Ερώτηση 1

Στην ερώτηση αυτή θα κάνετε διερευνητική ανάλυση (exploratory analysis) δεδομένων από το Yelp. Το Yelp είναι μια πλατφόρμα για κριτικές σε επιχειρήσεις (κυρίως εστιατόρια) που είχε μεγάλη επιτυχία πριν κάποια χρόνια στις ΗΠΑ και εξακολουθεί να είναι δημοφιλής. 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.

Θα χρησιμοποιήσετε το academic Yelp dataset που είναι μια συλλογή από δεδομένα που προσφέρει το Yelp για ερευνητικούς και εκπαιδευτικούς σκοπούς. Μπορείτε να το κατεβάσετε από εδώ. Είναι ένα μεγάλο dataset οπότε θα χρειαστείτε κάποιο χρόνο και χώρο για να το κατεβάσετε. Επίσης είναι κάτι που θα πρέπει να το έχετε υπόψη σας στην επεξεργασία (π.χ., δεν είναι εύκολο να φορτώσετε τα αρχεία στη μνήμη). Τα δεδομένα είναι σε json μορφή. Θα χρησιμοποιήσετε την βιβλιοθήκη json της Python για να τα επεξεργαστείτε.

Προεπεξεργασία δεδομένων: Από τα αρχεία που έχει στο συμπιεσμένο αρχείο που θα κατεβάσετε θα χρησιμοποιήσετε μόνο τα αρχεία yelp_academic_dataset_business.json (κυρίως αυτό) και yelp_academic_dataset_reviews.json. Από το πρώτο αρχείο θα κρατήσετε μόνο τις επιχειρήσεις που είναι στην πόλη της Φιλαδέλφειας (Philadelphia). Κρατήστε μόνο τις εγγραφές που δεν περιέχουν null τιμές. Από τα reviews κρατήστε μόνο αυτά που είναι για τις επιχειρήσεις στην Φιλαδέλφεια που επιλέξατε στο προηγούμενο βήμα. Σώστε αυτά τα δεδομένα σε csv αρχεία, ώστε να μην κάνετε την επεξεργασία πολλαπλές φορές. Για τα επόμενα βήματα θα ξεκινήσετε φορτώσετε τα δεδομένα από τα αρχεία που δημιουργήσατε.

Η άσκηση αποτελείται από τα παρακάτω κομμάτια. Ο στόχος είναι να υλοποιήσετε τα παρακάτω φορτώνοντας τα δεδομένα σε Pandas dataframes και χρησιμοποιώντας κατά κύριο λόγο μεθόδους της βιβλιοθήκης Pandas (συν δικές σας συναρτήσεις που θα εφαρμόσετε με apply).

Α. Στο κομμάτι αυτό μας ενδιαφέρει να καταλάβουμε την κατανομή που ακολουθεί ο αριθμός των reviews (review_count) που έχουν οι επιχειρήσεις (θα το πάρετε από το αρχείο business). Θα κάνετε τα εξής γραφήματα (plots):

1. Ένα ιστόγραμμα του review_count με 100 κάδους (bins) χρησιμοποιώντας έτοιμη συνάρτηση της βιβλιοθήκης Pandas
2. Ένα ιστόγραμμα του **λογαρίθμου** του review_count με 100 κάδους χρησιμοποιώντας πάλι μεθόδους της βιβλιοθήκης Pandas
3. Μια γραφική παράσταση που θα δείχνει την συνάρτηση πυκνότητας πιθανότητας του review_count που θα κατασκευάσετε εσείς. Θα σπάσετε το διάστημα των τιμών του review_count σε 100 ισομεγέθη διαστήματα. Στον X άξονα θα έχετε το κάτω άκρο του διαστήματος, και στον Y τον αριθμό των επιχειρήσεων με review_count σε αυτό το διάστημα. Θα κάνετε plot το Y ως προς το X παίρνοντας λογαριθμική κλίμακα και στους δύο άξονες.
4. Το Zipf plot της κατανομής των review_counts. Το Zipf plot κατασκευάζεται έχοντας στον Y άξονα τις τιμές (review_count στην περίπτωση μας) και στο X την τάξη (rank) των τιμών. Για παράδειγμα το μέγιστο review_count 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 2×2 και σχολιάστε την κατανομή

Σημείωση: Δεν υπάρχει σαφές συμπέρασμα για την κατανομή που ακολουθούν οι πόντοι αλλά μπορείτε να κάνετε κάποιες παρατηρήσεις για το σχήμα της κατανομής. Μπορείτε επίσης να προσθέσετε κάποιο δικό σας plot αν πιστεύετε ότι θα σας βοηθήσει. Τα βήματα 3,4 είναι πιο δύσκολο να υλοποιηθούν χρησιμοποιώντας Pandas (ειδικά το Βήμα 3), μπορείτε αν θέλετε να τα υλοποιήσετε μεταφέροντας τα δεδομένα σε λίστες.

