Project 2.2: Recommend a City

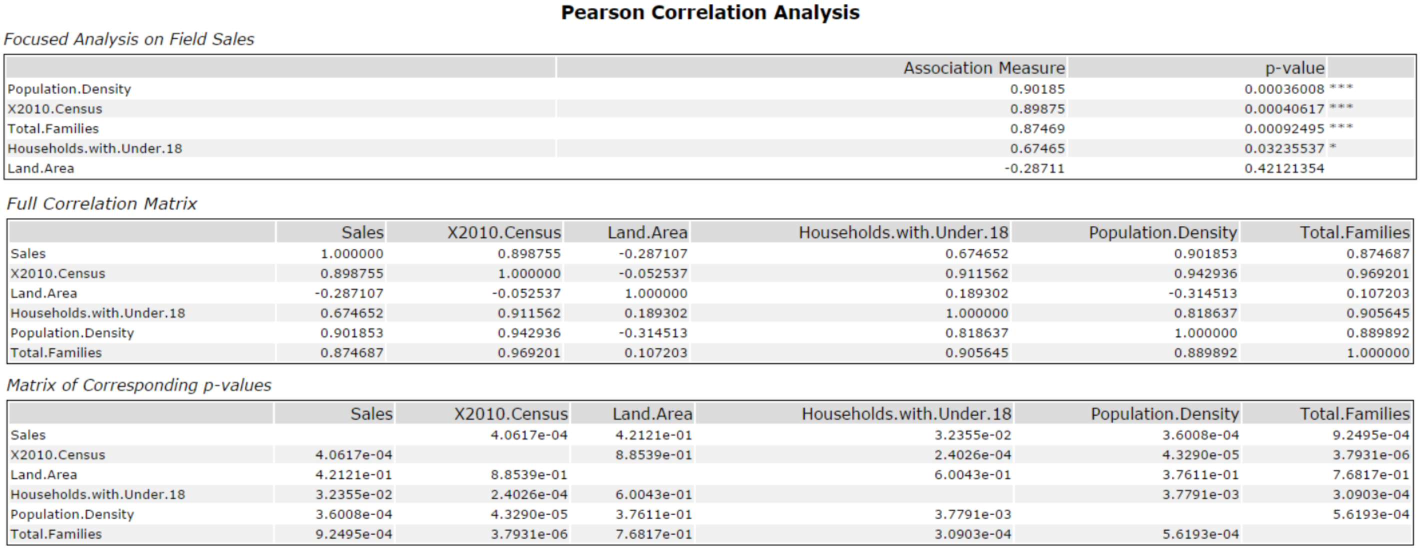
Complete each section. When you are ready, save your file as a PDF document and submit it here: https://classroom.udacity.com/nanodegrees/nd008/parts/3d606c26-cb8e-43af-9199-7e3577aa3392/project#

Note that this project is a continuation from Project 2.1: Data Cleanup. You must meet specifications for Project 2.1 before you can continue on with this Project 2.2

Step 1: Linear Regression

Create a linear regression model off your training set and present your model. Visualizations are highly encouraged in this section. (750 word limit)
At the minimum, answer these questions:

Having already created an analytical dataset, the relationship between all variables was examined using the Association Analysis tool on Alteryx. Particularly, focus was given to the relationship of the candidate predictor variables with the target variable (i.e. Sales).



Pearson Correlation Analysis

Focused Analysis on Field 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

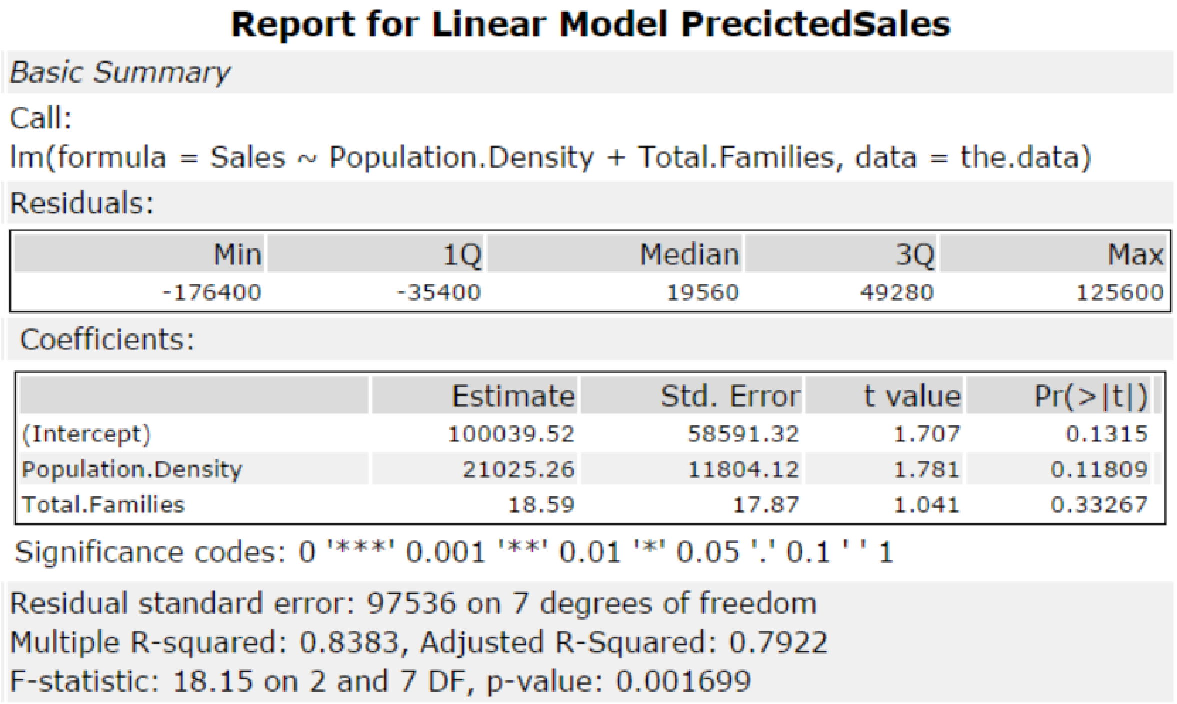
Full Correlation Matrix

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

Matrix of Corresponding p-values

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

Pearson Correlation analysis returns association measurements for two variables in vacuum, and although a model was built using Population Density and Total Families as predictor variables it was rejected because p-values were unacceptable and (adj.) R-squared relatively ok, please see below. Also, Population Density and 2010 Census are highly correlated so only one of them was used.



Report for Linear Model PrecictedSales

Basic Summary

Call:
lm(formula = Sales ~ Population.Density + Total.Families, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-176400	-35400	19560	49280	125600

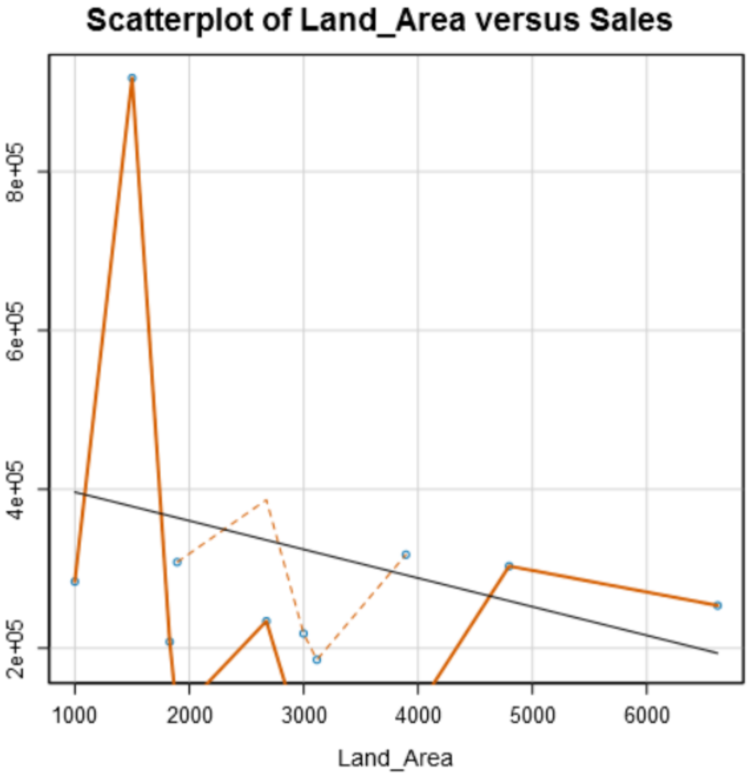
Coefficients:

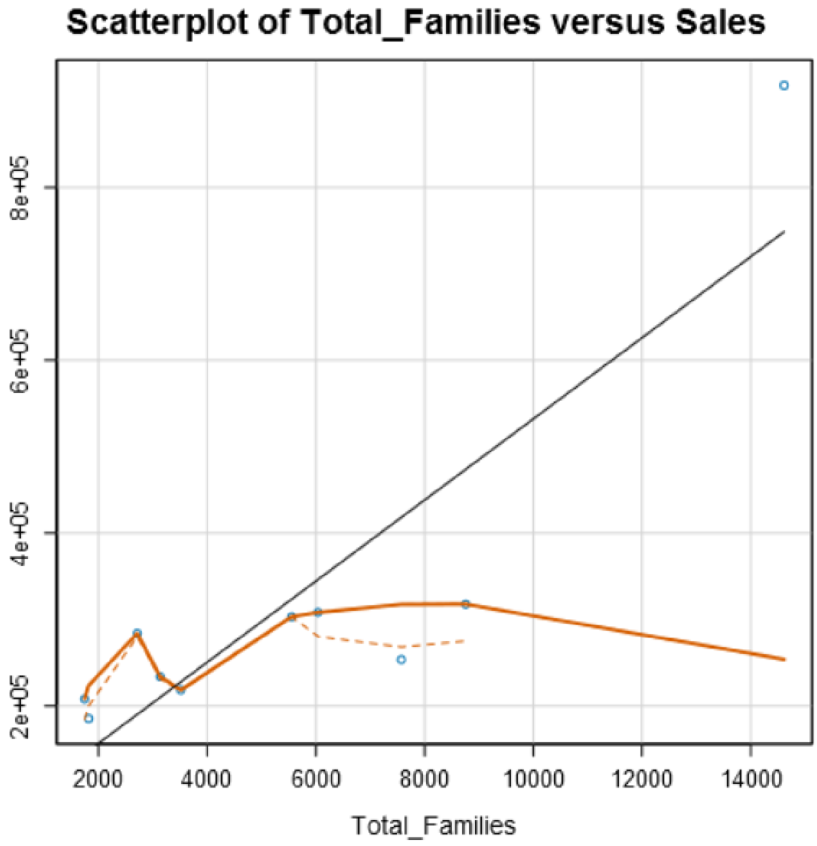
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	100039.52	58591.32	1.707	0.1315
Population.Density	21025.26	11804.12	1.781	0.11809
Total.Families	18.59	17.87	1.041	0.33267

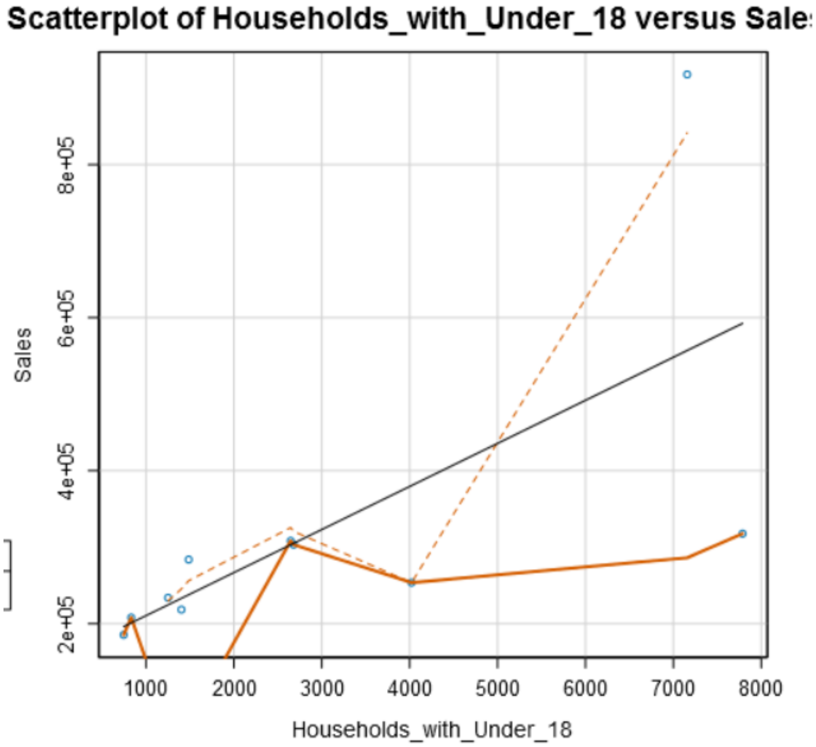
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 97536 on 7 degrees of freedom
Multiple R-squared: 0.8383, Adjusted R-Squared: 0.7922
F-statistic: 18.15 on 2 and 7 DF, p-value: 0.001699

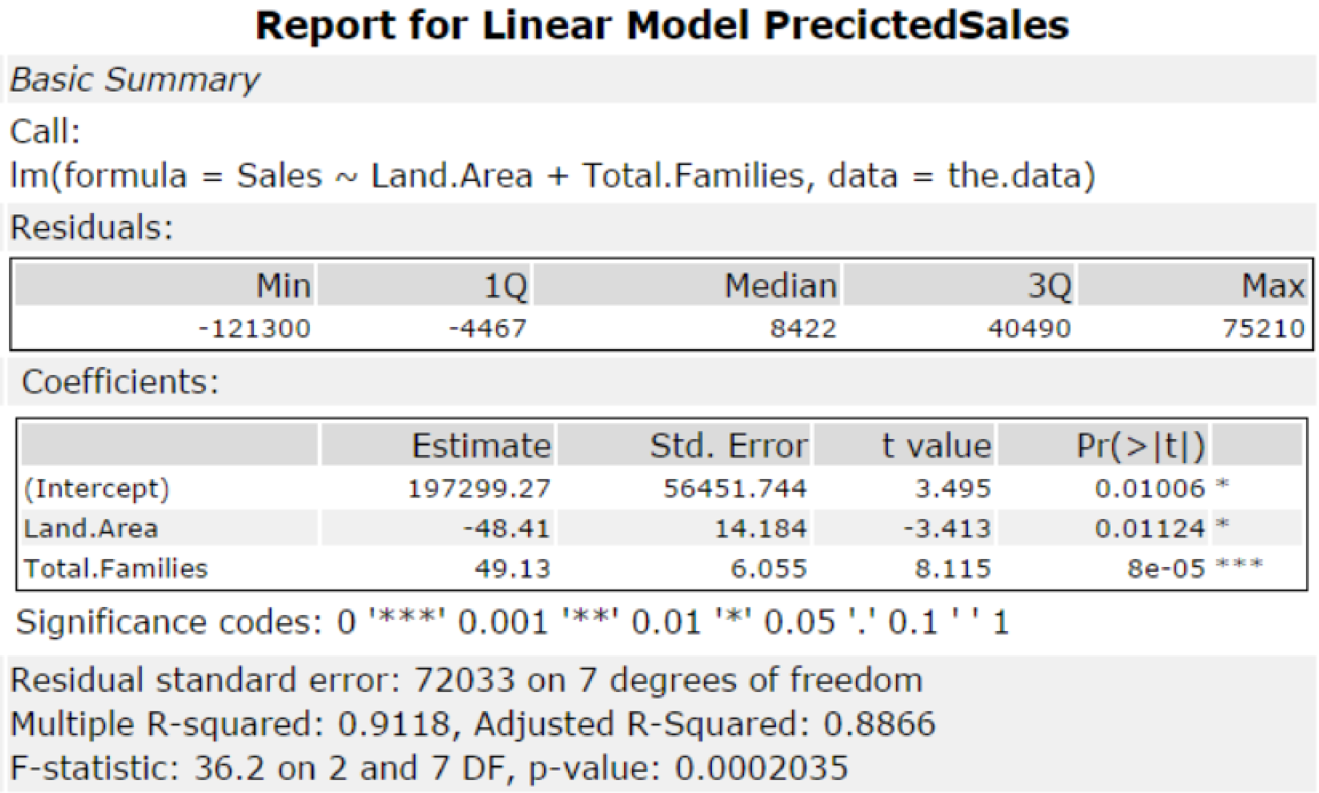
After this attempt, a new model was built based on Total Families and Land Area which does not appear to be correlated to Sales ‘in a vacuum’. First, a few scatterplots to show linear relationship.




