Deskriptivní statistika

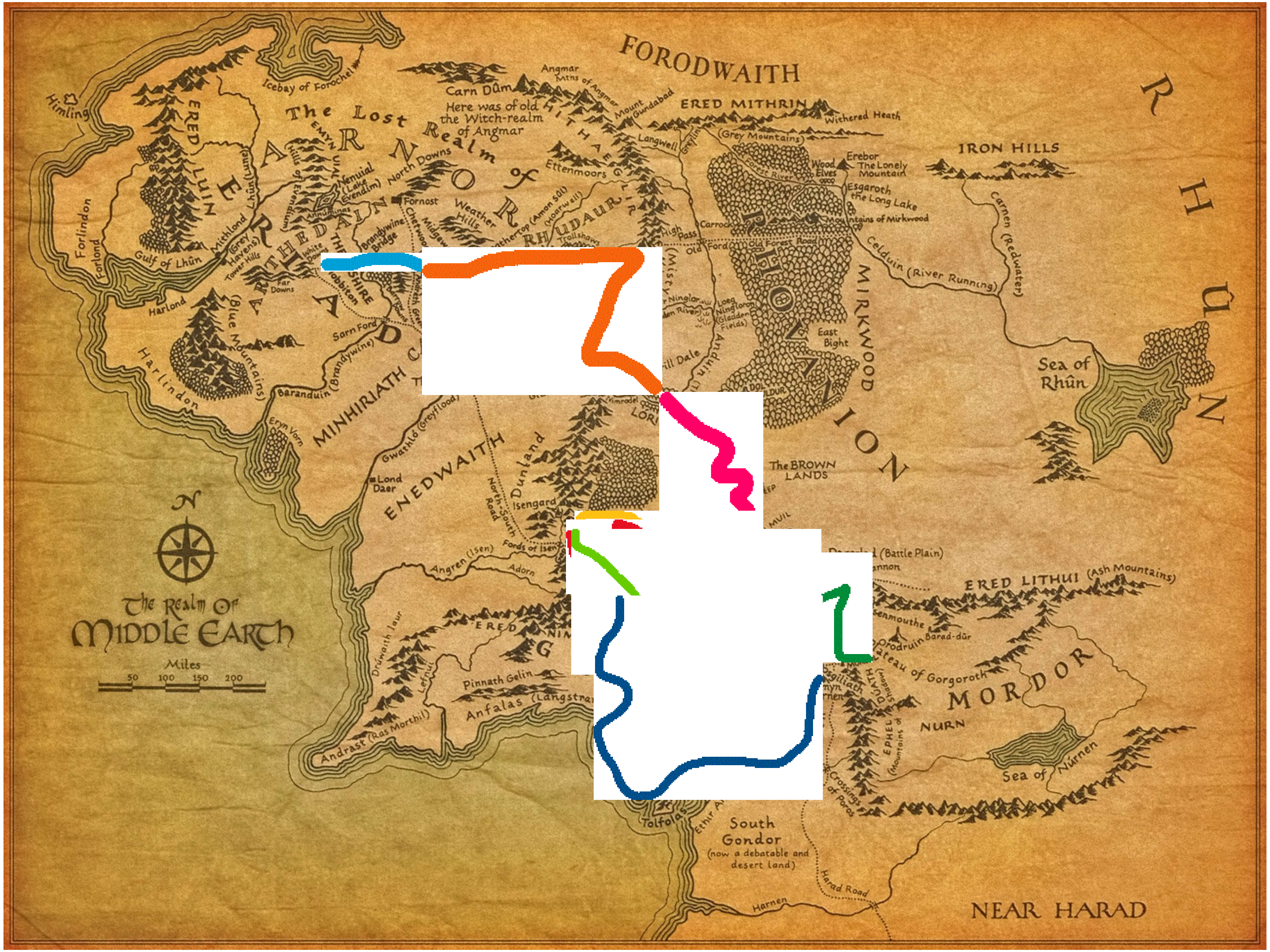
POLb1139 Statistické myšlení v sociálních vědách

Dnes se posouváme o krok dále

- Známe typy proměnných
- Máme data (vlastní sběr / jiným způsobem)

- Jak začít analýzu?

- Ideální první kroky:
  - Poznejte svá data – struktura, distribuce
  - Vizualizace dat


The Realm of Middle Earth
Miles
50 100 150 200
N
FORODWAITH
RHÛN
ERED MITHRIN
IRON HILLS
Withered Heath
Carn Dûm
Here was of old the Witch-realm of Angmar
The Lost Realm of ARNOR
ERIADOR
ERED LUIN
Forlindon
Harlindon
Gulf of Lhûn
Mithlond (Grey Havens)
Harlond
Forlond
Himling
Icebay of Forochel
Evendim
Fornost
North Downs
Weather Hills
Ettenmoors
RHUDAUR
Trollshaws
Carrock
Old Ford
Old Forest Road
RHOVANION
MIRKWOOD
Erebor The Lonely Mountain
Esgaroth the Long Lake
Celduin (River Running)
Carnen (Redwater)
Sea of Rhûn
East Bight
Anduin
The BROWN LANDS
MINHIRIATH
ENEDWAITH
Dunland
Isengard
Lond Daer
Gwathló (Greyflood)
Baranduin (Brandywine)
Sarn Ford
Eryn Vorn
Fords of Isen
Angren (Isen)
Adorn
Pinnath Gelin
Anfalas
Andrast
Druwaith Iaur
Lefnui
ERED LITHUI (Ash Mountains)
Dagorlad (Battle Plain)
Barad-dûr
Plateau of Gorgoroth
MORDOR
NURN
Sea of Nûrnen
EPHEL DUATH
Crossings of Poros
South Gondor (now a debatable and desert land)
Harad Road
Harnen
Tolfalas
NEAR HARAD

Deskriptivní analýza

- Explorace dat v rámci jedné proměnné
- Jednorozměrná analýza
- Cílem je popsat a porozumět datům
- Nehledáme rozdíly ani souvislosti mezi proměnnými
- Má smysl před vícerozměrnou analýzou (nebo i samostatně)
- Vizualizace

Deskriptivní analýza

- Záleží na úrovni proměnných podle měření
  - Různé typy proměnných poskytují různé možnosti
  - Kardinální > ordinální > nominální
  - SPSS vás zpravidla nezachrání (a neupozorní na očividný nesmysl)

- Prostor pro odhalování chyb (měření)
  - Identifikace odlehlých případů (outliers)
  - Identifikace chyb při vkládaní dat (pokud se dají jednoduše rozpoznat)

Kategorická data

- Nominální a ordinální proměnné
- Společné znaky a rozdíly
- Co je (a není) s nimi možné dělat?

- Primárně můžeme sledovat, kolik případů spadá do jednotlivých kategorií
- Jaká je distribuce hodnot napříč kategoriemi
- Numerické kódy pro jejich hodnoty mají pouze symbolický význam → důsledky?

