Aggregation Pippeline

Aldona Biewska

10 maja 2017

Dane

- *Zgony w Stanach Zjednoczonych w 2014 roku*.
- Plik DeathRecords.csv obrobiony przy pomocy skryptu w R *prepare.R*.
- Wybrane pliki:

Plik	**Liczba dokumentów**	**Czas importu**
DeathRecords_prepare.csv	2631171	4 min 40 s
EntityAxisCondition.csv	8052877	7min 41s
Icd10Code.csv	12131	3s

Przykładowy dokument z kolekcji *deaths*

```
{
  "_id" : ObjectId("58fcdd07bd26e96c4d2d9129"),
  "Id" : 1,
  "Sex" : "M",
  "Age" : 87,
  "MonthOfDeath" : 1,
  "Icd10Code" : "I64",
  "AgeType" : "Years",
  "Education" : "9 - 12th grade, no diploma",
  "MaritalStatus" : "Married",
  "DayOfWeekOfDeath" : "Wednesday",
  "Race" : "White",
  "MannerOfDeath" : "Natural",
  "ActivityCode" : "Not applicable"
}
```

Przykładowy dokument z kolekcji *conditions* i *icd10code*

conditions - odzwierciedlenie struktury aktu zgonu:

```
{
  "_id" : ObjectId("58fcdebfbd26e96c4d55c198"),
  "Id" : 1,
  "DeathRecordId" : 1,
  "Part" : 1,
  "Line" : 1,
  "Sequence" : 1,
  "Icd10Code" : "I64"
}
```

icd10code:

```
{
  "_id" : ObjectId("58fce11abd26e96c4dd0c1d1"),
  "Code" : "A00",
  "Description" : "Cholera"
}
```

Informacje

Agregacje w JS dostępne są *tutaj*.
Dodatkowo wszystkie aggregacje zostały napisane i uruchomione w R, dzięki czemu łatwiej było wykonać wszystkie wykresy oraz tabele *mongo.R*

Agregacja 1: Porównanie samobójstw w danym przedziale wiekowym

Czas wykonania: 4s 370ms

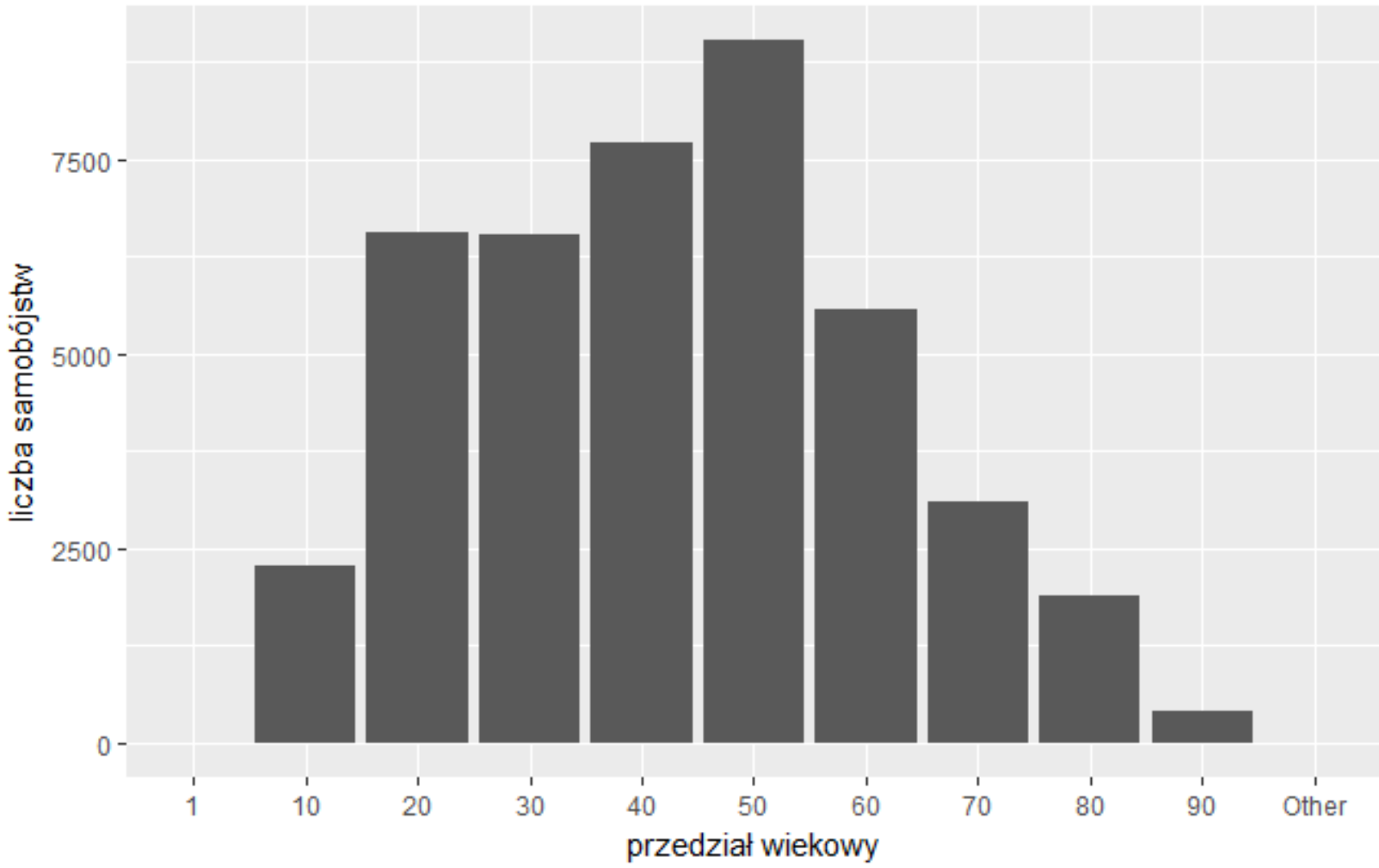
```
db.deaths.aggregate([
 {
   $match:{
    "MannerOfDeath":"Suicide",
    "AgeType":"Years"
   }
 },
 {
   $bucket:{
    groupBy:"$Age",
    boundaries:[1,10,20,30,
     40,50,60,70,80,90,100],
    default:"Other"
   }
 }
])
```

Wynik:

```
[
  { "_id":1,"count":3 },
  { "_id":10,"count":2270},
  { "_id":20,"count":6578},
  { "_id":30,"count":6526},
  { "_id":40,"count":7712},
  { "_id":50,"count":9032},
  { "_id":60,"count":5588},
  { "_id":70,"count":3115},
  { "_id":80,"count":1898},
  { "_id":90,"count":403},
  { "_id":"Other","count":7}
]
```

Przy pomocy *$bucket* i *$boundaries* określane są przedziały, np.: dla 1 przedziałem jest [1,9).

Agregacja 1: Porównanie samobójstw w danym przedziale wiekowym - Wykres



Agregacja 2: Wpływ edukacji i stanu cywilnego na żywotność kobiet i mężczyzn - krok 1

Wybranie typu wieku (w tym przypadku lata) oraz przedziału od 35 do 100.

```
var stage1 =  {
 $match: {
   "AgeType": "Years",
   "Education":{$ne: "NA"},
   $and: [{ "Age" : { $gt : 35 }}, { "Age" : { $lt : 100 }}]
 }
};
```

Agregacja 2: Wpływ edukacji i stanu cywilnego na żywotność kobiet i mężczyzn - krok 2

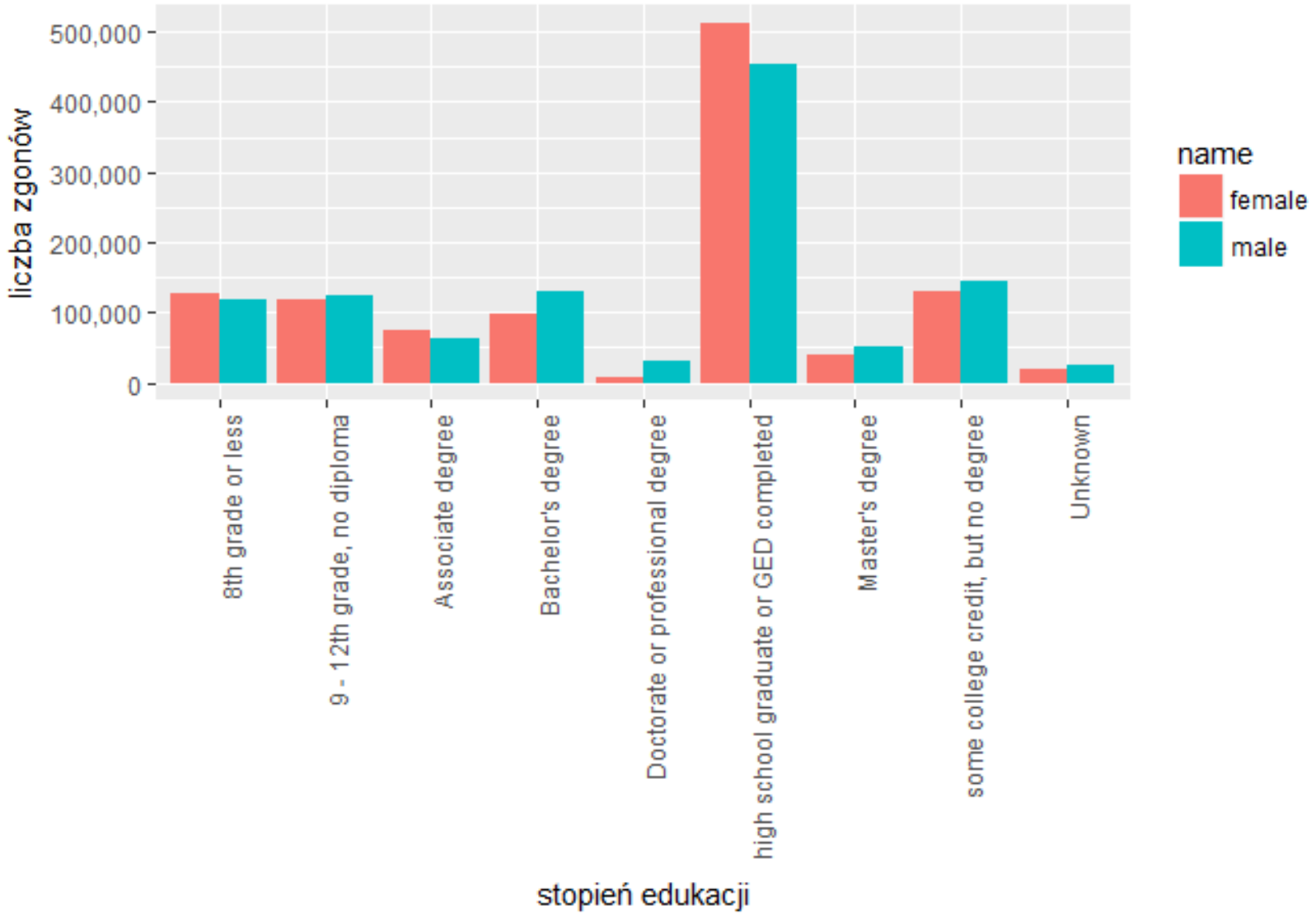
```
var stage2 =  {
 $facet: {
  Education: [
   { $group: {
       _id:{sex: "$Sex",edu: "$Education"},
       count: {$sum: 1}}},
   { $sort: {count: -1} },
   { $group: {
       _id: "$_id.sex",
       education: {
         $push:{range: "$_id.edu" , total:"$count"}}}}
  ],
  Marriage: [
   { $group: {
       _id:{sex: "$Sex", status: "$MaritalStatus" },
       count: {$sum: 1}} },
   { $sort: {count: -1} },
   { $group: {
       _id: "$_id.sex",
       marriage: {
         $push: {status: "$_id.status" , total:"$count"}}}}
  ]
 }
};
```

Agregacja 2: Wpływ edukacji i stanu cywilnego na żywotność kobiet i mężczyzn

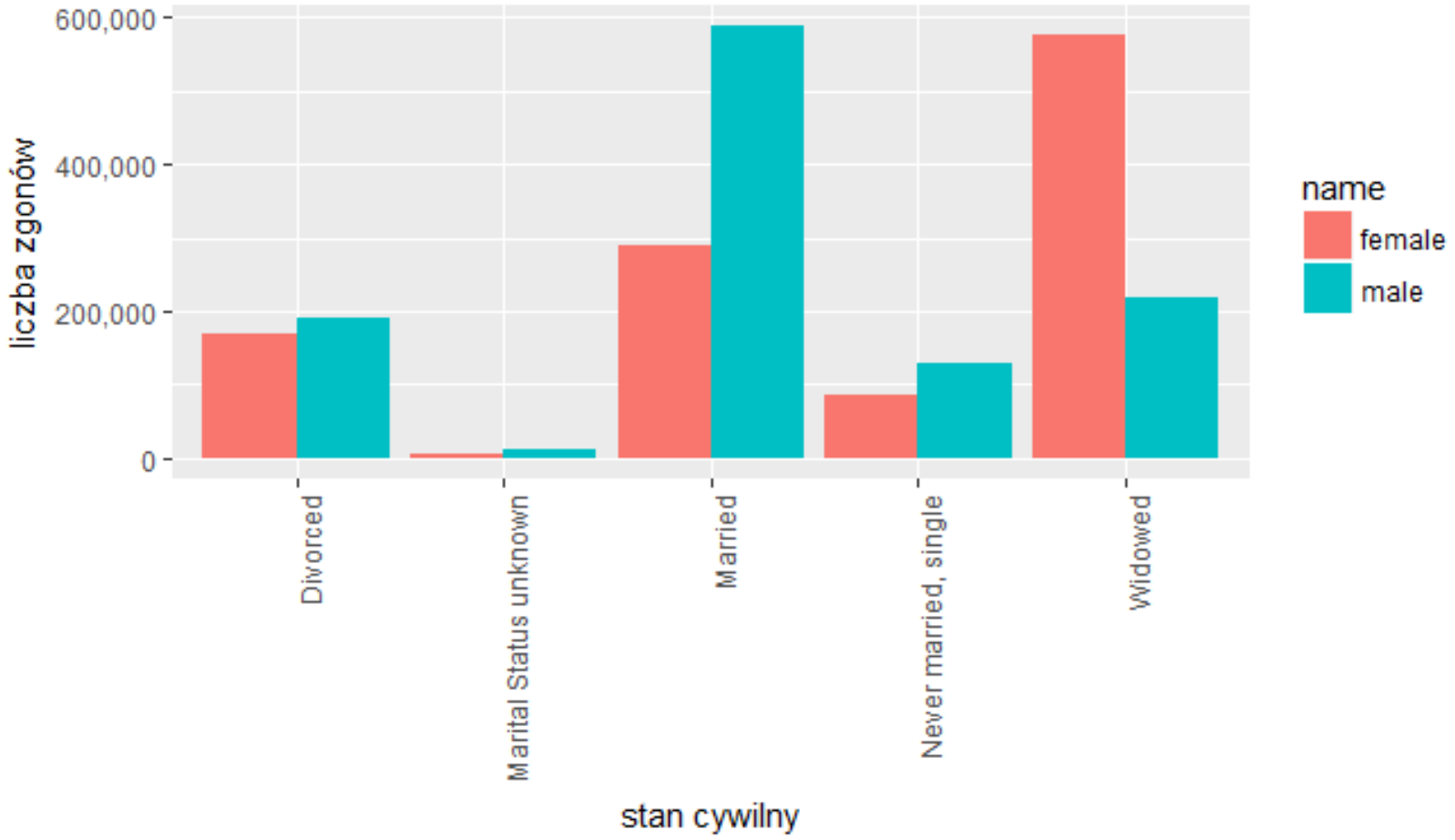
Czas wykonania: 11s 389ms Całość:

```
db.deaths.aggregate( [
 stage1,
 stage2
]);
```

