

Εξερεύνηση Δεδομένων

Μέρος 1

Fundamentals of Machine Learning for Predictive Data
Analytics

© John D. Kelleher and Brian Mac Namee and Aoife D'Arcy

Αθανάσιος Σάκκας, ΟΠΑ

- 1 **Η Αναφορά Ποιότητας Δεδομένων**
 - Μελέτη Περίπτωσης: Απάτη σε Ασφάλιση Αυτοκινήτου
- 2 **Γνωριμία με τα Δεδομένα**
- 3 **Εντοπισμός Ζητημάτων Ποιότητας Δεδομένων**
 - Μελέτη Περίπτωσης: Απάτη σε Ασφάλιση Αυτοκινήτου
- 4 **Αντιμετώπιση Ζητημάτων Ποιότητας Δεδομένων**
 - Αντιμετώπιση Ελλειπουσών Τιμών
 - Αντιμετώπιση Ακραίων Τιμών
 - Μελέτη Περίπτωσης: Απάτη σε Ασφάλιση Αυτοκινήτου
- 5 **Σύνοψη**

Η Αναφορά Ποιότητας Δεδομένων

- Μια αναφορά ποιότητας δεδομένων (data quality report) περιλαμβάνει πινακοποιημένες αναφορές που περιγράφουν τα χαρακτηριστικά κάθε feature σε έναν ABT, χρησιμοποιώντας τυπικά στατιστικά μέτρα **κεντρικής τάσης** και **διασποράς**.
- Οι πινακοποιημένες αναφορές συνοδεύονται από οπτικοποιήσεις δεδομένων:
 - Ένα **ιστόγραμμα** για κάθε συνεχές feature σε έναν ABT.
 - Ένα **ραβδόγραμμα** (bar plot) για κάθε κατηγορικό feature σε έναν ABT.

Πίνακας: Οι δομές των πινάκων που περιλαμβάνονται σε μία αναφορά ποιότητας δεδομένων για την περιγραφή (a) συνεχών features και (b) κατηγορικών features.

(α□) Συνεχή Χαρακτηριστικά (Features)

[illegible]

(β□) Κατηγορικά Χαρακτηριστικά (Features)

[illegible]

Μελέτη Περίπτωσης: Απάτη σε Ασφάλιση Αυτοκινήτου

Οι επόμενες διαφάνειες δείχνουν ένα τμήμα του ABT που έχει αναπτυχθεί για την ανίχνευση απάτης σε απαιτήσεις (claims) ασφάλισης αυτοκινήτου.

Αρχικά παρουσιάζεται ένα τμήμα του ABT που αναπτύχθηκε για αυτή τη λύση.

Πίνακας: Τμήματα του ABT για το πρόβλημα ανίχνευσης απάτης σε claims ασφάλισης αυτοκινήτου.

ID	Type	Inc.	Marital Status	Num Clmmts.	Injury Type	Hospital Stay	Claim Amnt.	Total Claimed	Num Claims	Num Soft Tiss.	% Soft Tiss.	Claim Amt Rcvd.	Fraud Flag
1	CI	0	Married	2	Soft Tissue	No	1,625	3250	2	2	1.0	0	1
2	CI	0		2	Back	Yes	15,028	60,112	1		0	15,028	0
3	CI	54,613		1	Broken Limb	No	-99,999	0	0	0	0	572	0
4	CI	0		4	Broken Limb	Yes	5,097	11,661	1	1	1.0	7,864	0
5	CI	0	Single	4	Soft Tissue	No	8869	0	0	0	0	0	1
6	CI	0		1	Broken Limb	Yes	17,480	0	0	0	0	17,480	0
7	CI	52,567		3	Broken Limb	No	3,017	18,102	2	1	0.5	0	1
8	CI	0		2	Back	Yes	7463	0	0	0	0	7,463	0
9	CI	0	Married	1	Soft Tissue	No	2,067	0	0	0	0	2,067	0
10	CI	42,300		4	Back	No	2,260	0	0	0	0	2,260	0
		:					:					:	
300	CI	0	Married	2	Broken Limb	No	2,244	0	0	0	0	2,244	0
301	CI	0		1	Broken Limb	No	1,627	92,283	3	0	0	1,627	0
302	CI	0		3	Serious	Yes	270,200	0	0	0	0	270,200	0
303	CI	0		1	Soft Tissue	No	7,668	92,806	3	0	0	7,668	0
304	CI	46,365	1	Back	No	3,217	0	0	0	0	1,653	0	
		:				:					:		
458	CI	48,176	Married	3	Soft Tissue	Yes	4,653	8,203	1	0	0	4,653	0
459	CI	0		1	Soft Tissue	Yes	881	51,245	3	0	0	0	1
460	CI	0	Divorced	3	Back	No	8,688	729,792	56	5	0.08	8,688	0
461	CI	47,371		1	Broken Limb	Yes	3,194	11,668	1	0	0	3,194	0
462	CI	0		1	Soft Tissue	No	6,821	0	0	0	0	0	1
		:					:					:	
491	CI	40,204	Single	1	Back	No	75,748	11,116	1	0	0	0	1
492	CI	0		1	Broken Limb	No	6,172	6,041	1	0	0	6,172	0
493	CI	0	Married	1	Soft Tissue	Yes	2,569	20,055	1	0	0	2,569	0
494	CI	31,951		1	Broken Limb	No	5,227	22,095	1	0	0	5,227	0
495	CI	0		2	Back	No	3,813	9,882	3	0	0	0	1
496	CI	0		1	Soft Tissue	No	2,118	0	0	0	0	0	1
497	CI	29,280	Married	4	Broken Limb	Yes	3,199	0	0	0	0	0	1
498	CI	0		1	Broken Limb	Yes	32,469	0	0	0	0	16,763	0
499	CI	46,683	Married	1	Broken Limb	No	179,448	0	0	0	0	179,448	0
500	CI	0		1	Broken Limb	No	8,259	0	0	0	0	0	1

Πίνακας: Μια αναφορά ποιότητας δεδομένων για τον ABT
ανίχνευσης απάτης σε claims ασφάλισης αυτοκινήτου

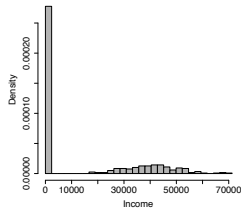
(α□) Συνεχή Χαρακτηριστικά (Features)

Feature	Count	% Miss.	Card.	Min	1 st Qrt.	Mean	Median	3 rd Qrt.	Max	Std. Dev.
Income	500	0.0	171	0.0	0.0	13,740.0	0.0	33,918.5	71,284.0	20,081.5
Num Claimants	500	0.0	4	1.0	1.0	1.9	2	3.0	4.0	1.0
Claim Amount	500	0.0	493	-99,999	3,322.3	16,373.2	5,663.0	12,245.5	270,200.0	29,426.3
Total Claimed	500	0.0	235	0.0	0.0	9,597.2	0.0	11,282.8	729,792.0	35,655.7
Num Claims	500	0.0	7	0.0	0.0	0.8	0.0	1.0	56.0	2.7
Num Soft Tissue	500	2.0	6	0.0	0.0	0.2	0.0	0.0	5.0	0.6
% Soft Tissue	500	0.0	9	0.0	0.0	0.2	0.0	0.0	2.0	0.4
Amount Received	500	0.0	329	0.0	0.0	13,051.9	3,253.5	8,191.8	295,303.0	30,547.2
Fraud Flag	500	0.0	2	0.0	0.0	0.3	0.0	1.0	1.0	0.5

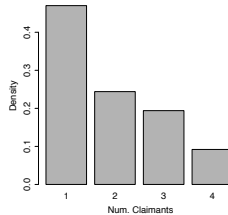
Πίνακας: Μια αναφορά ποιότητας δεδομένων για τον ABT
ανίχνευσης απάτης σε claims ασφάλισης αυτοκινήτου.

(α□) Κατηγορικά Χαρακτηριστικά (Features)

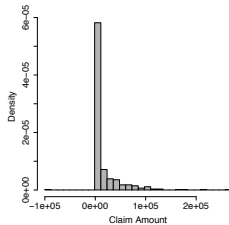
Feature	Count	% Miss.	Card.	Mode	Mode Freq.	Mode %	2 nd Mode	2 nd Mode Freq.	2 nd Mode %
Insurance Type	500	0.0	1	CI	500	1.0	--	--	--
Marital Status	500	61.2	4	Married	99	51.0	Single	48	24.7
Injury Type	500	0.0	4	Broken Limb	177	35.4	Soft Tissue	172	34.4
Hospital Stay	500	0.0	2	No	354	70.8	Yes	146	29.2



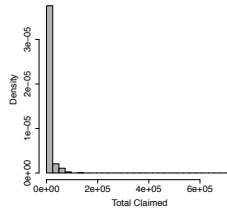
(α) Income



(β) Num Claimants

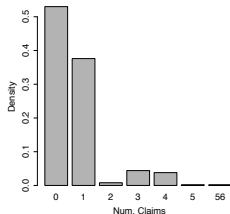


(γ) Claim Amount

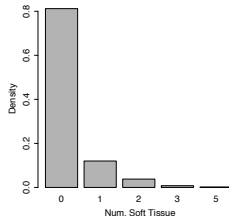


(δ) Total Claimed

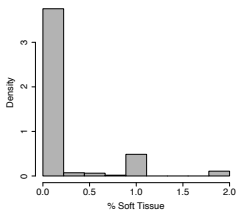
Σχήμα: Οπτικοποιήσεις των συνεχών και κατηγορικών features στον ABT ανίχνευσης απάτης σε claims ασφάλισης αυτοκινήτου.



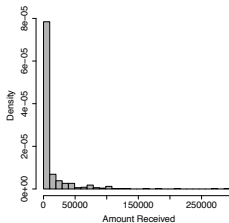
(α) Num Claims



(β) Num Soft Tissue

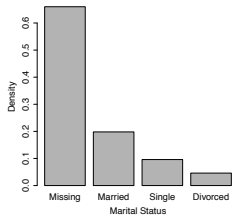


(γ) % Soft Tissue

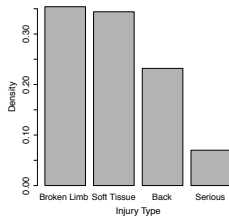


(δ) Amount Received

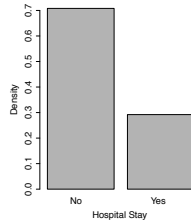
Σχήμα: Οπτικοποιήσεις των συνεχών και κατηγορικών features στον ABT ανίχνευσης απάτης σε claims ασφάλισης αυτοκινήτου του



(α) Marital Status

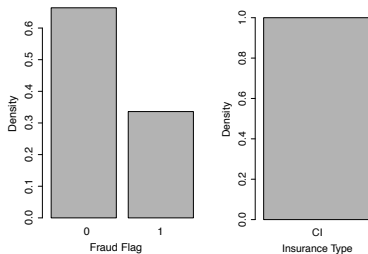


(β) Injury Type



(γ) Hospital Stay

Σχήμα: Οπτικοποιήσεις των συνεχών και κατηγορικών features στον ABT ανίχνευσης απάτης σε claims ασφάλισης αυτοκινήτου του Πίνακα 2 ^[7].



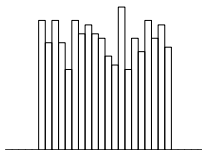
(α) Fraud Flag (β) Insurance Type

Σχήμα: Οπτικοποιήσεις των συνεχών και κατηγορικών features στον ABT ανίχνευσης απάτης σε claims ασφάλισης αυτοκινήτου.

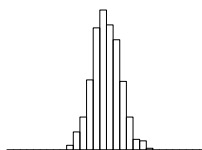
Γνωριμία με τα Δεδομένα

- Για τα κατηγορικά features, πρέπει να:
 - Εξετάζουμε το mode, το 2^{nd} mode, το ποσοστό του mode (mode %) και το ποσοστό του 2^{nd} mode (2^{nd} mode %), καθώς αυτά δείχνουν τα συχνότερα επίπεδα (levels) στα συγκεκριμένα features και αναδεικνύουν αν κάποιο level κυριαρχεί στο dataset.
- Για τα συνεχόμενα features πρέπει να:
 - Εξετάζουμε τον μέσο όρο (mean) και την τυπική απόκλιση (standard deviation) κάθε feature, ώστε να αποκτήσουμε αίσθηση της κεντρικής τάσης και της διασποράς των τιμών στο dataset για το συγκεκριμένο feature.
 - Εξετάζουμε τις ελάχιστες και μέγιστες τιμές (minimum και maximum) για να κατανοήσουμε το εύρος (range) τιμών που είναι δυνατό για κάθε feature.

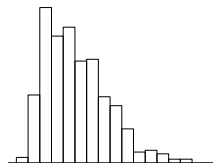
- Όταν δημιουργούμε ιστογράμματα (histograms) για features, υπάρχουν αρκετά συνηθισμένα και καλά κατανοητά σχήματα (shapes) που πρέπει να αναζητούμε.



(α) Ομοιόμορφη (Uniform)

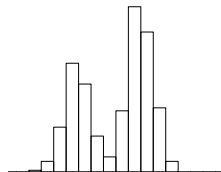
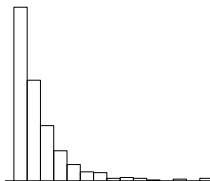
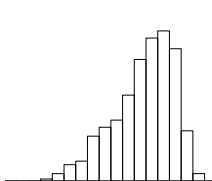


(β) Κανονική (Normal
Unimodal)



(γ) Μονοκορυφική (δεξιά ασύμ-
μετρη)

Σχήμα: Ιστογράμματα για διαφορετικά σύνολα δεδομένων, καθένα από τα οποία παρουσιάζει γνωστά και συνηθισμένα χαρακτηριστικά.

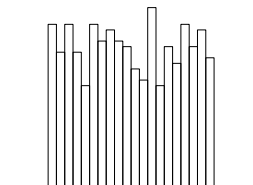


(α) Μονοκορυφική (αριστερά
ασύμμετρη)

(β) Εκθετική (Exponential)

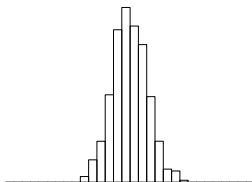
(γ) Πολυκορυφική (Multimodal)

Σχήμα: Ιστογράμματα για διαφορετικά σύνολα δεδομένων, καθένα από τα οποία παρουσιάζει γνωστά και συνηθισμένα χαρακτηριστικά.



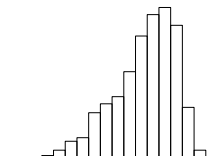
Ομοιόμορφη (Uniform)

- Μια ομοιόμορφη κατανομή (uniform distribution) δείχνει ότι ένα feature είναι εξίσου πιθανό να πάρει μια τιμή σε οποιοδήποτε από τα εύρη τιμών που εμφανίζονται.

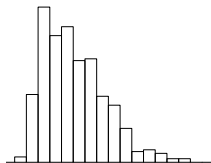


Κανονική (Normal / Unimodal)

- Features που ακολουθούν κανονική κατανομή (normal distribution) χαρακτηρίζονται από ισχυρή τάση προς μια κεντρική τιμή και συμμετρική διασπορά γύρω από αυτή.

