Identifying Suspicious Records in Household Data

A dissertation submitted for the Degree of Master of Computer Science

Miss.L.Namasivayam
University of ColomboSchool of Computing
2018

Identifying Suspicious Records in Household Data

Miss. L.Namasivayam
2018

Declaration

The thesis is my original work and has not been submitted previously for a degree at this or any other university/institute.

To the best of my knowledge it does not contain any material published or written by another person, except as acknowledged in the text.

Student Name: L. Namasivayam

Registration Number: 2014/mcs/051

Index Number: 14440514

_____

Signature: Date:

This is to certify that this thesis is based on the work of

Mr./Ms. L.Namasivayam

Under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by:

Supervisor Name: Dr. Rasika Dayarathna

_____

Signature: Date:

Acknowledgements

Firstly, I would like to express my endless gratitude to Dr Rasika Dayarathna, my supervisor, who has been helping me all the way along to achieve my degree. This thesis and research project would not have been possible without his professional guidance and consistent encouragement at each stage. His comments and advice are always inspiring when I face unexpected challenges in my study.

I would like to thank Mr. Sampath Deegalla, Senior lecturer at University of Peradeniya, who has been helping to finish my degree with his valuable advices and comments.

I would like to sincerely thank all the lecturers at the University Of Colombo School Of Computing for their valuable advices and comments given at various stages of this research. Without your support, I could not have completed this research with success.

I would like to thank to Director General and ICT staffs in department of Census and statistics, who has been helping to collect the data as much as possible.

I like to thank all my dear friends who were there around me, with best of their encouragement, suggestions and support throughout this research.

Finally, I would like to express my heartfelt thanks towards my lovely husband and my family for their support and encouragement through the many days and nights dedicated to the completion of this research.

Abstract

With growing advancement in the financial field, fraud is spreading all over the world, causing major financial losses. This thesis represents methods to detect suspicious records in household income and expenditure survey gathered by the census and statistics department in Sri Lanka. K-means, hierarchical clustering method and model base clustering method were used in this research to find the suspicious records in the in household income and expenditure dataset. Some statistical methods were also used to clarify the suspicious records. In the dataset when the adhoc expenses such as weddings /funerals for family members, Social activities / Ceremonies, gift, donation and purchased properties house exceed the limit, income is higher than the expenditure without proof those are considered as a suspicious record.

Table of Contents

List of Table

List of Figure

List of Abbreviations

DCS	Department of Census and Statistics
HIES	Household Income and Expenditure
PSU	Primary Sampling Units
SVM	Support Vector Machine
RBF	Radial Basis Function
K-NN	K-Nearest Neighbor
SAS	Statistical Analysis Software

Chapter 1

1 Introduction

The Department of Census and Statistics (DCS) conducts the Household Income and Expenditure Survey (HIES) under the National Household Survey Programme. The HIES had been conducted in combination with Labour Force Survey named as Labour Force and Socio-Economic Survey till 1990. DCS first initiated the HIES as a separate survey in 1990 and since then it has been continued once in every five years till 2006/07. In response to the rapidly changing economic conditions the DCS decided to conduct the HIES once in every three years starting from 2009/10 which enabled to monitor the income and spending patterns in the country far more frequently.

Generally the HIES is conducted over a period of 12 consecutive months to capture seasonal variations of income and expenditure patterns in Sri Lanka. The HIES 2012/13 is the eighth in its series. The field work of this survey was carried out during the period from July 2012 to June 2013. Household information is covering the demography, school education, health information, how the people spend money for food and nonfood, income and housing information. According to the information DCS is measure levels and changes in living conditions of the people, observe the consumption patterns and compute various other human development and socio economic indicators such as poverty, price indices etc.

Suspicious records are the records that are highly different from the normal records. Some people give the wrong information to the survey. Most of the people they don’t like to give the additional income. But here the Household Income and Expenditure questionnaire contains separate questions for income and expenditure. It means Income is ask like Income from agricultural activities, Income from Non - Agricultural activities, Income from other cash receipt during last calendar month and the Expenditure is like Household expenditure on Housing , Fuel & Light, Non- durable goods , Services & Consumer durable for main Household ,Weekly Consumptions on Food & Drink. We can calculate total income and expenditure if the difference is high it may be suspicious records.

Banks have systems to identify certain types of suspicious transactions. An individual opening an account with an unlikely bulk cash amount, especially if the person does not have a valid reason for opening the account or if their information is vague, might be identified as a suspicious activity. Another scenario would be if an individual withdraws / deposits a large amount of money. Transactions conducted with a relatively small amount but with high frequency can also be taken as a suspicious transaction. If they deposit more than one million rupees, the bank will check with their income and decide if this is suspicious or not. Abnormal transactions can be detected based on the customer's past behavior and this is done by comparing the old transactions with the recent ones. Unreasonable behaviors of the relevant customer when conducting a transaction (nervous, rushed, unconfident, etc.) and customers giving wrong information with respect to his/her identity, sources of income or businesses gives the impression of a probable fraudulent activity.

There are other suspicious financial transactions such as providing forged bill. Considering a real scenario, an official of a divisional secretariat in Gampaha district had refused to pay for a forged bill of Rs.800 to buy a kilo of tomatoes from a cooperative store that prepares relief food for flood victims, when the actual price of a kilo of tomatoes was only Rs.200. Another good example is some airline ticket prices being completely different for the same distance and location when compared to other airlines. If that difference is more than a reasonable amount, it can be a probable fraud.

Main goal in this research was to analyze data using data mining techniques to identify the suspicious records in the household data.

1.1 Statement of the problem

The Department of Census and Statistics gives accurate results based on collected data to the country. In the Household income and expenditure survey reveals that an average household received Rs. 45,878 per month at national level in 2012/13 in Sri Lanka, mean of income for every district, average expenditure per month and average monthly household expenditure by food and non food items. With the help of the survey results, the government administrative takes many decisions to develop the country. The department conducts surveys to collect the relevant information from people. Sometimes some people give incorrect information. When they analyze the data with wrong information they won't be able to give accurate results. To avoid this problem, incorrect information needs to be identified from collected information. In this research a suitable method is going to be proposed to identify the suspicious records.

The department of census and statistics measure levels and changes in living conditions of the people, observe the consumption patterns and compute various other human development and socio economic indicators such as poverty, price indices etc.

1.2 Objectives

The objective of this project is to identify the suspicious records in the Household income and expenditure dataset using clustering methods.

1.3 Scope

The scope of this project is to identify the suspicious records with the help of household income and expenditure real data set from the census and statistics.

1.4 Thesis Summary

There are six chapters in this thesis: Chapter 1 introduces the research and states the problem. Chapter 2 discusses the related work, supervised and unsupervised methods of fraud detection. The chapter 3 briefs about the data collection methods, collected data set, tool selection for running the algorithms and clustering algorithms. Chapter 4 contains evaluation and results that are obtained from this research. Chapter 5 discusses the conclusion and the future work.

Chapter 2

2 Literature Review

Fraudulent transactions are infrequent in nature making its detection quite challenging. They represent a very small fraction of activity within an organization. Another significant challenge is that a small percentage of activity can have a huge adverse effect on the financial wellbeing of the organization. The presence of the right tools and systems would help prevent this effect to a huge extent. Criminals are crafty. The fraud schemes used keep varying with time, each more advanced and hard to detect. However, technology makes it feasible to enhance modern detection methods in order to cater to newly developed fraudulent techniques.

Most organizations still use rule-based systems as their primary tool to detect fraud. Though it's true that rules be very helpful while revealing known patterns they aren't very useful in detecting unpredicted pattern-free methods or handling complicated new-age techniques. This is where machine learning becomes necessary for fraud detection [3].

Simply put, machine learning automates the extraction of known and unknown patterns from data. It expresses those patterns as either a formula or instruction set that can be applied to new and unseen data. The machine is fortunately, flexible and adaptive to new changes and well suited for even independent detection with no human interference. Machine learning techniques are categorized as supervised and unsupervised to detect fraud.

2.1 Related work

Several techniques have been used for the detection of fraud. Bank fraud can be analyzed using hybridization of two support vector machine classes namely, binary and single class. Bank fraud can Fraud by credit cards, Money laundering and Mortgage fraud. Credit card fraud merely requires some information, to make an online purchase using the target victim's card. From the historical database the attributes such as the frequency of use, shopping frequency, daily transactions and average number of consumption are calculated for every record in order to discover a pattern for the above mentioned fraudulent activity. For the detection of suspicious money laundering activities, the bank should come up with a threshold which is a predefined number beyond which it is considered an alleged case of fraudulence. Also an abnormal amount of cash being transacted, an alien source of transfer or a suspicious change of residence can be considered indices to detect money laundering. The indices used for detection of mortgage fraud are personal and professional information of customers as well as the presented mortgage.

This fraud detection system takes bank database and performs learning for extracting a model which represents the characteristics of the data. The model is used to evaluate new transactions. A transaction accepted by the model is added to the database to improve the model. Transactions rejected by the model (Suspicious) pass on to a manual check. If they are considered normal, they are executed and then added to the database otherwise the transaction is rejected (Djeffal Abdelhamid, 2014).

Credit card fraud can be defined as a wide-ranging term for theft & fraud committed using or involving a payment card, such as a credit card or debit card, are fraudulent source of funds in various kind of transaction. The purpose may be to obtain goodies without paying, or to obtain unauthorized funds from an account or to avail some kind of service. Decision tree, Genetic algorithm, Meta learning strategy, neural network, HMM are the presented methods used to detect credit card frauds. In contemplate system for fraudulent detection, artificial intelligence concept of Support Vector Machine (SVM) & decision tree is being used to solve the problem (Vijayshree B. &Nipane 2016).

Novel learning method, applying it on different data such as diabetes data, heart Data, satellite Data and Shuttle data which have two or multi class. These dataset classified using support vector machine with RBF kernel, Rule base classifier, K-NN classifier, Rule base classifier

with discretization and LTF classifier. Using support vector machine with RBF kernel is giving the best accuracy than others (Durgesh K. & Srivastava 2009).

Put a light on performance evaluation based on the correct and incorrect instances of data classification using Naïve Bayes and J48 classification algorithm. Naive Bayes algorithm is based on probability and j48 algorithm is based on decision tree. The paper sets out to make comparative evaluation of classifiers Naïve Bayes and J48 in the context of bank dataset to maximize true positive rate and minimize false positive rate of defaulters rather than achieving only higher classification accuracy using WEKA tool (Tina R. &Patil 2013).

The system is generate fraud transactions with a given sample dataset. If genetic algorithm is applied in bank for credit card fraud detection, the chance of fraud transactions can be predicted soon after credit card transactions is in process. Anti-fraud strategies are adopted to prevent banks from great losses before the transaction and thereby reduce risks (R. D. Patel & D. K. Singh) [7].

According to Vaishali (2014), Clustering approach is used for credit card fraud detection. Data is generated randomly for credit card and then K-means clustering algorithm is used for detecting whether the transaction is fraud or legitimate. Clusters are formed to detect fraud in credit card transaction which are low, high, risky and high risky.

Data is generated randomly using Microsoft SQL Server Management Studio. Then K-means clustering algorithm is applied to detect fraud using Visual Studio 2012 software. The data table includes transaction ID, transaction amount, transaction country, transaction date, credit card number, merchant category id, cluster id, "is fraud" and new transaction. K- Means clustering algorithm is applied on this data using .NET language on Visual Studio 2012. Four clusters are formed i.e. low, high, risky and high risky to detect the transaction whether it is fraud transaction or legitimate transaction. K-means clustering algorithm is a simple and efficient algorithm for credit card fraud detection.

Andrei Sorin Sabau (2012) said among partitioned clustering techniques, k-means clustering and its variants with Euclidian distance as dissimilarity metric are the most common used ones. Hierarchical clustering techniques come in second place being used in one quarter of the surveyed papers.

In the real-world application it is very difficult predict the number of clusters for the unknown domain data set. If the fixed number of cluster is very small then there is a chance of putting

dissimilar objects into same group and suppose the number of fixed cluster is large then the more similar objects will be put into different groups (Ahamed Shafeeq B M. & Hareesha K S, 2012) .

The proposed solution uses dynamic clustering of data with modified k-means algorithm. The proposed method works for both the cases i.e. for known number of clusters in advance as well as unknown number of clusters. The algorithm takes number of clusters (K) as the input from the user and the user has to mention whether the number of clusters is fixed or not. If the number of clusters fixed then it works same as K-means algorithm. Suppose the number of clusters is not fixed then the user has to give least possible number of clusters as an input. The K-means procedure is repeated by incrementing the number of clusters by one in each iteration until it reaches the cluster quality validity threshold.

The algorithm is developed and tested for efficiency of different data points in C language. The algorithm takes more computational time compared to the K-means algorithm for large dataset in some cases. The algorithm works same as K-means for the fixed number of clusters. For the unknown data set it starts with the minimum number of cluster given by the user and after the completion of every set of iteration, the algorithm checks for efficiency and it repeats by incrementing the number of cluster by 1 until it reaches the termination condition. The main drawback of the proposed approach is that it takes more computational time than the K-means for larger data sets.

2.2 Chapter summary

This chapter describes the supervised, unsupervised methods and classification algorithms, clustering methods. The related work that the researchers done about fraud detection. The following chapter will cover the methodology.

Chapter 3

3 Methodology

This chapter described the how the data set is collected and which tool is selected for analyze the data and the analyzed algorithms.

3.1 Data collection methods

Data collection is a process of collecting information from all the relevant sources to find answers to the research problem, test the hypothesis and evaluate the outcomes. Data collection methods can be divided into two categories namely the secondary methods of data collection and primary methods of data collection.

Secondary data is a type of data that has already been published in books, newspapers, magazines, journals, online portals etc. Primary data collection methods can be divided into two groups: quantitative and qualitative. Quantitative data collection methods are based on mathematical calculations of various formats. Methods of quantitative data collection and analysis include questionnaires with closed-ended questions, methods of correlation and regression, mean, mode, median etc. Qualitative research methods, on the contrary, do not involved numbers or mathematical calculations. Qualitative research is closely associated with words, sounds, feeling, emotions, colors and other elements that are non-quantifiable. [4]

The Department of census and statistics (DCS) collect the data by using quantitative data collection method. Based on the questionnaires the staff members visit the houses in Sri Lanka and ask the questions and fill the questionnaire. The gathered data being confidential is not published in the website. After the data is analyzed it is published as a final report. The department only gives the 25% of the data for external use. This data is usually given in order to help a number of students in need of it for their researches.

The Department of census and statistics (DCS) has a record of the accurate statistic results of the country. They conduct surveys and derive at the outcomes of domains such as population census, labor force survey, Household income and expenditure, tourism etc. For this research the data obtained via the household income and expenditure survey 2012/2013 is gathered from the department since it has financial records. The Sri Lanka Household Income and Expenditure Survey (HIES) is conducted by the Department of Census and Statistics under

the National Household Sample Survey Program. This survey provides information on household income and expenditure to measure the levels and changes in living conditions of the people. Data collected from this survey is used to observe the consumption patterns to compute various other socioeconomic indicators such poverty, price indices etc. The DCS, with the help of the rapid developments in the ICT, decided to conduct the HIES once in every three years in Sri Lanka. The HIES 2012/13 is the eighth survey in the HIES series. The field data collection of the survey was done within a period of twelve months starting from July 2012 to June 2013 and successfully covered all the districts. The HIES questionnaire consists of nine sections to collect household information covering the following areas:

Sections	Details
Demography	Name, sex, age, marital status, religion, main occupation, etc
School education	Name, school education, school type, grade in this year, Time taken to school for travelling, etc
Health	Did you visit Government hospital or Medical / Health Center, For what kind of treatment , etc
Food and non-food expenditure	Cereals, Vegetable, Meat, Fish, Dried Fish, Eggs, Sugar, Jaggery, Treacle, Drugs and Tobacco, Water bills, Fuel & Light (Average per month), Communication etc
Income	Wages / Salaries, Bonus, Arrears Payment, Income from agricultural activities, Income from Non - Agricultural activities, etc
Inventory of durable goods	Household Equipment, Agricultural Equipment, Banks (Government / Private)
Access to facilities in the area and debts of the households	Place of facilities and Distance from your house to the closest facility (Km)
Housing Information	Type of Structure, Number of bed rooms, Total floor area (Sq. feet), Main source of drinking water
Agriculture holdings and Livestock	Land ownership, Livestock (owned),

Table 1: Questionnaire sections and details

3.1.1 Data set

The department collects the data and enters those data into data entry form using CSPro software. It saves as a text file (CSV). In this dataset some sections are based on person wise such as demography, school and health are available on a per-person basis while details such as expenditure are based on a per-family basis. Since the research is based on financial records, financial data is considered to a larger extend.

In order to make use of this data set, the CSV format is converted into a Microsoft access file. For that each section converted into access table and then join as a one table for the analyze purpose. Note that all details are calculated for every family on a monthly basis. The field with the total expenditure is adjoined with the fields with the total income details. The Electricity bill and water bill are collected from Food and non-food expenditure sections. Since the Food expenditure is given in a weekly manner its per-month value is calculated and then added to the data base. Furthermore, fields such as total communication expenditure, other adhoc expenses, Household Services (Laundry Charges, grinding charges) and other individual expenditure (Provident fund /W. & O. P. fund, Insurance / Agrahara) are added to the table.

The total communication expenditure, household services and other individual expenditure is considered as well. The other adhoc expense is considered a one-time occurrence, hence being retained as a yearly event. After these manipulations, the generalized table is converted into excel. This data set contains 20280 records and 15 attributes. The attributes are district, ds division, sector, PSU, A0, serial number as ID fields and Household income per month, household expenditure per month, electricity bill, water bill, food expenditure, communication expenditure, adhoc expenditure, Household Services and other individual expenditure.

Fields	Description
District	District of the person district code 11 -92
Sector	Sector of the person
DS Division	DS division for the person
PSU	Primary sampling units (PSUs) are the census blocks selected for the survey.
A0	Identification Number
Serial Number	Household number It contains 1-10
Household Income	Monthly income for the family
Household expenditure	Monthly expenditure for the family
Electricity bill	Monthly electricity bill
Water bill	Monthly water bill
Food expenditure	Total food expenditure for family contains Cereals, Prepared food(bread), Vegetable, Meat, Fish, Dried fish, Eggs, Fats and oil, Sugar , Juggery & Treacle, Fruit
Communication expenditure	Monthly communication expenditure - Postal & Telegraph charges, Telephone charges (Domestic), E-mail / Internet charge.
Adhoc expenditure	Rarely expenditure for yearly- Expenditure on weddings /funerals for family members, Social activities / Ceremonies, Gift, Donation, Similar transfers, Purchased properties House
Household Services	Monthly household service - Laundry Charges, Grinding Charges, Charges for Drivers, Wagers to Servants /Chauffeurs
Other individual expenses	Monthly other expenditure - Insurance / Agrahara, Payment for debits, Provident fund /W. & O. P. fund.

Table 2: Selected field for analyze

3.2 Data Analysis

3.2.1 Tool selection

There are a lot of tools used for analyzing data in the data mining field. A significant share of people, who crunch numbers for a living, use Microsoft Excel or other spreadsheet programs like Google Sheets. Others use proprietary statistical software like SAS, Stata, or SPSS that they often first learned in school.

While Excel and SAS are powerful tools, they have serious limitations. Excel cannot handle datasets above a certain size, and does not easily allow for reproducing previously conducted analyses on new datasets. The main weakness of programs like SAS are that they were developed for very specific uses, and do not have a large community of contributors constantly adding new tools.

R and Python are the two most popular programming languages used by data analysts and data scientists. Both are free, open source, and were developed in the early 1990s—R for statistical analysis and Python as a general-purpose programming language. Python is better for data manipulation and repeated tasks, while R is good for ad hoc analysis and exploring datasets. R is also convenient for statistics-heavy projects and one-time dives into a dataset.

3.2.2 Algorithms used for supervised and unsupervised methods

Suspicious records can be identified by using supervised or unsupervised methods. First the plan was to proceed using the supervised method. If the suspicious records are identified from the data set by experts, it can be checked if the data is classified correctly using this algorithm in the supervised method. The German credit card sample dataset was obtained for this research. This dataset is used to judge a loan applicant by classifying the applicant into good or bad. The dataset has 1000 records. The attributes are Account Balance, Duration of Credit, Payment Status of Previous Credit, Purpose, Credit Amount and installment per cent. 67% of data is defined as trained data while the remaining 33% of data is considered as test data. This classification is done using Support vector machine, Naive Bayes theorem and Decision tree in r studio to obtain the accuracy. The real data set is obtained from the census department. Unfortunately there was no expert indication to identify the suspicious records in the data set. Hence, the unsupervised clustering methods were used to proceed further.

3.2.3 Algorithm selected for the analysis

Since the data set does not contain labels, we use clustering techniques. This research uses K-means, hierarchical clustering methods and model base clustering for data analysis.

K-Means

The K-means clustering algorithm uses iterative refinement to produce a final result. The algorithm inputs are the number of clusters K and the data set. The data set is a collection of features for each data point. The algorithms start with initial estimates for the K centroids, which can either be randomly generated or randomly selected from the data set. The algorithm then iterates between two steps as data assignment step and centroid update step. In the data assignment step each data point is assigned to its nearest centroid, based on the squared Euclidean distance. And another step is Centroid update this is done by taking the mean of all data points assigned to that centroid's cluster.

Here the cluster number is two so the data assigned to those two clusters.

```
> k.means.fit$centers
  DISTRICT   SECTOR DS      PSU       AO  SNUMBER    HIncPM    HExpPM Electricity_bill Water_bill
1 29.95588 1.573529 10 81.75000 39319.19 5.308824 252912.75 355005.21        2044.6029   448.5000
2 40.91327 1.840936 10 68.10375 49990.25 5.466802  44623.77  40466.94         750.8228   158.7814
  Food_Value Communication_Expen wedding_Expen OtherIndi_Expen Indi_Expen
1   5735.029           3087.6176    2833810.29       39824.824  1216.5294
2   3898.369            873.5224      28479.25        4368.651   618.2361
```

Figure 1 K-Means centers

The above image shows that in household income and expenditure dataset creates the center points for each cluster and each attribute.

```
> k.means.fit$cluster
   [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
  [52] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [103] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [154] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [205] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [256] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [307] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [358] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [409] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [460] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
```

Figure 2 Clusters

And then the each data goes to each cluster. Here the first data goes to second cluster likewise whole dataset is assigned to one of the two clusters.

This algorithm is guaranteed to converge to a result. The result may be a local optimum (i.e. not necessarily the best possible outcome), meaning that assessing more than one run of the algorithm with randomized starting centroids may give a better outcome.

Hierarchical clustering

K means clustering requires us to specify the number of clusters, and finding the optimal number of clusters can often be hard. Hierarchical clustering is an alternative approach which builds a hierarchy from the bottom-up, and doesn't require us to specify the number of clusters beforehand. The algorithm works as put each data point in its own cluster then identify the closest two clusters and combine them into one cluster and repeat till all the data points are in a single cluster. Once this is done, it is usually represented by a dendrogram like structure. [5]

The dendrogram is a visual representation of the compound correlation data. The individual compounds are arranged along the bottom of the dendrogram and referred to as leaf nodes. Compound clusters are formed by joining individual compounds or existing compound clusters with the join point referred to as a node [6]. At each dendrogram node there is a right and left sub-branch of clustered compounds. The vertical axis is labeled distance and refers to a distance measure between compounds or compound clusters.

