Project 2.1: Creating an Analytical Dataset

Step 1: Business and Data Understanding

Pawdacity is a leading pet store chain in Wyoming with 13 stores throughout the state. Pawdacity would like to expand and open a 14th store and we will perform an analysis to recommend the city for Pawdacity's newest store, based on predicted yearly sales.

In this project we are trying to gather, clean and blend datasets together and then deal with outliers.

Key Decisions:

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

   We are trying to predict the location of the 14th store based on predicted yearly sales.

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

   For predicting the location of the new store, we need data related to existing stores and data related to demographic information.

Step 2: Building the Training Set

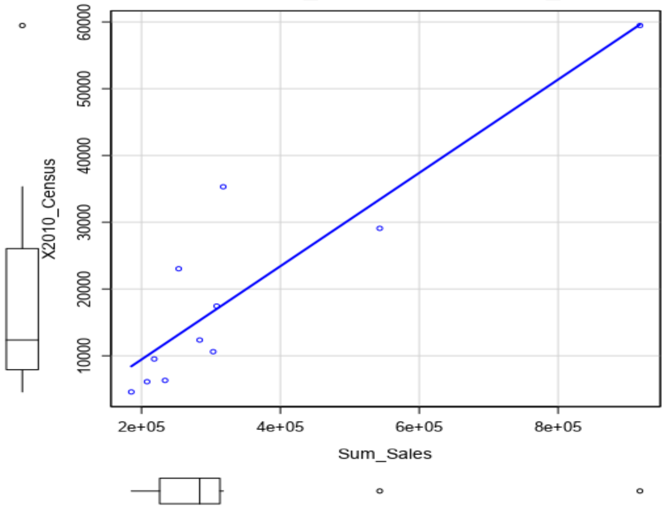
Column	**Sum**	**Average**
Census Population	*213,862*	*19442*
Total Pawdacity Sales	*3,773,304*	*343027.64*
Households with Under 18	*34,064*	*3096.73*
Land Area	*33,071*	*3006.49*
Population Density	*63*	*5.71*
Total Families	*62,653*	*5695.71*

Step 3: Dealing with Outliers

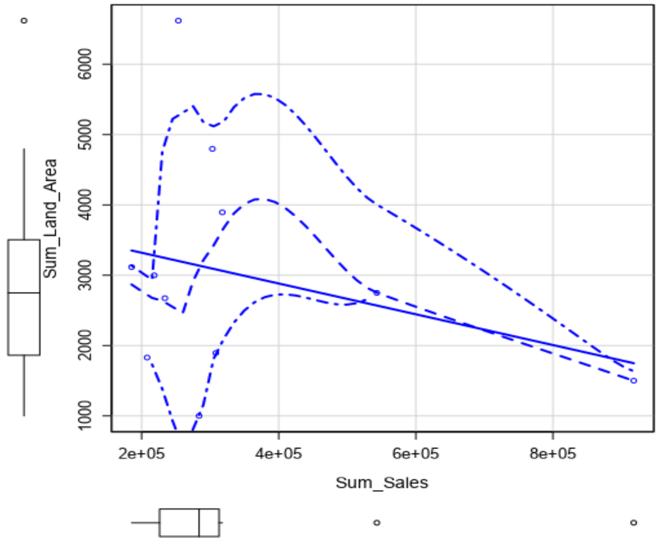
Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.

The scatterplots below show the relation between Pawdacity total sales and other variables. Based on the scatter plots below, Gillette and Cheyenne seem to be outliers as their sales data are higher than expected.

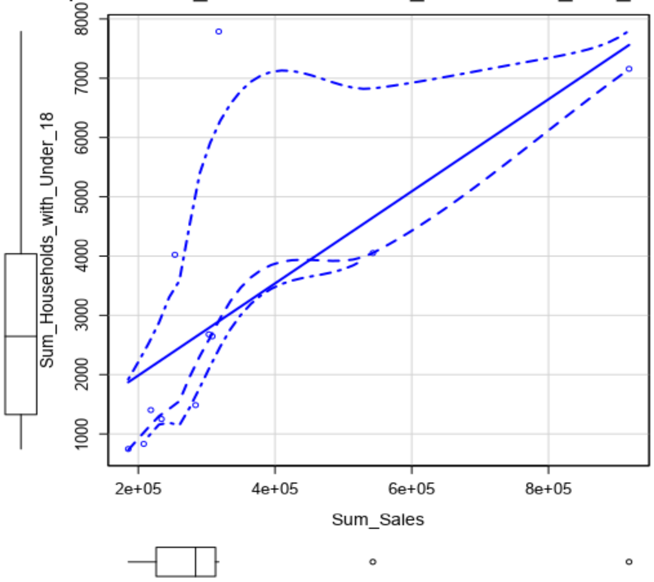
Scatterplot of Sum_Sales versus X2010_Census



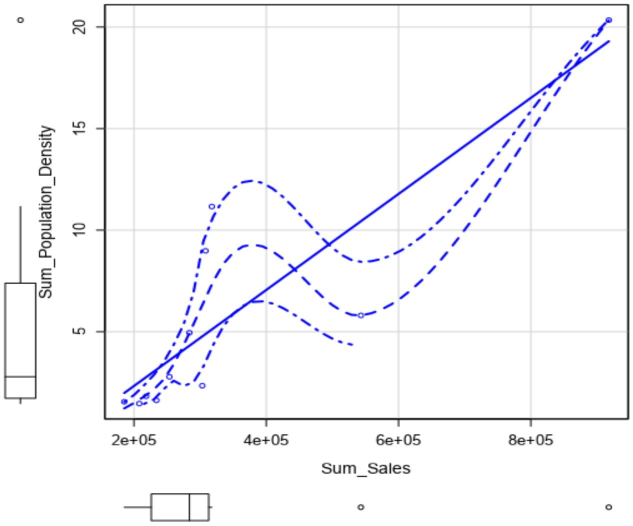
Scatterplot of Sum_Sales versus Sum_Land_Area

