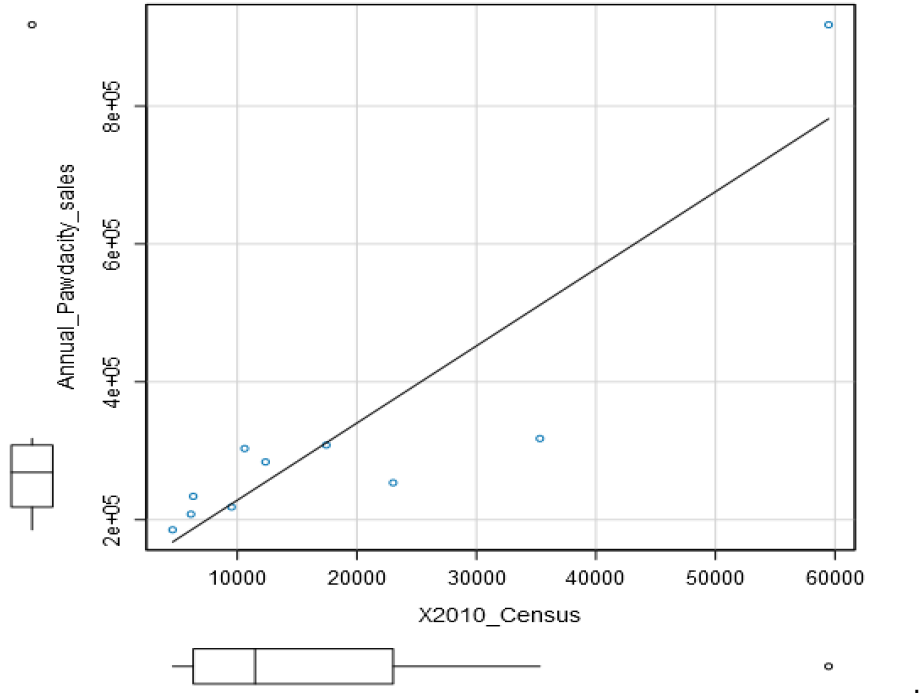
Project 2.2: Recommend a City

Step 1: Linear Regression

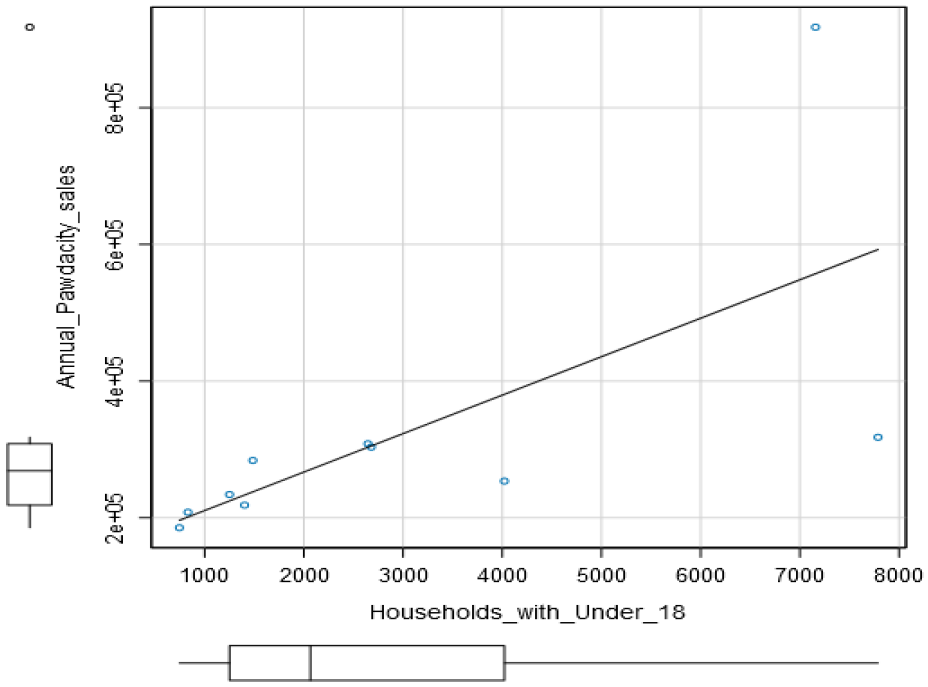
1. How and why did you select the predictor variables (see supplementary text) in your model? You must show that each predictor variable has a linear relationship with your target variable with a scatterplot.

