**Katedra obecné lingvistiky**
**Univerzita Palackého v Olomouci**

# Zápočtový projekt z predmetu DATA2:
## Dátová analýza reálnych viacrozmerných dát

**Jakub Čieško**
**2. roč. BŠ**
**Odbor: FIma-LHmi**
**2022/2023**

# Obsah

# 1. Úvod

## 1.1. Cieľ práce

Cieľom práce je popísať vlastnosti a entity z datasetu Titanic dataset, ktorý vznikol z údajov o pasažieroch Titanicu. Zámerom projektu je tiež odhaliť, ktoré vlastnosti mohli mať vplyv na šancu jednotlivých pasažierov prežiť potopenie Titanicu.

## 1.2. Popis dát

Dáta z Titanic datasetu sú zložené z informácií o 891 entitách – pasažieroch – popísaných pomocou 11 vlastností rôzneho druhu. Jednotlivé vlastnosti sú popísane v podnadpise 1.2.1. Základný popis a definícia štatistických znakov v datasete a v tabuľke 1.

### 1.2.1. Základný popis a definícia štatistických znakov v datasete

#### 1.2.1.1. Alternatívne vlastnosti

Dve vlastnosti, pohlavie (*Sex*) a informácia o prežití (*Survived*), sú alternatívne. Pohlavie je kódované pomocou hodnôt male (muž) a female (žena). Informácia o prežití pasažiera, *Survived*, naberá hodnoty 0 a 1 – 0 v prípade úmrtia pasažiera a 1 v prípade prežitia.

#### 1.2.1.2. Ostatné vlastnosti

Vek pasažiera (*Age*) je vlastnosť vyjadrujúca vek cestujúceho. V prípade detí mladších ako 1 rok je jeho hodnota vyjadrená číslom medzi 0 a 1. Pokiaľ je vek cestujúceho iba odhadovaný (nie je istý) jeho hodnota má tvar XX.5, kde XX označuje odhadovaný vek. Ináč je vek celočíselný.

Počet súrodencov alebo partnerov na palube Titanicu (*SibSp*) naberá celočíselné hodnoty. Za partnerov sú považovaní iba manželia.

Počet rodičov alebo detí na palube Titanicu (*Parch*) taktiež naberá iba hodnotu celých čísel. V prípade deti, ktoré cestovali bez rodičov, ale v sprievode opatrovateľa/ky, naberá tento znak hodnotu 0.

Výška cestovného (znak *Fare*) naberá hodnoty nezáporných reálnych čísiel a vyjadruje výšku pasažierom zaplateného cestovného.

Znak *Pclass* vyjadruje triedu lístka, ktorý mal daný cestujúci. Tento znak naberá 3 rôzne hodnoty: 1 – prvá trieda, 2 – druhá trieda, 3 – tretia trieda.

*Cabin* má hodnotu reťazca, stringu, a označuje číslo kajuty, v ktorej daný cestujúci cestoval.

Vlastnosť *Embarked* označuje prístav, v ktorom daný cestujúci, nastúpil na palubu Titanicu. Táto vlastnosť má 3 rôzne hodnoty: C – Cherbourg, Q – Queenstown, S – Southampton.

Znak *Name* označuje meno pasažiera a znak *Ticket* zas číslo lístka. Oba znaky majú typ reťazcov.

| | **Preklad** | **Typ** | **Hodnoty** |
|---|---|---|---|
| **Survived** | Prežitie | Celé číslo | 0, 1 |
| **Sex** | Pohlavie | Reťazec | male, female |
| **Age** | Vek | Nezáporné číslo | 0,42–80 |
| **SibSp** | Počet súrodencov/partnerov | Celé číslo | 0–8 |
| **Parch** | Počet rodičov/detí | Celé číslo | 0–6 |
| **Fare** | Výška cestovného | Nezáporné čísla | 0–512,3292 |
| **Pclass** | Trieda lístka | Celé číslo | 1, 2, 3 |
| **Cabin** | Číslo kajuty | Reťazec | – |
| **Embarked** | Prístav nalodenia | Reťazec | C, Q, S |
| **Name** | Meno | Reťazec | – |
| **Ticket** | Číslo lístka | Reťazec | – |

**Tabuľka 1**: Štatistické znaky v datasete, ich preklad, typ a rozsah hodnôt.

#### 1.2.2. Zdroj dát

Nakoľko sa jedná o veľmi známy dataset pre machine learning, dáta sú voľne dostupné na stránke: https://www.kaggle.com/competitions/titanic/data. Dataset dostupný na uvedenej adrese je rozdelený na dve časti – na trénovanie modelu a na jeho testovanie. V tomto projekte budeme však pracovať s nerozdelenou verziou z kurzu KOL/DATA1.

#### 1.2.3. Problémy datasetu

Najväčším problémom datasetu sú chýbajúce hodnoty vlastností pre niektorých pasažierov. Tomuto problému sa budeme venovať po základnom opise dát v časti 2.2.6. Chýbajúce údaje veku.

# 2. Analýza vlastností

## 2.1. Alternatívne vlastnosti

#### 2.1.1. Vlastnosť *Survived*

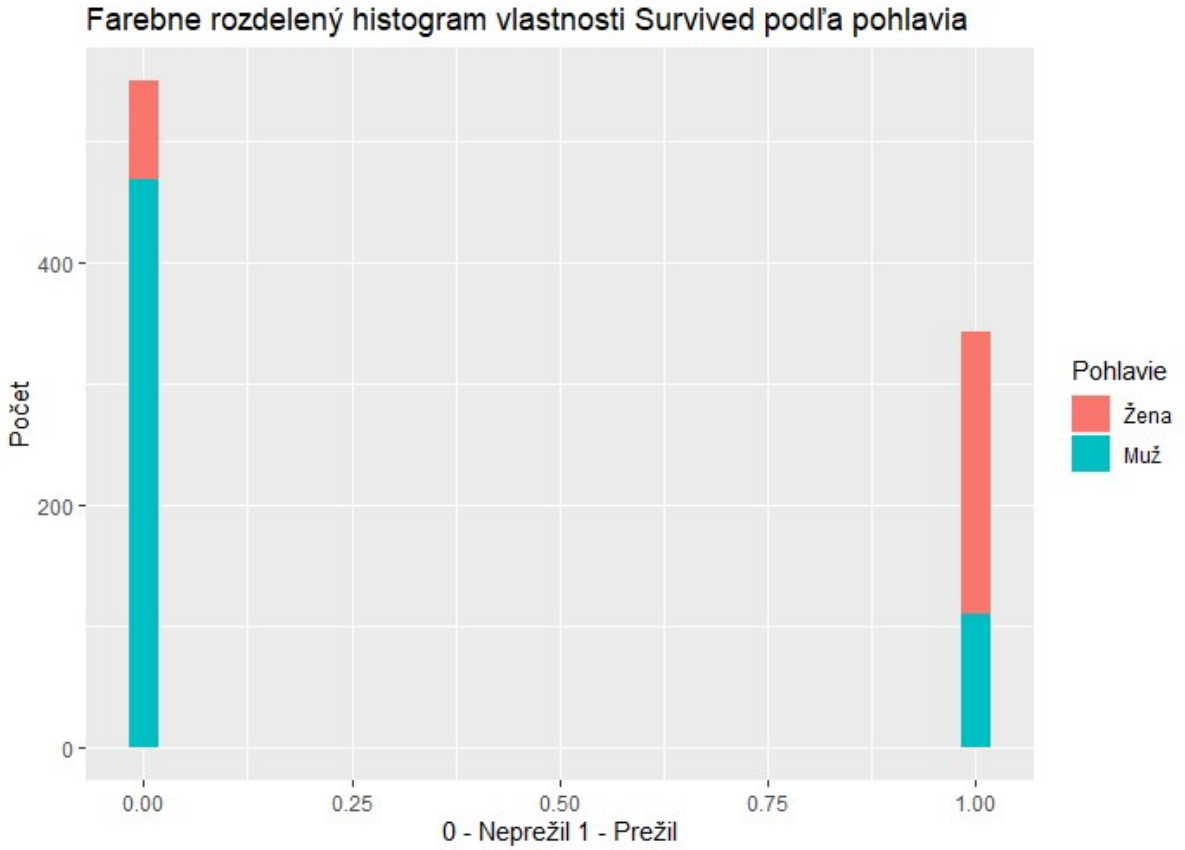
Vlastnosť *Survived* naberá dve hodnoty (0 a 1) a popisuje, či daný pasažier prežil potopenie Titanicu. V tabuľke 2 môžeme vidieť celkové počty preživších a tých, ktorí neprežili, spolu s ich percentuálnym zastúpením. Vidíme, že celkovo potopenie Titanicu prežilo 342 (38,38 %) pasažierov, zatiaľ čo zvyšných 549 (61,62 %) cestujúcich túto tragédiu neprežilo.

| Trieda | Počet | Percentuálne zastúpenie |
|---|---|---|
| **0 – neprežil** | 549 | 61,62 % |
| **1 – prežil** | 342 | 38,38 % |

**Tabuľka 2:** Počty tried štatistického znaku *Survived* s percentuálnym zastúpením.

2.1.2. Vlastnosť *Sex*

Na palube Titanicu bolo 577 mužov (64,76 % pasažierov) a 314 žien (35,24 %). Pokiaľ rozdelíme všetkých cestujúcich do dvoch skupín (podľa toho, či prežili, alebo neprežili nehodu), pomer mužov a žien sa zmení. Pomer žien medzi preživšími (68,13 %) je omnoho väčší než ich pomer v skupine všetkých pasažierov, zatiaľ čo pomer mužov je omnoho väčší medzi tými, ktorý tragédiu Titanicu neprežili – 85,25 %, ako môžeme vidieť v tabuľke 3 a na obrázku 1.

