Project 2.1: Data Cleanup

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? In this project we need to predict yearly sales of Pawdacity for different cities to determine which city should welcome the newest store in the state in the next year.

2. What data is needed to inform those decisions? To work on this project, we'll need data on the monthly sales for all the Pawdacity's stores in 2010. Additionally, the sales of all competitor stores of the last 12 months for different cities in the state, and also demographic data which encompass population numbers and some more specific data (e.g. land area, population density and so on) for different counties and cities in the state.

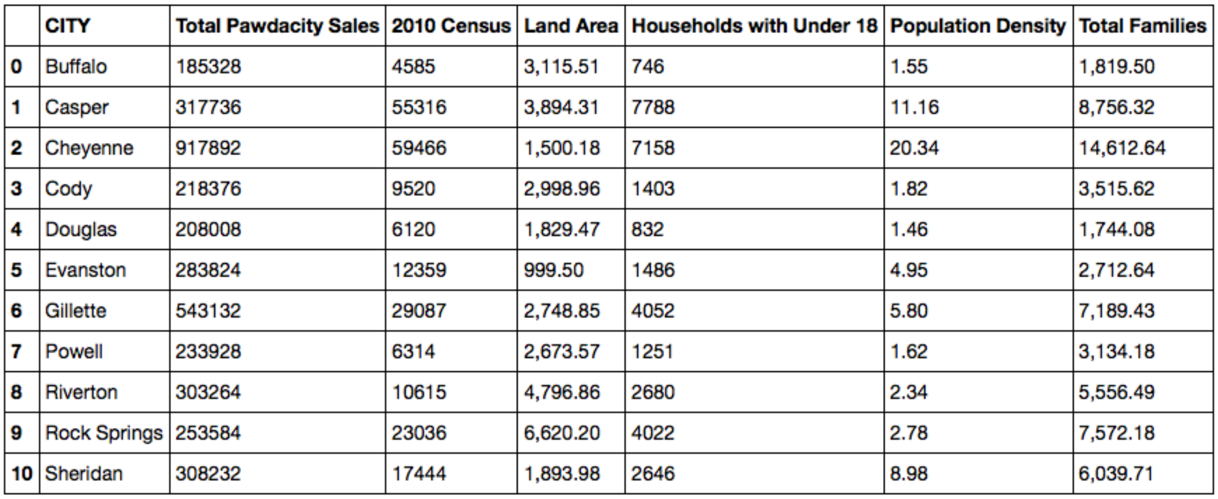
Step 2: Building the Training Set

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	*233,862 (OK)*	*31260.18*
Total Pawdacity Sales	*3,773,304 (OK)*	*343027.64*
Households with Under 18	*34,064 (OK)*	*3096.73*
Land Area	*33,071 (OK)*	*3006.49*
Population Density	*63 (OK)*	*5.71*
Total Families	*69,653* ***(NOT OK!)***	*5695.71*

Dear reviewer, I've tried very hard to see why the number of families is not matching. Would you please check out the code in https://github.com/jaycwb/udacity-bizanalytics-create-analytical-dataset to see what I'm missing?



	CITY	Total Pawdacity Sales	2010 Census	Land Area	Households with Under 18	Population Density	Total Families
0	Buffalo	185328	4585	3,115.51	746	1.55	1,819.50
1	Casper	317736	55316	3,894.31	7788	11.16	8,756.32
2	Cheyenne	917892	59466	1,500.18	7158	20.34	14,612.64
3	Cody	218376	9520	2,998.96	1403	1.82	3,515.62
4	Douglas	208008	6120	1,829.47	832	1.46	1,744.08
5	Evanston	283824	12359	999.50	1486	4.95	2,712.64
6	Gillette	543132	29087	2,748.85	4052	5.80	7,189.43
7	Powell	233928	6314	2,673.57	1251	1.62	3,134.18
8	Riverton	303264	10615	4,796.86	2680	2.34	5,556.49
9	Rock Springs	253584	23036	6,620.20	4022	2.78	7,572.18
10	Sheridan	308232	17444	1,893.98	2646	8.98	6,039.71

```
Total Pawdacity Sales has a sum of = 3773304
Total Pawdacity Sales has an average of = 343027.64
2010 Census has a sum of = 233862
2010 Census has an average of = 21260.18
Land Area has a sum of = 33071.380389
Land Area has an average of = 3006.49
Households with Under 18 has a sum of = 34064
Households with Under 18 has an average of = 3096.73
Population Density has a sum of = 62.8
Population Density has an average of = 5.71
Total Families has a sum of = 62652.79
Total Families has an average of = 5695.71
```

Step 3: Dealing with Outliers

Answer these questions

Are there any outliers in the training set? How many? I used Tukey's method to find outliers in this dataset. I've ran Tukey for each attribute solely, and the only City that popped out as an outliers multiple times as Cheyenne.

Which outlier have you chosen to remove or impute? Please provide a brief justification for each outlier on why you kept, removed, or imputed it. I don't think imputing the values of Cheyenne would be a good idea, since it was flagged as an outlier 4 times. If we proceeded with imputation of 4 values, this city would not really represent the reality, since the dataset has only 5 attributes (with the exception of the pawdacity sales, that will work as our target variable later on). This way, my suggestion would be towards removing it from the data set.

Note: Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.

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.