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 chain in Wyoming with 13 stores throughout the state. Pawdacity would like to know where to open the 14th store.
2. What data is needed to inform those decisions?
   The data required in order to inform this decision are ***city*****,** ***2010 census population*****,** ***Pawdacity sales in other stores*****,** ***competitor sales*****,** ***household 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

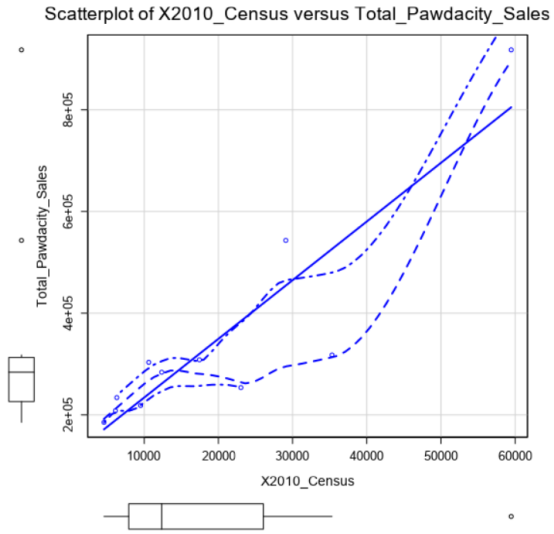
Column	**Sum**	**Average**
Census Population	*213,862*	*19442*
Total Pawdacity Sales	*3,773,304*	*343028*
Households with Under 18	*34,064*	*3097*
Land Area	*33,071*	*3006*
Population Density	*63*	*6*
Total Families	*62,653*	*5696*

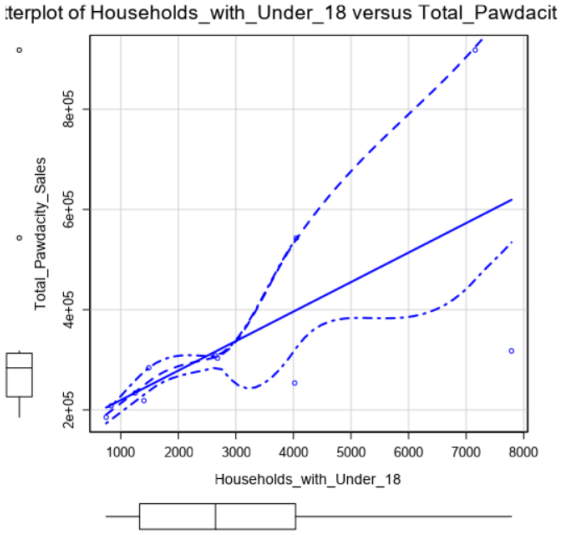
Step 3: Dealing with Outliers

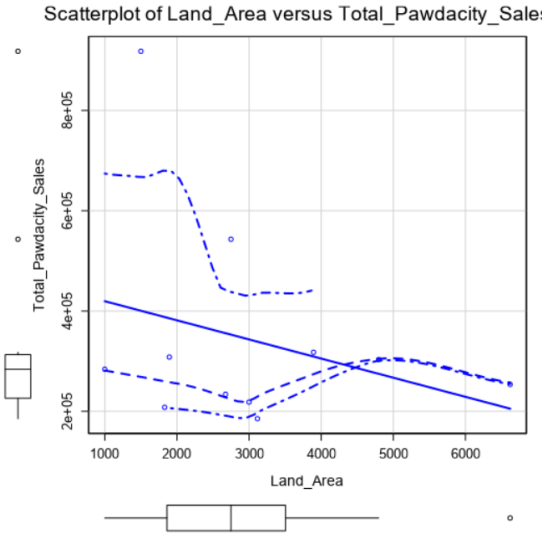
Answer these questions

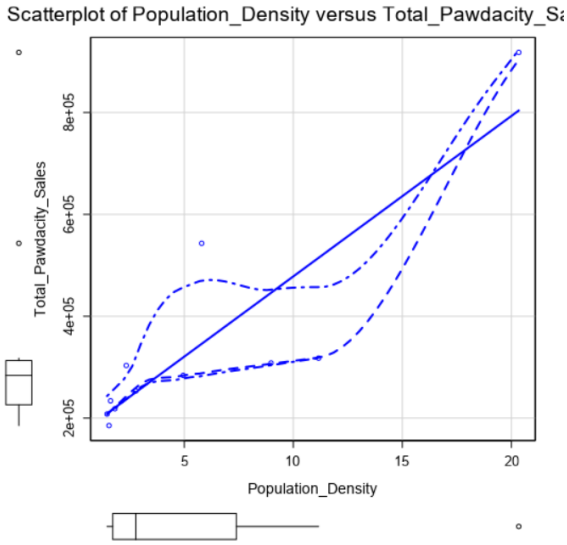
Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.

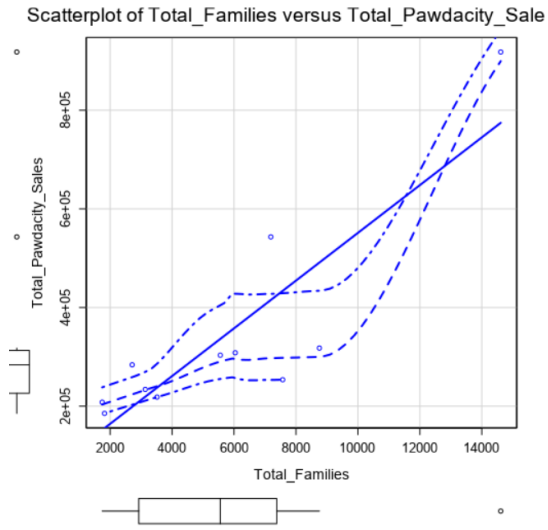
The scatter plots of Census Population, Households with Under 18, Land Area, Population Density, Total Families vs. Total Pawdacity Sales by each city:











The calculated Q1, Q3, IQR Upper and Lower Fence:

City	Total Pawdacity Sal	2010 Cens	Land Area	Households with Under 1	Population Densi	Total Famili
Q1	226152	7917	1861.721069	1327	1.720000029	2923.409912
Q3	312984	26061.5	3504.908325	4037	7.389999866	7380.805176
IQR	86832	18144.5	1643.187256	2710	5.669999838	4457.395264
Upper fence	443232	53278.25	5969.689209	8102	15.89499962	14066.89807
Lower fence	95904	-19299.75	-603.059814	-2738	-6.784999728	-3762.68298
Average	343028	19442	3006	3097	6	5696

It looks like the total Sales data of Cheyenne and Gillette city were higher than expected, due to its distance to the linear trending lines. However, then we might conclude that Gillete and Cheyenne are the outliner in this case.
Other than that, Gillete other data relating to population still seems correlated, then we can keep Gillete in this analysis.
Then I decided to remove the Cheyenne data and keep Gillete's for further analysis.

The Altertyx workflow: