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?

The decision needs to be made is, “what is the profitable city for Pawdacity company to open its newest store in?” and that will be done by predicting the yearly sales of each city in Wyoming.

2. What data is needed to inform those decisions?

We need to predict the yearly sales for each city. In order to do this, first we need to combine data from different datasets together. This data includes the monthly sales data, the population data, and the demographic data for each city in the state of Wyoming. Then, we need to aggregate the monthly sales of Pawdacity’s stores in each city in order to get the total yearly sales for each city which based on we will decide which city will be the profitable one to open the new store in.

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*	19,442
Total Pawdacity Sales	*3,773,304*	343,028
Households with Under 18	*34,064*	3,097
Land Area	*33,071*	3,006.45
Population Density	*63*	5.73
Total Families	*62,653*	5,695.73

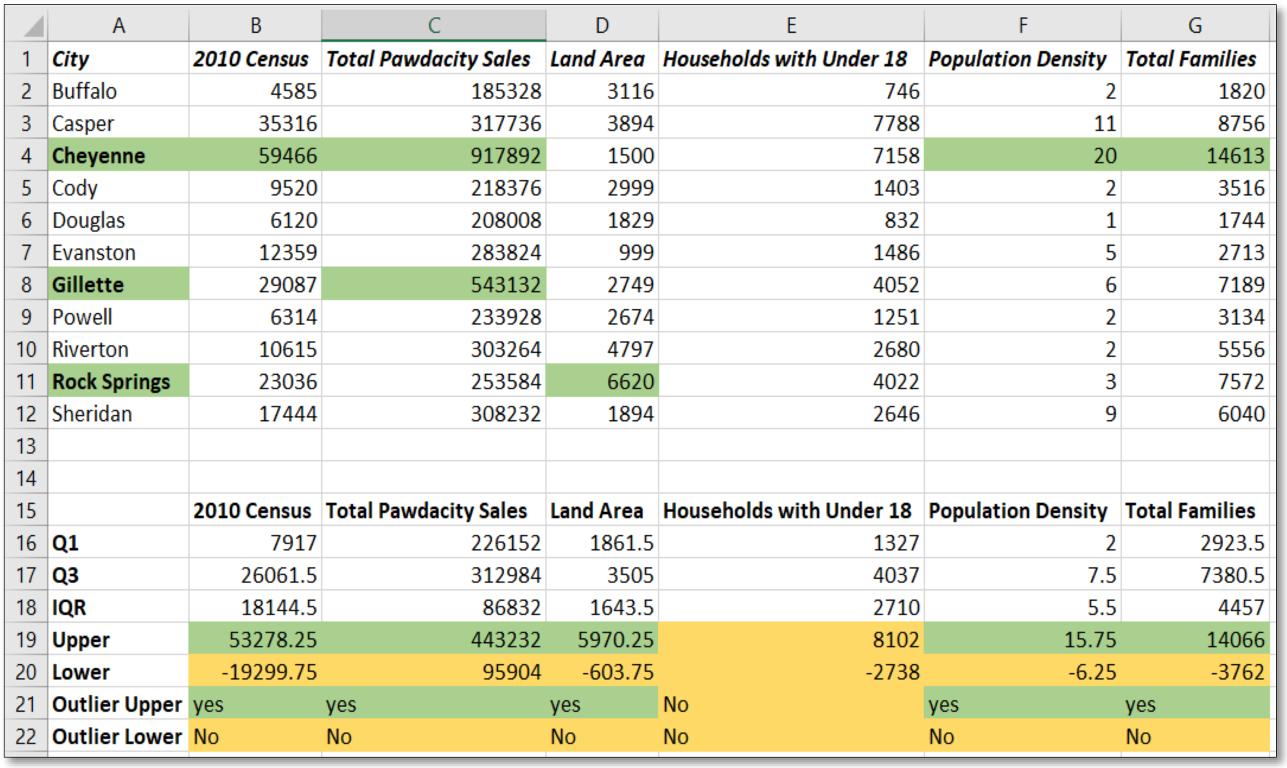
Step 3: Dealing with Outliers

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.

<u>First:</u> Are there any cities that are outliers in the training set?

To answer this question, we need to calculate the upper fence and the lower fence for each field in the dataset by using Excel. So, eventually any value above the upper or below the lower fences will be considered as an outlier.



