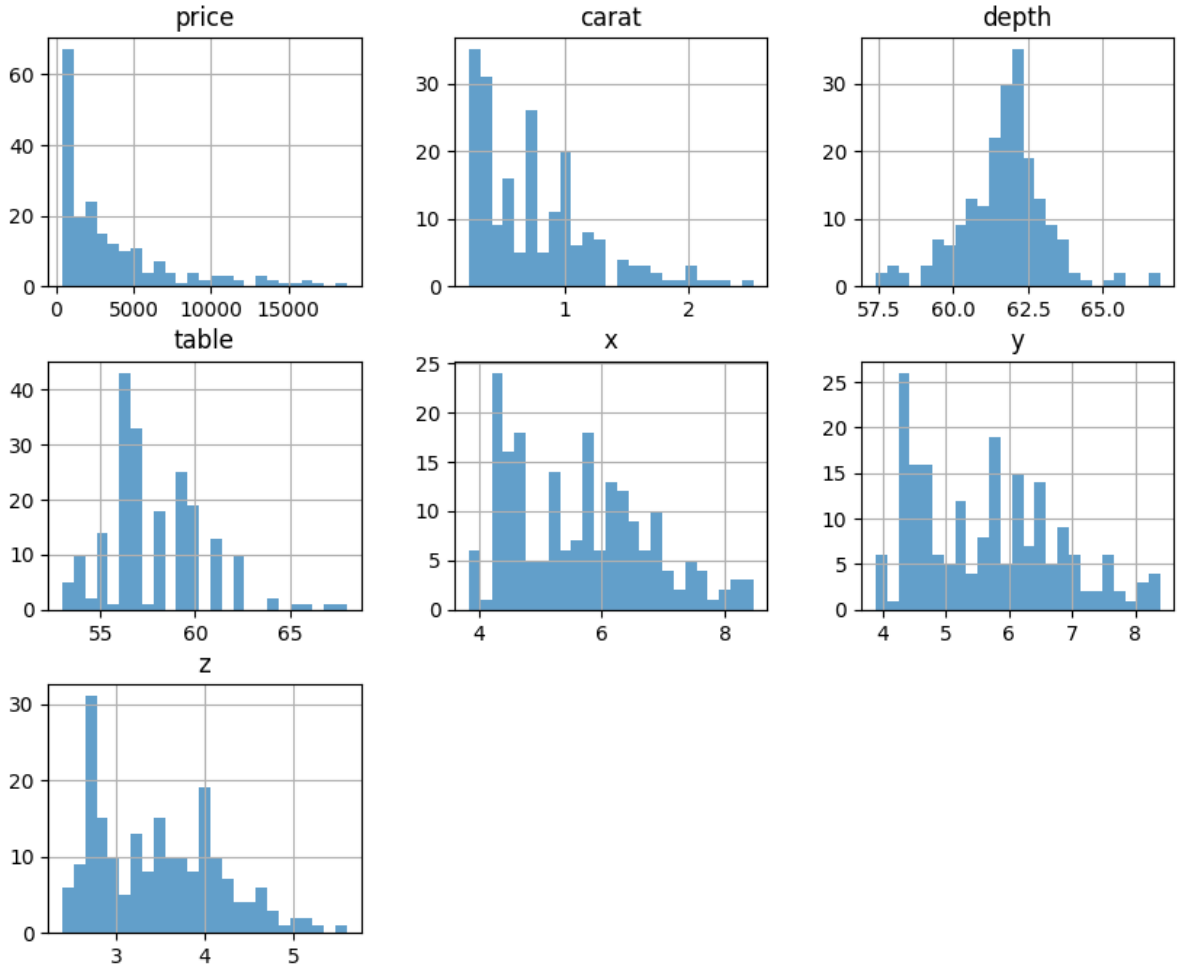
Semestrální práce SP2

Libor Pezinek

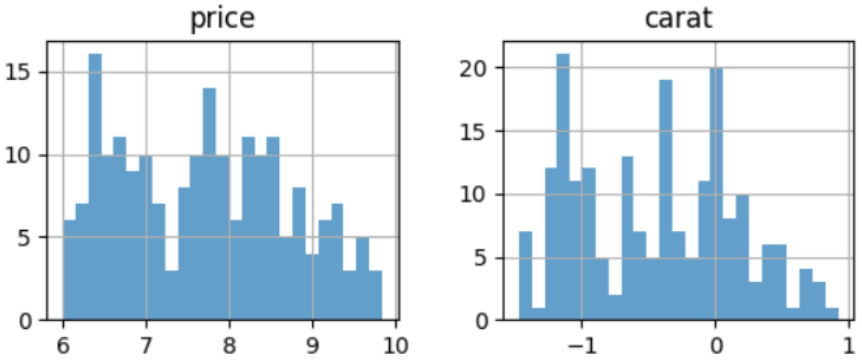
248222

1 Grafická analýza

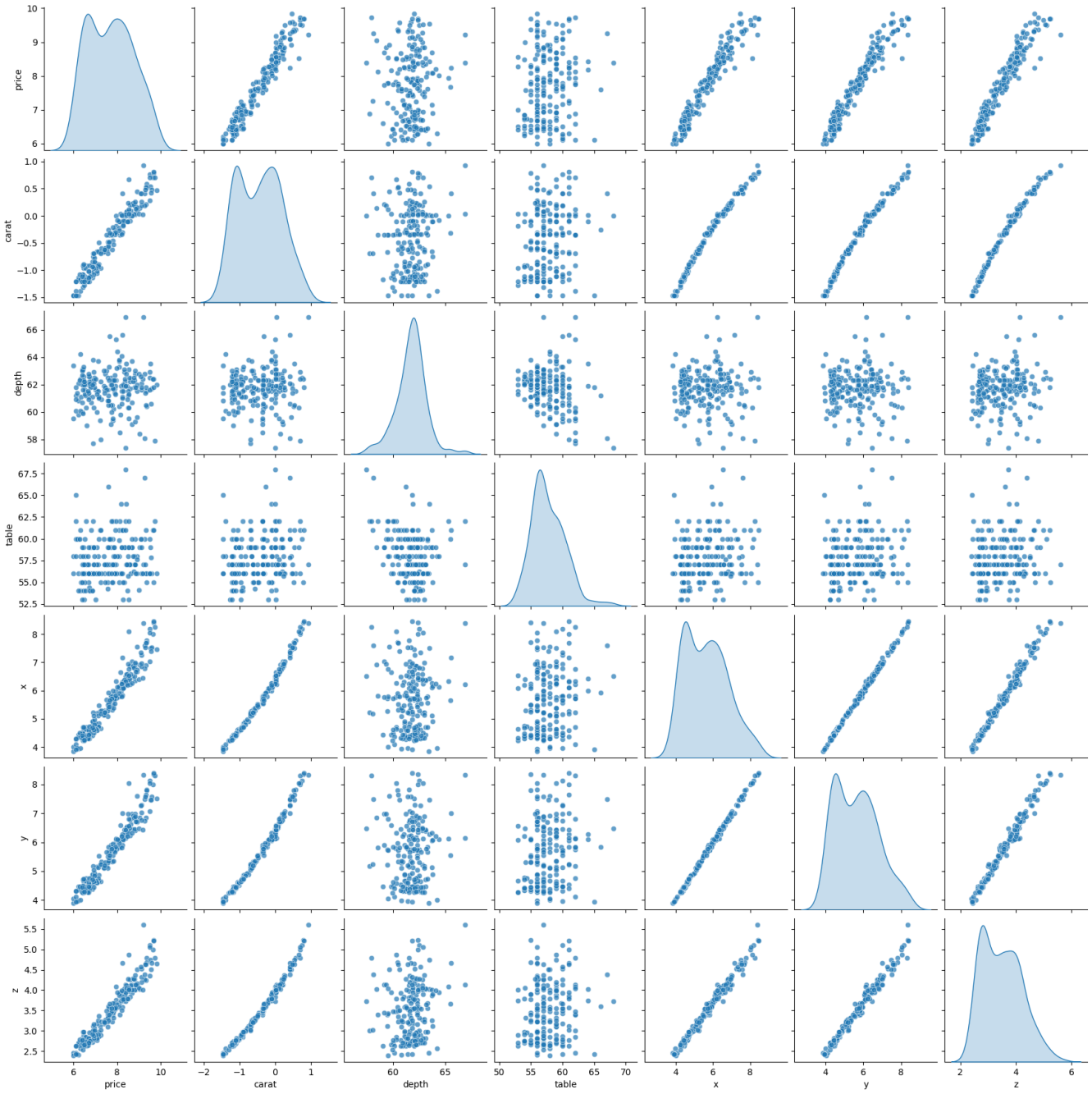
Histogramy spojitých proměnných



