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)

1.What decisions need to be made?

Which city is suitable for the 14th store. In other words, which city will have relative big demand(high sales volume).

2. What data is needed to inform those decisions?

- The monthly sales data for all of the Pawdacity stores for the year 2010
- Most current sales of all competitor stores
- 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

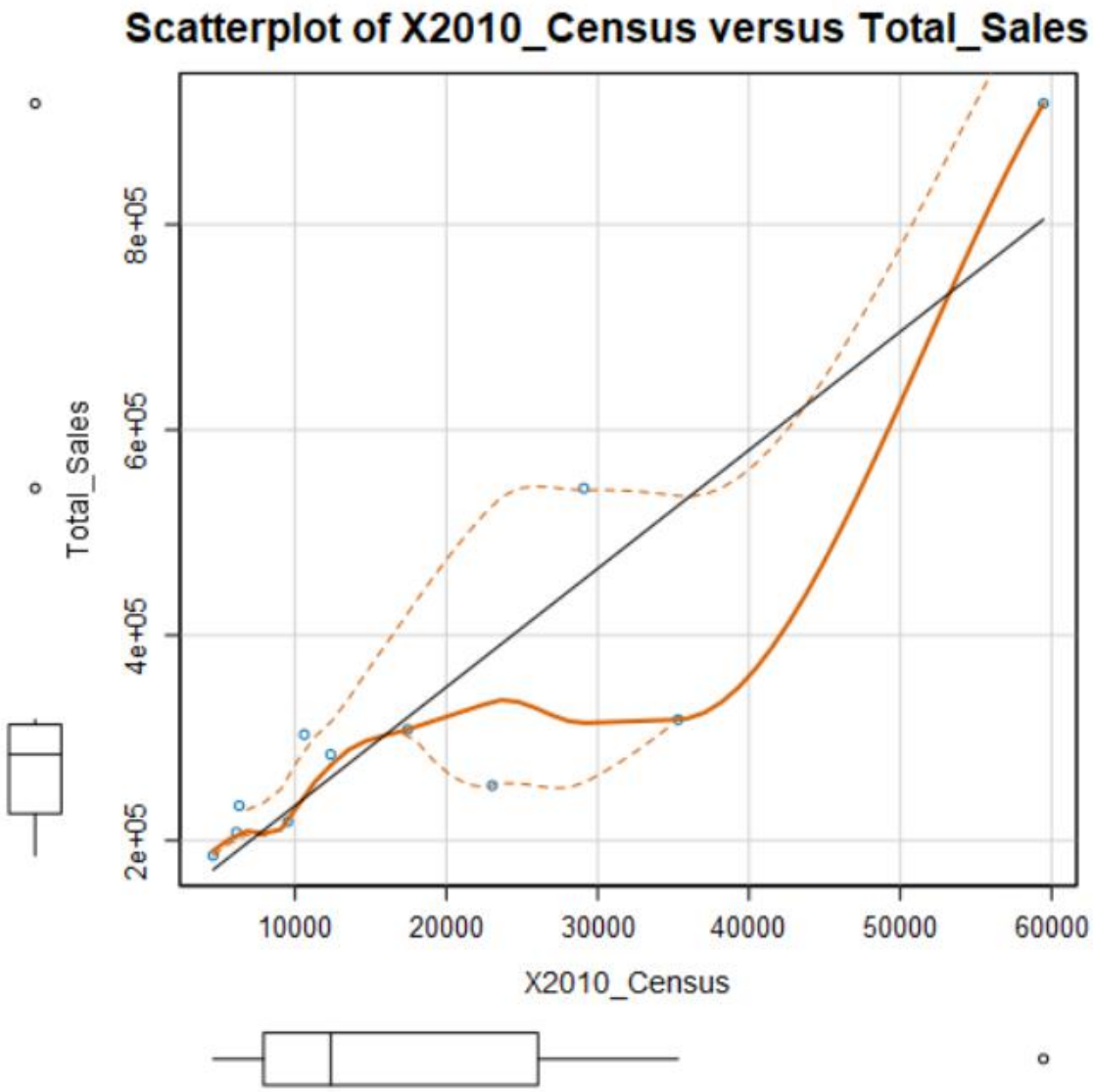
Build your training set given the data provided to you. The column sums of your dataset should match the sums in the table below.

