Project #2.1
Data Cleanup

Claudia Dai

22th September, 2019

1 Business and Data Understanding

Pawdacity, a leading pet store chain in Wyoming, needs help to make an informed decision on where to open its 14^{th} store based on predicted yearly sales. Hence, we must determine the best location to open the new store and whether the new store will be profitable to Pawdacity.

To approach this problem in a data-driven manner, we will collect data for each city in Wyoming with a Pawdacity store for the following:

- Total sales volume of Pawdacity in 2010
- 2010 census data on the city populations
- Amount of households with under 18
- Land area
- Population density
- Total number of families

This will help us gain some insights to how the Pawdacity stores are currently operating and we can determine whether some of the above features may contribute to the stores' profitability. Furthermore, we can add data on Pawdacity's competitors in order to see whether the sales volume of competing stores have a relationship with Pawdacity's sales volume. Finally, we can take a look at the cities in Wyoming which do not have a Pawdacity store yet, and compare the above mentioned variables for determining the best location for Pawdacity's new store.

2 Building the Training Set

The following actions were performed to clean the given four datasets before merging them:

Dataset	Actions performed
`p2-2010-pawdacity-monthly-sales.csv`	• None
`p2-partially-parsed-wy-web-scrape.csv`	• Removed HTML tags • Removed [square brackets] from 2010 Census column, and converted to numeric column • Split City|County column into two columns • Removed "?" from City and stripped white space
`p2-wy-453910-naics-data.csv`	• None • However, noticed that for some stores the SALES VOLUME column indicates a zero ("0"). If these rows are relevant later, would be good to check if they represent actual zero sales volumes or if they are placeholders for a missing value.
`p2-wy-demographic-data.csv`	• None

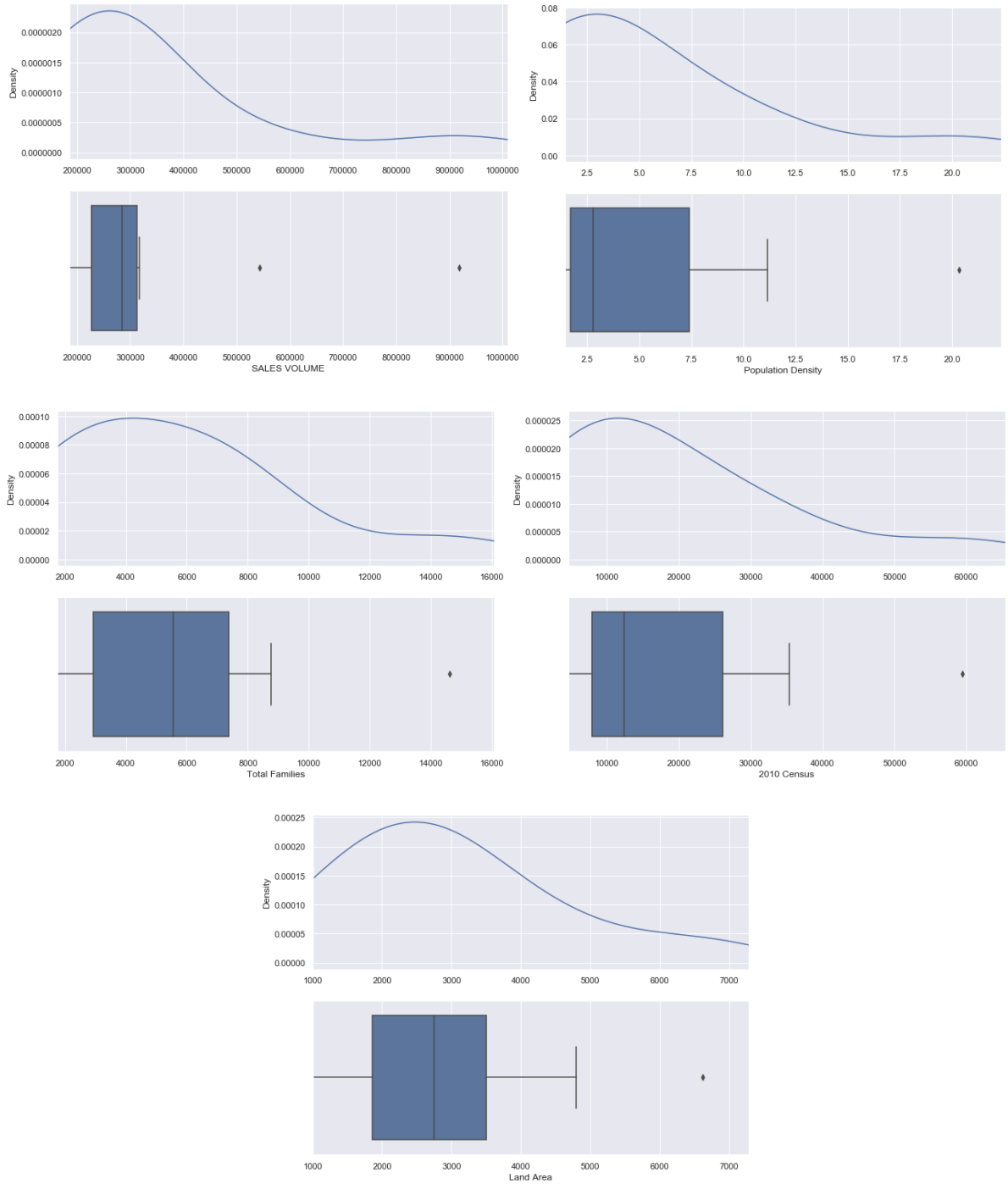
After cleaning the four datasets, I created a new dataset from Pawdacity's monthly sales dataset with the store name, the city (in order to be able to merge with the other datasets), and the total annual sales amount (which was calculated from the data given for 12 months). I did a left join of this dataset with the other datasets based on the common variable "City".

