ISyE 6740 – Summer 2022
Final Report

Team Name: Matthew Kesselman

Project Title: Deep Dive into the Restaurant Revitalization Fund

Contents

Problem Statement:

In 2021, in response to the growing crisis of restaurants facing COVID-related lockdowns and closures, the Small Business Administration (SBA) of the US government offered restaurants "funding equal to their pandemic-related revenue loss up to $10 million per business and no more than $5 million per physical location" in a Restaurant Revitalization Fund (RRF). This money was not required to be paid back, if recipients used it for "eligible uses" by March 11, 2023. The stated intent of RRF was to prioritize specific small businesses, these businesses included those owned 51% by:

- Women
- Veterans
- Socially and economically disadvantaged

Following the release of funds, the SBA provided a high-level summary report of total money spent in total, by state, by grant size, and by restaurant type[1], but failed to disclose if they were successful in their intent to prioritize "small businesses" or those run by women, veterans, and/or the socially and economically disabled. 101,000 restaurants were granted money, but the SBA received critiques as 265,000 others were turned away, seemingly in inconsistent fashion, as allocated money ran out, and as the SBA was also threatened with lawsuits from applicants outside priority groups (mainly white, male business owners)[2].

The data accessibility and the stated intentions of the RRF offers an appealing opportunity to examine government spending. This report will use a data-driven, computational approach on the RRF dataset to understand how funds were allocated and examine if they were allocated in a manner that fits the originally stated intent. Did the SBA successfully prioritize small businesses run by women, veterans, and the socially and economically disadvantaged? In a larger sense, this analysis will approach the on-going problem of responsible and transparent government spending: Are American tax dollars consciously being used/allocated? This deep dive will provide insights on how execution truly aligns to stated purpose when it comes to government funding.

Data Source:

The primary dataset used in this analysis is the RRF's raw grant data made accessible by the SBA[3]. This data includes 100,650 rows of approved businesses with 28 fields associated with the grant, the business, and the applicant. This dataset is supplemented with additional 2020 demographic data of median household income by zip code provided by the Census/ American Community Survey[4].

Field Name	Description	Example
LoanNumber	SBA Loan Number	1040219109
ApprovalDate	Approval Date	6/17/21
BusinessName	Business Name	shasbad INC
BusinessAddress	Business Address	14635 Beechnut St Ste B suit B
BusinessCity	Business City	Houston
BusinessState	Business State	TX
BusinessZip	Business Zip	77083
GrantAmount	Grant Amount	15049

[1] https://www.sba.gov/sites/default/files/2021-07/RRF_Report-508.pdf
[2] https://www.nytimes.com/2021/07/01/business/restaurant-revitalization-fund-sba.html
[3] https://data.sba.gov/dataset/rrf-foia
[4] https://www.census.gov/programs-surveys/acs/data/data-tables.html

FranchiseName	Franchise Name	
RuralUrbanIndicator	Rural or Urban Indicator. R=Rural; U=Urban	U
HubzoneIndicator	Hubzone Indicator. Y=Hubzone	N
CD	Congressional District	TX-09
grant_purp_cons_outdoor_seating	Purpose of Grant. 'Y' if box was checked on application	Y
grant_purpose_covered_supplier	Purpose of Grant. 'Y' if box was checked on application	Y
grant_purpose_debt	Purpose of Grant. 'Y' if box was checked on application	Y
grant_purpose_food	Purpose of Grant. 'Y' if box was checked on application	Y
grant_purpose_maintenance_indoor	Purpose of Grant. 'Y' if box was checked on application	Y
grant_purpose_operations	Purpose of Grant. 'Y' if box was checked on application	Y
grant_purpose_payroll	Purpose of Grant. 'Y' if box was checked on application	Y
grant_purpose_rent	Purpose of Grant. 'Y' if box was checked on application	Y
grant_purpose_supplies	Purpose of Grant. 'Y' if box was checked on application	Y
grant_purpose_utility	Purpose of Grant. 'Y' if box was checked on application	Y
LegalOrganizationType	Organization Type	Subchapter S Corporation
LMIIndicator	LMI Indicator. Y=LMI	N
SocioeconmicIndicator	Socially and economically disadvantaged individual(s) Indicator. 'Y' if box was checked on application	N
VeteranIndicator	Veteran Indicator. 'Y' if box was checked on application	N
WomenOwnedIndicator	Women Owned Business Indicator. 'Y' if box was checked on application	N
RestaurantType	Restaurant Type. All restaurant types checked on the application are listed.	Restaurant

Methodology

Given that this investigation is an exploratory look into the nuances of the Restaurant Revitalization Fund payout, this investigation will use appropriate exploratory approaches to dissect and better understand the program. These include:

- An unsupervised approach to determine if there are clusters of restaurants granted funds (and to examine if these clusters align with the stated SBA goals).
- A dimensional reduction (MDS) upon the dataset to visualize its many features in a lower-dimensional space
- Summary statistics on both the algorithmically defined clusters of businesses and exceptional outliers.

Data Preparation – linking household income by zip code

An important promise by the SBA was that the funds would be used to prioritize economically disadvantaged individuals. Considering the attribute corresponding to this goal on the main dataset ('SocioeconmicIndicator') was self-determined by the applicants of the grants, a third-party proxy data for economic advantage was added into the analysis. This was median household income by zip code from the American Community Survey as of 2020 which would offer some insight on the economic state of the restaurant's neighborhood.

Within the income dataset, there were 28 outlier zipcodes with household incomes labeled as "250,000+." These were cleaned to 250000. Additionally, there was one zipcode with a household income labeled as "2,500-." This zipcode

was removed from the data set. Upon linking the household income by zip code data set 776 restaurants were unable to be matched, so these were removed, rendering the usable data set 99,874 restaurants with 28 fields.

Data Preparation – using Gower's distance to make use of the dataset's binary and categorical features

The dataset features an abundance of binary and categorical features that provide rich information (ex SocioeconmicIndicator, VeteranIndicator, RestaurantType). In order to properly utilize these features, we will want define a distance function that can properly establish similarity (or dissimilarity) between these types of features. Simply encoding categorical variables as '1', '2', '3'… is not ideal as the relationship between the numbers likely does not correspond to the Euclidean distance between these numbers (i.e. an algorithm may take a 1 to indicate a "Restaurant" is more similar to a 2, a "Caterer," than a 3, a "Snack and Nonalcoholic Beverage Bar." To resolve this problem, we will approach this problem using Gower's distance, a similarity measure defined by J.C Gower[5].

Gower distance measures the dissimilarity of items (rows) with mixed numeric and non-numeric data. A value of 1 between two data points means terms have no similarity, while 0 means terms have absolute similarity.[6]

$$S_{\text{Gower}}(x_i, x_j) = \frac{\sum_{k=1}^{p} s_{ijk}\,\delta_{ijk}}{\sum_{k=1}^{p} \delta_{ijk}}$$

The above represents the dissimilarity score between two data points (x_i, x_j are data points composed of features k = 1… p). For each feature, a score is calculated (s_{ijk}). A quantity (δ_{ijk}) is also calculated with the values of 0, 1 depending on if features can be compared, or if they have different types. The overall score is the summation of the similarity of comparable features over the total number of comparable features.

The gower score for categorical variables is calculated by: $s_{ijk} = 1 x_{ik} = x_{jk}$ where the score is 1 if the categories of feature k are the same and 0 if they are not. (The distance/dissimilar metric is calculated by: $d_{\text{Gower}} = \sqrt{1 - S_{\text{Gower}}}$ which is why a score of 0 means absolute similarity). The gower score for numeric variables is calculated by $s_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{R_k}$. This is simply the Manhattan distance between two values normalized by R_k which is the range of the feature.

Data Preparation – filtering to useful features and down sampling the data

The size of the data set proved computationally difficult for Gower's algorithm (creating the dissimilarity matrix is $O(N^2)$ complex as every data point is compared to every other data point), so most trials down sampled to 10,000 restaurant samples to analyze. The dataset was then filtered down to key features of focus: indicators for rural/urban status, veteran status, socioeconomic status (self-reported), women-owned status, HH income (provided by the zip code data base), and most importantly, the grant amount.

Data Analysis – DBSCAN, MDS, Summary Statistics

To explore the data, an unsupervised clustering approach was needed. Gower's distance is a non-Euclidean measure, so the K-Means algorithm cannot be used. Instead, the DBSCAN algorithm will be used. DBSCAN is a density-based approach for discovering clusters proposed by Martin Ester, Hans-Peter Kriegel, Jiirg Sander, Xiaowei Xu in 1996[7].

