ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΠΡΟΓΡΑΜΜΑ ΠΡΟΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΜΑΙΟΣ 2024

‘ΚΡΑΤΗΣΕΙΣ ΞΕΝΟΔΟΧΕΙΟΥ’

Γουέρα Ευαγγελία-Ζωή

Εισαγωγή στις Πιθανότητες και Στατιστική με R

Διδάσκοντες: Δ.Καρλής, Β.Χασίωτης

Περιεχόμενα

1. ΕΙΣΑΓΩΓΗ.....3
2. ΑΝΑΛΥΣΗ 7

1. ΕΙΣΑΓΩΓΗ

Στο πλαίσιο αυτής της εργασίας εξετάζουμε ένα τυχαίο δείγμα κρατήσεων δωματίων σε ένα ξενοδοχείο.

Το πρώτο βήμα είναι η τροποποίηση των δεδομένων. Θεωρήσαμε σωστό τις μεταβλητές number of children, number of weekend.nights, number of week.nights, car parking space και repeated να τις μετατρέψουμε σε κατηγορικές και την μεταβλητή lead.time σε ποσοτική. Επίσης, παρατηρούμε την ύπαρξη μη διαθέσιμων τιμών (NA), οι οποίες θα χρειαστεί να αφαιρεθούν. Το επόμενο βήμα πριν ολοκληρώσουμε την τροποποίηση των δεδομένων είναι να εξετάσουμε κάθε μεταβλητή ξεχωριστά και να δούμε τις τιμές που περιέχει.

Για τις κατηγορικές μεταβλητές μπορούμε να δούμε τις συχνότητες με τις οποίες έχει εμφανιστεί η κατηγορία της κάθε μεταβλητής και τις σχετικές συχνότητες. Ενώ για τις μεταβλητές που είναι ποσοτικές μπορούμε να δούμε την μέση τιμή, τη διάμεσο , το πρώτο και τρίτο τεταρτημόριο, την μέγιστη τιμή και την ελάχιστη τιμή.

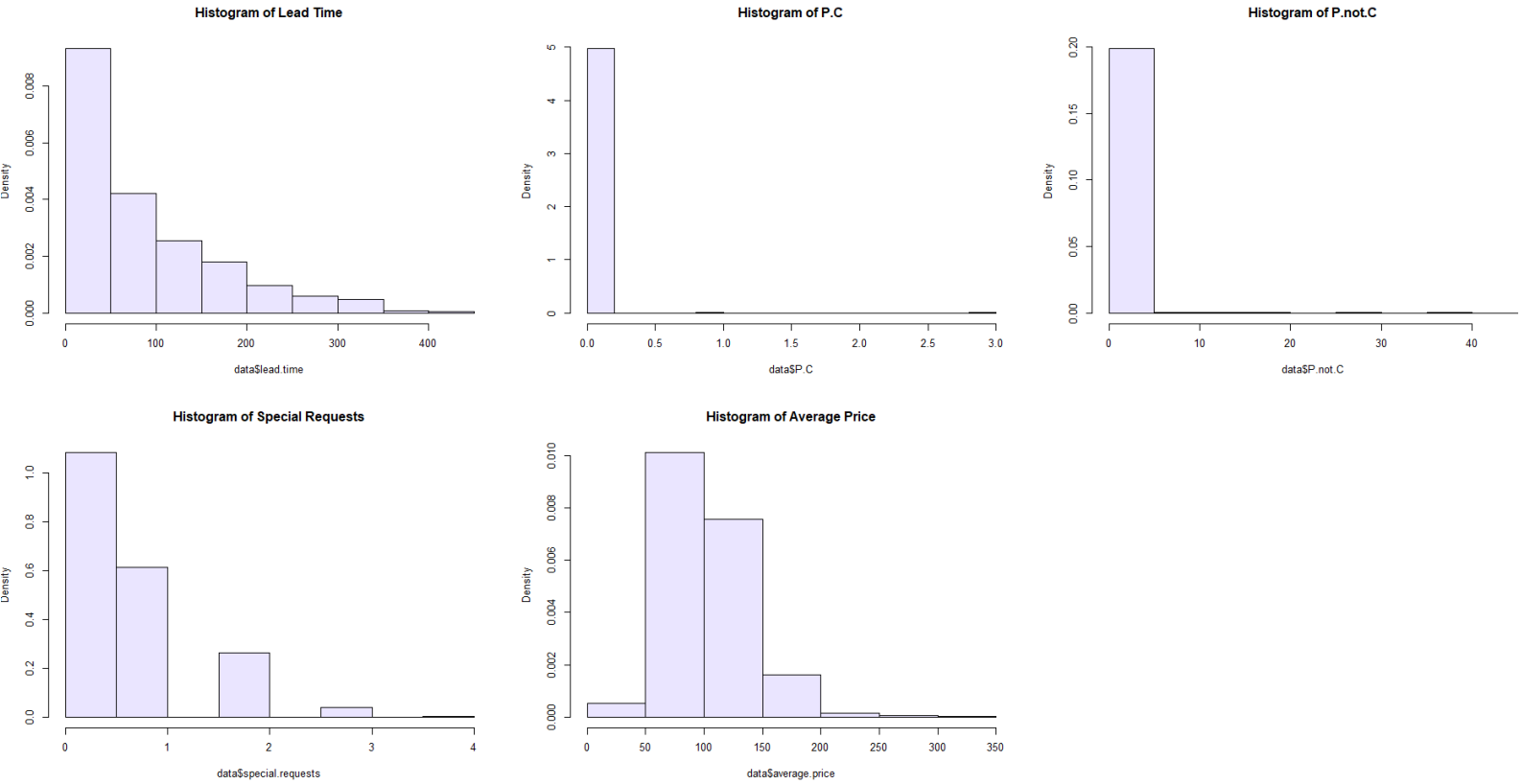
Ξεκινώντας με τις ποσοτικές παρατηρούμε ότι:

- Η μεταβλητή lead time έχει ελάχιστη τιμή το -99 , μια τιμή που είναι μη αναμενομένη αφού η μεταβλητή δηλώνει τον αριθμό ημερών μεταξύ της ημερομηνίας κράτησης και της ημερομηνίας άφιξης. Θεωρώντας την συγκεκριμένη τιμή ως μη λογική και ως τυπικό λάθος στην καταγραφή δεδομένων θα χρειαστεί να αφαιρέσουμε από όλο το σετ δεδομένων την παρατήρηση που αντιστοιχεί σε αυτήν την τιμή. Επίσης παρατηρούμε ότι το 25% των παρατηρήσεων παίρνουν τιμές μέχρι το 17 , το 50% μέχρι το 56 και το 75% μέχρι το 121 με μέγιστη τιμή το 418.
- Για τις μεταβλητές P.C και P.not.C παρατηρούμε ότι το 75% των παρατηρήσεων παίρνουν μηδενικές τιμές και η μεγαλύτερη τιμή τους είναι το 3 και το 41 αντίστοιχα.
- Η μεταβλητή average price παρατηρούμε ότι έχει μηδενική ελάχιστη τιμή. Συγκεκριμένα, παρατηρούμε ένα πλήθος τιμών, μέσα σε αυτό και οι μηδενικές τιμές, που δεν γνωρίζουμε κατά πόσο αντιστοιχεί στην μέση τιμή κρατήσεων, δηλαδή κατά πόσο έχει λογική ερμηνεία. Επειδή όμως δεν γνωρίζουμε αν μπορεί να θεωρηθεί ως σφάλμα καταχώρησης των δεδομένων, θα το κρατήσουμε προσπαθώντας να κατανοήσουμε την κατανομή της μεταβλητής. Επίσης παρατηρούμε ότι το 25% των παρατηρήσεων παίρνουν τιμές μέχρι το 80 , το 50% μέχρι το 97,75 και το 75% μέχρι το 121 με μέγιστη τιμή το 349,6.
- Για την μεταβλητή special requests παρατηρούμε ότι το 50% των παρατηρήσεων παίρνουν μηδενικές τιμές και το 75% τιμές μέχρι το 1 ενώ η μέγιστη τιμή του είναι το 4.

