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?
   Determine the Wyoming city where the new store for Pawdacity will be open based on the higher yearly sales.

2. What data is needed to inform those decisions?
   Yearly sales by city, population, population density, land area, total families can be used to predict yearly sales.

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*	*19,442.00*
Total Pawdacity Sales	*3,773,304*	*34,3027.64*
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

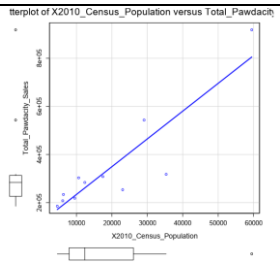
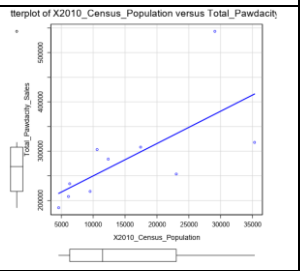
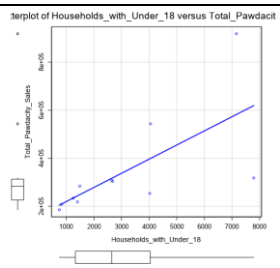
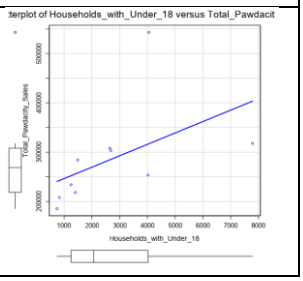
Answer these questions

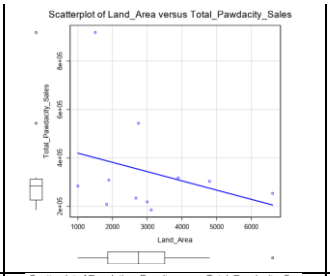
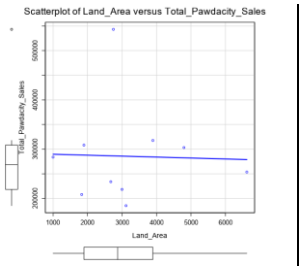
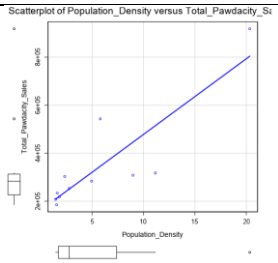
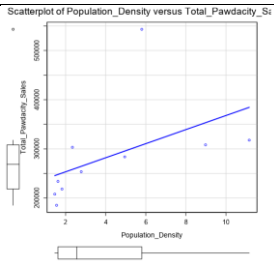
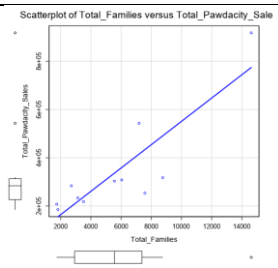
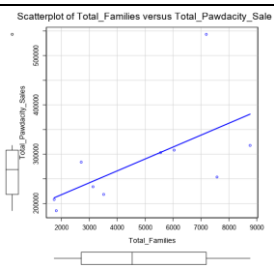
Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.

The record for Cheyenne city was removed because for the target variable and several predictor variables has values that were the higher differences compared against the upper fence from the quartile analysis.

