Project 2.2: Recommend a City

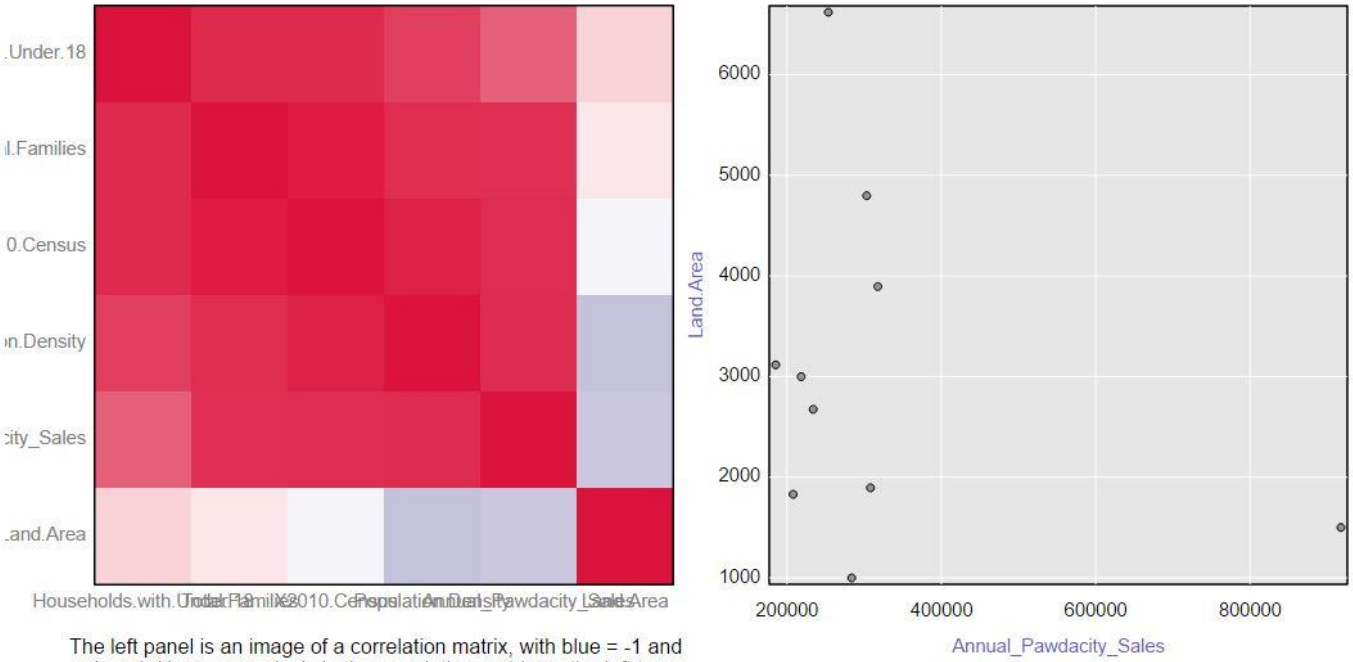
Step 1: Linear Regression

The target variable chosen is Annual_Pawdacity_Sales as this is the basis on which the new store location will be determined. The next step is to identify the predictor variables to build the regression model. The following variables are available for selection:

1. Land Area
2. Total Families
3. Households with Under 18
4. Population Density
5. 2010 Census

**2014 Estimate is not used as the target variable is 2010 sales so training a model on 2014 population estimates is not right. CITY variable is not considered as city name is not related to the sales.

After creating a correlation matrix, we find that Land Area is not highly correlated to the target variable. All the others i.e. Total Families, Households with Under 18, Population Density & 2010 Census are strongly correlated.



The left panel is an image of a correlation matrix, with blue = -1 and red = +1. Hover over pixels in the correlation matrix on the left to see the values; click to see the corresponding scatterplot on the right. The variables have been clustered based on degree of correlation, so that highly correlated variables appear adjacent to each other.

Hence, Land Area may be retained as one of the predictor variables. Now to determine which variable among the rest may be selected, the following steps are taken:

- 4 linear regression models can be created with land area as one variable one among the others as the other.

- So, 4 models with 2 predictor variables each are made and compared them with each other to see which one is the optimal model.
- The optimal model will be the one that has the highest adjusted R-squared value and all of its predictor variables will have significant p values.

