Creating an Analytical Dataset

Christopher Giler
Udacity Business Analyst Nanodegree

August 29, 2017
Project #2-1 / 2-2

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

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

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*

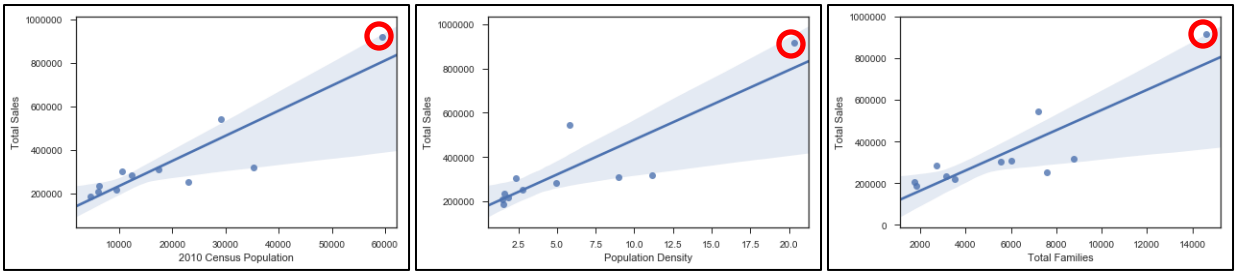
Step 3: Dealing with Outliers

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.

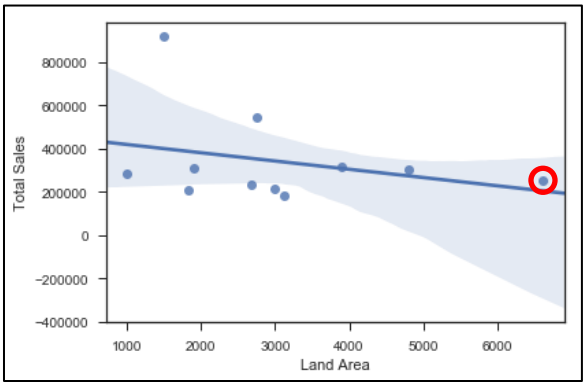
```
2010 Census Population exceeds upper fence in Cheyenne
Total Sales exceeds upper fence in Cheyenne
Population Density exceeds upper fence in Cheyenne
Total Families exceeds upper fence in Cheyenne
Total Sales exceeds upper fence in Gillette
Land Area exceeds upper fence in Rock Springs
```

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.



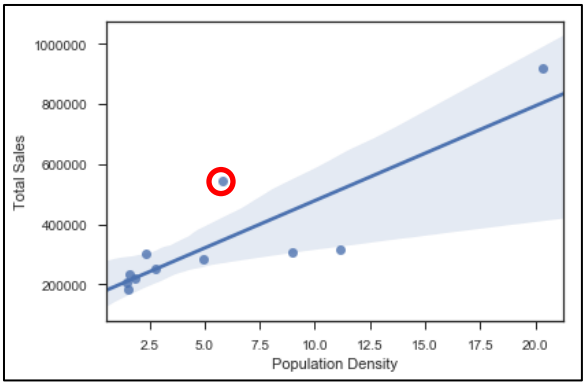
Total Sales vs. Population, Population Density, and Total Families; Cheyenne highlighted

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.



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.



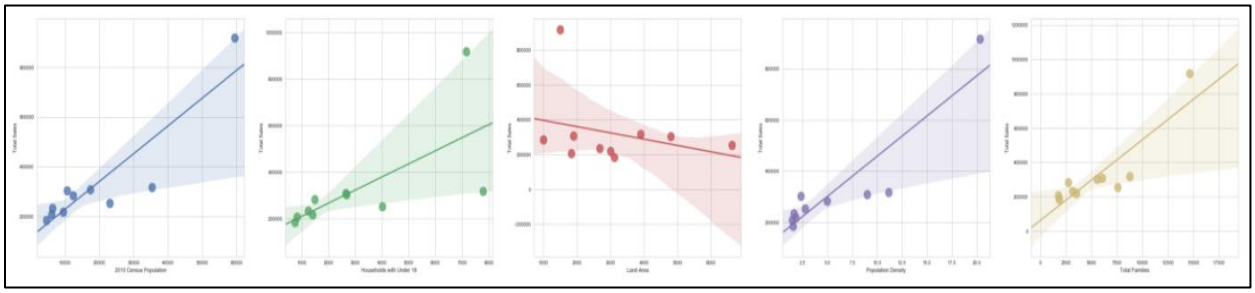
Total Sales vs. Population Density; Gillette highlighted

Step 4: Build a Linear Regression Model

Removing Gillete from the set of current Pawdacity store locations yields the following chart for our training dataset.

