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?

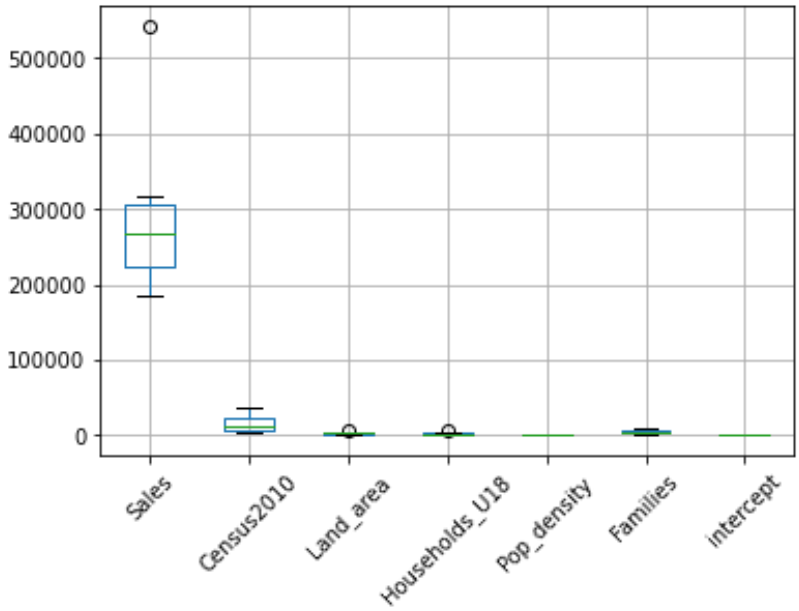
   City
   2010 census population
   Pawdacity sales in other stores
   competitor sales
   household with under 18
   land area
   population density
   total families

