Aggregation Pippeline

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 każdej 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"
}
```

conditions:

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

icd10:

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

Agregacje

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"
    }
 }
])
```

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

Otrzymany 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}
]
```

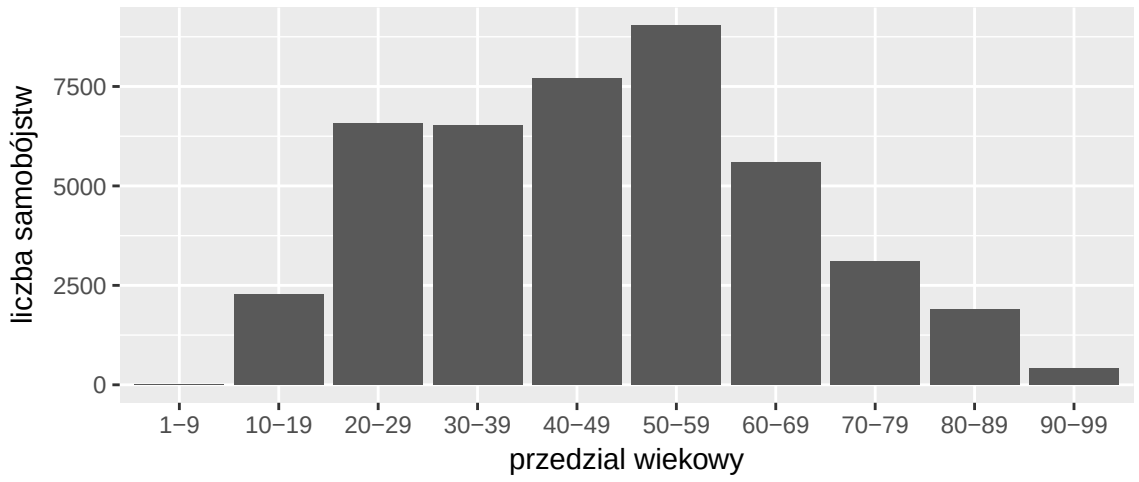


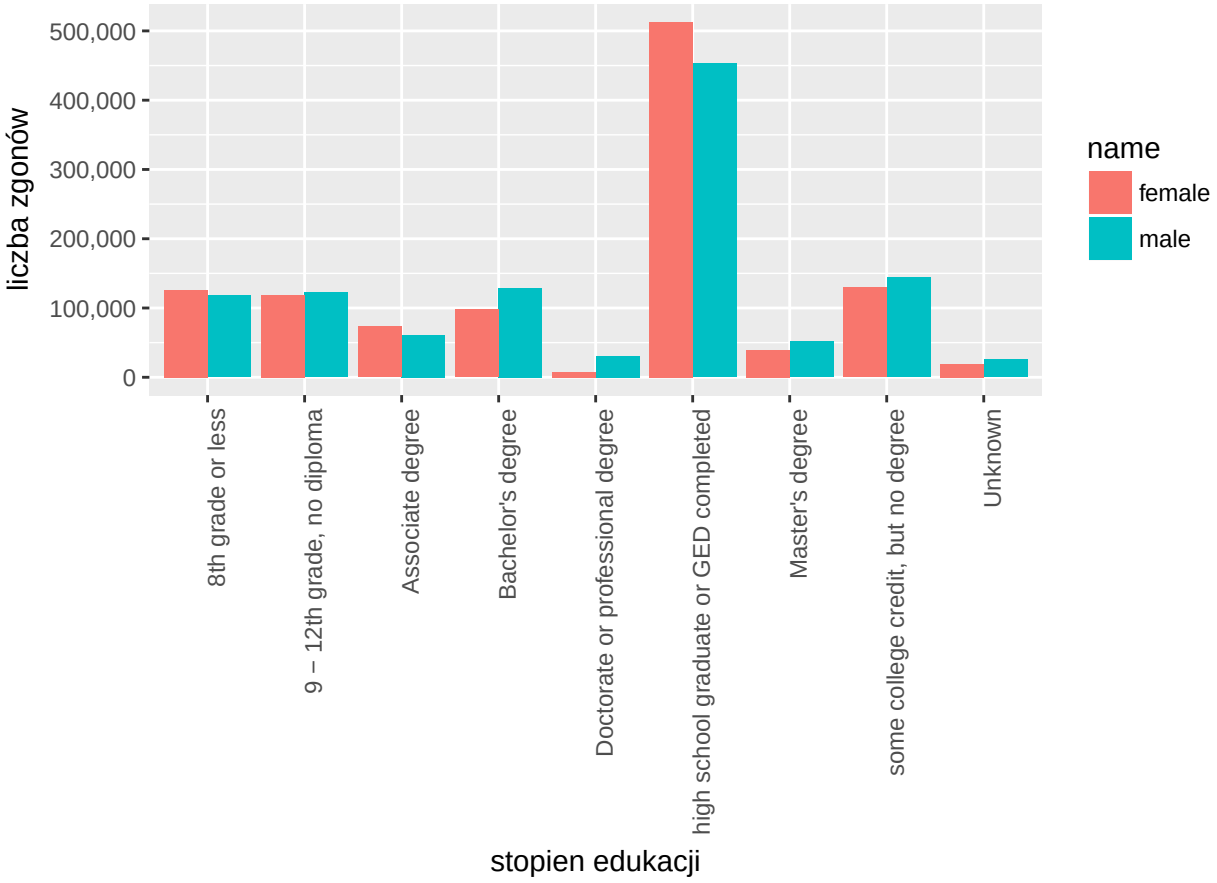
Diagram wykonany przy pomocy skryptu w R *plot1.R*.

2. Wpływ edukacji i stanu cywilnego na żywotność kobiet i mężczyzn.

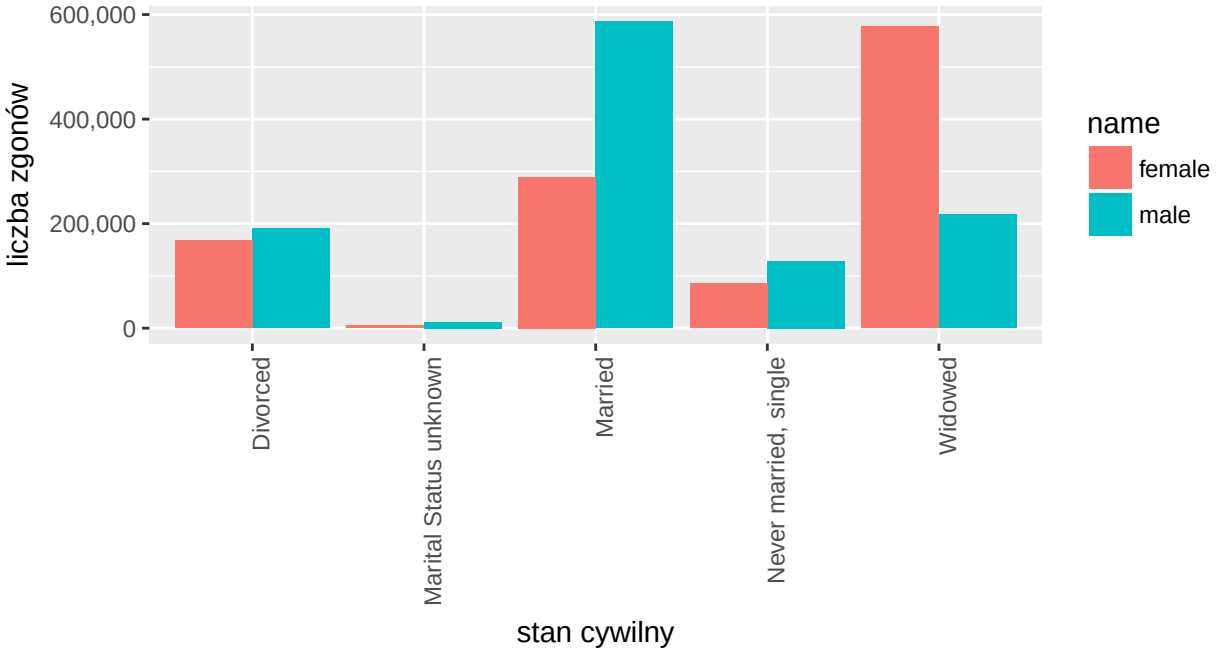
Czas wykonania: 11s 389ms

