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 manager needs to decision that predicts the most suitable city of the 14th store, based on predicted yearly sales from other stores.

2. What data is needed to inform those decisions?

   To conclude the useful decision, we need those following data:
   - All sales data of the Pawdacity stores for year 2010 in monthly based.
   - All competitor stores NAICS data on the most current their sales where total sales is equal to one year of sales.
   - A partially parsed data file that can be used for population numbers.
   - Demographic data like households with individuals under 18, land area, population density, and total families.

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

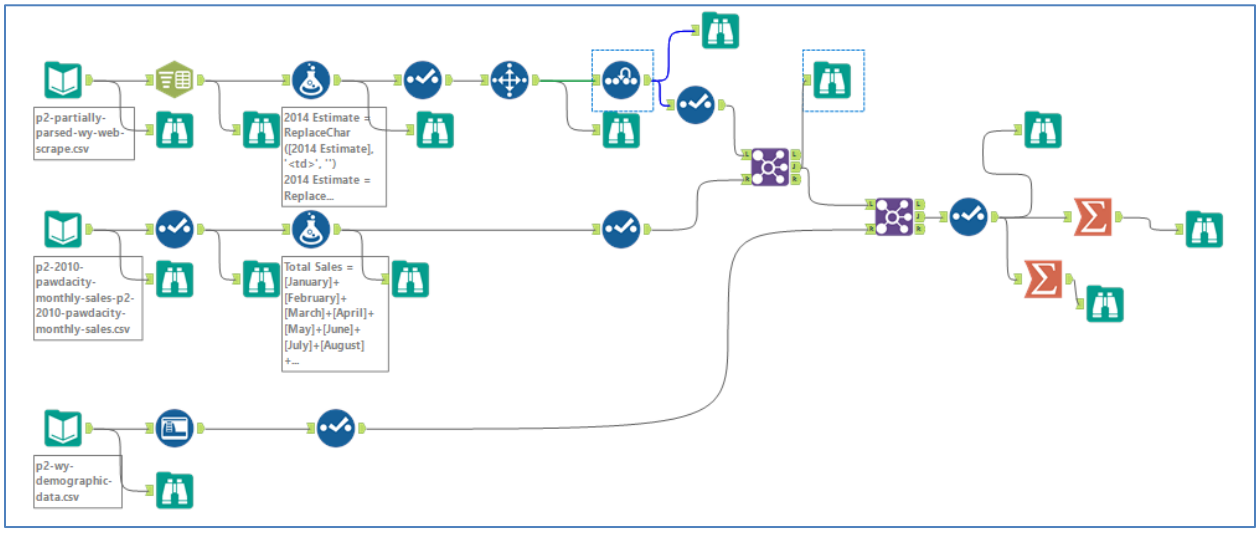
To calculate the average of columns, we need to do the following steps:

1- Cleaning and transforming city|country, 2010 Census columns in the partially-parsed-wy-web-scrape.csv.
2- Calculate the total sales for each store in pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales.csv .
3- Join the pervious data files with p2-wy-demographic-data.csv by joining city column to calculate the average of census population, total Pawdacity Sales, Households with under 18, land area, population density and total families columns.

Table 1: Sum and Averages of Dataset Columns

Column	Sum	Average
Census Population	*213,862*	*19,442*
Total Pawdacity Sales	*3,773,304*	*343,027*
Households with Under 18	*34,064*	*3,096.72*
Land Area	*33,071*	*3,006.48*
Population Density	*63*	*5.70*
Total Families	*62,653*	*5,695.70*

Figure 1: The workflow of project



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.

To identify if there are outliers, I create two tables which are the city statistics, and IQR table for the attributes.

Table 2: Cities statistics

City	Total Sales	Census Population	Land Area	Households with Under 18	Population Density	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 Springs	253584	23036	6620.202	4022	2.78	7572.18
Sheridan	308232	17444	1893.977	2646	8.98	6039.71

As the above table, there are many outlier values of the city in the dataset like Cheyenne, Gillette and Rock Spring. Also, the outlier values are colored by gray. Whereas the Cheyenne record (row) has 4 values above the upper fence, I decide to remove that record because of most value of its attributes is an outlier. This decision based on IQR table.

Table 3: IQR

Item	Total Sales	Census Population	Land Area	Households with Under 18	Population Density	Total Families
Q1	226152	7917	1861.721	1327	1.72	2923.41
Q2	283824	12359	2749	2646	3	5556
Q3	312984	26061.5	3504.908	4037	7.39	7380.805
IQR	86832	18144.5	1643.187	2710	5.67	4457.395
Lower fence	95904	-19299.8	-603.06	-2738	-6.785	-3762.68
Upper fence	443232	53278.25	5969.689	8102	15.895	14066.9

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.