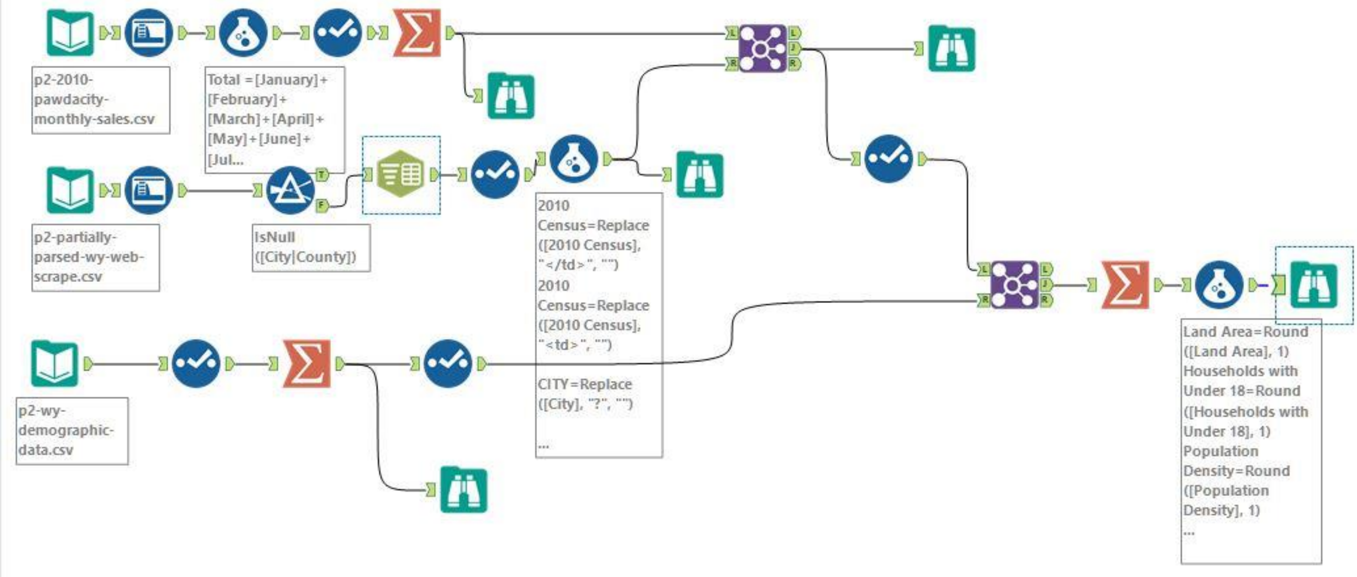
Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?
   a) We have a business problem which involves blending parsing and cleaning data in order to choose a city location for the new Pawdacity store.
   b) We have a lot of data in this problem. However, we need to properly format the data to form the training data set.
   c) Therefore, the data needs to be treated with appropriate data cleaning tools in proper order such as removing null values, removing random string characters, removing spaces, aggregating data and blending datasets to obtain the final training dataset.

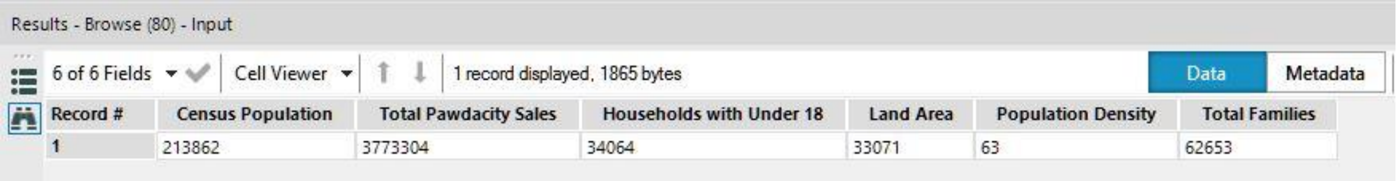
2. What data is needed to inform those decisions?
   a) Data on the consolidated sales of Pawdacity annually across the cities.
   b) Population and other demographic data consolidated according to the cities in which Pawdacity has operations.


p2-2010-pawdacity-monthly-sales.csv
Total =[January]+[February]+[March]+[April]+[May]+[June]+[Jul...
p2-partially-parsed-wy-web-scrape.csv
IsNull ([City|County])
T
F
2010 Census=Replace ([2010 Census], "</td>", "")
2010 Census=Replace ([2010 Census], "<td>", "")
CITY=Replace ([City], "?", "")
...
L
J
R
L
R
L
J
R
L
R
p2-wy-demographic-data.csv
Land Area=Round ([Land Area], 1)
Households with Under 18=Round ([Households with Under 18], 1)
Population Density=Round ([Population Density], 1)
...

.

