

Report - STAR TREK

(Princípy dátovej vedy)

Petra Nagyová, 3.DAV

2. januára 2026

Obsah

1	Cieľ projektu a kladenie otázok	2
1.1	Hlavné výskumné otázky	2
1.2	Definície pojmov	3
2	Dáta a dátové zdroje	5
2.1	Rozsah analyzovaných seriálov	6
2.2	Jednoznačná identifikácia epizód v rámci projektu	7
2.3	Primárny zdroj: IMDb Non-Commercial Datasets	7
2.4	Doplňujúce zdroje pre získanie dodatočných informácií	8
3	Zber a formátovanie dát	9
3.1	Kroky zberu (ETL)	10
3.2	Výstupné tabuľky	12
4	Čistenie a normalizácia	14
4.1	Normalizácia postáv, identifikátorov a chýbajúcich hodnôt	14
4.2	Výber "relevantných" postáv	15
5	Metriky popularity epizód	16
5.1	Popularita epizódy (cieľové premenné)	16
5.2	Návrhy metrík popularity	17
5.3	Výber najvhodnejšej metriky	20
6	Analytické metódy a príprava na modelovanie	21
6.1	Exploratívna analýza dát (EDA) s grafmi	22
6.2	Konfundéry (confounders)	32
6.3	Baseline (modely)	33
6.4	Modely s postavami (od jednoduchých po silné)	34
7	Implementácia modelov	36
7.1	Konštrukcia programov	36
7.2	Vyhodnotenie modelov	37
7.3	Grafy modelov	38

8 Interpretácia výsledkov	41
8.1 Kontext seriálu	44
8.2 Záver a odpovede na položené otázky	49

1 Cieľ projektu a kladenie otázok

Cieľom projektu je preskúmať, do akej miery prítomnosť konkrétnych postáv v epizódach seriálov *Star Trek* súvisí s tým, či sú dané epizódy fanúšikmi vnímané ako obľúbené alebo neobľúbené. Popularitu epizód budeme aproximovať kvantitatívnymi ukazovateľmi dostupnými v otvorených dátach (najmä hodnoteniami a počtom hlasov), pričom prítomnosť postáv budeme reprezentovať pomocou informácií o obsadení (cast) na úrovni jednotlivých epizód.

Zámerom nie je iba opísať, ktoré epizódy sú hodnotené najlepšie alebo najhoršie, ale analyzovať vzťah medzi “obsahom” epizódy (z pohľadu postáv) a jej prijatím publikom. Pri interpretácii je potrebné rátať s tým, že hodnotenia môžu byť ovplyvnené viacerými faktormi (napr. sezóna, rok vysielania, typ epizódy, zmeny v publiku či rozdiely medzi seriálmi). Preto budú otázky formulované tak, aby umožňovali porovnanie s jednoduchými baseline prístupmi a aby výsledky bolo možné overiť robustnými štatistickými a modelovacími metódami.

V tejto kapitole definujeme hlavné výskumné otázky a základné pojmy, ktoré budú používané v ďalších častiach práce, aby bolo jasné, čo presne v projekte rozumieme pod pojmi ako popularita epizódy, prítomnosť postavy a vplyv postavy.

1.1 Hlavné výskumné otázky

Predtým, ako začneme skúmať obsah a štruktúru epizód *Star Treku*, položíme si otázky, ktoré budeme na konci analýzy vedieť zodpovedať a podložiť výsledkami. Výskumné otázky rozdeľujeme do tematických kategórií pre lepšiu prehľadnosť; zároveň nám pomôžu navigovať analýzu takým smerom, aby sme vedeli každú z nich vyhodnotiť.

1. Vzťahy medzi postavami a popularitou epizód

- Sú epizódy s výskytom postavy X v priemere hodnotené inak než epizódy bez postavy X ?
- Ktoré postavy sú najviac spojené s najlepšie, resp. najhoršie hodnotenými epizódami?
- Sú epizódy bez “hlavného obsadenia” (t. j. s minimom hlavných postáv) systematicky hodnotené horšie?

2. Robustnosť a vplyv vonkajších faktorov

- Pretrváva asociácia postáv s popularitou naprieč seriálmi, sezónami, rokmi vysielania či poradím epizódy v sezóne? (napr. môže sa stať, že finálne epizódy sezón sú v priemere populárnejšie bez ohľadu na zostavu účinkujúcich postáv.)

- Je medzi metrikami popularity (najmä hodnotením a počtom hlasov) pozorovateľný vzťah, a do akej miery sa tieto metriky správajú konzistentne? Vieme navrhnúť vhodnejšiu metriku?

3. Porovnanie naprieč seriálmi a sezónami

- Je “vplyv” tej istej postavy konzistentný naprieč viacerými seriálmi/časťami Star Trekovej “franchise” (ak sa postava vyskytuje vo viacerých)?
- Dá sa identifikovať “era efekt” — teda iné preferencie publika v rôznych dekádach vysielania?

4. Predikcia a praktická využiteľnosť modelov

- O koľko sa zlepší predikcia popularity epizódy, keď pridáme informácie o postavách oproti baseline (len seriál/sezóna/rok)?
- Ktoré postavy sú najdôležitejšie prediktory (“features”) podľa interpretovateľných modelov (napr. regularizovaná regresia)?

5. Extrémy, “fan service” a anomálie

- Sú nízko hodnotené epizódy spojené s konkrétnymi postavami alebo skôr s “anomálnymi” epizódami (odlišné obsadenie, experimentálne diely)?
- Sú epizódy s návratom obľúbenej postavy (“guest star”) spojené s nárastom počtu hlasov?
- Vyskytujú sa “outliers” epizódy, kde sú postavy prítomné, ale hodnotenie je opačné než by model očakával?
- Je neobľúbenosť niektorých epizód viazaná na konkrétne obdobia/produkčné zmeny skôr než na postavy?

1.2 Definície pojmov

Súčasťou projektu budú pojmy, ktoré je najlepšie na začiatku upresniť, aby sme si zjednotili význam a kontext. Definujeme teda zopár skratiek či pojmov, ktoré neskôr budeme využívať už bez upresnenia/vysvetlenia, kvôli jednoduchosti a lepšej čitateľnosti textu.

Nech pojem **epizóda** (epizóda seriálu) označuje základnú jednotku analýzy. Každý záznam v dátach bude predstavovať konkrétnu epizódu s vlastnými *metadátami*, ako sú napr. názov (**ep_title**), sezóna, poradie v sezóne, dátum vysielania, režisér (**director**) atď., spolu s informáciami o obsadení (a teda postavami). Epizódy budeme označovať jednoznačným identifikátorom (**ep_id**) z databázy IMDb, pričom pre epizódy je užitočné aj prepojenie s identifikátorom seriálu (**series**).

Pojem **seriál** znamená konkrétny titul v rámci franchise Star Treku – označený podľa unikátnej skratky ako **series**. Často pri referovaní na konkrétny seriál použijeme skratku názvu, ktorá bude uvedená v neskoršej kapitole s výberom dát a dátových zdrojov. **Sezóna** je oficiálne číslovanie epizód v rámci daného seriálu (daná číslom / integerom **season**).

Poradie epizódy v sezóne (označené číslom / integerom `order_ep`) určuje, koľká v poradí je konkrétna epizóda v určitej sezóne.

Termín **franchise** označuje súbor všetkých zahrnutých Star Trek seriálov analyzovaných v tomto projekte. V texte bude explicitne uvedené, ktoré seriály budú relevantné, keďže rozsah môže ovplyvniť konečné výsledky a ich interpretáciu. Všeobecne ale budeme rozumieť pod pojmom franchise všetky seriály, ktoré sú dokončené a odvysielané (úplne, teda dokonca).

Dátum vysielania je dôležitý na časové porovnanie epizód a na identifikáciu trendov naprieč rokmi / dekadami. Ide o deň, v ktorý bola konkrétna epizóda prvýkrát odvysielaná na vybranej mediálnej platforme a budeme ju označovať ako `air_date`.

V projekte rozlišujeme medzi pojmami **postava** a **herec**. Herec je konkrétna osoba z reálneho života (identifikovaná ako `actor`), zatiaľ čo postava je meno/rola, ktorú herec v danej epizóde stvárňuje (označujeme `character`). Toto rozlíšenie je nesmierne dôležité v rámci našej analýzy, keďže chceme zistiť údaje o popularite postáv a nie hercov. Jeden herec môže však hrať viacero postáv a jedna postava môže byť naprieč celou franchise stváraná viacerými hercami.

Definujme **prítomnosť postavy v epizóde** ako binárnu informáciu – teda postava je buď v epizóde uvedená, alebo nie. Navyše (ak v dostupných dátach bude táto informácia k dispozícii) môžeme rozlišovať aj medzi prítomnosťou v hlavnom obsadení (tzv. `main cast`) oproti vedľajšiemu. Povedzme teda, že prítomnosť postavy v epizóde (označujeme ju ako `role`) bude:

$$\text{role} = \begin{cases} 2 & \text{hlavné obsadenie} \\ 1 & \text{vedľajšie obsadenie} \\ 0 & \text{postava nie je prítomná v epizóde} \end{cases}$$

Obsadenie alebo **cast** predstavuje zoznam hercov a ich rolí, ktorí sú v epizóde uvedení v titulkoch (alebo ich vieme vytiahnuť z nejakej databázy týkajúcej sa konkrétnej epizódy). Z dátového pohľadu ide o množinu záznamov priradených ku konkrétnej epizóde – teda o set trojíc (postava, herec, prítomnosť postavy) – budeme označovať (`cast = {(character, actor, role)}, ...`).

Vplyv postavy v tejto práci neoznačuje priamu príčinnú súvislosť, ale štatistickú asociáciu medzi prítomnosťou danej postavy v epizóde a zvolenou metrikou popularity epizódy. Inými slovami, ak budeme hovoriť, že postava má „pozitívny“ alebo „negatívny vplyv“, myslíme tým, že epizódy, v ktorých sa postava vyskytuje, majú po zohľadnení kontrolných premenných (napr. seriál, sezóna, rok vysielania) tendenciu dosahovať vyššie alebo nižšie hodnoty popularity v porovnaní s epizódami bez tejto postavy.

