Analisi del mercato immobiliare del Texas

Svolgi i punti uno alla volta e produci un **documento di testo word, pdf, HTML o markdown** in cui, per ogni punto, posso visualizzarne il codice, l'output di R e il tuo commento, spiegando ciò che hai fatto e il ragionamento.

Puoi consegnare anche il file.R per sicurezza, ma non deve essere obbligatorio da leggere per me per capire cosa hai fatto.

NOTA BENE: questo non è un progetto di programmazione, ma di statistica, e mi aspetto di leggere commenti e considerazioni statistiche per i vari passaggi e risultati!

Non preoccuparti se non sai come risolvere qualche punto: provaci, riprovaci e male che vada consegna anche solo il codice col tentativo. Altre considerazioni e analisi fuori dagli schemi sono assolutamente ben accette!

Usa il manuale integrato di R utilizzando la funzione "help(nomefunzione)" oppure "?nomefunzione", per capire come usare tutti i vari argomenti… e non dimenticarti che Google è tuo amico, anzi, è pieno di amici!

Non c'è alcun limite imposto per l'utilizzo di R e la fantasia! Se sei uno smanettone come me, divertiti a cercare pacchetti e funzioni che possano agevolarti il lavoro.

Se non sei un principiante della programmazione, sbizzarrisciti pure!

Io ti lascio comunque qualche file utile e ti consiglio di darci almeno un'occhiata prima di partire 😉

Progetto:

Importa il dataset "Real Estate Texas.csv", contenente dei dati riguardanti le vendite di immobili in Texas. Le variabili del dataset sono:

- city: città
- year: anno di riferimento
- month: mese di riferimento
- sales: numero totale di vendite
- volume: valore totale delle vendite in milioni di dollari
- median_price: prezzo mediano di vendita in dollari
- listings: numero totale di annunci attivi
- months_inventory: quantità di tempo necessaria per vendere tutte le inserzioni correnti al ritmo attuale delle vendite, espresso in mesi.

1. Indica il tipo di variabili contenute nel dataset. (Attenzione alle variabili che sottintendono il tempo e a come vengono trattate!)
2. Calcola Indici di posizione, variabilità e forma per tutte le variabili per le quali ha senso farlo, per le altre crea una distribuzione di frequenza. Commenta tutto brevemente.
3. Qual è la variabile con variabilità più elevata? Come ci sei arrivato? E quale quella più asimmetrica?
4. 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.
5. "Indovina" l'indice di gini per la variabile city.
6. 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?
7. Esiste una colonna col prezzo mediano, creane una che indica invece il prezzo medio, utilizzando le altre variabili che hai a disposizione
8. Prova a creare un'altra colonna che dia un'idea di "efficacia" degli annunci di vendita. Riesci a fare qualche considerazione?
9. Prova a creare dei summary(), o semplicemente media e deviazione standard, di alcune variabili a tua scelta, condizionatamente alla città, agli anni e ai mesi. Puoi utilizzare il linguaggio R di base oppure essere un **vero Pro** con il pacchetto dplyr. Ti lascio un suggerimento in pseudocodice, oltre al cheatsheet nel materiale:

```
dati %>%

 group_by(una o più variabili di raggruppamento) %>%

 summarise(nomecolonna1=funzione1(variabile da sintetizzare),

                                  nomecolonna2=funzione2(variabile da sintetizzare))
```

Sfruttando questa notazione puoi creare anche dei grafici super per la prossima sezione!

Da qui in poi utilizza ggplot2 per creare grafici fantastici! Ma non fermarti alla semplice soluzione del quesito, prova un po' a personalizzare i grafici utilizzando temi, colori e annotazioni, e aggiustando i vari elementi come le etichette, gli assi e la legenda.

Consiglio: Fai attenzione quando specifichi le variabili month e year tra le estetiche, potrebbe essere necessario considerarle come fattori.

1. Utilizza i boxplot per confrontare la distribuzione del prezzo mediano delle case tra le varie città. Commenta il risultato.
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?
3. 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.

4. Prova a creare un line chart di una variabile a tua scelta per fare confronti commentati fra città e periodi storici. Ti avviso che probabilmente all'inizio ti verranno fuori linee storte e poco chiare, ma non demordere.

Consigli: Prova inserendo una variabile per volta. Prova a usare variabili esterne al dataset, tipo vettori creati da te appositamente.

Se non riesci proprio a venirne a capo inizia lavorando su dataset ridotti, ad esempio prendendo in considerazione un solo anno o una sola città. Aiutati con il pacchetto dplyr:

```
dati2014 <- filter(dati, year==2014)

dati_Beaumont <- filter(dati, city==”Beaumont”)
```

…altrimenti usa la funzione facet_wrap()