Predicting New Store Location

Part 1 – Cleaning the Data.

Business decisions.

Pawdacity is a leading pet store chain in Wyoming with 13 stores throughout the state. This year, Pawdacity would like to expand and open a 14th store. The aim of this project is to perform analysis to recommend the city for Pawdacity's newest store, based on predicted yearly sales.

What decisions need to be made?

There are three sets of data:

p2-2010-pawdacity-monthly-sales.csv,
p2-partially-parsed-wy-web-scrape.csv,
p2-wy-453910-naics-data.csv.

We need to work out what data from the above files will be necessary to predict where our next store should be.

What data is needed to inform those decisions?

We will need to extract the following columns of data from the above files:

City
2010 Census Population
Total Pawdacity Sales
Households with under 18
Land Area
Population Density
Total Families

The data from the above fields will later be used to create a prediction model for the new store location.

The Dataset.

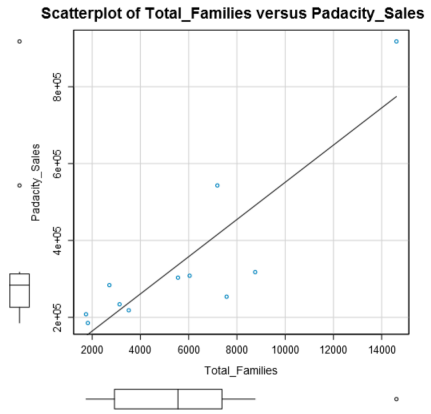
The below is a summary of the dataset.

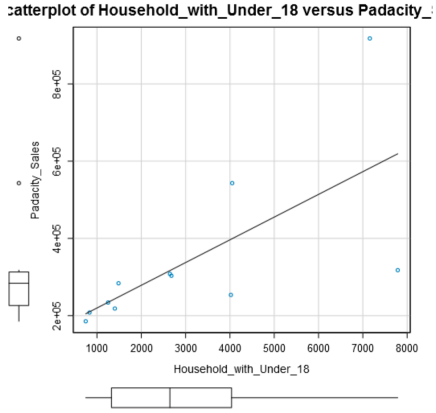
Column	Sum	Average
Census Population	213862	19442
Total Pawdacity Sales	3773304	343027.64
Households with Under 18	34064	3096.73
Land Area	33071	3006.49
Population Density	63	5.71

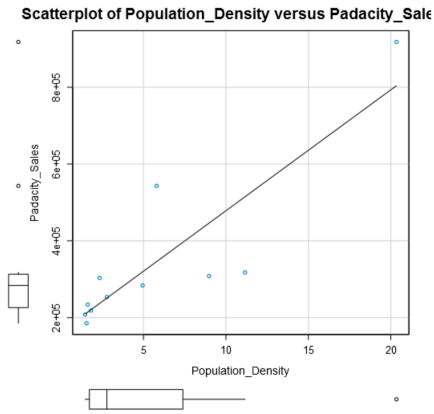
Total Families	62653	5695.71

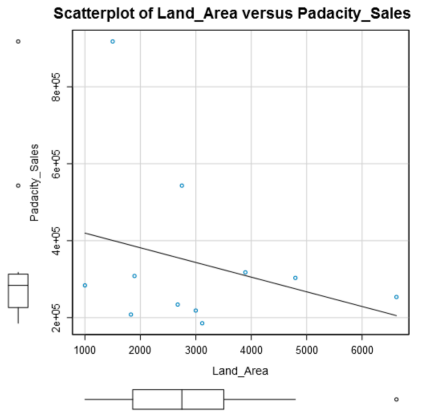
Outliers in the dataset.

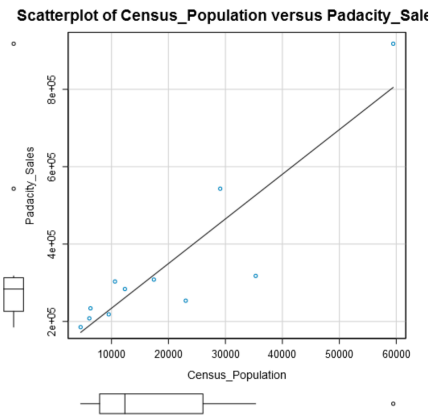
Below are scatter plots and boxplots of the dataset, with each potential predictor variable plotted against the Pawdacity Sales for that city.











