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 city to open a new Pawdacity store needs to be selected based on analyzing predicted sales. The ideal variables from the data files will need to be selected in order to build a model that can predict which city to open the new Padacity store in.

2. **What data is needed to inform those decisions?**

   Pawdacity monthly sales file as well as the partially parsed data file, and Demographic file are needed in order to build a data set. More specifically the city and total sales of each Pawdacity store are needed. In addition 2010 Census, Households with under 18 year olds, Land area, population density and total family data sets are needed in order to develop a working model.

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*	3006.45
Population Density	*63*	5.73
Total Families	*62,653*	5695.73

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.

	A	B	C	D	E	F	G	H
1	CITY	2010 Census	Total Pawdacity Sales	Households with Under 18	Land Area	Population Density	Total Families	
2	Buffalo	4585	185328	746	3116	2	1820	
3	Casper	35316	317736	7788	3894	11	8756	
4	Cheyenne	59466	917892	7158	1500	20	14613	
5	Cody	9520	218376	1403	2999	2	3516	
6	Douglas	6120	208008	832	1829	1	1744	
7	Evanston	12359	283824	1486	999	5	2713	
8	Gillette	29087	543132	4052	2749	6	7189	
9	Powell	6314	233928	1251	2674	2	3134	
10	Riverton	10615	303264	2680	4797	2	5556	
11	Rock Springs	23036	253584	4022	6620	3	7572	
12	Sheridan	17444	308232	2646	1894	9	6040	
13		1st Quartile	1st Quartile	1st Quartile	1st Quartile	1st Quartile	1st Quartile	
14		7917	226152	1327	1861.5	2	2923.5	
15		3rd Quartile	3rd Quartile	3rd Quartile	3rd Quartile	3rd Quartile	3rd Quartile	
16		26061.5	312984	4037	3505	7.5	7380.5	
17		IQR	IQR	IQR	IQR	IQR	IQR	
18		18144.5	86832	2710	1643.5	5.5	4457	
19		Upper	Upper	Upper	Upper	Upper	Upper	
20		53278.25	443232	8102	5970.25	15.75	14066	
21		Lower	Lower	Lower	Lower	Lower	Lower	
22		-19299.75	95904	-2738	-603.75	-6.25	-3762	
23								

I have Highlighted outliers in yellow as well as the upper/lower quartile that was crossed to define the outlier. I have chosen to remove the city of Gillette. Though Cheyenne has outliers across multiple variables it is a capital city which means larger variables. On the other hand Gillette has an outlier on total Pawdacity sales but no other outliers which would validate its occurrence such as an outlier population density to grant it merit.

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.