Recommend a City

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?
   In which city in Wyoming should pet store chain Pawdacity open a new store?

2. What data is needed to inform those decisions?
   We need data on expected yearly sales values for different cities in Wyoming to decide which city is best for opening a new pet store. The city that has the highest predicted yearly sales would be chosen.

   We need to build a regression model in order to predict yearly sales in different cities.

   Data on current yearly sales at 13 Pawdacity stores in various cities, census data for population and demographic data at these cities will be used as a training set for building the regression model.

Step 2: Building the Training Set

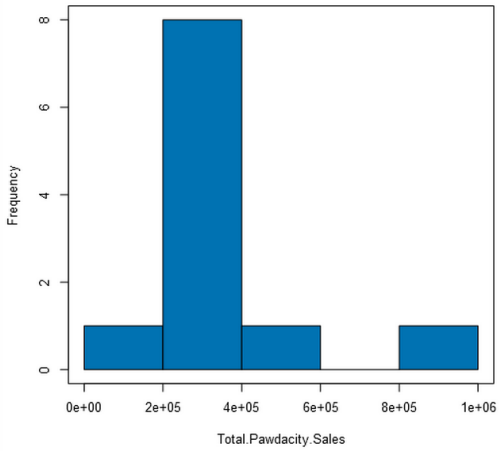
Column	Sum	Average
Census Population	*213,862*	19,442
Total Pawdacity Sales	*3,773,304*	343,027.64
Households with Under 18	*34,064*	3096.73
Land Area	*33,071*	3006.49
Population Density	*63*	5.7
Total Families	*62,653*	5695.71

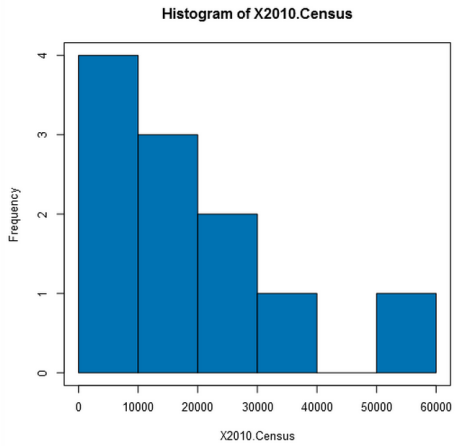
Step 3: Dealing with Outliers

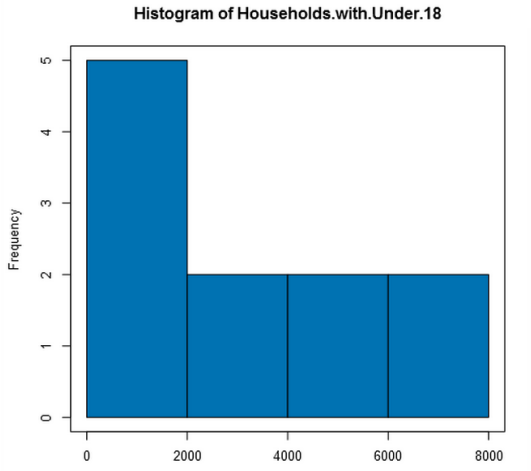
Are there any outliers in the training set? Which outlier have you chosen to remove or impute?

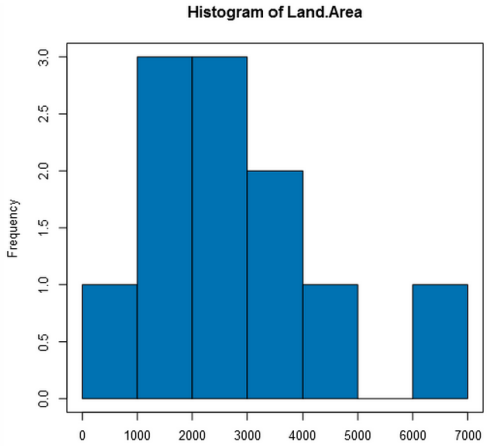
1. From the IQR calculations the outliers in each variable can be identified as:
   Pawdacity Sales > 443,232 (2 outlier Gillette: 543,132 and Cheyenne: 917,892)
   2010 Census > 53,278 (1 outlier Cheyenne: 59,466)
   Household with Under 18 > 8,102 (0 outlier)
   Land Area > 5970 (1 Outlier Rock Springs 6620)
   Population Density > 16 (1 outlier Cheyenne: 20.34)
   Total Families > 14,066 (1 outlier Cheyenne: 14,613)

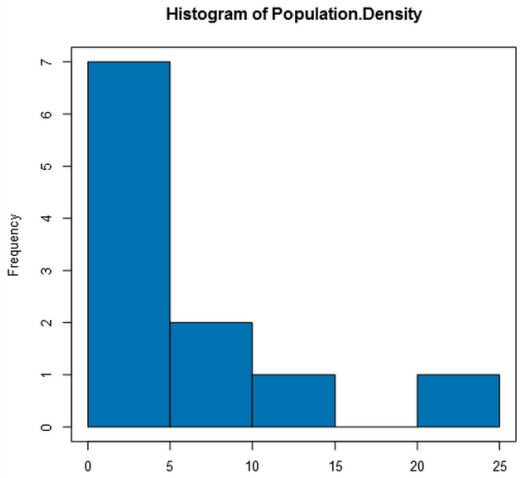
2. Histograms: Histograms also show the presence of outliers.

