

Co je statistika?

ÚVOD DO STATISTIKY



George Boorman

Curriculum Manager, DataCamp

Co je statistika?

- **Statistika** – praxe a studium sběru a analýzy dat
- Dvě hlavní odvětví statistiky:
 - **Popisná statistika** – popis nebo shrnutí dat
 - **Inferenční statistika** – sběr vzorku dat a aplikace výsledků na populaci, kterou vzorek reprezentuje

Statistika je všude!

- Sportovní statistiky



- Osobní finance



¹ Image credits: <https://unsplash.com/@jesusance>; <https://unsplash.com/@andretaissin>; <https://unsplash.com/@unarchive>

Co statistika dokáže?

- Umožňuje odpovídat na praktické otázky:
 - Jaký je průměrný plat v USA?
 - Kolik dotazů zákazníků firma pravděpodobně obdrží za týden?
- Má uplatnění napříč společnostmi:
 - Vývoj bezpečnějších výrobků, jako jsou auta nebo letadla
 - Pomáhá vládám pochopit potřeby populace
- Potvrzuje vědecké průlomy, například vakcíny proti Covid-19

¹ source: <https://www.bmj.com/content/373/bmj.n1088>

Omezení statistiky

- Statistika vyžaduje konkrétní, měřitelné otázky:
 - Je rocková hudba populárnější než jazz?
 - Žijí ženy v průměru déle než muži?
- Statistika nám nevysvětlí *proč* vztahy existují



¹ Image credit: <https://unsplash.com/@mohammadmetri>

Typy dat: numerická

- Spojitá data:
 - Ceny akcií

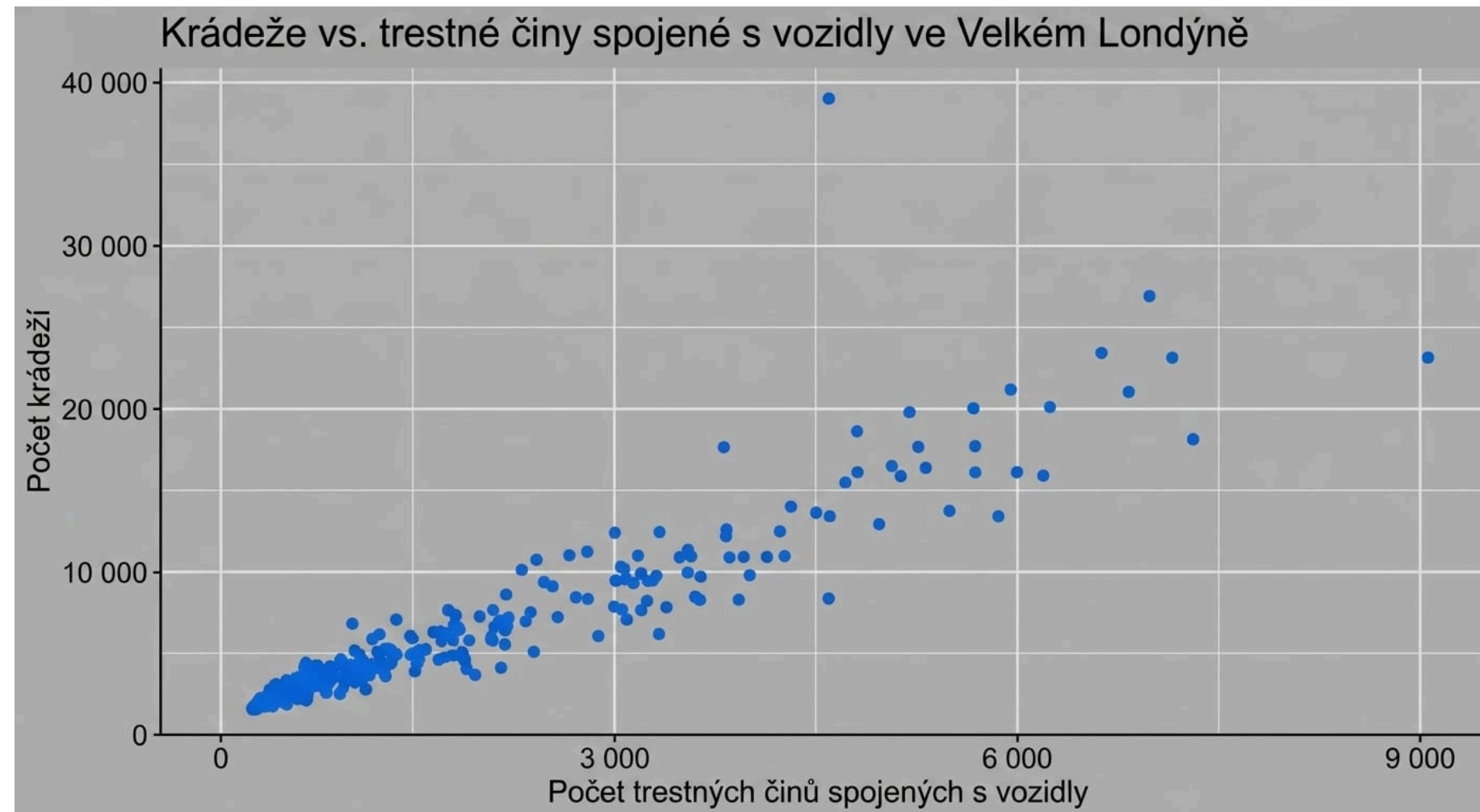
Akcie	Otevírací cena (\$)	Zavírací cena {\$}
Amazon.com, Inc	2328.14	2329.00
Apple Inc	156.77	157.04
Netflix Inc	188.32	188.75

- Intervalová/početní data:
 - Kolik šálků kávy lidé vypijí za den?