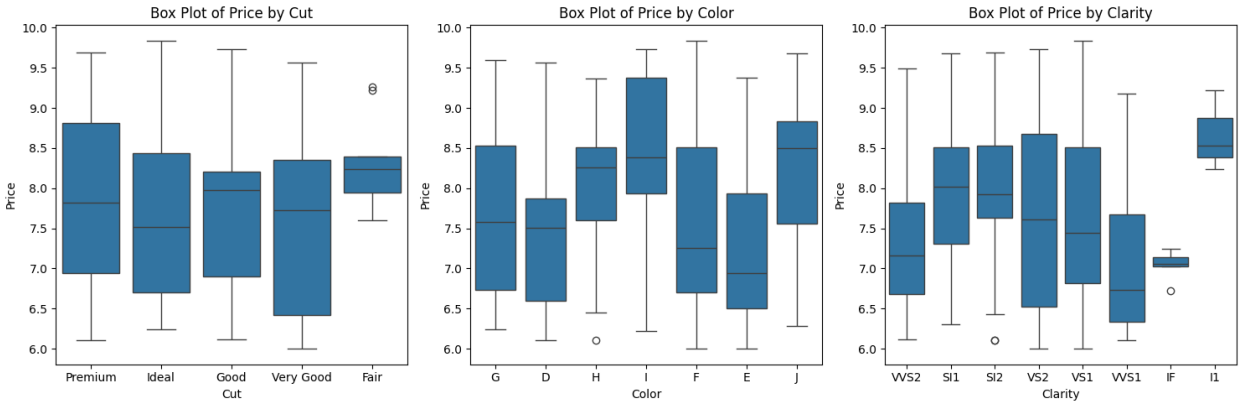
Spojité proměnné **price** a **carat** nemají normální rozdělení. Z tohoto důvodu tato data transformujeme, abychom je přiblížili normálnímu rozdělení. Transformaci provedeme logaritmováním.