Ako cieľovú premennú definujeme pojem **popularita epizódy** – označenie `popularity`. Budeme ju aproximovať metrikami dostupnými zo zdrojov ako je IMDb:

- *Hodnotenie* (označme ako `rating`) = priemerné hodnotenie používateľov na škále od 1 po 10 (kde 1 je najhoršie, 10 je najlepšie hodnotenie)

- *Počet hlasov* (označme ako `num_votes`) = počet používateľov, ktorí epizódu hodnotili

Tieto dve metriky reprezentujú odlišné aspekty popularity: hodnotenie skôr kvalitatívne hodnotenie publika, počet hlasov skôr mieru zapojenia alebo viditeľnosti / sledovanosti. Nutnosť normalizácie popularity (ak by sa teda mala skladať z hodnotenia a počtu hlasov) vyplýva z toho, že epizódy s väčším počtom hlasov by „prevažovali“ nad tými s menej hlasmi, čo by nevedlo k objektívnym výsledkom. Spôsobom, ako popularitu vypočítať a normalizovať sa budeme venovať v kapitole 4.1.

Zhrnutie definícií pojmov, ktoré budeme v projekte využívať, možno ilustrovať na nasledujúcom príklade – majme riadok v tabuľke, ktorý vyzerá nasledovne:

```
(series, ep_id, ep_title, season, order_ep,
    air_date, director, cast, popularity)
=
('DSN', 0040602, 'Rocks and Shoals', 6, 2, 1997-10-6, 'Michael Vejar',
{('Ben Sisko', 'Avery Brooks', 2), ('Constable Odo', 'Rene Auberjonois', 2),
 ('Lt. Cmdr. Worf', 'Michael Dorn', 1)}, (rating=8.5, num_votes=2600))
```

Takýto záznam označuje epizódu zo seriálu so skratkou názvu „DSN“ (ide o seriál „Deep Space Nine“), pričom je epizóde priradené unikátne id. Vieme vyčítať názov epizódy („Rocks and Shoals“) a určiť, že sa jedná o druhú epizódu šiestej sezóny, ktorá bola prvýkrát odvysielaná 6. októbra 1997 a režirovaná Michaelom Vejarom. Z obsadenia vidíme dve postavy aj s príslušnými hercami v hlavnom obsadení, ako aj jednu postavu s vedľajším obsadením. Keďže ešte nemáme určený vzorec na počítanie popularity ako jedného konkrétneho čísla, zo záznamu aspoň vieme, že hodnotenie epizódy dosiahlo 8.5 a epizóda bola hodnotená 2600 používateľmi.

2 Dáta a dátové zdroje

Táto kapitola opisuje, aké dáta sú potrebné na zodpovedanie výskumných otázok a v akej forme s nimi budeme pracovať. Keďže cieľom projektu je skúmať vzťah medzi prítomnosťou konkrétnych postáv v epizódach a popularitou jednotlivých epizód, potrebujeme:

- epizódové metadáta,
- informácie o obsadení na úrovni epizód,
- metriky popularity epizód.

K týmto údajom je zároveň potrebné mať stabilné identifikátory, aby bolo možné jednotlivé tabuľky spoľahlivo prepájať a vytvoriť reprodukovateľnú analytickú pipeline.

Základným objektom v dátach je epizóda. Pre každú epizódu budeme evidovať jej príslušnosť k seriálu (v rámci franchise), sezónu a poradie v sezóne, názov epizódy a časový údaj (rok, resp. dátum prvého vysielania). Na úrovni epizódy budeme ďalej potrebovať metriky popularity dostupné z verejných zdrojov, najmä priemerné hodnotenie používateľov a počet hlasov. Tieto ukazovatele budú predstavovať cieľové premenné v neskorších analýzach a modeloch.

Druhou kľúčovou skupinou dát sú informácie o postavách a hercoch. Pre každú epizódu potrebujeme zoznam postáv, ktoré sa v nej vyskytujú, a spôsob, ako túto prítomnosť kvantifikovať (napr. rozlíšenie medzi hlavným a vedľajším obsadením). Vzhľadom na to, že v dátach sa môžu vyskytovať rôzne zápisy toho istého mena postavy alebo rôzne typy rolí, budú tieto údaje ďalej čistené a normalizované, aby boli porovnateľné naprieč epizódami a seriálmi.

Z praktických dôvodov je vhodné od začiatku jasne vymedziť rozsah analyzovaných dát. V tomto projekte sa zameriame na **oficiálne televízne seriály** z franchise *Star Treku* a budeme pracovať s epizódami týchto titulov (bez filmov a bez neoficiálnych alebo fanúšikovských produkcií). Zahnuté seriály budú v texte explicitne uvedené a budú tvoriť fixnú množinu, s ktorou budeme pracovať vo všetkých krokoch analýzy. Takéto obmedzenie znižuje heterogenitu dát (napr. rozdiely medzi filmami a epizódami) a zároveň zjednodušuje interpretáciu výsledkov v jednotnom kontexte.

V ďalších podkapitolách následne predstavíme konkrétne dátové zdroje, z ktorých budú jednotlivé typy údajov pochádzať, a odôvodníme, prečo sú vhodné na riešenie stanovených výskumných otázok.

