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)

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

Answer:

Pawdacity is planning to open the 14th store for the pet store chain. The decision needs to be made is the location for the new store. The decision is based on the predicted yearly sales for each city. The city chosen should have the highest predicted sales.

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

Answer:

We need to calculate the predicted yearly sales of all new cities. To do that, we need to study past sales data to identify which variables or factors specific to a city that contributed to past sales, such as **Census population**, **the number of households with under 18 people**, **land area**, **population density**, and **the total number of families**. This will be the inputs for our training linear regression model. After that, we will need a new data set of new cities that the above-mentioned model can be applied to in order to calculate the predicted sales.
Both the training data set and the new data set need to be clean without any duplicates and have missing values and outliers properly treated.

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.

Answer:

I built the following workflow to merge 3 data sets and clean them.

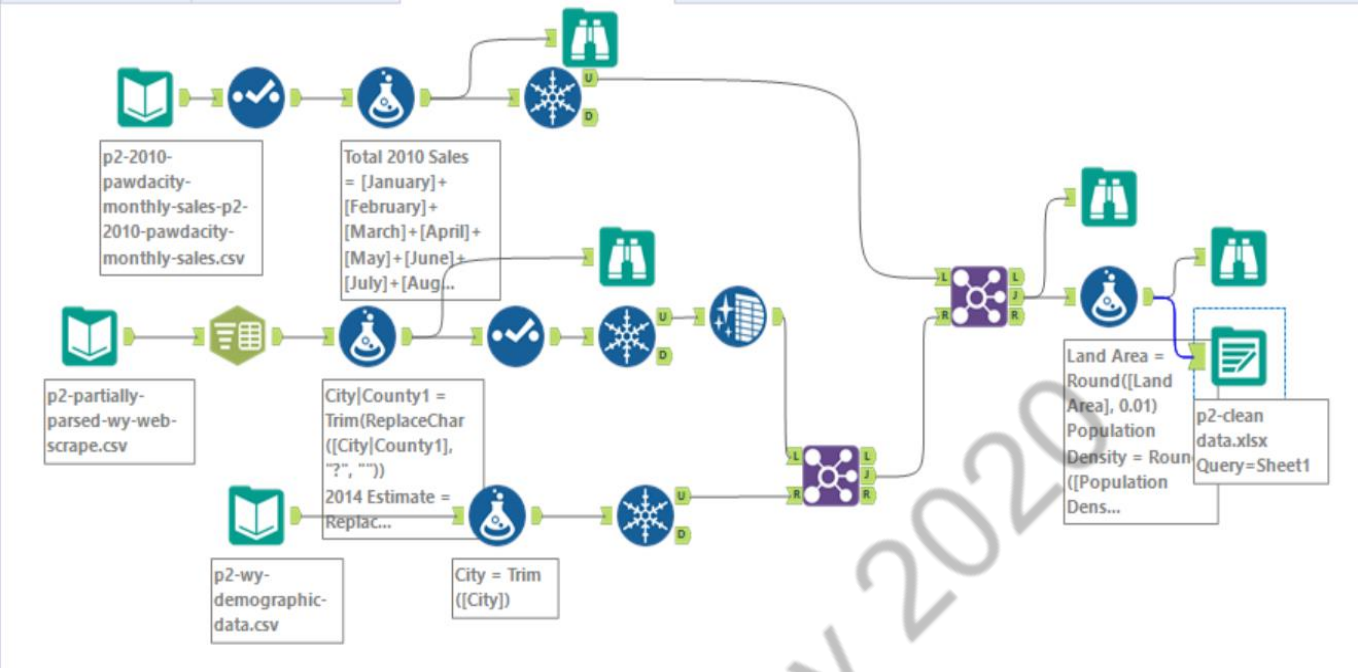


Figure 1 Data Preparing Workflow

Process explanation:

Pawdacity Monthly Sales data set

- Use the **Input Data** tool to upload the data set.
- Use the **Select** tool to change data type from **String** to **Float** for month sales fields because those are numerical variables.
- Use the **Formula** tool to calculate the **total 2010 Sales** from summing up 12 monthly sales fields. This will result in a new column called "**Total 2010 Sales**". In addition, to make sure that there are no trailing white spaces in the string values contained in the **city** column.
- Use the **Unique** tool to filter out all duplicate values.