Scatterplot of Total_Families versus Sales
2e+05
4e+05
6e+05
8e+05
2000
4000
6000
8000
10000
12000
14000
Total_Families


Scatterplot of Households_with_Under_18 versus Sale
Sales
2e+05
4e+05
6e+05
8e+05
1000
2000
3000
4000
5000
6000
7000
8000
Households_with_Under_18

More or less we can see there is a linear relationship between the candidate predictor variables and the target variable. HH under 18 was finally omitted as a predictor variable as it had a p-value greater than 0.05. The report is shown below, all variables achieve very low p-values and very good statistical significance levels as well as high adjusted R-square for the model overall, making them a good fit for the model.



Report for Linear Model PrecictedSales

Basic Summary

Call:
lm(formula = Sales ~ Land.Area + Total.Families, 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	*
Land.Area	-48.41	14.184	-3.413	0.01124	*
Total.Families	49.13	6.055	8.115	8e-05	***

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

The best linear equation to score the new cities for the pet store expansion is:

Predicted Sales = 197,299.27 – 48.41*[Land Area] + 49.13*[Total Families]

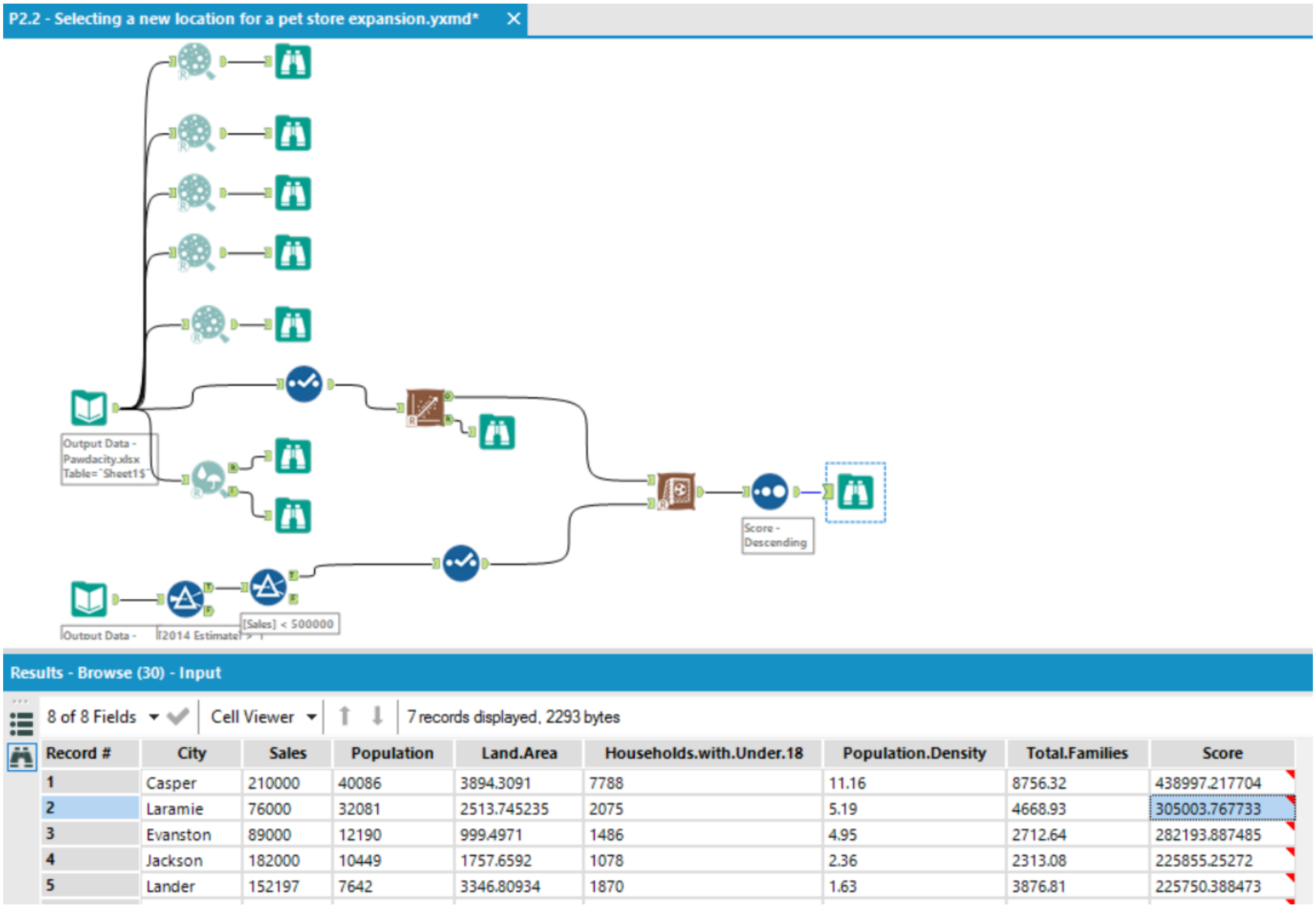
Step 2: Analysis

Use your model results to provide a recommendation. (500 word limit)

At the minimum, answer this question:

The scoring exercise indicates that the best option given the restriction imposed, e.g. no existing Pawdacity stores and population over 4,000, is **Laramie**. Predicted sales in Laramie are expected to be as high as $305,003.77 and this is the recommended city for our planned Pawdacity expansion.

Alteryx Workflow:



P2.2 - Selecting a new location for a pet store expansion.yxmd*

Results - Browse (30) - Input

8 of 8 Fields | Cell Viewer | 7 records displayed, 2293 bytes

Record #	City	Sales	Population	Land.Area	Households.with.Under.18	Population.Density	Total.Families	Score
1	Casper	210000	40086	3894.3091	7788	11.16	8756.32	438997.217704
2	Laramie	76000	32081	2513.745235	2075	5.19	4668.93	305003.767733
3	Evanston	89000	12190	999.4971	1486	4.95	2712.64	282193.887485
4	Jackson	182000	10449	1757.6592	1078	2.36	2313.08	225855.25272
5	Lander	152197	7642	3346.80934	1870	1.63	3876.81	225750.388473

Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.