2.1 Rozsah analyzovaných seriálov

V praxi zároveň treba počítať s tým, že neexistuje jediný dátový zdroj, ktorý by pre každú epizódu poskytoval kompletnú kombináciu všetkých požadovaných informácií (metadáta, obsadenie na úrovni epizód, metriky popularity a presné dátumy vysielania). Z toho dôvodu bude potrebné pracovať s prienikom dostupných dát a projektovo zvoliť také obmedzenia, ktoré zabezpečia konzistentnú a porovnateľnú analytickú vzorku.

Aby boli výsledky interpretovateľné v jednotnom kontexte a aby sa minimalizovala heterogenita medzi rôznymi formátmi (napr. filmy vs. epizódy), zameriame sa na **oficiálne seriály**¹ franchise *Star Trek*. V ďalšej analýze budeme pracovať s pevne definovanou množinou seriálov uvedených v Tabuľke 1. Tieto seriály budú predstavovať základné „jadro“ dát pre všetky kroky čistenia, agregácie a modelovania.

¹https://en.wikipedia.org/wiki/List_of_Star_Trek_television_series

Index	Skratka	Oficiálny názov seriálu
001	TOS	Star Trek: The Original Series
002	TAS	Star Trek: The Animated Series
003	TNG	Star Trek: The Next Generation
004	DSN	Star Trek: Deep Space Nine
005	VOY	Star Trek: Voyager
006	ENT	Star Trek: Enterprise
007	DIS	Star Trek: Discovery
008	STT	Star Trek: Short Treks
009	PIC	Star Trek: Picard
010	LD	Star Trek: Lower Decks
011	PRO	Star Trek: Prodigy
012	SNW	Star Trek: Strange New Worlds

Tabuľka 1: Zoznam analyzovaných seriálov a ich skratky používané v projekte

2.2 Jednoznačná identifikácia epizód v rámci projektu

Hoci IMDb poskytuje vlastné identifikátory epizód, v praxi môže byť pri prepájaní tabuliek a pri práci s viacerými zdrojmi užitočné zaviesť aj projektové jednoznačné označenie epizódy. Preto definujeme **projektové ID epizódy** (označenie `proj_ep_id`) ako reťazec, ktorý vznikne z troch častí:

1. trojmiestny index seriálu podľa Tabuľky 1,
2. dvojmiestne číslo sezóny,
3. dvojmiestne poradie epizódy v sezóne (všetko s doplnením núl zľava).

$$\text{proj_ep_id} = \text{SSS} \parallel \text{SS} \parallel \text{EE}$$

kde **SSS** je index seriálu (napr. 004 pre DSN), **SS** je sezóna (napr. 06) a **EE** je poradie epizódy v sezóne (napr. 02). Teda druhá epizóda šiestej sezóny seriálu DS9 bude mať projektové ID:

$$\text{proj_ep_id} = 0040602.$$

Toto ID budeme používať ako stabilný kľúč v rámci projektu (najmä v analytických tabuľkách a vizualizáciách), pričom pôvodné IMDb identifikátory budú v dátach zachované na účely spätného dohľadania zdrojov a overenia.

2.3 Primárny zdroj: IMDb Non-Commercial Datasets

Hlavným zdrojom dát pre tento projekt budú **IMDb Non-Commercial Datasets**², ktoré poskytujú konzistentnú tabuľkovú reprezentáciu titulov, epizód, osôb a základných hodnotiacich metrik. Tento zdroj je vhodný najmä preto, že obsahuje široké pokrytie seriálov aj jednotlivých epizód a zároveň používa stabilné identifikátory (pre tituly/epizódy aj pre osoby), čo umožňuje spoľahlivé prepojenie informácií naprieč rôznymi typmi údajov.

Z IMDb datasetov budeme primárne využívať tri typy informácií:

² <https://datasets.imdbws.com/>

- **Epizódové metadáta** — základné informácie o epizódach a ich väzba na nadradený seriál (napr. názov epizódy, sezóna, poradie v sezóne, príslušnosť k seriálu).
- **Obsadenie a osoby** — prepojenie epizód na hercov (osoby) a ich roly, ktoré nám umožní odvodiť prítomnosť postáv v jednotlivých epizódach a vytvoriť reprezentáciu epizóda–postava.
- **Metriky popularity** — najmä priemerné hodnotenie epizódy a počet hlasov, ktoré budú tvoriť cieľové premenné v ďalšej analýze.

IMDb dáta budú zároveň slúžiť ako „chrbtica“ celého datasetu: projektové `ep_id` bude odvodené z príslušnosti epizódy k seriálu a z jej pozície v rámci sezóny, pričom pôvodné IMDb identifikátory (titulu/epizódy a osoby) ostanú zachované ako referenčné kľúče. Vďaka tomu bude možné jednotlivé tabuľky konzistentne prepájať aj pri kombinovaní IMDb s doplnkovými zdrojmi a v prípade potreby spätne dohľadať konkrétne záznamy v pôvodnom zdroji.

