ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΕΠΙΧΕΙΡΗΜΑΤΙΚΗ ΕΥΦΥΪΑ

ΙΩΑΝΝΗΣ ΤΖΙΟΓΚΑΣ
ΑΜ: 20135

Περιγραφική Ανάλυση

Σκοπός της παρούσας εργασίας είναι η διερεύνηση των παραγόντων που ενδέχεται να επηρεάζουν τη διαμόρφωση της τιμής ενοικίασης ενός καταλύματος Airbnb στην Ελλάδα. Τα δεδομένα που χρησιμοποιήθηκαν και υπήρχαν διαθέσιμα από την ίδια την υπηρεσία για διάφορες πόλεις του κόσμου (http://insideairbnb.com/get-the-data.html) αφορούν την Αθήνα, τη Θεσσαλονίκη και την Κρήτη για τον μήνα Οκτώβριο 2020.

Οι παράγοντες που θα ελεγχθούν για το αν επηρεάζουν την τιμή ενοικίασης είναι:

- η περιοχή (Αθήνα, Θεσσαλονίκη, Κρήτη)
- το είδος καταλύματος
- ο αριθμός των καταλυμάτων που έχει ένας ιδιοκτήτης στην κατοχή του
- η επισκεψιμότητα που έχει κάθε κατάλυμα (προσέγγιση με τη μεταβλητή Reviews per month)
- οι διαθέσιμες ημέρες το χρόνο (προορισμοί για όλο τον χρόνο και εποχικοί προορισμοί)
- η απόσταση από κάποιο από τα Top 10 αξιοθέατα σύμφωνα με το TripAdvisor (μοντέλο με συνεχή μεταβλητή την ελάχιστη απόσταση από τα Top 10 αξιοθέατα)

Στη συνέχεια, θα γίνει ανάλυση ξεχωριστά στις τρεις περιοχές ώστε να ελεγχθούν επιπλέον κάποιοι τοπικοί παράγοντες. Αυτοί είναι:

- η γειτονιά (τα 7 Δημοτικά Διαμερίσματα (ΔΔ) για την Αθήνα, οι γειτονιές στις οποίες ανήκουν τα καταλύματα στη βάση δεδομένων της υπηρεσίας για τη Θεσσαλονίκη και οι 4 νομοί για την Κρήτη)
- η απόσταση από τα τοπικά μέρη ενδιαφέροντος (οι στάσεις του μετρό για την Αθήνα, τα σημαντικότερα μουσεία της πόλης για τη Θεσσαλονίκη και οι παραλίες με γαλάζιες σημαίες για την Κρήτη καθώς αποτελεί θερινό προορισμό κατά κύριο λόγο)

Διαστάσεις και επίπεδα

Χώρος

Νομος, περιοχή , στην Αθήνα Δημοτικά διαμερίσματα.

Καταλύματα
Είδος καταλύματος,ο αριθμός των καταλυμάτων που έχει ένας ιδιοκτήτης στην κατοχή του,η επισκεψιμότητα που έχει κάθε κατάλυμα (προσέγγιση με τη μεταβλητή Reviews per month), οι διαθέσιμες ημέρες το χρόνο (προορισμοί για όλο τον χρόνο και εποχικοί προορισμοί),η απόσταση από κάποιο από τα Top 10 αξιοθέατα σύμφωνα με το TripAdvisor (μοντέλο με συνεχή μεταβλητή την ελάχιστη απόσταση από τα Top 10 αξιοθέατα),
Στην Αθήνα:Απόσταση από τις στάσεις του μετρό στην Αθήνα, την τιμή του καταλύματος
Στην Θεσσαλονικη:Απόσταση από μουσεία της Θεσσαλονίκης
Στην Κρήτη: Απόσταση από παραλίες με γαλάζια σημαία

Μετρικές
Αθροίσματα
Μέσες τιμές
Ποσοστά
Καταμέτρηση
Χιλιομετρική απόσταση
Ελάχιστη τιμή

Ο αριθμός των καταλυμάτων Airbnb που είναι καταγεγραμμένα σε Αθήνα, Θεσσαλονίκη και Κρήτη είναι 29.601. Η μέση τιμή ενοικίασης ενός καταλύματος ανά ημέρα είναι 123,8 €, ενώ η διάμεση τιμή ενοικίασης είναι τα 57 €. Το εύρος των τιμών κυμαίνεται από 8 έως 20.653€, ωστόσο το 75% των καταλυμάτων ενοικιάζονται με τιμή έως 100 ευρώ ανά ημέρα. Οι τιμές διαφοροποιούνται ανάλογα με την περιοχή και οι κατανομές των τιμών παρουσιάζονται στη συνέχεια ξεχωριστά για Αθήνα, Θεσσαλονίκη και Κρήτη (Πίνακας 1).

Τα διαφορετικά καταλύματα διαφέρουν ως προς το είδος τους, την επισκεψιμότητά τους (η οποία μπορεί να προσεγγιστεί από τον αριθμό μηνιαίων κριτικών καθώς μπορούμε να υποθέσουμε πως τα καταλύματα με τις περισσότερες κριτικές έχουν και μεγαλύτερη επισκεψιμότητα) και τη διαθεσιμότητά τους κατά τη διάρκεια του χρόνου (Πίνακας 1). Κάθε ιδιοκτήτης έχει στην κατοχή του διαφορετικό αριθμό καταλυμάτων, με μέσο αριθμό καταλυμάτων ανά ιδιοκτήτη τα 14 καταλύματα, ενώ το 50% των ιδιοκτητών έχουν έως 3 καταλύματα.

Πίνακας 1: Περιγραφή της κατανομής των μεταβλητών ενδιαφέροντος για τα καταλύματα στην Ελλάδα και ανά περιοχή (Αθήνα, Θεσσαλονίκη, Κρήτη)

	Μέση τιμή (τυπική απόκλιση)	Διάμεση τιμή (25ο-75ο τετ.*)	Εύρος
Ελλάδα			
Τιμή ενοικίασης (€)	123,8 (471,0)	57,0 (38,0-100,0)	8,0-20.653,0
Μέσος αριθμός κριτικών ανά μήνα	1,0 (1,3)	0,5 (0,2-1,3)	0,0-14,2
Αριθμός καταλυμάτων ανά ιδιοκτήτη	14,3 (32,5)	3,0 (1,0-8,0)	1,0-190,0
Ετήσια διαθεσιμότητα (ημέρες)	250,9 (118,7)	295,0 (174,0-362,0)	0,0-365,0
Ελάχιστη απόσταση από Top10 αξιοθέατα (χλμ)	7,6 (10,4)	2,2 (0,6-11,5)	0,0-54,7
Αθήνα			
Τιμή ενοικίασης (€)	70,1 (416,0)	43,0 (30,0-65,0)	8,0-20.653
Μέσος αριθμός κριτικών ανά μήνα	1,6 (1,7)	1,0 (0,3-2,4)	0,0-14,2
Αριθμός καταλυμάτων ανά ιδιοκτήτη	9,4 (17,5)	2,0 (1,0-8,0)	1,0-101,0
Ετήσια διαθεσιμότητα (ημέρες)	254,1 (121,7)	316,0 (168,0-360,0)	0,0-365,0
Ελάχιστη απόσταση από Top10 αξιοθέατα (χλμ)	0,9 (0,7)	0,7 (0,4-1,1)	0,0-5,1
Θεσσαλονίκη			
Τιμή ενοικίασης (€)	58,1 (235,2)	40,0 (30,0-55,0)	9,0-8.000,0
Μέσος αριθμός κριτικών ανά μήνα	1,5 (1,5)	1,0 (0,4-2,0)	0,0-11,1
Αριθμός καταλυμάτων ανά ιδιοκτήτη	5,3 (9,6)	2,0 (1,0-5,0)	1,0-55,0
Ετήσια διαθεσιμότητα (ημέρες)	231,0 (127,4)	266,0 (118,0-358,0)	0,0-365,0
Ελάχιστη απόσταση από Top10 αξιοθέατα (χλμ)	1,1 (1,3)	0,5 (0,3-1,2)	0,0-10,8
Κρήτη			
Τιμή ενοικίασης (€)	161,5 (517,1)	71,0 (46,0-136,0)	9,0-8.000,0
Μέσος αριθμός κριτικών ανά μήνα	0,6 (0,7)	0,4 (0,1-0,8)	0,0-12,3
Αριθμός καταλυμάτων ανά ιδιοκτήτη	18,1 (39,4)	3,0 (1,0-9,0)	1,0-190,0
Ετήσια διαθεσιμότητα (ημέρες)	251,8 (115,6)	274,0 (179,0-363,0)	0,0-365,0
Ελάχιστη απόσταση από Top10 αξιοθέατα (χλμ)	12,1 (11,4)	9,7 (3,1-17,0)	0,0-54,7

