

Vysoká škola ekonomická v Praze  
Fakulta informatiky a statistiky



# **Modely logistické regrese v oblasti esportových dat**

## **BAKALÁŘSKÁ PRÁCE**

Studijní program: Aplikovaná informatika

Studijní obor: Aplikovaná informatika

Autor: Michal Lauer

Vedoucí práce: Ing. Zdeněk Šulc, Ph.D.

Praha, Duben 2022

## Prohlášení

Prohlašuji, že jsem bakalářskou práci *Modely logistické regrese v oblasti esportových dat* vypracoval samostatně za použití v práci uvedených pramenů a literatury.

V Praze dne DD. Dubna 2021

.....

Podpis studenta

---

## **Poděkování**

Rád bych poděkoval panu doktoru Zdenku Šulcovi, který mou bakalářskou práci podpořil a vedl, i přes odlišný studijní obor. Dále děkuji autorům citovaných knih za poskytnutou příležitost se ve logistických modelech zlepšit. Bez nich by se práce psala velmi složitě.

---

## Abstrakt

Cílem bakalářské práce je kvantitativně zanalyzovat esportové zápasy ze hry Counter-Strike: Global Offensive (CSGO) a predikovat výhry **hráčů a týmů**. Použitý datový soubor je z internetového portálu kaggle.com a obsahuje data od roku 2015 až do roku 2020. Práce je rozdělena do tří částí. V první části je představen esport jako takový, je zde shrnutý jeho vývoj a jsou zde definovány důležité pojmy a termíny. Druhá část obsahuje popis metod, které jsou pro analýzu a predikci použity. Predikce jsou založené na vícerozměrném logistickém regresním modelu. **V závěrečné praktické části jsou modely sestavené jak pro hráče, tak pro týmy a to přes různé kategorie map. Modely jsou následně vyhodnoceny pomocí matice záměn a významnost koeficientů je testována pomocí Waldova testu. Modely jsou následně interpretovány v kontextu esportu a reálného života.**

## Klíčová slova

model, logistická regrese, predikce, esport

---

## **Abstract**

- Bude přeložen po odsouhlasení abstraktu v češtině

## **Keywords**

model, logistic regression, prediction, esport

# Obsah

<b>1</b>	<b>Úvod</b>	<b>2</b>
<b>2</b>	<b>Představení esportu</b>	<b>4</b>
2.1	Historie esportu . . . . .	4
2.2	Zasazení do dnešní doby . . . . .	4
2.3	Představení titulu Counter-Strike: Global Offensive . . . . .	5
2.4	Propojení práce a titulu Counter-Strike: Global Offensive . . . . .	7
<b>3</b>	<b>Teoretická část</b>	<b>9</b>
3.1	Vizualizace dat . . . . .	9
3.1.1	Bodový graf . . . . .	9
3.1.2	Sloupcový graf . . . . .	10
3.1.3	Histogram . . . . .	10
3.1.4	Krabičkový graf . . . . .	12
3.1.5	Korelační matice . . . . .	14
3.2	Logistická regrese . . . . .	17
3.2.1	Interpretace parametrů . . . . .	17
3.2.2	Největší pravděpodobnost . . . . .	18
3.2.3	Křížová validace . . . . .	19
3.2.4	Matice záměn . . . . .	19
3.2.5	Testování hypotéz . . . . .	20
3.2.6	Waldův test . . . . .	20
<b>4</b>	<b>Praktická část</b>	<b>21</b>
4.1	Cíle analýzy . . . . .	21
4.2	Datové soubory . . . . .	22
4.2.1	soubor players.csv . . . . .	22
4.2.2	soubor results.csv . . . . .	23
4.2.3	Omezení datového souboru . . . . .	23
4.3	Průzkumová analýza dat . . . . .	24
4.3.1	Korelační matice . . . . .	25
4.3.2	Histogramy kvantitativních prediktorů . . . . .	26
4.3.3	Sloupcový graf výher přes počáteční stranu . . . . .	27
4.4	Modely pro výkon hráče . . . . .	27
4.4.1	Model pro mapu Mirage . . . . .	28
4.4.2	Matice záměn pro mapu Mirage . . . . .	28
4.4.3	Model pro mapu Vertigo . . . . .	29
4.4.4	Matice záměn pro mapu Vertigo . . . . .	31
4.4.5	Interpretace výsledků . . . . .	31

---

<b>5 Závěr</b>	<b>32</b>
5.1 Závěrečné vyhodnocení modelu . . . . .	32
5.2 Interpretace modelu do reálného světa . . . . .	32
5.3 Použití modelu v reálném světě . . . . .	32
5.4 Místo pro budoucí vylepšení . . . . .	32
<b>Seznam použité literatury</b>	<b>33</b>
<b>Seznam elektronických zdrojů</b>	<b>34</b>
<b>Seznam obrázků</b>	<b>35</b>
<b>Seznam tabulek</b>	<b>36</b>
<b>Seznam použitých zkratek</b>	<b>37</b>
<b>I Přílohy</b>	<b>38</b>
<b>A Datové soubory</b>	<b>39</b>
A.1 Transformovaný datový soubor players.csv . . . . .	39
A.2 Transformovaný datový soubor results.csv . . . . .	39

# 1. Úvod

Esport je označení pro elektronický sport. Obsahuje všechny důležité oblasti jako klasický sport (např. turnaje, trénování, investice, stadiony, či sázení) s tím rozdílem, že se hraje na nějakém zařízení (počítač, konzole, mobil). Je to jedno z nejrychleji rostoucích odvětví v dnešní době. V roce 2021 se tržní hodnota esportu pohybovala kolem jedné miliardy dolarů — skoro 50% nárůst oproti roku 2020. Lze předpovídat, že v roce 2024 esport překročí hodnotu 1,5 miliardy dolarů (Gough, 2021). Dalo by se spekulovat, že za takový velký nárůst je zodpovědná **pandemie koronaviru v letech 2019 - 2022. Kombinace rozvoje počítačových her a generace, která je na práci s počítačem zvyklá od mala, vzniklo výborně prostředí pro organický růst esportu.** Většina populace **byla** nucena zůstat doma a to otevřelo dveře se s esportem přirozeně seznámit a nějakým způsobem se ho účastnit (online divák, soutěžící, organizátor, fanoušek...). Hrají se různé kategorie her jako např. Multiplayer Online Battle Arena (MOBA)<sup>1</sup>, karetní hry, First-Person Shooter (FPS) či Battle Royale (BR).

Práce se zaměřuje primárně na esportový titul Counter-Strike: Global Offensive (CSGO). Je to jeden z nejdéle hraných esportových titulů, boří mnohé divácké rekordy<sup>2</sup> a je aktuálně nejhranějším FPS esport titulem. CSGO vyniká nejen detailní herní mechanikou, ale i bohatou a zajímavou historií. Hra je unikátní i tím, že obsahuje mnoho různých módů<sup>3</sup> a hráč může strávit mnoho hodin pouze objevováním komunitních serverů, hraním klasických zápasů či trénováním na offline mapách.

Finální cíl práce je vytvořit logistický regresní model, který předpovídá výsledek zápasů. Pro tvorbu kvalitního modelu bude kritické zvolit vhodné prediktory. Pro predikci jsou použity prediktory, které se nacházejí ve dvou samostatných datových souborech<sup>4</sup>, které podávají informace jak už o zápase (např. datum, výsledek zápasu, výsledek jednotlivých map, typ zápasu), hráčích (např. statistiky za zápas, statistiky za mapy, statistiky za team), tak o vývoji celého zápasu (především ekonomika týmu). V práci bude tedy vytvořeno více specializovaných modelů pro každý vybraný tým a následně je pro každý tým vybrán nejlepší model. Výsledné modely jsou v závěru mezi sebou porovnány.

Logistický model je preferován kvůli své **snadné** interpretaci a dobré aplikaci v reálném životě. Výsledky **představených modelů se dají využít v sázkových kancelářích** a lze předpovídat, kdo vyhraje zápas, kdo vyhraje mapu, jaký hráč bude mít nejlepší statistiky, či zda si hráč koupí určitou zbraň.

---

<sup>1</sup>tzn. MOBA, kde hráči hrají v jedné online aréně proti sobě

<sup>2</sup><https://www.invenglobal.com/articles/15619/csgo-major-breaks-viewership-records-overtakes-the-international>

<sup>3</sup>rozšíření, jak hru hrát. Každý mód má svá vlastní pravidla, mapy, či herní fanoušky

<sup>4</sup><https://www.kaggle.com/mateusdmachado/csgo-professional-matches>



Práce je rozdělená do tří částí. V první části je kladen důraz na esport, jeho vývoj, a na esportový titul CSGO. Jsou zde také představená pravidla, podle kterých se hra hraje. V druhé části jsou popsány popisné a statistické metody. Jsou zde definované grafické nástroje pro popis datasetu, logistický regresní model, a evaluační nástroje pro model. Třetí část se zaměřuje na praktickou tvorbu modelů, jejich interpretaci, a vzájemné porovnání.

## 2. Představení esportu

### 2.1 Historie esportu

I přes fakt, že esport není obecně známý pojem mezi širokou veřejností, má přes 70 let bohaté historie. Za jeho počátky by se daly považovat arkádové automaty, kde hráči z počátku soutěžili sami proti sobě. Největší rozvoj arkádových automatů se děl kolem 70 let minulého století. Nejen za tímto účelem byla 9. 2. 1982 založena Twin Galaxies National Scoreboard (TGNS). TGNS měla na starosti nejen udržování výsledkové tabulky (scoreboard), ale i tvorbu prvotních pravidel pro férovou hru. Za tímto účelem byla vydána kniha Twin Galaxies' Official Video Game & Pinball Book of World Records.

Na přelomu osmdesátých let minulého století se začal esport vyvíjet již více profesionálním směrem. V roce 1972 pořádala Stanfordská Universita historicky první esportový turnaj v arkádové hře Spacewar!. Výherce si mohl odnést předplatné magazínu Rolling Stones. Dále v roce 1983 byl založen první esportový profesionální tým, který se nacházel ve Spojených státech. Všechno toto se stalo díky podnikateli Walteru Day, který je zakladatel společnosti TGNS a založil již zmíněný první esportový tým. Ač se Walter považuje za jednoho z hlavních pionýrů esportu, v roce 2010 TGNS opustil kvůli své vášni pro hudbu.

Další důležitou kapitolou ve vývoji esportu je příchod internetu a výkonných počítačů. Hráči se dostali k rychlejším sestavám, stolní počítače se stali cenově dostupnějšími a díky tomu se zpřístupnili k více lidem. Klesala cena hardwaru, vývoj nové technologie a her se zrychloval. Díky rozvoji počítačových sítí se mohli hrát LAN<sup>1</sup> party či organizovat BYOC<sup>2</sup> turnaje. Dále už esport potřeboval jen čas na organický růst a dnes má tržní hodnotu přes jednu miliardu amerických dolarů (Gough, 2021), (Larch, 2019).

