Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Key Decisions:

1. **What decisions needs to be made?**

   The Decision which we need to made is **"In which city to open a new Pawdacity Store**".

2. **What data is needed to inform those decisions?**

   - The data needed to inform those decision is the current data from the 11 existing stores.
   - The Project details says to use Census Population, Total Pawdacity Sales, Households with under 18, Land Area, Population Density, Total Families.

Step 2: Building the Training Set

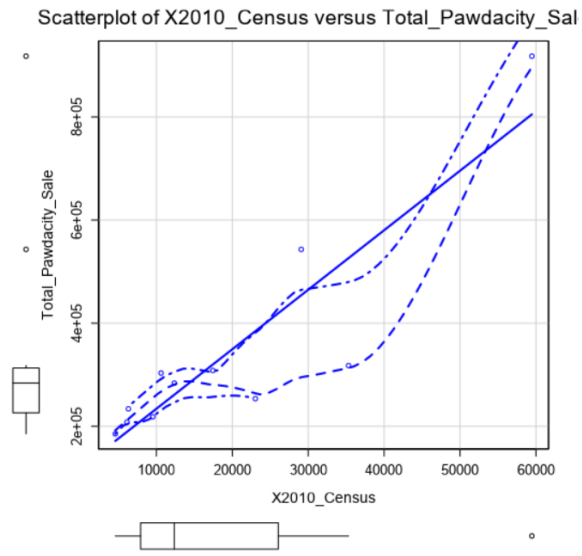
After performing the data cleansing with Alteryx on the given four datasets, the averages for the variables are mentioned below:

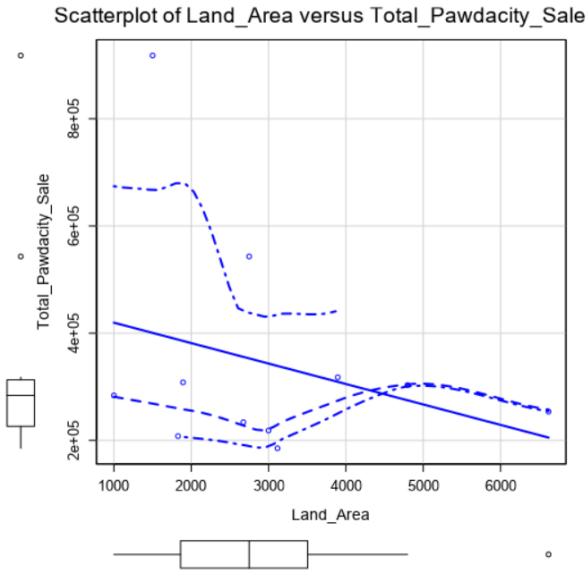
Column	Sum	Average
Census Population	**213,862**	**19,442**
Total Pawdacity Sales	**3,773,304**	**343,027.64**
Households with Under 18	**34,064**	**3,096.73**
Land Area	**33,071**	**3,006.49**
Population Density	**63**	**5.71**
Total Families	**62,653**	**5,695.71**

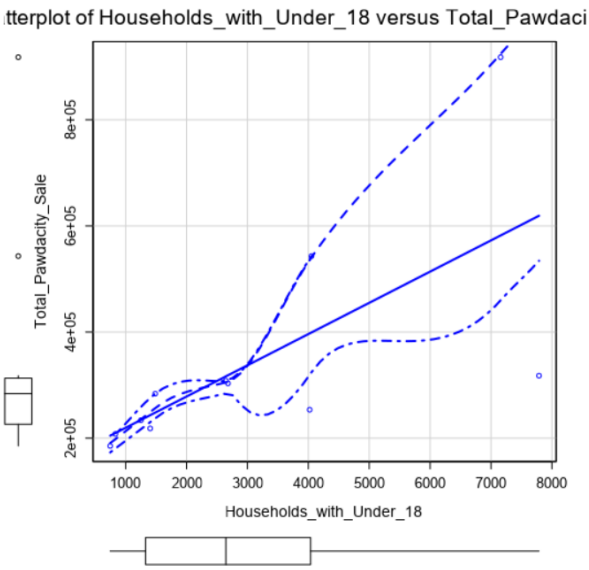
Step 3: Dealing with Outliers

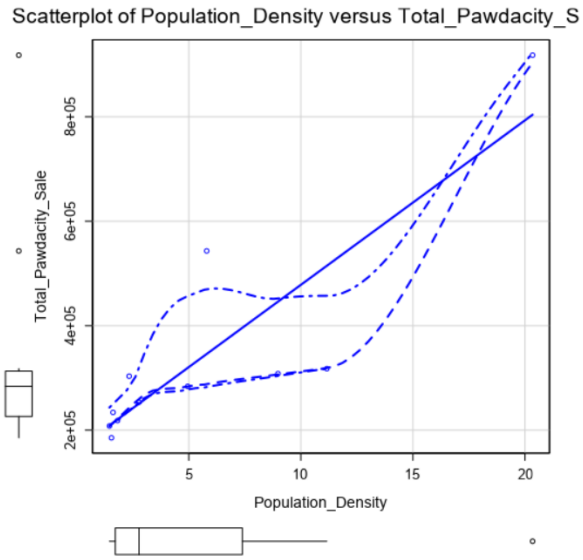
Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.