Najděte alespoň 5 nom. / ord. proměnných

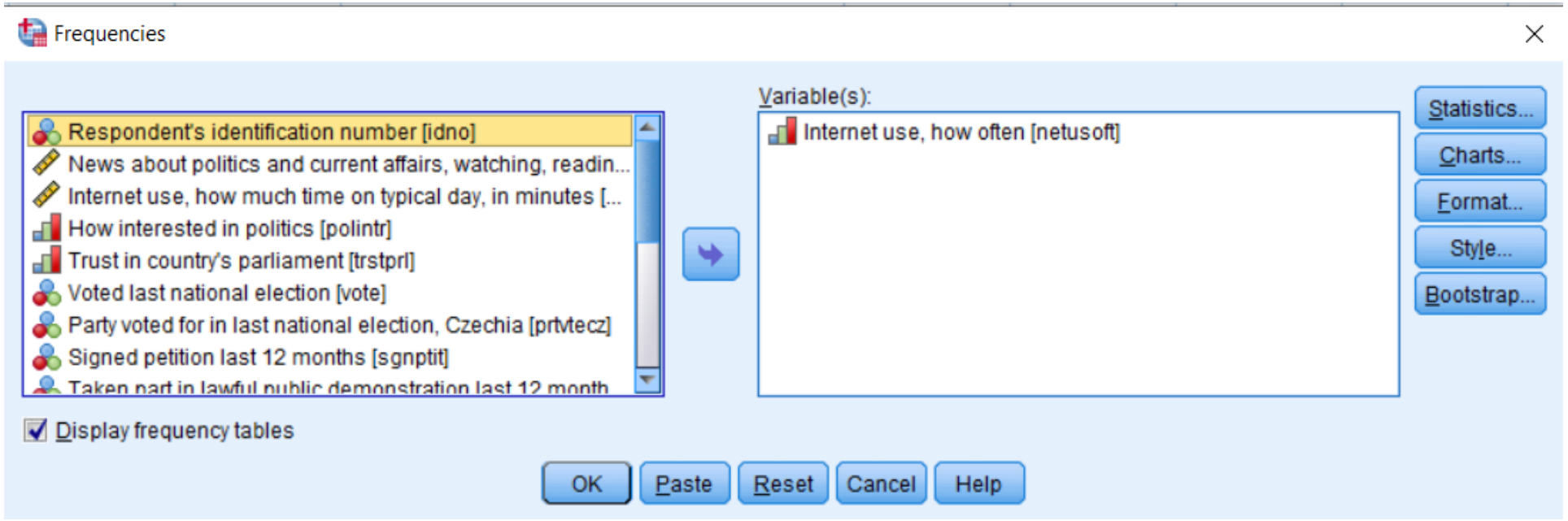
ESS9CZ.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align
1	[redacted]	Numeric	9	0	Respondent's identification number	None	None	11	Right
2	nwspol	Numeric	4	0	News about politics and current affairs, watchi...	{7777, Refus...	7777, 8888, ...	8	Right
3	[redacted]	Numeric	1	0	Internet use, how often	{1, Never}...	7, 8, 9	10	Right
4	netustm	Numeric	4	0	Internet use, how much time on typical day, i...	{6666, Not a...	6666 - 9999	9	Right
5	[redacted]	Numeric	1	0	How interested in politics	{1, Very inter...	7, 8, 9	9	Right
6	trstprl	Numeric	2	0	Trust in country's parliament	{0, No trust a...	77, 88, 99	9	Right
7	vote	Numeric	1	0	Voted last national election	{1, Yes}...	7, 8, 9	6	Right
8	prtvtecz	Numeric	2	0	Party voted for in last national election, Czechia	{1, KSČM}...	66 - 99	10	Right
9	sgnptit	Numeric	1	0	Signed petition last 12 months	{1, Yes}...	7, 8, 9	9	Right
10	pbldmn	Numeric	1	0	Taken part in lawful public demonstration last ...	{1, Yes}...	7, 8, 9	8	Right
11	lrscale	Numeric	2	0	Placement on left right scale	{0, Left}...	77, 88, 99	9	Right
12	[redacted]	Numeric	1	0	Gender	{1, Male}...	9	6	Right
13	agea	Numeric	4	0	Age of respondent, calculated	{999, Not ava...	999	6	Right
14	yrbrn	Numeric	4	0	Year of birth	{7777, Refus...	7777, 8888, ...	7	Right
15	marsts	Numeric	2	0	Legal marital status	{1, Legally m...	66 - 99	8	Right
16	eisced	Numeric	2	0	Highest level of education, ES - ISCED	{0, Not possi...	77, 88, 99	8	Right
17	eduyrs	Numeric	2	0	Years of full-time education completed	{77, Refusal}...	77, 88, 99	8	Right
18	[redacted]	Numeric	1	0	Ever had a paid job	{1, Yes}...	6 - 9	9	Right
19	hinctnta	Numeric	2	0	Household's total net income, all sources	{1, J - 1st de...	77, 88, 99	10	Right
20	[redacted]	String	5	0	Region	{99999, Not ...	99999	8	Left

Tabulka četností

- Analyze → Descriptive Statistics → Frequencies



Tabulka četností

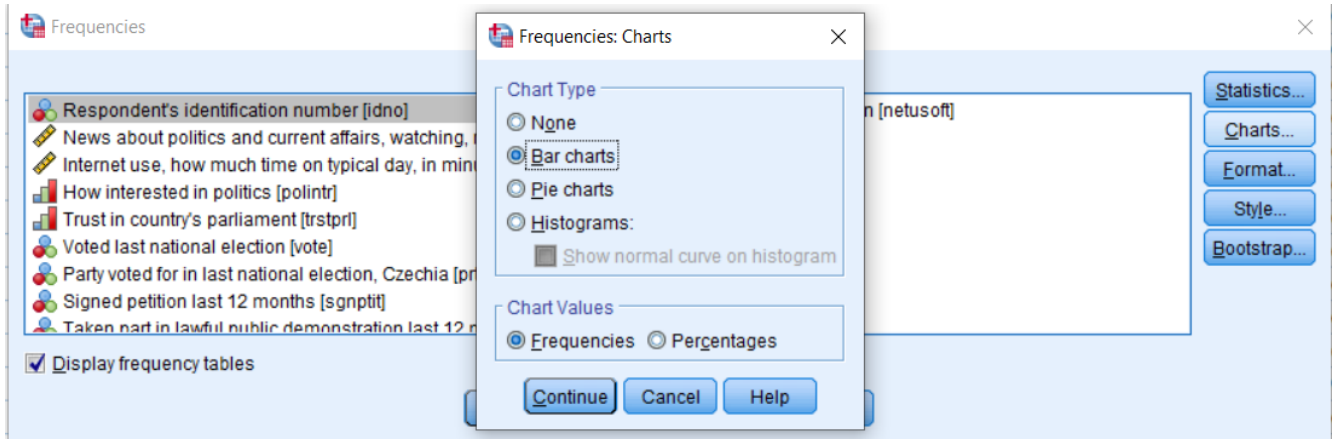
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Never	281	11,7	11,8	11,8
	Only occasionally	186	7,8	7,8	19,6
	A few times a week	269	11,2	11,3	30,8
	Most days	393	16,4	16,5	47,3
	Every day	1259	52,5	52,7	100,0
	Total	2388	99,6	100,0	
Missing	Don't know	10	,4		
Total		2398	100,0		

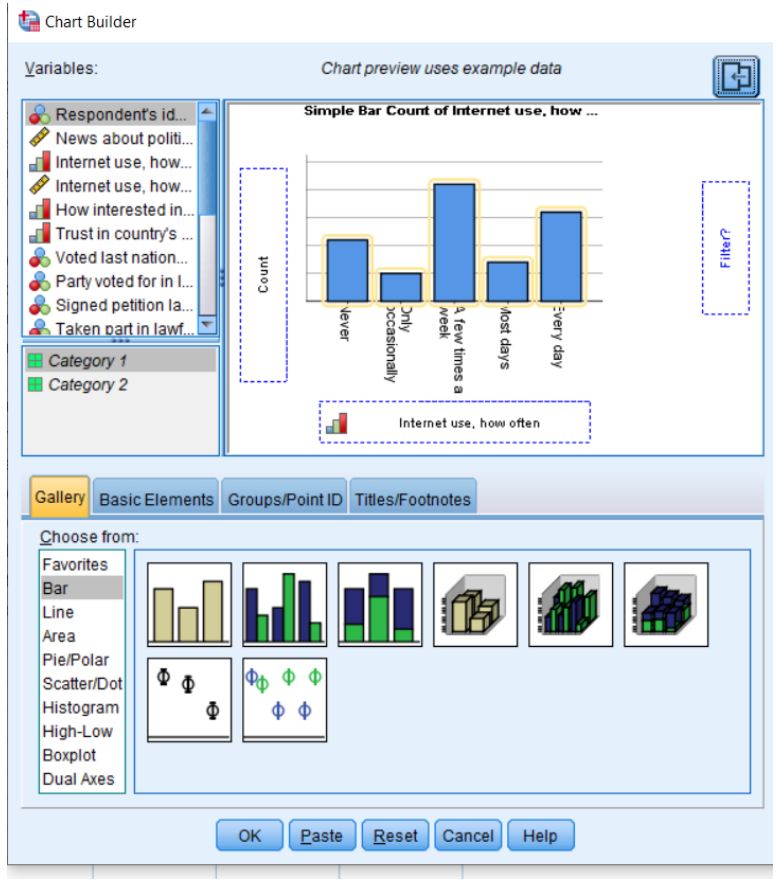
Vizualizace

- Základní pravidlo – nekomplikovat si grafy (život)
- Cíl – vizualizovat distribuci hodnot
- Plně postačí jednoduché sloupcové grafy (bar charts)
- Je zbytečné přidávat různé prvky typu 3D, kombinovat barvy, používat pro efekt koláčové grafy (pie charts) atd.

