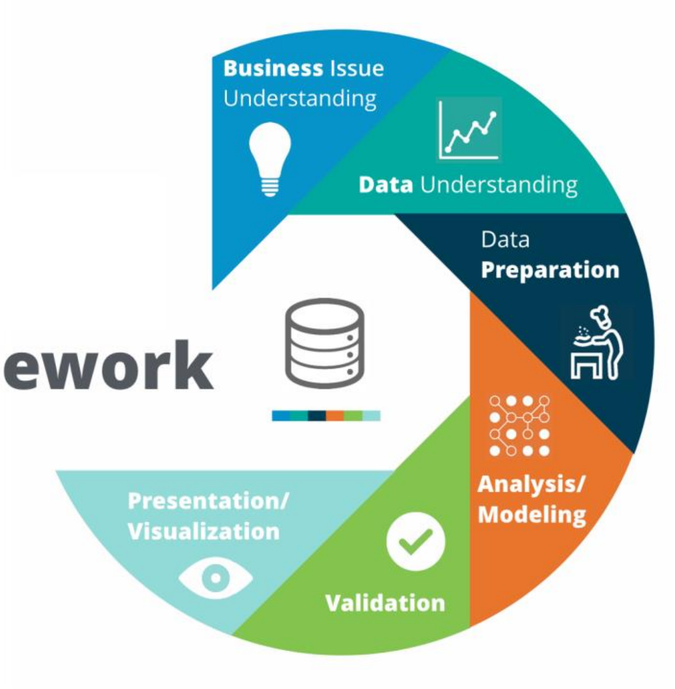
Project 2.1: Data Cleanup

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)

Pawdacity is looking to expand its business by opening its 14th store in Wyoming. Based on available data, the business specifically seeks to decide in which city in Wyoming would be best to open the new store based on predicted yearly sales. Available datasets include demographics of the 13 cities that Pawdacity already has business in, the monthly revenues of Pawdacity's stores and revenue data of competitors. Hence, this qualifies as a data rich problem and next step is to apply a suitable analysis methodology. However, this is not possible as the data is messy, meaning that we need to prepare the data and create an analytical dataset as per the CRISP-DM framework.



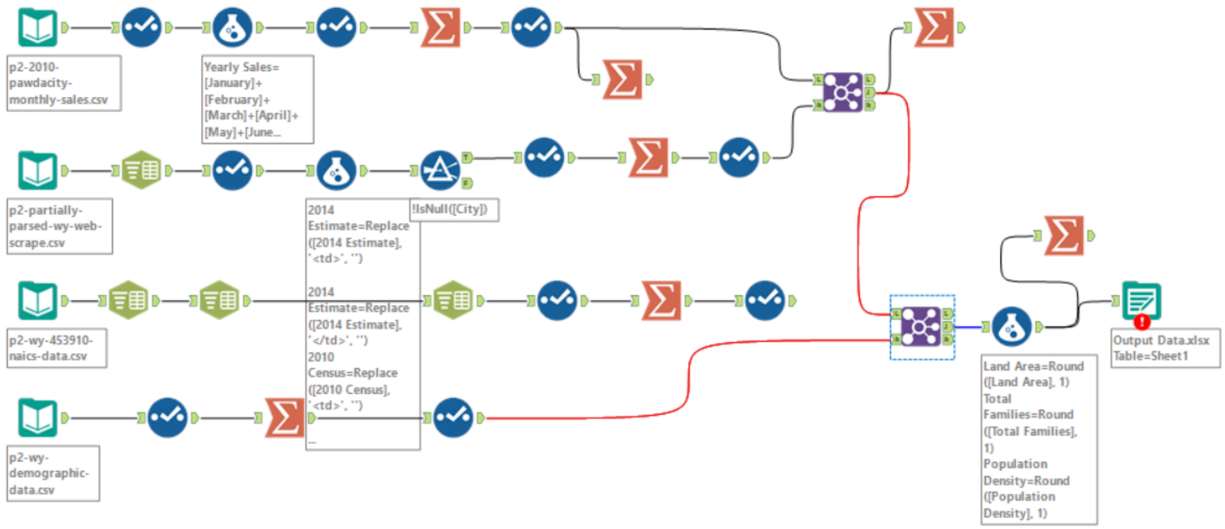
Step 2: Building the Training Set

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

The data was processed and an analytical dataset was prepared containing 11 records (cities) and data which are summarized below.

Column	Sum	Average
Census Population	213,862	19,442
Total Pawdacity Sales	$3,773,304	$34,3027.64
Households with Under 18	34,064	3,096.73
Land Area	33,071	3,006.45
Population Density	63	5.73
Total Families	62,653	5,695.73

Alteryx workflow:



p2-2010-pawdacity-monthly-sales.csv
p2-partially-parsed-wy-web-scrape.csv
p2-wy-453910-naics-data.csv
p2-wy-demographic-data.csv
T
F
L
J
R
Output Data.xlsx Table=Sheet1

Yearly Sales=[January]+[February]+[March]+[April]+[May]+[June...

!IsNull([City])

2014 Estimate=Replace([2014 Estimate], '<td>', '')
2014 Estimate=Replace([2014 Estimate], '</td>', '')
2010 Census=Replace([2010 Census], '<td>', '')
...

