Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Pawdacity, the leading pet store chain in Wyoming is looking to expand by opening a new store. Pawdacity currently has 13 stores across Wyoming. This purpose of this analysis is to recommend potential city for new store based on available data such as sales, demographic info etc.

Key Decisions:

Answer these questions

1. What decisions needs to be made?
   Of the data available, identify dataset (and data points) that will be useful for prediction. Identify location for 14th store.

2. What data is needed to inform those decisions?
   Following datasets will be needed –
   - p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales.csv
   - p2-wy-demographic-data.csv
   - p2-partially-parsed-wy-web-scrape.csv

Step 2: Building the Training Set

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**	**Sum from my Alteryx**	**Average from my Alteryx**
Census Population	*213,862*	213,862	19,442
Total Pawdacity Sales	*3,773,304*	3,773,304	343,028
Households with Under 18	*34,064*	34,064	3,097
Land Area	*33,071*	33,071	3,006
Population Density	*63*	63	6
Total Families	*62,653*	62,653	5,696

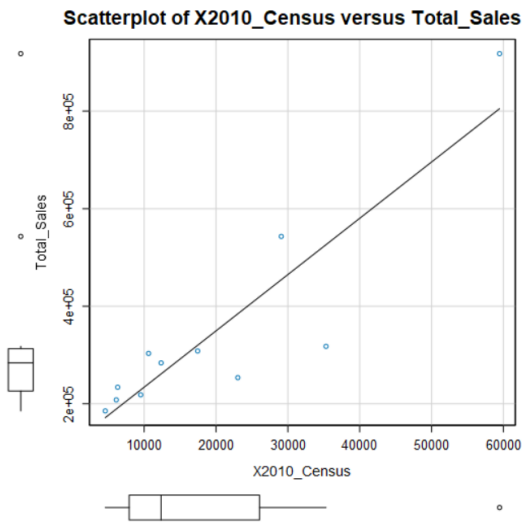
Step 3: Dealing with Outliers

Answer these questions

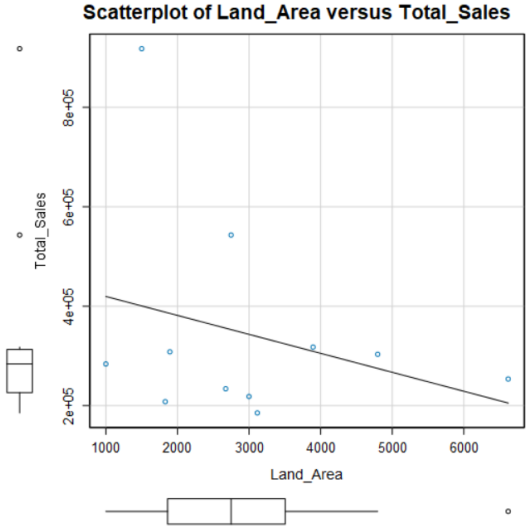
Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.

Scatterplots:

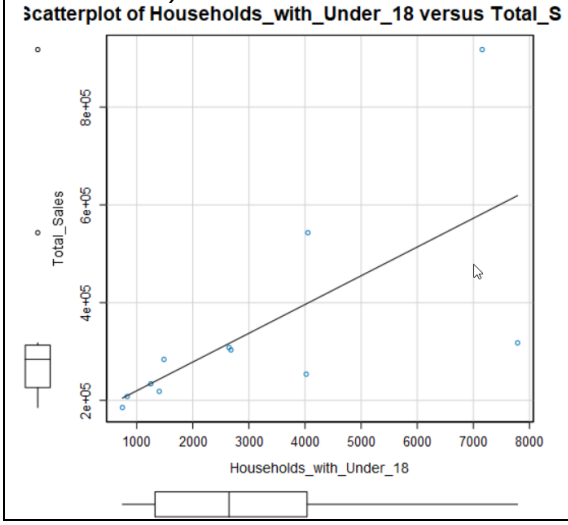
1) Census v/s Total Sales



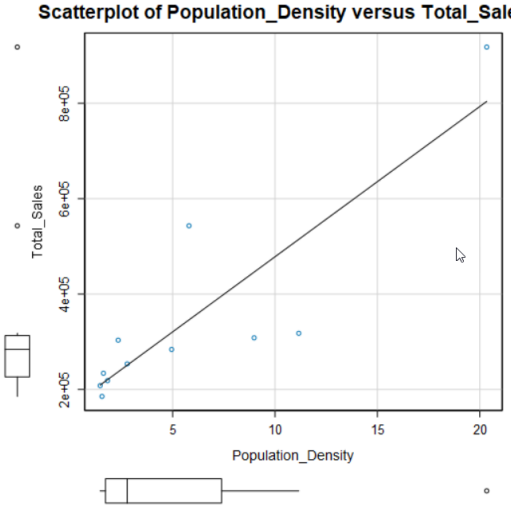
2) Land_Area v/s Sales[1]



