

Analýza dát

Prednáška 1

Základy analýzy dát

3 myšlienky na úvod

Čísla – v správnych rukách – majú moc meniť svet k lepšiemu. Dobrá štatistika nie je nečestná finta, ale nástroj, pomocou ktorého zbadáme to, čo by inak ostalo neviditeľné.

Tim Harford „Ako spočítať svet“

„Dáta ako terapia“, ak ich pochopíte, dosiahnete duševný pokoj. Pretože svet nie je taký dramatický, ako sa zdá.

Hans Rosling „Moc faktov“

Všetci sme nejako začínali. Nie je chyba robiť chyby. Naopak. Čím viac ich urobíš, tým rýchlejšie sa učíš, kadiaľ cesta k cieľu nevedie.

Veronika Hurdová „Agnes a ostrov stínů“

Študentský prieskum

- Prieskum sa uskutočnil medzi študentmi v úvodnom kurze analýzy dát. Nižšie je uvedených niekoľko **otázok** v prieskume a príslušné **premenné**, do ktorých boli uložené údaje z odpovedí:
- **pohlavie (gender)**: Aké je vaše pohlavie?
- **intro_extra**: Ste introvert alebo extrovert?
- **spánok (sleep)**: Koľko hodín v noci v priemere spíte?
- **čas spánku (bedtime)**: V akom čase zvyčajne chodíte spať?
- **krajiny (countries)**: Koľko krajín ste navštívili?
- **strach (dread)**: Na stupnici od 1 do 5, ako veľmi sa bojíte byť tu?

Dátová matica (dátový súbor)

- Odpovede zo študentského prieskumu boli zaznamenané vo forme dátovej matice, ktorej riadky predstavujú pozorovania (štatistické jednotky) a stĺpce jednotlivé premenné.

	<i>variable</i> ↓	<i>=premenná</i>			
Stu.	gender	intro_extra	...	dread	
1	male	extravert	...	3	
2	female	extravert	...	2	
3	female	introvert	...	4	←
4	female	extravert	...	2	<i>observation</i>
⋮	⋮	⋮	⋮	⋮	<i>=pozorovanie</i>
86	male	extravert	...	3	

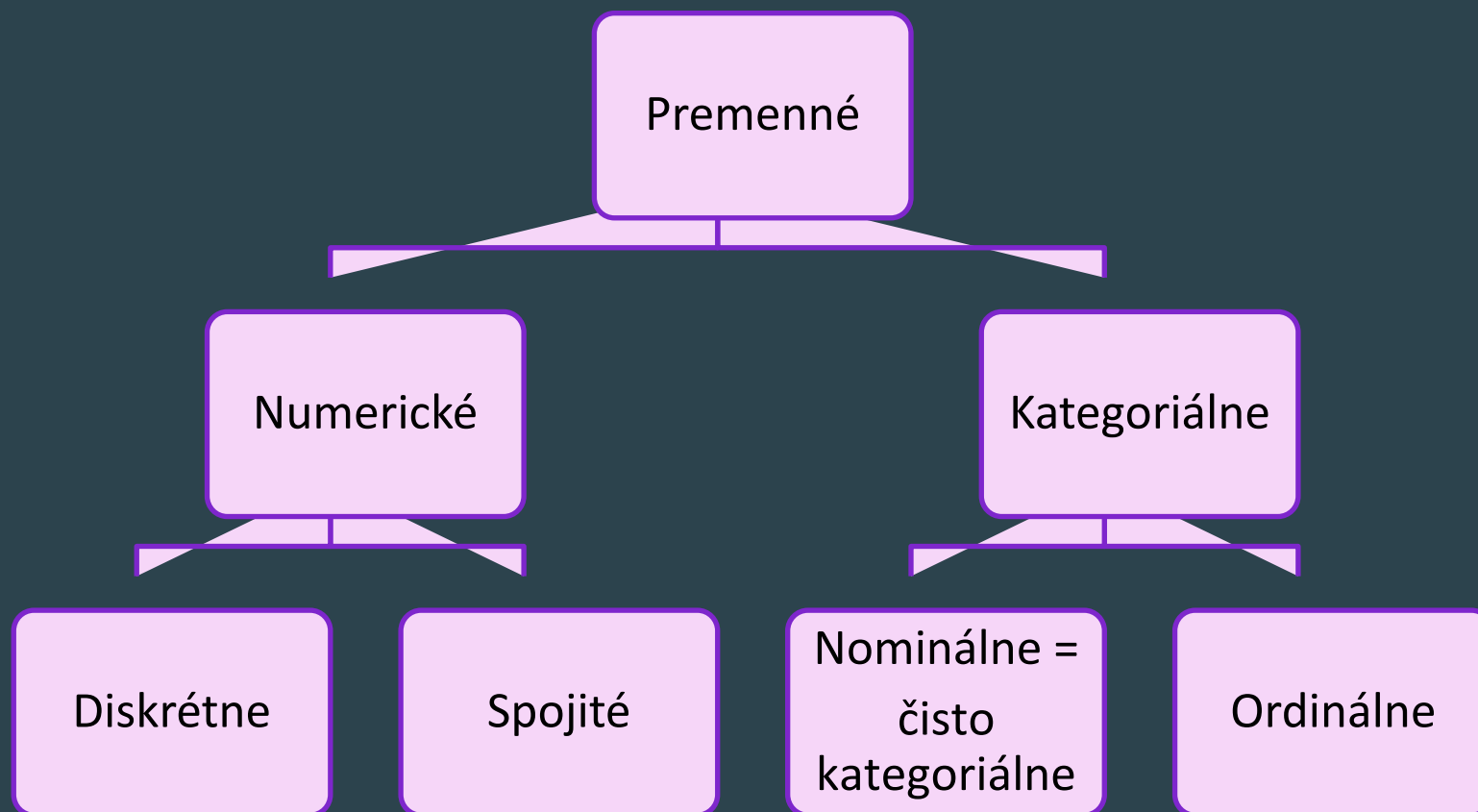
Každá premenná má svoje **pomenovanie** a svoje **hodnoty**; napr.

pomenovanie premennej: pohlavie
hodnoty premennej: muž, žena, iné

pomenovanie premennej: strach
hodnoty premennej: 1, 2, 3, 4, 5

pomenovanie premennej: spánok
hodnoty premennej: reálne čísla z intervalu $\langle 0, 24 \rangle$

Typy premenných



Typ premenných z prieskumu

- **pohlavie (gender):**
 - kategoriálna, nominálna
- **intro_extra:**
 - kategoriálna, nominálna
- **spánok (sleep):**
 - numerická, spojitá
- **čas spánku (bedtime):**
 - kategoriálna, ordinálna
- **krajiny (countries):**
 - numerická, diskretná
- **strach (dread):**
 - kategoriálna, ordinálna alebo numerická diskretná

Kategoriálne premenné

- **Kategoriálne (kvalitatívne, nominálne, slovné)** premenné vznikajú v tom prípade, keď jedincov zapojených do výskumu možno rozdeliť podľa nejakej vlastnosti do skupín. Skupinám možno priradiť číselnú hodnotu, ale nemá zmysel s nimi prevádzať numerické operácie.
 - **Nominálne** premenné: medzi vzniknutými skupinami nemá zmysel žiadne usporiadanie (pohlavie, národnosť, status).
 - **Ordinálne** premenné: medzi skupinami existuje zmysluplné usporiadanie (hodnotenie v škole, vojenské/akademické hodnosti, miera spokojnosti).

