

Luxury Retail RECOMMENDATION SYSTEM



Indice

1. Contesto

- 1.1. Azienda
- 1.2. Problema di Business
- 1.3. Dati a disposizione

2. Obiettivi dell'analisi

3. Tecnologie e Architettura di BI adottate

4. Fasi dell'analisi dei dati

- 4.1. Data Ingestion
- 4.2. Data Cleaning
- 4.3. Statistiche Descrittive
- 4.4. Algoritmo
- 4.5. Risultati

5. Visualizzazioni a supporto delle evidenze

6. Possibili sviluppi futuri

Contesto aziendale

Azienda del settore **Retail Fashion Luxury**

- **Fatturato medio** annuo: circa 1,3 miliardi € ; utili intorno ai 14,0 milioni €;
- **300 negozi** monomarca: Boutique (Retail e Travel Retail) ed Outlet, oltre a sito E-commerce
 - di gestione e proprietà diretta
 - distribuiti su tutti i continenti ad eccezione di alcune aree, come East Europa e Medio Oriente, dove sono presenti negozi Franchising (punti vendita monomarca con licenza di vendita);
- Collezioni Uomo, Donna e Bambino; 3 principali linee di produzione:
 - **Ready To Wear**: abbigliamento costituito da abiti realizzati e venduti finiti in taglie standard;
 - **Sartoriale o Made To Measure**: abiti maschili realizzati su misura del cliente con personalizzazioni;
 - **Alta Moda**: abito realizzato a mano, interamente artigianale e dai materiali pregiati;
- Tramite accordi licenziatari, l'azienda commercializza una linea **occhiali, profumi e makeup**;
- Negli ultimi tempi è stata creata una **linea food** ed una collezione **home**.



Problema di Business

Tra i diversi punti di un'ampia strategia di **revenue generation**, impostata dalla direzione Retail e da quella Commerciale, vi è quello di sviluppare e migliorare l'**up-sell** e il **cross-sell** di vendita

UP-SELLING

tecnica di vendita che mira ad offrire al consumatore qualcosa di maggior valore rispetto alla sua scelta d'acquisto iniziale, per **farlo spendere di più**.



CROSS-SELLING

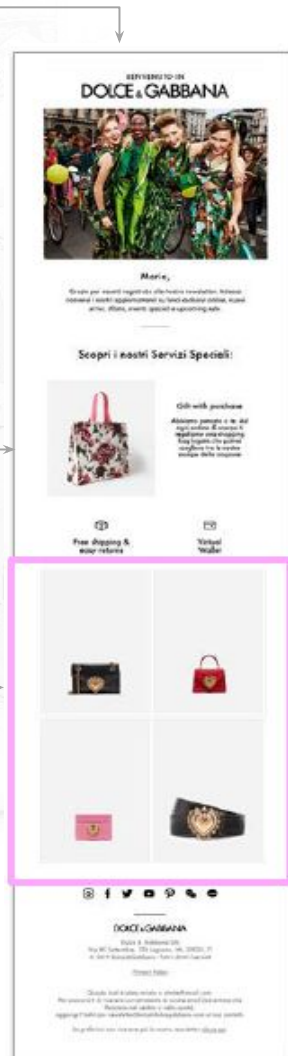
consiste nell'aumentare il valore dello scambio mettendo a disposizione prodotti in qualche modo collegati con la scelta d'acquisto iniziale, per rendere l'**offerta più ricca e completa**.



Esempio di campagna marketing di cross/up sell dopo un acquisto, per completare il look

Acquisto effettuato

Suggerimento di prodotti post acquisto



Problema di Business

Perché?

Analizzando i basket d'acquisto è emerso che nella maggior parte dei ticket è presente **una sola** categoria di prodotto o classe commerciale

Come?

Sviluppando di un sistema di **Product Recommendation** che restituisca i più probabili pattern d'acquisto

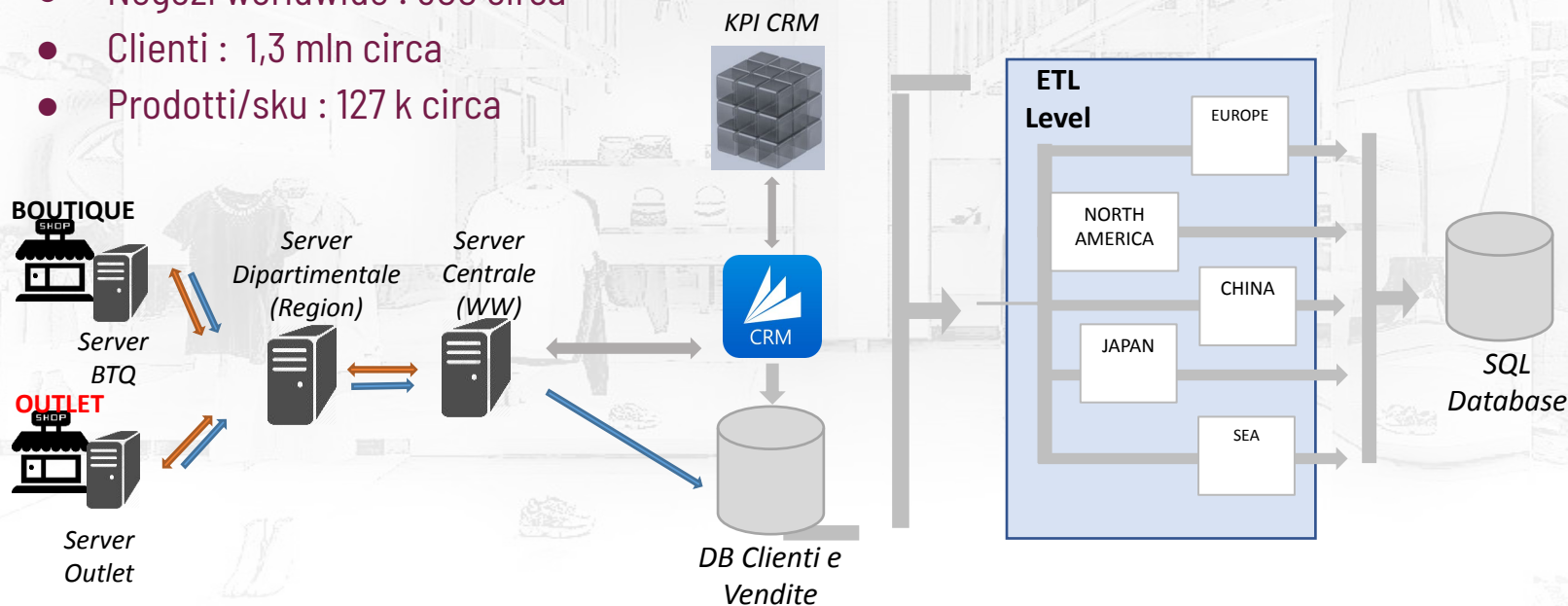
Per chi?

- **Ufficio Merchandising**: miglioramento dello sviluppo della collezione;
- **Ufficio Visual Merchandising**: miglioramento dell'esposizione del prodotto in negozio;
- **Ufficio Buyer**: ottimizzazione dei processi di acquisto dei negozi;
- **Ufficio CRM**: fornire contenuti di comunicazione rilevanti sugli acquisti passati per guidare cross sell o riattivazioni clienti;
- **Forza vendita**: supporto agli store manager e sales associati nella predisposizione dei look da proporre alla clientela.

Dati a disposizione

Dati consuntivi delle vendite estratti dal CRM aziendale (> 5 mln righe)

- Finestra temporale considerata: **2015-2021**
- Negozi worldwide : 300 circa
- Clienti : 1,3 mln circa
- Prodotti/sku : 127 k circa



Obiettivi dell'analisi

- supportare la forza vendita durante la cerimonia di vendita
- migliorare l'efficacia delle campagne di direct marketing
- supportare le strategie di product merchandising
- migliorare il processo di Demand Planning

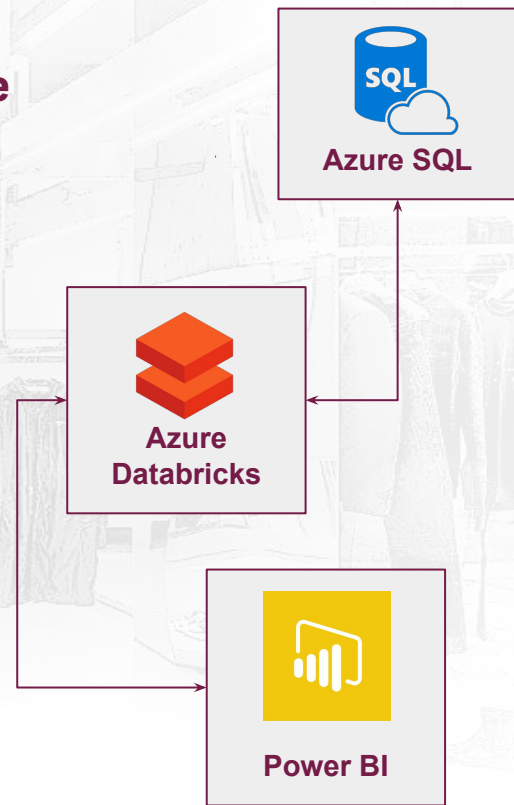
Tecnologie e Architetture di BI utilizzate

La scelta “Tecnologica” è ricaduta nella piattaforma cloud **Microsoft Azure**

Azure offre un’ampia selezione di servizi che vanno dall’elaborazione e archiviazione dei dati, fino all’analisi degli stessi attraverso strumenti di Business Intelligence e Business Analytics, tutti estremamente interconnessi.