---

[5] https://www.jstor.org/stable/2528823?origin=crossref; https://towardsdatascience.com/clustering-on-numerical-and-categorical-features-6e0ebcf1cbad

[6] https://ruivieira.dev/gower-distance.html

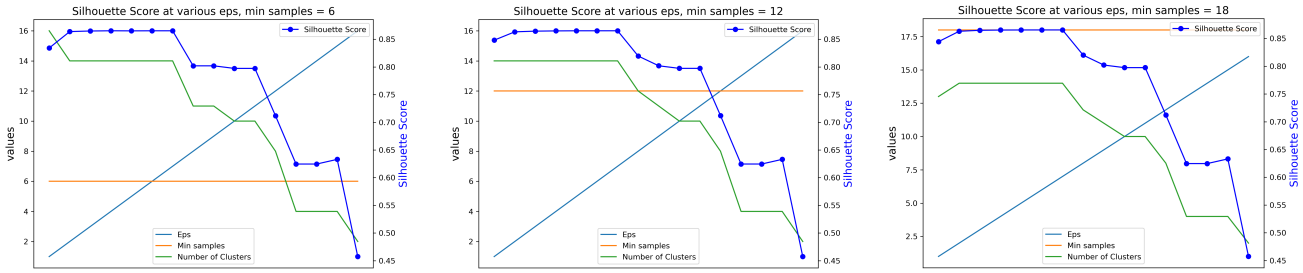
[7] https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf

DBSCAN is a more generic approach than K-Means which assumes clusters are convex. DBSCAN, on the other hand, can find clusters of any shape (including using this Gower dissimilarity geometry)[8]. DBSCAN does this by viewing areas of high density as clusters separated by areas of low density. Areas of low density, where there are insufficient numbers of near neighbors (based on a parameter definition) are considered "noise" and not grouped into a cluster.

DBSCAN requires two parameters: ε (eps), which is the maximum distance that is used to determine two points are "neighbors" of each other, and minPts which defines the minimum amount of neighbor data points needed to create a cluster. With 10,000 businesses (as the down sample) and 6 features, these parameters were determined by balancing the silhouette score of the clusters with the number of clusters. The silhouette score is calculated by the mean intra-cluster distance and the mean nearest-cluster distance that the sample is not a part of for each sample.[9]

$$(S = \frac{(b - a)}{max(a, b)})$$

Where *a* represents the mean distance between the observation and all other data points in the same cluster (mean intra-cluster distance) and *b* represents the mean distance from the observation and all other data points of the nearest cluster. A silhouette score ranges from -1 to 1 where 1 represents clusters that are dense and well-separated from other clusters, 0 which represents overlapping clusters, and -1 which represents values are likely assigned to wrong clusters[10].



These charts show silhouette scores at various epsilons with a fixed number of minPts (6, 12, 18). The aim was to maximize the silhouette score, while keeping the number of clusters manageable and interpretable. MinPts were fixed along these values (6, 12, 18) as multiples of the number of dimensions (6) that the data had[11]. After fitting for a cluster count less than 10, there was a significant drop off in the silhouette scores for all min samples, thus, an ε and minPts that provides the highest silhouette score with a cluster count of ten was chosen. This was ε =11 and minPts = 12, which yields a silhouette score of ~.797, 10 clusters, and 6 noise points that fit into no cluster.

To reduce the dimensionality of the data, a generic MDS (multi-dimensional scaling) was conducted on the dissimilarity matrix rather than a PCA analysis. This is because PCA analysis assume an Euclidean space where as for a more generic MDS implementation a precomputed dissimilarity matrix can be used. From the MDS analysis, the first two dimensions were taken to plot the data set in two dimensions. The graphical representation uses the labels determined by the DBSCAN cluster analysis. Summary statistics were performed on each cluster, looking for the size of the clusters (number of restaurants grouped in them), and the means, mins, and maxes of their various features.

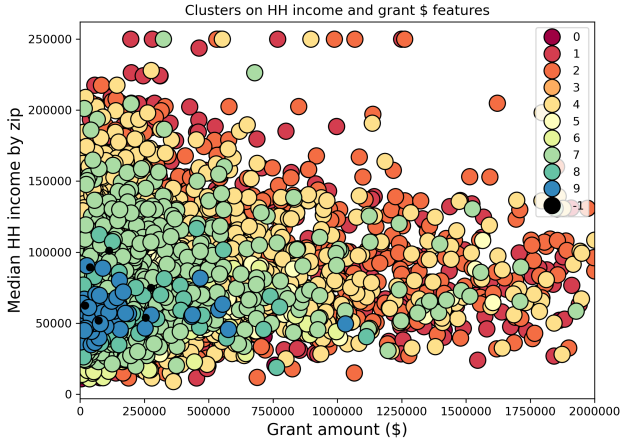
[8] https://scikit-learn.org/stable/modules/clustering.html#dbscan
[9] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html
[10] https://dzone.com/articles/kmeans-silhouette-score-explained-with-python-exam
[11] (Sander et al., 1998)