Β. Στο κομμάτι αυτό θα εξετάσουμε πως εξελίσσονται τα ratings των επιχειρήσεων στον χρόνο. Γι αυτή την ερώτηση θα χρησιμοποιήσετε το αρχείο με τα reviews απ' όπου θα πάρετε την ημερομηνία των reviews. Χρησιμοποιώντας την ημερομηνία του πρώτου review για κάθε επιχείρηση θα υπολογίσετε το μήνα στον οποίο έγινε το review. (Υποθέστε ότι ο μήνας έχει 30 μέρες. Το πρώτο review είναι την μέρα 0. Όσα reviews είναι λιγότερο από 30 μέρες από το πρώτο review είναι στον πρώτο μήνα, όσα είναι 30-59 μέρες είναι στον δεύτερο, κ.ο.κ.). Κάνετε μια γραφική παράσταση της μέσης τιμής του star rating σαν συνάρτηση του μήνα στον οποίο γράφηκε το review. Στη συνέχεια κάνετε το ίδιο γράφημα περιορίζοντας τα reviews σε αυτά που γράφτηκαν το πολύ 5 χρόνια μετά το πρώτο review. Τι παρατηρείτε στις γραφικές παραστάσεις? Δώστε μια πιθανή εξήγηση.

Στη γραφική παράσταση θα πρέπει να φαίνεται και το 95% confidence interval. Συνίσταται να χρησιμοποιήσετε την μέθοδο lineplot της Seaborn. Για την επεξεργασία των ημερομηνιών χρησιμοποιείστε μεθόδους της βιβλιοθήκης datetime. Για την εξαγωγή της ημερομηνίας του πρώτου review και τον υπολογισμό του μήνα του review χρησιμοποιείστε μεθόδους της βιβλιοθήκης Pandas.

Γ. Στο κομμάτι αυτό μας θα ερευνήσουμε αν υπάρχει συσχέτιση μεταξύ του πόσο ακριβή είναι μια περιοχή της πόλης και των επιχειρήσεων που βρίσκονται σε αυτή την περιοχή. Η υπόθεση είναι ότι οι ακριβές περιοχές είναι αυτές στις οποίες υπάρχουν πολλές επιχειρήσεις ή καλές επιχειρήσεις.

Για τον ορισμό της περιοχής θα χρησιμοποιήσουμε το πεδίο postal_code. Σας δίνεται επίσης και το tab-separated αρχείο RedfinPhila.tsv το οποίο περιέχει στατιστικά για real-estate τιμές για ακίνητα σε διαφορετικές περιοχές (zip/postal codes). Από αυτό το αρχείο θα κρατήσετε μόνο τις στήλες region (που περιέχει το zip code το οποίο είναι το ίδιο με το postal code) και median_sale_price που έχει την διάμεση τιμή πώλησης για διαφορετικού τύπου κατοικίες σε αυτή την περιοχή για διαφορετικές χρονικές περιόδους. Από αυτά τα δεδομένα θα υπολογίσετε για κάθε zip code την μέση τιμή του median_sale_price. Αυτή θεωρούμε ότι είναι η εκτίμηση της ακρίβειας/αξίας της περιοχής.

Χρησιμοποιώντας αυτά τα δεδομένα θα δημιουργήσετε ένα dataframe το οποίο θα έχει για κάθε zip code, την αξία της περιοχής με αυτό το zip code (μέσο median_sale_price), τον αριθμό των επιχειρήσεων σε αυτό το zip_code, και τη μέση τιμή των star ratings των επιχειρήσεων σε αυτό το zip code. Θα εξετάστε την συσχέτιση μεταξύ των τριών πεδίων. Δημιουργήστε ένα γράφημα με όλα τα scatter plots των πεδίων ανά δύο (χρησιμοποιήστε την μέθοδο pairplot της seaborn), και δύο 3×3 πίνακες με τα Pearson correlation coefficients και τα αντίστοιχα p-values. (Εναλλακτικά μπορείτε να παρουσιάσετε τα αποτελέσματα χρησιμοποιώντας heatmaps με τις τιμές). Σχολιάστε τα αποτελέσματα. Παρατηρείτε κάποια ενδιαφέρουσα συσχέτιση? Μια συσχέτιση είναι ενδιαφέρουσα αν έχει μεγάλο coefficient και p-value μικρότερο του 0.05.

Στη συνέχεια κάνετε τις ίδιες μετρήσεις για περιοχές με αξία μεγαλύτερη των 200,000 (αφαιρούμε τις φτηνές περιοχές). Τι παρατηρείτε? Σχολιάστε τα αποτελέσματα.

Ποιο είναι το συμπέρασμα για την αρχική μας υπόθεση?
Σημείωση: Η δημιουργία των dataframes θα πρέπει να γίνει με εντολές της βιβλιοθήκης Pandas. Μπορεί να σας φανεί χρήσιμη η εντολή value_counts.

