

ΑΝΑΛΥΤΙΚΗ ΔΕΔΟΜΕΝΩΝ ΜΕΣΩ WEKA ΚΑΙ SQL

- Data Mining software
- Query Language



ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

UNIVERSITY OF PIRAEUS

Γεράσιμος Ραζής (razis@uth.gr)





WEKA

Συλλογή από εργαλεία οπτικοποίησης και αλγορίθμους για ανάλυση δεδομένων και προγνωστική μοντελοποίηση

Εισαγωγή

- WEKA: **W**aikato **E**nvironment for **K**nowledge **A**nalysis
- Το WEKA είναι ένα λογισμικό για εξόρυξη δεδομένων (γραμμένο σε Java) το οποίο περιέχει υλοποιημένες μεθόδους για:
 - Προεπεξεργασία Δεδομένων
 - Ταξινόμηση
 - Συσταδοποίηση
 - Εύρεση Κανόνων Συσχέτισης
- Διαθέσιμο για εγκατάσταση από την ιστοσελίδα:
<http://www.cs.waikato.ac.nz/ml/weka/>

Περιβάλλον WEKA

- Δίνεται η δυνατότητα επιλογής ενός συνόλου δεδομένων για εφαρμογή τεχνικών αναφορικά με:
 - Preprocessing
 - Classification
 - Clustering
 - Association
 - Attributes selection
 - Visualization

Περιβάλλον WEKA

- Επιλέγοντας ένα σύνολο δεδομένων εμφανίζονται:
 - γραφικά τα δεδομένα για καθένα από τα γνωρίσματα
 - στατιστικές πληροφορίες
 - οι κλάσεις των δεδομένων (αν υπάρχουν)
 - τα δεδομένα που ανήκουν στην ίδια κλάση εμφανίζονται με το ίδιο χρώμα

Περιβάλλον WEKA

Weka 3.5.7 - Explorer

Program Applications Tools Visualization Windows Help

Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter

Choose None Apply

Current relation

Relation: iris
Instances: 150
Attributes: 5

Attributes

All None Invert Pattern

No.	Name
1	<input type="checkbox"/> sepallength
2	<input type="checkbox"/> sepalwidth
3	<input type="checkbox"/> petallength
4	<input type="checkbox"/> petalwidth
5	<input type="checkbox"/> class

Remove

Selected attribute

Name: sepallength
Missing: 0 (0%)
Distinct: 35
Type: Numeric
Unique: 9 (6%)

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Class: class (Nom) Visualize All

The histogram displays the distribution of the 'sepallength' attribute for three classes: blue, red, and cyan. The x-axis represents the value of sepallength, ranging from 4.3 to 7.9. The y-axis represents the count of instances for each class. The blue class has 16 instances, the red class has 30 instances, and the cyan class has 34 instances. The distribution shows that the cyan class has the highest frequency of instances, followed by the red class, and then the blue class.

Class	Count
Blue	16
Red	30
Cyan	34

Αρχεία .arff

- Τα αρχεία που περιέχουν το σύνολο δεδομένων πρέπει να έχουν συγκεκριμένο format με την επέκταση *.arff*
- Στον φάκελο *C:\Program Files\Weka-3-5\data* περιέχονται κάποια παραδείγματα τέτοιων αρχείων
- Δεδομένα μπορούν επίσης να δοθούν από ένα URL ή από μία SQL βάση δεδομένων

Παράδειγμα

@relation heart-disease-simplified

@attribute age numeric

@attribute sex { female, male }

@attribute chest_pain_type { typ_angina, asympt, non_anginal, atyp_angina }

@attribute cholesterol numeric

@attribute exercise_induced_angina { no, yes }

@attribute class { present, not_present }

@data

63,male,typ_angina,233,no,not_present

67,male,asympt,286,yes,present

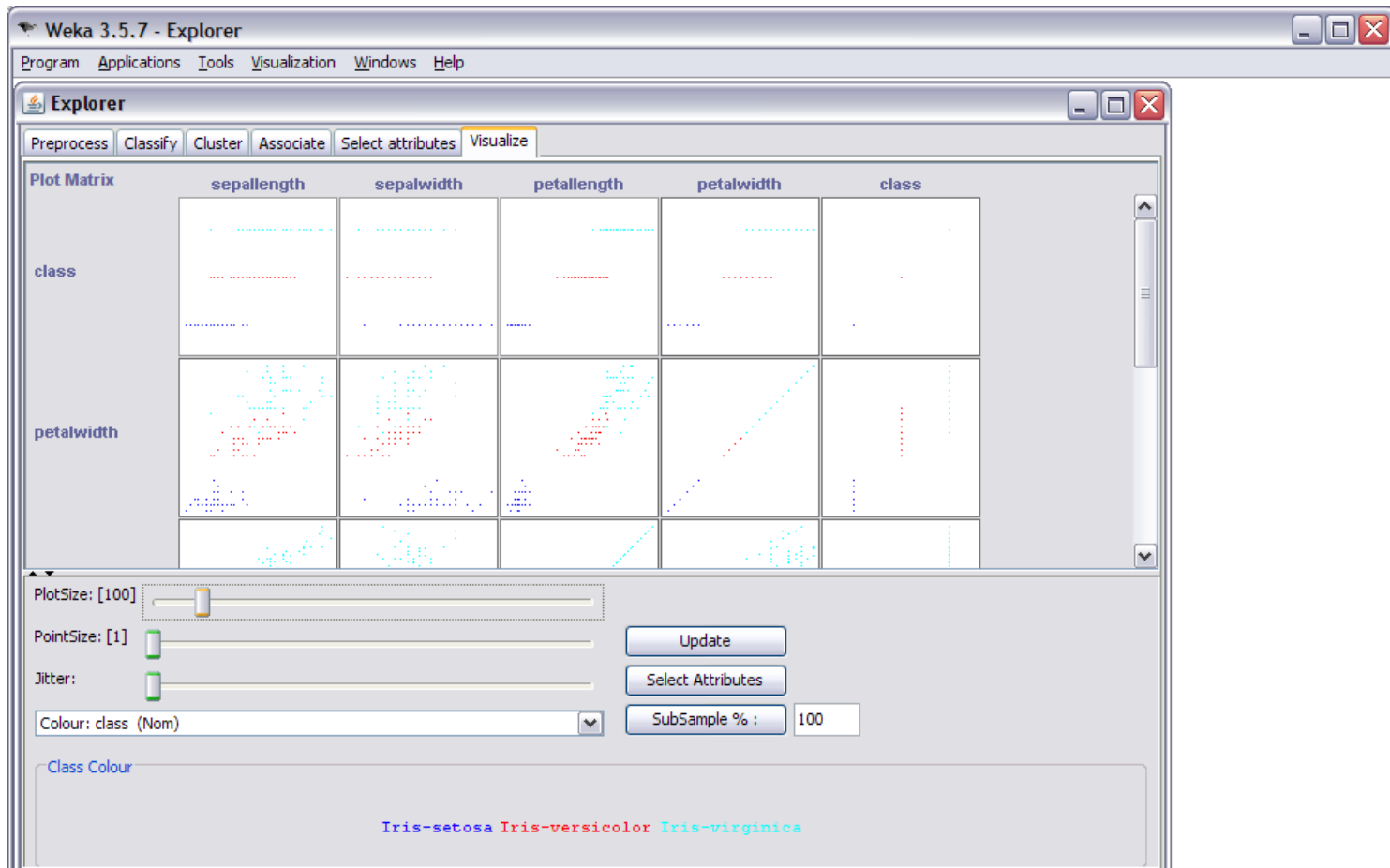
67,male,asympt,229,yes,present

38,female,non_anginal,?,no,not_present

...

Οπτικοποίηση Δεδομένων

Από την καρτέλα *Visualize* υπάρχει η δυνατότητα να εμφανιστεί η γραφική αναπαράσταση κάθε γνωρίσματος σε συνάρτηση με κάθε άλλο γνώρισμα



Συσταδοποίηση Δεδομένων

- Είναι δυνατόν να γίνει συσταδοποίηση σε ένα σύνολο δεδομένων
 - εύρεση ομάδων «όμοιων» δεδομένων
- Μπορεί να επιλεγεί ο αλγόριθμος συσταδοποίησης

Επιλογή Αλγορίθμου

- Οι διαθέσιμοι αλγόριθμοι συσταδοποίησης
 - Cobweb (ιεραρχική συσταδοποίηση)
 - DBSCAN
 - EM
 - Farthest First
 - OPTICS
 - SimpleKmeans (k-means)
 - Xmeans

Επιλογή Αλγορίθμου

The screenshot shows the Weka 3.5.7 Explorer interface. The 'Cluster' tab is active, and the 'SimpleKMeans -N 2 -S 10' algorithm is selected. The 'Cluster mode' section has 'Use training set' selected. The 'Clusterer output' window displays the following text:

```
kMeans
=====
Number of iterations: 7
Within cluster sum of squared errors: 62.1436882815797

Cluster centroids:

Cluster 0
Mean/Mode:  6.262  2.872  4.906  1.676  Iris-versicol
Std Devs:   0.6628 0.3328 0.8256 0.4248 N/A
Cluster 1
Mean/Mode:  5.006  3.418  1.464  0.244  Iris-setosa
Std Devs:   0.3525 0.381  0.1735 0.1072 N/A

Clustered Instances

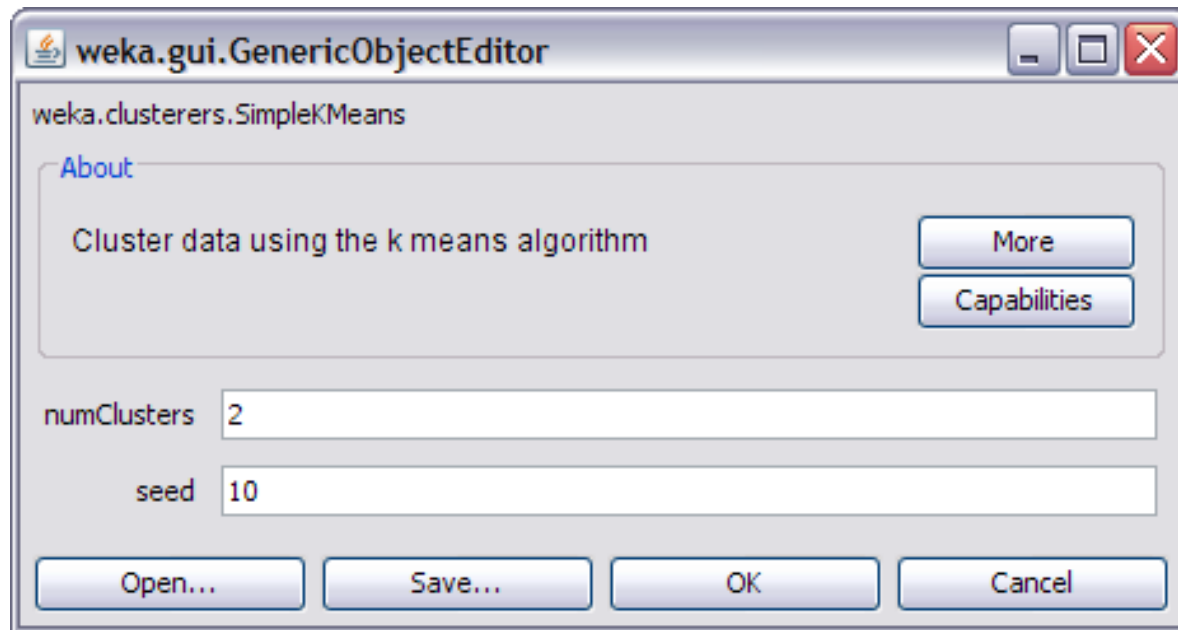
0    100 ( 67%)
1     50 ( 33%)
```

Two red arrows point to the 'SimpleKMeans' selection and the 'Std Devs' values in the output.

