Minseok (Richard) Park
Predictive Analytics for Business Nanodegree

Project 2.1: Data Cleanup

The Business Problem

Pawdacity is a leading pet store chain in Wyoming with 13 stores throughout the state. This year, Pawdacity would like to expand and open a 14th store. Your manager has asked you to perform an analysis to recommend the city for Pawdacity's newest store, based on predicted yearly sales.

Your first step in predicting yearly sales is to first format and blend together data from different datasets and deal with outliers.

You are given the following information to work with:

1. The monthly sales data for all of the Pawdacity stores for the year 2010.
2. NAICS data on the most current sales of all competitor stores where total sales is equal to 12 months of sales.
3. A partially parsed data file that can be used for population numbers.
4. Demographic data (Households with individuals under 18, Land Area, Population Density, and Total Families) for each city and county in the state of Wyoming. For people who are unfamiliar with the US city system, a state contains counties and counties contains one or more cities.

Business and Data Understanding

1. What decisions needs to be made?

The decisions that needs to be made through this analysis is to clean the given datasets, select necessary data, join them and remove outliers so that we can perform linear regression analysis and suggest a location for Pawdacity's new 14th pet store.

2. What data is needed to inform those decisions?

In order to suggest a location for Pawdacity's new 14th pet store, we would need a dataset consisting historical records of monthly sales and demographics of the existing stores in Wyoming, and total (monthly) sales of all the existing competitor stores in Wyoming.

Building the Training Set

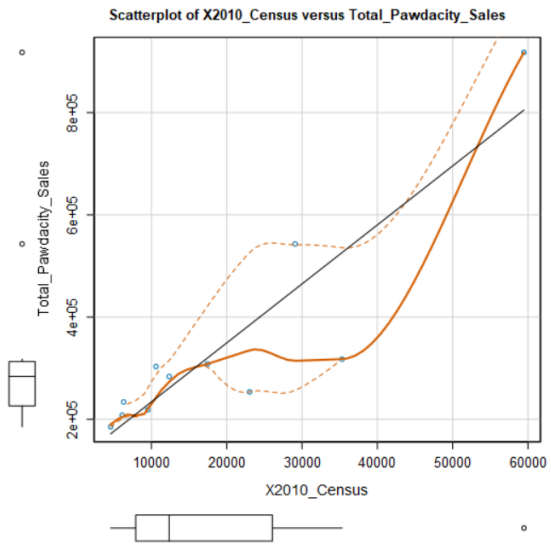
After the process of data cleaning (Used tools in Alteryx: 'Auto Fields', 'Formula', 'Select', 'Text to Columns', 'Filter', 'Data Cleansing', 'Join', 'Summarize'), I was able to retrieve total and average values for each column. Total (sum), and average values for each column are as follows:

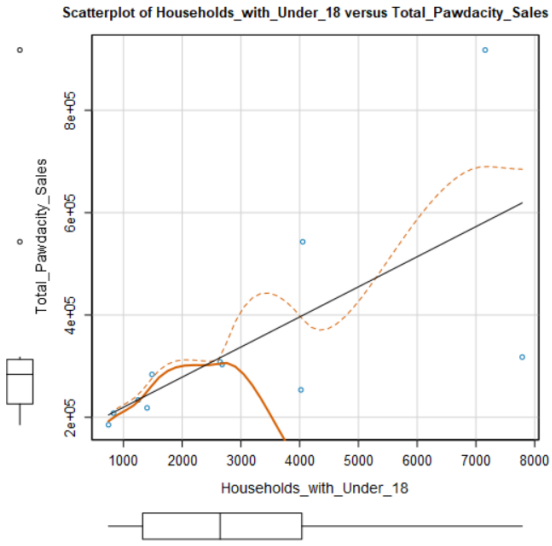
Minseok (Richard) Park
Predictive Analytics for Business Nanodegree

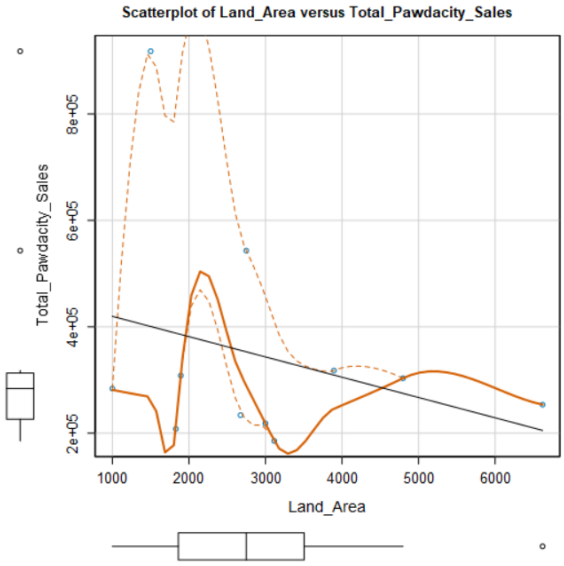
Column	Sum	Average
Census Population	*213,862*	***19,442***
Total Pawdacity Sales	*3,773,304*	***343,027.64***
Households with Under 18	*34,064*	***3,096.73***
Land Area	*33,071*	***3,006.49***
Population Density	*63*	***5.71***
Total Families	*62,653*	***5,695.71***

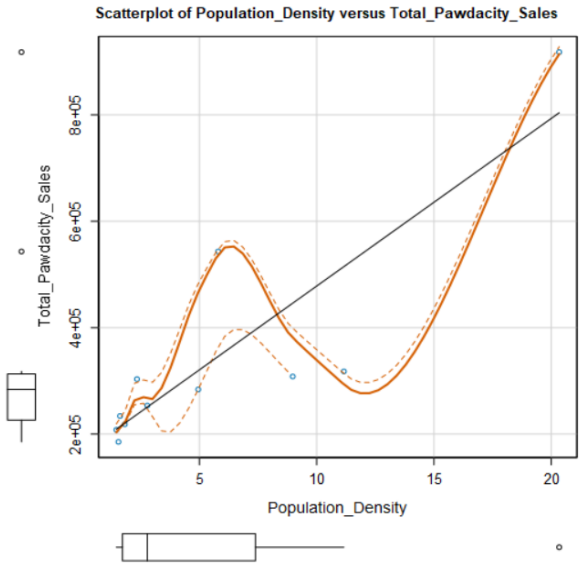
Dealing with Outliers

After cleaning the given datasets and put them into one whole dataset, scatterplots of each variables vs predictor variable ('Total Pawdacity Sales') was performed. The five scatterplots are shown below.

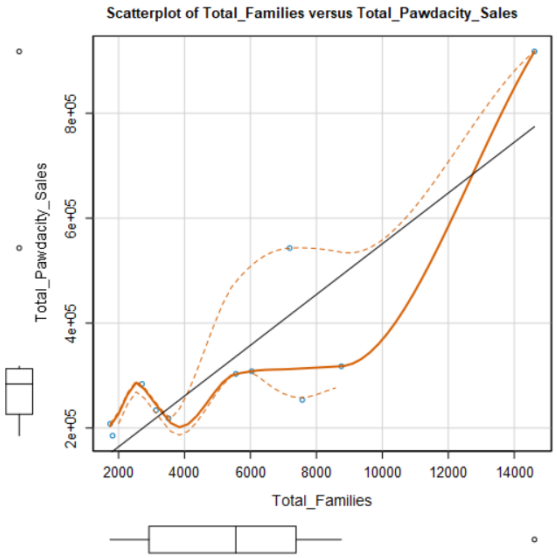




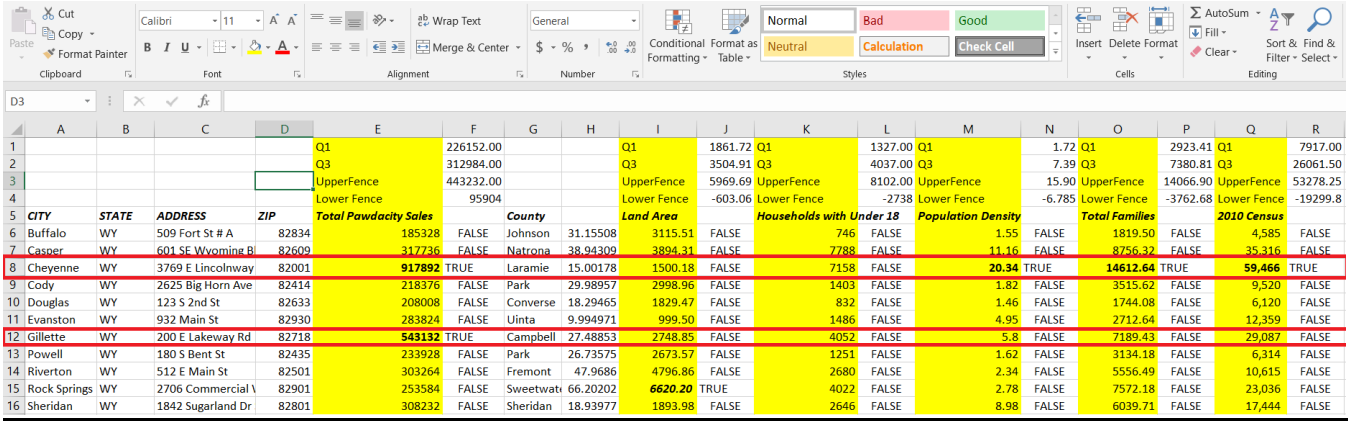




Minseok (Richard) Park
Predictive Analytics for Business Nanodegree



Based on the scatterplots of 5 explanatory variables vs Total Pawdacity Sales, there seem to be two possible outliers that have extremely high sales data. Therefore, final version of the data was exported in Microsoft Excel, and 'IQR', 'Upper' and 'Lower' fences were calculated to see possible outliers in detail. According to upper and lower fence restrictions, two observations seem to be outliers: City Cheyenne and City Gillette.



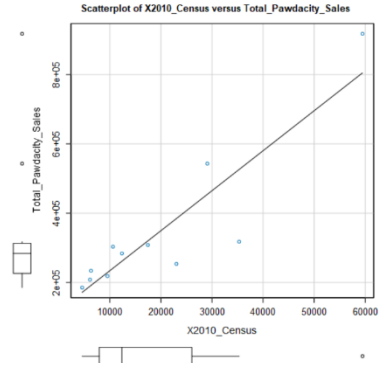
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1					Q1	226152.00			Q1	1861.72	Q1	1327.00	Q1	1.72	Q1	2923.41	Q1	7917.00
2					Q3	312984.00			Q3	3504.91	Q3	4037.00	Q3	7.39	Q3	7380.81	Q3	26061.50
3					UpperFence	443232.00			UpperFence	5969.69	UpperFence	8102.00	UpperFence	15.90	UpperFence	14066.90	UpperFence	53278.25
4					Lower Fence	95904			Lower Fence	-603.06	Lower Fence	-2738	Lower Fence	-6.785	Lower Fence	-3762.68	Lower Fence	-19299.8
5	***CITY***	***STATE***	***ADDRESS***	***ZIP***	***Total Pawdacity Sales***		***County***		***Land Area***		***Households with Under 18***		***Population Density***		***Total Families***		***2010 Census***	
6	Buffalo	WY	509 Fort St # A	82834	185328	FALSE	Johnson	31.15508	3115.51	FALSE	746	FALSE	1.55	FALSE	1819.50	FALSE	4,585	FALSE
7	Casper	WY	601 SE Wyoming Bl	82609	317736	FALSE	Natrona	38.94309	3894.31	FALSE	7788	FALSE	11.16	FALSE	8756.32	FALSE	35,316	FALSE
8	Cheyenne	WY	3769 E Lincolnway	82001	**917892**	TRUE	Laramie	15.00178	1500.18	FALSE	7158	FALSE	**20.34**	TRUE	**14612.64**	TRUE	**59,466**	TRUE
9	Cody	WY	2625 Big Horn Ave	82414	218376	FALSE	Park	29.98957	2998.96	FALSE	1403	FALSE	1.82	FALSE	3515.62	FALSE	9,520	FALSE
10	Douglas	WY	123 S 2nd St	82633	208008	FALSE	Converse	18.29465	1829.47	FALSE	832	FALSE	1.46	FALSE	1744.08	FALSE	6,120	FALSE
11	Evanston	WY	932 Main St	82930	283824	FALSE	Uinta	9.994971	999.50	FALSE	1486	FALSE	4.95	FALSE	2712.64	FALSE	12,359	FALSE
12	Gillette	WY	200 E Lakeway Rd	82718	**543132**	TRUE	Campbell	27.48853	2748.85	FALSE	4052	FALSE	5.8	FALSE	7189.43	FALSE	29,087	FALSE
13	Powell	WY	180 S Bent St	82435	233928	FALSE	Park	26.73575	2673.57	FALSE	1251	FALSE	1.62	FALSE	3134.18	FALSE	6,314	FALSE
14	Riverton	WY	512 E Main St	82501	303264	FALSE	Fremont	47.9686	4796.86	FALSE	2680	FALSE	2.34	FALSE	5556.49	FALSE	10,615	FALSE
15	Rock Springs	WY	2706 Commercial \	82901	253584	FALSE	Sweetwat	66.20202	***6620.20***	TRUE	4022	FALSE	2.78	FALSE	7572.18	FALSE	23,036	FALSE
16	Sheridan	WY	1842 Sugarland Dr	82801	308232	FALSE	Sheridan	18.93977	1893.98	FALSE	2646	FALSE	8.98	FALSE	6039.71	FALSE	17,444	FALSE

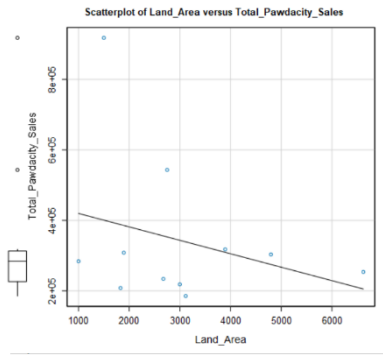
Because it is not easy to see if these points (Cheyenne and Gillette) are either outliers or abnormal points, I built two models excluding each city, and compared the scatterplots of the model with the model I attained originally. Please see below comparison of scatterplots with and without possible outlier.

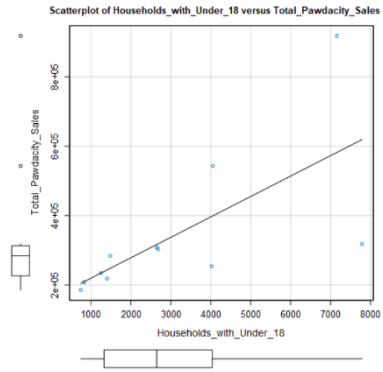
Minseok (Richard) Park
Predictive Analytics for Business Nanodegree

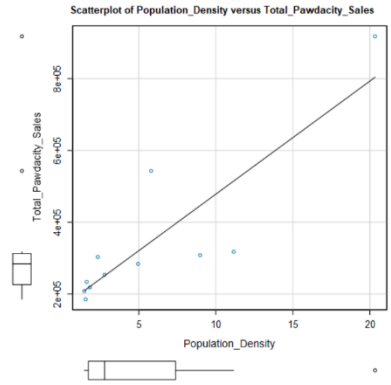
1. Cheyenne

Including City Cheyenne

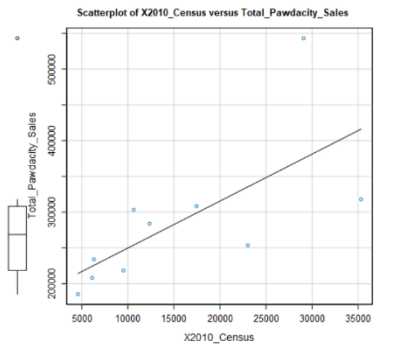


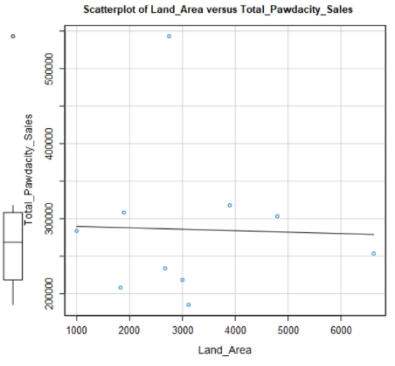


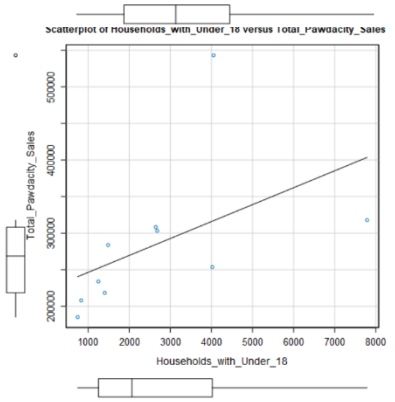


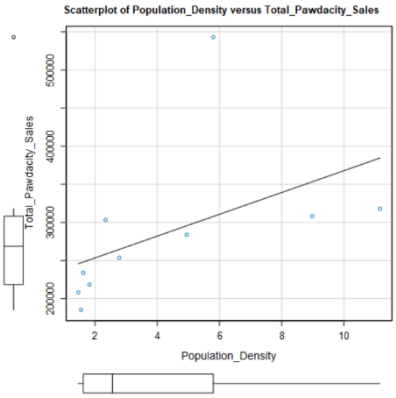


Excluding City Cheyenne

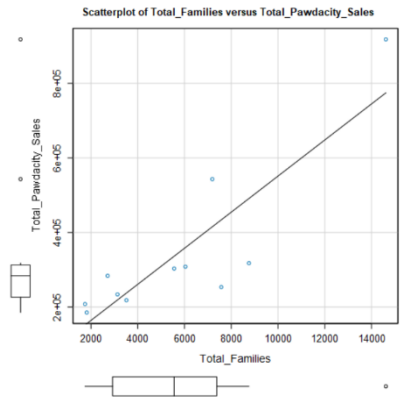


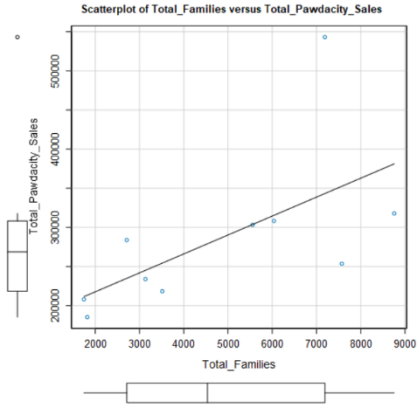






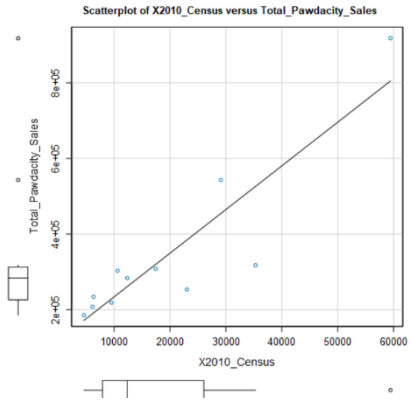
Minseok (Richard) Park
Predictive Analytics for Business Nanodegree

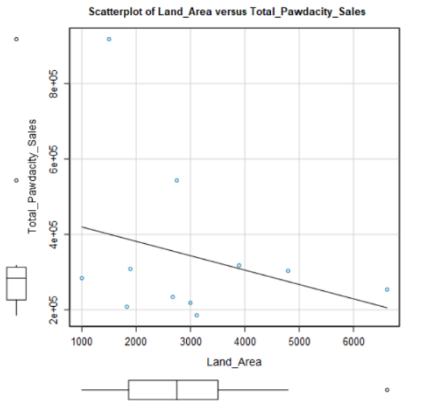


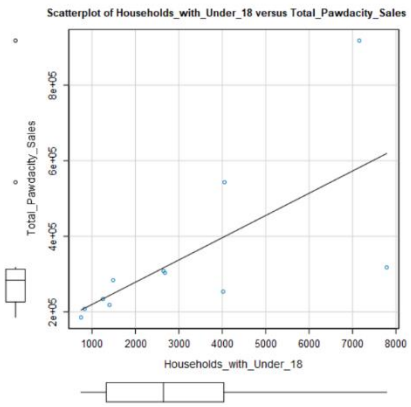


2. ***Gillette***

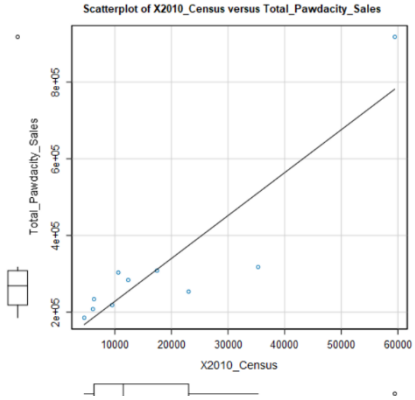
Including City of Gillette

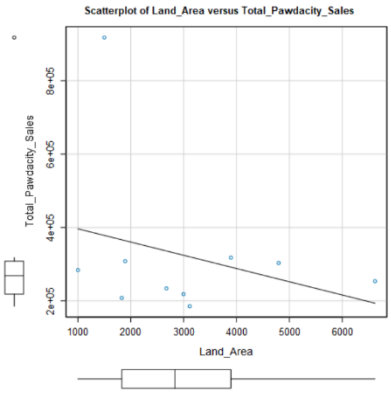


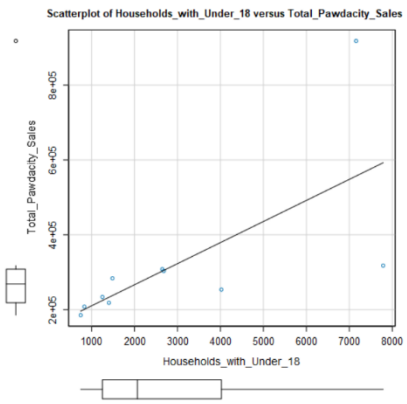




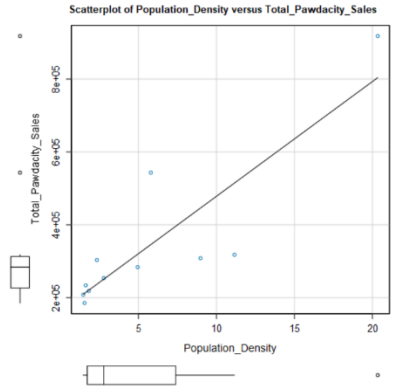
Excluding City of Gillette

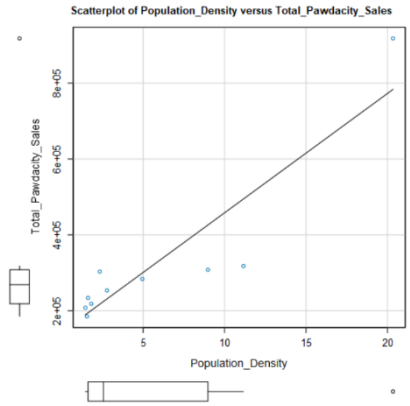


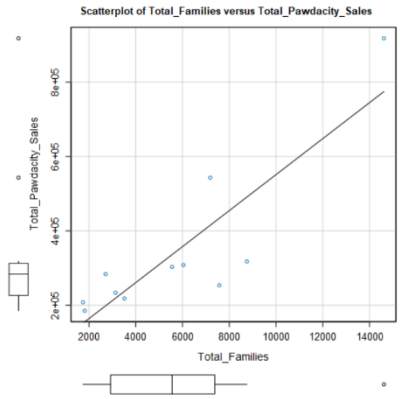


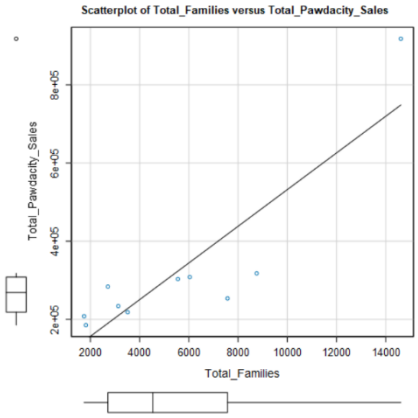


Minseok (Richard) Park
Predictive Analytics for Business Nanodegree









1. City of Cheyenne

Before looking at the scatterplots, this city seemed to be significant outlier. However, based on the comparisons of the scatterplots, it is noticeable that if we omit this observation, some of the relationships of the variables could change. For example, Land area vs Pawdacity Sales, we have significant decreasing slope if we keep the City of Cheyenne observation. However, if we decide to consider this data point as outlier and omit it, relationship between Land Area and Pawdacity Sales become rather a flat line. Therefore, we should also consider a possibility that this data point may be just an abnormal point.

2. City of Gillette

On the other hand, City of Gillette, could be considered as an outlier. Omitting this row of data does not affect the general relationship between explanatory variables and Total Sales that removing this observation does not significantly increase or decrease the slope of the line. Therefore, it can be legitimate to remove this observation from the dataset.

Minseok (Richard) Park
Predictive Analytics for Business Nanodegree

Alteryx Workflow