Numerické premenné

- **Numerické (kvantitatívne, číselné, kardinálne)** vznikajú zvyčajne ako výsledok merania (vek, váha, výška, IQ, krvný tlak, ale aj koľko strán prečítate zo svojej obľúbenej knihy, kým zaspíte 😊...), sú „citlivé“ na numerické operácie ako sčítanie, odčítanie, prípadne priemer
 - **Diskrétné:** udávajú počet nejakých položiek udávaný pomocou prirodzených čísel, pričom existuje zoznam možných hodnôt
 - **Spojité:** udávajú výsledok meraní, ich možné hodnoty sa nedajú vymenovať a udávame ich pomocou reálnych čísel

Kategoriálna premenná

Kategoriálna premenná – základné opisné štatistiky

- **Rozsah** = počet štatistických jednotiek v dátovom súbore označuje sa **n**
- **Absolútna početnosť** = počet štatistických jednotiek v príslušnej kategórii (označujeme **n_i** a udáva sa v prirodzených číslach);
- **Relatívna početnosť** = podiel počtu jedincov v príslušnej kategórii ku všetkým jedincom:
 - označuje sa **p_i** a udáva sa v reálnych číslach z intervalu $\langle 0, 1 \rangle$; **$p_i = \frac{n_i}{n}$**
 - Označuje **f_i** a udáva sa v percentách
- Údaje o absolútnej a relatívnej početnosti zaznamenávame zvyčajne do tzv. frekvenčnej tabuľky

Vizuálne spracovanie kategoriálnej premennej

Stĺpcový graf:

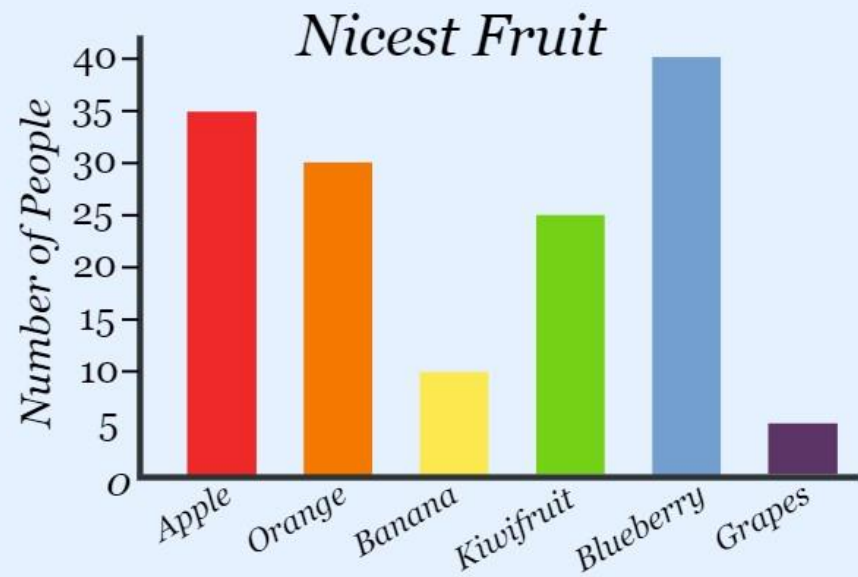
- na x-ovej osi názvy kategórií
- na y-ovej osi absolútna alebo relatívna početnosť príslušnej kategórie

Pruhový graf:

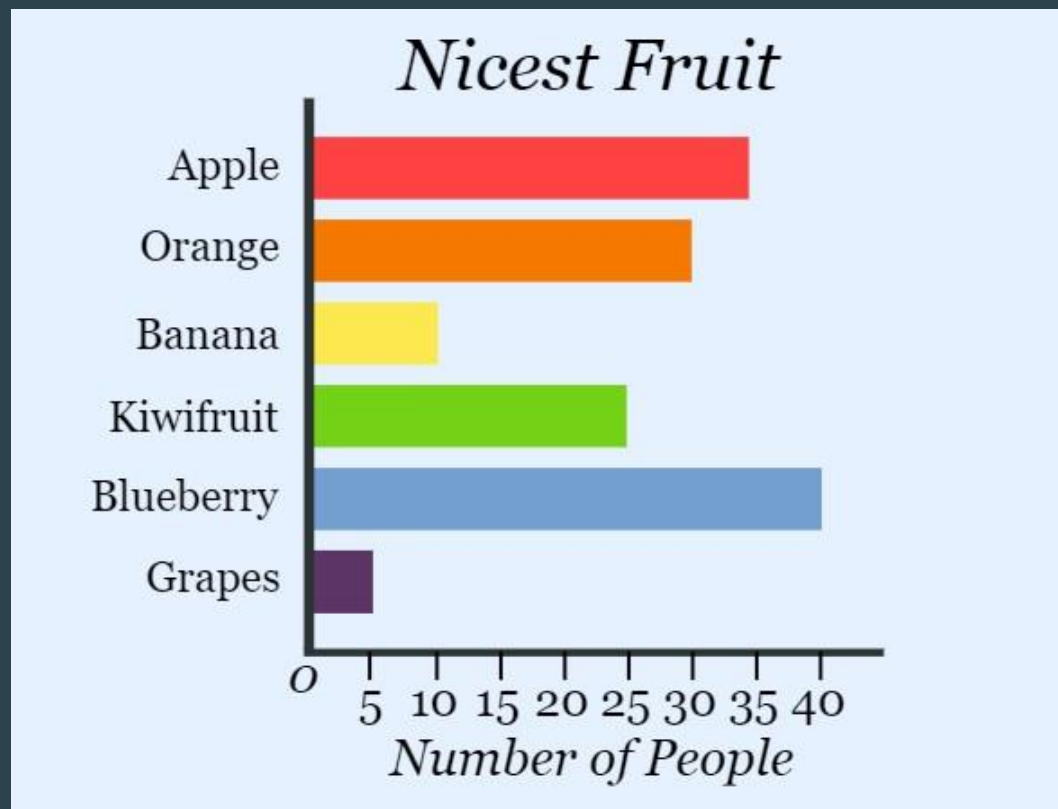
- na x-ovej osi je absolútna alebo relatívna početnosť;
- na y-ovej osi sú názvy kategórií

Koláčový graf:

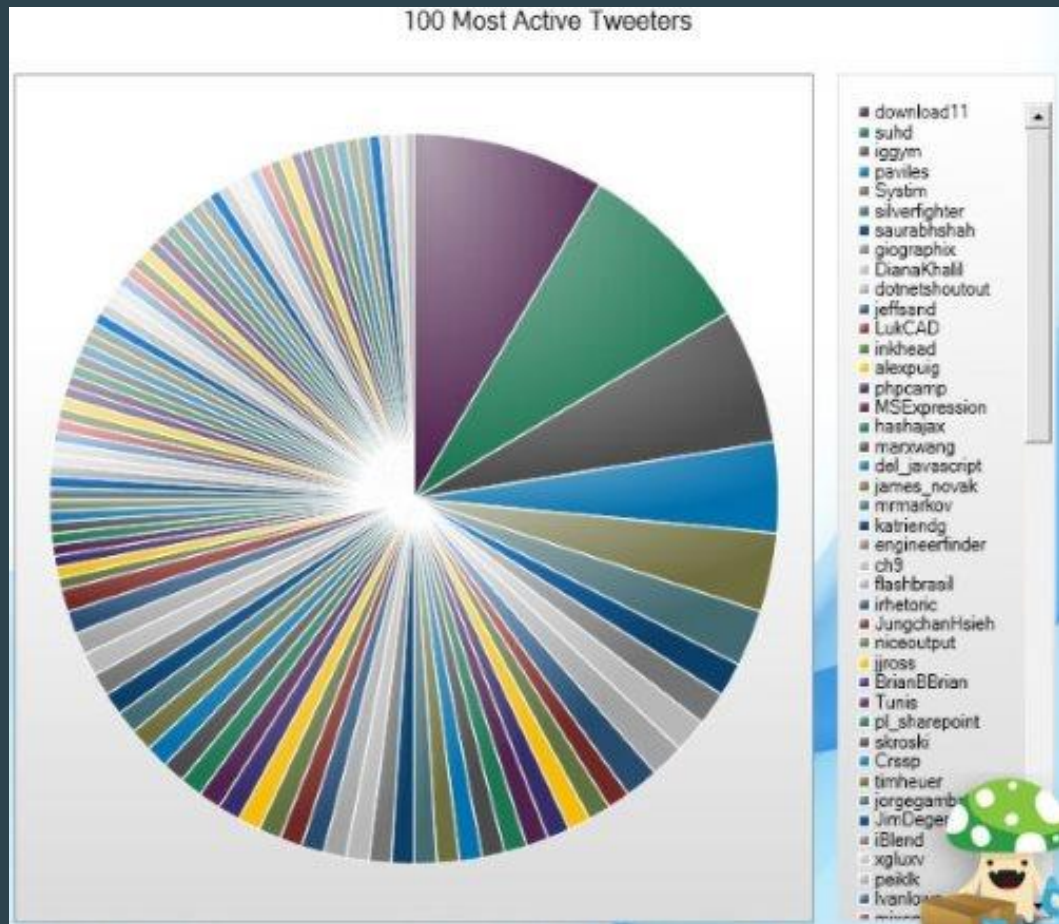
- vyjadruje rozdelenie celku na časti podľa relatívnej početnosti kategórií
- používame, ak nie je počet kategórií veľký a ak početnosti nie sú veľmi malé;
- radšej nie 3D variantu (skreslenie veľkosti jednotlivých častí)