Questa soluzione “all-in-one” che ci ha permesso di:

- creare e **condividere** la nostra base dati in Cloud
- sfruttare la **scalabilità** di calcolo utilizzando i notebook Databricks
- analizzare i dati ed i risultati ottenuti in Power BI



Fasi di analisi e rispettive tecnologie

DATA INGESTION



*DB relazionale in Cloud
(Azure)*

condivisione dei dati

DATA PROFILING



*Studio dei dati e
identificazione delle
problematiche
presenti*



DATA CLEANING



*Normalizzazione,
rimozione e/o
riclassificazione di
alcune variabili*

creazione tabella
OLAP



*Unione tabella dei fatti con
le tabelle delle dimensioni*

**STATISTICHE
DESCRITTIVE**



*Comprensione e
rappresentazione delle
relazioni tra variabili*

**ESECUZIONE
ALGORITMO**



*Sottomissione di diverse
configurazioni dello stesso
algoritmo*

valutazione dei
risultati



*Valutazione congiunta della
qualità delle metriche e
contestualizzazione*

DATA VISUALIZATION



*Dashboard a supporto delle
raccomandazioni fornite*

1. Contesto

2. Obiettivi

3. Tecnologie

4. Fasi dell'analisi

Data Ingestion



1) Estrazione CSV dal gestionale CRM (SAP)

- Negozi: 357
- Clienti: 1,30 mln
- Prodotti: 127,4 mila
- Righe scontrini: 5,86 mln

riconducibile a tipico **schema a stella** con la tabella dei fatti e le tabelle delle dimensioni

ANAGRAFICA NEGOZI

Store Code
Store Desc.
Region Code
Region Desc.
Sub Region Code
Sub Region Desc.
Country Code
Country Desc.
City

ANAGRAFICA CLIENTI

CRM Code
Nationality Group (CRM)
Residence Country Desc.
Birth Date
Customer Gender Code

TRANSAZIONI

CRM Code
Date
Ticket Prefix+Number
Model Part Color Code
Sold Qty
Sold SellOut Net (ExR SO)

ANAGRAFICA PRODOTTI

Brand Desc.
Brand Code
Retail Line Group
Line Desc.
Line Code
Commercial Class Desc. (It)
Commercial Class Code
Model Desc.
Model Code
Color Desc.
Color Code
Model Part Color Code
Article Desc.

Data Ingestion

2) Storage CSV nel Database in Cloud (Azure SQL Server)



Home page > Database SQL >


Database SQL

Università degli Studi di Milano-Bicocca (unimib.it...)

+ Crea Prenotazioni ...

Filtra per qualsiasi campo...

Nome ↑



Non ci sono elementi Database SQL da visualizzare

Se non è possibile visualizzare ciò che si sta cercando, provare a modificare i filtri.

[Altre informazioni](#)

[Crea Database SQL](#)

Crea database SQL

Microsoft

Selezionare la sottoscrizione per gestire le risorse distribuite e i costi. Usare i gruppi di risorse come le cartelle per organizzare e gestire tutte le risorse.

Sottoscrizione * Azure per studenti

Gruppo di risorse * Crea nuovo

Dettagli database

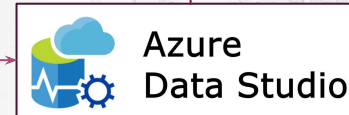
Immettere le impostazioni necessarie per questo database, ad esempio selezionare un server logico e configurare le risorse di calcolo e di archiviazione

Nome database * Project

Server * (nuovo) projectbida (Europa occidentale)

Usare il pool elastico SQL? ☐ Sì ☒ No

Calcolo e archiviazione * Utilizzo generico
Gen5, 2 vCore, 32 GB di archiviazione, ridondanza della zona disabilitata
[Configura database](#)



Step 2: Preview Data

This operation analyzed the input file structure to generate the preview below for up to 1000 rows.

Store_Code	Store_Desc	Region_Code	Region_Desc	Sub_Region_...	Sub_Region_...
010001	MILANO SPI...	EU	Europe	SE	Southern Europe
010002	MILANO VEN...	EU	Europe	SE	Southern Europe
010003	PORTO CERVO	EU	Europe	SE	Southern Europe
010004	FIRENZE	EU	Europe	SE	Southern Europe
010006	VERONA	EU	Europe	SE	Southern Europe
010008	ROMA COND...	EU	Europe	SE	Southern Europe
010009	FORTE DEI M...	EU	Europe	SE	Southern Europe
010010	PADOVA DG ...	EU	Europe	SE	Southern Europe
010011	VERONA	EU	Europe	SE	Southern Europe
010012	VERONA	EU	Europe	SE	Southern Europe

Import flat file wizard

Step 1: Specify Input File

Server the database is in *

projectbida.database.windows.net (projectbida)

Database the table is created in *

Project

Location of the file to be imported *

c:\Users\bragato\Desktop\DG\Anagrafica_Store.csv

[Browse](#)

New table name *

Anagrafica_Store

Table schema *

dbo

Data Ingestion

Condivisione del Database in Cloud, impostando gli accessi consentiti

Microsoft Azure

Home page > Project (projectbibda/Project) >

Impostazioni del firewall

projectbibda (SQL Server)

Salva Rimuovi + Aggiungi IP client

Criteri di connessione

Predefinito Proxy Reindirizzamento

Consenti alle risorse e ai servizi di Azure di accedere a questo server

Sì No

Indirizzo IP client 93.34.144.237

Nome regola	Indirizzo IP iniziale	Indirizzo IP finale
ClientIp-2021-8-3_9-17-39	93.34.144.237	93.34.144.237
Unimib1	54.190.17.122	54.190.17.122
Unimib2	18.237.152.94	18.237.152.94
Unimib3	34.222.171.45	34.222.171.45

Stringa di **connessione JDBC**, per collegare il Database in Cloud a Azure Databricks

Home page > Project (projectbibda/Project)

Project (projectbibda/Project) | Stringhe di connessione

Database SQL

Cerca (CTRL+ /)

ADO.NET **JDBC** ODBC PHP Go

JDBC (autenticazione SQL)

```
jdbc:sqlserver://projectbibda.database.windows.net:1433;database=Project;user=projectbibda@projectbibda;password=(your_password_here);encrypt=true;trustServerCertificate=false;hostNameInCertificate=*.database.windows.net;loginTimeout=30;
```

Scarica il driver JDBC per SQL server

Data Cleaning

Creazione area di lavoro
Azure Databricks

The screenshot shows the 'Clusters / Project' page in the Azure Databricks interface. The cluster is named 'Project' and is in a 'Stopped' state. The configuration includes:

- Cluster Mode: Single Node
- Databricks Runtime Version: 8.3 (includes Apache Spark 3.1.1, Scala 2.12)
- Autopilot Options: Terminate after 40 minutes of inactivity
- Node Type: Standard_DS3_v2 (14 GB Memory, 4 Cores)
- DBU / hour: 0.75

Buttons for 'Edit', 'Clone', and 'Restart' are visible at the top right of the cluster configuration area.

The screenshot shows the 'Project' overview page in the Azure Databricks interface. The page displays the project name 'Project' and its status 'Servizio Azure Databricks'. A search bar is present with the text 'Cerca (CTRL+/)'. Below the search bar, there are tabs for 'Panoramica', 'Log attività', 'Controllo di accesso (IAM)', and 'Tag'. The 'Informazioni di base' section shows the following details:

- Stato: Active
- Gruppo di risorse: Fabio
- Località: Europa occidentale
- Sottoscrizione: Azure per studenti
- ID sottoscrizione: c30dcd21-71dc-4698-af7b-8bfe23ceb3a3

Creazione **Cluster** "Project"

Data Ingestion

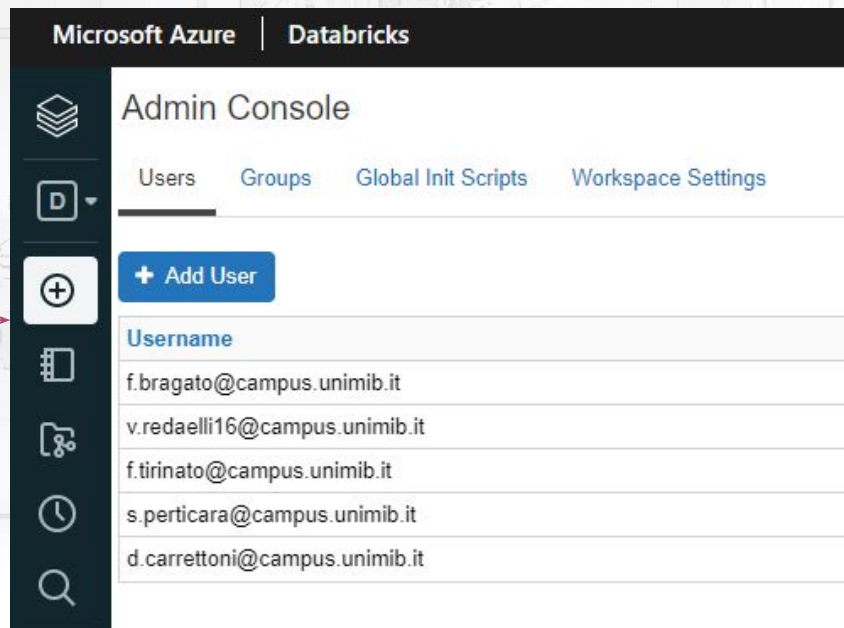
Data Cleaning

Istanza Databricks

Data Cleaning

Condivisione dell'ambiente

Azure Databricks (Cluster & Notebook)