Bodové grafy



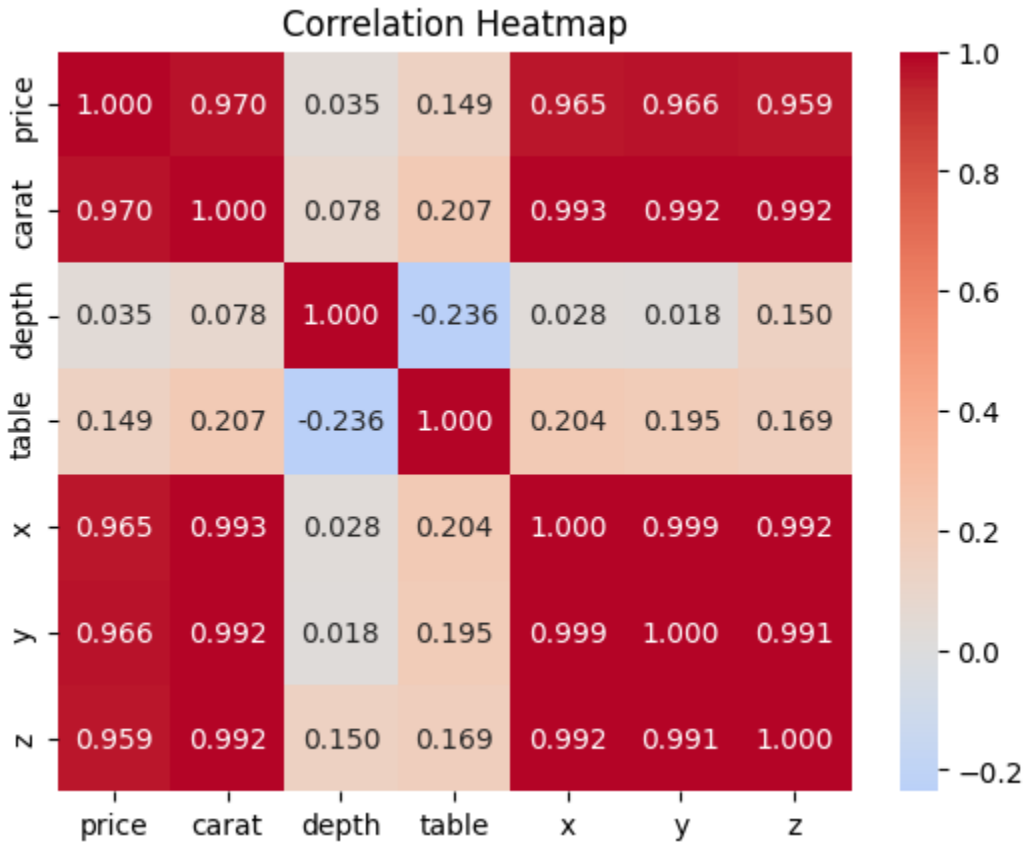
Krabicové diagramy proměnné price a kategoriálních proměnných



V diagramech je možné pozorovat výskyt odlehlých hodnot, které mohou způsobit nepřesnosti v modelu. Vzhledem k jejich nízkému počtu však bude jejich vliv minimální.

2 Korelační analýza proměnných

Výběrové korelace dvojic spojitých proměnných:



Testy hypotézy: Korelace je nulová

$\alpha = 0.05$

```
Proměnná 1| Proměnná 2| Korelace| p-hodnota| Interpretace
price| carat|  0.9700|  0.0000|  Zamítáme nulovou hypotézu, existuje korelace mezi proměnnými.
price| depth|  0.0351|  0.6219|  Nezamítáme nulovou hypotézu, korelace je nulová.
price| table|  0.1492|  0.0350|  Zamítáme nulovou hypotézu, existuje korelace mezi proměnnými.
price| x|  0.9649|  0.0000|  Zamítáme nulovou hypotézu, existuje korelace mezi proměnnými.
price| y|  0.9664|  0.0000|  Zamítáme nulovou hypotézu, existuje korelace mezi proměnnými.
price| z|  0.9588|  0.0000|  Zamítáme nulovou hypotézu, existuje korelace mezi proměnnými.
carat| depth|  0.0783|  0.2702|  Nezamítáme nulovou hypotézu, korelace je nulová.
carat| table|  0.2074|  0.0032|  Zamítáme nulovou hypotézu, existuje korelace mezi proměnnými.
carat| x|  0.9930|  0.0000|  Zamítáme nulovou hypotézu, existuje korelace mezi proměnnými.
carat| y|  0.9923|  0.0000|  Zamítáme nulovou hypotézu, existuje korelace mezi proměnnými.
carat| z|  0.9917|  0.0000|  Zamítáme nulovou hypotézu, existuje korelace mezi proměnnými.
depth| table|  -0.2362|  0.0008|  Zamítáme nulovou hypotézu, existuje korelace mezi proměnnými.
depth| x|  0.0281|  0.6929|  Nezamítáme nulovou hypotézu, korelace je nulová.
depth| y|  0.0183|  0.7966|  Nezamítáme nulovou hypotézu, korelace je nulová.
depth| z|  0.1496|  0.0345|  Zamítáme nulovou hypotézu, existuje korelace mezi proměnnými.
table| x|  0.2041|  0.0037|  Zamítáme nulovou hypotézu, existuje korelace mezi proměnnými.
table| y|  0.1954|  0.0056|  Zamítáme nulovou hypotézu, existuje korelace mezi proměnnými.
table| z|  0.1694|  0.0165|  Zamítáme nulovou hypotézu, existuje korelace mezi proměnnými.
x| y|  0.9989|  0.0000|  Zamítáme nulovou hypotézu, existuje korelace mezi proměnnými.
x| z|  0.9920|  0.0000|  Zamítáme nulovou hypotézu, existuje korelace mezi proměnnými.
y| z|  0.9907|  0.0000|  Zamítáme nulovou hypotézu, existuje korelace mezi proměnnými.
```

Výběrový parciální korelační koeficient price a table při fixních hodnotách carat a depth:

test hypotézy, že je tento parciální korelační koeficient nulový

```
Parciální korelační koeficient price a table při daném carat a depth: -0.2757
t-statistika: -4.0153
p-hodnota: 0.0001
Interpretace: Zamítnutí nulové hypotézy: Parciální korelační koeficient je nenulový.
```

3 Regresní analýza proměnné price v závislosti na ostatních proměnných

Dostatečný submodel jsme určili dopřednou selekcí:

Postupným přidáváním parametrů

```
Added carat, R^2: 0.9409, Adjusted R^2: 0.9406
Added clarity_SI2, R^2: 0.9483, Adjusted R^2: 0.9478
Added clarity_SI1, R^2: 0.9524, Adjusted R^2: 0.9517
Added color_I, R^2: 0.9548, Adjusted R^2: 0.9538
Added color_J, R^2: 0.9576, Adjusted R^2: 0.9565
Added color_H, R^2: 0.9594, Adjusted R^2: 0.9581
Added cut_Ideal, R^2: 0.9611, Adjusted R^2: 0.9597
Added depth, R^2: 0.9630, Adjusted R^2: 0.9614
Added clarity_VVS1, R^2: 0.9646, Adjusted R^2: 0.9630
Added clarity_VVS2, R^2: 0.9668, Adjusted R^2: 0.9650
Added clarity_IF, R^2: 0.9685, Adjusted R^2: 0.9666
Added color_G, R^2: 0.9702, Adjusted R^2: 0.9682
Added clarity_VS1, R^2: 0.9716, Adjusted R^2: 0.9696
Added clarity_VS2, R^2: 0.9857, Adjusted R^2: 0.9846
Added x, R^2: 0.9864, Adjusted R^2: 0.9853
Added cut_Premium, R^2: 0.9870, Adjusted R^2: 0.9858
Added cut_Very Good, R^2: 0.9875, Adjusted R^2: 0.9863
```

Tabulka regresních koeficientů a 95% intervalů spolehlivosti:

	coef	std err	t	P>|t|	[0.025	0.975]
carat	0.6762	0.082	8.269	0.000	0.515	0.838
clarity_SI2	1.1066	0.080	13.813	0.000	0.949	1.265
clarity_SI1	1.2701	0.080	15.816	0.000	1.112	1.429
color_I	-0.3949	0.036	-11.020	0.000	-0.466	-0.324
color_J	-0.4662	0.053	-8.717	0.000	-0.572	-0.361
color_H	-0.1286	0.031	-4.133	0.000	-0.190	-0.067
cut_Ideal	0.1023	0.030	3.430	0.001	0.043	0.161
depth	0.0477	0.004	11.165	0.000	0.039	0.056
clarity_VVS1	1.6339	0.089	18.292	0.000	1.458	1.810
clarity_VVS2	1.5756	0.085	18.596	0.000	1.408	1.743
clarity_IF	1.7048	0.102	16.773	0.000	1.504	1.905
color_G	-0.1045	0.025	-4.118	0.000	-0.155	-0.054
clarity_VS1	1.4939	0.083	18.072	0.000	1.331	1.657
clarity_VS2	1.3948	0.081	17.125	0.000	1.234	1.556
x	0.6585	0.042	15.616	0.000	0.575	0.742
cut_Premium	0.1008	0.032	3.104	0.002	0.037	0.165
cut_Very Good	0.0821	0.035	2.371	0.019	0.014	0.150

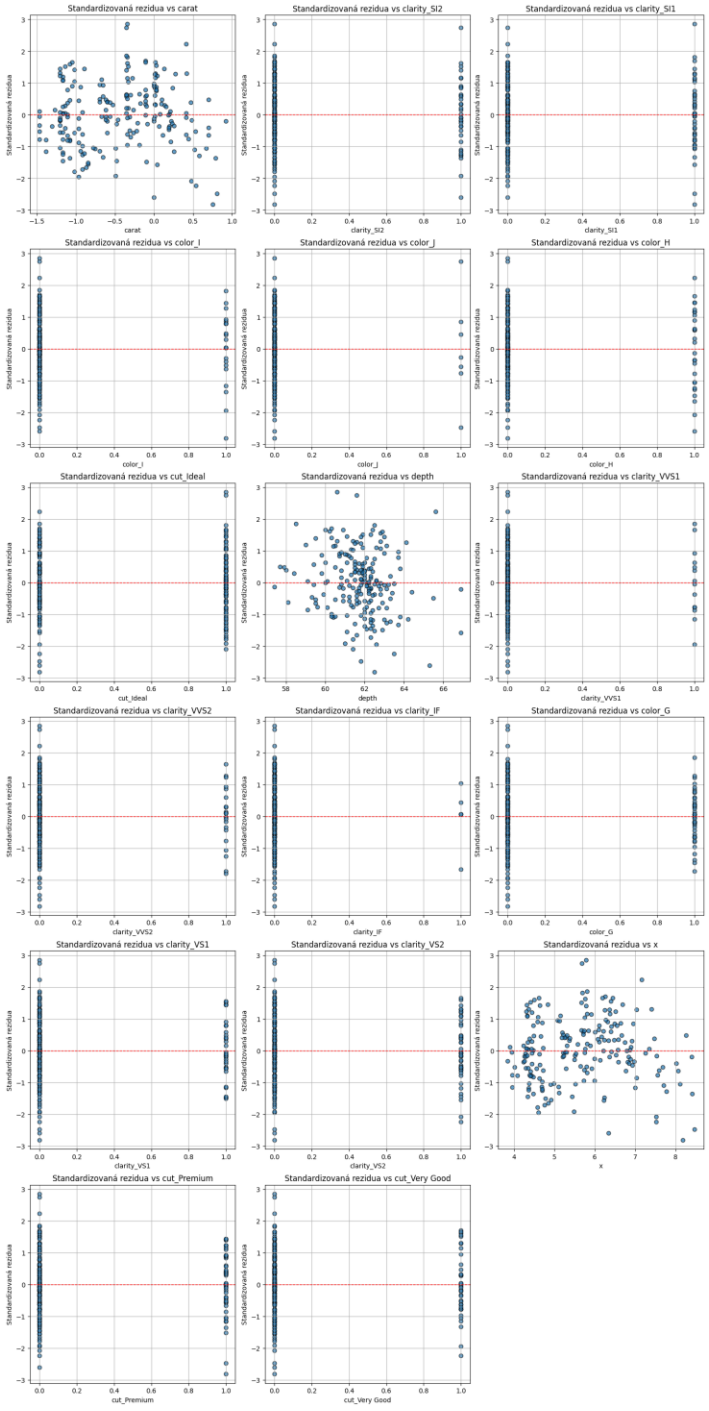
Nestranný odhad rozptylu pro daný submodel: $\sigma^2 = 0.0174556173$

