PROJECT: RECOMMEND A CITY BASED ON PREDICTED YEARLY SALES

PART 1: DATA CLEANUP

Step 1: Business and Data Understanding

Key Decisions:

1) **What is the Business Problem?**

- Pawdacity is the leading pet store chain in Wyoming with 13 stores throughout the state. This year Pawdacity would like to expand and open a 14th store. I have to perform an analysis to recommend the city for Pawdacity's newest store, based on predicted yearly sales.

2) **What decision needs to be made?**

- According to the business problem, we have to decide the 'Perfect location/city to open the Pawdacity's newest store location.

3) **What data is needed to inform the decisions?**

- To support a decision we need to predict sales of all the other cities where store has not opened yet by looking at the data of previous cities where store has already been opened and running well. Predicting sales depend on various factors like Pawdacity's past sales data, customer's demographics like, land area, total population, total families etc. Before solving any business problem, data gathering and data cleansing is required. We have relevant data, let's dive into data cleansing and data blending operations.

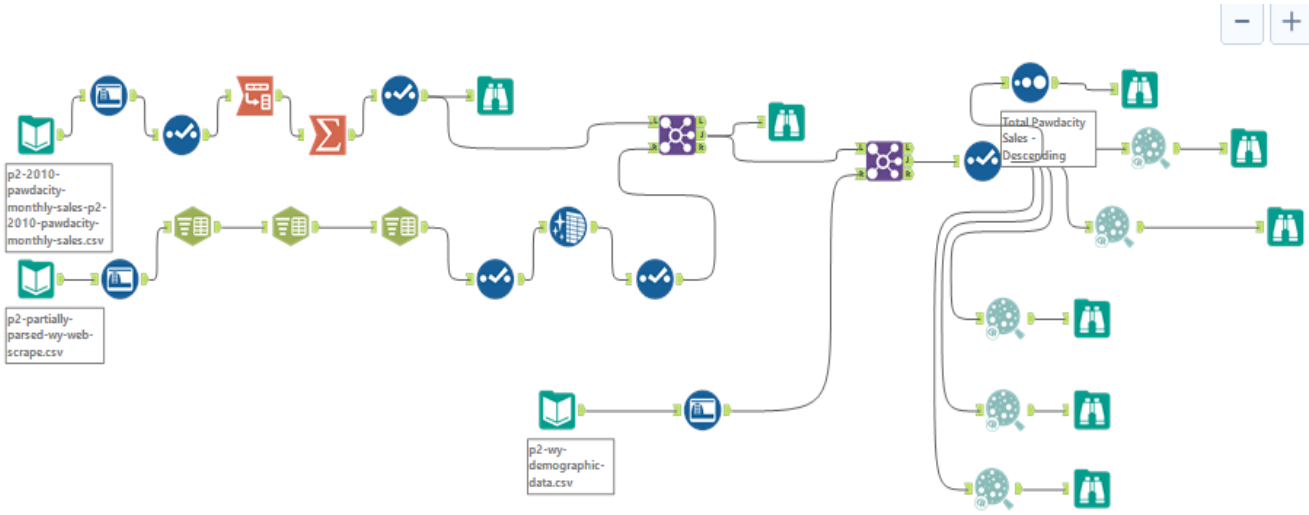
Step 2: Building the Training Set

We have 3 different data files named:

- Pawdacity's monthly sales of current stores in different cities.
- Pawdacity's web scrapped half parsed census population data.
- Pawdacity's demographic data.

Now, our objective is to solve all the data issues, cleansed it and merge all the data in 1 consolidated excel file. So, using Alteryx, I have performed data cleansing, data formatting and data blending and output it in new data file.

Below is the workflow provided, depicting the entire procedure.



Analysis:

COLUMN	SUM	AVERAGE
Census Population	213862	19442
Total Pawdacity Sales	3773304	343028
Households with under 18	34064	3097
Land Area	33071	3006
Population Density	63	6
Total Families	62653	5696

Step 3: Dealing with Outliers

Excel Analysis

CITY	2010 Census Population	Total Pawdacity Sales	Households with Under 18	Land Area	Population Density	Total Families
Cheyenne	59466	917892	7158	1500	20	14613
Gillette	29087	543132	4052	2749	6	7189
Casper	35316	317736	7788	3894	11	8756
Sheridan	17444	308232	2646	1894	9	6040
Riverton	10615	303264	2680	4797	2	5556
Evanston	12359	283824	1486	999	5	2713
Rock Springs	23036	253584	4022	6620	3	7572
Powell	6314	233928	1251	2674	2	3134
Cody	9520	218376	1403	2999	2	3516
Douglas	6120	208008	832	1829	1	1744
Buffalo	4585	185328	746	3116	2	1820
Average	**19442**	**343028**	**3097**	**3006**	**6**	**5696**
Total	**213862**	**3773304**	**34064**	**33071**	**63**	**62653**
Q1	9520	233928	1403	1894	2	3134
Q2	12359	283824	2646	2749	3	5556
Q3	26061.5	312984	4037	3505	7.5	7380.5
IQR	16541.5	79056	2634	1611	5.5	4246.5
Upper fence	50873.75	431568	7988	5921.5	15.75	13750.25
lower fence	-15292.25	115344	-2548	-522.5	-6.25	-3235.75

There are 3 cities **Cheyenne**, **Gillette**, and **Rock Springs** which can be considered as an outliers. So, as per my analysis, I will choose **Gillette** as an outlier because it skews high in sales and not in line with the linear relationship with other parameters.

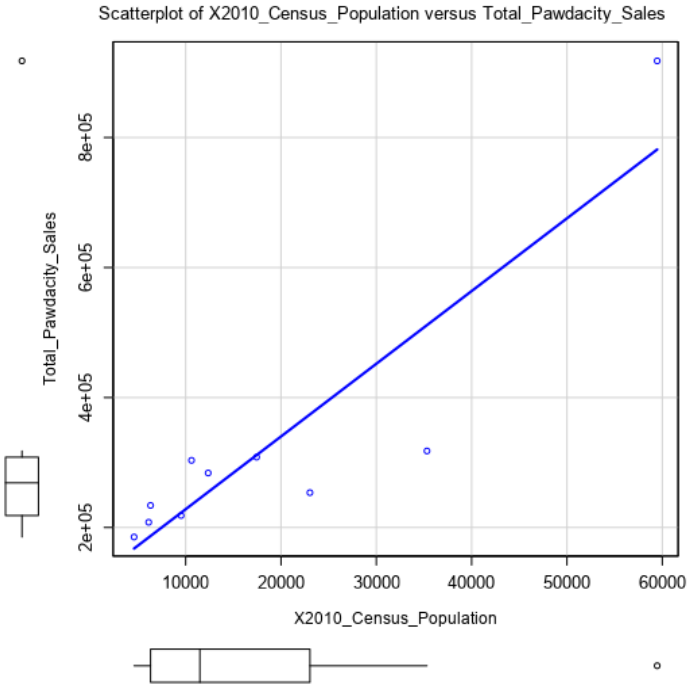
Whereas retaining **Cheyenne** can help in building robust model while comparing with the cities with larger population. Also, this is in line with linear relationship with other parameters, so keeping this in our data will not affect our analysis.

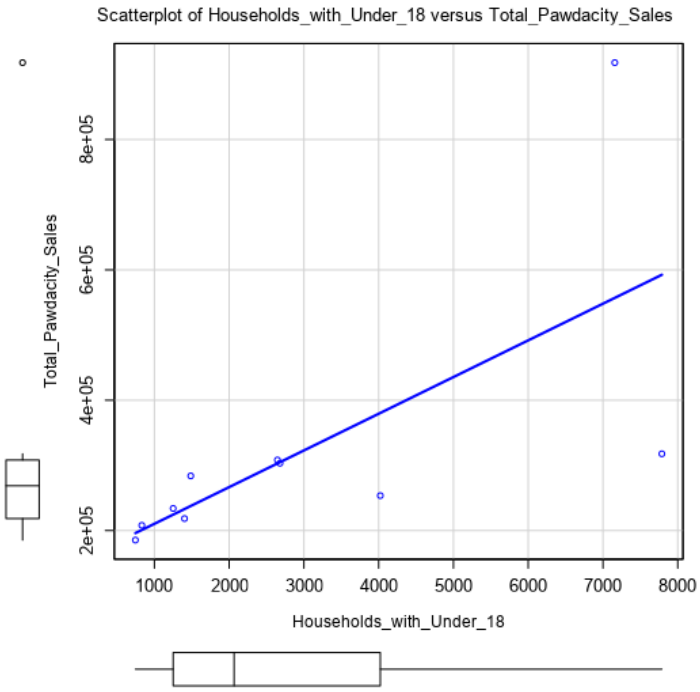
Since, we have small dataset so removing more than 1 outlier may result in inaccurate analysis.

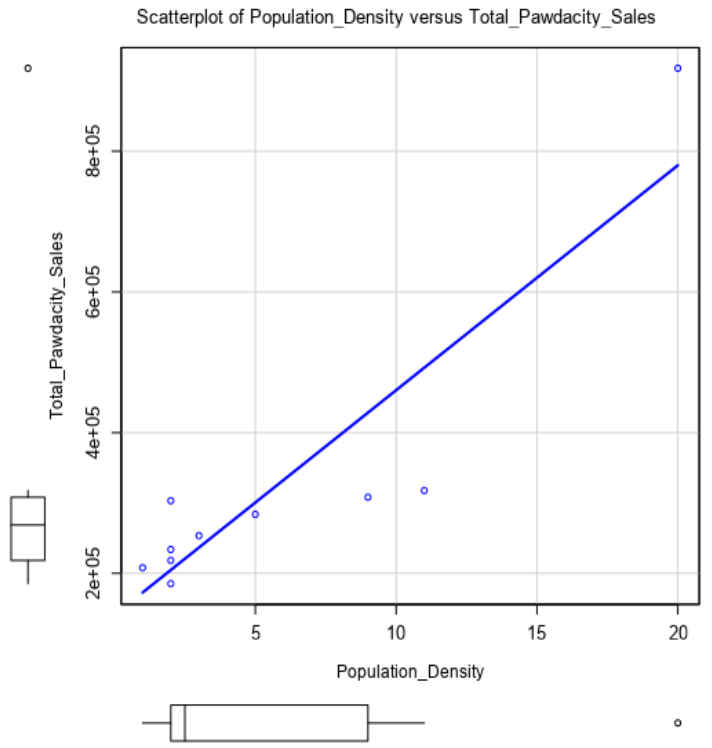
PART 2: RECOMMEND A CITY

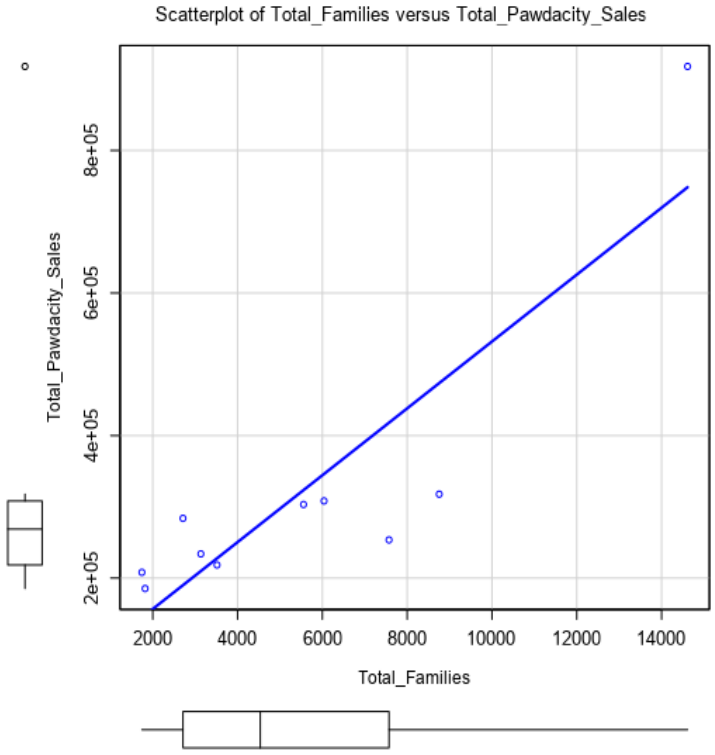
Step 4: Selecting Predictor Variables

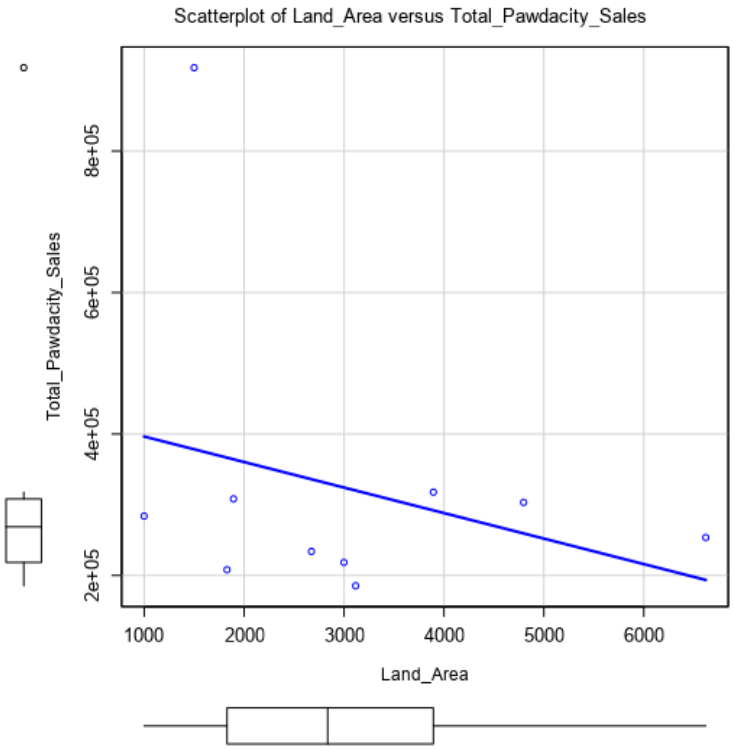
I first plotted each predictor variable against the target variable to check the correlation between them.






Scatterplot of Population_Density versus Total_Pawdacity_Sales
Total_Pawdacity_Sales
2e+05
4e+05
6e+05
8e+05
5
10
15
20
Population_Density


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


Scatterplot of Land_Area versus Total_Pawdacity_Sales
Total_Pawdacity_Sales
2e+05
4e+05
6e+05
8e+05
1000
2000
3000
4000
5000
6000
Land_Area

I can conclude all predictor variables are good potential predictor variables because they show a linear relationship between sales. I checked for correlations between my predictor variables to see if there is any possibility of multicollinearity in my dataset. Below is a table that shows the correlations between the different predictor variables:

Full Correlation Matrix

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

We can see that HHU18, Census, Families, and PDensity (Population Density) have strong correlations which each other. Land area however, is not as highly correlated. So I started by using land area as one predictor and then tested the four variables that are correlated.

I've found out that using land area and total families as the predictor variables produced the best model.

Basic Summary

Call:
lm(formula = Total.Pawdacity.Sales ~ Land_Area + Total_Families, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-121260	-4467	8422	40490	75208

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 degrees of freedom (DF), p-value 0.0002035

Reason why this model is good:

The p-values for land area and total families are both below 0.05 and the Multiple R-squared value is at .91 which is close to 1. This is model is a decent model.

Best Linear Regression Equation based on available data:

Y = 197299.27 – 48.41 * [Land_Area] + 49.13 * [Total_Families]

Step 5: ANALYSIS

Data cleaning and aggregation steps using Alteryx

I started with the Web Scraped Data from the Wyoming Wikipedia page, and used text to columns and select tools and the Data Cleansing to parse out the City, County, 2010 Census, and Estimate and remove all of the extra punctuation.

For the demographic data, I used the Auto-field tool to combine all of the numbers labeled as String fields. Before each join, I summarized the amounts by city to ensure that there were no duplicate city names within the data.

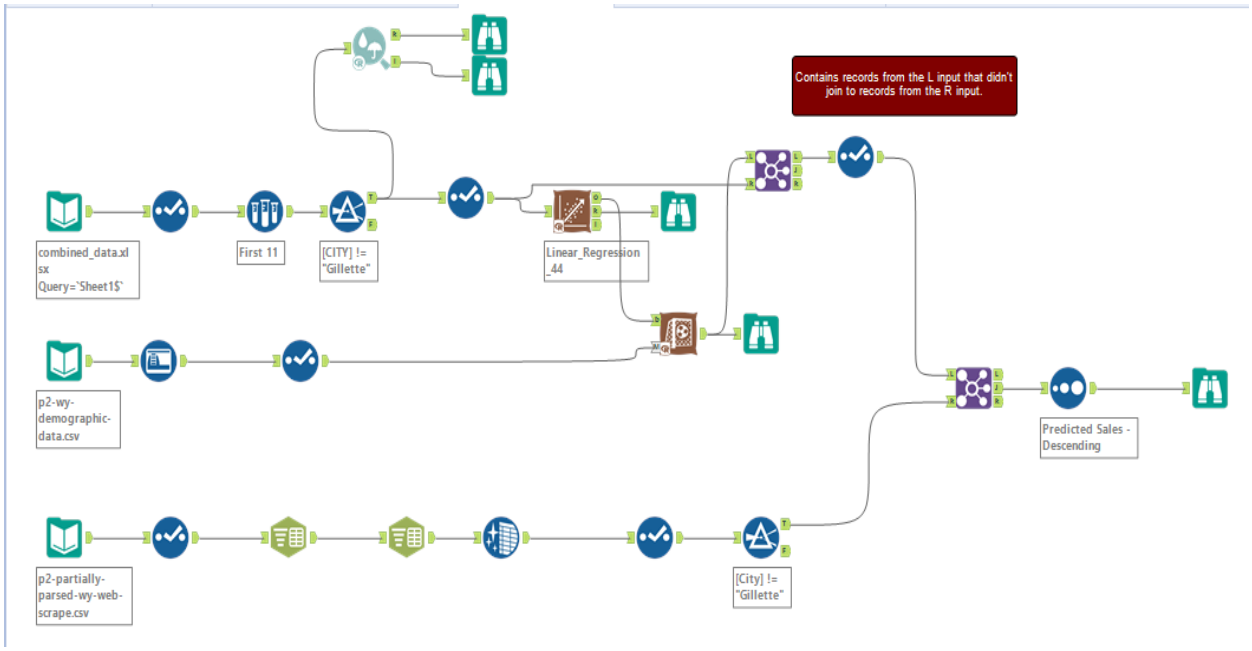
For Pawdacity sales file, I transposed the data to get City, Month, and Amount, and then summarized by City to get the total amount for each city.

From there, I created my data set and named it as combined_data.xlsx. Later I used this file to train my regression model.

Once the model was created, I applied the model to the demographic data file that contain cities other than those which are in Pawdacity Sales file.

Using score tool I predicted the sales. I took its output and joined to the ‘L’ input of join tool and ‘R’ input of join tool with combined data output. This excluded the cities where stores have already been opened.

I then applied the filters laid out in the project plan to come up with my list of possible cities, and sorted on the expected revenue to bring the best choice to the top.



Which City would I recommend?

After thorough analysis, I would recommend the city of **Laramie** with the predicted sales of **$305004**.