**Εργασία Στατιστικής με R**

Ματάκος Αλέξανδρος ΑΕΜ: 15777

**Σύνολο Δεδομένων:**

Arrests for Marijuana Possession

Περιγραφή

Δεδομένα σχετικά με την αστυνομική μεταχείριση ατόμων που συνελήφθησαν στο Τορόντο για απλή κατοχή μικρών ποσοτήτων μαριχουάνας. Τα δεδομένα αποτελούν μέρος ενός μεγαλύτερου συνόλου δεδομένων που παρουσιάζεται σε μια σειρά άρθρων στην εφημερίδα Toronto Star.

Πηγή: <https://vincentarelbundock.github.io/Rdatasets/doc/carData/Arrests.html>

Format

Data frame με 5226 παρατηρήσεις στις ακόλουθες 8 μεταβλητές:

**released**: αν απελευθερώθηκε ο κρατούμενος με κλήτευση ή όχι

**colour**: το χρώμα του κρατούμενου

**year:** 1997 έως 2002

**age:** η ηλικία του κρατούμενουσε έτη

**sex**: το φύλο του κρατούμενου

**employed**: αν ο κρατούμενος εργάζεται

**citizen**: αν ο άνθρωπος που συλλαμβάνεται είναι πολίτης ή όχι

**checks**: Πλήθος αστυνομικών βάσεων δεδομένων (προηγούμενων συλλήψεων, καθίρξεων, αν είναι σε αναστολή κλπ. – σύνολο 6) στις οποίες εμφανίζεται το όνομα του συλληφθέντος

Ερώτημα 1

released: Ποιοτική μεταβλητή με levels **No/Yes**

colour: Ποιοτική μεταβλητή με levels **Black/White**

year: Ποσοτική μεταβλητή, αριθμητικό διάνυσμα

age: Ποσοτική μεταβλητή, αριθμητικό διάνυσμα

sex: Ποιοτική μεταβλητή με levels **Female/Male**

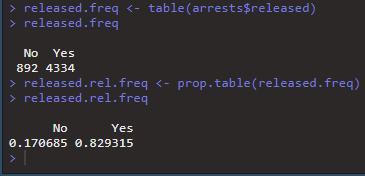
employed: Ποιοτική μεταβλητή με levels No/Yes

citizen: Ποιοτική μεταβλητή με levels **No/Yes**

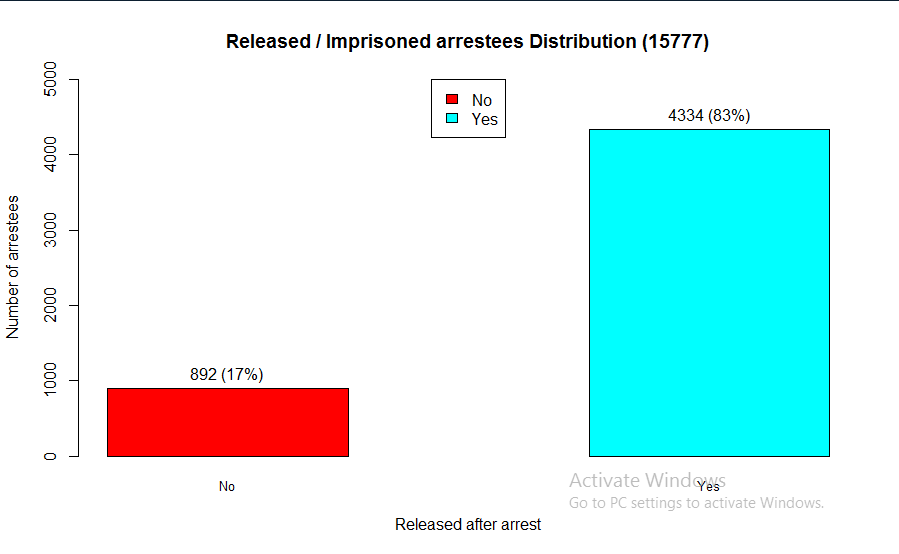
checks: Ποσοτική μεταβλητή, αριθμητικό διάνυσμα