Stípcový graf

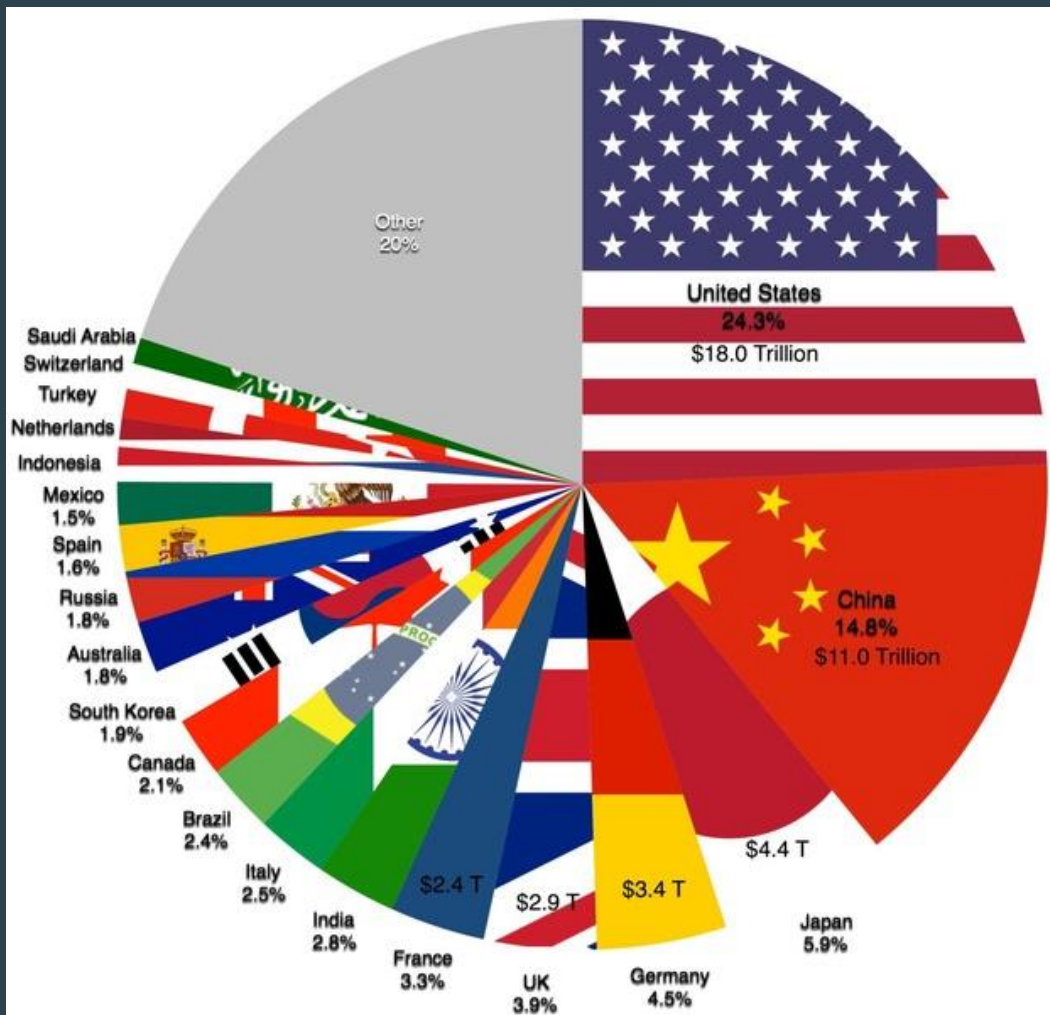


Pruhový graf

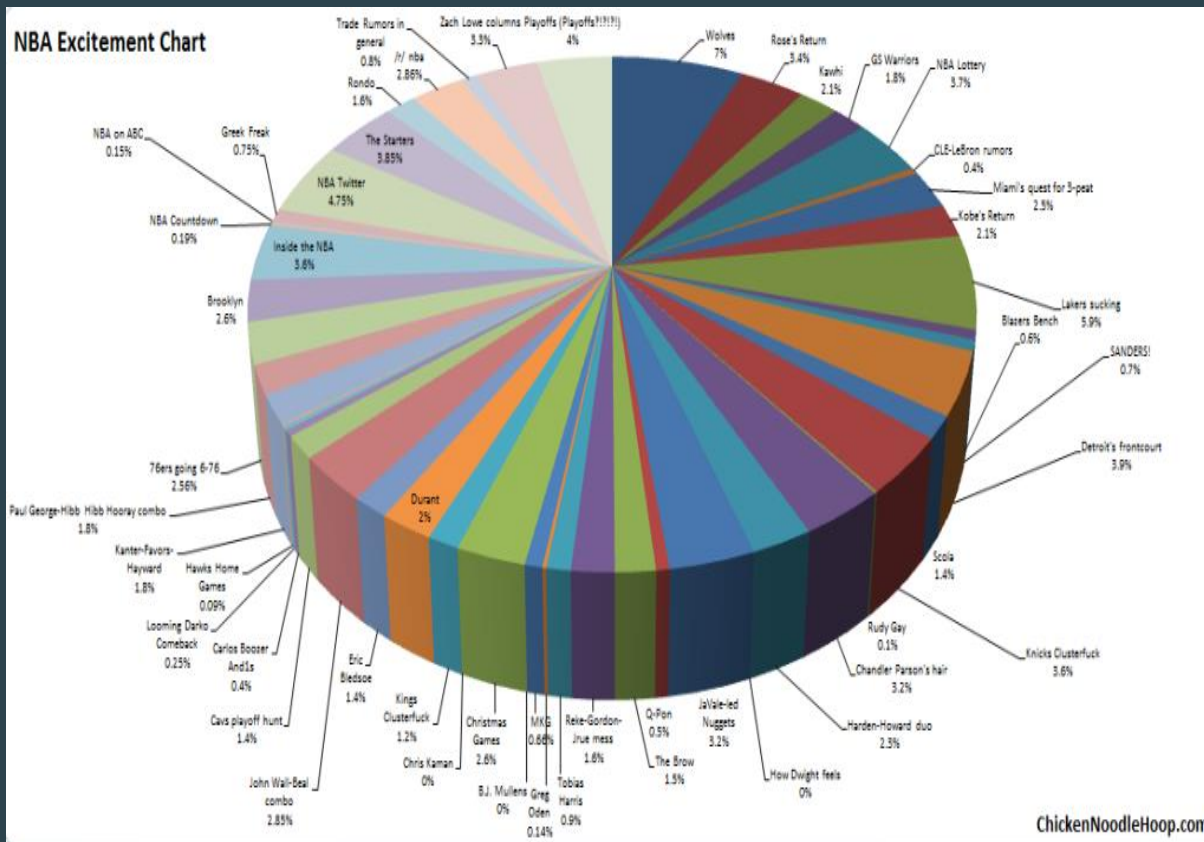


Koláčový graf

Koláčový graf



Koláčový graf



Ukážka spracovania kategoriálnej premennej

Numerická premenná

Opisné štatistiky pre numerické premenné

01

„Čísla“, ktoré dávajú obraz o všetkých hodnotách premennej na malom priestore;

02

Každá opisná štatistika má svoje výhody a nevýhody, treba ich poznať;

03

Obvykle je potrebná kombinácia viacerých, aby sme dostali komplexný obraz o premennej;

Opisné štatistiky pre numerické premenné

rozsah

minimum

maximum

variačné
rozpätie

Vizuálne spracovanie numerickej premennej (ak je v dátovom súbore málo hodnôt)

DOT PLOT = bodový graf

Dátový súbor

Movie Title	Profit
The Chicken and the Horse	50M
The Knight of Laughs	-10M
Criminal and Foreigner	10M
The Pig In the Hallway	0M
The Dinner Date	0M
Out of Time	20M
The Accidental Boss	20M
Confessions of a Biologist	20M
Funny You Should Ask	50M
Exactly	-10M

rozsah

$n = ?$

min = ?

max = ?

