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?

Pawdacity would like to open a 14th store with the help of the data on the previous 13 stores in the state of Wyoming. We will need to calculate the yearly sales for the stores to predict the sales for the new store. We have 3 different files that contain different information and we would need to join them and prepare a final dataset that will be used later for creating a linear regression model to reach a conclusion on as to where to open the 14th store and predict its yearly sales and such.

2. What data is needed to inform those decisions?

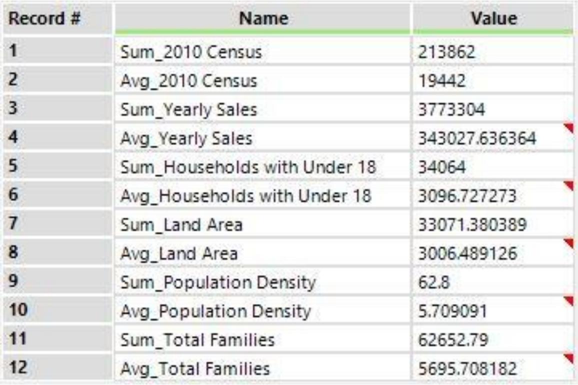
We will require the yearly sales of each store in the city/county, the census from the previous report (2010 census) to get an idea of the number of people visiting the stores. The land area of each city/county, the population density with respect to the land area, the total number of families and households with under 18 individuals. The above said data will give us a better understanding of how they are correlated to sales and what factors affect or contribute towards obtaining a good yearly sales standing.

Step 2: Building the Training Set

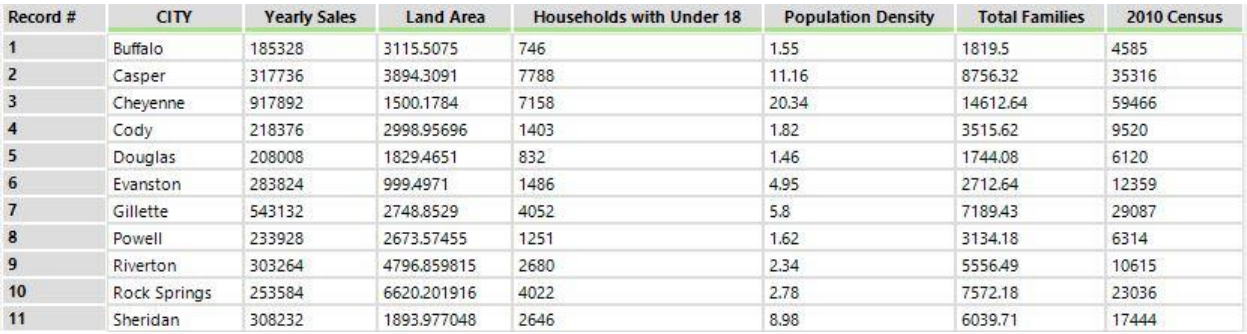
Build your training set given the data provided to you. Your column sums of your dataset should match the sums in the table below.

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*	*19442*
Total Pawdacity Sales	*3,773,304*	*343028*
Households with Under 18	*34,064*	*3097*
Land Area	*33,071*	*3006*
Population Density	*63*	*5.7*
Total Families	*62,653*	*5696*



Record #	Name	Value
1	Sum_2010 Census	213862
2	Avg_2010 Census	19442
3	Sum_Yearly Sales	3773304
4	Avg_Yearly Sales	343027.636364
5	Sum_Households with Under 18	34064
6	Avg_Households with Under 18	3096.727273
7	Sum_Land Area	33071.380389
8	Avg_Land Area	3006.489126
9	Sum_Population Density	62.8
10	Avg_Population Density	5.709091
11	Sum_Total Families	62652.79
12	Avg_Total Families	5695.708182



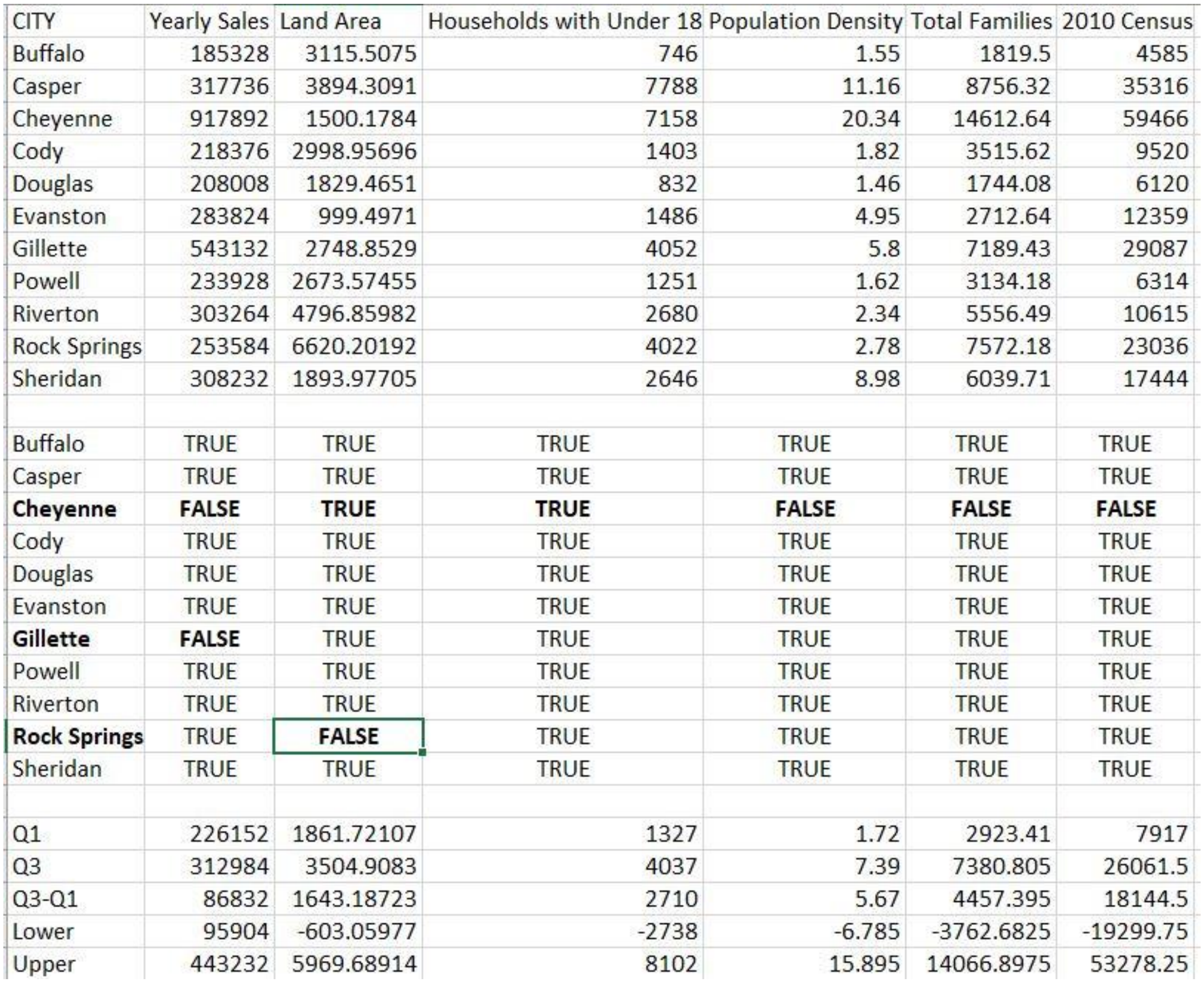
Record #	CITY	Yearly Sales	Land Area	Households with Under 18	Population Density	Total Families	2010 Census
1	Buffalo	185328	3115.5075	746	1.55	1819.5	4585
2	Casper	317736	3894.3091	7788	11.16	8756.32	35316
3	Cheyenne	917892	1500.1784	7158	20.34	14612.64	59466
4	Cody	218376	2998.95696	1403	1.82	3515.62	9520
5	Douglas	208008	1829.4651	832	1.46	1744.08	6120
6	Evanston	283824	999.4971	1486	4.95	2712.64	12359
7	Gillette	543132	2748.8529	4052	5.8	7189.43	29087
8	Powell	233928	2673.57455	1251	1.62	3134.18	6314
9	Riverton	303264	4796.859815	2680	2.34	5556.49	10615
10	Rock Springs	253584	6620.201916	4022	2.78	7572.18	23036
11	Sheridan	308232	1893.977048	2646	8.98	6039.71	17444

Step 3: Dealing with Outliers

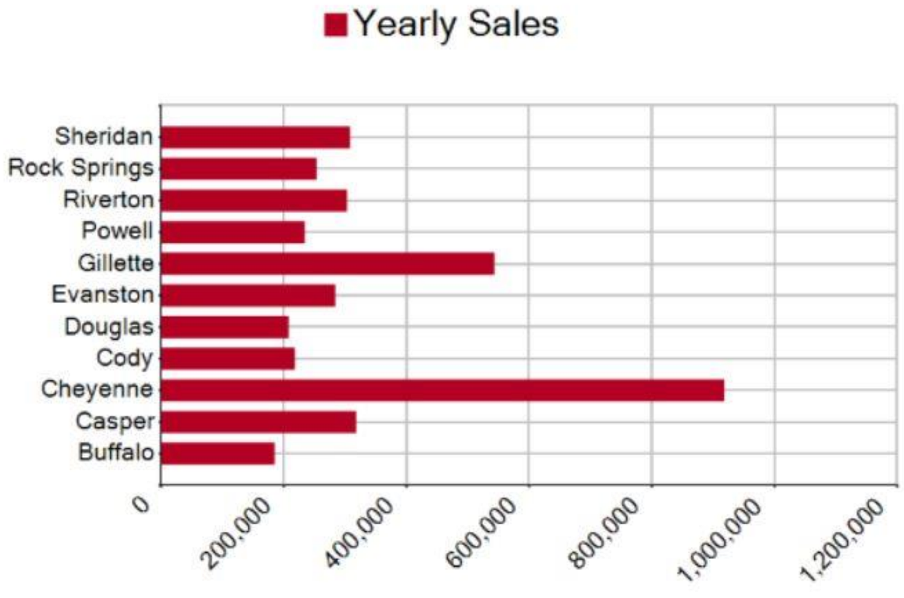
Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.

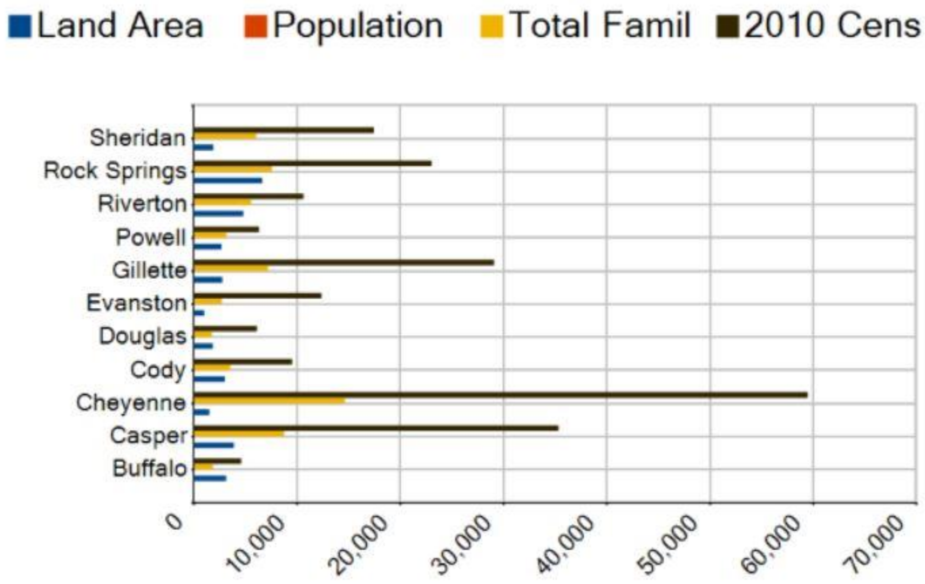
The City Cheyenne on closer inspection is considered an outlier.
Looking at the scatterplots between city(Cheyenne) and population density (20.34) *and* land area (1501) *and* yearly sales (917892).

For such a small area, it is very hard to think about a very high population density and high yearly sales. Therefore, we can regard this city as an outlier and remove this one city from the dataset before we submit it for linear regression and predicting model.



CITY	Yearly Sales	Land Area	Households with Under 18	Population Density	Total Families	2010 Census
Buffalo	185328	3115.5075	746	1.55	1819.5	4585
Casper	317736	3894.3091	7788	11.16	8756.32	35316
Cheyenne	917892	1500.1784	7158	20.34	14612.64	59466
Cody	218376	2998.95696	1403	1.82	3515.62	9520
Douglas	208008	1829.4651	832	1.46	1744.08	6120
Evanston	283824	999.4971	1486	4.95	2712.64	12359
Gillette	543132	2748.8529	4052	5.8	7189.43	29087
Powell	233928	2673.57455	1251	1.62	3134.18	6314
Riverton	303264	4796.85982	2680	2.34	5556.49	10615
Rock Springs	253584	6620.20192	4022	2.78	7572.18	23036
Sheridan	308232	1893.97705	2646	8.98	6039.71	17444
Buffalo	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Casper	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Cheyenne	**FALSE**	**TRUE**	**TRUE**	**FALSE**	**FALSE**	**FALSE**
Cody	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Douglas	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Evanston	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Gillette	**FALSE**	TRUE	TRUE	TRUE	TRUE	TRUE
Powell	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Riverton	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Rock Springs	TRUE	**FALSE**	TRUE	TRUE	TRUE	TRUE
Sheridan	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Q1	226152	1861.72107	1327	1.72	2923.41	7917
Q3	312984	3504.9083	4037	7.39	7380.805	26061.5
Q3-Q1	86832	1643.18723	2710	5.67	4457.395	18144.5
Lower	95904	-603.05977	-2738	-6.785	-3762.6825	-19299.75
Upper	443232	5969.68914	8102	15.895	14066.8975	53278.25

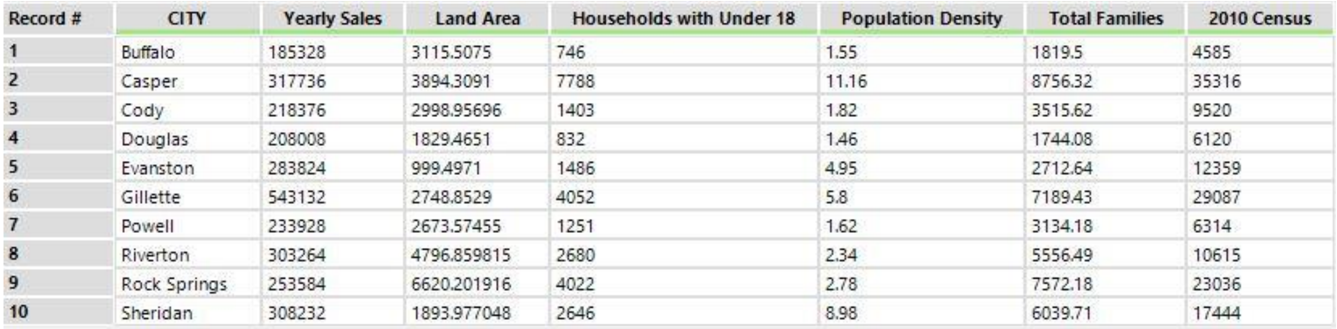




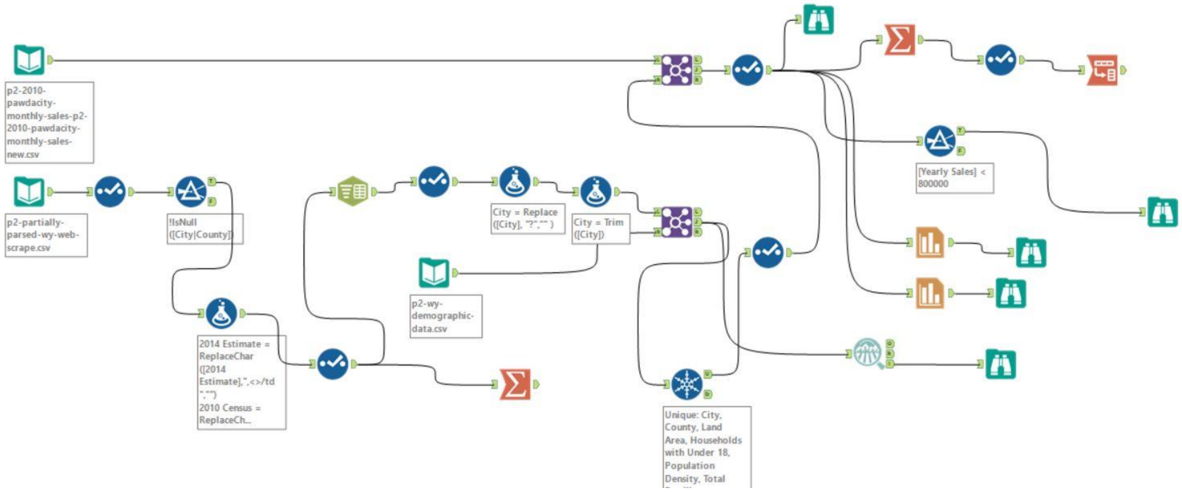
From the above table, we can see after calculating the IQR values for each individual variable, the major outliers are from cities Cheyenne and Gillette.

The City Cheyenne on closer inspection is considered an outlier.

Looking at the scatterplots between city(Cheyenne) and population density (20.34) *and* land area (1501) *and* yearly sales (917892).



Record #	CITY	Yearly Sales	Land Area	Households with Under 18	Population Density	Total Families	2010 Census
1	Buffalo	185328	3115.5075	746	1.55	1819.5	4585
2	Casper	317736	3894.3091	7788	11.16	8756.32	35316
3	Cody	218376	2998.95696	1403	1.82	3515.62	9520
4	Douglas	208008	1829.4651	832	1.46	1744.08	6120
5	Evanston	283824	999.4971	1486	4.95	2712.64	12359
6	Gillette	543132	2748.8529	4052	5.8	7189.43	29087
7	Powell	233928	2673.57455	1251	1.62	3134.18	6314
8	Riverton	303264	4796.859815	2680	2.34	5556.49	10615
9	Rock Springs	253584	6620.201916	4022	2.78	7572.18	23036
10	Sheridan	308232	1893.977048	2646	8.98	6039.71	17444


p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales-new.csv
p2-partially-parsed-wy-web-scrape.csv
!IsNull ([City|County])
2014 Estimate = ReplaceChar ([2014 Estimate],",<>/td ","")
2010 Census = ReplaceCh...
City = Replace ([City], "?","")
City = Trim ([City])
p2-wy-demographic-data.csv
Unique: City, County, Land Area, Households with Under 18, Population Density, Total
[Yearly Sales] < 800000