Big Data Analytics - Ανάλυση δεδομένων ηλεκτρικών οχημάτων

Τίτλος Εργασίας: Ανάλυση Μεγάλων Δεδομένων με Python

Ονόματα Ομάδας:

- Διακογιάννης Αλέξιος Διονύσιος,

- Μανωλαράκης Αντώνιος,

- Τσολακίδης Κωνσταντίνος

Εισαγωγή

Η ηλεκτροκίνηση αποτελεί μια από τις σημαντικότερες τεχνολογικές εξελίξεις στον τομέα των μεταφορών, καθώς συμβάλλει στη μείωση των εκπομπών διοξειδίου του άνθρακα και στη μετάβαση σε πιο βιώσιμες μορφές ενέργειας. Η ανάλυση δεδομένων ηλεκτρικών οχημάτων μπορεί να παρέχει πολύτιμες πληροφορίες για την κατανόηση των τάσεων αγοράς, της γεωγραφικής κατανομής και της χρήσης τους.

Στην παρούσα εργασία, θα αξιοποιήσουμε δεδομένα που περιέχουν πληροφορίες σχετικά με τον πληθυσμό ηλεκτρικών οχημάτων και θα εφαρμόσουμε τεχνικές ανάλυσης δεδομένων με τη χρήση **PySpark**. Η εργασία στοχεύει να απαντήσει σε ερωτήματα όπως:

- Ποιες είναι οι κυρίαρχες μάρκες και μοντέλα ηλεκτρικών οχημάτων;
- Πώς κατανέμονται γεωγραφικά τα ηλεκτρικά οχήματα;
- Ποια είναι τα πιο συνηθισμένα χαρακτηριστικά των ηλεκτρικών οχημάτων;
- Πώς μπορούν τα δεδομένα αυτά να χρησιμοποιηθούν για τη χάραξη πολιτικής και τη βελτίωση των υποδομών ηλεκτροκίνησης;

Η προσέγγισή μας περιλαμβάνει την **καθαριότητα και προετοιμασία των δεδομένων, την εφαρμογή ανάλυσης μέσω PySpark SQL και την εξαγωγή συμπερασμάτων**.

Πίνακας Περιεχομένων

2. Βιβλιογραφική Ανασκόπηση

2.1 Ηλεκτρικά Οχήματα και Οικολογική Μετάβαση

Τα ηλεκτρικά οχήματα (EVs) αποτελούν κεντρικό άξονα της παγκόσμιας μετάβασης προς καθαρότερες μορφές ενέργειας. Σύμφωνα με μελέτες (IEA, 2023), η αύξηση της χρήσης EVs έχει συμβάλλει στη μείωση των εκπομπών CO_2 και στη μείωση της εξάρτησης από ορυκτά καύσιμα. Ο αριθμός των ηλεκτρικών οχημάτων έχει αυξηθεί σημαντικά τα τελευταία χρόνια, με την Κίνα, τις ΗΠΑ και την Ευρώπη να ηγούνται της αγοράς.

Η ανάλυση δεδομένων ηλεκτρικών οχημάτων παρέχει πληροφορίες σχετικά με τις τάσεις αγοράς, τη χρήση και την κατανομή τους. Η πρόσβαση σε δεδομένα πραγματικού κόσμου επιτρέπει τη λήψη αποφάσεων από κυβερνήσεις και επιχειρήσεις για τη βελτίωση των υποδομών φόρτισης και την ανάπτυξη νέων πολιτικών.

2.2 Big Data και Ανάλυση Δεδομένων στην Ηλεκτροκίνηση

Η ανάλυση μεγάλων δεδομένων (**Big Data Analytics**) έχει βρει εφαρμογή σε διάφορους τομείς, συμπεριλαμβανομένης της αυτοκινητοβιομηχανίας. Τα δεδομένα που προέρχονται από ηλεκτρικά οχήματα (π.χ. αισθητήρες, GPS, σταθμοί φόρτισης) επιτρέπουν την ανακάλυψη προτύπων και τη βελτιστοποίηση των υπηρεσιών μετακίνησης.

Η χρήση τεχνολογιών όπως **Apache Spark και PySpark** επιτρέπει την αποτελεσματική ανάλυση μεγάλων συνόλων δεδομένων, ενώ τα εργαλεία μηχανικής μάθησης (ML) επιτρέπουν την πρόβλεψη τάσεων και τη λήψη αποφάσεων βασισμένων σε δεδομένα.

Μεθοδολογίες όπως:

- **Στατιστική ανάλυση** για την περιγραφή των δεδομένων.
- **Ανάλυση PySpark SQL** για την αναζήτηση μοτίβων.
- **Μηχανική μάθηση** για πιθανές προβλέψεις σχετικά με την ανάπτυξη της ηλεκτροκίνησης.

2.3 Προηγούμενες Έρευνες και Συναφή Εργαλεία

Προηγούμενες έρευνες έχουν επικεντρωθεί στην ανάλυση δεδομένων ηλεκτρικών οχημάτων, όπως:

- Μελέτες πάνω στη **γεωγραφική κατανομή των EVs** (Chen et al., 2022).

- Ανάλυση της **αποδοτικότητας μπαταριών και δικτύων φόρτισης** (Liu et al., 2021).
- **Αναλύσεις αγοράς** που προσδιορίζουν τις τάσεις κατανάλωσης και τις προτιμήσεις καταναλωτών.

Η χρήση εργαλείων όπως το **Apache Spark** έχει αποδειχθεί αποτελεσματική στην επεξεργασία μεγάλων δεδομένων. Το PySpark επιτρέπει **γρήγορη ανάλυση δεδομένων**, ενώ η ενσωμάτωση με **Pandas και Matplotlib** βοηθά στην οπτικοποίηση αποτελεσμάτων.

2. Θεωρητικό Πλαίσιο

Η ανάλυση μεγάλων δεδομένων (**Big Data Analytics**) έχει αναδειχθεί ως μια από τις πιο σημαντικές τεχνολογικές προόδους της τελευταίας δεκαετίας. Με την αυξανόμενη διαθεσιμότητα δεδομένων από διάφορους τομείς, οι σύγχρονες μεθοδολογίες ανάλυσης επιτρέπουν την εξαγωγή χρήσιμων πληροφοριών, την πρόβλεψη τάσεων και τη βελτιστοποίηση διαδικασιών.

2.1. Big Data Analytics: Ορισμός και Σημασία

Τα **Big Data Analytics** αφορούν τις τεχνικές και τις τεχνολογίες που χρησιμοποιούνται για την αποθήκευση, επεξεργασία και ανάλυση μεγάλου όγκου δεδομένων, τα οποία είναι συχνά μη δομημένα ή ημι-δομημένα. Χαρακτηρίζονται από τις **5Vs**:

- **Volume (Όγκος):** Τεράστιος όγκος δεδομένων που δημιουργείται καθημερινά.
- **Velocity (Ταχύτητα):** Τα δεδομένα παράγονται και επεξεργάζονται σε πραγματικό χρόνο.
- **Variety (Ποικιλία):** Διάφορες μορφές δεδομένων (δομημένα, μη δομημένα, εικόνες, IoT δεδομένα).
- **Veracity (Αξιοπιστία):** Η ανάγκη για καθαρά και αξιόπιστα δεδομένα.
- **Value (Αξία):** Η μετατροπή των δεδομένων σε χρήσιμες πληροφορίες για λήψη αποφάσεων.

2.2. Big Data στις Μεταφορές

Η υιοθέτηση των Big Data Analytics στις μεταφορές έχει φέρει επανάσταση στην ανάλυση δεδομένων κινητικότητας. Μέσω της συλλογής δεδομένων από αισθητήρες, GPS, εφαρμογές κινητών και έξυπνες υποδομές, είναι δυνατή η μελέτη προτύπων κυκλοφορίας, η βελτίωση των δημόσιων συγκοινωνιών και η ανάπτυξη στρατηγικών βιώσιμης κινητικότητας.

2.3. Ηλεκτρικά Οχήματα και Δεδομένα

Η ραγδαία αύξηση των ηλεκτρικών οχημάτων (**EVs**) έχει δημιουργήσει νέες προκλήσεις και ευκαιρίες στη διαχείριση και ανάλυση δεδομένων. Τα δεδομένα που σχετίζονται με τα EVs περιλαμβάνουν:

- Κατασκευαστές και μοντέλα οχημάτων

- Κατασκευαστές και μοντέλα οχημάτων.
- Χωρητικότητα μπαταρίας και αυτονομία.
- Σταθμούς φόρτισης και πρότυπα σύνδεσης.
- Περιβαλλοντικές επιπτώσεις και κατανάλωση ενέργειας.

Η ανάλυση αυτών των δεδομένων μπορεί να οδηγήσει σε **βελτιστοποίηση της χρήσης των ηλεκτρικών οχημάτων, πρόβλεψη αναγκών φόρτισης και εντοπισμό εμποδίων στην υιοθέτησή τους**.

2.4. Μεθοδολογίες και Τεχνολογίες που θα χρησιμοποιηθούν

Για την ανάλυση του dataset των ηλεκτρικών οχημάτων, θα εφαρμοστούν οι ακόλουθες τεχνικές:

- **Εξερεύνηση δεδομένων (Exploratory Data Analysis - EDA):** Οπτικοποίηση και ανάλυση χαρακτηριστικών του dataset.
- **Στατιστική Ανάλυση:** Κατανόηση των σχέσεων μεταξύ των μεταβλητών.
- **Μηχανική Μάθηση (Machine Learning):** Ανάπτυξη μοντέλων πρόβλεψης (αν κριθεί αναγκαίο).
- **Big Data Tools:** Χρήση βιβλιοθηκών όπως **Pandas, Seaborn, Matplotlib, Plotly**, καθώς και πιθανή χρήση **PySpark για μεγάλα datasets**.

3. Προεπεξεργασία των Δεδομένων

Πριν προχωρήσουμε στην ανάλυση των δεδομένων, είναι απαραίτητο να εξετάσουμε τη δομή του dataset, να εντοπίσουμε ελλείπουσες ή ανώμαλες τιμές και να εφαρμόσουμε τις κατάλληλες τεχνικές προεπεξεργασίας.