### 2.2 Zasazení do dnešní doby

V dnešní době je esport téměř miliardová záležitost. Díky pandemii, která trvá již od r. 2019, si esport ještě přilepšil. Dle průzkumu<sup>3</sup> z října roku 2020 si 73 % dotázaných myslelo, že se úroveň zájmu a obchodní činnost esportu v Q4 2020 a Q1 2021 zvětší. Respondenti, kteří se průzkumu zúčastnili, jsou považováni za experty v oblasti esportu. Tento průzkum byl následně podpořen růstem že tržní hodnoty esportu a mezi lety 2019 a 2020 vzrostla o téměř 50 % (Gough, 2021).

---

<sup>1</sup>Hráči hrají v jedné místnosti na lokální počítačové síti.

<sup>2</sup>z ang. Bring Your Own Computer, kde si hráči si na akci donesou vlastní počítač

<sup>3</sup><https://www.statista.com/statistics/1247902/covid-impact-esports-investments>

K takto prudkému růstu tržní hodnoty esportu z velké části přispěla právě pandemie. Mladá generace byla nucena zůstat doma, což dovolilo i esportem nedotčeným jedincům do tohoto světa proniknout. Větší zájem o esport přinesl i větší tržby herním studiím, která začala do esportových turnajů více investovat (Professeur, 2021), (liquipedia, 2021). S větším počtem diváku roste i marketingový potenciál, investiční příležitost a kariérní růst.

Druhý dominantní žánr je FPS. V této kategorii jsou nejvýznamnější hry CSGO a Valorant. V tomto žánru proti sobě hrají dva týmy, většinou složené z pěti hráčů. Každý hráč pak má v týmu různou roli, jako např. velitel či odstřelovač. Jeden tým má obvykle za úkol něco zničit (položit bombu, unést rukojmí) a druhý tým jim v tom musí zabránit (ochránit oblast proti bombě, záchrana rukojmí).

Poslední žánr který zmíním je Battle Royale (BR). V těchto hrách hraje buď každý hráč sám za sebe, ve dvojici, nebo ve skupině po čtyřech. Zde hráči padají na začátku kola na velkou mapu. Jejich úkolem je získat vybavení, aby mohl porazit ostatní hráče a kolo sami, nebo s týmem vyhrát. Nacházejí se zde různé role, avšak trochu rozdílné oproti žánru FPS. Hlavním titulem této kategorie je hra Fortnite, která žánru dominuje. Stal se z ní jak esportový titul, tak perfektní marketingové místo pro teenagery. Hráči si zde mohou koupit oblečky různých filmových či komiksových postav. Pokud vychází nový film, ve hře se může objevit „event“ (událost), který daný film propaguje. Toto lze vidět například na propagaci Avengers: Endgame<sup>4</sup>.

## 2.3 Představení titulu Counter-Strike: Global Offensive

CSGO, jak ho známe dnes, má bohatou a dlouhou historii. Ne vždy se to ovšem jmenovalo stejně. Úplně první iterace hry se jmenovala čistě Counter-Strike a byl to pouze mód<sup>5</sup> do hry Half-Life. Half-Life byl vyvinutí společností Valve, tehdy primárně společností zaměřenou na vývoj her. Mód byl vytvořen studenty vysoké školy, panem Minh Le a Jess Cliffe. Toto rozšíření začali programovat v roce 1999. Jelikož mód byl neoficiálním rozšířením, Valve o něj neprojevovalo veliký zájem. Až po pěti betaverzích hry Counter-Strike si společnost Valve všimla rozšíření, její komunity, ale především jejich autorů. Minh a Jess se v roce 2000 stali oficiálními zaměstnanci Valve a duševní vlastnictví módu prodali. Autoři, nově jako zaměstnanci Valve, roku 2000 vydávají první oficiální verzi hry Counter-Strike. I přes toto „oficiální“ datum vydání je většina komunity přesvědčena, že výročí má CSGO v den svého úplně první vydání, a to 18. června 1999.

<sup>4</sup>Trailer pro propagaci události: [https://www.youtube.com/watch?v=TanGK9o\\_d24](https://www.youtube.com/watch?v=TanGK9o_d24)

<sup>5</sup>upravení či rozšíření hry

Hra je z žánru FPS a hraje se primárně online proti skutečným hráčům. Counter-Strike se v herní komunitě rychle rozrostl díky své jednoduchosti. Hra se dá velmi dobře popsat pořekadlem „Lehké hrát, těžké vypilovat“. Hra má mechaniky<sup>6</sup>, které jsou lehké na pochopení, ale velmi těžké na vypilování k dokonalosti. Spolu s touto vlastností je hra vlastně velmi jednoduchá a hráč hraje buď za policisty, nebo za teroristy. Hráči tak mohli, a stále mohou, hru velmi lehce a rychle začít hrát, jelikož se tento formát od roku 2000 nijak extrémně nezměnil.

Hra tedy rostla zejména díky své komunitě. Hráči hru různě upravovali, přidávali další módy, typy her, zbraně, mapy či audiovizuální obsah. Tento trend se přenášel přes mnoho různých verzí hry. První velký „průlom“ udělala verze 1.6, tedy Counter-Strike 1.6. Ta vynikala jak esportem, tak komunitním obsahem. Jen v České a Slovenské republice bylo několik herních serverů, na kterých se mohlo sejít sta tisíce hráčů. Např. na česko-slovenském herním portálu kotelná hrálo celkem přes 1,5 milionu unikátních hráčů (csko, 2021). Hra byla populární nejen mezi obyčejnými hráči, ale i profesionály.

Counter-Strike 1.6 je pionýrem esportu pro FPS žánr. Za podpory Valve se hráli první major<sup>7</sup> turnaje, kde hráči mohli ukázat svůj um za tehdy relativně velkou sumu peněz. Hra se časem vyvíjela, hráči nalézali nové strategie či triky a Valve vydalo novou verzi — Counter-Strike: Source. Tato nová verze získala nepříliš pozitivní ohlas, jelikož velmi rozdělila herní komunitu. Představila nové mechaniky, staré mechaniky změnila a hráčům, zejména v esportu, se nechtělo učit něco úplně nového. Valve se rozhodlo sjednotit herní komunitu, a proto vydalo novou verzi hry s názvem CSGO

CSGO se snažilo sjednotit oba tábory z her Counter-Strike 1.6 a Counter-Strike: Source. Hra vyšla 21. srpna 2012 a z počátku nebyla tolik úspěšná, ale díky přidání různých skinů (Valve, 2013) na zbraně hra přilákala úplně nové publikum. Díky novému a velkému publiku se začali hrát menší esportové turnaje právě ve hře CSGO, ke kterým se později přidali i profesionálové z předchozích dvou verzí. Díky tomuto organickému růstu má Counter-Strike velmi silnou komunitu, která se o hru i nadále stará. I přes netradiční interakci mezi Valve a herní komunitou hra stále roste. CSGO se díky své dlouhé historii, bohaté komunitě a různým možnostem, jak hru hrát, dostala na špičku esportu. I přes několik titulů, které se s hrou snaží soutěžit, je hra stále největším a nejsledovanějším esport titulem v rámci FPS žánru (Henningson, 2020).

---

<sup>6</sup>herní prvky či unikátní vlastnosti

<sup>7</sup>turnaj pořádaný přímo Valve, který má největší prestiž

## 2.4 Propojení práce a titulu Counter-Strike: Global Offensive

Práce se zaměřuje na identifikování významných prediktorů a následně vytvoření regresního modelu. Před jakoukoliv prací s daty je ale nutné pochopit, jak se hra vlastně hraje a jaká jsou její pravidla. Ve hře CSGO hraje pět hráčů proti pěti (dále jen 5v5). Hra se většinou hraje online, avšak velké esportové turnaje se hrají offline, tedy v nějaké např. aréně. Hra má v základu 30 kol a po prvních patnácti se mění strany. Jedna strana jsou policisté (Counter-Terrorists či CT), kteří mají za úkol chránit „bomboviště“ - část mapy, která má vybuchnout. Naopak cíl Teroristů (T) je právě bombu položit a „bomboviště“ nechat vybuchnout. Vyhrává tým, který první vyhraje 16 kol. Pokud ovšem po první 30 kolech je stav nerozhodný, tedy 15:15, hraje se prodloužení. Tento formát není standardizovaný pro všechny turnaje, proto zmíním pouze pravidla, která se týkají turnajů od společnosti Valve (již zmíněné a nejvíc prestižní Majory). Zde se hraje prodloužení ve formát Bo6, tedy kdo první získá 4 body, vyhraje zápas. Takto může jít zápas teoreticky do nekonečna. Nejdelší semi-profesionální zápas, který se ovšem neodehrál na Majoru, se stal mezi týmem exCeL a XENEX(hltv.org, 2015). Zápas pokračoval do úctyhodných 88 kol.

V každém kole má tým určitý počet peněz. Každá hráč začíná polovinu (ted v první a šestnácté kolo) s \$800. Finance každého hráče pak záleží na mnoha faktorech, jako kolik vyhrál jeho tým kol v řadě, kolik nakoupil zbraní, kolik zabil nepřátel, kolik peněz dostane hráč za zabití či jak kolo skončí. V profesionálním týmu je velmi obtížné pracovat s financemi, jelikož všichni musí být v tomto ohledu jednotní. V tuto chvíli přichází na řadu tzn. In-Game Leader (velitel týmu). Tuto roli má většinou jeden hráč v každém týmu. Je to ta nejdůležitější role ze všech. Má na starosti např. finance týmu, rozhoduje kdy se koupí a kdy půjde tzn. eco (hráči nekoupí nic, aby ušetřili peníze), jaké se budou hrát mapy či jaká se půjde v daném kole strategie. V dnešní době k tomu In-Game Leader má i trenéra. Ten hru nehraje, ale pozoruje hráče a dává jim různé typy a triky.

Role trenéra není nijak silně definovaná a každý esportový tým má trochu jiného trenéra. V jednom případě může být trenér čistě jako podpora a pomáhá hráčům když se nedaří a řeší interní problémy. V jiném týmu může ovšem mít velký zásah do hry, pomáhat In-Game Leaderovi se strategiemi, obelstění soupeře či sledováním předchozích zápasů pro kontinuální zlepšování týmu. Další role v týmu jsou například Entry Fragger (má za úkol získat první zabití pro tým), support (podporuje svůj tým za pomoci různých granátů nebo se často pro svůj tým obětuje), AWP hráč (hráč je specifický tím, že hraje primárně s jednou zbraní) a Lurker (chodí po mapě sám a snaží se nepřítele odchytnout ze stran, které by nečekali)

Zápasy se pak hrají ve formátech „Best of“. Best of 3 například znamená, že se hrají tři mapy. Kdo první vyhraje dvě mapy, vyhrál celý zápas. Turnaje se pak odehrávají v tradičních formátech, jako je pavouk. Ten se charakterizuje tím, že vypadá jak pavučina, jde zleva doprava a každý tým může prohrát pouze jednou. Následně tu máme Upper/Lower bracket formát, který je v podstatě pavoučí formát, akorát jsou zde dvě „sítě“ a každý tým může prohrát maximálně jednou, jelikož druhá prohra znamená vyřazení z turnaje. Specifičtější formát pro CSGO je například swiss, který se počítá přes různé body a statistiky výsledných zápasů.

