

## 1. Cel projektu

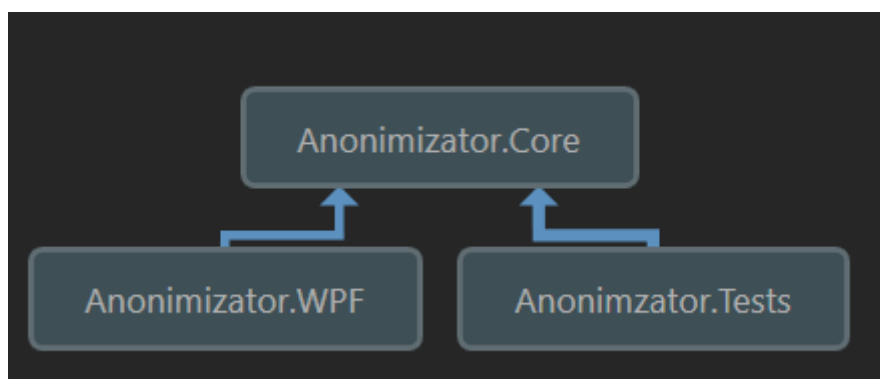
Celem projektu było zapoznanie się z tematem anonimizacji tabel danych oraz procesem odwrotnym, pozwalającym na pozyskanie pierwotnych danych z przetworzonych tabel. Ze względu na ograniczenia czasowe oraz eksperymentalną naturę projektu wszelkie rozwiązania zostały zaimplementowane pod określony schemat tabeli danych.

## 2. Opis rozwiązania

Zaimplementowano aplikację desktopową pozwalającą zarówno na przeprowadzenie anonimizacji danych wybranymi algorytmami jak i udostępniającą interfejs ułatwiający analizę potencjalnie zanonimizowanych tabel danych.

### 2.1. Architektura oprogramowania

Ze względu na stopniowy wzrost skomplikowania oraz zakresu projektu zdecydowano się na rozdzielenie warstwy prezentacyjnej od warstwy logiki zawierającej implementację wszelkich algorytmów. Dodatkowo, w celu ułatwienia testowania poprawności implementacji oraz automatyzacji testów regresyjnych istniejących funkcjonalności utworzono projekt z testami jednostkowymi.



Rysunek 1 Architektura rozwiązania

- W warstwie prezentacyjnej wykorzystano technologię Windows Presentation Foundation (WPF) wchodzącą w skład frameworka .NET 4.5. Wykorzystano wzorzec MVVM umożliwiający oddzielenie kodu źródłowego opisującego widok od danych jakie reprezentuje.
- Warstwa logiki (Core) została zaimplementowana używając klas wchodzących w zakres .NET Standard 2.0 co w praktyce umożliwia dołączanie jej do projektów zarówno używających .NET Frameworka jak i zbudowanych w .NET Core (obydwa implementują .NET Standard 2.0).
- W projekcie testowym wykorzystano bibliotekę MSTest. Testy zostały napisane dla każdego z zaimplementowanych algorytmów jak i dla


pomocniczych klas lub metod, które wyróżniały się potencjalnymi nietrywialnymi problemami.

## 2.2. Algorytmy

### a. Algorytmy anonimizacji dla jednej kolumny

Pierwszym etapem projektu było skonstruowanie algorytmów anonimizacji wybranej kolumny. Umożliwiły one przeprowadzenie procesu k-anonimizacji i jego wariantów dla PIDu składającego się z jednego atrybutu. W dalszym etapie wykorzystano różne kombinacje tych algorytmów dla anonimizacji wykonanej na złożonym pseudo identyfikatorze.

- **Supresja atrybutów** – odnosi się do usunięcia całej partii danych, w bazach danych nazywanej również “kolumną”, lub zastąpienia ich jedną wartością (np. “\*”).




Gender	Job	City	FirstName
M	Tancerz	Lipsko	C
M	Inzynier	Katowice	A
M	Programista	Lubin	A
M	Inzynier	Wroclaw	J
M	Inzynier	Wroclaw	K
M	Inzynier	Czestochowa	J
M	Programista	Sopot	V
M	Programista	Wroclaw	S
M	Programista	Siedlce	S
M	Programista	Wroclaw	F

Gender	Job	City	FirstName
*	Inzynier	Wroclaw	Jan
*	Inzynier	Katowice	Adam
*	Inzynier	Wroclaw	Kamil
*	Inzynier	Czestochowa	Jakub
*	Programista	Wroclaw	Szymon
*	Programista	Wroclaw	Filip
*	Malarz	Rybnik	Mikolaj
*	Malarz	Olawa	Julia
*	Tancerz	Olawa	Zofia
*	Muzyk	Wroclaw	Kacper
*	Muzyk	Wroclaw	Kacper

Rysunek 2 Anonimizacja kolumny z wykorzystaniem supresji atrybutów

- **Generalizacja wartości liczbowych** – celowe obniżenie precyzji danych poprzez zmianę dokładnych wartości liczbowych na przedziały w których się znajdują.

FirstName	Surname	Age
Wojciech	Kowalski	17
Szymon	Osowski	19
Szymon	Kowal	19
Filip	Nowak	22
Roman	Nowak	22
Daniel	Kowal	22
Cezary	Wojcik	23
Andrzej	Malinowski	23
Adam	Kowalski	23
Wiktor	Kowalski	23
Jozef	Kowalski	23
Krystian	Malinowski	24
Beata	Osowski	24



FirstName	Surname	Age
Jan	Nowak	15 - 17
Kamil	Malinowski	15 - 17
Jakub	Jagiel	15 - 17
Wojciech	Kowalski	15 - 17
Szymon	Osowski	19 - 22
Szymon	Kowal	19 - 22
Filip	Nowak	19 - 22
Roman	Nowak	19 - 22
Daniel	Kowal	19 - 22
Cezary	Wojcik	23 - 24
Jozef	Kowalski	23 - 24
Andrzej	Malinowski	23 - 24
Adam	Kowalski	23 - 24


Rysunek 3 Anonimizacja kolumny z wykorzystaniem generalizacji wartości liczbowych

- **Generalizacja wartości tekstowych** - grupowanie wartości tekstowych w zbiory elementów, których długości znajdują się w określonym przedziale.

FirstName	Sur	FirstName	Surna
Jan	Nov	3 - 4 letters	Nowal
Adam	Kov	3 - 4 letters	Kowal:
Kamil	Mal	3 - 4 letters	Szyma
Jakub	Jagi	3 - 4 letters	Kowal:
Szymon	Osc	3 - 4 letters	Kowal:
Filip	Nov	3 - 4 letters	Malinc
Mikolaj	Kov	3 - 4 letters	Kowal:
Julia	Woj	5 letters	Malinc
Zofia	Kov	5 letters	Jagiel
Kacper	Szy	5 letters	Nowal
Kacper	Szy	5 letters	Wnicil


Rysunek 4 Anonimizacja kolumny z wykorzystaniem generalizacji wartości tekstowych

- **Częściowe maskowanie wartości tekstowych** - grupowanie wartości tekstowych w zbiory zawierające jak najdłuższy wspólny początek. Jeśli wartość okazywała się zbyt unikalna (w zbiorze nie było wystarczająco elementów o takim samym początku aby spełniony został warunek K-anonimizacji) zamieniano ją na "\*". Elementów o wartości "\*" również musi być co najmniej K.