Model-based clustering

Both hierarchical clustering and k-means clustering use a heuristic approach to construct clusters, and do not rely on a formal model. Model-based clustering assumes a data model and applies an EM (expectation maximization) algorithm to find the most likely model components and the number of clusters. Significance of statistical distribution of variables in the dataset is the measure. When cluster number is defined as two it will cluster the dataset into two classes.

mc.R | HIES

Source on Save | Run | Source

```
1  library(MASS)
2  library(mclust)
3
4  mc <- Mclust(HIES, G=2)
5  summary(mc)
6  mc
7  mc$classification
8  out <- cbind(HIES,clusterNum = mc$classification)
9  write.table(out,file = "C:/Users/abcd/Desktop/res/Output2.csv", sep =",")
10
11
12
13
```

Figure 3 : Model Base Cluster

Console | Terminal

```
> mc <- Mclust(HIES, G=2)
fitting ...
  |======================================================================| 100%
> summary(mc)
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust VII (spherical, varying volume) model with 2 components:

 log.likelihood     n df       BIC       ICL
      -3579664 20280 35 -7159674 -7159776

Clustering table:
    1     2
17764  2516
> mc
'Mclust' model object: (VII,2)

Available components:
 [1] "call"           "data"           "modelName"      "n"
 [5] "d"              "G"              "BIC"            "bic"
 [9] "loglik"         "df"             "hypvol"         "parameters"
[13] "z"              "classification" "uncertainty"
> mc$classification
   [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1
  [40] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
  [79] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [118] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [157] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1
 [196] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [235] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

Figure 4: Output of Model Base Cluster

B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
DISTRICT	SECTOR	DS	PSU	A0	SNUMBER	HIncPM	HExpPM	Electricity	Water_bil	Food_Exp	Food_Val	Communi	adhoc_Expen	OtherIndi	Indi_Expe	cluster
82	2	10	15	89521	3	163844	11267.38	0	0	4364	1091	400	0	0	30	2
32	2	10	16	42426	5	18906.67	14703.45	65	0	4788	1197	0	104000	0	55	2
82	2	10	31	89937	9	129126.2	65133.71	625	350	10804	2701	500	0	20040	65	2
11	2	10	68	12988	7	156588.6	80481.64	650	110	15872	3968	550	508000	6100	75	2
71	2	10	2	79196	3	11204.83	21072.95	500	0	6508	1627	0	150000	0	80	2
81	2	10	12	86789	5	18668.33	21202.27	75	60	8308	2077	200	100000	50	120	2
82	2	10	31	89937	8	171288.3	25628.77	210	0	9520	2380	300	150000	60	120	2
44	2	10	34	56368	10	154016.7	12660.04	0	0	8760	2190	250	150	0	120	2
33	2	10	31	46721	10	185454.3	14533.12	150	150	8844	2211	400	3000	0	120	2
91	2	10	60	94407	3	186921.4	14287.07	100	0	8088	2022	100	1000	0	130	2
41	2	10	10	48421	2	13183.81	101904.8	400	0	8268	2067	300	1000500	0	135	2
72	2	10	61	83430	10	143070	50397.63	1400	0	30636	7659	500	7000	0	140	2
81	2	10	82	86747	7	13836.91	35289.61	100	180	10572	2643	200	211000	400	150	2
72	2	10	29	83605	7	147453.6	34810.45	676	398	13556	3389	250	0	1887	152	2
82	2	10	31	89937	7	181619	44542.34	650	275	18428	4607	1000	1300	10240	160	2
81	2	10	14	87637	4	158264.9	62804.17	300	150	13900	3475	800	404000	0	160	2
31	2	10	6	39800	4	17923.81	43207.96	360	0	18092	4523	675	120000	0	169	2
45	2	10	31	58435	9	19093.1	57510.55	0	0	6680	1670	0	600000	0	170	2
31	1	10	77	40333	10	17081.9	28618.4	236	537	9208	2302	0	127000	50	173	2
33	2	10	64	45193	10	131421.9	65478.34	0	0	8608	2152	1000	30000	50000	173	2
62	2	10	63	77037	7	148498.6	43749.99	1500	500	18992	4748	2000	26500	150	180	2
22	2	10	51	33370	7	134764.3	29478.1	200	0	7144	1786	1000	20000	1558	182	2
23	3	10	32	36636	2	172092.4	25544	0	0	18968	4742	0	2500	950	184	2
32	2	10	95	42132	3	18262.38	36276.16	179	0	7652	1913	100	123000	5500	185	2

Figure 5: Output data

3.2.4 Statistical method for Analysis

Some statistical methods were used to identify the outliers. Here Benford law, IQR method and descriptive analysis used for statistical analysis.

Benford law

Benford's law (first digit law or Benford's distribution), is a distribution that the first digits of many (but not all) data sets conform to. Benford's law can often be used as an indicator of fraudulent data, and can assist with auditing accounting data [16].

In data sets that obey the law, the number 1 appears as the most significant digit about 30% of the time, while 9 appears as the most significant digit less than 5% of the time.

Use the benford law for Household Income and Expenditure data using R. For that first load the data into R, then install the benford.analysis package to run the benford law. And then use the benford for each attributes for analyze the attributes contain suspicious records.

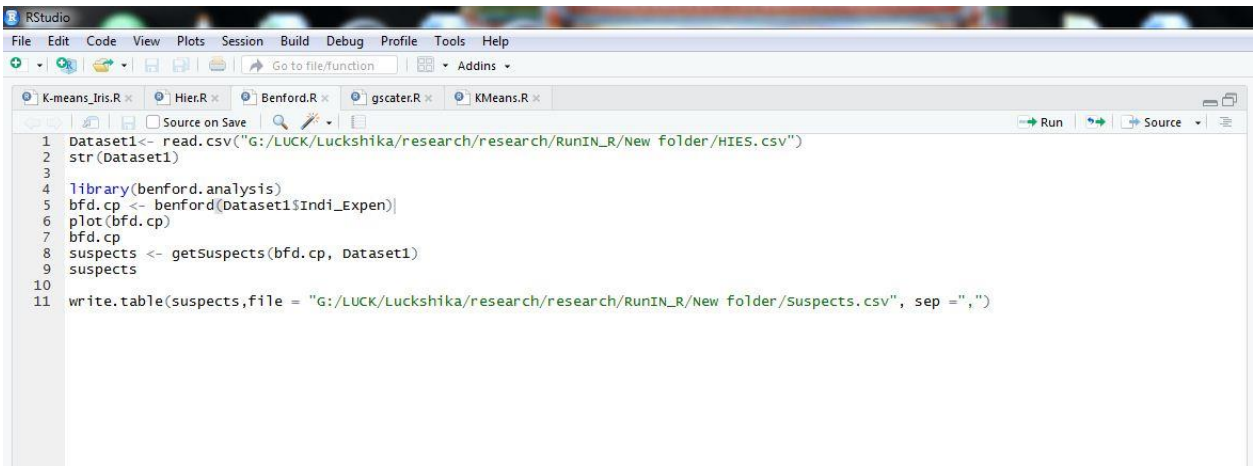


Figure 6 Benford code

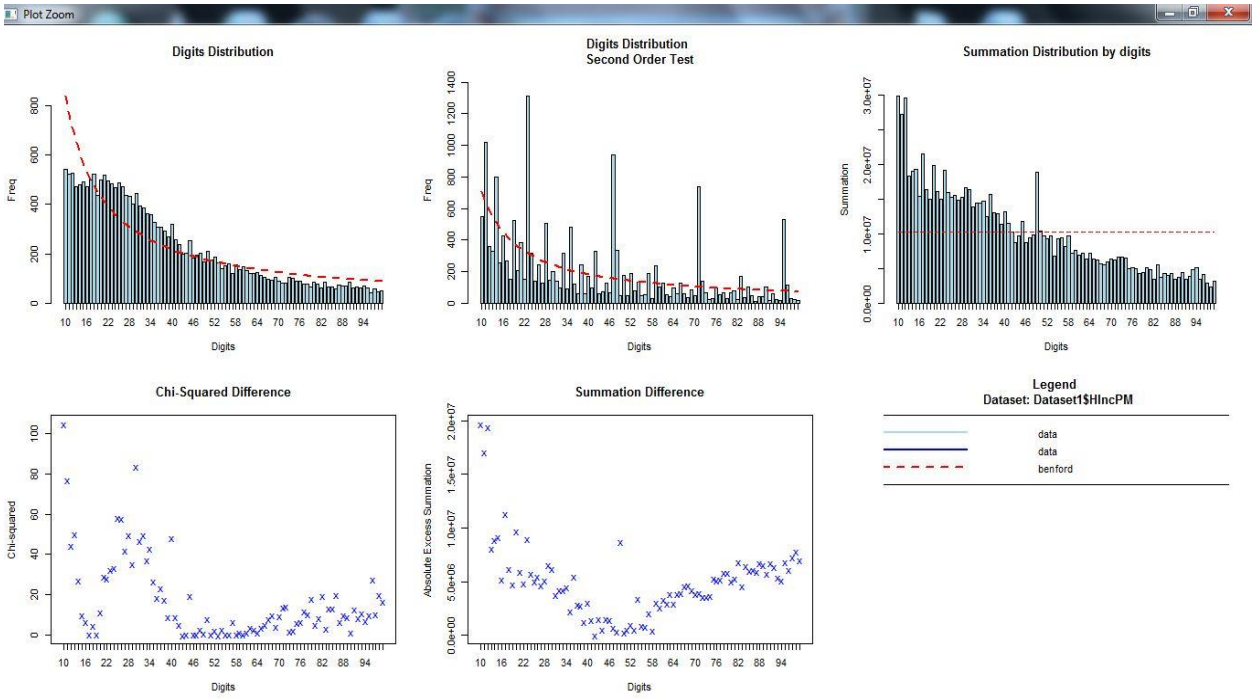


Figure 7 Benford Graph

