Project 2.1: Data Cleanup

By: Fairoza Amira Binti Hamzah

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?

   We need to predict which city is the best to open the 14th store based on the previous sales data of each city.

2. What data is needed to inform those decisions?

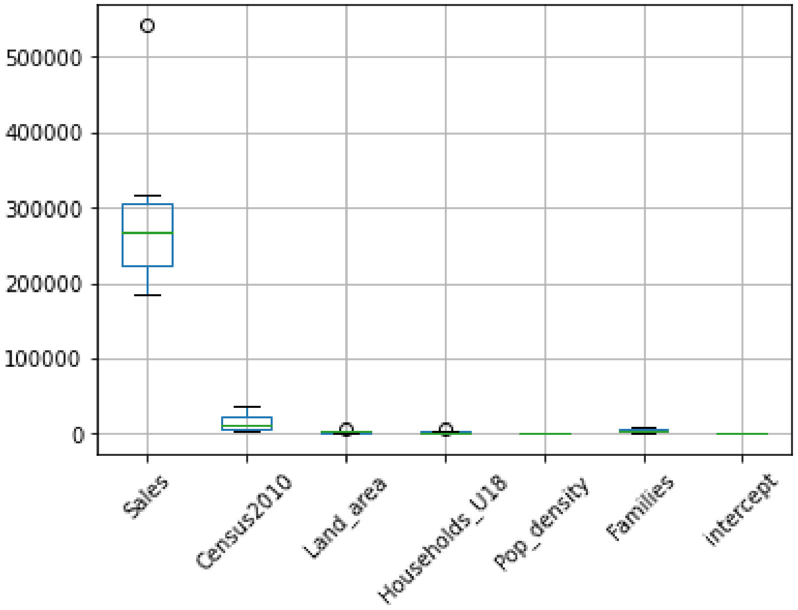
   City
   2010 census population
   Pawdacity sales in other stores
   competitor sales
   household with under 18
   land area
   population density
   total families

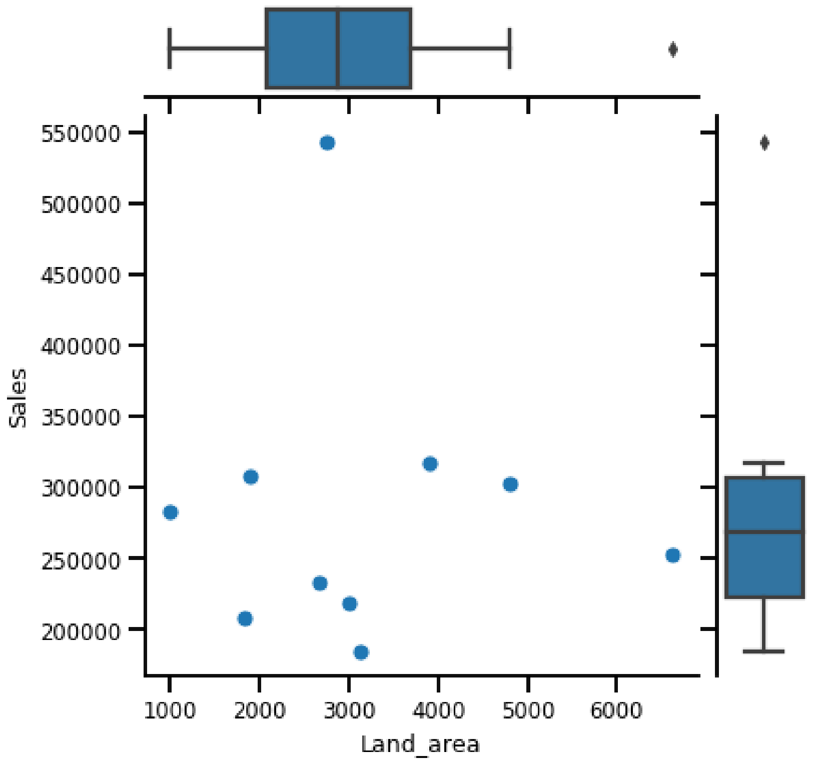
Step 2: Building the Training Set

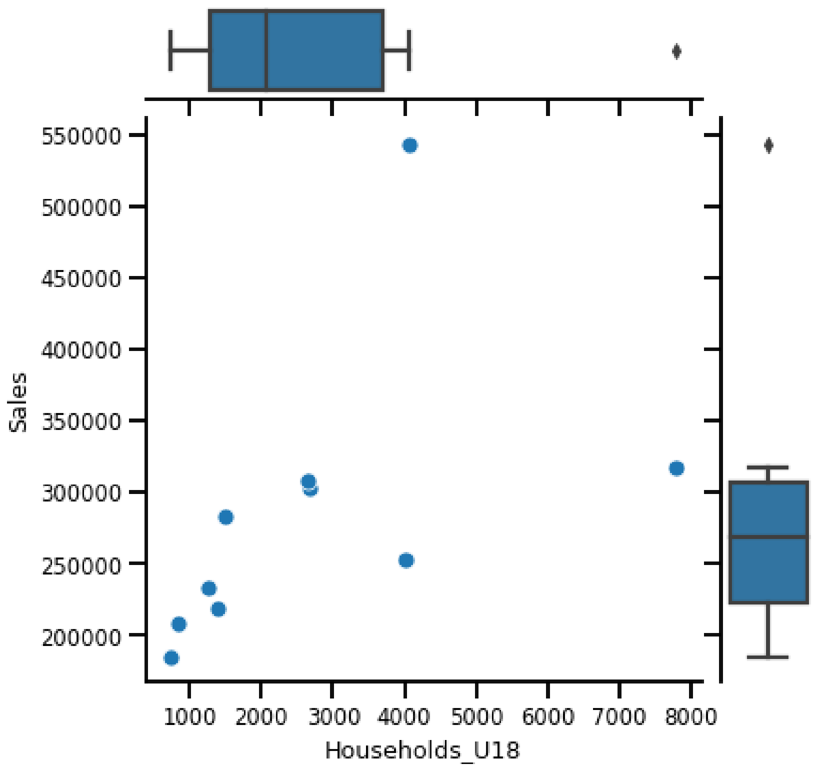
Column	**Sum**	**Average**
Census Population	*213,862*	*19442*
Total Pawdacity Sales	*3,773,304*	*343027.64*
Households with Under 18	*34,064*	*3096.73*
Land Area	*33,071*	*3006.49*
Population Density	*63*	*5.71*
Total Families	*62,653*	*5695.71*

