7-2021

Fraud Detection using Data Analytics

Ayesha Karmustaji
ak5721@rit.edu

Fraud Detection using Data Analytics

by

Ayesha Karmustaji

A Capstone Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Professional Studies:

Data Analytics

Department of Graduate Programs & Research

Rochester Institute of Technology
RIT Dubai
July, 2021

RIT

Master of Science in Professional Studies:

Data Analytics

Graduate Capstone Approval

Student Name**: Ayesha Karmustaji**

Graduate Capstone Title**: Fraud Detection using Data Analytics**

Graduate Capstone Committee:

Name: Dr. Sanjay Modak Date:

Chair of committee

Name: Dr. Ioannis Karamitsos Date:

Member of committee

Acknowledgments

I would like to express my gratitude towards Dr. Ioannis Karamitsos for guiding me throughout this project. I would also express my gratitude towards Dr. Sanjay Modak for giving me the opportunity to allow me to successfully work and conduct this project and for giving me a second chance when I needed it the most.

The following project was done under the supervision and guidance of Dr. Ioannis Karamitsos. I feel thankful to have worked with such a kind patient mentor, who provided me with all the necessary feedback and ensured that I follow the right track and guided me step by step to ensure the success of my work. The success of this project would not be achieved without the guidance of Dr. Ioannis. Therefore, I couldn’t thank him enough for supporting me through the completion of my work and guiding me every step of the way and for investing his time and effort to ensure that I succeed.

Abstract

At present, the biggest concern of every organization is to detect and control financial fraud. Tax frauds cause the loss of billions of dollars every year. As a result, data mining techniques are used to combat the growing problem of tax fraud. Tax evasions cause a reduction in revenue collection. It also has a bleak impact on government policies and budget. The goal of this study is to describe the use of data analytics tools to process and analyze tax data related to value-added tax evasions. This study is a conceptual perspective that provides a theoretical and methodological basis for data analytic application to detect evasions in taxation. This study will also explain that tax collecting companies can identify fraud detection through these methods and can save a lot of time and finances. Outlier analysis is applied to a wide variety of fields which include fraud detection, medical diagnosis, intrusion recognition, web analytics and fault identification. This project aims to detect tax fraud by using data analytics technique as the outlier analysis. In this project, an outlier detection technique used for identifying tax fraud, especially in VAT. Research on financial fraud detection has a long tradition. Recent theoretical developments have revealed that the clustering technique used for outlier detection is not only on point but also very cheap and helpful.

Keywords

Tax Evasion, VAT, Fraud Detection, Data Analytics, Outlier Detection

Table of Contents

List of Figures

List of Tables

CHAPTER 1

1.1 Introduction

Nowadays, Internet of Things (IoT) is used in all domains including pharmaceutical, sales, transportation, manufacturing and logistics. All payment systems are connected to the internet and are considered to be processed through online systems. Due to the technological advancements (internet and online banking, credit card payment systems) frauds are continuously increasing, which is a great loss in terms of business and revenue. This loss of revenue effects the overall economy of country. Fraudulent activities are taking place in many sectors of daily life such as telecom networks, mobile services, internet banking and electronic commerce. Therefore, it is extremely important to develop fraud detection techniques. Tax fraud includes wide range of activities which involve declaring false information, contradicting realities and attaining financial benefits regardless of the facts available in the legalized structure.

Tax fraud is a crime and the people who are responsible of tax fraud, will have to be punished. Although legal punishment is only given when fraud is detected by government and fraudster is summoned on corruption charges. Tax fraud has great impact on states, governments, organizations, companies, investors and employs. As the use of internet and mobile apps is growing and technology can be availed at lower rates, chances of fraudulent activities increased far more than the past. So, tax fraud can no longer be detected manually. It is necessary to use technology which is capable of doing job with accuracy and quality. The problem statement for this research is; “What method can be used to detect tax fraud?”

Taxes are the most vital source of generating revenues in every system. But the taxpayers are always ready to find different methods so that they can reduce the amount of taxes required to pay. The circumstances in which a person or business group purposefully and knowingly provides false information about their tax return to pay less amount of tax that they are required to pay is defined as tax fraud. It is an illegal act. Mostly tax fraud comprises of using incorrect social security number, asserting inaccurate deductions, showing personal expenditure as business expenditure and not to disclose correct income However, tax evasions drastically affect the economic conditions of a country. The most important goal and highest priority of tax offices all over the world is to detect and minimize tax evasions (frauds). Bank of America decided to pay $16.5 billion for solving and controlling financial frauds (Fox News, 2014). On the other hand, the European Union and rest of the countries all over the world are putting numerous efforts for minimizing tax frauds. These countries also trying hard to investigate the different types of tax frauds, reducing leakage of information and taking essential actions to find reasons behind tax evasions. It is also evident that in the domain of reputable organizations numerous individuals are responsible for tax frauds. The owners of these companies hide their annual incomes and assets. However, tax invasion is recently evident from off-shore financial institutions such as the Panama Papers and Luxemburg leaks.

The importance of VAT can be understood by the fact that it is implemented in 140 countries all over the world as compared to 47 countries that adopted VAT in 1990. Nowadays, low-income countries are adopting VAT at a fast pace (IMF 2011). The self-enforcement mechanism of VAT is very helpful as it is useful for buyers. As buyers can ask for a receipt when they buy any product thus, claiming a VAT refund for their purchases. Also, buyers can claim VAT refunds by keeping a record of purchases. On the other hand, sellers try to under-report their sales so they can pay less VAT. Another advantage of VAT is that it provides written evidence for every transaction and both buyer and seller keep the record of data.

Billions of euros are lost in VAT returns by EU member countries every year due to evasions in VAT and insufficient tax collection structure. This loss is termed as “VAT gap” and it is defined as the difference between expected VAT returns and the actual amount collected by VAT. These losses are not only due to tax evasion and avoidance but also due to financial liquidation, bankruptcy and mistaken calculations. Research done by the European Commission reported that the total VAT Gap for 26 EU countries calculated to be approximately EUR 193 billion in the year 2011 alone. Sweden had a tax gap of 8 % of total taxes in the year 2000 and the tax gap of the United Kingdom is almost the same size as the gap of Sweden and the U.S.

Global use of information technology and electronic communication in the business processes of companies has decreased human participation. As a result, fraudulent activities have increased. Scammers also take advantage of the latest technologies. In order to overcome these fraudulent activities, the same technological methodologies should be implemented. A large amount of data can be collected by analyzing the huge database of online transactions (availability of the electronic environment has made this collection easy). The development of new techniques, methods, approaches and tools made it easy to analyze the huge amount of data at low cost and thus, increase consumer satisfaction. The two most important elements to deal with fraud are the detection of fraud, and the prevention of fraud. There are several techniques, which are used to detect frauds by using direct and indirect methods. It is necessary to collect proof of evasion. The quality of evidence should be enough to execute the evaders. The process of tax fraud detection involves analysis and handling of a massive amount of data and information to find about actions of fraudulent. It needs fast, systematic and logical algorithms.

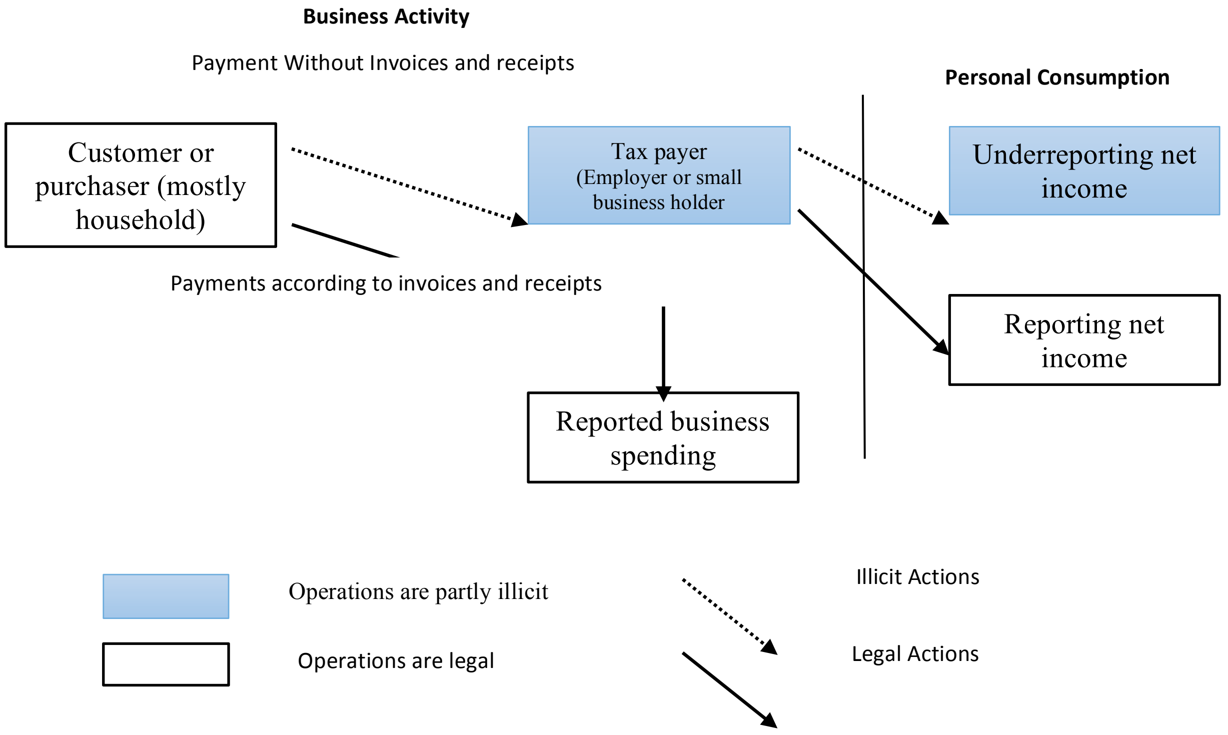

Business Activity
Payment Without Invoices and receipts
Personal Consumption
Customer or purchaser (mostly household)
Tax payer (Employer or small business holder)
Underreporting net income
Payments according to invoices and receipts
Reporting net income
Reported business spending
Operations are partly illicit
Operations are legal
Illicit Actions
Legal Actions

Figure 1. **Cycle of Tax Evasion.**

Tax analysts, government institutions and accounting firms can successfully use data analytics to find about the fraud. According to PwC global economic crime survey conducted in 2009, 30 per cent of companies are the victim of tax fraud (PricewaterhouseCoopers LLP 2009). Data analytics (DA) is termed as science which involves an investigation of data to make decisions in business based on results obtained. There are four types of data analytics methods: 1) prescriptive 2) predictive 3) descriptive 4) diagnostic. HMRC has reported that between April 2013 and April 2014 it was able to recover £2.6 billion by using data analytics tools, whose initial investment was £45 million. (United Kingdom Houses of Parliament 2014).

Controlling fraud is a business process that involves the reduction of fraud in an organization to a level that is acceptable. This process greatly depends on both people and the internet for its success. A fraud control plan involves the approaches which are programmed to prevent the fraud to take place in the first place. Once fraud occurs, strategies should be used to detect fraud as soon as possible. The response should be quick and appropriate after the detection of fraud. Frauds should be monitored, reported and evaluated to assure that responsibilities are fulfilled, and accountability is boosted. Prevention of fraud is a cost-efficient and effective strategy. As fraud occurrence is a fact and it can happen at any time. Data mining techniques are helpful in identifying fraudulent activities.

Tax authorities are progressively using big data & advance analytics techniques to perform audits and show trends and deviations, using new techniques which are rule-based monitoring, predictive analysis and outlier detection. The predictive analysis takes account of predictive tasks which make a prediction for every observation. Data analytics is also able to detect lapse in security systems by recognizing any problem that occurs during the analysis of a substantial amount of linear and nonlinear data received from different sources. Data analytics can also foresee fraud activity before it arises, which helps to find new approaches to address and manage illegal financial gains (Deloitte, 2014). Data analytics can also tackle tax frauds across the borders by dealing with unlawful investment transfers. Data analytics is also helpful to law agencies as it helps to share data effectively and quickly to detect tax fraud by analyzing and recognizing patterns (Houses of Parliament 2014). The modern age is an age of cybercrime, and the financial industry is more prone to tax frauds. Tax fraud activities have caused significant economic and financial losses. A considerable amount of transactions take place every single day. Complicated illegal activities cannot be identified by investigation of organized data such as credit card transactions (Gutierrez, Anzelde, and Gobenceaux, 2013). Data analytics can analyze both organized and unorganized data coming from different sources such as mobile devices and social media to detect tax frauds between accounts and find relation among sources and receiver.

Data is analyzed to reveal hidden anomalies and patterns to avoid losses and to warn companies to prevent frauds. Anything that is exceptional, unusual or abnormal, is called analytical anomaly. Also, any point which fall out of normal or new patterns is termed as analytical anomaly. Examples of analytical anomalies are:

- Inliers at points where they are not anticipated
- Outliers
    - Transactions at unexpected timings
    - Transactions are either too many or too less
    - Abnormal relationship between items
    - Unexplained items
    - Exceptional account balances
    - Inconsistencies
    - Large number of copies of item numbers
    - Unusual payment methods

Artificial intelligence and data mining techniques can assist tax administrators to take precautionary measures (for fraudulent activities) and to refine layout of tax design. Data mining involves investigation of large amount of data to get useful information by using neural networks, statistical techniques and artificial intelligence to discover hidden patterns and designs. Evaluation of every component of data by using analytical and logical reasoning is termed as data analytics.

“Importance of data is equivalent to oil in the 21st century”. In the growing age of cyber and electronic worlds data has become key asset for any industry and organization. Majority government bodies are making the data repositories based on the tax data. They try to keep the record of tax data generated by companies because in numerous countries, tax is the biggest source of revenue generation (IMF, 2017). Tax administrations have always tried to control false income reporting, misinformation about tax money, all types of tax frauds especially related to VAT (value added tax), PIT (personal income tax) and CIT (corporate income tax).

In this project an analytical framework is used, which incorporates use of machine learning technique, outlier model to detect the fraud in VAT.

1.2 Project goals

Although VAT has many advantages such as generating revenues, encourages businesses to control costs; easy to manage etc. but, VAT sometimes fail to perform in developing countries (Besley and Persson 2013). VAT is facing many challenges. One of them is fraud and evasion due to false invoicing. Sometimes taxpayers do not register themselves or there is under reporting of payable tax. Increase of fraud in VAT is becoming threatening with every passing day. There is a need to develop a technique to predict and control frauds in VAT. The motivation behind this research is to predict tax evasions to prevent losses and to stop frauds in VAT to generate more revenues, as these revenues are used for the well-being of public by the government. In time prediction, detection and prevention of tax fraud can boost the economy. The main goal of this research is to find analytical methods which can be used in fraud detection. Data analytics and data mining are two important fields which can be used for purpose of fraud detection.

To understand statement of the problem, following research questions are discussed

- What is financial tax fraud?
- What are the types of financial tax fraud?
- What are the reasons behind the tax fraud?
- What are the impacts and outcomes of tax fraud?
- What are the methods used to reduce tax fraud?
- How the use of technology in tax fraud detection can help?

1.3 Aims and objectives

Regardless of the fact that data analytic can play crucial role in detection of fraud, only 3% cases of fraud are analyzed using data analytics. In order to identify fraudulent activities, common and familiar models should not be followed. As, fraudsters also use new techniques and tool. This is the age of cybercrime, which involve online banking fraud, mobile banking fraud, ATM CARDS fraud, financial documentation fraud, etc. The reason behind the inadequate use of data analytics is that its effectiveness to fight fraudulent activities is not well known to the higher authorities (Phill Ostwalt 2016). Thus, the objectives for the capstone project are as follows:

1. Data analytics techniques should be implemented in financial sectors to prevent tax fraud.
2. To find and develop the algorithm for detection of tax fraud on tax on goods and services.
3. Implementation of algorithm
4. To assess the practicability and usefulness of proposed solution

1.4 Research methodology

Research methodology is based on the knowledge of financial fraud, its types and causes. Losses due to tax fraud are also discussed. Different techniques used for financial fraud detection are also explained. Implementation of these techniques in different countries and in different financial fields is presented. The data from the following source is collected: OECD Revenue Statistics 2020 http://oe.cd/revenue-statistics. This data involves taxes on personal income, profits, gains, payroll taxes, taxes on property and taxes on goods and services (VAT). This is the most important step because data is the basic building block of analysis in data analytics. Selection of data plays a vital role in building analytical models. All the data that is collected from federal OECD is then gathered and converted in a form of data mart. Data cleaning and data preprocessing is done as the first step of methodology. Methodology also involves outlier models, Z score, Gaussian distribution and k-means clustering technique for detection of anomalies in tax data. Clustering technique is implemented to tax fraud data and shows the results. In Clustering analysis similar data points are grouped together. Thus, one group contain similar data points or objects. Any data point or item which is different from points in a cluster are termed as outliers. Specific number of clusters is selected. Algorithm is run to detect outliers. After analysis results are presented.