Jméno	Šálků kávy za den
Jessica	4
Andrew	2
Penny	3

¹ Image credits: Stocks https://unsplash.com/@behy_studio

Vizualizace numerických dat



¹ Data source: Metropolitan Police Service, United Kingdom

Typy dat: kategorická

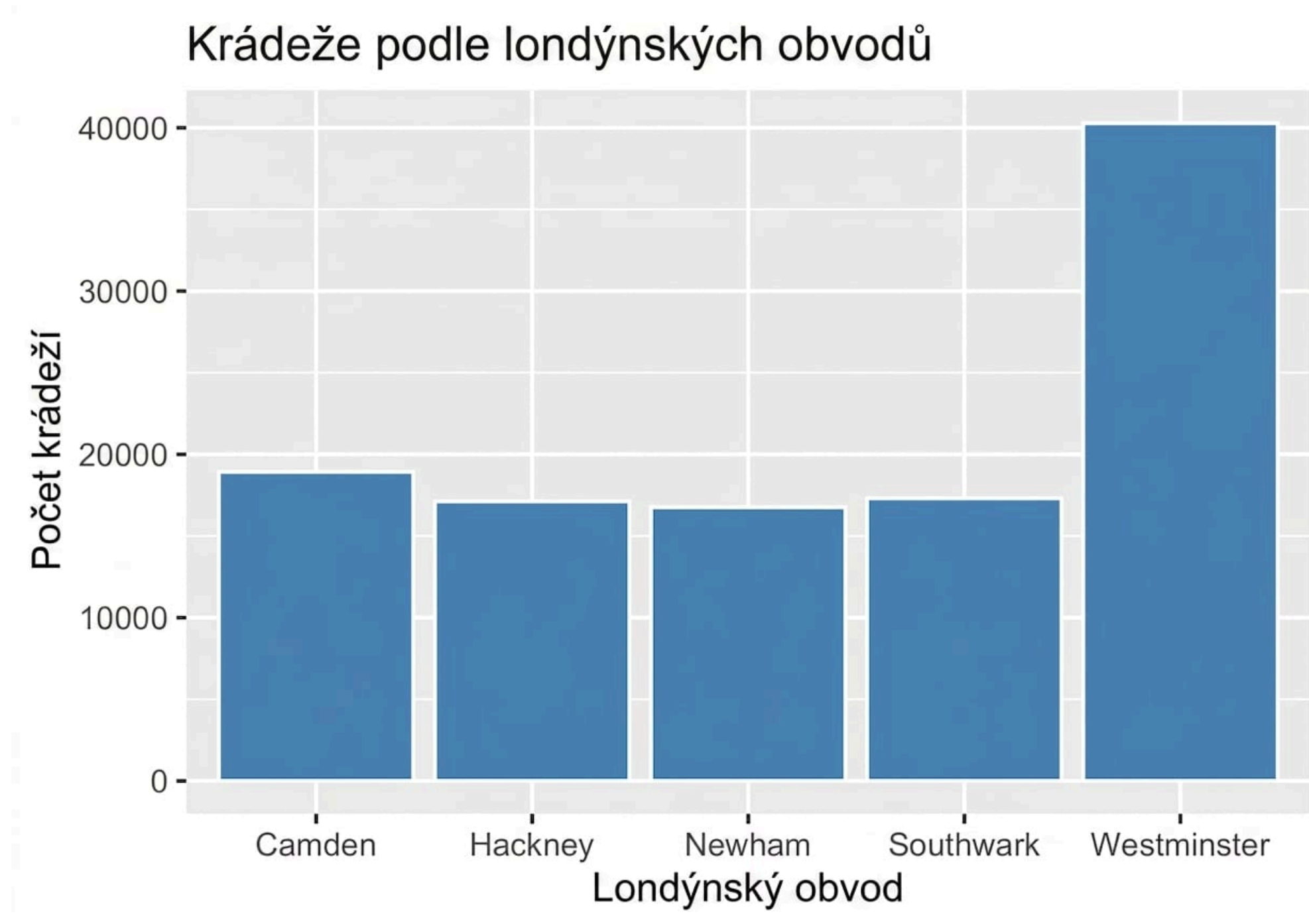
- Nominální data:
 - Barva očí

Jméno	Barva očí
Jessica	Hnědá
Adam	Zelená
Sarah	Modrá

- Ordinální data:
 - Jak silně souhlasíte s tím, že basketbal je nejlepší sport?
 - ☐ Rozhodně nesouhlasím
 - ☐ Spíše nesouhlasím
 - ☐ Ani souhlasím, ani nesouhlasím
 - ☒ Spíše souhlasím
 - ☐ Rozhodně souhlasím

¹ Image credit: https://unsplash.com/@mango_quan

Vizualizace kategorických dat



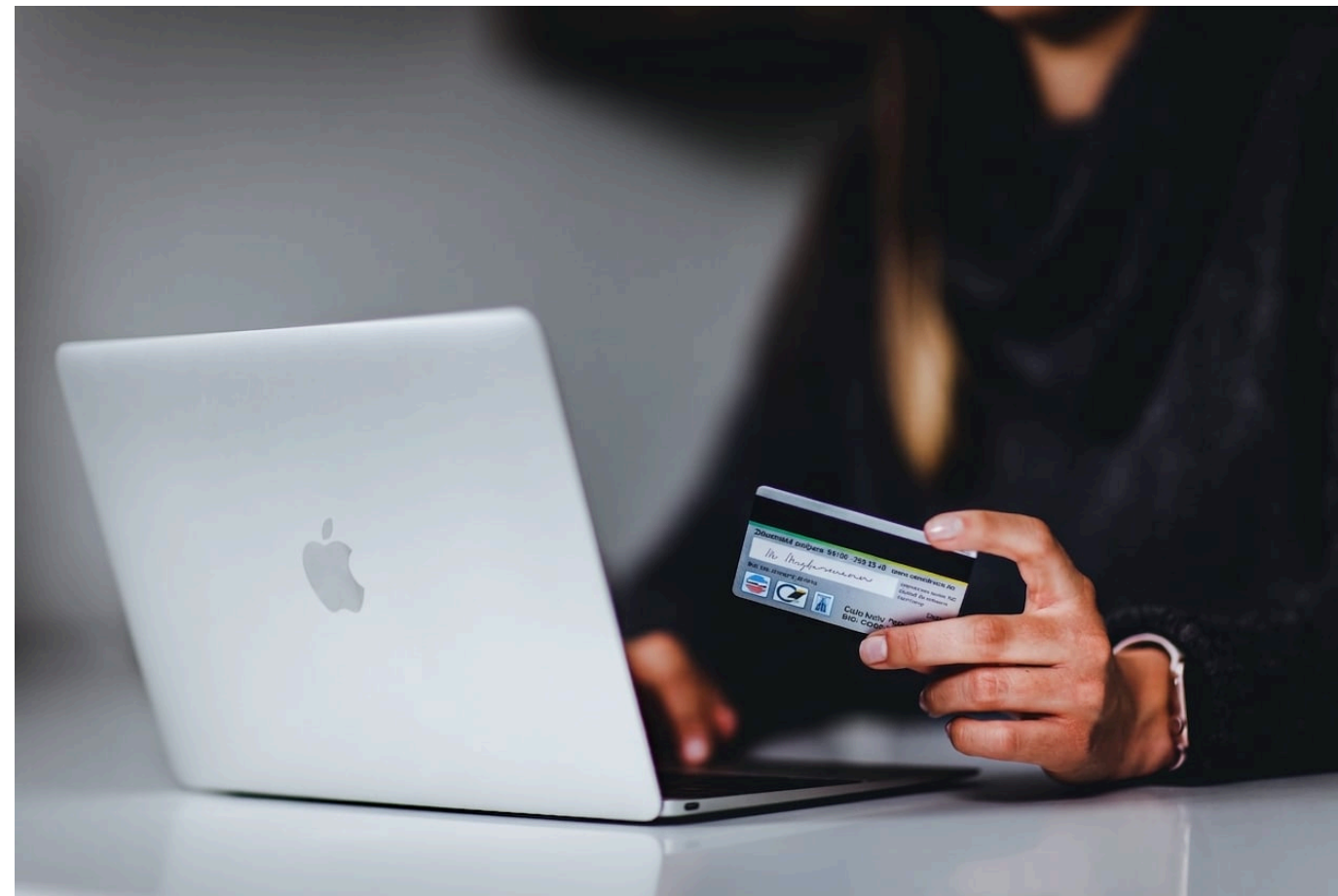
Popisná statistika

- *Popis* nebo *shrnutí* dat

Obvod	Počet krádeží	Podíl z celku
Westminster	40 278	36,48 %
Camden	18 928	17,15 %
Southwark	17 309	15,68 %
Hackney	17 121	15,51 %
Newham	16 762	15,18 %

Inferenční statistika

- Pomocí vzorku *vyvozujeme závěry* o populaci
- Kolik lidí nakoupí oblečení po reklamě na sociálních sítích?



Lass uns üben!

ÚVOD DO STATISTIKY

Míry středu

ÚVOD DO STATISTIKY



George Boorman

Curriculum Manager, DataCamp

Proč jsou míry středu užitečné?

- Průměrný počet prodejních objednávek za měsíc
- Typická cena domu
- Nejčastější barva vlasů



