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?

ANSWER: A decision needs to be made in which city to establish the 14th Pawdacity Store based on the yearly sales data available.

2. What data is needed to inform those decisions?

ANSWER: We need the sales data for existing stores, demographic data of existing and new cities like population, Number of families, Population density etc. We can also consider competitor sales as a metric to shortlist / finalize a new location

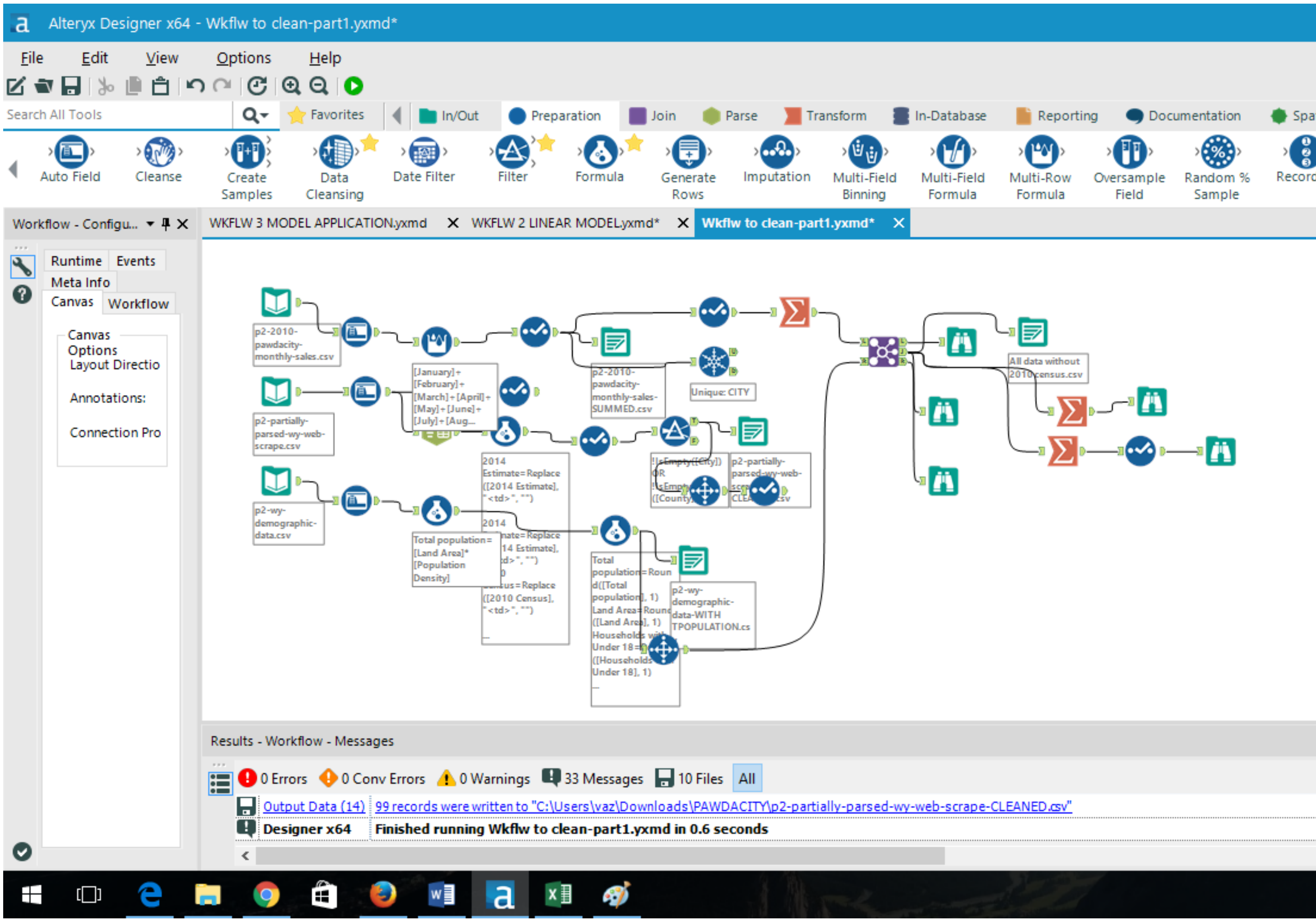
Step 2: Building the Training Set

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

Here is an image of the workflow , I built to achieve the above answers


Alteryx Designer x64 - Wkflw to clean-part1.yxmd*
File
Edit
View
Options
Help
Search All Tools
Favorites
In/Out
Preparation
Join
Parse
Transform
In-Database
Reporting
Documentation
Spa
Auto Field
Cleanse
Create Samples
Data Cleansing
Date Filter
Filter
Formula
Generate Rows
Imputation
Multi-Field Binning
Multi-Field Formula
Multi-Row Formula
Oversample Field
Random % Sample
Record
Workflow - Configu...
Runtime
Events
Meta Info
Canvas
Workflow
Canvas Options
Layout Directio
Annotations:
Connection Pro
WKFLW 3 MODEL APPLICATION.yxmd
WKFLW 2 LINEAR MODEL.yxmd*
Wkflw to clean-part1.yxmd*
p2-2010-pawdacity-monthly-sales.csv
p2-partially-parsed-wy-web-scrape.csv
p2-wy-demographic-data.csv
[January]+[February]+[March]+[April]+[May]+[June]+[July]+[Aug...
