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?

2. What data is needed to inform those decisions?

Answer:

Question 1

The aim of this project is to find a new city store location for business expansion. This decision will be based on the predicted yearly sales of the company.

Question 2

To make this decision, some data are needed to do this analysis. These are:

- Monthly sales data of Pawdacity stores
- The most current sales of all competitor stores
- Population numbers
- Demographic data of Wyoming

Then from those datasets we will find information about

- City
- 2010 Census Population
- Total Pawdacity Sales
- Households with under 18
- Land Area
- Population Density
- Total Families

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*	*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

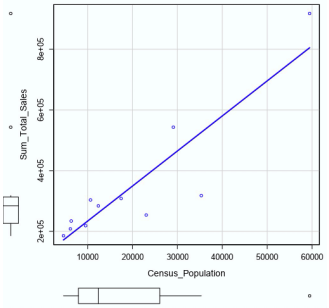
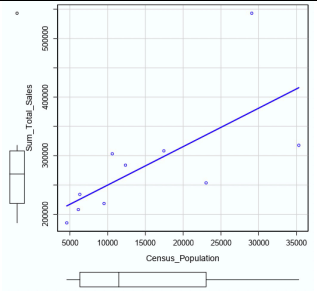
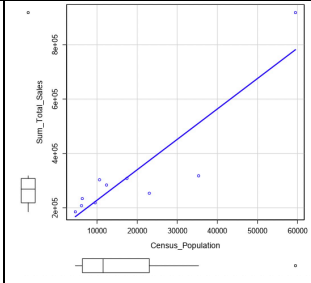
Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.

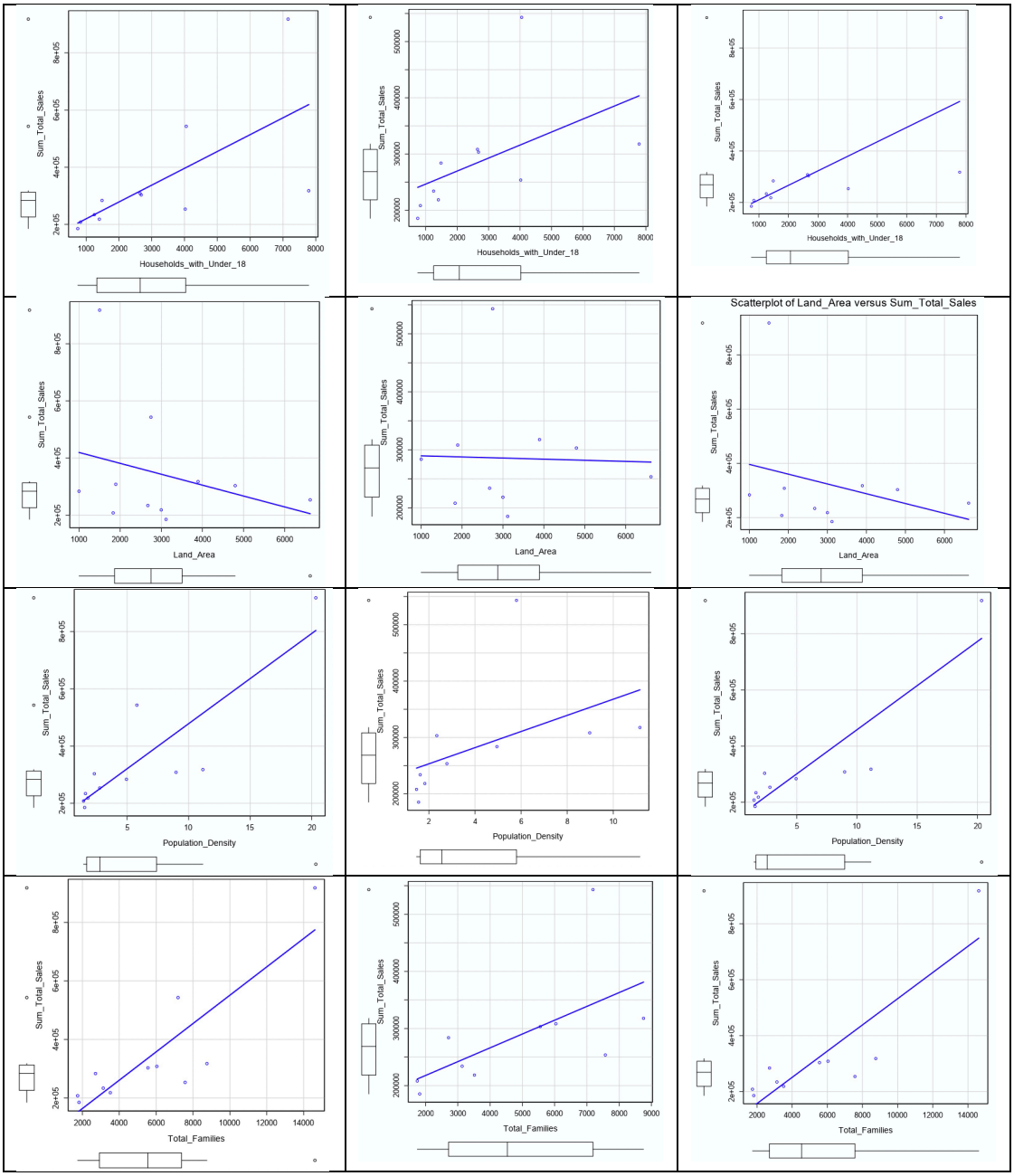
Answer:

Initially I used Excel to identify the outliers in our dataset. As a result, there are three city that identified as an outlier.