Results

Figure 1: DBSCAN Clustering with HH income and Grant Features

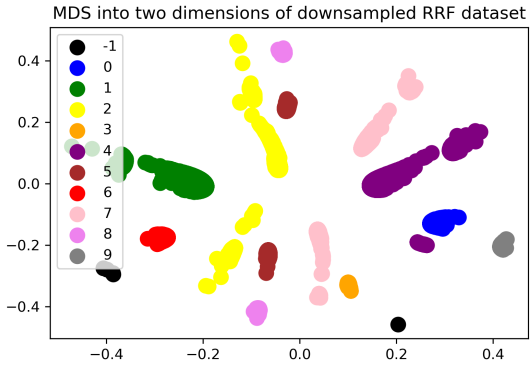
This figure shows the results of the DBSCAN clustering with ε =11 and minPts = 12 highlighting the two continuous factors of the six factors within the analysis (zip code Median HH income, serving as a proxy for income, and Grant amount). Clusters 1, 2, and 3, appear to be receiving a significantly higher grant amounts, however, at this zoomed out lens, the relationship between HH income and grant amount is less clear.

Figure 2: DBSCAN Clustering with HH income and Grant Features Zoomed in



This zoomed-in look of the DBSCAN clustering shows that cluster 9 appears to trend at the lowest grant amount and HH income, and cluster 8 to a lesser extent. Generally, there are tiers of HH income (1/2, 4/5, 6/7, 8/9). It is key to understand the nature, and size, of these different buckets.

Figure 3: MDS 2-dimensional representation of down sampled dataset



This is the 2-D MDS projection of the sampled dataset. The colored labeling corresponds to the DBSCAN cluster labels. Based on the projection, clusters 2 and 4 are similar to each other, clusters 7 and 4 are similar, 0 and 9 are similar. Clusters 0, 1, 3, 6, and 8 appear more unique.

These graphical representations of the data offer clues to the composition of the program. Examples include: cluster 2 and 4 are similar, but cluster 2 was given more money, cluster 6 and 9 were given similar-sized grants, but look to have significant differences (likely demographics). These intuitions can be compared against the average summary statistics found in each cluster.

Table 1: Summary statistics of different clusters

Cluster	Count of businesses	Mean Grant Amount	Min/max Grant Amount	Mean HH Income	% Urban (vs Rural)	% Veteran (vs not)	% Socio economic (vs not)	% women owned (vs not)
0	650	$131,000	$1,000 / $2,243,000	$58,000	0%	0%	0%	100%
1	2310	$287,000	$1,000 / $10,000,000	$80,000	100%	2%	100%	0%
2	2662	$377,000	$1,000 / $10,000,000	$81,000	100%	9%	0%	0%
3	75	$156,000	$1,000 / $1,298,000	$53,000	0%	0%	100%	100%
4	2670	$253,000	$1,000 / $6,720,000	$79,000	100%	4%	0%	100%
5	412	$203,000	$2,000 / $3,284,000	$62,000	0%	0%	0%	0%
6	236	$161,000	$3,000 / $2,278,000	$55,000	0%	0%	100%	0%
7	856	$227,000	$1,000 / $4,478,000	$77,000	100%	3%	100%	100%
8	66	$190,000	$2,000 / $1,113,000	$57,000	0%	100%	0%	0%
9	57	$142,000	$2,000 / $1,031,000	$57,000	0%	100%	0%	100%
Noise (-1)	6	$129,000	$20,000 / $276,000	$72,000	0%	100%	100%	17%
Full Population	99,874	$282,000	$1,000 / $10,000,000	$76,000	85%	6%	34%	44%

Clusters 1, 2, and 4 compose 76.4% of the sample, so understanding them is core to understanding the RRF population. Cluster 1 is self-indicated to be economically disadvantaged, however, the zip codes have the second highest Mean HH income of $80,000. Although the owners of the businesses could live in a different zip code and/or have a lower

income than the average HH in their area, it does bring into question what methodology/vetting when into the various statuses of the study. They also received the second highest average grant amount.

Cluster 4 is women owned, has the third highest mean HH income and is not self-indicated to be economically disadvantaged. It received the third highest grant amount. It has some veterans (4%, slightly less than the full population).

