Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (250 word limit)

Key Decisions:

Answer these questions

1. What decisions needs to be made?
   Determine the Wyoming city where the new store for Pawdacity will be open based on the higher yearly sales.

2. What data is needed to inform those decisions?
   Yearly sales by city, population, population density, land area, total families can be used to predict yearly sales.

Step 2: Building the Training Set

Build your training set given the data provided to you. Your column sums of your dataset should match the sums in the table below.

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	Average
Census Population	*213,862*	*19,442.00*
Total Pawdacity Sales	*3,773,304*	*34,3027.64*
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

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.

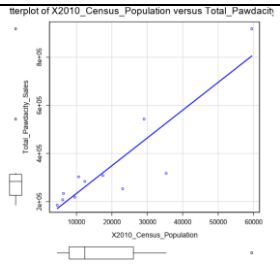
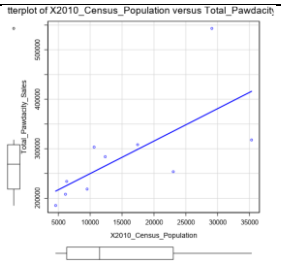
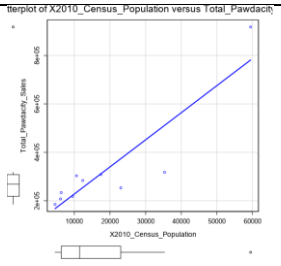
It has been removed the record for Gillette city, because it skews high in sales and potentially will affect the sales predictions, although not skew relative to the other data fields in the training set per the *Table 1 IQR analysis* shown below. And the record for Cheyenne city has been kept

because per the *table 2 for Scatter plots analysis shown* below it is in line with the linear relationship and removing it will affect the sales predictions considerably.

City	Total Pawdacity Sales	2010 Census Population	Households with Under 18	Land Area	Population Density	Total Families
Buffalo	185,328.0	4,585.0	746.0	3,115.5	1.6	1,819.5
Casper	317,736.0	35,316.0	7,788.0	3,894.3	11.2	8,756.3
Cheyenne	917,892.0	59,466.0	7,158.0	1,500.2	20.3	14,612.6
Cody	218,376.0	9,520.0	1,403.0	2,999.0	1.8	3,515.6
Douglas	208,008.0	6,120.0	832.0	1,829.5	1.5	1,744.1
Evanston	283,824.0	12,359.0	1,486.0	999.5	5.0	2,712.6
Gillette	543,132.0	29,087.0	4,052.0	2,748.9	5.8	7,189.4
Powell	233,928.0	6,314.0	1,251.0	2,673.6	1.6	3,134.2
Riverton	303,264.0	10,615.0	2,680.0	4,796.9	2.3	5,556.5
Rock Springs	253,584.0	23,036.0	4,022.0	6,620.2	2.8	7,572.2
Sheridan	308,232.0	17,444.0	2,646.0	1,894.0	9.0	6,039.7
QUARTILE	VALUE	VALUE	VALUE	VALUE	VALUE	VALUE
1	226,152.0	7,917.0	1,327.0	1,861.7	1.7	2,923.4
3	312,984.0	26,061.5	4,037.0	3,504.9	7.4	7,380.8
SUM	3,773,304.0	213,862.0	34,064.0	33,071.4	62.8	62,652.8
MEDIAN	283,824.0	12,359.0	2,646.0	2,748.9	2.8	5,556.5
IQR	86,832.0	18,144.5	2,710.0	1,643.2	5.7	4,457.4
UPPER FENCE	**443,232.0**	**53,278.3**	**8,102.0**	**5,969.7**	**15.9**	**14,066.9**
LOWER FENCE	**95,904.0**	**(19,299.8)**	**(2,738.0)**	**(603.1)**	**(6.8)**	**(3,762.7)**
QUARTILE	VALUE	VALUE	VALUE	VALUE	VALUE	VALUE
1	218,376.0	6,314.0	1,251.0	1,829.5	1.6	2,712.6
3	317,736.0	29,087.0	4,052.0	3,894.3	9.0	7,572.2
IQR	99,360.0	22,773.0	2,801.0	2,064.8	7.4	4,859.5
UPPER FENCE	466,776.0	63,246.5	8,253.5	6,991.6	20.0	14,861.5
LOWER FENCE	69,336.0	(27,845.5)	(2,950.5)	(1,267.8)	(9.4)	(4,576.7)
QUARTILE	VALUE	VALUE	VALUE	VALUE	VALUE	VALUE
1	226,152.0	7,917.0	1,327.0	1,861.7	1.7	2,923.4
3	312,984.0	26,061.5	4,037.0	3,504.9	7.4	7,380.8
IQR	86,832.0	18,144.5	2,710.0	1,643.2	5.7	4,457.4
UPPER FENCE	443,232.0	53,278.3	8,102.0	5,969.7	15.9	14,066.9
LOWER FENCE	95,904.0	(19,299.8)	(2,738.0)	(603.1)	(6.8)	(3,762.7)