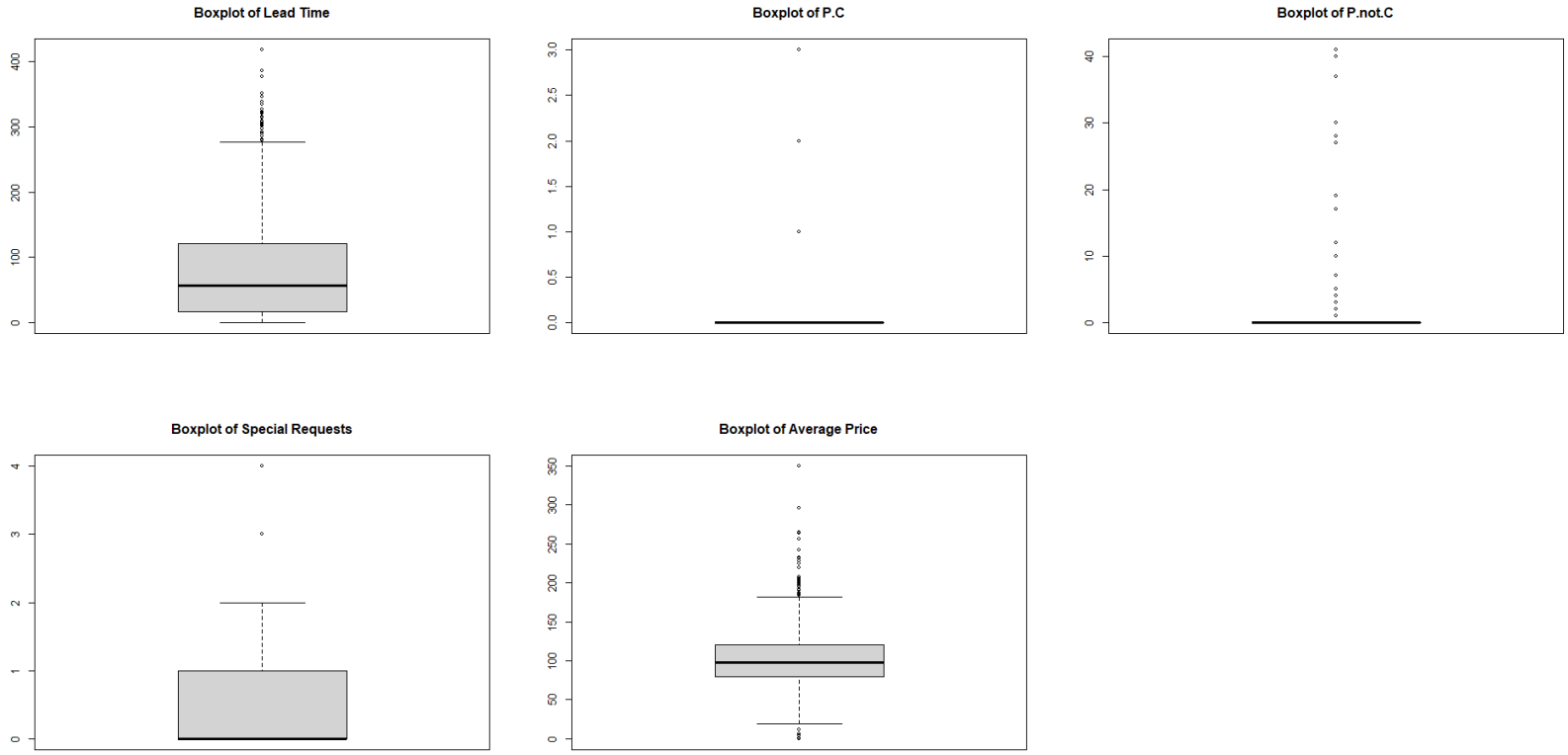
Από τα παραπάνω περιγραφικά μέτρα αλλά και από την χρήση διαγραμμάτων (Σχήμα 1, Σχήμα 2) παρατηρούμε την ύπαρξη ακραίων τιμών για όλες τις μεταβλητές όπως και την ασσυμετρία συμπεραίνοντας ότι καμία δεν προσεγγίζει την κανονική κατανομή.

Έπειτα, για τις κατηγορικές μεταβλητές παρατηρούμε ότι:

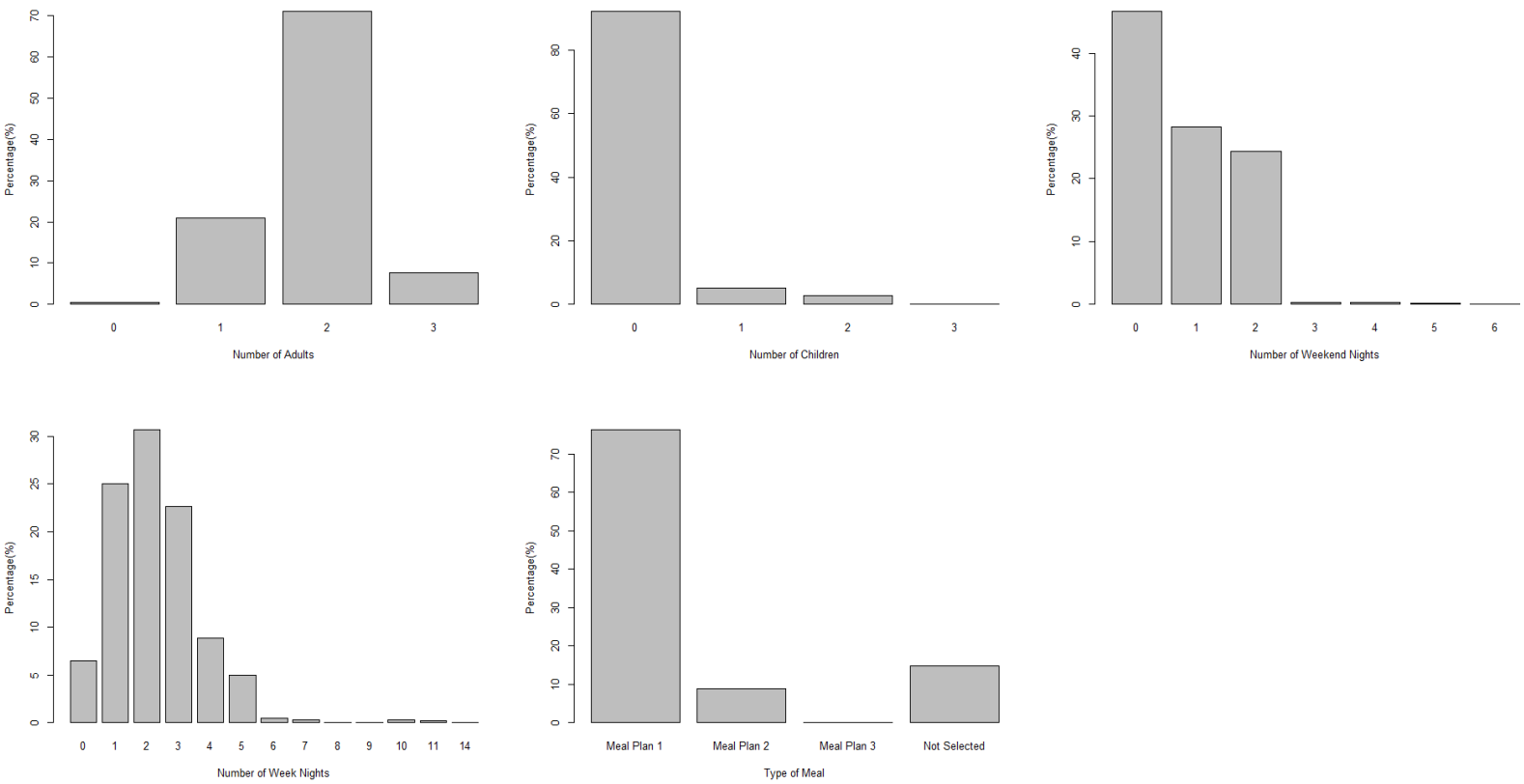
- Για την μεταβλητή number of adults, εμφανίζεται η κατηγορία FA έχοντας συχνότητα εμφάνισης μια φορά. Επειδή αυτή η τιμή δεν αντιπροσωπεύει τον αριθμό των ενηλίκων και ούτε ταιριάζει με τις υπόλοιπες κατηγορίες, θα την θεωρήσουμε ως μη λογική και ως τυπικό λάθος στην καταγραφή δεδομένων. Άρα, θα χρειαστεί να αφαιρέσουμε από όλο το σετ δεδομένων την παρατήρηση που αντιστοιχεί σε αυτήν την τιμή. Επίσης εμφανίζονται 11 παρατηρήσεις που αντιστοιχούν σε μηδέν αριθμό ενηλίκων. Αν δούμε πιο αναλυτικά αυτές τις 11 παρατηρήσεις βλέπουμε ότι οι 10 έχουν αριθμό ανήλικων ίσο με 2. Θα θεωρήσουμε ότι ενδεχομένως αυτές οι παρατηρήσεις να αντιστοιχούν σε εφήβους ή παιδιά κάτω των 18 που μπορούν να κάνουν κράτηση χωρίς την παρουσία κάποιου ενήλικα. Η παρατήρηση που θα αφαιρεθεί θα είναι η μια από τις 11, που αντιστοιχεί σε μηδέν αριθμό ενηλίκων και ανήλικων, ως λανθασμένη. Επίσης μπορεί αυτές οι 10 παρατηρήσεις να προέκυψαν από λάθος καταχώρηση της ηλικίας στο Online market type από το οποίο έγιναν.
- Για την μεταβλητή type of meal, εμφανίζονται οι κατηγορίες Nemo και Zari με συχνότητα εμφάνισης μια φορά όπως επίσης και μια κενή τιμή. Επειδή αυτές οι τιμές δεν αντιπροσωπεύουν τον τύπο γεύματος και ούτε ταιριάζουν με τις υπόλοιπες κατηγορίες, θα τις θεωρήσουμε ως μη λογικές και ως τυπικό λάθος στην καταγραφή δεδομένων. Άρα, θα χρειαστεί να αφαιρέσουμε από όλο το σετ δεδομένων τις παρατηρήσεις που αντιστοιχούν σε αυτές τις τιμές.
- Για την μεταβλητή type of room, εμφανίζονται δυο κενές τιμές τις οποίες θα χρειαστεί να αφαιρέσουμε. Άρα, θα αφαιρέσουμε από όλο το σετ δεδομένων τις παρατηρήσεις που αντιστοιχούν σε αυτές τις κενές τιμές.
- Για την μεταβλητή market segment type, εμφανίζονται οι κατηγορίες Tetaken και Aviation όπως επίσης και μια κενή τιμή. Επειδή αυτές οι τιμές δεν αντιπροσωπεύουν τον τύπο τμήματος της αγοράς και ούτε ταιριάζουν με τις υπόλοιπες κατηγορίες, θα τις θεωρήσουμε ως μη λογικές και ως τυπικό λάθος στην καταγραφή δεδομένων. Άρα, θα χρειαστεί να αφαιρέσουμε από όλο το σετ δεδομένων τις παρατηρήσεις που αντιστοιχούν σε αυτές τις τιμές.
- Για την μεταβλητή date of reservation, ελέγχοντας αν υπάρχει κάποια κενή τιμή, παρατηρήσαμε την εμφάνιση μιας τιμής την οποία θα χρειαστεί να αφαιρέσουμε από όλο το σετ δεδομένων. Επίσης αλλάζουμε το σύμβολο μεταξύ ημερών, μηνών και ετών ώστε όλες οι ημερομηνίες να έχουν τι ίδιο σύμβολο (‘/’). Μια επιπλέον τροποποίηση που κάνουμε κατά την διάρκεια των ερωτημάτων είναι να αφαιρέσουμε κάποιες ημερομηνίες που είναι λανθασμένες και κάποιες που ενώ είναι με άλλη μορφή, αλλάζοντας τες θα είναι λανθασμένες.



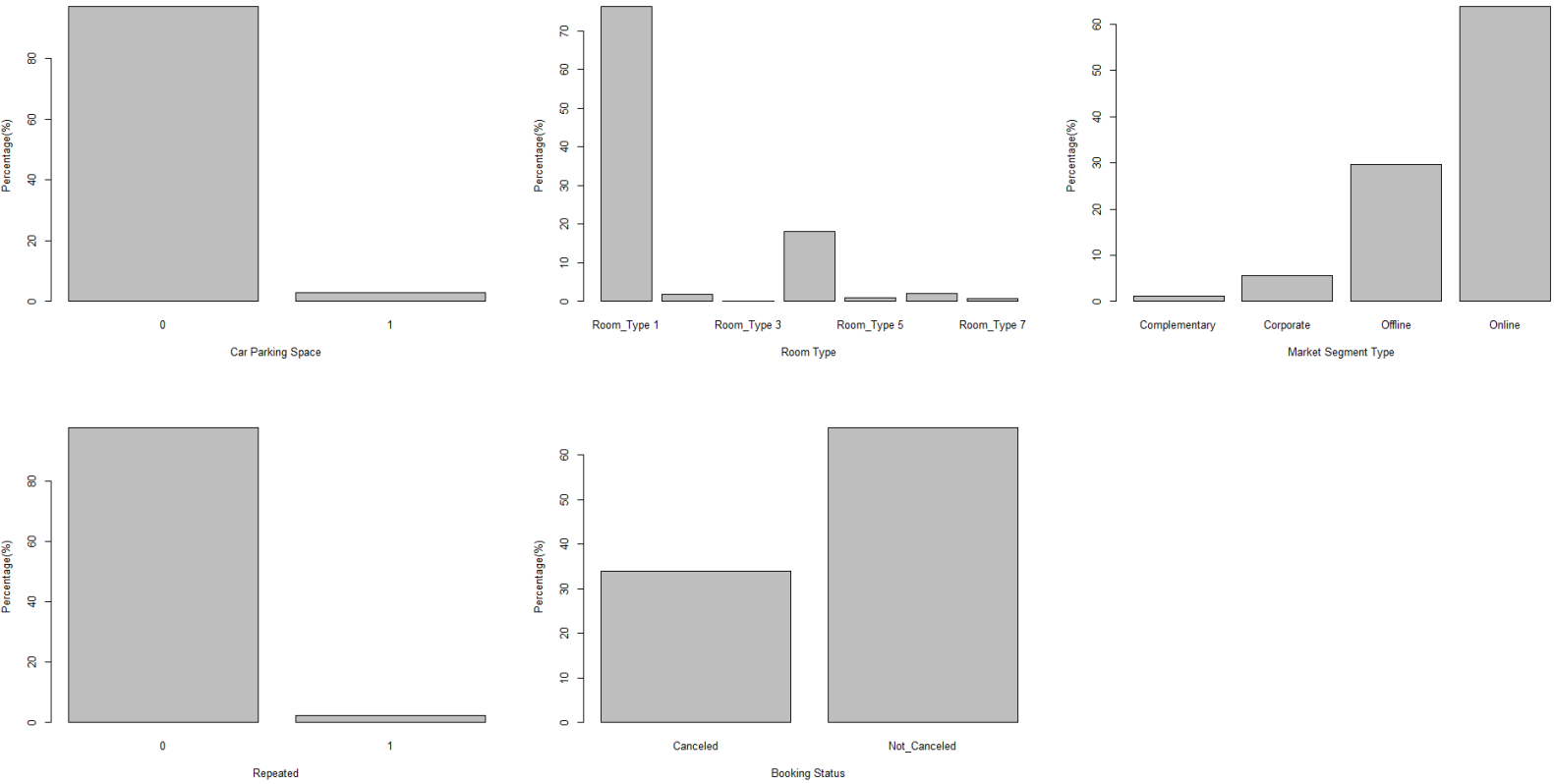
Σχήμα 1: Διαγράμματα πυκνότητας πιθανότητας για τις μεταβλητές lead time, P.C, P.not.C, special requests και average price



Σχήμα 2: Θηκόγραμμα για τις μεταβλητές lead time, P.C, P.not.C, special requests και average price



Σχήμα 3: Ραβδόγραμμα για τις μεταβλητές number of adults, number of children, number of weekend nights, number of week nights και type of meal

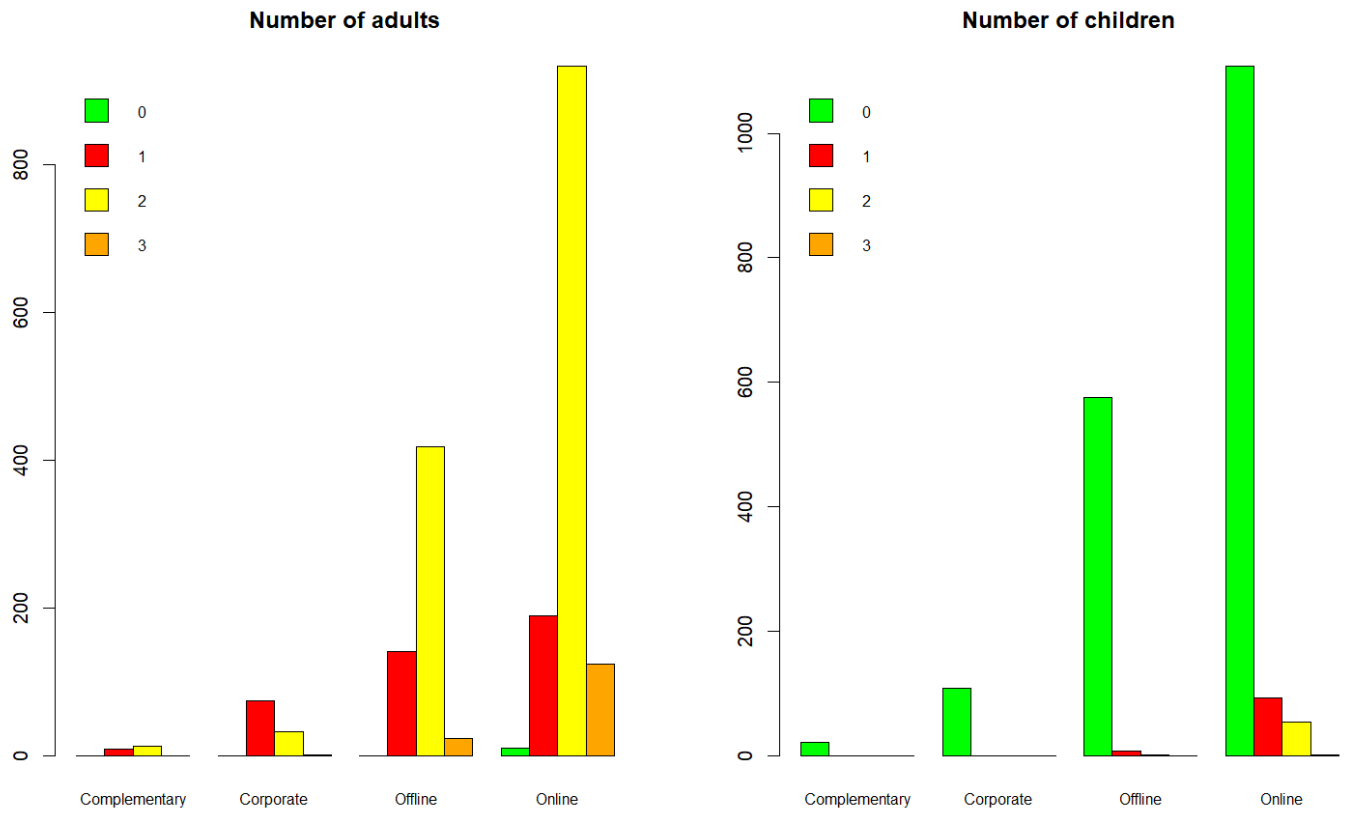


Σχήμα 4: Ραβδόγραμμα για τις μεταβλητές car parking space, room type, market segment type, repeated και booking status

2. ΑΝΑΛΥΣΗ

Ερώτημα 2

Μελετώντας τον αριθμό των ενηλίκων και των παιδιών σε κάθε τύπο τμήματος της αγοράς που σχετίζεται με την κράτηση, βρήκαμε τις αντίστοιχες συχνότητες για τα ζεύγη number of adults/market segment type και number of children/market segment type. Από τις συχνότητες αλλά και από τα παρακάτω διαγράμματα (Σχήμα 5) παρατηρούμε ότι οι online κρατήσεις είναι οι υψηλότερες τόσο για τον αριθμό των ενηλίκων όσο και για τον αριθμό των παιδιών. Οι offline κρατήσεις έχουν επίσης μεγάλο αριθμό κρατήσεων, κυρίως για κρατήσεις με 2 ενήλικες και σχεδόν καθόλου παιδιά. Οι corporate κρατήσεις αφορούν 1 ή 2 ενήλικες, χωρίς καθόλου παιδιά. Οι complementary κρατήσεις έχουν τις λιγότερες κρατήσεις συνολικά, με πολύ λίγους ενήλικες και καθόλου παιδιά.



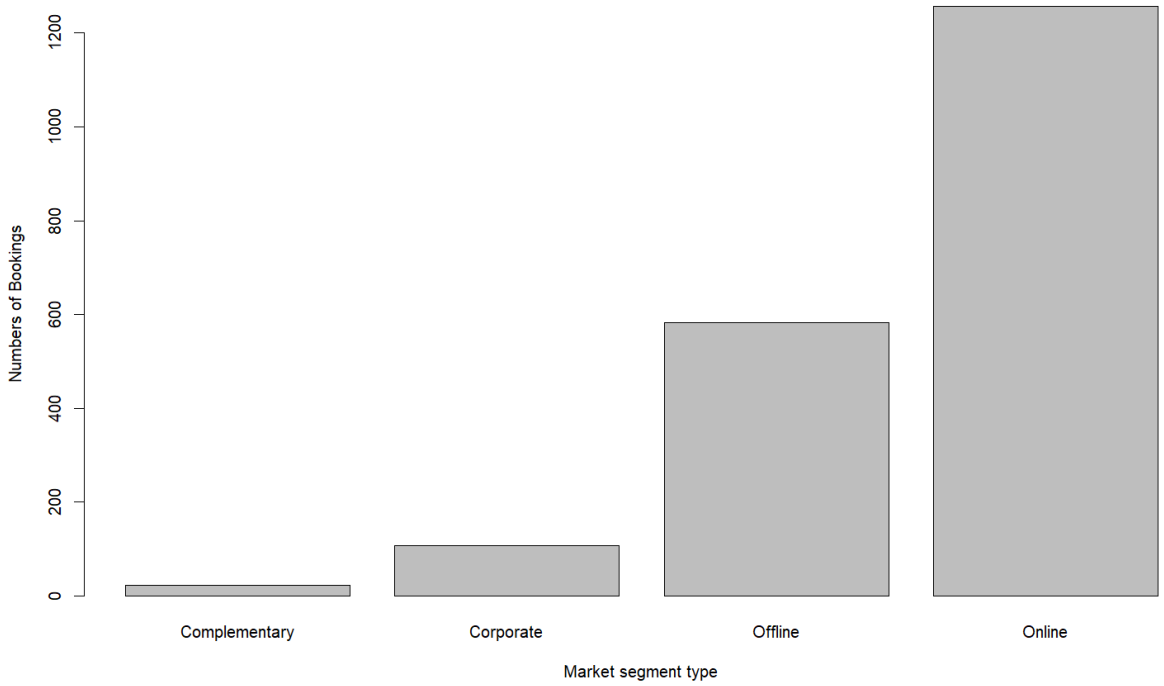
Σχήμα 5: Ραβδόγραμμα για τις συχνότητες των number of adults/market segment type και number of children/market segment type

Ερώτημα 3

Κάνοντας τους πίνακες συχνοτήτων για τις μεταβλητές number of weekend nights/market segment type και number of week nights/market segment type και βρίσκοντας τα περιθώρια αθροίσματα των στηλών συμπεραίνουμε ότι οι online κρατήσεις έχουν τις περισσότερες διανυκτερεύσεις Σαββατοκύριακου ενώ οι complementary κρατήσεις έχουν τις λιγότερες εβδομαδιαίες διανυκτερεύσεις.

Ερώτημα 4

Από τον πίνακα συχνοτήτων για την μεταβλητή market segment type αλλά και από το παρακάτω διάγραμμα παρατηρούμε ότι οι περισσότερες κρατήσεις έχουν γίνει online με συχνότητα 1257 κρατήσεις (Σχήμα 6).



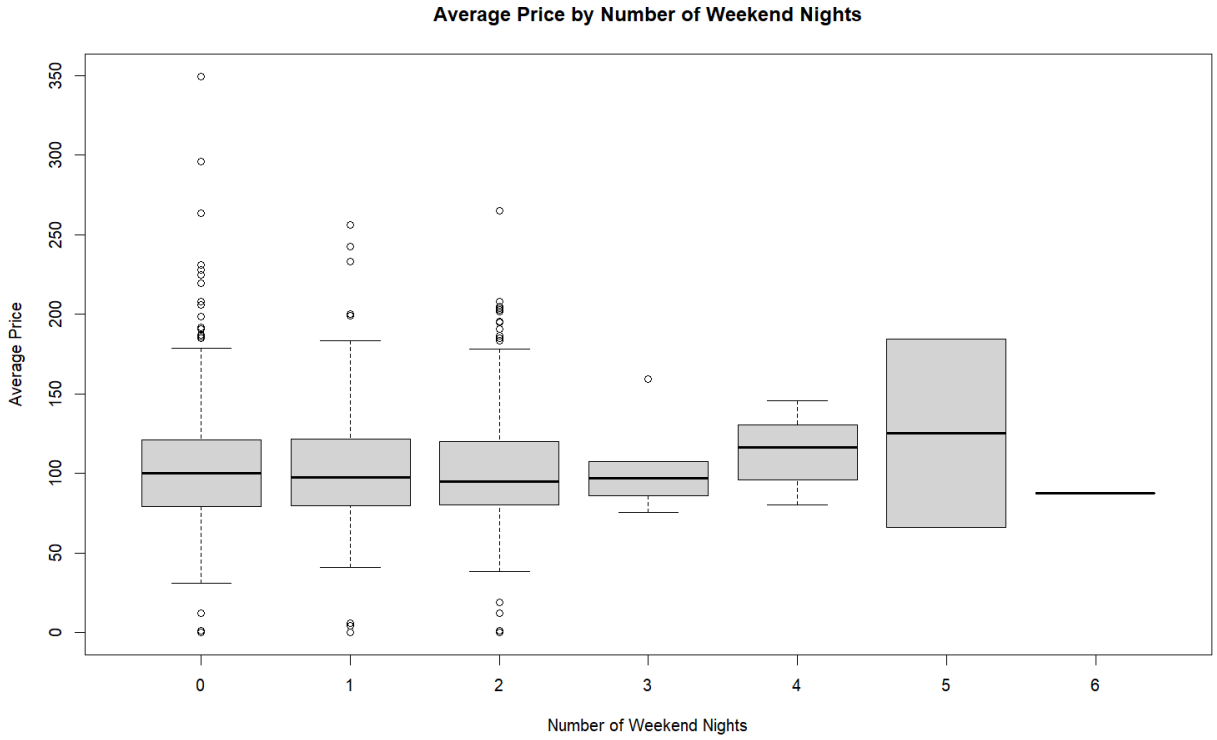
Σχήμα 6: Ραβδόγραμμα για τις συχνότητες της μεταβλητής market segment type

Ερώτημα 5

Από τα περιγραφικά μέτρα αλλά και από τα παρακάτω διαγράμματα (Σχήμα 9) παρατηρούμε ότι για αριθμό διανυκτερεύσεων από 0 έως 2 η κατανομή της μέσης τιμής είναι θετικά ασύμμετρη αφού η μέση τιμή είναι μεγαλύτερη από την διάμεσο. Από το εύρος των τιμών για 0 διανυκτερεύσεις η κατανομή παρουσιάζει μεγάλη μεταβλητότητα ενώ για αριθμό διανυκτερεύσεων 1 και 2 η μεταβλητότητα των τιμών είναι πιο μέτρια. Οι ακραίες τιμές τους υποδηλώνουν ότι ορισμένες κρατήσεις είναι

σημαντικά ακριβότερες. Για αριθμό διανυκτερεύσεων ίσο με 3 παρατηρούμε ότι η κατανομή της μέσης τιμής είναι θετικά ασύμμετρη με στενότερο εύρος ενώ για αριθμό διανυκτερεύσεων ίσο με 4 παρατηρούμε ότι η κατανομή είναι αρνητικά ασύμμετρη. Για αριθμό διανυκτερεύσεων ίσο με 5 παρατηρούμε μια συμμετρική κατανομή με μεγάλο ενδοτεταρτημοριακό εύρος. Τέλος για αριθμό διανυκτερεύσεων ίσο με 6 έχουμε την ύπαρξη μόνο μιας παρατήρησης.

Λόγω της διασποράς των τιμών και της παρουσίας ακραίων τιμών υπάρχει μεταβλητότητα στις μέσες τιμές κράτησης η οποία όμως τείνει να μειώνεται καθώς αυξάνεται ο αριθμός των διανυκτερεύσεων Σαββατοκύριακου.

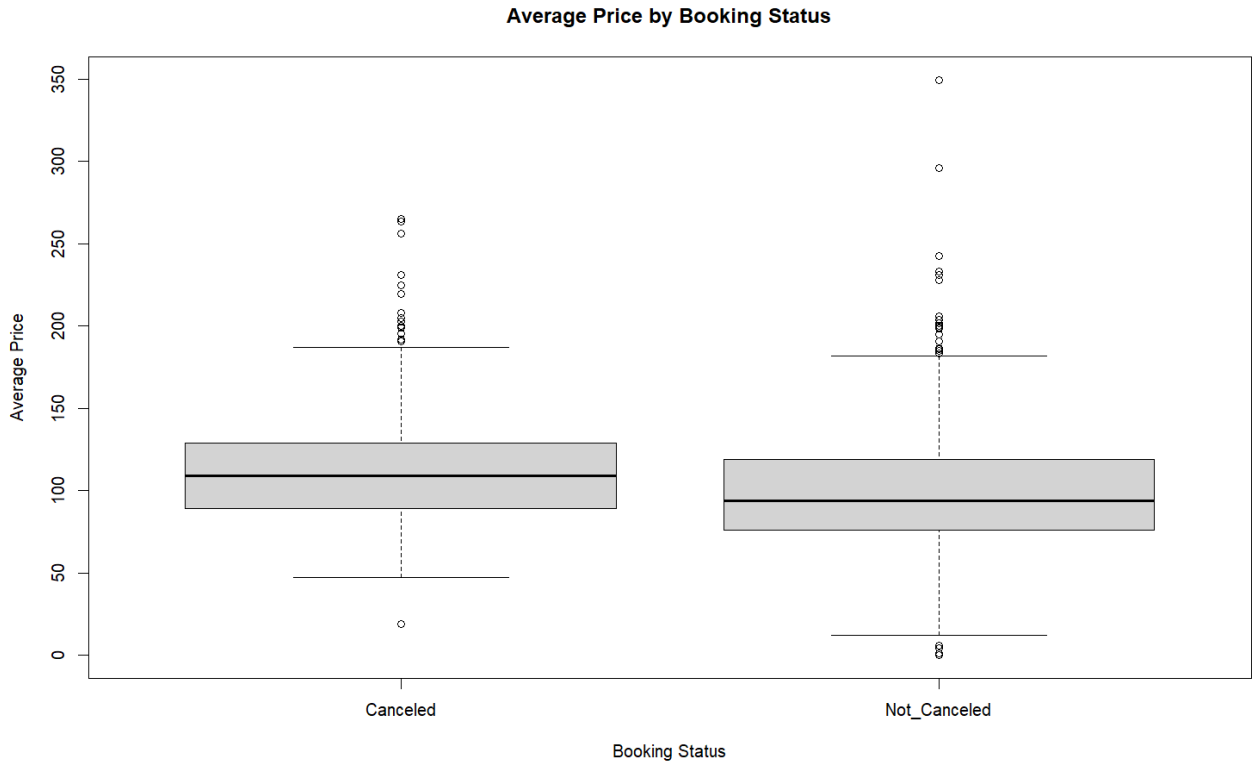


Σχήμα 7: Θηκόγραμμα για την μέση τιμή κρατήσεων με βάση τον αριθμό διανυκτερεύσεων Σαββατοκύριακου

Ερώτημα 6

Από τα περιγραφικά μέτρα αλλά και από τα παρακάτω διαγράμματα (Σχήμα 10) παρατηρούμε ότι οι ακυρωμένες κρατήσεις τείνουν να έχουν υψηλότερες μέσες τιμές παρατηρώντας το 25%, το 50% και το 75% των παρατηρήσεων όπως επίσης και τον μέσο όρο, σε σύγκριση με τις μη ακυρωμένες κρατήσεις. Ωστόσο, η μέγιστη τιμή είναι υψηλότερη για τις μη ακυρωμένες κρατήσεις. Άρα υπάρχουν κρατήσεις με πολύ υψηλές τιμές που δεν ακυρώνονται, γεγονός που επηρεάζει σημαντικά το εύρος και τη μεταβλητότητα των τιμών.

Συμπερασματικά, οι ακυρωμένες κρατήσεις έχουν γενικά υψηλότερες μέσες τιμές, όμως, οι μη ακυρωμένες κρατήσεις έχουν μεγαλύτερο εύρος μέσων τιμών με πιο έντονες ακραίες τιμές.



Σχήμα 8: Θηκόγραμμα για την μέση τιμή κρατήσεων με βάση την κατάσταση της κράτησης (ακυρωμένη ή μη ακυρωμένη)

Ερώτημα 7

Ελέγχουμε, αρχικά, την σχέση της μεταβλητής average price με τις μεταβλητές lead time, P.C, P.not.C και special requests εφαρμόζοντας μη παραμετρικό έλεγχο kendall, λόγω μεγαλύτερης αξιοπιστίας έναντι ακραίων τιμών και μη κανονικότητας. Βρίσκοντας τον αντίστοιχο συντελεστή συσχέτισης παρατηρούμε ότι οι τιμές είναι πολύ κοντά στο 0 υποδεικνύοντας για όλες τις μεταβλητές μια ασθενής συσχέτιση.

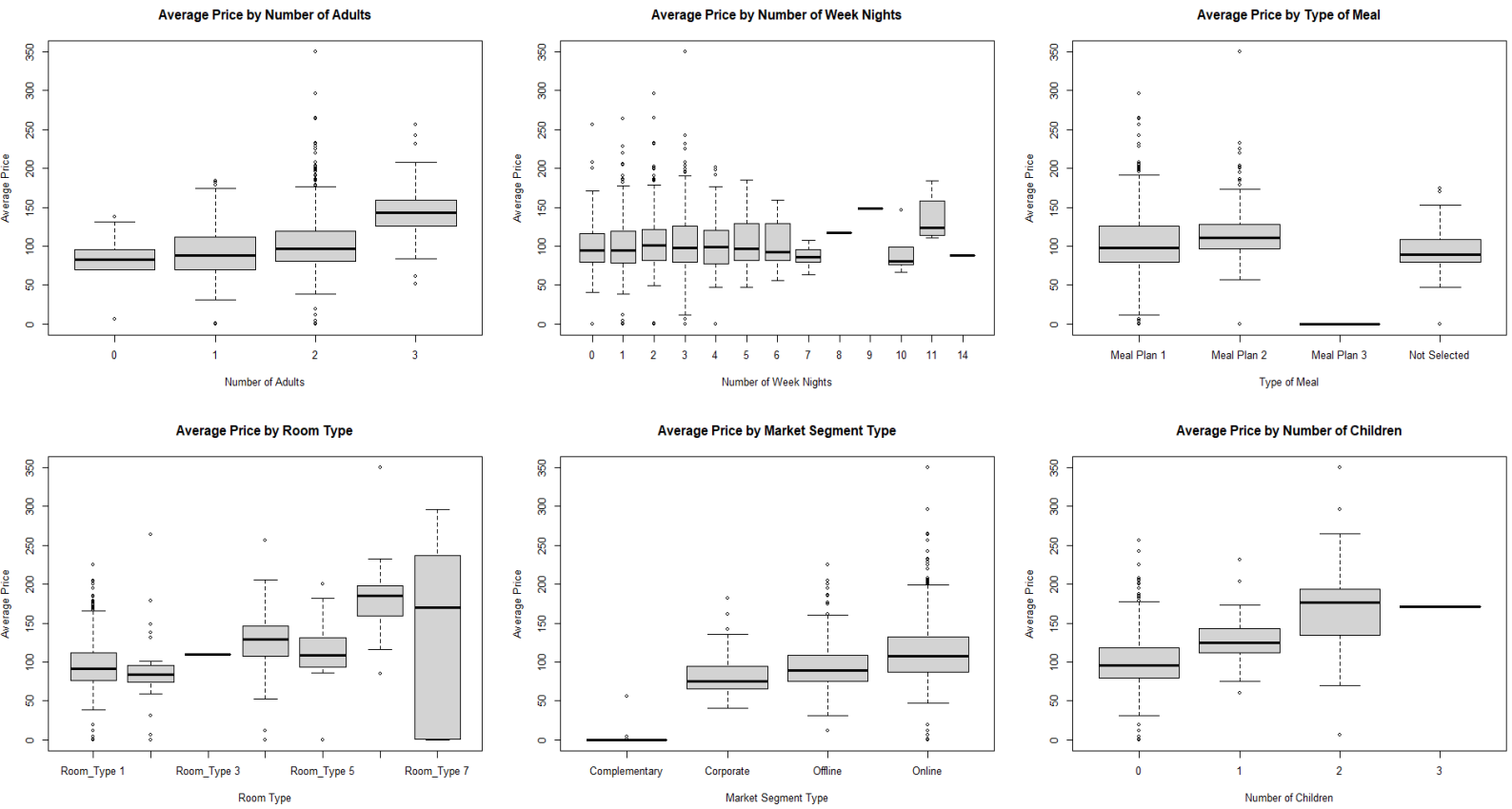
Επίσης, έχοντας ερευνήσει την σχέση της average price με τις μεταβλητές booking status και number of weekend nights στα παραπάνω ερωτήματα, θα μπορούσαμε να ερευνήσουμε και τις εξής σχέσεις:

- average price ~ number of adults
- average price ~ number of week nights
- average price ~ type of meal
- average price ~ room type
- average price ~ market segment type
- average price ~ number of children

Εφαρμόζοντας κατάλληλα διαγράμματα παρατηρούμε ότι:

- Για τον αριθμό ενηλίκων, η μέση τιμή φαίνεται να αυξάνεται όσο αυξάνεται ο αριθμός των ενηλίκων. Υπάρχει μια ανοδική τάση.
- Για τον αριθμό διανυκτερεύσεων την εβδομάδα υπάρχει μεταβλητότητα στη μέση τιμή. Οι μέσες τιμές φαίνεται να αυξομειώνονται σε διάφορους αριθμούς εβδομαδιαίων διανυκτερεύσεων.
- Για τον τύπο γεύματος παρατηρούμε διαφορετικές μέσες τιμές. Το γεύμα τύπου 1 και 2 έχουν παρόμοιες κατανομές τιμών, ενώ το γεύμα τύπου 3 έχει χαμηλότερο και στενότερο εύρος τιμών.
- Για τον τύπο δωματίου υπάρχουν σημαντικές διαφορές στις μέσες τιμές. Ο τύπος δωματίου 7 έχει σημαντικά μεγάλο εύρος μέσων τιμών, γεγονός που υποδηλώνει μεγάλη μεταβλητότητα.
- Για τον τύπο τμήματος της αγοράς υπάρχουν κάποιες διαφορές στις μέσες τιμές μεταξύ των τμημάτων της αγοράς. Παρατηρούμε για το κάθε τμήμα της αγοράς μια ανοδική τάση της μέσης τιμής αλλά και μια αύξηση της μεταβλητότητας των τιμών.
- Για τον αριθμό παιδιών, η μέση τιμή φαίνεται να αυξάνεται όσο αυξάνεται ο αριθμός των παιδιών.

Με βάση την παραπάνω ανάλυση, ο αριθμός των ενηλίκων, ο αριθμός των παιδιών και ο τύπος τμήματος της αγοράς μπορεί να έχουν θετική συσχέτιση με τη μέση τιμή κράτησης, ενώ ο αριθμός των διανυκτερεύσεων ανά εβδομάδα και ο τύπος δωματίου μπορεί να έχουν ασθενέστερη ή καθόλου συσχέτιση. Ο τύπος γεύματος φαίνεται να έχει μέτρια συσχέτιση (Σχήμα 9).



Σχήμα 9: Θηκόγραμμα για την μέση τιμή κρατήσεων με βάση τις μεταβλητές number of adults, number of week nights, type of meal, room type, market segment type και number of children

Ερώτημα 8

Από τον πίνακα συχνοτήτων για τις μεταβλητές booking status/market segment type παρατηρούμε ότι οι περισσότερες κρατήσεις τόσο για αυτές που ακυρώθηκαν όσο και για αυτές που δεν ακυρώθηκαν έγιναν online με συχνότητα 480 και 777 κρατήσεις αντίστοιχα.

Ερώτημα 9

Για να βρούμε τους τρεις τύπους δωματίου με το μεγαλύτερο πλήθος ειδικών αιτημάτων χωρίζουμε την μεταβλητή special requests ανά τύπο δωματίου με την χρήση της aggregate εντολής και αθροίζουμε τις τιμές της για κάθε τύπο δωματίου. Έπειτα, ταξινομούμε το πλαίσιο δεδομένων σε φθίνουσα σειρά με βάση τη στήλη των τιμών special requests και διαλέγουμε τις 3 πρώτες γραμμές. Το αρνητικό πρόσημο μέσα στην ταξινόμηση εξασφαλίζει ότι η ταξινόμηση θα γίνει με φθίνουσα σειρά. Ως αποτέλεσμα, παίρνουμε τους τρεις τύπους δωματίων με το μεγαλύτερο άθροισμα των special requests.

Room Type	Special Requests
Room_Type 1	**845**
Room_Type 4	**315**
Room_Type 6	**40**

Πίνακας 1.

Ακριβώς το ίδιο κάνουμε για να βρούμε τους τρεις τύπους δωματίου με το μεγαλύτερο πλήθος ημερών μεταξύ της ημερομηνίας κράτησης και της ημερομηνίας άφιξης. Χωρίζουμε την μεταβλητή lead time ανά τύπο δωματίου με την χρήση της aggregate εντολής και αθροίζουμε τις τιμές της για κάθε τύπο δωματίου. Έπειτα, ταξινομούμε το πλαίσιο δεδομένων σε φθίνουσα σειρά με βάση τη στήλη των τιμών lead time και διαλέγουμε τις 3 πρώτες γραμμές. Ως αποτέλεσμα, παίρνουμε τους τρεις τύπους δωματίων με το μεγαλύτερο άθροισμα τιμών της lead time.

Room Type	Lead Time
Room_Type 1	**132697**
Room_Type 4	**23880**
Room_Type 2	**3577**

Πίνακας 2.

Για το μεγαλύτερο πλήθος ενηλίκων που περιλαμβάνονται στην κράτηση χρειάζεται να φτιάξουμε τον πίνακα συχνοτήτων για τις μεταβλητές number of adults και room type, να αθροίσουμε τις συχνότητες ως προς τις στήλες, δηλαδή για κάθε τύπο δωματίου, και ύστερα να ταξινομήσουμε τα αθροίσματα σε φθίνουσα σειρά και να κρατήσουμε τις 3 πρώτες τιμές. Ως αποτέλεσμα, παίρνουμε τους τρεις τύπους δωματίων με το μεγαλύτερο άθροισμα ενηλίκων.

Room Type	Number of Adults
Room_Type 1	**1502**
Room_Type 4	**358**
Room_Type 6	**41**

Πίνακας 3.

Ερώτημα 10

Φτιάχνουμε μια συνάρτηση μέσα στην οποία ορίζουμε μια κενή λίστα με όνομα top_bookings για να αποθηκεύσει τις πιο ακριβές κρατήσεις για κάθε τύπο δωματίου. Έπειτα, βρίσκουμε όλους τους μοναδικούς τύπους δωματίων και τους αποθηκεύουμε στην μεταβλητή room_type. Μέσα σε μια επαναληπτική διαδικασία για κάθε τύπο δωματίου, κρατάμε από τα δεδομένα τις γραμμές για τις οποίες η μεταβλητή room type ταιριάζει με τον τρέχοντα τύπο δωματίου αποθηκεύοντας τες στη μεταβλητή room_data. Υστέρα, ελέγχουμε αν ο αριθμός των γραμμών στη μεταβλητή room_data

είναι μικρότερος από 5. Εάν υπάρχουν λιγότερες από 5 κρατήσεις για τον τρέχοντα τύπο δωματίου, δημιουργούμε ένα μήνυμα το οποίο θα εμφανίζεται αποθηκεύοντας το στο top_bookings. Εάν υπάρχουν 5 ή περισσότερες κρατήσεις για τον τρέχοντα τύπο δωματίου, ταξινομούμε τα room_data σε φθίνουσα σειρά με βάση την μεταβλητή average price και αποθηκεύουμε τα ταξινομημένα δεδομένα στη μεταβλητή sort. Στην συνέχεια, από τα ταξινομημένα δεδομένα sort κρατάμε τις 5 πρώτες τιμές της μεταβλητής Booking_ID και τις αποθηκεύουμε στη μεταβλητή top5. Αποθηκεύουμε την μεταβλητή top5 στη λίστα top_bookings. Αφού ολοκληρωθεί η επαναληπτική διαδικασία, η συνάρτηση θα επιστρέφει τη λίστα top_bookings.