1.4.1 Data preprocessing and cleaning

The techniques which are applied to the data set before application of data mining methods is termed as data processing. Data processing is a very important phase because it gives the final data set which is then used for data mining. Most commonly data is imperfect with inconsistency, noise, missing values and useless data. Also, data is growing at fast rate due to development in every field. Huge amount of data needs more systematic and organized techniques for analysis. Data mining algorithms are developed to process data. A powerful tool which is used to deal with complex and huge data is data preprocessing. It is a time-consuming process which involve many steps such as data preparation, data transformation, integration, cleaning and normalization. Complexity of data is reduced by data reduction techniques which involve feature selection, instance selection and discretization. As a result of application of above-mentioned techniques to the data, final data obtained is reliable and valid.

1.4.1.1 *Missing data and noise treatment*

Missing data and noise are most common problems faced during data preprocessing. A value may be missing due to fault in sampling or limitations of cost. If the missing values are not handled properly, it results in imperfect knowledge extraction and defective results. Statistical techniques are useful for dealing with missing data. There are two main approaches which are used to deal with noise are 1) data polishing methods 2) noise filter. Even the limited amount of noise correction proves to be beneficial. Hence, it is necessary to remove noise and cater missing data before applying data mining techniques.

1.4.1.2 *Data transformation:*

It is a process in which format of source data is changed to format of the desired destination. Attribute transformation and dimensionality reduction are the methods used for data transformation. Data transformation is done after preprocessing and cleaning data to further refine the data. In this step data is transformed efficiently according to required objectives. Data mapping is the first step of data transformation. A critical metadata is produced. After which the actual data conversion is done.

Transformation techniques used for data transformation are as follows:

- Smoothing: it removes noise from data.
- Aggregation: it helps in summarization and construction of data cube.
- Generalization: it gives the concept of hierarchy climbing.
- Normalization: it is scaled to fall within a small and specified range.
- Attribute/feature transformation.

1.4.1.3 *Dimensionality reduction:*

Dimensionality problem arises when the data becomes big and huge with the increase in the number of instances. As a result, the functioning of data mining algorithms is interrupted. Feature Selection (FS) and space transformation-based methods are used to solve this problem.

Feature selection: This process involves the removal of useless and needless information. This technique is used during initial phase of data collection, thus saving time and cost.

Space transformations: This technique creates a whole new set of features with the amalgamation of original and new features.

Instance reduction: This technique is used to reduce the effect of large datasets on the algorithms. The quality of data is not compromised and data is decreased by reducing instances or by creating new instances.

1.4.1.4 *Discretization:*

It involves treating continuous features as if they are categorical. It divides the numerical features into a limited number of intervals which are not duplicate.

1.4.1.5 *Description of input data*

Data plays a key role in any outlier detection technique. Generally, input data is collection or group of data instances. These instances are also termed as object, record, point, vector, pattern, event, case, sample, observation and entity (Tan, P.-N., Steinbach, M., Kumar 2005) Each data instance is explained by using a set of attributes. Attributes are also termed as variable, characteristic, feature, field and dimension. The attributes can be of many different types such as binary, categorical and continuous. Data instance which involve only one attribute is termed as univariate. Data instance which involve many attributes is termed as multivariate. Attributes can be of same type or a mixture of different data types in case of multivariate data.

Input data can also be classified on the basis of relationship between data instances. Most outlier detection techniques deal with the data in which it is assumed that there is no relationship between the data instances. Such data is called point data.

The most common types of structured data sources that are analyzed in taxation

- Transactions from banks
- Tax reports published annually and quarterly
- Audits
- Legalized organizations
- Electrically generated invoices and receipts

- Statements
- Personal income tax
- Unstructured data sources include:
- Personal information and messages data
- Emails
- Documentation
- Interpretation of receipts
- Activities on social media
- News from media

1.5 Limitations of the Study

As the clustering is unsupervised technique, there is no principle or rule present to verify the results. Also, the datasets of real life are quite different from the data distribution in clustering. K-means clustering algorithm is highly dependent on data set and its initial status which may result in insignificant solutions. Number of clusters has to be specified by the user before running the algorithm. Determining the ideal number of clusters is a challenging task because it require true and former knowledge about data which is not available at all occasions. So, the number of clusters should be predefined. It does not wot work well when the shapes of cluster are other than sphere, sizes of clusters are different and data is non numerical. As compared to other clustering techniques, k-means clustering is cost efficient and has low computational cost.

Following points should be taken into consideration before implementation of k-means algorithm:

- k-means algorithm should be implemented to scaled and consistent data.
- Sometimes the elbow method for selection of k numbers doesn’t work because error function start decreasing monotonically and become flat.
- More weight is assigned to huge clusters by k-means.
- It is assumed that the shape of clusters is sphere in k-means. It doesn’t work well with other shapes such as elliptical.
- k-means clustering cannot measure uncertainty for overlapping clusters.
- Data form different sources such as uniform distributions can be clustered by k-means.

CHAPTER 2 LITERATURE REVIEW

This chapter presents an overview of fraud types and different techniques used for detecting frauds, which are implemented using machine-learning methodologies.

Fraud in taxation is an important issue which results in significant loss of government revenues. As a result, tax rules become less effective and leads to injustice between evaders and honest tax payers (Alm, 2011). Fraud is defined by Association of Certified Fraud Examiners (ACFE) as "the unauthorized use of one's position of power to get personal benefits by misusing assets and resources of a company" (Investigating Fraudulent Acts, 2000). Illegality is the most common feature of all the tax fraud definitions. Aimed, legal or illegal and money-making intentions of a taxpayer results in tax evasion which in turn leads to decline or eradication of tax accountability. Other financial advantages are also obtained from tax evasions. The failure and insufficiency of tax accountability is also termed as tax fraud. Illegal criminal activities are common in financial sectors. Criminal acts occur at large scale and they are not unusual. 8% of the total world's economical wealth is lost due to tax evasion and frauds. (Zucman, 2013).

Tax administrations are constantly concerned about tax fraud and losses caused by tax evasions, especially with reference to developing countries. Taxes not only generate handsome amount of revenue for the government but also shows efficacy and dedication of government to carry out its administrative functions (Davia, Coggins, Wideman, & Kastantin, 2000).

Tax fraud is a complicated and problematic issue which needs to be resolved by using systematic, organized and correlated scientific approaches. One country is not capable of dealing with this issue alone. Tax fraud is an intentional act which has to be punished under criminal law. Tax fraud involves the situations in which a person provides false statements, documents and records. Sometimes taxpayers hide their income and assets from tax authorities so that they can pay less tax. Often taxpayers also take advantage of loopholes which are not intended by legislation (EC – Taxation and Customs Union, 2017).

Developments and growth in modern technology especially internet has resulted in increased financial fraud (Yeh I and Lien C-h 2009). Societal factors such as growing usage of credit cards and rise in expenditures result in rise in frauds. Often transaction activities don't seem related to tax fraud but, when tested through data analytics frauds are detected (Tatiana,2016). As the volume of data is increasing day by day, incorrect data cannot be identified by conventional methods which are based on linear data analysis. That's why data analytics is used to analyse nonlinear data. Also, it analyses the data that is not well defined and has a high correlation. Data analytics is a more robust and powerful technique. It has replaced manual methods of analysis. Thus, resulting in finding missing connections, validating fast reactions to frauds and controlling bigger loses before it is too late.

Different data mining techniques are useful for classification and these techniques can be implemented efficiently to large datasets. Thus, these data mining techniques has ability to deal with problem without having any prior information about problem statement (Ravisankar P, et al 2011). Data analytics is also able to detect lapse in security systems by recognizing any problem that occurs during the analysis of a substantial amount of linear and nonlinear data received from different sources. Data analytics can also foresee fraud activity before it arises, which helps to find new approaches to address and manage illegal financial gains (Deloitte, 2014). Data analytics can also tackle tax frauds across the borders by dealing with unlawful investment transfers. Data analytics is also helpful to law agencies as it helps to share data effectively and quickly to detect tax fraud by analyzing and recognizing patterns. The modern age is an age of cybercrime, and the financial industry is more prone to tax frauds. Tax fraud activities have caused significant economic and financial losses. A considerable amount of transactions take place every single day. Complicated illegal activities cannot be identified by investigation of organized data such as credit card transactions (Gutierrez, Anzelde, and Gobenceaux, 2013). Data analytics can analyze both organized and unorganized data coming from different sources such as mobile devices and social media to detect tax frauds between accounts and find relation among sources and receiver.

Economy can be classified into three groups: the shadow economy, the maintenance economy or household production and the market economy. Market economy belongs to people. Maintenance economy is main hub of human activity. Shadow economy is everything in between maintenance and market economy. Shadow economy is highly affected by tax fraud because people do not trust their governments and use their authority to get benefits in shadow economy (Peterson, Thießen and Wohlleben 2010)

VAT is the most effective way of earning tax revenues. The value added tax (VAT) is implemented in almost 130 countries during different phases of economic growth. It plays key role in generating revenues. But, like any other tax, VAT is also at a risk of fraud and evasion. VAT generate almost 45% of revenue which is equal to USD $18.7 billion dollars and produce over 400 million invoices a year, of which 56% are generated in paper format and 44% are generated as electronic format False invoicing is the most common type of fraud seen in VAT (Bergman, 2010).

2.1 Financial Fraud

Association of Certified Fraud Examiners (ACFE) defines financial fraud as a “well planned action which causes damage to the financial assets of people and deprive them from their money. This action is considered to fall in the category of fraudulent act (ACFE 2010).

2.2 Types of Financial Fraud

Economic experts have defined numerous categories of financial frauds which are defined in state of the art. Few of these types are defined in the next section.

2.2.1 Corporate Fraud

Corporate fraud is also known as financial statement fraud. Financial statement shows the financial state of an organization. Fraud in financial statements is performed in order to gain more profit from business by reducing mandatory payable taxes.

2.2.2 Credit Card Fraud

Credit card frauds has two types

1. Application fraud
2. Behavioral fraud

Application fraud: It includes acquiring of new cards by using false information or information of other people

Behavioral Fraud: It includes fake cards, stolen or misplaced card, personal mail theft and absence of card holder.

2.2.3 Security and Products Fraud

It includes investment fraud, goods fraud, foreign exchange fraud, security account fraud, manipulation of market fraud.

2.2.4 Vat Fraud

Different types of frauds in VAT are as follows:

2.2.5 MTIC Fraud and Carousel Fraud

MTIC stands for missing trader intra-community fraud. This is most common and simple type of fraud in VAT and can be easily controllable through using appropriate technology. It happens when a company or a business purchase goods without paying VAT. Later, VAT is collected on onward sale and then the company disappears without handing over the collected VAT. MTIC frauds are common in products which have high value and low volume such as cell phones and computer chips. Carousel fraud is same as MTIC fraud but it comprises of series of transactions which takes place between different companies in different countries. But they are managed by same groups comprising single person or persons. (Fabrizio Borselli 2008)

2.2.6 Cars related VAT fraud

It is also called car flipping. It is similar to MTIC and carousel fraud. Foreign registered luxury cars are bought and then sold at higher price in other countries. (Ainsworth, Richard 2007)

2.2.7 Frauds at Reduced Rates

Reduced rates are applied to specific good and services. Suppliers tries to get benefit from the reduced rates by selling banned or ineligible products.

2.2.8 Aircraft Leasing Frauds

Aircrafts are the major commodities of leasing business. Sometimes, an importer leases an aircraft for himself or for a family member. In such case value of VAT on leasing aircraft is much lower than value of VAT on import of aircraft. This falls under the category of tax avoidance.

2.2.9 Electronic-Commerce Related Fraud

It includes import fraud, e-commerce frauds when assistance is provided by electronically supplied services, frauds between European union states when exports and imports take place at a distance (C. Jennings 2010)

2.2.10 Money Laundering

It is a process in which criminals hide their earnings or income from illegal acts and convert their earnings into services and goods. Thus, money laundering is the conversion of dirty money into legal and clean assets (Ngai E, et al 2011).

2.2.11 Insurance fraud

It includes insurance procedure frauds such as billing fraud, claim fraud, ratings fraud etc. which can be done by workers, healthcare organizations, agents and dealers etc.

2.2.12 Mortgage Fraud

It includes property or mortgage document fraud. It is used to twist original value of property.

2.3 Causes of Tax Fraud

There are many factors that affect the occurrence of tax evasion or tax fraud. Following six areas are strongly influenced by the phenomena of tax fraud.

2.3.1.1 Financial factors

It includes factors like economic situation of a tax paying organization, the amount of payable tax, the likelihood of the detection of tax fraud, the amount of accreditation, common business conditions and business inactivity.

2.3.1.2 Societal Factors

It includes factors like unfair attitude of government towards taxpayers and their tax benefits, injustice in society and impartiality in state policies.

2.3.1.3 Legal Factors

It includes factors like lack of trust between government and public organizations, uncooperative attitude towards taking responsibility, complex and fluctuating tax rules and regulations and influence of taxation on economic events.

2.3.1.4 Demographic factors

It includes sex, age, marital status and education.

2.3.1.5 Psychological factors

It includes behavior towards legal laws and perception of nationality.

2.3.1.6 Moral factors

It includes behavior towards civil accountability, behavior towards taxation, principles, religion and habits.

Many tools are developed for the detection of financial frauds to solve the problem of frauds in taxes and to provide authentic and proven solutions to business organizations. Outlier detection process is normally used to detect financial frauds It is a data mining technique which help to discover outliers, hidden trends and patterns normally found in huge databases (Jin Y, Rejesus R, Little B. 2005).

2.4 Outcomes of tax fraud

2.4.1 Personal outcomes

People who violate tax laws and get involved in tax fraud will face negative outcomes. These outcomes consist of legal penalties and displeasure from the society. If an organization is involved in fraud, it may result in bankruptcy, loss of stockholder and bad impression. It may also result in reduced income, reduced benefits, less work or loss of work (Whiting, Hansen, McDonald, Albrecht, & Albrecht, 2012).

2.4.2 State outcomes

Governments pay huge amounts to detect tax frauds and fraudulent. If a government fail to detect tax evasion, then policies of governments are considered ineffective and unproductive (Wu, Ou, Lin, Chang, & Yen, 2012). It was estimated that almost twenty-eight billion dollars were lost due to VAT in 2009. Overall tax gap was around 345 billion dollars in US which is 16.3 % of the tax that needed to be collected in 2011.

2.4.3 Outcomes for organizations

Studies show that companies waste useful resources in avoiding tax rather than using them for making more products and getting more benefits (Wu et al., 2012). Other negative outcomes of financial fraud involve loss of investor's confidence in a company's financial statement and credibility of an organization. As a consequence, investments in the company are decreased thus lowering equity, less recourses and less budget or assets for operations and development. Stock markets are used as a measurement for the impact of tax fraud because the stock value of company get decreased if they have committed fraud. Fraudulent activity of single company results in overall market loss (Whiting et al., 2012).

2.5 Measures for reducing tax fraud

It is the duty of government to implement tax laws. Making tax laws is not useful until they are implemented accurately. Following measures can be taken to reduce tax fraud.

- Sometimes higher tax rates are cause of tax evasion so reduction in Marginal tax can help in reducing tax evasion.

- Number of tax audits should be increased.

- Imposing higher penalties for those who break the law.

- Higher fines should be implemented.

- Using data mining techniques.

2.6 Financial fraud detection techniques

Following are different tools and methods used for detection of financial fraud.

2.6.1 Unsupervised Techniques

These techniques focus mainly on inherent structures, correlations and connections.

1. Self-Organizing Maps (SOMs): It is a neural network (NL) technique. It uses learning function which is unsupervised. Data can be seen from high dimension to low dimension by the user (D.Olszewski, 2014)

2. Outlier detection (OD): It is unique and exceptional method as compared to usual methods. It is used to detect outliers or anomalies by using different procedures (N.Malini, M.Pushpa 2017).

3. Group method for data handling (GMDH): It is an analytical information and knowledge algorithm for design of complex systems. Highly complex data is handled by using this self-organizing approach (Ravisankar P, et al 2011).

4. Density based spatial clustering of applications and noise: It is a clustering algorithm which is derived from density. It is employed to detect outlier and groups of random shapes (Ravisankar P, et al 2011).

5. Association rule analysis (AR): It is used for transaction sets. When handling relational data, every bit of data is considered as a pair and transaction is considered as group. (D.Sa´nchez, M.A. Vila, et al 2009).

2.6.2 Predictive Techniques

In these techniques, analytical model is created to select objects of importance.

1.Bayesian belief network (BBN): It represent dependencies between subsets of attributes. It is a directed open chain graph, where every intersection shows a characteristic and each pointer indicate an anticipated dependency (E.Kirkos, et al 2007).

2. Bayesian slanted logit pattern (BSL): In this pattern unbalanced samples are used. Asymmetric links are used to measure the probability of xi = 0 and xi = 1 (Ll. Bermudez 2007)

3. BP (Bivariate Probit technique): This model is utilized when a bilateral index is obtained. Qualitative information is obtained in the form of an autonomous variable, which is the expected result. The possibility that the autonomous explanatory variable becomes remote cannot be dismissed, by controlling group of covariates (B.Bai, et al 2008).

4. Classification and Regression Tree (CART): This technique is different from traditional statistical methods. It is a programmed and distribution-free technique. It uses RPA (binary Recursive Partitioning Algorithm) for classification of instances. Instances are classified into a specified number of discontiguous areas. These areas refers to a terminating intersection of the regression tree (B.Bai, et al 2008).

5. Cost-sensitive decision tree (CSDT): It is a selection algorithm. It is used to detect malicious activities in transactions of credit cards. Decision tree algorithms use two splitting criteria 1) it is independent of costs and group classification 2) the cost is predetermined for a fixed ratio (N.Malini, M.Pushpa, 2017).

6. Decision Trees (DT): It is organized as model of tree. Every junction shows an experiment on a characteristic and every branch represent the results of the experiment. So, the model of tree divides findings into specified smaller groups or branches (E. Kirkos 2007).

7. Decision Trees C4.5: It is used to solve the problems which cannot be solved by using DT such as missing attribute values.

8. K-nearest neighbor (KNN): It uses supervised learning techniques. This method is mostly used for fraud detection especially credit card frauds (N.Malini, M.Pushpa, 2017).

9. Logistic Regression (LR): It involves a single dimension and common form of lineal model. It is not useful when predictable variable is in pair due to normality assumptions.

10. Neural Networks (NN): It is a highly developed technique with accepted concepts and specified programming fields. It involves interconnected processing units which consist of neurons. Each interconnection has a specific mathematical number, called "weight or density"(E. Kirkos 2007).

11. Naïve Bayes (NB): It is classification tool which use Bayes rules based on expectations. Each characteristic and label of a group is considered as arbitrary quantity, and it is assumed that the characteristics are autonomous. The naïve Bayes helps to identify a group for every new observation. Its probability increases when the estimated values of the characteristics are given (Ll. Bermudez 2007).

12. PNN (Probabilistic neural network): It is a neural network having linearization. It is also a pattern classification network. One pass training algorithm is used by PNN for categorization and plotting of data. Classical Bayes organizer is the basis of PNN (Ravisankar P 2011).

13. (SVM) Support Vector Machines: It uses a rectilinear model which is applied to stochastic class limits by plotting the contributing vectors stochastically. This process is done in high-magnitude attribute span. A quintessential separating subspace is built in the new span. (Ravisankar P 2011).

2.6.3 Artificial Intelligence Techniques

Complicated real-world problems are solved by using artificial intelligence techniques. These are some computer-based methodologies:

1. Artificial Immune System (AIS): The design principles for AIS are adapted from the human biological immune system to solve numerous problems (N. Wong et al 2012).

2. Artificial Immune Recognition System (AIRS): In this method both self/non-self units and locater units are illustrated as attribute vectors. ARB (Artificial Recognition Ball) is used to reduce redundancy, which is an illustration of same memory units (Ravisankar P 2011).

3. Artificial neural network (ANN): This network is created to copy the functions and practices of human brain. A NN (neural network) involves connections of basic objects called the elementary neuron (A.Mubalik (Mubarek) , E.Adali 2017).

4. Genetic programming (GP): It is a research method which belongs to the clan of progressive and developed data processing and computing. It is developed version of genetic algorithms (GA). GP produces a group of solutions at random and at initial stages. Then, the initial solutions are operated by genetic operators to produce new populations (Ravisankar P 2011).

5. Genetic Algorithm (GA): It is inspired from natural evolution. An initial population is developed from randomly generated rules (B. Hoogs 2007).

6. Hidden Markov Model (HMM): A hidden Markov model is a representation of the simplest dynamic Bayesian network. It is different from standard analytics by Markov because of having imperceptible phases (Y.Dai et al 2016).

7. ID3 (Iterative Dichotomiser 3): It deals with illustrative databases. Information is expressed as decision tree (N.Malini, M.Pushpa, 2017).

8. MLFF-NN (Multi-layer feed forward neural network): It is the simplest, customary and effective neural network structure. (Ravisankar P 2011).

9. (MPL) Multilayer Perception Algorithm: It is an artificial NN (neural network) technique. One of its advantage is that it is a nonparametric estimator which is very helpful for classifying and detecting intrusions (A.Mubalik (Mubarek) , E.Adali 2017).

*10. Parenclitic Network (PN)***:** It is a technique which is used for network reconstruction. It helps to focus on the contrast between one sample and a standard sets (Ravisankar P 2011).

2.6.4 Previous studies

Researchers have studied most of the modern technologies which help in finances and assistance such as credit cards, online transactions, internet banking, insurance policies, e-commerce marketing and telecommunication since long time. It is evident that these technologies must make use of anti-fraud techniques to reduce fraud and losses. Detection of fraud in tax is done by using many machine learning techniques, which helps as reference for solving fraud problems (Abdallah, Maarof and Zainal 2016).

Initially, fraud detection techniques targeted the use of statistical models such as logistic regression and NN (neural networks). Different algorithms are developed to detect financial frauds so that this problem can be addressed and solutions can be provided. Normally, tax frauds are detected by using outlier detection methods which is validated by data mining techniques. As a result, beneficial information is obtained which include trends, designs, patterns, connections and relations. (Jayakumar GDS, Thomas BJ 2013). Data mining is defined as “a process that make use of numerical, analytical, mathematical, artificial intelligence and machine learning techniques to draw out and recognize beneficial information in such a manner that knowledge can be obtained from a large database. Data mining is also termed as discovery of data or discovery of information. Financial frauds can be detected by using different techniques such as logistic regression, decision tree, support vector machine (SVM), neural network (NN), Naïve Bayes (NB), and Bayesian belief network (BBN) (Ngai et.al. 2011). The internal revenue service (IRS), is an organization deals with the administrative tasks regarding taxes in United States (US Government Accountability Office, 2004). IRS is using data mining techniques for detection of tax fraud, financial activities of scammers, identification of electronic fraud, recognizing exploitation of house tax and detection of frauds due to money laundering (OECD, 2004a; OECD, 2004b)

A large sample of labelled data from Mellon Bank was firstly used to implement NN (neural network) techniques in 1994 (Ghosh and Reilly 1994). A combination of decision tree and Boolean logic method was used to find the difference between normal and fraudulent transactions. Methods involving neural networks and Bayesian belief are also employed for detection of fraud. A fraud detection model was introduced which used Artificial Immune Systems to decrease the cost, reduce response time of system and improves accuracy. The fraud risk management model is introduced at Alibaba by using actual big data processing and intelligent risk framework (Chen, Tao, Wang et al. 2015).

Artificial intelligence and data mining techniques are used to investigate and process a large amount of data during audit planning. As a result, significant rules and patterns are recognized, which helps the organization in the prediction of tax frauds. This knowledge is used to increase profits or to decrease costs. The data mining techniques such as logistic models, artificial neural networks, the Bayesian network, and decision trees are commonly used for fraud detection. The supervised learning methods which use association rules to improve the tax evasion performance of VAT are applied in Taiwan (Wu et al. 2012). Fraud plans are identified by using association rules combined with two dimension-reduction methods thus creating a fraud scale to rate the taxpayers in brazil. A clustering algorithm named the SOM method is used to find about similar behavior of taxpayers. (González et al., 2013). The Cluster Analysis approach can also be used to manage data in an organized manner to get corresponding groups. This technique is also helpful in finding the companies which are at high risk of fraud detection (Dias et al., 2016). Data mining techniques such as Probabilistic Neural Networks (PNN), and Logistic Regression (LR) are used to find the companies that are involved in financial statement fraud). Data mining techniques are used in India to suspect probable tax evader by taking into consideration large data of tax returns in VAT business entities of 2003-2004. Prediction models are developed for gross profit, tax growth and income rate. Discriminant function and classification tree models is used to analyze a large group of taxpayers according to given data. This technique was appropriately efficient with a good strike rate and output.

Peru is the first country which employed neural networks, artificial intelligence and data mining techniques for fraud detection. Developed model was enhanced by the application of fuzzy rules and classification and regression trees (CART) in 2004 for selection of specific variables (Torgler, 2005).

Self-organizing maps were applied to data provided by Australian Taxation Office (ATO). In this context eighty-nine different features are taken into consideration, which involved income profile, market segments, collection or accumulation profile, tax avoidance plans involvement, etc. Higher accuracy in clustering was obtained by using Z-score and min-max normalization techniques. This method helped to detect unusual and abnormal clusters, which in turn identifies financial frauds (Williams et al. 2007). The relationship between payment of taxes and use of false statements was established for an audit selection strategy in Chile by using data mining techniques, including self-organizing maps, neural gas, and decision trees. This technique helped to detect doubtful activities at different phases of fraud detection (Gonzalez and Velasquez 2013). The Minnesota Department of Revenue (DOR) also used many data mining approaches for enhancement of audit selection task.

Almost 150 medium and large Chilean companies were surveyed for fraudulent activities in 2006. According to results 41% of companies suffered losses due to frauds (Chena, Huang, and Kuo 2009). Large amount of data is required to solve fraud detection problems. Large amount of data related to fraudulent transactions is processed which require statistical analysis. This analysis is done by using fast and efficient algorithms. In such cases data analytics yields appropriate tools and help to understand and interpret the process-taking place behind analysis of data.

The tax control program of Australian Tax Office employs risk models by using data mining techniques and statistical analysis to uncover unusual patterns and anomalies. The techniques used were logistic regression, decision trees and SVM (US Government Accountability Office, 2004). Self-organizing maps (SOM) are used to detect "Hot Spots" which are unusual small clusters and abnormal sub-populations. Clustering algorithms like k means, diagrams and visuals are detected during the procedure, which are easily understood by non-technical users (Denny and Christen 2007).

2.6.5 Outlier

n outlier is defined as "an observation that diverges so much from normal observations as to generates suspicion that it has generated through unusual process" (Hawkins 1980). An outlier deviates exceptionally from the other segments of data (Barnet and Lewis 1994). An outlier is also defined as "an observation which is conflicting with other data sets". Outlier is termed as anomaly, abnormal, unusual and divergent in data analytics. Deviation or anomaly can be defined as a stage in time where system behaves abnormal and unfamiliar which is different from expected and usual behavior (Ahmad, S., Lavin, A., Purdy, S., & Agha, Z. 2017). Outlier detection methods are applied to innumerable fields such as VAT fraud detection, credit card fraud detection, detection of measurement errors, clinical inquiries, analysis of athlete's performance, weather forecast, data obtained from sensors, checking for irregular voting, data cleaning, network interruption, geographical information and data mining (Penny and Jolliffe, 2001)

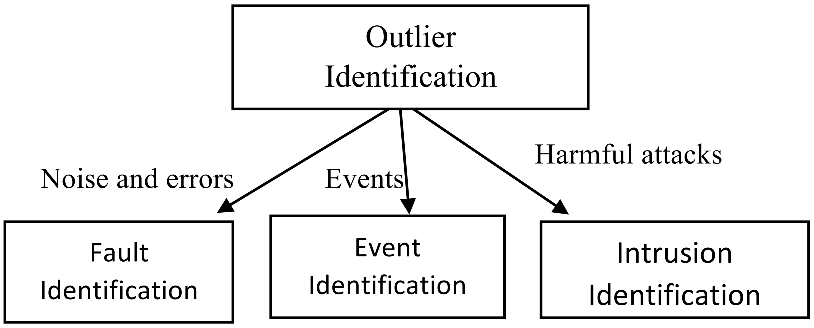


Figure 2 Outlier identification

2.6.6 Types of Outliers

Outliers can be classified into three categories. (Shin Ando 2007).

1. **Point Outliers:** If the single data point is exceptional or abnormal with respect to rest of the data, then this data point is called point outlier.

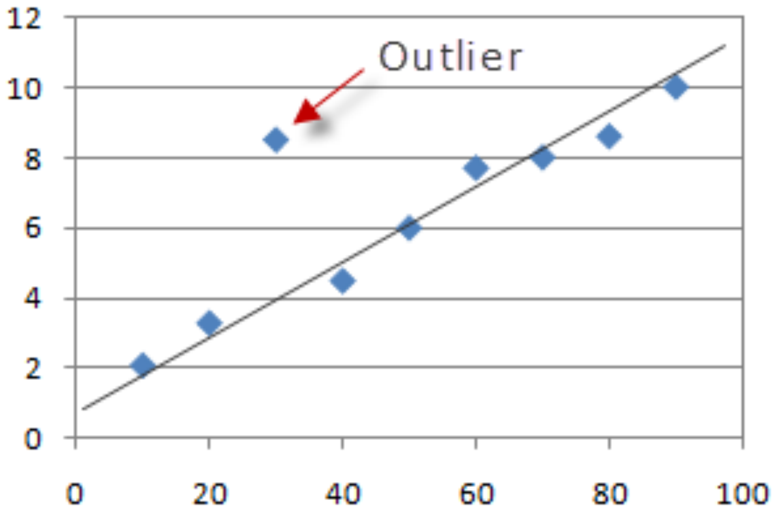


Figure 3. Point outlier (stack exchange)

2. **Contextual Outliers:** If an abnormal behavior occurs at certain instance in terms of specific circumstances, then it is called contextual outlier e.g. in case of credit card theft, unusual expenditures are detected.

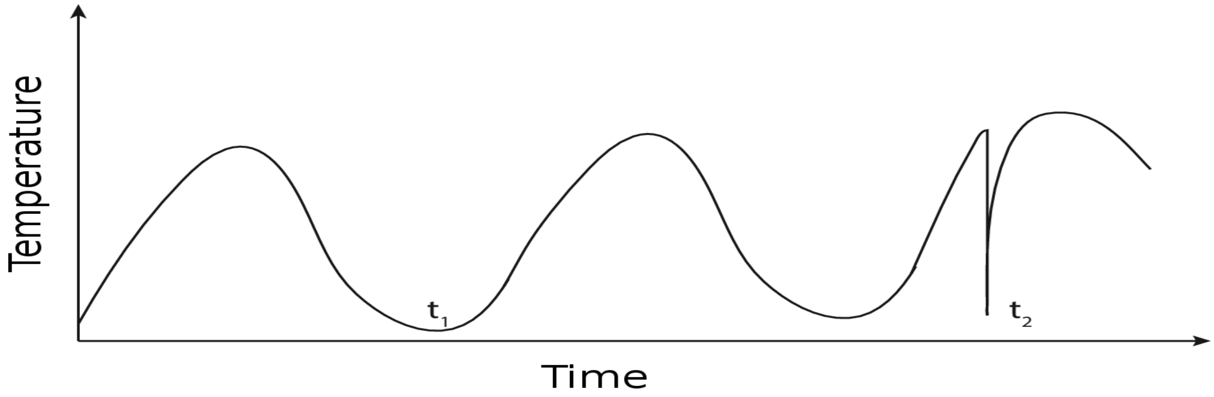


Figure 4. Contextual outlier (Source: Wikipedia)

3. **Collective Outliers:** If a collection of data points is abnormal as compared to complete data set then, it is called collective outliers.

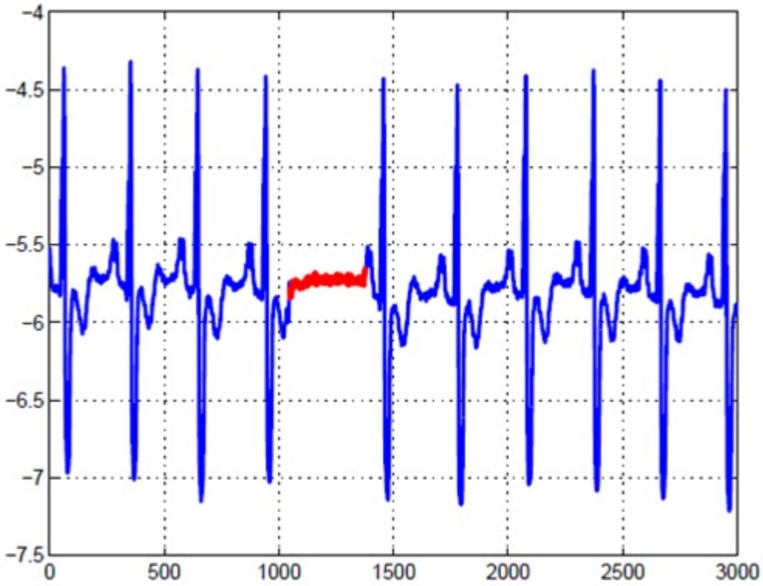


Figure 5. Collective outliers (research mining)

2.6.7 Clustering

The process of recognizing logical clusters and reasonable collection of items in a huge data set on the basis of resemblance behavior is called clustering e.g. Euclidean distance (Jain et al. 2000). Clustering is the chief procedure in AI (Artificial Intelligence). Clustering plays an important role in identification of patterns, designs, data mining and machine learning. Clustering technique is formulated on different types of items and contrast between them by using the functions of distance and density to make up the framework for classification (BianZhaoQi and ZhangXuegong 2000). Clustering technique is applied to many fields such as machine learning, image processing, color image estimation, data mining and compression (G. Hamerly 2003). A cluster can be recognized by its centroid. Also, data clustering is difficult task because clusters formed can be of many different shapes and sizes (Jain et al. 2000). Clustering techniques are basically of two types which are partition clustering and hierarchical clustering. A cluster tree or dendrogram is generated in hierarchical clustering by using merging methods or analytical splitting. Partition clustering break down the data set into specific number of clusters with specified centroids which are non-overlapping. The main goal of this algorithm is to decrease the value of SSE (square error function). Also, partition clustering is more accepted and approachable than hierarchical clustering because they are more robust in pattern identification and interpretation (Jain et al. 2000). The algorithm used for partition clustering is iterative.

