Wilson

Select New Pet Store Location

1. What decisions needs to be made?
   a. What city should pawdacity open its next store to maximize sales. Based on what information and data they have? What data is needed?

2. What data is needed to inform those decisions?
   b. City
   2010 Census Population
   Total Pawdacity Sales
   Households with Under 18
   Land Area
   Population Density
   Total Families

Data wrangled:

1. Pawdacity monthly sales
2. Census scraped data from Wyoming Wikipedia
3. Wyoming Demographic Data

Pawdacity monthly sales:To get Total pawdacity sales, I used Multi-Row Formula  to create a column that sums up all monthly sales for each Pawdacity store. Next I SUM total sales and grouped by City to get Total Sales from each city.

Census scraped data: In order to get 'City' I split City|County from each other using Text To Column  with a | delimiter. After renaming them it becomes a column with City and County. Next I cleaned 2010 Census Population by removing non-numeric characters like <td> and </td> using the Formula  and Data cleansing tool 

Here's the data before cleaning

City|County	2014 Estimate	2010 Census	2000 Census
Afton|Lincoln	<td>1,968</td>	<td>1,911</td>	<td>1,818</td>
Albin|Laramie	<td>185</td>	<td>181</td>	<td>120</td>
Alpine|Lincoln	<td>845</td>	<td>828</td>	<td>550</td>
Baggs|Carbon	<td>439</td>	<td>440</td>	<td>348</td>
Bairoil|Sweetwater	<td>107</td>	<td>106</td>	<td>97</td>
Bar Nunn|Natrona	<td>2,735</td>	<td>2,213</td>	<td>936</td>

Here's after cleaning

2014 Estimate	2010 Census	2000 Census	City	County
<td>1,968</td>	1911	<td>1,818</td>	Afton	Lincoln
<td>185</td>	181	<td>120</td>	Albin	Laramie
<td>845</td>	828	<td>550</td>	Alpine	Lincoln
<td>439</td>	440	<td>348</td>	Baggs	Carbon
<td>107</td>	106	<td>97</td>	Bairoil	Sweetwater
<td>2,735</td>	2213	<td>936</td>	Bar Nunn	Natrona

Demographic Data contains

1. Households under 18
2. Land Area
3. Population Density
4. Total Families

These 4 columns doesn't need any cleaning.

I used the Join Tool  to merge Total Pawdacity sales from the monthly sales data with City column from Census scraped data using 'City' as the common identifier.

Then I use Join Tool  again merging the initial Joined Data with Demographic Data with City as the common identifier.

Data is cleaned and joined properly with all the proper columns needed to find our next Pawdacity store.

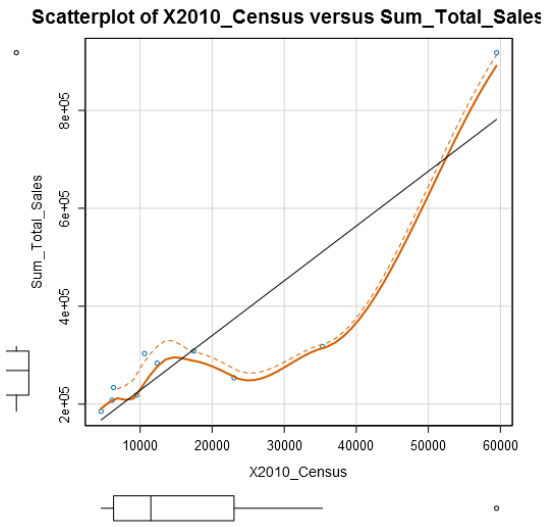
CITY	Sum_Total Sales	2010 Census	Land Area	Households with Under 18	Population Density	Total Families
Buffalo	185328	4585	3115.5075	746	1.55	1819.5
Casper	317736	35316	3894.3091	7788	11.16	8756.32
Cheyenne	917892	59466	1500.1784	7158	20.34	14612.64
Cody	218376	9520	2998.95696	1403	1.82	3515.62

1. 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.

