**Analisi Esplorativa del Mercato Immobiliare del Texas**

**Parte 1**

1. Ho importato il dataframe con la funzione “read.csv”, poi ho usato “attach” per riferirmi al nome delle colonne senza la notazione “nome\_dataframe$nome\_colonna”.
2. Tipo variabili:
   * 1. city: qualitativa nominale.
     2. year: qualitativa ordinale.
     3. month: qualitativa ordinale.
     4. sales: quantitativa discreta, scala rapporti.
     5. volume: quantitativa continua, scala rapporti.
     6. median\_price: quantitativa continua, scala rapporti.
     7. listings: quantitativa discreta, scala rapporti.
     8. month\_inventory: quantitativa discreta, scala rapporti.
3. Indici:
   * 1. Posizione: per le variabili qualitative city, year e month è stata calcolata solo la moda in quanto unico indice che ha senso calcolare.  
        Moda city: distribuzione equimodale. Valore della fequenza è 60.  
        Moda year: distribuzione equimodale. Valore della fequenza è 48.  
        Moda month: distribuzione equimodale. Valore della fequenza è 20.  
          
        Per le altre variabili, sono stati calcolati tutti gli indici di posizione, essendo quantitative per il calcolo della moda è stata calcolata prima la distribuzione in frequenza.  
        Inoltre si è sempre calcolato sia la media aritmetica che quella ponderata per mostrare le differenze dei due metodi.   
        Per variabili quantitative sales, volume, median\_price, listings, month\_inventory gli indici di posizione sono calcolati nel seguente modo(esempio per variabile sales, le altre seguono la stessa struttura):  
          
         1) divisione in classi tramite la funzione:  
         cut(sales, seq(min(sales), max(sales), (max(sales)-min(sales))/10 ))  
         come si nota le classi sono 10.  
          
         2) calcolo delle frequenze:  
         ni = table(sales\_div\_classi)

fi = table(sales\_div\_classi) / length(sales)

Ni = cumsum(table(sales\_div\_classi))

Fi = cumsum(table(sales\_div\_classi)) / length(sales)  
  
 3) costruzione della tabella di contingenza:  
 as.data.frame( cbind( ni, fi, Ni, Fi))  
  
 4) calcolo della moda, vengono proposti due metodi:  
 primo: table(sales\_div\_classi)

max(table(sales\_div\_classi))

secondo: max(sales\_distr\_freq["ni"])   
  
 5) calcolo mediana, vengono proposti due metodi:  
 primo: median(sort(sales))

secondo: x = sales\_distr\_freq["Fi"][[1]]

index = which(x >= 0.50)[1]

rownames(sales\_distr\_freq)[index]  
  
 6) calcolo quantili(min e max):

quantile(sort(sales))[c(1, 5)]  
  
 7) media:  
 (vengono calcolate sia l aritmetica che la ponderata, solo a scopo illustrativo)   
 aritmetica: mean(sales)

ponderata: seq(min(sales), max(sales), (max(sales)-min(sales))/10 ) + (((max(sales)-min(sales))/10) /2)

valori = valori[-11]

pesi = sales\_distr\_freq["ni"][[1]]

weighted.mean(valori, pesi)

Per vedere i risultati consultare direttamente il codice.

* + 1. Variabilità: per le variabili qualitative è stato calcolato solo l’indice di Gini con la funzione “indice\_gini” che si trova nel file Utils.R.  
       In particolare per queste 3 variabili l’indice di Gini sarà uguale a 1 in quanto sono equimodali.  
       Per le altre variabili oltre all’indice di Gini sono stati calcolati anche tutti gli altri indici di variabilità tramite la funzione “indici\_di\_variabilità”.  
       Come la moda, anche per il calcolo dell’indice di Gini è stata prima calcolata la distribuzione in frequenze quando la variabile è quantitativa.  
       Non ci sono particolari note da riportare per queste due funzioni, vedere direttamente il codice nel file Utils.R per maggiori dettagli e per i risultati ottenuti.
    2. Forma: gli indici di forma non sono stati calcolati per le variabili qualitative.  
       Per le restanti è stata usata la funzione “indici\_di\_forma” che consente di calcolare sia l’asimmetria che la curtosi di una distribuzione. Anch’essa è contenuta nel file Utils.R.

