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

Key Decisions:

Answer these questions

1. What decisions needs to be made?

Pawdacity, the leading pet store chain in Wyoming with 13 locations, would like to expand and open the 14th location. We need to conduct an analysis to recommend Pawdacity's newest store based on the projected annual sales.

2. What data is needed to inform those decisions?

The following data is needed

- The monthly sales data for all of the Pawdacity stores for the year 2010.
- NAICS data on the most current sales of all competitor
- Population of all the cities.
- Demographic data (Households with individuals under 18, Land Area, Population Density, and Total Families) for each city and county in the state of Wyoming.

Step 2: Building the Training Set

Based on the workflow that I generated in Alteryx, my dataset column amounts match the amounts in the table below. The screenshot below shows the 11 cities I get after the data is cleaned, and the SUM of each metric is at the bottom.

Record	Sales	2010 Census	Land Area	Households with Under 18	Population Density	Total Families
1	185328	4585	3115.5075	746	1.55	1819.5
2	317736	35316	3894.3091	7788	11.16	8756.32
3	917892	59466	1500.1784	7158	20.34	14612.64
4	218376	9520	2998.95696	1403	1.82	3515.62
5	208008	6120	1829.4651	832	1.46	1744.08
6	283824	12359	999.4971	1486	4.95	2712.64
7	543132	29087	2748.8529	4052	5.8	7189.43
8	233928	6314	2673.57455	1251	1.62	3134.18
9	303264	10615	4796.859815	2680	2.34	5556.49
10	253584	23036	6620.201916	4022	2.78	7572.18
11	308232	17444	1893.977048	2646	8.98	6039.71
12	3773304	213862	33071.380389	34064	62.8	62652.79

