K-Anonymization: Medical Data

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

Data Set

- Medical cost (insurance) dataset from kaggle (Choi, 2018)
- 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**
- Quasi-identifiers of the dataset: **age, sex, and region**.
- Sensitive data is: **BMI, children, smoker, and charges**,
  - can be used against an individual if their identity is released and not properly anonymized.

Problem

- Several quasi-identifiers have not been generalized or suppressed,
- Leaves many of the categories such as **age** and **gender** available to be connected to other databases and identify the individual
- If there is a leakage in this database, there can be several legal consequences, such as violating HIPAA.

Significance of Problem

- Re Identification of the person by a linkage attack.
- 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
- If the data is leaked, legal and governmental issues such as HIPAA violations may occur because the data was not properly anonymized

Examples of Medical Data Leakage

(*Healthcare Data Breach Statistics* 2020)

- Anthem Blue Cross (2015): 78.8 million patient records had been stolen. The cyber attack claimed highly sensitive data, including names, Social Security numbers, home addresses, and dates of birth
  - Held too much sensitive information without anonymization or masking of certain identifiers to protect patients

Examples of Medical Data Leakage

(*Healthcare Data Breach Statistics* 2020)

- American Collection Agency (2019): 26 million records had been hacked. Exposed information included: first and last name, date of birth, address, phone, date of service, provider, and balance information.
  - System also included credit card or bank account information that was provided by the consumer
  - Anonymization and protection techniques were not used to prevent identification and access to all of this sensitive information

Methods Used in Current Tutorials

Study 1

Anonymizing & Sharing Medical Text Records: study done in 2017 by Xiao-Bai Li & Jialun Qin

- Proposed new systematic approach to extract, cluster, & anonymize medical text records
- Key novel elements: “recursive partitioning method to cluster medical text records based on the similarity of the healthy and medical info & a value-enumeration method to anonymize potentially identifying information in the text data”

Study 1 ctd.

- Clustering approach is implemented w/ a recursive binary partitioning algorithm for controlling the level of disclosure risk
- Instead of generalization or noise perturbation, they propose a novel value-enumeration method which results in less information loss
- Utilizes both cluster-level & dataset-level anonymity

Study 2

Utility-Preserving Anonymization For Health Data Publishing: study done in 2017 by Hyukki Lee, Soohyung Kim, Jong Wook Kim, & Yon Dohn Chung

- 3 part method to preserve data utility:
  - Utility-preserving model
  - Counterfeit record insertion
  - Catalog of the counterfeit records
- Anonymization algorithm applies full-domain generalization

Analytics Methods Examined

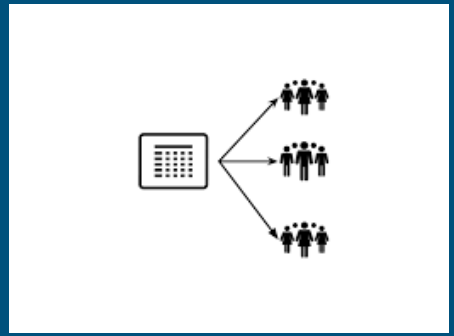
- Mondrian algorithm was used to implement k-anonymity in Python
  - This algorithm uses a greedy search algorithm to partition the data into smaller and smaller groups (N., 2018).
  - After partitioning the data, the values are aggregated and anonymization occurs.

Results

- Used spans function to calculate max-min for numerical columns and number of different values for categorical columns to create partitions of the dataset. (N., 2018)
- Implemented a split function that takes the partition and returns two partitions which split values above and below the median of the given partition (N., 2018)
- Anonymized the partitioned dataset using k-anonymity (k=3) and set the featured values to [‘age’, ‘bmi’, ‘children’] and the sensitive value to [‘charges’]
- Aggregated each column
- Generated k-anonymized dataset

	age	bmi	children	charges
24	18.000000	21.371250	0.0	1607.51010
25	18.000000	21.371250	0.0	1702.45530
26	18.000000	21.371250	0.0	13747.87235
27	18.000000	21.371250	0.0	14283.45940
33	18.000000	23.095000	0.0	1121.87390
...	...	...	...	...
978	63.333333	27.201667	0.0	29330.98315
979	63.333333	27.201667	0.0	29523.16560
1327	63.666667	31.775000	2.0	16069.08475

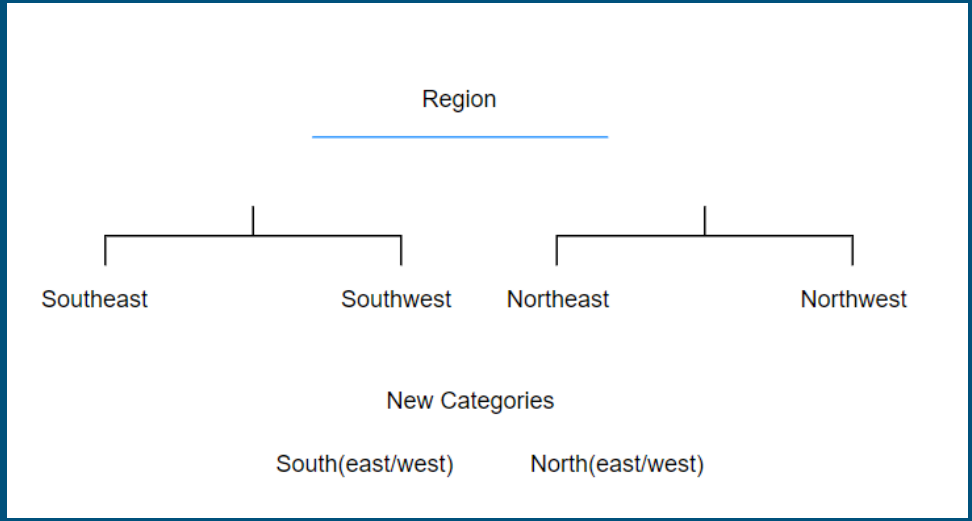
Analysis



- Dataset was generalized with k-anonymity (K=3)
- ‘age’ and ‘bmi’ were generalized to mean of partition, ‘charges’ column was swapped and unordered to make indistinguishable
- Loss of data occurred as the final anonymized output dropped our categorical variables [‘region’, ‘sex’, ’smoker’]
- Privacy increased due to generalization, utility decreased due to lack of variables

Improvements

- Generalizing region data to increase data utility as opposed to suppression



Discussion

- K-Means Clustering method provides higher utility/lower privacy
  - Since dataset pertains to personal health information, low privacy is not ideal as it allows for re-identification and linkage attacks
- Increasing K-anonymity values in Mondrian method caused data utility to decrease, making data difficult to use
- Keeping ‘charges’ unshuffled could improve utility

References

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

References ctd.

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/.