ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

«ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ»

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Εκτίμηση Αξίας Ακινήτων με τη χρήση Μηχανικής Μάθησης

Πανταζή Χριστίνα

Υποβλήθηκε ως προ-απαιτούμενο για την απόκτηση
του Μεταπτυχιακού Διπλώματος ειδίκευσης στα Πληροφοριακά Συστήματα.

Επιβλέπων καθηγητής
Ευάγγελος Καλαμπόκης

ΘΕΣΣΑΛΟΝΙΚΗ
2023

Περίληψη

Η παρούσα διπλωματική εργασία με τίτλο «Εκτίμηση Αξίας Ακινήτων με τη χρήση Μηχανικής Μάθησης» διερευνά τη δυνατότητα εφαρμογής της μηχανικής μάθησης στο μεσιτικό τομέα με σκοπό την πρόβλεψη της τιμής των ακινήτων στην πόλη της Νέας Υόρκης. Πιο συγκεκριμένα, επιχειρήθηκε η ανάλυση των δεδομένων και η δημιουργία ενός μοντέλου μηχανικής μάθησης αξιοποιώντας τα δεδομένα τα οποία αντλήθηκαν από την επίσημη κυβερνητική ιστοσελίδα της NYC (NYC website). Για την ανάλυση των ιστορικών δεδομένων για τα έτη 2021-2022, εφαρμόστηκε ο αλγόριθμος XGBoost της μηχανικής μάθησης.

Στο παρόν σύγγραμμα παρουσιάζεται η συμβολή της μηχανικής μάθησης στο μεσιτικό τομέα και η καταλληλότητα του προαναφερθέντος αλγορίθμου για την ανάλυση των δεδομένων. Έπειτα, παρουσιάζονται τα αποτελέσματα της εφαρμογής του μοντέλου πρόβλεψης αξίας ακινήτων, των οποίων το ποσοστό επιτυχίας ανέρχεται στα 66,691%. Ο απαραίτητος κώδικας γράφτηκε στη γλώσσα προγραμματισμού Python και ο αλγόριθμος που χρησιμοποιήθηκε βρίσκεται υλοποιημένος στη βιβλιοθήκη της Python, την Scikit-Learn.

Εν κατακλείδι, καθώς τα τελευταία χρόνια η ανάπτυξη της τεχνολογίας και της ψηφιοποίησης είναι ταχεία, οι προσεγγίσεις μηχανικής μάθησης θα μπορούσαν να αποβούν πολύτιμες στην περαιτέρω ακαδημαϊκή και επαγγελματική μακροοικονομική έρευνα.

Λέξεις κλειδιά: μηχανική μάθηση, μεσιτικός τομέας, ακίνητο, αλγόριθμος μηχανικής μάθησης, python, jupyter notebook, kaggle, xgboost

Περιεχόμενα

Λίστα εικόνων

1. Εισαγωγή

1.1. Το πρόβλημα και ο στόχος της διπλωματικής εργασίας

Η αγορά ακινήτων έχει σημαντικές οικονομικές επιπτώσεις στην ευημερία ενός έθνους. Τα οικιστικά ακίνητα παρέχουν καταφύγιο, εξασφαλίζουν την αποταμίευση των νοικοκυριών και αποτελούν έναν από τους κύριους μοχλούς στην οικονομία μέσω της χρηματοδότησης και των κατασκευών. Πιο συγκεκριμένα, στη Νέα Υόρκη, η οποία αποτελεί ένα από τα τρία επιτελικά κέντρα για την παγκόσμια οικονομία (μαζί με το Λονδίνο και το Τόκιο), η αγορά ακινήτων είναι μια σημαντική δύναμη στην οικονομία της πόλης, καθώς η συνολική αξία όλων των ιδιοκτησιών της πόλης ήταν 802,4 δισεκατομμύρια δολάρια το 2006. Το κτίριο της Time Warner Center, που είναι εισηγμένο στο χρηματιστήριο, είναι το ακίνητο με τη μεγαλύτερη αγοραία αξία της πόλης με 1,1 δισεκατομμύρια δολάρια το 2006. Στην πόλη της Νέας Υόρκης υπάρχουν μερικά από τα πιο πολύτιμα ακίνητα της Αμερικής αλλά και ολόκληρου του κόσμου. (*New York Travel website*).

Καθώς η τιμή των κατοικιών αυξάνεται ετησίως, η δημιουργία ενός συστήματος πρόβλεψης της τιμής των ακινήτων αναμένεται να βοηθήσει τους ανθρώπους που σχεδιάζουν να αγοράσουν ένα ακίνητο, καθώς γνωρίζοντας την τιμή στο μέλλον, τότε μπορούν να προγραμματίσουν καλά τη χρηματοδότησή τους. Επιπλέον, οι προβλέψεις τιμής ακινήτων είναι επίσης επωφελείς για τους επενδυτές ακινήτων, οι οποίοι επιθυμούν να γνωρίζουν την τάση των τιμών των κατοικιών σε μια συγκεκριμένη τοποθεσία. Ως εκ τούτου, η παροχή ακριβών προβλέψεων είναι εξίσου σημαντική για την κεντρική τράπεζα, τους επενδυτές ακινήτων και τους ιδιοκτήτες κατοικιών, όπως και για τους πολιτικούς φορείς λήψης αποφάσεων.

Σύμφωνα με τα παραδοσιακά στατιστικά μοντέλα, ο Δείκτης Τιμών Κατοικιών (ΔΤΚ) χρησιμοποιείται συνήθως για την εκτίμηση των μεταβολών στην τιμή των κατοικιών. Αποτελεί μια ποσοτική προσέγγιση που χρησιμοποιεί δεδομένα χρονοσειρών. Δεδομένου ότι η τιμή της κατοικίας συσχετίζεται στενά και με άλλους παράγοντες όπως η τοποθεσία, το μέγεθος του ακινήτου, ο όροφος, απαιτεί άλλες πληροφορίες εκτός από τον ΔΤΚ για την πρόβλεψη της μεμονωμένης τιμής κατοικίας.

Η παρούσα διπλωματική εργασία διερευνά την ικανότητα των μοντέλων μηχανικής μάθησης να προβλέπουν την αξία των ακινήτων με βάσει τα ποσοτικά και ποιοτικά χαρακτηριστικά τους. Η μηχανική μάθηση έχει προηγουμένως αυξήσει το πεδίο εφαρμογής της σε διάφορους τομείς, με μοναδικό σκοπό την αύξηση της αποτελεσματικότητας (Jung et al., 2018). Στην ιδανική περίπτωση, η μηχανική μάθηση είναι σε θέση να αναλύει δεδομένα ταχύτερα, φθηνότερα, πιο συστηματικά και να βρίσκει μη παρατηρήσιμες συσχετίσεις που το ανθρώπινο μάτι και οι παραδοσιακές στατιστικές μέθοδοι μπορεί να επιβλέπουν. Πιο συγκεκριμένα, για την παραγωγή προβλέψεων θα χρησιμοποιηθεί δείγμα δεδομένων για ακίνητα στην πόλη της Νέας Υόρκης μεταξύ των ετών 2021 και 2022.

1.2. Διάρθρωση διπλωματικής εργασίας

Το υπόλοιπο της διατριβής είναι οργανωμένο ως εξής:

Στην ενότητα 2, παρουσιάζουμε το υπόβαθρο και την πιο σχετική υπάρχουσα βιβλιογραφία. Στην ενότητα 3 περιγράφεται το σύνολο δεδομένων, καθώς και οι σχετικές προσαρμογές και παραδοχές. Στην ενότητα 4 παρουσιάζεται συνοπτικά η σχετική μεθοδολογία μηχανικής μάθησης. Η ενότητα 5 παρουσιάζει την εφαρμογή του μοντέλου και η ενότητα 6 περιγράφει τα τριμηνιαία και ετήσια αποτελέσματα από τις προβλέψεις. Η ενότητα 7 είναι διττή, καθώς εξετάζει τόσο τις επιδόσεις της μηχανικής μάθησης όσο και τις σχετικές πτυχές στην αγορά κατοικίας. Τελικά, η τελευταία ενότητα ολοκληρώνει τα ευρήματά μας

Τα υπόλοιπα κεφάλαια της εργασίας δομούνται ως εξής:

Το Κεφάλαιο 2 προσπαθεί να εντάξει τον αναγνώστη με όσο πιο συνοπτικό και κατανοητό τρόπο γίνεται, στις βασικές λειτουργίες και αρχιτεκτονικές της μηχανικής μάθησης. Στην ενότητα 2.1,

παρατίθενται οι πιο διαδεδομένοι αλγόριθμοι της μηχανικής μάθησης στη βιβλιογραφία και στη συνέχεια (ενότητα 2.2), παρουσιάζεται ο αλγόριθμος (XGBoost) που χρησιμοποιήθηκε για τη δημιουργία του παρόντος μοντέλου πρόβλεψης αξίας ακινήτων.

Στο Κεφάλαιο 3 παρουσιάζονται πληροφορίες σχετικά με την ανάπτυξη του μεσιτικού τομέα στην πόλη της Νέας Υόρκης με την πάροδο του χρόνου, καθώς και η συμβολή της στην οικονομία της πόλης.

Στο Κεφάλαιο 4 περιγράφονται όλα τα βιβλιογραφικά δεδομένα που αποκτήθηκαν για τους σκοπούς της εργασίας (ενότητα 4.1) και όλα απαραίτητα στάδια επεξεργασία που ακολουθήθηκαν σε αυτά (ενότητα 4.2). Στη συνέχεια, αναλύονται όλα τα δεδομένα που δημιουργήθηκαν κατά την εκπόνηση της εργασίας (ενότητα 4.3) και παρουσιάζεται η χρήση του αλγορίθμου με σκοπό τη διαμόρφωση του βέλτιστου μοντέλου πρόβλεψης (ενότητα 4.4).

Στο Κεφάλαιο 5 παρουσιάζονται αναλυτικά όλα τα αποτελέσματα που προκύπτουν στην παρούσα εργασία. Τα αποτελέσματα αυτά περιλαμβάνουν εικόνες, πίνακες αξιολόγησης αλλά και γραφήματα με τα οποία γίνεται ευκολότερα αντιληπτή η απόδοση του κάθε ταξινομητή στον αναγνώστη.

Τέλος, στο Κεφάλαιο 6 παρουσιάζονται συγκεντρωτικά τα συμπεράσματα που προκύπτουν από την ανάλυση των αποτελεσμάτων.

2. Γνωστικό υπόβαθρο

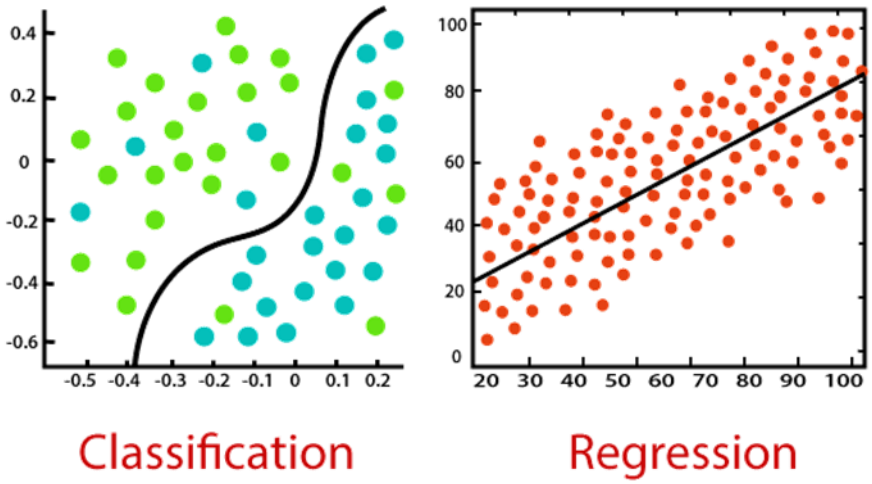
2.1. Μηχανική μάθηση

Μηχανική μάθηση είναι υπό πεδίο της επιστήμης των υπολογιστών, που αναπτύχθηκε από τη μελέτη της αναγνώρισης προτύπων και της υπολογιστικής θεωρίας μάθησης στην τεχνητή νοημοσύνη. Το 1959, ο Άρθουρ Σάμουελ ορίζει τη μηχανική μάθηση ως "Πεδίο μελέτης που δίνει στους υπολογιστές την ικανότητα να μαθαίνουν, χωρίς να έχουν ρητά προγραμματιστεί". Η μηχανική μάθηση διερευνά τη μελέτη και την κατασκευή αλγορίθμων που μπορούν να μαθαίνουν από τα δεδομένα και να κάνουν προβλέψεις σχετικά με αυτά. Τέτοιοι αλγόριθμοι λειτουργούν κατασκευάζοντας μοντέλα από πειραματικά δεδομένα, προκειμένου να κάνουν προβλέψεις βασιζόμενες στα δεδομένα ή να εξάγουν αποφάσεις που εκφράζονται ως το αποτέλεσμα.

Σύμφωνα με τον Tom M. Mitchell ένας πιο επίσημος ορισμός της μηχανικής μάθησης είναι ο εξής: «Ένα πρόγραμμα υπολογιστή λέγεται ότι μαθαίνει από εμπειρία E ως προς μια κλάση εργασιών T και ένα μέτρο επίδοσης P, αν η επίδοσή του σε εργασίες της κλάσης T, όπως αποτιμάται από το μέτρο P, βελτιώνεται με την εμπειρία E». (Βικιπαίδεια, 3 Μαΐου 2013).

Ο τομέας της Μηχανικής Μάθησης αναπτύσσει τρεις τρόπους μάθησης, ανάλογους με τους τρόπους με τους οποίους μαθαίνει ο άνθρωπος: επιβλεπόμενη μάθηση, μη επιβλεπόμενη μάθηση και ενισχυτική μάθηση. Πιο αναλυτικά:

- Επιβλεπόμενη Μάθηση (Supervised Learning) είναι η διαδικασία όπου ο αλγόριθμος κατασκευάζει μια συνάρτηση που απεικονίζει δεδομένες εισόδους (σύνολο εκπαίδευσης) σε γνωστές επιθυμητές εξόδους, με απώτερο στόχο τη γενίκευση της συνάρτησης αυτής και για εισόδους με άγνωστη έξοδο. Χρησιμοποιείται σε προβλήματα:
    - Κατηγοριοποίησης (Classification)
    - Παλινδρόμησης (Regression)



Εικόνα 1 Αναπαράσταση Κατηγοριοποίησης και Παλινδρόμησης (javaTpoint, n.d)

- Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning), όπου ο αλγόριθμος κατασκευάζει ένα μοντέλο για κάποιο σύνολο εισόδων υπό μορφή παρατηρήσεων χωρίς να γνωρίζει τις επιθυμητές εξόδους. Χρησιμοποιείται σε προβλήματα:
    - Ανάλυσης Συσχετισμών (Association Analysis)
    - Ομαδοποίησης (Clustering)

- Ενισχυτική Μάθηση (Reinforcement Learning), όπου ο αλγόριθμος μαθαίνει μια στρατηγική ενεργειών μέσα από άμεση αλληλεπίδραση με το περιβάλλον. Χρησιμοποιείται κυρίως σε προβλήματα Σχεδιασμού (Planning), όπως για παράδειγμα ο έλεγχος κίνησης ρομπότ και η βελτιστοποίηση εργασιών σε εργοστασιακούς χώρους.

2.1.1. Αλγόριθμοι μηχανικής μάθησης

Όπως προαναφέρθηκε, η μηχανική μάθηση είναι η μελέτη διάφορων αλγορίθμων, οι οποίοι έχουν το χαρακτηριστικό της βελτιστοποίησης μέσω ιστορικών δεδομένων και εμπειρίας και χρησιμοποιούνται για τη δημιουργία μοντέλων. Ένα μοντέλο μηχανικής μάθησης παρομοιάζεται με ένα λογισμικό σχεδιασμένο να αναγνωρίζει μοτίβα και συμπεριφορές, βάσει προηγούμενων εμπειριών ή δεδομένων. Με παρόμοιο τρόπο, οι αλγόριθμοι μηχανικής μάθησης εκπαιδεύονται σε μια ομάδα δεδομένων και ανακαλύπτουν μοτίβα δημιουργώντας μοντέλα, τα οποία χρησιμοποιούν αυτά τα μοτίβα και κάνουν προβλέψεις πάνω σε νέα δεδομένα.

Πιο συγκεκριμένα, στην επιβλεπόμενη μάθηση χρησιμοποιούνται τα παρακάτω μοντέλα:

1. Λογιστική Παλινδρόμηση: Το λογιστικό μοντέλο είναι ένα στατιστικό μοντέλο που μοντελοποιεί την πιθανότητα να λάβει χώρα ένα γεγονός. Χρησιμοποιείται σε προβλήματα κατηγοριοποίησης (0 ή 1).

2. Μηχανή διανυσματικής υποστήριξης. Οι μηχανές με διανύσματα υποστήριξης (δίκτυα διανυσμάτων υποστήριξης) χρησιμοποιούνται για την ανάλυση δεδομένων αναλύουν σε προβλήματα κατηγοριοποίησης. Δημιουργούν συντεταγμένες για κάθε αντικείμενα σε ένα n-διανυσματικό πεδίο και χρησιμοποιούν ώστε να ομαδοποιήσουν τα αντικείμενα με κοινά χαρακτηριστικά.

3. Naive Bayes: Ο Naive Bayes είναι ένας αλγόριθμος πιθανοτήτων που βασίζεται στην εφαρμογή του θεωρήματος bayes. Χρησιμοποιείται συνήθως για προβλήματα κατηγοριοποίησης, όπως για παράδειγμα στην αναγνώριση προσώπου, την πρόγνωση καιρού, την ιατρική διάγνωση, την ταξινόμηση ειδήσεων, την ανάλυση συναισθήματος.

4. Δέντρα αποφάσεων: Το μοντέλο δέντρου απόφασης είναι το μοντέλο υπολογισμού στο οποίο ένας αλγόριθμος αντιμετωπίζεται ως ένα δέντρο αποφάσεων, δηλαδή μια ακολουθία ερωτημάτων ή δοκιμών που γίνονται προσαρμοστικά, οπότε το αποτέλεσμα των προηγούμενων δοκιμών/ελέγχων μπορεί να επηρεάσει τη δοκιμή που πρόκειται να εκτελεστεί στη συνέχεια.

5. Γραμμική παλινδρόμηση: Η γραμμική παλινδρόμηση είναι μια προσέγγιση για τη μοντελοποίηση της σχέσης μεταξύ μιας βαθμωτής εξαρτημένης μεταβλητής Y και μίας ή περισσότερων επεξηγηματικών μεταβλητών (ή ανεξάρτητων μεταβλητών) X. Περίπτωση μιας επεξηγηματικής μεταβλητής ονομάζεται απλή γραμμική παλινδρόμηση.

6. KNN (K-Πλησιέτεροι Γείτονες): Η τεχνική k Πλησιέστεροι Γείτονες περιλαμβάνει την ομαδοποίηση των πλησιέστερων αντικειμένων σε ένα σύνολο δεδομένων και την εύρεση των πιο συχνών ή μέσων χαρακτηριστικών μεταξύ των αντικειμένων.

7. Τυχαίο δάσος: Ο τυχαίος ταξινομητής δασών μπορεί να χρησιμοποιηθεί για την επίλυση προβλημάτων παλινδρόμησης ή κατηγοριοποίησης. Ο αλγόριθμος τυχαίου δάσους αποτελείται από μια συλλογή δέντρων αποφάσεων και κάθε δέντρο αποτελείται από ένα δείγμα δεδομένων, το οποίο ονομάζεται δείγμα bootstrap, που διαμορφώνεται μέσω συνεχής αντικατάστασης δεδομένων από ένα σύνολο.

8. Boosting Algorithms: Οι αλγόριθμοι ενίσχυσης, όπως η μηχανή ενίσχυσης κλίσης, το XGBoost και το LightGBM, χρησιμοποιούν εκμάθηση συνόλου. Συνδυάζουν τις

προβλέψεις από πολλαπλούς αλγόριθμους (όπως δέντρα αποφάσεων) λαμβάνοντας παράλληλα υπόψη το σφάλμα από τον προηγούμενο αλγόριθμο.

Αντίθετα με την επιβλεπόμενη μάθηση, στη μη εποπτευόμενη μηχανική μάθηση τα μοντέλα που χρησιμοποιούνται είναι περιορισμένα, με πιο διαδεδομένα τα παρακάτω:

1. K-Means: Ο αλγόριθμος k-means έχει ως στόχο να διαχωρίσει n παρατηρήσεις σε K ομάδες, έτσι ώστε κάθε παρατήρηση να ανήκει στη συστάδα με το κοντινότερο μέσο, το οποίο χρησιμεύει ως ένα χαρακτηριστικό δείγμα της συστάδας.

2. Ιεραρχική ομαδοποίηση: Οι μέθοδοι αυτής της κατηγορίας λειτουργούν ιεραρχικά. Ξεκινούν χρησιμοποιώντας κάθε παρατήρηση σαν μια ομάδα και σε κάθε βήμα ενώνουν σε ομάδες τις παρατηρήσεις που βρίσκονται πιο κοντά.

Τέλος, υπάρχουν δύο κύριοι τύποι αλγορίθμων ενισχυτικής μάθησης:

1. Model-based αλγόριθμοι, οι οποίοι χρησιμοποιούν χρησιμοποιεί τη λειτουργία μετάβασης και ανταμοιβής για την εκτίμηση της βέλτιστης πολιτικής. Χρησιμοποιούνται σε σενάρια όπου υπάρχει πλήρη γνώση του περιβάλλοντος και του τρόπου με τον οποίο αυτό αντιδρά σε διαφορετικές ενέργειες.

2. Model-free αλγόριθμοι, οι οποίοι εντοπίζουν τη βέλτιστη πολιτική με πολύ περιορισμένη γνώση της δυναμικής του περιβάλλοντος. Δεν χρησιμοποιούν καμία λειτουργία μετάβασης/ανταμοιβής για να κρίνουν την καλύτερη στρατηγική. Η ενισχυτική μάθηση χωρίς μοντέλα θα πρέπει να εφαρμόζεται σε σενάρια που περιλαμβάνουν ελλιπείς πληροφορίες για το περιβάλλον.

2.1.2. Αλγόριθμος XGBoost

Ο XGBoost (Extreme Gradient Boosting) είναι μια βιβλιοθήκη ανοιχτού κώδικα που μπορεί να εκπαιδεύσει και να δοκιμάσει μοντέλα σε μεγάλες ποσότητες δεδομένων. Αποτελεί μία μέθοδο βελτιστοποίησης πρώτης τάξης με τη χρήση δέντρων αποφάσεων. Όπως προαναφέρθηκε, τα δένδρα αποφάσεων είναι ισχυροί ταξινομητές, οι οποίοι χρησιμοποιούν μια δομή δέντρου για να μοντελοποιήσουν τις σχέσεις μεταξύ των χαρακτηριστικών και των πιθανών αποτελεσμάτων.

Ο XGBoost ξεκίνησε αρχικά ως ερευνητικό έργο το 2014 και πλέον θεωρείται η κορυφαία βιβλιοθήκη μηχανικής μάθησης για προβλήματα παλινδρόμησης, ταξινόμησης και κατάταξης. αποτελεί μία μέθοδο βελτιστοποίησης πρώτης τάξης με τη χρήση δέντρων αποφάσεων. Όπως προαναφέρθηκε, τα δένδρα αποφάσεων είναι ισχυροί ταξινομητές, οι οποίοι χρησιμοποιούν μια δομή δέντρου για να μοντελοποιήσουν τις σχέσεις μεταξύ των χαρακτηριστικών και των πιθανών αποτελεσμάτων. Η τεχνική XGBoost έγινε ευρέως γνωστή τα τελευταία χρόνια μέσω των διαγωνισμών της Kaggle. Το Kaggle επιτρέπει στους χρήστες να βρίσκουν και να δημοσιεύουν σύνολα δεδομένων, να εξερευνούν και να δημιουργούν μοντέλα σε ένα διαδικτυακό περιβάλλον επιστήμης δεδομένων, να συνεργάζονται με άλλους data scientists και μηχανικούς μηχανικής μάθησης και να συμμετέχουν σε διαγωνισμούς για την επίλυση προκλήσεων της επιστήμης δεδομένων.

Αξίζει να σημειωθεί ότι τα μοντέλα μηχανικής μάθησης XGBoost έχουν γίνει δημοφιλή καθώς προσφέρουν τον κορυφαίο συνδυασμό απόδοσης πρόβλεψης και χρόνου επεξεργασίας σε σύγκριση με άλλους αλγόριθμους. Πιο αναλυτικά, κάποια από τα πλεονεκτήματα της χρήσης του είναι τα παρακάτω:

1. Ο XGBoost χρησιμοποιείται σε ένα ευρύ φάσμα εφαρμογών, συμπεριλαμβανομένης της επίλυσης προβλημάτων παλινδρόμησης, ταξινόμησης και κατάταξης
2. Αποτελεί μια βιβλιοθήκη που χτίστηκε από την αρχή για να είναι αποτελεσματική και ευέλικτη. Ο XGBoost έχει ενσωματωθεί με μια μεγάλη ποικιλία άλλων εργαλείων και πακέτων, όπως το scikit-learn για τους λάτρεις της Python και το caret για τους χρήστες R.

3. Υποστηρίζει τη δημιουργία δέντρων με υποδειγματοληπτημένα δεδομένα τόσο ως προς το σύνολο των παρατηρήσεων όσο και ως προς τα χαρακτηριστικά τους
4. Ένα άλλο χαρακτηριστικό του XGBoost είναι ότι χρησιμοποιεί L1 και L2 κανονικοποιήσεις, το οποίο καθιστά τη διαχείριση των δεδομένων ευκολότερη και πιο αξιόπιστη
5. Ο XGBoost χρησιμοποιεί αναγνωρισιμότητα cache, η οποία βοηθά στη μείωση της χρήσης μνήμης κατά την εκπαίδευση μοντέλων με μεγάλα σύνολα δεδομένων
6. Τέλος, προσφέρει υπολογιστικές δυνατότητες εκτός πυρήνα χρησιμοποιώντας δομές δεδομένων που βασίζονται σε δίσκο αντί για δομές δεδομένων στη μνήμη κατά τη διάρκεια της φάσης υπολογισμού.

Λόγω των παραπάνω, συμπεραίνουμε πως ο XGBoost προσφέρει τον κορυφαίο συνδυασμό απόδοσης πρόβλεψης και χρόνου επεξεργασίας, όπως προαναφέρθηκε, και θεωρείται από τους πιο αποτελεσματικούς αλγορίθμους machine learning. Έτσι, λοιπόν, στη συγκεκριμένη διπλωματική εργασία που έχει ως στόχο την πρόβλεψη της αξίας των ακινήτων με βάσει τα ποσοτικά και ποιοτικά χαρακτηριστικά τους, χρησιμοποιήθηκε η τεχνική XGBoost.2.1.3. Στοιχεία μηχανικής μάθησης

2.2. Χρήσιμα βοηθήματα

2.2.1. Η γλώσσα προγραμματισμού Python

Η Python είναι μια σύγχρονη και η πιο διαδεδομένη γλώσσα παγκοσμίως στις μέρες μας. Είναι μια αντικειμενοστραφής γλώσσα προγραμματισμού (object-oriented) υψηλού επιπέδου, με ενσωματωμένες δομές δεδομένων και δυναμικές ιδιότητες. Είναι η γλώσσα προγραμματισμού πίσω από γιγαντιαίες εφαρμογές όπως το YouTube, το Instagram, το Spotify, το Uber, το Netflix και πολλές άλλες εφαρμογές. Θεωρείται η πιο δημοφιλής γλώσσα εξαιτίας δύο βασικών χαρακτηριστικών:

1. Είναι πολύ απλή στην ανάγνωση και στη γραφή της. Η «απλότητά» της σε σχέση με άλλες γλώσσες προγραμματισμού, της δίνει προβάδισμα και την κάνει ιδανική και για προγραμματιστές που αρχίζουν τώρα.

2. Είναι μια γλώσσα προγραμματισμού με μεγάλη χρηστική αξία. Εφαρμόζεται σε πολλά πεδία, επιχειρήσεις και projects, από απλές εφαρμογές Web Development και πιο σύνθετα πεδία όπως το Data Science, το Artificial Intelligence και το Machine Learning.

Εικόνα 2 Χαρακτηριστικά της Python (SSDN Technologies, blog)

Οι κύριες βιβλιοθήκες στην Python που χρησιμοποιήθηκαν για την υλοποίηση του μοντέλου μας είναι οι εξής:

- Pandas: για τη σωστή διαχείριση του συνόλου δεδομένων

- NumPy: για τη διαχείριση μεγάλων, πολυδιάστατων πινάκων, μαζί με μια μεγάλη συλλογή μαθηματικών συναρτήσεων υψηλού επιπέδου
- Matplotlib: για την δημιουργία γραφημάτων – εικόνων
- Seaborn: ως επέκταση της παραπάνω για την καλύτερη και σωστή οπτικοποίηση των δεδομένων
- Sklearn: για την δημιουργία, την εκπαίδευση και την αξιολόγηση μοντέλων μηχανικής μάθησης και την δημιουργία προβλέψεων.

2.2.2. Το Jupyter Notebook

Το Jupyter Notebook, παλαιότερα γνωστό ως IPython Notebook, είναι ένα διαδραστικό υπολογιστικό περιβάλλον ανοιχτού κώδικα που επιτρέπει στους χρήστες να δημιουργούν, να εκτελούν και να μοιράζονται κώδικα ανάλυσης δεδομένων και μηχανικής εκμάθησης με συνεργατικό και αναπαραγώγιμο τρόπο.

Το Jupyter Notebook παρέχει ένα περιβάλλον εργασίας που βασίζεται σε πρόγραμμα περιήγησης καθιστώντας ικανή την εκτέλεσή του από σχεδόν κάθε ηλεκτρονικό υπολογιστή ανά τον κόσμο.

Τα βασικά χαρακτηριστικά των Jupyter Notebooks είναι τα εξής:

1. Σημειωματάρια: Ένα Jupyter Notebook αποτελείται από μια συλλογή κελιών που μπορούν να περιέχουν διάφορους τύπους περιεχομένου, όπως κώδικα, κείμενο, εξισώσεις και απεικονίσεις. Κάθε κελί μπορεί να εκτελεστεί ανεξάρτητα, επιτρέποντας μια διαδραστική και επαναληπτική διαδικασία κωδικοποίησης.
2. Εκτέλεση κώδικα: Τα Jupyter Notebooks υποστηρίζουν πολλές γλώσσες προγραμματισμού, συμπεριλαμβανομένων των Python, R και άλλων.
3. Υποστήριξη εμπλουτισμένου κειμένου και πολυμέσων: Τα σημειωματάρια υποστηρίζουν τη χρήση της Markdown, μιας ελαφριάς γλώσσας σήμανσης, για τη σύνταξη μορφοποιημένου κειμένου, τη δημιουργία επικεφαλίδων, λιστών, πινάκων και τη συμπερίληψη μαθηματικών εξισώσεων. Τα σημειωματάρια επιτρέπουν επίσης την ενσωμάτωση εικόνων, βίντεο και αλληλεπιδραστικών απεικονίσεων.
4. Οπτικοποίηση δεδομένων: Τα Jupyter Notebooks ενσωματώνονται καλά με δημοφιλείς βιβλιοθήκες οπτικοποίησης δεδομένων όπως Matplotlib, Seaborn και Plotly. Αυτό επιτρέπει στους χρήστες να δημιουργούν αλληλεπιδραστικά γραφήματα, γραφήματα και γραφήματα απευθείας μέσα στο περιβάλλον του σημειωματαρίου.
5. Συνεργασία και κοινή χρήση: Τα Jupyter Notebooks μπορούν εύκολα να μοιραστούν με άλλους εξάγοντάς τα σε διάφορες μορφές, όπως HTML, PDF ή απλές δέσμες ενεργειών Python. Οι χρήστες μπορούν επίσης να μοιράζονται σημειωματάρια μέσω πλατφορμών όπως το GitHub ή το Jupyter Notebook Viewer. Η συνεργατική επεξεργασία και ο έλεγχος έκδοσης είναι δυνατά μέσω υπηρεσιών όπως το JupyterHub και το JupyterLab.

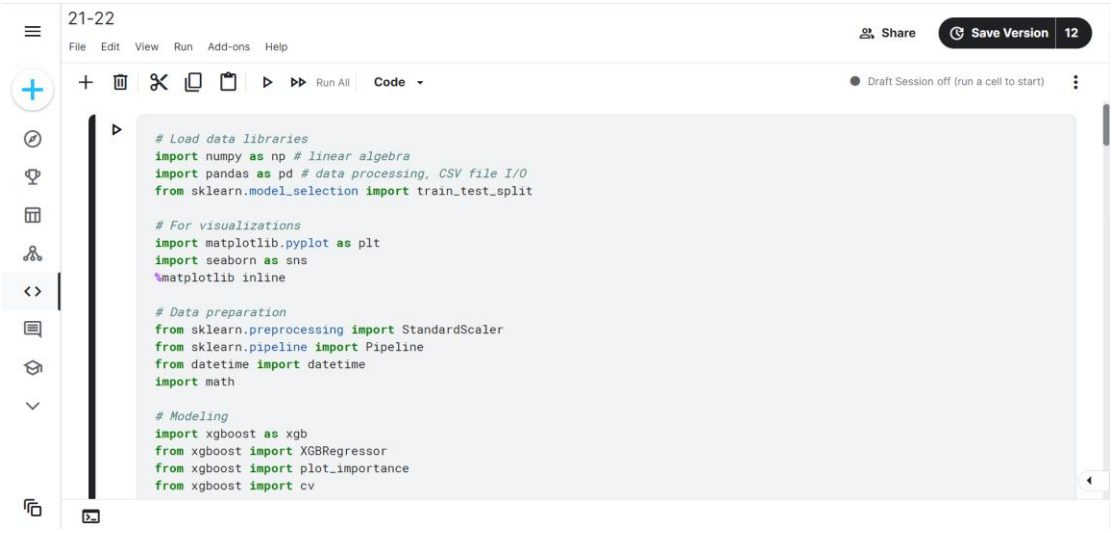
Τα Jupyter Notebooks έχουν κερδίσει δημοτικότητα στις κοινότητες επιστήμης δεδομένων και έρευνας λόγω της διαδραστικής φύσης τους, της υποστήριξης πολλαπλών γλωσσών προγραμματισμού και της δυνατότητας συνδυασμού κώδικα, τεκμηρίωσης και απεικονίσεων σε ένα ενιαίο έγγραφο. Παρέχουν ένα αποτελεσματικό περιβάλλον για την εξερεύνηση, ανάλυση και κοινή χρήση αναπαραγώγιμων ερευνητικών ροών εργασίας.

2.2.3. Η κοινότητα Kaggle

Το Kaggle είναι μια διαδικτυακή πλατφόρμα που έχει ως στόχο να προσελκύσει, να καλλιεργήσει, να εκπαιδεύσει και να προκαλέσει επιστήμονες δεδομένων από όλο τον κόσμο για την επίλυση προβλημάτων επιστήμης δεδομένων, μηχανικής μάθησης και προγνωστικής ανάλυσης.

Ιδρύθηκε το 2010 και εξαγοράστηκε από την Google το 2017 (Moyer, 2017). Το Kaggle επιτρέπει στους επιστήμονες δεδομένων και σε άλλους προγραμματιστές να συμμετέχουν στην εκτέλεση διαγωνισμών μηχανικής μάθησης, να γράφουν και να μοιράζονται κώδικα και να φιλοξενούν σύνολα δεδομένων.

Επιπρόσθετα, το Kaggle προσφέρει σημειωματάρια (notebooks), τα οποία αποτελούν ένα περιβάλλον υπολογιστικού νέφους (cloud) που επιτρέπει την ανάλυση δεδομένων. Τα Jupyter notebooks αποτελούνται από μια ακολουθία κελιών, όπου κάθε κελί μορφοποιείται είτε σε Markdown (για τη σύνταξη κειμένου) είτε στη γλώσσα προγραμματισμού Python για τη σύνταξη κώδικα.



Εικόνα 3 Jupyter Notebook στο γραφικό περιβάλλον της Kaggle (Kaggle)

Το Kaggle έχει γίνει μια δημοφιλής πλατφόρμα τόσο για αρχάριους όσο και για έμπειρους επιστήμονες δεδομένων, καθώς προσφέρει την δυνατότητα να βελτιώσουν τις δεξιότητές τους, να συνεργαστούν με συναδέλφους και να παρουσιάσουν τη δουλειά τους. Χρησιμεύει ως πολύτιμος πόρος για μάθηση, πειραματισμό και επίλυση πραγματικών προβλημάτων δεδομένων.

3. Μεσιτικός τομέας στην πόλη της Νέας Υόρκης

Η οικονομία της Νέας Υόρκης περιλαμβάνει τη μεγαλύτερη δημοτική και περιφερειακή οικονομία στις Ηνωμένες Πολιτείες. Η Νέα Υόρκη είναι ένας παγκόσμιος κόμβος επιχειρήσεων και εμπορίου, καθώς αποτελεί το κορυφαίο κέντρο τραπεζικής, χρηματοοικονομικής και επικοινωνίας στον κόσμο. Ο κλάδος των ακινήτων αποτελεί έναν από τους σημαντικότερος συντελεστές της οικονομίας στη Νέα Υόρκη. Σύμφωνα με μελέτες, η συνολική αξία όλων των ακινήτων της Νέας Υόρκης εκτιμήθηκε σε 1,072 τρισεκατομμύρια δολάρια για το οικονομικό έτος 2017.

Εικόνα 4 Η πόλη της Νέας Υόρκης (Mononews Οκτώβριος 2016)

Η αγορά ακινήτων της Νέας Υόρκης αποτελείται από μια τεράστια θάλασσα από πολυτελή διαμερίσματα, διαμερίσματα, σπίτια και αρχοντικά που βρίσκονται σε όλη τη Νέα Υόρκη και τα προάστια της. Μία από τις πιο πυκνοκατοικημένες και ελκυστικές πόλεις στον κόσμο προσφέρει τόσες πολλές επιλογές για αγοραστές και επενδυτές ακινήτων.

Η Νέα Υόρκη έχει ιστορικό ως μία από τις καλύτερες μακροπρόθεσμες επενδύσεις σε ακίνητα στις ΗΠΑ. Η αγορά ακινήτων της Νέας Υόρκης ανθεί χρόνο με το χρόνο. Σύμφωνα με μελέτες, οι τιμές των κατοικιών στη Νέα Υόρκη σχεδόν διπλασιάστηκαν την τελευταία δεκαετία. Με την προσφορά και τη ζήτηση να συνεχίζουν να ευνοούν τους πωλητές, οι τιμές συνεχίζουν να αυξάνονται χρόνο με το χρόνο. Παρ'όλα αυτά, υπάρχουν διάφοροι λόγοι, οι οποίοι μπορεί να οδηγήσουν τον πωλητή στο να θέσει την τιμή του ακινήτου σε χαμηλότερη τιμή όπως είναι η ανάγκη ρευστότητας, προσωπικοί λόγοι κ.α..

