SAA
School of Management

Università degli Studi di Torino

BUSINESS INTELLIGENCE E SISTEMI INFORMATIVI DIREZIONALI

Data Analysis & Visualization for Management and Machine Learning

Studente/i:

Vittorio Grasso 1036705

Docente:

Prof.re Giovanni Quattrone

Tutor:

Dott.re Michael Messina

Anno accademico

2022-2023

Introduzione

Lo scopo del progetto consiste nell'analisi e nella successiva rappresentazione grafica di un dataset xlsx derivante da un'estrazione da un più ampio database di proprietà di Cerved per conto di una SPA operante nell'ambito dell'ingrosso farmaceutico Europeo.

Nello specifico l'obiettivo è quello di pervenire ad un algoritmo di automazione del l'intero processo di analisi attraverso l'utilizzo di Python.

Il processo di analisi mira ad ottenere un cruscotto di indicatori inerenti l'andamento economico finanziario del portafoglio clienti (nello specifico le farmacie) su tutto il territorio Italiano. Ciò fatto nell'ottica di monitorare sia gli andamenti per ogni regione come per l'intera nazione, sia l'efficacia dell'engagement sul campo e del supporto economico e finanziario dato ai clienti.

Inoltre è stato inserito anche un algoritmo di regressione lineare per poter fare previsioni circa il valore assunto da un parametro al variare (tramite valori espliciti) degli altri parametri di ricerca, quindi misurarne il loro impatto. Ciò per constatare che differenti regioni hanno differenti regressioni e quindi differenti ripercussioni, ed essere in grado di pervenire ad una stima di esse.

La rilevanza del progetto consiste nella sua possibilità di:

- Poter essere applicato a qualsiasi dataset contenete bilanci
- Essere valido per analizzare n°+ x tuple e pervenire sempre allo stesso cruscotto di sintesi
- Poter essere adattato rapidamente in base ai parametri di analisi che si vogliono adoperare
- Essere standardizzato e automatizzato e quindi pervenire al risultato con una riduzione del tempo stimata in circa 83,33% (da 4h a 40min)
- Poter prendere in input qualsiasi tipo di file e ritornare come output sempre xlsx in modo da non creare barriere all'accesso e diffusione dei risultati.
- Poter essere utilizzato dal top management per pervenire ad una panoramica di sintesi, per poter meglio decidere dove concentrare l'attenzione e possibilmente ripetere l'intero processo ad una granularità differente fino a rintracciare il singolo bilancio problematico di interesse su decine migliaia.

Dati

Nello specifico i dati all'interno del database consistono in bilanci (valori di conto economico e stato patrimoniale) di 846 farmacie sparse su tutto il territorio Italiano. Il primo problema che si è dovuto affrontare è stata la ricerca di una disposizione dei dati consona a poter essere analizzati con tecniche di automatizzate. Si è passati quindi da una rappresentazione del bilancio con poste in verticale ad una a vettore orizzontale, tecnica anche detta anche pivoting, così che si è potuto avere negli attributi i nomi delle poste. Ciò ha fato si che si è potuto ottenere delle singole tuple contenenti un intero bilancio in quanto le poste sono uguali per legge.

Lo shape del database così consegnato risulta essere quindi di 846r x 41c = 34.686 a cui in fase di analisi sono stati calcolati ed aggiunti 15 attributi portandolo a 47.376

Il problema di adoperare bilanci depositati come database consiste nella bontà, veridicità e coerenza delle poste e dei valori iscritti, che com'è noto possono essere ribilanciati prima del deposito in modo da enfatizzare aspetti diversi. Ambiguo sotto il profilo giuridico (in quanto il valore finale è sempre lo stesso) ma di enorme peso nel calcolo di indicatori aggregati.

Ciò però non riduce la validità della ricerca in quanto se pur vero quando detto sopra è altrettanto vero che lo strumento che si costituirà serve proprio ad andare a ricercare quelle criticità in modo da poterle meglio estrapolare, attenzionare e comprendere lì dove si tratta di reali situazioni di rischio economico e finanziario oppure no. Inoltre ciò non intacca la bontà del modello creato in quanto la bontà dei risultati ottenuti è direttamente proporzionale alla bontà dei dati su cui esso lavora.

Metodo

Come si è detto sopra l'obbiettivo del progetto era pervenire ad una analisi e ad una successiva sintesi riguardante i focus elencati nell'introduzione. Nello specifico gli obbiettivi di indagine che si erano posti erano:

1. Capire la suddivisione regionale del portafoglio clienti (farmacie)
2. Analizzare ogni portafoglio regionale in base a 3 parametri calcolabili per ogni bilancio
3. Raggruppare i valori ottenuti dei parametri in 3 categorie di rischio:
   Red-Yellow-Green
4. Analizzare la distribuzione % regionale delle tre categorie in modo da comprendere quali siano le più virtuose e quali le meno
5. Pervenire ad una selezione ed estrazione dei bilanci su base nazionale che rientrano contemporaneamente in Red sui 3 parametri
6. Estrarre tutto in un file .xlsx
7. Visualizzare una comparativa fra la distribuzione nazionale del parametro e la rispettiva distribuzione % regionale
8. Visualizzare una comparativa fra la distribuzione nazionale del livello Red di rischiosità e la distribuzione regionale del valore medio di ogni parametro
9. Creare un algoritmo di regressione lineare ed addestrarlo in modo che predica, dati dei valori noti dei 3 parametri, un quarto, riportando anche l'R2 score.

Per riuscire a svolgere tutti i passaggi elencati si è scelto di utilizzare Python come strumento, essendo in grado di gestire una quantità decisamente maggiore di calcoli e metodi rispetto a excel. Ma si è scelto di rappresentare i risultati ottenuti all'interno di un file excel per garantire i massimi livelli di fruizione da parte di ogni utente.

