Creating an Analytical Dataset

Christopher Giler
Udacity Business Analyst Nanodegree

August 24, 2017
Project #2-1

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 currently has 13 stores open in the state of Wyoming, and is considering expanding by opening a 14th store within the state. A decision must be made to determine the best location to open the new store and whether or not the decision to expand will be profitable. Making these decisions requires compiling a dataset which can be used to project revenue for the new store for different cities in Wyoming

: Awesome: Yes, the main decision that Pawdacity needs to make is to select a city for its 14th store.

2. What data is needed to inform those decisions?

   These decisions will be driven by a dataset containing the following information for each city in Wyoming in which a Pawdacity store is currently operating:
   1. Total Pawdacity sales in 2010
   2. City population based on 2010 Census data
   3. Land area
   4. Population density
   5. Number of households with people under 18
   6. Total number of families

   In later steps, predictions will be made by considering data for other cities with potential to open a new store location:
   1. Market/sales data for other pet stores in the city
   2. Population and demographics for cities of interest

: Suggestion: Great job here! In addition we could also seek out the amount of traffic drivers to our current stores to understand what are the other landmarks or businesses that can drive foot traffic to our store. Understanding how far our competitors' stores are from the company's stores can help us model any traffic drivers as well. We also would like to gather data relative to our current local promotions and marketing budget spent per city on the company's current stores and would try to get information on the expected marketing money the company will spend to promote the new store.

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.

The dataset for current store locations is created by blending data from the following sources:

1. Monthly sales data for all Pawdacity stores in 2010
2. Population data from U.S. Census Bureau (2000, 2010, and estimated 2014)
3. Demographic data for each city and county in the state of Wyoming

The data is first aggregated by city, and then joined by city name. The resulting dataset contains 11 entries, with each entry representing a city in which Pawdacity currently operates. This dataset will be used as training data to build a predictive model for determining the new store location.

Column	Sum	Average
Census Population	*213,862*	*19,442.00*
Total Pawdacity Sales	*3,773,304*	*343,027.60*
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*	*5695.71*

: Awesome: The averages are correct! Great job!

Step 3: Dealing with Outliers

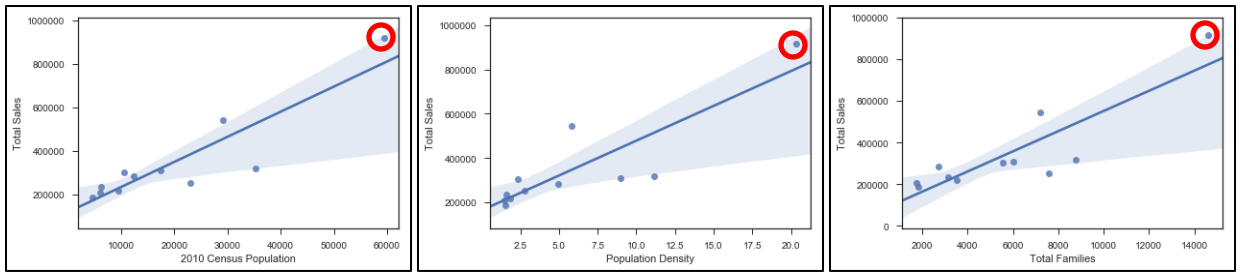
Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), ***you should only remove or impute one outlier****. Please explain your reasoning.*

The following outliers were detected in the new dataset using IQR methods for each feature.

```
2010 Census Population exceeds upper fence in Cheyenne
Total Sales exceeds upper fence in Cheyenne
Population Density exceeds upper fence in Cheyenne
Total Families exceeds upper fence in Cheyenne
Total Sales exceeds upper fence in Gillette
Land Area exceeds upper fence in Rock Springs
```

The first outlier is the city of **Cheyenne**, which exceeds the expected range for four features (2010 census population, population density, total families, and total sales). However, we would expect the state capital to be a more urban city compared to other cities within this dataset, so the high population density, total population, and total families all correlate with each other. A scatterplot of total sales against these three metrics also correlates with the linear relationship despite being an outlier in the data. This entry is left in to provide some additional robustness in the model for other more populated cities.

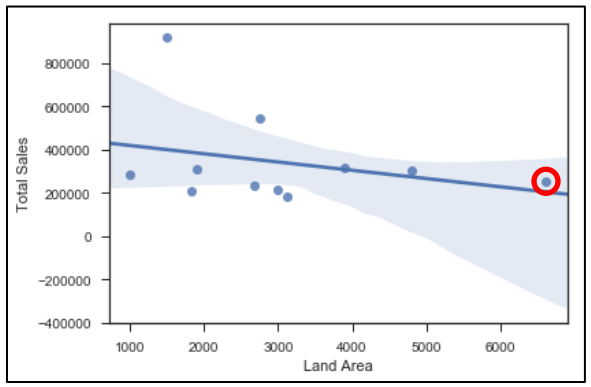
: Awesome: Yes, Cheyenne is just a big city because its numbers are above every other city in the training set on almost every data field. We can conclude that Cheyenne is not an anomaly, but just a big city given the other smaller cities in the available training set (n = 11) and would want to include this big city to have a more robust model so we can model any future cities with big numbers.



Total Sales vs. Population, Population Density, and Total Families; Cheyenne highlighted

Rock Springs is the second city which is considered an outlier based on land area. However, despite having a much higher land area compared to other cities in the dataset, it still correlates with a linear fit of the other data when looking at total sales against land area. Because it does not skew the data in this respect, this city will be left in the dataset as well.

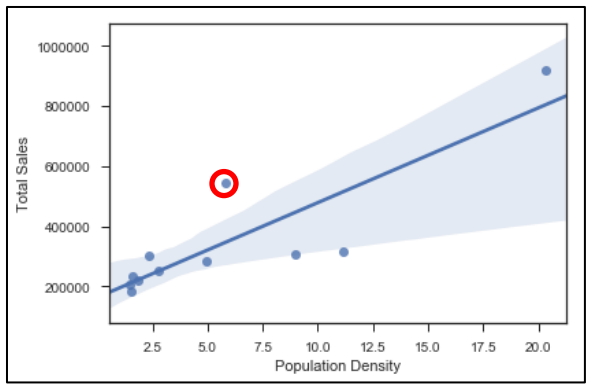
: Suggestion: Indeed,! Rock Springs is an outlier in the Land Area field as we can see from the table..To be more precise we suspect Land Area will not have a large effect on sales, therefore, we should leave it in.



Total Sales vs. Land Area; Rock Springs highlighted

Gillette is the third and final city which was determined to be an outlier based on total sales, despite being within range for all other features. Because we observe that sales for this city is an outlier which cannot be explained by any other outliers found in the population or demographics metrics, leaving this entry in the dataset has potential to skew any models trained on this data. Therefore, this city will be removed from the data before building our statistical model.

: Awesome: Indeed, Gillette is a true anomaly because its demographic numbers are within the expected range, yet the Pawdacity sales are really high, which doesn't make sense given the traditional understanding that if we have a higher number of people in an area, we should expect a bigger volume of sales, but Gillette is a small city with a very high amount of sales compared to the other cities in the training set. Therefore we should remove it.



Total Sales vs. Population Density; Gillette highlighted