Project2: Create an Analytical Dataset

Project2.1: Data Clean Up

PROJECT SUBMISSION WRITTEN BY CHIDIEBERE STEPHEN NWOSU

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 need to be made?

Answer:

The manager has asked me to expand and open a 14th pet store based on the predicted yearly sales from the datasets. The decision to be made is to cleanse, format and blend the different dataset together and deal with outlier before predicting the 14th store.

2. What data are needed to inform those decisions?

Answer:

The following data files are needed to inform those decisions:

- 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.

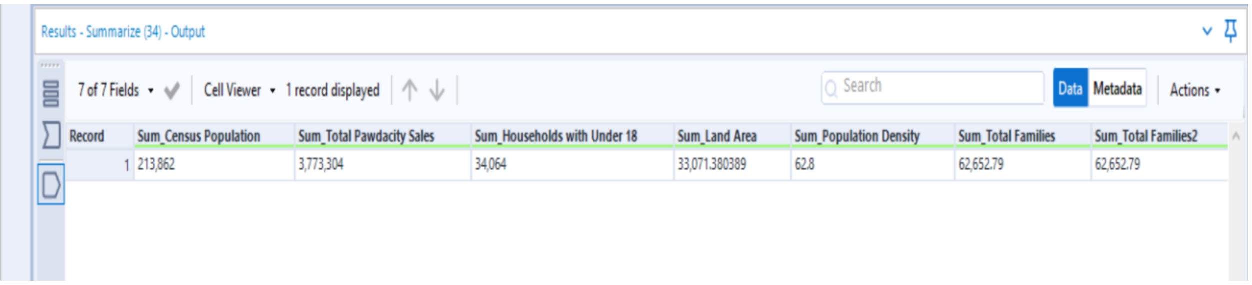
Step 2: Building the Training Set

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

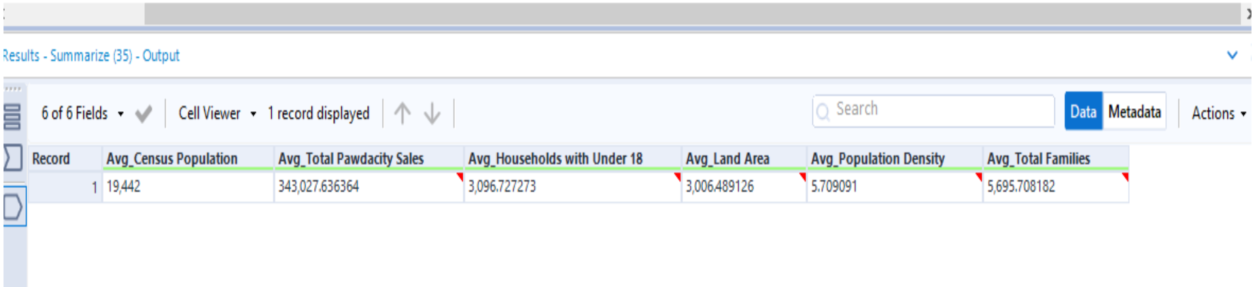
Column	**Sum**	**Average**
Census Population	213,862	19,442
Total Pawdacity Sales	3,773,304	*343,027.64*
Households with Under 18	34,064	*3,096.73*
Land Area	33,071	*3006.49*
Population Density	63	*5.71*
Total Families	62,653	*5,695.71*

Figure1



Record	Sum_Census Population	Sum_Total Pawdacity Sales	Sum_Households with Under 18	Sum_Land Area	Sum_Population Density	Sum_Total Families	Sum_Total Families2
1	213,862	3,773,304	34,064	33,071.380389	62.8	62,652.79	62,652.79

Figure2



Record	Avg_Census Population	Avg_Total Pawdacity Sales	Avg_Households with Under 18	Avg_Land Area	Avg_Population Density	Avg_Total Families
1	19,442	343,027.636364	3,096.727273	3,006.489126	5.709091	5,695.708182