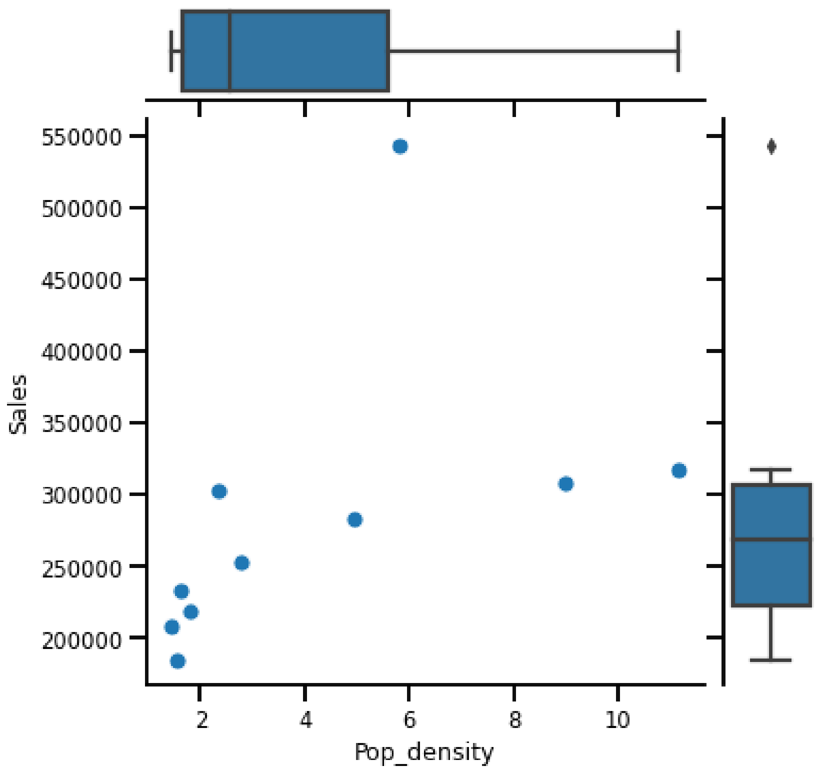
Step 3: Dealing with Outliers

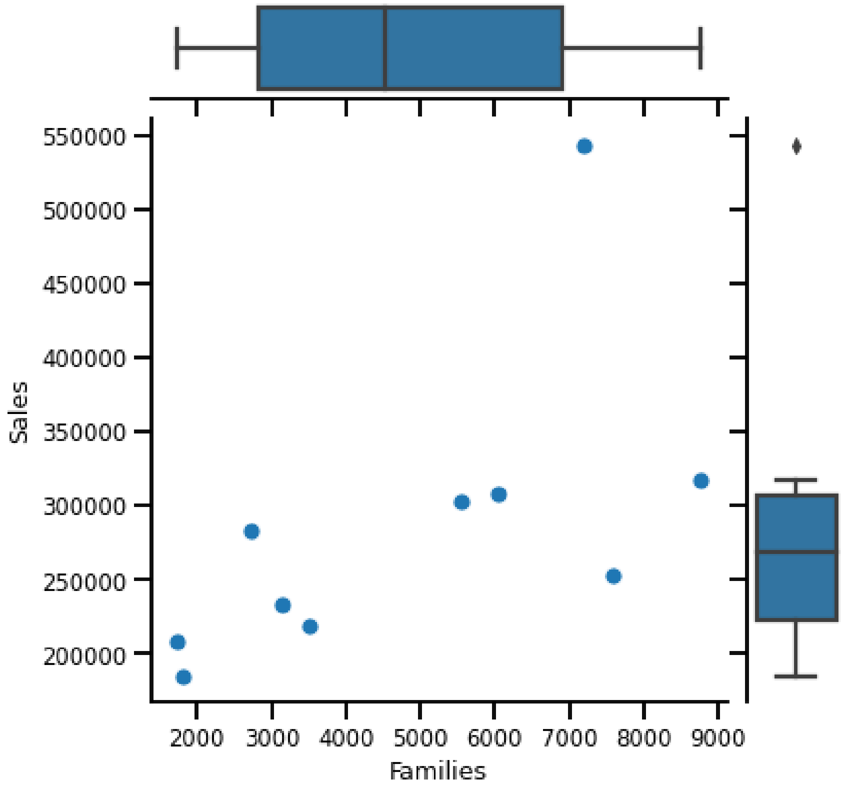
Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.

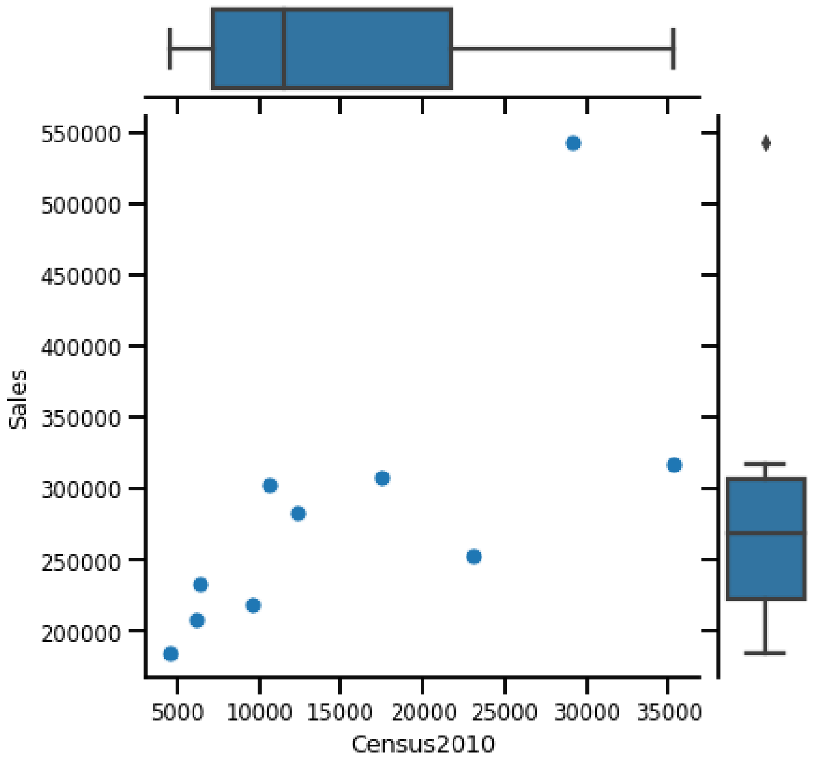




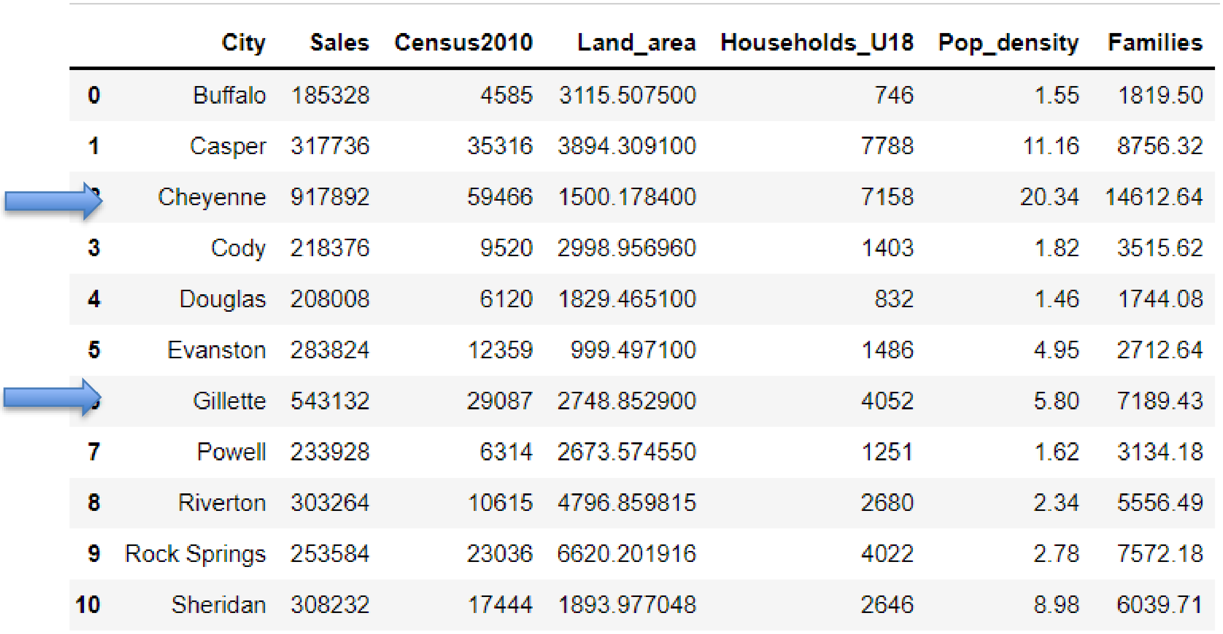

550000
500000
450000
400000
350000
300000
250000
200000
Sales
1000
2000
3000
4000
5000
6000
7000
8000
Households_U18


550000
500000
450000
400000
350000
300000
250000
200000
Sales
2
4
6
8
10
Pop_density





Let’s compare with the table below.



	City	Sales	Census2010	Land_area	Households_U18	Pop_density	Families
0	Buffalo	185328	4585	3115.507500	746	1.55	1819.50
1	Casper	317736	35316	3894.309100	7788	11.16	8756.32
2	Cheyenne	917892	59466	1500.178400	7158	20.34	14612.64
3	Cody	218376	9520	2998.956960	1403	1.82	3515.62
4	Douglas	208008	6120	1829.465100	832	1.46	1744.08
5	Evanston	283824	12359	999.497100	1486	4.95	2712.64
6	Gillette	543132	29087	2748.852900	4052	5.80	7189.43
7	Powell	233928	6314	2673.574550	1251	1.62	3134.18
8	Riverton	303264	10615	4796.859815	2680	2.34	5556.49
9	Rock Springs	253584	23036	6620.201916	4022	2.78	7572.18
10	Sheridan	308232	17444	1893.977048	2646	8.98	6039.71

There is an outlier based on the first figure of the boxplot, which is the Cheyenne city. We will remove Cheyenne from our data as it is too far from the normal data. There is also another outlier in the rest of the data plotted by using scatter plot and box plot, which is the Gillette city. Gillette would also be the outlier in this case when compared against all other cities due to its greatest distance from the linear trend. Since the relationships between Gillette's population related variables and total sales are still correlated, Gillette should be kept for prediction and analysis.