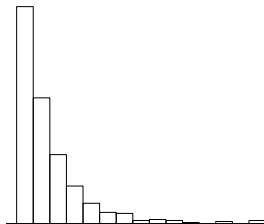


Μονοκορυφική (αριστερά ασύμμετρη)



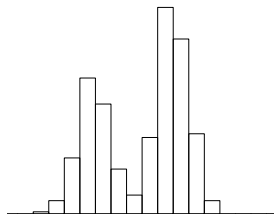
Μονοκορυφική (δεξιά ασύμμετρη)

- Η ασυμμετρία (skew) είναι απλώς μια τάση προς πολύ υψηλές (**right skew**) ή πολύ χαμηλές (**left skew**) τιμές.



Εκθετική (Exponential)

- Σε ένα feature που ακολουθεί **εκθετική κατανομή** (exponential distribution), η πιθανότητα εμφάνισης μικρού αριθμού χαμηλών τιμών είναι πολύ υψηλή, αλλά μειώνεται απότομα όσο οι τιμές αυξάνονται.



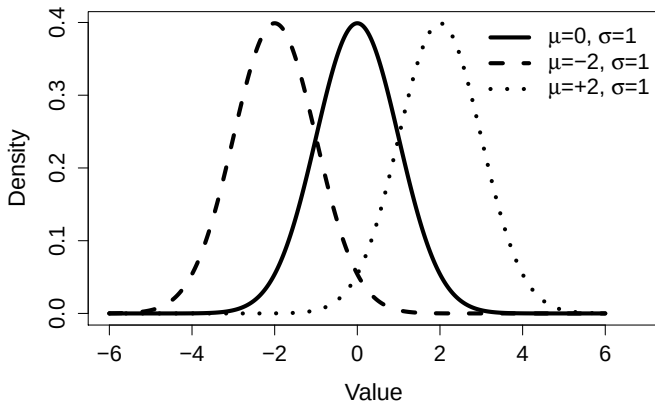
Πολυκορυφική (Multimodal)

- Ένα feature με **πολυκορυφική κατανομή** (multimodal distribution) έχει δύο ή περισσότερα πολύ συχνά εμφανιζόμενα εύρη τιμών, τα οποία είναι καθαρά διαχωρισμένα.

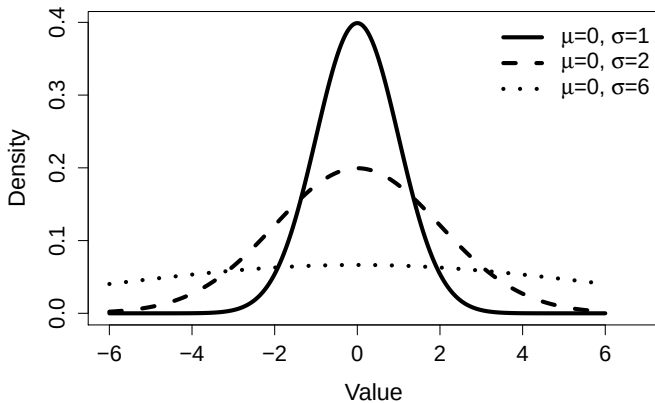
- Η συνάρτηση πυκνότητας πιθανότητας (probability density function) για την **κανονική** κατανομή (ή **Gaussian distribution**) είναι

$$N(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x - \mu)^2}{2\sigma^2}} \quad (1)$$

όπου x είναι οποιαδήποτε τιμή και οι μ και σ είναι παράμετροι που καθορίζουν το σχήμα της κατανομής: ο **μέσος όρος πληθυσμού** (population mean) και η **τυπική απόκλιση πληθυσμού** (population standard deviation).

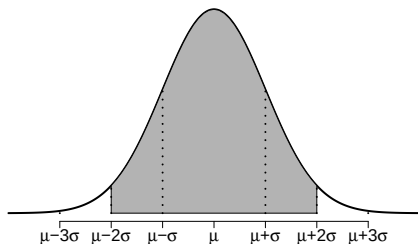


Σχήμα: Τρεις κανονικές κατανομές με διαφορετικούς μέσους όρους και ίδια τυπική απόκλιση.



Σχήμα: Τρεις κανονικές κατανομές με ίδιο μέσο όρο και διαφορετικές τυπικές αποκλίσεις.

- Ο κανόνας 68–95–99.7 είναι ένα χρήσιμο χαρακτηριστικό της κανονικής κατανομής.
- Ο κανόνας λέει ότι περίπου:
 - το 68% των παρατηρήσεων θα βρίσκεται εντός μίας σ από τη μ ,
 - το 95% των παρατηρήσεων θα βρίσκεται εντός δύο σ από τη μ ,
 - το 99.7% των παρατηρήσεων θα βρίσκεται εντός τριών σ από τη μ .



Σχήμα: Απεικόνιση του κανόνα ποσοστών 68–95–99.7 που η κανονική κατανομή ορίζει ως την αναμενόμενη κατανομή των παρατηρήσεων. Η γκρι περιοχή δείχνει το εύρος όπου αναμένεται να βρίσκεται το 95% των παρατηρήσεων.

Μελέτη Περίπτωσης: Απάτη σε Ασφάλιση Αυτοκινήτου

Εξετάστε την αναφορά ποιότητας δεδομένων για το σενάριο πρόβλεψης απάτης και σχολιάστε την κεντρική τάση και τη διασπορά (variation) για κάθε feature.

Εντοπισμός Ζητημάτων Ποιότητας Δεδομένων

- Ένα **ζήτημα ποιότητας δεδομένων** (data quality issue) ορίζεται χαλαρά ως οτιδήποτε *ασυνήθιστο* στα δεδομένα ενός ABT.
- Τα συνηθέστερα ζητήματα ποιότητας δεδομένων είναι:
 - **ελλείπουσες τιμές** (missing values),
 - **ανώμαλη πληθικότητα / καρδιναλιότητα** (irregular cardinality),
 - **ακραίες τιμές** (outliers).

- Τα ζητήματα ποιότητας δεδομένων που εντοπίζουμε από μια αναφορά ποιότητας δεδομένων είναι δύο τύπων:
 - Ζητήματα ποιότητας δεδομένων λόγω **μη έγκυρων δεδομένων** (invalid data).
 - Ζητήματα ποιότητας δεδομένων λόγω **έγκυρων δεδομένων** (valid data).

Πίνακας: Η δομή ενός πλάνου ποιότητας δεδομένων (data quality plan).

Feature	Ζήτημα Ποιότητας Δεδομένων	Πιθανές Στρατηγικές Αντιμετώπισης
-----	-----	-----
-----	-----	-----
-----	-----	-----
-----	-----	-----

Πίνακας: Το πλάνο ποιότητας δεδομένων για τον ABT πρόβλεψης απάτης σε ασφάλιση αυτοκινήτου.

Feature	Ζήτημα Ποιότητας Δεδομένων	Πιθανές Στρατηγικές Αντιμετω
Num Soft Tissue	Ελλείπουσες τιμές (2%)	
Claim Amount	Ακραίες τιμές (υψηλές)	
Amount Received	Ακραίες τιμές (υψηλές)	

Αντιμετώπιση Ζητημάτων Ποιότητας Δεδομένων

- Προσέγγιση 1: Απόρριψη (drop) οποιωνδήποτε features έχουν ελλείπουσες τιμές.
- Προσέγγιση 2: Εφαρμογή **complete case analysis**.
- Προσέγγιση 3: Κατασκευή ενός **missing indicator feature** από features με ελλείπουσες τιμές.

