**A close up of text

Description automatically generated**

**ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ**

**ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ**

**ΤΜΗΜΑ ΔΙΟΙΚΗΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ**

**ΕΠΙΧΕΙΡΗΜΑΤΙΚΗ ΕΥΦΥΪΑ ΚΑΙ ΑΝΑΛΥΣΗ ΜΕΓΑΛΩΝ ΔΕΔΟΜΕΝΩΝ**

**ΕΡΓΑΣΙΑ ΕΞΑΜΗΝΟΥ:**

**Data Warehousing and Data Mining**

**Επιμέλεια εργασίας:**

**Ιωάννης Καλαντζής (8200235)**

**Οδυσσέας**

**Καθηγητής: Χατζηαντωνίου Δαμιανός**

**Αθήνα, Ιανουάριος 2024**

**Data Extraction and Cleaning**

Για την εύρεση των δεδομένων μας, χρησιμοποιήσαμε τον ιστότοπο της Kaggle (<https://www.kaggle.com/>). Συγκεκριμένα, εργαστήκαμε πάνω σε ένα dataset, το οποίο αποτελείται από καταχωρήσεις κατοικιών-ακινήτων, στις Ηνωμένες Πολιτείες της Αμερικής. Ο σύνδεσμος για το συγκεκριμένο online dataset είναι ο ακόλουθος: <https://www.kaggle.com/datasets/austinreese/usa-housing-listings?fbclid=IwAR3-tif1QU43u4R1AbmmuK9kZboVGdYfgAG3AGTJK_yEm2b3aK_ZLHbhF6U>.

Το παρόν σετ δεδομένων αποτελείται από **385.000 instances** και **22 features**. Πιο συγκεκριμένα, τα δεδομένα που μας παρέχει το dataset είναι τα εξής:

1. **id**: Αναγνωριστικό για κάθε καταχώρηση.
2. **region**: Περιοχή του ακινήτου.
3. **price**: Τιμή ενοικίασης.
4. **type**: Τύπος του ακινήτου(π.χ. διαμέρισμα, σπίτι κλπ.).
5. **sqfeet**: Τετραγωνικά πόδια του ακινήτου.
6. **beds**: Αριθμός υπνοδωματίων.
7. **baths**: Αριθμός μπάνιων.
8. **cats\_allowed**: Εάν επιτρέπονται οι γάτες(δυαδική τιμή, 1: Ναι και 0: Όχι).
9. **dogs\_allowed**: Εάν επιτρέπονται οι σκύλοι(δυαδική τιμή, 1: Ναι και 0: Όχι).
10. **wheelchair\_access**: Εάν υπάρχει πρόσβαση για άτομα με αναπηρικά αμαξίδια(δυαδική τιμή, όπου 1: Ναι και 0: Όχι).
11. **electric\_vehicle\_charge**: Εάν υπάρχει φόρτιση για ηλεκτρικά οχήματα(δυαδική τιμή, 1: Ναι και 0: Όχι).
12. **comes\_furnished**: Εάν προσφέρεται επιπλωμένο(δυαδική τιμή, 1: Ναι και 0: Όχι).
13. **laundry\_options**: Επιλογή πλυντηρίου(π.χ. w/d in unit, w/d hookups, laundry on site κλπ.).
14. **parking\_options**: Επιλογές στάθμευση (π.χ. carport, attached garage, off-street parking κλπ.).
15. **smoking\_allowed**: Εάν επιτρέπεται το κάπνισμα(δυαδική τιμή, 1: Ναι και 0: Όχι).
16. **lat**: Γεωγραφικός πλάτος.
17. **long**: Γεωγραφικό μήκος.
18. **state**: Πολιτεία καταχώρησης.
19. **description**: Περιγραφή.
20. **image\_url**: Το URL της εικόνας της καταχώρησης.
21. **url**: Το URL της καταχώρησης.
22. **region\_url**: Το URL της περιοχής.

Προφανώς, προτού εισάγουμε τα δεδομένα μας στο Data Warehouse, πρέπει να εξετάσουμε τις τιμές των βασικών features που θα χρειαστούμε για την ανάλυσή μας, να αφαιρέσουμε εκείνα τα features που δεν θα μας ωφελήσουν περαιτέρω. Επιπλέον, θα αφαιρέσουμε ακραίες (outliers) και κενές (null) τιμές.

Για τον σκοπό αυτό χρησιμοποιήσαμε το Jupyter Notebook, ένα εργαλείο open-source, το οποίο είναι δημοφιλές για την ανάλυση δεδομένων και την διαχείρισή τους.

Αφού πλοηγηθήκαμε και εξοικειωθήκαμε με τα δεδομένα, τις στήλες και τις γραμμές του dataset μας, δημιουργήσαμε το αντίστοιχο DataFrame και ξεκινήσαμε την διαδικασία καθαρισμού και επεξεργασίας.

A screenshot of a computer

Description automatically generated

Στην συνέχεια, έπρεπε να επεξεργαστούμε και να αφαιρέσουμε τις κενές τιμές, δηλαδή όσες τιμές είναι ΝaN.

**A screenshot of a computer

Description automatically generated**

Από το παραπάνω αποτέλεσμα, παρατηρούμε ότι οι στήλες που έχουν κενές τιμές είναι οι στήλες laundry\_options, parking\_options, description, lat και long. Ας δούμε και το άθροισμα αυτών των τιμών για να μπορέσουμε να τις διαχειριστούμε αναλόγως.

**A screenshot of a computer

Description automatically generated**

Έτσι, επικεντρωθήκαμε στις στήλες laundry\_options και parking\_options, οι οποίες είχαν και τις περισσότερες τιμές NaN. Προκειμένου να μην διαγράψουμε όλες αυτές τις τιμές από το σετ δεδομένων μας, γιατί θα μίκραινε αρκετά σε μέγεθος, αποφασίσαμε να τις μετατρέψουμε σε ‘Unknown’. Τις στήλες long και lat, επειδή είχαν τόσο λίγες παρατηρήσεις με κενές τιμές, αποφασίσαμε να τις αφαιρέσουμε τελείως. Την στήλη με τις περιγραφές, την αφαιρούμε εξολοκλήρου από τα δεδομένα μας, αφού, δεν θα την χρειαστούμε περαιτέρω.

**A screenshot of a computer

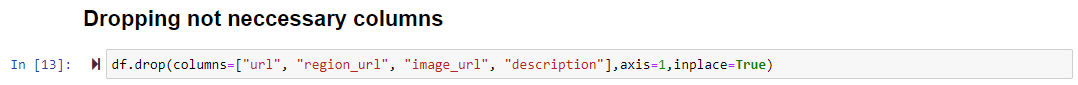
Description automatically generated**

Επιπλέον, πραγματοποιήσαμε ελέγχους στις στήλες που το θεωρήσαμε αναγκαίο, όπως για παράδειγμα στις στήλες με δυαδικές τιμές για να βεβαιωθούμε ότι όλες οι μεταβλητές με δυαδικές τιμές είναι πράγματι δυαδικές.

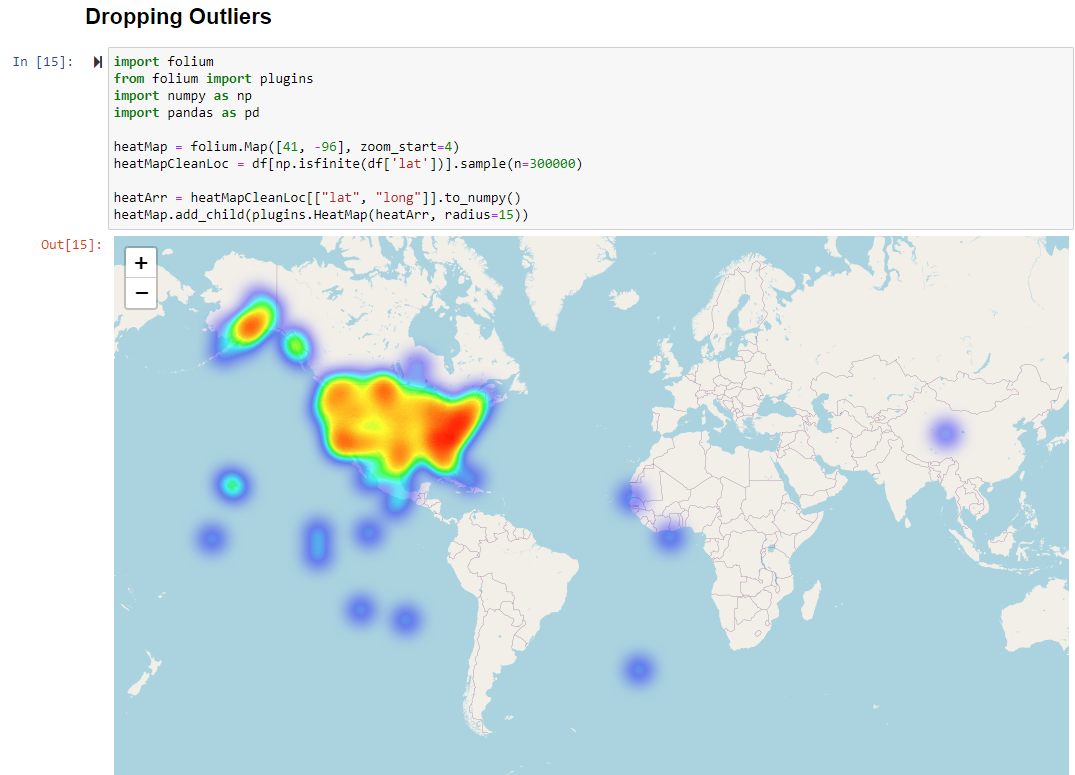
A computer screen shot of a computer code

Description automatically generated

Επίσης, αφαιρέσαμε στήλες οι οποίες, όπως αναφέραμε και παραπάνω, δεν θα ευνοήσουν την ανάλυσή μας.

****

Έπειτα, με την βοήθεια ενός HeatMap, όπως φαίνεται και παρακάτω, παρατηρήσαμε ότι μερικές συντεταγμένες δεν αποτυπώνονται στον χάρτη πάνω στις ΗΠΑ, αλλά σε διαφορετικές χώρες του κόσμου. Αυτό οφείλεται σε ορισμένες ακραίες τιμές, οι οποίες έχουν καταχωρηθεί στα δεδομένα μας λάθος και πρέπει να αφαιρεθούν.

****

Ουσιαστικά, η διαδικασία που ακολουθούμε βασίζεται στην απόρριψη όσων τιμών δεν βρίσκονται στα προσδοκόμενα σύνορα, με βάση το γεωγραφικό πλάτος και μήκος των ΗΠΑ.

**A computer screen shot of a computer code

Description automatically generated**

Πράγματι, αφαιρέθηκαν όλες οι καταχωρήσεις των οποίων οι γεωγραφικές συντεταγμένες δεν ενθυλακώνονταν στις ΗΠΑ, όπως φαίνεται παρακάτω.

**A map of the world

Description automatically generated**

Τέλος, θα αφαιρέσουμε τις ακραίες τιμές από τις βασικές μας στήλες με αριθμητικές τιμές. Αυτές είναι τα δωμάτια(beds), τα μπάνια(baths), η τιμή(price) και τα τετραγωνικά πόδια(sqfeet). Για τον σκοπό αυτό δημιουργήσαμε τον παρακάτω κώδικα. Η τεχνική που ακολουθήσαμε για την αφαίρεση των ακραίων τιμών βασίζεται στο Interquartile Range(IQR) και ονομάζεται IQR rule ή IQR method. Αυτή η τεχνική χρησιμοποιεί τα προκαθορισμένα όρια και , για να αποφασίσει ποιες τιμές θεωρούνται ακραίες.

**A screenshot of a computer program

Description automatically generated**

****

Τελικά, από το αρχικό σετ δεδομένων μας διαγράψαμε 42.000 γραμμές μέσω της εκαθάρησης και της επεξεργασίας.

**A white rectangular object with text

Description automatically generated with medium confidence**

**Data Mining Tasks**

Για την διαδικασία του Data Mining, αποφασίσαμε να εργαστούμε με τα open-source εργαλεία Jupyter Notebook και RapidMiner. Το Jupyter Notebook έχει χρησιμοποιηθεί και προηγουμένως στην παρούσα εργασία για τον καθαρισμό και την διαχείριση των δεδομένων, ενώ το RapidMiner είναι ένα επιπλέον open-source εργαλείο που χρησιμοποιείται για την ανάλυση δεδομένων και την εξόρυξη γνώσης. Οι λειτουργίες εξόρυξης δεδομένων που χρησιμοποιήθηκαν είναι η μέθοδος της παλινδρόμησης, η κατηγοριοποίηση, καθώς και η συσταδοποίηση.

1. **Μέθοδοι Παλινδρόμησης**

Για την πρώτη λειτουργία εξόρυξης, αποφασίσαμε να εστιάσουμε σε διάφορες μορφές παλινδρόμησης, δημιουργώντας προβλεπόμενες τιμές, και στις συσχετίσεις μεταξύ της τιμής μιας καταχώρησης και των παροχών της. Ας εξηγήσουμε, όμως, τον τρόπο με τον οποίο εργαστήκαμε.

Αρχικά, φορτώσαμε και διαβάσαμε σε ένα Jupyter Notebook αρχείο τα καθαρισμένα και επεξεργασμένα δεδομένα μας.

A screenshot of a computer

Description automatically generated

Ύστερα, προκειμένου να μπορέσουμε να εφαρμόσουμε τις τεχνικές παλινδρόμησης, θα χρειαστεί να προσαρμόσουμε τα δεδομένα μας καταλλήλως. Γι’ αυτόν τον σκοπό, θα χρησιμοποιήσουμε τη βιβλιοθήκη scikit-learn(sklearn) για την επεξεργασία των δεδομένων. Συγκεκριμένα, χρησιμοποιείται ο LabelEncoder από το sklearn.preprocessing για τη μετατροπή των *κατηγορικών* μεταβλητών σε *αριθμητικές* τιμές. Επίσης, αποθηκεύουμε τα αρχικά states σε έναν πίνακα, μαζί με τις καινούργιες αριθμητικές τιμές, γιατί θα τα χρειαστούμε στην συνέχεια της ανάλυσής μας.

A screen shot of a computer program

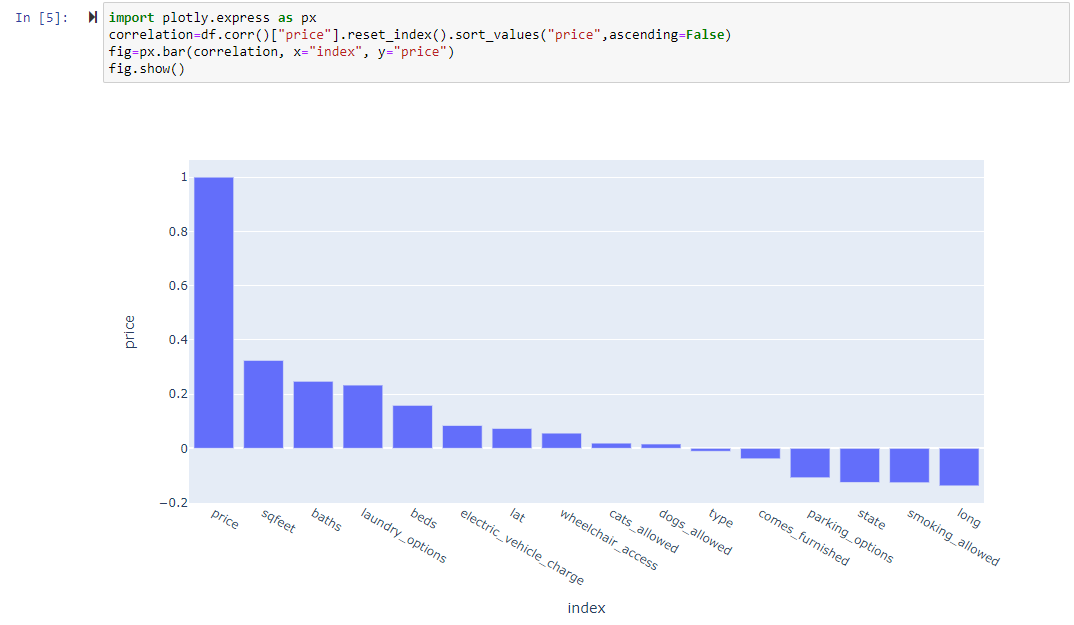
Description automatically generated

A white rectangular object with text

Description automatically generated



Στην συνέχεια, ο κώδικας χρησιμοποιεί τη βιβλιοθήκη Plotly Express για να δημιουργήσει ένα διάγραμμα με μπάρες που απεικονίζει τη συσχέτιση των χαρακτηριστικών με τη μεταβλητή "price", δηλαδή την τιμή κάθε καταχώρησης.

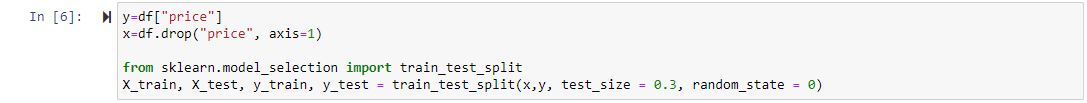


Από το παραπάνω διάγραμμα, συγκεντρώνουμε πληροφορία για την συσχέτιση μεταξύ της τιμής μίας καταχώρησης και των λοιπών παροχών της. Προκειμένου να μην εστιάσουμε ξεχωριστά σε κάθε μία από τις παροχές, μπορούμε να τις ομαδοποιήσουμε και να δημιουργήσουμε κατηγορίες ανάλογα με την ισχύ της συσχέτισης. Έτσι, έχουμε τις εξής κατηγορίες:

1. **Ισχυρή συσχέτιση**: Στην συγκεκριμένη κατηγορία εντάσσονται τα τετραγωνικά πόδια(sqfeet) του κτίσματος, τα baths και αν είναι διαθέσιμες οι επιλογές πλυντηρίου, όπου η τιμή της συσχέτισης είναι μεγαλύτερη από ().
2. **Μέτρια συσχέτιση**: Σε αυτή την κατηγορία εντάσσεται ο αριθμός των υπνοδωματίων, το γεωγραφικό πλάτος, η παροχή φορτιστή για ηλεκτρικά αυτοκίνητα, η προσβασιμότητα για άτομα με αναπηρίες, καθώς και η άδεια κατοχής κατοικίδιων. Η τιμή της συσχέτισης είναι μεγαλύτερη του και μικρότερη το ().
3. **Αρνητική συσχέτιση**: Στην συγκεκριμένη κατηγορία ενθυλακώνεται ο τύπος της καταχώρησης, η επίπλωση, το parking, η πολιτεία, η άδεια για κάπνισμα και το γεωγραφικό μήκος. Η τιμή της συσχέτισης φαίνεται να έχει τιμές μικρότερες του και μεγαλύτερες του ().

Προφανώς, η συσχέτιση δείχνει την σχέση που υπάρχει μεταξύ των εκάστοτε παροχών και της τιμής. Έτσι, το παραπάνω πόρισμα, είναι αρκετά ωφέλιμο για τους επενδυτές και τους κατόχους κατοικιών οι οποίοι επιθυμούν να δημιουργήσουν μια καταχώρηση και να νοικιάσουν τον χώρο τους. Γνωρίζοντας την συσχέτιση μεταξύ τιμών και παροχών, μπορούν να βελτιστοποιήσουν την ενοικίαση του διαμερίσματός τους, επιτρέποντας, για παράδειγμα, την κατοχή κατοικίδιων, ή εξοπλίζοντας τον χώρο με φορτιστή για ηλεκτρικά αμάξια(θετική συσχέτιση). Επίσης, μπορούν να εστιάσουν στην μετατροπή των παροχών οι οποίες μειώνουν την τιμή, όπως για παράδειγμα, να μην επιπλώνουν τον χώρο ή να μην επιτρέπουν το κάπνισμα(αρνητική συσχέτιση).

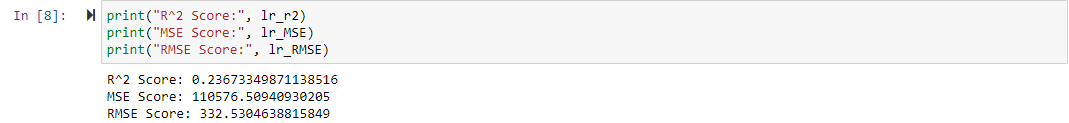
Τώρα, προχωρώντας με τις παλινδρομήσεις, ο παραπάνω κώδικας διαχωρίζει το σύνολο δεδομένων σε σύνολα εκπαίδευσης και ελέγχου, όπου οι μεταβλητές εξαρτώμενης και ανεξάρτητης μεταβλητής ορίζονται ως y και x αντίστοιχα. Στη συνέχεια, χρησιμοποιείται η συνάρτηση train\_test\_split από τη βιβλιοθήκη scikit-learn για να διαχωριστούν τα δεδομένα σε τυχαία εκπαιδευτικά και δοκιμαστικά σύνολα, όπου το 30% των δεδομένων προορίζεται για το δοκιμαστικό σύνολο, ενώ το random\_state ορίζει μια σταθερά για την αναπαραγωγή των αποτελεσμάτων.



Ο παρακάτω κώδικας εκπαιδεύει ένα μοντέλο *γραμμικής παλινδρόμησης* στα δεδομένα εκπαίδευσης και στη συνέχεια το χρησιμοποιεί για να προβλέψει τις τιμές στα δεδομένα ελέγχου, υπολογίζοντας το R^2, MSE και RMSE για να αξιολογήσει την απόδοση του μοντέλου.

A white background with black text

Description automatically generated



Θα δημιουργήσουμε τον αντίστοιχο πίνακα με τις κανονικές τιμές και τις προβλέψεις που προκύπτουν από το μοντέλο της παλινδρόμησης, καθώς και ένα διάγραμμα που θα αναπαριστά τις τιμές αυτές.

A screenshot of a computer

Description automatically generated

A screen shot of a graph

Description automatically generated

Τα αποτελέσματα των μοντέλων παλινδρόμησης θα εξεταστούν συνολικά αφού εφαρμοστούν όλα. Έτσι, θα παραθέσουμε τις αντίστοιχες τεχνικές, όπως ακολουθούνται και παραπάνω, για κάθε μοντέλο. Στην συνέχεια, θα χρησιμοποιήσουμε το *μοντέλο παλινδρόμησης XGBoost* (Extreme Gradient Boosting) για να εκπαιδεύσουμε ένα μοντέλο πρόβλεψης. Το XGBoost είναι ένα δημοφιλές μοντέλο μηχανικής μάθησης που χρησιμοποιείται για προβλέψεις σε προβλήματα παλινδρόμησης και ταξινόμησης.

A white background with black text

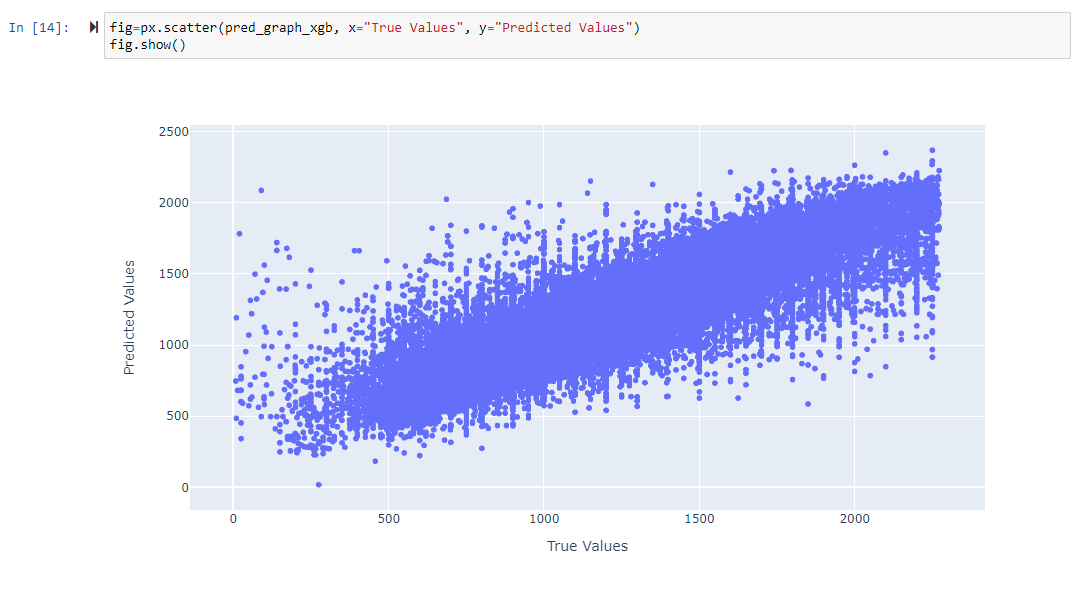
Description automatically generated

A white rectangular object with a black border

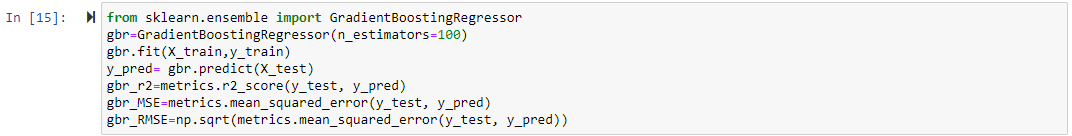
Description automatically generated

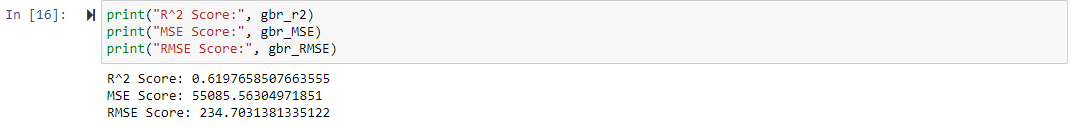
A screenshot of a computer

Description automatically generated



Στον παρακάτω κώδικα χρησιμοποιείται το *μοντέλο παλινδρόμησης Gradient Boosting*, συγκεκριμένα η κλάση GradientBoostingRegressor από τη βιβλιοθήκη scikit-learn, για να εκπαιδευτεί ένα μοντέλο πρόβλεψης. Το Gradient Boosting είναι μια τεχνική αναβάθμισης δέντρων αποφάσεων, όπου τα δέντρα εκπαιδεύονται στη σειρά και κάθε δέντρο επιδιορθώνει τα σφάλματα του προηγούμενου.





A screenshot of a computer

Description automatically generated

A screen shot of a graph

Description automatically generated

Τέλος, ο παραπάνω κώδικας χρησιμοποιεί το *μοντέλο παλινδρόμησης RandomForestRegressor* από τη βιβλιοθήκη scikit-learn για να εκπαιδεύσει ένα μοντέλο πρόβλεψης. Το RandomForestRegressor χρησιμοποιεί ένα σύνολο τυχαίων δέντρων αποφάσεων και συνδυάζει τις προβλέψεις τους για να παράγει μια τελική πρόβλεψη.

A white background with black text

Description automatically generated

A white background with a black and white flag

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

A screen shot of a graph

Description automatically generated

A screenshot of a computer code

Description automatically generated

Από τον τελευταίο πίνακα, ο οποίος έχει όλα τα δεδομένα από όλες τις παλινδρομήσεις, σύμφωνα με τις μετρικές , και που παρέχονται, το μοντέλο που θα ακολουθήσουμε είναι το Random Forest Regressor(RAN). Ο λόγος είναι ότι το Random Forest Regressor έχει το υψηλότερο R^2(0.902279), το χαμηλότερο MSE(14157.106064) και το χαμηλότερο RMSE(118.983638) σε σύγκριση με τα άλλα μοντέλα. Αυτό υποδεικνύει ότι το μοντέλο Random Forest Regressor έχει την καλύτερη ικανότητα πρόβλεψης της μεταβλητής εξαρτώμενης σε σχέση με τα άλλα μοντέλα.

Τελικά, υπολογίσαμε την ποσοστιαία διαφορά μεταξύ της πραγματικής τιμής και της προβλεπόμενης.

A screenshot of a computer

Description automatically generated

Εστιάζοντας στα τρία πρώτα και τα τρία τελευταία, παρατηρούμε τα εξής:

1. Missouri (MO): Η προβλεπόμενη τιμή είναι 5.26% υψηλότερη από την πραγματική τιμή.
2. West Virginia (WV): Η προβλεπόμενη τιμή είναι 3.49% υψηλότερη από την πραγματική τιμή.
3. Wyoming (WY): Η προβλεπόμενη τιμή είναι 2.52% υψηλότερη από την πραγματική τιμή.

Και τα τρία πρώτα παραδείγματα δείχνουν υψηλότερη πρόβλεψη από την πραγματική τιμή, πράγμα που μπορεί να είναι σημαντικό για έναν επενδυτή ή ιδιοκτήτη κατοικίας, καθώς μπορεί να επηρεάσει τις αποφάσεις τους σχετικά με την ενοικίαση του χώρου του.

1. Nebraska (NE): Η προβλεπόμενη τιμή είναι 0.36% χαμηλότερη από την πραγματική τιμή.
2. Wisconsin (WI): Η προβλεπόμενη τιμή είναι 0.41% χαμηλότερη από την πραγματική τιμή.
3. Alaska (AK): Η προβλεπόμενη τιμή είναι 0.71% χαμηλότερη από την πραγματική τιμή.

Τα τρία τελευταία παραδείγματα δείχνουν ελαφρά υποεκτίμηση από το μοντέλο, οπότε ένας επενδυτής ή ιδιοκτήτης κατοικίας θα μπορούσε να λάβει υπόψη αυτήν την πληροφορία όταν καθορίζει την τιμή που είναι διατεθειμένος να πληρώσει για την απόκτηση μίας κατοικίας ή να δεχτεί για ένα ακίνητο που του ανήκει στις συγκεκριμένες πολιτείες.

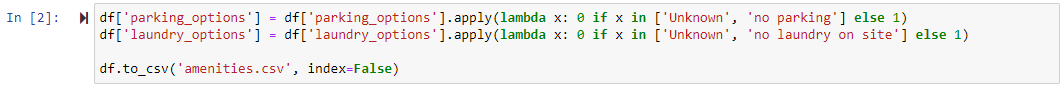
1. **Συσταδοποίηση**

Για την μέθοδο της συσταδοποίησης επιλέξαμε το open-source εργαλείο RapidMiner και δημιουργήσαμε την παρακάτω διαδικασία(process).

A screenshot of a computer

Description automatically generated

Ας εξηγήσουμε, λοιπόν, τον τρόπο με τον οποίο εργαστήκαμε στην συγκεκριμένη διαδικασία. Αρχικά, μέσω του Jupyter Notebook, δημιουργήσαμε ένα καινούργιο αρχείο μετατρέποντας όλες τις παροχές σε δυαδικές τιμές.



