DS 300 Project Report

Option #1: Design or Modify an Existing Method to Protect Data Privacy or Security

Members: Arwa Hararwala, Abhishek Mahajan, Ally Racho, & Tanya Wang

Table of Contents

Abstract:

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. Our study aims to find and implement an anonymization method through properly generalizing or suppressing the quasi-identifiers in the dataset in order to prevent linkage attacks and violations of privacy and security laws. We decided on the Mondrian Algorithm in order to implement k-anonymization on our dataset in Python. The algorithm utilizes a greedy search algorithm that allows for more desirable anonymizations than traditional exhaustive optimal algorithms for two single-dimensional models (N., 2018). Furthermore, it allows for multi-dimensional modeling, which is what's best for our specific dataset. The Mondrian Algorithm method allowed for much higher levels of privacy, but in return, took away a lot of the data utility since values of the sensitive attribute were shuffled and the quasi-identifiers were generalized. Our dataset contains a lot of unfiltered and unmasked information that could be leaked and taken advantage of by an adversary, so although k-mean clustering would achieve a higher data utility, the Mondrian Algorithm gives us the data privacy we need.

In conclusion, we believe that a combination of different algorithms may be the best choice when trying to preserve sensitive information in a dataset while also keeping data utility. The specific algorithms used will all depend on what kind of dataset we are dealing with and how much of a tradeoff between privacy and utility we are willing to make.

Data Set:

The data set that our group has chosen for this project is the medical cost (insurance) dataset from kaggle (Choi, 2018). This dataset contains 1338 rows of unique individuals and the corresponding columns to portray the different values that can affect one’s insurance charges. The columns that are included in this dataset are age, sex, bmi, children, smoker, region, charges. These columns show us the quasi-identifiers of an individual, which include age, sex, and region. The sensitive data would be the bmi, children, smoker, and charges, that are associated with an individual because it can be used against an individual if their identity is released and not properly anonymized.

Problem:

The problem with this dataset is that since several quasi-identifiers have not been generalized or suppressed, it leaves many of the categories such as age and gender available to be connected to other databases and identify the individual. Additionally, if there is a leakage in this database, there can be several legal consequences, such as violating HIPAA.

Significance of Problem:

Because the dataset pertains to medical cost 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 total amount of medical charges. Such information is beneficial to data snoopers and hackers wanting to learn more about an individual. 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.

One example was in 2019 when 78.8 million patient records had been stolen from the medical insurance company, Anthem Blue Cross. This cyber attack caused highly sensitive data such as social security number, phone number, address, name, and date of birth to be claimed and used by the attacker (Healthcare Data Breach Statistics, 2020). Anthem held too much information about their patients without proper anonymization, which allowed the attackers to easily find all of their information in one spot.

The American Collection Agency, a credit card company, had a similar data leak that occurred in 2015. In this hacking issue, 26 million records had been tampered with. The information that was taken included name, date of birth, address, phone, date of service, credit card and bank information (Healthcare Data Breach Statistics, 2020). Having all of these sensitive data values leaked to the attacker caused a ripple effect exposure of those customers and could have instilled a level of fear and distrust.

Overall, we can see that misusing and failing to provide their customers with the proper anonymity causes data leakage and hacking. Since the information is not properly masked the users are at a high risk of having their private information made public. With this project we would like to make sure to properly correct and mask the medical insurance data in order to secure the privacy of the customer, but still provide some utility from the data itself.

Project Objective:

The goal of this project is to be able to properly generalize or suppress the quasi-identifiers in the dataset in order to prevent linkage attacks and violations of privacy and security laws. We will be utilizing several different data anonymization tools and techniques in order to make sure that this data is properly anonymized so that adversaries cannot access this personal and private information. We will be using different methods within k-anonymization in order to evaluate which method will be best suited for our dataset and will allow us to find a good balance between privacy and utility. This is because if one of the methods prevents the data from being comprehensible and relevant, then it would lose its' purpose for users who would like to use it. Another concept that we will be looking into is changing the settings of the parameters in order to make sure to properly anonymize the data and provide proper utility.

