CPS Food Security – A Case Problem

Corey Waldner
School of Technology and Engineering, National University
Course code: TIM 8501
Dr. Holbert
January 10, 2023

CPS Food Security – A Case Problem

For the signature assignment of this course, we were assigned to perform an Exploratory Data Analysis (EDA) on a chosen dataset. The dataset that was chosen for analysis was the 2019 Current Population Survey (CPS) Food Security dataset, a well-known government initiative to gather official statistics on employment and unemployment in the United States. The CPS involves interviews with around 56,000 strategically selected households each month, representing the nation, individual states, and specific areas. The survey allows for reliable comparisons from month to month and year to year with minimal inconvenience to the participating households. Additionally, the survey collects demographic information critical for updating data obtained once a decade through the decennial census. The CPS provides detailed and supplementary data that serves diverse needs, providing valuable insights for government policymakers, legislators, and other users of labor market information to understand the nation's economic dynamics.

This particular research assignment involves the analysis of the CPS dataset and the selection and justification of our chosen EDA methods. The outcomes of several measures, such as Measures of Variability and Central Tendency, Frequency, Variance, and Standard Deviation, Outlier Detection and Distribution Modality, Univariate Analysis Methods, Cluster Analysis, and Data Grouping, are presented in the paper and Appendix. We also provide our observations and results. Finally, we discuss the statistical discoveries we have made and suggest subsequent actions that should be taken based on our analysis.

EDA Process

I accessed the CPS dataset and examined its structure to initiate the exploratory data analysis process. The column headers were initially unclear, and the values were mainly small numbers in the range of 1-10. I opted to use Python to analyze the dataset and utilized the `print(df.info())` and `print(df.head())` commands [A] to conduct my preliminary data exploration. According to the results, the dataset had 510 column headers with 138,964 rows of data, providing a total of 70,871,640 unique data points for exploratory analysis. However, due to the dataset's vast size and my computer's limited processing power, I chose to work with a smaller sample size, as analyzing every data point was not feasible.

After reviewing the technical documentation, I identified eleven columns that could potentially have correlated data and require further research. These columns include PRPERTYP, HETENURE, HEHOUSUT, HEFAMINC, HRNUMHOU, HRHTYPE, HUBUS, GEREG, PEHRFTPT, PRMJOCGR, and HESS3, and their meanings and possible values are available in the Appendix section B. The selected columns vary from person type, living situation, income, region, and occupation, and each column's values have distinct meanings, which need to be incorporated into the code. However, there were consistent values for each column: -1, indicating a "blank" entry, -2, indicating a "don't know" entry, and -3, indicating a "refused" entry.

Measures of Variability and Central Tendency

Measures of central tendency include mean, median, and mode, while measures of variability include standard deviation (or variance), the minimum and maximum values of the variables, kurtosis, and skewness (Mukhiya & Ahmed, 2020). To calculate this, I utilized Python's built-in Pandas module[C]. The resulting table is displayed below.

	Mean	Median	Mode	Minimum	Maximum	Range	IQR	Variance	Std Deviation	Kurtosis	Skewness
PRPERTYP	1.399319	2.0	2	-1	3	4	1.00	1.154137	1.074308	0.851649	-1.561872
HETENURE	1.027338	1.0	1	-1	3	4	1.00	0.782418	0.884544	0.996736	-1.033985
HEHOUSUT	1.252958	1.0	1	1	12	11	0.00	1.238790	1.113010	1.050553	5.114942
HEFAMINC	9.862677	12.0	-1	-1	16	17	8.00	3.524049	5.789996	0.666774	-0.836572
HRNUMHOU	2.813074	3.0	2	0	14	14	2.00	3.423706	1.850326	0.951813	0.675016
HRHTYPE	2.205190	1.0	1	0	10	10	3.00	4.895355	2.212545	0.224297	1.069760
HUBUS	1.418742	2.0	2	-3	2	5	1.00	1.219918	1.104499	1.367417	-1.700428
GEREG	2.743243	3.0	3	1	4	3	2.00	1.055093	1.027177	0.975519	-0.387483
PEHRFTPT	0.931493	-1.0	-1	-1	3	4	0.00	0.155551	0.394399	3.782530	5.839296
PRMJOCGR	0.508635	-1.0	-1	-1	7	8	2.25	4.326595	2.080047	0.578229	1.255107
HESS3	0.114411	-1.0	-1	-9	3	12	0.00	2.648964	1.627564	0.430771	1.090072

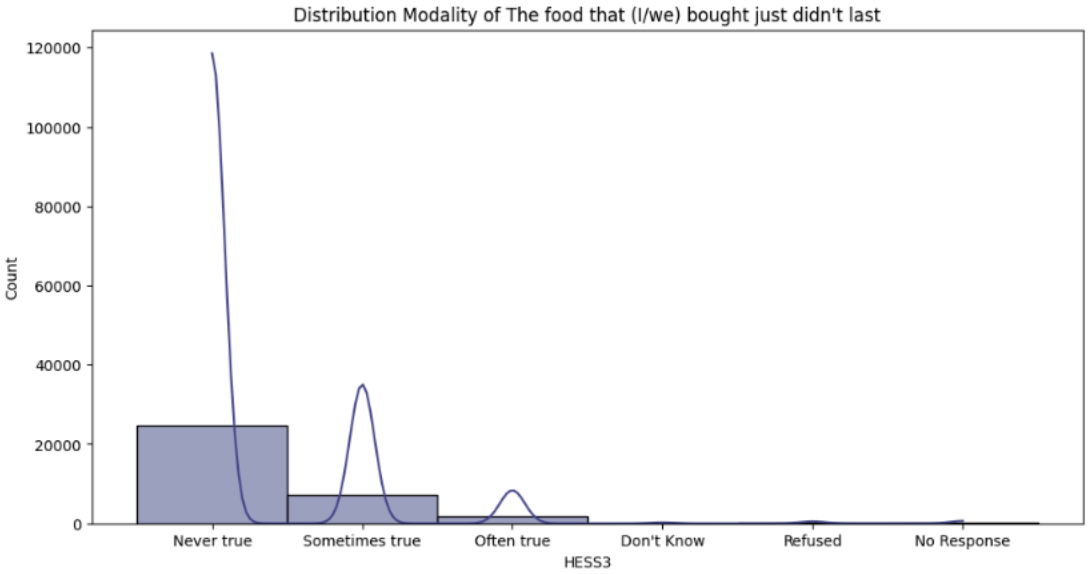
Frequency, Variance, and Standard Deviation

The variance measures the deviation from the mean. It is the average value of the squared difference between observed values and the mean. Standard deviation is the square root of the variance. Its unit is the same as for the original observations. This makes it easier for an analyst to evaluate the exact deviation from the mean. The lower value of standard deviation represents the lesser distance of observations from the mean; this means observations are less widely spread. The higher value of standard deviation represents a large distance of observations from the mean—that is, observations are widely spread (Avinash Navlani, 2019). As previously stated above, Python Pandas was used to compute the frequency, variance, and standard deviation. The sample output is demonstrated on the right-hand side of the display, while the remaining outputs for the column can be located in section D of the Appendix of this paper.

```
PRPERTYP Analysis:
------------------------------

Frequency for PRPERTYP:
PRPERTYP
 2    96697
-1    21032
 1    20806
 3      429
Name: count, dtype: int64
Variance: 1.1541372566861325
Standard Deviation: 1.0743078035116995
```

Outlier Detection and Distribution Modality

Outlier detection aims to discover objects that behave very differently from most objects. Outlier detection plays an important role in many aspects. In public safety, it helps to detect fraudulent acts; in medical diagnosis, it can predict diseases of concern; in network monitoring, it can be applied to data monitoring in wireless sensors; in industrial production, it can be used to identify defective products. Hawkins gave a generally accepted definition, "an is a deviation from other observations that leads to suspicion that it is produced by a different mechanism" (Zhengwei, Yangb, & Li, 2023). The concept of distribution modality pertains to the quantity of peaks or modes found in a probability distribution. It characterizes whether a distribution displays a single peak (unimodal), two peaks (bimodal), three peaks (trimodal), or more than three peaks (multimodal). As an example, this visualization showcases a trimodal distribution,



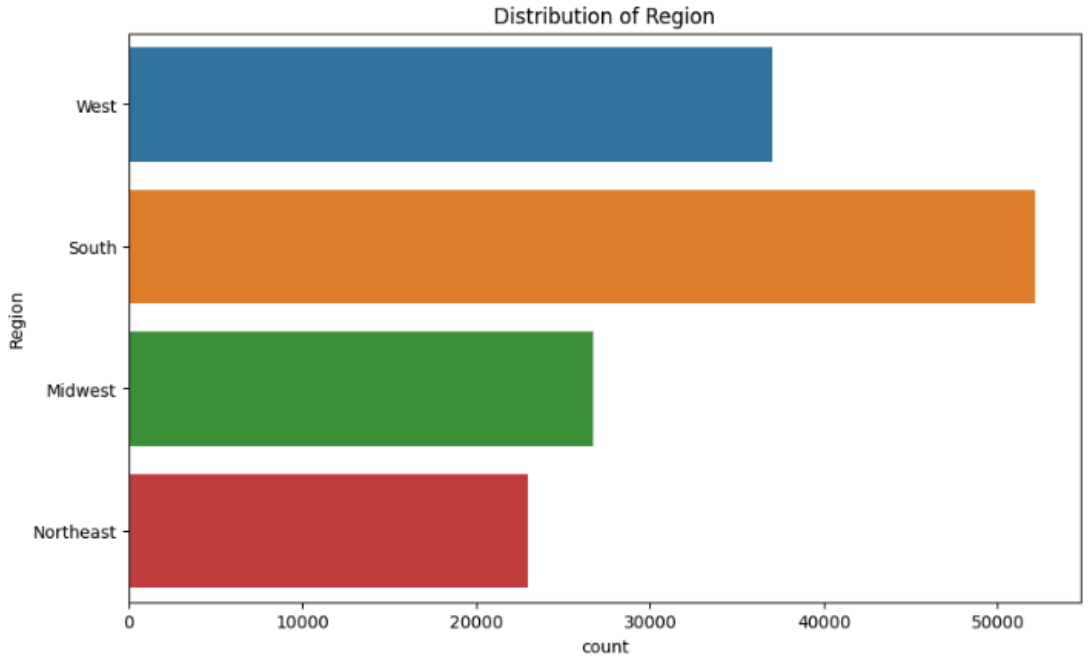
while the rest of the modality distributions are available in Appendix Section E.

As per the chosen dataset and columns, the criteria mentioned in the code did not detect any outliers. It is important to note that outliers are typically observed in numerical or continuous

data where certain values deviate significantly from the rest of the dataset. However, since categorical data is composed of distinct values, outliers are absent in this dataset.

Univariate Analysis Methods

The data set provided comprises the outcomes of a survey featuring 11 categorical data columns, which can be distinguished into two types - nominal and ordinal. Histograms serve as the most efficient way to represent the distribution and interrelationships among these categories. For additional visualizations, please refer to Appendix Section F of the document.



The above sections are categorized as Univariate Analysis, which refers to a statistical method utilized to examine and describe a single variable's distribution, central tendency, and variability. It involves analyzing one variable at a time, and various techniques can be employed to gain insights into the variable's properties. In this dataset, Descriptive Statistics were utilized, including Measures of Central Tendency (Mean, Median, and Mode) and Measures of

Variability (Range, Variance, and Standard Deviation). Additionally, Frequency Distribution was utilized, and histograms were chosen as the primary visualization due to the nature of the dataset.

Cluster Analysis and Data Grouping

Cluster analysis and data grouping are forms of multivariate methods that help uncover patterns and relationships within data. Cluster analysis examines and evaluates information focused on combining the information into groups based on how similar the individual facts are to one another. The groups, or clusters, of related items that are formed are helpful to the analyst in understanding the information. Cluster analysis can also be used to help summarize large amounts of information as part of another purpose, such as categorizing information to find the most relevant facts (Ungvarsky, 2023). The categorization of data based on multiple variables is commonly known as data grouping. This term encompasses various methods such as clustering, classification, and factor analysis.

The dataset was grouped based on Region[G] and analyzed for correlation between selected columns. The dataset was clustered using the k-modes[H] clustering technique, which is suitable for categorical data. K-means is a well-known and efficient clustering algorithm. It requires numerical data on a ratio scale, while in the real world, many times, categorical and mixed categorical and numerical data are present that need to be analyzed. Since the K-means algorithm always finds a local minimum and is efficient for large data sets, it is advisable to inherit its beneficial properties that are carried out by the K-modes algorithm. The algorithm uses the same cost function as the K-means, but a dissimilarity measure is applied instead of Euclidean distance, and modes instead of means are calculated on the resulted clusters in each

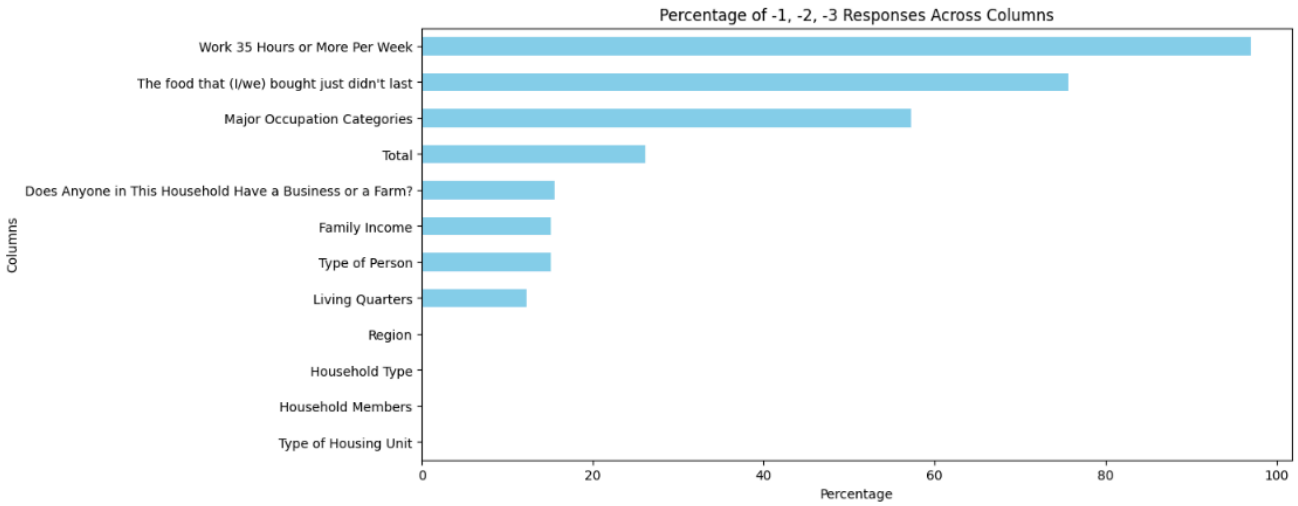
iteration throughout minimizing a cost function. Since the K-modes need fewer iterations to converge than the K-means, the modification to categorical data resulted in a faster algorithm (Tolner, Fegyverneki, Eigner, & Barta, 2021). The elbow method[I] was used to determine the optimal number of clusters required for analysis.

Statistical Findings, Insights, and Next Steps

As previously mentioned, the dataset I worked with was quite extensive. To be specific, I only analyzed 2% of the total columns and data points, which amounts to 1,528,604 data points out of 70,871,640. Despite the limited sample size, I gleaned valuable insights that could be useful for further examination. This highlights the importance of conducting EDA (Exploratory Data Analysis) in data analysis. This section is divided into two parts: firstly, the statistical findings I observed, and secondly, the recommended actions that should be taken after analyzing only 2% of the data.

Statical Findings

After conducting a comprehensive review of the data and results, it was discovered that a significant portion of the responses gathered from all the columns (26%) were either left blank,



marked as "don't know," or refused. These values were included to see the holistic picture of the data and to identify correlations with each respective column. However, concerning results were found for the question "PEHRFTPT - Do you work 35 hours or more a week?" as it had the highest total of non-descriptive responses, with nearly 97% of respondents not providing a response or answering with "don't know" or refusal.

Univariate Analysis

According to the results of a univariate analysis, it was observed that the majority of individuals either possess or pay for a single-unit dwelling, such as a house, apartment, or flat. The distribution of family income reveals that the highest frequency of responses, after excluding the blank responses, is a yearly family income of $150,000 or higher. It is commonly believed that a higher family income corresponds to a lower likelihood of experiencing food insecurity. However, since the CPS has been in operation for more than 50 years, while the Food Security Supplement has only recently been introduced, there may be flaws in the data collection process that could affect the accuracy of food security assessments in America. Additionally, if one assumes that if a family income is toward the lower end of the spectrum, they may not have a home or apartment available for an interview, resulting in an increased likelihood of food insecurity.