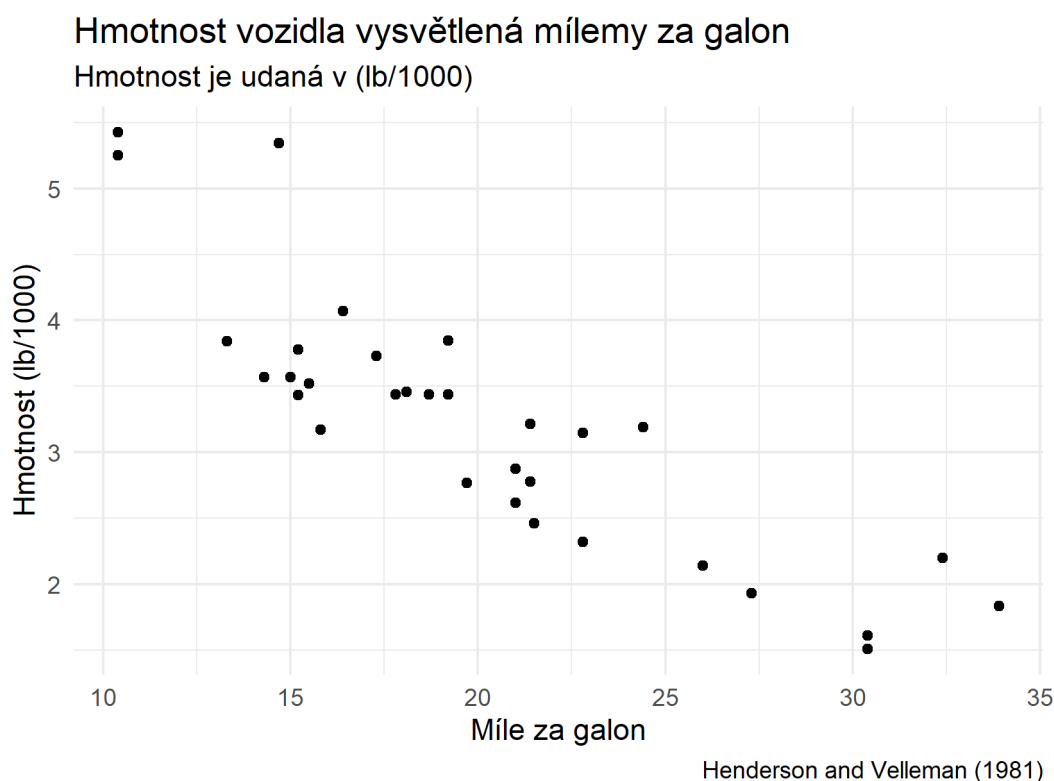
## 3. Teoretická část

V následující části jsou popsány jak teoretické metody pro vizualizaci dat, tak i tvar, forma a vyhodnocení logistického regresního modelu. Ke každé části, která se věnuje popisu dat pomocí nějakého grafu, je přidána praktická ukázka s popisem a praktickým vysvětlením. vhodné.

### 3.1 Vizualizace dat

#### 3.1.1 Bodový graf

Bodový graf slouží pro zobrazení vztahu dvou kvantitativních proměnných. Z pravidla se vysvětlovaná proměnná dává na osu Y, zatímco proměnná vysvětlující se nachází na ose X. Vysvětlovaná (nezávislá) proměnná je ta proměnná, která má být předvídaná. Vysvětlující proměnná se naopak snaží vysvětlovanou proměnnou předpovědět či popsat. **Zobrazením** vysvětlované a vysvětlující proměnné na bodovém grafu lze vidět např. sílu korelace nebo vztah mezi proměnnými (např. lineární, kvadratický, logaritmický).



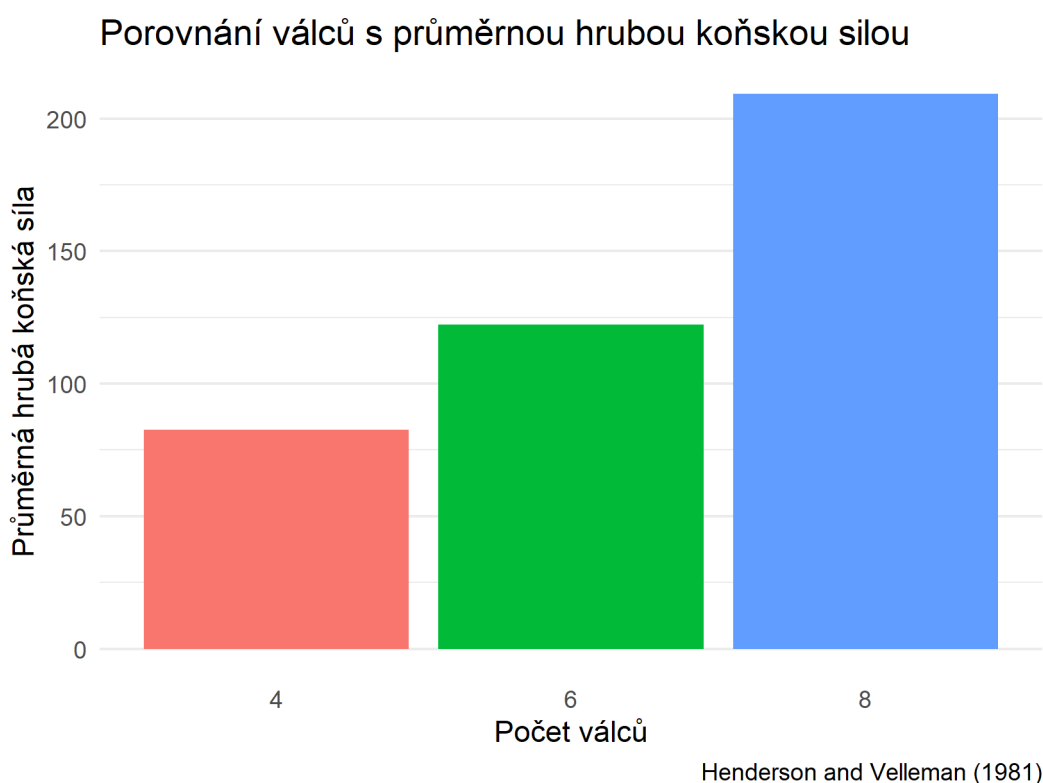
Obrázek 3.1: Bodový graf hmotnosti a míly za galon

Obrázek 3.1 zobrazuje negativní korelaci mezi hmotností vozidla a mílemi ujetými za galon.

### 3.1.2 Sloupcový graf

Sloupcový graf slouží k zobrazení **četností** kategorií. Na jednu osu (z pravidla osu X) se položí možné kategorie. Na druhou osu se pak položí sledovaná statistika. Sledovat můžeme nejen četnost, ale i průměr či kteroukoli jinou statistiku, kterou bude možné na ose zobrazit.

Nejčastěji se pomocí sloupcového grafu porovnává absolutní četnost přes kategorie. Řazení kategorií se dále odvíjí podle toho, zda je daná proměnná ordinální či nominální. V případě nominální proměnné se sloupce řadí podle absolutní četnosti, **a to od nejvyšší po nejnižší**. V případě ordinální proměnné se zachovává přirozené řazení. Příklad sloupcového grafu je zobrazen na obrázku 3.2, který porovnává průměrnou hrubou koňskou silou s počtem válců. Je na něm také vidět závislost hrubé koňské síly na počtu válců.

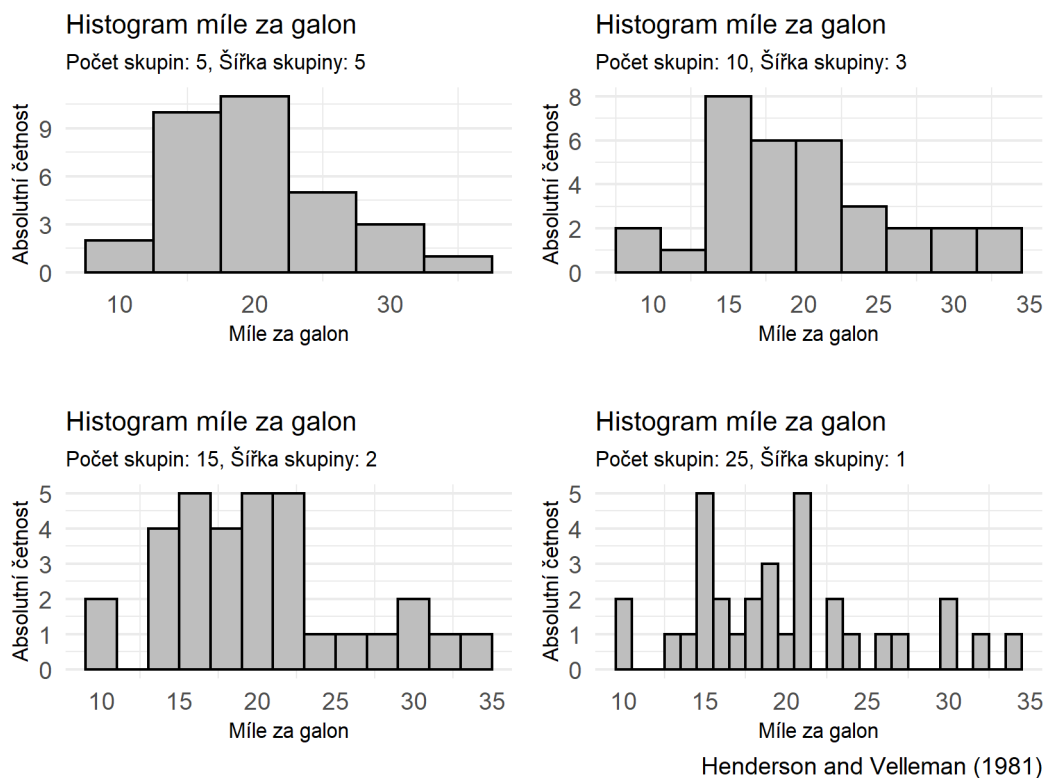


Obrázek 3.2: Sloupcový graf počtu válců a průměrné hrubé koňské síly

### 3.1.3 Histogram

Histogram je speciální typ sloupcového grafu. Hlavní rozdíl je v tom, že popisuje rozdělení spojité proměnné a mezi sloupci není žádná mezera. Pro histogram je třeba data sloučit do skupin o určité šířce. Správný výběr počtu skupin je kritický, jelikož může velmi silně ovlivnit interpretaci dat. Pokud se vybere příliš malý počet skupin, data se seskupí a může se ztratit důležitý vztah. Pokud se ovšem vybere moc velký počet skupin, v datech bude obtížné najít nějaký obecný vztah či trend, viz **obrázek 3.3**