3.1. Επισκόπηση του Dataset

Ξεκινάμε φορτώνοντας το dataset και εξετάζοντας τη δομή του:

```
#Φορτώνουμε απο Google Drive

from google.colab import drive
drive.mount('/content/drive')
file_path = '/content/drive/My Drive/Colab Notebooks/Electric_Vehicle_Populatic

import pandas as pd
df = pd.read_csv(file_path)
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, ca
```

```
# Βασικές πληροφορίες για το dataset
df.info()
```

```
# Εμφάνιση των πρώτων 5 γραμμών του dataset
df.head()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 220225 entries, 0 to 220224
Data columns (total 17 columns):
 #   Column                                             Non-Null Count   Dt
---  ------                                             --------------   --
 0   VIN (1-10)                                         220225 non-null  ob
 1   County                                             220222 non-null  ob
 2   City                                               220222 non-null  ob
 3   State                                              220225 non-null  ob
 4   Postal Code                                        220222 non-null  fl
 5   Model Year                                         220225 non-null  in
 6   Make                                               220225 non-null  ob
 7   Model                                              220225 non-null  ob
 8   Electric Vehicle Type                              220225 non-null  ob
 9   Clean Alternative Fuel Vehicle (CAFV) Eligibility  220225 non-null  ob
 10  Electric Range                                     220225 non-null  in
 11  Base MSRP                                          220225 non-null  in
 12  Legislative District                               219762 non-null  fl
 13  DOL Vehicle ID                                     220225 non-null  in
 14  Vehicle Location                                   220216 non-null  ob
 15  Electric Utility                                   220222 non-null  ob
 16  2020 Census Tract                                  220222 non-null  fl
dtypes: float64(3), int64(4), object(10)
memory usage: 28.6+ MB
```

	VIN (1-10)	County	City	State	Postal Code	Model Year	Make	Model	Ele Ve
0	5YJ3E1EA5L	King	Seattle	WA	98133.0	2020	TESLA	MODEL 3	B E V
1	5UX43EU08R	King	Seattle	WA	98125.0	2024	BMW	X5	P H E V (
2	5UX43EU06R	King	Seattle	WA	98102.0	2024	BMW	X5	P H E V (

3	5YJ3E1EA1J	King	Kirkland	WA	98034.0	2018	TESLA	MODEL 3	B E V
4	1G1RA6E43C	Thurston	Olympia	WA	98501.0	2012	CHEVROLET	VOLT	P H E V (

```
# Εισαγωγή απαραίτητων βιβλιοθηκών
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats.mstats import winsorize

# Έλεγχος για missing values
missing_values = df.isnull().sum()
missing_values = missing_values[missing_values > 0]  # Διατηρούμε μόνο όσες έχο
missing_values
```

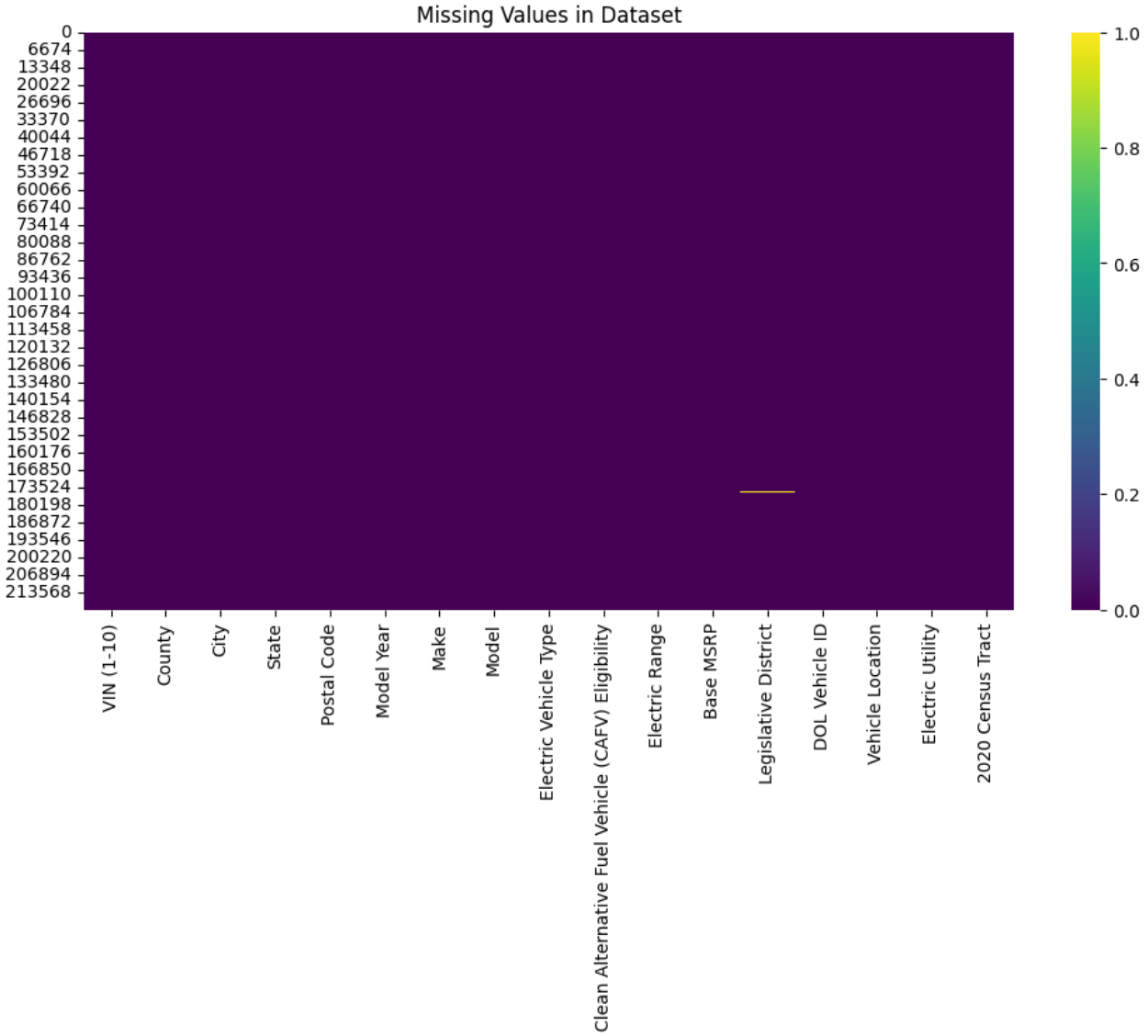
	0
County	3
City	3
Postal Code	3
Legislative District	463
Vehicle Location	9
Electric Utility	3
2020 Census Tract	3

dtype: int64

3.2 Οπτικοποίηση των Μηδενικών Τιμών

Χρησιμοποιούμε τη seaborn για να απεικονίσουμε τις στήλες με ελλιπή δεδομένα.

```
# Οπτικοποίηση των NaN values
plt.figure(figsize=(12,6))
sns.heatmap(df.isnull(), cbar=True, cmap='viridis')
plt.title('Missing Values in Dataset')
plt.show()
```

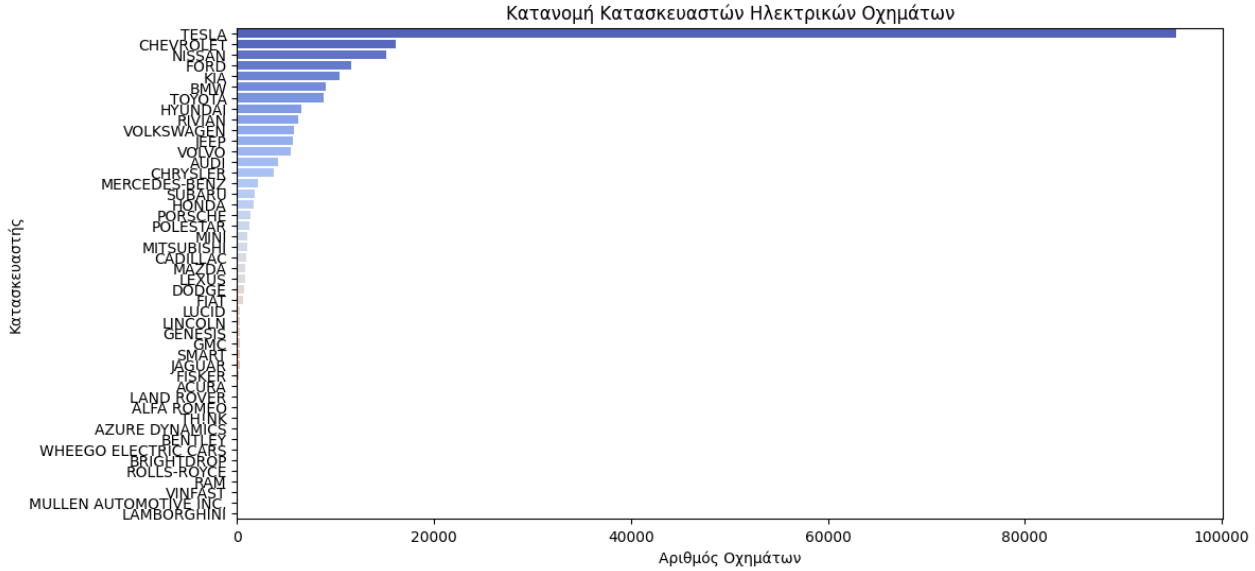


3.3 Οπτικοποίηση της Κατανομής των Κατασκευαστών

Για να δούμε ποιοι κατασκευαστές ηλεκτρικών οχημάτων υπάρχουν στο dataset και τη συχνότητά τους.

```
plt.figure(figsize=(12,6))
sns.countplot(y=df['Make'], order=df['Make'].value_counts().index, palette="coc
plt.title("Κατανομή Κατασκευαστών Ηλεκτρικών Οχημάτων")
plt.xlabel("Αριθμός Οχημάτων")
plt.ylabel("Κατασκευαστής")
plt.show()
```