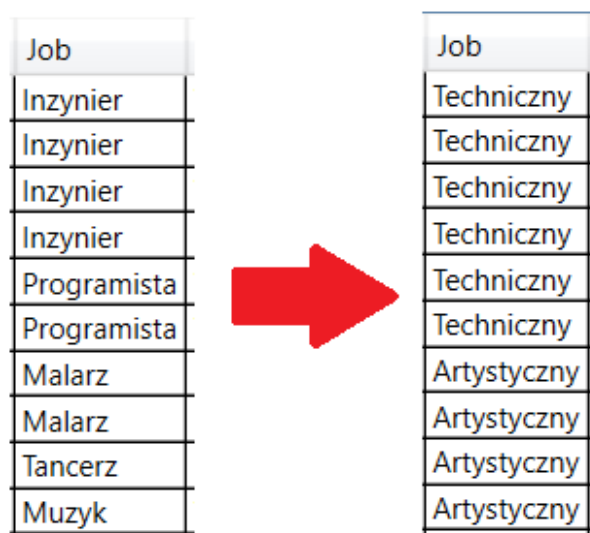
FirstName	Surname	Age		FirstName	Surname	Age
Dawid	Kowalski	60		Mikolaj	Kowa...	30
Wiktor	Nowak	90		Zofia	Kowa...	43
Szymon	Kowal	19		Dawid	Kowa...	60
Edward	Jagiel	45		Szymon	Kowa...	19
Cezary	Wojcik	23		Adam	Kowa...	11
Beata	Osowski	24		Wojciech	Kowa...	28
Adam	Kowalski	11		Wojciech	Kowa...	17
Kamil	Malin	32		Joanna	Kowa...	39
Krystian	Nowak	37		Daniel	Kowa...	22
Joanna	Szymanski	55		Jozef	Kowa...	23

Rysunek 5 Anonimizacja kolumny z wykorzystaniem częściowego maskowania wartości tekstowych

- **Generalizacja słownikowa** – celowe obniżenie precyzji danych poprzez zmianę szczegółowych wartości na wartości ze słownika o mniejszej dokładności jednak opisujące daną wartość. W tworzonym programie wykorzystane dwa słowniki: miast i zawodów. W przypadku słownika miast nazwy miast zamieniane były na nazwy województw lub państw. Nazwy zawodów natomiast były uogólniane z bardzo szczegółowych na np. zawód techniczny lub artystyczny.

City		City
Wroclaw		Dolnoslaskie
Katowice		Slaskie
Wroclaw		Dolnoslaskie
Czestochowa		Slaskie
Wroclaw		Dolnoslaskie
Wroclaw		Dolnoslaskie
Rybnik		Slaskie
Olawa		Dolnoslaskie
Olawa		Dolnoslaskie
Wroclaw		Dolnoslaskie

Rysunek 6 Anonimizacja kolumny za pomocą generalizacji słownikowej z wykorzystaniem słownika miast



Rysunek 7 Anonimizacja kolumny za pomocą generalizacji słownikowej z wykorzystaniem słownika zawodów

#### b. Algorytmy anonimizacji dla wybranego PID

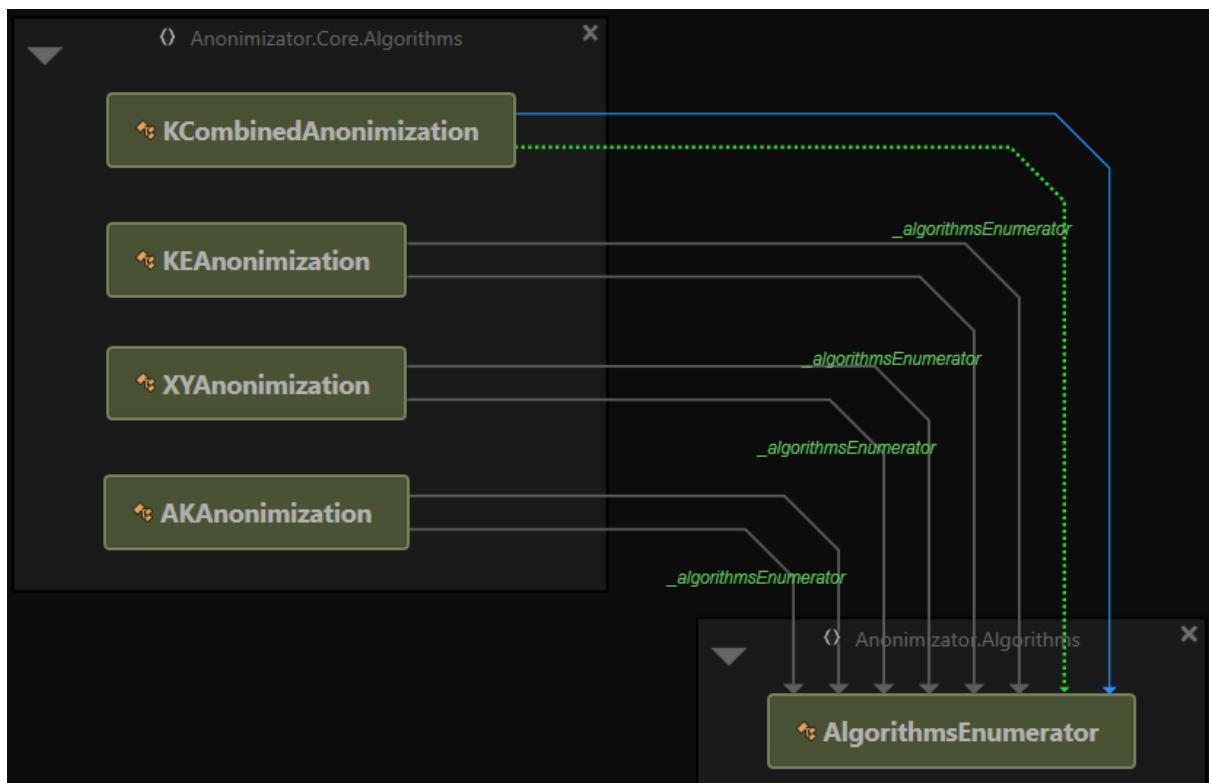
Do implementacji różnych wariantów K anonimizacji wykorzystano kombinację algorytmów opisanych powyżej. W celu znalezienia kombinacji spełniającej kryteria danego algorytmu przeszukiwano całą przestrzeń rozwiązań dla wybranego pseudo identyfikatora wykorzystując następujące algorytmy.