2.6.8 K-Means Clustering

The iterative K-means procedure is the most recognized method which is used all over the world. It is based on k-means clustering algorithm which was first presented by J.B. McQueen. It is also called as incremental algorithm. k-means clustering is unsupervised technique which is used in data mining and identification of patterns. K-means algorithm is based on minimization of squared errors and index of clustering execution index. (Jain A K, Dubes R C. 1998). K-means clustering always show convergence but it is bound to find the restricted lowest solution rather than comprehensive solution. So, it is difficult to find the most appropriate partition.

2.7 Use of Big Data in Tax Fraud Problems

Data analytics is beneficial for processing huge and complex data by applying specific tools, techniques and algorithms to discover anomalies, trends, models, irregularities, rules and patterns. This analysis and procedure cannot be done manually. Advanced data analytics not only deal with LTO (Large Taxpayer Offices) but it can also handle average and small and taxpayers. Following steps should be followed to effectively apply Data analytics.

Analytics Model should be built by using basic knowledge about the company and parameters and indicators used by industrial sector.

- Use appropriate technology which include algorithms, models, trends and proven ideas.
- Pre-processing the data to get high quality data. High quality data gives excellent results.
- Apply conventional fraud detection techniques of data analytics by changing dimensions.
- Fraud detection models and standards should be improved constantly.

Chapter 3 Data Analysis

3.1 Dataset Collection

Data is growing at exponential rate due to development in science, technology, engineering, business and World Wide Web. This huge amount of data cannot be handled manually. Also, this data is not well organized or systematic. Today, traditional methods are not enough for processing huge amount of data. These problems in data resulted in emergence of data mining and data science. The extraction of useful knowledge and information from immense and vast amount of data is called as Knowledge Discovery in databases (KDD) or Data Mining (Jaiwei Han and Micheline Kamber, 2011). Big Data can be defined as data having high volume, velocity and variety which require an efficient and high-quality processing.

The standard, quality, efficiency and performance of extracted data depend not only on the method used for data mining but also on the adequacy and quality of data. Many factors such as noise, missing values, inconsistency, useless data and huge sizes greatly affect the results of data mining techniques for extraction of data. High quality data results in high quality knowledge and vice versa. Data processing is a very important phase because it gives the final data set which is then used for data mining.

In this step, all the data sources from which tax evasion data is collected are identified. In this research data is collected from OECD. The data from the following source is collected: OECD Revenue Statistics 2020 http://oe.cd/revenue-statistics. This data involves taxes on personal income, profits, gains, payroll taxes, taxes on property and taxes on goods and services (VAT). This is the most important step because data is the basic building block of analysis in data analytics. Selection of data plays a vital role in building analytical models. All the data that is collected from federal OECD is then gathered and converted in a form of data mart.

Tax data is extracted, modified and stocked in data warehouse structure. Data is stored and controlled in multifaceted database. Data is made available to information technology experts for analysis. Analysis of data is done by using appropriate software. Tax structure in America is as follows:

Sr. No.	**Tax categories**	**Tax type**
1	Taxes on income, profits and capital gains	Income taxes and gains of companies
		Payroll taxes
		Personal income taxes and gains
		Withholding tax
2	Property taxes	Tax on property transfer
3	Consumption taxes	VAT
		Excise duties
4	Taxes on trade	Export duty
		Custom duty

Table 1. Tax structure in USA

Data is collected from different sources and then evaluated to find conclusion. So, data mining is a subgroup of data analytics. Data analytics is further divided into three categories:

EDA (exploratory data analysis)

It is the first stage when there is no knowledge available about data relationships. At this stage hypothesis is developed and new patterns related to different characteristics of data is explored. Mostly EDA techniques are visual and graphical. Statistical graphs are used to plot the data to get understanding of the data.

CDA (confirmatory data analysis)

At this stage testing is done to find either hypothesis is true or false. Results are then applied to complete data sets. Two types of relationships are proven which are cause-and-effect and casual. Commonly Online analytical processing (OLAP) tools are used in CDA.

QDA (qualitative data analysis)

In this case data mainly consist of images and text. So, the data is non quantitative or non-numerical data. It is mostly used in organizational audits of controls, procedures, plans, policies and processes.

Fundamental exploratory data analysis is implemented by using OLAP (online analytical processing), so that multidimensional data can be analyzed (e.g. roll up, drill down, slicing, dicing). Now the data is cleaned. This is done to remove all the conflicting data such as misplaced and duplicate data. Several other transformations are also taken into consideration such as binning, alphanumeric to numeric, geographical aggregation etc.

3.2 Outlier Detection

Outlier detection is an important and crucial activity in tax fraud detection as it plays important role in identifying and preventing fraudulent activities. Outliers emerges due to many reasons such as fault in system, human error, malicious behavior and fraudulent activities This model will work with irregular and exceptional data. As the data we are using is not typical. Predictive analytical models are very successful in detecting frauds because outlier model has the ability to analyze abnormal data that deviates from normal. Outliers models are very helpful in detecting frauds in transactions and credit cards because they can assess the amount of money lost, location, purchase history, time and the nature of the purchase. Outliers are the exceptional values that vary greatly from the other values in a data. They signify changes in measurement, errors in experiments or unconventionality. Outliers are categorized as univariate and multivariate. Univariate outliers are implemented in this research. Univariate outliers can be found when looking at a distribution of values in a single characteristic space. The univariate outlier is calculated for our variable known as amount of tax paid. Point outliers, contextual outliers, or collective outliers are the types of outliers which are used according to circumstances. Common errors of outliers that we are considering in this data set are data processing errors (data handling or data set unintended changes) and sampling errors (extracting or mixing data from inaccurate or various sources).

3.3 Approaches for outlier detection

Normally, three basic approaches are employed for detection of outliers which are as follows:

Approach 1

This approach is employed when there is no initial information available about data. It is similar to clustering which is unsupervised. The data is processed as statistical distribution by locating the farthest points thus, identifying outliers. It is assumed that mistakes and errors are far and different from normal data. Diagnosis and accommodation are two techniques used to deal with this approach. This approach is selected for my research.

Approach 2

This approach is employed when both normality and abnormality is present in data. It is similar to classification which is supervised. It requires the data which is labeled and tagged properly as normal data and abnormal data. Classifiers work best with the statistical data as in classification it is needed to build model from the basic rules.

Approach 3

This approach is employed mostly when data is normal or in very small number of problems where data is abnormal. It is similar to detection or identification which is supervised. This approach only takes into account the data which is labeled as normal. If a new data enters the system, it is recognized as normal if it is within the boundary and recognized as novel if it lies outside the boundary.

3.4 Univariate statistical model for fraud detection

Statistical models are the most initial algorithms that were used for fraud detection. This is the best approach for one dimensional and univariate data sets. Grubbs method is used for one-dimensional data sets in which Z score is calculated by finding the difference between the raw number and the mean value of sample and then dividing the difference by the standard deviation. (Grubbs 1969). Significance level of 1% to 5% is compared with the Z value calculated for the raw number. Data is used to define all the parameters or characters so there is no need for user parameters. Huge datasets result in model which is presentable statistically.

When the data is distributed normally, it is said to be symmetric. If there are nearly equally distributed high outliers or low outliers then, the effect of outliers is minimum on the mean because they counterbalance each other. Commonly outliers are detected using Z score e.g. if all the observations that are three standard deviations away from mean value of sample are excluded, then two authentic readings are deleted for every 100,000 finding. Outliers are found in even very small data sets. A decision criteria ά is set for selecting a starting point which help to tag an information as an outlier. The value α is then divided by two by assuming that outliers are equally distributed on each side of distribution. The value of decision criteria should be kept small.

It is assumed that the data set is distributed equally and autonomously. It is also assumed that the distribution of variables and type of outlier is already known. Outlier in this research are assumed as point outlier. A model is generated which recognize small number of results and observations which are illustrated arbitrarily from distribution G1, ……., Gr which is different as compared to selected target distribution F. Normal or Gaussian distribution N(μ, α2) is taken into account. Gaussian distribution model make use of MLE (Maximum Likelihood Estimates) to find out the variance and mean of the Gaussian distribution of the data set. The variable μ is termed as mean of the distribution. Variable α is termed as standard deviation and α2 is the variance of the distribution.

$$\mu = \mathrm{E}\ [x];$$

$$\sigma 2 = \text{variance}\ [x].$$

$$f(\mathrm{x}) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}} \left(\frac{x-\mu}{\sigma}\right) 2$$

Outlier is then identified as one lying inside outlier region. Limits of coefficient α are, 0 < α < 1.

Normalization constant is as follows:

$$\mathrm{Z} = \sqrt{2^2}$$

For distribution N (μ, σ2), α- outlier area is illustrated as:

$$\text{Outlier}\ (\alpha, \sigma 2, \mu) = \{x\colon |x - \mu| z_1 - \alpha 2^{\sigma}\}$$

Here “zr” is the “r” quantile of N (0, 1).

3.5 Z –Score

Z-Score is also termed as Z-values, Z-ratio, or Z. It is a statistical measurement of a number as compared to the mean of the group of numbers. In a standardized normal curve, these are the points along the base. A Z-value is 0 at the center point of the curve. Z-values to the left of 0 are negative and Z-values to the right of 0 are positive. If a Z-score is to the right of 0 center point, it is termed as above the mean and if Z-score is to the left of 0 center point, it is termed as below the mean. Standard deviations are used to measure the distance from the mean. If the Z-score is 0, it is called 0 standard deviations from the mean and it is equal to the mean.

Z score is standard score when observation is made by pinpointing that at how many standard deviations a data point is away from the sample’s mean. Z-score is calculated by finding the difference between the raw number and the mean value of sample and then dividing the difference by the standard deviation.

$$Z = \frac{(X-\mu)}{\sigma}$$

X represents the raw number. μ is the population mean and the standard deviation symbol is σ.

The Formula for σ standard deviation is as follows:

$$\sigma = \sqrt{\frac{\Sigma(x_i-\mu)^2}{N}}$$

Z-score is standardized all over the world. Z-score is standard score when observation is made by pinpointing that at how many standard deviations a data point is away from the sample's mean. It is irrelevant when comparing currencies such as Canadian dollars, U.S. dollars, euros, or British pounds. In fact, the unit may be measuring height, weight, education levels, or test scores. The Z-score is always referred to the mean that is the center or chosen as zero. The Z-score tells about the distribution of the numbers in a data set and call attention to the extremes.

According to Statistical theory, 99.7 percent of the time, the Z-score will be between –3.00 and +3.00. For 95 percent of the time, it will be between –2.00 and +2.00, and for 68 percent of the time, it will be between –1.00 and +1.00.

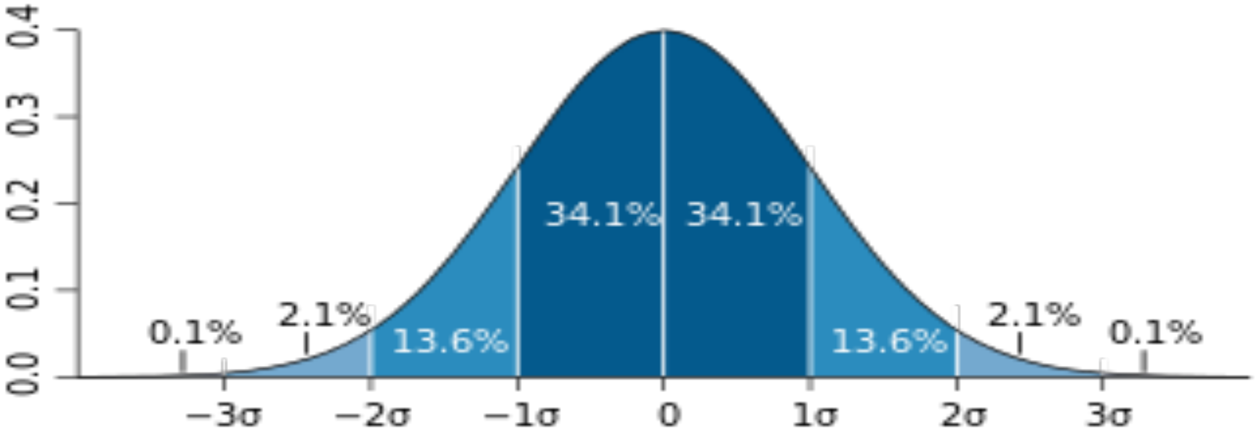


Figure 6. Standard deviation (Source: Wikipedia,2018)

