Lutwaffe

Descrizione

Il dataset riguarda luoghi e perdite dell'aviazione tedesca durante la seconda guerra mondiale.

Dataset

Il dataset è composto da due csv differenti: uno contenente le posizioni delle flotte aeree, ed uno contenente le variazioni numeriche delle flotte aeree. Entrambi i file contengono le seguenti colonne (che possono essere usate per unirli):

- **group_type**: tipologia del compito assegnato
- **squadron_number**: numero dello squadrone
- **group_number**: numero del gruppo
- **subgroup**: numero del sotto-gruppo come numero romano
- **date**: data (sempre a metà del mese)

Il file delle posizioni contiene anche le seguenti colonne:

- **lat**: latitudine
- **lon**: longitudine
- **locations**: descrizione della posizione

il file delle variazioni contiene anche le seguenti colonne:

- **total**: numero totale degli apparecchi all'inizio del mese
- **model**: modello e variante dell'apparecchio
- **add**: numero totale di apparecchi aggiunti nel mese
- **add_new**: numero di nuovi apparecchi aggiunti nel mese
- **add_maintenance**: numero di apparecchi aggiunti nel mese provenienti dalla manutenzione
- **add_other_units**: numero di apparecchi aggiunti nel mese provenienti da altre unità
- **lost**: numero totale di apparecchi persi nel mese
- **lost_enemy**: numero totale di apparecchi persi nel mese in combattimento
- **lost_accident**: numero totale di apparecchi persi nel mese per incidente
- **lost_maintenance**: numero totale di apparecchi persi nel mese perchè in manutenzione
- **lost_other_units**:numero totale di apparecchi persi nel mese perchè ceduti ad altre unità
- **total_eom**: numero totale di apparecchi disponibili alla fine del mese

Quesiti

Programmare le istruzioni pyspark per rispondere ai seguenti quesiti:

- Qual è la media delle flotte trimestre per trimestre del numero totale degli apparecchi?
- Qual è la media delle flotte della percentuale di apparecchi persi in combattimento?
- Qual è la varianza delle variazioni a livello mensile degli apparecchi, sul totale delle flotte? (non è permesso usare le funzioni pre-fabbricate per calcolare la varianza)
- Per ogni luogo censito (colonna locations) qual'è la distribuzione percentuale della tipologia di apparecchi (colonna group_type) che si sono trovati in quel luogo lungo tutto il corso del tempo?
- Qual è stato il mese più "letale" (quindi con più apparecchi persi in combattimento)?