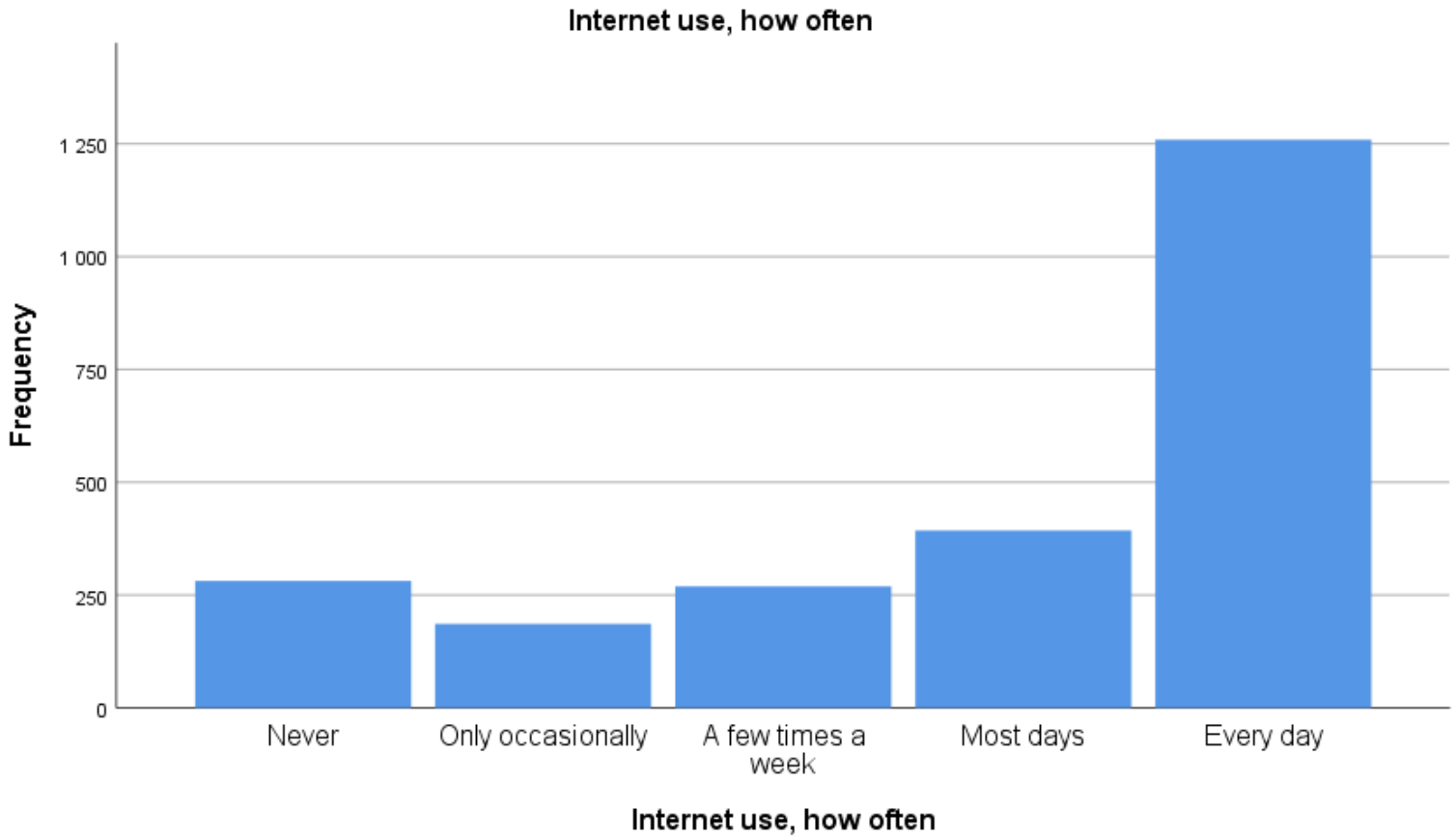
Vizualizace

- Analyze → Descriptive Statistics → Frequencies → Charts
- Graphs → Chart Builder → Bar…

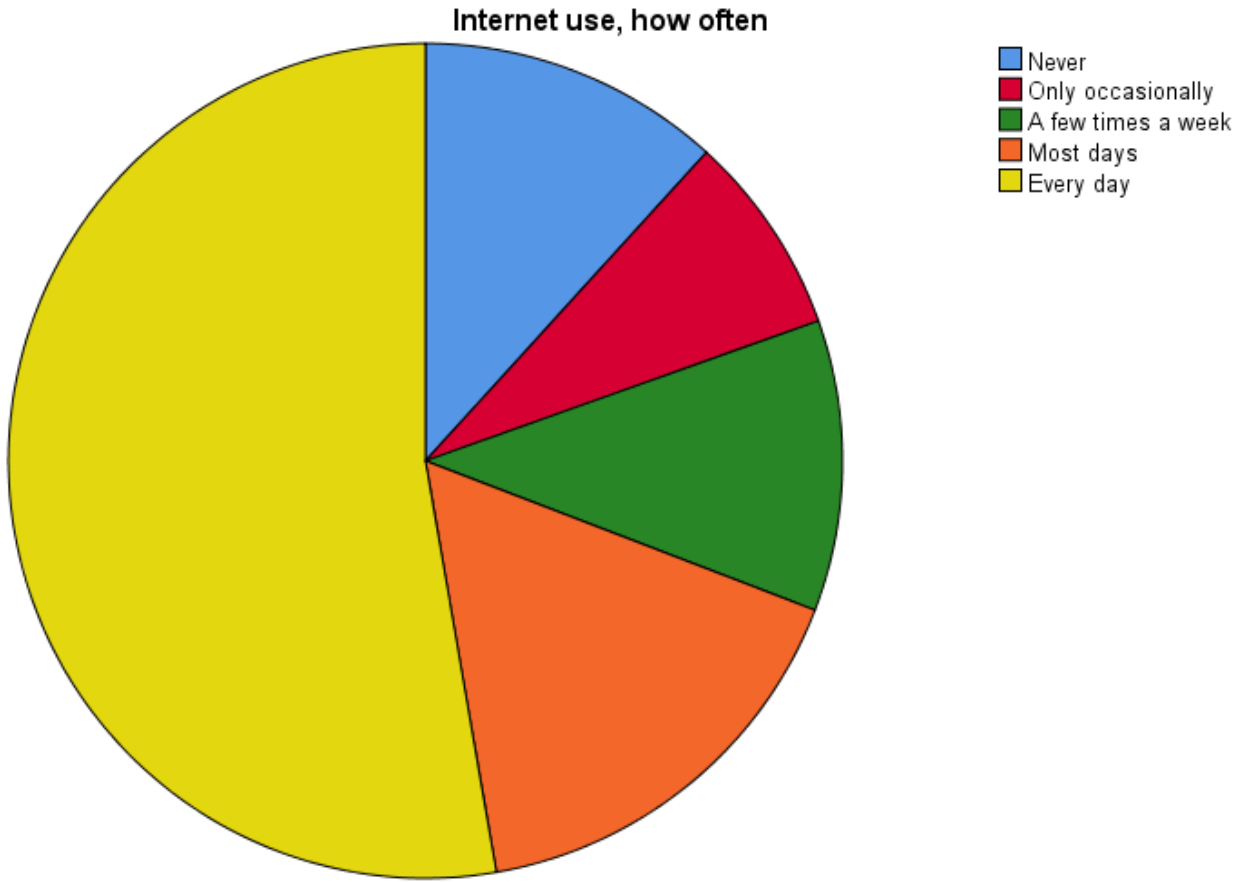


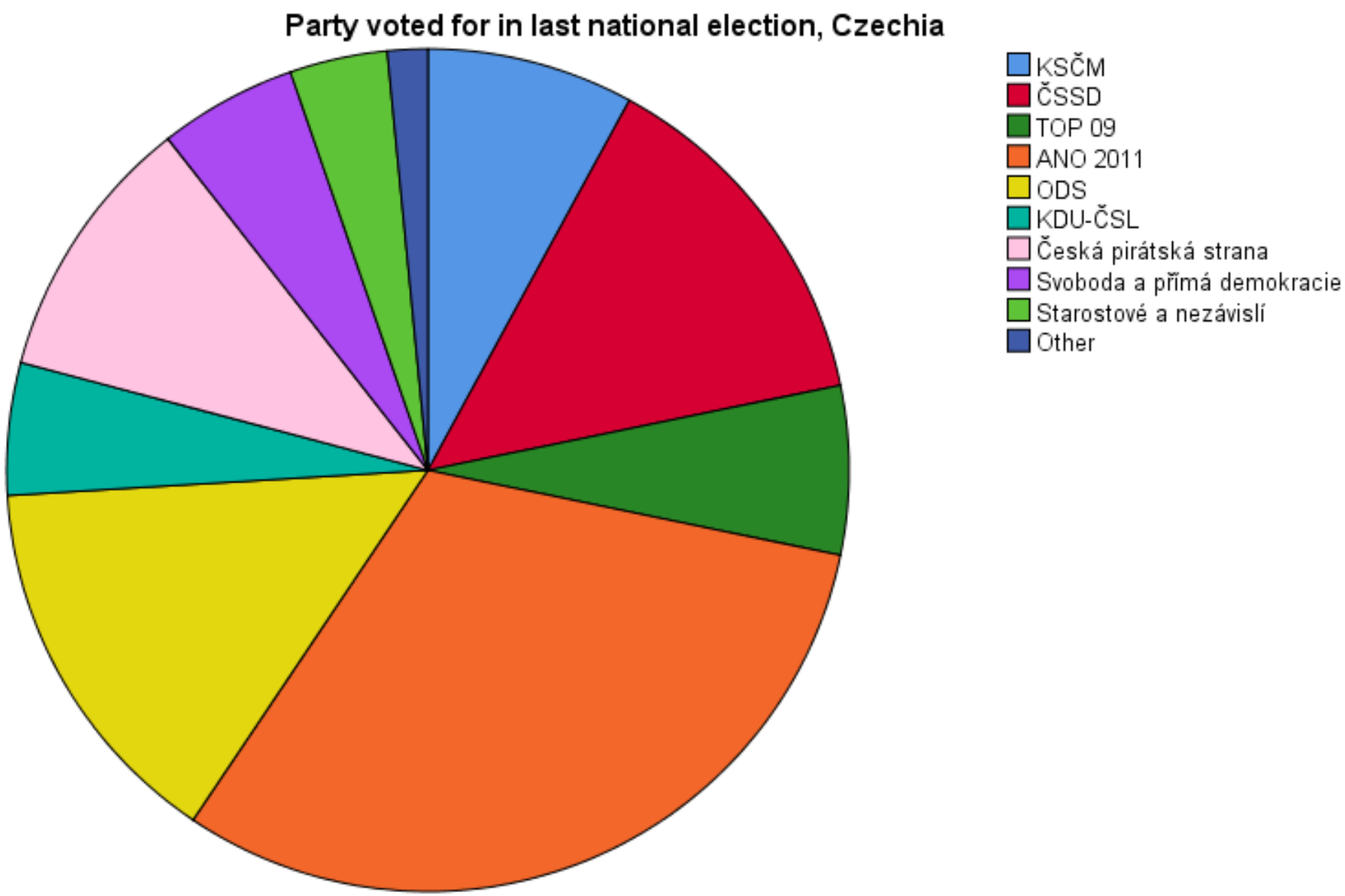


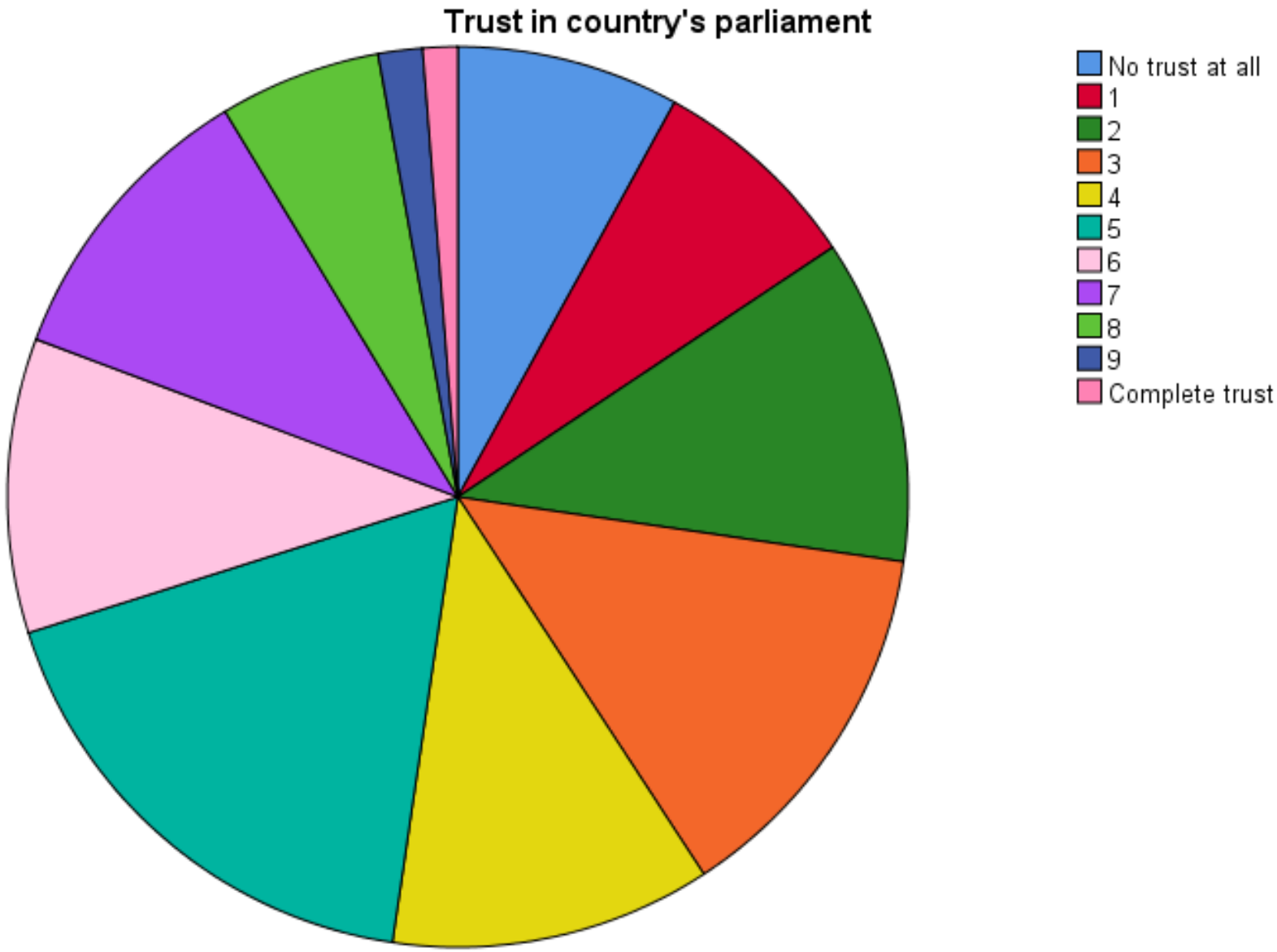
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Never	281	11,7	11,8	11,8
	Only occasionally	186	7,8	7,8	19,6
	A few times a week	269	11,2	11,3	30,8
	Most days	393	16,4	16,5	47,3
	Every day	1259	52,5	52,7	100,0
	Total	2388	99,6	100,0	
Missing	Don't know	10	,4		
Total		2398	100,0		



		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Never	281	11,7	11,8	11,8
	Only occasionally	186	7,8	7,8	19,6
	A few times a week	269	11,2	11,3	30,8
	Most days	393	16,4	16,5	47,3
	Every day	1259	52,5	52,7	100,0
	Total	2388	99,6	100,0	
Missing	Don't know	10	,4		
Total		2398	100,0		



