

Politecnico di Milano

Ing. Fisica

Corso di Statistica - Prof. Alessandro Toigo

Anno accademico 2017/2018

Diamanti: analisi statistica e deduzione di una legge di prezzo



Garbagnati Elisa

Grieci Valentina

Negretti Fabio

1. Introduzione

Il presente progetto vuole condurre un'analisi statistica di un dataset contenente informazioni (numeriche e categoriche) relative a 4500 diamanti. Segnaliamo che gli elementi di tale data-frame sono stati estratti randomicamente, attraverso opportuno algoritmo, da un ulteriore campione costituito da 53940 realizzazioni; si è ritenuta tale scelta auspicabile per la maggiore facilità di gestione di una mole ridotta di dati, congiunta ad uno sgravio sulle risorse computazionali del calcolatore e alla possibilità di eseguire determinati test statistici, altresì irrealizzabili.

Obiettivo principale del progetto è risalire ad una legge empirica di regressione per la deduzione del prezzo di un diamante a partire dagli attributi della gemma inclusi nel dataset. Si vogliono inoltre dimostrare alcune affermazioni, utili al nostro scopo e mostrare la veridicità (o meno) di assunti generali sul soggetto.

Per la gestione dei dati e la realizzazione dei grafici e modelli a seguire, ci siamo serviti del supporto computazionale del software statistico R.

Lasciamo, in calce al documento, la fonte del data-frame integrale con cui abbiamo operato, assieme agli indirizzi da cui sono state tratte le informazioni con cui si contestualizzerà a breve il soggetto dell'analisi.

2. Contenuto del dataset

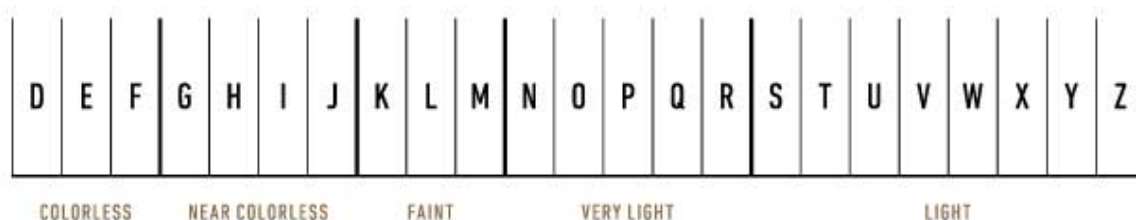
Come già menzionato, il data-frame si costituisce di 4500 righe, ciascuna associata al singolo esemplare di diamante; se ne presentano ora gli attributi, riportati ciascuno in una delle 10 colonne:

- *Carat*: è il peso del diamante espresso in carati [k], unità di misura correntemente usata nell'ambito della gioielleria, secondo l'equivalenza $1k = 0.2g$.
- *Cut*: qualità del taglio, da intendersi come variabile categorica che assume per valori le seguenti accezioni (in ordine crescente) di qualità; "FAIR"<"GOOD"<"VERY GOOD"<"PREMIUM"<"IDEAL".
- *Color*: colore del diamante, sempre espresso attraverso variabile categorica, che assume qui i possibili valori di "D">"E">"F">"G">"H">"I">"J"; dove sono stati indicati per primi i codici riferiti a gemme praticamente incolori (D, E, F), sino a giungere a quelle che presentano un'impercettibile punta di colore giallo.

Riportiamo di seguito, per chiarezza, la scala presentata dalla GIA, osservando come vi figurino numerosi altri codici, non presenti nel dataset a disposizione:

La preferenza generale ricade in maniera particolare su esemplari "colorless" o su tonalità cromatiche piuttosto singolari, da non essere presentate sulla scala di cui sotto.

GIA COLOR SCALE

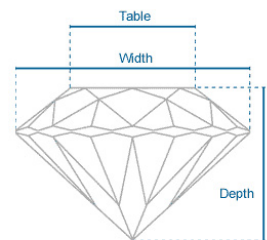


- *Clarity*: Indicatore qualitativo riferito alla presenza di particolari segni che caratterizzano la parte interna dell'esemplare ("inclusions"), determinate dalle estreme condizioni di pressione e calore con cui un diamante viene generato in natura. Quanto più la connotazione dell'indicatore è positiva, tanto più l'esemplare è privo di tali imperfezioni, con ovvie ricadute sul prezzo. Si riporta anche qui la classificazione GIA:

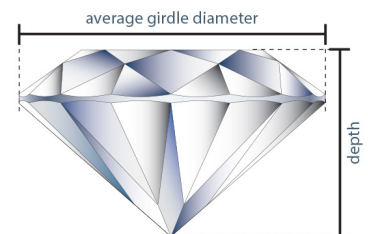
GIA CLARITY SCALE

FLAWLESS	INTERNALLY FLAWLESS	VVS ₁	VVS ₂	VS ₁	VS ₂	SI ₁	SI ₂	I ₁	I ₂	I ₃
		VERY VERY SLIGHTLY INCLUDED		VERY SLIGHTLY INCLUDED		SLIGHTLY INCLUDED		INCLUDED		

- *Table* (%): è la larghezza della tavola (“table”) del diamante espressa secondo valore percentuale rispetto al suo diametro medio. Si veda la figura:



- *Depth* (%): è la distanza tra la tavola e l’apice (nelle figure il punto estremo inferiore), divisa per il diametro medio (“average girdle diameter”).



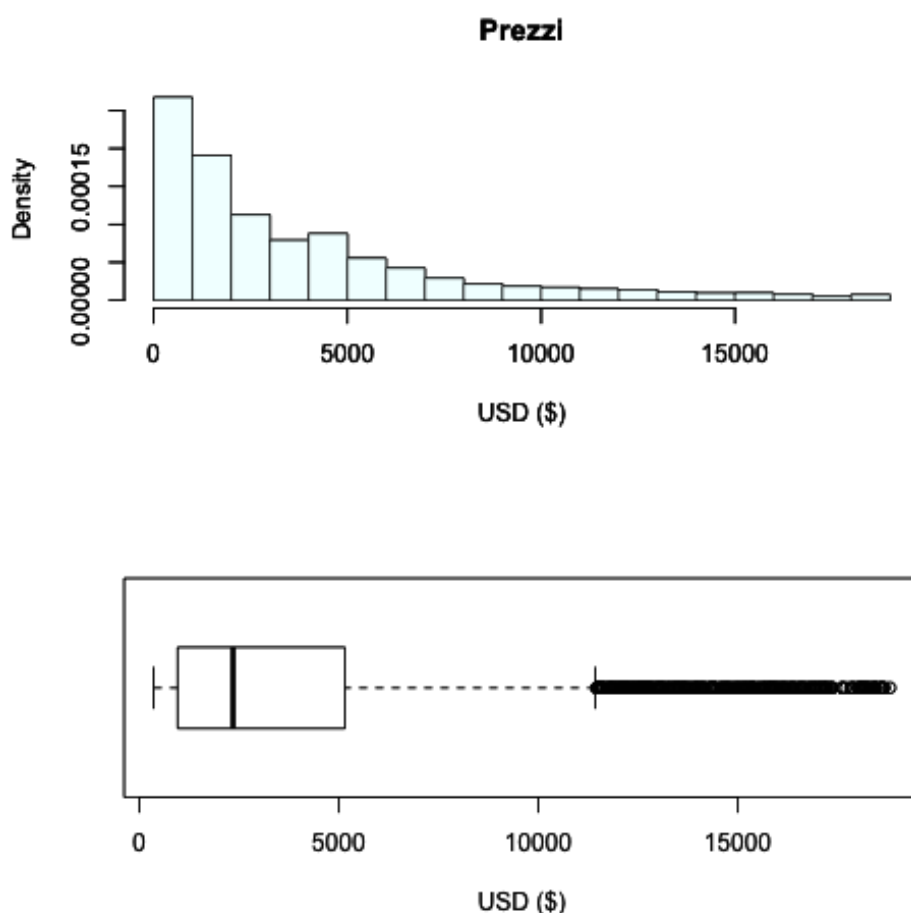
- *Price*: prezzo (US\$).
- *x*: lunghezza [mm].
- *y*: larghezza [mm].
- *z*: profondità [mm].

Si noti che sono già state fatte alcune assunzioni relative alle variazioni di prezzo, che risultano vere sull’attuale mercato e giustificate da opportuna fonte, dovute a determinati attributi del diamante, quali presenza di imperfezioni (clarity) o colore.