```
Remember: Real data will never conform perfectly to Benford's Law. You should not focus on p-values!> suspects
<- getSuspects(bfd.cp, Dataset1)
> suspects
      DISTRICT SECTOR DS PSU    A0 SNUMBER   HIncPM   HExpPM Electricity_bill Water_bill Food_Expenditure Food_
Value
   1:       91      2 10  58 92811       5 11535.71 10084.10             249          0             4956
 1239
   2:       32      2 10  66 43325      10  1090.00  7436.43              75         60             5788
 1447
   3:       91      3 10  72 91663       8 10300.00  5339.76               0          0             2372
  593
   4:       91      1 10  20 91891       7 11087.14 10404.29               0        100             7060
 1765
   5:       41      2 10  30 48527       5 11608.33 10847.14               0          0             8892
 2223
  ---
1063:       53      1 10  36 67784       2 10230.00 21267.14               0          0            17600
 4400
1064:       82      2 10  15 89521       4 10541.67 19363.33               0          0            13440
 3360
1065:       61      2 10 112 73824       4 11342.86  7865.48               0          0             5980
 1495
1066:       53      2 10  47 67936       5  1000.00  9634.48               0          0             5872
 1468
1067:       81      3 10  10 86030       6  1138.10  5485.02               0          0             4268
 1067
      Communication_Expen adhoc_Expen OtherIndi_Expen Indi_Expen
   1:                   0         500             50         10
   2:                   0           0              0         10
   3:                   0           0              0         10
   4:                   0           0              0         15
   5:                   0           0              0         20
  ---
1063:                1000           0              0          0
1064:                   0           0              0          0
1065:                   0           0              0          0
```

Figure 8 Benford Suspicious

For all attributes find the suspect data using benford law.

Box Plot

Boxplot is a method for graphically depicting groups of numerical data through their quartiles. Box plots may also have lines extending vertically from the boxes indicating variability outside the upper and lower quartiles. Using R can plot the data points and remove the outliers.

outliers = boxplot(HIES, plot=TRUE)$out

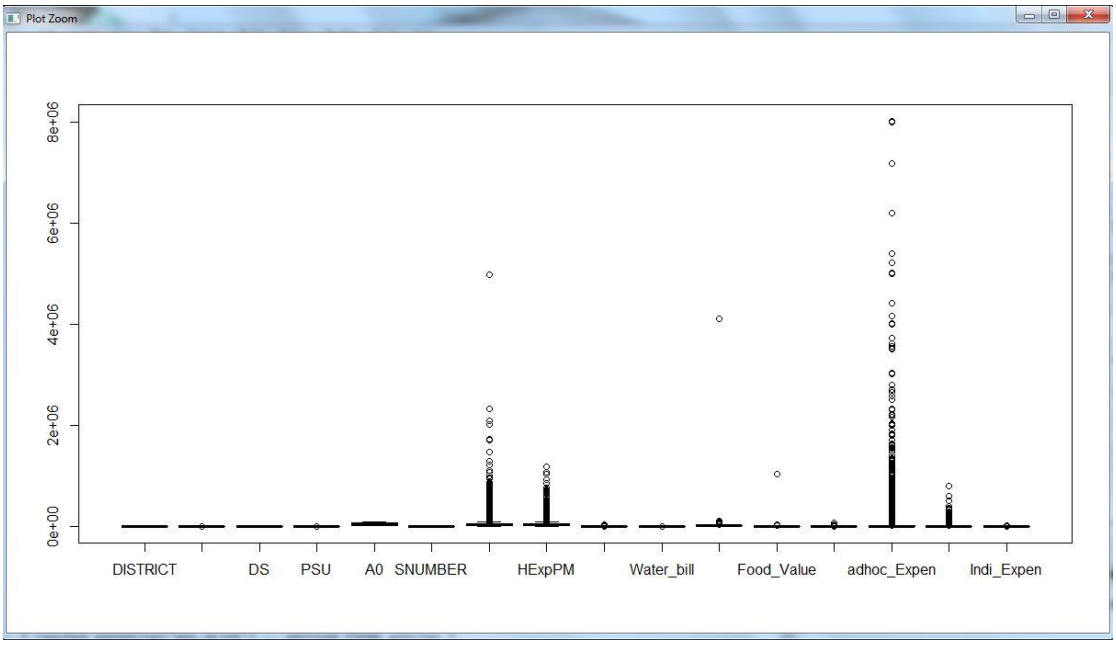


Figure 9 Boxplot outlier

After remove outlier

outliers1 = boxplot(HIES, plot=TRUE, outline=FALSE)$out

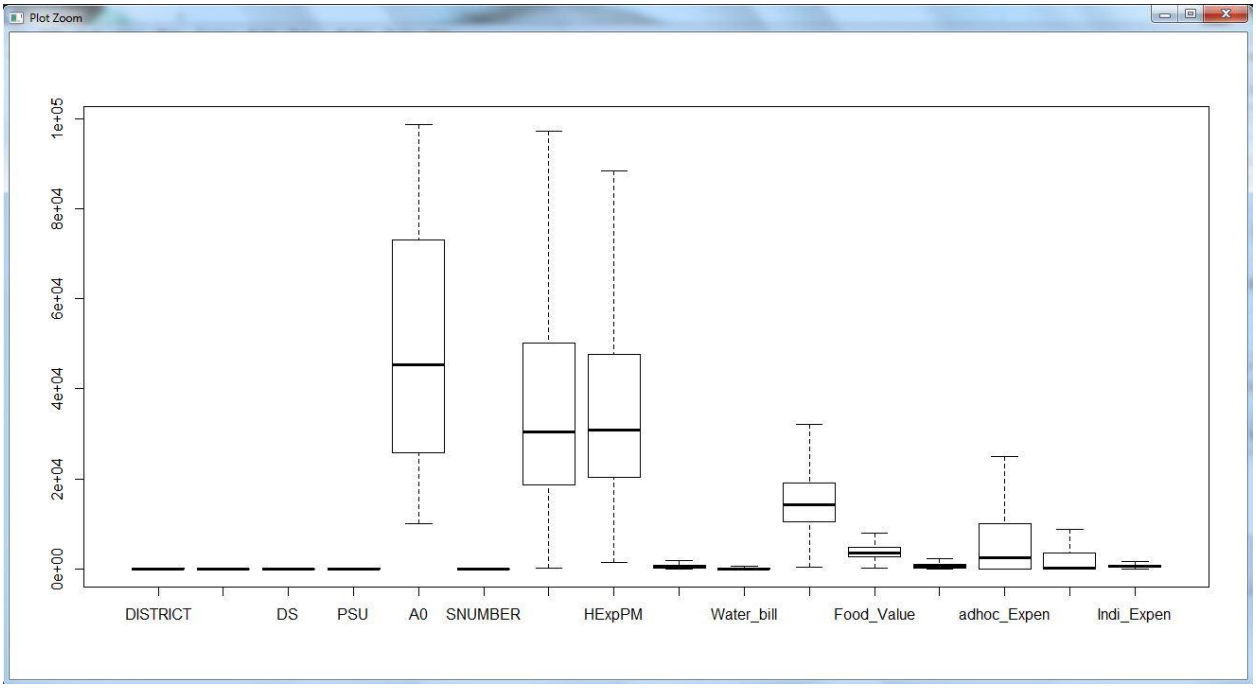


Figure 10 Outlier

Median and Interquartile Deviation Method (IQD)

For this outlier detection method, the median of the residuals is calculated, along with the 25th percentile (Q1) and the 75th percentile (Q3). The difference between the 25th and 75th percentile is the interquartile deviation (IQR). Then, the difference is calculated between each historical value and the residual median. If the historical value is a certain number of IQD away from the median of the residuals, that value is classified as an outlier.

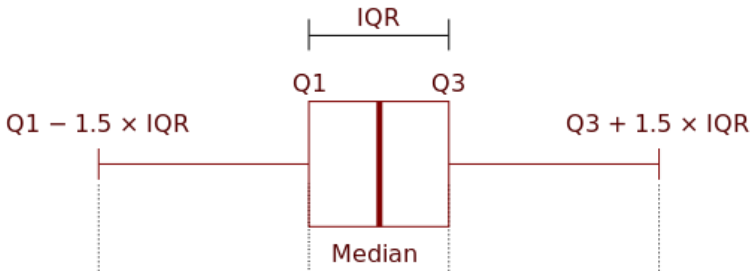


Figure 11: IQD method

IQR = Q3 − Q1

In this research, the attribute data which is away from Upper Threshold and beyond the Median or away from Lower Threshold and from the further side of the Median is taken as the outlier value for that attribute. If the data falls above mentioned range, the particular ID is coded as “1” for that attribute. If the data does not fall above mentioned range, the particular user ID is coded as “0” for that attribute.

When use the IQR method for attributes that found as outlier in boxplot. Those are Income, Expenditure, adhoc expenses, Food expenses, Individual, Other Expenses. Using IQR method category each record is outlier or not for each attributes. All attributes are found as an outlier for 41 records.

Total outlier in 6 attributes	Number of records
6	41
5	133
4	373
3	593

Table 3: IQR outlier

	A	B	C	D	E	F	G	H	I	J
1	UserId	HIncPM_Outlier_	HExpPM_Outlier_	Food_Expenditu	adhoc_Expen_Outlier_IQR_M	OtherIndi_Exper	Indi_Expen_Ou	Total_Oulier_IQR	Total_Oulier_IQR_Percentage	
2	4612	1	1	1	1	1	1	6	100	
3	4642	1	1	1	1	1	1	6	100	
4	4670	1	1	1	1	1	1	6	100	
5	4681	1	1	1	1	1	1	6	100	
6	4696	1	1	1	1	1	1	6	100	
7	4698	1	1	1	1	1	1	6	100	
8	4704	1	1	1	1	1	1	6	100	
9	4719	1	1	1	1	1	1	6	100	
10	4730	1	1	1	1	1	1	6	100	
11	4765	1	1	1	1	1	1	6	100	
12	4784	1	1	1	1	1	1	6	100	
13	4788	1	1	1	1	1	1	6	100	
14	4795	1	1	1	1	1	1	6	100	
15	4819	1	1	1	1	1	1	6	100	
16	4827	1	1	1	1	1	1	6	100	
17	4833	1	1	1	1	1	1	6	100	
18	4836	1	1	1	1	1	1	6	100	
19	4845	1	1	1	1	1	1	6	100	
20	4858	1	1	1	1	1	1	6	100	
21	4859	1	1	1	1	1	1	6	100	
22	4863	1	1	1	1	1	1	6	100	
23	9487	1	1	1	1	1	1	6	100	
24	9497	1	1	1	1	1	1	6	100	
25	9517	1	1	1	1	1	1	6	100	

Figure 12: IQR

After remove the outliers use the K-means clustering method for cluster the data set.

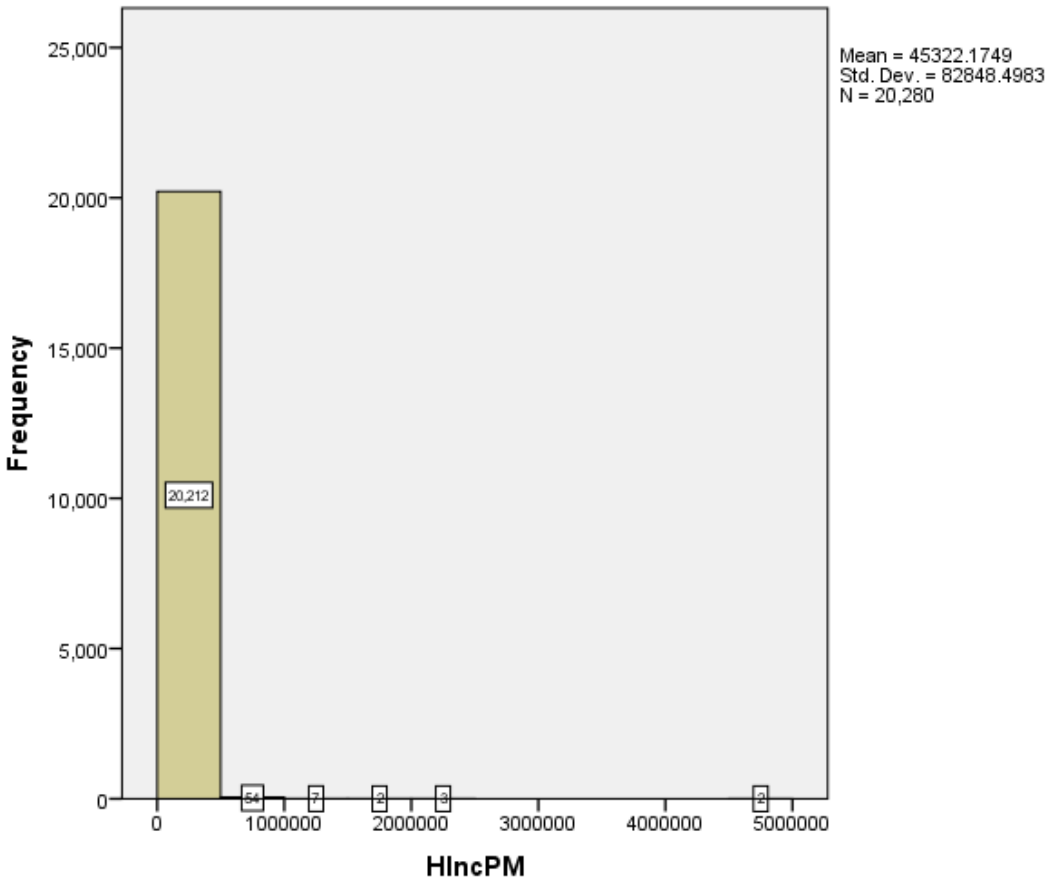
Descriptive Analysis

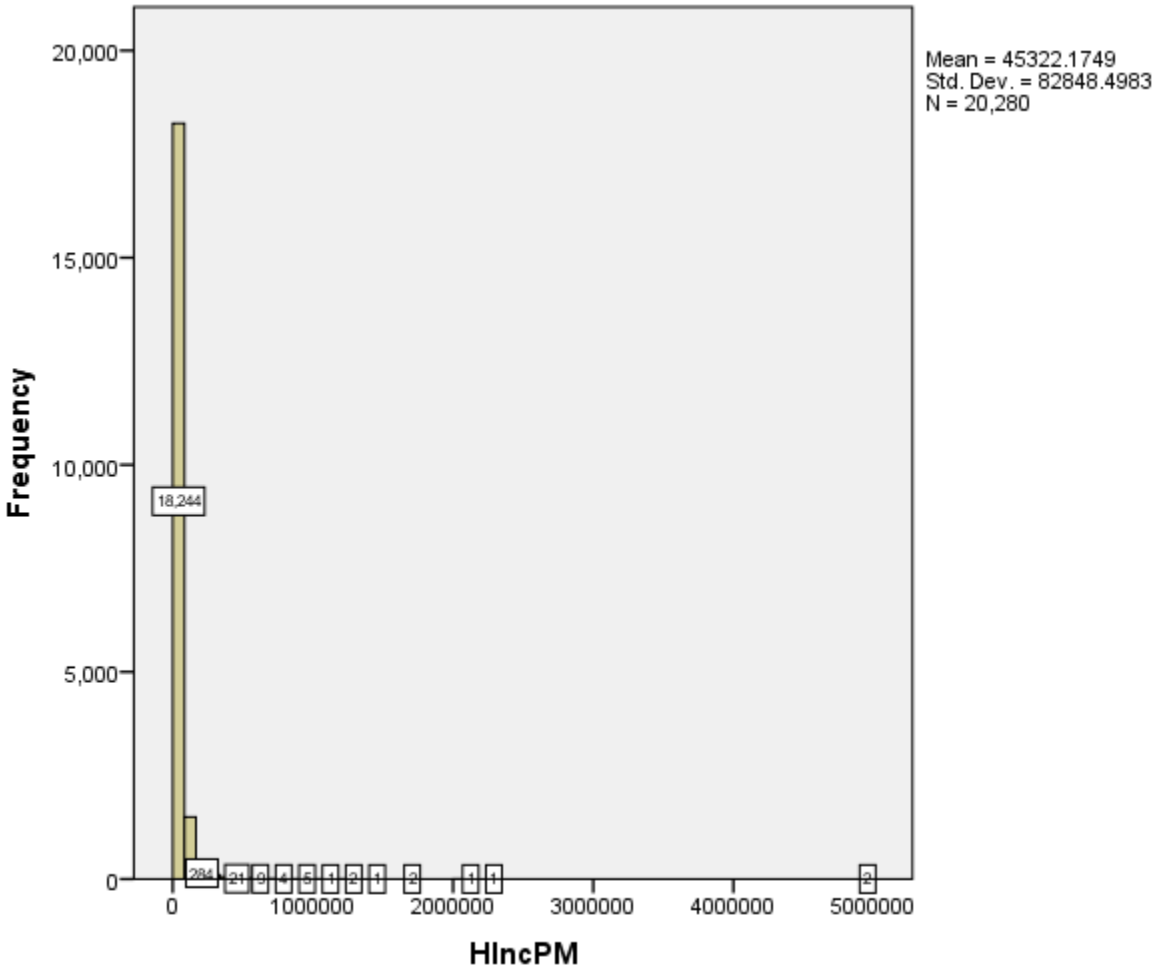
Descriptive analysis helps to identify how the each attributes change with the household members in household income and expenditure dataset. It will produce the minimum, maximum, mean, standard deviation for each attributes.

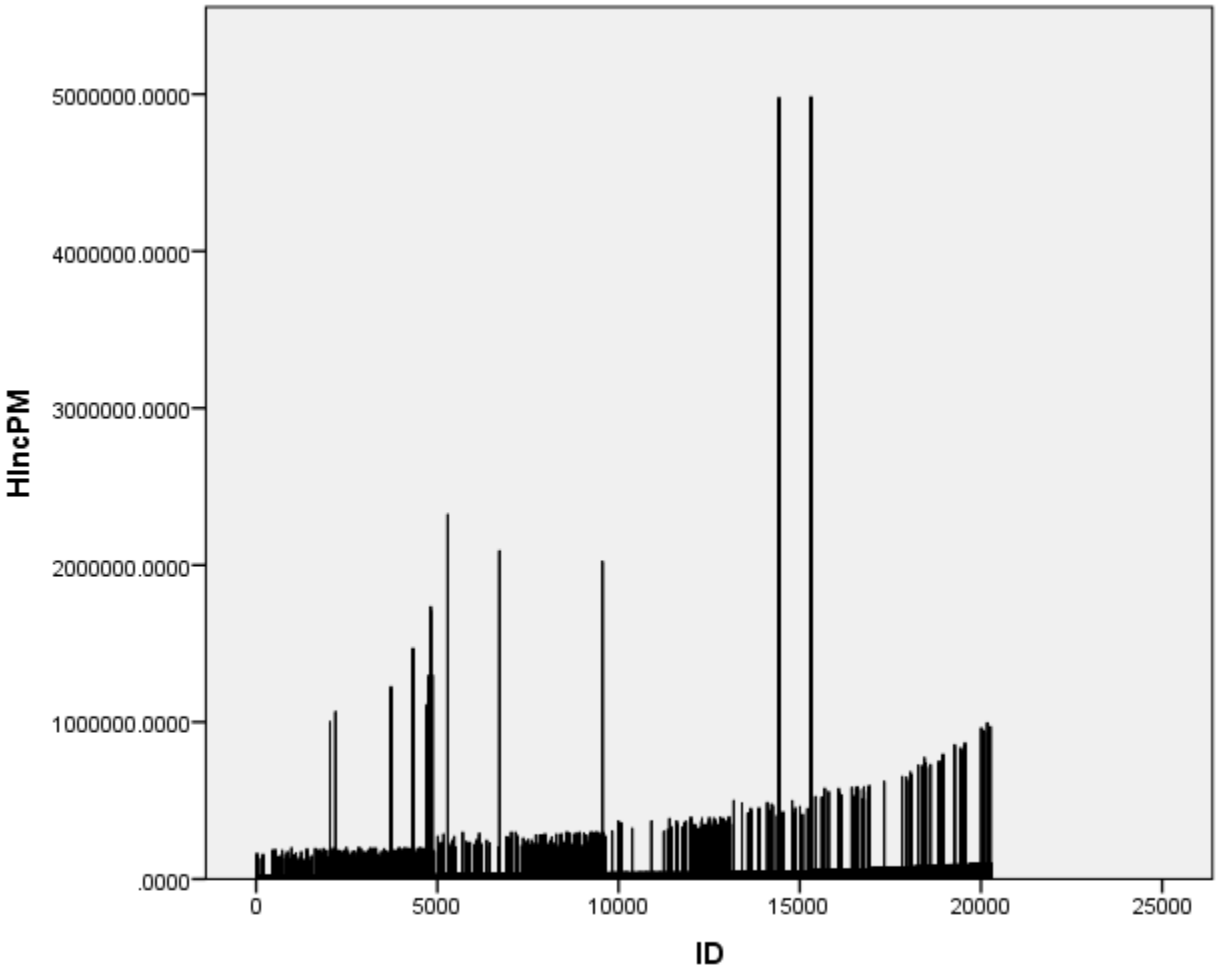
Household Income Per Month

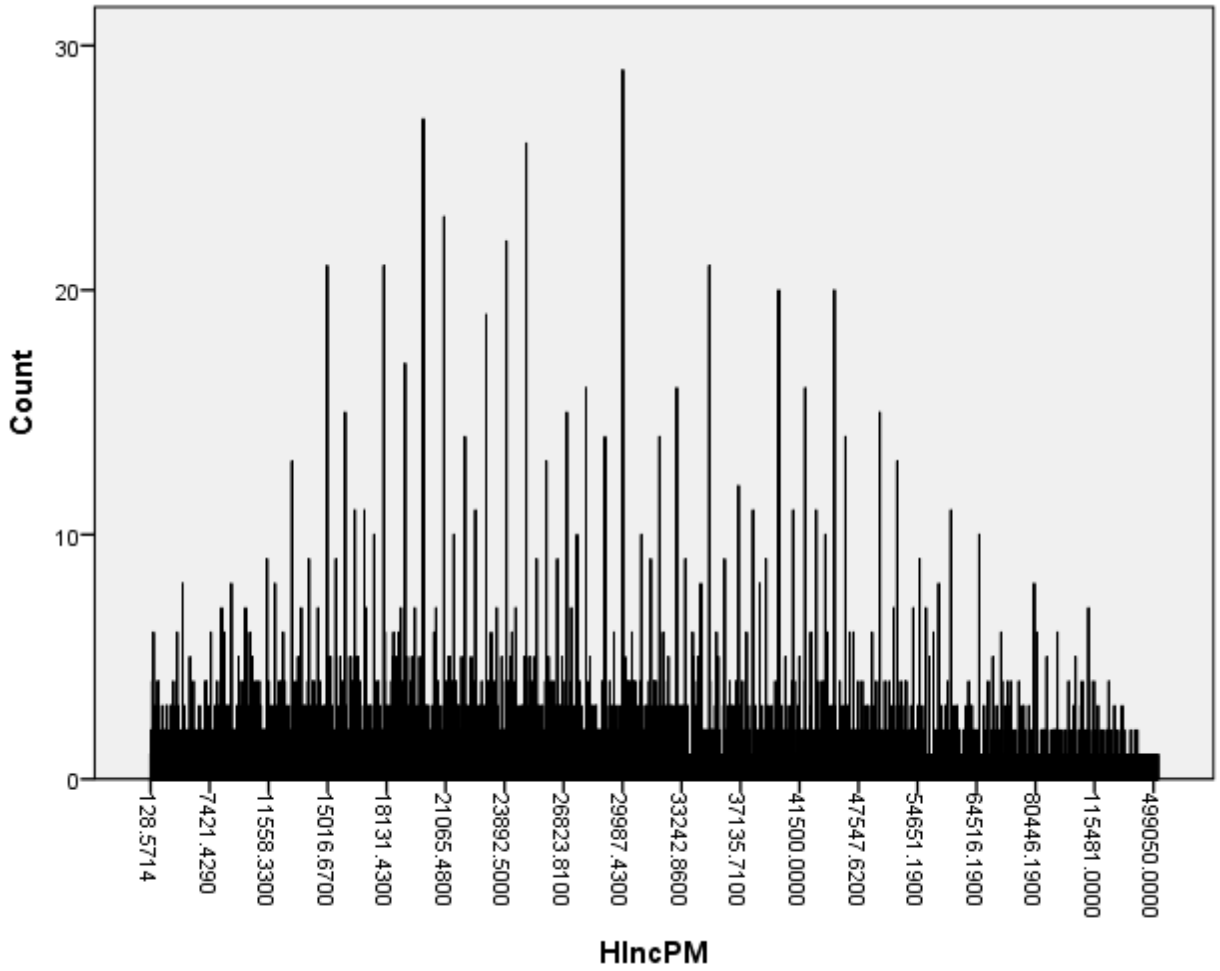
Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
HIncPM	20280	128.5714	4983000.0000	45322.174922	82848.4982522
Valid N (listwise)	20280				




Mean = 45322.1749
Std. Dev. = 82848.4983
N = 20,280
Frequency
20,000
15,000
10,000
5,000
0
18,244
284
0
1000000
2000000
3000000
4000000
5000000
HIncPM

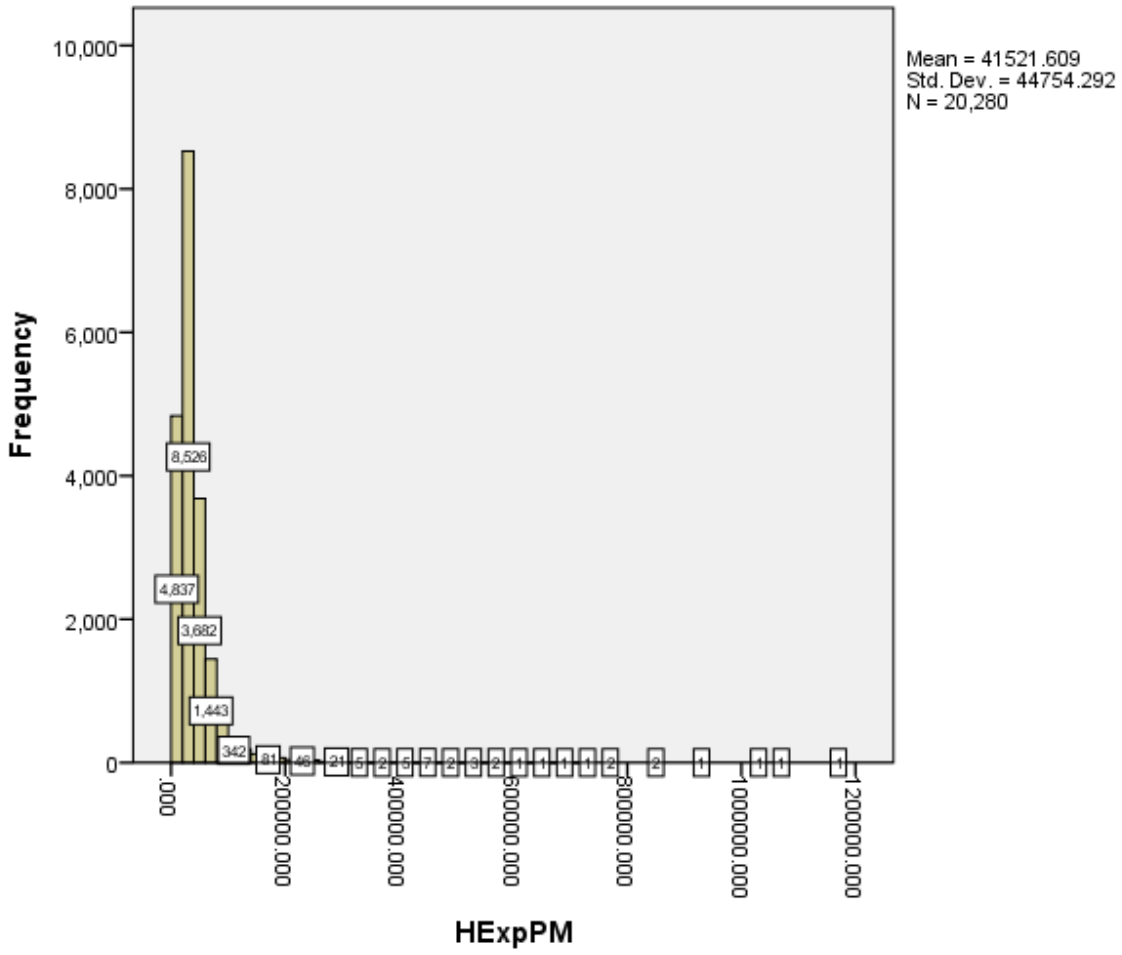

HIncPM
5000000.0000
4000000.0000
3000000.0000
2000000.0000
1000000.0000
.0000
0
5000
10000
15000
20000
25000
ID

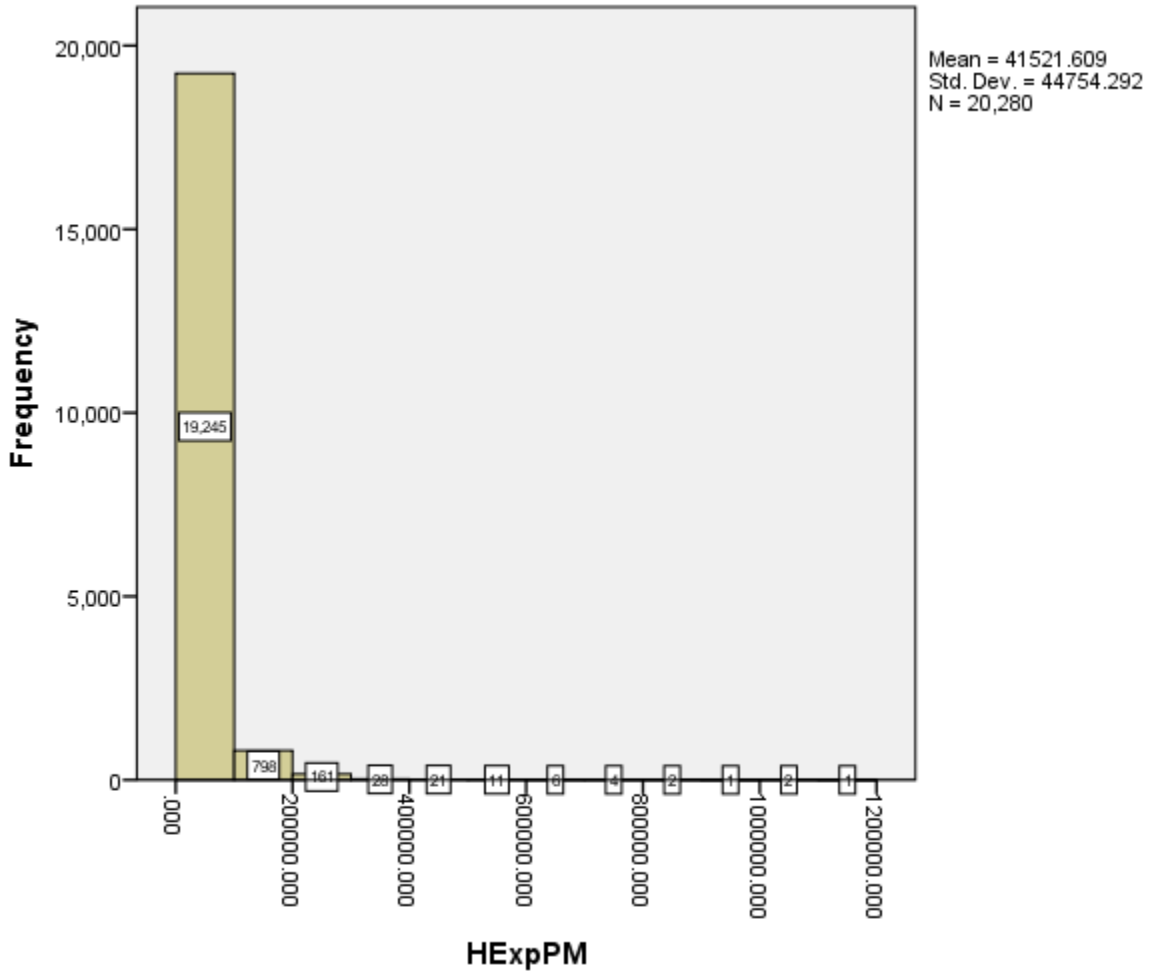


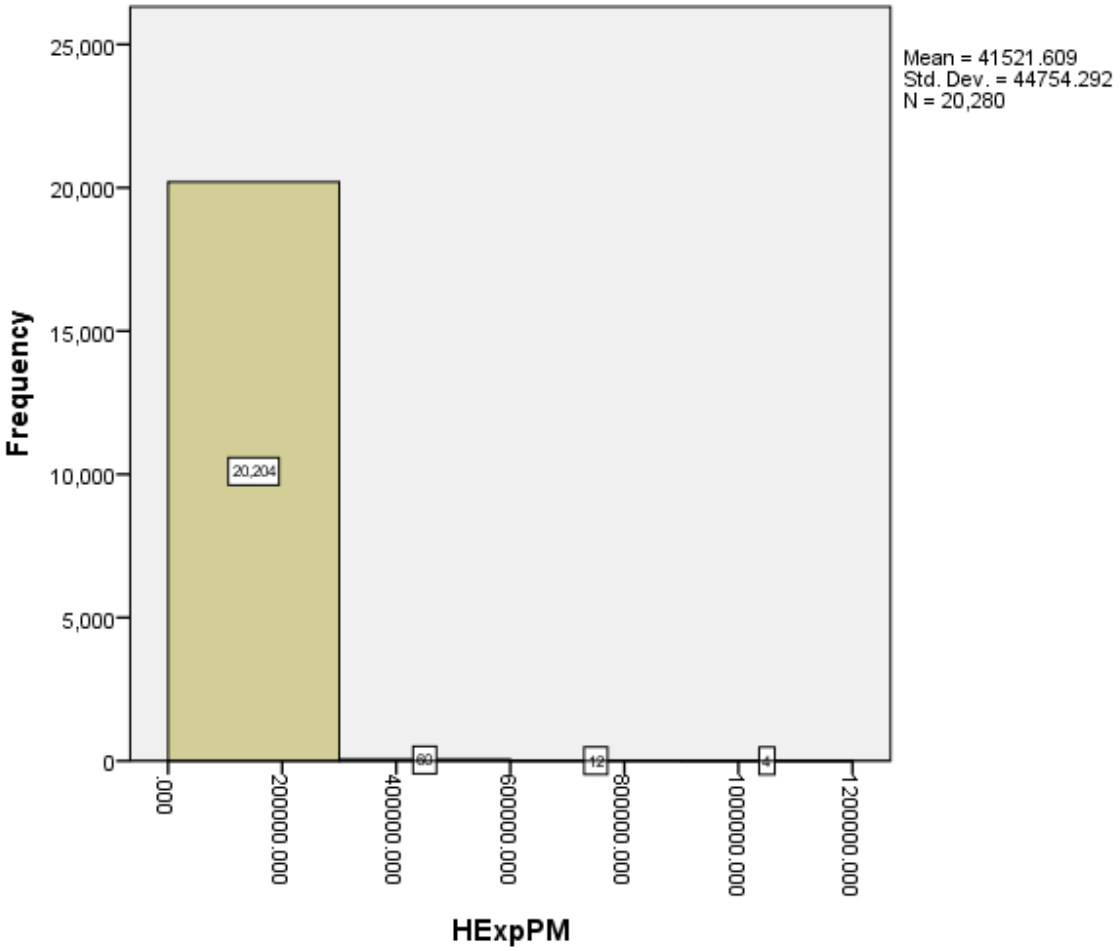
The minimum value for the HIncPM is 128.57 and the maximum value for the HIncPM is 4983000.00. There is a huge difference between the minimum value and the maximum value. The mean is 45322.17. The standard deviation is 82848.49. 99.7% of family's HIncPM value falls below 500000 and 0.3% of family's value falls above 500000. Based on the graph, the majority of family's HIncPM value is close to the mean. Minority of the family's (0.3%) values give significant different compared to the majority of family's values for the HIncPM attribute.

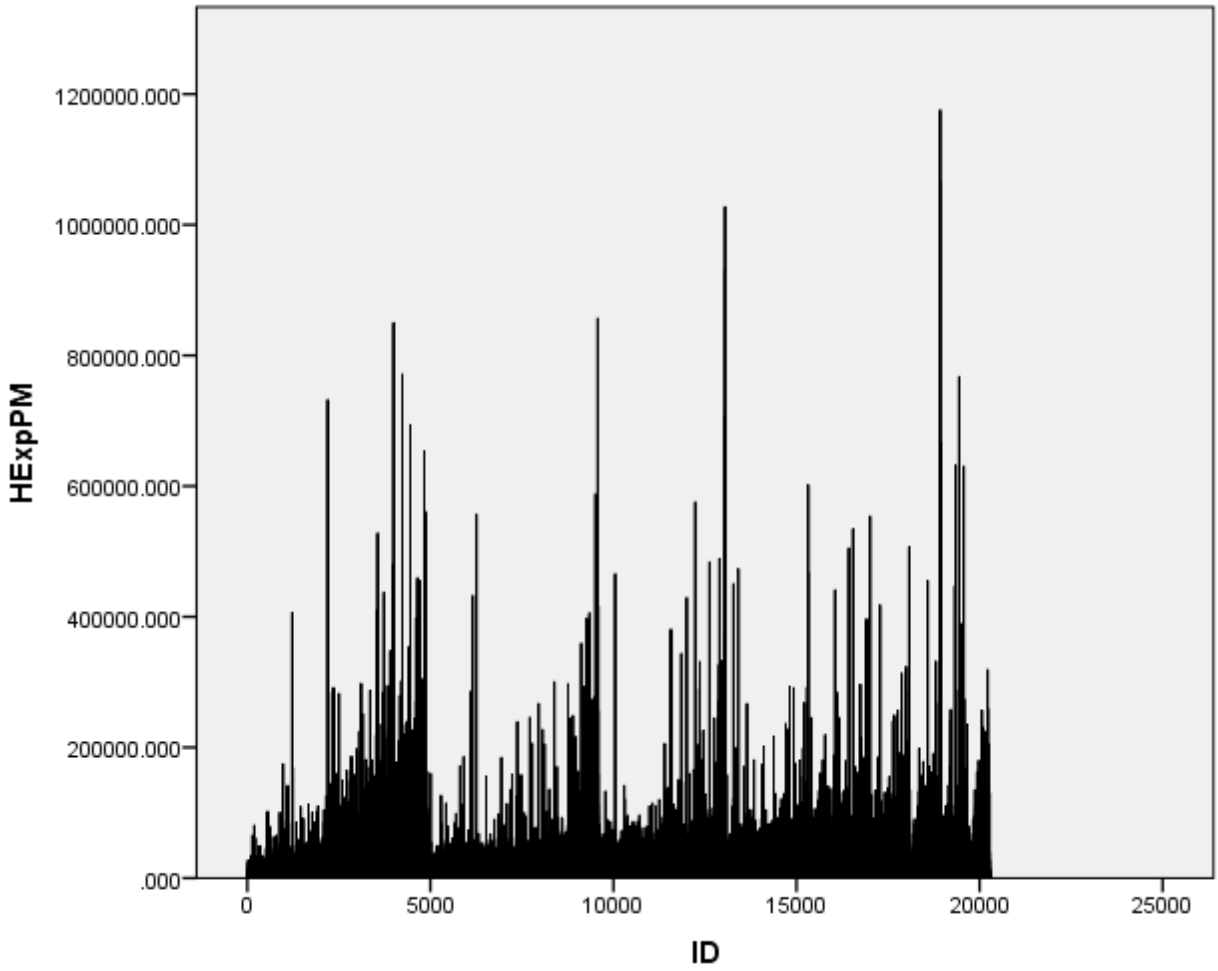
Household Expenditure per Month

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
HExpPM	20280	411.143	1175382.000	41521.60870	44754.291784
Valid N (listwise)	20280				




Mean = 41521.609
Std. Dev. = 44754.292
N = 20,280
Frequency
20,000
15,000
10,000
5,000
0
19,245
798
161
28
21
11
6
4
2
1
2
1
.000
200000.000
400000.000
600000.000
800000.000
1000000.000
1200000.000
HExpPM


Mean = 41521.609
Std. Dev. = 44754.292
N = 20,280
Frequency
25,000
20,000
15,000
10,000
5,000
0
20,204
60
12
4
.000
200000.000
400000.000
600000.000
800000.000
1000000.000
1200000.000
HExpPM

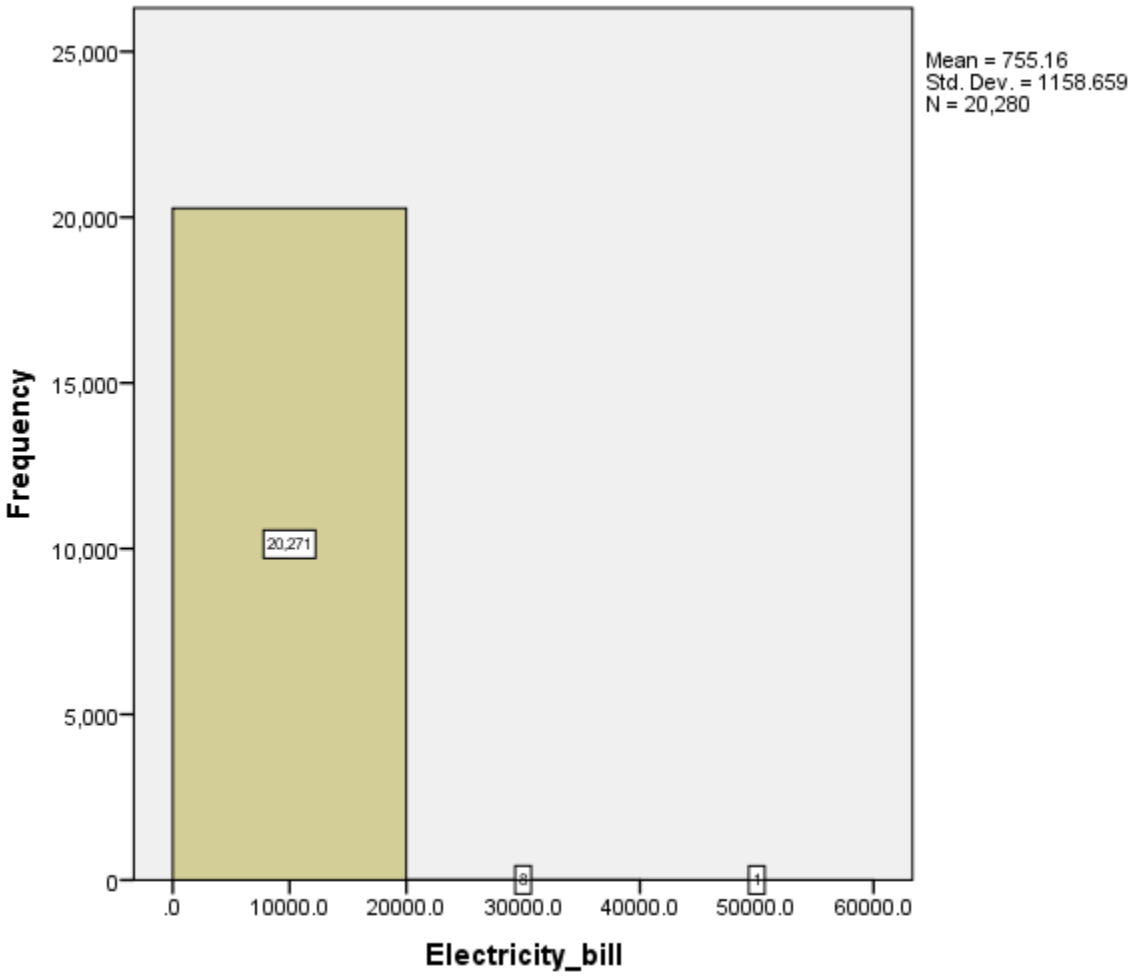


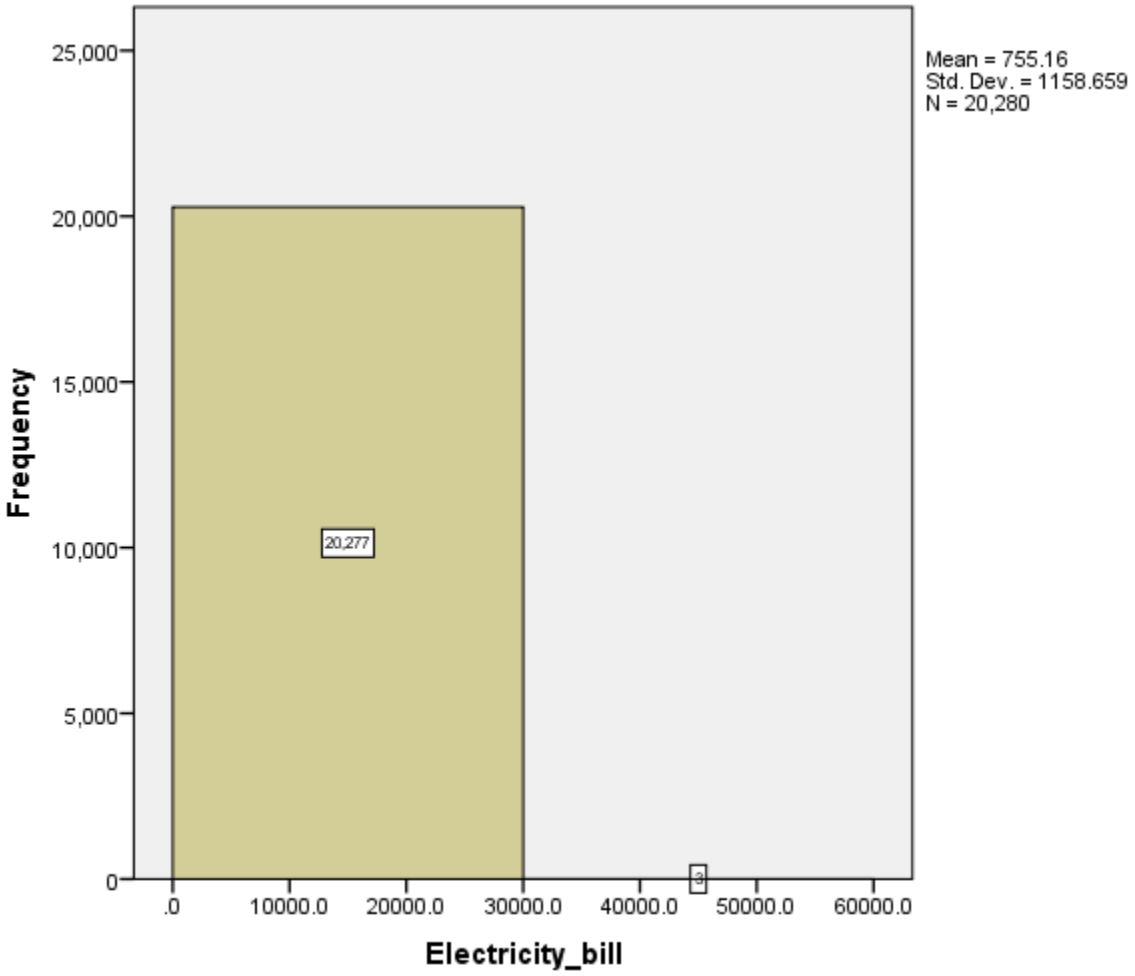