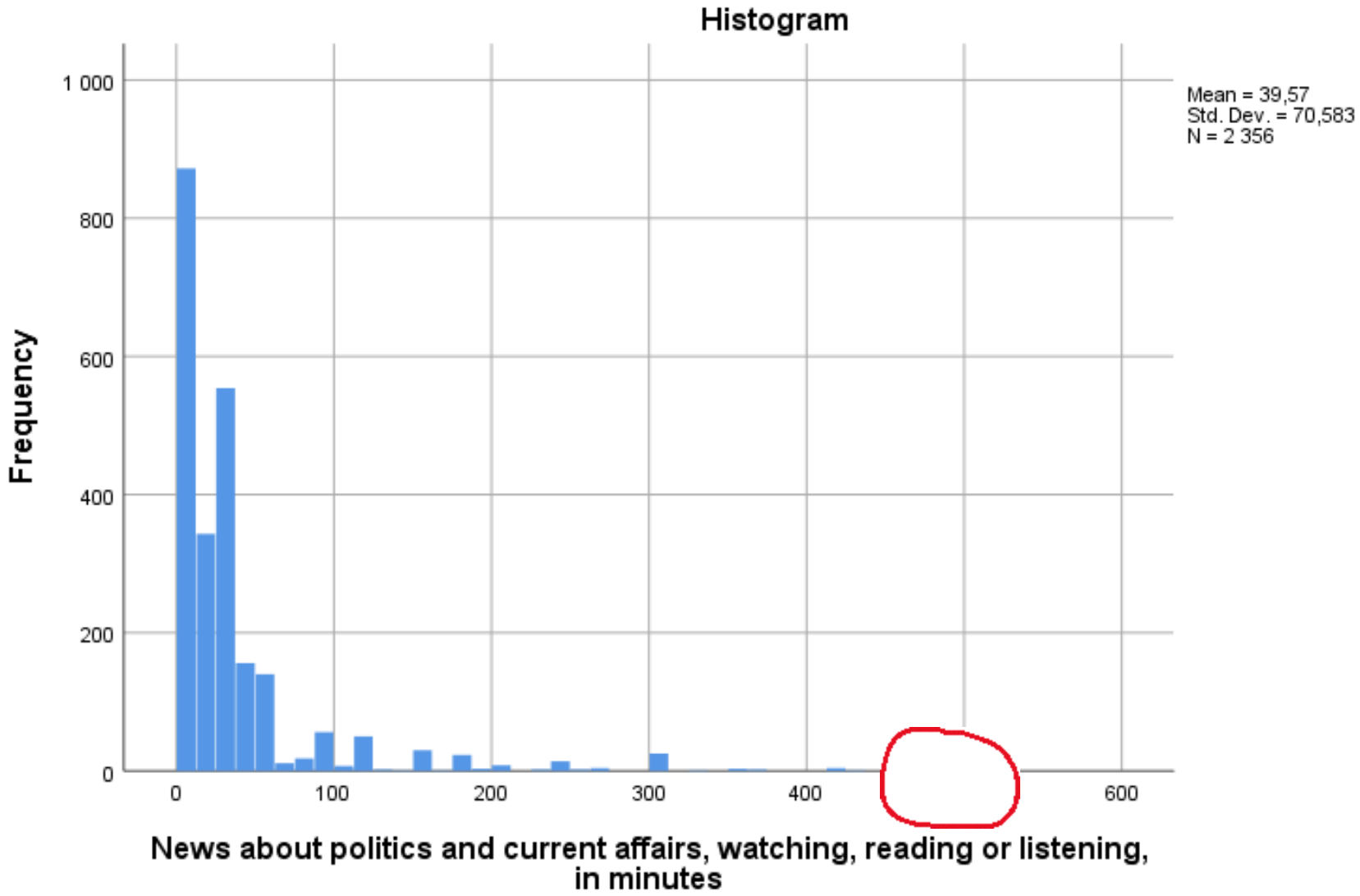

Party voted for in last national election, Czechia
KSČM
ČSSD
TOP 09
ANO 2011
ODS
KDU-ČSL
Česká pirátská strana
Svoboda a přímá demokracie
Starostové a nezávislí
Other


