Project: Predicting the Best Location for a New Store

Summary: A pet store chain's managers would like to add another store to the already existing 13 stores in Wyoming and they would like to do that based on the predicted yearly sales of their own branches. The business problem here is which city in Wyoming would be the best option for the new store to be located.

Step 1: Business and Data Understanding

A pet store chain Pawdacity managers would like to add another store to the already existing 13 stores in Wyoming and they would like to do that based on the predicted yearly sales of their own branches. To make the best prediction, I first predicted yearly sales for all the cities in Wyoming using the historical data and then based on this prediction choose the best performing city.

I created a regression equation to predict the sales and then apply the equation coefficients to the new dataset. I used the predictor variables to predict the sales of the potential stores. I have those predictor variables such as *2010 Census Population, Household with Under 18*, *Population Density, Land Area, and Total Families.* I also have the target variable *Total Pawdacity Sales.* Based on this dataset I will a regression equation and apply it to the new dataset.

Step 2: Building the Training Set

After cleaning and blending the data, I came up with 11 records. I also added to the table the average values I found.

Column	**Sum**	**Average**
Census Population	*213,862*	*19442*
Total Pawdacity Sales	*3,773,304*	*343027.63*
Households with Under 18	*34,064*	*3096.72*
Land Area	*33,071*	*3006.48*
Population Density	*63*	*5.70*
Total Families	*62,653*	*5695.70*

Step 3: Data preparation and Dealing with Outliers.

I started with the Web Scraped Data from the Wyoming Wikipedia page, and used text to columns and select tools and the Data Cleansing to parse out the City, County, 2010 Census, and 2014 Estimate and remove all the extra punctuation.
For the demographic data, I used the Auto-field tool to combine all the numbers labeled as String fields.

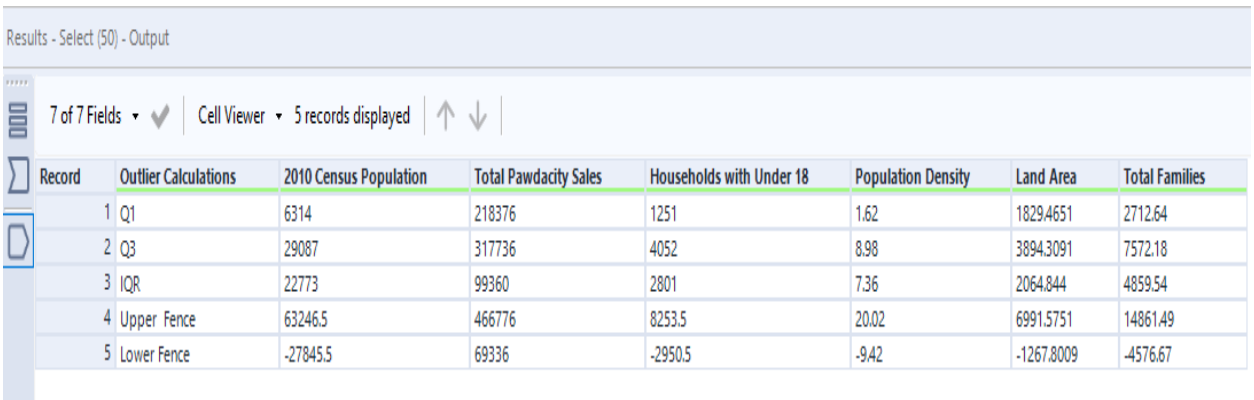
Before each join, I summarized the amounts by city to ensure that there were no duplicate city names within the data.
For Pawdacity sales file, I transposed the data to get City, Month, and Amount, and then summarized by City to get the total amount for each city.
From there, I created my data set used to train my regression model.