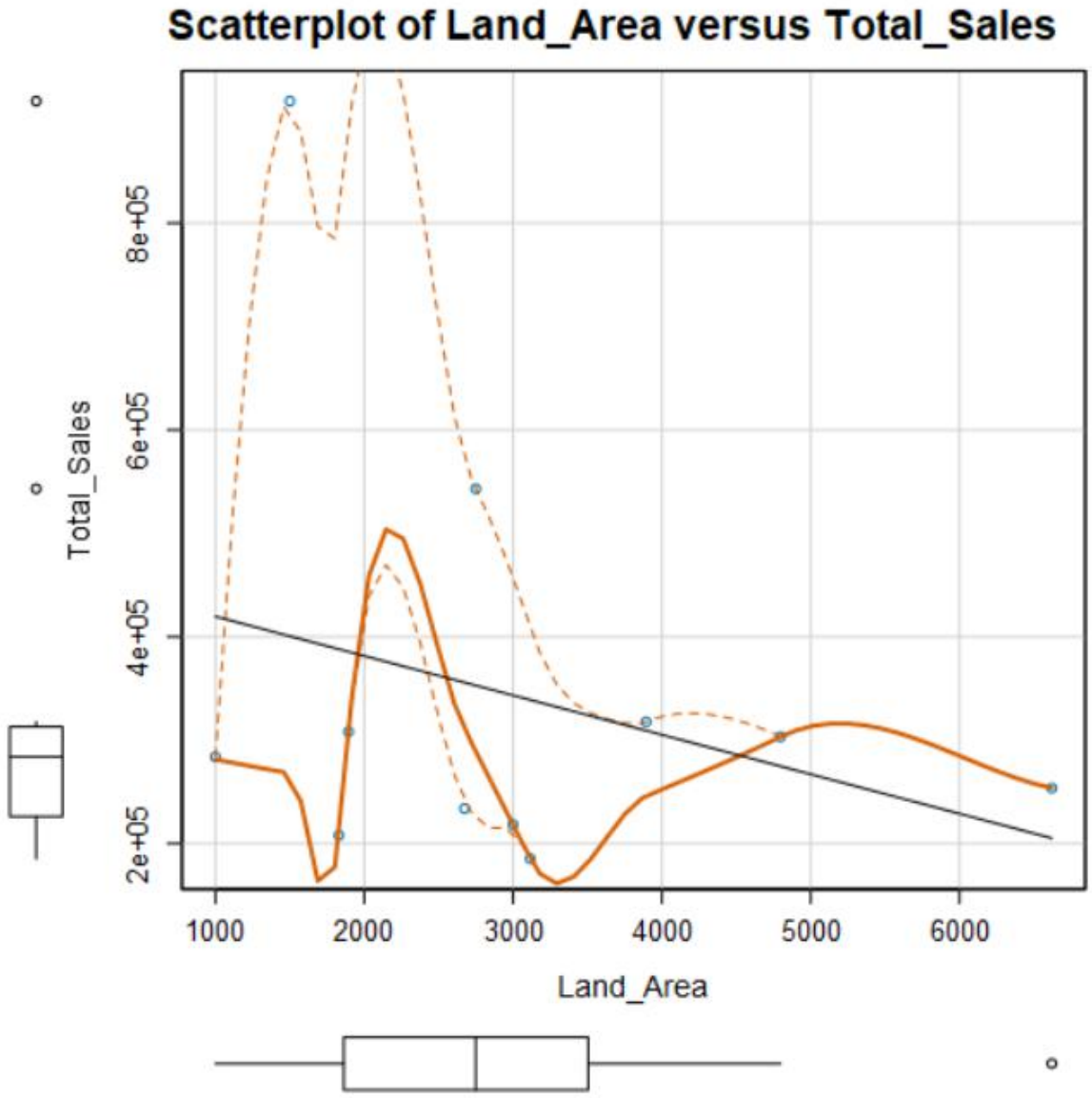
Column	Sum	Average
Census Population	213862	19442
Total Pawdacity Sales	3773304	343027.64
Households with Under 18	34064	3096.73
Land Area	33071	3006.49
Population Density	63	5.71
Total Families	62653	5695.71

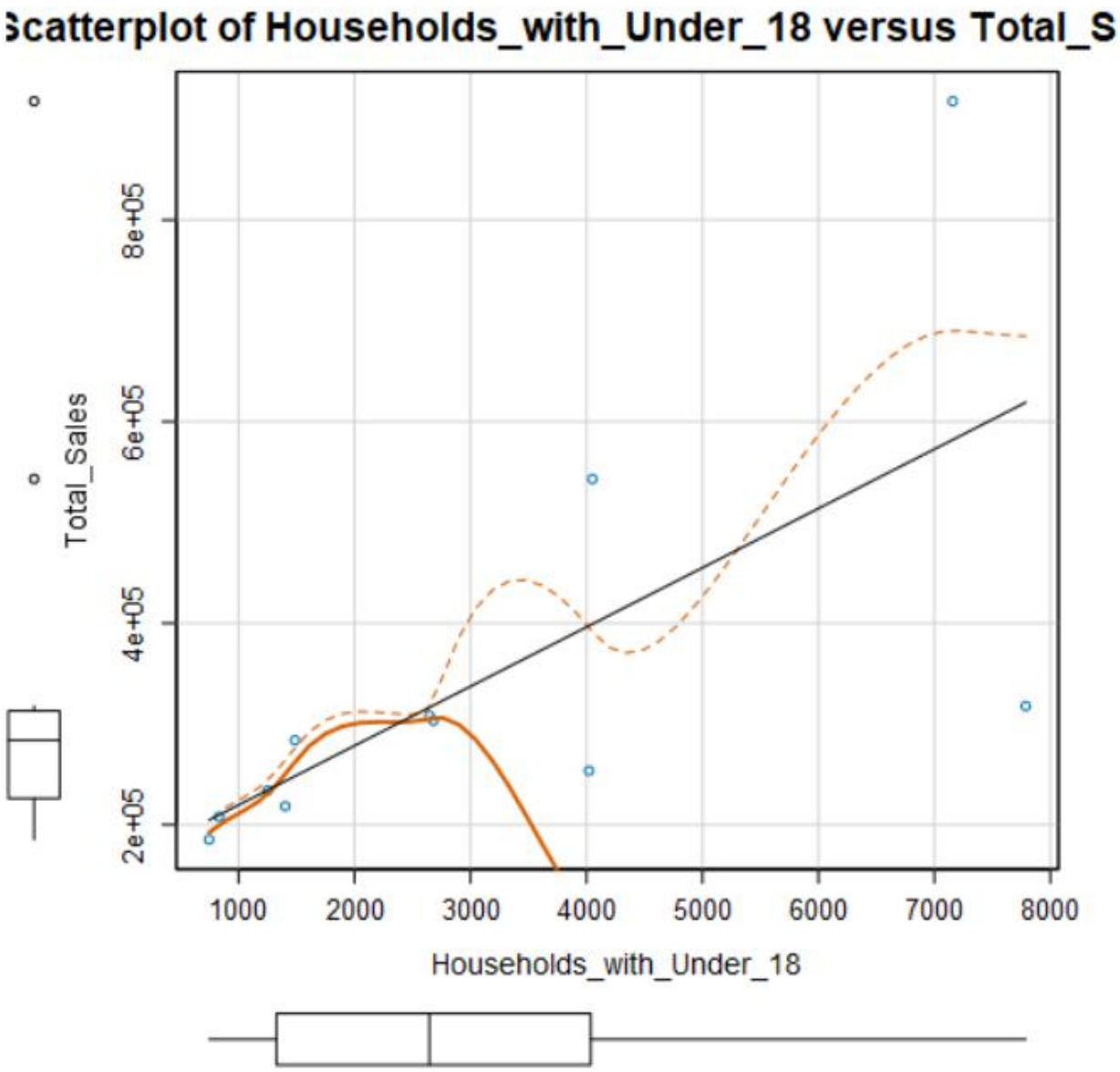
Step 3: Dealing with Outliers

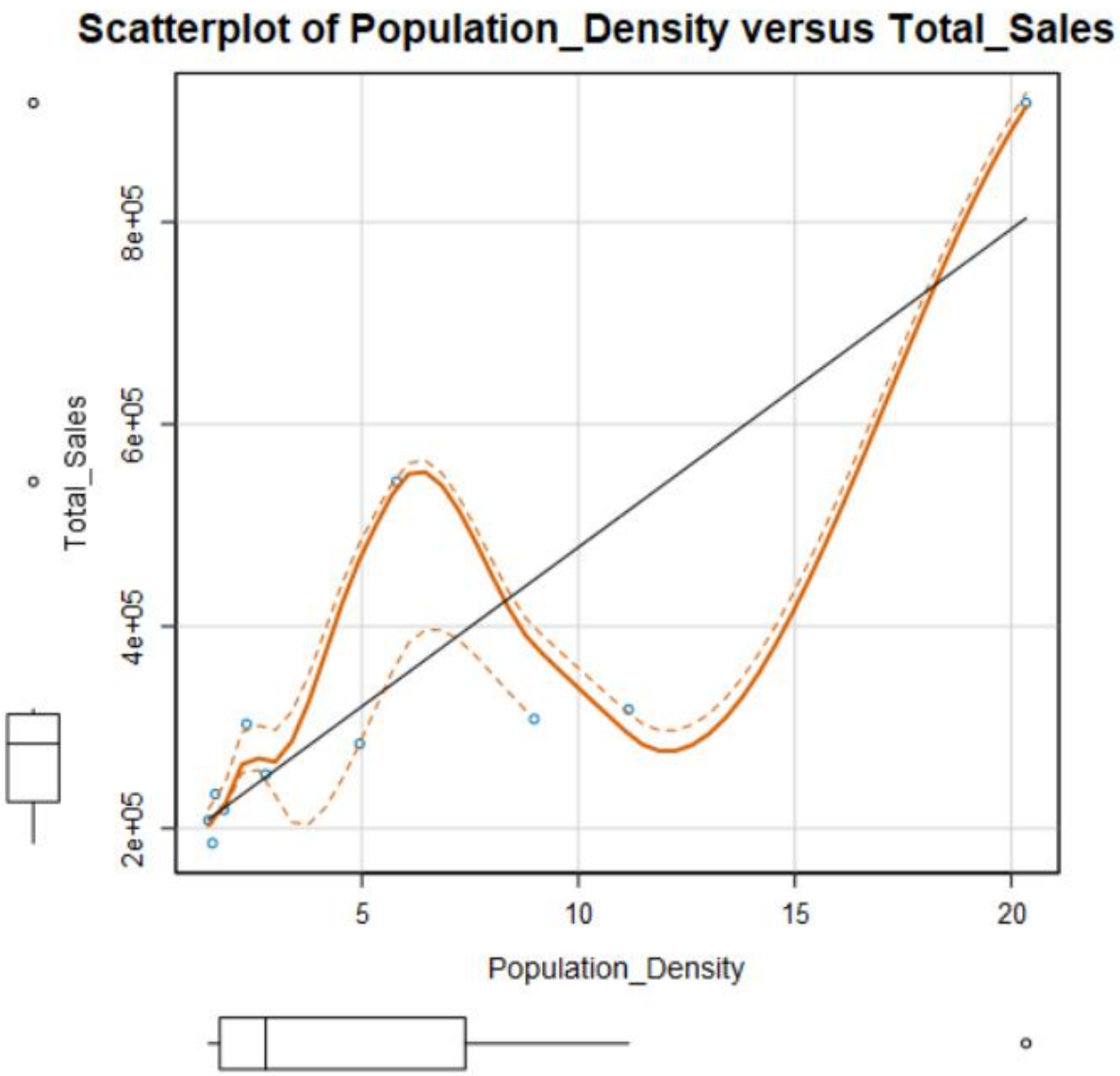
Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small dataset (11 cities), you