Trust in country's parliament
No trust at all
1
2
3
4
5
6
7
8
9
Complete trust

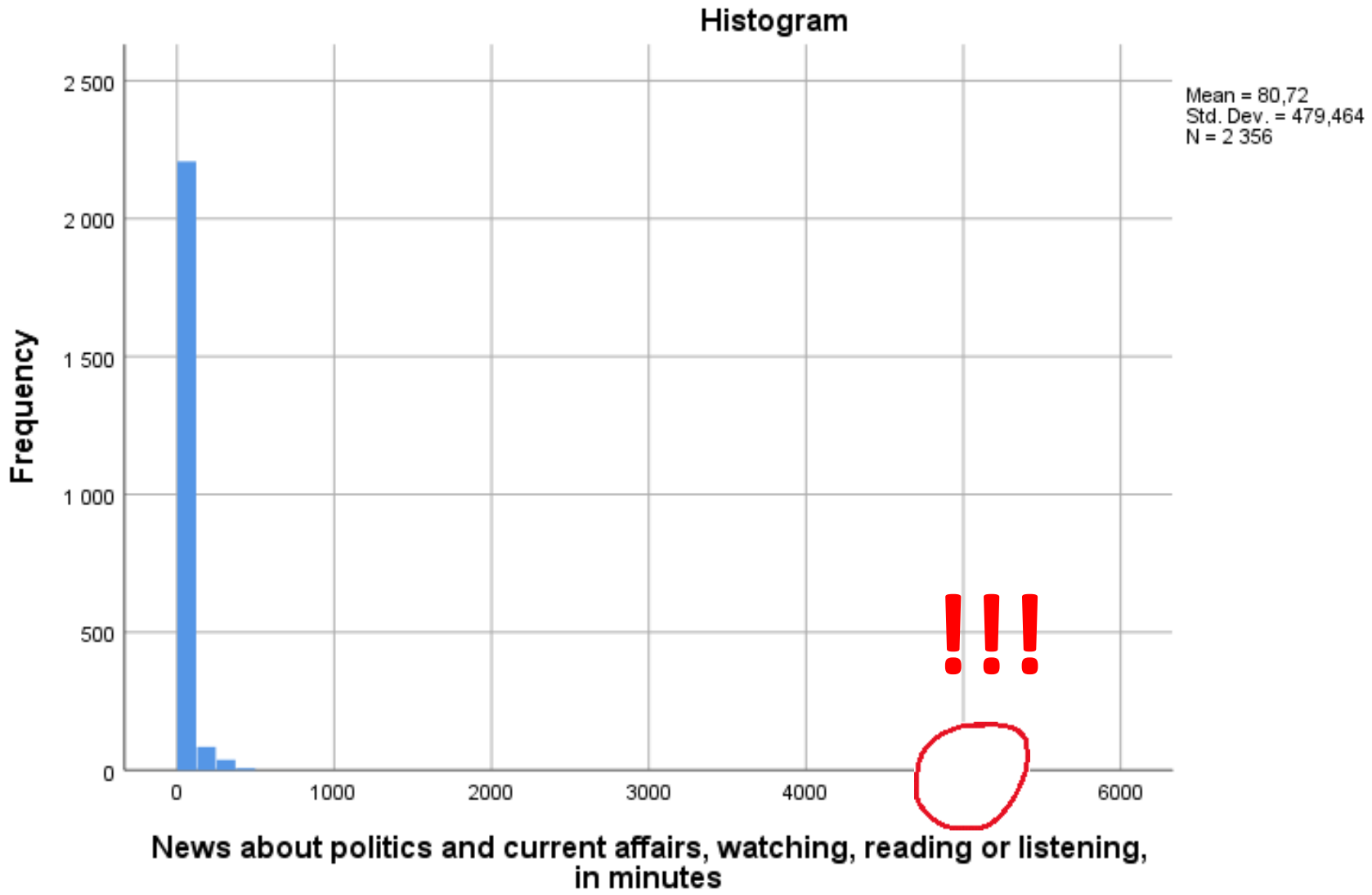
Kardinální proměnné

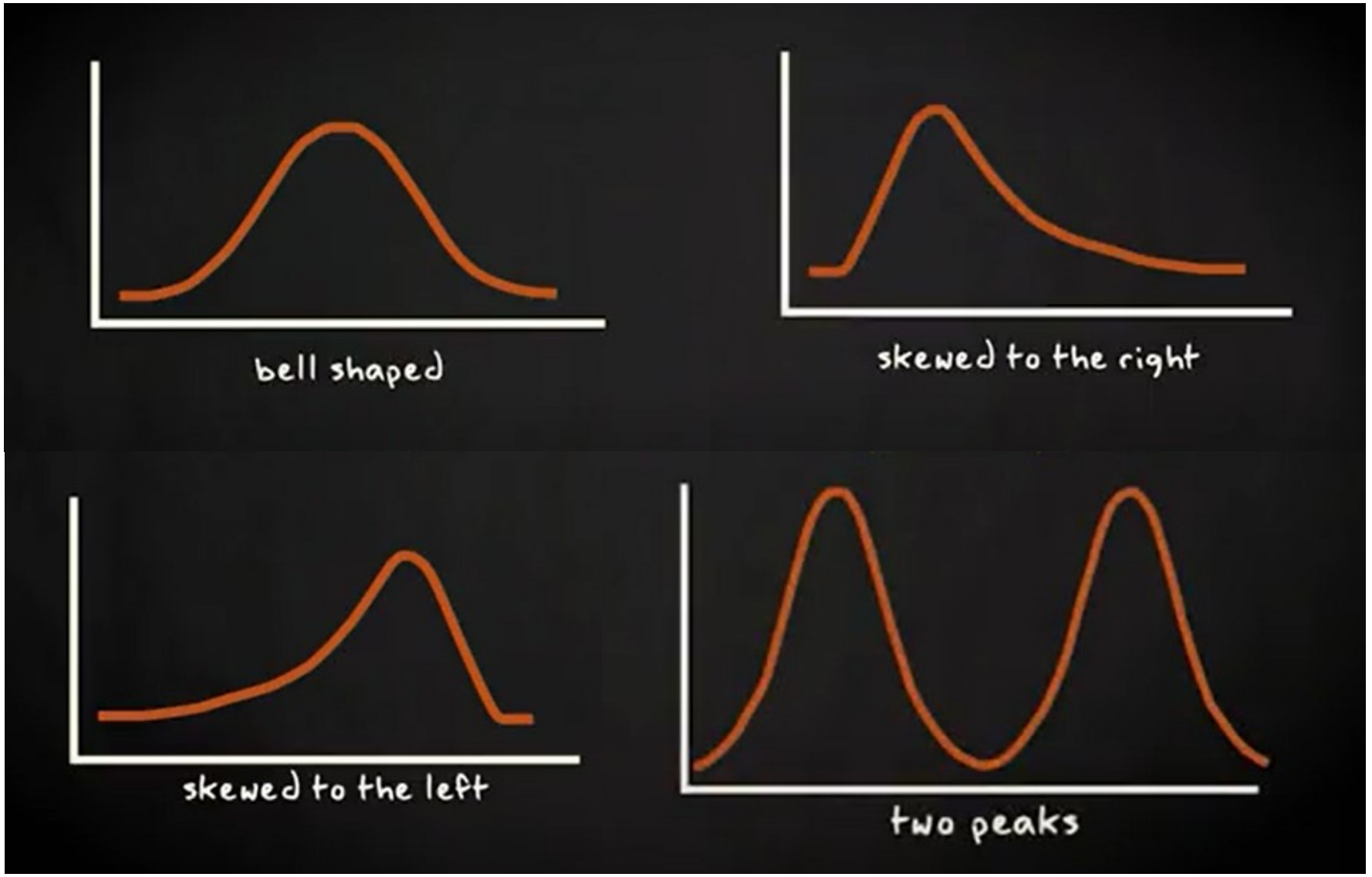
- Intervalové a poměrové (SPSS nerozlišuje – obě jsou *scale*)
- Numerické kódy (zpravidla) odpovídají reálným pozorovaným hodnotám
- Více možností jednorozměrné analýzy oproti nominálním a ordinálním proměnným
- Vizualizace – stejná pravidla

Histogram


Histogram
Mean = 39,57
Std. Dev. = 70,583
N = 2 356
Frequency
1 000
800
600
400
200
0
0
100
200
300
400
600
News about politics and current affairs, watching, reading or listening, in minutes

Histogram


Histogram
Mean = 80,72
Std. Dev. = 479,464
N = 2 356
Frequency
2 500
2 000
1 500
1 000
500
0
0
1000
2000
3000
4000
6000
!!!
News about politics and current affairs, watching, reading or listening, in minutes



- Více – přednáška 15.3. o normální distribuci

Míry centrální tendence

- Užitečné nástroje k lepšímu poznání našich dat
  - Modus, medián, průměr

- Použití závisí od typu proměnné:
  - Nominální – modus
  - Ordinální – modus, medián
  - Kardinální – modus, medián, průměr

Modus

- Nejčastější hodnota
- Frekvenční tabulka - nejvyšší hodnota
- Sloupcový graf / histogram - nejvyšší sloupec

- Využití při všech typech proměnných
- Modus nemusí být nutně pouze jeden (bimodální, multimodální distribuce)

Medián

- Středová hodnota, rozděluje dataset na dvě poloviny hodnot
  - Hodnota, pod kterou leží 50 % hodnot a nad kterou leží 50 % hodnot
  - V kategorických datech = mediánová kategorie (kumulativní četnost zahrnuje 50 % případů pod mediánem)
  - 50. percentil

- Postup:
  - Seřadíme hodnoty vzestupně
  - Najdeme tu, která leží uprostřed data setu (jednodušší pro matice s lichým počtem hodnot)

- Výhoda: je stabilní, není citlivý na extrémní hodnoty

Medián - příklad

- Počet hodin denně na sociálních sítích (9 lidí): 7, 0, 15, 8, 4, 6, 3, 10, 1
  - Seřazení → 0, 1, 3, 4, 6, 7, 8, 10, 15
  - Výběr hodnoty uprostřed (5. v pořadí) → **6**

- Co když máme sudý počet pozorování?
  - 8 lidí, stejný příklad: 7, 0, 15, 8, 4, 6, 3, 10
  - Seřazení → 0, 3, 4, 6, 7, 8, 10, 15
  - Medián je uprostřed dvou prostředních naměřených hodnot: (6+7)/2 = **6,5**

- Sudý a lichý počet – při velkém počtu dat je rozdíl věcně zanedbatelný

Průměr

- Aritmetický průměr = součet hodnot / počet případů
- Pouze u kardinálních proměnných
- Citlivý na extrémní hodnoty
- Průměrná mzda vs. mediánová mzda

Měsíčné příjmy hostů restaurace v tis. Kč

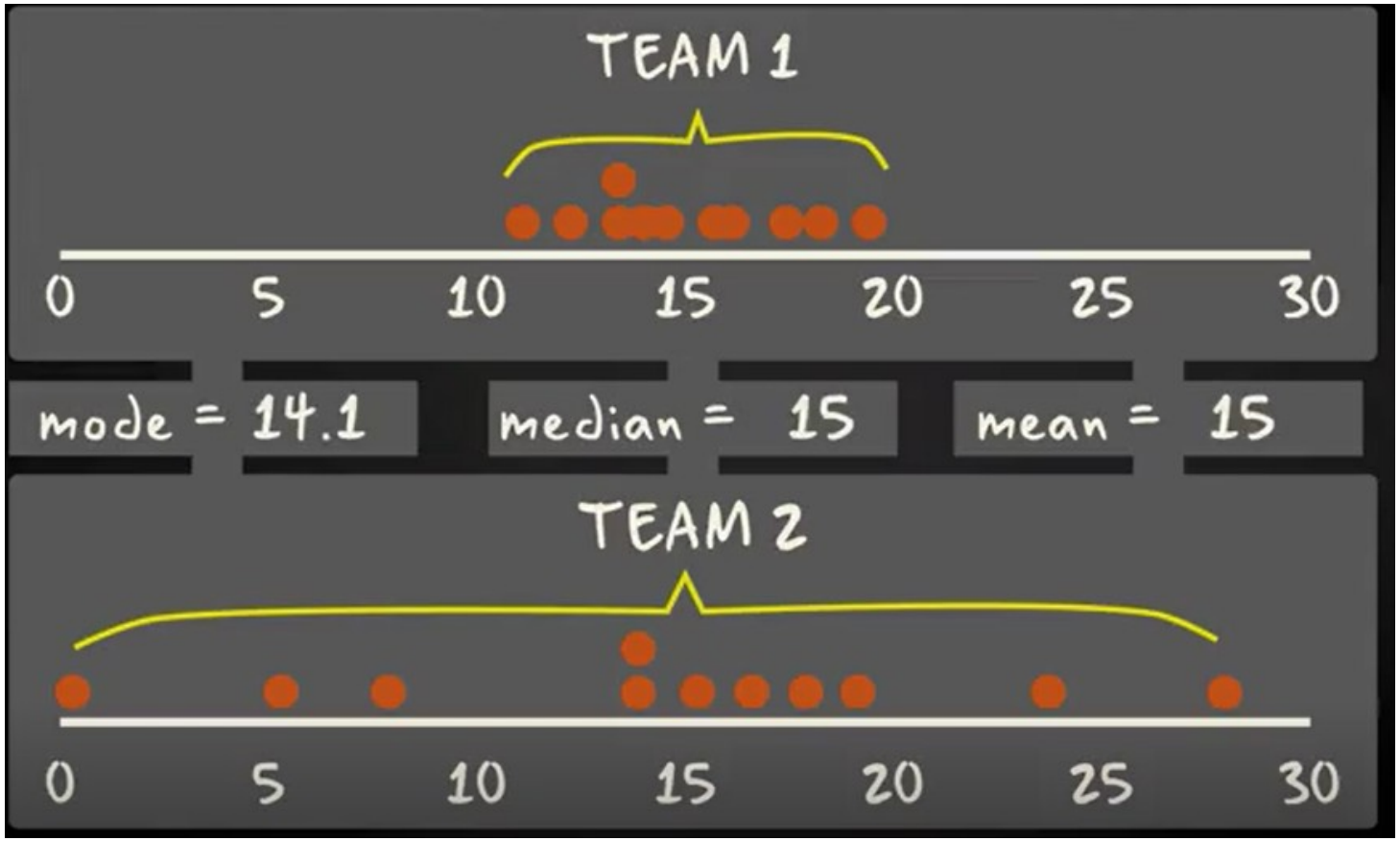
- **Příklad 1:**
  - 11 hostů: 20, 30, 35, 40, 45, 50, 55, 60, 70, 75, 80
  - Medián = 50k
  - Průměr = 50,9k

- **Příklad 2:**
  - 13 hostů: 20, 30, 35, 40, 45, 50, 55, 60, 70, 75, 80, 400, 450
  - Medián = 55k
  - Průměr = 108,5k

- **Příklad 3:**
  - Do restaurace vstoupí Elon Musk a Bill Gates
  - Medián = ?
  - Průměr = ?

