Project 2.1: Data Cleanup

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?

Ans 1: Pawdacity is an already established pet store in 13 locations in Wyoming state. An analysis is required to decide the 14th location for opening the new Pawdacity pet store based on the predicted yearly sales.

2. What data is needed to inform those decisions?

Ans 2: The data required to make the above mentioned decision is as follows:

- City wise monthly sales of all Pawdacity stores for the year 2010.
- Country and city wise census population data for the year 2010.
- Demographics data, consisting of the following information of every city and country in Wyoming State:
  - Population Density
  - Total number of families
  - Households with individuals under the age of 18 years
  - Land Area

All of this data is available in different data sources. We will have to merge the data sets based on one key data field **(City column)** and do necessary cleaning to be able to achieve the required training data set.

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.

I have used the following data sources to build my training data set:

- p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales.csv
- p2-partially-parsed-wy-web-scrape.csv
- p2-wy-demographic-data.csv

Data cleaning steps:

Data Source: p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales.csv

Step 1: Pivot the column wise monthly sales data into rows
Step 2: Group the data by cities
Step 3: Sum each's city sales data of all 12 months

Data Source: p2-partially-parsed-wy-web-scrape.csv

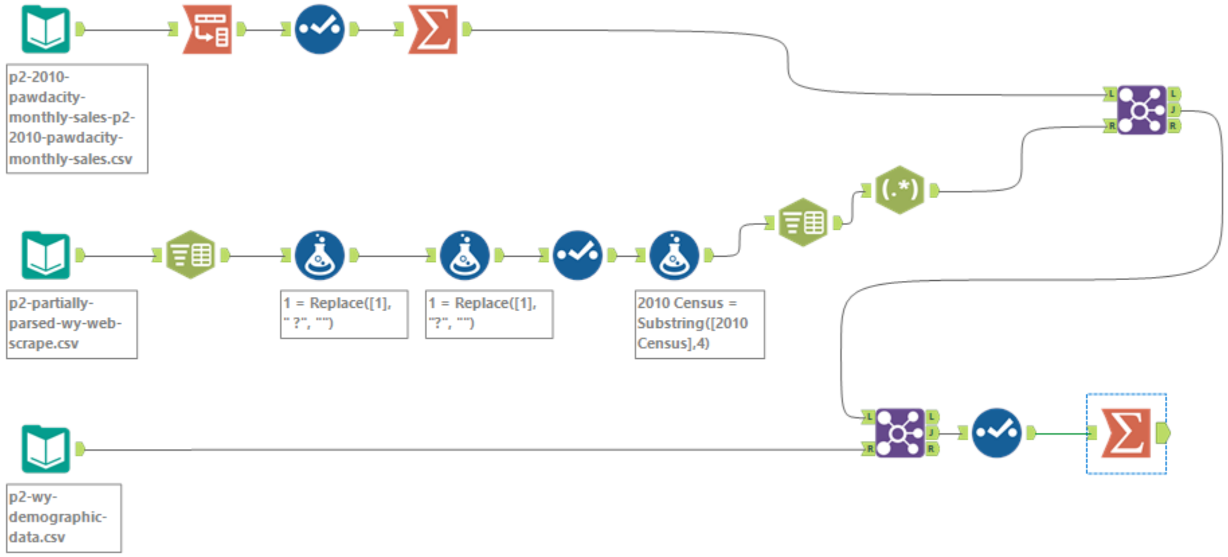
Step 1: Split the column "City|Country" on the basis of the delimiter "|"
Step 2: Clean the column containing city name by removing extra characters before and after any names.
Step 3: Extract the numbers given in the column "2010 Census" using the substring function and then splitting the column on the basis of the delimiter "<"
Step 4: Remove the commas in the new "2010 Census" column to be able to convert it in integer.

Data Blending:

Join all these datasets on the basis of one key parameter "City" to get a resultant dataset that could be used for training. The training dataset contains the following columns:

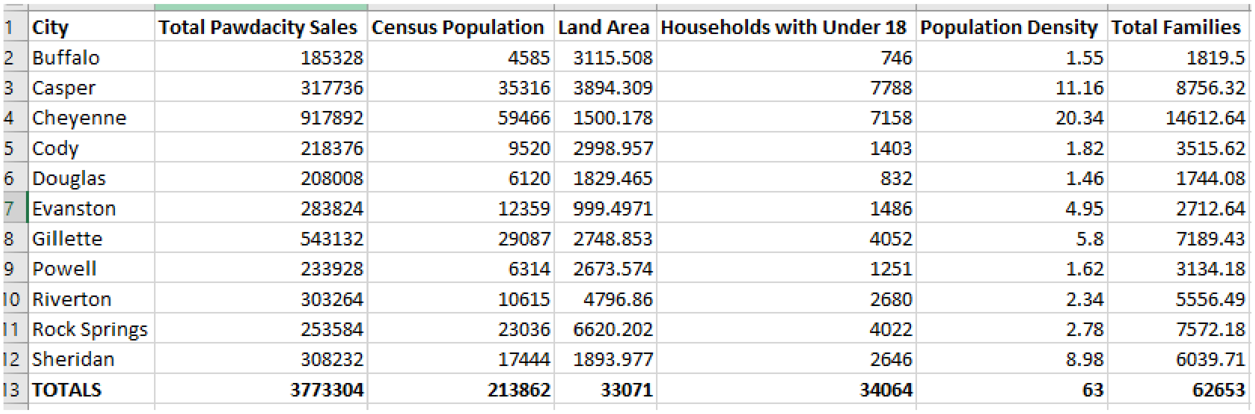
- Total Pawdacity Sales
- Census Population
- Land Area
- Households with Under 18
- Population Density
- Total Families

For reference, my Alteryx workflow looks like this:


p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales.csv
p2-partially-parsed-wy-web-scrape.csv
1 = Replace([1], " ?", "")
1 = Replace([1], "?", "")
2010 Census = Substring([2010 Census],4)
p2-wy-demographic-data.csv
(.*)
L
R
L
J
R
L
R
L
J
R

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

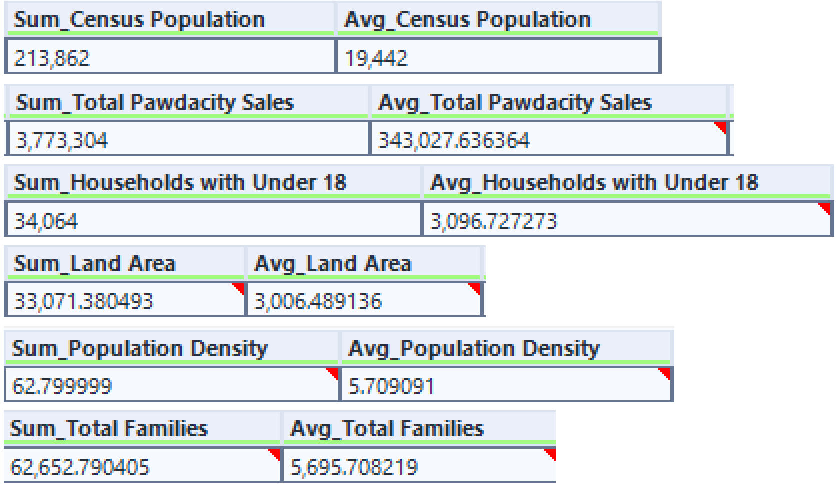
For reference, I have put the training dataset into an excel and summed all the columns to verify my working:



1	City	Total Pawdacity Sales	Census Population	Land Area	Households with Under 18	Population Density	Total Families
2	Buffalo	185328	4585	3115.508	746	1.55	1819.5
3	Casper	317736	35316	3894.309	7788	11.16	8756.32
4	Cheyenne	917892	59466	1500.178	7158	20.34	14612.64
5	Cody	218376	9520	2998.957	1403	1.82	3515.62
6	Douglas	208008	6120	1829.465	832	1.46	1744.08
7	Evanston	283824	12359	999.4971	1486	4.95	2712.64
8	Gillette	543132	29087	2748.853	4052	5.8	7189.43
9	Powell	233928	6314	2673.574	1251	1.62	3134.18
10	Riverton	303264	10615	4796.86	2680	2.34	5556.49
11	Rock Springs	253584	23036	6620.202	4022	2.78	7572.18
12	Sheridan	308232	17444	1893.977	2646	8.98	6039.71
13	**TOTALS**	**3773304**	**213862**	**33071**	**34064**	**63**	**62653**

Column	**Sum**	**Average**
Census Population	*213,862*	***19,422***
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***

**Screenshots from Alteryx*



Sum_Census Population	Avg_Census Population
213,862	19,442

Sum_Total Pawdacity Sales	Avg_Total Pawdacity Sales
3,773,304	343,027.636364

Sum_Households with Under 18	Avg_Households with Under 18
34,064	3,096.727273

Sum_Land Area	Avg_Land Area
33,071.380493	3,006.489136

Sum_Population Density	Avg_Population Density
62.799999	5.709091

Sum_Total Families	Avg_Total Families
62,652.790405	5,695.708219

Step 3: Dealing with Outliers

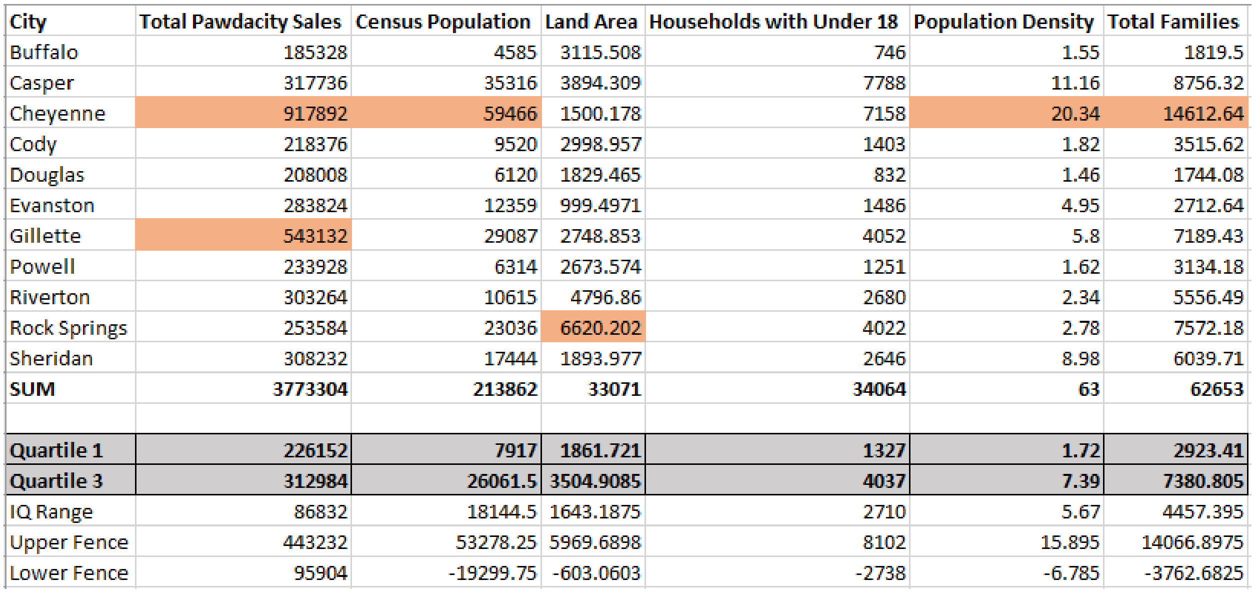
Answer these questions

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.

Ans 3: The training set that we created using multiple data sources and then blending and cleaning it in Alteryx, we checked the resultant dataset for any outliers using the following steps:

- Step 1: Calculate first quartile using the excel formula =QUARTILE.INC(array, 1)
- Step 2: Calculate third quartile using the excel formula =QUARTILE.INC(array, 3)
- Step 3: Calculate the interquartile range by finding out the difference between the value of third quartile and the first quartile
- Step 4: Calculate the upper fence by using the formula (Interquartile Range * 1.5) + Value of quartile 3
- Step 5: Calculate the lower fence by using the formula; Value of quartile 1 - (Interquartile Range * 1.5)
- Step 6: Identify the values in respective columns which are above its upper fence and below its lower fence.

I did the above explained steps in Excel and the highlighted values were identified as the outliers. I used conditional formatting in excel to highlight the outliers.



City	Total Pawdacity 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	2673.574	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
SUM	**3773304**	**213862**	**33071**	**34064**	**63**	**62653**
Quartile 1	**226152**	**7917**	**1861.721**	**1327**	**1.72**	**2923.41**
Quartile 3	**312984**	**26061.5**	**3504.9085**	**4037**	**7.39**	**7380.805**
IQ Range	86832	18144.5	1643.1875	2710	5.67	4457.395
Upper Fence	443232	53278.25	5969.6898	8102	15.895	14066.8975
Lower Fence	95904	-19299.75	-603.0603	-2738	-6.785	-3762.6825

As seen in the table above, the outliers are present in the data of the cities **Cheyenne, Gillette** and **Rock Springs.**

From the two options given to deal with the outliers, we would choose to remove record having outliers since imputation is not a viable option here.

We choose to remove Cheyenne only although Gillette and Rock Springs also contain outliers too. The rationale behind this decision is that since the data is already of only 11 rows, we can’t afford to delete a lot of rows and also, Cheyenne has the most outliers and the outliers of Gillette and Rock Springs are only one so we will keep them.