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 need to be made?

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

2. What data is needed to inform those decisions?

Data required in order to make an informed decision are given to us in the form of

1.The monthly sales data for all the Pawdacity stores for the year 2010.
2.NAICS data on the most current sales of all competitor stores where total sales are equal to 12 months of sales.
3.A partially parsed data file that can be used for population numbers.
4.Demographic data (Households with individuals under 18, Land Area, Population Density, and Total Families) for each city and county in the state of Wyoming. For people who are unfamiliar with the US city system, a state contains counties and counties contains one or more cities

Step 2: Building the Training Set

Build your training set given the data provided to you. The 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*	
Total Pawdacity Sales	*3,773,304*	
Households with Under 18	*34,064*	
Land Area	*33,071*	
Population Density	*63*	
Total Families	*62,653*	

After performing the data cleansing with Alteryx on the given four datasets, the averages for the variables are mentioned below. The Alteryx workflow for this project is in the image below.

Column	Sum	Average
2010 Census Population	**213,862**	**19442**
Total Pawdacity Sales	**3,773,304**	**343,027.63**
Households with Under 18	**34,064**	**3,096.72**
Land Area	**33,071**	**3,006.48**
Population Density	**63**	**5.70**
Total Families	**62,653**	**5695.70**

Table: Average of variables

My Alteryx workflow

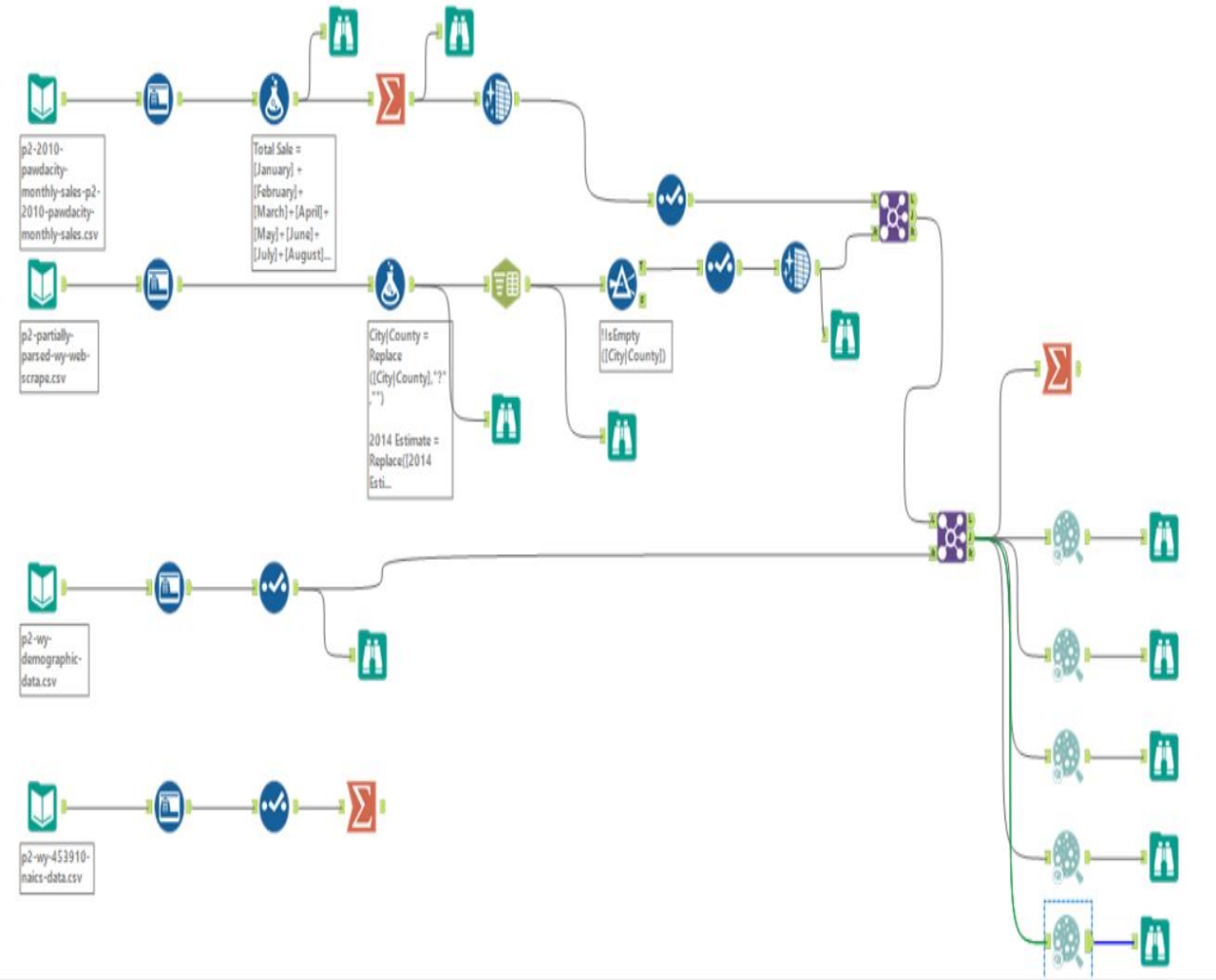
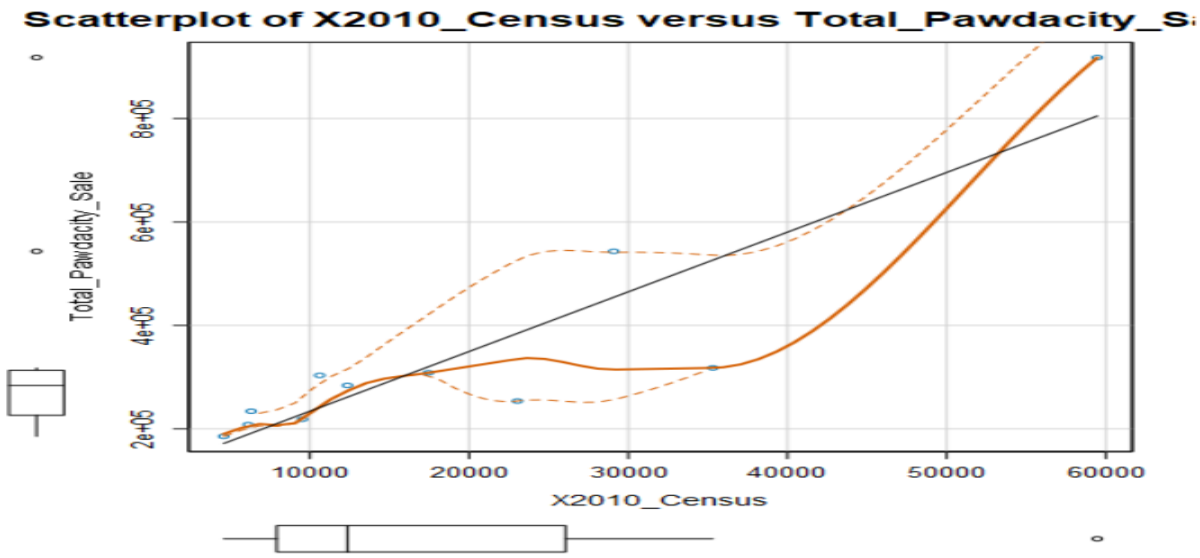


