

INDIAN INSTITUTE OF TECHNOLOGY, ROORKEE

Department of Computer Science

(2020-2022)

**CSN -505 Project Lab**

**“Privacy Preserving in Data Mining”**

**Submitted To:** **Submitted By:**

Dr. Durga Toshniwal Aayushi Jain (20535001)

**ACKNOWLEDGEMENT**

I am highly indebted to Dr. Durga Toshniwal (Professor) for their guidance and constant supervision as well as for providing necessary information regarding the project & also for her support in completing the project. I would like to express our gratitude towards members of INDIAN INSTITUTE OF TECHNOLOGY, ROORKEE for their kind cooperation and encouragement which helped in completion of this project.

**CONTENTS**

[**1.**](#_heading=h.gjdgxs) **Introduction 5**

[1.1 k-Anonymity Model 5](#_heading=h.30j0zll)

[1.2](#_heading=h.30j0zll) l-Diversity Model 5

[**2. Dataset**](#_heading=h.1fob9te) 6

[**3.Data Preprocessing**](#_heading=h.3znysh7) 6

[3.1 Missing Values Handling](#_heading=h.2et92p0) 6

[3.2 Categorical Variables Handling](#_heading=h.tyjcwt) 6

[**4. k-Anonymity Model**](#_heading=h.2xcytpi) 7

4.1 Mondrian’s Algorithm 7

4.2 Analysis after applying Mondrian’s Algorithm 15

4.3 Possible Attacks on K-Anonymity Model 15

**5**[**. l-Diversity Model**](#_heading=h.17dp8vu) 16

5.1 l-Diversity Principle 16

5.2 Implementation of l-diversity Model 17

**6**[**.**](#_heading=h.1ci93xb) **Attack on l-Diversity Model** 22

[**7. Significance of above mentioned models**](#_heading=h.1pxezwc) 22

**8. Code Snippet** 23

**9. Applications of PPDM** 29

**10. Future Work** 30

**11**[**. References**](#_heading=h.35nkun2) 30

**Table of Figures**

***Figure 1: Dataset used in the project…………………..…………………………………………….…… 6***

***Figure 2:Dataset after Preprocessing……..………………….………….………….…..…….………….. 7***

***Figure 3:*  *Span Function*  *………...……………….………….………….………….……..…………….8***

***Figure 4: Splitting Function………….………….………….…………...……….………………………..8***

***Figure 5: Partitioning Algorithm ……………….…..……………………...……..………………………9***

***Figure 6: Partitions for k=5…………..…………………….………...…………..……………………….9***

***Figure 7: Partitions’ Coordinates for k=5………..…………….……..…………………………………10***

***Figure 8: Plot for the partitions having k=5 for age versus bmi…….………..…………………………10***

***Figure 9: Partitions for k=20 ………………….…………………….………….……..…………………11***

***Figure 10: Partitions’ Coordinates for k=5 ………………….…………….…..……………………..…11***

***Figure 11: Plot for the partitions having k=20 for age versus bmi.…………………...……………..…12***

***Figure 12: Partitions for k=45 ………………….………………………………….………….……...…12***

***Figure 13: Partitions’ Coordinates for k=45 ………………….………….…………………………..…13***

***Figure 14: Plot for the partitions having k = 45 for age versus bmi…………….....…………………...13***

***Figure 15: Final anonymized dataset for k=5 ………………….………………..…………………..…14***

***Figure 16: Final anonymized dataset for k=20 ………………….……………...…………………..…14***

***Figure 17: Final anonymized dataset for k=45 ………………….……………....…………………..…15***

***Figure 18: Partitions for k=5 and l=5 ………………….……………………………………. ……..…17***

***Figure 19: Plot for the partitions having k = 5 and l =5 for age versus bmi…..…………….……..…18***

***Figure 20: Final Anonymized Dataset for k=5 and l=5 ………………….………………….……..…18***

***Figure 21:Partitions for k=20 and l=5 ………………….………………………………….………..…19***

***Figure 22: Plot for the partitions having k = 20 and l =5 for age versus bmi……………....……..…19***

***Figure 23: Final Anonymized Dataset for k=20 and l=5………………….……...…………..……..…20***

***Figure 24: Partitions for k=45 and l=5………………….……………………………..………..…..…20***

***Figure 25: Plot for the partitions having k = 45 and l =5 for age versus bmi……………………..…21***

***Figure 26: Final Anonymized Dataset for k=45 and l=5 …….………………………………...…..…21***

***Figure 27: Semantic Similarity in dataset ……………………….……………………..…………..…22***

***Figure 28: Code Snippet of k-Anonymity …………….………………………….………..………..…23***

***Figure 29: Code Snippet of l-diversity ………………….………………….……..………………..…26***

**1. Introduction**

The basic notion of information privacy is to have control over handling and collecting an

individual’s personal data. Collection of data from various sources may have many

advantages, but it may also lead to information leakage. To deal with information leakage,

methods have been proposed, which are known as **Privacy Preserving Data Mining**

**(PPDM)** Techniques. PPDM techniques work by modifying user’s original data. PPDM

techniques are designed in such a way so as to hide the user’s data, while maintain the

data utility. Protecting sensitive information as well as preserving data utility are both

extremely important in today’s world, especially with all the data being collected every

single day of our lives. I have implemented anonymization methods through properly

generalizing or suppressing the quasi-identifiers in the dataset in order to prevent linkage

attacks and violations of privacy and security laws.

Models implemented for anonymizing the dataset are mentioned below:

**1.1 k-Anonymity Model:**

k-anonymity model is a model which comes under data publishing privacy. If the identifier attributes of a record cannot be discriminated from k-1 records at the least, the dataset is said to be k-anonymous, i.e., any record in a dataset is similar to at least k other records. And an equivalence class is such a set of k records. I have implemented Mondrian Algorithm. The algorithm utilizes a greedy search algorithm that allows for more desirable anonymizations than traditional exhaustive optimal algorithms. Time Complexity of Mondrian’s Algorithm is O(nlogn).

**1.2 l-Diversity Model :**

Because of the limitations of the k-anonymity model, the l-diversity model was proposed. The l-diversity model is an expansion of the k-anonymity model, in the sense that it follows the l-diversity principle in each equivalence class. The l-diversity principle states that “in each equivalence class, at least l ‘well represented values’ exist for the sensitive attributes.”

# 2. Dataset

The dataset that I have chosen for this project is the Patient Disease dataset from Kaggle.

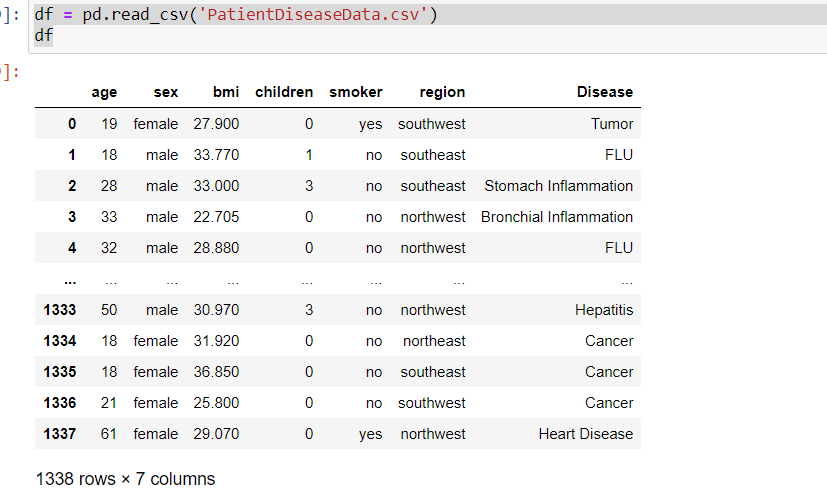
This dataset contains 1338 rows of unique individuals. The columns that are included in this

dataset are age, sex, bmi, children, smoker, region, Disease. Some columns show us the

quasi-identifiers of an individual, which include age, sex, bmi, children and region. The

sensitive data would be the smoker and disease that are associated with an individual because

it can be used against an individual if their identity is released and not properly anonymized.



**Figure 1: Dataset used in the project**

# 3. Data Preprocessing

**3.1 Missing Values Handling**

Missing values were deleted, because there were very few missing values. There were around 5 to 6 missing values which were only for the column named “Disease”.

### 3.2 Categorical variables Handling

To make sure available data would be usable for machine learning models, I decided to

map categorical variables to numerical values. Since the gender attribute in the dataset only

contained male & female we mapped 0 to male and 1 to female. Furthermore the data set also

had a binary column which categorized if a patient was a smoker or not. In our dataset, 0

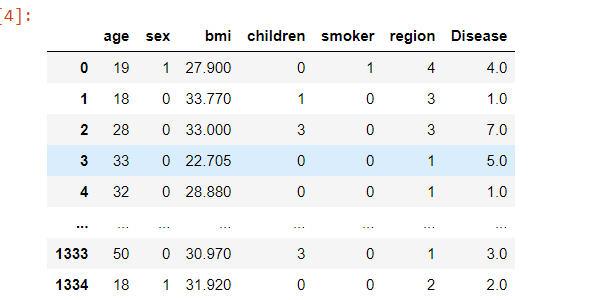
corresponded to not a smoker, and 1 corresponded to being smoker. Apart from that, disease

columns’ values are also mapped to numerical ones in the same way. Some of the columns

were incorrectly categorized as the wrong data type so we corrected this and ensured

that all categorical columns had a dtype of category. All other data was categorized as float64

or int64. After all of these steps our dataset was ready to be worked on.



**Figure 2: Dataset after Preprocessing**

# 4. k-Anonymity Model

# k-anonymity model is a model which comes under data publishing privacy. With respect to preserving sensitive attributes about person, k-anonymity is used as the key concept of the model. If the identifier attributes of a record cannot be discriminated from k-1 records at the least, the dataset is said to be k-anonymous, i.e., any record in a dataset is similar to at least k other records. And an equivalence class is such a set of k records. Using k-anonymity, it becomes difficult for a person to identify a person’s sensitive attribute because any record is similar to k-1 other records.