The minimum value for the HExpPM is 1411.14 and the maximum value for the HExpPM is 1175382.00. There is a huge difference between the minimum value and the maximum value. The mean is 41521.60. The standard deviation is 44754.29. 77.2% of family's HExpPM value falls below 50 000 and 22.3% of family's value falls above 50000 and below the 300000 and 0.37% of family's value falls above 300000. Based on the graph, the majority of family's HExpPM value is close to the mean. Minority of the family's values give significant different compared to the majority of family's values for the HExpPM attribute.

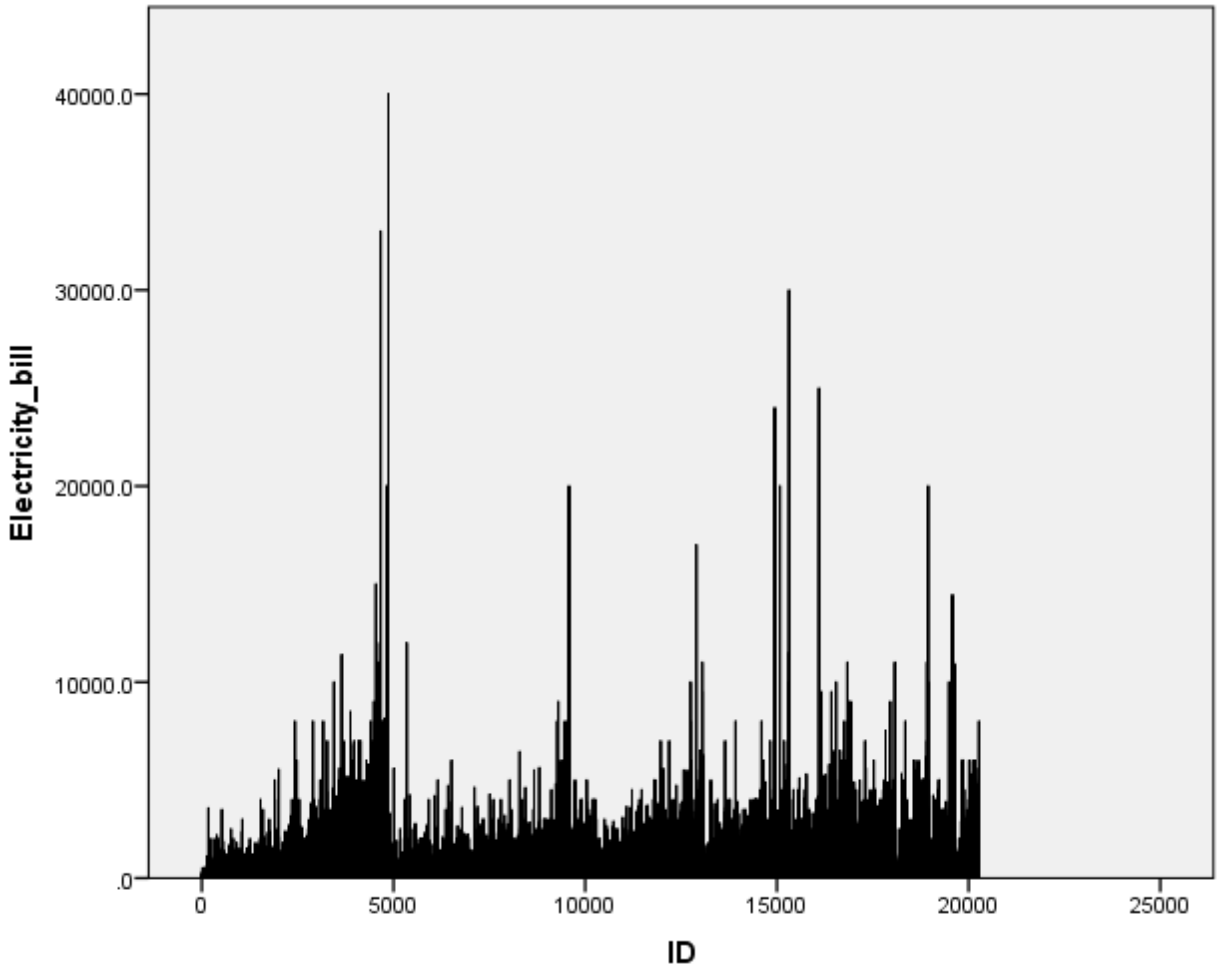
Electricity bill

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Electricity_bill	20280	.0	40000.0	755.161	1158.6585
Valid N (listwise)	20280				




Mean = 755.16
Std. Dev. = 1158.659
N = 20,280
Frequency
25,000
20,000
15,000
10,000
5,000
0
20,277
3
.0
10000.0
20000.0
30000.0
40000.0
50000.0
60000.0
Electricity_bill

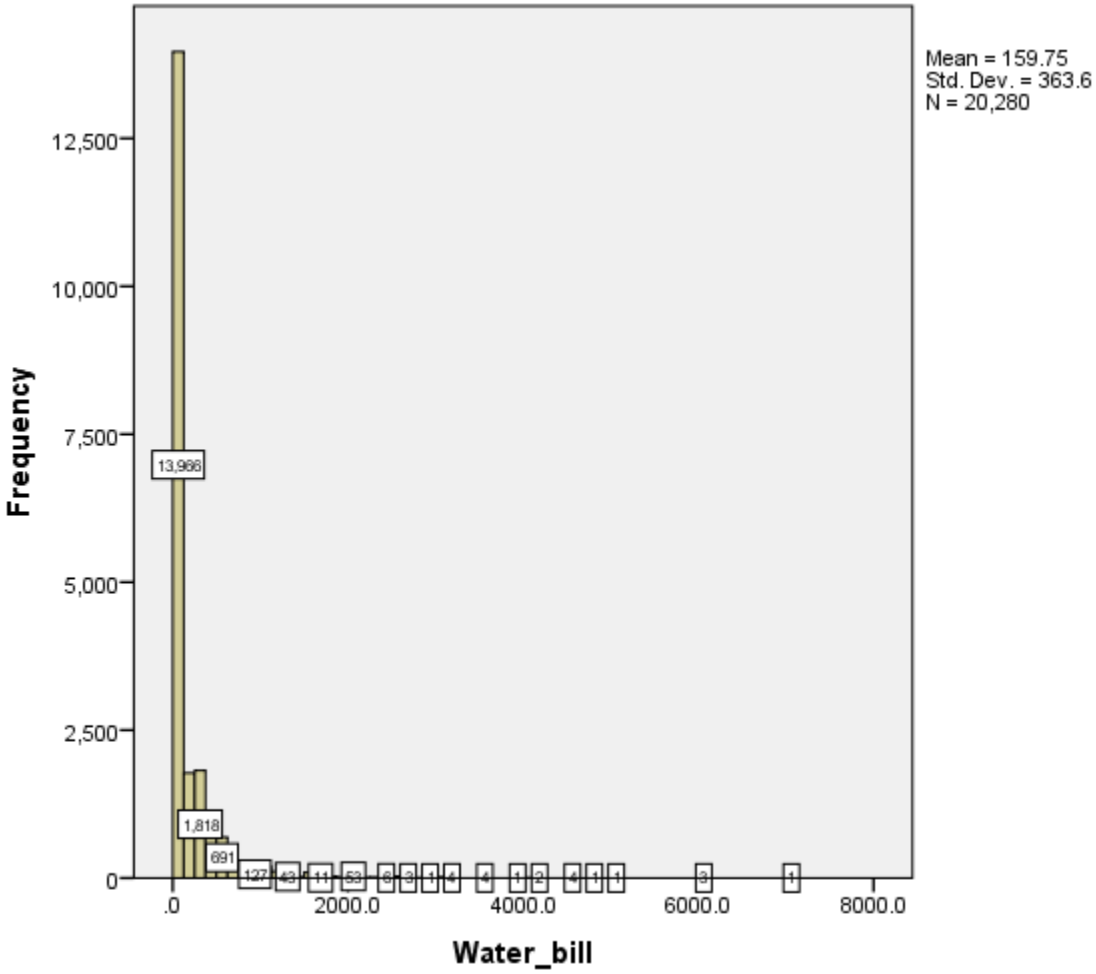


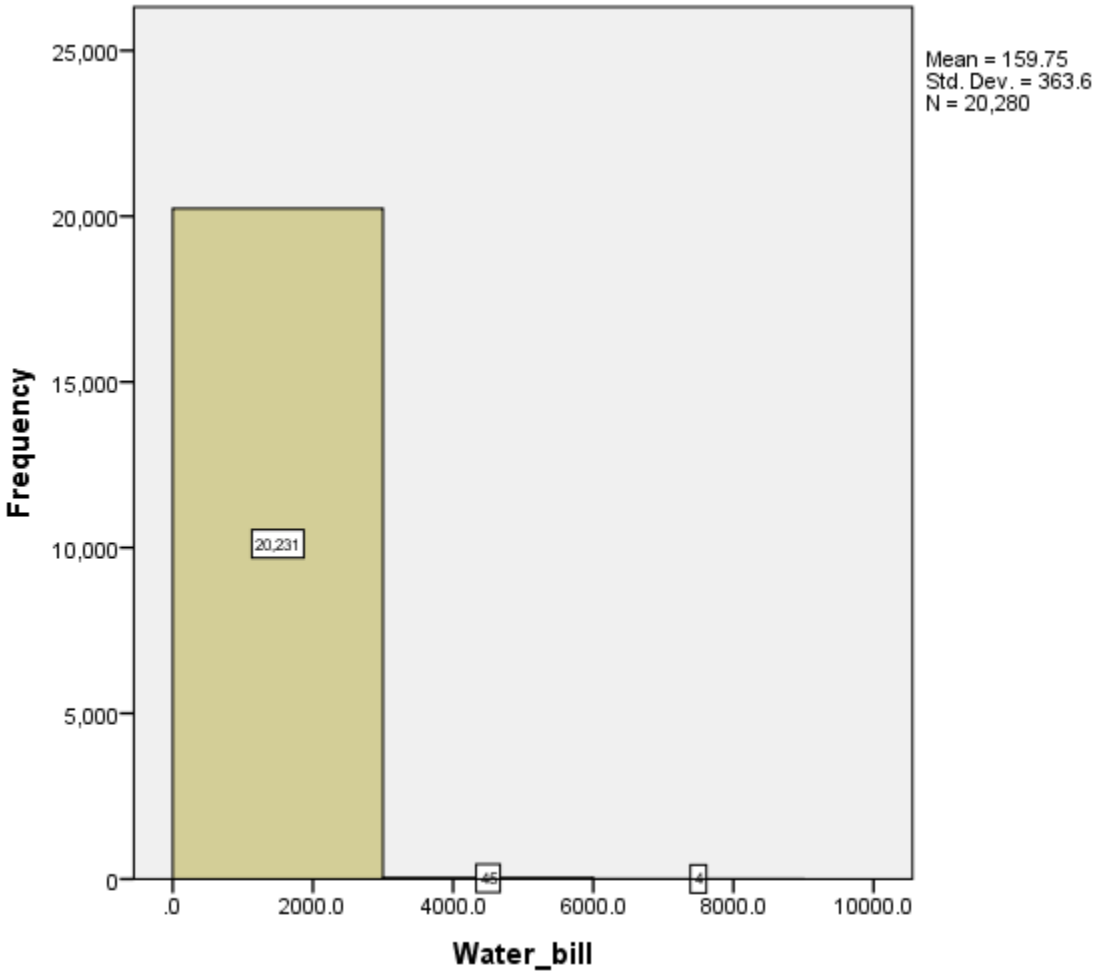
The minimum value for the electricity bill is 0.0 and the maximum value for the electricity bill is 40 000.00. There is a difference between the minimum value and the maximum value. The mean is 755.16. The standard deviation is 1158.65. 99.8% of family's electricity bill value falls below 10 000 and 0.2% of family's value falls above 10 000. Based on the graph, the majority of family's electricity bill value is close to the mean. Minority of the family's values give significant different compared to the majority of family's values for the electricity bill attribute.

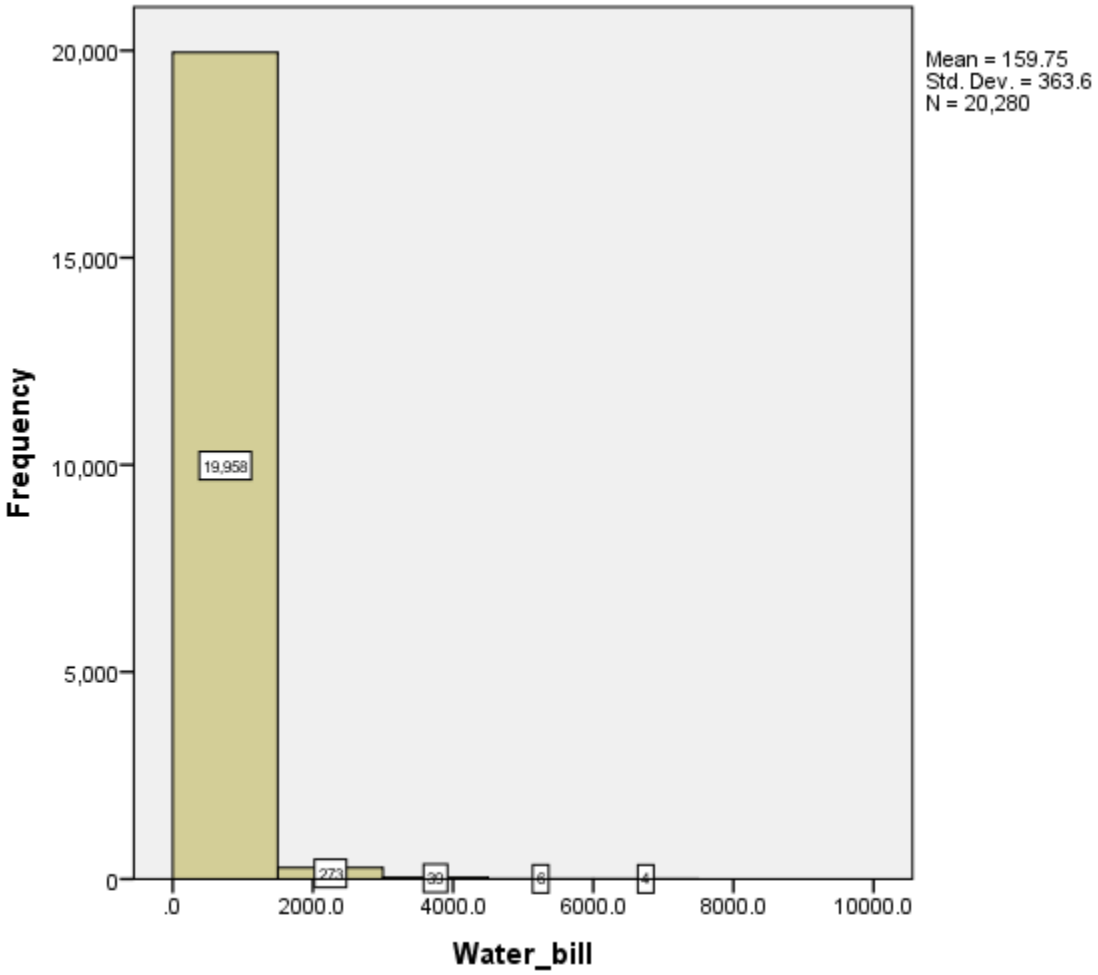
Water bill

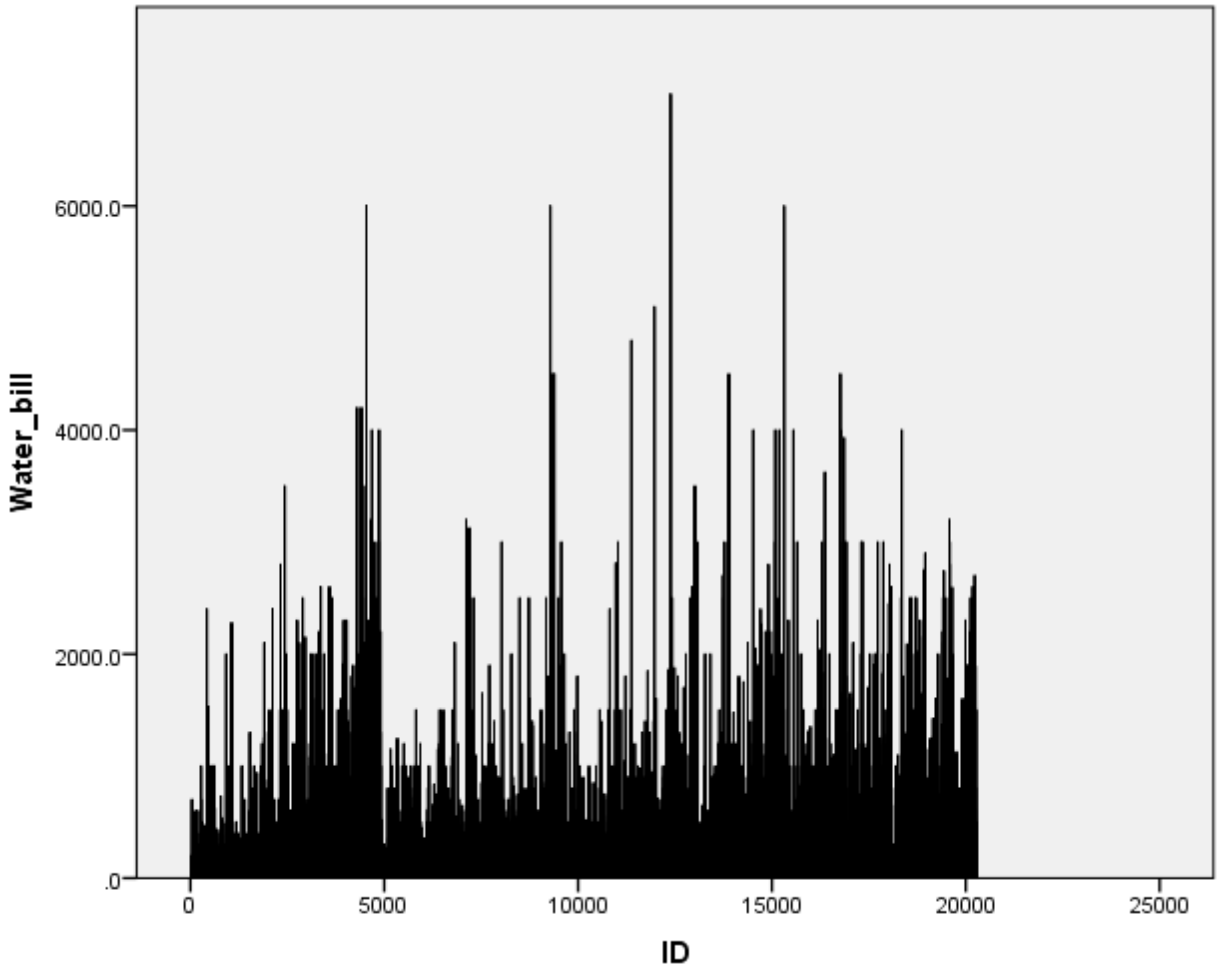
Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Water_bill	20280	.0	7000.0	159.753	363.5998
Valid N (listwise)	20280				




Mean = 159.75
Std. Dev. = 363.6
N = 20,280
Frequency
25,000
20,000
15,000
10,000
5,000
0
20,231
45
4
.0
2000.0
4000.0
6000.0
8000.0
10000.0
Water_bill


Mean = 159.75
Std. Dev. = 363.6
N = 20,280
Frequency
20,000
15,000
10,000
5,000
0
19,958
273
39
6
4
.0
2000.0
4000.0
6000.0
8000.0
10000.0
Water_bill

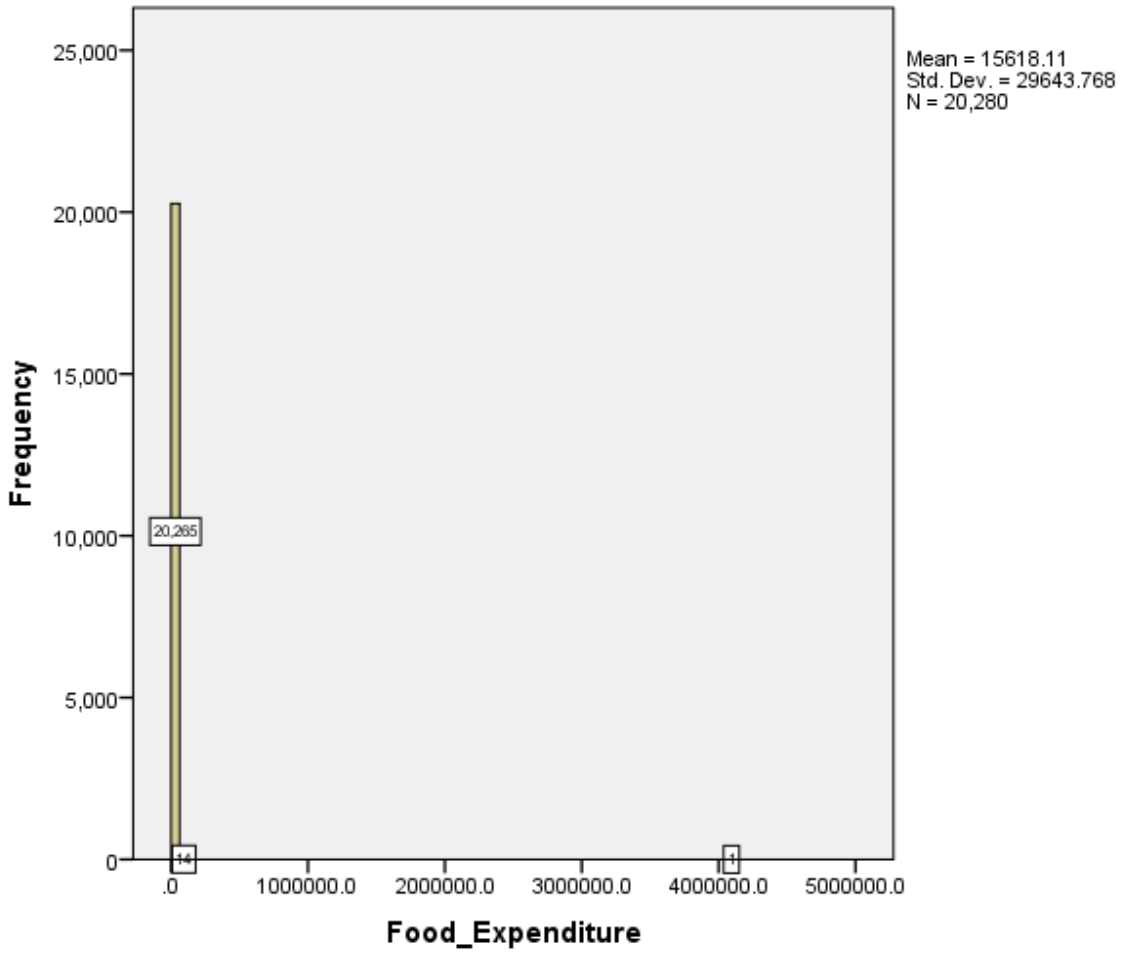


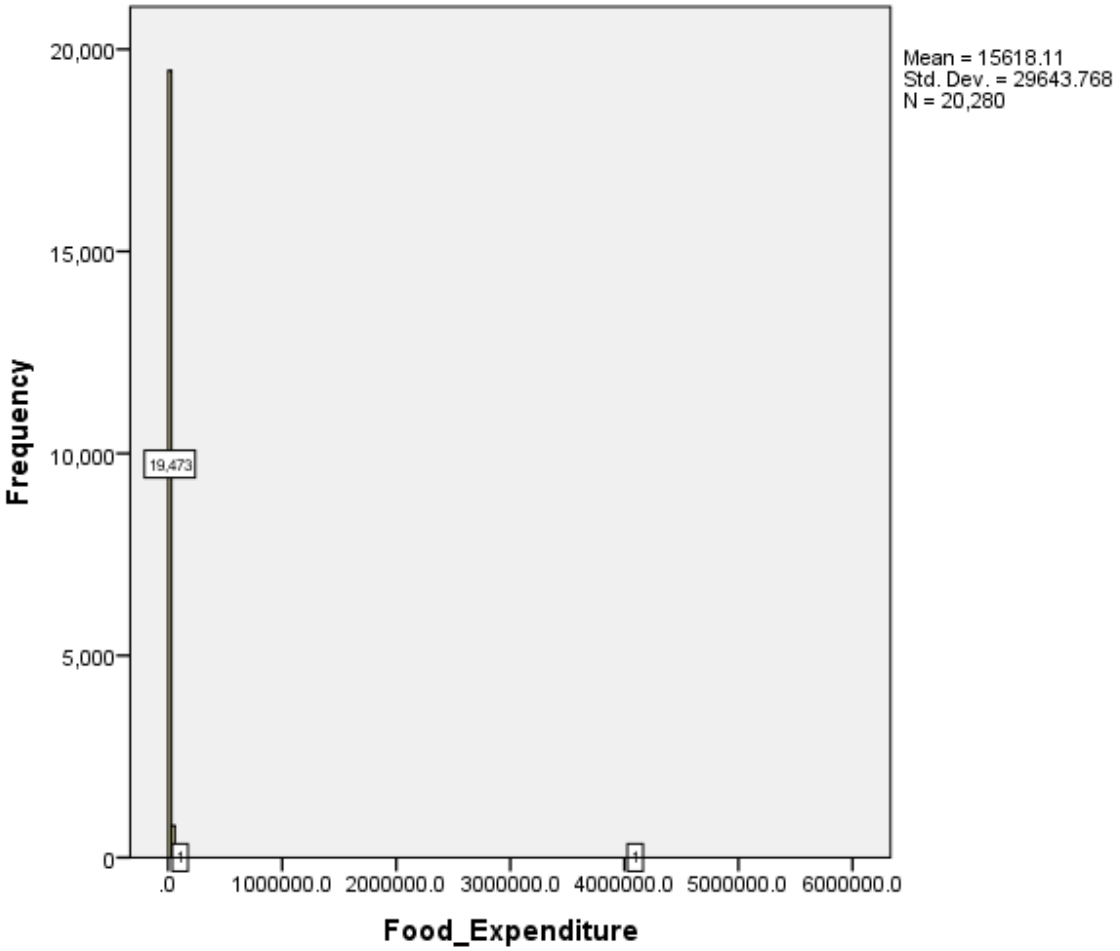
The minimum value for the water bill is 0.0 and the maximum value for the water bill is 7000.0. The mean is 159.75. The standard deviation is 363.6. 98.41% of family's water bill value falls below 1500 and 1.6% of family's value falls above 1500. Based on the graph, the majority of family's water bill value is close to the mean. Minority of the family's (1.6%) values give significant different compared to the majority of family's values for the water bill attribute.

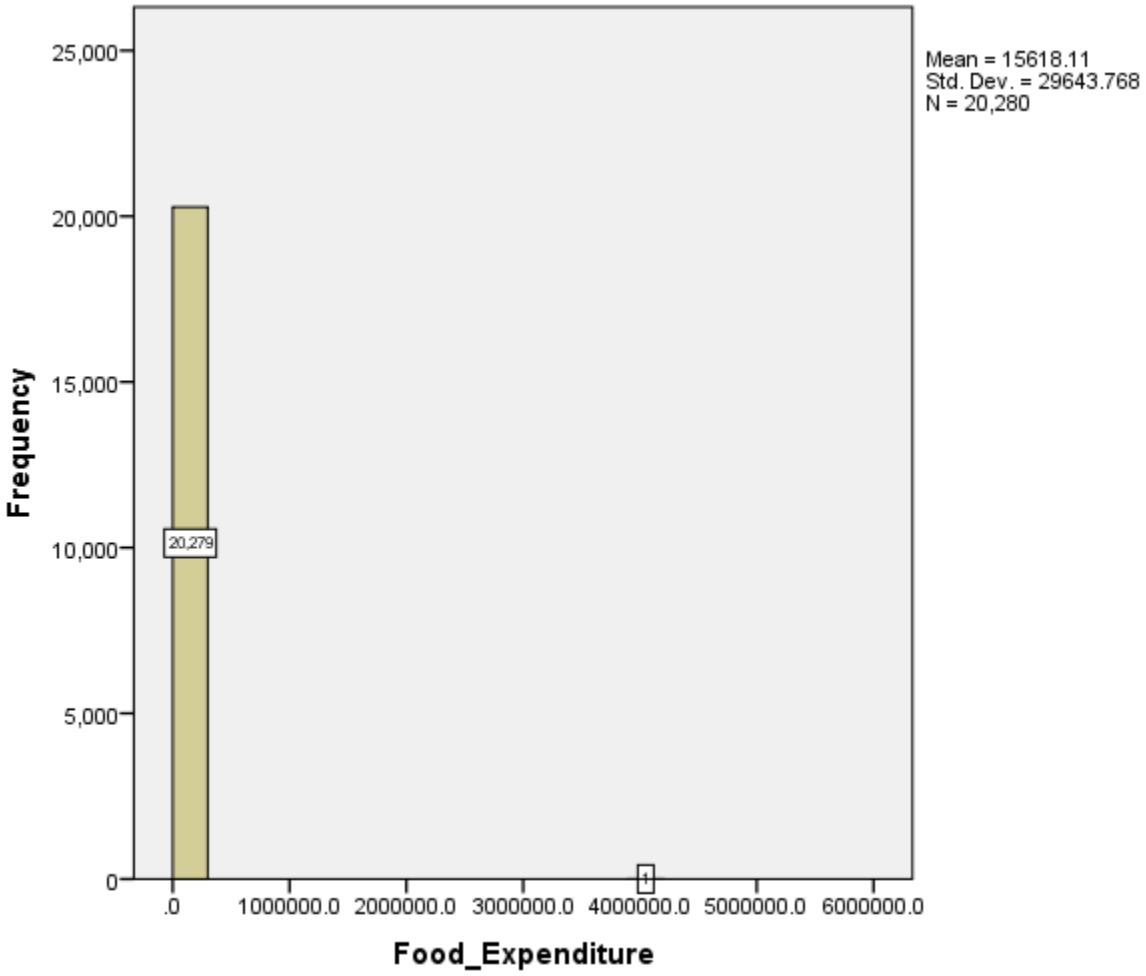
Food Expenditure

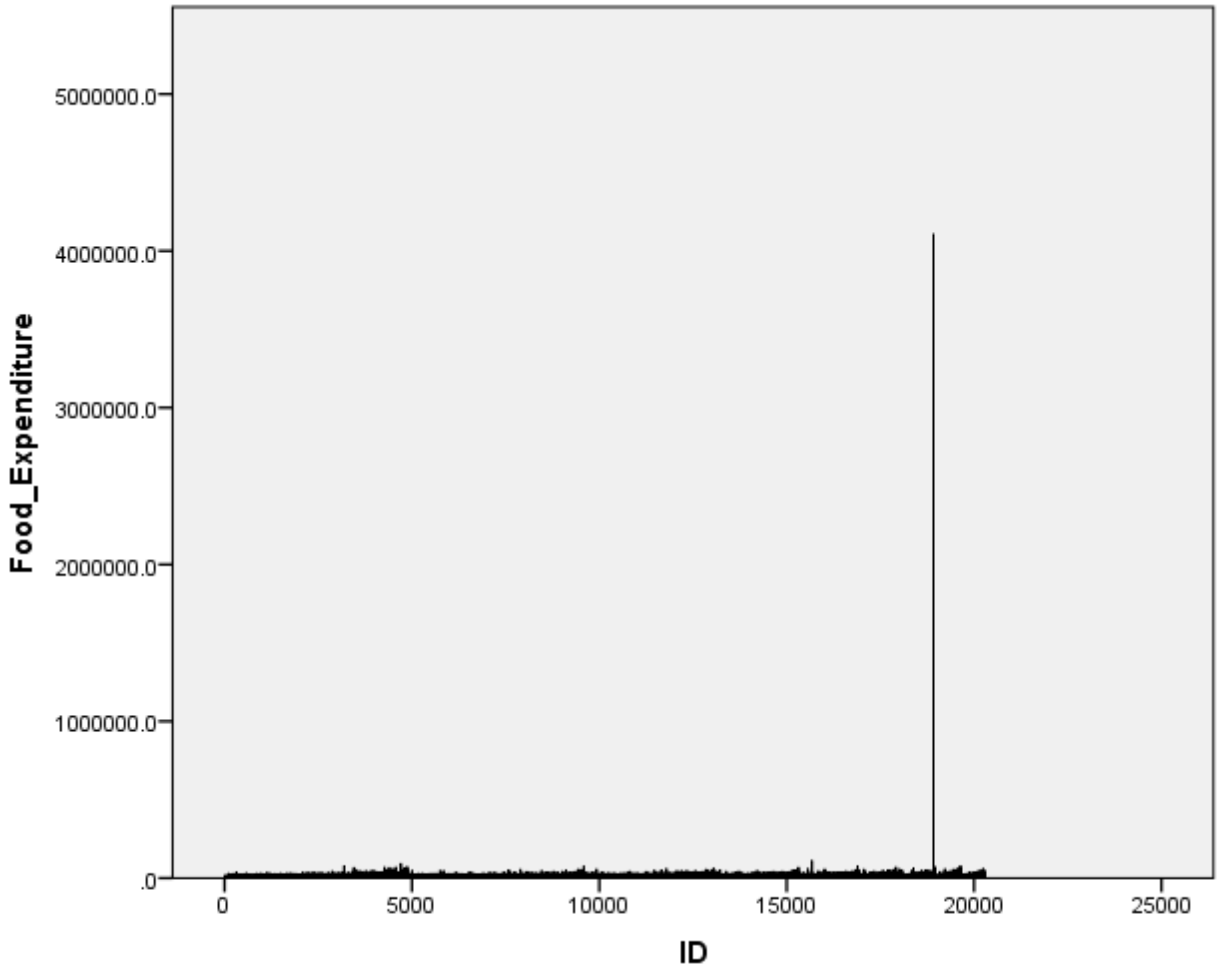
Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Food_Expenditure	20280	436.0	4104636.0	15618.110	29643.7677
Valid N (listwise)	20280				




Mean = 15618.11
Std. Dev. = 29643.768
N = 20,280
Frequency
20,000
15,000
10,000
5,000
0
19,473
.0
1000000.0
2000000.0
3000000.0
4000000.0
5000000.0
6000000.0
Food_Expenditure


Mean = 15618.11
Std. Dev. = 29643.768
N = 20,280
Frequency
25,000
20,000
15,000
10,000
5,000
0
20,279
1
.0
1000000.0
2000000.0
3000000.0
4000000.0
5000000.0
6000000.0
Food_Expenditure

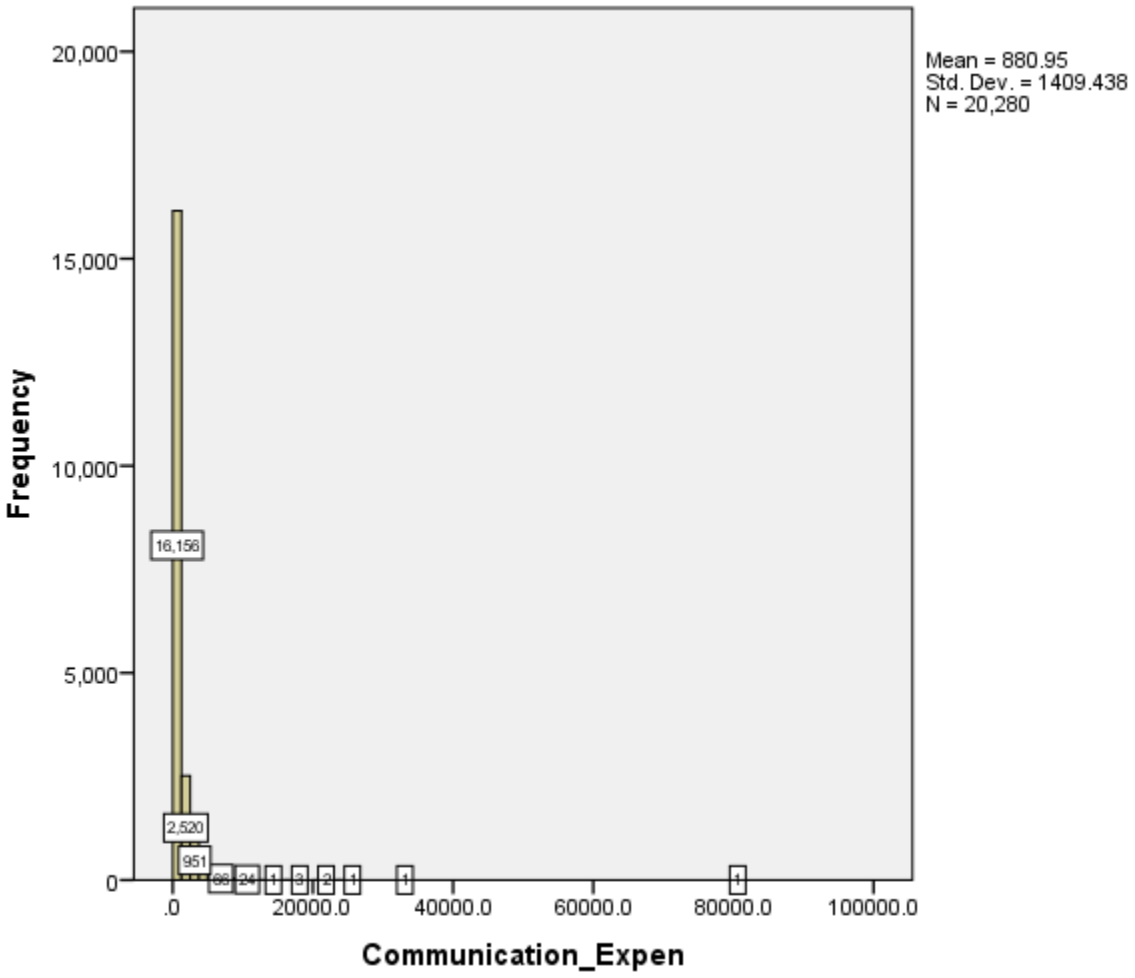


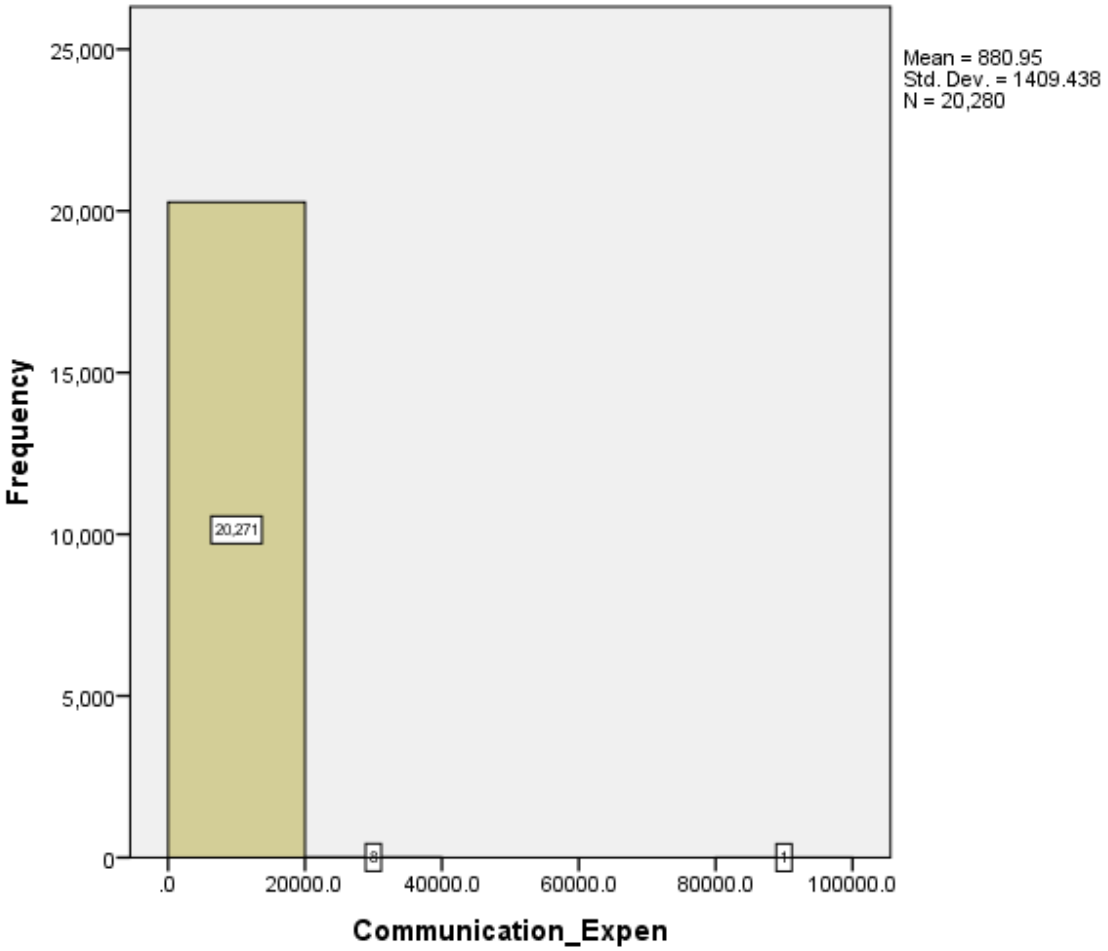
The minimum value for the food expenditure for household is 436.0 and the maximum value for the food expenditure is 4104636.0. The mean is 15618.11. The standard deviation is 29643.77. 99.9% of family's food expenditure value falls below 30000 and only one family exceed the limit that is 4s104636.0. Based on the graph, the majority of family's water bill value is close to the mean.

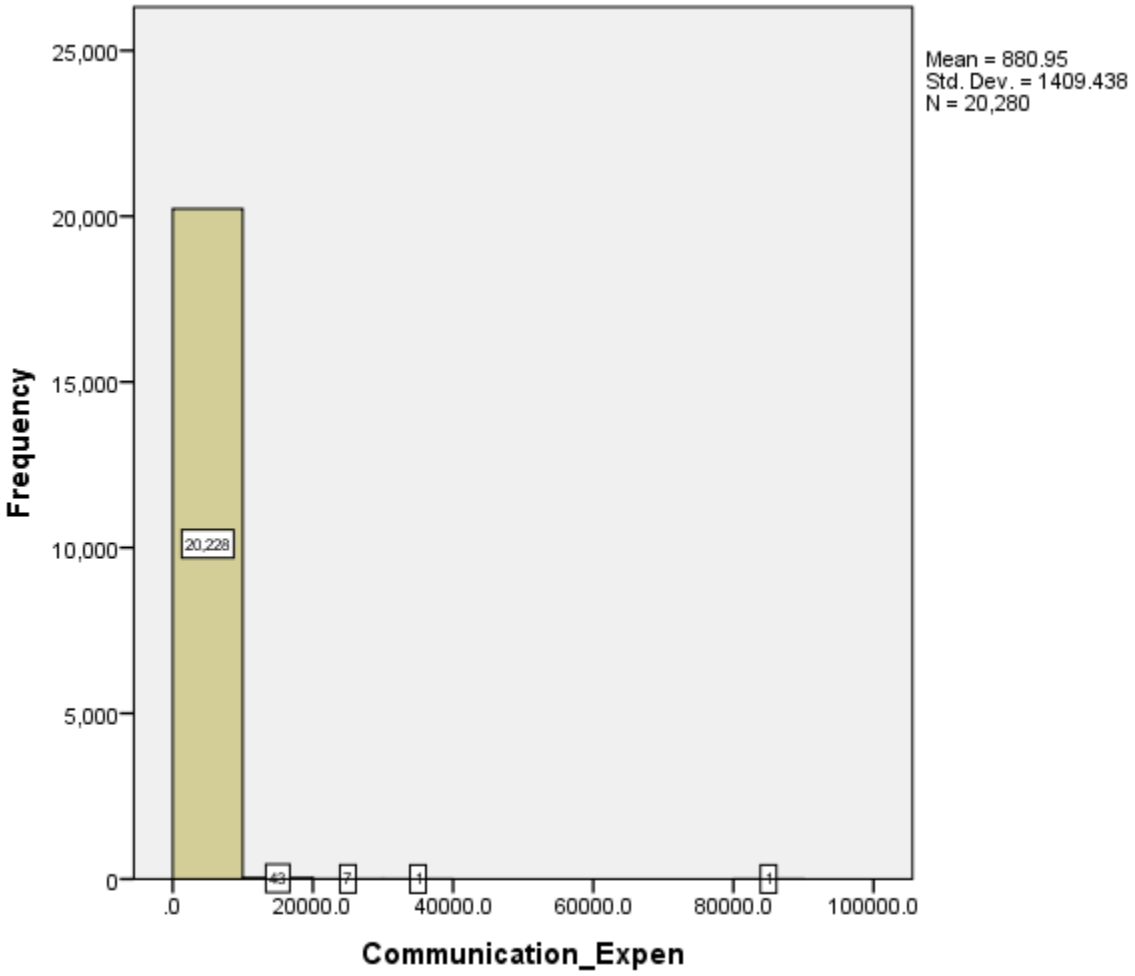
Communication Expenditure

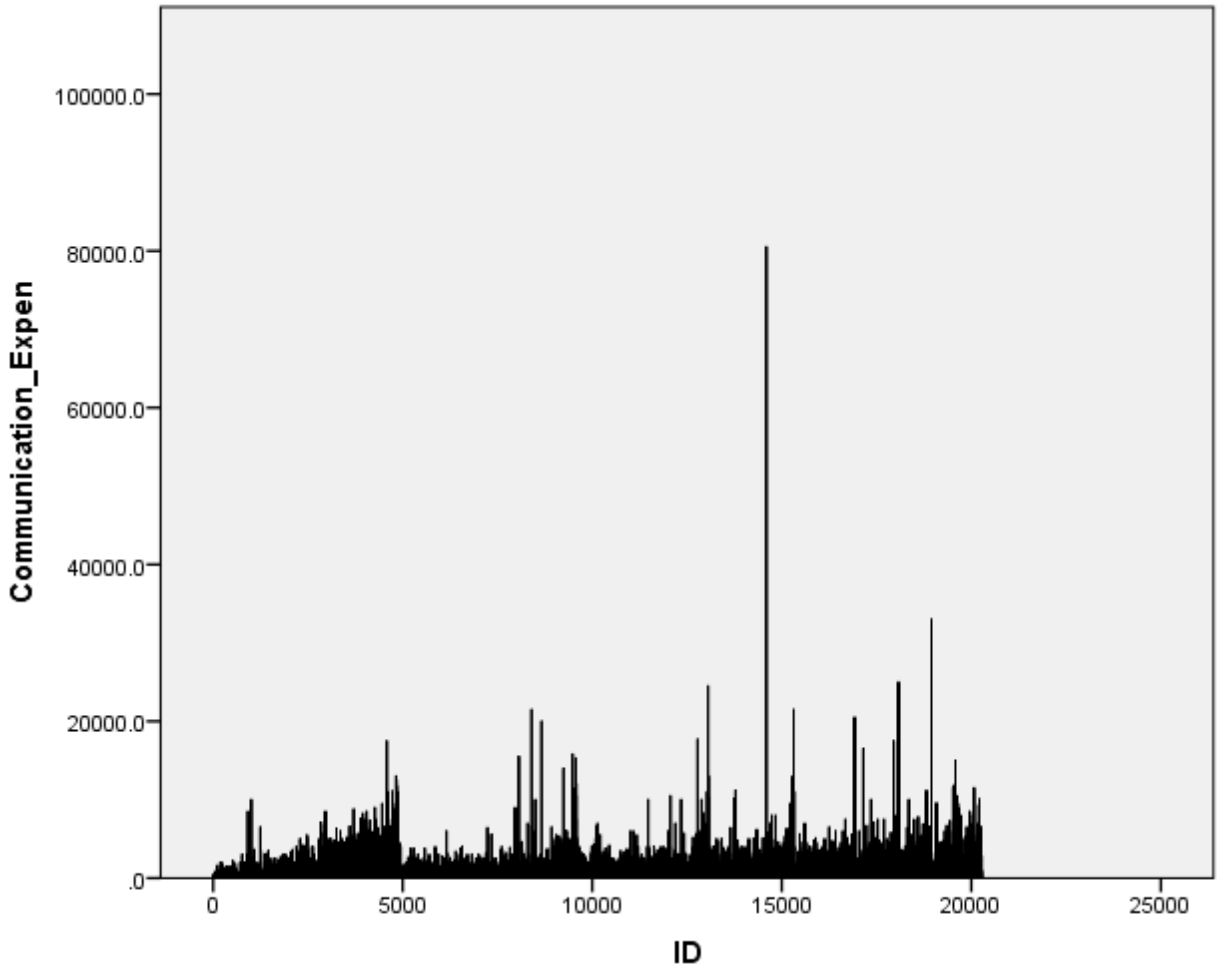
Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Communication_Expen	20280	.0	80500.0	880.946	1409.4383
Valid N (listwise)	20280				




Mean = 880.95
Std. Dev. = 1409.438
N = 20,280
Frequency
25,000
20,000
15,000
10,000
5,000
0
20,271
8
1
.0
20000.0
40000.0
60000.0
80000.0
100000.0
Communication_Expen


Mean = 880.95
Std. Dev. = 1409.438
N = 20,280
Frequency
25,000
20,000
15,000
10,000
5,000
0
20,228
43
7
1
1
.0
20000.0
40000.0
60000.0
80000.0
100000.0
Communication_Expen

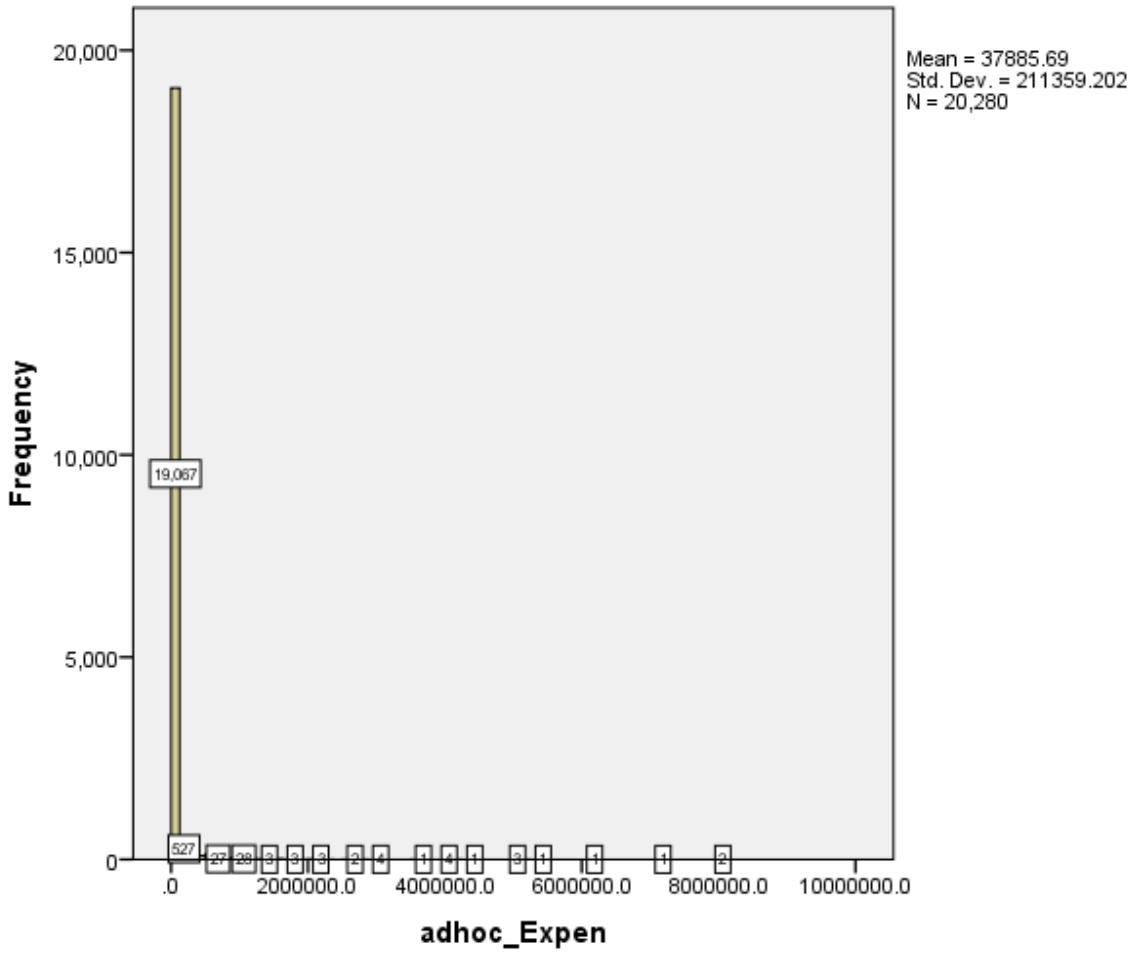


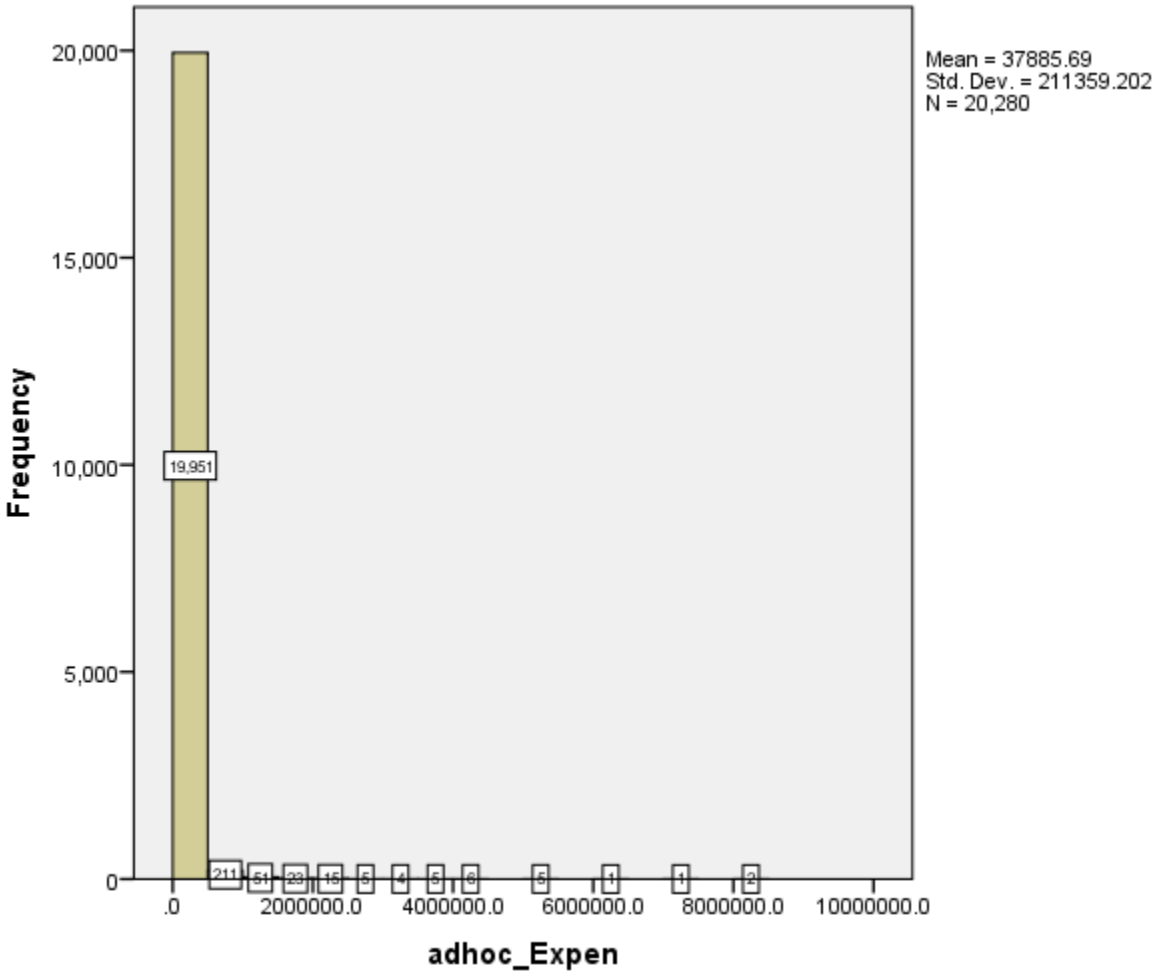
The minimum value for the communication expenditure is 0.0 and the maximum value for the communication expenditure is 80500.0. There is a huge difference between the minimum value and the maximum value. The mean is 880.95. The standard deviation is 1409.44. 99.74% of family's communication expenditure value falls below 10000 and 0.3% of family's value falls above 10 000. Minority of the family's (0.3%) values give significant different compared to the communication expenditure majority of family's values for the communication expenditure attribute.

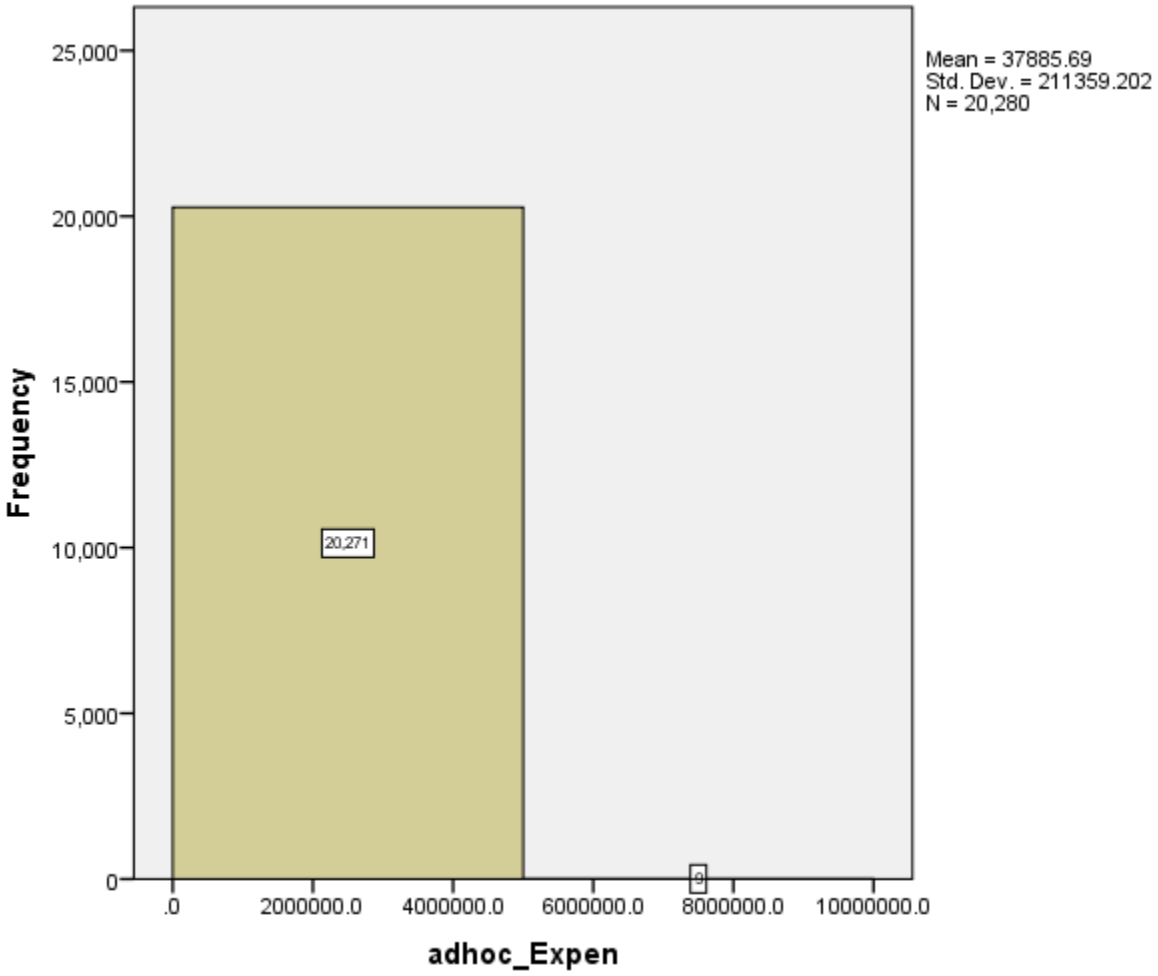
Adhoc_Expenses

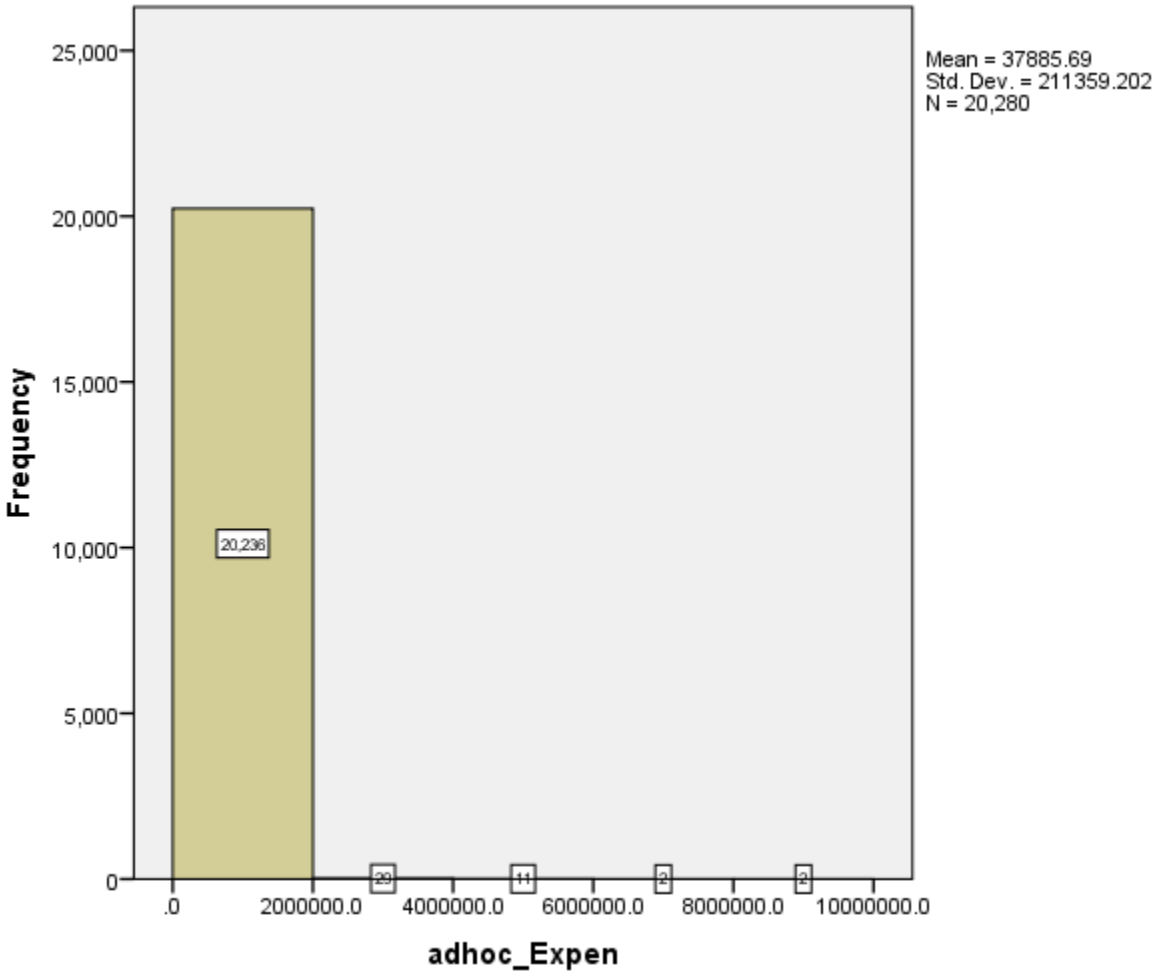
Descriptive Statistics

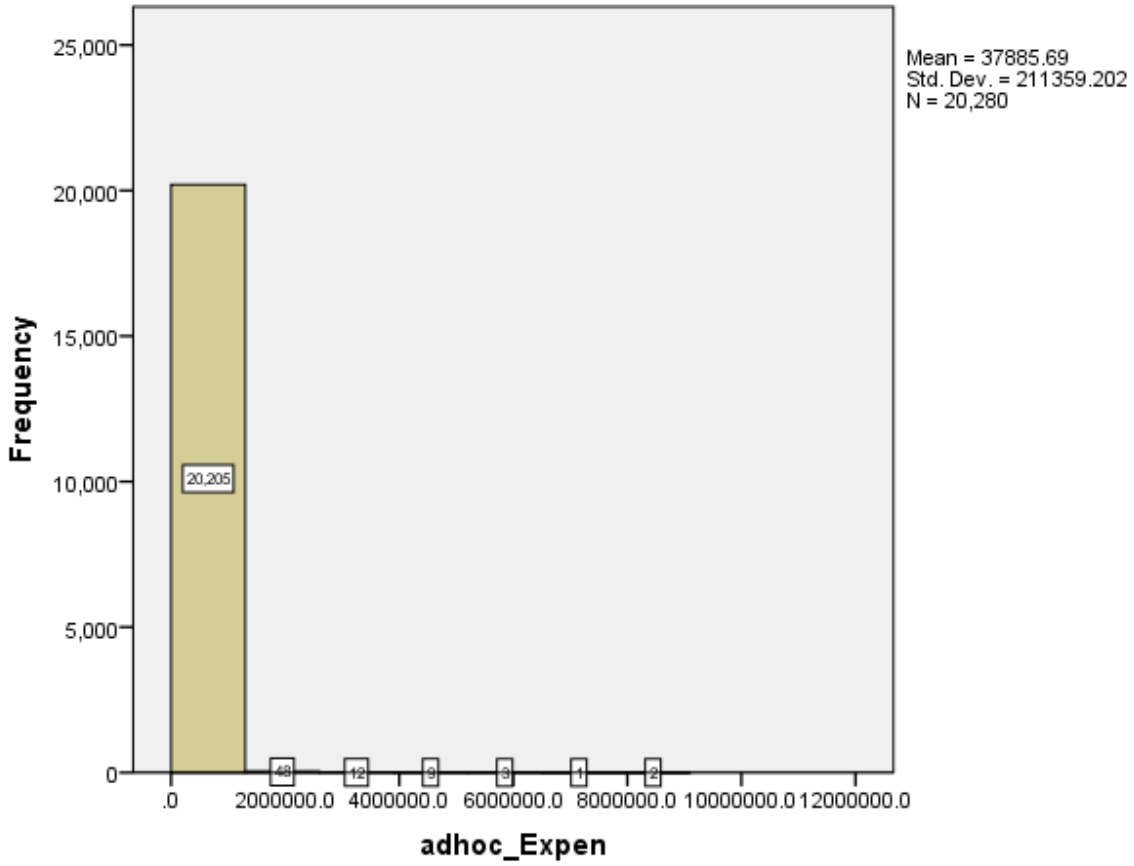
	N	Minimum	Maximum	Mean	Std. Deviation
adhoc_Expen	20280	.0	8017000.0	37885.685	211359.2020
Valid N (listwise)	20280				

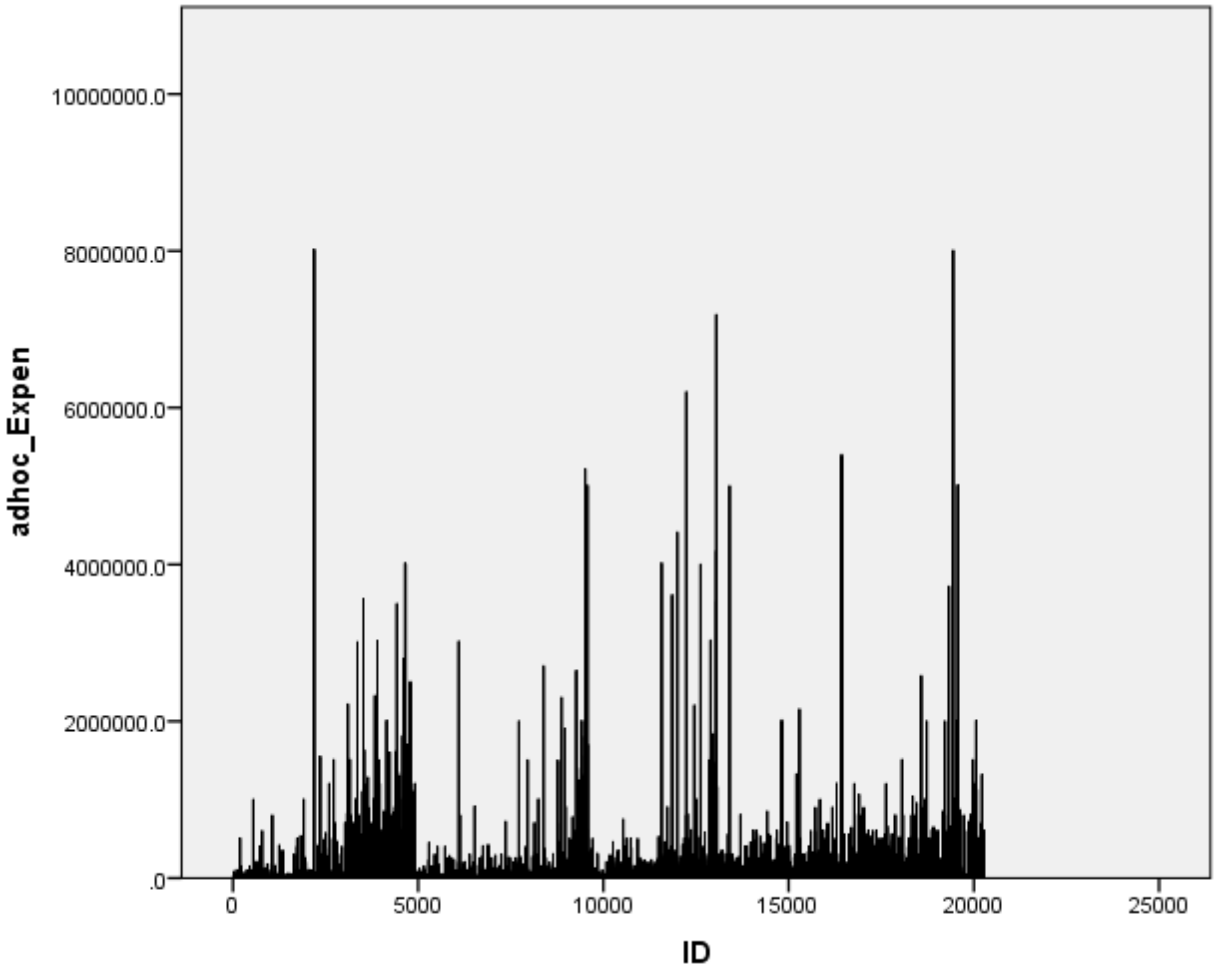



Mean = 37885.69
Std. Dev. = 211359.202
N = 20,280
Frequency
20,000
15,000
10,000
5,000
0
19,951
211
51
23
15
5
4
5
6
5
1
1
2
.0
2000000.0
4000000.0
6000000.0
8000000.0
10000000.0
adhoc_Expen


Mean = 37885.69
Std. Dev. = 211359.202
N = 20,280
Frequency
25,000
20,000
15,000
10,000
5,000
0
20,271
9
.0
2000000.0
4000000.0
6000000.0
8000000.0
10000000.0
adhoc_Expen


Mean = 37885.69
Std. Dev. = 211359.202
N = 20,280
Frequency
25,000
20,000
15,000
10,000
5,000
0
20,236
29
11
2
2
.0
2000000.0
4000000.0
6000000.0
8000000.0
10000000.0
adhoc_Expen


Mean = 37885.69
Std. Dev. = 211359.202
N = 20,280
Frequency
25,000
20,000
15,000
10,000
5,000
0
20,205
48
12
9
3
1
2
.0
2000000.0
4000000.0
6000000.0
8000000.0
10000000.0
12000000.0
adhoc_Expen

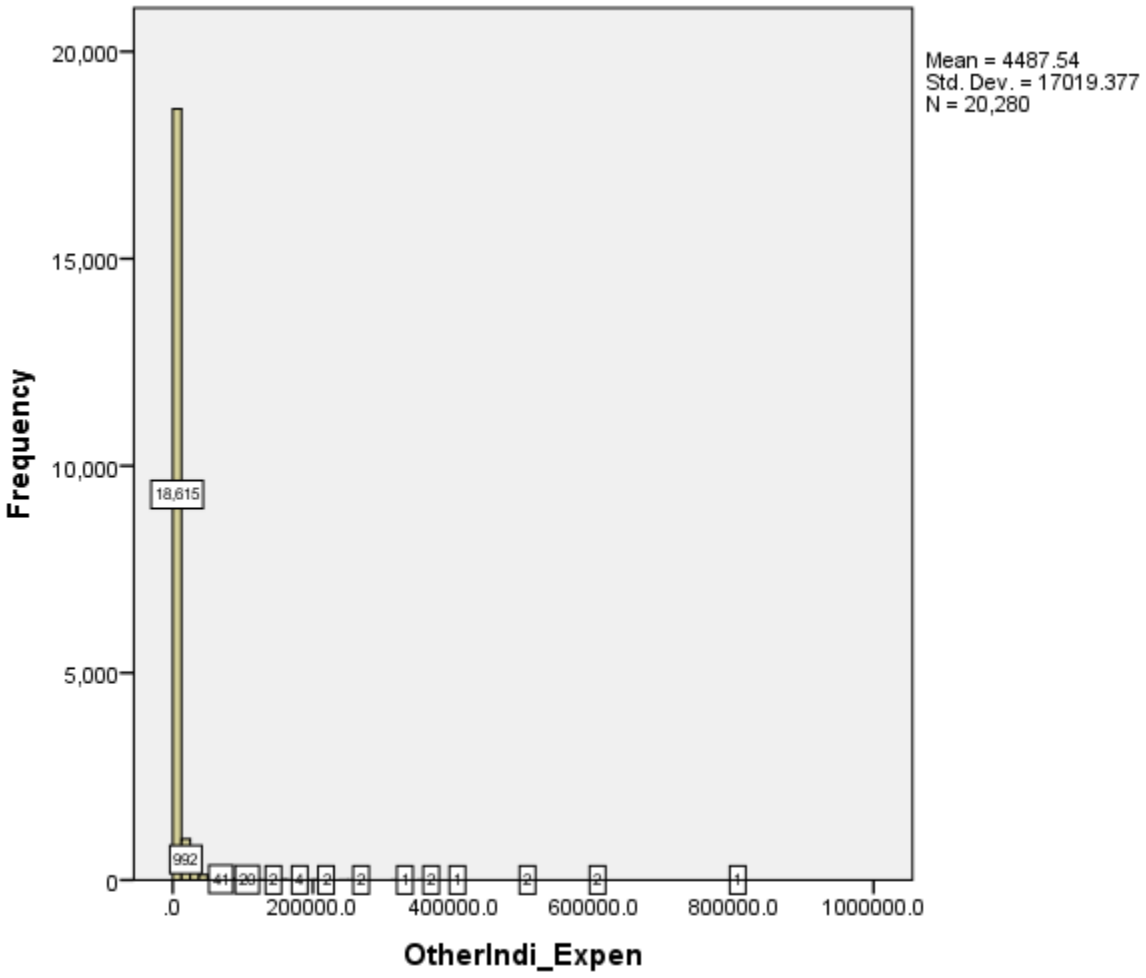


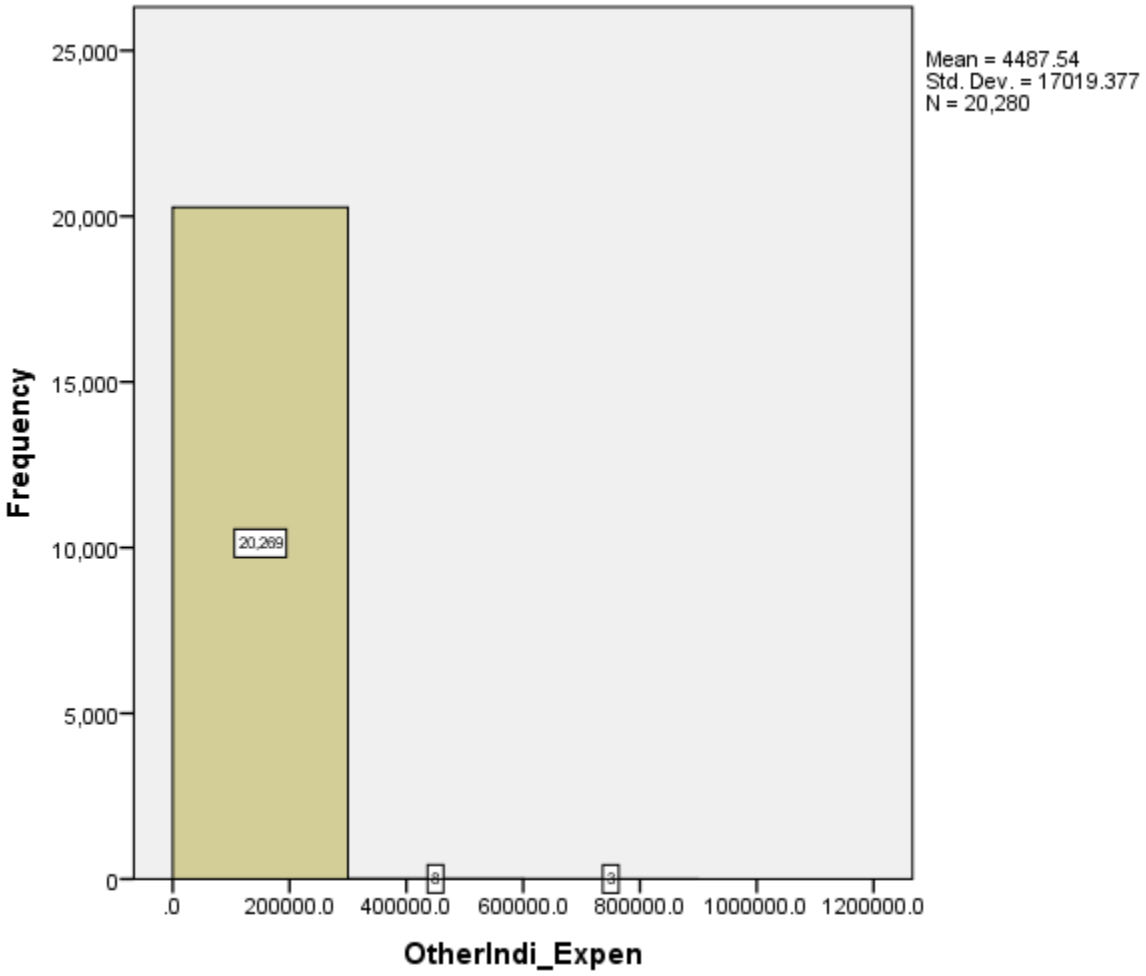
The minimum value for the adhoc expenses is 0.0 and the maximum value for the adhoc expenses is 8017000.0. There is a huge difference between the minimum value and the maximum value. The mean is 37885.68. The standard deviation is 211359.20. 99.43% of family's adhoc expenses value falls below 1000000 and 0.3% of family's value falls above 1000000 and below 2000000 and 0.2% of family's value falls above 2000000. Minority of the family's (0.2%) values give significant different compared to the majority of family's values for the adhoc expenses attribute.

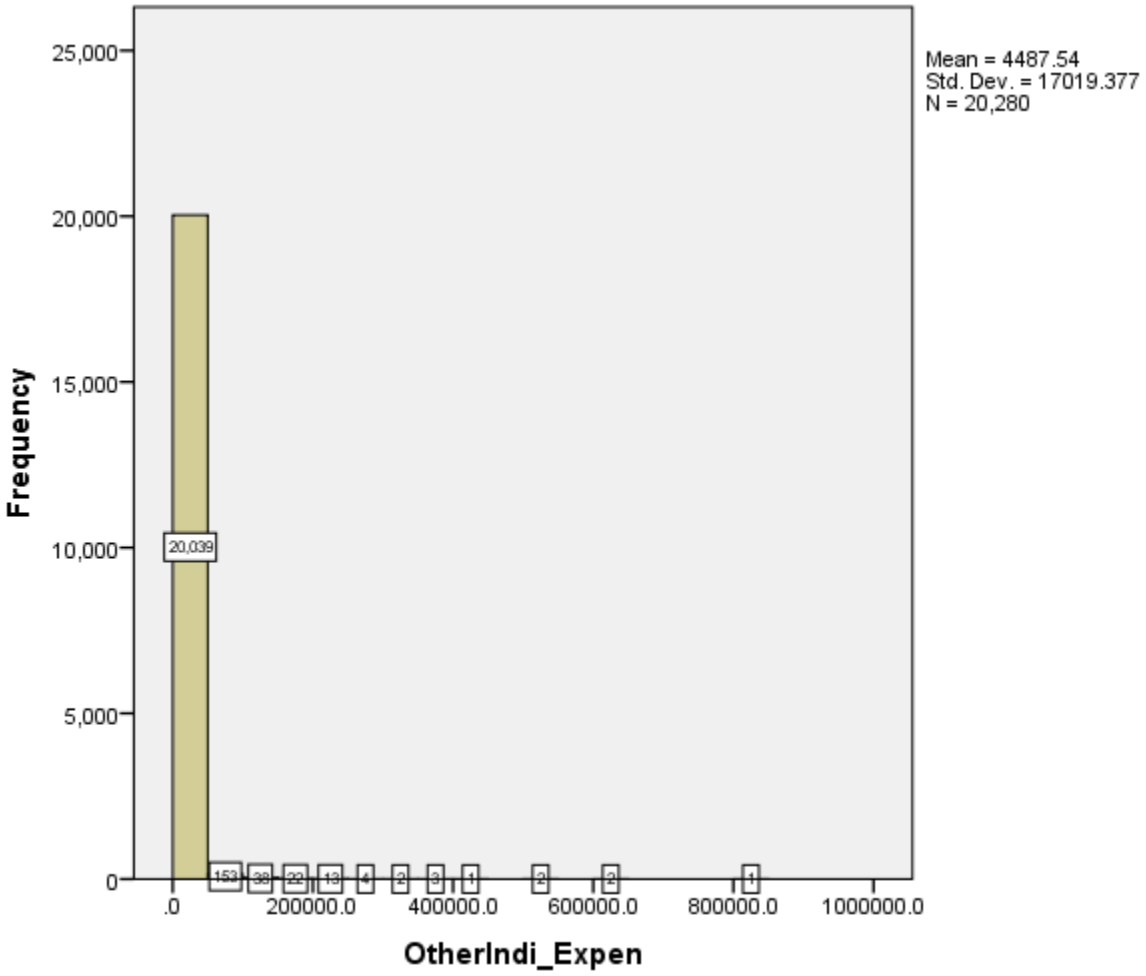
Other Individual Expenditure

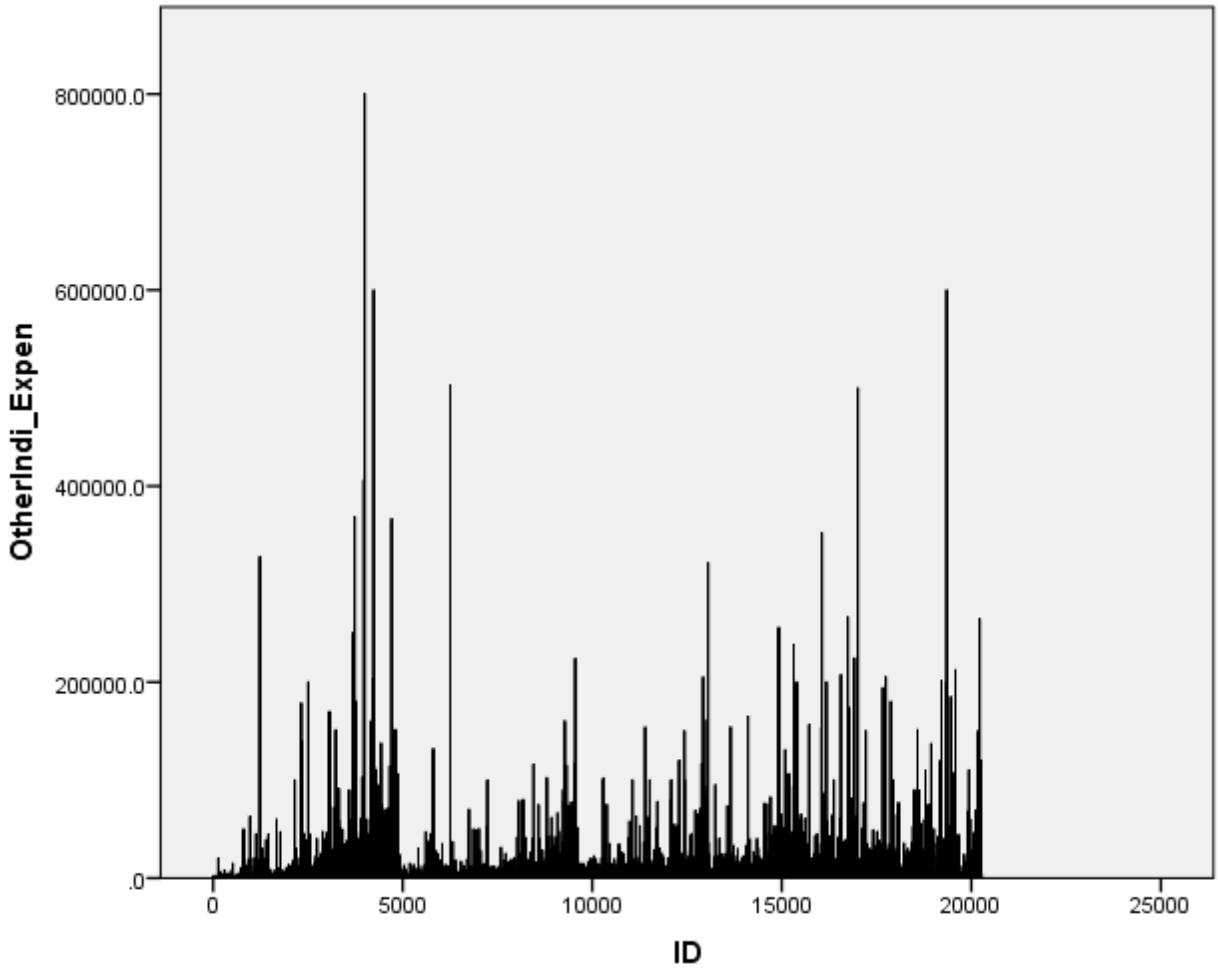
Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
OtherIndi_Expen	20280	.0	800000.0	4487.537	17019.3770
Valid N (listwise)	20280				




Mean = 4487.54
Std. Dev. = 17019.377
N = 20,280
Frequency
25,000
20,000
15,000
10,000
5,000
0
20,269
8
3
.0
200000.0
400000.0
600000.0
800000.0
1000000.0
1200000.0
OtherIndi_Expen


Mean = 4487.54
Std. Dev. = 17019.377
N = 20,280
Frequency
25,000
20,000
15,000
10,000
5,000
0
20,039
153
38
22
13
4
2
3
1
2
2
1
.0
200000.0
400000.0
600000.0
800000.0
1000000.0
OtherIndi_Expen

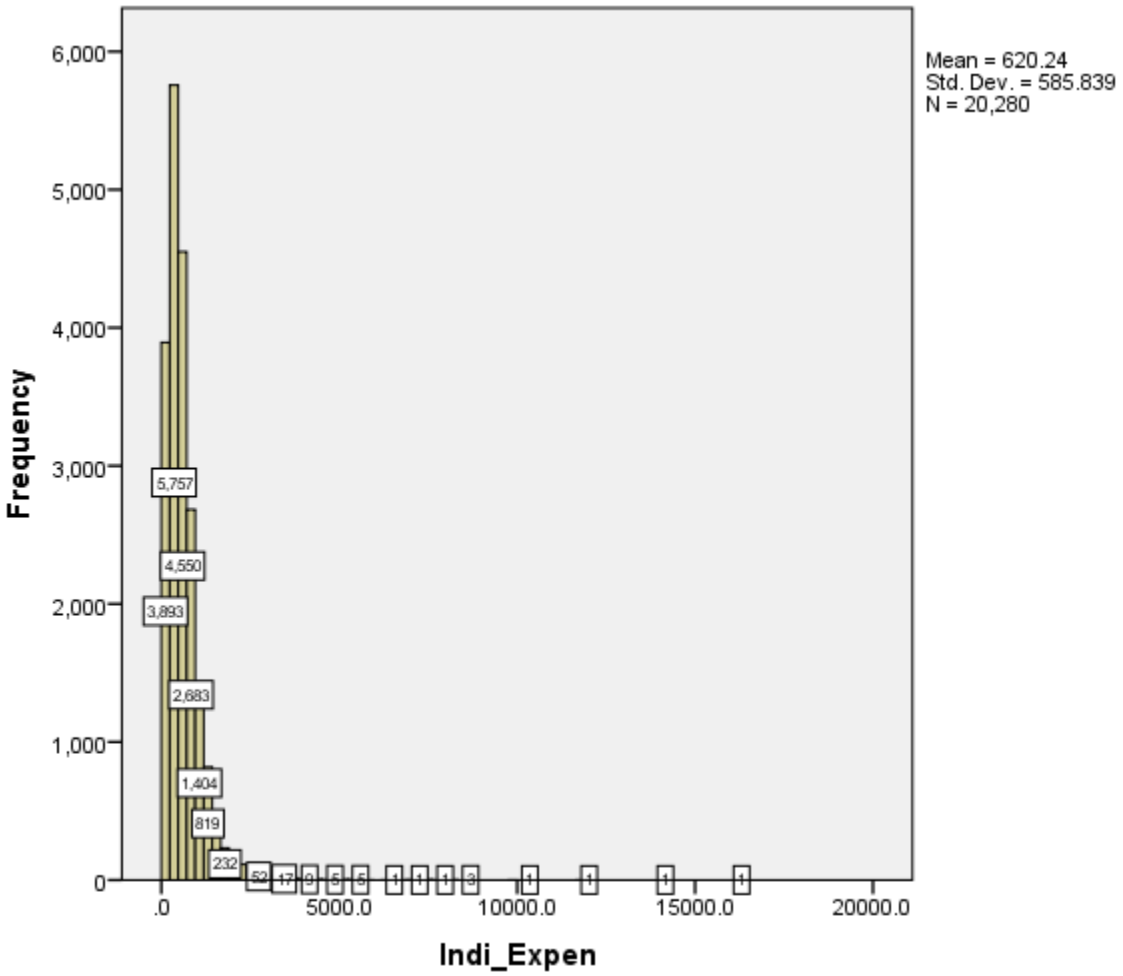


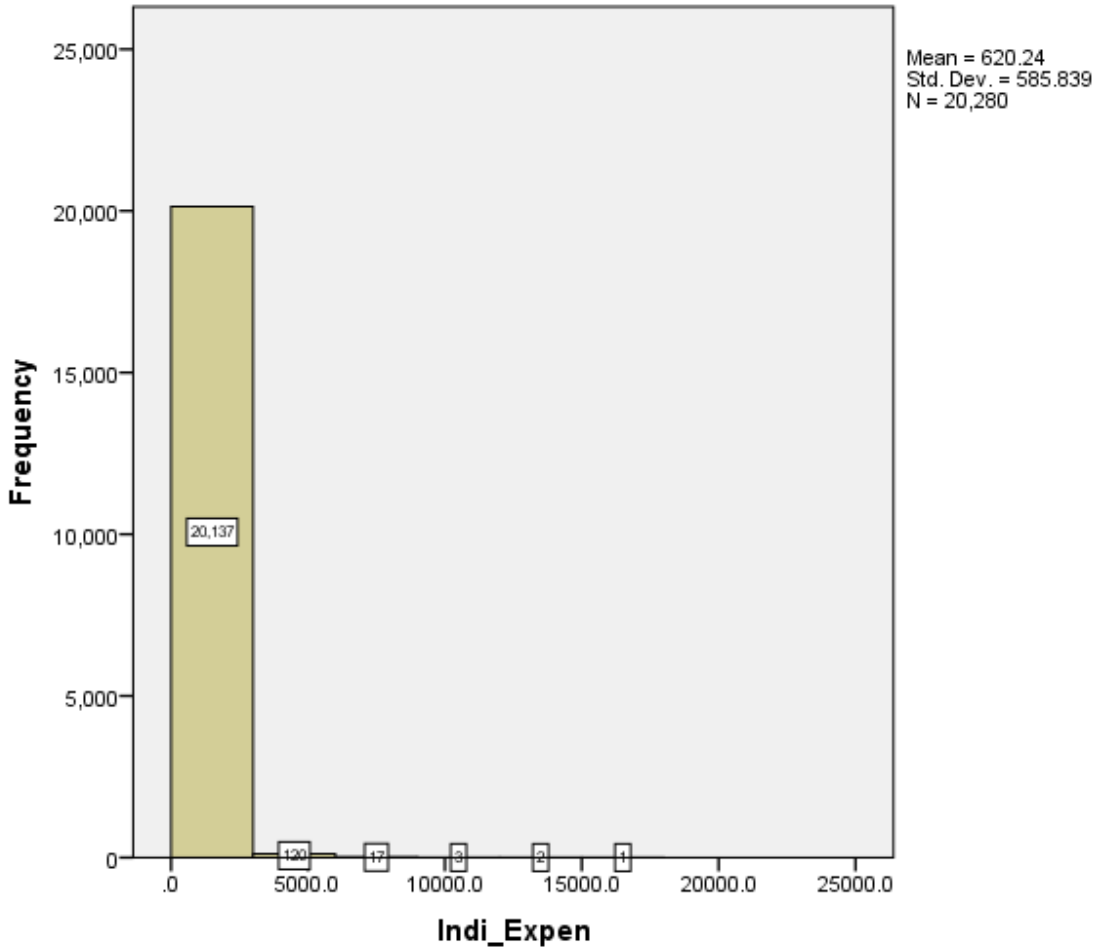