Interestingly, cluster 2 with the highest mean grant amount ($377,000) is nearly the largest cluster (only 8 restaurants behind the cluster 4) at 26.6% of the 10,000 restaurant sample, and has no veterans, no women, and no socioeconomically disadvantaged. Cluster 4 had no self-indicated socioeconomically disadvantaged but is women owned. They are both urban. This means that over 50% of the sample do not meet 2 of the 3 prioritized criteria. Cluster 2 also contains the highest max grant payout of $10,000,000. Relative to the full population, it can be clearly seen this cluster 2 only meets the stated goal of supporting veterans % of veterans relative to the full population (9% vs 6%), but it is important to note that veteran count still is quite small even in the full population (as reference, 7% of US population over 18 are veterans).

The remaining clusters received smaller grant amounts and are represented with lower HH income areas. Clusters 6, 7, and 3 together represented 11.7% of the sample and all were self-indicated socioeconomically disadvantaged (although cluster 7 still had a higher-than-average HH income for its zip code). By comparing to clusters given more grant money, a key question is introduced: what was the composition of 265,000 businesses turned away? Given that cluster 2, with no self-indicated economically disadvantage and no women, received the highest mean grant amounts, why were some of their funds not allocated the businesses that were turned away? Did those businesses not meet either the government's stated criteria or other criteria either obscured in the dataset and/or government methodology There is some suggestion, based on mean HH income, of a "richer get richer" effect or "too big to fail" with clusters 1 and 2 receiving the highest mean grants.

Discussion and future work

The results of this study cannot conclusively say whether the SBA is meeting their goals, or if the manner of fund distribution was consciously and equitably allocated. However, that is not a matter that can be concluded—it can only be *discussed*--with opinions deciding what arrangement is equitable and/or sensible. What can be concluded is that a significant portion of the allocated funds did not align with the SBA's stated priorities, but perhaps this composition of fund allocation had "sufficient" prioritization of marginalized groups to achieve an equitable balance with non-marginalized groups. To reach a more definitive conclusion, the matter would need to be discussed in the public forum, aided by the insights and analyses found in this study.

In terms of methodology, there are some areas where this study's specific analysis could be refined and modified:

1. Less down sampling when using a clustering algorithm. The RRF data set was robust with over one hundred thousand businesses to explore, however exploring such a robust dataset led to computational difficulties. With more computing power and/or time, this hurdle could be overcome and more of the data could be used to find more refined insights.
2. Experimentation with the number of clusters. Although ten clusters offered a clear "break off" point for the DBSCAN's silhouette score, this is because ten clusters meant the categorical variables could neatly be sorted and bucketed. Though this configuration offers an overall look into how the RRF was broken up by major category, by forcing fewer clusters, one could see where there was more overlap between demographic (women, socioeconomic, veteran) segments. Potential alternative parameters could be eps=14, minPts=12, which yielded 4 clusters and a silhouette score near .64 in the DBSCAN analysis.
3. Gower's distance is one methodology to turn categorical variables a geometry that is more readily analyzable, but it is not the only way. Or, as an alternative around the need to find alternative geometry than Euclidian, the self-reported categorical variables could be removed entirely from the analysis, and the focus could shift toward

additional datasets linked on zip code/city/state to provide more neutral third-party continuous features such as demographic variables.

On a larger scale (outside of the analysis conducted in this study), the RRF dataset has more to offer in other potential avenues of exploration. For instance, the focus of this approach was on the identity and status of the recipient of the grant (and the size of their grant), however, the data set also includes self-reported information on the nature of the grant's use (ex: 'grant_purpose_rent', 'grant_purpose_payroll', 'grant_purpose_utility'). Exploration of this portion of the dataset could offer insight into how the government perceived different businesses, if one need was deemed more "vital" than the other, or if some combination of nature of need and social/gender/veteran status saw an exceptional rise in funding.

Lastly, major outliers need to be explored and discussed. 64 businesses received grants of $10 million while there were over two hundred thousand other businesses turned away. The mean household income of these restaurant's/corporation's zip codes were $91,043, higher than the mean household income of any the DBSCAN analysis's clusters. Of the 64 businesses, 9 were owned by women, 7 were owned by veterans, and/or 12 marked for low "socio economic" status. By contrast, 34% of all businesses were marked with low "socio economic" status, 6% were marked with owned by veterans, and 44% of all businesses were marked with owned by women. Why did these specific businesses—DOJ DONUTS LLC, Piazza Romana LLC, TheBrickLLC as some examples—receive such outstanding grants? Did they save net more jobs than the other potential restaurants that were denied? What were the criteria of the decision? At minimum, it can be definitively concluded that these outlier examples did not prioritize the stated goal of supporting women and the socially and economically disadvantaged relative to the overall population by a significant degree (14% vs 44% for women, 19% vs 34% for socioeconomically disadvantaged), however these outliers are overrepresented with veterans (11% vs 6%).