CITY	Census_Population	Is_Census_Outliers	Sum_Total_Sales	Is_Saoes_Outliers	Households with Under 18	Is_Household_Outliers	Land Area	Is_Land_Outliers	Population Density	Is_Population_Outliers	Total Families	Is_Family_Outliers
Buffalo	4,585	no	185,328	no	746	no	3,116	no	1.55	no	1819.5	no
Casper	35,316	no	317,736	no	7,788	no	3,894	no	11.16	no	8756.32	no
Cheyenne	59,466	yes	917,892	yes	7,158	no	1,500	no	20.34	yes	14612.64	yes
Cody	9,520	no	218,376	no	1,403	no	2,999	no	1.82	no	3515.62	no
Douglas	6,120	no	208,008	no	832	no	1,829	no	1.46	no	1744.08	no
Evanston	12,359	no	283,824	no	1,486	no	999	no	4.95	no	2712.64	no
Gillette	29,087	no	543,132	yes	4,052	no	2,749	no	5.80	no	7189.43	no
Powell	6,314	no	233,928	no	1,251	no	2,674	no	1.62	no	3134.18	no
Riverton	10,615	no	303,264	no	2,680	no	4,797	no	2.34	no	5556.49	no
Rock Springs	23,036	no	253,584	no	4,022	no	6,620	yes	2.78	no	7572.18	no
Sheridan	17,444	no	308,232	no	2,646	no	1,894	no	8.98	no	6039.71	no
Outliers Calculation												
Q1	7917		226152		1327		1861.721074		1.72		2923.41	
Q3	26061.5		312984		4037		3504.9083		7.39		7380.805	
IQR	18144.5		86832		2710		1643.187226		5.67		4457.395	
Upper Fence	**-19299.75**		**95904**		**-2738**		**-603.059765**		**-6.785**		**-3762.6825**	
Lower Fence	**53278.25**		**443232**		**8102**		**5969.689139**		**15.895**		**14066.8975**	

Since only one city allowed to remove or impute, I will analyze the outlier one by one. First at Rock Springs is only slightly different in Land Area. Hence, I will keep this city. Next I will use scatter plot to see the trend of each variable compared to total sales before and after the outlier removal. Since this value will be our target variables.

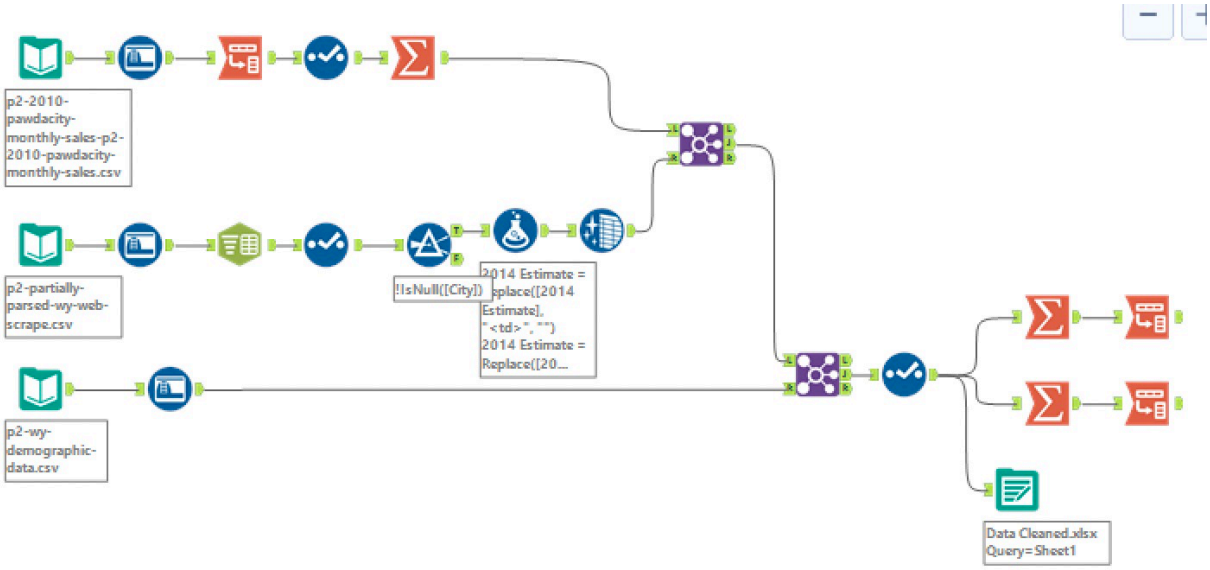
Original	Remove Cheyenne	Remove Gillete
		



By Examining the scatter plot results and data on the table, even though Cheyenne is become an outlier in for variables. The total sales values still in the same proportion with other variable like total population. It makes sense that Cheyenne store have higher value due to the higher population. So, I decided to keep Cheyenne on the dataset.

On the other hand, Gillete total sales is not inline with other variable values. This city has higher total sales even though only have small population. Hence, Gillete will be excluded from the dataset in order to build an unbiased model

Appendix



Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.