rozpätie

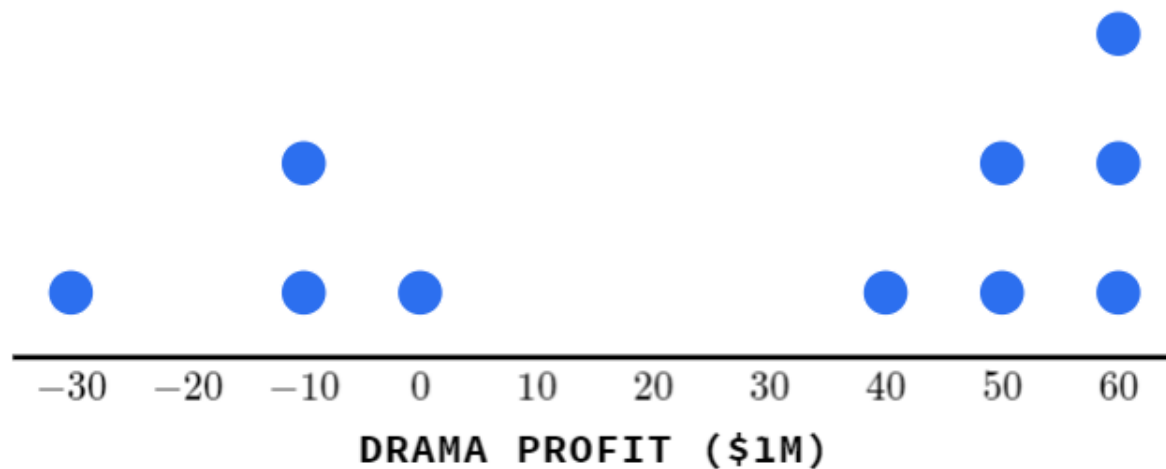
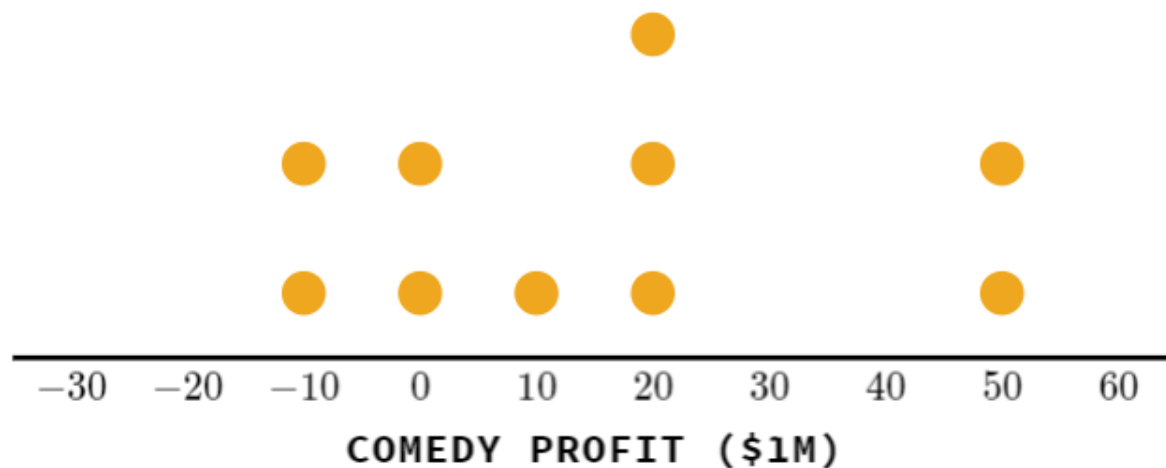
$R = ?$

Dot plot



Opisné štatistiky pre numerické premenné

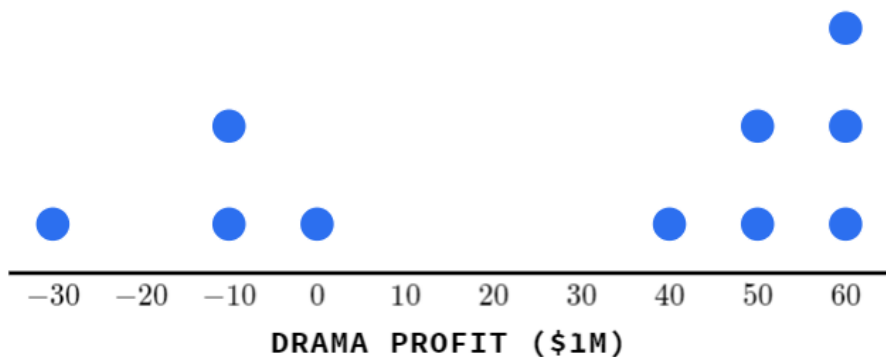
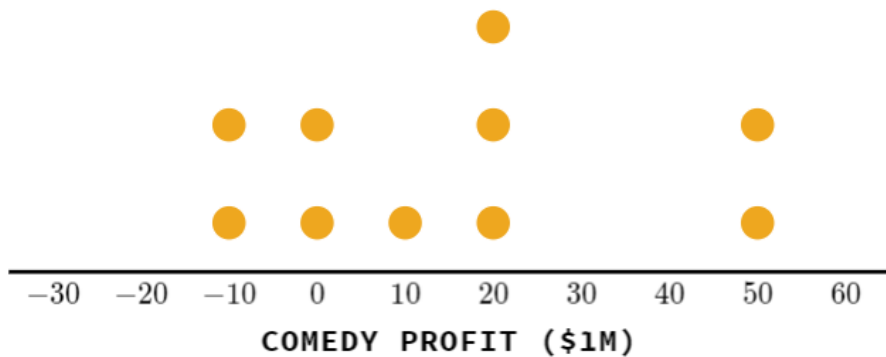
názov	označenie	vzorec	význam	funkcia v Exceli
rozsah	n		Počet štatistických jednotiek	count
minimum	min		Najmenšia hodnota	Min
maximum	max		Najväčšia hodnota	Max
Rozpätie (variačné rozpätie)	R	max-min	Rozdiel maximálnej a minimálnej hodnoty	Max -Min



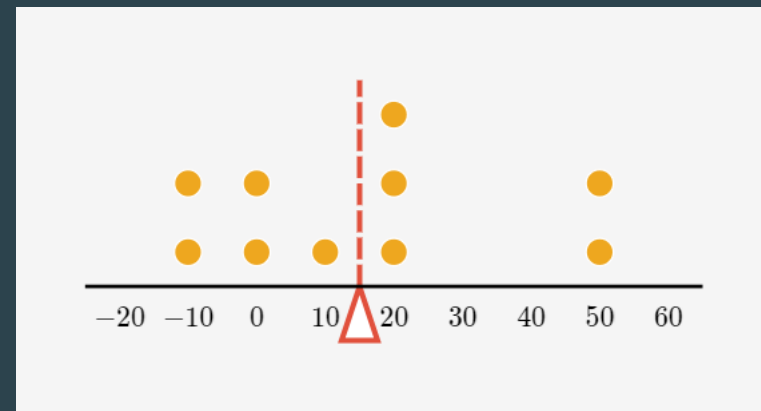
Ktorý zo žánrov má tendenciu mať väčší profit?

Prečo si to myslíte?

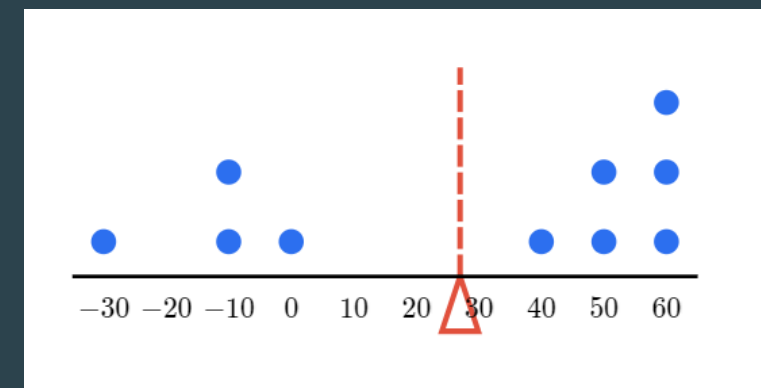
Aritmetický priemer = predstavuje ťažisko, bod rovnováhy pre hodnoty