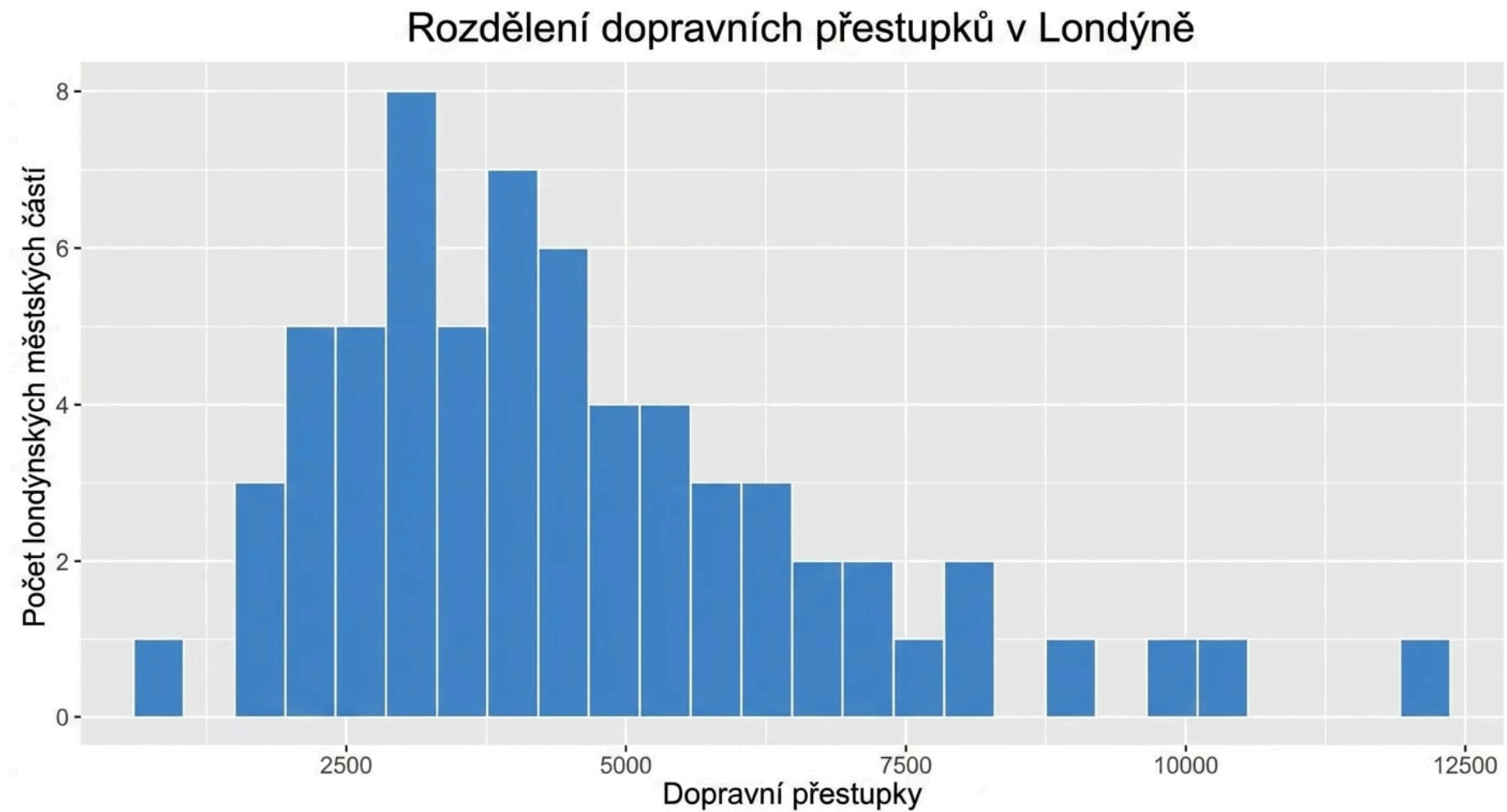
Kriminální data

Obvod	Vloupání	Loupež	Krádež	Trestné činy s vozidly	Celkem
Barking and Dagenham	2318	1265	6300	4668	14551
Barnet	5067	1369	9875	9841	26152
Bexley	1583	444	4500	4216	10743
Brent	3893	1650	10026	7739	23308
Bromley	3053	844	8635	6966	19498
...	...	...	...	...	...

Kriminální data

Obvod	Vloupání	Loupež	Krádež	Trestné činy s vozidly	Celkem
Barking and Dagenham	2318	1265	6300	4668	14551
Barnet	5067	1369	9875	9841	26152
Bexley	1583	444	4500	4216	10743
Brent	3893	1650	10026	7739	23308
Bromley	3053	844	8635	6966	19498
...	...	...	...	...	...

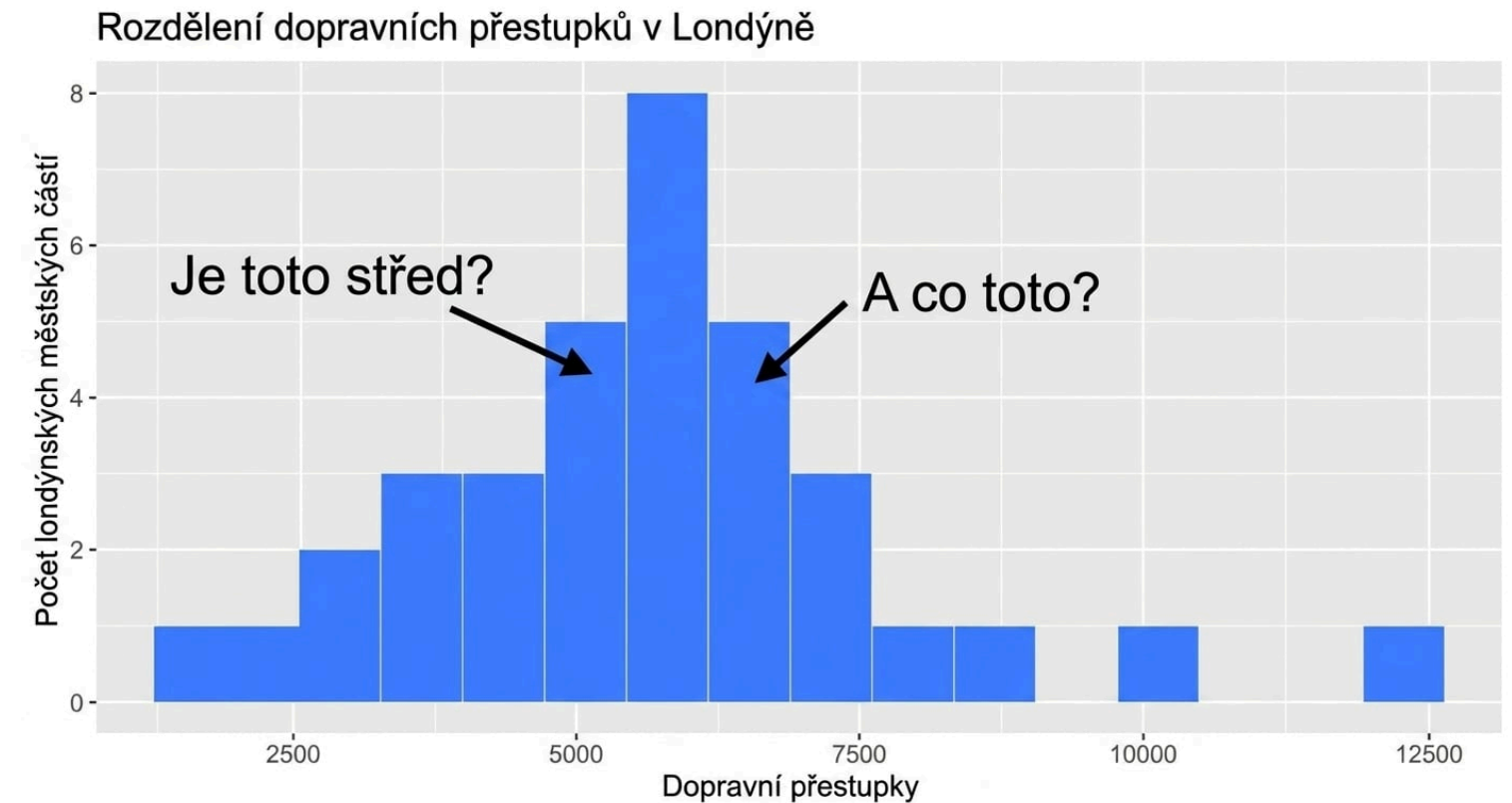
Histogramy



Jaký je typický počet trestných činů s vozidly v Londýně?

Kde leží střed dat?

- Průměr
- Medián
- Modus



Míry středu: průměr

$$\text{mean} = \frac{\text{Sum of all values}}{\text{Count of all values}}$$

- Průměrný počet vloupání na londýnský obvod:

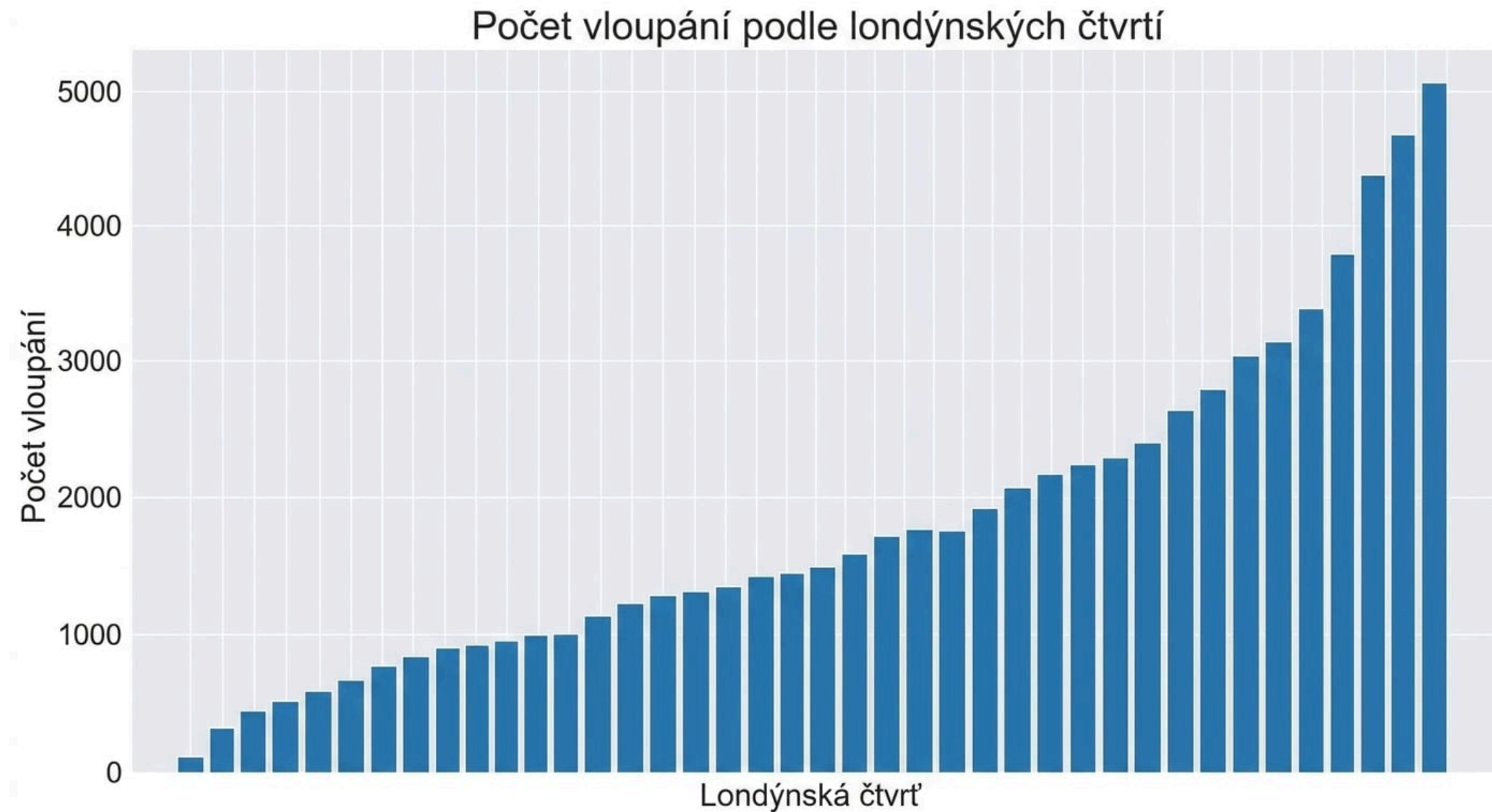
$$\text{Mean}(\text{Burglary}) = \frac{(2,318 + 5,067 + 1,583 + 3,893 + \dots)}{32} = 3,462.8125$$

Míry středu: průměr

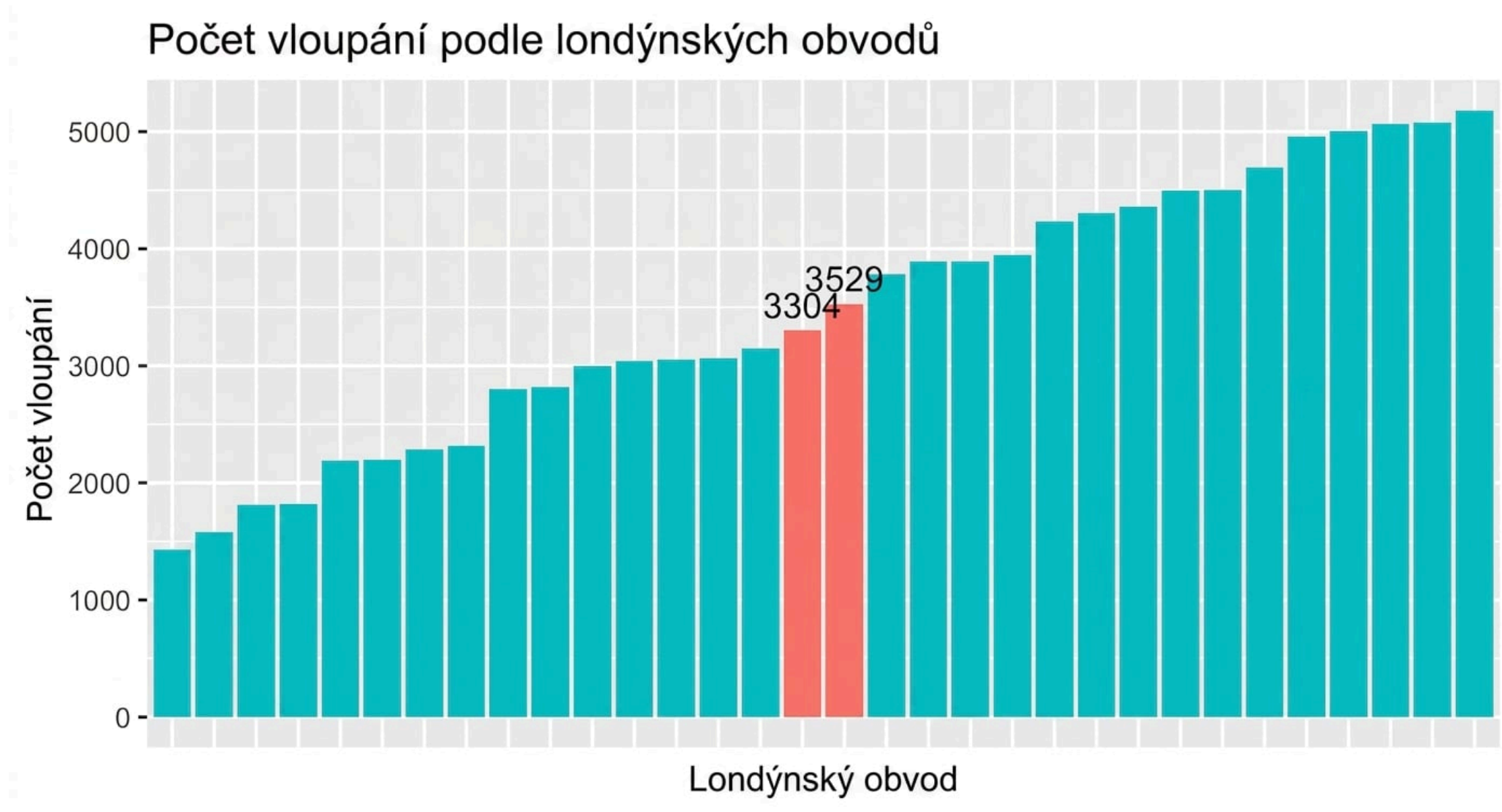
Typ trestného činu	Průměr
Vloupání	3358.06
Loupež	1450.85
Krádež	10642.21
Trestné činy s vozidly	6227.27
Celkem	47671.62

Míry středu: medián

- Medián – prostřední hodnota dané proměnné



Míry středu: medián



Míry středu: medián

$$\text{Median}(\text{Burglary}) = \frac{(3,304 + 3,529)}{2}$$

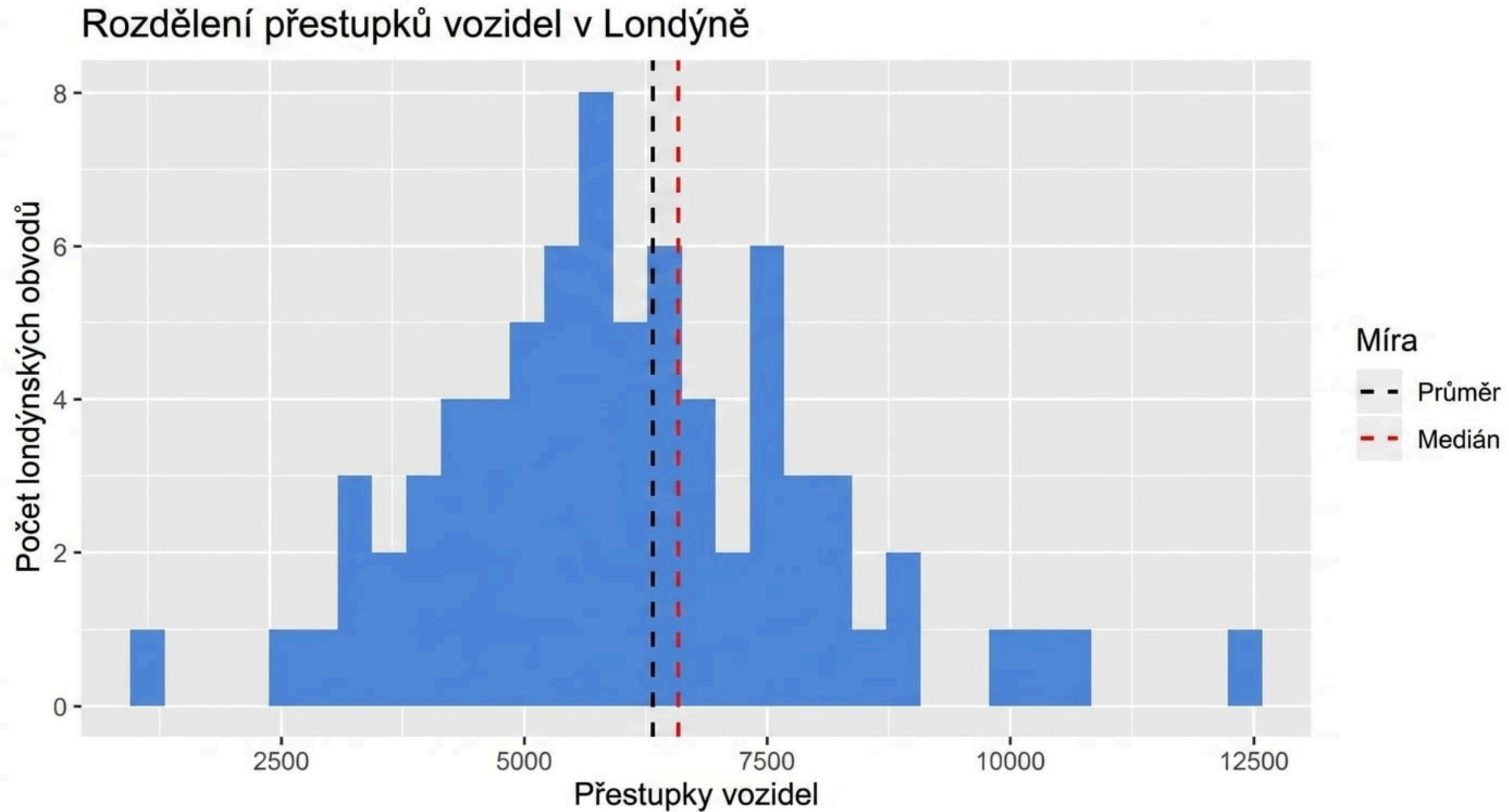
$$\text{Median}(\text{Burglary}) = 3,416.5$$