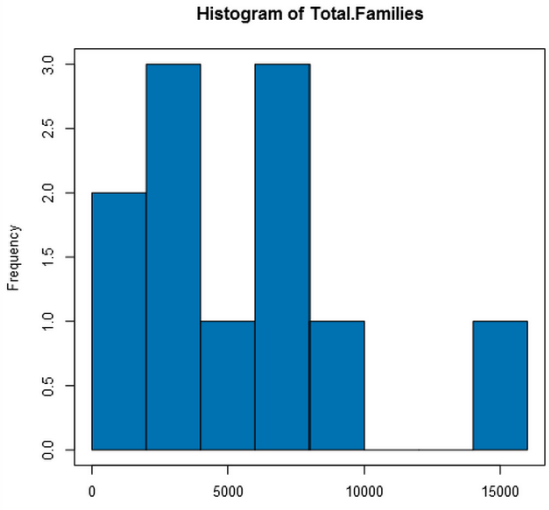










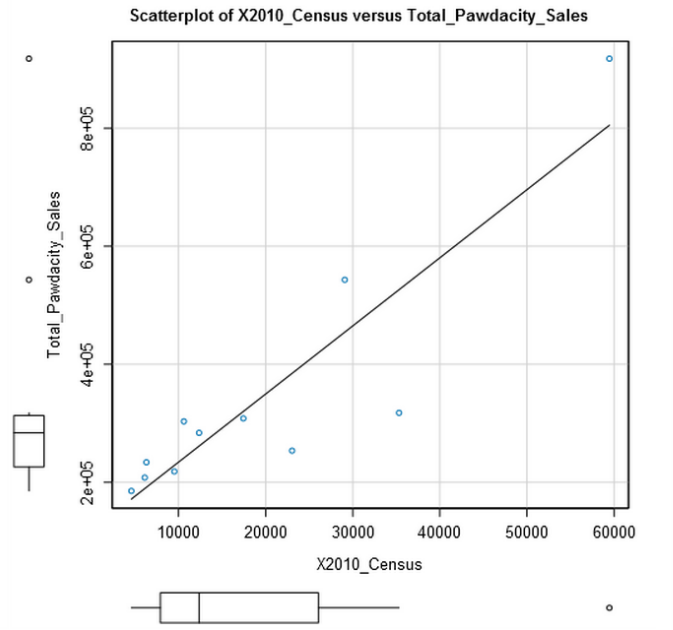


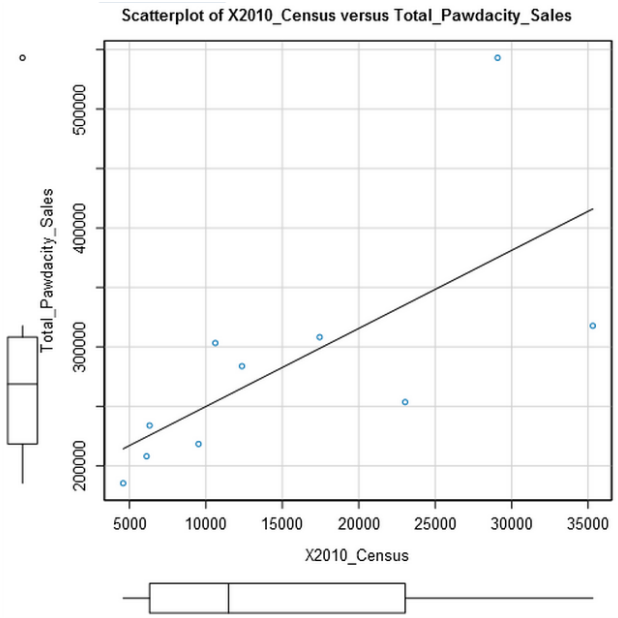
3. Scatterplots to identify outliers:

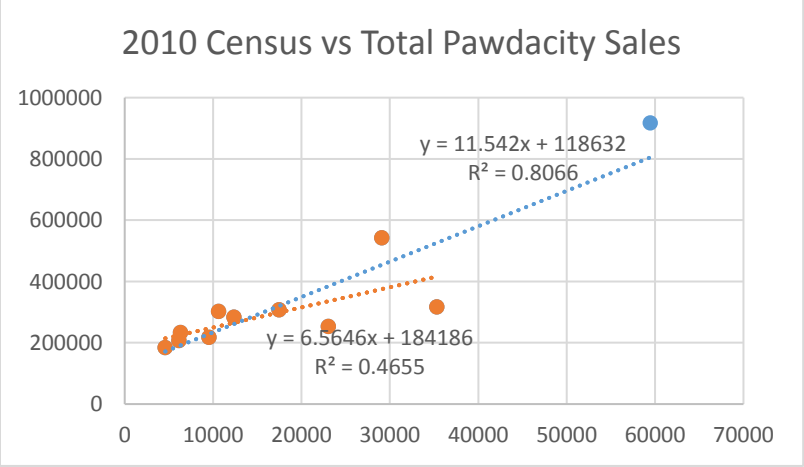
We show alteryx scatterplots with outliers and 1 extreme outlier in Pawdacity Sales removed. To quantitatively note the effect of removing outlier, the excel plots with slope values are also shown.

a) 2010 Census vs Total Pawdacity Sales:

Boxplots show 2 outliers in Total Pawdacity sales and 1 outlier in 2010 census data.

