**CDS 2017 Mid-Sem Examination: Group 13 report**

**Problem Statement:**

Identify possible locations for a new store of RandMart or in other words, What is the impact of store location on the store performance where performance is measured in terms of profitability which is calculated as Revenue – Expenses?

Note: We are not using % profit because if a store is located in a scarcely populated neighbourhood it is possible that it could have high % profit but from a company’s perspective the small amount of profit brought in will not do them much good. We will be more interested in the absolute profit.

**Objective:**

Our approach to this problem was two-pronged:

* Create a model to predict expected revenue or profit for existing stores in an area (zip + city)
* Simulate the case of opening a new store in each area independently and use the model to estimate the revenue/profit for that store

The stores with the maximum profit are our recommendations for possible high impact locations. The store performance (expected profit) will depend on the following factors:

* Demographic/location specific factors (Current account of burglary in area, Total number of households in the area, Average intracity mass transit fares, Indicator of competing store in neighbourhood, Population Density, Per Capita Income etc.)
* Store specific factors (No. of leads pursued by store, Whether store has front road opening, No. of parking spaces available, Carpet area of store, Indicator if a store is a pad site in front of a mall, Cannibalism)
* Expenses (Marketing Expense, Salaries, etc.)

Our objective is to find which of these factors are most significant and which of these have maximum influence on expected revenue/profit and combining this knowledge we can predict possible locations for a new store.

**Data Description:**

The file received had 2823 observations of 56 variables. However after removing the observations with missing values for number of employees and other variables we were left with 2802 observations.

It is important to note that the expected profit for a RandMart store could be affected by the presence of other RandMart stores in the vicinity. This effect is known as cannibalism and to account for this we created variables such as total number of stores at zip, city, state, division and region level denoting the total no. of RandMart stores in that zone.

The other important calculation is that of the expected profit. It is calculated as difference between Revenues and Expenses.

Expenses are split into 2 categories

1. Marketing expense: Here we are considering marketing expense incurred for 2016
2. Total Salary Expense: There are 2 types of employees i.e employees paid on an hourly basis and those paid on salary basis. We are given the no. of employees of each type as well as the average pay per employee of each type. Average pay per S type employee is given yearly while those of H type is on an hourly basis. One assumption we made is that the employees paid on an hourly basis work for an average of 40 hours in a week. This amounts to nearly 2100 hours per year. After converting hourly pay to yearly pay and adding the salary expense for both type of employees we got the total salary expense.

Note:

As we don’t have information regarding the timing of attrition as well as the post at which attrition occurred (Type H or Type S) we have disregarded it from our analysis.

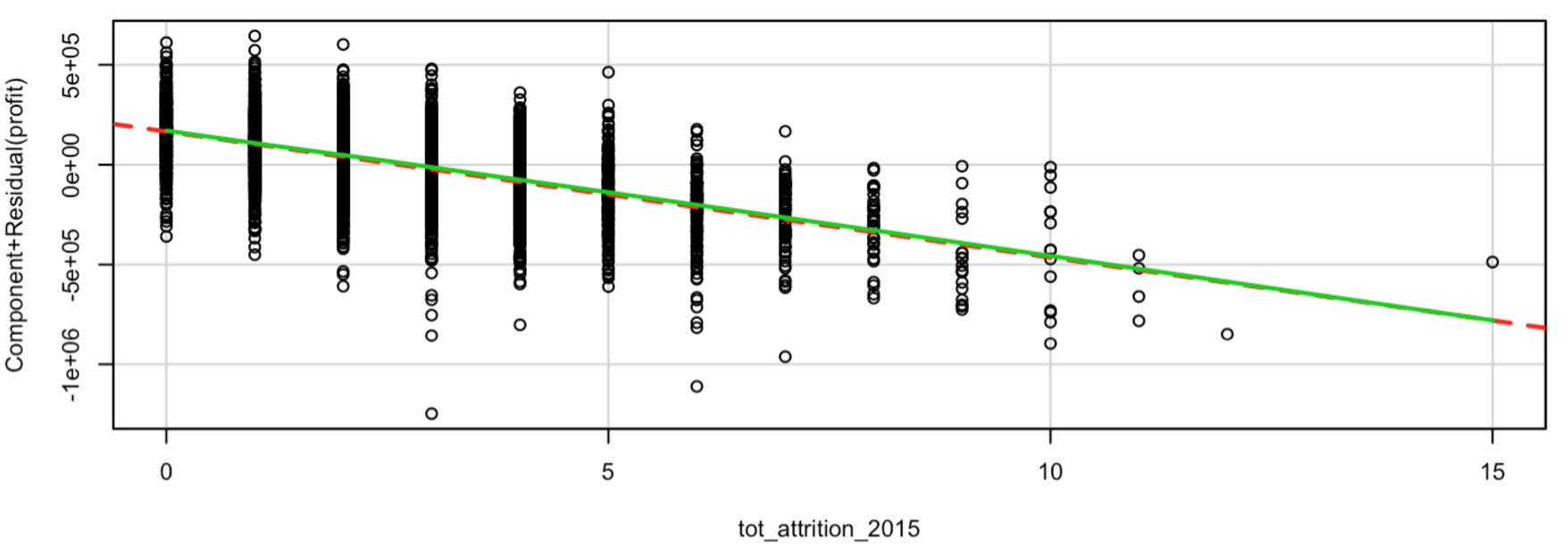
There were certain stores showing discrepancy in average pay per employee on salary basis i.e the values were 0 (must be missing), these were changed to reflect the median value of the average pay per employee on salary basis after removing the 0s i.e. 45,000

