Luyan Deng
April 23rd 2018
Project 2-1/2-2

]

Creating an Analytical Dataset
Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (250 word limit)

Key Decisions:

Answer these questions

1. **What decisions needs to be made?**

   Pawdacity, a pet store chain in Wyoming with 13 stores, is planning to open the 14th store. This report will analyze the relation between historical sales data and the city that a certain store located in, and thus recommend the city for Pawdacity's new store, which is most likely to generate best yearly sales in the future.

2. **What data is needed to inform those decisions?**

   The dataset requires some data including:
   1) All of the Pawdacity stores' sales in 2010
   2) NAICS data on the most current sales of all competitors in 12 months (each store's sales in cities)
   3) 2010's Census
   4) Demographic data in each city in Wyoming(including Households with Under 18, Land Area, population Density and Total families).

Step 2: Building the Training Set

Build your training set given the data provided to you. Your column sums of your dataset should match the sums in the table below.

The dataset for a new store is built from a set of data from the following sources:

1) 2010 Pawdacity monthly sales
2) US Census Bureau's population data
3) Demographic data for each city in Wyoming

The data is cleaned and aggregated by city. Three tables were joined by city and the result has 11 rows, representing the information of 11 cities that Pawdacity's stores currently operate. The dataset will be used to predict the sales of Pawdacity's new store in different cities. The following is a summary of the dataset.

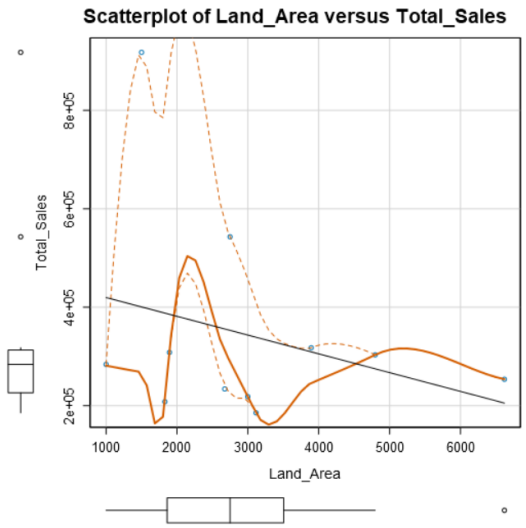
Luyan Deng
April 23rd 2018
Project 2-1/2-2

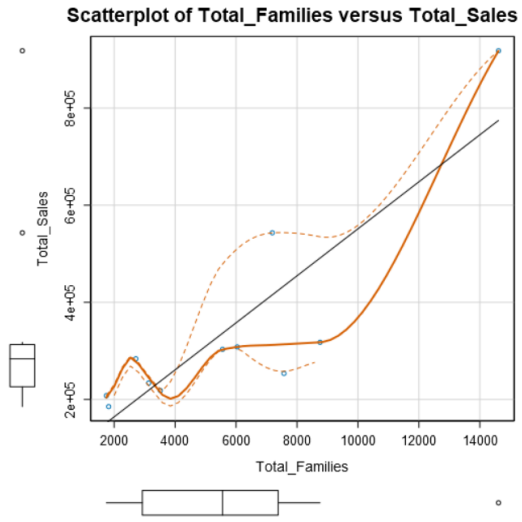
Column	Sum	Average
2010 Census Population	*213,862*	*19442*
Total Pawdacity Sales	*3,773,304*	*343027.64*
Households with Under 18	*34,064*	*3096.73*
Land Area	*33,071*	*3006.49*
Population Density	*63*	*5.71*
Total Families	*62,653*	*5695.71*

Step 3: Dealing with Outliers

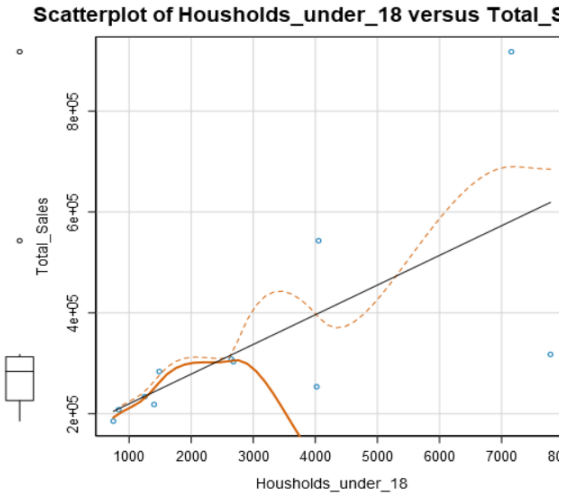
Answer these questions

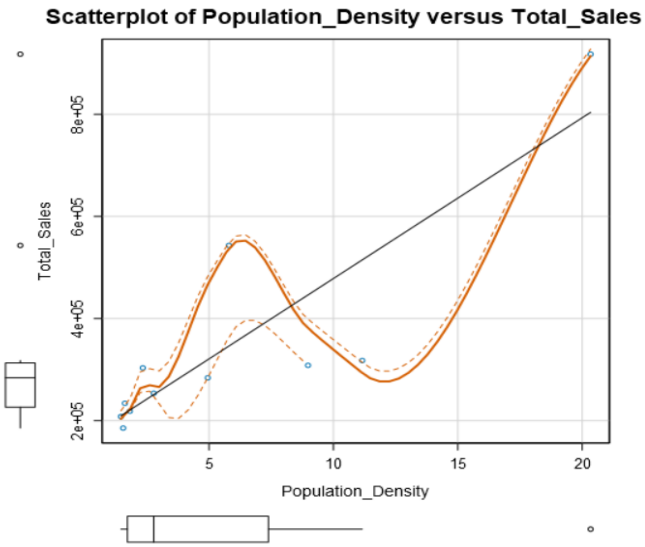
Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.

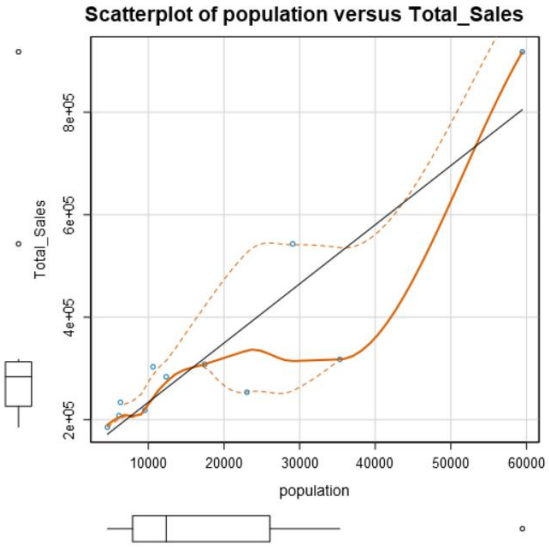




Luyan Deng
April 23rd 2018
Project 2-1/2-2







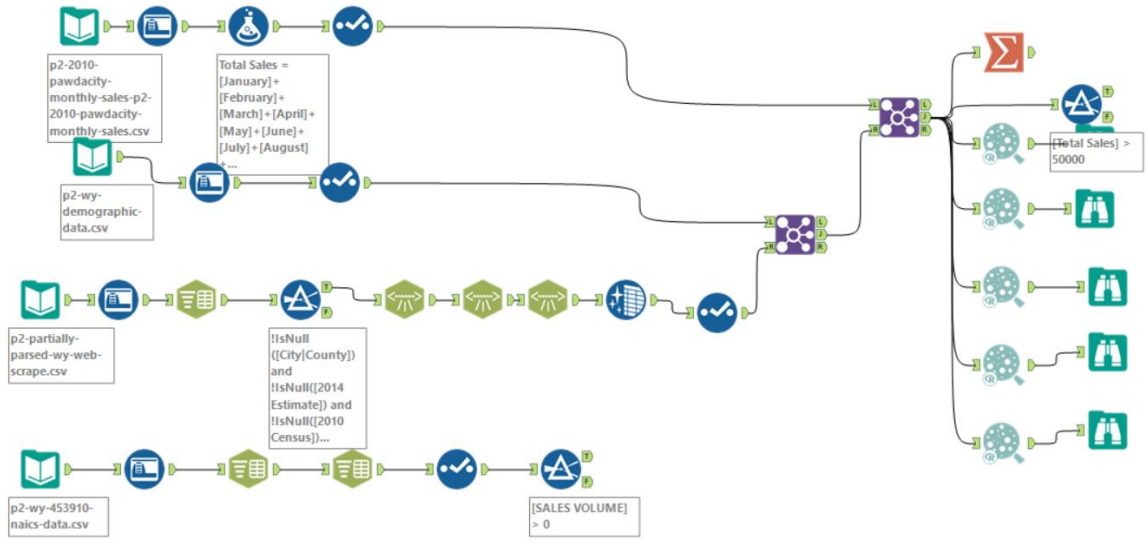
According to the five diagrams above, the first possible outlier for total sales is city of **Cheyenne,** sales data are higher than expected in Land Area-Total Sale, Total Families, population density. Because the area, population and population density are highly related. And the total sale and these variables do show a significant linear relationship with each of these variables. Therefore, the Cheyenne's record is kept in the dataset so as to predict the result of popular cities.

Luyan Deng
April 23rd 2018
Project 2-1/2-2

The second possible outlier is **Gillette,** for it has a great distance from the trend line in diagram land area, total families and population density for total sales. We also find that Gillette's sales cannot be explained by those variables and will skew the model that train the data. Therefore, Gillette is an outlier and should be removed,

Third possible outlier is **Rock Spring**. In land area versus Total sales, Rocks spring has much larger size of land. However, it still fits the linear model with total sales. Because it doesn't skew the model, it will be kept for prediction.

The following is the workflow to clean up data and observe data features.

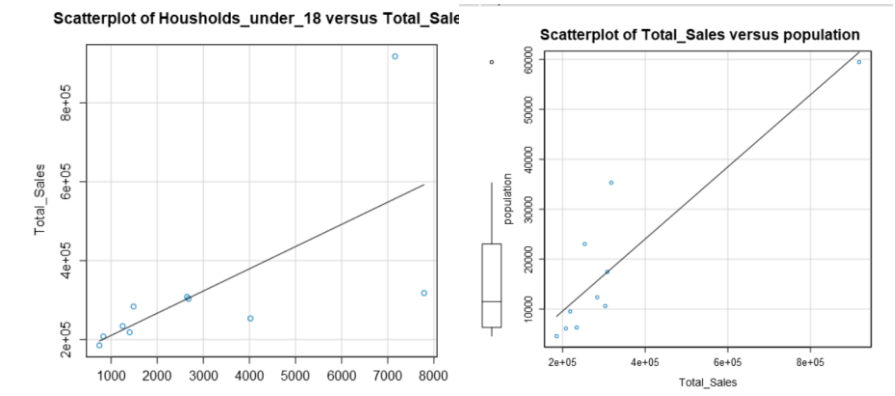


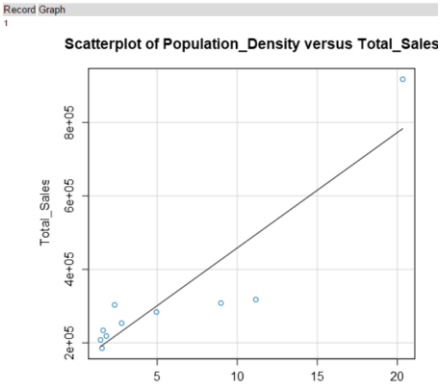
Luyan Deng
April 23rd 2018
Project 2-1/2-2

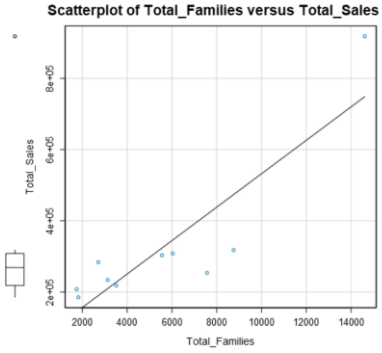
Project 2-2 Regression and Recommendation

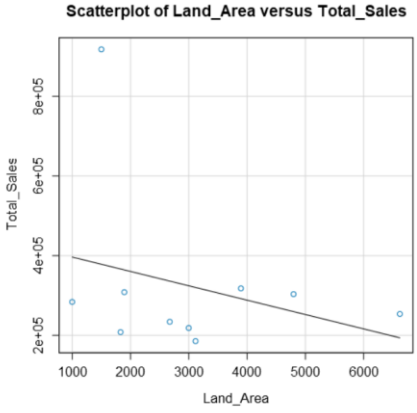
Step 4. Build a Linear regression model

After removing the city of Gillette from the dataset, there are 10 rows left. I built up the regression model between each variable and the total sales. Apart from Land Area, the other four variables show significant positive correlation with the total sales.









Luyan Deng
April 23rd 2018
Project 2-1/2-2

We then take a look at the Pearson correlation analysis between variables to see if they relate to each other. Apart from Land Area, variables (population density, population, total families, household under 18) seem to highly related to each other. We can get a clearer view from the correlation matrix, the red parts showing the significant correlations between two variables while the lighter boxes show weaker correlations.

Pearson Correlation Analysis

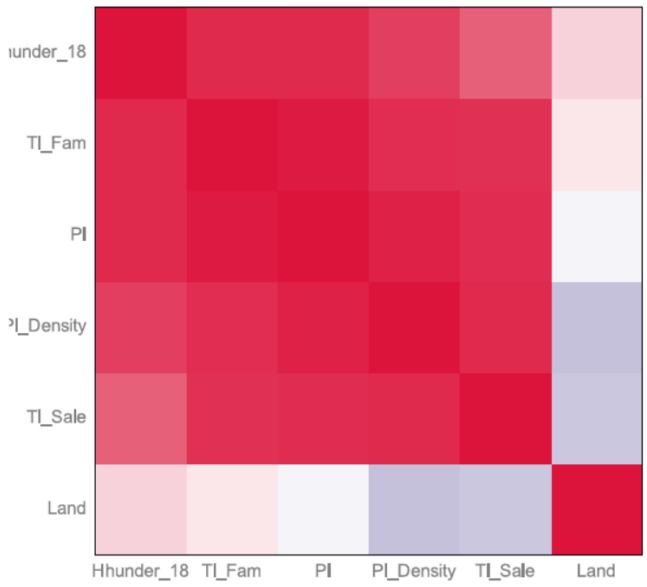
Focused Analysis on Field Tl_Sale

	Association Measure	p-value
Pl_Density	0.90618	0.00030227 ***
Pl	0.89875	0.00040617 ***
Tl_Fam	0.87466	0.00092561 ***
Hhunder_18	0.67465	0.03235537 *
Land	-0.28708	0.42126310

Full Correlation Matrix

	Tl_Sale	Land	Hhunder_18	Pl_Density	Tl_Fam	Pl
Tl_Sale	1.00000	-0.28708	0.67465	0.90618	0.87466	0.89875
Land	-0.28708	1.00000	0.18938	-0.31742	0.10730	-0.05247
Hhunder_18	0.67465	0.18938	1.00000	0.82199	0.90566	0.91156
Pl_Density	0.90618	-0.31742	0.82199	1.00000	0.89168	0.94439
Tl_Fam	0.87466	0.10730	0.90566	0.89168	1.00000	0.96919
Pl	0.89875	-0.05247	0.91156	0.94439	0.96919	1.00000

Correlation Matrix with ScatterPlot

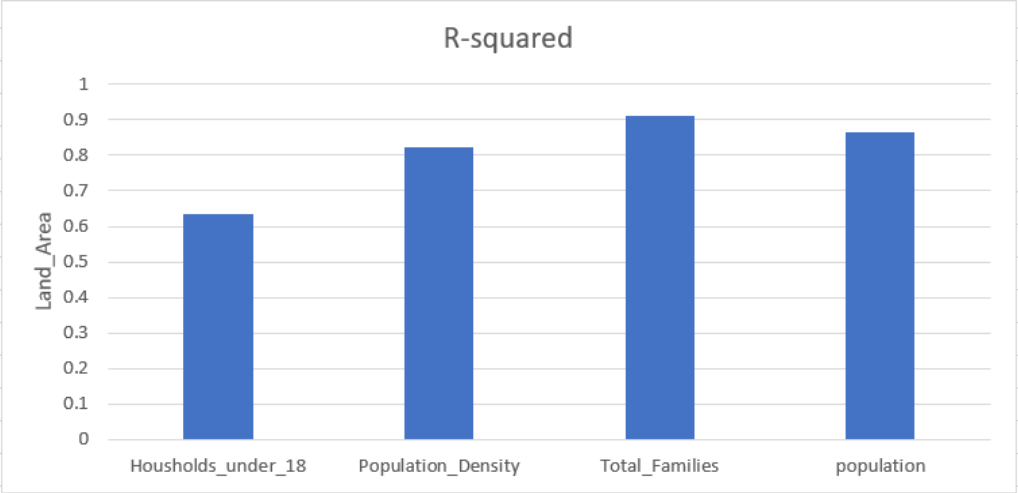


Luyan Deng
April 23rd 2018
Project 2-1/2-2

Therefore, we fit a regression model to the Land_Area and one of other variables, and compare those regression rereports. From Land_Area and Total_Families fit the regression best, with a R-squared 0.91. We select these two predictor variables for the linear regression

Based on the regression analysis, we can have the regression model:
Total_Sales =197330.41 -48.42*Land_Area + 49.14*Total_Families

As both p values are less than 0.05 and the R-squared is higher than 0.9, it is a valid model.



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

Type II ANOVA Analysis

Response: Total.Sales

	Sum Sq	DF	F value	Pr(>F)
Land_Area	60473052720.43	1	11.66	0.01123 *
Total_Families	341673845917.83	1	65.85	8e-05 ***
Residuals	36318449406.44	7		

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

Luyan Deng
April 23rd 2018
Project 2-1/2-2

Step 5: Make Prediction and Recommendation

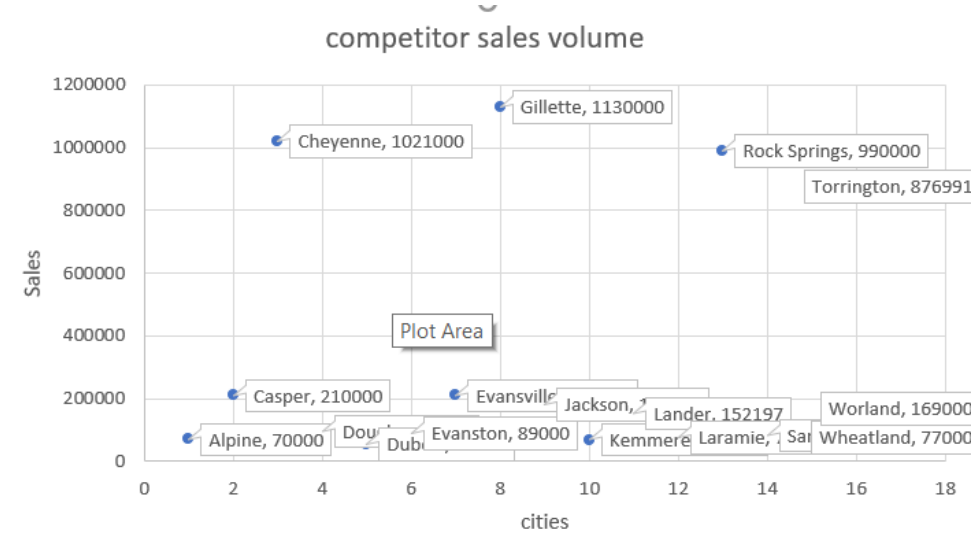
For choosing the optimal next Pawdacity store, there are two sets of data will be used:

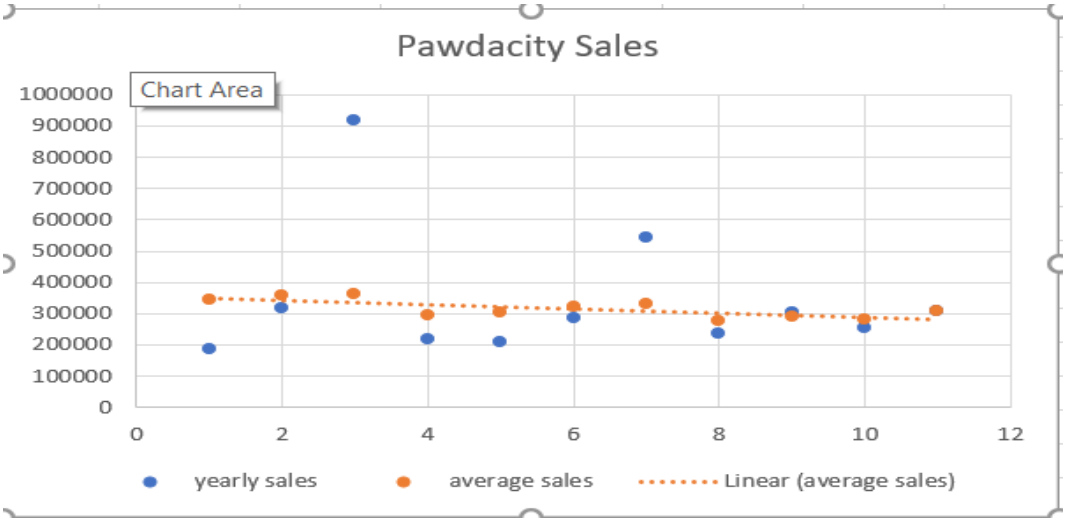
1) Demographic data for each city and county in the state of Wyoming population from US Census Bureau(2000,2010 and 2014 estimated)
2) NAICS Data for competitors' city, county and sales volume

We have a few criteria to pick the new city:

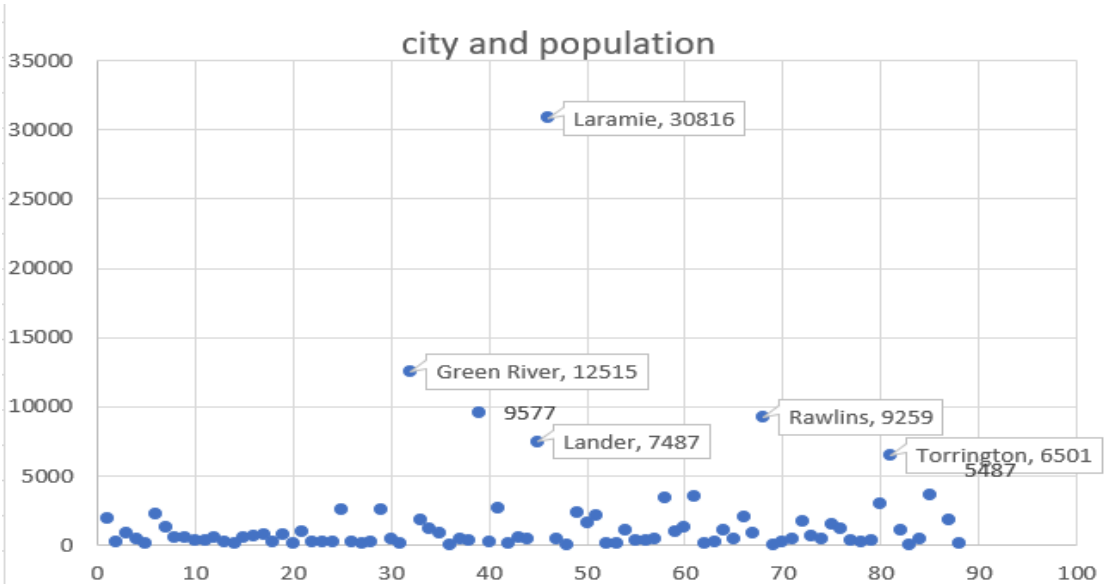
1) It must be a new city, that means the cities with Pawdacity stores will be excluded.
2) The population of the city (based on 2014 estimation) should be higher than 4500.
3) The competitor's sales volume should not higher than 500,000.
4) The predicted yearly sales must be above $200,000.
5) The city chosen has the highest predicted sales.

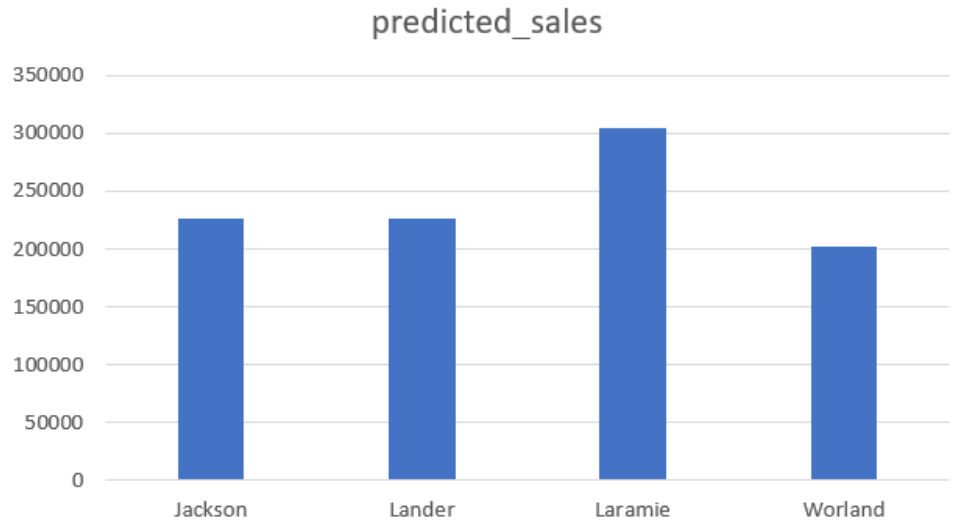
After cleaning the NAICS data, we join the demographic data and sorted out the cities that already have Pawdacity stores.





Luyan Deng
April 23rd 2018
Project 2-1/2-2





Based on the sorting criteria, we than have a predicted set of cities: Jackson, Lander, Larramie, Worland. Compared the four cities' predicted sales, Laramie has the highest predicted sales of $305081.8, and thus it is the best option for a new store.

Luyan Deng
April 23rd 2018
Project 2-1/2-2