Methods Used In Current Tutorials/Other Papers:

There have been articles and papers published surrounding the importance of anonymizing health information so that sensitive information will not be leaked, but also trying to maintain utility of the data for research purposes. According to an article in 2017, the need to worry about how our medical data is being processed and used is extremely crucial. There are for-profit companies using our anonymized medical data in another market and due to the advances in computing, it has become increasingly possible for outsiders to identify people, despite the anonymization, which puts many intimate secrets about our bodies and minds at risk (Tanner 2017). Anonymized medical files can be cross referenced with other sensitive files that hackers and thieves have obtained in recent years to re-identify people. Michelle De Mooy, director of the Privacy and Data Project at the Center for Democracy and Technology says that "traditional methods of anonymization from commercial entities, such as the use of patient identifiers, has also become more of a problem with the amount of data available about individuals" (Tanner 2017).

Some current privacy approaches for medical text data focus on the detection and removal of patient identifiers from the data, but this may be inadequate for protecting privacy or preserving data quality (Li & Qin 2017). A study done in 2017 proposed a new systematic approach to extract, cluster, and anonymize medical text records by integrating methods developed in both data privacy and health informatics fields (Li & Qin 2017). The key novel elements of the approach include a "recursive partitioning method to cluster medical text records based on the similarity of the health and medical information and a value-enumeration method to anonymize potentially identifying information in the text data" (Li & Qin 2017). They first identified the weaknesses in HIPAA Safe Harbor based de-identification and proposed clustering documents based on health and medical information, which can significantly increase the utility of the anonymized data. According to the study, the clustering approach is implemented with a recursive binary partitioning algorithm for controlling the level of disclosure risk (Li & Qin 2017). In class, we mentioned traditional methods like generalization or noise perturbation to anonymize the data after clustering, but the study proposes a novel value-enumeration method which results in a smaller information loss than traditional methods (Li & Qin 2017). It utilizes cluster-level and dataset-level anonymity for value enumeration and develops a drill-down method to further reduce information loss in the anonymized data (Li & Qin 2017).

In another study done in 2017 that focused on preserving the utility of anonymized medical data, they state that a very common practice for publishing privacy-preserving data is to anonymize the data before publishing, and therefore it satisfies privacy models such as k-anonymity (like what we learned in class). We also learned that generalization is often used, but it can inevitably cause information loss. This study proposed a three part method to preserve data utility: (1) utility-preserving model, (2) counterfeit record insertion, (3) catalog of the counterfeit records (Lee, Kim, Kim, Chung 2017). The anonymization algorithm they proposed using the methods applies full-domain generalization.

Analytic Methods Examined:

After evaluating the numerous methods that are available to anonymize datasets, our group has chosen to examine the Mondrian Algorithm in order to implement k-anonymization in Python. This algorithm uses a greedy search algorithm that is able to partition the data into smaller and smaller groups (N., 2018). Following the partitioning, the data values are aggregated and anonymized.

The reason we chose this method was because most often greedy search algorithms allow for more desirable anonymizations than exhaustive optimal algorithms for two single-dimensional models. Furthermore, it allows for multi-dimensional modeling, which will be best for our dataset.

Some of the other methods and algorithms that we have evaluated for k-anonymity was K-Means Clustering. This method allows us to minimize information loss and allow for better data utility. Since data records are naturally similar to each other clustering allows us to group them in equivalence classes (Byun, Kamra, Bertino, & Li, 2007). The smaller the clusters are, the less distortion is needed to the dataset, which can ensure proper utility as well as privacy. While we could have better utility the privacy may be hindered since there will not be as broad of a generalization that will be used between equivalence classes.

Taking these measures into consideration, we thought that the Mondrian Algorithm would be best suited to perform on our dataset. Since it is a greedy algorithm it will allow for higher privacy, while working to ensure proper utility as well. Furthermore, having an algorithm that is used on the data for anonymization could allow for better privacy on the dataset and be at a lower risk of being hacked, since the equivalence classes may be harder to predict and notice.