Δ. Στο κομμάτι αυτό μας ενδιαφέρει να μελετήσουμε αν υπάρχει διαφορά μεταξύ των star ratings διαφορετικών επιχειρήσεων με βάση την τιμή τους και την ενδυμασία που απαιτούν. Για το κομμάτι αυτό θα επικεντρωθείτε μόνο σε επιχειρήσεις που είναι εστιατόρια. Τα εστιατόρια είναι οι επιχειρήσεις που έχουν την λέξη Restaurants στο πεδίο categories. Δημιουργείστε ένα dataframe που έχει μόνο τα εστιατόρια από τις επιχειρήσεις που έχετε τώρα.

Από το πεδίο attributes, το οποίο είναι ένα dictionary, θα πάρετε το attribute 'RestaurantPriceRange2', το οποίο είναι μια τιμή μεταξύ 1 και 4 που μας λέει πόσο ακριβό είναι το εστιατόριο. Θα θεωρήσουμε φτηνά τα εστιατόρια με price range 1 ή 2, και ακριβά αυτά με price range 3, ή 4. Δημιουργήστε ένα pointplot που να δείχνει το μέσο star rating για φτηνά και ακριβά εστιατόρια και το 95% confidence interval. Τι παρατηρείτε με το μάτι? Στη συνέχεια χρησιμοποιείστε το t-test για να μετρήσετε αν η διαφορά που βλέπετε είναι στατιστικά σημαντική, και σχολιάστε το αποτέλεσμα.

Θα κάνετε την ίδια διαδικασία για το attribute 'RestaurantsAttire'. Στην περίπτωση αυτή θα έχουμε πάλι δύο τύπους εστιατορίων, αυτά που απαιτούν επίσημη ενδυμασία (το RestaurantsAttire περιέχει τα strings dressy ή formal) και αυτά στα οποία η ενδυμασία είναι ανεπίσημη (το RestaurantsAttire περιέχει το string casual). Δημιουργήστε πάλι ένα pointplot που να δείχνει το μέσο star rating για επίσημα και ανεπίσημα εστιατόρια και το 95% confidence interval. Τι παρατηρείτε με το μάτι? Στη συνέχεια χρησιμοποιείστε το t-test για να μετρήσετε αν η διαφορά που βλέπετε είναι στατιστικά σημαντική, και σχολιάστε το αποτέλεσμα.

Σημείωση: Η εντολή eval θα σας είναι χρήσιμη για να μετατρέψετε ένα string σε dictionary (μπορείτε να το κάνετε και json object). Αν για κάποια εστιατόρια δεν μπορείτε να πάρετε τις τιμές που θέλετε μπορείτε να τα αγνοήσετε. Η επεξεργασία των attributes θέλει λιγο προσοχή γιατί τα δεδομένα δεν είναι πάντα καθαρά. Διαβάστε για την poinplot εδώ (χρησιμοποιήστε την επιλογή join=False). Μπορείτε να κάνετε και ένα οριζόντιο plot.

Ε. Στο κομμάτι αυτό θα χρησιμοποιήσετε πάλι τα δεδομένα για τα εστιατόρια και θα εξετάσετε αν υπάρχει συσχέτιση μεταξύ διαφορετικών χαρακτηριστικών των εστιατορίων. Από τη στήλη Attribute θα κοιτάξουμε το πεδίο 'Ambience' (ατμόσφαιρα). Το πεδίο αυτό είναι ένα λεξικό με κλειδιά κάποιες προεπιλεγμένες κατηγορίες και τιμές True/False. Ένα εστιατόριο μπορεί να έχει πολλά από αυτά τα χαρακτηριστικά.

Θα εξετάσουμε αν τα μέρη τα οποία είναι καλά για ραντεβού είναι επίσης και «κυριλέ». Ένα εστιατόριο είναι καλό για ραντεβού αν έχει ατμόσφαιρα 'romantic' ή 'intimate'. Ένα εστιατόριο είναι «κυριλέ» αν έχει ατμόσφαιρα 'classy' ή 'upscale'. Δημιουργείστε το contingency table γι αυτά τα δύο χαρακτηριστικά (καταλληλόλητα για ραντεβού και το αν είναι «κυριλέ» το εστιατόριο). Χρησιμοποιήστε το χ^2-test για να εξετάσετε αν τα χαρακτηριστικά είναι ανεξάρτητα. Σχολιάστε τα αποτελέσματα σας. Στο σχολιασμό σας θα πρέπει επίσης να συγκρίνεται το δικό σας contingency table με αυτό που προκύπτει υποθέτοντας ότι τα δύο χαρακτηριστικά είναι ανεξάρτητα.

Σημείωση: Προσοχή γιατί το attribute Ambience δεν εμφανίζεται σε όλα τα εστιατόρια, και ακόμη και όταν εμφανίζεται σε κάποιες περιπτώσεις μπορεί να μην υπάρχει κάποιο από τα χαρακτηριστικά που μας ενδιαφέρουν, ή να έχει την τιμή None. Μπορείτε να θεωρήσετε τις None τιμές ως False, και να απορρίψετε όλες τις εγγραφές που δεν έχουν τις τιμές που χρειάζεστε.

(Bonus). 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.