ANALISI GERMAN CREDIT DATASET

Nicola Santillo

12 Gennaio 2023

1 Introduzione

Lo sviluppo dei consumi finanziari ha portato ad un aumento della domanda di finanziamenti. Tutte le volte che una banca riceve una richiesta di questo genere attiva un processo di *valutazione del merito creditizio.* Il suo scopo è quello di predire se il consumatore sarà capace di adempiere al proprio impegno di restituzione del prestito. In una fase preliminare, si costruisce la *storia creditizia* del richiedente unendo la sua documentazione anagrafica e reddituale con quella presente nelle banche dati. Tramite queste ultime gli istituti riescono a individuare se ci sono crediti pendenti o se ci sono stati ritardi negli ultimi pagamenti, delineando due profili: *cattivo pagatore* (se l'ipotesi precedente è vera) o *buon pagatore*. L'obiettivo è quello di ridurre al minimo l'emissione di prestiti rischiosi, massimizzando le possibilità di trarre profitto da buoni prestiti.

2 Dati

La valuta nel dataset è DM = Deutsch Mark, 1 EUR=1.95583 DEM.

Il dataset *German Credit* riporta 1000 osservazioni e 20 variabili, la 21-esima è il giudizio finale della banca (1= Good , 2= Bad).
I riscontri, di cui la banca ha tenuto conto, sono i seguenti:

- Status of existing checking account: Indica lo stato del conto corrente (CA)
    - A11 : CA < 0 DM
    - A12 : 0 <= CA < 200 DM
    - A13 : CA >= 200 DM / salary assignments for at least 1 year
    - A14 : no checking account

- Duration in month: Durata in mesi.
- Credit history: Indica la storia creditizia
    - A30 : no credits taken/ all credits paid back duly
    - A31 : all credits at this bank paid back duly,
    - A32 : existing credits paid back duly till now
    - A33 : delay in paying off in the past,
    - A34 : critical account/ other credits existing (not at this bank)

- Purpose: Scopo del credito
    - A40 : car (new)
    - A41 : car (used)
    - A42 : furniture/equipment
    - A43 : radio/television
    - A44 : domestic appliances
    - A45 : repairs
    - A46 : education
    - A47 : vacation
    - A48 : retraining

- Amount: Importo del credito

- Savings account/bonds: Indica il conto di risparmio/obbligazioni (SA)
    - A61 : SA < 100 DM
    - A62 : 100 <= SA < 500 DM
    - A63 : 500 <= SA < 1000 DM
    - A64 : SA >= 1000 DM
    - A65 : unknown/ no savings account

- Present employment since: Indica la durata dell'attuale occupazione (PE)
    - A71 : unemployed
    - A72 : PE < 1 year
    - A73 : 1 <= PE < 4 years
    - A74 : 4 <= PE < 7 years
    - A75 : PE >= 7 years
- Installment rate in percentage of disposable income: tasso di rata in percentuale

- Personal status and sex: Indica lo stato personale e il sesso
    - A91 : male : divorced/separated
    - A92 : female : divorced/separated/married
    - A93 : male: single
    - A95 : female : single

- Other debtors / guarantors: Indica la presenza di altri debitori
    - A101 : none
    - A102 : co-applicant
    - A103 : guarantor

- Present residence since: Indica da quanto tempo si è in quell'abitazione
- Property: Indica la proprietà di maggior valore del cliente
    - A121 : real estate
    - A122 : if not A121 : building society savings agreement/life insurance
    - A123 : if not A121/A122 : car or other, not in attribute 6
    - A124 : unknown / no property

- Age in years: Indica l'età
- Other installment plans: Indica se ci sono altri piani di rateizzazione e verso chi
    - A141 : bank
    - A142 : stores
    - A143 : none

- Housing: Indica in che modo è stata avuta l'abitazione
    - A151 : rent
    - A152 : own
    - A153 : for free

- Number of existing credits at this bank: Indica numero di crediti esistenti presso questa banca.
- Job: indica lo stato occupazionale
    - A171 : unemployed/ unskilled - non-resident

    - A172 : unskilled - resident
    - A173 : skilled employee / official
    - A174 : management/ self-employed/ highly qualified employee/ officer

- Number of people being liable to provide maintenance: Numero di persone tenute a provvedere al mantenimento
- Telephone: Indica se il cliente ha il numero di telefono registrato presso la banca
    - A191 : none
    - A192 : yes, registered under the customers name

- Foreign Worker: Indica se il cliente è un lavoratore straniero o no
    - A201 : yes
    - A202 : no

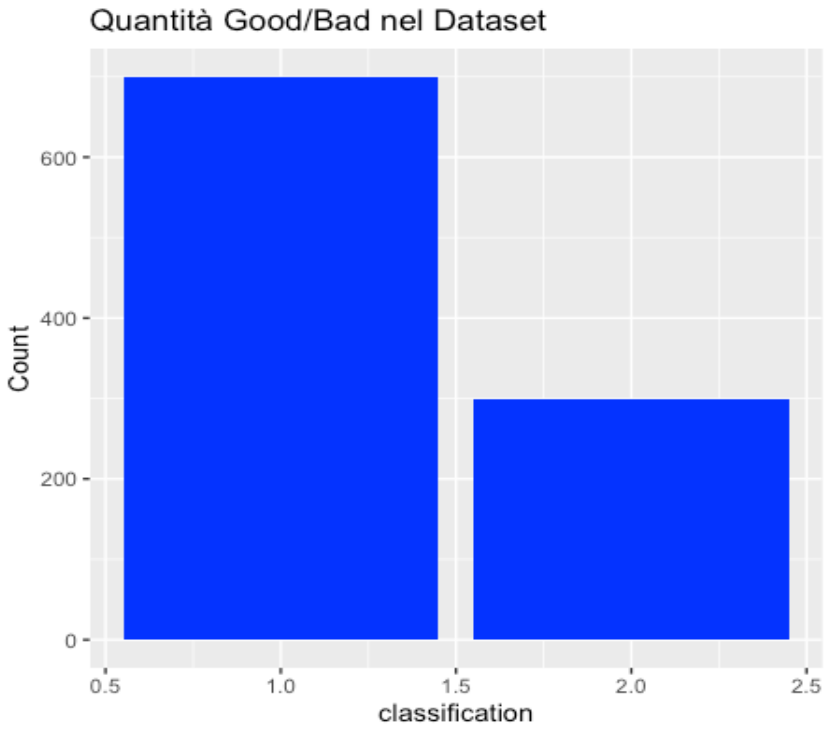
- Response: Indica la decisione della banca sulla richiesta di prestito da parte del cliente.

3 Analisi Descrittiva

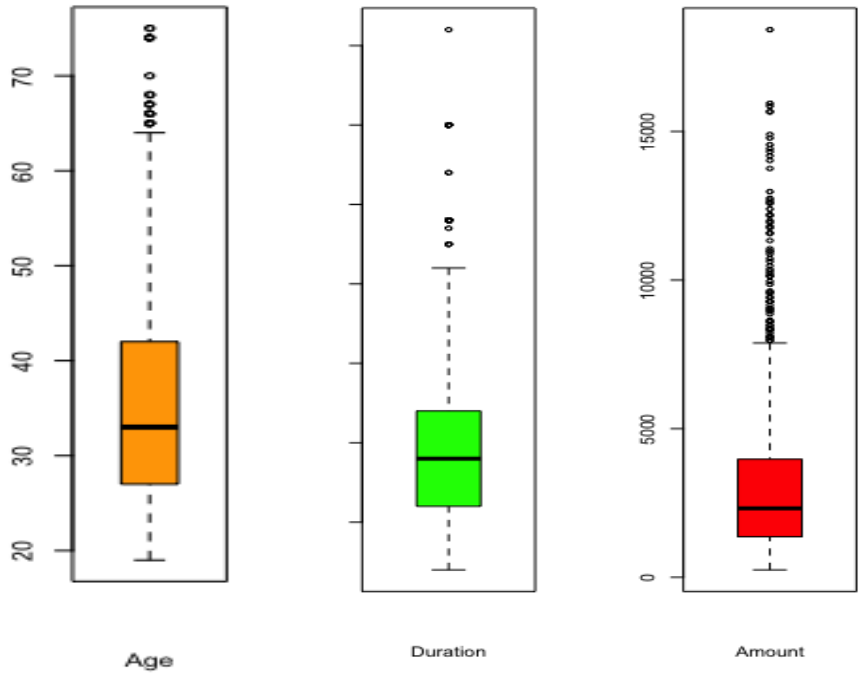
Il dataset è composto da 7 variabili quantitative e 13 variabili qualitative:

- *Variabili Quantitative:* Amount, Duration, Installment Rate, Present Residence, Age, Number of existing credits, Number of People being liable to provide maintenance.
- *Variabili Qualitative:* Status of existing checking account, Credit history, Purpose, Savings account/bonds, Present Employment, Sex, Other Debtor, Property, Other Installment Plans, Housing, Job, Telephone, Foreign Worker.