1. Qual è la variabile con variabilità più elevata? Come ci sei arrivato? E quale quella più asimmetrica?  
     
   La variabile con variabilità più elevata è il volume, conclusione fatta osservando il coefficiente di variazione delle diverse variabili. Ho utilizzato questo indice perchè permette di confrontare variabili provenienti da distribuzioni diverse.  
   La variabile più asimmetrica è ancora il volume. Per giungere a questa conclusione ho confrontato il valore assoluto dei vari indici di asimmetria delle variabili e prendendo quello più grande, ricordando che un indice di asimmetria pari a zero corrisponde a una simmetria perfetta, come quella della distribuzione normale.
2. Dividi una delle variabili quantitative in classi, scegli tu quale e come, costruisci la distribuzione di frequenze, il grafico a barre corrispondente e infine calcola l’indice di Gini.   
     
   La variabile scelta è il volume. Per ottenere la distribuzione in frequenze è stata prima usata la funzione “get\_distribuzione\_frequenze” del file Utils.R e dopo la funzione “barplot” per ottenere il grafico a barre.  
   Per l’indice di Gini vedere il punto 3.  
   Per maggiori dettagli vedere direttamente il codice.  
   Nota: ingrandire bene il grafico per vedere allineate in modo corretto il nome delle classi con le relative barre.
3. Indovina l’indice di gini per la variabile city.  
     
   l’indice di Gini è uguale a 1, infatti le classi hanno il massimo livello di omogeneità ovvero la distribuzione è equimodale.
4. Qual è la probabilità che presa una riga a caso di questo dataset essa riporti la città “Beaumont”? E la probabilità che riporti il mese di Luglio? E la probabilità che riporti il mese di dicembre 2012?  
     
   Essendo che ho a disposizione i dati posso usare l'approccio classico: numero di casi favorevoli / totale dei casi.  
   1. prob\_beaumont = table(city)["Beaumont"] / length(city)  
       prob\_beaumont = 0.25 infatti ho 4 valori possibili equiprobabili.
   2. prob\_luglio = table(month)[7] / length(month)

prob\_luglio = 0.0833 infatti ho 12 valori possibili equiprobabili.

* 1. prob\_dic\_2012 = sum( RealEstateTexax\_Dataframe[["month"]] == "12" & RealEstateTexax\_Dataframe[["year"]] == 2012 ) / nrow(RealEstateTexax\_Dataframe) prob\_dic\_2012 = 0.01666667, viene usata la probabilità risultante dall’intersezione di due eventi.

1. Esiste una colonna col prezzo mediano, creane una che indica invece il prezzo medio, utilizzando le altre variabili che hai a disposizione.  
     
   Il prezzo medio è stato ottenuto con la seguente istruzione:  
   RealEstateTexax\_Dataframe["mean\_price"] = volume/sales\*1000000
2. Prova a creare un’altra colonna che dia un’idea di “efficacia” degli annunci di vendita. Riesci a fare qualche considerazione?  
     
     
   L’indicatore creato è stato chiamato “Coefficiente di Conversione” e indica la percentuale di annunci che si sono convertiti in vendite:  
     
    RealEstateTexax\_Dataframe["conversion\_coefficient"] = sales/listings\*100  
     
   Dal quale è stato ricavato media e deviazione standard:  
     
    mean(coefficiente\_conversione)

sd(coefficiente\_conversione)  
  
Inoltre sono state fatte le seguenti osservazioni:

1) Si è riscontrato che la maggior efficacia degli annunci la si ha nel periodo estivo:  
  
dati\_raggruppati1 <- RealEstateTexax\_Dataframe[ c("month", "conversion\_coefficient")] %>%

group\_by(month) %>%

summarise(media\_conversione1 = mean(conversion\_coefficient, na.rm = TRUE))  
  
2) Si è riscontrato un trend positivo nell'efficacia degli annunci in base al tempo:

dati\_raggruppati2 <- RealEstateTexax\_Dataframe[ c("year", "conversion\_coefficient")] %>%

group\_by(year) %>%

summarise(media\_conversione2 = mean(conversion\_coefficient, na.rm = TRUE))

1. Prova a creare dei summary().  
     
   Oltre a quelli creati nel punto 9:  
     
   totale\_venduto\_città <- RealEstateTexax\_Dataframe %>%

group\_by(city) %>%

summarise(media\_fatturato = mean(volume, na.rm = TRUE))

totale\_venduto\_città  
  
Dalla quale si ricava la città con maggiore fatturato che è Tyler.

**Parte 2**

Nota: vengono qui riportate solo le considerazioni per non appesantire la lettura, per il codice guardare il file main2.R

1. Utilizza i boxplot per confrontare la distribuzione del prezzo mediano delle case tra le varie città. Commenta il risultato.  
     
   1.1) Le case con il median price più alto si trovano a Bryan-College Station e quelle con quello più basso a Wichita Falls.  
     
   1.2) Le case con i median price più variabili sembrano trovarsi a Wichita Falls, dato ricavato osservando la maggior ampiezza del range interquantile rispetto agli altri e anche dalla maggior lunghezza delle code, considerando anche gli outliers.  
   Al contrario, quelle con i median price meno variabili a Bryan-College Station.
2. Utilizza i boxplot o qualche variante per confrontare la distribuzione del valore totale delle vendite tra le varie città ma anche tra i vari anni. Qualche considerazione da fare?  
     
   2.1) il volume totale delle vendite a Wichita Falls è il più basso tra tutti e non presenta trend, si è infatti mosso all'interno di un range orizzontale. Anche la varianza è rimasta costante negli anni.

2.2) per le altre città invece si può notare un movimento direzionale, meno evidente per Beaumount, mentre molto in vista per le restanti 2.

2.3) Oltre ad avere un trend positivo per quanto riguarda il volume delle vendite , Bryan-College Station e Tyler presentano anche un trend positivo per quanto riguarda la variabilità, la quale infatti è aumentata al passare degli anni.

2.4) Dal grafico si può giungere quindi alla conclusione che negli ultimi anni l'interesse per l'acquisto di una casa a Bryan-College Station o a Tyler è aumentato in maniera significativa, leggermente a Beaumount e per nulla a Wichita falls.

1. Usa un grafico a barre sovrapposte per confrontare il totale delle vendite nei vari mesi, sempre considerando le città. Prova a commentare ciò che viene fuori. Già che ci sei prova anche il grafico a barre normalizzato. Consiglio: Stai attento alla differenza tra geom\_bar() e geom\_col(). PRO LEVEL: cerca un modo intelligente per inserire ANCHE la variabile Year allo stesso blocco di codice, senza però creare accrocchi nel grafico.

3.1) dal grafico a barre sovrapposte si può notare come le vendite nella città di Beaumount e Wichita Falls siano state più o meno costanti durante tutti i mesi.

3.2) mentre per la città di Bryan-College Station e (in maniera minore Tyler) sembra che le vendite siano maggiori durante i mesi primaverili ed estivi piuttosto che negli altri.

3.3) inoltre il numero totale di vendite è maggiore nella città di Tyler e a seguire in Bryan-College Station, Beaumount e Wichita Falls.

1. Prova a creare un line chart di una variabile a tua scelta per fare confronti commentati fra città e periodi storici.

4.1) Wichita Falls non ha subito incrementi nelle vendite negli anni e la sua variabilità è rimasta la stessa.

4.2) Beaumont ha visto un aumento nelle vendite col passare degli anni ma non della variabilita.

4.3) Al contratio invece Bryan-College Station, Tyler hanno visto, col passare degli anni, sia un aumento nella quantità delle vendite sia della loro stessa variabilità  
città con meno vendite (in media): Wichita Falls

città con più vendite (in media): Tyler