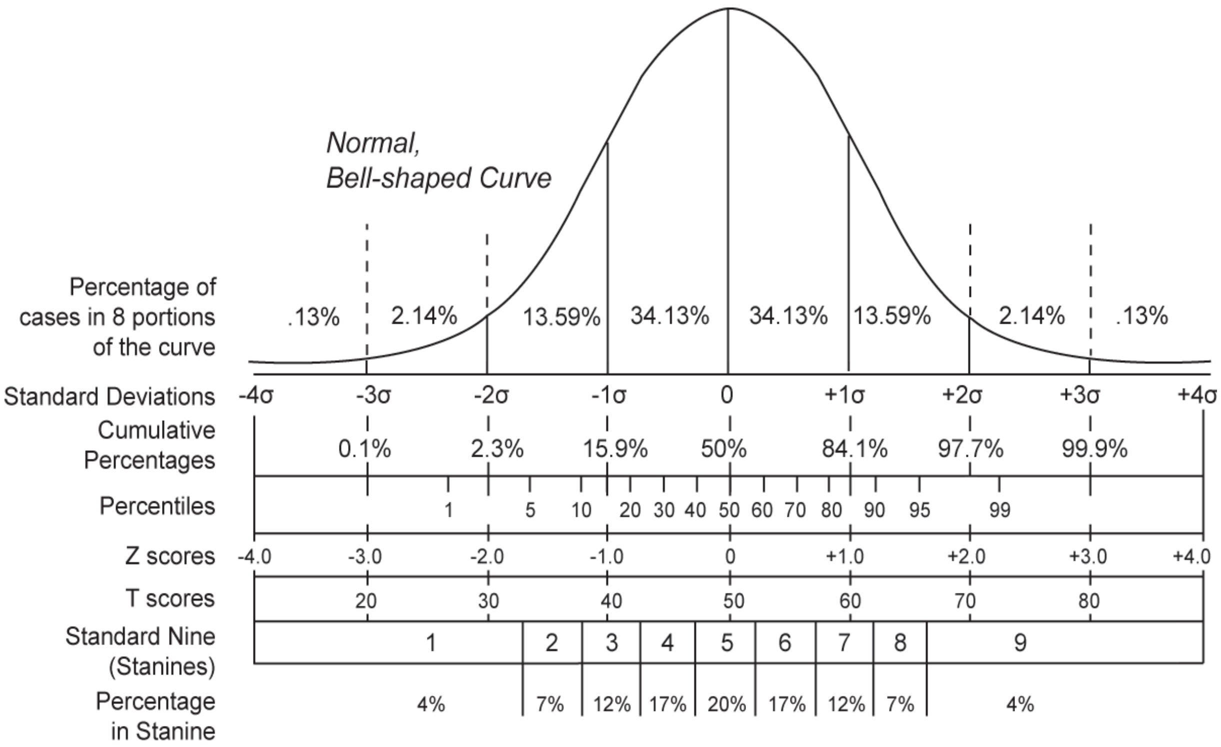
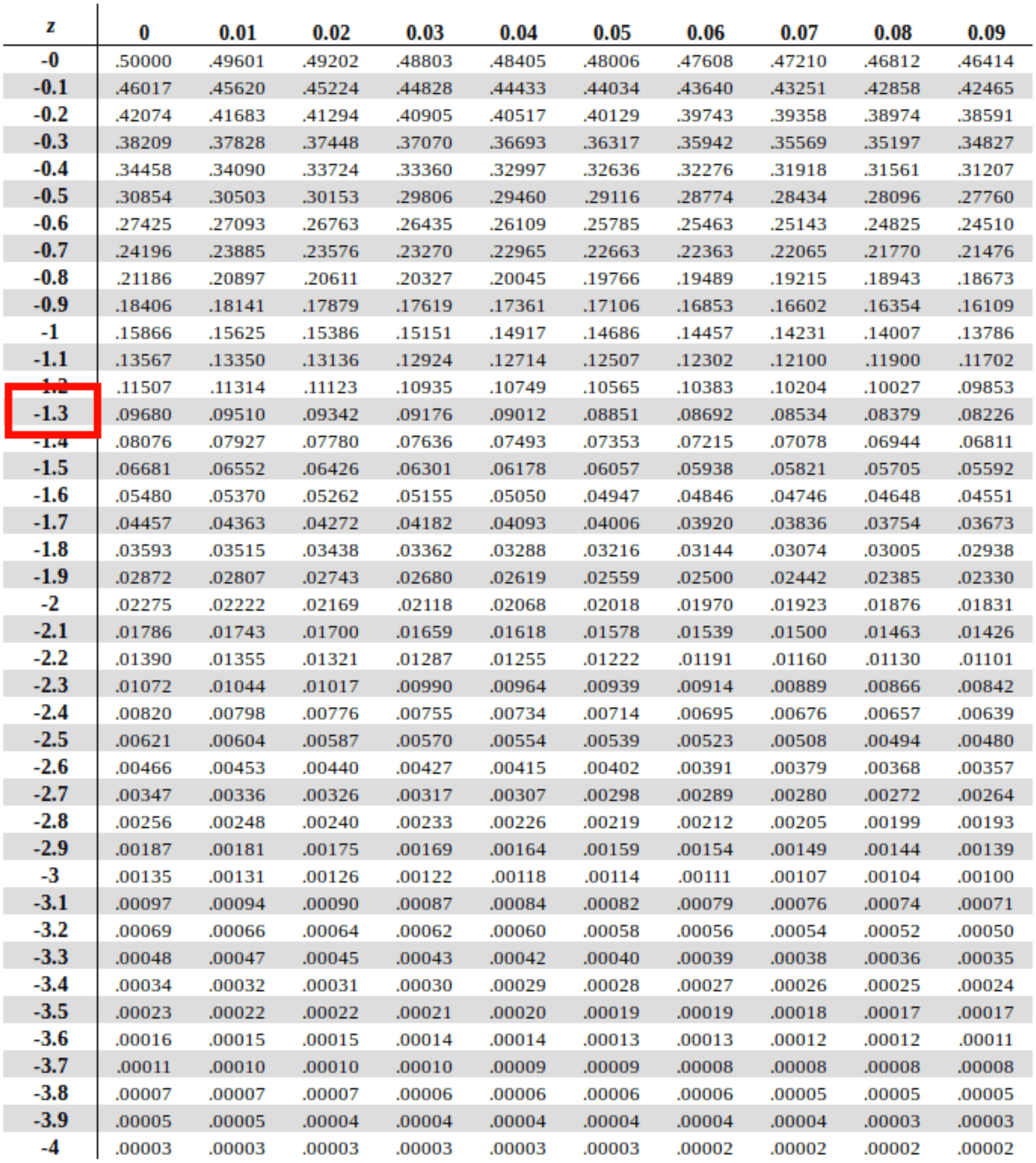
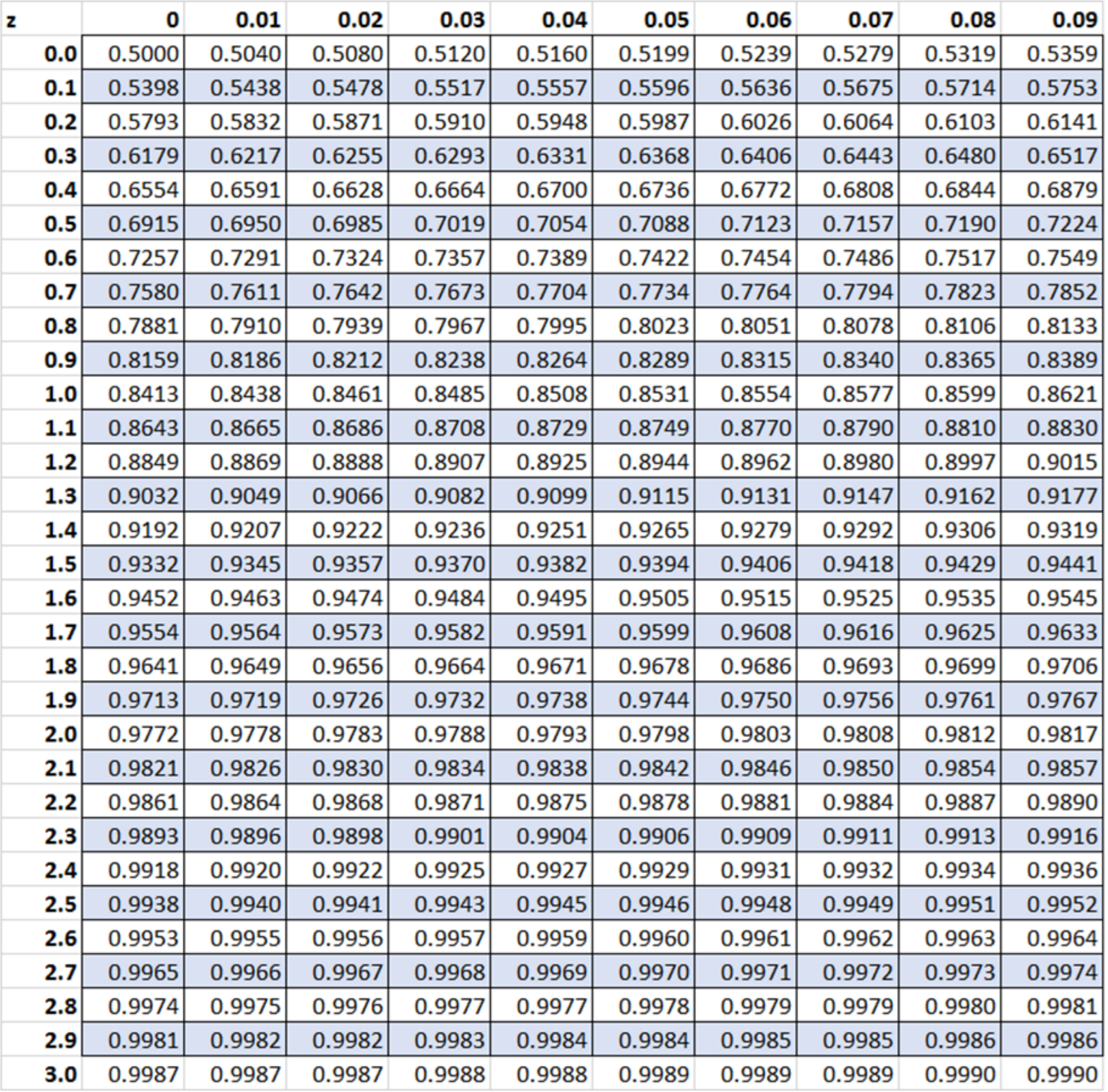


Figure 7. Normal bell-shaped curve (Image by Julie Bang © Investopedia 2019)



z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-0	.50000	.49601	.49202	.48803	.48405	.48006	.47608	.47210	.46812	.46414
-0.1	.46017	.45620	.45224	.44828	.44433	.44034	.43640	.43251	.42858	.42465
-0.2	.42074	.41683	.41294	.40905	.40517	.40129	.39743	.39358	.38974	.38591
-0.3	.38209	.37828	.37448	.37070	.36693	.36317	.35942	.35569	.35197	.34827
-0.4	.34458	.34090	.33724	.33360	.32997	.32636	.32276	.31918	.31561	.31207
-0.5	.30854	.30503	.30153	.29806	.29460	.29116	.28774	.28434	.28096	.27760
-0.6	.27425	.27093	.26763	.26435	.26109	.25785	.25463	.25143	.24825	.24510
-0.7	.24196	.23885	.23576	.23270	.22965	.22663	.22363	.22065	.21770	.21476
-0.8	.21186	.20897	.20611	.20327	.20045	.19766	.19489	.19215	.18943	.18673
-0.9	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109
-1	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786
-1.1	.13567	.13350	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702
-1.2	.11507	.11314	.11123	.10935	.10749	.10565	.10383	.10204	.10027	.09853
-1.3	.09680	.09510	.09342	.09176	.09012	.08851	.08692	.08534	.08379	.08226
-1.4	.08076	.07927	.07780	.07636	.07493	.07353	.07215	.07078	.06944	.06811
-1.5	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592
-1.6	.05480	.05370	.05262	.05155	.05050	.04947	.04846	.04746	.04648	.04551
-1.7	.04457	.04363	.04272	.04182	.04093	.04006	.03920	.03836	.03754	.03673
-1.8	.03593	.03515	.03438	.03362	.03288	.03216	.03144	.03074	.03005	.02938
-1.9	.02872	.02807	.02743	.02680	.02619	.02559	.02500	.02442	.02385	.02330
-2	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831
-2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
-2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-3	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-4	.00003	.00003	.00003	.00003	.00003	.00003	.00002	.00002	.00002	.00002

Table 2. Negative Z score table (Source: www.statology.org)



z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Table 3. Positive Z Score table (Source: www.statology.org)

3.6 Cluster based outlier detection

Clustering is a special method which is used to classify data into different categories, groups and clusters. Clusters can be defined as classes of data items which are closely related to other items in the cluster as compared to items in other clusters. The traditional K-means clustering technique is commonly used in outlier detection. This technique helps to identify the dense spots or areas in the data set and then density evaluation is carried out. After density evaluation, clusters are grouped according to size of cluster. The distance between every spot and its head cluster is evaluated which is termed as anomaly count. Local density cluster-based outlier factor, clustering based Gaussian outlier score and histogram-based outlier score are the well-known cluster-based outlier detection models. Clustering based Gaussian outlier score is employed in this research for the tax fraud detection.

An established segregating clustering algorithm which is used worldwide and has great number of applications is named as Standard k-means. It uses a repetitive resetting procedure to identify k-way strong clustering which help to decrease the imbalance between items of data and group of k cluster illustration. Every illustration is termed as centroid. Euclidean distance is used to measure the imbalance. SSE (Sum of squared errors) is defined as the sum of the squared Euclidean distances of each element to its nearest centroid. SSE (sum-of-squared error) is decreased between the items and cluster centroids {μ1, . . ., μr}

$$SSE(C) = \sum_{c=1}^{r} \sum_{xi \in Cj} \vee x_i - \mu_j \| 2$$

Where $\mu_j \frac{\sum_{x_i \in c_j} x_i}{c_j}$

k means algorithm is based on following main steps:

A data set is broken down into smaller, continuous and levelled subdivisions which are made up of k disjoint clusters donated as $R = \{R_1, \dots \dots, R_k\}$

A random preliminary cluster with cluster centers (μ1, …….μ_k) is created.

For each item $x_i \in x$

Calculate || $x_i - \mu_r$|| as 1 ≤c ≤k

Cluster is assigned x_i according to the closest cluster centroid.

Cluster centroids are rationalized.

The above procedure is repeated again until closing criteria is achieved.

CBLOF stands for Cluster-Based Local Outlier Factor. It uses the clusters to identify the areas with high density in a dataset. Density for every cluster is calculated by using Cluster-Based Local Outlier Factor. In the first step datasets are clustered by using k-means. Probing technique is used by Cluster-Based Local Outlier Factor (CBLOF) to categorize clusters into small and large. Then the distance of each data item is calculated from the cluster centroid to get the outlier score. Outlier score is the multiplied to the data items affiliated to the cluster.

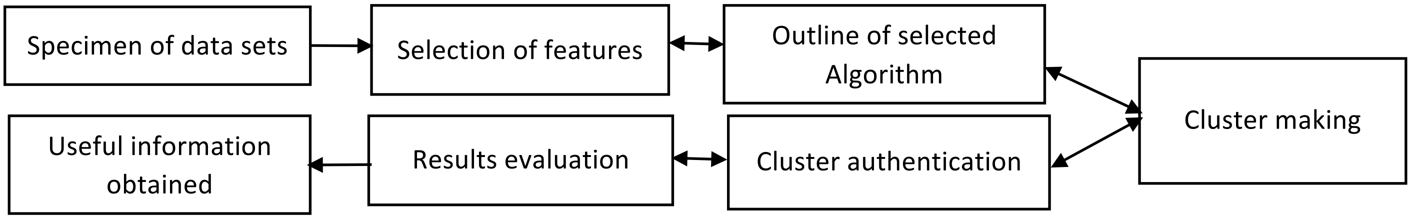


Figure 8. Cycle of clustering

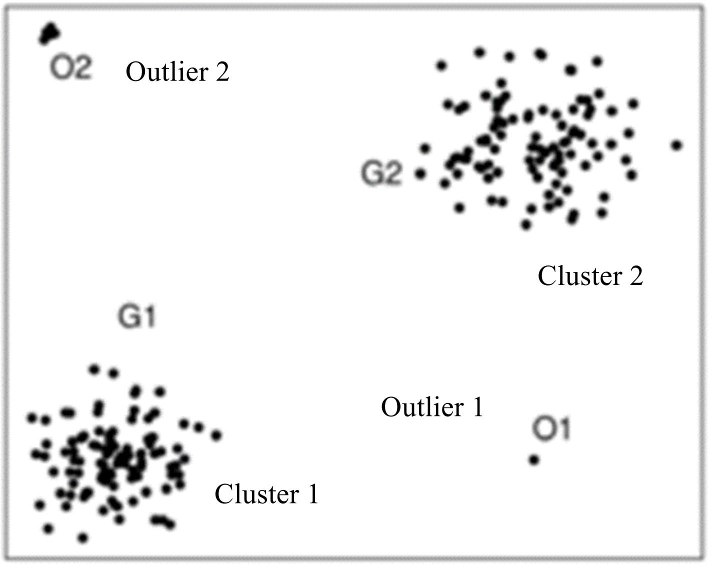


Figure 9. Output of clustering technique

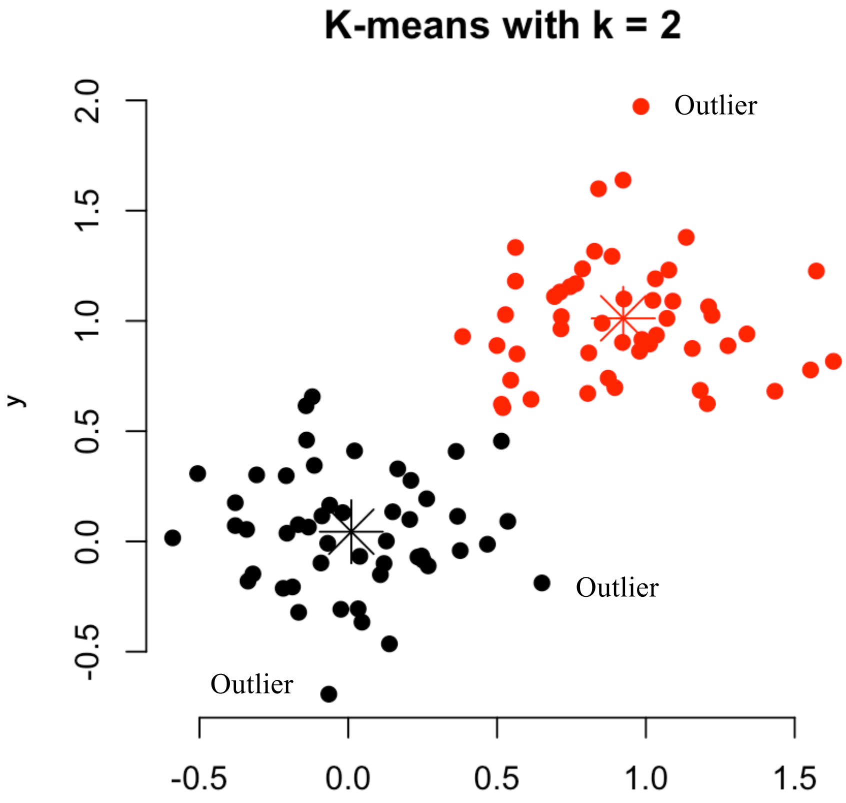


Figure 10. Clustering for k=2

3.7 Model selected analysis and result

The framework of unsupervised anomaly identification is the most accommodating and flexible system as it does not need labeling of data. As the approaches for outlier detection are illustrated in methodology section. Approach 1 which is about unsupervised anomaly detection, is taken into account here.

scikit-learn is the most powerful tool used for data cleaning. It is free and very successful machine learning library from python. It has very effective preprocessing tools for data transformation. scikit-learn also deal efficaciously with noise and missing data. It can fill in the missing values like mean, median, mode etc. by using Simple Imputer class which has many tools for attributing missing values. Scikit also present many advance clustering and classification algorithms. So, scikit is used for data cleaning.

