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 word limit)*

### Key Decisions:

*Answer these questions*

1. What decisions needs to be made?

* To perform an analysis to recommend the city for Pawdacity’s newest store, based on predicted yearly sales.
* Decide what fields to keep and whether they are statistically significant to the target variable.
* Identify outliers and decide if we need to delete, impute, or leave them unchanged.

1. What data is needed to inform those decisions?

* First, we need to have the target variable i.e, annual sales per city for training of the dataset. We are given monthly data for each city. We must find the aggregate value for the whole year.
* We need to figure out the predictor variables. I am choosing the predictor variables as Census population for 2010, Households with under 18, Land area, Population density, Total Families. They have been grouped by city.
* Since we are analyzing data from the year 2010, we need population for each city in that year. However, there can be different cases like first where population is distributed over a large area (less dense) or second the same population confined to a smaller area (highly dense). In second case one can expect more number of customers, hence preferable. Similarly, the Land Area can help reach a decision. If the city area is large, then for the same number of living in different cities, city with large area will be less dense and hence the customers may be lower reducing the sales.
* Generally, kids under 18years encourage parents to get pets. Hence, larger this population more likely larger will be the number of customers, thereby increasing the sales. Similarly, higher the number of families higher will be the sales.
* We can visualize these relationships by generating scatter plots between predictor variable and the target variable.
* We need to clean the data to extract the meaningful and usable form of data.

## 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* | *343027.63* |
| *Households with Under 18* | *34,064* | *3096.72* |
| *Land Area* | *33,071* | *3006.48* |
| *Population Density* | *63* | *5.7* |
| *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.

Answer:

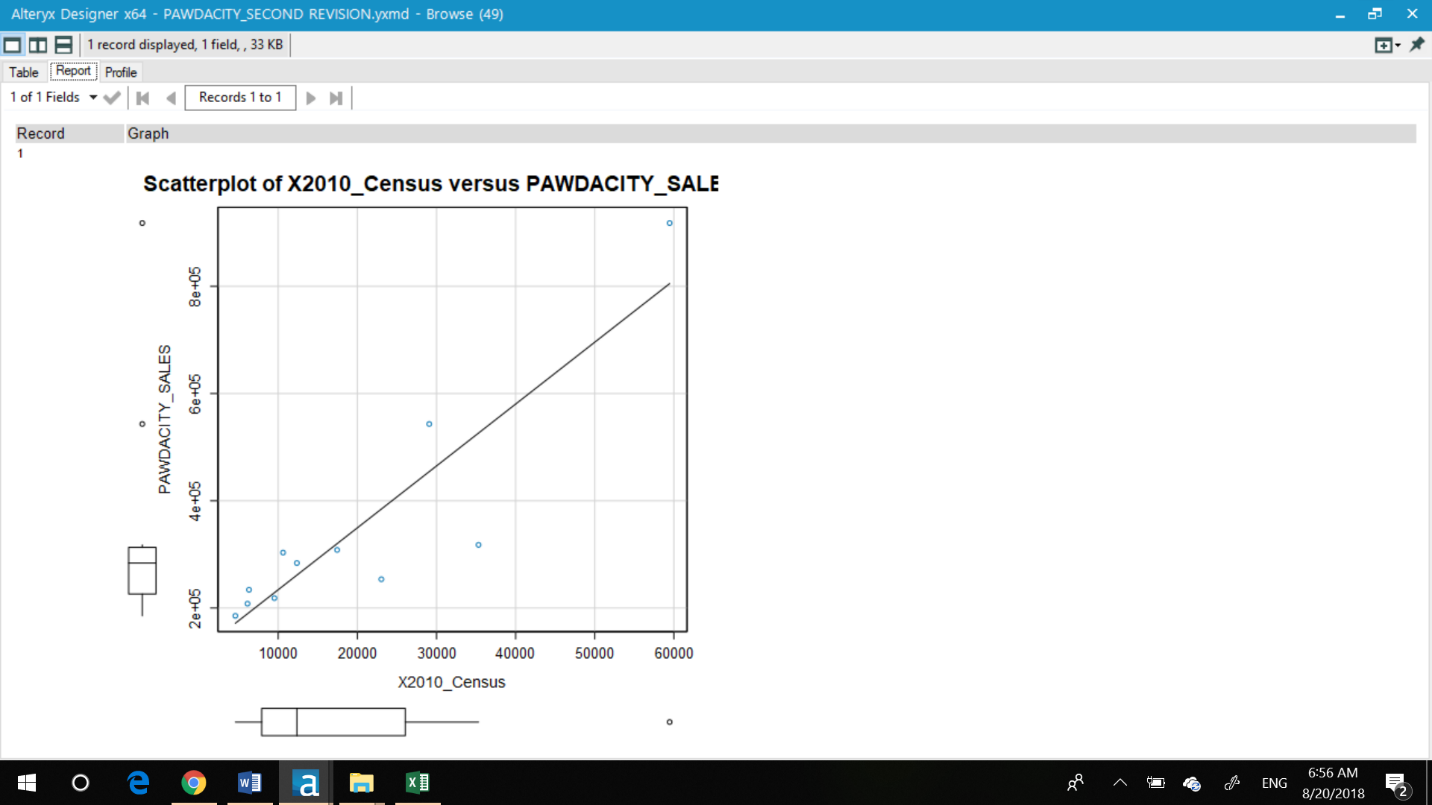
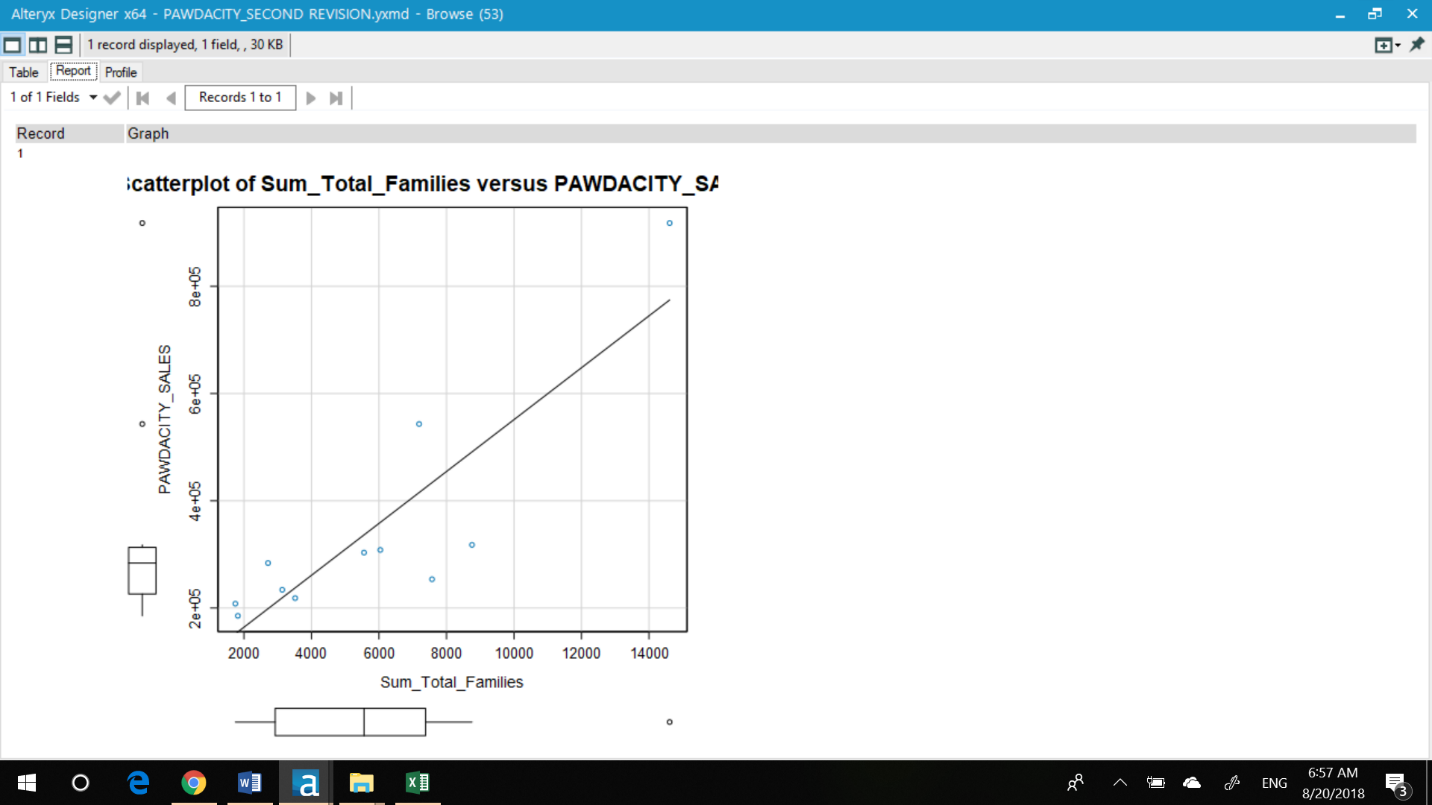
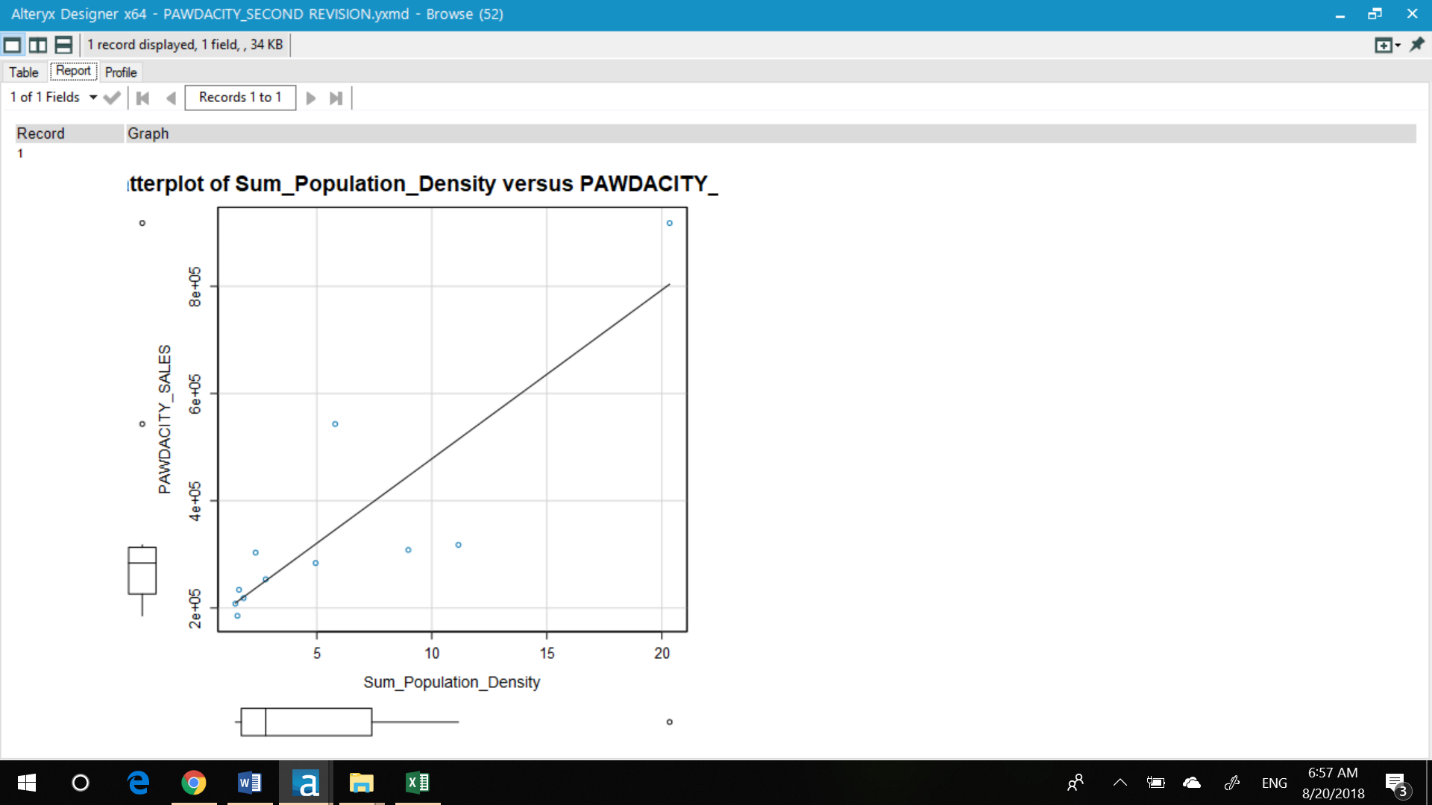