Zároveň však treba zdôrazniť, že IMDb datasety nemusia obsahovať všetky atribúty, ktoré sú pre projekt dôležité, najmä pokiaľ ide o **presné dátumy prvého vysielania** alebo niektoré špecifické metadáta vhodné na kontextovú interpretáciu epizód. Preto bude IMDb predstavovať primárny zdroj pre identifikáciu epizód, obsadenie a metriky popularity, zatiaľ čo vybrané doplnkové informácie budú doplnené z externých zdrojov tak, aby bola výsledná analytická vzorka čo najkompletnejšia a zároveň konzistentná.

2.4 Doplnujúce zdroje pre získanie dodatočných informácií

Keďže nie všetky požadované atribúty (najmä presné dátumy prvého vysielania) sú dostupné v jednom zdroji pre všetky epizódy, využijeme aj doplnkové zdroje. Primárnym doplnkovým zdrojom bude databáza *startrek-db*³, ktorá je dostupná vo forme SQLite databázy. Obsahuje metadáta pre epizódy a filmy v rámci franchise *Star Trek* a je určená na dotazovanie informácií ako zoznamy epizód, dátumy vysielania, „stardate“ a ďalšie súvisiace údaje.

Pre účely projektu je kľúčová najmä tabuľka `episode`, z ktorej vieme získať jednotnú a čitateľnú reprezentáciu epizódových metadát vo forme:

```
episode_id, title, airdate, remastered_airdate, season,
episode_number, production_code, stardate
```

Tieto atribúty budú využité najmä na doplnenie (resp. verifikáciu) časových údajov a na neskoršie analýzy trendov naprieč rokmi a dekadami. Databáza zároveň obsahuje aj ďalšie pomocné tabuľky súvisiace s médiami (napr. DVD/Blu-ray sety a ich obsah), ktoré môžu byť v prípade potreby využité pri interpretácii alebo pri doplnujúcich kontrolách konzistencie epizód (napr. pri neštandardných špeciáloch).

³https://fossil.2of4.net/_startrek-db/dir?ci=tip

Pri použití *startrek-db* budeme (rovnako ako pri ostatných zdrojoch) pracovať iba s epizódami patriacimi do nášho vopred definovaného rozsahu oficiálnych seriálov, aby boli výsledky konzistentné s vymedzením analyzovanej analytickej vzorky.

Po zjednotení základných dát (zoznam epizód, metadáta a obsadenie) budeme mať prehľad o všetkých epizódach aj o tom, ktoré postavy sa v nich vyskytujú. Na kvalitnejšiu **interpretáciu výsledkov** nám môžu pomôcť aj tzv. fanúšikovské a komunitné zdroje. Ide o webstránky a databázy vytvorené fanúšikmi alebo autormi recenzií, ktoré sú prirodzene subjektívne, no poskytujú kontext, ktorý z čisto kvantitatívnych dát často nevieme priamo vyčítať: napr. čo publikum považuje za silné/slabé stránky epizódy, aké sú opakujúce sa výhrady, alebo ktoré dejové línie a postavy vyvolávajú polarizáciu.

Takéto zdroje sú užitočné najmä pri vysvetľovaní „outlierov“ — napríklad ak by sme mali epizódu, ktorú model na základe obsadenia predikuje ako veľmi obľúbenú, no v skutočnosti má nízke hodnotenie (alebo naopak). V takýchto prípadoch vieme z diskusií a recenzií dohľadať možné dôvody odchýlky (napr. kontroverzný dej, nekonzistentné správanie postáv, produkčné rozhodnutia). Zároveň však tieto zdroje nebudeme používať ako primárne dáta pre modelovanie, ale ako doplnkový kvalitatívny materiál na interpretáciu.

Medzi fanúšikovské a komunitné zdroje, ktoré budú v projekte relevantné, patria:

- **Jammer’s Reviews**⁴ — v sekcii „Star Trek Reviews“ sú uvedené seriály aj epizódy, pričom ku každej epizóde je priradené hodnotenie (škála 1–5 hviezdíček od autora - Jamahl Epsicokhan) a často aj rozsiahla diskusia návštevníkov webu. Táto diskusia (aj zopár stoviek príspevkov na epizódu) môže poskytnúť argumenty a opakujúce sa témy vhodné na interpretáciu výsledkov.
- **Bjorn Munson – Every Episode of Every Star Trek Series, Ranked**⁵ — autor uvádza prehľad epizód naprieč seriálmi a postavami (vrátane stručného komentára), čo môže byť užitočné ako ďalší externý referenčný rámec pri kvalitatívnom porovnávaní vybraných epizód.
- **The Pensky Podcast**⁶ — podcast venovaný analýze epizód *Star Treku*, kde autori (Wes a Clay) diskutujú o detailoch jednotlivých epizód v približne hodinových epizódach. Na webe sa nachádzajú aj hodnotenia epizód (škála 1–5) s krátkym zôvodnením, ktoré môžu doplniť interpretáciu niektorých zistení.

