***Πρόβλημα 1.***

Το αρχείο «H.1.Mat_Tim.sav» περιέχει τις μεταβλητές id: αριθμός δοκιμίου, Time: χρόνος κατεργασίας και Material: υλικό με τρεις στάθμες Low, Medium και High. Στόχος της μελέτης είναι να βρεθεί το υλικό (ή τα υλικά) που χρειάζεται το μικρότερο χρόνο επεξεργασίας. Ο στατιστικός που το ανέλαβε ακολούθησε την παρακάτω διαδικασία την οποία πρέπει να επαναλάβετε με τις κατάλληλες εντολές και να σχολιάσετε όταν χρειαστεί.

1. Να δοθεί Πίνακας περιγραφικής στατιστικής της Time με το μέγεθος δείγματος ανά στάθμη (N), MIN, MAX, MEDIAN, MEAN και SD.
2. Να γίνει το θηκόγραμμα (boxplot) της Time με τις τρεις στάθμες του Material όπου τα παράτυπα σημεία (outliers) θα έχουν ως ετικέτες τον αριθμό γραμμής του αρχείου data.frame.
3. Να γίνει έλεγχος προσαρμογής στην κανονική κατανομή (Shapiro-Wilk) και το αντίστοιχο QQplot για κάθε στάθμη του Material και να παρουσιαστεί το αποτέλεσμα.
4. Να γίνει έλεγχος ομοιογένειας των διασπορών για τις τρεις στάθμες του Material με τους ελέγχους του Levene και του Bartlet και να γραφεί το αποτέλεσμα.
5. Σύμφωνα με τα αποτελέσματα των Ερ. 4 και 5, να αποφασίσετε το επόμενο βήμα: 5.1) Χρήση των διαδικασιών Box-Cox ή Spread and Level Plot για εύρεση του κατάλληλου μετασχηματισμού και πιθανή εξαίρεση των παράτυπων σημείων ή όχι αν οι προϋποθέσεις της κανονικής κατανομή και της ισότητας των διασπορών ικανοποιηθούν ή 5.2) Εξαίρεση των παράτυπων σημείων με χρήση των κατάλληλων ελέγχων και στη συνέχεια αν χρειαστεί εφαρμογή μετασχηματισμού ως αποτέλεσμα της διαδικασίας Box-Cox ή Spread and Level Plot.
6. Να γίνει ανάλυση διασποράς με έναν παράγοντα με χρήση της συνάρτησης aov και να παρουσιάσετε τον πίνακα ανάλυσης της διασποράς με καταγραφή και ερμηνεία του αποτελέσματος.

7. Να χρησιμοποιήσετε τη βιβλιοθήκη emmeans και να παρουσιάσετε τις κατά

ζεύγη συγκρίσεις μεταξύ των τριών σταθμών του Material ώστε να φανεί ποια ή ποιες στάθμες δίνουν το μικρότερο χρόνο.

8. Να γίνει έλεγχος προσαρμογής των υπολοίπων στην κανονική κατανομή καθώς και οπτικός έλεγχος (γραφική/ές παραστάσεις) για την προϋπόθεση της σταθερής διασποράς των υπολοίπων.

***Πρόβλημα 2.***

Το αρχείο «cars1920.txt» περιέχει 50 παρατηρήσεις αυτοκινήτων από τη δεκαετία του 1920 για τα οποία έχουν καταγραφεί οι μεταβλητές speed: ταχύτητα και distance: απόσταση που χρειάστηκε το αυτοκίνητο για να σταματήσει τρέχοντας με την καταγεγραμμένη ταχύτητα. Στόχος της μελέτης είναι να εκτιμηθεί με μια απλή (γραμμική) σχέση η επίδραση της ταχύτητας ενός αυτοκινήτου στην απόσταση που χρειάζεται το αυτοκίνητο για να σταματήσει. Ακολουθείστε τα επόμενα για να εξάγετε το τελικό συμπέρασμα για την ανάλυση των δεδομένων αυτών (χρήση της συνάρτησης lm).

1. Κάντε το διάγραμμα διασποράς και σχολιάστε σχετικά με τη γραμμική σχέση μεταξύ των δυο μεταβλητών.

2. Κάντε τη γραμμική παλινδρόμηση της απόστασης στην ταχύτητα.

3. Κάντε έλεγχο υπόθεσης για την προσαρμογή των υπολοίπων στην κανονική κατανομή καθώς και το QQplot. Να ελέγξετε με γραφικό τρόπο την καταλληλόλητα του μοντέλου

4. Κάντε τον έλεγχο για την αυτοσυσχέτιση πρώτης τάξης των υπολοίπων και γράψτε τα συμπεράσματά σας.

5. Βρείτε τις παρατηρήσεις που ενδεχομένως επηρεάζουν την εξίσωση της γραμμικής παλινδρόμησης.

6. Να ελέγξετε την ανεξάρτητη μεταβλητή για επαναλαμβανόμενες τιμές, να κάνετε έναν πίνακα συχνοτήτων για τις τιμές που επαναλαμβάνονται.

7. Να αντιμετωπίσετε το πρόβλημα των επαναλήψεων των τιμών της ανεξάρτητης μεταβλητής μελετώντας τη σημαντικότητα του «καθαρού σφάλματος».

8. Συγκρίνετε τα δυο μοντέλα και αποφασίστε ποιο είναι καλύτερο.

9. Να κάνετε έλεγχο των προϋποθέσεων για το νέο μοντέλο και να γράψετε την εξίσωσή του.

***Πρόβλημα 3.***

Στο αρχείο «ColorStudyCR.sav», περιέχονται οι μεταβλητές CR: παράμετρος του χρώματος (ποσοτική συνεχής) που καταγράφεται για κάθε δοκίμιο στο εργαστήριο, ο παράγοντας Material με τέσσερις στάθμες : Brx, Prt, Ktn και emx και τέλος ο παράγοντας Solution με τέσσερις στάθμες: Tea, Coffee, Wine και Aging. Στόχος της μελέτης είναι να εκτιμηθεί η σημαντικότητα της αλληλεπίδρασης και των κυρίων επιδράσεων των δυο παραγόντων και να βρεθεί το βέλτιστο μοντέλο που επηρεάζει την παράμετρο CR. Τα αποτελέσματα πρέπει να περιέχουν: 1) μεθοδολογία για τη στατιστική ανάλυση (ποιο μοντέλο χρησιμοποιήθηκε και πως ελέγχθηκαν οι προϋποθέσεις για την εγκυρότητά του, ποιες επιδράσεις βρέθηκαν στατιστικά σημαντικές και πως δόθηκε η εκτίμηση των σημαντικών επιδράσεων πχ με το 95% διάστημα εμπιστοσύνης, ποιο στατιστικό πρόγραμμα χρησιμοποιήθηκε και ποιες βιβλιοθήκες – αφού πρόκειται για την R και ποια επιλέχτηκε να είναι η στάθμη στατιστικής σημαντικότητας), 2) πίνακα περιγραφικής στατιστικής και κατάλληλη γραφική παράσταση, 3) πίνακα ανάλυσης διασποράς για τα μοντέλα που δοκιμάστηκαν αλλά δε βρέθηκαν σημαντικά και αποτελέσματα για τις συγκρίσεις που οδήγησαν στο τελικό μοντέλο, 4) πλήρη στοιχεία για το τελικό μοντέλο και κατά ζεύγη συγκρίσεις που οδήγησαν στο τελικό συμπέρασμα καθώς και τον έλεγχο για τις προϋποθέσεις που πρέπει να ικανοποιεί το τελικό μοντέλο.

***Πρόβλημα 4.***

Οι κάτω χώρες αντιμετωπίζουν πρόβλημα διάβρωσης του εδάφους και για το λόγο αυτό με τη βοήθεια δορυφόρου καταγράφεται για κάθε κτίριο η απόσταση επιλεγμένων σημείων από το έδαφος. Το παρόν αρχείο περιέχει για ένα δείγμα κτιρίων ηλικίας άνω των 150 χρόνων το ρυθμό βύθισης/έτος (<0) ή υπερύψωσης/έτος (>0), (αυτό συμβαίνει λόγω σφαλμάτων είτε στη μέτρηση είτε στην επεξεργασία) για 8 συνεχόμενα έτη. Για την ακρίβεια στο αρχείο «Def5.txt», περιέχονται οι μεταβλητές id: κωδικός κτιρίου, AGE: ηλικία κτιρίου, year.n: έτος καταγραφής (1 – 8) και DV: ρυθμός βύθισης ή ανύψωσης σε χιλιοστά ανά έτος. Ο στόχος είναι να εξαχθεί κάποιο συμπέρασμα σχετικά με τον ρυθμό βύθισης των κτιρίων της περιοχής. Για την ακρίβεια: 1) Να

μελετήσετε και να κάνετε στατιστική επεξεργασία των δεδομένων ώστε να εξάγετε με

αιτιολόγηση το συμπέρασμά σας (στο πρόβλημα αυτό θα απαντήσετε ελεύθερα χωρίς καθοδήγηση), και 2) Να καταγράψετε κάποιον/ους παράγοντες – μεταβλητές που εκτιμάτε ότι θα ήταν χρήσιμοι στο σχεδιασμό