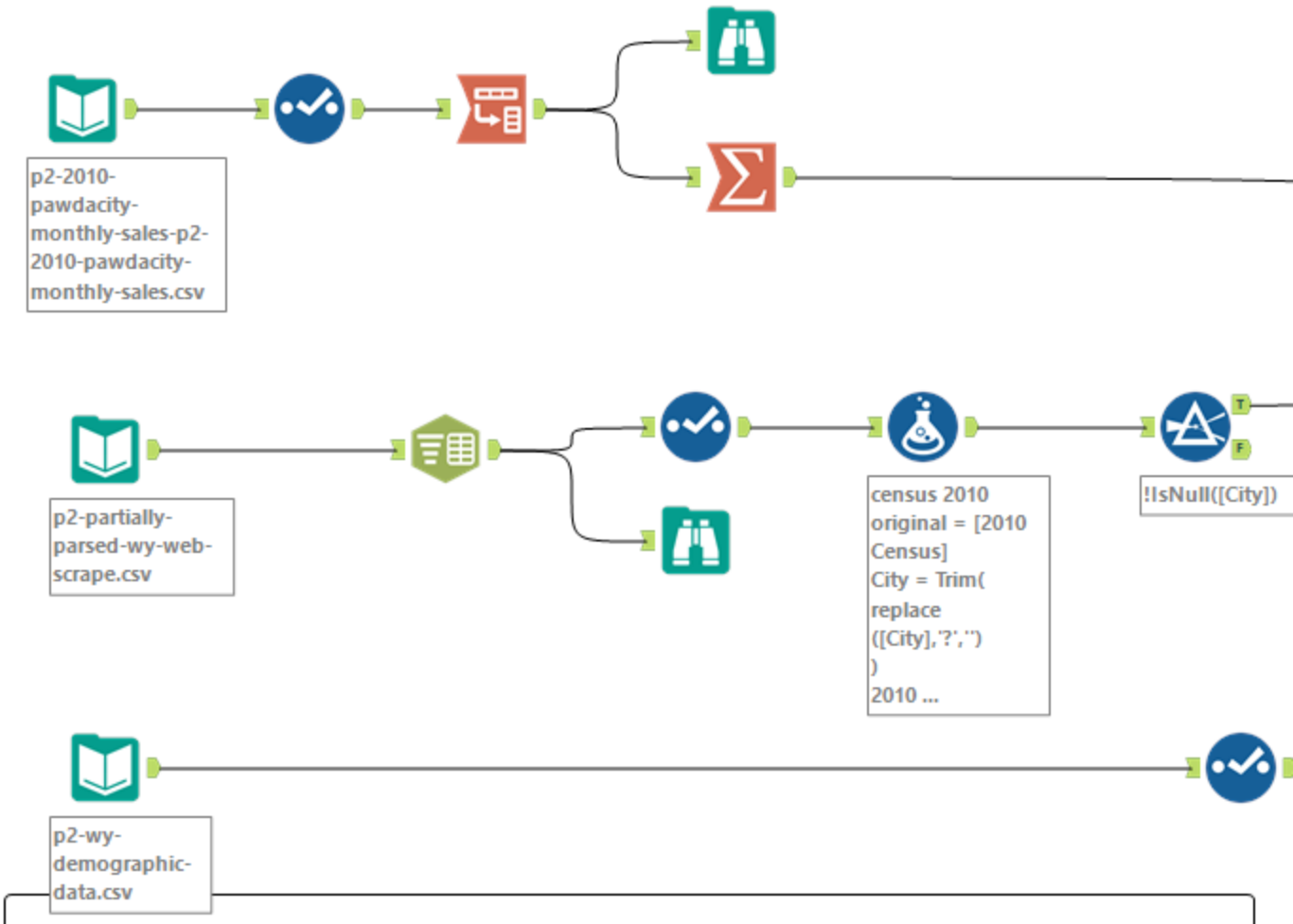
	City	Total Pawdacity Sales	2010 Census Population	Households with Under 18	Land Ar
2	Buffalo	185,328.0	4,585.0	746.0	3,11
3	Casper	317,736.0	35,316.0	7,788.0	3,89
4	Cheyenne	917,892.0	59,466.0	7,158.0	1,50
5	Cody	218,376.0	9,520.0	1,403.0	2,99
6	Douglas	208,008.0	6,120.0	832.0	1,82
7	Evanston	283,824.0	12,359.0	1,486.0	99
8	Gillette	543,132.0	29,087.0	4,052.0	2,74
9	Powell	233,928.0	6,314.0	1,251.0	2,67
10	Riverton	303,264.0	10,615.0	2,680.0	4,79
11	Rock Springs	253,584.0	23,036.0	4,022.0	6,62
12	Sheridan	308,232.0	17,444.0	2,646.0	1,89
13					
14	QUARTILE	VALUE	VALUE	VALUE	VALUE
15	1	226,152.0	7,917.0	1,327.0	1,86
16	3	312,984.0	26,061.5	4,037.0	3,50
17					
18	IQR	86,832.0	18,144.5	2,710.0	1,64
19	UPPER FENCE	443,232.0	53,278.3	8,102.0	5,96
20	LOWER FENCE	95,904.0	(19,299.8)	(2,738.0)	(60

Also, it was analyzed the impact of removing the outlier record with scatter plot for every predictor variable and target variable, and the slope was not affected considerably.

	With outlier	Without outlier
Census population		
Households with Under 18		

Land Area		
Population Density		
Total Families		

Alteryx solution



Data
p2-2010-pawdacity-monthly-sales.csv - This file contains all of the monthly sales for all Pawdacity stores for 2010.

p2-partially-parsed-wy-web-scrape.csv - This is a partially parsed data file that can be used for population numbers.

p2-wy-453910-naics-data.csv - NAICS data on the sales of all competitor stores where total sales is equal to 12 months of sales

p2-wy-demographic-data.csv - This file contains demographic data for each city and county in Wyoming.

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.