Create an Analytical Dataset

ansjayan@msn.com

## Step 1: Business and Data Understanding

# Pawdacity, the leading pet store chain in Wyoming is planning to open her 14th store in the state. Pawdacity wants to know, which city is best for the new store, which can be found by predicting yearly sales.

### Key Decisions:

*Answer these questions*

1. What decisions needs to be made?

Find the best city based on predicted yearly sales.

1. What data is needed to inform those decisions?

Data required: - Details of each city are required, which includes population, population density, land area, total families, homes with under 18 members, and pawdacity sales in each city.

## 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* |
| *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

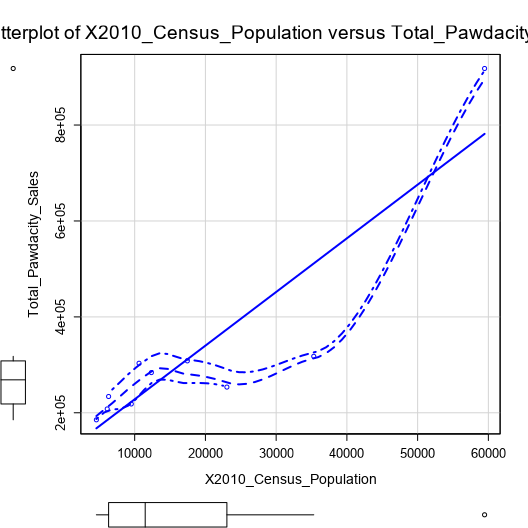
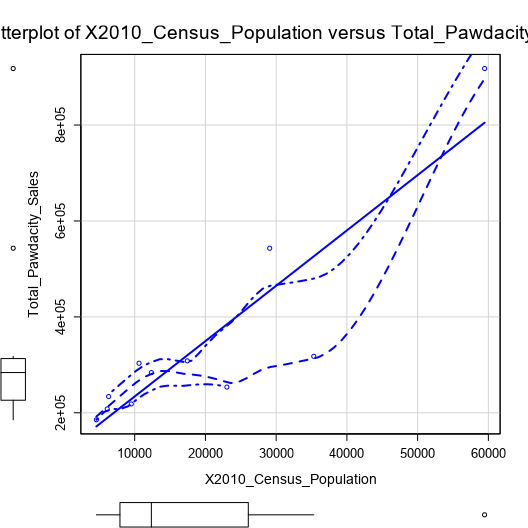
*Answer these questions*

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.



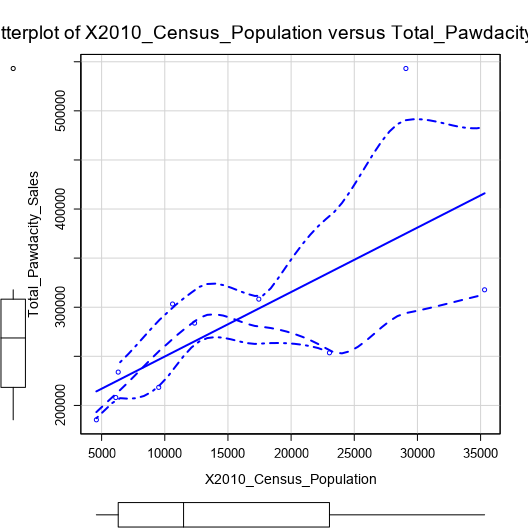
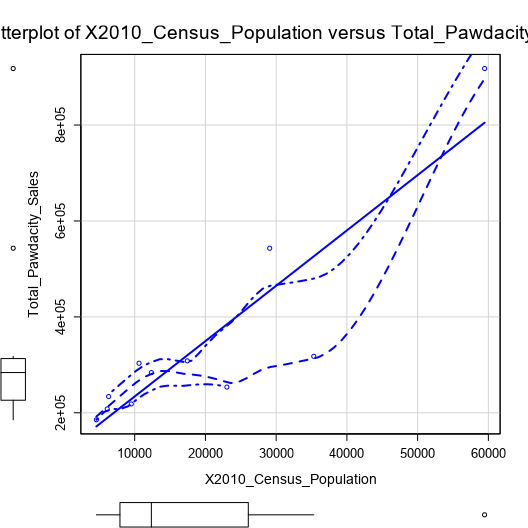
There are three outliers: - Gillette, Cheyenne and Rock Springs.

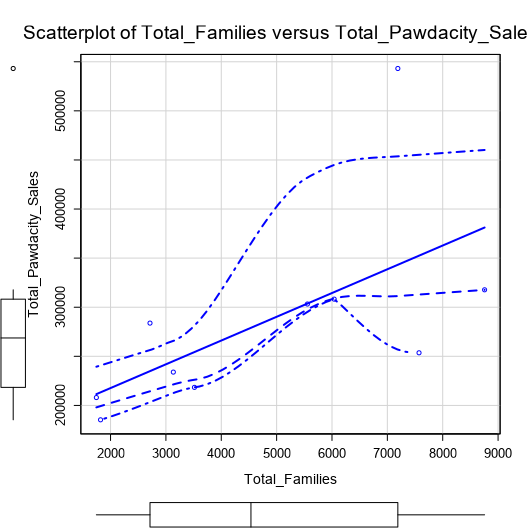
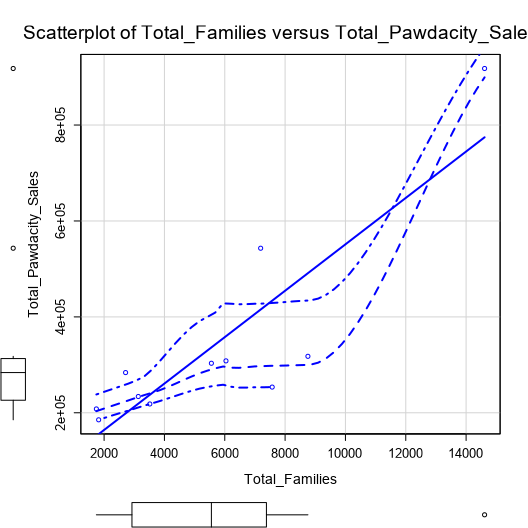
Outlier 1: - **Gillette**

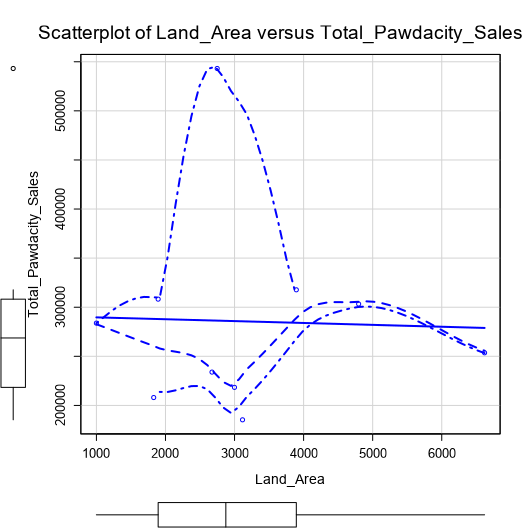
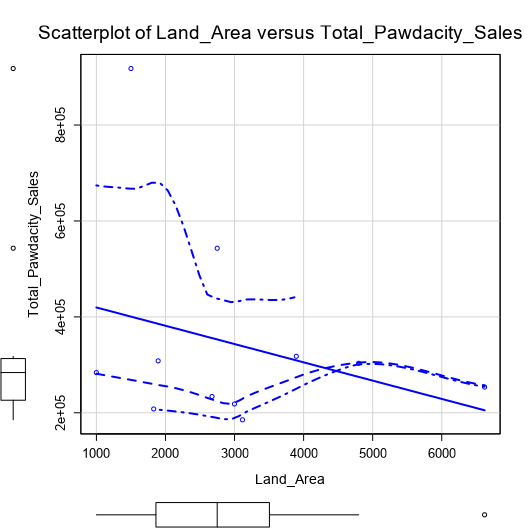


The left side graph is with Gillette and right side is without. Even though the slope remains same the plot shows that Gillette is above the regression line. So, the outlier Gillette can be removed.

Outlier 2: - **Cheyenne**

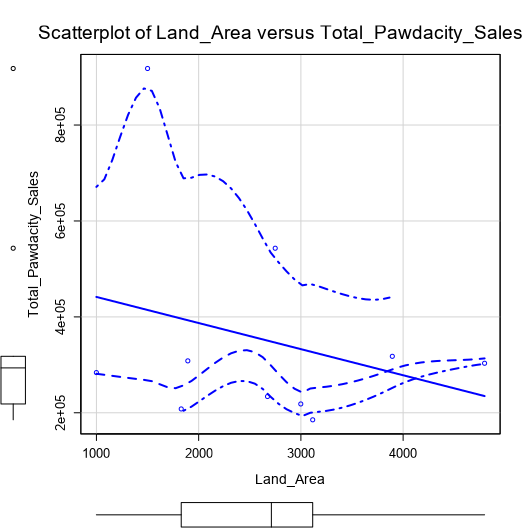
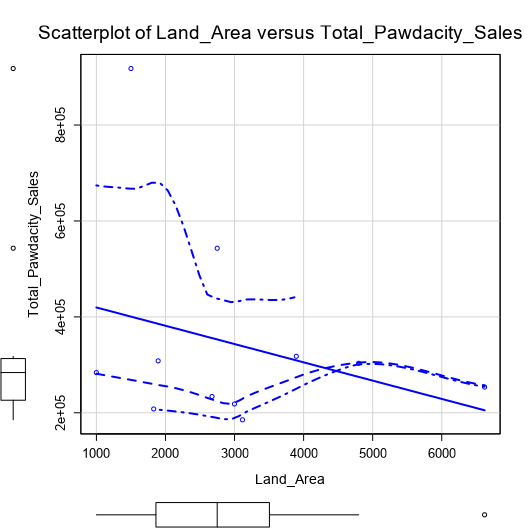




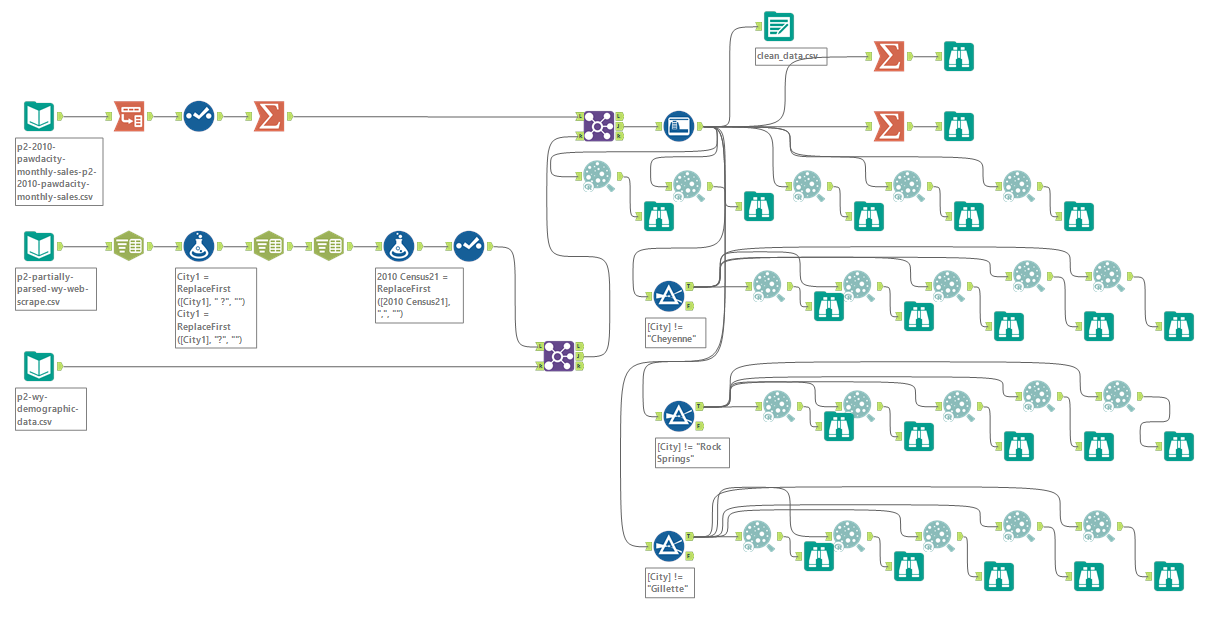


The left side graph is with Cheyenne and right side is without. Cheyenne, a big city, is an outlier to Sales, Population, Families, and density. The slope is changed when Cheyenne is removed. If we model any future big cities, keeping Cheyenne will be good. As we got only 11 rows in dataset, keep Cheyenne.

Outlier 3: - **Rock Springs**



The left side graph is with Rock Springs and right side is without. Rock Springs skew high in land area, but not with other variables. Rock Springs is an outlier but the slope stayed consistent with the plot without this outlier. So, models build with this outlier will be consistent like models without it and we have only 11 rows we can keep it.



Before you Submit

Please check your answers against the requirements of the project dictated by the [rubric](https://review.udacity.com/#!/rubrics/382/view) here. Reviewers will use this rubric to grade your project.