Vysoká škola ekonomická v Praze
Fakulta informatiky a statistiky

Modely logistické regrese v oblasti esportových dat

BAKALÁŘSKÁ PRÁCE

Studijní program: Aplikovaná informatika
Studijní obor: Aplikovaná informatika

Autor: Michal Lauer
Vedoucí práce: Ing. Zdeněk Šulc, Ph.D.

Praha, Duben 2022

Prohlášení

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

V Praze dne DD. Dubna 2021 .....

Podpis studenta

Poděkování

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

Abstrakt

Cílem bakalářské práce je kvantitativně zanalyzovat esportové zápasy ze hry **CSGO!** (**CSGO!**) a predikovat výhry vybraného teamu. Použitý datový soubor je z internetového portálu kaggle.com[1] a obsahuje data od roku 2017 až do roku 2020. Práce je rozdělena do tří částí. V první části je představen esport jako takový, je zde shrnutý jeho vývoj a jsou zde definovány důležité pojmy a termíny. Druhá část obsahuje popis metod, které jsou pro analýzu a predikci použity. Pro analýzu data setu jsou zobrazeny grafy jako boxplot či čárový graf. Predikce jsou založené na vícerozměrném logistickém regresním modelu. V závěrečné praktické části jsou metody použité pro analýzu data setu, predikci výhry daného teamu a model je vyhodnocen jak kvantitativně, tak i v kontextu reálného využití.

Klíčová slova

Model, logistická regrese, predikce, esport

[1] https://www.kaggle.com/mateusdmachado/csgo-professional-matches

Abstract

– Bude přeložen po odsouhlasení abstraktu v češtině

Keywords

Model, logistic regression, prediction, esport

Obsah

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 (**Gough2021**). Dalo by se spekulovat, že za takový velký nárůst je zodpovědná aktuální pandemie. Většina populace je nucena zůstat doma. Toto otevřelo dveře se s esportem přirozeně seznámit a nějakým způsobem se ho účastnit (online divák, soutěžící, organizátor, fanoušek...). Hrají se různé kategorie her jako např. střílečky, **MOBA!** (**MOBA!**)[1], karetní hry, **FPS!** (**FPS!**) či **BR!** (**BR!**).

Práce se zaměřuje primární na esportový titul **CSGO!** (**CSGO!**). Je to jeden z nejdéle hraných esportových titulů, boří mnohé divácké rekordy[2] a je aktuálně nejhranějším **FPS!** esport titulem. **CSGO!** vyniká nejen detailní herní mechanikou, ale i bohatou a zajímavou historií. Hra je unikátní i tím, že obsahuje mnoho různých módu[3] a hráč může strávit mnoho hodin pouze objevováním komunitních serverů, hraním klasických zápasů či trénováním na offline mapách.

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

Logistický model je preferován kvůli své lehké interpretaci a dobré aplikaci v reálném životě. Výsledky, statistiky a pravděpodobnosti mohou být použity např. v sázkových kancelářích, kdy se výsledky modelu dají využít na nejrůznější sázky a lze předpovídat, kdo vyhraje zápas, kdo vyhraje mapu, jaký hráč bude mít nejlepší statistiky, či zda si hráč koupí určitou zbraň.

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

[1] tzn. MOBA, kde hráči hrají v jedné online aréně proti sobě

[2] https://www.invenglobal.com/articles/15619/csgo-major-breaks-viewership-records-overtakes-the-international

[3] rozšíření, jak hru hrát. Každý mód má svá vlastní pravidla, mapy, či herní fanoušky

[4] https://www.kaggle.com/mateusdmachado/csgo-professional-matches

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 **TGNS!** (**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 prvních 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[1] party či organizovat BYOC[2] turnaje. Dále už esport potřeboval jen čas na organický růst a dnes má tržní hodnotu přes jednu miliardu amerických dolarů (**Gough2021**), (**Larch2019**).

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[3] 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 považování za experty v oblasti esportu. Tento průzkum byl následné podpořen růstem že tržní hodnoty esportu a mezi lety 2019 a 2020 vzrostla o téměř 50 % (**Gough2021**).

[1] Hráči hrají v jedné místnosti na lokální počítačové síti.
[2] z ang. Bring Your Own Computer, kde si hráči si na akci donesou vlastní počítač
[3] 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 (**Professeur2021**), (**liquipedia2021**). S větším počtem diváku roste i marketingový potenciál, investiční příležitost a kariérní růst.

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

Poslední žánr který zmíním je **BR!** (**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 trošku 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[4].

2.3 Představení titulu Counter-Strike: Global Offensive

CSGO!, jak ho známe dnes, má bohatou a dlouho historii. Ne vždy se to ovšem jmenovalo stejně. Úplně první iterace hry se jmenovala čistě Counter-Strike a byl to pouze mód[5] 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.

[4]Trailer pro propagaci události: https://www.youtube.com/watch?v=TanGK9o_d24

[5]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[6], 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 (**csko2021**). 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[7] turnaje, kde hráči mohli ukázat svůj um za tehdy relativně velkou sumu peněz. Hra se časem vyvíjela, hráči nalézali nové strategie či triky a Valve vydalo novou verzi — Counter-Strike: Source. Tato nová verze získala nepříliš pozitivní ohlas, jelikož velmi rozdělila herní komunitu. Představila nové mechaniky, staré mechaniky změnila a hráčům, zejména v esportu, se nechtělo učit něco úplně nového. Valve se rozhodlo sjednotit herní komunitu, a proto vydalo novou verzi hry s názvem **CSGO!**

CSGO! se snažilo sjednotit oba tábory z her Counter-Strike 1.6 a Counter-Strike: Source. Hra vyšla 21. srpna 2012 a z počátku nebyla tolik úspěšná, ale díky přidání různých skinů (**Valve2013**) 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 (**Henningson2020**).