- Η **αποκατάσταση/συμπλήρωση** (imputation) αντικαθιστά τις ελλείπουσες τιμές ενός feature με μια εύλογη εκτιμώμενη τιμή, βασισμένη στις τιμές που υπάρχουν.
- Η συνηθέστερη προσέγγιση είναι να αντικαθιστούμε τις ελλείπουσες τιμές ενός feature με ένα μέτρο κεντρικής τάσης του feature.
- Θα ήμασταν διστακτικοί να χρησιμοποιήσουμε imputation σε features με ελλείψεις άνω του 30% των τιμών τους και θα συνιστούσαμε έντονα να αποφεύγεται σε features με ελλείψεις άνω του 50%.

- Ο ευκολότερος τρόπος αντιμετώπισης ακραίων τιμών (outliers) είναι να χρησιμοποιήσουμε έναν **clamp transformation** που «κόβει» όλες τις τιμές πάνω από ένα άνω όριο και κάτω από ένα κάτω όριο, θέτοντάς τις ίσες με αυτά τα όρια, αφαιρώντας έτσι τις προβληματικές ακραίες τιμές.

$$a_i = \begin{cases} lower & \text{if } a_i < lower \\ upper & \text{if } a_i > upper \\ a_i & \text{otherwise} \end{cases} \quad (2)$$

όπου a_i είναι μια συγκεκριμένη τιμή του feature a , και τα $lower$ και $upper$ είναι τα κάτω και άνω όρια αντίστοιχα.

Μελέτη Περίπτωσης: Απάτη σε Ασφάλιση Αυτοκινήτου

Ποιες στρατηγικές αντιμετώπισης θα προτείνετε για τα ζητήματα ποιότητας δεδομένων που εντοπίστηκαν στον ABT απάτης σε ασφάλιση αυτοκινήτου;

Μελέτη Περίπτωσης: Απάτη σε Ασφάλιση Αυτοκινήτου

Πίνακας: Το πλάνο ποιότητας δεδομένων για τον ABT πρόβλεψης απάτης σε ασφάλιση αυτοκινήτου.

Feature	Ζήτημα Ποιότητας Δεδομένων	Πιθανή Αντιμετώπιση
Num Soft Tissue	Ελλείπουσες τιμές (2%)	Imputation (median: 0.0)
Claim Amount	Ακραίες τιμές (υψηλές)	Clamp transformation (manual: 0, 80 000)
Amount Received	Ακραίες τιμές (υψηλές)	Clamp transformation (manual: 0, 80 000)

Σύνοψη

- Τα βασικά αποτελέσματα (outcomes) της διαδικασίας **εξερεύνησης δεδομένων** (data exploration) είναι ότι ο/η αναλυτής/αναλύτρια πρέπει:
 - 1 Να έχει εξοικειωθεί με τα features του ABT, ειδικά με την κεντρική τους τάση, τη διασπορά και τις **κατανομές** τους.
 - 2 Να έχει εντοπίσει τυχόν **ζητήματα ποιότητας δεδομένων** στον ABT, ιδίως **ελλείπουσες τιμές, ανώμαλη καρδιναλιότητα** και **ακραίες τιμές**.
 - 3 Να έχει διορθώσει τα ζητήματα ποιότητας δεδομένων που οφείλονται σε **μη έγκυρα δεδομένα**.
 - 4 Να έχει καταγράψει τα ζητήματα ποιότητας δεδομένων που οφείλονται σε **έγκυρα δεδομένα** σε ένα **data quality plan**, μαζί με πιθανές στρατηγικές αντιμετώπισης.
 - 5 Να είναι βέβαιος/η ότι υπάρχει αρκετή ποσότητα δεδομένων καλής ποιότητας ώστε να συνεχιστεί το έργο.

Εξερεύνηση Δεδομένων Μέρος 2

Fundamentals of Machine Learning for Predictive Data Analytics

© John D. Kelleher and Brian Mac Namee and Aoife D'Arcy

Αθανάσιος Σάκκας, ΟΠΑ

1 Προχωρημένη Εξερεύνηση Δεδομένων

- Οπτικοποίηση Σχέσεων Μεταξύ Χαρακτηριστικών
- Μέτρηση Συνδιακύμανσης & Συσχέτισης

2 Προετοιμασία Δεδομένων

- Normalization
- Binning
- Sampling

3 Σύνοψη

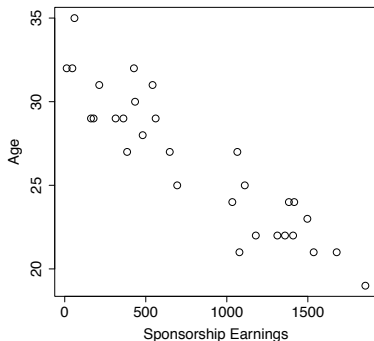
Προχωρημένη Εξερεύνηση Δεδομένων

ID	Position	Height	Weight	Career Stage	Age	Sponsorship Earnings	Shoe Sponsor
1	forward	192	218	veteran	29	561	yes
2	center	218	251	mid-career	35	60	no
3	forward	197	221	rookie	22	1,312	no
4	forward	192	219	rookie	22	1,359	no
5	forward	198	223	veteran	29	362	yes
6	guard	166	188	rookie	21	1,536	yes
7	forward	195	221	veteran	25	694	no
8	guard	182	199	rookie	21	1,678	yes
9	guard	189	199	mid-career	27	385	yes
10	forward	205	232	rookie	24	1,416	no
11	center	206	246	mid-career	29	314	no
12	guard	185	207	rookie	23	1,497	yes
13	guard	172	183	rookie	24	1,383	yes
14	guard	169	183	rookie	24	1,034	yes
15	guard	185	197	mid-career	29	178	yes
16	forward	215	232	mid-career	30	434	no
17	guard	158	184	veteran	29	162	yes
18	guard	190	207	mid-career	27	648	yes
19	center	195	235	mid-career	28	481	no
20	guard	192	200	mid-career	32	427	yes
21	forward	202	220	mid-career	31	542	no
22	forward	184	213	mid-career	32	12	no
23	forward	190	215	rookie	22	1,179	no
24	guard	178	193	rookie	21	1,078	no
25	guard	185	200	mid-career	31	213	yes
26	forward	191	218	rookie	19	1,855	no
27	center	196	235	veteran	32	47	no
28	forward	198	221	rookie	22	1,409	no
29	center	207	247	veteran	27	1,065	no
30	center	201	244	mid-career	25	1,111	yes

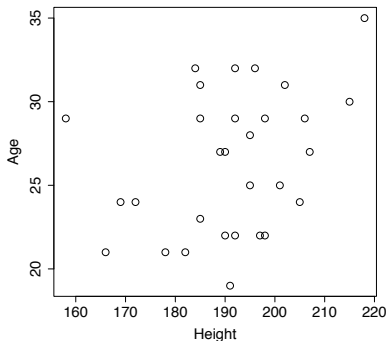
- Ένα **διάγραμμα διασποράς** (scatter plot) βασίζεται σε δύο άξονες: ο οριζόντιος άξονας αναπαριστά ένα χαρακτηριστικό και ο κάθετος άξονας ένα δεύτερο.
- Κάθε παρατήρηση (instance) στο σύνολο δεδομένων αναπαρίσταται ως ένα σημείο στο διάγραμμα, το οποίο καθορίζεται από τις τιμές της παρατήρησης για τα δύο χαρακτηριστικά.

Σχήμα: Παράδειγμα διαγράμματος διασποράς που δείχνει τη σχέση μεταξύ των χαρακτηριστικών Height και Weight από το dataset της επαγγελματικής ομάδας μπάσκετ στον Πίνακα 4 ^[4].