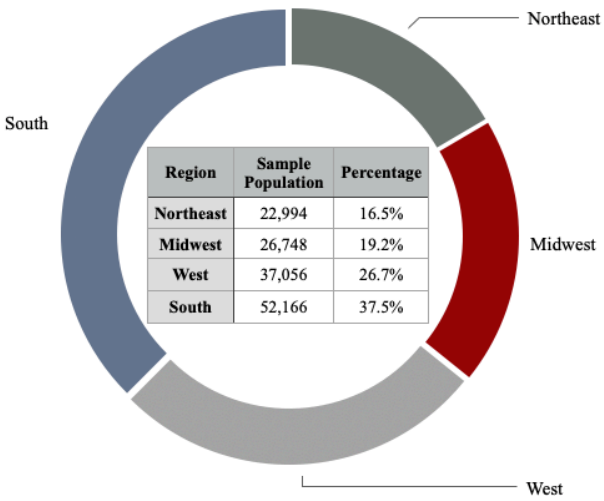
The findings indicate that the majority of households have two members, with the next most common size being four members. The primary family structure reported in the interviews was a married Male and Female. Based on this data, there is a high probability that households are married, and if they have children, it can be inferred that they would have two children. Additionally, there is a gradual decrease in frequency as household size occupants increases.

Based on the interviews conducted, it was observed that most of the respondents are not business or farm owners and are situated in the Southern Region. The highest percentage of occupations falls under the category of management, professional, and related occupations, followed by sales and office occupations. Furthermore, the majority of respondents stated that they do not have any concerns about their food consumption.

Multivariate Analysis

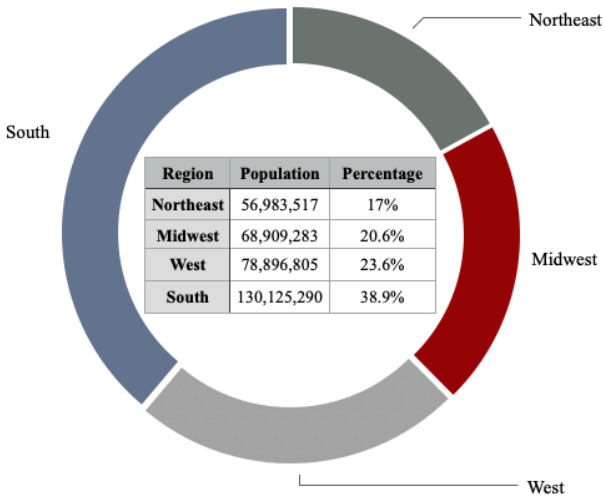
The analysis was conducted by grouping the data based on the Region. This grouping aimed to see how different regions of America differ and if there were meaningful insights into how Region plays a role. However, it was observed that the number of data points was disproportionately skewed towards the South, indicating that the results may not be representative of the entire population. While the South led in every category and result except one (where the West had more households with seven members), the multivariate analysis proved to be of little value due to the uneven distribution of data points. On cross-referencing the US Census Field with the 2023 data, we found that the percentage of data points belonging to the

2019 CPS Regional Data



Region	Sample Population	Percentage
Northeast	22,994	16.5%
Midwest	26,748	19.2%
West	37,056	26.7%
South	52,166	37.5%

2023 Census Regional Data



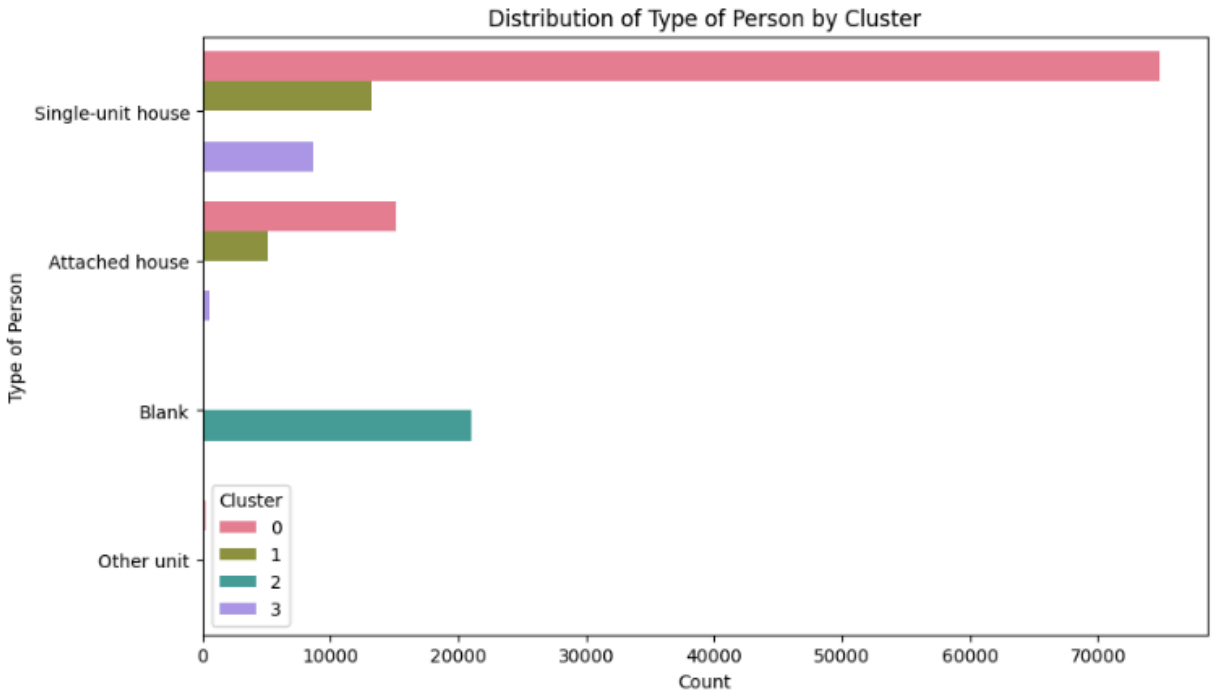
Region	Population	Percentage
Northeast	56,983,517	17%
Midwest	68,909,283	20.6%
West	78,896,805	23.6%
South	130,125,290	38.9%

Region was similar, indicating that the sampling size was accurate. (Source: United States Census Bureau - Growth, 2024).

The process of Cluster Analysis is not designed to enforce identical cluster assignments across all variables. However, it is possible to observe some consistency in clusters across multiple variables, indicating that the cluster contains data points that share similar modes for various categorical variables.

By adjusting the number of cluster points, it showed similar grouping methods throughout the first cluster assignment, which is marked as 0, indicating common patterns between the selected columns. Different cluster profiles were experimented with, including 3, 4, 5, and 8, with clustering with 4 being the most intriguing pattern observed.

With this pattern, the majority of fields containing data had the most predominant answer, instead of being blank, "don't know," or "refused." Cluster 0 is likely to be more internally consistent regarding that variable, as demonstrated below. Members of the cluster are similar to each other with respect to their responses or values for that field chosen.



Subsequent Actions

It is important to investigate the reasons behind the high rate of blank, don't know, and refused answers in certain columns during interviews. This information can shed light on the case at hand. For instance, if the question was asked in a way that didn't apply to certain fields, this could explain why some fields have a higher rate of blanks than other types of responses. Another test that needs to be conducted is to randomly select a large sample size of the 510 optional fields to see if other fields exhibit similar behavior. By examining both the interview process and the behavior of other fields, we can determine whether or not these fields should be included in further analysis.

Research needs to be conducted to understand why the majority of family incomes in the survey are $150,000 or more annually. The high frequency of such responses raises concerns about the reliability of the survey. It could be due to flaws in the survey methodology or the likelihood of respondents not answering truthfully. Therefore, the survey should be expanded to cover a wider range of family incomes to see if respondents are simply choosing the highest value possible.

If the Census Bureau aims to identify food security issues, it may need to adjust the target audience. While it's possible for families in the higher socioeconomic bracket to face such problems, it's generally understood that lower-income families are more vulnerable. Therefore, the research needs to delve deeper and determine at what level of family income food security concerns start to become an issue.

A more detailed analysis is required to determine if certain major occupational categories increase the likelihood of being unable to consistently provide food for the household. Analysis needs to identify if there are any patterns in professions that lead to higher rates of food

insecurity and conduct a correlation analysis with household income. It's possible that seasonal or commission-based jobs could have a high yearly income but lack regular pay, which may affect the ability to provide meals consistently.

During the EDA process, one of the most intriguing outcomes was the result of cluster analysis. We observed distinctive patterns within the 0 cluster that set it apart from others. The observed pattern is a defining characteristic representing a shared behavior, preference, or trait among the individuals in that cluster. Knowing that a cluster exhibits a dominant response allows for targeted insights and a more precise understanding of that group's characteristics. It is important to note that further exploration into these clusters could provide a treasure trove of targeted insights that could help in addressing the food security issue in America.

References

Avinash Navlani, A. F. (2019). *Python Data Analysis - Third Edition.* Packt Publishing.

Mukhiya, S. K., & Ahmed, U. (2020). *Hands-On Exploratory Data Analysis with Python.* Packt Publishing.

Tolner, F., Fegyverneki, S., Eigner, G., & Barta, B. (2021). Clustering based on Preferences with K-modes using Categorical Variables. *IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY), Subotica, Serbia*, pp. 55-60.

Ungvarsky, J. (2023). *Cluster analysis.* Salem Press Encyclopedia.

Zhengwei, Z., Yangb, G., & Li, Z. (2023). Outlier detection for incomplete real-valued data based on inner boundary. *Journal of Intelligent & Fuzzy Systems 44*, 3023–3041.

Appendix

[A]

```
import pandas as pd

#Load the dataset into a DataFrame
file_path = '/Users/cwaldner/Sandbox/TIM_8501/Week8/cps.csv'
df = pd.read_csv(file_path)

# Prints basic information
print(df.info())

# Prints the first 5 rows
print(df.head())
```

[B]

Name_of_Column	Friendly_Name	Possible_Values	Value_Meaning
PRPERTYP	Type of Person	1	Child household member (0-14 years old)
PRPERTYP	Type of Person	2	Adult civilian household member (15+ years old)
PRPERTYP	Type of Person	3	Adult Armed Forces household member (15+ years old)
HETENURE	Living Quarters	1	OWNED OR BEING BOUGHT BY A HH MEMBER
HETENURE	Living Quarters	2	RENTED FOR CASH
HETENURE	Living Quarters	3	OCCUPIED WITHOUT PAYMENT OF CASH RENT
HEHOUSUT	TYPE OF HOUSING UNIT	0	OTHER UNIT
HEHOUSUT	TYPE OF HOUSING UNIT	1	HOUSE, APARTMENT, FLAT
HEHOUSUT	TYPE OF HOUSING UNIT	2	HU IN NONTRANSIENT HOTEL, MOTEL, ETC.
HEHOUSUT	TYPE OF HOUSING UNIT	3	HU PERMANENT IN TRANSIENT HOTEL, MOTEL
HEHOUSUT	TYPE OF HOUSING UNIT	4	HU IN ROOMING HOUSE
HEHOUSUT	TYPE OF HOUSING UNIT	5	MOBILE HOME OR TRAILER W/NO PERM. ROOM ADDED
HEHOUSUT	TYPE OF HOUSING UNIT	6	MOBILE HOME OR TRAILER W/1 OR MORE PERM. ROOMS ADDED
HEHOUSUT	TYPE OF HOUSING UNIT	7	HU NOT SPECIFIED ABOVE
HEHOUSUT	TYPE OF HOUSING UNIT	8	QUARTERS NOT HU IN ROOMING OR BRDING HS
HEHOUSUT	TYPE OF HOUSING UNIT	9	UNIT NOT PERM. IN TRANSIENT HOTL, MOTL
HEHOUSUT	TYPE OF HOUSING UNIT	10	UNOCCUPIED TENT SITE OR TRLR SITE
HEHOUSUT	TYPE OF HOUSING UNIT	11	STUDENT QUARTERS IN COLLEGE DORM

HEHOUSUT	TYPE OF HOUSING UNIT	12	OTHER UNIT NOT SPECIFIED ABOVE
HEFAMINC	FAMILY INCOME	1	LESS THAN $5,000
HEFAMINC	FAMILY INCOME	2	5,000 TO 7,499
HEFAMINC	FAMILY INCOME	3	7,500 TO 9,999
HEFAMINC	FAMILY INCOME	4	10,000 TO 12,499
HEFAMINC	FAMILY INCOME	5	12,500 TO 14,999
HEFAMINC	FAMILY INCOME	6	15,000 TO 19,999
HEFAMINC	FAMILY INCOME	7	20,000 TO 24,999
HEFAMINC	FAMILY INCOME	8	25,000 TO 29,999
HEFAMINC	FAMILY INCOME	9	30,000 TO 34,999
HEFAMINC	FAMILY INCOME	10	35,000 TO 39,999
HEFAMINC	FAMILY INCOME	11	40,000 TO 49,999
HEFAMINC	FAMILY INCOME	12	50,000 TO 59,999
HEFAMINC	FAMILY INCOME	13	60,000 TO 74,999
HEFAMINC	FAMILY INCOME	14	75,000 TO 99,999
HEFAMINC	FAMILY INCOME	15	100,000 TO 149,999
HEFAMINC	FAMILY INCOME	16	150,000 OR MORE
HRNUMHOU	HOUSEHOLD MEMBERS	1	1 Member
HRNUMHOU	HOUSEHOLD MEMBERS	2	2 Members
HRNUMHOU	HOUSEHOLD MEMBERS	3	3 Members
HRNUMHOU	HOUSEHOLD MEMBERS	4	4 Members
HRNUMHOU	HOUSEHOLD MEMBERS	5	5 Members
HRNUMHOU	HOUSEHOLD MEMBERS	6	6 Members
HRNUMHOU	HOUSEHOLD MEMBERS	7	7 Members
HRNUMHOU	HOUSEHOLD MEMBERS	8	8 Members
HRNUMHOU	HOUSEHOLD MEMBERS	9	9 Members
HRNUMHOU	HOUSEHOLD MEMBERS	10	10 Members
HRNUMHOU	HOUSEHOLD MEMBERS	11	11 Members
HRNUMHOU	HOUSEHOLD MEMBERS	12	12 Members
HRNUMHOU	HOUSEHOLD MEMBERS	13	13 Members
HRNUMHOU	HOUSEHOLD MEMBERS	14	14 Members
HRNUMHOU	HOUSEHOLD MEMBERS	15	15 Members
HRNUMHOU	HOUSEHOLD MEMBERS	16	16 Members
HRHTYPE	HOUSEHOLD TYPE	0	NON-INTERVIEW HOUSEHOLD
HRHTYPE	HOUSEHOLD TYPE	1	HUSBAND/WIFE PRIMARY FAMILY (NEITHER AF)
HRHTYPE	HOUSEHOLD TYPE	2	HUSB/WIFE PRIM. FAMILY (EITHER/BOTH AF)
HRHTYPE	HOUSEHOLD TYPE	3	UNMARRIED CIVILIAN MALE-PRIM. FAM HHLDER

HRHTYPE	HOUSEHOLD TYPE	4	UNMARRIED CIV. FEMALE-PRIM FAM HHLDER
HRHTYPE	HOUSEHOLD TYPE	5	PRIMARY FAMILY HHLDER-RP IN AF, UNMAR.
HRHTYPE	HOUSEHOLD TYPE	6	CIVILIAN MALE PRIMARY INDIVIDUAL
HRHTYPE	HOUSEHOLD TYPE	7	CIVILIAN FEMALE PRIMARY INDIVIDUAL
HRHTYPE	HOUSEHOLD TYPE	8	PRIMARY INDIVIDUAL HHLD-RP IN AF
HRHTYPE	HOUSEHOLD TYPE	9	GROUP QUARTERS WITH FAMILY
HRHTYPE	HOUSEHOLD TYPE	10	GROUP QUARTERS WITHOUT FAMILY
HUBUS	DOES ANYONE IN THIS HOUSEHOLD HAVE A BUSINESS OR A FARM?	1	Yes
HUBUS	DOES ANYONE IN THIS HOUSEHOLD HAVE A BUSINESS OR A FARM?	2	No
GEREG	REGION	1	NORTHEAST
GEREG	REGION	2	MIDWEST
GEREG	REGION	3	SOUTH
GEREG	REGION	4	WEST
PEHRFTPT	WORK 35 HOURS OR MORE PER WEEK	1	Yes
PEHRFTPT	WORK 35 HOURS OR MORE PER WEEK	2	No
PEHRFTPT	WORK 35 HOURS OR MORE PER WEEK	3	Hours Vary
PRMJOCGR	MAJOR OCCUPATION CATEGORIES	1	Management, professional, and related occupations
PRMJOCGR	MAJOR OCCUPATION CATEGORIES	2	Service occupations
PRMJOCGR	MAJOR OCCUPATION CATEGORIES	3	Sales and office occupations
PRMJOCGR	MAJOR OCCUPATION CATEGORIES	4	Farming, fishing, and forestry occupations
PRMJOCGR	MAJOR OCCUPATION CATEGORIES	5	Construction, and maintenance occupations
PRMJOCGR	MAJOR OCCUPATION CATEGORIES	6	Production, transportation, and material moving occupations
PRMJOCGR	MAJOR OCCUPATION CATEGORIES	7	Armed Forces
HESS3	The food that (I/we) bought just didn't last, and (I/we) didn't have money to get more. Was that OFTEN, SOMETIMES, or NEVER true for (you/your household) in the last 12 months?	1	Often true
HESS3	The food that (I/we) bought just didn't last, and (I/we) didn't have money to get more. Was that OFTEN, SOMETIMES, or NEVER true for (you/your household) in the last 12 months?	2	Sometimes true
HESS3	The food that (I/we) bought just didn't last, and (I/we) didn't have money to get more. Was that OFTEN, SOMETIMES, or NEVER true for (you/your household) in the last 12 months?	3	Never true