$$\bar{x}_{comedy} = ?$$

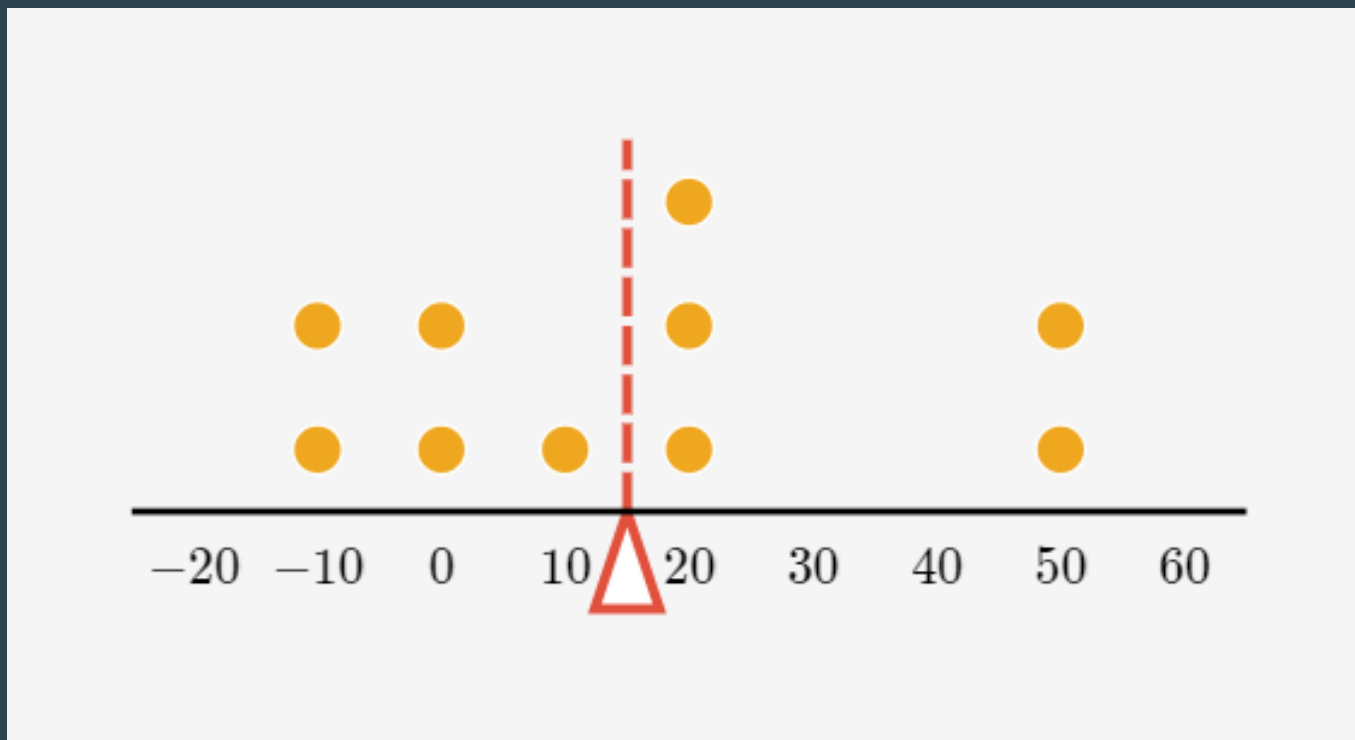


$$\bar{x}_{drama} = ?$$



Čo by sa stalo s aritmetickým priemerom, ak by pribudol 11.ty film s profitom 200?

Čo by sa stalo s aritmetickým priemerom, ak by pribudol 11.ty film so stratou 100M (profitom -100)?

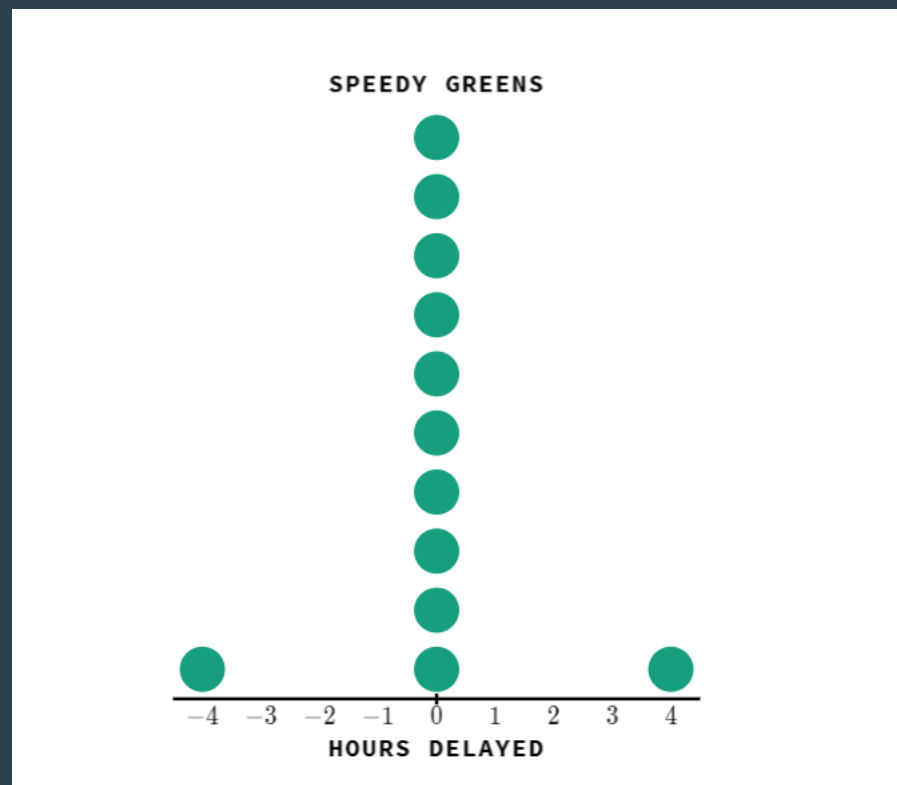
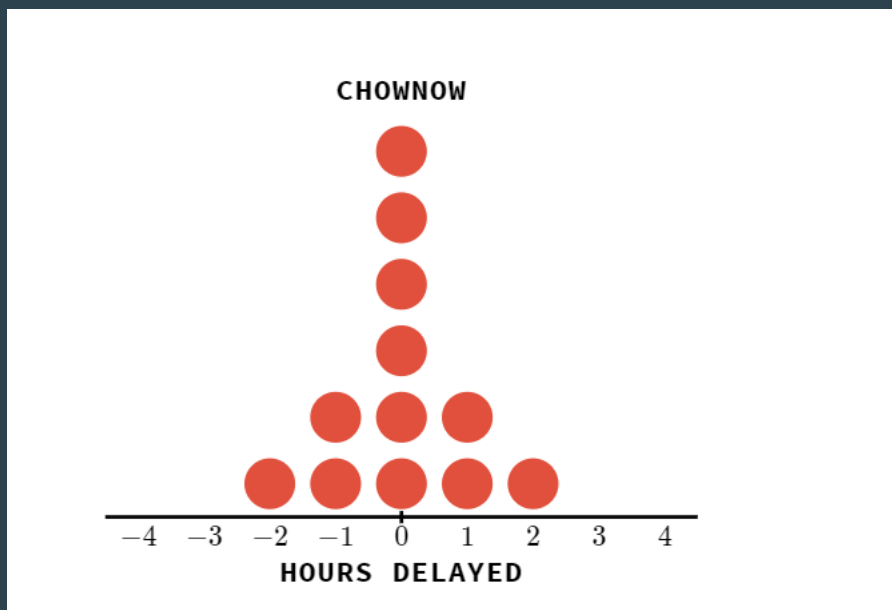
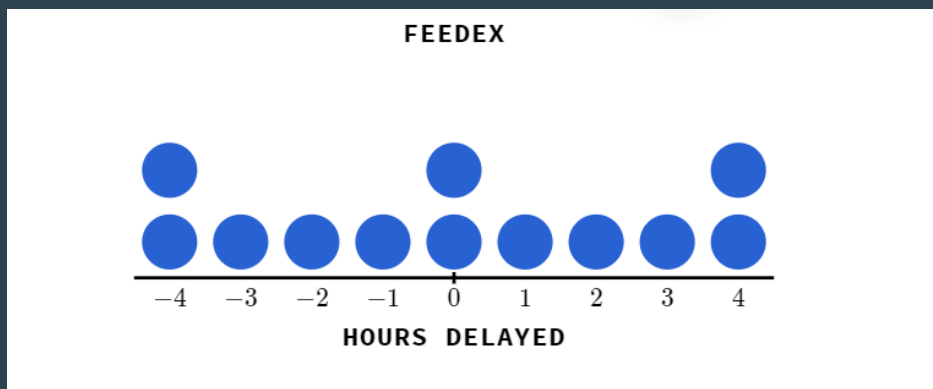