3 Zber a formátovanie dát

V tejto kapitole popisujeme proces získania a prípravy dát potrebných na analýzu vplyvu postáv na hodnotenia epizód seriálov *Star Trek*. Naším primárnym zdrojom sú

⁴<https://www.jammersreviews.com/>

⁵<https://www.bjornmunson.com/2020/09/07/every-episode-of-every-star-trek-series-ranked/>

⁶<https://thepenskypodcast.com/>

IMDb Non-Commercial Datasets, ktoré poskytujú informácie o tituloch, epizódach, hodnoteniach používateľov a základných metadátach. Keďže však IMDb neposkytuje pre všetky epizódy konzistentné a ľahko použiteľné údaje o prvom vysielaní (resp. tieto údaje nemusia byť priamo dostupné v jednom mieste), doplníme ich o externú databázu *startrek-db* (SQLite), ktorá obsahuje zjednotenú evidenciu epizód v rámci celej franchise a umožňuje jednoduché dotazovanie dátumov vydania a ďalších epizódových atribútov.

Cieľom zberu nie je len stiahnuť dáta, ale aj ich **normalizovať a preformátovať** do jednotnej, vopred definovanej schémy, ktorá bude následne slúžiť ako vstup do analytickej časti projektu. V praxi to znamená, že jednotlivé zdroje najprv spracujeme samostatne (extrakcia relevantných tabuliek a atribútov), následne ich prepojíme na úrovni epizód a vytvoríme finálne tabuľky.

Výstupom bude konzistentný dataset, v ktorom má každá epizóda jednoznačný identifikátor, priradené metadáta (ako séria, sezóna, číslo epizódy, dátum vysielania) a cieľové premenné reprezentujúce popularitu (najmä IMDb hodnotenie a počet hlasov), doplnené o informáciu o výskyte postáv potrebnú pre ďalšie modelovanie.

3.1 Kroky zberu (ETL)

Budeme pokračovať podľa štandardného postupu spracovania dát, tzv. **ETL** alebo **Extract–Transform–Load**, bežne využívaného v dátovej vede. Tento prístup je vhodný kvôli reprodukovateľnosti procesu: pri rovnakých vstupoch a použití rovnakých kódov/programov dostane ľubovoľný používateľ rovnaké výstupy.

Prvým krokom je teda „Extract“ – **získanie surových dát zo zdrojov**. V rámci tohto projektu to znamená stiahnutie IMDb datasetov z vyššie uvedeného odkazu², z čoho získame 5 súborov:

- `name.basics.tsv`
- `title.basics.tsv`
- `title.episode.tsv`
- `title.principals.tsv`
- `title.ratings.tsv`

IMDb Non-Commercial datasety sú distribuované ako tabuľky vo formáte TSV (tab-separated values), pričom každý riadok reprezentuje jeden záznam a jednotlivé stĺpce obsahujú atribúty daného typu entity. V rámci projektu pracujeme s piatimi uvedenými súborami (uloženými v priečinku `imdb_raw`), ktoré pokrývajú základné metadáta o tituloch, epizódach, ľuďoch, obsadení a hodnoteniach. Kľúčové je, že IMDb používa konzistentné interné identifikátory: `tconst` pre tituly (vrátane epizód) a `nconst` pre osoby.

Tieto súbory však obsahujú *všetky* filmy, seriály a epizódy zahrnuté a uvedené v rámci IMDb databáz, čo znamená, že manuálne vyhľadávanie Star Trekových seriálov a príslušných epizód či obsadení je nemožné (keďže jednoduchý príkaz `Ctrl+F` zapríčiní zamrznutie

počítača na niekoľko desiatok sekúnd — prehľadáva totiž približne 14 miliónov riadkov, a to iba v jednom súbore).

Preto použijeme kód napísaný v Pythone (priložený k projektu pod názvom `imdb_process_raw.py`). Kód implementuje časť *Extract* aj *Transform* nad surovými datasetmi — prefiltruje ich (špecifickým vyhľadaním iba malej množiny záznamov). Najprv skript pri načítaní ošetruje špecifické chýbajúce hodnoty IMDb, ktoré sú v TSV reprezentované reťazcom `\N`. Tieto hodnoty konvertuje na štandardnú reprezentáciu chýbajúcich údajov (napr. `NA`), čím zjednodušuje následné filtrovanie a joinovanie tabuliek.

Následne skript identifikuje relevantné seriály *Star Trek*. Keďže súbor `title.basics.tsv` je veľký (ako aj ostatné štyri), spracováva sa po častiach (**chunkovanie**), aby bola pamäťová náročnosť zvládnuteľná. Záznamy sa filtrujú na základe pevne definovaného zoznamu presných názvov (`EXACT_TITLES`) pre 12 seriálov. Osobitne je ošetrovaný pôvodný seriál *Star Trek: The Original Series* tak, aby sa vybral konkrétny záznam typu `tvSeries` so `startYear = 1966` (t.j. *The Original Series*), čím sa predchádza nejednoznačnosti pri tituloch s podobným názvom.

Po identifikácii seriálov skript získa všetky epizódy pomocou `title.episode.tsv`. Vyfiltruje len tie riadky, kde `parentTconst` patrí medzi vybrané seriály. Výsledkom je množina epizódových identifikátorov `tconst` doplnená o `seasonNumber` a `episodeNumber`, čo tvorí základnú „mapu“ epizód v rámci jednotlivých seriálov.