Οπτικοποίηση Σχέσεων Μεταξύ Χαρακτηριστικών



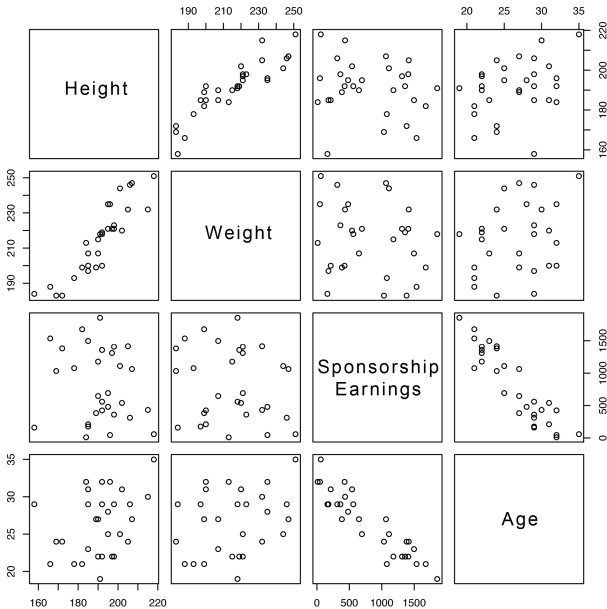
(α □)



(β □)

Σχήμα: Παραδείγματα διαγραμμάτων διασποράς που δείχνουν (a) την ισχυρή αρνητική συνδιακύμανση μεταξύ των Sponsorship Earnings και Age και (b) τα Height και Age από το dataset στον Πίνακα 4 ^[4].

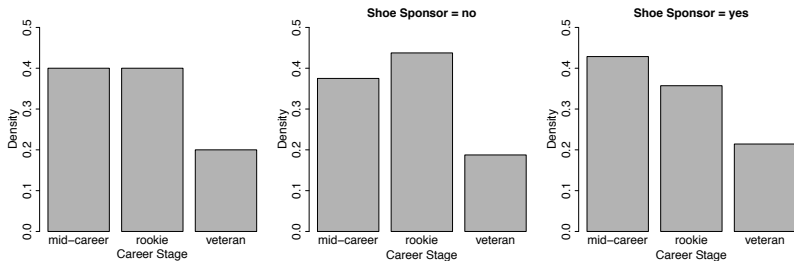
- Ένας **πίνακας διασποράς** (scatter plot matrix) ή **SPLOM** εμφανίζει διαγράμματα διασποράς για μια ολόκληρη συλλογή χαρακτηριστικών, οργανωμένα σε μορφή πίνακα.
- Είναι χρήσιμος για την εξερεύνηση σχέσεων μεταξύ ομάδων χαρακτηριστικών — π.χ. όλων των συνεχών χαρακτηριστικών ενός ABT.



Σχήμα: Πίνακας διασποράς (SPLOM) που παρουσιάζει διαγράμματα διασποράς των συνεχών χαρακτηριστικών.

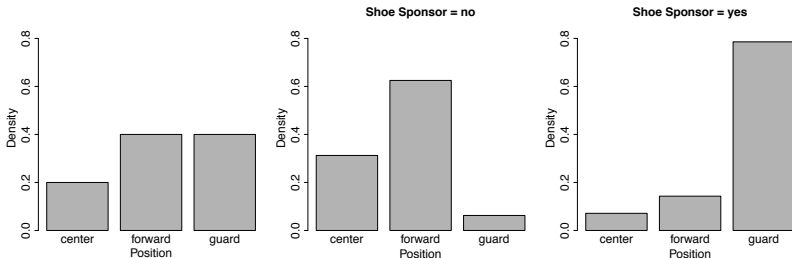
- Ο απλούστερος τρόπος για να οπτικοποιήσουμε τη σχέση μεταξύ δύο κατηγορικών μεταβλητών είναι να χρησιμοποιήσουμε μια συλλογή από ραβδογράμματα **small multiples**.

Οπτικοποίηση Σχέσεων Μεταξύ Χαρακτηριστικών



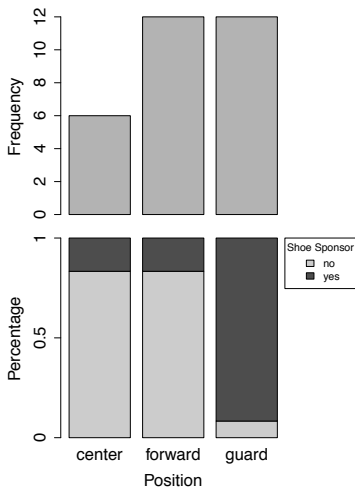
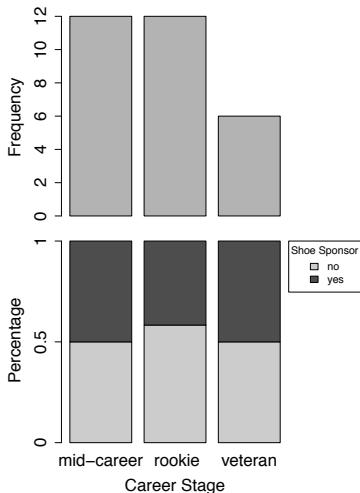
Σχήμα: Χρήση ραβδογραμμάτων small multiples για την απεικόνιση της σχέσης μεταξύ των Career Stage και Shoe Sponsor.

Οπτικοποίηση Σχέσεων Μεταξύ Χαρακτηριστικών



Σχήμα: Χρήση ραβδογραμμάτων small multiples για την απεικόνιση της σχέσης μεταξύ των Position και Shoe Sponsor.

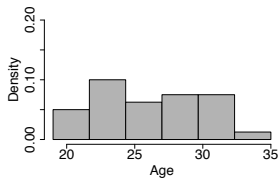
- Αν ο αριθμός των επιπέδων (levels) ενός από τα χαρακτηριστικά που συγκρίνονται δεν είναι μεγαλύτερος από τρία, μπορούμε να χρησιμοποιήσουμε **stacked bar plots** ως εναλλακτική των small multiples.



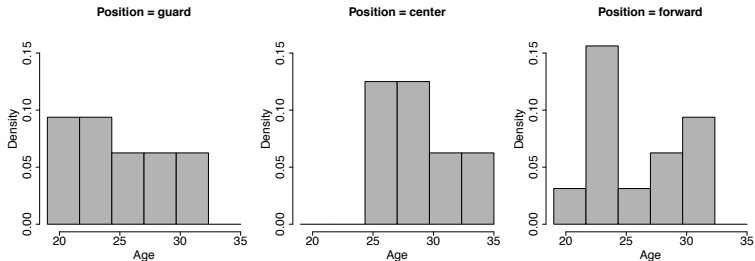
(α) Career Stage and Shoe (β) Position and Shoe Sponsor Sponsor

Σχήμα: Οπτικοποιήσεις με stacked bar plots.

- Για να οπτικοποιήσουμε τη σχέση μεταξύ ενός συνεχούς χαρακτηριστικού και ενός κατηγορικού χαρακτηριστικού, είναι χρήσιμη μια προσέγγιση **small multiples** όπου σχεδιάζουμε ένα ιστόγραμμα των τιμών του συνεχούς χαρακτηριστικού για κάθε επίπεδο του κατηγορικού χαρακτηριστικού.

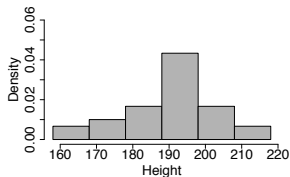


(α) Age

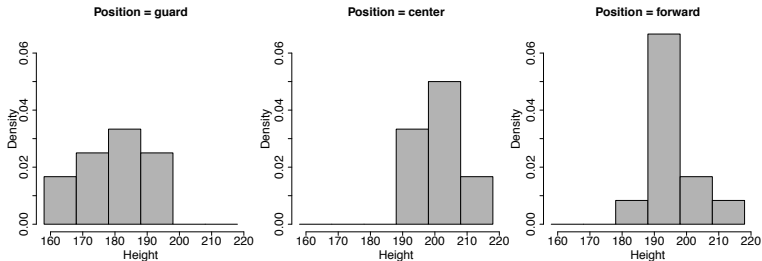


(β) Age and Position

Σχήμα: Χρήση small multiple ιστογραμμάτων για την απεικόνιση της σχέσης μεταξύ του Age και του Position.