Column	Sum	Average
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

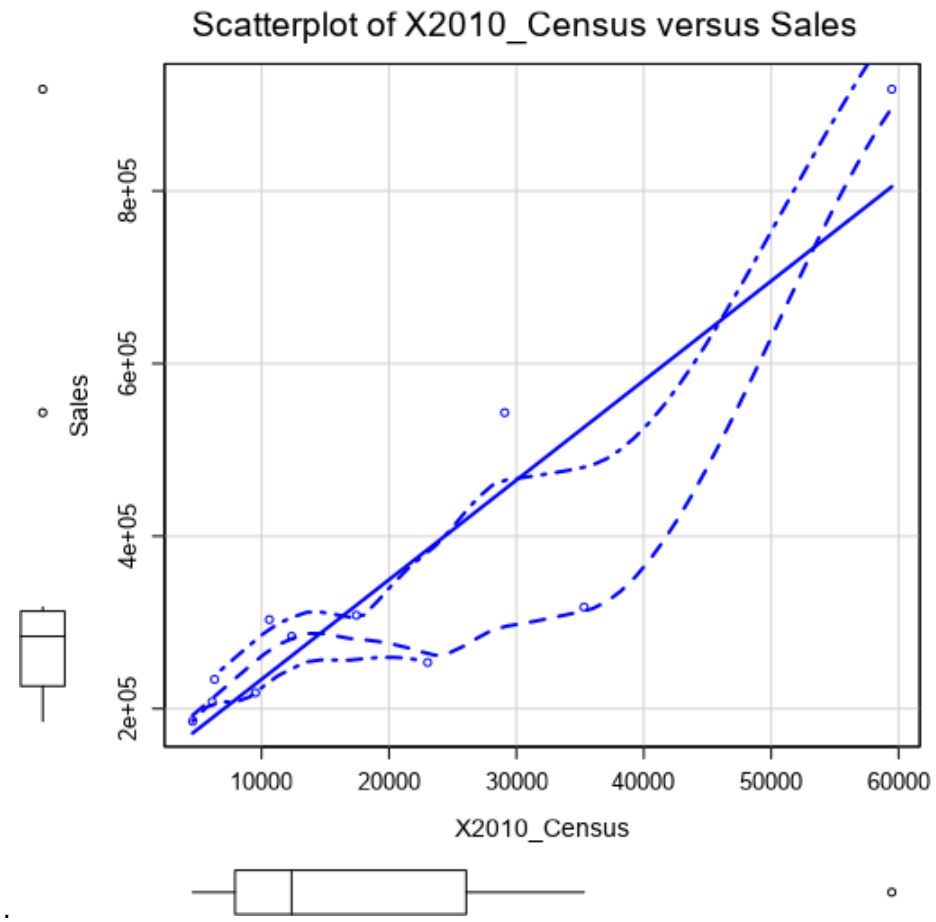
Step 3: Dealing with Outliers

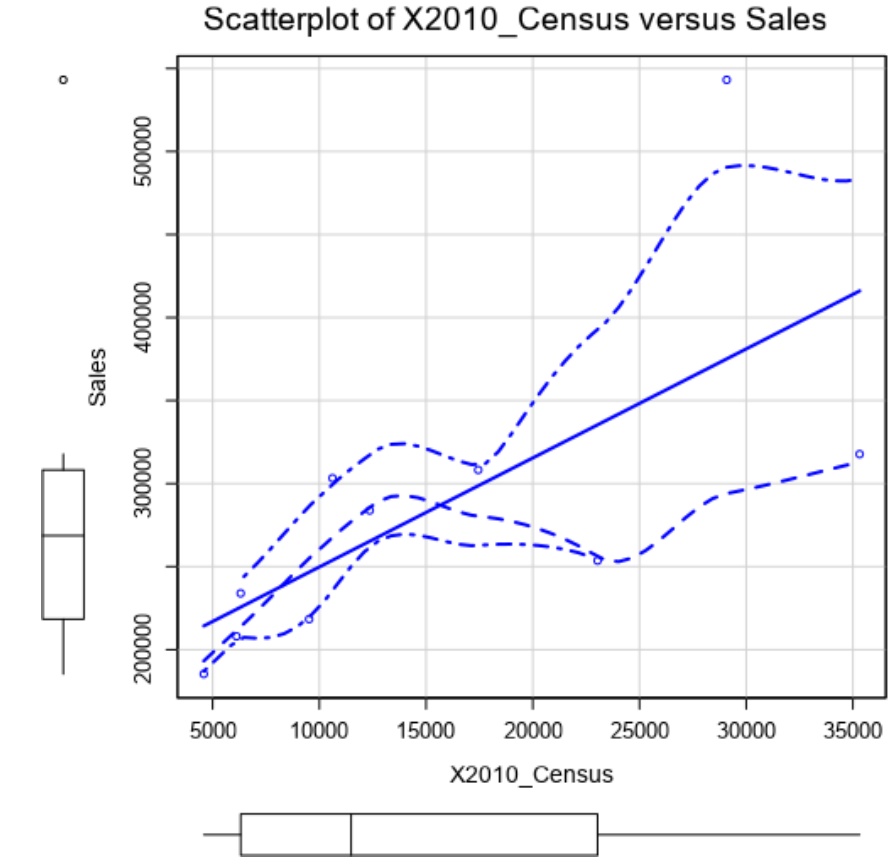
Answer these questions

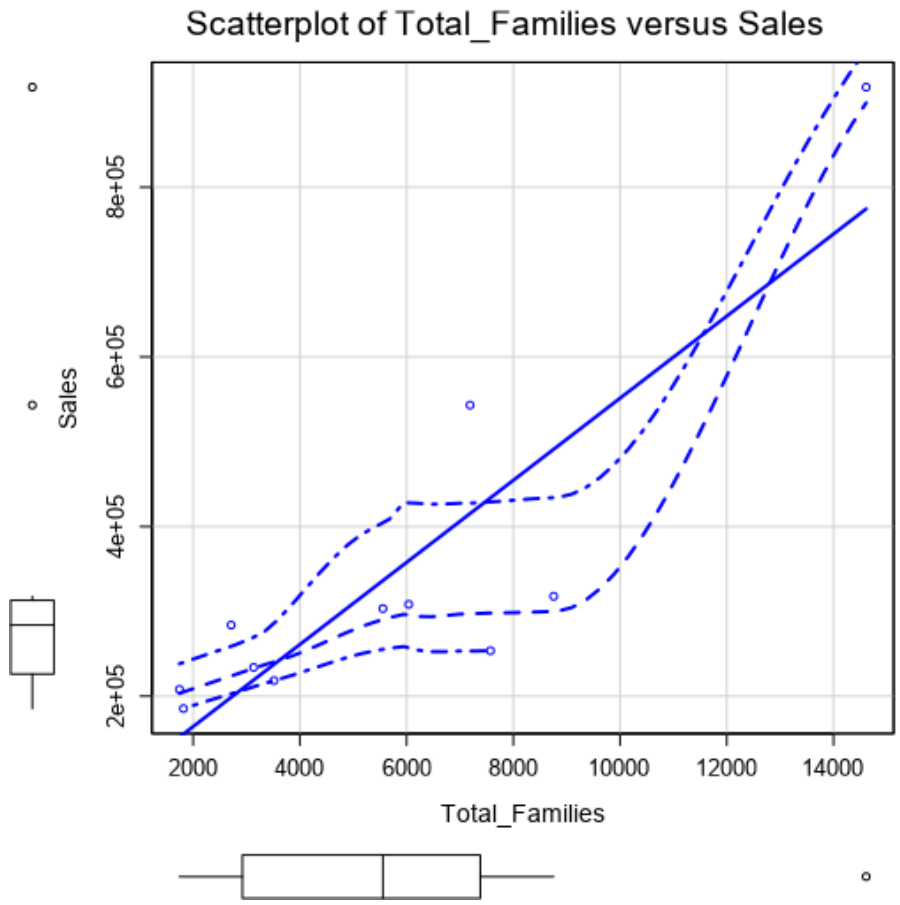
Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.