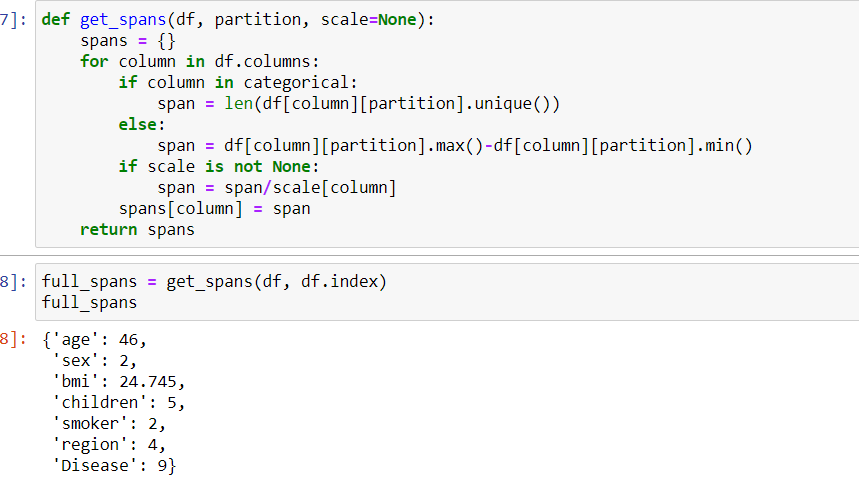
# 4.1 Mondrian Algorithm : I have used Mondrian Algorithm in order to implement k-anonymization on the dataset in Python. The algorithm utilizes a greedy search algorithm that allows for more desirable anonymizations than traditional exhaustive optimal algorithms Furthermore, it allows for multidimensional models, which is what’s best for our specific dataset.

1. The **span function** within the Mondrian Algorithm calculated the max-min for

numerical columns and the number of different values for categorical variables. This

function is calculated for all of the columns to help in figuring out the partitions of

the data.

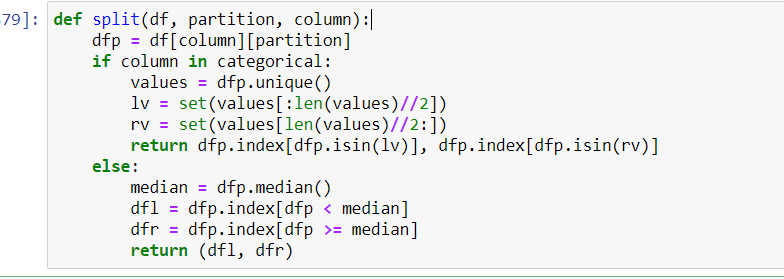


**Figure 3: Span Function**

1. Following the spanning function, I implemented the **split function** that takes in the

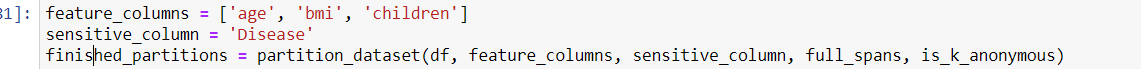
data frame’s given partition and returns two partitions which split the given partition

with the values above and below the median into two separate columns.



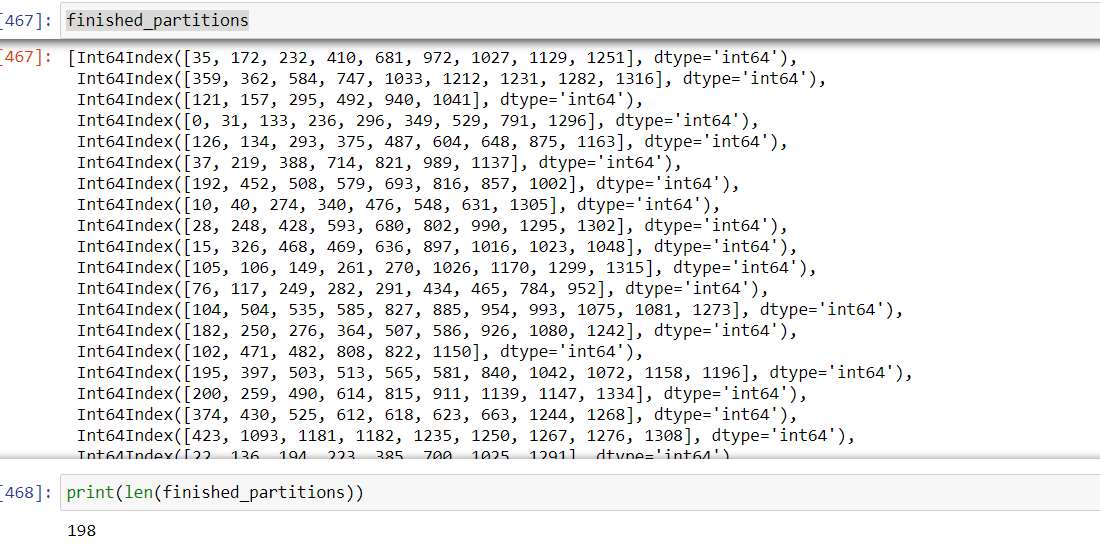
**Figure 4: Splitting Function**

1. The next step is to anonymize the partition the dataset using partitioning method which is being performed for 3 different values of k viz. We set the featured values to [‘age’, ‘bmi’, ‘children’] and the sensitive values to be [‘Disease’].



**Figure 5: Partitioning Algorithm**

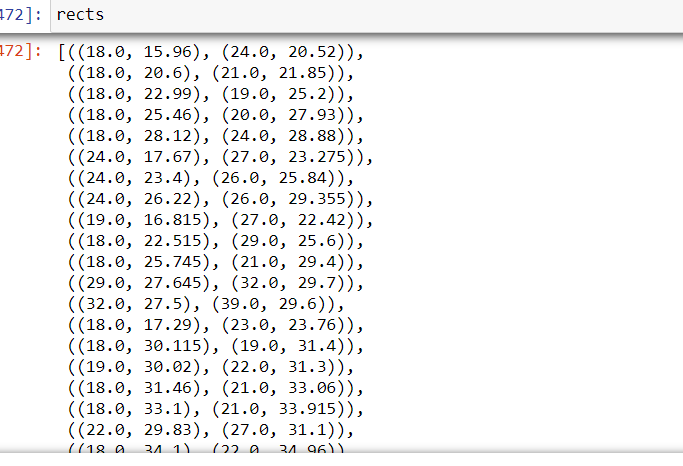
* 1. There are two type of partitioning available with us viz. Single Dimensional Partitioning and Multidimensional partitioning. I have applied multidimensional partitioning using feature column [‘age’, ‘bmi’, ‘children’].
  2. Final Partitions that I got after applying partitioning algorithm on different values of k are shown below. Also, partitions plotted between age and bmi as x coordinate and y coordinate respectively.
     1. **For k = 5:**
        1. List of list of partitions which contains indexing of rows is shown below:



**Figure 6: Partitions for k=5**

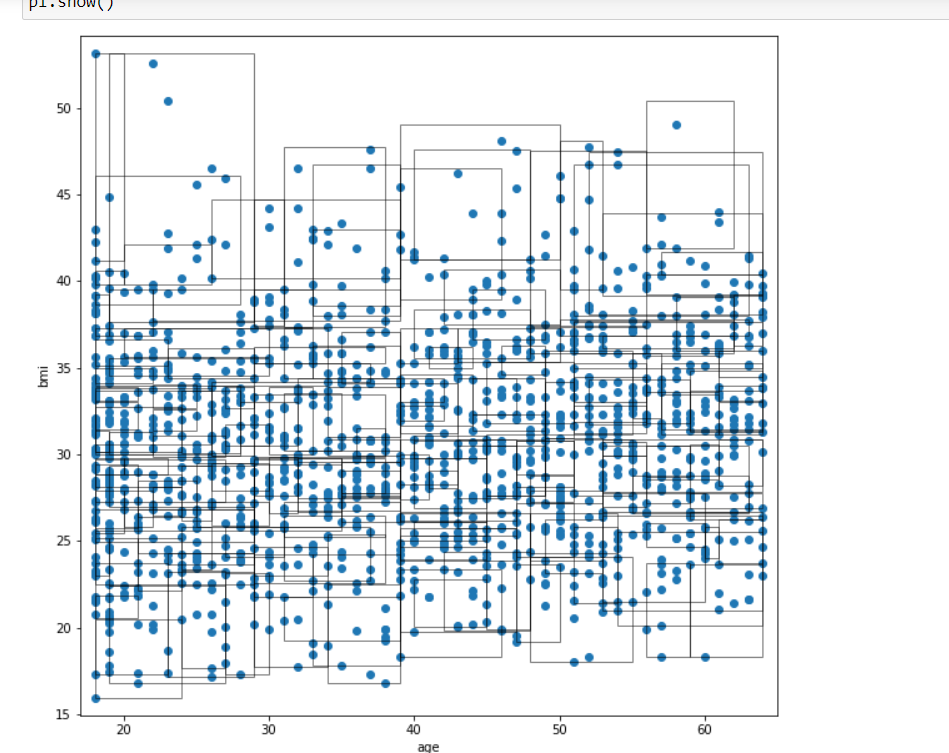
* + - 1. Partitions’ coordinates are shown below. These are in the form of

{(xl,yl),(xr,yr)} where x coordinates are for age and y are for bmi.



