Project
Preprocessing and Visualization

Το dataset με το οποίο θα ασχοληθείτε είναι το ακόλουθο:
US Accidents Dataset (https://www.kaggle.com/sobhanmoosavi/us-accidents).
Αποτελείται από περίπου 7,7 εκατομμύρια εγγραφές, και περιέχει πληροφορίες σχετικά με αυτοκινητιστικά ατυχήματα στις Η.Π.Α κατά το χρονικό διάστημα 02/2016–12/2023 (ανανεώνεται σε ετήσια βάση). Για τους σκοπούς της εργασίας θα επιλέξετε (με τυχαία δειγματοληψία) **100.000 εγγραφές**.
ΒΗΜΑ 1: Διαλέξτε **100.000** εγγραφές από το σύνολο των εγγραφών. Με αυτές θα κάνετε training & evaluation αργότερα. Φορτώστε το dataset με την βιβλιοθήκη pandas κάντε **visualize** του dataset αρχικά, να φαίνεται η αρχή (pandas :: **head()**) και το τέλος (pandas :: **tail()**).
ΒΗΜΑ 2: Καθαρίστε τα δεδομένα από ελλιπείς ή εσφαλμένες τιμές (**NaN**). Εφόσον αρχικά εντοπίσετε αυτές τις **NaN** τιμές βγάλτε τες από το dataset με έναν απ όλους τους διαθέσιμους τρόπους, μέσω της βιβλιοθήκης pandas:

1. **dropna()**: removes rows or columns with NaN or -inf values
2. **replace()**: replaces NaN and -inf values with a specified value
3. **interpolate()**: fills NaN values with interpolated values

ΒΗΜΑ 3: Μόλις αφαιρεθούν οι άκυρες τιμές κάντε οπτικοποίηση (pandas :: **describe()**) των **min, max, mean, std**, για κάθε ένα από τα χαρακτηριστικά (columns) του dataset.
ΒΗΜΑ 4: Βρείτε τη συσχέτιση των features μεταξύ τους κάνοντας χρήση του **correlation matrix** (pandas :: **corr()**). Κάντε plot το correlation matrix, μέσω της βιβλιοθήκης seaborn (**heatmap**).
ΒΗΜΑ 5: Οπτικοποιήστε τα δεδομένα σας μέσω **ιστογράμματος, boxplot, pairplot.**

ΜΠΟΡΕΙΤΕ ΝΑ ΒΡΕΙΤΕ ΤΑ ΠΑΝΤΑ ONLINE
ΧΡΗΣΙΜΟΠΟΙΗΣΤΕ ΟΠΟΙΑ ΒΙΒΛΙΟΘΗΚΗ ΣΑΣ ΒΟΛΕΥΕΙ ΚΑΙ ΚΑΤΑΝΟΗΣΤΕ ΤΙ ΚΑΝΕΤΕ ΚΥΡΙΩΣ