Project Creating Analytical Dataset

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made.

Key Decisions:

The company is looking to expand its operations and open the 14th store. The objective of the project is to identify and recommend location (city) for the newest store based on the predicted yearly revenue. So, we are tasked with collecting and cleaning data to create a test data set.

We would need historical sales data. To compare our business operation with our competitors, we need a competitive set. With a competitive set, we could identify potential markets we haven't penetrated or areas where our market share is low. To understand our customer base, we would need the census and demographic data.

Step 2: Building the Training Set

To build a predictor model I had to create training dataset which included the target variable Total Pawdacity Sales, and predictor variables: Census Population, Households with Under 18, Land Area, Population Density and Total Families.
I used power query to clean and joined the data at the city level granularity. To validate my data set I checked the sum of variables.

SUM

CITY	2010Census	Sale	HouseholdUnder18	LandArea	PopDensity	TotalFamilies
Buffalo	4,585.00	185,328.00	746.00	3116	2	1820
Casper	35,316.00	317,736.00	7,788.00	3894	11	8756
Cheyenne	59,466.00	917,892.00	7,158.00	1500	20	14613
Cody	9,520.00	218,376.00	1,403.00	2999	2	3516
Douglas	6,120.00	208,008.00	832.00	1829	1	1744
Evanston	12,359.00	283,824.00	1,486.00	999	5	2713
Gillette	29,087.00	543,132.00	4,052.00	2749	6	7189
Powell	6,314.00	233,928.00	1,251.00	2674	2	3134
Riverton	10,615.00	303,264.00	2,680.00	4797	2	5556
Rock Springs	23,036.00	253,584.00	4,022.00	6620	3	7572
Sheridan	17,444.00	308,232.00	2,646.00	1894	9	6040
Total	**213,862.00**	**3,773,304.00**	**34,064.00**	**33071**	**63**	**62653**

Average

CITY	2010Census	Sale	HouseholdUnder18	LandArea	PopDensity	TotalFamilies
Buffalo	4,585.00	185,328.00	746.00	3116	2	1820
Casper	35,316.00	317,736.00	7,788.00	3894	11	8756
Cheyenne	59,466.00	917,892.00	7,158.00	1500	20	14613
Cody	9,520.00	218,376.00	1,403.00	2999	2	3516
Douglas	6,120.00	208,008.00	832.00	1829	1	1744
Evanston	12,359.00	283,824.00	1,486.00	999	5	2713
Gillette	29,087.00	543,132.00	4,052.00	2749	6	7189
Powell	6,314.00	233,928.00	1,251.00	2674	2	3134
Riverton	10,615.00	303,264.00	2,680.00	4797	2	5556
Rock Springs	23,036.00	253,584.00	4,022.00	6620	3	7572
Sheridan	17,444.00	308,232.00	2,646.00	1894	9	6040
Total	**19,442.00**	**343,027.64**	**3,096.73**	**3006**	**6**	**5696**

Column	**Sum**	**Average**
Census Population	*213,862*	*19442*
Total Pawdacity Sales	*3,773,304*	*343027.64*
Households with Under 18	*34,064*	*3069.73*
Land Area	*33,071*	*3006.49*
Population Density	*63*	*5.71*
Total Families	*62,653*	*5695.71*

Step 3: Dealing with Outliers

For detecting outliers, I calculated the summary statistics for each variable and also plotted it on the graph for visualizing the data.

