Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Recommend a city for Pawdacity for store expansion based on predicted yearly sales.

1. What decisions needs to be made?

   Need to work what data from the supplied files (sales from other Pawdacity stores, local population and Wyoming demographic data) will be needed to predict the next city to open a store in.

2. What data is needed to inform those decisions?

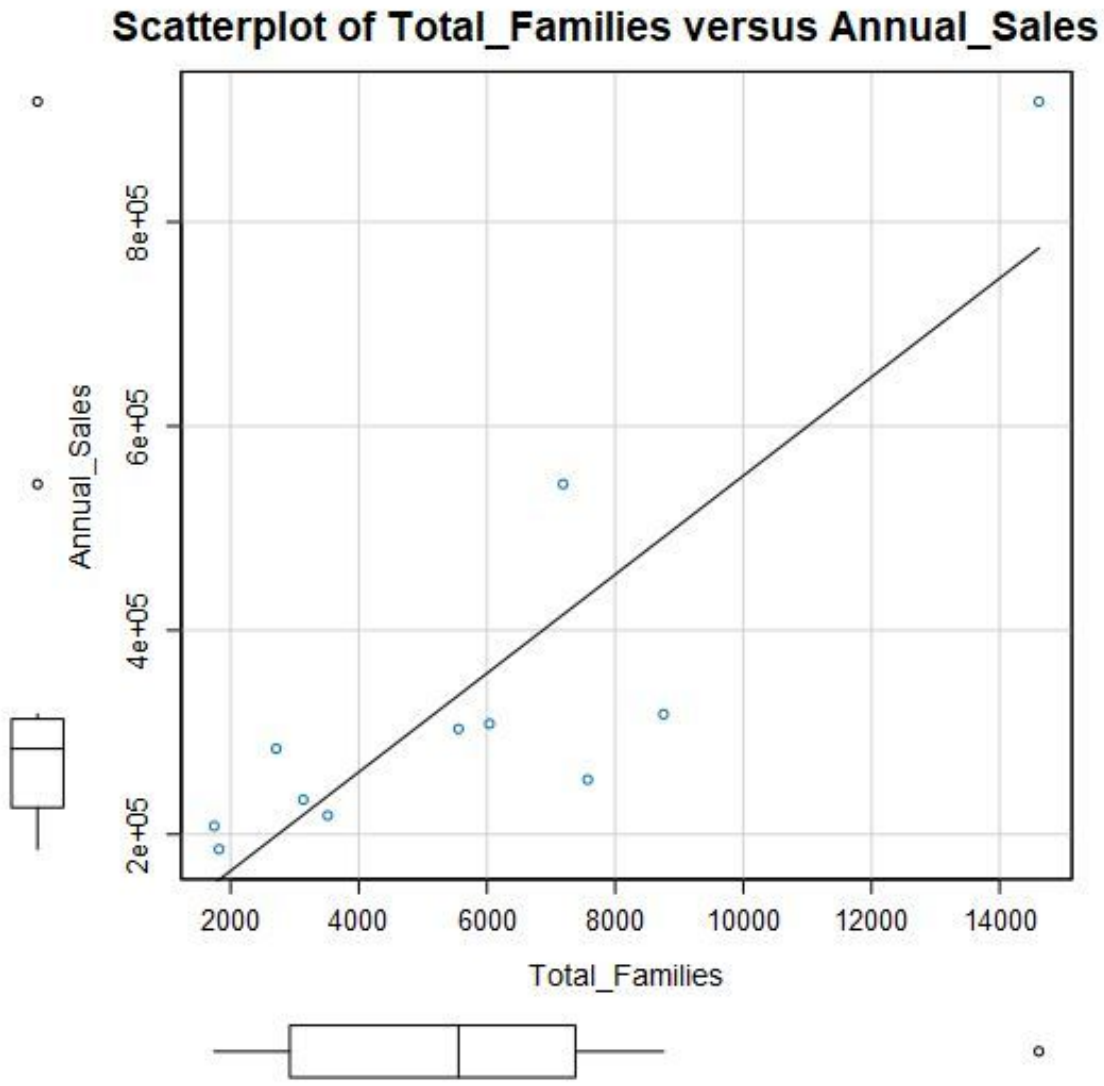
   We need to find the following data:

   City
   Census Population
   Total Pawdacity Sales
   Households under 18
   Land Area
   Population density
   Total Families

Step 2: Building the Training Set

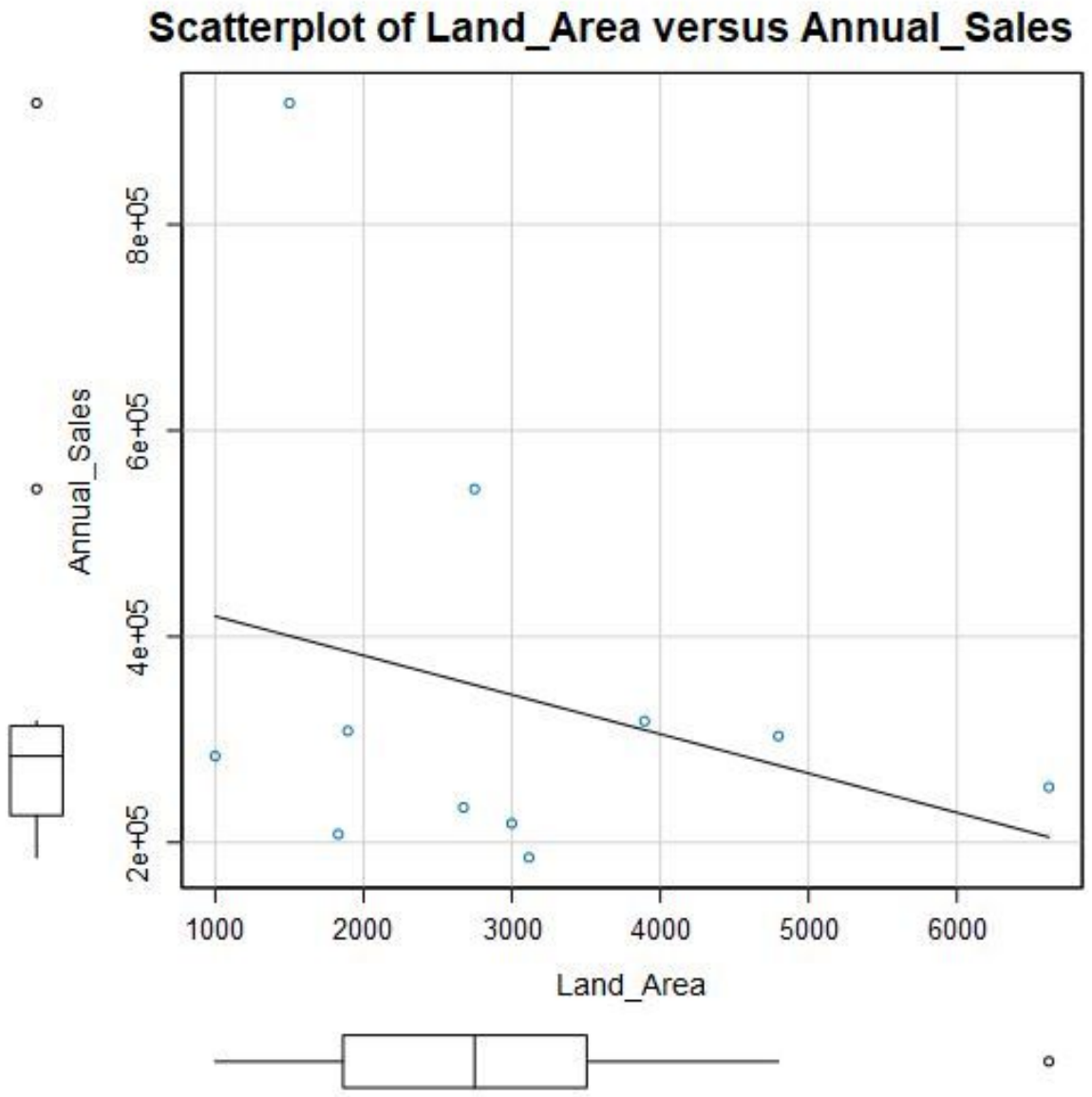
Column	Sum	Average
Census Population	*213,862*	*19,442*
Total Pawdacity Sales	*3,773,304*	*343,027.64*
Households with Under 18	*34,064*	*3,0976.73*
Land Area	*33,071*	*3006.49*
Population Density	*63*	*5.71*
Total Families	*62,653*	*5695.71*

