Project 2.1: Data Cleanup

Complete each section. When you are ready, save your file as a PDF document and submit it here: https://classroom.udacity.com/nanodegrees/nd008/parts/8d60a887-d4c1-4b0e-8873-b2f36435eb39/project

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?

   We need to recommend what city to choose if we want to open a new Pawdacity 14th pet store based on the predicted sales.

   If we want to make this decision we will need to find city with most potential customer with the least competition in proximity, and start a business there :D

2. What data is needed to inform those decisions?

   In order to conduct this analysis, we would need :
   1. The sales data of existing pawdacity store and location, including sales trend
   2. Competitor sales data, competition, and trend of sales of pet goods of competitor in each city.
   3. Demographic data of the state where pawdacity locates and the potential areas, such as land area, density, number of family, or number of potential customers, which can retrieved by number of pet owner in the city, or from pet census

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

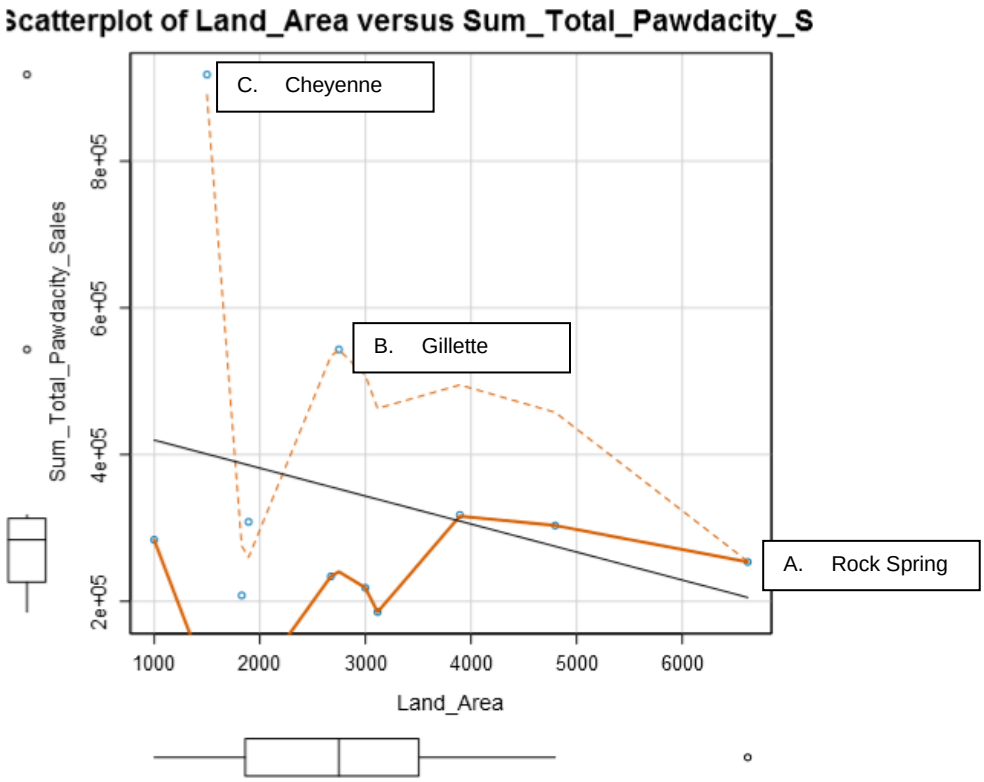
Column	Sum	Average
Census Population	***213,862***	***19,442***
Total Pawdacity Sales	***3,773,304***	***343,027.64***

Households with Under 18	***34,064***	***3,096.73***
Land Area	***33,071***	***3,006.45***
Population Density	***63***	***5.73***
Total Families	***62,653***	***5,695.73***

Step 3: Dealing with Outliers

Answer these questions

Are there any 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.

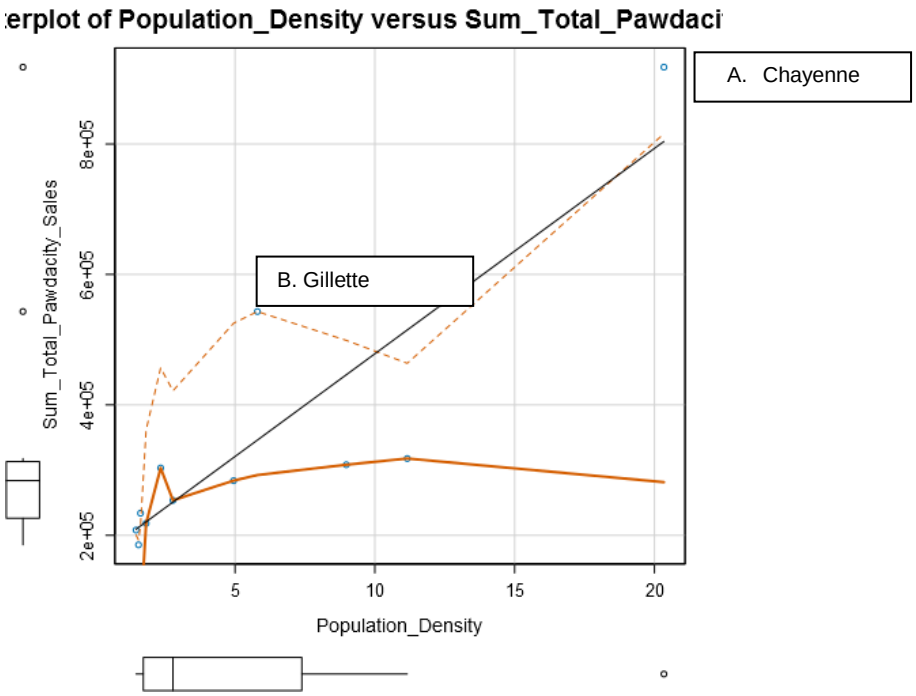


Based on regression analysis Land Area vs pawdacity sales, I can at least identify 3 potential outliers:

1st is Cheyenne, with the highest pawdacity sales (outliers in term of pawdacity sales)

2nd is Gillette (outliers in term of pawdacity sales also)

3rd Rock Spring (outliers in term of Land area, with average of pawdacity sales)



On the other regression analysis of population density, Chayenne and Gillette also came out as outliers again. This implies that both city have high sales much more than the other cities and can be considered as outliers. But If I can only remove one city,I would remove Gillette rather than Cheyenne as an outlier.
The reason behind this is while Cheyenne is an outlier in sales, it is also an outlier in density record that makes Cheyenne is a highly populated city with high sales, this makes sense, that Cheyenne as the capital and the most populous city in Wyoming has the highest sales. Meanwhile Gillette is an regularly sized city but with an outlier sales, so we should remove it because we are trying to predict sales in our regression

This particular reason also the reason why I didn't remove Rock Springs, because while rock springs, has outlier land area, the sales are still in the normal range, and since we are tryong to predict sales not land area, I would not remove Rock Springs

Result : Remove Gillette.

Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.