Ερώτημα 11

Για να βρούμε την πιο ακριβή κράτηση για κάθε τύπο γεύματος αρχικά θα χρειαστεί να χωρίσουμε τις τιμές της μεταβλητής average price με βάση τον κάθε τύπο γεύματος. Χρησιμοποιώντας την εντολή split θα πάρουμε μια λίστα η οποία θα περιλαμβάνει τις μέσες τιμές που αντιστοιχούν σε κάθε τύπο γεύματος. Έπειτα, μέσα σε αυτήν την λίστα θα εφαρμόσουμε την συνάρτηση function(x) max(x) οπού θα μας επιστρέψει την μέγιστη μέση τιμή για το κάθε στοιχείο της, δηλαδή για κάθε τύπο γεύματος. Για να αναζητήσουμε το booking id που αντιστοιχεί στον κάθε τύπο γεύματος και στην μέγιστη μέση τιμή τους, θα αναζητήσουμε από το σετ δεδομένων τις θέσεις γραμμών για τις οποίες ο τύπος γεύματος και η μέση τιμή αντιστοιχούν στις μέγιστες μέσες τιμές του κάθε γεύματος. Αφού βρούμε τις θέσεις, βρίσκουμε και τα αντίστοιχα booking id τα οποία τα βάζουμε σε ένα πλαίσιο δεδομένων μαζί με τους αντίστοιχους τύπους γεύματος. Ως αποτέλεσμα των παραπάνω έχουμε τα εξής:

	Booking Id
Meal Plan 1	**BID28235**
Meal Plan 2	**BID34307**
Meal Plan 3	**BID28595**
Not Selected	**BID33953**

Πίνακας 4.

Ερώτημα 12.

Εφαρμόζουμε ακριβώς τα ίδια βήματα με αυτά του προηγούμενου ερωτήματος μόνο που εδώ τυποποιούμε τις μέσες τιμές της μεταβλητής average price πριν ξεκινήσουμε να βρούμε την πιο ακριβή κράτηση για κάθε τύπο γεύματος. Τα αποτελέσματα που παίρνουμε σε αυτό το ερώτημα είναι τα εξης:

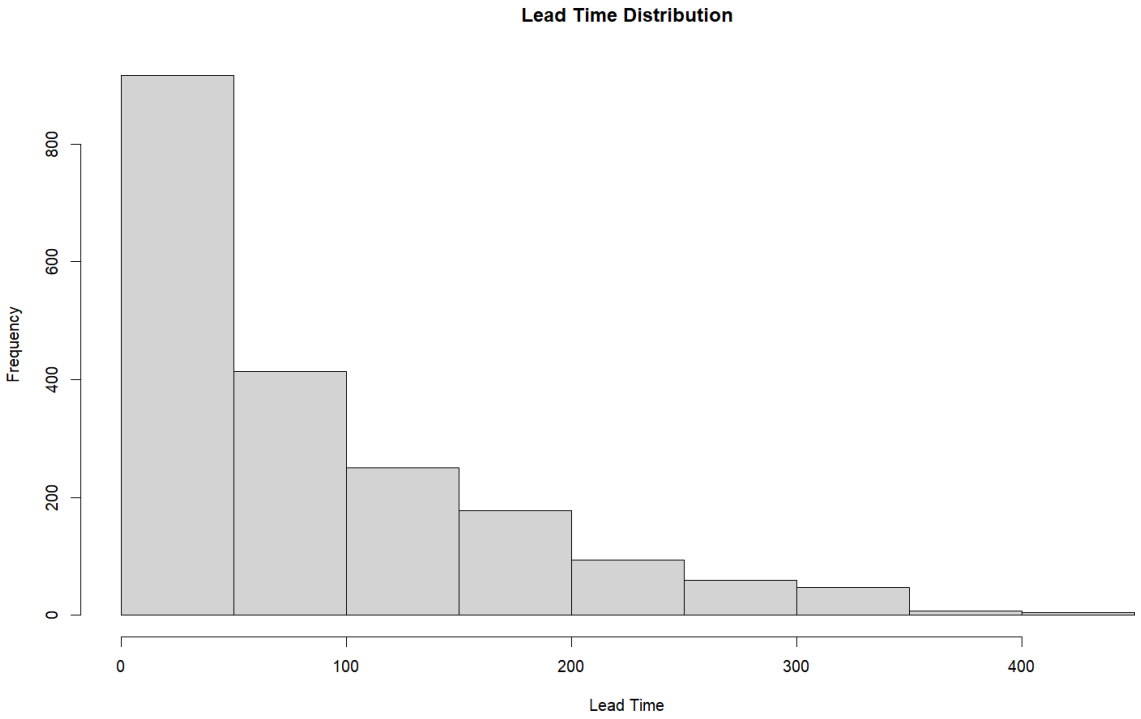
	Booking Id2
Meal Plan 1	**BID28235**
Meal Plan 2	**BID34307**
Meal Plan 3	**BID28595**
Not Selected	**BID33953**

Πίνακας 5.

Παρατηρούμε ότι παρόλο που τυποποιήσαμε τις μέσες τιμές των κρατήσεων, τα αποτελέσματα είναι ακριβώς τα ίδια με αυτά του προηγούμενου ερωτήματος.

Ερώτημα 13

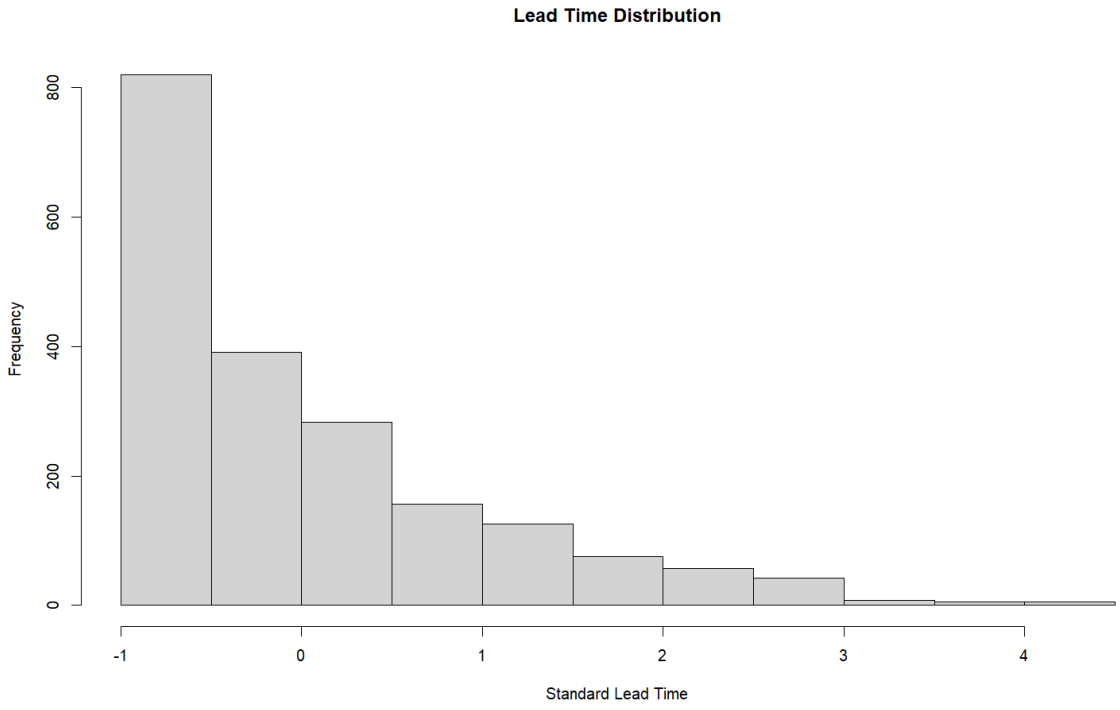
Υπολογίζοντας την ασυμμετρία και την κύρτωση για τον αριθμό των ημερών μεταξύ της ημερομηνίας κράτησης και της ημερομηνίας άφιξης παρατηρούμε ότι έχουμε θετική ασυμμετρία, άρα η δεξιά ουρά της κατανομής είναι μακρύτερη από την αριστερή, και ότι η κατανομή είναι πλατύκυρτη. Την κατανομή της μεταβλητής lead time μπορούμε να την παρατηρήσουμε και από το παρακάτω διάγραμμα καταλήγοντας στα ιδιά συμπεράσματα. Την θετική ασυμμετρία την γνωρίζαμε και από τον πίνακα περιγραφικών μέτρων που εφαρμόσαμε στην εισαγωγή αφού η μέση τιμή είναι μεγαλύτερη από την διάμεσο.



Σχήμα 7: Διάγραμμα πυκνότητας πιθανότητας για την μεταβλητή lead time

Ερώτημα 14

Τυποποιώντας την μεταβλητή lead time παρατηρούμε ότι αυξήθηκε ελάχιστα η τιμή της κύρτωσης, με την ασυμμετρία να παραμένει ίδια, χωρίς όμως να αλλάξει η κατανομή της μεταβλητής. Άρα η μεταβλητή παραμένει πλατύκυρτη έχοντας θετική ασυμμετρία.