- Gender - Supresja atrybutów
- City - Generalizacja słownikowa
- Job - Generalizacja słownikowa
- Surname - Generalizacja wartości tekstowych
- Firstname - Generalizacja wartości tekstowych
- Age - Generalizacja wartości liczbowych

W celu spełnienia warunku k-anonimizacji dla całego PIDu manipulowano wartościami algorytmów anonimizujących dla pojedynczych kolumn poprzez inkrementację wartości K dla każdego z atrybutów. Jeśli wartość K osiągnęła maksymalną wartość graniczną (20) używano “Supresji atrybutów”, która w praktyce znacząco obniża unikalność wartości PID.

#### \*Trudności implementacyjne

Rozwiązanie zostało zaprojektowane w możliwie modułowy sposób. Wyodrębniono klasę generującą kolejne kombinację algorytmów anonimizujących, dzięki czemu wszystkie inne klasy reprezentujące warianty k-anonimizacji mogły z niej korzystać.



Rysunek 8 Struktura zależności między wybranymi klasami

W związku z mnogością opcji jakie mogą zostać przyjęte przez klasę generującą kombinację algorytmów (*AlgorithmsEnumerator*) wykorzystano również kreatywny wzorec projektowy - budowniczy oraz idee fluent API. Dzięki takiemu rozwiązaniu dużo łatwiej jest interpretować jakie dokładnie opcje są ustawione w danym algorytmie. Dodawanie kolejnych opcji wiąże się z dodaniem metody do klasy buildera z odpowiednią nazwą co pozwala na swobodne rozszerzanie funkcjonalności bez obniżania czytelności kodu.

Listing 1 Wykorzystanie wzorca budowniczy

```
_algorithmsEnumerator = new AlgorithmsEnumeratorBuilder()
    .SetMaximumKParameter(100)
    .SetPID(_xExpressions.ToArray())
    .AddDictionary(p => p.City, cityDictionary)
    .AddDictionary(p => p.Job, jobDictionary)
    .Build();
```

Aby maksymalnie wykorzystać statyczny system typów wykorzystywany przez język C# zdecydowano się operować na typach *Expression<T>* oraz *Func<T>*. Nie wnikając w dalsze szczegóły implementacyjne, pozwoliło nam to na typowanie atrybutów wybranych jako PID poprzez wybieranie ich z obiektu reprezentującego wiersz tabeli.

Listing 2 Wybór kolumn należących do PID-a

```
var pid = new Expression<Func<Person, object>>[] { p => p.FirstName, p => p.Surname, p => p.Age };
```

Dodatkowym plusem takiego podejścia jest możliwość wykorzystania wielu ze stworzonych klas w kontekście anonimizacji tabeli o całkiem innej strukturze,

reprezentowanej przez inną klasę. Porównywanie identyczności obiektów typu Expression nie jest wbudowane w język, dlatego też należało je zaimplementować. Testy sprawdzające poprawność tego rozwiązania dołączono do projektu testowego.

- **K - anonimizacja**

Uwzględniając powyżej opisane rozwiązania, algorytm k-anonimizacji można zapisać w kilku liniach jednej metody.

Listing 3 Algorytm anonimizacji

```
public List<Person> GetAnonymizedData(IEnumerable<Person> people)
{
    if (people == null || !people.Any())
        return new List<Person>();

    var groups = GetGroupedPeople(people);

    foreach (var algorithms in _algorithmsEnumerator)
    {
        var anonymizedData = algorithms.Aggregate(people.Clone(), (acc, algo) => algo.GetAnonymizedData(acc));
        groups = GetGroupedPeople(anonymizedData);
        if (IsListAnonymized(groups))
            break;
    }

    return groups.SelectMany(x => x.People).ToList();
}
```

Gdzie warunek anonimizacji mówi o tym, że każda grupa utworzona przez unikalną wartość PID ma licznosc przynajmniej K.

Listing 4 Warunek spełnienia założeń algorytmu k-anonimizacji

```
private bool IsListAnonymized(IEnumerable<PeopleGroup<string>> groups)
{
    return groups.All(g => g.Count >= ParameterK);
}
```

- **K-(X, Y) - anonimizacja**

Metoda K-(X, Y) - anonimizacji jest uogólnieniem metody k-anonimizacji. Założeniem tej metody jest podział danych na dwa rozłączne zbiory kolumn. Metoda wymaga, aby dla każdej unikatowej wartości ze zbioru X występowało co najmniej k różnych wartości ze zbioru Y.

Z założeń wynika, że metoda k-anonimizacji jest szczególnym przypadkiem metody K-(X, Y) - anonimizacji została więc również w tym przypadku wykorzystana metoda *GetAnonymizedData*, a zmianie w porównaniu do innych opisanych metod uległa implementacja metody *IsListAnonymized*.

Listing 5 Warunek spełnienia założeń algorytmu K-(X, Y)-anonimizacji

```
private bool IsListAnonymized(IEnumerable<PeopleGroup<string>> groups)
{
    return groups.All(g =>
    {
        var yPropertiesGroups = g.People.GroupBy(p => p.GetPersonProperties(_yExpressions.ToArray())).ToList();
        return yPropertiesGroups.Count >= ParameterK;
    });
}
```

### • ( $\alpha$ , k) - anonimizacja

Metoda ( $\alpha$ , k) - anonimizacji jest szczególnym przypadkiem metody k-anonimizacji gdzie oprócz spełnienia warunku k-anonimizacji musi być równocześnie spełniony warunek  $\alpha$ -deasocjacji. Spełnienie warunku k-anonimizacji wymaga, aby liczba rekordów w grupach wydzielonych przez pseudo identyfikator była większa lub równa zadanej wartości parametru k. Spełnienie warunku  $\alpha$ -deasocjacji wymaga natomiast, żeby dla zadanej wartości wrażliwej prawdopodobieństwo jej wystąpienia w każdej z grup wyznaczonych przez pseudo identyfikator było mniejsze lub równe wartości parametru  $\alpha$ . Wartość wrażliwa to określona wartość atrybutu znajdującego się w jednej z kolumn.

Ze względu na podobieństwo pomiędzy algorytmami metoda *GetAnonymizedData* jest identyczna do przedstawionej dla algorytmu k-anonimizacji. Algorytm różni się natomiast implementacją metody *IsListAnonymized* sprawdzającej czy tabela spełnia warunki ( $\alpha$ , k) - anonimizacji.

Listing 6 Warunek spełnienia założeń algorytmu ( $\alpha$ , k)-anonimizacji

```
private bool IsListAnonymized(IEnumerable<PeopleGroup<string>> groups)
{
    var alphaParameterCondition = groups.All(g =>
    {
        var numberItems = g.People.Select(_selectedAttributeProperty).Count(p => p.ToString() == AttributeValue);
        var groupSize = g.People.Select(_selectedAttributeProperty).Count();
        return numberItems / (double)groupSize <= ParameterAlpha;
    });

    var kParameterCondition = groups.All(g => g.Count >= ParameterK);

    return alphaParameterCondition && kParameterCondition;
}
```

### • (k, e) - anonimizacja

Metoda (k, e) - anonimizacja jest wariantem metody k-anonimizacji skierowanym do ochrony wrażliwych danych liczbowych. Założeniem tej metody jest spełnienie równocześnie warunku k-anonimizacji i zapewnieniu maksymalnej różnicy wartości liczbowych w grupach wydzielonych zgodnie z założeniami k-anonimizacji wynoszącej przynajmniej tyle ile wynosi wartość parametru e.