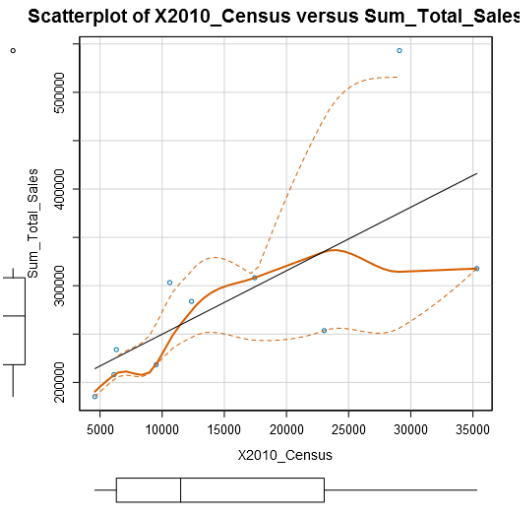
Yes…Cheyenne has 4 outliers, Gillette has 1 and Rock Springs has 1. I've chosen to delete city 'Gillette' from the dataset because its total sales is an outlier. Even though Cheyenne has 4 potential outliers, its total sales is justified by a larger census population, population density, and total family. Rock Springs land area is an outlier but its total sales, census population, population density and total families are within norms. I chose to delete it over imputing because it's easier to work with. I made sure it didn't mess with other data by testing

scatterplot models with and without Gillette in the dataset. Without Gillette linear regression between Sum Total Sales x 2010 Census improved. Without Cheyenne it worsened.

Without Gillette

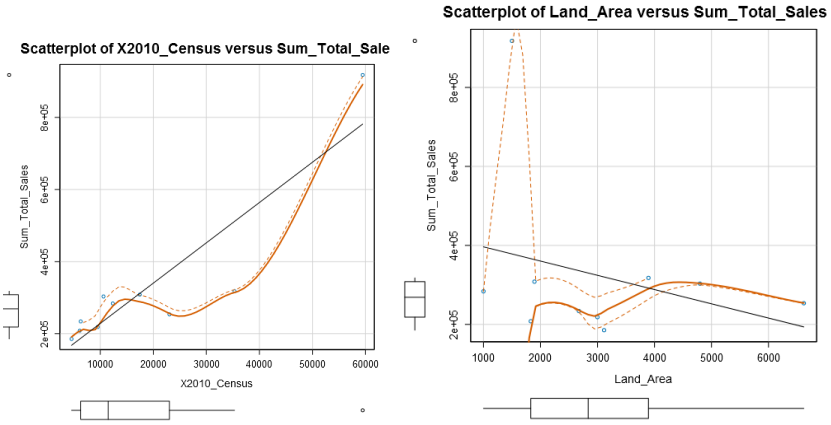


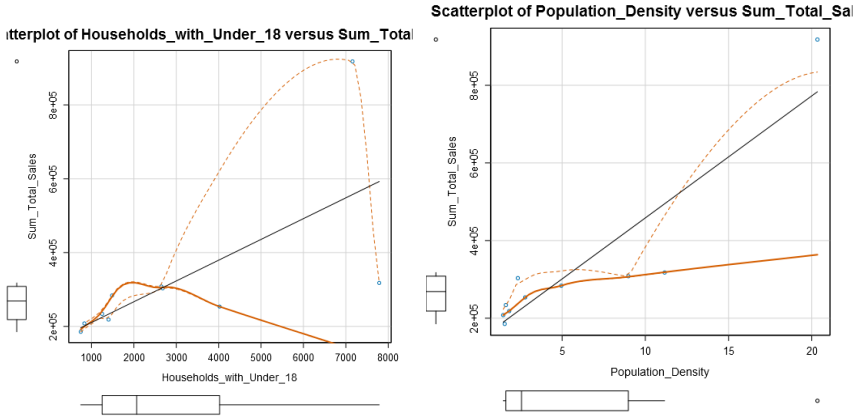
Without Cheyenne

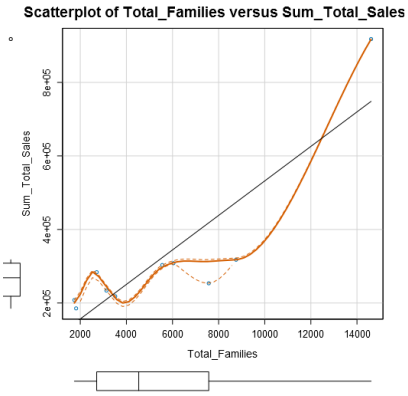


In order to select our new location pet store, we need to perform Linear Regression we need our Cleaned data and demographic data.