Table 1. IQR analysis

Also, it was analyzed the impact of removing the outlier record with scatter plot for every predictor variable and target variable, and the slope was not affected considerably.

	With outlier	Without Cheyenne outlier	Without Gillette outlier
Census population			

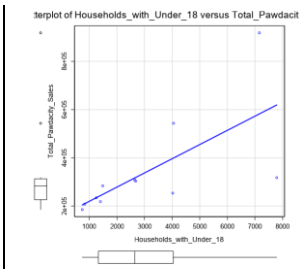
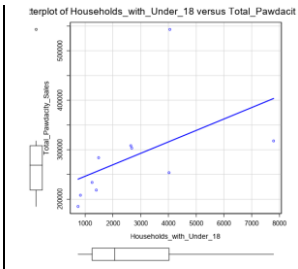
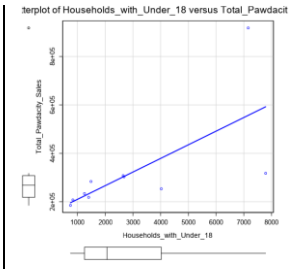
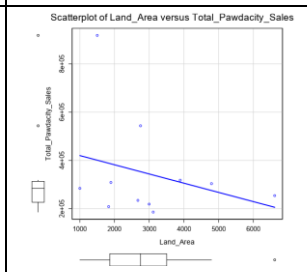
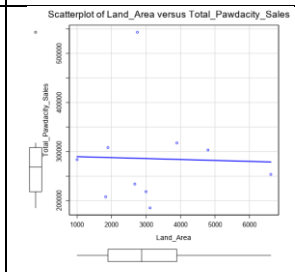
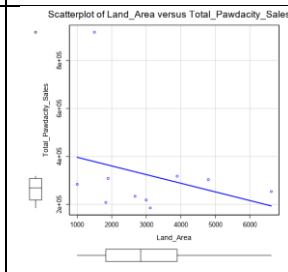
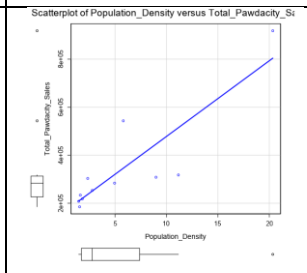
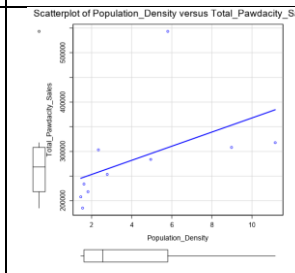
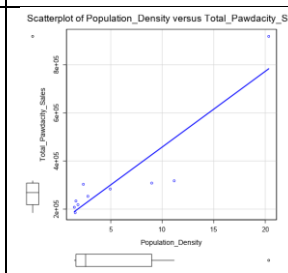
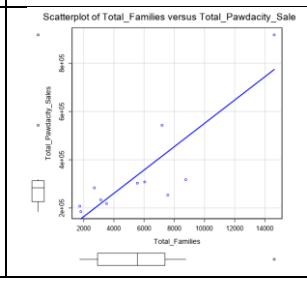
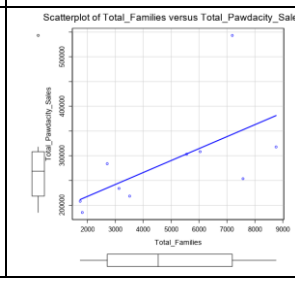
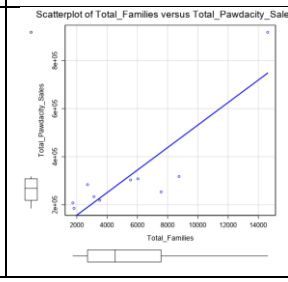
Households with Under 18			
Land Area			
Population Density			
Total Families			

Table 2. Scatter plots analysis

Alteryx solution