ONYEKABA NZUBECHUKWU JUDE

Project 2.1: Data Cleanup

Make a copy of this document. Complete each section. When you are ready, save your file as a PDF document and submit it here:
https://classroom.udacity.com/nanodegrees/nd008/parts/8d60a887-d4c1-4b0e-8873-b2f36435eb39/project

Step 1: Business and Data Understanding

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

Key Decisions:

Answer these questions

1. What decisions needs to be made?

A leading pet store chain in Wyoming, Pawdacity needs recommendation on where to open its fourteen (14th) store.

2. What data is needed to inform those decisions?

Some of the data required in order to inform this decision are 2010 census population, city, competitor sales, Pawdacity sales in other stores, household with under 18, population density, land area, 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.

In addition, provide the averages on your data set here to help reviewers check your work. You should round up to two decimal places, ex: 1.24

Column	Sum	Average
Census Population	*213,862*	*19442.00*
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*

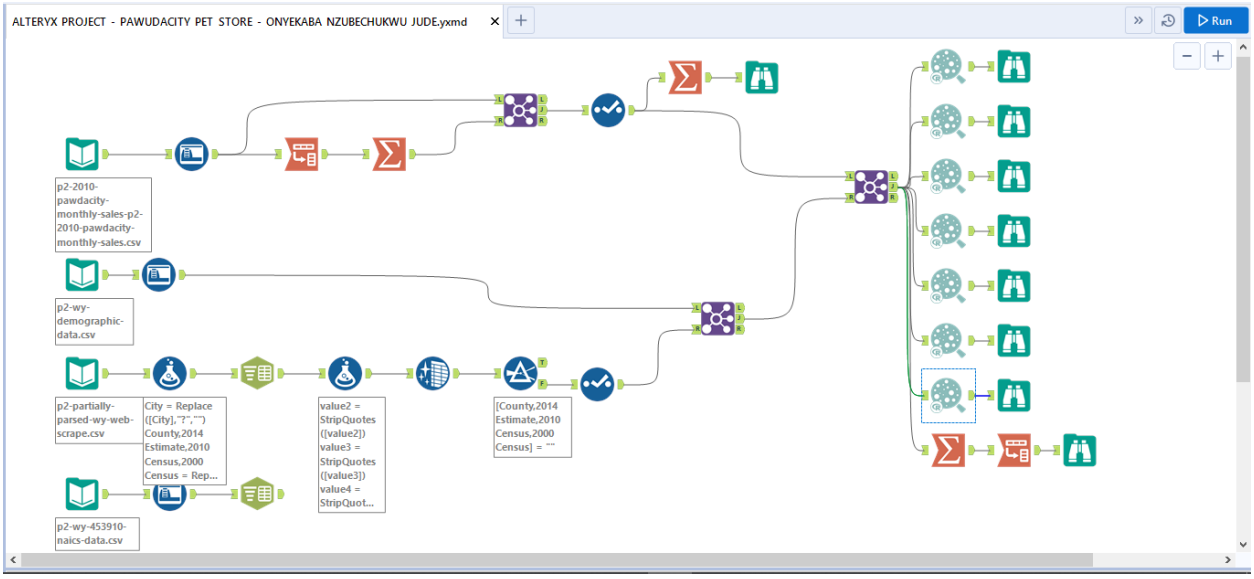
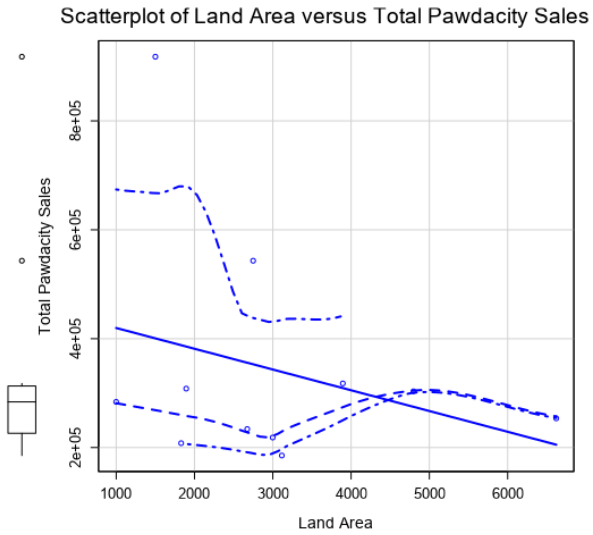


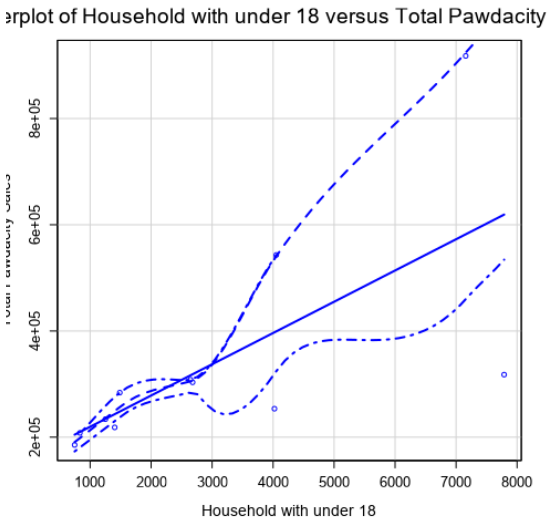
Figure 1: Workflow

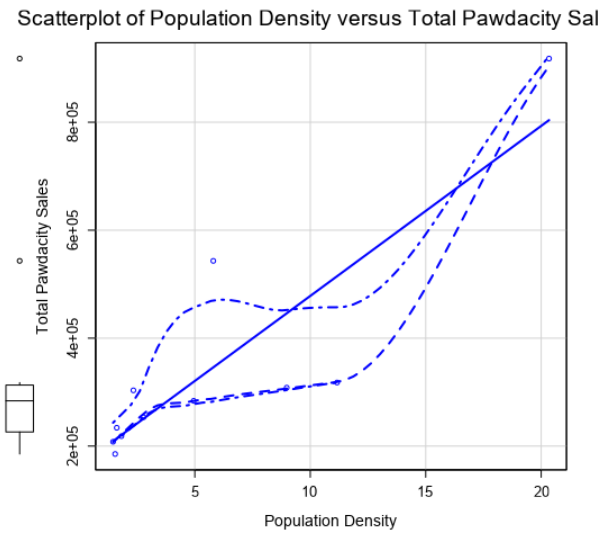
Step 3: Dealing with Outliers

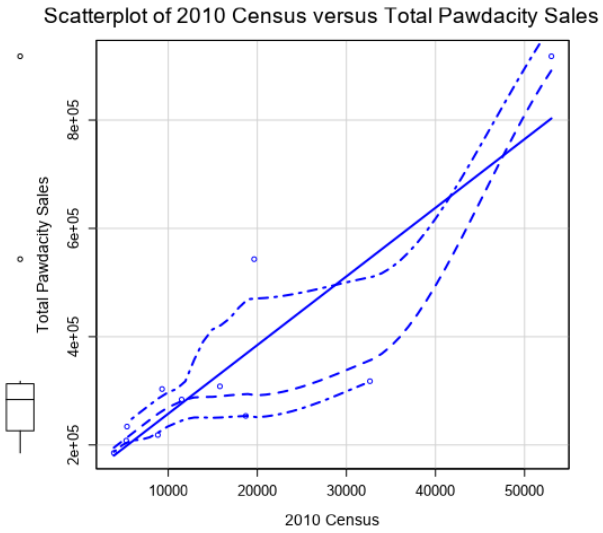
Answer these questions

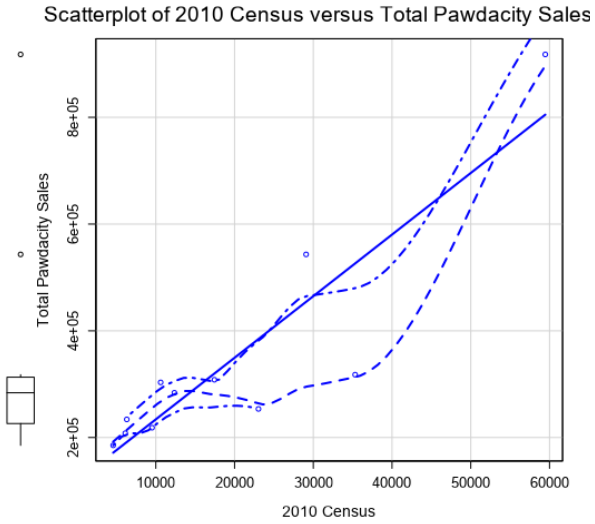
Are there any cities that are outliers in the training set?

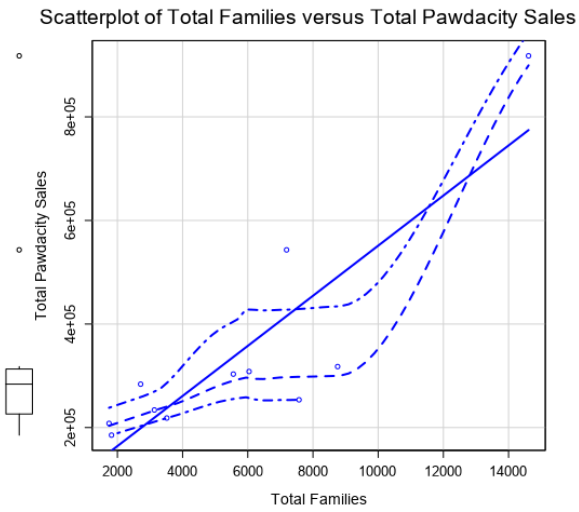












Based on the six (6) scatterplots above, there seem to be two outliers. The cities of Gillette and Cheyenne seem to be the outliers as their sales data are higher than expected.

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.

After making research, I was still not sure which is correct. I decided to Step-Wise Regression Tool in Alteryx to select the best predictor variables. I did Linear regression and step wise regression for the following cases:

- With both outliers
- Without only Cheyenne
- Without only Gillette

A. <u>With both outliers:</u>

This model was the worst of the three cases. The p values were above 0.05 showing that they were not statistically significant

Basic Summary

Call:

lm(formula = Total_Sales ~ Land.Area + Households.with.Under.18 + Population.Density + Total.Families + X2014.Estimate, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-48096	-30267	-5151	18014	71292

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	238561.544	64274.402	3.712	0.01383 *
Land.Area	-67.346	36.065	-1.867	0.12082
Households.with.Under.18	-45.595	31.706	-1.438	0.20993
Population.Density	-25887.442	17033.708	-1.520	0.18904
Total.Families	72.241	32.719	2.208	0.07831 .
X2014.Estimate	8.931	6.815	1.310	0.247

B. Without only Cheyenne:

This model was the best. All the p values are less than 0.05, and the star ratings were high for each variable.

Basic Summary

Call:

lm(formula = Total_Sales ~ Land.Area + Population.Density + Total.Families, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-53464	-29088	-13915	26096	69487

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	292680.20	47765.60	6.127	0.00086 ***
Land.Area	-91.83	23.21	-3.957	0.00748 **
Population.Density	-36987.13	12397.14	-2.984	0.02452 *
Total.Families	91.55	20.19	4.533	0.00396 **

C. Without only Gillette:

This model wasn't as good as the model without Cheyenne. Some p values were above 0.05 after using step regression.

Basic Summary

Call:

lm(formula = Total_Sales ~ Land.Area + Households.with.Under.18 + Total.Families + X2014.Estimate, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-79333	-22215	6996	21394	75337

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	167245.396	46261.739	3.615	0.0153 *
Land.Area	-27.234	18.254	-1.492	0.19592
Households.with.Under.18	-62.161	26.998	-2.302	0.06956 .
Total.Families	48.886	24.021	2.035	0.09747 .
X2014.Estimate	7.877	7.168	1.099	0.32189

Therefore, I have decided to drop the outlier - Cheyenne row as this gave me my best model.

Model:
Total Sales = 292680 - 91.83*Land Area – 36987.13*Population Density + 91.55*Family Size

Final Workflow: