Analýza dat v R
O závislostech, o vizualizaci (a možná i o jednom hadovi)

Mgr. Adéla Vrtková (adela.vrtkova@vsb.cz)

Lékařská fakulta, Ostravská univerzita

Katedra aplikované matematiky
Fakulta elektrotechniky a informatiky, VŠB – Technická univerzita Ostrava

CeBMI educational centre

Co-funded by the Erasmus+ Programme of the European Union

Obsah

- R? R!
  - Výhody, nevýhody a jedno naprosto subjektivní hodnocení
- O závislostech
- O vizualizaci
  - Seznamte se, ggplot2

Materiály naleznete na: https://github.com/AdelaVrtkova/PrednaskaR (Prednaska2022)

Co-funded by the Erasmus+ Programme of the European Union

R? R!

Výhody

- primárně „statistický“ jazyk
- open-source
- rozšiřující knihovny (packages)
- pro Windows, Linux, OS X
- velká komunita, série knih Use R!
- statistické metody, ML algoritmy
- knihovny pro vizualizaci dat
- knihovna pro interaktivní aplikace
- a další…

Nevýhody

- primárně „statistický“ jazyk
- ze začátku obtížnější osvojení
- riziko nekvalitních knihoven
- „překrývající se“ knihovny
- memory management
- rychlost
- zabezpečení (?)
- a další…

Co-funded by the Erasmus+ Programme of the European Union

O závislostech

Datový soubor obsahuje následující údaje o 600 pacientech:

- ID – jednoznačný identifikátor pacienta (celá čísla od 1 do 600)
- Pohlavi – pohlaví pacienta (muž, žena)
- BMI – Body Mass Index (kg/m^2)
- Mnozstvi_tuku – množství tuku v těle (v %)
- Obvod pasu – obvod pasu pacienta (v cm)
- Kvalita_spanku – hodnocení kvality spánku (dobrá, špatná)

Poznámka: Data jsou inspirována reálnými studiemi závislosti tělesné konstituce a kvality spánku, nicméně byla na základě této inspirace kompletně uměle vygenerována.

Co-funded by the Erasmus+ Programme of the European Union

O závislostech

ID	Pohlavi	BMI	Mnozstvi_tuku	Obvod_pasu	Kvalita_spanku
1	muž	29,81	22,66	90,1	špatná
2	žena	22,5	26,59	79,8	dobrá
3	muž	24,5	13,75	76,4	dobrá
4	žena	24,04	30,79	87,4	špatná
5	muž	22,56	16,7	83,7	dobrá
6	žena	19,98	26,18	83	dobrá
7	žena	23,61	35,59	84	dobrá
8	muž	20,85	2,77	72	dobrá
9	muž	26,95	21,29	97,5	špatná
...	...	...	...	...	....

O závislostech

ID	Pohlavi	BMI	Mnozstvi_tuku	Obvod_pasu	Kvalita_spanku
1	muž	29,81	22,66	90,1	špatná
2	žena	22,5	26,59	79,8	dobrá
3	muž	24,5	13,75	76,4	dobrá
4	žena	24,04	30,79	87,4	špatná
5	muž	22,56	16,7	83,7	dobrá
6	žena	19,98	26,18	83	dobrá
7	žena	23,61	35,59	84	dobrá
8	muž	20,85	2,77	72	dobrá
9	muž	26,95	21,29	97,5	špatná
...	...	...	...	...	...

1. Analyzujte závislost množství tělesného tuku a obvodu pasu.

O závislostech

ID	Pohlavi	BMI	Mnozstvi_tuku	Obvod_pasu	Kvalita_spanku
1	muž	29,81	22,66	90,1	špatná
2	žena	22,5	26,59	79,8	dobrá
3	muž	24,5	13,75	76,4	dobrá
4	žena	24,04	30,79	87,4	špatná
5	muž	22,56	16,7	83,7	dobrá
6	žena	19,98	26,18	83	dobrá
7	žena	23,61	35,59	84	dobrá
8	muž	20,85	2,77	72	dobrá
9	muž	26,95	21,29	97,5	špatná
...	...	...	...	...	...

2. Analyzujte závislost množství tělesného tuku a kvality spánku pacienta.

O závislostech

ID	Pohlavi	BMI	Mnozstvi_tuku	Obvod_pasu	Kvalita_spanku
1	muž	29,81	22,66	90,1	špatná
2	žena	22,5	26,59	79,8	dobrá
3	muž	24,5	13,75	76,4	dobrá
4	žena	24,04	30,79	87,4	špatná
5	muž	22,56	16,7	83,7	dobrá
6	žena	19,98	26,18	83	dobrá
7	žena	23,61	35,59	84	dobrá
8	muž	20,85	2,77	72	dobrá
9	muž	26,95	21,29	97,5	špatná
...	...	...	...	...	...

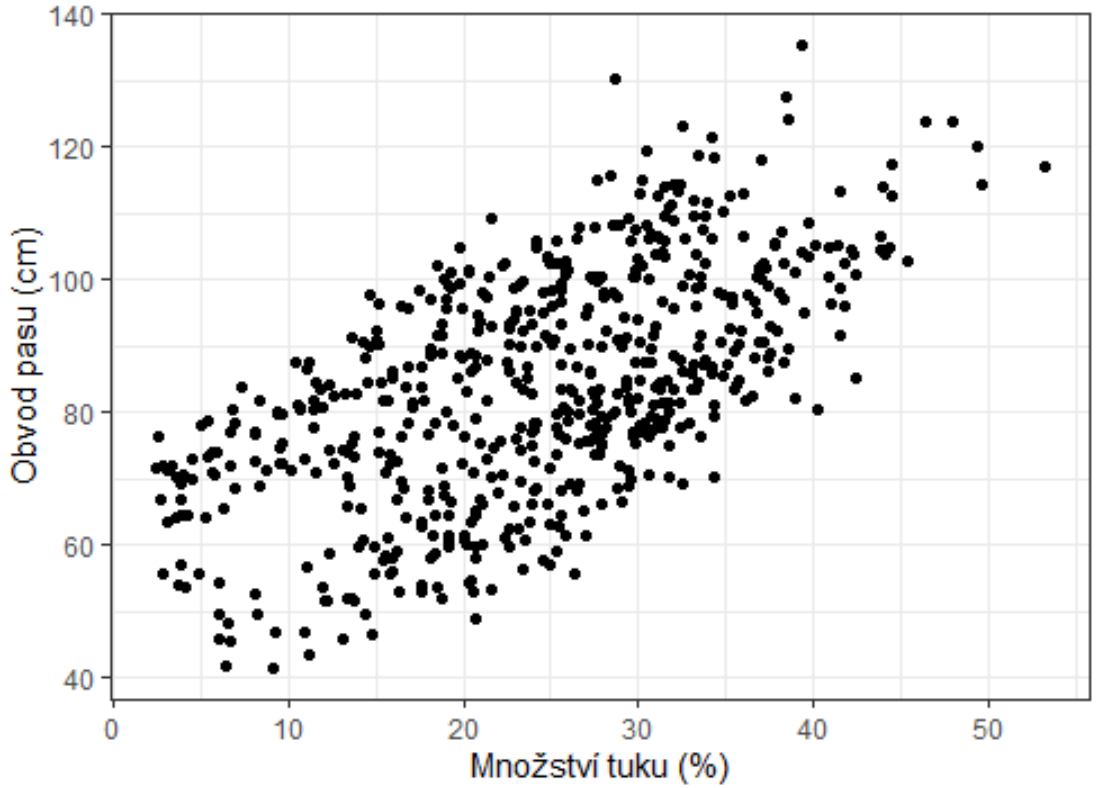
3. Analyzujte závislost pohlaví pacienta a kvality jeho spánku.

