Project 2.1: Data Cleanup

By: Anthony Nguyen

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?
   - The decision to select a suitable city for Pawdacity to open a new 14th store.

2. What data is needed to inform those decisions?
   - The following data is needed from the current 11 cities in which Pawdacity currently holds locations:
     - Land Area
     - Households With Under 18
     - Population Density
     - Total Families
     - 2010 Census
     - Total Pawdacity Sales
   - From this data, we determine which city holds outliers in which we can remove from the data set.

city	county	land area	household w/under18	pop density	total families	2010 census	jan	feb	mar	apr	may	june	july	aug	sept	oct	nov	dec
Buffalo	Johnson	3115.5075	746	1.55	1819.5	4585	16200	13592	14688	17064	18360	14040	12960	19224	15984	13592	13176	16848
Casper	Natrona	3894.3091	7788	11.16	8756.32	35316	29160	21600	27000	27648	29160	27216	25488	15704	22896	25272	28944	27648
Cheyenne	Laramie	1500.1784	7158	20.94	14612.64	59466	79920	70632	79056	77544	73656	77976	73872	77544	78516	74520	74736	79920
Cody	Park	2998.95696	1403	1.82	3515.62	9520	19440	15984	19008	18144	16632	17496	18792	20304	19224	18144	18576	16632
Douglas	Converse	1829.4651	832	1.46	1744.08	6120	16200	13392	14688	17064	18360	14040	12960	19224	15984	29808	17496	18792
Evanston	Uinta	999.4971	1486	4.95	2712.64	12359	24840	21168	21600	22248	24192	24624	25488	15704	22032	21168	25920	24840
Gillette	Campbell	2748.8529	4052	5.8	7189.43	29087	47520	41796	48384	47088	42336	41904	42120	47088	49032	48168	42984	44712
Powell	Park	2673.57455	1251	1.62	3134.18	8514	20520	17928	20304	21168	21600	17928	18144	18576	20304	21168	17496	18792
Riverton	Fremont	4796.859815	2680	2.34	5556.49	10615	27000	22032	28512	26784	25920	24192	25056	22896	25488	26352	26784	22248
Rock Springs	Sweetwater	6620.201916	4022	2.76	7572.18	23036	21600	19872	22248	20952	17496	24840	22464	21816	21384	20304	22032	18576
Sheridan	Sheridan	1895.977048	2646	8.98	6039.71	17444	27000	26352	28080	22032	21168	29376	25920	20304	33696	23760	25056	25488
TOTAL		33071.38039	34064	62.8	62652.79	213862	329400	284148	323568	317736	308880	313632	303264	318384	324540	322056	313200	314496
AVERAGES		3006.49	3096.73	5.71	5695.71	19442.00	29945.45	25831.64	29415.27	28885.09	28080.00	28512.00	27569.45	28944.00	29503.64	29277.82	28472.73	28590.55
	IQR	2064.844	2801	7.36	4859.54	22773	9720	10368	9504	9504	10800	11880	7776	6480	14472	9504	11448	9072
	1st quartile	1829.4651	1251	1.62	2712.64	6314	19440	15984	19008	18144	18360	17496	18144	19224	19224	20304	17496	18576
	Median	2748.8529	2646	2.78	5556.49	12359	24840	21168	22248	22032	21600	24624	25056	21816	22032	23760	25056	22248
	3rd quartile	3894.3091	4052	8.98	7572.18	29087	29160	26352	28512	27648	29160	29376	25920	25704	33696	29808	28944	27648
	Upper Fence	6991.5751	8253.5	20.02	14861.49	63246.5	43740	41904	42768	41904	45360	47196	37584	35424	55404	44064	46116	41256
	Lower Fence	-1267.8009	-2950.5	-9.42	-4576.67	-27845.5	4860	432	4752	5868	2160	-324	6480	9504	-2484	6048	324	4968
	Total Sales	3773304																
	Avg Total Sales	28585.64																
	OUTLIERS																	

Step 2: Building the Training Set

Column	Sum	Average
Census Population	*213,862*	*19,442.00*
Total Pawdacity Sales	*3,773,304*	*28,585.64*
Households with Under 18	*34,064*	*3,096.73*
Land Area	*33,071*	*3,006.49*
Population Density	*63*	*5.71*
Total Families	*62,653*	*5,695.71*

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.

- Yes, there are outliers in the training set = Cheyenne + Gillette
- The outlier chosen to remove = Cheyenne.
- Reasoning for removing Cheyenne is that nearly all of the variables are almost double or nearly meet the Upper Fence compared to the 2nd city that has variables outside the upper fence, Gillette.
    - Gillette is the 2nd city which has outliers in 1 variable, sales. 7/12 months of sales are outliers.
    - Cheyenne has outliers in 2 variables, population density + sales. 12/12 months of sales are outliers.