Πληροφορίες σχετικά με τα αποτελέσματα του clustering στα δεδομένα

Παράμετροι

- Οι τιμές των παραμέτρων συσταδοποίησης μπορούν να τροποποιηθούν
 - αριθμός clusters στον k-means
 - Eps και MinPts στον DBSCAN



Οπτικοποίηση Αποτελέσματος

Γραφική αναπαράσταση των δεδομένων με βάση τις ομάδες που έχουν προκύψει από το clustering

Weka 3.5.7 - Explorer

Program Applications Tools Visualization Windows Help

Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Clusterer

Choose SimpleKMeans -N 3 -S 10

Cluster mode

- Use training set
- Supplied test set Set...
- Percentage split % 66
- Classes to clusters evaluation (Nom) class
- Store clusters for visualization

Ignore attributes

Start Stop

Result list (right-click for options)

14:48:43 - SimpleKMeans

Clusterer output

Number of iterations: 3
Within cluster sum of squared errors: 7.817456892309574

Cluster centroids:

Cluster 0
Mean/Mode: 5.936 2.77 4.26 1.326 Iris-versicol
Std Devs: 0.5162 0.3138 0.4699 0.1978 N/A

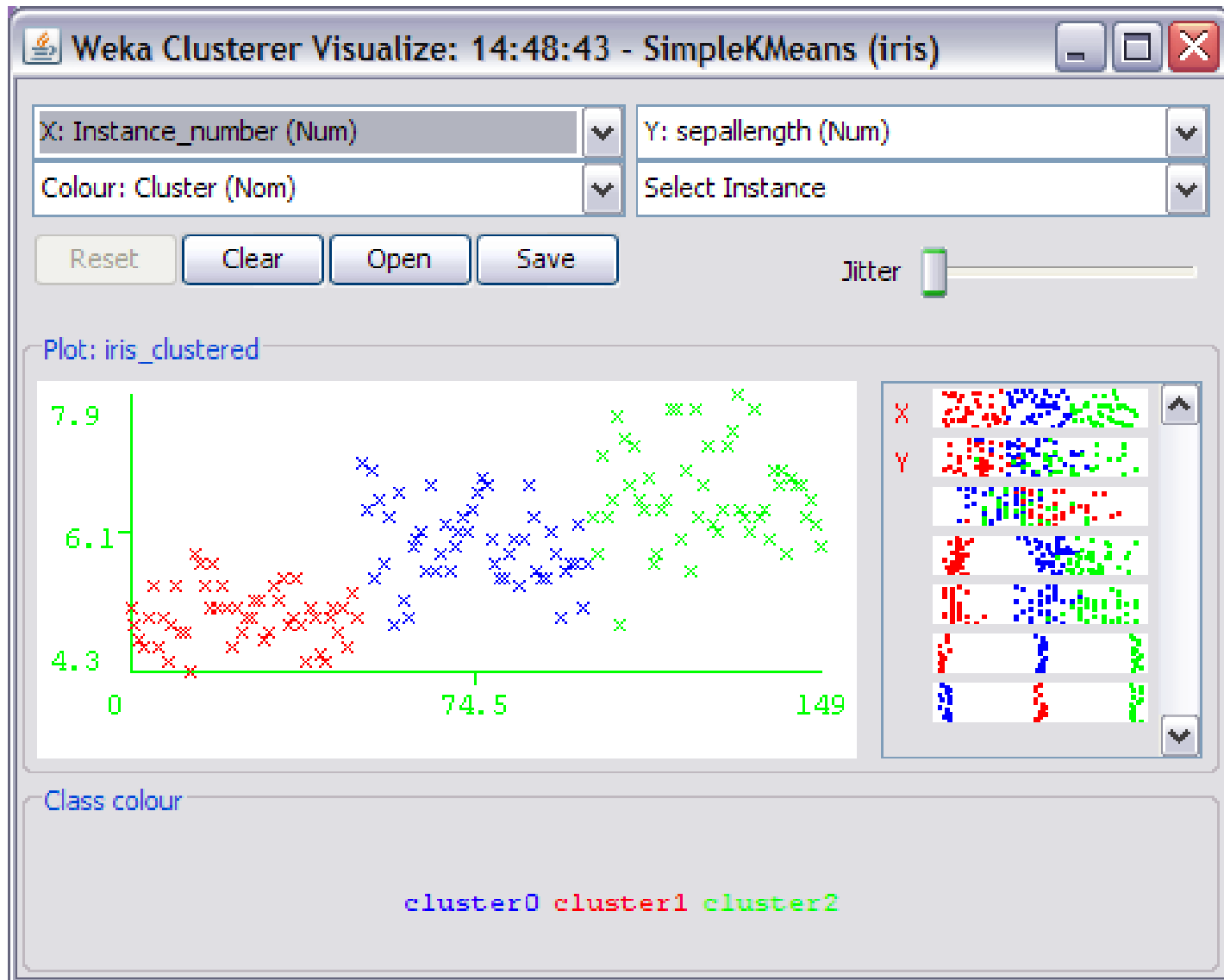
Cluster 1
Mean/Mode: 5.006 3.418 1.464 0.244 Iris-setosa
Std Devs: 0.3525 0.381 0.1735 0.1072 N/A

Cluster 2
Mean/Mode: 6.588 2.974 5.552 2.026 Iris-virginic
Std Devs: 0.6359 0.3225 0.5519 0.2747 N/A

Clustered Instances

0	50 (33%)
1	50 (33%)
2	50 (33%)

Οπτικοποίηση Αποτελέσματος





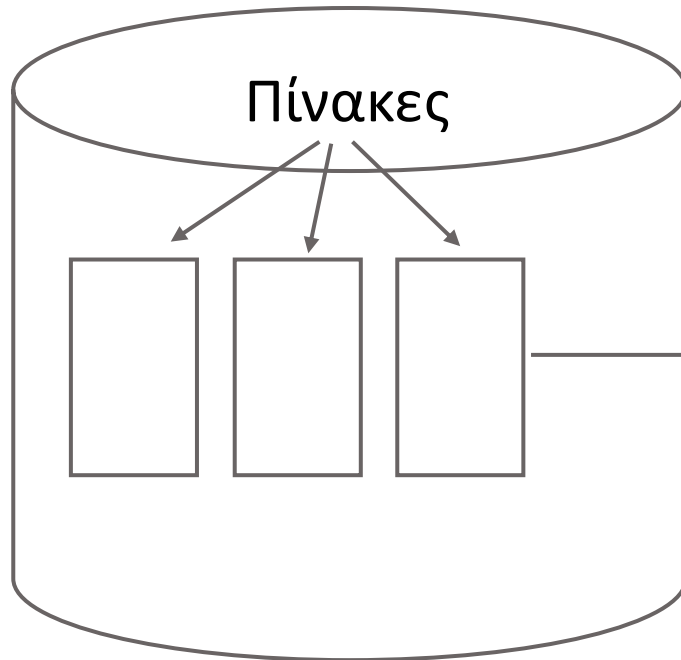
SQL

Η SQL μπορεί να δημιουργήσει πολύπλοκα μοντέλα και αναλύσεις δεδομένων γρήγορα και αποδοτικά