The screenshot shows the Microsoft Azure Databricks Admin Console. The top navigation bar includes 'Microsoft Azure' and 'Databricks'. The main header is 'Admin Console'. Below it, there are tabs for 'Users', 'Groups', 'Global Init Scripts', and 'Workspace Settings'. The 'Users' tab is selected. A '+ Add User' button is visible. Below the button, there is a table of users with the following entries:

Username
f.bragato@campus.unimib.it
v.redaelli16@campus.unimib.it
f.tirinato@campus.unimib.it
s.perticara@campus.unimib.it
d.carrettoni@campus.unimib.it

Data Ingestion

Data Cleaning

Istanza Databricks

Data Cleaning

Creazione Notebook in **Azure Databricks** e connessione via JDBC da **Azure SQL Server**

Step 1: Load Data

Cmd 5

Connessione JDBC da Azure SQL Server

```
1 driver = "com.microsoft.sqlserver.jdbc.SQLServerDriver"
2 url = "jdbc:sqlserver://projectbibda.database.windows.net;DatabaseName=Project"
3 table = "dbo.Anagrafica_Store"
4 user = "projectbibda"
5 password = "1234567890!"
6
7 Client = spark.read.format("jdbc")\
8   .option("driver", driver)\
9   .option("url", url)\
10  .option("dbtable", table)\
11  .option("user", user)\
12  .option("password", password)\
13  .load()
```

Data Cleaning

Transaction rows (Sales)

- Eliminazione delle "quotes"
- Estrazione dell'informazione "Store", estrapolando l'attributo dalla colonna scontrino (Ticket)
- Eliminazione degli importi negativi (n scontrini=104.985)

	Customer	Date	Ticket	Sku	Qty	Sales
1	"DG820948"	"06/04/18"	"1 01000101A 201800002081"	"F68B2TGDH57HHI85"	"-1"	"-995"
2	"DG820948"	"06/04/18"	"1 01000101A 201800002081"	"F68H8THS1RQHAM64"	"-1"	"-3.450"
3	"DG820948"	"06/04/18"	"1 01000101A 201800002081"	"F68P5TFS57MHCM58"	"-1"	"-1.650"
4	"DG820948"	"06/04/18"	"1 01000101A 201800002081"	"F69O9TFSFGGHNM62"	"-1"	"-1.350"
5	"DG820948"	"06/04/18"	"1 01000101A 201800002081"	"F69P1TFSEGIHAP58"	"-1"	"-1.450"



```
sales_view = sales_view.withColumn('ticket', regexp_replace('ticket', ' ', ''))
sales_view = sales_view.withColumn('substr_store', substring('ticket',0,2))

#case when: store from ticket
sales_view = sales_view.withColumn('store', when(sales_view.substr_store == "10", substring('ticket',2,6))
                                     .when(sales_view.substr_store == "1V", substring('ticket',3,6))
                                     .when(sales_view.substr_store == "10", substring('ticket',3,6))
                                     .when(sales_view.substr_store == "1R", substring('ticket',3,6))
                                     .otherwise("000000"))
```


Data Cleaning

Product / SKU

- Eliminazione caratteri e spazi indesiderati
- Sostituzione descrizioni non consistenti
- Eliminazione righe del campo "**Commercial Class Desc.**" pari a "NO VALUE", "NON DEFINITO", "NON DEFINITA", "OBJECTS", "OTHERS"
- Eliminazione righe del campo "**Line Desc.**" pari a "FOOD&BEVERAGE", "PUBBLICAZIONI", "LICENZIATARI", "LICENZE VARIE"
- Eliminazione righe del campo "**Model Desc.**" pari a "CANDELA"

E' stata inoltre creata una tabella di raccordo utile alla riclassificazione della singola SKU per determinare un livello di aggregazione idoneo per l'algoritmo.

Data Cleaning

Customer

Creazione del campo **"Generation"**, compilato calcolando la generazione di appartenenza del cliente ottenuta a partire dalla data di nascita:

→ Builders	'30 - '45
→ Baby Boomers	'46 - '65
→ Gen X	'66 - '80
→ Gen Y	'81 - '96
→ Gen Z	'97 - '03

Altri record sono stati riclassificati come **"Undefined"** a seconda se la Birth Date fosse pari a 01/01/01, data fittizia utilizzata per i clienti non disposti a fornire la propria data di nascita, o riferiti ad anni recenti (2021, 2020, 2019 ecc..)

Data Cleaning

Store

Estrazione di altri dettagli di carattere qualitativo dal campo **"Store Description"**

→ **Channel** : Boutique / Outlet

→ **Target**: Woman / Man / Kids

→ **Stato** : Chiuso / Aperto

"Store Code"	"Store Desc."	"Region Code"	"Region Desc."	"Sub Re
"010044"	"FORTE DEI MARMI KIDS CLOSED "	"EU"	"Europe"	"SE"
"010045"	"ROMA RINASCENTE UOMO TEMP"	"EU"	"Europe"	"SE" "So
"010046"	"ROMA RINASCENTE DONNA TEMP"	"EU"	"Europe"	"SE"
"010047"	"FORTE DEI MARMI"	"EU"	"Europe"	"SE" "Southern E
"010048"	"ROMA SPAGNA"	"EU"	"Europe"	"SE" "Southern Europ
"010049"	"MILANO SPIGA DONNA "	"EU"	"Europe"	"SE" "Southe

Tabella OLAP

Microsoft Azure | Databricks

2_ETL_DG_Databricks_finale (Python)

Project

Cmd 19

Creazione Tabella OLAP

```
1 from pyspark.sql import SQLContext
2 sqlContext = SQLContext(sc)
3
4 SQL_Transaction = "SELECT store||customer||day||model as Ticket_Row, st
5 store = store_cod left join customer on customer = customer_crm inner j
6
7 sales_dg = sqlContext.sql(SQL_Transaction)
8 #sales_dg.show(5)
```

sales dg: pyspark.sql.dataframe.DataFrame = [Ticket_Row: string, Ticket: string ... 39 more fields]

Command took 0.07 seconds ← by f.bragato@campus.unimib.it at 9/9/2021, 15:45:30 on Project

Efficienza Databricks

Tabelle Anagrafiche

Clienti: 1,30 mln record
Prodotti : 127 mila record
Negozi: 357 record

Tabella dei fatti

Vendite: 5,8 mln record

RISULTATO LEFT JOIN :

5.82 mln record
39 campi
< 1 secondo

Data Ingestion

Data Cleaning

Tabella OLAP

Statistiche Descrittive



1) **Connessione** da Azure Databricks a PowerBI

Recupera dati



Tutto

Azure

Tutto



Azure Databricks

Azure Databricks

Server Hostname ⓘ

adb-1306140956724568.8.azuredatabricks.net

HTTP Path ⓘ

sql/protocolv1/o/1306140956724568/0811-134915-fur834

▸ Advanced Options (facoltativo)

Modalità Connettività dati ⓘ

☒ Importa

☐ DirectQuery

Data Ingestion

Data Cleaning

Tabella OLAP

Descrittive

Configurazioni

Statistiche Descrittive



2) Importazione dati in PowerBI

Strumento di navigazione

Opzioni di visualizzazione ▾

- adb-1306140956724568.8.azuredatabricks.net...
- SPARK [1]
- default [1]
- ☒ olap_sales_table

olap_sales_table

customer	model	qty	sales	day	store
DG770347	A10005A169280049	1	416,5	22/05/2016	010023
DG4700824	A10005A169280049	1	297	10/10/2017	090003
DG4390293	A10005A169280049	1	297	25/09/2017	090003
DG4756731	A10005AC62680999	1	247	04/10/2017	090018
DG4976815	A10005AC62680999	1	247	01/10/2017	090018
DG4704069	A10005AC62680999	1	247	14/10/2017	090018
DG4700824	A10005AC62680999	1	247	10/10/2017	090003

Carica

olap_sales_table
5.366.490 righe da {"host":"adb-1306140956724568.8.azuredatabricks.net","httpPath":134915-fur834"}.