Sebbene si tratti di osservazioni qualitative e addirittura intuitive, vorremmo dimostrare tali affermazioni attraverso la realizzazione di un opportuno modello di descrizione del prezzo.

3. Statistica descrittiva

Si vuole iniziare con un'indagine esplorativa del dataset a disposizione per dedurne i caratteri generali e andamenti particolari. Essendo l'interesse ultimo uno studio sui prezzi, ne analizziamo innanzitutto la distribuzione:

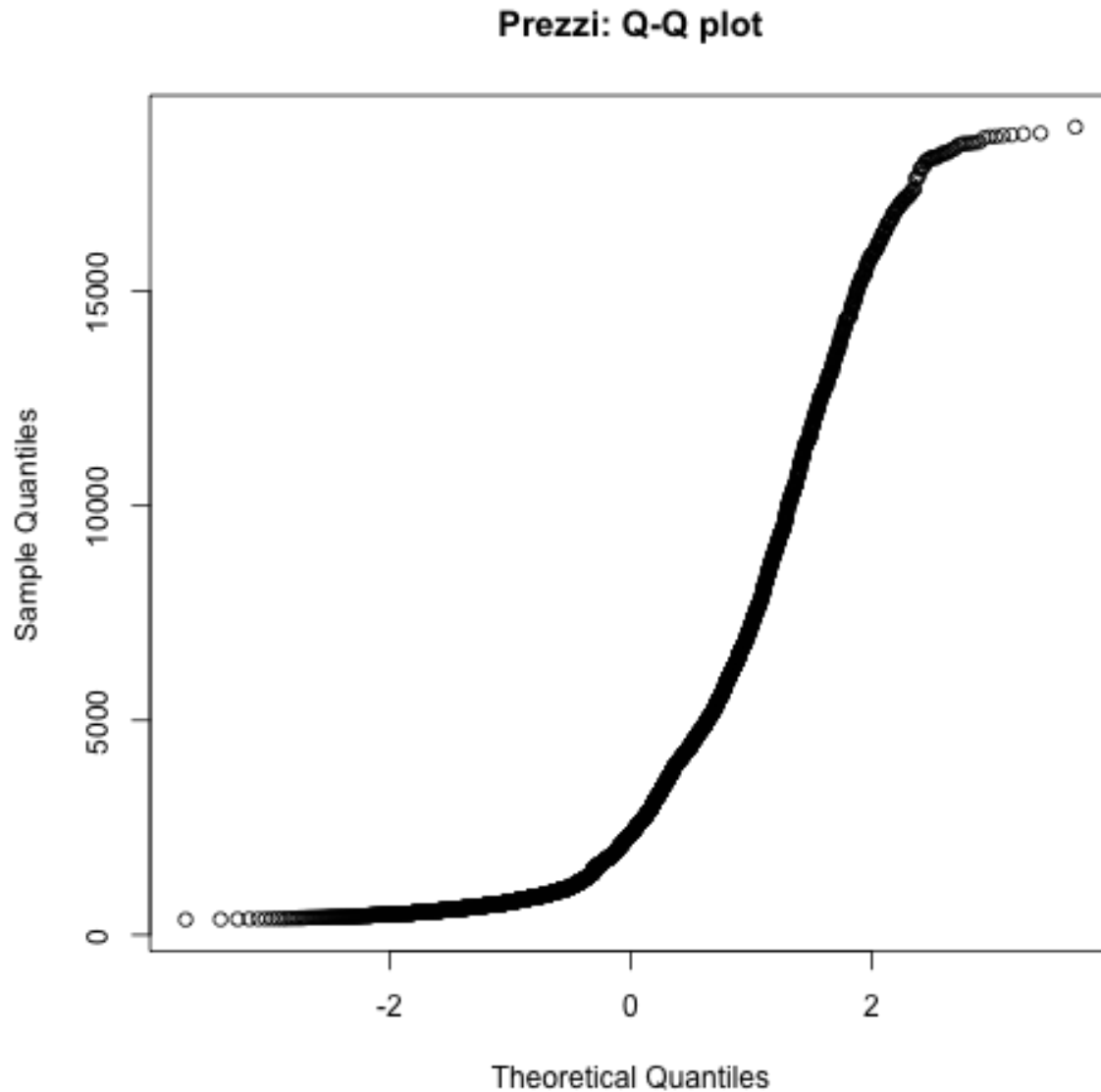


Si riscontra, già a partire dall'istogramma, una forte asimmetria dei dati, con presenza di una lunga coda a destra che ci viene confermata dall'osservazione degli indici di posizione riportati poco sotto (in particolare $\text{media} > \text{mediana}$, dove la media è ben più suscettibile alla presenza di valori "anomali") e dall'abbondante presenza di outliers nella regione destra del box plot; si deduce da qui una forte variabilità nel prezzo.

La maggior parte dei dati, in ogni caso, si concentra a sinistra, distribuendosi attorno ad una mediana di soli 2349 \$

```
summary(price)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   357    961    2349    3865    5154   18823
```

Riportiamo di seguito, unicamente per completezza (viste le evidenze fornite dai grafici sopra), il Q-Q plot dei prezzi, a riprova che è impossibile cogliere un andamento gaussiano nella distribuzione dei prezzi del presente dataset.



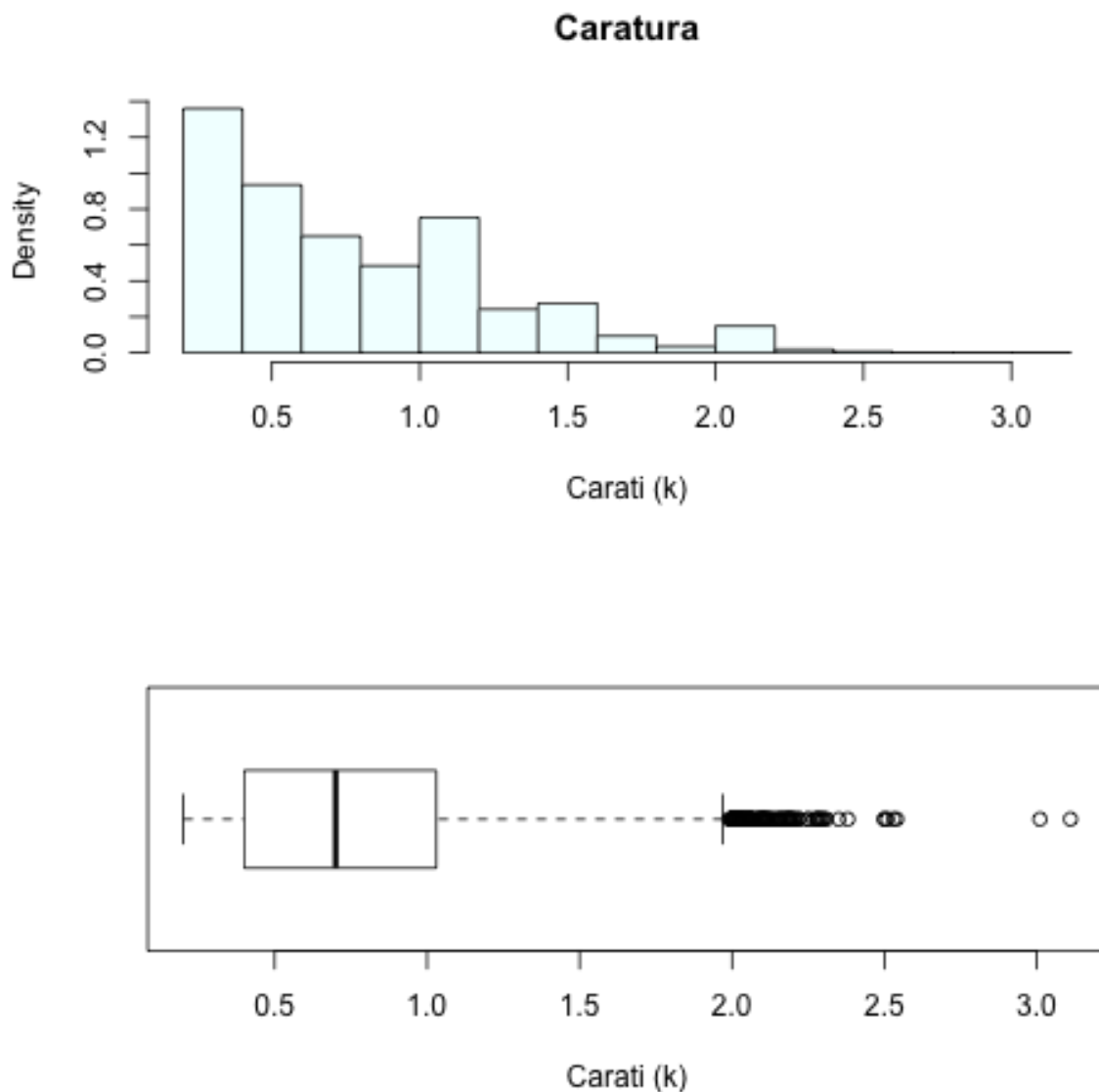
Ricordando che il dataset con cui operiamo è stato estratto casualmente da uno più ampio e dunque potrebbe non ricalcarne le tipicità, ci sentiamo comunque rassicurati dalla selezione di righe operata dal calcolatore, in quanto tale distribuzione e quelle a seguire si mostrano conformi all'andamento presentato dal campione integrale.