W przypadku tego algorytmu implementacja metody *GetAnonymizedData* jest identyczna do tej zaprezentowanej w listingu 3. Algorytm różni się natomiast implementacją metody *IsListAnonymized*. Wyliczana jest minimalna i maksymalna wartość liczbową w danej kolumnie, a następnie różnica wartości maksymalnej od minimalnej jest porównywana do wartości parametru e.

Listing 7 Warunek spełnienia założeń algorytmu (k, e)-anonimizacji



```
private bool IsListAnonymized(IEnumerable<PeopleGroup<string>> groups)
{
    var eParameterCondition = groups.All(g =>
    {
        var min = g.People.Select(_anonymizedProperty).Min(p => Convert.ToInt32(p.ToString()));
        var max = g.People.Select(_anonymizedProperty).Max(p => Convert.ToInt32(p.ToString()));
        return max - min >= ParameterE;
    });

    var kParameterCondition = groups.All(g =>
    {
        var uniqueValues = g.People.GroupBy(_anonymizedProperty).Count();
        return uniqueValues >= ParameterK;
    });

    return eParameterCondition && kParameterCondition;
}
```

### c. Narzędzia wspomagające proces deanonimizacji danych

W ramach projektu zostały zrealizowane zakładki wspomagające użytkownika w procesie deanonimizacji danych poprzez rozpoznanie potencjalnie zanonimizowanych kolumn oraz wyliczenie parametru K K-anonimizacji.

Trzy zakładki realizujące proces deanonimizacji danych:

- **Analiza danych**

Proces anonimizacji danych powoduje zmniejszenie różnorodności wartości atrybutów w poszczególnych kolumnach. W celu zobrazowania liczby poszczególnych wartości w konkretnych kolumnach została zastosowana metoda grupowania wartości po identycznych wartościach lub równej liczbie liter ciągów znaków. Do prezentacji wyników zostały wykorzystane wykresy słupkowe. Użytkownik ma możliwość sprawdzenia liczności grup dla każdej kolumny wprowadzonego zbioru danych.

- **Wyliczanie parametru K**

Po przeanalizowaniu liczności grup tworzonych przez wartości w kolumnach i wytypowaniu kolumn należących do potencjalnego PID-a użytkownik może wskazać kolumny i wyliczyć minimalny parametr K dla którego wybrane kolumny spełniają warunek K-anonimizacji.

- **Rozpoznawanie parametru K**

Uogólnieniem procesu wyliczania parametru K jest sprawdzenie wszystkich możliwych kombinacji kolumn mogących tworzyć PID i wyliczenie dla nich wartości parametru K dla którego spełniony zostanie warunek algorytmu K-anonimizacji. Wyniki przeprowadzonych obliczeń zostały zaprezentowane w formie tabeli.

### 3. Możliwe rozszerzenia

Aplikacja może zostać rozszerzona o kolejne algorytmy anonimizujące opisane w artykule [3] takie jak metoda l-dywersyfikacji, (X-Y) - dołączalności czy też t-bliskości.

Możliwe jest też dodanie nowych algorytmów anonimizujących wybraną kolumnę. Warto rozpatrzyć nowe sposoby na stworzenie algorytmów, które korzystają z wiedzy domenowej ekspertów. Wiele możliwości daje też sama manipulacja wartościami tekstowymi w celu ich częściowego maskowania. W przypadku naszego rozwiązania poszukiwaliśmy jedynie wspólny początek wyrazów.

Budowa aplikacji pozwala na zastosowanie innych enumeratorów generujących zestawy algorytmów anonimizujących. Ulepszenie aktualnego rozwiązania może polegać na zastosowaniu pewnego rodzaju heurystyk umożliwiających wydajniejsze przeszukiwanie przestrzeni rozwiązań.

Rozszerzenie narzędzi pozwalających na analizę danych może zawierać w sobie próbę automatycznego wykrycia PIDu z określoną wartością pewności tego wykrycia.

### 4. Instrukcja użytkownika

Aplikacja jest intuicyjna i umożliwia użytkownikowi sprawne przeprowadzenie anonimizacji oraz deanonimizacji. Program umożliwia użytkownikowi przeprowadzanie operacji anonimizacji i deanonimizacji na przygotowanym zbiorze danych lub wygenerowanie losowego zbioru danych o dowolnej liczbie rekordów. W zależności od zakładki aplikacja oferuje różne operacje na danych i różne sposoby parametryzowania tych operacji w zależności od stosowanego algorytmu. Dodatkowo użytkownik ma możliwość zapisania wyników przeprowadzonej anonimizacji do pliku o wprowadzonej nazwie.

Wyniki przeprowadzonych operacji są przechowywane w tymczasowym pliku co umożliwia użytkownikowi wczytywanie wyników operacji przeprowadzanych w różnych zakładkach aplikacji. Dane do tymczasowego pliku są zapisywane automatycznie po przeprowadzeniu dowolnej operacji anonimizacji. Funkcjonalność ta umożliwia również wykonywanie różnych rodzajów anonimizacji na tym samym zbiorze danych.

Przykładowo przeprowadzamy anonimizację z wykorzystaniem algorytmu  $(\alpha, k)$  - anonimizacji na kolumnach: City, FirstName i Surname, a następnie na wyniku przeprowadzonej anonimizacji przeprowadzamy anonimizację na kolumnie Job. Do wykonania tej operacji należy po przeprowadzeniu pierwszej anonimizacji przejść do zakładki "K-anonimizacja" i wcisnąć przycisk "Wczytaj dane tymczasowe".

W zakładkach służących do przeprowadzania deanonimizacji użytkownik ma możliwość wczytania dowolnego pliku csv z danymi. Ma to umożliwić sprawdzenie czy

wczytane dane są danymi zanonimizowanymi oraz które kolumny tworzą PID. Ze względu na założenia projektowe dane wczytywane do programu muszą być zgodne z przyjętą strukturą danych i zawierać kolumny o określonych w projekcie wartościach. Zaleca się wczytywanie danych wcześniej zapisanych z poziomu programu do pliku.

