Πρώτη Σειρά Ασκήσεων – Μέρος Β

Σε αυτή την άσκηση θα κάνετε διερευνητική ανάλυση (exploratory analysis) δημογραφικών δεδομένων (census data) από τις Ηνωμένες Πολιτείες. Ο στόχος είναι να κάνετε κάποιες μετρήσεις πάνω στα δεδομένα, να βρείτε ενδιαφέρουσες συσχετίσεις και να ερευνήσετε κάποιες υποθέσεις. Επίσης, να εξασκηθείτε με την χρήση των Pandas για ανάλυση δεδομένων.

Η άσκηση αποτελείται από τα παρακάτω κομμάτια/βήματα. Ο στόχος είναι να υλοποιήσετε τα παρακάτω χρησιμοποιώντας κατά κύριο λόγο μεθόδους της βιβλιοθήκης Pandas (συν δικές σας συναρτήσεις που θα εφαρμόσετε με apply), και τις συναρτήσεις που έχουμε δει για plotting και στατιστική ανάλυση.

Βήμα 0. Το αρχείο με το οποίο θα δουλέψετε είναι το Census_Data.csv. Στο αρχείο έχουμε τα εξής δεδομένα για κατοίκους των ΗΠΑ:

- age: Ακέραιος αριθμός με την ηλικία
- CoW (Category of Work): Ο τύπος εργασίας (δημόσιος υπάλληλος, ιδιωτικός υπάλληλος, κλπ)
- education: Το επίπεδο εκπαίδευσης
- marital: Η οικογενειακή κατάσταση
- occupation: Το επάγγελμα
- PoB (Place of Birth): Ο τόπος γέννησης
- hours: Οι ώρες εργασίας ανά εβδομάδα
- sex: Το φύλο
- race: Η φυλή
- state: Η πολιτεία κατοικίας
- income: Το ετήσιο εισόδημα.

Οι τιμές που έχουν τα attributes αυτή τη στιγμή είναι κωδικοποιημένες αριθμητικά. Π.χ., η πολιτεία της Alabama είναι το νούμερο 1, το φύλο Male είναι η τιμή 1, κλπ. Πριν ξεκινήσετε την επεξεργασία θα φέρετε τα δεδομένα σε μια μορφή που είναι πιο εύκολο να τα διαβάσετε και να τα παρουσιάσετε. Σας δίνεται το αρχείο Attribute_Values.csv το οποίο περιέχει τριάδες της μορφής (Attribute Name, Attribute Numeric Value, Attribute Value), όπου για κάθε Attribute έχει τις διαφορετικές (αριθμητικές) τιμές που παίρνει, και την αντίστοιχη String τιμή. Χρησιμοποιώντας το αρχείο, τροποποιήστε τις τιμές στο dataframe που φορτώσατε τα αρχικά δεδομένα για τις παρακάτω μεταβλητές ως εξής:

- CoW: Θα πρέπει να καταλήξετε με τους εξής τύπους εργαζομένων: Private Employee, Government Employee, Self-Employed, No pay, Unemployed.

- education: Κρατήστε την αριθμητική τιμή, αλλά μπορεί να χρειαστείτε και μια νέα στήλη με τις κατηγορίες εκπαίδευσης για τα Βήματα 6 και 7 Θα ομαδοποιήσετε τις κατηγορίες 1-15 ως no diploma, 16-17, high-school, 18-19 post-high-school, και τα υπόλοιπα τα κρατάτε ως έχουν.
- marital: Αντικαταστήστε με την οικογενειακή κατάσταση
- occupation: Κρατήστε την αριθμητική τιμή, αλλά μπορεί να χρειαστεί να την αντικαταστήσετε για το Βήμα 7. Μπορείτε να πάρετε γκρουπ από επαγγέλματα που αντιστοιχούν σε διαφορετικούς τομείς εργασίας (π.χ., οικονομικός τομέας, εκπαίδευση, κλπ)
- PoB: Μετατρέψετε τις τιμές σε US (αριθμητικές τιμές μεταξύ 1-99, εκτός της τιμής 60), Not US (τιμές μεγαλύτερες του 99 και η τιμή 60).
- sex: Αντικαταστήστε με το φύλο.
- race: Αντικαταστήστε με την φυλή
- state: Αντικαταστήστε με τα ονόματα των πολιτειών.

Χρησιμοποιήσετε λειτουργίες της βιβλιοθήκης Pandas για την μετατροπή των τιμών, οι οποίες είναι πολύ πιο γρήγορες. Θα σας βολέψει να φορτώσετε το αρχείο σε dataframe και να κάνετε κάποιες μετατροπές με λειτουργίες μεταξύ dataframes.

Σας δίνεται το αποτέλεσμα του Βήματος 0 στο αρχείο Census_Data_cleaned.csv

Βήμα 1. Στο κομμάτι αυτό μας ενδιαφέρει να καταλάβουμε την κατανομή που ακολουθεί το εισόδημα (income). Κάνετε τις εξής γραφικές παραστάσεις (plots):

1. Ένα ιστόγραμμα του income με 100 κάδους (bins) χρησιμοποιώντας έτοιμη συνάρτηση της βιβλιοθήκης Pandas
2. Δημιουργείστε μία νέα στήλη log_income με το **λογάριθμο** του income και κάνετε ένα ιστόγραμμα με 100 κάδους χρησιμοποιώντας πάλι την μέθοδο της βιβλιοθήκης Pandas
3. Δημιουργείστε εσείς ένα ιστόγραμμα με κάδους που το μέγεθος τους αυξάνεται εκθετικά (διπλασιάζονται). Ο πρώτος κάδος θα έχει μέγεθος 5000 (τα εισοδήματα μεταξύ [0,5000]), ο δεύτερος 10000 (τα εισοδήματα μεταξύ [5000,15000]), ο τρίτος 20000 (τα εισοδήματα μεταξύ [15000,35000]), κλπ. Κάνετε plot τον αριθμό των εγγραφών σε κάθε κάδο, ως προς το δεξί άκρο του κάδου σε λογαριθμική κλίμακα και στους δύο άξονες.
4. Υπολογίστε το cumulative frequency vector. Το διάνυσμα για κάθε εισόδημα (από το μικρότερο προς το μεγαλύτερο) κρατάει τον αριθμό των εγγραφών που έχουν **τουλάχιστον** τόσο εισόδημα (π.χ., για το μικρότερο εισόδημα είναι όλες οι εγγραφές). Κάνετε μια γραφική παράσταση του cumulative frequency ως συνάρτηση του εισοδήματος σε λογαριθμική κλίμακα και στους δύο άξονες.
5. Κάνετε το Zipf plot για το εισόδημα. Το Zipf plot κατασκευάζεται έχοντας στον Y άξονα τις τιμές (το εισόδημα στην περίπτωση μας) και στο X την τάξη (rank) των τιμών. Για παράδειγμα το μέγιστο εισόδημα έχει rank 1, το δεύτερο μεγαλύτερο 2, κλπ. Το plot θα είναι σε λογαριθμική κλίμακα και τους δύο άξονες.

Παρουσιάστε τα γραφήματα σας σε ένα grid και **σχολιάστε την κατανομή**.

Σημείωση: Για τα βήματα 3-5 μπορείτε αν θέλετε να τα υλοποιήσετε μεταφέροντας τα δεδομένα σε λίστες.

Βήμα 2. Στο κομμάτι αυτό θα εξετάσετε αν υπάρχει κάποια συσχέτιση μεταξύ του φύλου (sex), της φυλής (race), και της καταγωγής (PoB) και της κατηγορίας της εργασίας (CoW). Δημιουργείστε τα contingency tables μεταξύ αυτών των τριών ζευγαριών μεταβλητών (μπορείτε να τα αναπαραστήσετε με ένα heatmap) και κάνετε το χ^2-test για να διαπιστώσετε αν υπάρχει κάποια στατιστικά σημαντική συσχέτιση. Η ύπαρξη συσχέτισης είναι ενδεικτική αδικίας εφόσον αυτά τα χαρακτηριστικά δεν θα έπρεπε να καθορίσουν ποιος μπορεί να κάνει ποια δουλειά. Αξιολογήστε τα αποτελέσματα κοιτώντας τα p-value. Κοιτάξτε και επιμέρους ζευγάρια από τιμές με μεγάλο lift (λόγος της τιμής του ζευγαριού προς την αναμενόμενη τιμή – σας την δίνει το χ^2-τεστ). Σχολιάστε τα αποτελέσματα.