[6] herní prvky či unikátní vlastnosti

[7] turnaj pořádaný přímo Valve, který má největší prestiž

2.4 Propojení práce a titulu Counter-Strike: Global Offensive

Práce se zaměřuje na identifikování významných prediktorů a následně vytvoření regresního modelu. Před jakoukoliv prací s daty je ale nutné pochopit, jak se hra vlastně hraje a jaká jsou její pravidla. Ve hře **CSGO!** hraje pět hráčů proti pěti (dále jen 5v5). Hra se většinou hraje online, avšak velké esportové turnaje se hrají offline, tedy v nějaké např. aréně. Hra má v základu 30 kol a po prvních patnácti se mění strany. Jedna strana jsou policisté (Counter-Terrorists či CT), kteří mají za úkol chránit „bomboviště“ - část mapy, která má vybouchnout. Naopak cíl Teroristů (T) je právě bombu položit a „bomboviště“ nechat vybouchnout. Vyhrává tým, který první vyhraje 16 kol. Pokud ovšem po první 30 kolech je stav nerozhodný, tedy 15:15, hraje se prodloužení. Tento formát není standardizovaný pro všechny turnaje, proto zmíním pouze pravidla, která se týkají turnajů od společnosti Valve (již zmíněné a nejvíc prestižní Majory). Zde se hraje prodloužení ve formát Bo6, tedy kdo první získá 4 body, vyhraje zápas. Takto může jít zápas teoreticky do nekonečna. Nejdelší semi-profesionální zápas, který se ovšem neodehrál na Majoru, se stal mezi týmem exceL a XENEX(**hltv.org2015**). 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 trošku 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.

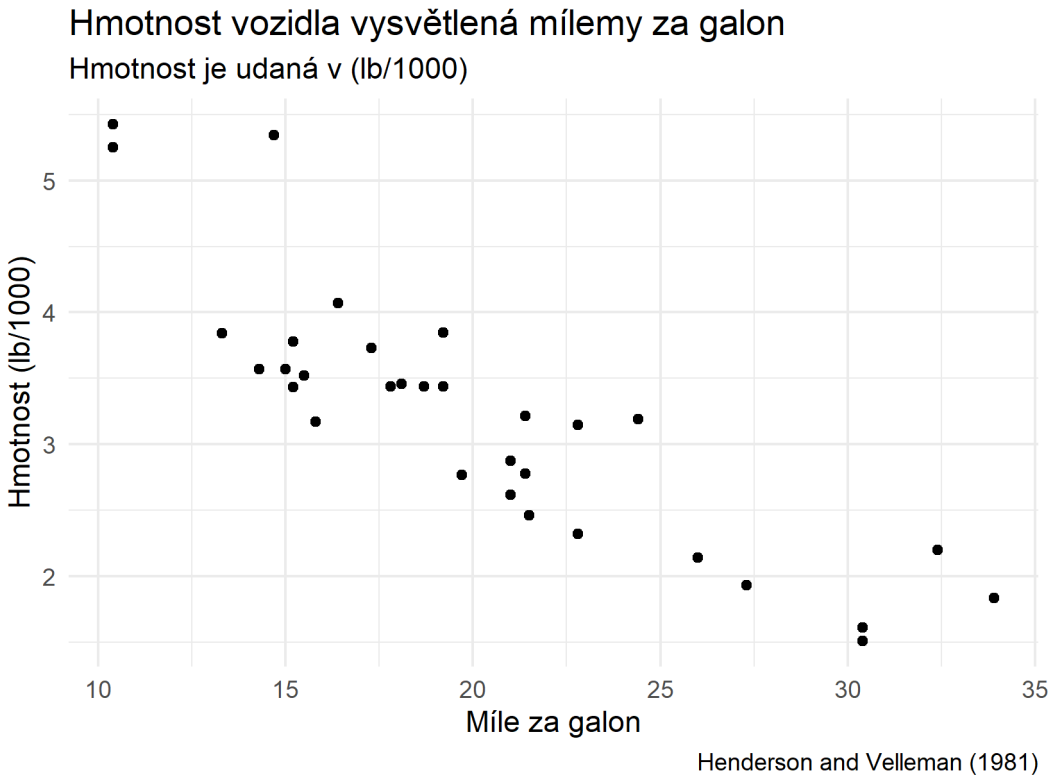
3. Teoretická část

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. Testovací citace: (**Hebak2015**), (**Kleinbaum2010**)

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. Propojením 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

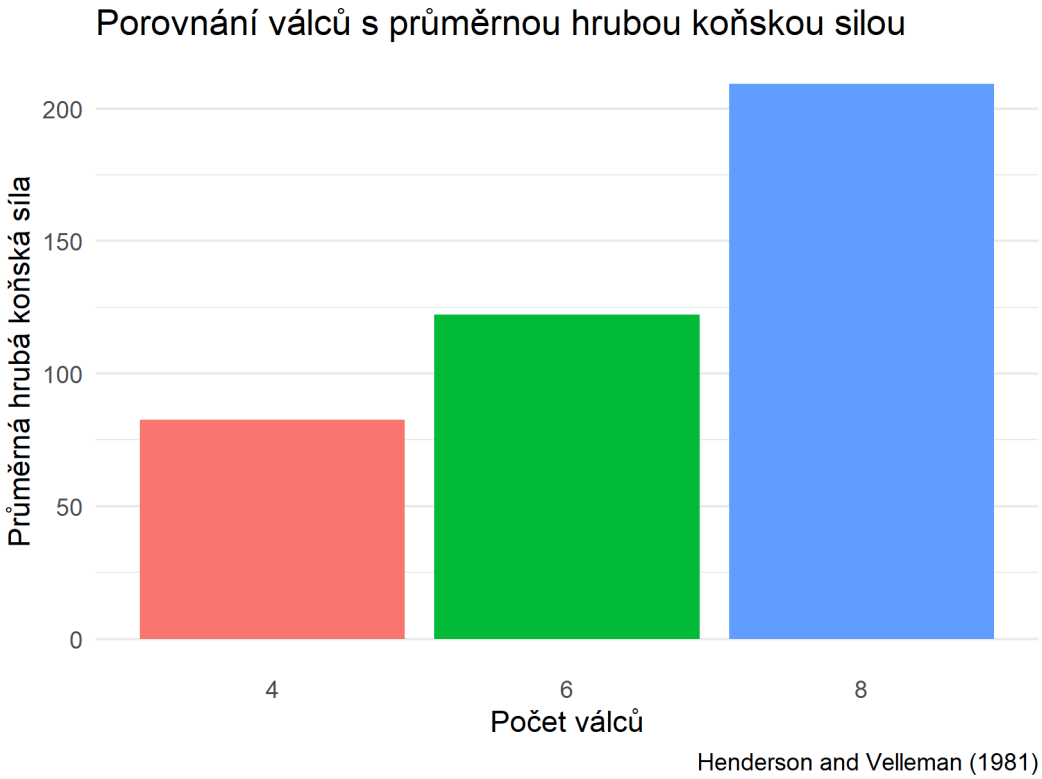
Obrázek **??** zobrazuje negativní korelaci mezi hmotností vozidla a mílemi ujetými za galon.

3.1.2 Sloupcový graf

Sloupcový graf slouží k zobrazení četnosti 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.

Nejčastější se pomocí sloupcového grafu porovnává absolutní četnost dané 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 dané kategorie. V případě ordinální proměnné se zachovává přirozené řazení.

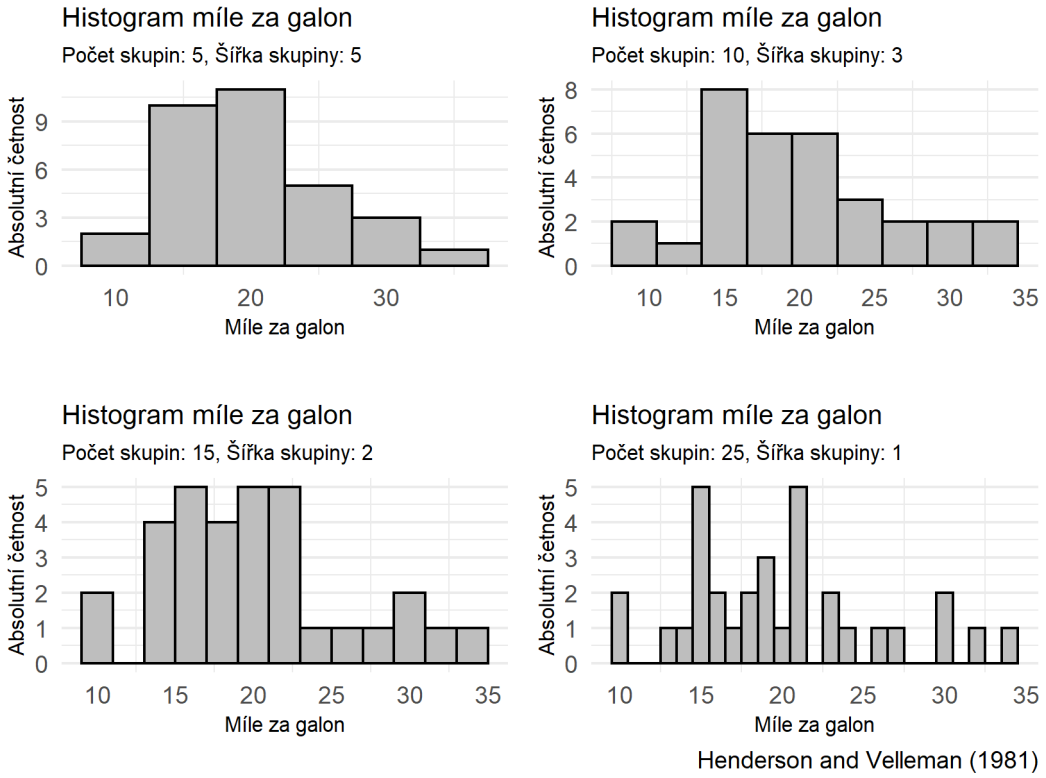
Příklad sloupcového grafu je zobrazen na obrázku, který porovnává průměrnou hrubou koňskou sílu s počtem válců. Je na něm také vidět vztah, kdy s vyšším počtem válců stoupá průměrná koňská síla.



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 *(bins)* 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.



Obrázek 3.3: Porovnání histogramů s různým počtem skupin

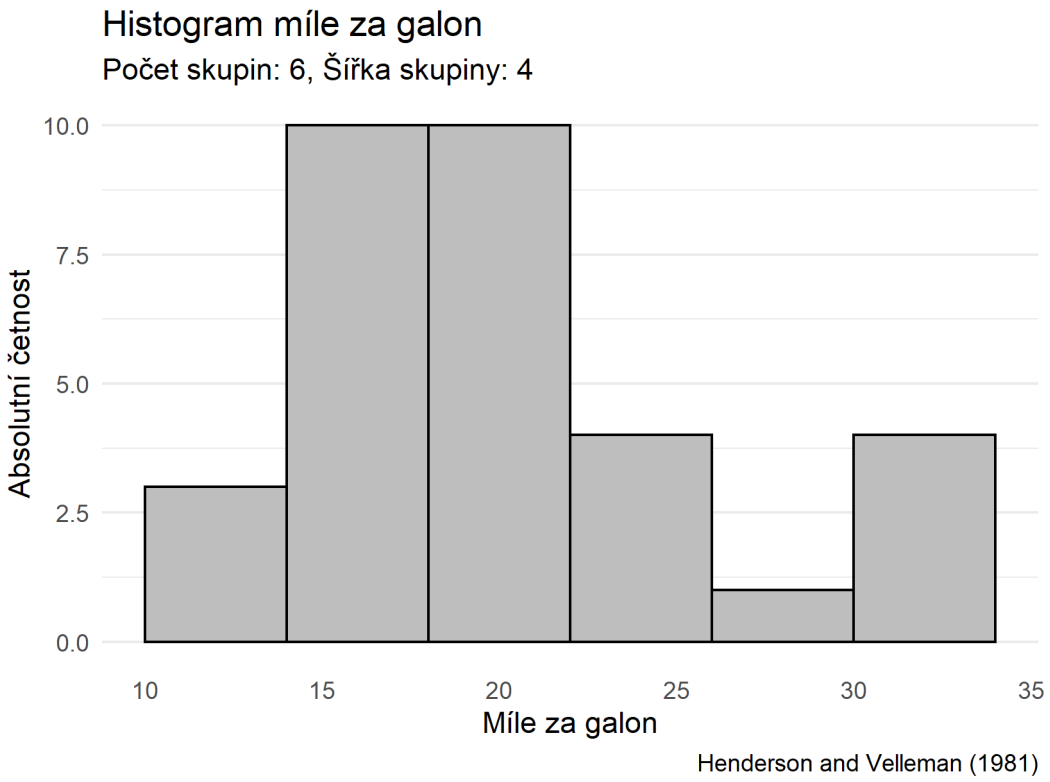
Pro vhodný počet skupin existuje mnoho způsobů. Nejznámější je takzvané Sturgesovo pravidlo, které se spočítá následujícím vztahem:

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

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:

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

kde x je zobrazovaná proměnná, k je počet skupin a w je výsledná šířka intervalu. Uplatněním rovnic **??** a **??** na datový soubor z obrázku **??** lze zobrazit representativnější sloupcový graf. Nutné je však podotknout, že není pravidlem se danými výpočty řídit a výsledný sloupcový graf je nutné přizpůsobit jednotlivým datům.



Obrázek 3.4: Histogram s počtem skupin dle Sturgesova pravidla

3.1.4 Krabičkový graf

Five-number summary

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

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

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

kde Q_0 a $Q_{1,00}$ označují minimum a maximum číselné řady. Kvartily Q_1, Q_2 (medián) a Q_3 jsou čísla, která rozděluji 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.

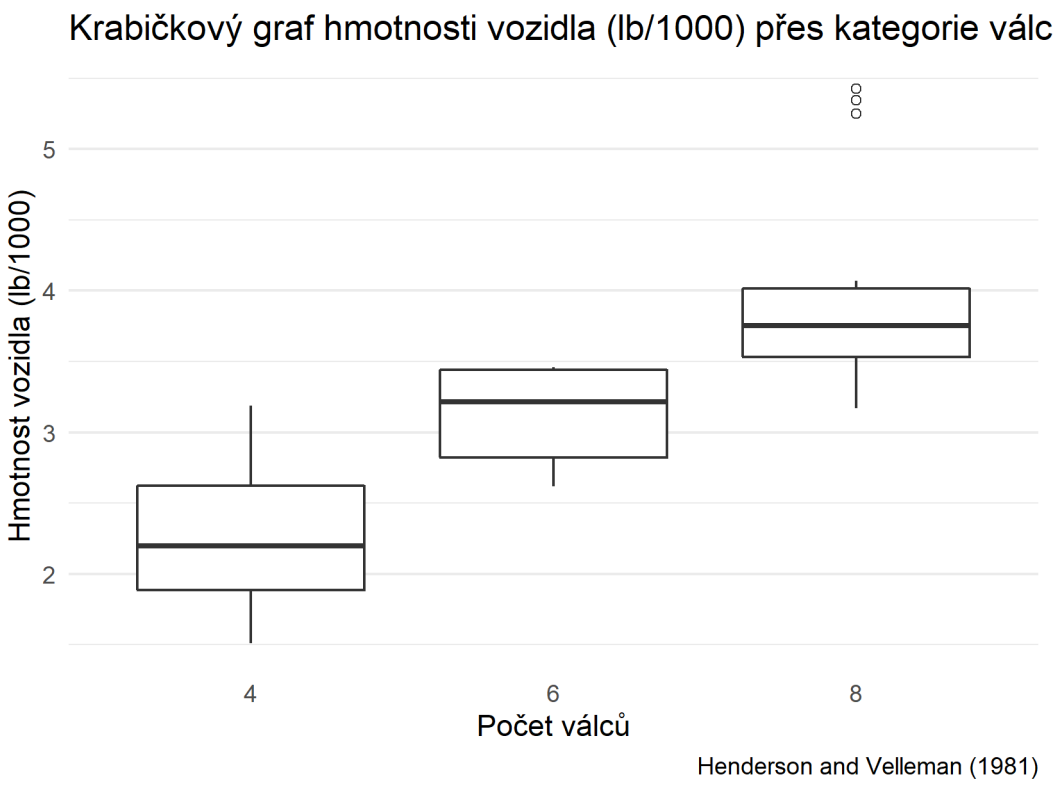
Krabičkový graf

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

$$\textit{Spodní hranice} = Q_1 - 1{,}5 * IQR \tag{3.3}$$

$$\textit{Horní hranice} = Q_3 + 1{,}5 * IQR \tag{3.4}$$

Hodnoty, které spadají do intervalu $\langle Q_1 - 1{,}5IQR; Q_1 - 3IQR\rangle$ a $\langle Q_3 + 1{,}5IQR, Q_3 + 3IQR\rangle$ se nazývají jako odlehlé. Hodnoty které leží mimo tento vztah, tedy hodnoty menší než $Q_1 - 3IQR$ nebo větší než $Q_3 + 3IQR$ se nazývají jako hodnoty extrémní a v krabičkovém grafu jsou z pravidla vyznačeny nějakým speciálním znakem, např. kolečkem. 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 **??** 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