Ďalej sa doplnia ratingy epizód. Zo súboru `title.ratings.tsv` sa ponechajú len tie záznamy, ktorých `tconst` zodpovedá vyfiltrovaným epizódam. Tým získame pre každú epizódu hodnotenie `averageRating` a počet hlasov `numVotes`, ktoré budú neskôr používané ako cieľové premenné popularity.

Ďalším krokom je extrakcia obsadenia a réžie epizód. Zo súboru `title.principals.tsv` sa ponechajú iba záznamy, ktoré patria k epizódam, a zároveň sa filtrujú kategórie tak, aby reprezentovali obsadenie a režiséra: `category` je obmedzené na `actor`, `actress`, `director`, `self`. Výsledkom je tabuľka väzieb epizóda–osoba (cez `tconst` a `nconst`) vhodná na ďalšie spracovanie výskytu postáv.

Nakoniec skript doplní mená osôb pre vyfiltrovaný cast. Z predchádzajúceho kroku sa vyberie množina všetkých potrebných `nconst` a zo súboru `name.basics.tsv` sa ponechajú iba záznamy pre tieto osoby. Tento krok výrazne zmenší objem dát a zároveň zachováva všetky informácie potrebné pre interpretáciu výsledkov (napr. mená hercov).

Výstupom skriptu je päť prefiltrovaných TSV tabuliek, ktoré sú už výrazne menšie než pôvodné surové datasety a tvoria vstup do ďalších krokov: `series.tsv`, `episodes.tsv`, `ratings.tsv`, `principals_cast.tsv`, `names.tsv` — všetky uložené v priečinku `imdb_processed`. Konkrétne počty uložených dát sú nasledovné:

Series:	12
Episodes:	958
Ratings rows:	942
Principals (cast) rows:	10738
Names:	1372

IMDb datasety sú vhodné na získanie hodnotení a obsadenia, avšak pre potreby projektu nám chýba jednotný a spoľahlivý atribút **presného dátumu prvého vysielania** pre všetky epizódy (a zároveň aj konzistentná štruktúra metadát naprieč celou franchise). Preto využívame doplnkový zdroj *startrek-db* (z vyššie uvedeného zdroja³) vo forme SQLite databázy, ktorá obsahuje zjednotený zoznam epizód a ich metadát pre celé univerzum *Star Trek*. Pre projekt je kľúčová najmä tabuľka **episode**, z ktorej vieme získať **airdate**, **season** a **episode_number**, a tabuľka **series**, ktorá poskytuje názvy seriálov a interval vysielania.

Databázu *startrek-db* preto spracovávame samostatným pythonovským skriptom - **startrek_db_process_raw.py**, pričom cieľom nie je exportovať celý obsah, ale iba minimálny výrez potrebný na prepojenie s IMDb (seriál, sezóna, číslo epizódy, názov epizódy, dátum vysielania, prípadne **stardate**).

Výstupom skriptu sú dva CSV súbory uložené v priečinku **startrek_processed/**. Prvý súbor **series.csv** obsahuje základný zoznam seriálov v tvare **series_id**, **series_title**, **series_begin**, **series_end**, teda jednoznačný identifikátor seriálu, jeho názov a interval vysielania. Druhý súbor **episodes.csv** obsahuje epizódy spolu s priradením k seriálu: **episode_id**, **series_id**, **series_title**, **season**, **episode_number**, **episode_title**, **airdate** (a doplnkovo aj **original_airdate**, **remastered_airdate**, **production_code**, **stardate**, **vignette**). Atribút **airdate** je určený ako preferovaný dátum vysielania (v prípade existencie využíva remastrovaný dátum), pričom dátumy sú uložené v štandardnom ISO formáte (YYYY-MM-DD). Tieto výstupy následne používame ako zdroj presných dátumov vysielania a na prepojenie s IMDb epizódami (najmä cez kombináciu **series_title**, **season** a **episode_number**); zároveň sú z exportu zámerne vylúčené série *Star Trek Continues* a *Star Trek: very Short Treks*, aby dataset zodpovedal oficiálnym seriálom analyzovaným v projekte.

3.2 Výstupné tabuľky

Cieľom tejto časti je spojiť prefiltrované IMDb dáta (epizódy, hodnotenia, osoby) s exportom z databázy *startrek-db* (najmä dátum vysielania) do jednotnej množiny tabuliek, ktoré budú priamo použiteľné pre ďalšiu analýzu a následné modelovanie. Implementácia je realizovaná s kódom **build_tables.py**.

Najskôr sa pre epizódy z IMDb doplní názov seriálu (cez **parentTconst** a tabuľku seriálov) a následne sa priradí projektová skratka seriálu podľa vopred definovaného mapovania (v kapitole 2.2 Jednoznačná identifikácia epizód v rámci projektu). Epizódy z IMDb sa potom spájajú s exportom zo *startrek-db* cez trojicu kľúčov

(**series**, **season**, **order_ep**),

kde **series** je skratka seriálu, **season** číslo sezóny a **order_ep** číslo epizódy v sezóne. Zo *startrek-db* týmto spôsobom získame najmä atribúty **ep_title** a **air_date**. Následne sa k epizódam z IMDb pripoja hodnotenia (**avg_rating**, **num_votes**) z nami vytvoreného súboru **ratings.tsv**.