First, to decide the potential predictor variables, we need to check which of the variables have linear relationship with the target variable, which is Annual Sales. To validate this, we generate scatterplots for each of the variable with the target variable.

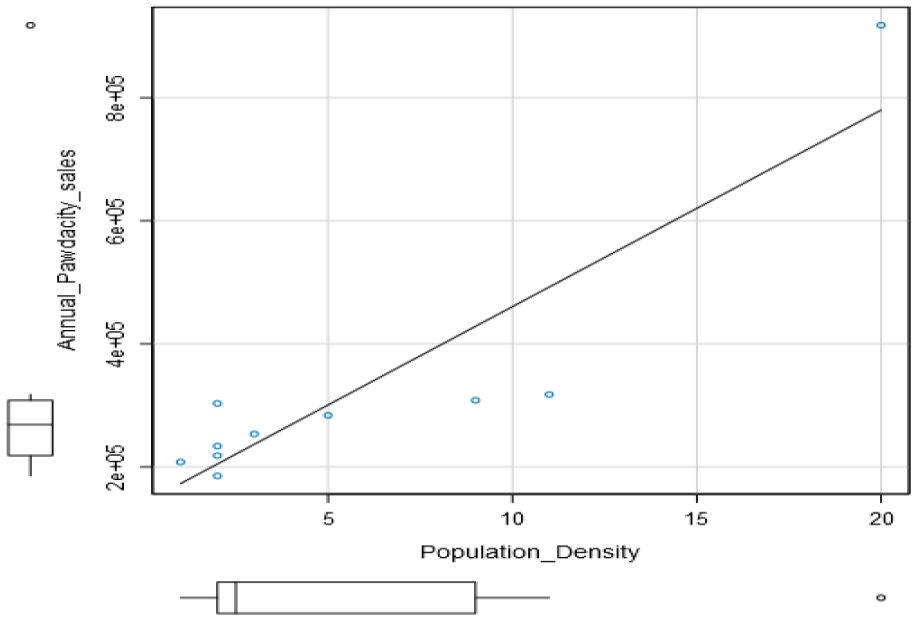
Scatterplot of 2010 Census Vs Total Sales



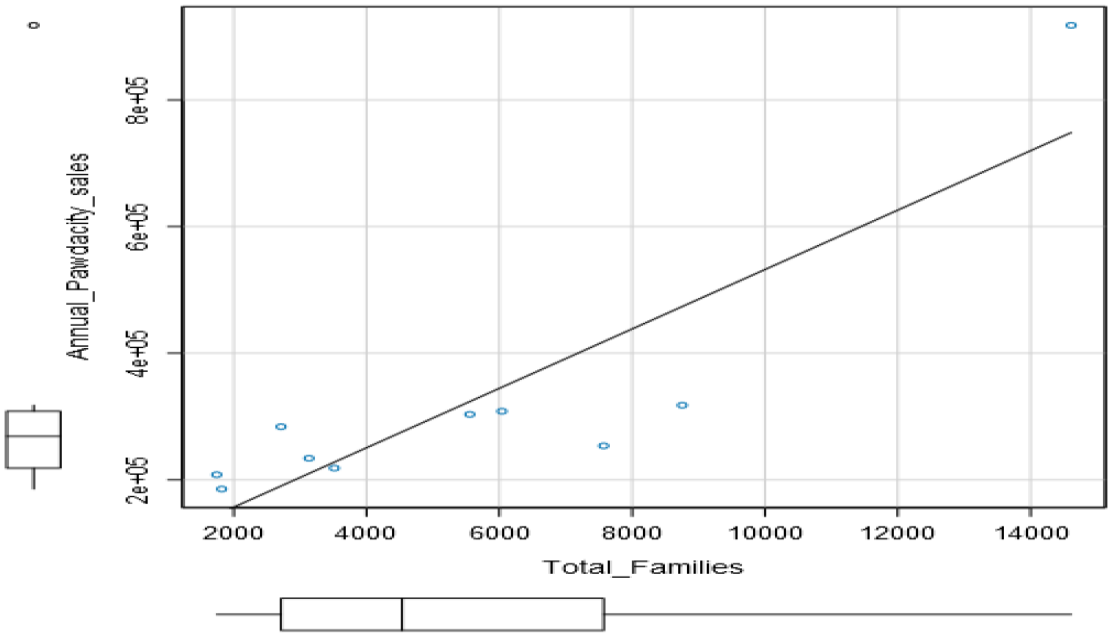
Scatterplot of Households with under 18 Vs Total Sales