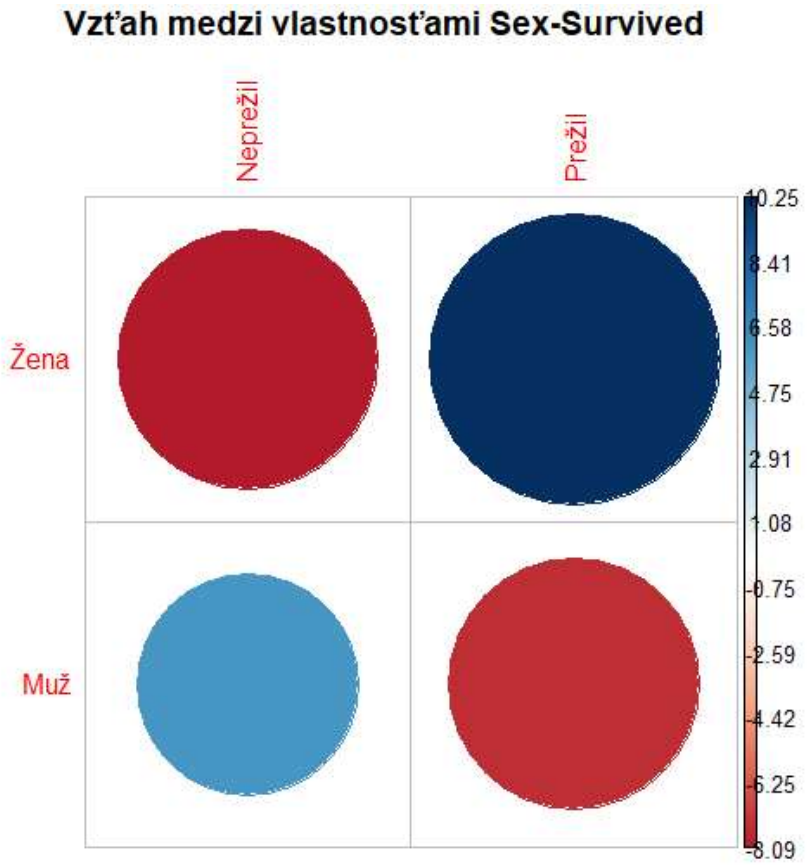



**Obrázok 1:** Farebne rozdelený histogram vlastnosti Survived podľa pohlavia.

| | Celkovo | Preživší | Obete |
|---|---|---|---|
| **Muži** | | | |
| **počet** | 577 | 109 | 468 |
| **podiel** | 64,76 % | 31,87 % | 85,25 % |
| **Ženy** | | | |
| **počet** | 314 | 233 | 81 |
| **podiel** | 35,24 % | 68,13 % | 14,75 % |

**Tabuľka 3**: Počet a podiel pohlaví medzi obeťami a preživšími potopenia Titanicu.

Na určenie vzťahu medzi pohlavím a prežitím pasažiera môžeme použiť tiež chí-kvadrátový test. Nulová hypotéza tohto testu znie, že obe veličiny (*Sex* a *Survived*) sú na sebe štatisticky nezávislé. V prípade tabuľky s rozmermi 2x2 (2 pohlavia x obeť/preživší) sa použije chí-kvadrátový test s 1 stupňom voľnosti. S pomocou príkazu chisq.test v programovacom jazyku R dostávame hodnotu štatistiky 260,72 a p-hodnoty menšiu než $2{,}2.10^{-16}$, a teda môžeme nulovú hypotézu zamietnuť. Mieru vzťahu medzi pohlavím a prežitím môžeme vidieť na obrázku 2. Modré odtiene znamenajú pozitívny vzťah premenných, červené negatívny. Z obrázku je jasné, že ženy majú väčšiu šancu mať hodnotu vlastnosti *Survived* 1 než muži, a muži majú väčšiu šancu mať hodnotu vlastnosti *Survived* 0 než ženy.




**Obrázok 2:** Vzťah a jeho miera medzi vlastnosťami *Sex* a *Survived*.

Celkovo tak tragédiu prežilo 18,89 % a neprežilo 81,11 % všetkých mužov. Zo všetkých žien potopenie Titanicu prežilo 74,20 % a neprežilo 25,80 %. Napriek ich celkovo nižšiemu zastúpeniu na palube Titanicu môžeme usúdiť, že ženy mali väčšiu šancu prežiť potopenie Titanicu než muži.

## 2.2. Ostatné vlastnosti

### 2.2.1. Vlastnosť *Pclass*

Triedy lístkov na Titanic boli 3. Najpočetnejšia bola najnižšia, tretia trieda. Druhá najpočetnejšia bola prvá trieda. A najmenej početná bola stredná, druhá trieda. Počty a podiely cestujúcich jednotlivých tried môžeme vidieť v tabuľke 4.

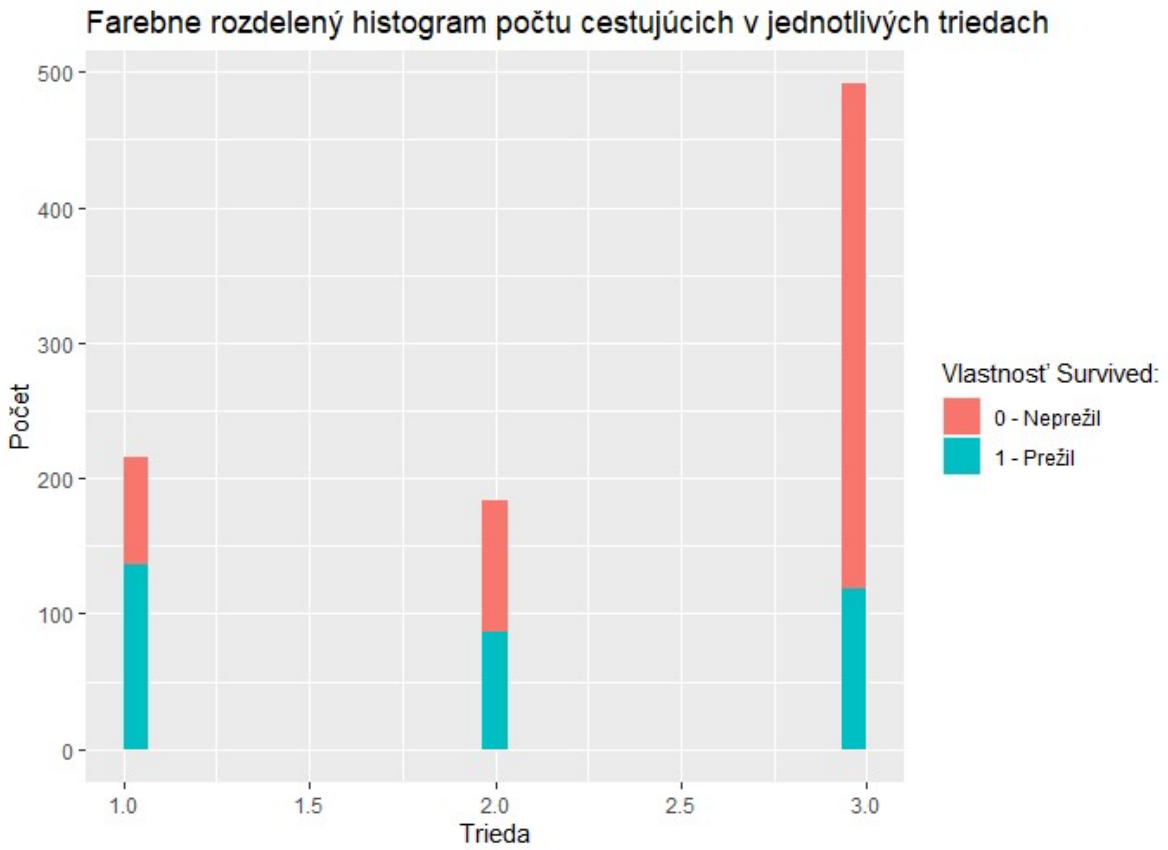
| | Počet | Percentuálne zastúpenie |
|---|---|---|
| **1. trieda** | 216 | 24,24 % |
| **2. trieda** | 184 | 20,65 % |
| **3. trieda** | 491 | 55,11 % |

**Tabuľka 4:** Počet a podiel cestujúcich jednotlivých tried.

Na obrázku 3 a v tabuľke 5 vidíme, že 62,96 % cestujúcich v prvej triede prežilo potopenie Titanicu, zatiaľ čo v prípade cestujúcich v druhej triede to bolo len 47,28 % a v tretej triede iba 24,24 %.

| | Preživší | Obete |
|---|---|---|
| **1. trieda** | 136 (62,96 %) | 80 (37,04 %) |
| **2. trieda** | 87 (47,28 %) | 97 (52,72 %) |
| **3. trieda** | 119 (24,24 %) | 372 (75,76 %) |