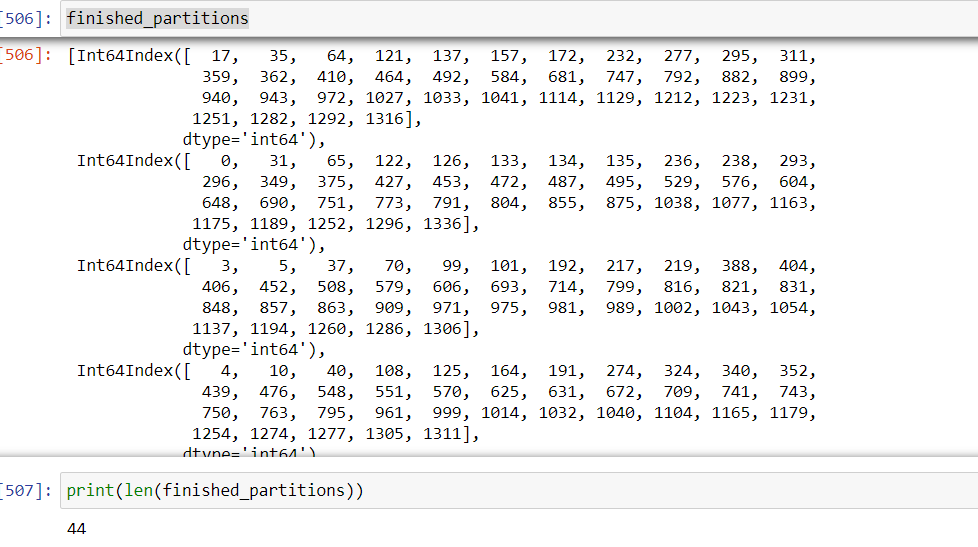
**Figure 7: Partitions’ Coordinates for k=5**

**3.2.1.3** Graph shown below contains the partitions having values for age versus bmi:



**Figure 8: Plot for the partitions k=5 for age versus bmi**

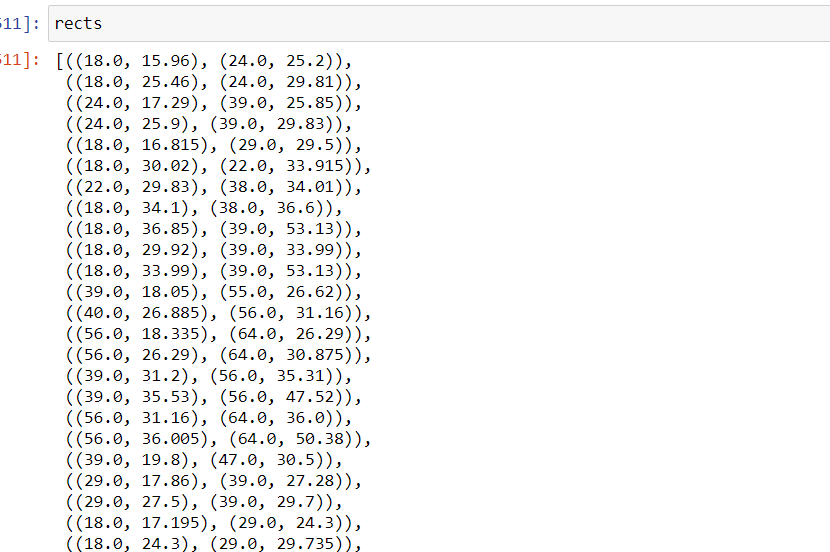
* + 1. **For k = 20:**
       1. List of list of partitions which contains indexing of rows is shown below:



**Figure 9:** **Partitions for k=20**

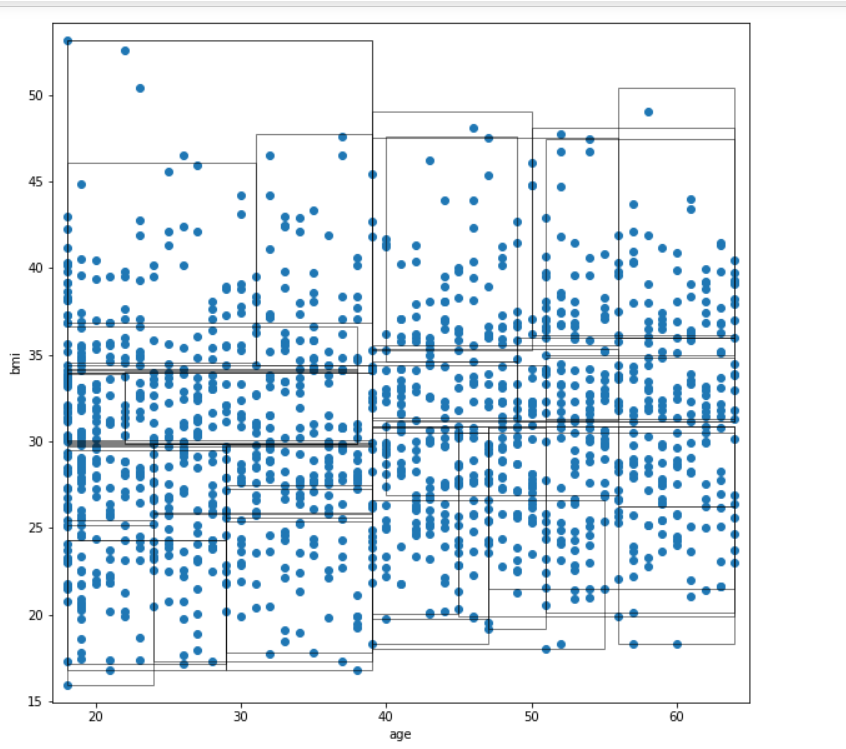
**3.2.2.2** Partitions’ coordinates are shown below. These are in the form of

{(xl,yl),(xr,yr)} where x coordinates are for age and y are for bmi.



**Figure 10:** **Partitions’ Coordinates for k=5**

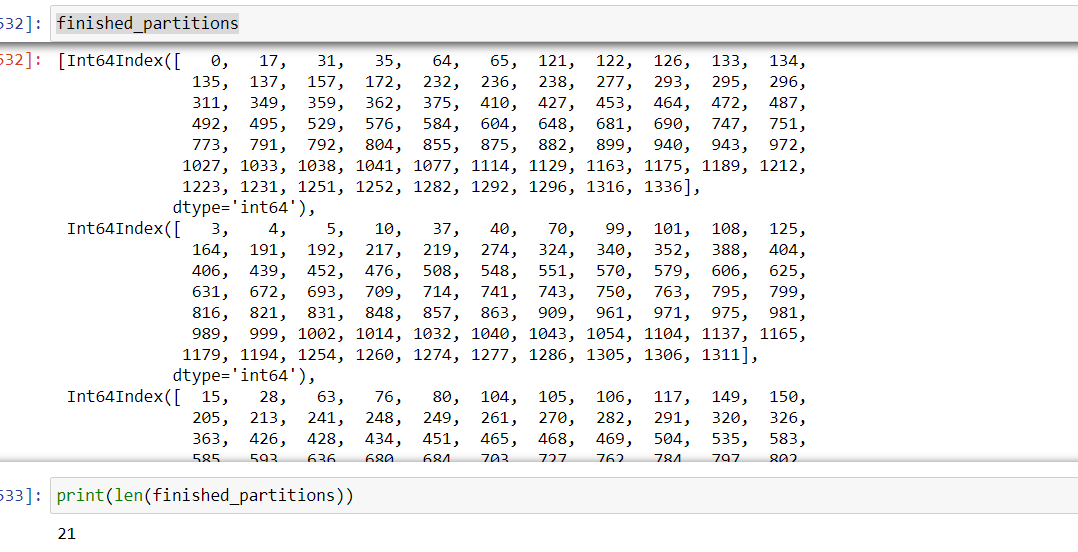
**3.2.2.3** Graph shown below contains the partition having values for age versus bmi:

****

**Figure 11: Plot for the partitions having k=20 for age versus bmi**

**3.2.3 For k = 45:**

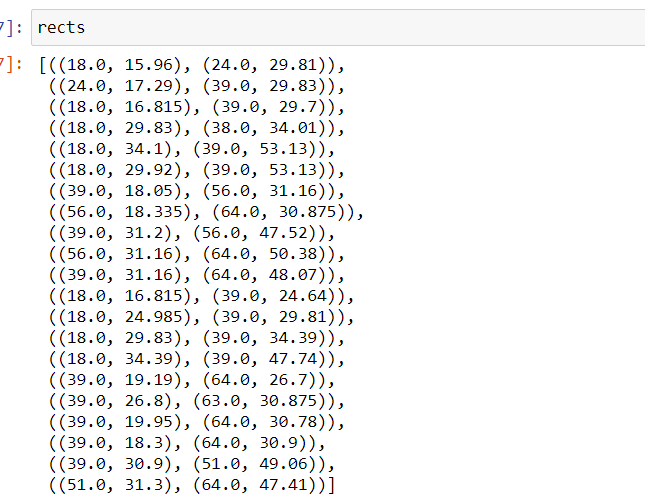
**3.2.3.1** List of list of partitions which contains indexing of rows is shown below:

****

**Figure 12: Partitions for k=45**

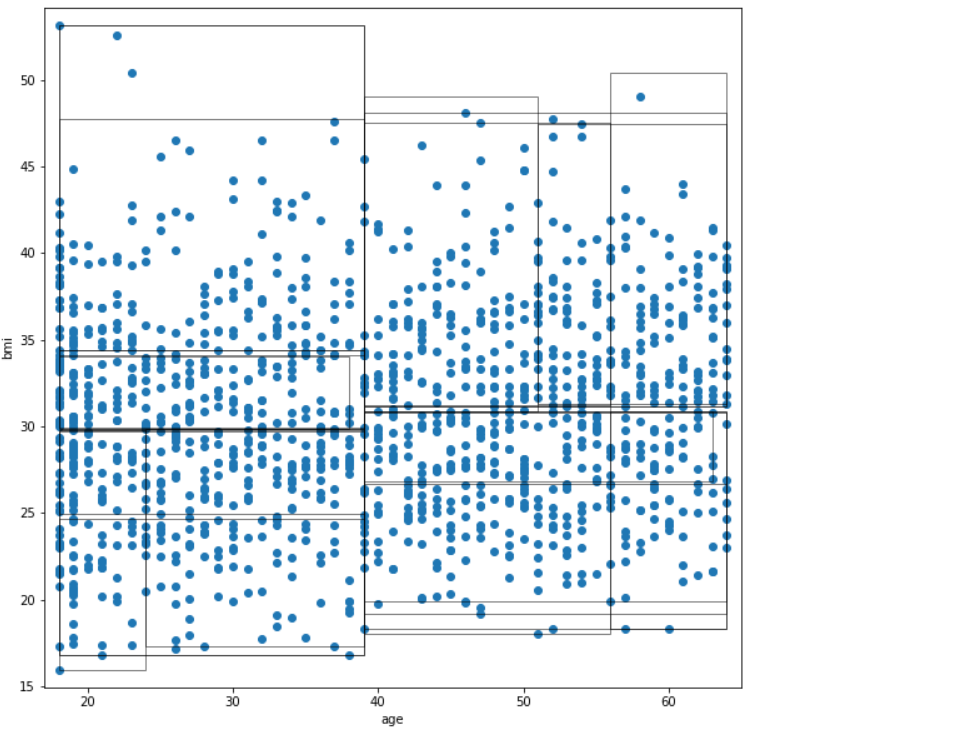
**3.2.3.2** Partitions’ coordinates are shown below. These are in the form of

{(xl,yl),(xr,yr)} where x coordinates are for age and y are for bmi.



**Figure 13**: **Partitions’ Coordinates for k=45**

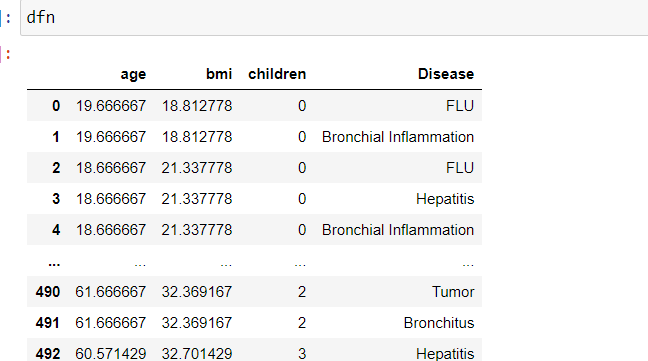
**3.2.3.3** Graph shown below contains the partition having values for age versus bmi:



**Figure 14: Plot for the partitions having k = 45 for age versus bmi**

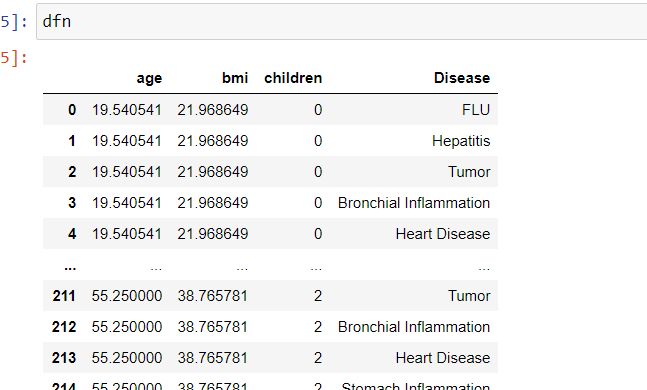
1. Now that we have generated an anonymized dataset, our next step is to aggregate each of the columns and be able to generate the final anonymized dataset. As we can see in the image below, the data before we used these anonymization techniques has a lot more columns and information about each of the individuals.

**4.1** For k = 5

****

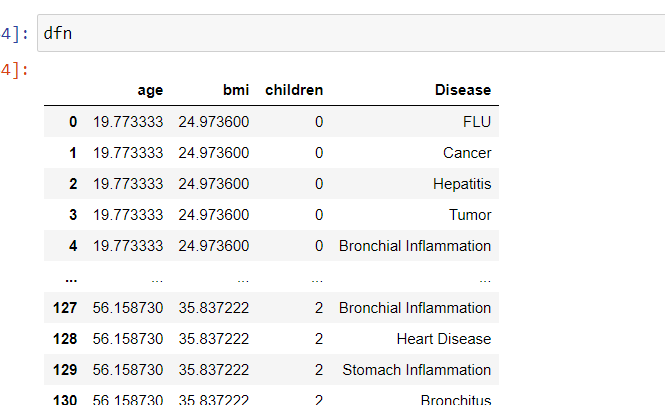
**Figure 15: Final anonymized dataset for k=5**

**4.2** For k = 20



**Figure 16:** **Final anonymized dataset for k=20**

**4.3** For k = 45

****

**Figure 17: Final anonymized dataset for k=45**

* 1. **Analysis after applying Mondrian’s Algorithm:**

1. TheMondrianalgorithm generalized our dataset over the calculated partitions with k-anonymity of k = 5, k = 20 , k = 45 . As we are increasing the value of k, data privacy is increasing but data utility is reducing.
2. Our dataset lost information variables [‘sex’, ‘smoker’, ‘region’] during the anonymization.
3. Some of the columns in the dataset such as age and bmi were generalized to be the mean value of their partition. This helped with making the entries indistinguishable and not as easily recognizable to an adversary.
4. Additionally, the sensitive data value, charges, was given a different value in comparison to the original data, in that the numbers were swapped and unordered so that an adversary could not fully comprehend the true values and link it to the individual.
5. This process, while increasing the privacy, does diminish some of the usability and utility of the dataset.
   1. **Possible attacks on k-Anonymity Model :**
6. **Homogeneity Attack:** This attack leverages the case where all the values for a

sensitive value within a set of *k* records are identical. In such cases, even though the

data has been *k*-anonymized, the sensitive value for the set of *k* records may be

exactly predicted.

1. **Background Knowledge Attack:** This attack leverages an association between one

or more quasi-identifier attributes with the sensitive attribute to reduce the set of

possible values for the sensitive attribute. Common known facts or background

knowledge can de-anonymize the identity of a person.

The above mentioned problems can be dealt by making the sensitive values within an

equivalence class more diverse, and this approach is the building block for the

l-diversity model.

# l-Diversity Model

Because of the limitations of the k-anonymity model, the l-diversity model was proposed. The l-diversity model is an expansion of the k-anonymity model, in the sense that it follows the l-diversity principle in each equivalence class.

Two types of information disclosure have been identified :

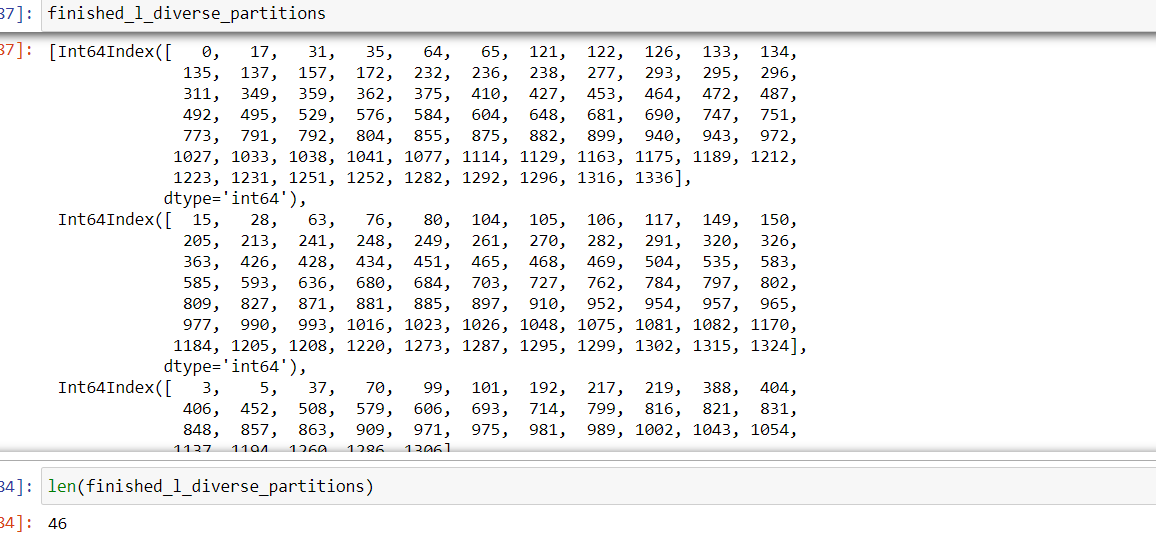
1. **Identity Disclosure:** Identity disclosure occurs when an individual is linked to a particular record in the table. Identity disclosure often leads to attribute disclosure. Once there is identity disclosure, an individual is re-identified and the corresponding sensitive values are revealed.
2. **Attribute Disclosure:** Attribute disclosure occurs when new information about some individuals is revealed using an attribute. Attribute disclosure can occur with or without identity disclosure. It has been recognized that even disclosure of false attribute information may cause harm.

While k-anonymity protects against identity disclosure, it is insufficient to prevent attribute disclosure. To address this limitation of k-anonymity, a new notion of privacy is introduced, called **l-diversity**.

* 1. **l-Diversity Principle:** The l-diversity principle states that “in each equivalence class, at least l ‘well represented values’ exist for the sensitive attributes.” A dataset is said to be l-diverse, if all the equivalence classes follow the property of l-diversity.

