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 location of new store based on previous yearly sales.

2. What data is needed to inform those decisions?

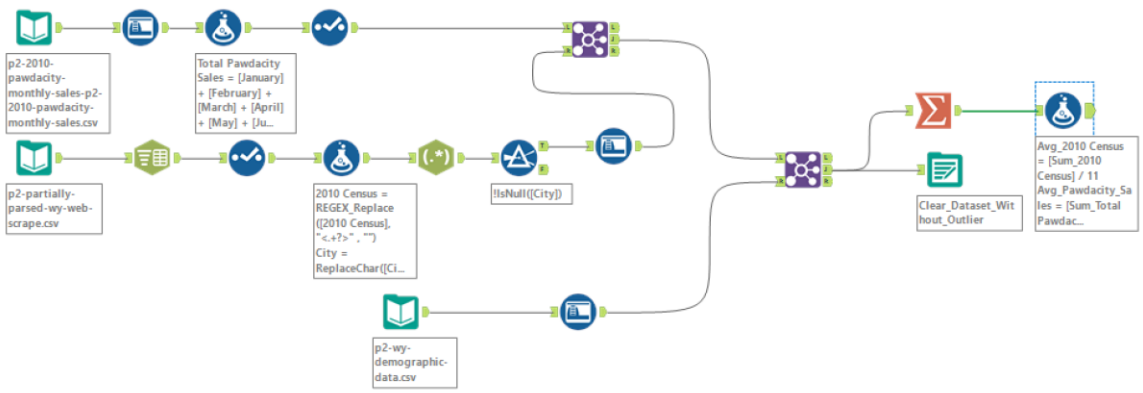
Demographic data for all stores.
Dataset for all existing stores.

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*	*343027.63*
Households with Under 18	*34,064*	*3096.73*
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 table below shows cities information:

City	Total Pawd	2010 Censi	Land Area	Household	Population	Total Families
Buffalo	185328	4585	3115.508	746	1.55	1819.5
Casper	317736	35316	3894.309	7788	11.16	8756.32
Cheyenne	917892	59466	1500.178	7158	20.34	14612.64
Cody	218376	9520	2998.957	1403	1.82	3515.62
Douglas	208008	6120	1829.465	832	1.46	1744.08
Evanston	283824	12359	999.4971	1486	4.95	2712.64
Gillette	543132	29087	2748.853	4052	5.8	7189.43
Powell	233928	6314	2673.575	1251	1.62	3134.18
Riverton	303264	10615	4796.86	2680	2.34	5556.49
Rock Sprin	253584	23036	6620.202	4022	2.78	7572.18
Sheridan	308232	17444	1893.977	2646	8.98	6039.71

The table below shows the quartiles for every data in the dataset:

Q1 (25th percentile)	7917	226152	1327	1861.5	2	2923.5
Q3 (7th percentile)	26061.5	312984	4037	3505	7.5	7380.5
IQR (Q3-Q1)	18144.5	86832	2710	1643.5	5.5	4457
Lower Q1-1.5xIQR	-19299.75	95904	-2738	-603.75	-6.25	-3762
Upper Q3+1.5xIQR	53278.25	443232	8102	5970.25	15.75	14066

We can observe that there are three cities we can considers as outliers:

1- **Cheyenne outliers' Fields:**
   - Total Pawdacity Sales.
   - 2010 Census.
   - Population Density.
   - Total Families.
2- **Rock Springs outliers' Field:**
   - Land Area.
3- **Gillette outliers' Field:**
   - Total Pawdacity Sales.

The city we exclude is **Cheyenne** since this city have larger values compared to the other cities. If we look to all outliers of **Cheyenne** they seem close to the upper fence except for the dield "total Pawdacity sales (917892)" which is much higher than the upper fence thus if we want to build I reliable model for predicting the location if new store we remove **Cheyenne.**
For **Gillette** and **Rock Springs** we keep them even with the outliers since we have small dataset, and they all have an outlier in one field only.

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.