Business and Data Understanding

What decisions need to be made?

Perform an analysis to recommend the city for Pawdacity's newest store

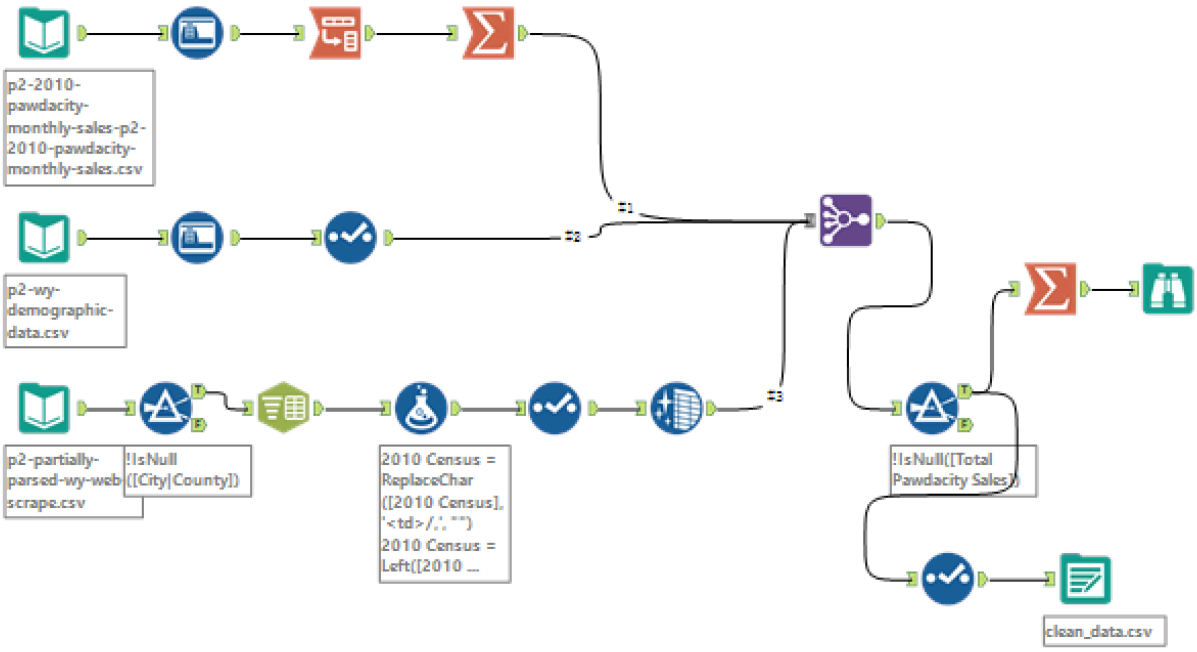
What data is needed to inform those decisions?

We need to predict the yearly sales for each city, and it can be calculated from Past Sales Data, Demographic Data of the cities, Population data, and the sales of competitor stores.

Analysis, Modeling, and Validation

Building the Training Set

Data cleaning



For Pawdacity sales, transpose tool is used to get *month* and *amount* field, then summarize by *city*.

For demographic data, select tool is used to pick fields we need.

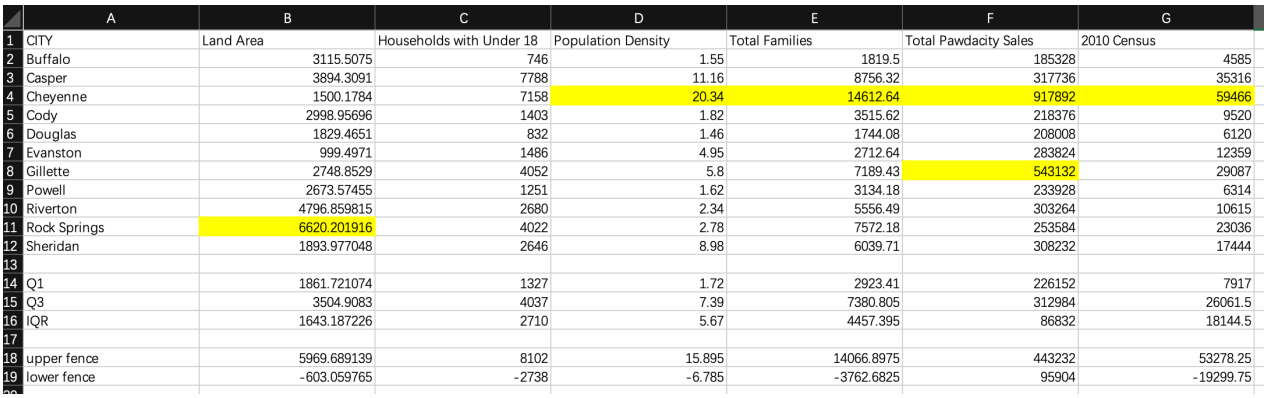
For population data, *text_to_column*, *left* , *replace* and *replacechar* function are used for data cleaning.

Then I use *joint_multiple* to join above three files, then save the data as *clean_ data.csv*.

For work check:

Column	Sum
Census Population	213862
Total Pawdacity Sales	3773304
Households with Under 18	34064
Land Area	33071
Population Density	62.8
Total Families	62653

Dealing with outliers

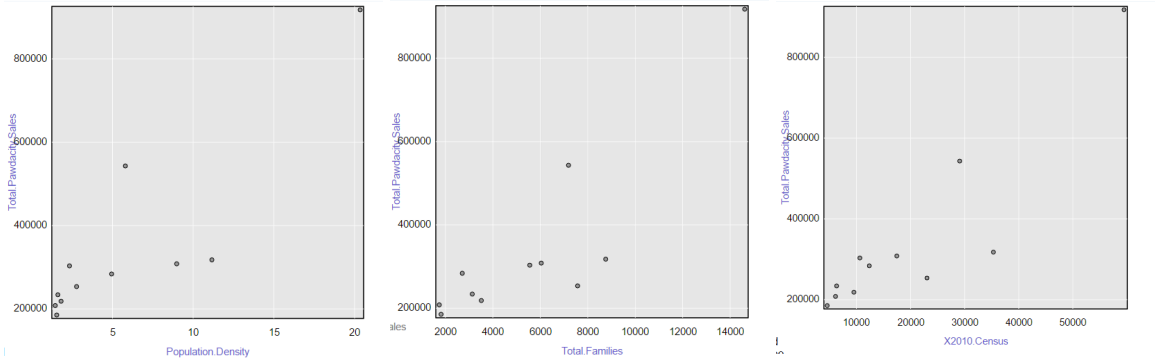


	A	B	C	D	E	F	G
1	CITY	Land Area	Households with Under 18	Population Density	Total Families	Total Pawdacity Sales	2010 Census
2	Buffalo	3115.5075	746	1.55	1819.5	185328	4585
3	Casper	3894.3091	7788	11.16	8756.32	317736	35316
4	Cheyenne	1500.1784	7158	20.34	14612.64	917892	59466
5	Cody	2998.95696	1403	1.82	3515.62	218376	9520
6	Douglas	1829.4651	832	1.46	1744.08	208008	6120
7	Evanston	999.4971	1486	4.95	2712.64	283824	12359
8	Gillette	2748.8529	4052	5.8	7189.43	543132	29087
9	Powell	2673.57455	1251	1.62	3134.18	233928	6314
10	Riverton	4796.859815	2680	2.34	5556.49	303264	10615
11	Rock Springs	6620.201916	4022	2.78	7572.18	253584	23036
12	Sheridan	1893.977048	2646	8.98	6039.71	308232	17444
13							
14	Q1	1861.721074	1327	1.72	2923.41	226152	7917
15	Q3	3504.9083	4037	7.39	7380.805	312984	26061.5
16	IQR	1643.187226	2710	5.67	4457.395	86832	18144.5
17							
18	upper fence	5969.689139	8102	15.895	14066.8975	443232	53278.25
19	lower fence	-603.059765	-2738	-6.785	-3762.6825	95904	-19299.75

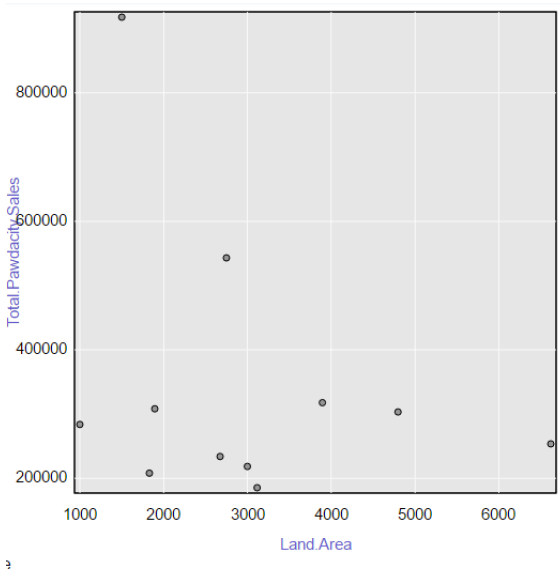
QUARTILE in Excel is used to identify outliers with IQR, numbers higher than the upper limit are highlighted in the table.

```
## We can also do it in python
# calculate quartile
first_quartile = data['field'].describe()['25%']
third_quartile = data['field'].describe()['75%']

# calculate IQR
iqr = third_quartile -first_quartile

# remove ouliters(if you are certain)
data = data [(data['field']>(first_quartile - 3*iqr)) & (data['field']
<(third_quartile + 3*iqr))]
```

City Cheyenne has 4 outliers in families, population and sale amounts, but I choose to keep it considering Cheyenne is center city with lots of people, high sales makes sense.



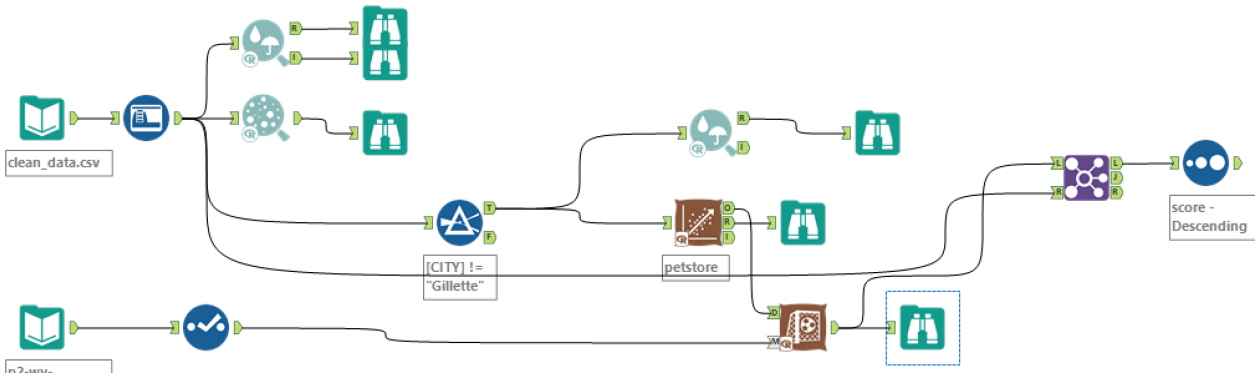
The above three figures show that the points of city Cheyenne are in line with the overall linear relationship.



Similarly, land area of Rock Springs also makes sense in *land_area-sales* relationship(shown in above figure).

Finally, Gillete's data is deleted considering its other fields are at their average while sales is abnormally high.

How and why did you select the predictor variables in your model?



Correlation is used to see if there is any possibility ofmulticollinearity in dataset.

Full Correlation Matrix

	Total.Pawdacity.Sales	Land.Area	Households.with.Under.18	Population.Density	Total.Families	X2010.Census
Total.Pawdacity.Sales	1.000000	-0.288898	0.676012	0.862894	0.864660	0.898099
Land.Area	-0.288898	1.000000	0.180704	-0.317244	0.099389	-0.061587
Households.with.Under.18	0.676012	0.180704	1.000000	0.815756	0.907242	0.911883
Population.Density	0.862894	-0.317244	0.815756	1.000000	0.884792	0.927702
Total.Families	0.864660	0.099389	0.907242	0.884792	1.000000	0.968005
X2010.Census	0.898099	-0.061587	0.911883	0.927702	0.968005	1.000000

So *Census*, *Families*,*Households with under 18* and *Population density* have strongcorrelations which each other, while *Land area* is not.

Land area and *Total families* are selected to bulid the final model.

```
Sales=197330.41-48.42*land.area+49.14*total.families
```

Explain why you believe your linear model is a good model.

Basic Summary

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

Residuals:

Min	1Q	Median	3Q	Max
-121300	-4453	8418	40490	75200

Coefficients:

	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	***

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

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 DF, p-value: 0.0002035

The p-values for *land area* and *total families* are both below 0.05 and the R-squared value is close to 1.

Conclusion

What is your recommendation?

When it comes to choose new city to open a store.

1. The new store should be located in a new city. That means there should be no existing stores in the new city.
2. The total sales for the entire competition in the new city should be less than $500,000
3. 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).
4. The predicted yearly sales must be over $200,000.
5. The city chosen has the highest predicted sales from the predicted set.

So four cities is selected:

City	2014 Estimate	Population VOLUME SALES	Land Area
Jackson	10,449	110000	1757
Lander	7,642	108197	3346
Laramie	32,081	76000	2513
Worland	5,366	100000	1294

Then linear model is used to to predict total sales.

City	2014 Estimate	Population VOLUME SALES	Land Area	Total Family	Predict Sales
Jackson	10,449	110000	1757	2313	225917.29
Lander	7,642	108197	3346	3876	225783.73
Laramie	32,081	76000	2513	4668	301036.47
Worland	5,366	100000	1294	1364	197701.89

So I would recommend the city of Laramie

Appendix

Difference between QUARTILE.EXC and QUARTILE.INC in Excel.

https://zhuanlan.zhihu.com/p/79461597

By 知乎专栏：小数据