Project 2.1: Data Cleanup

Business and Data Understanding

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

   Pawdacity is a leading pet store chain in Wyoming with 13 stores throughout the state. This year, Pawdacity would like to expand and open a 14th store. The aim of this project is to recommend the city for Pawdacity's newest store, based on predicted yearly sales.

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

   p2-2010-pawdacity-monthly-sales.csv,
   p2-partially-parsed-wy-web-scrape.csv,
   p2-wy-453910-naics-data.csv.

   We need to work out what data from the above files will be necessary to predict where our next store should be.

   We will need to extract the following columns of data from the above files:

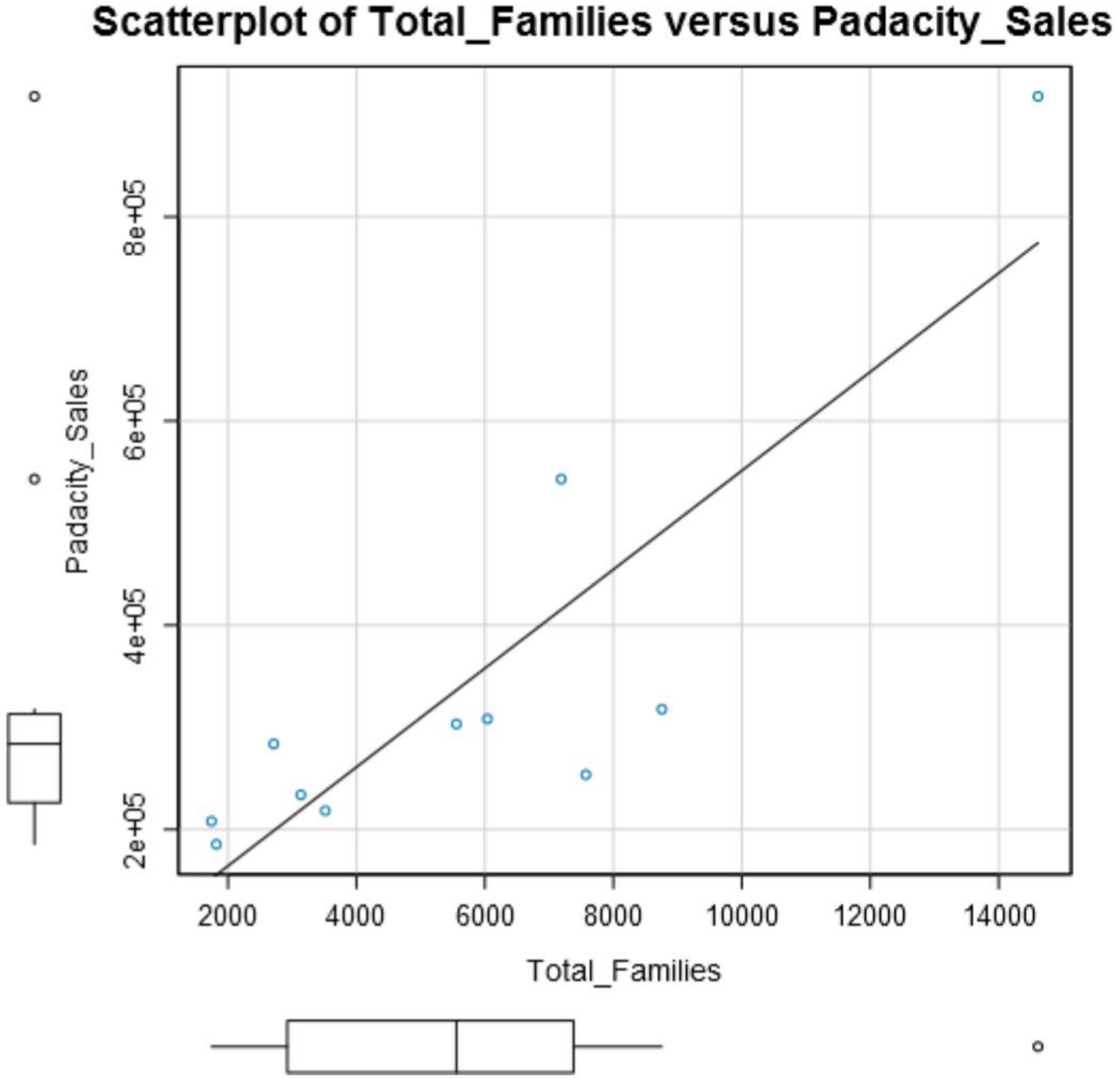
City
2010 Census Population
Total Pawdacity Sales
Households with under 18
Land Area
Population Density
Total Families