**Tabuľka 5:** Počet a podiel preživších a obetí v jednotlivých triedach.

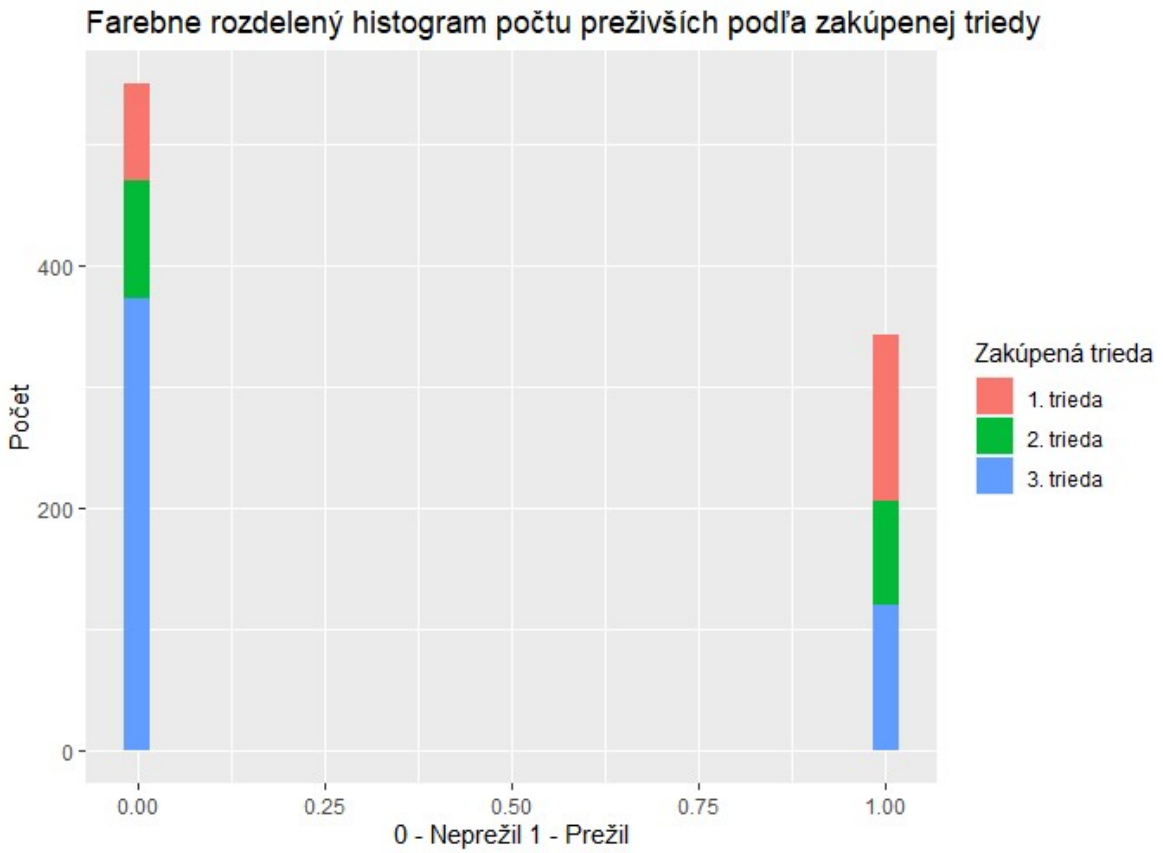



**Obrázok 3:** Farebne rozlíšený histogram počtu cestujúcich v jednotlivých triedach.

Pokiaľ porovnáme zastúpenie tried v dvoch skupinách cestujúcich rozdelených na základe hodnoty vlastnosti *Survived*, zistíme, že pôvodný pomer zastúpenia jednotlivých tried (24,24 : 20,65 : 55,11) je v týchto dvoch skupinách celkom odlišný, ako môžeme vidieť na obrázku 4 a v tabuľke 6.

| | Podiel medzi obeťami | Podiel medzi preživšími |
|---|---|---|
| **1. trieda** | 14,57 % | 39,77 % |
| **2. trieda** | 17,67 % | 25,44 % |
| **3. trieda** | 67,76 % | 34,80 % |

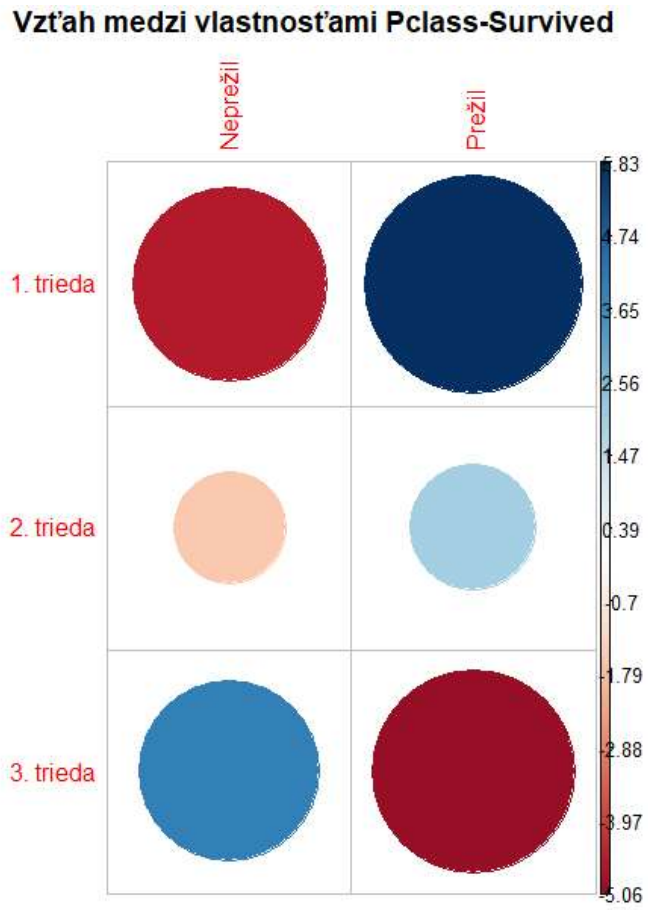
**Tabuľka 6:** Podiel cestujúcich jednotlivých tried medzi obeťami a preživšími.




**Obrázok 4:** Farebne rozlíšený histogram vlastnosti Survived podľa jednotlivých tried.

Zdá sa teda, nakoľko je pomer cestujúcich z 3. triedy medzi preživšími väčší než pomer cestujúcich z 2. triedy, že aj pravdepodobnosť toho, že ten, kto prežil potopenie Titanicu, bol cestujúcim 3. a nie 2. triedy, bude väčšia. Avšak najväčšiu šancu prežiť mali aj tak cestujúci 1. triedy.

Podobne ako v prípade zisťovania vzťahu medzi binárnymi vlastnosťami pohlavia a prežitia, môžeme aj v tomto prípade použiť chí-kvadrátový test. Nakoľko máme v tomto prípade 3 rôzne možnosti hodnoty vlastnosti *Pclass* a 2 rôzne možnosti hodnoty vlastnosti *Survived*, použijeme chí-kvadrátový test s 2 stupňami voľnosti. Hodnota vypočítanej štatistiky je potom 102,89 a p-hodnota je menšia než $2{,}2.10^{-16}$, a teda hypotézu nezávislosti vlastností *Survived* a *Pclass* môžeme zamietnuť. Silu vzťahu medzi týmito vlastnosťami môžeme vidieť na obrázku 5. Najsilnejší vplyv na hodnotu *Survived* majú, ako vidíme, hodnoty *Pclass* 1 a 3.

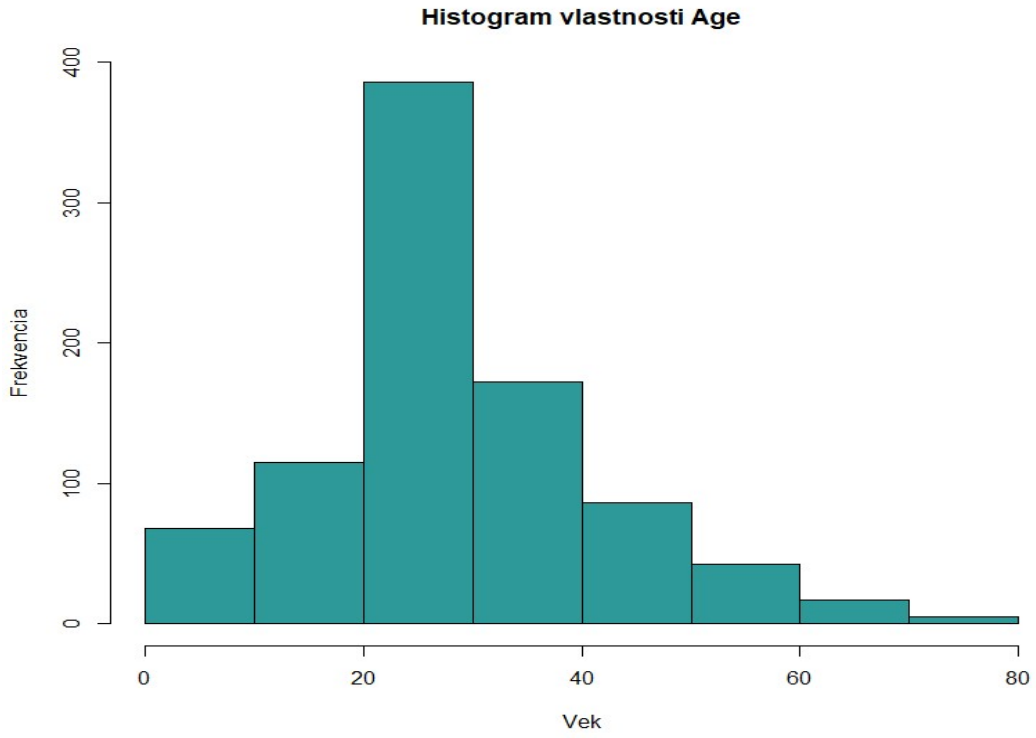



**Obrázok 5:** Vzťah medzi vlastnosťami *Pclass* a *Survived*.

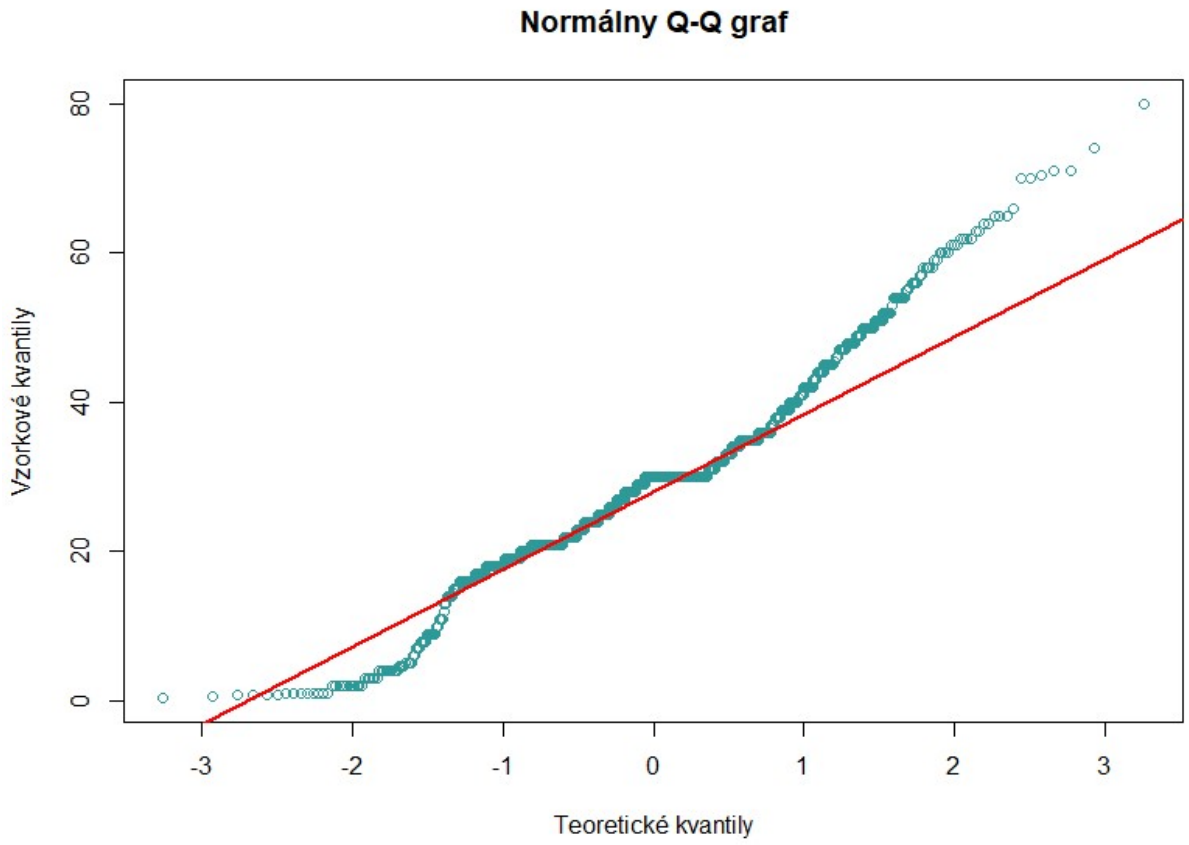
### 2.2.2. Vlastnosť *Age*

Vlastnosť *Age* má v Titanic datasete viaceré (173) neuvedené, chýbajúce hodnoty. V tejto časti sa budeme venovať deskripcii len tých dát, ktoré sú v datasete prítomné. Možné doplnenie chýbajúcich dát bude témou časti 2.2.6. Chýbajúce údaje veku.

Najstarší cestujúci Titanicu mal 80 rokov, najmladší zas 0,42 roku. Priemerný vek je 29,70 roku. Ostatné dôležité štatistiky sa nachádzajú v tabuľke 8. Histogram vlastnosti *Age* nájdeme na obrázku 6. Tvar histogramu by mohol napovedať, že sa jedná o vlastnosť s normálnym rozdelením. Avšak, keď vyhotovíme Q-Q graf (obrázok 7), zistíme, že čiaru normálneho rozdelenia vlastnosť *Age* kopíruje len zhruba. Po vypočítaní Shapiro-Wilkovho testu dostávame hodnotu štatistiky W = 0,98 a p-hodnotu = $7{,}83.10^{-8}$, a teda môžeme zamietnuť hypotézu, že by bol vek na Titanicu normálne rozdelený.




**Obrázok 6:** Histogram vlastnosti *Age*.

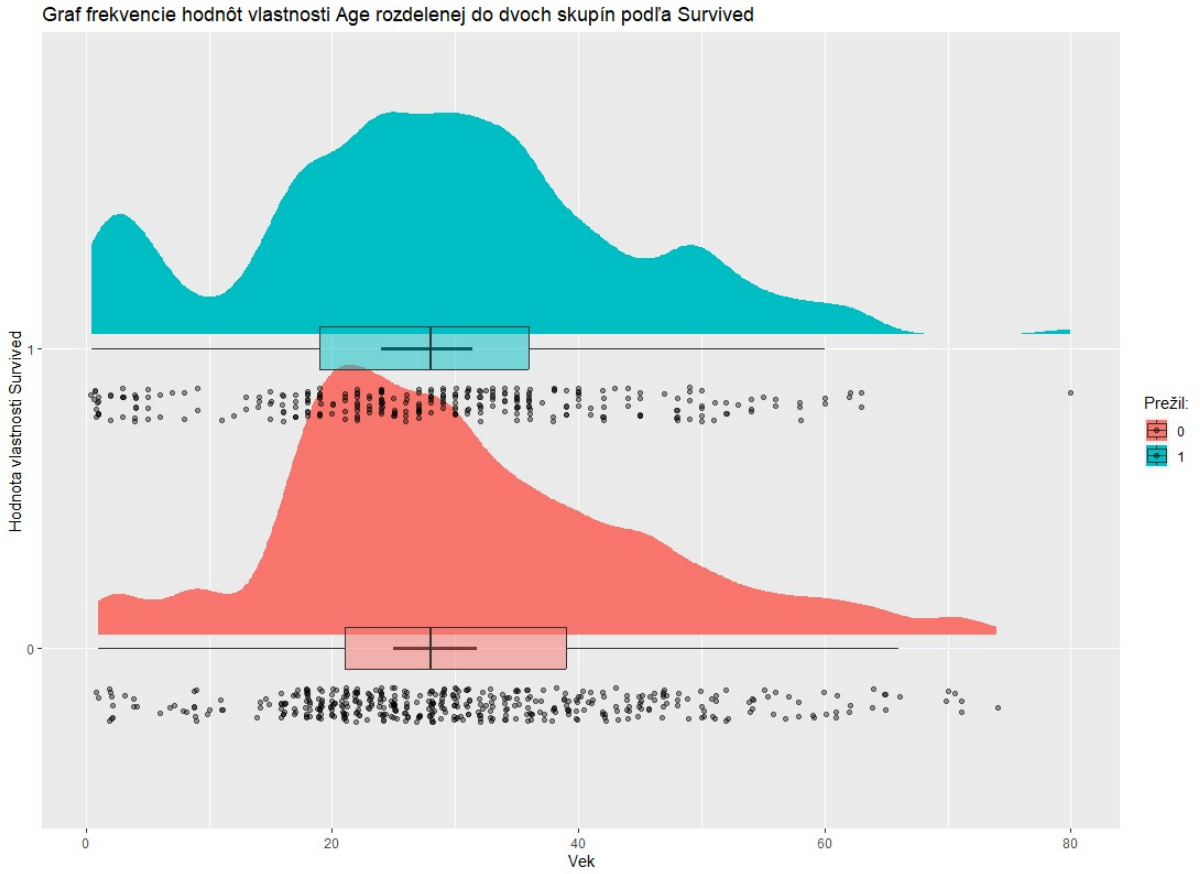



**Obrázok 7:** Normálny Q-Q graf vlastnosti *Age*.

Pokiaľ rozdelíme cestujúcich podľa hodnoty premennej *Survived*, zistíme, že cestujúci, ktorí prežili boli v priemere o niečo mladší, čo môžeme vidieť aj na obrázku 8, kde je modrý boxplot posunutý viac doľava. Presné hodnoty štatistík pre takto rozdelených cestujúcich môžeme vidieť v tabuľke 7.

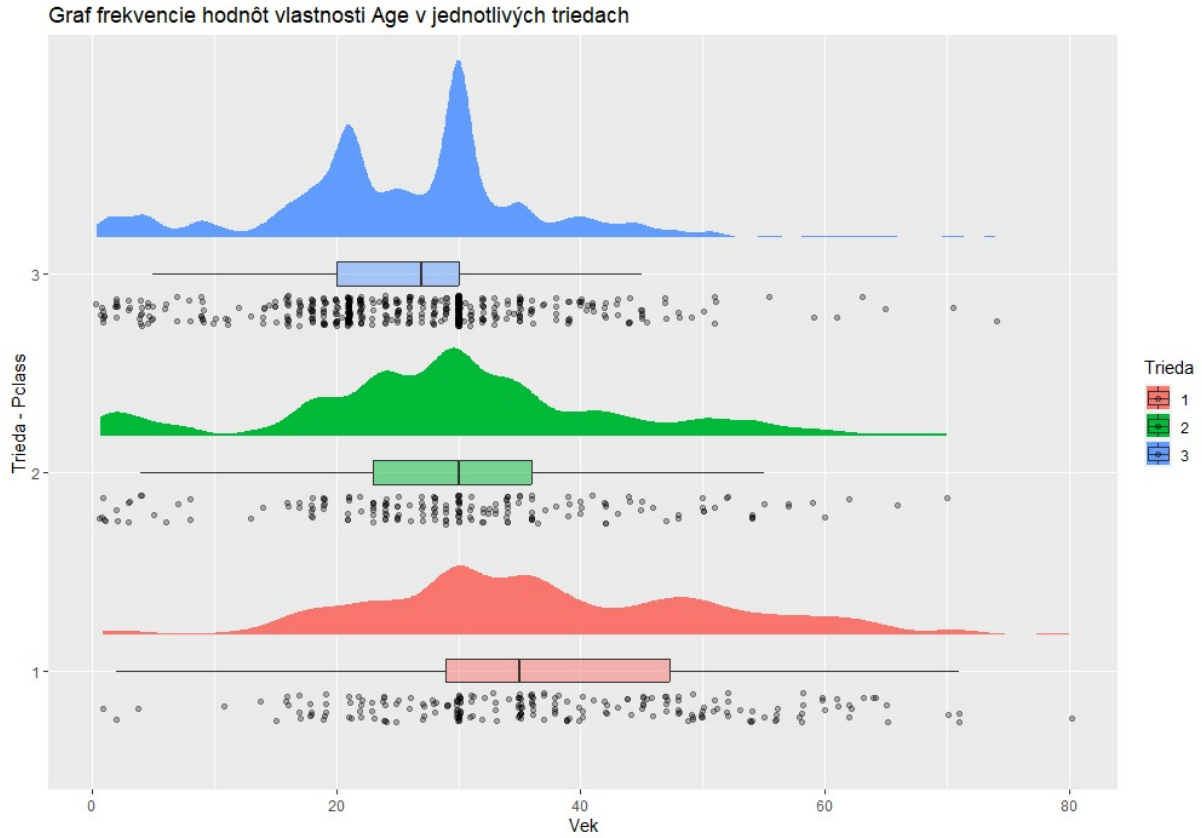
| | **Obete** | **Preživší** |
|---|---|---|
| **Min** | 1,00 | 0,42 |
| **Q1** | 21,00 | 19,00 |
| **Medián** | 28,00 | 28,00 |
| **Priemer** | 30,50 | 28,18 |
| **Q3** | 39,00 | 36,00 |
| **Max** | 74,00 | 80,00 |

