Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Pawdacity is a leading pet store chain in Wyoming with 13 stores throughout the state. We need to perform an analysis to recommend the city for Pawdacity newest store, based on predicted yearly sales and for that we will be needing historical data of the store and the Demographic data to perform our analysis.

Key Decisions:

1. What decisions needs to be made?
   Pet store chain has 13 stores throughout the state and would like to expand and open a 14th store. We need to perform an analysis to recommend the city for the newest store, based on predicted yearly sales.

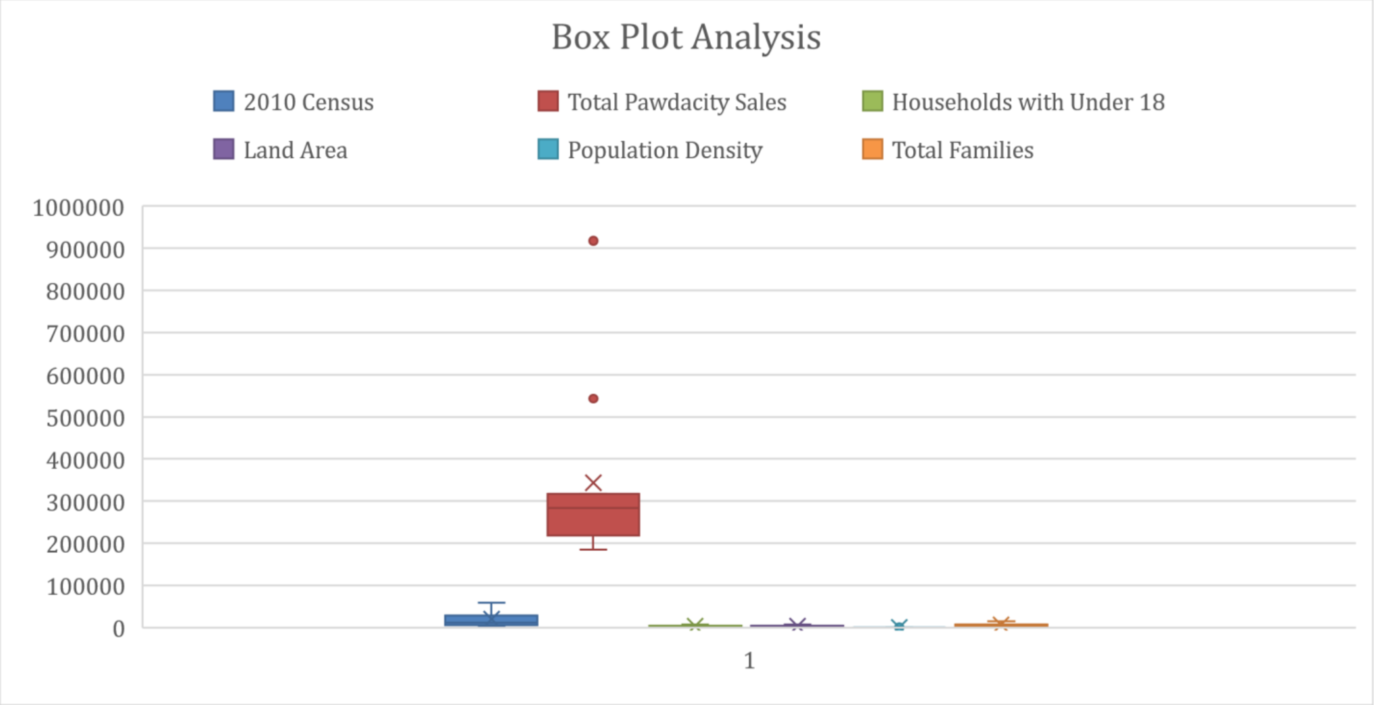
2. What data is needed to inform those decisions?
   We will be requiring the Sales data for all of the stores at city level and the Demographic data like Households, Land Area, Population Density and Total Families for each city to estimate and recommend the new location.