*τεταρτημόριο

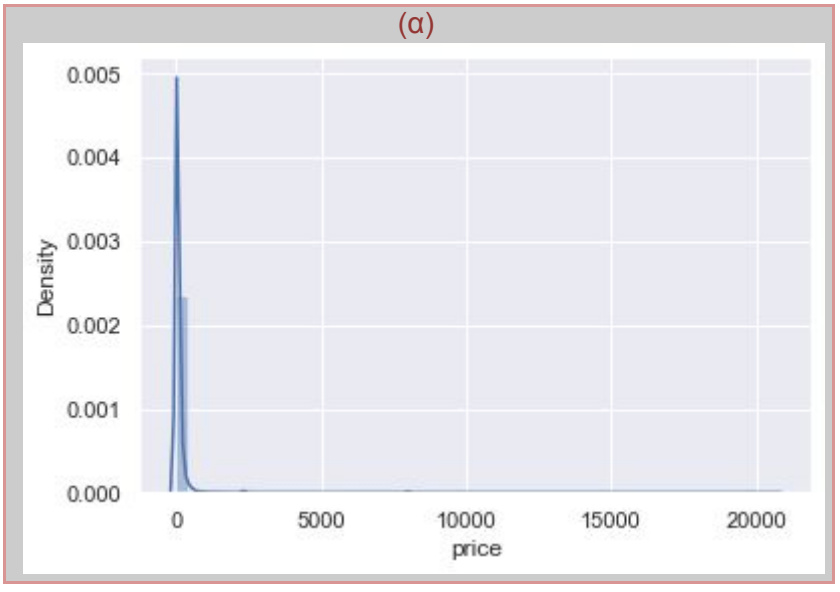
Όσον αφορά το είδος των καταλυμάτων, τα περισσότερα καταλύματα είναι ολόκληρα σπίτια/διαμερίσματα (88,3%), το 9,1% είναι ιδιωτικά δωμάτια ενώ μικρό ποσοστό είναι δωμάτια ξενοδοχείου (2,3%) και κοινόχρηστα δωμάτια (0,3%). Οι τιμές διαφοροποιούνται ανάλογα τον τύπο του καταλύματος (Πίνακας 2). Τα δωμάτια ξενοδοχείου έχουν κατά μέσο όρο υψηλότερη τιμή ενοικίασης και τα κοινόχρηστα δωμάτια την χαμηλότερη. 0RDXZC

Πίνακας 2: Μέσες και ενδιάμεσες τιμές ενοικίασης καταλυμάτων ανά τύπο καταλύματος στην Ελλάδα (Αθήνα, Θεσσαλονίκη, Κρήτη)

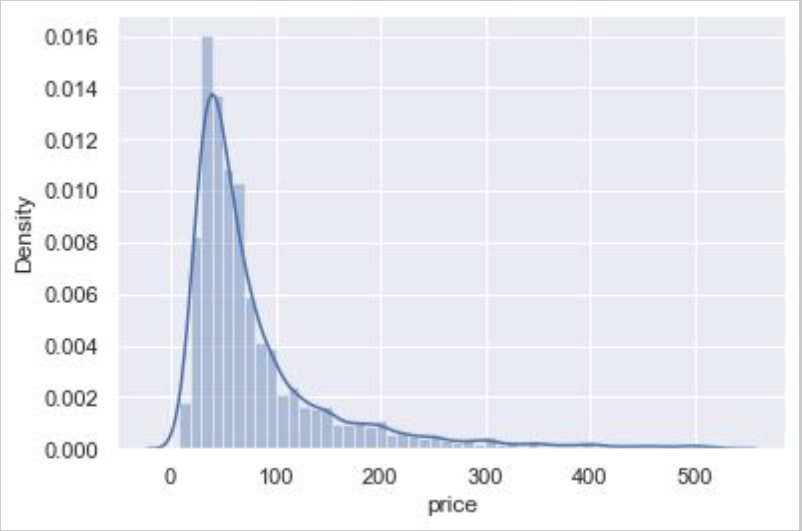
Τύπος καταλύματος	Διάμεση τιμή ενοικίασης (€)	Μέση τιμή ενοικίασης (€)
Ολόκληρο σπίτι/διαμέρισμα	59,0	114,0
Δωμάτιο ξενοδοχείου	72,0	405,7
Ιδιωτικό δωμάτιο	45,0	149,4
Κοινόχρηστο δωμάτιο	19,0	79,8

Η κατανομή της τιμής ενοικίασης στην Αθήνα παρουσιάζεται στο Γράφημα 1α . Υπάρχουν κάποιες ακραίες τιμές γι' αυτό και η κατανομή είναι θετικά ασύμμετρη. Κρατώντας το 99% των παρατηρήσεων (Γράφημα 1β), παρατηρούμε πως παραμένουν κάποιες ακραίες τιμές ωστόσο δεν παρατηρείται η ασυμμετρία στο βαθμό που παρατηρήθηκε έχοντας το σύνολο των παρατηρήσεων. Για να προσεγγίσει η κατανομή των τιμών την κανονική κατανομή ώστε να μπορέσουν να εφαρμοστούν μοντέλα παλινδρόμησης, θα χρησιμοποιηθούν οι λογαριθμισμένες τιμές ενοικίασης. Οι λογαριθμισμένες τιμές προσεγγίζουν αρκετά καλά την κανονική κατανομή (Γράφημα 1γ).

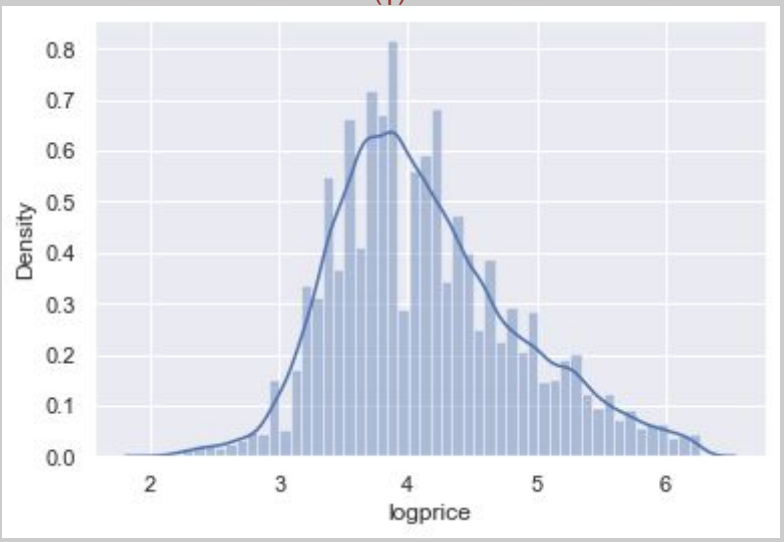
Γράφημα 1: Κατανομή των τιμών ενοικίασης στην Ελλάδα (Αθήνα, Θεσσαλονίκη, Κρήτη)



(β)


0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
Density
0
100
200
300
400
500
price

(γ)


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Density
2
3
4
5
6
logprice

Αθήνα:
Στην Αθήνα είναι καταγεγραμμένα 9455 καταλύματα. Η μέση τιμή ενοικίασης ενός καταλύματος ανά ημέρα είναι 70,14 €, ενώ η διάμεση τιμή ενοικίασης είναι τα 43 €. Το εύρος των τιμών κυμαίνεται από 8 έως 20.653€, ωστόσο το 75% των καταλυμάτων ενοικιάζονται με τιμή έως 65 ευρώ ανά ημέρα. Η τιμή διαφοροποιείται ανάλογα με το Δημοτικό Διαμέρισμα (ΔΔ) στο οποίο ανήκει το κατάλυμα, με τη μέση ημερήσια τιμή ενοικίασης να κυμαίνεται από 42 € (6ο ΔΔ) έως 85 € (3ο ΔΔ). Οι μέσες και διάμεσες ημερήσιες τιμές ενοικίασης ανά ΔΔ αναφέρονται στον Πίνακα 3.

Κατά μέσο όρο, η ελάχιστη απόσταση ενός καταλύματος από κάποιο μετρό είναι στα 0,44 χλμ και το 50% των καταλυμάτων απέχει από στάση μετρό απόσταση μικρότερη των 0,37 χλμ. Η ελάχιστη απόσταση ενός καταλύματος από στάση του μετρό είναι έως 2 χλμ.

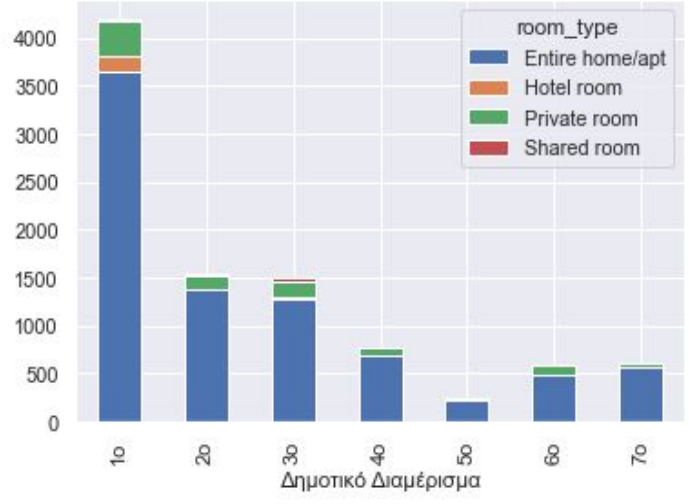
Όσον αφορά την ελάχιστη απόσταση των καταλυμάτων από κάποιο αξιοθέατο που ανήκει στα Top10 αξιοθέατα σύμφωνα με το Trip Advisor, η μέση απόσταση είναι 0,85 χλμ και κυμαίνεται από 18 μέτρα έως 5 χλμ.

Πίνακας 3: Μέσες και ενδιάμεσες τιμές ενοικίασης καταλυμάτων ανά Δημοτικό Διαμέρισμα στην Αθήνα

Δημοτικό Διαμέρισμα	Διάμεση τιμή ενοικίασης (€)	Μέση τιμή ενοικίασης (€)
1ο	50,0	75,7
2ο	40,0	74,2
3ο	41,0	85,0
4ο	33,0	51,8
5ο	31,5	43,4
6ο	32,0	42,3
7ο	39,0	46,8

Στο Γράφημα 2 βλέπουμε την κατανομή των καταλυμάτων στα ΔΔ ανά είδος. Τα περισσότερα καταλύματα βρίσκονται στο 1ο ΔΔ ενώ τα λιγότερα στο 5ο. Η πλειοψηφία των καταλυμάτων είναι ολόκληρα σπίτια.

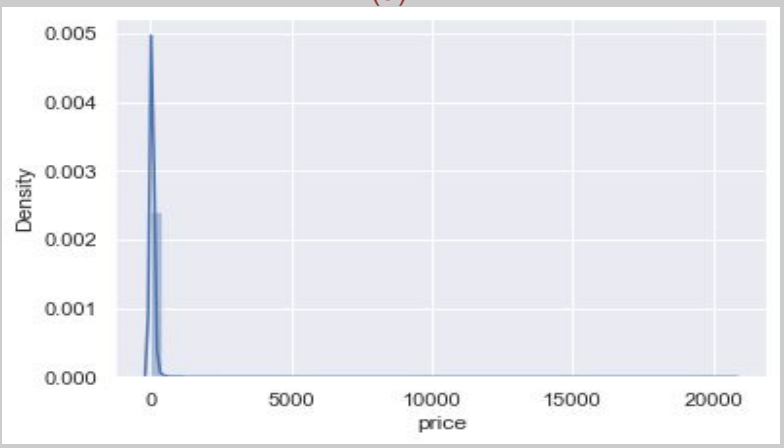
Γράφημα 2: Κατανομή καταλυμάτων ανά είδος και Δημοτικό διαμέρισμα στην Αθήνα



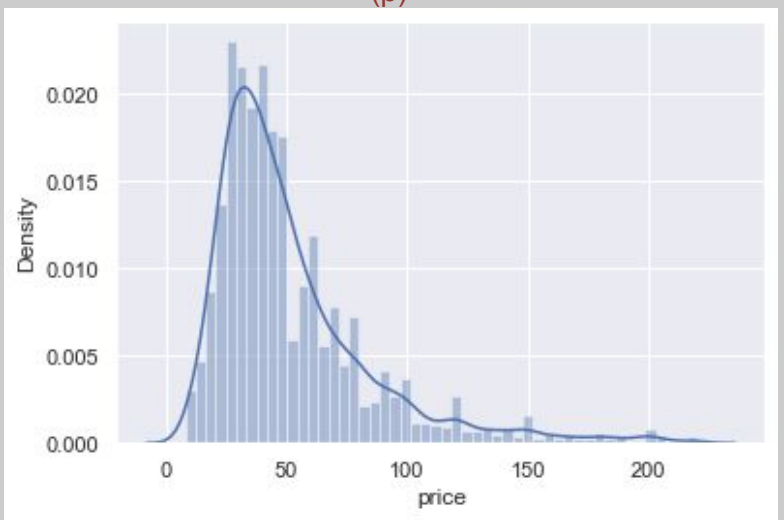
Η κατανομή της τιμής ενοικίασης στην Αθήνα παρουσιάζεται στο Γράφημα 3α . Υπάρχουν κάποιες ακραίες τιμές γι' αυτό και η κατανομή είναι θετικά ασύμμετρη. Κρατώντας το 99% των παρατηρήσεων (Γράφημα 3β), παρατηρούμε πως παραμένουν κάποιες ακραίες τιμές ωστόσο δεν παρατηρείται η ασυμμετρία στο βαθμό που παρατηρήθηκε έχοντας το σύνολο των παρατηρήσεων. Για να προσεγγίσει η κατανομή των τιμών την κανονική κατανομή ώστε να μπορέσουν να εφαρμοστούν μοντέλα παλινδρόμησης, θα χρησιμοποιηθούν οι λογαριθμισμένες τιμές ενοικίασης. Οι λογαριθμισμένες τιμές προσεγγίζουν αρκετά καλά την κανονική κατανομή (Γράφημα 3γ).

Γράφημα 3: Κατανομή των τιμών ενοικίασης στην Αθήνα

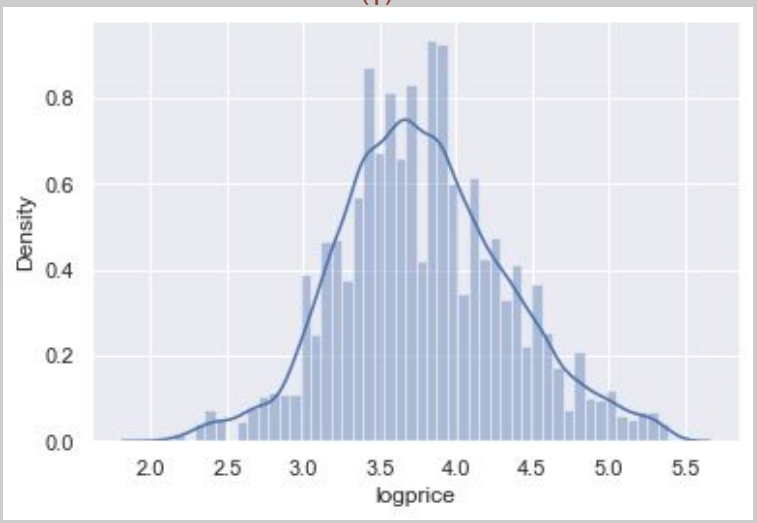
(α)



(β)



(γ)



Θεσσαλονίκη:
Στη Θεσσαλονίκη είναι καταγεγραμμένα 2441 καταλύματα. Η μέση τιμή ενοικίασης ενός καταλύματος ανά ημέρα είναι 58 €, ενώ η διάμεση τιμή ενοικίασης είναι τα 40 €.
Το εύρος των τιμών κυμαίνεται από 9 έως 8.000 €, ωστόσο το 75% των καταλυμάτων ενοικιάζονται με τιμή έως 55 ευρώ ανά ημέρα. Η τιμή διαφοροποιείται στις 7 περιοχές της πόλης, με τη μέση ημερήσια τιμή ενοικίασης να κυμαίνεται από 36 € (Αμπελόκηποι-Μενεμένη) έως 114 € (Πυλαία-Χορτιάτης). Οι μέσες και διάμεσες ημερήσιες τιμές ενοικίασης ανά περιοχή αναφέρονται στον Πίνακα 4.

Πίνακας 4: Μέσες και ενδιάμεσες τιμές ενοικίασης καταλυμάτων ανά περιοχή στη Θεσσαλονίκη

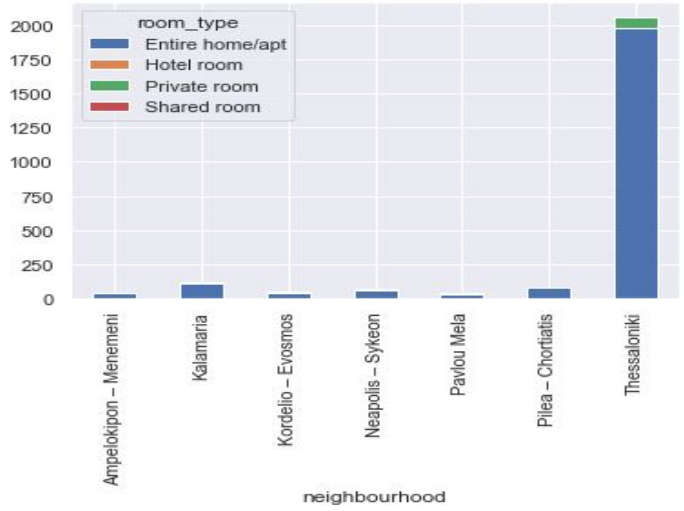
Περιοχή	Διάμεση τιμή ενοικίασης (€)	Μέση τιμή ενοικίασης (€)
Αμπελόκηποι- Μενεμένη	32,0	36,3
Καλαμαριά	47,5	55,3
Κορδελιό- Εύοσμος	36,5	54,0
Νεάπολη- Συκιές	32,0	39,0
Παύλου Μελά	37,5	46,4
Πυλαία- Χορτιάτης	50,0	114,2
Θεσσαλονίκη	40,0	57,4

Κατά μέσο όρο, η ελάχιστη απόσταση ενός καταλύματος από κάποιο μουσείο είναι στα 0,81 χλμ και το 50% των καταλυμάτων απέχει από μουσείο απόσταση μικρότερη των 0,36 χλμ. Η ελάχιστη απόσταση ενός καταλύματος από μουσείο φτάνει τα 11,7 χλμ.

Όσο αφορά την ελάχιστη απόσταση των καταλυμάτων από κάποιο αξιοθέατο που ανήκει στα Top10 αξιοθέατα σύμφωνα με το Trip Advisor, η μέση απόσταση είναι 1,1 χλμ και κυμαίνεται από 10 μέτρα έως 11 χλμ.

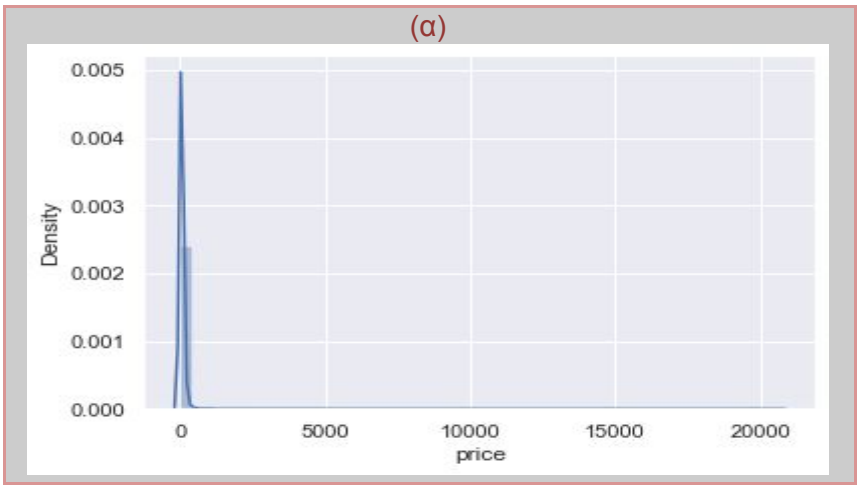
Στο Γράφημα 4 βλέπουμε την κατανομή των καταλυμάτων στις περιοχές ανά είδος. Τα περισσότερα καταλύματα βρίσκονται στο κέντρο της πόλης ενώ οι υπόλοιπες περιοχές έχουν λιγότερα από 200 καταλύματα. Η πλειοψηφία των καταλυμάτων είναι ολόκληρα σπίτια, όπως και στην Αθήνα, με ένα μικρό ποσοστό να είναι ιδιωτικά δωμάτια. Δεν υπάρχουν δωμάτια ξενοδοχείων που ενοικιάζονται σαν Airbnb ούτε και κοινόχρηστα δωμάτια.

Γράφημα 4: Κατανομή καταλυμάτων ανά είδος και περιοχή στη Θεσσαλονίκη

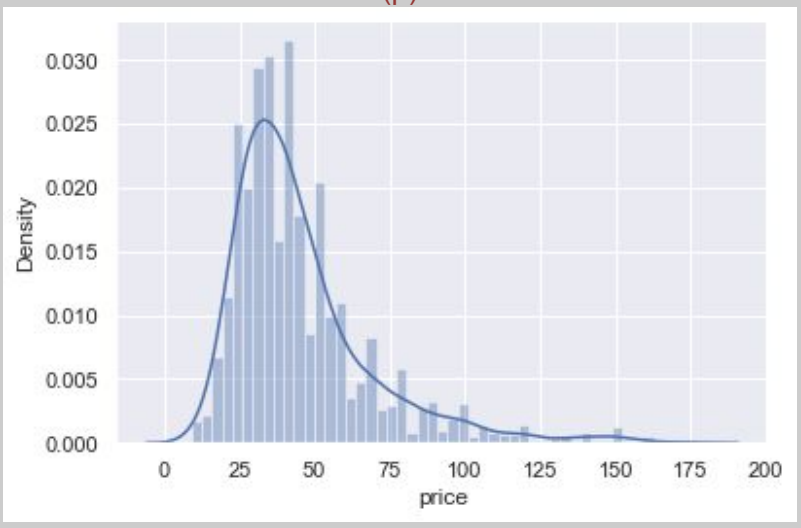


Η κατανομή της τιμής ενοικίασης στη Θεσσαλονίκη παρουσιάζεται στο Γράφημα 5α . Υπάρχουν κάποιες ακραίες τιμές γι' αυτό και η κατανομή είναι θετικά ασύμμετρη. Κρατώντας το 98% των παρατηρήσεων (Γράφημα 5β), παρατηρούμε πως παραμένουν κάποιες ακραίες τιμές ωστόσο δεν παρατηρείται η ασυμμετρία στο βαθμό που παρατηρήθηκε έχοντας το σύνολο των παρατηρήσεων. Για να προσεγγίσει η κατανομή των τιμών την κανονική κατανομή ώστε να μπορέσουν να εφαρμοστούν μοντέλα παλινδρόμησης, θα χρησιμοποιηθούν οι λογαριθμισμένες τιμές ενοικίασης. Οι λογαριθμισμένες τιμές προσεγγίζουν αρκετά καλά την κανονική κατανομή (Γράφημα 5γ).

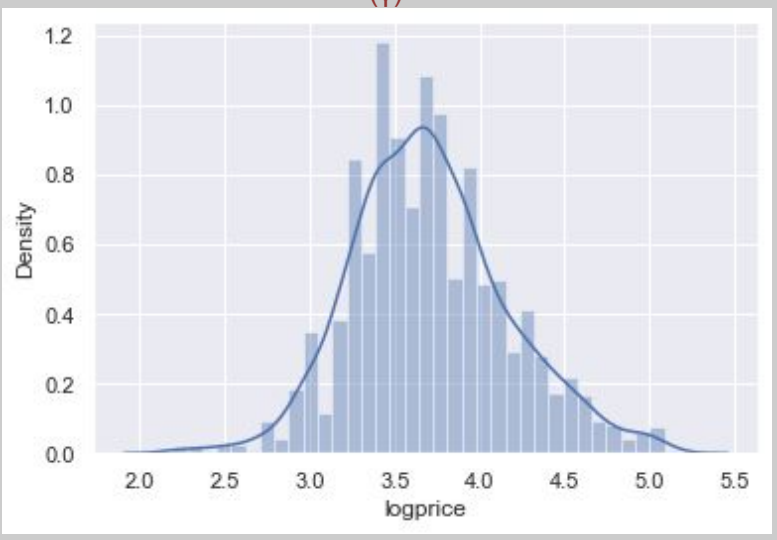
Γράφημα 5: Κατανομή τιμών ενοικίασης στη Θεσσαλονίκη



(β)


0.030
0.025
0.020
0.015
0.010
0.005
0.000
Density
0
25
50
75
100
125
150
175
200
price

(γ)


1.2
1.0
0.8
0.6
0.4
0.2
0.0
Density
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
logprice

Κρήτη:
Στην Κρήτη είναι καταγεγραμμένα 17705 καταλύματα. Η μέση τιμή ενοικίασης ενός καταλύματος ανά ημέρα είναι 161,5 €, ενώ η διάμεση τιμή ενοικίασης είναι τα 71 €.
Το εύρος των τιμών κυμαίνεται από 9 έως 8.000 €, ωστόσο το 75% των καταλυμάτων ενοικιάζονται με τιμή έως 55 ευρώ ανά ημέρα. Η τιμή διαφοροποιείται στους 4 νομούς του νησιού, με τη μέση ημερήσια τιμή ενοικίασης να κυμαίνεται από 143 € (Νομός Λασιθίου) έως 178 € (Νομός Ηρακλείου). Οι μέσες και διάμεσες ημερήσιες τιμές ενοικίασης ανά περιοχή αναφέρονται στον Πίνακα 5.

Πίνακας 5: Μέσες και ενδιάμεσες τιμές ενοικίασης καταλυμάτων ανά νομό στην Κρήτη

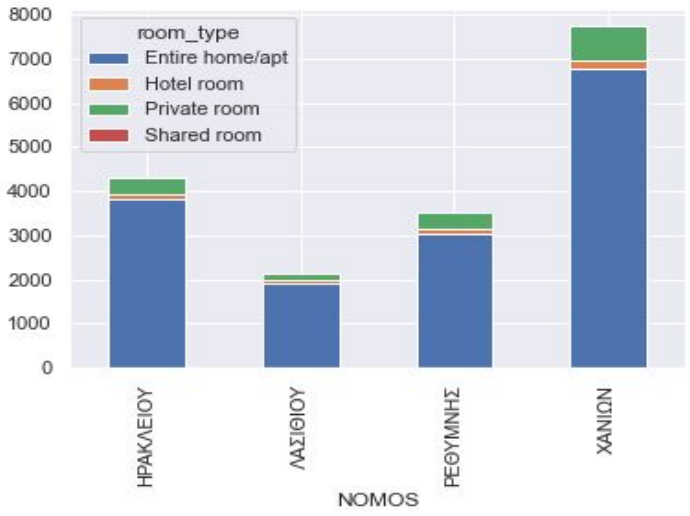
Νομός	Διάμεση τιμή ενοικίασης (€)	Μέση τιμή ενοικίασης (€)
Ηρακλείου	60,0	178,5
Λασιθίου	66,0	143,6
Ρεθύμνου	90,0	164,8
Χανίων	75,0	155,5

Κατά μέσο όρο, η ελάχιστη απόσταση ενός καταλύματος από κάποια παραλία με γαλάζια σημαία είναι στα 4,77 χλμ και το 50% των καταλυμάτων απέχει από παραλία απόσταση μικρότερη των 2,81 χλμ. Η ελάχιστη απόσταση ενός καταλύματος από παραλία είναι φτάνει έως και 58 χλμ.

Όσον αφορά την ελάχιστη απόσταση των καταλυμάτων από κάποιο αξιοθέατο που ανήκει στα Top10 αξιοθέατα σύμφωνα με το Trip Advisor, η μέση απόσταση είναι 12,05 χλμ και κυμαίνεται από 3 μέτρα έως 55 χλμ.

Στο Γράφημα 6 βλέπουμε την κατανομή των καταλυμάτων στους νομούς ανά είδος. Τα περισσότερα καταλύματα βρίσκονται στο νομό Χανίων ενώ τα λιγότερα στο νομό Λασιθίου. Η πλειοψηφία των καταλυμάτων είναι ολόκληρα σπίτια, όπως και στην Αθήνα και στη Θεσσαλονίκη, με ένα μικρό ποσοστό να είναι ιδιωτικά δωμάτια, δωμάτια ξενοδοχείου ή κοινόχρηστα δωμάτια.

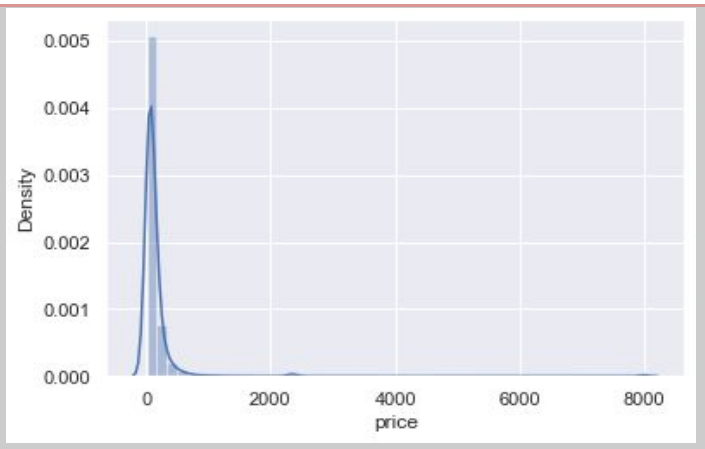
Γράφημα 6: Κατανομή καταλυμάτων ανά είδος και νομό στην Κρήτη



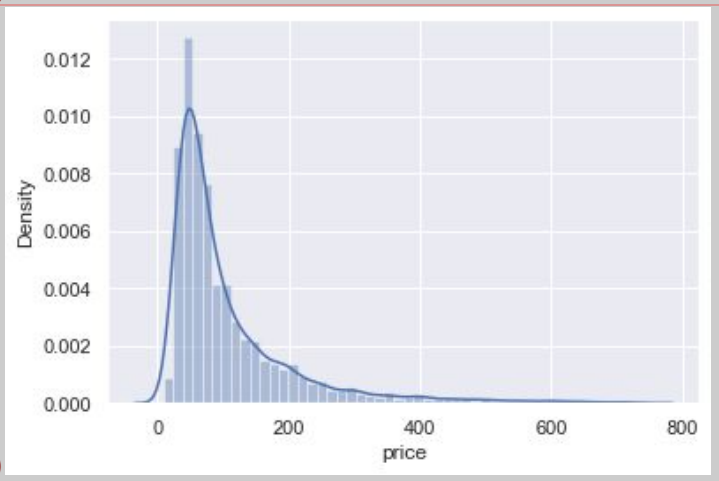
Η κατανομή της τιμής ενοικίασης στην Κρήτη παρουσιάζεται στο Γράφημα 7α. Υπάρχουν κάποιες ακραίες τιμές γι' αυτό και η κατανομή είναι θετικά ασύμμετρη. Κρατώντας το 98% των παρατηρήσεων (Γράφημα 7β), παρατηρούμε πως παραμένουν κάποιες ακραίες τιμές ωστόσο δεν παρατηρείται η ασυμμετρία στο βαθμό που παρατηρήθηκε έχοντας το σύνολο των παρατηρήσεων. Για να προσεγγίσει η κατανομή των τιμών την κανονική κατανομή ώστε να μπορέσουν να εφαρμοστούν μοντέλα παλινδρόμησης, θα χρησιμοποιηθούν οι λογαριθμισμένες τιμές ενοικίασης. Οι λογαριθμισμένες τιμές προσεγγίζουν αρκετά καλά την κανονική κατανομή (Γράφημα 7γ).

Γράφημα 7: Κατανομή τιμών ενοικίασης στην Κρήτη

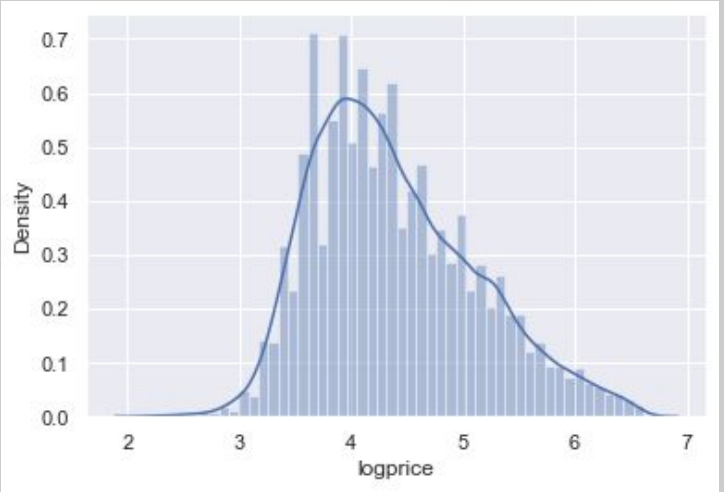
(α)



(β)



(γ)



Εξόρυξη Γνώσης

Α) ΠΑΛΙΝΔΡΟΜΗΣΗ

Για την πρόβλεψη της τιμής ενοικίασης ενός καταλύματος αρχικά χρησιμοποιήθηκαν μοντέλα παλινδρόμησης. Ως εξαρτημένη μεταβλητή επιλέχθηκε ο λογάριθμος της τιμής ενοικίασης του καταλύματος, για να ικανοποιούνται οι προϋποθέσεις του γραμμικού μοντέλου, δηλαδή η κανονικότητα των τιμών της εξαρτημένης μεταβλητής.

Ως ανεξάρτητες μεταβλητές στα μοντέλα με όλα τα καταλύματα επιλέχθηκαν οι:

- η ελάχιστη απόσταση από τα Top-10 αξιοθέατα
- ο αριθμός αξιολογήσεων ανά μήνα
- ο αριθμός των καταλυμάτων που έχει ένας ιδιοκτήτης στην κατοχή του
- οι διαθέσιμες ημέρες το χρόνο
- ο τύπος καταλύματος
- η περιοχή (διαχωρισμός με βάση το νομό: Νομός Αττικής, Νομός Θεσσαλονίκης, Νομός Ηρακλείου, Νομός Λασιθίου, Νομός Ρεθύμνου)

Ως ανεξάρτητες μεταβλητές στα μοντέλα με τα καταλύματα της Αθήνας επιλέχθηκαν οι:

- η ελάχιστη απόσταση από τα Top-10 αξιοθέατα
- ο αριθμός αξιολογήσεων ανά μήνα
- ο αριθμός των καταλυμάτων που έχει ένας ιδιοκτήτης στην κατοχή του
- οι διαθέσιμες ημέρες το χρόνο
- ο τύπος καταλύματος
- το δημοτικό διαμέρισμα
- η ελάχιστη απόσταση από στάση του μετρό

Ως ανεξάρτητες μεταβλητές στα μοντέλα με τα καταλύματα της Θεσσαλονίκης επιλέχθηκαν οι:

- η ελάχιστη απόσταση από τα Top-10 αξιοθέατα
- ο αριθμός αξιολογήσεων ανά μήνα
- ο αριθμός των καταλυμάτων που έχει ένας ιδιοκτήτης στην κατοχή του
- οι διαθέσιμες ημέρες το χρόνο
- τύπος καταλύματος
- γειτονία
- η ελάχιστη απόσταση από κάποιο μουσείο

Ως ανεξάρτητες μεταβλητές στα μοντέλα με τα καταλύματα της Κρήτης επιλέχθηκαν οι:

- η ελάχιστη απόσταση από τα Top-10 αξιοθέατα
- ο αριθμός αξιολογήσεων ανά μήνα
- ο αριθμός των καταλυμάτων που έχει ένας ιδιοκτήτης στην κατοχή του
- οι διαθέσιμες ημέρες το χρόνο
- ο τύπος καταλύματος
- ο νομός

Τα μοντέλα που εκπαιδεύτηκαν είναι:

1) **Multiple linear regression (MLG)**
   Η διαδικασία προσδιορισμού της σχέσης μιας μεταβλητής y (εξαρτημένη μεταβλητή που στη συγκεκριμένη περίπτωση είναι ο λογάριθμος της τιμής ενοικίασης) με κάποιες άλλες μεταβλητές x1, x2, …, xn (ανεξάρτητες). Ως ανεξάρτητες μεταβλητές χρησιμοποιήθηκαν οι μεταβλητές που προαναφέρθηκαν.
   Η αναμενόμενη τιμή της εξαρτημένης μεταβλητής μοντελοποιείται υποθέτοντας ότι υπάρχει γραμμική συνάρτηση με τις ανεξάρτητες μεταβλητές.
   Η μέθοδος εκτίμησης των συντελεστών παλινδρόμησης που χρησιμοποιήθηκε ήταν η μέθοδος των ελαχίστων τετραγώνων, που ελαχιστοποιεί το σφάλμα μεταξύ της εκτιμώμενης συνάρτησης και των πραγματικών δεδομένων.

2) **Random Forest regression (RFR)**
   Η περίπτωση αυτή είναι αποτέλεσμα bagging δέντρων αποφάσεων, για κάθε νέο σημείο που θέλουμε να προβλέψουμε επιλέγουμε το μέσο όρο των τιμών που δίνουν τα δέντρα που συνιστούν το δάσος. Είναι αντίστοιχη διαδικασία με αυτή της ταξινόμησης όμως κάθε περιοχή που στην ταξινόμηση αντιπροσωπεύει μια κλάση στην περίπτωση της παλινδρόμησης αντιπροσωπεύει μια τιμή και η τιμή αυτή είναι ο μέσος όρος των στοιχείων εκπαίδευσης τα οποία ανήκουν στην περιοχή αυτή.

3) **Bayes Regression (BR)**
   Στη συγκεκριμένη περίπτωση, υπάρχει η υπόθεση ότι οι συντελεστές παλινδρόμησης ακολουθούν κατανομή κανονική με μέση τιμή 0 και διασπορά λ. Οι παράμετροι λ και α (που χρησιμοποιείται για regularization ακολουθούν κατανομή γάμμα. Οι υπερπαράμετροι (παράμετροι που καθορίζουν τις κατανομές των α και λ) συνήθως επιλέγονται να είναι η πληροφοριακές και πήραν τις default τιμές (10^-6) (πολύ μικρές τιμές). Οι παράμετροι των κατανομών εκτιμώνται από τα δεδομένα και μεγιστοποιώντας την marginal log likelihood. Δίνει παρόμοια αποτελέσματα με την πολλαπλή γραμμική παλινδρόμηση καθώς οι συντελεστές που εκτιμώνται είναι πιο κοντά στο 0.

Τα μοντέλα εκπαιδεύτηκαν/επαληθεύτηκαν σε αναλογία 90/10. Για την αξιολόγηση των μοντέλων υπολογίστηκαν:

- Το μέσο απόλυτο σφάλμα (MAE):

Το MAE βρίσκει τις απόλυτες διαφορές της πραγματικής τιμής y και της προβλεπόμενης από το μοντέλο και παίρνει το μέσο όρο τους. Δεν λαμβάνει υπόψη αν η πρόβλεψη από το

μοντέλο είναι υψηλότερη ή χαμηλότερη από την πραγματική τιμή αλλά λαμβάνει υπόψη την απόλυτη διαφορά. Όσο χαμηλότερο το MAE τόσο πιο ακριβείς είναι οι προβλέψεις του μοντέλου. Μας λέει πόσο κατά μέσο όρο αποκλίνουν οι προβλέψεις από τις πραγματικές τιμές και έχει μονάδες μέτρησης τις μονάδες της Y (εδώ ευρώ). Για να αποφανθούμε για το μέγεθος του σφάλματος συγκρίνουμε το σφάλμα με τη μέση τιμή της y (εδώ τιμή ενοικίασης). Χρησιμοποιείται συνήθως όταν δεν υπάρχουν ακραίες τιμές στα δεδομένα.

- Το μέσο τετραγωνικό σφάλμα (MSE):

Το MSE βρίσκει τις τετραγωνικές διαφορές της πραγματικής τιμής y και της προβλεπόμενης από το μοντέλο και παίρνει το μέσο όρο τους. Αποτέλεσμα της τετραγώνισης είναι το ότι δίνει μεγαλύτερο βάρος στα μεγαλύτερα σφάλματα. Συνήθως, τα outliers δίνουν και μεγαλύτερα σφάλματα γιατί εκεί δεν κάνει καλό fit το μοντέλο. Αν υπάρχουν outliers πρέπει να εξετάζουμε το MSE. Αντίθετα με το MAE, δεν έχει τις μονάδες μέτρησης της y.

- Η ρίζα του μέσου τετραγωνικού σφάλματος (RMSE):

Το RMSE είναι η ρίζα του MSE. Αν μας ανησυχούν τα outliers και αν αυτά επηρεάζουν τις προβλέψεις , πρέπει να το κοιτάμε. Έχει τις ίδιες μονάδες μέτρησης της y. Όσο μικρότερο τόσο το καλύτερο.

Ο συντελεστής προσδιορισμού (R^2):
Ο συντελεστής αυτός μετράει πόση διακύμανση της εξαρτημένης μεταβλητής κατάφεραν να ερμηνεύσουν οι ανεξάρτητες μεταβλητές. Οι τιμές του συντελεστή προσδιορισμού R2 κυμαίνονται από το 0 ως το 1 και προφανώς όσο η τιμή πλησιάζει προς το 1 τόσο καλύτερη προσαρμογή έχει το μοντέλο.

Τα μέτρα αυτά υπολογίστηκαν και στο training dataset και στο testing dataset. Τα αποτελέσματα παρουσιάζονται στους παρακάτω πίνακες (Πίνακες 4-5).

Έγινε κανονικοποίηση των εξαρτημένων μεταβλητών πριν χρησιμοποιηθούν στα μοντέλα μας.

Παρατηρούμε πως ο αλγόριθμος Random Forest δίνει τα μικρότερα σφάλματα συγκριτικά με τους άλλους αλγορίθμους παλινδρόμησης που χρησιμοποιήθηκαν. Όσον αφορά τα απόλυτα σφάλματα, όλοι οι αλγόριθμοι δίνουν μικρά σφάλματα λαμβάνοντας υπόψη τις μέσες τιμές ενοικίασης κάθε πόλης και συνολικά. Οι πολλαπλή γραμμική παλινδρόμηση και η παλινδρόμηση κατά Bayes δίνουν παρόμοια αποτελέσματα, γεγονός που ήταν αναμενόμενο καθώς δεν είχαμε κάποια εκ των προτέρων γνώση για τις κατανομές των συντελεστών παλινδρόμησης που θα μπορούσαμε να αξιοποιήσουμε στην κατά Bayes παλινδρόμηση.

Σχετικά με τις μεταβλητές που εισήχθησαν στα μοντέλα, εξηγούν διαφορετικά ποσοστά της μεταβλητότητας των τιμών ενοικίασης ανάλογα την πόλη με βάση τα training sets. Οι συντελεστές R^2 που προέκυψαν από τον Random Forest είναι >85% ενώ οι υπόλοιποι συντελεστές είναι της τάξης του 20-25%. Στη Θεσσαλονίκη εκτιμήθηκε ο χαμηλότερος συντελεστής R^2, περίπου 10%. Αν λοιπόν θα έπρεπε να επιλέξουμε μεταξύ των αλγορίθμων που δοκιμάσαμε, ο αλγόριθμος Random Forest υπερτερεί, εκτιμώντας υψηλότερο R^2 και μικρότερα απόλυτα και σχετικά σφάλματα.

Πίνακας 4: Μέτρα αξιολόγησης αλγορίθμων στο training dataset

	MAE	MSE	RMSE	R^2
Συνολικά (Αθήνα, Θεσσαλονίκη και Κρήτη)				
MLG	46,20	8537,85	92,40	0,26
RFR	17,99	1977,70	44,47	0,89
BR	46,20	8537,85	92,40	0,26
Αθήνα				
MLG	20,26	1002,55	31,66	0,21
RFR	8,14	217,50	14,74	0,87
BR	20,27	1002,98	31,67	0,21
Θεσσαλονίκη				
MLG	14,21	420,71	20,51	0,10
RFR	5,70	88,35	9,40	0,85
BR	14,21	421,28	20,52	0,10
Κρήτη				
MLG	56,12	9668,86	98,33	0,19
RFR	20,76	1993,63	44,65	0,89
BR	56,12	9670,54	98,33	0,19

Πίνακας 5: Μέτρα αξιολόγησης αλγορίθμων στο testing dataset

	MAE	MSE	RMSE	R^2
Συνολικά (Αθήνα, Θεσσαλονίκη και Κρήτη)				
MLG	45,38	8475,90	92,06	0,27
RFR	41,42	6937,27	83,29	0,39
BR	45,38	8475,90	92,06	0,27
Αθήνα				
MLG	19,23	863,17	29,38	0,24
RFR	18,44	783,74	28,00	0,31
BR	19,23	863,58	29,39	0,24
Θεσσαλονίκη				
MLG	13,67	395,63	19,89	0,15
RFR	13,00	369,43	19,22	0,21
BR	13,67	396,88	19,92	0,14
Κρήτη				
MLG	57,53	9601,22	97,98	0,18
RFR	47,79	7108,16	84,30	0,37
BR	57,52	9601,88	97,98	0,18

Β) ΤΑΞΙΝΟΜΗΣΗ

Προκειμένου να εκπαιδευτεί έναν κατηγοριοποιητή που θα προβλέπει την κατηγορία τιμής ενός καταλύματος, οι τιμές ενοικίασης διαχωρίστηκαν σε 5 μη επικαλυπτόμενες κατηγορίες ως εξής:

Κατηγορία 1: τιμή ενοικίασης<20^{ο} εκατοστημόριο της κατανομής
Κατηγορία 2: τιμή ενοικίασης≥20^{ο} και <40^{ο} εκατοστημόριο της κατανομής
Κατηγορία 3: τιμή ενοικίασης≥40^{ο} και <60^{ο} εκατοστημόριο της κατανομής
Κατηγορία 4: τιμή ενοικίασης≥60^{ο} και <80^{ο} εκατοστημόριο της κατανομής
Κατηγορία 5: τιμή ενοικίασης ≥80^{ο} εκατοστημόριο της κατανομής

Χρησιμοποιήθηκαν οι ίδιες μεταβλητές που χρησιμοποιήθηκαν και στην παλινδρόμηση και οι εξής τεχνικές:

1) **Δέντρα απόφασης-Decision trees (DT)**
   Το μοντέλο που εκπαιδεύεται έχει τη μορφή δέντρου. Οι εσωτερικοί κόμβοι αντιστοιχούν σε κάποιο γνώρισμα και τα φύλλα αντιστοιχούν σε κλάσεις. Ένας κόμβος διαχωρίζεται (split) σε δύο ή περισσότερους με βάση μια συνθήκη ελέγχου. Πλεονέκτημα αποτελεί το γεγονός ότι είναι μια μη παραμετρική προσέγγιση, δηλαδή δε στηρίζεται σε υπόθεση εκ των προτέρων γνώσης σχετικά με τον τύπο της κατανομής πιθανότητας που ικανοποιεί η κλάση ή τα άλλα γνωρίσματα. Κάθε πλειάδα προς κατηγοριοποίηση πρέπει να περάσει από το δένδρο. Η διαδικασία αυτή παίρνει χρόνο ανάλογο με το ύψος του δένδρου. Ωστόσο, ένα από τα βασικά μειονεκτήματα τους είναι ότι δεν μπορούν να χειριστούν συνεχή δεδομένα. Επίσης, τα δένδρα απόφασης προϋποθέτουν ότι ο χώρος του πεδίου διαιρείται σε ορθογώνιες περιοχές. Επιπρόσθετα, το φαινόμενο της υπερπροσαρμογής είναι πιθανό να εμφανιστεί στα δένδρα απόφασης αφού αυτό δημιουργείται βάσει των δεδομένων εκπαίδευσης. Τέλος, τα δένδρα απόφασης δε λαμβάνουν υπόψη τις πιθανές συσχετίσεις που υπάρχουν μεταξύ των χαρακτηριστικών.