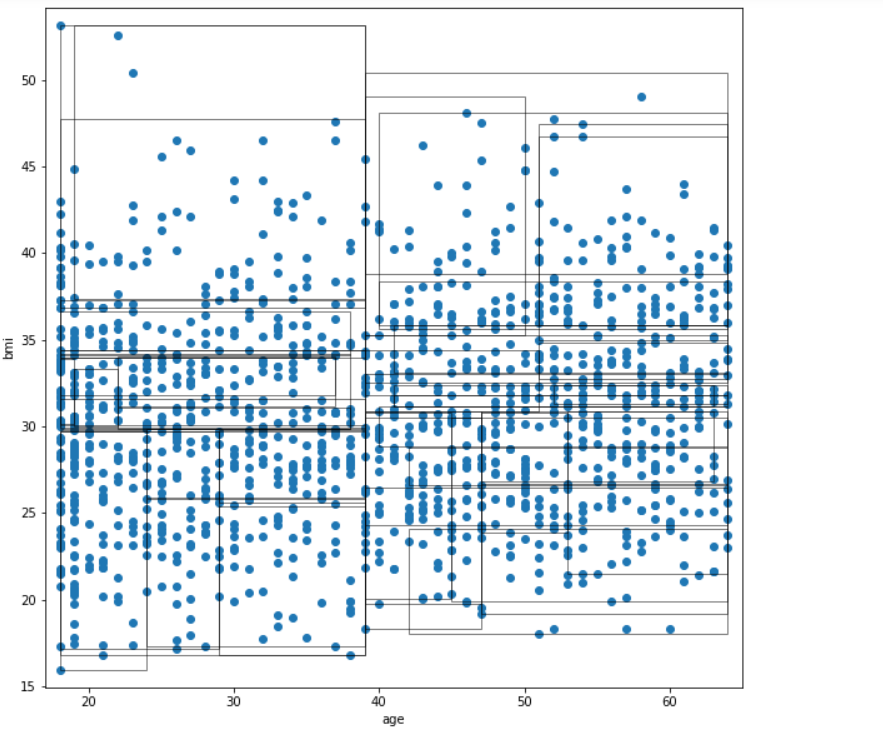
In general, ‘well-represented’ means there are l-distinct values for a sensitive attribute in an equivalence class.

* 1. **Implementation of l-Diversity Model:** 
     1. Following the spanning and splitting method, partition\_dataset method which is used for partitioning also keep a check on the sensitive attribute. There should be l-distinct values for a sensitive attribute. I have chosen l=5. This value is chosen randomly.
        1. **For k = 5**
           1. List of list of partitions which contains indexing of rows is shown below:



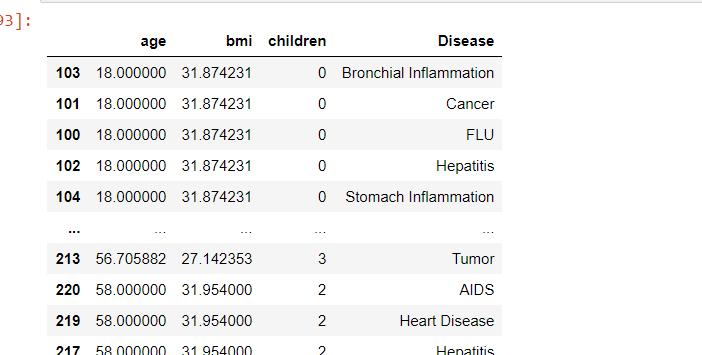
**Figure 18:** **Partitions for k=5 and l=5**

**5.2.1.1.2** Graph shown below contains the partition having values for age versus bmi:



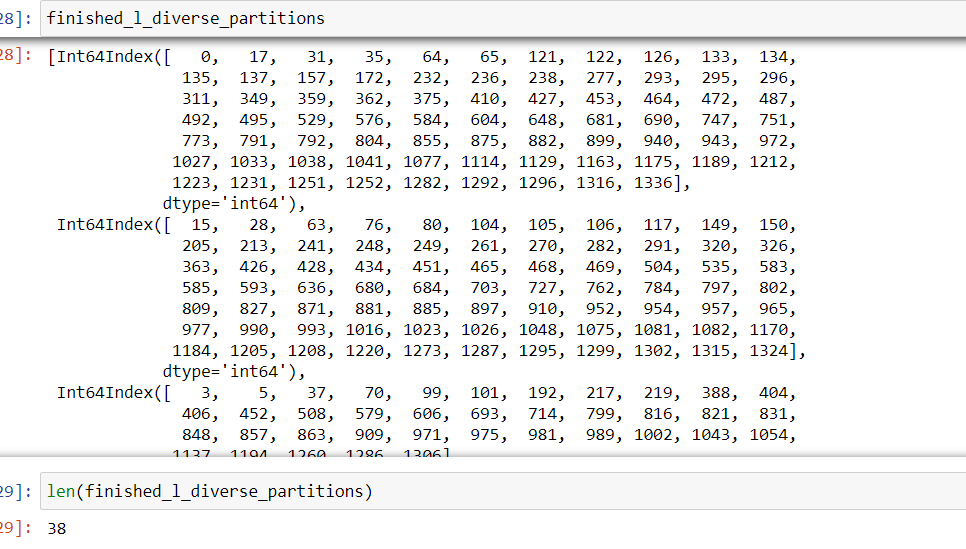
**Figure 19:** **Plot for the partitions having k = 5 and l =5 for age versus bmi**

* + - * 1. Final Anonymized dataset is shown below :



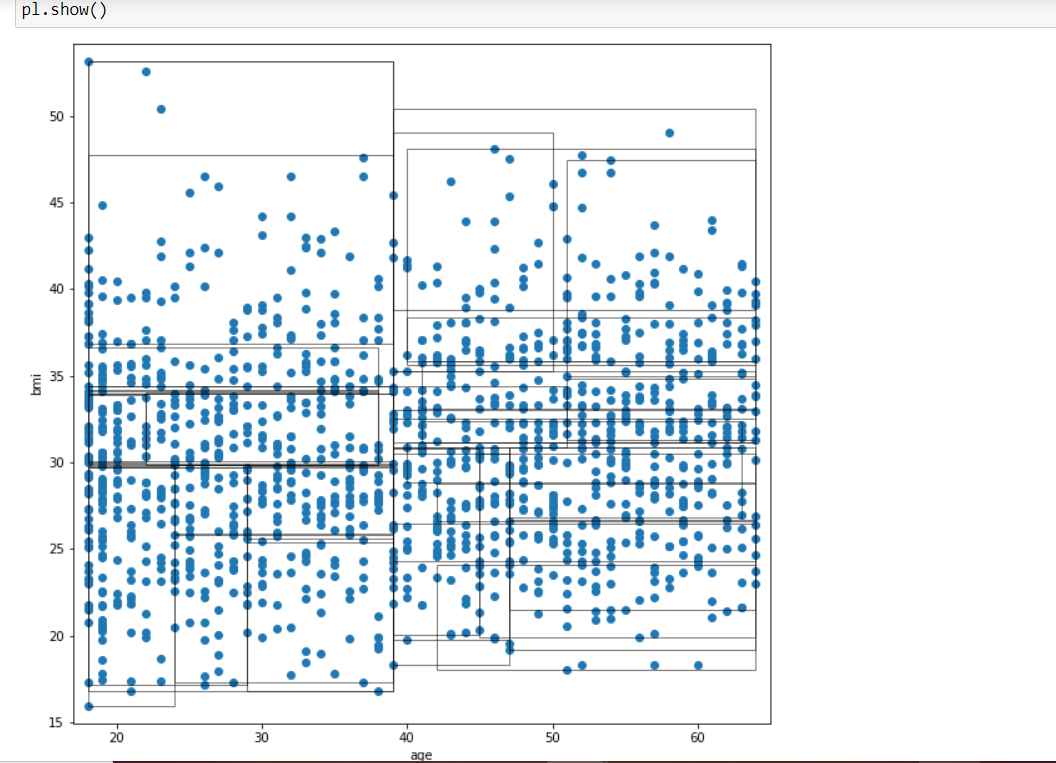
**Figure 20: Final Anonymized Dataset for k=5 and l=5**

* + - 1. **For k = 20:**
         1. List of list of partitions which contains indexing of rows is shown below:



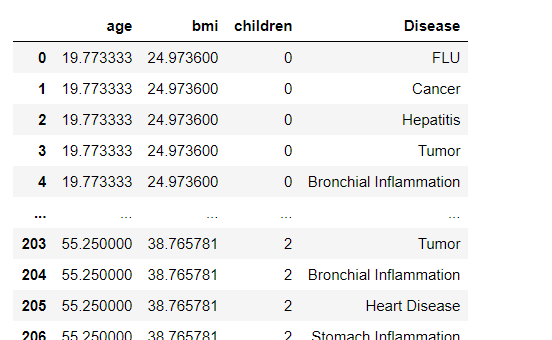
**Figure 21:** **Partitions for k=20 and l=5**

**5.2.1.2.2** Graph shown below contains the partition having values for age versus bmi:

****

**Figure 22: Plot for the partitions having k = 20 and l =5 for age versus bmi**

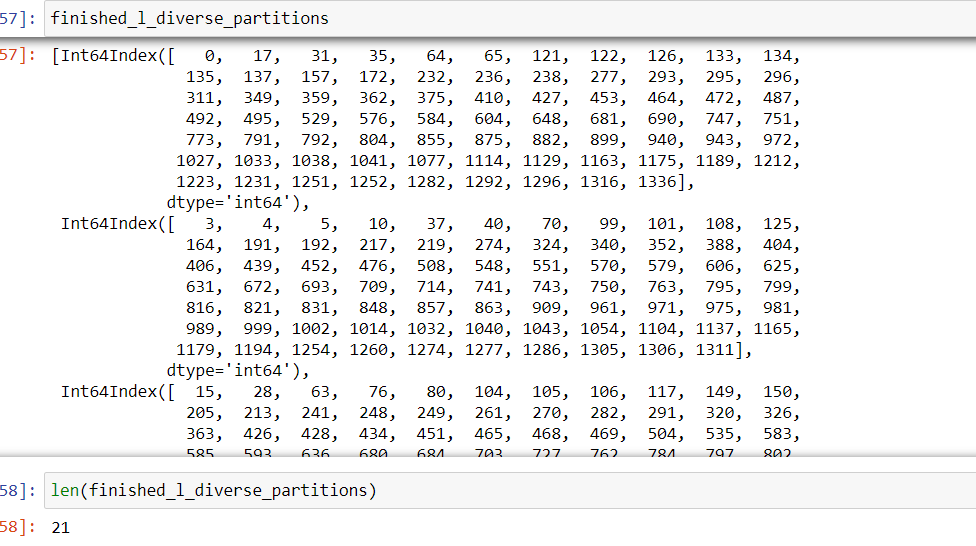
* + - * 1. Final Anonymized dataset is shown below :



