Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

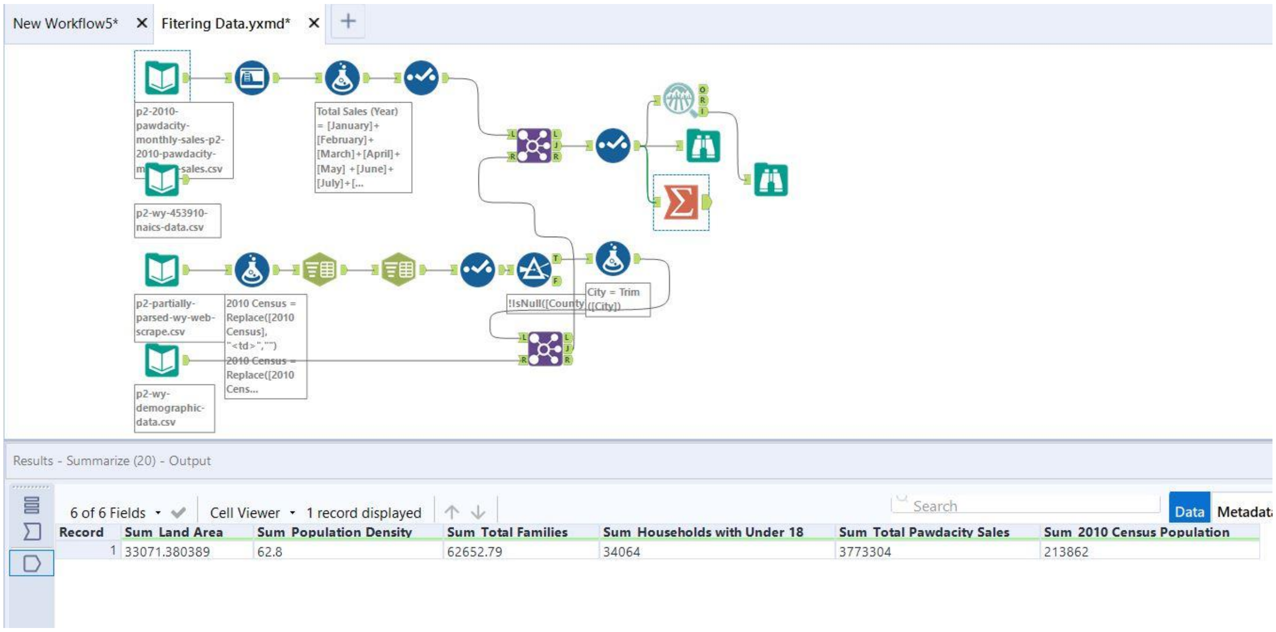
Key Decisions:

1. What decisions needs to be made?
2. What data is needed to inform those decisions?

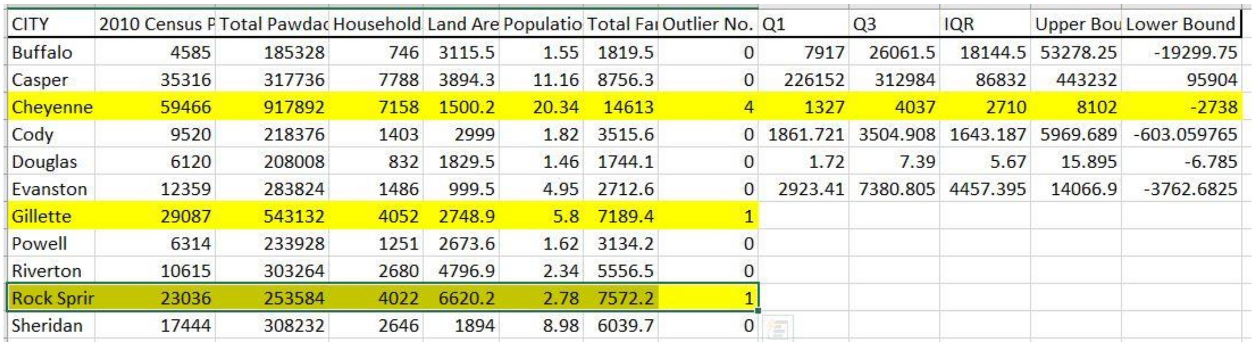
The following project showcases a detailed report of a leading pet store chain called Pawdacity who intends to open a new store in United States. This study highlights an overview into how the past sales data and other datasets are leveraged to identify the most lucrative location for opening a new outlet. The datasets needed to frame this decision are the population of the customers in the area with a demographic understanding of the regions, sales for the Pawdacity store for the past year, and competitor sales metrics (Serving in the same Domain).

Step 2: Building the Training Set

Column	**Sum**	**Average**
Census Population	*213,862*	*19442*
Total Pawdacity Sales	*3,773,304*	*3,43,027.64*
Households with Under 18	*34,064*	*3,096.73*
Land Area	*33,071.38*	*3,006.49*
Population Density	*62.8*	*5.70*
Total Families	*62,652.79*	*5,695.71*


New Workflow5*
Fitering Data.yxmd*
p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-m... sales.csv
Total Sales (Year) = [January]+[February]+[March]+[April]+[May] +[June]+[July]+[...
p2-wy-453910-naics-data.csv
p2-partially-parsed-wy-web-scrape.csv
2010 Census = Replace([2010 Census],"<td>","")
2010 Census = Replace([2010 Cens...
p2-wy-demographic-data.csv
!IsNull([County
City = Trim ([City])
Results - Summarize (20) - Output
6 of 6 Fields
Cell Viewer
1 record displayed
Search
Data
Metadata

Record	Sum Land Area	Sum Population Density	Sum Total Families	Sum Households with Under 18	Sum Total Pawdacity Sales	Sum 2010 Census Population
1	33071.380389	62.8	62652.79	34064	3773304	213862



CITY	2010 Census P	Total Pawdac	Household	Land Are	Populatio	Total Fa	Outlier No.	Q1	Q3	IQR	Upper Bou	Lower Bound
Buffalo	4585	185328	746	3115.5	1.55	1819.5	0	7917	26061.5	18144.5	53278.25	-19299.75
Casper	35316	317736	7788	3894.3	11.16	8756.3	0	226152	312984	86832	443232	95904
Cheyenne	59466	917892	7158	1500.2	20.34	14613	4	1327	4037	2710	8102	-2738
Cody	9520	218376	1403	2999	1.82	3515.6	0	1861.721	3504.908	1643.187	5969.689	-603.059765
Douglas	6120	208008	832	1829.5	1.46	1744.1	0	1.72	7.39	5.67	15.895	-6.785
Evanston	12359	283824	1486	999.5	4.95	2712.6	0	2923.41	7380.805	4457.395	14066.9	-3762.6825
Gillette	29087	543132	4052	2748.9	5.8	7189.4	1					
Powell	6314	233928	1251	2673.6	1.62	3134.2	0					
Riverton	10615	303264	2680	4796.9	2.34	5556.5	0					
Rock Sprir	23036	253584	4022	6620.2	2.78	7572.2	1					
Sheridan	17444	308232	2646	1894	8.98	6039.7	0					

Step 3: Dealing with Outliers

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 three cities that have outliers are:

- Cheyenne
- Gillette
- Rock Springs

The most significant outlier in this particular dataset is the Cheyenne city as it exceeds the limits set under the upper or lower bound in four distinct categories mentioned below,

- 2010 Census Population
- Total Pawdacity Sales
- Population Density
- Total Families

This particular issue may skew the linear regression model enough to provide falsified results when modeling. Thus, eradicating this row might give us an ideal insight into the best possible location for our new store. The other way around this issue might be to truncate the categorical data up to the closest lower or upper fence.