should only remove or impute one outlier. Please explain your reasoning.

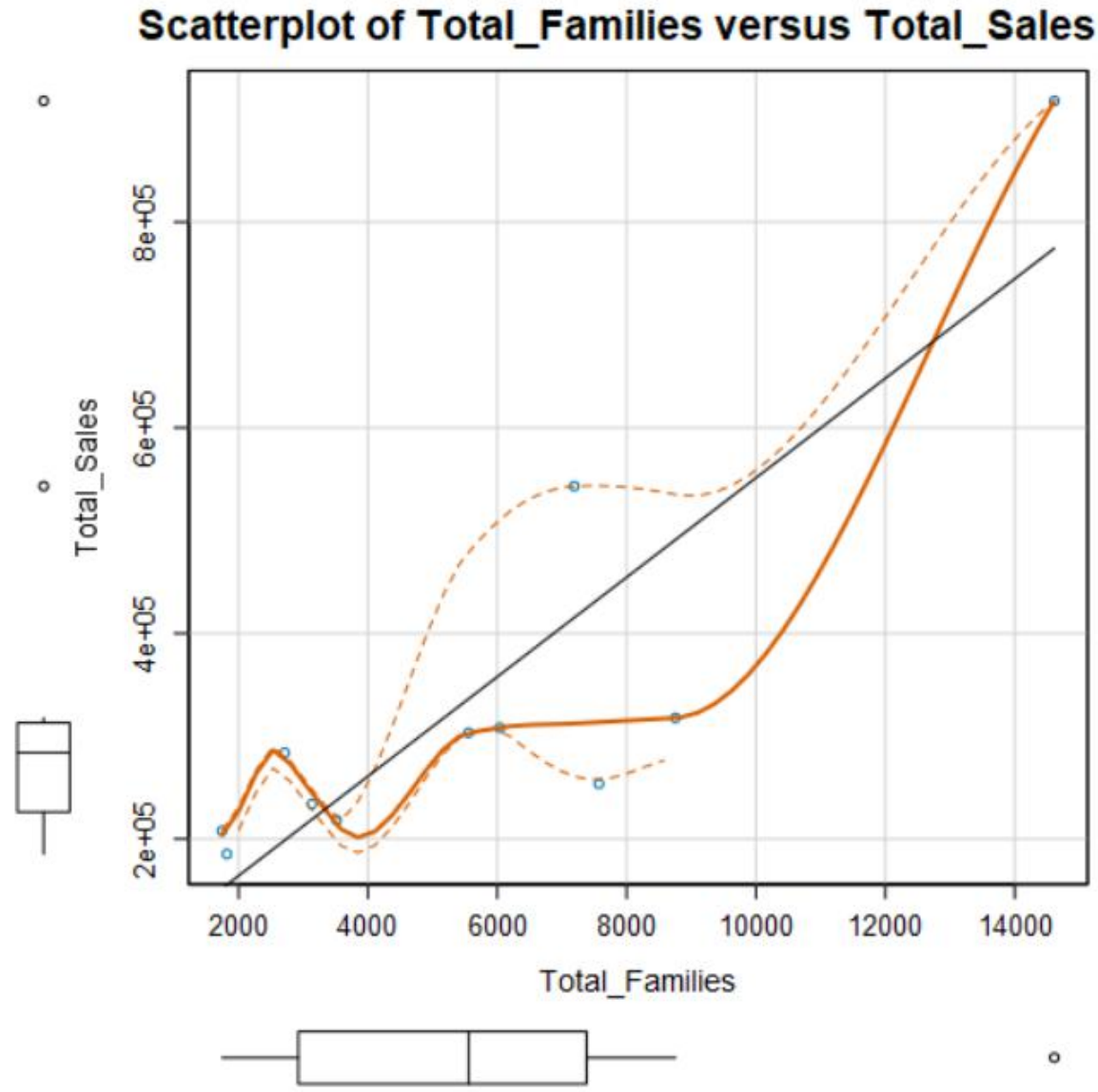
Below is the scatter plots of Total_Sales versus different variables.