Funkcje dostępne w większości zakładek:

- zapis do pliku wyników anonimizacji poprzez przycisk “Zapisz”
- przywrócenie danych sprzed wykonywania operacji anonimizacji poprzez przycisk “Przywróć dane”
- przywrócenie danych tymczasowych zapisanych po wykonaniu anonimizacji w innym oknie poprzez przycisk “Przywróć dane tymczasowe”

## 4.1. Anonimizacja kolumn

Gender	Job	City	FirstName	Surname	Age
M	Inżynier	Wrocław	Jan	Nowak	15
M	Inżynier	Katowice	Adam	Kowalski	11
M	Inżynier	Wrocław	Kamil	Malinowski	15
M	Inżynier	Częstochowa	Jakub	Jagiel	15
M	Programista	Wrocław	Szymon	Osowski	19
M	Programista	Wrocław	Filip	Nowak	22
M	Malarz	Rybnik	Mikolaj	Kowalski	30
K	Malarz	Olawa	Julia	Wojcik	40
K	Tancerz	Olawa	Zofia	Kowalczyk	43
M	Muzyk	Wrocław	Kacper	Szymanski	50
M	Muzyk	Wrocław	Kacper	Szymanski	51
M	Muzyk	Wrocław	Damian	Nowakowski	53
M	Muzyk	Swidnica	Dawid	Kowalski	60
M	Muzyk	Rybnik	Wiktoria	Nowak	90
M	Programista	Siedlce	Szymon	Kowal	19
M	Programista	Glubczyce	Edward	Jagiel	45
M	Muzyk	Namyslow	Cezary	Wojcik	23
K	Programista	Wyszkow	Beata	Osowski	24
M	Programista	Lubin	Adam	Kowalski	11
M	Programista	Zabrze	Kamil	Malin	32
M	Muzyk	Olawa	Krystian	Nowak	37
K	Muzyk	Lubin	Joanna	Szymanski	55
M	Muzyk	Ciechanow	Roman	Nowak	22

Rysunek 9 Zakładka do supresji atrybutów kolumn

Zakładka umożliwia zastosowanie supresji atrybutów dla konkretnych kolumn poprzez wybór kolumny z combobox-a i wciśnięcie przycisku “Maskuj znaki”. Dostępne są również opcje “Zapisz”, “Przywróć dane” i “Przywróć dane tymczasowe”.

## 4.2. K-anonimizacja kolumn

Gender	Job	City	FirstName	Surname	Age
M	Inzynier	Wroclaw	Jan	Nowak	15
M	Inzynier	Katowice	Adam	Kowalski	11
M	Inzynier	Wroclaw	Kamil	Malinowski	15
M	Inzynier	Czestochowa	Jakub	Jagiel	15
M	Programista	Wroclaw	Szymon	Osowski	19
M	Programista	Wroclaw	Filip	Nowak	22
M	Malarz	Rybnik	Mikolaj	Kowalski	30
K	Malarz	Olawa	Julia	Wojcik	40
K	Tancerz	Olawa	Zofia	Kowalczyk	43
M	Muzyk	Wroclaw	Kacper	Szymanski	50
M	Muzyk	Wroclaw	Kacper	Szymanski	51
M	Muzyk	Wroclaw	Damian	Nowakowski	53
M	Muzyk	Swidnica	Dawid	Kowalski	60
M	Muzyk	Rybnik	Wiktor	Nowak	90
M	Programista	Siedlce	Szymon	Kowal	19
M	Programista	Glubczyce	Edward	Jagiel	45
M	Muzyk	Namyslow	Cezary	Wojcik	23
K	Programista	Wyszkow	Beata	Osowski	24
M	Programista	Lubin	Adam	Kowalski	11
M	Programista	Zabrze	Kamil	Malin	32
M	Muzyk	Olawa	Krystian	Nowak	37
K	Muzyk	Lubin	Joanna	Szymanski	55
M	Muzyk	Ciechanow	Roman	Nowak	22

Parametr K: 1  
Kolumna: Age

Anonimizuj Zapisz  
Przywróć dane Przywróć dane tymczasowe

Rysunek 10 Zakładka do k-anonimizacji kolumn

Zakładka umożliwia zastosowanie algorytmu k-anonimizacji do konkretnych kolumn poprzez wybór kolumny z combobox-a i wprowadzenie żadanego parametru k. Anonimizacja następuje po wciśnięciu przycisku “Anonimizuj”. Zaleca się stosowanie anonimizacji jednokrotnie na danej kolumnie. Dostępne są również opcje “Zapisz”, “Przywróć dane” i “Przywróć dane tymczasowe”.

## 4.3. PID K-anonimizacja

Gender	Job	City	FirstName	Surname	Age
M	Inzynier	Wroclaw	Jan	Nowak	15
M	Inzynier	Katowice	Adam	Kowalski	11
M	Inzynier	Wroclaw	Kamil	Malinowski	15
M	Inzynier	Czestochowa	Jakub	Jagiel	15
M	Programista	Wroclaw	Szymon	Osowski	19
M	Programista	Wroclaw	Filip	Nowak	22
M	Malarz	Rybnik	Mikolaj	Kowalski	30
K	Malarz	Olawa	Julia	Wojcik	40
K	Tancerz	Olawa	Zofia	Kowalczyk	43
M	Muzyk	Wroclaw	Kacper	Szymanski	50
M	Muzyk	Wroclaw	Kacper	Szymanski	51
M	Muzyk	Wroclaw	Damian	Nowakowski	53
M	Muzyk	Swidnica	Dawid	Kowalski	60
M	Muzyk	Rybnik	Wiktor	Nowak	90
M	Programista	Siedlce	Szymon	Kowal	19
M	Programista	Glubczyce	Edward	Jagiel	45
M	Muzyk	Namyslow	Cezary	Wojcik	23
K	Programista	Wyszkow	Beata	Osowski	24
M	Programista	Lubin	Adam	Kowalski	11
M	Programista	Zabrze	Kamil	Malin	32
M	Muzyk	Olawa	Krystian	Nowak	37
K	Muzyk	Lubin	Joanna	Szymanski	55
M	Muzyk	Ciechanow	Roman	Nowak	22

Parametr K: 0

Anonimizuj Zapisz  
Przywróć dane Przywróć dane tymczasowe

Rysunek 11 Zakładka do k-anonimizacji z wykorzystaniem PID-a

Zakładka umożliwia zastosowanie algorytmu k-anonimizacji dla konkretnych kolumn tworzących PID. Kolumny tworzące PID można wybrać w prawym panelu

zakładki. Anonimizacja następuje po wciśnięciu przycisku “Anonimizuj”. Użytkownik może również wybrać parametr K anonimizacji. Dostępne są również opcje “Zapisz”, “Przywróć dane” i “Przywróć dane tymczasowe”.

