

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 Zdeňku Šulcovi, který mou bakalářskou práci podpořil a vedl, i přes odlišný studijní obor. Dále děkuji autorům citovaných knih, článků a softwaru za poskytnutou příležitost se v logistických modelech zlepšit a napsat na toto téma bakalářskou práci.

---

## Abstrakt

Esport je sport ve virtuálním světě, který se od počátku dvacátého prvního století rozrůstá mezi novou generací, která vyrůstala ve světě počítačů, mobilů a technologií. Esport nejsou jen klasické sporty jako fotbal, hokej či rugby, ale soutěžit se může i v různých počítačích, mobilních či konzolových hrách. Tato práce je zaměřená na jednu z počítačových her, a to na hru Counter-Strike: Global Offensive (CSGO). Téma je aktuální zejména díky tomu, že je esport ve světě relativně nová záležitost a neustále se vyvíjí. Na to se musí přizpůsobit například sázkové kanceláře, které využívají analýzu esportových dat pro nabízení mnoha různých kurzů.

Cílem bakalářské práce je kvantitativně zanalyzovat esportové zápasy ze hry CSGO, predikovat výhry hráčů a týmů a zjistit, jaké prediktory jsou pro výhru zápasu statisticky významné. Použitý datový soubor je z internetového portálu kaggle.com a obsahuje data od roku 2015 až do roku 2020.

Teoretická část práce se zabývá představením a historií esportu a esportové hry CSGO. Teoretická část je pak zaměřená na analýzu dat v programovacím jazyku R. Predikce výher je založena na logistickém vícerozměrném modelu a k jeho vyhodnocení je použita matice záměn a statistiky z ní vypočítané. Pro určení významnosti prediktorů je použit Waldův test.

Výsledek práce jsou logistické modely, které jsou schopné predikovat výhru hráče či týmu podle různých charakteristik. Zároveň jsou identifikované významné prediktory, které výhru zápasu ovlivňují. Toto zjištění by bylo možné použít např. pro stanovení kurzu sázkovou kanceláří na výhru či prohru daného hráče či týmu.

## Klíčová slova

model, logistická regrese, predikce, esport

---

## Abstract

Přeložit před odevzdáním

## Keywords

model, logistic regression, prediction, esport

# Obsah

<b>1</b>	<b>Úvod</b>	<b>2</b>
<b>2</b>	<b>Představení esportu</b>	<b>3</b>
2.1	Historie esportu . . . . .	3
2.2	Zasazení do dnešní doby . . . . .	3
2.3	Představení titulu Counter-Strike: Global Offensive . . . . .	4
2.4	Propojení práce a titulu Counter-Strike: Global Offensive . . . . .	6
<b>3</b>	<b>Teoretická část</b>	<b>8</b>
3.1	Vizualizace dat . . . . .	8
3.1.1	Bodový graf . . . . .	8
3.1.2	Sloupcový graf . . . . .	9
3.1.3	Histogram . . . . .	10
3.1.4	Krabičkový graf . . . . .	11
3.1.5	Korelační matice . . . . .	13
3.2	Logistická regrese . . . . .	15
3.2.1	Interakce mezi prediktory . . . . .	15
3.2.2	Interpretace parametrů . . . . .	15
3.2.3	Maximální pravděpodobnost . . . . .	16
3.2.4	Křížová validace . . . . .	17
3.2.5	Matice záměn . . . . .	17
3.2.6	Testování hypotéz . . . . .	18
3.2.7	Waldův test . . . . .	19
<b>4</b>	<b>Teoretická část</b>	<b>20</b>
4.1	Cíle analýzy . . . . .	20
4.2	Příprava dat . . . . .	21
4.2.1	Soubor players.csv . . . . .	21
4.2.2	Soubor results.csv . . . . .	22
4.2.3	Datový soubor pro modelování . . . . .	23
4.2.4	Agregovaný datový soubor . . . . .	23
4.2.5	Trénování a validace modelů . . . . .	24
4.3	Průzkumová analýza dat . . . . .	25
4.3.1	Korelační matice . . . . .	25
4.3.2	Histogramy kvantitativních prediktorů . . . . .	26
4.3.3	Míra výhry pro nejlepší týmy . . . . .	26
4.4	Predikce výhry hráčů na různých mapách . . . . .	28
4.4.1	Model pro mapu Mirage . . . . .	28
4.4.2	Model pro mapu Vertigo . . . . .	30
4.4.3	Interpretace a porovnání modelů . . . . .	31

---

4.5	Predikce výhry týmu . . . . .	33
4.5.1	Celkový model . . . . .	34
4.5.2	Model pro tým Astralis . . . . .	35
4.5.3	Model pro tým Sprout . . . . .	36
4.5.4	Interpretace a porovnání modelů . . . . .	38
<b>5</b>	<b>Závěr</b>	<b>40</b>
	<b>Seznam použitého softwaru</b>	<b>41</b>
	<b>Seznam použité literatury</b>	<b>42</b>
	<b>Seznam elektronických zdrojů</b>	<b>43</b>
	<b>Seznam obrázků</b>	<b>44</b>
	<b>Seznam tabulek</b>	<b>45</b>
	<b>Seznam použitých zkratk</b>	<b>47</b>
<b>I</b>	<b>Přílohy</b>	<b>48</b>
<b>A</b>	<b>Datové soubory</b>	<b>49</b>
A.1	Transformovaný datový soubor players.csv . . . . .	49
A.2	Transformovaný datový soubor results.csv . . . . .	49
A.3	Spojený datový soubor . . . . .	49
A.4	Agregovaný datový soubor . . . . .	50
<b>B</b>	<b>Modely, matice záměn a statistiky pro individuální hráče</b>	<b>51</b>
B.1	Mapa Cache . . . . .	51
B.2	Mapa Cobblestone . . . . .	52
B.3	Mapa Dust2 . . . . .	53
B.4	Mapa Inferno . . . . .	54
B.5	Mapa Nuke . . . . .	55
B.6	Mapa Overpass . . . . .	56
B.7	Mapa Train . . . . .	57
<b>C</b>	<b>Kompletní kód při tvorbě praktické části bakalářské práce</b>	<b>58</b>

# 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. 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 výborné 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...).

Práce se zaměřuje primárně na esportový titul *Coutner-Strike: Global Offensive (CSGO)*. Je to jeden z nejdéle hraných esportových titulů, boří mnohé divácké rekordy a je aktuálně nejhranějším *First-Person Shooter (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>1</sup> a hráč může strávit mnoho hodin pouze objevováním komunitních serverů, hraním klasických zápasů či trénováním na offline mapách.

Finálním cílem práce je vytvořit vícerozměrný logistický regresní model, který předpovídá výsledky zápasů. Pro tvorbu kvalitního modelu bude kritické zvolit vhodné charakteristiky hráčů a jednotlivých týmů. Pro predikci jsou použity charakteristiky, které se nacházejí ve dvou samostatných datových souborech, které podávají informace jak už o zápase (např. datum, výsledek zápasu, výsledek jednotlivých map, typ zápasu) a hráčích (např. charakteristiky za zápas).

Predikovat výhru hráče a týmu je důležité zejména v oblasti sázení, která je s esportem úzce spojená. Přesné predikce a kvalitní modely sázkovým kancelářím umožňují stanovovat výhodné a profitabilní kurzy. Kurzy pak nemusí být pouze na výhru či prohru, ale např. na charakteristiky hráčů či hranou mapu. Logistický model je pro predikci výhry či prohry preferován kvůli své snadné interpretaci.

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 datového souboru, logistického regresního modelu, a evaluační nástroje pro model. Třetí část se zaměřuje na praktickou tvorbu modelů, jejich interpretaci, a vzájemné porovnání.

---

<sup>1</sup>rozšíření, jak hru hrát. Každý mód má svá vlastní pravidla, mapy, či herní fanoušky



## 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, 2022), (Larch, 2022).

### 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, 2022).

---

<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, 2022), (liquipedia.net, 2021). S větším počtem diváku roste i marketingový potenciál, investiční příležitost a kariérní růst.

Jedním z dominantních žánrů je žánr 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).

Druhý významný a zajímavý žánr je žánr 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 kotelna hrálo celkem přes 1,5 milionu unikátních hráčů (csko.cs, 2022). 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. V dnešní době majory trhají světové rekordy a kouká se na ně miliony diváků<sup>8</sup>. 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

<sup>8</sup><https://www.invenglobal.com/articles/15619/csgo-major-breaks-viewership-records-overtakes-the-international>

## 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 (Professeur, 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 z leva 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

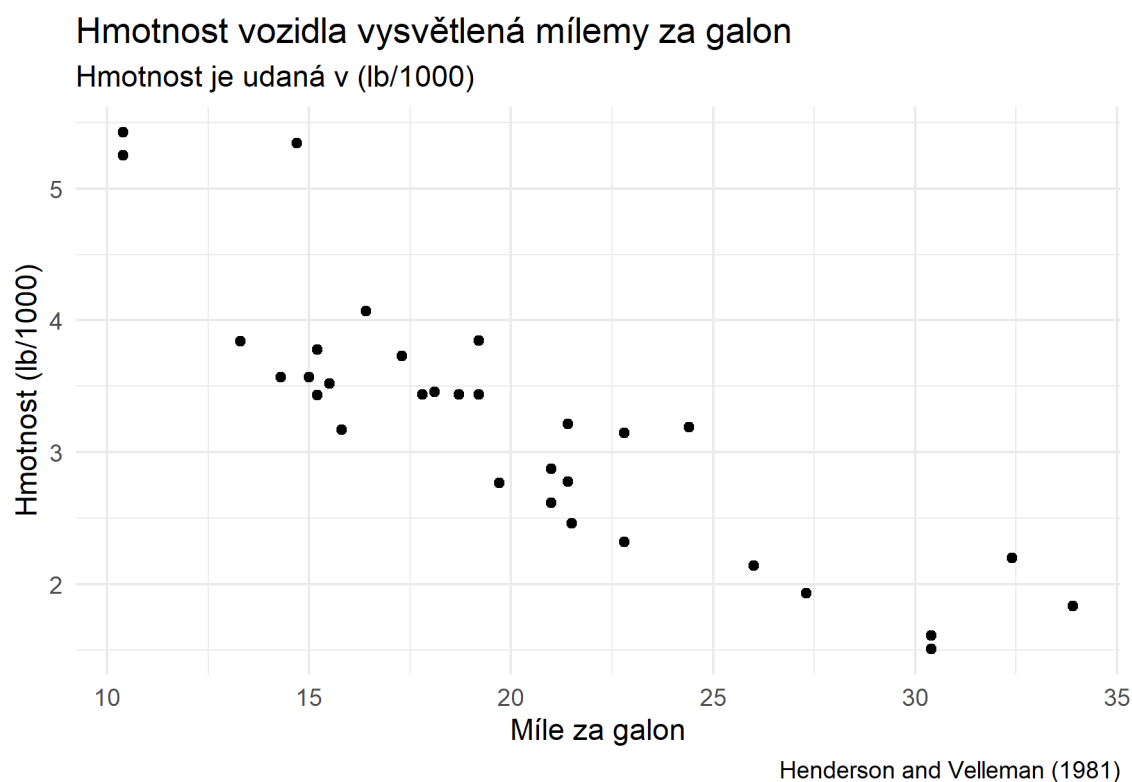
V následující části jsou popsány jak teoretické metody pro vizualizaci dat, tak i tvar, forma a vyhodnocení logistického regresního modelu. Ke každé části, která se věnuje popisu dat pomocí nějakého grafu, je přidána praktická ukázka s popisem a praktickým vysvětlením. vhodné.

### 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

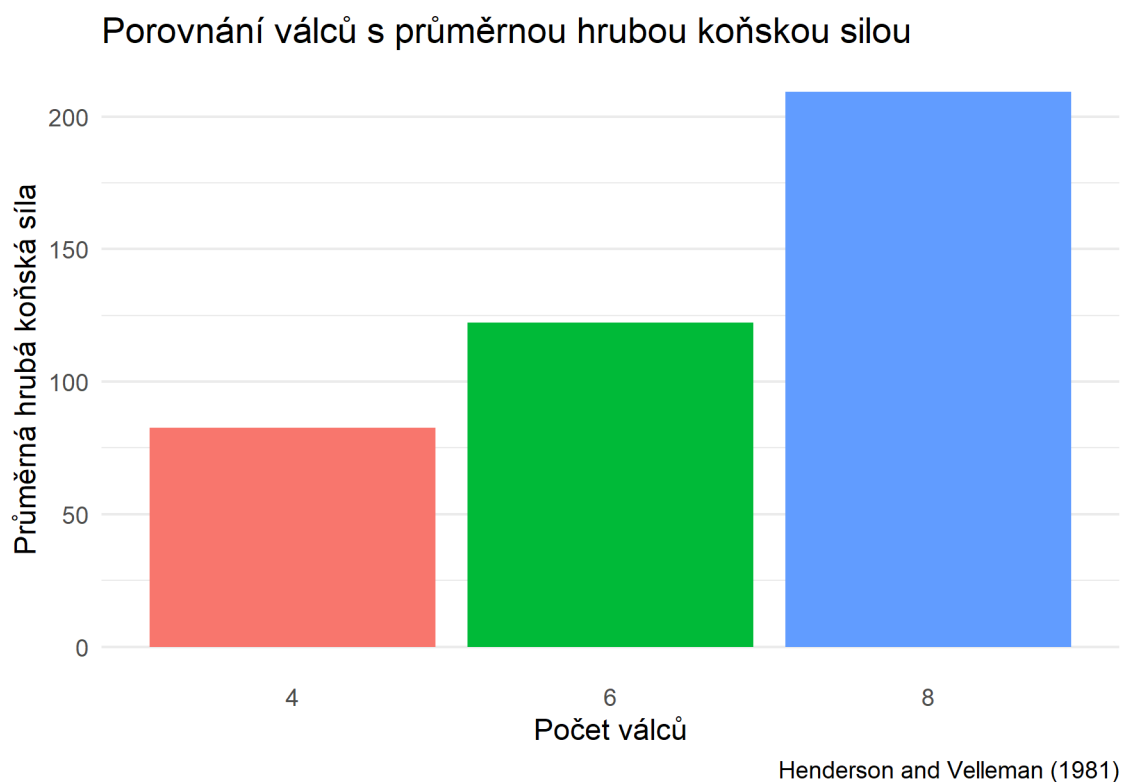
Obrázek 3.1 zobrazuje negativní korelaci mezi hmotností vozidla a mílemi ujetými za galon.