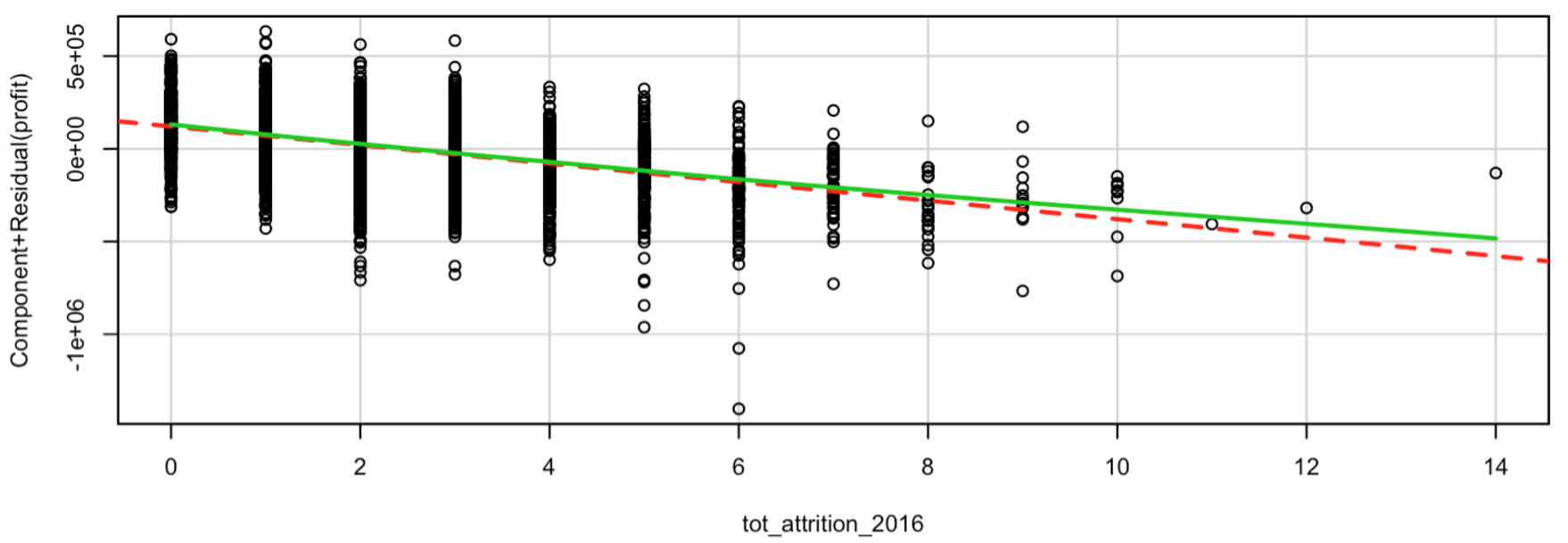
Exploratory Data Analysis:

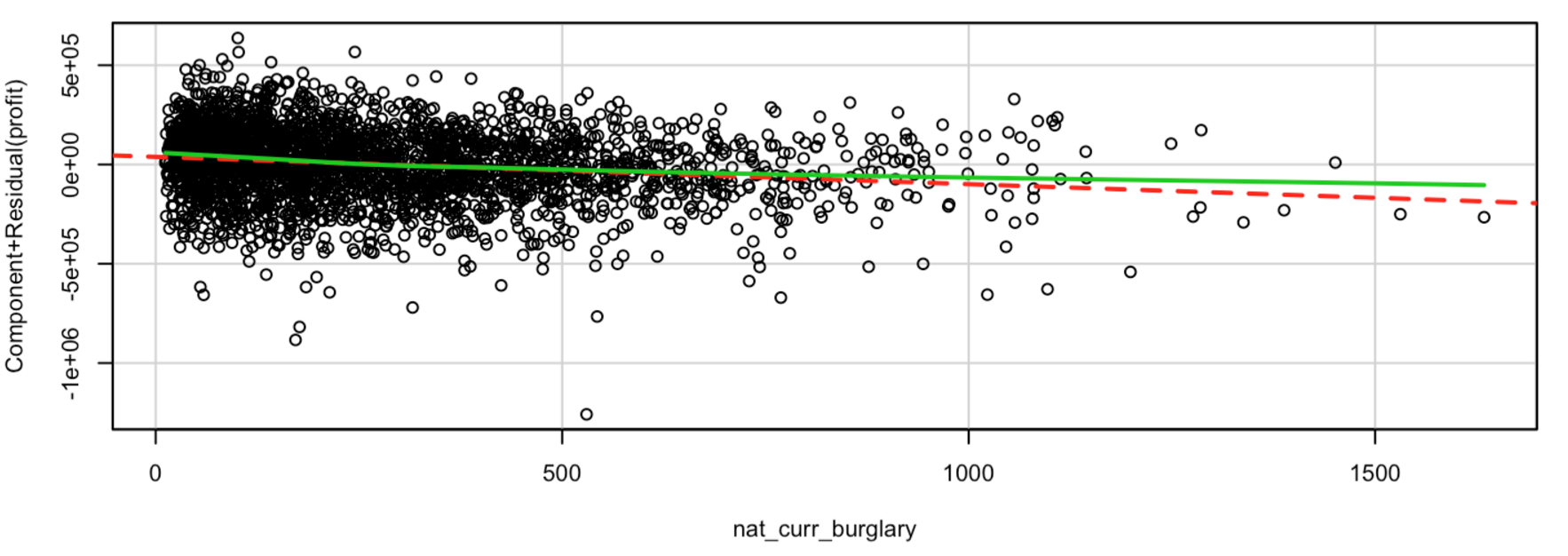
First we tried to identify the pairwise correlations between predictor variables using the *corrplot* function of R. It revealed the following correlations to be significant (>90%)

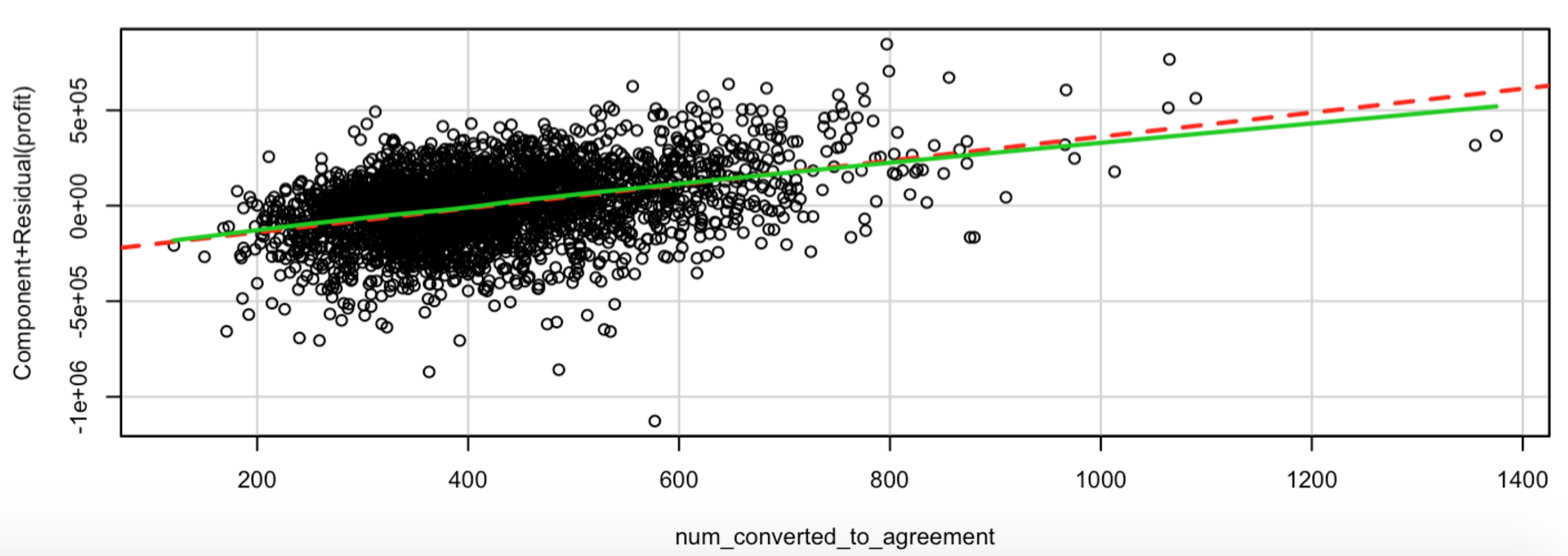
1. Current account of burglary in the area with past account of burglary in the area
2. Current account of robbery in the area with past account of robbery in the area
3. Current account of motor vehicle theft in the area with past account of motor vehicle theft in the area