Obrázek 3.3: Porovnání histogramů s různým počtem skupin

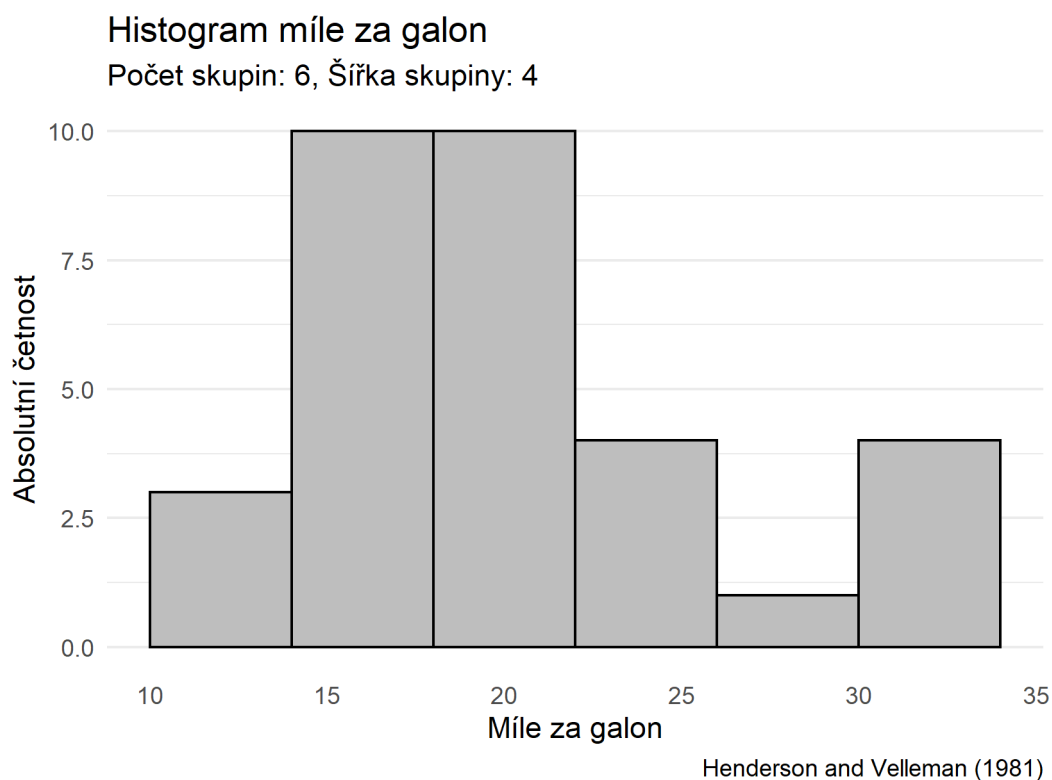
Pro vhodný počet skupin existuje mnoho způsobů. Nejznámější je takzvané Sturgesovo pravidlo, které se spočítá následujícím vztahem:

$$k \doteq 1 + 3,3 * \log_{10}(n) \quad (3.1)$$

kde  $k$  je přibližný počet skupin a  $n$  je počet pozorování. Druhý parametr, který je pro tvorbu histogramu potřeba, je šířka skupiny. Ta by měla být ideálně stejná pro všechny skupiny. Pokud tomu tak není, histogram může být zavádějící a čtenář mu nemusí plně rozumět. Pro vypočtení počtu skupin má šířka skupiny následující tvar:

$$w = \frac{\max(x) - \min(x)}{k} \quad (3.2)$$

kde  $x$  je zobrazovaná proměnná,  $k$  je počet skupin a  $w$  je výsledná šířka intervalu. Nutné je však podotknout, že není pravidlem se danými výpočty řídit a výsledný sloupcový graf je nutné přizpůsobit **konkrétnímu datovému souboru**.



Obrázek 3.4: Histogram s počtem skupin dle Sturgesova pravidla

### 3.1.4 Krabičkový graf

#### Five-number summary

Five-number summary je číselná tabulka, která pomocí pěti různých čísel shrnuje seřazenou číselnou řadu. Základní statistický nástroj pro vytvoření takové tabulky jsou kvantily. Hodnota  $P$ -tého percentilu označuje číslo, které rozděluje seřazenou číselnou řadu na dva intervaly. První interval obsahuje  $P * 100\%$  číselné řady a druhý analogicky  $(1 - P) * 100\%$ . Různé hodnoty percentilů mohou mít specifitější pojmenování a značí se  $Q_P$ . Percentil  $P = 0,5$  se označuje jako medián a rozděluje seřazenou číselnou řadu na polovinu. Percentily, kde  $P = 0,25$  nebo  $P = 0,75$ , se označují jako kvartily a značí se  $Q_1$  a  $Q_3$ . Oba tyto typy kvartilů jsou použité při tvorbě Five-number summary tabulky. Jako příklad je uvedena následující tabulka

$Q_0(Q_0)$	$Q_{0,25}(Q_1)$	$Q_{0,50}$	$Q_{0,75}(Q_3)$	$Q_{1,00}$
1,513	2,58125	3,325	3,61	5,424

Tabulka 3.1: Five-number summary tabulka hmotnosti vozidla (lb/1000)

kde  $Q_0$  a  $Q_{1,00}$  označují minimum a maximum číselné řady. Kvartily  $Q_1$ ,  $Q_2$  (medián) a  $Q_3$  jsou čísla, která rozdělují časovou řadu na čtvrtiny. V prvním případě, tedy  $Q_1 = Q_{0,25}$ , je 25% čísel menší než 1,513 a 75% dat větší. Pro kvantil  $Q_3 = Q_{0,75}$  je 75% čísel menších než 3,61 a 25% větších.  $Q_{0,50}$  označuje medián.

## Krabičkový graf

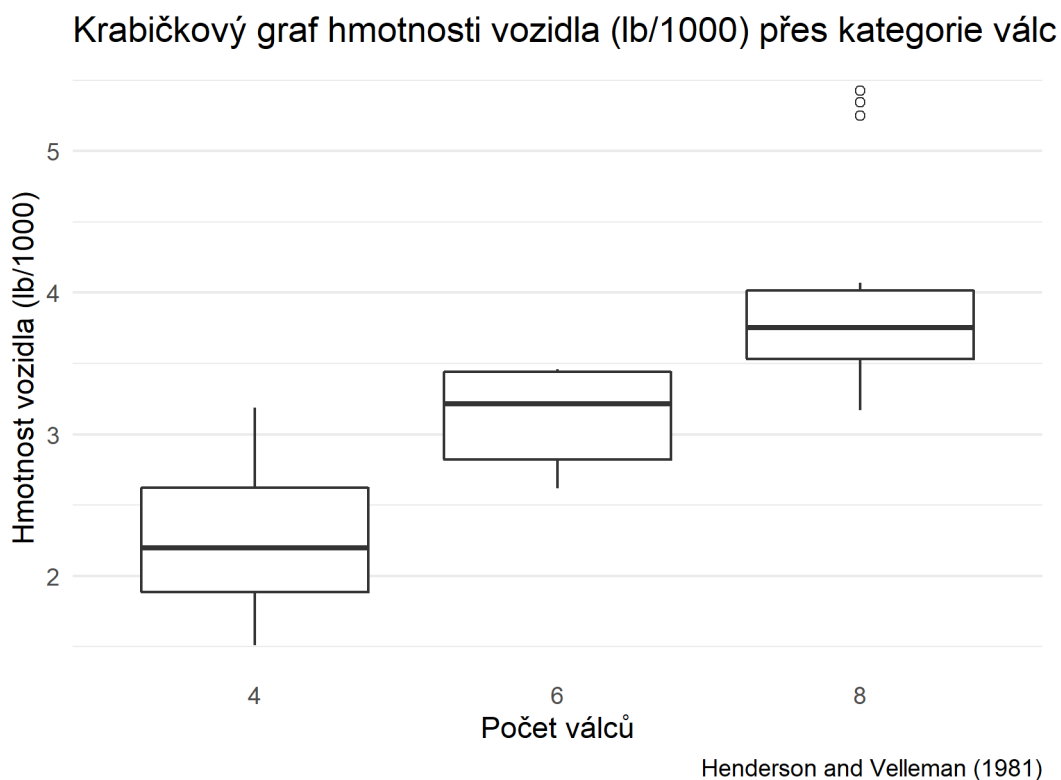
Krabičkový graf je grafické zobrazení a rozšíření Five-number summary tabulky. Kromě grafického zobrazení pěti kvantilů ukazuje odlehlé a extrémní hodnoty. V Krabičkovém grafu se také nachází obdélník, který ukazuje mezikvartilové rozpětí (IQR), tedy prostředních 50 % dat. V obdélníku se také nachází černá čára, která značí medián. Z prostředního obdélníku vedou oběma směry čáry, jejichž konce značí hranici pro odlehlá pozorování. Pokud datový soubor neobsahuje žádná odlehlá pozorování, konec těchto čar značí minimum a maximum datového souboru. Pozorování, která jsou buď větší než horní hranice, nebo menší než spodní hranice, označujeme jako **odlehlá** nebo extrémní.

$$\text{Spodní hranice} = Q_1 - 1,5 * IQR \quad (3.3)$$

$$\text{Horní hranice} = Q_3 + 1,5 * IQR \quad (3.4)$$

Hodnoty, které spadají do intervalu  $\langle Q_1 - 1,5IQR; Q_1 - 3IQR \rangle$  a  $\langle Q_3 + 1,5IQR, Q_3 + 3IQR \rangle$  se nazývají jako odlehlé. Hodnoty které leží mimo tento vztah, tedy hodnoty menší než  $Q_1 - 3IQR$  nebo větší než  $Q_3 + 3IQR$  se nazývají jako hodnoty extrémní. **Odehlá pozorování se v krabičkovém grafu většinou značí kolečkem, zatím co pozorování extrémní hvězdičkou.**

Díky grafickému zobrazení lze lehce porovnávat rozdělení jedné vysvětlované kvantitativní proměnné tříděné přes několik kategorií.



Obrázek 3.5: Krabičkový graf hmotnosti auta pro různý počet válců

Průhledná kolečka v obrázku 3.5 v kategorii osmi válců značí odlehlé hodnoty, t.j. hodnoty v intervalu  $\langle Q_3 + 1,5IQR, Q_3 + 3IQR \rangle$ .

### 3.1.5 Korelační matice

#### Korelace

Korelace popisuje směr a sílu vztahu mezi dvěma proměnnými  $X$  a  $Y$ . Značí se  $r$  a nabývá hodnot  $\langle -1; 1 \rangle$ .