[C]

```
import pandas as pd
```

```
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import kurtosis, skew

# Load the Dataset
file_path = '/Users/cwaldner/Sandbox/TIM_8501/Week8/cps.csv'
df = pd.read_csv(file_path)

# Selected Columns for EDA
columns_for_eda = ['PRPERTYP', 'HETENURE', 'HEHOUSUT', 'HEFAMINC', 'HRNUMHOU',
'HRHTYPE', 'HUBUS', 'GEREG', 'PEHRFTPT', 'PRMJOCGR', 'HESS3']

# Subset DataFrame with selected columns
df_subset = df[columns_for_eda]

# Measures of Central Tendency
mean_values = df_subset.mean()
median_values = df_subset.median()
mode_values = df_subset.mode().iloc[0]  # Select the first row from the mode DataFrame

# Measures of Variability
range_values = df_subset.max() - df_subset.min()
iqr_values = df_subset.quantile(0.75) - df_subset.quantile(0.25)
variance_values = df_subset.var()
std_deviation_values = df_subset.std()

# Additional Measures
min_values = df_subset.min()
max_values = df_subset.max()
kurtosis_values = df_subset.kurtosis()
skewness_values = df_subset.skew()

# Store statistics in a DataFrame
statistics_df = pd.DataFrame({
    'Mean': mean_values,
    'Median': median_values,
    'Mode': mode_values,
    'Minimum': min_values,
    'Maximum': max_values,
    'Range': range_values,
    'IQR': iqr_values,
    'Variance': variance_values,
    'Std Deviation': std_deviation_values,
    'Kurtosis': kurtosis_values,
    'Skewness': skewness_values
})

# Print the DataFrame
print(statistics_df)

# Plot histograms for numerical columns
for column in df_subset.select_dtypes(include=['int64', 'float64']).columns:
    plt.figure(figsize=(10, 6))
    sns.histplot(df_subset[column], kde=True)
    plt.title(f'Histogram of {column}')
    plt.show()
```

[D]

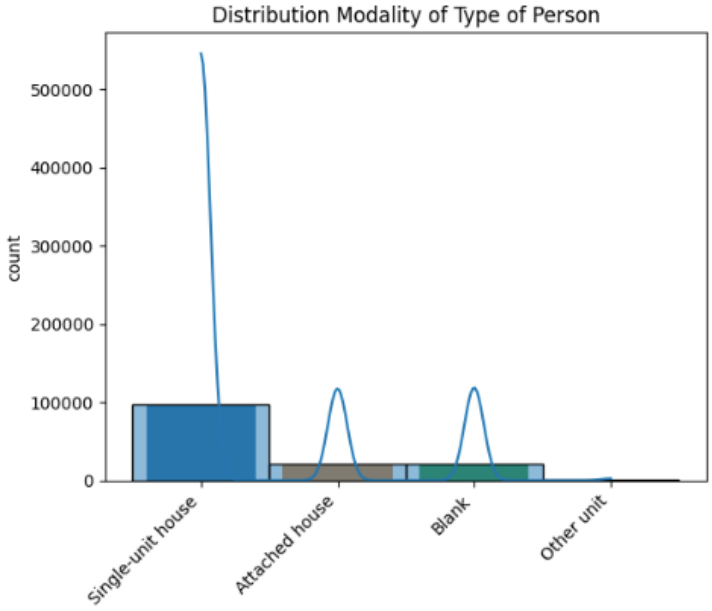
```
PRPERTYP Analysis:
--------------------------------

Frequency for PRPERTYP:
PRPERTYP
 2    96697
-1    21032
 1    20806
 3      429
Name: count, dtype: int64
Variance: 1.1541372566861325
Standard Deviation: 1.0743078035116995

--------------------------------

HETENURE Analysis:
--------------------------------

Frequency for HETENURE:
HETENURE
 1    85327
 2    35239
-1    17059
 3     1339
Name: count, dtype: int64
Variance: 0.7824179305191393
Standard Deviation: 0.8845439110180677

--------------------------------

HEHOUSUT Analysis:
--------------------------------

Frequency for HEHOUSUT:
HEHOUSUT
1     131228
5       5943
6        894
12       287
10       275
7        128
2         92
4         58
3         34
8         24
9          1
Name: count, dtype: int64
Variance: 1.2387904292954408
Standard Deviation: 1.113009626775726

--------------------------------
```

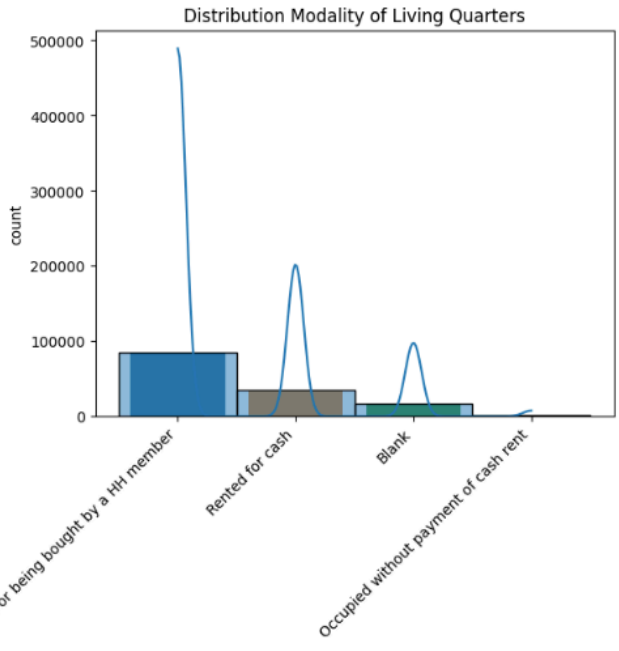
```
HRHTYPE Analysis:
--------------------------------

Frequency for HRHTYPE:
HRHTYPE
1     70274
0     21032
4     16870
7     11166
6     10685
3      7569
2      1211
10       50
8        40
5        34
9        33
Name: count, dtype: int64
Variance: 4.895355001634943
Standard Deviation: 2.2125449151678125

--------------------------------

HUBUS Analysis:
--------------------------------

Frequency for HUBUS:
HUBUS
 2    102221
-1     21032
 1     15192
-3       410
-2       109
Name: count, dtype: int64
Variance: 1.2199177444450315
Standard Deviation: 1.1044988657509032

--------------------------------

GEREG Analysis:
--------------------------------

Frequency for GEREG:
GEREG
3    52166
4    37056
2    26748
1    22994
Name: count, dtype: int64
Variance: 1.0550931186010253
Standard Deviation: 1.0271772576342533

--------------------------------
```

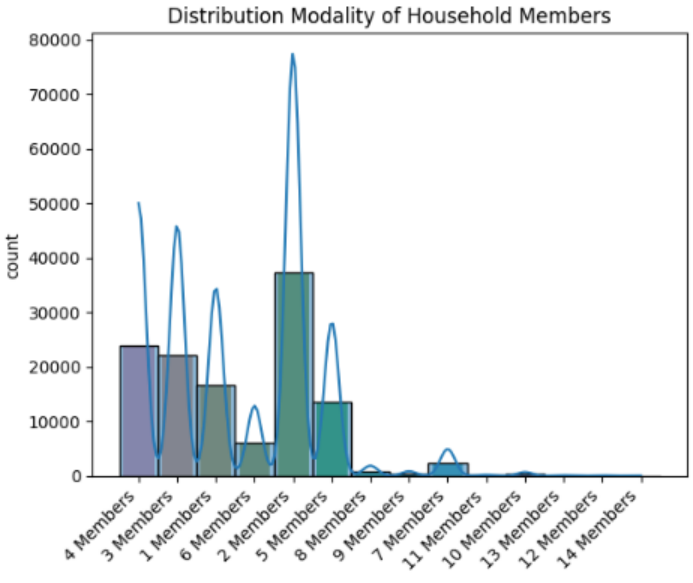
```
HEFAMINC Analysis:
--------------------------------

Frequency for HEFAMINC:
HEFAMINC
-1     21032
 16    17998
 15    17927
 14    15779
 13    12065
 12     9552
 11     8970
 9      6267
 10     6041
 7      5071
 8      4936
 6      3574
 4      2441
 1      2315
 5      2237
 3      1647
 2      1112
Name: count, dtype: int64
Variance: 33.52404927058201
Standard Deviation: 5.789995619219587

--------------------------------

HRNUMHOU Analysis:
--------------------------------

Frequency for HRNUMHOU:
HRNUMHOU
2     37219
4     23981
3     22068
1     16594
0     15189
5     13560
6      6166
7      2364
8       896
9       401
10      334
11       89
13       52
12       37
14       14
Name: count, dtype: int64
Variance: 3.4237056436411213
Standard Deviation: 1.8503258209410367

--------------------------------
```

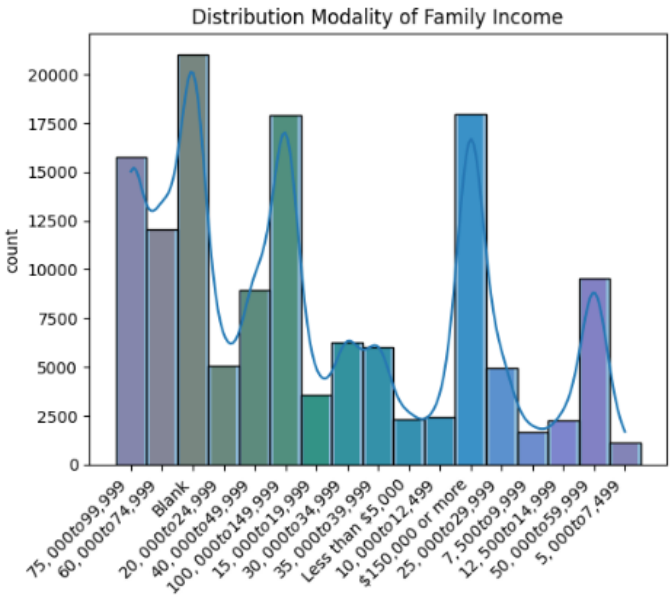
```
PEHRFTPT Analysis:
--------------------------------

Frequency for PEHRFTPT:
PEHRFTPT
-1    134740
 1      3158
 2      1060
 3         6
Name: count, dtype: int64
Variance: 0.15555085758284512
Standard Deviation: 0.3943993630609019

--------------------------------

PRMJOCGR Analysis:
--------------------------------

Frequency for PRMJOCGR:
PRMJOCGR
-1    79564
 1    24659
 3    12614
 2     9753
 6     6860
 5     5005
 4      503
 7        6
Name: count, dtype: int64
Variance: 4.326595119592869
Standard Deviation: 2.080046903219461

--------------------------------

HESS3 Analysis:
--------------------------------

Frequency for HESS3:
HESS3
-1    104899
 3     24725
 2      7320
 1      1741
-9       128
-3       102
-2        49
Name: count, dtype: int64
Variance: 2.648963973708463
Standard Deviation: 1.6275638155563863

--------------------------------
```

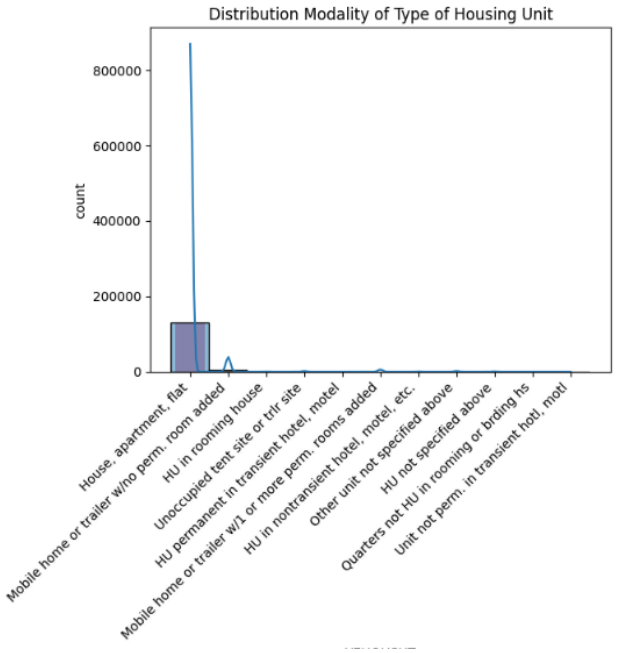
[E]

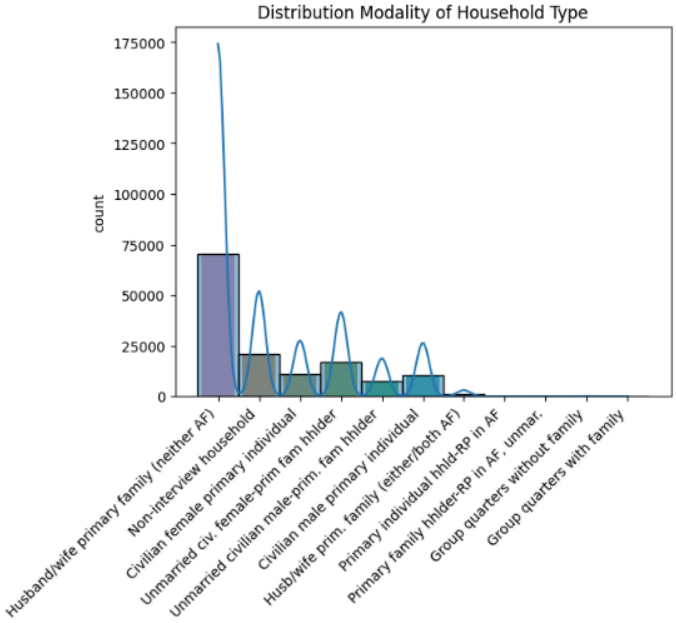

Distribution Modality of Type of Person
count
500000
400000
300000
200000
100000
0
Single-unit house
Attached house
Blank
Other unit

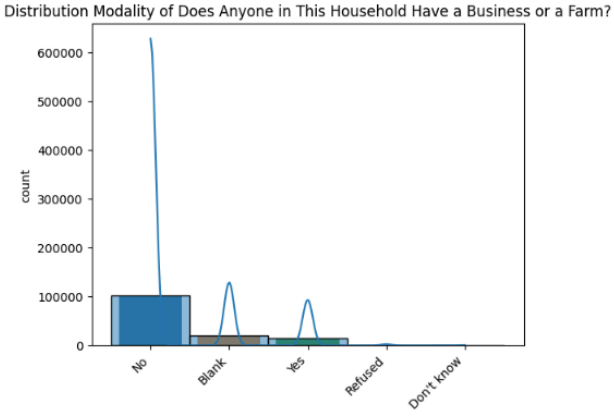

Distribution Modality of Living Quarters
count
500000
400000
300000
200000
100000
0
Owned or being bought by a HH member
Rented for cash
Blank
Occupied without payment of cash rent

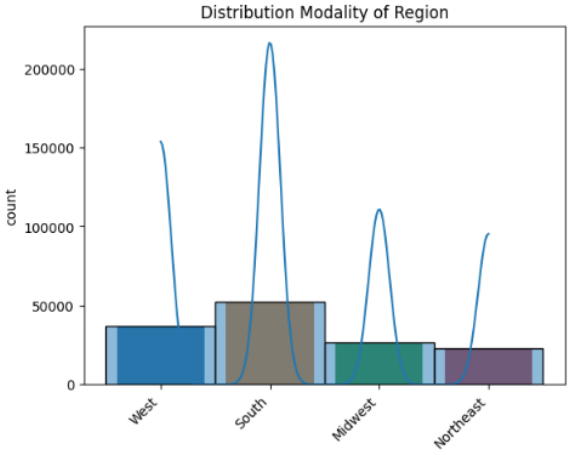

Distribution Modality of Household Members
count
80000
70000
60000
50000
40000
30000
20000
10000
0
4 Members
3 Members
1 Members
6 Members
2 Members
5 Members
8 Members
9 Members
7 Members
11 Members
10 Members
13 Members
12 Members
14 Members

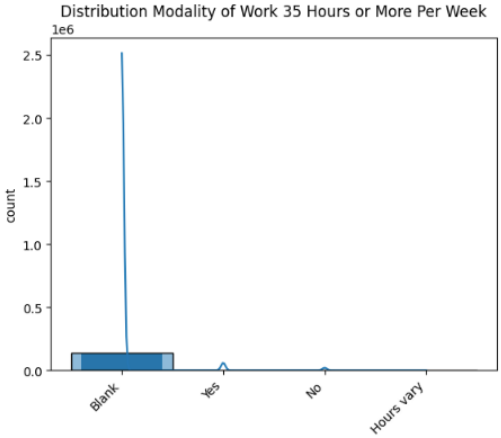

Distribution Modality of Family Income
count
20000
17500
15000
12500
10000
7500
5000
2500
0
75,000to99,999
60,000to74,999
Blank
20,000to24,999
40,000to49,999
100,000to149,999
15,000to19,999
30,000to34,999
35,000to39,999
Less than $5,000
10,000to12,499
$150,000 or more
25,000to29,999
7,500to9,999
12,500to14,999
50,000to59,999
5,000to7,499