O závislostech

ID	Pohlavi	BMI	Mnozstvi_tuku	Obvod_pasu	Kvalita_spanku
1	muž	29,81	22,66	90,1	špatná
2	žena	22,5	26,59	79,8	dobrá
3	muž	24,5	13,75	76,4	dobrá
4	žena	24,04	30,79	87,4	špatná
5	muž	22,56	16,7	83,7	dobrá
6	žena	19,98	26,18	83	dobrá
7	žena	23,61	35,59	84	dobrá
8	muž	20,85	2,77	72	dobrá
9	muž	26,95	21,29	97,5	špatná
...	...	...	...	...	...

1. Analyzujte závislost množství tělesného tuku a obvodu pasu.

O závislostech



O závislostech

- **Závislost dvou kvantitativních proměnných**
  - ✓ bodový graf
  - ✓ vhodný korelační koeficient (příp. jeho test významnosti)

Co-funded by the Erasmus+ Programme of the European Union

O závislostech

Pearsonův korelační koeficient

- ✓ míra lineární závislosti
- ✓ nabývá hodnot z intervalu $\langle -1, 1 \rangle$
- ✓ je-li nulový, proměnné nejsou lineárně závislé
- ✓ předpokladem souvisejícího statistického testu je normalita obou proměnných
- ✓ citlivý na odlehlá pozorování

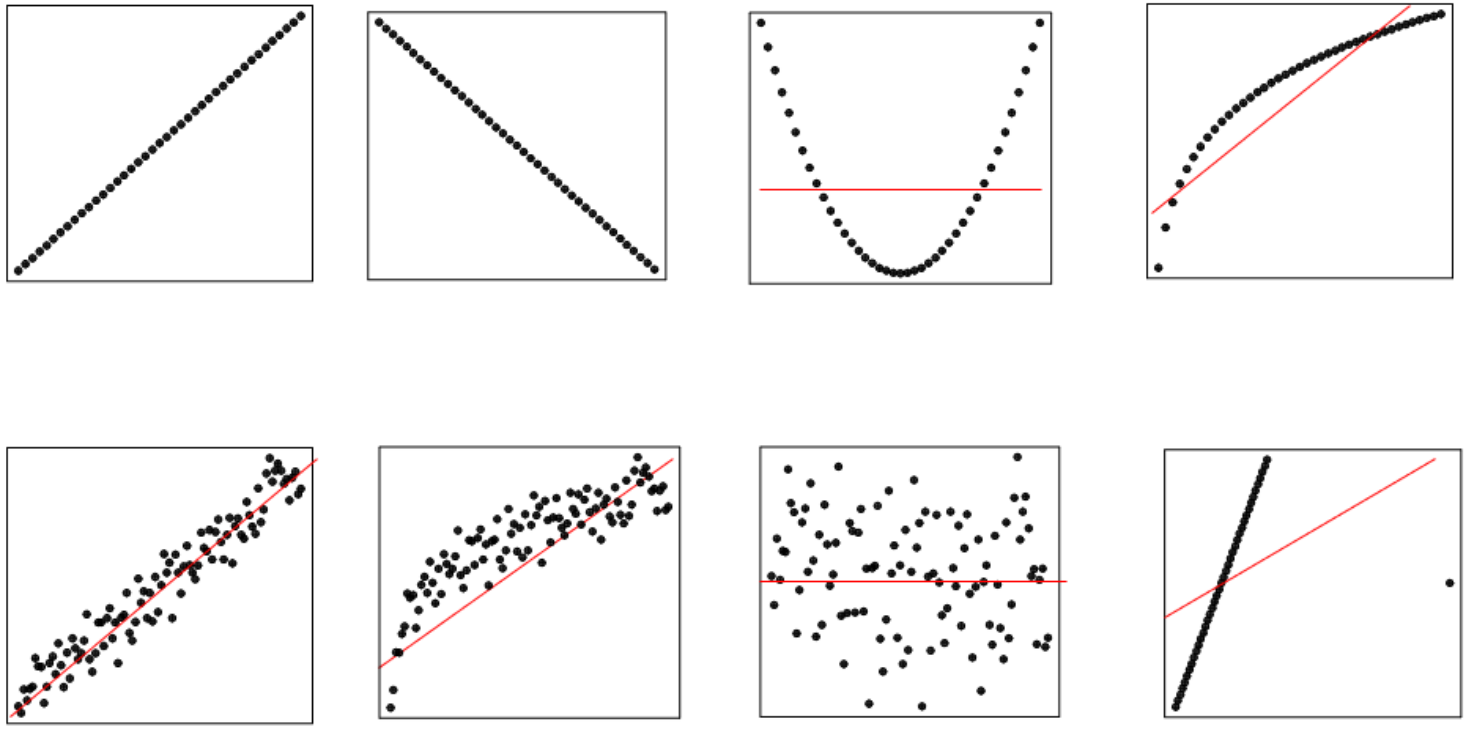
Spearmanův korelační koeficient

- ✓ míra monotónní závislosti
- ✓ nabývá hodnot z intervalu $\langle -1, 1 \rangle$
- ✓ je-li nulový, mezi proměnnými není monotónní závislost
- ✓ spadá pod neparametrické metody
- ✓ robustní vůči odlehlým pozorováním

Korelace neznamená kauzalitu!

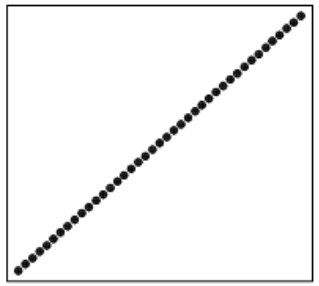
Nekorelovanost obecně neznamená „úplnou“ nezávislost!

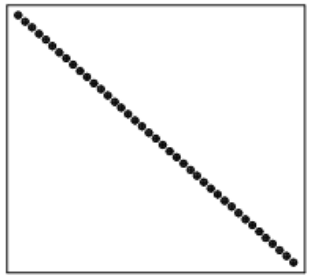
O závislostech

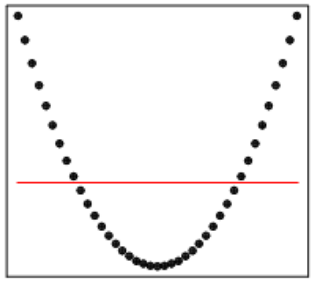


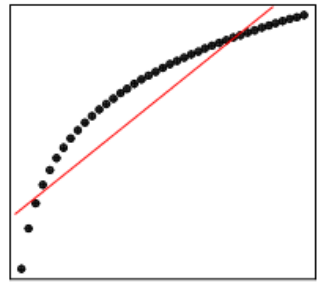
Zdroj: Litschmannová Martina, Statistické myšlení (přednáška), Katedra aplikované matematiky, Fakulta elektrotechniky a informatiky, VŠB – Technická univerzita Ostrava

O závislostech

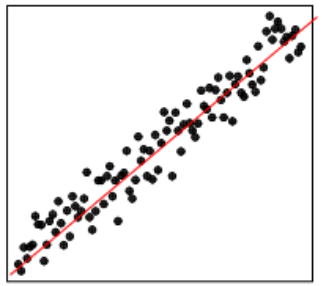


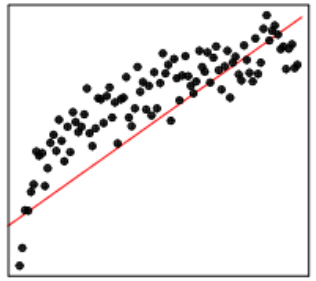


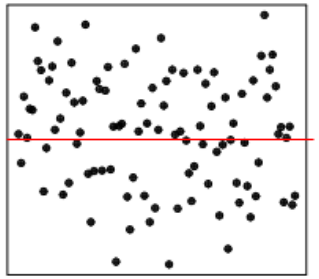


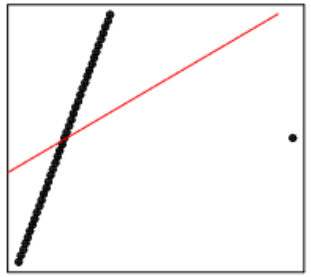


$\rho(X,Y) = 1,000$
$\rho_S(X,Y) = 1,000$



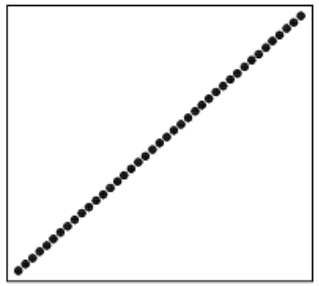




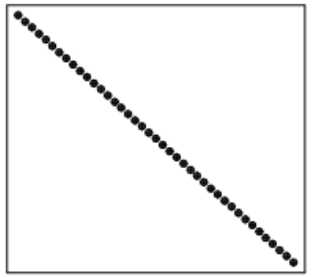


Zdroj: Litschmannová Martina, Statistické myšlení (přednáška), Katedra aplikované matematiky, Fakulta elektrotechniky a informatiky, VŠB – Technická univerzita Ostrava

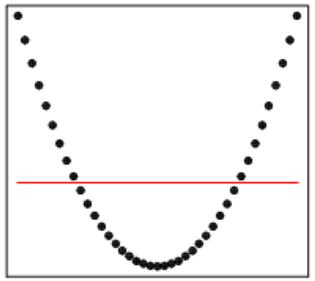
O závislostech

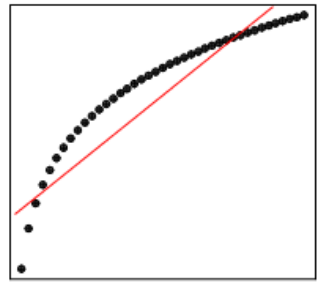


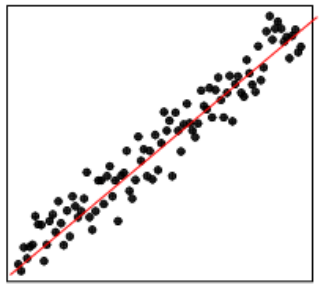
$\rho(X,Y) = 1,000$
$\rho_S(X,Y) = 1,000$

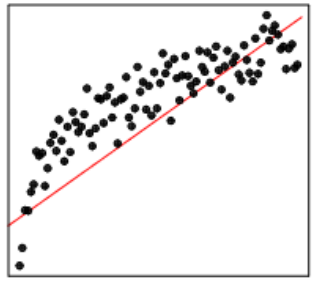


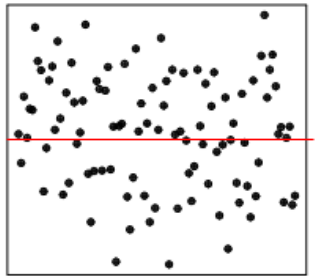
$\rho(X,Y) = -1,000$
$\rho_S(X,Y) = -1,000$

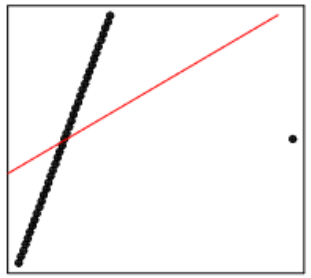






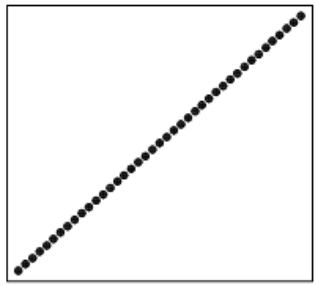




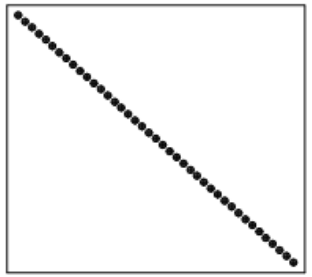


Zdroj: Litschmannová Martina, Statistické myšlení (přednáška), Katedra aplikované matematiky, Fakulta elektrotechniky a informatiky, VŠB – Technická univerzita Ostrava

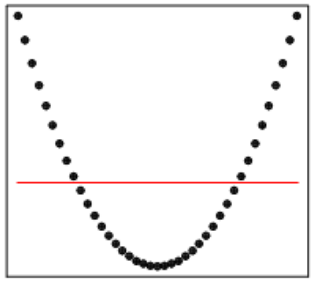
O závislostech



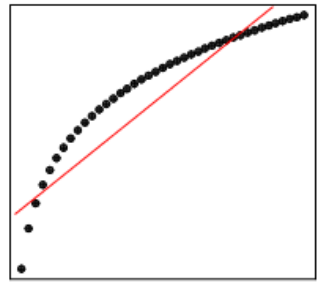
$\rho(X,Y) = 1,000$
$\rho_S(X,Y) = 1,000$

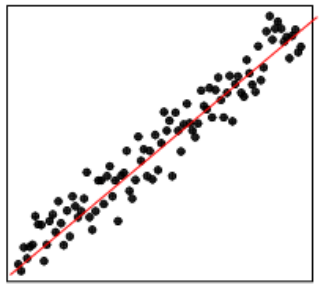


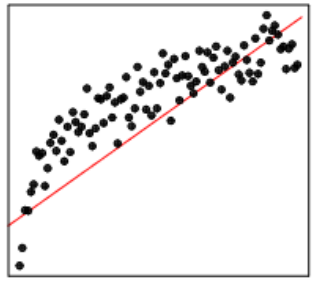
$\rho(X,Y) = -1,000$
$\rho_S(X,Y) = -1,000$

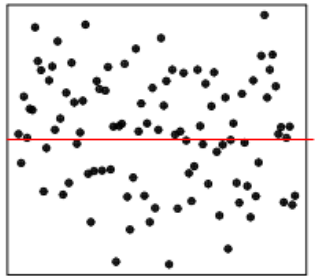


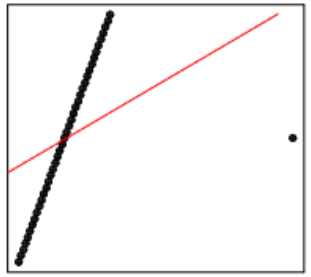
$\rho(X,Y) = 0,000$
$\rho_S(X,Y) = 0,000$





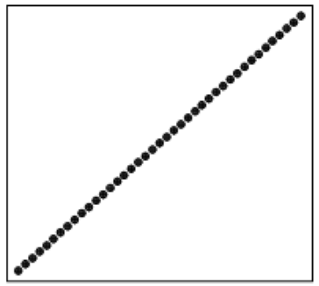




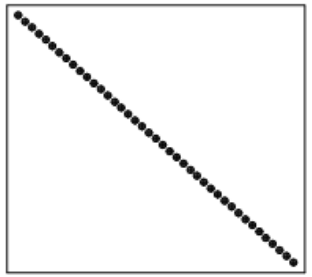


Zdroj: Litschmannová Martina, Statistické myšlení (přednáška), Katedra aplikované matematiky, Fakulta elektrotechniky a informatiky, VŠB – Technická univerzita Ostrava

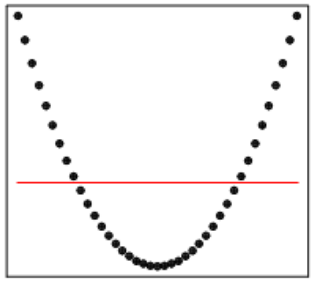
O závislostech



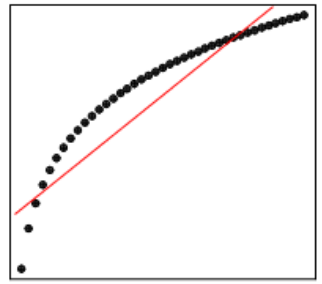
$\rho(X,Y) = 1,000$
$\rho_S(X,Y) = 1,000$



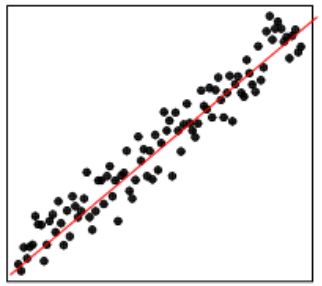
$\rho(X,Y) = -1,000$
$\rho_S(X,Y) = -1,000$

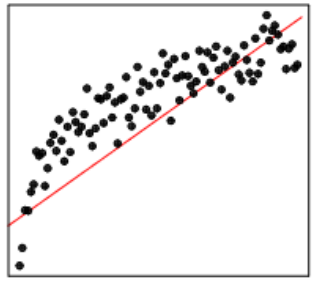


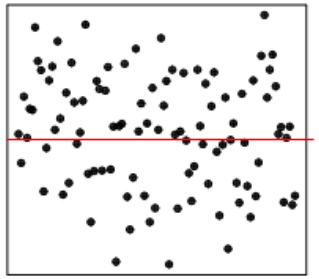
$\rho(X,Y) = 0,000$
$\rho_S(X,Y) = 0,000$

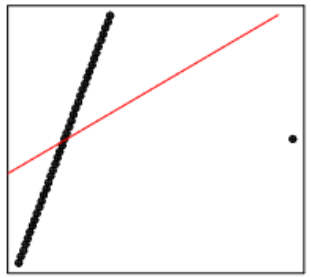


$\rho(X,Y) = 0,934$
$\rho_S(X,Y) = 1,000$



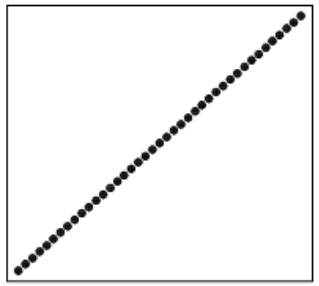




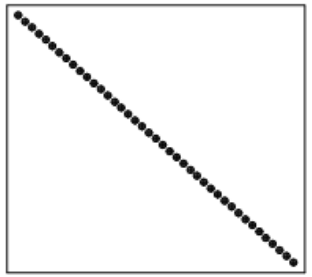


Zdroj: Litschmannová Martina, Statistické myšlení (přednáška), Katedra aplikované matematiky, Fakulta elektrotechniky a informatiky, VŠB – Technická univerzita Ostrava

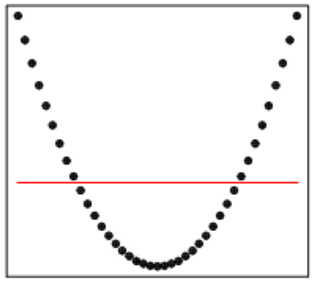
O závislostech



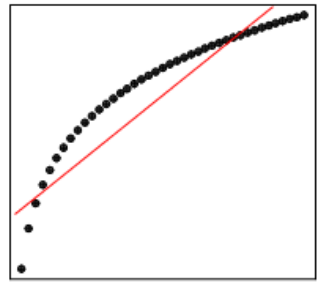
$\rho(X,Y) = 1,000$
$\rho_S(X,Y) = 1,000$



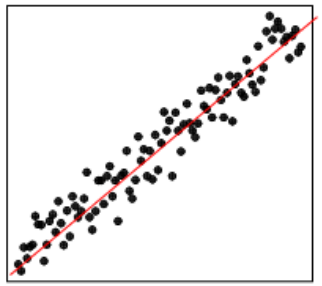
$\rho(X,Y) = -1,000$
$\rho_S(X,Y) = -1,000$



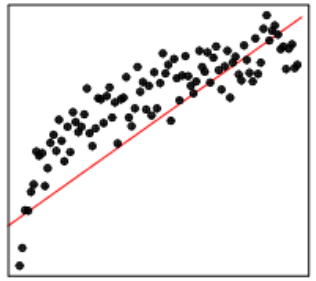
$\rho(X,Y) = 0,000$
$\rho_S(X,Y) = 0,000$

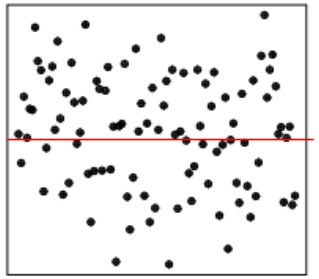


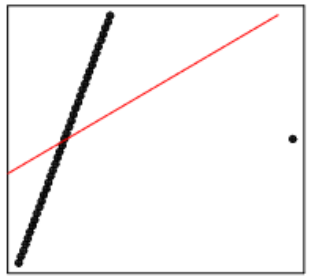
$\rho(X,Y) = 0,934$
$\rho_S(X,Y) = 1,000$



$\rho(X,Y) = 0,967$
$\rho_S(X,Y) = 0,981$

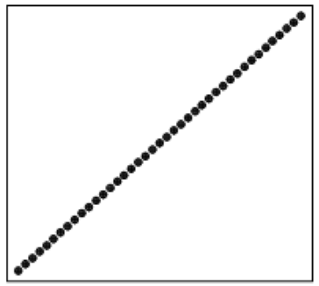




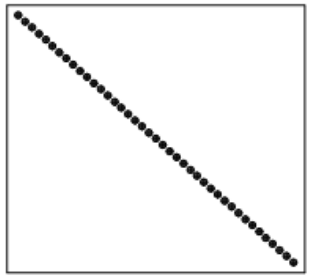


Zdroj: Litschmannová Martina, Statistické myšlení (přednáška), Katedra aplikované matematiky, Fakulta elektrotechniky a informatiky, VŠB – Technická univerzita Ostrava

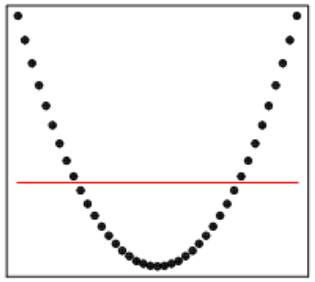
O závislostech



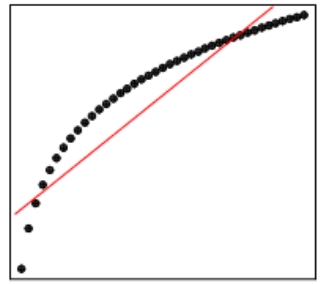
$\rho(X,Y) = 1,000$
$\rho_S(X,Y) = 1,000$



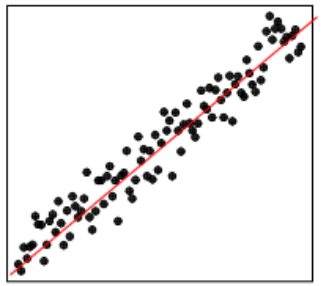
$\rho(X,Y) = -1,000$
$\rho_S(X,Y) = -1,000$



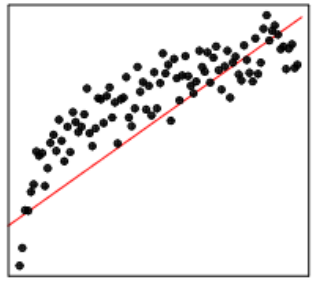
$\rho(X,Y) = 0,000$
$\rho_S(X,Y) = 0,000$



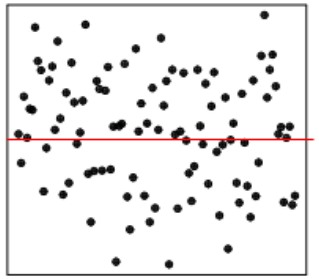
$\rho(X,Y) = 0,934$
$\rho_S(X,Y) = 1,000$

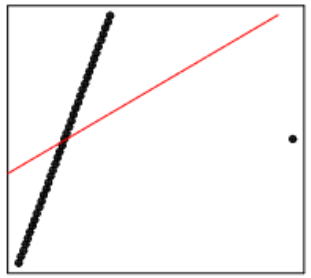


$\rho(X,Y) = 0,967$
$\rho_S(X,Y) = 0,981$



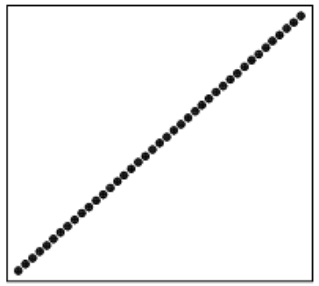
$\rho(X,Y) = 0,857$
$\rho_S(X,Y) = 0,893$



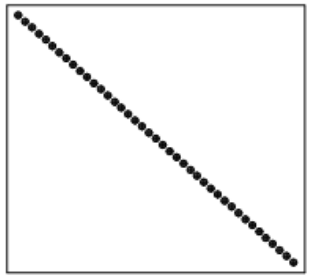


Zdroj: Litschmannová Martina, Statistické myšlení (přednáška), Katedra aplikované matematiky, Fakulta elektrotechniky a informatiky, VŠB – Technická univerzita Ostrava

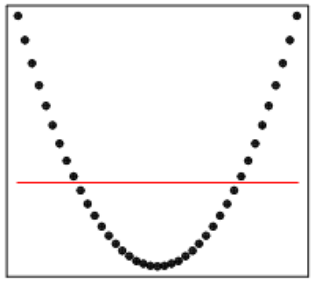
O závislostech



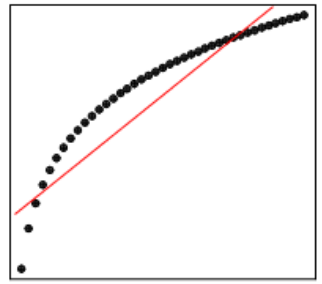
$\rho(X,Y) = 1,000$
$\rho_S(X,Y) = 1,000$



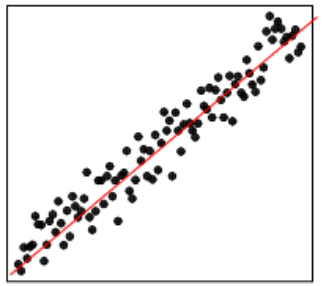
$\rho(X,Y) = -1,000$
$\rho_S(X,Y) = -1,000$



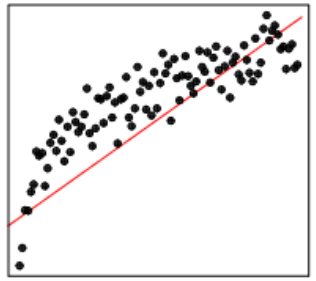
$\rho(X,Y) = 0,000$
$\rho_S(X,Y) = 0,000$



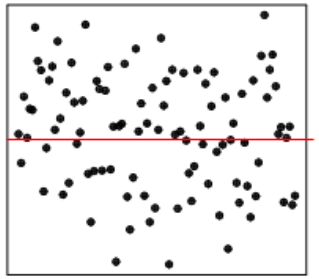
$\rho(X,Y) = 0,934$
$\rho_S(X,Y) = 1,000$



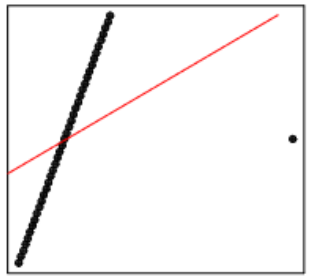
$\rho(X,Y) = 0,967$
$\rho_S(X,Y) = 0,981$