Figure3

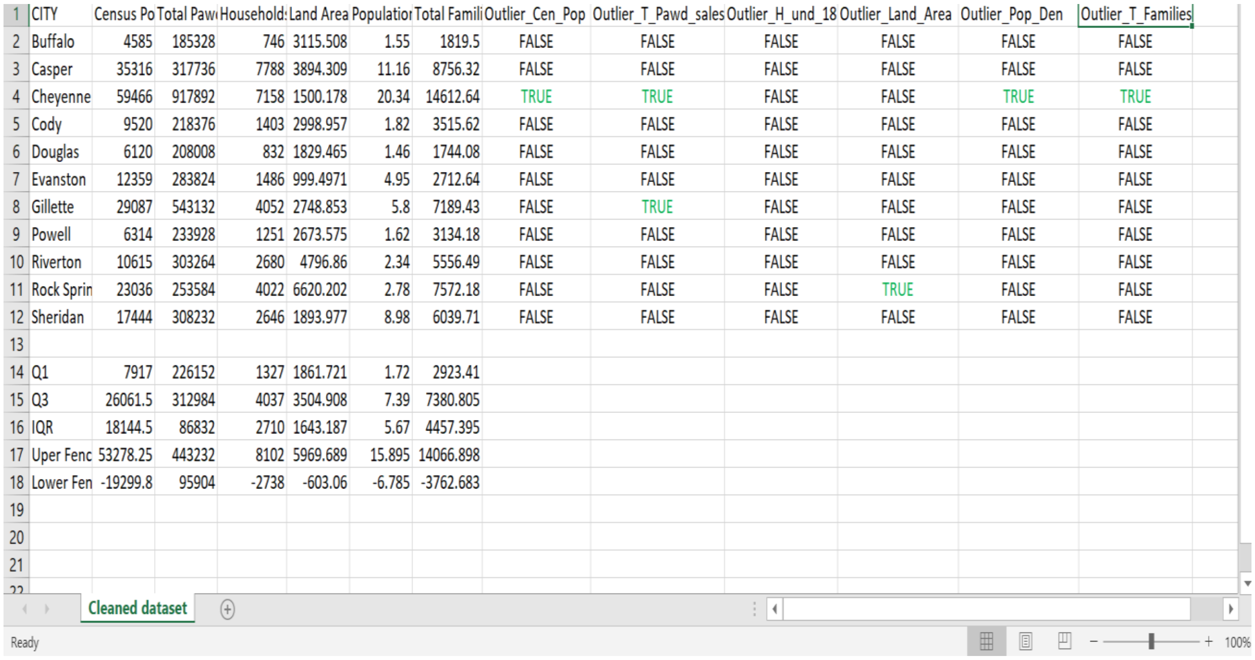
Step 3: Dealing with Outliers

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.

Answer:

There are 3 cities with outlier in the training set and they are; Cheyenne, Gillette and Rock Spring.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	CITY	Census Po	Total Paw	Household	Land Area	Populatior	Total Famili	Outlier_Cen_Pop	Outlier_T_Pawd_sales	Outlier_H_und_18	Outlier_Land_Area	Outlier_Pop_Den	Outlier_T_Families
2	Buffalo	4585	185328	746	3115.508	1.55	1819.5	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
3	Casper	35316	317736	7788	3894.309	11.16	8756.32	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
4	Cheyenne	59466	917892	7158	1500.178	20.34	14612.64	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE
5	Cody	9520	218376	1403	2998.957	1.82	3515.62	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
6	Douglas	6120	208008	832	1829.465	1.46	1744.08	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
7	Evanston	12359	283824	1486	999.4971	4.95	2712.64	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
8	Gillette	29087	543132	4052	2748.853	5.8	7189.43	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
9	Powell	6314	233928	1251	2673.575	1.62	3134.18	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
10	Riverton	10615	303264	2680	4796.86	2.34	5556.49	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
11	Rock Sprin	23036	253584	4022	6620.202	2.78	7572.18	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
12	Sheridan	17444	308232	2646	1893.977	8.98	6039.71	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
13													
14	Q1	7917	226152	1327	1861.721	1.72	2923.41						
15	Q3	26061.5	312984	4037	3504.908	7.39	7380.805						
16	IQR	18144.5	86832	2710	1643.187	5.67	4457.395						
17	Uper Fenc	53278.25	443232	8102	5969.689	15.895	14066.898						
18	Lower Fen	-19299.8	95904	-2738	-603.06	-6.785	-3762.683						
19													
20													
21													

Figure4; shows calculations of Outlier, 1st quartile, 2nd quartile, interquartile, lower and upper fence.

From figure3 above the table indicates the 3 outliers with the color green representing true for outliers.

Which outlier have you chosen to remove or input?