### 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 nejnížší. V případě ordinální proměnné se zachovává přirozené řazení. Příklad sloupcového grafu je zobrazen na obrázku

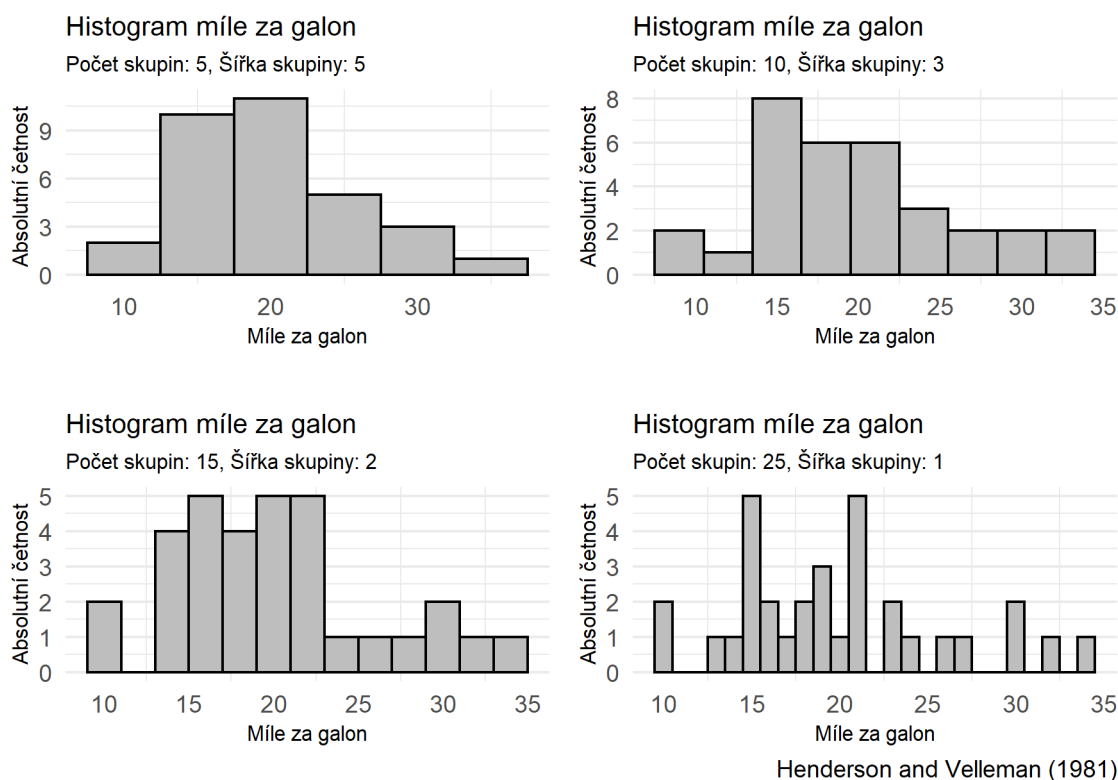
3.2, který porovnává průměrnou hrubou koňskou silou s počtem válců. Je na něm také vidět závislost hrubé koňské síly na počtu válců.



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 počítá následujícím vztahem:

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

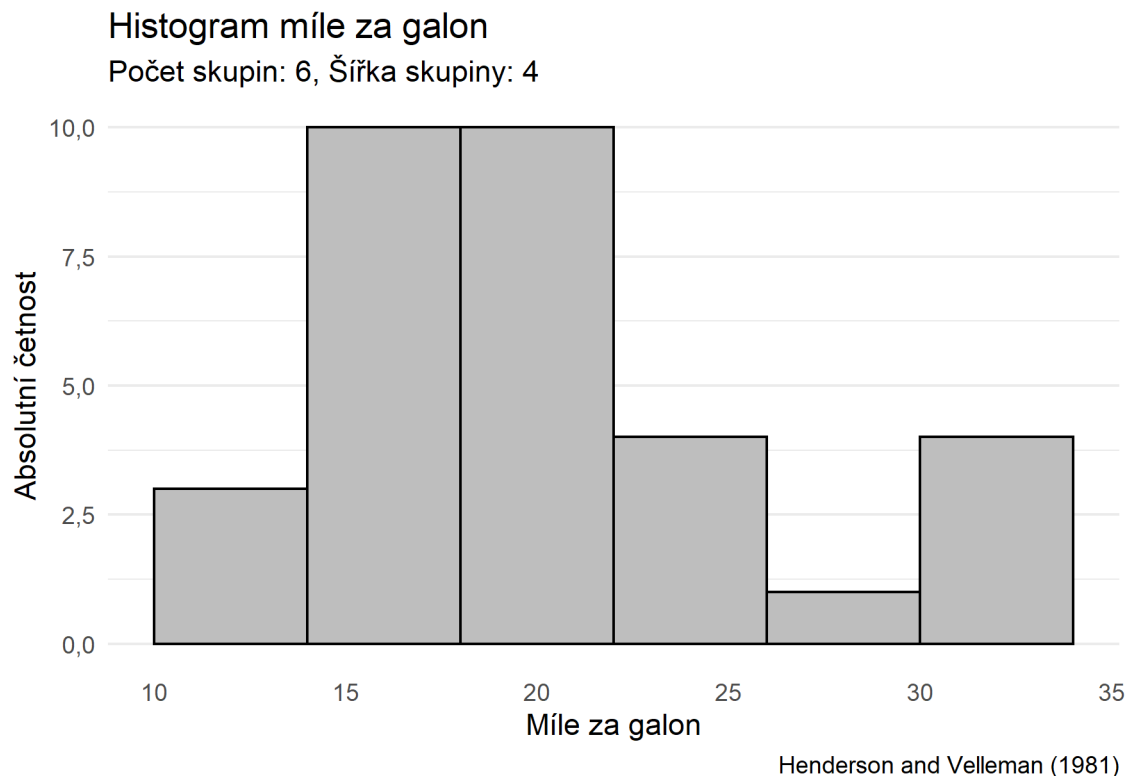
kde  $k$  je přibližný počet skupin a  $n$  je počet pozorování. Druhý parametr, který je pro tvorbu histogramu potřeba, je šířka skupiny. Ta by měla být ideálně stejná pro všechny skupiny. Pokud tomu tak není, histogram může být zavádějící a čtenář mu nemusí plně rozumět. Pro vypočtení počtu skupin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

Obrázek 3.4 ukazuje histogram proměnné míle za galon. Počet sloupců je vypočítán podle Sturgesova pravidla 3.1.

### 3.1.4 Krabičkový graf

#### Pětičíselné shrnutí

Pětičíselné shrnutí 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 tabulky.

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

Tabulka 3.1: pětičíselné shrnutí hmotnosti vozidla (lb/1000)

Příkladem pětičíselné shrnutí je tabulka 3.1,

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 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 pětičíselné shrnutí. 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.

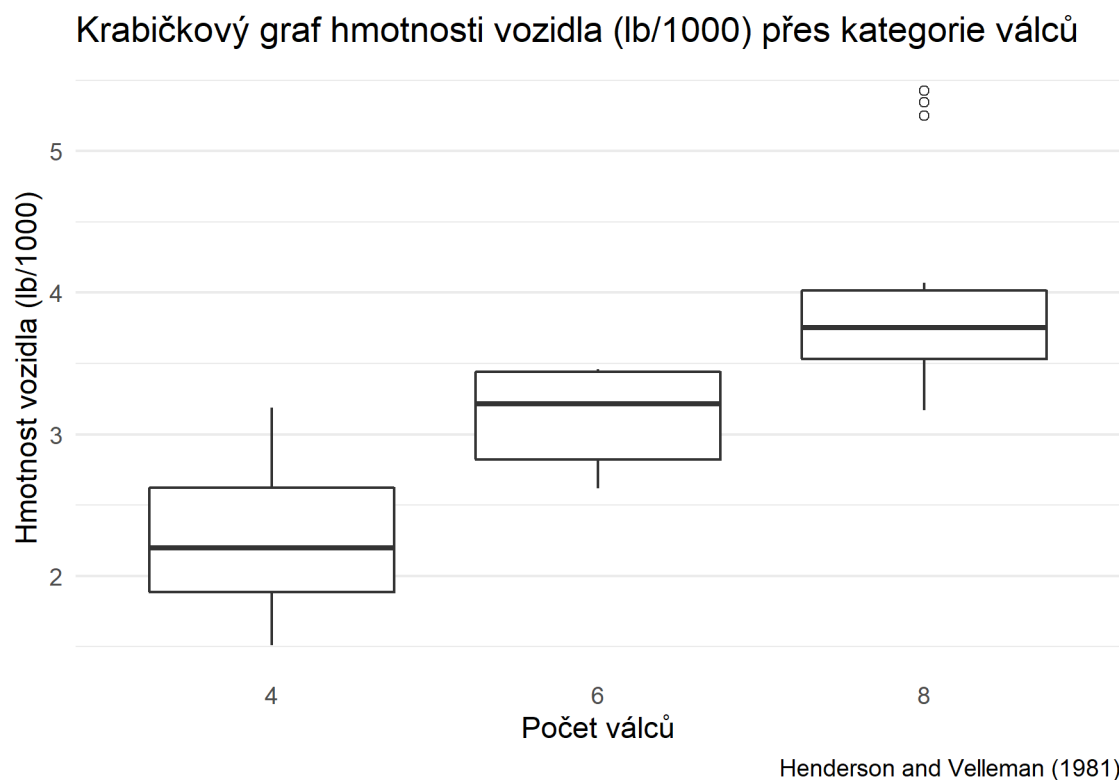
$$\text{Spodní hranice} = Q_1 - 1,5 * IQR$$

$$\text{Horní hranice} = Q_3 + 1,5 * IQR$$

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.

Odlehlá 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  $\langle -1; 1 \rangle$ .

$$r = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{(\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n})(\sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n})}} \quad (3.3)$$

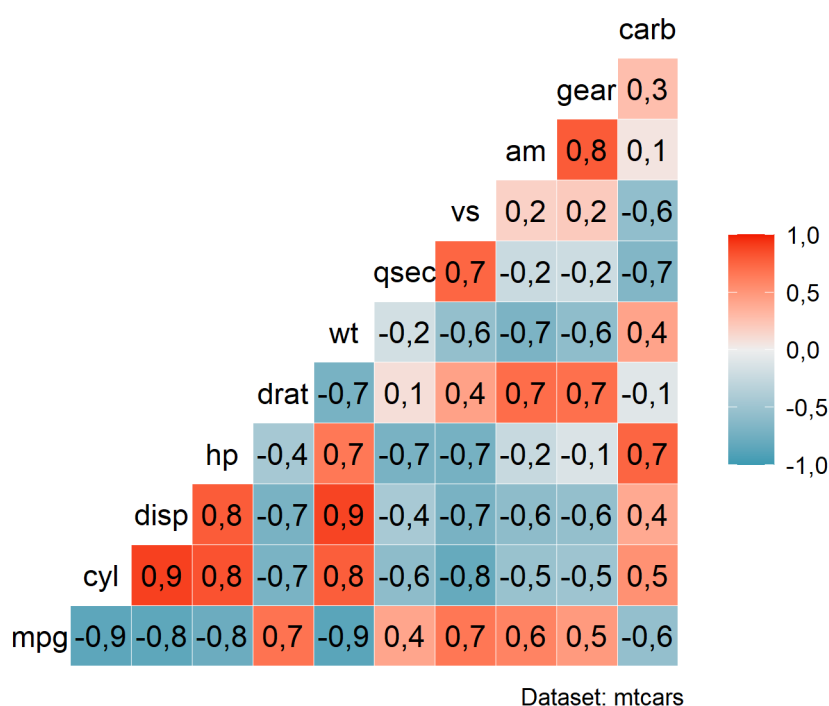
Čím větší je absolutní hodnota korelace mezi proměnnou  $X$  a  $Y$ , tím lépe lze pomocí jedné proměnné vysvětlit proměnnou druhou. Kladnost, případně zápornost korelace pak značí směr vztahu, tedy zda proměnná  $X$  s růstem proměnné  $Y$  klesá či stoupá. Pokud je korelace záporná, tedy  $r < 0$ , 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 neznamena kauzalitu. Pokud existuje kladná korelace mezi proměnnou  $X$  a  $Y$ , neznamena to, že růst jedné proměnné způsobí růst druhé proměnné.

## Korelační matice

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

$$f(z) = \frac{1}{1 + e^{-z}}. \quad (3.4)$$

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  $\alpha$ .

$$\begin{aligned} z &= \alpha + \beta_1 X_1 + \dots + \beta_2 X_2 + \beta_k X_k \\ &= \alpha + \sum_{i=1}^k \beta_i X_i \end{aligned} \quad (3.5)$$

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

$$P(Y = 1 \mid X_1, X_2, \dots, X_k) = \frac{1}{1 + e^{-\left(\alpha + \sum_{i=1}^k \beta_i X_i\right)}}, \quad (3.6)$$

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

### 3.2.1 Interakce mezi prediktory

Prediktory mohou na výslednou pravděpodobnost působit i dohromady, tzn. interakcí. Jako příklad lze uvést model, který predikuje *nehodu* podle číselného prediktoru *rychlost* a dichotomického prediktoru *dálnice*. Šance na nehodu může být ovlivněná jak rychlostí, tak tím, zda je řidič na dálnici.

### 3.2.2 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

$$\text{šance} = e^{\beta_i}. \quad (3.7)$$

Šance je podíl dvou pravděpodobností. Pokud bychom měli šanci jevu A oproti jevu B 2 : 1, značí to, že výskyt jevu A je dvakrát pravděpodobnější než 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  $Y = 1$ . Pokud je naopak nižší, pravděpodobnost se zmenšuje. Pokud je potřeba interpretovat pravděpodobnost jako šanci, použije se logitová funkce

$$\text{šance jevu } A = \frac{p}{1-p}, \quad (3.8)$$

kde  $p$  je pravděpodobnost výskytu jevu A. Pokud je vhodné šanci interpretovat spíše v procentech, šanci lze převést pomocí následujícího vztahu

$$\text{Pravděpodobnost jevu } A = \frac{\text{šance jevu } A}{1 + \text{šance jevu } A}, \quad (3.9)$$

### 3.2.3 Maximální pravděpodobnost

Parametry logistického modelu v rovnici 3.6 jsou pouze teoretické a je třeba je odhadnout. Již vypočtené odhady se proto neznačí pouze  $\beta_i$ , ale  $\hat{\beta}_i$ . Pro odhad parametrů se při logistické regresi používá metoda maximální věrohodnosti. Pro výpočet maximální věrohodnosti se používá věrohodnostní funkce  $L(\theta)$  kde  $\theta$  jsou parametry logistického modelu  $\alpha, \beta_1, \dots, \beta_k$ . Pro logistickou regresi má věrohodnostní funkce tvar

$$L(\theta) = \prod_{i=1}^{m_1} P(X_i) \prod_{i=m_1+1}^n (1 - P(X_i)), \quad (3.10)$$

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 3.4. 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 et al., 2010).

$$\frac{\partial L(\theta)}{\partial \beta_i} = 0 \quad (3.11)$$

### 3.2.4 Křížová validace

Při tvorbě logistického 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 které nebyl natrénován. Druhá množina dat se většinou nazývá *validační* nebo *testovací*. Pokud je následně model použit na *testovací* množinu dat a výsledky jsou vyhodnocené např. pomocí matice záměn, jsou zachycené variace, na které model není připraven a lze tak objektivněji určit kvalitu modelu. Způsobů, jak datový soubor rozdělit, je mnoho. Existuje například *k*-fold validace, kdy se trénovací množina dat rozdělí na *k* náhodných podmnožin. Jedna podmnožina dat se použije pro validaci a zbylých *k* – 1 podmnožin se použije pro trénování. Celý proces se opakuje *k* krát, tedy každá podmnožina bude právě jednou použita pro testování. Výsledné statistiky lze zprůměrovat a použít jako hodnocení daného modelu.

### 3.2.5 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 záměn 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  $n$ .

$$Přesnost = \frac{\text{Skutečně pozitivní} + \text{Skutečně negativní}}{n} \quad (3.12)$$

Přesnost říká, jaké procento objektů bylo klasifikováno správně. Pokud je ovšem poměr původních pozitivních a negativních hodnot velmi nevyrovnaný, tato statistika není vhodná. V případě velké nevyrovnanosti predikovaných hodnot přesnost zkresluje schopnost modelu predikovat méně zastoupenou predikovanou hodnotu. 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, neboli

$$Senzitivita = \frac{\text{Skutečně pozitivní}}{\text{Skutečně pozitivní} + \text{Falešně pozitivní}}. \quad (3.13)$$

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 a je definovaná jako

$$Specifická = \frac{\text{Skutečně negativní}}{\text{Skutečně negativní} + \text{Falešně negativní}}. \quad (3.14)$$

### 3.2.6 Testování hypotéz

Cílem testovací hypotézy je zjistit, zda neznámý parametr  $\theta$  patří do nějaké prostoru  $\Omega_0$ . V případě testování parametrů logistického modelu bude neznámý parametr  $\theta$  roven parametru  $\beta_i$ . Prostor  $\Omega_0$  je pak populace. Testuje se nulová hypotéza, značená  $H_0$ , která říká, že parametr  $\theta$  do prostoru  $\Omega_0$  patří. Proti ní je postavená hypotéza alternativní, značená  $H_1$ , která tvrdí, že parametr  $\theta$  do prostoru  $\Omega_0$  nepatří. Pro testování je nutné zvolit parametr  $\alpha$ , který značí maximální hodnotu chyby prvního druhu. Chyba prvního druhu stanovuje, jaká je pravděpodobnost, že se zamítne testovaná nulová hypotéza  $H_0$  za předpokladu, že je  $H_0$  pravdivá (Härdle et al., 2015).

$$P(\text{Zamítnutí } H_0 | H_0 \text{ je platná}) = \alpha \quad (3.15)$$

Existuje také chyba druhého typu  $\beta$ , která značí pravděpodobnost nezamítnutí neplatné nulové hypotézy.

$$P(\text{Nezamítnutí } H_0 | H_0 \text{ je neplatná}) = \beta \quad (3.16)$$



Chyby nelze eliminovat a je nutné je s nimi při testování hypotézy počítat. Pro parametr  $\alpha$  se ustálily hodnoty 0.1, 0.05 a 0.01.