Σχήμα 8: Διάγραμμα πυκνότητας πιθανότητας για την τυποποιημένη μεταβλητή lead time

Ερώτημα 15

Για να βρούμε ένα διάστημα που περιλαμβάνει τουλάχιστον το 93.75% των εβδομαδιαίων διανυκτερεύσεων που περιλαμβάνονται στην κράτηση αρκεί να βρούμε τον αριθμό των διανυκτερεύσεων όπου η αθροιστική αναλογία είναι μικρότερη ή ίση με 0,9375. Βρίσκουμε τον πίνακα συχνοτήτων για την μεταβλητή number of week nights και το αθροιστικό άθροισμα των ποσοστών κάθε τιμής για τις αντίστοιχες σχετικές συχνότητες της μεταβλητής. Δηλαδή, μετατρέπουμε τον πίνακα σε αναλογίες και υπολογίζουμε το αθροιστικό άθροισμα αυτών των αναλογιών. Έπειτα εφαρμόζουμε την συνθήκη που αναφέραμε για τον αριθμό των διανυκτερεύσεων. Δηλαδή, βρίσκουμε τον αριθμό των διανυκτερεύσεων που αντιστοιχούν για αθροιστική αναλογία μικρότερη ή ίση του 0,9375. Τέλος, από τον αριθμό των εβδομαδιαίων διανυκτερεύσεων διαλέγουμε τον μικρότερο και μεγαλύτερο αριθμό.

Number of Week Nights	
‘0’	**0.06497462**
‘1’	**0.31522843**
‘2’	**0.62182741**
‘3’	**0.84873096**
‘4’	**0.93705584**

Πίνακας 6: Αριθμός εβδομαδιαίων διανυκτερεύσεων για αθροιστική αναλογία μικρότερη ή ίση του 0,9375

Ερώτημα 16

Για να βρούμε το σύνολο της τιμής των κρατήσεων που πραγματοποιήθηκαν ανά μήνα θα χρειαστεί αρχικά να ορίσουμε ως μεταβλητή dates την μεταβλητή date of reservation ως αντικείμενο ημερομηνιών με μορφή μήνας/μέρα/έτος. Υστέρα, θα ορίσουμε την μεταβλητή months που θα επιστρέφει από τις παραπάνω ημερομηνίες μόνο τους μήνες. Βλέποντας τις τιμές για την μεταβλητή months παρατηρούμε κάποιες Na τιμές. Για τις αντίστοιχες θέσεις, λοιπόν, μελετάμε την μεταβλητή των ημερομηνιών και παρατηρούμε ότι τιμές αυτές θα χρειαστεί να αφαιρεθούν όπως είχαμε αναφέρει αναλυτικά και στην εισαγωγή. Αφού έχουν γίνει όλες οι απαραίτητες τροποποιήσεις, χωρίζουμε τις μέσες τιμές κρατήσεων με βάση τους μήνες των ημερομηνιών εμφανίζοντας τα αποτέλεσμα σε μια λίστα και με την χρήση της sapply θα αθροίσουμε για κάθε στοιχείο της, δηλαδή για κάθε μήνα, τις μέσες τιμές που περιλαμβάνουν. Ενναλακτικά, θα μπορούσαμε να πάρουμε τις μέσες τιμές για κάθε μήνα και να τις αθροίσουμε μόνο με την χρήση της aggregate εντολής.

Months	Average Price
01	**4792.62**
02	**7478.13**
03	**12424.80**
04	**15834.22**
05	**15505.02**
06	**21460.64**
07	**16397.14**
08	**22660.82**
09	**28056.44**
10	**27125.34**
11	**16322.46**
12	**14528.21**

Πίνακας 7.

Ερώτημα 17

Δημιουργούμε μία ακολουθία αριθμών με βήμα 5 που θα έχει τιμές από το 1 μέχρι το 23 και ορίζουμε την μεταβλητή average_price_missing ως την μεταβλητή average price, δηλαδή ως την μέση τιμή κρατήσεων. Στις θέσεις που αντιστοιχούν με τα νούμερα τις ακολουθίας, δηλώνουμε τις τιμές της μεταβλητής average_price_missing ως Na τιμές. Υστέρα, συγκρίνοντας τα περιγραφικά μέτρα για την νέα μεταβλητή average_price_missing και για την average price παρατηρούμε μια ελάχιστη μείωση στον μέσο όρο η οποία δεν θεωρείτε σημαντική (από 102.84 σε 102.83). Άρα η ύπαρξη των 5 Na τιμών δεν επηρέασαν σημαντικά την κατανομή της μέσης τιμής κρατήσεων.

Ερώτημα 18

Ορίζουμε την μέγιστη και ελάχιστη τιμή της average_price_missing δηλώνοντας την ύπαρξη Na τιμών με την χρήση της παραμέτρου na.rm = TRUE και ύστερα δημιουργούμε μια ακολουθία τυχαίων αριθμών από ομοιόμορφη κατανομή όπου θα παίρνει τιμές από την ελάχιστη μέχρι και την μέγιστη τιμή της average_price_missing με πλήθος όσες και οι Na τιμές της. Στην συνέχεια, ορίζουμε την μεταβλητή average_price_new ως την μεταβλητή average_price_missing και δηλώνουμε στις θέσεις όπου υπάρχουν οι Na τιμές της average_price_missing, τις τιμές της ακολουθίας. Άρα, η νέα μεταβλητή average_price_new θα περιλαμβάνει όλες τις τιμές της average_price_missing αλλά στις θέσεις των ελλειπουσών τιμών θα περιλαμβάνει τις νέες τιμές τις ακολουθίας.

Ερώτημα 19

Αρχικά, από το σετ δεδομένων αποθηκεύουμε όλες τις γραμμές για τις οποίες το average price είναι μεγαλύτερο ή ίσο του 50 και ορίζουμε την νέα μεταβλητή average_price_group για την οποία αν οι μέσες τιμές κράτησης είναι μεγαλύτερες ή ίσες του 50 και μικρότερες του 110 , θα παίρνει την τιμή ‘Low’ , αν οι μέσες τιμές κράτησης είναι μεγαλύτερες ή ίσες του 110 και μικρότερες του 170 θα παίρνει την τιμή ‘Medium’, αλλιώς θα παίρνει την τιμή ‘High’. Δηλώνουμε την νέα μεταβλητή ως κατηγορική. Για να βρούμε πόσες είναι οι προηγούμενες κρατήσεις που ακυρώθηκαν από τον πελάτη πριν από την τρέχουσα κράτηση και πόσες είναι οι προηγούμενες κρατήσεις που δεν ακυρώθηκαν από τον πελάτη πριν από την τρέχουσα κράτηση για κάθε κατηγορία της μεταβλητής average_price_group αρκεί να πάρουμε τις τιμές της μεταβλητής P.C, και της P.not.C αντίστοιχα, που θα αντιστοιχούν σε κάθε κατηγορία της average_price_group με την χρήση της εντολής aggregate και ορίζοντας μέσα σε αυτήν μια κατάλληλη συνάρτηση να αθροίσουμε τις τιμές για κάθε κατηγορία. Μετά την διαδικασία αυτή, παρατηρούμε τα εξής αποτελέσματα:

Average Price Group	P.not.C
High	**0**
Low	**265**
Medium	**3**

Πίνακας 8.

Average Price Group	P.C
High	**0**
Low	**20**
Medium	**0**

Πίνακας 9.

Ερώτημα 20

Αρχικά, από το σετ δεδομένων αποθηκεύουμε όλες τις γραμμές για τις οποίες το average price είναι μεγαλύτερο ή ίσο του 100. Για το νέο σετ δεδομένων θα βρούμε τον πίνακα συχνοτήτων για τις μεταβλητές number of adults, number of children και booking status, όπου θα χωρίζεται σε 2 μέρη καθώς έχουμε 3 μεταβλητές. Έπειτα, από τον πίνακα συχνοτήτων θα ερευνήσουμε αν υπάρχει κάποια συχνότητα που να αντιστοιχεί για αριθμό ενηλίκων ίσο με '1', για αριθμό παιδιών ίσο με '2' και για τύπο κράτησης ίσο με 'Not_Canceled'. Επίσης θα ερευνήσουμε αν υπάρχει κάποια συχνότητα που να αντιστοιχεί για αριθμό ενηλίκων ίσο με '1', για αριθμό παιδιών ίσο με '3' και για τύπο κράτησης ίσο με 'Not_Canceled'. Και στα 2 ενδεχόμενα το αποτέλεσμα μας είναι μηδενικό άρα δεν υπάρχουν πελάτες που να έχουν κάνει κράτηση που δεν ακυρώθηκε, με 1 ενήλικο μέλος, με 2 ή 3 παιδιά , και με μέση τιμή κράτησης μεγαλύτερη ή ίση του 100. Συμπερασματικά, κανένας πελάτης δεν θα λάβει δωρεάν κράτηση.