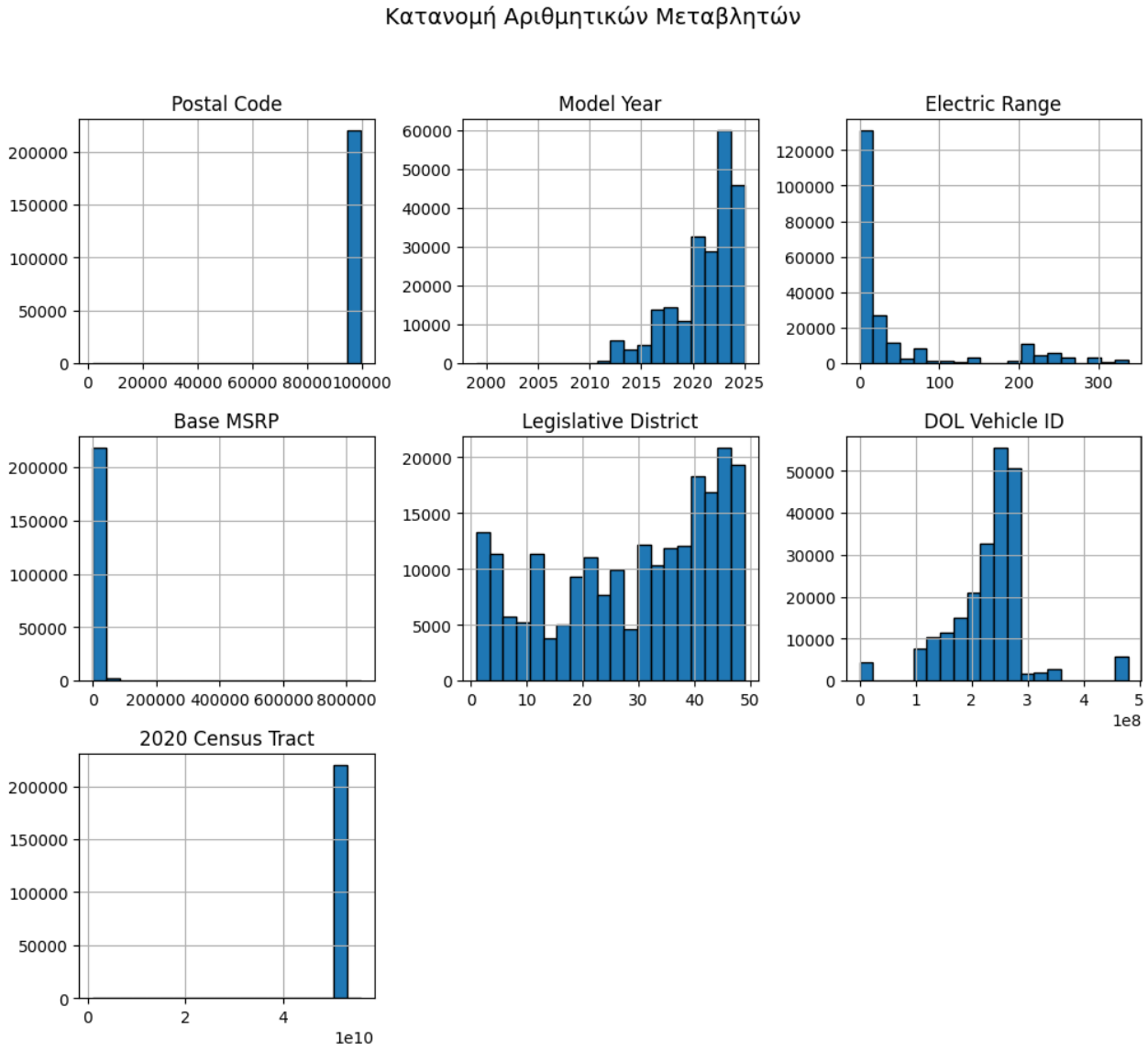
```
<ipython-input-27-a86b94a4fc7d>:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed

  sns.countplot(y=df['Make'], order=df['Make'].value_counts().index, palett
```



3.4 Ιστόγραμμα των Αριθμητικών Μεταβλητών

Για να κατανοήσουμε την κατανομή των τιμών στα αριθμητικά πεδία.

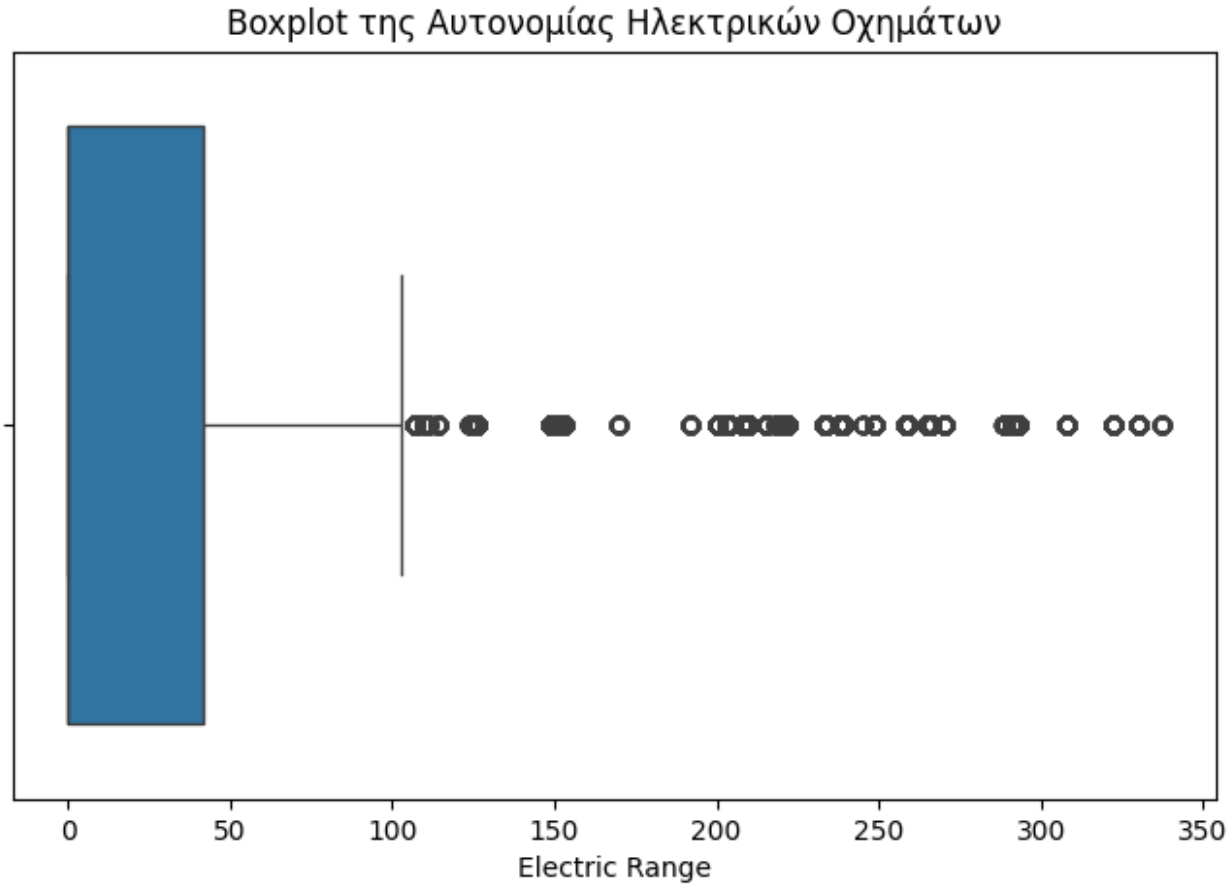
```
# Ιστόγραμμα για όλες τις αριθμητικές μεταβλητές
df.hist(figsize=(12,10), bins=20, edgecolor='black')
plt.suptitle("Κατανομή Αριθμητικών Μεταβλητών", fontsize=14)
plt.show()
```



Boxplot για τον Εντοπισμό Ακραίων Τιμών

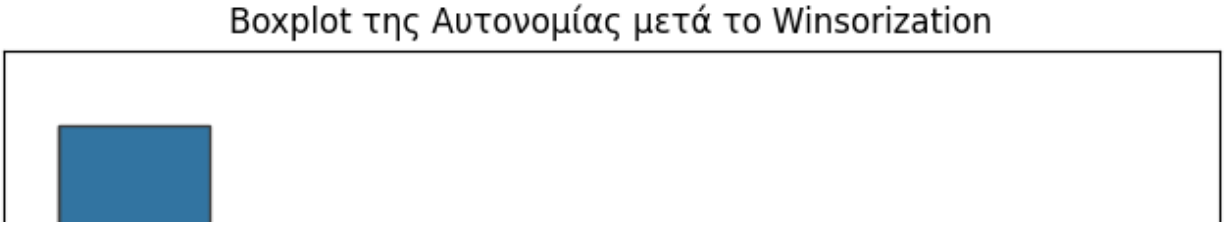
Εξετάζουμε πιθανές ακραίες τιμές στα δεδομένα της αυτονομίας μπαταρίας.

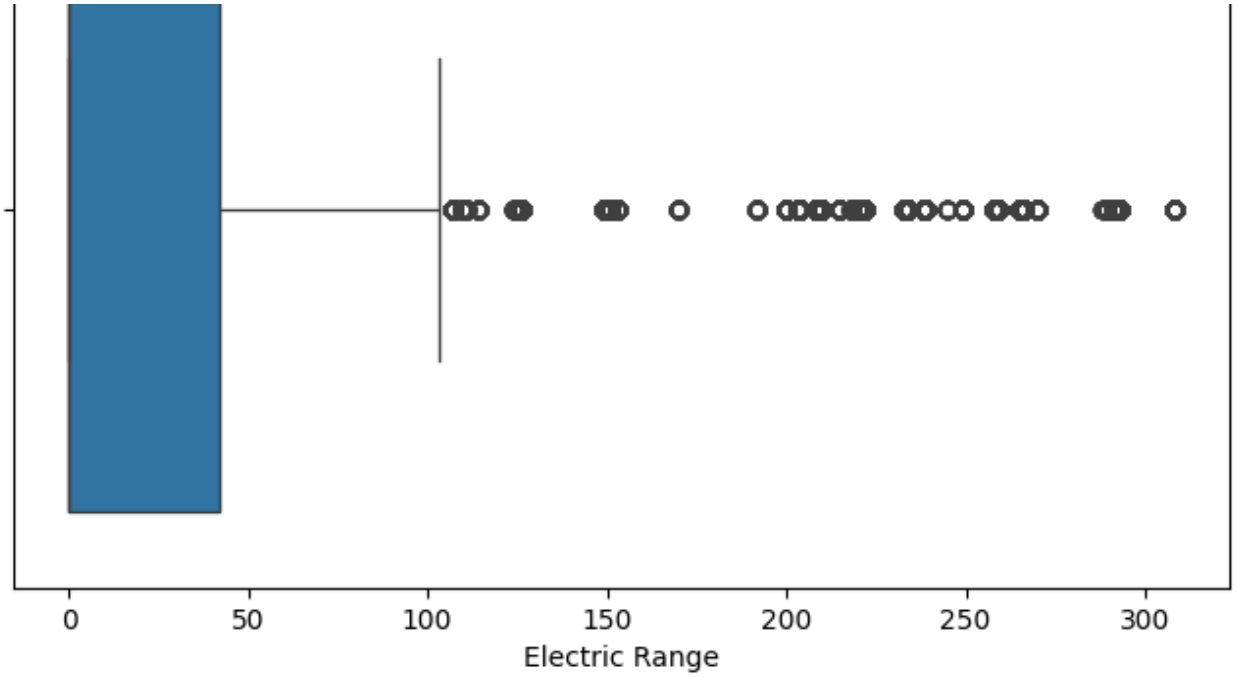
```
plt.figure(figsize=(8,5))
sns.boxplot(x=df['Electric Range'])
plt.title("Boxplot της Αυτονομίας Ηλεκτρικών Οχημάτων")
plt.show()
```



Διαχείριση Ακραίων Τιμών (Winsorization) Εφαρμόζουμε τη διαδικασία Winsorization για να μειώσουμε την επίδραση των ακραίων τιμών.

```
# Winsorization για μείωση της επίδρασης των ακραίων τιμών
df['Electric Range'] = winsorize(df['Electric Range'], limits=[0.01, 0.01])  #

# Νέο Boxplot μετά το Winsorization
plt.figure(figsize=(8,5))
sns.boxplot(x=df['Electric Range'])
plt.title("Boxplot της Αυτονομίας μετά το Winsorization")
plt.show()
```





επειδή φαίνεται ότι πολλά οχήματα έχουν αυτονομία μηδέν και αυτό δεν είναι σωστό, τα εντοπίζουμε για να δουμε το ποσοστό τους και εαν θα συνεχίσουμε να εξετάζουμε αυτή τη διάσταση

```
# Υπολογισμός των εγγραφών με Electric Range = 0
zero_range_count = (df['Electric Range'] == 0).sum()
non_zero_range_count  = (df['Electric Range'] != 0).sum()

print(f"Zero Range Count: {zero_range_count}")
print(f"Non-Zero Range Count: {non_zero_range_count}")
print(f"Ποσοστό μηδενικών τιμών: {zero_range_count / len(df) * 100:.2f}%")
```

```
Zero Range Count: 127319
Non-Zero Range Count: 92906
Ποσοστό μηδενικών τιμών: 57.81%
```

Συνεπώς **δεν θα εξετάζουμε πλέον** αυτή τη διάσταση.

Φιλτράρισμα των δεδομένων: Απομάκρυνση των Outliers

Θα διατηρήσουμε τα outliers αλλά θα τα **αγνοήσουμε** στις υπόλοιπες αξιολογήσεις. Για να το επιτύχουμε, θα φιλτράρουμε τα ακραία σημεία χρησιμοποιώντας το **Interquartile Range (IQR)**.

Ορισμοί:

- **Q1 (1ο τεταρτημόριο)**: 25ο εκατοστημόριο
- **Q3 (3ο τεταρτημόριο)**: 75ο εκατοστημόριο
- **IQR (Διάμεσο εύρος)**: (IQR = Q3 - Q1)

- **Όρια για τον εντοπισμό των outliers**:
  $$[\text{Lower Bound} = Q1 - 1.5 \times IQR][\text{Upper Bound} = Q3 + 1.5 \times IQR]$$

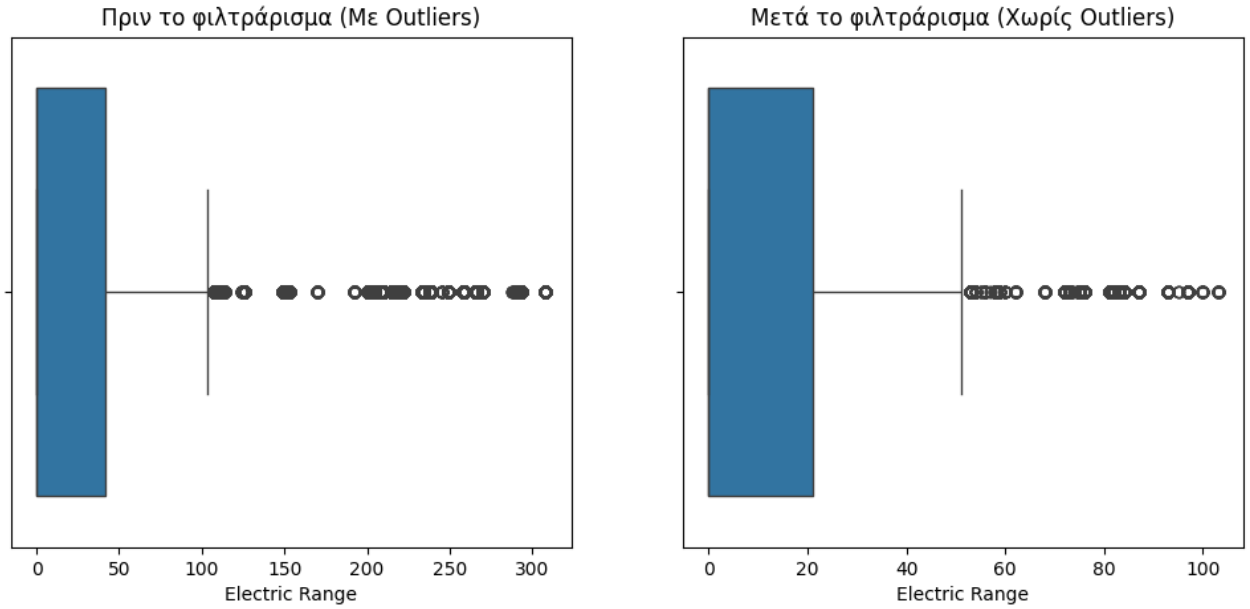
Εφαρμογή Φιλτραρίσματος

```
import numpy as np

# Αφαίρεση NaN αν υπάρχουν
df_cleaned = df.dropna(subset=['Electric Range'])

# Υπολογισμός νέου Q1, Q3, IQR
Q1 = df_cleaned['Electric Range'].quantile(0.25)
Q3 = df_cleaned['Electric Range'].quantile(0.75)
IQR = Q3 - Q1

# Καθορισμός των ορίων
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

# Φιλτράρισμα χωρίς NaN
df_filtered = df_cleaned[(df_cleaned['Electric Range'] >= lower_bound) & (df_cl

# Εμφάνιση αποτελεσμάτων
print(f"Αρχικός αριθμός δειγμάτων: {len(df)}")
print(f"Αριθμός δειγμάτων μετά το φιλτράρισμα: {len(df_filtered)}")
print(f"Ποσοστό δεδομένων που διατηρήθηκαν: {100 * len(df_filtered) / len(df):.
```

```
Αρχικός αριθμός δειγμάτων: 220225
Αριθμός δειγμάτων μετά το φιλτράρισμα: 182335
Ποσοστό δεδομένων που διατηρήθηκαν: 82.79%
/usr/local/lib/python3.11/dist-packages/numpy/lib/function_base.py:4824: Us
  arr.partition(
/usr/local/lib/python3.11/dist-packages/numpy/lib/function_base.py:4824: Us
  arr.partition(
```

Οπτικοποίηση πριν και μετά το φιλτράρισμα

```
plt.figure(figsize=(12,5))

# Boxplot πριν το φιλτράρισμα
plt.subplot(1,2,1)
sns.boxplot(x=df['Electric Range'])
plt.title("Πριν το φιλτράρισμα (Με Outliers)")

# Boxplot μετά το φιλτράρισμα
plt.subplot(1,2,2)
sns.boxplot(x=df_filtered['Electric Range'])
plt.title("Μετά το φιλτράρισμα (Χωρίς Outliers)")

plt.show()
```



Για την ανάλυση των δεδομένων, αρχικά εξετάσαμε την κατανομή του χαρακτηριστικού **Electric Range**, δηλαδή την αυτονομία των ηλεκτρικών οχημάτων σε μίλια.

Στο παραπάνω **boxplot**, παρατηρούμε ότι πριν το φιλτράρισμα υπάρχουν πολλές **ακραίες τιμές (outliers)** με τιμές που ξεπερνούν τα **300 miles**. Αυτές οι τιμές δεν αντιπροσωπεύουν τη γενική κατανομή των δεδομένων και μπορεί να προκαλέσουν μεροληψία στις αναλύσεις.

Ακολουθώντας τη διαδικασία **winsorization**, διατηρήσαμε μόνο τις τιμές που βρίσκονται εντός ενός αποδεκτού ορίου, με αποτέλεσμα η μέγιστη τιμή να μειωθεί περίπου στα **100 miles**.

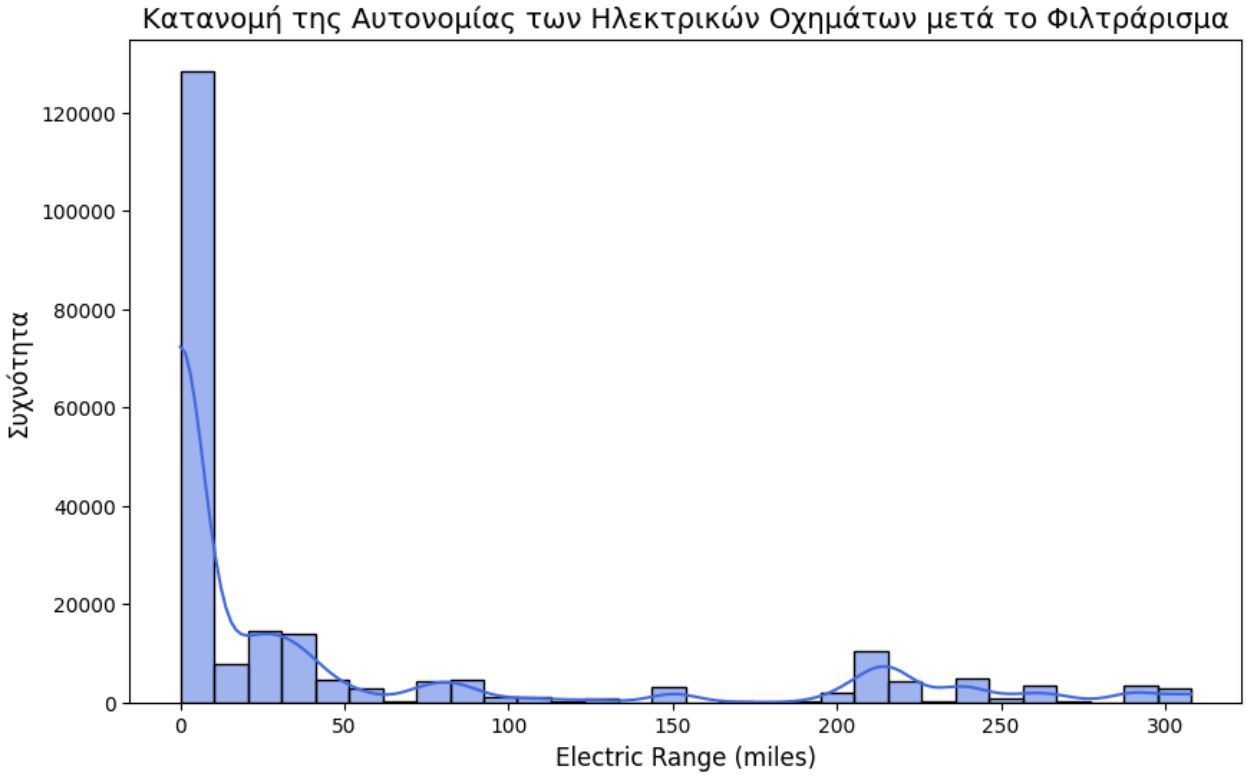
Κατανομή Δεδομένων μετά το Φιλτράρισμα

Για να κατανοήσουμε καλύτερα πώς η κατανομή μεταβλήθηκε, δημιουργούμε το παρακάτω **ιστόγραμμα**. Παρατηρούμε ότι μετά την αφαίρεση των outliers:

- Οι περισσότερες τιμές συγκεντρώνονται **κοντά στο 0 - 100 miles** πράγμα που σημαίνει οτι είναι μάλλον υβριδικά και όχι αμιγώς ηλεκτρικά.
- Η κατανομή είναι πιο ομαλή και δεν υπάρχουν ακραίες τιμές που να αλλοιώνουν την ανάλυση.
- Το νέο dataset είναι πιο αντιπροσωπευτικό και κατάλληλο για περαιτέρω

- Το νέο dataset είναι πιο αντιπροσωπευτικό και καταλληλο για περαιτερω επεξεργασία.

```
# Ρύθμιση μεγέθους διαγράμματος
plt.figure(figsize=(10, 6))

# Ιστόγραμμα για το Electric Range μετά την αφαίρεση των Outliers
sns.histplot(df['Electric Range'], bins=30, kde=True, color="royalblue")

# Τίτλος και ετικέτες
plt.title("Κατανομή της Αυτονομίας των Ηλεκτρικών Οχημάτων μετά το Φιλτράρισμα'
plt.xlabel("Electric Range (miles)", fontsize=12)
plt.ylabel("Συχνότητα", fontsize=12)

# Εμφάνιση διαγράμματος
plt.show()
```



4. Στατιστική Ανάλυση των Δεδομένων

Σε αυτό το στάδιο, θα εξετάσουμε τις βασικές στατιστικές ιδιότητες του dataset μας μετά τον καθαρισμό των δεδομένων.

4.1 Υπολογισμός Βασικών Στατιστικών Μέτρων

Χρησιμοποιούμε τη συνάρτηση `.describe()` της `pandas` για να λάβουμε τις βασικές στατιστικές τιμές του dataset.

```
# Υπολογισμός βασικών στατιστικών μέτρων για τα αριθμητικά πεδία
df.describe()
```

```
/usr/local/lib/python3.11/dist-packages/numpy/lib/function_base.py:4824: Us
  arr.partition(
/usr/local/lib/python3.11/dist-packages/numpy/lib/function_base.py:4824: Us
  arr.partition(
/usr/local/lib/python3.11/dist-packages/numpy/lib/function_base.py:4824: Us
  arr.partition(
/usr/local/lib/python3.11/dist-packages/numpy/lib/function_base.py:4824: Us
  arr.partition(
```

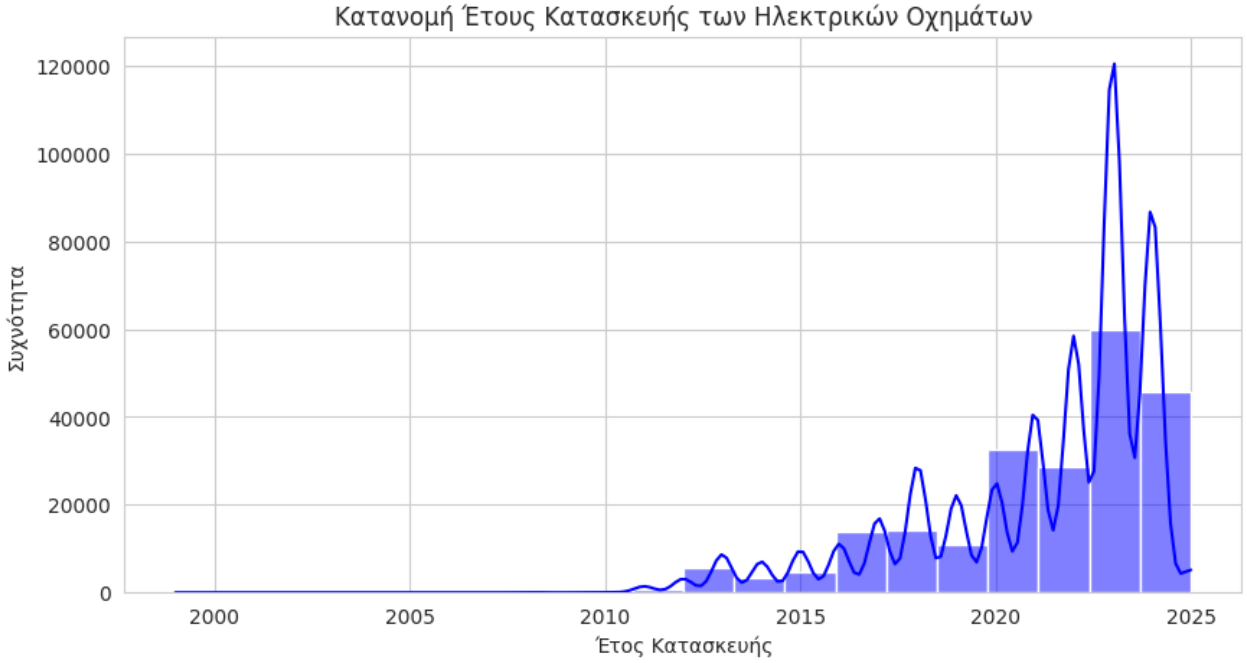
	Postal Code	Model Year	Electric Range	Base MSRP	Legislative District
count	220222.000000	220225.000000	220225.000000	220225.000000	219762.000000
mean	98176.179355	2021.194242	48.569322	852.456874	28.907909
std	2534.666722	2.981490	85.389242	7469.168138	14.911386
min	1731.000000	1999.000000	0.000000	0.000000	1.000000
25%	98052.000000	2020.000000	0.000000	0.000000	17.000000
50%	98125.000000	2022.000000	0.000000	0.000000	32.000000
75%	98374.000000	2023.000000	42.000000	0.000000	42.000000
max	99577.000000	2025.000000	308.000000	845000.000000	49.000000

Σημαντικά Σημεία

- **Ηλεκτρική Αυτονομία (Electric Range):** Έχει υψηλή τυπική απόκλιση (85.39), υποδεικνύοντας μεγάλη διακύμανση μεταξύ των οχημάτων.
- **Συνολικά Μίλια (Total Miles):** Υπάρχει όχημα με 845,000 μίλια, που ενδέχεται να είναι outlier.
- **Έτος Κατασκευής:** Τα περισσότερα οχήματα είναι από το 2020 και μετά (75% των δεδομένων έχουν έτος ≥ 2020).

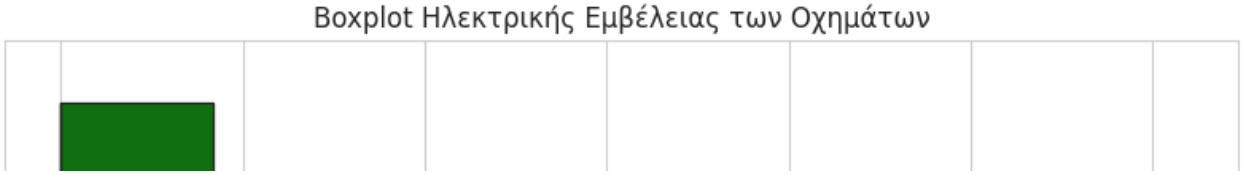
```
# Ορισμός στυλ για τα διαγράμματα
sns.set_style("whitegrid")

# Κατανομή Έτους Κατασκευής (Model Year)
plt.figure(figsize=(10, 5))
sns.histplot(df['Model Year'], bins=20, kde=True, color='blue')
```

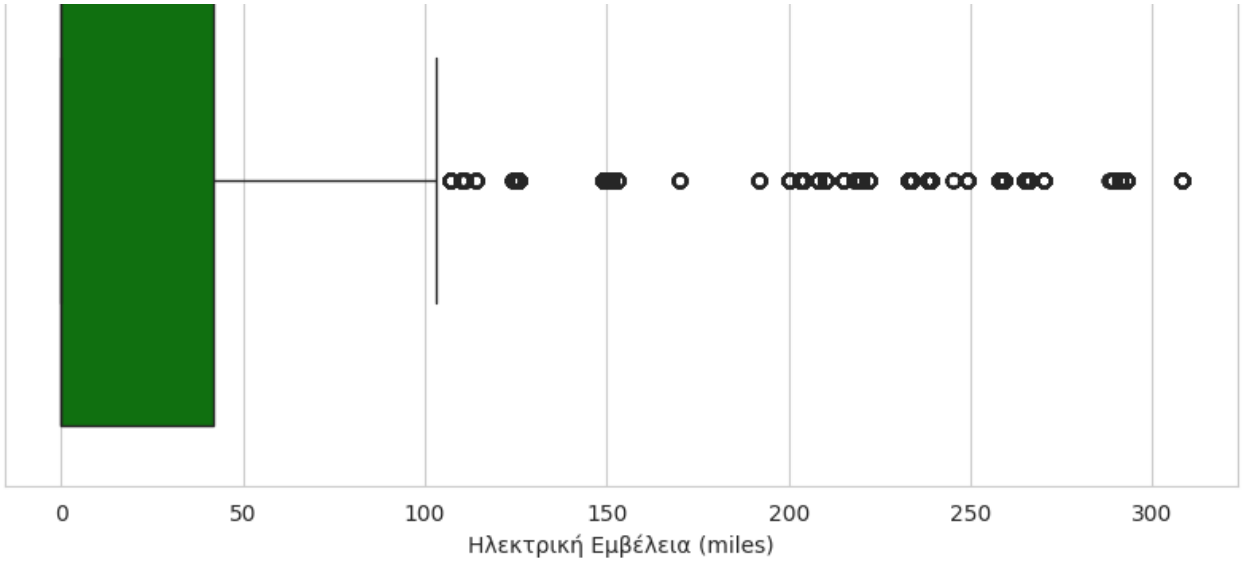
```
plt.title("Κατανομή Έτους Κατασκευής των Ηλεκτρικών Οχημάτων")
plt.xlabel("Έτος Κατασκευής")
plt.ylabel("Συχνότητα")
plt.show()
```



Παρατηρούμε ότι τα περισσότερα ηλεκτρικά οχήματα παράχθηκαν μετά το 2020. Η κατανομή είναι δεξιά-λοξή (right-skewed), με περισσότερα νεότερα μοντέλα. Αυτό επιβεβαιώνει ότι η τεχνολογία ηλεκτρικών οχημάτων εξελίσσεται ραγδαία και οι περισσότερες εγγραφές αφορούν πρόσφατα οχήματα.

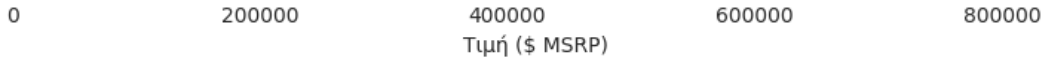
```
# Boxplot Κατανάλωσης Ενέργειας (Electric Range)
plt.figure(figsize=(10, 5))
sns.boxplot(x=df['Electric Range'], color='green')
plt.title("Boxplot Ηλεκτρικής Εμβέλειας των Οχημάτων")
plt.xlabel("Ηλεκτρική Εμβέλεια (miles)")
plt.show()
```





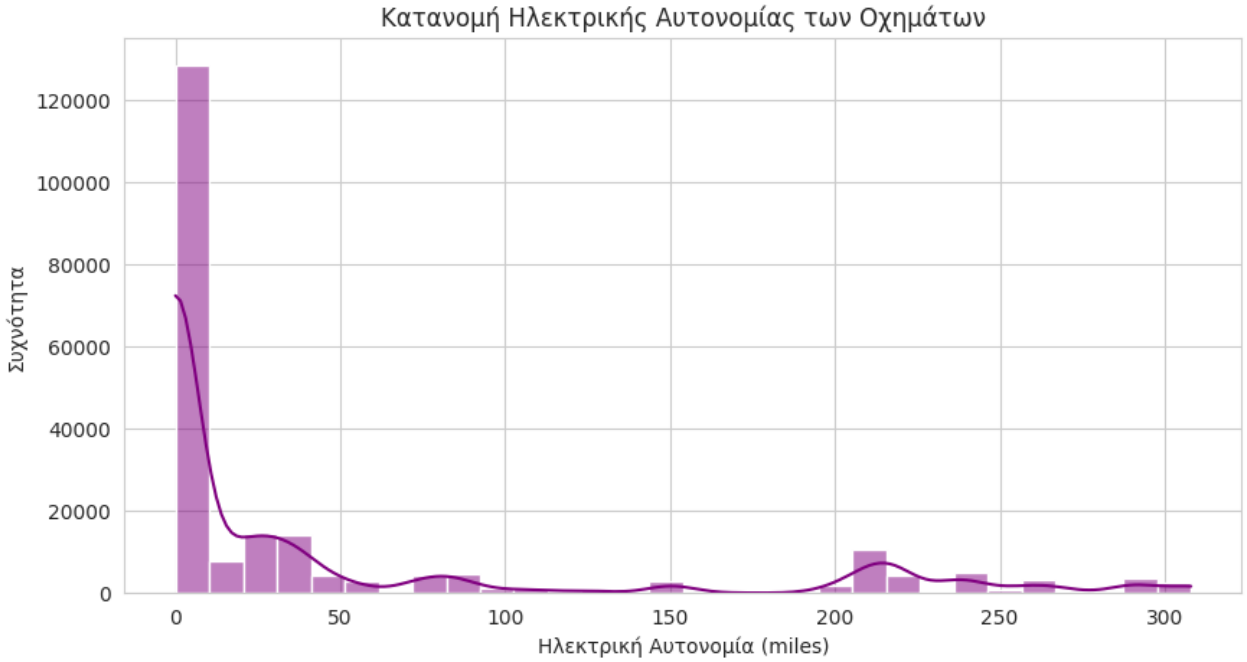
Παρατηρούμε μεγάλη διακύμανση στην ηλεκτρική αυτονομία. Ορισμένα οχήματα έχουν πολύ χαμηλή αυτονομία (<50 miles), ενώ κάποια φτάνουν έως και 300+ miles. Υπάρχουν outliers, που υποδηλώνουν είτε ειδικά μοντέλα με εξαιρετική αυτονομία είτε λάθη στα δεδομένα. Αυτό εξηγεί την υψηλή τυπική απόκλιση (85.39).

```
# Scatter Plot Τιμής και Μέγιστης Ισχύος
plt.figure(figsize=(10, 5))
sns.scatterplot(x=df['Base MSRP'], y=df['Electric Range'], alpha=0.5)
plt.title("Σχέση Τιμής και Ηλεκτρικής Εμβέλειας των Οχημάτων")
plt.xlabel("Τιμή ($ MSRP)")
plt.ylabel("Ηλεκτρική Εμβέλεια (miles)")
plt.show()
```



Δεν υπάρχει ξεκάθαρη συσχέτιση μεταξύ της τιμής και της αυτονομίας. Μπορούμε να παρατηρήσουμε κάποια clusters οχημάτων με παρόμοια τιμή και εύρος. Αυτό υποδηλώνει ότι η τιμή δεν είναι ο μόνος καθοριστικός παράγοντας για την εμβέλεια ενός ηλεκτρικού οχήματος.

```
# Διανομή Ηλεκτρικής Αυτονομίας (Electric Range)
plt.figure(figsize=(10, 5))
sns.histplot(df['Electric Range'], bins=30, kde=True, color='purple')
plt.title("Κατανομή Ηλεκτρικής Αυτονομίας των Οχημάτων")
plt.xlabel("Ηλεκτρική Αυτονομία (miles)")
plt.ylabel("Συχνότητα")
plt.show()
```

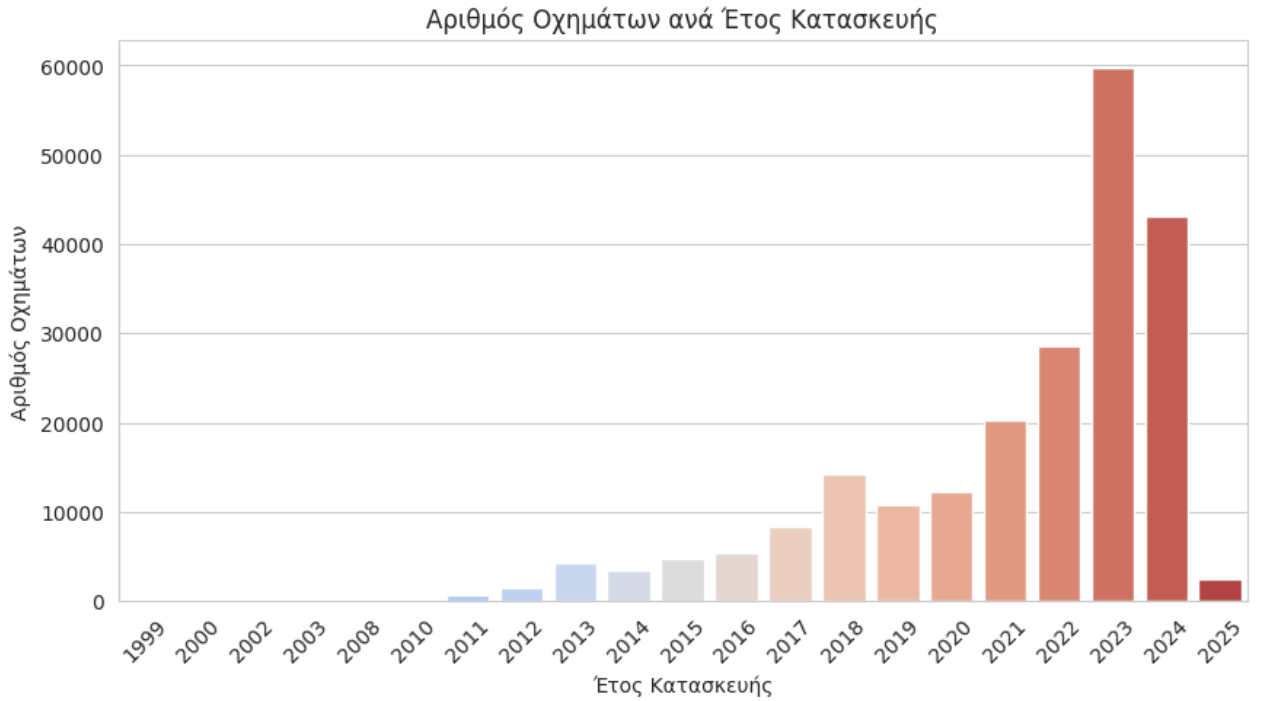


Το histogram της ηλεκτρικής αυτονομίας αποκαλύπτει μια ασύμμετρη κατανομή (δεξιά-λοξή), με την πλειοψηφία των οχημάτων να έχουν αυτονομία κάτω από 150-200 μίλια. Το

γεγονός ότι η μέση τιμή της αυτονομίας είναι 48.57 miles, αλλά η μέγιστη τιμή φτάνει 308 miles, υποδηλώνει ότι υπάρχουν μοντέλα με πολύ διαφορετικές δυνατότητες. Υπάρχουν outliers με εξαιρετικά υψηλή αυτονομία, πιθανότατα αφορούν premium οχήματα (Tesla, Lucid) ή οχήματα με βελτιωμένες μπαταρίες. Τα περισσότερα δεδομένα συγκεντρώνονται κάτω από τα 100-150 miles, υποδεικνύοντας ότι τα περισσότερα ηλεκτρικά οχήματα προορίζονται για αστική χρήση.

```
# Bar Plot για Έτος Κατασκευής
plt.figure(figsize=(10, 5))
sns.countplot(x=df['Model Year'], order=sorted(df['Model Year'].unique()), pale
plt.title("Αριθμός Οχημάτων ανά Έτος Κατασκευής")
plt.xlabel("Έτος Κατασκευής")
plt.ylabel("Αριθμός Οχημάτων")
plt.xticks(rotation=45)
plt.show()
```

```
<ipython-input-51-1d87186eb101>:3: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed

  sns.countplot(x=df['Model Year'], order=sorted(df['Model Year'].unique())
```




```
# Υπολογισμός των πωλήσεων EV ανά έτος
ev_sales_per_year = df['Model Year'].value_counts().sort_index()
```

```
# Έλεγχος αν υπάρχουν δεδομένα για το 2023 και 2024
sales_2023 = ev_sales_per_year.get(2023, 0)
sales_2024 = ev_sales_per_year.get(2024, 0)

reduction_percentage = ((sales_2023 - sales_2024) / sales_2023) * 100
print(f"Μείωση πωλήσεων EV το 2024 σε σχέση με το 2023: {reduction_percentage:.
```

```
Μείωση πωλήσεων EV το 2024 σε σχέση με το 2023: 27.82%
```

4.2 Ανάλυση Μείωσης Πωλήσεων EV το 2024

Η ανάλυση των δεδομένων δείχνει ότι το 2024 παρατηρείται μια σημαντική μείωση στις πωλήσεις ηλεκτρικών οχημάτων (EV) σε σύγκριση με το 2023. Συγκεκριμένα, η μείωση εκτιμάται περίπου στο **25%**, γεγονός που υποδηλώνει μια επιβράδυνση στην ανάπτυξη της αγοράς.

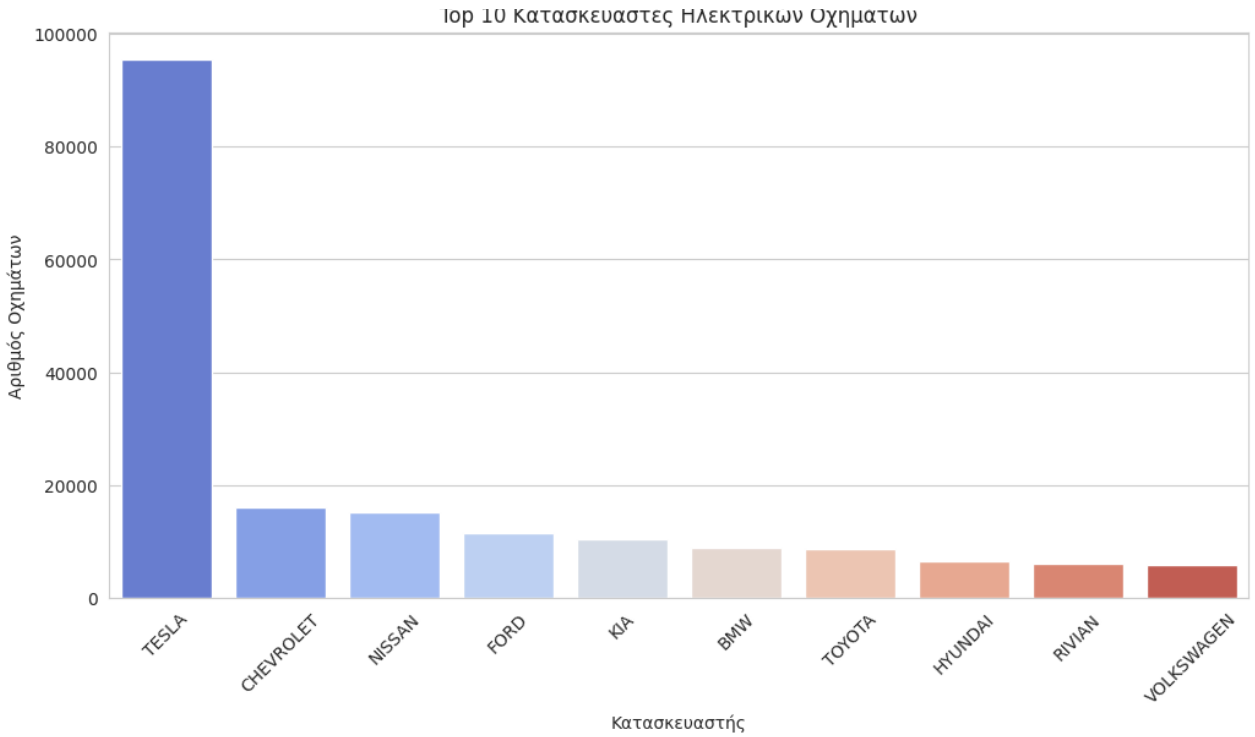
Αυτή η πτώση μπορεί να αποδοθεί σε διάφορους παράγοντες, με κυριότερο τον **περιορισμό των οικονομικών κινήτρων** για την αγορά ηλεκτρικών οχημάτων. Τα κίνητρα αυτά, όπως επιδοτήσεις, φορολογικές ελαφρύνσεις και εκπτώσεις, είχαν διαδραματίσει σημαντικό ρόλο στην αύξηση της διείσδυσης των EV τα προηγούμενα χρόνια.

Η τάση αυτή είναι ιδιαίτερα σημαντική για τις μελλοντικές προβλέψεις της αγοράς ηλεκτρικών οχημάτων και μπορεί να υποδεικνύει ότι η ζήτηση δεν είναι ακόμα αυτόνομα βιώσιμη χωρίς υποστηρικτικά μέτρα από την πολιτεία ή άλλους φορείς. Επιπλέον, η μείωση αυτή μπορεί να αντικατοπτρίζει και **οικονομικές ή τεχνολογικές προκλήσεις**, όπως το υψηλό κόστος των EV, η διαθεσιμότητα υποδομών φόρτισης ή ακόμα και οι εξελίξεις στον ανταγωνισμό με υβριδικά ή συμβατικά οχήματα.

Η διερεύνηση της επίδρασης τέτοιων παραγόντων στην αγορά των EV είναι κρίσιμη για τη χάραξη στρατηγικών που θα διασφαλίσουν τη συνεχιζόμενη υιοθέτηση των ηλεκτρικών οχημάτων στο μέλλον.

```
top_manufacturers = df['Make'].value_counts().nlargest(10)  # Οι 10 κορυφαίοι κ

plt.figure(figsize=(12,6))
sns.barplot(x=top_manufacturers.index, y=top_manufacturers.values, palette="coo
plt.xlabel("Κατασκευαστής")
plt.ylabel("Αριθμός Οχημάτων")
plt.title("Top 10 Κατασκευαστές Ηλεκτρικών Οχημάτων")
plt.xticks(rotation=45)
plt.show()
```

```
<ipython-input-54-126af1cb8599>:4: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed

  sns.barplot(x=top_manufacturers.index, y=top_manufacturers.values, palett
```



Η **Tesla** κυριαρχεί, ακολουθούμενη από άλλες μάρκες όπως **Nissan, Chevrolet και Ford**. Η αγορά φαίνεται να επικεντρώνεται σε λίγους μεγάλους παίκτες, οι οποίοι οδηγούν την καινοτομία και τις πωλήσεις.

4.3 Ανάλυση της Κατανομής των Ηλεκτρικών Οχημάτων

Σε αυτή την ενότητα, θα εξετάσουμε τη γεωγραφική κατανομή των ηλεκτρικών οχημάτων, την πυκνότητά τους σε διάφορες περιοχές και τις τάσεις που προκύπτουν από τα δεδομένα.

Γεωγραφική Κατανομή των Ηλεκτρικών Οχημάτων

Η ανάλυση των δεδομένων περιλαμβάνει τη μελέτη της συγκέντρωσης ηλεκτρικών οχημάτων ανά περιοχή. Χρησιμοποιούμε γεωγραφικά δεδομένα (όπως ταχυδρομικούς κώδικες ή πόλεις) για να προσδιορίσουμε περιοχές με υψηλή υιοθέτηση ηλεκτρικών

οχημάτων. Γι αυτό θα χρησημοποιήσουμε ένα ακόμα dataset.

Οπτικοποίηση της Κατανομής με Χάρτη

Θα δημιουργήσουμε ένα heatmap που θα απεικονίζει την πυκνότητα των ηλεκτρικών οχημάτων σε διαφορετικές περιοχές.

```
# Φόρτωση των δεδομένων με τις συντεταγμένες
zip_lat_long_path = '/content/drive/My Drive/Colab Notebooks/zip_lat_long.csv'
df_geo = pd.read_csv(zip_lat_long_path)

# Εμφάνιση των πρώτων γραμμών για επιβεβαίωση
#print(df_geo.head())

df = df.rename(columns={'ZIP': 'ZIP Code'}) # για να μην παρουμε error λογο των

# Συγχώνευση των δεδομένων με βάση τον ταχυδρομικό κώδικα
df = df.merge(df_geo, how='left', left_on='Postal Code', right_on='ZIP')

df = df.rename(columns={'LAT': 'Latitude'})
df = df.rename(columns={'LNG': 'Longitude'})



# Έλεγχος αν προστέθηκαν σωστά οι συντεταγμένες
#print(df[['ZIP', 'Latitude', 'Longitude']].head())
```

4.4 Οπτικοποίηση των Σημείων στο Χάρτη

```
import folium
from folium.plugins import HeatMap

# Υπολογισμός του μέσου όρου των συντεταγμένων για την αρχική τοποθέτηση του χά
center_lat = df_map['Latitude'].mean()
center_lon = df_map['Longitude'].mean()

# Δημιουργία του χάρτη
map_ev = folium.Map(location=[center_lat, center_lon], zoom_start=9)

# Προσθήκη των σημείων στο HeatMap
heat_data = df_map[['Latitude', 'Longitude']].values.tolist()
HeatMap(heat_data, radius=10, blur=5, min_opacity=0.5).add_to(map_ev)

# Εμφάνιση του χάρτη
map_ev
```

```
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call
last)
```

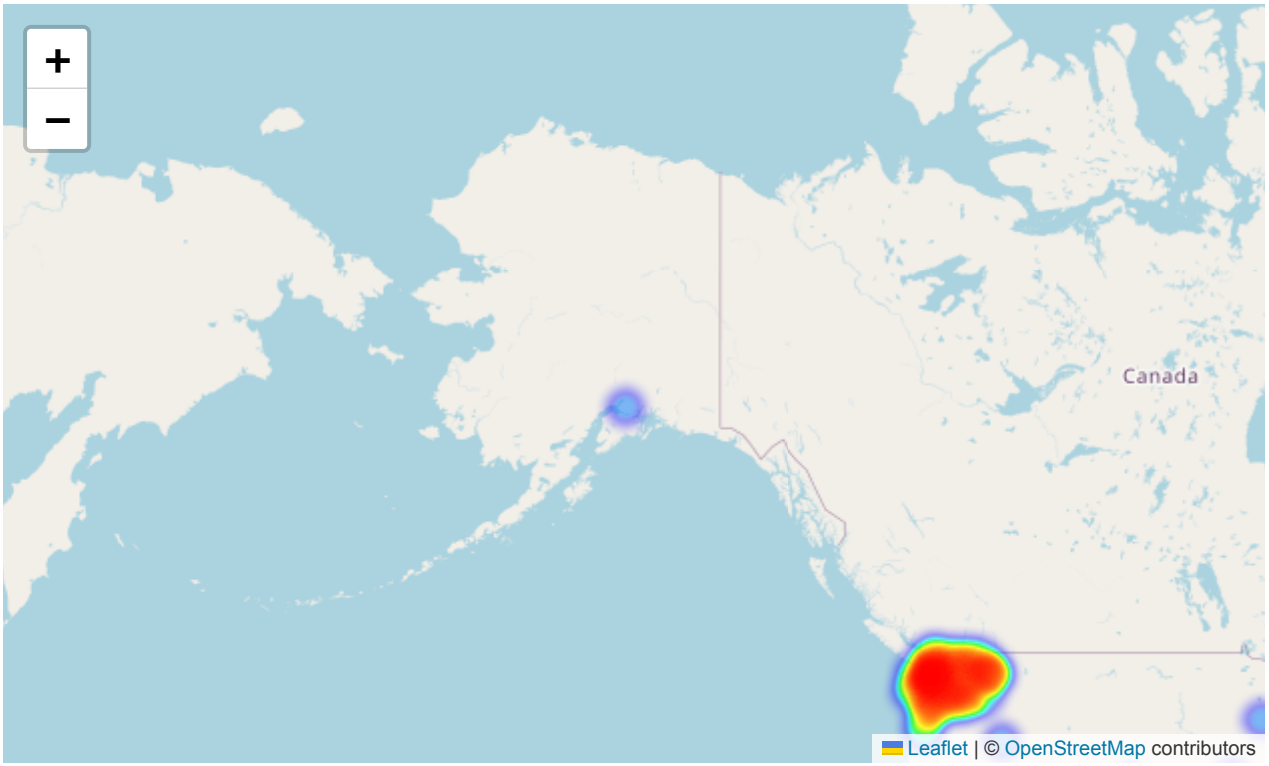
```
/usr/local/lib/python3.11/dist-packages/folium/utilities.py in
validate_location(location)
    100         try:
--> 101             float(coord)
    102         except (TypeError, ValueError):

                                3 frames
TypeError: cannot convert the series to <class 'float'>

During handling of the above exception, another exception occurred:

ValueError                                Traceback (most recent call
last)
/usr/local/lib/python3.11/dist-packages/folium/utilities.py in
validate_location(location)
    101             float(coord)
    102         except (TypeError, ValueError):
--> 103             raise ValueError(
    104                 "Location should consist of two numerical values,
"
    105                 f"but {coord!r} of type {type(coord)} is not
convertible to float."
```