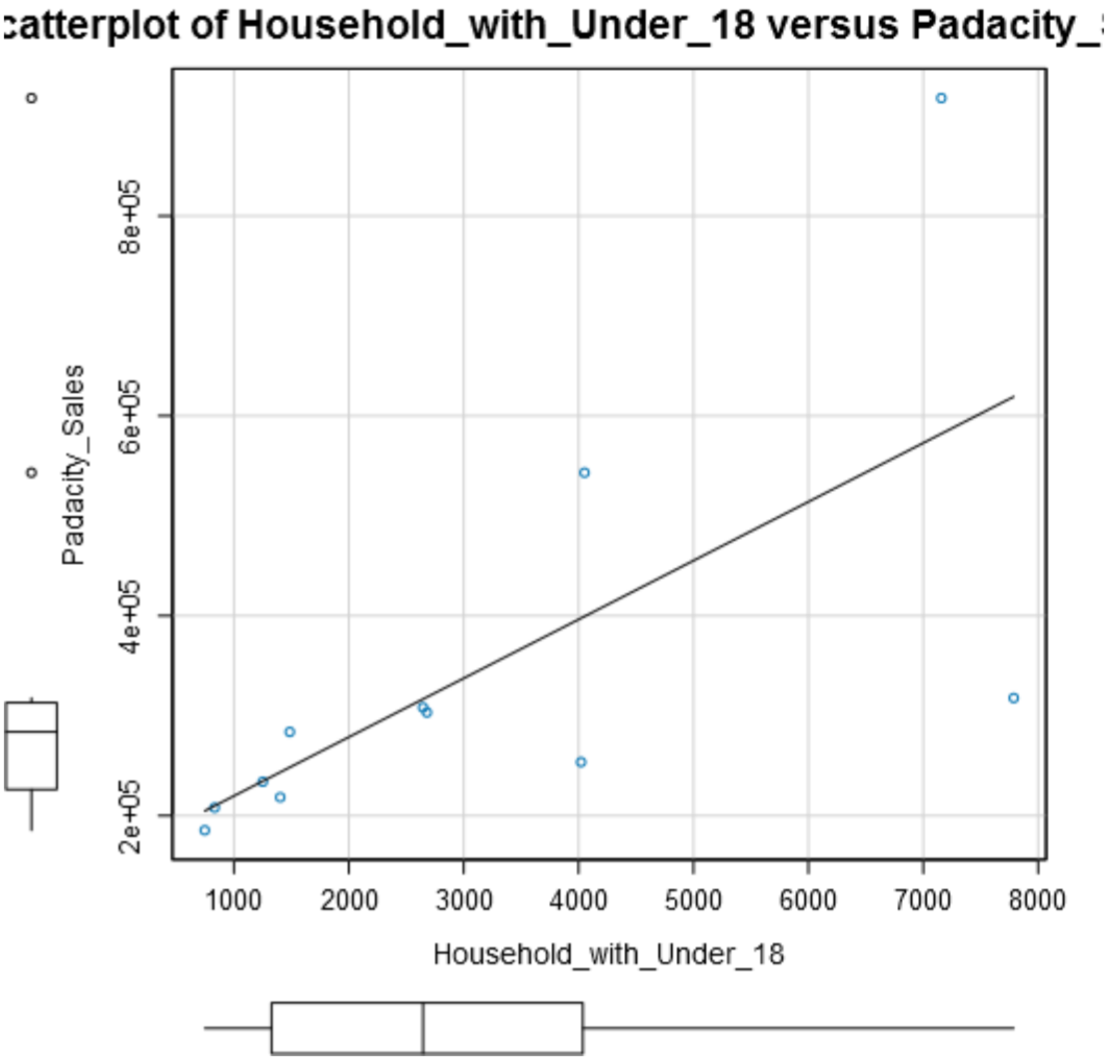
Building the Training Set

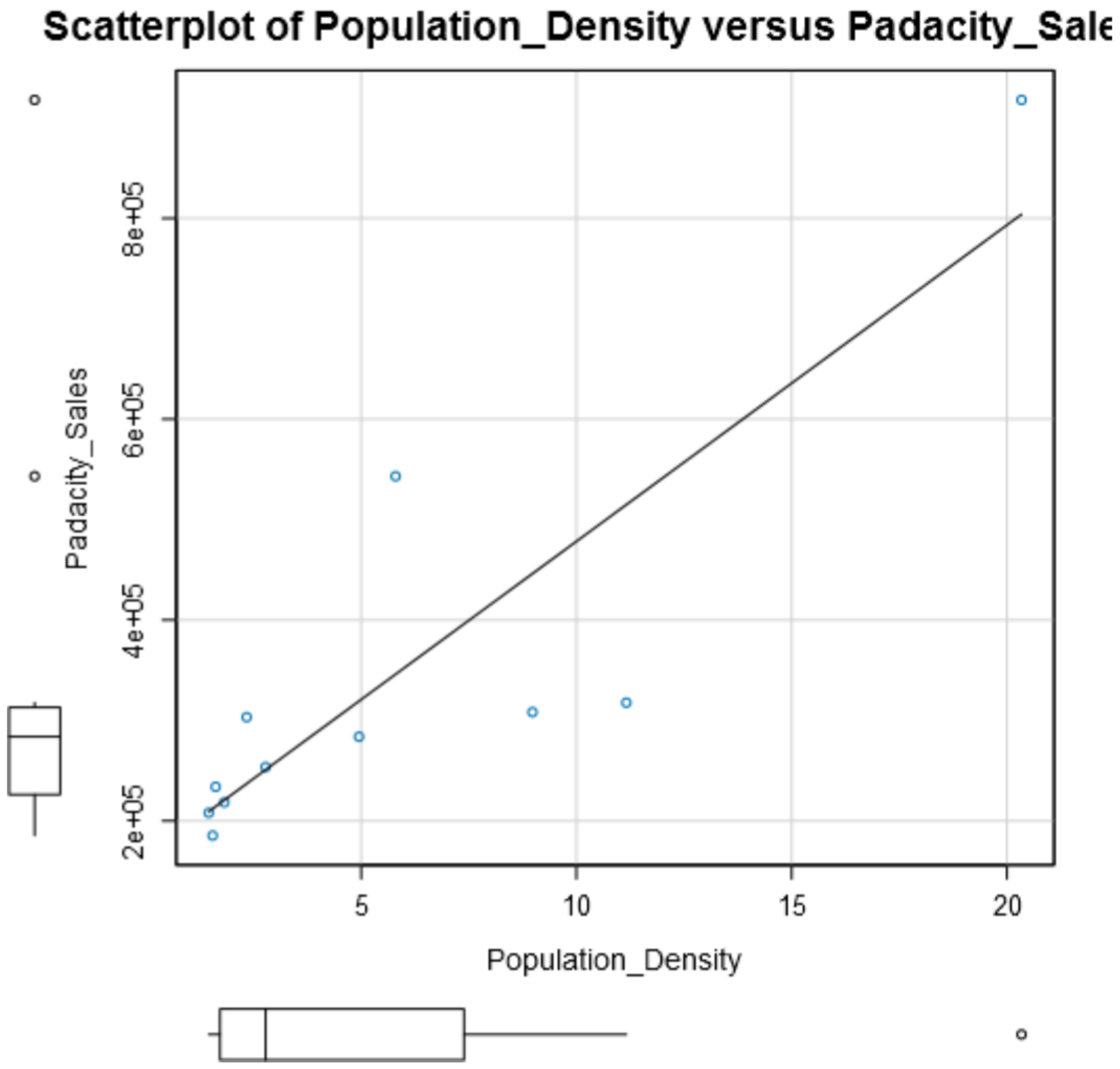
Column	**Sum**	**Average**
Census Population	*213,862*	*19,442*
Total Pawdacity Sales	*3,773,304*	*3,43,027.64*
Households with Under 18	*34,064*	*3,096.73*
Land Area	*33,071*	*3,006.49*
Population Density	*63*	*5.71*
Total Families	*62,653*	*5,695.71*