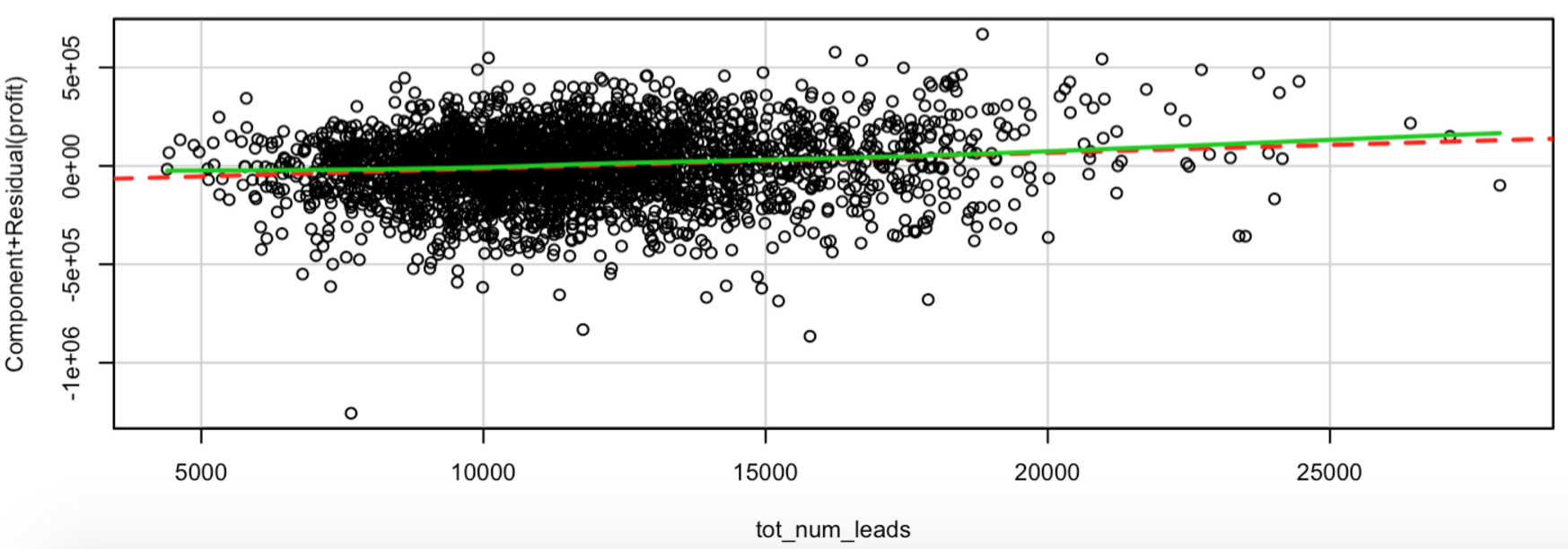
Attached below are some of the scatter plots between the response and different predictor variables. These indicate if there exists a linear relationship between the two or in case a transformation would be required for the predictor variables. Looking at the plots below it appears that the assumption that the relationship between the response and predictor variable will be linear is likely to hold true.







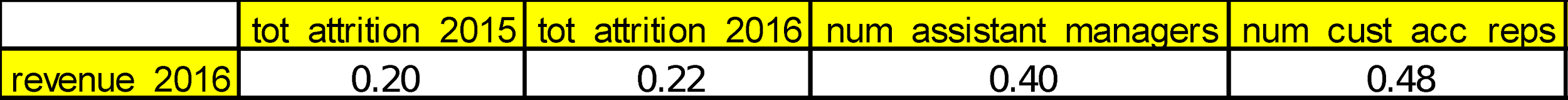




The South region and Texas (TX) state have the most no. of stores followed by Midwest region and New York (NY) state

|  |  |
| --- | --- |
|  |  |

Shown below is a snapshot of the correlations of revenue with different predictor variables. Number of employees of type H has the highest correlation with revenue.