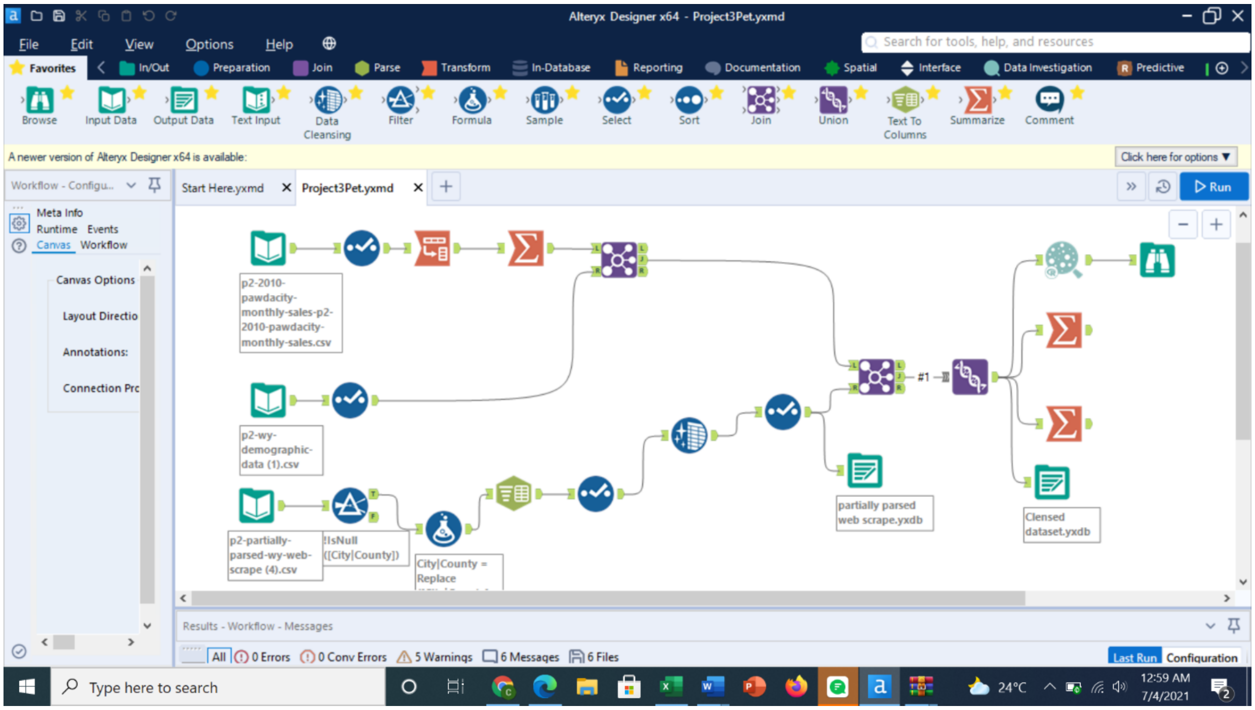
After taking a closer look at the data above I observed that there are 3 cities causing outlier accordingly.

1. Cheyenne city is observed to have outlier cutting across all the fields except Household with under 18 and Land Area. The city is in high population density, extremely high in total sale but small land mass area. Cheyenne city, is causing inconsistency and biasness. It may skew our model from what is reasonable or acceptable, this city should be removed.
2. Gillette city: Gillette city has an extreme high value in sales. This is reasonable because it shows consistent reasonable high values in household under 18, population density and Total families. It has only an outlier in total sales. Therefore, I decided to keep it.
3. Rock Spring: this city has an outlier in land area, it is very big city. It has consistent data across the fields so it is vital for our model in predicting sales. The fact is our sample size is small and if we decide to remove this city, we'll lose a very important information for our model. The caveat here is, land area does not impact on sales, yet I'll keep this city.

Justifications

I may justify removing Cheyenne because, it is inconsistent, skew high in sales and different from other cities or justify keeping it because, it shows a linear relationship (linear regression).

Gillette: I justify keeping Gillette because, it is an outlier in one field or justify removing it because, it does not skew relative to other field except in total sales.


Alteryx Designer x64 - Project3Pet.yxmd
File Edit View Options Help
Search for tools, help, and resources
Favorites In/Out Preparation Join Parse Transform In-Database Reporting Documentation Spatial Interface Data Investigation Predictive
Browse
Input Data
Output Data
Text Input
Data Cleansing
Filter
Formula
Sample
Select
Sort
Join
Union
Text To Columns
Summarize
Comment
Click here for options ▼
Workflow - Configu...
Meta Info
Runtime Events
Canvas Workflow
Canvas Options
Layout Directio
Annotations:
Connection Prc
Start Here.yxmd
Project3Pet.yxmd
Run
p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales.csv
p2-wy-demographic-data (1).csv
p2-partially-parsed-wy-web-scrape (4).csv
!IsNull ([City|County])
City|County = Replace
L
J
R
T
F
#1
partially parsed web scrape.yxdb
Clensed dataset.yxdb
Results - Workflow - Messages
All
Last Run Configuration
Type here to search
24°C
12:59 AM
7/4/2021

A newer version of Alteryx Designer x64 is available:

0 Errors 0 Conv Errors 5 Warnings 6 Messages 6 Files

Figure5; Work Flow