**DISCORSO BEA DELLA CHIAMATA DELL'ALTRO GIORNO**

00:00 —> 06:43 **Presentazione**
Cosa contiene il dataset? La time-series settimanale delle vendite delle top 20 birre (specifichiamo che sono top in termini di volumi venduti e non di valore in sé del prodotto), da febbraio 2019 a febbraio 2024. Abbiamo le vendite sia in volume che in valore e tutti i vari tipi di sconto che sono stati applicati. Noi usiamo i vari tipi di sconto come variabili binarie, li abbiamo ricevuti come numero di vendite per ogni tipologia ma l'abbiamo convertito in un indicatore se lo sconto è stato applicato o meno. Ogni birra è indicata dal proprio brand (Moretti, Ichnusa, ecc…) e dal gruppo "economico" a cui appartiene. I valori mancanti che abbiamo trovato sono stati sostituiti con un'interpolazione lineare tra quello successivo e quello precedente. Per Peroni e Peroni Nastro Azzurro sono stati identificati degli outlier che sono stati tenuti per non falsare troppo l'indagine e mantenere l'integrità del dataset originale. In particolare possiamo guardare i due plot in alto a destra nel primo scaffale dove c'è una overview generale del dataset. Le vendite sono in volume rispetto a sconto o non sconto dove ogni riga (colore) corrisponde a un prodotto e questi sono suddivisi per vendor. Nel caso Peroni possiamo notare propriamente questo comportamento strano all'inizio del 2024 dove le vendite sono veramente basse rispetto a quelle del prodotto.

06.43 —> 11:30 **Research Question**
La prima research question è relativa a come le varie promozioni influiscono sui volumi venduti. Per rispondere a questa domanda usiamo l'ANOVA (on Sales volume by promotion types) che ci serve a dire che le vendite rispetto ai vari tipi di promozione sono diverse. L'altro metodo è il valore dei coefficienti della regressione per questi tipi di sconto. Se guardiamo nell'ultimo scaffale c'è il dotplot sui 95% CI di Beta dove, in particolare, compaiono i tre tipi di discount che ci interessano maggiormente (display, pricereduction, flyer). Tutti e tre assumono volumi positivi e quello con i valori maggiori è PriceReduction che ci indica come avere questa tipologia di sconto ci aiuta maggiormente ad avere un aumento di vendite in volume.
La seconda domanda di ricerca è legata a se i diversi prodotti e brand hanno dei comportamenti diversi. Abbiamo il MANOVA test, il clustering tra prodotti Leader e Follower all'interno del mercato e infine la regressione (in particolare il modello mixed model, guardare l'ultima immagine del 95% CI rispetto al prodotto tra intercetta e prezzi scontati/non).

11:30 —> 15:23 **ANOVA e MANOVA**
Sopra abbiamo i boxplot delle vendite in volume (scala logaritmica) e della percentuale di sconto (colonna che ci siamo trovati noi avendo le vendite in volume e in valore con/senza promozione e totale). In particolare il MANOVA viene fatto per prodotto e brand. Viene fuori che la cosa migliore è quella di rifiutare l'ipotesi nulla e quindi c'è differenza sia nelle vendite in volume sia nelle percentuali di sconto in entrambi i raggruppamenti brand e prodotto. Le nostre risposte sono affette da entrambe le categorizzazioni. Questo risponde alla seconda domanda di ricerca ma ci aiuta quando facciamo il linear mixed model perchè se diciamo che la percentuale di sconto varia sia per brand che per prodotto allora ci è consentito di utilizzare nella random slope del modello le covariate relative al prezzo scontato e non scontato. Grazie a queste due covariate random otterremo infatti un modello migliore rispetto a quello con solo l'intercetta o solo una delle due covariate random. Per l'ANOVA rispetto al tipo di promozione bisogna fare attenzione che non è detto che una promozione stia andando a escludere l'altra e quindi ci potrebbero essere le varie combinazioni che corrispondono a 14. Con il boxplot delle vendite in volume in scala logaritmica rispetto alle varie promozioni affermiamo che rifiutiamo l'ipotesi nulla quindi effettivamente c'è una differenza molto significativa nella scelta di sconto per le vendite in volume. Anche se usiamo la scala logaritmica sulla nostra variabile di risposta, non in tutti i gruppi abbiamo l'ipotesi di normalità e anche l'uguaglianza della covarianza (o varianza per ANOVA) è violata in alcuni casi. Li riportiamo come debolezze del progetto infatti.

15:23 —> 19:07 **Cluster**
Ci serve per rispondere alla seconda domanda perchè vediamo come si comportano leader e follower in termine di prodotto e lo useremo come covariata nella regressione. Usiamo un algoritmo k-means con k=2 per dividere nelle due categorie. Facciamo tre tipologie diverse di cluster. Per la prima facciamo le somme annuali di volume e value sales per ogni prodotto in modo da analizzare ogni prodotto rispetto a ogni annata visualizzata. Tre birre (Ichnusa, Moretti ed Heineken, l'unica a scendere relativamente come valore negli anni) restano sempre Leader mentre la Peroni da 66 cl già dal secondo anno diventa una follower (evidenziata nel grafico). La distanza

utilizzata dovrebbe essere quella euclidea. Questo comportamento della Peroni è indubbiamente legato alla presenza degli outlier di cui parlavamo prima. Il secondo tipo di cluster è il “partitional clusterting”, un k-means applicato su tutta la time-series, da cui risulta che la Peroni in realtà fa parte dei Leader ma ha un calo drastico nell’ultimo periodo. L’ultimo tipo di cluster tiene in conto la somma su tutti gli anni e ci dice che la Moretti, la Heineken e l’Ichnusa sono i prodotti dominanti. Useremo questa classificazione che sembra la più coerente e più robusta rispetto al discorso degli outlier.