**Model:**

As the problem instructs us to predict locations for a new store rather than the store specifications we could have only predicted the expected profit based on demographic/location specific explanatory variables however this would give an incomplete and incorrect picture and the estimates of the β parameters could be vastly different had we included the store-specific variables as well. So we have taken both demographic as well as store-specific explanatory variables.

The cleaned dataset was split into training (75%) and testing (25%) for model building purposes. Our initial approach was to calculate profit for existing stores using actual revenue and estimated expense (calculated as discussed above in Data Description), and model this with the other available explanatory variables. This model gave us very poor results. As estimated profit had a high correlation with revenue (> 0.5), we finally modeled revenue at a zip and city level. Below is the approach in detail:

* Model Revenue against explanatory variables at a zip and city level on the training dataset using repeated 10-fold cross-validation. We used **elastic-net linear regression** for the final model.
* Validate model on the testing dataset once to check generalisation of model
* Predict revenue on the full dataset
* Simulate the case for adding one new store for each zip and city, keeping location and census factors constant and incrementing store specific parameters i.e. number of employees, leads, etc. by the average in that area
* Predict new revenue using the model
* Increment in revenue is the revenue gained from opening a new store in that area
* Deduct estimated expense from revenue to get the estimated profit for a new store in each area
* Areas with new store showing maximum profit were provided as recommendations

Model Parameters chosen via grid search and repeated 10-fold cross-validation:

* alpha (mixing parameter) = 0.04
* lambda (regularisation constant) = 0.1

Validation metrics:

|  |  |  |
| --- | --- | --- |
| **Metric** | **Train** | **Test** |
| **R-squared** | 0.86 | 0.84 |
| **RMSE** | 125154.6 | 136066.2 |

Explanatory variables used to predict Revenue in final model:

|  |  |  |
| --- | --- | --- |
| **Variable name** | **Description** | **Coefficients** |
| Intercept | Intercept for the model | -7,74,015.62 |
| square\_feet | Predictor variable giving the total carpet area of the stores for given zip code and city | 14.60 |
| num\_cust\_acc\_reps | Predictor variable giving the total number of customer representatives at the stores for a given city and zip code | 6,528.75 |
| num\_store\_managers | Predictor variables giving the total number of store managers for a store for a given city and zip code | -5,335.34 |
| num\_emp\_pay\_type\_h | Predictor variable giving the total number of employees with hourly pay of a stores in a given city and zip code | 70,567.16 |
| nat\_curr\_robbery | Predictor variable giving the total current account of robbery in areas for a given zip code and city | 127.44 |
| nat\_curr\_burglary | Predictor variable giving the total current account of burglary in areas for a given zip code and city | -38.69 |
| avg\_pay\_rate\_pay\_type\_s\_mutated | Predictor variable giving the average pay per employee on salary basis of a given zip code and city (imputed with median) | 13.49 |
| avg\_pay\_rate\_pay\_type\_h | Predictor variable giving the average pay per employee on salary basis of a given zip code and city | 16,530.54 |
| num\_converted\_to\_agreement | Predictor variable giving the total number of leads converted to agreement of the stores for a given zip code and city | 811.06 |
| num\_stores\_in\_city | Predictor variable giving the total number of stores in a given city and zip code | -1,090.49 |
| population\_density | A derived predictor variable computed from the total area and the population of the area | 2.37 |
| OR | One-hot predictor variable indicating whether the given city and zip code belongs to state ‘OR’ | -1,89,354.49 |
| WA | One-hot predictor variable indicating whether the given city and zip code belongs to state ‘WA’ | -60,686.71 |
| other | One-hot predictor variable indicating whether the given city and zip code belongs to state other than ‘WA’ or ‘OR’ | 39,005.77 |
| census\_region\_northeast | One-hot predictor variable indicating whether the given city and zip code belongs to region ‘Northeast’ | 36,687.02 |
| census\_region\_midwest | One-hot predictor variable indicating whether the given city and zip code belongs to region ‘Midwest’ | -13,141.26 |

**Final Location Recommendation:**

State City Zip Expected Profit from new store  
CA DELANO 93215 720177.19  
CA MODESTO 95351 701599.94  
CA OROVILLE 95966 655511.02

VA COLLINSVILLE 24078 626512.22  
PA JOHNSTOWN 15904 619363.57  
CA HANFORD 93230 609204.35  
CA YUBA CITY 95993 581684.99  
MD BALTIMORE 21237 560399.40

CA REDDING 96002 530265.19

TN JACKSON 38301 528216.39