Step 2: Building the Training Set

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*	3006.46
Population Density	*63*	5.73
Total Families	*62,653*	5695.73

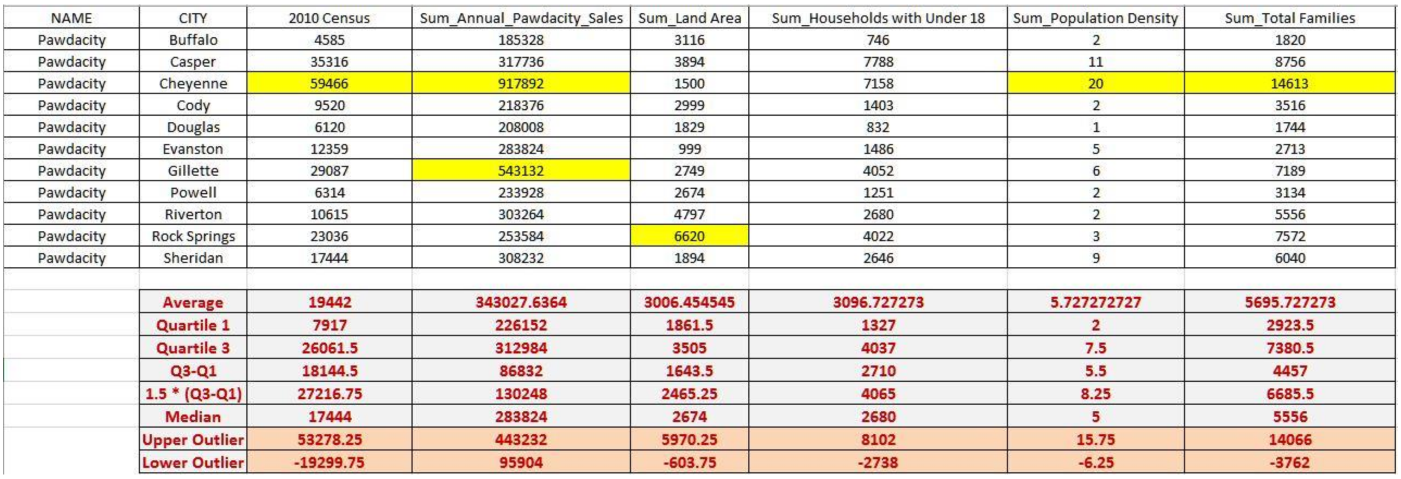


Results - Browse (80) - Input

6 of 6 Fields | Cell Viewer | 1 record displayed, 1865 bytes | Data | Metadata

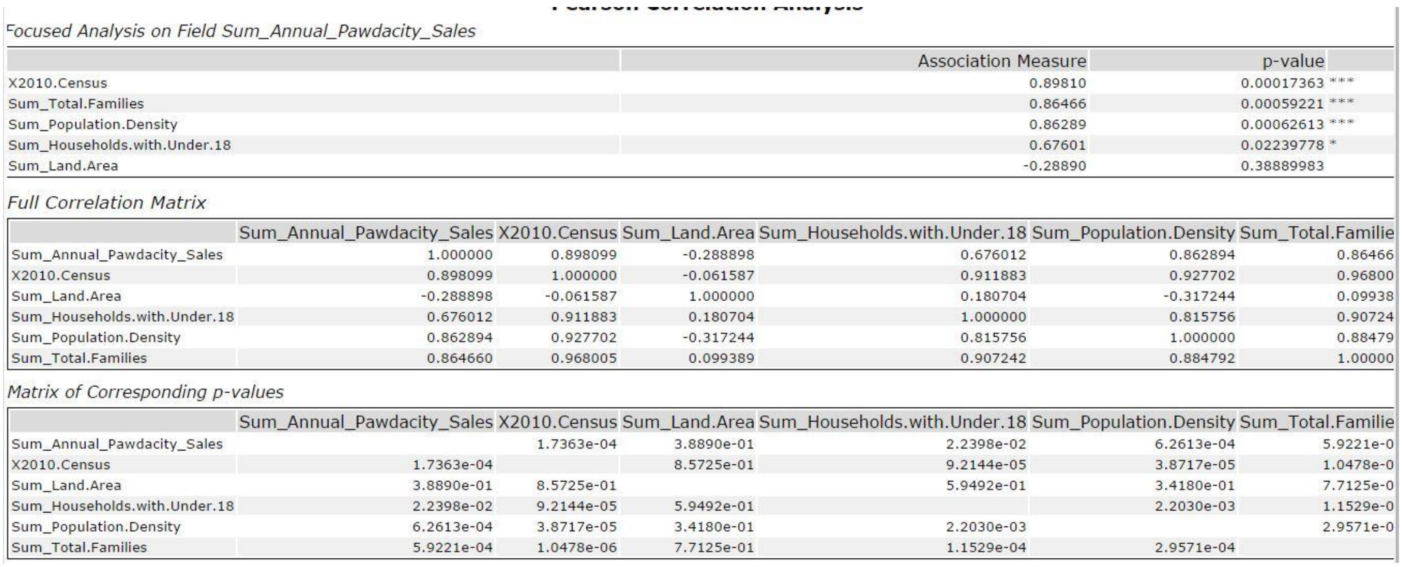
Record #	Census Population	Total Pawdacity Sales	Households with Under 18	Land Area	Population Density	Total Families
1	213862	3773304	34064	33071	63	62653

