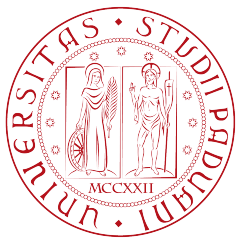
Lancio sul Mercato di un Nuovo Set LEGO®: Analisi del Prezzo

Annachiara Dal Colle, Chiara Micheletti, Andrea Rossi

Università degli Studi di Padova

28 Maggio 2020



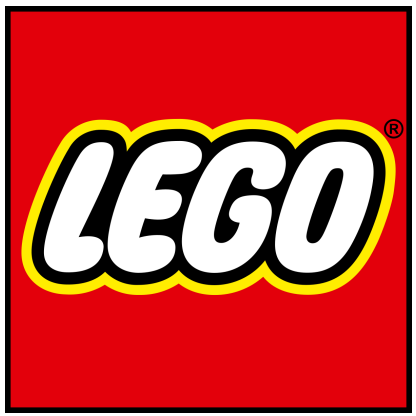
Sommario

1. Dataset
2. Domanda di Business
3. Analisi Esplorative
4. Modelli
5. Conclusioni

Dataset

Dataset

Il dataset oggetto d'analisi è stato scaricato dalla piattaforma `kaggle`. Ciascuna delle 12261 osservazioni fa riferimento a un preciso set dell'azienda danese **LEGO®**, riportandone alcune caratteristiche.

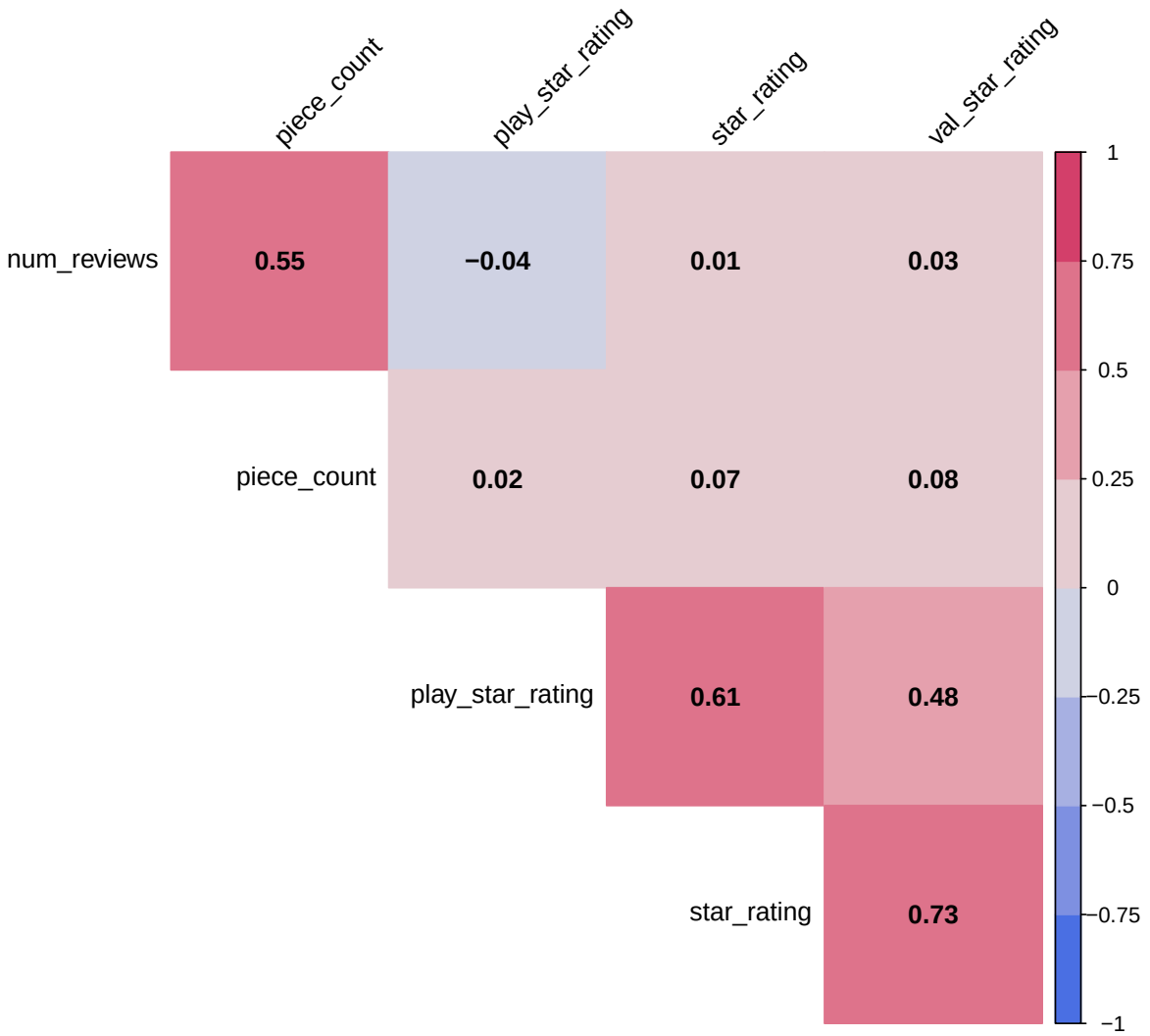


Variabili

Le variabili a disposizione sono:

- **list_price**
- ages
- review_difficulty
- prod_id
- prod_desc
- prod_long_desc
- num_reviews
- set_name
- theme_name
- star_rating
- play_star_rating
- val_star_rating
- country

Variabili


piece_count
play_star_rating
star_rating
val_star_rating
num_reviews
0.55
−0.04
0.01
0.03
piece_count
0.02
0.07
0.08
play_star_rating
0.61
0.48
star_rating
0.73
1
0.75
0.5
0.25
0
−0.25
−0.5
−0.75
−1

Assunzioni

Non avendo a disposizione informazioni dettagliate sulle variabili del dataset è stato assunto che:

- *star_rating* fosse la **soddisfazione complessiva** media assegnata dal consumatore al set (in stelline, da 1 a 5);
- *play_star_rating* fosse il livello medio di **giocabilità** assegnato dallo stesso campione al set (in stelline, da 1 a 5).

Struttura

Alcune righe del dataset fanno riferimento allo stesso set: cambiano solo la nazione (*country*) e il prezzo (*list_price*, espresso in USD) a cui il set viene venduto in tale nazione.

Sono presenti 21 stati, quelli in cui LEGO® ha uno store online, con l'esclusione della Corea del Sud.

I set distinti sono 744.

Dati Mancanti

- 1620 osservazioni sono state rimosse per poter procedere con la modellazione, perché presentavano dati mancanti per le variabili *star_rating* e *play_star_rating*;

Dati Mancanti

- 1620 osservazioni sono state rimosse per poter procedere con la modellazione, perché presentavano dati mancanti per le variabili *star_rating* e *play_star_rating*;
- Per 3 prodotti non risultava *theme_names*: l'informazione è stata reperita direttamente dal dataset e poi aggiunta manualmente;

Dati Mancanti

- 1620 osservazioni sono state rimosse per poter procedere con la modellazione, perché presentavano dati mancanti per le variabili *star_rating* e *play_star_rating*;
- Per 3 prodotti non risultava *theme_names*: l'informazione è stata reperita direttamente dal dataset e poi aggiunta manualmente;
- 2055 set non avevano assegnato un livello di difficoltà (*review_difficulty*): guardando a che tema appartenevano o il numero di pezzi che li componevano, è stato imputato;

Ricodificazione livelli

- La variabile *ages*, che presentava 31 categorie, è stata raggruppata in: "<6", "6-11", "12+".

Ricodificazione livelli

- La variabile *ages*, che presentava 31 categorie, è stata raggruppata in: "<6", "6-11", "12+".
- La variabile *theme_name*, che presentava 40 categorie, è stata ricodificata accorpando i temi affini creando nuove modalità, come per esempio "Dinosaurs" e "Action & Super Heroes". Sono risultati 21 livelli finali.

Ricodificazione livelli

- La variabile *ages*, che presentava 31 categorie, è stata raggruppata in: "<6", "6-11", "12+".
- La variabile *theme_name*, che presentava 40 categorie, è stata ricodificata accorpando i temi affini creando nuove modalità, come per esempio "Dinosaurs" e "Action & Super Heroes". Sono risultati 21 livelli finali.
- La variabile *review_difficulty* presentava 5 gradi di difficoltà: "Very Easy", "Easy", "Average", "Challenging" e "Very Challenging". Si è deciso di inglobare "Very Challenging" in "Challenging".

Domanda di Business

Domanda di Business

La domanda di business a cui si è provato a rispondere è:

Comprendere la relazione che esiste tra il **prezzo** di un set e le sue caratteristiche, per poter prevedere il prezzo di un **nuovo set LEGO®** da lanciare sul mercato.

Domanda di Business

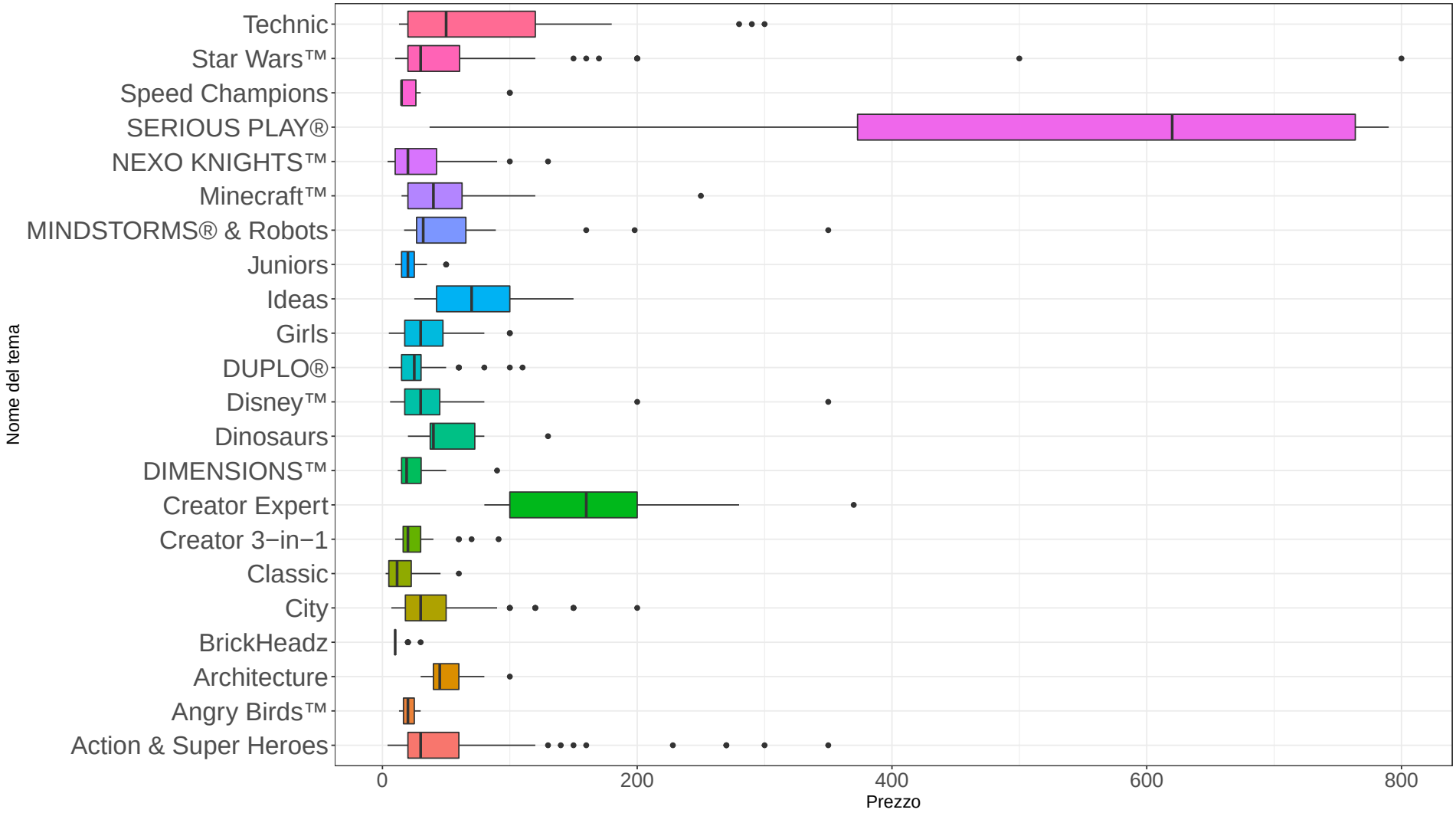
Si è immaginato che, in occasione degli 800 anni dell'Ateneo, l'Università di Padova avesse deciso di proporre un gadget speciale per i suoi utenti: un set LEGO®.

Tale set è...

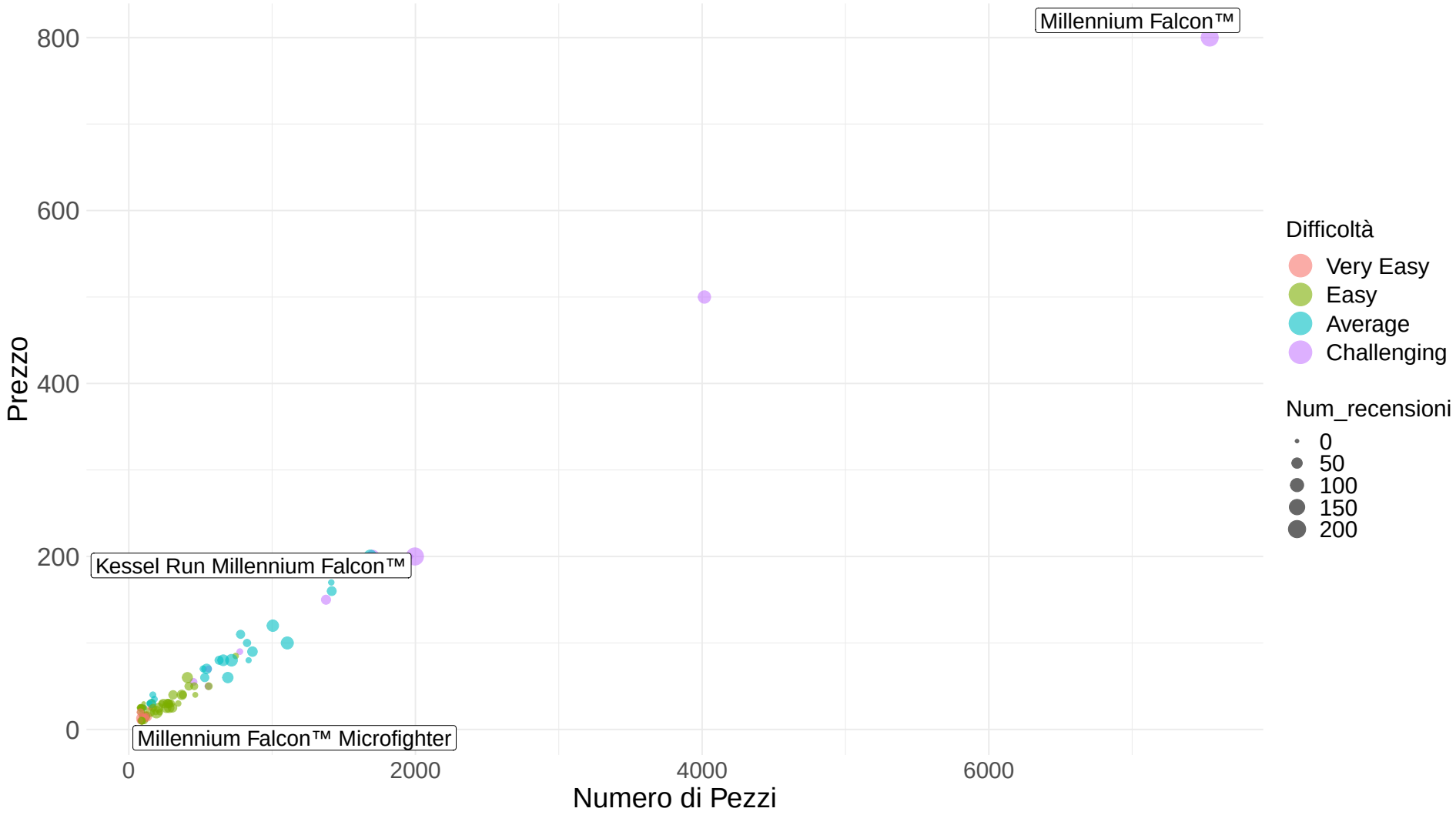
Santa Caterina^{TM}

Analisi Esplorative

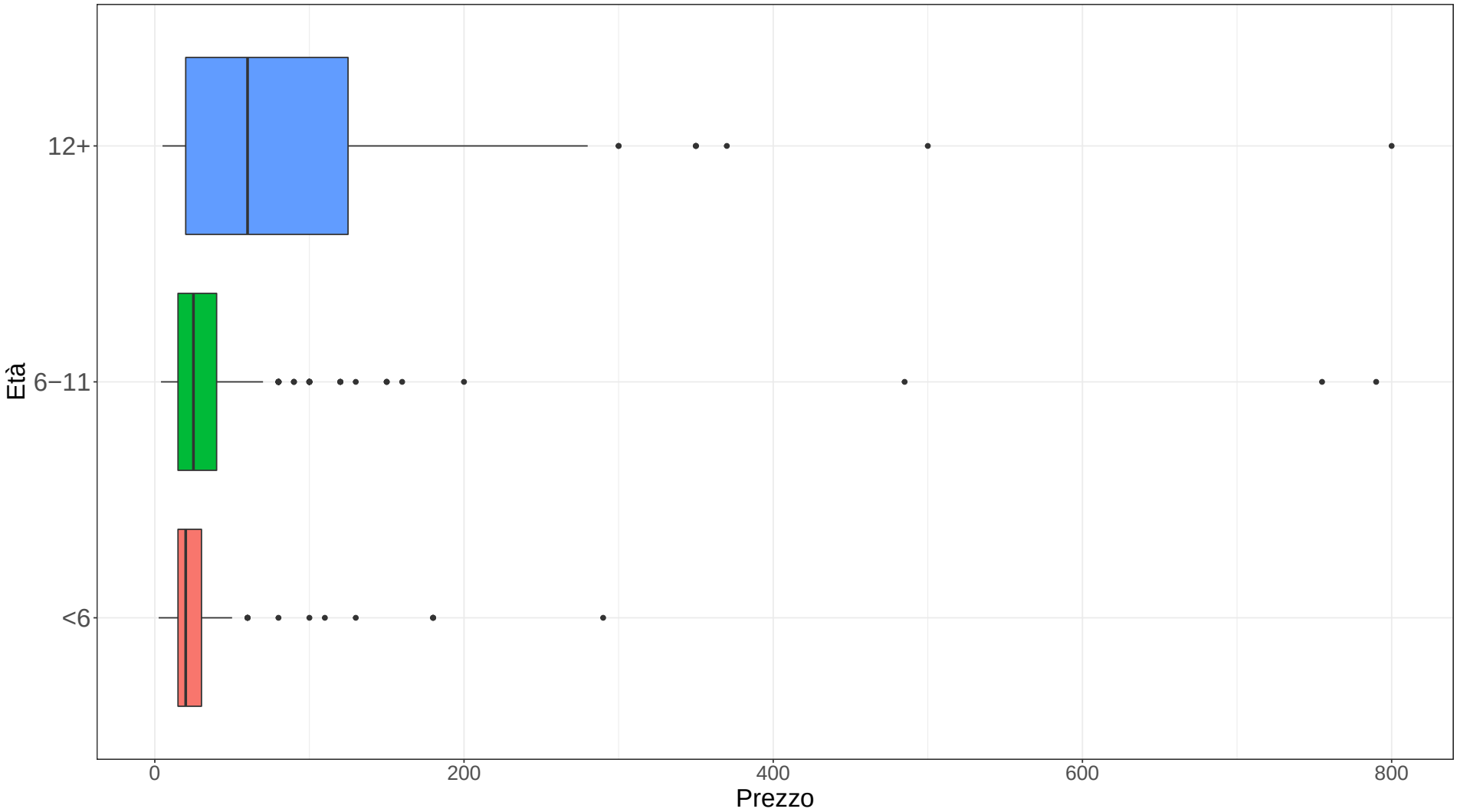
Variabile *theme_name*



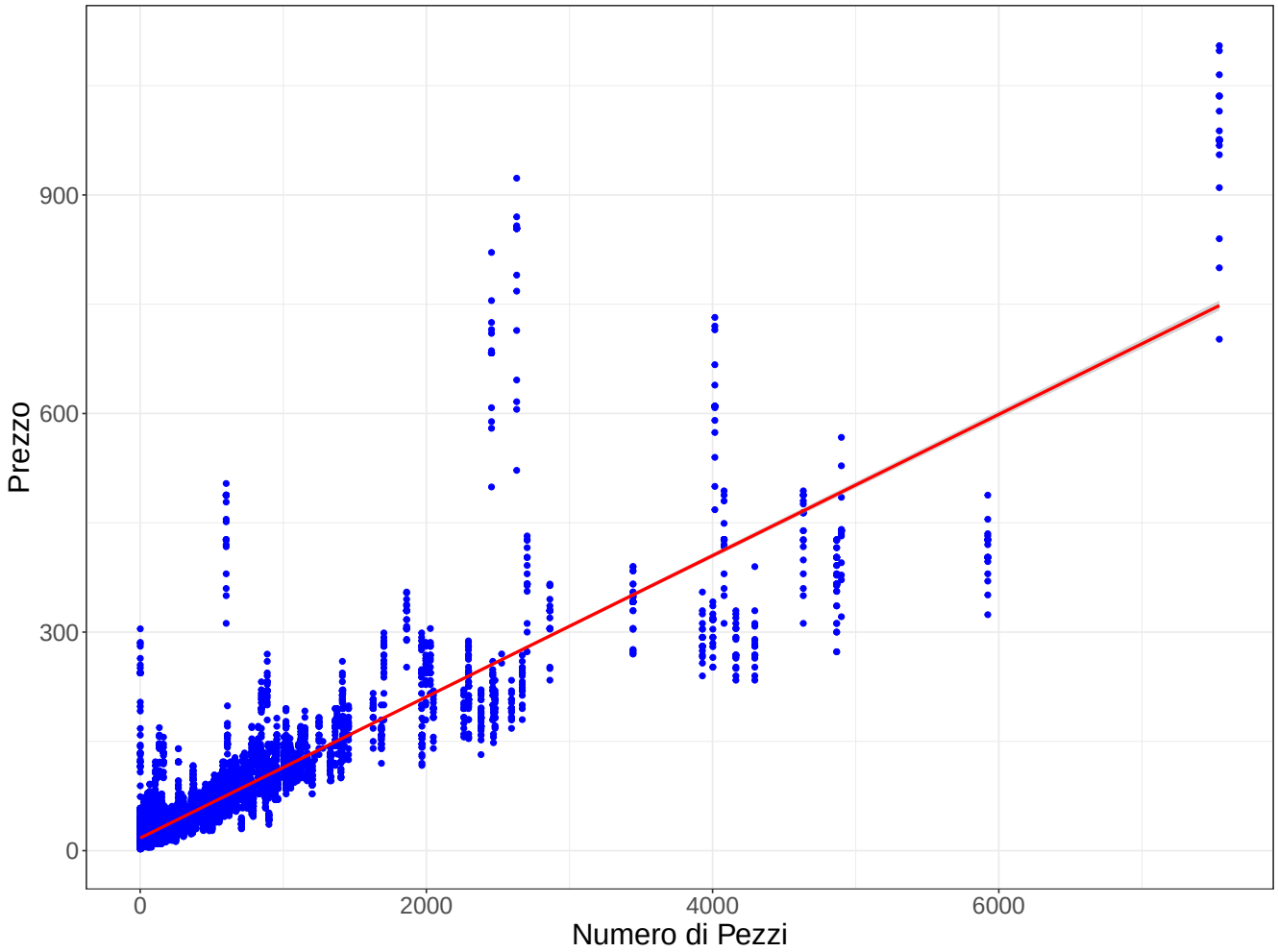
Il Millennium Falcon™



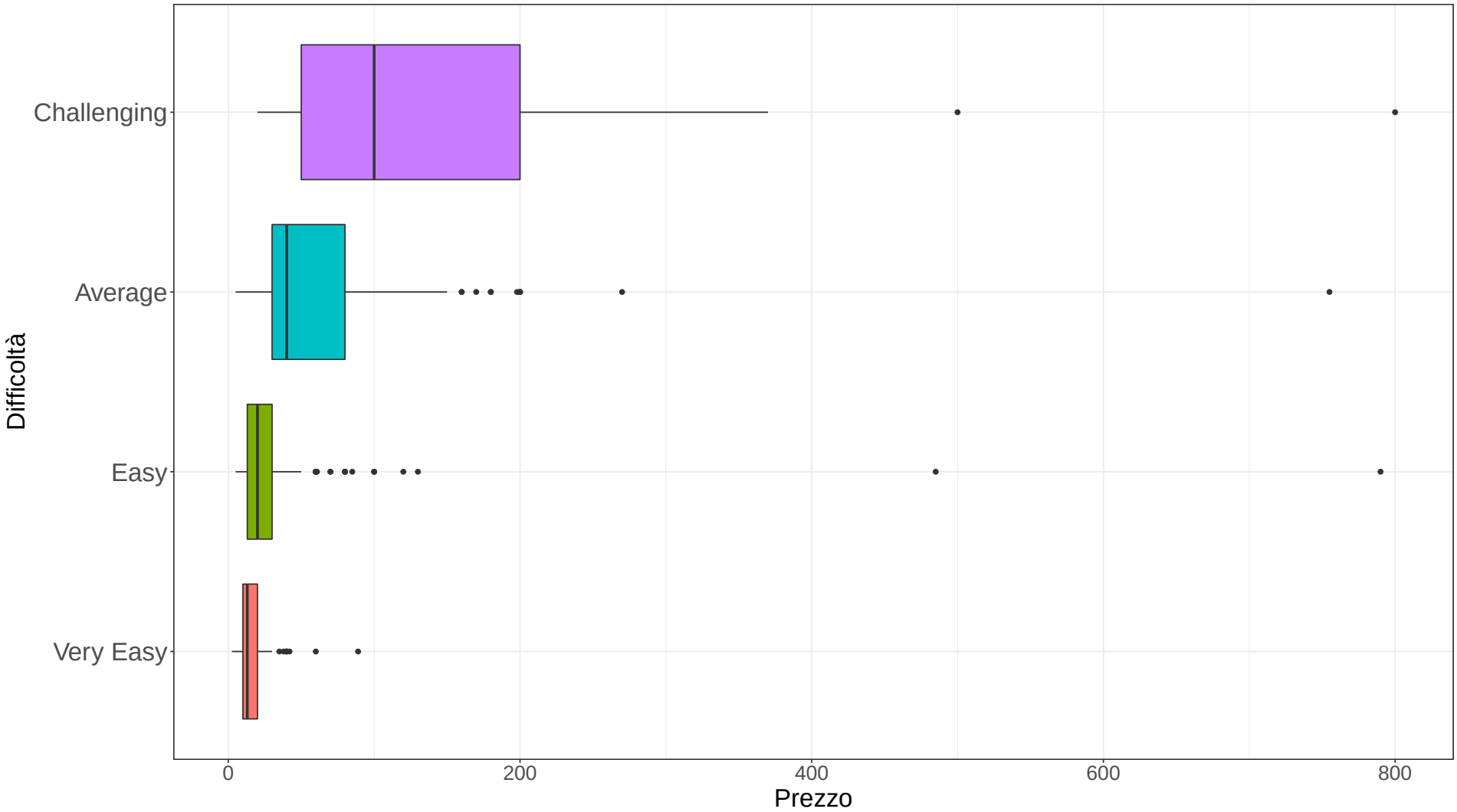
Variabile *ages*



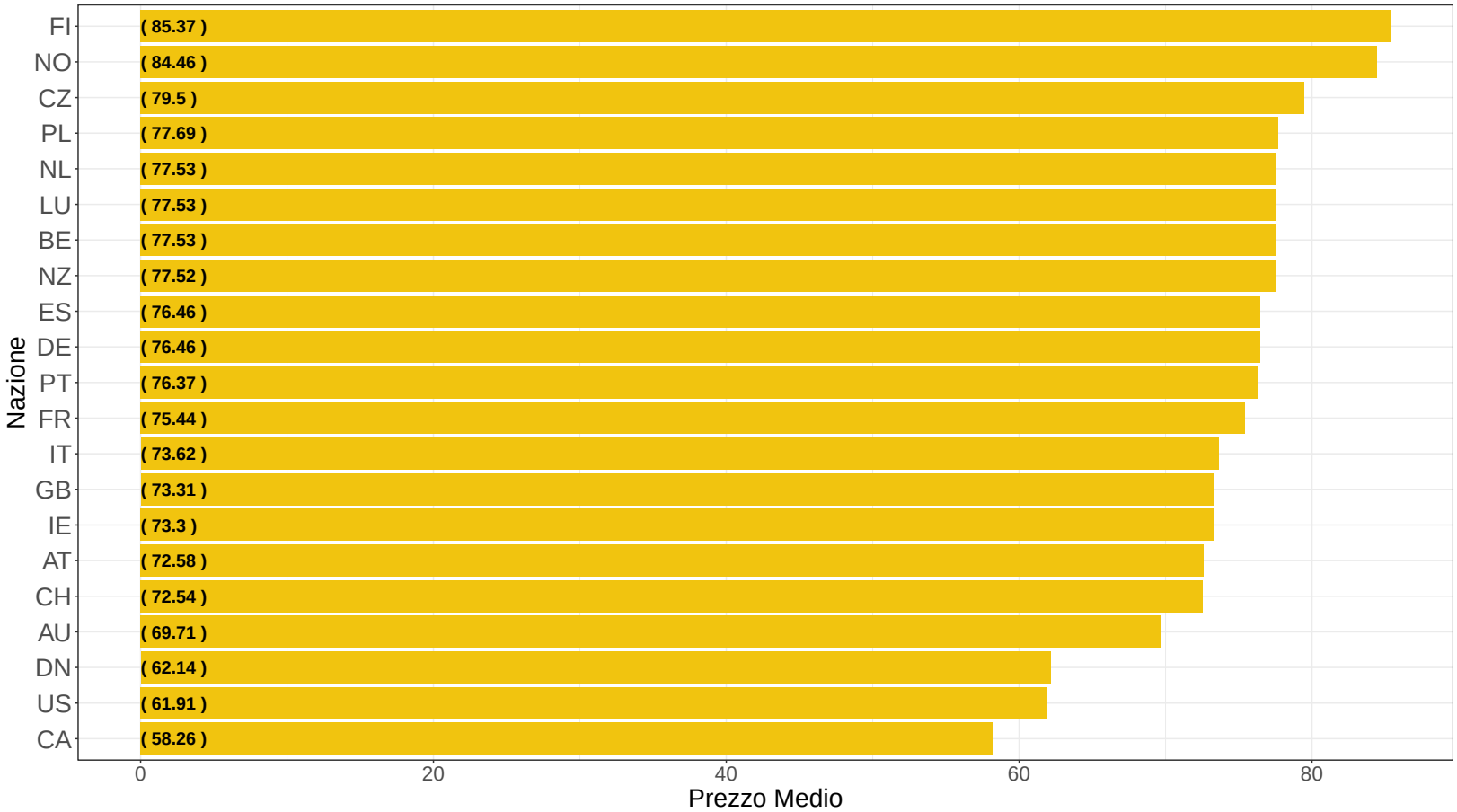
Variabile *piece_count*



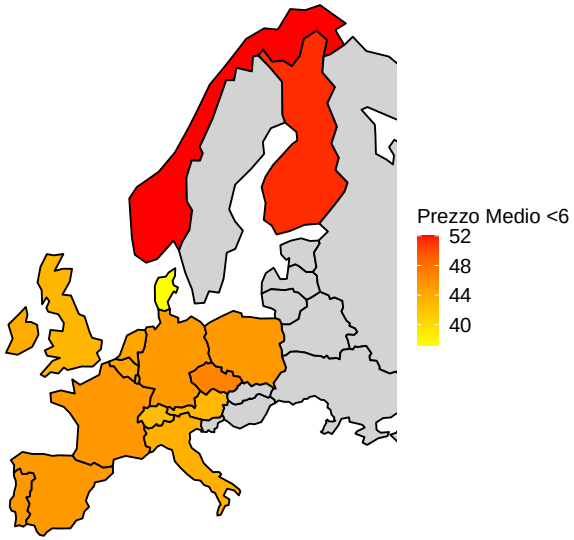
Variabile *review_difficulty*

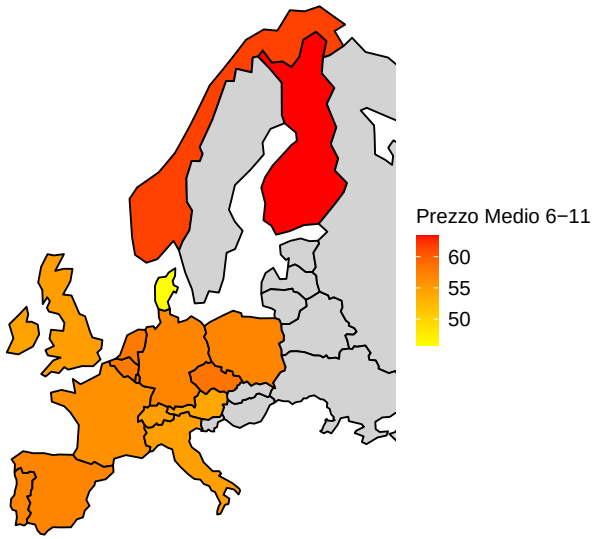


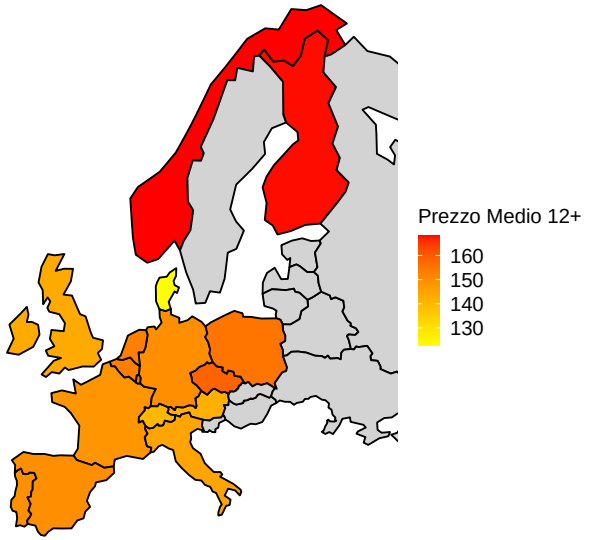
Variabile *country*



Il caso europeo: prezzo medio per classe d'età







Santa Caterina™

Le analisi esplorative hanno permesso di ottenere una panoramica abbastanza completa degli attributi di un set LEGO®. Si è quindi ipotizzato che il set di Santa Caterina™ avesse le seguenti caratteristiche:

Santa Caterina™

Le analisi esplorative hanno permesso di ottenere una panoramica abbastanza completa degli attributi di un set LEGO®. Si è quindi ipotizzato che il set di Santa Caterina™ avesse le seguenti caratteristiche:

- **Fascia d'età**: 12+

Santa Caterina™

Le analisi esplorative hanno permesso di ottenere una panoramica abbastanza completa degli attributi di un set LEGO®. Si è quindi ipotizzato che il set di Santa Caterina™ avesse le seguenti caratteristiche:

- **Fascia d'età**: 12+
- **Tema**: Creator Expert

Santa Caterina™

Le analisi esplorative hanno permesso di ottenere una panoramica abbastanza completa degli attributi di un set LEGO®. Si è quindi ipotizzato che il set di Santa Caterina™ avesse le seguenti caratteristiche:

- **Fascia d'età**: 12+
- **Tema**: Creator Expert
- **Numero di pezzi**: 4000

Santa Caterina™

Le analisi esplorative hanno permesso di ottenere una panoramica abbastanza completa degli attributi di un set LEGO®. Si è quindi ipotizzato che il set di Santa Caterina™ avesse le seguenti caratteristiche:

- **Fascia d'età**: 12+
- **Tema**: Creator Expert
- **Numero di pezzi**: 4000
- **Difficoltà**: Challenging

Santa Caterina™

Le analisi esplorative hanno permesso di ottenere una panoramica abbastanza completa degli attributi di un set LEGO®. Si è quindi ipotizzato che il set di Santa Caterina™ avesse le seguenti caratteristiche:

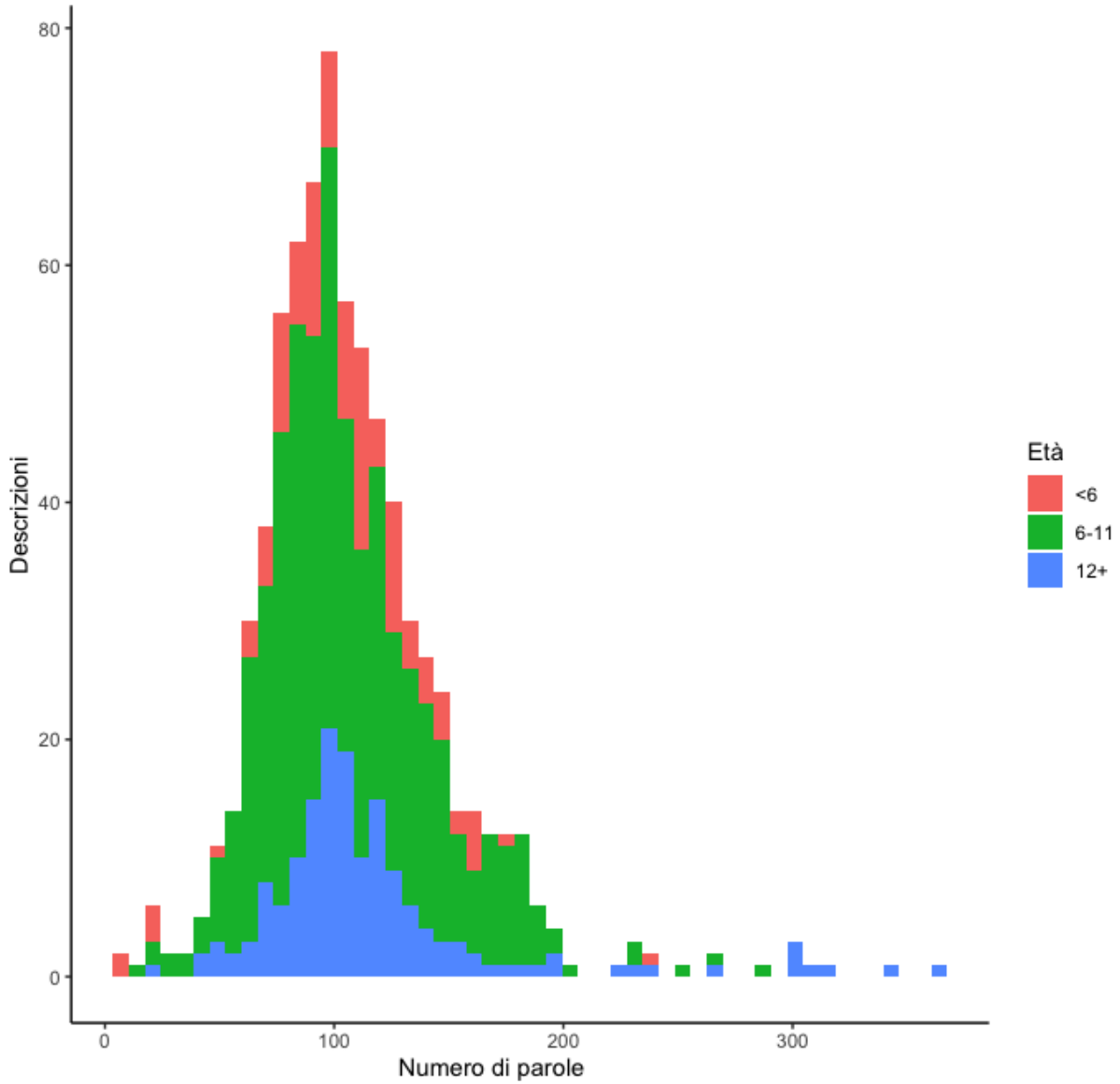
- **Fascia d'età**: 12+
- **Tema**: Creator Expert
- **Numero di pezzi**: 4000
- **Difficoltà**: Challenging
- **Paese**: Italia

Text Mining

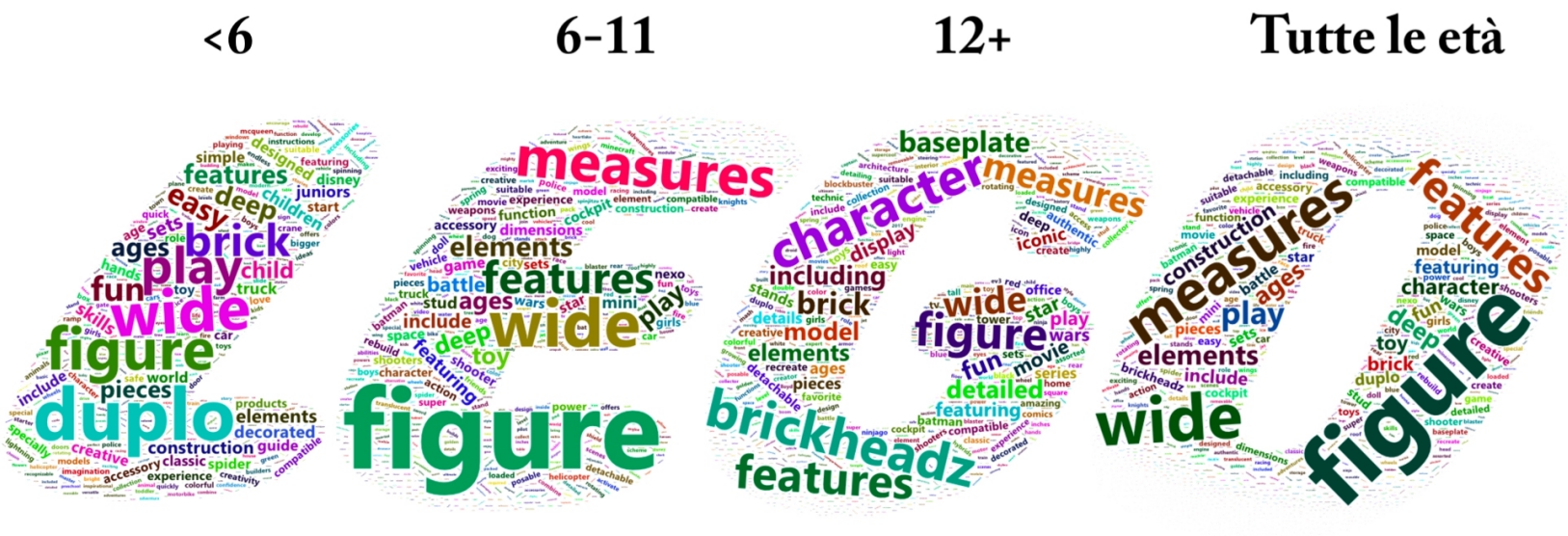
Ogni set LEGO® necessita di un'accurata descrizione confacente alle caratteristiche del prodotto e del cliente a cui è destinato.

Per questo motivo si è deciso di analizzare lunghezza e contenuto della variabile *prod_long_desc*, in base alle diverse fasce d'età.

Text Mining


80
60
40
20
0
Descrizioni
0
100
200
300
Numero di parole
Età
<6
6-11
12+

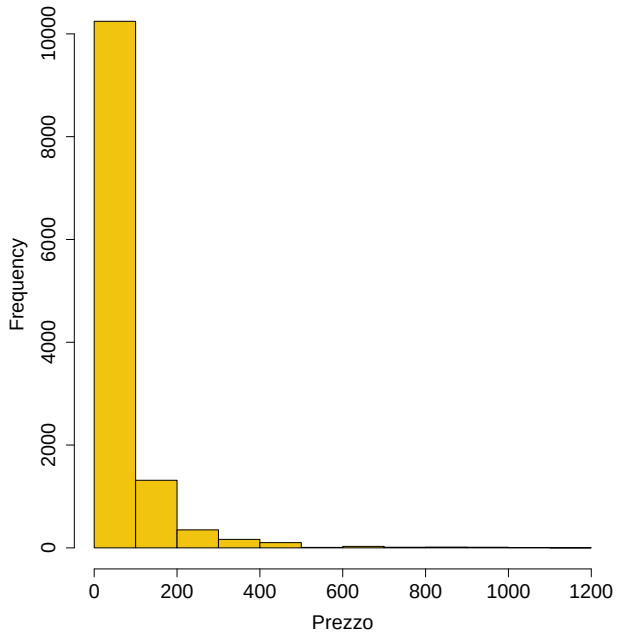
Wordcloud

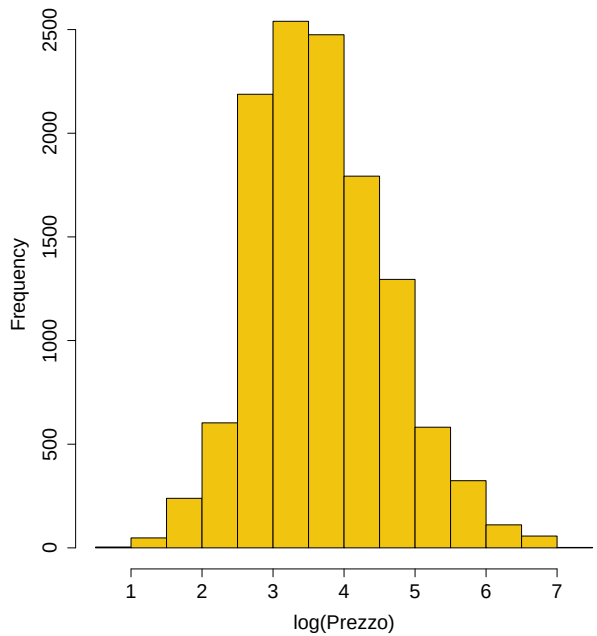


Modelli

Introduzione

Per poter procedere con l'implementazione dei modelli è stata apportata una trasformazione logaritmica alla variabile risposta, vista la netta asimmetria che presentava la sua distribuzione.





Introduzione

- Per la stima dei modelli non sono state prese in considerazione le variabili relative alla descrizione e all'identificazione del prodotto.

Introduzione

- Per la stima dei modelli non sono state prese in considerazione le variabili relative alla descrizione e all'identificazione del prodotto.

- Per evitare di ridurre ulteriormente il numero di variabili esplicative si è deciso di tenere comunque conto delle variabili relative ai consumatori.

Introduzione

- Per la stima dei modelli non sono state prese in considerazione le variabili relative alla descrizione e all'identificazione del prodotto.
- Per evitare di ridurre ulteriormente il numero di variabili esplicative si è deciso di tenere comunque conto delle variabili relative ai consumatori.
- Le variabili quantitative sono state standardizzate, data la presenza di range molto diversi.

Introduzione

- Per la stima dei modelli non sono state prese in considerazione le variabili relative alla descrizione e all'identificazione del prodotto.
- Per evitare di ridurre ulteriormente il numero di variabili esplicative si è deciso di tenere comunque conto delle variabili relative ai consumatori.
- Le variabili quantitative sono state standardizzate, data la presenza di range molto diversi.
- Si è deciso di procedere attraverso stima-verifica, con un *train set* contenente il 60% delle osservazioni ed un *test set* contenente il restante 40%.

Regressione lineare

Coefficients:	Estimate	Std. Error	t value	$Pr(>\|t\|)$	
(Intercept)	4.146434	0.284182	14.591	$<$ 2e-16	***
ages6-11	0.605744	0.054717	-11.071	$<$ 2e-16	***
ages12+	0.674073	0.052549	-5.216	1.89e-07	***
num_reviews	-0.022256	0.008468	-2.628	0.008601	**
piece_count	0.505620	0.009905	51.047	$<$ 2e-16	***
play_star_rating	0.159329	0.014695	10.842	$<$ 2e-16	***
review_difficultyEasy	0.363591	0.021221	17.133	$<$ 2e-16	***
review_difficultyAverage	0.894236	0.023218	38.515	$<$ 2e-16	***
review_difficultyChallenging	0.898615	0.033111	27.140	$<$ 2e-16	***
star_rating	-0.274719	0.016858	-16.296	$<$ 2e-16	***

Commenti

- I coefficienti di *ages* sono significativi, crescono al crescere della fascia d'età a cui è rivolto il set e sono tutti positivi, coerentemente con il fatto che la categoria di riferimento è "6+";
- I coefficienti di *review_difficulty* sono significativi, crescono al crescere della difficoltà e sono tutti positivi, coerentemente con il fatto che la categoria di riferimento è "Very Easy";

Regressione lineare

Coefficients:	Estimate	Std. Error	t value	$Pr(>\|t\|)$	
theme_nameArchitecture	0.039864	0.275646	0.145	0.885015	
theme_nameMINDSTORMS® & Robots	0.659876	0.275690	2.394	0.016716	*
theme_nameBrickHeadz	-0.582443	0.273648	-2.128	0.033340	*
theme_nameCity	0.180001	0.272153	0.661	0.508382	
theme_nameJuniors	-0.663871	0.277776	-2.390	0.016880	*
theme_nameClassic	-1.050680	0.278397	-3.774	0.000162	***
theme_nameCreator 3-in-1	-0.155369	0.272997	-0.569	0.569292	
theme_nameCreator Expert	-0.059872	0.275225	-0.218	0.827794	
theme_nameAction & Super Heroes	0.081720	0.271815	0.301	0.763696	
theme_nameDIMENSIONS™	0.115583	0.274134	0.422	0.673310	
theme_nameDisney™	0.133642	0.273947	0.488	0.625680	
theme_nameDUPLO®	-0.151133	0.277343	-0.545	0.585819	
theme_nameGirls	-0.020081	0.272512	-0.074	0.941261	
theme_nameIdeas	0.243416	0.277622	0.877	0.380634	
theme_nameDinosaurs	0.688989	0.277203	2.486	0.012963	*
theme_nameMinecraft™	0.261041	0.273947	0.953	0.340683	
theme_nameNEXO KNIGHTS™	0.088925	0.277361	0.321	0.748517	
theme_nameSERIOUS PLAY®	0.979667	0.280660	3.491	0.000485	***
theme_nameSpeed Champions	-0.020135	0.274090	-0.073	0.941441	
theme_nameStar Wars™	0.310451	0.272001	1.141	0.253765	
theme_nameTechnic	0.101550	0.272987	0.372	0.709909	

Regressione lineare

Coefficients:	Estimate	Std. Error	t value	$Pr(>\|t\|)$	
countryAU	-0.087319	0.038379	-2.275	0.022930	*
countryBE	0.080447	0.038786	2.074	0.038108	*
countryCA	-0.282633	0.035051	-8.064	8.82e-16	***
countryCH	-0.020828	0.037376	-0.557	0.577378	
countryCZ	0.094406	0.038887	2.428	0.015224	*
countryDE	0.025269	0.038474	0.657	0.511335	
countryDN	-0.122233	0.038087	-3.209	0.001337	**
countryES	0.039301	0.038820	1.012	0.311394	
countryFI	0.177805	0.038548	4.612	.06e-06	***
countryFR	0.016851	0.038223	0.441	0.659329	
countryGB	0.003076	0.039267	0.078	0.937558	
countryIE	-0.002059	0.038388	-0.054	0.957227	
countryIT	-0.024652	0.038810	-0.635	0.525326	
countryLU	0.040991	0.039512	1.037	0.299579	
countryNL	0.052912	0.037952	1.394	0.163308	
countryNO	0.165726	0.039509	4.195	2.77e-05	***
countryNZ	0.086219	0.038403	2.245	0.024798	*
countryPL	0.018140	0.038907	0.466	0.641054	
countryPT	0.043260	0.038387	1.127	0.259811	
countryUS	-0.254936	0.035104	-7.262	4.27e-13	***

Commenti

- L'appartenenza del set a un certo tema (con riferimento a *Angry Birds*^{TM}) può essere in alcuni casi significativa (*SERIOUS PLAY*® e *Classic* hanno rispettivamente il coefficiente massimo e il coefficiente minimo);
- Guardando i coefficienti di *country* (in cui la nazione di riferimento è l'Austria) si ritrova quanto visto con le analisi esplorative. I coefficienti significativi con segno negativo e modulo più elevato si trovano proprio in Stati Uniti, Canada e Danimarca.

Commenti

- L'appartenenza del set a un certo tema (con riferimento a *Angry Birds*^{TM}) può essere in alcuni casi significativa (*SERIOUS PLAY*® e *Classic* hanno rispettivamente il coefficiente massimo e il coefficiente minimo);
- Guardando i coefficienti di *country* (in cui la nazione di riferimento è l'Austria) si ritrova quanto visto con le analisi esplorative. I coefficienti significativi con segno negativo e modulo più elevato si trovano proprio in Stati Uniti, Canada e Danimarca.

Il modello lineare, nella sua semplicità, è riuscito comunque a cogliere in modo sensato alcuni aspetti del problema rivelandosi un buon punto di partenza.

Modelli stimati e MSE

Modello	MSE
Random Forest	0.027
Gradient Boosting	0.028
Rete Neurale	0.035
Albero di Regressione	0.048
GAM	0.105
MARS	0.129
Bagging	0.142
Modello Lineare	0.218