Data Wrangling:

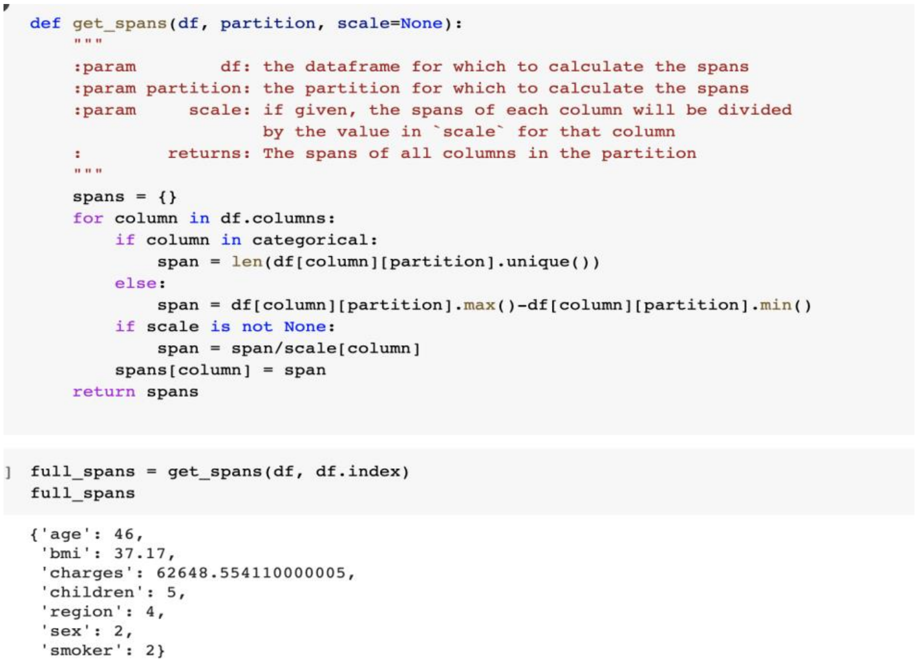
In order to be able to apply the Mondrian Algorithm to our dataset, we need to first go through some data wrangling steps in order to ensure that the data is concise, not missing any values, & is all the same format for our algorithm to work effectively.

All missing values were taken care of using mean imputation. To make sure our data would be usable for machine learning models we 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 a 0 corresponded to False/not a smoker, and 1 corresponded to being True/smoker. 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.

Results:

After performing the Mondrian Algorithm to implement K-Anonymity in Python, there are several changes that occurred in the dataset in order to properly mask the dataset and protect the privacy while allowing for utility of the data.

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 figure partitions of the data.

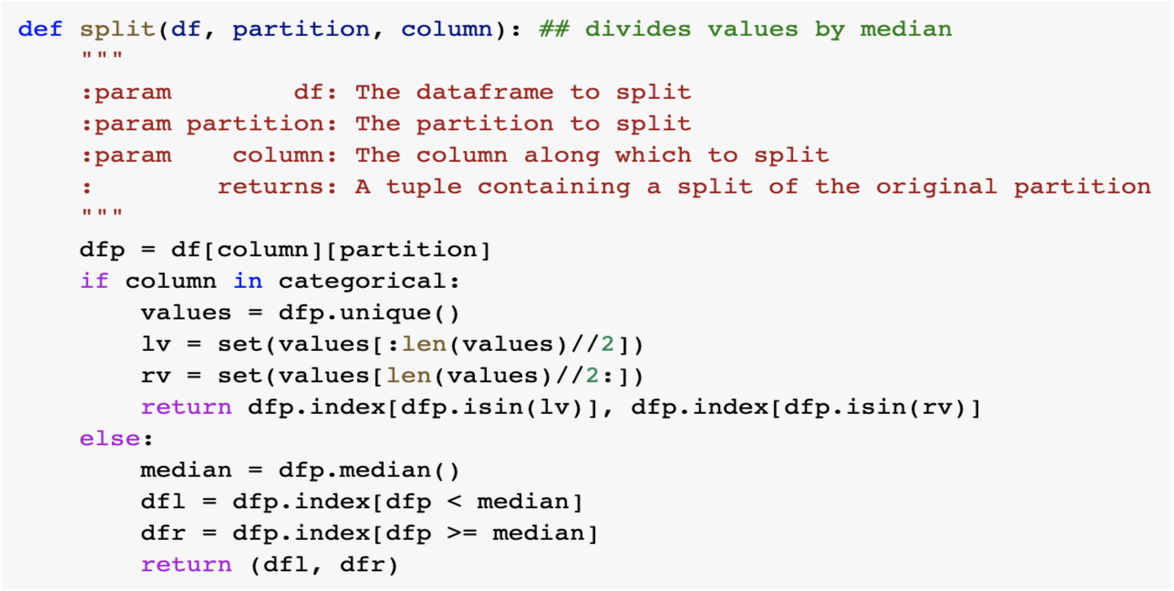


```
def get_spans(df, partition, scale=None):
    """
    :param        df: the dataframe for which to calculate the spans
    :param partition: the partition for which to calculate the spans
    :param     scale: if given, the spans of each column will be divided
                      by the value in `scale` for that column
    :        returns: The spans of all columns in the partition
    """
    spans = {}
    for column in df.columns:
        if column in categorical:
            span = len(df[column][partition].unique())
        else:
            span = df[column][partition].max()-df[column][partition].min()
        if scale is not None:
            span = span/scale[column]
        spans[column] = span
    return spans
```

```
] full_spans = get_spans(df, df.index)
  full_spans
```