Míry středu: modus

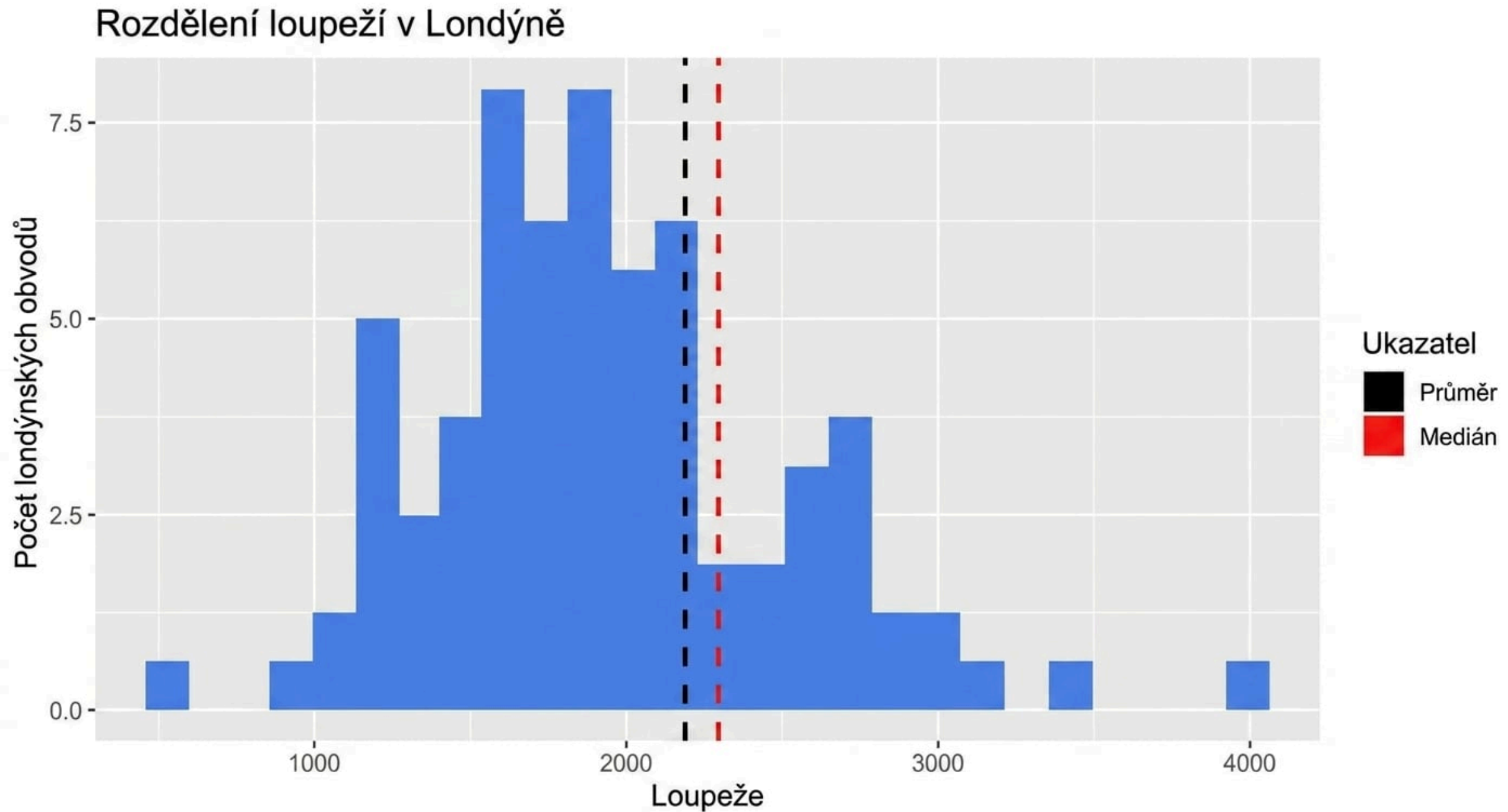
- Modus – *nejčastější hodnota*

Typ trestného činu	Počet
Vloupání	110 810
Loupež	47 877
Krádež	350 025
Trestné činy s vozidly	205 337

Která míra je vhodná?



Která míra je vhodná?



Pojďme si procvičit!