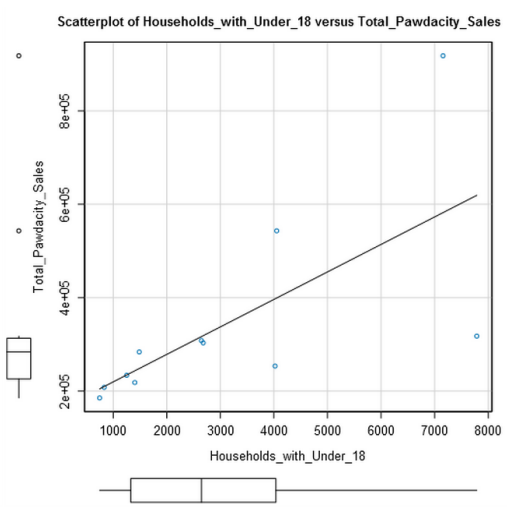


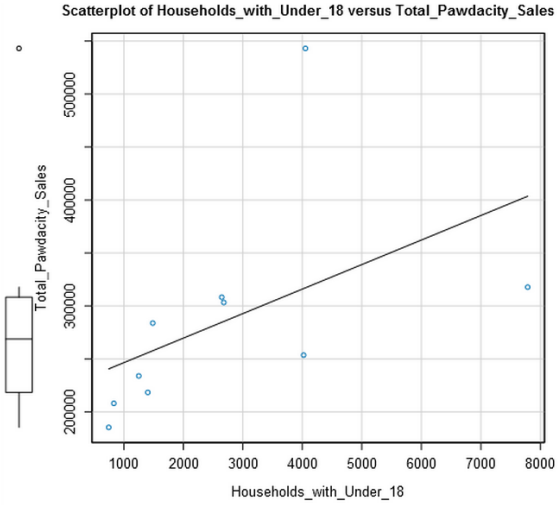


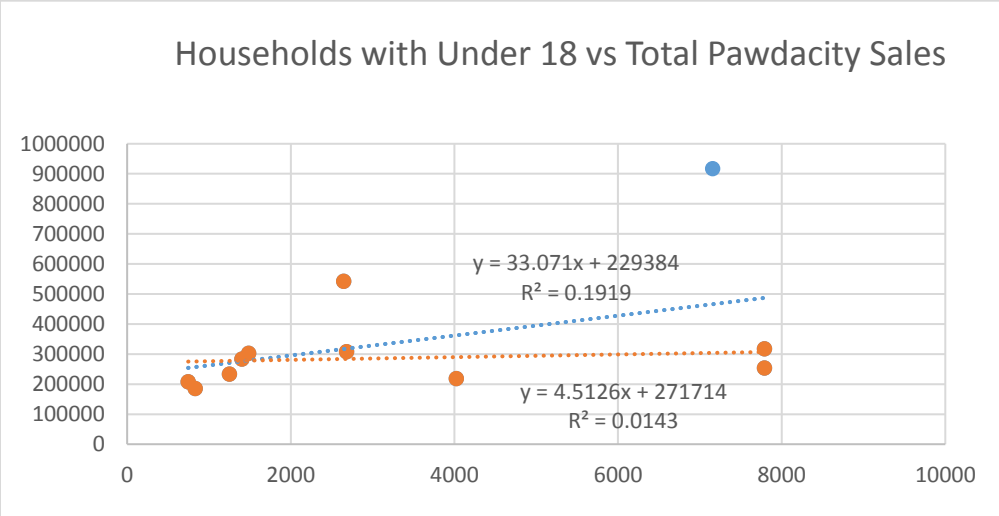


Blue: with outlier
Orange: 1 outlier removed

b) Households with individual under 18 vs Total Pawdacity Sales:
Boxplots show 0 outliers in Households with individuals under 18 data