Opisné štatistiky pre numerické premenné

ARITMETICKÝ PRIEMER

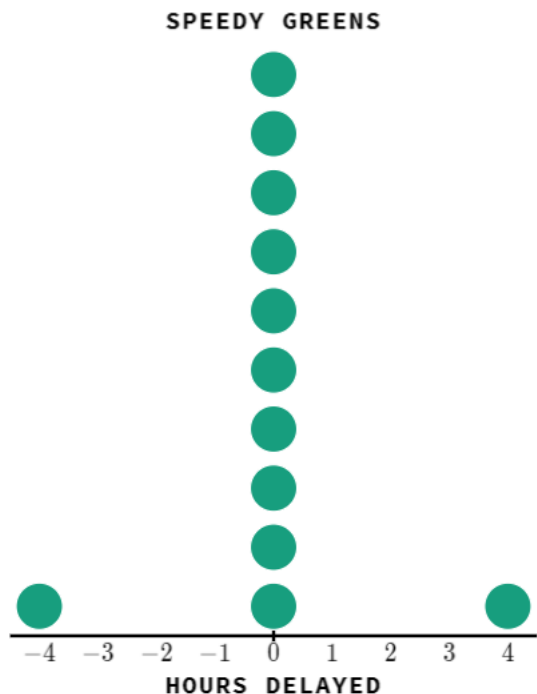
názov	označenie	vzorec	význam	funkcia v Exceli
Aritmetický priemer	$\bar{x}$	$\frac{\sum_{i=1}^n x_i}{n}$	Stredná hodnota	Average

- číslo, ktoré udáva stred údajov;
- za istých okolností môže byť užitočný a dá predstavu o všetkých údajoch v súbore;
- získame ho, ak sčítame všetky hodnoty a vydelíme počtom štatistických jednotiek pre skúmanú numerickú premennú
- číslo veľmi citlivé na extrémne odchýlené hodnoty



Porovnajme rozpätie a aritmetický priemer u všetkých troch donáškových služieb. V čom sa líšia?

Skúsme vymyslieť opisnú štatistiku, ktorá by merala, ako sú hodnoty rozmiestnené v priestore a ako ďaleko sú od priemeru...



Opisné štatistiky pre numerické premenné

ROZPTYL a SMERODAJNÁ ODCHÝLKA

názov	označenie	vzorec	význam	funkcia v Exceli
rozptyl	s^2	$\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$	Priemer druhých mocnín rozdielov	var
smerodajná odchýlka	s	$\sqrt{s^2}$	Odmocnina z rozptylu	Stdev.p

- Smerodajná (štandardná) odchýlka:
 - Číslo, ktoré udáva, ako veľmi sa údaje sústredili okolo aritmetického priemeru
 - Čím je odchýlka menšia, tým sú údaje viac nahusto okolo aritmetického priemeru
- Rozptyl
 - predstavuje iným spôsobom vyjadrenú mieru variability premennej.
 - Jeho hodnotu získame tak, že vypočítame priemer druhých mocnín rozdielov medzi hodnotami premennej a aritmetickým priemerom

Date	Shoppers
Feb. 1	20
Feb. 2	22
Feb. 3	19
Feb. 4	20
Feb. 5	18
Feb. 6	22
Feb. 7	23
Feb. 8	18
Feb. 9	19
Feb. 10	26
Feb. 11	21
Feb. 12	22
Feb. 13	23
Total	273

Aký bol priemerný počet zákazníkov na deň v prvých 13 dňoch februára?

Čo sa stane, ak na Valentína (14.2.) kvôli snehovej búrke nepríde nikto ? Aký bude priemerný počet zákazníkov?

Čo sa stane, ak na Valentína (14.2.) príde 147 zákazníkov? Aký bude priemerný počet zákazníkov?

Skúsme vymyslieť opisnú štatistiku, ktorá by nebola tak veľmi senzitívna na extrémne hodnoty...

1.krok = usporiadajme si hodnoty

Date	Shoppers ↕
Feb. 5	18
Feb. 8	18
Feb. 3	19
Feb. 9	19
Feb. 1	20
Feb. 4	20
Feb. 11	21
Feb. 2	22
Feb. 6	22
Feb. 12	22
Feb. 7	23
Feb. 13	23
Feb. 10	26

2.krok = nájdime hodnotu v strede

Date	Shoppers ↕
Feb. 5	18
Feb. 8	18
Feb. 3	19
Feb. 9	19
Feb. 1	20
Feb. 4	20
Feb. 11	21
Feb. 2	22
Feb. 6	22
Feb. 12	22
Feb. 7	23
Feb. 13	23
Feb. 10	26

Čo ak je hodnôt párny počet?

Date	Shoppers ↕
Feb. 5	18
Feb. 8	18
Feb. 3	19
Feb. 9	19
Feb. 1	20
Feb. 4	20
Feb. 11	21
Feb. 2	22
Feb. 6	22
Feb. 12	22
Feb. 7	23
Feb. 13	23
Feb. 10	26
Feb. 14	147

Date	Shoppers ↕
Feb. 5	18
Feb. 8	18
Feb. 3	19
Feb. 9	19
Feb. 1	20
Feb. 4	20
Feb. 11	21
Feb. 2	22
Feb. 6	22
Feb. 12	22
Feb. 7	23
Feb. 13	23
Feb. 10	26
Feb. 14	147

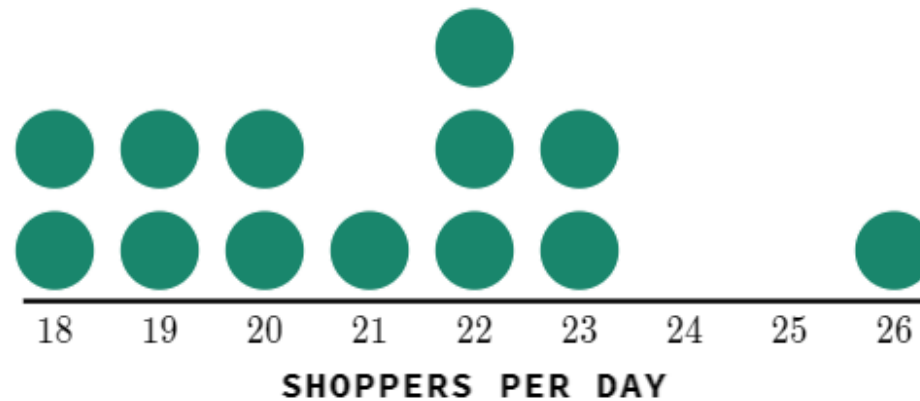
Opisné štatistiky pre numerické premenné

MEDIÁN

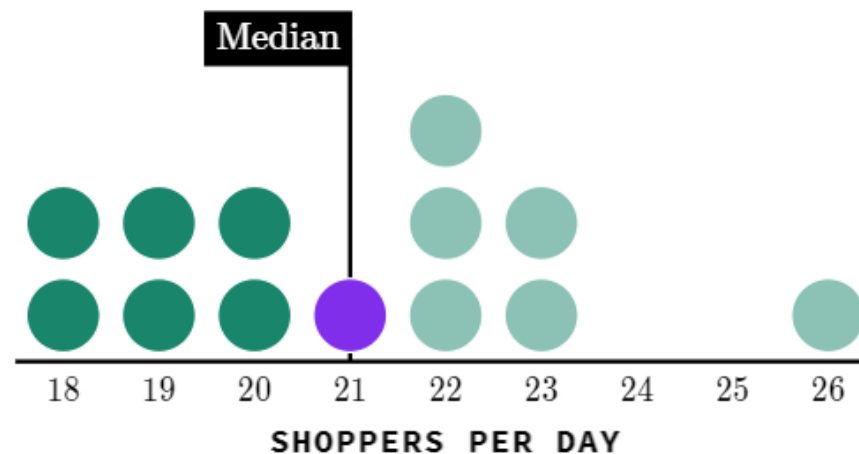
názov	označenie	vzorec	význam	funkcia v Exceli
medián	medián		stredná hodnota	median

- Číslo, ktoré udáva stred údajov, ale iným spôsobom ako aritmetický priemer;
- Získame ho tak, že si všetky údaje usporiadame vzostupne a hodnota, ktorá sa nachádza v strede usporiadaných dát je medián (v prípade párneho počtu dát, urobíme aritmetický priemer susedných stredných hodnôt)

