CenterCheck Project Report

**Part A**

Methodology and Justification

1. Reduce number of columns in the raw financial dataset so each row represents a single retail location.
2. Identify and address the instances in which the same combinations of name, city, state, and address had contradictory postal codes.
3. Turn the following columns to lowercase, remove special characters, and trim whitespace: city, state, postal code, address, and name.
4. Make a new column “city\_state\_code” in both datasets that is a concatenation of the city, state, and postal code.
5. Make a new column “alt\_names” in the business locations dataset that lists all names referring to the same business entity id. Although this step never got leveraged, I included to demonstrate my consideration of edge cases.
6. Joined the raw financial dataset and business locations dataset on the new variable called “city\_state\_code” and address and saved any matches to a new data frame. My reasoning here is that rows with matching city, state, postal code, and address are the most pristine matches so we want to prioritize this.
7. Of the remaining unjoined data, join on city\_state\_code and name. The reason I considered this less pristine than the previous join was the case of a chain retailer having multiple locations within the same city\_state\_code. Its very plausible but did not occur in our data. My plan for this instance would have been to use an approximate matching algorithm on the address to determine which was the more correct match. The other instance that I was prepared for was no match at all due to the variation in name within a business entity id. My plan would have been to search through the “alt\_names” column I made for these cases. I believe this joining system accounts for most reasonable edge cases.
8. To ensure all rows were correctly matched, I used an approximate match algorithm on the two address columns. After a quick inspection I observed that even the lowest similarity scores seemed to be correct matches.

Assumptions

1. I assumed that all instances of postal code in the business locations dataset were correct when it differed from the postal code in the raw financial dataset.

**Part B**

Methodology and Justification

Estimating the percent of cash revenue in a city based on the population of unbanked, underbanked, and fully banked households:

* The location of each retailer was valuable information and my first thought was to investigate a metric explaining the affluency of the area. I decided to use unbanked and underbanked household rates because I thought it would explain the usage of cash better than other poverty statistics like income or welfare programs. There seem to be other cultural factors at play determining if homes will rely on the banking system and card transactions. A household being unbanked logically leads me to believe that they will be using cash as having a bank account is crucial for being able to pay with cards.
* For each city, I have the proportion of the population making up my three groups (unbanked, underbanked, and fully banked). I also need to know the proportion of all spend is cash for each respective group. I had to derive this using the percent of total transactions made across these three groups which is not the same since we know cash transactions are more likely to be smaller than card transactions. Using a scaler based on average size of cash transaction and average size of card transaction, as well as a scaler I would later solve for to scale the cash revenue rate back to my known statistic of 12% countrywide, I solved for the percent cash spend of each group. Although inexact, I believe this is a very reasonable approach to refine the statistics I found to better apply to the situation in question.
* I brought in data of the unbanked and underbanked rates in all cities above 10,000 population and joined based on city and state to my existing data. Any city that did not find a match inherited the statistics of the state it was in. 115,735 cities had a match while 33,136 took on the statistics of their state.
* Using these unbanked and underbanked percentages, I took whatever portion was left and used it as the fully banked population. I multiplied the proportion in each group by the proportion they spend in cash and added together to estimate what the total proportion of cash revenue would be for the given area.

Proportion of cash per area = percent unbanked\*.35 + percent underbanked\*.18 + fully banked\*.09

* Subtract the previous calculation from one along with .001 representing 1% customer financing to derive the percent of card used.
* Create a multiplier that is 1/(the percent of revenue that is card). This when multiplied by the raw revenue represents the estimated total revenue.

Using square footage of the retailer to adjust cash usage

* The type of retailer also must play a huge role in the cash usage but we unfortunately do not have this information. My thought was that the square footage of a retailer gives a lot of information about retailer type and more specifically the size of the purchase. I came across many statistics noting that the larger purchase sizes are much more likely to use card.
* After researching different typical sizes of stores. I bucketed the square footage into a new column representing retailer type.
* Like the challenge outlined in the unbanked predictor, I needed the percent of cash revenue for each of these retailer types but only had the percent of transactions. After being stuck at this step trying to find the right information, I decided to make some loose estimation as to what percent cash revenue each retailer had.
* The same methodology was used to derive a multiplier as previously stated, but this time I made a second column with an adjusted multiplier by dividing the original by (1/national average percent of revenue in card). Since both multipliers account for a portion of assumed cash, I needed to scale this one to not double count the standard amount of cash usage. Both multipliers were above 1.0 which did not make sense to use them both. After the division step however, some multipliers fall below 1.0 and some above which is how we expect our multiplier to work.
* Finally, I multiplied the unbanked multiplier by the new adjusted store type multiplier to get a combined multiplier which got multiplied to the raw revenue for our final adjusted revenue.

Assumptions

1. The first assumption I made was that the only payment types not included in our current revenue totals were cash and consumer financing based on the breakdown given in my code. More specifically, I assumed that all transactions classified as “Digital Wallet” by Capital One were already included.
2. I assumed that the 1% of payments that were consumer financing did not vary across any of the predictors I looked at. This is obviously an oversimplification and with a little more research, I could have used a lot of the same methodology to better estimate this category. My reasoning was like my first assumption in that I did not want to bite off more than I could chew and predict many things poorly instead of one thing well.
3. I assumed that all three groups (unbanked, underbanked, and fully banked) spend money at the same rate. This is probably very incorrect as I would infer unbanked and underbanked households have a much lower disposable income and spend less money.
4. I assumed that for cities that did not find a match (under 10,000 population) they had the same unbanked and underbanked rates as their state. This is a generalization and probably could have been better estimated with data about more rural areas.
5. I concretely assigned retailer types strictly based on the square footage. This is an overgeneralization and can be incorrect in many instances.
6. I assumed the percentages of cash revenue for each retailer type based on the percent of transactions in cash. I reasoned that the retailers with larger purchase sizes would be the most different since we can infer that the larger transactions are going to be in card. As a result, I kept the convenience store and fast-food categories relatively similar. There is an overall lack of methodology on this step