Step 3: Dealing with Outliers



Q1_Total Families	Q3_Total Families	IQR_families	upper_fence_families	lower_fence_families
2923.41	7380.81	4457.40	14066.90	-3762.68

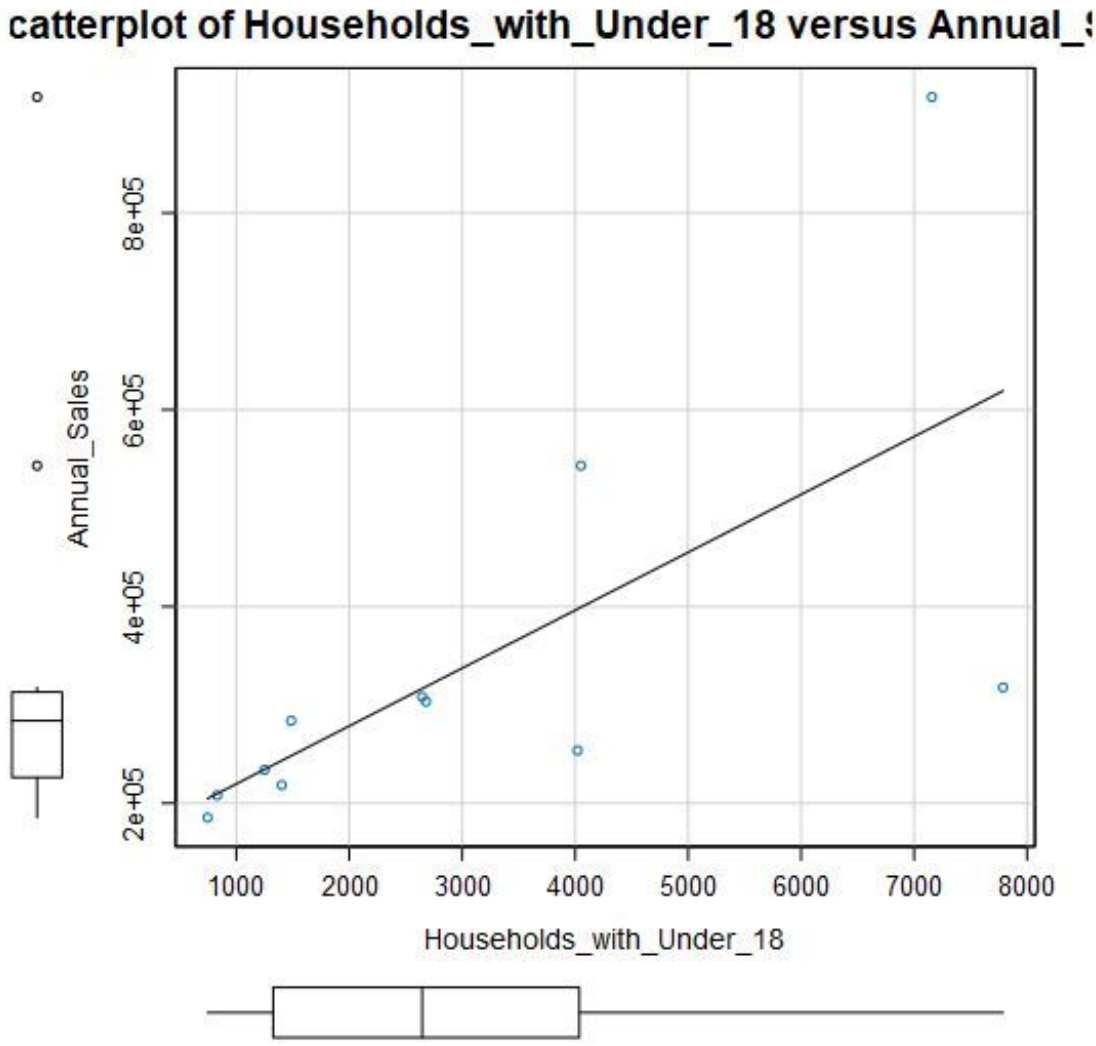
Cheyenne is above the upper quartile (at 14612.64). The plot is only just above the IQR Q3 and does follow the trend line upwards but is quite higher than the trend line so should be classified as an outlier. Removing it will also lower angle of the trend line.