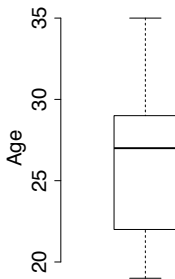
(α) Height



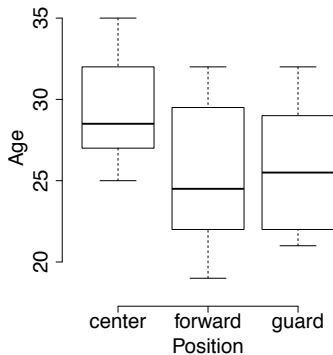
(β) Height and Position

Σχήμα: Χρήση small multiple ιστογραμμάτων για την απεικόνιση της σχέσης μεταξύ του Height και του Position.

- Μια δεύτερη προσέγγιση για να οπτικοποιήσουμε τη σχέση μεταξύ ενός κατηγορικού και ενός συνεχούς χαρακτηριστικού είναι να χρησιμοποιήσουμε μια συλλογή από box plots.
- Για κάθε επίπεδο του κατηγορικού χαρακτηριστικού σχεδιάζουμε ένα box plot των αντίστοιχων τιμών του συνεχούς χαρακτηριστικού.

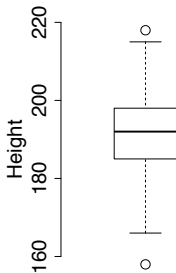


(α) Age

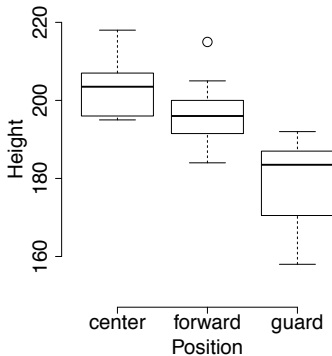


(β) Age and Position

Σχήμα: Χρήση box plots για την απεικόνιση της σχέσης μεταξύ του Age και του Position.



(α) Height



(β) Height and Position

Σχήμα: Χρήση box plots για την απεικόνιση της σχέσης μεταξύ του Height και του Position.

- Εκτός από την οπτική επιθεώρηση των scatter plots, μπορούμε να υπολογίσουμε τυπικά (formal) μέτρα της σχέσης μεταξύ δύο συνεχών χαρακτηριστικών χρησιμοποιώντας **συνδιακύμανση** (covariance) και **συσχέτιση** (correlation).
- Για δύο χαρακτηριστικά, a και b , σε dataset με n παρατηρήσεις, η **δειγματική συνδιακύμανση** (sample covariance) μεταξύ a και b είναι:

$$\text{cov}(a, b) = \frac{1}{n-1} \sum_{i=1}^n ((a_i - \bar{a}) \times (b_i - \bar{b})) \quad (1)$$

όπου a_i και b_i είναι οι τιμές των χαρακτηριστικών a και b για την i -οστή παρατήρηση και $\bar{a}$ και $\bar{b}$ είναι οι δειγματικοί μέσοι όροι των a και b .

- Οι τιμές της συνδιακύμανσης ανήκουν στο διάστημα $[-\infty, \infty]$:
 - αρνητικές τιμές υποδηλώνουν αρνητική σχέση,
 - θετικές τιμές υποδηλώνουν θετική σχέση,
 - τιμές κοντά στο μηδέν υποδηλώνουν μικρή ή μηδενική σχέση.

Υπολογισμός συνδιακύμανσης μεταξύ του Height και των Weight, Age από το dataset των παικτών μπάσκετ.

	Height		Weight		$(h - \bar{h}) \times$	Age		$(h - \bar{h}) \times$
ID	(h)	$h - \bar{h}$	(w)	$w - \bar{w}$	$(w - \bar{w})$	(a)	$a - \bar{a}$	$(a - \bar{a})$
1	192	0.9	218	3.0	2.7	29	2.6	2.3
2	218	26.9	251	36.0	967.5	35	8.6	231.3
3	197	5.9	221	6.0	35.2	22	-4.4	-26.0
4	192	0.9	219	4.0	3.6	22	-4.4	-4.0
5	198	6.9	223	8.0	55.0	29	2.6	17.9
				...				
26	191	-0.1	218	3.0	-0.3	19	-7.4	0.7
27	196	4.9	235	20.0	97.8	32	5.6	27.4
28	198	6.9	221	6.0	41.2	22	-4.4	-30.4
29	207	15.9	247	32.0	508.3	27	0.6	9.5
30	201	9.9	244	29.0	286.8	25	-1.4	-13.9
Mean	191.1		215.0			26.4		
Std Dev	13.6		19.8			4.2		
Sum					7,009.9			570.8

Υπολογισμός συνδιακύμανσης μεταξύ του Height και των Weight, Age από το dataset των παικτών μπάσκετ.

$$\begin{aligned}\text{cov}(\text{Height}, \text{Weight}) &= \frac{7,009.9}{29} = 241.72 \\ \text{cov}(\text{Height}, \text{Age}) &= \frac{570.8}{29} = 19.7\end{aligned}$$

- Η **συσχέτιση** (correlation) είναι μια κανονικοποιημένη μορφή της συνδιακύμανσης και παίρνει τιμές από -1 έως $+1$.
- Η συσχέτιση μεταξύ δύο χαρακτηριστικών a και b δίνεται από:

$$corr(a, b) = \frac{cov(a, b)}{sd(a) \times sd(b)} \quad (2)$$

όπου $cov(a, b)$ είναι η συνδιακύμανση και $sd(a)$, $sd(b)$ οι τυπικές αποκλίσεις των a και b αντίστοιχα.

- Οι τιμές της συσχέτισης ανήκουν στο διάστημα $[-1, 1]$:
 - τιμές κοντά στο -1 δείχνουν πολύ ισχυρή αρνητική συσχέτιση,
 - τιμές κοντά στο $+1$ δείχνουν πολύ ισχυρή θετική συσχέτιση,
 - τιμές κοντά στο 0 δείχνουν ανυπαρξία συσχέτισης.
- Χαρακτηριστικά χωρίς συσχέτιση λέγονται **ανεξάρτητα** (independent).

Υπολογισμός συσχέτισης μεταξύ του Height και των Weight, Age από το dataset των παικτών μπάσκετ.

$$\text{corr}(\text{Height}, \text{Weight}) = \frac{241.72}{13.6 \times 19.8} = 0.898$$

$$\text{corr}(\text{Height}, \text{Age}) = \frac{19.7}{13.6 \times 4.2} = 0.345$$

- Στα περισσότερα ABTs υπάρχουν πολλά συνεχή χαρακτηριστικά, μεταξύ των οποίων θέλουμε να εξερευνήσουμε σχέσεις.
- Δύο χρήσιμα εργαλεία είναι ο πίνακας συνδιακύμανσης (covariance matrix) και ο πίνακας συσχέτισης (correlation matrix).