Step 2: Building the Training Set

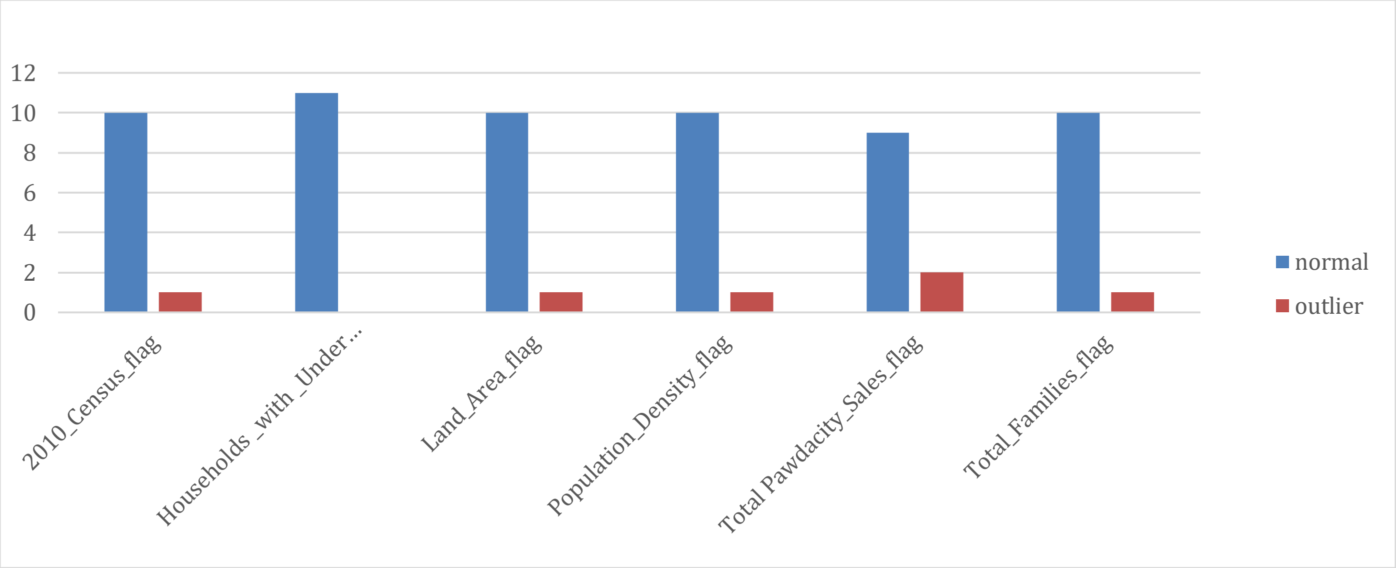
Column	**Sum**	**Average**
Census Population	*213,862*	19442
Total Pawdacity Sales	*3,773,304*	343027.64
Households with Under 18	*34,064*	3096.73
Land Area	*33,071*	3006.49
Population Density	*63*	5.71
Total Families	*62,653*	5695.71

Step 3: Dealing with Outliers

<u>Outlier Analysis</u>

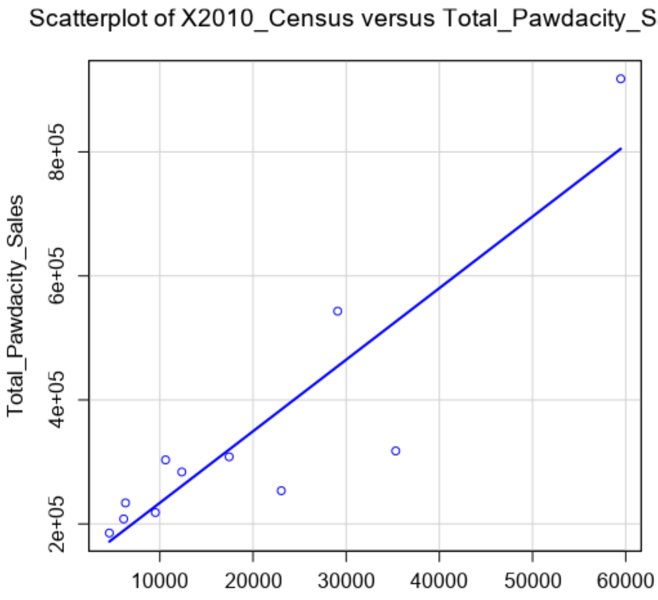


Sum of Count	Column Labels		
Row Labels	**normal**	**outlier**	**Grand Total**
2010_Census	10	1	11
Households _with _Under _18	11		11
Land_Area	10	1	11
Population_Density	10	1	11
Total Pawdacity_Sales	9	2	11
Total_Families	10	1	11
Grand Total	**60**	**6**	**66**

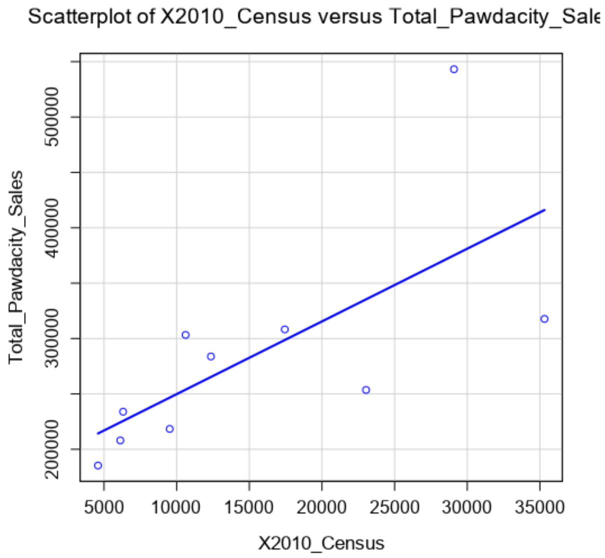


CITY	NORMAL	OUTLIER
Cheyenne	**2**	**4**
Gillette	**5**	**1**
Rock Springs	**5**	**1**
Buffalo	6	0
Casper	6	0
Cody	6	0
Douglas	6	0
Evanston	6	0
Powell	6	0
Riverton	6	0
Sheridan	6	0

