Title - Creating an Analytical Dataset

Description - Building cleaned dataset for model building

Problem Statement - Understand the business needs and then build datasets according to those needs, such as removing null values and taking appropriate action for outliers.

Approach and Technique Used - Exploratory method for understanding data, then coming up with a dataset suitable for business problems.

Tool - Alteryx

What I Learned - Understanding business problem
Having a holistic approach to dataset building method
Building visualization for a better understanding of data
Handling null values and outliers

This project is part of the **Predictive Analytics for Business** NanoDegree program offered by Udacity.

Project Start:

Step 1: Business and Data Understanding

1. What decisions need to be made?

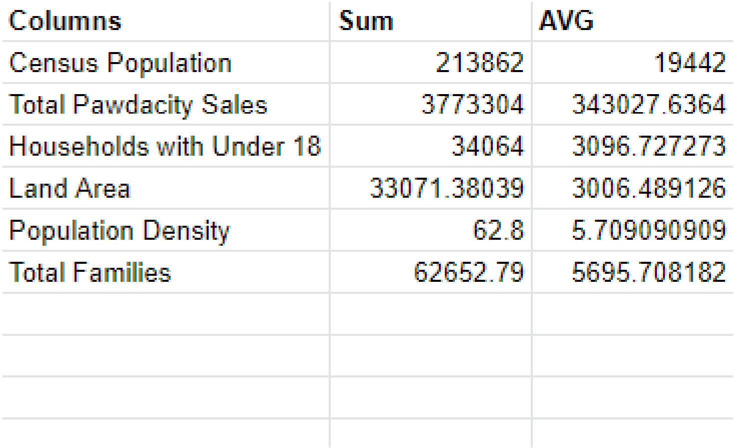
We need to find a location to open the 14th store in Wyoming based on predicted yearly sales data.

2. What data is needed to inform those decisions?

Pawda City past sales data and Wyoming demographics data.

Step 2: Building the Training Set

This is the SUM and AVG of all Fields came up with.



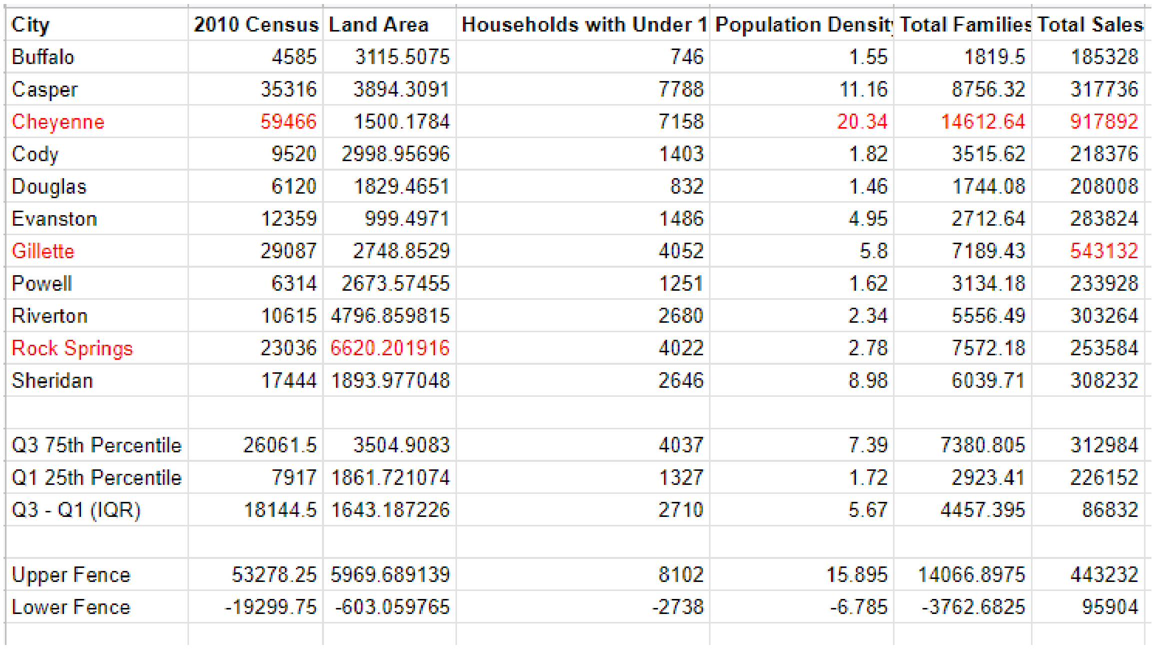
Columns	Sum	AVG
Census Population	213862	19442
Total Pawdacity Sales	3773304	343027.6364
Households with Under 18	34064	3096.727273
Land Area	33071.38039	3006.489126
Population Density	62.8	5.709090909
Total Families	62652.79	5695.708182

Step 3: Dealing with Outliers

I found three outliers using the IQR method; the first one is Cheyenne City. It has 4 outliers. 2010 census population, population density, total families, and total sales. In my justification deletion of this row would be best because it has many outliers.

