Χωρική ανάλυση και μοντελοποίηση

Διερευνητική Ανάλυση Χωρικών Δεδομένων

Exploratory Spatial Data Analysis (ESDA)

Κ. Ποϊραζίδης

Η παρουσίαση Βασίστηκε κυρίως στο βιβλίο Spatial Analysis – Methods and Practice, Grekousis 2020

Διερευνητική Ανάλυση Χωρικών Δεδομένων (ESDA)

- Τι θα μάθουμε
  - Την έννοια της Διερευνητικής Χωρικής Ανάλυσης Δεδομένων
  - Χωρικά στατιστικά και η σημαντικότητα τους στην ανάλυση χωρικών δεδομένων
  - Ανάλυση μονομεταβλητών / διμεταβλητών δεδομένων
  - Απλά εργαλεία της Διερευνητικής Χωρικής Ανάλυσης Δεδομένων (ιστογράμματα- θηκογράμματα, άλλες τεχνικές), για ενδότερη διερεύνηση των χωρικών βάσεων δεδομένων
  - Συσχετίσεις
  - Κανονικοποιήσεις δεδομένων
  - Η σημαντικότητα της κανονικής κατανομής στην κλασσική στατιστική και ενσωμάτωση της στη χωρική ανάλυση

Διερευνητική Ανάλυση Χωρικών Δεδομένων (ESDA)

- Τι είναι η ESDA?

- Μια συλλογή οπτικών και αριθμητικών μεθόδων ανάλυσης χωρικών δεδομένων για:

  - Εφαρμογή κλασσικής μη χωρικών στατιστικών που συνδέονται δυναμικά με χωρικά αντικείμενα

  - Προσδιορισμός χωρικών αλληλεπιδράσεων, σχέσεων και προτύπων

Διερευνητική Ανάλυση Χωρικών Δεδομένων (ESDA)

- Που χρησιμοποιείται;
  - Περιγραφή και σύνοψη χωρικών κατανομών
  - Εντοπισμός προτύπων ενδιαφέροντος
  - Οπτικοποίηση χωρικών κατανομών
  - Εξέταση χωρικής αυτοσυσχέτισης
  - Αναγνώριση χωρικών ακραίων τιμών
  - Προσδιορισμός συστάδων (clusters)
  - Αναγνώριση θερμών και ψυχρών θέσεων (hot – cold spots)

Διερευνητική Ανάλυση Χωρικών Δεδομένων (ESDA)

- **Περιγραφική Στατιστική και ESDA**

- Περιγραφή της βάσης δεδομένων = 1ο βήμα στην ανάλυση
  - Κατανόηση της μεταβλητότητας των δεδομένων
  - Προσδιορισμός πιθανών λαθών

- Περιγραφικά στατιστικά = summary statistics

Διερευνητική Ανάλυση Χωρικών Δεδομένων (ESDA)

- **Περιγραφική Στατιστική και ESDA**

- Τα περιγραφικά στατιστικά είναι χρήσιμα για τον υπολογισμό τιμών όπως η μέση τιμή και η τυπική απόκλιση => κατανομή των δεδομένων
  - Κατανόηση της μεταβλητότητας των δεδομένων, αλλά όχι της χωρικής κατανομής
  - Αλλά μέσω κυρίως της χωρικής αυτοσυσχέτισης => ανακάλυψη προτύπων
- Σύνδεση ιστογραμμάτων/ θηκογραμμάτων/ scatter plots με τη χωρική θέση τους

Διερευνητική Ανάλυση Χωρικών Δεδομένων (ESDA)

- **Εργαλεία ESDA και Περιγραφικής Στατιστικής**

  για οπτικοποίηση μονομεταβλητών χωρικών δεδομένων

- Χωροπληθείς χάρτες (*choropleth maps*)
- Κατανομή συχνοτήτων και ιστογράμματα
- Μέτρηση του κέντρου, διασποράς και σχήματος της κατανομής
- Μέτρηση της τάσης / Μεταβλητότητας
- Αναγνώριση ακραίων τιμών
- Θηκογράμματα – Normal QQ Plot

Εργαλεία ESDA και Περιγραφικής Στατιστικής

- **Χωροπληθείς χάρτες (choropleth maps)**

- Ορισμός

  - Οι χωροπληθείς χάρτες είναι θεματικοί χάρτες ομαδοποίησης που εστιάζουν στην ταξινόμηση ενός συνόλου δεδομένων σε διακριτές ομάδες, ανάλογα με την κατανομή που εμφανίζουν οι τιμές των δεδομένων.

  - Στους χωροπληθείς χάρτες η οπτική μεταβλητή που χρησιμοποιείται είναι η ένταση μιας απόχρωσης

  - Βασικό κριτήριο η ομοιογένεια μεταξύ των δεδομένων που κατατάσσονται σε κάθε ομάδα

Εργαλεία ESDA και Περιγραφικής Στατιστικής

- **Χωροπληθείς χάρτες (choropleth maps)**

- Γιατί χρησιμοποείται

  - Γραφική απεικόνιση της χωρικής κατανομής των τιμών μιας μεταβλητής

  - Ένας χάρτης choropleth παρέχει έναν διαισθητικό τρόπο οπτικοποίησης του τρόπου με τον οποίο μια συγκεκριμένη μεταβλητή (**όπως πυκνότητα πληθυσμού, εισόδημα** κ.λπ.) θα μπορούσε να ποικίλλει σε διαφορετικές γεωγραφικές περιοχές.

Εργαλεία ESDA και Περιγραφικής Στατιστικής

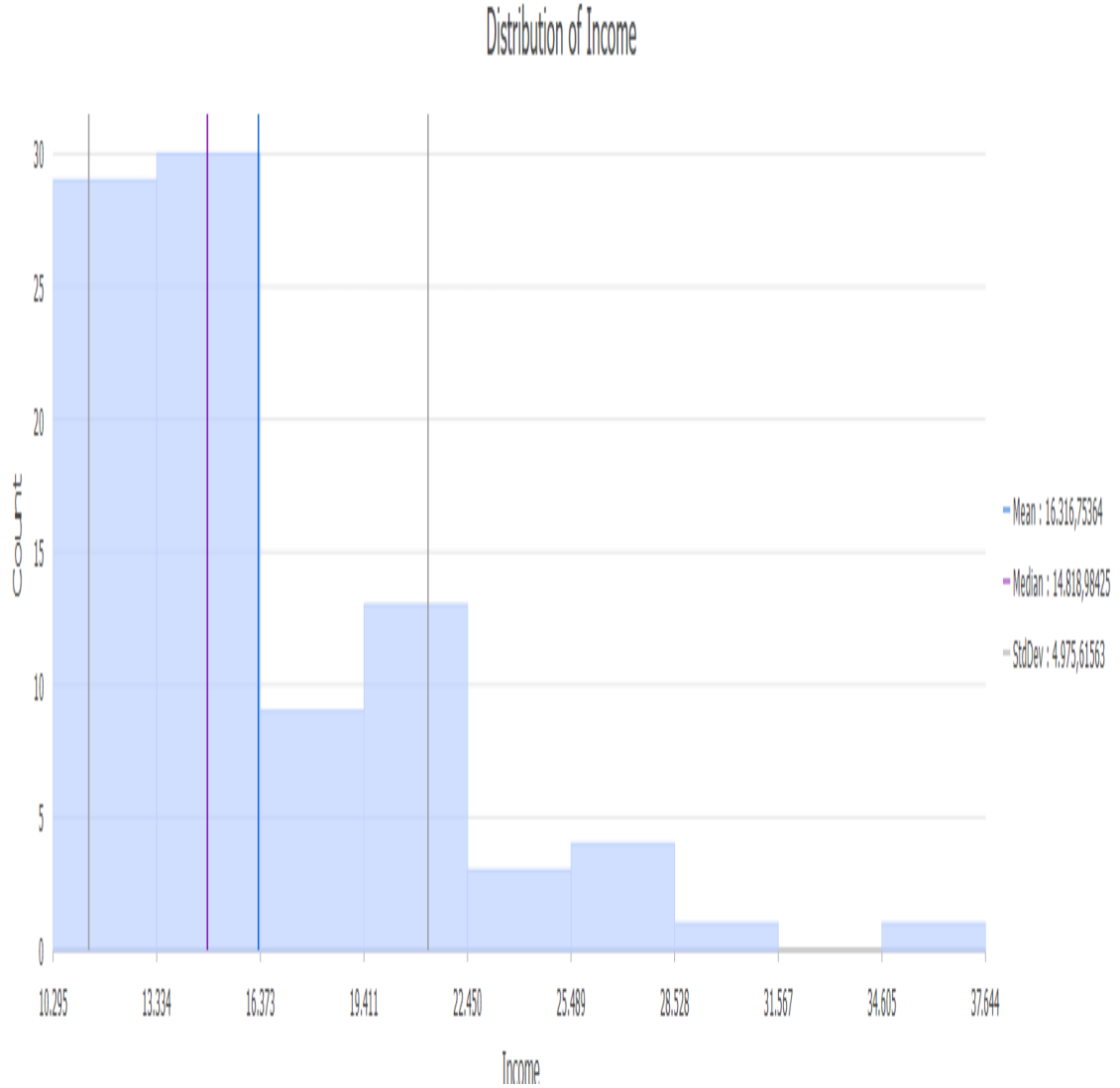
- **Χωροπληθείς χάρτες (choropleth maps)**

Statistics

	Dataset
Mean	16.316,8
Median	14.819,0
Std. Dev.	4.975,6
Rows	90
Count	90
Nulls	0
Min	10.294,9
Max	37.644,1
Sum	1.468.507,8
Skewness	1,63
Kurtosis	6,1

Εργαλεία ESDA και Περιγραφικής Στατιστικής

- **Χωροπληθείς χάρτες (choropleth maps)**



Εργαλεία ESDA και Περιγραφικής Στατιστικής

- **Χωροπληθείς χάρτες (choropleth maps)**
  - Τα είχαμε δει και με τους χάρτες της Ελλάδος.
  - **Ένα από τα σημαντικότερα προβλήματα** στη δημιουργία χωροπληθών χαρτών είναι η διαδικασία ομαδοποίησης των πολυπληθών δεδομένων σε ένα πεπερασμένο αριθμό διαδοχικών ομάδων ώστε να διαφοροποιηθούν οπτικά με ευκρίνεια από το μέσο απόδοσης, εφαρμόζοντας την οπτική μεταβλητή της έντασης μιας απόχρωσης.
  - Κατά την ομαδοποίηση τα δεδομένα πρέπει να ταξινομούνται με τέτοιο τρόπο, ώστε οι τιμές τους να παρουσιάζουν αφενός ομοιογένεια μέσα στις ομάδες, αφετέρου σημαντικές διαφορές μεταξύ των ομάδων.

Εργαλεία ESDA και Περιγραφικής Στατιστικής

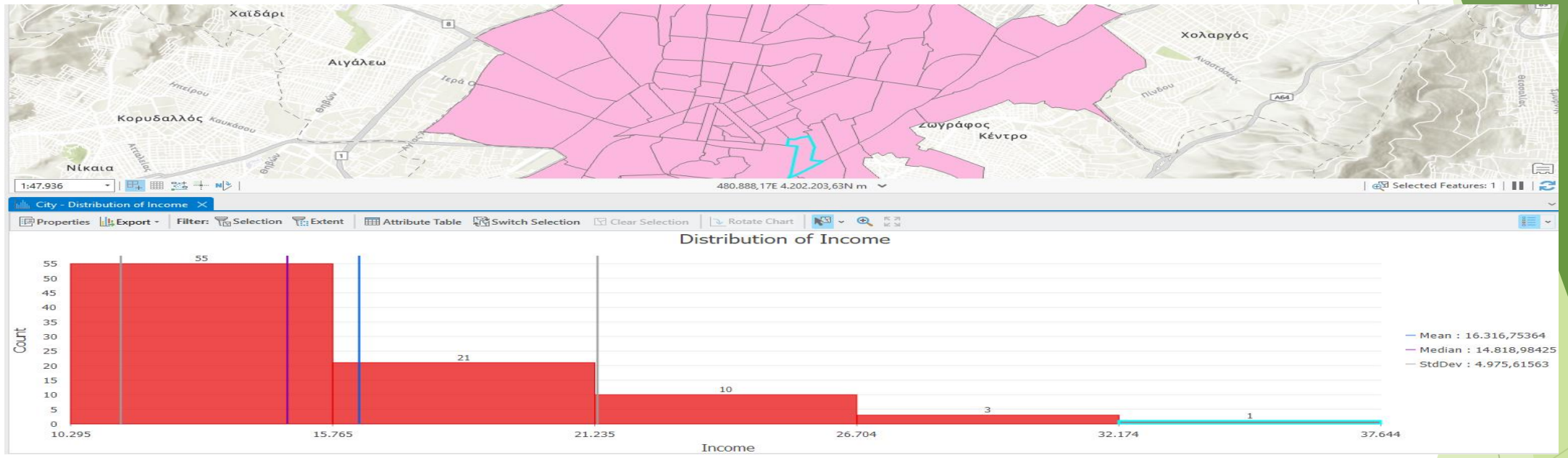
- **Χωροπληθείς χάρτες (choropleth maps)**

- Ένα από τα σημαντικότερα προβλήματα στη δημιουργία χωροπληθών χαρτών είναι η διαδικασία ομαδοποίησης των πολυπληθών δεδομένων σε ένα πεπερασμένο αριθμό διαδοχικών ομάδων ώστε να διαφοροποιηθούν οπτικά με ευκρίνεια από το μέσο απόδοσης, εφαρμόζοντας την οπτική μεταβλητή της έντασης μιας απόχρωσης.

- Κατά την ομαδοποίηση τα δεδομένα πρέπει να ταξινομούνται με τέτοιο τρόπο, ώστε οι τιμές τους να παρουσιάζουν αφενός ομοιογένεια μέσα στις ομάδες, αφετέρου σημαντικές διαφορές μεταξύ των ομάδων.

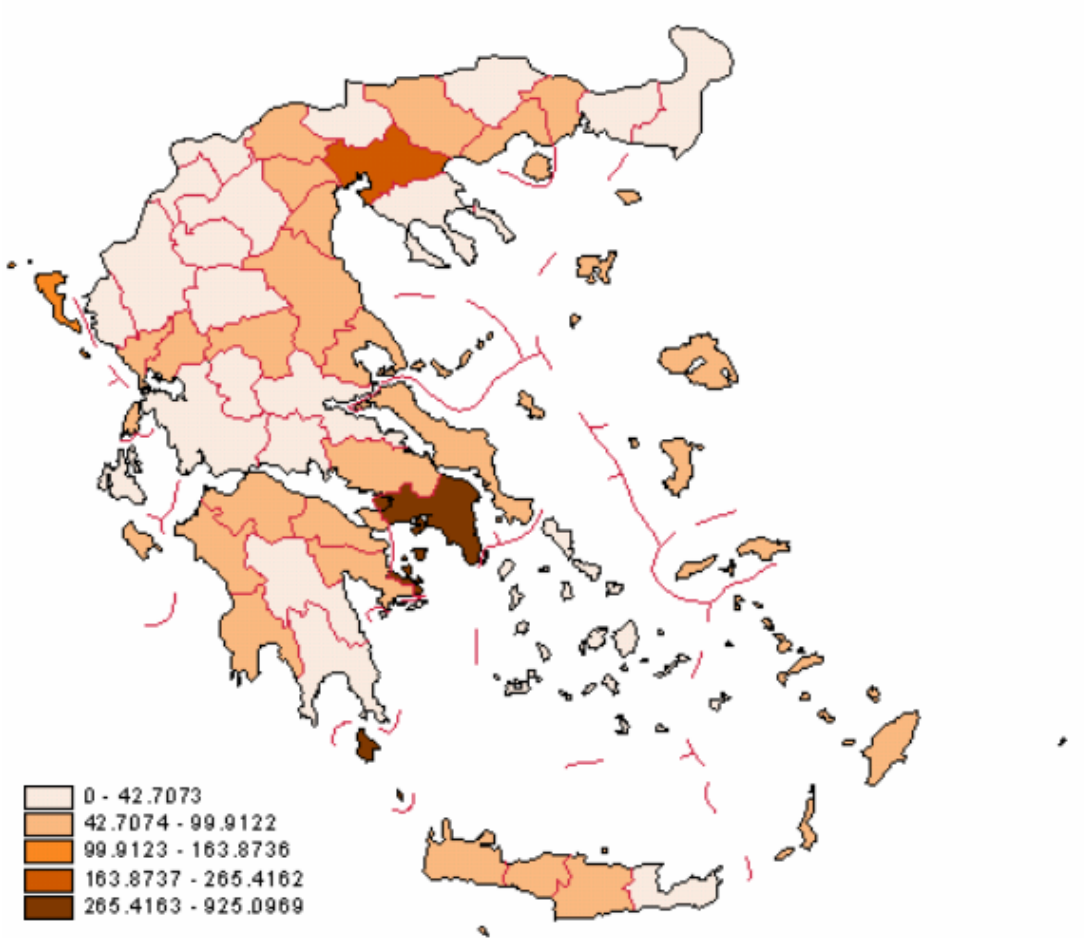
Εργαλεία ESDA και Περιγραφικής Στατιστικής

- **Χωροπληθείς χάρτες (choropleth maps)**

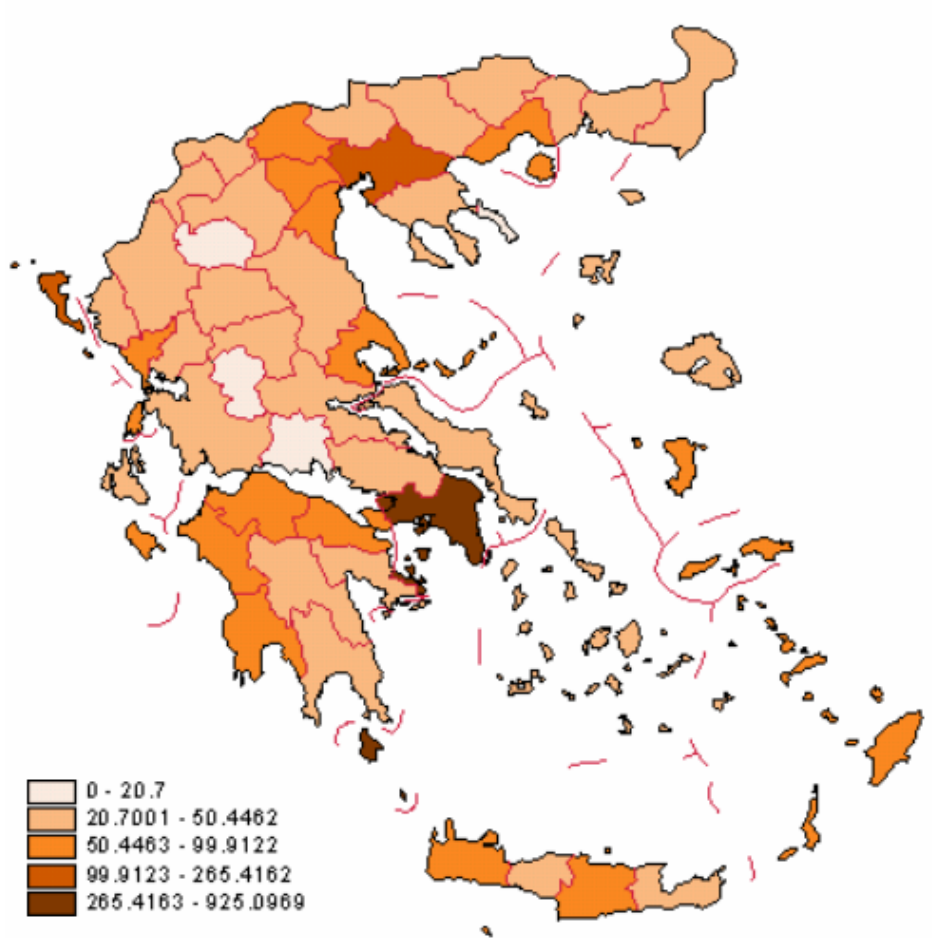


Εργαλεία ESDA και Περιγραφικής Στατιστικής

- **Χωροπληθείς χάρτες (choropleth maps)**



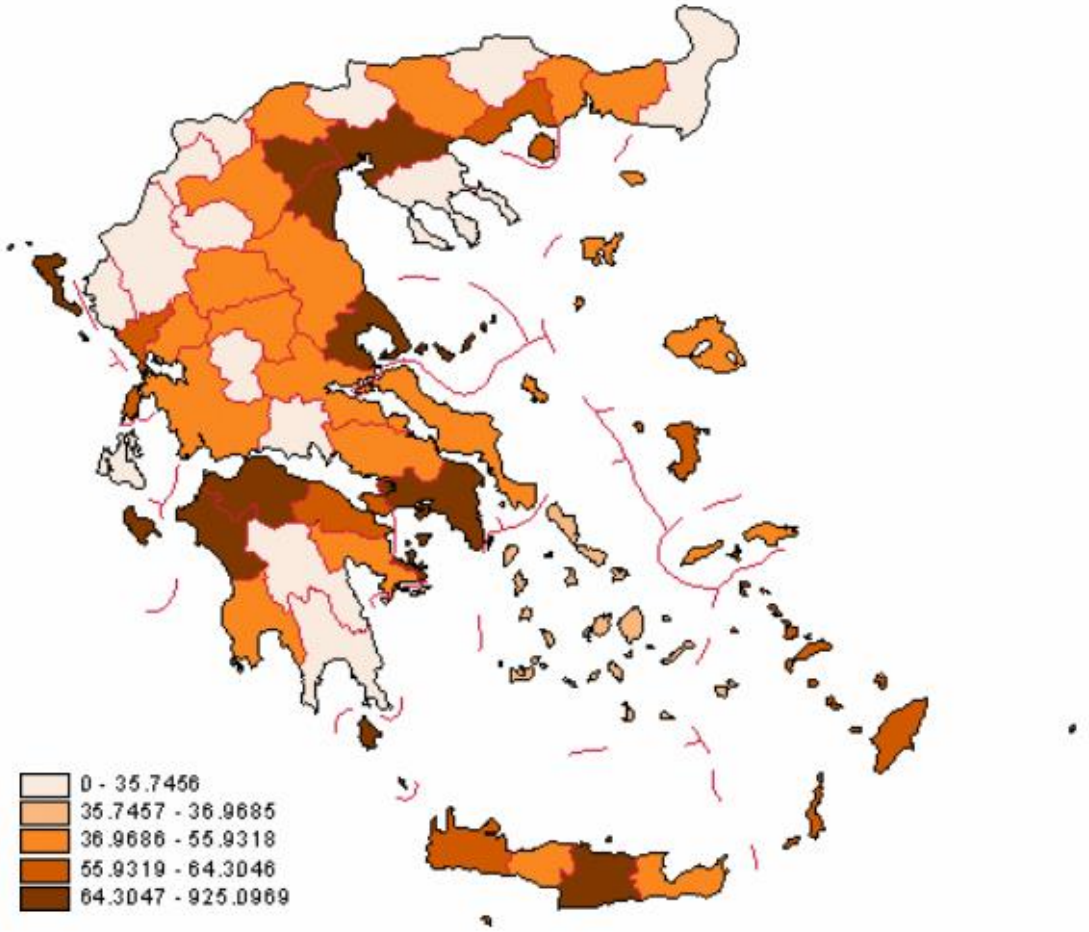
***Χάρτης 3**: Μέθοδος βέλτιστης προσαρμογής της απόλυτης απόκλισης (GADF)*



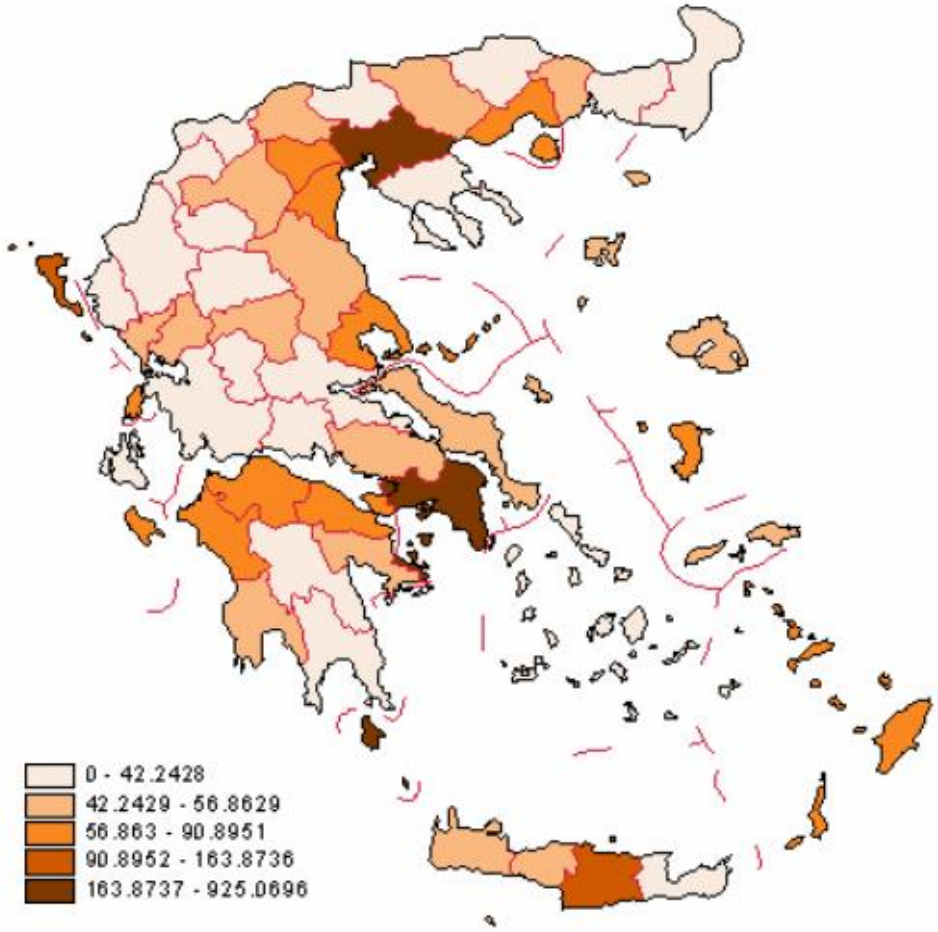
***Χάρτης 4**: Μέθοδος των φυσικών διακοπών*

Εργαλεία ESDA και Περιγραφικής Στατιστικής

- **Χωροπληθείς χάρτες (choropleth maps)**



***Χάρτης 5**: Μέθοδος κανονικής τμηματοποίησης*



***Χάρτης 6**: Μέθοδος του αλγόριθμου SOM*

Διερευνητική Ανάλυση Χωρικών Δεδομένων (ESDA)

- **Εργαλεία ESDA και Περιγραφικής Στατιστικής**

  για οπτικοποίηση μονομεταβλητών χωρικών δεδομένων

- Χωροπληθείς χάρτες (*choropleth maps*)
- Κατανομή συχνοτήτων και ιστογράμματα
- Μέτρηση του κέντρου, διασποράς και σχήματος της κατανομής
- Μέτρηση της τάσης / Μεταβλητότητας
- Αναγνώριση ακραίων τιμών
- Θηκογράμματα – Normal QQ Plot

Εργαλεία ESDA και Περιγραφικής Στατιστικής

- **Κατανομή συχνοτήτων και ιστογράμματα**

  Ορισμός

- Τρεις έννοιες
  - **Συχνότητα**: πόσες φορές εμφανίζεται μια κατηγορία στο dataset
  - **Σχετική συχνότητα**: Η αναλογία (%) των παρατηρήσεων που ανήκουν σε μια κατηγορία. Χρησιμοποιείται για την κατανόηση πως ένα δείγμα (ή πληθυσμός) κατανέμεται ανάμεσα στο εύρος των τιμών = συχνότητα / n
  - **Αθροιστική σχετική συχνότητα**: Σε κάθε γραμμή προστίθεται η σχετική συχνότητα από την τιμή αυτή και πάνω. Μας λέει το ποσοστό του πληθυσμού ως εκείνο το σημείο.

Διερευνητική Ανάλυση Χωρικών Δεδομένων (ESDA)

- **Εργαλεία ESDA και Περιγραφικής Στατιστικής**

  για οπτικοποίηση μονομεταβλητών χωρικών δεδομένων

- Χωροπληθείς χάρτες (*choropleth maps*)
- Κατανομή συχνοτήτων και ιστογράμματα
- Μέτρηση του κέντρου, διασποράς και σχήματος της κατανομής
- Μέτρηση της τάσης / Μεταβλητότητας
- Αναγνώριση ακραίων τιμών
- Θηκογράμματα – Normal QQ Plot

Εργαλεία ESDA και Περιγραφικής Στατιστικής

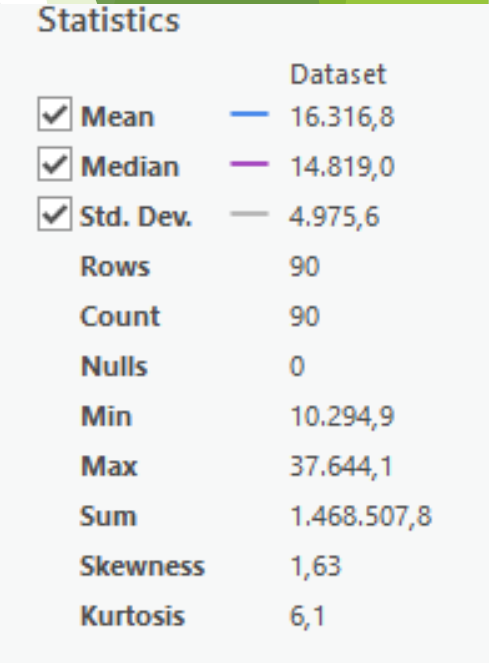
- **Μέτρηση του κέντρου, διασποράς και σχήματος της κατανομής**

- Η μέτρηση της κεντρικής τάσης, παρέχει πληροφορίες για το που βρίσκεται το κέντρο μιας κατανομής.

  - **Μέση** (απλή αριθμητική μέση τιμή)

  - **Διάμεσος** (διαμερισμός στη μέση ανάμεσα στις μικρότερες και μεγαλύτερες τιμές.

  - **Σχήμα κατανομής**
    - **Λοξότητα - Skewness** = μέτρηση της ασυμμετρίας της κατανομής γύρω από το μέσο,
    - **Κύρτωση - Kurtosis** = ο βαθμός της κορύφωσης ή ομαλότητας μιας κατανομής τιμών



Διερευνητική Ανάλυση Χωρικών Δεδομένων (ESDA)

- **Εργαλεία ESDA και Περιγραφικής Στατιστικής**

  για οπτικοποίηση μονομεταβλητών χωρικών δεδομένων

- Χωροπληθείς χάρτες (*choropleth maps*)
- Κατανομή συχνοτήτων και ιστογράμματα
- Μέτρηση του κέντρου, διασποράς και σχήματος της κατανομής
- Μέτρηση της τάσης / Μεταβλητότητας
- Αναγνώριση ακραίων τιμών
- Θηκογράμματα – Normal QQ Plot

Εργαλεία ESDA και Περιγραφικής Στατιστικής

- **Μέτρηση της τάσης / Μεταβλητότητας**

- Η μέτρηση της τάσης /μεταβλητότητας, παρέχει πληροφορίες πόσο οι τιμές μιας μεταβλητής διαφέρουν μεταξύ τους και σε σχέση με τη μέση τιμή.

  - **Εύρος**
  - **Απόκλιση από το μέσο (deviation)**
  - **Διακύμανση (variation)**
  - **Τυπική απόκλιση**
  - **Εκατοστημόρια, Τεταρτημόρια, Ποσοστιαία (Percentiles, Quartilies and Quantiles)**

Διερευνητική Ανάλυση Χωρικών Δεδομένων (ESDA)

- **Εργαλεία ESDA και Περιγραφικής Στατιστικής**

  για οπτικοποίηση μονομεταβλητών χωρικών δεδομένων

- Χωροπληθείς χάρτες (*choropleth maps*)
- Κατανομή συχνοτήτων και ιστογράμματα
- Μέτρηση του κέντρου, διασποράς και σχήματος της κατανομής
- Μέτρηση της τάσης / Μεταβλητότητας
- Αναγνώριση ακραίων τιμών
- Θηκογράμματα – Normal QQ Plot

Εργαλεία ESDA και Περιγραφικής Στατιστικής

- **Αναγνώριση ακραίων τιμών**
  - Οι ακραίες τιμές μιας κατανομής
- **Γιατί είναι χρήσιμες**
  - Λάθος μετρήσεις
  - Οι ακραίες τιμές παραμορφώνουν πολλά στατιστικά αποτελέσματα
  - Μπορεί να κρύβουν σημαντική πληροφορία που αξίζει να ερευνηθεί

Διερευνητική Ανάλυση Χωρικών Δεδομένων (ESDA)

- **Εργαλεία ESDA και Περιγραφικής Στατιστικής**

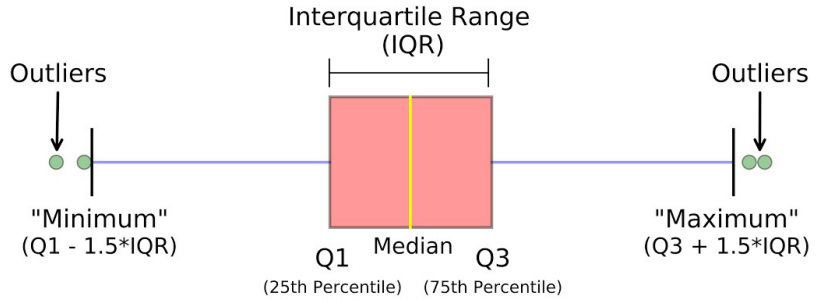
  για οπτικοποίηση μονομεταβλητών χωρικών δεδομένων

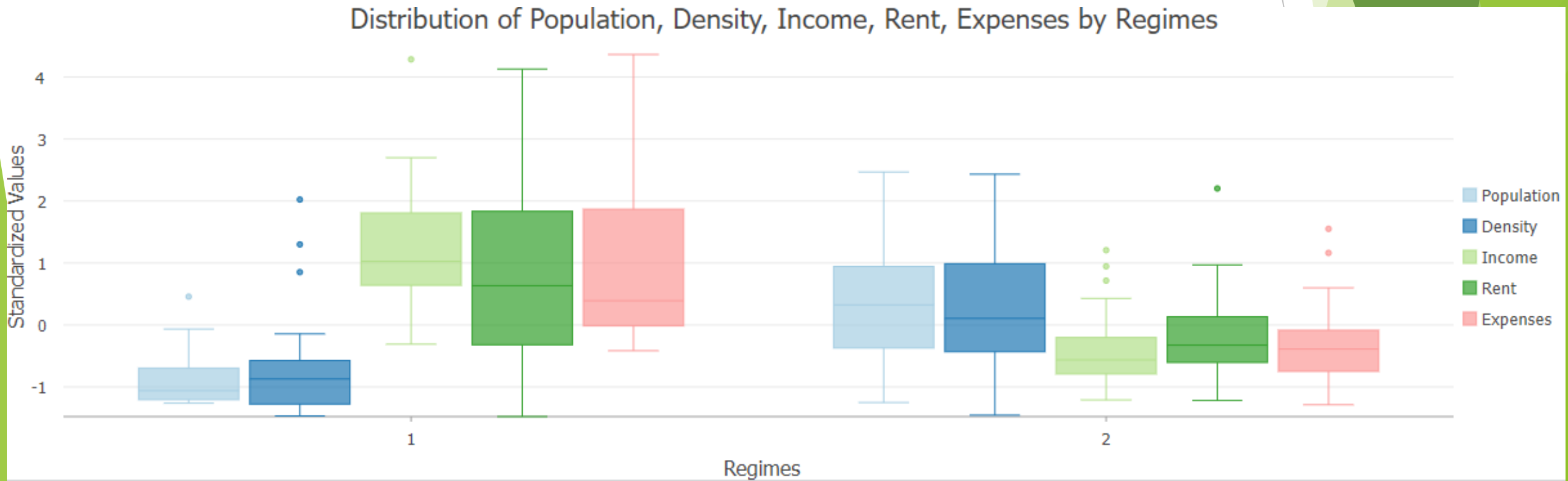
- Χωροπληθείς χάρτες (*choropleth maps*)
- Κατανομή συχνοτήτων και ιστογράμματα
- Μέτρηση του κέντρου, διασποράς και σχήματος της κατανομής
- Μέτρηση της τάσης / Μεταβλητότητας
- Αναγνώριση ακραίων τιμών
- Θηκογράμματα – Normal QQ Plot

Εργαλεία ESDA και Περιγραφικής Στατιστικής

- **Θηκογράμματα**

Ορισμός: η γραφική αναπαράσταση της κύριας περιγραφικής στατιστικής μιας κατανομής



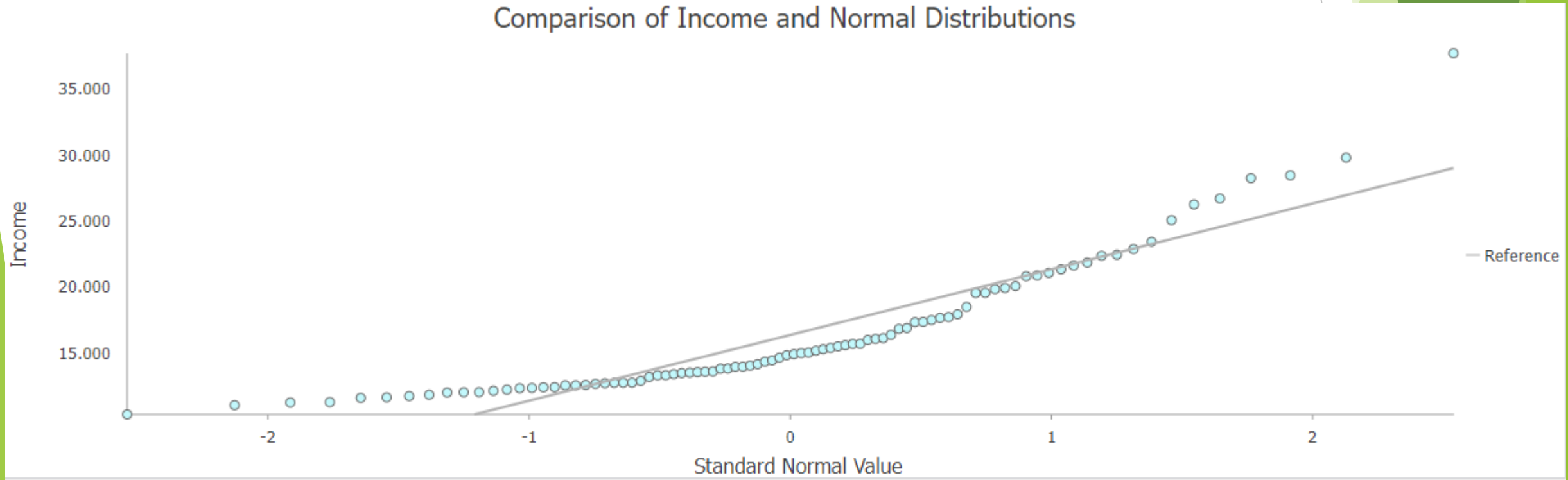


Εργαλεία ESDA και Περιγραφικής Στατιστικής

- **Normal QQ Plot**

Ορισμός: είναι η γραφική τεχνική που σχεδιάζει δεδομένα με βάση μια θεωρητική κανονική κατανομή

- Που χρειάζεται: Για την αναγνώριση αν τα δεδομένα ακολουθούν κανονική κατανομή

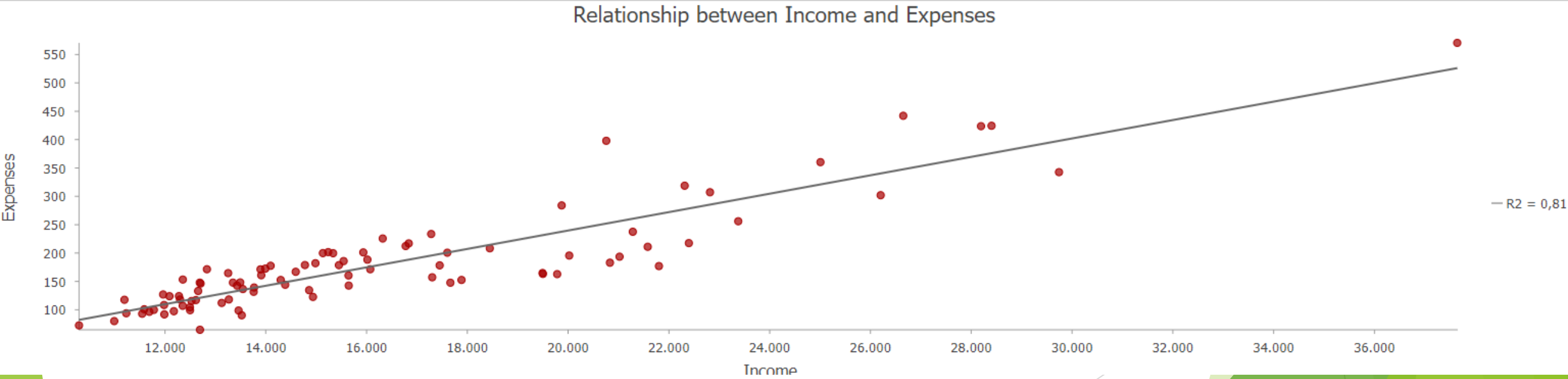


Εργαλεία ESDA και Περιγραφικής Στατιστικής

- **Διμεταβλητές Αναλύσεις**
  - Scatter plot - Scatter plot matrix
  - Covariance, Variance – Covariance matrix
  - Correlation coefficient
  - Pairwise correlation
  - General QQ plot

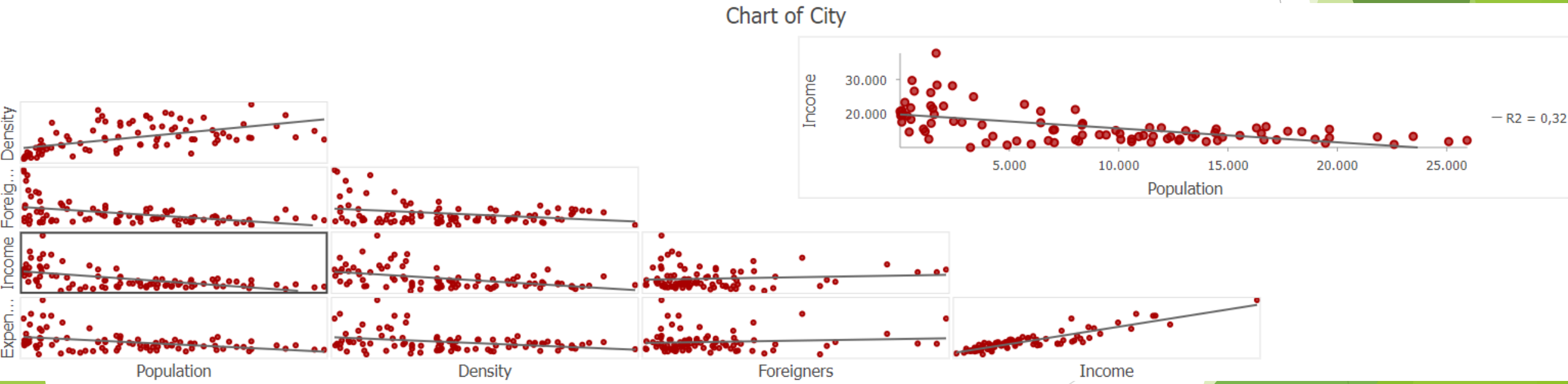
Εργαλεία ESDA και Περιγραφικής Στατιστικής

- **Scatter plot - Scatter plot matrix**
  - Χρησιμοποιείται για την αναγνώριση της σχέσης ανάμεσα σε δύο μεταβλητές και ανίχνευση πιθανών ακραίων τιμών



Εργαλεία ESDA και Περιγραφικής Στατιστικής

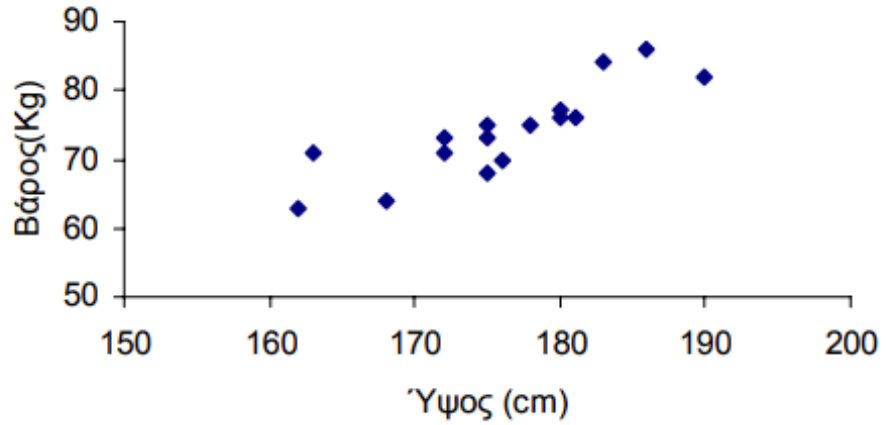
- **Scatter plot - Scatter plot matrix**
  - Χρησιμοποιείται για την αναγνώριση της σχέσης ανάμεσα σε δύο μεταβλητές και ανίχνευση πιθανών ακραίων τιμών



Συσχέτιση

Για τη εύρεση της σχέσης δύο (ή και περισσοτέρων) μεταβλητών υπάρχουν δείκτες, που υπολογίζουν κάποιο συντελεστή που προσδιορίζει το είδος (+ ή -) και το βαθμός της συσχέτισης των μεταβλητών.

Ένας απλός τρόπος για να αποκτήσουμε μια πρώτη ιδέα για το αν και πώς δυο μεταβλητές συσχετίζονται, είναι να κατασκευάσουμε το διάγραμμα διασποράς (*Scatter Diagram*)



Πόσο ισχυρή είναι όμως αυτή η συσχέτιση;

Πώς μπορεί να μετρηθεί;

Συντελεστής Γραμμικής Συσχέτισης του Pearson

Ο δειγματικός συντελεστής γραμμικής συσχέτισης του Pearson συμβολίζεται με r και ορίζεται από τον τύπο:

$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

όπου,

$$s_{xy} = Cov(X,Y) = \frac{\sum_{i=1}^{v}(x_i - \bar{x})\cdot(y_i - \bar{y})}{v-1} = \frac{\sum_{i=1}^{v} x_i y_i - v\cdot\bar{x}\cdot\bar{y}}{v-1}$$

$$s_x = \sqrt{\frac{1}{v-1}\sum_{i=1}^{v}(x_i - \bar{x})^2} \quad \text{και} \quad s_y = \sqrt{\frac{1}{v-1}\sum_{i=1}^{v}(y_i - \bar{y})^2}$$

Επομένως

$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\sum_{i=1}^{v}(x_i - \bar{x})\cdot(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{v}(x_i - \bar{x})^2}\cdot\sqrt{\sum_{i=1}^{v}(y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{v} x_i y_i - v\cdot\bar{x}\cdot\bar{y}}{\sqrt{\sum_{i=1}^{v} x_i^2 - v\cdot\bar{x}^2}\cdot\sqrt{\sum_{i=1}^{v} y_i^2 - v\cdot\bar{y}^2}}$$

Ο συντελεστή γραμμικής συσχέτισης r δίνει ένα μέτρο του μεγέθους της γραμμικής συσχέτισης μεταξύ δύο μεταβλητών. Το εύρος των τιμών που μπορεί να πάρεις είναι από -1 έως 1 [-1, 1]

Συσχέτιση

Αν $r = \pm 1$ υπάρχει **τέλεια γραμμική** συσχέτιση.
Αν $-0,3 \le r < 0,3$ **δεν υπάρχει γραμμική** συσχέτιση. Αυτό, όμως, δεν σημαίνει ότι δεν υπάρχει άλλου είδους συσχέτιση μεταξύ των δύο μεταβλητών.
Αν $-0,5 < r \le -0,3$ ή $0,3 \le r < 0,5$ υπάρχει **ασθενής γραμμική** συσχέτιση.
Αν $-0,7 < r \le -0,5$ ή $0,5 \le r < 0,7$ υπάρχει **μέση γραμμική** συσχέτιση.
Αν $-0,8 < r \le -0,7$ ή $0,7 \le r < 0,8$ υπάρχει **ισχυρή γραμμική** συσχέτιση.
Αν $-1 < r \le -0,8$ ή $0,8 \le r < 1$ υπάρχει **πολύ ισχυρή γραμμική** συσχέτιση.

Θετικές τιμές του r δεν δηλώνουν κατ' ανάγκη και υψηλότερο βαθμό συσχέτισης απ΄ ότι αν ήταν αρνητικές!!!

Ο βαθμός γραμμικής συσχέτισης καθορίζεται από την απόλυτη τιμή του r και όχι από το πρόσημό του. Το τελευταίο καθορίζει μόνο το είδος της σχέσης (αν είναι θετική ή αρνητική)…

Δηλ. η τιμή r = -0.9 δείχνει ισχυρότερη γραμμική συσχέτιση από την τιμή r = 0.7, ενώ οι τιμές r = -0.5 και r = 0.5 δείχνουν τον ίδιο.

Τι σημαίνει όμως «το είδος της σχέσης..?»

Συσχέτιση

Συντελεστής Γραμμικής Συσχέτισης του Spearman

Όταν οι μεταβλητές είναι τακτικής κλίμακας χρησιμοποιείται ο μη παραμετρικός συντελεστής συσχέτισης Spearman

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}.$$

- Έχει παρόμοιες ιδιότητες με αυτόν του Pearson
- Χρησιμοποιείται συχνά όταν έχουμε ποιοτικά διατάξιμα χαρακτηριστικά (ή απαντήσεις στην κλίμακα 1-2-3-4-5 – καθόλου, λίγο, μέτρια, πολύ, πάρα πολύ)
- Μπορεί να χρησιμοποιηθεί και στα ποσοτικά, κυρίως σε **περιπτώσεις που δεν έχουμε συμμετρικές κατανομές** *(ωπ..μου θυμίζει κάτι από το προηγούμενο μάθημα?)*

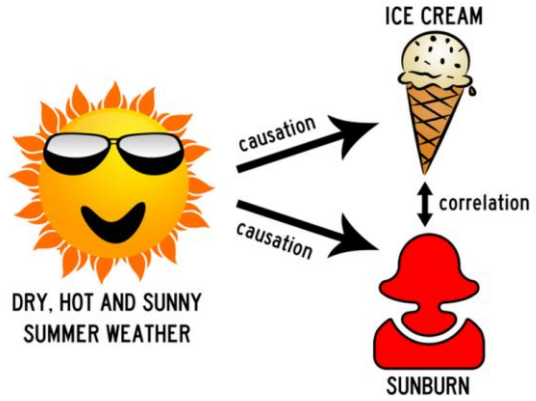
Συσχέτιση

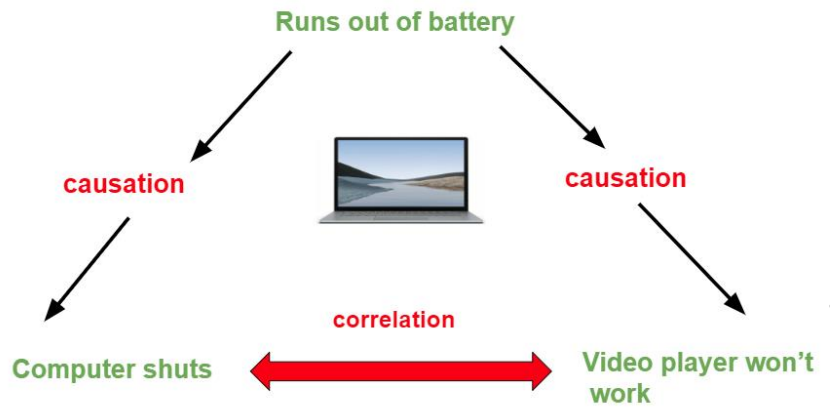
Συσχέτιση δε σημαίνει αιτιότητα

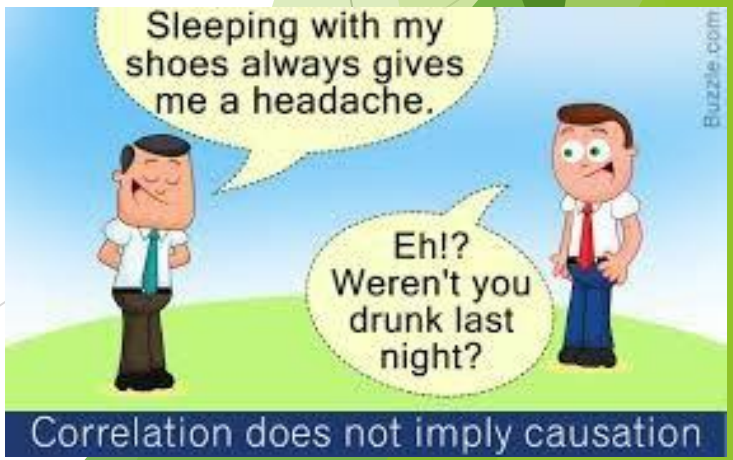
Όταν σε μια μη πειραματική έρευνα (δειγματοληψία) δύο μεταβλητές Χ και Υ βρίσκονται συσχετισμένες αυτό σημαίνει μόνο ότι οι μεταβλητές αυτές συνδέονται με κάποια σχέση. Δε συνεπάγεται, κατ' ανάγκη, αιτιότητα. Οι δύο μεταβλητές μπορεί βεβαία να συνδέονται με σχέση αιτιότητας, μπορεί όμως, όχι.

- *Παρατηρήθηκε ότι το ύψος των μαθητών ενός σχολείου, ηλικίας 6 έως 13 ετών, έχει ισχυρή θετική γραμμική συσχέτιση με την αντιληπτική ικανότητα των μαθητών. Προφανώς η αντιληπτική ικανότητα των μαθητών δεν επηρεάζεται από το ύψος τους. Απλώς τόσο η πνευματική όσο και η φυσική ανάπτυξη των μικρών μαθητών επηρεάζονται παράλληλα από άλλους παράγοντες.*

- *Παρατηρήθηκε ότι οι πωλήσεις ταχύπλοων στο Sidney είχαν, για μια μακρά περίοδο, ισχυρή θετική συσχέτιση με τις πωλήσεις έγχρωμων τηλεοράσεων στη Melbourne. Προφανώς, τόσο οι πωλήσεις ταχύπλοων όσο και οι πωλήσεις έγχρωμων τηλεοράσεων ήταν συνάρτηση γενικότερων ευνοϊκών οικονομικών παραγόντων.*







Στατιστική Σημαντικότητα

Στατιστική σημαντικότητα (ή ένα στατιστικά σημαντικό αποτέλεσμα) επιτυγχάνεται όταν το p-value είναι μικρότερο από το επίπεδο σημαντικότητας.

Ως θέμα της ορθής επιστημονικής μεθόδου-πρακτικής, ένα επίπεδο σημαντικότητας επιλέγεται πριν τη συλλογή δεδομένων και συνήθως βρίσκεται στο 0,05 (5%). Άλλα επίπεδα σημαντικότητας (π.χ., 0,01) δύνανται να βρουν εφαρμογή, ανάλογα με τον τομέα μελέτης.

Values of p	Inference
$p > 0.10$	No evidence against the null hypothesis.
$0.05 < p < 0.10$	Weak evidence against the null hypothesis
$0.01 < p < 0.05$	Moderate evidence against the null hypothesis
$0.05 < p < 0.001$	Good evidence against null hypothesis.
$0.001 < p < 0.01$	Strong evidence against the null hypothesis
$p < 0.001$	Very strong evidence against the null hypothesis

Στατιστική Σημαντικότητα

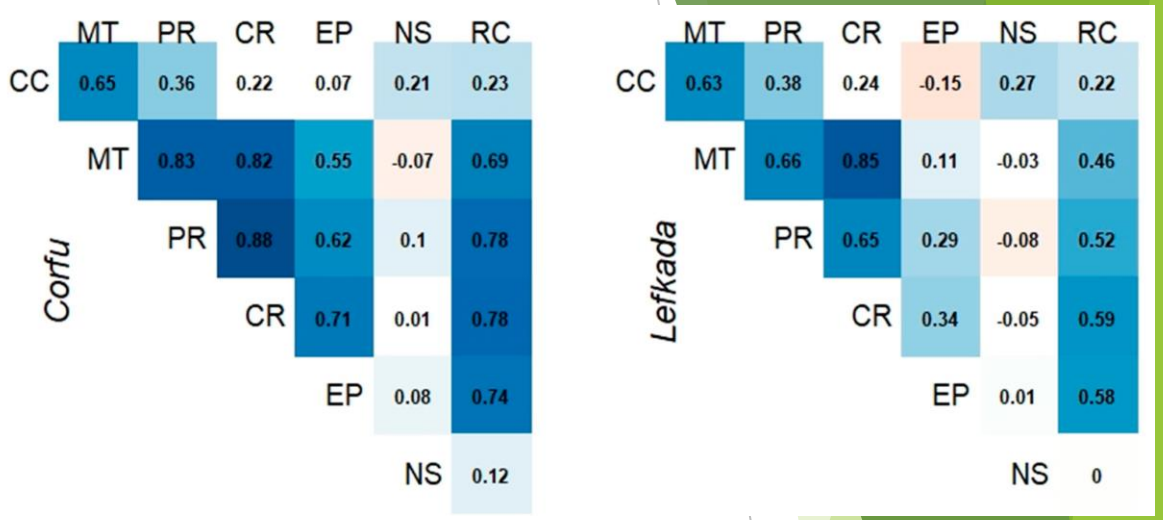
Η στατιστική σημαντικότητα είναι θεμελιώδης για την δοκιμή στατιστικής υπόθεσης. Σε κάθε πείραμα ή παρατήρηση που περιλαμβάνει τη σύνταξη ενός δείγματος από έναν πληθυσμό, υπάρχει πάντα η πιθανότητα ότι ένα παρατηρούμενο αποτέλεσμα θα συνέβαινε λόγω σφάλματος δειγματοληψίας μόνο.

Αλλά αν η p-τιμή είναι μικρότερη από το επίπεδο σημαντικότητας (π.χ., $p < 0,05$), **τότε ο ερευνητής μπορεί να συμπεράνει ότι η παρατηρούμενη επίδραση αντανακλά στην πραγματικότητα τα χαρακτηριστικά του πληθυσμού και όχι μόνο δειγματοληπτικό σφάλμα.** Ένας ερευνητής μπορεί στη συνέχεια να αναφέρει ότι το αποτέλεσμα επιτυγχάνει στατιστική σημαντικότητα!

Η σημερινή έννοια της στατιστικής σημαντικότητας ξεκίνησε με τον Ronald Fisher, όταν αναπτύχθηκαν δοκιμές στατιστικών υποθέσεων με βάση p-value στις αρχές του 20ου αιώνα. Ήταν οι Jerzy Neyman και Egon Pearson οι οποίοι αργότερα συνέστησαν το επίπεδο σημαντικότητας να οριστεί εκ των προτέρων, πριν από κάθε συλλογή δεδομένων.

Συσχέτιση

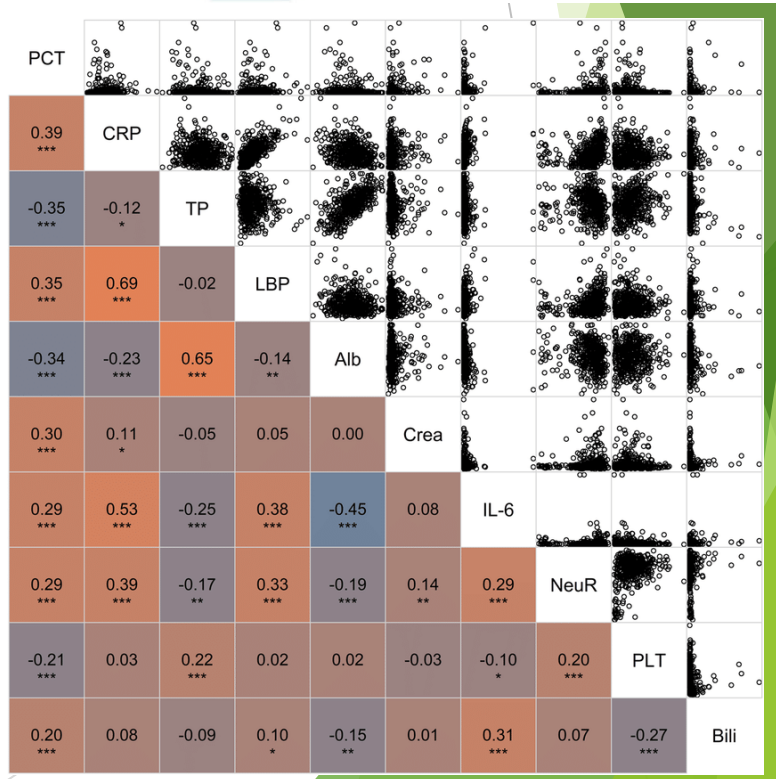
	Hours spent studying	Exam score	IQ score	Hours spent sleeping	School rating
Hours spent studying	1.00	0.82	0.48	-0.22	0.36
Exam score	0.82	1.00	0.33	-0.04	0.23
IQ score	0.08	0.33	1.00	0.06	0.02
Hours spent sleeping	-0.22	-0.04	0.06	1.00	0.12
School rating	0.36	0.23	0.02	0.12	1.00



Correlations

		reading score	writing score	math score	science score	female
reading score	Pearson Correlation[a]	1	.597**	.662**	.630**	-.053
	Sig. (2-tailed)[b]	.	.000	.000	.000	.455
	N[c]	200	200	200	200	200
writing score	Pearson Correlation	.597**	1	.617**	.570**	.256**
	Sig. (2-tailed)	.000	.	.000	.000	.000
	N	200	200	200	200	200
math score	Pearson Correlation	.662**	.617**	1	.631**	-.029
	Sig. (2-tailed)	.000	.000	.	.000	.680
	N	200	200	200	200	200
science score	Pearson Correlation	.630**	.570**	.631**	1	-.128
	Sig. (2-tailed)	.000	.000	.000	.	.071
	N	200	200	200	200	200
female	Pearson Correlation	-.053	.256**	-.029	-.128	1
	Sig. (2-tailed)	.455	.000	.680	.071	.
	N	200	200	200	200	200

**. Correlation is significant at the 0.01 level (2-tailed).



Εργαλεία ESDA και Περιγραφικής Στατιστικής

- **ΣΤΟ ΕΠΟΜΕΝΟ (24/11/2022)**

- **ΑΝΑΛΥΣΗ ΓΕΩΓΡΑΦΙΚΩΝ ΚΑΤΑΝΟΜΩΝ ΚΑΙ ΣΗΜΕΙΑΚΩΝ ΠΡΟΤΥΠΩΝ**

Εργαλεία ESDA και Περιγραφικής Στατιστικής

- **ΩΡΑ ΓΙΑ ΕΦΑΡΜΟΓΗ**
  - **Έχετε όλοι τον φάκελο “Data\Data_GREKOUSIS” που έχει διάφορα shp αρχεία.σ**
  - **Να έχετε αυτό τον φάκελο αποθηκευμένο στο δίσκο του υπολογιστή σας**

Εργαλεία ESDA και Περιγραφικής Στατιστικής

Σκοπός της εργασίας:

Ένας επενδυτής ψάχνει το κατάλληλο μέρος για να εγκαταστήσει ένα coffee μαγαζί και απευθύνεται σε εμάς (ως εταιρία) για συμβουλές και διαβούλευση.

Με εξαίρεση τα κόστη ενοικίασης, ο επενδυτής ενδιαφέρεται να βρει την κατάλληλη γειτονιά που να πληροί τα ακόλουθα χαρακτηριστικά:

1. Καθώς το μαγαζί θα είναι premium, θα πρέπει στην περιοχή οι κάτοικοι να έχουν **υψηλό εισόδημα**
2. Το μαγαζί να είναι σε περιοχή **με χαμηλή εγκληματικότητα**
3. Η ομάδα στόχοπς θα πρέπει να θέλει να ξοδεύει περισσότερα χρήματα από το μέσο όρο, συμπεριλαμβάνοντας έξοδα για καφέ. Θα πρέπει να βρεθούν οι high spenders
4. Οι κοινωνικο-οικονομικοί παράγοντες (drivers) πίσω από αυτούς τους ανθρώπους θα πρέπει να προσδιοριστούν

Εργαλεία ESDA και Περιγραφικής Στατιστικής

Σκοπός της εργασίας:

Το τελευταίο χαρακτηριστικό δεν σχετίζεται άμεσα με την εύρεση της κατάλληλης θέσης.

Εστιάζει στον προσδιορισμό των χωρικών σχέσεων που θα χρησιμοποιηθούν στη μοντελοποίηση. Για εκπαιδευτικούς λόγους, θα γίνει μόνο με βάση κοινωνικο-οικονομικές μεταβλητές

Θα κινηθούμε στην ανάλυση τριών σταδίων

What -> Where -> How/Why

- Η περιοχή μελέτης είναι η πόλη των Αθηνών και τα κοινωνικο-οικονομικά δεδομένα είναι της απογραφής του 2011.

Εργαλεία ESDA και Περιγραφικής Στατιστικής

- **Χωροπληθείς χάρτες (choropleth maps)**
  - **ΩΡΑ ΓΙΑ ΕΦΑΡΜΟΓΗ**
    - **Ας ξεκινήσουμε με το ArcPro**

Άσκηση 1: Γνωριμία με τα δεδομένα και την περιοχή μελέτης

Εργαλεία που θα χρησιμοποιήσουμε: Zoom tools, Table of content, Attribute table, Symbology, Normalization

Αναμενόμενα αποτελέσματα: Choropleth maps, Mapping ratios

Εργαλεία ESDA και Περιγραφικής Στατιστικής

- **Χωροπληθείς χάρτες (choropleth maps)**
  - **ΩΡΑ ΓΙΑ ΕΦΑΡΜΟΓΗ**
    - **Ας ξεκινήσουμε με το ArcPro**

Άσκηση 1. Γνωριμία με τα δεδομένα και την περιοχή μελέτης

Στόχος: Χαρτογράφηση του πληθυσμού και πως μπορεί να υπολογιστεί και αποδοθεί η πληθυσμιακή πυκνότητα

Εργαλεία ESDA και Περιγραφικής Στατιστικής

- **Χωροπληθείς χάρτες (choropleth maps)**
  - **ΩΡΑ ΓΙΑ ΕΦΑΡΜΟΓΗ**
    - **Βήματα - tasks**

Χαρτογράφηση του πληθυσμού

- Χαρτογράφηση του πληθυσμού σε πέντε κατηγορίες με βάση τα διαστήματα

  > **2000 > 5000 > 10.000 > 15.000 > 30.000 >**

- Color ramp: **Yellow to Brown**

- Number format: **2 decimals**

Εργαλεία ESDA και Περιγραφικής Στατιστικής

- **Χωροπληθείς χάρτες (choropleth maps)**
  - **ΩΡΑ ΓΙΑ ΕΦΑΡΜΟΓΗ**
    - **Βήματα - tasks**

Υπολογισμός και χαρτογράφηση της πληθυσμιακής πυκνότητας

- Κανονικοποίηση (normalization) του πληθυσμού με βάση την έκταση (area)
- Χαρτογράφηση του πληθυσμού σε τέσσερις κατηγορίες με βάση τα νέα διαστήματα

  > 0,01 > 0,02 > 0,03 > 1 (pop/sqm) = 1 άνθρωπος / 100m2
- Color ramp: **Light green to Dark green**
- **Να σωθεί ως αρχείο Layers για να διατηρηθεί το sympology**

Εργαλεία ESDA και Περιγραφικής Στατιστικής

- **Χωροπληθείς χάρτες (choropleth maps)**
  - **ΩΡΑ ΓΙΑ ΕΦΑΡΜΟΓΗ**
    - **Βήματα - tasks**

Υπολογισμός και χαρτογράφηση της πληθυσμιακής πυκνότητας

- **Ερμηνεύστε το αποτέλεσμα του χωροπληθή χάρτη**
- **Πειραματιστείτε με άλλα παράγωγα της πληθυσμιακής πυκνότητας και κανονικοποίησης των δεδομένων**

Εργαλεία ESDA και Περιγραφικής Στατιστικής

- **Χωροπληθείς χάρτες (choropleth maps)**
  - **ΩΡΑ ΓΙΑ ΕΦΑΡΜΟΓΗ**
    - **Ας ξεκινήσουμε με το ArcPro**

Στόχοι

1. **Να βρεθούν περιοχές με υψηλό εισόδημα**
2. **Ν**

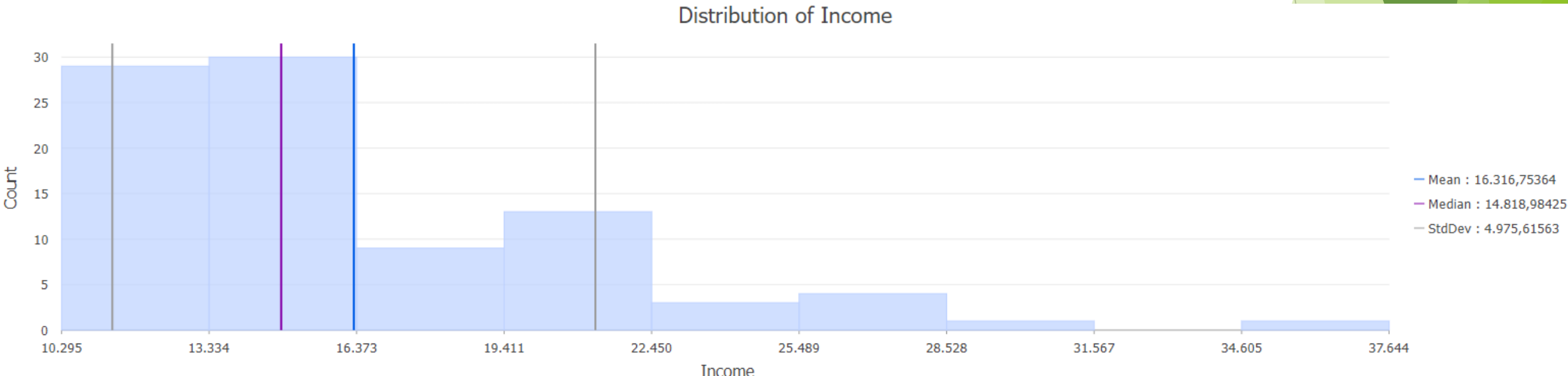
Εργαλεία ESDA και Περιγραφικής Στατιστικής

- **Χωροπληθείς χάρτες (choropleth maps)**
  - **ΩΡΑ ΓΙΑ ΕΦΑΡΜΟΓΗ**
    - **Ας ξεκινήσουμε με την ArcPro**
      - **Με βάση το αρχείο City, δημιουργήστε ένα χωροπληθή χάρτη με βάση το εισόδημα και τα κριτήρια:**
        - **Color Ramp: Yellow to Brown**
        - **Classes 4: Break values >15.000, >20.000, >25.000 > 40.000**
    - **Εξηγήστε τα αποτελέσματα**
    - **Βρείτε την περιοχή με εισόδημα > 20.000**

Εργαλεία ESDA και Περιγραφικής Στατιστικής

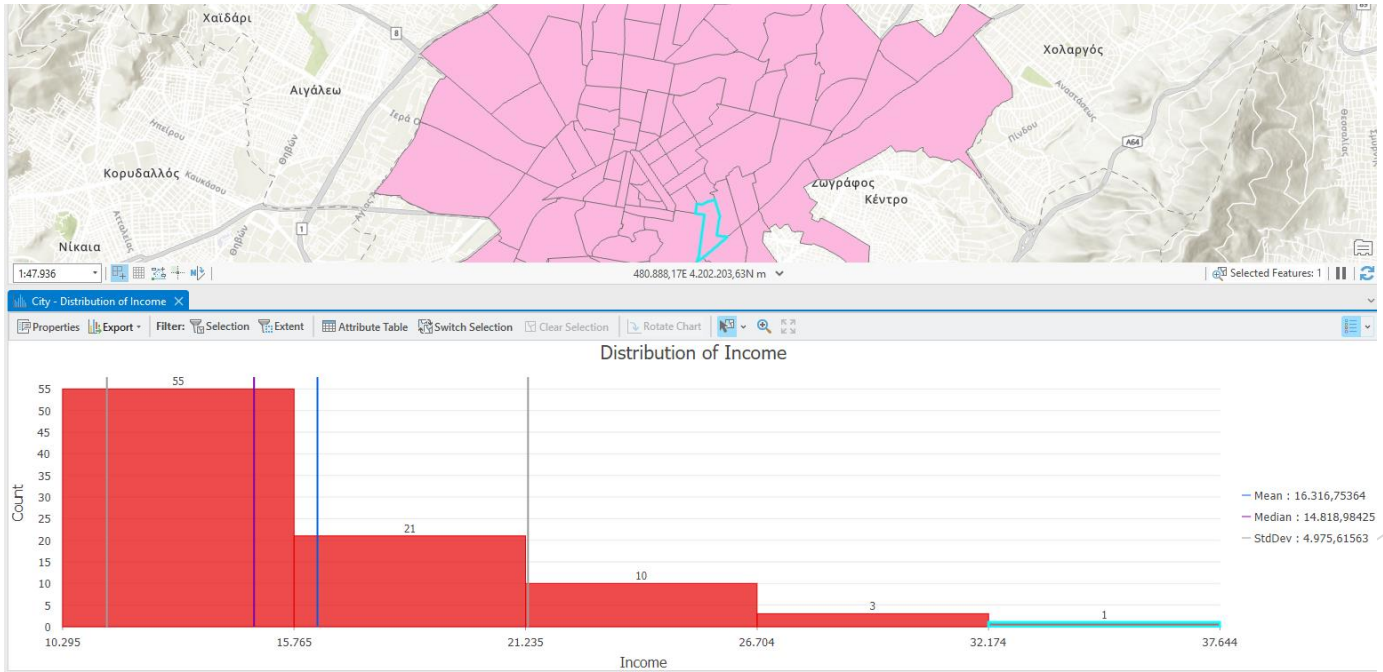
- **Ιστογράμματα**
  - Με βάση το αρχείο City

διαμορφώστε το σχετικό ιστόγραμμα του εισοδήματος, σε 5 bins και εμφανίστε τις στατιστικές τιμές «μέση τιμή, διάμεσος και τυπική απόκλιση)



Εργαλεία ESDA και Περιγραφικής Στατιστικής

- **Ιστογράμματα**
  - Πειραματίστε και με τα άλλα πεδία και αφού τα εξάγετε ως γραφικά αρχεία, περιγράψτε την πόλη των Αθηνών με βάση αυτά τα ιστογράμματα.
  - Με τη βοήθεια του ιστογράμματος του εισοδήματος, προσδιορίστε την περιοχή με το υψηλότερο εισόδημα και εξάγετε αυτή την περιοχή ως νέο αρχείο.



Εργαλεία ESDA και Περιγραφικής Στατιστικής

- Πειραματίστε και με τα στατιστικά και περιγράψτε τα δεδομένα σας