### 3.2.7 Waldův test

Koeficienty v logistickém regresním modelu nemusí být statisticky významné. Pokud je prediktor nevýznamný, znamená to, že není významný při predikci prediktoru. K otestování významnosti prediktoru lze použít 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 roven 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  $S_{\hat{\beta}_i}$

$$Z = \frac{\hat{\beta}_i}{S_{\hat{\beta}_i}}. \quad (3.17)$$

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

## 4. Teoretická část

Cílem praktické části je prozkoumat dva typy modelů. První typ modelu bude predikovat výhru zápasu pro jednotlivé hráče a druhý typ pro jednotlivé týmy. Hlavním cílem modelů pro hráče 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 týmu. Pro agregaci bude použit buď aritmetický nebo geometrický průměr. Všechny modely jsou vyhodnoceny pomocí testovacího datového souboru a matice záměn. Pro tvorbu modelů a veškerou práci s daty je použit programovací jazyk R (Team, 2022) a vývojové prostředí RStudio (RStudio, 2022). Zdrojový kód celé analýzy je v příloze C

### 4.1 Cíle analýzy

Cílem analýzy je vytvořit logistické modely pro předpověď výhry jak pro individuální hráče, tak pro dva referenční týmy nebo všechny týmy. Jelikož se charakteristiky hráčů i zápasů mohou měnit dle mapy, na které se zápas odehrává, modely pro individuální hráče jsou tříděné přes kategorie map. Modely pro předpověď výhry týmu používají interakci mezi mapou a začínající stranou.

Pro hráče budou sestaveny logistické modely z celého datového souboru a pro každou mapu zvlášť. Modely budou následně vyhodnoceny pomocí statistik vycházejících z matice záměn a bude porovnána významnost jejich parametrů. Modely budou porovnané na vybraných dvou mapách, a to na mapě Mirage a Vertigo. Mirage je klasický 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 toho, jaké prediktory jsou pro model významné.

Modely pro týmy budou sestavené pro dva vybrané referenční týmy a pro všechny týmy. První tým, pro který se vytvoří model, bude tým Astralis. Ten je dlouhodobě považován za jeden z nejlepších týmu na světě, vyhrál několik prestižních Majorů a v grafu 4.3 je vidět, že má v datovém souboru největší míru výhry. Druhým týmem bude německý celek Sprout. Tým Sprout se řadí v době extrakce dat k profesionálnímu týmu s průměrným třicátým místem na světovém žebříčku. 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 modelem, který je sestaven na celém datovém souboru. 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.

## 4.2 Příprava dat

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 jednoho datového souboru, který obsahuje charakteristiky všech hráčů v právě jednom zápase, potřebné informace o zápase a výsledek (zda hráč zápas vyhrál či nikoliv). 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 zbylých dvou souborů (*picks.csv*, *economy.csv*) proto v bakalářské práci použit není.

### 4.2.1 Soubor *players.csv*

Datový soubor *players.csv* obsahuje charakteristiky jednotlivých hráčů v daném zápase. Původní datový soubor obsahuje 101 proměnných a 383 317 pozorování. V původním datovém souboru se jeden řádek (pozorování) rovná charakteristikám jednoho hráče za celý zápas, který se může odehrávat až na třech mapách. Pro potřeby bakalářské práce je tak nutné získat charakteristiky hráčů za jednotlivé mapy. Proto je původní datový soubor transformován do podoby, kde se jedno pozorování rovná charakteristikám právě jednoho hráče na právě jedné mapě, a to bez ohledu na to, 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.

Z datového souboru jsou odstraněny záznamy o mapě Default. Ta značí automatickou výhru pro jeden team, například díky formátu turnaje. V souboru je také možné narazit na tým, pro který za zápas na jedné mapě nehrálo právě 5 hráčů. Více než pět hráčů mohlo na mapě hrát v případě použitého náhradníka. Může se také stát, že hraje méně hráčů, pokud se jeden například nestihne dostavit. V obou případech jsou data pro danou mapu v zápase odstraněna.

---

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

Transformovaný a očištěný datový soubor má 10 proměnných a 640 225 pozorování. Příklad jednotlivých pozorování v transformovaném datovém souboru je v příložené tabulce A.1. Interpretace charakteristik 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 charakteristik 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 19 proměnných a 45 773 pozorování. Datový soubor *results.csv* obsahuje na rozdíl od datového souboru *players.csv* jedno chybné pozorování. Dle něho hrál tým sám proti sobě, což nedává věcný a logický smysl. Jelikož je zápas na webovém portálu zadán správně, nejspíše se jedná o neznámou chybu, která nastala při exportu dat z webového portálu. Zároveň jsou vybrané pouze proměnné, které pro predikci lze využít. Dále je zde rozdělený tým Astralis do dvou týmů: týmu Astralis a tým „?“. Důvod rozdělení je ten, že historicky hrál tým bez organizace a označoval se jako „?“. Po vytvoření organizace Astralis jsou hráči řazeni pod tuto organizaci. Jelikož se tímto týmem zabývá specializovaný logistický model, je označení týmu sjednoceno pod název Astralis.

Po transformacích vznikne tabulka o 7 proměnných a 91 502 pozorování. Každé pozorování identifikuje výsledek právě 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>

---

<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>

### 4.2.3 Datový soubor pro modelování

Pro modelování je nutné vytvořit jeden datový soubor, na kterém bude model sestaven. Datový model je získán spojením představených souborů *players.csv* a *results.csv*. Datové soubory jsou spojené pomocí proměnných *match\_id*, *team* a *map*. Při propojování souborů jsou smazané záznamy, které nejsou reprezentovány v obou souborech. To znamená, že charakteristiky hráčů v zápase, pro který není výsledek, jsou smazané. Stejně jsou smazané zápasy, pro které nejsou zadány charakteristiky hráčů. Spojený datový soubor je pak na úrovni jednotlivých hráčů a ukazuje jejich charakteristiky v právě jednom zápase na právě jedné mapě. Příklad spojeného datového souboru je v příložené tabulce A.3.

### 4.2.4 Agregovaný datový soubor

Spojený datový soubor obsahuje charakteristiky hráčů na úrovni jednotlivých zápasů. To je vhodné pro první logistický model, který se zabývá predikcí a identifikací významných prediktorů pro individuální hráče. Pro predikci výhry týmu a identifikaci významných prediktorů pro tým je nutné data agregovat. Pro agregaci charakteristik, u kterých to dává věcný smysl, je použit aritmetický průměr. Pro charakteristiky, u kterých dává věcný smysl násobení, je použit průměr geometrický. Jako příklad lze uvést charakteristika počtu zabití (prediktor *kills*), která je agregovaná pro daný tým jako průměrný počet zabití (prediktor *mean\_kills*) hráčů týmu na dané mapě. Pro charakteristiku procent zabití do hlavy (prediktor *hs*) je použit geometrický průměr a v modelu je použitý průměr procenta zabití do hlavy (prediktor *mean\_hs*). Příklad agregovaných dat je v příložené tabulce A.4. Popis agregovaných prediktorů je pak v následující tabulce.

Tabulka 4.1: Přehled agregací charakteristik pro daný tým na dané mapě v daném zápase

Prediktor	Agregace	Popis
<i>mean_kills</i>	Aritmetický průměr	Průměrný počet zabití
<i>mean_assists</i>	Aritmetický průměr	Průměrný počet asistencí
<i>mean_deaths</i>	Aritmetický průměr	Průměrný počet smrtí
<i>mean_hs</i>	Geometrický průměr	Průměrné procento zabití do hlavy
<i>mean_fkdiff</i>	Aritmetický průměr	Průměrný rozdíl mezi prvním zabitím a první smrtí

#### 4.2.5 Trénování a validace modelů

Pro křížovou validaci modelů se vytvoří náhodné rozdělení dat v poměru 8:2. Pro trénování modelu bude použito 80 % z datového souboru, na validaci je použito zbylých 20 %. Natrénované modely jsou vyhodnoceny pomocí Waldova testu, který je v každé tabulce v posledním sloupci  $Pr(> |z|)$ . Pro určení významnosti je použita hladina významnosti  $\alpha = 0.05$ . Případné nevýznamné prediktory jsou pak z modelu odstraněny a model je natrénován znovu na stejné trénovací množině. Pro validaci modelu je vytvořena matice záměn, která popisuje přesnost predikcí modelu na validačních datech. K následnému vyhodnocení modelů jsou použity statistiky Přesnost, Senzitivita a Specificita.

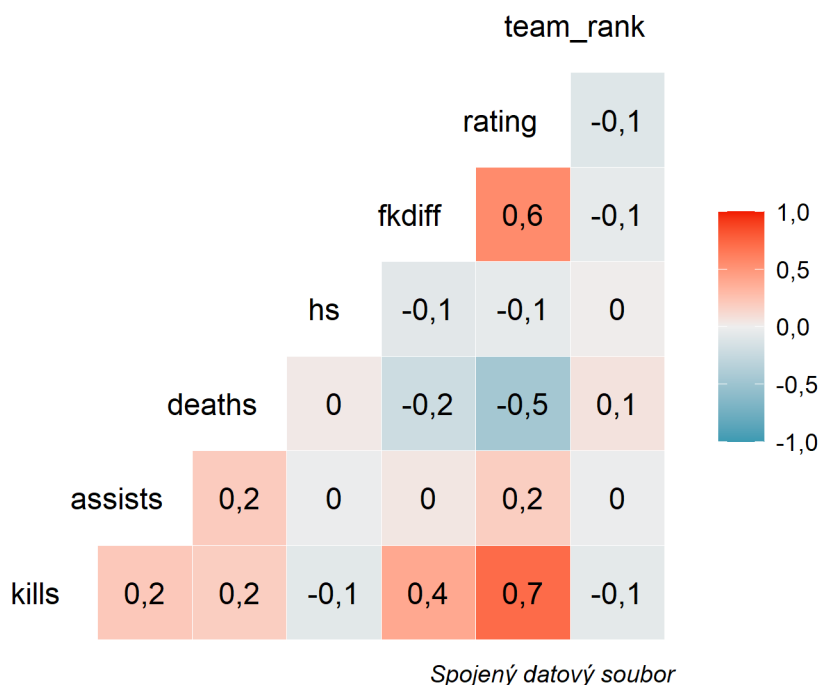
## 4.3 Průzkumová analýza dat

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.

### 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.

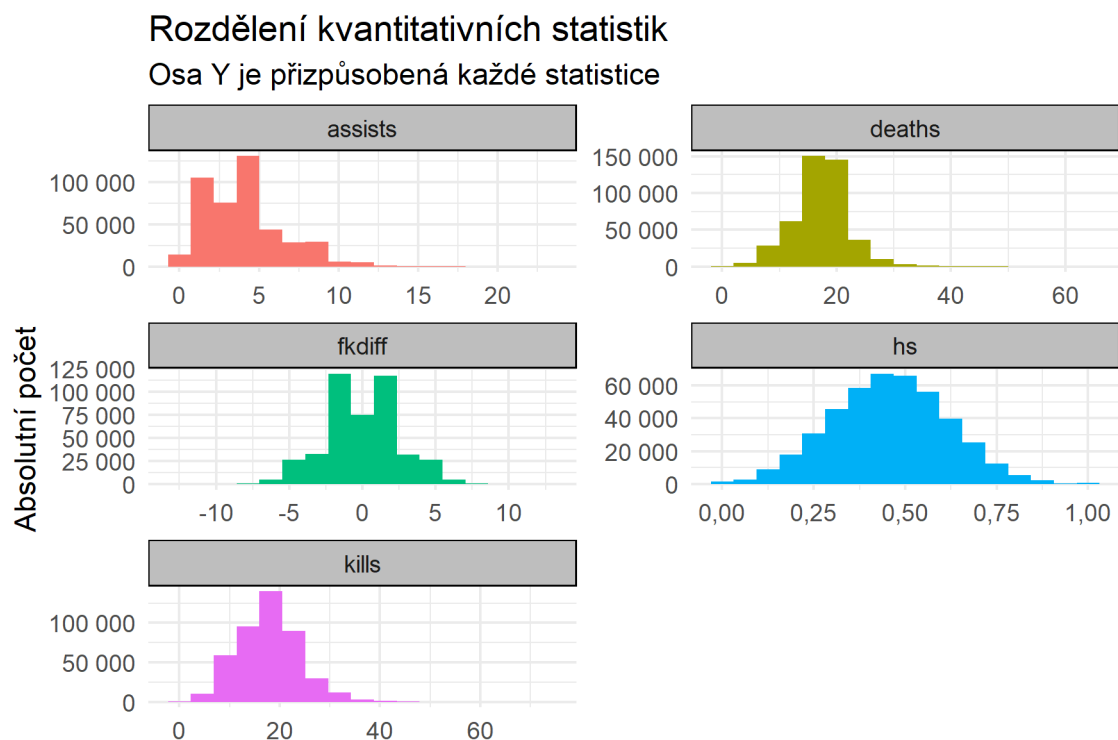
Korelační matice kvantitativních prediktorů



Obrázek 4.1: Korelační matice

Z korelační matice 4.1 lze vyčíst, že korelace mezi rankem týmu a charakteristikami hráčů se blíží nule. Z toho plyne, že neexistuje lineární závislost mezi výkonem hráče a rankem týmu. Zároveň je vidět silná korelace mezi prediktorem *rating* a prediktory *fkdif*, *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



*Spojení souboru players.csv a results.csv*

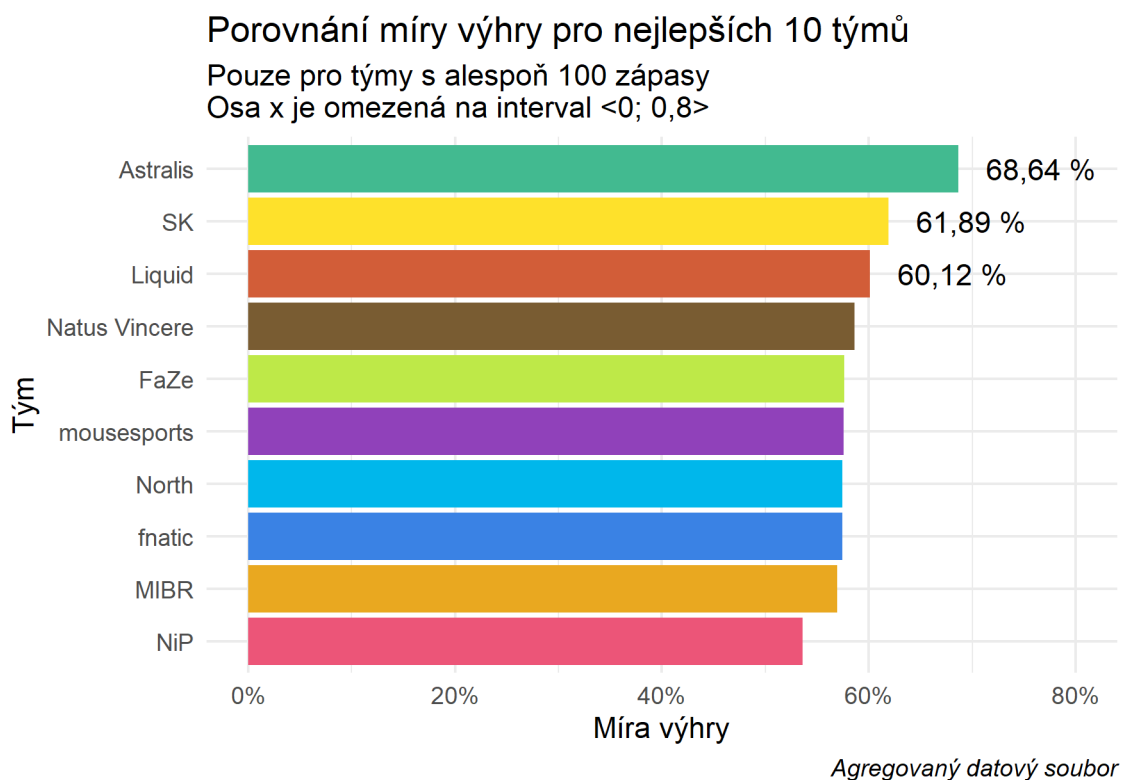
Obrázek 4.2: Histogram prediktorů

Histogramy prediktorů z obrázku 4.2 ukazují, že prediktory *kills*, *deaths* a *hs* mají normální rozdělení a v prediktorech se nenachází mnoho extrémních hodnot. Prediktor *fkdifff* má bimodální rozdělení. Prediktor *assists* je zešíkmení doprava, což značí velké množství odlehlých či extrémních hodnot. Pro logistickou regresi není předpoklad normálního rozdělení prediktorů, a proto analýza slouží čistě k získání povědomí o tom, jakých hodnot každý prediktor nabývá a jaké je jejich rozdělení.

### 4.3.3 Míra výhry pro nejlepší týmy

Pro výběr do logistického modelu je nutné vybrat referenční tým, který je považován za jeden z nejlepších.





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

Obrázek 4.3 ukazuje, že tým Astralis má v datovém souboru nejlepší míru výhry. Rozdíl mezi mírou výhry Astralis a týmu SK je necelých 7 procent. Z tohoto důvodu je tým Astralis použit při analýze dat jako jeden z referenčních týmů.

## 4.4 Predikce výhry hráčů na různých mapách

Cílem modelů je predikovat výhru zápasu pro jednotlivé hráče na různých mapách a identifikovat významné prediktory s nimi spojené. Model lze využít na profesionální úrovni k identifikaci toho, jaká charakteristika nejvíce přispívá k výhře či prohře hráče. Prediktory se týkají pouze výkonu jednotlivých hráčů, model tedy pro předpověď výhry hráče nepoužívá charakteristiky spoluhráčů. K modelování je použit spojený datový soubor, který je detailně popsán v sekci 4.2.3. Obecně by šlo model zapsat následovně:

$$P(1|X_{kills}, X_{assists}, X_{deaths}, X_{fkdiff}, X_{starting_{ct}}) = \frac{1}{1 + e^{-z}}$$
$$z = \beta_0 + \beta_1 * X_{kills} + \beta_2 * X_{assists} + \beta_3 * X_{deaths} +$$
$$+ \beta_4 * X_{fkdiff} + \beta_5 * X_{starting_{ct}}.$$

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, jaký prediktor hraje významnou roli při predikování výhry hráče na různé mapě.

### 4.4.1 Model pro mapu Mirage

Tabulka 4.2: Výstup z programu R pro model hráčů na mapě Mirage

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	2,3954	0,0599	39,9740	0,0000
kills	0,1807	0,0021	87,4819	0,0000
assists	0,3031	0,0047	64,8606	0,0000
deaths	-0,3661	0,0032	-115,9604	0,0000
hs	-0,1957	0,0634	-3,0854	0,0020
fkdiff	0,0171	0,0046	3,6881	0,0002
starting_ct	-0,1995	0,0199	-10,0442	0,0000

Tabulka 4.2 představuje model pro mapu Mirage se všemi významnými prediktory. Prediktory *kills*, *assists* a *fkdiff* šanci na výhru hráče zvyšují. Naopak prediktory *deaths* a *starting\_ct* šanci snižují. Model lze také zapsat jako přepis funkce.

$$P(1|X_{kills}, X_{assists}, X_{deaths}, X_{fkdiff}, X_{starting_{ct}}) = \frac{1}{1 + e^{-z}}$$
$$z = 2,3954 + 0,1807 * X_{kills} + 0,3031 * X_{assists} - 0,3661 * X_{deaths} -$$
$$- 0,1957 * X_{hs} + 0,0171 * X_{fkdiff} - 0,1995 * X_{starting_{ct}}$$
(4.1)

Ze získaných koeficientů lze získat změnu šance. Změnu šance s každým dalším zabitím lze získat umocnění parametru  $\beta_1 = 0,1807$  na konstantu  $e$ .

$$\begin{aligned} \text{Změna šance} &\sim e^{\beta_1} \\ &\sim e^{0,1807} \\ &\sim 1.1980 \end{aligned} \tag{4.2}$$

S každým dalším zabitým hráčem se zvyšuje šance na výhru hráče zhruba 1,2 krát. U prediktoru *deaths* je parametr  $\beta_3$  záporný. To znamená, že se šance na výhru snižuje.

$$\begin{aligned} \text{Změna šance} &\sim e^{\beta_3} \\ &\sim e^{-0,3661} \\ &\sim 0.6934 \end{aligned} \tag{4.3}$$

Parametr lze interpretovat tak, že s každou další hráčovou smrtí se jeho šance na výhru sníží o zhruba 31 %  $((1 - 0.6934) * 100)$ .

### Matice záměn pro mapu Mirage

Na predikci je použita validační podmnožina a model z tabulky 4.2.

Tabulka 4.3: Matice záměn pro model hráčů na mapě Mirage

	Původní pozitivní	Původní negativní
Pozitivní predikce	7037	1938
Negativní predikce	1618	6804

Tabulka 4.4: Vybrané statistiky z matice záměn pro model hráčů na mapě Mirage

statistika	hodnota
Přesnost	0,7956
Senzitivita	0,7783
Specifická	0,8131

Z tabulky 4.3 je vidět, že model predikoval správně 7 037 výher ( $\sim 81,31\%$ ) a 6 804 proher ( $\sim 77,83\%$ ). Celkově model určil správně 13 841 objektů ( $\sim 79,56\%$ ). Statistiky z matice záměn jsou zobrazené v tabulce 4.4.

#### 4.4.2 Model pro mapu Vertigo

Model pro mapu Vertigo má identické vstupní prediktory, podle kterých se predikuje proměnná *map\_winner*, jako model pro mapu Mirage.

Tabulka 4.5: Výstup z programu R pro model hráčů na mapě Vertigo

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	1,8170	0,2199	8,2645	0,0000
kills	0,1826	0,0081	22,6135	0,0000
assists	0,2983	0,0174	17,1061	0,0000
deaths	-0,3508	0,0120	-29,2434	0,0000
hs	0,2077	0,2500	0,8306	0,4062
fkdiff	0,0092	0,0180	0,5099	0,6101
starting_ct	-0,0021	0,0773	-0,0265	0,9788

Pro mapu Vertigo jsou v modelu nevýznamné parametry. Statisticky nevýznamné prediktory *hs*, *fkdiff* a *starting\_ct* jsou z modelu odebrány a model je znovu natrénován.

Tabulka 4.6: Výstup z programu R pro optimalizovaný model hráčů na mapě Vertigo

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	1,9090	0,1803	10,5872	0,0000
kills	0,1841	0,0074	24,8957	0,0000
assists	0,2981	0,0174	17,1172	0,0000
deaths	-0,3521	0,0116	-30,2299	0,0000

Pro hráče jsou na mapě Vertigo významné pouze prediktory *kills*, *assists* a *deaths*. Ostatní prediktory *hs*. Optimalizovaný model lze zapsat do následující rovnice:

$$P(1|X_{kills}, X_{assists}, X_{deaths}) = \frac{1}{1 + e^{-z}} \quad (4.4)$$

$$z = 1,9090 + 0,1841 * X_{kills} + 0,2981 * X_{assists} - 0,3521 * X_{deaths}$$

Každé hráčovo zabití vede ke zvýšení šance na výhru na mapě Vertigo o zhruba 20 %. Každá další hráčova smrt vede ke snížení šance na výhru o zhruba 30 %. Matice záměn i statistiky modelu jsou potom vyhodnocené na optimalizovaném modelu pomocí validačních dat.

Tabulka 4.7: Matice záměn pro model hráčů na mapě Vertigo

	Původní pozitivní	Původní negativní
Pozitivní predikce	455	113
Negativní predikce	131	411

Tabulka 4.8: Vybrané statistiky z matice záměn pro modelu hráčů na mapě Vertigo

statistika	hodnota
Přesnost	0,7802
Senzitivita	0,7844
Specifická	0,7765

Z tabulky 4.8 je vidět, že model predikoval správně 455 výher ( $\sim 77,65\%$ ) a 411 proher ( $\sim 78,44\%$ ). Celkově model určil správně 866 objektů ( $\sim 78,02\%$ ). Vybrané statistiky jsou vyobrazené v tabulce 4.8.

#### 4.4.3 Interpretace a porovnání modelů

Na mapě Mirage jsou pro hráče významné všechny prediktory. Nejvíce pozitivní vliv má na výhru hráče počet asistencí a největší negativní vliv na výhru hráče má pak počet smrtí. Významnost prediktorů *fkdiff* a *assists* naznačuje, že je důležitá souhra hráčů a zkušenější týmy mají na mapě výhodu. Pokud hráči hrají spolu, mohou se při zabíjení nepřátel doplňovat (prediktor *assists*). Zároveň je důležité, aby se v zápase hráči navzájem podporovali a mohli získat první zabití v daném kole (prediktor *fkdiff*). Pokud hráči začínají mapu Mirage na straně Counter-Terroristů, je jejich šance na výhru nižší.

Pro mapu Vertigo jsou významné pouze prediktory *kills*, *assists* a *deaths*. Pro hráče tedy není důležité, jak přesně střílí (prediktor *hs*), jak dobře se tým podporuje na začátku kola (prediktor *fkdiff*) ani na jaké straně hráč začíná (prediktor *starting\_ct*). Největší vliv na výhru zde má prediktor *assists*, největší vliv na výhru má pak prediktor *deaths*. Porovnání obou statistik obou modelů je pak v následující tabulce.

Tabulka 4.9: Porovnání statistik pro model hráčů na mapě Mirage a Vertigo

statistika	Mapa Mirage	Mapa Vertigo
Přesnost	0,7956	0,7802
Senzitivita	0,7783	0,7844
Specifická	0,8131	0,7765

Mapa Mirage se hodí nejvíce na predikci proher díky své vysoké specifické. Rozdíly mezi statistikami pro mapu Vertigo jsou pak nepatrné a pohybují se okolo 1 %. Největší rozdíl mezi modely je ve Specifické, která je pro model na mapě Mirage větší o zhruba 3,5 procentních bodů.

Modely, matice záměn i vybrané statistiky z matic záměn pro ostatní mapy jsou v příloze B. Zajímavé je, že pro všechny ostatní modely jsou významné všechny prediktory. To z mapy Vertigo, která má významné pouze 3 prediktory, dělá poněkud unikátní mapu. Z tohoto výsledku lze usoudit, že pro všechny mapy, které se hrají už nějakou dobu, je důležitý jak výkon hráče, tak sehranost týmu. Jelikož je mapa Vertigo nová, je na začátku svého vývoje a pro hráče je stále relativně neznámá. Jelikož je na mapě Vertigo méně významná sehranost, zkušenosti a spolupráce v týmu, měli by jí preferovat nové či semi-profesionální týmy. Mapu také mohou využít profesionální týmy, které spoléhají spíše na individuální výkon hráčů, než na týmové strategie.

## 4.5 Predikce výhry týmu

Cílem modelů je predikovat výhru týmu na základě agregovaných charakteristik hráčů za tým na mapě. Tento typ modelu se hodí zejména v sázkových kancelářích. Celkový model je také vhodný při predikci výhry týmu, pro který není dostatek dat. V tomto případě by nebylo možné vytvořit separátní model a predikci výhry lze tak určit pomocí jednoho celkového modelu. Díky referenčním modelům lze určit šanci, že určitý tým vyhraje. To umožní kancelářím stanovit výdělečný kurz pro zápas. Matematický přepis modelu je následující:

$$\begin{aligned}
 P(1|X_{mean\_kills}, X_{mean\_assists}, X_{mean\_deaths}, X_{mean\_hs}, X_{mean\_fkdiff}, X_{team\_rank}, \\
 X_{mapCache*starting\_ct}, X_{mapCobblestone*starting\_ct}, X_{mapDust2*starting\_ct}, X_{mapInferno*starting\_ct}, \\
 X_{mapMirage*starting\_ct}, X_{mapNuke*starting\_ct}, X_{mapOverpass*starting\_ct}, X_{mapTrain*starting\_ct}, \\
 X_{mapVertigo*starting\_ct}) = \frac{1}{1 + e^{-z}} \\
 z = \beta_0 + \beta_1 * X_{mean\_kills} + \beta_2 * X_{mean\_assists} + \beta_3 * X_{mean\_deaths} + \\
 + \beta_4 * X_{mean\_hs} + \beta_5 * X_{mean\_fkdiff} + \beta_6 * X_{team\_rank} + \\
 + \beta_7 * X_{mapCache*starting\_ct} + \beta_8 * X_{mapCobblestone*starting\_ct} + \\
 + \beta_9 * X_{mapDust2*starting\_ct} + \beta_{10} * X_{mapInferno*starting\_ct} + \\
 + \beta_{11} * X_{mapMirage*starting\_ct} + \beta_{12} * X_{mapNuke*starting\_ct} + \\
 + \beta_{13} * X_{mapOverpass*starting\_ct} + \beta_{14} * X_{mapTrain*starting\_ct} + \\
 + \beta_{15} * X_{mapVertigo*starting\_ct}.
 \end{aligned}
 \tag{4.5}$$

Prediktory *mean\_kills*, *mean\_assists*, *mean\_deaths*, *mean\_hs*, *mean\_fkdiff* jsou průměrné charakteristiky hráčů v týmu na dané mapě v daném zápase a lze očekávat, že šanci na výhru ovlivňují. Dále do modelu vstupuje prediktor *team\_rank*, který značí rank týmu v daném zápase. Model následně obsahuje interakci mezi proměnnou *map* a *starting\_ct*. Interakce je z toho důvodu, že vliv počáteční strany může být na každé mapě jiný. Přesněji jsou prediktory definované v sekci 4.2.4

#### 4.5.1 Celkový model

Tabulka 4.10: Výstup z programu R pro model týmu pro všechny týmy

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	0,3423	0,1645	2,0814	0,0374
mean_kills	1,3688	0,0144	95,2199	0,0000
mean_assists	0,1321	0,0160	8,2750	0,0000
mean_deaths	-1,3958	0,0141	-98,8032	0,0000
mean_hs	-0,4997	0,2511	-1,9900	0,0466
mean_fkdiff	-0,0741	0,0196	-3,7900	0,0002
team_rank	-0,0009	0,0003	-3,0148	0,0026
starting_ct:mapCache	-0,0240	0,0855	-0,2802	0,7793
starting_ct:mapCobblestone	-0,3211	0,0974	-3,2976	0,0010
starting_ct:mapDust2	-0,2051	0,0897	-2,2860	0,0223
starting_ct:mapInferno	-0,2993	0,0708	-4,2266	0,0000
starting_ct:mapMirage	-0,2362	0,0640	-3,6919	0,0002
starting_ct:mapNuke	-0,1404	0,0915	-1,5347	0,1248
starting_ct:mapOverpass	-0,4307	0,0793	-5,4315	0,0000
starting_ct:mapTrain	-0,2341	0,0741	-3,1602	0,0016
starting_ct:mapVertigo	0,2211	0,2231	0,9909	0,3217

Parametry celkového modelu se nachází v tabulce 4.10. První vytvořený model je sestavený na celém trénovacím datovém souboru. Pro model jsou významné všechny prediktory bez interakce, jmenovitě *mean\_kills*, *mean\_assists*, *mean\_deaths*, *mean\_hs*, *mean\_fkdiff* a *team\_rank*. Interakce mezi prediktory *map* a *starting\_ct* není významná u map Vertigo, Cache a Nuke.

$$\begin{aligned}
P(1|X_{mean\_kills}, X_{mean\_assists}, X_{mean\_deaths}, X_{mean\_hs}, X_{mean\_fkdiff}, X_{team\_rank}, \\
X_{mapCache*starting\_ct}, X_{mapCobblestone*starting\_ct}, X_{mapDust2*starting\_ct}, X_{mapInferno*starting\_ct}, \\
X_{mapMirage*starting\_ct}, X_{mapNuke*starting\_ct}, X_{mapOverpass*starting\_ct}, X_{mapTrain*starting\_ct}, \\
X_{mapVertigo*starting\_ct}) = \frac{1}{1 + e^{-z}} \\
z = 0,3423 + 1,376 * X_{mean\_kills} + 0,1321 * X_{mean\_assists} - 1,3958 * X_{mean\_deaths} - \\
- 0,4997 * X_{mean\_hs} - 0,0741 * X_{mean\_fkdiff} - 0,0009 * X_{team\_rank} - \\
- 0,0240 * X_{mapCache*starting\_ct} - 0,3211 * X_{mapCobblestone*starting\_ct} - \\
- 0,2051 * X_{mapDust2*starting\_ct} - 0,2993 * X_{mapInferno*starting\_ct} - \\
- 0,2362 * X_{mapMirage*starting\_ct} - 0,1404 * X_{mapNuke*starting\_ct} - \\
- 0,4307 * X_{mapOverpass*starting\_ct} - 0,2341 * X_{mapTrain*starting\_ct} + \\
+ 0,2211 * X_{mapVertigo*starting\_ct}
\end{aligned}
\tag{4.6}$$



Agregované charakteristiky hráče *mean\_kills* a *mean\_assists* šanci na výhru týmu zvyšují. Pokud se průměr zabitých nepřátel za tým zvýší o jednotku, šance na výhru týmu se zvýší zhruba 3,93 krát. Pokud se průměr smrtí hráčů za tým zvýší o jednotku, šance na výhru se sníží o zhruba 75 %. Všechny statisticky významné interakce mezi prediktory *map* a *starting\_ct* naznačují, že je pro tým nevýhodné začínat na straně Counter-Terroristů. Jejich šance na výhru se vždy sníží, a to nejvíce na mapě Overpass, kde je zhoršení o zhruba 35 %.

## Matice záměn pro obecný model

Tabulka 4.11: Matice záměn pro model týmu pro všechny týmy

	Původní pozitivní	Původní negativní
Pozitivní predikce	8321	517
Negativní predikce	459	8337

Tabulka 4.12: Vybrané statistiky z matice záměn pro model týmu pro všechny týmy

statistika	hodnota
Přesnost	0,9447
Senzitivita	0,9416
Specifická	0,9477

Dle matice 4.11 model úspěšně predikoval 8 321 výher ( $\sim 94,77\%$ ) a 8 337 proher ( $\sim 94,16\%$ ). Celkem model predikoval správně 16 658 objektů ( $\sim 94,47\%$ ). Všechny výkonnostní statistiky z matice záměn jsou identické v řádu setin procent. Z tabulky 4.12 nelze jednoznačně určit, zda je tým vhodnější na predikci výher či proher.

### 4.5.2 Model pro tým Astralis

Tabulka 4.13: Výstup z programu R pro optimalizovaný pro model týmu pro tým Astralis

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	1,5840	0,9507	1,6661	0,0957
mean_kills	1,3331	0,1328	10,0345	0,0000
mean_deaths	-1,4069	0,1380	-10,1935	0,0000

Po odstranění nevýznamných prediktorů lze získat logistický model, který je popsán v tabulce 4.13. Pro tým Astralis jsou významné pouze prediktory *mean\_kills* a *mean\_deaths*.

$$P(1|X_{mean\_kills}, X_{mean\_deaths}, X_{team\_rank}) = \frac{1}{1 + e^{-z}} \quad (4.7)$$

$$z = 1,9185 + 1,3452 * X_{mean\_kills} - 1,4180 * X_{mean\_deaths} - 0,1120 * X_{team\_rank}$$

S každým dalším průměrným zabitím (prediktor *mean\_kills*) se zvýší šance týmu Astralis na výhru zhruba 3,84 krát. S každou další průměrnou smrtí (prediktor *mean\_deaths*) se šance na výhru týmu sníží o zhruba 76 %. Před začátkem zápasu, tedy při nulových prediktorech *mean\_kills* a *mean\_deaths*, je šance na výhru týmu Astralis zhruba 6,81 větší, než jeho prohra.

### Matice záměn pro tým Astralis

Tabulka 4.14: Matice záměn pro model týmu pro tým Astralis

	Původní pozitivní	Původní negativní
Pozitivní predikce	44	7
Negativní predikce	6	118

Tabulka 4.15: Vybrané statistiky z matice záměn pro model týmu pro tým Astralis

statistika	hodnota
Přesnost	0,9257
Senzitivita	0,9440
Specifická	0,8800

Z matice záměn v tabulce 4.14 lze vidět, že model predikoval správně 44 výher ( $\sim 88,00\%$ ) a 118 proher ( $\sim 94,40\%$ ). Celkem optimální model predikoval správně 162 objektů ( $\sim 92,57\%$ ). Statistiky jsou zobrazené v tabulce 4.15.

### 4.5.3 Model pro tým Sprout

Tým Sprout byl v době extrakce dat čistě německý tým a patřil k průměrným profesionálnímu týmům. Na žebříčku týmů se obvykle řadil kolem třicátého místa.

Tabulka 4.16: Výstup z programu R pro optimalizovaný pro model týmu pro tým Sprout

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-2,7474	1,3338	-2,0598	0,0394
mean_kills	1,5741	0,1887	8,3426	0,0000
mean_deaths	-1,4307	0,1670	-8,5664	0,0000

Tabulka 4.16 obsahuje pouze významné prediktory pro model týmu Sprout. Významé prediktory pro tým jsou pouze *mean\_kills* a *mean\_deaths*. Matematický zápis modelu je v následujícím přepisu.

$$P(1|X_{mean\_kills}, X_{mean\_deaths}, X_{team\_rank}) = \frac{1}{1 + e^{-z}} \quad (4.8)$$

$$z = -2,7474 + 1,5741 * X_{mean\_kills} - 1,4307 * X_{mean\_deaths}$$

S každým dalším průměrným zabitím (prediktor *mean\_kills*) se šance na výhru týmu zvýší zhruba 4,83 krát. S každou další průměrnou smrtí (prediktor *mean\_deaths*) se šance na výhru sníží o zhruba 76 %. Ještě před začátkem zápasu, tedy při nulovém průměrném počtu zabití a smrtí, je šance na výhru týmu Sprout zhruba 6 %.

Tabulka 4.17: Matice záměn pro model týmu pro tým Sprout

	Původní pozitivní	Původní negativní
Pozitivní predikce	47	1
Negativní predikce	4	71

Tabulka 4.18: Vybrané statistiky z matice záměn pro model týmu pro tým Sprout

statistika	hodnota
Přesnost	0,9593
Senzitivita	0,9861
Specificita	0,9216

Z matice záměn v tabulce 4.17 lze vidět, že model predikoval správně 47 výher ( $\sim 92,16\%$ ) a 71 proher ( $\sim 98,61\%$ ). Celkem optimální model predikoval správně 118 objektů ( $\sim 95,93\%$ ). Statistiky jsou zobrazené v tabulce 4.18.

#### 4.5.4 Interpretace a porovnání modelů

Pro celkový model, tedy model pro predikci výhry jakéhokoliv týmu, jsou významné všechny průměrné charakteristiky hráčů. Nejvíce šanci na výhru zvyšuje každé další průměrné zabití. Šanci na prohru pak nejvíce zvyšuje každá další průměrná smrt. Z modelu lze usoudit, že při určování šance výhry na určitých mapách je významné, na jaké straně tým začíná. Pokud tým začíná na straně Counter-Terroristů, jeho šance na výhru je na mapách Cobblestone, Dust 2, Inferno, Mirage, Overpass a Train nižší. Na mapách Cache, Nuke a Vertigo je pak začínající strana nevýznamná. Statistiky z matice záměn jsou velmi podobné a proto se model hodí jak na predikci výhry, tak prohry. Na rozdíl od modelů pro tým Astralis a Sprout je model velmi komplexní a obsahuje 15 prediktorů. To lze odůvodnit tím, že je model ovlivněn neprofesionálními týmy, pro které je mnoho prediktorů významných z důvodu nezkušenosti. Nezkušené týmy mezi sebou nejsou vůbec vyrovnané a proto je pro ně důležité získat jakoukoli výhodu proti nepřátelskému týmu. Profesionální týmy jsou naopak velmi vyrovnané a proto je nejdůležitější to, jak dobře budou hráči týmu hrát.

Model pro tým Astralis naznačuje, že je pro tým významný pouze průměrný počet zabití a průměrný počet úmrtí. Výhru týmu neovlivňuje průměrný počet asistencí, průměrné procento zabití do hlavy, průměrný rozdíl prvního zabití a úmrtí, rank týmu v okamžik zápasu a počáteční strana na jakékoliv mapě. Jelikož se tým Astralis řadí mezi jedny z nejlepších týmů na světě, jeho šance na výhru před začátkem zápasu je poměrně vysoká. To potvrzuje jak logistický model, tak průzkumová analýza dat. Vytvořený model pro tým Astralis je vhodný spíše na predikci výhry než prohry.

Významné prediktory pro tým Sprout jsou pouze průměrný počet zabití a průměrný počet smrtí. Výhru týmu tedy neovlivňuje průměrný počet asistencí, průměrné procento zabití do hlavy, průměrný rozdíl prvního zabití a úmrtí, rank týmu v okamžik zápasu a počáteční strana na jakékoliv mapě. Tým se nepovažuje za jeden z nejlepších na světě, a proto má před začátkem zápasu větší šanci na prohru. Vliv průměrného počtu zabití je větší než vliv průměrného počtu smrtí a proto by měl tým více riskovat a hrát spíše na velký počet průměrného zabití. Tento model je nejvíce vhodný na predikci výhry týmu a méně vhodný na predikci prohry týmu.

Tabulka 4.19: Porovnání statistik pro týmové modely

statistika	Celkový model	Tým Astralis	Tým Sprout
Přesnost	0,9447	0,9257	0,9593
Senzitivita	0,9416	0,9440	0,9861
Specifická	0,9477	0,8800	0,9216

Oba dva referenční modely mají stejné významné prediktory, a to *mean\_kills* a *mean\_deaths*. Z modelů lze usoudit, že nejlepší tým na světě a průměrný profesionální tým má stejné významné prediktory. Tabulka 4.19 ukazuje porovnání mezi jednotlivými statistikami přes vybrané modely. Model pro tým Sprout má největší přesnost i senzitivitu, zatímco celkový model má největší specifitu. Ze statistik vychází, že nejméně spolehlivě předpovídá model pro referenční tým Astralis.

Rozdíl mezi referenčními modely a celkovým modelem by bylo možné vysvětlit např. vlivem neprofesionálních týmů. Amatérské týmy musí využít každé možné výhody, kterou mohou získat, aby vyhráli. Z tohoto důvodu lze usoudit, že je pro ně mnoho prediktorů významných. Když proti sobě hrají profesionální týmy, jejich rozdíl je méně patrný a je proto nejvíce signifikantní výkon jednotlivých hráčů.

## 5. Závěr

Esport neustále roste a rozvíjí se do odvětví jako sázení, analytika či podnikání. Jedná se o nový trend, který zajímá nejvíce mladou populaci a hráče her. Díky propojení esportu a sázkových kanceláří vzniká potřeba pro kvalitní a přesné modely, díky kterým lze výsledky zápasů předpovídat a stanovovat tak vhodné kurzy. V profesionální scéně lze modely využít na identifikaci významných prediktorů pro výhru zápasu. Hráči a týmy se pak mohou pomoci dat a modelů mohou soustředit pouze na charakteristiky, které jejich šanci na výhru zvyšují.

Vytvořené modely v bakalářské práci byli sestavené jak pro individuální hráče, tak pro dva referenční týmy a pro všechny týmy v datovém souboru. Z modelů je patrné, že nejdůležitější prediktory se týkají počtu zabití nepřátel a počtu úmrtí hráče. Ty hrají významnou roli jak v modelech pro hráče, tak pro tým. Rozdíl ve významnosti prediktorů se nachází mezi profesionálními týmy a amatérskými týmy. Z modelů lze odvodit důraz na individuální charakteristiky hráčů v profesionálních týmech. V týmech amatérských lze pak pozorovat důraz na různé charakteristiky hráčů a na charakteristiky zápasu.

Model by bylo možné rozšířit více podrobnými a specifickými charakteristikami hráčů. Modely zároveň nepočítají s tím, že se hráči v týmu mohou měnit a každý hráč má jiné charakteristiky a délku hraní v týmu. Předpoklad pro aktuální modely je takový, že je známé, na jaké mapě se hraje a jaký tým začíná na jaké straně. V praxi je tento fakt známý až velmi krátce před začátkem zápasu. K modelům by bylo proto vhodné přidat predikci mapy, která se bude mezi dvěma týmy hrát.

# Seznam použitého softwaru

RSTUDIO, 2022. *RStudio / Open source & professional software for data science teams* [online]. Osobní počítač: RStudio. 2022.02.2+485 [cit. 2022-04-30]. Dostupné z: <https://www.rstudio.com/>.

TEAM, R Core, 2022. *R: The R Project for Statistical Computing* [online]. Osobní počítač: R Project. Ver. 4.2.0 [cit. 2022-04-30]. Dostupné z: <https://www.r-project.org/>.

# Seznam použité literatury

- CSKO.CS, 2022. *Kotelna - Contents* [online] [cit. 2022-04-24]. Dostupné z: <https://stats.csko.cz/statsx/hlstats.php>.
- GOUGH, Christina, 2022. *Global eSports market revenue 2024* [Statista] [online] [cit. 2022-04-24]. Dostupné z: <https://www.statista.com/statistics/490522/global-esports-market-revenue/>.
- HÄRDLE, Wolfgang; SIMAR, Léopold, 2015. *Applied multivariate statistical analysis*. 4. ed. Heidelberg Berlin: Springer. ISBN 978-3-662-45170-0.
- HENNINGSON, Joakim, 2020. *The history of Counter-Strike* [Red Bull] [online] [cit. 2022-04-24]. Dostupné z: <https://www.redbull.com/se-en/history-of-counterstrike>.
- KLEINBAUM, David G.; KLEIN, Mitchel, 2010. *Logistic regression: a self-learning text*. Third Edition. Ve spol. s RIHL PRYOR, Erica. New York Dordrecht Heidelberg London: Springer. Statistics for Biology and Health. ISBN 978-1-4939-3697-7.
- LARCH, Florian, 2022. *History of eSports: How it all began* [online] [cit. 2022-04-24]. Dostupné z: <https://www.ispo.com/en/markets/history-origin-esports>.
- LIQUIPEDIA.NET, 2021. *PGL Major Stockholm 2021* [Liquipedia Counter-Strike Wiki] [online] [cit. 2022-04-24]. Dostupné z: <https://liquipedia.net/counterstrike/PGL/2021/Stockholm>.
- PROFESSEUR, 2015. *HLTV.org - The home of competitive Counter-Strike* [HLTV.org] [online] [cit. 2022-04-24]. Dostupné z: <https://www.hltv.org/matches/2295340/xenex-vs-excel-esl-uk-premiership-season-1>.
- PROFESSEUR, 2022. *ESEA increase prize pool and number of seasons for 2021; simplify path to Pro League* [HLTV.org] [online] [cit. 2022-04-24]. Dostupné z: <https://www.hltv.org/news/30926/esea-increase-prize-pool-and-number-of-seasons-for-2021-simplify-path-to-pro-league>.
- RSTUDIO, 2022. *RStudio / Open source & professional software for data science teams* [online]. Osobní počítač: RStudio. 2022.02.2+485 [cit. 2022-04-30]. Dostupné z: <https://www.rstudio.com/>.
- TEAM, R Core, 2022. *R: The R Project for Statistical Computing* [online]. Osobní počítač: R Project. Ver. 4.2.0 [cit. 2022-04-30]. Dostupné z: <https://www.r-project.org/>.
- VALVE, 2013. *Counterstrike: Global Offensive - Arms Deal* [online] [cit. 2022-04-24]. Dostupné z: <http://counter-strike.net/armsdeal>.



# Seznam elektronických zdrojů

- CSKO.CS, 2022. *Kotelna - Contents* [online] [cit. 2022-04-24]. Dostupné z: <https://stats.csko.cz/statsx/hlstats.php>.
- GOUGH, Christina, 2022. *Global eSports market revenue 2024* [Statista] [online] [cit. 2022-04-24]. Dostupné z: <https://www.statista.com/statistics/490522/global-esports-market-revenue/>.
- HENNINGSON, Joakim, 2020. *The history of Counter-Strike* [Red Bull] [online] [cit. 2022-04-24]. Dostupné z: <https://www.redbull.com/se-en/history-of-counterstrike>.
- LARCH, Florian, 2022. *History of eSports: How it all began* [online] [cit. 2022-04-24]. Dostupné z: <https://www.ispo.com/en/markets/history-origin-esports>.
- LIQUIPEDIA.NET, 2021. *PGL Major Stockholm 2021* [Liquipedia Counter-Strike Wiki] [online] [cit. 2022-04-24]. Dostupné z: <https://liquipedia.net/counterstrike/PGL/2021/Stockholm>.
- PROFESSEUR, 2015. *HLTV.org - The home of competitive Counter-Strike* [HLTV.org] [online] [cit. 2022-04-24]. Dostupné z: <https://www.hltv.org/matches/2295340/xenex-vs-excel-esl-uk-premiership-season-1>.
- PROFESSEUR, 2022. *ESEA increase prize pool and number of seasons for 2021; simplify path to Pro League* [HLTV.org] [online] [cit. 2022-04-24]. Dostupné z: <https://www.hltv.org/news/30926/esea-increase-prize-pool-and-number-of-seasons-for-2021-simplify-path-to-pro-league>.
- VALVE, 2013. *Counterstrike: Global Offensive - Arms Deal* [online] [cit. 2022-04-24]. Dostupné z: <http://counter-strike.net/armsdeal>.

# Seznam obrázků

3.1	Bodový graf hmotnosti a míly za galon . . . . .	8
3.2	Sloupcový graf počtu válců a průměrné hrubé koňské síly . . . . .	9
3.3	Porovnání histogramů s různým počtem skupin . . . . .	10
3.4	Histogram s počtem skupin dle Sturgesova pravidla . . . . .	11
3.5	Krabičkový graf hmotnosti auta pro různý počet válců . . . . .	13
3.6	Korelační matice . . . . .	14
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	pětičíselné shrnutí hmotnosti vozidla (lb/1000) . . . . .	12
3.2	Matice záměn . . . . .	17
4.1	Přehled agregací charakteristik pro daný tým na dané mapě v daném zápase . .	23
4.2	Výstup z programu R pro model hráčů na mapě Mirage . . . . .	28
4.3	Matice záměn pro model hráčů na mapě Mirage . . . . .	29
4.4	Vybrané statistiky z matice záměn pro model hráčů na mapě Mirage . . . . .	29
4.5	Výstup z programu R pro model hráčů na mapě Vertigo . . . . .	30
4.6	Výstup z programu R pro optimalizovaný model hráčů na mapě Vertigo . . . . .	30
4.7	Matice záměn pro model hráčů na mapě Vertigo . . . . .	30
4.8	Vybrané statistiky z matice záměn pro modelu hráčů na mapě Vertigo . . . . .	31
4.9	Porovnání statistik pro model hráčů na mapě Mirage a Vertigo . . . . .	31
4.10	Výstup z programu R pro model týmu pro všechny týmy . . . . .	34
4.11	Matice záměn pro model týmu pro všechny týmy . . . . .	35
4.12	Vybrané statistiky z matice záměn pro model týmu pro všechny týmy . . . . .	35
4.13	Výstup z programu R pro optimalizovaný pro model týmu pro tým Astralis . . .	35
4.14	Matice záměn pro model týmu pro tým Astralis . . . . .	36
4.15	Vybrané statistiky z matice záměn pro model týmu pro tým Astralis . . . . .	36
4.16	Výstup z programu R pro optimalizovaný pro model týmu pro tým Sprout . . . .	36
4.17	Matice záměn pro model týmu pro tým Sprout . . . . .	37
4.18	Vybrané statistiky z matice záměn pro model týmu pro tým Sprout . . . . .	37
4.19	Porovnání statistik pro týmové modely . . . . .	38
A.1	Záznam z transformovaného datového souboru players.csv . . . . .	49
A.2	Příklad záznamu z transformovaného datového souboru results.csv . . . . .	49
A.3	Příklad záznamu z transformovaného datového souboru results.csv . . . . .	49
A.4	Agregovaná data pro týmy za zápas a mapu . . . . .	50
B.1	Výstup z programu R pro model hráčů na mapě Cache . . . . .	51
B.2	Matice záměn pro model hráčů na mapě Cache . . . . .	51
B.3	Vybrané statistiky z matice záměn pro model hráčů na mapě Cache . . . . .	51
B.4	Výstup z programu R pro model hráčů na mapě Cobblestone . . . . .	52
B.5	Matice záměn pro model hráčů na mapě Cobblestone . . . . .	52
B.6	Vybrané statistiky z matice záměn pro model hráčů na mapě Cobblestone . . . .	52
B.7	Výstup z programu R pro model hráčů na mapě Dust2 . . . . .	53
B.8	Matice záměn pro model hráčů na mapě Dust2 . . . . .	53
B.9	Vybrané statistiky z matice záměn pro model hráčů na mapě Dust2 . . . . .	53
B.10	Výstup z programu R pro model hráčů na mapě Inferno . . . . .	54
B.11	Matice záměn pro model hráčů na mapě Inferno . . . . .	54

B.12 Vybrané statistiky z matice záměn pro model hráčů na mapě Inferno . . . . .	54
B.13 Výstup z programu R pro model hráčů na mapě Nuke . . . . .	55
B.14 Matice záměn pro model hráčů na mapě Nuke . . . . .	55
B.15 Vybrané statistiky z matice záměn pro model hráčů na mapě Nuke . . . . .	55
B.16 Výstup z programu R pro model hráčů na mapě Overpass . . . . .	56
B.17 Matice záměn pro model hráčů na mapě Overpass . . . . .	56
B.18 Vybrané statistiky z matice záměn pro model hráčů na mapě Overpass . . . . .	56
B.19 Výstup z programu R pro model hráčů na mapě Train . . . . .	57
B.20 Matice záměn pro model hráčů na mapě Train . . . . .	57
B.21 Vybrané statistiky z matice záměn pro model hráčů na mapě Train . . . . .	57

# 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
2309869	9859	EURONICS	Mirage	13	4	19	0,5385	-2	0,7800
2322900	15631	FURIA	Nuke	18	1	14	0,5000	2	1,1400
2304505	7205	DarkPassage	Cache	15	3	21	0,2667	-1	0,8400
2301029	3997	PENTA	Cache	21	3	17	0,4762	1	1,2800
2333140	10264	OpTic	Dust2	5	1	16	0,4000	0	0,3800
2320084	9032	Astralis	Overpass	16	9	21	0,5625	-1	1,0600

## 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
2017-03-29	2309175	New4	Inferno	0	0	123
2017-09-09	2314493	FaZe	Inferno	1	1	8
2019-04-22	2332676	Epsilon	Mirage	1	0	22
2019-07-29	2335195	North	Nuke	1	1	12
2019-10-30	2337316	Windigo	Vertigo	0	1	40
2019-10-24	2337128	Evil Geniuses	Dust2	1	1	3

## A.3 Spojený datový soubor

Tabulka A.3: Příklad záznamu z transformovaného datového souboru results.csv

date	match_id	team	map	map_winner	starting_ct	team_rank	player_id	k
2017-05-17	2310786	ENCE	Cache	0	1	214	684	
2016-07-24	2303591	Chiefs	Dust2	0	0	88	9636	
2019-03-23	2331595	Liquid	Overpass	1	0	2	8738	
2019-12-22	2338373	Vitality	Mirage	1	1	8	8184	
2019-12-09	2338387	OFFSET	Inferno	1	0	73	11205	
2020-01-24	2339089	AUGUST	Dust2	0	0	114	10870	

## A.4 Agregovaný datový soubor

Tabulka A.4: Agregovaná data pro týmy za zápas a mapu

match_id	map	team	mean_kills	mean_assists	mean_deaths	mean_hs	mean_fkdif	map_winner	starting_ct	team_rank
2302745	Train	Virtus.pro	10,8000	2,0000	19,8000	0,3860	-2,0000	0,0000	0,0000	6,0000
2306295	Train	Natus Vincere	14,6000	2,2000	18,4000	0,3823	-0,8000	0,0000	1,0000	4,0000
2338225	Dust2	PC419	13,0000	2,4000	19,8000	0,4057	0,0000	0,0000	0,0000	172,0000
2334630	Dust2	Nordavind	16,4000	5,0000	20,4000	0,5492	0,6000	0,0000	1,0000	54,0000
2313618	Train	Adaptation	14,8000	4,0000	19,4000	0,4322	-1,4000	0,0000	0,0000	121,0000
2333754	Vertigo	Singularity	17,4000	4,8000	7,0000	0,4448	2,0000	1,0000	1,0000	52,0000



## B. Modely, matice záměn a statistiky pro individuální hráče

### B.1 Mapa Cache

Tabulka B.1: Výstup z programu R pro model hráčů na mapě Cache

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	2,4497	0,0826	29,6427	0,0000
kills	0,1808	0,0029	62,6847	0,0000
assists	0,3007	0,0064	46,9248	0,0000
deaths	-0,3727	0,0045	-83,5807	0,0000
hs	-0,2503	0,0901	-2,7771	0,0055
fkdiff	0,0308	0,0065	4,7475	0,0000
starting_ct	-0,0441	0,0278	-1,5863	0,1127

Tabulka B.2: Matice záměn pro model hráčů na mapě Cache

	Původní pozitivní	Původní negativní
Pozitivní predikce	3733	989
Negativní predikce	804	3464

Tabulka B.3: Vybrané statistiky z matice záměn pro model hráčů na mapě Cache

statistika	hodnota
Přesnost	0,8006
Senzitivita	0,7779
Specifická	0,8228

## B.2 Mapa Cobblestone

Tabulka B.4: Výstup z programu R pro model hráčů na mapě Cobblestone

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	2,7568	0,0981	28,1043	0,0000
kills	0,1916	0,0034	56,1622	0,0000
assists	0,2972	0,0073	40,4559	0,0000
deaths	-0,3955	0,0054	-73,9026	0,0000
hs	-0,5199	0,1065	-4,8835	0,0000
fkdiff	0,0137	0,0076	1,8163	0,0693
starting_ct	-0,0165	0,0324	-0,5093	0,6105

Tabulka B.5: Matice záměn pro model hráčů na mapě Cobblestone

	Původní pozitivní	Původní negativní
Pozitivní predikce	2809	769
Negativní predikce	630	2622

Tabulka B.6: Vybrané statistiky z matice záměn pro model hráčů na mapě Cobblestone

statistika	hodnota
Přesnost	0,7952
Senzitivita	0,7732
Specifická	0,8168

## B.3 Mapa Dust2

Tabulka B.7: Výstup z programu R pro model hráčů na mapě Dust2

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	2,1869	0,0886	24,6769	0,0000
kills	0,1806	0,0031	58,3519	0,0000
assists	0,2608	0,0063	41,5863	0,0000
deaths	-0,3664	0,0047	-77,1522	0,0000
hs	0,2304	0,0921	2,5016	0,0124
fkdiff	-0,0090	0,0069	-1,3053	0,1918
starting_ct	-0,1337	0,0293	-4,5554	0,0000

Tabulka B.8: Matice záměn pro model hráčů na mapě Dust2

	Původní pozitivní	Původní negativní
Pozitivní predikce	3142	934
Negativní predikce	754	2988

Tabulka B.9: Vybrané statistiky z matice záměn pro model hráčů na mapě Dust2

statistika	hodnota
Přesnost	0,7841
Senzitivita	0,7619
Specificita	0,8065

## B.4 Mapa Inferno

Tabulka B.10: Výstup z programu R pro model hráčů na mapě Inferno

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	2,3654	0,0661	35,7613	0,0000
kills	0,1845	0,0023	78,5946	0,0000
assists	0,2848	0,0047	60,7672	0,0000
deaths	-0,3772	0,0036	-106,0105	0,0000
hs	-0,2167	0,0701	-3,0896	0,0020
fkdiff	0,0208	0,0052	4,0137	0,0001
starting_ct	-0,2086	0,0220	-9,4669	0,0000

Tabulka B.11: Matice záměn pro model hráčů na mapě Inferno

	Původní pozitivní	Původní negativní
Pozitivní predikce	5837	1660
Negativní predikce	1304	5618

Tabulka B.12: Vybrané statistiky z matice záměn pro model hráčů na mapě Inferno

statistika	hodnota
Přesnost	0,7944
Senzitivita	0,7719
Specifická	0,8174

## B.5 Mapa Nuke

Tabulka B.13: Výstup z programu R pro model hráčů na mapě Nuke

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	2,6884	0,0899	29,9096	0,0000
kills	0,1903	0,0031	61,1310	0,0000
assists	0,3234	0,0075	43,3763	0,0000
deaths	-0,3862	0,0048	-79,8802	0,0000
hs	-0,3237	0,0973	-3,3261	0,0009
fkdiff	0,0146	0,0070	2,1036	0,0354
starting_ct	-0,2529	0,0297	-8,5059	0,0000

Tabulka B.14: Matice záměn pro model hráčů na mapě Nuke

	Původní pozitivní	Původní negativní
Pozitivní predikce	3336	880
Negativní predikce	691	3107

Tabulka B.15: Vybrané statistiky z matice záměn pro model hráčů na mapě Nuke

statistika	hodnota
Přesnost	0,8040
Senzitivita	0,7793
Specifická	0,8284

## B.6 Mapa Overpass

Tabulka B.16: Výstup z programu R pro model hráčů na mapě Overpass

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	2,5688	0,0756	33,9813	0,0000
kills	0,1847	0,0027	69,6590	0,0000
assists	0,2877	0,0055	52,1250	0,0000
deaths	-0,3800	0,0041	-92,3994	0,0000
hs	-0,1816	0,0824	-2,2041	0,0275
fkdiff	0,0216	0,0059	3,6595	0,0003
starting_ct	-0,3085	0,0253	-12,1919	0,0000

Tabulka B.17: Matice záměn pro model hráčů na mapě Overpass

	Původní pozitivní	Původní negativní
Pozitivní predikce	4387	1209
Negativní predikce	1035	4246

Tabulka B.18: Vybrané statistiky z matice záměn pro model hráčů na mapě Overpass

statistika	hodnota
Přesnost	0,7937
Senzitivita	0,7784
Specifická	0,8091

## B.7 Mapa Train

Tabulka B.19: Výstup z programu R pro model hráčů na mapě Train

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	2,4844	0,0697	35,6607	0,0000
kills	0,1854	0,0025	75,5635	0,0000
assists	0,3137	0,0056	55,6218	0,0000
deaths	-0,3768	0,0038	-99,6224	0,0000
hs	-0,1471	0,0736	-1,9980	0,0457
fkdiff	0,0156	0,0055	2,8470	0,0044
starting_ct	-0,2711	0,0236	-11,5127	0,0000

Tabulka B.20: Matice záměn pro model hráčů na mapě Train

	Původní pozitivní	Původní negativní
Pozitivní predikce	5194	1323
Negativní predikce	1130	5060

Tabulka B.21: Vybrané statistiky z matice záměn pro model hráčů na mapě Train

statistika	hodnota
Přesnost	0,8070
Senzitivita	0,7927
Specificita	0,8213

## **C. Kompletní kód při tvorbě praktické části bakalářské práce**



# Analýza dat pro Bakalářskou práci

Michal Lauer

2022-05-01

## Funkce a globální nastavení

Nastavení generování bloků kódu a výstupu z print funkcí, načtení knihoven, a načtení funkcí.