Agregacja 2: Wpływ edukacji i stanu cywilnego na żywotność kobiet i mężczyzn - Wykres



Agregacja 2: Wpływ edukacji i stanu cywilnego na żywotność kobiet i mężczyzn - Wykres



Agregacja 3: Porównanie ilościowe i procentowe zgonów w wyniku zabójstwa w wybranych rasach - agregacja pomocnicza krok 1

Zliczono liczbę wszystkich zgonów według wybranych ras, przy czym określone rasy szczegółowe przypisano do rasy *Yelllow*. Rasa *Black* i *White* była już określona w kolekcji. Rasy dodawane są do nowego pola *Race_new*.

```
var stage1 =  {
 $project: {
   "Id": 1,
   "MannerOfDeath": 1,
   "Race_new": {
    $switch: {
      branches: [
       { "case": { $eq: ["$Race","White"]},then: "White"},
       { "case": { $eq: ["$Race","Black"]},then: "Black"},
       { "case": { $or: [
         { $eq: ["$Race","Chinese"]},
         { $eq: ["$Race","Japanese"]},
         { $eq: ["$Race","Asian Indian"]},
         { $eq: ["$Race","Korean"]},
         { $eq: ["$Race","Other Asian or Pacific Islander"]},
         { $eq: ["$Race","American Indian (...)"]},
         { $eq: ["$Race","Vietnamese"]},
         { $eq: ["$Race","Guamanian"]}]},
        then: "Yellow"}],
      default: "Did not match"}}}};
```

Agregacja 3: Porównanie ilościowe i procentowe zgonów w wyniku zabójstwa w wybranych rasach - agregacja pomocnicza krok 2,3,4

Grupowanie zgonów według ras, odpowiednie wyświetlenie wyników:

```
var stage2={
 $match: { "Race_new": {$ne: "Did not match"}}
};
var stage3={
 $group: { "_id": { "race": "$Race_new"},"count": {$sum: 1}}
};
var stage4={
 $project: { "_id": 0,"race": "$_id.race","count": $count}
};
```

Agregacja 3: Porównanie ilościowe i procentowe zgonów w wyniku zabójstwa w wybranych rasach - agregacja pomocnicza

Całość:

```
var count =
 db.deaths.aggregate([
  stage1,
  stage2,
  stage3,
  stage4
]);
```

Otrzymany wynik:

```
{ "race" : "Yellow",
  "count" : 56205 }
{ "race" : "Black",
  "count" : 309504 }
{ "race" : "White",
  "count" : 2241510 }
```

Przypisanie wyników zmiennym:

```
var white,yellow,black;
count.forEach(function(record) {
 if (record.race=="White") white=record.count;
 if (record.race=="Black") black=record.count;
 if (record.race=="Yellow") yellow=record.count;
});
```

Agregacja 3: Porównanie ilościowe i procentowe zgonów w wyniku zabójstwa w wybranych rasach - agregacja właściwa krok 1,2,3

Krok 1 taki jak w agregacji pomocniczej.
Krok 2,3 wybieranie rodzaju śmierci jako zabójstwo i grupowanie według ras.

```
var stage2_2 = {
 $match: {
  "MannerOfDeath":"Homicide",
  "Race_new":{$ne: "Did not match"} }
};
```

```
var stage3_1 = {
 $group: {_id:{race: "$Race_new"}, count: {$sum: 1} }
};
```

Agregacja 3: Porównanie ilościowe i procentowe zgonów w wyniku zabójstwa w wybranych rasach - agregacja właściwa krok 4