Test hypotézy, že dvě úrovně zvolené kategoriální proměnné mají stejný efekt:

```
clarity_SI2 clarity_SI1| Waldova statistika: 29.6074| p-hodnota: 0.0000| Odmítám. H0: Úrovně mají různé efekty.
clarity_SI2 clarity_VVS1| Waldova statistika: 133.9523| p-hodnota: 0.0000| Odmítám. H0: Úrovně mají různé efekty.
clarity_SI2 clarity_VVS2| Waldova statistika: 148.3252| p-hodnota: 0.0000| Odmítám. H0: Úrovně mají různé efekty.
clarity_SI2 clarity_IF| Waldova statistika: 79.1315| p-hodnota: 0.0000| Odmítám. H0: Úrovně mají různé efekty.
clarity_SI2 clarity_VS1| Waldova statistika: 126.6649| p-hodnota: 0.0000| Odmítám. H0: Úrovně mají různé efekty.
clarity_SI2 clarity_VS2| Waldova statistika: 88.2323| p-hodnota: 0.0000| Odmítám. H0: Úrovně mají různé efekty.
clarity_SI1 clarity_VVS1| Waldova statistika: 71.0077| p-hodnota: 0.0000| Odmítám. H0: Úrovně mají různé efekty.
clarity_SI1 clarity_VVS2| Waldova statistika: 69.1289| p-hodnota: 0.0000| Odmítám. H0: Úrovně mají různé efekty.
clarity_SI1 clarity_IF| Waldova statistika: 42.7045| p-hodnota: 0.0000| Odmítám. H0: Úrovně mají různé efekty.
clarity_SI1 clarity_VS1| Waldova statistika: 47.3938| p-hodnota: 0.0000| Odmítám. H0: Úrovně mají různé efekty.
clarity_SI1 clarity_VS2| Waldova statistika: 18.8233| p-hodnota: 0.0000| Odmítám. H0: Úrovně mají různé efekty.
clarity_VVS1 clarity_VVS2| Waldova statistika: 1.5719| p-hodnota: 0.2115| Neodmítám H0: Úrovně mají stejný efekt.
clarity_VVS1 clarity_IF| Waldova statistika: 1.0045| p-hodnota: 0.3175| Neodmítám H0: Úrovně mají stejný efekt.
clarity_VVS1 clarity_VS1| Waldova statistika: 10.1672| p-hodnota: 0.0017| Odmítám. H0: Úrovně mají různé efekty.
clarity_VVS1 clarity_VS2| Waldova statistika: 31.9017| p-hodnota: 0.0000| Odmítám. H0: Úrovně mají různé efekty.
clarity_VVS2 clarity_IF| Waldova statistika: 3.6967| p-hodnota: 0.0561| Neodmítám H0: Úrovně mají stejný efekt.
clarity_VVS2 clarity_VS1| Waldova statistika: 4.5985| p-hodnota: 0.0333| Odmítám. H0: Úrovně mají různé efekty.
clarity_VVS2 clarity_VS2| Waldova statistika: 25.5497| p-hodnota: 0.0000| Odmítám. H0: Úrovně mají různé efekty.
clarity_IF clarity_VS1| Waldova statistika: 10.3379| p-hodnota: 0.0015| Odmítám. H0: Úrovně mají různé efekty.
clarity_IF clarity_VS2| Waldova statistika: 22.8340| p-hodnota: 0.0000| Odmítám. H0: Úrovně mají různé efekty.
clarity_VS1 clarity_VS2| Waldova statistika: 9.8065| p-hodnota: 0.0020| Odmítám. H0: Úrovně mají různé efekty.
color_I color_J| Waldova statistika: 1.4702| p-hodnota: 0.2269| Neodmítám H0: Úrovně mají stejný efekt.
color_I color_H| Waldova statistika: 42.3433| p-hodnota: 0.0000| Odmítám. H0: Úrovně mají různé efekty.
color_I color_G| Waldova statistika: 58.7314| p-hodnota: 0.0000| Odmítám. H0: Úrovně mají různé efekty.
color_J color_H| Waldova statistika: 34.9097| p-hodnota: 0.0000| Odmítám. H0: Úrovně mají různé efekty.
color_J color_G| Waldova statistika: 42.9501| p-hodnota: 0.0000| Odmítám. H0: Úrovně mají různé efekty.
color_H color_G| Waldova statistika: 0.5240| p-hodnota: 0.4701| Neodmítám H0: Úrovně mají stejný efekt.
cut_Ideal cut_Premium| Waldova statistika: 0.0032| p-hodnota: 0.9548| Neodmítám H0: Úrovně mají stejný efekt.
cut_Ideal cut_Very Good| Waldova statistika: 0.5558| p-hodnota: 0.4569| Neodmítám H0: Úrovně mají stejný efekt.
cut_Premium cut_Very Good| Waldova statistika: 0.3744| p-hodnota: 0.5414| Neodmítám H0: Úrovně mají stejný efekt.
```

4 Regresní diagnostika

Grafy standardizovaných reziduí v závislosti na doprovodných proměnných:



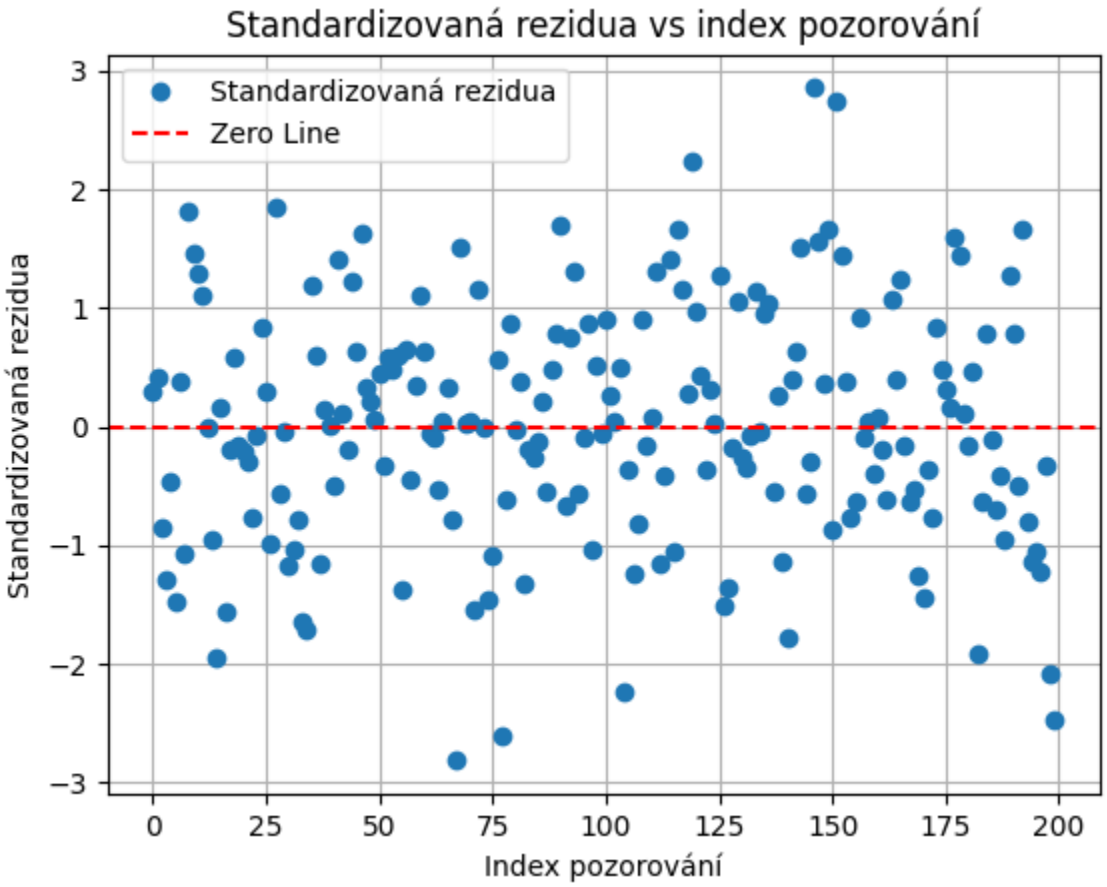
Předpoklady modelu: Rezidua by měla být rozdělena okolo nuly náhodně (tedy neměla by být vidět závislost), což je podle grafů splněno.