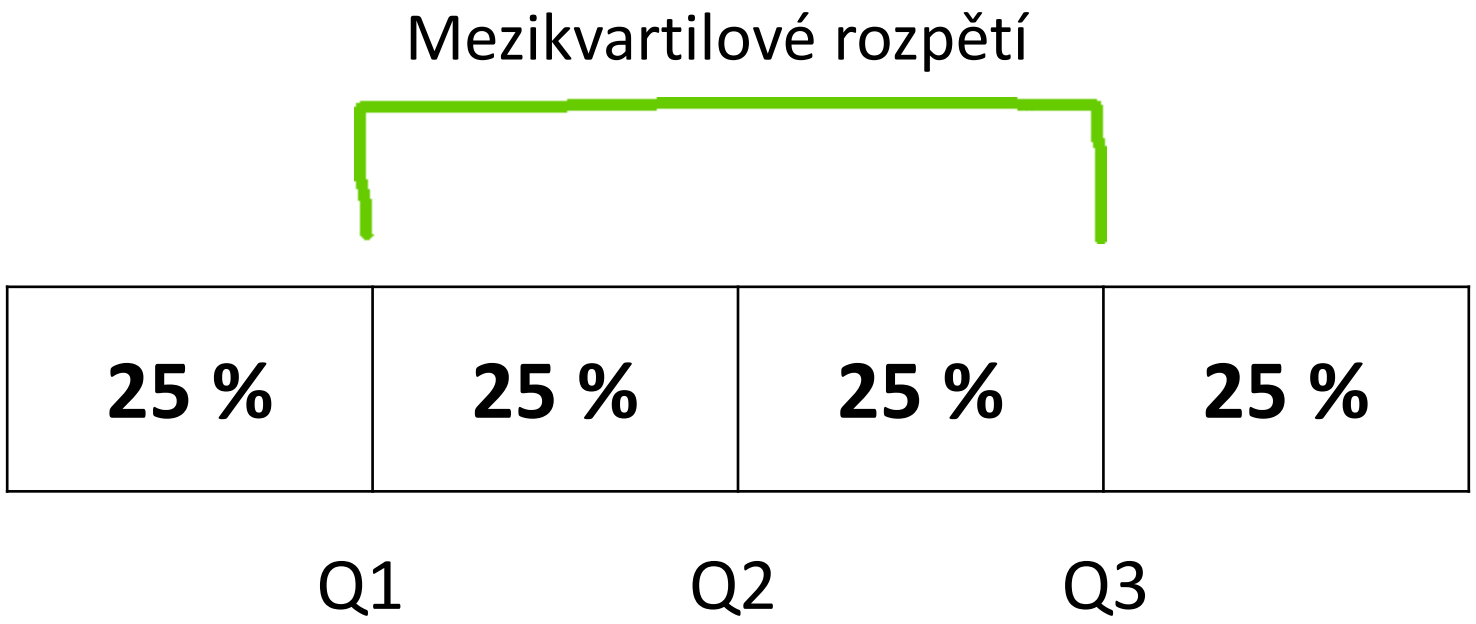
Míry centrální tendence

- Užitečné ukazatele, někdy však nemusí stačit
- Např. dva soubory dat mají stejné průměry, ale ve skutečnosti se dost odlišují
- Důležité je znát i míru rozptýlení dat (dispersion)



Mezikvartilové rozpětí

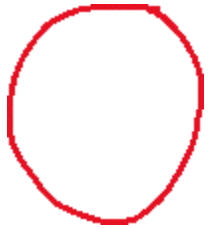
- Interquartile range (IQR)
- Umožňuje snížit citlivost na odlehlé případy
- Kvartily – hodnoty, které rozdělují soubor dat na 4 stejně velké skupiny
- První kvartil (Q1), druhý kvartil (Q2), třetí kvartil (Q3)



- **IQR = Q3 – Q1**

- Co je Q2?