Q1_Land Area	Q3_Land Area	IQR_land_area	upper_fence_land_area	lower_fence_land_area
1861.72	3504.91	1643.19	5969.68	-603.05

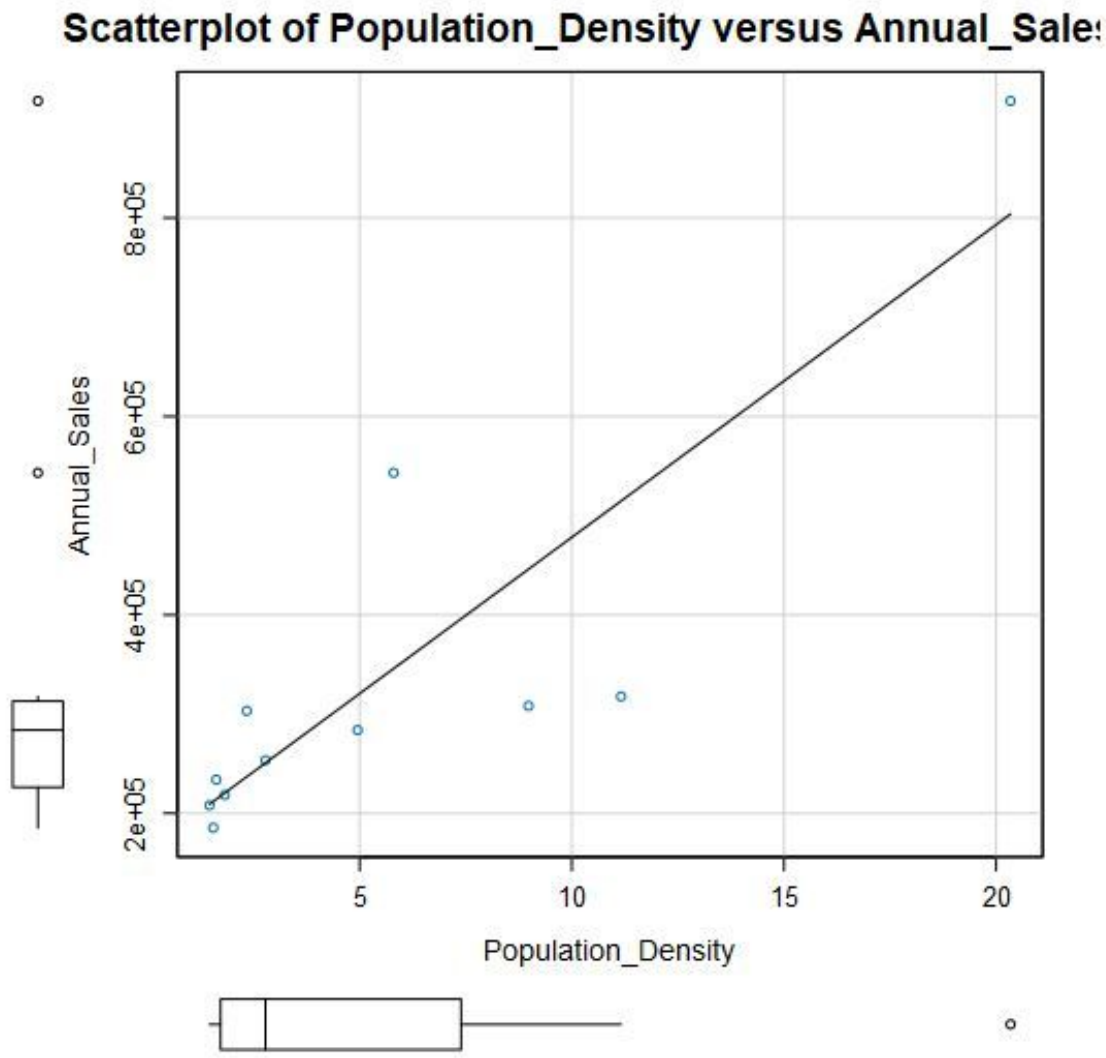
Rock Springs is above the upper quartile range (at 6620.02) outside of the upper fence. The plot follows the downward trend so should not be classified as an outlier.

Cheyenne however is the plot outlier here, with 1500 land area and 917892 in sales.



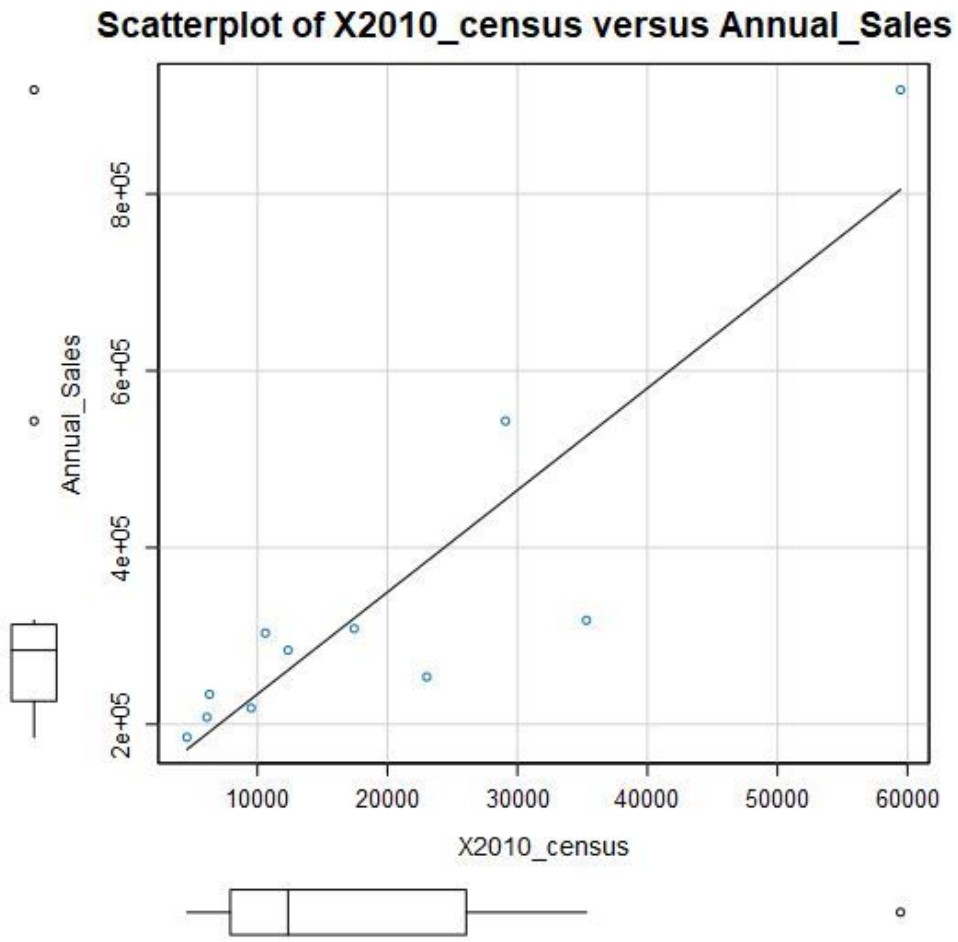
Q1_Households with Under 18	Q3_Households with Under 18	IQR_household_18	upper_fence_household	lower_fence_household
1327	4037	2710	8102	-2738

From a IQR Q3 point of view there are no outliers. However there is one plot way above the trend line to the right. That is Cheyenne, it is the outlier of this graph.



Q1_Population Density	Q3_Population Density	IQR_density	upper_fence_density	lower_fence_density
1.72	7.39	5.67	15.90	-6.79

Cheyenne has a population density of 20.34. It's above the trend line and the IQR Q3 so is an outlier here.



Q1_2010 census	Q3_2010 census	IQR_census	Upper_fence_census	lower_fence_census
7917	26061.5	18144.5	53278.25	-9072.25

Cheyenne has a value of 59466. It's above the trend line and the IQR Q3 so is an outlier here.

Q1_Annual Sales	Q3_Annual Sales	IQR_sales	upper_fence_sales	lower_fence_sales
226152	312984	86832	443232	95904

Cheyenne has sales of 917892 and Gillette has sales of 543132, both above the IQR Q3. So both are outliers here.

After looking at the scatter graphs of annual sales plotted against possible predictor variables one city has repeated outlier data. That is Cheyenne.

My recommendation would be to remove Cheyenne as it is an outlier on every plot. However I would be reluctant as we only have a small data set. We could possibly impute the data on Cheyenne further.