

**ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ**

**ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ**

**ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ**

**ΠΜΣ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΚΑΙ ΥΠΗΡΕΣΙΕΣ**

**ΚΑΤΕΥΘΥΝΣΗ: ΠΡΟΗΓΜΕΝΑ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ**

**Εργασία Μαθήματος:**

**Αποθήκες Δεδομένων και Επιχειρηματική Ευφυία**

**Σχεδίαση και υλοποίηση αποθήκης δεδομένων**

**Δημήτριος Τσέλιος Βασίλειος Ζώης**

**ΜΕ 2059 ΜΕ 2044**

**Επιβλέπουσα Καθηγήτρια:**

Μαρία Χαλκίδη, Αναπληρώτρια Καθηγήτρια

|  |  |
| --- | --- |
|  |  |

**ΠΕΙΡΑΙΑΣ**

**Ιούνιος 2021**

**ΠΕΡΙΛΗΨΗ**

Η παρούσα εργασία έχει σαν στόχο τη δημιουργία μίας αποθήκης δεδομένων (ΑΔ) και ειδικότερα την κατασκευή κύβων από μεγάλες βάσεις δεδομένων (ΒΔ) και την αναλυτική επεξεργασία των δεδομένων του κύβου (OLAP). Ειδικότερα, για την ανάπτυξη της Αποθήκης Δεδομένων χρησιμοποιήθηκε ένα σύνολο δεδομένων με πληροφορίες και δεδομένα για τον Covid-19 από τον ιστότοπο <https://www.ecdc.europa.eu/en/publications-data/data-national-14-day-notification-rate-covid-19>. Έγινε σχεδιασμός της ΒΔ για την εισαγωγή των δεδομένων σε μία σχεσιακή Βάση Δεδομένων. Στη συνέχεια επιλέχθηκαν τα δεδομένα που χρησιμοποιήθηκαν για αναλυτική επεξεργασία, και προχωρήσαμε στην διαδικασία της προ επεξεργασίας (καθαρισμός, μετασχηματισμός) με τη βοήθεια του εργαλείου Excel. Πάνω στα δεδομένα που επιλέχθηκαν στο προηγούμενο βήμα, έγινε εισαγωγή τους στη Βάση Δεδομένων που σχεδιάστηκε, με τα δεδομένα να είναι έτοιμα για εισαγωγή στην Αποθήκη Δεδομένων. Στη συνέχεια, ορίστηκε μια αποθήκη δεδομένων – κύβος – με τις κατάλληλες διαστάσεις, ιεραρχίες και μέτρα. Η ΑΔ ακολουθεί ως μοντέλο το σχήμα αστέρα (star schema). Ακόμη, παρουσιάζονται τα ερωτήματα που θέσαμε στον κύβο, τα συμπεράσματα που οδηγούμαστε, καθώς επίσης και οπτικοποιημένα δεδομένα των απαντήσεων που λάβαμε από τον κύβο για σαφέστερα συμπεράσματα. Τέλος, εφαρμόστηκαν τεχνικές εξαγωγής κανόνων συσχέτισης από το σύνολο δεδομένων, περιγράφονται τα αποτελέσματα και τα μέτρα ενδιαφέροντος που χρησιμοποιήθηκαν (support, confidence, lift κλπ) για την επιλογή των κανόνων.

Περιεχόμενα

[**Περίληψη 2**](#_Toc65017075)

[**Περιεχόμενα 3**](#_Toc65017076)

[**1. Εισαγωγή 4**](#_Toc65017077)

[**2. Πηγή Δεδομένων 4**](#_Toc65017078)

[**3. Σχεδιασμός και Αρχιτεκτονική Αποθήκης Δεδομένων 7**](#_Toc65017079)

[**4. Αναλυτική Επεξεργασία-OLAP 9**](#_Toc65017079)

**5. Εξαγωγή Κανόνων Συσχέτισης (Assosiation Rules)………………………………………………….14**

**5.1 Εξαγωγή και αξιολόγηση itemsets………………….…………………………………………….15**

**5.2 Εξαγωγή και αξιολόγηση Κανόνων Συσχέτισης……………………………………..………17**

**6. Βιβλιογραφία………………………………………………………………………………………………………….19**

1. **ΕΙΣΑΓΩΓΗ**

Στην επιστήμη των υπολογιστών, μια αποθήκη δεδομένων (ΑΔ) είναι μια βάση δεδομένων που χρησιμοποιείται για την αναφορά και ανάλυση. Τα δεδομένα που είναι αποθηκευμένα στην αποθήκη φορτώνονται από το λειτουργικό σύστημα. Τα δεδομένα μπορεί να περάσουν μέσα από ένα λειτουργικό χώρο αποθήκευσης δεδομένων για τις πρόσθετες εργασίες πριν χρησιμοποιηθούν στις αποθήκες δεδομένων για την υποβολή ερωτήσεων.

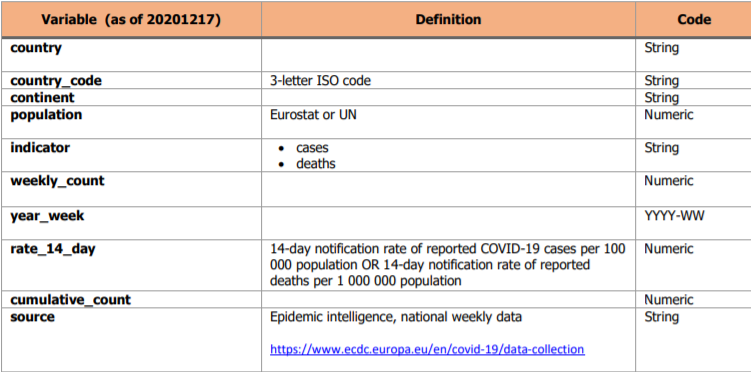
Μία αποθήκη δεδομένων διατηρεί τις λειτουργίες της σε τρία στρώματα: σταδιοποίηση, την ένταξη και την πρόσβαση. Η σταδιοποίηση χρησιμοποιείται για την αποθήκευση πρώτων δεδομένων με σκοπό τη χρήση τους από προγραμματιστές. Η ένταξη δεδομένων χρησιμοποιείται για την ενσωμάτωση δεδομένων ώστε να έχουν ένα επίπεδο αφαίρεσης από τους χρήστες. Το στρώμα πρόσβασης είναι για να εξάγουν τα επιθυμητά δεδομένα για τους χρήστες.

Οι αποθήκες δεδομένων μπορούν να υποδιαιρεθούν σε εξειδικευμένα υποσύνολα δεδομένων, τα λεγόμενα data marts. Τα data marts εφοδιάζουν επιμέρους υποσύνολα δεδομένων από μια αποθήκη.

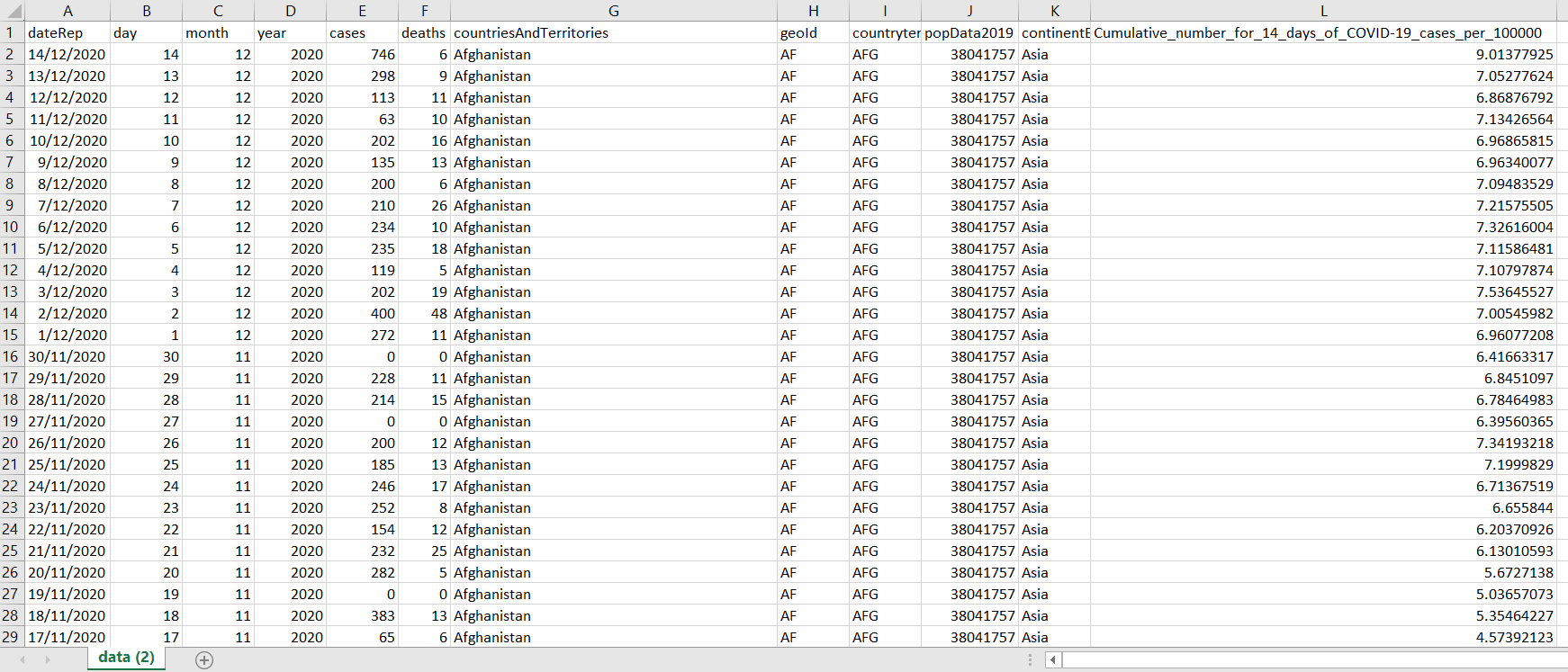
Ο ορισμός της αποθήκης δεδομένων εστιάζει στην αποθήκευση δεδομένων. Η κύρια πηγή των δεδομένων ξεδιαλύνεται, μεταμορφώνεται, κατηγοριοποιείται και διατίθεται για χρήση από τους διαχειριστές και άλλους επαγγελματίες για την εξόρυξη δεδομένων, διαδικτυακής αναλυτικής επεξεργασίας, έρευνα αγοράς και υποστήριξης αποφάσεων. Ωστόσο, τα μέσα για την ανάκτηση και την ανάλυση δεδομένων, την εξαγωγή, μετατροπή και φόρτωση των δεδομένων, καθώς και η διαχείριση του λεξικού δεδομένων θεωρούνται επίσης ουσιώδεις συνιστώσες ενός συστήματος αποθήκευσης δεδομένων. Πολλές αναφορές για την αποθήκευση δεδομένων χρησιμοποιούν αυτό το ευρύτερο πλαίσιο. Έτσι, ένας γενικός ορισμός για την αποθήκευση δεδομένων περιλαμβάνει εργαλεία επιχειρηματικής ευφυΐας, εργαλεία για την εξαγωγή, μετατροπή και φόρτωση δεδομένων στο χώρο αποθήκευσης, καθώς και τα εργαλεία για τη διαχείριση και ανάκτηση των μεταδεδομένων.