ÚVOD DO STATISTIKY

Míry rozptylu

ÚVOD DO STATISTIKY

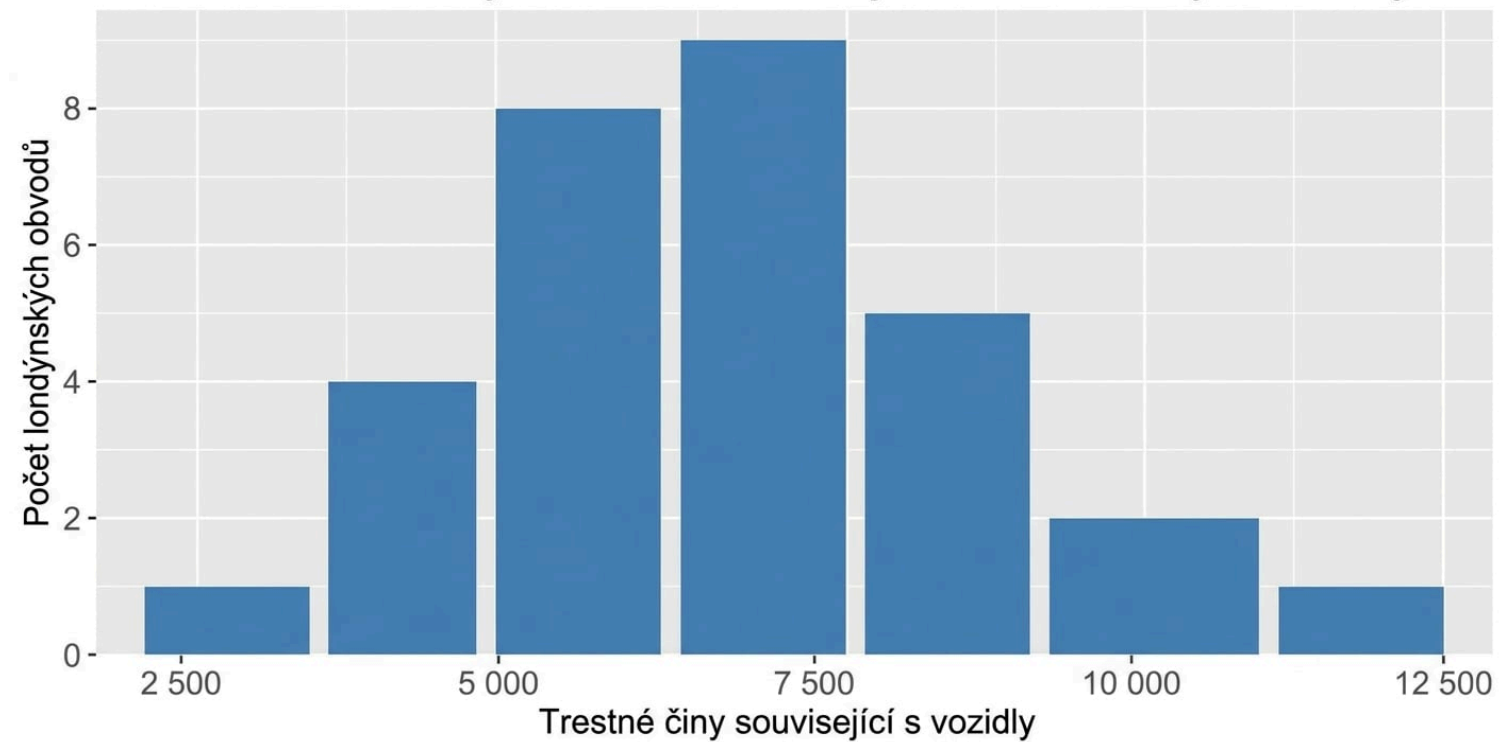


George Boorman

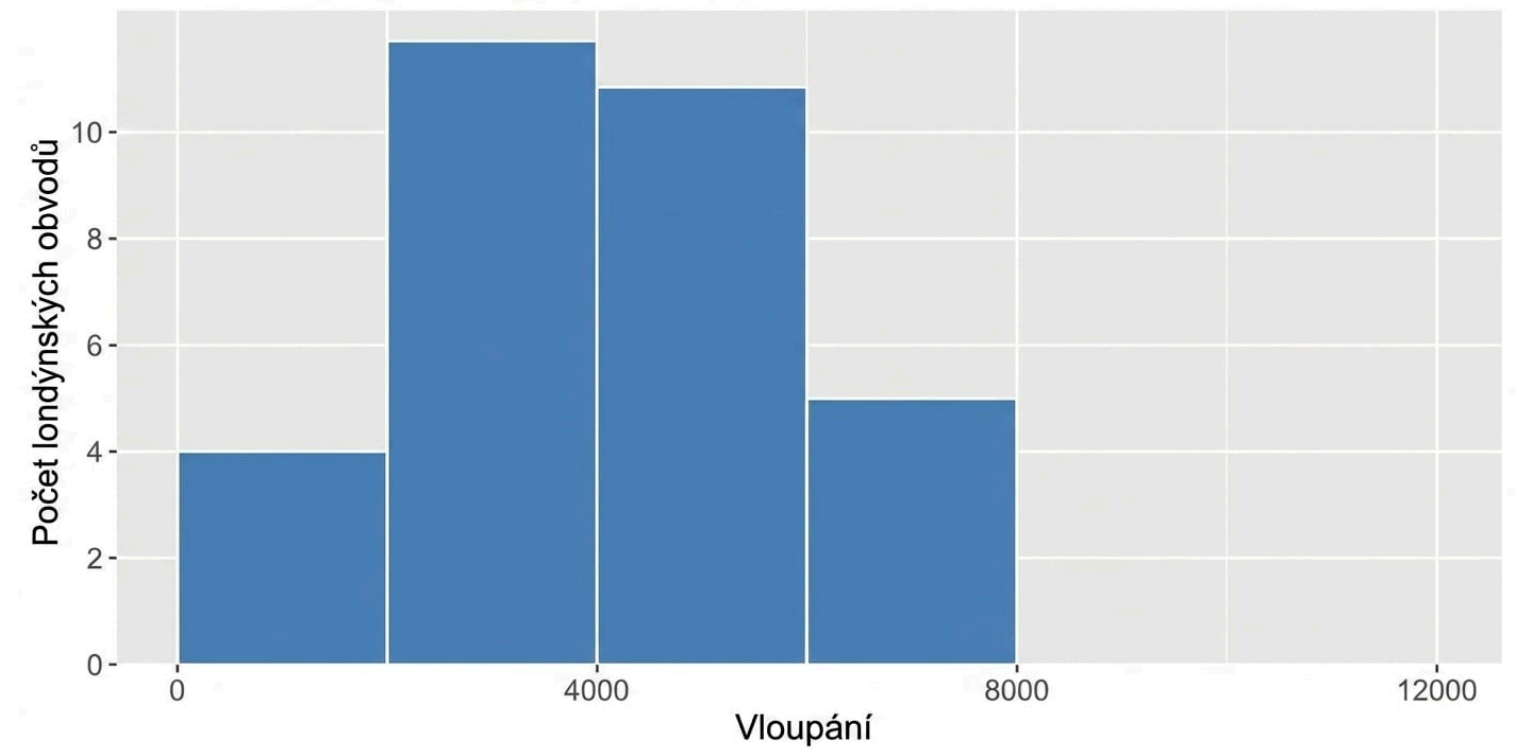
Curriculum Manager, DataCamp

Co je rozptyl?

Rozdělení trestných činů souvisejících s vozidly v Londýně



Rozdělení počtu vloupání v Londýně



Proč je rozptyl důležitý?

- Rozptyl měří variabilitu dat
- Trička obvykle stojí 30 \$
 - Mohou stát 10–200 \$
 - Jaká je pravděpodobnost, že tričko bude stát 30 \$?
- Pokud by trička stála 20–50 \$
 - Změnila by se tím pravděpodobnost, že jedno bude stát 30 \$?



¹ Image credit: <https://unsplash.com/@uyk>

Rozsah

$range = maximum - minimum$

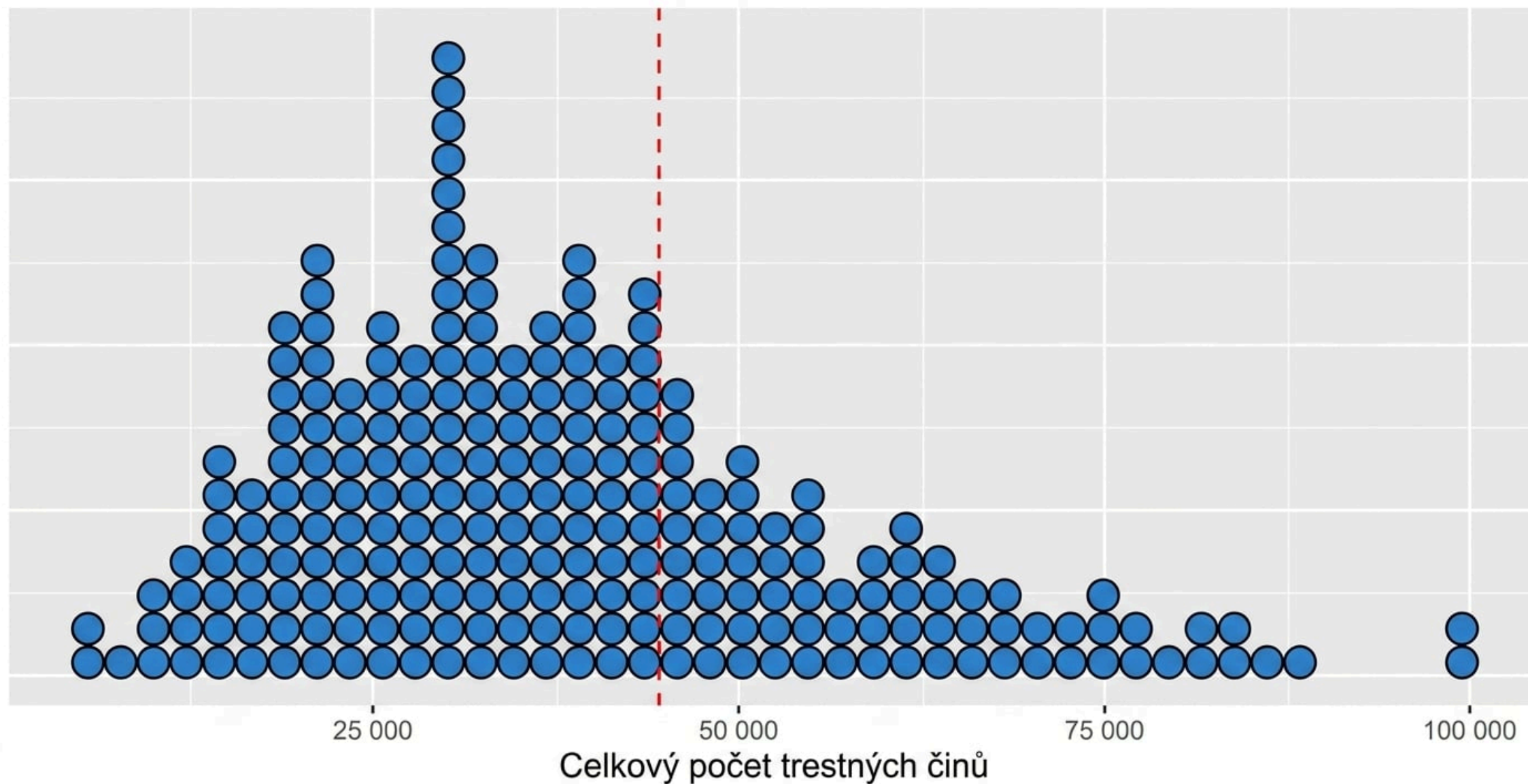
$range(Burglaries) = 5183 - 1432$

$range(Burglaries) = 3751$

Borough	Vloupání
Tower Hamlets	5 183
Hackney	5 079
Barnet	5 067
...	...
Sutton	1 815
Bexley	1 583
Kingston upon Thames	1 432