- Ο πίνακας συνδιακύμανσης (συνήθως συμβολίζεται ως Σ) για ένα σύνολο συνεχών χαρακτηριστικών $\{a, b, \dots, z\}$ δίνεται από:

$$\sum_{\{a,b,\dots,z\}} = \begin{bmatrix} \text{var}(a) & \text{cov}(a, b) & \dots & \text{cov}(a, z) \\ \text{cov}(b, a) & \text{var}(b) & \dots & \text{cov}(b, z) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(z, a) & \text{cov}(z, b) & \dots & \text{var}(z) \end{bmatrix} \quad (3)$$

- Αντίστοιχα, ο **πίνακας συσχέτισης** (correlation matrix) είναι μια κανονικοποιημένη εκδοχή του πίνακα συνδιακύμανσης και δείχνει τη συσχέτιση για κάθε ζεύγος χαρακτηριστικών:

$$correlation\ matrix_{\{a,b,...,z\}} = \begin{bmatrix} corr(a, a) & corr(a, b) & \cdots & corr(a, z) \\ corr(b, a) & corr(b, b) & \cdots & corr(b, z) \\ \vdots & \vdots & \ddots & \vdots \\ corr(z, a) & corr(z, b) & \cdots & corr(z, z) \end{bmatrix} \quad (4)$$

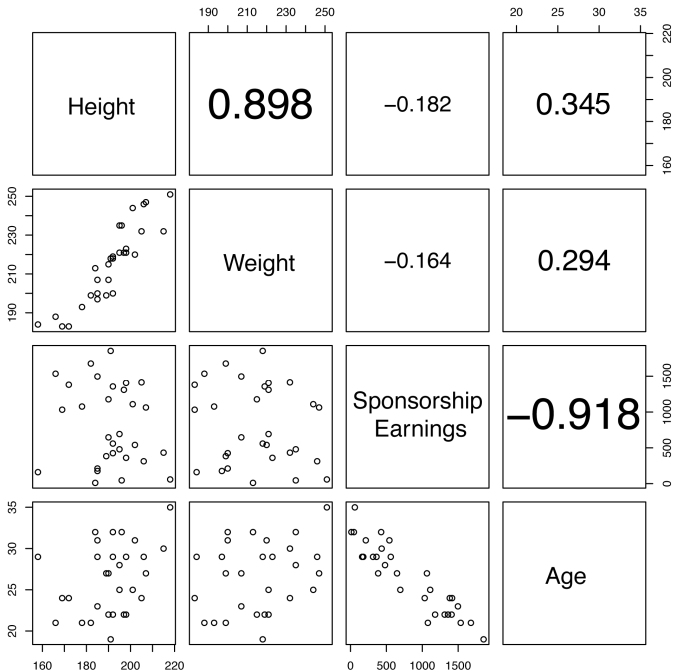
- Υπολογισμός πίνακα συνδιακύμανσης για τα Height, Weight, Age από το dataset των παικτών μπάσκετ.

$$\sum_{\langle \text{Height}, \text{Weight}, \text{Age} \rangle} = \begin{bmatrix} 185.128 & 241.72 & 19.7 \\ 241.72 & 392.102 & 24.469 \\ 19.7 & 24.469 & 17.697 \end{bmatrix}$$

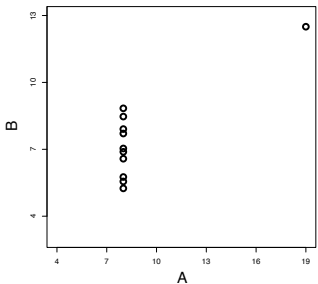
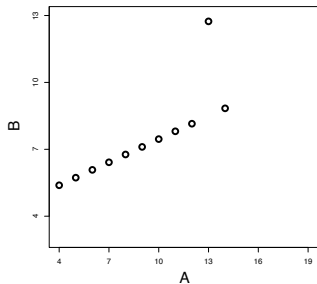
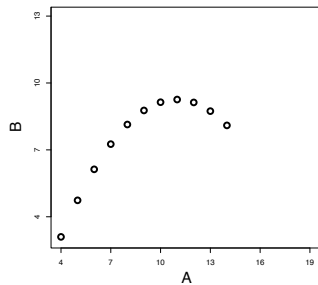
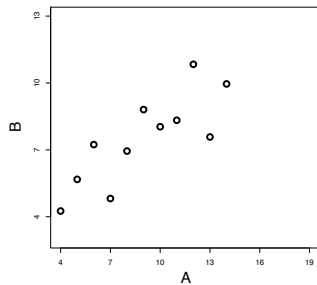
- Υπολογισμός πίνακα συσχέτισης για τα Height, Weight, Age από το dataset των παικτών μπάσκετ.

$$\text{correlation matrix}_{<Height,Weight,Age>} = \begin{bmatrix} 1.0 & 0.898 & 0.345 \\ 0.898 & 1.0 & 0.294 \\ 0.345 & 0.294 & 1.0 \end{bmatrix}$$

- Ο πίνακας διασποράς (SPLOM) είναι ουσιαστικά μια οπτικοποίηση του πίνακα συσχέτισης.
- Αυτό γίνεται πιο εμφανές αν συμπεριλάβουμε τους συντελεστές συσχέτισης στα κελιά πάνω από τη διαγώνιο.



- Η συσχέτιση είναι ένα καλό μέτρο σχέσης μεταξύ δύο συνεχών χαρακτηριστικών, αλλά δεν είναι σε καμία περίπτωση τέλειο.
- Ορισμένοι περιορισμοί της συσχέτισης φαίνονται καθαρά στο κλασικό παράδειγμα **Anscombe's quartet** του **Francis Anscombe**.



- Ίσως το πιο σημαντικό που πρέπει να θυμόμαστε σχετικά με τη συσχέτιση είναι ότι **η συσχέτιση δεν συνεπάγεται απαραίτητα αιτιότητα.**

Προετοιμασία Δεδομένων

- Ορισμένες τεχνικές προετοιμασίας δεδομένων αλλάζουν τον τρόπο αναπαράστασης των δεδομένων, απλώς για να τα κάνουν πιο συμβατά με συγκεκριμένους αλγορίθμους μηχανικής μάθησης.
 - Normalization
 - Binning
 - Sampling

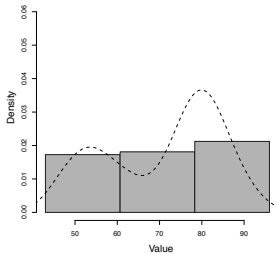
- Οι τεχνικές **Normalization** μπορούν να χρησιμοποιηθούν ώστε ένα συνεχές χαρακτηριστικό να «μετασχηματιστεί» και να πέσει μέσα σε ένα προκαθορισμένο εύρος τιμών, διατηρώντας τις σχετικές διαφορές μεταξύ των τιμών του χαρακτηριστικού.

Το αποτέλεσμα της κανονικοποίησης (normalising) ενός μικρού δείγματος των Height και Sponsorship Earnings από το dataset της επαγγελματικής ομάδας μπάσκετ.

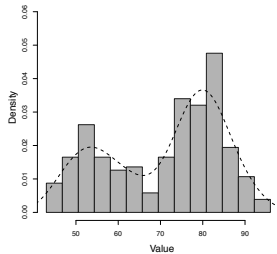
	Height			Sponsorship Earnings		
	Values	Range	Standard	Values	Range	Standard
	192	0.500	-0.073	561	0.315	-0.649
	197	0.679	0.533	1,312	0.776	0.762
	192	0.500	-0.073	1,359	0.804	0.850
	182	0.143	-1.283	1,678	1.000	1.449
	206	1.000	1.622	314	0.164	-1.114
	192	0.500	-0.073	427	0.233	-0.901
	190	0.429	-0.315	1,179	0.694	0.512
	178	0.000	-1.767	1,078	0.632	0.322
	196	0.643	0.412	47	0.000	-1.615
	201	0.821	1.017	1111	0.652	0.384
Max	206			1,678		
Min	178			47		
Mean	193			907		
Std Dev	8.26			532.18		

- Το **Binning** μετατρέπει ένα συνεχές χαρακτηριστικό σε κατηγορικό.
- Για να κάνουμε binning, ορίζουμε μια σειρά από διαστήματα τιμών (bins) για το συνεχές χαρακτηριστικό, τα οποία αντιστοιχούν στα επίπεδα του νέου κατηγορικού χαρακτηριστικού.
- Θα παρουσιάσουμε δύο από τις πιο δημοφιλείς μεθόδους ορισμού bins:
 - **equal-width binning**
 - **equal-frequency binning**

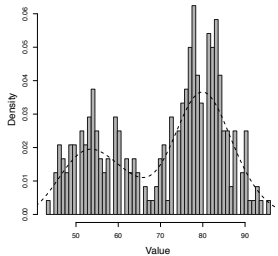
- Η επιλογή του αριθμού bins μπορεί να είναι δύσκολη. Το βασικό trade-off είναι:
 - Αν ορίσουμε πολύ λίγα bins, μπορεί να χάσουμε σημαντική πληροφορία.
 - Αν ορίσουμε πάρα πολλά bins, μπορεί να έχουμε πολύ λίγες παρατηρήσεις ανά bin ή ακόμη και κενά bins.