Below is a summary of the dataset, with a further analysis of the interquartile ranges for the variables and their subsequent upper fence which for this project will be [1.5 * Interquartile Range] + 3rd Quartile.

I will look into values that are above the “Upper Fence” for each variable.

Name	Min	Max	Median	Mean	Std. Dev.
Census_Population	4585.00	59466.00	12359.00	19442.00	16616.02
Household_with_Under_18	746.00	7788.00	2646.00	3096.73	2453.00
Land_Area	999.50	6620.20	2748.85	3006.49	1617.46
Padacity_Sales	185328.00	917892.00	283824.00	343027.64	213538.71
Population_Density	1.46	20.34	2.78	5.71	5.85
Total_Families	1744.08	14612.64	5556.49	5695.71	3816.05

Census_Population_IQR	Padacity_Sales_IQR	Household_with_Under_18_IQR	Land_Area_IQR	Population_Density_IQR	Total_Families_IQR
18144.50	86832.00	2710.00	1643.19	5.67	4457.40
Census_Population_Upper_Fence	**Padacity_Sales_Upper_Fence**	**Household_with_Under_18_Upper_Fence**	**Land_Area_Upper_Fence**	**Population_Density_Upper_Fence**	**Total_Families_Upper_Fence**
53278.25	443232.00	8102.00	5969.69	15.90	14066.90

The list below indicates max points above that of their respective “Upper Fence”:

Census Population for Cheyenne
Land Area for Rock Springs
Population Density for Cheyenne
Total Families for Cheyenne
Pawdacity Sales for Gillette and Cheyenne

Below is a summary of the Pearson Correlation calculated from the predictor variables and the target variable which in this instance is Pawdacity Sales.

Pearson Correlation Analysis

Focused Analysis on Field Padacity_Sales

	Association Measure	p-value
Census_Population	0.89810	0.00017363***
Total_Families	0.86466	0.00059221***
Population_Density	0.86289	0.00062613***
Household_with_Under_18	0.67601	0.02239778*
Land_Area	-0.28890	0.38889983

Currently, the outliers I need to investigate are Cheyenne City for Census Population, Land Area, Population Density, Rock Springs for Land Area and Pawdacity sales for Gillette.

The scatterplot for Land Area vs Sales would indicate to me that Rock Springs follows the downward direction of the line of best fit for that plot with sales roughly inline with other sales values in that plot.

Cheyenne on the other hand has two stores and their data is aggregated in this analysis which could cause it to be an outlier, however since we are looking at where to place the new store, we should look at this data at a city level. This would mean that Cheyenne justifiably is s city that produces higher sales to warrant two stores.

Gillette also has two stores, however looking through the other categories Gillette's data looks relatively with in our outlier range except for its sales. There doesn't seem to be a good reason for this based on the small amount of information that I know.

My recommendation here would be to keep Cheyenne and Rock Springs as I believe their data looks to be appropriate. Gillette however is harder to explain and it would be best to remove this city totally from our data set, however I am reluctantly removing Gillette due to the fact we already have a small amount of data.

Creating the model.

Below is the final dataset used for the regression model.

City	Census_Population	Household_with_Under_18	Land_Area	Padacity_Sales	Population_Density	Total_Families
Buffalo	4585	746	3115.5075	185328	1.55	1819.5
Casper	35316	7788	3894.3091	317736	11.16	8756.32
Cheyenne	59466	7158	1500.1784	917892	20.34	14612.64
Cody	9520	1403	2998.95696	218376	1.82	3515.62
Douglas	6120	832	1829.4651	208008	1.46	1744.08
Evanston	12359	1486	999.4971	283824	4.95	2712.64
Powell	6314	1251	2673.57455	233928	1.62	3134.18
Riverton	10615	2680	4796.859815	303264	2.34	5556.49
Rock Springs	23036	4022	6620.201916	253584	2.78	7572.18
Sheridan	17444	2646	1893.977048	308232	8.98	6039.71

Selecting the predictor variables.

