Project 2.1: Data Cleanup

Alhanoof Alyabes
3/25/2018

Business and Data Understanding

We need to Predict yearly sales for all cities, to make decision and recommendations about the prober place where to open new pet store in wyoming We need different data to inform those decisions, which are: city, 2010 census population, pawdacity sales, household with under 18, land areas, population density and total families data

Building the Training Set

By Using different tools like: select, filter, formula join and summarized on the given data sets, the average of the variables were calculated on table (1), and illustrated in workflow at the bottom of the page figure (2).

Column	Sum	Average
Census Population	213,862	**19442**
Total Pawdacity Sales	3,773,304	**343027.63**
Households with Under 18	34,064	**3096.72**
Land Area	33,071	**3006.48**
Population Density	63	**5.7**
Total Families	62,653	**5695.7**

Table (1) sum and average for each variable

Dealing with Outliers

By using scatterplot for each variable as shown figure (1), and calculated IQR by excel, two cities identified as outliers, Cheyenne and Gillette.

Cheyenne has upper fences in three variables: total sales=917892, population density =20.34, total families=14612.64. While Gillette has one upper fence with sales =543132.

Since we have small dataset (11) and Cheyenne is big city which have outliers value in many fields, they are in line with what is expected, we can keep it for further analysis.

However, Gillette total sales dataset are not related with other variables, so we can remove it to build unbiased linear regression model.

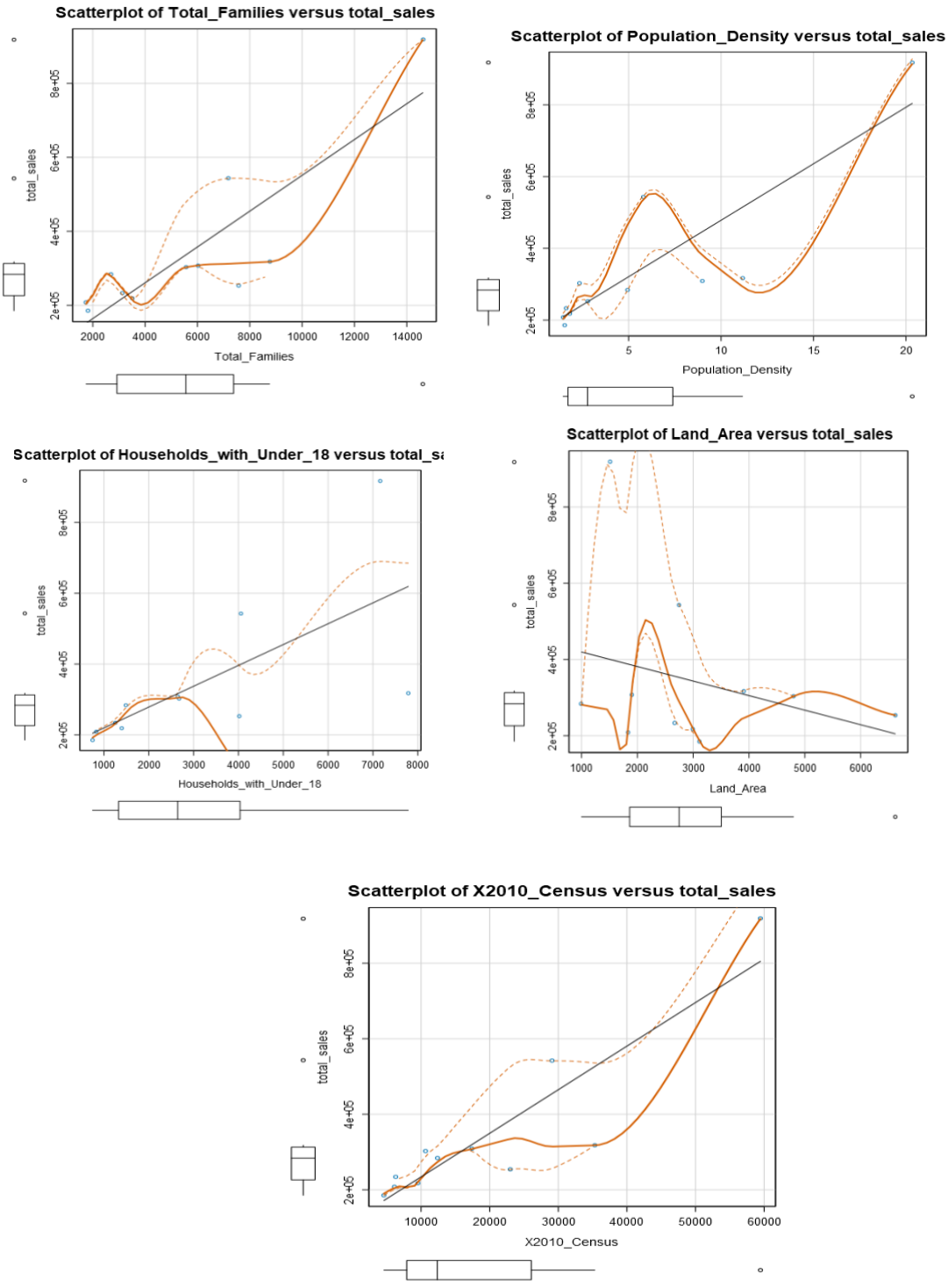

figure (1) scatterplots for each variable

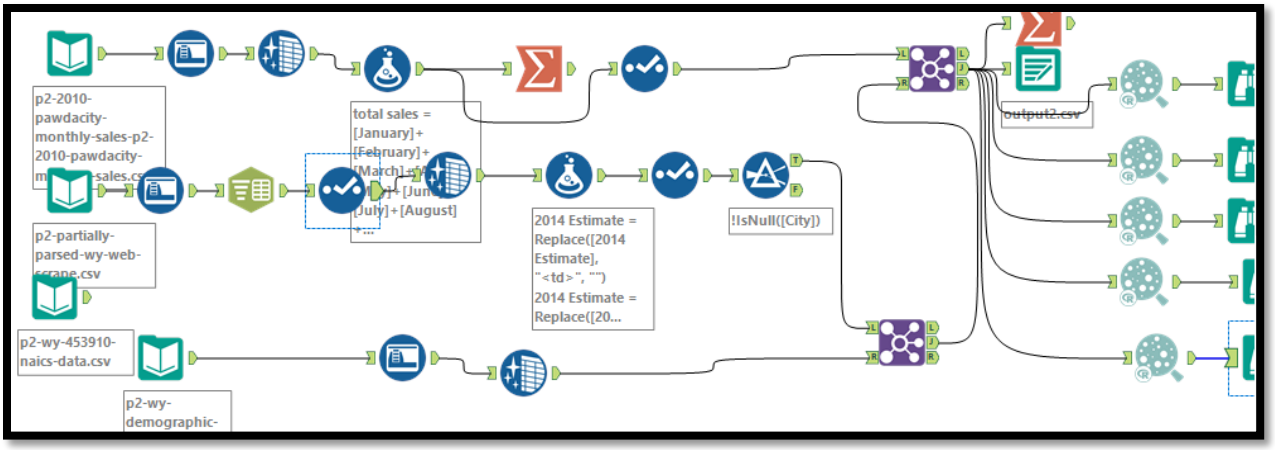

figure (2) alteryx workflow