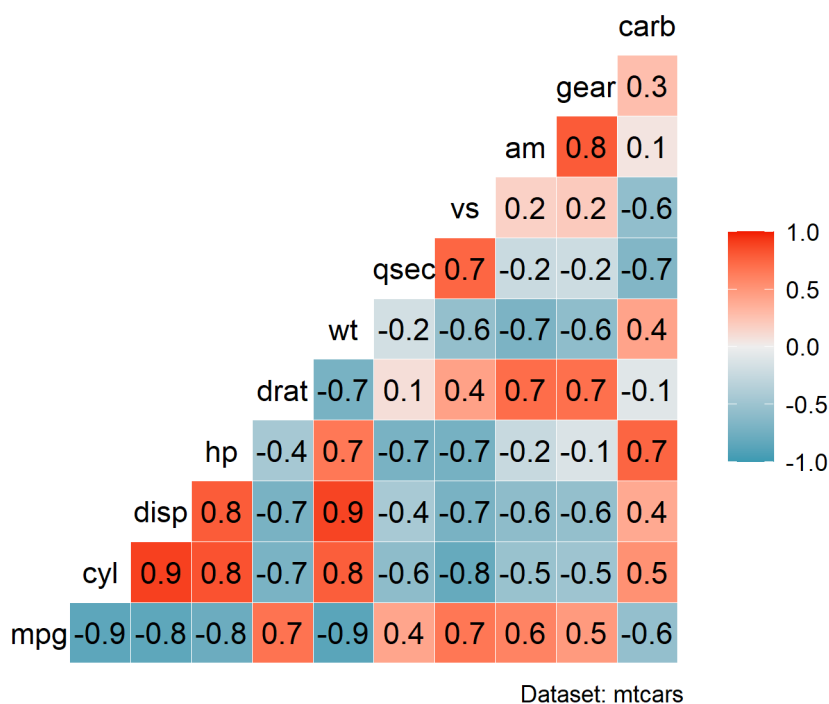
$$r = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{(\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n})(\sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n})}} \quad (3.5)$$

Čím větší je absolutní hodnota korelace mezi proměnnou  $X$  a  $Y$ , tím lépe lze pomocí jedné proměnné vysvětlit proměnnou druhou. Kladnost, případně zápornost korelace značí pak směr vztahu. Pokud je korelace záporná, tedy  $r < 0$ , s růstem jedné proměnné klesá proměnná druhá. Naopak při kladné korelace, tedy  $r > 0$ , s růstem jedné proměnné roste i druhá. Pokud se korelace  $r$  vychází kolem nuly, neexistuje lineární závislost mezi proměnnou  $X$  a  $Y$ . Důležité je také podotknout, že korelace neznamená kauzalitu. Pokud existuje kladná korelace mezi proměnnou  $X$  a  $Y$ , neznamená to, že růst jedné proměnné způsobil růst druhé proměnné.

## Korelační matice

Korelační matice je nástroj, díky kterému lze zobrazit korelaci mezi více jak dvěma páry proměnných. Matice může být zobrazená pomocí grafu. Korelační matice je velmi užitečná v regresní analýze kvůli předpokladu nezávislosti vysvětlujících proměnných v analýze. Pokud jsou při tvorbě modelu prediktory korelované, vzniká problém tzn. multikolinearity. Při multikolinearitě se zhoršuje přesnost a vypovídací hodnota koeficientů (Kleinbaum, 2010). V takovém případě je potřeba zvýšit počet pozorování nebo z modelu určité prediktory odebrat.

### Korelace mezi kvantitativními proměnnými



Obrázek 3.6: Korelační matice

Graf korelační matice může mít mnoho podob. V příkladu obrázku 3.6 je zobrazená korelační matice jako teplotní mapa. Z obrázku je možné pozorovat vysokou pozitivní korelaci mezi páry proměnných *cyl*, *disp* a *hp*. Naopak skoro žádná korelace není mezi proměnnou *qsec* a proměnnou *drat*. Korelační matice je zároveň symetrická, jelikož korelace mezi  $X$  a  $Y$  je stejná jako korelace mezi  $Y$  a  $X$ . Díky této vlastnosti je možné zobrazit pouze část korelační matice pod úhlopříčkou bez ztráty jakékoliv informace.

## 3.2 Logistická regrese

Logistická regrese je způsob, jak popsat vztah mezi jedním či několika prediktory a jednou binární vysvětlovanou proměnnou. K tomu slouží spojovací funkce, která transformuje lineární kombinaci prediktorů na index  $z$ . V případě logistické regrese se tato funkce nazývá logistická a je definovaná jako

$$f(z) = \frac{1}{1 + e^{-z}}. \quad (3.6)$$

Obor hodnot funkce je interval  $\langle 0, 1 \rangle$ . Proměnná  $z$  je lineární kombinace prediktorů  $X_1, X_2, \dots, X_k$ , jejich koeficientů  $\beta_1, \beta_2, \dots, \beta_k$  a parametru  $\alpha$ .

$$\begin{aligned} z &= \alpha + \beta_1 X_1 + \dots + \beta_2 X_2 + \beta_k X_k \\ &= \alpha + \sum_{i=1}^k \beta_i X_i \end{aligned} \quad (3.7)$$

Mějme tedy binární vysvětlovanou proměnnou  $Y$ , u které hodnota 1 značí výskyt jevu. Pravděpodobnost, že jev nastane vzhledem k definovaným prediktorům lze zapsat jako

$$P(Y = 1 \mid X_1, X_2, \dots, X_k) = \frac{1}{1 + e^{-\left(\alpha + \sum_{i=1}^k \beta_i X_i\right)}}, \quad (3.8)$$

kde  $\alpha$  a  $\beta_i$  jsou parametry odhadnuté z datového souboru. (Kleinbaum, 2010)

### 3.2.1 Interpretace parametrů

Parametry  $\alpha$  a  $\beta_i$  značí logaritmus šance.  $\alpha$  je logaritmus šance v případě, že všechny prediktory jsou teoreticky rovné 0. Parametr  $\beta_i$  značí logaritmus šance pro prediktor  $X_i$ . V případě, že všechny prediktory jsou konstantní a prediktor  $X_i$  se změní o jednotku, přirozený logaritmus šance se změní o  $\beta_i$ . Toto lze pozorovat například u binárních prediktorů, kdy typicky přítomnost daného prediktoru, značená jedničkou, změní výslednou šanci právě o odhadnutý parametr  $\beta$ . Pro přechod z přirozeného logaritmu šance na šanci lze využít vztahu

$$\text{šance} = e^{\beta_i}. \quad (3.9)$$

Šance je podíl dvou pravděpodobností. Pokud bychom měli šanci jevu A oproti jevu B 2 : 1, značí to, že výskyt jevu A je dvakrát tak pravděpodobný jako výskyt jevu B a jev A se vyskytuje ve  $\frac{2}{3}$  případů. Šance  $e^{\beta_i}$  tedy značí vztah mezi prediktorem  $X_i$  a vysvětlovanou proměnnou  $Y$ . Pokud je šance kladná, značí to, že s vyšší hodnotou prediktoru  $X_i$  se zvyšuje šance že  $P(Y = 1)$ . Pokud je naopak nižší, pravděpodobnost se zmenšuje. Pokud je potřeba interpretovat pravděpodobnost jako šanci, použije se logitová funkce

$$\text{šance jevu } A = \frac{p}{1 - p} \quad (3.10)$$

kde  $p$  je pravděpodobnost výskytu jevu A.

### 3.2.2 Největší pravděpodobnost

Parametry logistického modelu v rovnici 3.8 jsou pouze teoretické a je třeba je odhadnout. Již vypočtené odhady se proto neznačí pouze  $\beta$ , ale  $\hat{\beta}$ . Pro odhad parametrů se při logistické regresi používá metoda maximální věrohodnosti. Pro výpočet maximální věrohodnosti se počítá věrohodnostní funkce  $L(\theta)$  kde  $\theta$  jsou parametry logistického modelu  $\alpha, \beta_1, \dots, \beta_k$ . Pro logistickou regresi má věrohodnostní funkce tvar

$$L(\theta) = \prod_{i=1}^{m_1} P(X_i) \prod_{i=m_1+1}^n 1 - P(X_i), \quad (3.11)$$

kde  $n$  je počet pozorování a  $m_1$  je počet příznivých ( $Y = 1$ ) jevů. Funkce předpokládá, že datový soubor je seřazen tak, že prvních  $m_1$  výskytů jsou jevy příznivé.  $P(X_i)$  poté značí logistickou funkci 3.6. Pro vypočtení optimálního parametru  $\beta_i$  je nutné vypočítat maximum funkce  $L(\theta)$  vzhledem k parametru  $\beta_i$ . Parametr  $\beta_i$  lze tedy získat derivací funkce  $L(\theta)$  vzhledem k parametru  $\beta_i$ . (Kleinbaum, 2010)

$$\frac{\partial L(\theta)}{\partial \beta_i} = 0 \quad (3.12)$$



### 3.2.3 Křížová validace

Při **tvorbě** logistické modelu může dojít k takzvanému přeučení modelu. To znamená, že výsledný model je velmi přizpůsobený na data, ze kterých byl vytvořen, a nebude připravený na náhodnou variaci, která může v nových datech nastat. Z tohoto důvodu se datový soubor rozděluje na dvě podmnožiny. Jedna podmnožina, většinou zvaná *trénovací*, slouží k sestavení a natrénování modelu. **Model se pak otestuje** na druhé množině dat, na kterou nebyl natrénován. Druhá množina se většinou nazývá *validační* nebo *testovací*. **Způsobů, jak datový soubor** rozdělit, je mnoho. Existuje například *k-fold* validace, kdy se trénovací množina dat rozdělí na *k* podmnožin. Pro každou podmnožinu se pak vytvoří model, který se následně validuje na validační množině dat. Model s nejlepším výkonem je následně vybrán a použit.

### 3.2.4 Matice záměn

Matice záměn je nástroj pro vyhodnocení predikcí modelu. Matice je o velikosti  $2 \times 2$ . Pro potřeby logistické regrese se matice skládá ze dvou řádků a dvou sloupců. **Ve sloupcích se nachází původní hodnoty, tedy hodnoty, které chceme předpovídat. Ve řádcích se pak nachází předpovědi z modelu.**

		Původní pozitivní	Původní negativní
		1	0
Pozitivní predikce	1	Skutečně pozitivní	Falešně pozitivní
Negativní predikce	0	Falešně negativní	Skutečně negativní

Tabulka 3.2: Matice záměn

Pro sestavení matice je potřeba množina dat, u kterých známe vysvětlovanou proměnnou. Na datech pak provedeme predikci, díky čemuž získáme predikované hodnoty. Porovnáním původních a predikovaných hodnot vznikne matice 3.2. Každá ze čtyř vnitřních buněk má vlastní označení a interpretaci:

- **Skutečně pozitivní** — počet správných predikcí, které byly rovné jedné.
- **Falešně pozitivní** — počet predikcí rovných jedné, kde byla původní hodnota rovná nule.
- **Skutečně negativní** — počet správných predikcí, které byly rovné nule.
- **Falešně negativní** — počet predikcí rovných nule, kde byla původní hodnota rovná jedné.

Z matice lze následně vypočítat mnoho statistik. Pro vyhodnocení regresního modelu lze použít např. přesnost, která se vypočítá jako počet všech správných predikcí nad počtem všech provedených predikcí *n*.

$$Přesnost = \frac{\text{Skutečně pozitivní} + \text{Skutečně negativní}}{n} \quad (3.13)$$

Přesnost říká, jaké procento **objektů** bylo klasifikováno správně. Pokud je ovšem poměr původních pozitivních a negativních hodnot velmi nevyrovnaný, tato statistika není vhodná. Toto se může stát například v lékařství při identifikaci nemocného pacienta. Zde většinou dochází k velkému nepoměru mezi počtem nemocných a počtem zdravých. V takovém případě lze použít statistiku zvanou senzitivita. Ta se rovná poměru správných pozitivních predikcí a úhrnu všech pozitivních predikcí

$$Senzitivita = \frac{\text{Skutečně pozitivní}}{\text{Skutečně pozitivní} + \text{Falešně pozitivní}} \quad (3.14)$$

Senzitivita tedy určuje poměr správně klasifikovaných pozitivních případů a všech pozitivně klasifikovaných případů. Pokud by bylo vhodné preferovat spíše negativní klasifikace, tedy zdravé pacienty, lze použít statistiku zvanou specifická.

$$Specifická = \frac{\text{Skutečně negativní}}{\text{Skutečně negativní} + \text{Falešně negativní}} \quad (3.15)$$

### 3.2.5 Testování hypotéz

...

### 3.2.6 Waldův test

Waldův test ověřuje, zda je parametr  $\beta_i$  v populaci významný či nikoliv. Definice testu hypotézy je tedy:

$H_0$  : Koeficient  $\beta_i$  je rovný nule.

$H_A$  : Koeficient  $\beta_i$  je různý od nuly.

Pro vyhodnocení hypotézy se používá **testové kritérium**  $Z$ , který se vypočítá jako poměr testovaného parametru  $\beta_i$  a směrodatné chyby koeficientu  $S_{\hat{\beta}_i}$

$$Z = \frac{\hat{\beta}_i}{S_{\hat{\beta}_i}}. \quad (3.16)$$

**Testové kritérium**  $Z$  má za platnosti nulové hypotézy normální  $N(0, 1)$  a její mocnina,  $Z^2$ , má chi-kvadrát rozdělení s jedním stupněm volnosti (Kleinbaum, 2010).

## 4. Praktická část

Cílem praktické části je prozkoumat dva typy modelů. První typ modelu bude predikovat výhru zápasu pro jednotlivé hráče přes kategorie map. Hlavním cílem modelu bude identifikovat významné charakteristiky hráčů na vybraných mapách a následně rozdíly interpretovat. V druhém modelu bude zkoumaná výhra týmu v zápase. V tomto typu modelu se jako prediktory použijí agregované charakteristiky hráčů jednotlivých týmů. Pro agregaci bude použit buď aritmetický nebo geometrický průměr.

V této části se dále nachází popis dat, transformace dat, a tvorba logických regresních modelů. Nejprve jsou představené datové soubory, se kterými se pracuje. Následně jsou použité grafy, které jsou představené v sekci 3.1. Pomocí vizualizace dat lze představit proměnné, které do logistického modelu budou vstupovat. Také lze díky grafům získat povědomí, jak datový soubor vypadá a jaké mají proměnné rozdělení. Poté jsou vytvořené logistické regresní modely, jejich výstup je interpretován a různé modely jsou mezi sebou porovnány.

### 4.1 Cíle analýzy

Cílem analýzy je vytvořit logistické modely pro předpověď výhry jak pro individuální hráče, tak pro tým. Jelikož se charakteristiky hráčů i zápasů mohou měnit dle mapy, na které se zápas odehrává, všechny modely budou tříděné přes kategorie map.

Pro hráče budou sestaveny logistické modely z celého datového souboru přes kategorie map. Modely budou následně porovnány pomocí matice záměn a bude porovnána významnost jejich parametrů. Modely budou porovnané na vybraných dvou mapách, a to Mirage a Vertigo. Mirage je klasická mapa, která je ve hře od jeho vydání. Vertigo je naopak nejnovější přírůstek do profesionální scény a hráči, v době extrakce datového souboru, mapu ještě plně strategicky neznali. Modely budou porovnány podle přesnosti predikce a podle různé významnosti prediktorů.

Modely pro týmy budou sestavené prve z celého datového souboru a následně pro dva vybrané týmy. První tým, pro který se vytvoří model, bude tým Astralis. Ten je považován za velmi profesionální a v době extrakce datového souboru patří mezi ty nejlepší týmy na světě. Druhým týmem bude český celek eSuba. eSuba je jeden z nejdéle hrajících celků v České republice. Modely jsou vytvořené pouze pro dva týmy z toho důvodu, že je celkový počet týmů velmi vysoký a bylo by komplikované porovnávat všechny týmy naráz. Z toho důvodu se budou hledat významné rozdíly mezi dvěma referenčními modely a jedním celkovým. Mezi modely bude porovnána jak přesnost predikce, tak významnost parametrů. Jelikož je původní spojený datový soubor na úrovni hráčů, charakteristiky hráčů se musí agregovat na úroveň týmů a zápasů. Pro tuto agregaci bude použit aritmetický nebo geometrický průměr.

## 4.2 Datové soubory

Dataset<sup>1</sup> obsahuje čtyři **datové soubory**, které popisují zápasy ve hře CSGO. K potřebám této bakalářské práce budou použity pouze soubory *players.csv* a *results.csv*. Datové soubory jsou následně spojeny do jedné tabulky, která obsahuje charakteristiky všech hráčů v právě jednom zápase, potřebné informace o zápase a výsledek (zda hráč zápas vyhrál či nikoliv). Data jsou následně dle potřeby sjednocena pro týmy v zápase. Charakteristiky se agregují pomocí aritmetického nebo geometrického průměru. Zbylé dva soubory obsahují informace, které jsou již z probíhajících zápasů a z volby map. Tyto informace pro predikci výhry ještě před začátkem zápasu nelze využít. Žádný z těchto dvou souborů proto v práci není použit.

### 4.2.1 soubor players.csv

**Datový soubor** *players.csv* obsahuje **charakteristiky** jednotlivých hráčů v daném zápase. Původní datový soubor obsahuje 101 proměnných a 379 680 pozorování. V původním datovém souboru se jeden řádek (pozorování) rovná statistikám jednoho hráče za celý zápas, **který se může odehrávat až na třech mapách**. Pro potřeby bakalářské práce je tak nutné získat statistiky hráčů za jednotlivé mapy. Proto je původní datový soubor transformován do podoby, kde se jedno pozorování rovná statistikám právě jednoho hráče na právě jedné mapě, a to bez ohledu, kolik map se v daném zápase hrálo. Jinak řečeno, transformovaný datový soubor nebere v potaz, zda se daná mapa hrála jako první, druhá, či třetí. Transformovaný dataset má 10 proměnných a 643 620 pozorování. Příklad jednotlivých pozorování v transformovaném datovém souboru je v příložené tabulce A.1.

---

<sup>1</sup><https://www.kaggle.com/datasets/mateusdmachado/csgo-professional-matches>

Transformovaný **datový soubor** má 10 proměnných. Interpretace je následující:

- **match\_id** — identifikátor zápasu
- **player\_id** — identifikátor hráče
- **team** — jméno týmu
- **map** — název hrané mapy
- **kills** — počet zabití hráče v zápase na dané mapě
- **assists** — počet asistencí hráče v zápase na dané mapě
- **deaths** — počet smrtí hráče v zápase na dané mapě
- **hs** — procento zabití, které lze označit jako headshot<sup>2</sup>
- **fkdiff** — rozdíl, kolikrát hráč zabil jako první nepřítele versus kolikrát byl zabit jako první
- **rating** — shrnutí více statistik za jeden zápas do jednoho ukazatele výkonu<sup>3</sup>

#### 4.2.2 soubor results.csv

Druhý datový soubor, který je pro analýzu použit, obsahuje výsledky daných zápasů. Soubor se původně skládá z 45 773 záznamů a 19 proměnných. Dataset obsahuje na rozdíl od datového souboru *players.csv* jedno chybné pozorování. Dle něho hrál tým sám proti sobě, což nedává smysl. Jelikož je zápas na webovém portálu zadán správně, nejspíše se jedná o neznámou chybu, která nastala při exportu dat z webového portálu.

Po transformacích vznikne tabulka o 8 proměnných a 91 436 záznamech. Každé pozorování identifikuje výsledek jednoho týmu v jednom zápase na jedné mapě. Příklad je zobrazen v příložené tabulce A.2. Jednotlivé proměnné lze interpretovat následovně:

- **date** — datum, kdy se hrál zápas
- **match\_id** — identifikátor zápasu
- **team** — jméno týmu
- **map** — název hrané mapy
- **map\_winner** — binární značení, zda tým vyhrál (1) či prohrál (0)
- **starting\_ct** — binární značení, zda tým začal zápas na straně Counter-Terroristů (1) či Terroristů (0)
- **team\_rank** — rank týmu v okamžik, kdy se zápas hrál<sup>4</sup>
- **run\_mean\_3\_months** — klouzavý průměr týmu za poslední tři měsíce

#### 4.2.3 Omezení datového souboru

---

<sup>2</sup>hráč zabil nepřítele střelou do hlavy

<sup>3</sup><https://www.hltv.org/news/20695/introducing-rating-20>

<sup>4</sup><https://www.hltv.org/news/16061/introducing-csgo-team-ranking>

Všechny datové soubory obsahují pozorování o zápasech a charakteristikách hráčů od konce roku 2015 do začátku roku 2020. Jelikož pro finální modely je nutné datové soubory *players.csv* a *results.csv* sjednotit, může se stát, že se vytvoří zápas bez hráčů. Může také nastat situace, kdy zápas nebude mít přiřazených právě deset různých hráčů. To může nastat když tým nemá v zápase právě 5 hráčů. Tým může mít méně než 5 hráčů z toho důvodu, že je např. amatérsky<sup>5</sup>. Více hráčů může hrát za tým v případě, že tým použil náhradníka. Zároveň je možné, že kvůli historickým změnám ve hře a na webovém portálu nebude možné získat všechny potřebné charakteristiky hráčů. Také je nutné kontrolovat, zda tým nehrál sám proti sobě. V případě všech zmíněných chyb jsou chybné záznamy odstraněny.