Prima di iniziare, viene riportato il numero di valori Good (1) e il numero di valori Bad(2) presenti nel dataset. Attraverso questo grafico potremo capire se la Banca abbia avuto un comportamento rigido o meno nei confronti dei 1000 clienti.

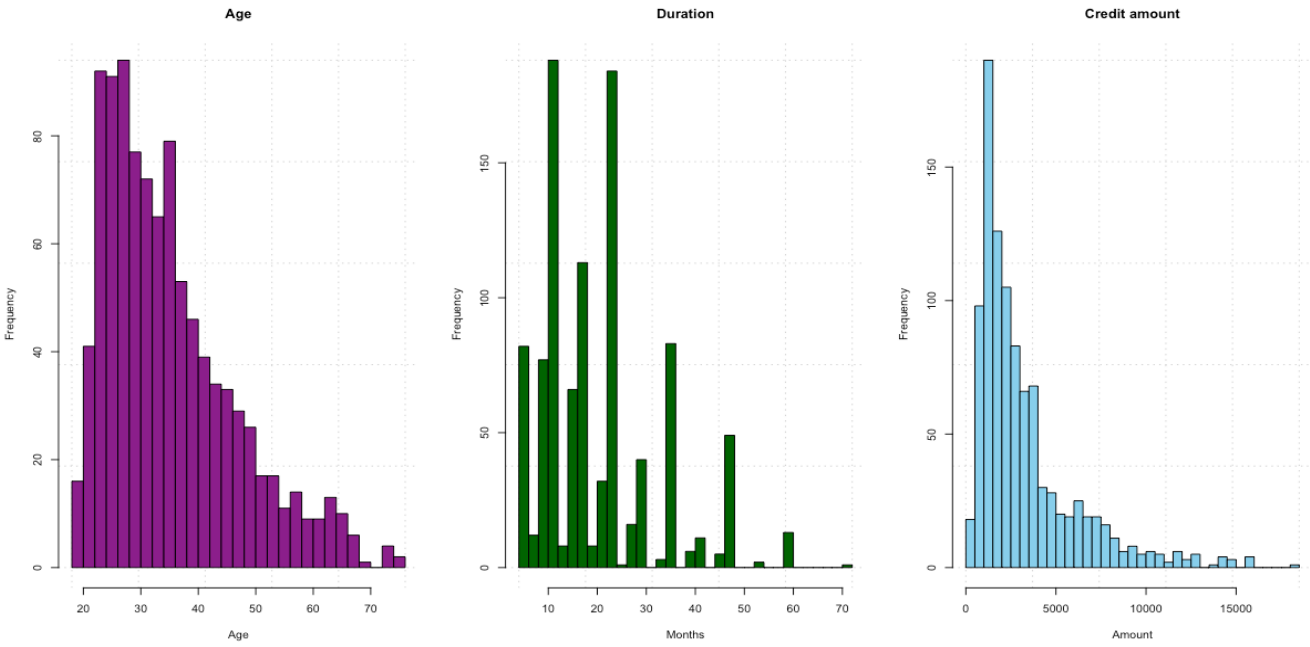


I valori indicano come ci siano stati per la maggior parte dei pareri positivi rispetto a quelli negativi. Ci aspettiamo pertanto, che nella maggior parte dei casi, la banca accetti le richieste.
L'esplorazione dei dati è iniziata dalle variabili quantitative ritenute più significative, rappresentate mediante boxplot.



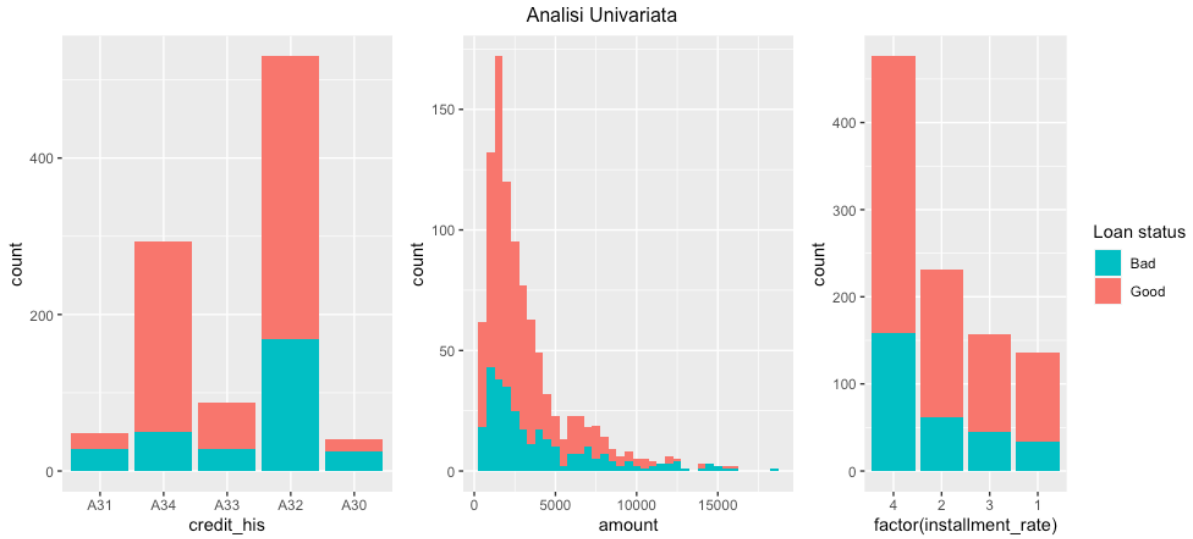
Age, Duration ed Amount hanno una distribuzione positivamente asimmetrica. Ciò significa che la Banca deve aspettarsi richieste di piccoli finanziamenti da clienti giovani con una durata del credito di circa 0-25 mesi.

Successivamente le tre variabili sono state rappresentate mediante istogramma per avere una conferma di quanto detto precedentemente.

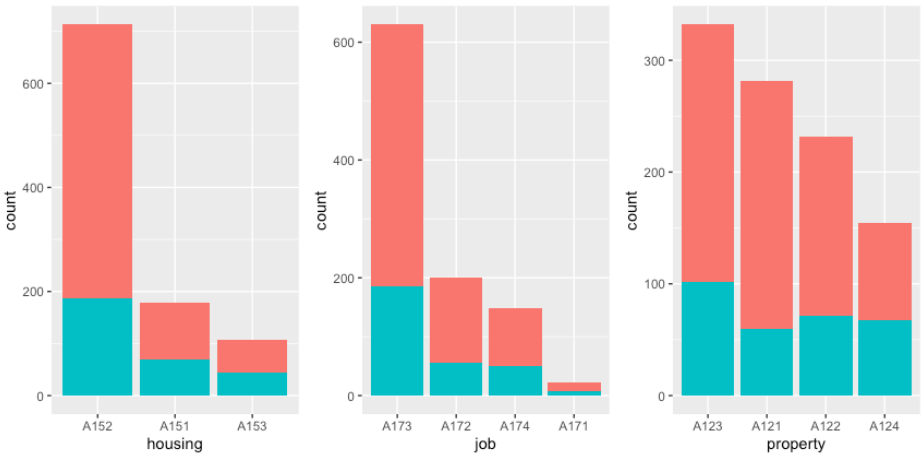


Partendo dall'età possiamo notare come tra i 20 e i 40 anni, con un picco sui 30, ci sia la maggior richiesta di finanziamenti. Infatti in questa fascia, i giovani stanno diventando adulti e pertanto stanno mettendo le fondamenta per un futuro solido. Per farlo c'è bisogno di molta liquidità, che molto spesso non hanno a disposizione. Curioso è il fatto che l'ammontare del prestito sia tra i 0 e i 5000 (non una cifra altissima) e con un rientro previsto, per la maggior parte, entro i 30 mesi.