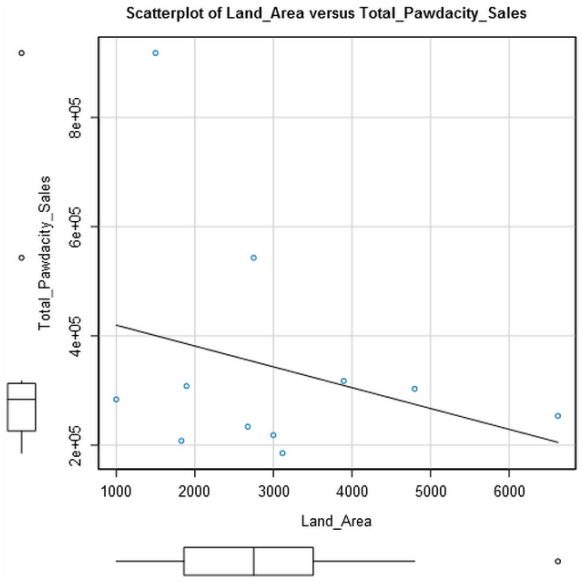


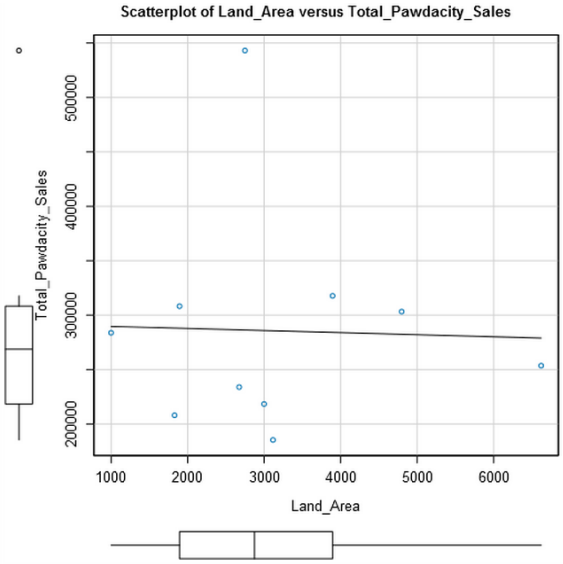


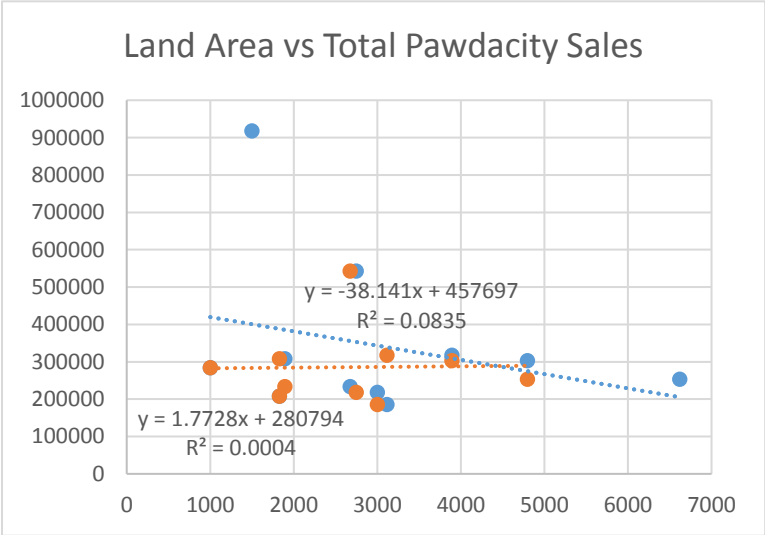


Blue: with outlier
Orange: 1 outlier removed

c) Land Area vs Total Pawdacity Sales:
Land Area data shows 1 outlier

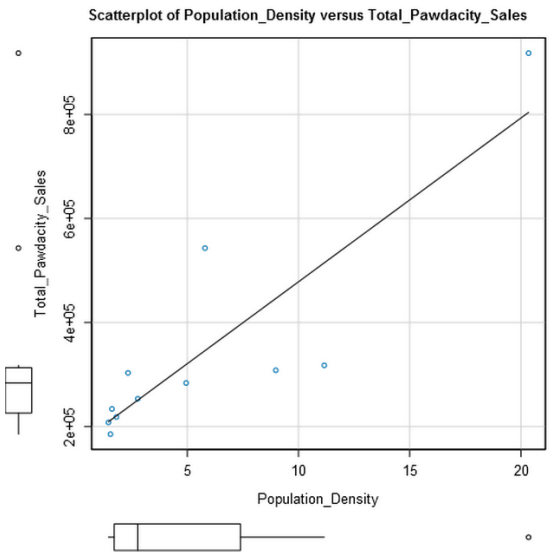


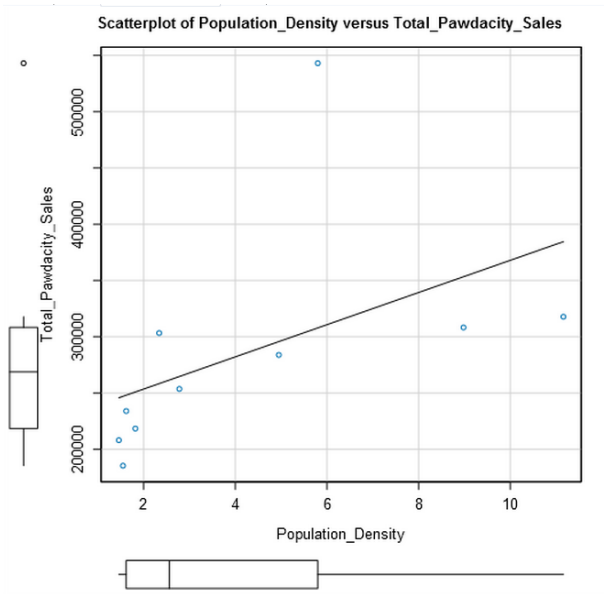


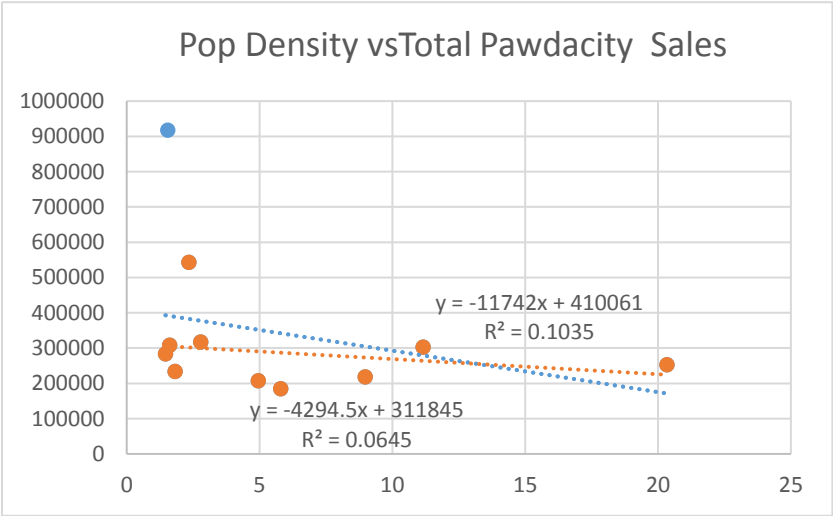


Blue: with outlier
Orange: 1 outlier removed

d) Population Density vs Total Pawdacity Sales:
Population Density boxplot shows 1 outlier

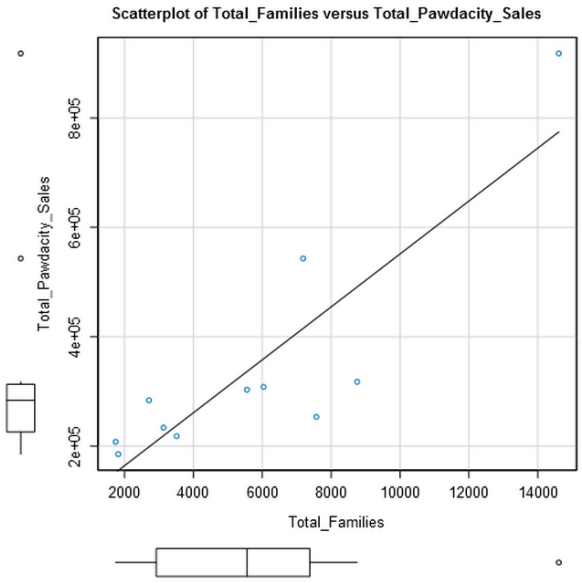


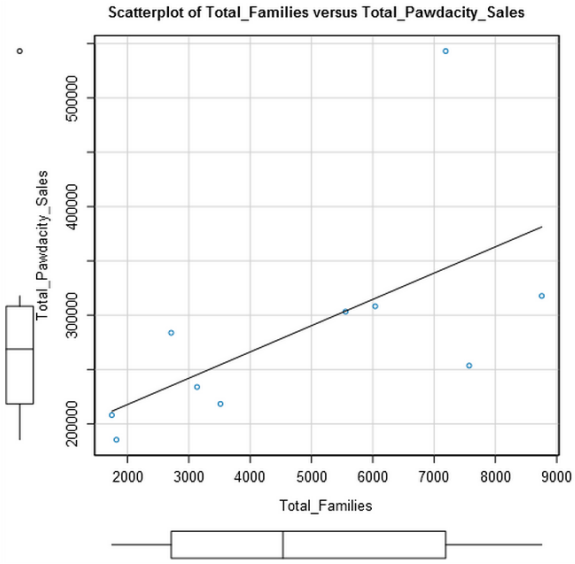


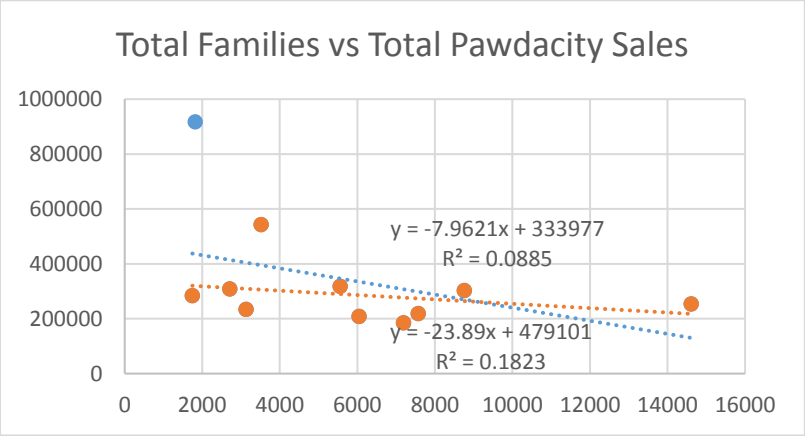


Blue: with outlier
Orange: 1 outlier removed

e) Total Families vs Total Pawdacity Sales:
Total Families boxplot shows 1 outlier







Blue: with outlier
Orange: 1 outlier removed

Boxplots & Scatterplots confirm the number of outliers calculated using the inter quartile range.

When the largest outlier in Total Pawdacity Sales at Cheyenne: 917,892 is removed, scatterplots show that slopes of the lines change drastically. All variables are very sensitive to the presence or absence of outliers in Total Pawdacity Sales.

The record corresponding to Total Pawdacity Sales for Cheyenne: 917,892 also corresponds to the outlier values of 2010 Census (Cheyenne: 59,466), Population Density (Cheyenne: 20.34) and Total Families (Cheyenne: 14,613).

Due to the added advantage of simultaneously removing outliers in many variables, we drop the Total Pawdacity Sales =917,892.

