STEP 1: DATA AND BUSINESS UNDERSTANDING

Key Decisions:

Answer these questions

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

The main objective of this project is to build an analytical dataset by combining all the three datasets provided with us. The resulted dataset will then be used to perform analysis to recommend a perfect city for pawdacity to open their new pet store. This recommendation will be made with the help of predictive modelling method built on top of this resulted dataset, hence, this dataset must be suitable to use for building a predictive modelling algorithm.

In order to build this final dataset, we need to ensure that we have all the necessary datum which corresponds to the sales occurs in the previous years. Data's such as geography - where the shops are located, population in that geographical area, since this is gonna be a pet shop - let's assume that kids are more interested in having pets than grown ups, so we need a data of the population of 18 minus people living in that geographical area. Since we are asked to do a city level analysis, we need to have a city wise geographical and population data than country wise data. Furthermore, as we are going to recommend the location based on the sales data, we must also need a previous years sales data as our target data.

STEP 2: BUILDING A TRAINING DATASET

Build your training set given the data provided to you. Your column sums of your dataset should match the sums in the table below.

	A	B	C	D	E	F	G
1	City	Land Area	Households with Under	Population Density	Total Families	Census Population	Total_Sales
2	Buffalo	3115.507568	746	1.549999952	1819.5	4585	185328
3	Casper	3894.309082	7788	11.15999985	8756.320313	35316	317736
4	Cheyenne	1500.178345	7158	20.34000015	14612.63965	59466	917892
5	Cody	2998.957031	1403	1.820000052	3515.620117	9520	218376
6	Douglas	1829.465088	832	1.460000038	1744.079956	6120	208008
7	Evanston	999.4970703	1486	4.949999809	2712.639893	12359	283824
8	Gillette	2748.852783	4052	5.800000191	7189.430176	29087	543132
9	Powell	2673.574463	1251	1.620000005	3134.179932	6314	233928
10	Riverton	4796.859863	2680	2.339999914	5556.490234	10615	303264
11	RockSprings	6620.202148	4022	2.779999971	7572.180176	23036	253584
12	Sheridan	1893.977051	2646	8.979999542	6039.709961	17444	308232
13							
14	SUM OF ALL COLUMNS	33071	34064	63	62653	213862	3773304
15	AVERAGE OF ALL COLUMNS	3006.489136	3096.727273	5.709090861	5695.708219	19442	343027.6364
16							
17	OUTLIER INFORMATION:						
18	OUTLIER INFORMATION	Land Area	Households with Under	Population Density	Total Families	Census Population	Total_Sales
19	IQR 1	1861.721069	1327	1.720000029	2923.409912	7917	226152
20	IQR 3	3504.908325	4037	7.389999866	7380.805176	26061.5	312984
21	IQR	1643.187256	2710	5.669999838	4457.395264	18144.5	86832
22	UPPER FENCE	5969.689209	8102	15.89499962	14066.89807	53278.25	443232
23	LOWER FENCE	-603.0598145	-2738	-6.784999728	-3762.682983	-19299.75	95904

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*	19442
Total Pawdacity Sales	*3,773,304*	343027.6364
Households with Under 18	*34,064*	3097
Land Area	*33,071*	3006.489136
Population Density	*63*	5.709090861
Total Families	*62,653*	5695.708219

STEP 3: DEALING WITH OUTLIERS:

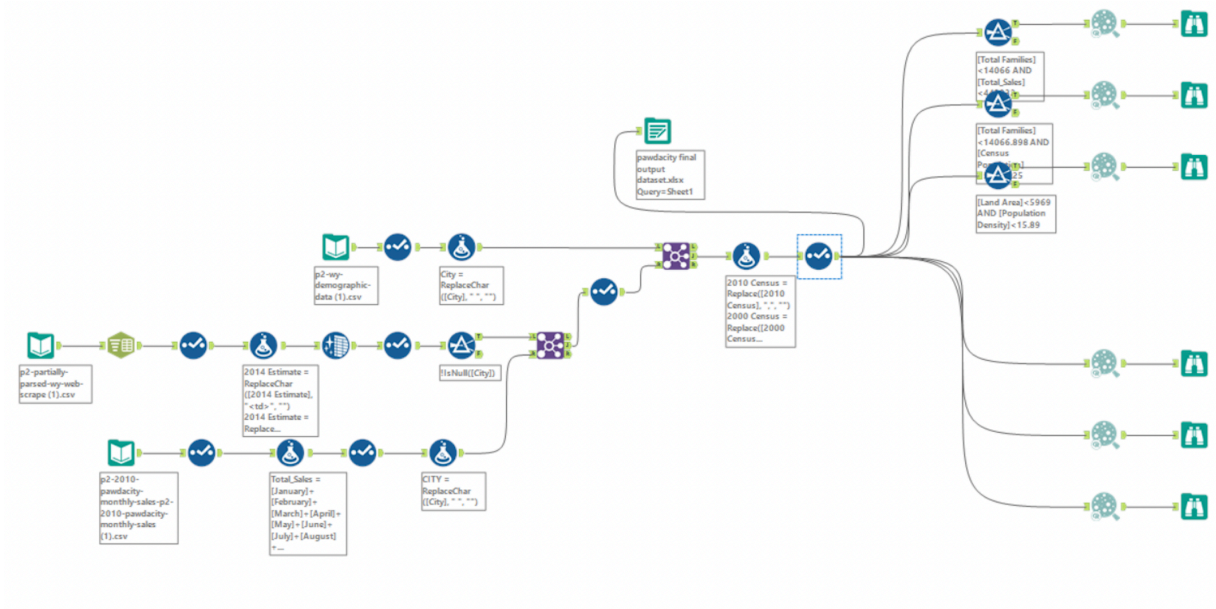
Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), you should only remove or impute one outlier. Please explain your reasoning.

By following the IQR method given in the material, I found out there are three cities which has an outlier values in them which might cause our analyses to fail to recommend a perfect cities for Pawdacity to open their new pet shops. The list of three cities which has an outlier values in them are

- Cheyenne.
- Gillette and
- RockSprings.

Since the dataset we have in our hand is small, we cannot simply remove all the three cities with outlier values in them. If we do so, we have to face a severe data loss issues after that. Hence we have to choose one city to remove from the dataset by leaving the other two as they are. As Gillette and RockSprings have just one outlier columns in them and also since those outliers are acceptable considering their nature total_sales and land area, I decided to go with removing the city Cheyenne from the dataset since it has large number of outlier columns in it.

ALTERYX WORKFLOW:



Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.