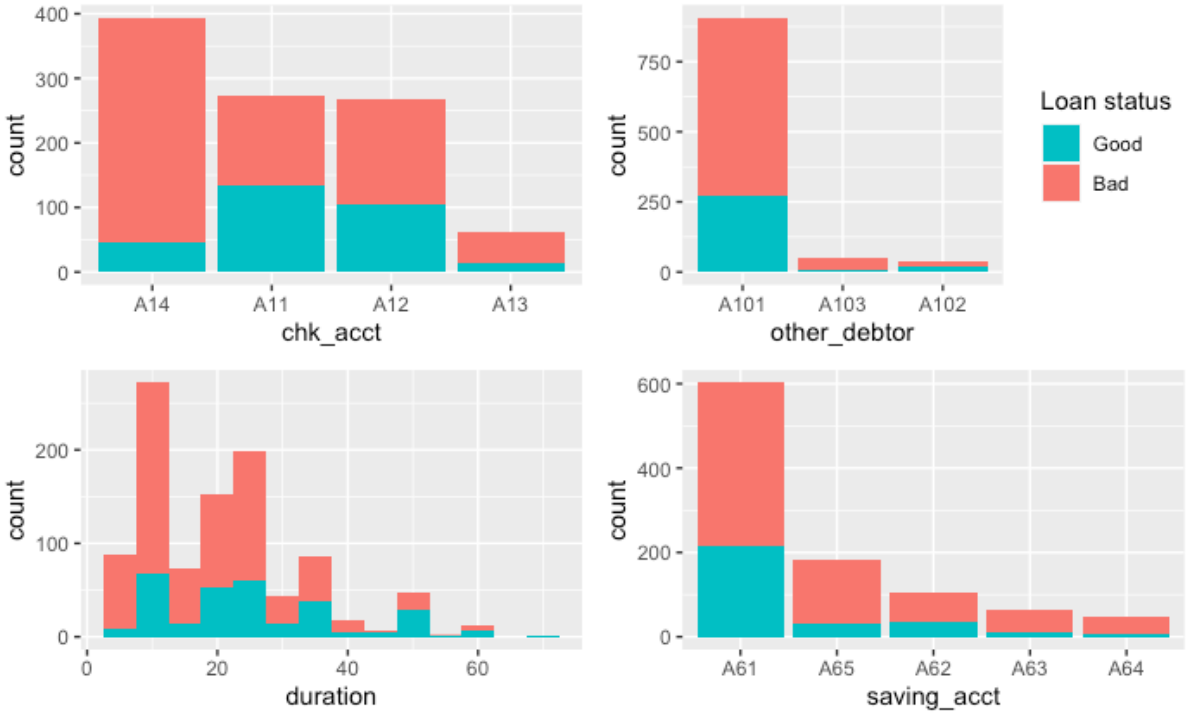
Andiamo a vedere che relazione c'è tra le variabili, ritenute interessanti, e la variabile response.



Se i clienti hanno già richiesto crediti, nella maggior parte dei casi la banca ha accettato la loro richiesta. In particolare, il numero maggiore di richieste proviene da clienti hanno già un credito, che stanno pagando regolarmente(A32).
Il prestito viene accettato specialmente per valori molto bassi. Tra 0 e 5000 la probabilità che il prestito venga accettato è molto alta. Anche se con poche osservazioni, più la richiesta aumenta più la risposta è negativa.
Rate per restituire il credito: Per la maggior parte di osservazioni le rate sono tra 1 e 4. Senza alcuna particolarità, ossia tutte le 'tranche' vengono accettate nella maggior parte dei casi.



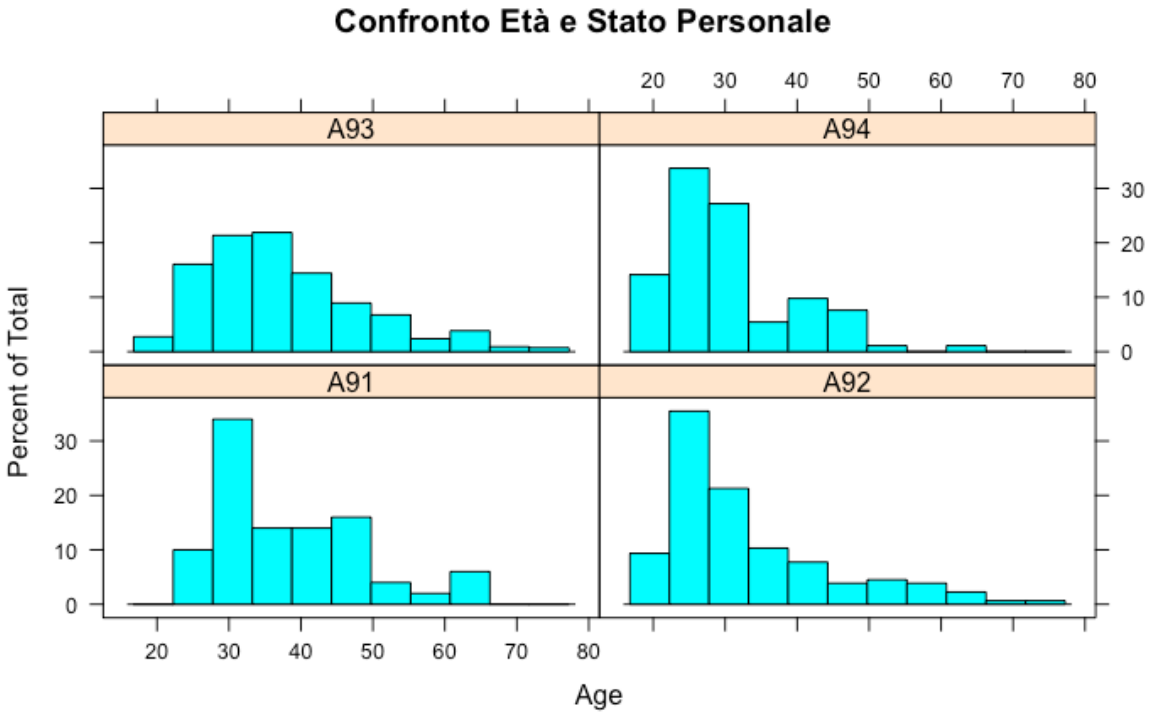
Housing: La casa è di proprietà(A151). I prestiti hanno avuto una buona risposta in relazione alle osservazioni ottenute. Solamente un terzo sono state ritenute bad. Alloggio gratuito (A153), in questo caso le risposte potrebbero dipendere da altri fattori. In quanto non c'è una netta differenza
Casa in affitto (A152). rispetto a A153 c'è qualche osservazione in più ritenuta good, ma anche qui altri fattori potrebbero essere essenziali per avere una buona proposta.
Property: Le proprietà del cliente non cambiano di molto la decisione dell'Istituto. Solamente se non si sanno/o non ci sono (A124), la decisione è a metà. Pertanto, in questo caso, serviranno altri riscontri.
Per quanto riguarda *Job* sottolineiamo solamente A171 ossia lo stato di disoccupato. Al netto delle poche osservazioni, la banca ha accettato più richieste di quante rifiutate. Probabilmente grazie a garanti o proprietà possedute.



Conto correnti (chk_acct): Se il richiedente ha meno di 0 DM (A11) le osservazioni del nostro dataset ci dicono che la percentuale di bad è maggiore rispetto agli altri. A partire da chi ha 0 questa percentuale diminuisce e da 200 in poi la risposta è nella maggior parte dei casi positiva. Chi non ha un conto corrente con la Banca (A14) ha nella maggior parte dei casi una valutazione positiva.
Buoni risparmi (saving_acct): Se <100 DM (A61) molto spesso le richieste vengono accolte, ma ci sono anche casi (un terzo) in cui la richiesta è stata rifiutata.
Per il resto dei casi la risposta è stata per lo più positiva, andando avanti con l'aumentare del numero di bonds.
Durata del Prestito (duration): Il prestito, in mesi, viene accettato se è tra gli 0 e 20 mesi. Andando avanti ci sono sempre meno richieste ma anche meno accettate. Magari il soggetto che richiede una durata così lunga non ha i requisiti giusti per la banca (proprietà, conto corrente ecc)

Crediti verso altri debitori (other_debtor): Se non esistono altri debitori (A101) è facile che venga accettato. Se il richiedente è affiancato da co-richiedente(A102) o da garante(A103), allora, anche se le osservazioni sono poche, vengono spesso accettate quelle con garante, mentre con co-richiedente la probabilità è più bassa.
Cos'è che pesa sulla risposta della banca? Attraverso queste domande, metteremo in relazione diverse variabili al fine di poter tirare fuori alcune conclusioni.