The 2nd outlier at Total Pawdacity Sales for Gillette: 543,132, does not appear to be extreme. To not lose information from the very small data set, we leave this data point in.

For the above same reasons the outlier in Land Area for Rock Springs at 6620 is also not removed.

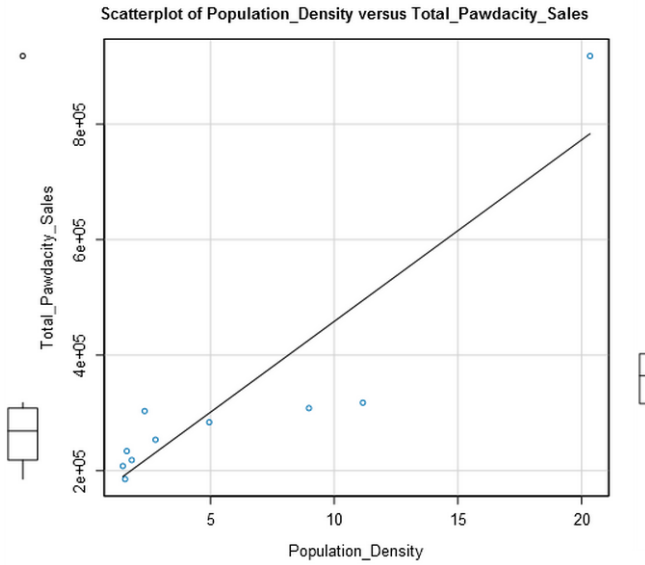
Final Training Data Set

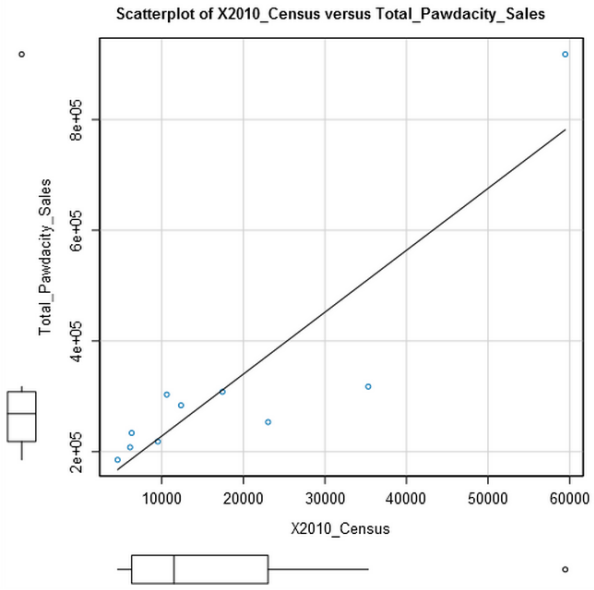
City	2010 Census	Total Pawdacity Sales	Households with Under 18	Land Area	Population Density	Total Families
Buffalo	4585	185328	746	3115.508	1.55	1819.5
Douglas	6120	208008	832	1829.465	1.46	1744.08
Cody	9520	218376	1403	2998.957	1.82	3515.62
Powell	6314	233928	1251	2673.575	1.62	3134.18
Rock Springs	23036	253584	4022	6620.202	2.78	7572.18
Evanston	12359	283824	1486	999.4971	4.95	2712.64
Riverton	10615	303264	2680	4796.86	2.34	5556.49
Sheridan	17444	308232	2646	1893.977	8.98	6039.71
Casper	35316	317736	7788	3894.309	11.16	8756.32
Gillette	29087	543132	4052	2748.853	5.8	7189.43

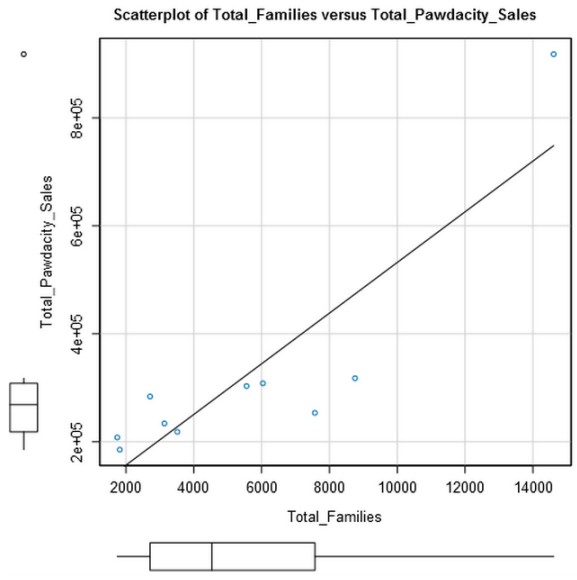
Step 4: Linear Regression

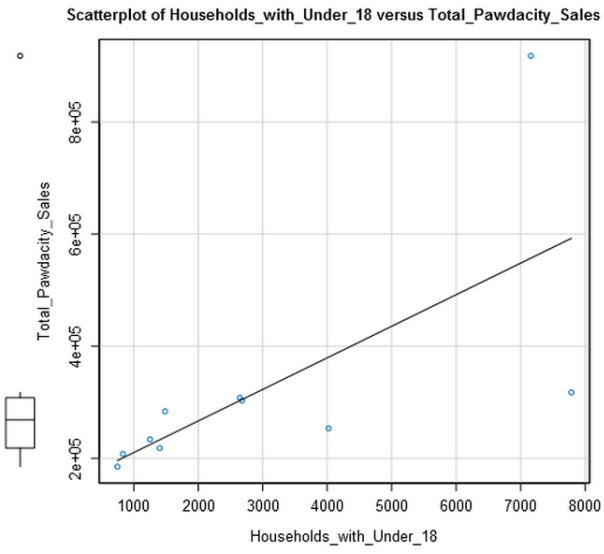
1. How and why did you select the predictor variables (see supplementary text) in your model? You must show that each predictor variable has a linear relationship with your target variable with a scatterplot.