Data Ingestion

Data Cleaning

Tabella OLAP

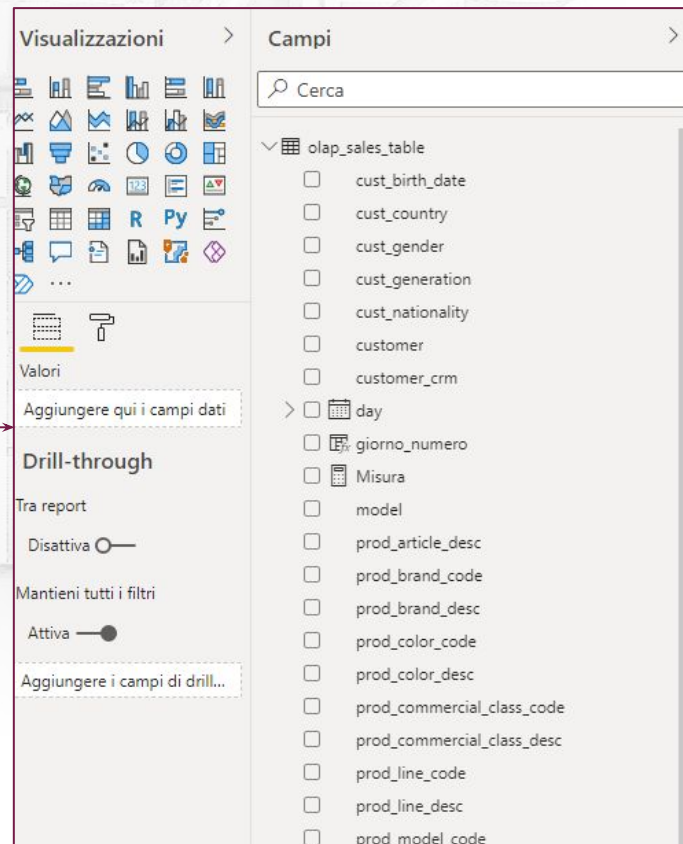
Descrittive

Configurazioni

Statistiche Descrittive



3) Dati disponibili per l'analisi in **Power BI**



Data Ingestion

Data Cleaning

Tabella OLAP

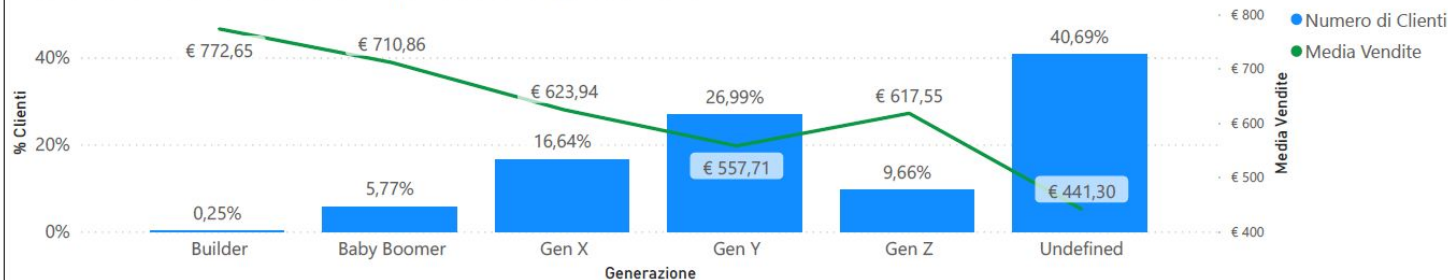
Descrittive

Configurazioni

Statistiche Descrittive

CLIENTI

Media vendite e numero clienti per classe di generazione



Totale clienti nel dataset:
1,30 mln

- di cui Maschi: 46,97 %
- di cui Femmine: 52,46%
- altro: 0,57%

Legenda generazioni

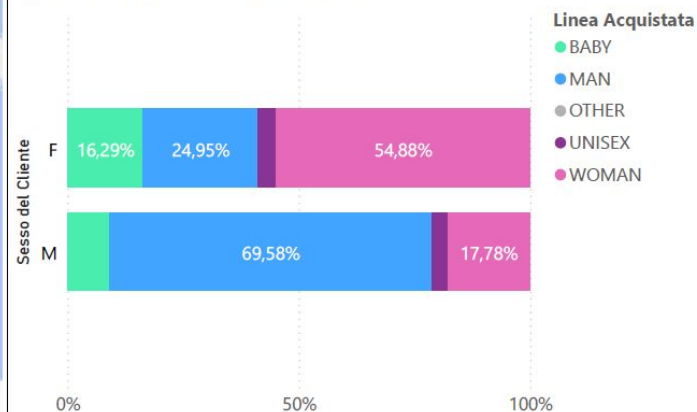
'30 - '45: Builder
'46 - '65: Baby Boomer
'66 - '80: Gen X
'81 - '95: Gen Y
'96 - '03: Gen Z

Undefined:
anno di nascita missing o
non plausibile

Distribuzione clienti per Paese



Linea acquistata per Genere



➤ In media i **"Builders"** hanno importi di spesa più alti rispetto alle altre generazioni; tuttavia è la generazione meno rappresentata tra i clienti.

➤ Le **donne** hanno un target di acquisto misto mentre l'uomo acquista per lo più capi maschili.

Data Ingestion

Data Cleaning

Tabella OLAP

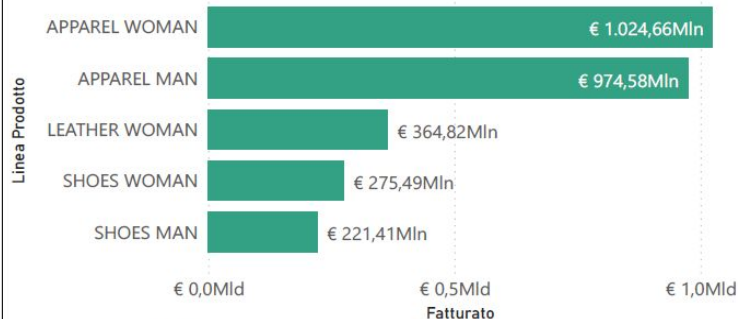
Descrittive

Clienti

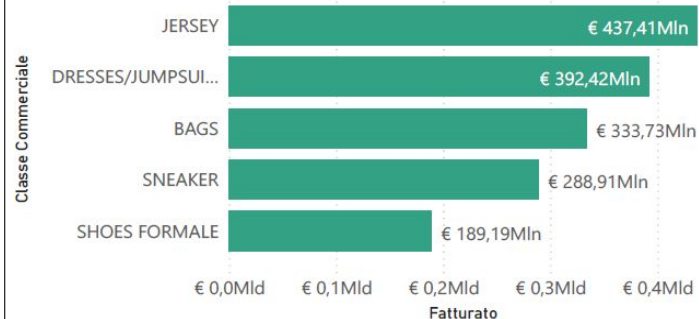
Statistiche Descrittive

PRODOTTI

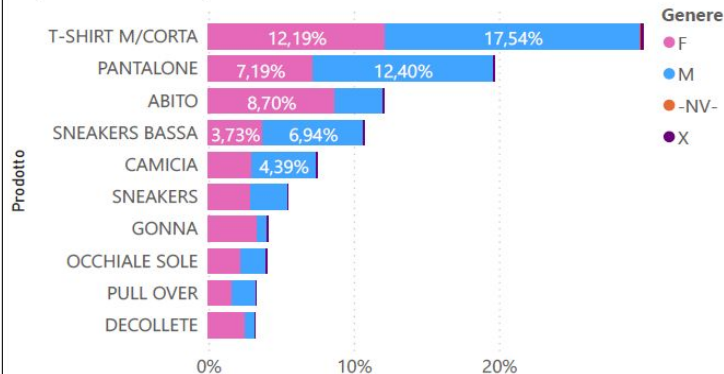
Top 5 Linee prodotto per Fatturato



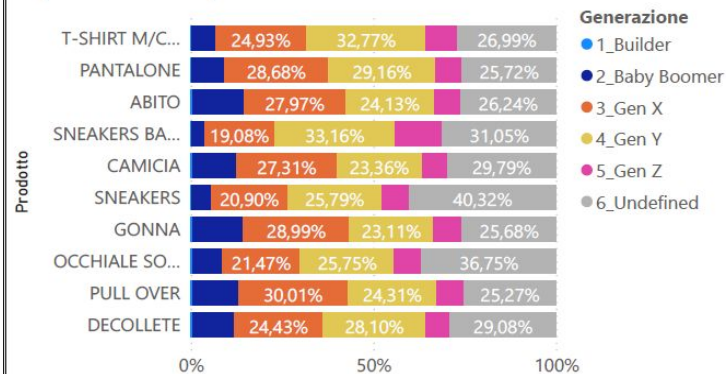
Top 5 Classi Commerciali per Fatturato



Top 10 Prodotti per Genere del Cliente



Top 10 Prodotti per Generazione



- L'abbigliamento **femminile** costituisce la principale fonte di fatturato.
- Le maglie e a seguire gli abiti e le borse rappresentano le classi commerciali più presenti per fatturato.

Totali

Linee prodotto: 28

Classi Commerciali: 124

Prodotti: 2,2 k

Articoli: 127,2 k

Legenda generazioni

'30 - '45: Builder

'46 - '65: Baby Boomer

'66 - '80: Gen X

'81 - '95: Gen Y

'96 - '03: Gen Z

Undefined:

anno di nascita *missing* o non plausibile

Data Ingestion

Data Cleaning

Tabella OLAP

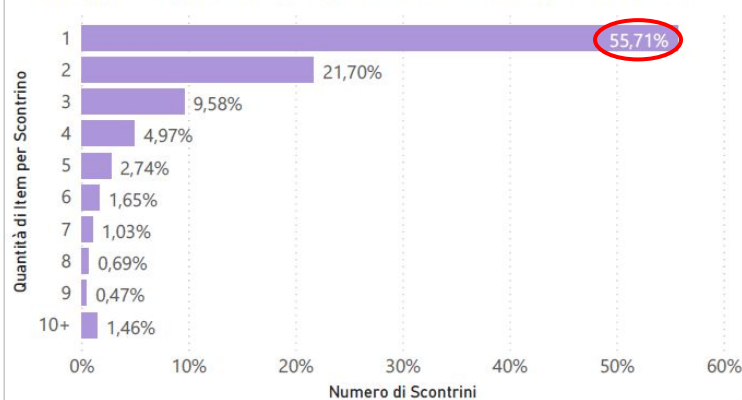
Descrittive

Prodotti

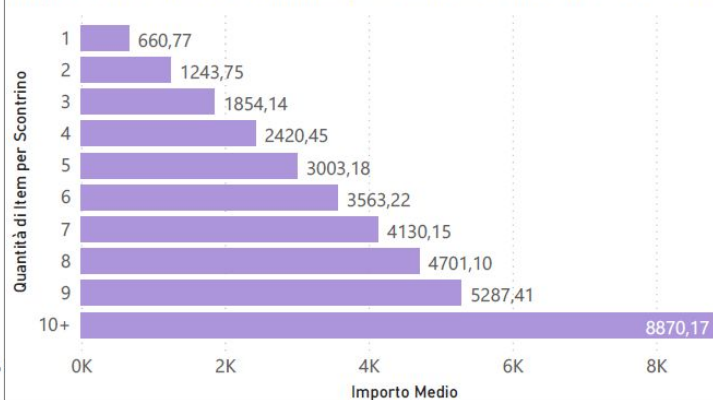
Statistiche Descrittive

SCONTRINI / NEGOZI

Conteggio degli scontrini per numero di item contenuti



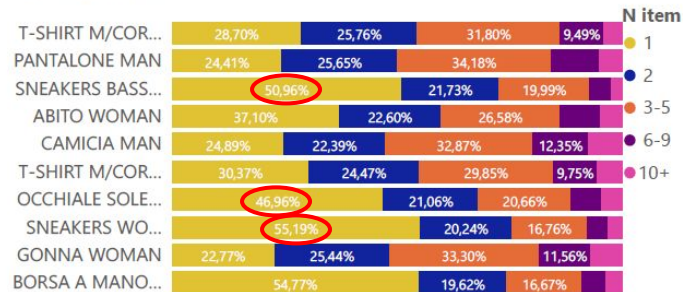
Importo medio di scontrino per numero di item contenuti



Numero di scontrini per Country



Conteggio degli scontrini per i top10 prodotti per numero di item contenuti



- Il 55% degli scontrini contiene **un solo item**.
- L'importo medio di uno scontrino è pari a **660€**; all'aumentare del numero degli item contenuti l'importo medio aumenta proporzionalmente.
- **In Europa** è presente la maggior parte dei negozi (circa il 40% del totale) seguito dalla **Cina** (25%).
- Le **sneakers** e gli **occhiali da sole** si presentano più frequentemente negli scontrini composti da un solo item (prodotti stand-alone).

Data Ingestion

Data Cleaning

Tabella OLAP

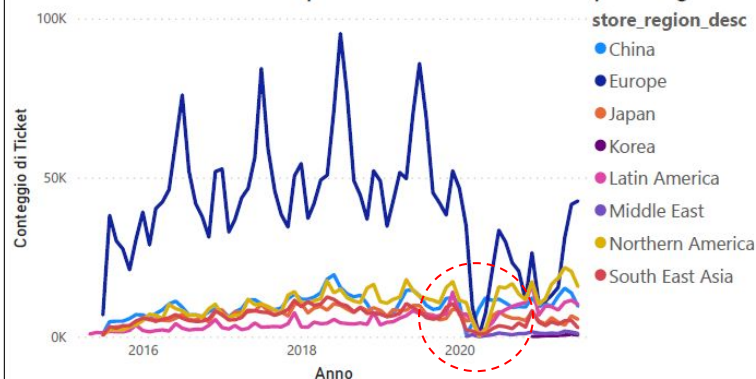
Descrittive

Scontrini/Negozi

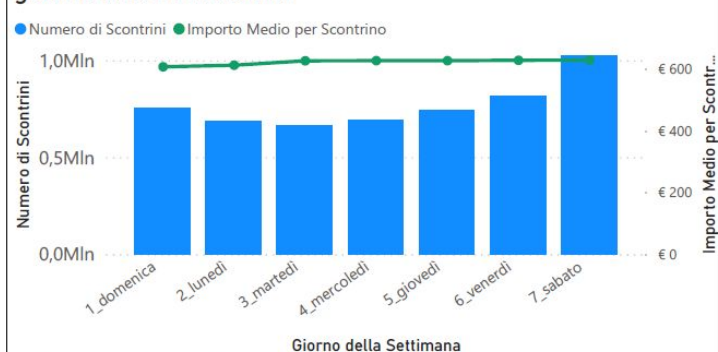
Statistiche Descrittive

ASSE TEMPORALE

Distribuzione dei Ticket per Annomese distinto per Region



Numero di scontrini e importo medio per scontrino per giorno della settimana



➤ La distribuzione temporale dei ticket risente della **stagionalità**: numero di ticket più elevato nei mesi estivi (sconti, capi meno costosi...).

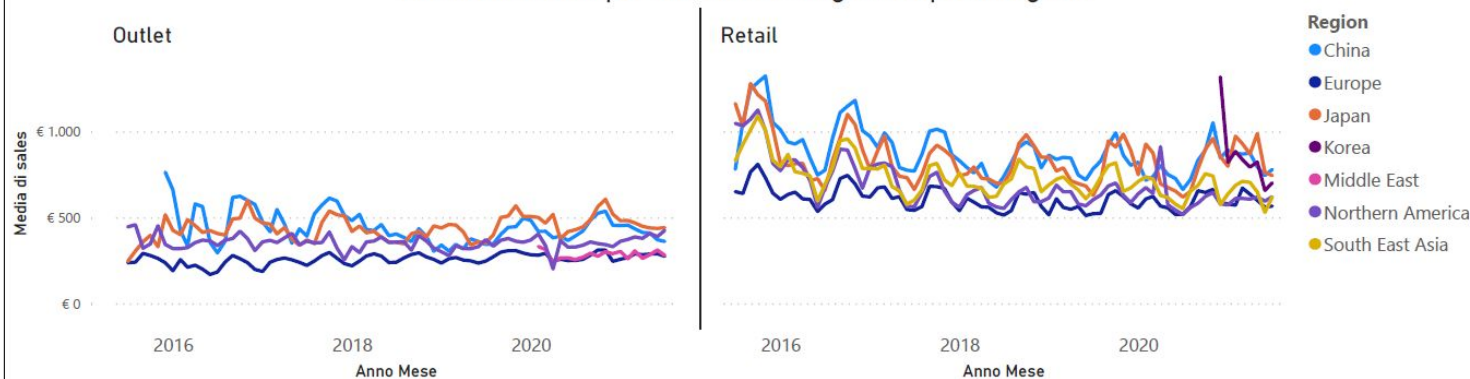
➤ **Marzo 2020**: riduzione del numero di scontrini in tutte le Region.

➤ Numero di scontrini in **Europa** ampiamente maggiori di quelli registrati in altre Region. La maggior parte dei negozi sono, infatti, in Europa.

➤ Vendite concentrate nel **weekend** (soprattutto Sabato) ma con un importo medio di scontrino costante.

➤ Fatturato **Outlet** < Fatturato **Boutique**: coerente con i prezzi medi Outlet inferiori ai prezzi medi in Boutique.

Fatturato medio per Annomese, Region e tipo di negozio



Data Ingestion

Data Cleaning

Tabella OLAP

Descrittive

Asse temporale

Market Basket Analysis

Obiettivi

- estrarre informazioni sul **comportamento d'acquisto** dei clienti
- attivare diverse **azioni di marketing** a seconda dei risultati, quali:

→ **cross-selling** proattivo

la vendita di prodotti o servizi aggiuntivi correlati al prodotto acquistato dal cliente o per il quale il cliente ha espresso interesse

ottimizzazione delle promozioni

"effetto leva" : promozione di un prodotto legato a molti altri

organizzazione della disposizione
dei prodotti nel punto vendita

gestione del magazzino

Data Ingestion

Data Cleaning

Tabella OLAP

Descrittive

Algoritmo: MBA

Obiettivi

Market Basket Analysis

Definizioni

- **ITEM:** singolo elemento presente nello scontrino
- **TRANSAZIONI:** scontrini
- **REGOLA:** siano A e B due itemset definiti da una partizione di un insieme di item I.
Se vale la seguente regola associativa: $A \Rightarrow B$
allora A è detto *antecedente* della regola e B *conseguente* della regola.

NB!: Non viene considerata la quantità acquistata di questi prodotti, ma solo la loro presenza/assenza nello scontrino.

Market Basket Analysis

Metriche principali (1)

SUPPORT $\in [0,1]$

- di un item A : frequenza relativa ovvero (n. transazioni contenenti A) / (n. transazioni totali)
- di una regola $A \Rightarrow B$: frequenza relativa della regola nell'insieme delle transazioni, cioè la frequenza relativa delle transazioni che contengono A e B .

Scontrino	Itemset
#1	Sneaker
#2	Abito
#3	Sneaker, T-Shirt

Item	Support
Sneaker	$2/3 = 0,66$
Abito	$1/3 = 0,33$
T-shirt	$1/3 = 0,33$
Sneaker, T-Shirt	$1/3 = 0,33$

Valori soglia tipici: **min Support** $\in [0.02, 0.1]$



alta: pochi frequent itemsets e poche regole, che però accadono spesso

bassa: molti frequent itemset e molte regole, che perlopiù accadono raramente

Il support è l'indice più semplice in quanto è la percentuale di transazioni che include due specifici prodotti

Market Basket Analysis

Metriche principali (2)

CONFIDENCE $\in [0,1]$

- rappresenta la proporzione di transazioni contenenti A che contengono anche B
- stima di una probabilità condizionata: $\text{support}(A \cup B) / \text{support}(A)$

Scontrino	Itemset
#1	Sneaker
#2	Abito
#3	Sneaker, T-Shirt

Regola	Confidence
Sneaker \Rightarrow T-Shirt	$(1/3) / (2/3) = 0,5$
T-Shirt \Rightarrow Sneaker	$(1/3) / (1/3) = 1$

Valori soglia tipici: **min Confidence** $\in [0,7, 0,9]$



alta: poche regole, molto forti

bassa: molte regole, di cui molte incerte

La confidence esprime l'accuratezza della regola e aiuta a comprendere la direzione del cross selling.



Regola valida

Support(regola) > **min Supp** and
Confidence(regola) > **min Conf**

Data Ingestion

Data Cleaning

Tabella OLAP

Descrittive

Algoritmo: MBA

Metriche

Market Basket Analysis

Metriche principali (3)

LIFT $\in [0, +\infty)$

confronto tra la probabilità di trovare A e B insieme nel carrello e la probabilità di trovarli nel carrello nell'ipotesi di indipendenza: $\text{confidence}(A \Rightarrow B) / \text{support}(B)$



>1: associazione **positiva** (A e B si verificano assieme più spesso di quanto non si verificherebbero se fossero associati casualmente)

=1: associazione **casuale** (A e B indipendenti)

<1: associazione **negativa** (A e B si sostituiscono, l'acquisto dell'uno con ogni probabilità fa evitare l'acquisto dell'altro)

Scontrino	Itemset
#1	Sneaker
#2	Abito
#3	Sneaker, T-Shirt

Regola	Lift
Sneaker \Rightarrow T-Shirt	$0,5/0,33 = 1,51$
T-Shirt \Rightarrow Sneaker	$0,5/0,66 = 0,75$

Market Basket Analysis

Algoritmo Apriori

Approccio per livelli per generare le regole di associazione:

- estrazione di regole con $\text{Supp} > \text{minSupp}$ e $\text{Conf} > \text{minConf}$ e con *conseguente* composto di **un solo elemento**;
- queste regole vengono poi usate per generare nuove regole candidate.

1. Identificazione Frequent Itemset

$\{a,b\}$ è **superset** di $\{a\}$; $\{a,b\}$ è **subset** di $\{a,b,d\}$

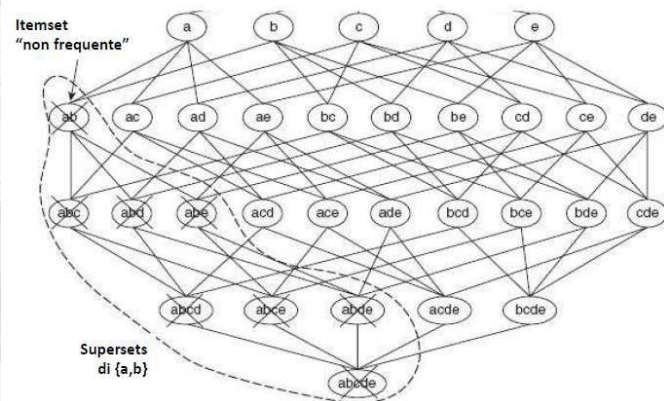
- Se un itemset $\{a,b\}$ è frequente allora anche tutti i suoi subset lo sono $\{a\}$

⇒ $\text{supp}(\{a\}) \geq \text{supp}(\{a,b\})$.

- Se un itemset $\{a,b\}$ non è frequente allora anche tutti i suoi superset $\{a,b,d\}$ non lo sono.

2. Generazione Regole Associate

- Risultato: regole che sopravvivono alle condizioni ⇒ regole valide (alto support e alta confidence)
- Risparmio di tempo e risorse ⇒ algoritmo "sgrava" il computer con una computazione intelligente



NB: La regola, inoltre, deve essere interessante per l'utilizzatore.

ES: "se partorisce è femmina" ⇒ ha confidence 1 ma non dice nulla di interessante!!

Market Basket Analysis

Fasi operative

1. Ristrutturazione del dataset : record = scontrino {itemset}

```
id;Ticket;items;  
1;000000DG38809962018-06-12;['PROFUMI UNISEX', 'PANTALONE MAN'];  
2;000000DG14428422019-05-17;['CAMICIA WOMAN', 'ABITO WOMAN'];
```



2. Eliminazione degli scontrini con 1 solo item

3. Settaggio soglie minime (arbitrarie, molto basse)

- min Supp = 0,1 %
- min Conf = 1,0 %



4. Algoritmo eseguito per granularità crescente di "item" (PySpark.ml.fpm)

5. Estrazione regole e valutazione congiunta delle metriche

classe commerciale
prodotto
prodotto + genere



```
## Unità: scontrino (cliente-giorno-negozio) , item : model desc 3 (prodotto riclassificato + MAN/WOM/BABY/)  
#raggruppo  
groups_model_3 = sales_dg.groupby(['Ticket']).agg(F.collect_set('prod_model_desc_3').alias('items'))  
#tolgo i record con solo 1 item  
groups_model_3= groups_model_3.select('*',size('items').alias('n_items'))  
groups_model_3_no_singoli = groups_model_3.filter(groups_model_3.n_items > 1)
```

eliminazione degli
scontrini con 1 solo item

Data Ingestion

Data Cleaning

Tabella OLAP

Descrittive

Algoritmo: MBA

Fasi

Market Basket Analysis

Risultati



```
from pyspark.ml.fpm import FPGrowth
fpGrowth = FPGrowth(itemsCol="items", minSupport=0.001, minConfidence=0.01)
model = fpGrowth.fit(groups)

# Display frequent itemsets.
fp = model.freqItemsets
#fp.sort(fp.freq.desc()).show(n=20)

#! salvare questi output in csv !
regole=model.associationRules

#### regole con dataset filtrato : item > 1 ####

fpGrowth = FPGrowth(itemsCol="items", minSupport=0.001, minConfidence=0.01)
model = fpGrowth.fit(groups_no_singoli)

# Display frequent itemsets.
fp = model.freqItemsets
#fp.sort(fp.freq.desc()).show(n=20)

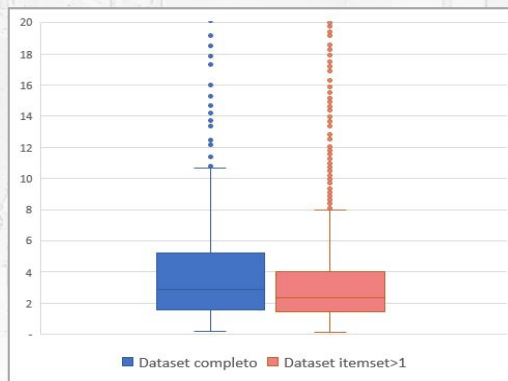
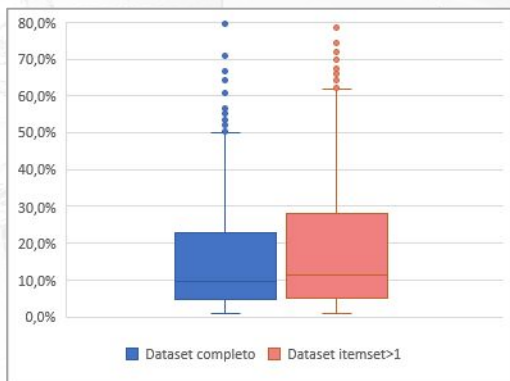
#! salvare questi output in csv !
regole_no_singoli=model.associationRules
#regole.sort(regole.support.desc()).show()
```

prova 1 :
dataset completo

prova 2 :
dataset itemset > 1

	n scontrini	n regole estratte
dataset completo	2,65 mln	577
dataset itemset > 1	1,05 mln	2.184

vengono estratte più regole in quanto mediamente i Support si alzano e quindi le regole candidate ad essere valide sono maggiori



Confidence

Lift

Supp

dataset	Min	1Q	Mediana	3Q	Max
completo	1,00%	4,60%	9,61%	22,92%	79,54%
itemset>1	1,00%	5,16%	11,20%	27,92%	79,54%

dataset	Min	1Q	Mediana	3Q	Max
completo	0,23	1,56	2,88	5,23	116,40
itemset>1	0,16	1,45	2,37	4,05	63,41

dataset	Min	1Q	Mediana	3Q	Max
completo	0,10%	0,12%	0,16%	0,25%	3,86%
itemset>1	0,10%	0,12%	0,16%	0,27%	9,75%

Data Ingestion

Data Cleaning

Tabella OLAP

Descrittive

Algoritmo: MBA

Risultati

Market Basket Analysis

Top 5 Regole Woman

	antecedente	conseguente	confidence	lift	support
1	FELPA CON ZIP WOMAN	PANTALONE WOMAN	66,5%	14,793	0,25%
2	POLO MAN, T-SHIRT M/CORTA WOMAN	T-SHIRT M/CORTA MAN	57,2%	2,420	0,12%
3	T-SHIRT M/CORTA WOMAN, PANTALONE MAN	T-SHIRT M/CORTA MAN	55,9%	2,366	0,44%
4	TOP WOMAN, PANTALONE WOMAN, GONNA WOMAN	ABITO WOMAN	54,9%	4,869	0,11%
5	BLOUSE WOMAN, PANTALONE WOMAN, GONNA WOMAN	ABITO WOMAN	54,5%	4,828	0,10%

➤ Le regole con confidence oltre al 50% rappresentano un subset di item con support molto basso, quindi raramente questi item si trovano tutti su un unico scontrino

➤ Le clienti che acquistano collana e orecchini rispetto a quelle che acquistano almeno la collana rappresentano il 37,4%.