Below is a table of all the variables and their Pearson correlation.

Pearson Correlation Analysis

Focused Analysis on Field Padacity_Sales

	Association Measure	p-value
Population_Density	0.90618	0.00030227***
Census_Population	0.89875	0.00040617***
Total_Families	0.87466	0.00092561***
Households_With_Under_18	0.67465	0.03235537*
Land_Area	-0.28708	0.42126310

Full correlation matrix.

Full Correlation Matrix

	Padacity_ Sales	Census_ Population	Households_ With_Under_18	Land_ Area	Population_ Density	Total_ Families
Padacity_Sales	1.00000	0.89875	0.67465	-0.28708	0.90618	0.87466
Census_ Population	0.89875	1.00000	0.91156	-0.05247	0.94439	0.96919
Households_ With_Under_18	0.67465	0.91156	1.00000	0.18938	0.82199	0.90566
Land_Area	-0.28708	-0.05247	0.18938	1.00000	-0.31742	0.10730
Population_ Density	0.90618	0.94439	0.82199	-0.31742	1.00000	0.89168
Total_Families	0.87466	0.96919	0.90566	0.10730	0.89168	1.00000

Matrix of p-values for Predictor Variables.

Matrix of Corresponding p-values

	Padacity_ Sales	Census_ Population	Households_ With_Under_18	Land_ Area	Population_ Density	Total_ Families
Padacity_Sales		4.0617e-04	3.2355e-02	4.2126e-01	3.0227e-04	9.2561e-04
Census_ Population	4.0617e-04		2.4026e-04	8.8554e-01	3.9116e-05	3.7982e-06
Households_ With_Under_18	3.2355e-02	2.4026e-04		6.0028e-01	3.5227e-03	3.0883e-04
Land_Area	4.2126e-01	8.8554e-01	6.0028e-01		3.7148e-01	7.6796e-01
Population_ Density	3.0227e-04	3.9116e-05	3.5227e-03	3.7148e-01		5.2748e-04
Total_Families	9.2561e-04	3.7982e-06	3.0883e-04	7.6796e-01	5.2748e-04	

The full correlation matrix shows good correlation between predictor variables, Census_Population, Households_with_Under_18, Population_Density and Total_Families. There may be some multicollinearity here.

Land_Area does not show great correlation with the other predictor variables so I will start by running a regression with Land_Area and add other predictor variables to the regression.

Report for Linear Model New_Store_Prediction___LM

Basic Summary

Call:

lm(formula = Padacity_Sales ~ Land_Area, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-158400	-110900	-78940	39380	539600

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	432401.73	146208.93	2.9574	0.01822*
Land_Area	-36.07	42.56	-0.8477	0.42126

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

Residual standard error: 217368 on 8 degrees of freedom

Multiple R-squared: 0.08241, Adjusted R-Squared: -0.03228

F-statistic: 0.7185 on 1 and 8 DF, p-value: 0.4213

Type II ANOVA Analysis

Response: Padacity_Sales

	Sum Sq	DF	F value	Pr(>F)
Land_Area	33949588837.33	1	0.72	0.42126
Residuals	377992295324.27	8		

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

R-Squared for linear model between Sales vs Land_Area = **0.08241**

Report for Linear Model New_Store_Prediction___LM

Basic Summary

Call:

lm(formula = Padacity_Sales ~ Census_Population + Land_Area, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-165000	-28630	-9045	30190	120300

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	210872.04	69180.625	3.048	0.01863*
Census_Population	11.03	1.728	6.383	0.00037***
Land_Area	-30.23	17.443	-1.733	0.12668

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

Residual standard error: 88974 on 7 degrees of freedom

Multiple R-squared: 0.8655, Adjusted R-Squared: 0.827

F-statistic: 22.52 on 2 and 7 DF, p-value: 0.0008928

Type II ANOVA Analysis

Response: Padacity_Sales

	Sum Sq	DF	F value	Pr(>F)
Census_Population	322578046861.07	1	40.75	0.00037***
Land_Area	23777499407.94	1	3	0.12668
Residuals	55414248463.21	7		