	2010 Census Population	Total Sales	Households with Under 18	Land Area	Population Density	Total Families
CITY						
Buffalo	4585.0	185328.0	746.0	3115.507500	1.55	1819.50
Casper	35316.0	317736.0	7788.0	3894.309100	11.16	8756.32
Cheyenne	59466.0	917892.0	7158.0	1500.178400	20.34	14612.64
Cody	9520.0	218376.0	1403.0	2998.956960	1.82	3515.62
Douglas	6120.0	208008.0	832.0	1829.465100	1.46	1744.08
Evanston	12359.0	283824.0	1486.0	999.497100	4.95	2712.64
Powell	6314.0	233928.0	1251.0	2673.574550	1.62	3134.18
Riverton	10615.0	303264.0	2680.0	4796.859815	2.34	5556.49
Rock Springs	23036.0	253584.0	4022.0	6620.201916	2.78	7572.18
Sheridan	17444.0	308232.0	2646.0	1893.977048	8.98	6039.71

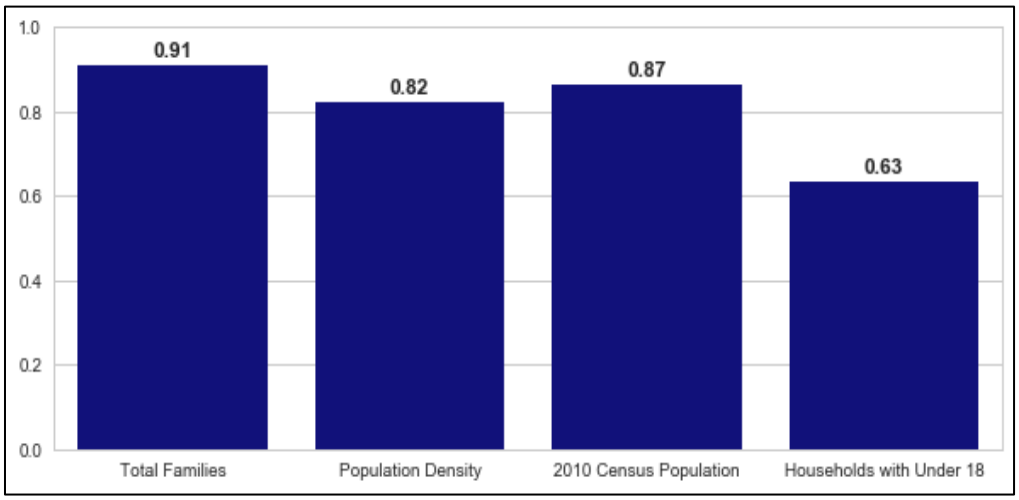
The first step in developing the linear model is to determine which of the available metrics would work well as predictor variables. In order to be an effective predictor, the variable itself should have some linear correlation with the value we are trying to predict (in this case 'Total Sales'). Analyzing each metric against total sales and applying a linear regression least squares fit yields the following:



We see that each of the potential metrics has a fairly linear relationship with total sales, so we cannot eliminate any metric at this step. However, intuition leads us to believe some of the metric variables themselves may be correlated with each other (for example, population vs total families or population density). To double-check this, we generate a correlation table for each of the metrics being considered.

	2010 Census Population	Total Sales	Households with Under 18	Land Area	Population Density	Total Families
2010 Census Population	1.000000	0.898755	0.911562	-0.052470	0.944389	0.969190
Total Sales	0.898755	1.000000	0.674652	-0.287078	0.906180	0.874663
Households with Under 18	0.911562	0.674652	1.000000	0.189376	0.821986	0.905660
Land Area	-0.052470	-0.287078	0.189376	1.000000	-0.317419	0.107304
Population Density	0.944389	0.906180	0.821986	-0.317419	1.000000	0.891680
Total Families	0.969190	0.874663	0.905660	0.107304	0.891680	1.000000

While land area does not seem to be affected by any other variable in the dataset, the others appear to be strongly correlated with each other. For the linear model, we should only select one of these variables as a predictor. In order to make this selection, we fit a linear model to a pair of variables (land area + one of the other variables in question) and compare the R^2 score to make our selection.



We can see that using the number of total families in the city allows us to model the data most effectively, with an R^2 score of 0.91. We select total families and land area as our two predictor variables for the linear model. Performing a fit over these predictors yields the following equation:

```
Predicted Sales = 197330.41 + (-48.42)*(Land Area) + (49.14)*(Total Families)
```

Step 5: Perform the Analysis

After fitting the model to our dataset, we can begin considering data for potential cities to open the next Pawdacity storefront. Two sets of data will be used to perform this analysis:

1. Population data from U.S. Census Bureau (2000, 2010, and estimated 2014)
2. Demographic data for each city and county in the state of Wyoming
3. NAICS data on the most current sales of all competitor stores where total sales is equal to 12 months of sales. A sample of this data is shown below:

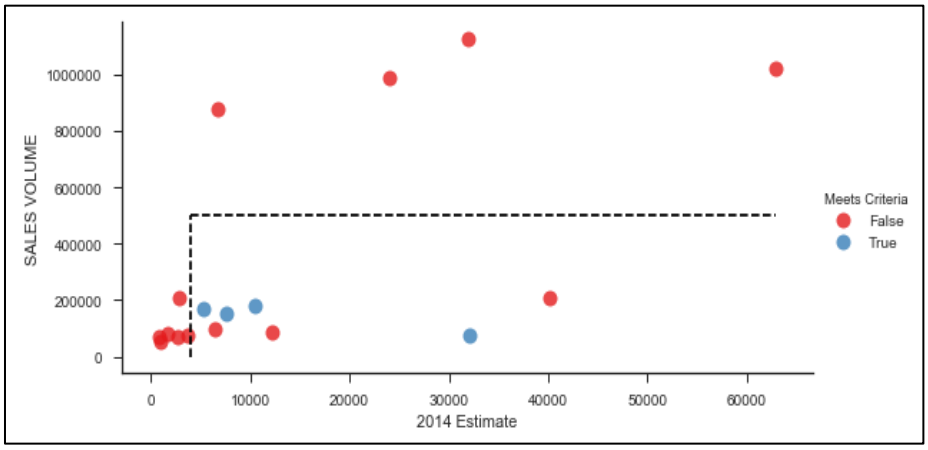
	BUSINESS NAME	SALES VOLUME	CASS_LastLine	City
0	Mile High Mobile Pet LLC	300000	Cheyenne, WY 82007-3528	Cheyenne
1	Pets City Inc	640000	Cheyenne, WY 82009-4851	Cheyenne
2	Petco Animal Sups Stores Inc	0	Cheyenne, WY 82009-4945	Cheyenne
3	Pet-A-Care	81000	Cheyenne, WY 82009-1009	Cheyenne
4	Muddy Paws Pet Salon	76000	Laramie, WY 82070-8979	Laramie

The NAICS data is aggregated first by city, and the sum of sales volume is calculated for each city. This table of aggregated data is joined by city to the population and demographics tables. From here, we consider limitations on potential cities where the new store could be opened:

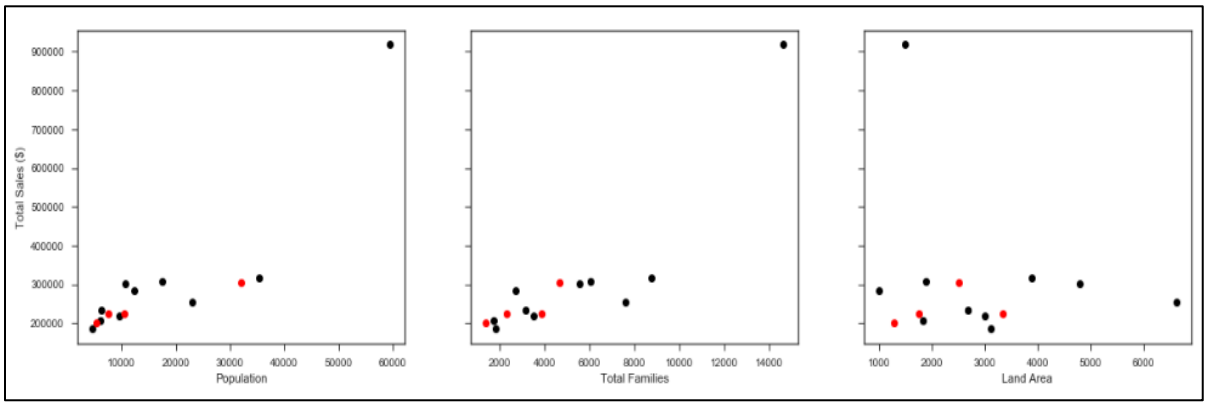
1. The new store should be located in a new city.
2. The total sales for the entire competition in the new city should be less than $500,000
3. The new city where you want to build your new store must have a population over 4,000 people (based on 2014 US Census)
4. The predicted yearly sales must be over $200,000
5. The city chosen has the highest predicted sales from the predicted set

Data is plotted below, with thresholds for city population and competitor sales volume also highlighted. Blue markers in this plot represent the four potential cities which meet the criteria 1-3 as listed above. The cities meeting these criteria are:

- Jackson
- Lander
- Laramie
- Worland



Sales predictions are made based on our trained linear model for these four cities of interest. We can see below that plotting sales predictions against population, total families, and land area for these four cities, along with total sales from our training data, shows that our predictions following similar trends to the training set.



Predicted yearly sales for the four cities in question are shown below. The predicted sales for each of the four stores are above $200,000, so all cities meet item #4 in our selection criteria. Based on current demographic, population, and competitor sales data, we can confidently recommend Pawdacity to expand their stores to **Laramie, WY**. Our predicted yearly sales volume in 2014 for this store is **$305,013.88**.