(ϵ) 3 bins

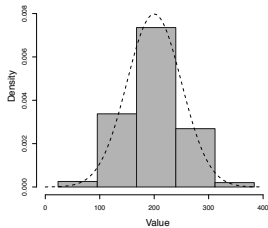


(ϵ) 14 bins

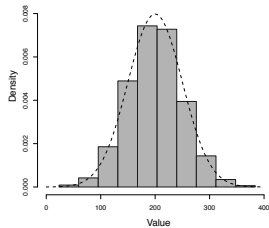


(ζ) 60 bins

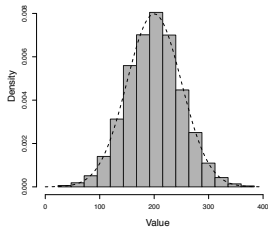
- Ο αλγόριθμος equal-width binning χωρίζει το εύρος τιμών (range) του χαρακτηριστικού σε b bins, το καθένα με πλάτος $\frac{range}{b}$.



(η) 5 Equal-width bins

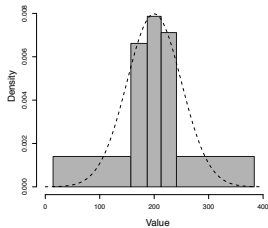


(θ) 10 Equal-width bins

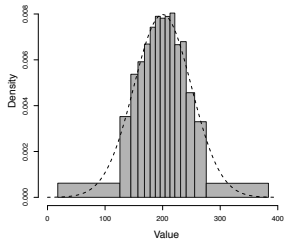
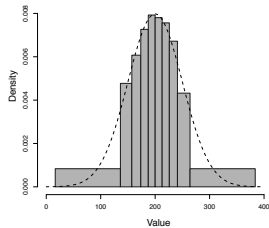


(ι) 15 Equal-width bins

- Το **equal-frequency binning** πρώτα ταξινομεί τις τιμές του συνεχούς χαρακτηριστικού σε αύξουσα σειρά και στη συνέχεια τοποθετεί ίσο αριθμό παρατηρήσεων σε κάθε bin, ξεκινώντας από το bin 1.
- Ο αριθμός παρατηρήσεων ανά bin είναι το πλήθος των παρατηρήσεων δια του αριθμού των bins, b .



(1α□) 5 Equal-frequency (1β□) 10 Equal-frequency
bins bins



(1γ□) 15 Equal-frequency
bins

- Μερικές φορές το dataset είναι τόσο μεγάλο ώστε δεν χρησιμοποιούμε όλα τα διαθέσιμα δεδομένα στο ABT και αντ' αυτού **δειγματοληπτούμε** (sample) ένα μικρότερο ποσοστό.
- Πρέπει να είμαστε προσεκτικοί ώστε τα δείγματα να παραμένουν αντιπροσωπευτικά του αρχικού dataset και να μη εισαχθεί ακούσια **μεροληψία** (bias).
- Συνηθισμένες μορφές δειγματοληψίας είναι:
 - **top sampling**
 - **random sampling**
 - **stratified sampling**
 - **under-sampling**
 - **over-sampling**

- Το **top sampling** επιλέγει απλώς το «κορυφαίο» $s\%$ των παρατηρήσεων ενός dataset για να δημιουργήσει δείγμα.
- Ενέχει σοβαρό κίνδυνο εισαγωγής bias, επειδή το δείγμα επηρεάζεται από οποιαδήποτε ταξινόμηση/σειρά στο αρχικό dataset.
- Προτείνουμε να αποφεύγεται το top sampling.

- Η προτεινόμενη προεπιλογή μας, το **random sampling**, επιλέγει τυχαία ένα ποσοστό $s\%$ των παρατηρήσεων από ένα μεγάλο dataset ώστε να δημιουργήσει ένα μικρότερο σύνολο.
- Είναι συνήθως καλή επιλογή, γιατί η τυχαιότητα βοηθά στο να αποφευχθεί η εισαγωγή bias.

- Το **stratified sampling** διασφαλίζει ότι οι σχετικές συχνότητες των επιπέδων ενός συγκεκριμένου **stratification feature** διατηρούνται στο δείγμα.
- Για να εκτελέσουμε stratified sampling:
 - χωρίζουμε τις παρατηρήσεις σε ομάδες (strata), όπου κάθε ομάδα περιέχει μόνο παρατηρήσεις με ένα συγκεκριμένο επίπεδο του stratification feature,
 - επιλέγουμε τυχαία $s\%$ των παρατηρήσεων από κάθε stratum,
 - συνδυάζουμε τις επιλογές για να πάρουμε συνολικό δείγμα $s\%$ του αρχικού dataset.

- Το **under-sampling** ξεκινά χωρίζοντας το dataset σε ομάδες, όπου κάθε ομάδα περιέχει μόνο παρατηρήσεις με ένα συγκεκριμένο επίπεδο του χαρακτηριστικού που θέλουμε να under-sample.
- Το πλήθος παρατηρήσεων στη *μικρότερη* ομάδα γίνεται ο στόχος (under-sampling target size).
- Κάθε ομάδα με περισσότερες παρατηρήσεις δειγματοληπτείται τυχαία κατάλληλα, ώστε να φτάσει το target size.
- Οι ομάδες συνδυάζονται για να σχηματίσουν το τελικό under-sampled dataset.

- Το **over-sampling** αντιμετωπίζει το ίδιο ζήτημα, αλλά αντίστροφα.
- Μετά τον χωρισμό σε ομάδες, το πλήθος παρατηρήσεων στη *μεγαλύτερη* ομάδα γίνεται ο στόχος (over-sampling target size).
- Για κάθε μικρότερη ομάδα δημιουργούμε δείγμα αυτού του μεγέθους χρησιμοποιώντας **random sampling with replacement**.
- Τα δείγματα συνδυάζονται για να σχηματίσουν το τελικό over-sampled dataset.

Σύνοψη

- Τα βασικά αποτελέσματα της διαδικασίας **εξερεύνησης δεδομένων** (data exploration) είναι ότι ο/η αναλυτής/ρια θα πρέπει:
 - 1 Να έχει εξοικειωθεί με τα χαρακτηριστικά του ABT, ιδίως με τις κεντρικές τάσεις, τις διακυμάνσεις και τις **κατανομές** (distributions).
 - 2 Να έχει εντοπίσει **ζητήματα ποιότητας δεδομένων**, ειδικά **missing values**, **irregular cardinality** και **outliers**.
 - 3 Να έχει διορθώσει ζητήματα ποιότητας που οφείλονται σε **invalid data**.
 - 4 Να έχει καταγράψει ζητήματα ποιότητας που οφείλονται σε **valid data** σε ένα **data quality plan** μαζί με πιθανές στρατηγικές αντιμετώπισης.
 - 5 Να είναι βέβαιος/η ότι υπάρχει αρκετό καλής ποιότητας data ώστε να συνεχιστεί το έργο.