Proponiamo lo stesso approccio per analizzare gli altri attributi presentati nel par2:

circa la distribuzione delle carature, osserviamo una tendenza che è simile a quella già vista per i prezzi, con dati molto diradati sulla destra, con conseguente lunga coda e distribuzione evidentemente non gaussiana.

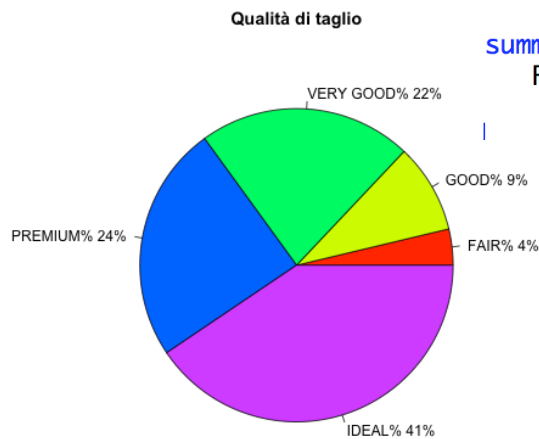
```
summary(carat)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.2000 0.4000 0.7000 0.7892 1.0300 3.1100
```

Si riportano sotto i soliti istogramma e box-plot a riprova di quanto asserito; per non appesantire ulteriormente la trattazione di inutili grafici, riteniamo poco adeguato l'introduzione di un Q-Q plot che rimarchi quanto già è adeguatamente mostrato dai prossimi grafici, circa la non gaussianità.



Prima di fornire un'interpretazione ad alcune delle interazioni più significative tra le variabili in gioco, si fornisce un'idea delle caratteristiche di colore, impurità (clarity) e taglio del campione dato.

Circa il taglio:



`summary(cut)`

Fair	Good	Very Good	Premium	Ideal
168	415	992	1100	1825

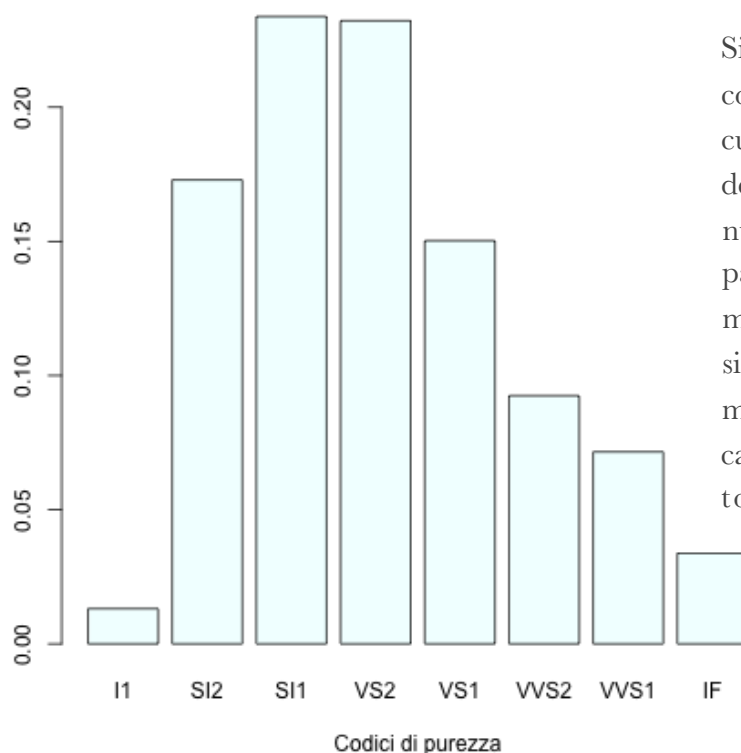
Si osserva dal grafico a torta che prevalgono fortemente diamanti che presentano la massima qualità possibile di taglio. Da una mera analisi descrittiva, tuttavia, operata su un campione peraltro non necessariamente rappresentativo della popolazione globale di diamanti lavorati, non ci sentiamo di affermare con certezza (almeno per ora) che si tratti della lavorazione “standard” per tali gemme.

Circa il grado di impurità:

`summary(clarity)`

I1	SI2	SI1	VS2	VS1	VVS2	VVS1	IF
59	778	1052	1045	676	416	322	152

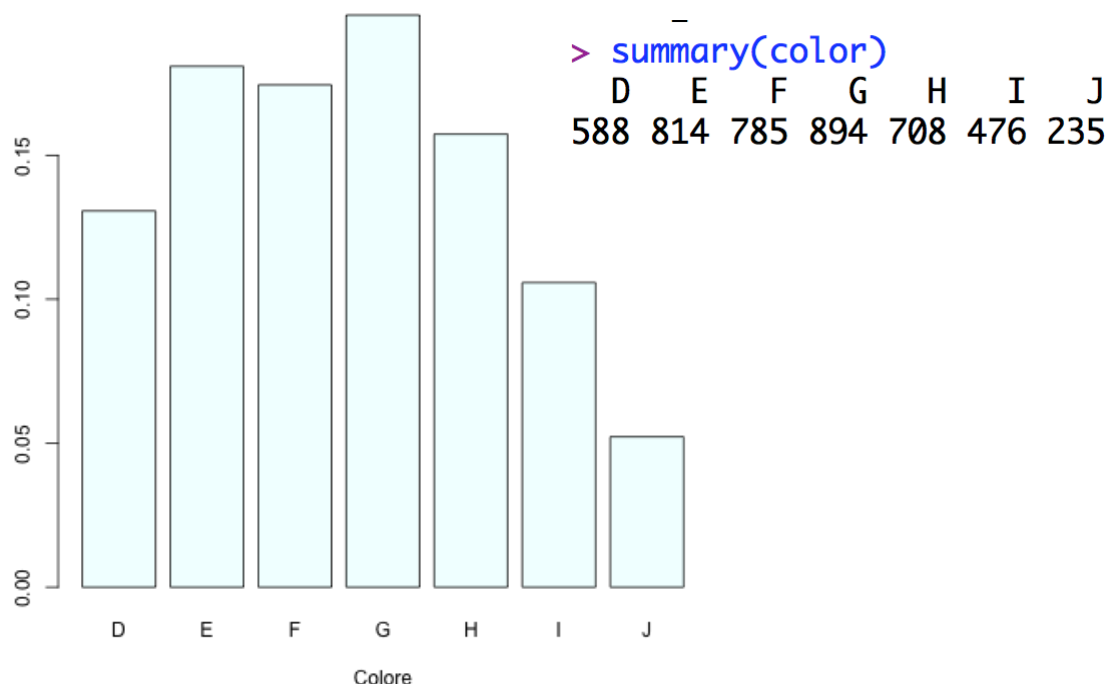
Proponiamo un bar-plot che ponga sull’asse delle ascisse, in ordine crescente di bontà, i codici di purezza, scelta a nostro parere più leggibile per un numero tanto alto di classi, di cui è fondamentale tenere conto della gerarchia:



Si osservi come, concordemente con l’assunto empirico secondo cui sono più rari gli esemplari denotati da presenza minima o nulla di impurità, la maggior parte dei diamanti cade nella metà sinistra del bar-plot, dove è sicuramente contenuta la mediana. In questo senso, il campione sembra rispecchiare la totalità della popolazione, sebbene con proporzioni forse diverse da quelle reali.