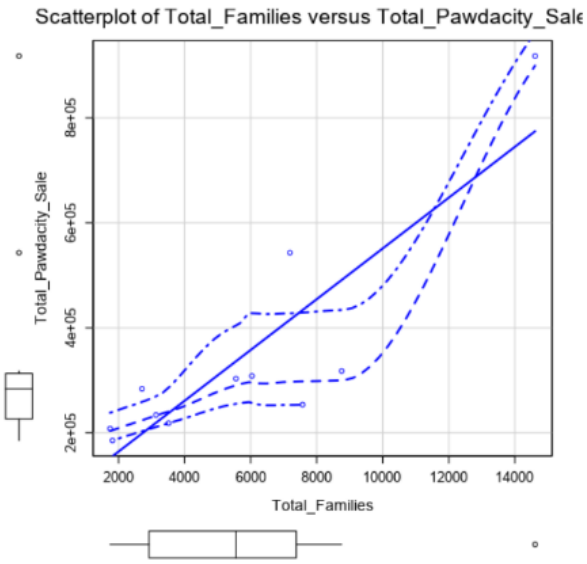
For better Understanding of Outliers, I have used Scatter Plot of Total Pawdacity sales Vs the other given variables:


Scatterplot of X2010_Census versus Total_Pawdacity_Sal
Total_Pawdacity_Sale
2e+05
4e+05
6e+05
8e+05
10000
20000
30000
40000
50000
60000
X2010_Census


Scatterplot of Land_Area versus Total_Pawdacity_Sale
Total_Pawdacity_Sale
2e+05
4e+05
6e+05
8e+05
1000
2000
3000
4000
5000
6000
Land_Area


tterplot of Households_with_Under_18 versus Total_Pawdaci
Total_Pawdacity_Sale
2e+05
4e+05
6e+05
8e+05
1000
2000
3000
4000
5000
6000
7000
8000
Households_with_Under_18


Scatterplot of Population_Density versus Total_Pawdacity_S
Total_Pawdacity_Sale
2e+05
4e+05
6e+05
8e+05
5
10
15
20
Population_Density


Scatterplot of Total_Families versus Total_Pawdacity_Sale
Total_Pawdacity_Sale
2e+05
4e+05
6e+05
8e+05
2000
4000
6000
8000
10000
12000
14000
Total_Families

	A	B	C	D	E	F	G	H
1	City	County	Total Pawdacity Sale	2010 Census	Land Area	Households with Under 18	Population Density	Total Families
2	Buffalo	Johnson	185328	4585	3115.5075	746	1.55	1819.5
3	Casper	Natrona	317736	35316	3894.3091	7788	11.16	8756.32
4	Cheyenne	Laramie	917892	59466	1500.1784	7158	20.34	14612.64
5	Cody	Park	218376	9520	2998.95696	1403	1.82	3515.62
6	Douglas	Converse	208008	6120	1829.4651	832	1.46	1744.08
7	Evanston	Uinta	283824	12359	999.4971	1486	4.95	2712.64
8	Gillette	Campbell	543132	29087	2748.8529	4052	5.8	7189.43
9	Powell	Park	233928	6314	2673.57455	1251	1.62	3134.18
10	Riverton	Fremont	303264	10615	4796.859815	2680	2.34	5556.49
11	Rock Springs	Sweetwater	253584	23036	6620.201916	4022	2.78	7572.18
12	Sheridan	Sheridan	308232	17444	1893.977048	2646	8.98	6039.71
13								
14	Q1		226152	7917	1861.721074	1327	1.72	2923.41
15	Q3		312984	26061.5	3504.9083	4037	7.39	7380.805
16	IQR		86832	18144.5	1643.187226	2710	5.67	4457.395
17	Upper Fence		443232	53278.25	5969.689139	8102	15.895	14066.8975
18	Lower Fence		95904	-19299.75	-603.059765	-2738	-6.785	-3762.6825
19								

From the Scatter Plot above and the Data Extracted, some of the observations are:

- Even though Cheyenne is flagged out as outlier but it is big city compared to the rest of the cities. Cheyenne's values for the different fields are larger in comparison with other cities even though it has a smaller number of stores.
- Comparing Gillette's with rest of the cities, its total sales are not in the proportion with other demographic fields such as population. If a city has a large sale, we would expect it to have a large population to drive those sales which isn't the case.
- For Land Area, Rock Spring was an outlier.

So, after considering all of the above point I will Suggest to remove Gillette from the dataset in order to get unbiased model.

My Alteryx Work Flow: