2

Δώστε μια σύντομη περιγραφή από που προέρχονται τα δεδομένα και πόσες περιπτώσεις περιέχονται.

Τα δεδομένα προέρχονται από την επίσημη ιστοσελίδα της κυβέρνησης της πολιτείας της Ουάσιγκτον (Washington) https://catalog.data.gov/dataset/electric-vehicle-population-data
και αφορούν τον πληθυσμό των ηλεκτρικών οχημάτων στην πολιτεία. Το σύνολο των δεδομένων περιέχει πληροφορίες για κάθε καταγεγραμμένο ηλεκτρικό όχημα, συμπεριλαμβανομένων χαρακτηριστικών όπως η μάρκα, το μοντέλο, το έτος κατασκευής, ο τύπος του οχήματος, η εμβέλεια της μπαταρίας (Electric Range), η τιμή καταλόγου (Base MSRP), καθώς και η παροχή ηλεκτρικής ενέργειας (Electric Utility).

Η αρχική βάση δεδομένων περιλαμβάνει συνολικά **210,165** περιπτώσεις (εγγραφές).

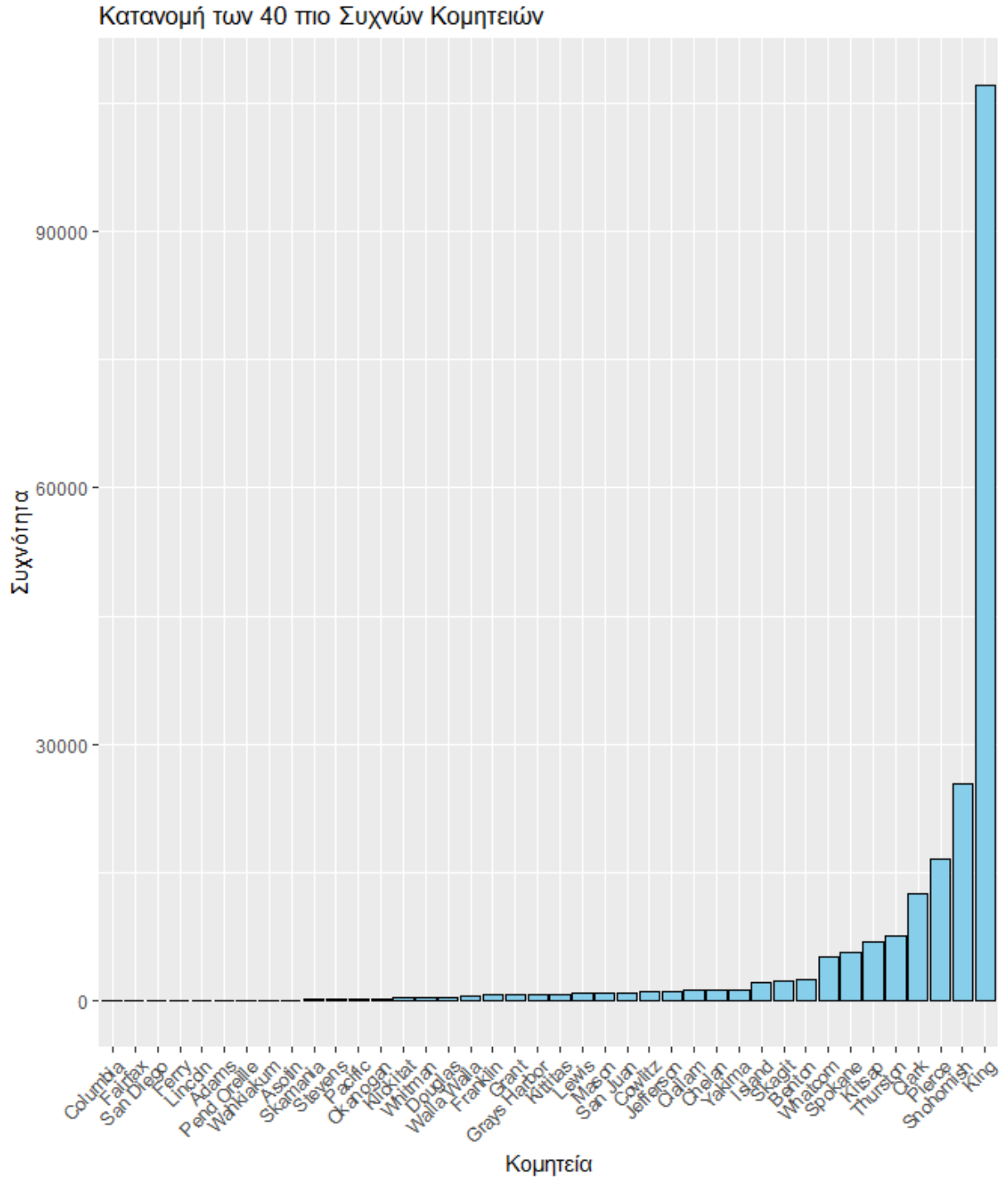
```
> colnames(cleaned_data)
 [1] "County"
 [2] "City"
 [3] "Model Year"
 [4] "Make"
 [5] "Model"
 [6] "Electric Vehicle Type"
 [7] "Clean Alternative Fuel Vehicle (CAFV) Eligibility"
 [8] "Electric Range"
 [9] "Base MSRP"
[10] "Electric Utility"
>
```

b. Ποιες είναι κατηγορικές και ποιες ποσοτικές μεταβλητές; Δώστε μια σύντομη περιγραφή κάθε μιας από αυτές (ή ορισμένων εάν είναι πάρα πολλές).

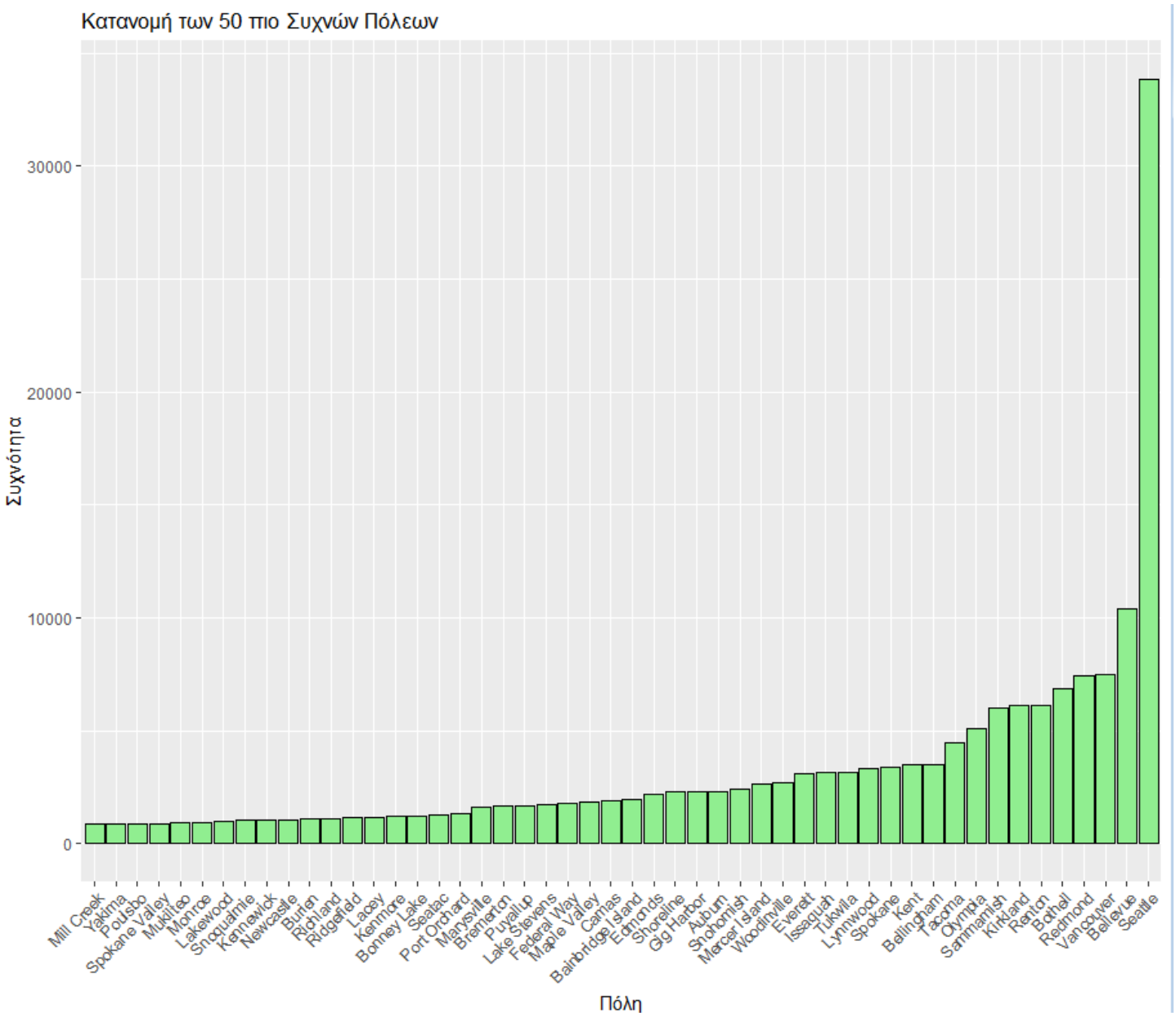
Κατηγορικές Μεταβλητές:	
County	Η περιοχή/κομητεία όπου καταχωρείται το όχημα. Κατηγορική μεταβλητή με ονομαστικές τιμές.
City	Η πόλη όπου καταχωρείται το όχημα. Κατηγορική μεταβλητή με ονομαστικές τιμές.
Make	Η μάρκα του οχήματος (π.χ., Tesla, Nissan). Κατηγορική μεταβλητή με ονομαστικές τιμές.
Model	Το μοντέλο του οχήματος (π.χ., Model S, Leaf). Κατηγορική μεταβλητή με ονομαστικές τιμές.
Electric Vehicle Type	Ο τύπος του ηλεκτρικού οχήματος (π.χ., BEV, PHEV). Κατηγορική μεταβλητή με ονομαστικές τιμές.
Clean Alternative Fuel Vehicle (CAFV) Eligibility	Δείχνει εάν το όχημα πληροί τα κριτήρια για εναλλακτικά καύσιμα. Δυαδική κατηγορική μεταβλητή.
Electric Utility	Ο πάροχος ηλεκτρικής ενέργειας για το όχημα. Κατηγορική μεταβλητή με ονομαστικές τιμές.

Ποσοτικές Μεταβλητές:	
Model Year	Το έτος κατασκευής του οχήματος. Ποσοτική μεταβλητή (διακριτή)
Electric Range	Η εμβέλεια του οχήματος με ηλεκτρική ενέργεια. Ποσοτική μεταβλητή (συνεχής).
Base MSRP	Η τιμή καταλόγου του οχήματος. Ποσοτική μεταβλητή (συνεχής).

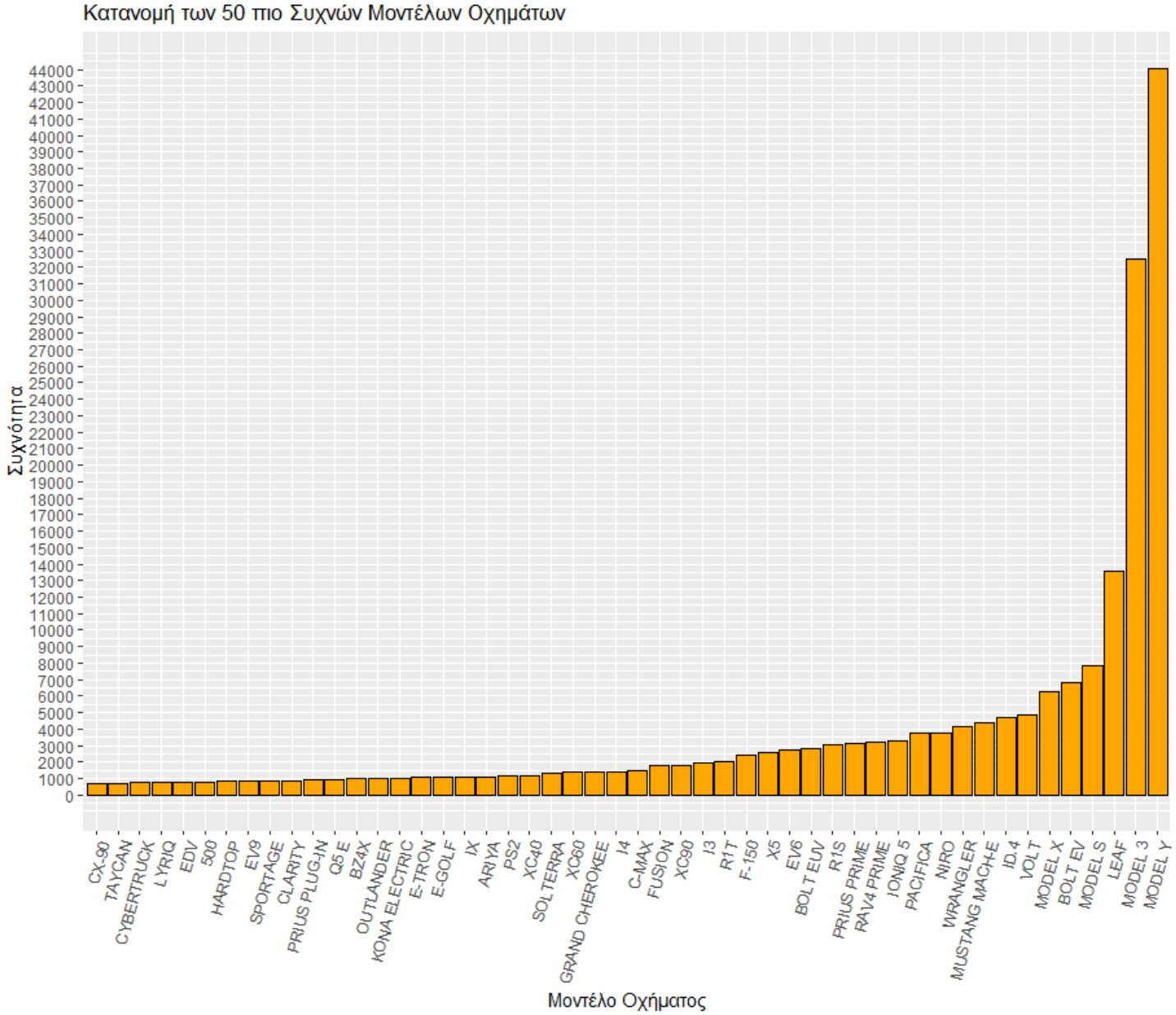
c. Δώστε τις κατανομές των μεταβλητών σε γραφική μορφή. Σχολιάστε τη μορφή των κατανομών, πιθανούς λόγους που έχουν αυτή τη μορφή, την ύπαρξη ατυπικών σημείων (outliers) κτλ.



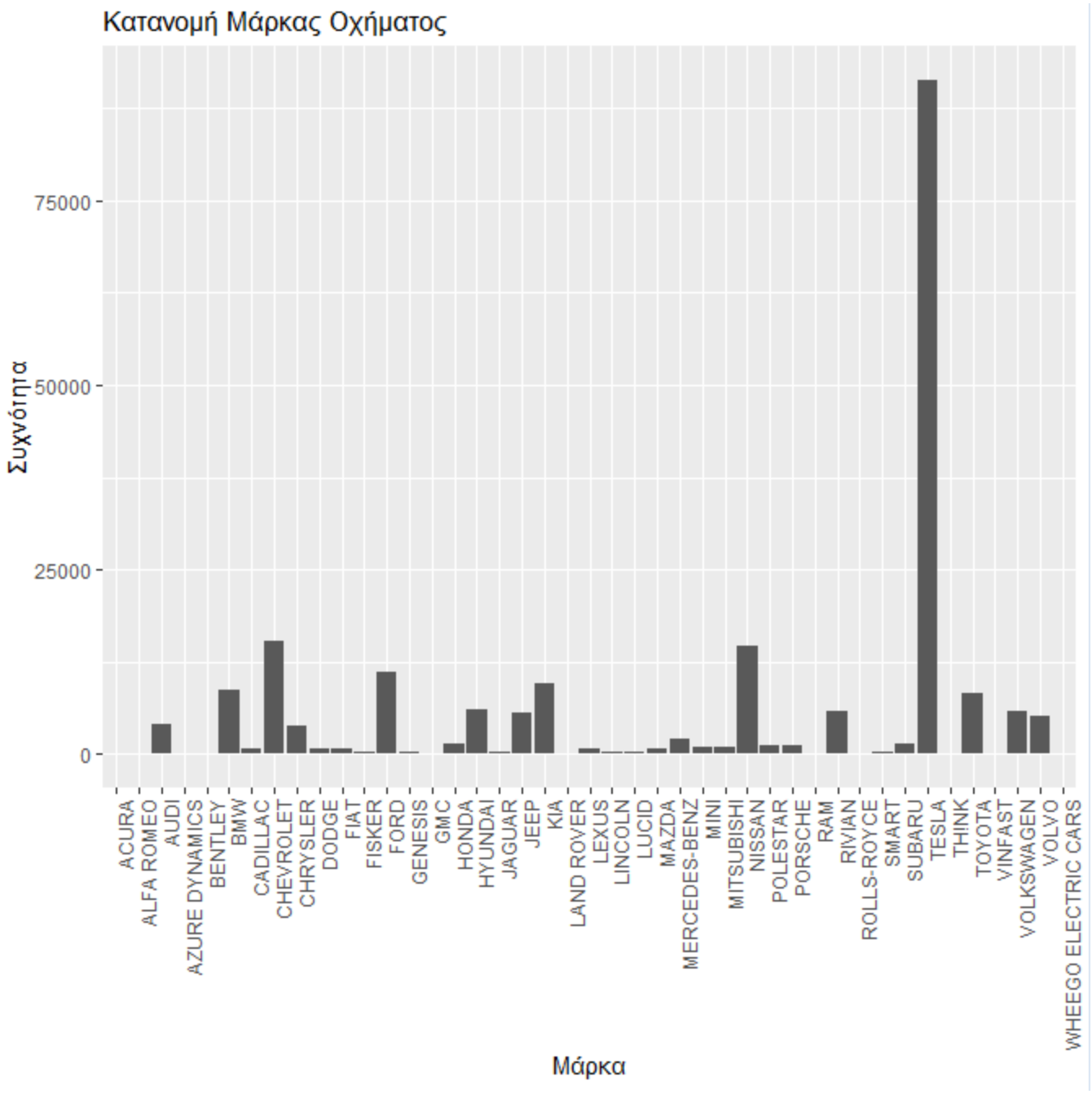
Η κατανομή των εγγραφών ανά κομητεία είναι έντονα **ασύμμετρη προς τα δεξιά**, με τις περισσότερες κομητείες να έχουν χαμηλή συχνότητα εγγραφών, ενώ λίγες, όπως οι **King** και **Snohomish**, ξεχωρίζουν με πολύ υψηλές τιμές. Αυτό υποδηλώνει ότι οι εγγραφές ηλεκτρικών οχημάτων είναι συγκεντρωμένες κυρίως σε αστικά κέντρα με μεγάλο πληθυσμό, ενώ οι μικρότερες, αγροτικές κομητείες παρουσιάζουν χαμηλότερη υιοθέτηση. Αν και δεν υπάρχουν κλασικά outliers, οι υψηλές τιμές για συγκεκριμένες κομητείες επηρεάζουν σημαντικά τη μορφή της κατανομής.



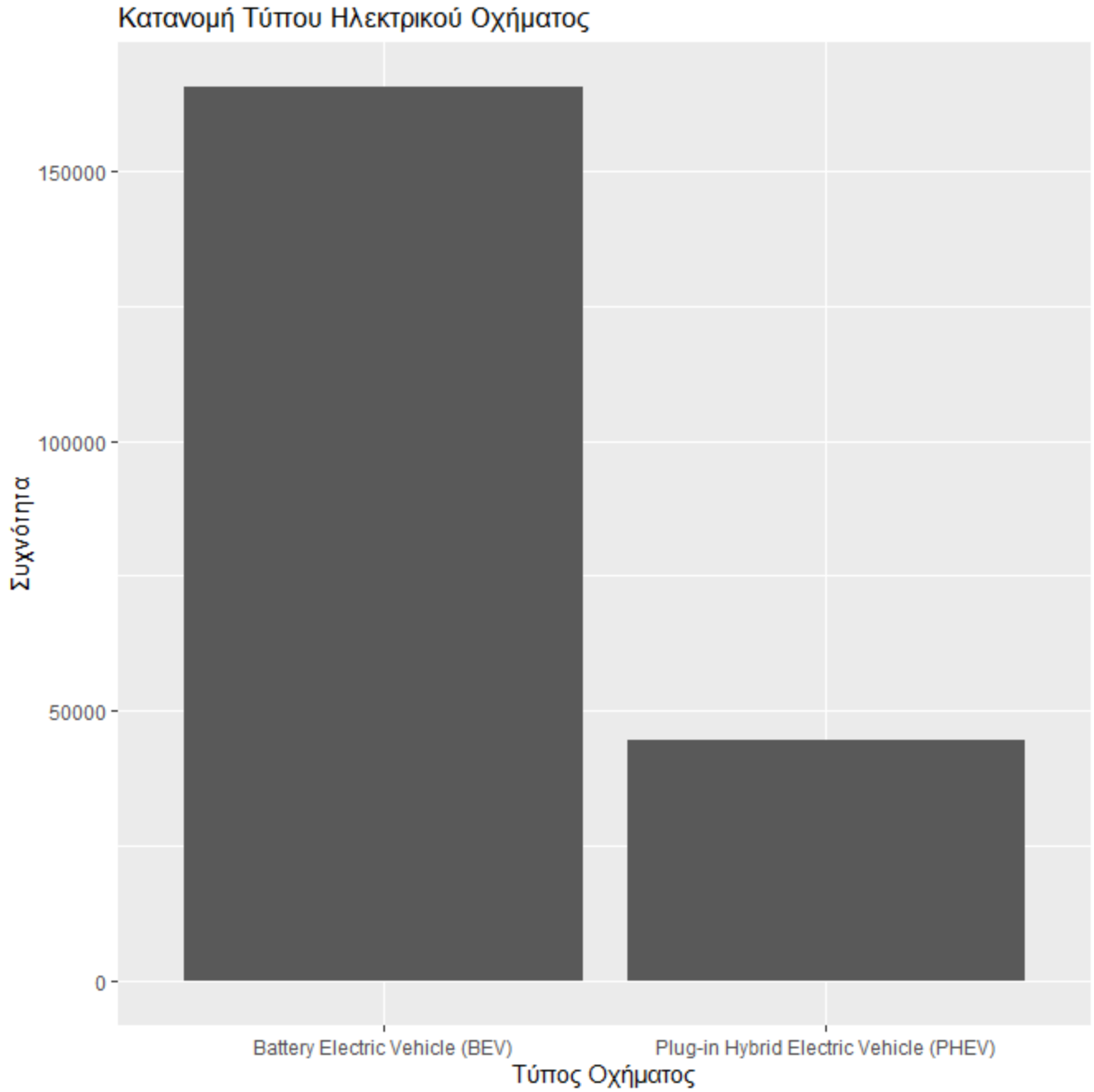
Η κατανομή των 50 πιο συχνών πόλεων είναι επίσης **ασύμμετρη προς τα δεξιά**, με τη συντριπτική πλειοψηφία των πόλεων να έχουν χαμηλές συχνότητες εγγραφών, ενώ η πόλη του **Seattle** κυριαρχεί με σημαντικά υψηλότερο αριθμό. Αυτό υποδηλώνει ότι η υιοθέτηση ηλεκτρικών οχημάτων συγκεντρώνεται σε μεγάλες αστικές περιοχές, όπως το Seattle, που είναι το μεγαλύτερο αστικό κέντρο στην περιοχή, με πιθανή ύπαρξη καλύτερης υποδομής και υψηλότερης περιβαλλοντικής συνείδησης. Δεν παρατηρούνται κλασικά outliers, αλλά η τεράστια διαφορά μεταξύ Seattle και των υπόλοιπων πόλεων καθιστά τη διανομή ασύμμετρη και υποδεικνύει σημαντική συγκέντρωση εγγραφών σε συγκεκριμένα αστικά κέντρα.



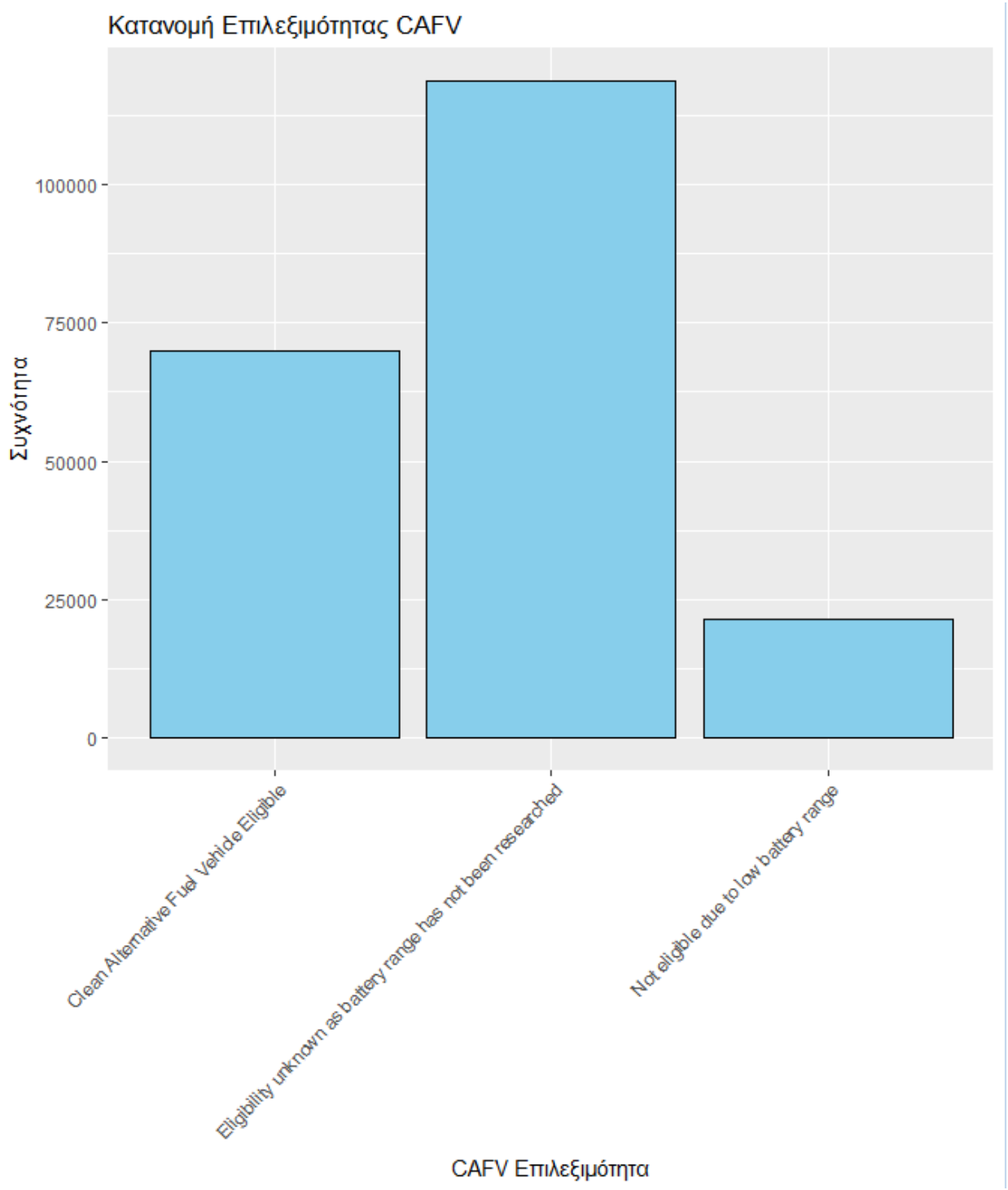
Η κατανομή των 50 πιο συχνών μοντέλων οχημάτων είναι έντονα **ασύμμετρη προς τα δεξιά**, με λίγα μοντέλα να έχουν εξαιρετικά υψηλή συχνότητα. Τα μοντέλα **Model Y** και **Model 3** της Tesla ξεχωρίζουν με διαφορά, υποδηλώνοντας την κυριαρχία της Tesla στην αγορά ηλεκτρικών οχημάτων. Τα περισσότερα υπόλοιπα μοντέλα έχουν σχετικά χαμηλή συχνότητα, κάτι που δείχνει ότι η αγορά είναι συγκεντρωμένη σε λίγα δημοφιλή μοντέλα. Δεν παρατηρούνται κλασικά outliers, αλλά η τεράστια διαφορά στη δημοτικότητα μεταξύ των κορυφαίων μοντέλων και των υπόλοιπων κάνει τη διανομή έντονα ασύμμετρη. Αυτή η τάση πιθανώς αντικατοπτρίζει την προτίμηση των καταναλωτών για τα πιο δημοφιλή και επιτυχημένα μοντέλα της Tesla



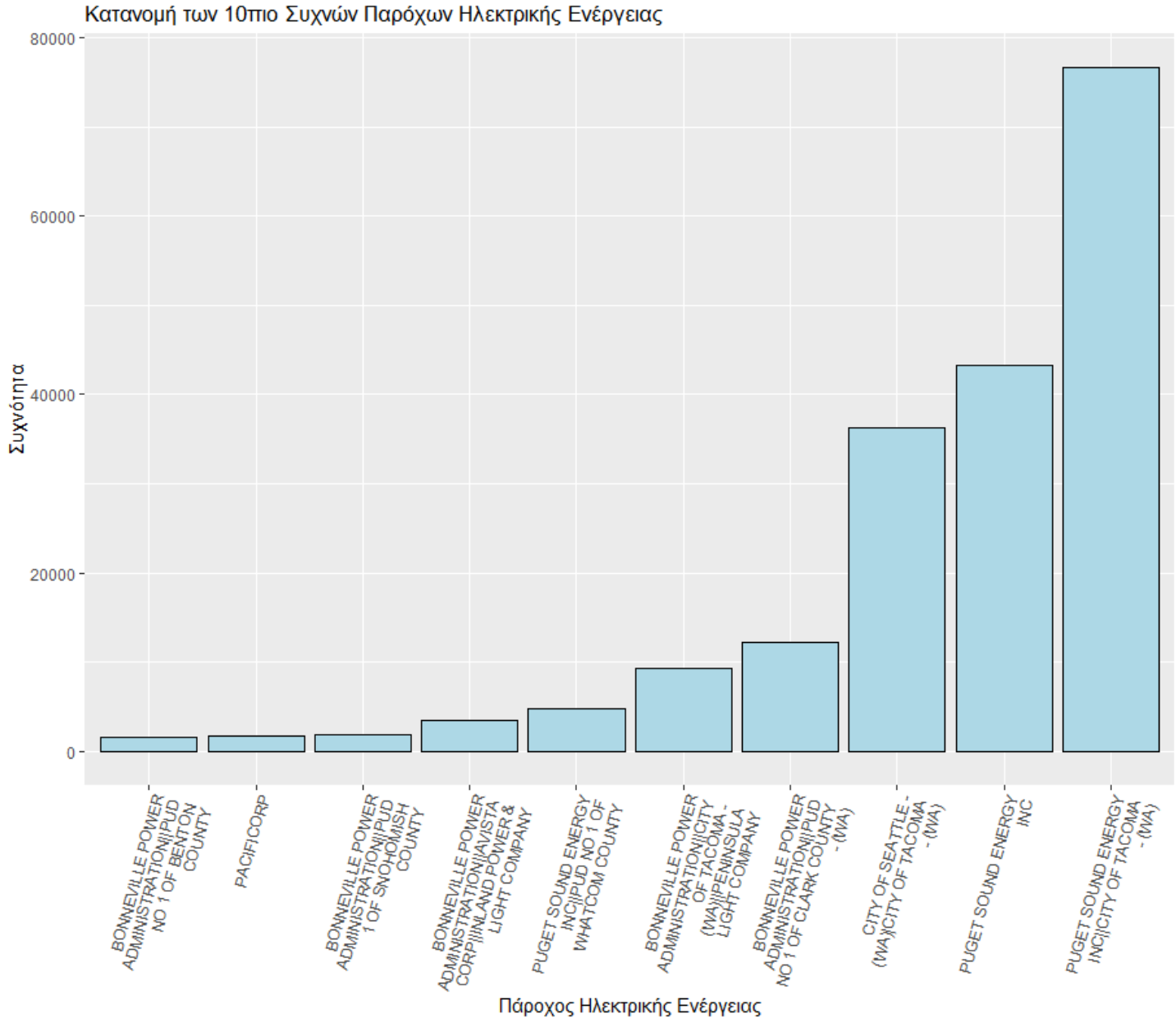
Η κατανομή της μάρκας οχήματος είναι έντονα **ασύμμετρη**, με την Tesla να κυριαρχεί σαφώς σε σχέση με όλες τις άλλες μάρκες. Η διαφορά αυτή είναι ιδιαίτερα εμφανής, καθώς η συχνότητα των οχημάτων της Tesla είναι πολύ υψηλότερη από οποιαδήποτε άλλη μάρκα, γεγονός που υποδεικνύει την ηγετική θέση της στην αγορά ηλεκτρικών οχημάτων. Οι υπόλοιπες μάρκες έχουν σχετικά χαμηλές και ομοιόμορφες συχνότητες, χωρίς να παρατηρούνται ιδιαίτερα outliers, εκτός από την εξαιρετικά υψηλή συχνότητα της Tesla. Η ασυμμετρία αυτή αντικατοπτρίζει την προτίμηση των καταναλωτών για τα οχήματα της Tesla και τη μεγάλη δημοτικότητά της στον τομέα των ηλεκτρικών οχημάτων.



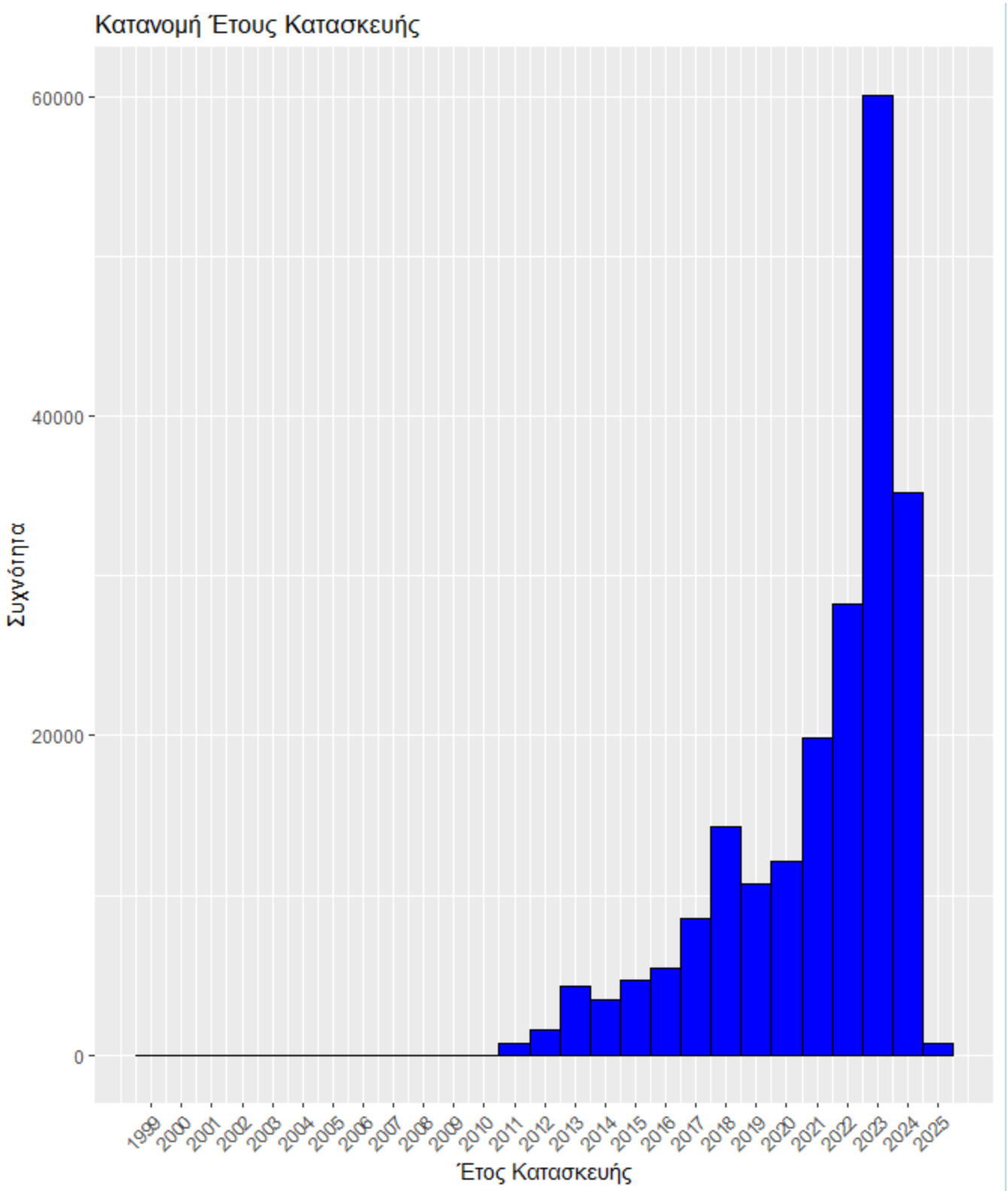
Η κατανομή των τύπων ηλεκτρικών οχημάτων δείχνει σαφή υπεροχή των **Battery Electric Vehicles (BEV)** σε σχέση με τα **Plug-in Hybrid Electric Vehicles (PHEV)**. Τα BEVs έχουν σχεδόν τριπλάσιο αριθμό εγγραφών σε σύγκριση με τα PHEVs, γεγονός που υποδηλώνει την αυξημένη προτίμηση των καταναλωτών για πλήρως ηλεκτρικά οχήματα. Η τάση αυτή μπορεί να οφείλεται στη μεγαλύτερη αυτονομία και την αυξημένη αποδοτικότητα των BEVs, καθώς και στις φορολογικές ελαφρύνσεις και κίνητρα που προσφέρονται συχνότερα για τα πλήρως ηλεκτρικά οχήματα. Δεν υπάρχουν outliers σε αυτήν την κατανομή, αλλά η σαφής διαφορά μεταξύ των δύο τύπων οχημάτων καταδεικνύει τη σημαντική προτίμηση προς τα BEVs στην αγορά.



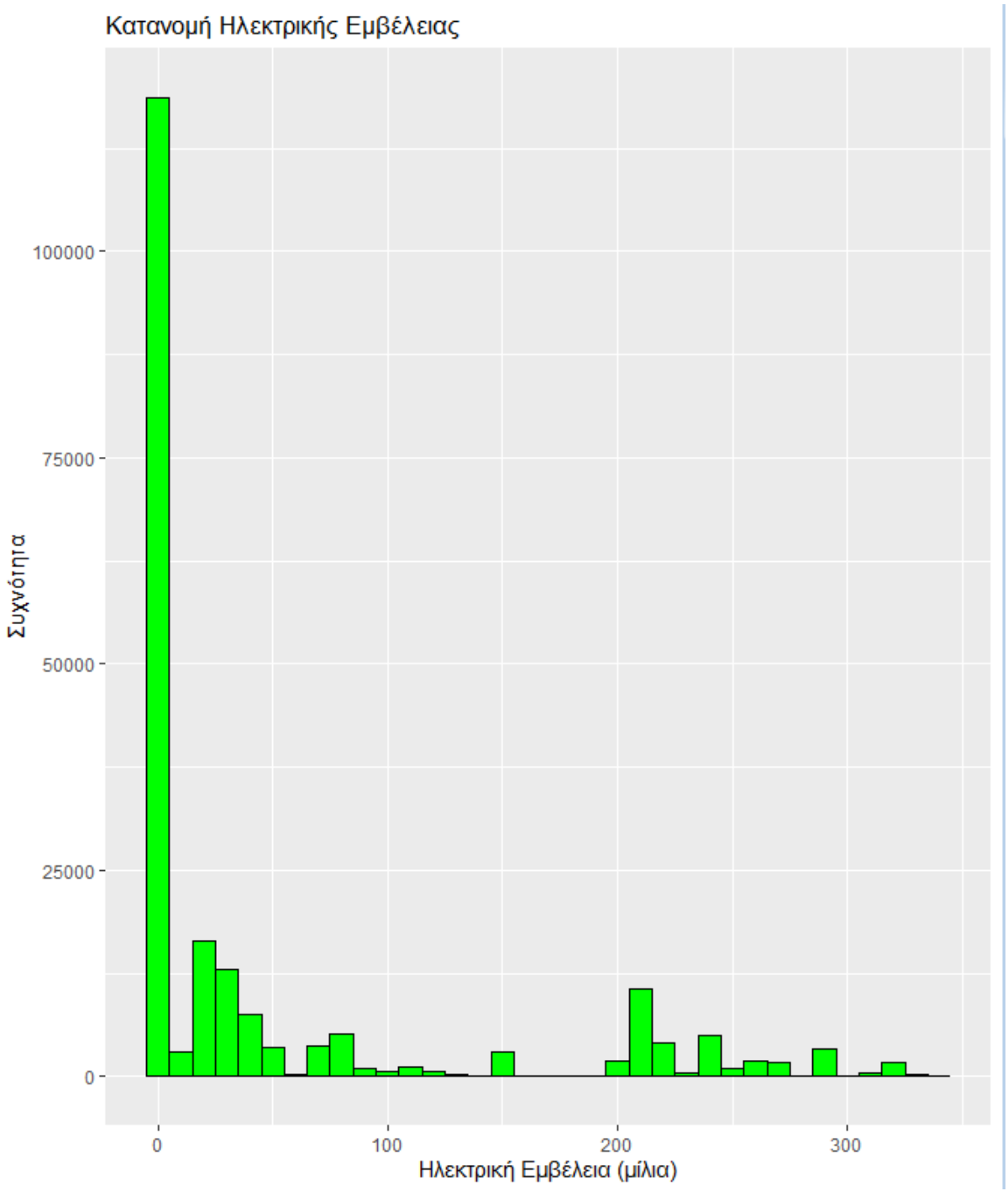
Η κατανομή της επιλεξιμότητας για το **Clean Alternative Fuel Vehicle (CAFV)** δείχνει τρεις κατηγορίες: τα οχήματα που είναι επιλέξιμα, τα οχήματα με άγνωστη επιλεξιμότητα (λόγω μη καταγεγραμμένης αυτονομίας) και τα οχήματα που δεν είναι επιλέξιμα λόγω χαμηλής αυτονομίας. Παρατηρούμε ότι η μεγαλύτερη ομάδα αφορά οχήματα με άγνωστη επιλεξιμότητα, πιθανώς λόγω έλλειψης πληροφοριών για την αυτονομία τους. Ακολουθούν τα επιλέξιμα οχήματα, τα οποία δείχνουν αυξημένη υιοθέτηση εναλλακτικών καυσίμων. Η μικρότερη ομάδα είναι τα μη επιλέξιμα οχήματα, κάτι που μπορεί να οφείλεται στη χαμηλή αυτονομία τους, καθιστώντας τα μη κατάλληλα για προγράμματα επιδότησης. Η κατανομή δεν παρουσιάζει εμφανή outliers, αλλά η διαφορά μεταξύ των κατηγοριών υποδεικνύει πιθανή έλλειψη επαρκών δεδομένων για την αυτονομία ορισμένων οχημάτων.



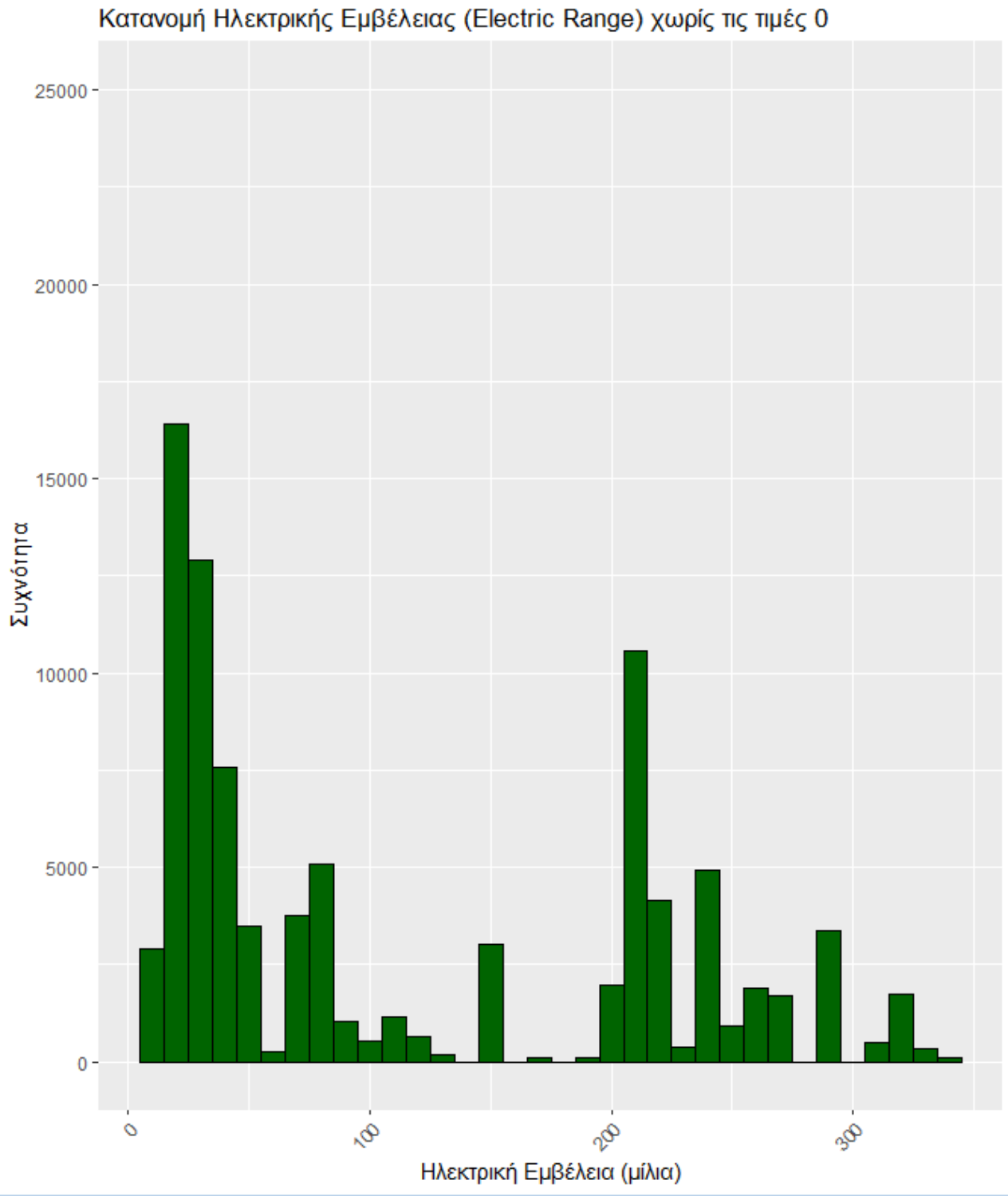
Η κατανομή των 10 πιο συχνών παρόχων ηλεκτρικής ενέργειας δείχνει μεγάλη διαφοροποίηση στη συχνότητα, με τον **Puget Sound Energy, Inc** να κυριαρχεί ως ο πιο κοινός πάροχος, ακολουθούμενος από την **City of Seattle**. Η σαφής υπεροχή αυτών των δύο παρόχων πιθανώς οφείλεται στη μεγάλη κάλυψη που προσφέρουν σε αστικές περιοχές με υψηλή συγκέντρωση ηλεκτρικών οχημάτων. Οι υπόλοιποι πάροχοι έχουν σημαντικά χαμηλότερες συχνότητες, κάτι που υποδηλώνει ότι εξυπηρετούν μικρότερες ή πιο εξειδικευμένες αγορές. Η κατανομή είναι **ασύμμετρη προς τα δεξιά**, χωρίς εμφανή outliers, αλλά η μεγάλη διαφορά μεταξύ των κορυφαίων παρόχων και των υπόλοιπων αντικατοπτρίζει την κυριαρχία λίγων μεγάλων παρόχων στην παροχή ηλεκτρικής ενέργειας για τα ηλεκτρικά οχήματα.



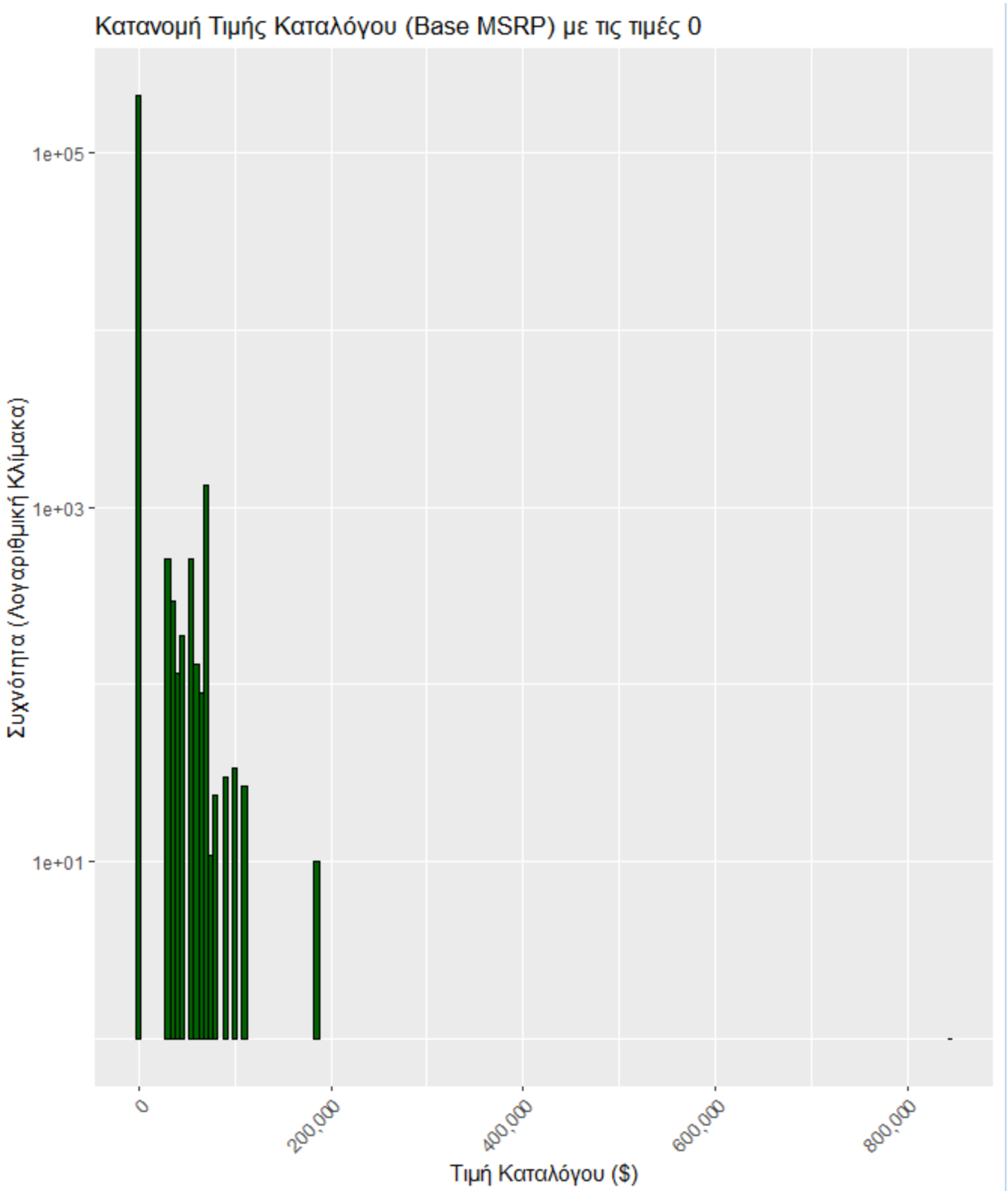
Η κατανομή του έτους κατασκευής είναι έντονα **ασύμμετρη προς τα δεξιά**, με τον αριθμό των εγγραφών να αυξάνεται σημαντικά μετά το 2018. Παρατηρούμε μια απότομη αύξηση τα τελευταία χρόνια, με τις κορυφαίες εγγραφές να προέρχονται από τα έτη 2021, 2022, και 2023. Αυτό δείχνει την αυξανόμενη υιοθέτηση ηλεκτρικών οχημάτων, καθώς η τεχνολογία βελτιώνεται και οι καταναλωτές στρέφονται περισσότερο προς πιο πρόσφατα μοντέλα. Δεν παρατηρούνται εμφανή outliers, αλλά η συγκέντρωση των εγγραφών στα τελευταία χρόνια αντικατοπτρίζει την τάση της αγοράς και τη στροφή προς τα νεότερα, πιο αποδοτικά ηλεκτρικά οχήματα.



Η κατανομή της ηλεκτρικής εμβέλειας είναι έντονα **ασύμμετρη προς τα αριστερά**, με μια μεγάλη κορυφή στις τιμές κοντά στο 0. Αυτό δείχνει ότι υπάρχουν πολλές εγγραφές με μηδενική ή πολύ χαμηλή αυτονομία, πιθανώς επειδή περιλαμβάνονται οχήματα χωρίς καταγεγραμμένη εμβέλεια ή παλαιότερα μοντέλα με περιορισμένη απόδοση. Εκτός από αυτή την αιχμή, η κατανομή δείχνει πολλαπλές μικρότερες κορυφές, ειδικά γύρω από τις 100 και 200 μίλια, που αντιπροσωπεύουν πιο σύγχρονα ηλεκτρικά οχήματα με μεγαλύτερη αυτονομία. Υπάρχουν ακραίες τιμές (outliers) με εμβέλεια πάνω από 300 μίλια, που πιθανότατα αντιστοιχούν σε οχήματα υψηλής απόδοσης. Συνολικά, η κατανομή αντανακλά την ποικιλία στην τεχνολογία των ηλεκτρικών οχημάτων και την εξέλιξη της αυτονομίας τους.

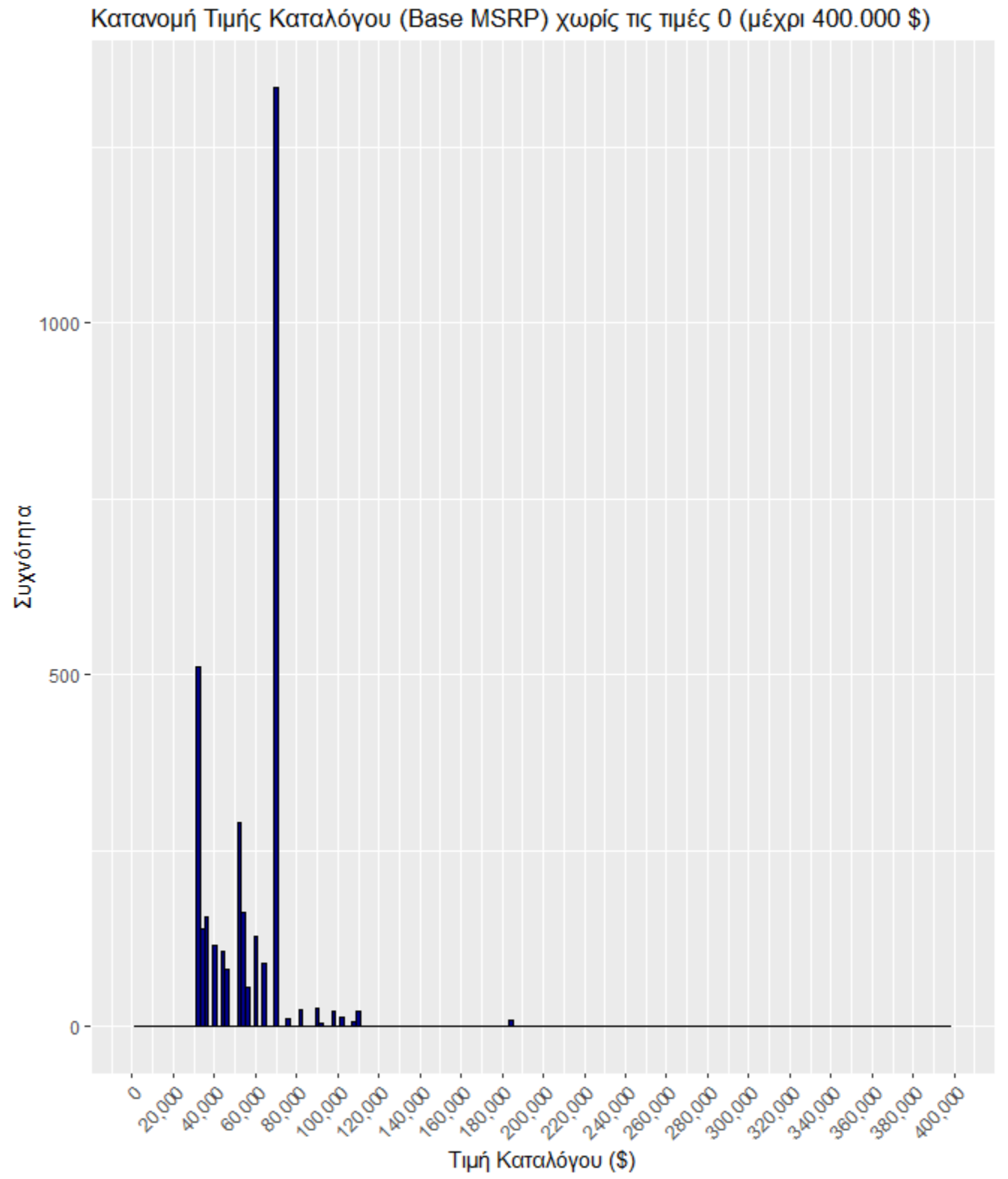


Η κατανομή της ηλεκτρικής εμβέλειας, αφού αφαιρέθηκαν οι τιμές 0, δείχνει μια πιο καθαρή εικόνα των πραγματικών δυνατοτήτων των ηλεκτρικών οχημάτων. Η κατανομή είναι **πολυτροπική**, με τρεις κύριες κορυφές: γύρω από τις 50 μίλια, 100 μίλια και 200 μίλια. Η κορυφή στα 50 μίλια πιθανόν αντιστοιχεί σε plug-in υβριδικά οχήματα με μικρότερη αυτονομία, ενώ η κορυφή στα 200 μίλια υποδεικνύει πλήρως ηλεκτρικά οχήματα με μεγαλύτερη αυτονομία. Παρατηρούμε επίσης μια μικρή κορυφή μετά τις 300 μίλια, η οποία αντιστοιχεί σε οχήματα υψηλής απόδοσης. Η κατανομή δεν δείχνει ακραίες τιμές (outliers), αλλά η ύπαρξη πολλαπλών κορυφών υποδηλώνει την ποικιλία στην τεχνολογία και τα χαρακτηριστικά των οχημάτων. Η διαφορά στις κορυφές μπορεί να αντικατοπτρίζει τις διαφορετικές κατηγορίες οχημάτων (PHEV vs. BEV) και τις εξελίξεις στην τεχνολογία μπαταριών.



Η κατανομή της τιμής καταλόγου (Base MSRP) με τις τιμές 0 περιλαμβανόμενες είναι **έντονα ασύμμετρη προς τα αριστερά**, όπως φαίνεται και από τη λογαριθμική κλίμακα στον άξονα y. Υπάρχει μια πολύ μεγάλη συγκέντρωση οχημάτων με τιμή καταλόγου κοντά στο 0, που πιθανότατα αντιπροσωπεύει εγγραφές χωρίς τιμή ή με μη καταγεγραμμένη τιμή MSRP. Μετά τη μεγάλη αιχμή κοντά στο 0, παρατηρούμε μερικές κατανομές σε υψηλότερες τιμές, με τη συντριπτική πλειοψηφία να κυμαίνεται κάτω από τις 100.000 δολάρια. Υπάρχουν επίσης μερικές ακραίες τιμές (outliers) που φτάνουν ή και ξεπερνούν τις 800.000 δολάρια, πιθανότατα για οχήματα πολυτελείας ή υψηλής απόδοσης. Η χρήση λογαριθμικής κλίμακας βοηθά να δούμε καλύτερα την

κατανομή στις χαμηλότερες τιμές και να διαχειριστούμε την έντονη ασυμμετρία λόγω των εξαιρετικά υψηλών τιμών.



Η κατανομή της τιμής καταλόγου (Base MSRP) χωρίς τις τιμές 0 και περιορισμένη μέχρι τα 400.000 δολάρια είναι έντονα **ασύμμετρη προς τα δεξιά**, με τη μεγάλη πλειοψηφία των τιμών να συγκεντρώνονται κάτω από τις 100.000 δολάρια. Παρατηρούμε μια απότομη αιχμή κοντά στις 50.000 δολάρια, που πιθανότατα αντιπροσωπεύει την τυπική τιμή για τα πιο δημοφιλή ηλεκτρικά οχήματα. Μετά από αυτό το σημείο, η συχνότητα μειώνεται απότομα, ενώ υπάρχουν μερικές σποραδικές τιμές μέχρι και τα 400.000 δολάρια, που

πιθανότατα αντιστοιχούν σε οχήματα υψηλότερης κατηγορίας ή πολυτελείας. Δεν παρατηρούνται εμφανή outliers εντός αυτού του εύρους, αλλά η ασυμμετρία καταδεικνύει την επικράτηση των οχημάτων μεσαίας κατηγορίας τιμής στην αγορά.

d. Για κάθε ποσοτική μεταβλητή, υπολογίστε

α) τη μέση τιμή και τυπική απόκλιση, και

β) τη σύνοψη των πέντε αριθμών. Σχολιάστε την καταλληλότητα των α), β) για κάθε μεταβλητή.

Model Year:
mean = 2021.049
sd = 2.988941

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1999    2019    2022    2021    2023    2025
```

Electric Range:
mean = 50.60224
sd = 86.97321

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    0.0     0.0     0.0    50.6    42.0   337.0       5
```

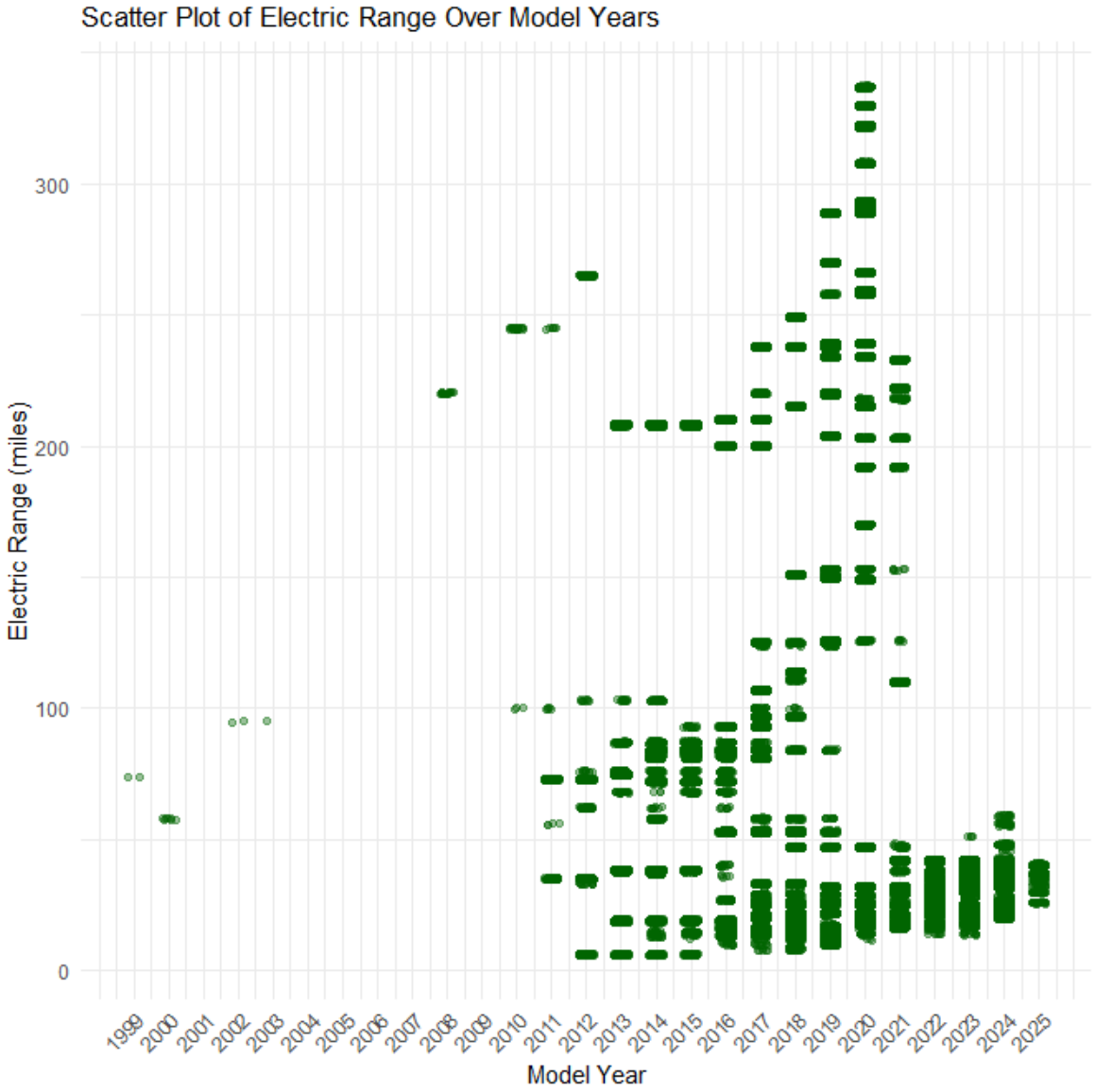
Base MSRP:
mean = 897.6769
sd = 7653.589