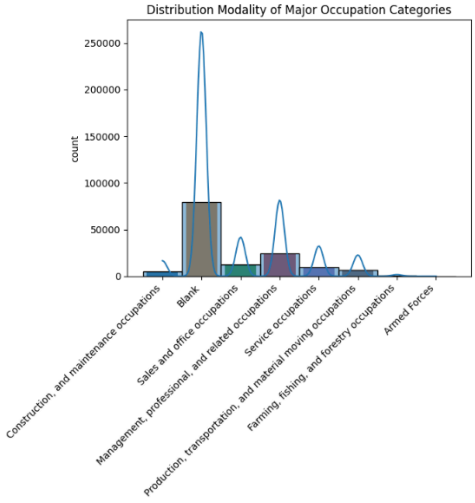

Distribution Modality of Type of Housing Unit
count
800000
600000
400000
200000
0
House, apartment, flat
Mobile home or trailer w/no perm. room added
HU in rooming house
Unoccupied tent site or trlr site
HU permanent in transient hotel, motel
Mobile home or trailer w/1 or more perm. rooms added
HU in nontransient hotel, motel, etc.
Other unit not specified above
HU not specified above
Quarters not HU in rooming or brding hs
Unit not perm. in transient hotl, motl

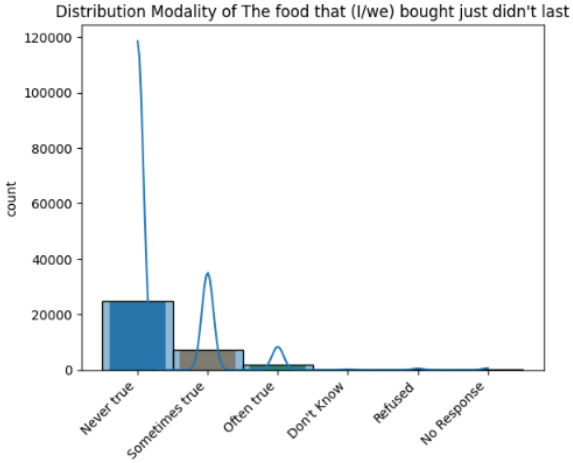

Distribution Modality of Household Type
count
175000
150000
125000
100000
75000
50000
25000
0
Husband/wife primary family (neither AF)
Non-interview household
Civilian female primary individual
Unmarried civ. female-prim fam hhlder
Unmarried civilian male-prim. fam hhlder
Civilian male primary individual
Husb/wife prim. family (either/both AF)
Primary individual hhld-RP in AF
Primary family hhlder-RP in AF, unmar.
Group quarters without family
Group quarters with family


Distribution Modality of Does Anyone in This Household Have a Business or a Farm?
count
600000
500000
400000
300000
200000
100000
0
No
Blank
Yes
Refused
Dont know

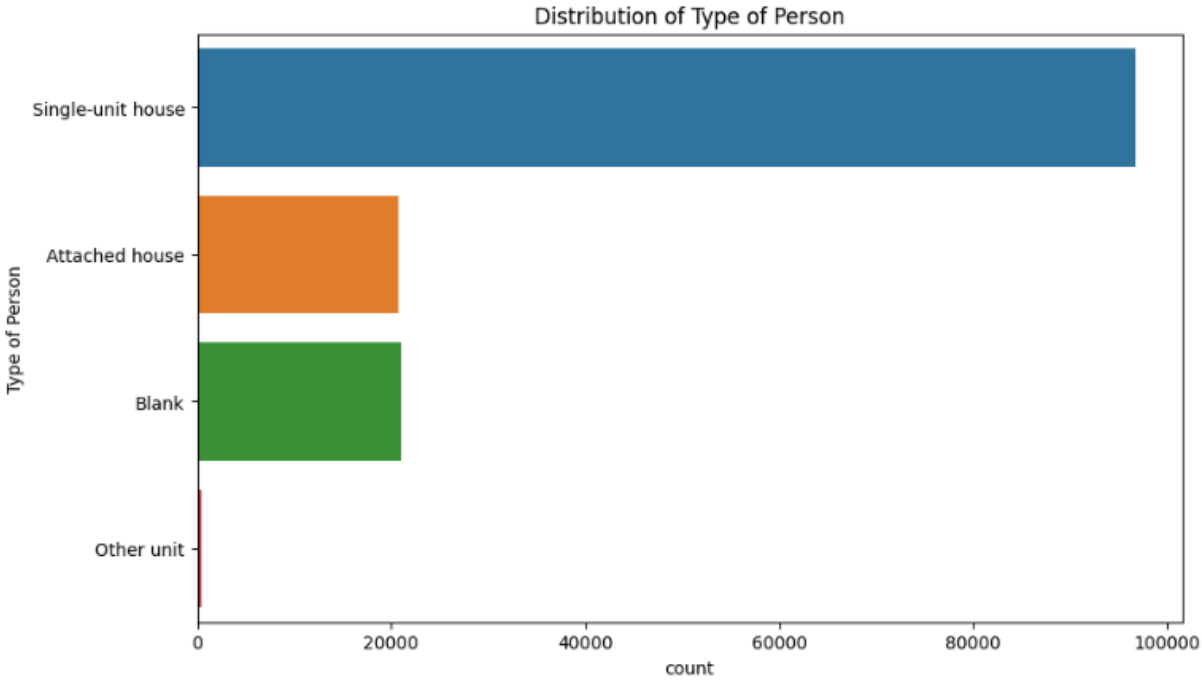

Distribution Modality of Region
count
200000
150000
100000
50000
0
West
South
Midwest
Northeast

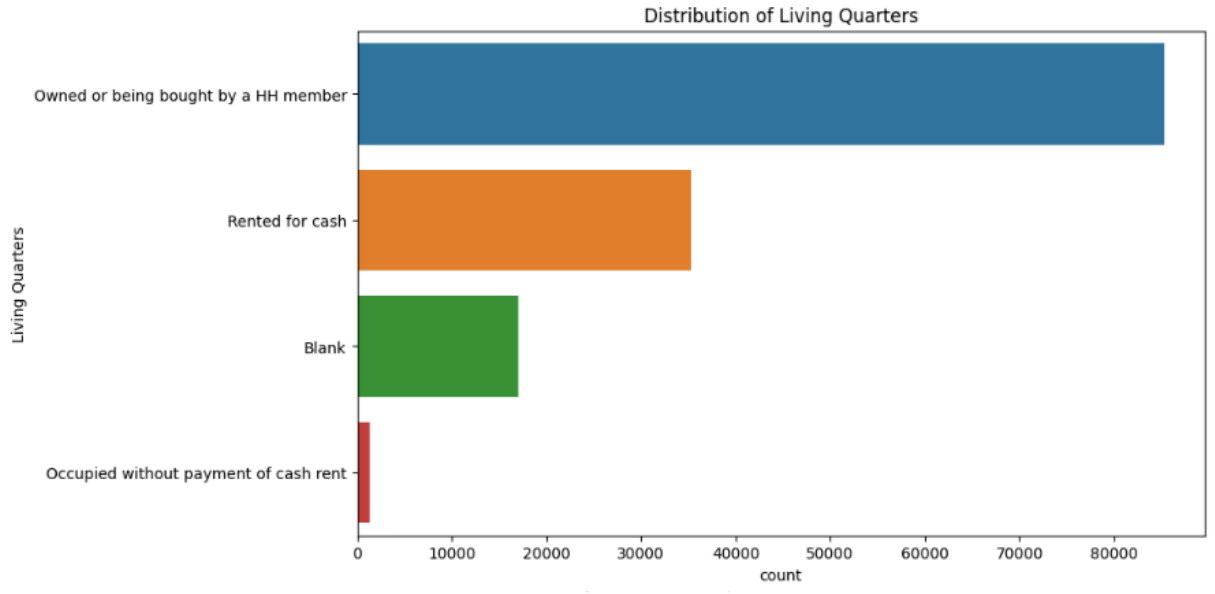

Distribution Modality of Work 35 Hours or More Per Week
1e6
count
2.5
2.0
1.5
1.0
0.5
0.0
Blank
Yes
No
Hours vary

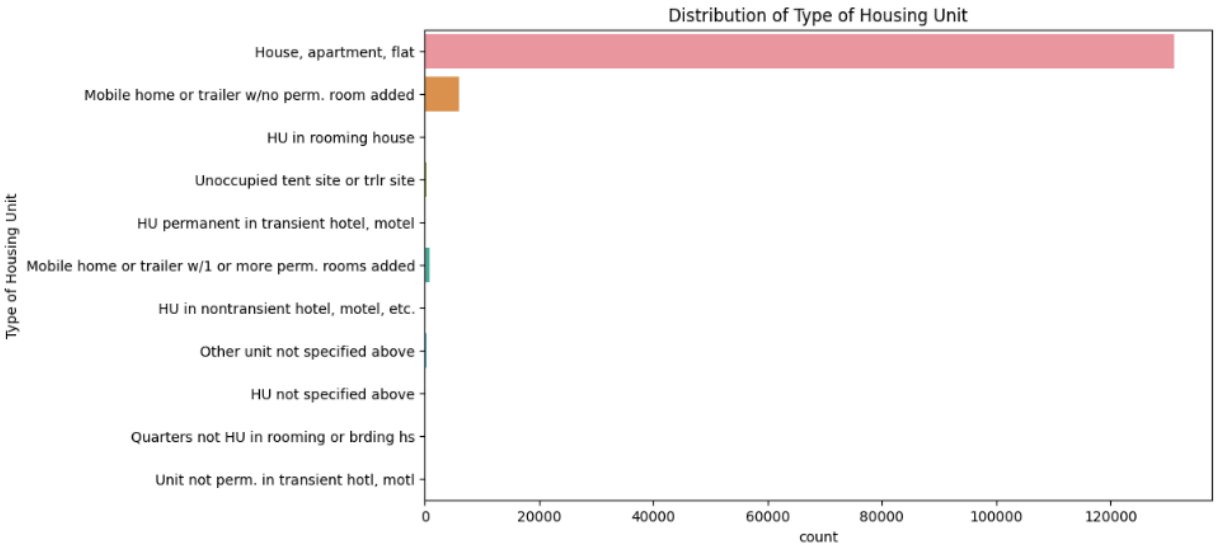

Distribution Modality of Major Occupation Categories
count
250000
200000
150000
100000
50000
0
Construction, and maintenance occupations
Blank
Sales and office occupations
Management, professional, and related occupations
Service occupations
Production, transportation, and material moving occupations
Farming, fishing, and forestry occupations
Armed Forces

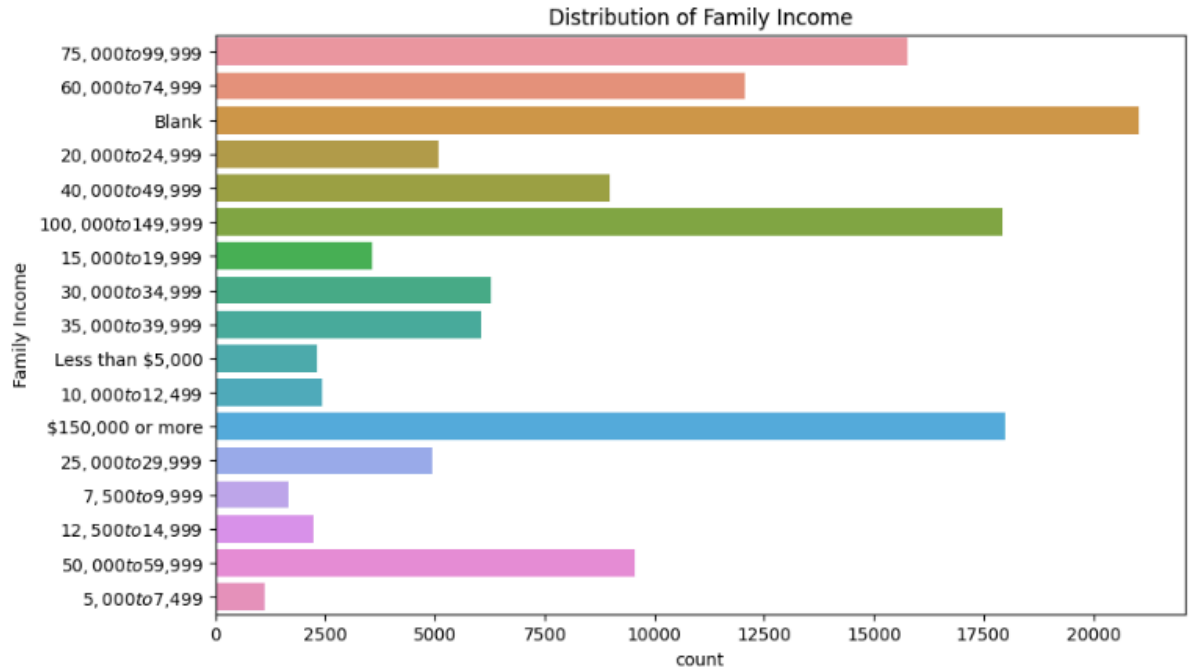

Distribution Modality of The food that (I/we) bought just didn't last
count
120000
100000
80000
60000
40000
20000
0
Never true
Sometimes true
Often true
Don't Know
Refused
No Response

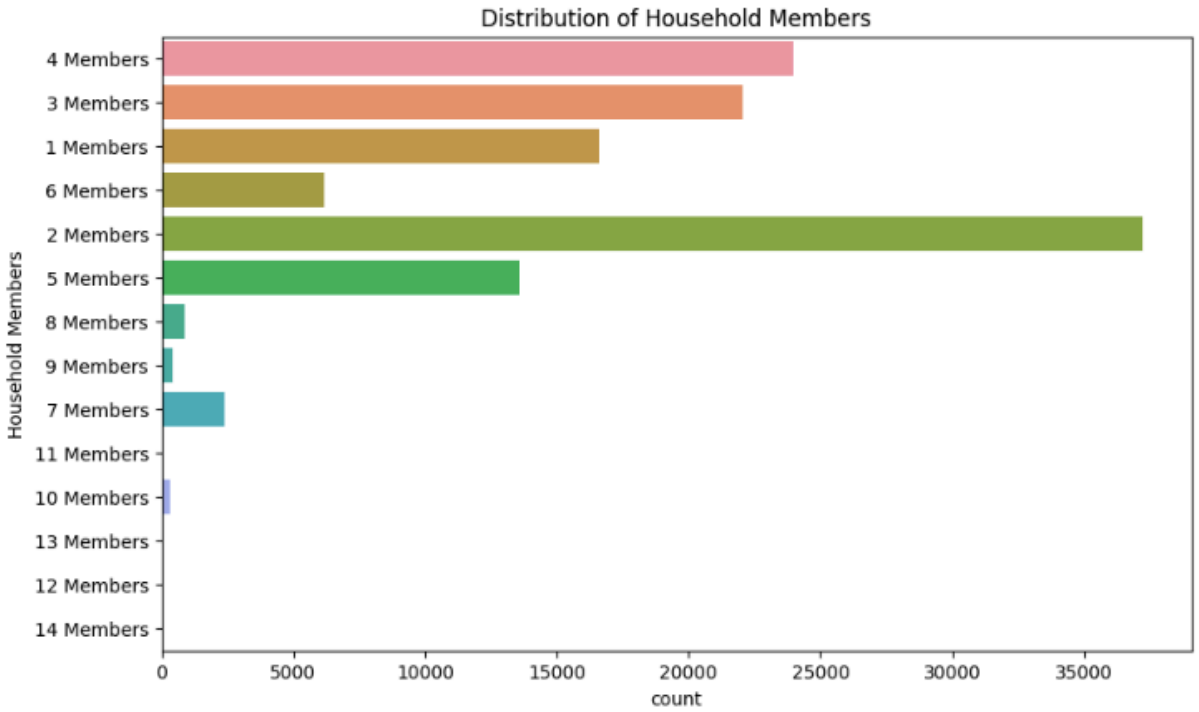
[F]

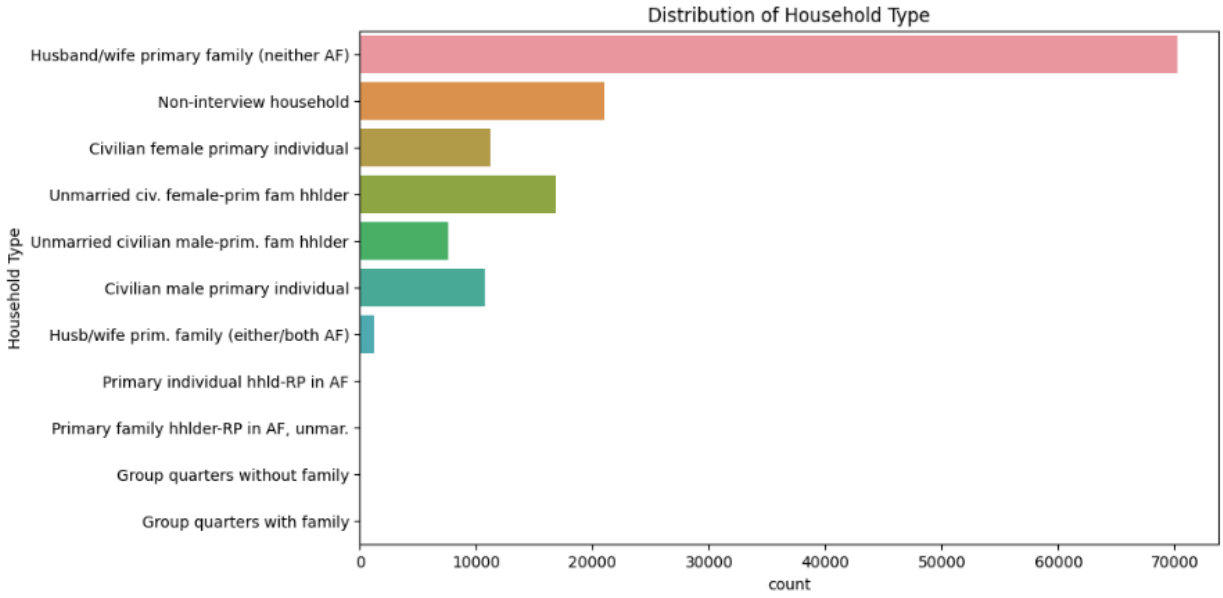

Distribution of Type of Person
Type of Person
Single-unit house
Attached house
Blank
Other unit
0
20000
40000
60000
80000
100000
count

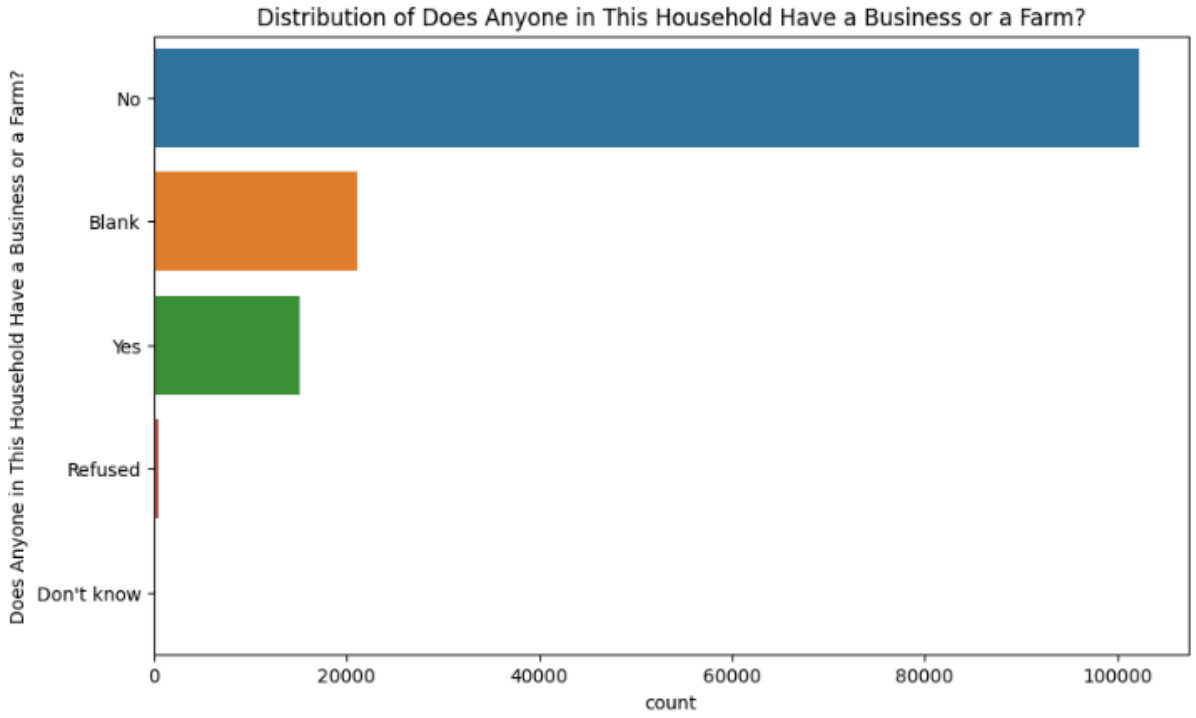

Distribution of Living Quarters
Owned or being bought by a HH member
Rented for cash
Blank
Occupied without payment of cash rent
Living Quarters
0
10000
20000
30000
40000
50000
60000
70000
80000
count

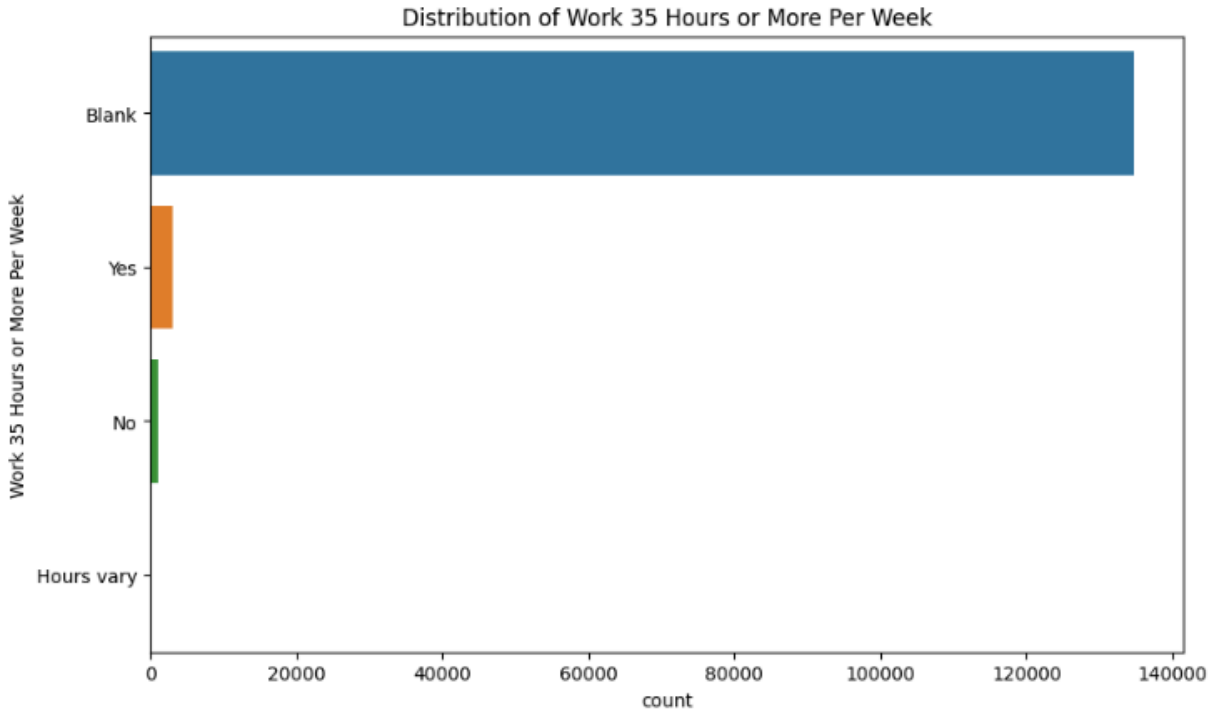

Distribution of Type of Housing Unit
House, apartment, flat
Mobile home or trailer w/no perm. room added
HU in rooming house
Unoccupied tent site or trlr site
HU permanent in transient hotel, motel
Mobile home or trailer w/1 or more perm. rooms added
HU in nontransient hotel, motel, etc.
Other unit not specified above
HU not specified above
Quarters not HU in rooming or brding hs
Unit not perm. in transient hotl, motl
Type of Housing Unit
0
20000
40000
60000
80000
100000
120000
count

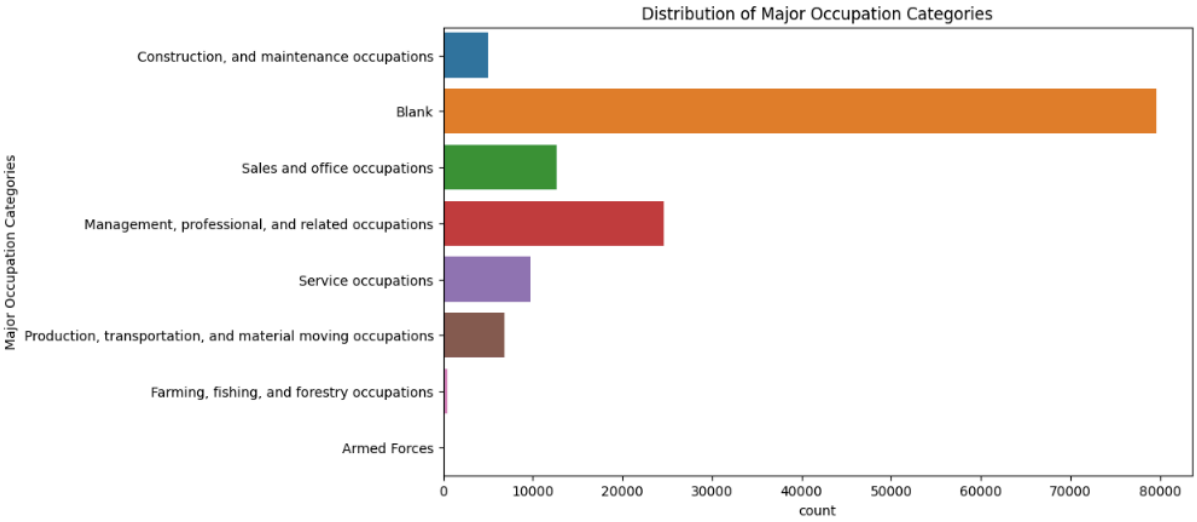

Distribution of Family Income
Family Income
75,000to99,999
60,000to74,999
Blank
20,000to24,999
40,000to49,999
100,000to149,999
15,000to19,999
30,000to34,999
35,000to39,999
Less than $5,000
10,000to12,499
$150,000 or more
25,000to29,999
7,500to9,999
12,500to14,999
50,000to59,999
5,000to7,499
0
2500
5000
7500
10000
12500
15000
17500
20000
count

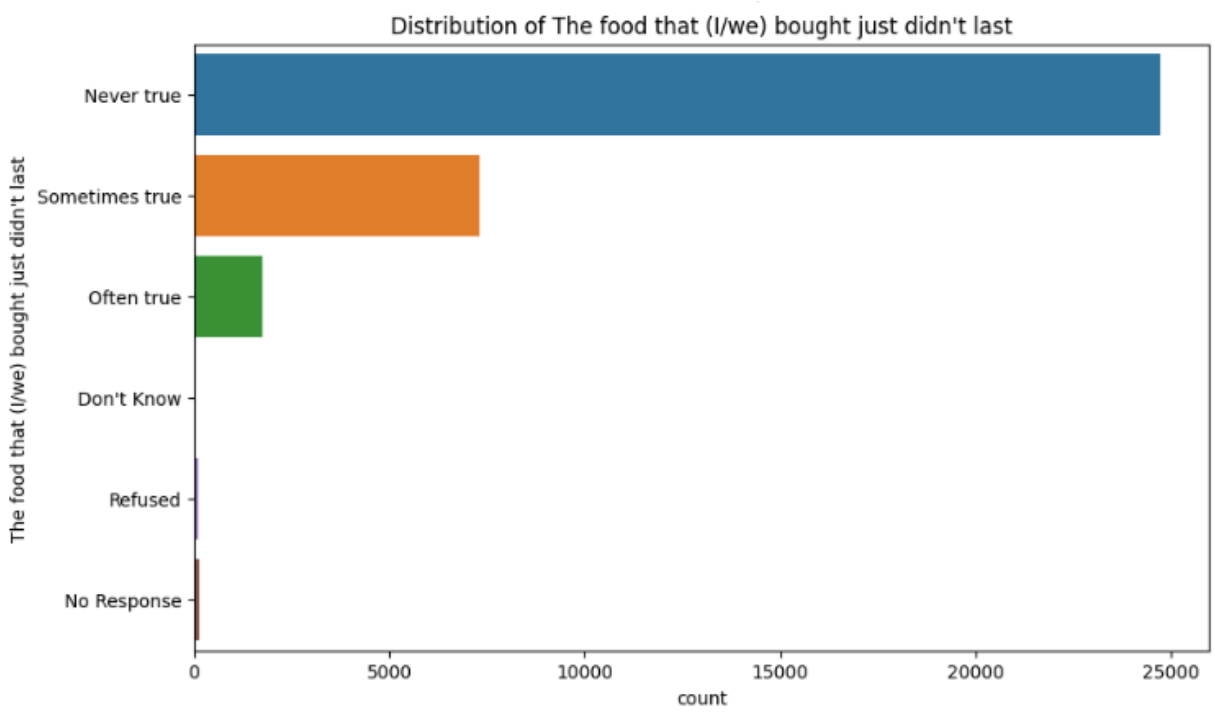

Distribution of Household Members
Household Members
4 Members
3 Members
1 Members
6 Members
2 Members
5 Members
8 Members
9 Members
7 Members
11 Members
10 Members
13 Members
12 Members
14 Members
0
5000
10000
15000
20000
25000
30000
35000
count


Distribution of Household Type
Household Type
Husband/wife primary family (neither AF)
Non-interview household
Civilian female primary individual
Unmarried civ. female-prim fam hhlder
Unmarried civilian male-prim. fam hhlder
Civilian male primary individual
Husb/wife prim. family (either/both AF)
Primary individual hhld-RP in AF
Primary family hhlder-RP in AF, unmar.
Group quarters without family
Group quarters with family
0
10000
20000
30000
40000
50000
60000
70000
count

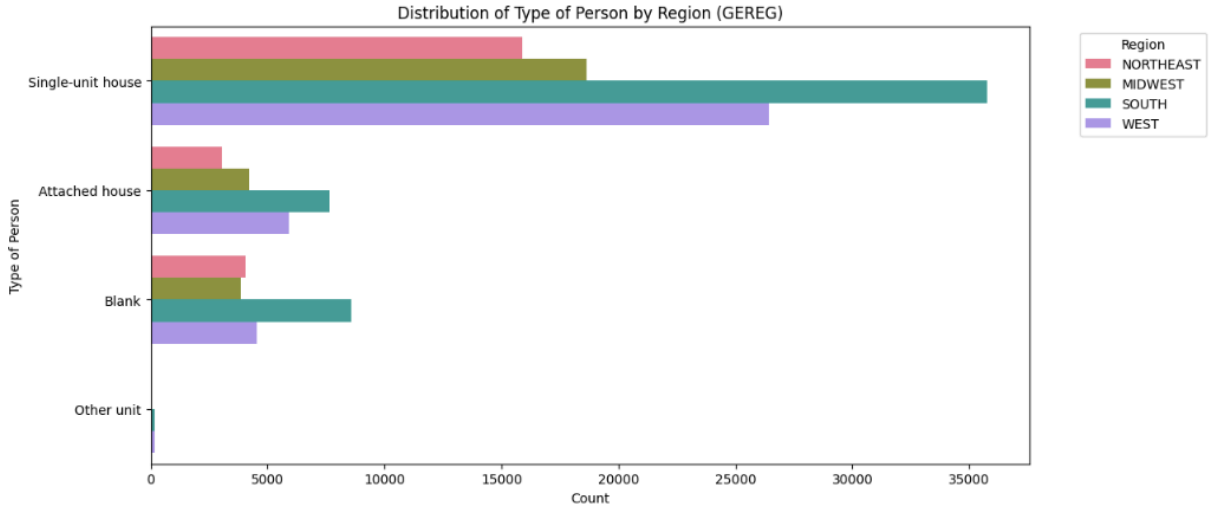

Distribution of Does Anyone in This Household Have a Business or a Farm?
Does Anyone in This Household Have a Business or a Farm?
No
Blank
Yes
Refused
Don't know
0
20000
40000
60000
80000
100000
count

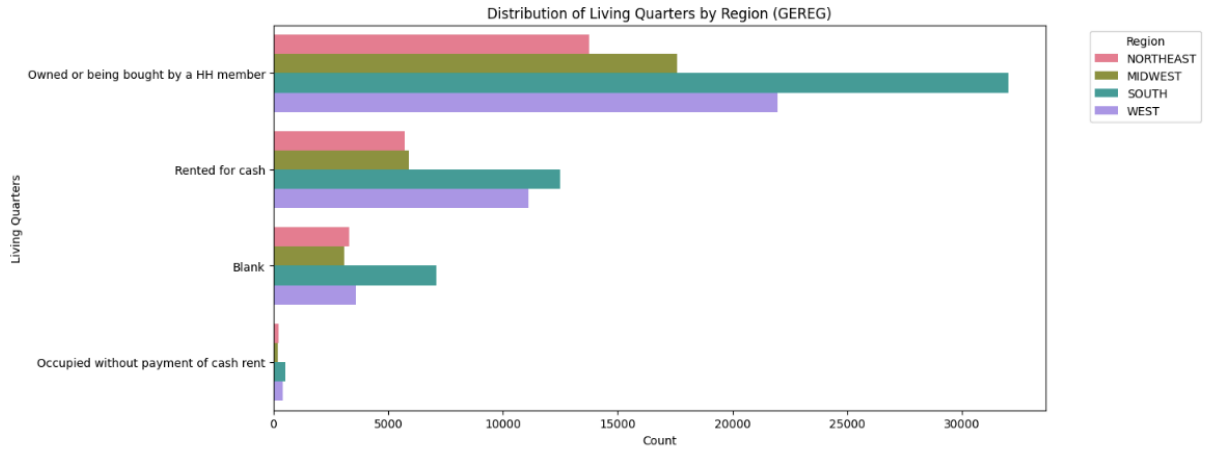

Distribution of Work 35 Hours or More Per Week
Work 35 Hours or More Per Week
Blank
Yes
No
Hours vary
0
20000
40000
60000
80000
100000
120000
140000
count

