The first and foremost step to approach any Data Science problem is to understand the business requirement and data. Let's understand the business requirement before moving further.

Business and Data Understanding

1. **What is the business requirement?**
   Pawdacity is planning to open its 14th store. Considering yearly sales as the selection category, a new city should be identified for opening a new store.

2. **What data is needed to make a prediction?**
   We need to come up with a model, which can predict yearly sales for new cities. So, we need our store's annual sales data and data that influence the annual sales of the store in that city. Demographic information like Land area, Population density, Total Families, Households under 18, Average income, the population of individual ethnic groups will be helpful. Besides, factors driving sales like the marketing budget of the store, the number of competitor stores will be beneficial in building a good model.
   To predict the yearly sales for new cities, we need a similar set of demographic data, planned marketing budget for the new cities. But unfortunately, we don't have few data like average income and marketing budget. Including them in future will help us to build a better model.

Before further analysis, we must clean the data and remove outliers. In the previous project, we didn't clean nor check for outliers because it was a cleaned dataset.

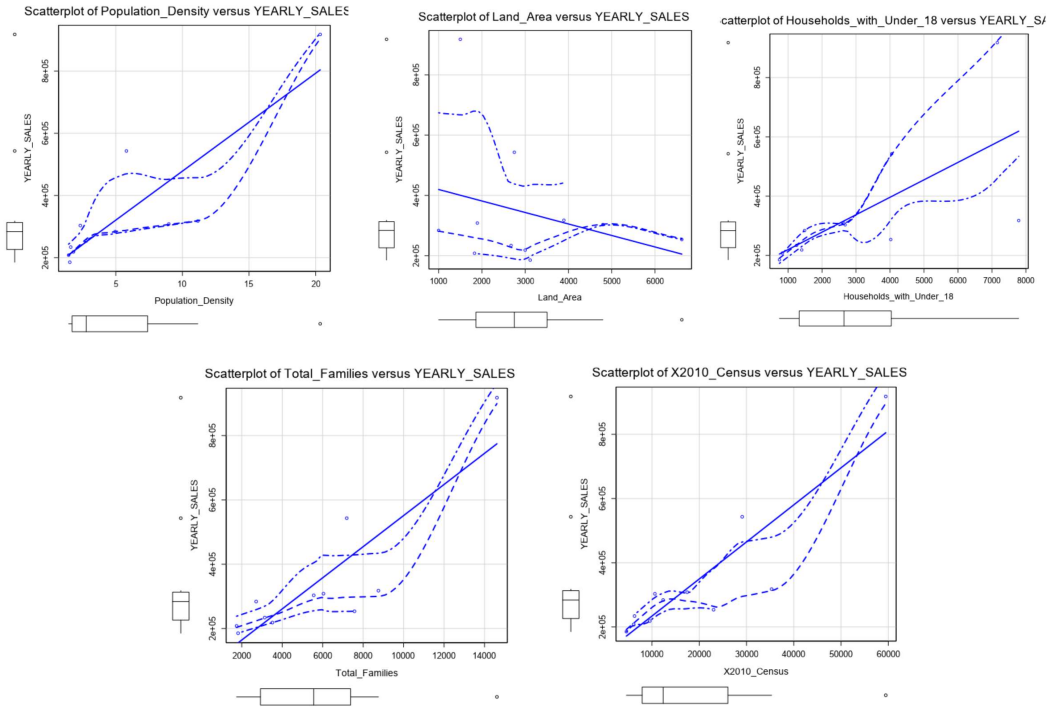
Data Wrangling

Data wrangling is the process of cleaning, structuring and enriching raw data into the desired format for better decision making in less time.

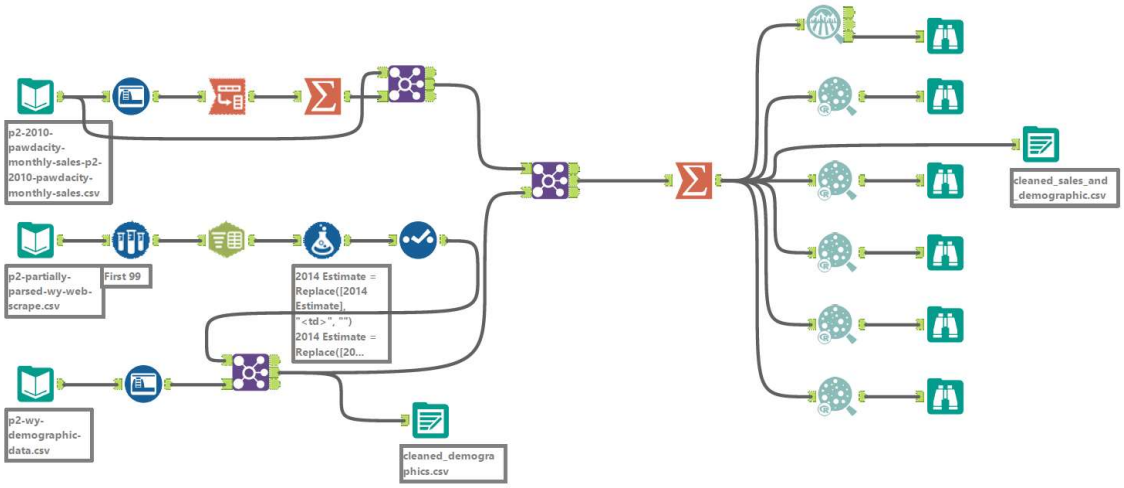
In our project, I have checked and removed the null records from all data sources. Most of our scrapped data should be cleaned before using it for further analysis. Hence, I have parsed and cleaned the scrapped population data and changed the field name into the required formats. Finally, merged it with demographic data and created a new file named cleaned_demographics.csv.

Our sales data has monthly sales data, but we need yearly sales for our model. Hence, I have processed the data and made it as annual sales. Finally, I merged it with the newly created demographic data to form a combined sales and demographic data file named cleaned_sales_and_demographics.csv

Having outliers in the data will have a high negative impact on the model's performance, so it must be removed before modeling. Considering yearly sales, both"Cheyenne" and "Gillette" are outliers to our dataset. But while looking closely at the dataset, we can see that all variables are skewed for "Cheyenne," which shows that it is a large city; hence it will be useful in designing a model, which may be used to predict sales for large cities. But in the case of "Gillette," we can see that other variables are within the range (while looking for outliers using IQR), while only the sales variable is out of range. Hence, we can remove "Gillette" from the dataset.

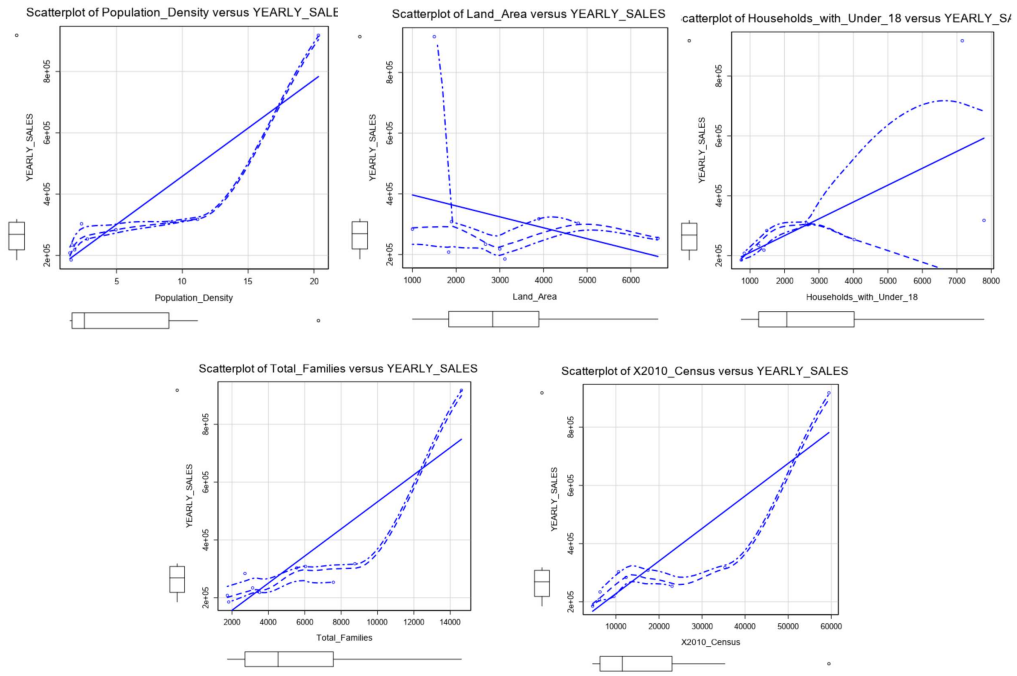


Data Wrangling – Alteryx Workflow



Analysis

Before modeling, we must analyze the dataset. Plotting each predictor variable with the target variable using scatterplot shows that all variables are correlated with the target variable.



I checked for correlations between my predictor variables to see any possibility of multicollinearity in my dataset. We can see that HHU18, Census, Families, and PDensity (Population Density) have strong correlations which each other. The land area, however, is not as highly correlated. So I started by using the 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.

Full Correlation Matrix

	YEARLY.SALES	Population.Density	Land.Area	Households.with.Under.18	Total.Families	X2010.Census
YEARLY.SALES	1.00000	0.90618	-0.28708	0.67465	0.87466	0.89875
Population.Density	0.90618	1.00000	-0.31742	0.82199	0.89168	0.94439
Land.Area	-0.28708	-0.31742	1.00000	0.18938	0.10730	-0.05247
Households.with.Under.18	0.67465	0.82199	0.18938	1.00000	0.90566	0.91156
Total.Families	0.87466	0.89168	0.10730	0.90566	1.00000	0.96919
X2010.Census	0.89875	0.94439	-0.05247	0.91156	0.96919	1.00000

Modeling and Validation

We have identified the optimal predictor variables, and the next step is to build the model. Usually, we will perform steps like scaling and one hot encoding before modeling, but Alteyrx takes care of that. So let us create a Linear regression model with these predictor variables.

After building a model, it is crucial to validate the model. Upon validating the model, we can see that the model has an adjusted R-squared value of 0.8866. It's important to note that the r-squared value represents the variation in the target variable that the model captures. Generally, we consider models with an adjusted r-squared >0.7 to be acceptable. In this case, about 89% of the variation is accounted for, leaving 11% to be explained by variables outside the model.

Multiple R-squared: 0.9118, Adjusted R-Squared: 0.8866

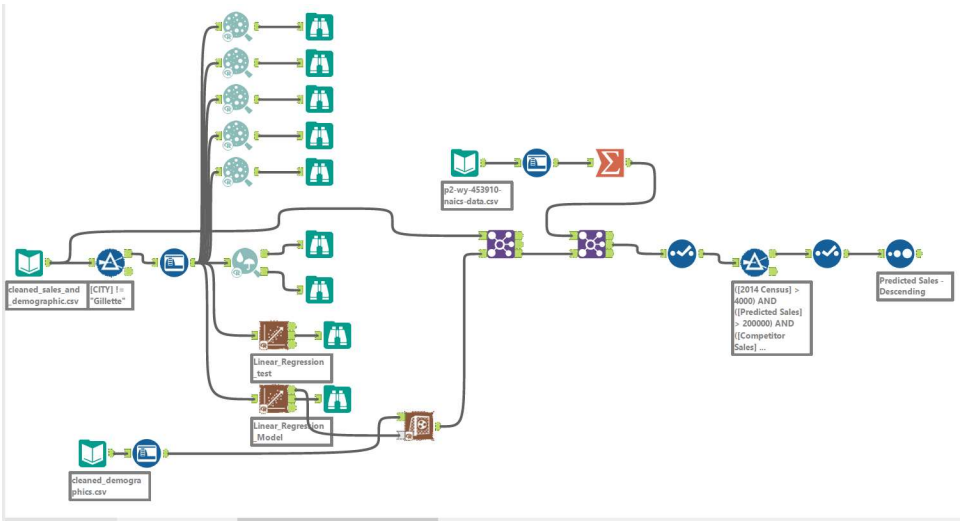
From the model, we can get the below regression equation:

Yearly_Sales = 197330.41 – 48.42 * Land_Area + 49.14 8 Total_Families

Results

We can use the cleaned city data on the model to predict yearly sales for every city. And then, we can use other factors to shortlist the best city. We have to ensure that the new city doesn't have any Pawdacity store, and the population must be greater than 4000 based on the 2014 census. Also, predicted yearly sales must be greater than $200,000, and the competitor's sales should not exceed $500,000. Based on the above filtering, we are getting four cities.

Alteryx Workflow



Recommendation

With detailed analysis, we build a model that can account for an 89% variance of the yearly sales from a city. Land Area and Total families are the only variables that have a good impact on sales, and we should work with our marketing team to include the marketing budget for existing stores and planned budget for the new store.

We have shortlisted four cities, and among that selected one city for the new store. **We predicted that Lamarine would be the best city for the new store, with the expected sales of $305,013.**

City	Predicted Sales
Laramie	305013.881671
Jackson	225870.8236
Lander	225751.400203
Worland	201700.325919