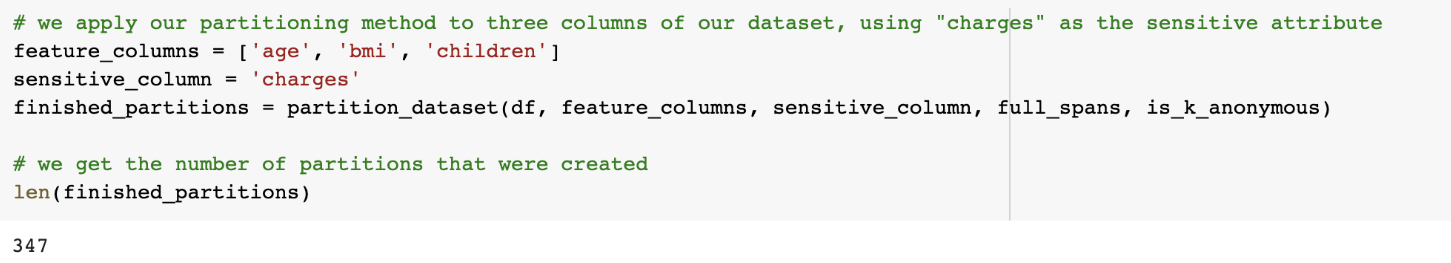
```
{'age': 46,
 'bmi': 37.17,
 'charges': 62648.554110000005,
 'children': 5,
 'region': 4,
 'sex': 2,
 'smoker': 2}
```

Following the spanning function, we implemented the split function that takes in our 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.



```
def split(df, partition, column): ## divides values by median
    """
    :param        df: The dataframe to split
    :param partition: The partition to split
    :param    column: The column along which to split
    :        returns: A tuple containing a split of the original partition
    """
    dfp = df[column][partition]
    if column in categorical:
        values = dfp.unique()
        lv = set(values[:len(values)//2])
        rv = set(values[len(values)//2:])
        return dfp.index[dfp.isin(lv)], dfp.index[dfp.isin(rv)]
    else:
        median = dfp.median()
        dfl = dfp.index[dfp < median]
        dfr = dfp.index[dfp >= median]
        return (dfl, dfr)
```

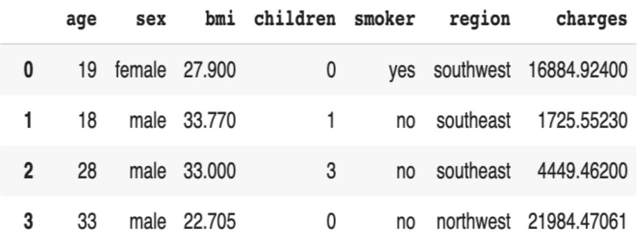
Our next step was to anonymize the partitioned dataset using K-anonymity. This was done using a value of K=3. We set the featured values to [‘age’, ‘bmi’, ‘children’] and the sensitive values to be [‘charges’].



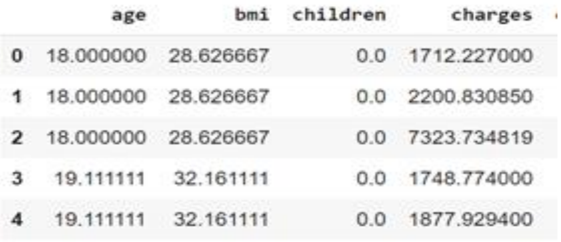
```
# we apply our partitioning method to three columns of our dataset, using "charges" as the sensitive attribute
feature_columns = ['age', 'bmi', 'children']
sensitive_column = 'charges'
finished_partitions = partition_dataset(df, feature_columns, sensitive_column, full_spans, is_k_anonymous)

# we get the number of partitions that were created
len(finished_partitions)
```

```
347
```

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.