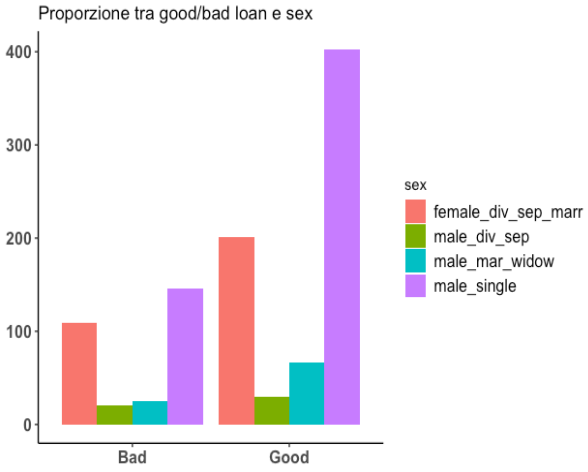
Stato personale e Età. *Che tipo di clienti hanno richiesto il prestito alla banca? E a che età?*



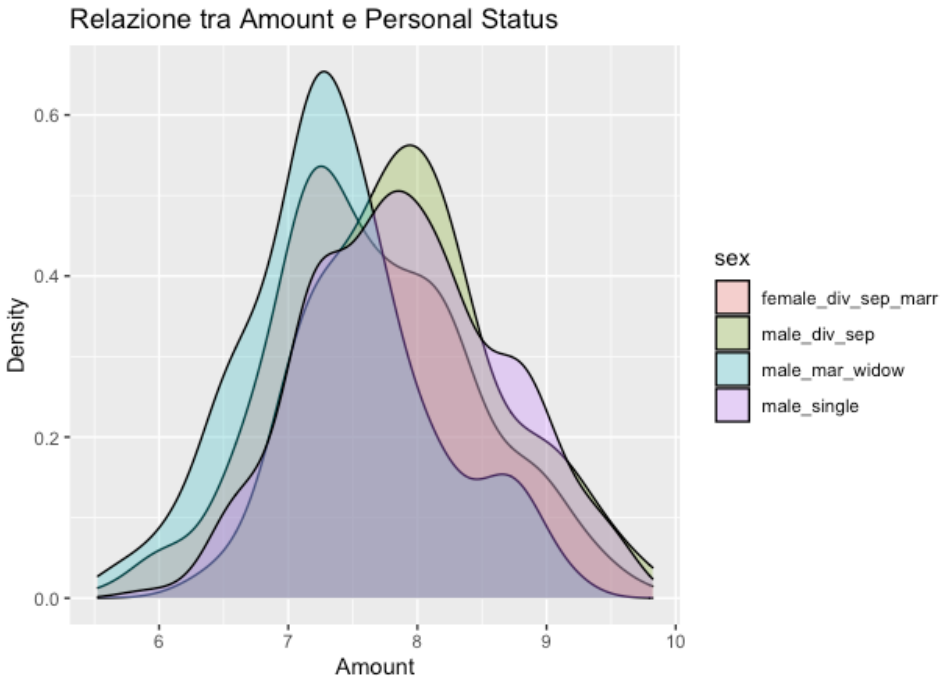
Le richieste di finanziamenti arrivano da 4 gruppi di persone: Maschi divorziati/separati(A91), Femmine divorziate/sposate/separate (A92), Maschi single (A93), Maschi sposati/vedovi (A94). Un grande numero di richieste, in relazione al numero di osservazioni, arrivano da donne non single intorno ai 30 anni, dagli uomini divorziati o separati o sposati/vedovi sempre nello stesso range di età.
Quindi un gran numero di persone giovani chiedono finanziamenti alla Banca.

Personal Status e Response. *Quanto incide lo stato di una persona sulla risposta?*

In realtà, molto poco. Infatti possiamo notare come c'è una netta differenza tra quelli accettati e non. L'unica non così netta è per i maschi che sono divorziati o separati.

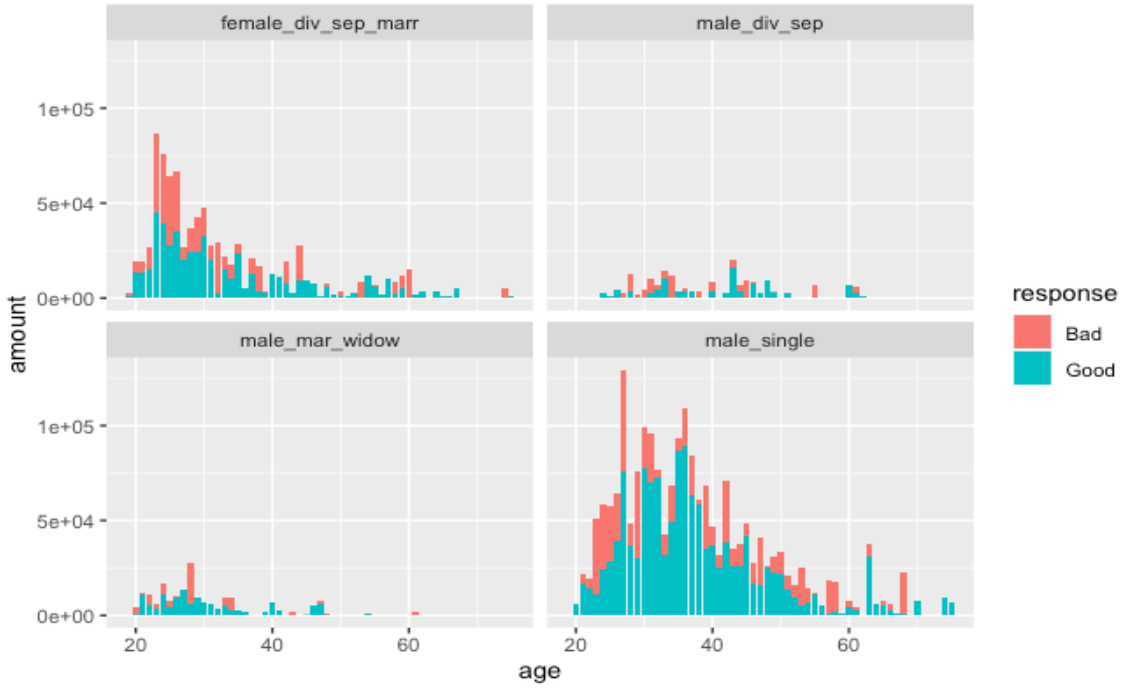


Chi è che chiede più prestiti? E chi richiede una cifra più alta?



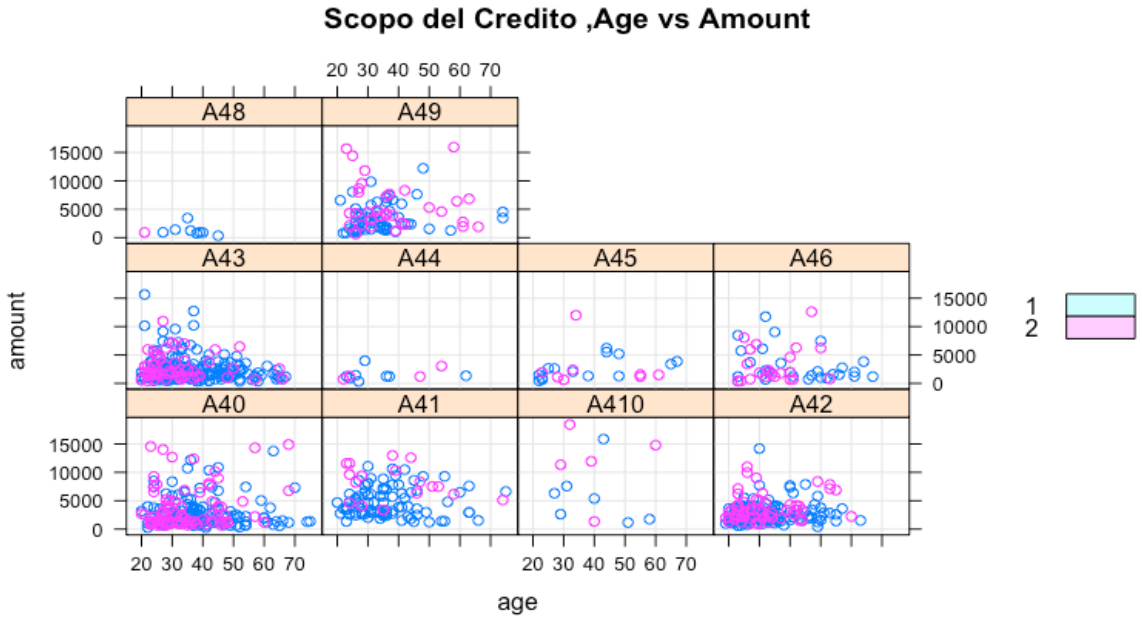
Da questo grafico a densità possiamo notare che le maggiori richieste derivano da Maschi Sposati o Vedovi, ma non hanno richieste esigenti. Dovute, per esempio, a poca liquidità e spese imminenti. Mentre chi richiede un prestito maggiore sono gli Uomini separati o divorziati.

Quale individuo ha più probabilità di avere una risposta positiva, in relazione all'età e alla quantità di denaro richiesta?



Da questo grafico possiamo notare quanto essere giovani influisca su una risposta negativa delle Banca. Infatti sia per maschi single che per femmine divorziate o sposate è molto difficile farsi accettare una richiesta di finanziamento, se hanno tra i 20 e i 30 anni.

Qual'è il reale scopo dei prestiti, e come vengono valutati?



La maggior parte dei prestiti vengono richiesti per comprarsi una nuova auto, per arredamento, televisione o radio, o per business (A40, A42,A43, A49). I richiedenti sono principalmente giovani.

4 Selezioni Delle Variabili

Lo scopo della selezione delle variabili è quello di identificare il miglior sottoinsieme di predittori tra molte variabili da includere in un modello. Dobbiamo quindi trovare le variabili necessarie tra l'insieme completo di variabili eliminando sia le variabili irrilevanti (variabili che non influenzano la variabile dipendente), sia le variabili ridondanti (variabili che non aggiungono nulla alla variabile dipendente). Tutto ciò al fine di ricavare il modello più parsimonioso possibile.

Sono state prese tutte le variabili quantitative del dataset, e su di loro sono stati fatti girare 2 modelli: *vscc e clustvarsel (anche backward).*

Le variabili selezionate sono:

- vscc: *n_people , n_credits, amount ,age ,present_resid ,installment_rate*
- clustvarsel : *n_credits, installment_rate ,duration,amount, age*
- clustvarsel (direction: backward): *duration, amount, installment_rate, age, n_credits*

Successivamente sono stati creati 3 dataset, ognuno costituito dalle variabili selezionate da ogni modello, e confrontate attraverso l'*adjustedRandIndex.*
L'indice più grande, è quello del metodo clustvarsel. Pertanto le variabili selezionate sono: *n_credits, installment_rate, duration, amount, age*.

Da ora in avanti, il dataset che verrà preso in considerazione sarà senza le variabili n_people (Number of people being liable for maintenance) e pres_resid (Present Residence)

5 Cluster Analysis

La Cluster Analysis ha l'obiettivo di trovare dei gruppi/classi (cluster) che siano i più simili tra di loro. La peculiarità di questa analisi è quella di avere dei gruppi in cui la *varianza within* sia la più piccola possibile e la *varianza between* la più grande. Quindi avremo classi il più possibili omogenee all'interno e il più differenti possibile tra di loro.

Per l'analisi del German Credit Dataset sono state utilizzate solo alcune variabili quantitative: Duratio, Age, Credit Amount. Successivamente il nuovo set di dati è stato scalato, a cui sono stati applicati diversi metodi di clustering.
Per quanto riguarda i *metodi gerarchici,* dove si parte da un gruppo per ogni unità fino a che ogni unità si trovi in un solo gruppo, sono stati usati:

- Ward
- Ward con distanza di Minkowski
- Ward con distanza di Manhattan
- Single Linkage

- Complete Linkage

Osservazione: La *distanza* è lo scarto tra due unità definendo una matrice di dissimilarità/prossimità

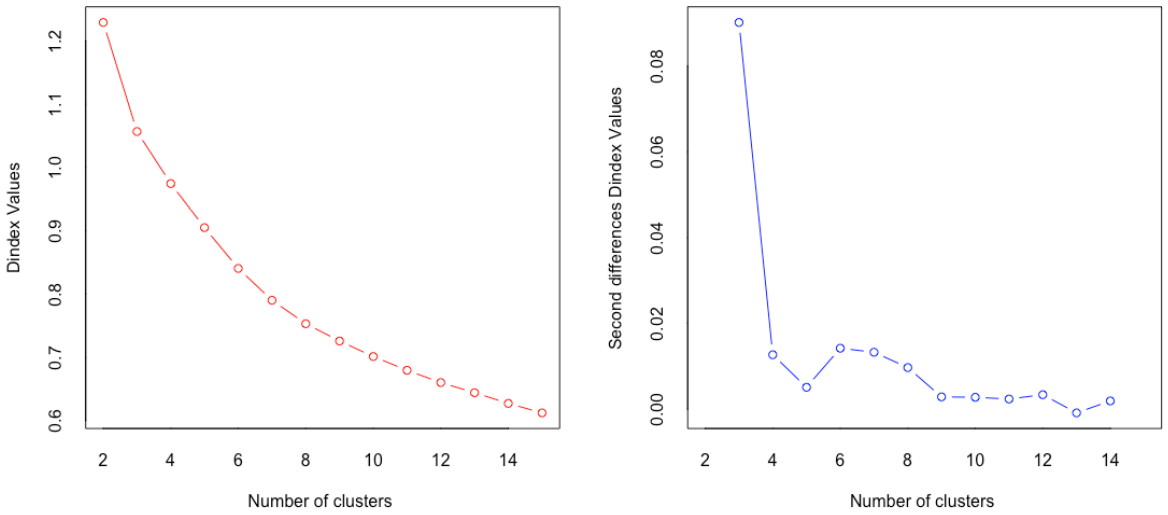
Per i *metodi a partizioni*, basati sulle distanze rispetto alle medie, è stato utilizzato il metodo K-means.

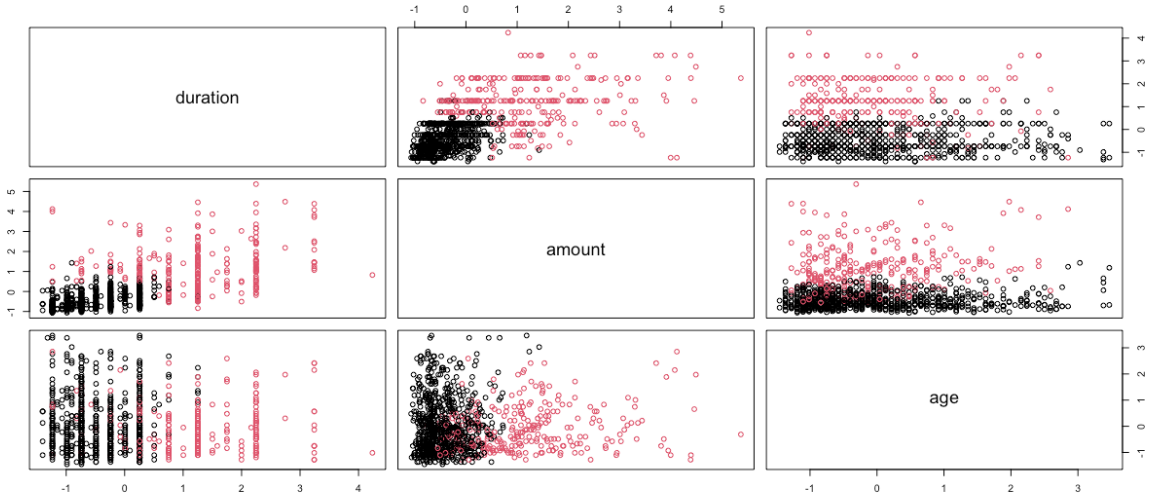
Di seguito, i risultati ottenuti:

Metodi Utilizzati	Indice di Calinski-Harabasz	Numero di Cluster
Ward	543.9664	2
Ward con distanza di Minkowski	543.9664	2
Ward con distanza di Manhattan	527.9957	3
Single Linkage	8.9553	2
Complete Linkage	481.5231	2

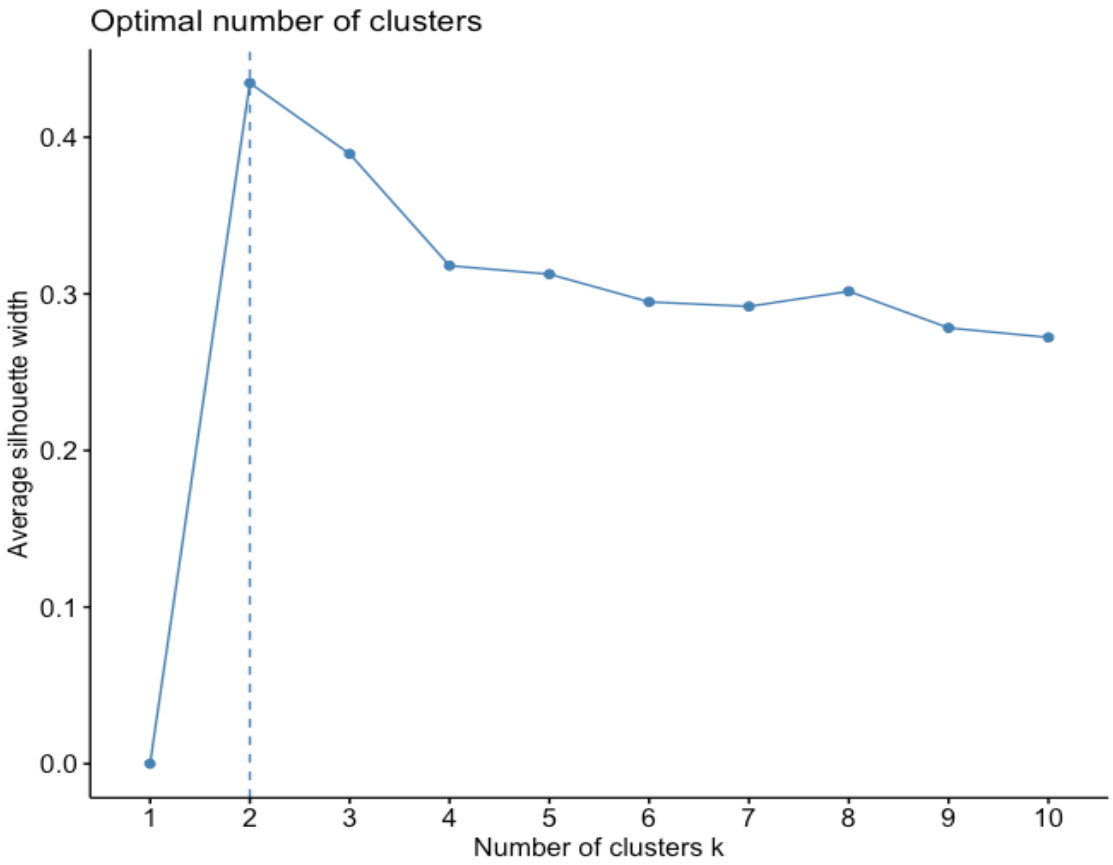
K-Means	508.5921	3

Dagli esiti ricavati vengono confrontati gli indici di Calinski-Harabasz, che si basano sul rapporto tra la varianza within e la varianza between.
Il valore più alto, 543.9664, corrisponde al metodo Ward che indica 2 cluster.

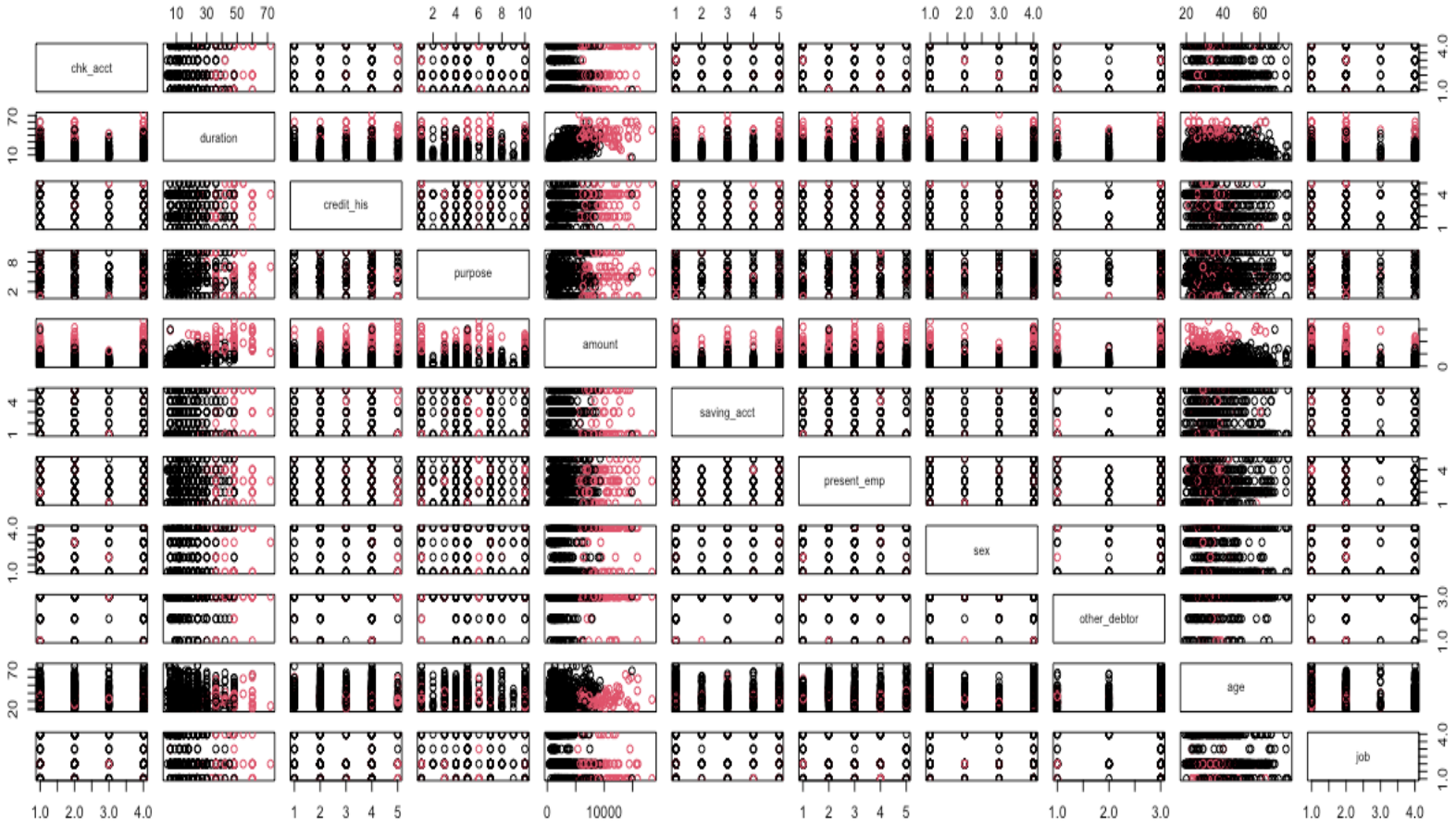




E’ stata fatta un’ultima verifica, per accertare il corretto numero di cluster, utilizzando il Metodo della Silhouette. Anche in questo caso il risultato è di 2 cluster.