The minimum value for the other individual expenses is 0.0 and the maximum value for the other individual expenses is 800000.0. There is a huge difference between the minimum value and the maximum value. The mean is 4487.54. The standard deviation is 17019.38. 99.94% of family's other individual expenses value falls below 200000 and 0.06% of family's value falls above 200000. Minority of the family's (0.06%) values give significant different compared to the majority of family's values for the other individual expenses attribute.

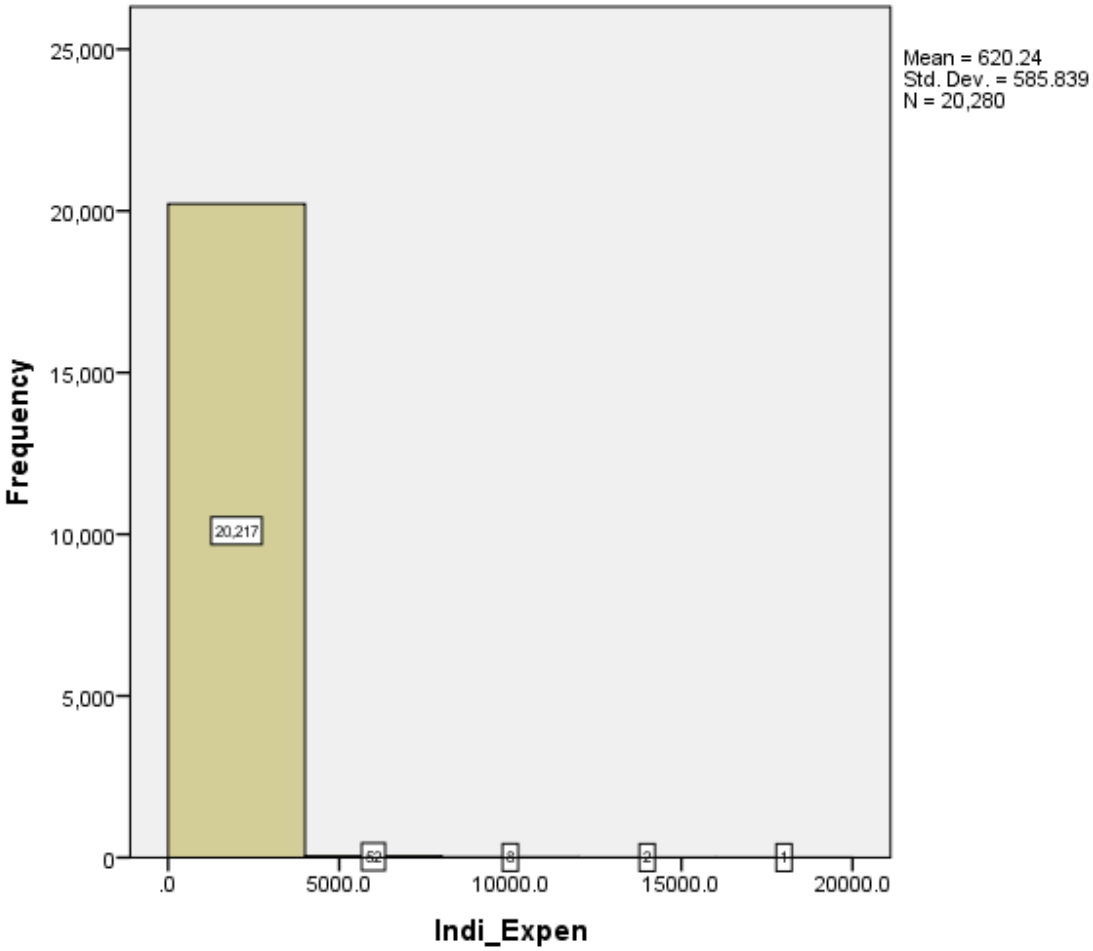
Individual expenses

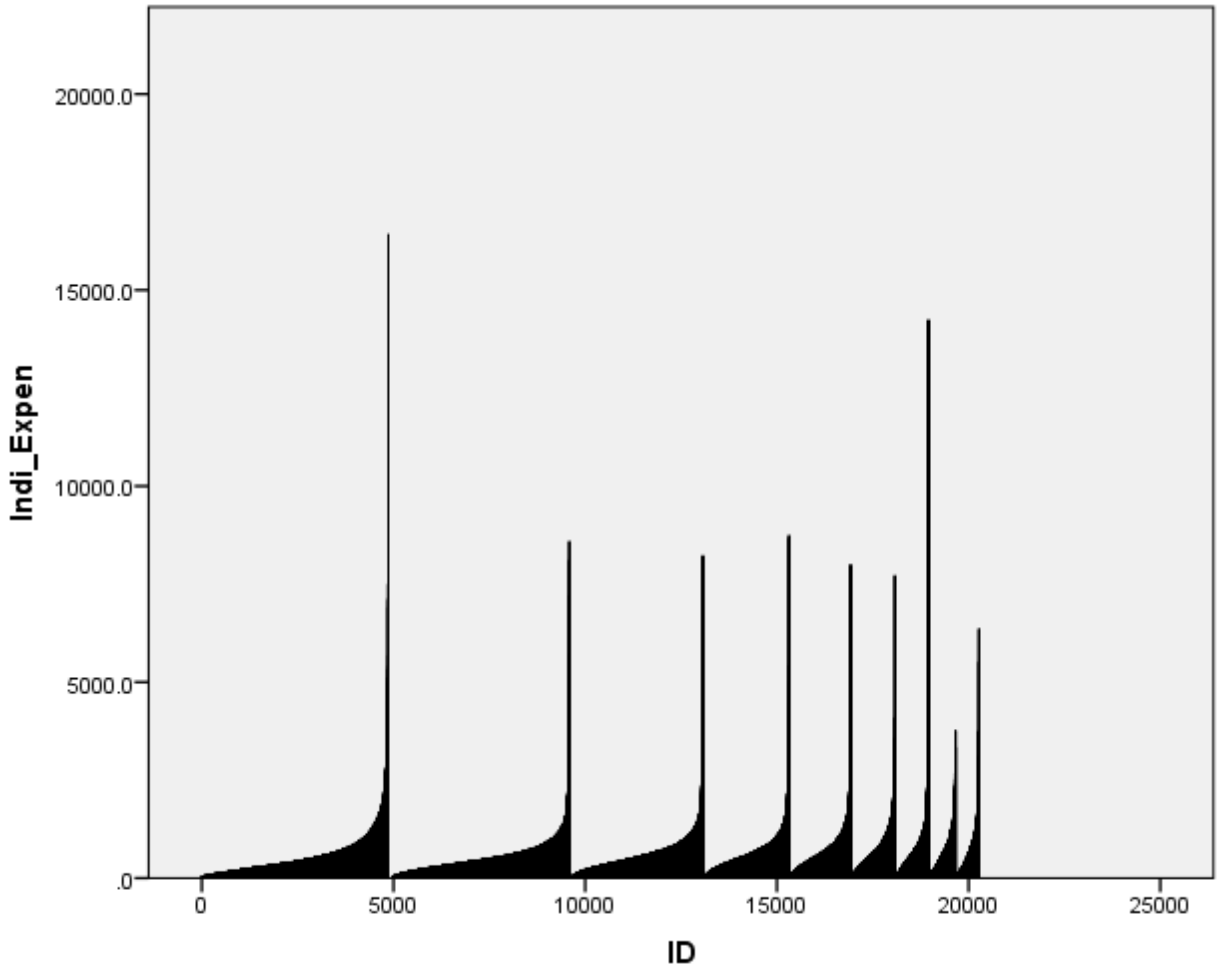
Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
ndi_Expen	20280	.0	16410.0	620.242	585.8387
Valid N (listwise)	20280				




Mean = 620.24
Std. Dev. = 585.839
N = 20,280
Frequency
25,000
20,000
15,000
10,000
5,000
0
20,137
120
17
3
2
1
.0
5000.0
10000.0
15000.0
20000.0
25000.0
Indi_Expen


Mean = 620.24
Std. Dev. = 585.839
N = 20,280
Frequency
25,000
20,000
15,000
10,000
5,000
0
20,217
62
8
2
1
.0
5000.0
10000.0
15000.0
20000.0
Indi_Expen



The minimum value for the individual expenses is 0.0 and the maximum value for the individual expenses is 16410.0. The mean is 620.24. The standard deviation is 585.84. 99.80% of family's individual expenses value falls below 5000 and 0.2% of family's value falls above 5000. Minority of the family's (0.2%) values give significant different compared to the majority of family's values for the individual expenses attribute.

3.3 Chapter summary

This chapter has described the data collection from real world and selected the suitable dataset and converting dataset data to needed formats. Algorithms and tools selection for execution and analysis of selected dataset data attributes and preprocessing the data are further explained in this chapter. Using the descriptive analysis income, expenditure, adhoc expenditure and food expenditure attributes are helps to identify suspicious records. The following chapter will cover the next step: how to execute those algorithms in dataset and the results.

Chapter 4

4 Evaluation and Results

The Department of Census and Statistics gives accurate results based on collected data to the country. This department conducts surveys to collect the relevant information from people. Sometimes some people give incorrect information. When they analyze the data with wrong information they won't be able to give accurate results. The clustering method is proposed to identify the suspicious records in the House hold income and expenditure survey. K-means, model base method and Hierarchical clustering methods have been used on House hold income and expenditure (2012/2013) data set to cluster the data into normal records and suspicious records.

The analysis of this dataset occurs with the help of R programming using K-means. R studio and R should be installed in the system for this purpose. After this is complete, load the dataset in the CSV format in R studio. K is assigned a value of 2 and the data is clustered using K-means. Then the center point of each cluster is calculated and then each data point is assigned to the relevant cluster. All the data points are then labeled cluster wise.

The hierarchical clustering method is used for label the data. The data set is imported into R studio and then the hierarchical clustering is run using hclust. Hclust requires the data to be provided in the form of a distance matrix. This can be done only using dist. By default, the complete linkage method is used. When the clustered data points are plotted into dendrogram, to store the record in RAM a minimum of 8GB is needed.

The model -based clustering also used for label the data. The data set is imported into R studio and then the model-based clustering is run using mclust. Clustered the data set into 2 classes and labeled.

4.1 Result of clustering using K-Means

The K value is selected as 2 and the real data set is run in R using K-means. It divides 20280 records, resulting in 20212 records in one cluster and 68 records in another. When analyzing the cluster with 68 records the adhoc expenditure is more for all data. Calculate the mean value for each cluster and find the difference for which attributes are most different from the

dataset. And draw the boxplot for each cluster to identify the outliers. The adhoc expenditure include expenditure due to weddings /funerals for family members, Social activities / Ceremonies, Gift, Donation, Similar transfers or Purchased properties. When these values are more than 1,500,000 it goes to one cluster. When analyzed those data with demographic and non food expenditure data they spend that much of money for a reason. They spent for their family member's wedding, for social activities. So we didn't take those as suspicious records. But some records they spent huge amount for adhoc expenses like 8,017,000 those are identified as suspicious records.

Field	Mean of Cluster 1	Mean of Cluster2	Difference (2-1)
Household Income Per Month	44623.77	252912.7	208288.93
Household expenditure Per Month	40466.94	355005.2	314538.26
Electricity bill	750.8228	2044.603	1293.7802
Water bill	158.7814	448.5	289.7186
Food Expenses	15593.48	22940.12	7346.52
Communication Expenses	873.5224	3087.618	2214.0956
Adhoc Expenses	28479.25	2833810	2805330.75
Other individual expenses	4368.651	39824.82	35456.169
Individual expenses	618.2361	1216.529	598.2929

Table 4: Mean value for 2 Clusters

The mean value is more different in Income per month, Expenditure per month, Adhoc expenses and other individual expenses.

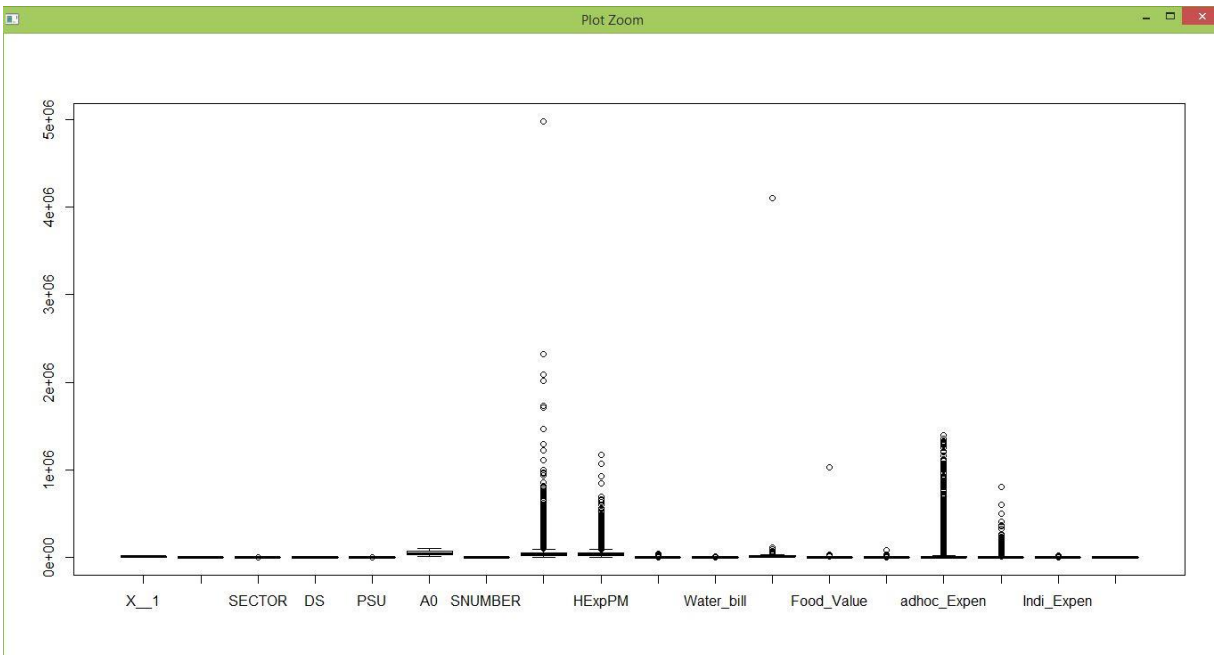


Figure 13 Boxplot for normal data

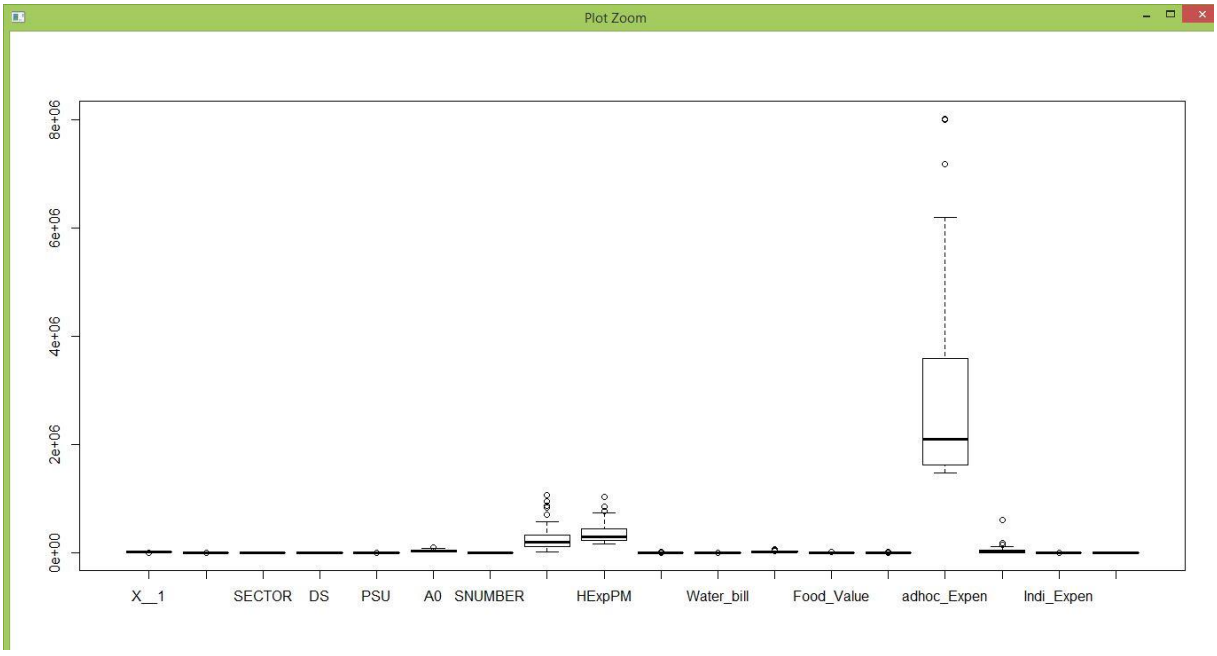


Figure 14: Boxplot for suspicious records

After remove the outlier using IQR method and apply the K-means and cluster the data set into two groups. It gives 20052 records in one group and 54 records in another group.

The adhoc expenses (wedding) are more than 1395000 as suspicious records with confirmation.

4.2 Result for clustering using Hierarchical clustering

When the data set is run using hierarchical clustering, a dendogram is obtained. In hierarchical clustering, the clustering can be stopped at any desired level (or clusters) using cutree function.

When run the Hierarchical clustering for HIES data set it produce the dendrogram. Because of the large data set it will not produce the clear dendrogram. The below image represent that dendrogram.

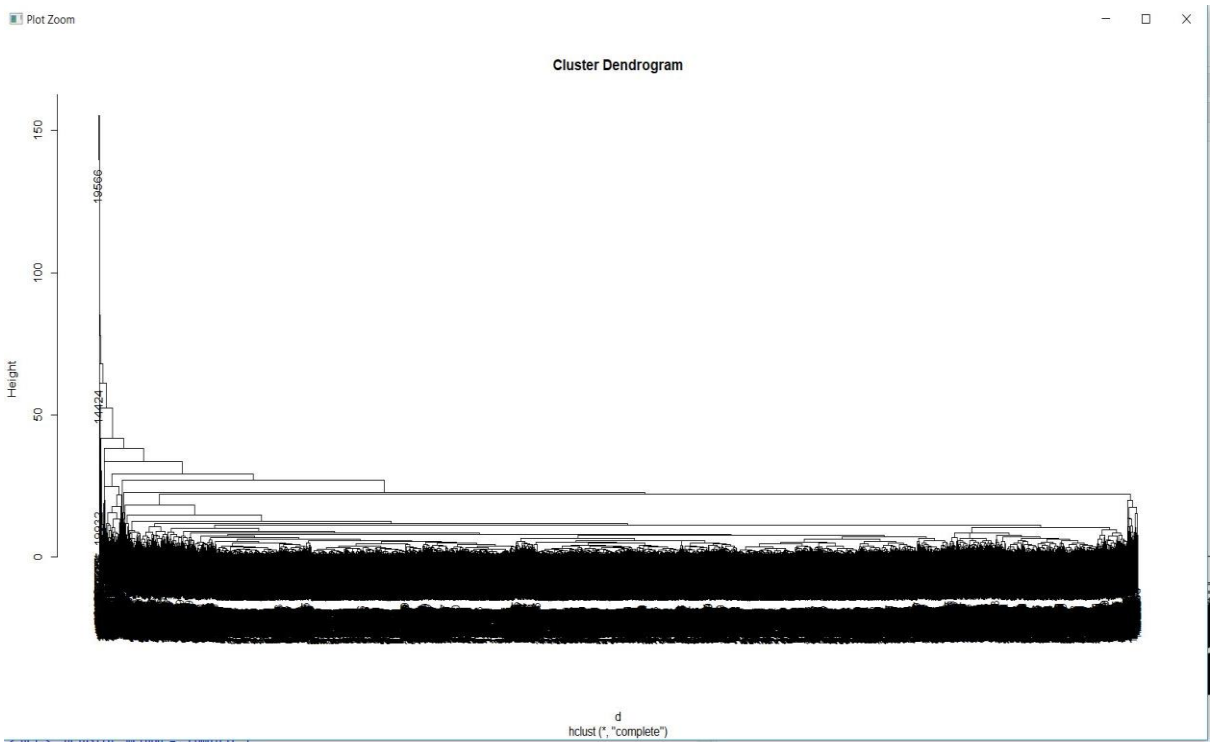


Figure 15 Hierarchical for real dataset

4.3 Result of Model base clustering

When the dataset was run using Model base clustering with two classes, 2516 records goes to one cluster and 17 764 records goes to another cluster. 2516 records include high adhoc expenses. These records were already clustered into one class by k-means. And, high expenses with zero adhoc are also included in the 2516 records.

B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
DISTRICT	SECTOR	DS	PSU	A0	SNUMBER	HIncPM	HExpPM	Electricity	Water_bil	Food_Exp	Food_Val	Communi	adhoc_Expen	OtherIndi	Indi_Expe	cluster
82	2	10	15	89521	3	163844	11267.38	0	0	4364	1091	400	0	0	30	2
32	2	10	16	42426	5	18906.67	14703.45	65	0	4788	1197	0	104000	0	55	2
82	2	10	31	89937	9	129126.2	65133.71	625	350	10804	2701	500	0	20040	65	2
11	2	10	68	12988	7	156588.6	80481.64	650	110	15872	3968	550	508000	6100	75	2
71	2	10	2	79196	3	11204.83	21072.95	500	0	6508	1627	0	150000	0	80	2
81	2	10	12	86789	5	18668.33	21202.27	75	60	8308	2077	200	100000	50	120	2
82	2	10	31	89937	8	171288.3	25628.77	210	0	9520	2380	300	150000	60	120	2