**Tabuľka 7:** Hodnoty štatistík vlastnosti *Age* rozdelenej podľa hodnoty *Survived*.




**Obrázok 8:** Graf frekvencie hodnôt vlastnosti *Age* rozdelenej do dvoch skupín podľa *Survived*.

Ak rozdelíme cestujúcich do skupín podľa toho, akou triedou cestovali, dostávame výsledok zobrazený na obrázku 9. Zdá sa, že priemerný vek cestujúcich v lepších triedach je vyšší než tých v horších. Na prvý pohľad sa môže zdať, že sa toto zistenie popiera so záverom z časti venovanej vlastnosti *Pclass* – väčšiu šancu na prežitie mali cestujúci prvej triedy. Možným vysvetlením však môže byť to, že cestujúci z horších tried mali šancu na prežitie vyššiu iba kvôli tomu, že boli mladší, a teda v lepšej fyzickej kondícii.

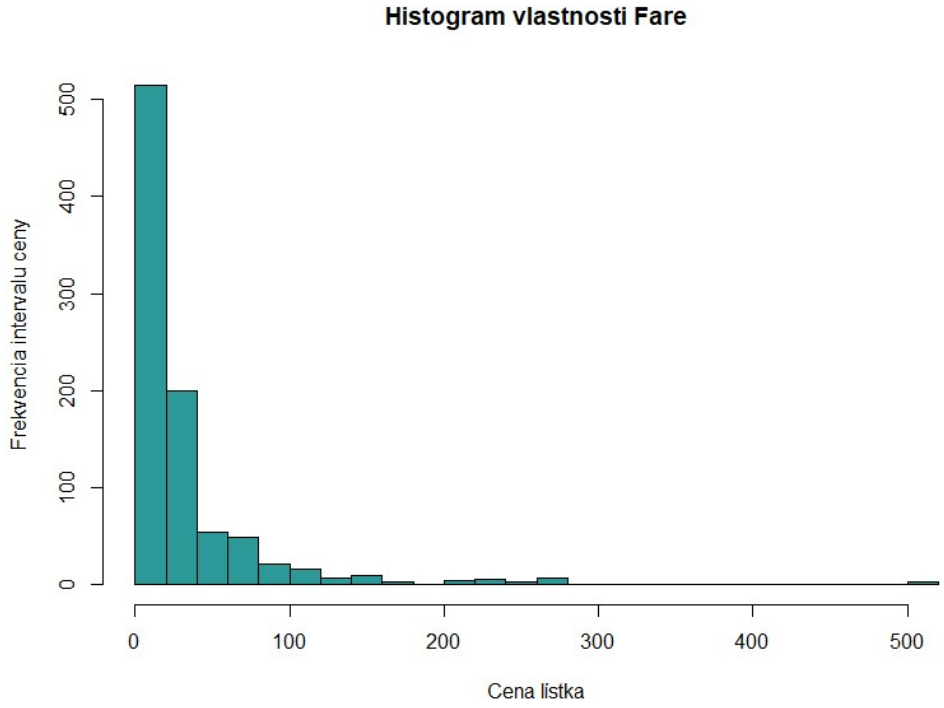



**Obrázok 9:** Graf frekvencie hodnôt vlastnosti *Age* v jednotlivých triedach.

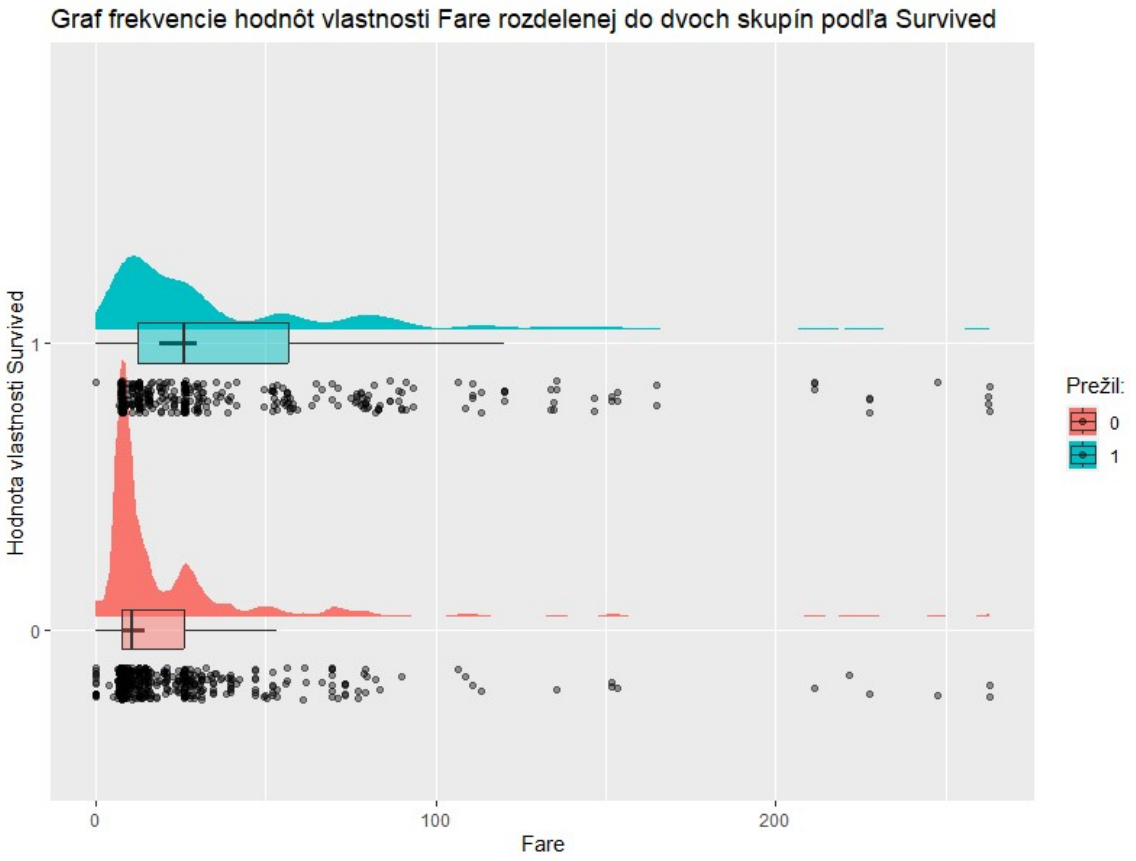
### 2.2.3. Vlastnosť *Fare*

Hodnoty štatistík vlastnosti *Fare* môžeme nájsť v tabuľke 8. Histogram tejto vlastnosti zas nájdeme na obrázku 10.

Pokiaľ rozdelíme cestujúcich do dvoch skupín podľa hodnoty znaku *Survived*, zistíme, že cestujúci, ktorí prežili, zaplatili v priemere viac peňazí, než tí, ktorí zomreli (viď obrázok 11). Tento fakt môže súvisieť s tým, čo sme pozorovali pri vlastnosti *Pclass* – cestujúci, ktorí cestovali v lepšej triede, mali väčšiu šancu na prežitie než tí, ktorí cestovali za menej peňazí – v nižšej triede.