I built scatterplot for each predictor variable.







Predictor variables are quality variables since they have a linear relationship with target variable.

I used Association Analysis to see the correlation between each predictor variable as shown below.

Full Correlation Matrix

	Sum_Total.Sales	X2010.Census	Land.Area	Households.with.Under.18	Population.Density	Total.Families
Sum_Total.Sales	1.00000	0.89875	-0.28708	0.67465	0.90618	0.87466
X2010.Census	0.89875	1.00000	-0.05247	0.91156	0.94439	0.96919
Land.Area	-0.28708	-0.05247	1.00000	0.18938	-0.31742	0.10730
Households.with.Under.18	0.67465	0.91156	0.18938	1.00000	0.82199	0.90566
Population.Density	0.90618	0.94439	-0.31742	0.82199	1.00000	0.89168
Total.Families	0.87466	0.96919	0.10730	0.90566	0.89168	1.00000

2010 census, household under18, population density and total families highly correlation but land area isn't.

Using land area as a predictor variable to test with others, I see that land area and total families produced the best linear regression model.

Min	1Q	Median	3Q	Max
-121300	-4453	8418	40490	75200

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	197330.41	56449.000	3.496	0.01005	*
Land.Area	-48.42	14.184	-3.414	0.01123	*
Total.Families	49.14	6.055	8.115	8e-05	***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 72030 on 7 degrees of freedom
Multiple R-squared: 0.9118, Adjusted R-Squared: 0.8866
F-statistic: 36.2 on 2 and 7 DF, p-value: 0.0002035

Type II ANOVA Analysis

Response: Sum_Total.Sales

	Sum Sq	DF	F value	Pr(>F)	
Land.Area	60473052720.43	1	11.66	0.01123	*
Total.Families	341673845917.83	1	65.85	8e-05	***
Residuals	36318449406.44	7			

This is a great model because p-values for both variables are below 0.05, with Adj R-Squared .88 which is close to 1.

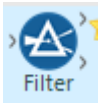
Linear regression equation: Y = 197,330 – 48.4*[Land Area] + 49.14*[Total Families]

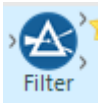
Here are the criteria's given to you in choosing the right city:

1. The new store should be located in a new city. That means there should be no existing stores in the new city.
2. The total sales for the entire competition in the new city should be less than $500,000
3. The new city where you want to build your new store must have a population over 4,000 people (based upon the 2014 US Census estimate).
4. The predicted yearly sales must be over $200,000.
5. The city chosen has the highest predicted sales from the predicted set.

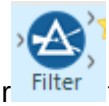
First I listed all the cities Pawdacity is doing business in per Criteria 1.

1. Buffalo
2. Casper
3. Cheyenne
4. Cody
5. Douglas
6. Evanston
7. Gillette
8. Powell
9. Riverton
10. Rock Springs
11. Sheridan

Next I took our competitors data, aggregated their yearly sales and grouped by cities. Then I filter 

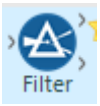
to show all cities with less than $500,000 sales per Criteria 2. I also filtered 

out competitor data with the same cities Pawdacity is already doing business in.

For 2014 Census we look at our Wyoming Census data earlier. We cleaned 2014 Estimates and Text to Column  City|County. Next we filter 

to show all 2014 Estimates Population > 4000 per Criteria 3.

Now I use Join Tool  to join the cities that fulfill our above criteria using competitor data and Wyoming census population data.

Here's our potential store locations after filtering 

Criteria 1, 2 and 3.

2014 Estimate	City	Sum_SALES VOLUME
10449	Jackson	182000
7642	Lander	152197
32081	Laramie	76000
5366	Worland	169000

Here’s the 4 stores with their predicted sales score.

City	County	Land.Area	Households.with.Under.18	Population.Density	Total.Families	Score
Laramie	Albany	2513.745235	2075	5.19	4668.93	305013.881671
Jackson	Teton	1757.6592	1078	2.36	2313.08	225870.8236
Lander	Fremont	3346.80934	1870	1.63	3876.81	225751.400203
Worland	Washakie	1294.105755	595	2.18	1364.32	201700.325919

The store that fulfills Criteria 4 and 5 is Laramie with $305,014 predicted sales.