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.
(Hi, as I am not from business background, please guide me thoroughly how to tackle these kinds of problems effectively if wrong. Thank you.)

1. What decisions needs to be made?

Aim: To expand and open a 14th store for Pawdacity, a leading pet store chain in Wyoming.
Decision: To perform an analysis to recommend the city for Pawdacity's newest store, based on predicted yearly sales.

2. What data is needed to inform those decisions?

Yearly sales.
Given the available data which act as predictor variables, they are ***2010 Census Population, Households with Under 18, Land Area, Population Density and 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.

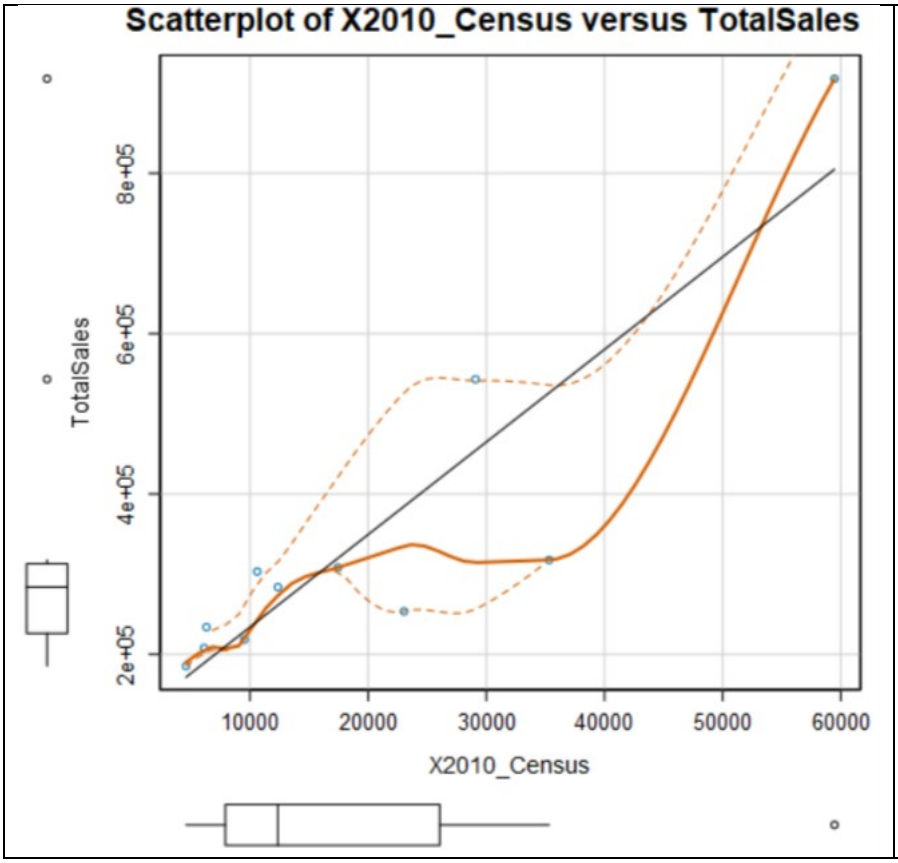
In addition provide the averages on your data set here to help reviewers check your work. You should round up to two decimal places, ex: 1.24

Column	Sum	Average
2010 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

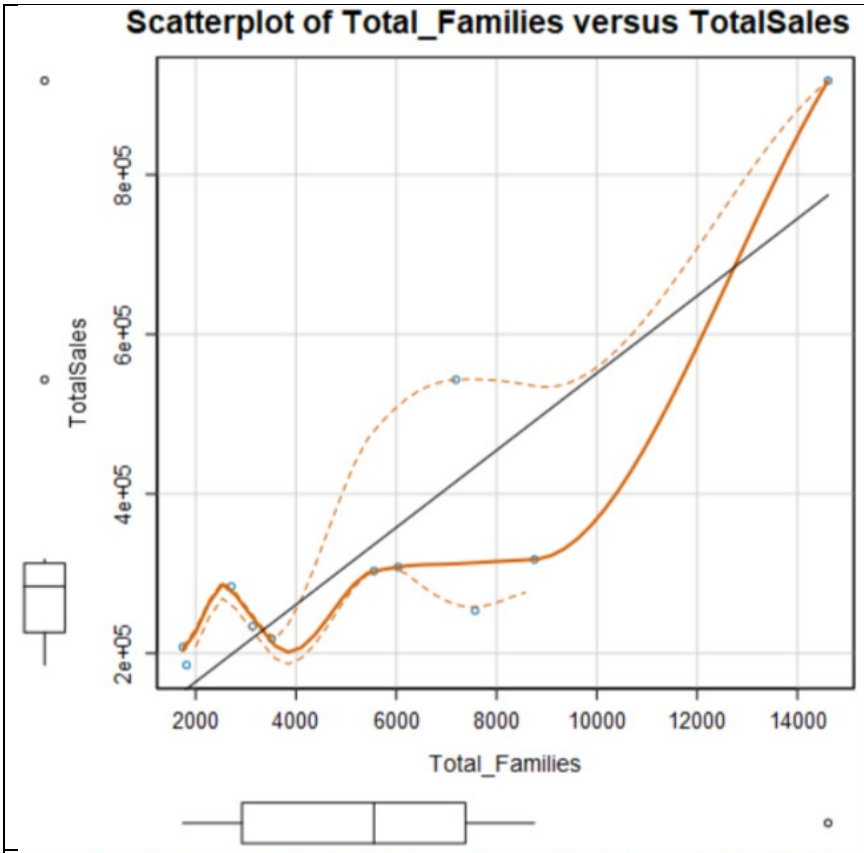
Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.