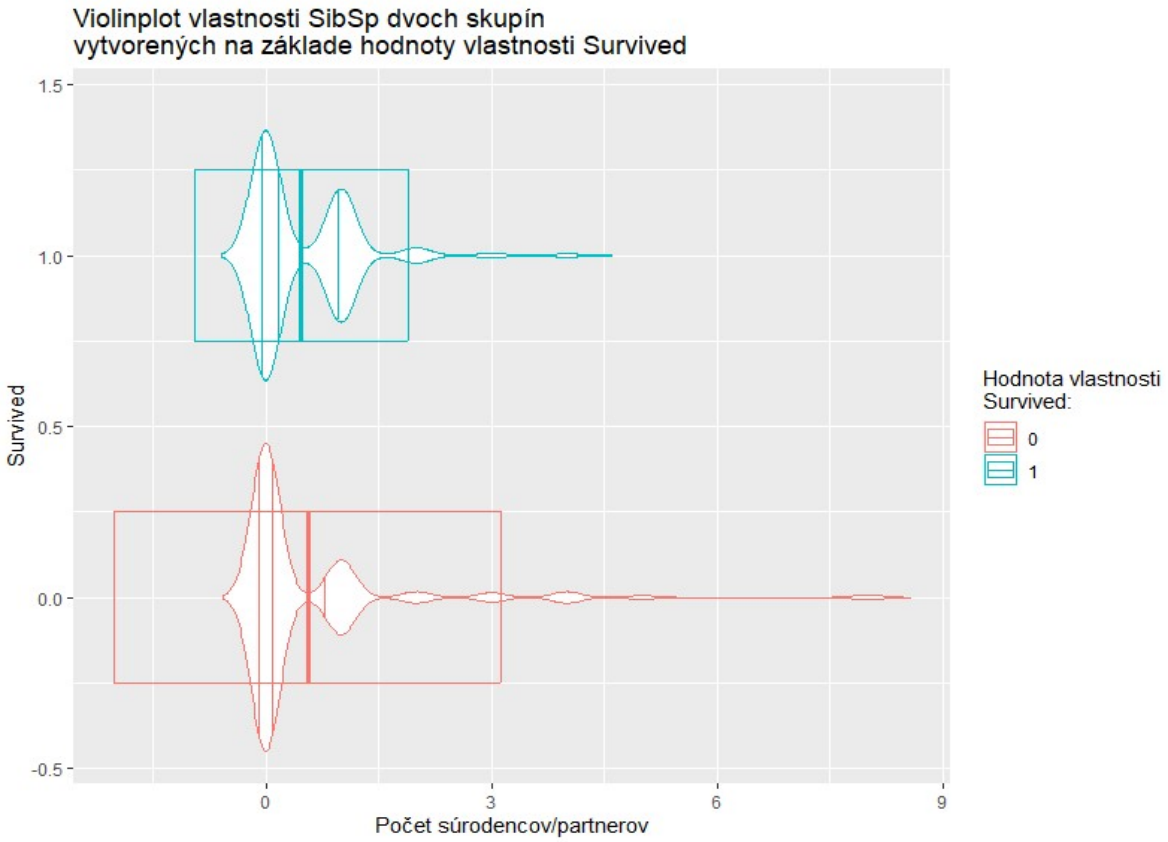

**Obrázok 10:** Histogram vlastnosti *Fare*. Veľkosť jednotlivých intervalov je 20.




**Obrázok 11:** Graf frekvencie hodnôt vlastnosti Fare dvoch skupín pasažierov podľa hodnoty znaku *Survived*.

### 2.2.4. Vlastnosť *SibSp*

Vlastnosť *SibSp* vyjadruje počet súrodencov a partnerov. Hodnoty jednotlivých štatistík sú uvedené v tabuľke 8. V prípade, že rozdelíme cestujúcich do skupín podľa toho, či prežili potopenie Titanicu, alebo nie, zistíme, že v priemere mali pasažieri, ktorí prežili, na palube menej súrodencov či partnerov, ako to môžeme vidieť aj na obrázku 12, na ktorom je vyobrazený violinplot vlastnosti *SibSp* dvoch skupín vytvorených na základe hodnoty vlastnosti *Survived*. Na grafe sú zvýraznené aj hodnoty prvého, tretieho kvartilu, medianu, priemeru a smerodajnej odchýlky. Tento rozdiel sa však nezdá ako významný.



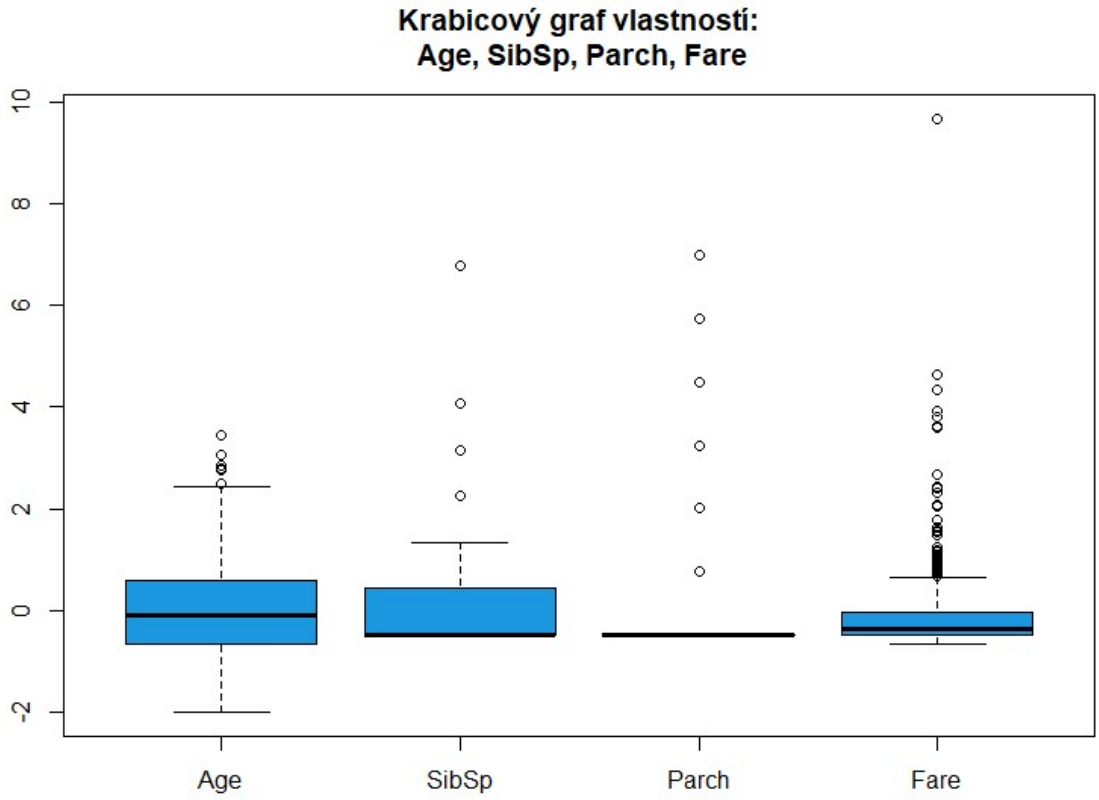

**Obrázok 12:** Violinplot vlastnosti *SibSp* dvoch skupín vytvorených na základe hodnoty vlastnosti *Survived*.

### 2.2.5. Vlastnosť *Parch*

Štatistický znak *Parch* zastupuje počet rodičov a detí jednotlivých pasažierov. Hodnoty jednotlivých štatistík spolu s ostatnými znakmi sú uvedené v tabuľke 8. Na obrázku 13 môžeme vidieť boxplot všetkých štatistických znakov skúmaných v tejto časti. Dáta boli pomocou z-skóre pred vyobrazením preškálované. Hodnota osi y teda nie je v jednotkách jednotlivých znakov, ale v smerodajnej odchýlke.

| | Age | SibSp | Parch | Fare |
|---|---|---|---|---|
| **Min** | 0,42 | 0,00 | 0,00 | 0,00 |
| **Max** | 80,00 | 8,00 | 6,00 | 512,33 |
| **Priemer** | 29,70 | 0,523 | 0,38 | 32,20 |
| **Medián** | 28,00 | 0,00 | 0,00 | 14,45 |
| **SD** | 14,53 | 1,10 | 0,81 | 49,69 |
| **Q1** | 20,12 | 0,00 | 0,00 | 7,91 |
| **Q3** | 38,00 | 1,00 | 0,00 | 31,00 |
| **NA** | 177 | | | |

**Tabuľka 8**: Popisné štatistiky datasetu.




**Obrázok 13:** Krabicový graf vlastností: *Age*, *SibSp*, *Parch*, *Fare*.

### 2.2.6. Chýbajúce údaje veku

Nakoľko vek môže hrať dôležitú úlohu pri tom, či daný pasažier prežil potopenie Titanicu, je pomerne veľkým problémom, že v tomto datasete nie je udaný nijaký vek pri 177 pasažieroch. Preto sa pokúsime aspoň nejakým spôsobom rekonštruovať vek pasažierov pred tým, ako vytvoríme model lineárnej regresie.

Jednoduchým riešením by bolo doplniť na všetky chýbajúce pozície hodnotu priemerného veku. Avšak nakoľko je priemerný vek 29,70, a medzi tými, ktorí nemajú uvedenú hodnotu veku, sú často viaceré vekové skupiny (napr. deti), ktoré určite neobsahujú človeka, ktorý by mohol mať takýto vek, mohla by táto metóda príliš ovplyvniť výsledky.

Použijeme preto trochu jemnejšiu metódu. Keďže deti, ktoré sú na palube Titanicu, majú vo svojom mene uvedené slovo Master, môžeme rozdeliť všetkých cestujúcich bez hodnoty veku na dve skupiny – na deti a zvyšok. Následne tým zo skupiny detí pridelíme priemerný vek detí (pasažierov so slovom Master v mene) ako ich nový vek.

Skupinu ostatných pasažierov potom rozdelíme na základe titulu, ktorý majú uvedený vo svojom mene, na viaceré skupiny. V každej tejto skupine potom nájdeme medián premennej *Age*. Následne každému z ostatných pasažierov s hodnotou veku NA pridelíme vek podľa titulu, ktorý má uvedený vo svojom mene. Touto metódou potom získavame dataset, v ktorom má každý pasažier pridelený vek, ktorý nie je tak veľmi skreslený, ako by mohol byť v prípade, že by sme postupovali prvou spomenutou metódou. Hodnoty štatistík takto upravenej premennej *Age* uvádzame v tabuľke 9.

| | **Min** | **Max** | **Priemer** | **Medián** | **Q1** | **Q3** | **SD** |
|---|---|---|---|---|---|---|---|
| **Pred úpravou** | 0,42 | 80,00 | 29,70 | 28,00 | 20,12 | 38,00 | 14,53 |
| **Po úprave** | 0,42 | 80,00 | 29,38 | 30 | 21,00 | 35,00 | 13,24 |

**Tabuľka 9**: Popisné štatistiky premennej *Age* pred a po úprave.

# 3. Lineárna regresia

Na vytvorenie modelu lineárnej regresie použijeme upravené dáta. Popis úpravy dát sa nachádza v časti 2.2.6. Chýbajúce údaje veku.