Step 3: Dealing with Outliers



NAME	CITY	2010 Census	Sum_Annual_Pawdacity_Sales	Sum_Land Area	Sum_Households with Under 18	Sum_Population Density	Sum_Total Families
Pawdacity	Buffalo	4585	185328	3116	746	2	1820
Pawdacity	Casper	35316	317736	3894	7788	11	8756
Pawdacity	Cheyenne	59466	917892	1500	7158	20	14613
Pawdacity	Cody	9520	218376	2999	1403	2	3516
Pawdacity	Douglas	6120	208008	1829	832	1	1744
Pawdacity	Evanston	12359	283824	999	1486	5	2713
Pawdacity	Gillette	29087	543132	2749	4052	6	7189
Pawdacity	Powell	6314	233928	2674	1251	2	3134
Pawdacity	Riverton	10615	303264	4797	2680	2	5556
Pawdacity	Rock Springs	23036	253584	6620	4022	3	7572
Pawdacity	Sheridan	17444	308232	1894	2646	9	6040
	Average	**19442**	**343027.6364**	**3006.454545**	**3096.727273**	**5.727272727**	**5695.727273**
	Quartile 1	**7917**	**226152**	**1861.5**	**1327**	**2**	**2923.5**
	Quartile 3	**26061.5**	**312984**	**3505**	**4037**	**7.5**	**7380.5**
	Q3-Q1	**18144.5**	**86832**	**1643.5**	**2710**	**5.5**	**4457**
	1.5 * (Q3-Q1)	**27216.75**	**130248**	**2465.25**	**4065**	**8.25**	**6685.5**
	Median	**17444**	**283824**	**2674**	**2680**	**5**	**5556**
	Upper Outlier	**53278.25**	**443232**	**5970.25**	**8102**	**15.75**	**14066**
	Lower Outlier	**-19299.75**	**95904**	**-603.75**	**-2738**	**-6.25**	**-3762**

Yes, there are outliers present which are highlighted in yellow in the above image file.
An association analysis using Alteryx was, using Sum of Annual Pawdacity sales as the target variable, to determine the statistical significance of the association of sales with the other variables.



Pearson Correlation Analysis

Focused Analysis on Field Sum_Annual_Pawdacity_Sales

	Association Measure	p-value
X2010.Census	0.89810	0.00017363 ***
Sum_Total.Families	0.86466	0.00059221 ***
Sum_Population.Density	0.86289	0.00062613 ***
Sum_Households.with.Under.18	0.67601	0.02239778 *
Sum_Land.Area	-0.28890	0.38889983

Full Correlation Matrix

	Sum_Annual_Pawdacity_Sales	X2010.Census	Sum_Land.Area	Sum_Households.with.Under.18	Sum_Population.Density	Sum_Total.Familie
Sum_Annual_Pawdacity_Sales	1.000000	0.898099	-0.288898	0.676012	0.862894	0.86466
X2010.Census	0.898099	1.000000	-0.061587	0.911883	0.927702	0.96800
Sum_Land.Area	-0.288898	-0.061587	1.000000	0.180704	-0.317244	0.09938
Sum_Households.with.Under.18	0.676012	0.911883	0.180704	1.000000	0.815756	0.90724
Sum_Population.Density	0.862894	0.927702	-0.317244	0.815756	1.000000	0.88479
Sum_Total.Families	0.864660	0.968005	0.099389	0.907242	0.884792	1.00000

Matrix of Corresponding p-values

	Sum_Annual_Pawdacity_Sales	X2010.Census	Sum_Land.Area	Sum_Households.with.Under.18	Sum_Population.Density	Sum_Total.Familie
Sum_Annual_Pawdacity_Sales		1.7363e-04	3.8890e-01	2.2398e-02	6.2613e-04	5.9221e-0
X2010.Census	1.7363e-04		8.5725e-01	9.2144e-05	3.8717e-05	1.0478e-0
Sum_Land.Area	3.8890e-01	8.5725e-01		5.9492e-01	3.4180e-01	7.7125e-0
Sum_Households.with.Under.18	2.2398e-02	9.2144e-05	5.9492e-01		2.2030e-03	1.1529e-0
Sum_Population.Density	6.2613e-04	3.8717e-05	3.4180e-01	2.2030e-03		2.9571e-0
Sum_Total.Families	5.9221e-04	1.0478e-06	7.7125e-01	1.1529e-04	2.9571e-04	

As we can see, the variables Sum_under18 and total land area are less significant variables and hence outlier values here will not skew the predictive model that includes it. Hence, Rock Springs field need not be imputed. Now, among Cheyenne and Gillette, we find that in Cheyenne, all values are in outlier range except for the two above mentioned less significant variables. In Gillette, only the final sum of sales in in the outlier range. Hence, it is likely that the values for Cheyenne are significant and correlated while for Gillette, the total sales may be not related significantly to the population metrics. Hence, Gillette field data may be imputed as outlier.