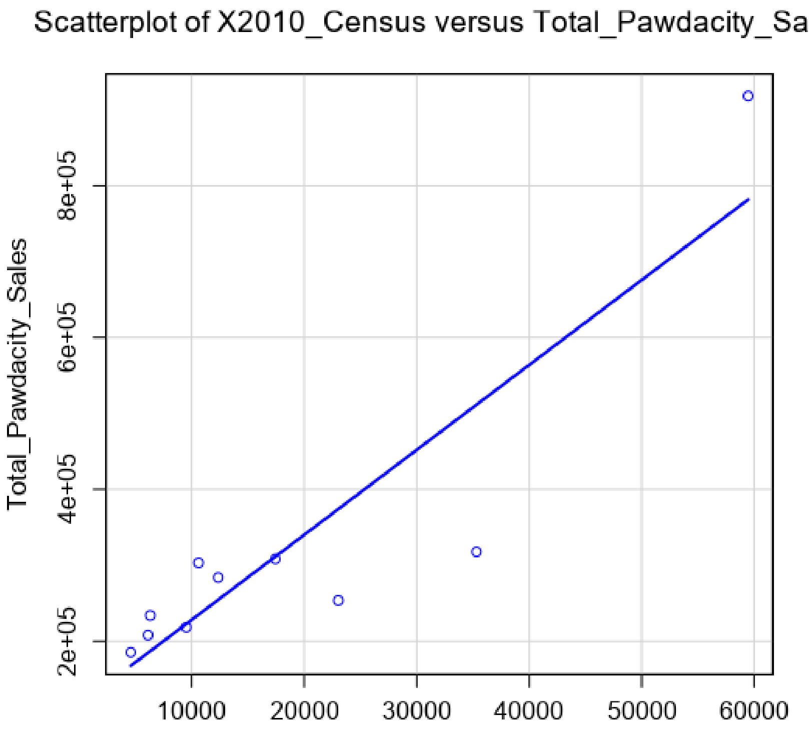
as we can see there are 3 cities where the outliers are showing up , so the idea is to drop one city at a time them build a regression equation with sales as target and compare all three models scatter plot t find which city to drop as per initial observation by the table city Cheyenne should drop because it offers only two normal records were the outlier contribution are 4 records (>50%).



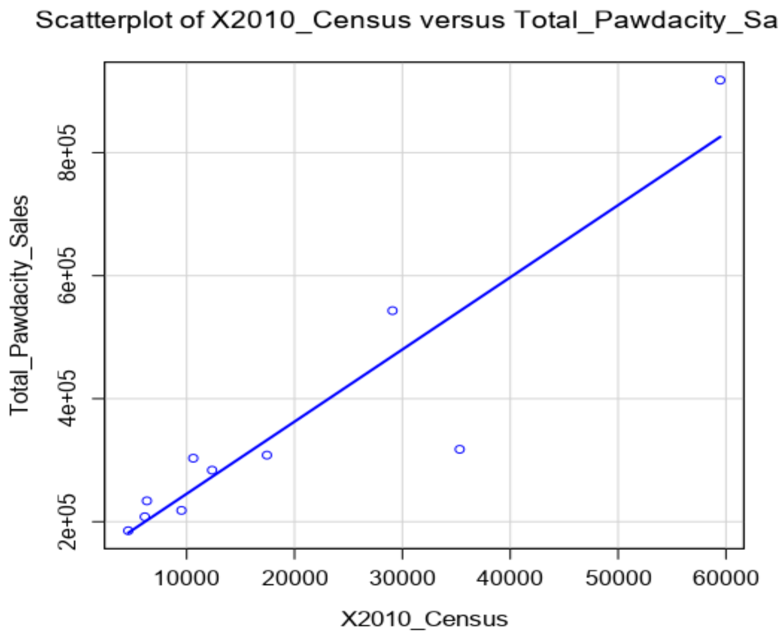
This scatter plot with Regression Line shows all the points where x-axis is 2010_census and y-axis is sales (With Outliers)



This scatter plot with Regression Line shows all the points where x-axis is 2010_census and y-axis is sales (Without Cheyenne)



This scatter plot with Regression Line shows all the points where x-axis is 2010_census and y-axis is sales (Without Gillette)



This scatter plot with Regression Line shows all the points where x-axis is 2010_census and y-axis is sales (Without rocks spring)

Outlier Analysis Insights

1

CITY	NORMAL	OUTLIER
Cheyenne	2	4
Gillette	5	1
Rock Springs	5	1

There are three cities with outlier records identified using box plot analysis and IQR method

2 By analysing the above table , we observer that city Cheyenne is more likely to remove because it has more outliers and less normal records than any other city.

3 after making scatter plot with regression line we can conclude that our initial observation to drop city Cheyenne is right as displayed in point 3 of outlier analysis. We can observe that slop has significantly dropped where in all other plot there is no significantly observable change of slop.

Imputation of outlier will not help in this scenario as the outlier are way to out from the normal records.