Zda se fakticky jedná o homoskedasticitu nebo heteroskedasticitu zjistíme pomocí Breusch-Paganova testu.

$\alpha = 0.05$

```
Breusch-Pagan statistika: 19.1152, p-hodnota: 0.3220, Rezidua jsou homoskedastická.
```

Graf standardizovaných reziduí v závislosti na indexu pozorování:

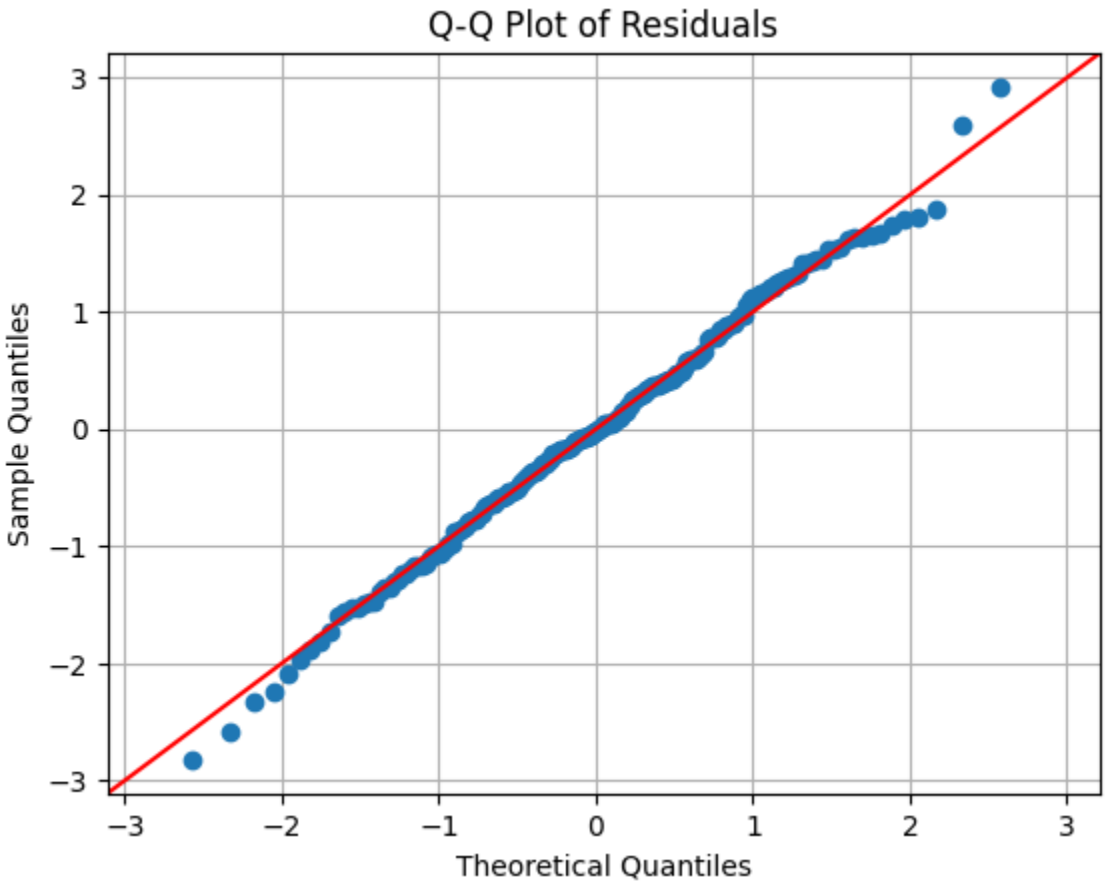


Nezávislost reziduí zjistíme Durbin-Watsonovým testem:

```
Durbin-Watson statistika: 1.7363, Rezidua jsou nezávislá.
```

Nezávislost, protože D.-W. statistika je blízká 2

Normalita reziduí



Normalitu reziduí zjistíme Shapiro-Wilkovým testem:

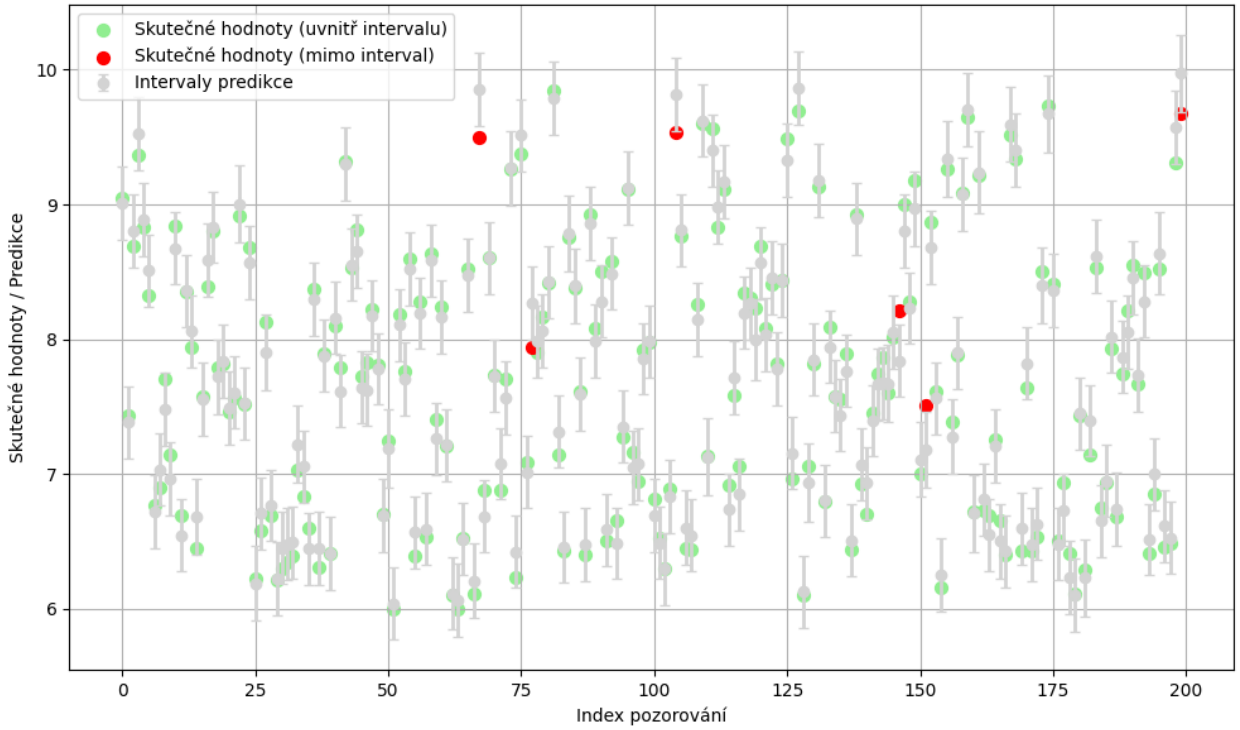
```
Shapiro-Wilk statistika: 0.9963, p-hodnota: 0.9107, Rezidua sledují normální rozdělení.
```

Předpověď a predikční interval pro hodnotu doprovodných proměnných prvního měření:

```
Skutečná hodnota: 9.0479
Predikovaná hodnota: 9.0107
Interval predikce: [8.7361, 9.2853]
```

5 Závěr

Graf predikcí ostatních pozorování:



Tento graf je v práci „navíc“, ale hezky ilustruje fakt, že model byl dobře zvolen, protože většina hodnot se nachází v intervalu predikce. Autor uznává, že se nápadem na graf inspiroval u kolegy, ovšem i přes to se ho zde rozhodl pro zajímavost vykreslit.

Práce byla zpracována v Pythonu.

Nejdříve jsme provedli grafickou analýzu, z které jsme zjistili, že je zapotřebí pracovat s transformovanými daty, protože některé spojité proměnné neměly samy o sobě normální rozdělení. Dále jsme sestavili krabicové diagramy, ze kterých jasně vyplynula existence odlehlých hodnot, které však příliš neovlivnily přesnost modelu.

V korelační analýze jsme si potvrdili, že cena diamantu je přímo úměrná jeho velikosti a karátu. Dále jsme zjistili, že výběrový parciální korelační koeficient price a table při fixních hodnotách carat a depth je nenulový a záporný. Proměnné carat a depth tedy negativně ovlivňují proměnné price a table.

V regresní analýze jsme nalezli vhodný submodel, který pokryl přes 98% původních dat. Pomocí Breusch-Paganova testu jsme ověřili homoskedasticitu reziduí. Nezávislost reziduí jsme ověřili nejdříve graficky a poté Durbin-Watsonovým testem. Zjistili jsme nezávislost, protože D.-W. statistika je blízká 2. Dále jsme ověřili normalitu reziduí pomocí Q-Q plotu a Shapiro-Wilkovým testem. Nakonec jsme stanovili předpověď a predikční interval pro hodnotu doprovodných proměnných prvního měření.