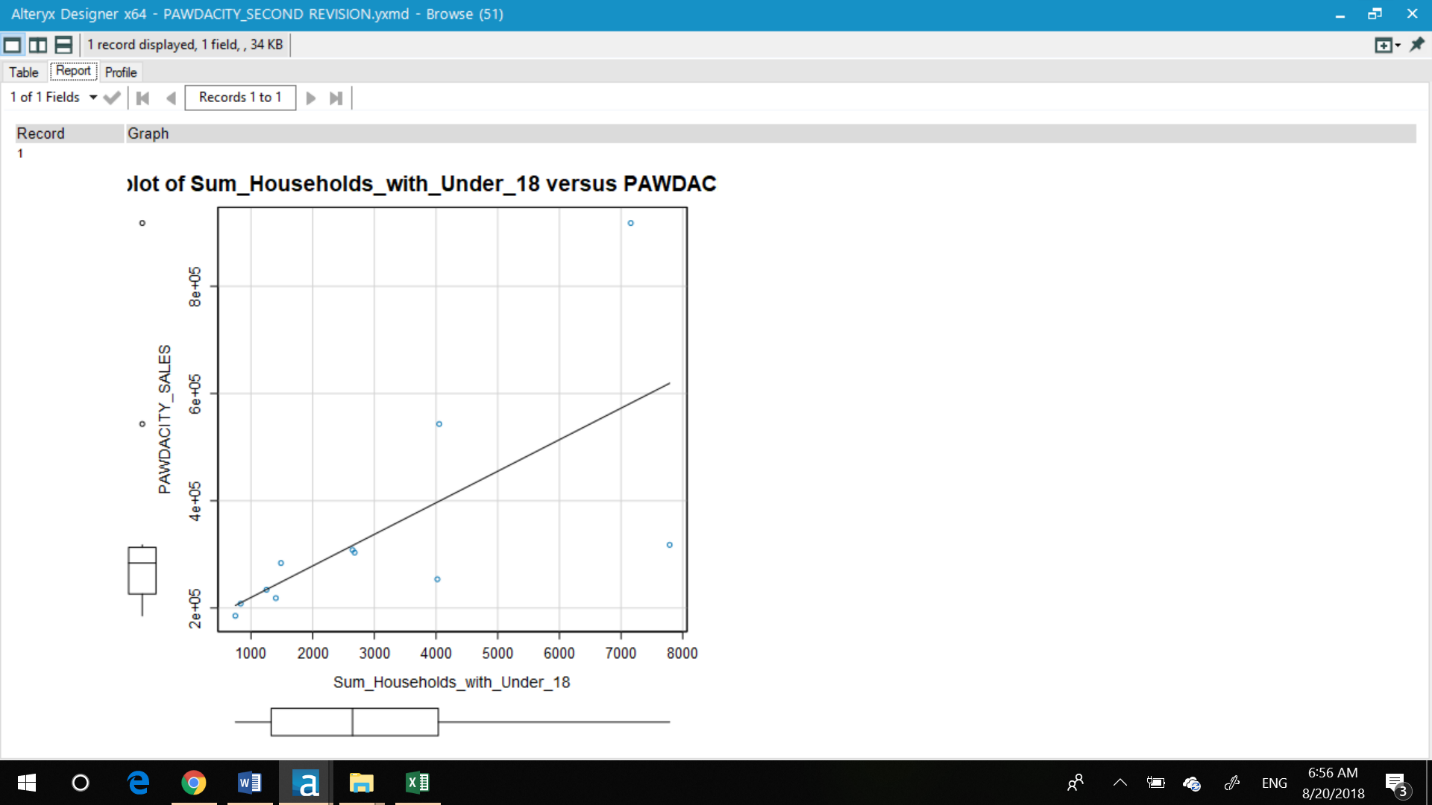
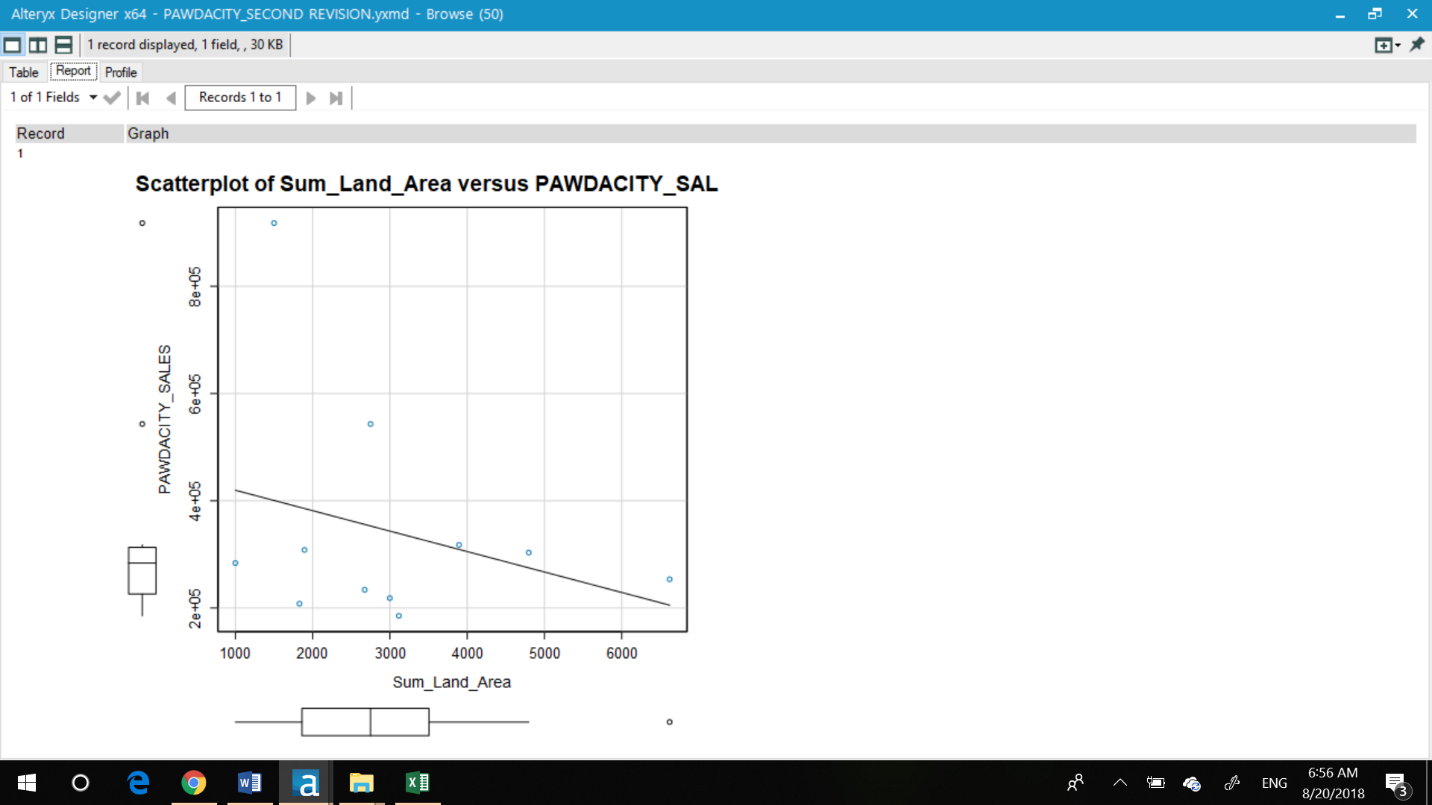
The outliers are highlighted in red (please refer the table below). Also, please refer the scatter plots below for seeing the correlation between predictor and target variables. As seen from the scatter plots only land area vs target variable has negative slope. All others have positive slope.