This is the generated scatterplot for **Yearly Sales (y-axis) against 2010_Census_Population (x-axis).**

From the graph, we can see that there is one outlier that has a value near 60,000 for 2010_Census_Population.

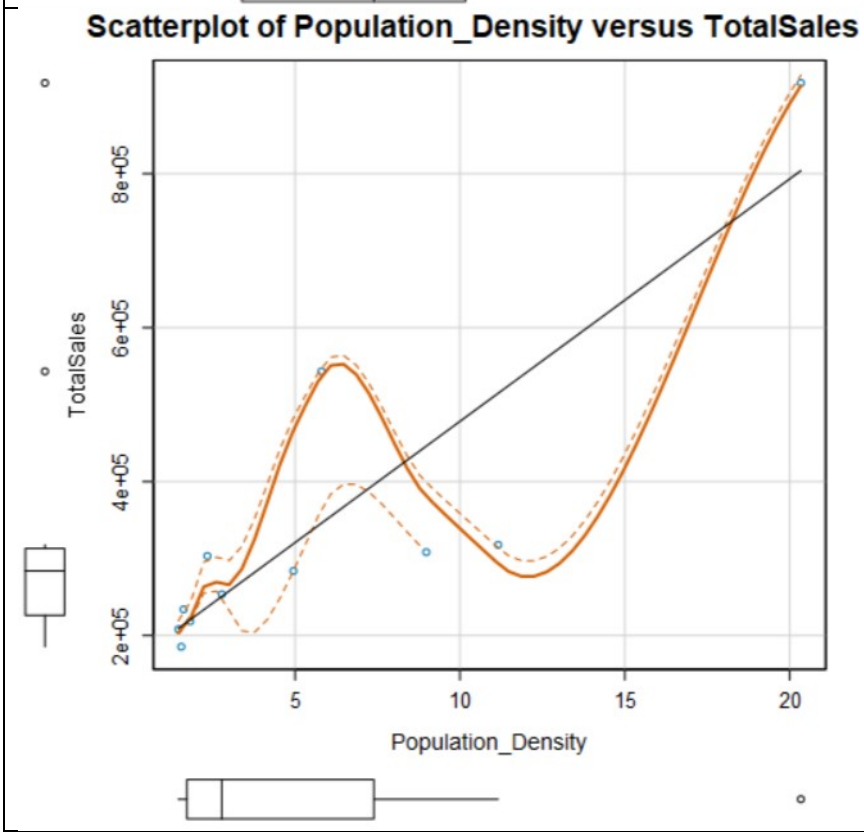
This outlier belongs to **Cheyenne City**.



This is the generated scatterplot for **Yearly Sales (y-axis) against Total Families (x-axis).**

From the graph, we can see that there is one outlier that has a value more than 14,000 for Total Families.

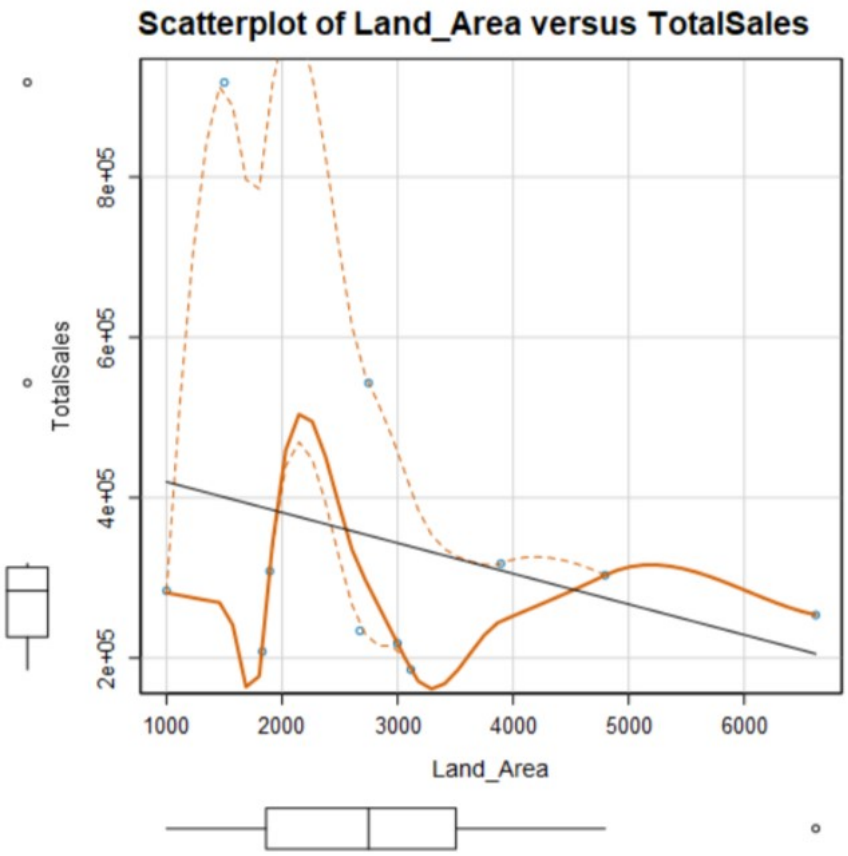
This outlier belongs to **Cheyenne City**.



This is the generated scatterplot for **Yearly Sales (y-axis) against Population Density (x-axis).**

From the graph, we can see that there is one outlier that has a value more than 20 for Population Density.

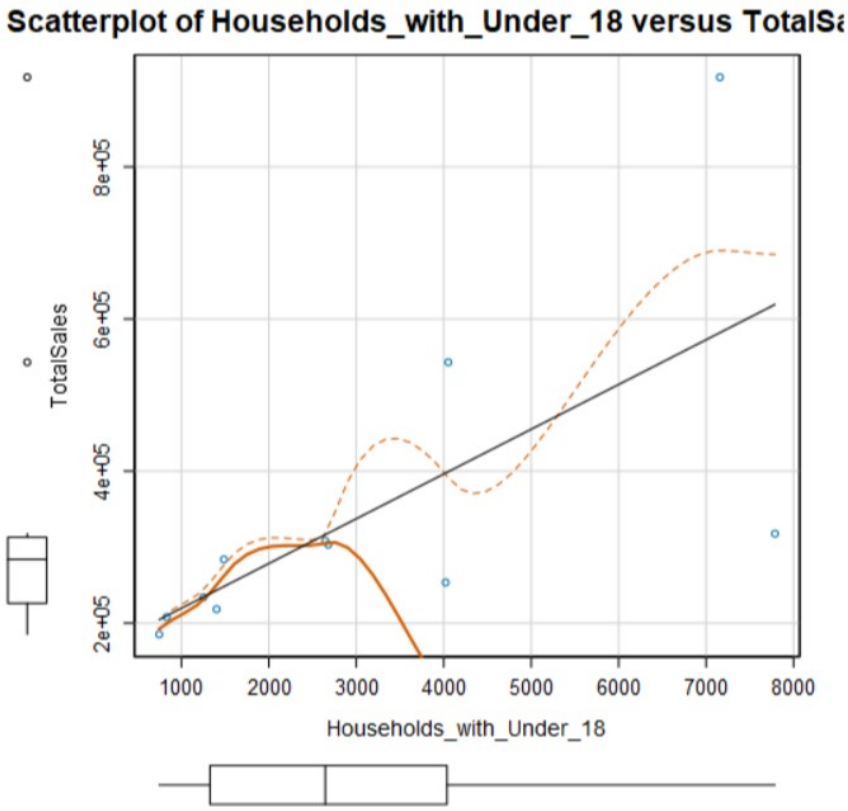
This outlier belongs to **Cheyenne City**.



This is the generated scatterplot for **Yearly Sales (y-axis) against Land Area (x-axis).**

From the graph, we can see that there is one outlier that has a value more than 6,000 for Land Area.

This outlier belongs to **Rock Springs City**.



This is the generated scatterplot for **Yearly Sales (y-axis) against Households with under 18 (x-axis).**

From the graph, we can see that there is no outlier.

I would choose to remove outlier **Cheyenne City** only since our data is small (11 cities). It is

very much away from the upper fence as compared to **Rock Springs City**.