Εισαγωγή στις Βάσεις Δεδομένων

Στις σχεσιακές ΒΔ τα δεδομένα αποθηκεύονται σε πίνακες

Αναπαράσταση μοντέλου σχεσιακής ΒΔ (RDBM)



Βάση Δεδομένων

Στήλες

	A	B	C	D	E
1	Model	mpg	cyl	disp	hp
2	Mazda RX4	21.0	6	160.0	110
3	Mazda RX4 Wag	21.0	6	160.0	110
4	Datsun 710	22.8	4	108.0	93
5	Hornet 4 Drive	21.4	6	258.0	110
6	Hornet Sportabout	18.7	8	360.0	175
7	Valiant	18.1	6	225.0	105
8	Duster 360	14.3	8	360.0	245
9	Merc 240D	24.4	4	146.7	62
10	Merc 230	22.8	4	140.8	95
11	Merc 280	19.2	6	167.6	123
12	Merc 280C	17.8	6	167.6	123
13	Merc 450SE	16.4	8	275.8	180
14	Merc 450SL	17.3	8	275.8	180
15	Merc 450SLC	15.2	8	275.8	180
16	Cadillac Fleetwood	10.4	8	472.0	205
17	Lincoln Continental	10.4	8	460.0	215
18	Chrysler Imperial	14.7	8	440.0	230
19	Fiat 128	32.4	4	78.7	66
20	Honda Civic	30.4	4	75.7	52
21	Toyota Corolla	33.9	4	71.1	65
22	Toyota Corona	21.5	4	120.1	97

Εγγραφές

Πίνακας

Το dataset του ναυαγίου του Τιτανικού

Dataset: 12 στήλες, 891 εγγραφές

Table: titanic_passengers

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter
1	1	No	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.25	NULL	S
2	2	Yes	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.0	1	0	PC 17599	71.2833	C85	C
3	3	Yes	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 31...	7.925	NULL	S
4	4	Yes	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1	C123	S
5	5	No	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.05	NULL	S
6	6	No	3	Moran, Mr. James	male	NULL	0	0	330877	8.4583	NULL	Q
7	7	No	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
8	8	No	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.075	NULL	S
9	9	Yes	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NULL	S
10	10	Yes	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NULL	C
11	11	Yes	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549	16.7	G6	S
12	12	Yes	1	Bonnell, Miss. Elizabeth	female	58.0	0	0	113783	26.55	C103	S
13	13	No	3	Saunderscock, Mr. William Henry	male	20.0	0	0	A/5. 2151	8.05	NULL	S
14	14	No	3	Andersson, Mr. Anders Johan	male	39.0	1	5	347082	31.275	NULL	S
15	15	No	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14.0	0	0	350406	7.8542	NULL	S

Αυτό είναι μόνο ένα δείγμα (~1,6%) του συνόλου των επιβατών

Δημιουργία ιστοριών δεδομένων

DATA



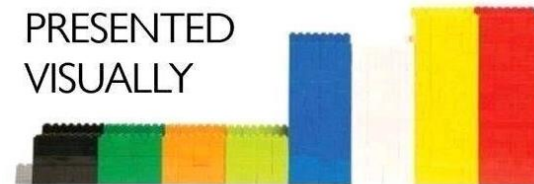
SORTED



ARRANGED



PRESENTED
VISUALLY



EXPLAINED
WITH A STORY



Δημιουργία ιστοριών δεδομένων

DATA STORYTELLING

FOR EFFECTIVE DATA DRIVEN DECISIONS.

Narrative Approach

Using Narrative approach for telling the story behind million rows of data.

Understanding Data

Data Storytellers are honed to develop a deep understanding data.

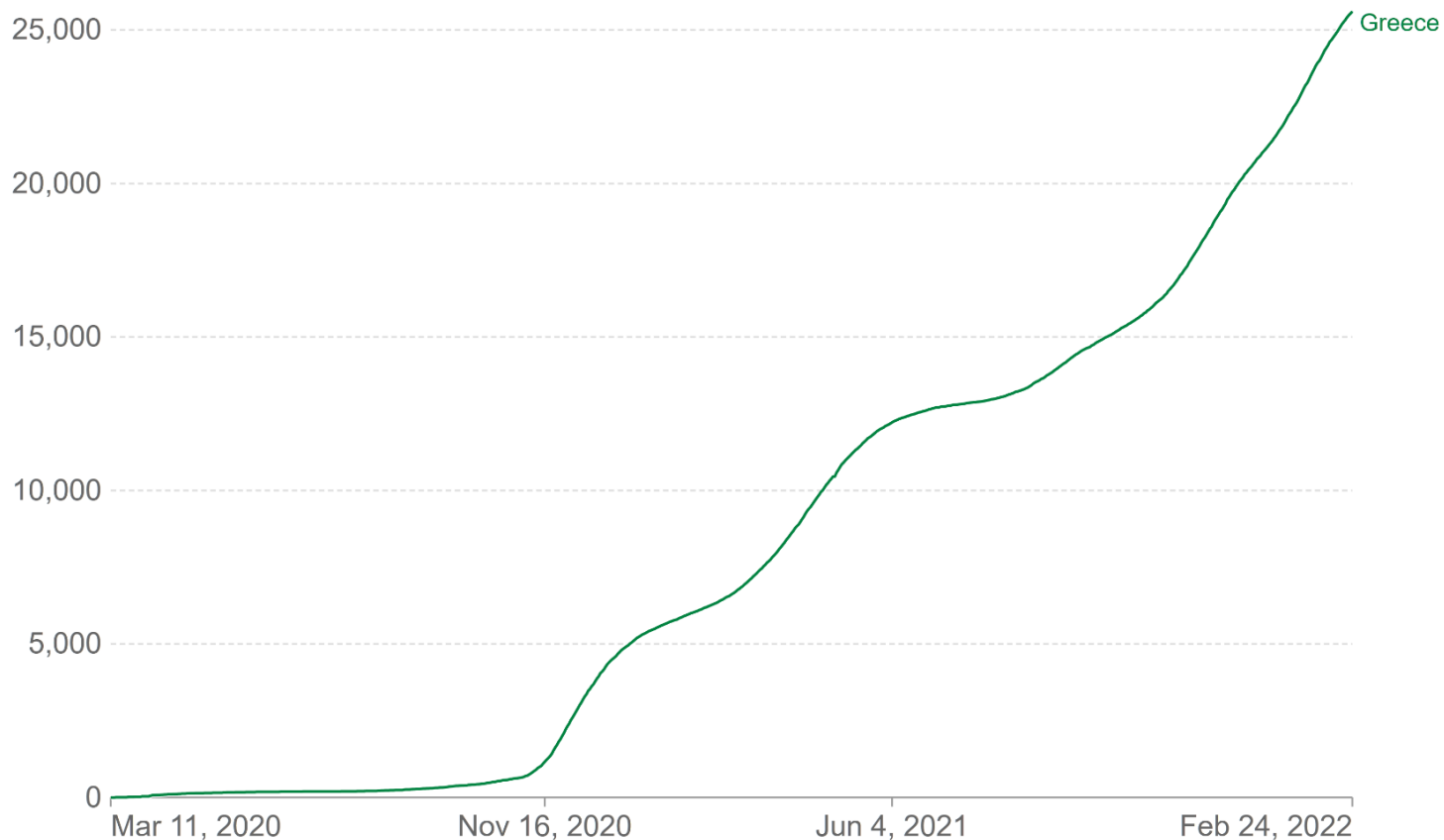
Effective Visualization

Using effective Visualizations to represent your data.

Δημιουργία ιστοριών δεδομένων

Cumulative confirmed COVID-19 deaths

For some countries the number of confirmed deaths is much lower than the true number of deaths. This is because of limited testing and challenges in the attribution of the cause of death.



Source: Johns Hopkins University CSSE COVID-19 Data

CC BY

Δημιουργία ιστοριών δεδομένων

Cumulative confirmed COVID-19 deaths

Our World
in Data

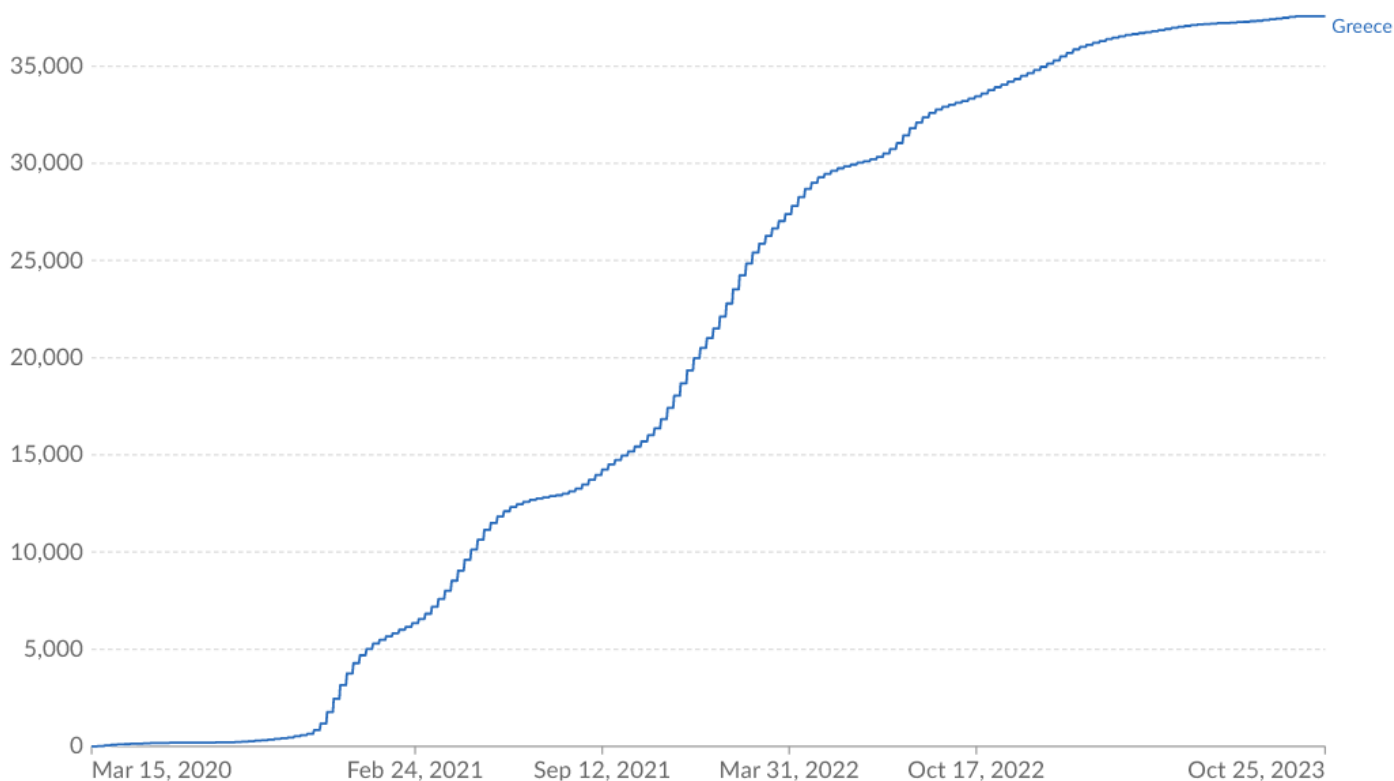
Due to varying protocols and challenges in the attribution of the cause of death, the number of confirmed deaths may not accurately represent the true number of deaths caused by COVID-19.

Table

Map

Chart

Settings



Mar 15, 2020



Oct 25, 2023

Δημιουργία ιστοριών δεδομένων

Daily new confirmed COVID-19 deaths per million people

7-day rolling average. Due to varying protocols and challenges in the attribution of the cause of death, the number of confirmed deaths may not accurately represent the true number of deaths caused by COVID-19.

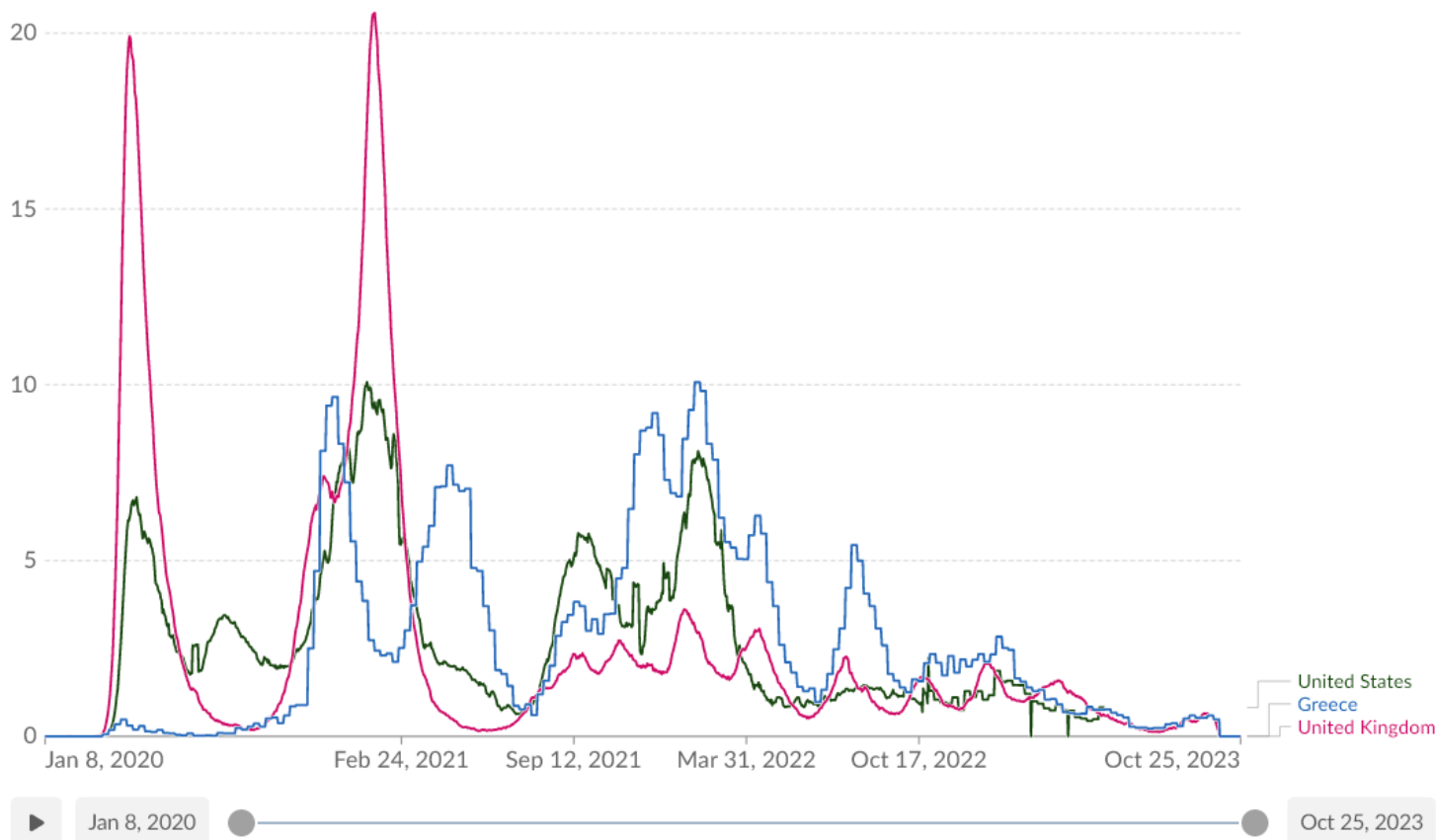
Our World
in Data

Table

Map

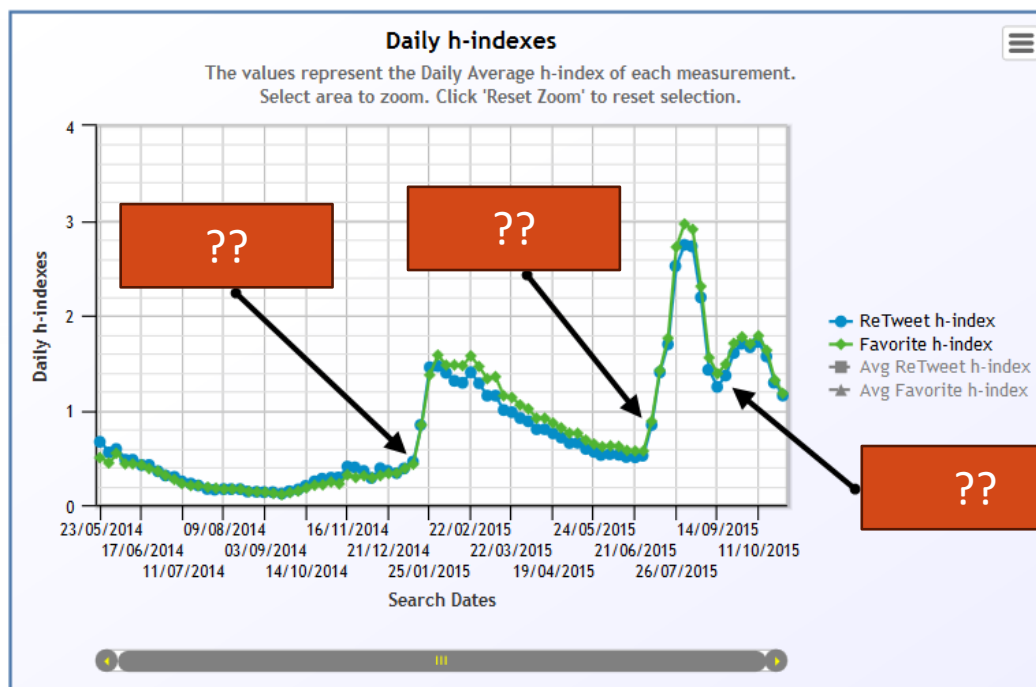
Chart

Settings



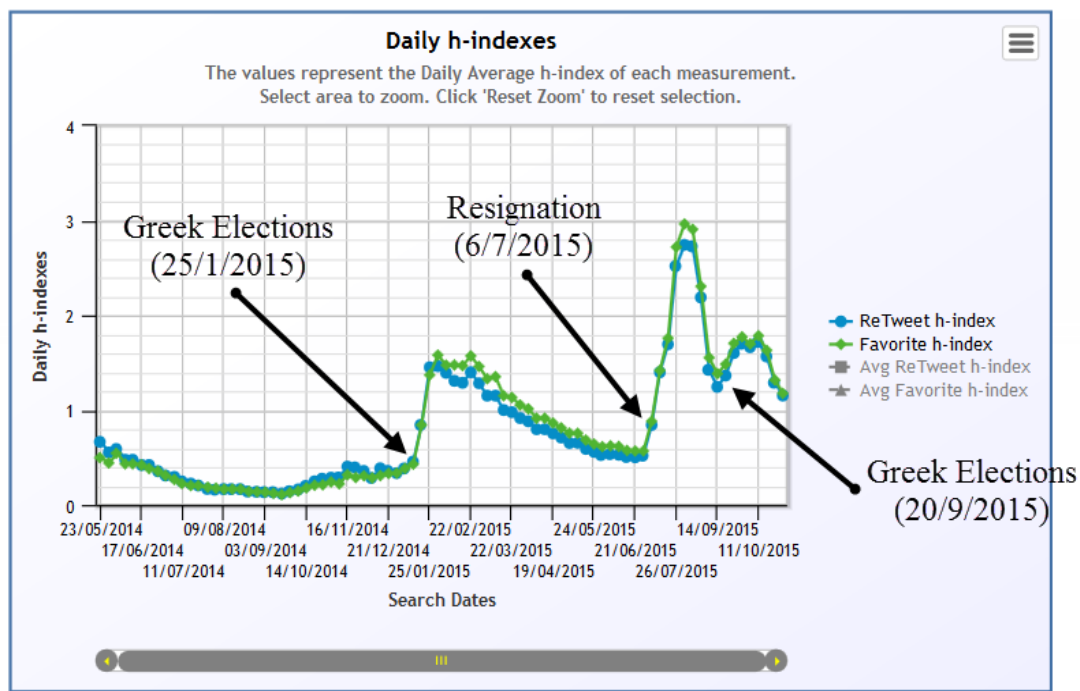
Δημιουργία ιστοριών δεδομένων

History of Account @yanisvaroufakis



Δημιουργία ιστοριών δεδομένων

History of Account @yanisvaroufakis



Ναυάγιο του Τιτανικού: Ιστορίες δεδομένων από την καταστροφή

Table: titanic_passengers

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter
1	1	No	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.25	NULL	S
2	2	Yes	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.0	1	0	PC 17599	71.2833	C85	C
3	3	Yes	3	Heikinen, Miss. Laina	female	26.0	0	0	STON/O2. 31...	7.925	NULL	S
4	4	Yes	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1	C123	S
5	5	No	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.05	NULL	S
6	6	No	3	Moran, Mr. James	male	NULL	0	0	330877	8.4583	NULL	Q
7	7	No	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
8	8	No	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.075	NULL	S
9	9	Yes	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NULL	S
10	10	Yes	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NULL	C
11	11	Yes	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549	16.7	G6	S