3) Household v/s Total Sales



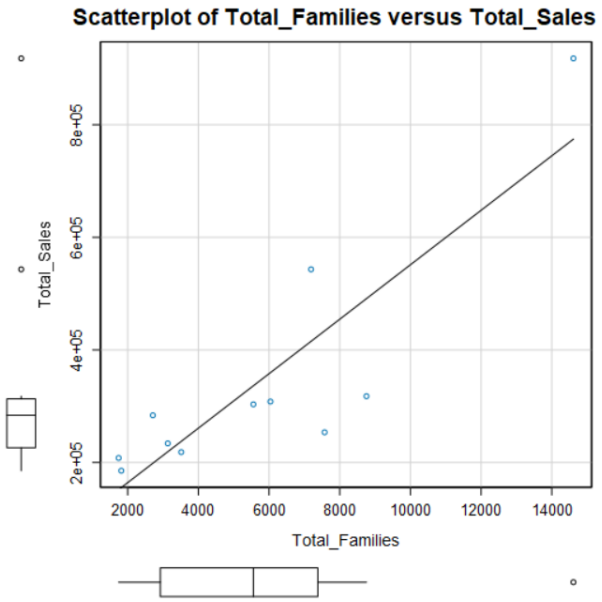
4) Population Density v/s Total Sales[2]



[1] Land Area has negative correlation with sales. The data indicates that increase in land area decreases sales.

[2] City of Cheyenne has significantly high population density.

5) Total Families v/s Total Sales



The Table below shows comparison of values with Upper fence for each variable. Data blending and IQR related calculations were performed in Alteryx. The values exceeding Upper fence are highlighted in red indicating potential outliers in the given category. E.g. city of Cheyenne has potential outliers in Total Pawdacity Sales, 2010 Census, Population Density and Total families.

Upper Fence	443,232.00	53,278.25	5,969.69	8,102.00	15.90	14,066.90
CITY	**Total_Sales (Pawdacity Sales)**	**2010 Census**	**Land Area**	**Households with Under 18**	**Population Density**	**Total Families**
Buffalo	185,328.00	4,585.00	3,115.51	746.00	1.55	1,819.50
Casper	317,736.00	35,316.00	3,894.31	7,788.00	11.16	8,756.32
Cheyenne	917,892.00	59,466.00	1,500.18	7,158.00	20.34	14,612.64
Cody	218,376.00	9,520.00	2,998.96	1,403.00	1.82	3,515.62
Douglas	208,008.00	6,120.00	1,829.47	832.00	1.46	1,744.08
Evanston	283,824.00	12,359.00	999.50	1,486.00	4.95	2,712.64
Gillette	543,132.00	29,087.00	2,748.85	4,052.00	5.80	7,189.43
Powell	233,928.00	6,314.00	2,673.57	1,251.00	1.62	3,134.18

Riverton	303,264.00	10,615.00	4,796.86	2,680.00	2.34	5,556.49
Rock Springs	253,584.00	23,036.00	6,620.20	4,022.00	2.78	7,572.18
Sheridan	308,232.00	17,444.00	1,893.98	2,646.00	8.98	6,039.71

Outliers analysis using IQR

1. **Cheyenne** – Outlier when it comes to Total Sales, population (2010 Census), population density and total families. However, higher total sales can be attributed to the higher population and comparatively high number of families in the city. Additionally, population density for Cheyenne is significantly high (20.34) compared to upper fence and average population density. We can consider potentially imputing population density of Cheyenne with average population density.

2. **Gillette** – Outlier when it comes to Total Sales, which can be justified by higher population.

3. **Rock Springs –** Outlier when it comes to Land Area.

Additional justification post annotations provided by reviewer:

As explained above, city of Cheyenne, Gillette and Rock Springs are clear outliers based on IQR analysis. The dataset to be analyzed has only 11 rows, therefore removing 3 may not provide enough training data for predicting new 14th store location. Gillette and Rock Springs are outliers in one field therefore these outliers should not be removed from the dataset. Gillette's higher sales can be justified by higher population of the city. For city of Rock Springs is outlier when it comes to Land Area which suggest negative correlation therefore, we probably will not use this Land Area for prediction in future.

For city of Cheyenne, even though its outlier when it comes to sales, this behavior can be attributed to higher population and higher number of families in the Cheyenne. The population density of Cheyenne is significantly high which is well above upper fence as well as average of 5.71. Imputing will still maintain positive and analogous correlation.