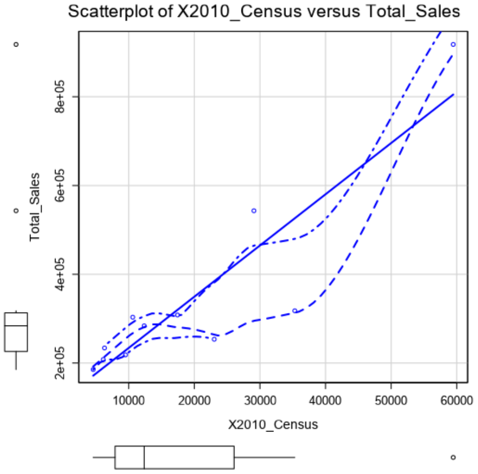
The Second outlier Is Gillette City's total sales. In my justification, it has only one outlier so it can be truncated.

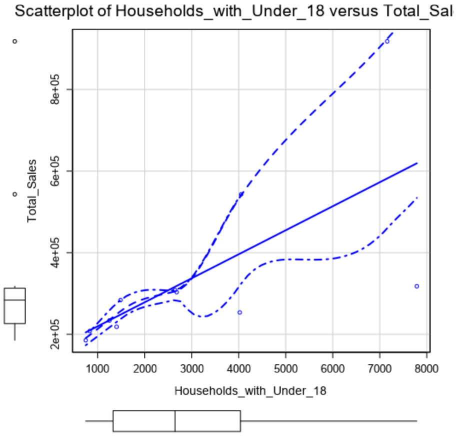
The 3rd outlier is the Rock Springs land area. In my justification, it has only one outlier so it can be truncated.

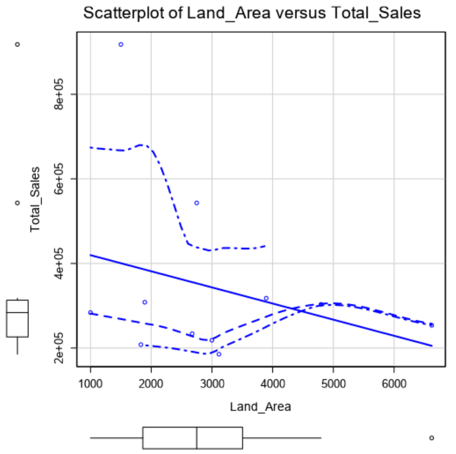


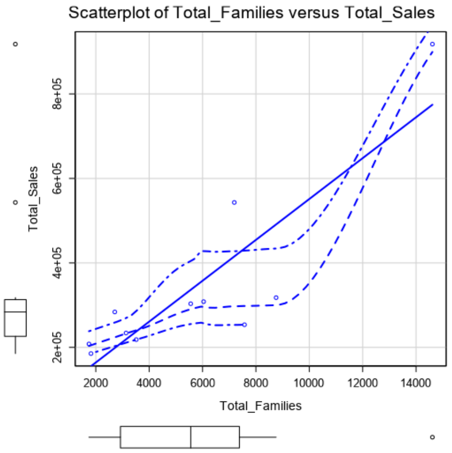
City	2010 Census	Land Area	Households with Under 1	Population Densit	Total Families	Total Sales
Buffalo	4585	3115.5075	746	1.55	1819.5	185328
Casper	35316	3894.3091	7788	11.16	8756.32	317736
Cheyenne	59466	1500.1784	7158	20.34	14612.64	917892
Cody	9520	2998.95696	1403	1.82	3515.62	218376
Douglas	6120	1829.4651	832	1.46	1744.08	208008
Evanston	12359	999.4971	1486	4.95	2712.64	283824
Gillette	29087	2748.8529	4052	5.8	7189.43	543132
Powell	6314	2673.57455	1251	1.62	3134.18	233928
Riverton	10615	4796.859815	2680	2.34	5556.49	303264
Rock Springs	23036	6620.201916	4022	2.78	7572.18	253584
Sheridan	17444	1893.977048	2646	8.98	6039.71	308232
Q3 75th Percentile	26061.5	3504.9083	4037	7.39	7380.805	312984
Q1 25th Percentile	7917	1861.721074	1327	1.72	2923.41	226152
Q3 - Q1 (IQR)	18144.5	1643.187226	2710	5.67	4457.395	86832
Upper Fence	53278.25	5969.689139	8102	15.895	14066.8975	443232
Lower Fence	-19299.75	-603.059765	-2738	-6.785	-3762.6825	95904