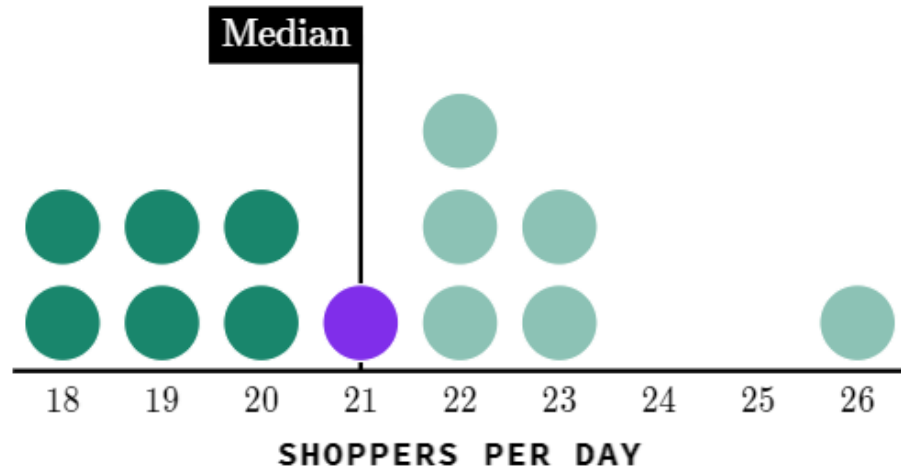
Date	Shoppers 
Feb. 5	18
Feb. 8	18
Feb. 3	19
Feb. 9	19
Feb. 1	20
Feb. 4	20
Feb. 11	21
Feb. 2	22
Feb. 6	22
Feb. 12	22
Feb. 7	23
Feb. 13	23
Feb. 10	26



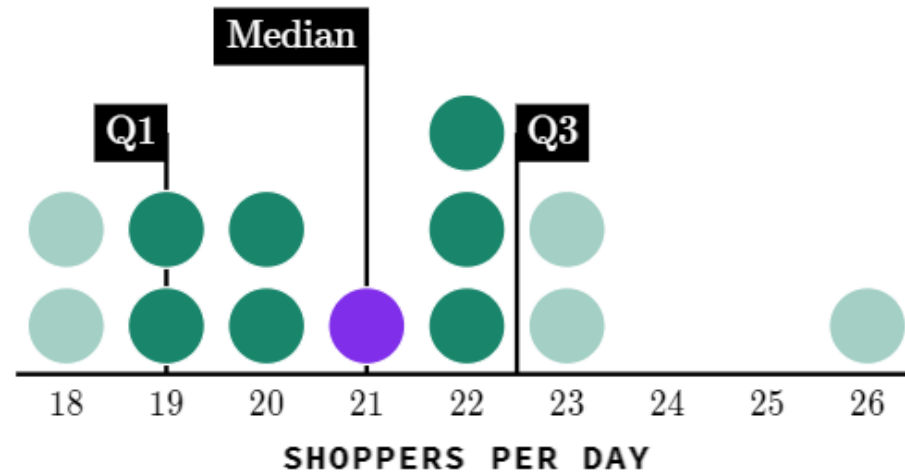
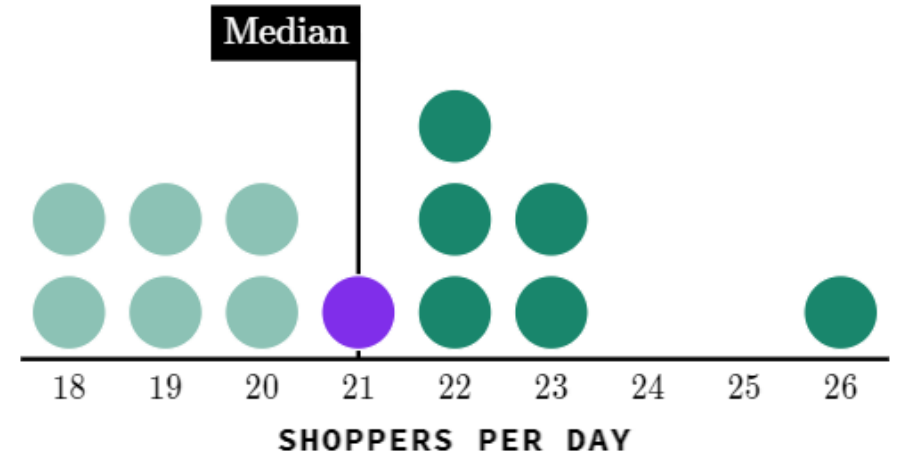
Kde bude medián?



Aký bude medián prvej (ľavej)
polovice hodnôt?



Aký bude medián druhej (pravej)
polovice hodnôt?



Opisné štatistiky pre numerické premenné

KVARTILY A MEDZIKVARTILOVÉ ROZPÄTIE

názov	označenie	vzorec	význam	funkcia v Exceli
kvartily	Q_1, Q_3		Rozdeľujú hodnoty po štvrtinách	quartile
medzikvartilové rozpätie	IQR	$Q_3 - Q_1$	stredných 50% hodnôt	

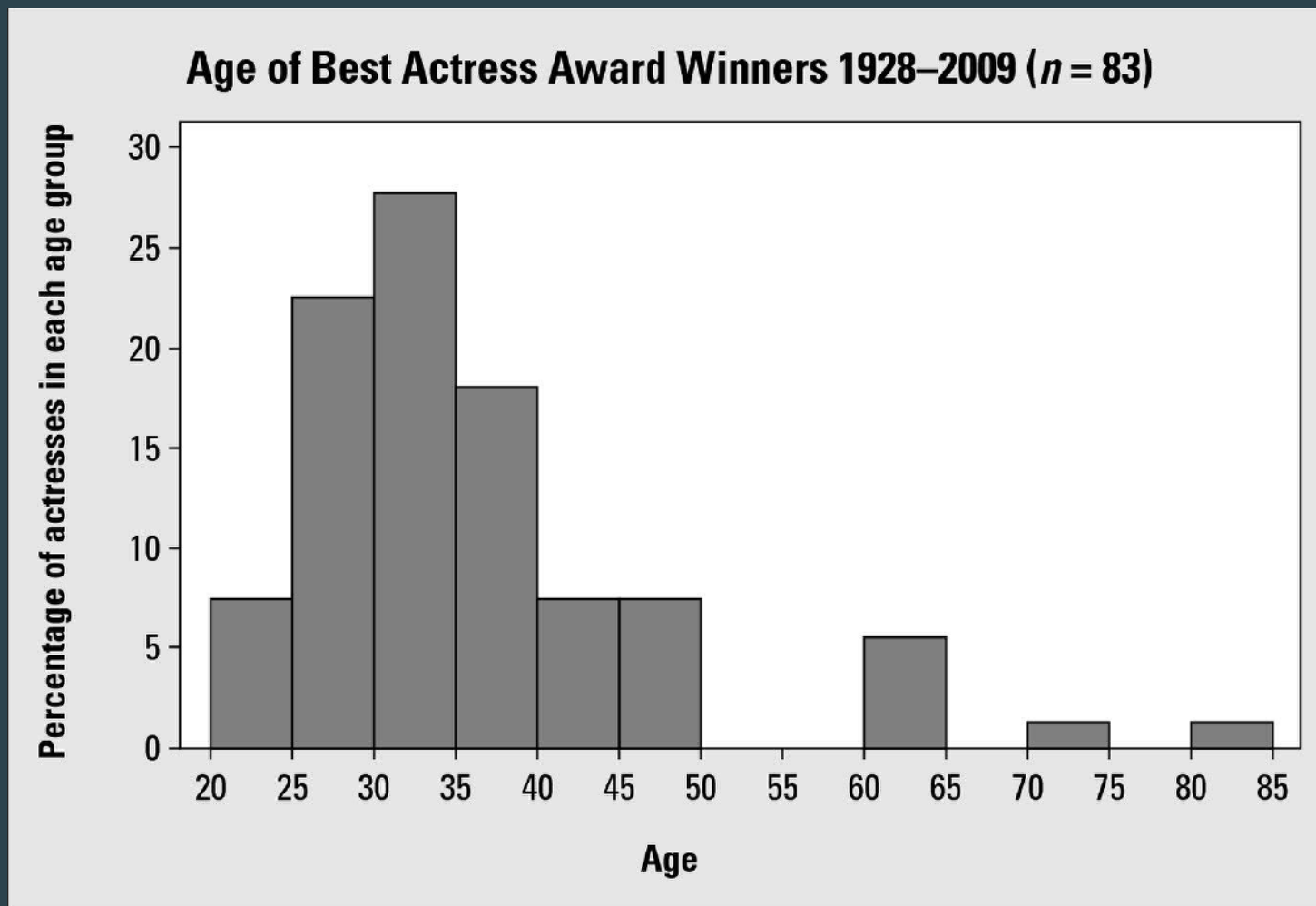
Grafické znázornenie numerických premenných