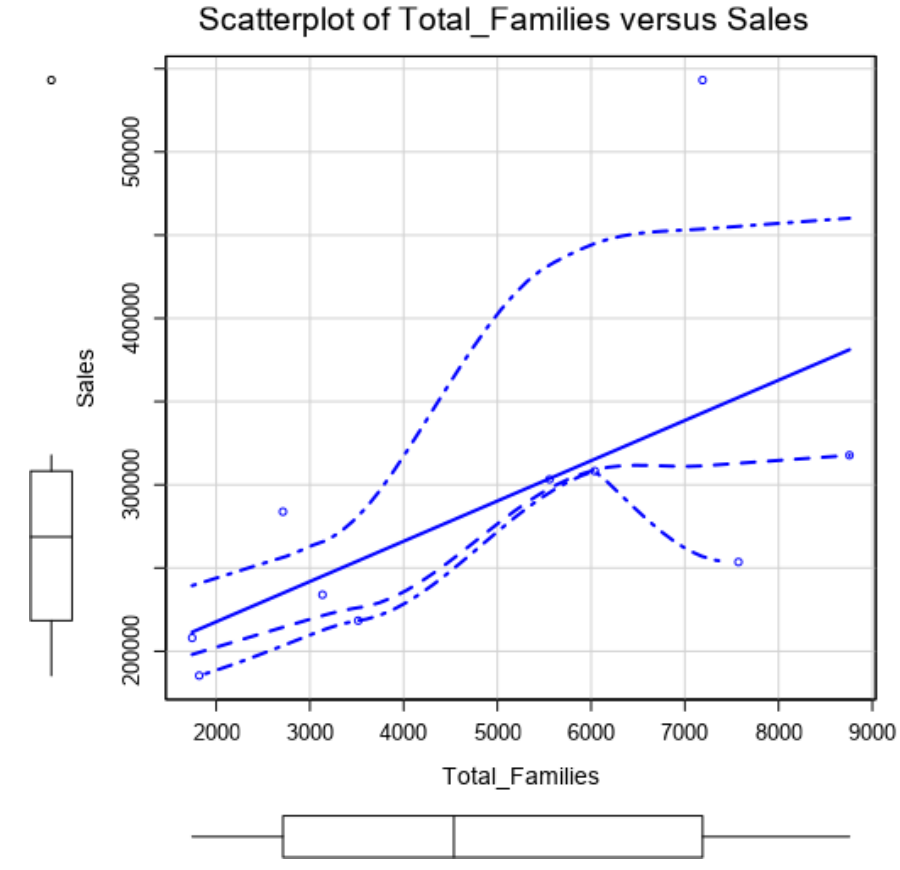
Calculated 1st quartile Q1 and 3rd quartile Q3 of the dataset using excel.
Calculated the Interquartile Range: IQR which is denoted by diff below
Upper Fence = Q3 + 1.5 * IQR
Lower Fence = Q1 – 1.5 * IQR
Comparing each value with upper fence and lower fence I have highlighted the outliers.