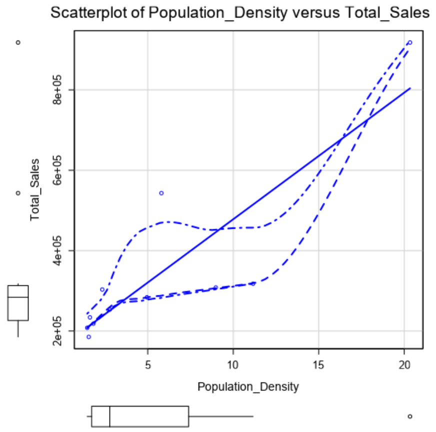
Scatter Plots for all other variables.



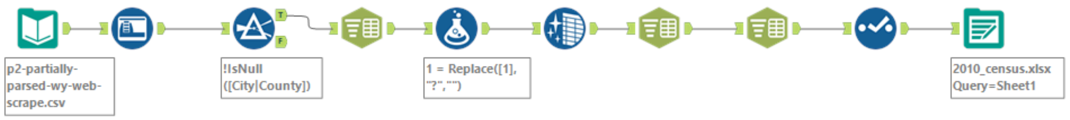




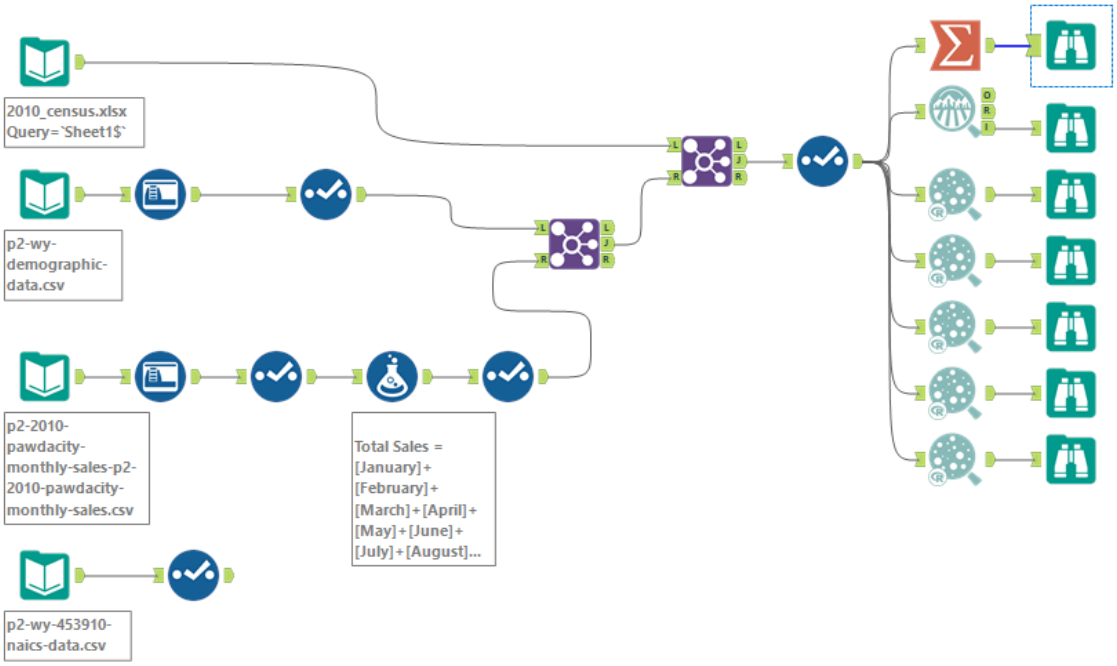




Alteryx Workflow 1


p2-partially-parsed-wy-web-scrape.csv
!IsNull ([City|County])
T
F
1 = Replace([1], "?","")
2010_census.xlsx Query=Sheet1

Alteryx Workflow 2


2010_census.xlsx
Query=`Sheet1$`
p2-wy-demographic-data.csv
p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales.csv
Total Sales = [January]+[February]+[March]+[April]+[May]+[June]+[July]+[August]...
p2-wy-453910-naics-data.csv
L
J
R
L
J
R
L
R
L
R
O
R
I