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 needs to be made?

   Pawdacity is a leading pet store and would like to expand its business by opening its 14th store in Wyoming. This expansion needs to be determined in which city within Wyoming will be best suited for Pawdacity using the demographic dataset.
   We need to see if the data is in proper format, is missing values, is clean or needs to be cleaned, needs blending using joins or unions and needs to be arranged in according to the need.

2. What data is needed to inform those decisions?

   The following data is needed to inform the decisions – 2010 Census Population, Total Pawdacity Sales, Household with under 18, Land Area, Population Density, Total Families, Sales per month in Pawdacity and competitor sales.

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*	*19,442*
Total Pawdacity Sales	*3,773,304*	*34,3027.64*
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

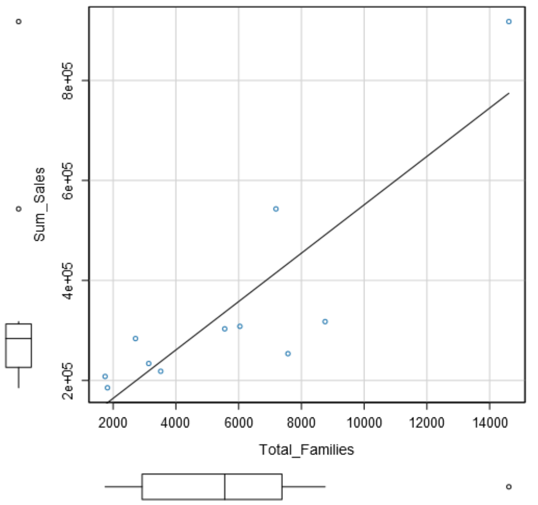
Answer these questions

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.

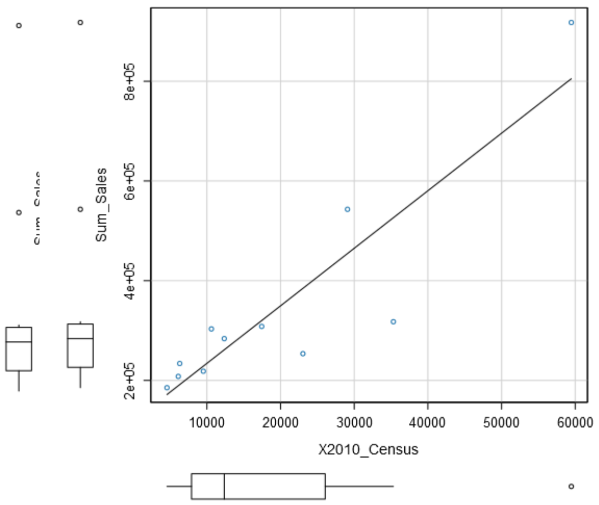
Answer. As we can see from the Scatterplots shown below, Gillette and Cheyenne are the cities which are outliers in the training data as their sales are higher than the expected sales and do not belong in the range.
But, when we extend the data (extrapolate), Cheyenne sales tend to fall in the expected range. Also, Cheyenne has high population density, thus high sales are justified by this. So, we are left with Gillette for which the data is not as expected and will be the outlier in this case. Gillette is a true anomaly as it demographics are in expected range but Pawdacity sales are really high which doesn't

follow the trend of higher number of people, bigger volume of sales as Gillette is a small city yet has higher sales. Thus, we have to remove it.

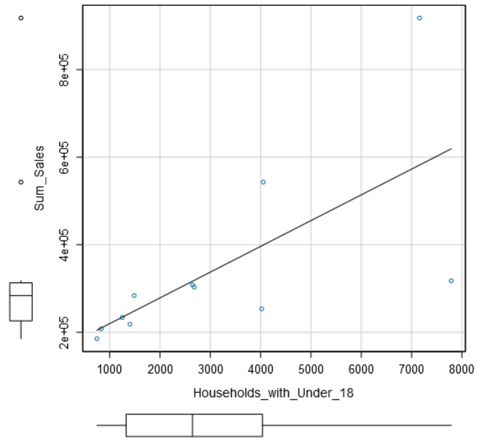
Scatterplot of Total_Families versus Sum_Sales



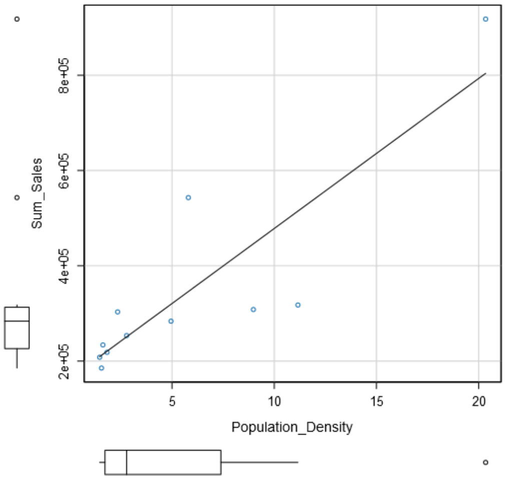
Scatterplot of X2010_Census versus Sum_Sales