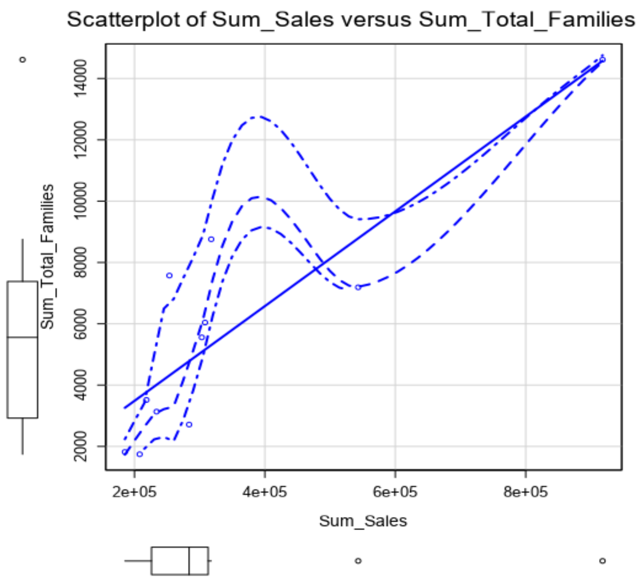


catterplot of Sum_Sales versus Sum_Households_with_Unde



Scatterplot of Sum_Sales versus Sum_Population_Densit

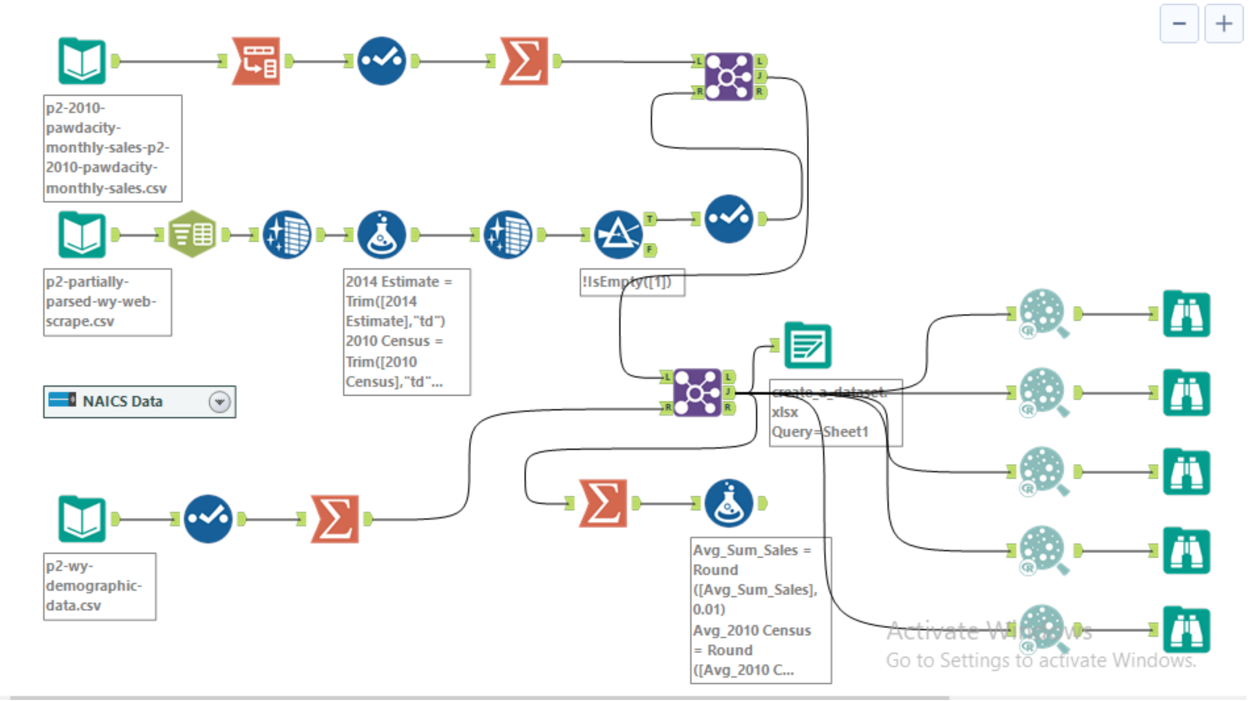




We calculated the IQR using excel and found the below values for upper fence and lower fence.
Upper Fence: 32888.7581
Lower Fence: -17349.9835

Only **Gillette (543132)** and **Cheyenne (917892)** have sales values that exceed the upper fence. Hence, we should remove these outliers.

Alteryx workflow is attached:


−
+
p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales.csv
L
R
L
J
R
p2-partially-parsed-wy-web-scrape.csv
2014 Estimate = Trim([2014 Estimate],"td") 2010 Census = Trim([2010 Census],"td"...
T
F
!IsEmpty([1])
NAICS Data
create_a_dataset.xlsx Query=Sheet1
p2-wy-demographic-data.csv
Avg_Sum_Sales = Round ([Avg_Sum_Sales], 0.01) Avg_2010 Census = Round ([Avg_2010 C...
Activate Windows
Go to Settings to activate Windows.