Solo il modello predittivo non è stato riportato in excel in quanto devono essere inserite le variabili in input come dataframe e deve essere eseguito lo script necessariamente dal programma stesso (ma verrà comunque riportato il codice per fare entrambe le cose)

Primo punto: Capire la suddivisione regionale del portafoglio clienti

Sul database importato è stato inizialmente fatto un raggruppamento contato per ricavare il numero di farmacie presenti in ogni regione. La somma è stata fatta per l'attributo "partita iva" mentre il raggruppamento ovviamente per regione. Si è scelto l'attributo partita iva perché si presta per le sue caratteristiche intrinseche ad essere una chiave primaria.

```
[114]: # ------------------ PORTAFOGLIO FARMACIE ---------------------- #
       ITA_NPharma = db.groupby("REGIONE")["PARTITA_IVA"].count().reset_index().rename(columns = {"PARTITA_IVA": "N°Farmacie"})
       ITA_NPharma.head(20)
```

[114]:

	REGIONE	N°Farmacie
0	Abruzzo	27
1	Basilicata	2
2	Calabria	37
3	Campania	42
4	Emilia-Romagna	59
5	Friuli-Venezia Giulia	15
6	Lazio	95
7	Liguria	10
8	Lombardia	200
9	Marche	31
10	Molise	4
11	Piemonte	23
12	Puglia	34
13	Sardegna	24
14	Sicilia	95
15	Toscana	63
16	Trentino-Alto Adige	8
17	Umbria	14
18	Valle d`Aosta	3
19	Veneto	59

Secondo e Terzo punto: Analizzare ogni portafoglio regionale in base a 3 parametri calcolabili per ogni bilancio - Raggruppare i valori ottenuti dei parametri in 3 categorie di rischio: Red-Yellow-Green

I parametri su cui basare l'analisi che sono stati scelti sono:

ROS (Return on sales) - MOL% (Margine operativo Lordo) - Rapporto Debiti/Ricavi

Il primo è già presente come valore in database mentre gli altri due vengono calcolati a partire da altri valori calcolati, con il seguente script:

```
db["Ricavi_Riproporz"] = (db["RICAVI"] / db["Durata operativa bilancio (mesi)"]) * 12
db["Acquisti_Riproporz"] = (db["ACQUISTI_DI_MATERIE"] / db["Durata operativa bilancio (mesi)"]) * 12
db["Costo_del_Venduto"] = db["Acquisti_Riproporz"] - db["correzione segno rimanenze"]
db["MOL"] = db["Ricavi_Riproporz"] - db["Costo_del_Venduto"]
db["MOL%"] = (db["MOL"] / db["Ricavi_Riproporz"])* 100
db["Rapporto_debiti/Ricavi_Riproporz"] = db["DEBITI"] / db["Ricavi_Riproporz"]
```

Per creare le categorie di rischio in cui far ricadere i valori dei parametri trovati si è ovviamente dovuto costituire una serie di intervalli di valori, nello specifico:

ROS:

- Green: X > 10
- Yellow: 5 < X <= 10
- Red: X <= 5

MOL%:

- Green: X > 30
- Yellow: 0 < X <= 30
- Red: X < 0

Rapporto debiti/ricavi:

- Green: X < 0.2
- Yellow: 0.2 < X < 1
- Red: X > 1

A questo punto si sono potuti impostare gli script che hanno permesso di inserire gli stati della variabile presa ad oggetto in una delle 3 categorie. Gli script sono i seguenti:

```
if (db["MOL%"].min() <= 0) == True:
    red_MOL_percent = db[db["MOL%"] <= 0]
if (db["MOL%"].max() > 30) == True:
    green_MOL_percent = db[db["MOL%"] > 30]
yellow_MOL_percent = db[(db["MOL%"]<=30) & (db["MOL%"]>0)]
```

nb. Questo è stato ripetuto per i 3 parametri presi in esame con i rispettivi valori di intervallo. La scelta possibile del numero di parametri è teoricamente infinita.

Quarto punto: Analizzare la distribuzione % regionale delle tre categorie in modo da comprendere quali siano le più virtuose e quali le meno

Per analizzare la distribuzione % del parametro nelle tre categorie su ogni regione si è svolto un procedimento complesso su diversi passaggi per queso non viene riportato qui l'itero script. Nel dettaglio è stata fatto un secondo raggruppamento contato solo nella variabili contenti i bilanci con il nostro parametro ricercato. Successivamente sono stati agganciati i raggruppamenti con un merge con in più la variabile contente il numero totale di bilanci per regione. Infine si sono inseriti i tre nuovi attributi andando a calcolare il valore della suddivisione % sul totale ed estratta la partizione contenete solo i valori di nostro interesse.

Viene di seguito riportato il risultato finale:

[116]:

	REGIONE	debiti/ricavi >1	0.2< debiti/ricavi <1	debiti/ricavi <0.2
0	Abruzzo	29.629630	40.740741	29.629630
1	Basilicata	0.000000	100.000000	0.000000
2	Calabria	16.216216	75.675676	8.108108
3	Campania	26.190476	64.285714	9.523810
4	Emilia-Romagna	20.338983	54.237288	25.423729
5	Friuli-Venezia Giulia	6.666667	33.333333	60.000000
6	Lazio	16.842105	68.421053	14.736842
7	Liguria	20.000000	50.000000	30.000000
8	Lombardia	17.000000	61.500000	21.500000
9	Marche	29.032258	61.290323	9.677419
10	Molise	0.000000	75.000000	25.000000
11	Piemonte	17.391304	60.869565	21.739130
12	Puglia	14.705882	58.823529	26.470588
13	Sardegna	16.666667	66.666667	16.666667
14	Sicilia	15.789474	71.578947	12.631579
15	Toscana	15.873016	63.492063	20.634921
16	Trentino-Alto Adige	12.500000	75.000000	12.500000
17	Umbria	7.142857	71.428571	21.428571
18	Valle d`Aosta	0.000000	33.333333	66.666667
19	Veneto	23.728814	54.237288	22.033898