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

Adjusted R-Squared for linear model between Sales vs Land_Area vs Census_Population = **0.827**

Report for Linear Model New_Store_Prediction___LM

Basic Summary

Call:

lm(formula = Padacity_Sales ~ Households_With_Under_18 + Land_Area, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-260700	-50920	-1834	47390	249800

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	297611.68	107140.63	2.778	0.02739*
Households_With_Under_18	63.09	19.44	3.245	0.01415*
Land_Area	-54.07	29.28	-1.847	0.10727

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

Residual standard error: 146831 on 7 degrees of freedom

Multiple R-squared: 0.6336, Adjusted R-Squared: 0.529

F-statistic: 6.054 on 2 and 7 DF, p-value: 0.02976

Type II ANOVA Analysis

Response: Padacity_Sales

	Sum Sq	DF	F value	Pr(>F)
Households_With_Under_18	227077058908.59	1	10.53	0.01415*
Land_Area	73529107680.71	1	3.41	0.10727
Residuals	150915236415.68	7		

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

Adjusted R-Squared for linear model between Sales vs Land_Area vs Households_With_Under_18 = **0.529**

Report for Linear Model New_Store_Prediction___LM

Basic Summary

Call:

lm(formula = Padacity_Sales ~ Land_Area + Population_Density, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-177100	-13380	17900	34970	134600

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.435e+05	87450.27	1.641189	0.14476
Land_Area	7.846e-02	21.18	0.003704	0.99715
Population_Density	3.145e+04	5848.33	5.377362	0.00103**

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

Residual standard error: 102588 on 7 degrees of freedom

Multiple R-squared: 0.8212, Adjusted R-Squared: 0.7701

F-statistic: 16.07 on 2 and 7 DF, p-value: 0.002419

Type II ANOVA Analysis

Response: Padacity_Sales

	Sum Sq	DF	F value	Pr(>F)
Land_Area	144414.54	1	0	0.99715
Population_Density	304321939965.4	1	28.92	0.00103**
Residuals	73670355358.88	7		

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

Adjusted R-Squared for linear model between Sales vs Land_Area vs Population_Density = **0.7701**

Report for Linear Model New_Store_Prediction___LM

Basic Summary

Call:

lm(formula = Padacity_Sales ~ Land_Area + 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*
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: Padacity_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		

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

Adjusted R-Squared for linear model between Sales vs Land_Area vs Total_Families = **0.8866**

Report for Linear Model New_Store_Prediction___LM

Basic Summary

Call:

lm(formula = Padacity_Sales ~ Census_Population + Land_Area + Total_Families, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-110000	-4750	10180	41560	75240

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	196536.22	60172.001	3.2662	0.01711*
Census_Population	-3.21	7.855	-0.4087	0.69697
Land_Area	-53.55	19.644	-2.7262	0.03436*
Total_Families	62.78	33.998	1.8465	0.11434