➤ Le clienti che acquistano acquistano orecchini e collana su quelle che acquistano almeno gli orecchini sono il 16,7%.

E' ragionevole ritenere più adeguato raccomandare orecchini a chi sta acquistando una collana che viceversa.

	antecedente	conseguente	confidence	lift	support
1	COLLANA WOMAN	ORECCHINI WOMAN	37,4%	28,292	0,22%
2	ORECCHINI WOMAN	COLLANA WOMAN	16,7%	28,292	0,22%
3	FELPA CON ZIP WOMAN	PANTALONE WOMAN	66,5%	14,793	0,25%
4	PANTALONE WOMAN	FELPA CON ZIP WOMAN	5,5%	14,793	0,25%
5	CANOTTA WOMAN, GONNA WOMAN	CARDIGAN C/BOTTONI WOMAN	31,2%	13,020	0,14%

Data Ingestion

Data Cleaning

Tabella OLAP

Descrittive

Algoritmo: MBA

Regole Woman

Market Basket Analysis

Top 5 Regole Woman

	antecedente	conseguente	confidence	lift	support
1	GONNA WOMAN	ABITO WOMAN	30,9%	2,740	2,11%
2	ABITO WOMAN	GONNA WOMAN	18,7%	2,740	2,11%
3	T-SHIRT M/CORTA WOMAN	T-SHIRT M/CORTA MAN	22,1%	0,937	1,64%
4	DECOLLETE WOMAN	ABITO WOMAN	34,6%	3,068	1,55%
5	ABITO WOMAN	DECOLLETE WOMAN	13,8%	3,068	1,55%

➤ Ogni 100 transazioni effettuate, 2 contengono una gonna ed un abito.

Rappresentando la solidità della regola ci da indicazioni sul potenziale economico delle regole che coinvolgono la Gonna e l'Abito.

Data Ingestion

Data Cleaning

Tabella OLAP

Descrittive

Algoritmo: MBA

Regole Woman

Market Basket Analysis

Top 5 Regole Man

	antecedente	conseguente	confidence	lift	support
1	GIACCA MAN, CAMICIA MAN, T-SHIRT M/CORTA MAN	PANTALONE MAN	74,4%	3,414	0,14%
2	GIUBBOTTO MAN, CAMICIA MAN, T-SHIRT M/CORTA MAN	PANTALONE MAN	71,9%	3,298	0,11%
3	PULL OVER MAN, SNEAKERS BASSA MAN, T-SHIRT M/CORTA MAN	PANTALONE MAN	71,1%	3,265	0,10%
4	FELPA CAPPUCCIO MAN, CAMICIA MAN, T-SHIRT M/CORTA MAN	PANTALONE MAN	70,9%	3,253	0,13%
5	FELPA CAPPUCCIO MAN, POLO MAN, PANTALONE MAN	T-SHIRT M/CORTA MAN	70,6%	2,989	0,11%



	antecedente	conseguente	confidence	lift	support
1	SLIP MEDIO MAN	SLIP BRANDO MAN	61,9%	63,409	0,30%
2	SLIP BRANDO MAN	SLIP MEDIO MAN	31,1%	63,409	0,30%
3	PAPILLON MAN, CAMICIA MAN	ABITO MAN	55,2%	23,385	0,23%
4	POCHETTE MAN	CRAVATTA MAN	34,0%	22,855	0,20%
5	CRAVATTA MAN	POCHETTE MAN	13,4%	22,855	0,20%



➤ Le regole riportate di seguito risultano poco frequenti (Support bassi) ma molto accurate (Confidence alta) e con un forte legame tra i prodotti (lift>1).

➤ Risultano regole **complesse** con combinazioni di quattro prodotti, che compongono quasi un *total look*; questo spiega anche i bassi livelli di Support.

➤ Osservando le prime due regole per ordinamento di Lift, sembra più probabile che acquistando uno Slip Medio si acquisti uno Brand che non viceversa.

I due prodotti vengono acquistati insieme circa 63 volte di più di quanto accadrebbe nell'ipotesi di indipendenza.

Data Ingestion

Data Cleaning

Tabella OLAP

Descrittive

Algoritmo: MBA

Regole Man

Market Basket Analysis

Top 5 Regole Man

	antecedente	conseguente	confidence	lift	support
1	PANTALONE MAN	T-SHIRT M/CORTA MAN	44,7%	1,894	9,75%
2	T-SHIRT M/CORTA MAN	PANTALONE MAN	41,3%	1,894	9,75%
3	SNEAKERS BASSA MAN	T-SHIRT M/CORTA MAN	38,1%	1,612	3,87%
4	T-SHIRT M/CORTA MAN	SNEAKERS BASSA MAN	16,4%	1,612	3,87%
5	CAMICIA MAN	PANTALONE MAN	41,7%	1,916	3,37%



Un lift di 1,89 significa che i consumatori acquistano i prodotti insieme circa 2 volte di più di quanto accadrebbe se ci fosse indipendenza tra i due prodotti.

Data Ingestion

Data Cleaning

Tabella OLAP




Descrittive

Algoritmo: MBA




Regole Man

Market Basket Analysis

Top 5 Regole Baby

	antecedente	conseguente	confidence	lift	support
1	BERMUDA BABY MALE, PANTALONE BABY MALE	T-SHIRT M/CORTA BABY MALE 	79,54%	18,72	0,27%
2	BERMUDA BABY MALE, T-SHIRT M/CORTA BABY FEMALE	T-SHIRT M/CORTA BABY MALE 	78,57%	18,49	0,13%
3	BERMUDA BABY MALE, ABITO BABY FEMALE	T-SHIRT M/CORTA BABY MALE	75,23%	17,70	0,10%
4	BERMUDA BABY MALE, CAPPuccio CON ZIP BABY MALE	T-SHIRT M/CORTA BABY MALE	75,05%	17,66	0,14%
5	T-SHIRT M/LUNGA BABY MALE, CAPPuccio CON ZIP BABY MALE	PANTALONE BABY MALE 	74,33%	21,75	0,13%

Anche qui come visto precedentemente regole complesse, o con molti item, sono accompagnate da livelli di support relativamente bassi. Ma gli indicatori di accuratezza (confidence) e legame (lift) suggeriscono una bontà della relazione dei prodotti.

	antecedente	conseguente	confidence	lift	support
1	FELPA CON ZIP BABY FEMALE	PANTALONE BABY FEMALE 	65,09%	36,81	0,22%
2	PANTALONE BABY FEMALE	FELPA CON ZIP BABY FEMALE 	12,46%	36,81	0,22%
3	BLOUSE BABY FEMALE	GONNA BABY FEMALE 	35,42%	30,46	0,14%
4	GONNA BABY FEMALE	BLOUSE BABY FEMALE	12,02%	30,46	0,14%
5	CAPPuccio CON ZIP BABY FEMALE	PANTALONE BABY FEMALE	53,16%	30,07	0,24%

Importante osservare che livelli di Lift alti possono essere causati da un livello basso di Support del conseguente e/o da una confidence dei due prodotti alta (e.g 65,09%).

Data Ingestion

Data Cleaning

Tabella OLAP

Descrittive

Algoritmo: MBA

Regole Baby

Market Basket Analysis

Top 5 Regole Baby

	antecedente	conseguente	confidence	lift	support
1	PANTALONE BABY MALE	T-SHIRT M/CORTA BABY MALE 	43,3%	10,191	1,48%
2	T-SHIRT M/CORTA BABY MALE	PANTALONE BABY MALE 	34,8%	10,191	1,48%
3	BERMUDA BABY MALE	T-SHIRT M/CORTA BABY MALE	74,3%	17,486	0,80%
4	T-SHIRT M/CORTA BABY MALE	BERMUDA BABY MALE	18,8%	17,486	0,80%
5	CAPPUCCIO CON ZIP BABY MALE	PANTALONE BABY MALE	62,5%	18,278	0,78%

Tranne che per la regola 5 siamo di fronte a combinazioni inverse degli stessi prodotti.

Sembrerebbe più probabile che all'acquisto di un Pantalone o di un Bermuda segua l'acquisto di una T-Shirt, e che ad una Felpa con Zip segua un Pantalone.

L'acquisto congiunto di questi prodotti avviene tra le 11 e le 18 volte di più di quanto non accadrebbe per caso.