44	2	10	34	56368	10	154016.7	12660.04	0	0	8760	2190	250	150	0	120	2
33	2	10	31	46721	10	185454.3	14533.12	150	150	8844	2211	400	3000	0	120	2
91	2	10	60	94407	3	186921.4	14287.07	100	0	8088	2022	100	1000	0	130	2
41	2	10	10	48421	2	13183.81	101904.8	400	0	8268	2067	300	1000500	0	135	2
72	2	10	61	83430	10	143070	50397.63	1400	0	30636	7659	500	7000	0	140	2
81	2	10	82	86747	7	13836.91	35289.61	100	180	10572	2643	200	211000	400	150	2
72	2	10	29	83605	7	147453.6	34810.45	676	398	13556	3389	250	0	1887	152	2
82	2	10	31	89937	7	181619	44542.34	650	275	18428	4607	1000	1300	10240	160	2
81	2	10	14	87637	4	158264.9	62804.17	300	150	13900	3475	800	404000	0	160	2
31	2	10	6	39800	4	17923.81	43207.96	360	0	18092	4523	675	120000	0	169	2
45	2	10	31	58435	9	19093.1	57510.55	0	0	6680	1670	0	600000	0	170	2
31	1	10	77	40333	10	17081.9	28618.4	236	537	9208	2302	0	127000	50	173	2
33	2	10	64	45193	10	131421.9	65478.34	0	0	8608	2152	1000	30000	50000	173	2
62	2	10	63	77037	7	148498.6	43749.99	1500	500	18992	4748	2000	26500	150	180	2
22	2	10	51	33370	7	134764.3	29478.1	200	0	7144	1786	1000	20000	1558	182	2
23	3	10	32	36636	2	172092.4	25544	0	0	18968	4742	0	2500	950	184	2
32	2	10	95	42132	3	18262.38	36276.16	179	0	7652	1913	100	123000	5500	185	2

Figure 16 Output of Model Based Clustering

Those data were analyzed with demographic, non food expenditure and food expenditure. In some cases, the total family member is 2 or 3, but their income is more than 400,000. These are unbelievable. The records which have higher expenditure than income and the records which have high income with very lower expenditure for food expenses and electricity bill were identified as suspicious records.

4.4 Chapter summary

This chapter has described the clustering methods which are used in research and how these algorithms are executed among the dataset and has presents experimental test observations and results acquired by using clustering methods executions.

Chapter 5

5 Conclusion and Future Work

5.1 Conclusion

After analyzing the Household income and expenditure data set with K-means, hierarchical clustering method and model base clustering method, for K-means the number of clusters have to be defined initially but it takes very less time to cluster a large dataset. But the hierarchical method automatically clusters the dataset in to groups but it takes some time to group data and more memory is needed to run this algorithm. When the adhoc expenses is more without unbelievable reason and expenditure is higher than the income for a family is identify as suspicious records using K-means. When using the model base clustering it identify the suspicious records that already identified using K-means and the income and expenditure different is higher than normal, when food expenditure is high or low compare with income. Those types of records identified as suspicious records from the household dataset.

For further confirmation checked with the demographic data those are how many family member in a house, if they spend for wedding check the age group, when the water bill is 0 then they use well water. Finally got some suspicious records from the data set based on these attributes.

5.2 Future work

The current data set analysis is based on sections such as food expenditure, communication expenditure etc. Instead, if smaller segments of these are considered as the attributes itself such as food expenditure for meat, food expenditure for fish etc, the identification of suspicious records would be more accurate. And this dataset is collected in 2012/2013. When collect the next survey to analyze can get the accurate suspicious record.

Reference

[1] “Report to the nations on occupational fraud and abuse,” [Online]. Available: http://www.acfe.com/rttn-introduction.aspx. [Accessed Mar. 20,2017].

[2] “Data science.com” [Online]. Available: https://www.datascience.com/blog/supervised-and-unsupervised-machine-learning-algorithms. [Accessed May. 21,2017].

[3]”SAS The power to know” [Online]. Available: https://www.sas.com/en_us/insights/articles/risk-fraud/fraud-detection-machine-learning.html. [Accessed Sep. 15, 2017].

[4] “Research Methodology”.[Online]. Available: https://research-methodology.net/research-methods/data-collection. [Accessed Jan. 20 2018].

[5] R-bloggers. (2018, Feb, 2). Retrieved from https://www.r-bloggers.com/hierarchical-clustering-in-r-2/

[6]”nonlinear”,[Online]. Available : http://www.nonlinear.com/support/progenesis/comet/faq/v2.0/dendrogram.aspx. [Accessed Feb. 15,2018].

[7] Y. Sahin & E. Duman, “Detecting credit card fraud by decision trees & support vector machines,” *Proceeding of the International MultiConfrence of Engineers & Computer Scientist, vol. I, 2011.*

[8] Ahamed Shafeeq B M and Hareesha K S, "Dynamic Clustering of Data with Modified K-Means Algorithm",2012.

[9] Durgesh K. Srivastava, Lekha Bhambhu, “Data Classification Using Support Vector Machine”,2009.

[10] P. Ravisankar, &V. Ravi, & G. Raghava Rao & I. Bose, ”Detection of financial statement fraud and feature selection using data mining techniques*”*.

[11] Vaishali, “Fraud Detection in Credit Card by Clustering Approach,” *International Journal of Computer Applications (0975 – 8887),* vol.98, 2014.

[12] Andrei Sorin SABAU, "Survey of Clustering based Financial Fraud Detection Research", *Informatica Economică,* vol 16, 2012.

[13] Oyelade, O. J &Oladipupo, O. O &Obagbuwa, I. C, "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance", *(IJCSIS) International Journal of Computer Science and Information Security*, vol 7, 2010.

[14] Manpreet kaur &Usvir Kaur, "Comparison Between K-Mean and Hierarchical Algorithm Using Query Redirection", *International Journal of Advanced Research in Computer Science and Software Engineering,* vol 3, 2013.

[15]Adrian Banarescu,"Detecting and Preventing Fraud with Data Analytics", *Procedia Economics and Finance,* 2015.

[16]"Statistical Consultants Ltd "[Online].Available: http://www.statisticalconsultants.co.nz/blog/benfords-law-and-accounting-fraud-detection.html. [Accessed June. 28, 2018].

Appendix A: Output using K-means and Hierarchical Clustering

K-Means

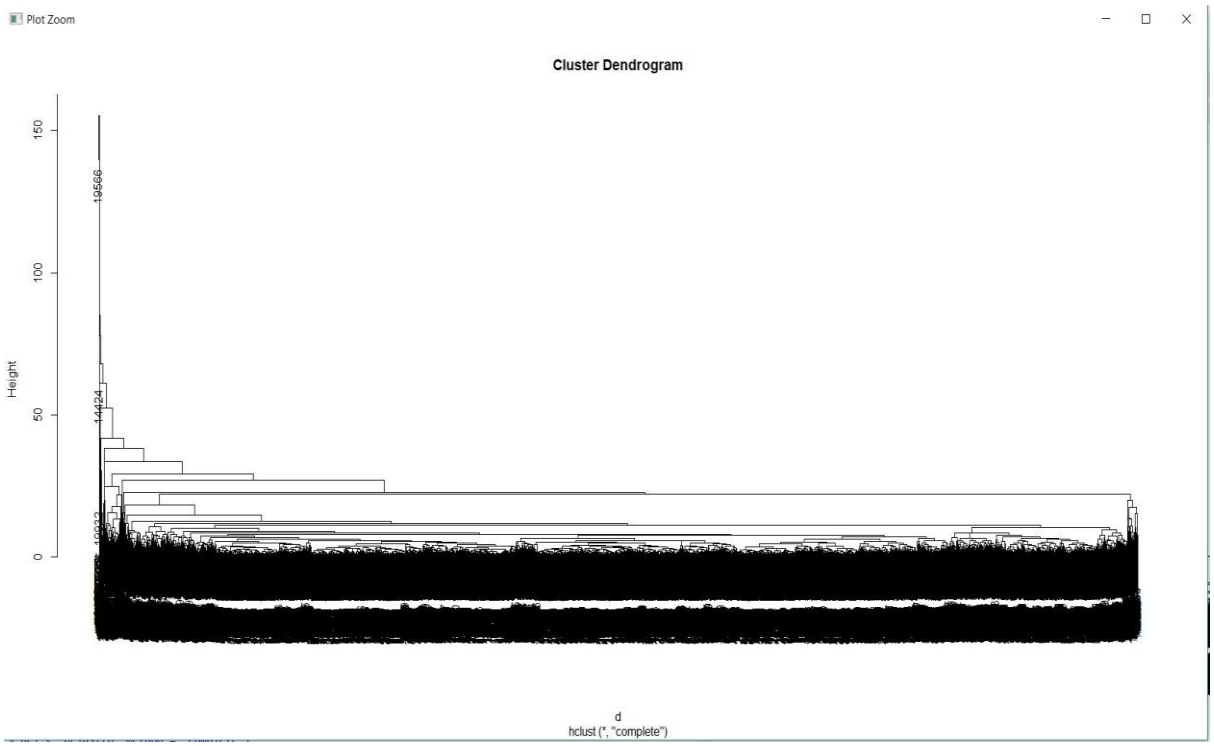
```
> library(readxl)
> HIES <- read_excel("F:/Luck/research/Datasets/HIES.xlsx")
> View(HIES)
> head(HIES)
# A tibble: 6 x 15
  DISTRICT SECTOR    DS   PSU    A0 SNUMBER    HIncPM    HExpPM Electricity_bill Water_bill Food_Value
     <dbl>  <dbl> <dbl> <dbl> <dbl>   <dbl>     <dbl>     <dbl>            <dbl>       <dbl>      <dbl>
1       32      2    10     3 42559       3  1467.143  1411.143               60           0        109
2       22      1    10    15 33040      10 30085.710 61056.910             2500        1200       9033
3       91      2    10    58 92811       5 11535.710 10084.100              249           0       1239
4       62      2    10    25 76982       2 16048.670 12305.890                0           0       2331
5       32      2    10    86 42457       5 28367.990 15560.340              300           0       3004
6       31      2    10    95 38173       9  8394.810  7561.476               76           0       1460
# ... with 4 more variables: Communication_Expen <dbl>, wedding_Expen <dbl>, OtherIndi_Expen <dbl>,
#   Indi_Expen <dbl>
> k.means.fit <- kmeans(HIES, 2) # k = 2
> attributes(k.means.fit)
$names
[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"
[8] "iter"         "ifault"