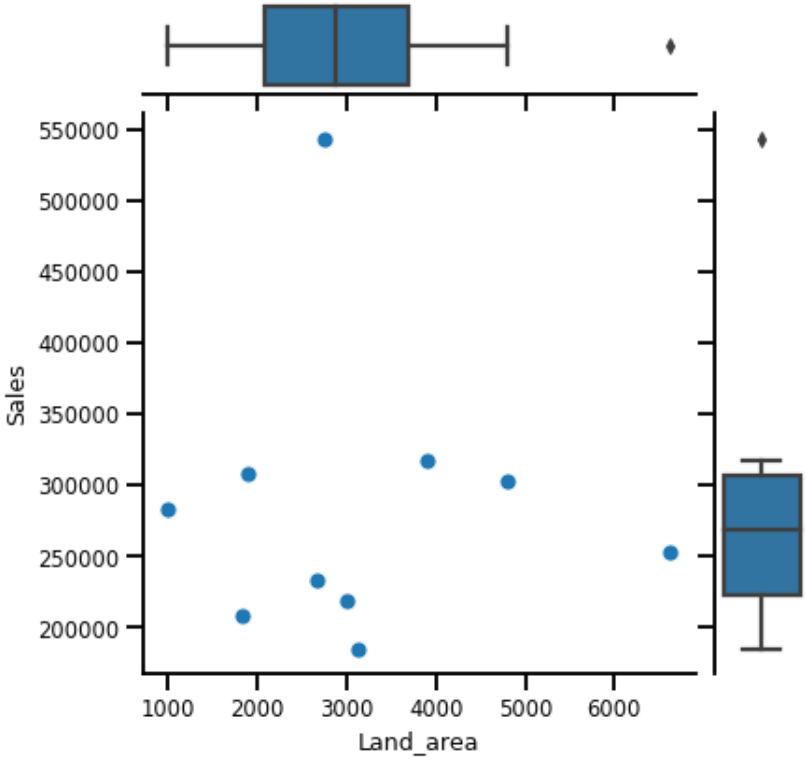
Step 2: Building the Training Set

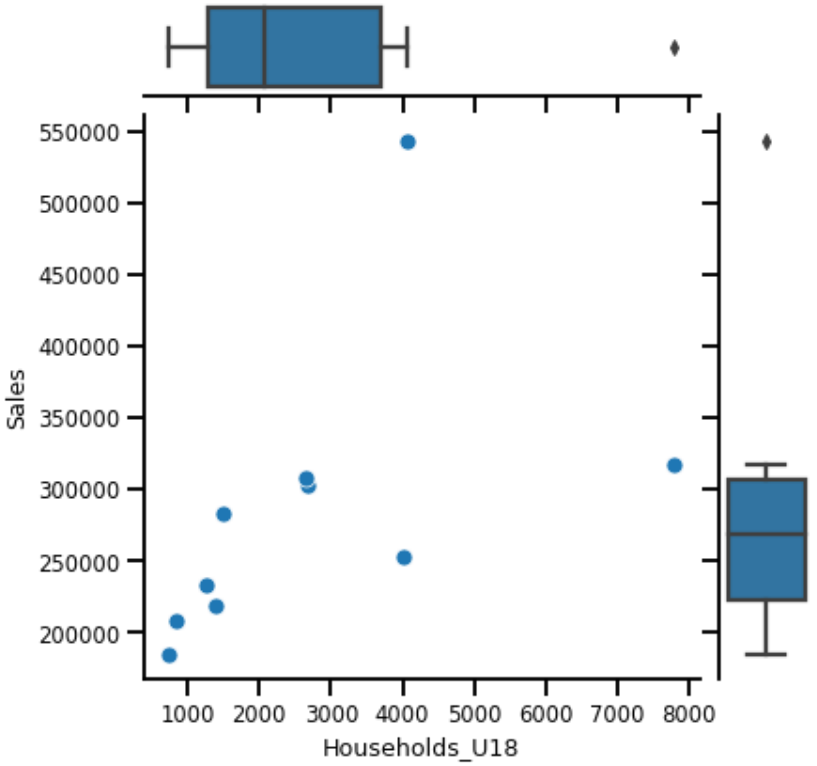
Column	Sum	Average
Census Population	*213,862*	*19442*
Total Pawdacity Sales	*3,773,304*	*343027.64*
Households with Under 18	*34,064*	*3096.73*
Land Area	*33,071*	*3006.49*
Population Density	*63*	*5.71*
Total Families	*62,653*	*5695.71*