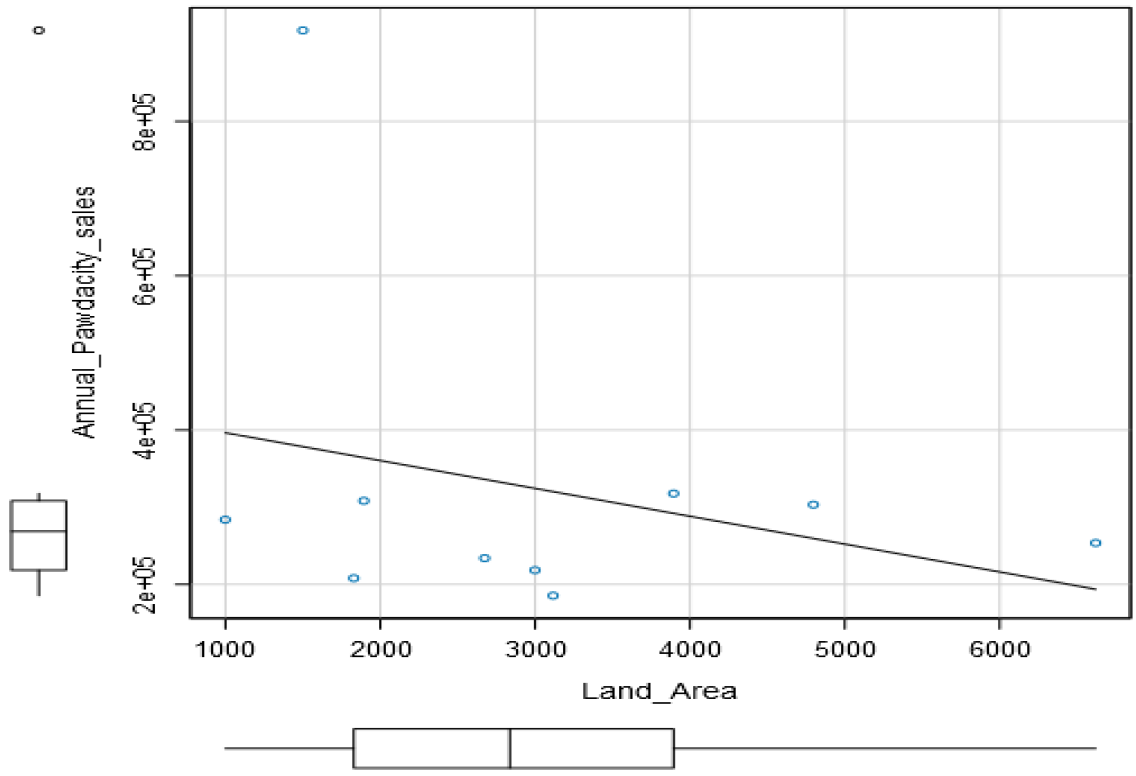
Scatterplot of Population Density Vs Total Sales



Scatterplot of Total Families Vs Total Sales



Scatterplot of Land Area Vs Total Sales



After analyzing the scatterplots, we can see that all the variables have a linear relationship with the target variables. Land are having a negative linear relationship with Total Sales, it just means that when value of one of them increases, the value of other decreases which indicates that there definitely is some relationship among them. So, *we can consider Population Density, 2010 Census, Total Families, Households with under 18 and Land Area as potential predictor variables*.