To identify the outliers, I used the IQR method. I calculated the 1st quartile Q1 and 3rd quartile Q3 using QUARTILE.EXC function in excel. Then I calculated the Interquartile Range for each record: IQR = Q3 - Q1. Then I added the (1.5 & IQR) with Q3 to get the upper fence. I subtract product (1.5 & IQR) from Q1 to get the lower fence.

Below are findings regarding outliers.

- *2010 Census Population* variable does not have any outlier. Upper Fence: 63246.5.
- *Total Pawdacity Sales* variable seems to have 2 outliers as the upper fence value for this variable is 466766 and there are two cities exceeding this sale threshold. Cheyenne being 917892 and Gillette being 543132.
- *Household with Under 18* variable does not have any outlier. Upper Fence: 8253.5.
- *Population Density* variable has one outlier. The upper fence value for this variable is 20.02 and the city Cheyenne seems to be an outlier with the value of 20.34.
- *Land Area* variable does not have any outlier. Upper Fence: 6991.57.
- *Total Families* variable does not have any outlier. Upper fence: 14861.49.

I prefer to remove the city of Cheyenne as it has an outlier in both Pawdacity sales (which is the target variable) and in population density which is a potential predictor variable.
Cheyenne's sales value is extremely higher than the upper fence value and it is likely to significantly bias the regression line.
Regression analysis is extremely sensitive to such extreme outliers. Therefore, it needs to be removed from the dataset.
I would not prefer to remove Gillette even though its sales value also seems to be an outlier because I do not have so many records and this value is just slightly over the upper fence. Therefore, I would rather keep the city of Gillette for my regression analysis.

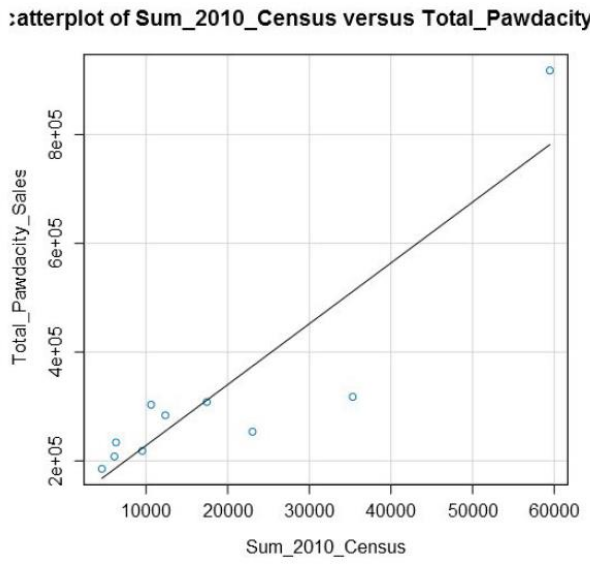

Results - Select (50) - Output

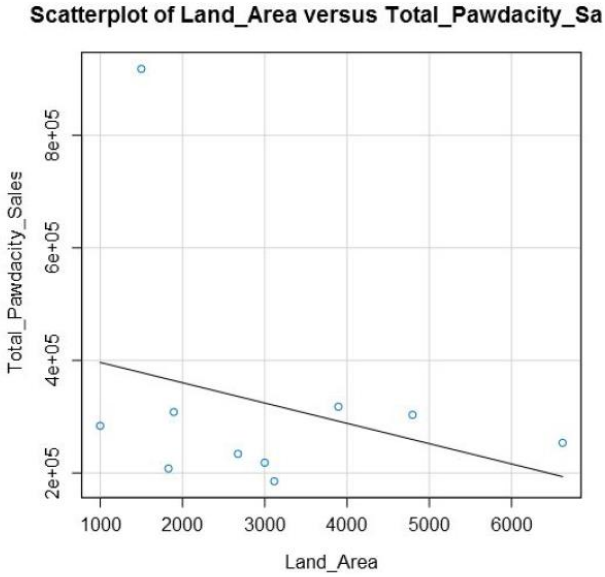
7 of 7 Fields ▾ | Cell Viewer ▾ 5 records displayed

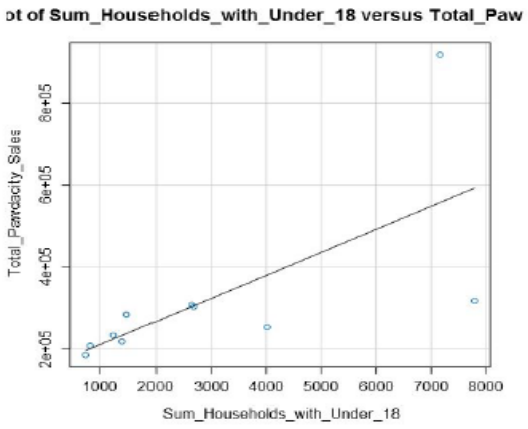
Record	Outlier Calculations	2010 Census Population	Total Pawdacity Sales	Households with Under 18	Population Density	Land Area	Total Families
1	Q1	6314	218376	1251	1.62	1829.4651	2712.64
2	Q3	29087	317736	4052	8.98	3894.3091	7572.18
3	IQR	22773	99360	2801	7.36	2064.844	4859.54
4	Upper Fence	63246.5	466776	8253.5	20.02	6991.5751	14861.49
5	Lower Fence	-27845.5	69336	-2950.5	-9.42	-1267.8009	-4576.67

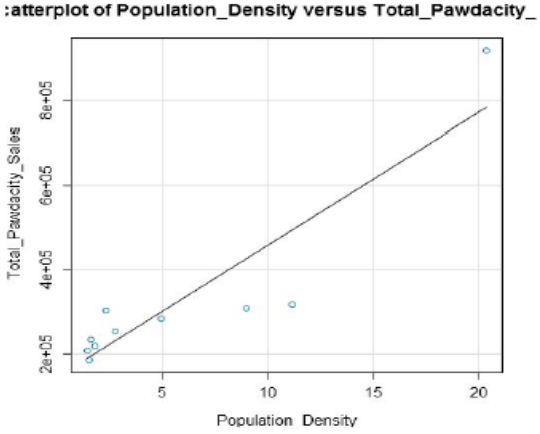
Step 4: Linear Regression

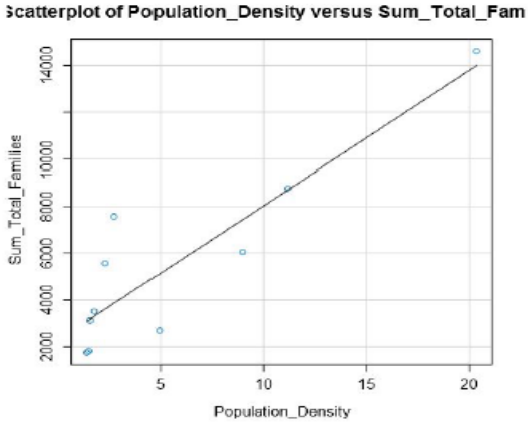
I first plotted each predictor variable against my target variable:











I can conclude all predictor variables are good potential predictor variables because they show a linear relationship between sales.

I checked for correlations between my predictor variables to see if there is any possibility of multicollinearity in my dataset. Below is a table that shows the correlations between the different predictor variables:

FieldName	Total Pawdacity Sales	Sum_2010 Census	Land Area	Sum_Households with Under 18	Population Density	Sum_Total Families
Total Pawdacity Sales	1.0000					
Sum_2010 Census	0.8988	1.0000				
Land Area	-0.2871	-0.0525	1.0000			
Sum_Households with Under 18	0.6747	0.9116	0.1894	1.0000		
Population Density	0.9062	0.9444	-0.3174	0.8220	1.0000	
Sum_Total Families	0.8747	0.9692	0.1073	0.9057	0.8917	1.0000

We can see that HHU18, Census, Families, and PDensity (Population Density) have strong correlations which each other. Land Area however, is not highly correlated. Therefore, I started by using land area as one predictor and then tested the four variables that are correlated. I found that using land area and total families as the predictor variables produced the best model.

Step 5: Analysis

Basic Summary

Call:
lm(formula = Total.Pawdacity.Sales ~ Land.Area + Sum_Total.Families, data = the.data)

Residuals:

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	*
Sum_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

As can be seen from the regression analysis below, the p-values for land area and total families are both below 0.05 and the Multiple R-squared value is at .91 which is close to 1. This model is a decent model.
The best linear regression equation based on the available data is shown be
low:

Y = 197,330 – 48.42 * [Land Area] + 49.14 * [Total Families]

Step 6: Conclusion and Recommendation

Once the model was created, by the means of Score Tool, I applied the model to the cities that were not already in the Pawdacity Sales file.

I then applied the filters laid out in the project plan to come up with my list of possible cities, and sorted on the expected revenue to bring the best choice to the top.

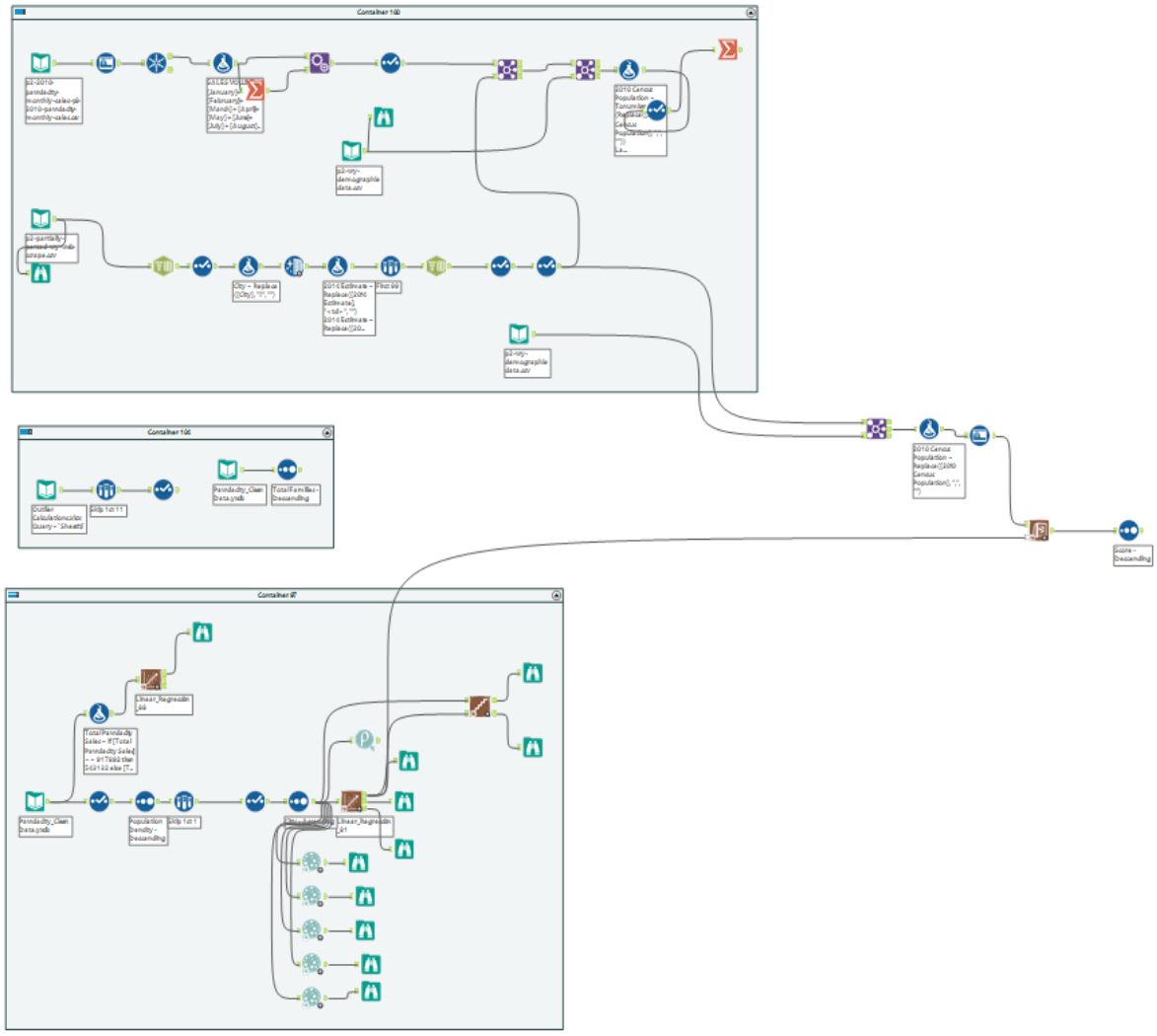
Results - Sort (111) - Output

8 of 8 Fields | Cell Viewer | 99 records displayed

Record	X2010.Census.Population	City	County	Households.with.Under.18	Land.Area	Population.Density	Total.Families	Score
1	59466	Cheyenne	Laramie	7158	1500.1784	20.34	14612.64	767670.913729
2	29087	Gillette	Campbell	4052	2748.8529	5.8	7189.43	498595.390053
3	30816	Laramie	Albany	2075	2513.745235	5.19	4668.93	390040.588115
4	35316	Casper	Natrona	7788	3894.3091	11.16	8756.32	326728.593752
5	3461	Mills	Natrona	563	281.40226	0.81	1138.55	324208.370352
6	17444	Sheridan	Sheridan	2646	1893.977048	8.98	6039.71	322919.774466