Circa il colore:

In tal caso, la distribuzione dei colori risulta piuttosto omogenea; si osserva solamente un minimo significativo per il codice “J”: trattandosi di una sfumatura che si allontana piuttosto significativamente dal “colorless” (“D”), per andare ad assumere, a tratti, una tonalità impercettibilmente gialla chiara, risulta essere poco significativa nell’ambito del mercato.

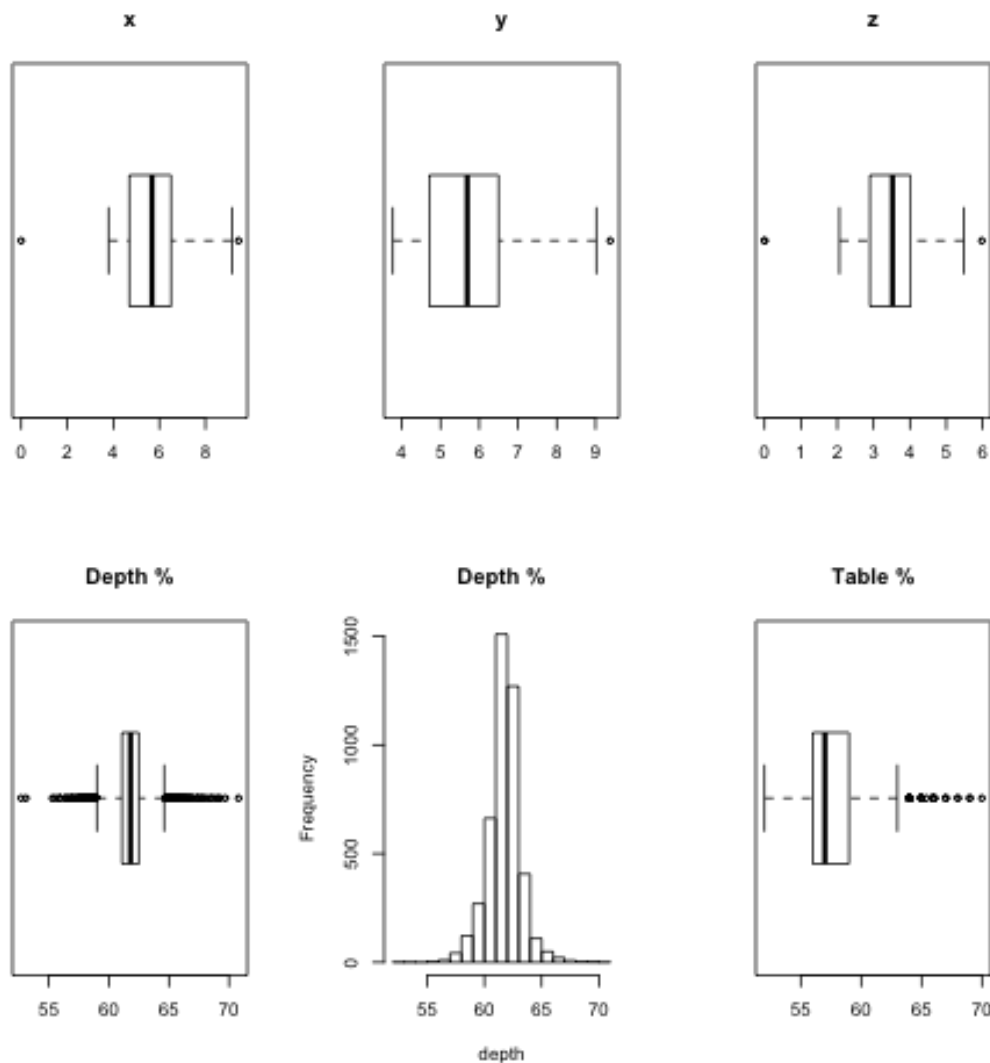


Forniamo ora, a carrellata, i box plot relativi agli attributi x, y, z, depth % e table % in modo da fornire almeno un’idea qualitativa della loro distribuzione, senza concentrarsi su ulteriori dettagli. Tale scelta è dettata, a nostro parere, dalla minore significatività dei suddetti caratteri ai fini del proseguimento della trattazione (test di ipotesi), anche se svolgono anch’essi un ruolo determinante come predittori nell’ambito della regressione.

Circa le 3 dimensioni, si evince una coda destra, per quanto non eccessivamente marcata, sebbene, per il resto, tali distribuzioni sono piuttosto regolari, con una mediana che cade quasi esattamente a metà del range interquantilico individuato dal rettangolo; segnaliamo una blandissima presenza di outlier.

La distribuzione di **depth %** risulta pressoché simmetrica, come testimoniato dall’istogramma riportato a lato del box plot e dalle due lunghe code di outliers; nonostante la simmetria, non si individua alcun carattere di gaussianità (p-value SW < 10⁻¹⁶).

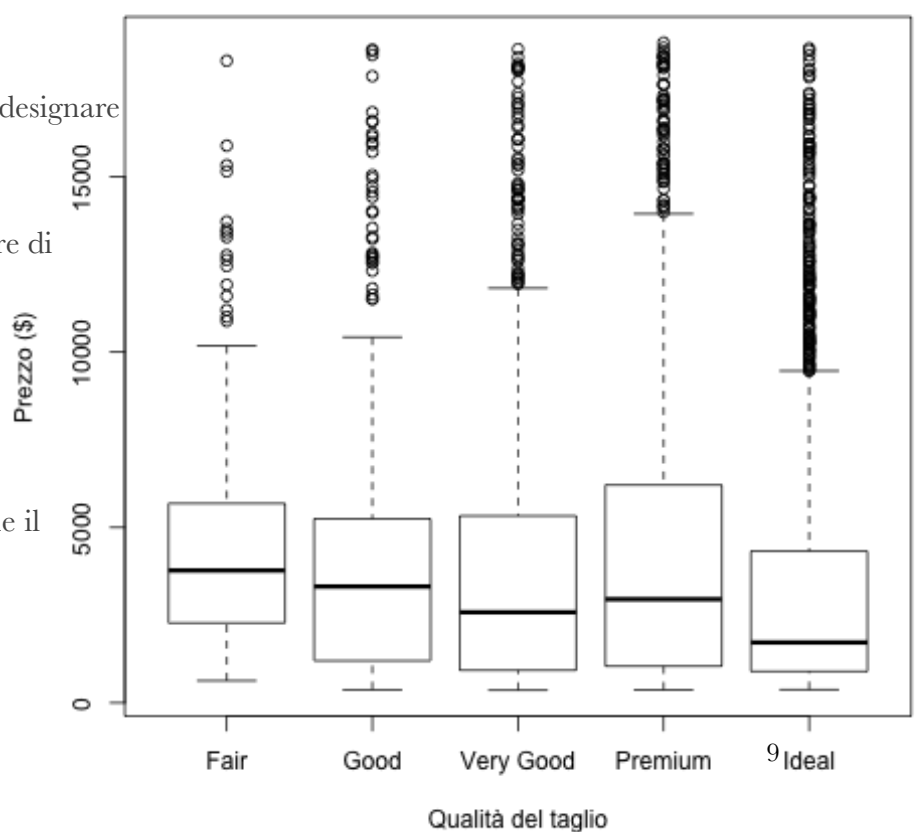
Il box plot di **table %** fa dedurre una distribuzione particolarmente concentrata verso sinistra (si veda la posizione della mediana), con una lunga coda verso destra, riconoscibile dalla presenza modesta di outliers.



Si vuole ora cominciare a dedurre qualche relazione, almeno qualitativamente, tra le variabili che definiscono un diamante: in primis, è interessante confrontare i prezzi di diamanti caratterizzati da qualità di taglio diverse (si veda il grafico sotto) in quanto, a meno delle caratteristiche intrinseche del prodotto grezzo, l'operazione di taglio è da ritenersi un processo laborioso che contribuisce, qualora ben eseguito, a determinare una lievitazione dei prezzi.

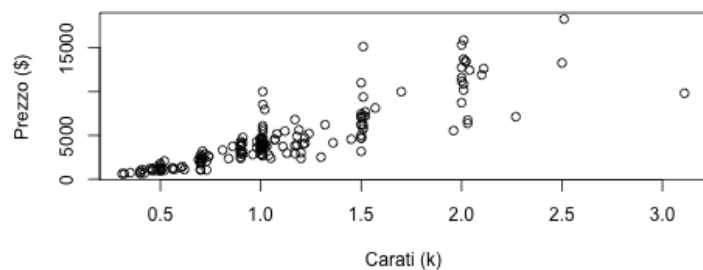
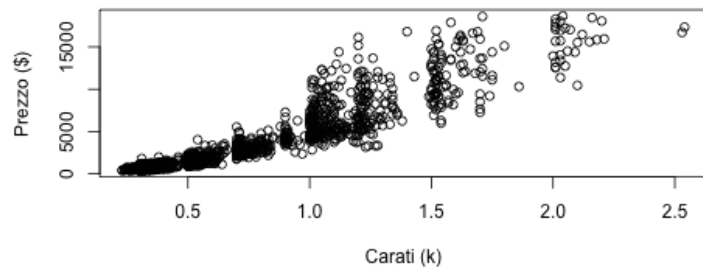
Si osserva tuttavia, che è difficile designare la qualità del taglio come unica discriminante tra prezzi: infatti, i box plot relativi alle diverse fatture di taglio ricoprono, grossomodo, lo stesso range di prezzo, con presenza significativa di outlier aventi la medesima estensione.

Paradosso a cui cercheremo di fornire risposta, ad esempio, è che il prezzo medio di un diamante di taglio "ideal" risulti di molto



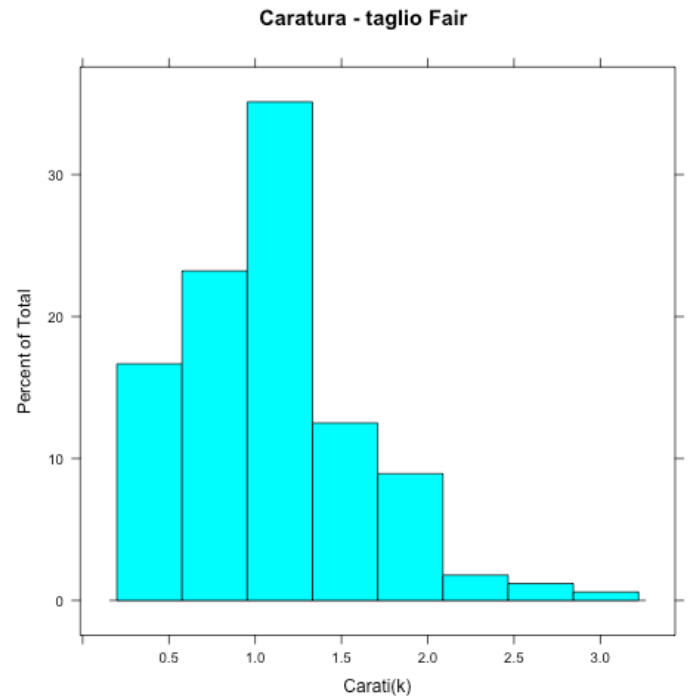
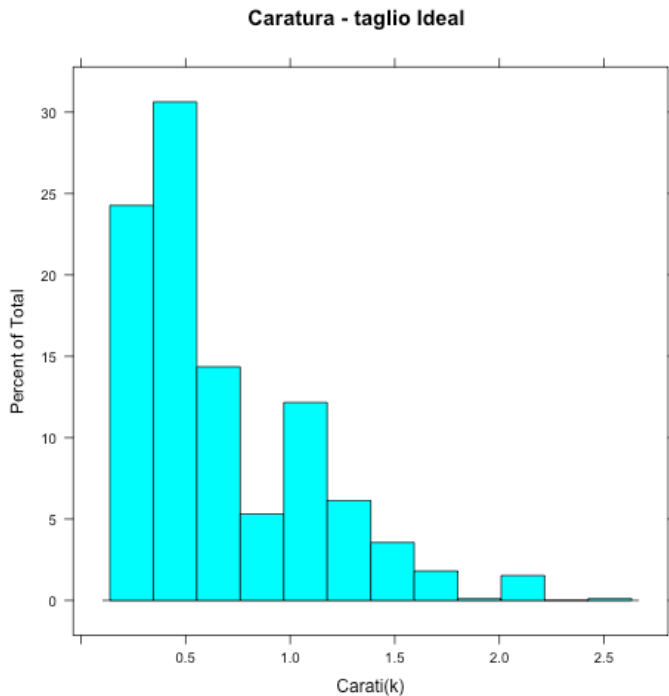
inferiore ad uno di taglio “fair” (si è calcolato, rispettivamente 3277.985\$ vs 4616.929\$) e lo stesso vale, come da grafico, per le rispettive mediane.

Motiviamo quanto osservato prendendo in considerazione i seguenti grafici, che evidenziano grossomodo una relazione lineare, seppur approssimativa, tra prezzo e caratura; ciò vale sia per il campione di esemplari di taglio “ideal” (sopra) che per il campione “fair” (sotto):



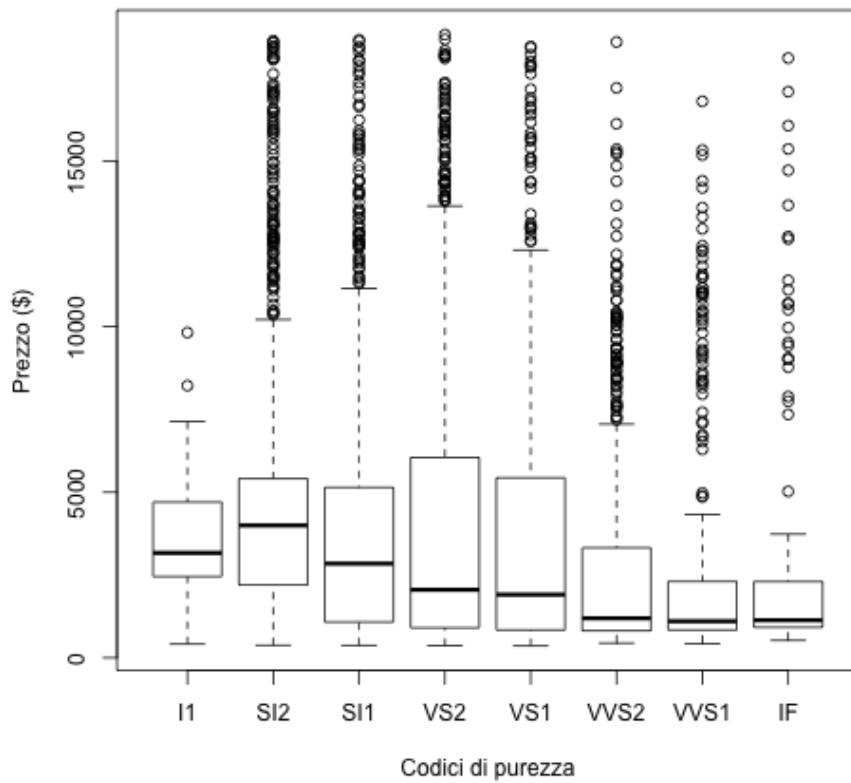
Consultiamo allora la distribuzione delle carature per ambo i campioni: qualora esse risultino differenti (in particolare, per campione “fair” si vorrebbe una maggiore concentrazione di carature verso la destra del grafico), allora la caratura è un fattore evidente in grado di spiegare uno scostamento di prezzi altrimenti ingiustificato.

Seguono le due distribuzioni:



Come atteso, il campione di diamanti “fair” presenta esemplari più massivi e abbiamo visto come la massa (caratura) concorra in maniera approssimativamente lineare all’incremento del prezzo.

Un ragionamento simile può essere portato avanti per una qualsiasi altra coppia di variabili (prezzo; ...); si prenda ad esempio il seguente grafico, che mostra l'interazione tra prezzo e grado di purezza del diamante:



Anche in questo caso, non si individuano scostamenti significativi tra i range di prezzo occupati dai box-plot: dunque si rende ancora più evidente come, per poter descrivere il prezzo di un diamante in funzione dei suoi attributi fisici, non sia sufficiente impiegare una sola variabile alla volta.

4. Test di ipotesi

Proponiamo alcuni test di ipotesi per rispondere a questioni inerenti il soggetto della nostra discussione:

1. Come è noto, accanto al mercato del diamante, se ne accosta un altro altrettanto remunerativo, quello dell'oro, che si configura come bene di rifugio tanto quanto il diamante.