### Nastavení

```
knitr::opts_chunk$set(message = F)

options("OutDec" = ",")
options("xtable.include.rownames" = F)
options("xtable.table.placement" = "H")
options("xtable.caption.placement" = "top")
options("xtable.format.args" = list(decimal.mark = ","))
options("digits" = 4)
```

### Knihovny

```
library(readr)
library(dplyr)
library(stringr)
library(tidyr)
library(ggplot2)
library(GGally)
library(caret)
library(xtable)
library(rsample)
library(palettter)

theme_set(theme_minimal() +
  theme(
    plot.caption = element_text(face = "italic")
  ))
```

### Globální proměnné

```

# Seznam map, které jsou potřebné pro analýzy dat
maps <- c(
  "Dust2", "Inferno", "Vertigo", "Overpass", "Nuke", "Mirage",
  "Train", "Cache", "Cobblestone"
)

# Adresář, kam ukládat grafy
PICTURES_DIR <- "../bakalarska-prace/obrazky/"
if (!dir.exists(PICTURES_DIR)) dir.create(PICTURES_DIR)

# Adresář, kam ukládat vytvořené latex tabulky
TABLES_DIR <- "../bakalarska-prace/tabulky/"
if (!dir.exists(TABLES_DIR)) dir.create(TABLES_DIR)

# Adresář, kam ukládat tabulky vytvořených modelů
MODELS_DIR <- "../bakalarska-prace/modely/"
if (!dir.exists(MODELS_DIR)) dir.create(MODELS_DIR)

# Adresář, kam ukládat latex tabulky pro matice záměn
MATRIXES_DIR <- "../bakalarska-prace/matice/"
if (!dir.exists(MATRIXES_DIR)) dir.create(MATRIXES_DIR)

# Adresář, kam ukládat latex tabulky pro výsledné statistiky z
# matice záměn
STATS_DIR <- "../bakalarska-prace/statistiky/"
if (!dir.exists(STATS_DIR)) dir.create(STATS_DIR)

```

## Uložení grafu

```

#' Uložení grafu jako png
#'
#' @param .graph Graf z knihovny ggplot
#' @param .file Soubor, do kterého obrázek uložit
save_bc_picture <- function(.graph, .file) {

  # Vytvoření cesty pro soubor
  .file = paste(PICTURES_DIR, .file, sep = "/")

  # Uložení obrázku
  ggsave(
    .file,
    plot = .graph,
    bg = "white",
    width = 150,
    height = 105,
    units = 'mm'
  )
}

```

## Uložení tabulky

```
#' Uložení libovolné tabulky do LaTeX souboru
#'  
#' @param .tibble Tabulka  
#' @param .file Soubor, do kterého uložit tabulku  
#' @param .cap Caption  
#' @param .lab Label  
#' @param .scale Poměr tabulky  
save_bc_table <- function(.tibble, .file, .cap, .lab,  
                           .scale = 1) {  
  
  # Pokud není zadáný label, použije se jako label název souboru  
  # bez přípony  
  if (is.na(.lab)) {  
    .lab <- sprintf("tab:%s", str_extract(.file, "^.*(?:\\.tex)"))  
  }  
  
  # Protože xtable neumí vypsat datum, převede se na text  
  if ("date" %in% colnames(.tibble)) {  
    .tibble <- .tibble |>  
      mutate(  
        across(  
          .cols = "date",  
          .fns = as.character))  
  }  
  
  # Vytvoření pojmenované tabulky  
  xtable <- xtable(  
    .tibble,  
    caption = .cap,  
    label = .lab,  
    digits = 4  
  )  
  
  # V xtable existuje bug, kde při globálním nastavení decimálního  
  # oddělovače na ',' se vypíše .scale s čárkou a LaTeX hází error  
  options(OutDec = ".")  
  
  # Vytvoření cesty pro soubor  
  .file = paste(TABLES_DIR, .file, sep = "/")  
  
  # Uložení tabulky do souboru  
  print(xtable,  
        file = .file,  
        scalebox = .scale  
  )  
  
  # Znovunastavení decimálního oddělovače na ','  
  options(OutDec = ",")  
}
```

## Uložení modelu do tabulky

```
#' Uložení logistického modelu do LaTeX souboru ve formě tabulky
#'  
#' @param .model Logistický model  
#' @param .file Soubor, do kterého tabulku uložit  
#' @param .cap Caption  
#' @param .lab Label  
save_bc_model <- function(.model, .file, .cap, .lab) {  
  
  # Pokud není zadáný label, použije se jako label název souboru  
  # bez přípony  
  if (is.na(.lab)) {  
    .lab <- sprintf("tab:%s", str_extract(.file, "^.*(?!\\.tex)"))  
  }  
  
  # Vytvoření pojmenované tabulky  
  xtable <- xtable(  
    .model,  
    caption = .cap,  
    label = .lab,  
    digits = 4  
  )  
  
  # Pokud není zadáný label, vytvoří se tabulka bez labelu a  
  # captionu  
  if (is.null(.lab)) {  
    xtable <- xtable(.model,  
                      digits = 4 )  
  }  
  
  # Vytvoření cesty pro soubor  
  .file = paste(MODELS_DIR, .file, sep = "/")  
  
  # Uložení tabulky do souboru  
  print(xtable,  
        file = .file,  
        include.rownames = T  
  )  
}
```

## Uložení matice záměn do tabulky

```
#' Uložení matice záměn do LaTeX souboru  
#'  
#' @param .table Tabulka z matice záměn  
#' @param .file Soubor, do tabulku z matice záměn uložit  
#' @param .cap Caption  
#' @param .lab Label  
save_bc_matrix <- function(.table, .file, .cap, .lab) {
```

```

# Přeložen názvů sloupečků
rownames = c("Pozitivní predikce", "Negativní predikce")
colnames = c("Původní pozitivní", "Původní negativní")

dimnames(.table) <- list(
  rownames,
  colnames
)

# Pokud není zadáný label, použije se jako label název souboru
# bez přípony
if (is.na(.lab)) {
  .lab <- sprintf("tab:%s", str_extract(.file, "^.*(?=\\.tex)"))
}

# Vytvoření pojmenované tabulky
xtable <- xtable(
  .table,
  caption = .cap,
  label = .lab,
  digits = 4
)

# Pokud není zadáný label, vytvoří se tabulka bez labelu a
# captionu
if (is.null(.lab)) {
  xtable <- xtable(.table,
    digits = 4)
}

# Vytvoření cesty pro soubor
.file = paste(MATRIXES_DIR, .file, sep = "/")

# Uložení tabulky do souboru
print(xtable,
  file = .file,
  include.rownames = T
)
}

```

## Uložení statistik z matice záměn do tabulky

```

#' Uložení statistik z matice záměn do LaTeX souboru
#'
#' @param .matrix Matice záměn
#' @param .file Soubor, do kterého statistiky uložit
#' @param .cap Caption matice
#' @param .lab Label matice
save_bc_matrix_output <- function(.matrix, .file, .cap, .lab) {

  # Spojení statistik

```

```

.matrix <- bind_rows(
  tibble(
    statistika = names(.matrix$overall),
    hodnota = .matrix$overall
  ),
  tibble(
    statistika = names(.matrix$byClass),
    hodnota = .matrix$byClass
  )
) |>
# Získání relevantních statistik a přeložení statistik
filter(statistika %in% c("Accuracy", "Sensitivity", "Specificity")) |>
mutate(statistika = c("Přesnost", "Senzitivita", "Specifická"))

# Pokud není zadáný label, použije se jako label název souboru
# bez přípony
if (is.na(.lab)) {
  .lab <- sprintf("tab:%s", str_extract(.file, "^.*(?:\\.tex)"))
}

# Vytvoření pojmenované tabulky
xtable <- xtable(
  .matrix,
  caption = .cap,
  label = .lab,
  digits = 4
)

# Pokud není zadáný label, vytvoří se tabulka bez labelu a
# captionu
if (is.null(.lab)) {
  xtable <- xtable(.matrix,
    digits = 4 )
}

# Vytvoření cesty pro soubor
.file = paste(STATS_DIR, .file, sep = "/")

# Uložení tabulky do souboru
print(xtable,
  file = .file
)
}

```

## Vytvoření matice záměn

```

#' Získání matice záměn
#'
#' @param .model Model pro matici záměn
#' @param .split Split z knihovny rsample

```

```

#'
#' @return Matice záměn z knihovny caret
get_matrix <- function(.model, .split) {
  confusionMatrix(
    # Predikce na testovací množině dat
    data = predict(object = .model,
                   newdata = testing(.split),
                   type = "response") |>
    round() |>
    as.factor(),
    # Původní hodnoty z testovací množiny dat
    reference = testing(.split)$map_winner |>
    as.factor(),
    positive = "1"
  )
}

```

## Získání charakteristik pro hráče na určité mapě

```

#' Funkce pro extrakci charakteristik hráče na mapě v jedno zápase
#'
#' @param tbl datový soubor players.csv
#' @param map číslo mapy pro extrakci charakteristik
#'
#' @return tabulka s charakteristikami hráčů na dané mapě
clear_player_map <- function(.tbl, .map) {
  m_ <- sprintf("m%i_", .map)
  .tbl |>
  select(match_id, player_id, team,
         # Vybrání mapy a statistik podle pro i-tou mapu
         map = sprintf("map_%i", .map),
         matches(sprintf("^m%i_", .map))
  ) |>
  # Získání pouze agregovaných statistik za obě strany
  # e.g. vyloučení m1_kills_ct, m1_kills_t a ponechání pouze m1_kills
  select(!matches("_c?t$")) |>
  drop_na(map) |>
  # Sjednocení názvů, e.g. m1_kills -> kills
  rename_with(
    .cols = starts_with(m_),
    .fn = ~ str_remove_all(.x, m_)
  ) |>
  # Odstranění sloupečků, které nejsou kompletní
  # kvůli historii vývoje cs:go a hltv.org
  select(-c(kdiff, flash_assists, kast, adr)) |>
  # konverze HS na relativní jednotky
  mutate(hs = ifelse(kills == 0, 0, hs / kills))
}

```

### Sjednocení výsledků pro daný tým

```
#' Získání charakteristik týmu v daném zápase  
#'  
#' @param .tbl Datový soubor results.csv  
#' @param i Číslo týmu pro získání charakteristik  
#'  
#' @return Tabulka s charakteristikami týmu v zápase  
clear_results_team <- function(.tbl, .i) {  
  .tbl |>  
    # Vybrání relevantních charakteristik pro daný tým  
    select(  
      date, match_id, team = matches(sprintf("team_%i", .i)),  
      map, map_winner, starting_ct,  
      team_rank = matches(sprintf("rank_%i", .i))  
    ) |>  
    # Získání pouze relevantních map  
    filter(map %in% maps) |>  
    mutate(  
      map_winner = as.numeric(map_winner == .i),  
      starting_ct = as.numeric(starting_ct == .i)  
    )  
}
```

## Geometrický průměr

```
#' Získání geometrického průměru pomocí log transformace
#'  
#'  
#'  
#'  
#'  
geom_mean <- function(x) {  
  exp(sum(log(x)) / length(x))  
}
```

## Datový soubor players.csv

## Nahrání dat

```
players.raw <- read_csv("data/players.csv", show_col_types = F)
```

## Čištění dat

```
players <-  
  # Získání charakteristik pro první mapu  
  clear_player map(players.raw, 1) |>
```



```

# Získání charakteristik pro druhou mapu
rbind(clear_player_map(players.raw, 2)) |>
# Získání charakteristik pro třetí mapu
rbind(clear_player_map(players.raw, 3)) |>
filter(map %in% maps) |>
drop_na() |>
# Odstranění map, kde není právě pět hráčů
group_by(match_id, team, map) |>
filter(n() == 5) |>
ungroup() |>
mutate(
  match_id = as.integer(match_id),
  player_id = as.integer(player_id),
  kills = as.integer(kills),
  assists = as.integer(assists),
  deaths = as.integer(deaths),
  fkdifff = as.integer(fkdifff)
)

```

## Uložení LaTeX tabulky

```

set.seed(12345)
players |>
  slice_sample(n = 6) |>
  save_bc_table(
    .tibble = _,
    .file = "players_csv_transformovano.tex",
    .cap = "Záznam z transformovaného datového souboru players.csv",
    .lab = NA,
    .scale = 0.90
  )

```

## Datový soubor results.csv

### Nahrání dat

```

results.raw <- read_csv("data/results.csv", show_col_types = F)

```

### Čištění dat

```

results <- results.raw |>
  select(date, team_1, team_2,
         map = `_map`, map_winner,
         starting_ct, rank_1, rank_2, match_id) |>
  # Smazání záznamu ke tým hrál sám proti sobě
  # Chybná extrakce dat

```

```

filter(match_id != 2314674)

results <-
  # Získání charakteristik pro první tým
  clear_results_team(.tbl = results, 1) |>
  # Získání charakteristik pro druhý tým
  rbind(clear_results_team(.tbl = results, 2)) |>
  mutate(
    match_id = as.integer(match_id),
    map_winner = as.integer(map_winner),
    starting_ct = as.integer(starting_ct),
    team_rank = as.integer(team_rank)
  ) |>
  mutate(team = ifelse(team == "?", "Astralis", team))

```

## Uložení LaTeX tabulky

```

set.seed(9574)
results |>
  slice_sample(n = 6) |>
  save_bc_table(
    .tibble = _,
    .file = "results_csv_transformovano.tex",
    .cap = "Příklad záznamu z transformovaného datového souboru results.csv",
    .lab = NA,
    .scale = 1
  )

```

## Spojený datový soubor

### Spojení dat

```

data <- results |>
  inner_join(players,
    by = c("match_id", "team", "map"))

```

## Uložení LaTeX tabulky

```

set.seed(9574)
data |>
  slice_sample(n = 6) |>
  save_bc_table(
    .tibble = _,
    .file = "data_spojena.tex",
    .cap = "Příklad záznamu z transformovaného datového souboru results.csv",
    .lab = NA,
  )