Finally, I obtained the same sums as were given, and the following averages:

Column	Sum	Average
Census Population	213,862	**19442.00**
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**

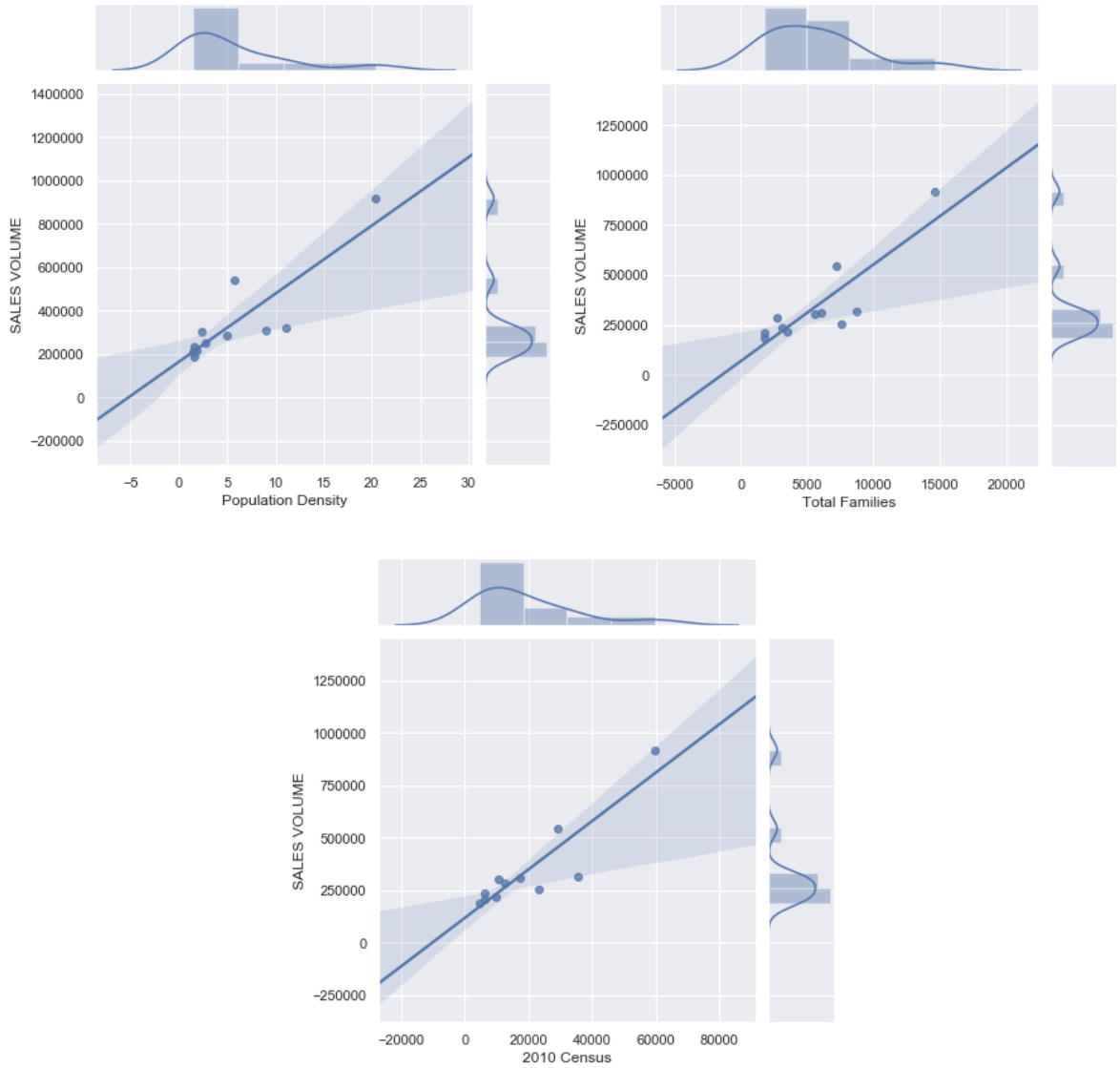
3 Dealing with Outliers

Three outliers were determined based on IQR: **Rock Springs, Cheyenne** and **Gilette.**

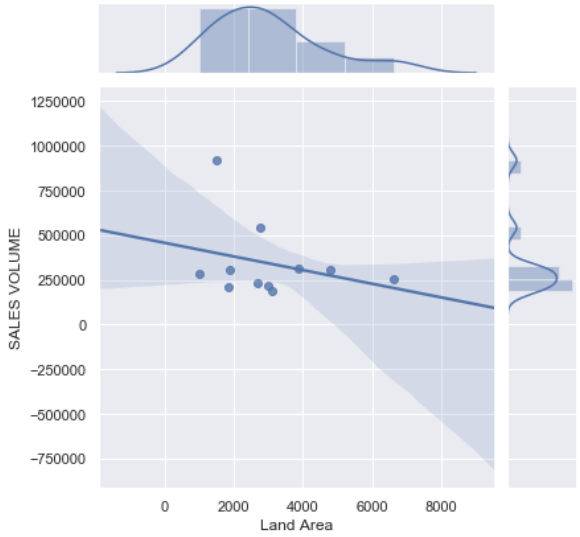


Cheyenne has significantly higher sales than the other cities and also seems to be an outlier when it comes to population density, total families, and 2010 census population. However, since Cheyenne is Wyoming's capital city, common sense would let us expect Cheyenne to have these outlier values as compared to the other cities.

When looking at the regression scatter plots for sales volume vs. population density, total families, and 2010 census population, Cheyenne falls within the confidence intervals. Hence, we will leave Cheyenne in the dataset to ensure our model stays robust and informative for bigger, more populated cities.



Rock Springs is an outlier when it comes to land area. If we take a look at the scatterplot of sales volume vs. land area, Rock Springs still has a linear fit with the other cities, and thus, we will also keep this city.



Finally, Gilette has significantly higher sales than the other cities and is thus, an outlier. As opposed to Cheyenne and Rock Springs, whose outliers could be explained by their nature or their correlation relationship, Gilette's high sales amount cannot directly be explained with the given data. Thus, we will remove this outlier in order to avoid a skewed model as a result.