Mezikvartilové rozpětí - postup

1 2,7 4,3 8,9 11,4  19,0 25,1 31,2 32,8 65,4

Najdeme Q2

Mezikvartilové rozpětí - postup

Najdeme Q1 a Q3

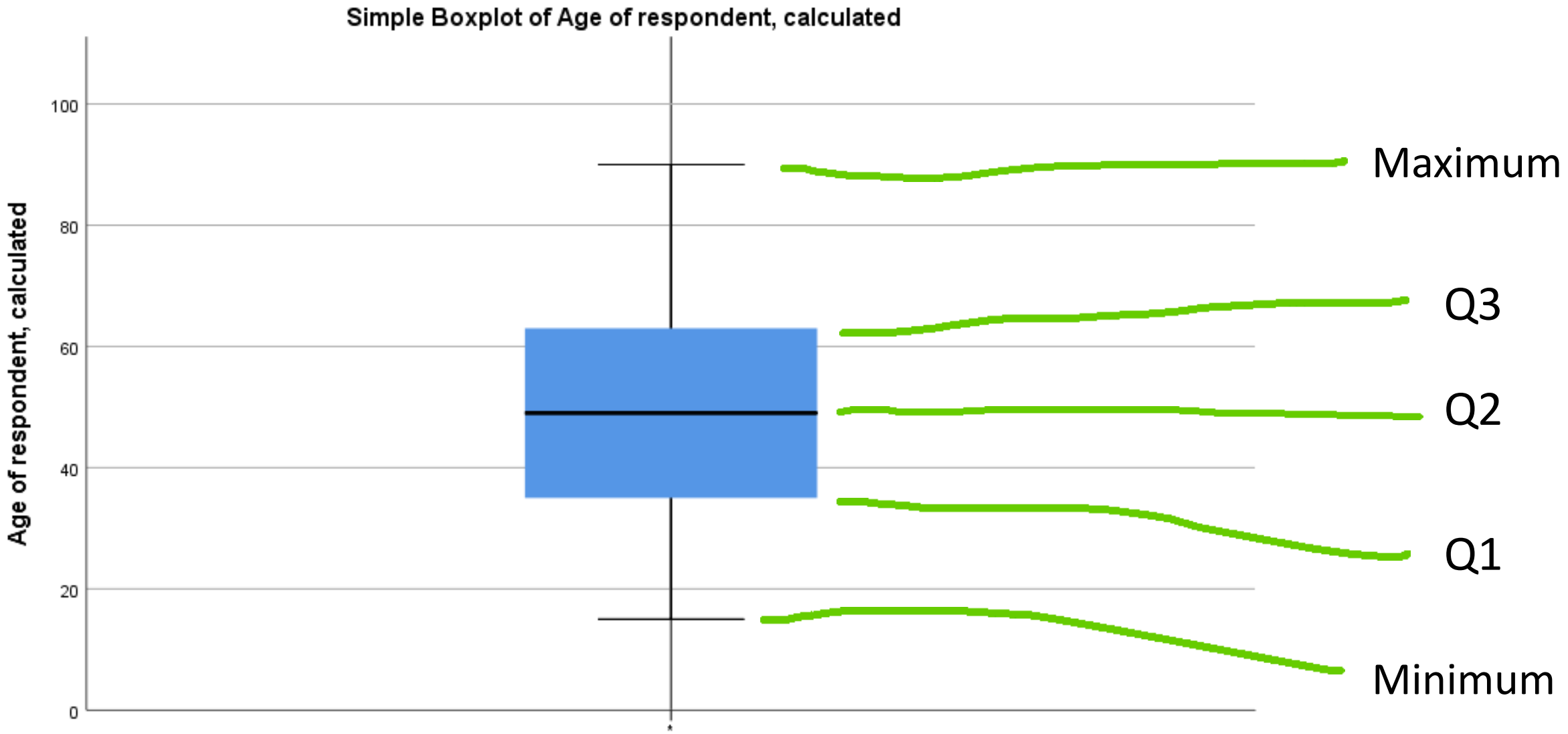
IQR = Q3 – Q1 = 31,2 – 4,3 = **26,9**

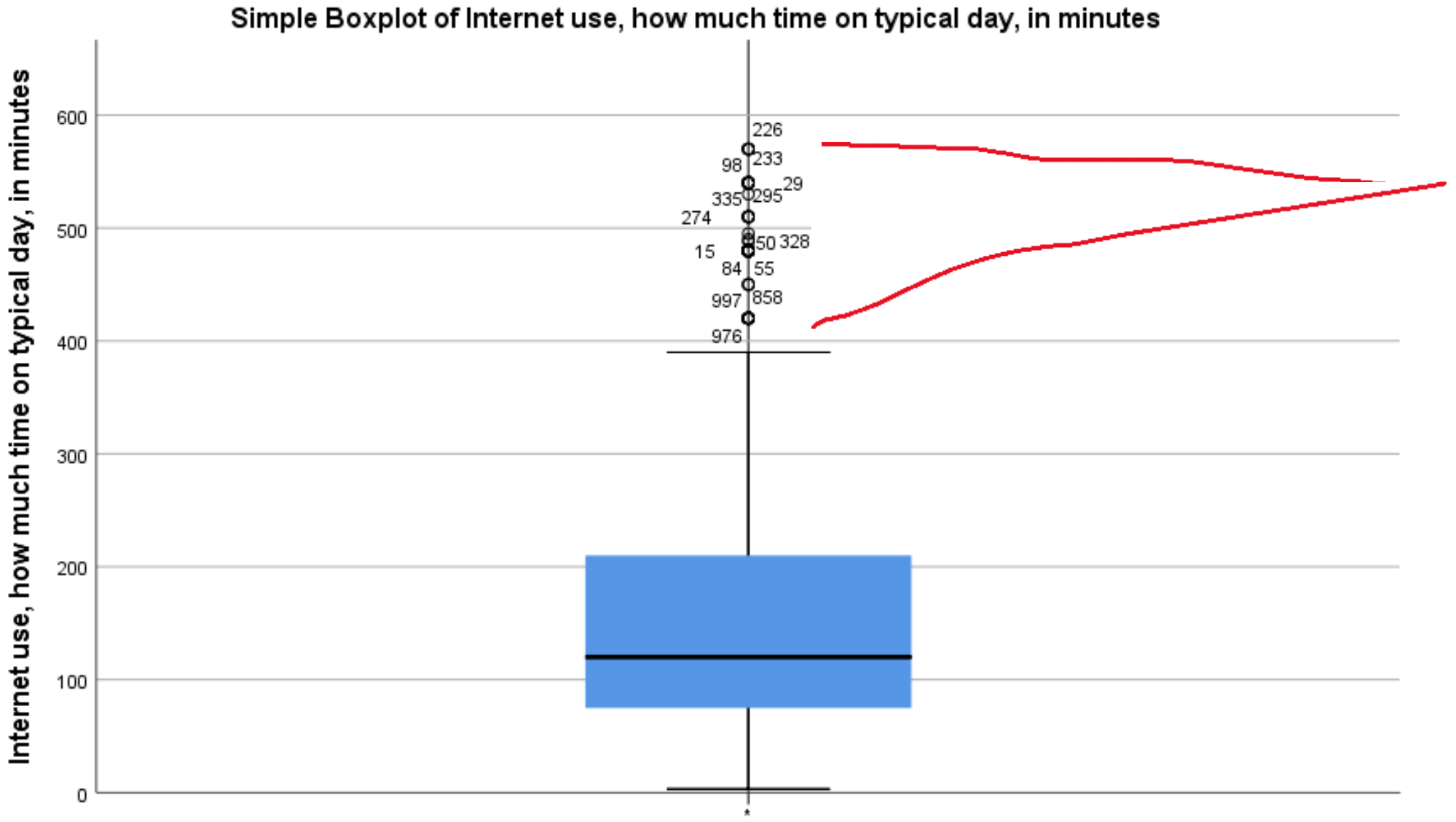
Odlehlé případy (outliers)

- < Q1 – 1,5*IQR
- > Q3 + 1,5*IQR

- Outliers leží za těmito hodnotami
- Vhodné poznat pro určité druhy analýzy (vliv na výsledky)

- Spočítaní nebo vizualizace pomocí krabicového grafu (boxplot)

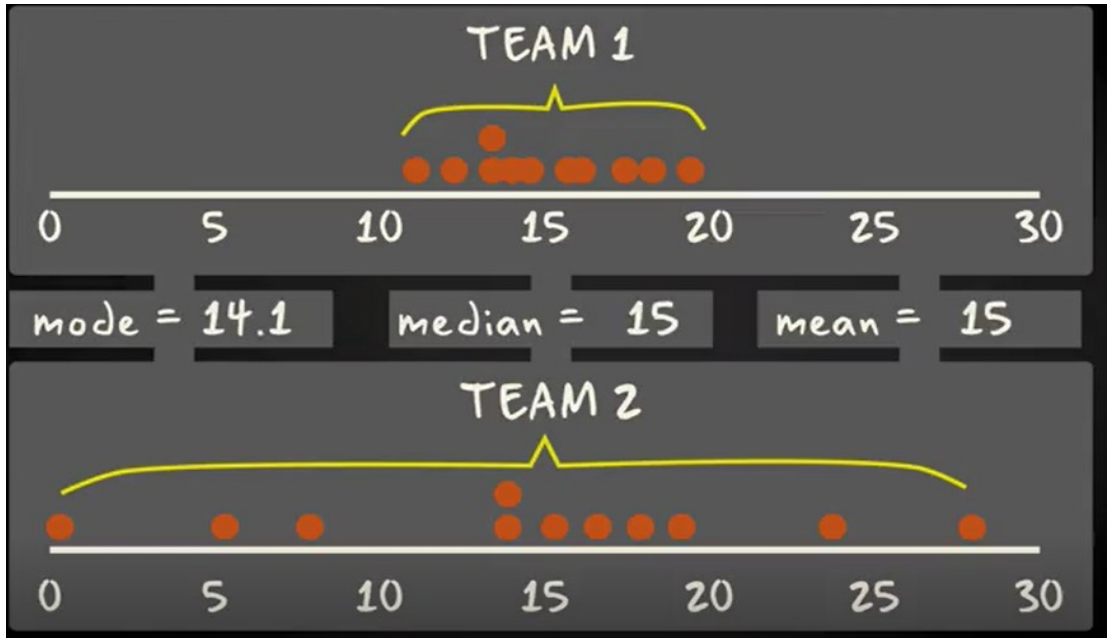

Simple Boxplot of Age of respondent, calculated
Age of respondent, calculated
100
80
60
40
20
0
Maximum
Q3
Q2
Q1
Minimum


Simple Boxplot of Internet use, how much time on typical day, in minutes
Internet use, how much time on typical day, in minutes
600
500
400
300
200
100
0
226
98
233
29
335
295
274
15
50
328
84
55
997
858
976
*

Outliers

Rozptyl (variance)

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$



- Vyšší hodnota indikuje více rozptýlená data
- Nevýhoda – uvádí se v jednotkách proměnné ale na druhou
- Řešení – směrodatná odchylka (standard deviation - SD)

Rozptyl

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

Hráči	x	x - x̄	$(x - \bar{x})^2$
Hráč 1	0	-15	$(-15)^2 = 225$
Hráč 2	24,1	9,1	82,81
Hráč 3	5,6	-9,4	88,36
Hráč 4	14,1	-0,9	0,81
Hráč 5	17,2	2,2	4,84
Hráč 6	8,7	-6,3	39,69
Hráč 7	19,2	4,2	17,64
Hráč 8	14,1	-0,9	0,81
Hráč 9	27,7	12,7	161,29
Hráč 10	15	0	0
Hráč 11	19,3	4,3	18,49
		0	639,74

$\bar{x} = 15$

$\Sigma(x - \bar{x})^2$ suma čtverců

n-1 = 10

639,74/10 = 63,97

Směrodatná odchylka (standard deviation)

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

- Odmocnina rozptylu
- Větší hodnota sm. odchylky = větší variabilita v datech

- Příklad fotbalového týmu:
  - Rozptyl = 63,97
  - Směrodatná odchylka = 8