Nb. la somma row wise da sempre 100%
Nbb. Ciò è stato fatto per tutti e 3 i parametri

Quinto punto: Pervenire ad una selezione ed estrazione dei bilanci su base nazionale che rientrano contemporaneamente in Red sui 3 parametri

Per fare questo si è dovuto procedere ad una query OLAP in cui si è ulteriormente chiesto di creare uno script che estraesse solo i bilanci che ricadessero contemporaneamente in Red sui 3 parametri, prescindendo la distribuzione regionale e privilegiando la possibilità di avere i dati in dettaglio per poter entrare nel merito. Ciò è stato fatto nel seguente modo:

```
dbb = red_debiti_su_Ricavi_Riproporz[red_debiti_su_Ricavi_Riproporz.PARTITA_IVA.isin(red_ROS["PARTITA_IVA"])]
dbb1 = red_ROS[red_ROS.PARTITA_IVA.isin(red_MOL_percent["PARTITA_IVA"])]
dbK = dbb[dbb.PARTITA_IVA.isin(dbb1["PARTITA_IVA"])]
```

Sesto punto: Estrarre tutto in un file .xlsx

Avendo detto in fase preliminare che il risultato dell'analisi deve essere fruibile da chiunque era stata posta la necessità che i risultati dell'analisi fossero inseriti in un file excel, all'interno del quale per comodità è stata costruita anche la corrispettiva visualizzazione grafica. Con questo si ha un unico pacchetto informativo altamente user-friendly e personalizzabile in base alle esigenze specifiche di ogni manager. Anche in questo caso la scelta inerente alla tipologia di file per l'estrazione è puramente discrezionale e nello specifico fatta per l'espressa esigenza del manager. Ciò vale a dire che il risultato dell'analisi può essere riportato in qualsiasi formato si desideri (così com'è possibile inserire i dati in input da qualsiasi formato in cui sia possibile archiviare i dati in modo strutturato.)

L'estrazione è stata fatta con il seguente script:

```
[141]: with pd.ExcelWriter("Italy_Situation.xlsx") as writer:
           Italy_MOL_percent_complessive.to_excel(writer, sheet_name = "MOL%", index=False)
           Italy_ROS_complessive.to_excel(writer, sheet_name = "ROS", index=False)
           Italy_debiti_su_Ricavi_complessive.to_excel(writer, sheet_name = "Debiti su Ricavi", index=False)
           dbK.to_excel(writer, sheet_name = "WARNING", index = False)
           Italy_mean_complessive.to_excel(writer, sheet_name = "Italy_mean_complessive.xlsx", index=False)
```

Questo è il risultato dell'estrazione in excel dei dati di sintesi dell'analisi, nello specifico qui viene riportata solo la tabella inerente al MOL% per esigenze di spazio. In allegato il file .xlsx completo.

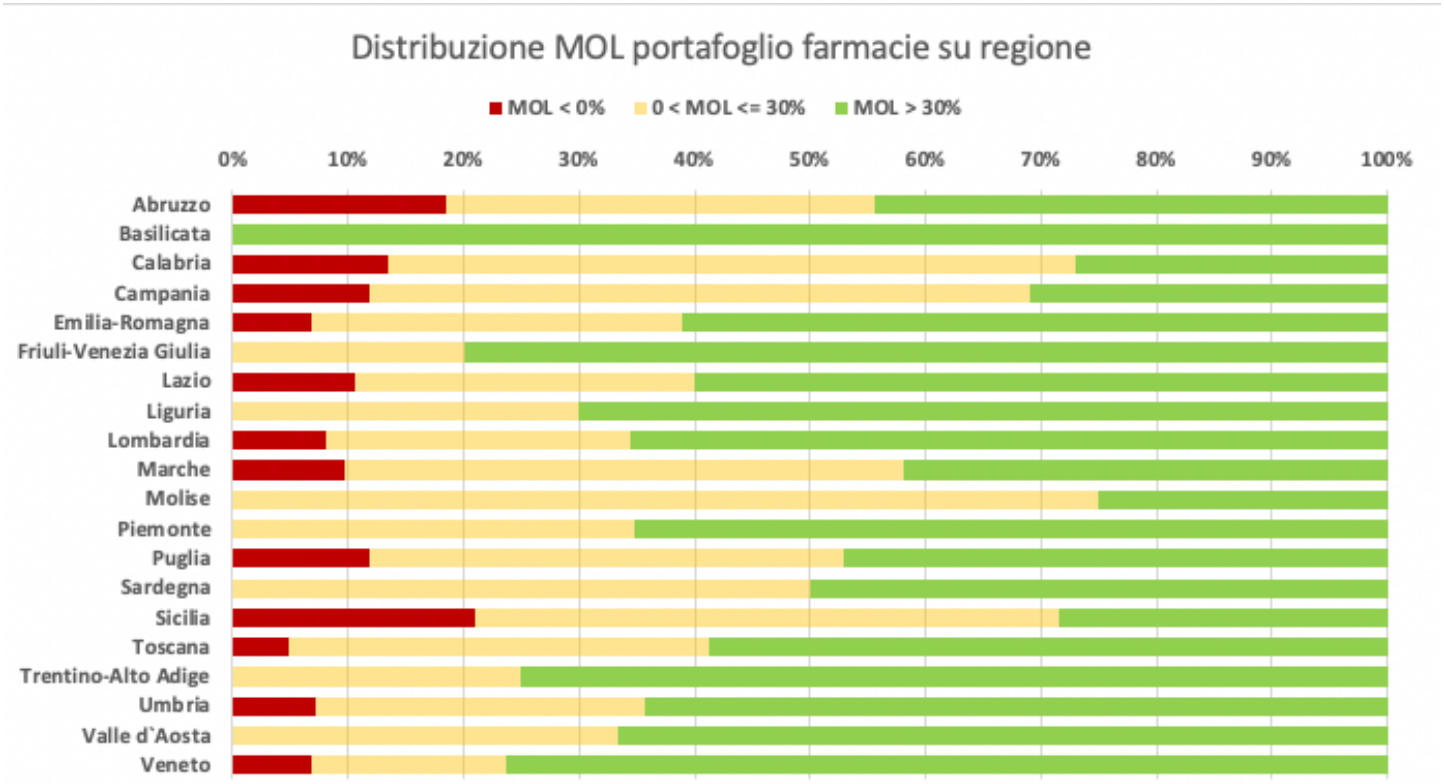
REGIONE %	MOL < 0%	0 < MOL <= 30%	MOL > 30%	Media regionale MOL
Abruzzo	18,52	37,04	44,44	33,09
Basilicata	0,00	0,00	100,00	34,21
Calabria	13,51	59,46	27,03	-1,05
Campania	11,90	57,14	30,95	23,70
Emilia-Romagna	6,78	32,20	61,02	30,46
Friuli-Venezia Giulia	0,00	20,00	80,00	35,55
Lazio	10,53	29,47	60,00	20,00
Liguria	0,00	30,00	70,00	27,02
Lombardia	8,00	26,50	65,50	24,96
Marche	9,68	48,39	41,94	-29,52
Molise	0,00	75,00	25,00	31,05
Piemonte	0,00	34,78	65,22	30,33
Puglia	11,76	41,18	47,06	17,69
Sardegna	0,00	50,00	50,00	30,49
Sicilia	21,05	50,53	28,42	-32,04
Toscana	4,76	36,51	58,73	29,08
Trentino-Alto Adige	0,00	25,00	75,00	32,74
Umbria	7,14	28,57	64,29	28,49
Valle d`Aosta	0,00	33,33	66,67	39,21
Veneto	6,78	16,95	76,27	7,95
Valore Medio Italiano	6,52	36,60	56,88	20,67