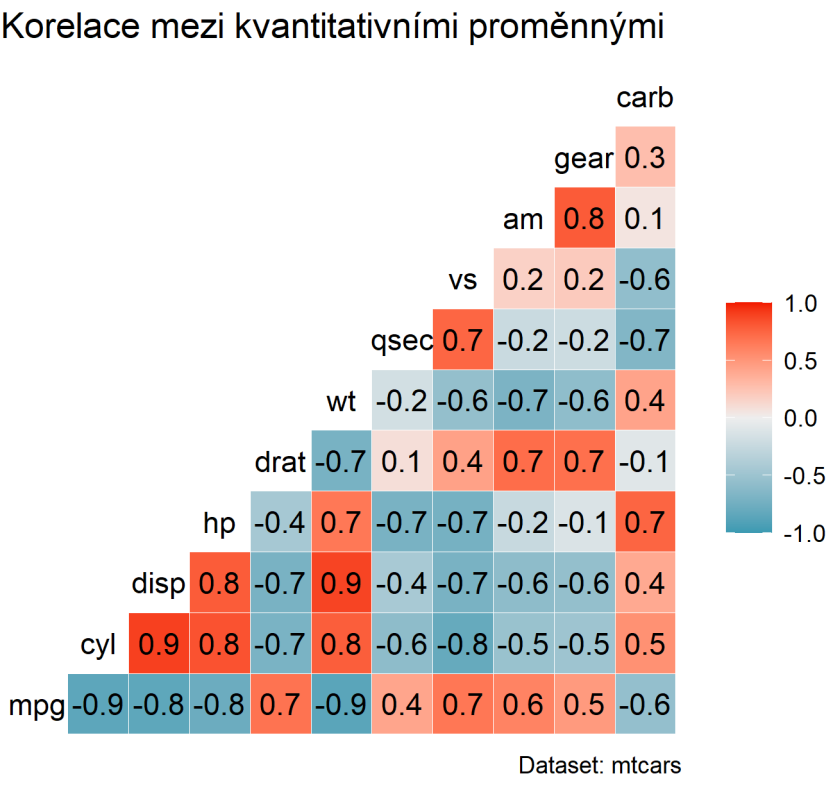
Korelace popisuje směr a sílu vztahu mezi dvěma proměnnými x a y. Značí se r a nabývá hodnot $\langle -1; 1\rangle$.

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

Čím větší je absolutní hodnota korelace mezi proměnnou x a y, tím lépe lze pomocí jedné proměnné vysvětlit druhou proměnnou. Kladnost, případně zápornost korelace značí pak směr vztahu. Pokud je korelace záporná, tedy $r < 0$, s růstem jedné proměnné klesá proměnná druhá. Naopak při kladné korelace, tedy $r > 0$, s růstem jedné proměnné roste i druhá. Pokud by se například změřila výška žáku ve třídě, kde proměnná x bude jejich naměřená výška v centimetrech a proměnná y jejich naměřená výška v metrech, korelace mezi proměnnou x a y bude rovna jedné. To je z toho důvodu, že výška v centimetrech lze perfektně vysvětlit výškou v měřenou v metrech. To platí i naopak.

Korelační matice

Korelační matice je nástroj, díky kterému lze zobrazit korelaci mezi více jak dvěma páry proměnných. Matice může být zobrazená jako tabulka nebo graf. Korelační matice je velmi užitečná v regresní analýze kvůli předpokladu nezávislosti proměnných. Pokud jsou při tvorbě modelu prediktory korelované, vzniká problém tzn. multikolinearity. Pokud jsou prediktory vysoce korelované, zhoršuje to přesnost a vypovídací hodnotu koeficientů. (**Kleinbaum2010**) V takovém případě je např zvýšit počet pozorování nebo z modelu určité modelované prediktory odebrat.



Obrázek 3.6: Korelační matice

Graf korelační matice může mít mnoho podob. V příkladu obrázku **??** 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

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

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

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

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

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

kde α a β_i jsou parametry odhadnuté z datového souboru.

3.2.1 Interpretace parametrů

Parametry α a β_i značí logaritmus šance. α je logaritmus šance v případě, že všechny prediktory jsou teoreticky rovné 0. Parametr β_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 B_i. Toto lze pozorovat například u binárních prediktorů, kdy typicky přítomnost daného prediktoru, značená jedničkou, změní výslednou šanci právě o odhadnutý parametr β. Pro přechod z přirozeného logaritmu šance na šanci lze využít vztahu

$$\check{s}ance \ = e^{\beta_i}. \tag{3.9}$$

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

$$\textit{šance jevu } A = \frac{p}{1-p} \tag{3.10}$$

kde p je pravděpodobnost výskytu jevu A.