$class
[1] "kmeans"

> k.means.fit$centers
  DISTRICT   SECTOR DS      PSU       A0  SNUMBER    HIncPM    HExpPM Electricity_bill Water_bill Food_Value
1 29.95588 1.573529 10 81.75000 39319.19 5.308824 252912.75 355005.21        2044.6029   448.5000   5735.029
2 40.91327 1.840936 10 68.10375 49990.25 5.466802  44623.77  40466.94         750.8228   158.7814   3898.369
  Communication_Expen wedding_Expen OtherIndi_Expen Indi_Expen
1           3087.6176    2833810.29        39824.824  1216.5294
2            873.5224      28479.25         4368.651   618.2361
```

```
> k.means.fit$centers
  DISTRICT   SECTOR DS      PSU       A0  SNUMBER    HIncPM    HExpPM Electricity_bill Water_bill Food_Value
1 29.95588 1.573529 10 81.75000 39319.19 5.308824 252912.75 355005.21        2044.6029   448.5000   5735.029
2 40.91327 1.840936 10 68.10375 49990.25 5.466802  44623.77  40466.94         750.8228   158.7814   3898.369
  Communication_Expen wedding_Expen OtherIndi_Expen Indi_Expen
1           3087.6176    2833810.29        39824.824  1216.5294
2            873.5224      28479.25         4368.651   618.2361
> k.means.fit$cluster
   [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
  [52] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [103] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [154] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [205] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [256] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [307] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [358] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [409] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [460] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [511] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [562] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [613] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [664] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [715] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [766] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [817] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [868] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [919] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [970] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [ reached getOption("max.print") -- omitted 19280 entries ]
> k.means.fit$size
[1]    68 20212
```

Hierarchical clustering



Model Based Clustering