**Figure 23:** **Final Anonymized Dataset for k=20 and l=5**

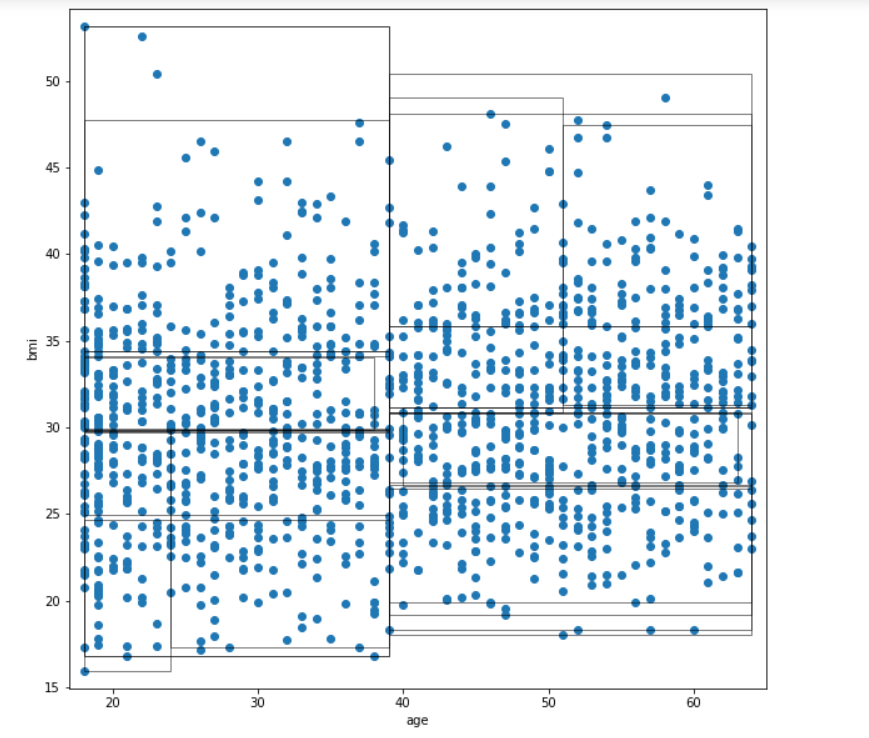
* + - 1. **For k = 45**

**5.2.1.3.1** List of list of partitions which contains indexing of rows is shown below:

****

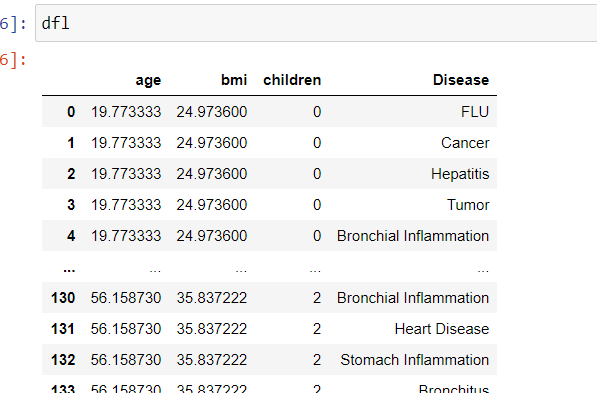
**Figure 24: Partitions for k=45 and l=5**

**5.2.1.3.2** Graph shown below contains the partition having values for age versus bmi:

****

**Figure 25: Plot for the partitions having k = 45 and l =5 for age versus bmi**

* + - * 1. Final Anonymized dataset is shown below :



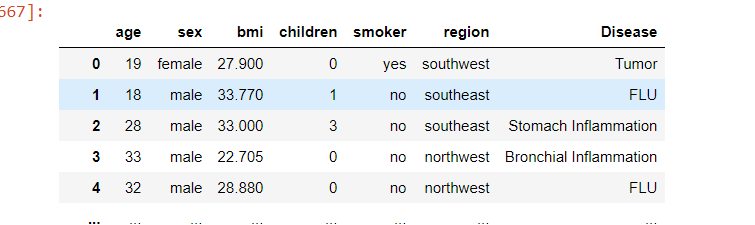
**Figure 26:** **Final Anonymized Dataset for k=45 and l=5**

# 6. Attack on l-Diversity Model

**Similarity attack :** When the sensitive attribute values in an equivalence class are distinct

but semantically similar, an adversary can learn important information.

There may be occurrence leakage of sensitive information because while l-diversity requirement ensures “diversity” of sensitive values in each group, it does not take into account the semantical closeness of these values. This attack can be overcome using t-closeness model. For Eg; in our dataset, we have stomach inflammation and bronchial inflammation. These are semantically similar but L- diversity model considers these values to be different while creating l-distinct values in an equivalence class.



**Figure 27: Semantic Similarity in dataset**

# 7. Significance of above mentioned models :

1. Since this dataset pertains to patient disease data records of individuals, If this data

was to be leaked, many individual’s personal information is at risk. This can lead to

reidentification of the person by a linkage attack. This is significant because the dataset

includes sensitive data such as the individual’s number of children, if they are a

smoker, and the disease with which they are suffering from. Such information is

beneficial to data snoopers and hackers wanting to learn more about an individual.

1. Another reason this problem is significant is because if the data is leaked, legal and

governmental issues such as HIPAA violations may occur because the data was not

properly anonymized. There have been countless medical data leaks in the past and

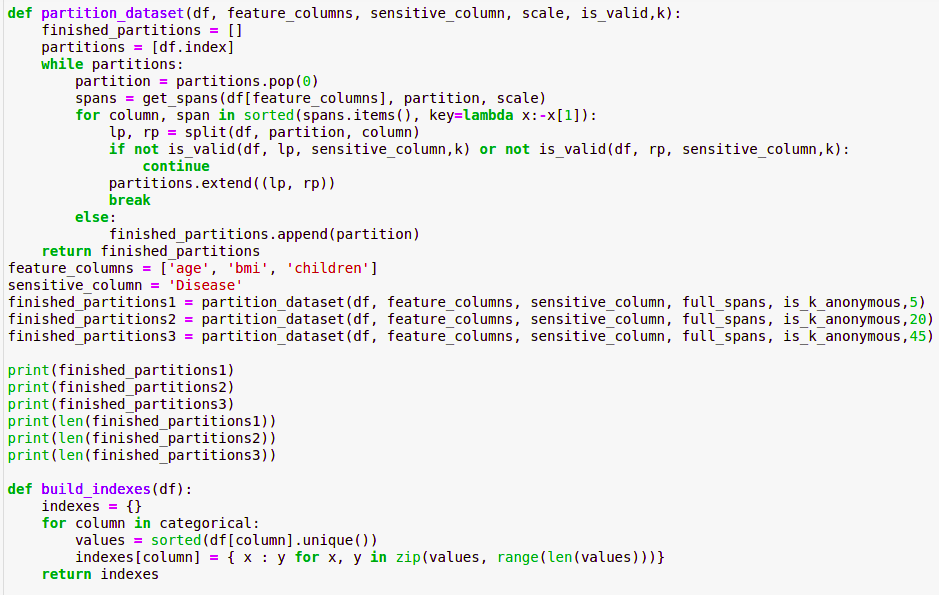
they have caused numerous individuals’ private information to be violated and used by