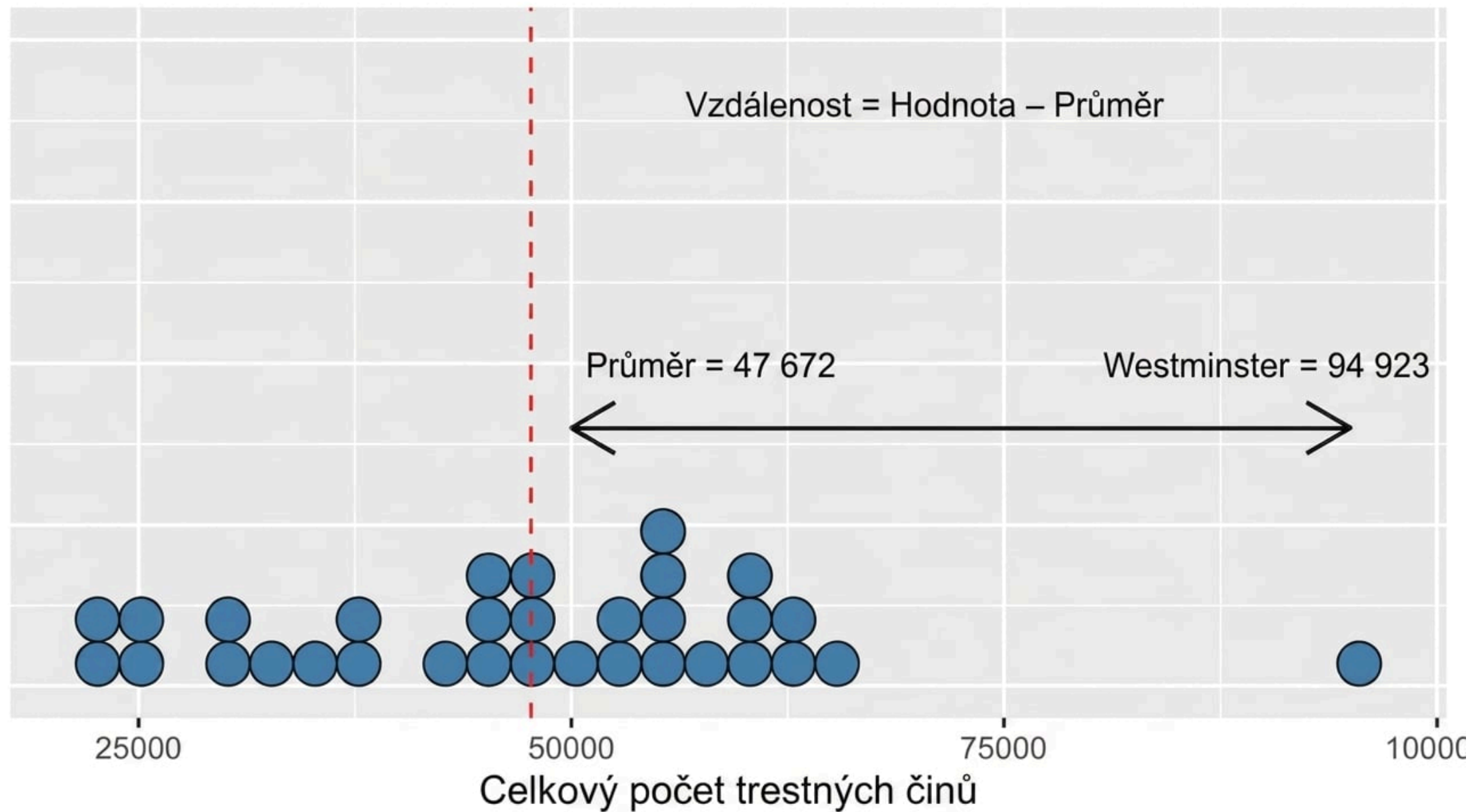
Rozptyl

Celkový počet trestných činů podle londýnských čtvrtí



Rozptyl

Celkový počet trestných činů podle londýnských obvodů



Rozptyl

Borough	Celková kriminalita	Průměr	Vzdálenost
Barking and Dagenham	37 939	47 672	-9 733
Barnet	52 421	47 672	4 749
Bexley	29 285	47 672	-18 387
Brent	55 465	47 672	7 793
Bromley	42 982	47 672	-4 690
Camden	54 806	47 672	7 134
...	...	...	...
Celkem	1 525 492	1 525 492	0

Rozptyl

Borough	Celková kriminalita	Průměr	Vzdálenost	Vzdálenost2
Barking and Dagenham	37 939	47 672	-9 733	94 731 289
Barnet	52 421	47 672	4 749	22 553 001
Bexley	29 285	47 672	-18 387	338 081 769
Brent	55 465	47 672	7 793	60 730 849
Bromley	42 982	47 672	-4 690	21 996 100
Camden	54 806	47 672	7 134	50 893 956
...	...	...	...	...
Celkem	1 525 492	1 525 492	0	7 509 750 824

Rozptyl

$$\text{variance}(\text{total crime}) = \frac{7,509,750,824}{32}$$

$$\text{variance}(\text{total crime}) = 234,679,713$$

Směrodatná odchylka

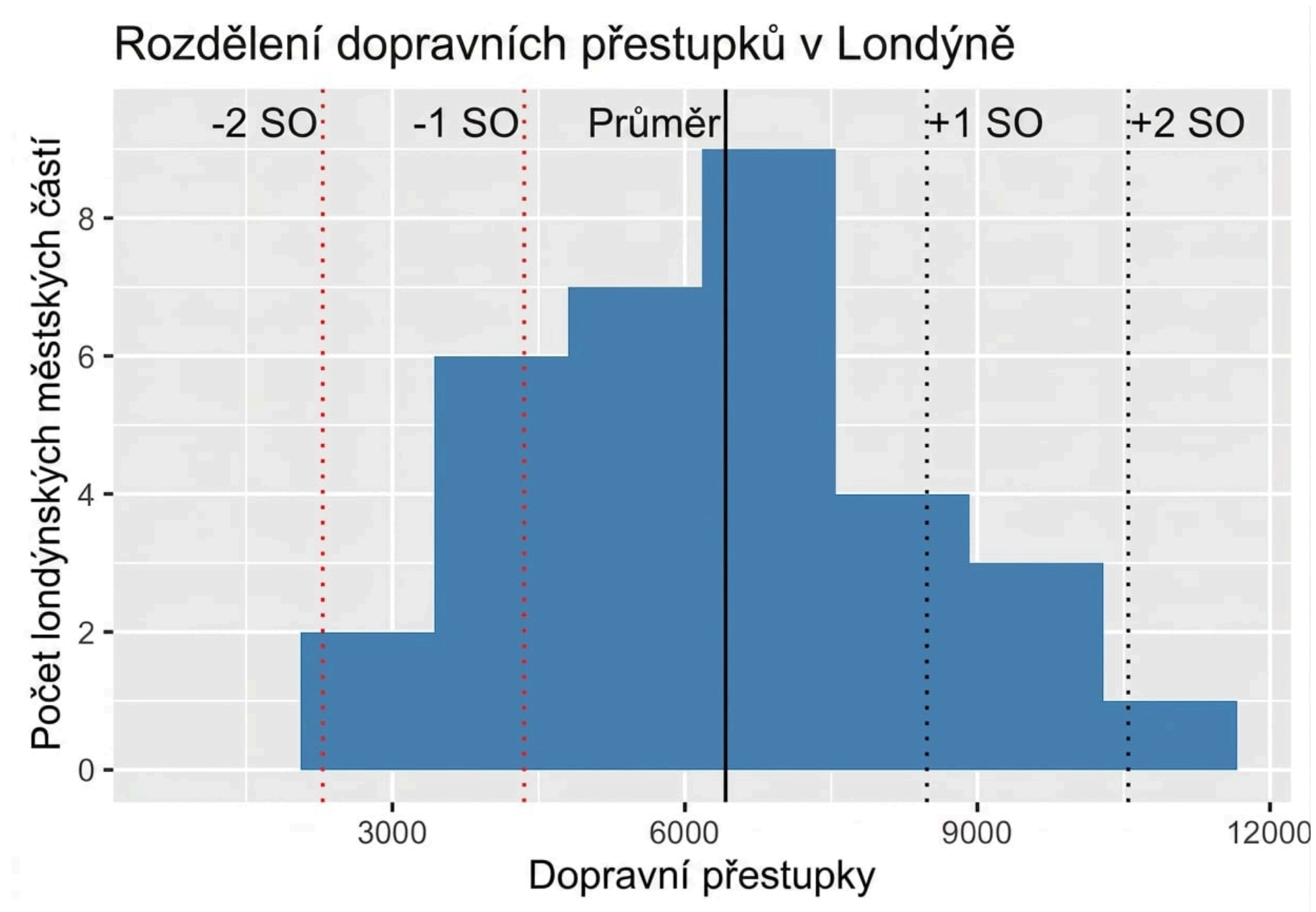
$$\text{standard deviation}(\text{total crime}) = \sqrt{\text{variance}(\text{total crime})}$$

$$\text{standard deviation}(\text{total crime}) = \sqrt{(234,679,713)}$$

$$\text{standard deviation}(\text{total crime}) = 15,319.26$$

- Směrodatná odchylka blízko nuly = data seskupená kolem průměru

Směrodatná odchylka v histogramu



Kvartily

- Kvartily:
 - rozdělení dat do čtyř stejných částí

Kriminalita	0 %	25 %	50 %	75 %	100 %
Vloupání	1 432,00	2 681,75	3 416,50	4 392,00	5 183,00
Loupeže	363,00	895,75	1 354,50	1 976,50	4 156,00
Krádeže	4 090,00	7 739,75	9 624,00	12 059,00	40 278,00
Krádeže vozidel	2 143,00	4 838,25	6 424,50	7 520,75	11 292,00