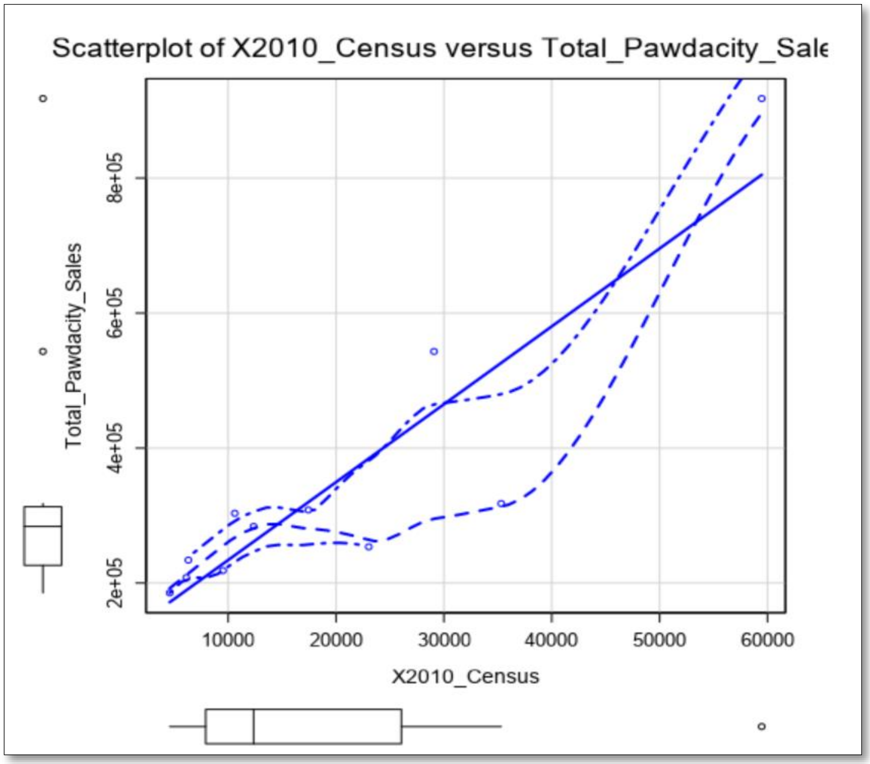
	A	B	C	D	E	F	G
1	***City***	***2010 Census***	***Total Pawdacity Sales***	***Land Area***	***Households with Under 18***	***Population Density***	***Total Families***
2	Buffalo	4585	185328	3116	746	2	1820
3	Casper	35316	317736	3894	7788	11	8756
4	**Cheyenne**	59466	917892	1500	7158	20	14613
5	Cody	9520	218376	2999	1403	2	3516
6	Douglas	6120	208008	1829	832	1	1744
7	Evanston	12359	283824	999	1486	5	2713
8	**Gillette**	29087	543132	2749	4052	6	7189
9	Powell	6314	233928	2674	1251	2	3134
10	Riverton	10615	303264	4797	2680	2	5556
11	**Rock Springs**	23036	253584	6620	4022	3	7572
12	Sheridan	17444	308232	1894	2646	9	6040
13							
14							
15		**2010 Census**	**Total Pawdacity Sales**	**Land Area**	**Households with Under 18**	**Population Density**	**Total Families**
16	**Q1**	7917	226152	1861.5	1327	2	2923.5
17	**Q3**	26061.5	312984	3505	4037	7.5	7380.5
18	**IQR**	18144.5	86832	1643.5	2710	5.5	4457
19	**Upper**	53278.25	443232	5970.25	8102	15.75	14066
20	**Lower**	-19299.75	95904	-603.75	-2738	-6.25	-3762
21	**Outlier Upper**	yes	yes	yes	No	yes	yes
22	**Outlier Lower**	No	No	No	No	No	No