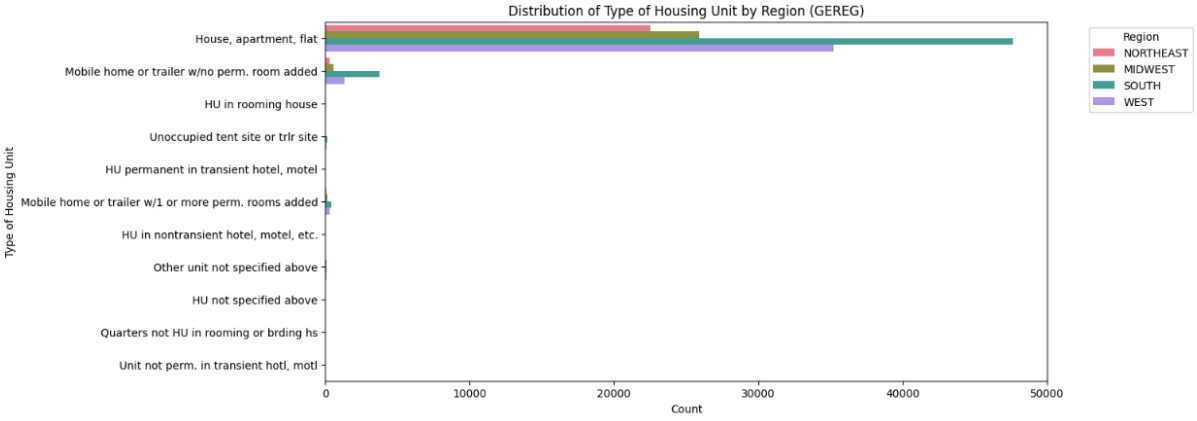

Distribution of Major Occupation Categories
Major Occupation Categories
Construction, and maintenance occupations
Blank
Sales and office occupations
Management, professional, and related occupations
Service occupations
Production, transportation, and material moving occupations
Farming, fishing, and forestry occupations
Armed Forces
0
10000
20000
30000
40000
50000
60000
70000
80000
count

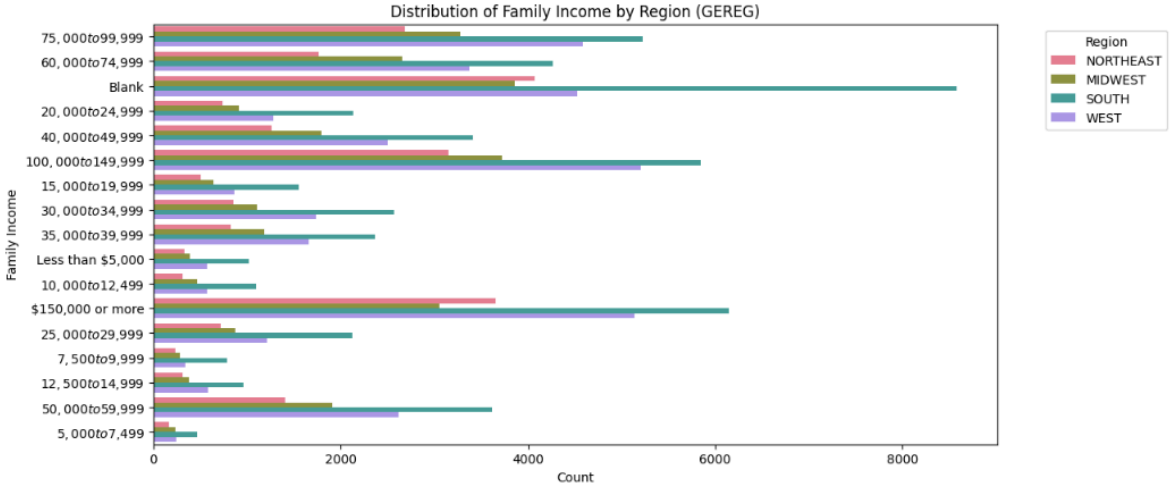

Distribution of The food that (I/we) bought just didn't last
Never true
Sometimes true
Often true
Don't Know
Refused
No Response
The food that (I/we) bought just didn't last
0
5000
10000
15000
20000
25000
count

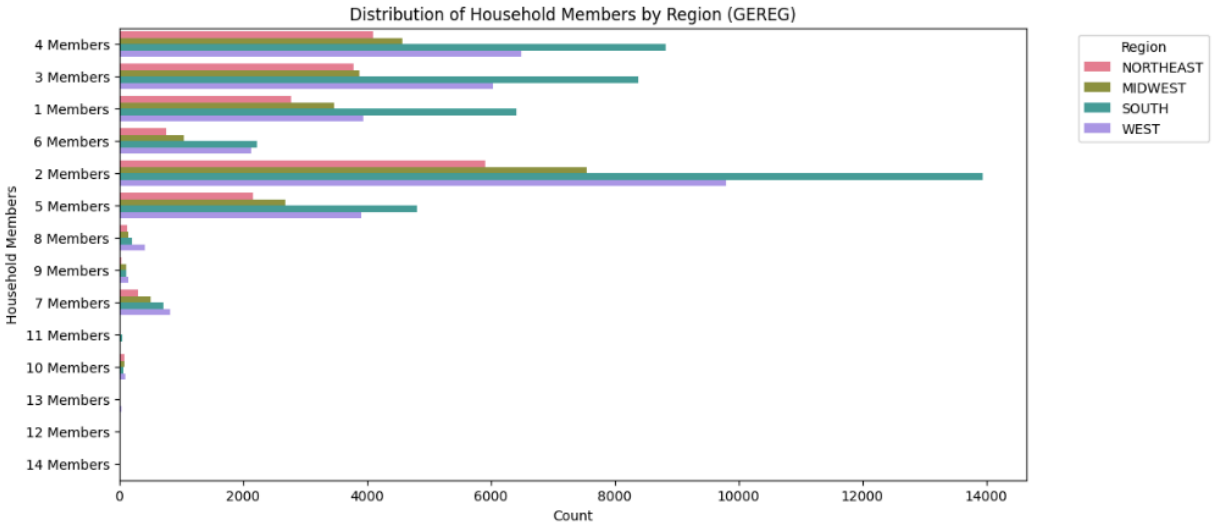
[G]

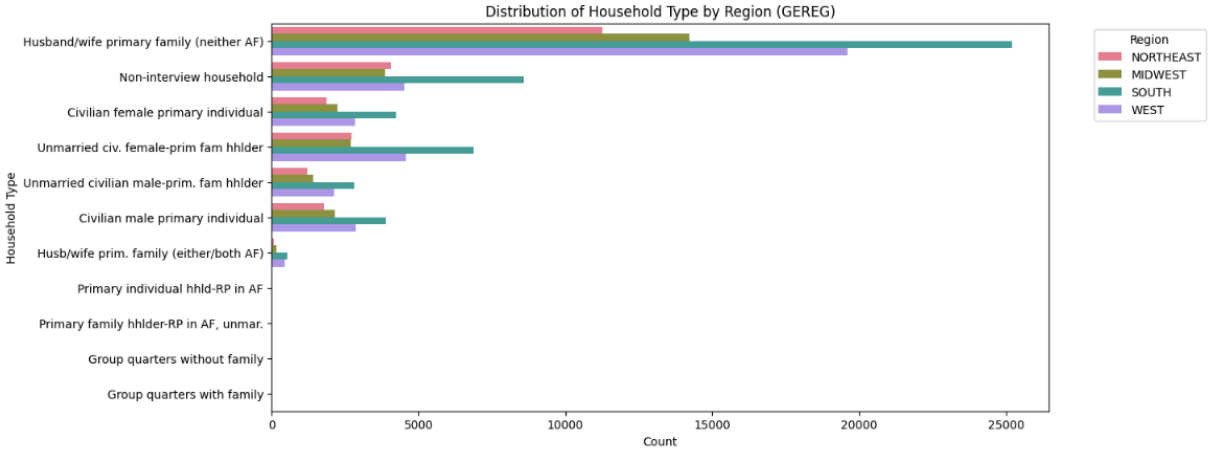

Distribution of Type of Person by Region (GEREG)
Single-unit house
Attached house
Blank
Other unit
Type of Person
0
5000
10000
15000
20000
25000
30000
35000
Count
Region
NORTHEAST
MIDWEST
SOUTH
WEST

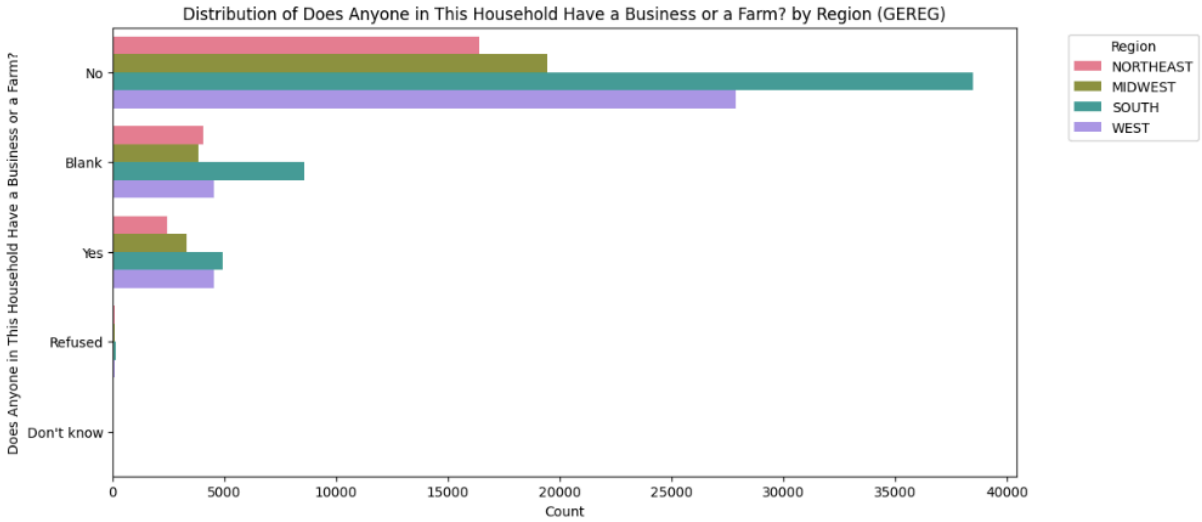

Distribution of Living Quarters by Region (GEREG)
Living Quarters
Owned or being bought by a HH member
Rented for cash
Blank
Occupied without payment of cash rent
0
5000
10000
15000
20000
25000
30000
Count
Region
NORTHEAST
MIDWEST
SOUTH
WEST

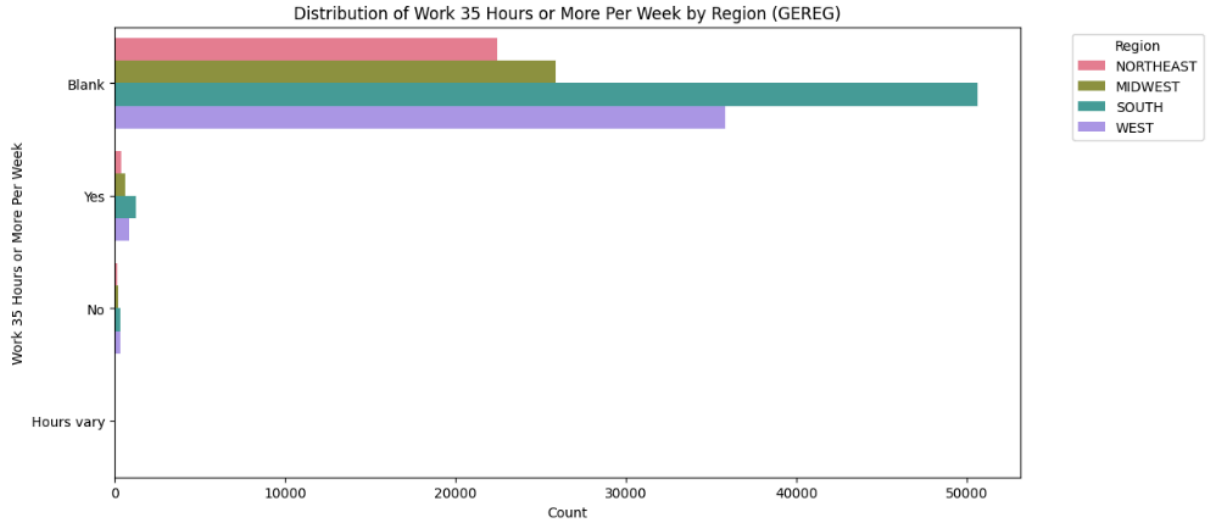

Distribution of Type of Housing Unit by Region (GEREG)
Type of Housing Unit
House, apartment, flat
Mobile home or trailer w/no perm. room added
HU in rooming house
Unoccupied tent site or trlr site
HU permanent in transient hotel, motel
Mobile home or trailer w/1 or more perm. rooms added
HU in nontransient hotel, motel, etc.
Other unit not specified above
HU not specified above
Quarters not HU in rooming or brding hs
Unit not perm. in transient hotl, motl
0
10000
20000
30000
40000
50000
Count
Region
NORTHEAST
MIDWEST
SOUTH
WEST

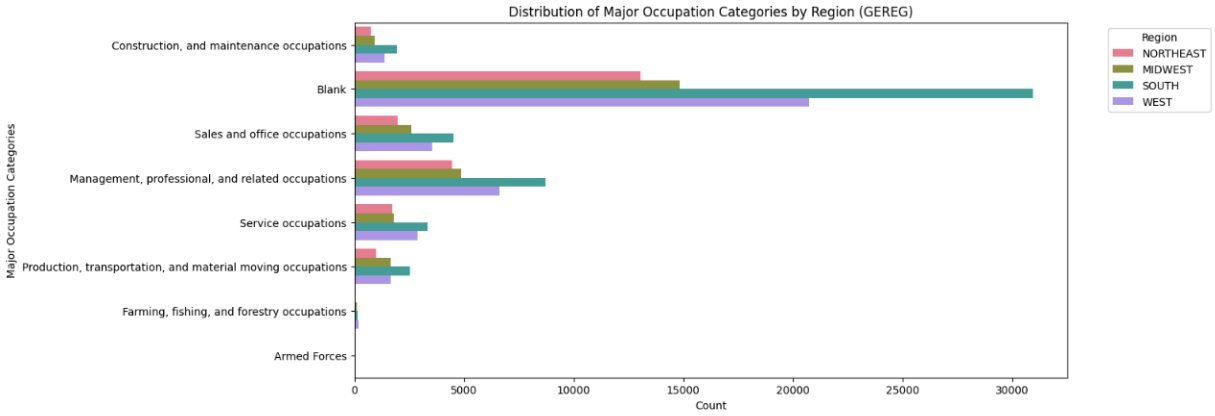

Distribution of Family Income by Region (GEREG)
Family Income
75,000to99,999
60,000to74,999
Blank
20,000to24,999
40,000to49,999
100,000to149,999
15,000to19,999
30,000to34,999
35,000to39,999
Less than $5,000
10,000to12,499
$150,000 or more
25,000to29,999
7,500to9,999
12,500to14,999
50,000to59,999
5,000to7,499
0
2000
4000
6000
8000
Count
Region
NORTHEAST
MIDWEST
SOUTH
WEST

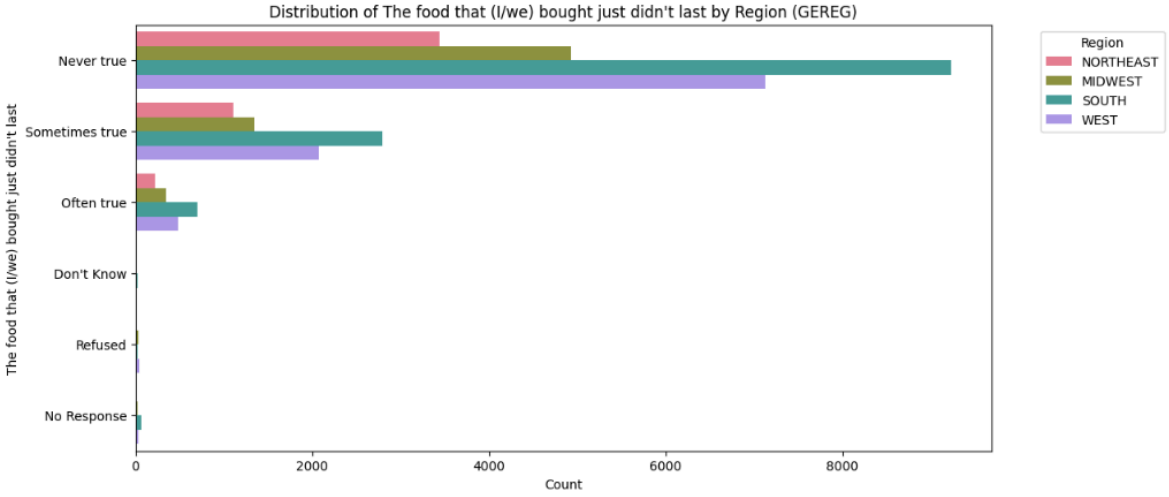

Distribution of Household Members by Region (GEREG)
Household Members
4 Members
3 Members
1 Members
6 Members
2 Members
5 Members
8 Members
9 Members
7 Members
11 Members
10 Members
13 Members
12 Members
14 Members
0
2000
4000
6000
8000
10000
12000
14000
Count
Region
NORTHEAST
MIDWEST
SOUTH
WEST