3.2.2 Největší pravděpodobnost

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

$$L(\theta) = \Pi_{l=1}^{m_1} P(X_i) \Pi_{l=m_1+1}^{n} 1 - P(X_i), \tag{3.11}$$

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

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

3.2.3 Matice záměn

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

		Pozitivní predikce	Negativní predikce
		1	0
Původní pozitivní	1	Skutečně pozitivní	Falešně negativní
Původní negativní	0	Falešně pozitivní	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 **??**. Každá ze čtyř vnitřních buněk má vlastní označení a interpretaci:

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

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

$$\mathit{Přesnost} = \frac{\text{Skutečně pozitivní} + \text{Skutečně negativní}}{n} \tag{3.13}$$

Přesnost říká, kolik predikcí bylo klasifikováno správně. Pokud je ovšem poměr původních pozitivních a negativních hodnot velmi nevyrovnaný, tato statistika není vhodná. Toto se může stát například v lékařství při identifikaci nemocného pacienta. Zde je důležitější úspěšně predikovat výskyt nemoci, který se ovšem v datovém souboru často objevuje v menším poměru. V tomto případě lze použít statistiku zvanou senzitivita. Ta se rovná poměru správných pozitivních predikcí a úhrnu všech pozitivních predikcí

$$\mathit{Senzitivita} = \frac{\text{Skutečně pozitivní}}{\text{Skutečně pozitivní} + \text{Falešně pozitivní}} \tag{3.14}$$

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

$$\textit{Specificita} = \frac{\text{Skutečně negativní}}{\text{Skutečně negativní} + \text{Falešně negativní}} \tag{3.15}$$

Nanesením sensitivity **??** a specificity **??** lze vytvořit tzn. ROC křivku.

3.2.4 Testování hypotéz

... Dopsat

3.2.5 Waldův test

Waldův test ověřuje, zda je parametr β_i v populaci významný či nikoliv. Definice testu hypotézy je tedy

H_0 : Koeficient β_i je rovný nule
H_A : Koeficient β_i je různý od nuly.

Pro vyhodnocení hypotézy se používá kritická hodnota Z, který se vypočítá jako poměr testovaného parametru $S_{\hat{\beta}_i}$ a směrodatné chyby koeficientu β_i

$$Z = \frac{\hat{\beta}_i}{S_{\hat{\beta}_i}}. \tag{3.16}$$

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

4. Praktická část

V této části se nachází popis dat, transformace dat, a tvorba logistického regresního modelu. Nejprve jsou představené datové soubory, se kterými se pracuje. Následně jsou použité grafy, které jsou představené v sekci **??**. Pomocí vizualizace dat lze představit proměnné, které by mohli do logistického modelu vstupovat. Také lze díky grafům získat intuici, jak datový soubor vypadá a jaké výsledky je možné očekávat. Poté jsou vytvořené logistické regresní modely, jejich výstup je interpretován a různé modelu jsou mezi sebou porovnány. V závěru každé podsekce se nachází zamyšlení, jak by daný model mohl být vylepšen a jaké je jeho použití v reálném životě.

4.1 Datové soubory

Dataset[1] obsahuje čtyři souboru, které popisují zápasy ve hře **CSGO!**. K potřebám této bakalářské práce budou použity pouze soubory *players.csv* a *results.csv*. Ostatní dva soubory obsahují informace, které nezle k predikci v případě této bakalářské práce použít.

4.1.1 soubor players.csv

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

[1] https://www.kaggle.com/datasets/mateusdmachado/csgo-professional-matches

Transformovaný dataset má 10 proměnných, které unikátně identifikují statistiky každého hráče na určité mapě v jednom zápase. Interpretace je následující:

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

4.1.2 soubor results.csv

Druhý datový soubor, který je pro analýzu použit, obsahuje výsledky daných zápasů. Dataset se původně skládá z 45 773 řádků a 19 proměnných. Dataset obsahuje na rozdíl od datového souboru *players.csv* jeden chybný záznam. Dle něho hrál tým sám proti sobě, což nedává smysl. Jelikož je zápas na webovém portálu zadán správně, nejspíše se jedná o neznámou chybu, která nastala při exportu dat z webového portálu. Po transformacích vznikne tabulka o 8 proměnných a 91 436 řádcích. Každé pozorování identifikuje výsledek jednoho týmu v jednom zápasu na jedné mapě. Příklad je zobrazen v přiložené tabulce **??**. 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-Teroristů (1) či Teroristů (0)
- **team_rank** — rank týmu v okamžik, kdy se zápas hrál[4]
- **run_mean_3_months** — klouzavý průměr týmu za poslední tři měsíce

[2] hráč zabil nepřítele střelou do hlavy

[3] https://www.hltv.org/news/20695/introducing-rating-20

[4] https://www.hltv.org/news/16061/introducing-csgo-team-ranking

4.1.3 Omezení datasetu

Dataset obsahuje pozorování o zápasech a statistikách od konce roku 2015 do začátku roku 2020. Díky velkému počtu záznamů je možné chybné či neúplné záznamy smazat. To může nastat např. když tým má méně než 5 hráčů nebo když zápas nemá v každém týmu právě 5 hráčů. Tým může mít méně než 5 hráčů z toho důvodu, že je např. amatérsky[5]. Pokud má hráčů více, vyberou se hráči dle délky svého působení v týmu. Tým může mít v zápasu více než pět hráčů, pokud použijí náhradníka. Také jsou smazané záznamy, které neobsahují všechny potřebné statistiky. To může nastat u velmi historických zápasů, kde výpočet daných statistik ještě nebyl možný.

[5]neprofesionální, tím pádem nemusí mít všichni hráči na webovém portálu založená profil

4.2 Cíl analýzy

Cílem analýzy je vytvořit logistické regresní modely, které předpovídají pravděpodobnost výhry daného hráče či týmu. Pro předpověď se použijí statistiky hráčů za každou mapu a výsledné týmové statistiky za každou mapu. Statistiky hráčů, které jsou označovaný za prediktory jsou: *kills*, *assists*, *deaths*, *hs*, *fkdiff* a *rating*. Za týmové statistiky jsou pak považovány proměnné *starting_ct* a *run_mean_3_months*. Lze očekávat, že prediktory *kills*, *assists*, *hs*, *rating* a *fkdiff* budou šanci na výhru zvyšovat. Naopak prediktor *deaths* by měl výslednou šanci na výhru snížit.

Nehodila by se tato část spíše někde v úvodu bakalářské práce? První typ logistického modelu bude predikovat výhru zápasu na určité mapě podle individuálních statistik hráče a týmových prediktorů. Nebere se v potaz výkon spoluhráčů ani průměrný rank týmu za poslední tři měsíce. Cílem modelu je identifikovat, která statistika hráčů má na určité mapě největší vliv na výhru.

Druhý model se bude soustředit na predikci výhry týmu. K tom využije agregované statistiky hráčů, kteří jsou součástí týmu. Může se stát, že za tým historicky hrálo více hráčů. V případě, že tým vymění jednoho hráče za druhého, bude za tým historicky hrát 6 hráčů. V takovém případě model počítá s hráči, kteří za tým hrají nejdelší dobu. V případě schody se vezme hráč, který hrál v týmu jako poslední. Agregace statistik je spočítaná jako aritmetický průměr přes všechny zápasy, kde daná pětice za tým hraje. Dále jsou v modelu prediktory *starting_ct* a *run_mean_3_months*. Cíl druhého modelu je zjistit, jaké prediktory jsou pro předpověď výhry týmu významné. Model je vytvořen jak pro celý datový soubor, tak pro jeden vybraný tým.

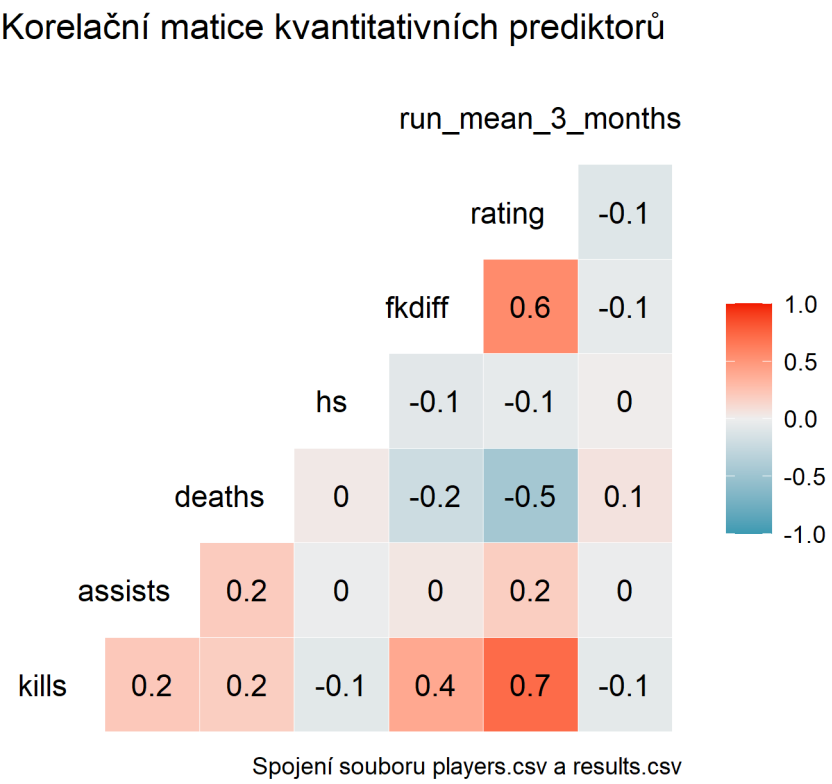
Poslední třetí logistický model je vytvořen pro jeden specifický tým, který bude předpovídat pravděpodobnost výhry proti historicky nejlepším třiceti týmům. Cílem tohoto modelu je zjistit, jak se liší významnost prediktorů, pokud tým hraje pouze proti těm nejlépe hodnocením týmům.

4.3 Průzkumová analýza dat

Průzkumová analýza slouží k vizualizaci prediktorů, hledání různých vztahů 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 nebyli lineárně závislé. Kombinaci korelací mezi kvantitativními prediktory lze zjistit z korelační matice.