#### 4.4. K-(X, Y)-anonimizacja

Gender	Job	City	FirstName	Surname	Age
M	Inzynier	Wroclaw	Jan	Nowak	15
M	Inzynier	Katowice	Adam	Kowalski	11
M	Inzynier	Wroclaw	Kamil	Malinowski	15
M	Inzynier	Czestochowa	Jakub	Jagiel	15
M	Programista	Wroclaw	Szymon	Osowski	19
M	Programista	Wroclaw	Filip	Nowak	22
M	Malarz	Rybnik	Mikolaj	Kowalski	30
K	Malarz	Olawa	Julia	Wojcik	40
K	Tancerz	Olawa	Zofia	Kowalczyk	43
M	Muzyk	Wroclaw	Kacper	Szymanski	50
M	Muzyk	Wroclaw	Kacper	Szymanski	51
M	Muzyk	Wroclaw	Damian	Nowakowski	53
M	Muzyk	Swidnica	Dawid	Kowalski	60
M	Muzyk	Rybnik	Wiktor	Nowak	90
M	Programista	Siedlce	Szymon	Kowal	19
M	Programista	Glubczyce	Edward	Jagiel	45
M	Muzyk	Namyslow	Cezary	Wojcik	23
K	Programista	Wyszkow	Beata	Osowski	24
M	Programista	Lubin	Adam	Kowalski	11
M	Programista	Zabrze	Kamil	Malin	32
M	Muzyk	Olawa	Krystian	Nowak	37
K	Muzyk	Lubin	Joanna	Szymanski	55
M	Muzyk	Ciechanow	Roman	Nowak	22

X: Age, City, FirstName, Surname, Job, Gender  
 Y: Age, City, FirstName, Surname, Job, Gender

Przywróć dane    Przywróć dane tymczasowe    Zapisz    Parametr K: 0    Anonimizuj

Rysunek 12 Zakładka do K-(X, Y)-anonimizacji z wykorzystaniem PID-a

Zakładka umożliwia zastosowanie algorytmu K-(X,Y)-anonimizacji. Użytkownik ma możliwość wprowadzenia wartości parametru K oraz wybór kolumn tworzących zbiory X i Y. Zbiory X i Y są wybierane analogicznie do wybierania kolumn tworzących PID i użytkownik ma możliwość wyboru wielu kolumn należących do danego zbioru. Anonimizacja następuje po wciśnięciu przycisku “Anonimizuj”. Użytkownik może również wybrać parametr K anonimizacji. Dostępne są również opcje “Zapisz”, “Przywróć dane” i “Przywróć dane tymczasowe”.

## 4.5. ( $\alpha$ , k)-anonimizacja

Anonimizator

K-anonimizacja

A-K Anonimizacja

Anonimizacja

Analiza

PID K-Anonimizacja

Wyliczanie parametru K

K-(X-Y) Anonimizacja

Generator danych

K-E Anonimizacja

Rozpoznawanie parametru K

Gender	Job	City	FirstName	Surname	Age
M	Inzynier	Wroclaw	Jan	Nowak	15
M	Inzynier	Katowice	Adam	Kowalski	11
M	Inzynier	Wroclaw	Kamil	Malinowski	15
M	Inzynier	Czestochowa	Jakub	Jagiel	15
M	Programista	Wroclaw	Szymon	Osowski	19
M	Programista	Wroclaw	Filip	Nowak	22
M	Malarz	Rybnik	Mikolaj	Kowalski	30
K	Malarz	Olawa	Julia	Wojcik	40
K	Tancerz	Olawa	Zofia	Kowalczyk	43
M	Muzyk	Wroclaw	Kacper	Szymanski	50
M	Muzyk	Wroclaw	Kacper	Szymanski	51
M	Muzyk	Wroclaw	Damian	Nowakowski	53
M	Muzyk	Swidnica	Dawid	Kowalski	60
M	Muzyk	Rybnik	Wiktor	Nowak	90
M	Programista	Siedlce	Szymon	Kowal	19
M	Programista	Glubczyce	Edward	Jagiel	45
M	Muzyk	Namyslow	Cezary	Wojcik	23
K	Programista	Wyszkow	Beata	Osowski	24
M	Programista	Lubin	Adam	Kowalski	11
M	Programista	Zabrze	Kamil	Malin	32
M	Muzyk	Olawa	Krystian	Nowak	37
K	Muzyk	Lubin	Joanna	Szymanski	55
M	Muzyk	Ciechanow	Roman	Nowak	22

Przywróć dane

Przywróć dane tymczasowe

Wartość atrybutu:

Parametr K:

0

Parametr alpha:

0

Anonimizuj

Zapisz

PID:

Age

City

FirstName

Surname

Job

Gender

Attribute:

Age

City

FirstName

Surname

Job

Gender

Rysunek 13 Zakładka do ( $\alpha$ , k)-anonimizacji

Zakładka umożliwia zastosowanie algorytmu ( $\alpha$ , k)-anonimizacji. W prawym panelu zakładki można wybrać kolumny należące do PID-a oraz kolumnę, której wartość podana w polu "Wartość atrybutu" będzie występować w danej kolumnie z prawdopodobieństwem nie większym niż parametr alpha. W dolnym panelu zakładki można wprowadzić wartości parametru K, parametru alpha i atrybutu. Anonimizacja następuje po wciśnięciu przycisku "Anonimizuj". Dostępne są również opcje "Zapisz", "Przywróć dane" i "Przywróć dane tymczasowe".

