

Project 2.1: Data Cleanup

Structure of Report

1. [Business and Data Understanding](#)
 - a. [Goal of project](#)
 - b. [Dataset overview](#)
 - b. [Key decision](#)
2. [Data Cleanup](#)
 - a. [Building the Training Set](#)
 - b. [Dealing with Outliers](#)
3. [Analysis, Modeling, and Validation](#)
 - a. [Analysis](#)
 - b. [Modeling](#)
 - c. [Validation](#)
4. [Conclusions](#)

1. Business and Data Understanding

a. Goal of project

- In this project, we will In the first part, we blend and format data, deal with outliers to make a clean dataset. After that, we use the cleaned up dataset to create a linear regression model. We have to choose which predictor variable(s) are the important for the regression model using Person correlation.
- In this project I use mainly Alteryx to deal with data, but sometime also Python.
- Scenario
 - 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 manager has asked to perform an analysis to recommend the city for Pawdacity's newest store, based on predicted yearly sales.
 - Given:
 - The monthly sales data for all of the Pawdacity stores for the year 2010.
 - A partially parsed data file that can be used for population numbers.
 - Demographic data (Households with individuals under 18 , Land Area , Population Density , and Total Families) for each city and county in the state of Wyoming. ##### b. Datasets overview
 - p2-2010-pawdacity-monthly-sales.csv :This file contains all of the monthly sales for all Pawdacity stores for 2010.
 - p2-partially-parsed-wy-web-scrape.csv : This is a partially parsed data file that can be used for population numbers.
 - p2-wy-demographic-data.csv : This file contains demographic data for each city and county in Wyoming. ##### c. Key Decisions:
- What decisions needs to be made?

We need to find out which city should be a new place to open new pet store, based on yearly sales.

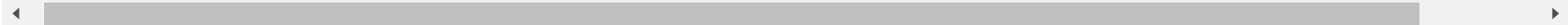
- What data is needed to inform those decisions?

We need consider some predictor variables such as Land Area , Population Density , and Total Families in order to predict revenues for stores expected opening at cities. After that we will decide which city will be choosen to open store.

2. Data Cleanup

a. Building the Training Set

	NAME	ADDRESS	CITY	STATE	ZIP	January	February	March	April	May	June	July	August	September	October	November
0	Pawdacity	509 Fort St # A	Buffalo	WY	82834	16200	13392	14688	17064	18360	14040	12960	19224	15984	13392	13176
1	Pawdacity	601 SE Wyoming Blvd Unit 252	Casper	WY	82609	29160	21600	27000	27648	29160	27216	25488	25704	22896	25272	28944
2	Pawdacity	3769 E Lincolnway	Cheyenne	WY	82001	79920	70632	79056	77544	73656	77976	73872	77544	78516	74520	74736
3	Pawdacity	2625 Big Horn Ave	Cody	WY	82414	19440	15984	19008	18144	16632	17496	18792	20304	19224	18144	18576
4	Pawdacity	123 S 2nd St	Douglas	WY	82633	16200	13392	14688	17064	18360	14040	12960	19224	15984	29808	17496



We need to check Dataset `p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales.csv` to ensure the data type suitable to . In addition we need to create a new column `Total_Sale` in order to make analysis based on predicted yearly sales.

Dataset 2 `p2-partially-parsed-wy-web-scrape.csv` :

	City County	2014 Estimate	2010 Census	2000 Census
0	Afton Lincoln	<td>1,968</td>	<td>1,911</td>	<td>1,818</td>
1	Albin Laramie	<td>185</td>	<td>181</td>	<td>120</td>
2	Alpine Lincoln	<td>845</td>	<td>828</td>	<td>550</td>
3	Baggs Carbon	<td>439</td>	<td>440</td>	<td>348</td>
4	Bairoil Sweetwater	<td>107</td>	<td>106</td>	<td>97</td>

Dataset `p2-partially-parsed-wy-web-scrape.csv` has some obvious problems with data, we need to parse column `City|County` into `City` and `County`. All of characters such as `<td>` and `</td>` or `,` need to be delete. Data type need to be suitable to make analysis.