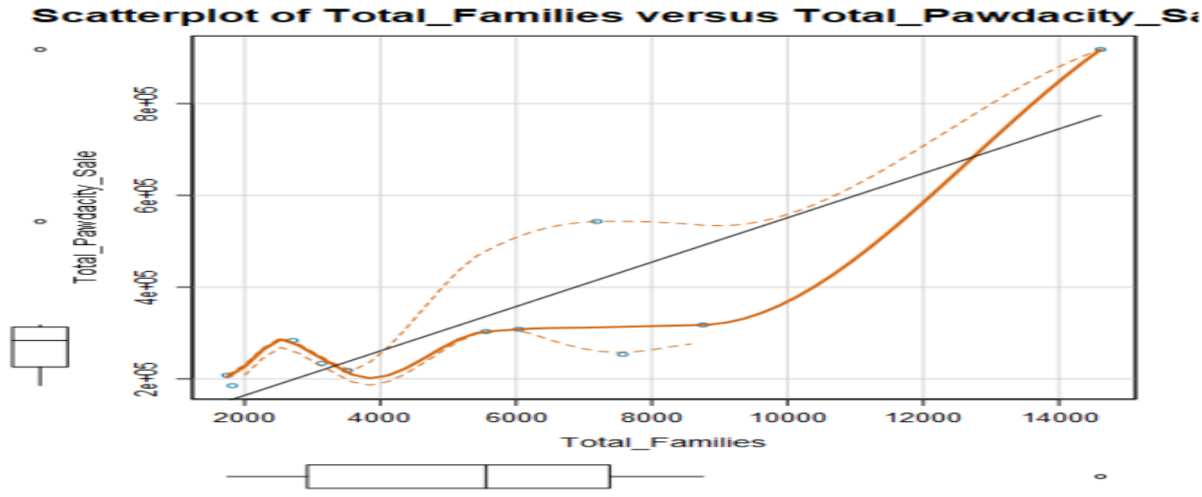
Fig: Alteryx Workflow

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.