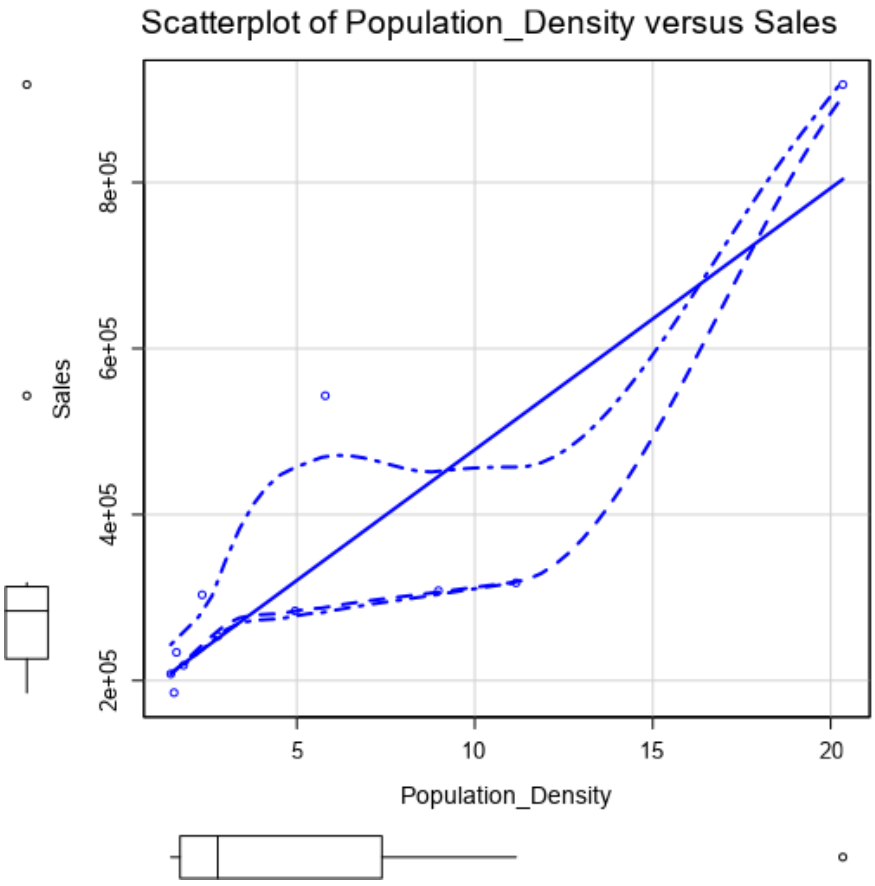
City	Sales	Census	Area	Household Under 18	Density	Families
Buffalo	185328	4585	3115.508	746	1.55	1819.5
Casper	317736	35316	3894.309	7788	11.16	8756.32
Cheyenne	917892	59466	1500.178	7158	20.34	14612.64
Cody	218376	9520	2998.957	1403	1.82	3515.62
Douglas	208008	6120	1829.465	832	1.46	1744.08
Evanston	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
INC	226152	7917	1861.721	1327	1.72	2923.41
EXC	312984	26061.5	3504.908	4037	7.39	7380.805
DIFF	86832	18144.5	1643.187	2710	5.67	4457.395
Upper Fence	443232	53278.25	5969.689	8102	15.895	14066.9
Lower Fence	95904	-19299.8	-603.06	-2738	-6.785	-3762.68

