



Διαχείριση Νομικών βάσεων δεδομένων με τη χρήση της R

Στεφανία Συρσίρη

ΠΕΡΙΕΧΟΜΕΝΑ

ΣΥΝΤΟΜΟΓΡΑΦΙΕΣ.....	2
ΚΕΦΑΛΑΙΟ ΠΡΩΤΟ	3
Νομικές βάσεις δεδομένων	3
Διευθετήσεις απαιτήσεων κατά της πόλης του Όστιν	3
Τα δεδομένα της βάσης	3
Το στοιχείο του χρόνου.....	4
Περαιτέρω διερεύνηση και εμφύτευση δεδομένων.....	8
Ανάλυση κειμένου για εξόρυξη δεδομένων	12

ΣΥΝΤΟΜΟΓΡΑΦΙΕΣ

ΕΕ : Ευρωπαϊκή Ένωση

ΗΠΑ : Ηνωμένες Πολιτείες της Αμερικής

λ.χ. : λόγου χάρη

μ.μ. : μετά μεσημβρίας

ΝΠΙΔ : Νομικό Πρόσωπο Ιδιωτικού Δικαίου

π.μ. : προ μεσημβρίας

ΚΕΦΑΛΑΙΟ ΠΡΩΤΟ

Νομικές βάσεις δεδομένων

Διευθετήσεις απαιτήσεων κατά της πόλης του Όστιν

Η πόλη του Όστιν (Austin) στο Τέξας της Αμερικής παρέχει ανοικτού τύπου δεδομένα για μελέτη προς το κοινό για διάφορες υπηρεσίες και λειτουργίες της. Μεταξύ αυτών είναι και δεδομένα από το Νομικό Τμήμα της τα οποία αφορούν σε διευθετήσεις απαιτήσεων ζημιωθέντων πολιτών και μη κατά της πόλης και τα οποία θα αναλύσουμε στην παρούσα έκθεση. Συγκεκριμένα πρόκειται για περιστατικά τα οποία έχουν ήδη εξεταστεί από το Νομικό τμήμα και έχουν καταλήξει σε κάποιον διακανονισμό.

Τα δεδομένα της βάσης

Αρχικά, γίνεται ανάκτηση των δεδομένων του Νομικού Τμήματος, και έπειτα πραγματοποιείται έλεγχος στη δομή και το περιεχόμενο του πλαισίου στο οποίο βρίσκονται, για να αποκτήσουμε μια πρώτη εικόνα σχετικά με τη διερεύνηση που θα μπορούσαμε να κάνουμε με ό,τι είναι διαθέσιμο, καθώς και με τις διορθώσεις που ενδεχομένως χρειάζονται, ώστε να καταστούν δυνατοί οι υπολογισμοί και τα γραφήματα που μας ενδιαφέρουν.

```
legal <- read.csv("https://query.data.world/s/5tpkrphqrkh5chf6okk4d2zi2tuqyo", header
= TRUE, stringsAsFactors = FALSE)
str(legal)      # structure of data frame

## 'data.frame':    159 obs. of  7 variables:
## $ Claim.Name      : chr  "Valdez, Michael" "Caballero, Rosita
" "Castillo, Angela" "Cutean, Nicki" ...
## $ Incident.Date   : chr  "02/28/2015 12:00:00 AM" "06/20/2014
12:00:00 AM" "04/02/2015 12:00:00 AM" "04/15/2015 12:00:00 AM" ...
## $ Department      : chr  "Austin Fire Dept" "Austin Police De
partment" "Austin Police Department" "Austin Police Department" ...
## $ Location.of.Incident..if.applicable.: chr  "Bulebell and Marigold" "Webberville
" "8th Street" "2nd Street & Guadalupe" ...
## $ Amount          : chr  "$700.00" "$7000.00" "$947.17" "$113
3.42" ...
## $ Category        : chr  "00 Auto" "00 Auto" "00 Auto" "00 Au
to" ...
## $ Disposition.Type : chr  "Claim Paid" "Claim Paid" "Claim Pai
d" "Claim Paid" ...

attach(legal)  # make the data easier to handle
```

Όπως φαίνεται, πρόκειται για ένα πλαίσιο δεδομένων με 159 εγγραφές (τα καταγεγραμμένα περιστατικά αξιώσεων), και με 7 μεταβλητές, οι οποίες είναι οι εξής:

1. ονοματεπώνυμο ζημιωθέντος,
2. ημερομηνία συμβάντος,
3. τμήμα διευθύνσεως της πόλης,
4. τοποθεσία (στην οποία έλαβε χώρα το συμβάν),
5. ποσό αποζημίωσης (σε δολάρια),
6. (ασφαλιστική) κατηγορία,
7. είδος διευθέτησης.

Η δομή των δεδομένων φαίνεται πως είναι ενιαία για όλες τις μεταβλητές και συγκεκριμένα σε μορφή κειμένου (*character* ή “chr”), κάτι που σημαίνει ότι θα χρειαστεί μετασχηματισμός κάποιων από αυτές, ούτως ώστε να είναι εφικτοί οι υπολογισμοί που θα πραγματοποιηθούν στη συνέχεια. Επί παραδείγματι, μεταβλητές που αποτελούνται από αριθμούς θα λάβουν αριθμητική μορφή (χρήσιμη για την εύρεση στατιστικών) και κάποιες άλλες μεταβλητές ως κατηγορηματικές (*factor*) θα αξιοποιηθούν συνδυαστικά, ώστε να γίνουν άλλοι υπολογισμοί, όπως θα δούμε αναλυτικά.

Το στοιχείο του χρόνου

Οι καταγραφές των συμβάντων αντιστοιχούν σε μια πλήρη ημερομηνία και ώρα, στοιχεία χρήσιμα για να απαντηθούν διάφορες ερωτήσεις όπως ποια είναι η συνολική περίοδος που κατεγράφησαν περιστατικά, πως έχουν κατανεμηθεί, πότε ήταν η μεγαλύτερη καταγραφή περιστατικών ημερολογιακά, αλλά και ωρολογιακά.