## 4.3 Průzkumová analýza dat

Průzkumová analýza vizualizuje prediktory, hledá různé vztahy a rozdělení proměnných. Díky průzkumu lze určit, které proměnné není vhodné použít pro tvorbu logistického regresního modelu, např. kvůli problému multikolinearity.

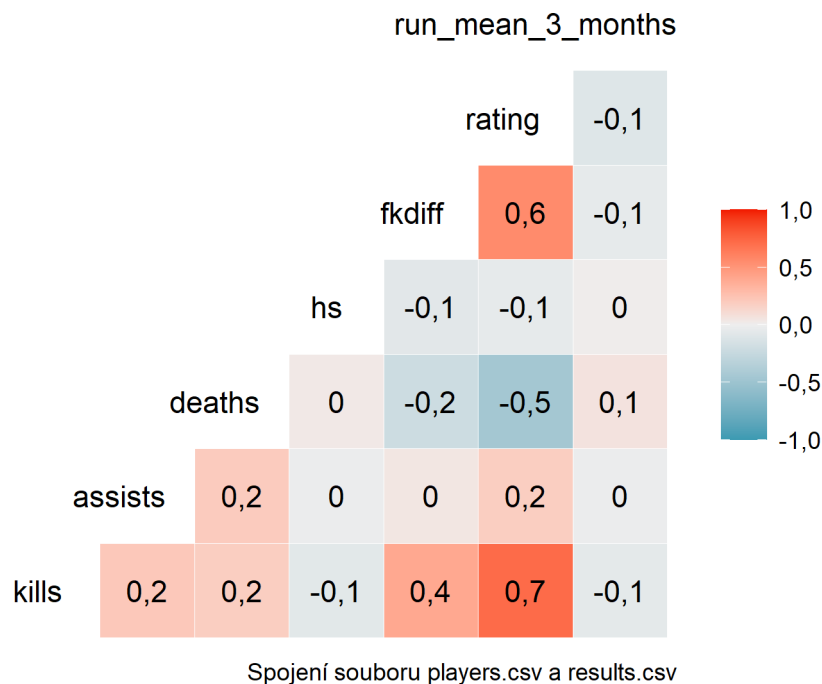
---

<sup>5</sup>neprofesionální, tím pádem nemusí mít všichni hráči na webovém portálu založená profil

### 4.3.1 Korelační matice

Pro logistickou regresi je důležité, aby prediktory nebyly lineárně závislé. Přehled korelací mezi kvantitativními prediktory lze zjistit z korelační matice.

#### Korelační matice kvantitativních prediktorů

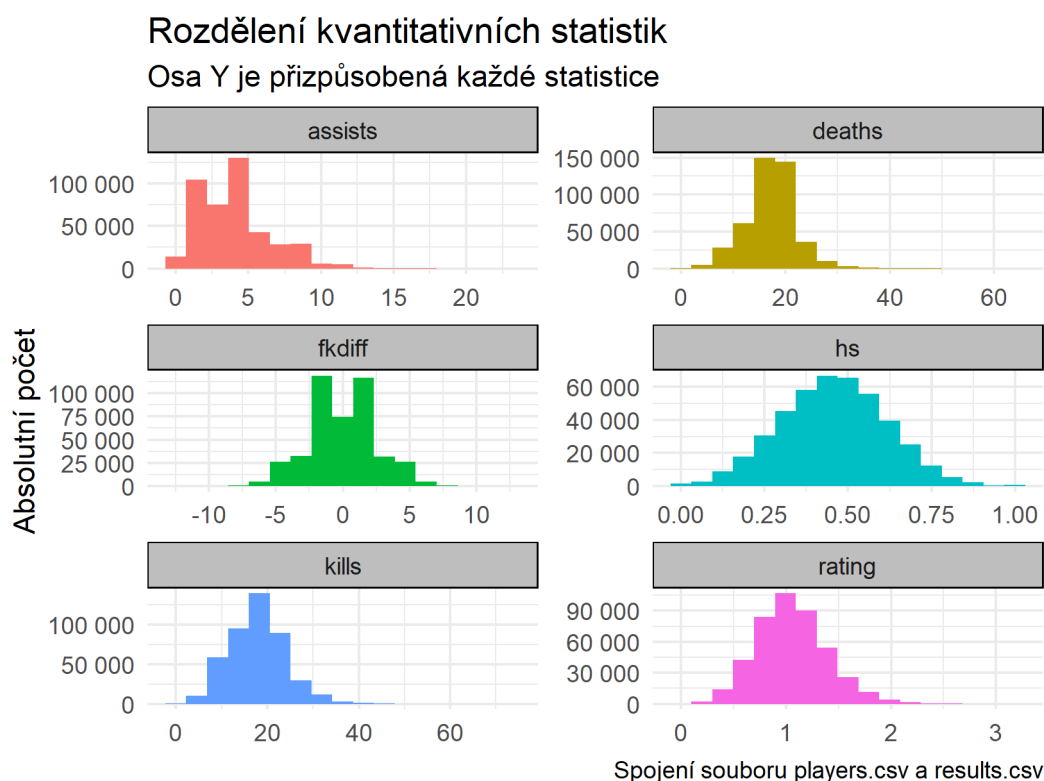


Obrázek 4.1: Korelační matice

Z korelační matice 4.1 lze vyčíst, že korelace mezi průměrem týmu za poslední tři měsíce a statistikami hráčů se blíží nule. Z toho plyne, že neexistuje lineární závislost mezi výkonem hráčů a průměrným rankem týmu za poslední tři měsíce. Zároveň je vidět silná korelace mezi prediktorem *rating* a prediktory *fkdiff*, *deaths* a *kills*. Jelikož by díky vysoké korelaci prediktorů vznikl problém multikolinearity, prediktor *rating* ve finálních modelech není použit.

### 4.3.2 Histogramy kvantitativních prediktorů

Histogramy kvantitativních prediktorů umožní zobrazit jejich rozdělení.



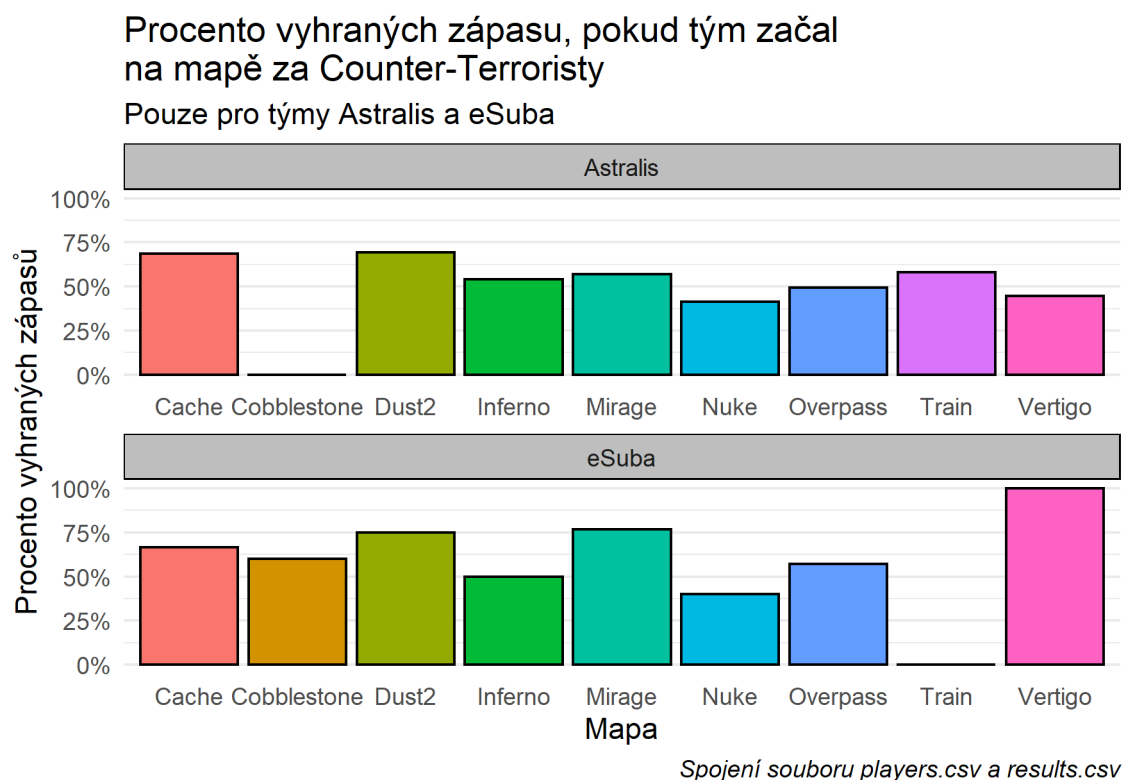
Obrázek 4.2: Histogram prediktorů

Histogram prediktorů *rating*, *hs*, *kills* a *deaths* má normální rozdělení a v proměnné se nena-  
chází mnoho extrémních hodnot. Prediktor *fkdif* má bimodální rozdělení. Prediktor *assists*  
je **zešíkmení** doprava, což značí velké množství odlehých či extrémních hodnot. **Pro logistic-**  
**kou regresy není předpoklad normálního rozdělení prediktorů. Proto analýza slouží hlavně k**  
**získání povědomí o tom, jakých hodnot každý prediktor nabývá a jaké je jejich rozdělení.**



### 4.3.3 Sloupcový graf výher přes počáteční stranu

Každá mapa funguje a vypadá jinak, což ovlivňuje mimo jiné i možné strategie. Proto jiné mapy mohou více vyhovovat jiným týmům. Toto způsobuje, že tým může mít vyšší procento výhry na mapě, pokud začíná na straně Counter-Terroristů.



Obrázek 4.3: Procento vyhraných zápasů na dané mapě za stranu Counter-Terroristů

Z obrázku 4.3 je patrné, že týmu Astralis nevyhovuje začínat mapu Cobblestone na straně Counter-Terroristů. Tým eSuba nemá žádný vyhraný zápas při začínající straně Counter-Terroristů na mapě Train. Díky obrázku je vidět, že šanci na výhru ovlivňuje interakce mezi mapou a začínající stranou. Logistický model by měl tento vztah reflektovat.

## 4.4 Modely pro výkon hráče

Cílem modelu je predikovat výhru zápasu pro jednotlivé hráče a identifikovat významné prediktory na odlišných mapách. Prediktory se týkají pouze výkonu jednotlivých hráčů, model tedy pro předpověď výhry hráče nepoužívá statistiky spoluhráčů. Pro porovnání jsou vybrány mapy Mirage a Vertigo. Mapa Mirage je jednou z nejvíce tradičních map a mapa Vertigo je naopak nejnovější přídavek do hry. Díky rozdílným modelům bude možné zkoumat, na čem pravděpodobnost výhry na mapách záleží. Pro vytvoření logistického modelu je použito 80% náhodně vybraných pozorování. Zbýlých 20% je použito pro validaci a vytvoření matice záměn.