Vogliamo, in particolare, verificare la veridicità (in particolare smentire) del seguente assunto, in cui ci siamo imbattuti durante la ricerca di informazioni:

“Il prezzo di un diamante di un solo carato è superiore rispetto a quello di 1kg di oro”

Prima di impostare un opportuno test di ipotesi, conviene chiarire le modalità con cui intendiamo approcciare tale assunto. Chiaramente, si dovranno considerare esemplari del dataset con una massa espressa in carati pari a 1k; inoltre, per “oro” intendiamo l'elemento (quasi) puro, adottato nella realizzazione di lingotti, piuttosto che quello comunemente adoperato in gioielleria/orologeria, che viene definito “usato”, in quanto mescolato con altri metalli a fornire una lega.

Prendiamo il prezzo dell'oro al grammo, da proporzionare rispetto ad 1kg nell'istante della produzione del documento, eliminando di fatto l'elemento di variabilità dovuto alle fluttuazioni giornaliere. Riscontriamo dunque, per 1kg di oro, $\mu_0 = 39920$ \$.

Circa la popolazione di diamanti da impiegare nel test, contiamo $n=138$ esemplari da esattamente 1k per i quali, il test di SW (si veda a lato) giunge alla conclusione forte che il campione (relativamente ai prezzi) non è gaussiano.

```
> shapiro.test(price[which(carat==1)])
```

Shapiro-Wilk normality test

```
data: price[which(carat == 1)]  
W = 0.84731, p-value = 1.198e-10
```

Per lo svolgimento di tale test di ipotesi, è necessaria l'applicazione del TLC (essendo il campione sufficientemente numeroso), che garantisce, in via approssimata, la gaussianità dello stimatore della media della distribuzione, diciamo \bar{X}_{138} (media campionaria) e conseguentemente della quantità pivotale usata nel test:

$$Z_0 = \frac{\bar{X}_{138} - \mu_0}{S_{138}} \sqrt{n}$$

dove S_{138} è la deviazione standard campionaria del campione di prezzi.

Sia dunque:

$$H_0: \mu = \mu_0$$

$$H_1: \mu < \mu_0$$

con statistica test Z_0 sopra riportata e $RC = Z_0 < -z_{1-\alpha}$.

Dall'elaborazione, il p-value α_0 del test risulta $< 2.2 * 10^{-16}$; segue pertanto, per un valore così basso, che vi è fortissima evidenza empirica per la validità di H_1 . La conclusione a cui siamo giunti è evidentemente forte: possiamo decretare con ottima sicurezza che l'enunciato riscontrato durante la nostra navigazione in rete è falso.

2. Riprendendo un'osservazione vista nel par. 2, circa la qualità del taglio, vogliamo dimostrare, attraverso opportuno test di ipotesi, il seguente asserto, di fatto deducendo una qualità della popolazione globale a partire dal campione dato:

“Più del 40% dei diamanti lavorati presenta un taglio di tipo “ideal” “

L'applicazione del seguente test di ipotesi per campione Bernoulliano numeroso risulta legittima essendo la numerosità del campione $n=4500 > 30$; l'i-esima v.a. (iid rispetto alle altre) del campione Bernoulliano assume valore 1 qualora il taglio sia “ideal” e 0 altrimenti.

Sia:

$$H_0: p \leq p_0 = 0.4$$

$$H_1: p > p_0$$

Ci si serve della quantità pivotale $Z_0 = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)}} \sqrt{n}$

dove $\hat{p} = \frac{n_{ideal}}{n} = 0.41$, proporzione campionaria degli esemplari “ideal”.

Con $RC = Z_0 > z_{1-\alpha}$

Riportiamo di seguito l'esito del test, ponendo particolare attenzione sul p-value:

1-sample proportions test without continuity correction

```
data: 1825 out of 4500, null probability 0.4
X-squared = 0.5787, df = 1, p-value = 0.2234
alternative hypothesis: true p is greater than 0.4
95 percent confidence interval:
 0.3935765 1.0000000
sample estimates:
      p
0.4055556
```

Per il valore di p-value estratto, non è presente sufficiente evidenza empirica per accettare H_1 : sarebbe eccessiva la probabilità di commettere errore di I specie qualora si faccia il contrario.

In definitiva, meno del 40% dei diamanti lavorati presentano qualità del taglio “ideal”, contrariamente a quanto mostrato dall’analisi descrittiva del nostro campione, dove figurava invece una $\hat{p} = 0.41$.

Quanto segue è un IC(0.95) bilatero circa la proporzione di esemplari di taglio “ideal”,

$$(\mathbf{0.3956299 : 0.4243701})$$

che si pone numericamente a cavallo tra quanto espresso dalle ipotesi H_0 ed H_1 del test precedente: individuato $p_0 = 0.40$ all’interno del dato intervallo, la regione alla destra di tale valore pertiene alla proposizione sostenuta dall’ipotesi alternativa (essendo $p > 0.4$), viceversa, alla sua sinistra si colloca la regione relativa all’ipotesi nulla ($p \leq 0.4$).

Supponendo ora di accettare la conclusione debole a cui si giunge dall’analisi del p-value del precedente test, e dunque accettando la validità di H_0 , si è in grado di risalire ad una restrizione dell’intervallo di confidenza sopra calcolato, del tipo $(0.3956299 : 0.40]$.

5. Regressione lineare multipla

Si vuole ora dedurre un modello di regressione lineare che possa spiegare in maniera sufficiente la variabilità del prezzo, variabile dipendente (Y), in funzione di tutti o solo alcuni dei predittori finora presentati (variabili dipendenti x_1, x_2 , ecc...).

Una prima iterazione del modello è stata, per semplicità, la legge lineare multipla che include tutti i predittori: mostriamo di seguito l'esito dell'elaborazione del calcolatore, senza riportare esplicitamente per esteso la legge, vista la presenza di numerosi predittori categorici che complicano la scrittura.

Call:

```
lm(formula = price ~ x + y + z + depth + table + color + clarity +  
    cut + carat)
```

Residuals:

Min	1Q	Median	3Q	Max
-10686.9	-579.2	-171.2	365.0	9936.5

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4030.737	1612.402	2.500	0.012461 *
x	-988.367	181.644	-5.441	5.57e-08 ***
y	-525.985	146.006	-3.602	0.000319 ***
z	428.823	238.890	1.795	0.072711 .
depth	-87.584	19.909	-4.399	1.11e-05 ***
table	-19.973	9.882	-2.021	0.043316 *
colorE	-167.102	60.283	-2.772	0.005595 **
colorF	-342.001	61.145	-5.593	2.36e-08 ***
colorG	-509.884	60.165	-8.475	< 2e-16 ***
colorH	-1006.837	63.359	-15.891	< 2e-16 ***
colorI	-1462.623	70.343	-20.793	< 2e-16 ***
colorJ	-2340.837	88.210	-26.537	< 2e-16 ***
clarityIF	5177.542	176.286	29.370	< 2e-16 ***
claritySI1	3408.537	152.263	22.386	< 2e-16 ***
claritySI2	2491.467	152.748	16.311	< 2e-16 ***
clarityVS1	4329.930	155.113	27.915	< 2e-16 ***
clarityVS2	4079.464	152.639	26.726	< 2e-16 ***
clarityVVS1	4887.251	162.826	30.015	< 2e-16 ***
clarityVVS2	4656.853	159.779	29.146	< 2e-16 ***
cutGood	799.012	107.789	7.413	1.47e-13 ***
cutIdeal	1103.040	107.197	10.290	< 2e-16 ***
cutPremium	1029.231	102.530	10.038	< 2e-16 ***
cutVery Good	1021.267	102.992	9.916	< 2e-16 ***
carat	11882.279	180.563	65.807	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1111 on 4476 degrees of freedom