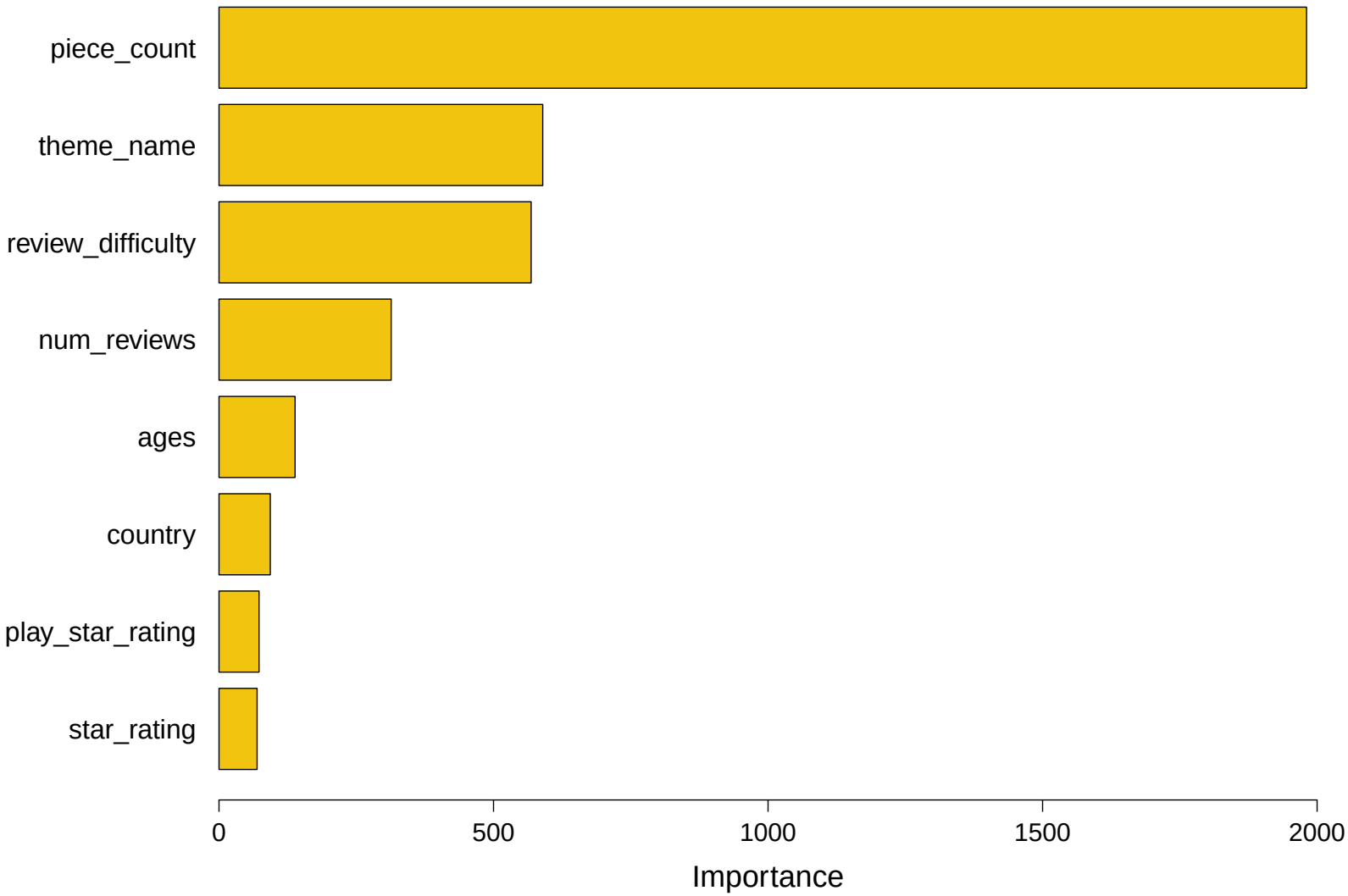
Tabella: Modelli stimati e relativi MSE

Modello scelto

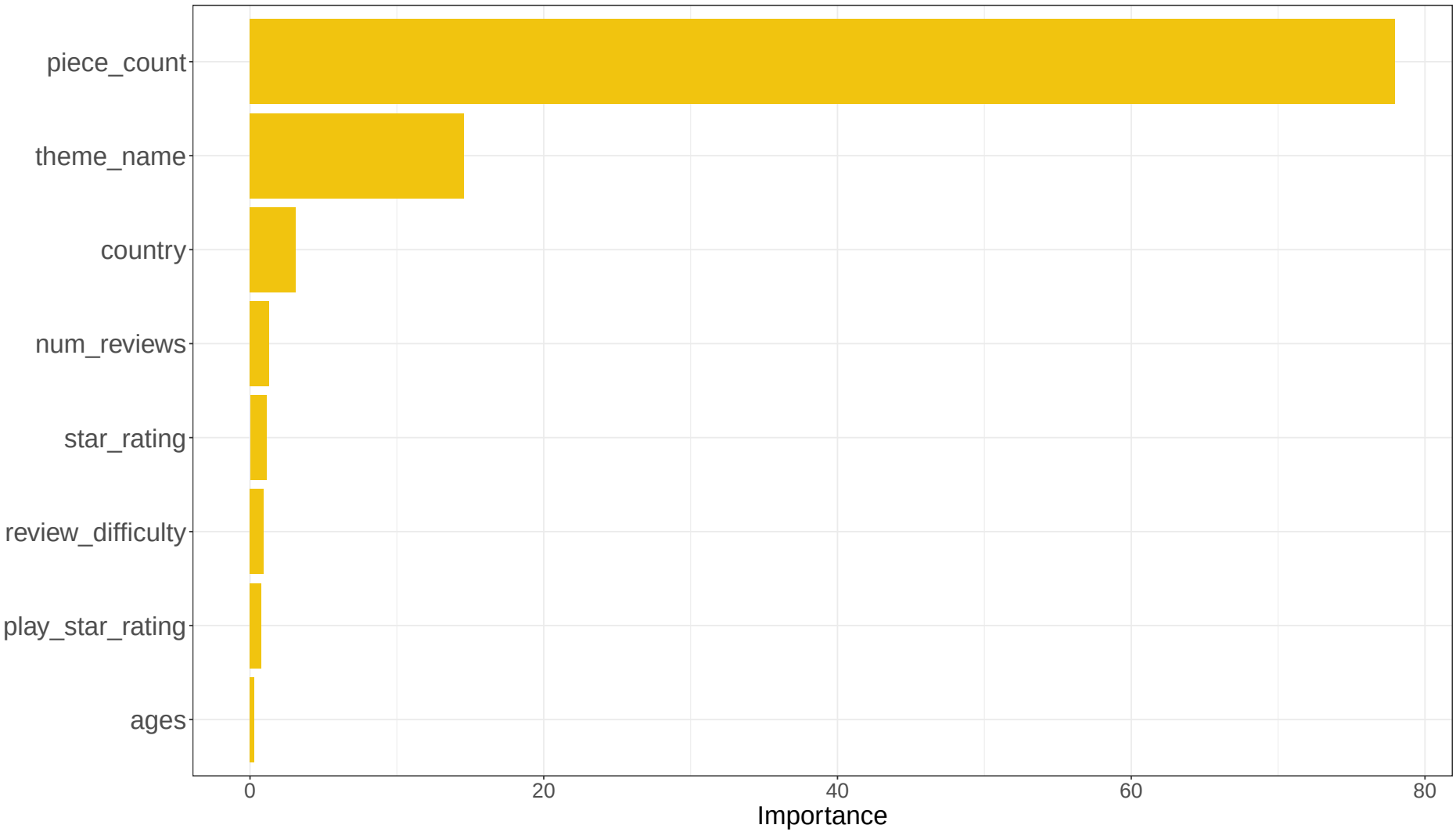
Il modello migliore, in termini di MSE, risulta essere il Random Forest.

Vista l'alta interpretabilità, di seguito verranno presentati il grafico di importanza relativa delle variabili ottenuto attraverso questo modello, seguito da quello del Gradient Boosting e dall'Albero di Regressione.

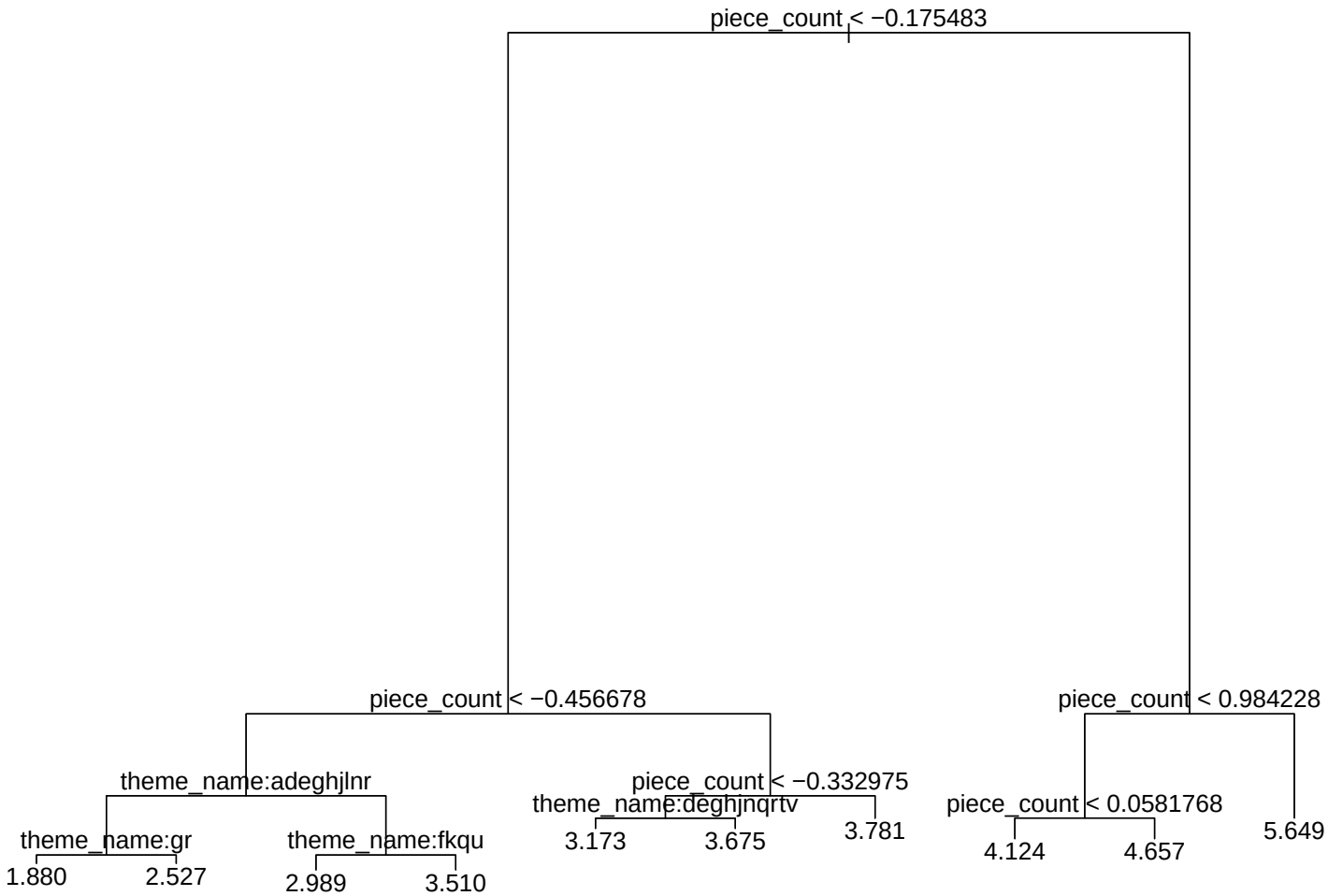
Random Forest



Gradient Boosting



Albero di Regressione



Conclusioni

Conclusioni

- Nel lanciare un nuovo set di LEGO® sul mercato, per determinarne il prezzo, bisogna tenere conto di più aspetti, di cui il principale è certamente il numero di pezzi.

Conclusioni

- Nel lanciare un nuovo set di LEGO® sul mercato, per determinarne il prezzo, bisogna tenere conto di più aspetti, di cui il principale è certamente il numero di pezzi.
- L'elemento immediatamente successivo da considerare è il tema cui si è deciso di far appartenere il set.

Conclusioni

- Nel lanciare un nuovo set di LEGO® sul mercato, per determinarne il prezzo, bisogna tenere conto di più aspetti, di cui il principale è certamente il numero di pezzi.
- L'elemento immediatamente successivo da considerare è il tema cui si è deciso di far appartenere il set.
- Difficoltà e fascia d'età sono meno rilevanti, probabilmente perché derivano dalle decisioni precedenti.

Conclusioni

- Nel lanciare un nuovo set di LEGO® sul mercato, per determinarne il prezzo, bisogna tenere conto di più aspetti, di cui il principale è certamente il numero di pezzi.
- L'elemento immediatamente successivo da considerare è il tema cui si è deciso di far appartenere il set.
- Difficoltà e fascia d'età sono meno rilevanti, probabilmente perché derivano dalle decisioni precedenti.
- I prezzi nei vari stati sono abbastanza omogenei, bisogna fare più attenzione in Danimarca, Stati Uniti e Canada, dove i prezzi sono più bassi e in Finlandia e Norvegia, dove sono invece più elevati.

Conclusioni

Che prezzo assegnare quindi al nuovo set da lanciare?

Conclusioni

Che prezzo assegnare quindi al nuovo set da lanciare?

Con il modello scelto, il prezzo previsto per il set Santa Caterina^{TM} è pari a **275.99 USD**.

(In caso qualcuno fosse interessato, si tratta di circa 250 €).

Grazie per l'attenzione!