Distribution of Household Type by Region (GEREG)
Household Type
Husband/wife primary family (neither AF)
Non-interview household
Civilian female primary individual
Unmarried civ. female-prim fam hhlder
Unmarried civilian male-prim. fam hhlder
Civilian male primary individual
Husb/wife prim. family (either/both AF)
Primary individual hhld-RP in AF
Primary family hhlder-RP in AF, unmar.
Group quarters without family
Group quarters with family
0
5000
10000
15000
20000
25000
Count
Region
NORTHEAST
MIDWEST
SOUTH
WEST

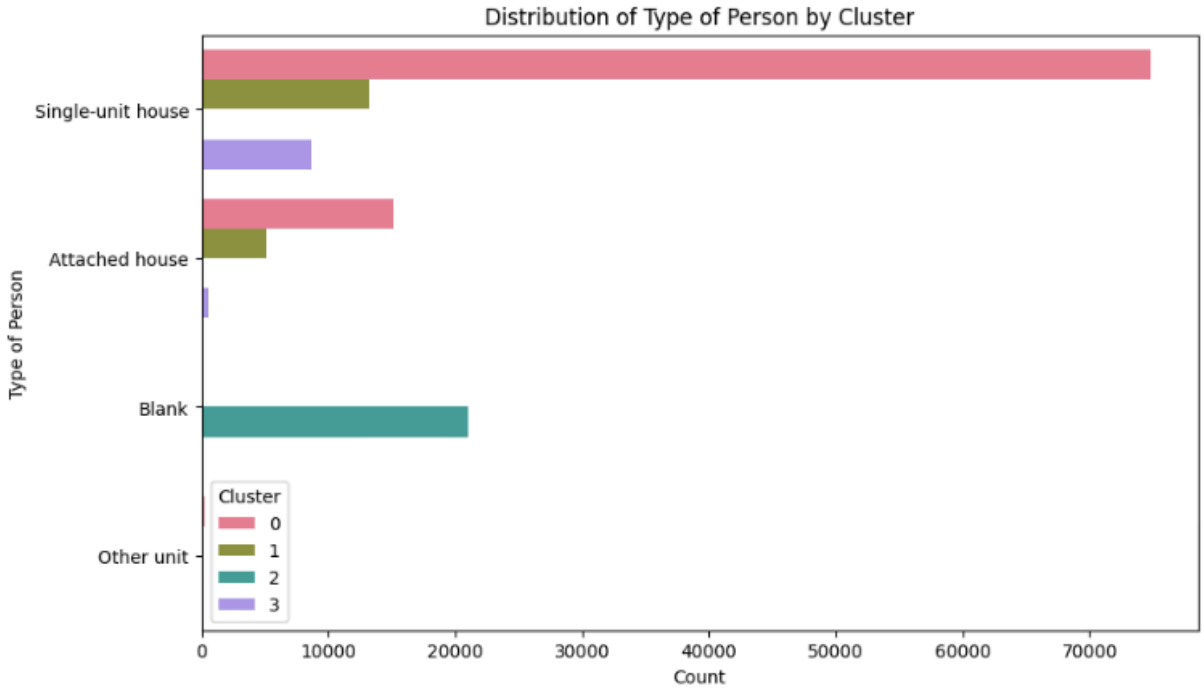

Distribution of Does Anyone in This Household Have a Business or a Farm? by Region (GEREG)
Does Anyone in This Household Have a Business or a Farm?
No
Blank
Yes
Refused
Don't know
0
5000
10000
15000
20000
25000
30000
35000
40000
Count
Region
NORTHEAST
MIDWEST
SOUTH
WEST

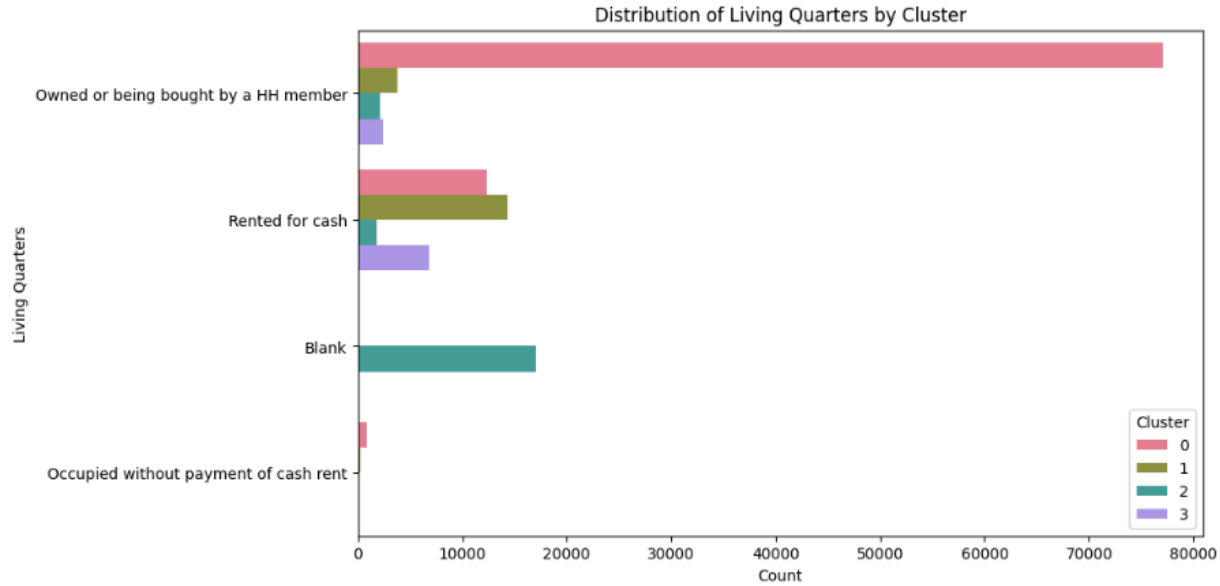

Distribution of Work 35 Hours or More Per Week by Region (GEREG)
Work 35 Hours or More Per Week
Blank
Yes
No
Hours vary
0
10000
20000
30000
40000
50000
Count
Region
NORTHEAST
MIDWEST
SOUTH
WEST

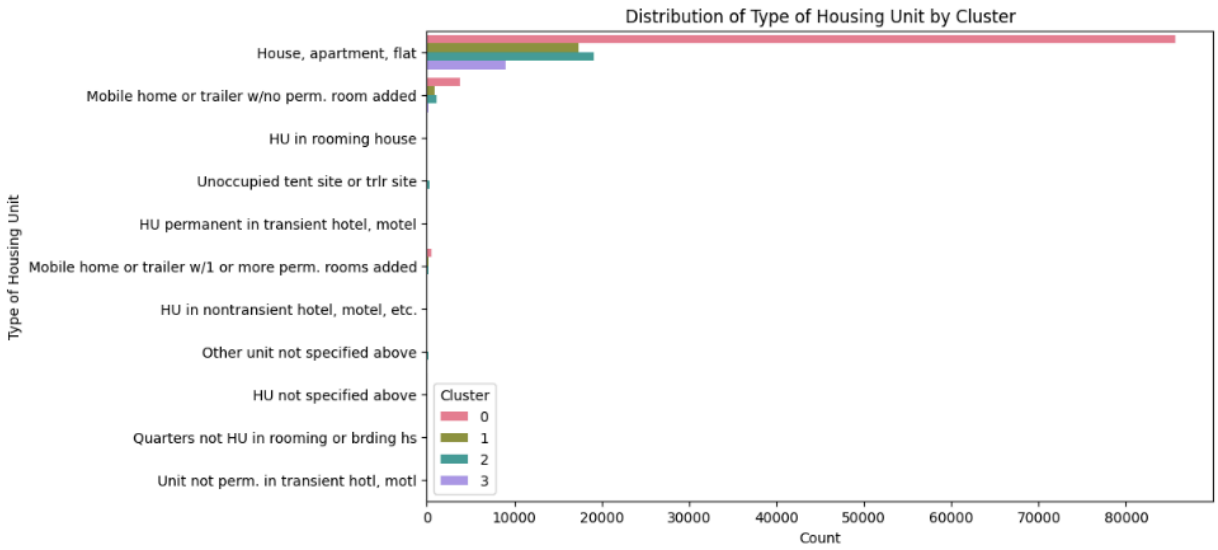

Distribution of Major Occupation Categories by Region (GEREG)
Major Occupation Categories
Construction, and maintenance occupations
Blank
Sales and office occupations
Management, professional, and related occupations
Service occupations
Production, transportation, and material moving occupations
Farming, fishing, and forestry occupations
Armed Forces
0
5000
10000
15000
20000
25000
30000
Count
Region
NORTHEAST
MIDWEST
SOUTH
WEST

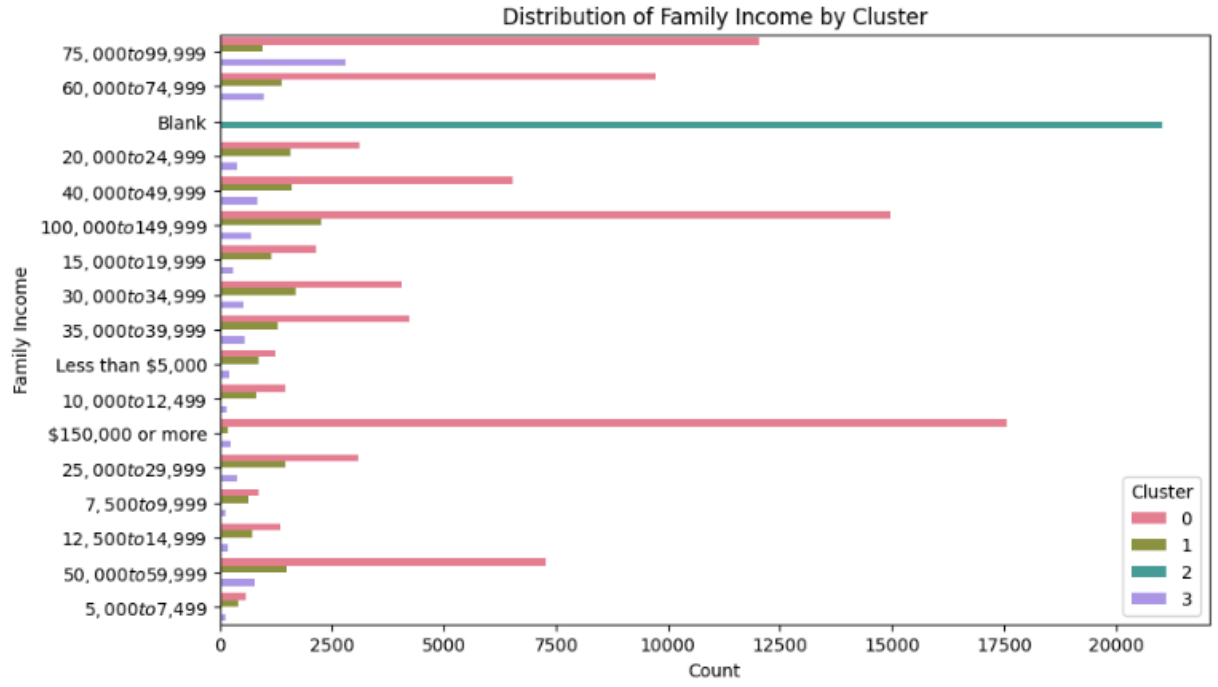

Distribution of The food that (I/we) bought just didn't last by Region (GEREG)
The food that (I/we) bought just didn't last
Never true
Sometimes true
Often true
Don't Know
Refused
No Response
0
2000
4000
6000
8000
Count
Region
NORTHEAST
MIDWEST
SOUTH
WEST

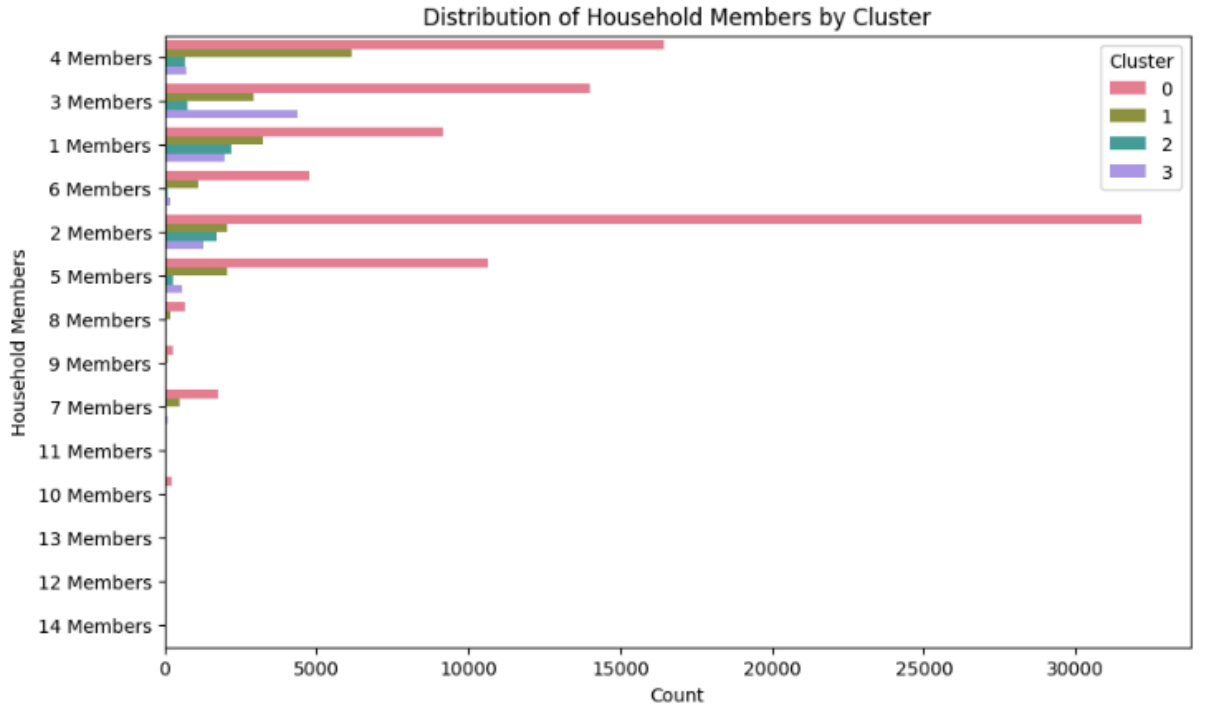
[H]

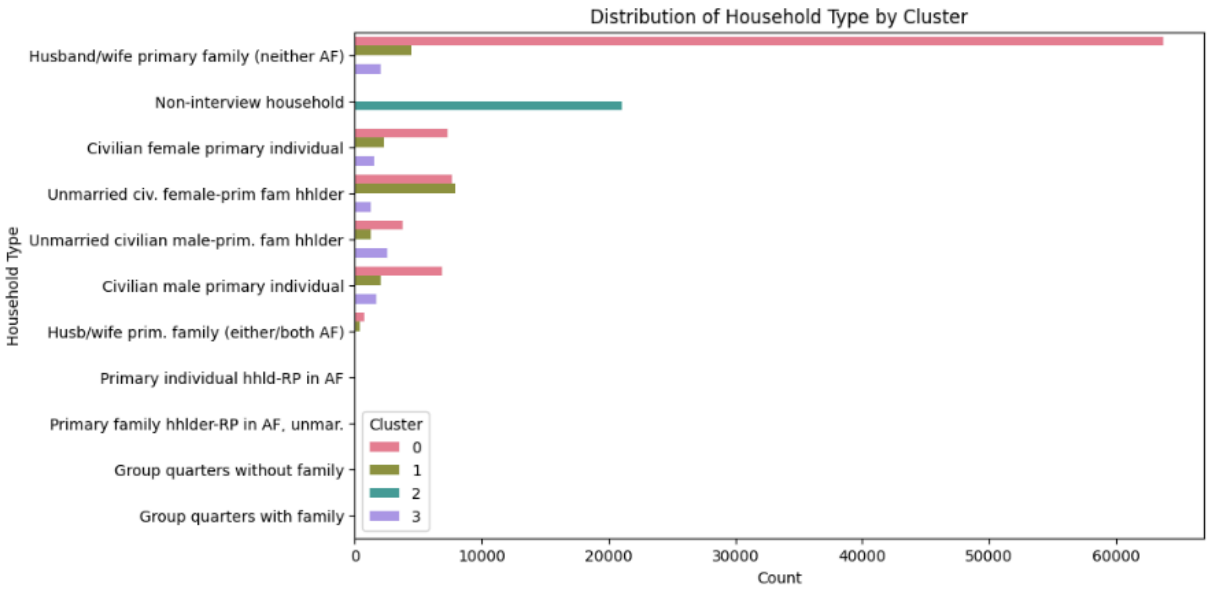

Distribution of Type of Person by Cluster
Type of Person
Single-unit house
Attached house
Blank
Other unit
Cluster
0
1
2
3
0
10000
20000
30000
40000
50000
60000
70000
Count