Cheyenne has outlier in all the fields except for Land Area and House hold under 18. The Land area value (1500) for Cheyenne is very near to the minimum value for that column (999). As we mentioned the smaller the land area the larger will be the population density. Here the total population is an outlier too and that explains why the population density and sum\_total families are outliers too. Hence, Cheyenne’s data doesn’t distort the outcome of prediction. I will keep Cheyenne’s data as it is for this reason.

If we leave Gillette’s sales as it is then it will not be the right indication of target variable as none of the predictor variables are outliers. Higher sales woud be explained better by higher population density etc. Hence, we need to impute the sales value of Gillette.

Rock Springs has outlier in land area. But if we look at the scatter plot of sum land area and total sales, then it looks that removing this outlier will have no significant effect on the sales. We can leave this unchanged.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Record# | Business Name | City | Pawdacity Sales | 2010 Census | Sum\_Land\_Area | Household under 18 | Sum Population density | Sum Total Families |
| 1 | Pawdacity | Buffalo | 185328 | 4585 | 3116 | 746 | 1.55 | 1820 |
| 2 | Pawdacity | Casper | 317736 | 35316 | 3894 | 7788 | 11.16 | 8756 |
| 3 | Pawdacity | Cheyenne | 917892 | 59466 | 1500 | 7158 | 20.34 | 14613 |
| 4 | Pawdacity | Cody | 218376 | 9520 | 2999 | 1403 | 1.82 | 3516 |
| 5 | Pawdacity | Douglas | 208008 | 6120 | 1829 | 832 | 1.46 | 1744 |
| 6 | Pawdacity | Evanston | 283824 | 12359 | 999 | 1486 | 4.95 | 2713 |
| 7 | Pawdacity | Gillette | 543132 | 29087 | 2749 | 4052 | 5.8 | 7189 |
| 8 | Pawdacity | Powell | 233928 | 6314 | 2674 | 1251 | 1.62 | 3134 |
| 9 | Pawdacity | Riverton | 303264 | 10615 | 4797 | 2680 | 2.34 | 5556 |
| 10 | Pawdacity | Rock Springs | 253584 | 23036 | 6620 | 4022 | 2.78 | 7572 |
| 11 | Pawdacity | Sheridan | 308232 | 17444 | 1894 | 2646 | 8.98 | 6040 |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| minimum value |  |  | 185328 | 4585 | 999 | 746 | 1.46 | 1744 |
| Q1 |  |  | 226152 | 7917 | 1862 | 1327 | 1.72 | 2923 |
| Q2 (Median) |  |  | 283824 | 12359 | 2749 | 2646 | 2.78 | 5556 |
| Q3 |  |  | 312984 | 26061.5 | 3505 | 4037 | 7.39 | 7381 |
| Maximum value |  |  | 917892 | 59466 | 6620 | 7788 | 20.34 | 14613 |
| Q3-Q1 |  |  | 86832 | 18144.5 | 1643 | 2710 | 5.67 | 4457 |
| Q1-1.5(Q3-Q1) |  |  | 95904 | -19300 | -602.5 | -2738 | -6.79 | -3763 |
| Q3+1.5(Q3-Q1) |  |  | 443232 | 53278.3 | 5970 | 8102 | 15.9 | 14067 |
|  |  |  |  |  |  |  |  |  |



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.