Il risultato della divisione in 2 gruppi del nostro dataset è il seguente

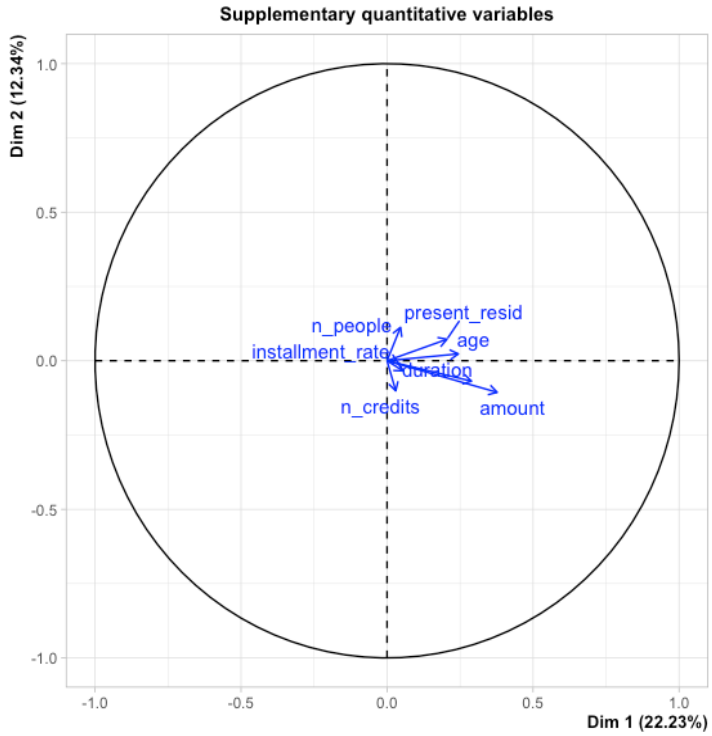


Dal grafico possiamo quindi dedurre che all'aumentare della quantità di denaro richiesta alla Banca, aumentano anche le rate del finanziamento. Inoltre che più il cliente è anziano e più il denaro richiesto è basso.

6 Multiple Correspondence Analysis (MCA)

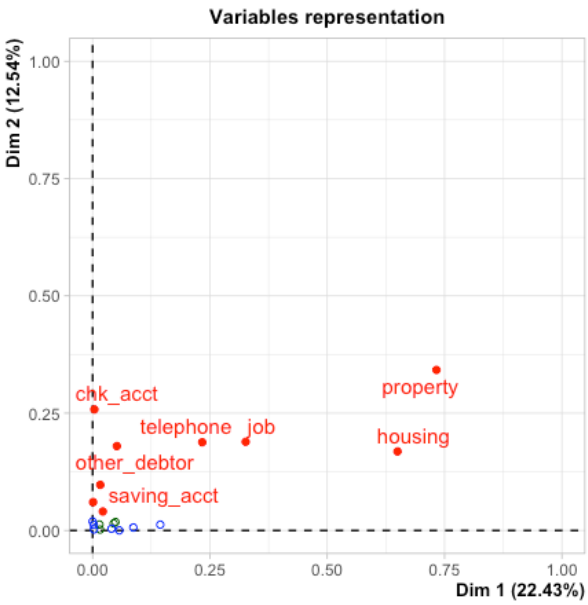
Tramite questa analisi esploreremo potenziali relazioni che ci sono nei dati tra gli individui, per identificare individui con profili simili e relazioni che ci possono essere tra le variabili qualitative. Pertanto si analizzerà se c'è una relazione fra le categorie delle diverse variabili qualitative andando a misurare un'eventuale associazione tra le variabili, facendo riferimento alle categorie.

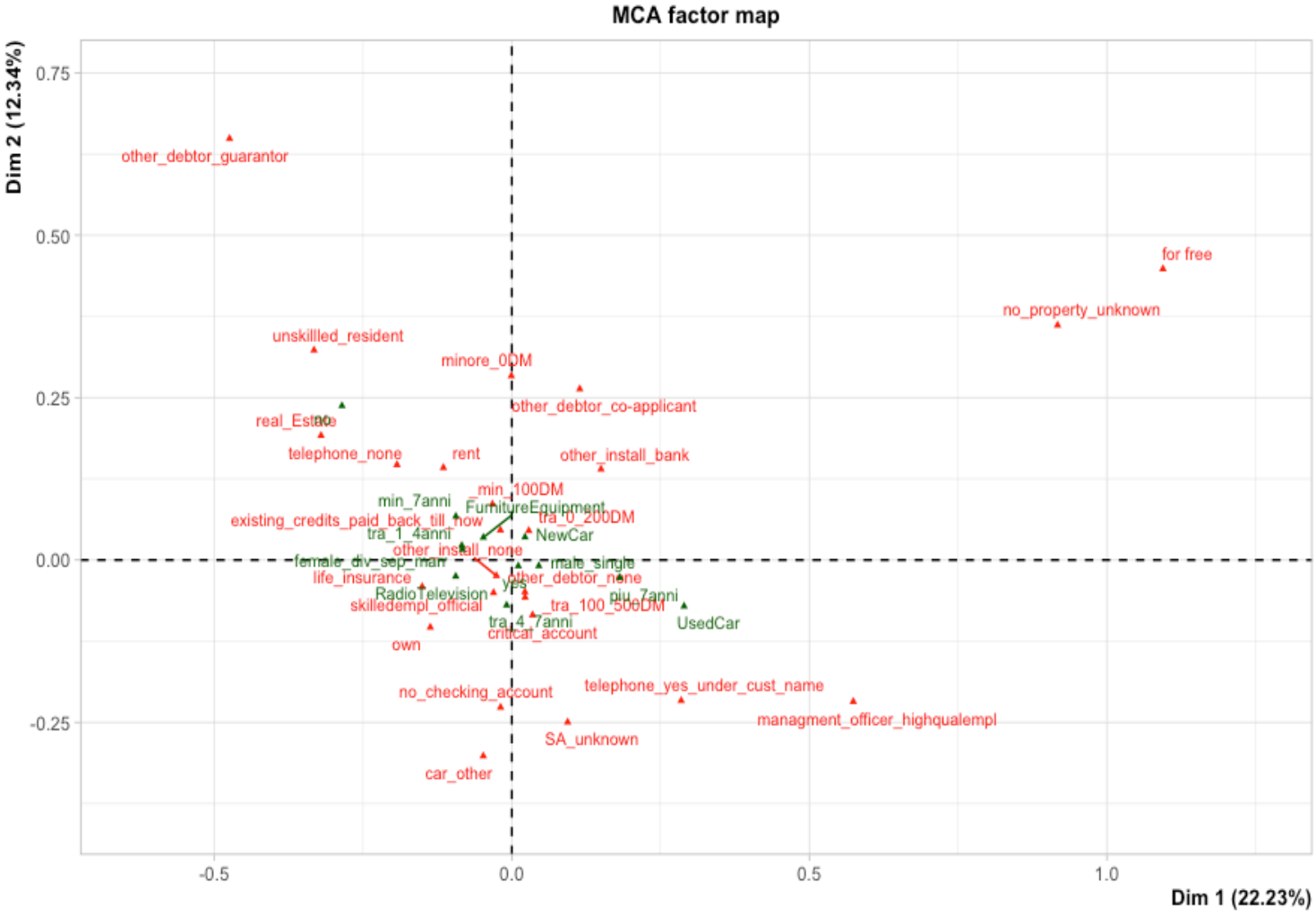
E' stata utilizzata la funzione *MCA* della libreria FactoMineR, dove abbiamo inserito al suo interno il Dataset senza la colonna della variabile response. Come risposta abbiamo ricevuto i seguenti grafici :



Questo grafico ci mostra le relazioni che ci sono tra tutte le variabili quantitative supplementari del nostro dataset. Possiamo notare come tutte siano raggruppate insieme, pertanto sono correlate positivamente tra loro. Inoltre amount è quella più lontana dall'origine e quindi quella meglio rappresentata.

Qui invece troviamo le variabili che più spiegano le dimensioni. Infatti, la *dimensione 1* è spiegata maggiormente dalle variabili *property* e *housing*. Per quanto riguarda la *dimensione 2* la varianza è spiegata da chk_acct (cheking account) e dal resto del gruppo in basso a sinistra.