Μετατρέπεται, καταρχάς, η μεταβλητή που αφορά στις ημερομηνίες τις οποίες έλαβε χώρα έκαστο συμβάν σε έναν άλλον τύπο (POSIXlt), ώστε να προβούμε σε υπολογισμούς επί των χρονικών περιόδων. Για να γίνει με ακρίβεια η μετατροπή ακολουθείται πιστά η μορφή της ημερομηνίας και της ώρας, ακριβώς όπως αναγράφονται.

```
# from character to date and time
Incident.Date <- as.POSIXlt(Incident.Date, tz = "GMT", "%m/%d/%Y %I:%M:%S %p")
head(Incident.Date, 10); class(Incident.Date)

## [1] "2015-02-28 GMT" "2014-06-20 GMT" "2015-04-02 GMT" "2015-04-15 GMT"
## [5] "2015-04-08 GMT" "2014-12-05 GMT" "2015-01-19 GMT" "2015-02-28 GMT"
## [9] "2015-01-19 GMT" "2015-04-24 GMT"
```

```
## [1] "POSIXlt" "POSIXt"

# min(start) and max(end) date
range(Incident.Date)

## [1] "2011-02-04 GMT" "2015-05-04 GMT"
```

Ως πρώτη ημερομηνία στο σύστημα είναι καταγεγραμμένη η 4η Φεβρουαρίου του 2011 (2011-02-04) και επομένως έχουμε διαθέσιμα δεδομένα μόνο από το 2011 (μπορεί να οφείλεται σε εσωτερικό μετασχηματισμό / συγχωνεύσεις τμημάτων ή σε αναβάθμιση του τεχνικού εξοπλισμού) και έπειτα η τελευταία ημερομηνία είναι 4η Μαΐου του 2015 (2015-05-04) και γίνεται αντιληπτό ότι τα δεδομένα αποτελούν τη φωτογραφία μιας συγκεκριμένης χρονικής περιόδου στο παρελθόν και δεν είναι ιδιαίτερα πρόσφατα.

```
days <- as.numeric(max(Incident.Date) - min(Incident.Date))
months <- round(days/30)
hours <- days*24
cat("The total time period of recorded incidents is",
    months, "months or", days, "days or", hours, "hours.")

## The total time period of recorded incidents is 52 months or 1550 days or 37200 hours.
```

Υπολογίζεται η διαφορά μεταξύ της τελευταίας και της πρώτης ημερομηνίας, για να λάβουμε το συνολικό διάστημα σε μέρες. Με τη διαφορά αυτή υπολογίζεται το ίδιο διάστημα σε μήνες και σε ώρες. Γίνεται αναφορά, τελικώς, σε μια περίοδο τεσσάρων (4) περίπου ετών σύμφωνα με τα δεδομένα, η οποία διαφορετικά απολογείται σε 52 μήνες, ή 1550 ημέρες, ή 37,200 ώρες.

```
median(Incident.Date)

## [1] "2014-10-08 GMT"
```

Το δε μεσοδιάστημα αυτής της περιόδου φαίνεται πως είναι η 23η Οκτωβρίου του 2014, αρκετά μακριά από το 2011 που προαναφέραμε, γεγονός που αποτελεί ένδειξη ότι ίσως υπήρξε μεγάλη αύξηση των περιστατικών ή ότι υπήρχε σωστότερη ενημέρωση και υποστήριξη για τα άτομα που έφεραν έννομο συμφέρον ή και καλύτερη οργάνωση και πιο μεθοδική και σωστή διευθέτηση των περιστατικών.

```
Incident.Date[round(length(Incident.Date)/2)]

## [1] "2014-10-23 GMT"
```

Έπειτα αν ελέγξουμε μέχρι ποια ημερομηνία έχουν καταγραφεί τα μισά περιστατικά, διακρίνουμε ότι αυτή είναι η 23η Οκτωβρίου του 2014, μια ημερομηνία σχετικά στο τέλος του 2014. Τα υπόλοιπα μισά περιστατικά, το 50% έχει καταγραφεί μετά από αυτήν την ημερομηνία, άρα και είναι ως επί το πλείστον συσσωρευμένο στο 2015. Όσον αφορά συγκεκριμένη μέρα και

όχι κάποιο διάστημα, τα περισσότερα περιστατικά έχουν καταγραφεί την 28η Αυγούστου του 2014 και ήταν τέσσερα (4).

```
which.max(table(factor(Incident.Date)))      # index with max count

## 2014-08-28
##          43

table(factor(Incident.Date))["2014-08-28"]  # count of previous factor level

## 2014-08-28
##          4
```

Για την ώρα των περιστατικών δημιουργούμε έναν πίνακα με κριτήριο αν η ώρα είναι μικρότερη από 12 σε 24ωρο ρολόι, ώστε να ταξινομηθεί σε π.μ. και μ.μ.

```
# if smaller than 12, then it's AM, else it's PM
ampm <- factor(ifelse(Incident.Date$hour < 12, "AM", "PM"))
ampm

## [1] AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM
## [26] AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM
## [51] AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM
## [76] AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM
## [101] AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM
## [126] AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM AM
## [151] AM AM AM AM AM AM AM AM AM AM
## Levels: AM
```