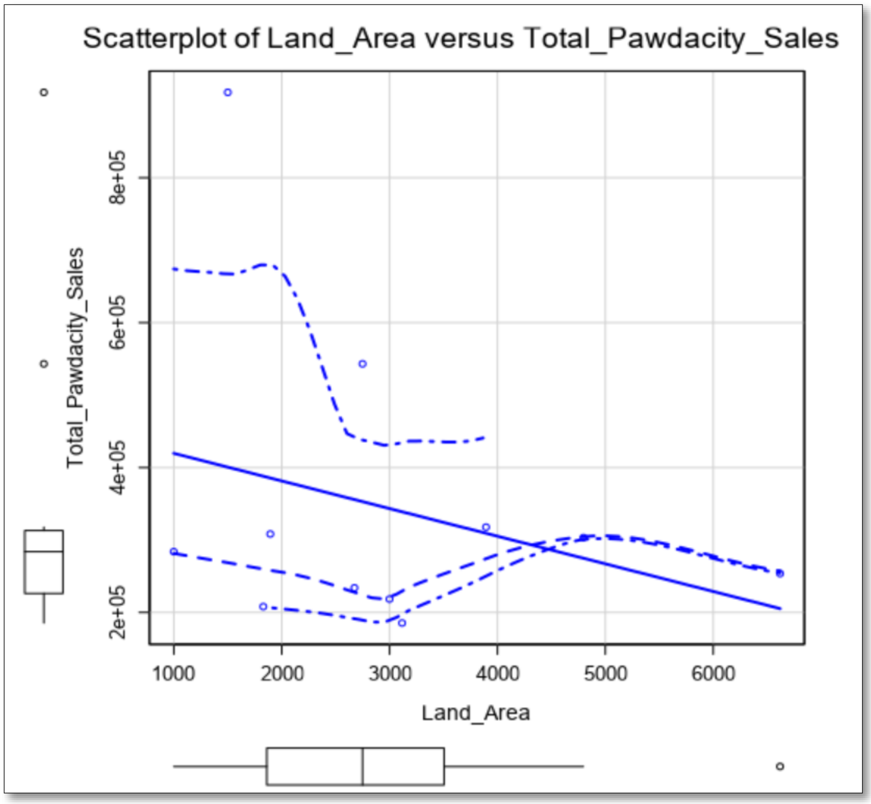
As shown in the above table, there are 3 outliers cities which are: Cheyenne, Gillette, and Rock Springs. Cheyenne city is an outlier in 4 fields which are: 2010 Census, Total Pawdacity Sales, Population Density, and Total Families. While, Gillette city is an outlier in only one field which is: Total Pawdacity Sales. Whereas, Rock Springs city is an outlier in only one field too which is: Land Area.

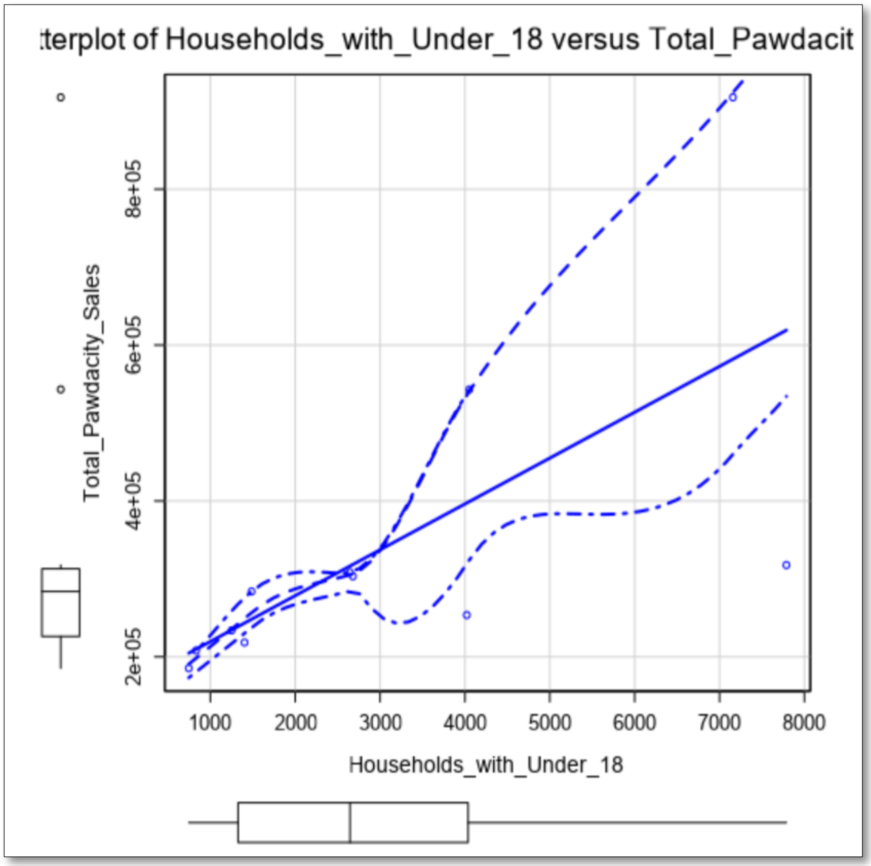
<u>Second: </u>Which outlier have you chosen to remove or impute? explain your reasoning.

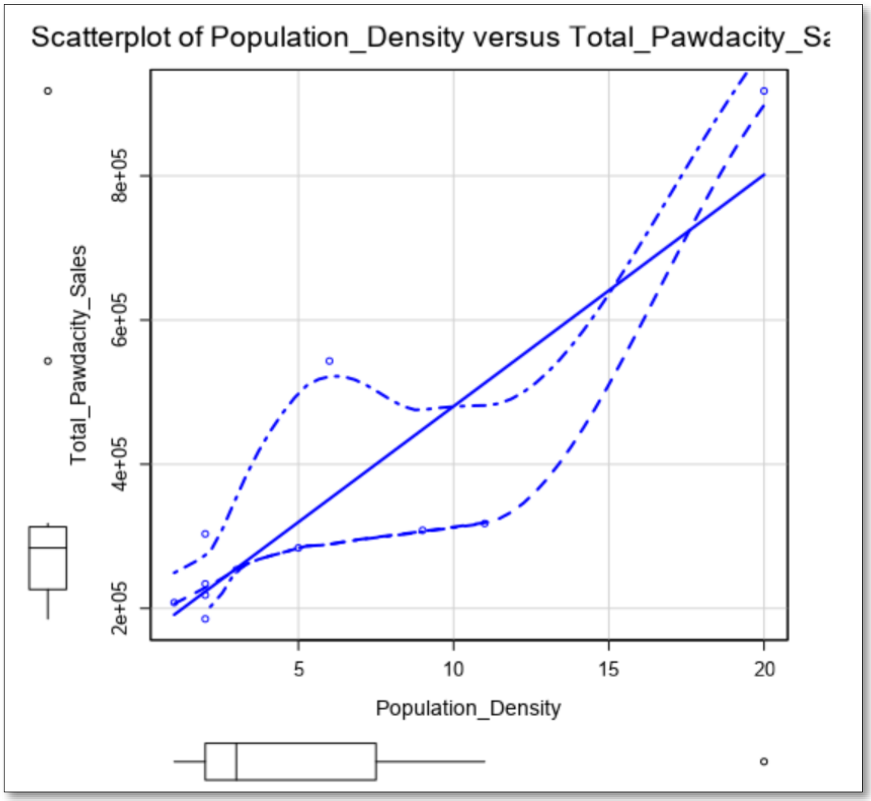
After doing a scatterplot analysis (see the below charts), I chose to remove Gillette outlier rather than Cheyenne outliers or Rock Springs outlier, because Gillette outlier skews high in sales, and does not skew relative to the other sales data in the training-set. Whereas, I chose to retain Cheyenne outliers and Rock Springs outlier because they are in line with the linear relationship. Furthermore, even though Cheyenne has an outlier in Total Pawdacity Sales, 2010 Census, Population Density and Total Families, but still it makes sense, because if there are a lot of families and high population density in a city, definitely there will be more sales. So, we shouldn't remove this city outliers, since it might be the recommended city that we are looking for. Whereas, Gillette's other fields are at their average except Total Pawdacity Sales' field. So, we should remove this city outlier since it doesn't make sense.

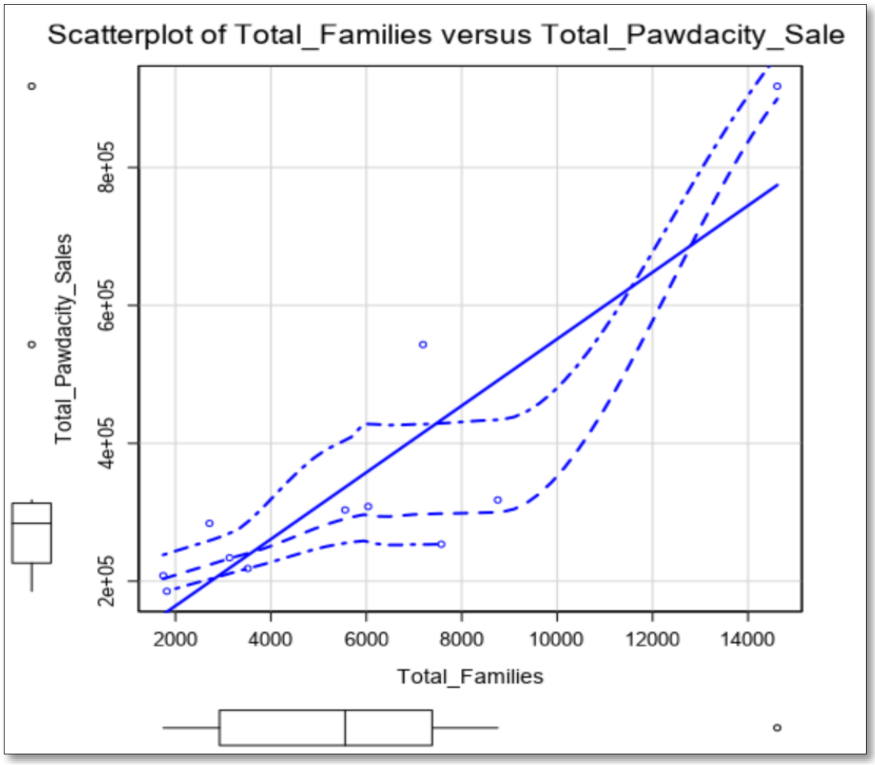
The following 5 scatterplots show the outliers in each predictive variable **before** removing Gillette's outlier form the target variable:



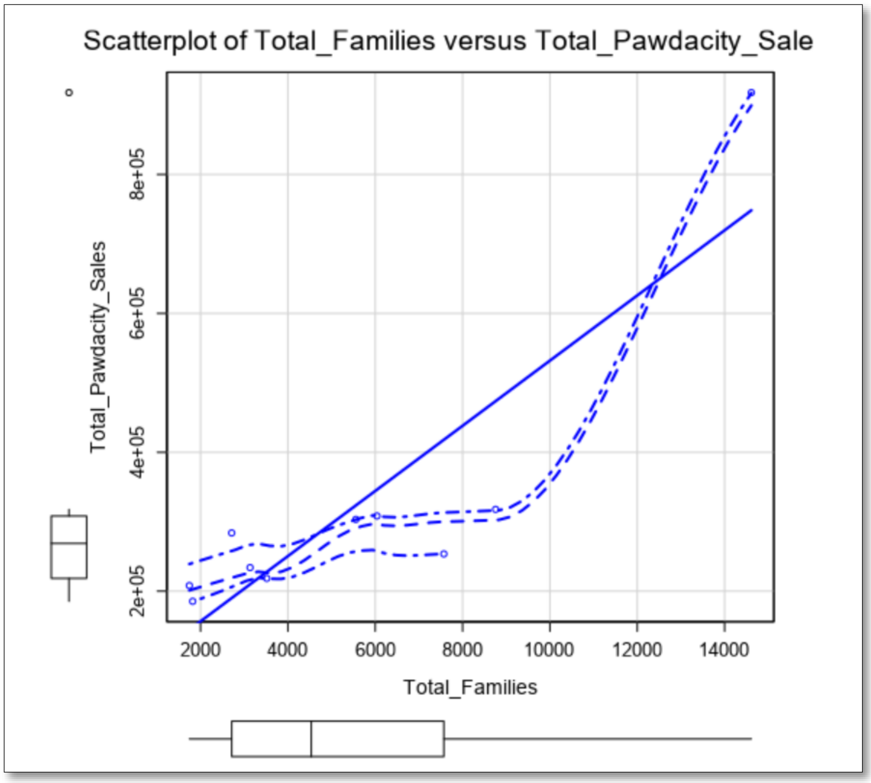

Scatterplot of Land_Area versus Total_Pawdacity_Sales
Total_Pawdacity_Sales
8e+05
6e+05
4e+05
2e+05
1000
2000
3000
4000
5000
6000
Land_Area


tterplot of Households_with_Under_18 versus Total_Pawdacit
Total_Pawdacity_Sales
8e+05
6e+05
4e+05
2e+05
1000
2000
3000
4000
5000
6000
7000
8000
Households_with_Under_18

