Project 2.1: Data Cleanup

https://classroom.udacity.com/nanodegrees/nd008/parts/8d60a887-d4c1-4b0e-8873-b2f36435eb39/project

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?

The location of Pawdacity's 14th store needs to be decided.

2. What data is needed to inform those decisions?

In order to decide which location is the best possible for Pawdacity's new store, the following data will be needed:

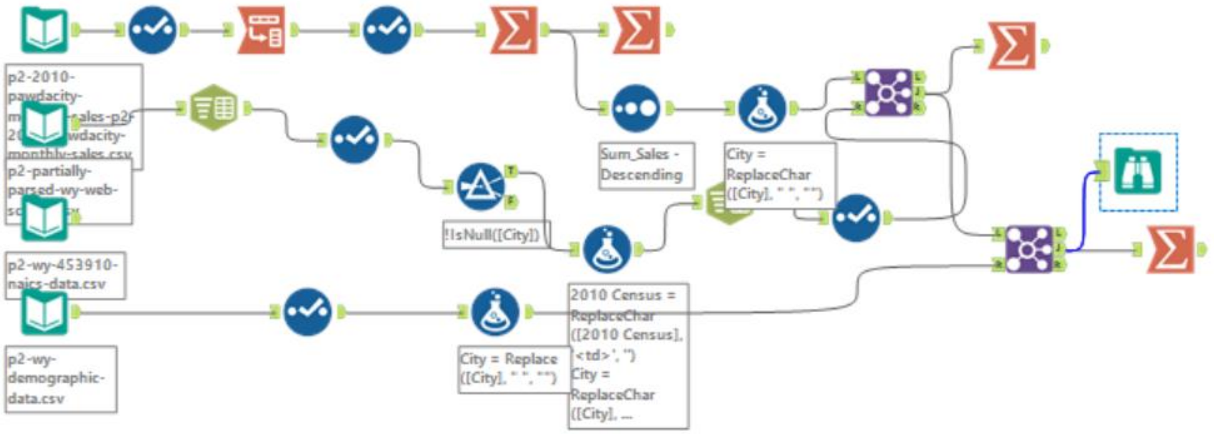
○ *Store locations of the company's **competitors***
○ ***Monthly sales data** for all of the Pawdacity stores for the year 2010*
○ ***Population** data*
○ ***Demographic** data for each city and county in the state of Wyoming*

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.)

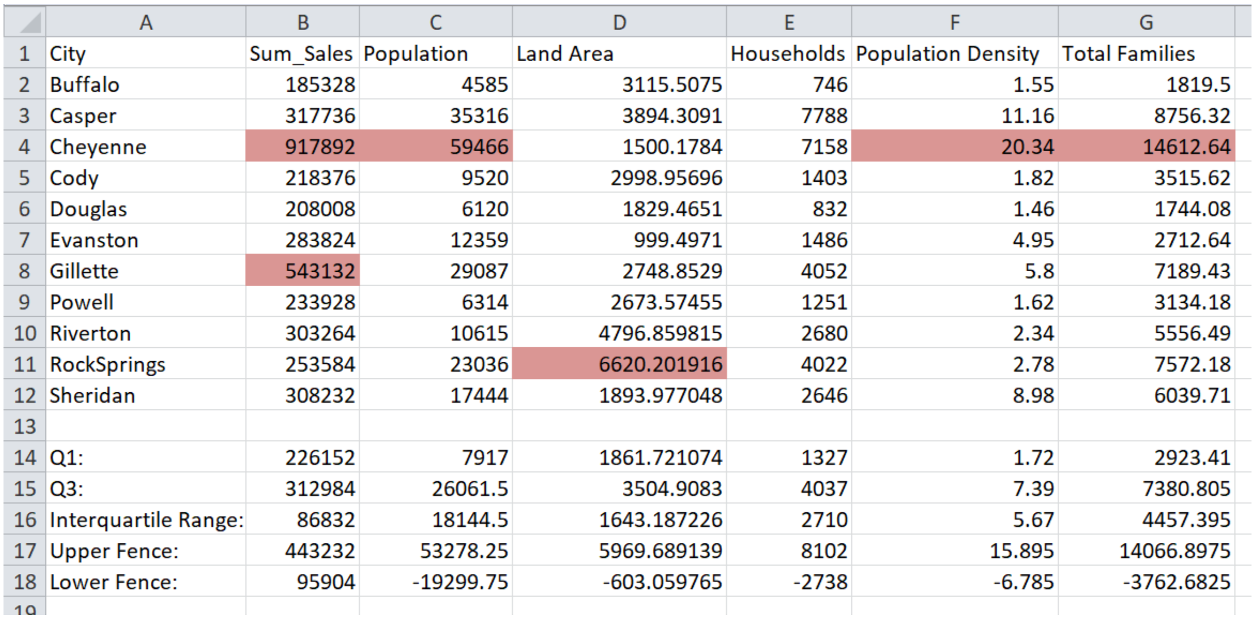
(In addition provide the averages on your data set here to help reviewers check your work. You should round up to two decimal places, ex: 1.24)

Column	Sum	Average
City		
Census Population	*213,862*	*19,442*
Total Pawdacity Sales	*3,773,304*	*343,027.64*
Households with Under 18	*34,064*	*3,096.73*
Land Area	*33,071*	*3,006.49*
Population Density	*63*	*5.71*
Total Families	*62,653*	*5,695.71*


p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales.csv
p2-partially-parsed-wy-web-scrape.csv
p2-wy-453910-naics-data.csv
p2-wy-demographic-data.csv
!IsNull([City])
Sum_Sales - Descending
City = ReplaceChar ([City], " ", "")
2010 Census = ReplaceChar ([2010 Census], '<td>', '')
City = ReplaceChar ([City], ...
City = Replace ([City], " ", "")
T
F
L
J
R

Step 3: Dealing with Outliers

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
1	City	Sum_Sales	Population	Land Area	Households	Population Density	Total Families
2	Buffalo	185328	4585	3115.5075	746	1.55	1819.5
3	Casper	317736	35316	3894.3091	7788	11.16	8756.32
4	Cheyenne	917892	59466	1500.1784	7158	20.34	14612.64
5	Cody	218376	9520	2998.95696	1403	1.82	3515.62
6	Douglas	208008	6120	1829.4651	832	1.46	1744.08
7	Evanston	283824	12359	999.4971	1486	4.95	2712.64
8	Gillette	543132	29087	2748.8529	4052	5.8	7189.43
9	Powell	233928	6314	2673.57455	1251	1.62	3134.18
10	Riverton	303264	10615	4796.859815	2680	2.34	5556.49
11	RockSprings	253584	23036	6620.201916	4022	2.78	7572.18
12	Sheridan	308232	17444	1893.977048	2646	8.98	6039.71
13							
14	Q1:	226152	7917	1861.721074	1327	1.72	2923.41
15	Q3:	312984	26061.5	3504.9083	4037	7.39	7380.805
16	Interquartile Range:	86832	18144.5	1643.187226	2710	5.67	4457.395
17	Upper Fence:	443232	53278.25	5969.689139	8102	15.895	14066.8975
18	Lower Fence:	95904	-19299.75	-603.059765	-2738	-6.785	-3762.6825
19							

Identifying outliers in the data helps us understand how vulnerable our model would be to a small set of observations. In all the cases above, the data itself could actually be valid and accurate, or it could be erroneous information. In either case, what we handle the outliers will depend on the purpose of the analysis.

After extracting the dataset from Alteryx, with all the information needed, I opened it using MS Excel to calculate the IQR (Interquartile Range). Cells highlighted with light red color, are below

Q1 or above Q3. I would remove ***Cheyenne*** *city, as it looks like most of its numbers are out of the two fences.*