Dealing with Outliers

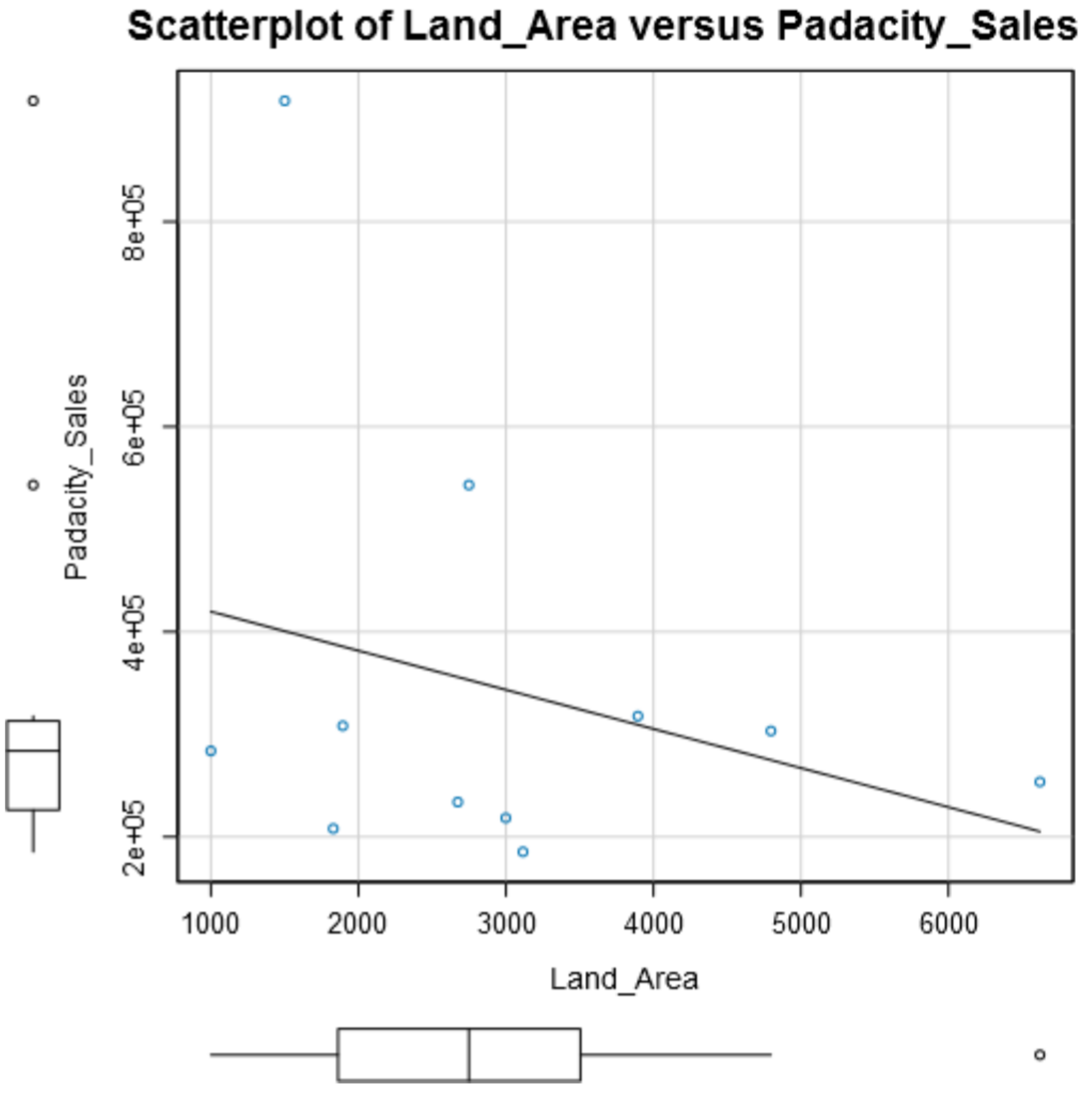
Below are scatterplots of each potential predictor variable against Pawdacity sales:

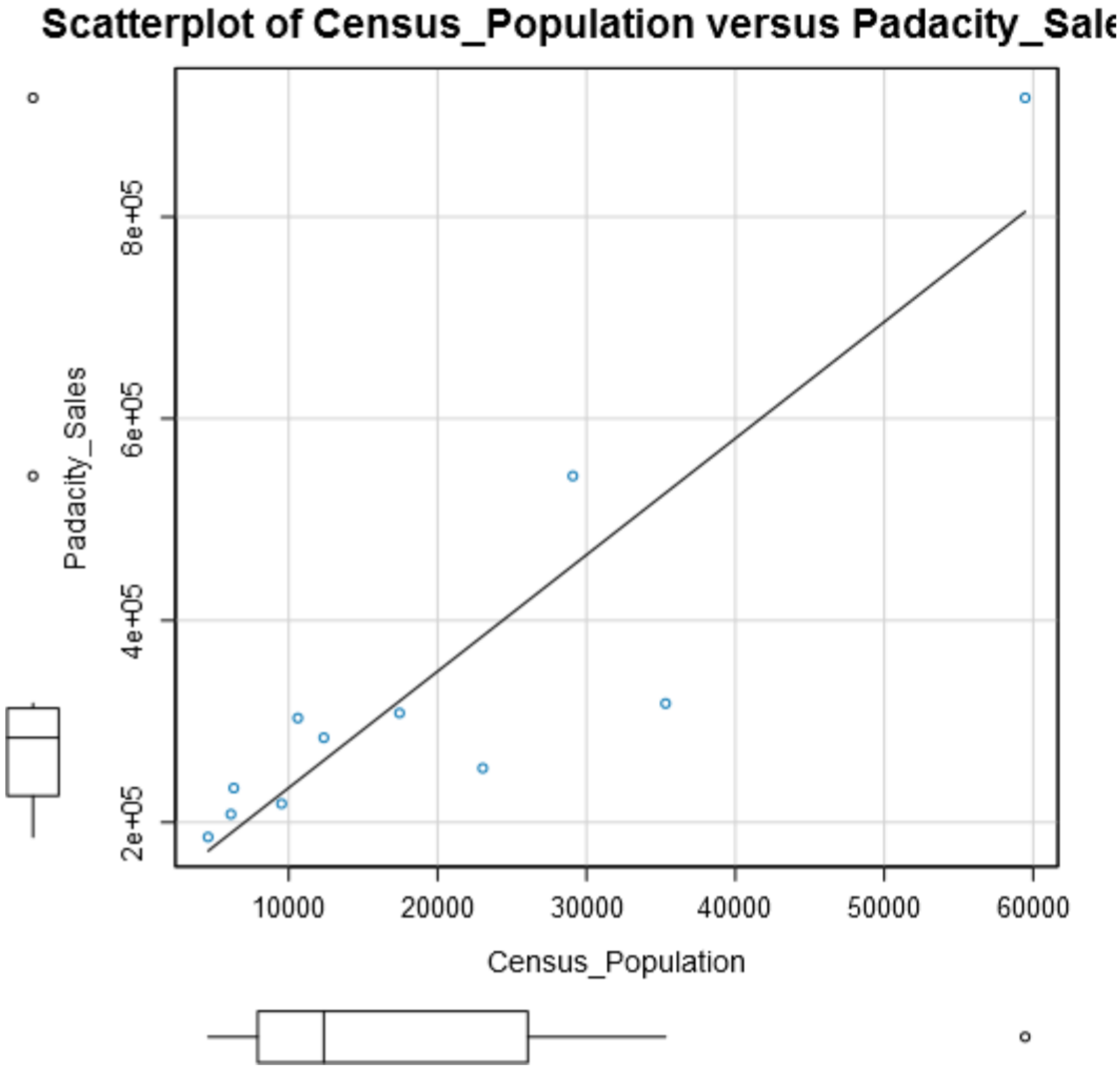



catterplot of Household_with_Under_18 versus Padacity_
Padacity_Sales
8e+05
6e+05
4e+05
2e+05
1000
2000
3000
4000
5000
6000
7000
8000
Household_with_Under_18



Scatterplot of Population_Density versus Padacity_Sale


Scatterplot of Land_Area versus Padacity_Sales
Padacity_Sales
2e+05
4e+05
6e+05
8e+05
1000
2000
3000
4000
5000
6000
Land_Area



By applying the IQR method of finding out the Upper Fence for each variable and identifying the outliers:

Census_Population_IQR	Padacity_Sales_IQR	Household_with_Under_18_IQR	Land_Area_IQR	Population_Density_IQR	Total_Families_IQR
18144.50	86832.00	2710.00	1643.19	5.67	4457.40
Census_Population_Upper_Fence	**Padacity_Sales_Upper_Fence**	**Household_with_Under_18_Upper_Fence**	**Land_Area_Upper_Fence**	**Population_Density_Upper_Fence**	**Total_Families_Upper_Fence**
53278.25	443232.00	8102.00	5969.69	15.90	14066.90

This provides us with the following potential outliers: Cheyenne City for Census Population, Land Area, Population Density, and Pawdacity Sales; Rock Springs for Land Area; Pawdacity sales for Gillette.

I feel confident in dismissing Rock Springs as it seems to follow the general downward trend of the line that fits the data points.

With Cheyenne City, the outlier behavior can be explained by the fact that they have 2 stores (which contributes to the excess), and that this behavior is spread across multiple variables. So, the excess sales is justifiable.

The same is not true for Gillette. They have 2 stores as well, but only its sales show outlier behavior, the rest being well within the expected range. There doesn’t seem to be a good enough reason for this, and hence I would remove this city from the dataset for further analysis.