Histogram:

- na x-ovej osi je rozsah rozdelený na polouzavreté intervaly (bez prázdnych miest);
- Na y-ovej osi absolútna alebo relatívna početnosť

Boxplot (škatuľkový/škatuľový graf)

Histogram - príklad



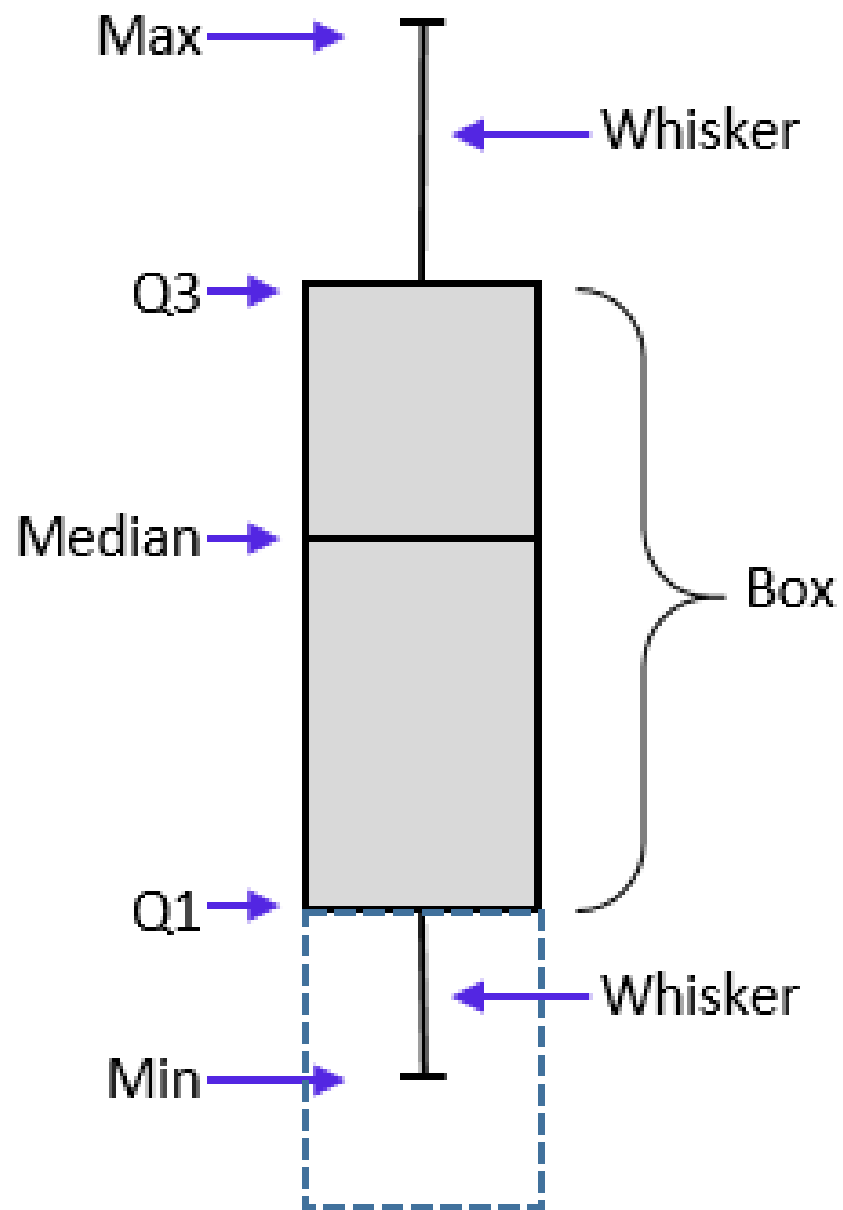
Ukážka spracovania numerickej premennej pomocou
histogramu

Boxplot (škatuľkový/škatuľový graf)

Jednorozmerný graf, ktorý je založený na 5-bodovom zhrnutí = 5 opisných štatistík numerickej premennej:

- Minimum
- Q_1
- Medián
- Q_3
- Maximum

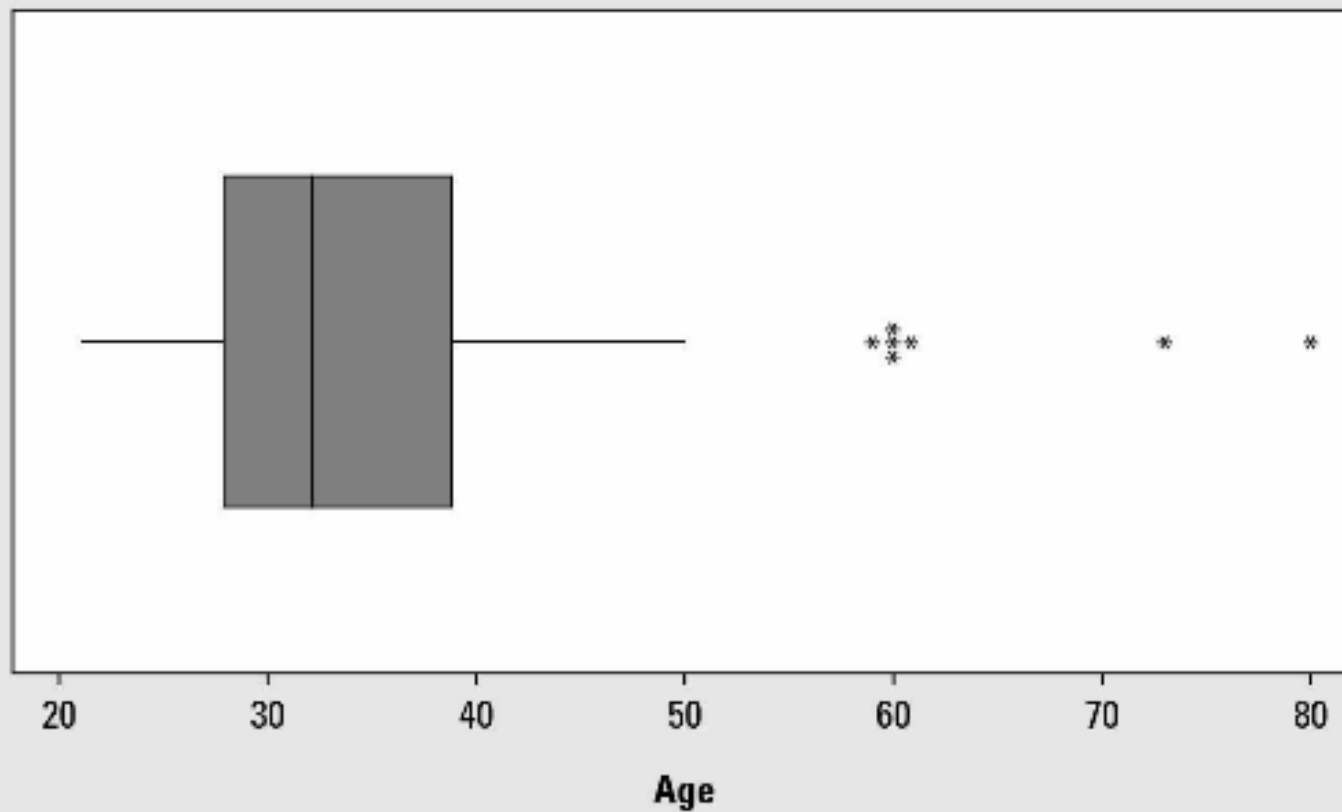
Forma: vertikálna, horizontálna



Boxplot

1. 5 bodové zhrnutie;
2. Číselná os s 5 bodmi;
3. Obdĺžnik medzi Q_1 a Q_3
4. Medián v obdĺžniku
5. Naznačenie „fúzikov“: fúzík je ukončený v min, resp. v max, ak $\min > 1,5 \text{ IQR}$ alebo $\max < 1,5 \text{ IQR}$
6. Prípadné extrémne odchýlené hodnoty (outliers)

Age of Best Actress Award Winners 1928–2009 ($n = 83$)



Boxplot -
příklad

Ukážka spracovania numerickej premennej pomocou
boxplotu