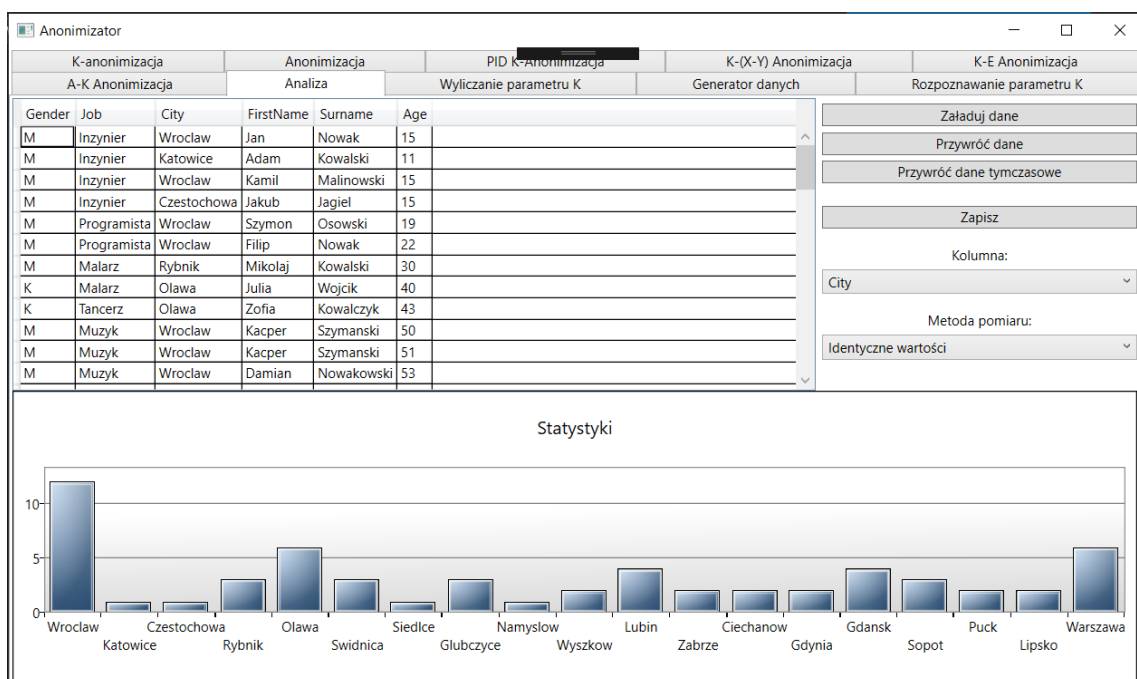
## 4.6. (K, E)-anonimizacja

Gender	Job	City	FirstName	Surname	Age
M	Inzynier	Wroclaw	Jan	Nowak	15
M	Inzynier	Katowice	Adam	Kowalski	11
M	Inzynier	Wroclaw	Kamil	Malinowski	15
M	Inzynier	Czestochowa	Jakub	Jagiel	15
M	Programista	Wroclaw	Szymon	Osowski	19
M	Programista	Wroclaw	Filip	Nowak	22
M	Malarz	Rybnik	Mikolaj	Kowalski	30
K	Malarz	Olawa	Julia	Wojcik	40
K	Tancerz	Olawa	Zofia	Kowalczyk	43
M	Muzyk	Wroclaw	Kacper	Szymanski	50
M	Muzyk	Wroclaw	Kacper	Szymanski	51
M	Muzyk	Wroclaw	Damian	Nowakowski	53
M	Muzyk	Swidnica	Dawid	Kowalski	60
M	Muzyk	Rybnik	Wiktor	Nowak	90
M	Programista	Siedlce	Szymon	Kowal	19
M	Programista	Glubczyce	Edward	Jagiel	45
M	Muzyk	Namyslow	Cezary	Wojcik	23
K	Programista	Wyszkow	Beata	Osowski	24
M	Programista	Lubin	Adam	Kowalski	11
M	Programista	Zabrze	Kamil	Malin	32
M	Muzyk	Olawa	Krystian	Nowak	37
K	Muzyk	Lubin	Joanna	Szymanski	55
M	Muzyk	Ciechanow	Roman	Nowak	22

Rysunek 14 Zakładka do (K, E)-anonimizacji

Zakładka umożliwia zastosowanie algorytmu (K, E)-anonimizacji. W prawym panelu zakładki można wybrać kolumny należące do PID-a. W dolnym panelu można wybrać wartości parametru K i parametru E. Anonimizacja następuje po wciśnięciu przycisku “Anonimizuj”. Dostępne są również opcje “Zapisz”, “Przywróć dane” i “Przywróć dane tymczasowe”.

## 4.7. Analiza



Rysunek 15 Analiza danych

Zakładka umożliwia analizę statystyk danych tabelarycznych. W prawym panelu można wybrać kolumnę, której ma dotyczyć wykres wyświetlający się w

dolnym panelu oraz metodę pomiaru. Do wyboru są dwie metody pomiaru: “Identyczne wartości” i “Wartości o równej długości”. Wybór parametrów analizy automatycznie przegenerowuje wykres. Dostępne są również opcje “Zapisz”, “Przywróć dane”, “Załaduj dane” i “Przywróć dane tymczasowe”. Opcja “Załaduj dane” umożliwia wczytanie danych z pliku csv do grida i przeprowadzenie na nich analiz.

Zakładka ta umożliwia wybranie na podstawie statystyk kolumn, które poprzez małe zróżnicowanie wartości są potencjalnymi kolumnami należącymi do PID-a.

## 4.8. Wyliczanie parametru K

Gender	Job	City	FirstName	Surname	Age
M	Inzynier	Wroclaw	Jan	Nowak	15
M	Inzynier	Katowice	Adam	Kowalski	11
M	Inzynier	Wroclaw	Kamil	Malinowski	15
M	Inzynier	Czestochowa	Jakub	Jagiel	15
M	Programista	Wroclaw	Szymon	Osowski	19
M	Programista	Wroclaw	Filip	Nowak	22
M	Malarz	Rybnik	Mikolaj	Kowalski	30
K	Malarz	Olawa	Julia	Wojcik	40
K	Tancerz	Olawa	Zofia	Kowalczyk	43
M	Muzyk	Wroclaw	Kacper	Szymanski	50
M	Muzyk	Wroclaw	Kacper	Szymanski	51
M	Muzyk	Wroclaw	Damian	Nowakowski	53
M	Muzyk	Swidnica	Dawid	Kowalski	60
M	Muzyk	Rybnik	Wiktor	Nowak	90
M	Programista	Siedlce	Szymon	Kowal	19
M	Programista	Glubczyce	Edward	Jagiel	45
M	Muzyk	Namyslow	Cezary	Wojcik	23
K	Programista	Wyszkow	Beata	Osowski	24
M	Programista	Lubin	Adam	Kowalski	11
M	Programista	Zabrze	Kamil	Malin	32
M	Muzyk	Olawa	Krystian	Nowak	37
K	Muzyk	Lubin	Joanna	Szymanski	55
M	Muzyk	Ciechanow	Roman	Nowak	22
K	Programista	Gdynia	Julia	Wojcik	35
M	Muzyk	Gdansk	Wojciech	Kowal	28
M	Programista	Sopot	Wojciech	Kowalski	17
M	Muzyk	Puck	Damian	Malinowski	26
K	Tancerz	Glubczyce	Joanna	Kowal	39
M	Tancerz	Lipso	Grzegorz	Malinowski	10
K	Programista	Warszawa	Zofia	Nowak	30
M	Straznik wiezienny	Gdansk	Daniel	Kowal	22
M	Policjant	Sopot	Jozef	Kowalski	23
M	Straznik wiezienny	Puck	Kamil	Malinowski	24
K	Historyk	Glubczyce	Kamila	Kowal	25

PID:

- Age
- City
- FirstName
- Surname
- Job
- Gender

Wylicz parametr K

1

Załaduj dane

Przywróć dane

Przywróć dane tymczasowe

Rysunek 16 Wyznaczanie parametru K

Zakładka umożliwia wyliczanie parametru K wprowadzonego zbioru danych. W prawym panelu zakładki można wybrać kolumny, które stworzą PID, dla którego zostanie wyliczony parametr K. Wyliczanie parametru K następuje po wciśnięciu przycisku “Wylicz parametr K”. Dostępne są również opcję “Załaduj dane”, “Przywróć dane” i “Przywróć dane tymczasowe”.

Opcja “Załaduj dane” umożliwia pobranie danych do programu z wybranego pliku csv z dysku.

Zakładka umożliwia użytkownikowi sprawdzenie wartości parametru K dla kolumn, które uważa za mogące stanowić PID zanonimizowanych danych.