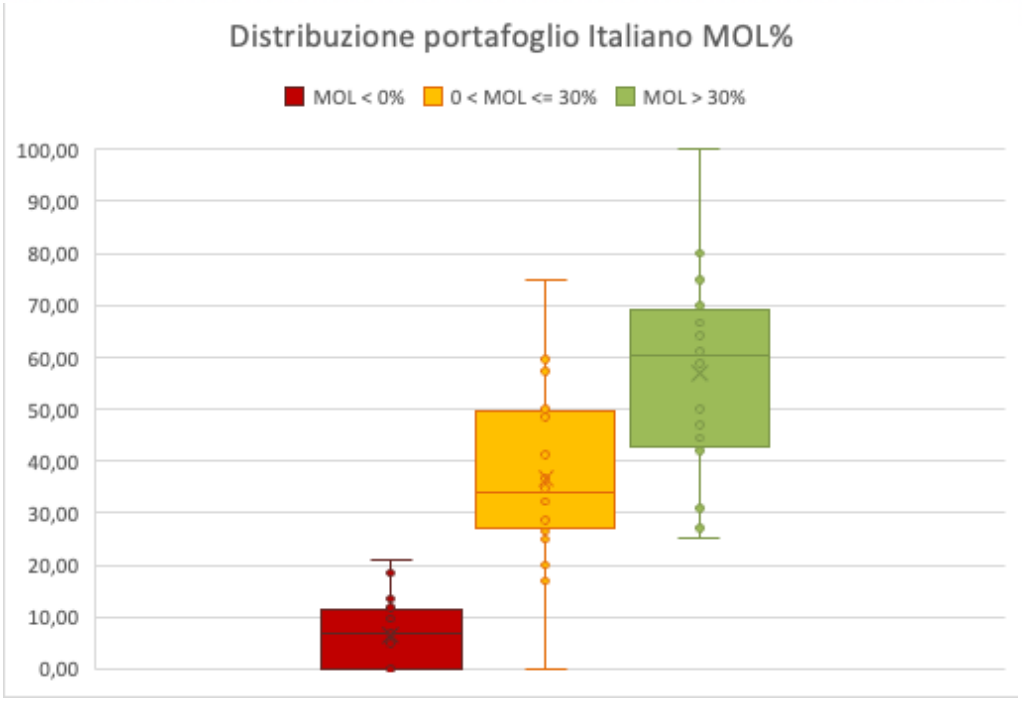
Questa tabella deve essere letta row wise e nello specifico ad esempio la regione "Abruzzo" ha il 18,52% delle farmacie sul territorio che presentano valori in bilancio che hanno portato ad un calcolo del MOL% inferiore allo 0%, un 37,04% di farmacie che presentano un MOL% compreso tra 0 e 30% e un 44,44% di farmacie che presentano un MOL% superiore al 30%. Il tutto a fronte di una media regionale di MOL%, basato sul totale delle farmacie in portafoglio sul territorio regionale, di 33,09.

Ciò porta alla conclusione che, se inizialmente si potrebbe pensare che la distribuzione è ottima (in quanto la % più elevata è sul Green), tuttavia avendo una media regionale al disopra del nostro migliore intervallo ci si concentra più sul problema che si abbia una valore di Red secondo in ordine di grandezza rispetto a tutta l'Italia. In conclusione la regione per i valori che potrebbe mettere in campo sta sottoperformando.

Settimo punto: Visualizzare una comparativa fra la distribuzione del parametro e la rispettiva distribuzione % regionale

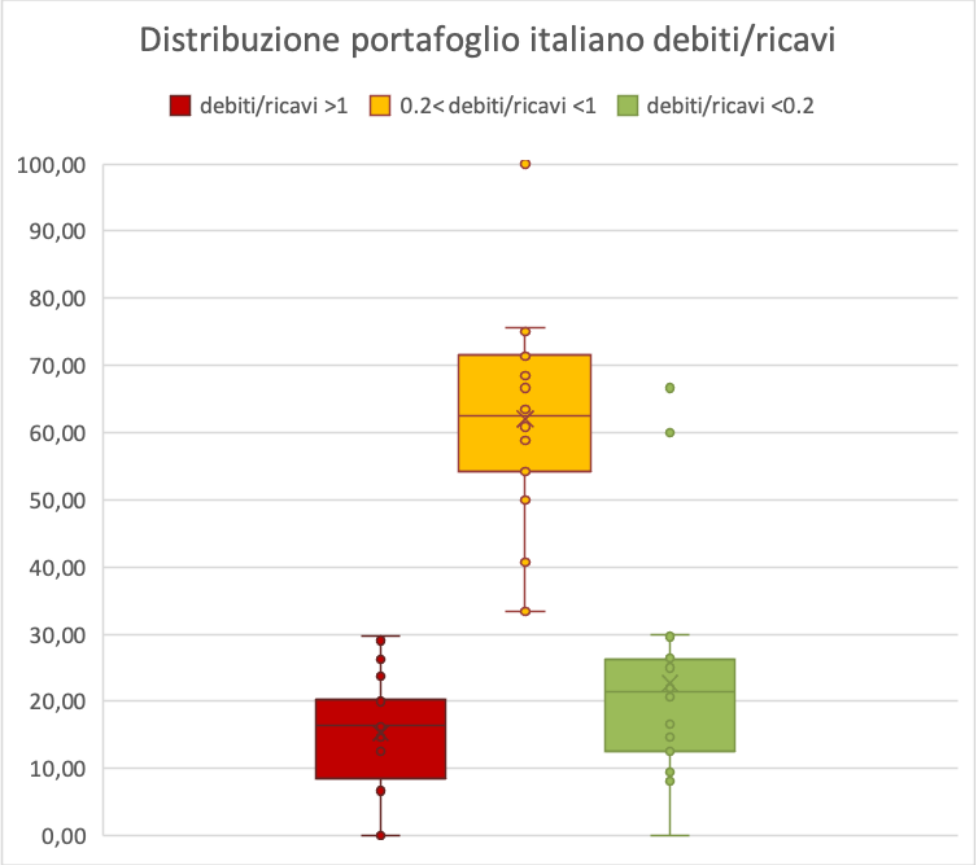
A questo punto partendo dai valori in tabella si possono creare interessanti rappresentazioni grafiche che ci danno il pregio di riuscire ad avere una visione d'insieme immediata, come anche di poter andare ad analizzare nel dettaglio eventuali comparative (dato anche dalla possibilità di essere inserito a fianco della tabella nel file excel). Le prime due rappresentazioni grafiche sono così proposte:





Già su di questi due grafici potrebbero essere svolti molti ragionamenti come ad esempio che il trend italiano riguardante la variabile ricercata è positivo in quanto si ha un box green più ampio rispetto agli altri due e con una mediana di 60% di quota posseduta di farmacie con un MOL% superiore al 30%. Ma ciò è dato dal fatto che ci sono regioni "virtuose" che trainano la distribuzione e compensano in parte le regioni "meno virtuose".

Situazione ad esempio radicalmente diversa per quanto riguarda il parametro debiti/ricavi

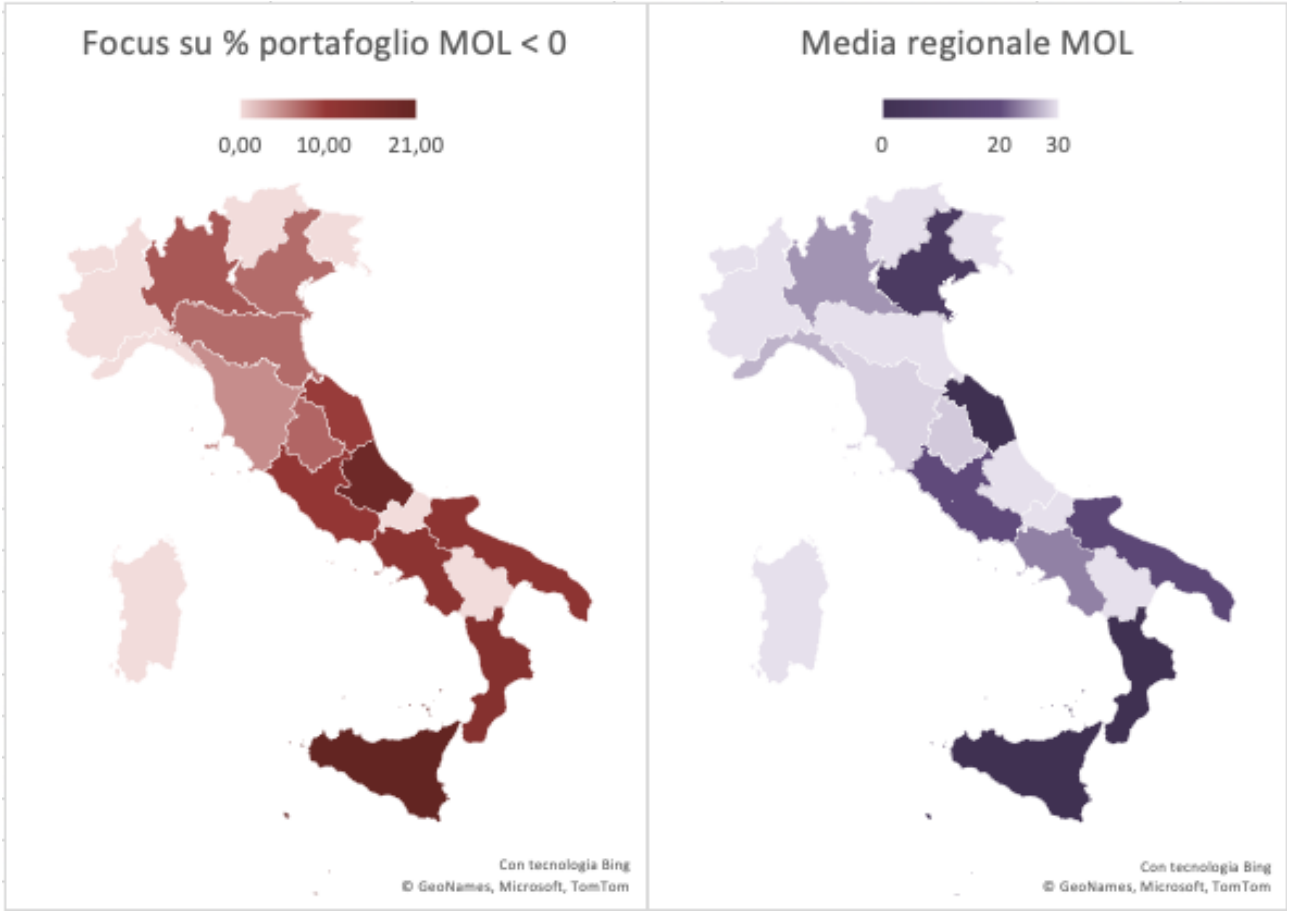


Qui, anche se potrebbe sembrare un problema, va considerato che una certa % di indebitamento è fisiologica e sintomo di una buona gestione. Nello specifico si è scelto appositamente di creare degli intervalli "estremi" ma già un valore di 0.5 viene considerato ottimo perché sta a significare che i ricavi coprono per due volte il totale dei debiti. Anche in questo caso la scelta dei valori di intervallo dei parametri di ricerca è stata fatta su richiesta manager per soddisfare le sue esigenze informative. Interessante notare qui la presenza di outlier.

Ottavo punto: Visualizzare una comparativa fra la distribuzione nazionale del livello Red di rischiosità e la distribuzione nazionale del valore medio dello stesso

Anche in questo caso possiamo ricorrere alla rappresentazione grafica per rispondere all'esigenza informativa prendendo i dati sempre dalla tabella di riferimento inerente al ROS% (stesso procedimento è applicato agli altri parametri analizzati).

Avendo una categoriazzazione geografica e una esigenza comparativa in questo caso su piano nazionale si è ritenuto opportuno e più efficace adoperare una Choropleth map, con il seguente risultato:



nb. La scala gradiente dei due colori è uguale

Il primo grafico riporta un livello di "virtuosità della regione" in proporzione alla % di farmacie presenti su quel territorio con un MOL% < 0.

Il secondo grafico riporta un livello di "virtuosità della regione" in proporzione alla media regionale di MOL% della regione stessa.

Ciò da un grosso aiuto nel visualizzare a colpo d'occhio le regioni che hanno una gradazione diversa del proprio colore, sia in rapporto al grafico in se, ma anche e soprattuto in relazione al grafico accanto.

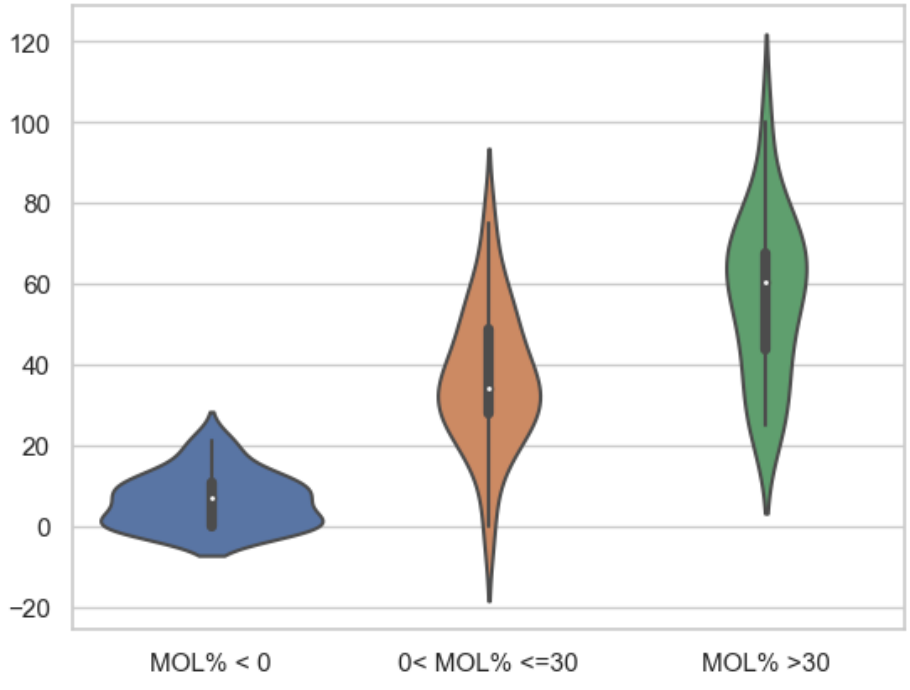
E ciò perché se esiste una differenza di gradazione della stessa regione da un grafico all'altro abbiamo una incoerenza che può rappresentare una virtuosità o una problematica.

Nello specifico in questa comparazione esistono 4 possibilità per ogni confronto della stessa regione:

- ***Entrambe chiare***: coerenza, alto MOL% medio della regione e quindi bassa % di portafoglio con MOL% < 0
- ***Entrambe scure***: coerenza, esattamente opposto al primo caso
- ***Sinistra chiara e destra scura***: incoerenza positiva, abbiamo una regione con una media MOL% molto bassa ma un portafoglio con MOL% < 0 molto piccolo.
- ***Sinistra scura e destra chiara***: incoerenza negativa, esattamente opposto al terzo caso.