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.

Having selected the potential predictor variables, now we check for the correlation among the variables. To do this, we use Association analysis tool and generate the Pearson Correlation Analysis report as shown below. We can see that Population Density, 2010 Census, Total Families are strongly correlated with the target variable based on the p-value. Households with under 18 is also correlated to the target variable which is the Annual Sales. The stars next to the p-value also suggest the same.

Land Area is the one which is not correlated to the target variable.

Pearson Correlation Analysis

Focused Analysis on Field Annual_Pawdacity_sales

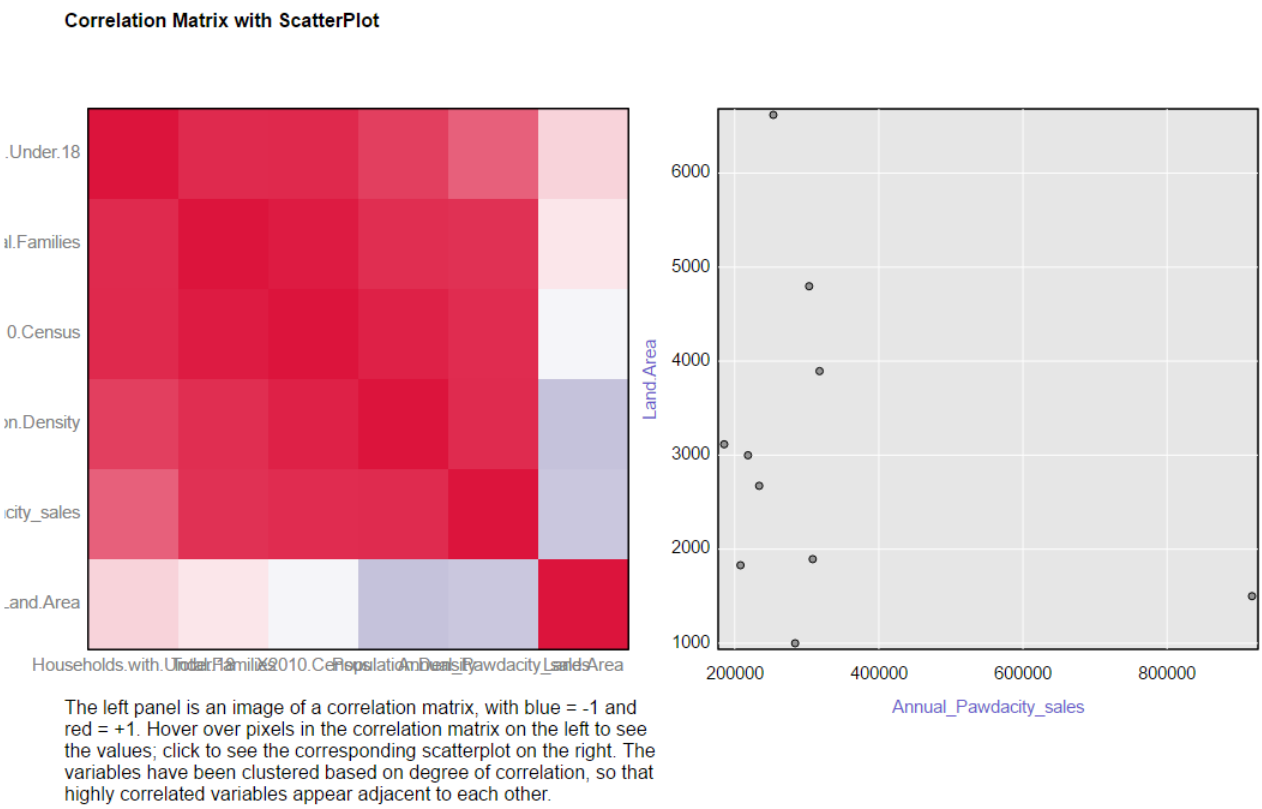
	Association Measure	p-value	
Population.Density	0.90185	0.00036008	***
X2010.Census	0.89875	0.00040617	***
Total.Families	0.87469	0.00092495	***
Households.with.Under.18	0.67465	0.03235537	*
Land.Area	-0.28711	0.42121354	

