Report

Team #101

I. Data Visualization

The data provided contains different types of business settlements in different zip code regions with their sales/revenue/receipts in different ranges and the corresponding number of settlements to each revenue range. At first, we use a map to visualize the frequency of different states and zip code regions in the US. So higher the frequency means that there are more different types of establishments inside the region. We can see that the state California, Texas, and Florida has more settlements. It may because these are large states and they have more business. When view the frequency map of zip code, California, South side of Florida, and the north east coast of US has a higher amount of different types of establishments.

II. Market Potential

In order to determine a list of zip code that is prioritized, we analyzed the marketing potential of each regions. The market potential is under the assumption that people live in a zip code region will spend their money on the business in the same region. Therefore, if there is a lower settlement's revenue to resident's income rate, then there is a higher market potential. However, if the market potential is too low, then probably people live in that region does not have a great desire on consuming, so the best choices of business are regions with the middle 50% of market potential. Since the dataset does not provide all the information we need, we use outside resources of number of people in different income range in zip code to help determine market potential. We use the median of each range to multiply population in that range and sum together to get the total income of each zip code region, and perform similar approach on the business' sale to get total sale in each zip code region and then merge the two datasets together to calculate variable "marketPotential". The "marketPotential" is calculated by the "totalSales" divided by the "totalIncome". Then we output the list of zip code region that is the middle 50% of the log transformed "marketPotential". There are 41 states included in the middle 50% of marketPotential, which reveals that the opportunities of business are not concentrated in a specific state or region, instead, it is widespread around the US. The top three states that have the most zip code regions inside are Texas, New York, and California. The result of that actually matches our impression of there are lots of opportunities inside these states. Further analysis shows that the number of population and settlements and market potential are highly correlated. Which makes sense since more population will generate more income, and the same for the number of business. We also find that the sales will decrease with a high marketPotential, which suggest that the regions with high marketPotential is saturated.

III. Businesses Proposal

We also analyze the NAICS.display-label by bar plots. It seems like the different types of business are not distributed evenly. We then find that the top five most common business in the US are Retail trade, Gasoline stations, Gasoline stations with convenience stores, Food and beverage stores, and Grocery stores. When we try to find the top five most common business for the places with middle 50% market potential, the result changes a lot, as the top five becomes Miscellaneous store retailers, Supermarkets and other grocery (except convenience) stores, Grocery stores, nonstore retailers, and Building material and supplies dealers. We think that it may because some of the original common business are saturated, so the new types of business have higher market potential. So we will recommend investors to focus on the less saturated types of business, which we also output the ten types of business with the highest frequency in the middle 50% market potential regions.

IV. Important Attributes

We also want to use Random Forest to classify the zip code regions so business can find out which region is a better choice and which features have significantly influences. To build the Random Forest model, we add several variables that are relative to taxes from another outside dataframe. Since we do have lots of variables, we perform the ANOVA test first to remove the variables that are not significant to zip code at all to have a better performance on later model. After cleaning, there are 16 independent features. Because the dataset is too large to run on our computer, we randomly sample out 50% of data to build and test the model. We split the sampled subset into train, validation and test dataset, each contains 20%, 20%, and 60% of the sampled dataset respectively. The output of the random forest with a 3-folds cross validation shows an average of accuracy rate equals 61.31%, and perform the model on testing dataset, we have an accuracy rate of 34%, which are lower than our expectations, but it may be due to the fact that the zip codes are all unique, so it is hard for the model to classify data into categories that are not in the training dataset. The Random Forest model also produce a list of the significance of variables. From the result list, we can find out that the variables estNotEntireYear, totalSales, estEntireYear, AGI_STB (Size of adjusted gross income), A18450 (State and local general sales tax amount), and N18300 (Number of returns with taxes paid) are relatively more important to identify the zip code region. To be straight forward, income, sale, tax are important factors in our model.