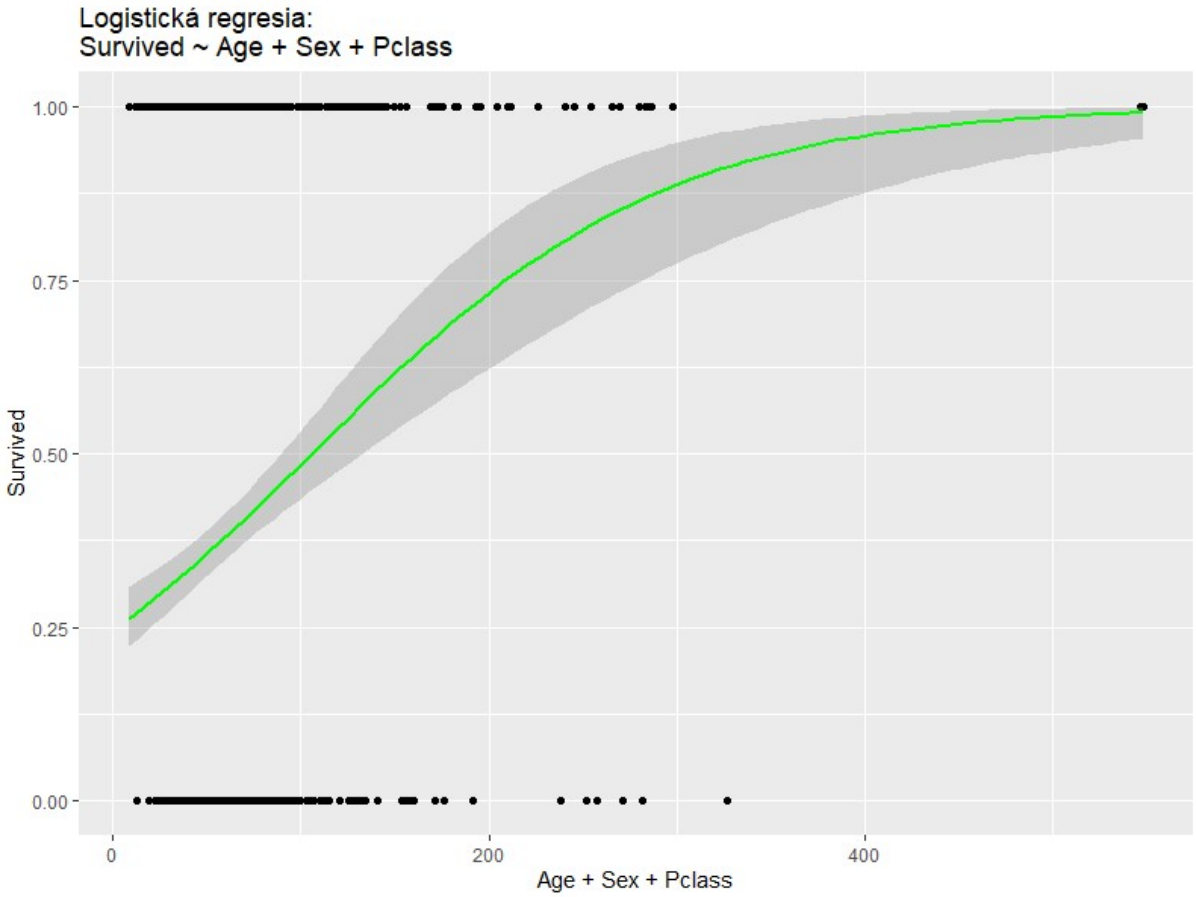
## 3.1. Model *Survived ~ Age + Sex + Pclass*.

Metóda na odhad hodnoty binárnej premennej na základe hodnôt iných vlastností sa nazýva logistická regresia. Model logistickej regresie vytvárame pomocou známych javov, ktoré môžu ovplyvniť výslednú hodnotu alternatívnej premennej. V našom prípade sme si zvolili trojicu premenných, ktoré by mohli mať vplyv na výslednú hodnotu vlastnosti *Survived*: *Age*, *Sex*, *Pclass*. To, že tieto vlastnosti môžu mať vplyv na hodnotu cieľovej premennej, sme zistili v našej analýze. Hodnotu *Fare*, ktorá taktiež mohla byť vhodným kandidátom sme vynechali, pretože už hodnota *Pclass* nesie nejakú informáciu o tom, čo vyjadruje aj *Fare*, sociálnom statuse cestujúceho. Pomocou týchto vlastností a logistickej regresie sa teda pokúsime odhadnúť hodnotu alternatívnej premennej *Survived*.

Výsledné koeficienty tohto modelu logistickej regresie môžeme vidieť v tabuľke 10. Graf logistickej regresie je vyobrazený na obrázku 14. Koeficienty logistickej regresie udávajú zmenu log-pravdepodobnosti nárastu vlastnosti *Survived* o jednotku – napr. jednotková zmena premennej *Age* zmenší log-pravdepodobnosť o 0,03.

| | **Odhad** | **Štandardná chyba** | **z-hodnota** | **p-hodnota** |
|---|---|---|---|---|
| **Priesečník** | 4,80 | 0,45 | 10,65 | $1{,}79.10^{-26}$ |
| **Age** | -0,03 | 0,01 | -4,76 | $1{,}98.10^{-6}$ |
| **Sex** | -2,59 | 0,19 | -13,88 | $8{,}72.10^{-44}$ |
| **Pclass** | -1,20 | 0,12 | -9,89 | $4{,}45.10^{-23}$ |

**Tabuľka 10**: Koeficienty logistickej regresie.




**Obrázok 14:** Logistická regresia: Survived ~ Age + Sex + Pclass.

Po aplikovaní tohto modelu na dáta dostávame confussion matrix (chybovú maticu) v tabuľke 11. Pričom hodnota správnosti (accuracy) je 79,35 % a hodnota precíznosti (precision) 75,16 %. Úplnosť (recall) je 69 %, špecificita 85,79 % a F-miera (F-score) 71,95 %.

| | **Skutočný výsledok 0** | **Skutočný výsledok 1** |
|---|---|---|
| **Odhadovaný výsledok 0** | 471 | 106 |
| **Odhadovaný výsledok 1** | 78 | 236 |

**Tabuľka 11**: Confussion matrix modelu logistickej regresie.

# 4. Analýza entít

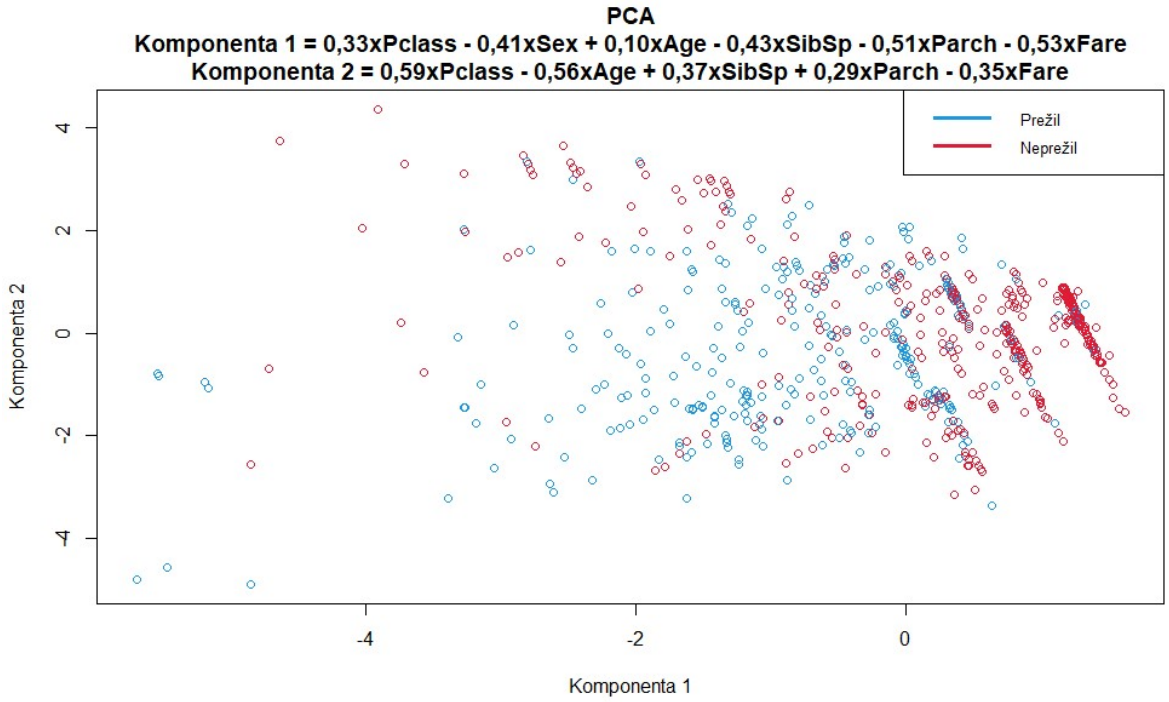
V časti analýza entít sa budeme venovať redukovaniu rozmerov dát pomocou metódy analýzy hlavných komponentov (časť 4.1. PCA), analýze zhlukov (časť 4.2. Hierarchické zhlukovanie) a analýze dát pomocou multidimenzionálneho škálovania (časť 4.3. MDS).

## 4.1. PCA

Pri analýze hlavných komponentov sa hľadajú lineárne kombinácie pôvodných premenných, ktoré zachovávajú čo najväčší podiel informácií, vyjadrený pomocou rozptylu, a zároveň majú

menši alebo nanajvýš rovnaký počet dimenzií ako pôvodné dáta. Jednotlivé syntetické komponenty majú potom jasne daný význam pomocou tzv. loadings.

Nakoľko jednotky jednotlivých vlastností sú určite rôzne, musíme pred vykonaním PCA dáta štandardizovať, preškálovať pomocou z-skóre. Po vykonaní PCA zistíme, že na dosiahnutie 95% pôvodnej informácie nám z pôvodných 7 vlastností (*Pclass*, *Sex*, *Age*, *SibSp*, *Parch*, *Fare* a *Embarked*) stačí iba 6 premenných. Pričom prvé dva hlavné komponenty s najväčšou smerodajnou odchýlkou zachycujú 50% pôvodných informácií. Práve tieto dve komponenty použijeme aj na vyobrazenie dát na obrázku 15. Jednotlivé body sú zafarbené podľa ich hodnoty znaku *Survived*. Modré body – preživší – sa vo väčšej miere vyskytujú v zápornej časti osi x, čo môže byť spôsobené napríklad tým, že medzi preživšími je viac žien (vyššia hodnota *Sex*) alebo bohatších ľudí (väčšia hodnota *Fare*). Čo sa týka polohy modrých bodov vzhľadom k osi y nie je medzi ňou a polohou červených bodov väčší rozdiel. Zdá sa, že pomocou PCA nemôžeme dostatočne od seba odlíšiť preživších a obete. Na obrázku 15 sú v podnadpise uvedené aj hodnoty loadings.




**Obrázok 15:** Zobrazenie dát Titanic datasetu pomocou dvoch hlavných komponentov.

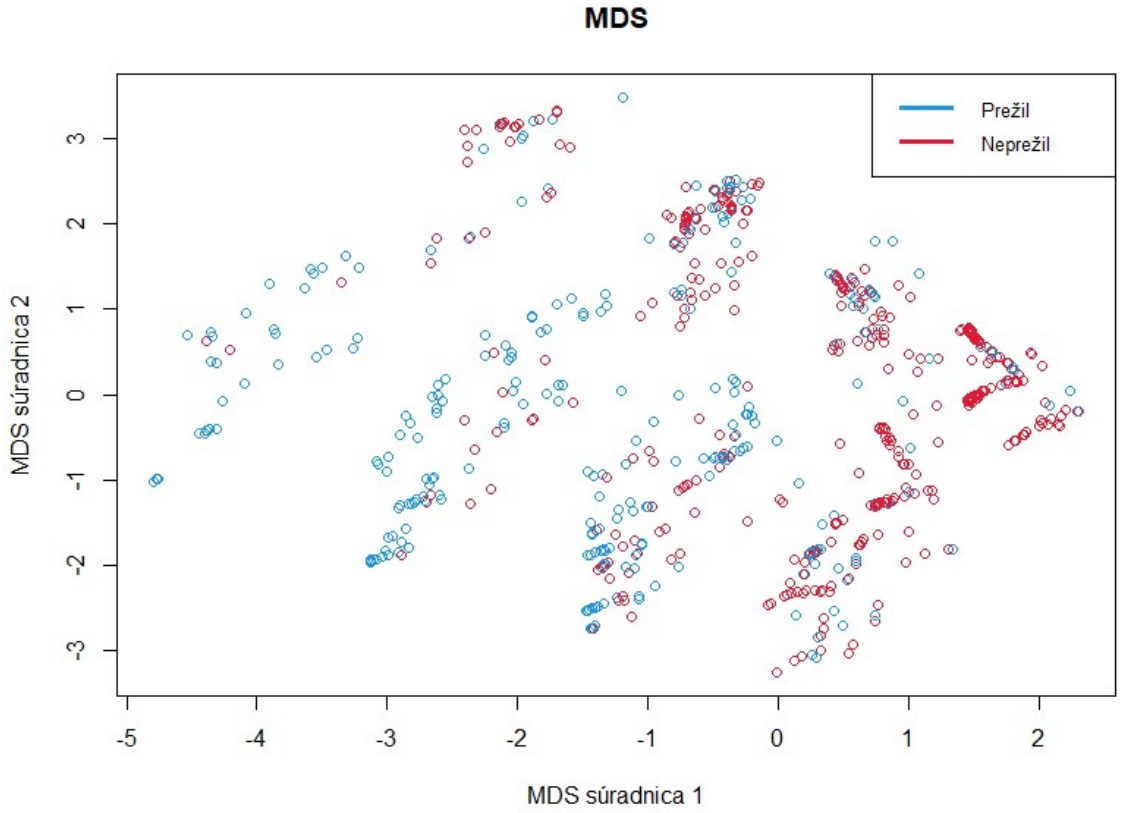
## 4.2. Hierarchické zhlukovanie

Rovnako ako neskôr na MDS, aj na hierarchické zhlukovanie použijeme párovú maticu vzdialeností. Aby sme s úplnou istotou mohli určiť, do ktorého z dvoch zhlukov budú jednotlivé dáta spadať, použijeme ako kritérium vzdialenosť priemerov clusterov, čiže metódu *average linkage*. Nakoľko sa snažíme nájsť v dátach dve skupiny pasažierov – podľa hodnoty znaku *Survived* – zvolili sme si ako cieľový počet zhlukov 2 zhluky. Takto určenú príslušnosť ku zhluku sme potom vyobrazili pomocou MDS a farieb na spodnej polovici obrázku 17. Ako vidíme, po porovnaní výsledkov hierarchického zhlukovania s hodnotou znaku *Survived* nemôžeme povedať, že by sa nám podarilo nájsť dostatočne presný spôsob vytvorenia zhlukov, ktoré by odpovedali skupinám pasažierov s rovnakou hodnotou cieľovej vlastnosti *Survived*.

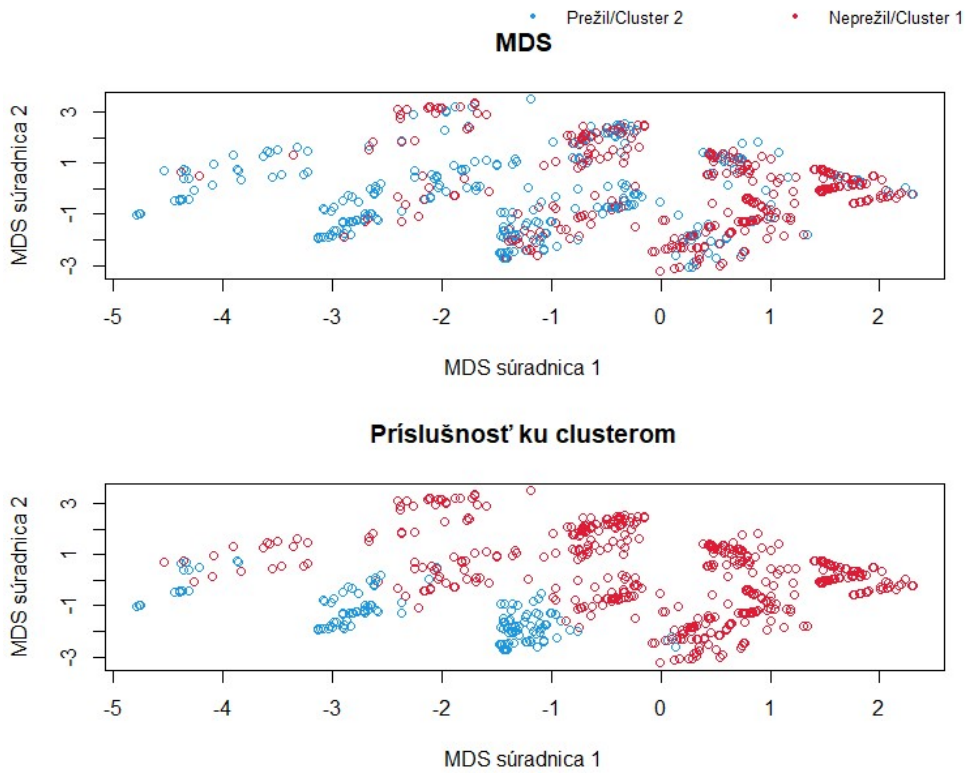
## 4.3. MDS

Vstupom pre metódu MDS je párová matica vzdialeností. Keďže nás zaujímajú absolútne rozdiely v jednotlivých vlastnostiach, použijeme ako metriku vzdialenosti canberrskú vzdialenosť, ktorá je váženou manhattanskou vzdialenosťou. Nakoľko však naše dáta majú rôzne jednotky, musíme ich najprv preškálovať pomocou z-skóre.

Po preškálovaní dát a prevedení MDS dostávame podiel pôvodného rozptylu (GOF – goodness of fit) dvoch vzniknutých rozmerov: 0,27 a 0,38. Dáta vykreslené pomocou týchto rozmerov môžeme vidieť na obrázku 16, na ktorom sú jednotlivé body zafarbené podľa ich hodnoty vlastnosti *Survived*. Nakoľko súradnice MDS sú syntetické a nie je jasné, ako zodpovedajú pôvodným vlastnostiam datasetu, nemôžeme úplne presne interpretovať zdanlivo väčšiu koncentráciu modrých bodov – preživších – v ľavej časti grafu.




**Obrázok 16:** Dvojrozmerné vyobrazenie viacrozmerných dát z datasetu Titanic pomocou metódy MDS.




**Obrázok 17:** Vyobrazenie príslušnosti k dvom zhlukom pomocou MDS a porovnanie s hodnotami štatistického znaku *Survived*.

# 5. Záver

V tejto práci sme sa pokúsili vykonať analýzu dát z datasetu Titanic s cieľom nájsť vlastnosti, ktoré mali najväčší vplyv na šancu prežiť potopenie Titanicu.

Najprv sme sa venovali jednotlivým vlastnostiam, ktoré tento dataset obsahuje, uviedli sme hodnoty viacerých štatistík, prípadne dané dáta vyobrazili pomocou grafov. Z našej analýzy vyplynulo, že najväčšiu šancu prežiť potopenie Titanicu mali najmä, po prvé, ženy, po druhé, bohatší cestujúci, resp. tí, ktorí cestovali vyššou triedou, a nakoniec, mladší pasažieri.

Ďalej sme sa pokúsili o analýzu entít a ich vzťahov pomocou metódy PCA, MDS a hierarchického zhlukovania. Nanešťastie sa nám však nepodarilo nájsť dostatočne vhodné zobrazenie dát (v prípade PCA a MDS) či ich zaradenie do zhlukov (v prípade hierarchického zhlukovania), pri ktorom by bolo možné veľmi jednoducho odlíšiť od seba pasažierov Titanicu na základe toho, či nehodu prežili alebo nie. Možným riešením by avšak mohlo byť zohľadnenie iných štatistických znakov ako tých, ktoré sme brali do úvahy v tejto práci, ich výhodnejšie a rozumnejšie spracovanie, alebo dokonca vytvorenie úplne nových vlastností.