Project 2.1: Data Cleanup

Make a copy of this document. 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 need to be made?

   I need to perform an analysis to recommend the city for Pawdacity's newest store, based on predicted yearly sales. Before predicting yearly sales, I formatted and blended together data from different datasets, and then dealt with outliers.

2. What data is needed to inform those decisions?

   To properly build the model and select predictor variables, I created a dataset using the following data:
   (a) Cities
   (b) 2010 census population numbers in each city
   (c) Total Pawdacity sales in 2010 in each city
   (d) The number of households with under 18 in each city
   (e) Land area in each city
   (f) Population density in each city
   (g) Total families in each city

   This dataset is my training set and will help me build a regression model in order to predict sales in the Practice Project in the next lesson.

   (b) and (f) could be useful as denser areas would have more customers, and (d) and (g) would allow me to detect areas with more potential customers.

   The following data could also be helpful if available:
   - Traffic drivers to our current stores and how far competitors' stores are from the company's stores → to understand what are the other landmarks or businesses which can affect foot traffic to the store.

- Competitor sales → to see whether there is significant customer demand in the city, and whether competition may pose a risk if Pawdacity decides to open a store there.
- Our current local promotions and marketing budget spent per city on the current stores.
- Expected marketing funds 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.

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*	*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*

Step 3: Dealing with Outliers

Answer these questions

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.

I used the IQR method for each attribute (Census Population, Total Pawdacity Sales, Households with Under 18, Land Area, Population Density and Total Families) of the cities to see which cities have outlier values for any of these.

City	Census Population	Total Pawdacity Sales	Households with Under 18	Land Area	Population Density	Total Families
Buffalo	4,585	185,328	746	3,115.51	1.55	1,819.50
Casper	35,316	317,736	7,788	3,894.31	11.16	8,756.32
Cheyenne	59,466	917,892	7,158	1,500.18	20.34	14,612.64
Cody	9,520	218,376	1,403	2,998.96	1.82	3,515.62

Douglas	6,120	208,008	832	1,829.47	1.46	1,744.08
Evanston	12,359	283,824	1,486	999.50	4.95	2,712.64
Gillette	29,087	543,132	4,052	2,748.85	5.80	7,189.43
Powell	6,314	233,928	1,251	2,673.57	1.62	3,134.18
Riverton	10,615	303,264	2,680	4,796.86	2.34	5,556.49
Rock Springs	23,036	253,584	4,022	6,620.20	2.78	7,572.18
Sheridan	17,444	308,232	2,646	1,893.98	8.98	6,039.71
Q1	6,314	218,376	1,251	1,829.47	1.62	2,712.64
Q3	29,087	317,736	4,052	3,894.31	8.98	7,572.18
IQR	22,773	99,360	2,801	2,064.84	7.36	4,859.54
Upper Fence	63,247	466,776	8,254	6,991.58	20.02	14,861.49
Lower Fence	-27,846	69,336	-2,951	-1,267.80	-9.42	-4,576.67

The table above shows that Cheyenne is an outlier in Total Pawdacity Sales and Population Density, and Gillette is an outlier in Total Pawdacity Sales. Even though Cheyenne has the most fields in which it is an outlier, its values in Census Population, Households with Under 18 and Total Families are quite high and nearly reach the Upper Fence. Cheyenne appears to be a city which has a large population, a large number of families and households with under 18, and high population density. Therefore, high sales figures can be expected, and Cheyenne can be kept in the dataset.

On the other hand, Gillette has moderate Census Population, Households with Under 18, Population Density and Total Families, but its sales figures are very high. This means that other attributes in the dataset are not able to explain the high sales figures. Hence Gillette can be removed.

Note that there are only 11 cities. Therefore, each city which we remove has a huge impact on our model. For example, if there were 100 cities in the dataset, we could remove all of the outliers; the size of the dataset has a huge influence on our decision of removing only one city.

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.