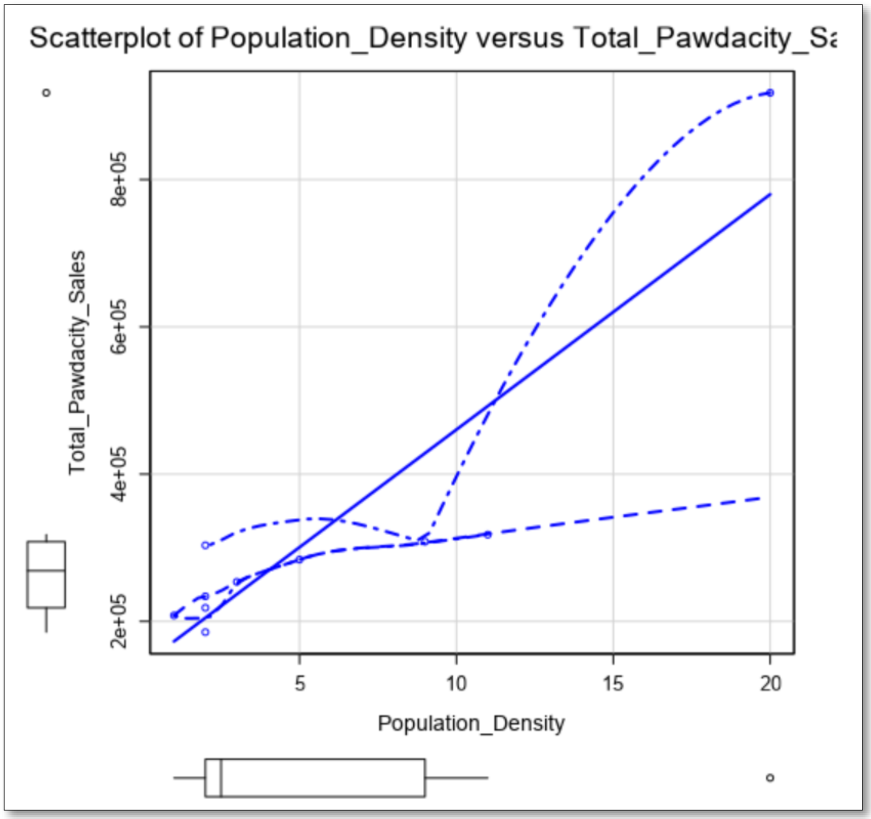

Scatterplot of Population_Density versus Total_Pawdacity_Sa
Total_Pawdacity_Sales
2e+05
4e+05
6e+05
8e+05
5
10
15
20
Population_Density


Scatterplot of Total_Families versus Total_Pawdacity_Sale
Total_Pawdacity_Sales
2e+05
4e+05
6e+05
8e+05
2000
4000
6000
8000
10000
12000
14000
Total_Families

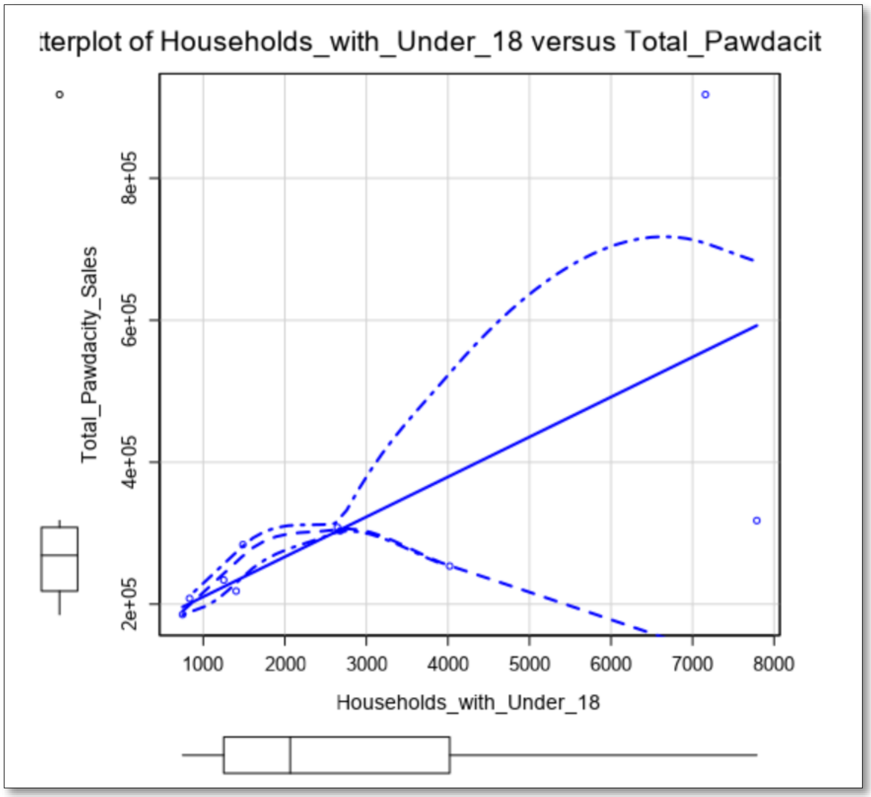
The following 5 scatterplots show what happened **after** removing Gillette's outlier:



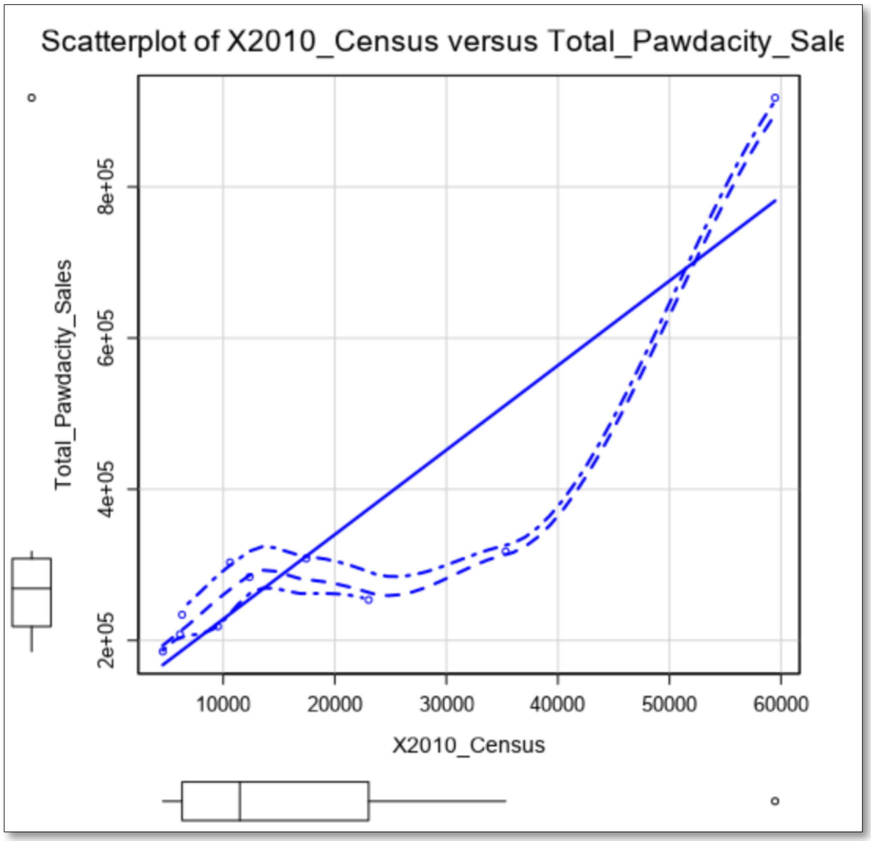
As you can see above, the outlier in the Total_Families' predictive variable has been gone after removing the Gillette's outlier from the Total_Pawdacity_Sales' target variable.



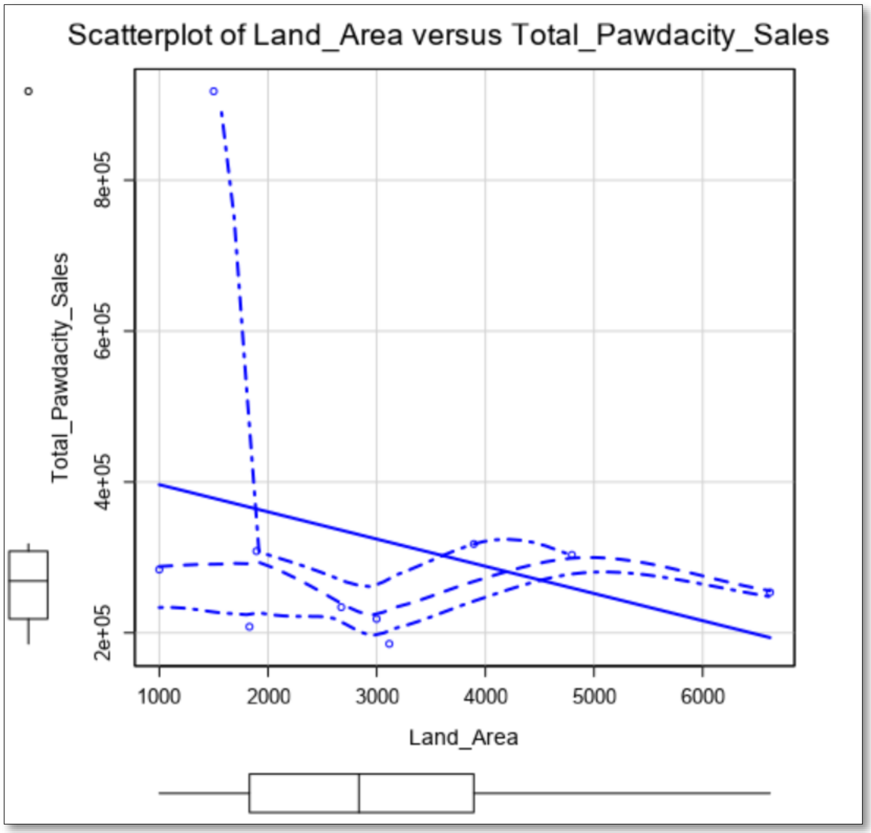
As you can see above, there is an outlier in the Population_Density’s predictive variable which is also an outlier for the Total_Pawdacity_Sales’ target variable. This makes sense, since we would expect the relationship to behave this way. Based on the fitted line, the outlier is in line with the relationship, so I will keep it in.



As you can see above, there is no outliers in the Households_with_Under_18 predictive variable which is the same before removing the Gillette's outlier.