7	2544	Evansville	Natrona	432	215.894254	0.62	873.51	313132.805578
8	2213	Bar Nunn	Natrona	417	208.57322	0.6	843.88	310538.696931
9	1807	Wright	Campbell	386	262.008785	0.55	685.27	294625.223098
10	1129	Pine Bluffs	Laramie	227	47.611047	0.65	463.76	290154.859891
11	23036	Rock Springs	Sweetwater	4022	6620.201916	2.78	7572.18	289557.599522
12	12515	Green River	Sweetwater	2113	3477.361206	1.46	3977.4	287024.11092
13	404	Midwest	Natrona	63	31.419439	0.09	127.12	284317.677624
14	301	Burns	Laramie	60	12.671352	0.17	123.43	282208.999974
15	195	Edgerton	Natrona	30	15.175894	0.04	61.4	281783.415806
16	181	Albin	Laramie	37	7.685902	0.1	74.87	281107.005908
17	757	Dayton	Sheridan	141	100.916016	0.48	321.81	280458.724603
18	855	Ranchester	Sheridan	167	119.853656	0.57	382.2	280453.778144
19	336	Superior	Sweetwater	56	91.40807	0.04	104.55	279360.001415
20	106	Bairoil	Sweetwater	18	29.45983	0.01	33.7	279322.40889
21	142	Clearmont	Sheridan	25	18.047953	0.09	57.55	279301.148816
22	451	Wamsutter	Sweetwater	84	138.488732	0.06	158.4	279151.587279
23	139	Granger	Sweetwater	23	38.270246	0.02	43.77	279135.273157
24	157	Frannie	Park	32	67.600917	0.04	79.25	278675.054055

I filtered my cities according to the given the criteria in the project and found that the best option was the city of Laramie. Then using the predicted measures of Laramie in the regression equation I calculated the predicted revenue of the new store. In conclusion I would recommend the city of Laramie with a predicted sales of $305,014.

Below are the Alteryx workflows for the whole analyses:



PREPARED BY: KADIR AKYUZ