Jason Grenig Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Pawdacity is a pet store chain in Wyoming with 13 stores and is looking to add a 14th location. In order to determine which new city would be the most profitable to expand into, we need to build a training dataset to fuel a regression model, which will predict the best possible new location.

Key Decisions:

1. What decisions needs to be made?
   We need to determine which new city will be the most profitable for Pawdacity to expand into. In order to reach a decision, we'll need to build a training set, clean and blend the given data, and then create a regression model (which is beyond the scope of this project).

2. What data is needed to inform those decisions?
   To properly build the model, and select predictor variables, we need t build a dataset with the following columns:
   - City
   - 2010 Census Population
   - Total Pawdacity Sales
   - Households with Under 18
   - Land Area
   - Population Density
   - Total Families

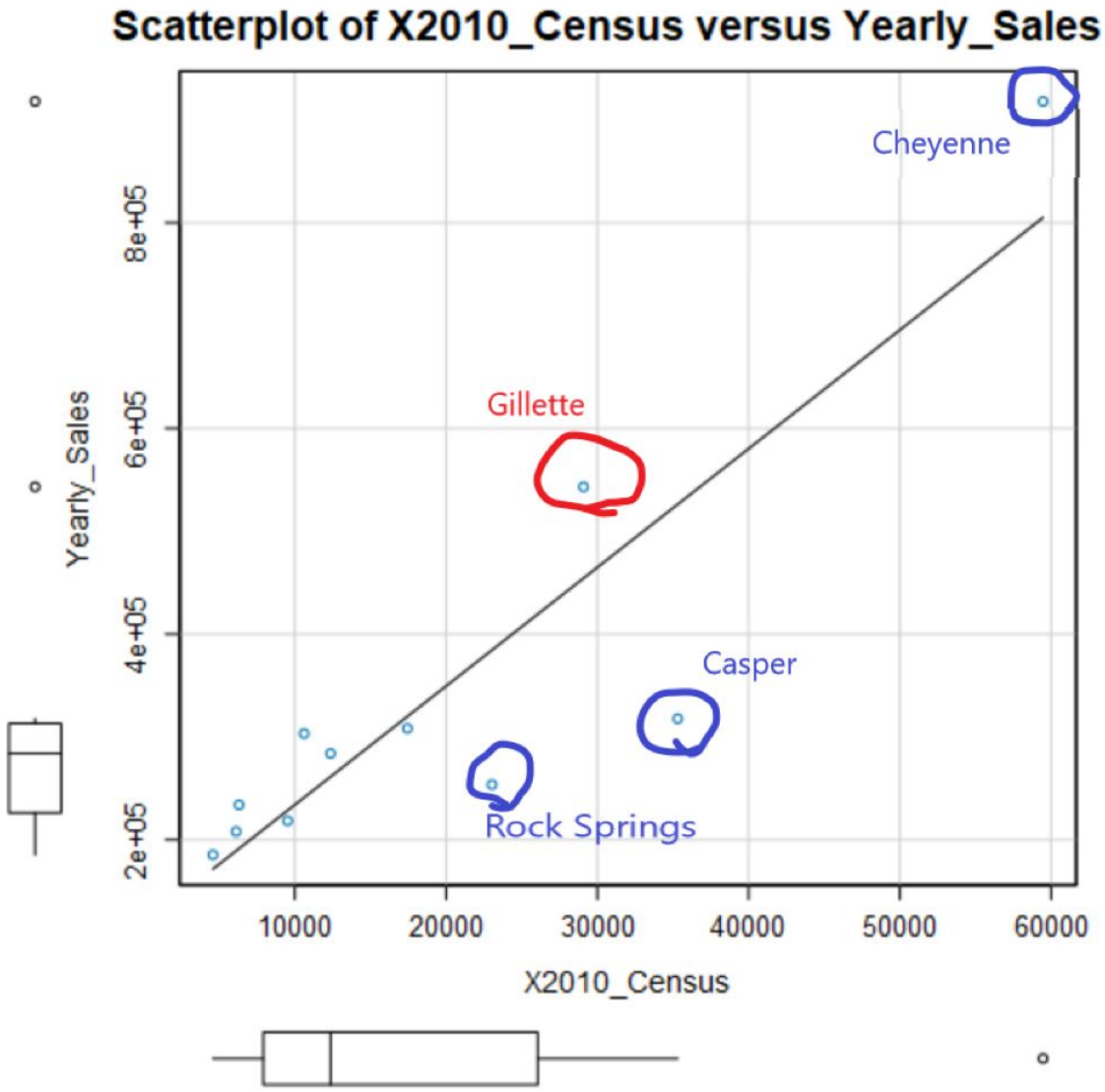
Step 2: Building the Training Set

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.

I decided to remove the city of Gillette because its yearly sales total was extraordinarily high, relative to its size.



Let's compare Gillette against its nearest population neighbors, Casper and Rock Springs:
Casper - Census Population = 35,316; Yearly Sales = $317,736
Gillette - Census Population = 29,087; Yearly Sales = $543,132
Rock Springs - Census Population = 23,036; Yearly Sales = $253,854

We can see that yearly sales for Gillette are nearly 1.75 times that of Casper, despite having a population that's 18% less than Casper. The sales data for Gillette could skew our model, so it should be removed from the training data.

Below is a screen shot of my Alteryx workflow, including summaries with and without the city of Gillette removed from the training data: