

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 decision for Pawdacity was to choose the city for its 14th store based on the data of census population, total sales, households under 18, land area, population density and total families.

2. What data is needed to inform those decisions?

3 datasets were required for the decision:

- ***Monthly sales for all Pawdacity stores for 2010***
- ***Partially parsed data file for population census***
- ***Demographic data for each city in 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
Census Population	213,862	19442
Total Pawdacity Sales	3,773,304	343027.636
Households with Under 18	34,064	3096.727
Land Area	33,071	3006.490
Population Density	63	5.709
Total Families	62,653	5695.708

Step 3: Dealing with Outliers

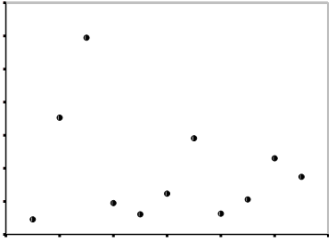
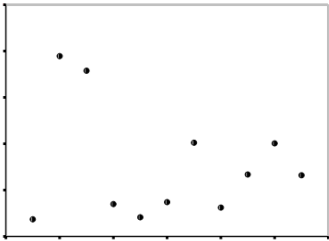
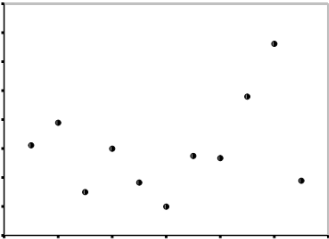
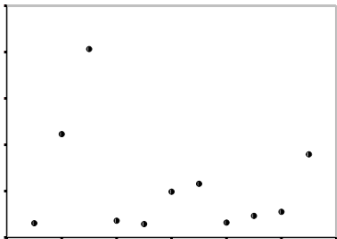
Answer these questions

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.

I picked out the following outliers in the data based on the IQR values and the Field summary report. I found 1 outlier in Gillette in Sales, 1 outlier in Rock Springs in Land Area, and outliers in 4 fields of Cheyenne City out of 6 fields. I imputed Cheyenne sales value by the mean of the column, as it is the most important variable in terms of financial decision and also difference of its actual value and the upper quartile for Cheyenne city is the highest compared to the 4 other outliers in the Cheyenne city.

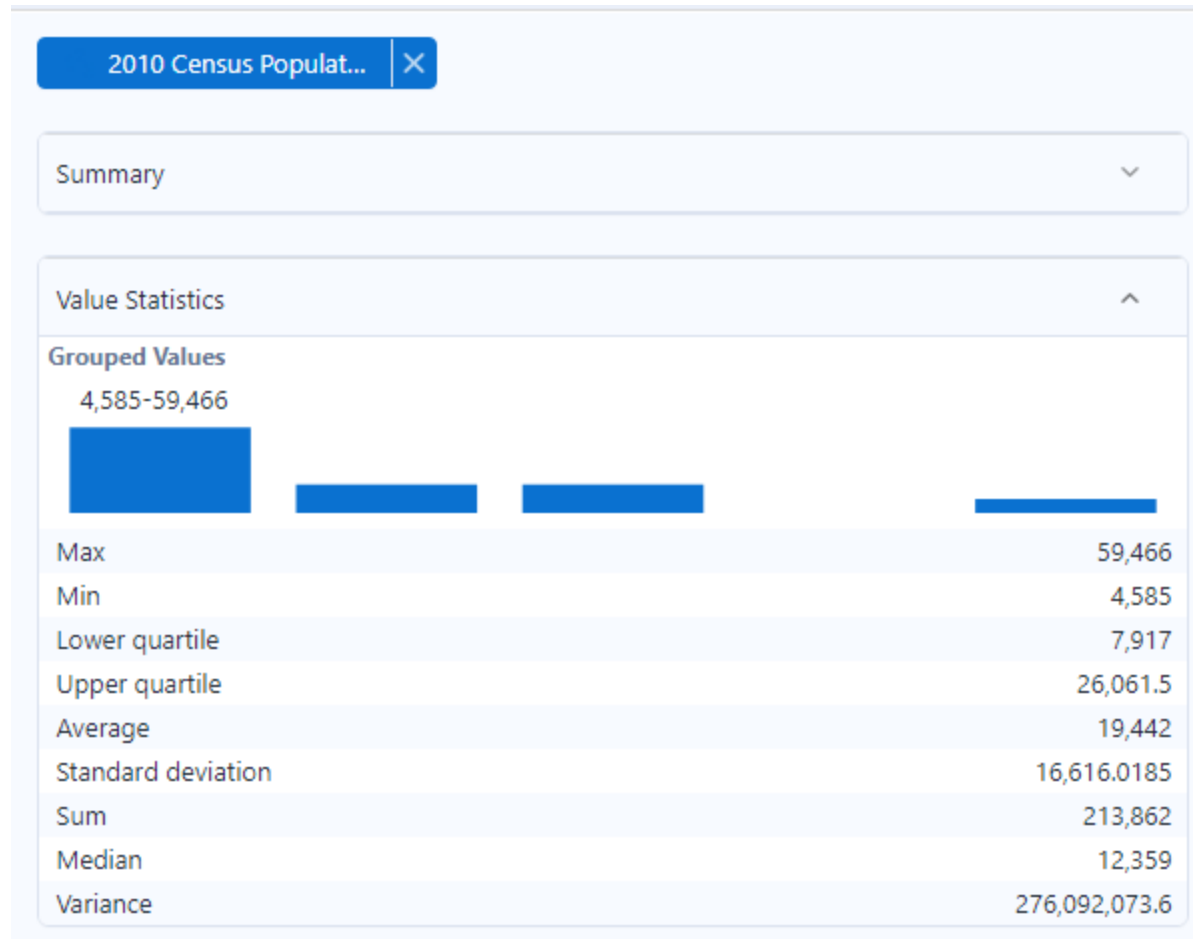
City	2010 Census Population	Land Area	Households with Under 18	Population Density	Total Families	Sum_Sales
Buffalo	4585	3115.5075	746	1.55	1819.5	185328
Casper	35316	3894.3091	7788	11.16	8756.32	317736
Cheyenne	59466	1500.1784	7158	20.34	14612.64	917892
Cody	9520	2998.95696	1403	1.82	3515.62	218376
Douglas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

Field Summary Report

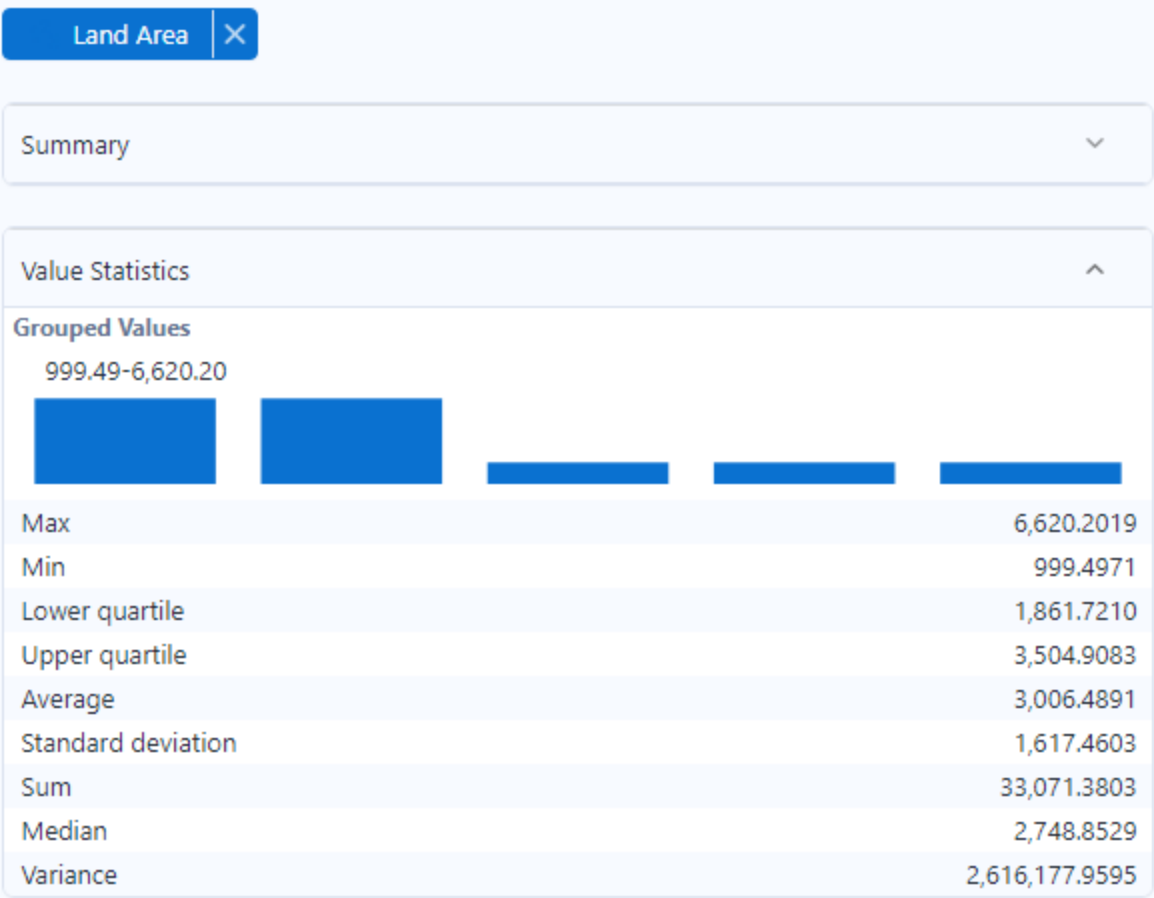
Name	Plot	% Missing	Unique Values	Min	Mean	Median	Max	Std Dev	Remarks
2010 Census Population		0.0%	11	4,585.000	19,442.000	12,359.000	59,466.000	16,616.019	
Households with Under 18		0.0%	11	746.000	3,096.727	2,646.000	7,788.000	2,453.003	
Land Area		0.0%	11	999.497	3,006.489	2,748.853	6,620.202	1,617.460	
Population Density		0.0%	11	1.460	5.709	2.780	20.340	5.850	

Sum_Sales	0.0%	11	185,328.000	343,027.636	283,824.000	917,892.000	213,538.712
Total Families	0.0%	11	1,744.080	5,695.708	5,556.490	14,612.640	3,816.050

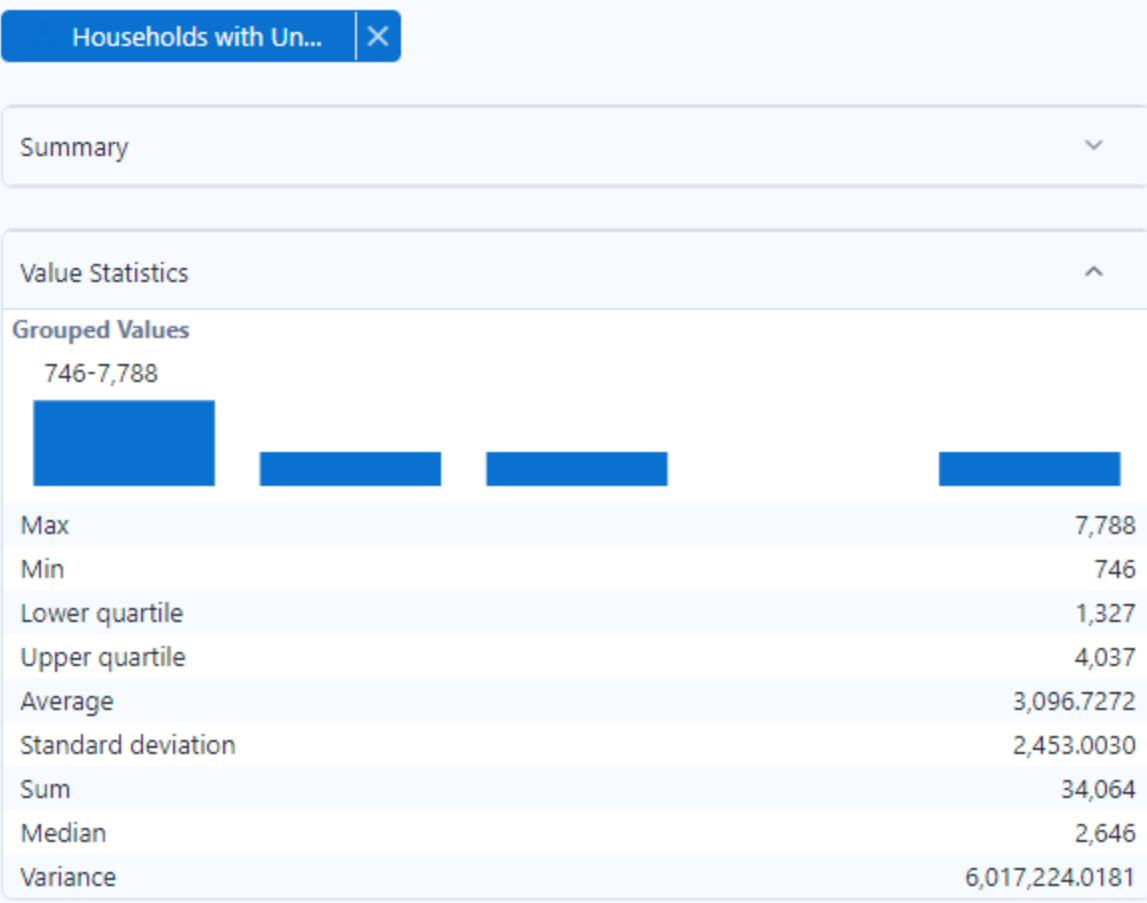
IQR of Census Population



IQR of Land Area



IQR of Household under 18



IQR of population density

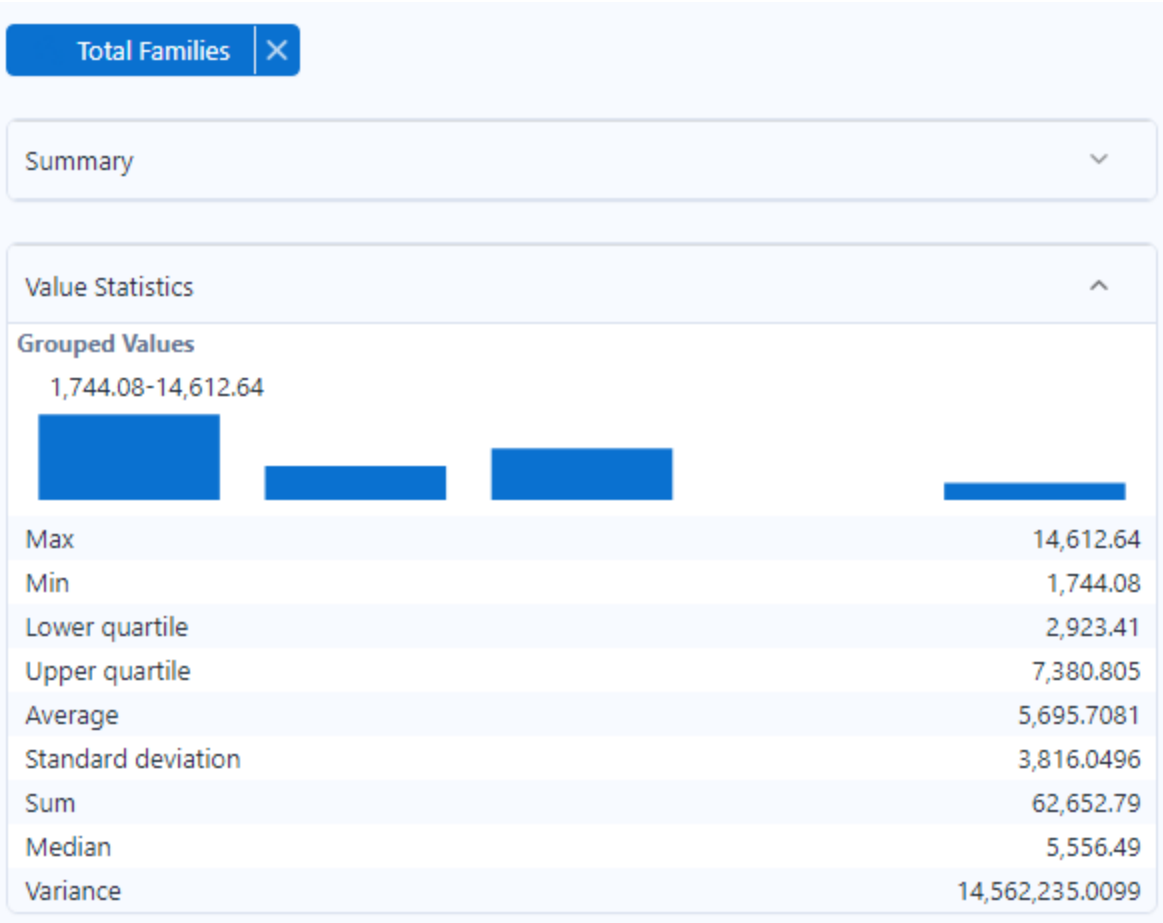
Population Density ✕

Summary ▼

Value Statistics ^

Grouped Values	
1.46-20.34	
	
Max	20.34
Min	1.46
Lower quartile	1.72
Upper quartile	7.39
Average	5.7090
Standard deviation	5.8496
Sum	62.8
Median	2.78
Variance	34.2188

IQR of total families



IQR of total sales