Βήμα 3. Στο κομμάτι αυτό θα εξετάσετε αν τα παραπάνω attributes (sex, race, PoB) έχουν κάποια συσχέτιση με το εισόδημα (income). Ο στόχος μας είναι να εξετάσουμε αν υπάρχει στατιστικά σημαντική διαφορά στις μέσες τιμές του εισοδήματος μεταξύ των φύλων, φυλών, ή καταγωγής.

Για το στόχο αυτό, για κάθε attribute δημιουργείστε ένα point plot (linestyle='none') που θα έχετε την μέση τιμή για κάθε brand και το 95% confidence interval. Προσπαθήστε να εξάγετε κάποιο συμπέρασμα αρχικά από το plot. Στη συνέχεια για κάθε ζευγάρι από τιμές χρησιμοποιείστε το t-test για να εξετάσετε αν υπάρχει στατιστικά σημαντική διαφορά στις μέσες τιμές των εισοδημάτων. Τα γραφήματα σας θα πρέπει να είναι σε ένα grid 1X3.

Σημείωση: Για την φυλή χρησιμοποιείστε το setting plt.xticks(rotation=90) ώστε να βγουν κάθετα τα labels στον X άξονα.

Περιγράψετε τα αποτελέσματα και τα συμπεράσματα σας.

Βήμα 4. Στο κομμάτι αυτό θα εξετάσετε αν υπάρχει συσχέτιση μεταξύ του εισοδήματος και κάποιων attributes που παίρνουν συνεχείς τιμές. Κάνετε scatter plots του education, του age και του hours με το income, και το log_income για την πολιτεία της California. Παρουσιάστε τα γραφήματα σας σε ένα grid 2X3.

Υπολογίστε για κάθε ζευγάρι το Pearson Correlation Coefficient και το αντίστοιχο p-value. Σχολιάστε τα αποτελέσματα και τις διαφορές όταν χρησιμοποιείτε το εισόδημα και το λογάριθμο του εισοδήματος.

Για το education για να δείτε καλύτερα τη σχέση του με το εισόδημα, κάνετε ένα γράφημα του μέσου εισοδήματος (για την California) ως προς το education, με 95%-confidence interval (χρησιμοποιήστε την lineplot της seaborn). Σχολιάστε το αποτέλεσμα.

Βήμα 5. Σας δίνεται το αρχείο voting-2020.csv που έχει τα αποτελέσματα των εκλογών στις ΗΠΑ το 2020. Τα αποτελέσματα είναι ανά πολιτεία και έχουν το ποσοστό που πήρε κάθε ο κάθε υποψήφιος. Χρησιμοποιώντας γραφήματα και στατιστικά τεστ εξετάστε την υπόθεση ότι οι πολιτείες με χαμηλό μέσο εισόδημα ψήφισαν Trump, και οι πολιτείες με υψηλό μέσο επίπεδο εκπαίδευσης ψήφισαν Biden. Για τον υπολογισμό μέσων τιμών ανά πολιτεία καθώς και την συγχώνευση των δύο datasets θα πρέπει να χρησιμοποιήσετε εντολές της βιβλιοθήκης Pandas.

Bonus: Απεικονίστε πάνω στον χάρτη τα αποτελέσματα των εκλογών, το μέσο εισόδημα, και το μέσο επίπεδο εκπαίδευσης, και συγκρίνετε οπτικά τους χάρτες.

Βήμα 6. Διατυπώστε μια δική σας υπόθεση και εξετάστε την χρησιμοποιώντας τα δεδομένα. Η υπόθεση σας θα πρέπει να είναι κάτι μη τετριμμένο και να την εξετάσετε χρησιμοποιώντας (και) κάποιο στατιστικό τεστ.

Παραδώσετε ένα Notebook το οποίο θα περιέχει τον κώδικα για την επεξεργασία των δεδομένων, τις γραφικές παραστάσεις και τους υπολογισμούς που κάνατε, καθώς και τις παρατηρήσεις και τα συμπεράσματα σας. Βάλετε headers ώστε να ξεχωρίζουν τα διαφορετικά κομμάτια της άσκησης.