Ερώτημα 2

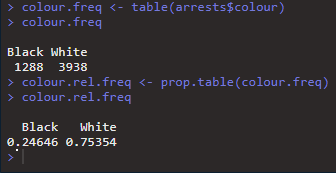
Πίνακες συχνοτήτων / σχετικών συχνοτήτων της μεταβλητής **released**: 15777



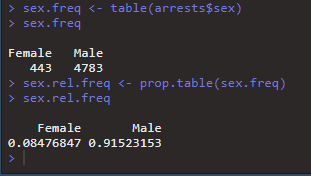
Ραβδόγραμμα για τη μεταβλητή released:



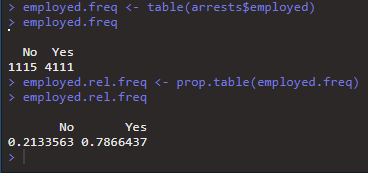
Πίνακες συχνοτήτων / σχετικών συχνοτήτων της μεταβλητής **colour:** 15777



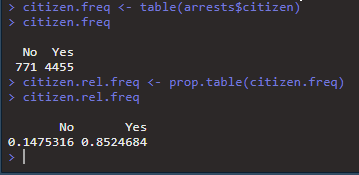
Πίνακες συχνοτήτων / σχετικών συχνοτήτων της μεταβλητής **sex:** 15777



Πίνακες συχνοτήτων / σχετικών συχνοτήτων της μεταβλητής **employed**: 15777

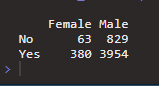


Πίνακες συχνοτήτων / σχετικών συχνοτήτων της μεταβλητής **citizen**: 15777

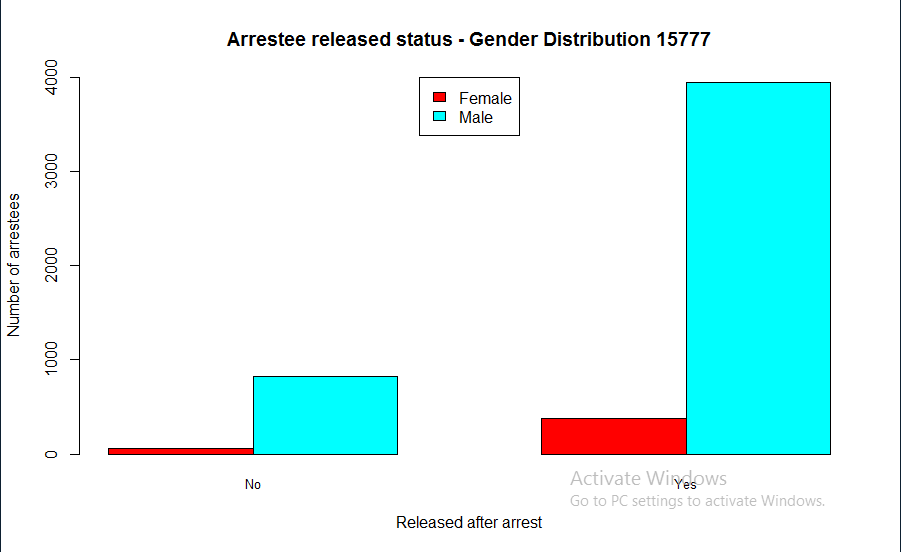


Ερώτημα 3

Cross-tabulation matrix της μεταβλητής released ως προς το φύλο (μεταβλητή sex). Επέλεξα αυτές τις μεταβλητές για να εξετάσω αν υπάρχει κάποιο bias στη μεταχείριση των κρατούμενων λόγω φύλου.

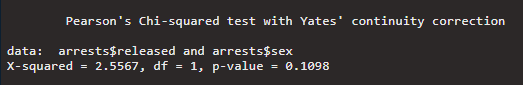
 15777

Ομαδοποιημένο ραβδόγραμμα της μεταβλητής released ως προς το φύλο (μεταβλητή sex)