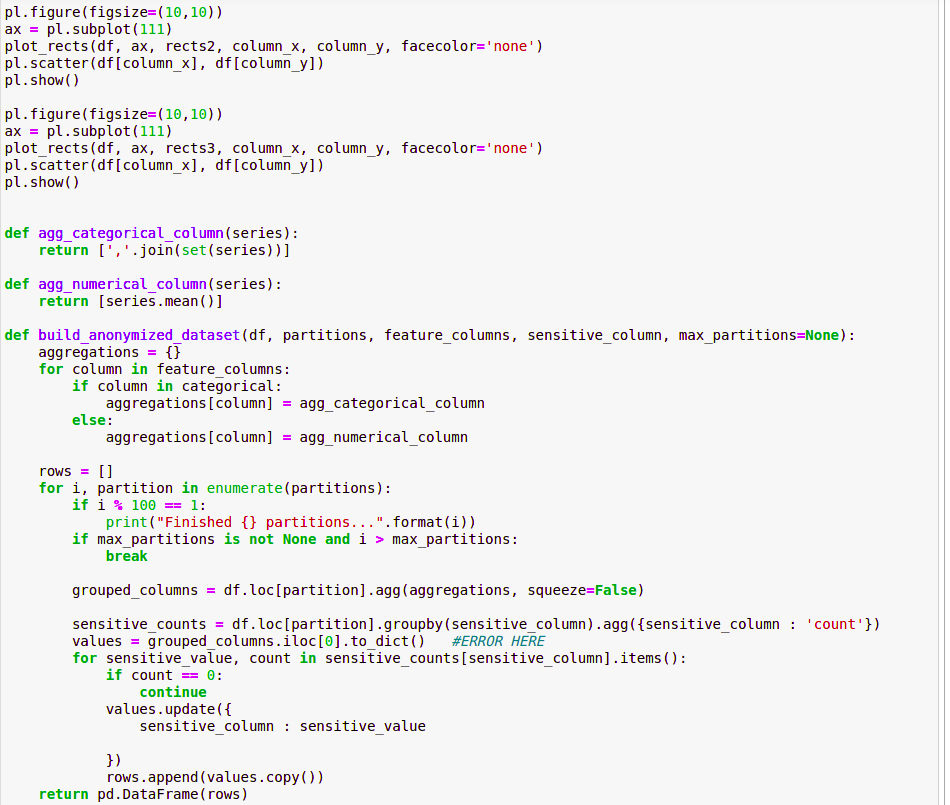
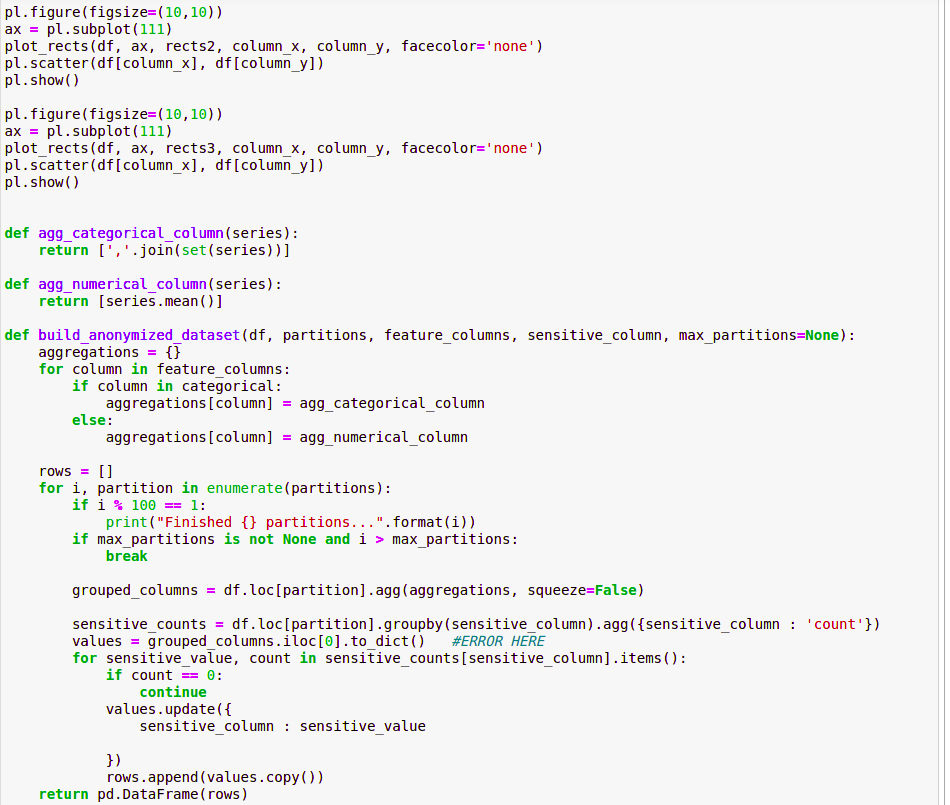
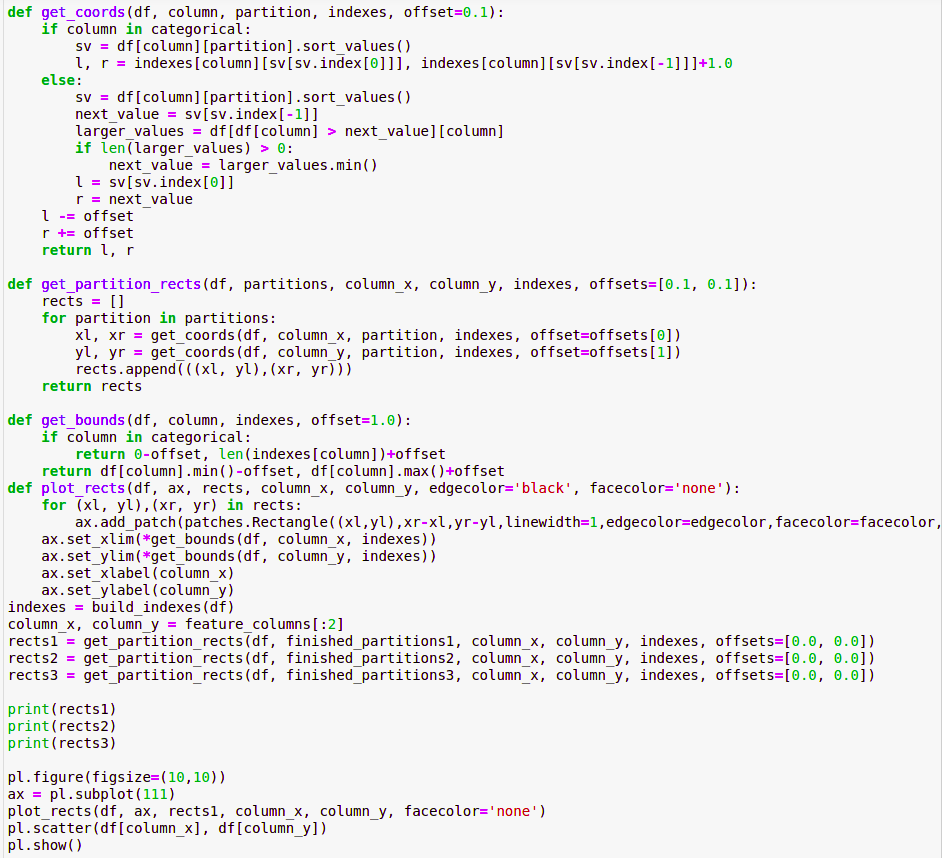
an adversary. This causes several issues and loss of customers for many businesses.