12	12	Yes	1	Bonnell, Miss. Elizabeth	female	58.0	0	0	113783	26.55	C103	S
13	13	No	3	Saunderscock, Mr. William Henry	male	20.0	0	0	A/5. 2151	8.05	NULL	S
14	14	No	3	Andersson, Mr. Anders Johan	male	39.0	1	5	347082	31.275	NULL	S
15	15	No	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14.0	0	0	350406	7.8542	NULL	S

Ποιες κατηγορίες επιβατών ήταν στον Τιτανικό;

```
SELECT DISTINCT Pclass AS unique_classes FROM titanic_passengers
```

	unique_classes
1	2
2	3
3	1

Από πόσα διαφορετικά μέρη επιβιβάστηκαν οι επιβάτες;

```
SELECT DISTINCT Embarked AS unique_places FROM titanic_passengers
```

	unique_places
1	S
2	C
3	Q
4	NULL

C = Cherbourg;
Q = Queenstown;
S = Southampton;

Περισσότερες ερωτήσεις που πρέπει να απαντηθούν...

- Πόσοι επέζησαν και πόσοι όχι;
- Πώς σχετίζεται η επιβίωση με το φύλο και την ηλικία των επιβατών;

```
SELECT COUNT(PassengerId) AS survived_passengers  
FROM titanic_passengers  
WHERE Survived = 'Yes'
```

	survived_passengers
1	342

```
SELECT COUNT(PassengerId) AS non_survived_passengers  
FROM titanic_passengers  
WHERE Survived = 'No'
```

	non_survived_passengers
1	549

Περισσότερα φίλτρα στο φύλο και την ηλικία

```
SELECT COUNT(PassengerId) AS survived_passengers_male  
FROM titanic_passengers  
WHERE Survived = 'Yes' AND Sex = 'male' AND  
(Age>16 OR Age IS NULL)
```

	survived_passengers_male
	87

```
SELECT COUNT(PassengerId) AS survived_passengers_female  
FROM titanic_passengers  
WHERE Survived = 'Yes' AND Sex = 'female' AND  
(Age>16 OR Age IS NULL)
```

	survived_passengers_female
	200

```
SELECT COUNT(PassengerId) AS survived_children  
FROM titanic_passengers  
WHERE Survived = 'Yes' AND Age<=16
```