3.8 Standard Deviation

It is a statistical term which calculate the distribution of data with respect to its mean. It is calculated as square root of variance. Farthest the data from mean, higher the deviation in the data and vice versa.

$$\sigma = \sqrt{\frac{\Sigma(x_i-\mu)}{N}}$$

Where,

σ = Population standard deviation

N = the size of the population

x_i = Each value from the population

μ = the population mean

For detection of outliers in fraud mean and standard deviation is calculated and then compared. The default value is specified as 3. Data which is placed outside three standard deviations is termed as outliers. As a general rule, z-scores which is lower than -1.96 or higher than 1.96 are assumed to be unusual and worth studying. That is, they are considered statistically significant outliers.

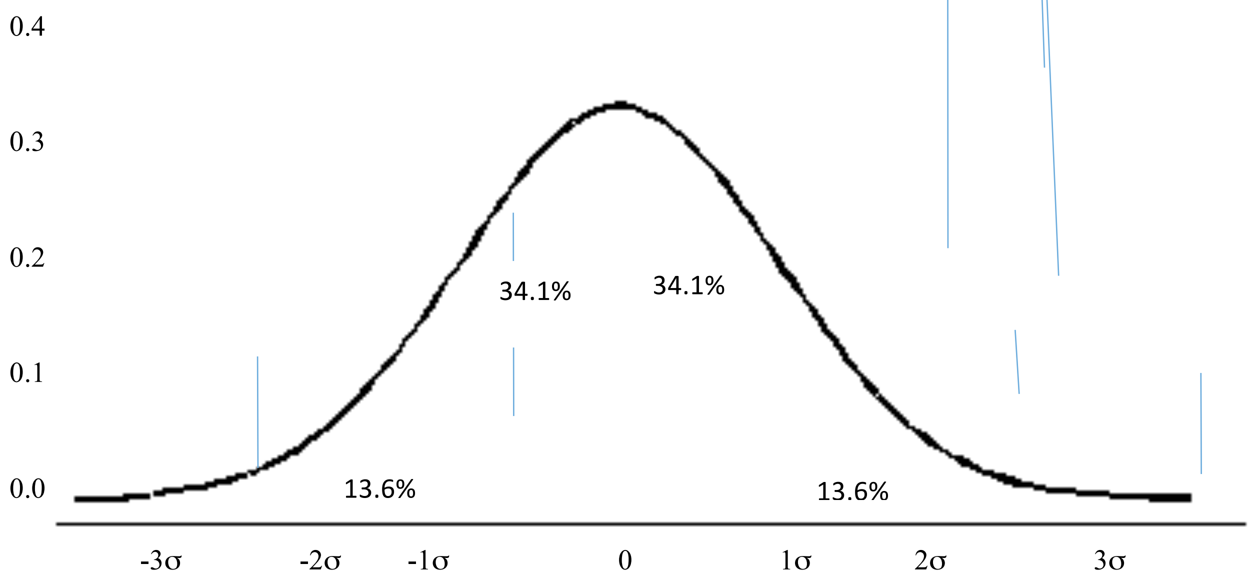


Figure 11 Standard deviation (Wikipedia 2018)

3.9 Languages and libraries used for implementation

Python: is a high-level, powerful and demonstrative programming language. It has strong focus on technology. Python has many useful and helpful libraries like NumPy, Pandas, Matplotlib, and scikit-learn. Structured data can be handled easily by using Pandas and NumPy. Python is also very famous in machine learning.

Pandas: are used for data transformation and data shaping. It is a python package which help to deal with missing values, conflicting and duplicate data. It also detects types of data. Data types which are incorporated by Pandas are float, int, bool, time delta[ns], category, datetime64[ns], and object.

scikit-learn: it is also very effective, useful and free library of python. It deals with clustering algorithms. Outliers can be detected by using anomaly detection algorithms such as local outlier factor.

K means clustering is performed in python. It is very important to understand the algorithm as it is the most crucial step in implementation process. In the first step number of clusters is selected which is k. k centroids are selected. Centroids are representative data elements which are at the center of cluster. The two fundamental components of k mean algorithm are expectation and maximization. In the expectation step each data element is assigned to its closest centroid and in maximization step mean of all the data elements is calculated for every cluster and new centroid is determined.

Algorithm for k means

- Select k number of clusters. K is 3 here.
- Boot up k centroids. Centroids are at first selected at random.
- Different points are assigned to their closest centroid
- Recalculate the cluster centroids.
- Repeat the above process.
- The process of expectation is done by assigning each data element to its closest centroid.
- Maximization is done by calculating mean of data elements in each cluster. New centroid for each cluster is determined

- The process is repeated until position of centroid do not changes.

K-means algorithm aims to reduce the sum of total distance between the elements and their corresponding centroid of the cluster. SSE is thus calculated within centroid.

Selection of k number for clustering

The selection of number of clusters k is based on the concept that how adequately a model performs with specific number of k clusters as clustering play important role in subsequent model presentation. Following two methods ae commonly used for selection of k.

Elbow method

This method of selecting k number of clusters is based on SSE (sum of squared distance) between data items and their corresponding centroids. The value of k is selected at a point when sum of squared distance (SSE) start flattening by creating an elbow. Sometimes it become difficult to select appropriate value of k from graph because the graph starts decreasing monotonically and there is no elbow formation at all.

Silhouette analysis

This analysis is based on extent of segregation between clusters. Average distance is calculated from all data items in the same cluster as ai. Average distance is calculated from all data points to nearest cluster as bi.

Calculate the coefficient as

$$\frac{b^i - a^i}{max\,(a^i, b^i)}$$

- Range of values is (-1, 1)
- If the value is 0 or greater, than the sample is closer to the nearby clusters.
- It the value is 1or greater, than the sample distant from the nearby clusters.
- It the value is -1 or greater, than the sample is appointed to the incorrect clusters.

Therefore, it is needed that the value of coefficient should be higher and closer to 1 to get good clustering. The thickness of the silhouette plot Indicates the size of each cluster.

Criteria for stopping K-means algorithm

There are three criteria:

- Position of centroid does not change.
- Points do no not change their position and remain in same cluster.
- Iterations have reached to their maximum number.

Inertia determines the distance between the points or elements inside the cluster. It actually evaluates the total sum of distance of all the points from center of cluster within the cluster. This value is calculated for all the clusters.

SSE (sum of squared errors) is calculated to find the grade and quality of clustering procedure by comparing repetitive processes. SSE value depiction of errors. So k-means try to minimize this error.

K-means algorithm appear to be nondeterministic when actuated randomly because if the same algorithm is run on same dataset, it will result in different clustering configuration every time. Many initializations are run until the one with minimum SSE is achieved. Python help in powerful implementation of k-means clustering by using machine learning package scikit-learn which is shown as sklearn.cluster.KMeans.

The code used here uses Python Packages and Python is installed with Anaconda.

Required packages are installed by using this code

Shell

```
(base) $ conda install matplotlib numpy pandas seaborne scikit-learn ipython

(base) $ conda install -c conda-forge kneed
```

ipython console or Jupyter Notebook can be followed by using this code. This code will result in import of modules and libraries which are required for implementation.

Python

1. *import libraries*

2. *import pandas as pa*
3. *import numpy as nm*
4. *import random as ra*
5. *import matplotlib. pyplot as mt_plt*
6. *import KneeLocator from kneed*
7. *import make_blobs from sklearn.datasets*
8. *import K Means from sklearn.cluster*
9. *import silhouette_score from sklearn.metrics*
10. *import Standard Scale from sklearn. Preprocessing*
11. *import seaborn as sns; sns.set()*
12. *import sqrt, random, array, from numpy*
13. *import load_tax from from sklearn.datasets*

Specimen of dataset presented

CSV file is studied, and first five rows of data are analyzed

Tax data = pd. read_csv ('clustering.csv')

data.head ()

Statistics of data data.describe () data_statistics.py

ax structure n USA %	**1999**	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**2016**	**2017**	**2018**	**2019**
axes on ncome, rofits and apital gains f which	30	31	32	35	35	36	36	49	50	45	45
ersonal ncome, rofits and ains	22	25	29	26	23	28	28	39	38	39	41
orporate ncome and ains	3	2	2	1	3	4	5	5	5	6	4
ocial ecurity ontributions	12	15	17	10	11	13	20	25	29	23	25

ayroll taxes	-	-	-	-	-	-	-	-	-	-	-
axes on roperty[3]	7	8	10	18	12	12	18	20	19	16	12
axes on oods and ervices	7	8	9	12	13	10	11	15	14	16	18
f which AT	-	-	-	-	-	-	-	-	-	-	-
ther	-	-	-	-	-	-	-	-	-	-	-
OTAL	81	89	95	100	97	117	104	100	100	100	100
ean	-2.54328	-6.77432	-3.54287	-2.87432	-4.54321	-1.79532	-4.45532	-8.43876	-6.85391	-1.96432	-1.74974
td	0.689543	0.775436	0.876956	0.987231	1.345236	0.457895	2.453212	0.983467	1.665743	2.984373	3.785432

Table 4. OECD Revenue Statistics 2020 http://oe.cd/revenue-statistics

								Unit US Dollar, Millions
Year	**1999**	**2000**	**2001**	**2002**	**2016**	**2017**	**2018**	**2019**
Total tax revenue	2,690,195	2,200,519	2,884,730	2,733,431	4,837,980	5,225,865	5,031,748	5,243,793
1000 Taxes on income, profits and capital gains	1,322,433	1,453,865	1,393,729	1,196,508	2,00,410	2,342,790	2,274,596	2,379,105
1100 individual income tax	1,099,080	1,224,538	1,227,295	1,039,681	2,312,410	2,342,790	2,274,596	2,379,105
1110 On income and profits	987,881	1,088,967	1,155,475	984,030	1,807,258	1,845,601	1,858,365	1,972,648
individual income tax federal	812,132	895,677	945,209	795,413	1,458,301	1,485,100	1,480,259	1,568,604
Individual income tax state and local govt.	175,749	193,290	210,267	188,617	348,958	360,501	378,105	404,044
1120 On capital gains	111,199	135,571	71,819	55,651	139,785	191,210	210,457	201,335
1200 Corporate income tax	223,353	229,327	166,434	156,827	365,367	305,979	205,774	205,121

Table 5. OECD Revenue Statistics 2020 http://oe.cd/revenue-statistics

Data should be represented in such a format that it is easy to use. In this research tax data is presented in the form of tables. Tables show the values of tax implemented by government, tax paid and payable tax. Also, fake values are added for implementation of clustering technique. Otherwise, there will be no outliers detected. Source of data is https://stats-3.oecd.org/.

Two variables are taken from data which are tax on goods and services and personal income tax

```
X = data [["VAT", "personal Income tax"]]
```

Visualize data points

```
plt.scatter(X["personal income tax"],X["taxes on goods and services"],c='blue green red')

plt.xlabel('total payable tax')

plt.ylabel('tax paid (In Thousands)')

plt.show()
```

Data can be generated from above GIF by using make_blobs (). It uses following parameters.
n_samples represent the total number of samples which are generated.
centers represent the number of centers generated.
cluster_std represent the standard deviation.
make_blobs() restore assembly of two values:
A two-dimensional NumPy array is shown with the x- and y-values for every sample in a dataset.
A one-dimensional NumPy array is shown which have the cluster labels for every sample in a dataset.
Number of cluster k is selected and random centroid is selected for each cluster.

```
#number of clusters
K=3

# Select random observation as centroids
Centroids = (X.sample(n=K))
plt.scatter(X["total payable tax"],X["tax paid"],c=' blue green, red')
plt.scatter(Centroids["total payable tax"],Centroids["tax paid``"],c=' blue, green, red')
plt.xlabel('total payable tax ')
plt.ylabel( tax paid(In Thousands)')
plt.show()
```

synthetic data and labels are generated as follows:

Python

```
features, X, y_true =make _blobs
true_labels = make_blobs( n_samples=400,  centers=3, cluster_std=3.75,  random_state=50 )
plt.scatter(X[:, 0], X[:, 1], s=60);
```

random_state is left as the default value.

the first five elements for each of the variables is generated by make_blobs() as :

python

```
 features [:5]
      array ([[ 8.76055834,   2.97521122],
          [ -8.72359667, 12.37461902],
          [ -5.71320683, -7.24645821],
          [-11.75175012, -11.65153587],
          [ -7.541037137, -5.64471371]])
 true_labels[:5]
       array ([0, 1, 2, 2, 2])
```

In the next step data is fitted to scaled_features. This command will help to perform 400 iterations for every run.

```
In [5]:  scaler = StandardScaler()
         scaled_features = scaler.fit_transform(features)
         kmeans.fit(scaled_features)
  scaled_features
array ([[ 1.23181128, 1.24614252],
      [-2.43608622, 2.31135734],
      [-0.42234162, -0.46682165],
      [-1.90246492, -1.34562519],
      [-2.30914611, -2.46243432]])
```

The framework used here is as follows:

Initialization technique is controlled by init. init is set to “random” for implementation of k means algorithm.

n_clusters is used to fix the value of k. This is the most important step.

Number of initializations are fixed by using n_init. In the default settings of scikit-learn k means algorithm is performed 10 times until the lowest value of SSE is obtained.

max_iter is used to fix the number of maximum iterations which are performed for every initialization of the k-means algorithm.

Python

```
kmeans = KMeans (init="random",   n_clusters=3,  n_init=10, max_iter=400,
                                random_state=50)
```

when the k-means call . fit(): , initialization will run with the lowest value of SSE which is key feature of k means.

Python

The lowest SSE value

```
 kmeans.inertia_
80.97061105829653
```

Final locations of the centroid

```
  kmeans.cluster_centers_
array ([[ 2.26529386, 1.23255248],
       [-1.35018975, 2.15489876],
       [-1.62947803, -2.39562737]])
```

The number of iterations required to converge

```
   kmeans. n_iter_
```

*At the end the clustering data is stored as one-dimensional NumPy array in kmeans.labels_*

Python

```
kmeans.labels_[:5]
array([0, 2, 1, 2, 2], dtype=int30)
```

we selected k=3. As a result, three centroids are selected randomly for three clusters.
The tax data is used to calculate value of SSE for different values of k. the point is noted where the curve is flattened and elbow is established.

```
# Kmeans algorithm is run to get the index of data items clusters. So the elbow method is used to get the value of k number of clustering.
SSE = [] list_k = list(range(1, 10))
for k in list_k:
    km = KMeans(n_clusters=k)
    km.fit(X_std)
    sse.append(km.inertia_)

# Plot sse against k
plt.figure(figsize=(6, 6))
plt.plot(list_k, sse, '-o')
plt.xlabel(r'Number of clusters *k*')
plt.ylabel('Sum of squared distance');
```