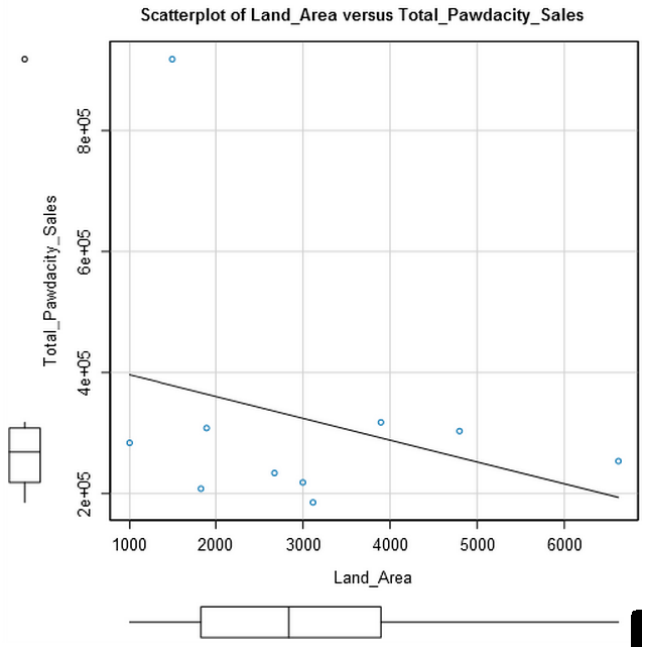
The scatterplots display the linear relation of target variable Total Pawdacity Sales with variables Population Density, 2010 Census and Total Families and Households under 18 and Land Area.











The linear trends of the scatterplots imply that all 5 variables, Population Density, 2010 Census and Total Families and Households under 18 and Land Area are likely predictors for target Total Pawdacity Sales.

The Pearson analysis for association between the target variable Total Pawdacity Sales with possible predictor in decreasing order of importance is shown below.

Pearson Correlation Analysis

Focused Analysis on Field Total.Pawdacity.Sales

	Association Measure	p-value	
Population.Density	0.90618	0.00030227	***
X2010.Census	0.89875	0.00040617	***
Total.Families	0.87466	0.00092561	***
Households.with.Under.18	0.67465	0.03235537	*
Land.Area	-0.28708	0.42126310	

The association measure with corresponding p values show that Total Pawdacity Sales is highly correlated to Population Density, 2010 Census and Total Families.

Full Correlation Matrix

	Total.Pawdacity.Sales	X2010.Census	Households.with.Under.18	Land.Area	Population.Density	Total.Families
Total.Pawdacity.Sales	1.00000	0.89875	0.67465	-0.28708	0.90618	0.87466
X2010.Census	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 Corresponding p-values

	Total.Pawdacity.Sales	X2010.Census	Households.with.Under.18	Land.Area	Population.Density	Total.Families
Total.Pawdacity.Sales		4.0617e-04	3.2355e-02	4.2126e-01	3.0227e-04	9.2561e-04
X2010.Census	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 correlation between predictors in table above shows that Population Density, 2010 Census and Total Families and Households with under 18 are highly inter correlated with association measure >= 0.9.

The corresponding p values show that these correlations are significant.

To avoid collinearity issues, we should retain only any one of the highly correlated variables.

To decide which from the 4 collinear variables Population Density, 2010 Census, Total Families and Households under 18 has to be kept as a predictor, we look at linear regression models with each combination and decide the best one based on adjusted R square values and p values.

The low association measure of Land Area to other predictors makes it a good candidate predictor variable.

2. Explain why you believe your linear model is a good model. You must justify your reasoning using the statistical results that your regression model created. . For each variable you selected, please justify how each variable is a good fit for your model by using the p-values and R-squared values that your model produced.

Results of 4 models with each of the 4 collinear predictors are shown below

Linear regression model (a): Population Density and Land Area.

Report for Linear Model PetSales

Basic Summary

Call:
lm(formula = Total.Pawdacity.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: Total.Pawdacity.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

The adjusted R squared value is 0.77 but the p values for the intercept and coefficient for Land Area are not significant with values greater than 0.05.

The linear regression model (b): 2010 Census and Land Area

Report for Linear Model PetSales

Basic Summary

Call:
lm(formula = Total.Pawdacity.Sales ~ X2010.Census + 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	*
X2010.Census	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: Total.Pawdacity.Sales

	Sum Sq	DF	F value	Pr(>F)	
X2010.Census	322578046861.06	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

The adjusted R squared value is 0.827 but the p values for the coefficient for Land Area is not significant with values greater than 0.05.

Linear regression model (c): Households under 18 and Land Area

Report for Linear Model PetSales

Basic Summary

Call:
lm(formula = Total.Pawdacity.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: Total.Pawdacity.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

The adjusted R squared value is low at 0.529 but the p values for the coefficient for Land Area is not significant, with value greater than 0.05.

Linear regression model (d): Total Families and Land Area

Report for Linear Model PetSales

Basic Summary

Call:
lm(formula = Total.Pawdacity.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: Total.Pawdacity.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

The adjusted R squared value for model is 0.8866. Thus 89% of variation in the Total Pawdacity Sales data can be explained by the model.

The p values for coefficients of intercept and Total Families and Land Area are significant with values much less than 0.05.

Comparing the adjusted R-squared values for different models

Model(a)	Population Density and Land Area	Adjusted R square: 0.7701
Model(b)	2010 Census and Land Area	Adjusted R square: 0.827
Model(c)	Households under 18 and Land Area	Adjusted R square: 0.529
Model(d)	Total Families and Land Area	Adjusted R square: 0.8866

Model (d) with Total Families and Land Area as predictors has the lowest adjusted R-square value.

We thus pick this linear model as a training model to predict the pet store sales in different cities in Wyoming.

3. What is the best linear regression equation based on the available data? Each coefficient should have no more than 2 digits after the decimal (ex: 1.28)

The linear regression equation is

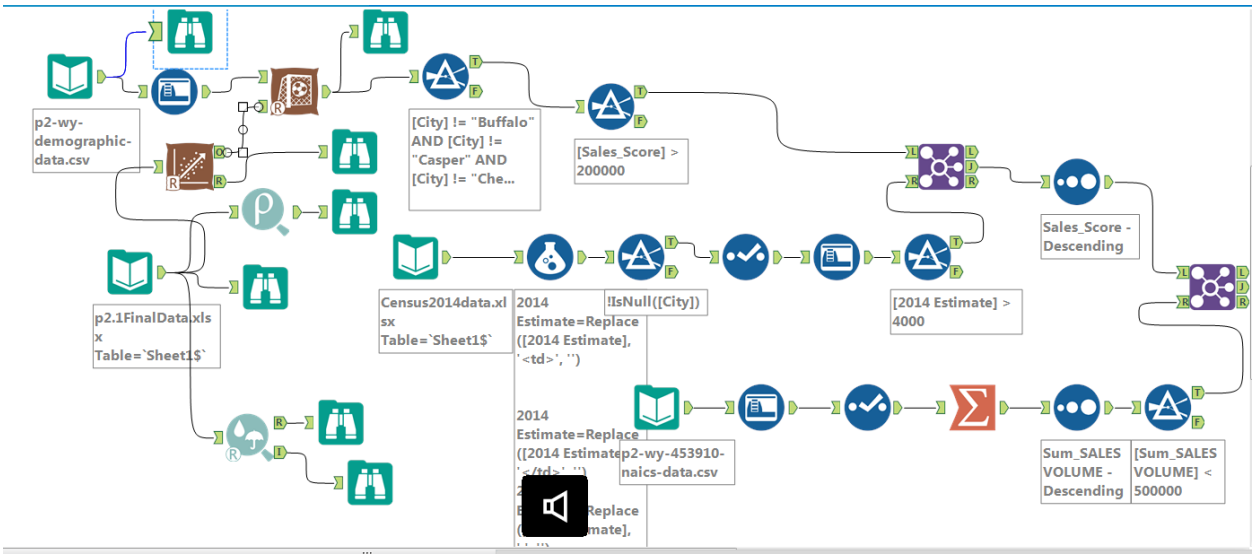
Total Pawdacity Sales = 197330.41 – 48.42 Land Area + 49.14 Total Families

The linear regression equation shows that for decrease in Land Area by 1 sq.mile, the Total Pawdacity Sales decreases by $48.42 and for increase in total Families by 1 the the Total Pawdacity Sales increases by $49.14.

Step 5: Analysis

1. Which city would you recommend and why did you recommend this city?

The workflow to find the best city for Pawdacity to open a new pet store is shown below.



From the analysis, 4 cities satisfy the required conditions:

1. The new store should be located in a city where there are no current Pawdacity stores.
2. The new city has a population over 4,000 people (based upon the 2014 US Census estimate).
3. The predicted yearly sales must be over $200,000.
4. The total sum of sales of competitors in the city is less than $500,000

Record #	City	County	Land.Area	Households.with.Under.18	Population.Density	Total.Families	Sales_Score	2014 Estimate	Sum_SALES VOLUME
1	Jackson	Teton	1757.6592	1078	2.36	2313.08	225870.8236	10449	182000
2	Lander	Fremont	3346.80934	1870	1.63	3876.81	225751.400203	7642	152197
3	Laramie	Albany	2513.745235	2075	5.19	4668.93	305013.881671	32081	76000
4	Worland	Washakie	1294.105755	595	2.18	1364.32	201700.325919	5366	169000

A final condition that the city chosen has the highest predicted sales from the predicted set means we choose Laramie.

Thus our recommendation is, Pawdacity should open a new store in Lamarie, WY where the predicted total sales value is $ 305,013.88.

This is a new city where Pawdacity does not have an existing store, has population of 32,081 according to 2014 census, the predicted sales are over $200,000 and total sales of all other competitors is $76,000.