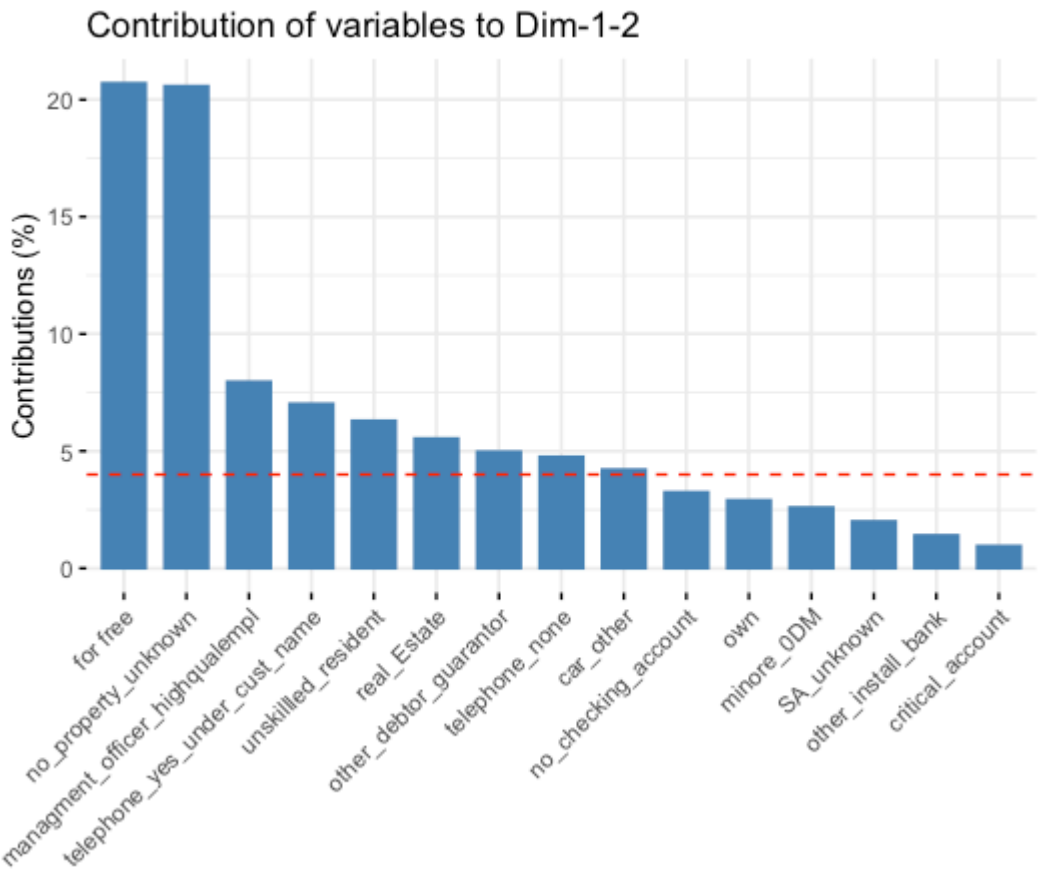


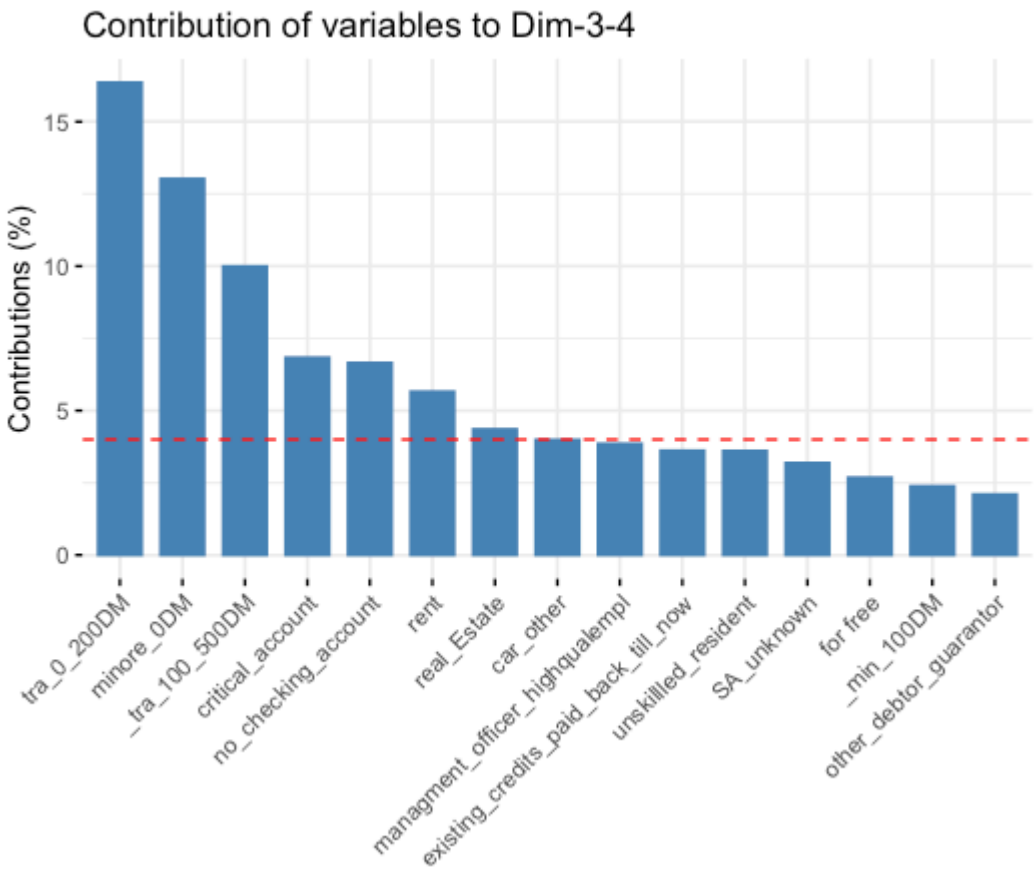


Da questa rappresentazione siamo in grado di conoscere che tipi di clienti abbiano richiesto finanziamenti alla Banca, in modo da poter stilare più profili. Così facendo l'Istituto sarà a conoscenza dei motivi dietro alla richiesta di prestito, e se da questo tipo di clientela può trarre profitto dal prestito erogato.

- Chi ha un lavoro da più di 7 anni, è di sesso Maschile ed è single. E' un lavoratore straniero che non ha altri finanziamenti attivi né presso la stessa banca né verso altre. Possiede anche buoni risparmi fino a 200DM e la sua richiesta di finanziamento è per comprarsi una nuova auto.

- Un lavoratore inesperto, che ha un'assicurazione sulla vita e una casa di proprietà richiede finanziamenti per comprarsi televisione/radio.
- Chi ha iniziato a lavorare tra gli 1 e i 4 anni, è di sesso Femminile è può essere divorziata/separata o single. Ha già crediti esistenti verso la banca, ma che finora sono stati pagati regolarmente. La sua richiesta di finanziamento è per l'acquisto di mobili o arredi per la casa. Probabilmente si tratta di una famiglia trasferitasi da altre parti, che ha chiesto un secondo prestito dopo quello per aver comprato casa.
- Coloro che sono stati definiti cattivi pagatori dalla banca (critical account), lavorano da almeno 4 anni e possiedono buoni risparmio tra i 100 e 500DM. La decisione della banca potrebbe essere stata presa prima di iniziare a lavorare, magari dovuta ad una situazione debitoria pregressa.
- Chi necessita di un garante o di un 'co-applicant' , è un soggetto che ha già altri crediti verso la Banca. Il suo conto corrente è pari a 0DM.





Andando a studiare in dettaglio il contributo delle variabili alle dimensioni, notiamo come l'apporto maggiore nella *dimensione 1:2* viene dato da: *for_free, no_property_unknown*. Ciò il profilo di una persona che vive in una casa gratis e che non ha altre proprietà intestate o sconosciute alla banca.
Per quanto riguarda la *dimensione 3:4* invece l'apporto maggiore viene da: *tra_0_200DM e minore_0DM*, ossia chi ha nel conto corrente una cifra minore di 200DM.

7 Conclusioni

Alla fine di questa breve analisi, possiamo concludere che la Banca valuta più negativamente le richieste di finanziamento di persone di età tra i 20 e i 30 anni o che hanno poco/valori negativi sul conto corrente. Pertanto sarà difficile, per maschi single, femmine divorziate/sposate/separate o lavoratori inesperti ricevere il credito richiesto se rientrano in una di queste categorie. Invece i clienti anziani sono valutati più positivamente anche perché mediamente richiedono prestiti più bassi.