1. **Πηγή Δεδομένων:**

Η πηγή δεδομένων είναι βασική προϋπόθεση για τη σύσταση και τη δημιουργία μιας Αποθήκης Δεδομένων. Για την ανάπτυξη της Αποθήκης Δεδομένων χρησιμοποιήθηκε ένα σύνολο δεδομένων με πληροφορίες και δεδομένα για τον Covid-19 από τον ιστότοπο <https://www.ecdc.europa.eu/en/publications-data/data-national-14-day-notification-rate-covid-19>. Το σύνολο δεδομένων περιέχει πληροφορίες σχετικά με το ποσοστό των θετικών κρουσμάτων και θανάτων σε ημερήσια βάση, ανά χώρα και ήπειρο. Επιπλέον, υπολογίζεται το ποσοστό ανά 100.000 κατοίκους, 14 ημερών για τα αναφερόμενα κρούσματα ανά εκατομμύρια πληθυσμού ανά ημέρα και χώρα.

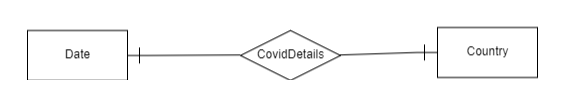


Πίνακας 1: Συνοπτικές πληροφορίες του συνόλου δεδομένων.



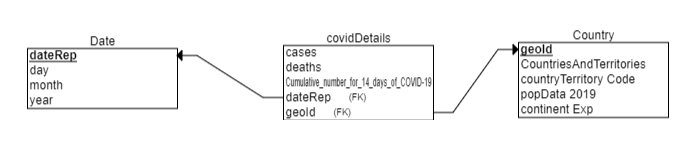
Πίνακας 2: Συνοπτικές πληροφορίες του συνόλου δεδομένων.

Για να επιτευχθεί η ανάλυση του συνόλου δεδομένων, είναι αναγκαίο να γίνουν έλεγχοι και μετασχηματισμοί ούτως ώστε να προετοιμάσουμε τα δεδομένα για την ανάλυση. Με τη βοήθεια του EXCEL διαπιστώθηκε για το σύνολο δεδομένων ότι δεν περιείχε ελλειπείς τιμές ή ακραίες τιμές για τη διαδικασία του καθαρισμού των δεδομένων (Data cleaning). Στη συνέχεια, έγιναν κάποιοι μετασχηματισμοί στα δεδομένα, ώστε να μετατραπούν σε σχεσιακούς πίνακες και να τα εισάγουμε στη βάση δεδομένων. Αρχικά δημιουργήθηκε το μοντέλο Οντοτήτων-Συσχετίσεων (ΕR model) για να διαπιστώσουμε τις οντότητες και τις σχέσεις μεταξύ των οντοτήτων, οι οποίες περιγράφουν το σύνολο δεδομένων.



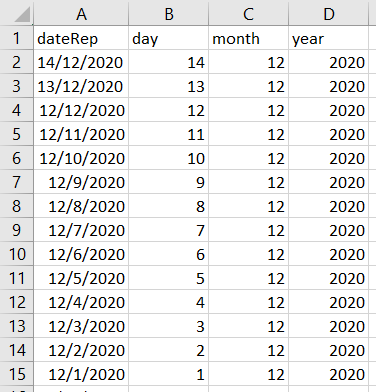
Πίνακας 3: Μοντέλο Οντοτήτων-Συσχετίσεων.

Ακολούθως, το μοντέλο Οντοτήτων-Συσχετίσεων, μετατράπηκε σε σχεσιακό (Relational model), ούτως ώστε να περιγράψουμε τις σχέσεις των οντοτήτων με τα κλειδιά τους.



Πίνακας 4: Σχεσιακό μοντέλο.

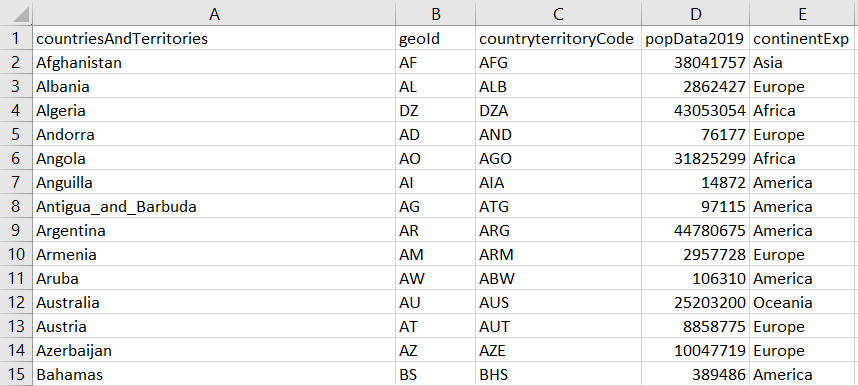
Εφόσον ολοκληρώθηκε η θεωρητική επεξεργασία του συνόλου για τη μετατροπή του σε σχεσιακή βάση δεδομένων, στα πλαίσια του μετασχηματισμού, έγινε διαχωρισμός των στηλών του αρχικού συνόλου και εισήχθησαν σε τρία νέα αρχεία αντίστοιχα, τύπου csv. Στο κάθε αρχείο έγινε ξεχωριστά η απαραίτητη επεξεργασία.



Πίνακας 5:Δεδομένα πίνακα Date.

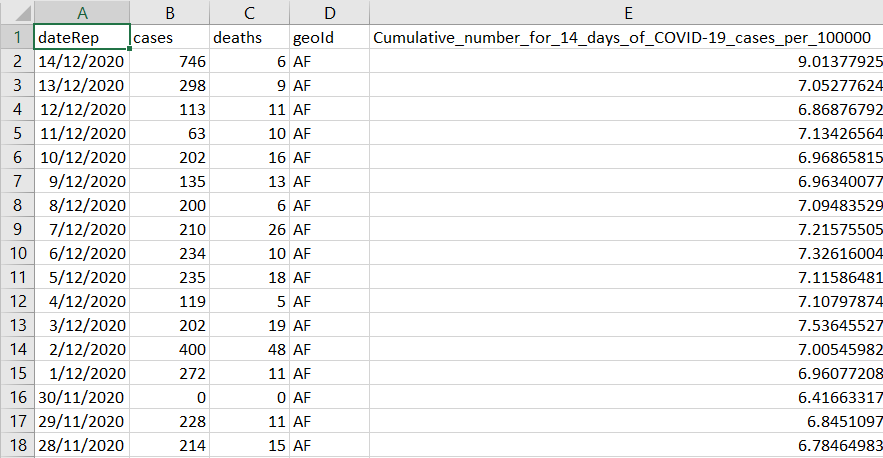
Ο παραπάνω πίνακας περιέχει τις ημερομηνίες και εφόσον η σχέση του με τον πίνακα covidDetails είναι ένα προς πολλά (1:Ν), διαγράφησαν οι διπλότυπες ημερομηνίες.

Αντίστοιχα και ο πίνακας Country που περιέχει λεπτομέρειες για τις χώρες και τις ηπείρους που ανήκουν, έχει σχέση ένα προς πολλά με τον πίνακα covidCountry, συνεπώς διαγράφησαν οι διπλότυπες τιμές.



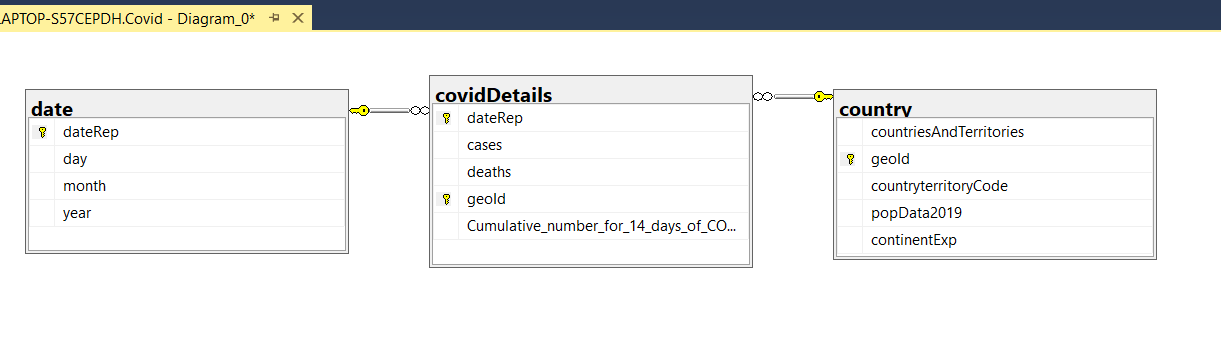
Πίνακας 6:Δεδομένα πίνακα covidCountry.

Ο πίνακας covidDetails αποτελεί ενδιάμεσο πίνακα, και συνδέει τους παραπάνω πίνακες. Περιέχει τα κλειδιά των πινάκων date και country ως ξένα κλειδιά (FK), καθώς επίσης και τα αναφερόμενα κρούσματα, τους αναφερόμενους θανάτους και το ποσοστό ανά 100.000 κατοίκους, 14 ημερών για τα αναφερόμενα κρούσματα ανά εκατομμύρια πληθυσμού ανά ημέρα και χώρα.



Πίνακας 6:Δεδομένα πίνακα covidDetails.

Τα δεδομένα πλέον είναι έτοιμα να εισαχθούν στη σχεσιακή βάση δεδομένων, ως σχεσιακοί πίνακες. Κατασκευάστηκε η βάση δεδομένων covid στον MSSQL Server και στη συνέχεια έγινε εισαγωγή των csv αρχείων. Οι πίνακες που δημιουργήθηκαν για το συγκεκριμένο σύνολο δεδομένων είναι τρείς, όσα και τα csv αρχεία που προέκυψαν στο στάδιο της επεξεργασίας και του μετασχηματισμού των δεδομένων. Οι τρείς πίνακες, dbo.country, dbo.covidDetails, dbo.date απεικονίζονται στην παρακάτω εικόνα.



Πίνακας 7: Physical model of Covid dataset.

**3.Σχεδιασμός και Αρχιτεκτονική Αποθήκης Δεδομένων**

Μια αποθήκη δεδομένων περιέχει πολλά αντικείμενα βάσης δεδομένων όπως πίνακες, προβολές, αποθηκευμένες διαδικασίες, λειτουργίες και ούτω καθεξής. Αποτελεί μια ΒΔ στήριξης αποφάσεων που διατηρείται ξεχωριστά από τη ΒΔ ενός οργανισμού/επιχείρησης. Επιπλέον, επικεντρώνεται στη μοντελοποίηση και ανάλυση δεδομένων με κύριο στόχο τη λήψη αποφάσεων σε βάθος χρόνου.

Ο σχεδιασμός της αποθήκης δεδομένων με τις κατάλληλες σχεδιαστικές τεχνικές υποστηρίζει την πολυδιάστατη ανάλυση δεδομένων. Για τη μοντελοποίηση της, έχουν αναπτυχθεί διάφορα μέτρα και διαστάσεις, όπως το:

* Star Schema
* Snowflake Schema,
* Fact Constellations (Galaxy Schema).

Για την ανάπτυξη της αποθήκης δεδομένων, χρησιμοποιήθηκε το εργαλείο SSDT. Δημιουργήθηκε το project covidAnalysis και έγινε εισαγωγή της ΒΔ που δημιουργήθηκε στο προηγούμενο βήμα. Για την κατασκευή του κύβου, επιλέχθηκαν οι πίνακες dbo.country, dbo.covidDetails, dbo.date μέσω των οποίων ορίστηκε το Fact table και τα Dimensions.

Η δομή των πινάκων κατέστησε αναγκαία την επιλογή του Star schema για την αρχιτεκτονική της Αποθήκης Δεδομένων και τη δομή του κύβου.

**Star Schema**: Το Star Schema μοιάζει με ένα αστέρι στο οποίο ο Fact Table ενεργεί ως άξονας καθώς βρίσκεται στο κέντρο, ενώ πολλαπλές διαστάσεις (Dimensions) συνδέονται με τον Fact table σε μορφή αστεριού με Foreign Keys. Ένα απλό Star Schema έχει συνήθως ένα Fact Table και πολλές διαστάσεις, αλλά τα σύνθετα σχήματα αστεριών μπορούν να αποτελούνται από περισσότερα.

**Fact table:**  Ο Fact table αποτελείται από παραμέτρους, τα measures και τα foreign keys. Τα measures αποτελούνται από αριθμητικές τιμές που μπορούν να μετρηθούν ή να μετρήσουν, ενώ τα foreign keys αποτελούν τη σύνδεση με τους πίνακες διαστάσεων. Τα measures μπορούν να χρησιμοποιηθούν για ανάλυση ερωτημάτων. Το Fact table αποτελεί ο πίνακας covidDetails.

**Dimensions:** Οι πίνακες διαστάσεων αποτελούνται από περιγραφικές τιμές. Κάθε πίνακας διάστασης έχει το δικό του πρωτεύον κλειδί. Οι πίνακες διάστασης είναι οι dim.country, dim.date.

**dim.country:**

Το Country dimension αποτελείται από τα χαρακτηριστικά geoId (Primary Key), CountriesAndTerritories, CountryTerritoryCode, popData2019, continent Exp. Αποτελείται από περιγραφικές τιμές σχετικά με τις χώρες, τον πληθυσμό τους και την ήπειρο την οποία ανήκουν.

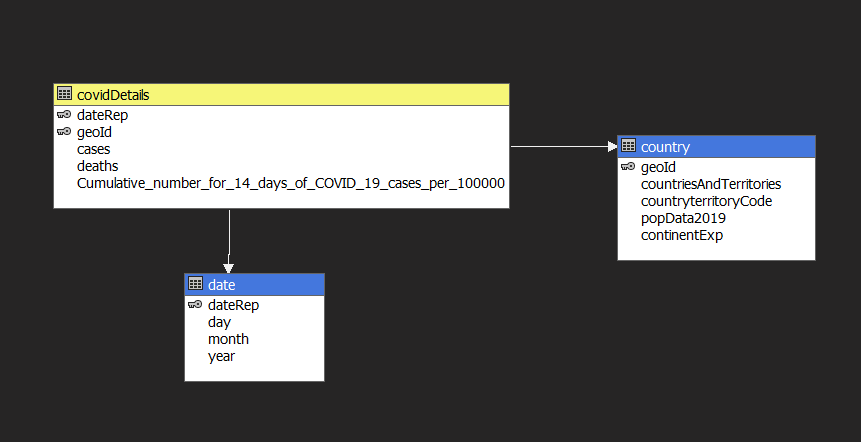
**dim.date:**

Το date dimension αποτελείται από τα χαρακτηριστικά dateRep(Primary Key), day, month, year. Αποτελείται από ημερομηνίες, ημέρες, μήνες και έτη.

**covidDetails:**

Το covidDetails αποτελείται από τα χαρακτηριστικά dateRep (Foreign Key), geoId(ForeignKey),cases,deaths,Cumulative\_number\_for\_14\_days\_of\_Covid\_10\_cases\_per\_100000.

**Star Schema of covidAnalysis:**



Πίνακας 8: Star schema του Covid Analysis.

**Ιεραρχίες Διαστάσεων:** Οι διαστάσεις αποτελούν τις παραμέτρους του προβλήματος που καλείται να λύσει ο κύβος, έτσι μπορούμε να πούμε ότι κάθε κύβος περιγράφει τους άξονες για τη μοντελοποίηση των ερωτημάτων ενός προβλήματος. Έτσι κρίνεται αναγκαία η δημιουργία ιεραρχιών στις διαστάσεις καθώς μπορούν εύκολα να διακρίνονται τα επίπεδα λεπτομέρειας των αξόνων που θέτουμε στα ερωτήματα. Κατασκευάστηκε μια σειρά ιεραρχιών σε κάθε διάσταση.

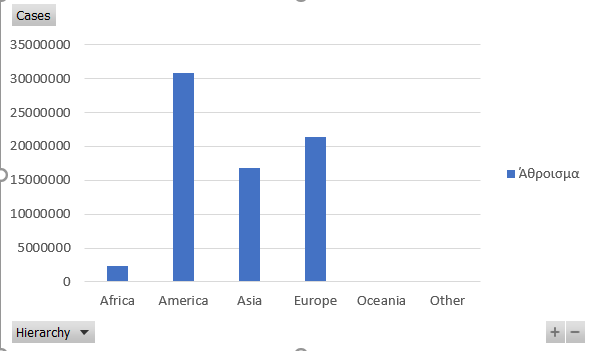
Hierarchy country: Στην ιεραρχία της διάστασης country τέθηκε πρώτα η ήπειρος (Continent Exp) και μετά η χώρα (Country).

Hierarchy date: Στην ιεραρχία της διάστασης date τέθηκαν πρώτα τα έτη (Year), μετά οι μήνες (Month) και τέλος οι μέρες (Day).

**4.Αναλυτική Επεξεργασία -OLAP:**

Πλέον η κατασκευή του κύβου έχει ολοκληρωθεί. Στη συνέχεια θα πρέπει να πραγματοποιηθεί πολυδιάστατη (multidimensional) αναπαράσταση των δεδομένων με τη βοήθεια του Analysis Servises (SSAS), καθώς και η οπτικοποίηση των αποτελεσμάτων για σαφέστερη απεικόνιση τους, με τη βοήθεια του Excel. Η δομή του κύβου σχεδιάστηκε με σκοπό να εξυπηρετήσει τα ερωτήματα που θέλουμε να θέσουμε στο σύνολο των δεδομένων, τα οποία είναι ενημερωτικά και μπορούν να καταστούν χρήσιμα για στατιστικά στοιχεία, για την πορεία και την εξέλιξη της πανδημίας αλλά και για στρατηγικές διαχείρισης της, λαμβάνοντας πληροφορίες για τους αναφερόμενους θανάτους και θετικά κρούσματα ανά χώρα και Ήπειρο, καθώς επίσης και για στατιστικά στοιχεία που αφορούν τον αριθμό των αναφερόμενων κρουσμάτων για 14 ημέρες ανά 100.000 πληθυσμού μιας χώρας.

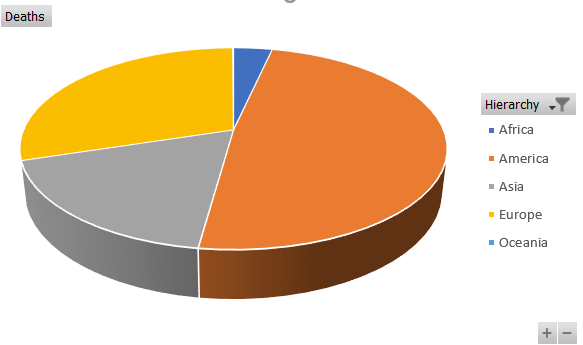
**Case study 1:**Αριθμός κρουσμάτων ανά Ήπειρο, τη χρονική περίοδο 2019-2020.  
Αυτό το ερώτημα τέθηκε περισσότερο για να έχουμε μια πρώτη εικόνα για τα δεδομένα μας, καθώς επίσης και για να κατανοήσουμε την εξάπλωση του Cocid-19 ανά Ήπειρο.



Πίνακας 9: Αριθμός κρουσμάτων ανά Ήπειρο τη χρονική περίοδο 2019-2020 .

**Case study 2:**

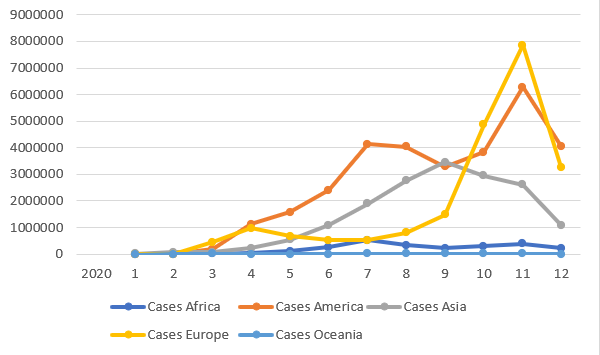
Αριθμός αναφερόμενων θανάτων ανά Ήπειρο, τη χρονική περίοδο 2019-2020.   
Αντίστοιχα με το παραπάνω ερώτημα, μπορούμε να κατανοήσουμε την επίπτωση που είχαν τα κρούσματα κάθε Ηπείρου για την κάθε χώρα, μαθαίνοντας πόσα από αυτά τα κρούσματα κατέληξαν σε θάνατο. Όπως ήταν αναμενόμενο, η Αμερική εξακολουθεί να έχει το μεγαλύτερο ποσοστό θανάτων, καθώς την περίοδο 2019-2020 ανέφερε τα περισσότερα κρούσματα. Θα είχε αξία στην ανάλυση και την αποτίμηση της, εάν παρατηρούσαμε κάποια Ήπειρο η οποία παρόλο που είχε περισσότερα κρούσματα από κάποια άλλη, χάριν σε κάποιο μαζικό στρατηγικό σχέδιο, να κατάφερνε να περιόριζε τον αριθμό των θανάτων, περισσότερο από άλλες. Κάτι τέτοιο όμως δεν παρατηρείται στη δική μας περίπτωση.



Πίνακας 10: Αριθμός θανάτων ανά Ήπειρο τη χρονική περίοδο 2019-2020 .

**Case study 3:**

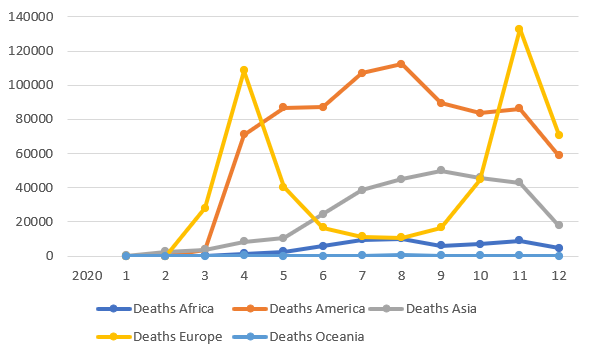
Το επόμενο ερώτημα ήταν ειδικότερο και είχε να κάνει με την πορεία των κρουσμάτων της κάθε Ηπείρου, ανά μήνα. Το ερώτημα αυτό μπορεί να φανεί χρήσιμο, καθώς κάθε Ήπειρος, κλήθηκε σε διαφορετικά χρονικά διαστήματα να αντιμετωπίσει την άνοδο των κρουσμάτων. Είναι χρήσιμο να παρατηρήσουμε το χρονικό διάστημα που τα κρούσματα βρίσκονται σε υψηλά επίπεδα, την τάση τους και το κατά πόσο ένα στρατηγικό πλαίσιο διαχείρισης της πανδημίας, μπορεί να περιορίσει τον αριθμό των κρουσμάτων ανά μήνα. Τέλος, μπορεί να παρατηρηθεί ότι οι μήνες τους οποίους κάθε Ήπειρος έχει υψηλό αριθμό κρουσμάτων αλλάζουν. Αυτό μπορεί να προκύπτει από το ότι τους καλοκαιρινούς μήνες η Ευρώπη είναι τουριστικός προορισμός και προσελκύει ανθρώπους από διαφορετικές Ηπείρους. Έτσι η εκθετική τάση των κρουσμάτων που φαίνεται να ξεκινάει από τον 7ο μήνα του έτους, φαίνεται να αυξάνει με εκθετικούς ρυθμούς τους επόμενους μήνες.



Πίνακας 11: Αριθμός κρουσμάτων ανά Ήπειρο τους μήνες του 2020.

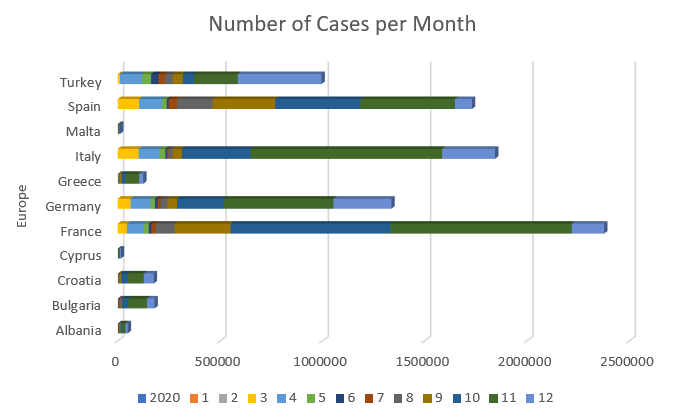
**Case study 4:**

Ως αποτέλεσμα των κρουσμάτων, ακολουθεί η άνοδος των θανάτων. Το επόμενο ερώτημα που θέσαμε στο σύνολο δεδομένων, ήταν για να διαπιστώσουμε, εάν η εκθετική αύξηση των κρουσμάτων που ξεκίνησε τους καλοκαιρινούς μήνες στην Ευρώπη, είχε σαν μοιραίο αποτέλεσμα την άνοδο των αναφερόμενων θανάτων. Όπως επιβεβαιώθηκε και διαγραμματικά από το ερώτημά μας, ένα μήνα αργότερα από την εμφάνιση της αυξητικής τάσης των κρουσμάτων, φαίνεται να ξεκίνησε η εκθετική αύξηση των θανάτων. Έτσι από τον 8ο μήνα στην Ευρώπη, παρατηρούμε το αποτέλεσμα που είχε η αύξηση των κρουσμάτων.