Scatterplot of Land_Area versus Total_Sales
Total_Sales
2e+05
4e+05
6e+05
8e+05
1000
2000
3000
4000
5000
6000
Land_Area


Scatterplot of Households_with_Under_18 versus Total_S
Total_Sales
2e+05
4e+05
6e+05
8e+05
1000
2000
3000
4000
5000
6000
7000
8000
Households_with_Under_18


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



Below is a summary of the dataset, with a further analysis of the interquartile ranges for the variables and their subsequent upper fence which for this project will be [1.5 * Interquartile Range] + 3rd Quartile and lower fence will be 1st Quartile - [1.5 * Interquartile Range].

I will look into values that are above the "Upper Fence" for each variable.

Name	Min	Max	Median	Mean	Std. Dev.
Census_Population	4585.00	59466.00	12359.00	19442.00	16616.02
Household_with_Under_18	746.00	7788.00	2646.00	3096.73	2453.00
Land_Area	999.50	6620.20	2748.85	3006.49	1617.46
Total_Sales	185328.00	917892.00	283824.00	343027.64	213538.71

Population_Density	1.46	20.34	2.78	5.71	5.85
Total_Families	1744.08	14612.64	5556.49	5695.71	3816.05

Census_Population_IQR	Total_Sales_IQR	Household_with_Under_18_IQR	Land_Area_IQR	Population_Density_IQR	Total_Families_IQR
18144.50	86832.00	2710.00	1643.19	5.67	4457.40
Census_Population_Upper_Fence	**Padacity_Sales_Upper_Fence**	**Household_with_Under_18_Upper_Fence**	**Land_Area_Upper_Fence**	**Population_Density_Upper_Fence**	**Total_Families_Upper_Fence**
53278.25	443232.00	8102.00	5969.69	15.90	14066.90

The list below indicates max points above that of their respective “Upper Fence”:

Census Population for Cheyenne
Land Area for Rock Springs
Population Density for Cheyenne
Total Families for Cheyenne
Total Sales for Gillette and Cheyenne

As the analysis above shown, Cheyenne, Gillette and Rock Springs are outliner. I choose to remove Cheyenne and leave Rock Springs and Gillette based on several reasons.

Firstly, Cheyenne outlines in many cases like Population_Density, Total_Families and Total_Sales.

Secondly, the scatterplot for Land Area vs Sales indicates that Rock Springs follows the downward direction of the line of best fit for that plot with sales roughly inline with other sales values in that plot.

Thirdly, Gillette outlines regard to Total_Sales because there are two stores. However looking through the other categories Gillette’s data looks relatively with in our outlier range except for its sales. There doesn’t seem to be a good reason to remove it.