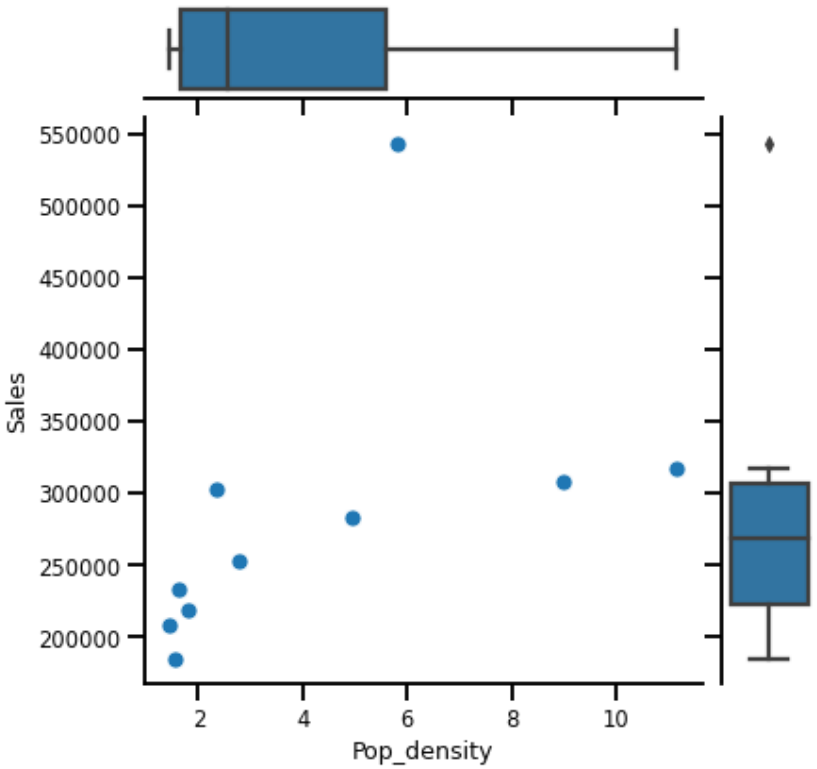
Step 3: Dealing with Outliers

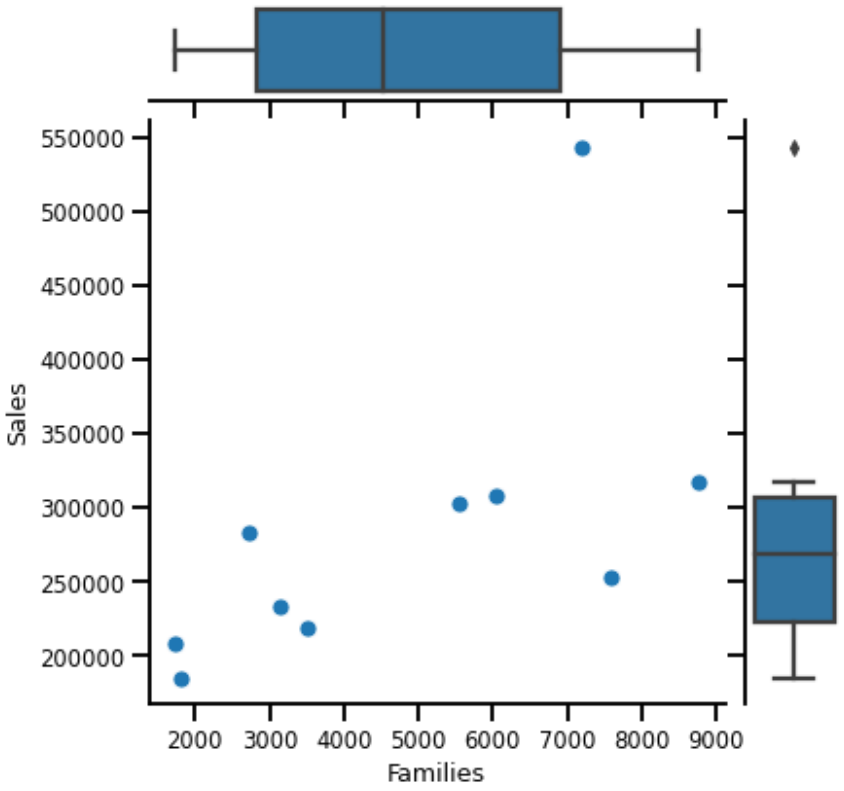
Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.

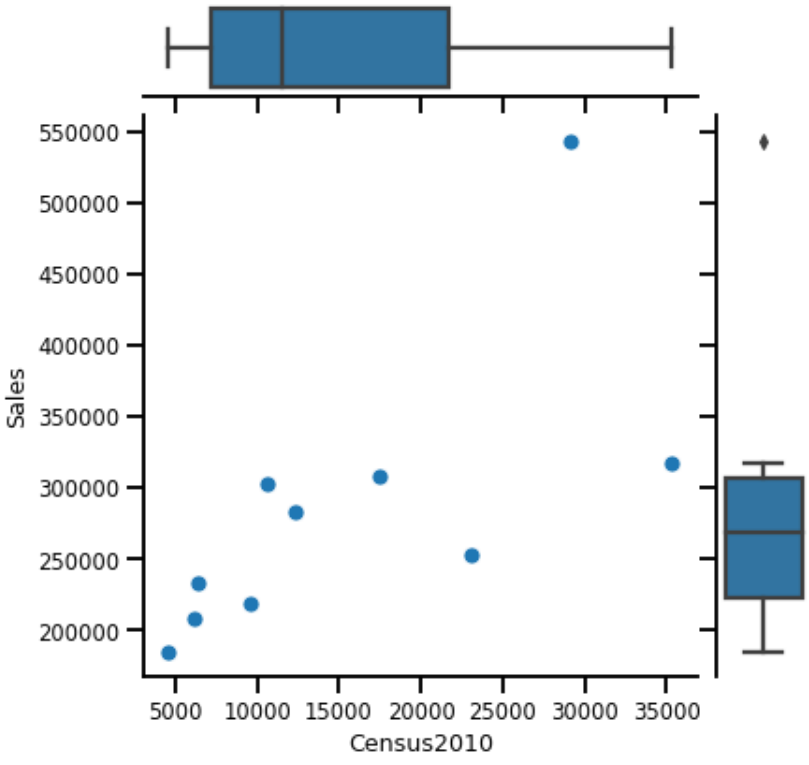





550000
500000
450000
400000
350000
300000
250000
200000
Sales
1000
2000
3000
4000
5000
6000
7000
8000
Households_U18


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


550000
500000
450000
400000
350000
300000
250000
200000
Sales
2000
3000
4000
5000
6000
7000
8000
9000
Families


550000
500000
450000
400000
350000
300000
250000
200000
Sales
5000
10000
15000
20000
25000
30000
35000
Census2010

There are 1 outlier in all the data plotted by using scatter plot and box plot, which is the Gillette city. Thus, Gillette would 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.