Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Key Decisions:

1. **What decisions needs to be made?**
   A pet store named pawdacity from Wyoming need to find out where to open the 14th store in the state.
2. **What data is needed to inform those decisions?**
   Some data like the city population, population density, sales in other stores, competitor sales, total families, household size are needed.

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.

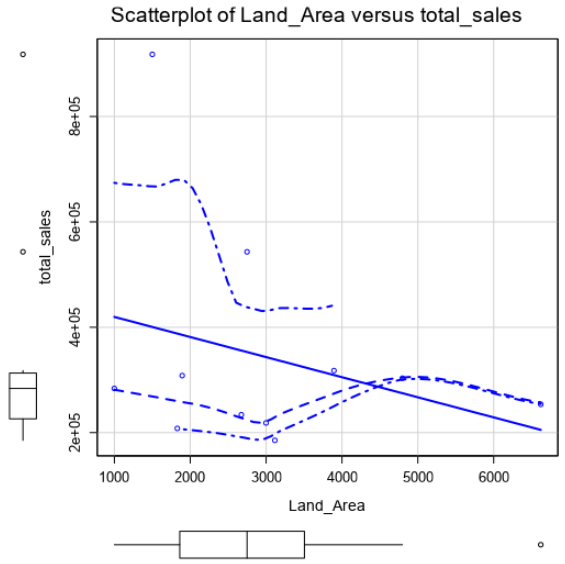
In addition provide the averages on your data set here to help reviewers check your work. You should round up to two decimal places, ex: 1.24

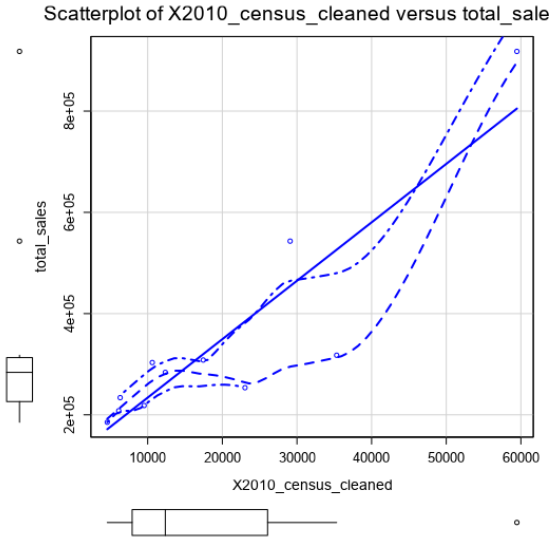
Column	Sum	Average
Census Population	*213,862*	*19442*
Total Pawdacity Sales	*3,773,304*	343027.64
Households with Under 18	*34,064*	*3096.73*
Land Area	*33,071*	*3006.49*
Population Density	*63*	*5.71*
Total Families	*62,653*	*5695.71*

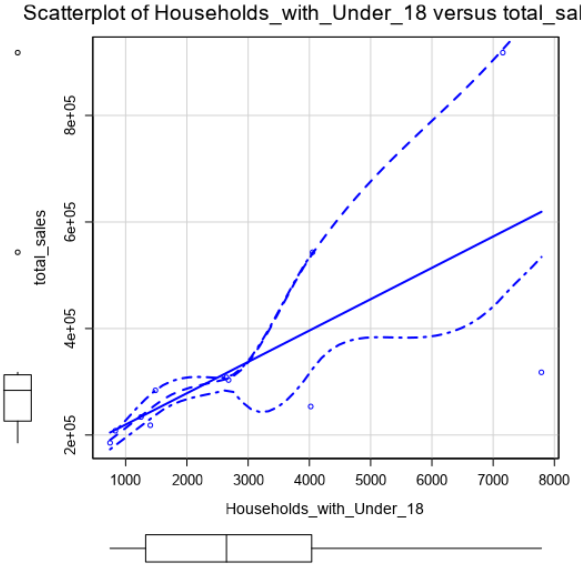
Step 3: Dealing with Outliers

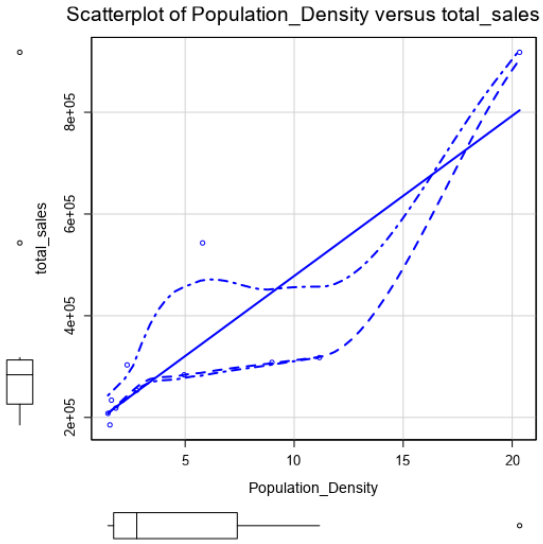
Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.