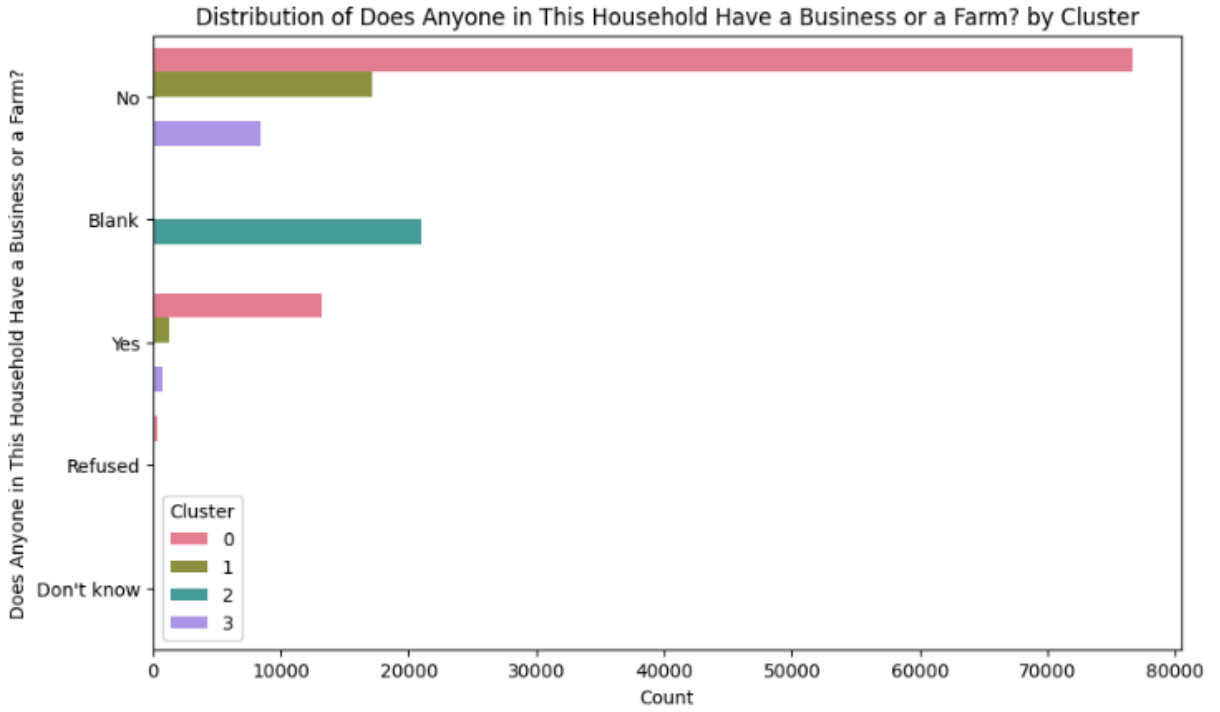

Distribution of Living Quarters by Cluster
Living Quarters
Owned or being bought by a HH member
Rented for cash
Blank
Occupied without payment of cash rent
Cluster
0
1
2
3
0
10000
20000
30000
40000
50000
60000
70000
80000
Count

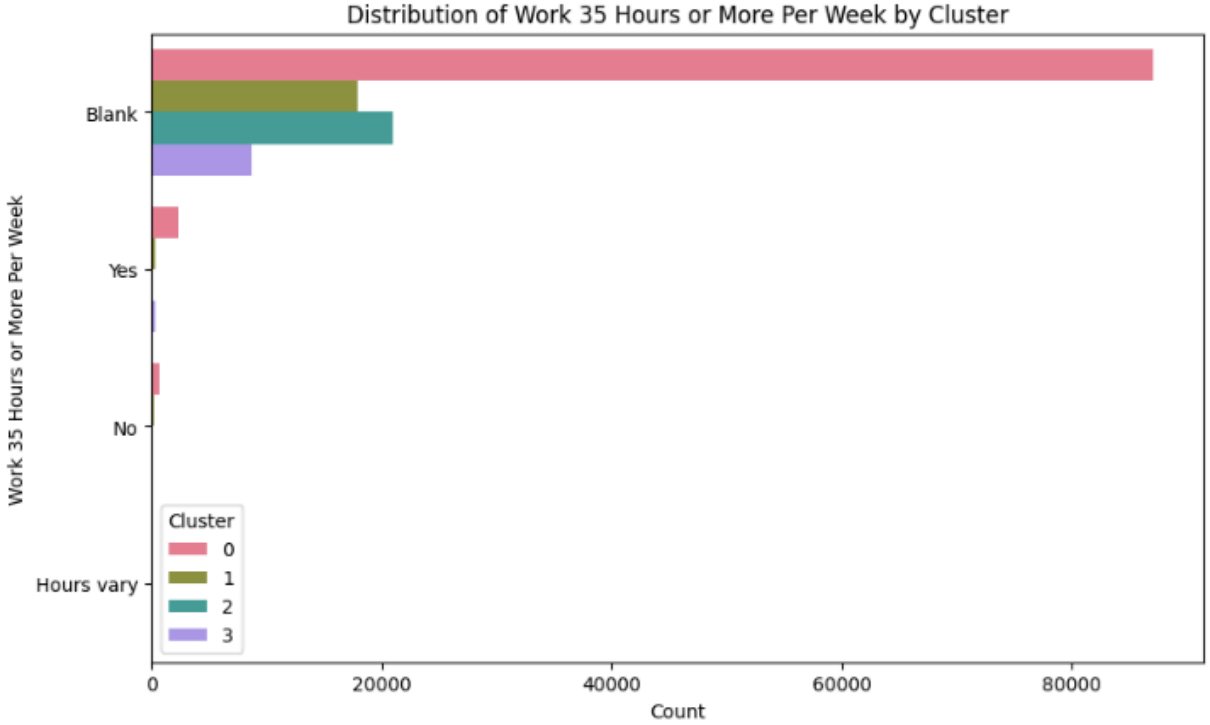

Distribution of Type of Housing Unit by Cluster
Type of Housing Unit
House, apartment, flat
Mobile home or trailer w/no perm. room added
HU in rooming house
Unoccupied tent site or trlr site
HU permanent in transient hotel, motel
Mobile home or trailer w/1 or more perm. rooms added
HU in nontransient hotel, motel, etc.
Other unit not specified above
HU not specified above
Quarters not HU in rooming or brding hs
Unit not perm. in transient hotl, motl
Cluster
0
1
2
3
0
10000
20000
30000
40000
50000
60000
70000
80000
Count

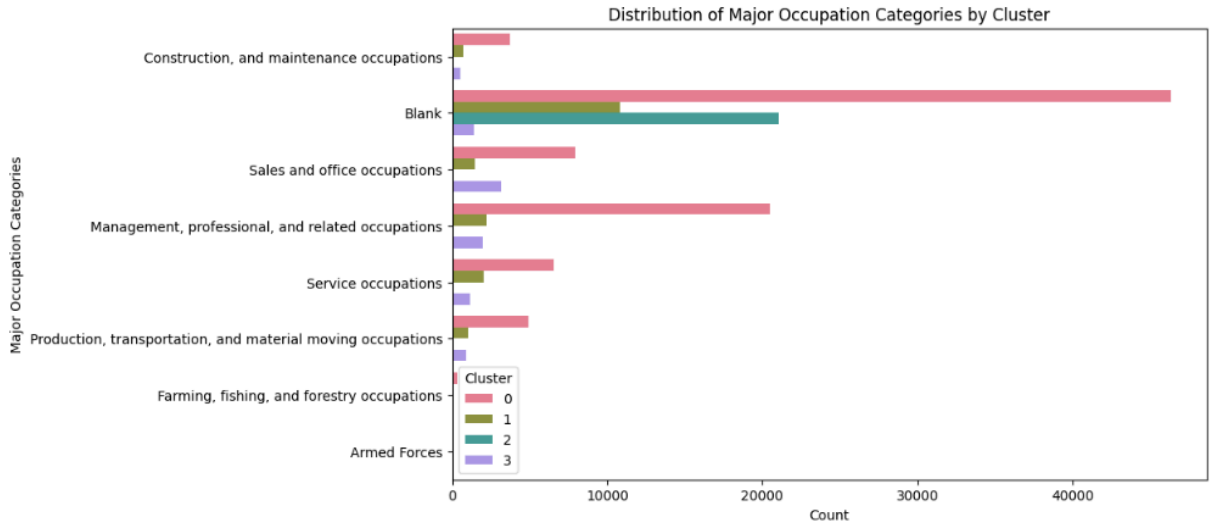

Distribution of Family Income by Cluster
Family Income
75,000to99,999
60,000to74,999
Blank
20,000to24,999
40,000to49,999
100,000to149,999
15,000to19,999
30,000to34,999
35,000to39,999
Less than $5,000
10,000to12,499
$150,000 or more
25,000to29,999
7,500to9,999
12,500to14,999
50,000to59,999
5,000to7,499
Cluster
0
1
2
3
0
2500
5000
7500
10000
12500
15000
17500
20000
Count

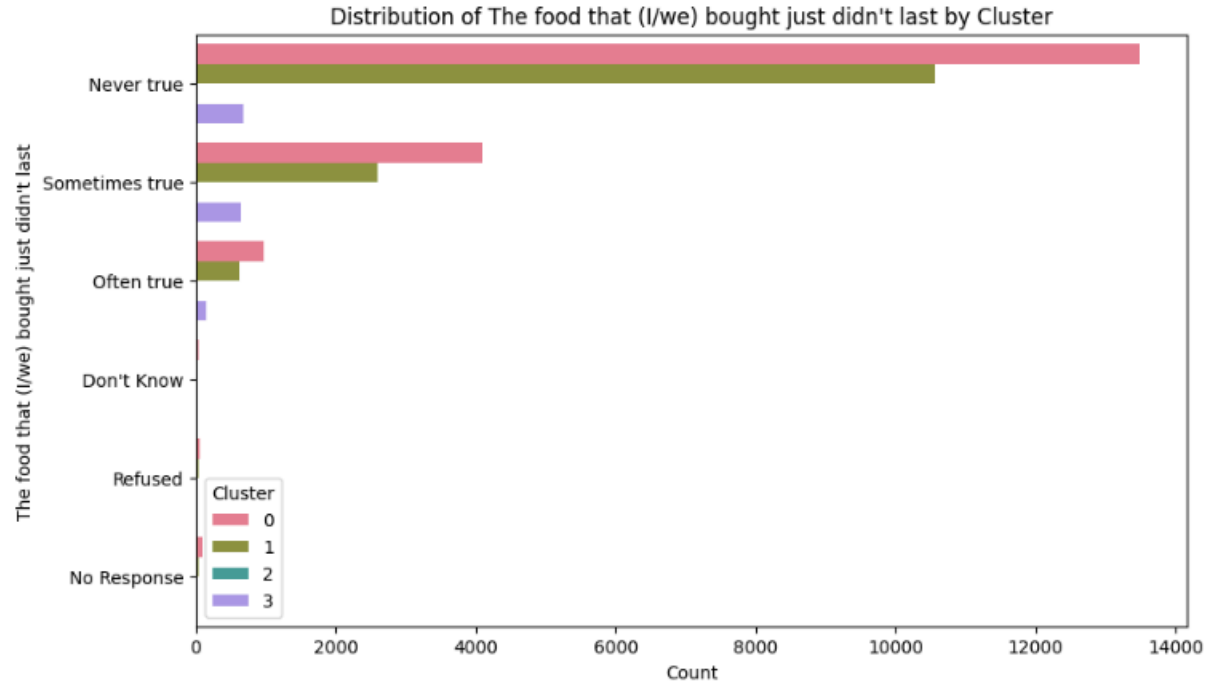

Distribution of Household Members by Cluster
Household Members
4 Members
3 Members
1 Members
6 Members
2 Members
5 Members
8 Members
9 Members
7 Members
11 Members
10 Members
13 Members
12 Members
14 Members
Cluster
0
1
2
3
0
5000
10000
15000
20000
25000
30000
Count


Distribution of Household Type by Cluster
Household Type
Husband/wife primary family (neither AF)
Non-interview household
Civilian female primary individual
Unmarried civ. female-prim fam hhlder
Unmarried civilian male-prim. fam hhlder
Civilian male primary individual
Husb/wife prim. family (either/both AF)
Primary individual hhld-RP in AF
Primary family hhlder-RP in AF, unmar.
Group quarters without family
Group quarters with family
Cluster
0
1
2
3
0
10000
20000
30000
40000
50000
60000
Count

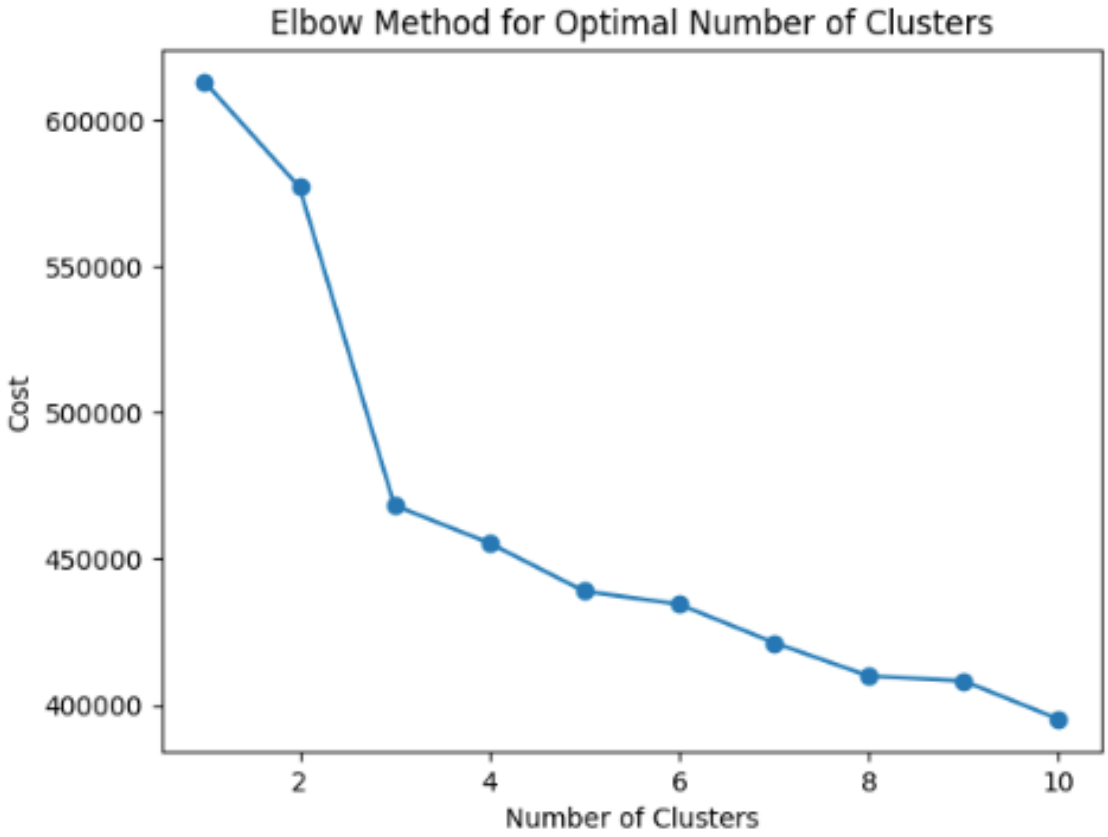

Distribution of Does Anyone in This Household Have a Business or a Farm? by Cluster
Does Anyone in This Household Have a Business or a Farm?
No
Blank
Yes
Refused
Don't know
Cluster
0
1
2
3
0
10000
20000
30000
40000
50000
60000
70000
80000
Count


Distribution of Work 35 Hours or More Per Week by Cluster
Work 35 Hours or More Per Week
Blank
Yes
No
Hours vary
Cluster
0
1
2
3
0
20000
40000
60000
80000
Count


Distribution of Major Occupation Categories by Cluster
Major Occupation Categories
Construction, and maintenance occupations
Blank
Sales and office occupations
Management, professional, and related occupations
Service occupations
Production, transportation, and material moving occupations
Farming, fishing, and forestry occupations
Armed Forces
Cluster
0
1
2
3
0
10000
20000
30000
40000
Count


Distribution of The food that (I/we) bought just didn't last by Cluster
The food that (I/we) bought just didn't last
Never true
Sometimes true
Often true
Don't Know
Refused
No Response
Cluster
0
1
2
3
0
2000
4000
6000
8000
10000
12000
14000
Count

[I]


Elbow Method for Optimal Number of Clusters
Cost
600000
550000
500000
450000
400000
2
4
6
8
10
Number of Clusters