Full Correlation Matrix

	Annual_Pawdacity_sales	Total.Families	Households.with.Under.18	X2010.Census	Land.Area	Population.Density
Annual_Pawdacity_sales	1.000000	0.874687	0.674652	0.898755	-0.287107	0.901853
Total.Families	0.874687	1.000000	0.905645	0.969201	0.107203	0.889892
Households.with.Under.18	0.674652	0.905645	1.000000	0.911562	0.189302	0.818637
X2010.Census	0.898755	0.969201	0.911562	1.000000	-0.052537	0.942936
Land.Area	-0.287107	0.107203	0.189302	-0.052537	1.000000	-0.314513
Population.Density	0.901853	0.889892	0.818637	0.942936	-0.314513	1.000000

Matrix of Corresponding p-values

	Annual_Pawdacity_sales	Total.Families	Households.with.Under.18	X2010.Census	Land.Area	Population.Density
Annual_Pawdacity_sales		9.2495e-04	3.2355e-02	4.0617e-04	4.2121e-01	3.6008e-04
Total.Families	9.2495e-04		3.0903e-04	3.7931e-06	7.6817e-01	5.6193e-04
Households.with.Under.18	3.2355e-02	3.0903e-04		2.4026e-04	6.0043e-01	3.7791e-03
X2010.Census	4.0617e-04	3.7931e-06	2.4026e-04		8.8539e-01	4.3290e-05
Land.Area	4.2121e-01	7.6817e-01	6.0043e-01	8.8539e-01		3.7611e-01
Population.Density	3.6008e-04	5.6193e-04	3.7791e-03	4.3290e-05	3.7611e-01	



With the reports and the correlation matrix, we see that the variables Households under 18, Total Families, 2010 Census and Population Density have strong correlation among them which is known as multicollinearity or collinearity. This means that we should not use more than one of these in the regression model failing which there would be issues while predicting the results.

From the Pearson report and the Correlation matrix, we see that Land Area is not highly correlated with the rest of the variables which means, it is the only variable that is not duplicate. Also, the correlation between Total Sales and Land Area is very low and in the scatterplot, there are no patterns and total randomness, meaning that the movement in Land Area field does not affect the Sales. This is exactly what we are wanting in our predictor variables and hence Land Area will be the first choice for the predictor variables.

Now, we can select only one among the remaining variables as there is Multicollinearity as explained above. To determine which among the remaining can be selected as a predictor variable, 4 linear regression models are created with the below variables.

a. Land Area vs Total Families
b. Land Area vs Households with under 18
c. Land Area vs 2010 Census
d. Land Area vs Population Density

Land Area vs Total Families

Report

Report for Linear Model Linear_Regression_50

Basic Summary

Call:
lm(formula = Annual_Pawdacity_sales ~ Total.Families + Land.Area, data = inputs$the.data)

Residuals:

Min	1Q	Median	3Q	Max
-121300	-4467	8422	40490	75210

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	197299.27	56451.744	3.495	0.01006	*
Total.Families	49.13	6.055	8.115	8e-05	***
Land.Area	-48.41	14.184	-3.413	0.01124	*

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

Residual standard error: 72033 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: Annual_Pawdacity_sales

	Sum Sq	DF	F value	Pr(>F)	
Total.Families	341664344221.7	1	65.85	8e-05	***
Land.Area	60453713643.39	1	11.65	0.01124	*
Residuals	36321013347.65	7			

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

Land Area vs Households with under 18

Report

Report for Linear Model Linear_Regression_52

Basic Summary

Call:
lm(formula = Annual_Pawdacity_sales ~ Households.with.Under.18 + Land.Area, data = inputs$the.data)

Residuals:

Min	1Q	Median	3Q	Max
-260700	-50940	-1822	47370	249800

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	297599.42	107142.82	2.778	0.02739	*
Households.with.Under.18	63.09	19.44	3.245	0.01415	*
Land.Area	-54.06	29.28	-1.847	0.1073	

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

Residual standard error: 146836 on 7 degrees of freedom
Multiple R-squared: 0.6336, Adjusted R-Squared: 0.5289
F-statistic: 6.053 on 2 and 7 DF, p-value: 0.02977

Type II ANOVA Analysis

Response: Annual_Pawdacity_sales

	Sum Sq	DF	F value	Pr(>F)	
Households.with.Under.18	227060622780.56	1	10.53	0.01415	*
Land.Area	73519609307.61	1	3.41	0.1073	
Residuals	150924734788.78	7			

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

Land Area vs 2010 Census

Report

Report for Linear Model Linear_Regression_58

Basic Summary

Call:
lm(formula = Annual_Pawdacity_sales ~ X2010.Census + Land.Area, data = inputs$the.data)

Residuals:

Min	1Q	Median	3Q	Max
-165000	-28640	-9055	30210	120300

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	210859.24	69183.929	3.048	0.01864 *
X2010.Census	11.03	1.728	6.383	0.00037 ***
Land.Area	-30.23	17.444	-1.733	0.12674

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

Residual standard error: 88979 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: 0.0008931

Type II ANOVA Analysis

Response: Annual_Pawdacity_sales

	Sum Sq	DF	F value	Pr(>F)
X2010.Census	322565140615.9	1	40.74	0.00037 ***
Land.Area	23771530917.7	1	3	0.12674
Residuals	55420216953.45	7		

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

Land Area vs Population Density

Report

Report for Linear Model Linear_Regression_60

Basic Summary

Call:
lm(formula = Annual_Pawdacity_sales ~ Land.Area + Population.Density, data = inputs$the.data)

Residuals:

Min	1Q	Median	3Q	Max
-174000	-19140	15940	33000	137800

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.426e+05	89629.86	1.59073	0.1557
Land.Area	-4.829e-01	21.62	-0.02234	0.9828
Population.Density	3.191e+04	6095.31	5.23567	0.0012 **

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

Residual standard error: 104805 on 7 degrees of freedom
Multiple R-squared: 0.8134, Adjusted R-Squared: 0.76
F-statistic: 15.25 on 2 and 7 DF, p-value: 0.002809

Type II ANOVA Analysis

Response: Annual_Pawdacity_sales

	Sum Sq	DF	F value	Pr(>F)
Land.Area	5481963.41	1	0	0.9828
Population.Density	301096993203.78	1	27.41	0.0012 **
Residuals	76888364365.56	7		

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

Based on the above Linear Regression Models, Land Area and Total Families is the best model. Moreover, when we think about the model the size of the area and the number of families make a very good sense as two predictors because we normally associate pets with one family, not with one person. So, we would like to use the number families across the size of the area to predict the sales. Also, the model that uses these two variables produces the higher R-squared values because both variables have p-values that are lower than 0.05 within the model making them statistically significant. Variables that are highly correlated with each other like 2010 Census and Households with under 18 produce higher R-squared values, but that does not make the model better since using two variables that measure nearly the same thing does not make sense and considering the fact that Multicollinearity might skew the prediction results we ignore them. With all these considerations, Land Area and Total Families is the best model.

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)

   Sales = 197299.27 + 49.13*(Total Families) – 48.41*(Land Area)

Step 2: Analysis

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

   The city that is selected should satisfy the following conditions.

   - The new store should be located in a new city. That means there should be no existing stores in the new city.
   - The total sales for the entire competition in the new city should be less than $500,000
   - 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).
   - The predicted yearly sales must be over $200,000.
   - The city chosen has the highest predicted sales from the predicted set.

   Data was processed to match the above criteria. Though there were a couple of cities with the predicted sales over $200,000, as per the last criteria, the highest predicted sales from the Predicted set is for the city LARAMIE ($304994.15). So, the recommendation for Pawdacity to open its 14th store in a new city would be ***LARAMIE***.