Dataset 3 `p2-wy-demographic-data.csv`

	City	County	Land Area	Households with Under 18	Population Density	Total Families
0	Laramie	Albany	2513.745235	2075	5.19	4668.93
1	Rock River	Albany	200.444000	165	0.41	372.30
2	Basin	Big Horn	543.951304	250	0.66	566.43
3	Burlington	Big Horn	137.646214	63	0.17	143.34
4	Byron	Big Horn	252.489592	116	0.31	262.93

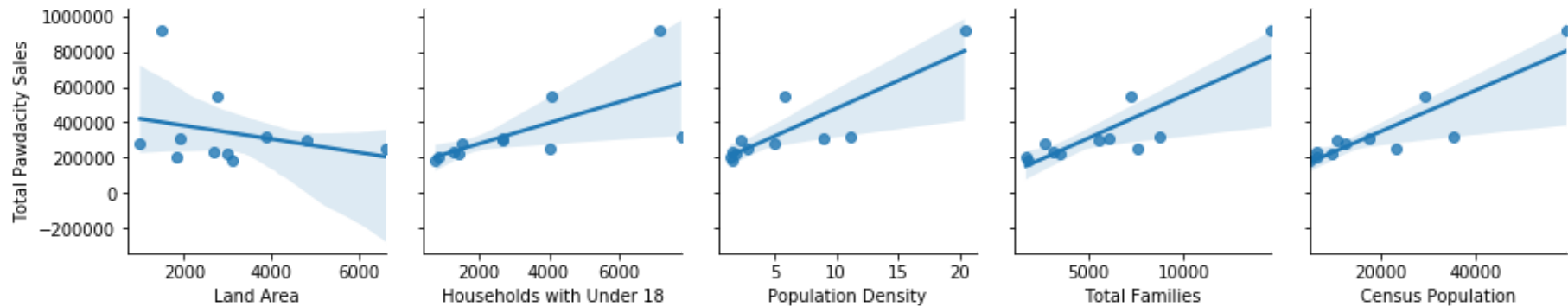
Dataset 3, generally is good. But we still need to make all data in exact types in order to make analysis easily and no mistakes occurred.

After processing data at the first step, we will have a raw dataset with the `sum` and `average` of all numeric columns like:

	sum	average
Land Area	33071.38	3006.49
Households with Under 18	34064.00	3096.73
Population Density	62.80	5.71
Total Families	62652.79	5695.71
Census Population	213862.00	19442.00
Total Pawdacity Sales	3773304.00	343027.64

b. Dealing with Outliers

The pairplot with kind="reg" following shows the linear relationship between variables in the dataset. With the plot we can see how good is the linear relationship between the predictor variables and target variable. Also, how is the outliers in the dataset can be seen in the plot. The target variable is Total Pawdacity Sales and predictor variables are Land Area , Households with Under 18 , Population Density , Total Families , Census Population



We use the method IQR to identify the outlier. We have:

- `Q1 = df.quantile(0.25)`
- `Q3 = df.quantile(0.75)`
- `IQR = Q3-Q1`

- So the upper bound = $Q3 + 1.5 \cdot IQR$
- And the low bound = $Q1 - 1.5 \cdot IQR$