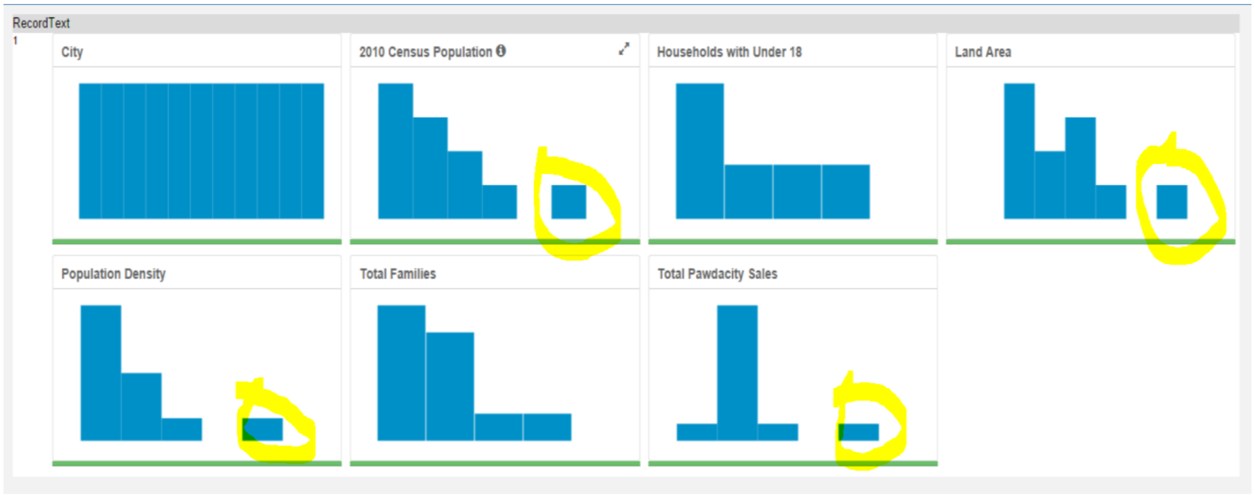
Land Area=Round([Land Area], 1)
Total Families=Round([Total Families], 1)
Population Density=Round([Population Density], 1)

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 first step to identifying possible outliers is a brief graphical summary of the distribution of data. Using the Field Summary tool:



The highlighted bars correspond to ‘extraordinary’ values for Cheyenne in Population, Population Density and Total Pawdacity Sales. Also for Rock Springs in Land Area. It is deemed that, Cheyenne being a bigger city, it is normal for these magnitudes of variables to stand out and are not really an outlier. If removed, it can be argued that the model may be limited to small and medium cities missing an opportunity to expand to a major city in Wyoming. Hence, Cheyenne will not be removed from our dataset.

A sounder method is to calculate the interquartiles Q1 and Q3 and subsequently the upper and lower fences. Having done this, we notice that Gillette’s total sales are also an outlier being twice as high as for cities with the same population. Measuring the impact of deleting the Gillette records:

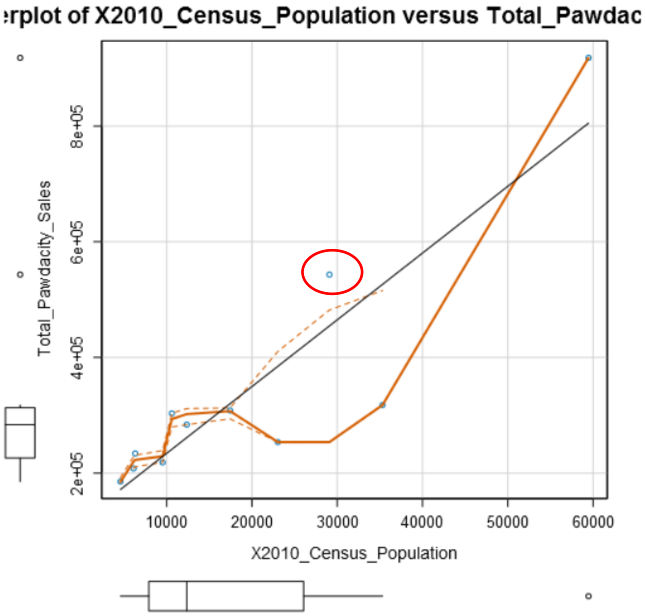


Figure 1 - Scatterplot including Gillette

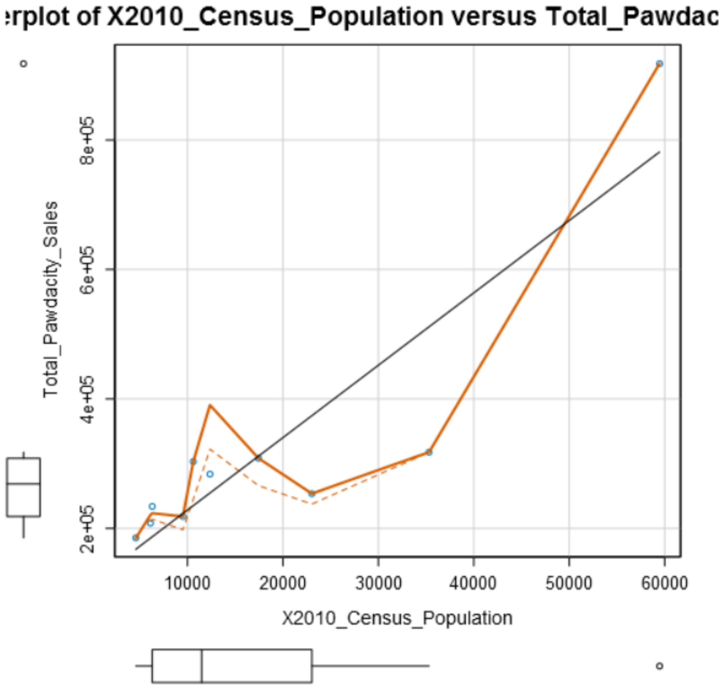


Figure 2 - Gillette sales not included in the plot

Removing Gillette as an outlier slightly decreases the slope of the fitted line but does not change the plot significantly. Also best practices recommend that regression models should have at least 15 points, so it was decided that Gillette data will not be removed but kept as is for our modelling and analysis purposes.

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.