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
     0.0      0.0      0.0    897.7      0.0 845000.0        5
```

e. Επιλέξτε δύο μεταβλητές και διερευνήστε τη σχέση τους. Εάν θεωρήσετε ότι υπάρχει σχέση, αυτή είναι αιτιατή ή όχι; Σχολιάστε αναλόγως.

Ιδέα:

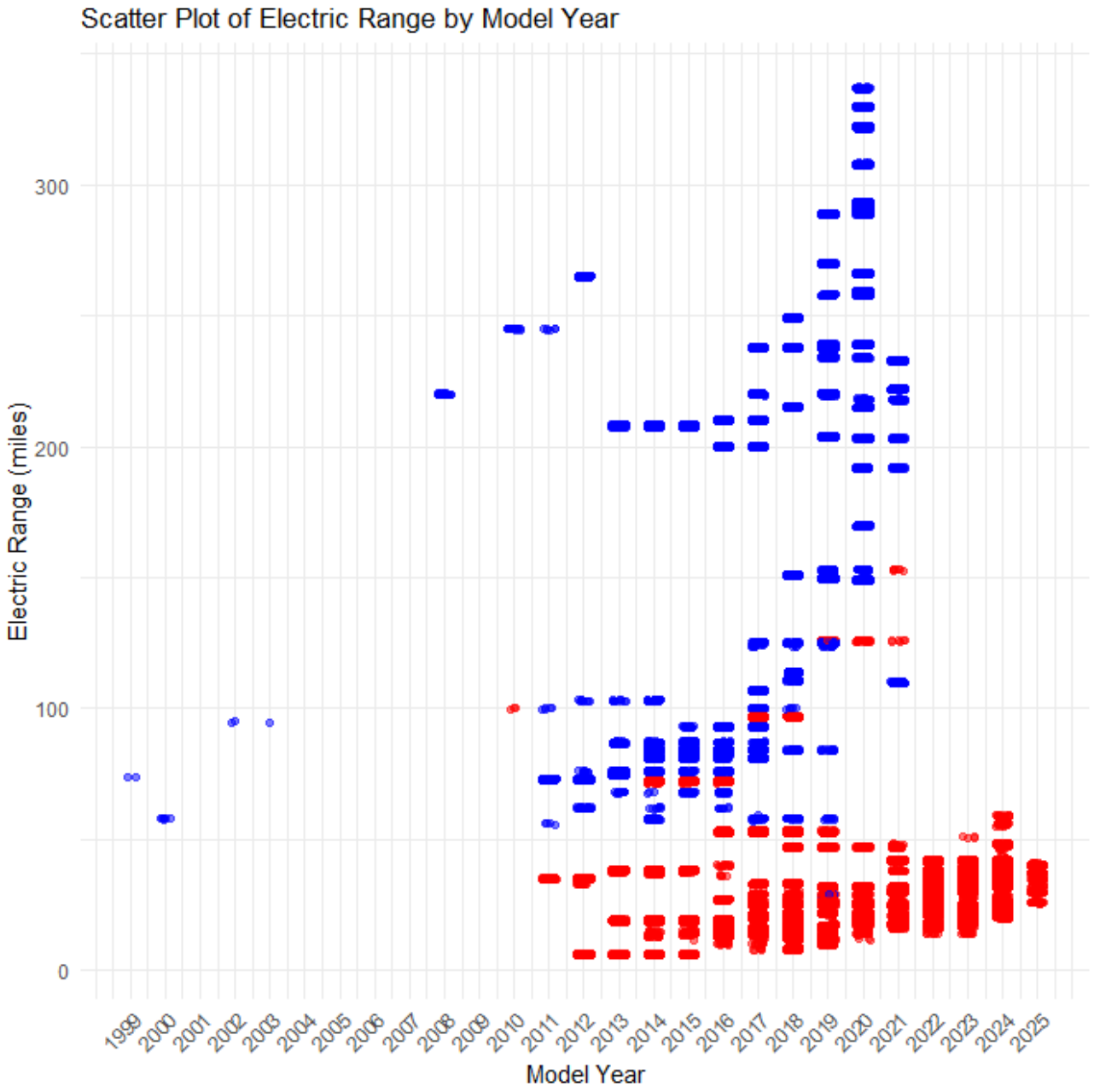
- **Το γράφημα δείχνει την ηλεκτρική εμβέλεια των οχημάτων ανά έτος κατασκευής, χωρίς διάκριση μεταξύ BEVs και PHEVs.**
- **Περιμένουμε τα νεότερα μοντέλα να έχουν μεγαλύτερη εμβέλεια λόγω της βελτίωσης της τεχνολογίας μπαταριών.**



Παρατηρήσεις:

- **Παρατηρούμε πολλά σημεία με χαμηλή εμβέλεια στα πιο πρόσφατα έτη κατασκευής, κάτι που δεν είναι αναμενόμενο.**
- **Πιθανή εξήγηση: Αυτά τα σημεία πιθανόν αντιστοιχούν σε PHEVs, τα οποία έχουν μικρότερη αυτονομία λόγω της μικρότερης μπαταρίας τους.**

- **Στο επόμενο βήμα, θα διαχωρίσουμε τα BEVs από τα PHEVs για να δούμε εάν αυτή η υπόθεση ισχύει.**

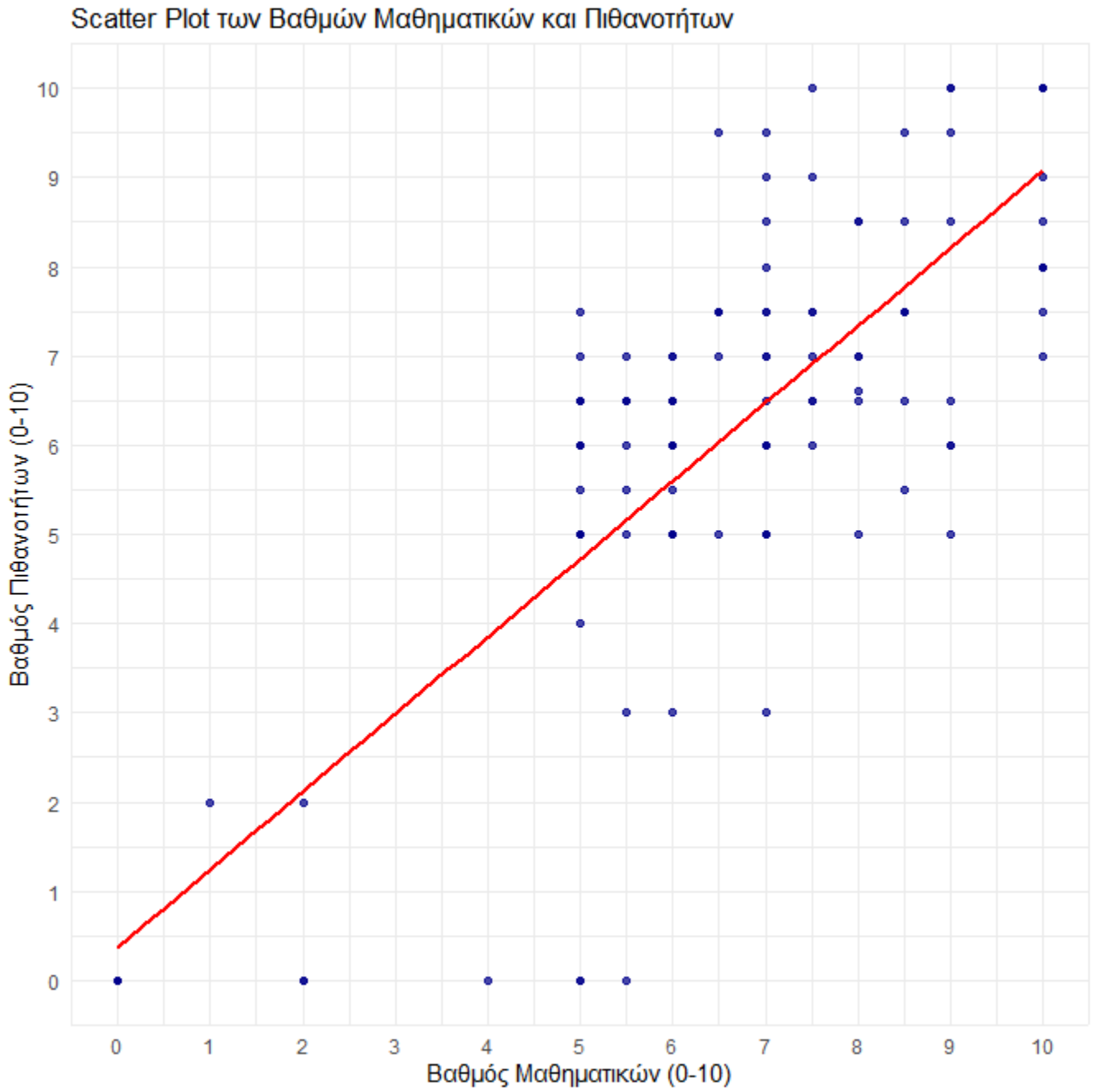


Παρατηρήσεις:

- **Το διάγραμμα δείχνει δύο τάσεις στην ηλεκτρική εμβέλεια:**
  - **Μπλε σημεία: BEVs (Battery Electric Vehicles), με αυξανόμενη εμβέλεια που φτάνει πάνω από 300 μίλια στα πιο πρόσφατα έτη.**
  - **Κόκκινα σημεία: PHEVs (Plug-in Hybrid Electric Vehicles), με σταθερά χαμηλότερη εμβέλεια, κάτω από 100 μίλια.**
- **Στα πιο πρόσφατα έτη, παρατηρείται αυξημένη δημοτικότητα των PHEVs με μικρή εμβέλεια, κάτι που επηρεάζει τη συνολική πτώση στην κατανομή της εμβέλειας. Αυτό πιθανόν οφείλεται στη ζήτηση για οικονομικά, επαναφορτιζόμενα οχήματα κατάλληλα για αστική χρήση.**

3.

a. Δώστε το scatterplot και σχολιάστε τη μορφή, κατεύθυνση και δύναμη της σχέσης των δύο μεταβλητών.
b.Υπολογίστε τον συντελεστή συσχέτισης και εκτελέστε γραμμική παλινδρόμηση ελαχίστων τετραγώνων.



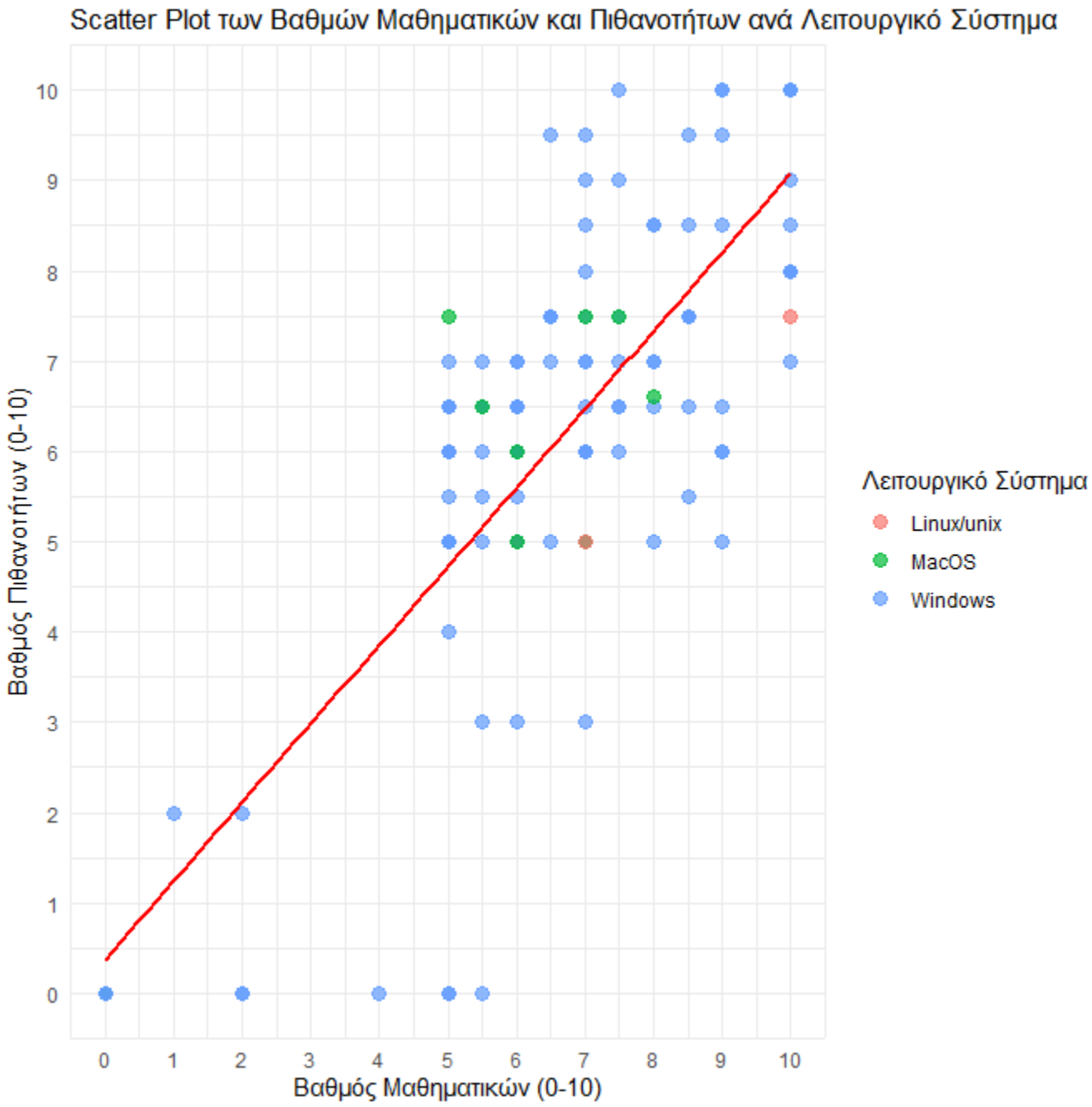
Επεξήγηση Διαγράμματος

- Το διάγραμμα δείχνει τη σχέση μεταξύ των βαθμών στα Μαθηματικά (x-άξονας) και των βαθμών στις Πιθανότητες (y-άξονας). Κάθε μπλε κουκκίδα αντιπροσωπεύει έναν μαθητή και τη βαθμολογία του στα δύο μαθήματα.
- Η κόκκινη γραμμή είναι η **γραμμή παλινδρόμησης**, που μας δείχνει την τάση των δεδομένων. Υπολογίστηκε χρησιμοποιώντας τη συνάρτηση **geom_smooth()**, που προσαρμόζει μια ευθεία που αντιπροσωπεύει την καλύτερη δυνατή γραμμική σχέση μεταξύ των δύο μεταβλητών.

Παρατηρήσεις

- Βλέπουμε ότι υπάρχει θετική συσχέτιση: όσο καλύτερος είναι ο βαθμός στα Μαθηματικά, τόσο καλύτερος είναι και στις Πιθανότητες. Η γραμμή τείνει προς τα πάνω, κάτι που το δείχνει ξεκάθαρα.
- Υπάρχουν όμως και μερικά σημεία που ξεφεύγουν από την τάση, δηλαδή μαθητές που πήγαν καλά στα Μαθηματικά αλλά όχι στις Πιθανότητες (ή το αντίστροφο). Αυτά είναι τα **ατυπικά σημεία** (outliers).

Γενικά, φαίνεται ότι οι μαθητές που έχουν καλή κατανόηση στα Μαθηματικά τείνουν να αποδίδουν καλά και στις Πιθανότητες, κάτι που ήταν αναμενόμενο.Με λίγα λόγια, το διάγραμμα δείχνει ότι οι δύο βαθμολογίες σχετίζονται αρκετά και η γραμμή παλινδρόμησης βοηθά στο να καταλάβουμε τη γενική κατεύθυνση αυτή της σχέσης.



Από το διάγραμμα παρατηρούμε ότι οι μαθητές που χρησιμοποιούν **Windows** φαίνονται να έχουν γενικά καλύτερους βαθμούς στις Πιθανότητες σε σύγκριση με τους χρήστες **MacOS** και **Linux/Unix**, οι οποίοι φαίνεται να έχουν πιο συγκεντρωμένα αποτελέσματα γύρω από τον μέσο όρο.