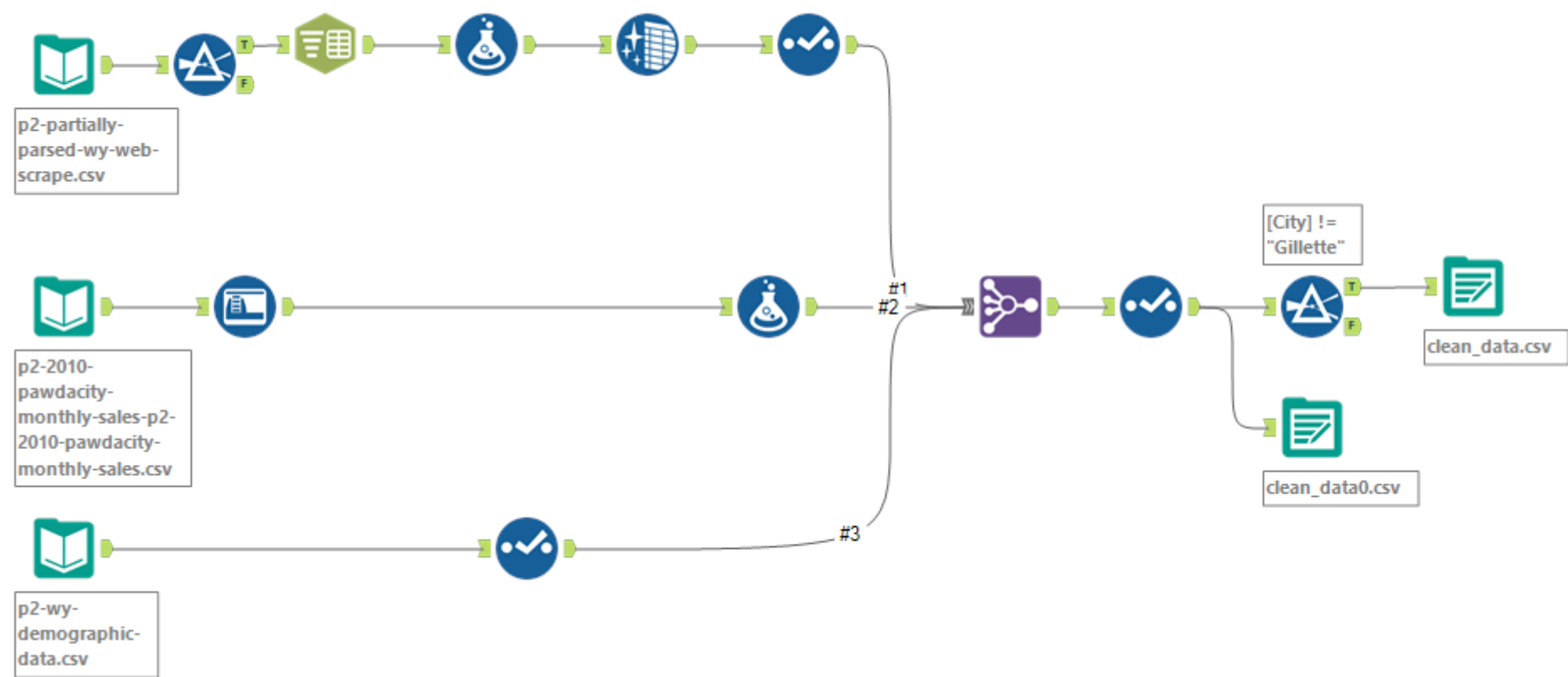
So the point outside the upper bound and lower bound will be considered as outliers. Using such method we will have outliers like following:

	NAME	City	County	ADDRESS	Land Area	Households with Under 18	Population Density	Total Families	Census Population	Total Pawdacity Sales
2	Pawdacity	Cheyenne	Laramie	3769 E Lincolnway	1500.178400	7158	20.34	14612.64	59466	917892
6	Pawdacity	Gillette	Campbell	200 E Lakeway Rd	2748.852900	4052	5.80	7189.43	29087	543132
9	Pawdacity	Rock Springs	Sweetwater	2706 Commercial Way	6620.201916	4022	2.78	7572.18	23036	253584

The problem is that here we dealt with a very small dataset, when we delete or remove all outliers, the dataset will be changed alot and maybe cause a bad analysis as well as a wrong prediction. So we need to consider only one outlier need to be remove or impute.

All in all, `Gillette` seems to be outside the regression model of dataset and the relationship between it and other values is unclear. Although `Cheyenne` and `Rock Springs` are also outliers because in some columns its values are not normal, but in general, its relation with other variables can be explained. So we will impute `Gillette` out of dataset.

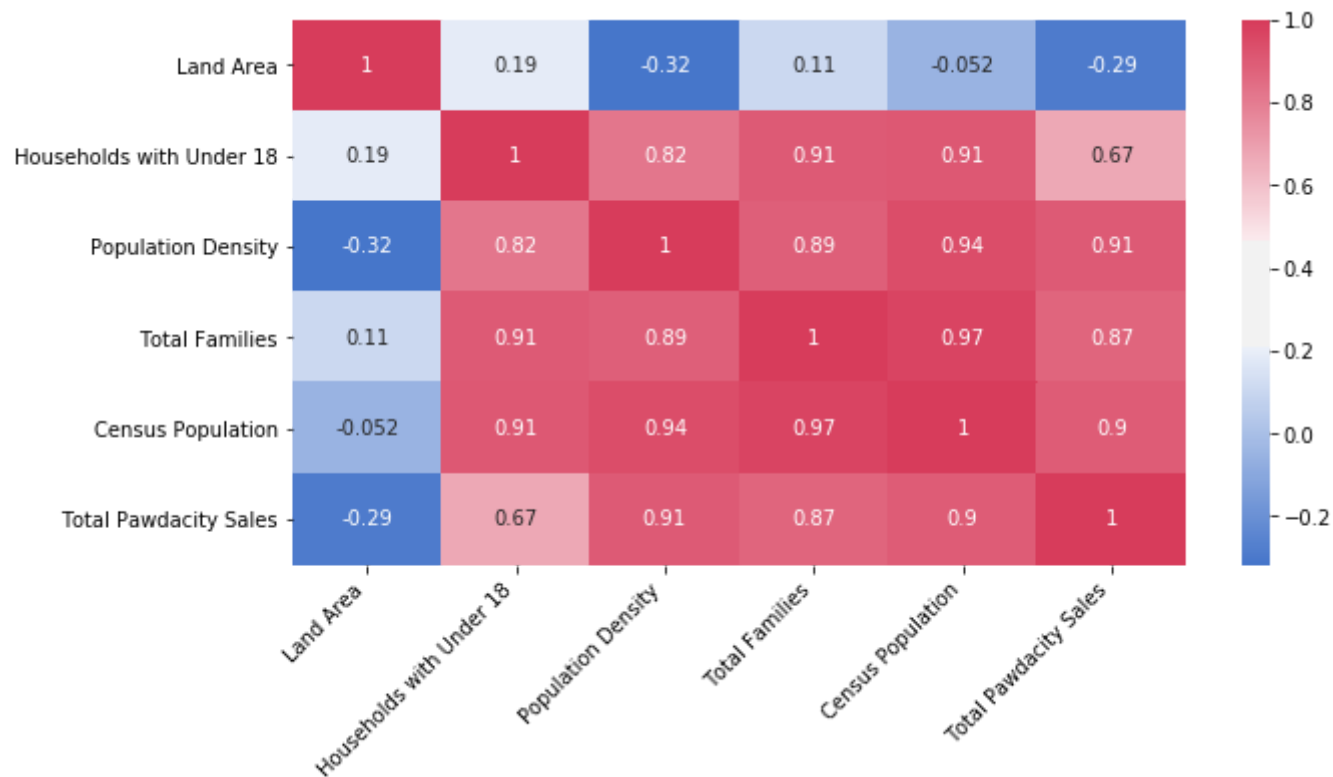
We use Alteryx to clean the data. The workflow is following:



3. Analysis, Modeling, and Validation

a. Analysis

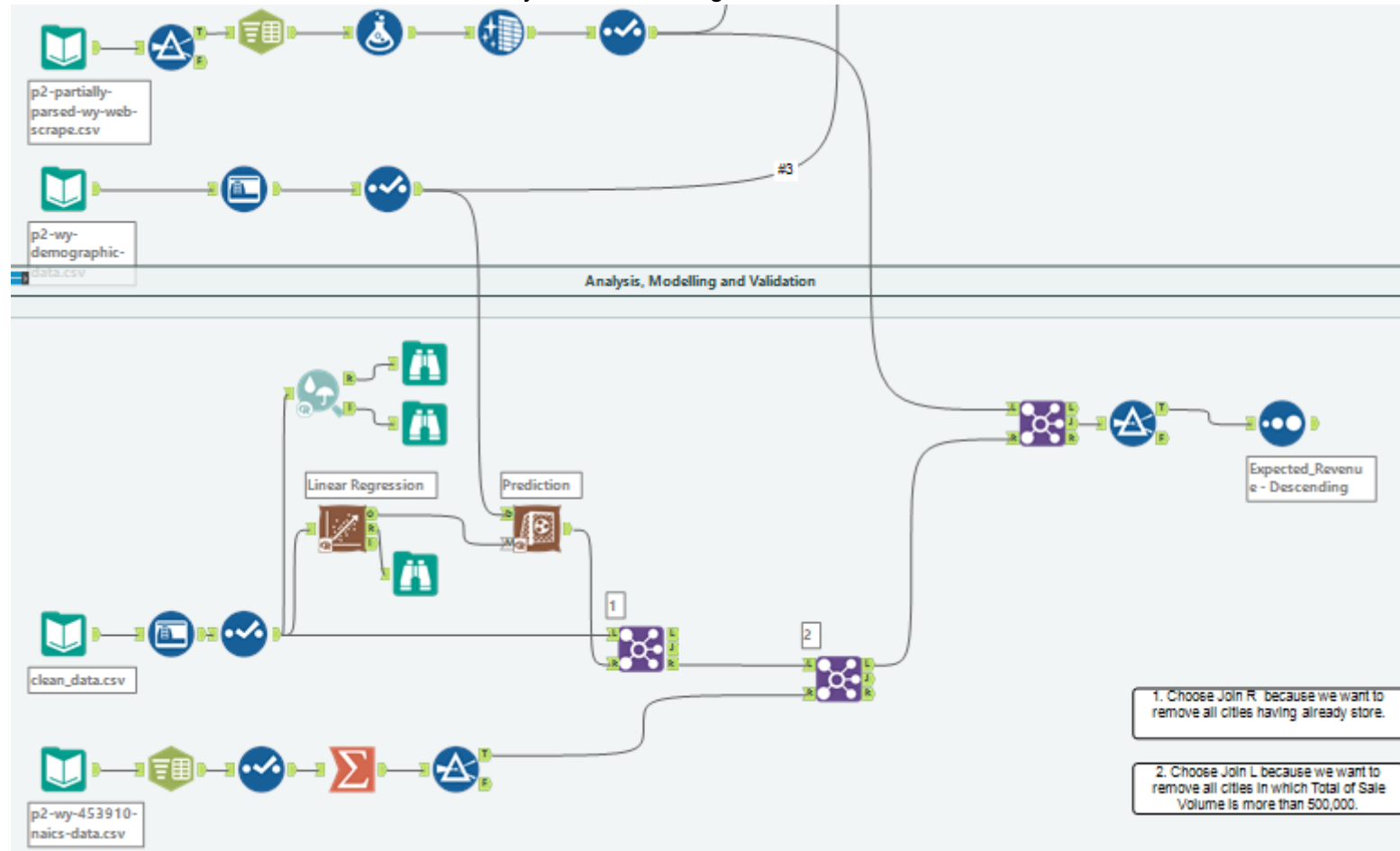
- There are some conditions from manager for choosing a city to open new store:
 - The new store should be located in a new city. That means there should be no existing stores in the new city.
 - The total sales for the entire competition in the new city should be less than \$500,000
 - The new city 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.
- At first, we need to find important predictor variables to use in the regression model with `Total Pawdacity Sales` as the target variable. The following is the heatmap to see the relationships between two variables.



