1. THE DATASET (file Preprocessing.R)

Presentazione del dataset originale proposto da Azzone, facciamo una selezione focalizzandoci sulle 20 birre più vendute in termini di volume, da febbraio 2019 a febbraio 2024. La nostra analisi si baserà su questo dataset, per capire come si comportano i diversi brand e prodotti ma anche come influiscono i tipi di promozione. Qui spiegherei anche gli outlier sulle peroni e peroni nastro azzurro, che però non abbiamo voluto modificare imputandoci altri valori né abbiamo voluto togliere, per non falsare troppo la nostra indagine.

2. ANOVA (file ANOVA.R)

2 diversi test: un MANOVA sulle vendite in volume e lo sconto rispetto al prodotto, per vedere se variano per prodotto, e un un MANOVA sulle vendite in volume e lo sconto rispetto al brand, per vedere se variano per brand. Un ultimo ANOVA è stato fatto per valutare l'impatto delle diverse tipologie di promozione sulle vendite in volume, da notare che i diversi tipi di sconto sono in tutto 14 perché a volte i singoli tipi di sconto sono applicati in coppia o di più. Spieghiamo i risultati e specifichiamo che le assunzioni (normality e equivalence of variances) molto spesso non sono rispettate.

3. CLUSTER (file Cluster.R)

Goal del processo di cluster è capire quali sono i leader e quali sono i follower di mercato a livello di prodotto, per le vendite in volume e in prezzo. Per capirlo è stato fatto K-means su: dati aggregati su ogni anno (somma per prodotto), dati aggregati sul totale degli anni (sempre somma), sulle time-series. Vediamo che si creano due cluster, uno con le vendite più alte e uno con le vendite più basse. Overall (quindi sulla time series), risultano come leader 4 prodotti, tuttavia quando andiamo ad analizzare la clusterizzazione per anno e su tutti gli anni dei dati aggregati, quelli che rimangono sempre e vincono sugli altri sono 3: Moretti 66 cl, Ichnusa non filtrata 50 cl e Heineken 66 cl. Useremo questo risultato come covariata per la regressione.

4. REGRESSIONE E LMM (file Regression.R)

Questo è il goal finale del nostro progetto, serve sia per capire come influiscono i diversi tipi di sconto (alcuni sono significativi e altri no, dai confidence interval dei beta vediamo anche quelli che hanno valori maggiori di altri) ma anche per capire quali sono in generale i fattori che influenzano le vendite in volume. A questo scopo, abbiamo introdotto nuove covariate, ispirandoci ai Marketing Mix Modeling: prezzo scontato, prezzo non scontato, dummy che ci dice quando è estate, dummy che ci dice quando è vacanza, dummy leader/follower che vene dal cluster. La variabile risposta è stata trasformata in logaritmo a causa del suo andamento (mettere istogramma, magari sovrapposto per far vedere la differenza prima e dopo). Attraverso un processo di selezione delle variabili tramite i p-value dei coefficienti, il modello finale contiene le seguenti covariate: …., abbiamo anche aggiunto una variabile chiamata ‘Volumi.bassi’ per cercare di tenere in considerazione e attenuare l’effetto degli outlier che dicevamo nella prima parte. Presentazione delle assunzioni, da mettere nelle weakness la non normalità dei dati (p-value dello shapiro test bassissimo, dal qqplot vediamo infatti delle code molto pesanti). Il modello è stato validato tenendo il 30% delle osservazioni (che corrisponde all’ultimo periodo temporale) come test set, mettiamo il grafico delle predizioni (con prediction interval, che è quello che presenteremmo a un nostro eventuale stakeholder) della Moretti da 66 cl (che è la top birra venduta). Mettere plot per i confidence intervals dei beta.
A questo punto iniziamo il discorso sull’implementazione dei LMM, infatti vedendo il boxplot dei residui rispetto a brand e prodotto vediamo che le varianze sono molto diverse. Vale la pena provare a implementare due LMM diversi, che verranno poi confrontati, grazie anche ai risultati che abbiamo spiegato prima dei test ANOVA sulle vendite in volume e sugli sconti. Infatti il primo step è quello di mettere solo l’intercetta dipendere dal prodotto / brand, poi abbiamo anche incluso una random slope legata alle covariate sui prezzi in sconto e non in sconto. Questi due risultano essere i modelli migliori, risultato convalidato da un test anova. Come ultimo step, abbiamo confrontato i due lmm finali (uno per prodotto e uno per brand), tramite un test anova,

che ci conferma che il modello lmm rispetto a prodotto è il migliore. Mettere tabella di mae e mse per i 3 modelli (lm, lmm prodotto, lmm brand) e le predizioni per la moretti, magari anche di AIC dei 3 modelli per vedere quale modelli è effettivamente il migliore.

Nel file Analisi_esplorativa.R ci sono alcuni plot dei nostri dati, capire quali mettere ed eventualmente farne altri rispetto a quello che vogliamo far vedere.