Πίνακας 12: Αριθμός θανάτων ανά Ήπειρο τους μήνες του 2020.

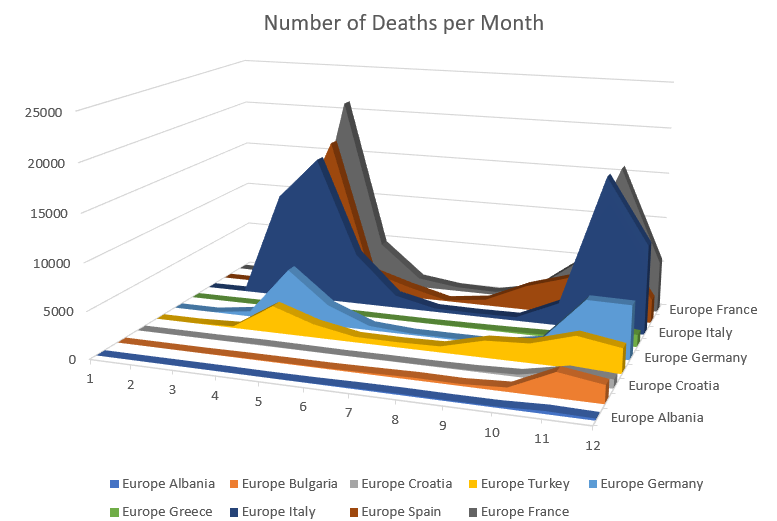
**Case study 5:**Στα επόμενα ερωτήματα εμβαθύναμε περισσότερο στην Ευρώπη. Πιο συγκεκριμένα, αναζητήσαμε τον αριθμό των κρουσμάτων για συγκεκριμένες χώρες της Ευρώπης, ανάμεσά τους και η Ελλάδα. Θέλαμε να παρατηρήσουμε τον τρόπο που διαμοιράστηκαν τα κρούσματα στις χώρες της Ευρώπης, τις χρονικές περιόδους που φαίνεται να αυξάνονται, καθώς επίσης και το μερίδιο που καταλαμβάνουν στο συνολικό αριθμό των κρουσμάτων της Ευρώπης.



Πίνακας 13: Αριθμός κρουσμάτων σε χώρες της Ευρώπης.

**Case study 6:**

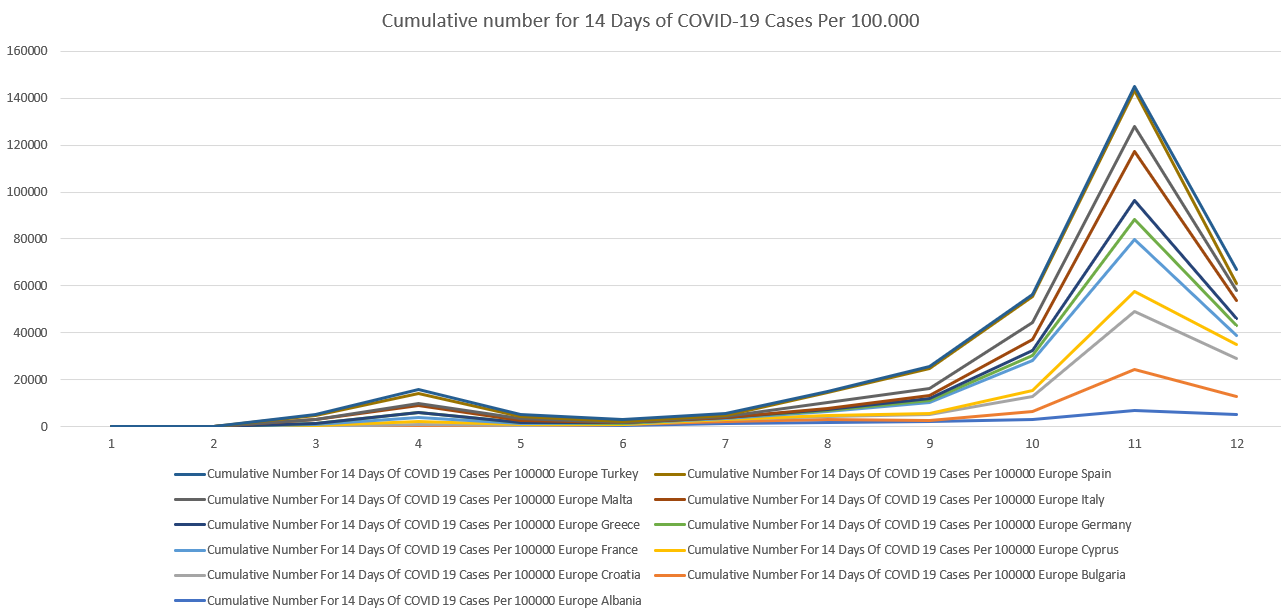
Στο επόμενο ερώτημα, προσπαθήσαμε να παρατηρήσουμε τους μήνες στους οποίους τα κρούσματα φαίνονται να αυξάνουν και εκείνους που περιορίζονται. Όπως μπορεί να παρατηρηθεί από την παρακάτω οπτικοποίηση των δεδομένων, η αρχή της πανδημίας σήμανε για τις χώρες της Ευρώπης σήμα κινδύνου καθώς η ραγδαία άνοδος των θανάτων ακολουθούσε ανοδική τάση. Ακολούθησαν πολιτικές αυστηρού Lock-down, για το σύνολο των χωρών, κάτι που φαίνεται να λειτούργησε αποτελεσματικά στη διαχείρισή τους. Όμως τέτοιες αυστηρές στρατηγικές δε μπορούν να διαρκέσουν για μεγάλα χρονικά διαστήματα. Έτσι κατά τους καλοκαιρινούς μήνες, φαίνεται να υπήρξε άνοδος του αριθμού των θανάτων λόγω της χαλάρωσης των μέτρων αυτών. Οι καλοκαιρινοί μήνες όμως είχαν σαν αποτέλεσμα το δεύτερο κύμα της πανδημίας, τη νέα επικίνδυνη άνοδο των αναφερόμενων θανάτων και έτσι ακολούθησε το δεύτερο Lock-down.



Πίνακας 14: Αριθμός θανάτων σε χώρες της Ευρώπης.

**Case study 7:**

Η στήλη Cumulative number for 14 Days of Covid-19 per 100.000 ,μετράει τον αριθμό των κρουσμάτων των χωρών, ανά 100.000 κατοίκους. Δηλαδή δείχνει το βαθμό που έχουν προσβληθεί οι χώρες από τον Covid-19 με βάση έναν σταθερό αριθμό πληθυσμού. O αριθμός αυτός προκύπτει λαμβάνοντας υπόψιν τα ενεργά κρούσματα από τη χρονική στιγμή που ανακοινώνονται έως και τις επόμενες 13 ημέρες όπου ο ιός παραμένει ενεργός. Συνεπώς μπορούμε να παρατηρήσουμε, να αναζητήσουμε και να συγκρίνουμε χώρες μεταξύ τους σχετικά με την εξάπλωση του Covid-19 με βάση τα ενεργά κρούσματα που έχουν ανά ημέρα ή μήνα. Αυτό το ερώτημα μπορεί να δώσει αξιόπιστες απαντήσεις και συμπεράσματα σχετικά με την εξέλιξη της πανδημίας και το βαθμό τον οποίο έχει προσβληθεί μια χώρα από τον Covid-19. Είναι χρήσιμο για την κατανόηση και την πλήρη εικόνα της πορείας των κρουσμάτων, βοηθώντας αρμόδιες επιτροπές να γνωρίζουν και να λαμβάνουν έγκαιρα μέτρα πρόληψης και διαχείρισης.



Πίνακας 15: Aθροιστικός αριθμός κρουσμάτων για 14 ημέρες χωρών της Ευρώπης.

**5.Εξαγωγή Κανόνων Συσχετίσεων (Association Rules):**

Η ανακάλυψη Κανόνων Συσχέτισης είναι μία από τις βασικότερες εργασίες Εξόρυξης Δεδομένων. Αντικείμενό της είναι η ανακάλυψη και διατύπωση σχέσεων-κανόνων, οι οποίες υπάρχουν στα δεδομένα. Οι σχέσεις αυτές προκύπτουν από τη συχνή ταυτόχρονη εμφάνιση τιμών δεδομένων. Ο βασικός αλγόριθμος εντοπισμού συχνών στοιχειοσυνόλων είναι ο Apriori. Οι κανόνες που παράγονται ικανοποιούν τα κριτήρια της υποστήριξης και της εμπιστοσύνης, δεν είναι σίγουρο όμως ότι είναι και ενδιαφέροντες. Ένα πρόσθετο στατιστικό μέτρο αξιολόγησης των κανόνων είναι το Lift, το οποίο αποδίδει τον βαθμό και το είδος συσχέτισης μεταξύ των γεγονότων του κανόνα. Εξόρυξη κανόνων μπορεί να γίνει όχι μόνο από βάσεις δεδομένων συναλλαγών, αλλά και από σχεσιακές βάσεις δεδομένων ή από πολυδιάστατους κύβους. Οι κανόνες αυτοί μπορεί να περιέχουν πολλά κατηγορήματα. Στην περίπτωση αυτή καλούνται πολυδιάστατοι. Εάν έχουν καθοριστεί ιεραρχίες εννοιών, μπορούν να δημιουργηθούν κανόνες που να αφορούν διαφορετικά επίπεδα γενίκευσης. Η χρήση διαφορετικών τιμών ελάχιστης υποστήριξης (minimum support) για τα διαφορετικά επίπεδα γενίκευσης διευκολύνει τη δημιουργία περιορισμένου αριθμού κανόνων για τα ανώτερα επίπεδα, καθώς και τον εντοπισμό ισχυρών κανόνων στα κατώτερα επίπεδα. Ένα σημαντικό πρόβλημα είναι η εξόρυξη κανόνων από αριθμητικά δεδομένα. Για την αντιμετώπιση αυτού του προβλήματος έχουν προταθεί διάφορες μέθοδοι, οι περισσότερες από τις οποίες εφαρμόζουν τη διακριτοποίηση, σε συνδυασμό με άλλες πρόσθετες τεχνικές. Με τη χρήση περιορισμών κατά την εξόρυξη των κανόνων επιτυγχάνεται επικέντρωση της ανάλυσης σε εντοπισμένα ζητήματα, αλλά και ταυτόχρονη μείωση του χώρου αναζήτησης λύσεων. Η δημιουργία κανόνων-συσχετίσεων έγινε με τη βοήθεια του εργαλείου MS SSDT. Πιο συγκεκριμένα, στο Mining Structures, επιλέχθηκε η δημιουργία Microsoft Association Rules, από τα υπάρχοντα δεδομένα του Covid-19.

Για την ανακάλυψη Κανόνων Συσχέτισης, εντοπίζουμε όλους τους κανόνες που έχουν υποστήριξη( support) και εμπιστοσύνη ίση ή μεγαλύτερη από τις προκαθορισμένες τιμές. Οι κανόνες αυτοί θεωρούνται ισχυροί. Ορισμένοι πρόσθετοι όροι των Κανόνων Συσχέτισης είναι οι ακόλουθοι:

* Στοιχειοσύνολο (Itemset) ονομάζεται ένα σύνολο από στοιχεία.
* k-Στοιχειοσύνολο (k-Itemset) είναι ένα στοιχειοσύνολο που περιέχει k στοιχεία.
* Συχνότητα Εμφάνισης (frequency ή support count ή count) ενός στοιχειοσυνόλου είναι το πλήθος των συναλλαγών που περιέχουν το στοιχειοσύνολο
* Υποστήριξη (support) ενός στοιχειοσυνόλου είναι το ποσοστό των συναλλαγών που περιέχουν το στοιχειοσύνολο
* Συχνό στοιχειοσύνολο είναι εκείνο το στοιχειοσύνολο του οποίου η υποστήριξη είναι ίση ή μεγαλύτερη από την ελάχιστη υποστήριξη, που ορίζει ο χρήστης.

Ο αλγόριθμος Assosiation Rules εκτελείται σε δύο στάδια. Στο πρώτο στάδιο, γίνονται οι υπολογισμοί για να επιλεγούν τα Itemsets που εμφανίζονται με τη μεγαλύτερη συχνότητα. Στο δεύτερο στάδιο, δημιουργούνται οι κανόνες συσχέτισης με βάση τις συχνότητες των itemsets. Ας υποθέσουμε ότι έχουμε τις χώρες Α, Β και τα σύνολα {Α} και {Α, Β}.

1. Στο πρώτο στάδιο, γίνονται οι υπολογισμοί, για να επιλεγούν τα itemsets που εμφανίζονται με τη μεγαλύτερη συχνότητα με την εύρεση του Support που έχουν είτε μεμονωμένα αντικείμενα είτε μετέπειτα συνδυασμοί αυτών.

* Support({A}): Το σύνολο των εγγραφών στις οποίες εμφανίζεται η χώρα Α (item {A}).
* Support({A, B}): Το σύνολο των εγγραφών στις οποίες εμφανίζονται οι χώρες Α και Β μαζί (itemset {A, B}).

Αν Support >= MINIMUM\_SUPPORT, τότε το itemset γίνεται αποδεκτό.

2. Στο δεύτερο στάδιο, δημιουργούνται οι κανόνες συσχέτισης με βάση τη πιθανότητα (probability) ή, αλλιώς, εμπιστοσύνη (confidence). Έστω ότι ελέγχουμε αν οι θάνατοι της χώρας Β εξαρτάται από τους θανάτους της χώρας Α. Ο τύπος του probability είναι ο ακόλουθος:   
Probability (A=>B) = Probability (B|A) = Support ({A, B})/Support ({A})  
Αν Probability >= MINIMUM\_PROBABILITY, ο κανόνας είναι ισχυρός.  
Τονίζεται ότι ο δείκτης probability μάς βοηθά να ελέγξουμε αν ένας κανόνας είναι «ισχυρός». Όμως, ένας κανόνας μπορεί να είναι ισχυρός (να έχει υψηλό probability/confidence) και να είναι «παραπλανητικός». Δηλαδή, να υπάρχει αρνητική συσχέτιση μεταξύ των στοιχείων του κανόνα. Για τον εντοπισμό «παραπλανητικών κανόνων» υπάρχει ο δείκτης σημαντικότητας (importance) που αξιολογεί τους κανόνες συσχέτισης. Ο τύπος του importance δίνεται παρακάτω:

Importance (A=>B) = log (Probability (Β|A)/Probability (B| not A))  
Αν importance = 0, δεν υπάρχει καμία συσχέτιση μεταξύ Α και Β.  
Αν importance < 0, probability(Β) μειώνεται αν το Α είναι αληθές.  
Αν importance > 0, probability(Β) αυξάνεται αν το Α είναι αληθές.

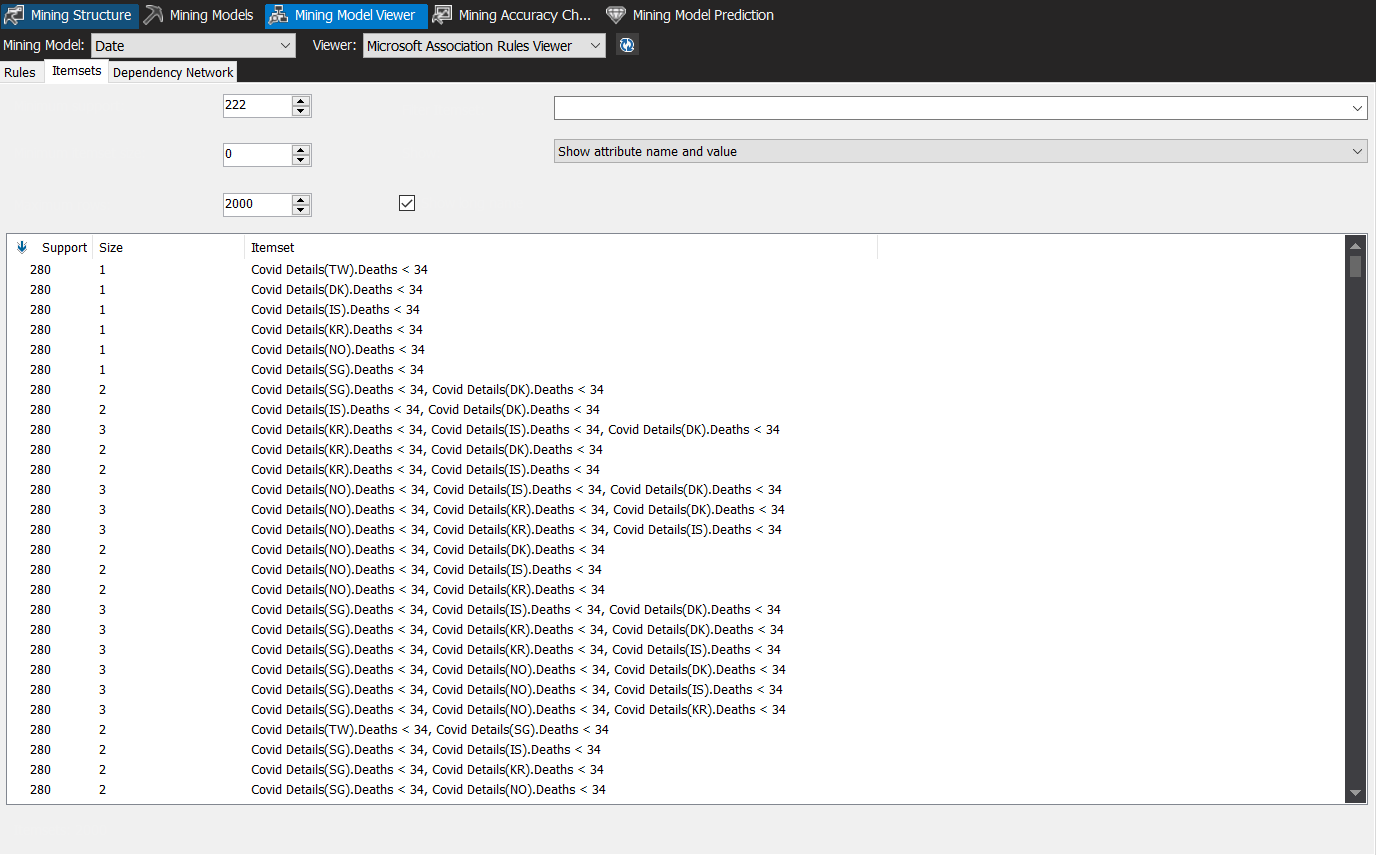
**5.1. Εξαγωγή και αξιολόγηση Itemsets.**

Για να αξιολογήσουμε τα Itemsets, επιλέγουμε την καρτέλα Mining Model Viewer και κατόπιν, την καρτέλα Itemsets, ώστε να εμφανιστούν τα itemsets με τη μεγαλύτερη συχνότητα που έχουν δημιουργηθεί από τον αλγόριθμο association rules. Όπως φαίνεται στoν Πίνακα 16,εμφανίζεται ένας πίνακας με τρεις στήλες.

1. Η στήλη **Support** εμφανίζει τη συχνότητα του κάθε itemset. Η ελάχιστη τιμή που παρατηρούμε σ ’αυτήν την στήλη καθορίζεται από την τιμή της παραμέτρου **MINIMUM\_SUPPORT**. Αν η τιμή της παραμέτρου είναι πολύ μικρή, είναι πιθανό να εμφανιστούν πολλά itemsets.
2. Η στήλη **Size** απεικονίζει το πλήθος των αντικειμένων που συγκροτούν το itemset. Η μέγιστη τιμή εξαρτάται από την τιμή της παραμέτρου MAXIMUM\_ITEMSET\_SIZE. Η προεπιλεγμένη τιμή της παραμέτρου είναι 3 και, επομένως, το μέγιστο πλήθος των ταινιών που περιέχει ένα itemset δεν ξεπερνάει τις 3.
3. Η στήλη **Itemset** περιέχει τα αντικείμενα (items) από τα οποία αποτελείται το κάθε itemset.

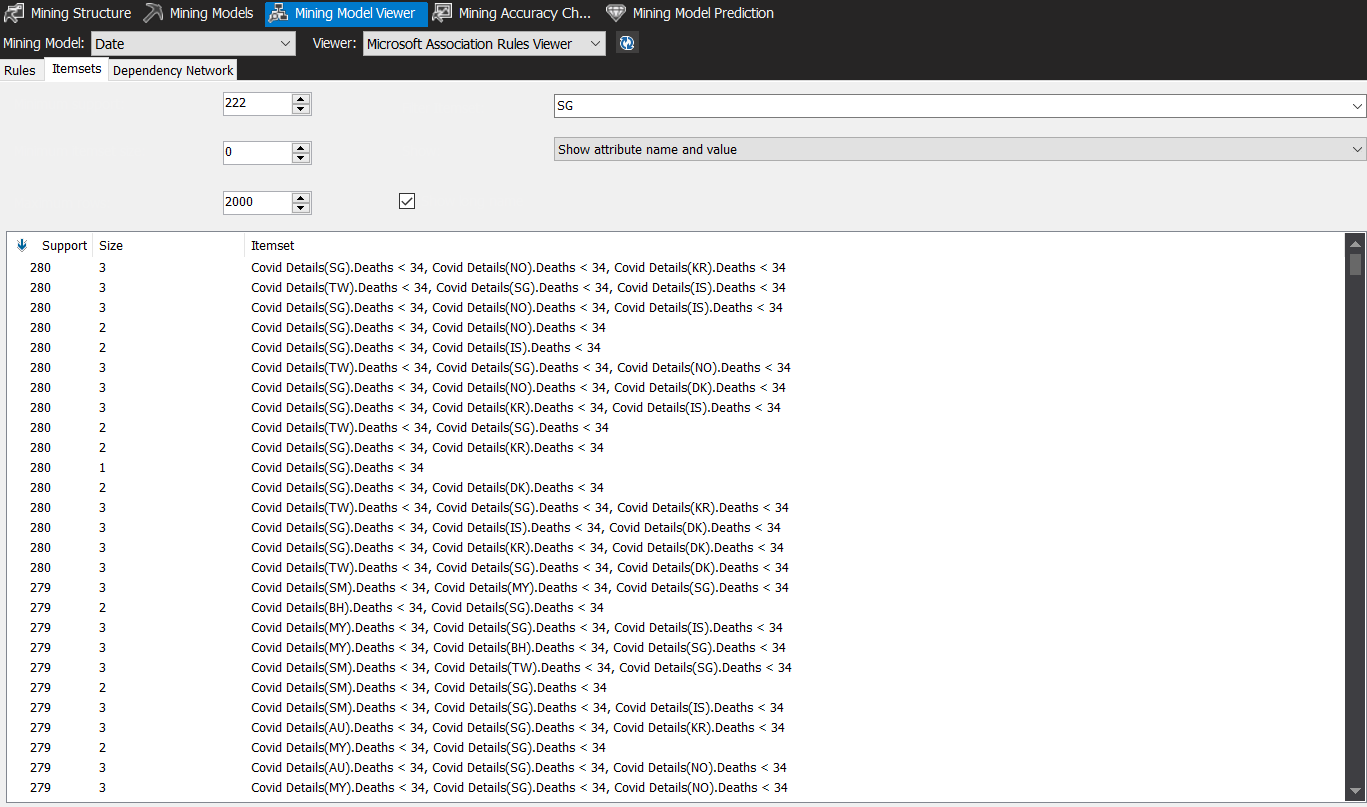
Επιπλέον, παρατηρούμε, όπως φαίνεται στην ίδια Εικόνα, ότι υπάρχουν διάφορα πεδία με τα οποία μπορούμε να παραμετροποιήσουμε τα αποτελέσματα. Για παράδειγμα, στη συγκεκριμένη περίπτωση:

* Στο πεδίο **Minimum** **Support** καθορίζεται ο ελάχιστος αριθμός εμφανίσεων που απαιτούμε να έχει το κάθε itemset. H τιμή του εδώ είναι 222.
* Στο πεδίο **Minimum itemset size** καθορίζεται το ελάχιστο πλήθος των items που συγκροτούν ένα itemset.
* Στο πεδίο **Maximum rows** καθορίζουμε το πλήθος των Itemsets. Η τιμή που δίνουμε εδώ ισούται με 2000 και περιορίζει την τιμή της παραμέτρου **MAXIMUM\_ITEMSET\_COUNT**.
* Στο πεδίο **Filter Itemset** καθορίζουμε το item με βάση το οποίο θέλουμε να φιλτράρουμε τα itemsets.
* Στο πεδίο **Show** επιλέγουμε τον τρόπο με τον οποίο θα εμφανίζονται τα itemsets(π.χ. show attribute name and value).



Πίνακας 16: Εξαγωγή itemsets.

Στη συνέχεια, ας υποθέσουμε ότι θέλουμε να βρούμε τα itemsets που εμπεριέχουν τη χώρα Σανγκάη (SG) και, επίσης, να βλέπουμε μόνο τα names των attributes. Όπως φαίνεται στoν Πίνακα 17, συμπληρώνουμε στο πεδίο Filter Itemset το όνομα της χώρας, δηλαδή SG. Επίσης, επιλέγουμε στο dropbox Show την επιλογή Show attribute name only. Όπως φαίνεται στην ίδια Εικόνα, βλέπουμε πλέον αποκλειστικά itemsets που εμπεριέχουν τη χώρα Σανγκάη (SG).



Πίνακας 17: Εξαγωγή itemsets που αφορούν τη χώρα Σανγκάη (SG) .