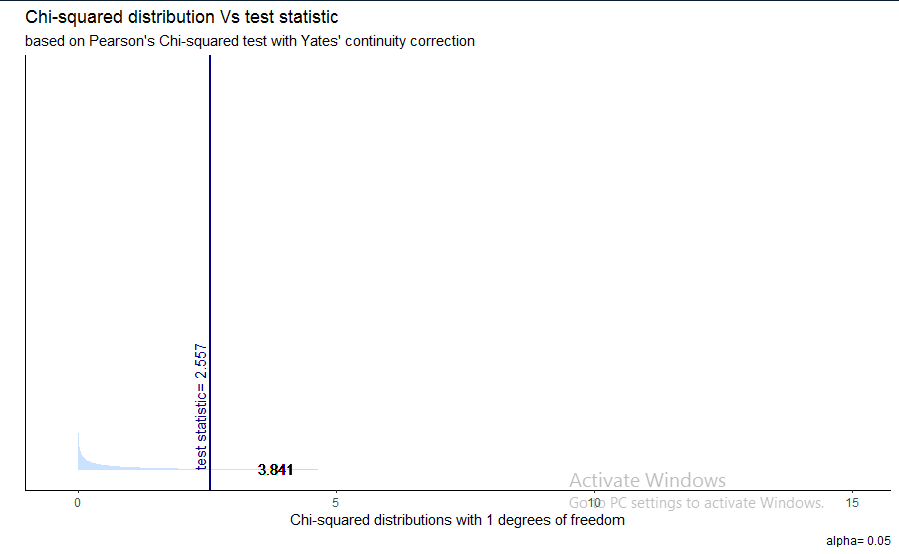
Chi-squared contingency table test

Η0: Οι δύο μεταβλητές είναι ανεξάρτητες Η1: Οι δύο μεταβλητές είναι εξαρτημένες



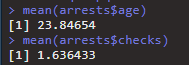
Με τιμή p-value = 0.1 και α = 0.05 δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση ότι είναι ανεξάρτητες, αλλά αν έπρεπε να βγάλουμε κάποιο συμπέρασμα θα λέγαμε ότι κατά πάσα πιθανότητα οι δύο μεταβλητές φαίνεται να είναι εξαρτημένες.

Οπτικοποίηση του αποτελέσματος του ελέγχου με gginference 15777



Ερώτημα 4

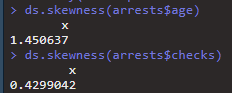
Υπολογισμός στατιστικών για τις ποσοτικές μεταβλητές age, checks.

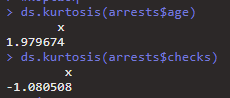
Μέση τιμή: 

Διάμεσος: 

Τυπική απόκλιση: 

Διασπορά: 

Λοξότητα: 

Κύρτωση: 

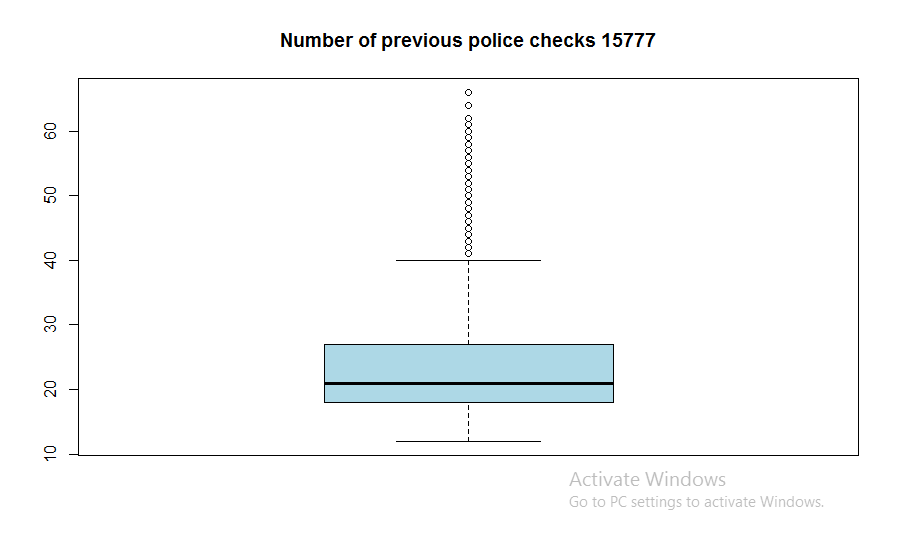
Εύρος: 

Ενδοτεταρτημοριακό εύρος: 

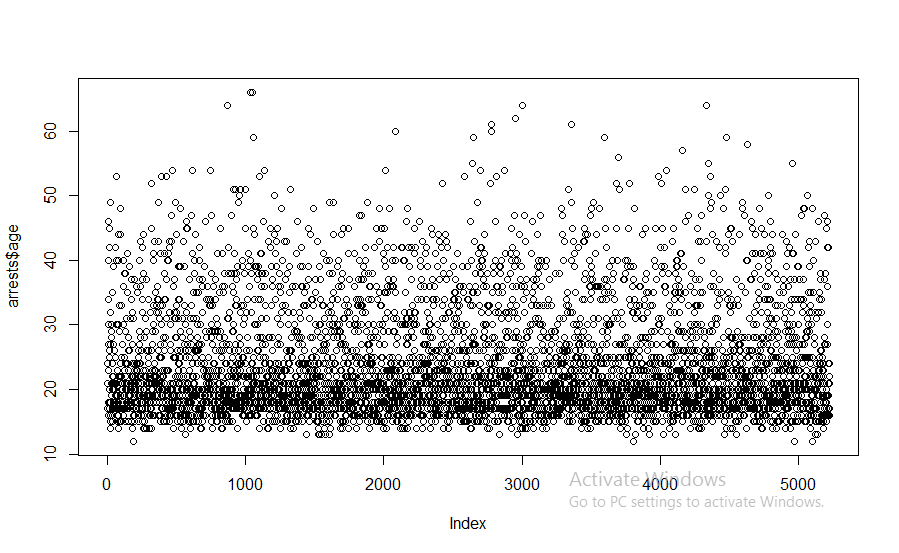
Θηκόγραμμα για τη μεταβλητή age:



Θηκόγραμμα για τη μεταβλητή checks:



Scatter plot για τη μεταβλητή age:

15777

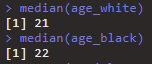
Οι παρατηρήσεις συγκεντρώνονται στις μικρότερες ηλικίες και να αραιώνουν όσο ανεβαίνουμε σε μεγαλύτερες ηλικίες.

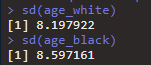
Φαίνεται να ακολουθούν κανονική κατανομή με λοξότητα στραμμένη προς τις μεγαλύτερες ηλικίες.

Ερώτημα 5

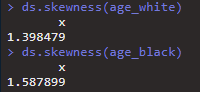
Υπολογισμός των στατιστικών της μεταβλητής age ως προς τις κατηγορίες της μεταβλητής colour (White/Black).

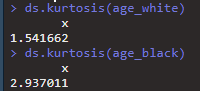
Μέση τιμή: 

Διάμεσος: 

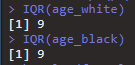
Τυπική απόκλιση: 

Διασπορά: 

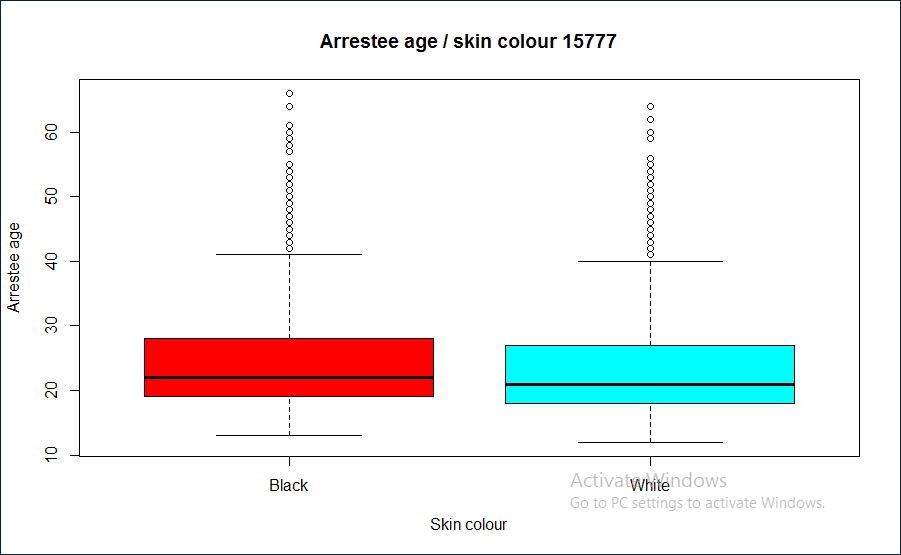
Λοξότητα: 