- With the heatmap, it is clear that Land Area has a weak correlation with other variables. It is good to choose it as 1 predictor variable.
- For all others, we can see that all have almost the same correlation. Logically, we can see that Households with Under 18 and Total Families are 1 variable. We will choose Total Families as 1 predictor variable because it includes information of Households with Under 18. In the comparison with other variable, the Total Families gains a better influence on Total Pawdacity Sales so we choose it as the second predictor variable in the linear regression model.
- In conclusion, we will do the linear regression to predict Total Pawdacity Sales with 2 variables: Land Area & Total Families

b. Modeling

After choosing predictor variables to the model. We will run Alteryx and have a regression liner:



- The linear regression will be:

$$[\text{Total Pawdacity Sales}] = 197,330 - 48.42 [\text{Land Area}] + 49.14 [\text{Total Families}]$$

c. Validation

Record	Report																								
1	Report for Linear Model Linear_Regression_78																								
2	<i>Basic Summary</i>																								
3	Call: lm(formula = Total.Pawdacity.Sales ~ Land.Area + Total.Families, data = the.data)																								
4	Residuals:																								
5	<table><tr><th>Min</th><th>1Q</th><th>Median</th><th>3Q</th><th>Max</th></tr><tr><td>-121261</td><td>-4453</td><td>8418</td><td>40491</td><td>75205</td></tr></table>	Min	1Q	Median	3Q	Max	-121261	-4453	8418	40491	75205														
Min	1Q	Median	3Q	Max																					
-121261	-4453	8418	40491	75205																					
6	Coefficients:																								
7	<table><tr><th></th><th>Estimate</th><th>Std. Error</th><th>t value</th><th>Pr(> t)</th><th></th></tr><tr><td>(Intercept)</td><td>197330.41</td><td>56449.000</td><td>3.496</td><td>0.01005</td><td>*</td></tr><tr><td>Land.Area</td><td>-48.42</td><td>14.184</td><td>-3.414</td><td>0.01123</td><td>*</td></tr><tr><td>Total.Families</td><td>49.14</td><td>6.055</td><td>8.115</td><td>8e-05</td><td>****</td></tr></table> <p>Significance codes: 0 '****' 0.001 '***' 0.01 '**' 0.05 '.' 0.1 ' ' 1</p>		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	****
	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	****																				
8	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 degrees of freedom (DF), p-value 0.0002035																								

With the report, we can see that linear regression is quite good to predict the revenue of expected stores because all of predictor variables are statistical significant with $p_value < 0.05$. In addition Adjusted R-Square = 0.8866 shows that the regression model is also good.

4.Conclusion:

With all given information, we will have the top 10 expected result in the following table:

	City	County	Land.Area	Population.Density	Households.with.Under.18	Total.Families	Census Population	2014 Estimate	Sum_Sale Volume of Rivals	Expected_Rever
0	Laramie	Albany	2513.745235	5.19	2075	4668.93	30816	32081	76000.0	305013.8816
1	Jackson	Teton	1757.659200	2.36	1078	2313.08	9577	10449	182000.0	225870.8236
2	Lander	Fremont	3346.809340	1.63	1870	3876.81	7487	7642	152197.0	225751.4002
3	Green River	Sweetwater	3477.361206	1.46	2113	3977.40	12515	12630	NaN	224372.0015
4	Worland	Washakie	1294.105755	2.18	595	1364.32	5487	5366	169000.0	201700.3259
5	Rawlins	Carbon	5322.661628	1.32	1307	2722.43	9259	9227	NaN	73349.5674

The Table shows the cities which are statisfied all conditions from manager to open new store. In details, inthese cities, the total sales for the entire competition is less than \$500,000, with a population over 4,000 people (based upon the 2014 US Census estimate), without existing stores.

Recomendation:

So the city which should be chosen to open new store is **Laramie in Albany** with population based on 2014 Estimate of 32,081 and the highest Expected_Revenue of \$305,013.88.