Όλες οι ώρες είναι σε εικοσιτετράωρο ρολόι είναι μικρότερες από 12, και κατ' επέκταση είναι προ μεσημβρίας και δεν έχει καταχωρηθεί καθόλου το «PM». Για να είναι μετά στον πίνακα, το προσθέτουμε χειροκίνητα.

```
levels(ampm)[2] <- "PM"
levels(ampm)

## [1] "AM" "PM"

which.max(table(ampm))

## AM
## 1

which.max(table(Department, ampm))

## [1] 3

rownames(table(Department, ampm))[3]

## [1] "Austin Police Department"

addmargins(table(Department, ampm))
```

```
##
## Department
## (WPD) Watershed Protection and Development Review
## Austin Fire Dept
## Austin Police Department
## Austin Resource Recovery
## Austin Water
## Aviation
## Code Compliance Department
## EMS
## FSD - Building Services
## Health and Human Services
## Library
## Parks and Recreation
## Planning and Development Review
## Public Works
## Transportation
## Sum
```

	ampm		
	AM	PM	Sum
(WPD) Watershed Protection and Development Review	1	0	1
Austin Fire Dept	7	0	7
Austin Police Department	40	0	40
Austin Resource Recovery	23	0	23
Austin Water	24	0	24
Aviation	3	0	3
Code Compliance Department	1	0	1
EMS	10	0	10
FSD - Building Services	1	0	1
Health and Human Services	5	0	5
Library	1	0	1
Parks and Recreation	22	0	22
Planning and Development Review	3	0	3
Public Works	16	0	16
Transportation	2	0	2
Sum	159	0	159

```
round(addmargins(prop.table(table(Department, ampm)))*100,1)

##
## Department
## (WPD) Watershed Protection and Development Review
## Austin Fire Dept
## Austin Police Department
## Austin Resource Recovery
## Austin Water
## Aviation
## Code Compliance Department
## EMS
## FSD - Building Services
## Health and Human Services
## Library
## Parks and Recreation
## Planning and Development Review
## Public Works
## Transportation
## Sum
```

	ampm		
	AM	PM	Sum
(WPD) Watershed Protection and Development Review	0.6	0.0	0.6
Austin Fire Dept	4.4	0.0	4.4
Austin Police Department	25.2	0.0	25.2
Austin Resource Recovery	14.5	0.0	14.5
Austin Water	15.1	0.0	15.1
Aviation	1.9	0.0	1.9
Code Compliance Department	0.6	0.0	0.6
EMS	6.3	0.0	6.3
FSD - Building Services	0.6	0.0	0.6
Health and Human Services	3.1	0.0	3.1
Library	0.6	0.0	0.6
Parks and Recreation	13.8	0.0	13.8
Planning and Development Review	1.9	0.0	1.9
Public Works	10.1	0.0	10.1
Transportation	1.3	0.0	1.3
Sum	100.0	0.0	100.0

Τα περισσότερα περιστατικά έχουν καταγραφεί «πρωί» και μάλιστα στο Αστυνομικό Τμήμα του Όστιν (40 εγγραφές). Στον πίνακα φαίνεται ότι για την ακρίβεια κανένα περιστατικό δεν συνέβη μετά τις 12μ.μ. Μάλιστα, μπορούμε να δούμε ότι όλα τα περιστατικά έχουν δηλωθεί ως 12π.μ. (00:00) επειδή ίσως τότε καταχωρούνται στο σύστημα με αυτοματοποιημένο τρόπο. Επομένως, δεν γνωρίζουμε την ακριβή ώρα που συνέβη κάποιο περιστατικό.

Περαιτέρω διερεύνηση και εμφύτευση δεδομένων

Παρακάτω θα δούμε πιο διαγραμματικά τι συνέβη στα τμήματα της πόλης του Όστιν με τη χρήση πίνακα και ραβδογράμματος.

```
Department <- factor(Department)

# make level name shorter so that it later fits in the graph
levels(Department)[1] <- "WPD"
table <- sort((table(Department, useNA = "ifany")))
# name of table
names(dimnames(table)) <- "Department"

length(table)

## [1] 15

addmargins(table)

## Department
##
##          WPD          Code Compliance Department
##          1          1
##      FSD - Building Services          Library
##          1          1
##      Transportation          Aviation
##          2          3
## Planning and Development Review          Health and Human Services
##          3          5
##      Austin Fire Dept          EMS
##          7          10
##      Public Works          Parks and Recreation
##          16          22
##      Austin Resource Recovery          Austin Water
##          23          24
##      Austin Police Department          Sum
##          40          159

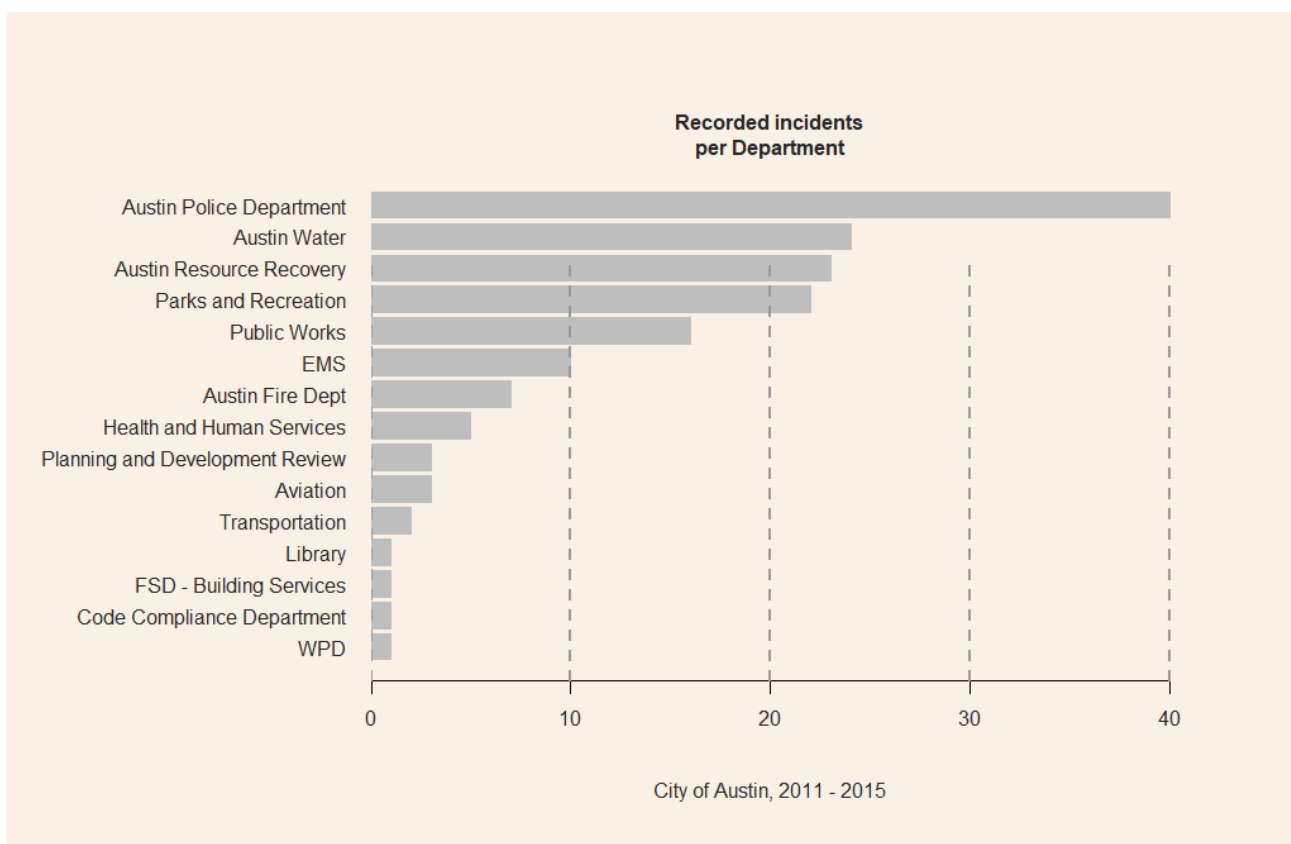
# adjust margins so that the text on the axis fits in the graph
par(mar = c(7,15,10,5), bg="linen")
# add bar chart
barplot(table,
  #orientation
  horiz = TRUE,
  las = 1,
  xlim = c(0, max(table(Department))),
  ylim = range(pretty(c(0, Department))),
  #text
  main = "Recorded incidents\nper Department",
  sub = "City of Austin, 2011 - 2015",
```



```

#size
  cex.main = 1,
#colors
  col.axis = "gray15",
  col.main = "gray15",
  col.sub = "gray15",
  col = "gray",
  border = "gray"
)
# add grid
grid(nx = NULL, ny = NA, lwd = 2, lty = 2, col = "gray60")

```



Συγκεκριμένα, μπορούμε να διακρίνουμε και πάλι ότι τα περισσότερα περιστατικά έχουν καταγραφεί στο Αστυνομικό Τμήμα του Όστιν. Ακολουθούν τα Austin Water με 24 εγγραφές, Austin Resource Recovery με 23, τα Parks and Recreation με 22. Έπειτα τα Public Works με 16 εγγραφές, το EMS με 10, το Πυροσβεστικό τμήμα με μόλις 7 και το Health and Human Services με 5. Παρακάτω έχει γίνει ακόμη μικρότερη καταγραφή περιστατικών, έτσι ώστε έχουν μια

συχνότητα λιγότερη από 1 περιστατικό ανά έτος και κατ' επέκταση έχουν μικρότερο ενδιαφέρον.

Στη συνέχεια, θα ελέγξουμε σε ποια κατηγορία ασφαλιστικού προγράμματος απαντώνται τα μεγαλύτερα ποσά συνολικά και κατά μέσο όρο. Εφόσον έχουμε δει ότι τα ποσά έχουν μορφή χαρακτήρα («λέξης»), θα αφαιρέσουμε το σύμβολο του δολαρίου και θα μετατρέψουμε τη μεταβλητή σε αριθμητική.

```
# remove "$" symbol and turn character to numeric
Amount <- as.numeric(substring(Amount, 2, nchar(Amount)))
head(Amount); class(Amount)

## [1] 700.00 7000.00 947.17 1133.42 13663.31 565.88

## [1] "numeric"
```

Με μια γρήγορη σύνοψη των περιγραφικών στατιστικών παρατηρούμε ότι υπάρχει μία εγγραφή NA (δηλαδή που δεν είναι διαθέσιμη), άρα σε κάποιο περιστατικό δεν καταχωρήθηκε το ποσό της αποζημίωσης. Παράλληλα, φαίνεται ότι το μεγαλύτερο ποσό αποζημίωσης ανέρχεται στα 155,000.00 δολάρια, ποσό ύποπτα υψηλό που μπορεί να αποτελεί ένδειξη για κάποιο τυπογραφικό λάθος ή για μια πολύ ειδική και σπάνια περίπτωση. Καθώς το ποσό αυτό επηρεάζει αρκετά τον μέσο όρο και είναι πιθανό να έχουμε παραποιημένα αποτελέσματα, θα προβούμε σε κάποιες τροποποιήσεις, για να έχουμε όσο γίνεται πιο σωστά και ακριβή αποτελέσματα.

```
summary(Amount)
```

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
##	50.0	587.3	1235.7	3643.2	3182.4	155000.0	1

Αρχικά, σχετικά με τη μη διαθέσιμη τιμή, την αντικαθιστούμε με τη διάμεσο των ποσών αποζημίωσης, για να μην εμποδίσει τους υπολογισμούς μας. Εφόσον υπάρχουν ύποπτα υψηλές τιμές και ο μέσος όρος έχει μεγάλη απόκλιση από τη διάμεσο, συμπεραίνουμε ότι η κατανομή των δεδομένων φέρει μια λοξότητα και για αυτόν τον λόγο προτιμάται η αντικατάσταση με τη διάμεσο και όχι με τον μέσο όρο. Όσον αφορά τα 155,000.00 δολάρια (και άλλα ύποπτα υψηλά ποσά που ενδέχεται να υπάρχουν), θα αντικαταστήσουμε τις ενδεχόμενες ακραίες τιμές που υπάρχουν με τον κανόνα του «Tukey», με την οποία θα αντικαταστήσουμε με τη διάμεσο τις τιμές που έχουν απόκλιση από το 75% των παρατηρήσεων όσο είναι 1.5 φορές το ενδοτεταρτημοριακό εύρος των ποσών. Αξίζει να σημειωθεί ότι βάσει του κανόνα ως ακραίες τιμές θεωρούνται και αυτές που είναι 1.5 φορές το ενδοτεταρτημοριακό εύρος μακριά από το 25% των παρατηρήσεων. Αυτό το νοητό περιθώριο χαρακτηρίζεται ως «Tukey's fences», αλλά στην προκειμένη περίπτωση δεν θα χρησιμοποιήσουμε καθόλου τον δεύτερο φράκτη, καθώς δεν υπάρχουν καθόλου αρνητικές τιμές και δεν νοείται να υπάρχουν αρνητικές τιμές σε ποσά απαιτήσεων (δεν αξιώνει κάποιος αρνητική αποζημίωση, κάτω του μηδενός).

```
# NA treatment
Amount[is.na(Amount)] <- median(Amount, na.rm = TRUE)

# outlier treatment
Q3 <- unname(quantile(Amount)[4])
outliers <- which(Amount > Q3 + 1.5 * IQR(Amount))
for(i in 1:length(Amount[outliers])){
  Amount[outliers][i] <- median(Amount)
}

summary(Amount)

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      50.0   588.2  1235.7  1651.1  1991.4  7000.0
```

Μετά τις τροποποιήσεις μπορούμε να δούμε κατ' αρχήν ότι ο μέσος όρος (\$ 1651.1) πλέον βρίσκεται σε πολύ μικρή απόκλιση από τη διάμεσο (\$ 1235.7) και, επιπλέον, ότι συνολικά και κατά μέσο όρο τα μεγαλύτερα ποσά απαντώνται στην κατηγορία 00 Auto με συνολικό ποσό \$ 239,726.7...

```
which.max(tapply(Amount, Category, sum)) # which index

## 00 Auto
##      1

tapply(Amount, Category, sum)["00 Auto"] # how much based on previous result

## 00 Auto
## 239726.7
```

...και κατά μέσο όρο \$ 1,887.61.

```
which.max(tapply(Amount, Category, mean))

## 00 Auto
##      1

tapply(Amount, Category, mean)["00 Auto"]

## 00 Auto
## 1887.612
```

Όπως είναι φυσικό, το ποσό των \$ 155,000.00 θα άλλαζε σημαντικά τα αποτελέσματα όσον αφορά και το σύνολο και τον μέσο όρο και θα καταλήγαμε απευθείας.

Ανάλυση κειμένου για εξόρυξη δεδομένων

Δεδομένου ότι έχουμε στη διάθεσή μας και τα ονοματεπώνυμα κάθε ζημιωθέντος προσώπου μπορούμε, επιπλέον, να δούμε τη συχνότητα εμφάνισης ενός (βαπτιστικού) ονόματος στη λίστα των ατόμων που έχουν αιτηθεί αποζημίωσης από την πόλη. Κάποια από τα ονόματα με μια ματιά είναι ως εξής:

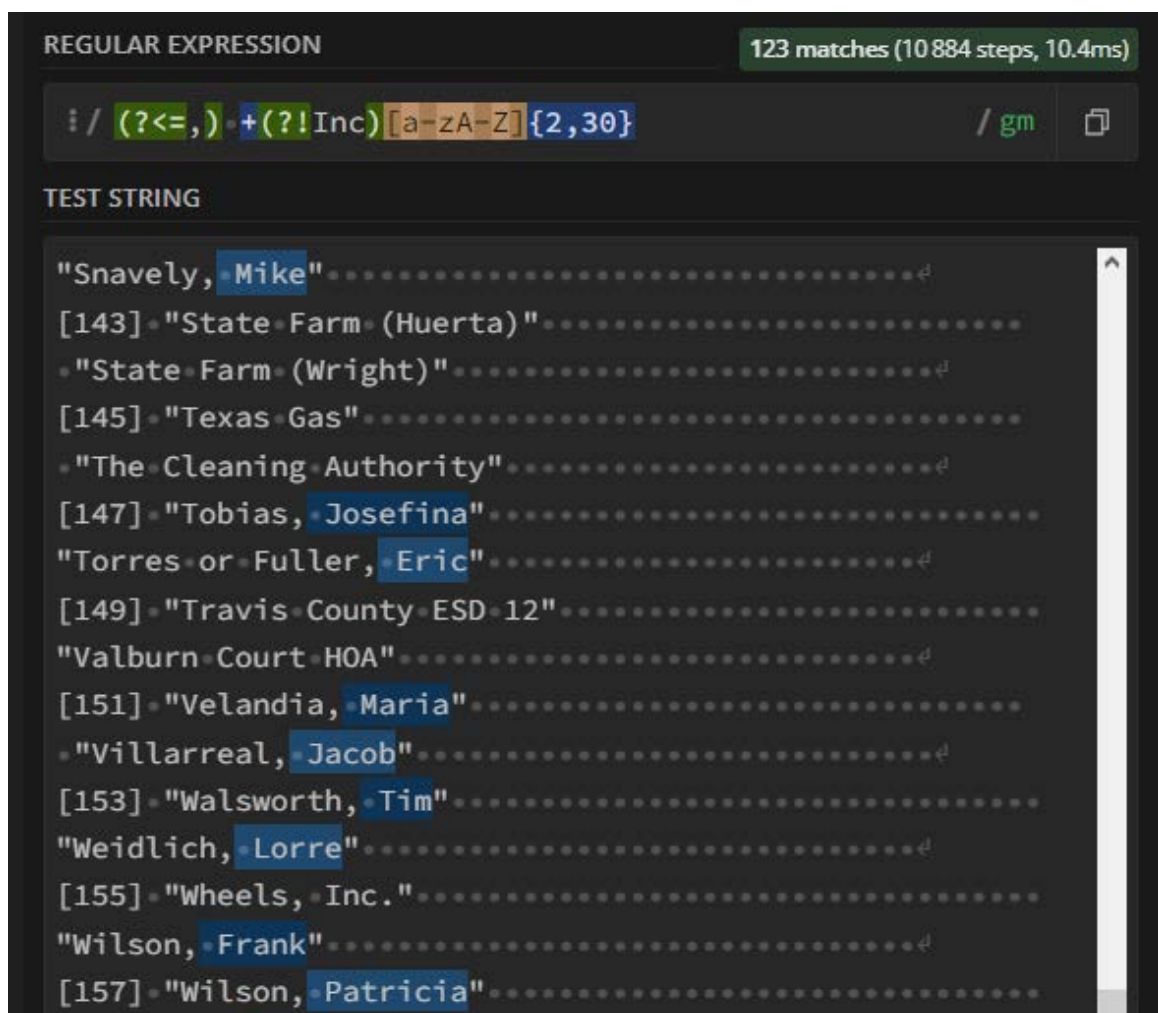
```
head(Claim.Name,35)

## [1] "Valdez, Michael"
## [2] "Caballero, Rosita"
## [3] "Castillo, Angela"
## [4] "Cutean, Nicki"
## [5] "Ferrovial/ Toyota Lease Trust"
## [6] "Hardwick, Andrew"
## [7] "Hardy, Ryan"
## [8] "Henna Chevrolet"
## [9] "Hertz (Hardy)"
## [10] "Kempter, Anne"
## [11] "Kenyon, Rachel"
## [12] "Lopez de Araya Bengoa, Imanol (MINOR INVOLVED)"
## [13] "Scott, Nadia"
## [14] "Aleman, Nelda"
## [15] "Blair, Jackie"
## [16] "Grays, Jessica"
## [17] "Katz, Lauren"
## [18] "Rayburn, Charles"
## [19] "USAA (Butler, Bill)"
## [20] "Allstate (Pitts)"
## [21] "AT&T (150280)"
## [22] "AT&T (150480)"
## [23] "Dawson, Jennifer"
## [24] "Reed Hawkins, Tam"
## [25] "Rodriguez, Gina"
## [26] "Saucedo, Blondena"
## [27] "Texas Gas (150257)"
## [28] "Seton (Brackenridge)"
## [29] "Stringer, John"
## [30] "Guo, Sue"
## [31] "Washington, Cynthia"
## [32] "Garcia, Miranda"
## [33] "Goodall, Jannette"
## [34] "Herbert, Robert"
## [35] "Martinez, Rudy"
```

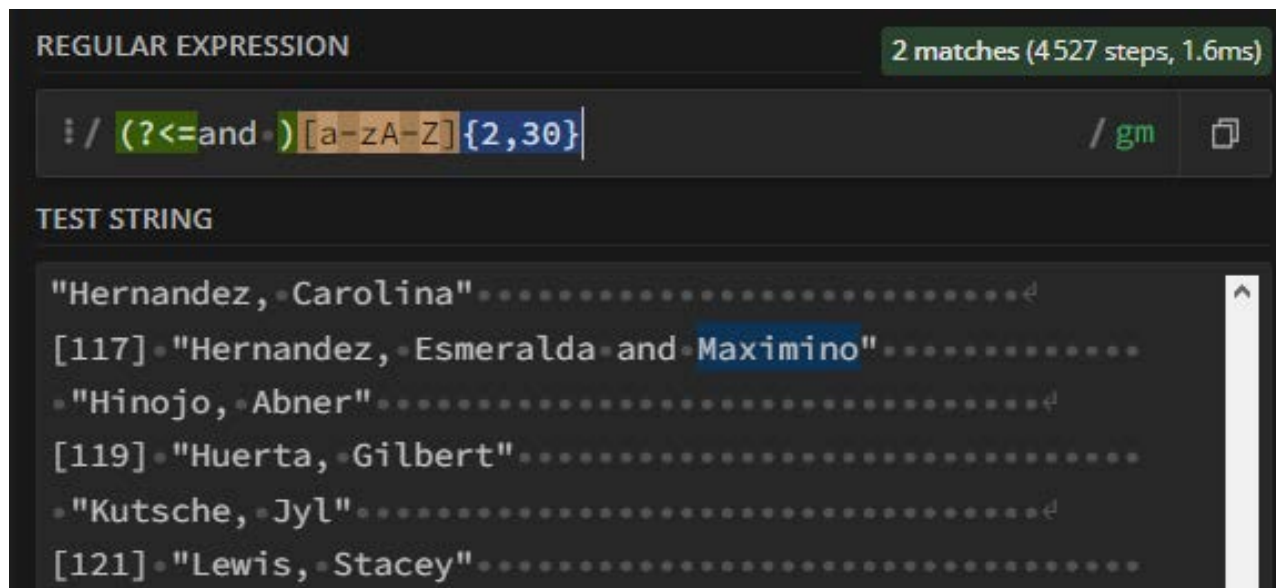
Όπως φαίνεται δεν υπάρχει μια ξεκάθαρη συνεπής δομή. Σε μία μόνο στήλη είναι μαζί το όνομα, το επώνυμο και διάφορες σημειώσεις. Ενίοτε, όταν πρόκειται για εμπλοκή ανήλικου ή

όταν πρόκειται για ΝΠΙΔ. Χρειάζεται να «τραβήξουμε» τα βαπτιστικά ονόματα, τα οποία, ωστόσο, διαφέρουν σε θέση (μπορεί να βρίσκονται μέσα σε μια παρένθεση) και στο μήκος (ένα όνομα μπορεί να είναι πολύ μικρό, ενώ ένα άλλο μεγαλύτερο και επομένως δεν μπορούμε να θέσουμε εξ αρχής πόσους χαρακτήρες θα τραβήξουμε από τον πίνακα. Αυτό το οποίο φαίνεται πως είναι κοινό σε όλες τις εγγραφές είναι ότι πάντοτε όταν υπάρχει ονοματεπώνυμο, προηγείται το επίθετο του φυσικού προσώπου, έπειτα υπάρχει κόμμα, κενό, και ακολουθεί το βαπτιστικό του όνομα. Επομένως θα εκμεταλλευτούμε το μοτίβο «κόμμα, κενό, ομάδα χαρακτήρων (όνομα)» για να απομονώσουμε κάθε ομάδα χαρακτήρων πριν από την οποία προηγούνται αυτά.

Ελέγχοντας τον πίνακα με τα ονόματα στο σύνολό του, γίνεται αντιληπτό ότι και πάλι υπάρχουν κάποιες μικρές εξαιρέσεις. Εμφανίζεται στα ονόματα και η λέξη «Inc», καθώς επίσης αποκλείονται κάποια ονόματα τα οποία δεν ακολουθούν κόμμα και κενό, αλλά τη λέξη «and», γιατί είναι ζευγαράκι με άλλο όνομα.



The screenshot shows a web-based Regular Expression tool. At the top, it says "REGULAR EXPRESSION" and "123 matches (10 884 steps, 10.4ms)". The input field contains the regex pattern: `/(?<=,)+(?!Inc)[a-zA-Z]{2,30}/gm`. Below the input field, there is a "TEST STRING" section. The test string contains several lines of text, each representing a record. The records are: "Snavely, Mike", "[143] State Farm (Huerta)", "State Farm (Wright)", "[145] Texas Gas", "The Cleaning Authority", "[147] Tobias, Josefina", "Torres or Fuller, Eric", "[149] Travis County ESD 12", "Valburn Court HOA", "[151] Velandia, Maria", "Villarreal, Jacob", "[153] Walsworth, Tim", "Weidlich, Lorre", "[155] Wheels, Inc.", "Wilson, Frank", and "[157] Wilson, Patricia". The matches are highlighted in blue, showing the pattern successfully extracting names and company names from the test string.



Καταληκτικά, με τη χρήση «regular expressions» ζητάμε από το πρόγραμμα να μας επιστρέψει μια ομάδα χαρακτήρων (κεφαλαίων ή πεζών) από το A έως το Z με μήκος το λιγότερο 2 χαρακτήρες (για το μικρότερο πιθανό όνομα) και έως 30 χαρακτήρες. Η ομάδα αυτή να ακολουθεί το σύμβολο του κόμματος και έπειτα ένα ή περισσότερα κενά και παράλληλα να μην είναι η λέξη «Inc».

```
# find a group of characters from A to Z, that is not "Inc", minimum 2 characters with
# a max of 30 that follows the pattern ", "
list1 <- regmatches(Claim.Name, regexpr("(?<=,) +(?!(Inc))[a-zA-Z]{2,30}", Claim.Name,
perl = TRUE))
list1
```

##	[1]	" Michael"	" Rosita"	" Angela"	" Nicki"	" Andrew"
##	[6]	" Ryan"	" Anne"	" Rachel"	" Imanol"	" Nadia"
##	[11]	" Nelda"	" Jackie"	" Jessica"	" Lauren"	" Charles"
##	[16]	" Bill"	" Jennifer"	" Tam"	" Gina"	" Blondena"
##	[21]	" John"	" Sue"	" Cynthia"	" Miranda"	" Jannette"
##	[26]	" Robert"	" Rudy"	" Rebecca"	" Raymond"	" Laurel"
##	[31]	" Bob"	" Dorcas"	" Christie"	" Daniel"	" Max"
##	[36]	" Sapna"	" Cassandra"	" Debra"	" Brian"	" Jennifer"
##	[41]	" Aditya"	" Catina"	" Laura"	" Danielle"	" Katrina"
##	[46]	" Dylan"	" Mauricio"	" James"	" Nick"	" Mike"
##	[51]	" Paul"	" Jennifer"	" Raymond"	" Maureen"	" Jon"
##	[56]	" Gilbert"	" Lisa"	" Kristy"	" Juan"	" Eduardo"
##	[61]	" Anna"	" Ana"	" Maria"	" Stephanie"	" Craig"
##	[66]	" Nicole"	" Matthew"	" Orlando"	" Ellen"	" William"
##	[71]	" Robert"	" Dennis"	" Marley"	" Ronnie"	" Marissa"
##	[76]	" Jessica"	" Delia"	" Bonnie"	" Andre"	" Brien"
##	[81]	" Paula"	" Alfredo"	" Sheri"	" Jesse"	" Debra"
##	[86]	" Nicole"	" Husain"	" Jose"	" Mary"	" William"

```
## [91] " Carolina" " Esmeralda" " Abner"      " Gilbert"  " Jyl"
## [96] " Stacey"    " Nancy"     " Maricela" " Alan"     " Luis"
## [101] " Bianca"   " Laura"     " Odalys"   " Patrick"  " Fernando"
## [106] " Dolores"  " Jonathan"  " Sean"     " Matthew"  " Phillip"
## [111] " Diana"    " Pauline"   " Mike"     " Josefina" " Eric"
## [116] " Maria"    " Jacob"     " Tim"      " Lorre"    " Frank"
## [121] " Patricia" " Tim"       " Sean"
```