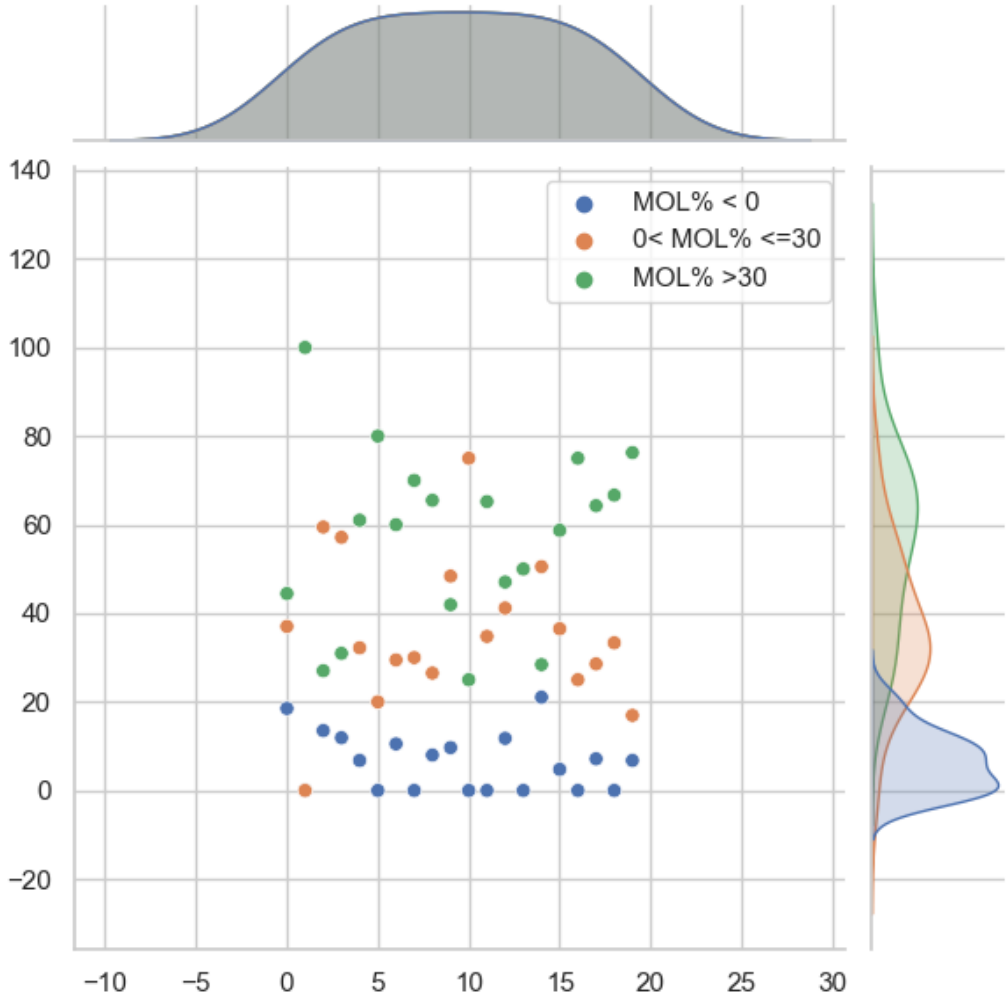
Si ricorda sempre che entrambi i grafici sono posti accanto alla tabella di riferimento, in modo da riuscire immediatamente ad ottenere il dato numerico richiamato dal grafico.

Extra: sotto vengono riportati due grafici generati all'interno di Python attraverso la libreria Seaborn sempre dalla stessa tabella di riferimento. L'interpretazione è lasciata al lettore



ViolinPlot

JointPlot



Nono punto: Creare un algoritmo di regressione lineare ed addestrarlo in modo che predica, dati dei valori noti dei 3 parametri, un quarto, riportando anche l'R2 score.

A questo punto resta da implementare l'algoritmo di predizione per poter riuscire ad avere un valore ipotetico ignoto (Y) correlato a tre variabili note (X1-X2-X3). Questo processo è ovviamente applicabile a priori di qualsiasi valore associamo alle nostre variabili e quindi è generalizzabile a qualsiasi quesito di indagine.

Le variabili prese in esame in questo caso sono ovviamente i nostri tre parametri (MOL%, ROS e Debiti/Ricavi) che andranno a fittare i valori in X, mentre per la correlazione si è scelto il ROI come valore di Y. Questo perché il ROI, acronimo di Return on Investment dipende (anche ma non solo) dal livello di redditività delle vendite e dalla gestione delle passività a breve e medio-lungo.

Si è scelto innanzi tutto di creare un database che contenesse solo i valori d'indagine dell'intero territorio italiano, per fare ciò è servita solo una riga di codice. Successivamente su questa partizione è stata posta una query per ridurre la granularità ed estrarre solo i valori inerenti alla regione di analisi. Anche in questo caso il livello di granularità dell'indagine è discrezionale, si potrebbe svolgere anche su tutta l'Italia quanto sulla singola provincia o città, salvo scendere a compromessi con la bontà della predizione data da un basso R^2 (0.15 sull'intera nazione).

```
IT_machine = db[["REGIONE","MOL%","R.O.S. (Return on Sales)", "Rapporto_debiti/Ricavi_Riproporz","R.O.I. Cerved (Return on Investment)"]]
ROI_machine = IT_machine.query('REGIONE == "Piemonte"')
ROI_machine.head(4)
```

	REGIONE	MOL%	R.O.S. (Return on Sales)	Rapporto_debiti/Ricavi_Riproporz	R.O.I. Cerved (Return on Investment)
45	Piemonte	38.577489	2.76	1.886959	0.36
110	Piemonte	17.330771	3.05	1.201086	3.91
133	Piemonte	9.229094	-8.66	1.128113	-2.86
143	Piemonte	32.307224	6.87	1.100484	6.73

Successivamente si vanno a fittare i valori d'interesse nelle rispettive variabili:

```
X = ROI_machine[["MOL%","R.O.S. (Return on Sales)", "Rapporto_debiti/Ricavi_Riproporz"]]
Y = ROI_machine["R.O.I. Cerved (Return on Investment)"]
```

Adesso si deve importare ed addestrare un modello predittivo che opera una regressione lineare multivariata in cui andremo ad inserire i nostri valori:

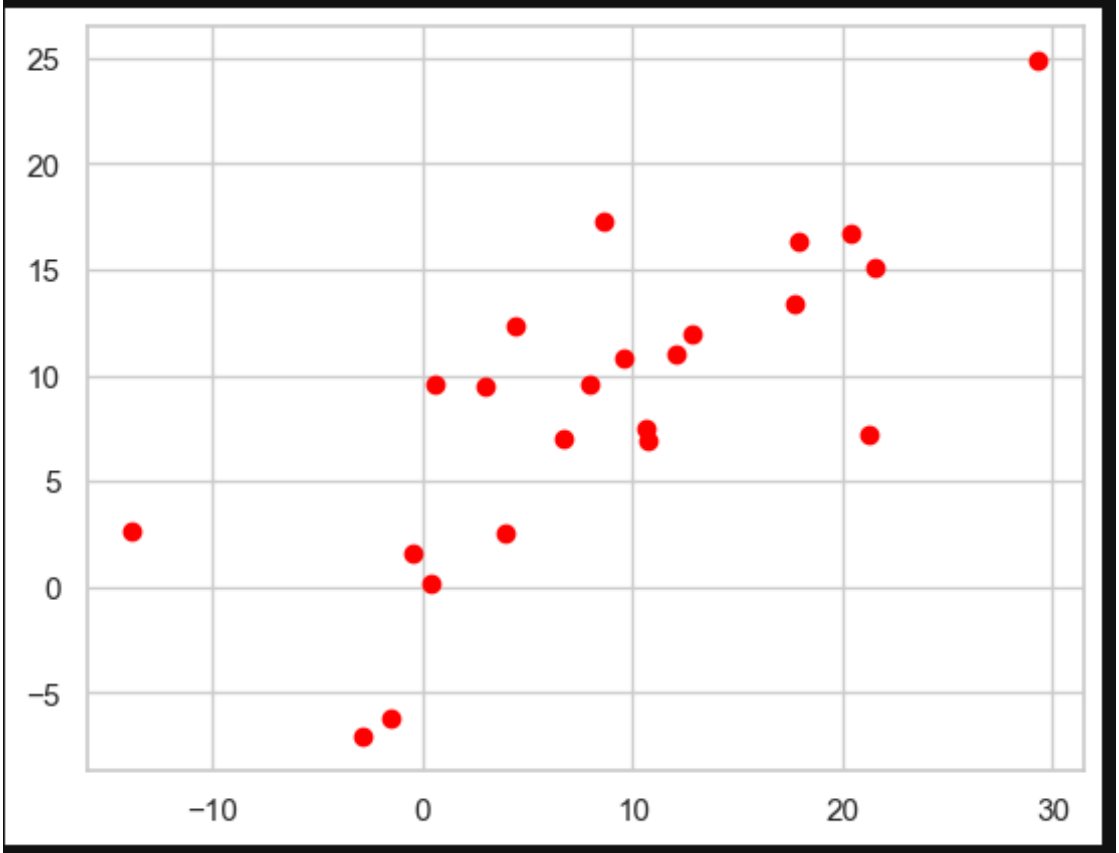
```
from sklearn.linear_model import LinearRegression

lr = LinearRegression()
lr.fit(X, Y)
```

A questo punto si deve valutare la bontà della corrispondenza e quindi se la nostra regressione ha avuto successo. Di norma per fare questo si potrebbe calcolare solo R^2 però volendo si può creare uno scatterplot per visualizzare la corrispondenza.

```
y_train_predict = lr.predict(X)
plt.scatter(Y, y_train_predict, color="Red")
```

```
<matplotlib.collections.PathCollection at 0x7fd245be93a0>
```




```
from sklearn.metrics import r2_score
r2_score(Y, y_train_predict)
```

```
0.5767463886537473
```

Di norma uno score di 0.57 è considerato medio-basso però, considerando che siamo consapevoli che la variazione del ROI dipende anche da altri fattori che non sono presi in calcolo in questo modello, e che il modello si basa su una regressione multivariata e non univariata è stato valutato sufficientemente accettabile come valore. Si potrebbe anche adoperare una regressione lineare univariata correlando il MOL% al ROS che hanno una correlazione calcolata nel nostro database di 0.76 ma si è ritenuta una ipotesi troppo scontata che esista una correlazione fra questi due parametri.

A questo punto non resta che creare un dataframe dove esplicitare i valori di X desiderati ed adoperare il modello addestrato per ottenere il valore di Y, così:

```
x_target = pd.DataFrame(
    {"MOL%" : [20.67],
     "R.O.S. (Return on Sales)" : [1.20],
     "Rapporto_debiti/Ricavi_Riproporz": [0.91]
    })

x_target #SONO I VALORI MEDI ITALIANI
```

	MOL%	R.O.S. (Return on Sales)	Rapporto_debiti/Ricavi_Riproporz
0	20.67	1.2	0.91

```
y_pred_ROI = lr.predict(x_target)
y_pred_ROI
```

```
array([2.69874072])
```

2.70 è il valore di ROI di una ipotetica farmacia situata in Piemonte che ha i valori dei parametri espressi nel dataframe in base al modello addestrato.

Ovviamente più dati abbiamo e più essi sono correlati migliore è l'affidabilità della predizione, ma non è mai una certezza, tuttalpiù la più realistica ipotesi possibile.