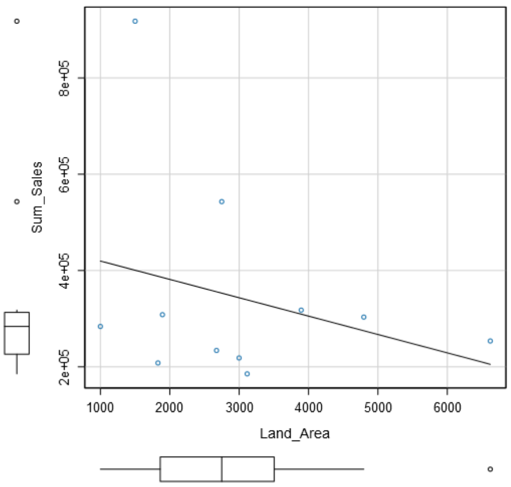
Scatterplot of Households_with_Under_18 versus Sum_Sa



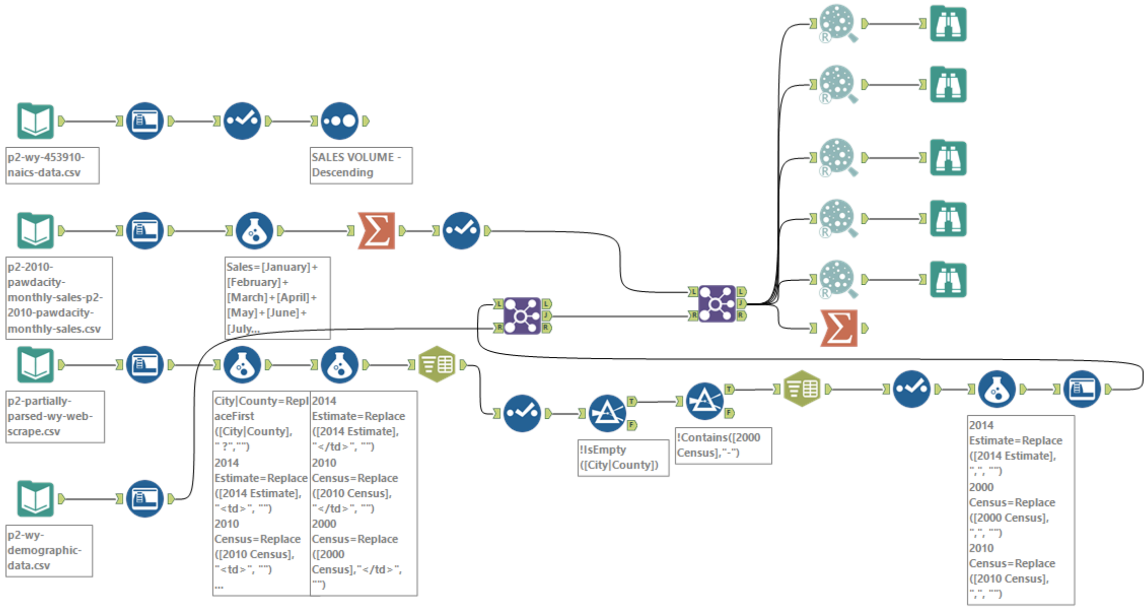
Scatterplot of Population_Density versus Sum_Sales



Scatterplot of Land_Area versus Sum_Sales



Workflow Diagram


p2-wy-453910-naics-data.csv
SALES VOLUME - Descending
p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales.csv
Sales=[January]+[February]+[March]+[April]+[May]+[June]+[July...
p2-partially-parsed-wy-web-scrape.csv
City|County=ReplaceFirst([City|County], " ?","")
2014 Estimate=Replace([2014 Estimate], "<td>", "")
2010 Census=Replace([2010 Census], "<td>", "")
...
2014 Estimate=Replace([2014 Estimate], "</td>", "")
2010 Census=Replace([2010 Census], "</td>", "")
2000 Census=Replace([2000 Census],"</td>", "")
p2-wy-demographic-data.csv
!IsEmpty([City|County])
!Contains([2000 Census],"-")
2014 Estimate=Replace([2014 Estimate], ",", "")
2000 Census=Replace([2000 Census], ",", "")
2010 Census=Replace([2010 Census], ",", "")