Έτσι λοιπόν, η δημιουργία ενός συστήματος πρόβλεψης της αξίας των ακινήτων με τη χρήση μηχανικής μάθησης θα βοηθήσει τόσο τους ιδιώτες όσο και τους επενδυτές ακινήτων στην πόλη της Νέας Υόρκης να αναγνωρίσουν ευκαιρίες και να επωφεληθούν από την αγορά ενός ακινήτου ακόμη και σε περιόδους οικονομικής ύφεσης.

4. Εφαρμογές Μηχανικής Μάθησης στο μεσιτικό τομέα

Η Μηχανική Μάθηση βρίσκει μεγάλη εφαρμογή σε πεδία που αφορούν την καθημερινή μας ζωή με αποτέλεσμα να βελτιώνεται η ποιότητα των υπηρεσιών στον κάθε τομέα. Έτσι, λοιπόν, η χρήση του δε θα μπορούσε να απουσιάζει από το μεσιτικό τομέα, καθώς η ακριβής εκτίμηση της αξίας των ακινήτων είναι ένα σημαντικό πρόβλημα για πολλά ενδιαφερόμενα μέρη, συμπεριλαμβανομένων ιδιοκτητών κατοικιών, αγοραστών κατοικιών, πρακτόρων, πιστωτών και επενδυτών.

Προηγούμενες μελέτες σχετικά με την πρόβλεψη των τιμών των κατοικιών με τεχνικές μηχανικής μάθησης έχουν συμβάλει σημαντικά στον τομέα. Αρκετοί ερευνητές έχουν διερευνήσει παρόμοια θέματα, εστιάζοντας σε διαφορετικές περιοχές και χρησιμοποιώντας διάφορα σύνολα δεδομένων και μεθοδολογίες. Αυτή η ενότητα παρουσιάζει μια ανασκόπηση ορισμένων αξιοσημείωτων μελετών που έχουν διερευνήσει την πρόβλεψη των τιμών των κατοικιών χρησιμοποιώντας μηχανική μάθηση σε διαφορετικά πλαίσια. Ενδεικτικά αναφέρονται οι εξής:

1. "Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics " από τους Antipov, Evgeny και Pokryshevskaya, Elena (2010): Αυτή η μελέτη επικεντρώθηκε στη χρήση της τεχνικής τυχαίου δάσους (RF) με μηχανική μάθηση για την εκτίμηση ακινήτων που βρίσκονται στην Αγία Πετρούπολη (Ρωσία). Τα αποτελέσματα έδειξαν ότι αυτή η τεχνική είναι εξαιρετικά αποτελεσματική και παράγει αξιόπιστα αποτελέσματα ακόμη και με ελλιπή δεδομένα, ακραίες τιμές, κατηγορικές μεταβλητές και υψηλή ετεροσκεδαστικότητα στα δεδομένα.
2. "Prediction of House Price Using XGBoost Regression Algorithm" από τους J.Avanijaa, Gurram Sunitha b, K.Reddy Madhavi c, Padmavathi Korad και R.Hitesh Sai Vittal (2021): Αυτή η έρευνα χρησιμοποίησε την τεχνική του XGBoost για να προβλέψει τις τιμές των κατοικιών στην περιοχή της Έιμς, στην Αϊόβα. Με βάσει τα αποτελέσματα, οι συγγραφείς κατέληξαν στο συμπέρασμα ότι ο XGBoost είναι ένας από τους καλύτερους αλγόριθμους παλινδρόμησης, καθώς βοηθά στην ικανοποίηση των αναγκών των πελατών αυξάνοντας την ακρίβεια της επιλογής των ακινήτων και μειώνοντας τον κίνδυνο για τους πελάτες να επενδύσουν σε ακίνητα.
3. "Housing Price Prediction Using Machine Learning Algorithms in COVID-19 Times" από τους Raul-Tomas Mora-Garcia, Maria-Francisca Cespedes-Lopez και Raul Perez-Sanchez (2022): Αυτή η μελέτη χρησιμοποίησε διάφορους αλγόριθμους μηχανικής μάθησης, συμπεριλαμβανομένου του τυχαίου δάσους και της ενίσχυσης κλίσης, για να προβλέψει τις τιμές των κατοικιών, και να ποσοτικοποιήσει τον αντίκτυπο της πανδημίας COVID-19 στις τιμές των κατοικιών στην πόλη της Ισπανίας, Αλικάντε. Τα αποτελέσματα της έρευνας έδειξαν ότι οι αλγόριθμοι μηχανικής μάθησης παρουσιάζουν μεγαλύτερη αποδοτικότητα σε σχέση με τα παραδοσιακά γραμμικά μοντέλα, επειδή είναι καλύτερα προσαρμοσμένοι στις μη γραμμικότητες σύνθετων δεδομένων, όπως τα δεδομένα της αγοράς ακινήτων.

Οι προαναφερθείσες μελέτες καταδεικνύουν την ευρεία εφαρμογή τεχνικών μηχανικής μάθησης στην πρόβλεψη των τιμών των κατοικιών σε διάφορες χώρες και περιοχές. Η παρούσα διπλωματική εργασία στοχεύει στην ενίσχυση της υπάρχουσας βιβλιογραφίας με τη διεξαγωγή μελέτης με την τεχνική μηχανικής μάθησης XGBoost για την πρόβλεψη των τιμών των κατοικιών στην πόλη της Νέας Υόρκης.

5. Πηγή Δεδομένων

Η παρούσα εργασία επικεντρώνεται στην συλλογή και ανάλυση πραγματικών δεδομένων που αφορούν κύρια χαρακτηριστικά κατοικιών και τα οποία αναφέρονται στο ευρύτερο γεωγραφικό αστικό τμήμα της πόλης της Νέας Υόρκης για τα έτη 2021-2022. Τα δεδομένα που χρησιμοποιήθηκαν αντλήθηκαν από την επίσημη κυβερνητική ιστοσελίδα της NYC (*NYC website*). Αυτός ο ιστότοπος χρησιμεύει ως ένας ολοκληρωμένος πόρος για τους κατοίκους, τις επιχειρήσεις και τους επισκέπτες για πρόσβαση σε πληροφορίες και υπηρεσίες που σχετίζονται με διάφορες πτυχές της ζωής στη Νέα Υόρκη. Στη NYC ιστοσελίδα, το Υπουργείο Οικονομικών παραθέτει κάθε χρόνο τα ακίνητα που πωλήθηκαν το τελευταίο δωδεκάμηνο στη Νέα Υόρκη για τις φορολογικές κατηγορίες 1, 2 και 4. Αυτά τα αρχεία περιλαμβάνουν:

- τη γειτονιά
- τον τύπο του κτιρίου
- τα τετραγωνικά μέτρα του ακινήτου, κ.α.

Για να γίνει ευκολότερη η κατανόηση του μοντέλου και των δεδομένων που αναλύθηκαν παρακάτω παρουσιάζονται αναλυτικά τα χαρακτηριστικά των ακινήτων που αποτυπώνονται στο δείγμα μας:

- Borough: Το όνομα του δήμου στον οποίο βρίσκεται το ακίνητο
- Neighborhood: Το όνομα της γειτονιάς κατά τη διάρκεια της αποτίμησης των ακινήτων. Ενδέχεται να υπάρχουν μικρές διαφορές στις συνοριακές γραμμές γειτονιάς και ορισμένες υπογειτονίες να μην περιλαμβάνονται.
- Building Class Category: Αυτό το πεδίο στο τύπο κατοικίας. Πιο αναλυτικά, στα δεδομένα υπάρχουν οι παρακάτω κατηγορίες:
    A. Μονοκατοικία
    B. Μεζονέτα
    C. Διαμέρισμα χωρίς σκάλες
    D. Διαμέρισμα με σκάλες
    E. Αποθήκη
    F. Εργοστάσια και Βιομηχανικά κτίρια
    G. Γκαράζ
    H. Ξενοδοχεία
    I. Νοσοκομεία και εγκαταστάσεις υγείας
    J. Θέατρα
    K. Κτίρια καταστημάτων
    L. Lofts
    M. Θρησκευτικές εγκαταστάσεις
    N. Άσυλα
    O. Κτίρια γραφείων

P. Εσωτερικές εγκαταστάσεις Δημοσίων Συγκεντρώσεων και Πολιτισμού

Q. Υπαίθριες εγκαταστάσεις αναψυχής

R. Διαμέρισμα σε συγκρότημα

S. Κατοικίες προς μίσθωση

T. Εγκαταστάσεις μεταφορών

U. Κτίρια δημόσιων υπηρεσιών

V. Οικόπεδα

W. Εκπαιδευτικές εγκαταστάσεις

Y. Κυβερνητικά/Δημοτικά κτίρια

Z. Άλλες κατηγορίες ακινήτων

- Block: Η μεταβλητή αυτή αναφέρεται στις διευθύνσεις που περιγράφουν τη θέση του δρόμου ενός ακινήτου, το οικοδομικό τετράγωνο και το οικόπεδο.
- Lot: Μια φορολογική παρτίδα είναι μια υποδιαίρεση ενός φορολογικού μπλοκ και αντιπροσωπεύει τη μοναδική τοποθεσία του ακινήτου.
- Address: Η οδός του ακινήτου όπως αναγράφεται στο Αρχείο Πωλήσεων. Ο αριθμός διαμερίσματος περιλαμβάνεται στο πεδίο διεύθυνσης.
- Zip code: Ταχυδρομικώς Κώδικας
- Residential units: Ο αριθμός των οικιστικών διαμερισμάτων στο ακίνητο
- Commercial units: Ο αριθμός των εμπορικές διαμερισμάτων στο ακίνητο
- Total units: Σύνολο διαμερισμάτων στο ακίνητο
- Land square feet: Το εμβαδόν του ακινήτου σε τετραγωνικά πόδια
- Gross square feet: Το συνολικό εμβαδόν όλων των ορόφων ενός κτιρίου όπως μετράται από τις εξωτερικές επιφάνειες των εξωτερικών τοίχων του κτιρίου, συμπεριλαμβανομένου του εμβαδού και του χώρου εντός κτιρίου
- Year built: Το έτος που κατασκευάστηκε το ακίνητο
- Building Class at time of sale: Ο τύπος της κατοικίας (όπως αναφέρθηκαν παραπάνω) κατά τη στιγμή της πώλησης
- Tax Class at time of sale: Η φορολογική κατηγορία του ακινήτου (Κλάσεις 1, 2, 3 και 4), με βάση τη χρήση του:

  Κλάση 1: Ακίνητα που είναι για οικιστική χρήση

  Κλάση 2: Ακίνητα που είναι κατά κύριο λόγο οικιστικά, όπως συνεταιρισμοί και συγκυριαρχίες.

  Κλάση 3: Ακίνητα με εξοπλισμό που ανήκει σε εταιρεία φυσικού αερίου, τηλεφώνου ή ηλεκτρικής ενέργειας.

  Κλάση 4: Όλα τα άλλα ακίνητα που δεν περιλαμβάνονται στις κατηγορίες 1,2 και 3, όπως γραφεία, εργοστάσια, αποθήκες, κτίρια γκαράζ κ.λπ.

- Sale price: Τιμή πώλησης
- Sale Date: Ημερομηνία πώλησης

6. Προ-επεξεργασία και ανάλυση δεδομένων στην Python

6.1. Προετοιμασία δεδομένων

Προτού προχωρήσουμε στην δημιουργία του μοντέλου μηχανικής μάθησης πρέπει να επεξεργαστούμε τα δεδομένα μας ώστε να φτάσουν στην κατάλληλη μορφή για την στατιστική ανάλυση. Τα βασικά Στάδια κατά την προετοιμασία των δεδομένων είναι τα εξής:

1. Καθαρισμός των Δεδομένων (cleaning)
   a. Ελλιπείς τιμές (missing values)
   b. Μη σωστά δεδομένα – ακραίες τιμές (outliers)
2. Μετασχηματισμός των δεδομένων (transformation)
   a. Μετατροπή δεδομένων σε άλλους τύπους
   b. Μείωση αριθμού δεδομένων ή πεδίων (δηλ. μείωση στηλών και δειγματοληψία

1.α) Τα αίτια για την ύπαρξη ελλιπών τιμών είναι συνήθως τα ακόλουθα:

- Λάθος κατά την εισαγωγή των δεδομένων
- Ευαίσθητα δεδομένα, τα οποία δεν πρέπει να εμφανιστούν

Και οι τρόποι διαχείρισης των τιμών αυτών:

- Διαγραφή των εγγραφών με ελλιπείς στοιχεία
- Συμπλήρωση των τιμών μέσω υπολογισμού του μέσου όρου της μεταβλητή - χαρακτηριστικού. Σε περιπτώσεις που υπάρχουν κατηγορηματικές μεταβλητές και δε γνωρίζουμε την τιμή, αντικαθιστούμε με 0

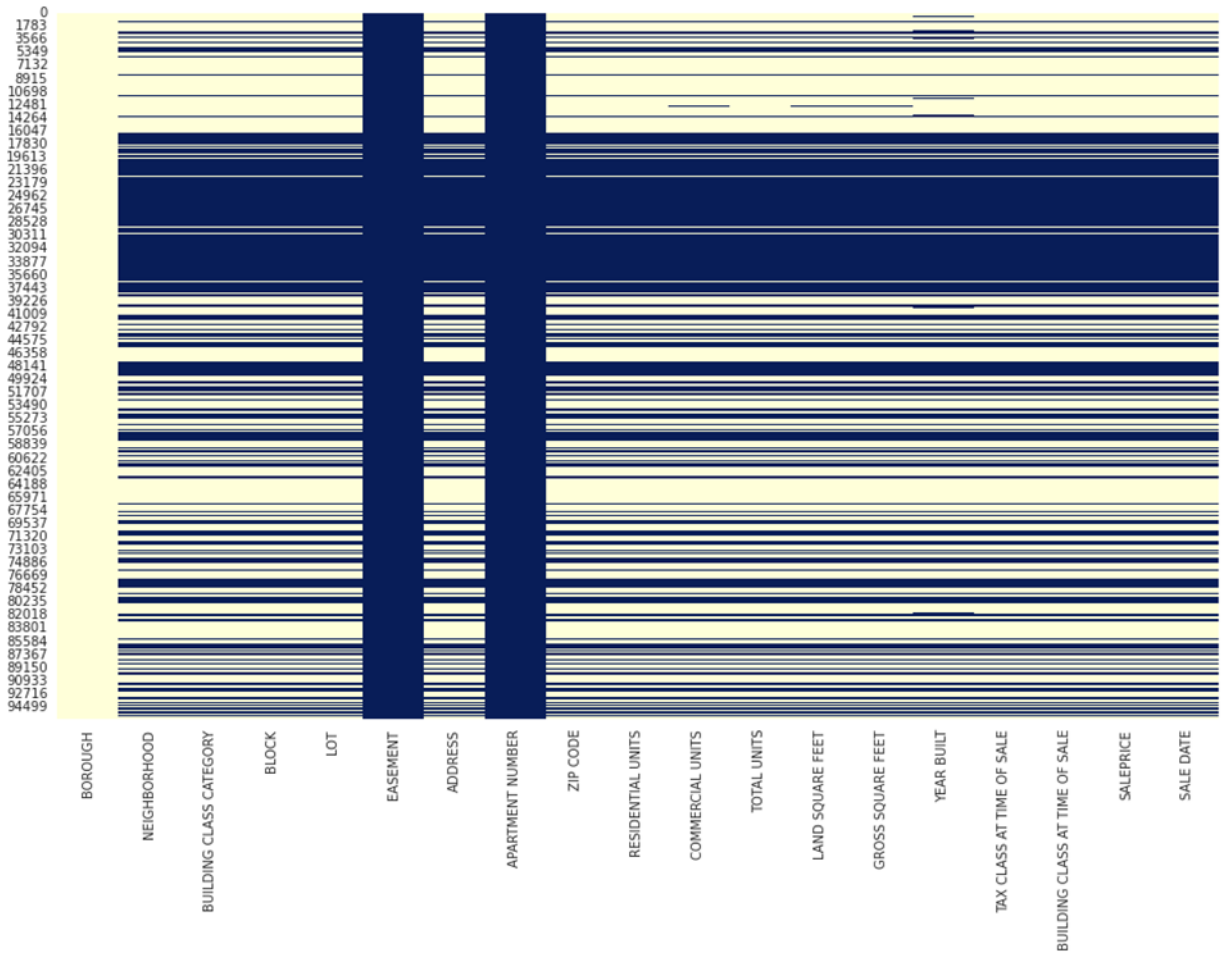
1.β) Ακραίες τιμές ή ακρότατα ονομάζονται οι μη ομαλές τιμές, οι οποίες δεν ακολουθούν την κατανομή των άλλων τιμών. Συνήθως, τα ακρότατα εμφανίζονται σε περιπτώσεις λάθους εισαγωγής δεδομένων ή λόγω κάποιου περίεργου γεγονότος το οποίο χρήζει περαιτέρω ανάλυσης (π.χ. στις περιπτώσεις οικονομικής απάτης).

2.α) Μέσω των μετασχηματισμών επιτυγχάνεται η μετατροπή των δεδομένων από μια μορφή σε μια άλλη. Για παράδειγμα μπορούμε να μετατρέψουμε ένα δεδομένα που έχει τύπο ακέραιο (integer), σε τύπο δεκαδικό (float).

2.β) Η μείωση των δεδομένων επιτυγχάνεται τη δειγματοληψία. Η μέθοδος αυτή χρησιμοποιείται όταν υπάρχει πολύ μεγάλο σύνολο δεδομένων, καθώς παίρνοντας ένα αντιπροσωπευτικό δείγμα οδηγούμαστε στα ίδια αποτελέσματα σε σχέση με την ανάλυση όλου του πληθυσμού. Αντίθετα, η μείωση των πεδίων επιτυγχάνεται με τη δημιουργία νέων χαρακτηριστικών, για παράδειγμα ενοποίηση δύο στηλών, ή με την επιλογή συγκεκριμένων πεδίων που συμβάλλουν στην προβλεπτική ικανότητα του μοντέλου.

Για την προετοιμασία των δεδομένων μας, πραγματοποιήσαμε τα παρακάτω βήματα:

1. Σκοπός του μοντέλου μας είναι να προβλέψει την τιμή πώλησης ενός ακινήτου με βάσει τα χαρακτηριστικά του κατά την περίοδο της πώλησής του. Έτσι, λοιπόν, αφαιρέσαμε από τα δεδομένα μας τις στήλες (TAX CLASS AT PRESENT και BUILDING CLASS AT PRESENT) που παρουσιάζουν τωρινές πληροφορίες, δηλαδή μετά την πώληση του ακινήτου.
2. Καθώς έχουμε εισάγει τα δεδομένα μας, ελέγχουμε αν υπάρχουν NA values, δηλαδή τιμές που λείπουν. Βρίσκουμε, λοιπόν, ότι υπάρχει ο παρακάτω αριθμός ελλιπών τιμών σε κάθε μία από τις μεταβλητές μας:



BOROUGH	0
NEIGHBORHOOD	45774
BUILDING CLASS CATEGORY	45774
BLOCK	45774
LOT	45774
EASEMENT	96262
ADDRESS	45774
APARTMENT NUMBER	96205
ZIP CODE	45781
RESIDENTIAL UNITS	45854
COMMERCIAL UNITS	45859
TOTAL UNITS	45790
LAND SQUARE FEET	45923
GROSS SQUARE FEET	45923
YEAR BUILT	47494
TAX CLASS AT TIME OF SALE	45774
BUILDING CLASS AT TIME OF SALE	45774
SALEPRICE	45774
SALE DATE	45774

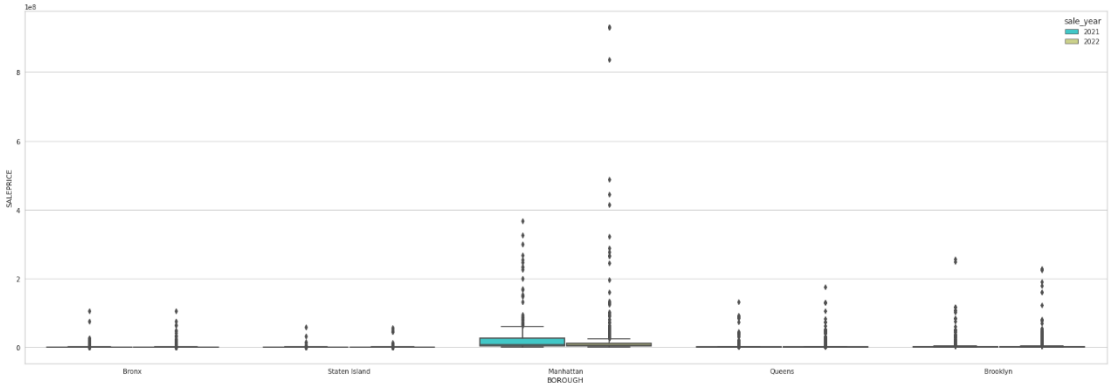
Εικόνα 5 Ελλιπείς τιμές

Εικόνα 6 Κατανομή ελλιπών τιμών

Στην παραπάνω εικόνα παρατηρούμε την κατανομή αυτών των ιδιαίτερων τιμών.

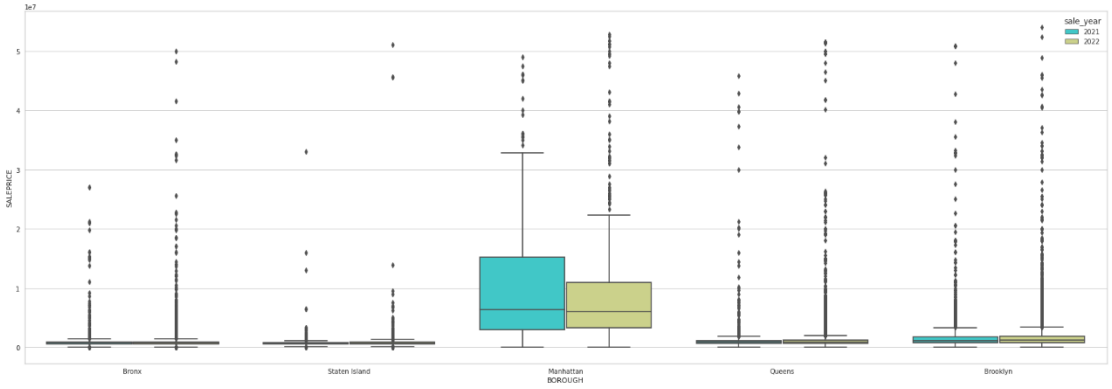
Στο παράδειγμά μας, όλες οι μεταβλητές με ποσοστό ελλειπουσών τιμών άνω του 99% αφαιρέθηκαν, και έπειτα όλες οι παρατηρήσεις με έστω και μια ελλείπουσα τιμή αφαιρέθηκαν.

3. Το επόμενο βήμα κατά την προετοιμασία των δεδομένων είναι να μετατρέψουμε τον τύπο των μεταβλητών σε όσες δεν ανταποκρίνεται η πραγματική τους σημασία στον σωστό τύπο. Συγκεκριμένα, όσες μεταβλητές περιέχουν μόνο ακέραιους αριθμούς, μετατρέπονται σε integer και όσες μεταβλητές αποτελούνται από ακολουθίες χαρακτήρων, μετατρέπονται σε string ως είθισται στη γλώσσα Python.
4. Στη συνέχεια, ελέγχουμε αν υπάρχουν εγγραφές όπου η τιμή ακινήτου, τα μεικτά και καθαρά τετραγωνικά μέτρα του ακινήτου και η ημερομηνία κατασκευής του είναι ίσα με το 0. Διαπιστώνουμε, πως υπάρχουν 64.732 εγγραφές όπου η τιμή ακινήτου είναι ίση με το 0, 48.075 εγγραφές για τα μεικτά τετραγωνικά μέτρα, 45.926 για τα καθαρά τετραγωνικά μέτρα και 47.494 για το έτος κατασκευής. Επιλέξαμε να διαγράψουμε τις ελλείπουσες τιμές και τις αντίστοιχες παρατηρήσεις, καθώς μετά από ανάλυση διαπιστώσαμε πως οι εγγραφές αυτές επηρεάζουν την προβλεπτική ικανότητα του μοντέλου μας.
5. Έπειτα, ελέγχουμε αν υπάρχουν outliers, δηλαδή ακραίες τιμές στα δεδομένα μας. Βρίσκουμε, λοιπόν, ότι υπάρχει η παρακάτω κατανομή τιμών στην μεταβλητή που θέλουμε να προβλέψουμε (τιμή ακινήτου - SALEPRICE):



Εικόνα 7 Ακραίες τιμές στη μεταβλητή SALEPRICE

Όπως μπορούμε να καταλάβουμε και από το διάγραμμα, κατά προσέγγιση, οι ακραίες τιμές παρουσιάζονται μετά το 5. Στο συγκεκριμένο πρόβλημα, τα ακρότατα δεν εμφανίζονται λόγω λάθους εισαγωγής δεδομένων ή λόγω κάποιου περίεργου γεγονότος. Στο μεσιτικό τομέα, οι τιμές των ακινήτων διαμορφώνονται ανάλογα τα ποιοτικά και ποσοτικά χαρακτηριστικά. Έτσι, λοιπόν, στο παράδειγμά μας, οι ακραίες τιμές είναι τα ακίνητα «πολυτελείας». Για να αποκλείσουμε, λοιπόν, τον αριθμό των ακραίων τιμών στα δεδομένα μας, αφαιρούμε τις εγγραφές όπου η τιμή του ακινήτου είναι μεγαλύτερη ή ίση του 55.000.000 (137 εγγραφές). Μετά τον μετασχηματισμό των δεδομένων, η κατανομή τιμών στην μεταβλητή της τιμής διαμορφώνεται ως εξής:



Εικόνα 8 Κατανομή τιμών μετά τον μετασχηματισμό - SALEPRICE

Βλέποντας το δεύτερο διάγραμμα, διαπιστώνουμε πως οι τιμές των ακινήτων στις πέντε περιφέρειες της Νέας Υόρκης (Bronx, Staten Island, Manhattan, Brooklyn και Queens) παρουσιάζουν καλύτερη κατανομή, μετά το μετασχηματισμό.

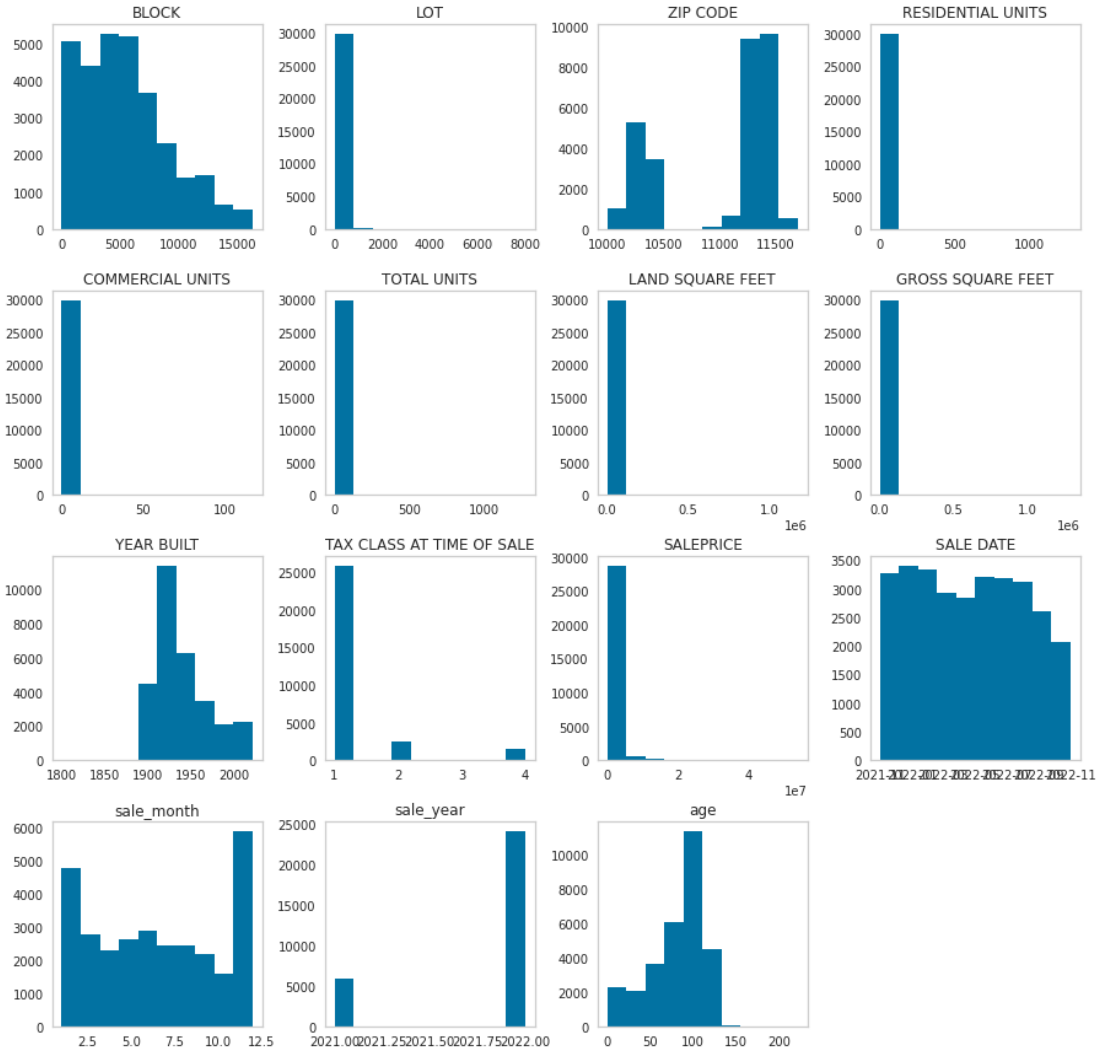
Με την εκτέλεση των προαναφερθέντων βημάτων, τα δεδομένα μας είναι πλέον έτοιμα για ανάλυση.

6.2. Στατιστική Ανάλυση Δεδομένων

Για καλύτερη κατανόηση των δεδομένων μας, προχωρήσαμε στην γραφική απεικόνισή τους. Παρακάτω παρουσιάζεται η σχέση μεταξύ των διάφορων μεταβλητών στο πρόβλημά μας, καθώς και οι συσχετίσεις τους.

1. **Ιστογράμματα (Histograms)**

Τα ιστογράμματα για τις αριθμητικές μεταβλητές του προβλήματός μας είναι τα εξής:



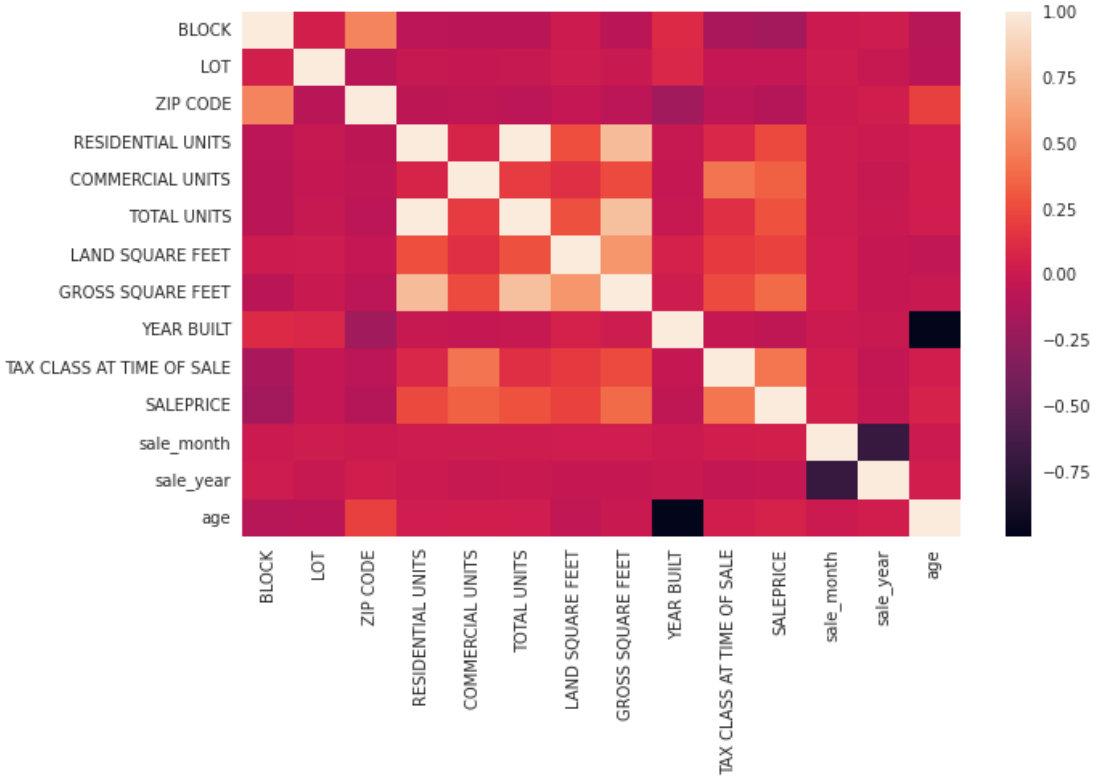
Εικόνα 9 Ιστογράμματα

Ένα ιστόγραμμα είναι ένας τύπος γραφήματος που έχει ευρείες εφαρμογές στα στατιστικά στοιχεία. Τα ιστογράμματα παρέχουν μια οπτική ερμηνεία των αριθμητικών δεδομένων υποδεικνύοντας τον αριθμό των σημείων δεδομένων που βρίσκονται μέσα σε ένα εύρος τιμών. Σχηματίζεται από παρακείμενα ορθογώνια. Η επιφάνεια κάθε ορθογωνίου είναι μέτρο της συχνότητας εμφάνισης της

συγκεκριμένης περιοχής τιμών ενώ το ύψος του ισούται με το λόγο της συχνότητας προς το εύρος των τιμών που αντιπροσωπεύει το ορθογώνιο. Πρόκειται για τη συνηθέστερη επιλογή γραφικής παράστασης συνεχών μεταβλητών. (Wikipedia, 2012).

2. Χάρτης Θερμότητας (Heatmap)

Το παρακάτω διάγραμμα, αποτελεί τον χάρτη για την αναπαράσταση των συσχετίσεων μεταξύ των διάφορων μεταβλητών στο πρόβλημά μας.



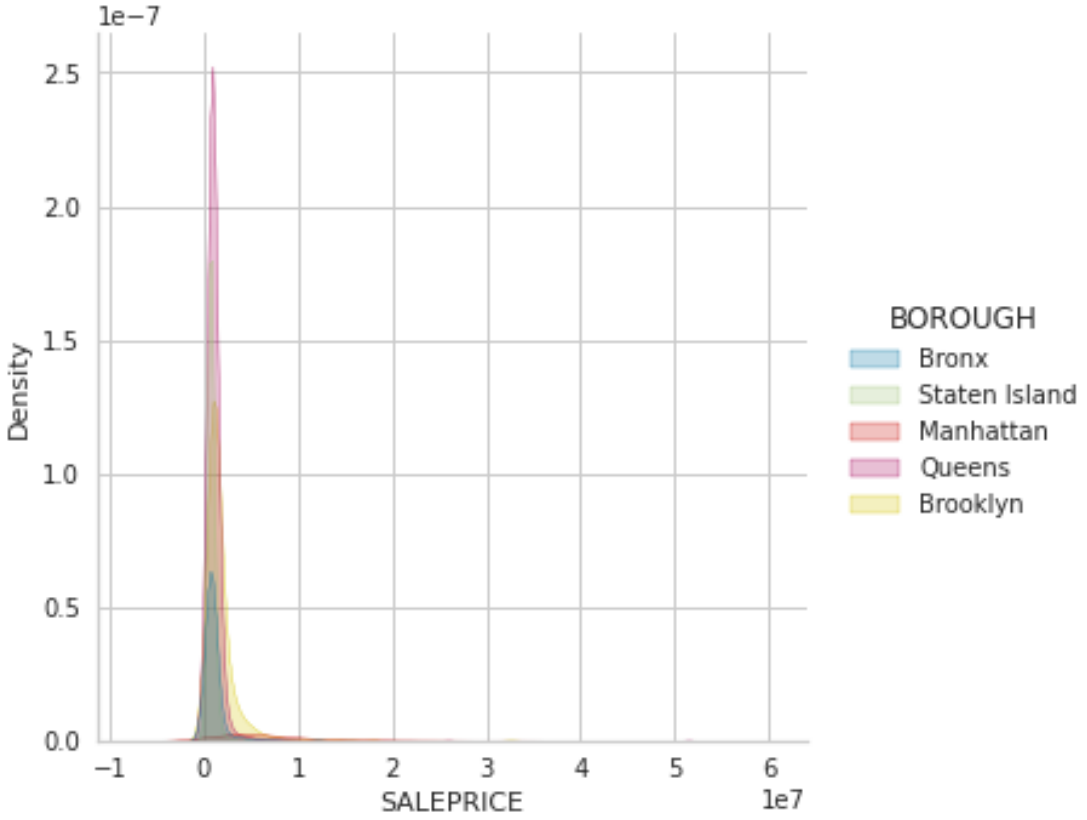
Εικόνα 10 Χάρτης Θερμότητας

Ο Heatmap αποτελεί έναν πολύ σημαντικό πίνακα ή χάρτη ο οποίος χρησιμοποιείται για να υπολογίσουμε σε ποσοστά τις συσχετίσεις μεταξύ διαφορετικών μεταβλητών και τις αναπαριστά με γραφικό και ζωντανό τρόπο (Wikipedia, 2022). Το χρώμα του κελιού διαμορφώνεται ανάλογα με την εξάρτηση-σχέση μεταξύ των εμπλεκόμενων μεταβλητών. Όπως παρατηρούμε και στον χάρτη θερμότητας του παραδείγματός μας, τα χρώματα των κελιών ξεκινούν από μπεζ, το οποίο δείχνει ότι υπάρχει θετική συσχέτιση μεταξύ των δύο χαρακτηριστικών. Αν η συσχέτιση είναι ίση με το 0, αυτό απεικονίζεται με κόκκινο χρώμα στο κελί και σημαίνει πως υπάρχει καμία εξάρτηση μεταξύ των δύο μεταβλητών. Τέλος, συνεχίζοντας προς τα κάτω, παρατηρούμε πως με μαύρο κελί παρουσιάζεται το αρνητικό αποτέλεσμα, το οποίο εκφράζει την αρνητική συσχέτιση των δύο χαρακτηριστικών.

Αυτή η απεικόνιση καθιστά τους χάρτες θερμότητας συσχέτισης ιδανικούς για ανάλυση δεδομένων, καθώς καθιστά τα μοτίβα εύκολα αναγνώσιμα και επισημαίνει τις διαφορές και τις διακυμάνσεις στα ίδια δεδομένα.

3. Γράφημα Πυκνότητας (Density plot)

Στο παρακάτω γράφημα έχουμε ένα συγκριτικό γράφημα πυκνότητας των τιμών των ακινήτων (Saleprice) ανάλογα με την περιοχή της Νέας Υόρκης στην οποία βρίσκονται (Borough):



Εικόνα 11 Γράφημα Πυκνότητας

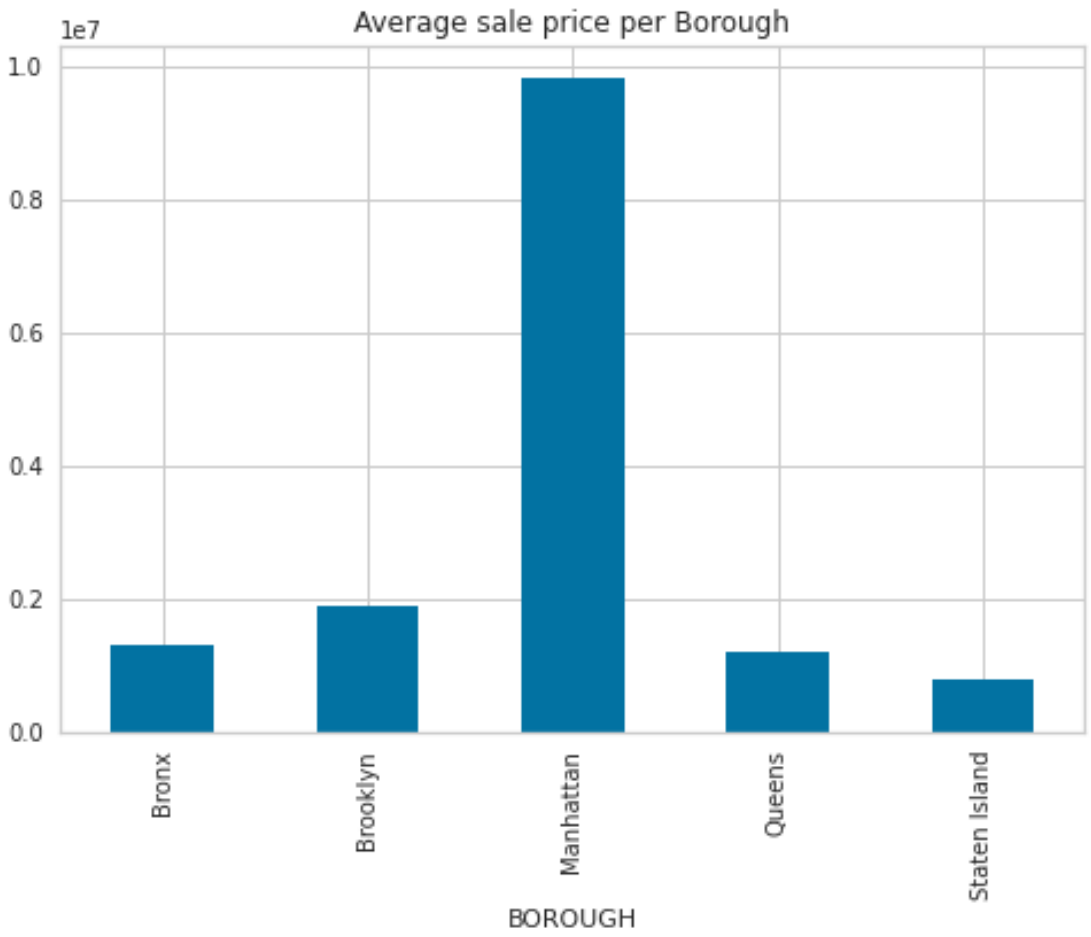
Το γράφημα πυκνότητας μπορεί να θεωρηθεί ως προέκταση του ιστογράμματος. Αποτελεί ένα στατιστικό γράφημα που μας παρέχει πληροφορίες σχετικά με την κατανομή μιας αριθμητικής μεταβλητής πάνω σε ένα (άλλο) συνεχές αριθμητικό διάστημα (π.χ. χρόνος). Απεικονίζει την κατανομή των δεδομένων σε μια δεδομένη περίοδο και οι κορυφές δείχνουν πού συγκεντρώνονται οι τιμές.

4. Ραβδογράμματα

Ένα γράφημα ράβδων ή ένα ραβδόγραμμα είναι ένα γράφημα που αντιπροσωπεύει την κατηγορία δεδομένων με ορθογώνιες ράβδους με μήκη και ύψη που είναι ανάλογες με τις τιμές που αντιπροσωπεύουν. Ένας από τους άξονες του γραφήματος αντιπροσωπεύει τις συγκεκριμένες κατηγορίες που συγκρίνονται, ενώ ο άλλος άξονας αντιπροσωπεύει τις μετρούμενες τιμές που αντιστοιχούν σε αυτές τις κατηγορίες.

Εικόνα 12

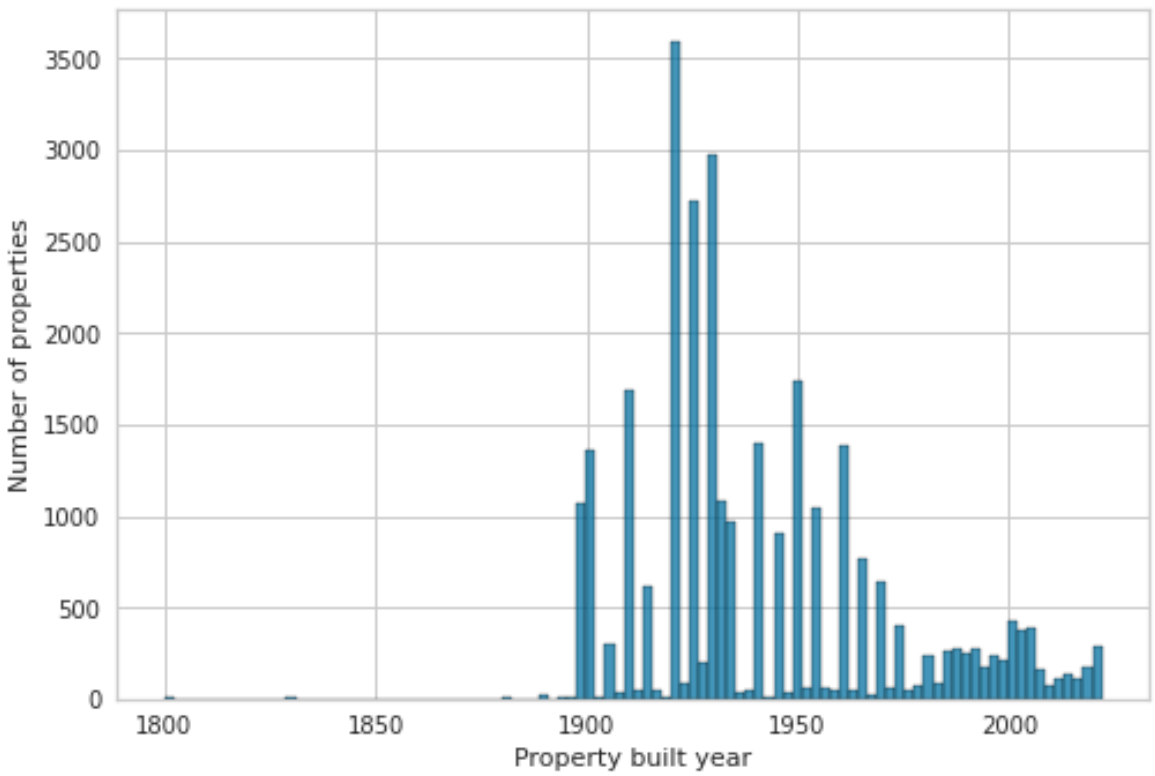
Στο παραπάνω ραβδόγραμμα απεικονίζονται οι 5 περιοχές της Νέας Υόρκης που περιλαμβάνονται στα δεδομένα μας ανά αριθμό πωλήσεων ακινήτων. Όπως καταλαβαίνουμε, τα περισσότερα ακίνητα κατά τη διάρκεια των ετών 2021 και 2022, πουλήθηκαν στην περιοχή του Queens, με περισσότερες από 11.000 πωλήσεις. Στη συνέχεια, η δεύτερη πιο δημοφιλής περιοχή για αγορά ακινήτου ήταν το Brooklyn και τελευταίο με τις λιγότερες πωλήσεις για τα έτη 2021-2022 ήταν η περιοχή του Manhattan.



Εικόνα 13

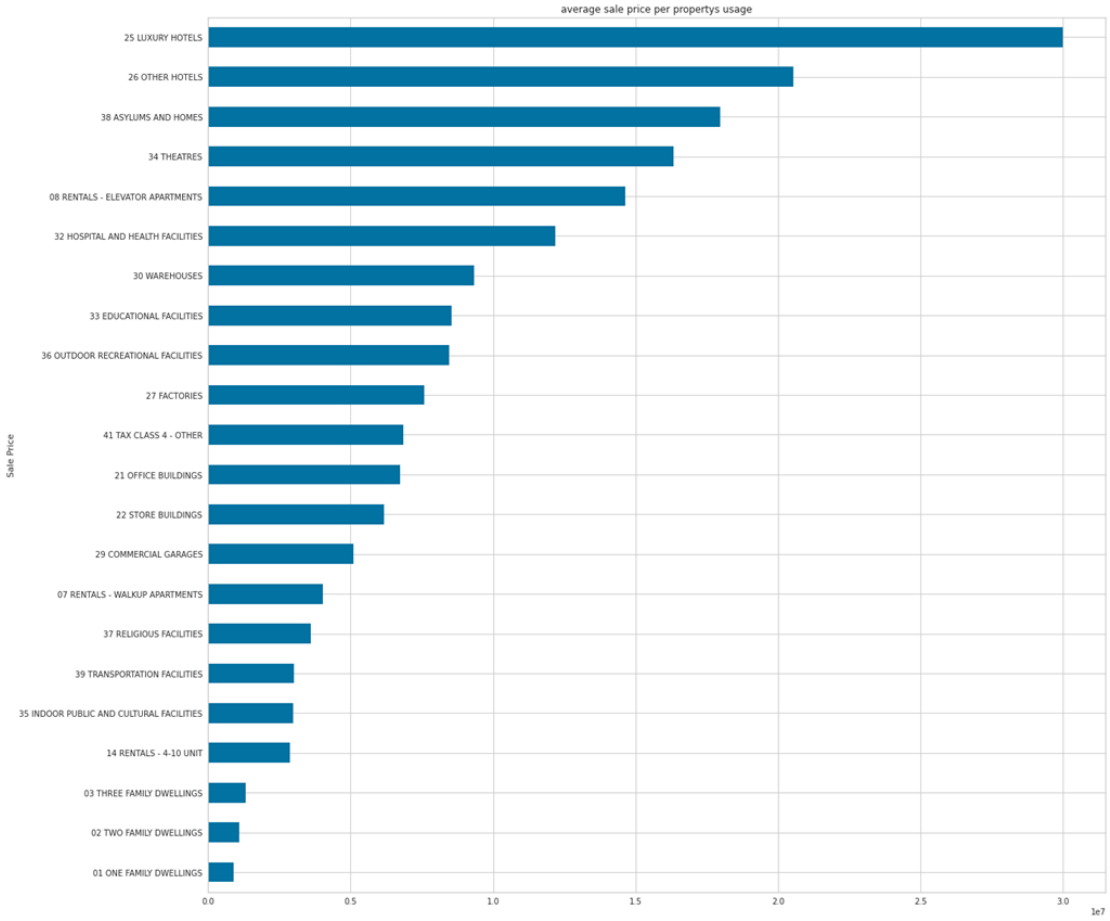
Στο παραπάνω γράφημα απεικονίζονται ξανά οι 5 περιοχές της Νέας Υόρκης ανά μέσο όρο τιμής πώλησης ακινήτων. Όπως καταλαβαίνουμε, το μεγαλύτερο μέσο όρο τιμής πώλησης ακινήτων για τα έτη 2021-2022, είχε η περιοχή του Manhattan , με σχεδόν 1 εκατομμύριο δολάρια, ενώ το τις χαμηλότερες τιμές στα ακίνητα είχε η περιοχή του State Island, με την περιοχή του Queens να ακολουθεί στην κατάταξη.

Θα μπορούσαμε να υποθέσουμε, λοιπόν, πως το Manhattan είχε τις λιγότερες αγορές ακινήτων στην περίοδο που αναλύουμε, εξαιτίας των υψηλών τιμών που παρουσιάζουν τα ακίνητα στην περιοχή αυτή.



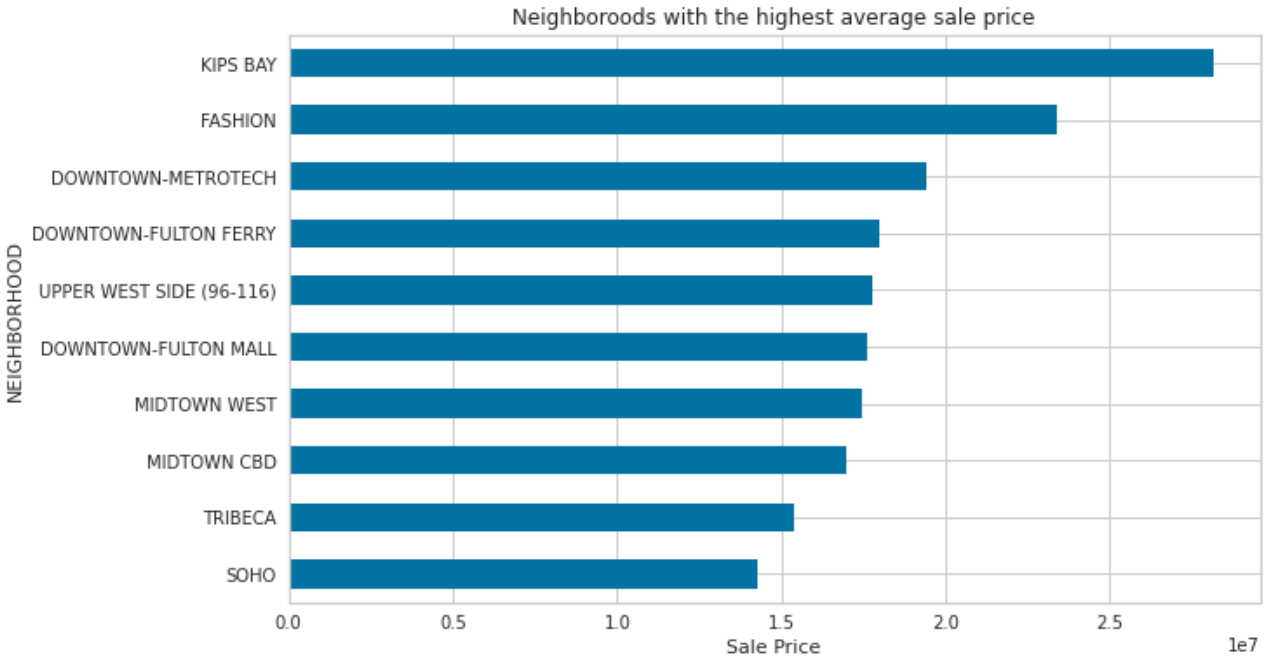
Εικόνα 14 Έτος κατασκευής ακινήτου

Στο παραπάνω ραβδόγραμμα απεικονίζονται τα έτη κατασκευής των ακίνητων που συμπεριλαμβάνονται στα δεδομένα μας. Παρατηρούμε, λοιπόν, πως το μεγαλύτερο ποσοστό ακινήτων (περισσότερα από 15.000 ακίνητα) χτίστηκε ανάμεσα στο 1900 με 1950, το οποίο δικαιολογείται με βάσει τα ιστορικά γεγονότα, καθώς η Νέα Υόρκη έγινε το κέντρο του διεθνούς εμπορίου και της βιομηχανίας στις αρχές του 20ού αιώνα και μετά την οικονομική κρίση του 1929 (κραχ), άρχισαν να χτίζονται οι πρώτοι ουρανοξύστες (1933 και έπειτα).



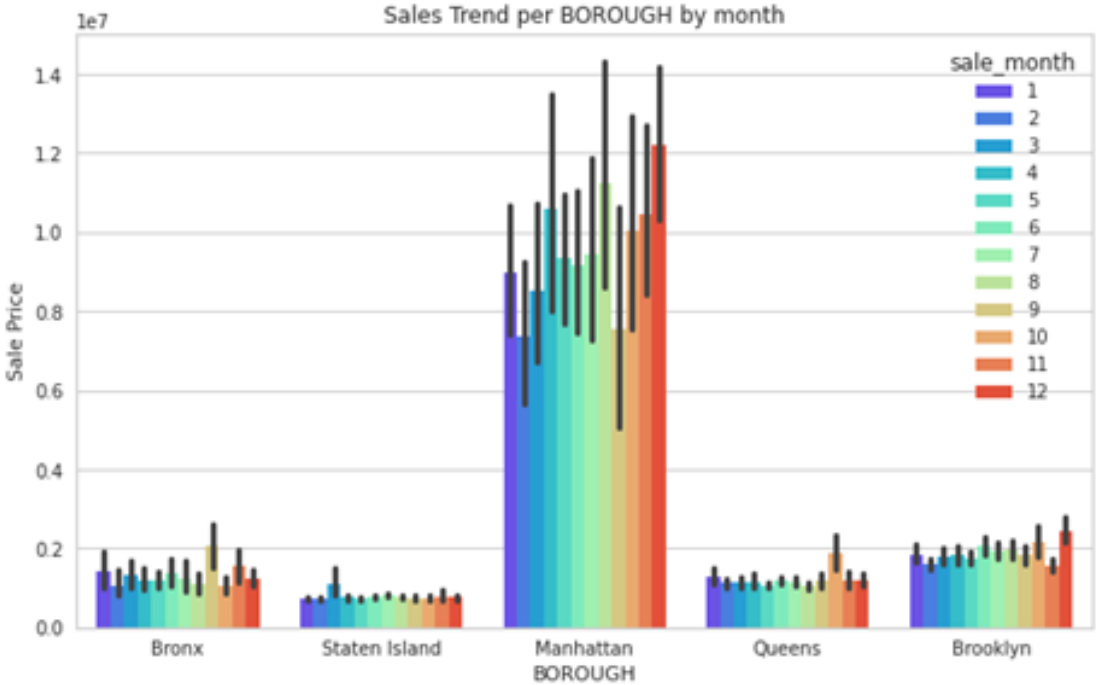
Εικόνα 15 Μέση τιμή πώλησης ανά κατηγορία ακινήτου

Στο παραπάνω ραβδόγραμμα απεικονίζεται η μέση τιμή πώλησης ανά κατηγορία ακινήτου. Παρατηρούμε, λοιπόν, πως τα ξενοδοχεία κατέχουν την πρώτη θέση με μέσο όρο τιμής πώλησης 3 εκατομμύρια δολλάρια για τα ξενοδοχεία πολυτελείας και πάνω από 2 εκατομμύρια δολάρια τα υπόλοιπα ξενοδοχεία. Αξιοσημείωτο είναι το γεγονός πώς από τους διάφορους τύπους κατοικιών την χαμηλότερη μέση τιμή πώλησης παρουσιάζουν οι μονοκατοικίες με λιγότερο από 200 χιλιάδες δολάρια.



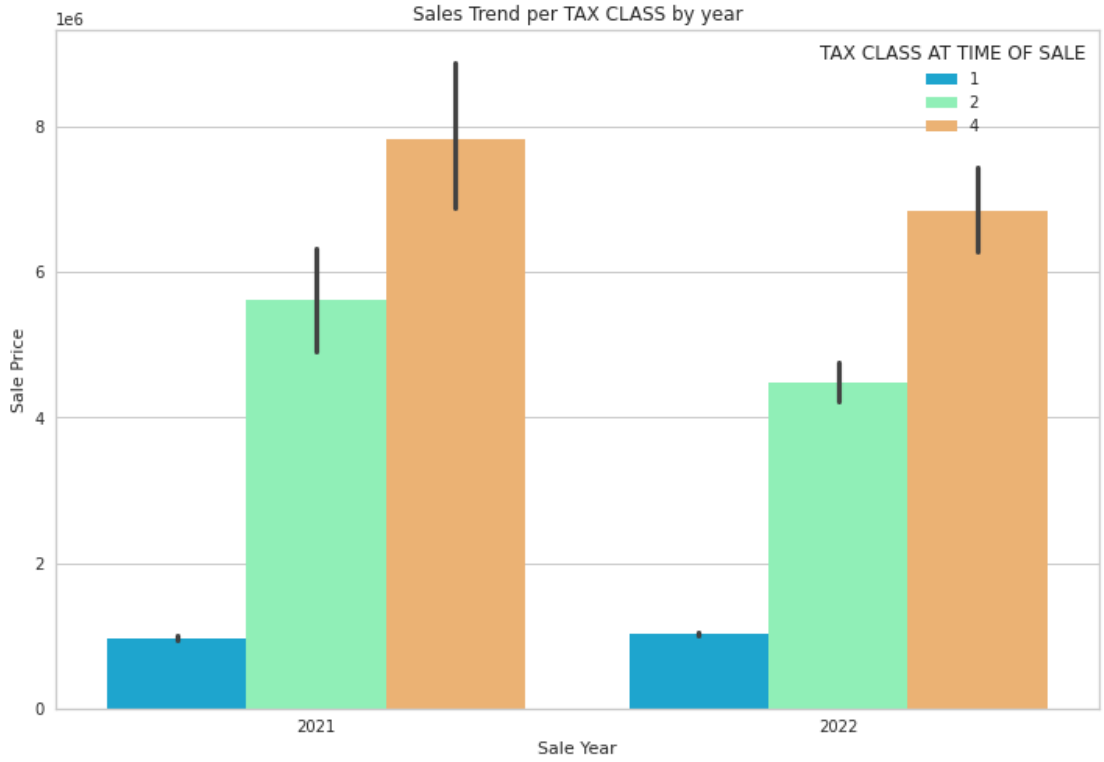
Εικόνα 16

Στο παραπάνω ραβδόγραμμα απεικονίζονται οι 10 γειτονιές της Νέας Υόρκης με το μεγαλύτερο μέσο όρο τιμής πώλησης ακινήτου. Παρατηρούμε πως η περιοχή Kips Bay κατέχει την πρώτη θέση με μέση τιμή πώλησης περίπου 2,75 εκατομμύρια δολάρια. Η γειτονία αυτή βρίσκεται στην ανατολική πλευρά της περιοχής του Μανχάταν και θεωρείται από τα καλύτερα μέρη για να ζεις στην Νέα Υόρκη, γεγονός που δικαιολογεί και τις υψηλές τιμές ακινήτων. Επίσης, παρατηρούμε πως και οι 10 γειτονίες που κατέχουν το μεγαλύτερο μέσο όρο τιμής πώλησης ακινήτου βρίσκονται στην περιοχή του Μανχάταν, το οποίο θεωρείται αναμενόμενο αποτέλεσμα καθώς το Μανχάταν θεωρείται το κέντρο της Νέας Υόρκης.



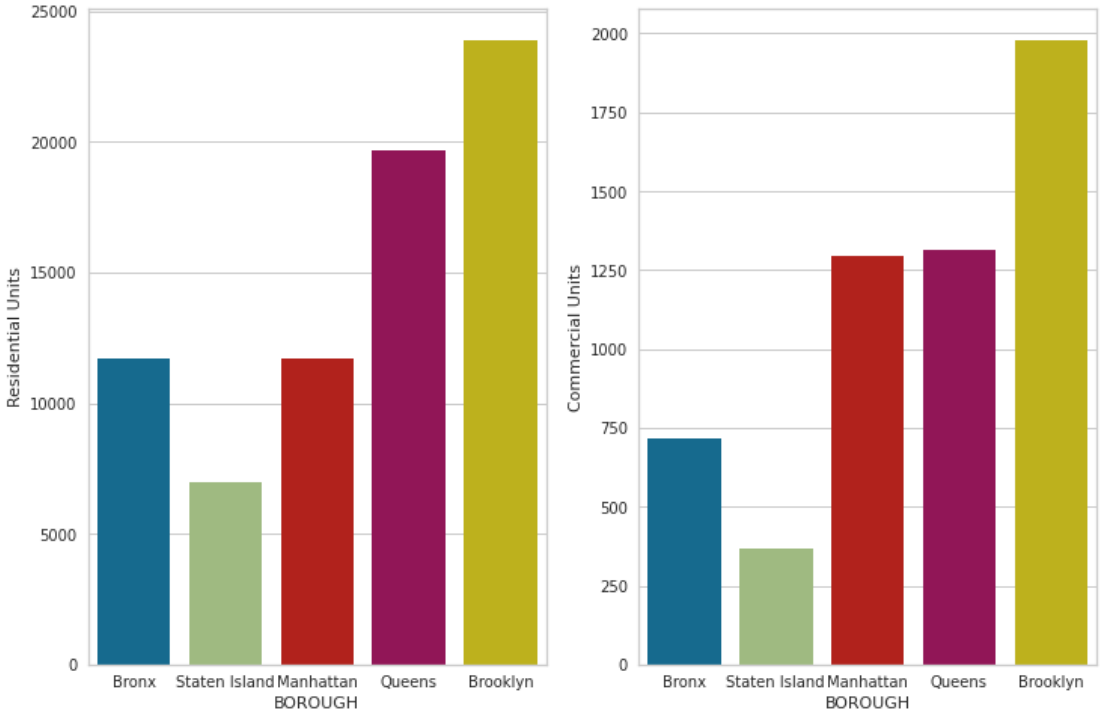
Εικόνα 17

Στο παραπάνω ραβδόγραμμα απεικονίζεται η τάση πώλησης ακινήτων στις 5 περιοχές της Νέας Υόρκης που περιλαμβάνονται στα δεδομένα μας ανά μήνα πωλήσεις. Παρατηρούμε, λοιπόν, πως τα ακριβότερα ακίνητα του δείγματος μας πουλήθηκαν στο Μανχάταν, με τις καλύτερες πωλήσεις βάσει τιμής πώλησης να έχουν πραγματοποιηθεί το μήνα Δεκέμβριο. Αντίθετα, η περιοχή με τις χαμηλότερες τιμές πωλήσεις είναι το Staten Island, το οποίο παρουσιάζει σταθερότητα τιμών στην πώληση ακινήτων, κοντά στα 100 χιλιάδες δολάρια.



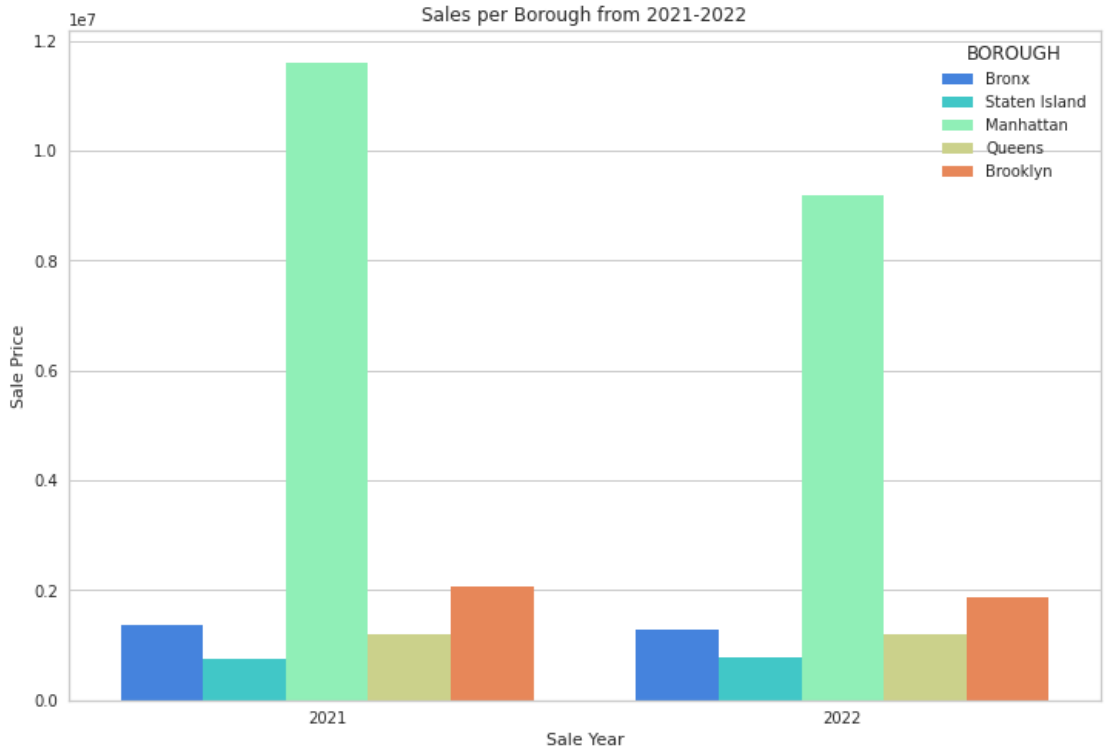
Εικόνα 18

Στο παραπάνω ραβδόγραμμα απεικονίζεται η τάση πώλησης ακινήτων για τα έτη 2021-2022 ανάλογα τη φορολογική κλάση που ανήκε το ακίνητο κατά τη στιγμή της πώλησης. Παρατηρούμε, λοιπόν, πως τα ακριβότερα ακίνητα του δείγματος μας που πουλήθηκαν τόσο το έτος 2021 όσο και το 2022 ανήκαν στην 4η φορολογική κατηγορία, η οποία περιλαμβάνει ακίνητα που δεν προορίζονται για οικιστική χρήση, όπως γραφεία, βιβλιοθήκες, γκαράζ κ.α. Με βάσει το ραβδόγραμμα, το 2021 η τιμή πώλησης για τα ακίνητα αυτής της κατηγορίας έφτασε τα 8 εκατομμύρια δολάρια, ενώ το 2022 περίπου τα 7 εκατομμύρια.



Εικόνα 19 Residential Units VS Commercial Units

Στο αριστερό ραβδόγραμμα απεικονίζεται το σύνολο των οικιστικών διαμερισμάτων στα ακίνητα που ανήκουν σε κάθε μία από τις 5 περιοχές της Νέας Υόρκης που περιλαμβάνονται στα δεδομένα μας. Αντίθετα, στο δεξί ραβδόγραμμα απεικονίζεται το σύνολο των εμπορικών διαμερισμάτων στα ακίνητα που ανήκουν σε κάθε μία αυτές τις 5 περιοχές. Συγκρίνοντας τα δύο ραβδογράμματα παρατηρούμε πως τα ακίνητα του δείγματός μας περιλαμβάνουν περισσότερα οικιστικά διαμερίσματα παρά εμπορικά. Το Brooklyn, που κατέχει την πρώτη θέση σε αριθμό τόσο οικιστικών αλλά και εμπορικών διαμερισμάτων, παρουσιάζει σύμφωνα με τις πωλήσεις του το 2021-2022 περίπου 24 χιλιάδες οικιστικά διαμερίσματα, ενώ μόλις 2 χιλιάδες εμπορικά.

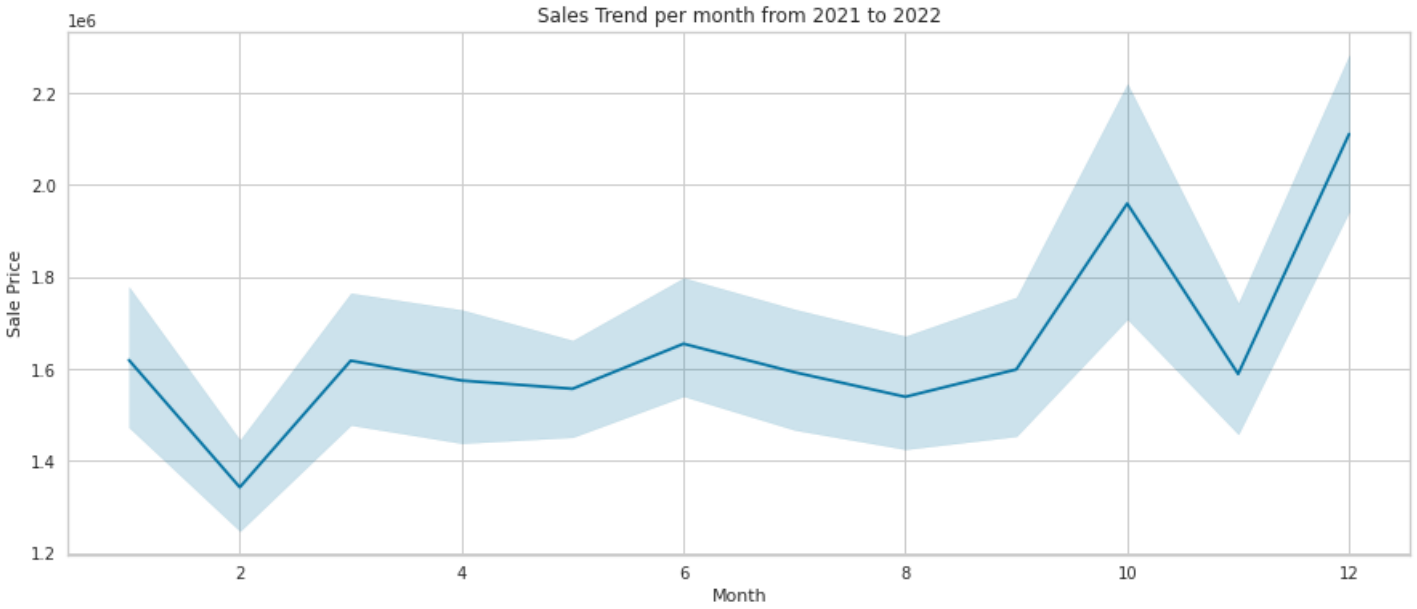


Εικόνα 20

Στο παραπάνω ραβδόγραμμα απεικονίζονται οι πωλήσεις κατά τη διάρκεια των ετών 2021-2022 ανά περιοχή. Παρατηρούμε, λοιπόν, πως οι υψηλότερες πωλήσεις ακινήτων, τόσο κατά τη διάρκεια του έτους 2021 όσο και το 2022, πραγματοποιήθηκαν στην περιοχή του Μανχάταν. Το 2021, οι τιμές πώλησης για το Μανχάταν έφτασαν τα 1.2 εκατομμύρια δολάρια και το 2022 τα 900 χιλιάδες, με το Brooklyn να ακολουθεί με περίπου 200 χιλιάδες δολάρια και στα δύο έτη.

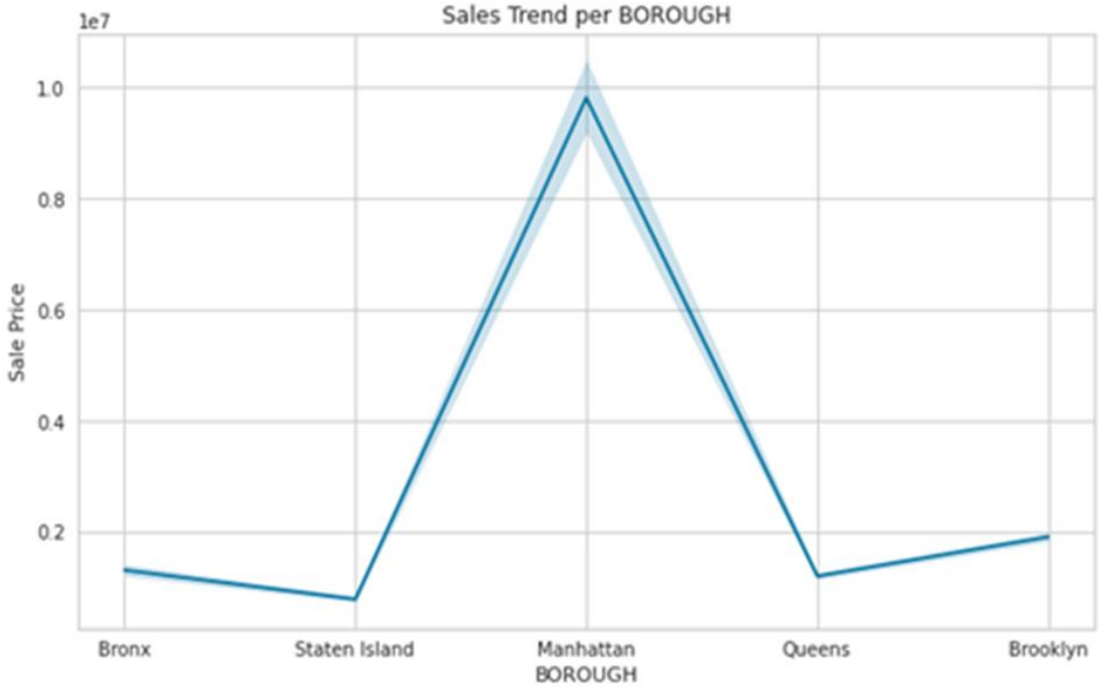
5. Γραμμικά γραφήματα (Lineplots)

Στα μαθηματικά, ένα γραμμικό διάγραμμα είναι μια γραφική αναπαράσταση δεδομένων, συνήθως, ενός μικρού συνόλου δεδομένων. Είναι επίσης γνωστό ως διάγραμμα κουκκίδας. Στο διάγραμμα γραμμών, τα δεδομένα απεικονίζονται σε μια γραμμή αριθμών χρησιμοποιώντας σύμβολα για τη συχνότητα.



Εικόνα 21

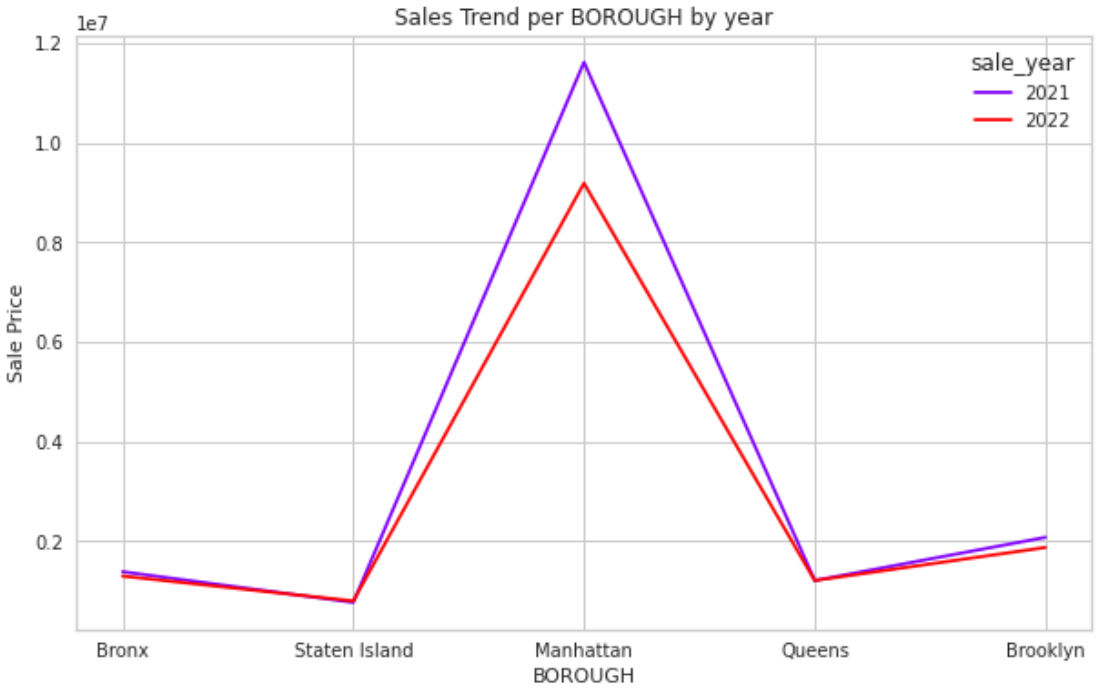
Στο παραπάνω γραμμικό γράφημα απεικονίζεται η τάση πώλησης ακινήτων κατά τη διάρκεια των ετών 2021-2022 ανά μήνα. Παρατηρούμε, λοιπόν, πως τα ακριβότερα ακίνητα του δείγματος μας για τα έτη 2021-2022, πουλήθηκαν το μήνα Δεκέμβριο, φτάνοντας πάνω από 2 εκατομμύρια δολάρια. Αντίθετα, η χαμηλότερη τιμή πώλησης παρατηρήθηκε το Φεβρουάριο με τιμή περίπου στα 1,3 εκατομμύρια δολάρια. Σύμφωνα με το γράφημα μας, η τάση πώλησης ακινήτων φαίνεται να έχει ανοδική πορεία η οποία τείνει να συνεχιστεί και το 2023.



Εικόνα 22

Στο παραπάνω γραμμικό γράφημα απεικονίζεται η τάση πώλησης ακινήτων ανά περιοχή. Παρατηρούμε, λοιπόν, πως τα ακριβότερα ακίνητα του δείγματος μας πουλήθηκαν στην περιοχή του Μανχάταν, όπως προαναφέρθηκε στο αντίστοιχο ραβδόγραμμα, φτάνοντας κατά μέσο όρο την τιμή

πώλησης του 1 εκατομμυρίου δολαρίων. Αντίθετα, η χαμηλότερη τιμή πώλησης παρατηρήθηκε στο Staten Island με τιμή περίπου στα 100 χιλιάδες δολάρια.



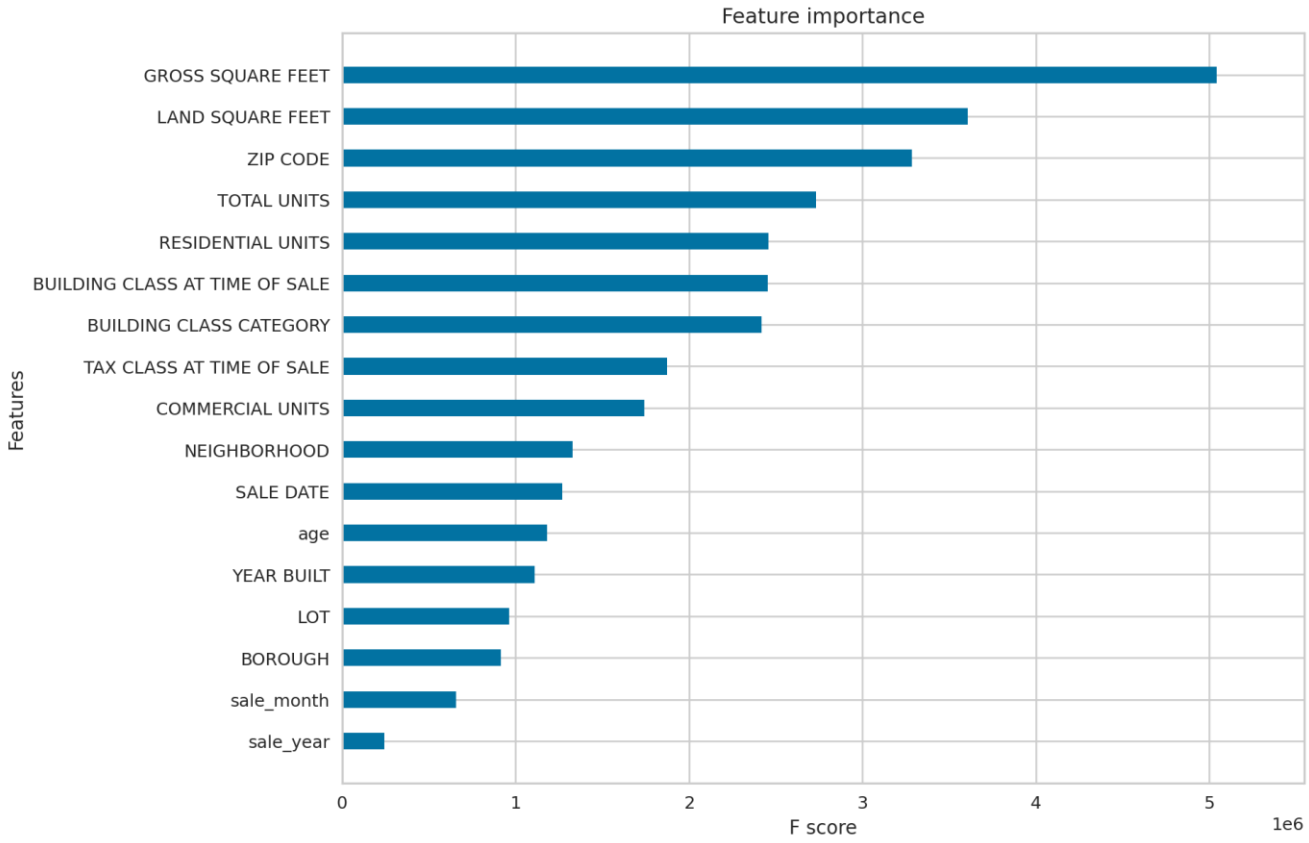
Εικόνα 23

Το παραπάνω διάγραμμα αποτελεί εξέλιξη του προηγούμενου, καθώς απεικονίζεται τα ίδια στοιχεία αλλά διαχωρισμένα σε δύο ομάδες δεδομένων για τα έτη 2021 και 2022. Όπως συμπεράναμε και προηγουμένως, τα ακριβότερα ακίνητα του δείγματος μας πουλήθηκαν στην περιοχή του Μανχάτταν φτάνοντας κατά μέσο όρο την τιμή πώλησης του 1,2 εκατομμυρίου δολαρίων για το 2021 και των 900 χιλιάδων δολαρίων για το 2022. Η χαμηλότερη τιμή πώλησης όπως παρατηρήθηκε στο Staten Island παρέμεινε σταθερή και στα δύο έτη, με τιμή περίπου στα 100 χιλιάδες δολάρια.

6.3. Χρήση Αλγορίθμου XGBoost

Προκειμένου να χρησιμοποιηθούν όλες οι μεταβλητές που απαρτίζουν τα δεδομένα μας στη δημιουργία του μοντέλου με τη χρήση του αλγορίθμου XGBoost, ακόμα και των κατηγορικών, δηλαδή των μεταβλητών εκείνων οι οποίες δεν εκφράζουν ποσότητα, αλλά ποιότητα, είναι αναγκαίο να υποβληθούν σε περαιτέρω επεξεργασία:

1. Επιλογή κατηγορηματικών μεταβλητών: Κάθε κατηγορική μεταβλητή αποτελείται από μοναδικές τιμές. Μια κατηγορηματική μεταβλητή λέγεται ότι έχει υψηλή πληθικότητα όταν υπάρχουν πάρα πολλές από αυτές τις μοναδικές τιμές. Πριν προχωρήσουμε στο επόμενο βήμα που είναι το Label encoding, επιλέξαμε να αφαιρέσουμε από το σύνολο των δεδομένων μας τις μεταβλητές: ADDRESS και BLOCK, οι οποίες παρουσιάζουν μεγάλο αριθμό μοναδικών τιμών. Σε τέτοιες περιπτώσεις το Label encoding γίνεται μεγάλο πρόβλημα, αφού έχουμε ξεχωριστή στήλη για κάθε μοναδική τιμή (υποδεικνύοντας την παρουσία ή την απουσία της) στην κατηγορική μεταβλητή.
2. Label encoding – κωδικοποίηση «ετικετών»: Χρησιμοποιώντας τη βιβλιοθήκη Label Encoding της python, μετατρέψαμε όλες τις μεταβλητές μας σε τύπο “integer”. Σκοπός αυτού του βήματος είναι η μετατροπή των μη-ακέραιων μεταβλητών σε μια μορφή δεδομένων που υποστηρίζεται από τα μοντέλα μηχανικής μάθησης με την τεχνική της κωδικοποίησης, δηλαδή κωδικοποιούνται σε αντίστοιχες αριθμητικές με την αντιστοίχιση κάθε τιμή της κατηγορικής μεταβλητής σε μια αριθμητική.



3. Μετασχηματισμός τετραγωνικής ρίζας στη μεταβλητή y: Με σκοπό τη μείωση της ετεροσκεδαστικότητας των υπολειμμάτων που παρατηρείται στη συγκεκριμένη περίπτωση γραμμικής παλινδρόμησης, προχωρήσαμε στο μετασχηματισμό τετραγωνικής ρίζας στη μεταβλητή SALESPRICE.
4. Διαχωρισμός δεδομένων σε train data και test data: Για το συγκεκριμένο μοντέλο χρησιμοποιήσαμε για εκπαίδευση το 80% των δεδομένων και για δοκιμή το υπόλοιπο 20%.
5. Hypertuning: Στη μηχανική μάθηση, το hypertuning είναι το πρόβλημα της επιλογής ενός συνόλου παραμέτρων που βελτιστοποιούν την απόδοση και την ικανότητα πρόβλεψης ενός μοντέλου μηχανικής μάθησης. Στην γλώσσα προγραμματισμού Python, υπάρχουν διάφορες τεχνικές εύρεσης βέλτιστων παραμέτρων, όπως η μέθοδος GridSearchCV και RandomizedSearchCV. Για την υλοποίηση της παρούσας εργασίας χρησιμοποιήθηκε η μέθοδος GridSearchCV.

Παρακάτω παρατίθεται ο πίνακας σημαντικότητας που παρήγαγε το εξεταζόμενο μοντέλο για τις μεταβλητές του δείγματος. Για τη δημιουργία του συγκεκριμένου ραβδογράμματος χρησιμοποιήθηκε η μέθοδος “gain”, η οποία υποδηλώνει τη σχετική συνεισφορά της κάθε μεταβλητής στο παραγόμενο μοντέλο υπολογίζοντας τη συνεισφορά κάθε χαρακτηριστικού σε κάθε δέντρο απόφασης που δημιουργήθηκε για το μοντέλο. Παρατηρούμε, λοιπόν, πως τη μεγαλύτερη συνεισφορά στη παραγόμενο μοντέλο είχε η μεταβλητή “GROSS SQUARE FEET”, γεγονός που υποδεικνύει πως το συνολικό εμβαδόν ολόκληρου του κτιρίου συμβάλλει σημαντικά στην διαμόρφωση της τιμής πώλησης του ακινήτου. Έπειτα, ακολουθούν τα χαρακτηριστικά “LAND SQUARE FEET” και “ZIP CODE” που κατέχουν αξιοσημείωτη επιρροή στη διαμόρφωση της τιμής πώλησης των ακινήτων.

Εικόνα 24

Για την αξιολόγηση του μοντέλου παλινδρόμησης μηχανικής μάθησης χρησιμοποιήσαμε τη μετρική αξιολόγησης:

- R2 - squared (τετραγωνικό R): Το τετραγωνικό R – squared αναπαριστά το ποσοστό της διακύμανσης στην εξαρτημένη μεταβλητή. Οι τιμές αυτής της μετρικής είναι μικρότερες από τη μονάδα πάντα (παίρνει τιμές $0 \leq x \leq 1$).

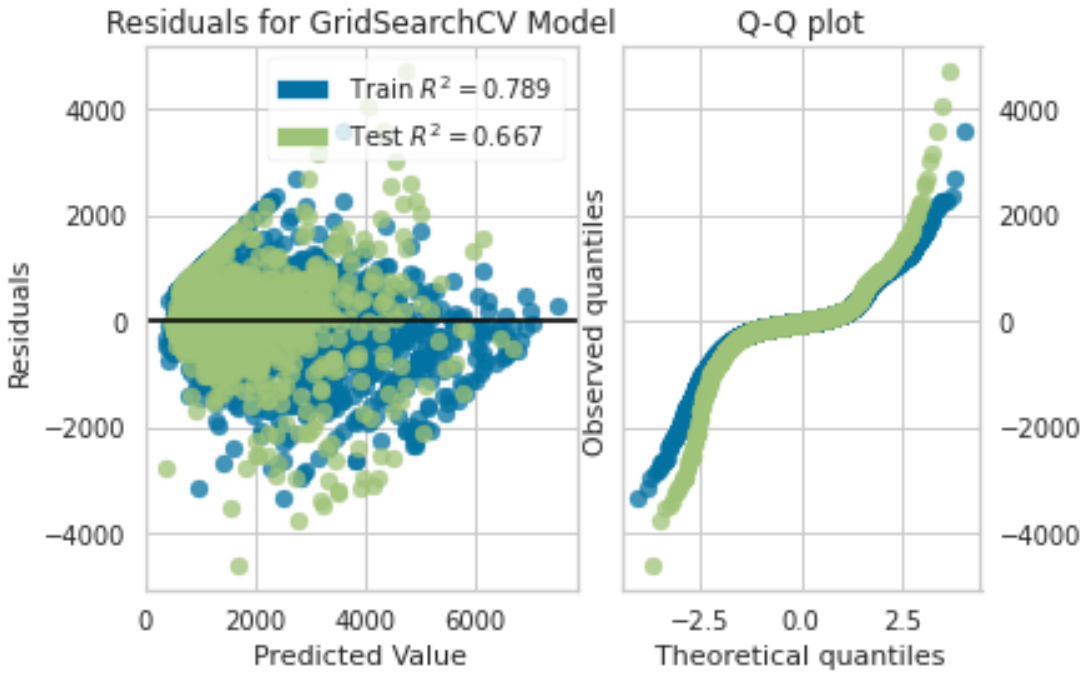
$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \overline{y})^2}$$

7. Αποτελέσματα

Τα αποτελέσματα τόσο για το υποσύνολο εκπαίδευσης (train data), όσο και για το υποσύνολο δοκιμής (test data), είναι τα εξής:

	Train Data	**Test Data**
R^2	0,78915	0,66691

Τα αποτελέσματα παρουσιάζονται και γραφικά στο παρακάτω Γράφημα Καταλοίπων:



Εικόνα 25 Residual Plot

Στα παραπάνω Γραφήματα Καταλοίπων φαίνεται η σχέση των καταλοίπων με την προβλεπόμενη τιμή (αριστερά), και ο βαθμός κανονικότητας τους (δεξιά). Από τα δύο γραφήματα φαίνεται ότι το μοντέλο προσαρμόζεται καλά στα δεδομένα.

8. Συμπεράσματα

Το πρόβλημα που διερευνήθηκε στην παρούσα διπλωματική εργασία ήταν η δημιουργία ενός συστήματος πρόβλεψης της τιμής των ακινήτων με τη βοήθεια μοντέλων μηχανικής μάθησης. Ο αλγόριθμος που επιλέχθηκε είναι ο XGBoost, ο οποίος αποτελεί μία από τις πιο αποδοτικές επιλογές, όσον αφορά την επίλυση του συγκεκριμένου προβλήματος και την διενέργεια προβλέψεων στην τιμή σπιτιών στην πόλη της Νέας Υόρκης.

Εξετάζοντας το διάγραμμα σπουδαιότητας χαρακτηριστικών με βάση το κέρδος, διαπιστώθηκε ότι οι μεταβλητές GROSS SQUARE FEET, LAND SQUARE FEET και ZIP CODE είναι τα τρία χαρακτηριστικά που συνεισέφεραν περισσότερο στο κάθε δέντρο απόφασης που δημιουργήθηκε για το μοντέλο. Στο ίδιο διάγραμμα, μπορεί να παρατηρηθεί πως στα 17 καλύτερα χαρακτηριστικά υπάρχουν και 3 που δεν υπήρχαν στα αρχικά σύνολα δεδομένων. Αυτά τα χαρακτηριστικά, δημιουργήθηκαν κατά την προ-επεξεργασία των δεδομένων και είναι τα εξής:

- age: Η ηλικία του κτιρίου κατά τη στιγμή της πώλησης, η οποία υπολογίστηκε ως το υπόλοιπο της αφαίρεσης του έτους κατασκευής από το έτος πώλησης.
- sale_month: Ο μήνας κάθε αγοράς, ο οποίος υπολογίστηκε από την ημερομηνία πώλησης του ακινήτου.
- sale_year: Το έτος κάθε αγοράς, το οποίο υπολογίστηκε από την ημερομηνία πώλησης του ακινήτου.

Στο πρόβλημα αυτό, επιλέχθηκε να χρησιμοποιηθεί το R2 - squared για την αξιολόγηση του μοντέλου παλινδρόμησης μηχανικής μάθησης που δημιουργήθηκε. Το αποτέλεσμα που σημειώθηκε για το υποσύνολο δοκιμής ήταν 0,66691. Στην προβλεπτική ικανότητα του μοντέλου, εκτός από τον αλγόριθμο ταξινόμησης XGBoost, συνέβαλλαν και άλλοι μέθοδοι, όπως το GridSearchCV που χρησιμοποιήθηκε για το Hypertuning και η τεχνική K-fold, η οποία ορίστηκε στο 5 και επέτρεψε το διαχωρισμό των δεδομένων μας σε 5 υποσύνολα, με σκοπό να αξιολογηθεί η προβλεπτική ικανότητα του εξεταζόμενου μοντέλου σε άγνωστα δεδομένα.

Διαπιστώνουμε, λοιπόν, πως η δημιουργία ενός αποδοτικού μοντέλου με ισχυρή προβλεπτική ικανότητα είναι αποτέλεσμα ενός συνόλου ενεργειών επεξεργασίας δεδομένων και υπολογιστικών μεθόδων, τα οποία σε συνδυασμό μπορούν να αποφέρουν ακριβή αποτελέσματα.

Η επιστήμη της Μηχανικής Μάθησης είναι ένας διαρκώς αναπτυσσόμενος κλάδος που διερευνά τη μελέτη και την κατασκευή αλγορίθμων που μπορούν να μαθαίνουν από τα δεδομένα και να κάνουν προβλέψεις σχετικά με αυτά (Wikipedia, 2013). Είναι σημαντικό να σημειωθεί πως δεν είναι όλοι οι αλγόριθμοι παλινδρόμησης κατάλληλοι για την επίλυση όλων των προβλημάτων και την επιτυχή πρόβλεψη τιμών κάποιων χαρακτηριστικών. Η επιλογή του κατάλληλου αλγορίθμου είναι αποτέλεσμα μελέτης, εφαρμογής διάφορων αλγορίθμων και υπολογιστικών τεχνικών, αξιολόγησης και σύγκρισης αποτελεσμάτων.