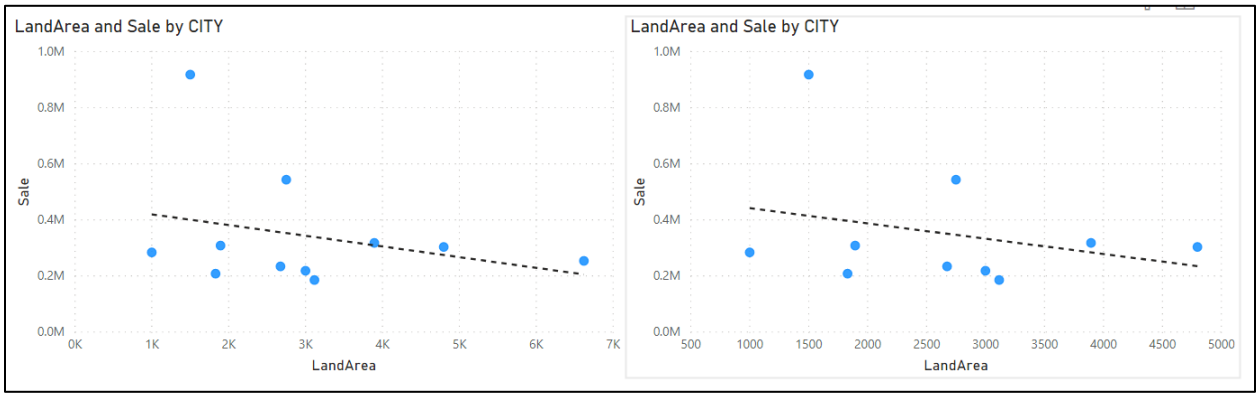
Sale	HouseholdUnder18	LandArea	PopDensity	TotalFamilies	2010Census	CITY
185328	746	3116	2	1820	4585	Buffalo
317736	7788	3894	11	8756	35316	Casper
917892	7158	1500	20	14613	59466	Cheyenne
218376	1403	2999	2	3516	9520	Cody
208008	832	1829	1	1744	6120	Douglas
283824	1486	999	5	2713	12359	Evanston
543132	4052	2749	6	7189	29087	Gillette
233928	1251	2674	2	3134	6314	Powell
303264	2680	4797	2	5556	10615	Riverton
253584	4022	6620	3	7572	23036	Rock Springs
308232	2646	1894	9	6040	17444	Sheridan

Column1	Sale	HouseholdUnder18	LandArea	PopDensity	TotalFamilies	2010Census
Q1	226152	1327	1861.5	2	2923.5	7917
Q3	312984	4037	3505	7.5	7380.5	26061.5
IQR	86832	2710	1643.5	5.5	4457	18144.5
UpperFence	443232	8102	5970.25	15.75	14066	53278.25
LowerFence	95904	-2738	-603.75	-6.25	-3762	-19299.75

Based on the Matrix below, we shortlisted three cities which had outliers; Rock Springs, Gillette and Cheyenne. Note all the graphs on the right are excluding the outlier.

Rock Spring

Rock Springs had an outlier in the land area variable.



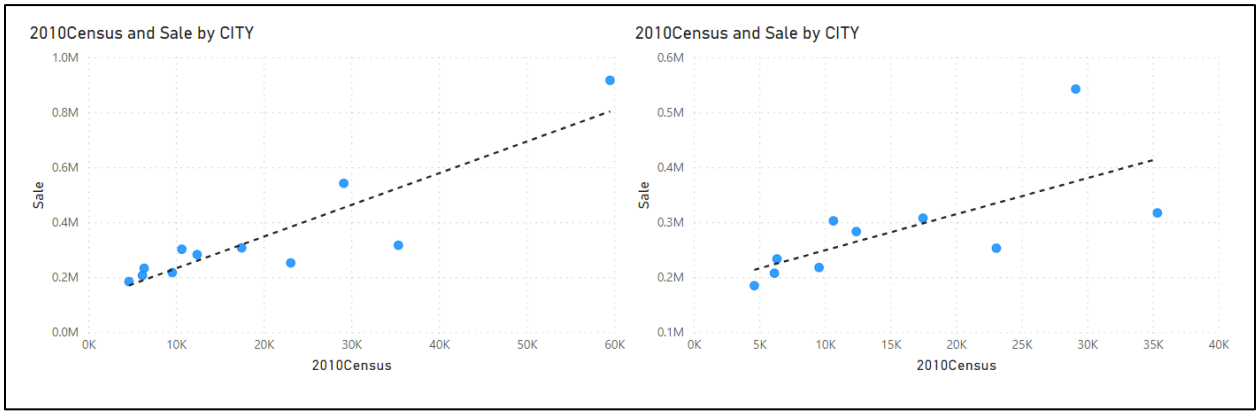
We have excluded the Rock Springs from the chart in the right. As you can see, the magnitude of the slop change does not warrant us to remove the city from the data set.

Cheyenne

The outliers are distributed across multiple variables for the city of Cheyenne; namely Population Density, Total Families and Census. We should drop this city based on the number of variables having an outlier. But to be safe we are also going to look at the graphs. And come to a conclusion of excluding the city Cheyenne from the dataset,

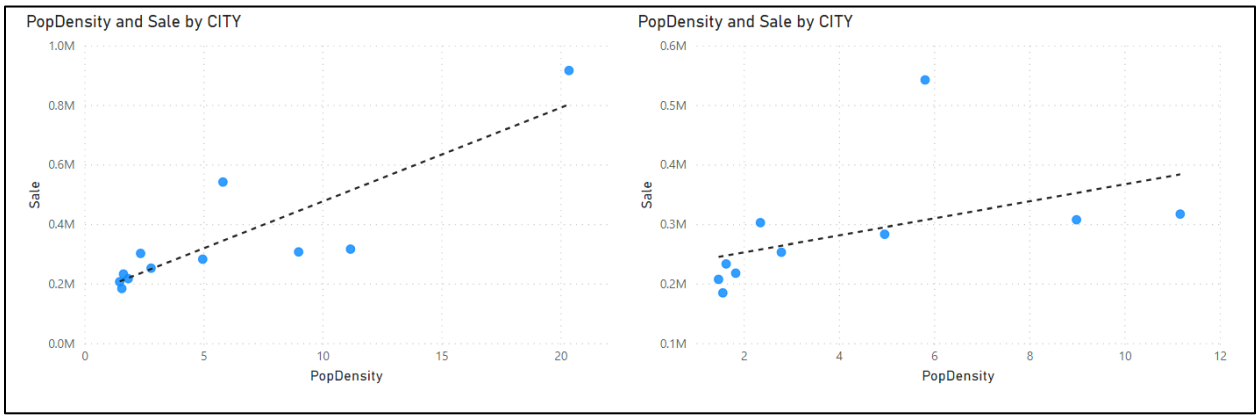
Census

The slope of the line decreases when the city is removed from the dataset.



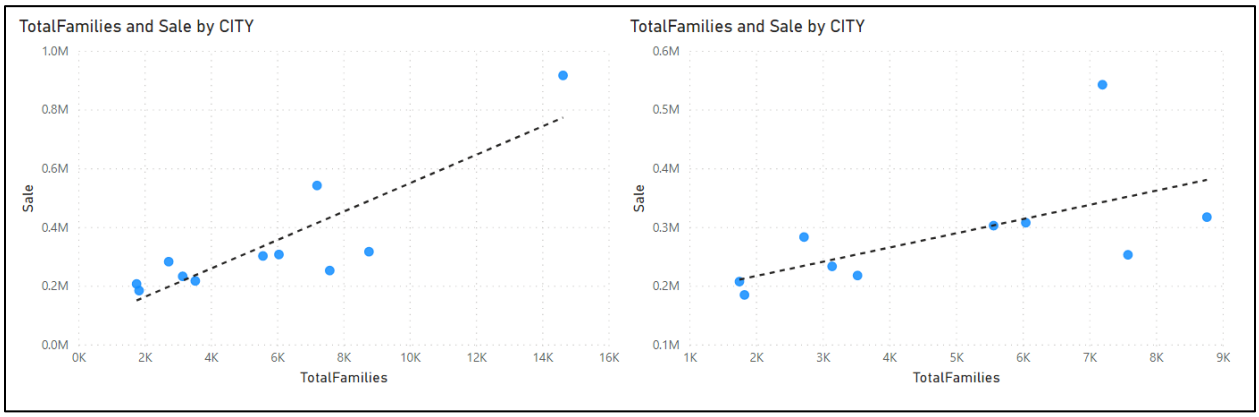
Population Density

The slope of the line decreases when the city is removed from the dataset.



Families

The slop of the line decreases when the city is removed from the dataset.



Thus, after extrapolating the graphs and data we have taken a call to remove the city of Cheyenne from the dataset.