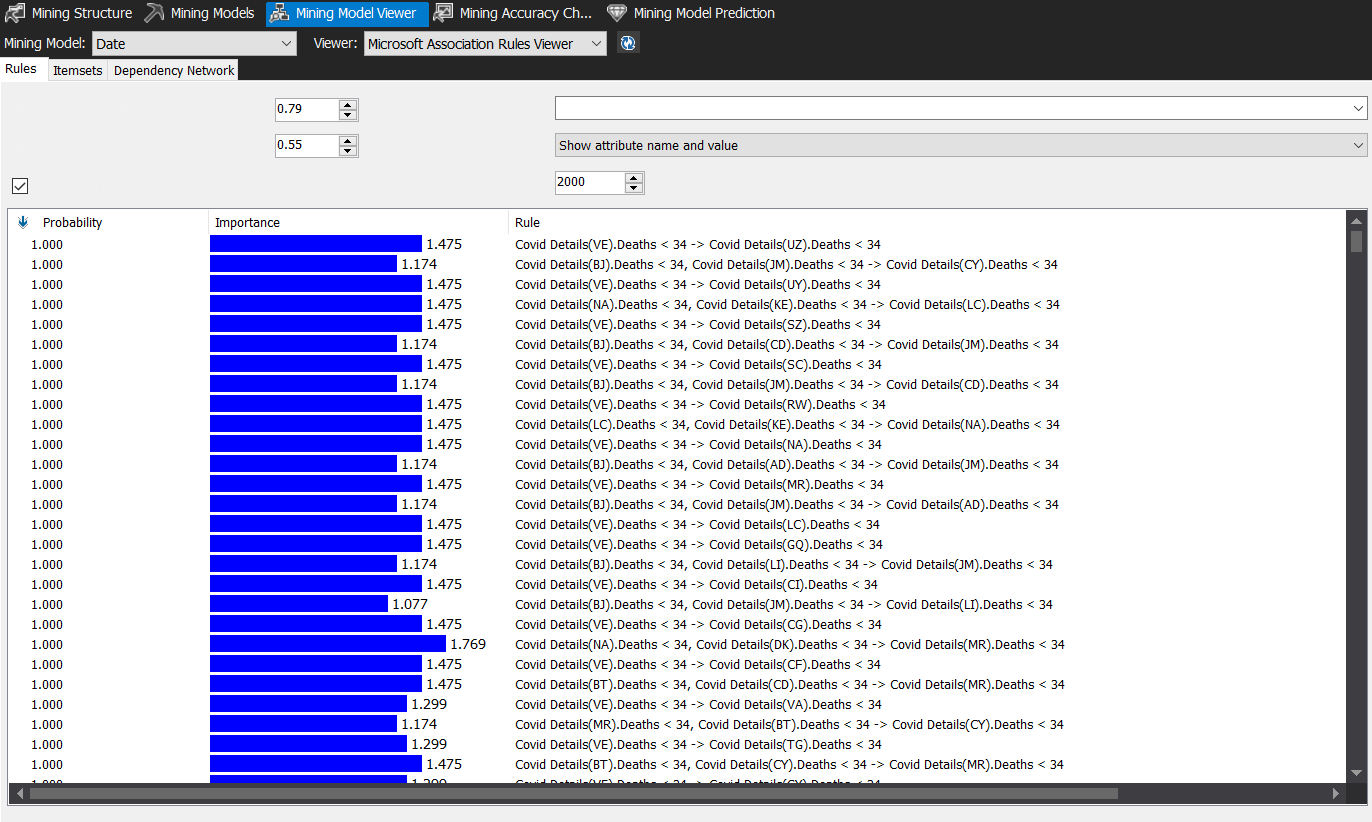
**5.2 Εξαγωγή και αξιολόγηση Κανόνων Συσχέτισης**

Για να αξιολογήσουμε τους κανόνες συσχέτισης, επιλέγουμε την καρτέλα Mining Model Viewer και,κατόπιν,την καρτέλα Rules, ώστε να εμφανιστούν οι κανόνες συσχέτισης με τη μεγαλύτερη πιθανότητα/εμπιστοσύνη (probability/confidence) και βαθμό σημαντικότητας (importance). Εφαρμόσθηκε στα δεδομένα διακριτοποίηση και στη συνέχεια επιλέχθηκε το 80% του συνόλου δεδομένων για σύνολο εκπαίδευσης, ενώ το 20% για σύνολο ελέγχου. Όπως φαίνεται στον Πίνακα 18, εμφανίζεται ένας πίνακας με τρεις στήλες.

1. Στην στήλη **Probability** εμφανίζεται ο βαθμός εμπιστοσύνης/πιθανότητας του κανόνα. Όλα τα ποσοστά που βλέπουμε είναι πάνω από 0.79, καθώς η επιλεγμένη τιμή της παραμέτρου **MINIMUM\_PROBABILITY** είναι 0.79.
2. Στην στήλη **Importance** εμφανίζεται ο βαθμός σημαντικότητας του κανόνα συσχέτισης.
3. Στην στήλη **Rule** απεικονίζονται οι κανόνες που τελικά παρήγαγε ο αλγόριθμος. Κάθε κανόνας αποτελείται από το αριστερό και το δεξιό μέρος και έχει την παρακάτω μορφή:

Σύνολο μεταβλητών που εξαρτούν τις μεταβλητές του δεξιού μέλους Σύνολο μεταβλητών που εξαρτώνται από τις μεταβλητές του αριστερού μέλους.

Έτσι, οι θάνατοι μιας χώρας που βρίσκεται στο δεξιό μέρος του κανόνα εξαρτάται από τους θανάτους της χώρας ή των χωρών που βρίσκονται στο αριστερό μέρος του.

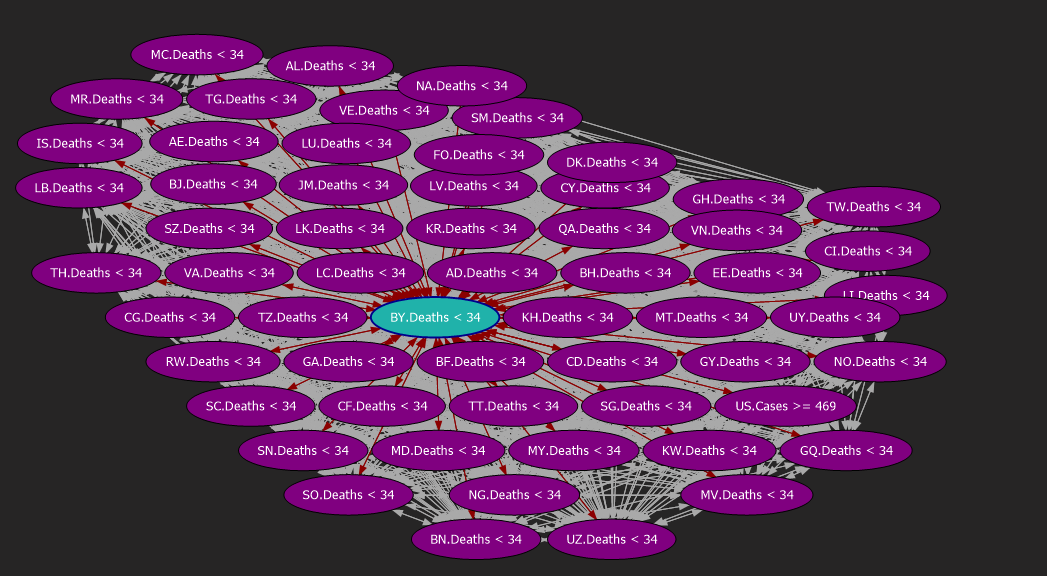


Πίνακας 18: Κανόνες συσχέτισης θανάτων μεταξύ γειτονικών χωρών.

Στη συνέχεια, για να δούμε καλύτερα τις συσχετίσεις που έχει βρει ο αλγόριθμος, ανοίγουμε την καρτέλα Dependency Network του Mining Model Viewer, όπως φαίνεται στον Πίνακα 19. Παρατηρούμε ότι:

* Κάθε κόμβος αντιπροσωπεύει ένα ξεχωριστό item.
* Δύο ή περισσότερα items που συνδέονται με μια γραμμή σχηματίζουν ένα itemset.
* Το βέλος μεταξύ δύο items συμβολίζει την ύπαρξη κανόνα.
* Η γραμμή κύλισης στο αριστερό μέρος του παραθύρου ελέγχει τον δείκτη importance score.

Εάν κάνουμε κλικ σ ’έναν κόμβο, βλέπουμε τους υπόλοιπους κόμβους που σχετίζονται με αυτόν. Για παράδειγμα, κάνοντας κλικ σένα node (π.χ. BY.Deaths<34), βλέπουμε όλους τους nodes που σχετίζονται με αυτόν. Με διαφορετικά χρώματα βλέπουμε αυτόν που επιλέξαμε, τους nodes που εξαρτώνται απ’ αυτόν και τους nodes απ’τους οποίους αυτός εξαρτάται.



Πίνακας 19: Dependency Network των κανόνων συσχέτισης θανάτων μεταξύ γειτονικών χωρών.

# Βιβλιογραφία

Han J, K. M. (2006). *Data Mining: Concepts and Techniques.* 2nd Edition: Morgan Kaufmann.

Lawrence Corr, J. S. (2011). *Agile Data Warehouse Design: Collaborative Dimensional Modeling, from Whiteboard to Star Schema.* DecisionOne Press.

Rainardi, V. (2008). *Building a Data Warehouse: With Examples in SQL Server.* 1st ed. Edition: Apress.

Ralph Kimball, M. R. (2013). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling.* 3rd Edition: Wiley.

Παναγιώτης Συμεωνίδης, Α. Γ. (2015). *Βάσεις, Αποθήκες καιΕξόρυξη Δεδομένωνμε τον SQL ServerΕργαστηριακός Οδηγός.* Ελληνικα Ακαδημαικά Ηλεκτρονικά Συγγράμματα και Βοηθήματα.