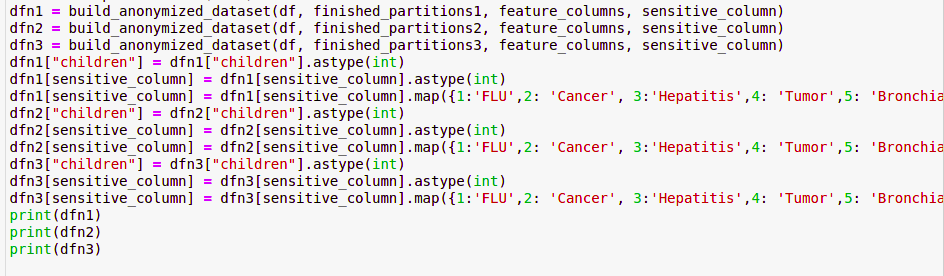
**8.Code Snippet:**

**8.1 Figure 28: Code Snippet for k-Anonymity Model**



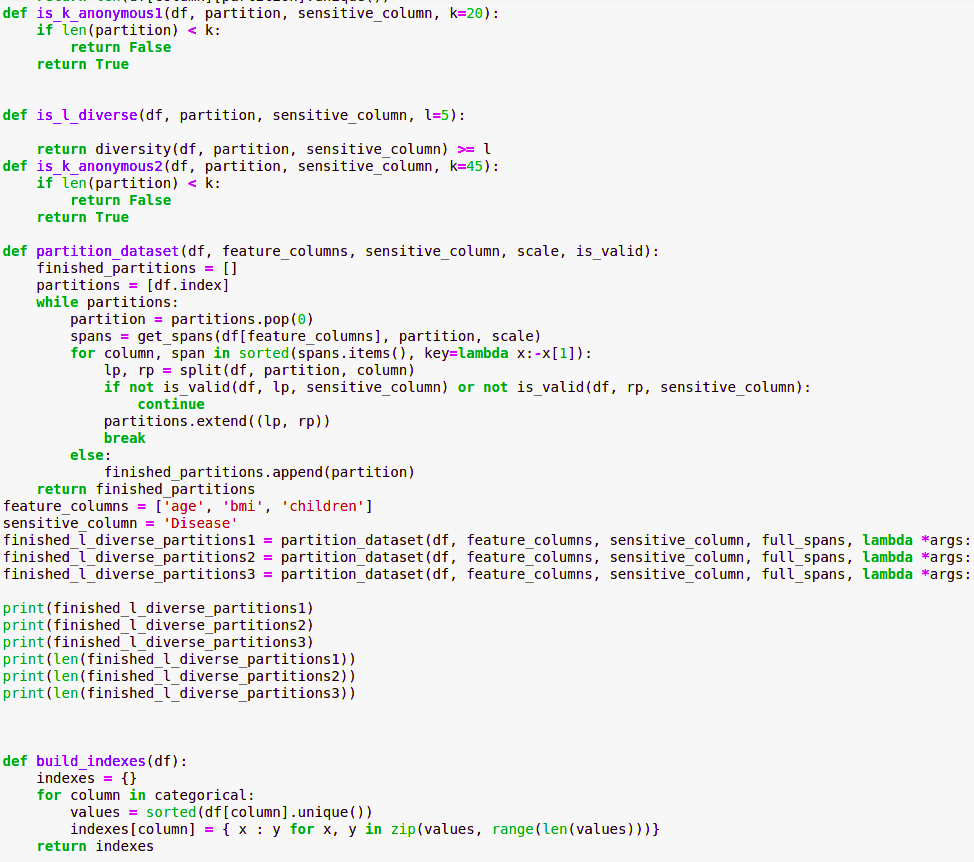


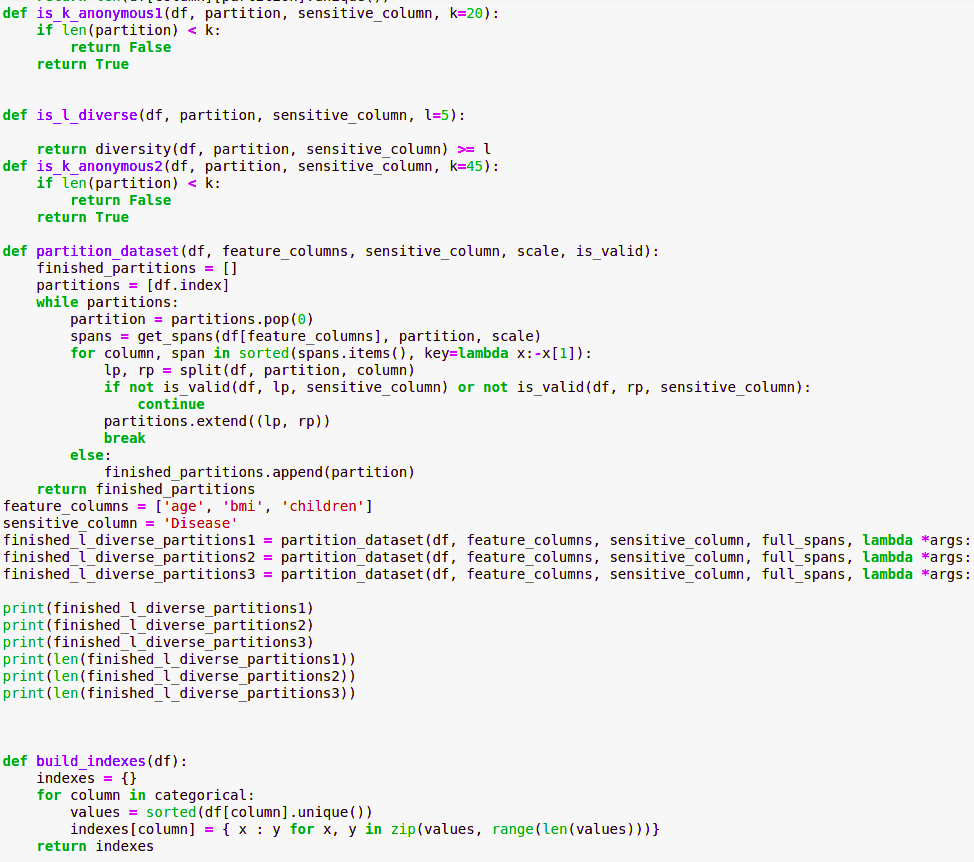


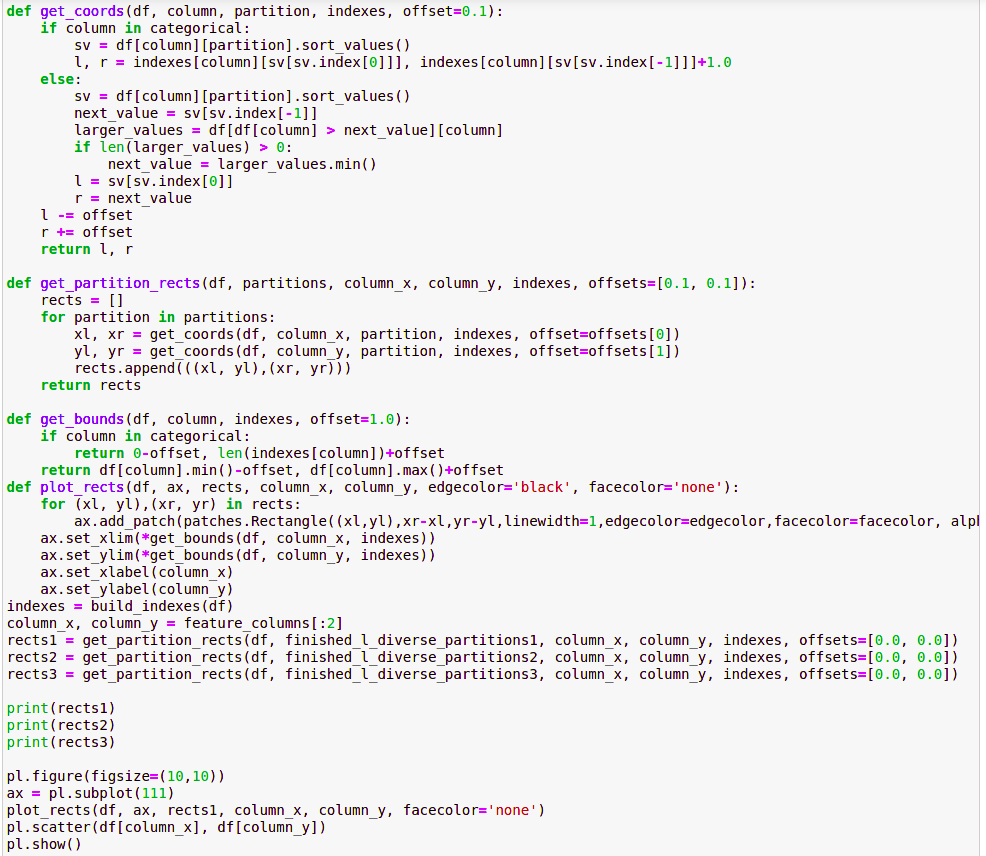


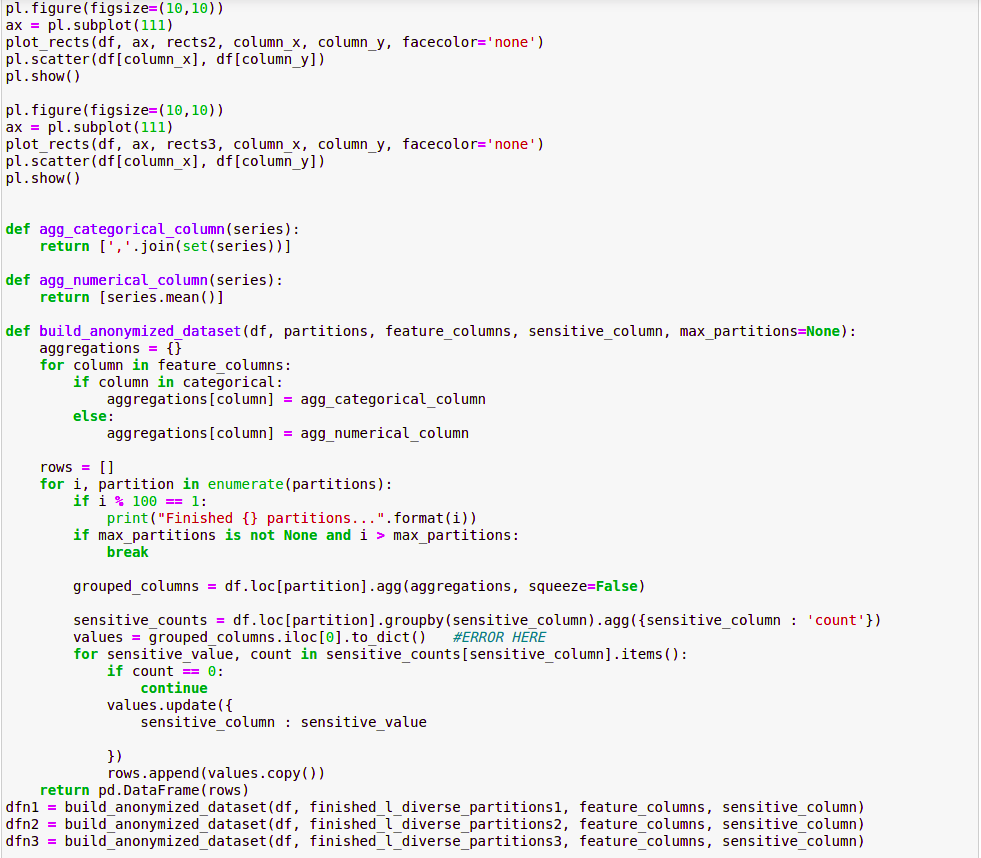
**8.2 Figure 29: Code Snippet for l-diversity model**

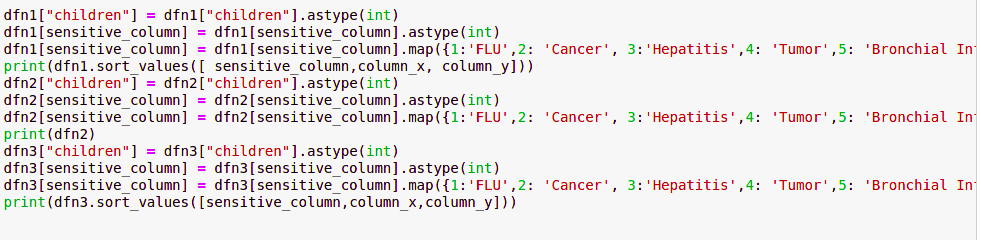
****

****

****

****

****

****

# 9.Applications of PPDM:

1. **PPDM in Cloud:** Cloud is a distributed infrastructure with great storage and

computation capabilities that is accessible through the network, anytime and anywhere.

Therefore, applications (or services) that collect, store and analyse large data quantities

often require the cloud. However, entities need to either trust cloud providers with

data, or to apply techniques that protect data while stored and/or during distributed

computation.

1. **PPDM in E-Health:** Health records are considered to be extremely private, as much of

this data is considered sensitive. However, the increase in the amount of data,combined

with the favourable properties of the cloud has led health services to store and

exchange medical records through this infrastructure . Thus, to protect from

unwanted disclosures, privacy-preserving approaches are considered.

1. **PPDM in location based services:** Technologies such as GPS have a gained a great

importance in recent times, as they allow to gain highly accurate location information.

But this comes with a disadvantage too. The location information can be used to keep a

track of a user’s activities, and discover important sensitive information about him, for

example, his workplace or his home. Also, since the information generally has

patterns, it may be used to disclose a person’s identity. Thus PPDM techniques are

applies to maintain the anonymity of users, and try to prevent the disclosure of their

sensitive information.

# 10. Future work :

1. More work needs to be done on optimal selection of values of k and l for k-anonymity

model and l- diversity model in order to hide user’s data, while maximizing the data

utility.

1. Several methods have been proposed, in order to solve the problem of privacy

preserving data mining such as k-anonymity model, l-diversity model, t-closeness

model etc. but because of the simplicity, the k-anonymity model and l- diversity is

the most worked-upon model, and many variations are tried to make this model

robust. But this does not solve the problem completely. Hence more work needs to be done in this area.

# 11. References:

* Mendes, Ricardo, and João P. Vilela. "Privacy-preserving data mining: methods, metrics, and applications." IEEE, 2017
* Kristen LeFevre, David J. DeWitt Raghu Ramakrishnan. “Mondrian Multidimensional K-Anonymity”
* Ashwin Machanavajjhala, Johannes Gehrke. “l-Diversity: Privacy Beyond k-Anonymity”
* Ninghui Li, Tiancheng Li. “t-Closeness: Privacy Beyond k-Anonymity and l-Diversity”
* Dataset: <https://www.kaggle.com/gabrielmakwana/patientdiseasedata>