3.10 Visualization of results

3.10.1 Visualization of Data

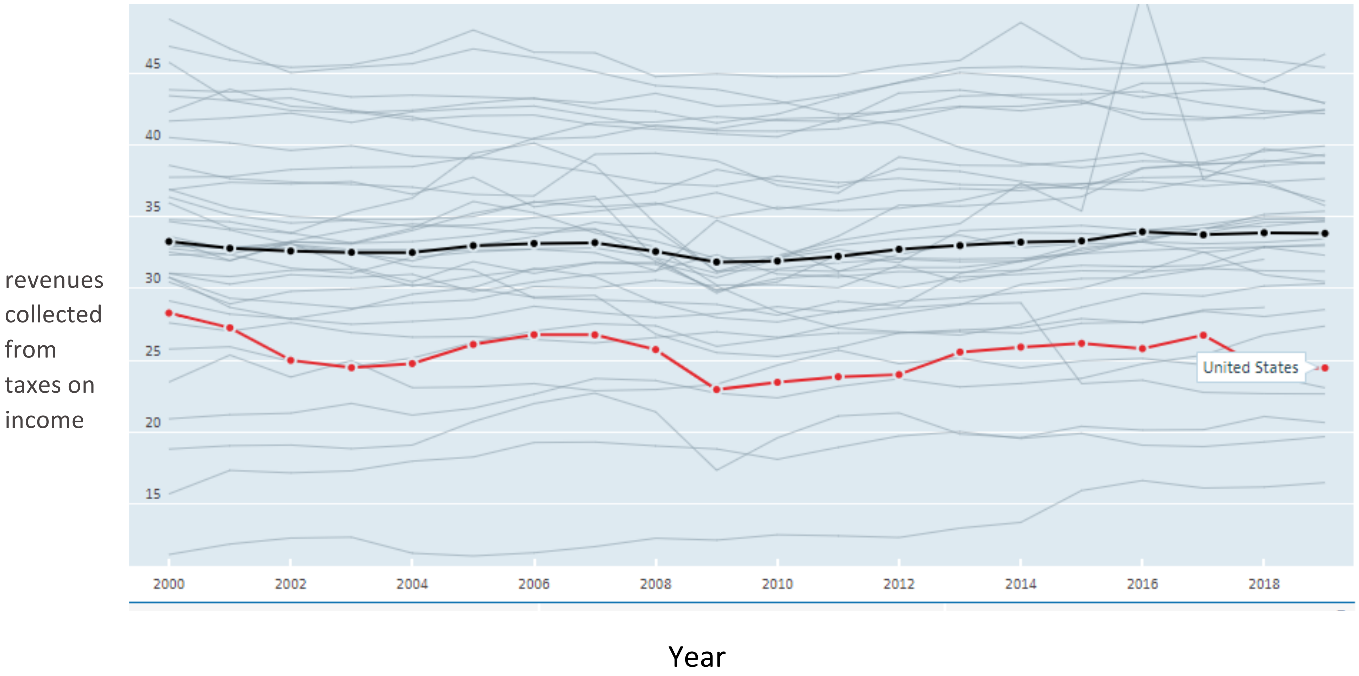
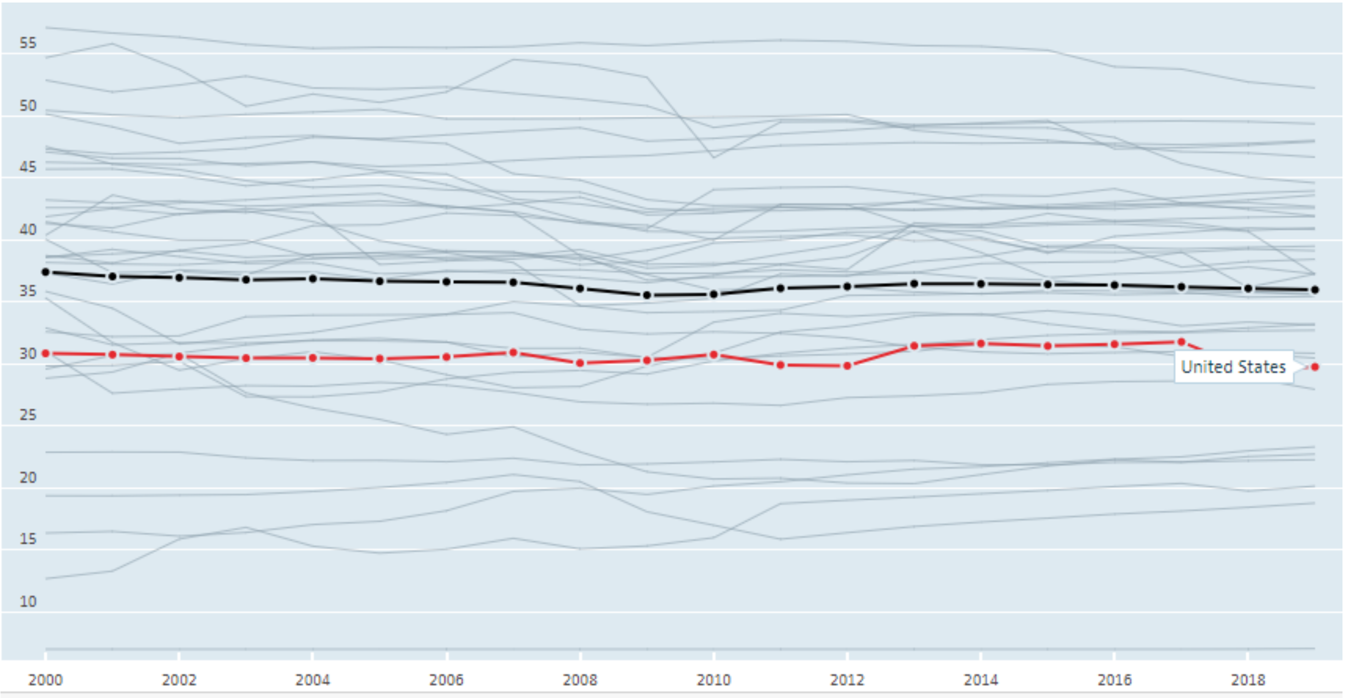


Figure 12 Revenues collected every year

Tax revenue comprise of the revenues collected from taxes on income and profits, social security contributions, taxes on goods and services, payroll taxes, taxes on the property and transfer of property. Total tax revenue is expressed as percentage of GDP which shows production of country which depend on tax collected by the government through taxes. The above graph indicates the relationship between revenues collected every year in USA. Red line shows the changes taking place in revenue collection over the years. It shows how revenue collected changes every year. Black line shows the OECD average.

Tax paid on personal income



Year

Figure 13 taxpaid on personal income every year

Tax on personal income can be defined as the taxes collected on the net income which include gross income excluding permissible tax relaxations. It also includes capital gains of individuals. This graph indicates the relationship between personal income tax collected during different years. Red line shows the changes taking place in tax paid on personal income over the years. Black line shows the OECD average.

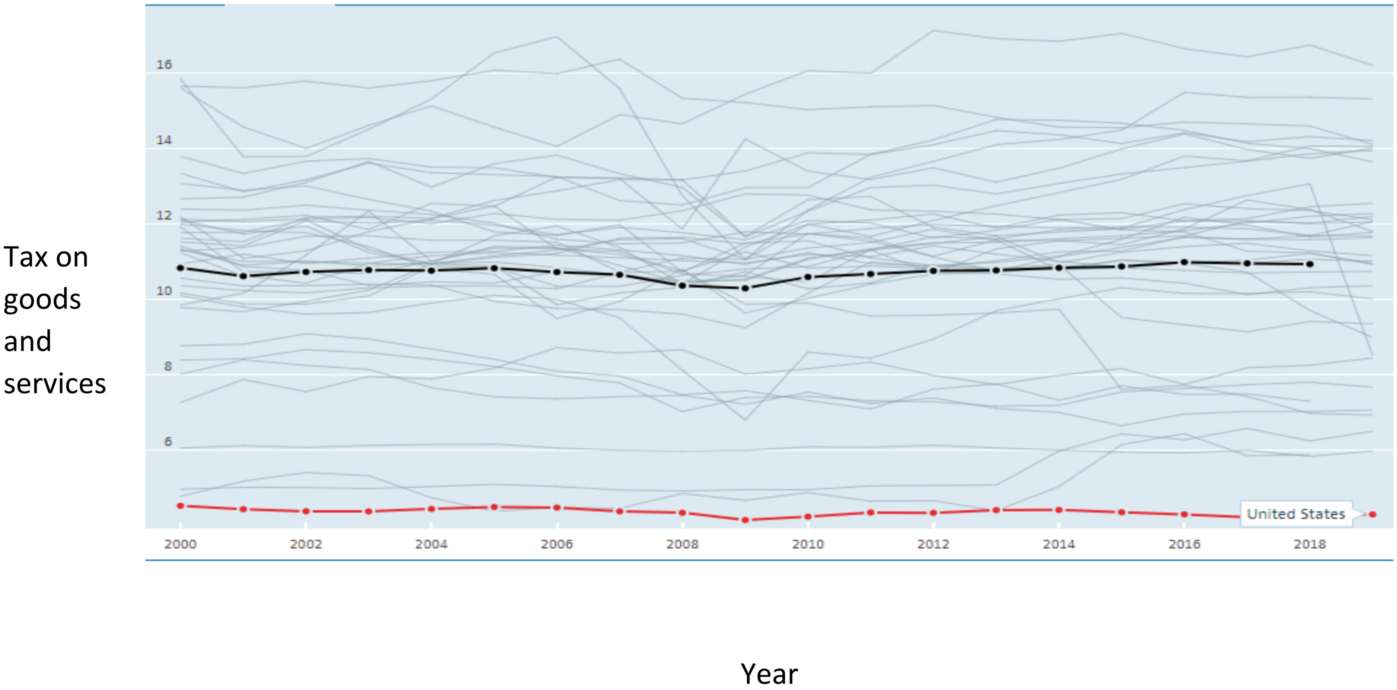


Figure 14 Tax on goods and services every year

Tax on goods and services includes all taxes collected from the production, extraction, sale, transfer, delivery of goods, and the rendering of services, or on the use of goods or permission to use goods or to perform activities. They mainly comprise of value added taxes and sales taxes. The above graph shows relationship between taxes on goods and services over the years. USA has the lowest taxes on goods and services. Red line shows the changes taking place in taxes on goods and services collection over the years. Black line shows the OECD average.

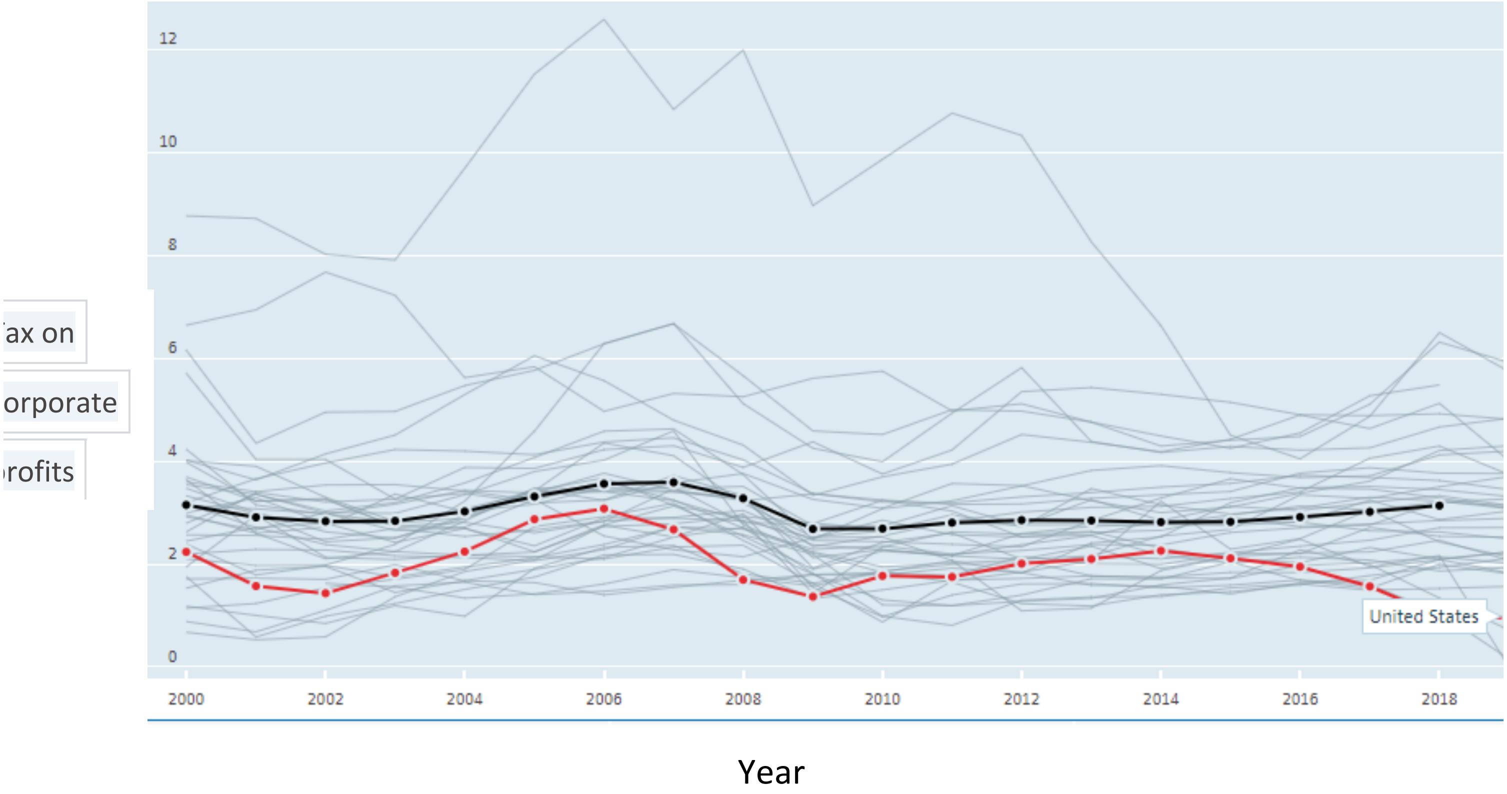


Figure 15 Tax on corporate profits

Tax on corporate profits is defined as taxes collected on the total profits of enterprises and companies. This graph is plotted between taxes on corporate profits collected over the years. Red line indicates the taxes on corporate profits in USA. Black line shows the OECD average

3.10.2 Visualization of Cluster Analysis

At first elbow method is used to determine value of k which turn out to be k=3 and initialization is done. Clustering classifies the 20 years' tax data into three clusters.

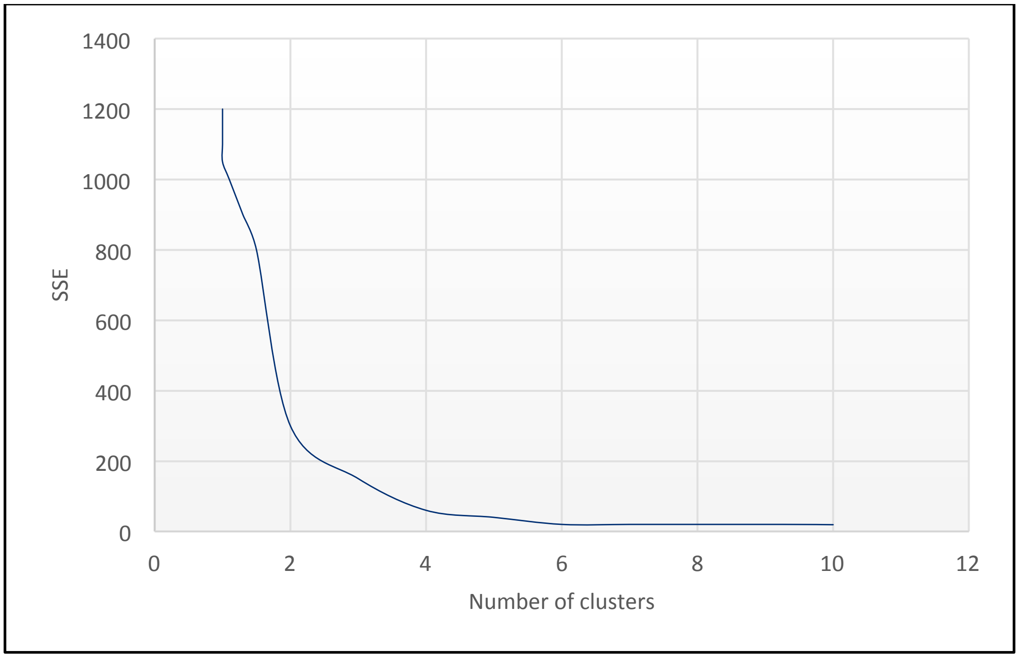


Figure 16 Elbow method for k=3

With every iteration position of centroid for each cluster keep on changing until the lowest value of SSE is attained. At this point centroids are converged. They no more change their position. Final results of clustering and iterations are shown in figure.

Number of clusters	3
Variables standardized	Yes

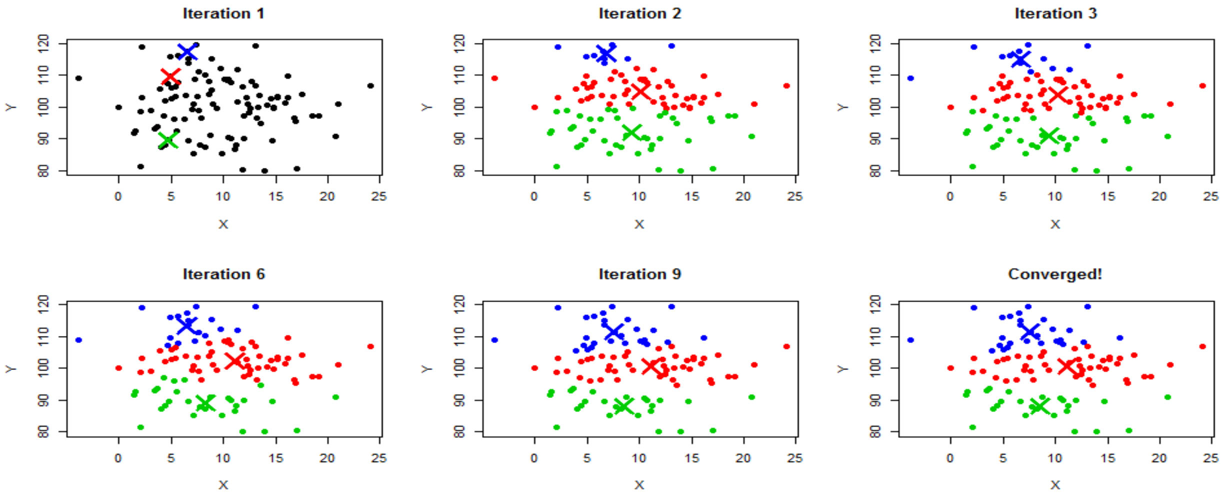
	observations	Sum of squares within clusters	Average distance from centroid	Maximum distance from centroid	color
Cluster1	10	1.453	0.45	0.775	blue
Cluster2	20	7.789	0.987	1.892	Red
Cluster3	32	13.456	1.453	1.345	green

Table 6 Final Partition results

	Cluster1	Cluster2	Cluster3
Cluster1	0.0000	2.5438	5.2341
Cluster2	2.5438	0.0000	3.4978
Cluster3	5.2341	3.4987	0.0000

Table 7 Centroids of clusters

At first iteration three centroids are located which are blue red green. In the second iteration three clusters are formed with three centroids as is shown in the following figure.