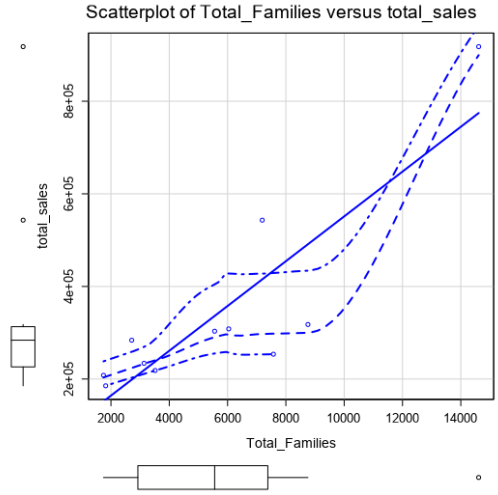
See the 5 scatter plots along with box plots of variables below. In boxplot of total sales, we saw two data points are outliers. Then from Land_area vs total sales plot, we can see that two cities with area 1500 and 2600, these two cities sales are more deviate above expected range, i.e. much higher sales for relatively small land area. These two cities are Gillette and Cheyenne. When plot total sales versus other remaining variables, we see that only one city is beyond expected range for variable 2010_census data, population density and total families. This city is Gillette. The city of Gillette is only correlated for households_with_under_18, therefore it makes sense for us to remove city Gillette as an outlier.



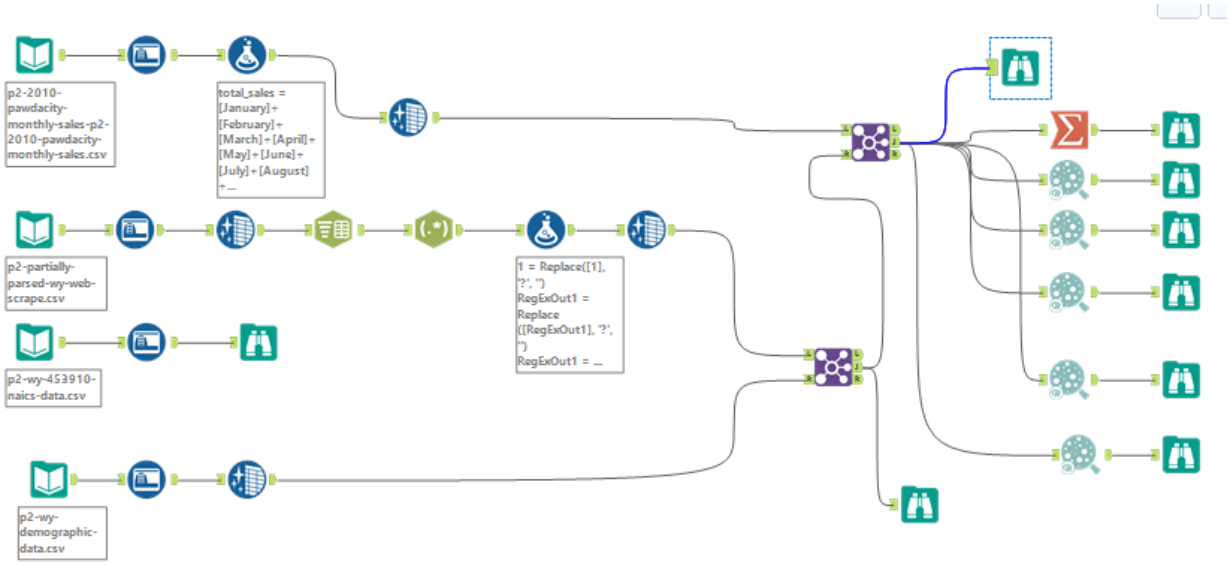








I also attach workflow:


p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales.csv
total_sales = [January]+[February]+[March]+[April]+[May]+[June]+[July]+[August]+...
p2-partially-parsed-wy-web-scrape.csv
1 = Replace([1], '?', '')
RegExOut1 = Replace([RegExOut1], '?', '')
RegExOut1 = ...
p2-wy-453910-naics-data.csv
p2-wy-demographic-data.csv