Κύρτωση: 

Εύρος: 

Ενδοτεταρτημοριακό εύρος: 

Συγκριτικό θηκόγραμμα της μεταβλητής age ως προς τις κατηγορίες της μεταβλητής colour:

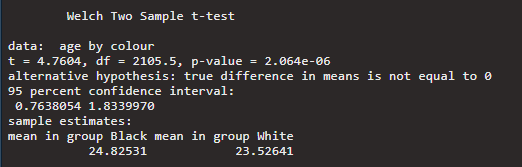


Η κατανομή ακολουθεί κανονική κατανομή με λοξότητα στραμμένη προς τα δεξιά (μεγαλύτερες ηλικίες) και για τις δύο κατηγορίες.

Ερώτημα 6

Θα χρησιμοποιήσουμε πάλι τις μεταβλητές age, colour. Εφόσον το πλήθος του δείγματος ειναι αρκετά μεγάλο(5200) και έχουμε ήδη δει ότι η μεταβλητή age ακολουθεί κανονική κατανομή μπορούμε να χρησιμοποιήσουμε t-test για τη σύγκριση των μέσων τιμών για τις κατηγορίες White/Black.

15777

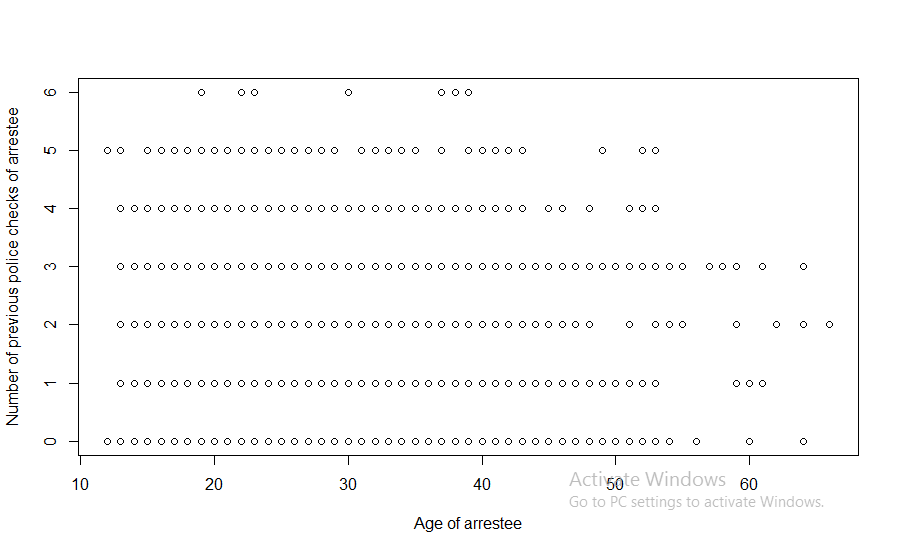


Άρα συμπεραίνουμε ότι οι μέσες τιμές είναι κατά πάσα πιθανότητα ίσες

Ερώτημα 7

Εξετάζουμε αν υπάρχει κάποια γραμμική συσχέτιση μεταξύ των μεταβλητών age και checks:

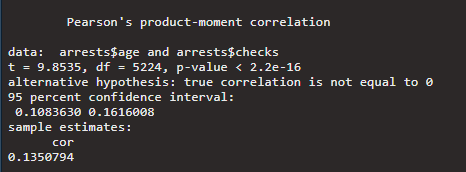
15777



#(Σχόλιο): Έπειτα προσθέτοντας τη γραμμή παλινδρόμισης δεν άλλαξε κάτι στο σχήμα. Ίσως είναι πολύ μεγάλο το δείγμα για μια τέτοιο σύγκριση μέσω της εντολής plot() ?

Συντελεστής συσχέτισης του Pearson:

15777



Οι δύο μεταβλητές δεν φαίνεται να είναι συσχετισμένες.