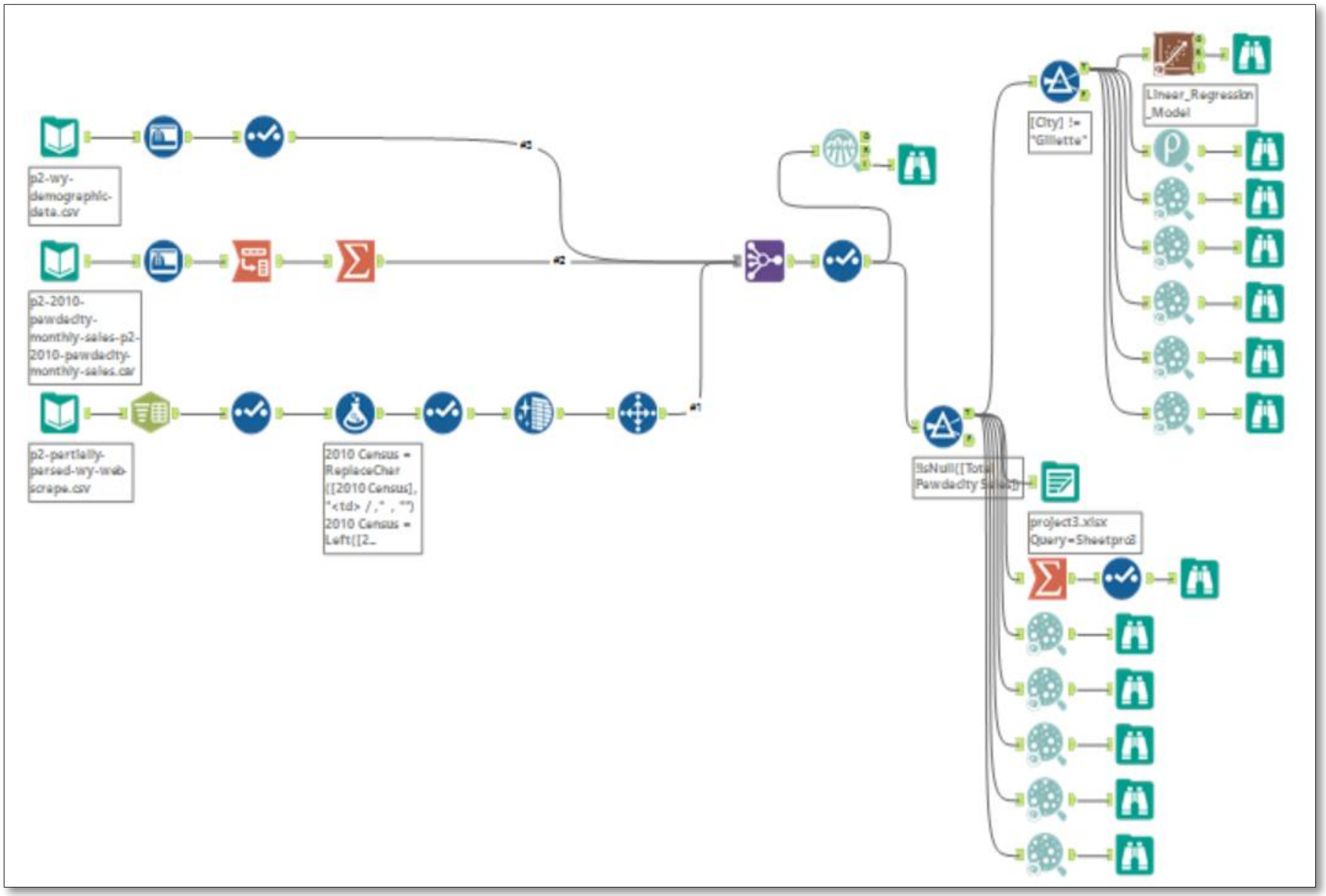
As you can see above, the outlier in the 2010_Census's predictive variable which is also an outlier for the Total_Pawdacity_Sales' target variable. This makes sense, since we would expect the relationship to behave this way. Based on the fitted line, the outlier is in line with the relationship, so I will keep it in.



As you can see above, the outlier in the Land_Area's predictive variable has been gone after removing the Gillette's outlier from the Total_Pawdacity_Sales's target variable.

To sum up, removing the Gillette city's outlier is the best choice for the dataset, since it minimized the number of the outliers in the dataset resulting in removing 2 other outliers which are the Total Families and the Land Area outliers.

Alteryx’s workflow:


p2-wy-demographic-data.csv
p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales.csv
p2-partially-parsed-wy-web-scrape.csv
2010 Census = ReplaceChar([2010 Census], "<td> /," , "") 2010 Census = Left([2…
#1
#2
#3
[City] != "Gillette"
Linear_Regression_Model
!IsNull([Total Pawdacity Sales])
project3.xlsx Query=Sheetpro3

Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.