	survived_children
	55

```
SELECT COUNT(PassengerId) AS  
non_survived_passengers_male FROM titanic_passengers  
WHERE Survived = 'No' AND Sex = 'male' AND  
(Age>16 OR Age IS NULL)
```

	non_survived_passengers_male
	439

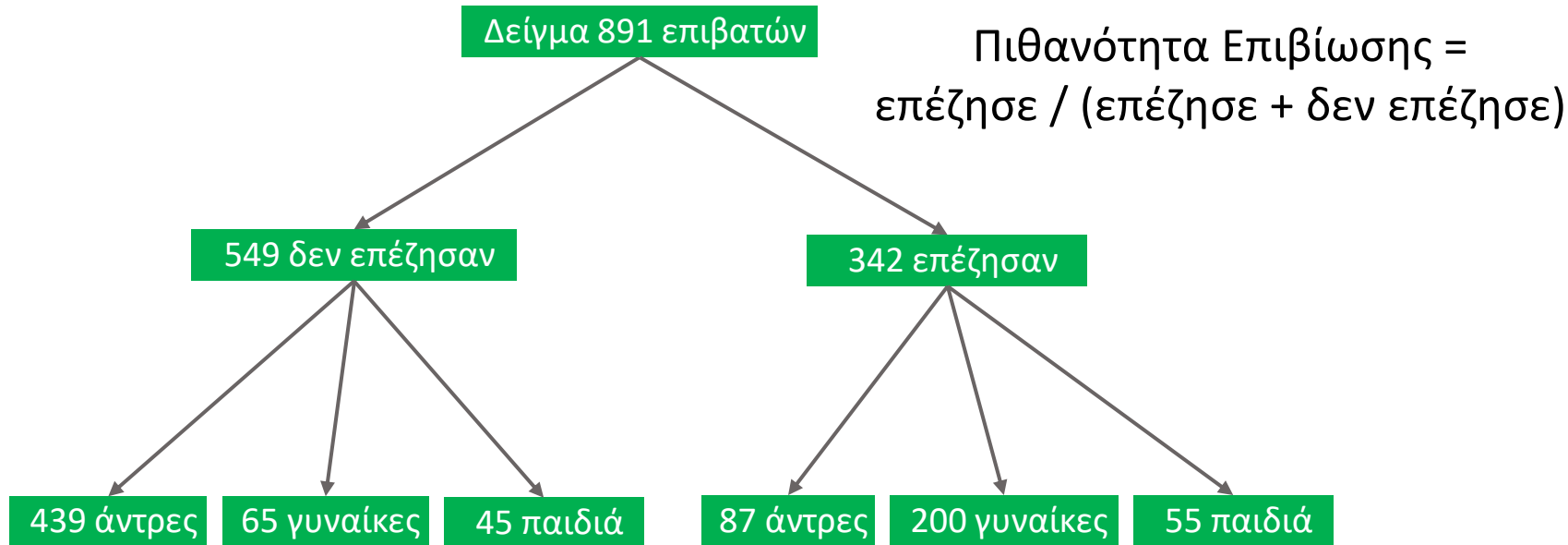
```
SELECT COUNT(PassengerId) AS  
non_survived_passengers_female FROM titanic_passengers  
WHERE Survived = 'No' AND Sex = 'female' AND  
(Age>16 OR Age IS NULL)
```

	non_survived_passengers_female
	65

```
SELECT COUNT(PassengerId) AS non_survived_children  
FROM titanic_passengers  
WHERE Survived = 'No' AND Age<=16
```

	non_survived_children
	45

Υπολογίζοντας την πιθανότητα επιβίωσης



- Πιθανότητα επιβίωσης ανδρών: $87/(87+439) = 0,165 \rightarrow 16,5\%$
- Πιθανότητα επιβίωσης γυναικών: $200/(200+65) = 0,754 \rightarrow 75,4\%$
- Πιθανότητα επιβίωσης παιδιών: $55/(55+45) = 0,55 \rightarrow 55\%$

Σχετίζοντας το ποσοστό επιβίωσης με την κατηγορία επιβατών

Αριθμός επιζώντων / μη επιζώντων ανά κατηγορία επιβατών και φύλο:

Άνδρες επιζώντες

```
SELECT Pclass AS passenger_class, COUNT(PassengerId) AS passengers
FROM titanic_passengers
WHERE Survived = 'Yes' AND Sex = 'male' AND (Age>16 OR Age IS NULL)
GROUP BY Pclass
```

passenger_class	passengers
1	42
2	8
3	37

Άνδρες μη επιζώντες

passenger_class	passengers
1	77
2	89
3	273

Γυναίκες επιζήσασες

```
SELECT Pclass AS passenger_class, COUNT(PassengerId) AS passengers
FROM titanic_passengers
WHERE Survived = 'Yes' AND Sex = 'female' AND (Age>16 OR Age IS NULL)
GROUP BY Pclass
```

passenger_class	passengers
1	86
2	60
3	54

Γυναίκες μη επιζήσασες

passenger_class	passengers
1	2
2	6
3	57

- Πιθανότητα επιβίωσης γυναικών στην Α' θέση: $86/(86+2) \rightarrow 97,7\%$
- Πιθανότητα επιβίωσης ανδρών στην Γ' θέση: $37/(37+273) \rightarrow 11,9\%$
- Πιθανότητα επιβίωσης στην Γ' θέση: $37+54/(37+273+54+57) \rightarrow 21,6\%$

Συμπεράσματα

- Οι περισσότερες γυναίκες Α' θέσης επέζησαν (97,7%)
- Τα παιδιά είχαν ένα μέσο ποσοστό επιβίωσης (55%)
- Οι άνδρες είχαν πολύ υψηλότερη σχετική πιθανότητα θανάτου από τις γυναίκες (+58,9%)
- Οι άνδρες της Γ' θέσης είχαν πολύ μεγάλη πιθανότητα θανάτου (88,1%)

Η αντίληψη της «τυχειότητας» εξαρτάται από το επίπεδο παρατήρησης



Όσο περισσότερο εμβαθύνουμε σε ένα πρόβλημα / κατάσταση / παρατήρηση, τόσο λιγότερο θεωρούμε τα αποτελέσματα τυχαία