Τώρα έχουμε καταλήξει με μια λίστα με ονόματα που μπροστά τους έχουν ακόμη κενό. Επαναλαμβάνεται η διαδικασία ζητώντας μόνο χαρακτήρες.

```
# find a group of characters from A to Z, minimum 2 characters with a max of 30
list1 <- regmatches(list1, regexpr("[a-zA-Z]{2,30}", list1, perl = TRUE))
list1

## [1] "Michael" "Rosita" "Angela" "Nicki" "Andrew" "Ryan"
## [7] "Anne" "Rachel" "Imanol" "Nadia" "Nelda" "Jackie"
## [13] "Jessica" "Lauren" "Charles" "Bill" "Jennifer" "Tam"
## [19] "Gina" "Blondena" "John" "Sue" "Cynthia" "Miranda"
## [25] "Jannette" "Robert" "Rudy" "Rebecca" "Raymond" "Laurel"
## [31] "Bob" "Dorcas" "Christie" "Daniel" "Max" "Sapna"
## [37] "Cassandra" "Debra" "Brian" "Jennifer" "Aditya" "Catina"
## [43] "Laura" "Danielle" "Katrina" "Dylan" "Mauricio" "James"
## [49] "Nick" "Mike" "Paul" "Jennifer" "Raymond" "Maureen"
## [55] "Jon" "Gilbert" "Lisa" "Kristy" "Juan" "Eduardo"
## [61] "Anna" "Ana" "Maria" "Stephanie" "Craig" "Nicole"
## [67] "Matthew" "Orlando" "Ellen" "William" "Robert" "Dennis"
## [73] "Marley" "Ronnie" "Marissa" "Jessica" "Delia" "Bonnie"
## [79] "Andre" "Brien" "Paula" "Alfredo" "Sheri" "Jesse"
## [85] "Debra" "Nicole" "Husain" "Jose" "Mary" "William"
## [91] "Carolina" "Esmeralda" "Abner" "Gilbert" "Jyl" "Stacey"
## [97] "Nancy" "Maricela" "Alan" "Luis" "Bianca" "Laura"
## [103] "Odalys" "Patrick" "Fernando" "Dolores" "Jonathan" "Sean"
## [109] "Matthew" "Phillip" "Diana" "Pauline" "Mike" "Josefina"
## [115] "Eric" "Maria" "Jacob" "Tim" "Lorre" "Frank"
## [121] "Patricia" "Tim" "Sean"
```

Στη συνέχεια δεν παραλείπεται να συμπερίληψη στα ονόματα (ομάδες χαρακτήρων) και αυτών που ακολουθούν το μοτίβο «and (κενό)», ούτως ώστε να καταλήξουμε με έναν πλήρη πίνακα.

```
# find a group of characters from A to Z, minimum 2 characters with a max of 30, that
# follows the pattern "and "
list2 <- regmatches(Claim.Name, regexpr("(?<=and ) [a-zA-Z]{2,30}", Claim.Name, perl =
TRUE))
list2

## [1] "Marilu" "Maximino"
```

```

firstnames <- factor(c(list1, list2))
firstnames

## [1] Michael Rosita Angela Nicki Andrew Ryan Anne
## [8] Rachel Imanol Nadia Nelda Jackie Jessica Lauren
## [15] Charles Bill Jennifer Tam Gina Blondena John
## [22] Sue Cynthia Miranda Jannette Robert Rudy Rebecca
## [29] Raymond Laurel Bob Dorcas Christie Daniel Max
## [36] Sapna Cassandra Debra Brian Jennifer Aditya Catina
## [43] Laura Danielle Katrina Dylan Mauricio James Nick
## [50] Mike Paul Jennifer Raymond Maureen Jon Gilbert
## [57] Lisa Kristy Juan Eduardo Anna Ana Maria
## [64] Stephanie Craig Nicole Matthew Orlando Ellen William
## [71] Robert Dennis Marley Ronnie Marissa Jessica Delia
## [78] Bonnie Andre Brien Paula Alfredo Sheri Jesse
## [85] Debra Nicole Husain Jose Mary William Carolina
## [92] Esmeralda Abner Gilbert Jyl Stacey Nancy Maricela
## [99] Alan Luis Bianca Laura Odalys Patrick Fernando
## [106] Dolores Jonathan Sean Matthew Phillip Diana Pauline
## [113] Mike Josefina Eric Maria Jacob Tim Lorre
## [120] Frank Patricia Tim Sean Marilu Maximino
## 110 Levels: Abner Aditya Alan Alfredo Ana Andre Andrew Angela Anna ... William

```

Εν τέλει τα ονόματα στο σύνολό τους είναι 110, συγκριτικά με τις 159 εγγραφές που ξέρουμε ότι υπάρχουν. Εφόσον ξέρουμε ότι κάποιο/α από αυτά επαναλαμβάνεται/ονται προχωράμε στην εύρεση του ονόματος με τη μεγαλύτερη συχνότητα εμφάνισης.

```
nlevels(firstnames)
```

```
## [1] 110
```

Το όνομα αυτό είναι το «Jennifer» και έχει εμφανιστεί στις εγγραφές μόνο 3 φορές. Παρακάτω μπορούμε να δούμε ότι κάθε φορά πρόκειται για διαφορετικό πρόσωπο και επομένως σε συνδυασμό με τη συχνότητα εμφάνισης πρόκειται απλώς για σύμπτωση.

```
which.max(table(firstnames))
```

```
## Jennifer
```

```
## 48
```

```
table(firstnames)["Jennifer"]
```

```
## Jennifer
```

```
## 3
```

```
# get observations that include the string "Jennifer"
```

```
Claim.Name[grep("Jennifer", Claim.Name)]
```

```
## [1] "Dawson, Jennifer" "Burrough, Jennifer" "Liu, Jennifer"
```

```
detach(legal)
```