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

Residual standard error: 76741 on 6 degrees of freedom

Multiple R-squared: 0.9142, Adjusted R-Squared: 0.8713

F-statistic: 21.32 on 3 and 6 DF, p-value: 0.001335

Type II ANOVA Analysis

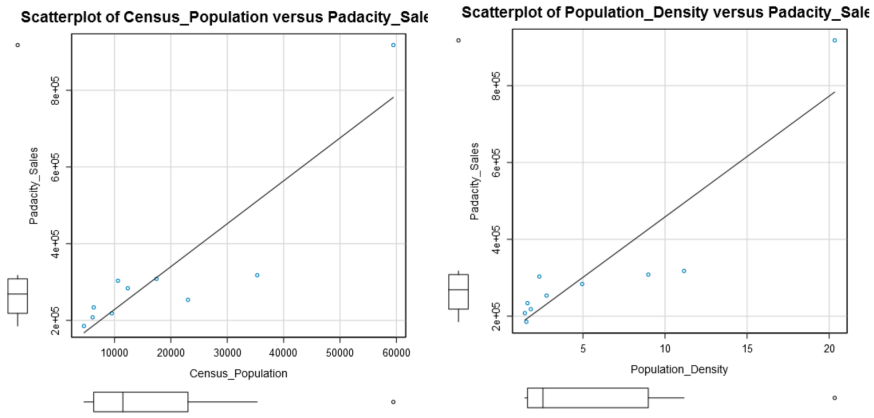
Response: Padacity_Sales

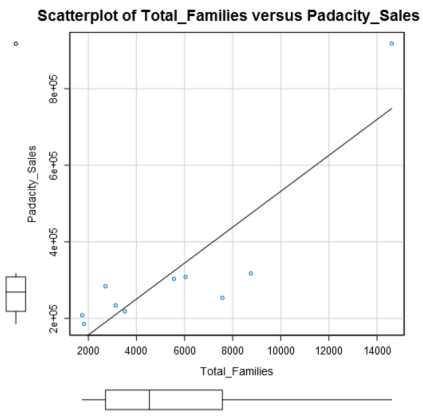
	Sum Sq	DF	F value	Pr(>F)
Census_Population	983564136.27	1	0.17	0.69697
Land_Area	43768907210.74	1	7.43	0.03436*
Total_Families	20079363193.04	1	3.41	0.11434
Residuals	35334885270.17	6		

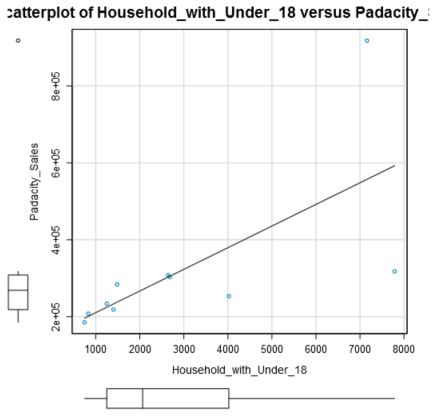
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

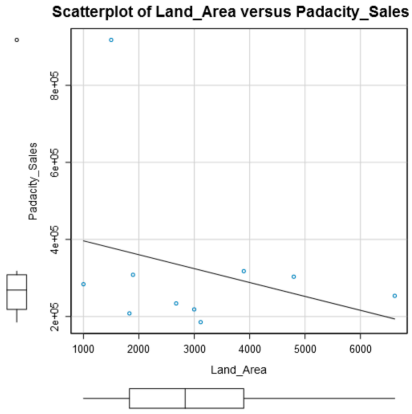
Adjusted R-Squared for linear model between Sales vs Land_Area vs Total_Families vs Census_Population = **0.8713**

Below are the scatter plots of the predictor variables vs our target variable (Pawdacity_Sales).









The scatter plots above give a good representation of the linearity between the target variable and its respective predictor variable.

Starting with Land_Area as a predictor variable (R-Squared = 0.08241) and adding the other variables, I can see that the largest jump in R-Squared comes from Land_Area and Total_Families (adjusted r-squared = 0.8866)

I will use Land_Area and Total_Families as my predictor variables for my linear model.

Below is the summary of the multilinear regression model.

Report for Linear Model New_Store_Prediction___LM

Basic Summary

Call:

lm(formula = Padacity_Sales ~ Land_Area + 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*
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: Padacity_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		

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

From the summary, the equation for the linear regression model is:

Y (Pawdacity_Sales) = 197330.41 – 48.42(Land_Area) + 49.14(Total_Families)

Final recommendation.

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.

With the required criteria, I would recommend Laramie City. Laramie City does not currently contain a store, has an estimated census population for 2014 of 32,081 and predicted sales of **$305,013.88**.

Below is a summary of the final possibilities for a new store with the highlighted row as the recommendation.

City	2014_Census_Pop_Est	Total_Families	Score
Laramie	32081.00	4668.93	305013.88
Torrington	6736.00	2548.50	245081.79
Jackson	10449.00	2313.08	225870.82
Lander	7642.00	3876.81	225751.40
Green River	12630.00	3977.40	224372.00
Worland	5366.00	1364.32	201700.33