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?**

Determine which city is the most appropriate for opening a new store for Pawdacity in terms of predicted yearly sales based on the data we have from the other 13 stores.

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

We would require historical data with yearly sales from the 13 stores Pawdacity has as well population numbers and demographical data to find out what type of customer we can address to in that particular region. Additionally, which region and type of persons are the most responsive to our products.

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*	***19,442***
Total Pawdacity Sales	*3,773,304*	***343,027.63***
Households with Under 18	*34,064*	***3,096.72***
Land Area	*33,071*	***3,006.45***
Population Density	*63*	***5.72***
Total Families	*62,653*	***5,695.72***

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.

City	Sales Volume	County	2014 Estimate	2010 Census	2000 Census	Land Area	Households with	Population D	Total Families
Buffalo	185328	Johnson	4615	4585	3900	3115.5075	746	1.55	1819.5
Casper	317736	Natrona	40086	35316	32644	3894.3091	7788	11.16	8756.32
Cheyenne	917892	Laramie	62845	59466	53011	1500.1784	7158	20.34	14612.64
Cody	218376	Park	9740	9520	8835	2998.95696	1403	1.82	3515.62
Douglas	208008	Converse	6423	6120	5288	1829.4651	832	1.46	1744.08
Evanston	283824	Uinta	12190	12359	11507	999.4971	1486	4.95	2712.64
Gillette	543132	Campbell	31971	29087	19646	2748.8529	4052	5.8	7189.43
Powell	233928	Park	6407	6314	5373	2673.57455	1251	1.62	3134.18
Riverton	303264	Fremont	10953	10615	9310	4796.859815	2680	2.34	5556.49
Rock Sprir	253584	Sweetwat	24045	23036	18708	6620.201916	4022	2.78	7572.18
Sheridan	308232	Sheridan	17916	17444	15804	1893.977048	2646	8.98	6039.71
	226152		8081.5	7917	7104	1861.721074	1327	1.72	2923.41
	312984		28008	26061.5	19177	3504.9083	4037	7.39	7380.805
	86832		19926.5	18144.5	12073	1643.187226	2710	5.67	4457.395
	443232		57897.75	53278.25	37286.5	5969.689139	8102	15.895	14066.8975
	95904		-21808.25	-19299.75	-11005.5	-603.059765	-2738	-6.785	-3762.6825

I would choose the city of **Cheyenne** because it contains multiple outliers.
All the values colored in red are outliers based on the calculations of IQR provided by you.

Steps followed:
1 . Calculate 1st quartile Q1 and 3rd quartile Q3 of the dataset. You can use the Excel function QUARTILE.INC or QUARTILE.EXC

2 . Calculate the Interquartile Range: IQR = Q3 - Q1

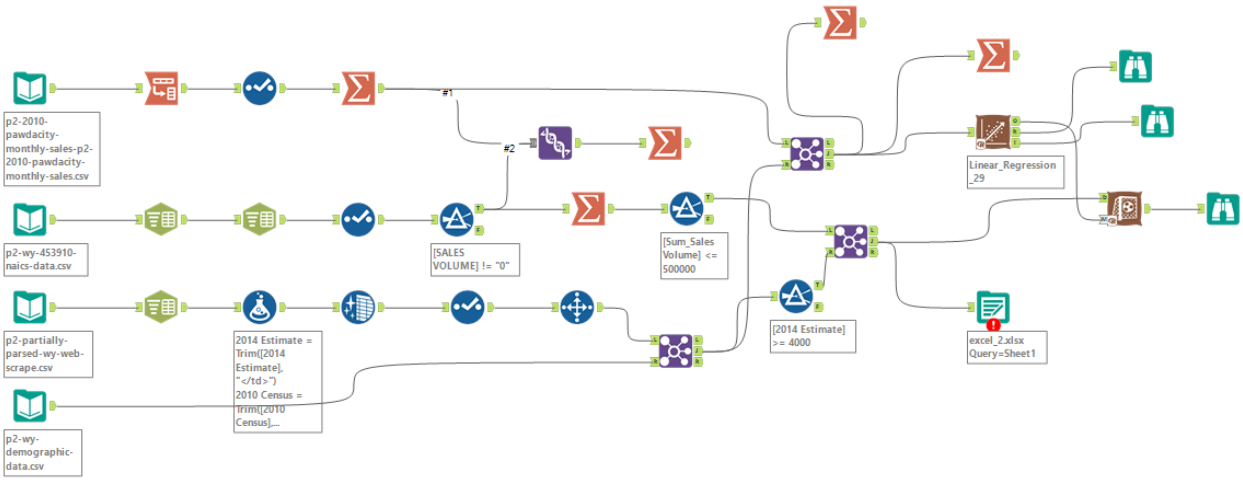
3 . Add 1.5 *IQR to Q3 to get the upper fence: Upper Fence = Q3 + 1.5* IQR
4 . Subtract 1.5 *IQR to Q1 to get the lower fence: Lower Fence = Q1 - 1.5* IQR
5 . Values above the Upper Fence and values below the Lower Fence are outliers

The results are on the lower base of the table and with these results I compared the values in the table and signaled the outliers.

The outlier from Gillette city can be an exception. Taking into consideration the numerous population, the sales volume might be appropriate. By the same token, I think it would be wiser to impute this value.

As for Rock Springs, the land of this city might be simply bigger than the others.

I would keep both outliers (impute sales in Gillette) and delete the city of Cheyenne from the analysis.



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.