2) **Support Vector Machines (SVM)**
   Μια νέα μέθοδος κατηγοριοποίησης για γραμμικά και μη γραμμικά δεδομένα. Χρησιμοποιεί μη γραμμική απεικόνιση για να μετασχηματίσει τα αρχικά δεδομένα εκπαίδευσης σε υψηλότερες διαστάσεις. Με τη νέα διάσταση, αναζητεί για το βέλτιστο γραμμικό υπερεπίπεδο που διαχωρίζει τα δεδομένα. Με την κατάλληλη μη γραμμική απεικόνιση σε μια ικανοποιητική υψηλότερη διάσταση, δεδομένα δύο κατηγοριών μπορούν πάντοτε να διαχωρίζονται από ένα υπερεπίπεδο. Η μέθοδος βρίσκει το υπερεπίπεδο χρησιμοποιώντας support vectors και όρια που ορίζονται από αυτούς.

3) **k-πιο κοντινοί γείτονες- k- Nearest Neighbors (KNN)**
   Ευρεία χρησιμοποιούμενη τεχνική κατηγοριοποίησης που βασίζεται στη χρήση μέτρων βασισμένων στην απόσταση είναι αυτή των K - πλησιέστερων γειτόνων (K-nearest neighbors KNN). Η τεχνική KNN προϋποθέτει ότι το σύνολο εκπαίδευσης δεν περιλαμβάνει μόνο τα δεδομένα αλλά επίσης και την επιθυμητή κατηγοριοποίηση για κάθε στοιχείο. Αυτό έχει σαν αποτέλεσμα τα δεδομένα εκπαίδευσης να αποτελούν το μοντέλο κατηγοριοποίησης. Όταν πρόκειται να γίνει μια κατηγοριοποίηση για ένα νέο στοιχείο, πρέπει να καθοριστεί η απόσταση του από κάθε στοιχείο του συνόλου εκπαίδευσης. Μόνο οι K κοντινότερες εκχωρήσεις στο σύνολο εκπαίδευσης λαμβάνονται υπόψη στη συνέχεια. Το νέο στοιχείο τοποθετείται στην κατηγορία που περιέχει τα

περισσότερα στοιχεία από το σύνολο των Ê κοντινότερων στοιχείων. Είναι απλή στη χρήση και στην υλοποίηση. Είναι ανθεκτική στα θορυβώδη δεδομένα εκπαίδευσης, ειδικά αν το αντίστροφο τετράγωνο της σταθμισμένης απόστασης χρησιμοποιείται ως μέτρο απόστασης. Επίσης, είναι πιο αποτελεσματική εάν ο αριθμός των δεδομένων εκπαίδευσης (δειγμάτων) είναι μεγάλος.

4) **Naïve Bayes Classifier (NB)**
Ένας κατηγοριοποιητής που βασίζεται στην στατιστική: κάνει πρόβλεψη, π.χ., προβλέπει την πιθανότητα ένα δείγμα να ανήκει σε κάποια κλάση. Στηρίζεται στο θεώρημα του Bayes. Ο naïve Bayesian classifier έχει απόδοση συγκρίσιμη με τα δέντρα απόφασης. Η απόδοση αυτού του είδους κατηγοριοποίησης είναι αρκετά υψηλή και χαρακτηρίζεται από την μεγάλη ταχύτητα της διαδικασίας κατηγοριοποίησης σε μεγάλες Βάσεις Δεδομένων.

Η πιστότητα (accuracy) των παραπάνω αλγορίθμων παρουσιάζεται συνολικά και ανά περιοχή στον Πίνακα 6.

Πίνακας 6: Πιστότητας (accuracy) των τεχνικών ταξινόμησης

	Πιστότητα			
	Συνολικά	**Αθήνα**	**Θεσσαλονίκη**	**Κρήτη**
DT	**0,39**	0,31	0,31	**0,37**
SVM	0,35	**0,34**	0,28	0,30
KNN	**0,39**	0,31	**0,34**	0,36
NB	0,32	0,20	0,27	0,24

Ανάλογα με την ανάλυση που πραγματοποιήθηκε σε κάθε πόλη, βλέπουμε πως διαφορετικός αλγόριθμος έχει το καλύτερο ποσοστό πιστότητα. Στην Αθήνα καλύτερη απόδοση έχει ο αλγόριθμος SVM, στη Θεσσαλονίκη ο KNN ενώ στην Κρήτη ο DT. Στην ανάλυση με όλες τις πόλεις, καλύτερη απόδοση έχουν οι αλγόριθμοι DT και KNN. Σε όλες τις περιπτώσεις, η πιστότητα του μοντέλου είναι περίπου 35%.

Γ) ΑΛΛΕΣ ΤΕΧΝΙΚΕΣ ΑΝΑΛΥΣΗΣ

1. ΣΥΣΤΑΔΟΠΟΙΗΣΗ

Για την εύρεση συστάδων με κοινά γνωρίσματα αντλούμενα από το υπάρχον dataset θα μπορούσαν να αξιοποιηθούν τεχνικές συσταδοποίησης. Για παράδειγμα θα μπορούσαμε να δημιουργήσουμε συστάδες κλάσεων τιμών ενοικίασης με βάση τις predicted τιμές ενοικίασης ή την πραγματική βαθμολογία (σε αστέρια) του κάθε καταλύματος.

Ο αλγόριθμος που θα χρησιμοποιούσαμε θα ήταν ο k-means. Ως αντιπροσωπευτικό σημείο κάθε συστάδας θα θεωρούσαμε τον trimmed μέσο όρο τιμών ώστε να αφαιρέσουμε τις ακραίες τιμές. Πριν την είσοδο θα κανονικοποιούσαμε τις τιμές των γνωρισμάτων ώστε να είναι στην ίδια κλίμακα. Έτσι τα γνωρίσματα θα είχαν μέση τιμή 0 και τυπική απόκλιση 1. Τα παραγόμενα αποτελέσματα θα ήταν οι συστάδες τιμών με βάση την βαθμολογία ή τις predicted τιμές ενοικίασης. Από την απεικόνιση των ομάδων σε ένα γράφημα θα μπορούσαμε να δούμε ομάδες με χαμηλή κλάση τιμής και υψηλή βαθμολογία ή υψηλή predicted τιμή ενοικίασης. Τέτοιες ομάδες θα χαρακτηρίζονταν value for money καθώς θα συνδύαζαν τη χαμηλή τιμή ενοικίασης με την καλή βαθμολογία ή την υψηλή τιμή ενοικίασης που έχει προβλεφθεί από τα μοντέλα που έχουμε εφαρμόσει (το οποίο θα σήμαινε ότι ενοικιάζεται σε χαμηλότερη τιμή απ' ότι θα μπορούσε με βάση τα γνωρίσματα του). Αυτό θα μπορούσε να βοηθήσει τον πελάτη ώστε να επιλέξει καταλύματα που προσφέρουν χαμηλή τιμή και παροχή υπηρεσιών υψηλής ποιότητας.

2. ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΗΣ

Για την Εύρεση συσχετίσεων της κλάσης τιμής με ένα ή περισσότερα γνωρίσματα (πχ την περιοχή (νομό), το είδος του διαμερίσματος, την κατηγορία βαθμολογίας (π.χ 5 κατηγορίες βαθμολογίας (βαθμολόγηση με αστέρια)) και το μέγεθος του καταλύματος (ταξινόμηση σε k κατηγορίες ανάλογα με τα τετραγωνικά)) θα μπορούσαν να αξιοποιηθούν τεχνικές κανόνων συσχέτισης, όπως ο αλγόριθμος apriori.

Πριν την είσοδο, θα έπρεπε να μετασχηματιστούν οι συμβολοσειρές σε σύνολα σημείων. Ως παραγόμενο αποτέλεσμα θα είχαμε τα πλέον συχνά σύνολα που θα περιείχαν την κλάση τιμής και ένα ή περισσότερα γνωρίσματα. Με τον τρόπο αυτό θα μπορούσε να εξαχθεί πληροφορία σχετικά με τα γνωρίσματα που συναντώνται σε καταλύματα με συγκεκριμένη κλάση τιμής και έτσι νέοι ιδιοκτήτες καταλυμάτων θα μπορούσαν να ορίσουν την τιμή ενοικίασης ενός καταλύματός τους ή παλαιοί ιδιοκτήτες να αναδιαμορφώσουν την τιμή που έχουν ήδη ορίσει, ώστε να είναι ανταγωνιστικές με βάση τις παροχές τους.

Υλοποίηση σε Python.

https://github.com/tziojo/airbnb_