```
Console   Terminal
~/
>
> mc <- Mclust(HIES, G=2)
fitting ...
  |======================================================================| 100%
> summary(mc)
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust VII (spherical, varying volume) model with 2 components:

 log.likelihood     n df       BIC       ICL
      -3579664 20280 35 -7159674 -7159776

Clustering table:
    1     2
17764  2516
> mc
'Mclust' model object: (VII,2)

Available components:
 [1] "call"           "data"           "modelName"      "n"
 [5] "d"              "G"              "BIC"            "bic"
 [9] "loglik"         "df"             "hypvol"         "parameters"
[13] "z"              "classification" "uncertainty"
> mc$classification
   [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1
  [40] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
  [79] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [118] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [157] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1
 [196] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [235] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
DISTRICT	SECTOR	DS	PSU	A0	SNUMBER	HIncPM	HExpPM	Electricity	Water_bil	Food_Exp	Food_Val	Communi	adhoc_Expen	OtherIndi	Indi_Expe	cluster
82	2	10	15	89521	3	163844	11267.38	0	0	4364	1091	400	0	0	30	2
32	2	10	16	42426	5	18906.67	14703.45	65	0	4788	1197	0	104000	0	55	2
82	2	10	31	89937	9	129126.2	65133.71	625	350	10804	2701	500	0	20040	65	2
11	2	10	68	12988	7	156588.6	80481.64	650	110	15872	3968	550	508000	6100	75	2
71	2	10	2	79196	3	11204.83	21072.95	500	0	6508	1627	0	150000	0	80	2
81	2	10	12	86789	5	18668.33	21202.27	75	60	8308	2077	200	100000	50	120	2
82	2	10	31	89937	8	171288.3	25628.77	210	0	9520	2380	300	150000	60	120	2
44	2	10	34	56368	10	154016.7	12660.04	0	0	8760	2190	250	150	0	120	2
33	2	10	31	46721	10	185454.3	14533.12	150	150	8844	2211	400	3000	0	120	2
91	2	10	60	94407	3	186921.4	14287.07	100	0	8088	2022	100	1000	0	130	2
41	2	10	10	48421	2	13183.81	101904.8	400	0	8268	2067	300	1000500	0	135	2
72	2	10	61	83430	10	143070	50397.63	1400	0	30636	7659	500	7000	0	140	2
81	2	10	82	86747	7	13836.91	35289.61	100	180	10572	2643	200	211000	400	150	2
72	2	10	29	83605	7	147453.6	34810.45	676	398	13556	3389	250	0	1887	152	2
82	2	10	31	89937	7	181619	44542.34	650	275	18428	4607	1000	1300	10240	160	2
81	2	10	14	87637	4	158264.9	62804.17	300	150	13900	3475	800	404000	0	160	2
31	2	10	6	39800	4	17923.81	43207.96	360	0	18092	4523	675	120000	0	169	2
45	2	10	31	58435	9	19093.1	57510.55	0	0	6680	1670	0	600000	0	170	2
31	1	10	77	40333	10	17081.9	28618.4	236	537	9208	2302	0	127000	50	173	2
33	2	10	64	45193	10	131421.9	65478.34	0	0	8608	2152	1000	30000	50000	173	2
62	2	10	63	77037	7	148498.6	43749.99	1500	500	18992	4748	2000	26500	150	180	2
22	2	10	51	33370	7	134764.3	29478.1	200	0	7144	1786	1000	20000	1558	182	2
23	3	10	32	36636	2	172092.4	25544	0	0	18968	4742	0	2500	950	184	2
32	2	10	95	42132	3	18262.38	36276.16	179	0	7652	1913	100	123000	5500	185	2

Appendix B: Source Code

Using K-means

```
library(readxl)

HIES <- read_excel("F:/Luck/research/Datasets/HIES.xlsx")

View(HIES)

head(HIES)

k.means.fit<-kmeans(HIES,2)

attributes(k.means.fit)

k.means.fit$centers

k.means.fit$cluster

k.means.fit$size

out <- cbind(HIES, clusterNum = k.means.fit$cluster)

out

write.table(out,file="F:/Luck/research/RunIN_R/Outputnn.csv",sep=",")
```

Using Hierarchical Clustering

```
library(readxl)
HIES <- read_excel("F:/Luck/research/Datasets/HIES.xlsx")
View(HIES)
HIES1 <- scale(HIES)

d <- dist(HIES1, method = "euclidean")

hc1 <- hclust(d, method = "complete")

plot(hc1)

clusterCut <- cutree(hc1, 2)

table(clusterCut)
```

Model Based Clustering

```
library(readxl)
HIES <- read_excel("F:/Luck/research/Datasets/HIES.xlsx")
library(MASS)

library(mclust)

mc<-Mclust(HIES,G=2)

summary(mc)

mc

mc$classification

out<- cbind(HIES,clusterNum=mc$classification)

write.table(out,"C:/Users/ICTO/Desktop/research/mclus_sus.csv",sep=",")
```

Appendix C: Questionnaire

Household Income and Expenditure Survey Questionnaire is added from next page