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:

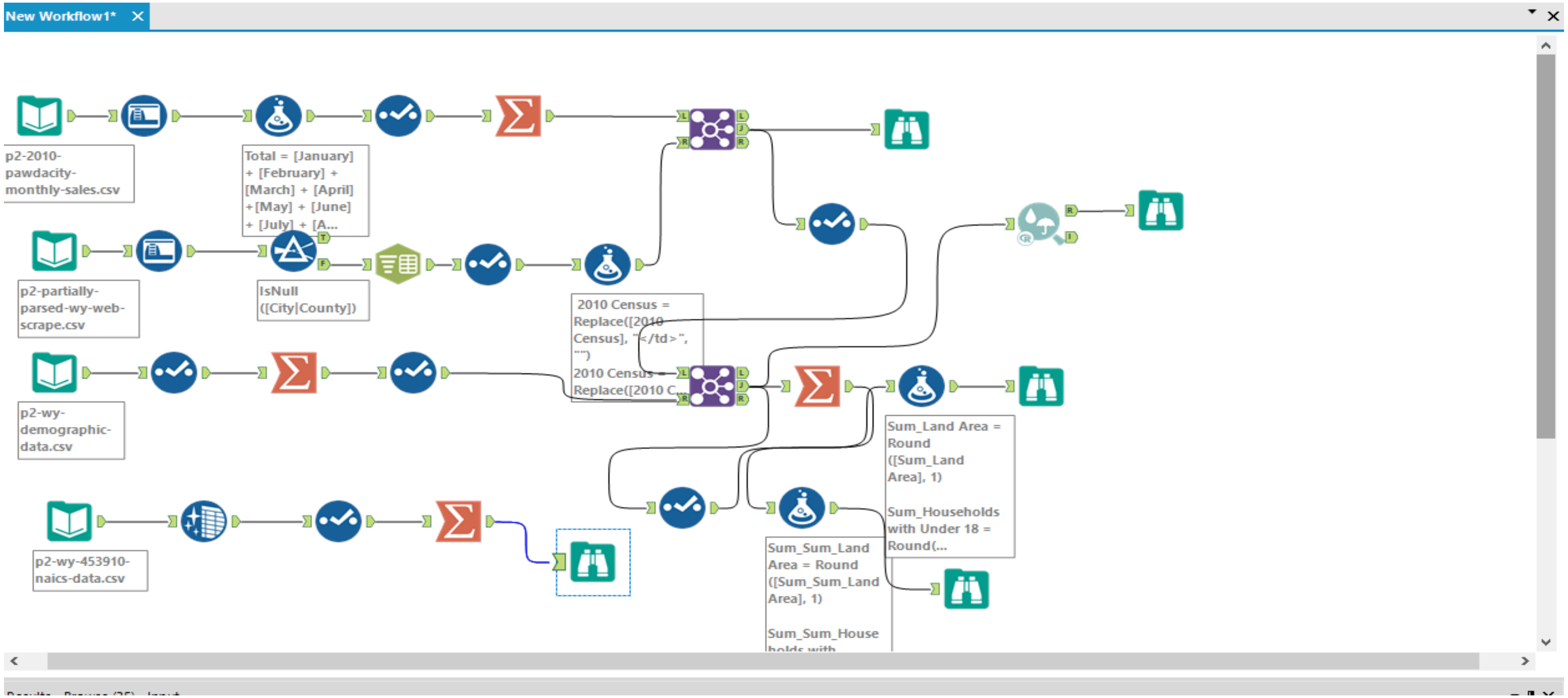
   - Understand the business data to facilitate in choosing the location of the city for the new pawdacity store to build 14th store (best location).

2. What data is needed to inform those decisions?

Answer:

- The total pawdacity sales across the cities
- Population density and census
- The demographic data of the cities where the Pawdacity has been performing to compete in sales revenue and favourable location to do business.
- Market data for the competitors

Step 2: Building the Training Set


New Workflow1*
p2-2010-pawdacity-monthly-sales.csv
p2-partially-parsed-wy-web-scrape.csv
p2-wy-demographic-data.csv
p2-wy-453910-naics-data.csv
T
F
L
J
R
D
I

Total = [January] + [February] + [March] + [April] +[May] + [June] + [July] + [A...