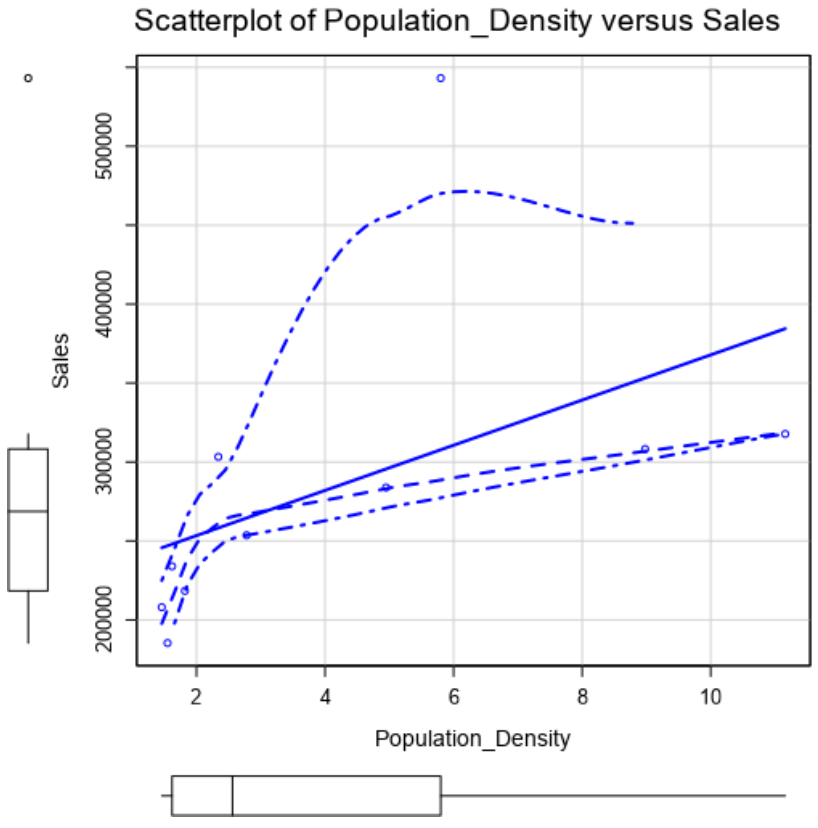

Scatterplot of X2010_Census versus Sales
Sales
2e+05
4e+05
6e+05
8e+05
10000
20000
30000
40000
50000
60000
X2010_Census


Scatterplot of X2010_Census versus Sales
Sales
200000
300000
400000
500000
5000
10000
15000
20000
25000
30000
35000
X2010_Census


Scatterplot of Total_Families versus Sales
Sales
2e+05
4e+05
6e+05
8e+05
2000
4000
6000
8000
10000
12000
14000
Total_Families

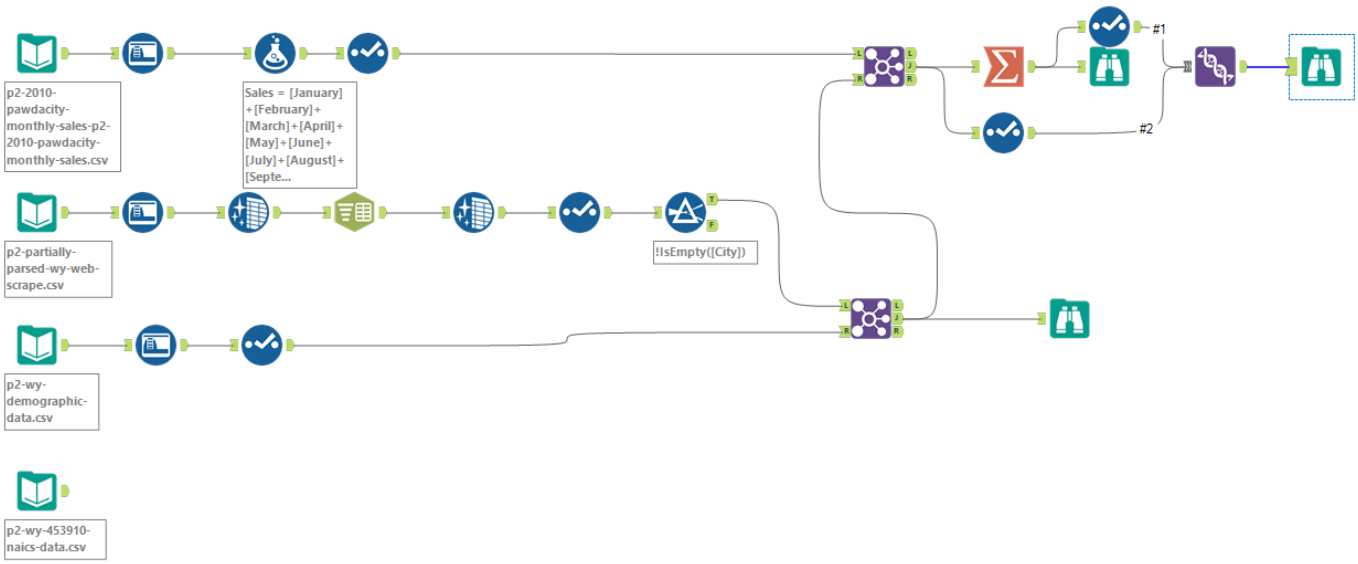

Scatterplot of Total_Families versus Sales
Sales
200000
300000
400000
500000
2000
3000
4000
5000
6000
7000
8000
9000
Total_Families


Scatterplot of Population_Density versus Sales
Sales
2e+05
4e+05
6e+05
8e+05
5
10
15
20
Population_Density


Scatterplot of Population_Density versus Sales
Sales
200000
300000
400000
500000
2
4
6
8
10
Population_Density

Seeing all the graph and just remove outlier Cheyenne city we see the difference. Therefore we should remove the row Cheyenne City.

Alteryx Workflow



Bipin Kumar Sultania