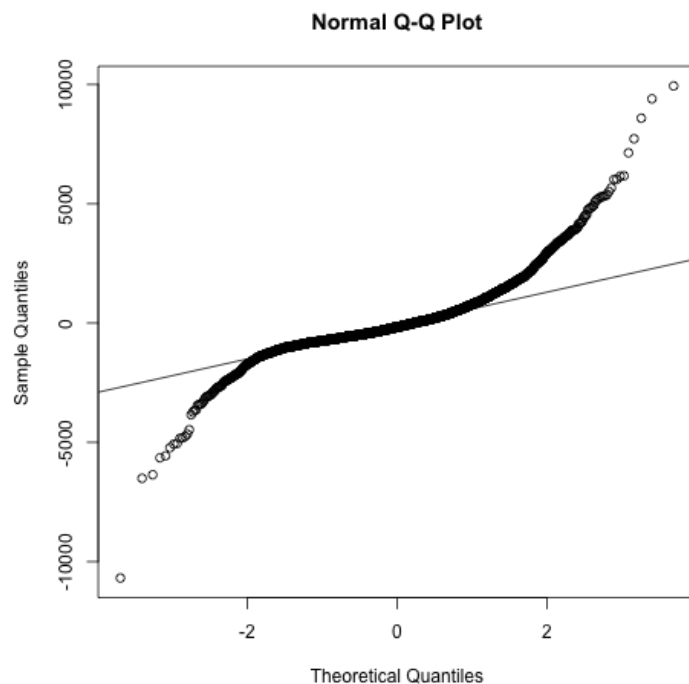
Multiple R-squared: 0.9217, Adjusted R-squared: 0.9213

F-statistic: 2291 on 23 and 4476 DF, p-value: < 2.2e-16

Ad un primo sguardo, il modello risulta sicuramente migliorabile: per quanto il coefficiente di determinazione R^2_{adjusted} risulti buono, tanto che il modello lineare spiega il 92.13% di

variabilità del prezzo, mostriamo alcune perplessità sulla significatività del predittore z , che mostra un $p\text{-value}=7.2711\%$ non sufficientemente basso.

La vera problematica inerente al modello in questione riguarda, tuttavia, il mancato rispetto dell'ipotesi di gaussianità dei residui; ciò si palesa dall'osservazione di un opportuno Q-Q plot in cui mettiamo a confronto i quantili teorici della normale standard con quelli provenienti dalle realizzazioni del dataset. Segue il grafico:



Sicuramente, non si necessita di un test di Shapiro-Wilks per constatare che occorre individuare un modello più adeguato: a tale scopo, ci è d'aiuto osservare la presenza di due code, una sinistra e l'altra, ben più pesante, a destra. L'andamento dei quantili nella parte centrale del grafico è, invece, buono, con la curva ben aderente alla retta congiungente il primo con il terzo quantile della normale standard.

Riteniamo allora opportuno operare una trasformazione logaritmica dei dati, che possa accomodare la pesante coda di destra e rendere un Q-Q plot che rispetti l'ipotesi di gaussianità dei residui.

Presentiamo allora il seguente modello, di seguito commentato in maniera più estensiva del primo; come prima, per semplicità abbiamo incluso tutti i predittori, per poi constatare se vi fosse la necessità di sfoltirli. Si è messo in relazione il logaritmo naturale del prezzo, $\log(\text{price})$, con tutti i predittori a disposizione, senza applicare alcuna funzione componente su questi ultimi. Ancora una volta, qualora se ne fosse mostrata la necessità, sarebbe stato opportuno tentare ulteriori iterazioni del modello modificando anche la forma con cui compaiono le variabili indipendenti nell'espressione.

```
Call:
lm(formula = I(log(price))) ~ x + y + z + color + clarity + carat +
  depth + cut + table)

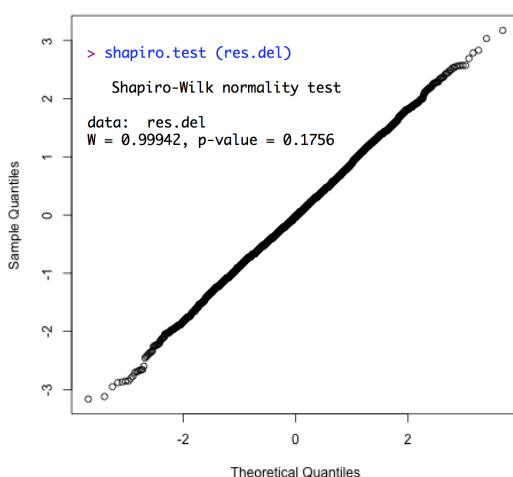
Residuals:
    Min       1Q   Median       3Q      Max
-0.47801 -0.08663 -0.00378  0.08578  0.46319

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.505141   0.331654 -10.569 < 2e-16 ***
x             0.683105   0.049813  13.713 < 2e-16 ***
y             0.497152   0.050791   9.788 < 2e-16 ***
z             0.377769   0.078066   4.839 1.35e-06 ***
colorE       -0.050593   0.007168  -7.058 1.95e-12 ***
colorF       -0.103410   0.007278 -14.209 < 2e-16 ***
colorG       -0.156883   0.007159 -21.915 < 2e-16 ***
colorH       -0.248831   0.007539 -33.007 < 2e-16 ***
colorI       -0.368116   0.008365 -44.009 < 2e-16 ***
colorJ       -0.498121   0.010493 -47.473 < 2e-16 ***
clarityIF     1.051494   0.021267  49.441 < 2e-16 ***
claritySI1    0.536912   0.018314  29.316 < 2e-16 ***
claritySI2    0.376084   0.018364  20.479 < 2e-16 ***
clarityVS1    0.736815   0.018681  39.442 < 2e-16 ***
clarityVS2    0.687723   0.018390  37.396 < 2e-16 ***
clarityVVS1   0.965124   0.019590  49.266 < 2e-16 ***
clarityVVS2   0.877320   0.019235  45.611 < 2e-16 ***
carat        -1.048875   0.022086 -47.490 < 2e-16 ***
depth         0.046460   0.005015   9.264 < 2e-16 ***
cutGood       0.103164   0.013048   7.906 3.32e-15 ***
cutIdeal      0.174137   0.012883  13.517 < 2e-16 ***
cutPremium    0.154604   0.012235  12.636 < 2e-16 ***
cutVery Good  0.134388   0.012545  10.713 < 2e-16 ***
table         0.009190   0.001176   7.812 6.95e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1318 on 4463 degrees of freedom
Multiple R-squared:  0.9828, Adjusted R-squared:  0.9827
F-statistic: 1.11e+04 on 23 and 4463 DF,  p-value: < 2.2e-16
```

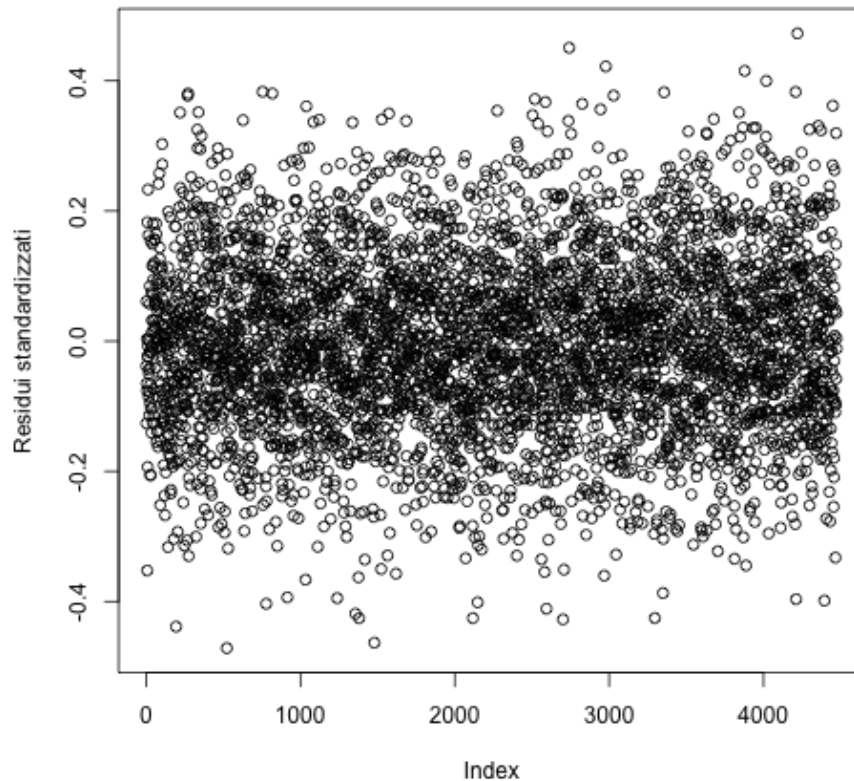
Prima di commentare la significatività del modello, peraltro migliorata, ci soffermiamo ancora una volta sulle ipotesi di applicabilità dello stesso, portando a giustificazione il grafico quantile-quantile per i residui, con annesso p-value del test di SW:

Il miglioramento è evidente: per quanto il Q-Q plot non sia perfetto sulle code, la tendenza osservata in precedenza si è ridotta drasticamente e



l'andamento della curva nella parte centrale è praticamente di tipo rettilineo; ci saremmo aspettati un valore di p-value SW un poco più alto, sebbene il valore attuale è sufficiente per non escludere la possibilità che il campione di residui in questione sia approssimativamente gaussiano.