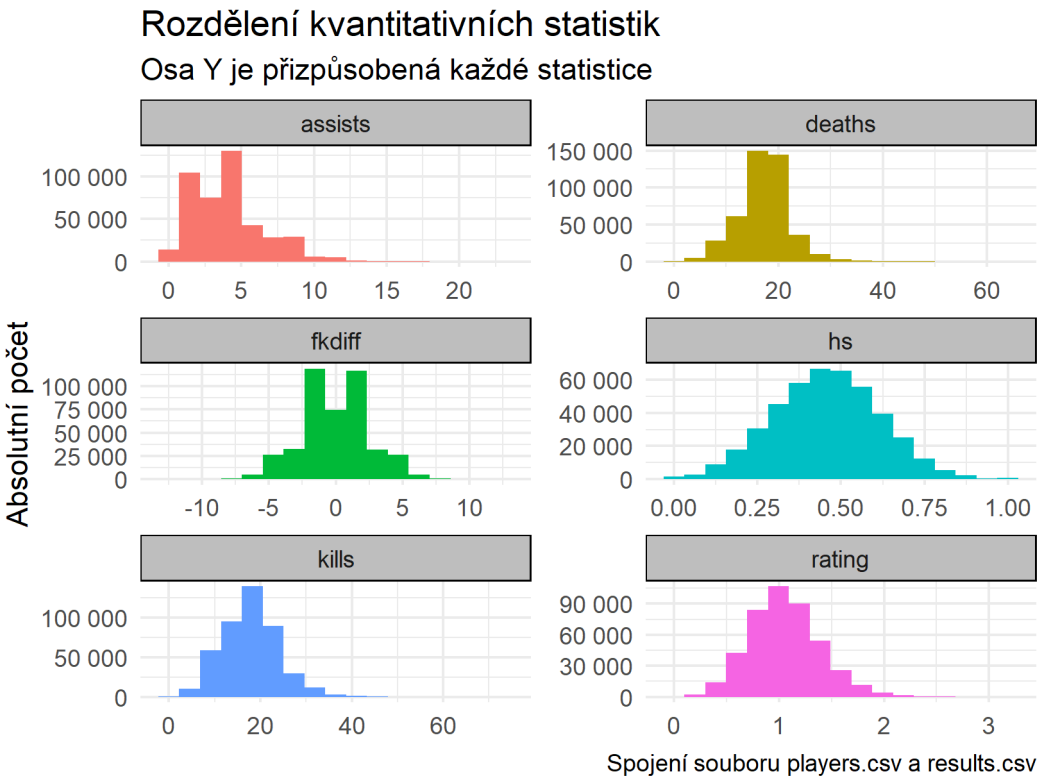


Obrázek 4.1: Korelační matice

Z korelační matice **??** lze vyčíst, korelace se blíží nule mezi průměrem týmu za poslední tři měsíce a agregovanými statistikami hráčů. Zároveň je vidět středně silná korelace mezi prediktorem *rating* a statistikami *fkdiff*, *deaths* a *kills*. Z tohoto důvodu je prediktor *rating* z tvorby modelů vyloučen.

4.3.2 Histogram kvantitativních prediktorů

Histogram kvantitativních prediktorů umožní zobrazit rozdělení prediktorů.



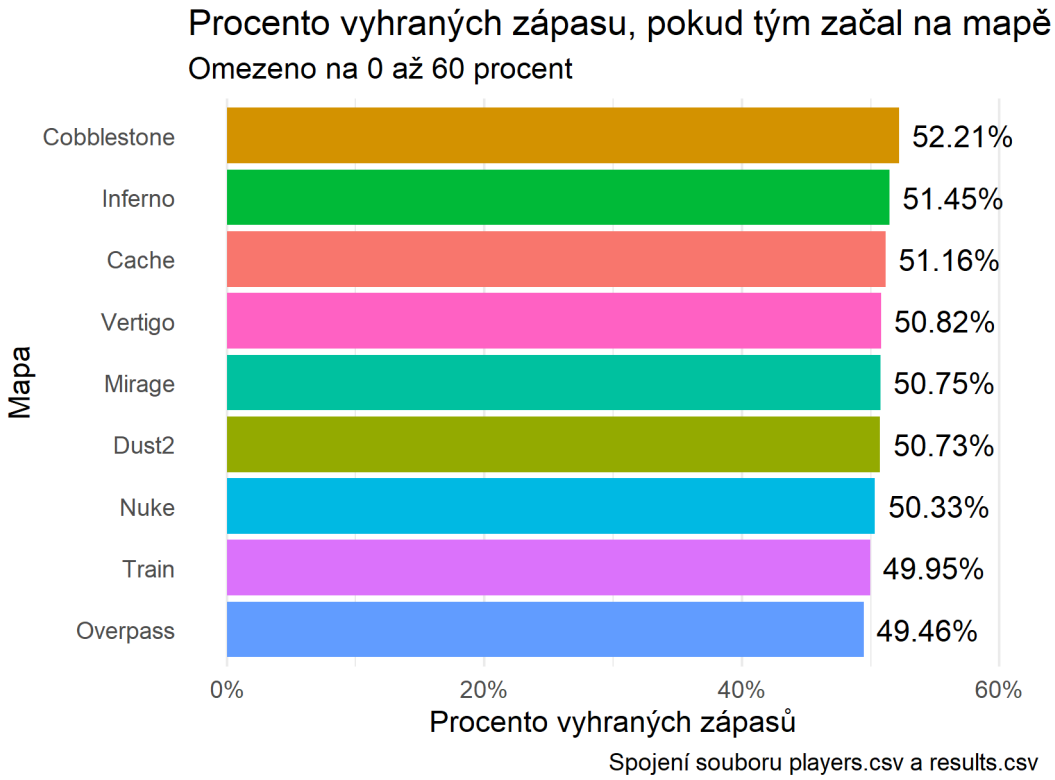
Obrázek 4.2: Histogram prediktorů

Histogram prediktorů *rating*, *hs*, *kills* a *deaths* má normální rozdělení a v proměnné se nenachází mnoho extrémních hodnot. Prediktor *fkdiff* má bimodální[6] rozdělení. Prediktor *assists* je sklonění doprava, což značí velké množství odlehlých či extrémních hodnot.

[6] rozdělení má dva vrcholy

4.3.3 Sloupcový graf výher přes počáteční stranu

Strana, na které tým začíná, může ovlivnit výsledek zápasu. Ovlivnění navíc může být rozdílné podle toho, na jaké mapě se zápas odehrává.



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

Počáteční strana má největší vliv na mapě Cobblestone, kde necelých 52 procent týmů, co začalo na straně Counter-Teroristů, mapu vyhrálo. Sloupcový graf naopak naznačuje, že je pro týmy nevýhodné začínat na mapě Overpass za stranu Counter-Teroristů. Historicky necelých čtyřicet-devět procent týmu, co na mapě začalo za Counter-Teroristy, mapu prohrálo.

5. Závěr

... Uzavření bakalářské práce

5.1 Závěrečné vyhodnocení modelu

... Výsledné vyhodnocení modelu pomocí všech statistik

5.2 Interpretace modelu do reálného světa

... Přenesení modelu do reálného světa

5.3 Použití modelu v reálném světě

... Použití modelu v reálném světě

5.4 Místo pro budoucí vylepšení

...

Seznam obrázků

Seznam tabulek

Seznam použitých zkratek

CSGO Coutner-Strike: Global Offensive
BR Battle Royale
MOBA Multiplayer online battle arena
FPS First-person shooter
TGNS Twin Galaxies National Scoreboard

Část I

Přílohy

A. Datové soubory

A.1 Transformovaný datový soubor players.csv

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

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

A.2 Transformovaný datový soubor results.csv

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

date	match_id	team	map	map_winner	starting_ct	team_rank	run_mean_3_months
2015-12-07	2299762	?	Overpass	1	1	2	2