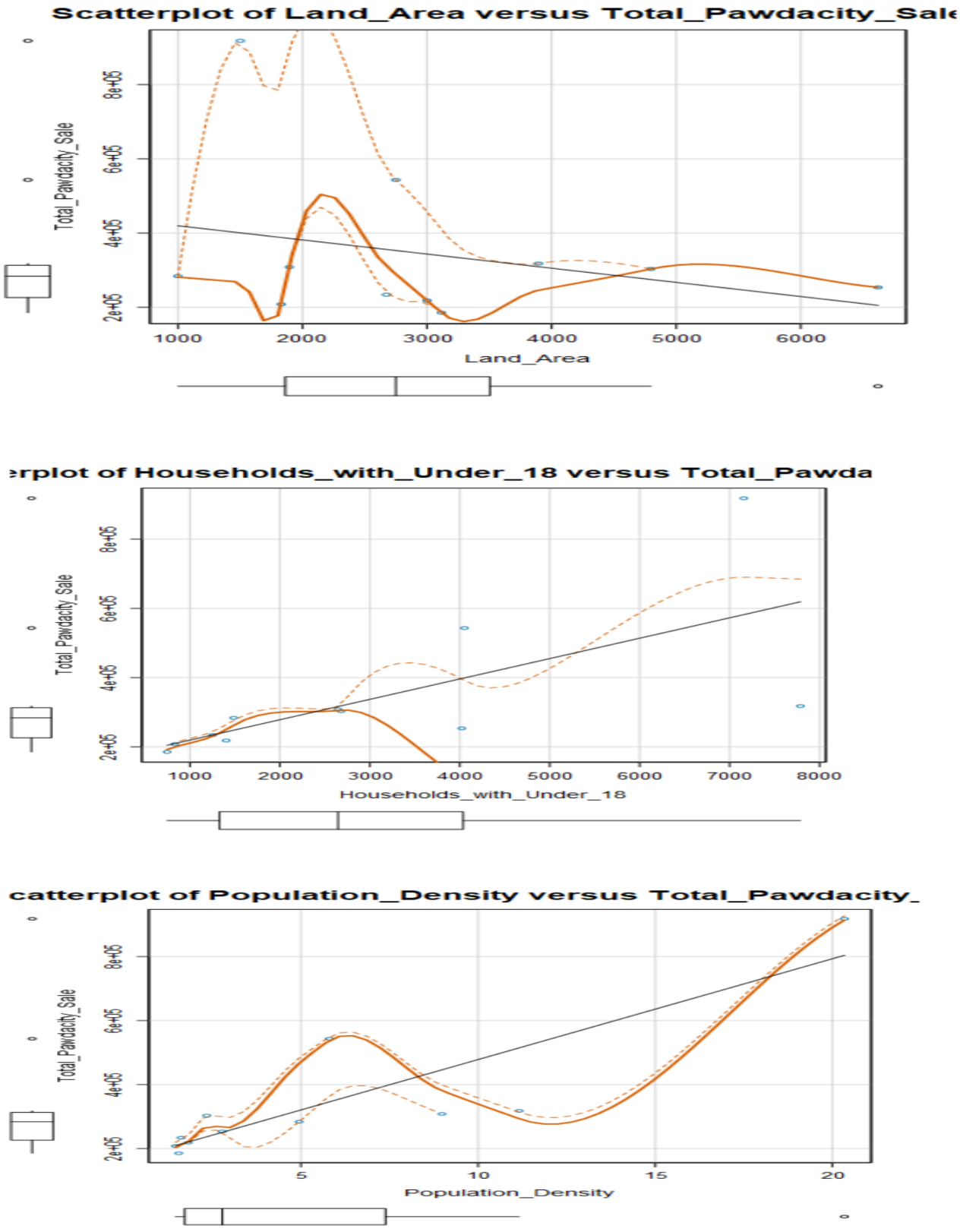


Figure: Scatterplots of Total Pawdacity sales Vs the other given variables

From the scatterplots above, and the data extracted, the city of **Gillette** and **Cheyenne** seems to be the possible outliers as their sales data are higher than the other cities. Few points to keep in mind is that these cities also have a higher number of stores and a larger population as well. 29087 and 59466 respectively. This could be a logical reason as to why these cities have higher sales compared to the rest cities of Wyoming. Other factors such as median income for the population of the cities are not available. The dataset given to us is limited and small. If I had more data to collaborate with I might have ignored these as outliers. However, since that is not the case, and from the plot data available, I believe Gillette's population and total sales are still correlated. I would keep Gillette for analysis as, without it, I believe I will have a skewed result going forward.