Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?

    - Pawdacity, a leading pet store chain in Wyoming, needs recommendation on where to open its 14th store.
    - 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?

    - Total Sales data of Pawdacity annually across the cities.
    - Population, Land Area, Total Families and other demographic data according to the cities in which Pawdacity has operations.

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*	**3006.49**
Population Density	*63*	**5.71**
Total Families	*62,653*	**5695.71**

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 present which are highlighted in yellow.

CITY	Sum_Total Sales	Sum_Land Area	Sum_Households with Under 18	Sum_Population Density	Sum_Total Families	Sum_2010 Census
Buffalo	185328	3115.5075	746	1.55	1819.5	4585
Casper	317736	3894.3091	7788	11.16	8756.32	35316
Cheyenne	917892	1500.1784	7158	20.34	14612.64	59466
Cody	218376	2998.95696	1403	1.82	3515.62	9520
Douglas	208008	1829.4651	832	1.46	1744.08	6120
Evanston	283824	999.4971	1486	4.95	2712.64	12359
Gillette	543132	2748.8529	4052	5.8	7189.43	29087
Powell	233928	2673.57455	1251	1.62	3134.18	6314
Riverton	303264	4796.859815	2680	2.34	5556.49	10615
Rock Springs	253584	6620.201916	4022	2.78	7572.18	23036
Sheridan	308232	1893.977048	2646	8.98	6039.71	17444
Average	343027.6364	3006.489126	3096.727273	5.709090909	5695.708182	19442
Quartile 1	226152	1861.721074	1327	1.72	2923.41	7917
Quartile 3	312984	3504.9083	4037	7.39	7380.805	26061.5
Q3-Q1	86832	1643.187226	2710	5.67	4457.395	18144.5
1.5*(Q3-Q1)	130248	2464.780839	4065	8.505	6686.0925	27216.75
Median	283824	2748.8529	2646	2.78	5556.49	12359
Upper Outlier	443232	5969.689139	8102	15.895	14066.8975	53278.25
Lower Outlier	95904	-603.059765	-2738	-6.785	-3762.6825	-19299.75

- 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 variable.

Pearson Correlation Analysis

Focused Analysis on Field Total.Sales

	Association Measure	p-value
X2010.Census	0.89810	0.00017363 ***
Total.Families	0.86466	0.00059221 ***
Population.Density	0.86289	0.00062613 ***
Households.with.Under.18	0.67601	0.02239778 *
Land.Area	-0.28890	0.38889985

Full Correlation Matrix

	Total.Sales	Land.Area	Households.with.Under.18	Population.Density	Total.Families	X2010.Census
Total.Sales	1.000000	-0.288898	0.676012	0.862894	0.864660	0.898099
Land.Area	-0.288898	1.000000	0.180704	-0.317244	0.099389	-0.061587
Households.with.Under.18	0.676012	0.180704	1.000000	0.815756	0.907242	0.911883
Population.Density	0.862894	-0.317244	0.815756	1.000000	0.884792	0.927702
Total.Families	0.864660	0.099389	0.907242	0.884792	1.000000	0.968005
X2010.Census	0.898099	-0.061587	0.911883	0.927702	0.968005	1.000000

Matrix of Corresponding p-values

	Total.Sales	Land.Area	Households.with.Under.18	Population.Density	Total.Families	X2010.Census
Total.Sales		3.8890e-01	2.2398e-02	6.2613e-04	5.9221e-04	1.7363e-04
Land.Area	3.8890e-01		5.9492e-01	3.4180e-01	7.7125e-01	8.5725e-01
Households.with.Under.18	2.2398e-02	5.9492e-01		2.2030e-03	1.1529e-04	9.2143e-05
Population.Density	6.2613e-04	3.4180e-01	2.2030e-03		2.9571e-04	3.8717e-05
Total.Families	5.9221e-04	7.7125e-01	1.1529e-04	2.9571e-04		1.0478e-06
X2010.Census	1.7363e-04	8.5725e-01	9.2143e-05	3.8717e-05	1.0478e-06	

- The variables **Households.with.under.18** and **Land.Area** are less significant variables and hence outlier values here will not affect the predictive model that includes it.
  Hence**, Rock Springs** field need not be imputed.

- We find that in **Cheyenne**, all values are in outlier range except for the two above mentioned less significant variables.
  Hence, it is likely that the values for **Cheyenne** are significant and correlated.

- In **Gillette**, only **Total Sales** is in the outlier range.Therefore, the total sales may be not related significantly to the population metrics.
  Hence, **Gillette** field data may be imputed as outlier.