1. Land Area + Total Families

Report for Linear Model P_2.2_Linear_Regression

Basic Summary

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

2. Land Area + 2010 Census

Report for Linear Model P_2.2_Linear_Regression

Basic Summary

Call:
lm(formula = Sales ~ X2010.Census + Land.Area, data = 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: 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

3. Land Area + Population Density

Report for Linear Model P_2.2_Linear_Regression

Basic Summary

Call:
lm(formula = Sales ~ Land.Area + Population.Density, data = 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: 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

4. Land Area + Households with Under 18

Report for Linear Model P_2.2_Linear_Regression

Basic Summary

Call:
lm(formula = Sales ~ Households.with.Under.18 + Land.Area, data = 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

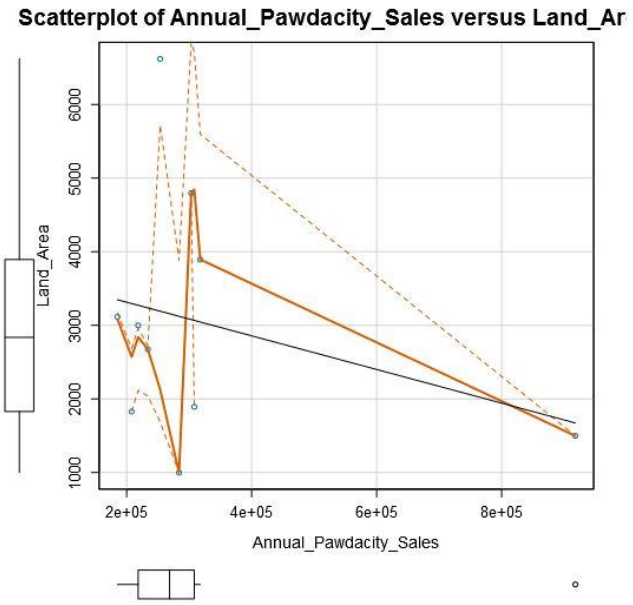
Response: Sales

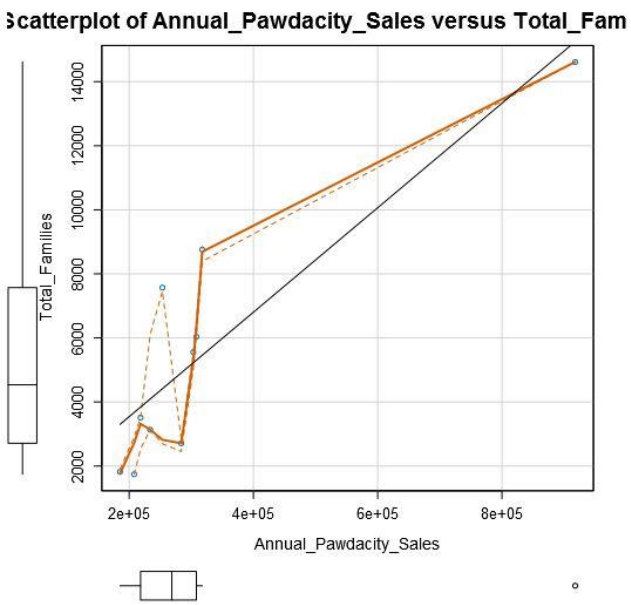
	Sum Sq	DF	F value	Pr(>F)	
Households.with.Under.18	227060622780.56	1	10.53	0.01415	*
Land.Area	73519609307.61	1	3.41	0.1073	
Residuals	150924734788.78	7			

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

As we can see, only the (Land Area + Total Families) model has R-square values above 0.9 and both the variables as significant. So it chosen as the optimal model.

Scatterplot analysis of the two predictor variables chosen i.e. Land Area & Total Families are linearly related to the target variable as shown below:





The best linear regression equation based on the available data is as follows:

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 following criteria's are given to choose the right city are:

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

   Accordingly, the data was processed and scored to determine the predicted sales in the new city using the linear regression model built up. After processing along these data only two new cities were found to have predicted sales > $200000 as follows

Record #	City	Predicted_Sales
1	Jackson	205555
2	Laramie	300929
3	Worland	205778

   Hence the new city where the 14th Pawdacity store should be opened is **LARAMIE**.