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 needs to be made?
   A decision about the most appropriate location for a new store needs to be made.

2. What data is needed to inform those decisions?
   Some of the data that is needed to inform our decision include:
   - Sales from both existing stores of Pawdacity and also competitors.
   - The size and number of competitive stores in selected area.
   - Information about population's size, growth, and segmentations.
   - The availability of facilities for pets in selected area.

: Awesome: Right! The main decision here is that the company wants to determine whether the expected profit from these customers exceeds $10,000 and then decide to send the catalog out to these customers or not.

: Awesome: Correctly explained in depth.

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.

After working on the data, I ended up with this training set which helped me to answer and match the next table.

City	2010 Census Population	Total Pawdacity Sales	Households with Under 18	Land Area	Population Density	Total Families
Buffalo	4585	185328	746	3115.5075	1.55	1819.5
Casper	35316	317736	7788	3894.3091	11.16	8756.32
Cheyenne	59466	917892	7158	1500.1784	20.34	14612.64
Cody	9520	218376	1403	2998.95696	1.82	3515.62
Douglas	6120	208008	832	1829.4651	1.46	1744.08
Evanston	12359	283824	1486	999.4971	4.95	2712.64
Gillette	29087	543132	4052	2748.8529	5.8	7189.43
Powell	6314	233928	1251	2673.57455	1.62	3134.18
Riverton	10615	303264	2680	4796.859815	2.34	5556.49
Rock Springs	23036	253584	4022	6620.201916	2.78	7572.18
Sheridan	17444	308232	2646	1893.977048	8.98	6039.71

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 **(see the attached excel file for more info)**

Column	Sum	Average

two decimal places, ex: 1.24 **(see the attached excel file for more info)**

Column	Sum	Average
Census Population	*213,862*	**19442**
Total Pawdacity Sales	*3,773,304*	**343027.63**
Households with Under 18	*34,064*	**3096.72**
Land Area	*33,071*	**3006.48**
Population Density	*63*	**5.70**
Total Families	*62,653*	**5695.70**

: Awesome: Correctly identified and justified all the variables required.

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.

Yes, there are two cities that are outliers in the training set; Cheyenne and Gillette. By conducting two different experiments, I would choose to remove the outlier data of the Gillette city. The essential reasons are that Cheyenne is not an outlier but rather a big urban city where high population density results in possible huge sales. In addition, Gillette is a true outlier as its demographic numbers are within the standard range, yet the sales of Pawdacity are high, and thereby, we could conclude that Gillette is an outlier.

: Awesome:
- Cheyenne has huge sales but as you also saw it has high population density.
- Which means that the high sales are justified by the high population density.
- Cheyenne is indeed a big urban city.
- We can conclude that Cheyenne is not an anomaly, but just a big city given the other smaller cities in the available training set and would want to include this big city to have a more robust model so we can model any future cities with big numbers.

: Awesome: 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.

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.