19:07 —> 37:48 **Regression Model**
Facciamo un modello lineare per vedere quali sono i fattori che influenzano le nostre vendite in volume e quali sono i tipi di promozione che hanno un beta più alto e quindi le più utili da ricercare. Inseriamo delle nuove variabili seguendo il marketing mixing modeling tra cui il prezzo scontato e non scontato, una variabile binaria che ci dice quando è estate, quando è vacanza e quando un prodotto è leader/follower seguendo il cluster precedente. Nel plot sulla destra c’è la distribuzione dei sales volume e utilizziamo una trasformazione logaritmica per migliorarne l’andamento (procedimento fatto anche su prezzo scontato e non).
Tramite il p-value dei coefficienti notiamo come alcuni possano essere rimossi e quello riportato sul cartellone è il modello più adatto con prezzo scontato e non, discount flyer, display, pricereduction e le variabili relative a isSummer, isLeader e LowVolumes che è stata aggiunta dopo per cercare di mitigare l’effetto degli outlier. Infatti, vediamo dal plot dei residui (in particolare il primo per l’omoschedasticità) notiamo che abbiamo dei punti sulla sinistra che non ci piacciono che sono quei volumi molto bassi visti in precedenza che si comportano da outlier. Purtroppo non abbiamo normalità nei residui per delle code molto pesanti e un valore di p-value molto basso. R2 del 65% abbastanza bene.
Se plottiamo i residui rispetto a prodotto e brand vediamo come le varianze sono molto diverse e questo ci porta a pensare che possiamo fare due modelli con effetti misti diversi, uno per i brand e uno per i prodotti. I migliori modelli sono quelli con random slope anche le variabili relative al prezzo scontato e non scontato oltre che l’intercetta, ovviamente. Nella tabellina c’è il PVRE del linear mixed model sia per prodotto e brand e quindi gran parte della variabilità di risposta della nostra risposta è spiegata da questi random.
Il modello LMM by Product risulta essere il migliore sia dall’ANOVA test tra i vari modelli sia dalla Pinball Loss che raggiunge una media più bassa rispetto agli altri.
Dato che il linear mixed model per prodotto risulta essere il migliore, decidiamo di fare una predizione sul test set (ultimo 20% del periodo) rispetto alla Moretti da 66 cl che è il nostro top prodotto. Vediamo che la predizione è buona perchè segue abbastanza l’andamento dei dati di train ma non riesce a prendere bene i picchi (non stessa intensità). Il MAE e MSE su questo test set sono inseriti come metriche che ci danno un buon andamento del modello.
Quali sono i vari fattori che influenzano le nostre vendite?
Siccome il coefficiente su IsLeader è positivo su tutti e 3 i modelli e questo ci dice che i prodotti Leader del mercato vendono più dei follower e questo ci riproduce bene le dinamiche di mercato per il modello. Le 3 covariate di sconto F, D, PR sono quelle che hanno anche da questo plot una buona influenza e sono le tre più significative di tutti gli sconti. Il coefficiente di IsSummer è positivo e quindi questo evidenza come in estate ci sia una vendita maggiore del prodotto e un conseguente aumento positivo di vendite. LowVolumes sul CI ci indica che se i volumi sono bassi chiaramente va in contrasto rispetto alla risposta, ci dice che se il volume è basso allora la risposta andrà in contrario. Notiamo che il prezzo scontato ha un coefficiente negativo e quindi più è basso il prezzo più è alto il volume di vendita. Dall’altra parte però ci chiediamo “perchè il coefficiente sul prezzo non scontato è positivo? Allora più è alto il prezzo più è alto il volume di vendita?”. Ci siamo dati la spiegazione che questo tipo di prodotti non avrà mai dei prezzi altissimi perchè comunque sono di supermercato e quindi anche se non scontati vengono venduti comunque bene.
Ultimo plot presenta i random effect per il modello LMM by product nel CI 95% tra intercetta e prezzo scontato e non. Si può far vedere come la Corona ha una stima dell’intercetta molto alta mentre quella della Dreher è molto basso. Questo si può interpretare che la Corona è meno influenzata dagli sconti e, per quanto non sia Leader del mercato, viene venduta a prescindere dalle promozioni e dai fattori mentre la Dreher è quella più influenzata da ciò. Se vediamo il random effect del discounted price della Dreher si nota che se effettivamente ha un prezzo in sconto più basso allora viene più venduta.

37:48 —> 39:06 **Conclusions**

Le conclusioni danno una risposta molto secca sulle nostre domande di ricerca. Le promozioni hanno un impatto forte sui sales volumes come evidenziato dall'ANOVA test, in particolare le strategie F, PR, D sono le principali. I Leader e Followers hanno dei comportamenti diversi e in particolare nei modelli lineari notiamo che le promozioni guidano le crescite delle vendite in particolare quando sono legate a trend di stagioni o in particolare il brand in sé (quindi il prodotto).