```

```
.scale = 1  
)
```

## Agregovaný datový soubor

### Agregace

```
data.aggregated <- data |>  
  group_by(match_id, map, team) |>  
  summarise(  
    mean_kills = mean(kills),  
    mean_assists = mean(assists),  
    mean_deaths = mean(deaths),  
    mean_hs = geom_mean(hs),  
    mean_fkdiff = mean(fkdiff),  
    # Průměr z vektoru stejných čísel c je c  
    map_winner = mean(map_winner),  
    starting_ct = mean(starting_ct),  
    team_rank = mean(team_rank),  
    .groups = "drop"  
  )
```

### Uložení LaTeX tabulky

```
set.seed(6587)  
data.aggregated |>  
  slice_sample(n = 6) |>  
  save_bc_table(  
    .tibble = _,  
    .file = "data_agregovana.tex",  
    .cap = "Agregovaná data pro týmy za zápas a mapu",  
    .lab = NA,  
    .scale = 0.5  
  )
```

## Průzkumová analýza dat

### Kvantitativní proměnné

```
g <- data |>  
  select(kills, assists, deaths, hs, fkdiff, rating, team_rank) |>  
  ggcorr(hjust = 0.8, size = 4, label = T) +  
  labs(  
    title = "Korelační matice kvantitativních prediktorů",  
    caption = "Spojený datový soubor"
```

```
)

save_bc_picture(.graph = g,
                .file = "prediktory_corr_matice.png")
```

## Histogram prediktorů

```
bin_count <- round(1 + 3.3*log10(nrow(data)/5))
g <- data |>
  pivot_longer(
    cols = c(kills, assists, deaths, hs, fkdifff),
    names_to = "stats",
    values_to = "vals"
  ) |>
  select(stats, vals) |>
  mutate(bins = 1 + 3.3*log10(n()/6),
         bins = 1) |>
  ggplot(aes(x = vals, fill = stats)) +
  facet_wrap(vars(stats), ncol = 2, scales = "free") +
  geom_histogram(bins = bin_count, show.legend = F) +
  scale_y_continuous(labels = scales::number) +
  theme(
    axis.title.x = element_blank(),
    strip.background = element_rect(fill = "grey")
  ) +
  labs(
    title = "Rozdělení kvantitativních statistik",
    subtitle = "Osa Y je přizpůsobená každé statistice",
    y = "Absolutní počet",
    caption = "Spojení souboru players.csv a results.csv"
  )

save_bc_picture(.graph = g,
                .file = "histogram_prediktoru.png")
```

## Míra výhry proměnných

```
data.plot <- data.aggregated |>
  semi_join(
    data.aggregated |>
      group_by(team) |>
      filter(n() > 100) |>
      summarise(mtr = mean(team_rank)) |>
      slice_min(order_by = mtr, n = 10),
    by = "team"
  ) |>
  group_by(team) |>
  summarise(mmw = mean(map_winner)) |>
  arrange(desc(mmw))
```

```

g <- ggplot(data = data.plot,
  aes(x = reorder(team, mmw), y = mmw, fill = team)) +
  geom_col(show.legend = F) +
  geom_text(data = filter(data.plot, mmw > 0.6),
    aes(label = paste0(round(mmw * 100, 2), " %"),
      y = mmw + 0.08)) +
  scale_fill_manual(values = paletteer_d("basetheme::deepblue")) +
  scale_y_continuous(labels = scales::percent, limits = c(0, 0.8)) +
  coord_flip() +
  labs(
    title = "Porovnání míry výhry pro nejlepších 10 týmů",
    subtitle = "Pouze pro týmy s alespoň 100 zápasy\n0sa x je omezená na interval <0; 0,8>",
    x = "Tým",
    y = "Míra výhry",
    caption = "Agregovaný datový soubor"
  )
)

save_bc_picture(.graph = g,
  .file = "mira_vyhry_tymu.png")

```

## Modely pro hráče - Mirage

### Optimalizovaný model

```

set.seed(547)
mirage.data <- data |>
  filter(map == "Mirage")

mirage.split <- initial_split(mirage.data, prop = 0.8)

mirage.model <- glm(
  formula = map_winner ~ kills + assists + deaths + hs +
    fkdiff + starting_ct,
  data = training(mirage.split),
  family = binomial
)

```

### Uložení LaTeX tabulky

```

save_bc_model(
  .model = mirage.model,
  .file = "Mirage_model.tex",
  .cap = "Výstup z programu R pro model hráčů na mapě Mirage",
  .lab = NA
)

```

## Matice záměn

```
mirage.matrix <- get_matrix(  
  .model = mirage.model,  
  .split = mirage.split  
)
```

## Uložení LaTeX tabulky

```
save_bc_matrix(  
  .table = mirage.matrix$table,  
  .file = "Mirage_matice.tex",  
  .cap = " Matice záměn pro model hráčů na mapě Mirage",  
  .lab = NA  
)  
  
save_bc_matrix_output(  
  .matrix = mirage.matrix,  
  .file = "Mirage_stats.tex",  
  .cap = "Vybrané statistiky z matice záměn pro model hráčů na mapě Mirage",  
  .lab = NA  
)
```

## Modely pro hráče - Vertigo

### Neoptimalizovaný model

```
set.seed(547)  
vertigo.data <- data |>  
  filter(map == "Vertigo")  
  
vertigo.split <- initial_split(vertigo.data, prop = 0.8)  
  
vertigo.model <- glm(  
  formula = map_winner ~ kills + assists + deaths + hs +  
    fkdif + starting_ct,  
  data = training(vertigo.split),  
  family = binomial  
)
```

## Uložení LaTeX tabulky

```
save_bc_model(  
  .model = vertigo.model,  
  .file = "Vertigo_model.tex",
```

```
.cap = "Výstup z programu R pro model hráčů na mapě Vertigo",
.lab = NA
)
```

## Optimalizovaný model

```
set.seed(547)
vertigo.data <- data |>
  filter(map == "Vertigo")

vertigo.split <- initial_split(vertigo.data, prop = 0.8)

vertigo.model <- glm(
  formula = map_winner ~ kills + assists + deaths,
  data = training(vertigo.split),
  family = binomial
)
```

## Uložení LaTeX tabulky

```
save_bc_model(
  .model = vertigo.model,
  .file = "Vertigo_model_opt.tex",
  .cap = "Výstup z programu R pro optimalizovaný model hráčů na mapě Vertigo",
  .lab = NA
)
```

## Matice záměn

```
vertigo.matrix <- get_matrix(
  .model = vertigo.model,
  .split = vertigo.split
)
```

## Uložení LaTeX tabulky

```
save_bc_matrix(
  .table = vertigo.matrix$table,
  .file = "Vertigo_matice_opt.tex",
  .cap = "Matice záměn pro model hráčů na mapě Vertigo",
  .lab = NA
)

save_bc_matrix_output(
  .matrix = vertigo.matrix,
```

```

.file = "Vertigo_stats_opt.tex",
.cap = "Vybrané statistiky z matice záměn pro modelu hráčů na mapě Vertigo",
.lab = NA
)

```

## Porovnání statistik obou modelů

```

# Vytvoření statistik z matice záměn pro mapu Mirage
m1 <-
  bind_rows(
    tibble(
      statistika = names(mirage.matrix$overall),
      `Mapa Mirage` = mirage.matrix$overall
    ),
    tibble(
      statistika = names(mirage.matrix$byClass),
      `Mapa Mirage` = mirage.matrix$byClass
    )
  ) |>
  filter(statistika %in% c("Accuracy", "Sensitivity", "Specificity")) |>
  mutate(statistika = c("Přesnost", "Senzitivita", "Specifická"))

# Vytvoření statistik z matice záměn pro mapu Vertigo
m2 <-
  bind_rows(
    tibble(
      statistika = names(vertigo.matrix$overall),
      `Mapa Vertigo` = vertigo.matrix$overall
    ),
    tibble(
      statistika = names(vertigo.matrix$byClass),
      `Mapa Vertigo` = vertigo.matrix$byClass
    )
  ) |>
  filter(statistika %in% c("Accuracy", "Sensitivity", "Specificity")) |>
  mutate(statistika = c("Přesnost", "Senzitivita", "Specifická"))

m12 <-
  cbind(
    m1,
    m2 |> select(-statistika)
  )

save_bc_table(
  .tibble = m12,
  .file = "Mirage_Vertigo_porovnani.text",
  .cap = "Porovnání statistik pro model hráčů na mapě Mirage a Vertigo",
  .lab = NA
)

```



## Uložení ostatních map

```
set.seed(962147)
# Uložení modelů, matic záměn a statistik pro zbývající mapy
for (.map in setdiff(maps, c("Mirage", "Vertigo"))){

  .data <- data |>
    filter(map == .map)

  .split <- initial_split(.data, prop = 0.8)

  .model <- glm(
    formula = map_winner ~ kills + assists + deaths + hs +
      fkdiff + starting_ct,
    data = training(.split),
    family = binomial
  )

  save_bc_model(
    .model = .model,
    .file = sprintf("%s_model.tex",
      .map),
    .cap = sprintf(
      "Výstup z programu R pro model hráčů na mapě %s",
      .map),
    .lab = NA
  )

  .matrix <- get_matrix(
    .model = .model,
    .split = .split
  )

  save_bc_matrix(
    .table = .matrix$table,
    .file = sprintf(
      "%s_matice.tex",
      .map),
    .cap = sprintf(
      "Matice záměn pro model hráčů na mapě %s",
      .map),
    .lab = NA
  )

  save_bc_matrix_output(
    .matrix = .matrix,
    .file = sprintf("%s_stats.tex",
      .map),
    .cap = sprintf("Vybrané statistiky z matice záměn pro model hráčů na mapě %s",
      .map),
    .lab = NA
  )
}
```

## Model pro všechny týmy

### Optimalizovaný model

```
set.seed(4872)
global.split <- initial_split(data.aggregated, prop = 0.8,
                              strata = map)

global.model <-
  glm(
    formula = map_winner ~ mean_kills + mean_assists + mean_deaths +
      mean_hs + mean_fkdiff + team_rank + starting_ct:map,
    data = training(global.split),
    family = binomial
  )
```

### Uložení LaTeX tabulky

```
save_bc_model(
  .model = global.model,
  .file = "Global_model.tex",
  .cap = "Výstup z programu R pro model týmu pro všechny týmy",
  .lab = NA
)
```

### Matice záměn

```
global.matrix <- get_matrix(
  .model = global.model,
  .split = global.split
)
```

### Uložení LaTeX tabulky

```
save_bc_matrix(
  .table = global.matrix$table,
  .file = "Global_matice.tex",
  .cap = "Matice záměn pro model týmu pro všechny týmy",
  .lab = NA
)

save_bc_matrix_output(
  .matrix = global.matrix,
  .file = "Global_stats.tex",
  .cap = "Vybrané statistiky z matice záměn pro model týmu pro všechny týmy",
  .lab = NA
)
```

## Model pro tým Astralis

### Neotimalizovaný model

```
set.seed(417)
astralis.split <- initial_split(data.aggregated |>
                                filter(team == "Astralis"),
                                prop = 0.8,
                                strata = "map")

astralis.model <-
  glm(
    formula = map_winner ~ mean_kills + mean_assists + mean_deaths +
      mean_hs + mean_fkdif + team_rank + starting_ct:map,
    data = training(astralis.split),
    family = binomial
  )
```

### Optimalizovaný model

```
set.seed(487)
astralis.split <- initial_split(data.aggregated |>
                                filter(team == "Astralis"),
                                prop = 0.8,
                                strata = "map")

astralis.model <- glm(
  formula = map_winner ~ mean_kills + mean_deaths,
  data = training(astralis.split),
  family = binomial
)
```

### Uložení LaTeX tabulky

```
save_bc_model(
  .model = astralis.model,
  .file = "Astralis_model_opt.tex",
  .cap = "Výstup z programu R pro optimalizovaný pro model týmu pro tým Astralis",
  .lab = NA
)
```

### Matice záměn

```
astralis.matrix <- get_matrix(
  .model = astralis.model,
  .split = astralis.split
)
```

## Uložení LaTeX tabulky

```
save_bc_matrix(
  .table = astralis.matrix$table,
  .file = "Astralis_matice_opt.tex",
  .cap = "Matice záměn pro model týmu pro tým Astralis",
  .lab = NA
)

save_bc_matrix_output(
  .matrix = astralis.matrix,
  .file = "Astralis_stats_opt.tex",
  .cap = "Vybrané statistiky z matice záměn pro model týmu pro tým Astralis",
  .lab = NA
)
```

## Model pro tým Sprout

### Neotimalizovaný model

```
set.seed(8741)
sprout.split <- initial_split(data.aggregated |>
  filter(team == "Sprout"),
  prop = 0.8,
  strata = "map")

sprout.model <-
  glm(
    formula = map_winner ~ mean_kills + mean_assists + mean_deaths +
      mean_hs + mean_fkdiff + team_rank + starting_ct:map,
    data = training(sprout.split),
    family = binomial
  )
```

### Optimalizovaný model

```
set.seed(326598)
sprout.split <- initial_split(data.aggregated |>
  filter(team == "Sprout"),
  prop = 0.8,
```

```

                                strata = "map")

sprout.model <- glm(
  formula = map_winner ~ mean_kills + mean_deaths,
  data = training(sprout.split),
  family = binomial
)

```

## Uložení LaTeX tabulky

```

save_bc_model(
  .model = sprout.model,
  .file = "Sprout_model_opt.tex",
  .cap = "Výstup z programu R pro optimalizovaný pro model týmu pro tým Sprout",
  .lab = NA
)

```

## Matice záměn

```

sprout.matrix <- get_matrix(
  .model = sprout.model,
  .split = sprout.split
)

```

## Uložení LaTeX tabulky

```

save_bc_matrix(
  .table = sprout.matrix$table,
  .file = "Sprout_matice_opt.tex",
  .cap = "Matice záměn pro model týmu pro tým Sprout",
  .lab = NA
)

save_bc_matrix_output(
  .matrix = sprout.matrix,
  .file = "Sprout_stats_opt.tex",
  .cap = "Vybrané statistiky z matice záměn pro model týmu pro tým Sprout",
  .lab = NA
)

```

## Porovnání statistik týmových modelů

```

# Vytvoření statistik z matice záměn pro všechny týmy
m1 <-
  bind_rows(
    tibble(
      statistika = names(global.matrix$overall),
      `Celkový model` = global.matrix$overall
    ),
    tibble(
      statistika = names(global.matrix$byClass),
      `Celkový model` = global.matrix$byClass
    )
  ) |>
  filter(statistika %in% c("Accuracy", "Sensitivity", "Specificity")) |>
  mutate(statistika = c("Přesnost", "Senzitivita", "Specifická"))

# Vytvoření statistik z matice záměn pro tým Astralis
m2 <-
  bind_rows(
    tibble(
      statistika = names(astralis.matrix$overall),
      `Tým Astralis` = astralis.matrix$overall
    ),
    tibble(
      statistika = names(astralis.matrix$byClass),
      `Tým Astralis` = astralis.matrix$byClass
    )
  ) |>
  filter(statistika %in% c("Accuracy", "Sensitivity", "Specificity")) |>
  mutate(statistika = c("Přesnost", "Senzitivita", "Specifická"))

# Vytvoření statistik z matice záměn pro tým Sprout
m3 <-
  bind_rows(
    tibble(
      statistika = names(sprout.matrix$overall),
      `Tým Sprout` = sprout.matrix$overall
    ),
    tibble(
      statistika = names(sprout.matrix$byClass),
      `Tým Sprout` = sprout.matrix$byClass
    )
  ) |>
  filter(statistika %in% c("Accuracy", "Sensitivity", "Specificity")) |>
  mutate(statistika = c("Přesnost", "Senzitivita", "Specifická"))

m123 <-
  cbind(
    m1,
    m2 |> select(-statistika),
    m3 |> select(-statistika)
  )

save_bc_table(

```

```
.tibble = m123,  
.file = "Tymy_porovnani.text",  
.cap = "Porovnání statistik pro týmové modely",  
.lab = NA  
)
```