Kvartily

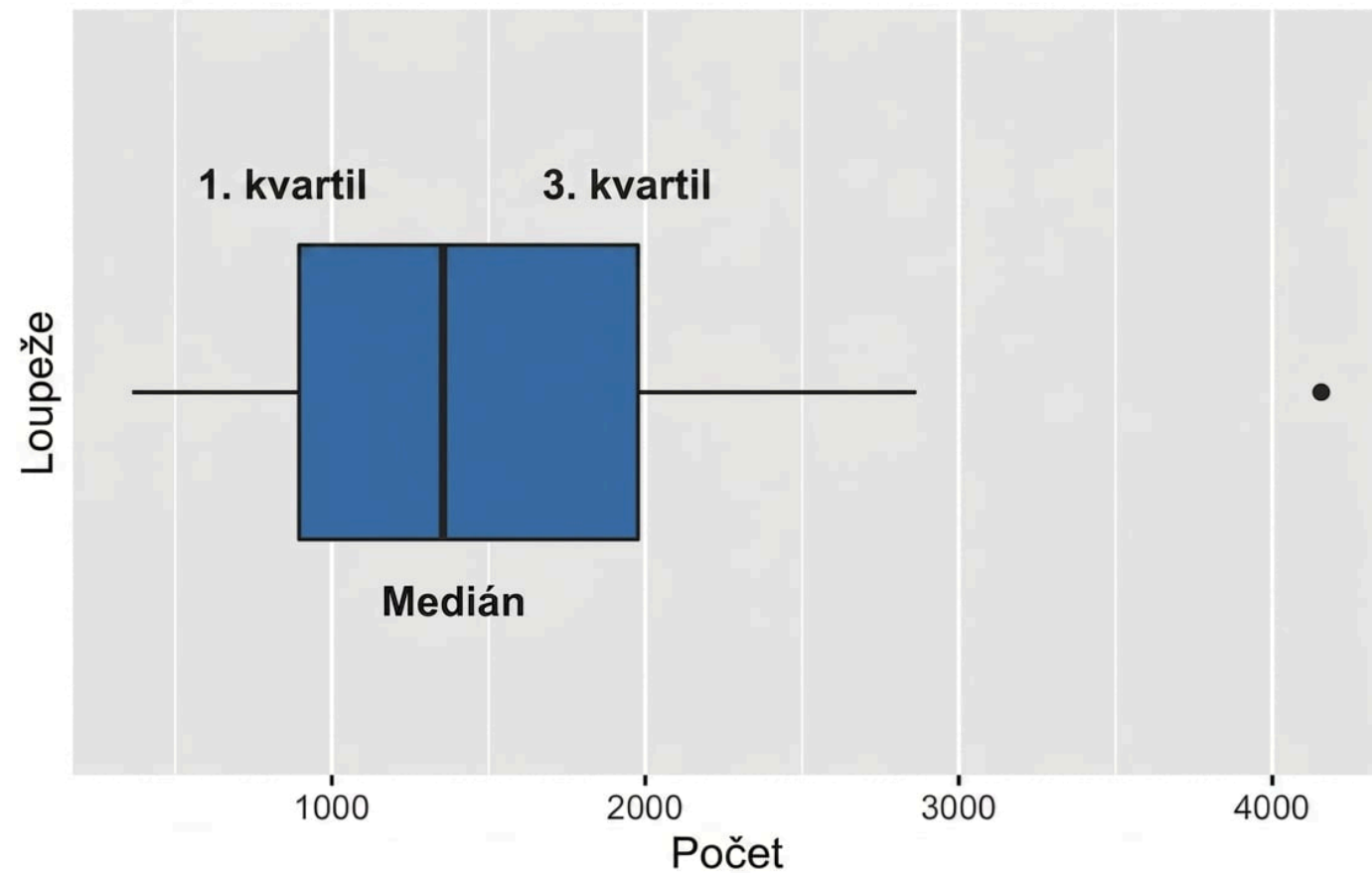
- Kvartily:
 - rozdělení dat do čtyř stejných částí

Kriminalita	0 %	25 %	50 %	75 %	100 %
Vloupání	1 432,00	2 681,75	3 416,50	4 392,00	5 183,00
Loupeže	363,00	895,75	1 354,50	1 976,50	4 156,00
Krádeže	4 090,00	7 739,75	9 624,00	12 059,00	40 278,00
Krádeže vozidel	2 143,00	4 838,25	6 424,50	7 520,75	11 292,00

- Druhý kvartil (50 %) = medián

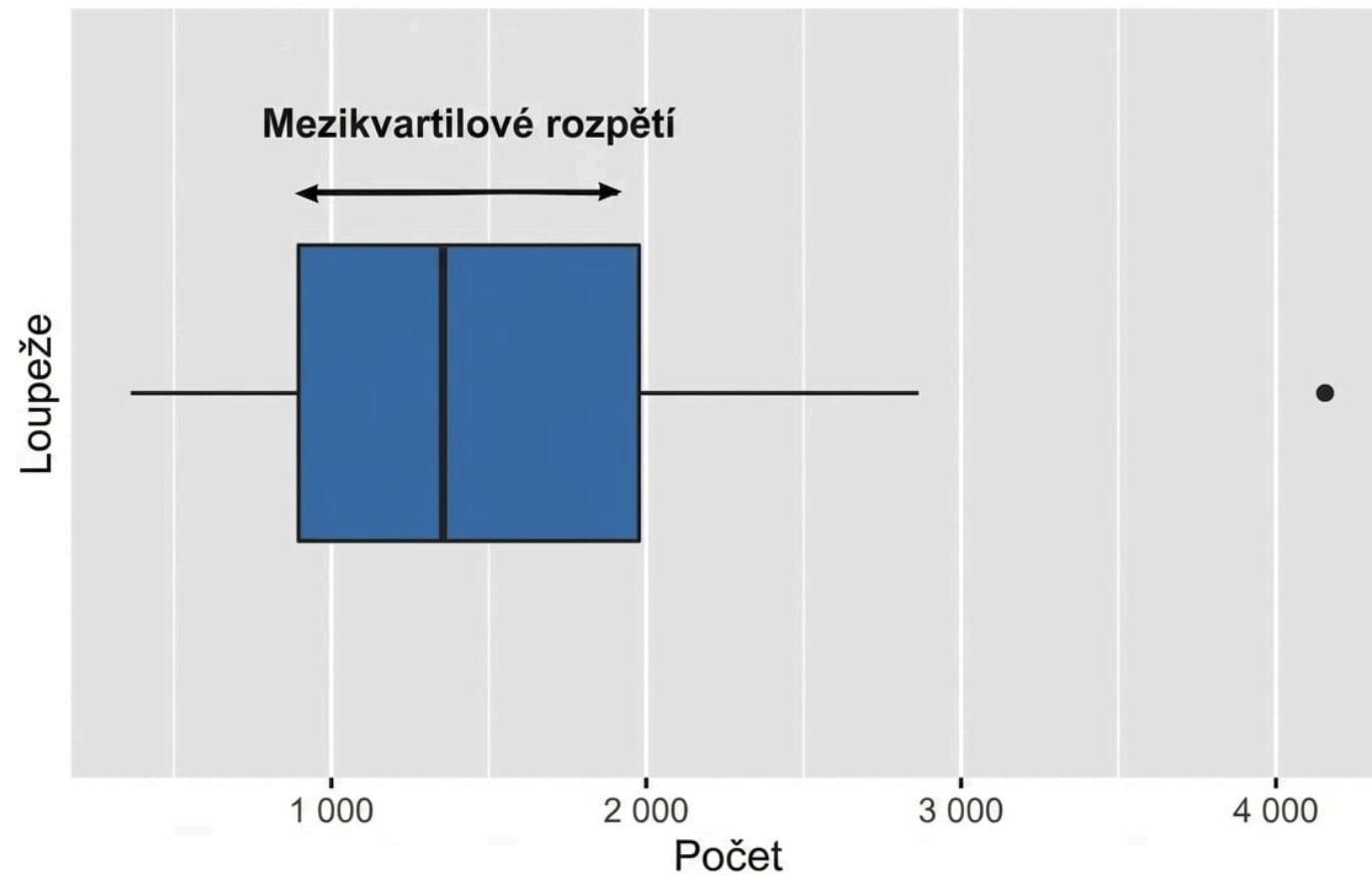
Krabicové grafy

Krabicový graf loupeží v londýnských obvodech



Mezikvartilní rozpětí (IQR)

Krabicový graf loupeží v londýnských obvodech



$$\text{IQR} = 3. \text{ kvartil} - 1. \text{ kvartil}$$

$$\text{IQR} = 1976,50 - 895,75$$

$$\text{IQR} = 1080,75$$

- IQR je méně ovlivněno krajními hodnotami

Pojďme procvičovat!

ÚVOD DO STATISTIKY