$\rho(X,Y) = 0,857$
$\rho_S(X,Y) = 0,893$

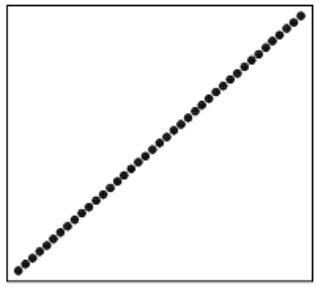


$\rho(X,Y) = -0,143$
$\rho_S(X,Y) = -0,178$

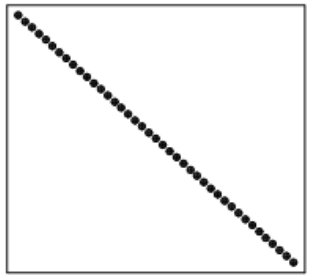


Zdroj: Litschmannová Martina, Statistické myšlení (přednáška), Katedra aplikované matematiky, Fakulta elektrotechniky a informatiky, VŠB – Technická univerzita Ostrava

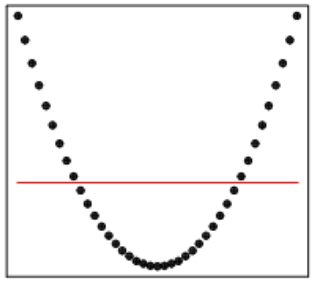
O závislostech



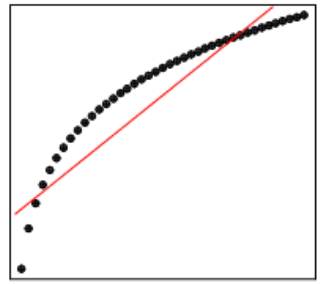
$\rho(X,Y) = 1,000$
$\rho_S(X,Y) = 1,000$



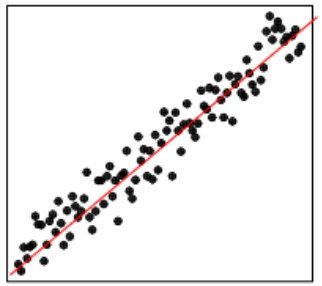
$\rho(X,Y) = -1,000$
$\rho_S(X,Y) = -1,000$



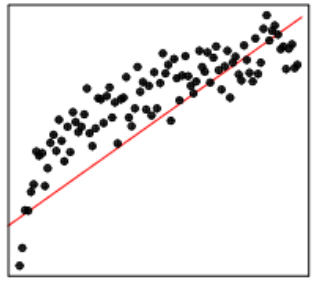
$\rho(X,Y) = 0,000$
$\rho_S(X,Y) = 0,000$



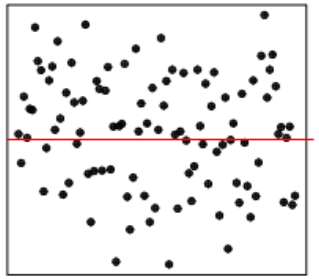
$\rho(X,Y) = 0,934$
$\rho_S(X,Y) = 1,000$



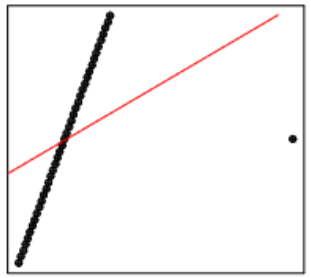
$\rho(X,Y) = 0,967$
$\rho_S(X,Y) = 0,981$



$\rho(X,Y) = 0,857$
$\rho_S(X,Y) = 0,893$



$\rho(X,Y) = -0,143$
$\rho_S(X,Y) = -0,178$



$\rho(X,Y) = 0,608$
$\rho_S(X,Y) = 0,911$

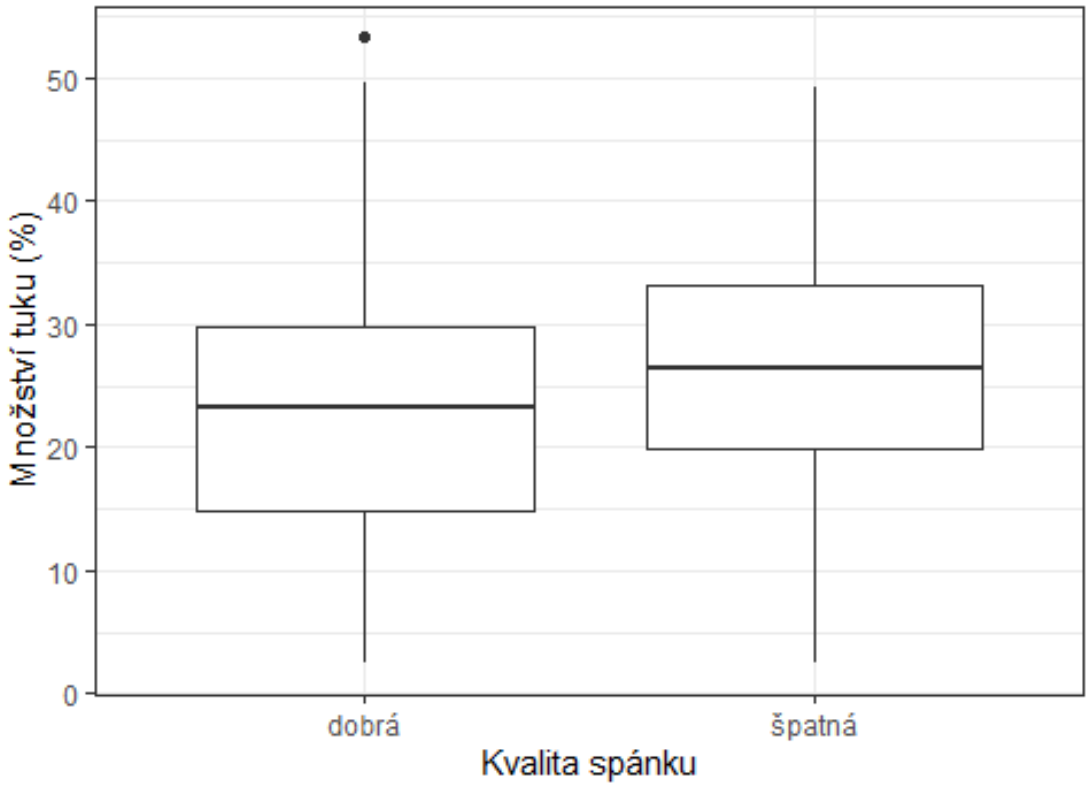
Zdroj: Litschmannová Martina, Statistické myšlení (přednáška), Katedra aplikované matematiky, Fakulta elektrotechniky a informatiky, VŠB – Technická univerzita Ostrava

O závislostech

ID	Pohlavi	BMI	Mnozstvi_tuku	Obvod_pasu	Kvalita_spanku
1	muž	29,81	22,66	90,1	špatná
2	žena	22,5	26,59	79,8	dobrá
3	muž	24,5	13,75	76,4	dobrá
4	žena	24,04	30,79	87,4	špatná
5	muž	22,56	16,7	83,7	dobrá
6	žena	19,98	26,18	83	dobrá
7	žena	23,61	35,59	84	dobrá
8	muž	20,85	2,77	72	dobrá
9	muž	26,95	21,29	97,5	špatná
...	...	...	...	...	...

2. Analyzujte závislost množství tělesného tuku a kvality spánku pacienta.

O závislostech



O závislostech

- **Závislost dvou kvantitativních proměnných**
  - ✓ bodový graf
  - ✓ vhodný korelační koeficient (příp. jeho test významnosti)

- **Závislost kvantitativní a kvalitativní proměnné**
  - ✓ vícenásobný krabicový graf (příp. sada histogramů)
  - ✓ srovnání číselných charakteristik kvantitativní proměnné vypočtených pro každou variantu kvalitativní proměnné (příp. vhodné statistické testy)

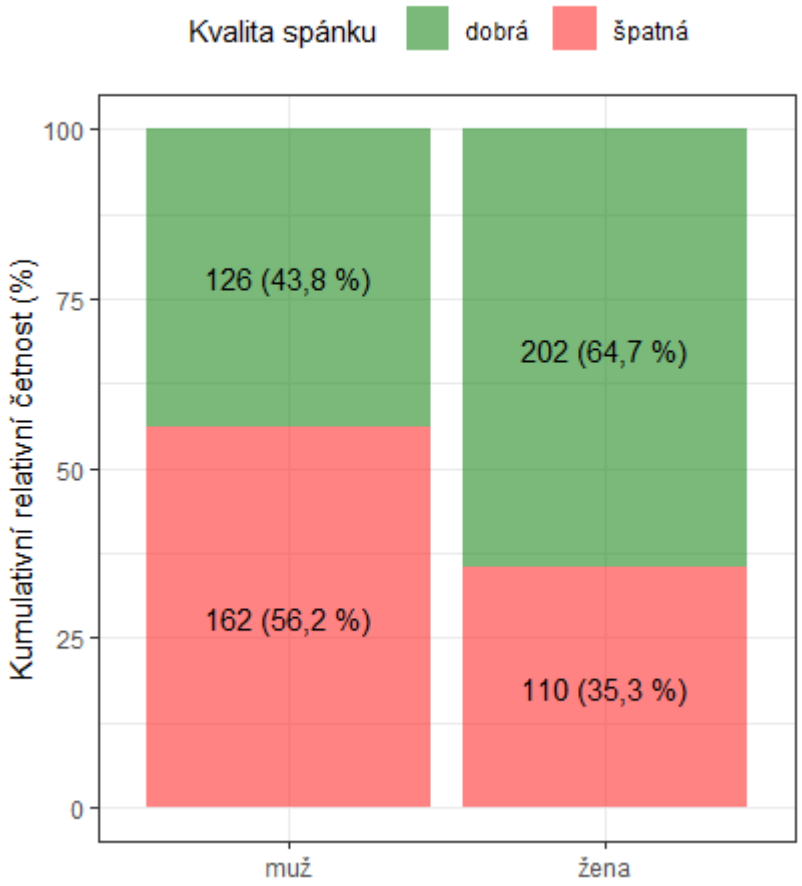
Co-funded by the Erasmus+ Programme of the European Union

O závislostech

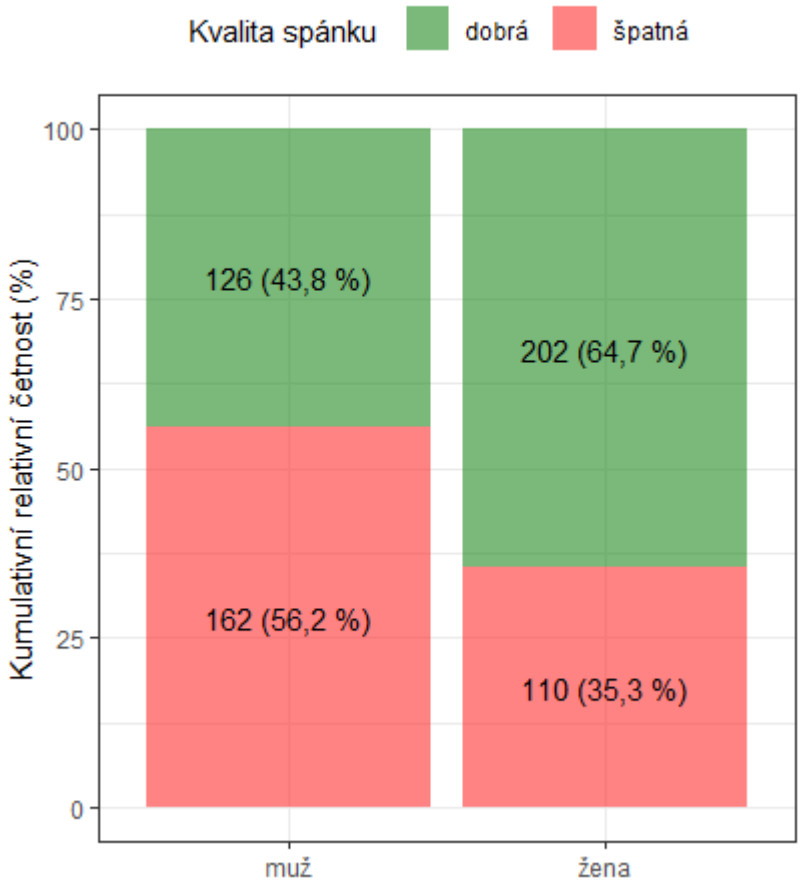
ID	Pohlavi	BMI	Mnozstvi_tuku	Obvod_pasu	Kvalita_spanku
1	muž	29,81	22,66	90,1	špatná
2	žena	22,5	26,59	79,8	dobrá
3	muž	24,5	13,75	76,4	dobrá
4	žena	24,04	30,79	87,4	špatná
5	muž	22,56	16,7	83,7	dobrá
6	žena	19,98	26,18	83	dobrá
7	žena	23,61	35,59	84	dobrá
8	muž	20,85	2,77	72	dobrá
9	muž	26,95	21,29	97,5	špatná
...	...	...	...	...	...

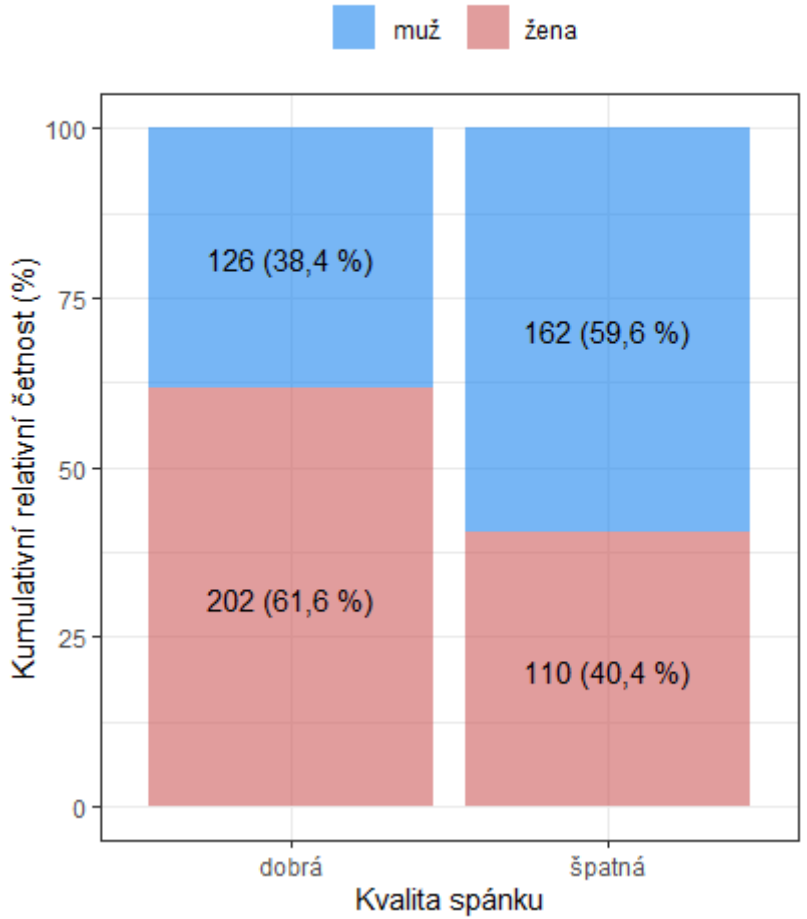
3. Analyzujte závislost pohlaví pacienta a kvality jeho spánku.

O závislostech



O závislostech





O závislostech

- **Závislost dvou kvantitativních proměnných**
  - ✓ bodový graf
  - ✓ vhodný korelační koeficient (příp. jeho test významnosti)

- **Závislost kvantitativní a kvalitativní proměnné**
  - ✓ vícenásobný krabicový graf, sada histogramů
  - ✓ srovnání číselných charakteristik kvantitativní proměnné vypočtených pro každou variantu kvalitativní proměnné (příp. vhodné statistické testy)

- **Závislost dvou kategoriálních proměnných**
  - ✓ kontingenční tabulka s vhodnými relativními četnostmi (řádkové/sloupcové)
  - ✓ 100% skládaný sloupcový graf, mozaikový graf (příp. vhodný statistický test)

Co-funded by the Erasmus+ Programme of the European Union

O vizualizaci

Seznamte se, ggplot2

Zdroj: https://rpubs.com/collnell/ggplot2 [cit. 11. 4. 2021]

O vizualizaci

Seznamte se, ggplot2

- Nejprve definujeme **"estetiku" (aesthetics - aes)**:
  - Důležitá část, kde specifikujeme proměnnou na ose x a/nebo na ose y.
  - Lze ale i určit parametr, který ovlivní velikost (size) nebo barvu (color) vykreslených objektů (např. bodů).
  - Dalšími parametry v estetice jsou - fill, linetype, label, shape a další...

- Následuje určení **"geometrie" (geometries - geom_???)**:
  - Tato část definuje, jak se mají data znázornit.
  - Např. jako body (geom_point), čáry (geom_line), krabicové grafy (geom_boxplot), sloupcové grafy (geom_bar),...
  - Je třeba uvážit typ dat a na základě toho, jakou chceme informaci předat, zvolit geometrii.
  - Různé "geom" lze i kombinovat, má-li to smysl.

Co-funded by the Erasmus+ Programme of the European Union

O vizualizaci

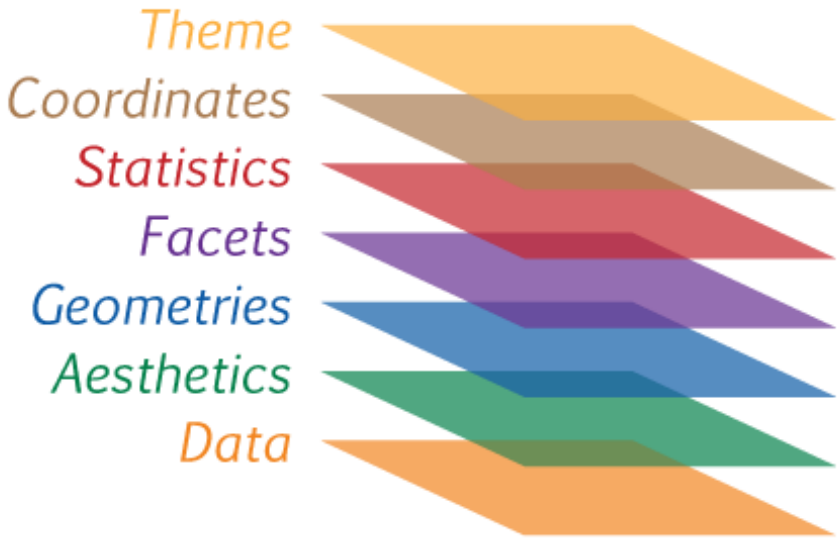
Seznamte se, ggplot2

- V dalších vrstvách lze nastavovat:
  - rozdělení na tzv. **"facets"** (mřížka grafů),
  - zakomponování "statistiky" - vrstva **"statistics"** - např. přidání trendu, vykreslení průměru jako bodu, přidání korelačního koeficientu apod.,
  - vrstvu **"coordinates"** měřítka os, změnit je na logaritmické, apod.,
  - změnit vzhled pomocí definovaných grafických témat - vrstva **"theme"** - případně si nastavit své vlastní.

Knihovnu ggplot2 lze používat s řadou rozšiřujících knihoven: např. ggpubr, GGally, ggTimeSeries, ggmosaic, ggpol, ggnewscale, ggExtra, plotROC, survminer a mnoho dalších! (https://exts.ggplot2.tidyverse.org/gallery/)

O vizualizaci

Seznamte se, ggplot2




```
ggplot(data,
       aes(x = Kvalita_spanku,
           y = Obvod_pasu))+
  geom_boxplot()+
  stat_summary(geom = "point",
               fun = "mean",
               color = "blue",
               shape = 3)+
  scale_y_continuous(breaks = seq(40, 140, 20))+
  labs(x = "Kvalita spánku",
       y = "Obvod pasu (cm)")+
  theme_bw()
```

Zdroj: https://rpubs.com/collnell/ggplot2 [cit. 11. 4. 2021]

Disclaimer

- Creation of this teaching material was supported by project No. 612462-EPP-1-2019-1-SK-EPPKA2-KA "University-Industry Educational Centre in Advanced Biomedical and Medical Informatics" co-funded by the Erasmus+ Programme of the European Union.
- This teaching material is licensed under a Creative Commons Attribution 4.0 International License.

Děkuji za pozornost

Mgr. Adéla Vrtková

adela.vrtkova@vsb.cz

Co-funded by the Erasmus+ Programme of the European Union