Population Data Set

- Use the **Input Data** tool to upload the data set.
- Use the **Text to Column** tool to parse the column **City|County** into two separate columns: **City|County1** and **City|County2.**
- Use the Formula tool to clean up unwanted characters and replace them with an empty character, denoted by **""**. Formulas used are:

Trim(ReplaceChar([City|County1], "?", ""))

ReplaceChar([2014 Estimate], "<td>,</td>", "")

ReplaceChar([2010 Census], "<td>,</td>", "")

ReplaceChar([2000 Census], "<td>,</td>", "")

- Use the **Select** tool to change data type from V_String to Float for numerical fields such as **2000 Census**, **2010 Census**, **2014 Estimate**. In addition, rename columns: **City|County1** to **City** and **City|County2** to **County**.
- Use the **Unique** tool to filter out all duplicate values
- Use the **Data Cleansing** tool to remove all **Null rows**.

Demographic Data Set

- Use the **Input Data** tool to upload the data set
- In the **Formula** tool's configuration pane, use the formula Trim([city]) to remove all the empty spaces in the string values.
- Use the **Unique** tool to filter out all duplicate values

Merge Data Set

- Use the **Join** tool to join the population data set with the demographic data set using the same field, i.e. **city.**
- Afterwards**,** connect the **J node** in the output of the joined data set with an input of another **Join** tool to join it with the **Pawdacity monthly sales data set**. This join is also based on the shared field: **city**.

Formatted the merged data set

- In the **Formula** tool's configuration pane, use the formula Round([variable name]), 0.01) to round the values to the nearest two decimal points.

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

Answer:

Column	Sum	Average
Census Population	213,862	19,442
Total Pawdacity Sales	3,773,304	343,027.64
Households with Under 18	34,064	3,096.73
Land Area	33,071	3,006.49
Population Density	63	5.71
Total Families	62,653	5,695.71

Step 3: Dealing with Outliers

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:

To check for outliers, I used the Inter-Quartile Range (IQR) method. Afterwards, I use scatterplots to study the impact of keeping and removing the outlier(s). As the result, I determined that Cheyenne is the outlier that needs to be treated.

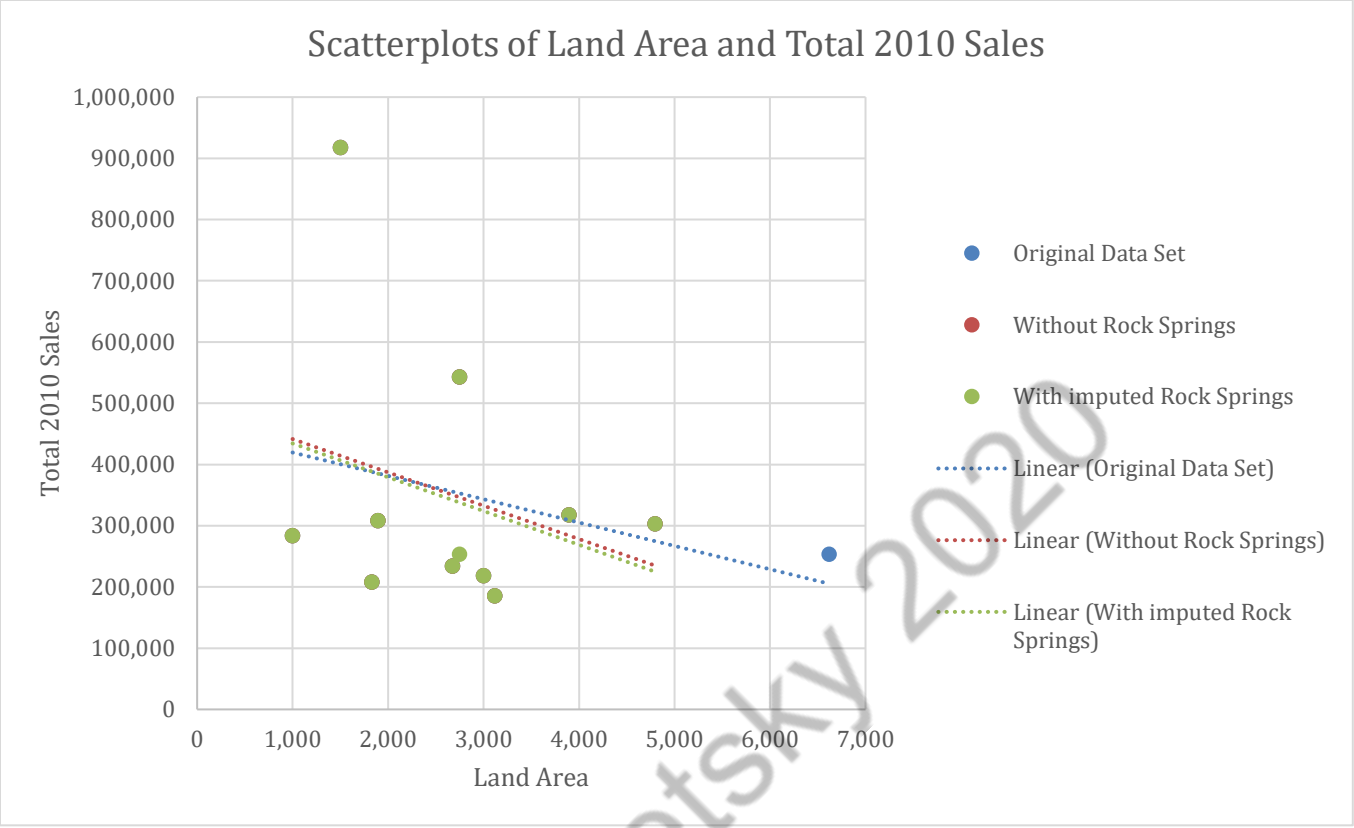
The output of IQR calculation is illustrated in the following table, with outlier values being highlighted in red:

CITY	Total 2010 Sales	2010 Census	Land Area	Households with Under 18	Population Density	Total Families
Buffalo	185328	4585	3115.51	746	1.55	1819.50
Casper	317736	35316	3894.31	7788	11.16	8756.32
Cheyenne	917892	59466	1500.18	7158	20.34	14612.64
Cody	218376	9520	2998.96	1403	1.82	3515.62
Douglas	208008	6120	1829.47	832	1.46	1744.08
Evanston	283824	12359	999.50	1486	4.95	2712.64
Gillette	543132	29087	2748.85	4052	5.80	7189.43
Powell	233928	6314	2673.57	1251	1.62	3134.18
Riverton	303264	10615	4796.86	2680	2.34	5556.49
Rock Springs	253584	23036	6620.20	4022	2.78	7572.18
Sheridan	308232	17444	1893.98	2646	8.98	6039.71
Lower Fence	95904	-19299.75	-603.05	-2738	-6.78	-3762.68
1st Quartile	226152	7917	1861.72	1327	1.72	2923.41
Median	283824	12359	2748.85	2646	2.78	5556.49
3rd Quartile	312984	26061.5	3504.91	4037	7.39	7380.81
Upper Fence	443232	53278.25	5969.69	8102	15.89	14066.90
interquartile	86832	18144.5	1643.19	2710.00	5.67	4457.40

As can be seen from the above table, there are three cities that contain outliers: Gillette, Rock Springs, and Cheyenne.

Rock Springs:

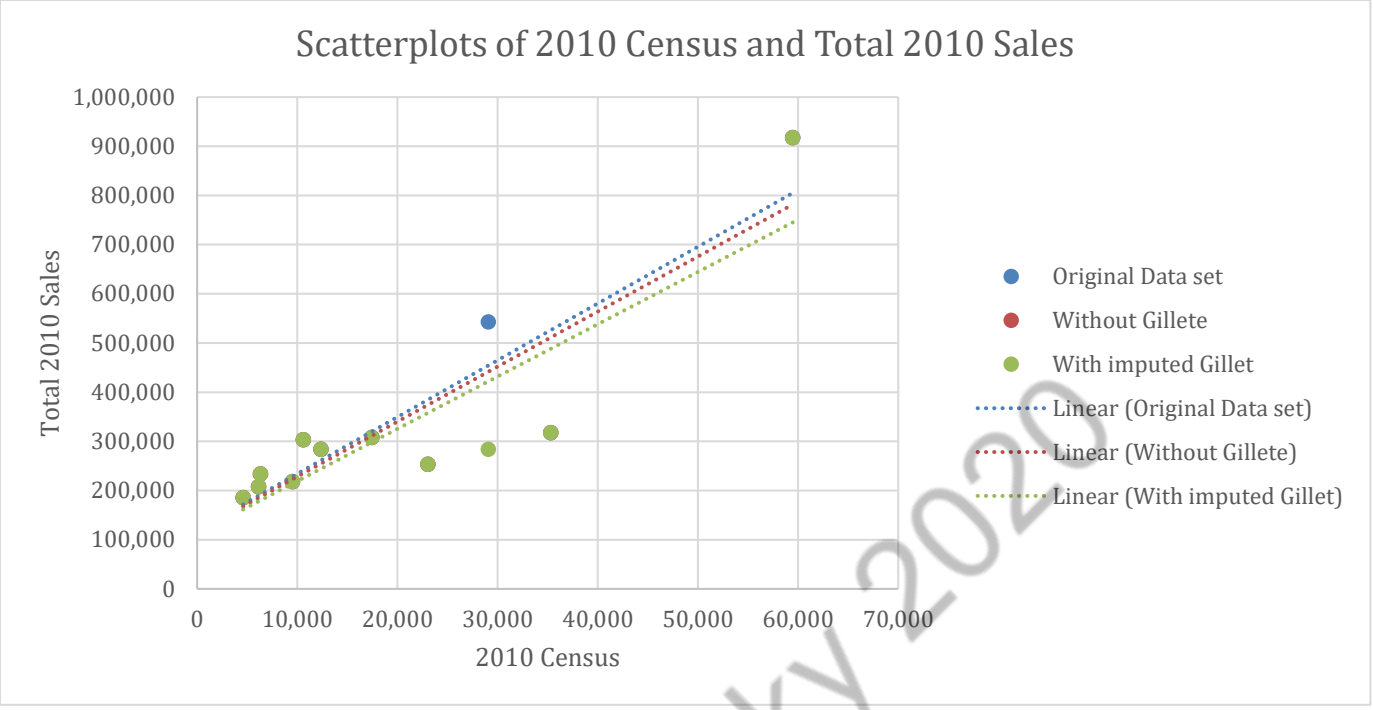
Rock Springs has only the **Land Area** value that is an outlier. However, the value is not too far away from the rest of the values and this is a small data set of only 11 records. Therefore, I would not worry about this record impacting my model and I will keep Rock Springs as is. The following graph confirms that:



I plotted 3 scatterplots of Land Area and Total Sales for three scenarios: 1) using the original data set, 2) removing the land area value of Rock Springs, and 3) imputing the value with the median of the field. It is clear that the outlier does not impact the model much since all three linear lines have roughly the same slope.

Gillet:

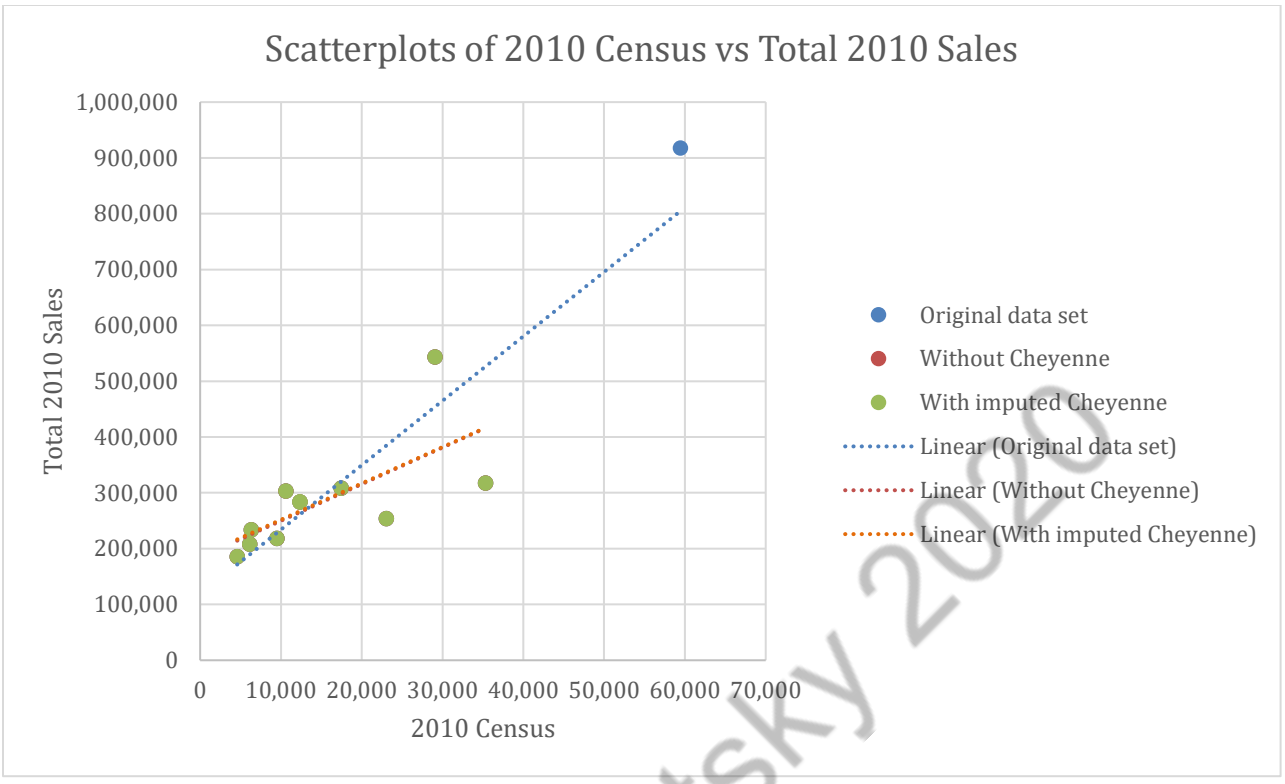
Gillet has only the **Total 2010 Sales** value that is an outlier. However, the value is not too far away from the rest of the sales values and this is a small data set of only 11 records. Therefore, I will keep Gillet as is. The following graph confirms that:

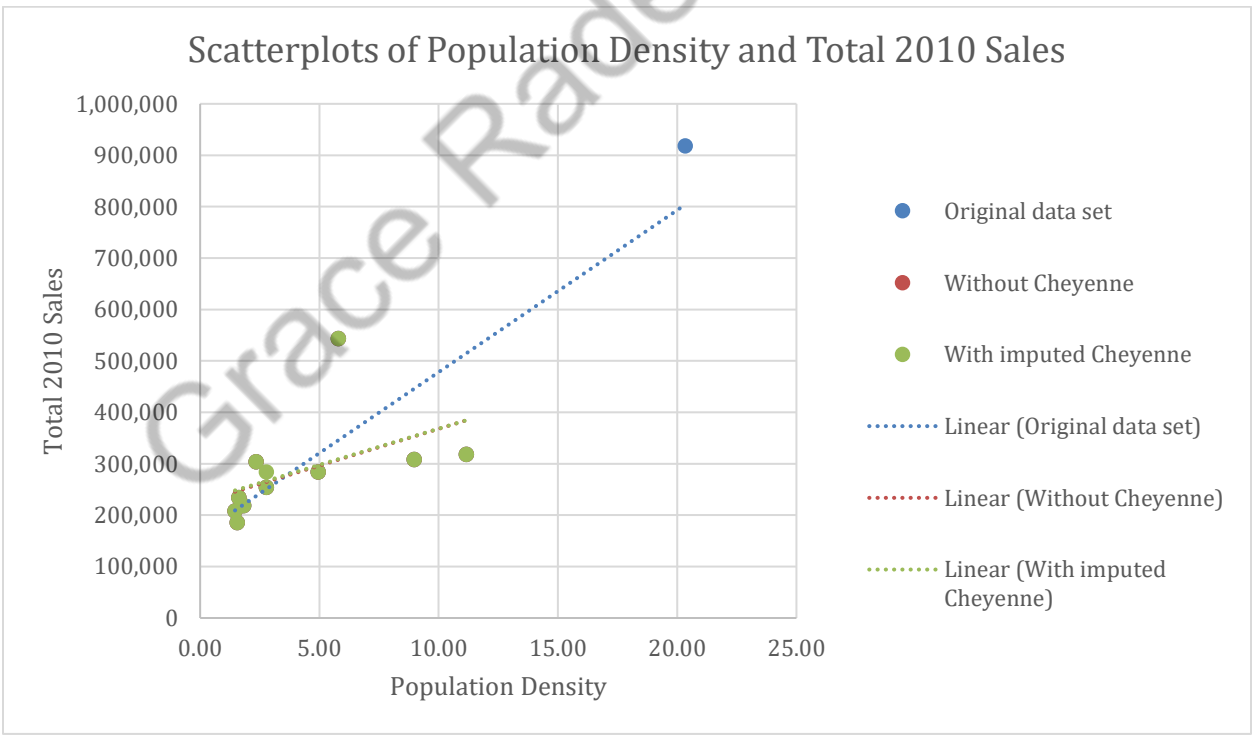


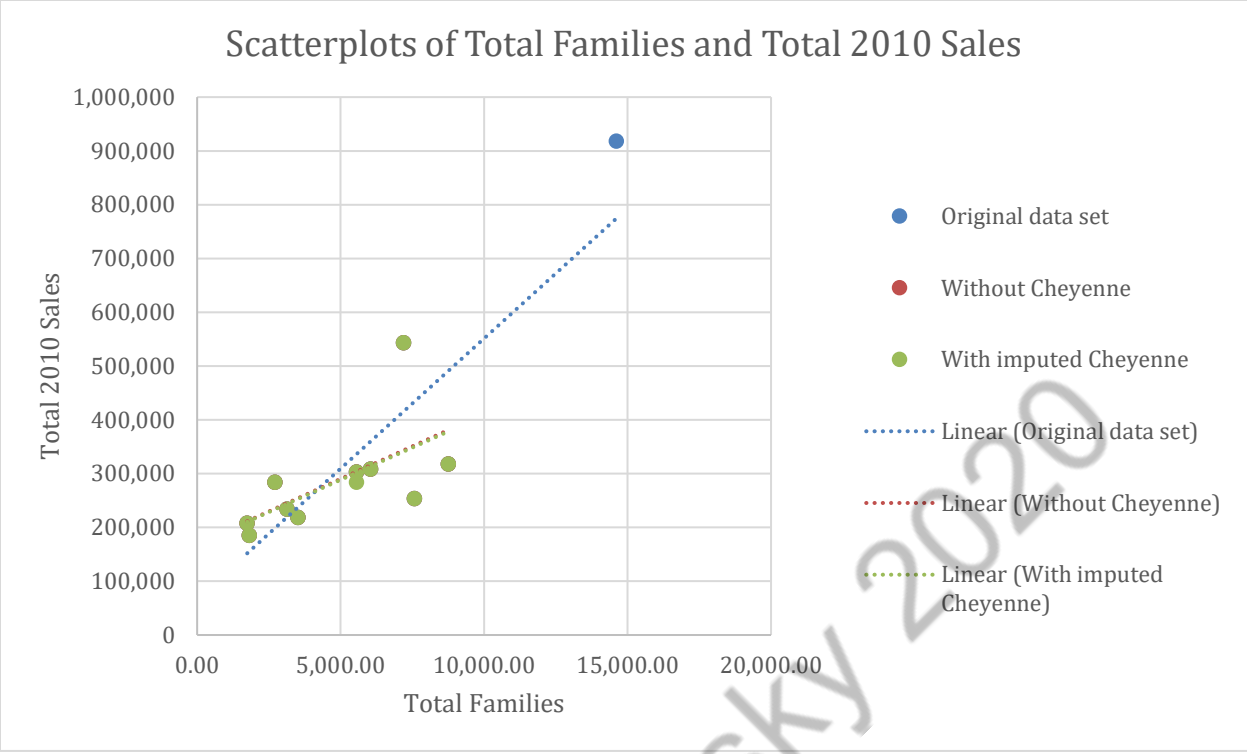
Similar to the scatterplots of Rock Springs, all three corresponding linear lines are roughly the same with similar slopes. Therefore, the outlier of Gillet will not likely impact our analytical model.

Cheyenne:

The city of Cheyenne has outliers across three fields: **Total 2010 Sales**, **2010 Census**, **Population Density**, and **Total Families**. Therefore, there is a strong potential that Cheyenne might impact the model. I will confirm that with the following three graphs, each plotting three scatterplots for 3 scenarios:







As can be seen from all three graphs above, the outliers cause the slopes of the linear lines to be steeper than when outliers are removed or imputed. This means that our predicted sales values are likely to be inflated if we keep this outlier in our model. In addition, the linear lines for the "without Cheyenne" scenario and the "with imputed Cheyenne" scenario are the same so I can choose either removing Cheyenne or imputing Cheyenne.