Vogliamo ora verificare l'ipotesi di indipendenza relativa ai residui: a tal scopo presentiamo di seguito lo scatterplot dei residui standardizzati:



Ci possiamo ritenere soddisfatti anche in questo caso; si osservi che i residui si distribuiscono a nuvola attorno all'asse orizzontale 0, senza individuare alcuno specifico pattern: confermiamo le ipotesi di omoschedasticità e indipendenza dei residui.

Inoltre, non si individua alcun punto dello scatterplot giacere al di fuori dell'intervallo $[-2,2]$.

Si noti che tali considerazioni vengono fatte a seguito della rimozione di un numero di osservazioni pari a 13 (lo si conferma attraverso il numero di gradi di libertà nel riassunto del modello di regressione, poco sopra), che sono risultate come outliers ai fini del modello, determinando nel nostro caso un crollo nel valore del p-value di SW. Abbiamo ritenuto che tali valori non fossero degni di nota e nemmeno utili a spiegare il modello in particolari condizioni limite, dunque abbiamo proceduto alla loro eliminazione.

Si vuole ora commentare la bontà del modello nei termini della sua capacità di spiegare la variabilità della variabile dipendente Y a partire dai predittori dati e della sua significatività globale.

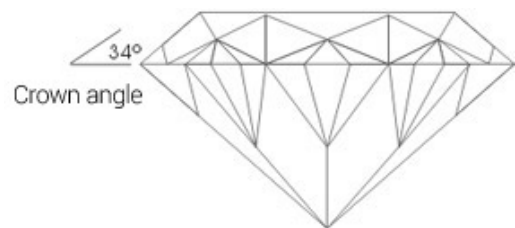
Circa il primo punto, segnaliamo un $R^2_{\text{adjusted}}=0.9827$, vicinissimo a 1 pertanto ottimale e addirittura molto simile al valore di $R^2_{\text{multiple}}=0.9828$: ciò implica che, nell'includere un

numero così elevato di predittori, non siamo caduti nella pratica dell'overfitting. In caso contrario, infatti, si sarebbe registrato un forte scostamento tra i due valori di R squared, essendo quello “aggiustato” pensato proprio per punire l'abuso di predittori.

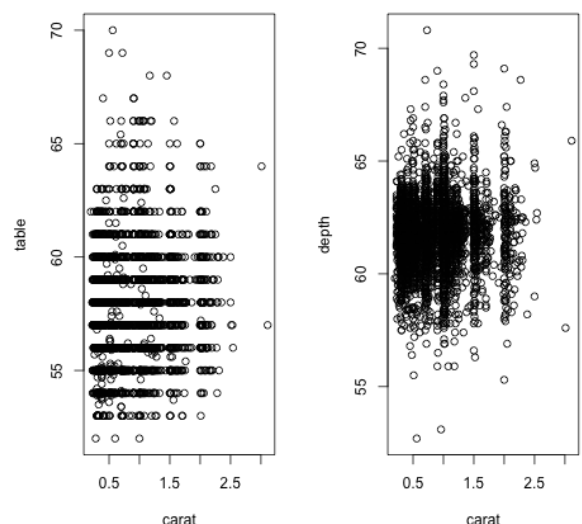
Per quanto riguarda la significatività globale, siamo rassicurati dai p-value relativi ai test di significatività singoli di tutte le variabili in questione, bassissimi, corredati da un p-value F-statistic altrettanto basso, che conferma l'adeguatezza nella scelta dei predittori e ci rassicura ulteriormente sulla bontà di R^2 .

Un motivo di dubbio può essere derivante dalla presenza di un numero di predittori superiore a 3 per descrivere geometricamente il diamante: sono state incluse x, y, z, depth% e table%, senza che l'algoritmo rilevasse alcuna ridondanza. In particolare, è legittimo domandarsi come mai includere depth% e table%, funzioni delle stesse x, y e z senza pagarne in termini di significatività globale. Una risposta plausibile risiede nella geometria complessa del singolo diamante, come testimoniato dalle figure nel par 2: qualora la sua forma fosse stata un parallelepipedo, allora sarebbero state sufficienti le sole 3 dimensioni per poterlo descrivere totalmente ed univocamente. Al contrario forme più complesse beneficiano di un numero sempre crescente di grandezze geometriche (purché non siano semplicemente combinazione lineare l'una dell'altra) che le descrivano, in quanto caratterizzate da un numero di gradi di libertà decisamente superiore a 3.

Possiamo supporre ad esempio, a partire da tale riflessione, che, qualora fossimo in possesso dei dati relativi all'angolo di corona (si veda figura), la percentuale di variabilità spiegata dal modello della Y aumenterebbe.



Allo stesso modo, notiamo l'indipendenza tra depth% (o table%) rispetto a carat: ciò è evidente tenendo conto che i due predittori “geometrici” sono forniti in forma percentuale e dunque slegati rispetto ad una grandezza di massa, come esplicitato dalla coppia di grafici a lato.






Consideriamo anche la relazione $x*y*z$ - carat o semplicemente x - carat (e anche y - carat, z - carat), ovviamente lineare e cubica rispettivamente (confermato da opportuno plot, omissis), in quanto trattasi di interazioni tra volume / singola dimensione e massa, che potrebbero aggiungere un fattore di ridondanza nel modello vista la dipendenza tra predittori.

Tale problema, a seguito dell'elaborazione software non si pone più: infatti, pur sperimentando altri modelli di tipo logaritmico privati di una (o alcune) delle grandezze x, y, z e *carat*, non sembra esserne uno in grado di giustificare altrettanto bene la variabilità dei prezzi; si sono ottenuti, infatti, risultati ove sia il coefficiente di determinazione, sia la significatività globale sono peggiorati.

Proviamo infine a mettere alla prova il modello realizzato all'atto pratico:

si è scelto su un sito referenziato di vendita online di diamanti, un esemplare dalle seguenti caratteristiche.

	Shape	Carat	Color	Clarity	Cut	Depth	Table	Polish	Symmetry	Measurement	Cert	Price
	 Round	0.50	F	SI2	Very good	64.1	54	Excellent	Very Good	5.00X5.03X3.21	GIA	US\$1,040

Notiamo che il modello è applicabile in quanto sono noti tutti i dati in input necessari (trascuriamo tuttavia, essendo assenti nel nostro modello, “polish” e “simmetry”, mentre “cert” indica che quanto sostenuto nello specchietto informativo è certificato dalla GIA, massima autorità nell'ambito; supponiamo che quest'ultimo aspetto non concorra all'aumento del prezzo, mentre gli altri due potrebbero giustificare eventuali inesattezze della seguente stima).

Ricaviamo un opportuno intervallo di predizione di livello $\gamma = 0.90$ e verifichiamo se il prezzo effettivo, indicato sopra, vi cada all'interno; in tal modo potremo constatare, seppur grossolanamente, la veridicità del modello.

Dall'elaborazione risulta un $IP_{\text{price}}(0.90) =$

```
> predict_0.90
[1] 900.671 1390.982
```

che è tale per cui il prezzo effettivo di 1040 \$ vi appartiene. Si tratta di un intervallo piuttosto ampio, ma abbastanza plausibile per la deduzione del prezzo del dato esemplare, anche considerando una politica di prezzo che può differire da rivenditore a rivenditore. Infine, tenendo conto della vicinanza (in quanto a valori assunti dai diversi attributi) dell'osservazione precedente a una generica realizzazione con la quale è stato costruito il modello, siamo sufficientemente confidenti della plausibilità della previsione.

6. Bibliografia e risorse

- Data-frame reperito presso: <https://www.kaggle.com/shivam2503/diamonds>
- Informazioni tecniche su terminologia e classificazione diamanti: <https://www.gia.edu>
- Supporto computazionale: <https://www.r-project.org>