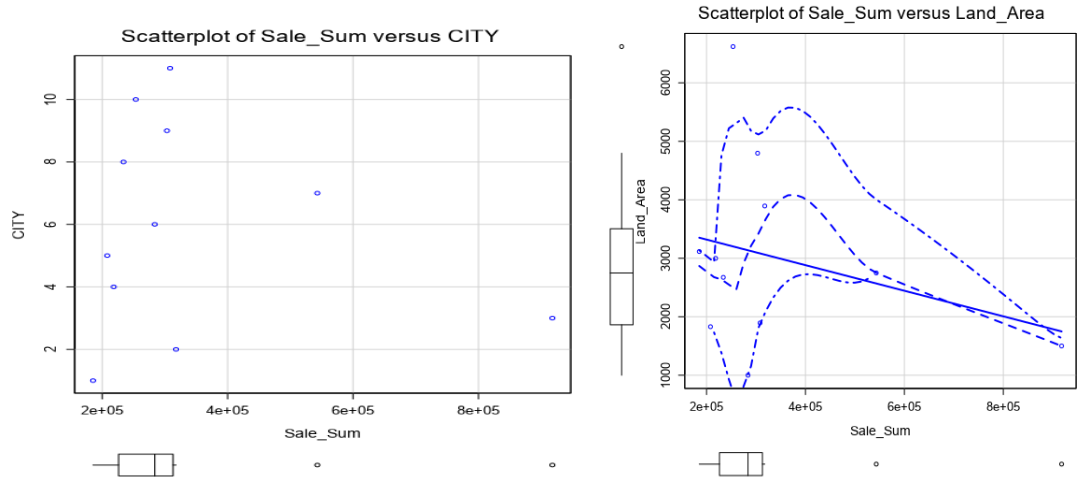
Column	Sum	Average
Census Population	*213,862*	*19442*
Total Pawdacity Sales	*3,773,304*	*343027.64*
Households with Under 18	*34,064*	*3096.73*
Land Area	*33,071*	*3006.49*
Population Density	*63*	*5.71*
Total Families	*62,653*	*5695.71*

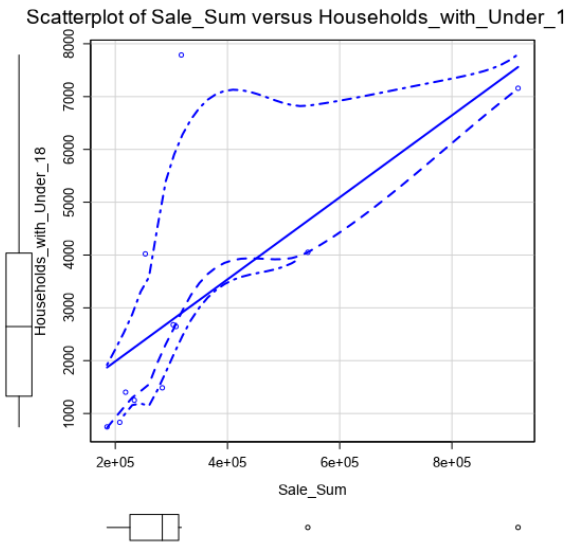
Step 3: Dealing with Outliers

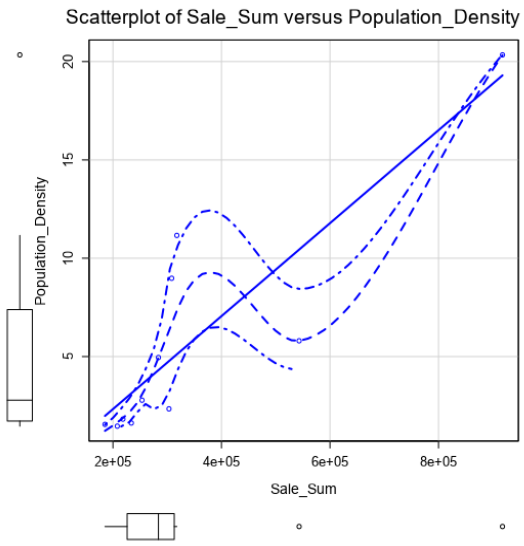
Answer these questions

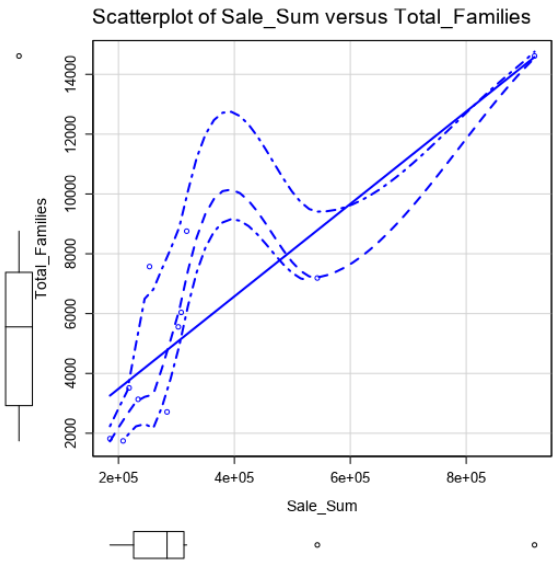
Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.

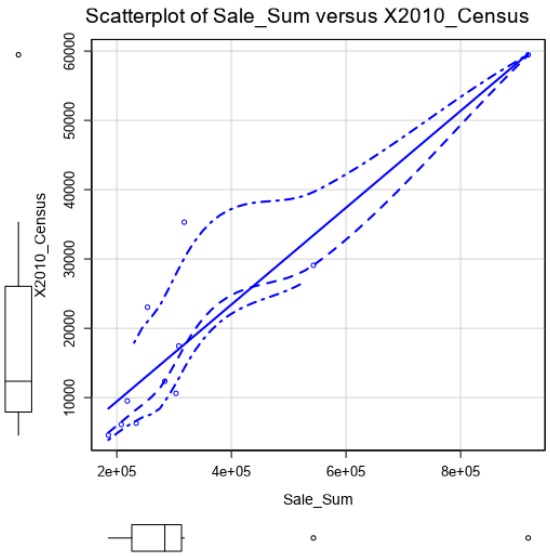
Based on the scatterplots and boxplot below, Gillette would be the outlier in this case when compared against all other cities due to its greatest distance from the linear trend. So we will remove this city as outlier.




Scatterplot of Sale_Sum versus Households_with_Under_1
Households_with_Under_18
1000
2000
3000
4000
5000
6000
7000
8000
2e+05
4e+05
6e+05
8e+05
Sale_Sum


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


Scatterplot of Sale_Sum versus Total_Families
Total_Families
2000
4000
6000
8000
10000
12000
14000
2e+05
4e+05
6e+05
8e+05
Sale_Sum


Scatterplot of Sale_Sum versus X2010_Census
X2010_Census
10000
20000
30000
40000
50000
60000
2e+05
4e+05
6e+05
8e+05
Sale_Sum