Παράρτημα Α’ - Κώδικας

```
# Load data libraries
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O
from sklearn.model_selection import train_test_split
# For visualizations
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
# Data preparation
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import Pipeline
from datetime import datetime
import math
# Modeling
import xgboost as xgb
from xgboost import XGBRegressor
from xgboost import plot_importance
from xgboost import cv
# Evaluation
from sklearn.model_selection import KFold, learning_curve
from sklearn.metrics import r2_score
from sklearn.model_selection import GridSearchCV
from yellowbrick.datasets import load_concrete
from yellowbrick.regressor import ResidualsPlot
from matplotlib import pyplot
# First, let's load the data
df = pd.read_csv('/kaggle/input/datasetsnewyork/rollingsales.csv')
# Let's see the first 5 rows
df.head()
# Deleting columns: TAX CLASS AT PRESENT and BUILDING CLASS AT PRESENT, since we are no
t interested in "present" details
del df['TAX CLASS AT PRESENT']
del df['BUILDING CLASS AT PRESENT']
# Rename column SALE PRICE to SALEPRICE
df.rename(columns = {'SALE PRICE':'SALEPRICE'}, inplace=True)
# Let's see the details of our data
df.info()
#find all NaN values
nan_count = df.isna().sum()
print (nan_count)
#Heatmap of missing values
plt.figure(figsize=(16,10))
sns.heatmap(df.isnull(), cbar=False, cmap="YlGnBu")
plt.show()
df.dropna(axis=1, thresh=int(0.01*df.shape[0]), inplace=True)
df.dropna(axis=0, how='any', inplace=True)
# Convert features that contain only integer numbers to type: integer
df['SALEPRICE'] = df['SALEPRICE'].astype(int)
df['TOTAL UNITS'] = df['TOTAL UNITS'].astype(int)
df['RESIDENTIAL UNITS'] = df['RESIDENTIAL UNITS'].astype(int)
df['COMMERCIAL UNITS'] = df['COMMERCIAL UNITS'].astype(int)
df['YEAR BUILT'] = df['YEAR BUILT'].astype(int)
df['GROSS SQUARE FEET']=df['GROSS SQUARE FEET'].astype(int)
df['LAND SQUARE FEET']=df['LAND SQUARE FEET'].astype(int)
df['TAX CLASS AT TIME OF SALE']=df['TAX CLASS AT TIME OF SALE'].astype(int)
df['ZIP CODE']=df['ZIP CODE'].astype(int)
# Convert the column SALE DATE from object to datetime
df['SALE DATE'] = pd.to_datetime(df['SALE DATE'])
```

```
# Create two new columns with the month and year for each sale
df['sale_month']= df['SALE DATE'].dt.month
df['sale_year']= df['SALE DATE'].dt.year
# Create one column showing the age of the building at the time of the sale
df['age']= df['sale_year']-df['YEAR BUILT']
df.info()
df.head()
df[df['SALEPRICE']==0].value_counts().count()
df[df['GROSS SQUARE FEET']==0].value_counts().count()
df[df['LAND SQUARE FEET']==0].value_counts().count()
df[df['YEAR BUILT']==0].value_counts().count()
# Keep only the records where SALEPRICE is more than 0
df = df[df['SALEPRICE']!=0]
df.head()
# Keep only records where GROSS SQUARE FEET is not 0
df = df[df['GROSS SQUARE FEET']!=0]
df.info()
# Keep only records where LAND SQUARE FEET is not 0
df = df[df['LAND SQUARE FEET']!=0]
df.info()
df = df[df['YEAR BUILT']!=0]
df.info()
#BOROUGH is A digit code for the borough the property is located in; in order these are Manhattan (
1), Bronx (2), Brooklyn (3), Queens (4), and Staten Island (5)
df['BOROUGH'].value_counts().count()
df['BOROUGH'] = df['BOROUGH'].astype(str)
df['BOROUGH'].replace({'1':'Manhattan','2':'Bronx','3':'Brooklyn','4':'Queens','5':'Staten Island'},inplac
e=True)
df.head()
df.info()
# Convert text features to type:string
df['NEIGHBORHOOD'] = df['NEIGHBORHOOD'].astype(str)
df['BUILDING CLASS CATEGORY'] = df['BUILDING CLASS CATEGORY'].astype(str)
df['BUILDING CLASS AT TIME OF SALE'] = df['BUILDING CLASS AT TIME OF SALE'].astype(
str)
df['ADDRESS'] = df['ADDRESS'].astype(str)
# Lets' see the Saleprice per Borough for each year
plt.figure(figsize=(30,10))
sns.boxplot(x="BOROUGH",y="SALEPRICE",hue='sale_year',data=df, palette='rainbow')
df['SALEPRICE'].describe()
# Removing outliers
df[df['SALEPRICE'] >= 55000000].value_counts().count()
df = df[df['SALEPRICE']<55000000]
plt.figure(figsize=(30,10))
sns.boxplot(x="BOROUGH",y="SALEPRICE",hue='sale_year',data=df, palette='rainbow')
plt.figure(figsize=(20,20))
displot = sns.displot(df, x='SALEPRICE',hue='BOROUGH',kind='kde',fill=True)
histogram_df = df.hist(figsize = (15,15), grid=False)
df['BOROUGH'].value_counts().plot.bar()
plt.title("Number of property sales by Borough");
df['SALEPRICE'].groupby(df['BOROUGH']).mean().plot.bar()
plt.title("Average sale price per Borough");
sns.histplot(df['YEAR BUILT'],bins=100);
plt.ylabel("Number of properties")
plt.xlabel("Property built year");
plt.figure(figsize=(20,20))
df['SALEPRICE'].groupby(df['BUILDING CLASS CATEGORY']).mean().sort_values().plot.barh()
plt.ylabel("Sale Price")
plt.title('average sale price per propertys usage');
plt.figure(figsize=(10,6))
```

```
df['SALEPRICE'].groupby(df['NEIGHBORHOOD']).mean().sort_values()[234:244].plot.barh()
plt.title('Neighboroods with the highest average sale price')
plt.xlabel("Sale Price");
plt.figure(figsize=(15,6))
sns.lineplot(x='sale_month',y='SALEPRICE',data=df)
plt.title('Sales Trend per month from 2021 to 2022')
plt.ylabel('Sale Price')
plt.xlabel('Month')
plt.show();
plt.figure(figsize=(10,6))
sns.lineplot(x='BOROUGH',y='SALEPRICE',data=df)
plt.title('Sales Trend per BOROUGH')
plt.ylabel('Sale Price')
plt.show();
plt.figure(figsize=(10,6))
sns.barplot(x='BOROUGH', y='SALEPRICE', hue='sale_month', data=df, palette='rainbow');
plt.title('Sales Trend per BOROUGH by month')
plt.ylabel('Sale Price')
plt.show()
plt.figure(figsize=(10,6))
sns.heatmap(df.corr());
plt.figure(figsize=(12,8))
plt.subplot(1, 2, 1)
sns.barplot(x="BOROUGH", y="RESIDENTIAL UNITS", data=df, estimator=sum, ci=None)
plt.ylabel('Residential Units')
plt.subplot(1, 2, 2)
sns.barplot(x="BOROUGH", y="COMMERCIAL UNITS", data=df, estimator=sum, ci=None)
plt.ylabel('Commercial Units');
plt.figure(figsize=(10,6))
sns.lineplot(x="BOROUGH",y="SALEPRICE",hue='sale_year',data=df, palette='rainbow',ci=None)
plt.title('Sales Trend per BOROUGH by year');
plt.ylabel('Sale Price');
plt.figure(figsize=(15,6))
plt.subplot(1, 2, 1)
plt.title('Sales Trend for Commercial Units by year')
sns.lineplot(x="COMMERCIAL UNITS",y="SALEPRICE",hue='sale_year',data=df, palette='rainbow'
)
plt.ylabel('Sale Price')
plt.xlabel('Commercial Units')
plt.subplot(1, 2, 2)
plt.title('Sales Trend for Residential Units by year')
sns.lineplot(x="RESIDENTIAL UNITS",y="SALEPRICE",hue='sale_year',data=df, palette='rainbow')
plt.ylabel('Sale Price')
plt.xlabel('Residential Units');
plt.figure(figsize=(12,8))
plt.title('Sales Trend per TAX CLASS by year')
sns.barplot(x="sale_year",y="SALEPRICE",hue='TAX CLASS AT TIME OF SALE',data=df[df['TAX
CLASS AT TIME OF SALE']!= ' '], palette='rainbow');
plt.xlabel('Sale Year')
plt.ylabel('Sale Price');
plt.subplots(figsize=(12,8))
sns.barplot(x='sale_year', y='SALEPRICE', hue='BOROUGH', data=df, palette='rainbow', ci=None)
plt.title('Sales per Borough from 2021-2022')
plt.ylabel('Sale Price')
plt.xlabel('Sale Year');
print(df.apply(lambda col: len(col.unique())))
#Choose categorical variables to drop
df = df.drop("ADDRESS", axis='columns')
df = df.drop("BLOCK", axis='columns')
from sklearn import preprocessing
```

```
from sklearn.preprocessing import LabelEncoder
le= preprocessing.LabelEncoder()
df['BOROUGH'] = le.fit_transform(df['BOROUGH'])
df['NEIGHBORHOOD'] = le.fit_transform(df['NEIGHBORHOOD'])
df['BUILDING CLASS CATEGORY'] = le.fit_transform(df['BUILDING CLASS CATEGORY'])
df['LOT'] = le.fit_transform(df['LOT'])
df['BUILDING CLASS AT TIME OF SALE'] = le.fit_transform(df['BUILDING CLASS AT TIME O
F SALE'])
df['SALE DATE'] = le.fit_transform(df['SALE DATE'])
df.info()
import numpy as np
df['SALEPRICE'] = np.sqrt(df['SALEPRICE'])
# Split the data into train and test
X, y = df.drop('SALEPRICE', axis=1), df.SALEPRICE
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
paramGrid = {
      "objective": ['reg:squarederror'],
      "subsample": [0.2],
      "colsample_bylevel": [0.1],
      "eta": [0.04],
      "max_depth": [3],
      "min_child_weight": [4],
      "n_estimators": [5000]
       }
model = xgb.XGBRegressor()
kfold = KFold(n_splits=5)
results = GridSearchCV(model, paramGrid, cv=kfold)
xg_reg = results.fit(X_train, y_train)
print(xg_reg.best_params_)
xg_reg.score(X_test, y_test)
best = xg_reg.best_estimator_
%matplotlib inline
fig, ax = pyplot.subplots(figsize=(10, 8), dpi=180)
plot_importance(best, height=0.4, importance_type='gain', max_num_features=30, show_values=False,
ax=ax)
pyplot.show()
# Predict on train data
preds = xg_reg.predict(X_train)
r2_train_xgbr = r2_score(y_train, preds)
print('R2 Score: ', r2_train_xgbr)
# Predict on test data
pred_test_xgbr = xg_reg.predict(X_test)
print(pred_test_xgbr)
r2_test_xgbr = r2_score(y_test, pred_test_xgbr)
print('R2 Score: ', r2_test_xgbr)
visualizer = ResidualsPlot(xg_reg, hist=False, qqplot=True)
visualizer.fit(X_train, y_train)
visualizer.score(X_test, y_test)
visualizer.show();
```

Βιβλιογραφία

Γνωστικό υπόβαθρο

1) Wikipedia, «Μηχανική μάθηση» , wikipedia, 3 Μαΐου 2013.
   Διαθέσιμο:
   https://el.wikipedia.org/wiki/Μηχανική_μάθηση
2) V7, «The Beginner's Guide to Deep Reinforcement Learning [2023]», 2 Φεβρουάριος 2023.
   Διαθέσιμο:
   https://www.v7labs.com/blog/deep-reinforcement-learning-guide
3) Simplilearn, «What is XGBoost? An Introduction to XGBoost Algorithm in Machine Learning», 23 Φεβρουάριος 2023.
   Διαθέσιμο:
   https://www.simplilearn.com/what-is-xgboost-algorithm-in-machine-learning-article
4) nvidia, «XGBoost», 23 Φεβρουάριος 2023.
   Διαθέσιμο:
   https://www.nvidia.com/en-us/glossary/data-science/xgboost/
5) Karriera.gr, «Να γιατί η Python είναι τόσο δημοφιλής», 21 Απριλίου 2021.
   Διαθέσιμο:
   https://resources.kariera.gr/el/blog/why-python-is-popular/
6) Kaggle, «What is Kaggle, Why I Participate, What is the Impact?», 2018.
   Διαθέσιμο:
   https://www.kaggle.com/getting-started/44916
7) Kaggle, «How to Use Kaggle».
   Διαθέσιμο:
   https://www.kaggle.com/docs/notebooks
8) Greek for Greeks, «Density Plots with Pandas in Python», 26 Νοεμβρίου 2022.
   Διαθέσιμο:
   https://www.geeksforgeeks.org/density-plots-with-pandas-in-python/

9) EFerrit, «Τι είναι ένα ιστόγραμμα;», 2023.
Διαθέσιμο:
https://el.eferrit.com/τι-είναι-ένα-ιστόγραμμα
10) Greek for Greeks, «How to create a seaborn correlation heatmap in Python?», 12 Νοεμβρίου 2020.
Διαθέσιμο:
https://www.geeksforgeeks.org/how-to-create-a-seaborn-correlation-heatmap-in-python/
11) Wikipedia, «Διάγραμμα», Wikipedia, 12 Ιουλίου 2007.
Διαθέσιμο:
https://el.wikipedia.org/wiki/Διάγραμμα
12) Serdar Yegulalp (2019), «What is Jupyter Notebook? Data analysis made easier», InfoWorld.
Διαθέσιμο:
https://www.infoworld.com/article/3347406/what-is-jupyter-notebook-data-analysis-made-easier.html

Μεσιτικός τομέας

1) Barnes, « Γιατί πρέπει να χρησιμοποιήσετε ένα μεσιτικό γραφείο για να σας καθοδηγήσει στην αγορά ακινήτων της Νέας Υόρκης;», 2023.
Διαθέσιμο:
https://barnes-newyork.com/el/tag/broker/
2) Crosby, N. Jackson, C., Orr, A., (2016), «Refining the real estate pricing model», Journal of Property Research
Διαθέσιμο:
https://www.tandfonline.com/doi/full/10.1080/09599916.2016.1237539?casa_token=-SaE34CeHDAAAAAA%3AQBbMCGxYugjAj_YlpAkSvEvSU0_dOFylHABIhOoKPERATvVl6-lBqKy_P3CRgu_HI5vD0UxTaZJv
3) NYC, Department of Finance, «Rolling Sales Data», 2023.
Διαθέσιμο:
Rolling Sales Data (nyc.gov)
4) New York Travel, «Η Οικονομία στη Νέα Υόρκη», 2023.
Διαθέσιμο:
https://newyorktravel.gr/η-οικονομία-στη-Νέα-Υόρκη
5) Bagnoli, C., & Smith, H. (1998). The theory of fuzzy logic and its application to real estate valuation. Journal of Real Estate Research, 16
6) Lentz, G. H., & Wang, K. (1998). Residential appraisal and the lending process: A survey of issues. Journal of Real Estate Research, 15(1/2)
7) Shiller R J. Understanding Recent Trends in House Prices and Home Ownership. Proceedings – Economic Policy Symposium - Jackson Hole (2007)
8) Joep Steegmans, Wolter Hassink. Financial Position and House Price Determination: An Empirical Study of Income and Wealth Effects. Journal of Housing Economics (2017)

Έρευνες με τη χρήση αλγορίθμων Τυχαίου Δάσους

1) Conway, D., Jennifer, E., (2018), «Artificial intelligence and machine learning: current application in real estate», Massachusetts Institute of Technology, Center for Real Estate, Programs in Real Estate Development, Vol. 3, 113-117.
2) Abigail Bola Adetunji, Oluwatobi Noah Akande, Funmilola Alaba Ajala, Ololade Oyewo, Yetunde Faith Akande, Gbenle Oluwadara (2022) «House Price Prediction using Random Forest Machine Learning Technique»
Διαθέσιμο:
https://www.sciencedirect.com/science/article/pii/S1877050922001016
3) Antipov, Evgeny και Pokryshevskaya, Elena (2010), «Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics»
Διαθέσιμο:

https://mpra.ub.uni-muenchen.de/27645/1/MPRA_paper_27645.pdf
4) Rinabi Tanamal, Nathalia Minoque, Trianggoro Wiradinata, Yosua Soekamto, Theresia Ratih (2023), «House Price Prediction Model Using Random Forest in Surabaya City»
Διαθέσιμο:
https://www.temjournal.com/content/121/TEMJournalFebruary2023_126_132.pdf
5) Santosh Yadaw (2020), «Predicting Housing Prices Using Scikit-Learn's Random Forest Model»
Διαθέσιμο:
https://towardsdatascience.com/predicting-housing-prices-using-a-scikit-learns-random-forest-model-e736b59d56c5
6) Raul-Tomas Mora-Garcia, Maria-Francisca Cespedes-Lopez και Raul Perez-Sanchez (2022), «Housing Price Prediction Using Machine Learning Algorithms in COVID-19 Times», MDPI
Διαθέσιμο:
https://www.mdpi.com/2073-445X/11/11/2100#B5-land-11-02100

Έρευνες με τη χρήση αλγορίθμων Boosting
1) J.Avanijaa, Gurram Sunitha b, K.Reddy Madhavi c, Padmavathi Korad και R.Hitesh Sai Vittal (2021), «Prediction of House Price Using XGBoost Regression Algorithm», Research Gate.
Διαθέσιμο:
https://www.researchgate.net/publication/350810698_Prediction_of_House_Price_Using_XGBoost_Regression_Algorithm
2) Byeonghwa, P., Jae Kwon, B. (2015), «Using machine learning algorithms for housing price prediction: The case of Fairfax Country housing data», Expert Systems with applications, Vol. 42, No. 6, 2928-2934.
3) Anders Hjort (2021), «House price prediction with gradient boosted trees under different loss functions»
Διαθέσιμο:
https://www.tandfonline.com/doi/full/10.1080/09599916.2022.2070525
4) B.Vijay Kumar, B.Ashritha, CH.Teja, M.Vineeth (2020), «HOUSE PRICE PREDICTION USING GRADIENT BOOST REGRESSION MODEL»
Διαθέσιμο:
https://www.ijrar.org/papers/IJRAR2001569.pdf
5) Ismail Ibrahim (2021), «Advanced Machine Learning Algorithms for House Price Prediction: Case Study in Kuala Lumpur»
Διαθέσιμο:
https://thesai.org/Downloads/Volume12No12/Paper_91-Advanced_Machine_Learning_Algorithms.pdf
6) N. Ragapriya, T. Ananth Kumar, R. Parthiban, P. Divya, S. Jayalakshmi, D. Raghu Raman (2023), «Machine Learning Based House Price Prediction Using Modified Extreme Boosting»
Διαθέσιμο:
https://ajast.net/data/uploads/83465.pdf
7) Marco Febriadi Kokasih, Adi Suryaputra Para (2020), «Property Rental Price Prediction Using the Extreme Gradient Boosting Algorithm», International Journal of Informatics and Information System, Vol. 3, No. 2, pp. 54-59
8) Anders Hjort, Johan Pensar, Ida Scheel, Dag Einar Sommervoll (2022), «House price prediction with gradient boosted trees under different loss functions»
Διαθέσιμο:
https://www.duo.uio.no/bitstream/handle/10852/94434/4/10-1080-09599916-2022-2070525.pdf