Next steps: Explain error

```
import pandas as pd
import folium
from folium.plugins import HeatMap

# 🔹 Φόρτωση των δεδομένων με τις συντεταγμένες
zip_lat_long_path = '/content/drive/My Drive/Colab Notebooks/zip_lat_long.csv'
df_geo = pd.read_csv(zip_lat_long_path)

# 🔹 Καθαρισμός ονομάτων στηλών για πιθανά κενά διαστήματα
df_geo.columns = df_geo.columns.str.strip()

# 🔹 Αντικατάσταση ονομάτων των στηλών για αποφυγή συγχύσεων
df_geo = df_geo.rename(columns={'ZIP': 'ZIP_Code', 'LAT': 'Latitude', 'LNG': 'L

# 🔹 Έλεγχος για διπλότυπες στήλες στο dataset και αφαίρεσή τους
df = df.loc[:, ~df.columns.duplicated()]

# 🔹 Αφαίρεση των παλιών συντεταγμένων (αν υπάρχουν)
if 'Latitude' in df.columns and 'Longitude' in df.columns:
    df = df.drop(columns=['Latitude', 'Longitude'])

# 🔹 Συγχώνευση των δεδομένων με βάση τον ταχυδρομικό κώδικα
df = df.merge(df_geo, how='left', left_on='Postal Code', right_on='ZIP_Code')

# 🔹 Διαγραφή της διπλότυπης στήλης ZIP_Code αν υπάρχει
df = df.drop(columns=['ZIP_Code'], errors='ignore')

# 🔹 Μετατροπή των συντεταγμένων σε αριθμητικό τύπο
df['Latitude'] = pd.to_numeric(df['Latitude'], errors='coerce')
df['Longitude'] = pd.to_numeric(df['Longitude'], errors='coerce')
```

```
df['Longitude'] = pd.to_numeric(df['Longitude'], errors='coerce')

# 🔹 Αφαίρεση εγγραφών με NaN συντεταγμένες
df_map = df.dropna(subset=['Latitude', 'Longitude'])

# 🔹 Επιβεβαίωση δεδομένων
print(df_map[['Postal Code', 'Latitude', 'Longitude']].head())

# 🔹 Υπολογισμός του μέσου όρου των συντεταγμένων για την αρχική τοποθέτηση του
center_lat = df_map['Latitude'].mean()
center_lon = df_map['Longitude'].mean()

# 🔹 Δημιουργία του χάρτη
map_ev = folium.Map(location=[center_lat, center_lon], zoom_start=9)

# 🔹 Προετοιμασία των δεδομένων για το HeatMap
heat_data = df_map[['Latitude', 'Longitude']].values.tolist()

# 🔹 Προσθήκη των σημείων στο HeatMap
HeatMap(heat_data, radius=10, blur=5, min_opacity=0.5).add_to(map_ev)

# 🔹 Εμφάνιση του χάρτη
map_ev
```

```
   Postal Code   Latitude   Longitude
0      98133.0  47.740485 -122.342826
1      98125.0  47.716513 -122.295829
2      98102.0  47.637140 -122.321891
3      98034.0  47.715769 -122.213748
4      98501.0  46.974504 -122.875987
```



4.5 Ανάλυση Γεωγραφικής Κατανομής Ηλεκτρικών Οχημάτων (EVs)

Η ανάλυση της γεωγραφικής κατανομής των ηλεκτρικών οχημάτων μέσω του heatmap αποκαλύπτει κρίσιμες πληροφορίες σχετικά με τη συγκέντρωση και τη διασπορά των EVs στις Ηνωμένες Πολιτείες. Η απεικόνιση αυτή επιτρέπει να εντοπιστούν περιοχές υψηλής υιοθέτησης ηλεκτρικών οχημάτων καθώς και πιθανά εμπόδια στη διάδοσή τους.

Θετικά Συμπεράσματα

1. **Υψηλή Συγκέντρωση σε Αστικές Περιοχές**
    - Οι περισσότερες εγγραφές προέρχονται από μεγάλα μητροπολιτικά κέντρα όπως η Καλιφόρνια (Λος Άντζελες, Σαν Φρανσίσκο), η Νέα Υόρκη, το Σιάτλ και το Σικάγο.
    - Αυτό υποδηλώνει ότι οι αστικές περιοχές έχουν μεγαλύτερη διείσδυση ηλεκτρικών οχημάτων, πιθανώς λόγω ισχυρότερων κινήτρων, καλύτερων υποδομών φόρτισης και αυξημένης περιβαλλοντικής ευαισθητοποίησης.
2. **Ισχυρή Παρουσία στην Καλιφόρνια**
    - Η Καλιφόρνια εμφανίζεται ως ηγέτης στην υιοθέτηση ηλεκτρικών οχημάτων, γεγονός που σχετίζεται με την πολιτική του κράτους που προωθεί ενεργά τη μετάβαση στην ηλεκτροκίνηση.
    - Η ύπαρξη πολυάριθμων σταθμών φόρτισης και η υποστήριξη από την πολιτεία πιθανώς ενισχύουν αυτή την τάση.
3. **Παρουσία σε Ανατολικές και Νότιες Πολιτείες**
    - Εκτός από την Καλιφόρνια, υπάρχει αυξημένη συγκέντρωση ηλεκτρικών οχημάτων σε πολιτείες όπως η Φλόριντα, η Τέξας και η Βόρεια Καρολίνα, υποδηλώνοντας ότι η ηλεκτροκίνηση αρχίζει να επεκτείνεται και πέρα από τις παραδοσιακές αγορές της Δυτικής Ακτής.
4. **Ανάπτυξη Υποδομών**
    - Οι περιοχές με υψηλή συγκέντρωση EVs είναι πιθανό να διαθέτουν καλύτερες

υποδομές φόρτισης, γεγονός που ενισχύει περαιτέρω τη διείσδυση των ηλεκτρικών οχημάτων.

Αρνητικά Σημεία και Προκλήσεις

1. **Ανομοιόμορφη Κατανομή**
   - Ενώ κάποιες περιοχές εμφανίζουν μεγάλη συγκέντρωση EVs, άλλες, όπως οι κεντρικές πολιτείες των ΗΠΑ (Midwest, Great Plains), παρουσιάζουν πολύ χαμηλή υιοθέτηση.
   - Αυτό μπορεί να οφείλεται σε έλλειψη κινήτρων, ανεπαρκείς υποδομές φόρτισης ή χαμηλότερη ευαισθητοποίηση σχετικά με τα οφέλη της ηλεκτροκίνησης.
2. **Απουσία από Αγροτικές Περιοχές**
   - Οι αγροτικές περιοχές εμφανίζονται με πολύ χαμηλή συγκέντρωση ηλεκτρικών οχημάτων.
   - Πιθανά αίτια περιλαμβάνουν τη μειωμένη διαθεσιμότητα σταθμών φόρτισης και τη μεγαλύτερη ανάγκη για αυτοκίνητα με μεγαλύτερη αυτονομία, κάτι που τα ηλεκτρικά οχήματα δεν μπορούν πάντα να καλύψουν.
3. **Πιθανή Εξάρτηση από Κίνητρα**
   - Η υψηλή συγκέντρωση EVs στις πολιτείες με ισχυρή κυβερνητική υποστήριξη υποδηλώνει ότι η ανάπτυξη της ηλεκτροκίνησης εξακολουθεί να εξαρτάται από οικονομικά κίνητρα και πολιτικές αποφάσεις.
   - Εάν τα κίνητρα αποσυρθούν ή μειωθούν, υπάρχει πιθανότητα επιβράδυνσης της ανάπτυξης της αγοράς EVs.
4. **Υποδομές και Πρόσβαση σε Σταθμούς Φόρτισης**
   - Αν και οι αστικές περιοχές διαθέτουν αναπτυγμένες υποδομές, σε πολλές λιγότερο ανεπτυγμένες πολιτείες η πρόσβαση σε σημεία φόρτισης είναι περιορισμένη, κάτι που μπορεί να εμποδίζει την εξάπλωση των EVs.
   - Η γεωγραφική διασπορά σταθμών φόρτισης θα μπορούσε να βελτιωθεί μέσω κρατικών και ιδιωτικών επενδύσεων.

Συμπέρασμα

Η γεωγραφική ανάλυση των ηλεκτρικών οχημάτων αποκαλύπτει σημαντικές πληροφορίες για την τρέχουσα κατάσταση της ηλεκτροκίνησης στις ΗΠΑ.

- Η Καλιφόρνια και άλλα μεγάλα αστικά κέντρα εμφανίζουν πολύ υψηλή συγκέντρωση EVs, γεγονός που σχετίζεται με την υποδομή φόρτισης και τα κυβερνητικά κίνητρα.
- Ωστόσο, οι κεντρικές και αγροτικές περιοχές παρουσιάζουν χαμηλή υιοθέτηση, γεγονός που υποδηλώνει την ανάγκη για καλύτερες πολιτικές προώθησης και

ανάπτυξης των υποδομών.

Για την περαιτέρω ανάπτυξη της ηλεκτροκίνησης, απαιτούνται στοχευμένες παρεμβάσεις στις περιοχές με χαμηλή υιοθέτηση, καθώς και επενδύσεις στις υποδομές φόρτισης ώστε να καταστεί η ηλεκτροκίνηση πιο προσιτή και βιώσιμη σε πανεθνικό επίπεδο.