## 4.9. Rozpoznawanie parametru K

Anonimizator

K-anonimizacja

A-K Anonimizacja

Anonimizacja

Analiza

PID K-Anonimizacja

Wyliczanie parametru K

K-(X-Y) Anonimizacja

Generator danych

K-E Anonimizacja

Rozpoznawanie parametru K

Gender	Job	City	FirstName	Surname	Age		K	Columns	
K	Pracownik	Polska	3 - 5 letters	5 letters	24 - 30		60	City	
K	Pracownik	Polska	6 - 8 letters	5 letters	24 - 30		60	Job	
K	Pracownik	Polska	6 - 8 letters	5 letters	31 - 90		60	City, Job	
K	Pracownik	Polska	6 - 8 letters	5 letters	31 - 90		27	FirstName	
K	Pracownik	Polska	6 - 8 letters	5 letters	31 - 90		27	City, FirstName	
K	Pracownik	Polska	6 - 8 letters	5 letters	31 - 90		27	FirstName, Job	
K	Pracownik	Polska	3 - 5 letters	6 - 7 letters	31 - 90		27	City, FirstName, Job	
K	Pracownik	Polska	3 - 5 letters	6 - 7 letters	31 - 90		16	Gender	
K	Pracownik	Polska	3 - 5 letters	6 - 7 letters	31 - 90		16	City, Gender	
K	Pracownik	Polska	3 - 5 letters	6 - 7 letters	31 - 90		16	Job, Gender	
K	Pracownik	Polska	3 - 5 letters	6 - 7 letters	24 - 30		16	City, Job, Gender	
K	Pracownik	Polska	3 - 5 letters	6 - 7 letters	31 - 90		14	Age	
K	Pracownik	Polska	3 - 5 letters	9 letters	31 - 90		14	Age, City	
K	Pracownik	Polska	3 - 5 letters	9 letters	31 - 90		14	Age, Job	
K	Pracownik	Polska	6 - 8 letters	9 letters	31 - 90		14	Age, City, Job	
K	Pracownik	Polska	6 - 8 letters	9 letters	31 - 90		9	Surname	
M	Pracownik	Polska	3 - 5 letters	10 letters	10 - 23		9	City, Surname	
M	Pracownik	Polska	6 - 8 letters	10 letters	10 - 23		9	Surname, Job	
M	Pracownik	Polska	6 - 8 letters	10 letters	10 - 23		9	City, Surname, Job	
M	Pracownik	Polska	3 - 5 letters	10 letters	24 - 30		7	FirstName, Gender	
M	Pracownik	Polska	6 - 8 letters	10 letters	24 - 30		7	City, FirstName, Gender	
M	Pracownik	Polska	6 - 8 letters	10 letters	24 - 30		7	FirstName, Job, Gender	
M	Pracownik	Polska	6 - 8 letters	10 letters	24 - 30		7	City, FirstName, Job, Gender	
M	Pracownik	Polska	3 - 5 letters	10 letters	31 - 90				

☒ Wyświetlaj K = 1

Przywróć dane

Przywróć dane tymczasowe

Wyznacz parametry K

Rysunek 17 Analiza parametru K

Zakładka służy do wyliczanie parametru K wprowadzonego zbioru danych dla wszystkich kombinacji kolumn. W prawym panelu zakładki można wybrać opcję wyświetlania kolumn, które tworzą PID, dla którego wyliczona wartość parametru K jest równa 1. Wyliczanie parametru K następuje po wciśnięciu przycisku “Wyznacz parametry “K”. Dostępne są również opcje “Przywróć dane” i “Przywróć dane tymczasowe”.

Zakładka wspomaga użytkownika w wyborze kolumn mogących należeć do PID-a zanonimizowanych danych.

## 4.10. Generator danych

Gender	Job	City	FirstName	Surname	Age
K	Programista	Sopot	Beata	Wozniak	58
K	Matematyk	Prudnik	Michalina	Wojcik	25
M	Malarz	Rybnik	Filip	Kowalski	56
K	Informatyk	Bytom	Barbara	Wozniak	55
M	Inzynier	Ostroleka	Mariusz	Wozniak	45
M	Straznik graniczny	Kluczbork	Marcin	Kowalski	43
K	Policjant	Glubczyce	Hanna	Gombrowicz	56
M	Muzyk	Wroclaw	Kacper	Mazur	52
K	Architekt	Gdynia	Iwona	Wojcik	61
M	Matematyk	Klodzko	Wojciech	Wojcik	36
K	Polonista	Namyslow	Magda	Abramczyk	15
K	Informatyk	Cieszyn	Danuta	Kowalczyk	24
M	Tancerz	Jelenia Gora	Mariusz	Szymanski	18
M	Informatyk	Pszczyna	Zbigniew	Wozniak	21
K	Programista	Jelenia Gora	Halina	Malinowski	22
M	Architekt	Swidnica	Jakub	Kowal	40
M	Programista	Olawa	Marcel	Malinowski	44
K	Strazak	Gdansk	Martyna	Kowal	41
M	Strazak	Olawa	Kamil	Szymanski	16
M	Informatyk	Brzeg	Damian	Wojcik	44

Rysunek 18 Generator danych losowych

Zakładka służąca do tworzenia losowego zbioru danych o zadanej wielkości. Po wprowadzeniu wielkości nowego zbioru danych i wciśnięciu przycisku “Generuj dane” zostaną utworzone nowe dane na podstawie wykorzystywanych w programie słowników. Dane zostaną zapisane jako dane tymczasowe i będzie je można wczytać w dowolnej zakładce poprzez wciśnięcie przycisku “Przywróć dane tymczasowe”. Zakładka wyświetla również podgląd wygenerowanych danych w formie grida.

## 5. Wnioski

W ramach projektu utworzono w pełni funkcjonalną aplikację pozwalającą anonimizować oraz analizować tabele danych. Przedstawione rozwiązanie pozwala na wiele rozszerzeń dzięki skupieniu się na modularności implementacji. Cały proces rozwoju projektu był przyrostowy. Rozpoczęto od anonimizacji wybranej kolumny, następnie algorytmów działających na całym pseudo identyfikatorze, kończąc na analizie zanonimizowanych danych. Podczas implementacji napotkano wiele problemów, które udało się w satysfakcjonujący sposób rozwiązać.

## 6. Bibliografia

- [1][https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Guide-to-Anonymisation\\_v1-\(250118\).pdf](https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Guide-to-Anonymisation_v1-(250118).pdf)
- [2]<http://biblioteka.pmwsz.opole.pl/download/attachment/11580/3-problemy-anonimizacji-dokumentow-medycznych-czesc-1.pdf>
- [3]<http://biblioteka.pmwsz.opole.pl/1205/arkadiusz-liber-problemy-anonimizacji-dokumentow-medycznych-czesc-2-anonimizacja-zaawansowana-oraz-sterowana-przez-posiadacza-danych-wrazliwych.html>
- [4][https://pl.wikipedia.org/wiki/Anonimizacja\\_danych](https://pl.wikipedia.org/wiki/Anonimizacja_danych)
- [5]<https://lexdigital.pl/co-to-jest-anonimizacja-i-pseudonimizacja>
- [6]<https://www.cs.sfu.ca/~wangk/pub/FWCY10csur.pdf>