	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061



	age	bmi	children	charges
0	18.000000	28.626667	0.0	1712.227000
1	18.000000	28.626667	0.0	2200.830850
2	18.000000	28.626667	0.0	7323.734819
3	19.111111	32.161111	0.0	1748.774000
4	19.111111	32.161111	0.0	1877.929400

Before

After

In the next section of the project, we will discuss and piece together the positives and negatives of this result, the data utility/privacy trade-off, and the aggregation/change in the data values themselves. The full code can be referenced in the appendix.

Analysis:

The Mondrian algorithm generalized our dataset over the calculated partitions with k-anonymity of 3 (K=3). Unfortunately, our dataset lost information variables [‘sex’, ‘smoker’, ‘region’] during the anonymization. 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. 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. This process, while increasing the privacy, does diminish some of the usability and utility of the dataset.

K-anonymity did increase our privacy since variables have been so generalized that individuals are more unidentifiable. High privacy is ideal for health insurance information to ensure prevention against linkage attacks of individuals and reidentification. When testing different methods and k values on our dataset we found that higher k values caused data utility to decrease even more and data became very difficult to work with. When compared, the K-means clustering method has higher data utility and lower privacy. When dealing with health insurance records, privacy is ideal and therefore the team believes the Mondrian algorithm with a k value of three is more ideal than other methods and higher k values.

Discussion & Conclusion:

After thoroughly reviewing the results and making an analysis, we can see that there were several positive and negative aspects of the Mondrian Algorithm. This method allowed for much higher levels of privacy, but also took away a lot of utility of the data since values of the sensitive attribute were shuffled and the quasi-identifiers were generalized. When comparing this model to the K-Means clustering algorithm, we can see that the K-Means clustering allowed for a higher data utility and lower levels of generalization. While this is a perk in the privacy-utility tradeoff concerns, we thought the dataset contained too much unfiltered and unmasked information that an adversary could take to their advantage of if we were to use K-Means clustering.

Some of the improvements we could have made to our model was to make sure that the sensitive data value, ‘charges’ did not have a shuffling of their values and were consistent with the original dataset. This provides us better data utility while still preserving privacy since all the other identifying attributes were generalized or suppressed.

Overall, we felt as if we were able to generate a dataset that was able to provide both privacy and utility for the user. This project was able to teach us how hard it is to be able to decide how to best mask a dataset to provide not only privacy, but also utility. It was able to point us to the several possibilities of algorithms that can be used to conduct k-anonymization and how the optimal method and algorithm varies based on the dataset at hand and the requirements for privacy and utility that one is trying to achieve.

Work Plan:

The project was split up into writing a proposal, going through the technical coding aspects, then creating presentation slides and writing up a final report to wrap up all aspects of the project.

The project proposal was divided up into sections so that each team member could contribute equally. Each person worked on 2-3 sections in order to produce the proposal and discussed certain aspects so that we could create a cohesive assignment.

In order to ensure each part of the rest of the project is done correctly, we had two of our members take on the responsibility of coding the data and implementing an anonymization method, while the other two members researched resources and started creating the presentation slides. Once the two group members were done working on the analytics, the other two members compiled sections from the proposal and newly written sections to create a final report. We all met as a group multiple times to discuss progress throughout the process so that everyone understood what was going on and if there were any obstacles. Overall, each team member was assigned sections to work on in order to integrate the work and produce a cohesive final report.

Some important deadlines we kept in mind were December 3rd, when the project slides were due, December 10th, the day we were presenting first, and December 14th, the day the final report is due on Canvas. Knowing these deadlines helped our team generate an effective work plan and we were able to stay on top of the work we needed to accomplish.