Έτσι, ανακτούμε τα δεδομένα μας και τα φορτώνουμε στην συγκεκριμένη διαδικασία. Ύστερα, επιλέγουμε τις κατάλληλες μεταβλητές, μέσω του Operator Select Attributes, που θα μας δώσουν πληροφορία για τις καταχωρήσεις. Οι μεταβλητές αυτές, όπως φαίνεται και παρακάτω, εστιάζουν στις παροχές των καταχωρήσεων.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Στην συνέχεια, πραγματοποιούμε ένα k-Means Clustering(συσταδοποίηση), και περνάμε τα αποτελέσματα αυτά στους Operators: Performance και Cluster Model Visualizer. Τελικά, επικεντρωνόμαστε στα παρακάτω αποτελέσματα:

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

Όπως παρατηρούμε, οι συστάδες(clusters) με τα περισσότερα items είναι το cluster 0, το cluster 1, το cluster 3, το cluster 4 και το cluster 9. Αναλύοντας τα ποσοστά εμφάνισης τα οποία παρουσιάζονται παραπάνω, καταλήγουμε στα εξής:

1. **Barren Listings**(Cluster 0): Σε αυτή την συστάδα ανήκουν οι καταχωρήσεις που δεν προσφέρουν καμία παροχή. Το ποσοστό των συγκεκριμένων καταχωρήσεων ανέρχεται στο .
2. **Pet-Friendly Haven with Laundry & Parking Amenities** (Cluster 1): Σε αυτή την συστάδα ανήκουν όλοι οι χώροι που παρέχουν την άδεια κατοχής σκύλου και γάτας, καθώς και πλυντήριο ρούχων και θέση πάρκινγκ. Ποσοστό εμφάνισης: .
3. **Pet-Friendly Nest with Laundry Option & ΝΟ Parking** (Cluster 3): Η ακριβώς ίδια κατηγορία με προηγουμένως, απλά στην συγκεκριμένη δεν συμπεριλαμβάνεται η θέση πάρκινγκ. Ποσοστό εμφάνισης: .
4. **Pet Heaven** (Cluster 4): Η συστάδα ενθυλακώνει καταχωρήσεις που συμπεριλαμβάνουν σε ποσοστό 100% την δυνατότητα κατοχής γάτας και σκύλου. Ποσοστό εμφάνισης: 7%.
5. **Convenient Laundry & Parking Hub** (Cluster 9): Η συστάδα περιλαμβάνει τις παροχές του πλυντηρίου και της θέσης πάρκινγκ. Το ποσοστό εμφάνισης είναι 9.1%.

Για να αναλύσουμε και την επιχειρηματική αξία, την οποία μας προσφέρουν οι παραπάνω συστάδες, καταλήγουμε στις ακόλουθες πληροφορίες για κάποιον επενδυτή ή κάτοχο ακινήτου:

* **Ανάλυση Παροχών**: Είναι αρκετά εύκολο να διαπιστωθούν οι παροχές εκείνες που εμφανίζονται στις περισσότερες συστάδες. Για παράδειγμα, φαίνεται ότι η παροχή πλυντηρίου εμφανίζεται σε πολλά clusters και συνδυάζεται ποικιλοτρόπως με άλλες παροχές.
* **Ζήτηση για Κατοικίδια**: Μια ακόμα παρατήρηση είναι αυτή της ζήτησης χώρων στους οποίους επιτρέπονται τα κατοικίδια ζώα. Φαίνεται, από τα αποτελέσματά μας, ότι η συγκεκριμένη παροχή εμφανίζεται στις περισσότερες συστάδες με μεγάλα ποσοστά εμφάνισης.
* **Συνολική Αξία**: Παρατηρούμε ότι αρκετές παροχές δεν συμπεριλαμβάνονται στις κύριες συστάδες. Οπότε, προκειμένου να αυξήσει την ζήτηση και την τιμή του ακινήτου του, κάποιος ιδιοκτήτης θα μπορούσε να συμπεριλάβει και άλλες παροχές, όπως να βελτιώσει την προσβασιμότητα για άτομα σε αναπηρικά αμαξίδια ή να εξοπλίσει τον χώρο με κάποιο φορτιστή για ηλεκτρικά αυτοκίνητα.
* **Ανάλυση Ανταγωνισμού**: Συνυφασμένος με την συνολική αξία του ακινήτου είναι ο ανταγωνισμός. Ένας ιδιοκτήτης μπορεί να παρατηρήσει τις σχέσεις μεταξύ των παροχών και τις ελλείψεις που υπάρχουν στην αγορά, δημιουργώντας νέες επενδυτικές ευκαιρίες.