Iteration 1

Iteration 2

Iteration 3

Iteration 6

Iteration 9

Converged!

Figure 17 Iterations for k=3

After nine iterations the centroids are the same and at this point centroids are converged. They no more change their position.

In the following table, the different variables in terms of clusters are presented.

Variables	Cluster 1	Cluster 2	Cluster 3	Grand centroid
Personal income tax	1.1432	0.4453	-0.8972	0.0000
Taxes on goods and services	1.1952	0.5432	-0.5156	0.0000
Total payable tax	1.9342	0.7432	-0.7845	0.0000

Table 8 Distance between centroids of clusters

CHAPTER 4 CONCLUSION

4.1 Conclusion

Fraud detection is the vital part of modern and growing financial sector. Frauds not only cause huge financial losses but also stigmatize the reputation of an organization. Nowadays, fraud risks of different forms, shapes and dimensions are emerging and the chances of committing fraud are increasing due to huge developments in technology. In this thesis, both statistical and computational techniques are used to detect fraud in taxation.

Data mining techniques are very helpful for detecting fraud in financial sectors. Data can be analyzed from many different perspectives to classify it and to compile the results in the form of relationships and patterns by using data mining software. Thus, useful information is extracted. Financial fraud is the biggest issue in this modern economic world which is growing with every passing day. So, it is very important to detect fraud, but it is very difficult to deal with large amount of data. Hence, many data mining techniques are employed for fraud detection. The areas which are highly affected by frauds include Insurance, Banking, Health and Financial Statement Frauds. Data mining is a process which involve analysis of large amount of data to get results from different patterns, trends, graphs and correlations by using technologies, statistical and mathematical methods. Several techniques are implemented.

In this research, detection of fraud in the field of taxation is discussed. This paper represents different types of financial tax fraud, reasons of tax fraud, impacts and outcomes of tax fraud, methods to reduce tax fraud, use of technology in tax fraud detection, need for fraud detection systems, different current fraud detection techniques, and the chances of future works.

The main goal of this study is to apply technological methods in the field of tax fraud detection. Data mining is used in number of fields to get useful information from the data. Technology is growing at a very fast pace. The key contribution of this work is to provide solution for fraud detection by using new technologies. Data mining is growing and evolving with every passing day. Organizations use these powerful data mining tools for processing huge amount of data to get useful knowledge. Modern technology has revolutionized due to robust processing power, hard disk storage capacity and statistical tools. As a result, precision of analysis has greatly increase with a decrease in cost.

Any criminal act which result in personal and financial benefits is termed as fraud. An Illegal and unlawful intention to get gains by using unauthorized and wrongful measures and methods is termed as financial fraud. Frauds greatly effect economy, industry and daily lives of a common man by destabilizing business, industries, economy and by changing daily life cost of a person. Conventional methods used for detection of fraud include auditing, which is inefficient to deal with large amount of data. But modern computer intelligence systems help to identify even the smallest anomaly. The two main aspects of fraud are use of deceitfulness or cheating and an effort to conceal or camouflage corruption. It is not necessary that the gains obtained by fraud are materials or financial benefits. They might also be intangible. The benefits can also be gained by third party instead of the person who is guilty of fraud.

Fraud control and prevention cannot be achieved by single isolated effort. Also, by controlling specific areas or activities, fraud cannot be controlled. It needs the collaboration and participation of members of whole institution. Fraud detection and prevention is a continuous struggle. Frauds can be of many different types and its detection is becoming difficult with every passing day due to development in technology. Fraud has become business nowadays. Fraudsters earn millions of dollars as a result of fraudulent activities. Therefore, prevailing tax fraud detection techniques which do not involve implementation of data analytics to huge tax data are insufficient. Also, these techniques are not able to detect fraud accurately. Application of efficient, scalable, structured and extensive tax fraud detection techniques by using data analytics has become critical to detect frauds and to prevent losses.

Fraud detection and prevention systems must be capable of detecting fraud in real-time by improving the credibility of the huge dataset. Behavior of user should be analyzed to find hidden patterns and correlations. Selection of appropriate machine learning technique depend on many factors such as problem identification, type of data, size of dataset, and resources. Machine learning algorithms need well defined and well-prepared dataset. Machine learning algorithms need significant expertise to build, understand and implement complex and powerful algorithms. Therefore, companies hire data science experts for the purpose of fraud detection. This help to expedite development and growth.

Commonly, rule-based approaches use algorithms which are implemented manually. They need adjustment of situations and can barely identify obvious relationships and patterns. On the other hand, machine learning algorithms detect hidden correlations between behavior of user and fraudulent activity. Machine learning approaches helps to develop algorithms which can deal with huge datasets and different variables. These techniques process the data at higher speed and need less manual effort. They also detect fraud activities automatically with a smaller number of confirmation steps and measures. Fraud detection approaches involve anomaly identification. Huge data sets are divided into two sets which are outliers and normal classification. Outliers are data points which vary greatly from normal data and indicate fraudulent activities.

Outlier detection is very useful for detection of frauds and intrusions and analysis of customer behavior. Outlier detection involves detection of anomaly and data points which are extremely different or conflicting from the remaining data set. Wide range of fields use outlier detection techniques such as quality control, fault diagnosis, intrusion detection, web analytics, and medical diagnosis. The goal of outlier detection is to expose those data points which are abnormal and dissimilar to remaining data set. At the same time these outliers contain useful information for fraud detection. Although, these data points comprise of very small percentage as compared to total data. Identification of outliers is very critical and crucial for fraud detection. The patterns have greatly evolved and changed as a result of fraudulent activities. The reason behind this change in patterns is that the fraudsters keep on inventing new ways and designs to spam online systems and people.

It is difficult to label data manually. Unlabeled data is processed by using unsupervised learning techniques and data is grouped into number of clusters to find hidden patterns, designs and correlations between data points. In this way not only the data labelling becomes accurate but also frauds can be detected easily.

k-means clustering is the most common type of partition clustering technique. This is most frequently used technique. It is a unique way of understanding huge datasets. Clustering is employed to find outliers in many fields but it is widely adopted in the field of fraud detection. In this research a method is presented for detection of outliers, which is k-means clustering. A clustering analysis is used in which cluster is run by changing characteristics of clusters such as by taking different values of k in k-means. Clusters are formed and points are detected which are distinct from rest of the data set. The historical data of tax record by oe.cd is used. k-means clustering method is very simple and easy to implement. It can scale to huge data, thus organizing data with gigantic volumes in the form of clusters. Clustering technique work very well when the shape of cluster is spherical. Clustering data can also be further used by other artificial intelligence techniques. Clustering help to extract new and useful information by presenting different recurring patterns, hidden rules and topics. According to this research k-means clustering is expected to be very effective and efficient in identifying tax fraud and its prevention.

This methodology is implemented to data obtained from OECD. Technique of clustering is applied to the 20-year tax data of USA. Clustering involve grouping of same classes of items. In this method clusters are made for different taxable incomes and taxes paid. The algorithm used by k-means clustering has many advantages such as efficiency, brevity and rapidity. The results of k-means algorithm greatly depend on initial clusters and their centroids. With each iteration the centroids change their position. If the initial cluster centroid is not accurate then the algorithm falls into minimization point locally. Also, k-means clustering does not allow the far away data points to be part of cluster even though they have same properties as that of data points inside cluster.

4.2 Recommendations

k-means can also be combined with other techniques to get better results e.g. combination of k-means and linear classifiers in the field of natural language processing (NLP) for semi-supervised applications. With the help of auto encoders and RBM (restricted Boltzmann machines), the efficiency and performance of k-means clustering can be greatly improved.

Bibliography

Alm, J. (2011), "Measuring, explaining and controlling tax evasion: lessons from theory, experiments, and field studies", International Tax and Public Finance, 19 (1), 54- 77.

Zucman, Gabriel. 2013. "The Missing Wealth of Nations: Are Europe and the US net Debtors or net Creditors?" Quarterly Journal of Economics, 128(3), 1321–1364

Davia, H. R., Coggins, P., Wideman, J., & Kastantin, J. (2000). Accountant's guide to fraud detection and control (2nd ed.). Wiley

Bergman, M. (2010). Tax evasion and the rule of law in Latin America: The political culture of cheating and compliance in Argentina and Chile. Penn State University Press.

Wu, R. S., Ou, C. S., Lin, H. Y., Chang, S. I., and Yen, D. C. (2012), "Using data mining technique to enhance tax evasion detection performance", Expert Systems with Applications, 39(10), 8769-877

González González, P. C., and Velásquez, J. D. (2013), "Characterization and detection of taxpayers with false invoices using data mining techniques", Expert Systems with Applications, 40(5), 1427-143

Dias, A., Pinto, C., Batista, J. and Neves, M.E, (2016), "Signaling Tax Evasion, Financial Ratios and Cluster Analysis", Working Paper 51, OBEGEF, Observatorio de Economia y Gestáo de Fraud, Coimbra

Williams, G.J., Christen, P., et al (2007).: Exploratory multilevel hot spot analysis: Australian taxation office case study. In: Proceedings of the sixth Australasian conference on Data mining and Analytics-Volume 70. pp. 77–84. Australian Computer Society, Inc. (2007)

Fabrizio Borselli, (2008) ''Pragmatic Policies to Tackle VAT Fraud in the European Union,'' Int. VAT Monitor (Sept./ Oct. 2008)

Ainsworth, Richard. (2007). UK Car-Flipping: The VAT Fraud Market-Place and Certified Solutions.

C. Jennings, (2010) The EU VAT System – Time for a New Approach, 21 Intl. VAT Monitor 4, p. 257 (2010), Journals IBFD; and Ainsworth, supra n. 19.

Phill Ostwalt, (2016) Global Data & Analytics, Trusted Analytics article series, KPMG International, July 2016

Chena, H., Huang, S., & Kuo, C. (2009). Using the artificial neural network to predict fraud litigation: Some empirical evidence from emerging markets. Expert Systems with Applications, 36, 1478–1484.

Fox News. (2014) Bank of America pays $16.5 bn to settle financial fraud case. Retrieved from Fox News Latino: http://latino.foxnews.com/latino/news/2014/08/21/bank-america-pays165-bn-to-settle-financial-fraud-case/, 2014

Jayakumar GDS, Thomas BJ. (2013) A New Procedure of Clustering based on Multivariate Outlier Detection. Journal of Data Science 2013; 11: 69-84

Ngai E, Hu Y, Wong Y, Chen Y, Sun X. (2011) The Application of Data Mining Techniques in Financial Fraud Detection: A Classification Framework and an Academic Review of Literature. Decision Support Systems 2011; 50: 559–569.

US Government Accountability Office (2004). Datamining: Agencies have taken key steps to protect privacy in selected efforts, but significant compliance issues remain. GAO Press

OECD (2004a). Compliance risk management, managing and improving tax compliance. forum on tax administration compliance subgroup. Centre for Tax Policy and Administration. OECD Press.

OECD (2004b). Compliance risk management, audit case selection systems. forum on tax administration compliance subgroup. Centre for Tax Policy and Administration. OECD Press

Denny, W., & Christen, P. (2007). Exploratory multilevel hot spot analysis: Australian taxation office case study. In Conferences in research and practice in information technology (Vol. 70, pp. 73–80). CRPIT Press

Torgler, B. (2005). Tax morale in Latin America. Public Choice, 122, 133–157

Barnett, V. and Lewis, T. (1994). Outliers in Statistical Data. John Wiley

Fawcett T., and Provost F (1997), "Adaptive fraud detection," Data-mining andKnowledge Discovery, 1(3), 291–316.

Penny K. I. and Jolliffe I. T. (2001), A comparison of multivariate outlier detection methods for clinical laboratory safety data, The Statistician 50(3), 295-308.

Ahmad, S., Lavin, A., Purdy, S., & Agha, Z. (2017). Unsupervised real-time anomaly detection for streaming data. Neurocomputing, 262, 134-147

D.Olszewski, (2014) Fraud detection using self-organizing map visualizing the user profiles, Elsevier, Knowledge-Based Systems, Volume 70, p324- 333 (2014)

D.Sa´nchez, M.A. Vila, L. Cerda, J.M. Serrano, (2009) Association rules applied to credit card fraud detection, Elsevier, Expert Systems with Applications, Volume 36, Issue 2, Part 2, p3630-3640(2009)

E.Kirkos, C.Spathis, Y.Manolopoulos, (2007) Data Mining techniques for the detection of fraudulent financial statements, Elsevier, Expert Systems with Applications Volume 32, Issue 4, p995- 1003(2007)

Ll. Bermudez, J.M. Perez, M. Ayusoc , E. Gomez , F.J. Vazquez, (2007) A Bayesian dichotomous model with asymmetric link for fraud in insurance, Elsevier, p779- 786(2007)

B.Bai, J.Yen, X.Yang, (2008) False financial statements: characteristics of china's listed companies and cart detection approach, International Journal of Information Technology & Decision Making

N.Malini, M.Pushpa, (2017) Analysis on Credit Card Fraud Identification Techniques based on KNN and Outlier Detection, IEEE, Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), Third International Conference (2017)

N. Wong, P.Ray, G.Stephens, L.Lewis, (2012) Artificial immune systems for the detection of credit card fraud: an architecture, prototype and preliminary results, Information Systems Journal, Volume22, Issue1, p53-76(2012)

A.Mubalik (Mubarek) , E.Adali, (2017) Multilayer Perception Neural network technique for fraud detection, IEEE, Computer Science and Engineering (UBMK), International Conference, p383-387 (2017)

IMF (2011) Revenue Mobilization in Developing Countries, Policy Paper, Washington DC: International Monetary Fund

Shin Ando, (2007) Clustering Needles in a Haystack: An Information Theoretic Analysis of Minority and Outlier Detection, Proceedings of the 2007 Seventh IEEE International Conference on Data Mining, p.13-22, October 28-31, 2007

Besley, T. and Persson, T. (2013) 'Taxation and Development', in A.J. Auerbach, R. Chetty, M. Feldstein and E. Saez (eds), Handbook of Public Economics, Vol. 5, Amsterdam: Elsevier B.V.: 51-110

Grubbs, F. E. (1969), 'Procedures for detecting outlying observations in samples'. Technometrics 11, 1–21.

Wikipedia contributors. Standard deviation — Wikipedia, the free encyclopedia, 2018. [Online; accessed 7-November-2018]. ix, 16

Yeh I and Lien C-h (2009) The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications 36, 2473-80.

Ravisankar P, Ravi V, Raghava Rao G, and Bose I (2011) Detection of financial statement fraud and feature selection using data mining techniques. Decision Support Systems 50, 491-500.

BianZhaoQi and ZhangXuegong (2000) Pattern Recognition Beijing Tsinghua University Press 2000.

JAIN A K, DUBES R C.(1998) Algorithms for clustering data[M].New Jersey:Prentice-Hall,1988.

G. Hamerly. (2003) Learning Structure and Concepts in Data using Data Clustering, PhD Thesis. University of California, San Diego, 2003.

Abdallah, A.; Maarof, M. A.; Zainal, A. (2016): Fraud detection system: a survey. Journal of Network and Computer Applications, vol. 68, no. 6, pp. 90-113.

Ghosh, S; Reilly, D. L. (1994): Credit card fraud detection with a neural-network. In IEEE System Sciences Proceedings of the Twenty-Seventh Hawaii International Conference, pp. 621-630.

Chen, J.; Tao, Y.; Wang, H.; Chen, T. (2015): Big data based fraud risk management at Alibaba. The Journal of Finance and Data Science, vol. 1, no. 1, pp. 1-10

Tatiana Tropina (2016), Do Digital Technologies Facilitate Illicit Financial Flows, World Bank, http://documents.worldbank.org/ curated/en/896341468190180202/pdf/102953-WP-Box394845B-PUBLIC-WDR16-BP-Do-Digital-Technologies-Facilitate-IllicitFinancial-Flows-Tropina.pdf.

Deloitte, Insight on Financial Crime: Challenges Facing Financial Institutions, (2014), http://www2.deloitte.com/content/dam/Deloitte/global/Documents/ Risk/gx-cm-insight_on_financial_crime.pdf; Trend Micro, Addressing Big Data Security Challenges,5;

Gutierrez, Anzelde, and Gobenceaux, Risk and Reward, 10; IBM, Combat Credit Card FraudwithBigData,2013,http://www.intel.de/content/dam/www/public/us/en/documents/white-papers/combat-credit-card-fraud-with-big-data-whitepaper.pdf, pg. 2.

EC – Taxation and Customs Union. 2017. "The fight against tax fraud and tax evasion - Time to get the missing part back". Accessed January 9, 2017. https://ec.europa.eu/taxation_customs/fightagainst-tax-fraud-tax-evasion/missing-part_en