```
db.deaths.aggregate([
 {
    $match:{
       "AgeType":"Years",
       "Education":{ $ne:"NA" },
       $and:[
          {"Age":{$gt:35}},
          {"Age":{$lt:100}}
       ]
    }
 },
 {
    $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"} }
          }
        }
      ]
    }
 }
]);
```

Porównanie zgonów według płci i edukacji:



Porównanie zgonów według płci i stany cywilnego:



Diagramy wykonano przy pomocy skryptów w R: *plot2.R*, *plot3.R*

3. Najczęstsze czynniki pośrednie przyczyniające się do głownej przyczyny śmierci.

Czas wykonania: 1min 200s

```
db.conditions.createIndex( { DeathRecordId: 1 } )
```

Do aggregacji jako główną przyczynę śmierci wybrano: **I469 - Cardiac arrest, cause unspecified (Zatrzymanie krążenia, nieokreślone)**.

```
db.conditions.aggregate([
 {
    $group:{
      _id:"$DeathRecordId",
      Other_conditions:{ $push:"$$ROOT" }
    }
 },
 {
    $match:{
      "Other_conditions":{
         "$elemMatch":{ "Part":1,"Line":1,"Icd10Code":"I469"}
      }
    }
 },
 {  $unwind:"$Other_conditions" },
 {
    $match:{
      "Other_conditions.Icd10Code":{ $ne:"I469" },
      "Other_conditions.Part":2
    }
 },
 {
    $group:{
      _id:"$Other_conditions.Icd10Code",
      count:{$sum:1}
    }
 },
 {  $sort:{count:-1 } },
 {  $limit:10 },
 {
    $lookup:{
      from:"icd10",
      localField:"_id",
      foreignField:"Code",
      as:"Icd10Description"
    }
 },
 {  $unwind:"$Icd10Description"},
 {
    $project:{
      "_id":0,
      "code":"$_id",
      "description":"$Icd10Description.Description",
      "count_of_cases":"$count"
    }
 }
],{ allowDiskUse: true } )
```

Główna przyczyna śmierci: **I469 - Cardiac arrest, cause unspecified (Zatrzymanie krążenia, nieokreślone)**

Poniżej tabele przedstawiające pośrednie przyczyny śmierci związane z główną przyczyną, kolejno w języku angielskim i polskim:

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
I48	Atrial fibrillation and flutter	10858
I251	Atherosclerotic heart disease	9503
I500	Congestive heart failure	8252
E119	Non-insulin-dependent diabetes mellitus: Without complications	6896
N189	Chronic kidney disease, unspecified	6761

code	description_PL	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
I48	Migotanie i trzepotanie przedsionków	10858
I251	Choroba serca w przebiegu miażdżycy	9503
I500	Niewydolność serca zastoinowa	8252
E119	Cukrzyca insulinoniezależna (bez powikłań)	6896
N189	Przewlekła niewydolność nerek, nieokreślona	6761

4. Porównanie ilościowe i procentowe zgonów w wyniku zabójstwa, pogrupowane według wybranych ras.

Najpierw zliczono liczbę wszystkich zgonów według wybranych ras, przy czym określone rasy szczegółowe zgrupowano do rasy *Yelllow*. Rasa *Black* i *White* była już określona w kolekcji.

```
db.deaths.aggregate([
 {
    $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 (includes Aleuts and Eskimos)" ] },
                        { $eq: [ "$Race", "Vietnamese" ] },
                        { $eq: [ "$Race", "Guamanian" ] },
                    ] }, then: "Yellow" }
          ],
              default: "Did not match"
       }
     }
   }
 },
 { $match: { "Race_new":{$ne: "Did not match" } } },
 { $group: { _id:{race: "$Race_new"}, count: {$sum: 1} } },
 { $project: { "_id":0,"race":"$_id.race","count":"$count"  } }
]);
```

Otrzymany wynik:

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

Następnie właściwa agregacja, uwzględniająca jaki procent stanowi zgon w wyniku zabójstwa wobec wszystkich zgonów danej rasy:

Czas wykonania: 3s 602ms

```
db.deaths.aggregate([
 {
    $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 (includes Aleuts and Eskimos)" ] },
                        { $eq: [ "$Race", "Vietnamese" ] },
                        { $eq: [ "$Race", "Guamanian" ] },
                    ] }, then: "Yellow" }
          ],
              default: "Did not match"
       }
     }
   }
 },
 {  $match: {"MannerOfDeath":"Homicide","Race_new":{$ne: "Did not match"} } },
 {  $group: {_id:{race: "$Race_new"}, count: {$sum: 1} } },
 {
    $project: { "percentage":
    {
      $switch: {
        branches: [
           { case: { $eq: [ "$_id.race","White" ] },
                     then: {"$multiply":[{"$divide":[100,2241510]},"$count"]} },
           { case: { $eq: [ "$_id.race","Black" ] },
                     then: {"$multiply":[{"$divide":[100,309504]},"$count"]} },
           { case: { $eq: [ "$_id.race","Yellow" ] },
                     then: {"$multiply":[{"$divide":[100,56205]},"$count"]} },
       ],
      default: "0"
     }
    },"_id":0,"count":"$count","race":"$_id.race"
   }
  }
]);
```

Wynik:

percentage	count	race
0.800640512409928	450	Yellow
2.6277527915632755	8133	Black
0.3606051277933179	8083	White

Agregacje w JS dostępne są tutaj. Dodatkowo dwie pierwsze aggregacje zostały napisane również w Pythonie.

Skrypty w Pythonie wymagają drivera mongodb: pymongo.

Uśrednione czasy dwóch pierwszych agregacji:

	JS	Python
Agregacja 1	4s 370ms	5s 22ms
Agregacja 2	11s 389ms	12s 41ms

Agregacje zostały zapisane do pliku .json:

```
mongo --quiet agg[1-4].js > result[1-4].json
```

Zmiana .json na .csv za pomocą polecenia:

```
type result[1-4].json | json2csv -f nazwy_kolumn -o result[1-4].csv
```

Otrzymane pliki .json i .csv tutaj.