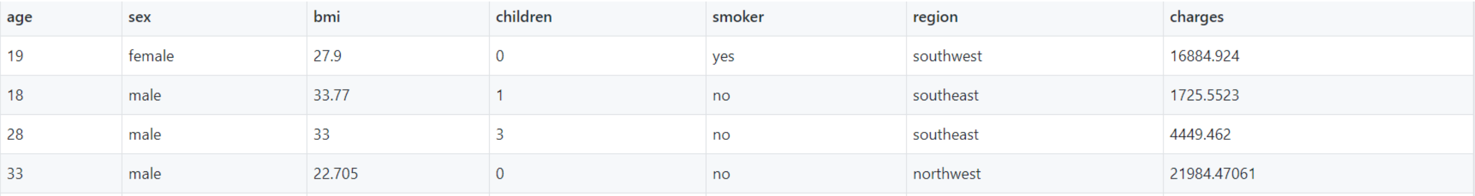
Appendix

Complete Implementation of Mondrian Algorithm in Python:

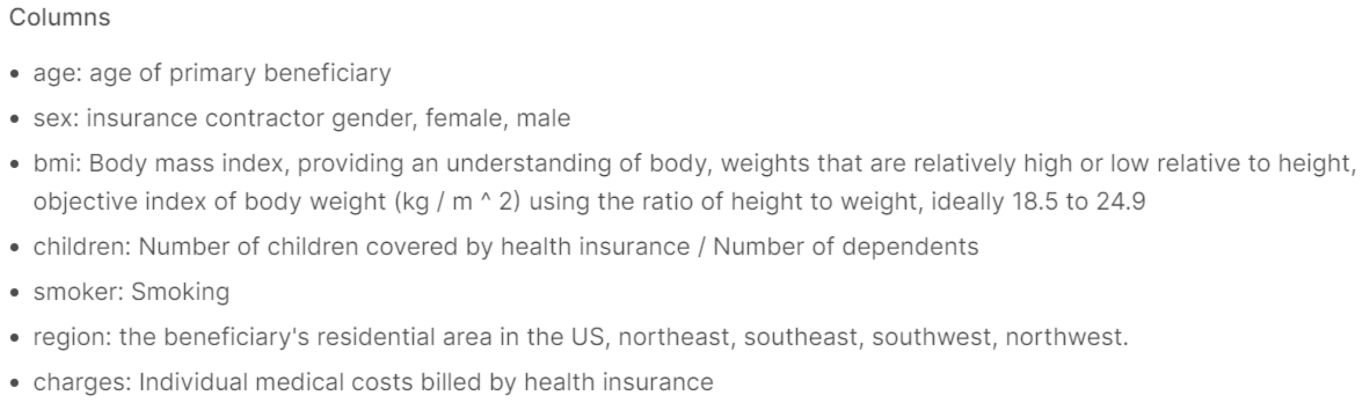
The following code was implemented and configured to fit our needs of k-anonymity. The original code was found on GitHub (N., 2018).

https://github.com/allyracho/DS-Final-Project/blob/main/projectcode.py

Dataset Info



age	sex	bmi	children	smoker	region	charges
19	female	27.9	0	yes	southwest	16884.924
18	male	33.77	1	no	southeast	1725.5523
28	male	33	3	no	southeast	4449.462
33	male	22.705	0	no	northwest	21984.47061



Columns

- age: age of primary beneficiary
- sex: insurance contractor gender, female, male
- bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
- children: Number of children covered by health insurance / Number of dependents
- smoker: Smoking
- region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- charges: Individual medical costs billed by health insurance

References

Byun, J., Kamra, A., Bertino, E., & Li, N. (2007, April 09). Efficient k-Anonymization Using Clustering Techniques. Retrieved November 30, 2020, from https://link.springer.com/chapter/10.1007/978-3-540-71703-4_18

Choi, M. (2018, February 21). Medical Cost Personal Datasets. Retrieved October 13, 2020, from https://www.kaggle.com/mirichoi0218/insurance

Healthcare Data Breach Statistics. (2020, February 18). Retrieved November 25, 2020, from https://www.hipaajournal.com/healthcare-data-breach-statistics/

Lee, H., Kim, S., Kim, J.W., & Chung, Y.D. (1970, January 01). Utility-preserving anonymization for health data publishing. Retrieved October 13, 2020, from https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-017-0499-0

Li, X. B., & Qin, J. (2017). Anonymizing and Sharing Medical Text Records. *Information Systems research : ISR*, *28*(2), 332–352. https://doi.org/10.1287/isre.2016.0676

N. (2018). K-Anonymity. Retrieved November 23, 2020, from https://github.com/Nuclearstar/K-Anonymity

Tanner, A. (2017, January 11). The Hidden Trade in Our Medical Data: Why We Should Worry. Retrieved October 13, 2020,from https://www.scientificamerican.com/article/the-hidden-trade-in-our-medical-data-why-we-should-worry/.