#### 4.4.1 Model pro mapu Mirage

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	2,446	0,060	40,666	0,000
kills	0,181	0,002	87,335	0,000
assists	0,302	0,005	64,062	0,000
deaths	-0,369	0,003	-115,755	0,000
hs	-0,134	0,064	-2,098	0,036
fkdiff	0,015	0,005	3,209	0,001
starting_ct	-0,227	0,020	-11,370	0,000

Tabulka 4.1: Výpis z funkce glm() v jazyku R

Pro model jsou významné všechny prediktory. S rostoucím počtem zabití, asistencí a statistikou fkdiff má hráč větší šanci mapu vyhrát. Naopak s rostoucím počtem smrtí a statistikou hs, šance na výhru klesá. Mapa je pro stranu Counter-Terroristů nevýhodná, a pokud hráč začne mapu na této straně, jeho šance na výhru klesá. Největší vliv má na pravděpodobnost výhry počet úmrtí hráče. Přepis funkce pak vypadá následovně

$$P(1|X_{kills}, X_{assists}, X_{deaths}, X_{hs}, X_{fkdiff}, X_{starting_{ct}}) = \frac{1}{1 + e^{-z}} \quad (4.1)$$

$$z = 2,446 + 0,181 * X_{kills} + 0,302 * X_{assists} - 0,369 * X_{deaths} +$$

$$- 0,134 * X_{hs} + 0,015 * X_{fkdiff} - 0,227 * X_{starting_{ct}}$$

#### 4.4.2 Matice záměn pro mapu Mirage

Predikce jsou provedené na validačním podmnožině z datového souboru.

	Původní pozitivní	Původní negativní
Pozitivní predikce	7022	1928
Negativní predikce	1615	6709

Tabulka 4.2: Matice záměn pro mapu Mirage

Model predikoval správně 7 022 výher a 6 709 proher. Z matice záměn lze dále vypočítat následující statistiky:

statistika	hodnota
Přesnost	0,795
Senzitivita	0,777
Specifická	0,813

Tabulka 4.3: Statistiky pro mapu Mirage z matice záměn

Model určuje zhruba 79,5 % dat správně. Specifická je o zhruba 4 % body vyšší než senzitivita, což značí, že je model úspěšnější při predikci proher než výher.

#### 4.4.3 Model pro mapu Vertigo

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	1,753	0,221	7,937	0,000
kills	0,183	0,008	22,703	0,000
assists	0,301	0,018	16,924	0,000
deaths	-0,351	0,012	-29,083	0,000
hs	0,207	0,251	0,826	0,409
fkdiff	0,008	0,018	0,460	0,645
starting_ct	0,089	0,077	1,159	0,247

Tabulka 4.4: Výpis z funkce glm() v jazyku R

Pro hráče jsou na mapě Vertigo významné pouze prediktory *kills*, *assists* a *deaths*. Ostatní prediktory *hs*, *fkdiff* a *starting\_ct* jsou pro model nevýznamné. Nevýznamnost lze interpretovat tak, že pro hráče není důležité, na jaké straně mapu začnou hrát (prediktor *starting\_ct*), ani jak přesně střílí (prediktor *HS*). Zde má na výhru hráče největší negativní vliv počet úmrtí. Po vyřazení nevýznamných prediktorů má model následující významné parametry

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	1,893	0,181	10,475	0,000
kills	0,184	0,007	24,804	0,000
assists	0,301	0,018	16,944	0,000
deaths	-0,352	0,012	-29,996	0,000

Tabulka 4.5: Výpis z funkce glm() v jazyku R

Přepis funkce je pak následující

$$P(1|X_{kills}, X_{assists}, X_{deaths}) = \frac{1}{1 + e^{-z}} \quad (4.2)$$

$$z = 1,893 + 0,184 * X_{kills} + 0,301 * X_{assists} - 0,352 * X_{deaths}$$

Každé zabití zvyšuje šanci na výhru o zhruba 6.63. Naopak každá smrt hráče šanci na výhru snižuje zhruba 0.70.

#### 4.4.4 Matice záměn pro mapu Vertigo

Matice záměn pro logistický model sestavený pro mapu Vertigo je následující

	Původní pozitivní	Původní negativní
Pozitivní predikce	415	110
Negativní predikce	136	441

Tabulka 4.6: Matice záměn pro mapu Vertigo

Model úspěšně predikoval 415 výher a 441 proher. Z matice záměn pak lze získat výkonnostní statistiky.

statistika	hodnota
Přesnost	0,777
Senzitivita	0,800
Specifická	0,753

Tabulka 4.7: Statistiky pro mapu Vertigo z matice záměn

Na nových datech model určil správně zhruba 77,7% pozorování. Model má 80% senzitivitu, což je více než jeho specifická o zhruba 5 %. Model je tedy spolehlivější pro předpověď výhry hráče.

#### 4.4.5 Interpretace výsledků

Model pro mapu Mirage je přesnější a je spolehlivější pro předpověď prohry hráče. Model pro mapu Vertigo má menší přesnost o zhruba 2 procentní body a je vhodnější pro predikování výhry se 80 % senzitivitou.

Pro logistický model mapy Mirage jsou významné všechny zvolené prediktory. Největší vliv na výhru hráče na mapě Mirage má počet jeho smrtí, kdy s každou další smrtí se šance na výhru zmenšuje o zhruba 0,69. Pro mapu Vertigo jsou významné pouze prediktory *kills*, *assists* a *deaths* a největší vliv na výhru hráče má zde také počet smrtí.

Rozdíl by mohl být vysvětlen právě stářím jednotlivých map. Jelikož je mapa Mirage už tradiční mapou, mnoho hráčů na ní zná různé triky a strategie. To vede k tomu, že každý malý detail hraje velkou roli a všechny vybrané prediktory jsou pro model významné. Naopak mapa Vertigo je relativně nová a pro hráče je nejdůležitější individuální výkon. Jelikož hráči mapu tak dobře strategicky neznají, jsou významné pouze statistiky jejich hráčů.

## 5. Závěr

... Uzavření bakalářské práce

### 5.1 Závěrečné vyhodnocení modelu

... Výsledné vyhodnocení modelu pomocí všech statistik

### 5.2 Interpretace modelu do reálného světa

... Přenesení modelu do reálného světa

### 5.3 Použití modelu v reálném světě

... Použití modelu v reálném světě

### 5.4 Místo pro budoucí vylepšení

...

# Seznam použité literatury

Kleinbaum, David (2010). *Logistic regression : a self-learning text*. 3. vyd. Springer. ISBN: 978-1-4939-3697-7.

# Seznam elektronických zdrojů

- csko (2021). „základní statistiky“. In: *csko.cz*. URL: <https://stats.csko.cz/statsx/hlstats.php>.
- Gough, Christina (srp. 2021). „esports market revenue worldwide from 2019 to 2024“. In: *statista*. URL: <https://www.statista.com/statistics/490522/global-esports-market-revenue/>.
- Henningson, Joakim (2020). „the history of counter-strike“. In: *redbull*. URL: <https://www.redbull.com/se-en/history-of-counterstrike>.
- hltv.org (2015). „xenex vs. excel at esl uk premiership season 1“. In: *hltv.org*. URL: [https://www.hltv.org/matches/2295340/xenex-vs-excel-esl-uk-premiership-season-1?\\_\\_cf\\_chl\\_jschl\\_tk\\_\\_=82S\\_Uc\\_dNU8PM71eUY1VKzNUZkp5\\_ArTb69qteh2wBI-1641459294-0-gaNycGzNChE](https://www.hltv.org/matches/2295340/xenex-vs-excel-esl-uk-premiership-season-1?__cf_chl_jschl_tk__=82S_Uc_dNU8PM71eUY1VKzNUZkp5_ArTb69qteh2wBI-1641459294-0-gaNycGzNChE).
- Larch, Florian (led. 2019). „the history of the origin of esports“. In: *ispo*. URL: <https://www.ispo.com/en/markets/history-origin-esports>.
- liquipedia (2021). „pgl major stockholm 2021“. In: *liquipedia*. URL: <https://liquipedia.net/counterstrike/PGL/2021/Stockholm>.
- Professeur (2021). „esea increase prize pool and number of seasons for 2021; simplify path to pro league“. In: *hltv*. URL: <https://www.hltv.org/news/30926/esea-increase-prize-pool-and-number-of-seasons-for-2021-simplify-path-to-pro-league>.
- Valve (2013). „counter-strike: the arms deal update“. In: URL: <http://counter-strike.net/armsdeal>.



# Seznam obrázků

3.1	Bodový graf hmotnosti a míly za galon . . . . .	9
3.2	Sloupcový graf počtu válců a průměrné hrubé koňské síly . . . . .	10
3.3	Porovnání histogramů s různým počtem skupin . . . . .	11
3.4	Histogram s počtem skupin dle Sturgesova pravidla . . . . .	12
3.5	Krabičkový graf hmotnosti auta pro různý počet válců . . . . .	14
3.6	Korelační matice . . . . .	15
4.1	Korelační matice . . . . .	25
4.2	Histogram prediktorů . . . . .	26
4.3	Procento vyhraných zápasů na dané mapě za stranu Counter-Terroristů . . . . .	27

# Seznam tabulek

3.1	Five-number summary tabulka hmotnosti vozidla (lb/1000) . . . . .	12
3.2	Matice záměn . . . . .	19
4.1	Výpis z funkce glm() v jazyku R . . . . .	28
4.2	Matice záměn pro mapu Mirage . . . . .	29
4.3	Statistiky pro mapu Mirage z matice záměn . . . . .	29
4.4	Výpis z funkce glm() v jazyku R . . . . .	29
4.5	Výpis z funkce glm() v jazyku R . . . . .	30
4.6	Matice záměn pro mapu Vertigo . . . . .	31
4.7	Statistiky pro mapu Vertigo z matice záměn . . . . .	31
A.1	Záznam z transformovaného datového souboru players.csv . . . . .	39
A.2	Příklad záznamu z transformovaného datového souboru results.csv . . . . .	39

# Seznam použitých zkratek

**CSGO** Counter-Strike: Global Offensive

**BR** Battle Royale

**MOBA** Multiplayer Online Battle Arena

**FPS** First-Person Shooter

**TGNS** Twin Galaxies National Scoreboard

**Část I**

**Přílohy**

# A. Datové soubory

## A.1 Transformovaný datový soubor players.csv

Tabulka A.1: Záznam z transformovaného datového souboru players.csv

match_id	player_id	team	map	kills	assists	deaths	hs	fkdiff	rating
2339385	8738	Liquid	Overpass	15	3	12	0.6	3	1.32

## A.2 Transformovaný datový soubor results.csv

Tabulka A.2: Příklad záznamu z transformovaného datového souboru results.csv

date	match_id	team	map	map_winner	starting_ct	team_rank	run_mean_3_months
2018-02-18	2319765	AGO	Mirage	1	1	20	20,96552