Obliczenie jaki procent stanowią zgony w wyniku zabójstwa wobec wszystkich zgonów:

```
var stage4_2 = {
 $project: {
  "percentage of all deaths":
   {$let: {
     vars: { "white": white, "black":black,"yellow":yellow },
     in: { $switch: {
            branches: [
       {case: { $eq: [ "$_id.race","White" ] },
           then: {
            "$multiply":[{"$divide":[100,"$$white"]},"$count"]}},
       {case: { $eq: [ "$_id.race","Black" ] },
           then: {
            "$multiply":[{"$divide":[100,"$$black"]},"$count"]}},
       {case: { $eq: [ "$_id.race","Yellow" ] },
           then: {
            "$multiply":[{"$divide":[100,"$$yellow"]},"$count"]}},
       ],
       default: "0"}}}},
  "_id":0,"count":"$count","race":"$_id.race"}};
```

Agregacja 3: Porównanie ilościowe i procentowe zgonów w wyniku zabójstwa w wybranych rasach - Wyniki

Czas wykonania: 3s 602ms Całość:

```
db.deaths.aggregate([
 stage1,
 stage2_2,
 stage3_2,
 stage4_2
]);
```

percentage homicide - all deaths	count	race
0.800640512409928	450	Yellow
2.6277527915632755	8133	Black
0.3606051277933179	8083	White

Agregacja 4: Najczęstsze czynniki pośrednie przyczyniające się do głównej przyczyny śmierci

Główna przyczyna: I469 - Cardiac arrest, cause unspecified (Zatrzymanie krążenia)
Dla każdego zgonu tworzę listę przyczyn (2-3); następnie wyszukuję rekordy, w których na liście jako główna przyczyna znajduje się kod *I469* (5-6); wybieram pośrednie przyczyny (8-10) i grupuję (11); łącze z tabelą *icd10*, aby wyodrębnić nazwy chorób (14-18); odpowiednio wyświetlam (19-25).

```
1    db.conditions.aggregate([
2     { $group: {
3        _id : "$DeathRecordId",
4        Other_conditions: { $push: "$$ROOT" }}},
5     { $match: {
6        "Other_conditions":{"$elemMatch":{"Part":1,"Line":1,"Icd10Code":"I469"}}}},
7     { $unwind : "$Other_conditions" },
8     { $match: {
9        "Other_conditions.Icd10Code":{$ne:"I469"},
10       "Other_conditions.Part":2}},
11    { $group: {_id: "$Other_conditions.Icd10Code", count: {$sum: 1}}},
12    { $sort: {count: -1}},
13    { $limit : 5},
14    { $lookup: {
15       from:"icd10",
16       localField:"_id",
17       foreignField:"Code",
18       as:"Icd10Description"}},
19    { $unwind : "$Icd10Description" },
20    { $project : {
21       "_id":0,
22       "code":"$_id",
23       "description":"$Icd10Description.Description",
24       "count_of_cases":"$count"}}
25   ],{allowDiskUse: true});
```

Agregacja 4: Najczęstsze czynniki pośrednie przyczyniające się do głównej przyczyny śmierci - Wyniki po angielsku

code	description	count
I10	Essential (primary) hypertension	28771
F179	Mental and behavioural disorders due to use of tobacco: Unspecified mental and behavioural disorder	25154
E149	Unspecified diabetes mellitus: Without complications	14413
J449	Chronic obstructive pulmonary disease, unspecified	11588
F03	Unspecified dementia	11165

Agregacja 4: Najczęstsze czynniki pośrednie przyczyniające się do głównej przyczyny śmierci - Wyniki po polsku

code	description	count
I10	Nadciśnienie samoistne (pierwotne)	28771
F179	Zaburzenia psychiczne i zaburzenia zachowania spowodowane paleniem tytoniu (zaburzenia psychiczne i zaburzenia zachowania, nieokreślone)	25154
E149	Cukrzyca nieokreślona (bez powikłań)	14413
J449	Nieokreślona przewlekła zaporowa choroba płuc	11588
F03	Otępienie bliżej nieokreślone	11165