ΑΝΑΦΟΡΑ ΤΕΛΙΚΟΥ PROJECT

1. Δεν βρήκα κάποια επιλογή για download στο github (άραγε υπάρχει;), οπότε πήρα το dataset με copy-paste.
2. Η πρώτη μου προσπάθεια για φιλτράρισμα ήταν λάθος, καθώς δοκίμασα το:

USE liquorSales;

SELECT\*

FROM finance\_liquor\_sales

WHERE date BETWEEN 2016 AND 2019;

Το οποίο δεν έδωσε αποτέλεσμα, καθώς ήθελε να προσδιορίσω τη χρονιά, από το date (YYYY-MM-DD). Όποτε μετά δοκίμασα και έσωσα ως csv το αποτέλεσμα αυτού του query:

USE liquorSales;

SELECT\*

FROM finance\_liquor\_sales

WHERE YEAR (date) >=2016

AND YEAR (date) <= 2019;

1. Η δυσκολία, φυσικά και λόγω απειρίας, ήταν ο ορισμός της σωστής σειράς των data και η επιλογή του σωστού function. Ενώ είναι ξεκάθαρο πχ. για το πρώτο ζητούμενο ότι πρέπει να δουλέψουμε με zip\_code και bottles\_sold, ανέτρεξα πάλι πίσω στις σημειώσεις για να δω ποιο πρέπει να μπει πρώτο, και μετά αναρωτήθηκα αν το sum() θα είναι η σωστή επιλογή, αλλά στα περισσότερα παραδείγματα και σημειώσεις με αυτό δουλεύουμε. Δεν είχα αρκετή γνώση των υπόλοιπων functions και τι μπορεί να κάνουν. Νομίζω μόνο το count() συνάντησα σε μια άσκηση. Μια άλλη απορία ήταν σε ποια περίπτωση/γραμμή κώδικα πρέπει (αν πρέπει) να μπαίνει το reset\_index. Το αποτέλεσμα βγήκε μετά από πολλές αλλαγές και πειράματα.
2. Το δεύτερο ζητούμενο με δυσκόλεψε περισσότερο απ’ όσο περίμενα. Αρχικά με μπέρδεψε και η εκφώνηση, θα μπορούσε να βελτιωθεί η διατύπωση (νόμιζα ήταν ένα ενιαίο query με το πρώτο). Όπως και με το πρώτο ζητούμενο, η δυσκολία μου δεν ήταν στη λογική/μαθηματικά, αλλά πως θα διατυπωθεί σωστά σε κώδικα. Σε προτεινόμενες εναλλακτικές (stackoverflow, chatgpt) είδα χρήση του idmax(0) και συνάρτησης lambda, αλλά καθώς δεν έχει νόημα το copy-paste, αλλά μια λύση σύμφωνα με τις προσωπικές μου γνώσεις έως τώρα, δεν είμαι σίγουρη αν χρησιμοποίησα τον βέλτιστο κώδικα. Περίμενα ότι με το:  
   **PercentageOfSales** = MostPopularOfStore/TotalStoreSales \* 100   
   το print θα μου έδινε το ζητούμενο, αλλά έβλεπα ότι σε ένα κατάστημα έδινε λάθος ποσοστό και κάποια καταστήματα εμφανιζόντουσαν 2 φορές. Έπρεπε να μπει και το:  
   **PercentageOfSalesPerStore** = PercentageOfSales.groupby('store\_number').head(1)  
   για να δώσει σωστό output.   
   Ακόμα έχω απορία γιατί έπρεπε να ξανακάνω groupby, όταν και το MostPopularOfStore είχε γίνει sorted και το TotalStoreSales είχε γίνει groupby('store\_number'), πριν τα χρησιμοποιήσω για το PercentageOfSales.
3. Το scatter δεν με δυσκόλεψε καθόλου.
4. Θέματα με τη δημιουργία του bar:  
   Πριν ξεκινήσω με το bar, οι 2 τελευταίες γραμμές του κώδικα που είχα για το 2ο ζητούμενο ήταν:  
     
   PercentageOfSales = MostPopularOfStore/TotalStoreSales \* 100

PercentageOfSalesPerStore = PercentageOfSales.groupby('store\_number').head(1). reset\_index()  
  
‘Όταν δοκίμασα PercentageOfSalesPerStore.plot(kind='bar') μου έδωσε εντελώς λάθος αποτέλεσμα. Δεν έδινε καν ποσοστά στον y άξονα, ενώ ο x άξονας έπαιρνε το index value.  
Μετά δοκίμασα PercentageOfSales.plot(kind='bar') και ενώ οι άξονες ήταν σωστοί (ποσοστό και κατάστημα), είχε φυσικά τα λάθος δεδομένα που συνάντησα και όταν έτρεχα το print(PercentageOfSales).  
Μόνο όταν αφαίρεσα το reset\_index() πήρα σωστό γράφημα. Δεν βρήκα όμως τρόπο στον x άξονα να δίνει το store number, αντί για το store name (που διαβάζεται και πολύ δύσκολα).