IsNull ([City|County])

2010 Census = Replace([2010 Census], "</td>", "")

2010 Census = Replace([2010 C...

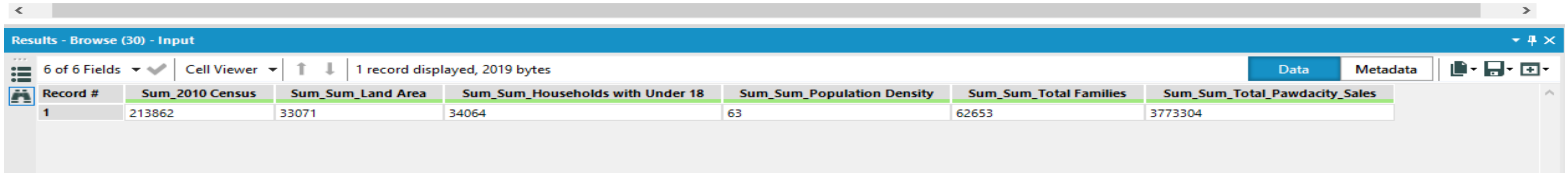
Sum_Land Area = Round ([Sum_Land Area], 1)

Sum_Households with Under 18 = Round(...

Sum_Sum_Land Area = Round ([Sum_Sum_Land Area], 1)

Sum_Sum_Households with

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	3006.45
Population Density	63	5.73
Total Families	62,653	5695.73



Record #	Sum_2010 Census	Sum_Sum_Land Area	Sum_Sum_Households with Under 18	Sum_Sum_Population Density	Sum_Sum_Total Families	Sum_Sum_Total_Pawdacity_Sales
1	213862	33071	34064	63	62653	3773304

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:

NAME	CITY	Sum_Total_Pawdacity_Sales	2010 Census	Sum_Land Area	Sum_Households with Under 18	Sum_Population Density	Sum_Total Families
Pawdacity	Buffalo	185328	4585	3116	746	2	1820
Pawdacity	Casper	317736	35316	3894	7788	11	8756
Pawdacity	Cheyenne	917892	59466	1500	7158	20	14613
Pawdacity	Cody	218376	9520	2999	1403	2	3516
Pawdacity	Douglas	208008	6120	1829	832	1	1744
Pawdacity	Evanston	283824	12359	999	1486	5	2713
Pawdacity	Gillette	543132	29087	2749	4052	6	7189
Pawdacity	Powell	233928	6314	2674	1251	2	3134
Pawdacity	Riverton	303264	10615	4797	2680	2	5556
Pawdacity	Rock Springs	253584	23036	6620	4022	3	7572
Pawdacity	Sheridan	308232	17444	1894	2646	9	6040
	Average	**343027.64**	**19442**	**3006.45**	**3096.73**	**5.73**	**5695.73**
	Q1	**226152**	**7917**	**1861.5**	**1327**	**2**	**2923.5**
	Q3	**312984**	**26061.5**	**3505**	**4037**	**7.5**	**7380.5**
	Q3-Q1	**86832**	**18144.5**	**1643.5**	**2710**	**5.5**	**4457**
	1.5*(Q3-Q1)	**130248**	**27216.75**	**2465.25**	**4065**	**8.25**	**6685.5**
	Upper Outlier	**443232**	**53278.25**	**5970.25**	**8102**	**15.75**	**14066**
	Lower outlier	**95904**	**-19299.75**	**-603.75**	**-2738**	**-6.25**	**-3762**

Yes, the blue colour on the above table signifies the outliers. With the help of the association analysis tool ensures the sum_Pawdacity_sales as the target variable in order to decide the statistical significance of the association between sales and other variables.

I realised that Cheyenne, Gillette and Rock Springs have higher magnitude of variables than the other cities when it comes to land area, pollution density, total families and total Pawdacity_Sales. Therefore I have choosing **Cheyenne, Gillette and Rock Springs** as outliers.

Cheyenne is the largest city among the other outliers (**Gillette** and **Rock Springs**) with a sum_population density of 20. It could be observed that the sum_total pawadcity sales and sum_population density shows more correlation or relationship and shows an effort not to skew data. Therefore retaining the dataset will be good enough to help in future modelling of other big cities. Again the **Rock Springs** city does not skew data when it comes to sales but it has limited data therefore I will keep it.

Based on the interquartile range (IQR), it seems only one of the outliers (**Gillette**) will be remove. Since the main focus of the analysis, is on the outlier sales but it could be seen that the population is comparatively small. As a results, it will be difficult to generate huge sale revenues.