Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Help leading pet store, Pawdacity to open its 14th store in state of Wyoming. To make decision we need to get data from city, 2010 census population, sales in other stores, avg sales, household with under 18. Count of total families and population density.

Step 2: Building the Training Set

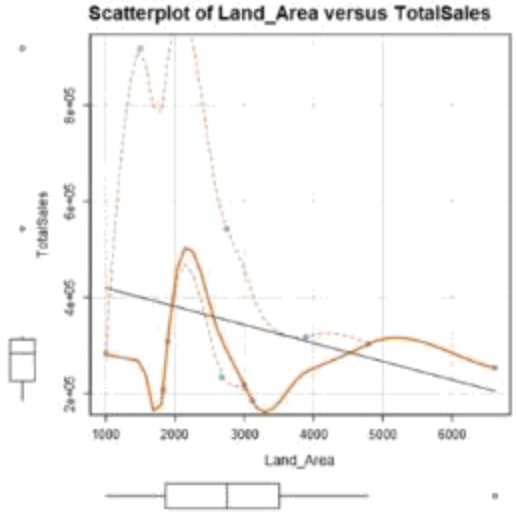
Build your training set given the data provided to you. Your column sums of your dataset should match the sums in the table below.

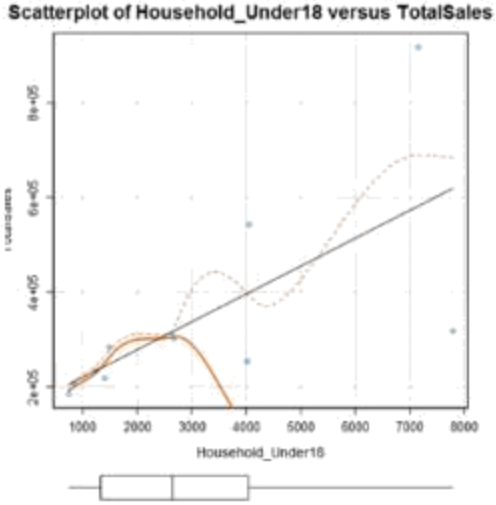
In addition provide the averages on your data set here to help reviewers check your work. You should round up to two decimal places, ex: 1.24

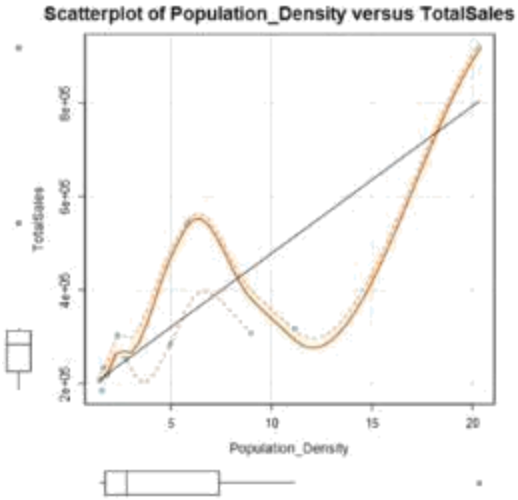
Column	Sum	Average
Census Population	*213,862*	*19442*
Total Pawdacity Sales	*3,773,304*	*343028*
Households with Under 18	*34,064*	*3096.8*
Land Area	*33,071*	*3006.4*
Population Density	*63*	*5.8*
Total Families	*62,653*	*5695.8*

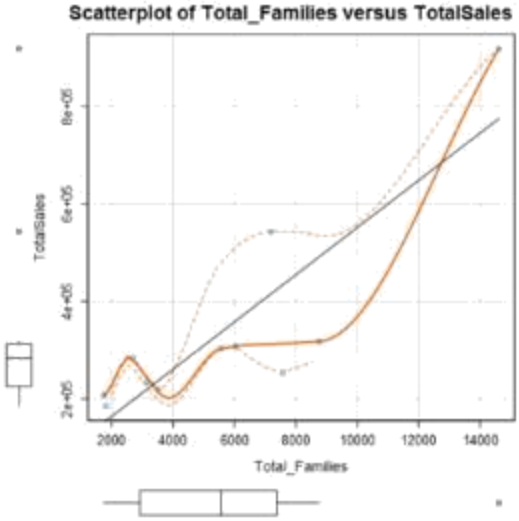
Step 3: Dealing with Outliers

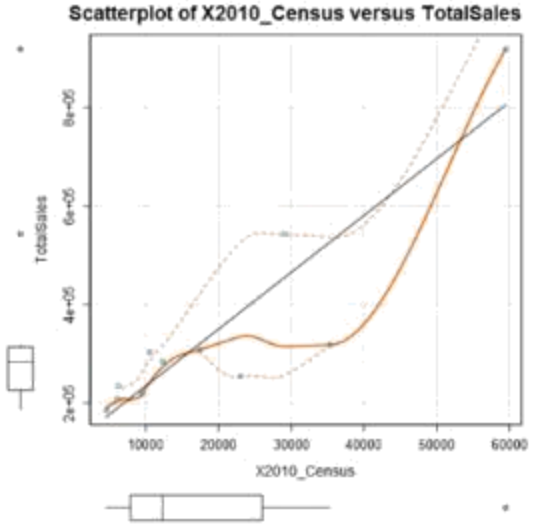
Answer these questions


Scatterplot of Land_Area versus TotalSales
TotalSales
Land_Area


Scatterplot of Household_Under18 versus TotalSales
Household_Under18


Scatterplot of Population_Density versus TotalSales
TotalSales
Population_Density


Scatterplot of Total_Families versus TotalSales
TotalSales
Total_Families


Scatterplot of X2010_Census versus TotalSales
TotalSales
X2010_Census

Gillette city and Cheyenne city has the outliners as we can see that sales data are too high in the scatterplots. Population of Gillette city and total sales are still relevant. Gillette will be a outliner when you compared it with all cities from document because of long distance from the linear trend. From Scatterplot we can see the data of Cheyenne that falls within the expected range when extrapolated. So now, Relationship between population related variables of Gillette and total sales are correlated, that’s why it should be in analysis.