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?

   The company needs to find out the new location taking into consideration the annual sales.

2. What data is needed to inform those decisions?

   We need to find out the annual sale for Pawdacity. We need to summarize data on the city level.

   We will need total pollution data of Wyoming for all the cities for existing market as well as future market. For this starting with the web scrapped data of Wyoming data, we will extract 2010 census population data.
   We would need Pawdacity sales data for each city.
   We would also need the demographic data for all the city which will give also the total land area, population density, total families and household under 18.

   We could also use some additional data such as data regarding competition location and sales, details of marketing budget etc.

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*
Total Pawdacity Sales	*3,773,304*	*343027.63*
Households with Under 18	*34,064*	*3096.72*

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.

The data for Cheyenne city can be considered outlier and can be removed. The data does not fall under the upper and lower fences and the value is too high and does not fit with other data in the set. Cheyenne seems to be a big city, in midst of a dataset that contains small and medium sized cities. It has multiple outlier fields and even its other field values, are unlike the other cities in the dataset. Therefore, it can possibly skew our predictor model and thus, its removal from the dataset is justified.

Though Cheyenne has been considered as an outlier and can be removed, but other than that there is one more outlier which is Gillette. From Table 1. we can see that the Cheyenne has an outlier value for three fields (population density, total families and pawdacity sales value) where as for Gillette the outlier is only for one field (pawdacity sales value). Therefore by keeping Gillette the analysis will not skew as compared to if we keep Cheyenne. And since the dataset is very small, we cannot afford to drop more than one city from the dataset. Therefore Gillette is retained in the dataset.

Table 1: City data along with upper and lower fence.

City	Sum_Land Area	Sum_Households with Under 18	Sum_Population Density	Sum_Total Families	Sum_Pawdacity sales	Sum_2010 Census21
Buffalo	3115.5075	746	1.55	1819.5	185328	4585
Casper	3894.3091	7788	11.16	8756.32	317736	35316
Cheyenne	**1500.1784**	**7158**	**20.34**	**14612.64**	**917892**	**59466**
Cody	2998.95696	1403	1.82	3515.62	**218376**	9520
Douglas	1829.4651	832	1.46	1744.08	208008	6120
Evanston	999.4971	1486	4.95	2712.64	283824	12359
Gillette	**2748.8529**	**4052**	**5.8**	**7189.43**	**543132**	**29087**
Powell	2673.57455	1251	1.62	3134.18	233928	6314
Riverton	4796.859815	2680	2.34	5556.49	303264	10615
Rock Springs	6620.201916	4022	2.78	7572.18	253584	23036
Sheridan	1893.977048	2646	8.98	6039.71	308232	17444
Upper fence	5969.68913	8102	15.895	14066.8975	443232	53278.25
Lower fence	-603.059765	-2738	-6.785	-3762.6825	95904	-19299.75

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.