Data Ingestion

Data Cleaning

Tabella OLAP

Descrittive

Algoritmo: MBA

Regole Baby

Filtro di selezione
sui prodotti per
Antecedent e/o
Consequent

Report Tabellare
combinazione dei
prodotti e relative
misure di Confidence
e Support

Grafo

- nodo: prodotto
- dimensione nodo: Confidence

MBA Report

Data Source: Internal Data (Business + CRM)

Data Update: July 2021

Filter Antecedent

Selezioni multiple

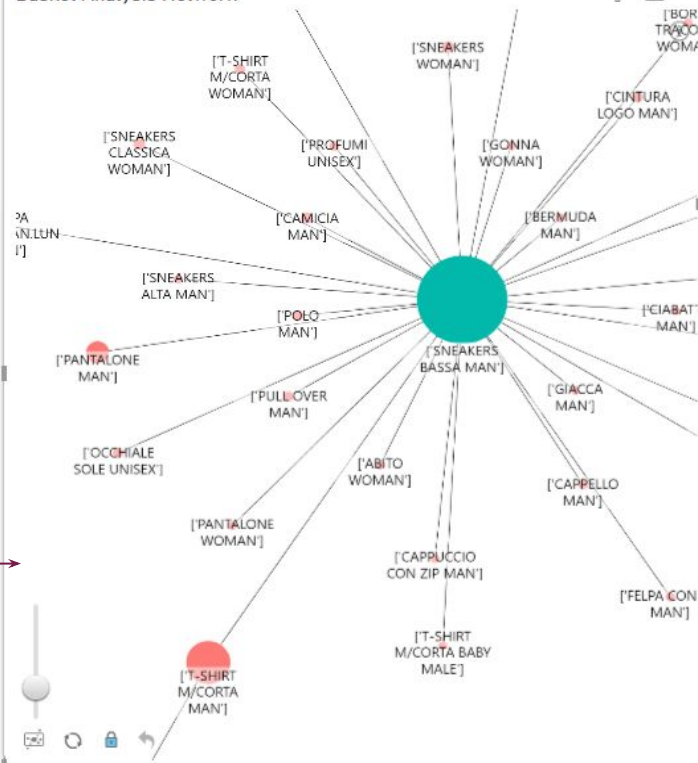
Filter Consequent

Tutte

Basket Analysis Detail

Antecedent	Consequent	Confidence	Support
['SNEAKERS BASSA MAN']	['T-SHIRT M/CORTA MAN']	38,0%	3,87%
['SNEAKERS ALTA MAN']	['T-SHIRT M/CORTA MAN']	34,0%	0,12%
['SNEAKERS ALTA MAN']	['SNEAKERS BASSA MAN']	30,0%	0,11%
['SNEAKERS BASSA MAN']	['PANTALONE MAN']	29,0%	2,99%
['SNEAKERS BASSA MAN']	['SNEAKERS BASSA WOMAN']	9,0%	0,90%
['SNEAKERS BASSA MAN']	['FELPA CAPPUCCIO MAN']	7,0%	0,67%
['SNEAKERS BASSA MAN']	['SNEAKERS WOMAN']	7,0%	0,69%
['SNEAKERS BASSA MAN']	['CAMICIA MAN']	6,0%	0,63%
['SNEAKERS BASSA MAN']	['FELPA GIROC.MAN.LUN MAN']	6,0%	0,56%
['SNEAKERS BASSA MAN']	['POLO MAN']	6,0%	0,56%
['SNEAKERS BASSA MAN']	['SNEAKERS CLASSICA WOMAN']	6,0%	0,60%
['SNEAKERS BASSA MAN']	['OCCHIALE SOLE UNISEX']	5,0%	0,49%
['SNEAKERS BASSA MAN']	['CINTURA LOGO MAN']	4,0%	0,45%
['SNEAKERS BASSA MAN']	['GIUBBOTTO MAN']	4,0%	0,38%
['SNEAKERS BASSA MAN']	['PULL OVER MAN']	4,0%	0,36%
['SNEAKERS BASSA MAN']	['T-SHIRT M/CORTA WOMAN']	4,0%	0,45%
['SNEAKERS BASSA MAN']	['CAPPELLO MAN']	3,0%	0,32%
['SNEAKERS BASSA MAN']	['GIUBBOTTO CON ZIP MAN']	3,0%	0,29%
['SNEAKERS BASSA MAN']	['PROFUMI UNISEX']	3,0%	0,31%
['SNEAKERS BASSA MAN']	['REGULAR BOXER UNDERWARE MAN']	3,0%	0,33%
['SNEAKERS BASSA MAN']	['ABITO WOMAN']	2,0%	0,20%
['SNEAKERS BASSA MAN']	['BERMUDA MAN']	2,0%	0,20%

Basket Analysis Network



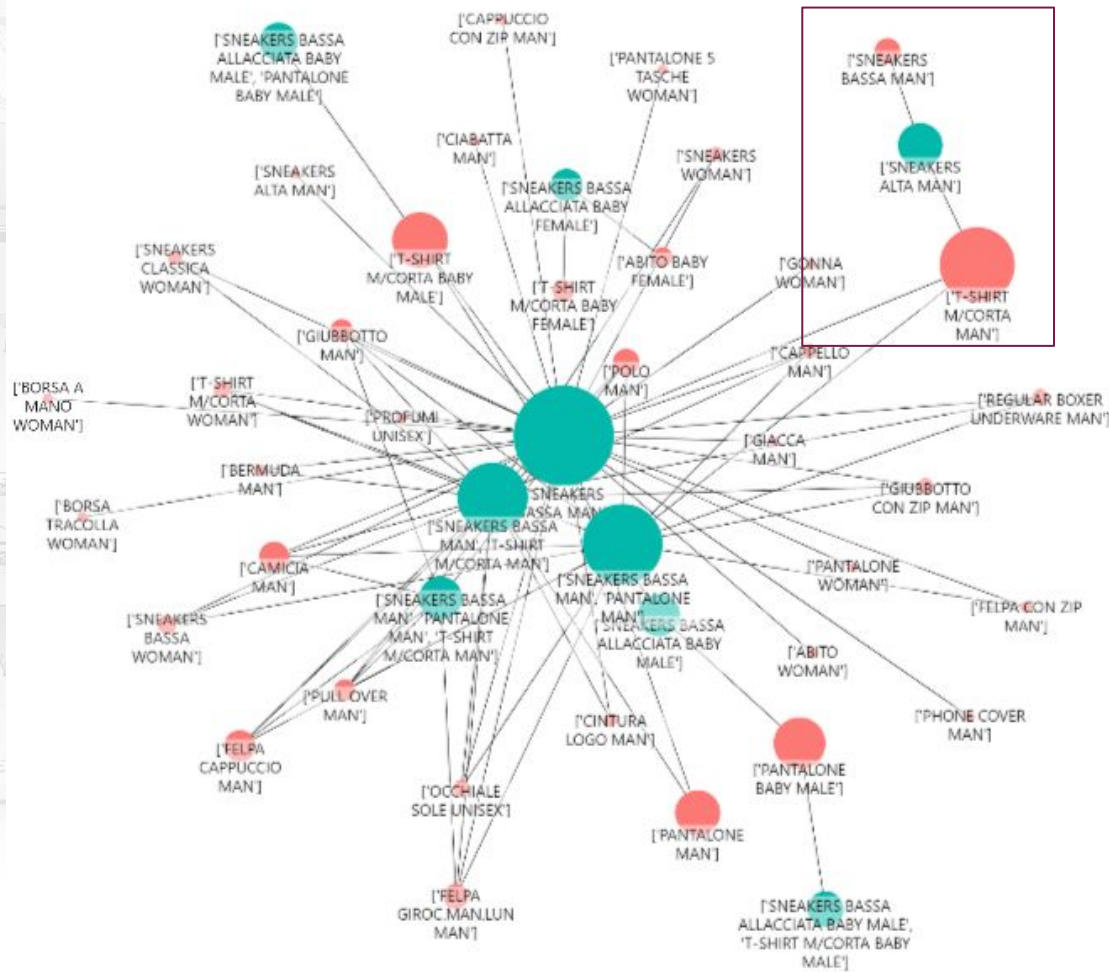
Ulteriori Insight

Collegamento del prodotto
"Sneakers", in qualità di antecedente,
con i conseguenti.

Nodi rossi: conseguenti con più alta Confidence

- T-shirt manica corta uomo
- Pantalone man

La Sneaker bassa con t-shirt manica
corta Man è a sua volta legata a
sneaker alta man



Possibili sviluppi futuri

- Possibilità di integrazione degli insight risultanti dalla MBA con algoritmi facenti parte della famiglia dei **Collaborative Filter** , che utilizzano quale metrica la distanza tra i prodotti.
- Calcolo computazionale molto più **complesso**
- Sempre da considerare il trade-off **costi-benefici**

P.IVA 02298700010

DOCUMENTO COMMERCIALE
di vendita o prestazione

DESCRIZ. IVA Prezzo(€)

1 X	100,00	
STAMPANTE	EPSON STYLUS C	
	22%	100,00
TOT.COMPLESSIVO		100,00
di cui IVA		18,03

Francesco Tirinato.....	€ 1.000
Fabio Bragato.....	€ 1.000
Diego Carrettoni.....	€ 1.000
Redaelli Valeria.....	€ 1.000
Perticarà Sophia.....	€ 1.000

GRAZIE E ARRIVEDERCI