p2-2010-pawdacity-monthly-sales-SUMMED.csv
Unique: CITY
2014 Estimate=Replace ([2014 Estimate], "<td>", "")
Total population=[Land Area]*[Population Density]
!IsEmpty([City]) OR !IsEmpty([County])
p2-partially-parsed-wy-web-scrape-CLEANED.csv
All data without 2010census.csv
Total population=Round([Total population], 1) Land Area=Round([Land Area], 1) Households with Under 18=Round([Households with Under 18], 1)
p2-wy-demographic-data-WITHTPOPULATION.csv
Results - Workflow - Messages
0 Errors
0 Conv Errors
0 Warnings
33 Messages
10 Files
All

Output Data (14) 99 records were written to "C:\Users\vaz\Downloads\PAWDACITY\p2-partially-parsed-wy-web-scrape-CLEANED.csv"

Designer x64 Finished running Wkflw to clean-part1.yxmd in 0.6 seconds

Step 3: Dealing with Outliers

Answer these questions

Are there any 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.

Answer : Please note I do not have geographical knowledge of the cities provided. I am an Indian national. Therefore, I do not know, which of the given names is a city / town / village etc. Therefore, logical reasoning for considering an outlier has not been applied. I have relied only on numbers.

Based on the Interquartile range calculations, “Cheyenne” seems to be the outlier in question here. The reasoning is as follows:

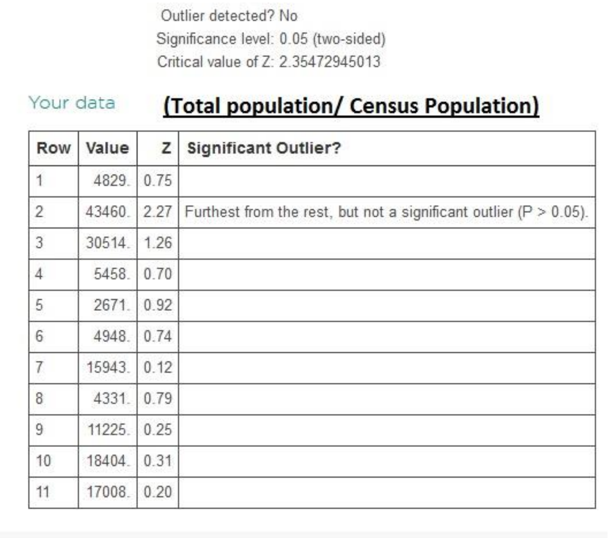
I calculated the outliers based on the interquartile range method.

I found three possible outliers

CITY	**OUTLIER VALUE UNDER THE CATEGORY**
CHEYENNE	Population Density & Total Sales
CASPER	Total Population
GILLETTE	Total Sales

To decide the answer, I used an online calculator to calculate z-score (through Grubbs test). This test showed the significance of an outlier:

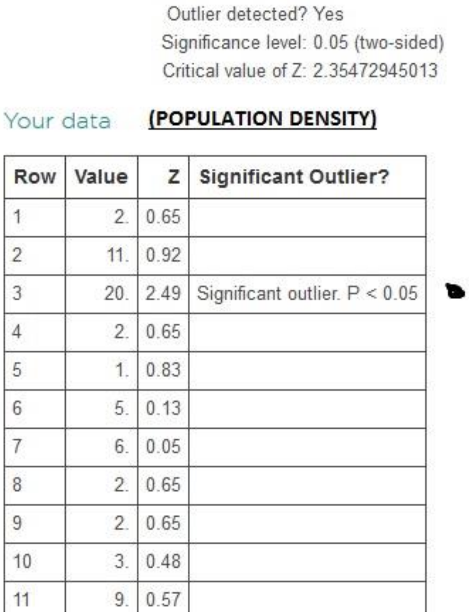
Based on the results, (Images attached Below), I concluded Cheyenne is the outlier.



Outlier detected? No
Significance level: 0.05 (two-sided)
Critical value of Z: 2.35472945013

Your data **(Total population/ Census Population)**

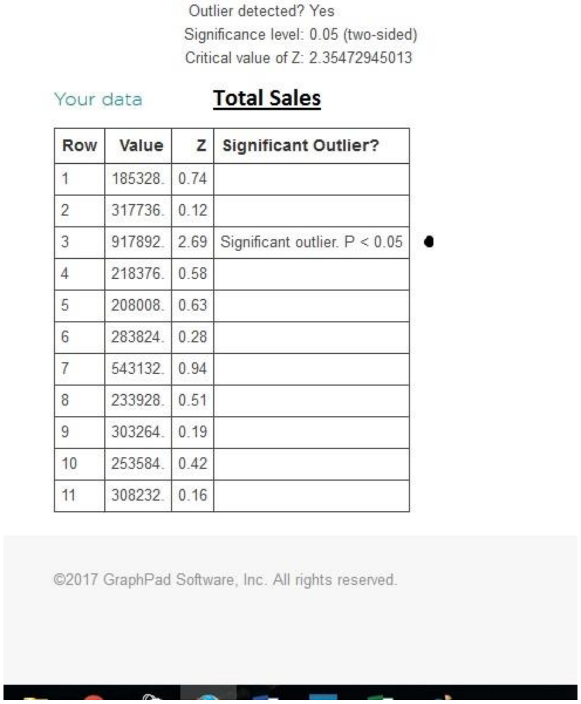
Row	Value	Z	Significant Outlier?
1	4829.	0.75	
2	43460.	2.27	Furthest from the rest, but not a significant outlier (P > 0.05).
3	30514.	1.26	
4	5458.	0.70	
5	2671.	0.92	
6	4948.	0.74	
7	15943.	0.12	
8	4331.	0.79	
9	11225.	0.25	
10	18404.	0.31	
11	17008.	0.20	



Outlier detected? Yes
Significance level: 0.05 (two-sided)
Critical value of Z: 2.35472945013

Your data **(POPULATION DENSITY)**

Row	Value	Z	Significant Outlier?
1	2.	0.65	
2	11.	0.92	
3	20.	2.49	Significant outlier. P < 0.05
4	2.	0.65	
5	1.	0.83	
6	5.	0.13	
7	6.	0.05	
8	2.	0.65	
9	2.	0.65	
10	3.	0.48	
11	9.	0.57	



Outlier detected? Yes
Significance level: 0.05 (two-sided)
Critical value of Z: 2.35472945013

Your data **Total Sales**

Row	Value	Z	Significant Outlier?
1	185328.	0.74	
2	317736.	0.12	
3	917892.	2.69	Significant outlier. P < 0.05
4	218376.	0.58	
5	208008.	0.63	
6	283824.	0.28	
7	543132.	0.94	
8	233928.	0.51	
9	303264.	0.19	
10	253584.	0.42	
11	308232.	0.16	

Based on the above identification of Cheyenne as the outlier, it is best to remove the city from the dataset . Imputation of data is not an option here at all, as complete data is available, there are no missing values as such and as we are predicting sales and the presence of an outlier will skew the further predictions we will make based on this dataset.

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.