Vytvárame teda tri finálne tabuľky:

- `episode` (1 riadok = 1 epizóda):

```
proj_ep_id, ep_id, series, ep_title, season, order_ep,
air_date, avg_rating, num_votes, popularity
```

Stĺpec `popularity` je zatiaľ ponechaný prázdny a dopĺňa sa až v neskoršej fáze (definícia metriky `popularity`).

- `director` (väzba epizóda–režisér):

```
proj_ep_id, ep_id, name
```

Z tabuľky `principals_cast.tsv` sa vyberú záznamy s `category = director` a `nconst` sa doplní meno cez `names.tsv`.

- `cast` (väzba epizóda–herec/postava):

```
proj_ep_id, nconst, name, character_name,
billing_order, is_main_cast
```

Z `principals_cast` sa ponechajú kategórie `actor`, `actress`, `self`. Meno herca sa doplní cez `names.tsv`. Pole `character_name` sa získava parsovaním atribútu `characters` (ak je dostupný), pričom sa používa prvý uvedený názov postavy. Atribút `billing_order` je odvodený z `ordering`. Premenná `is_main_cast` je heuristicky určená ako `billing_order ≤ 10`.

Výsledné tabuľky sa ukladajú v dvoch formách: ako CSV súbory do priečinka `project_tables_csv` a zároveň do SQLite databázy `project.db`. Pre efektívnejšie dotazovanie sa následne vytvoria indexy nad kľúčmi `ep_id`, `proj_ep_id` a `nconst`.

Na kontrolu konzistencie výslednej tabuľky `episode` sme po zostavení výstupov vypísali chýbajúce hodnoty v kľúčových atribútoch (`proj_ep_id`, `ep_title`, `air_date`, `avg_rating`, `num_votes`). Identifikované nezhody sme následne riešili dvomi spôsobmi: doplnením mapovania názvov seriálov na presné skratky používané v projekte (aby sa `series` zhodovalo so schémou v Tabuľke 1), a manuálnou korekciou malého počtu epizód, ktoré sa nedali spoľahlivo spárovať cez kľúč (`series`, `season`, `order_ep`).

Neúspešné párovania so *startrek-db* vznikali najmä v prípadoch dvojdielných alebo dvojdielkových epizód, kde môže byť číslovanie epizód v databázach posunuté (napr. *The Way of the Warrior* v DS9 je dvojdielková epizóda). Druhou skupinou boli epizódy z budúcich sezón, ktoré sa v čase zostavovania datasetu ešte nevyskytovali v *startrek-db* (napr. *Strange New Worlds* sezóna 4 je plánovaná na rok 2026), preto ich nechávame v datasete, ale s chýbajúcimi metadátami, aby sme nemuseli tieto epizódy úplne zahodiť.

4 Čistenie a normalizácia

Táto kapitola opisuje kroky, ktorými zjednocujeme a stabilizujeme spracované dáta tak, aby boli vhodné na porovnateľnú analýzu naprieč seriálmi a sezónami. Aj po extrakcii a spojení zdrojov sa v dátach vyskytujú nejednoznačnosti (napr. rôzne formáty názvov seriálov, duplicitné záznamy v obsadení, chýbajúce atribúty pri niektorých epizódach alebo rozdielne číslovanie epizód medzi zdrojmi). Cieľom čistenia je odstrániť zjavné chyby, obmedziť šum a zabezpečiť konzistentné identifikátory a kategórie, pričom sa snažíme minimalizovať manuálne zásahy a zachovať autentickosť dát.

V rámci projektu sa čistenie sústreďuje najmä na tri oblasti:

1. normalizáciu epizód a ich identifikátorov (napr. konzistentná skratka seriálu a `proj_ep_id`),
2. normalizáciu postáv a osôb v obsadení (napr. parsovanie a zjednotenie atribútu `character_name`, odstránenie duplikátov, definícia `is_main_cast`),
3. ošetrovanie chýbajúcich hodnôt a neprepojených epizód (napr. epizódy bez `air_date` alebo bez ratingu)

Výsledkom je sada tabuliek so stabilnou schémou, na ktorých vieme jednoznačne definovať metriky „vplyvu“ postáv a aplikovať analytické modely.

4.1 Normalizácia postáv, identifikátorov a chýbajúcich hodnôt

V tejto fáze projektu vykonávame dodatočné čistenie a kontrolu konzistencie nad výstupnými tabuľkami. Používame dva skripty: `normalize_cast.py` (normalizácia postáv a zjednotenie obsadenia) a `normalization_control.py` (kontrola identifikátorov epizód a diagnostika chýbajúcich hodnôt).

Program `normalize_cast.py` pracuje so vstupom `cast.csv` a vytvára dve reprezentácie názvu postavy: pôvodný reťazec `character_name_raw` (pre spätnú kontrolu) a znorlizovaný tvar `character_name_norm`. Normalizácia slúži najmä na odstránenie syntaktických rozdielov, ktoré by zbytočne rozdeľovali tú istú postavu na viac kategórií. Konkrétne:

- odstraňuje (koncovú) interpunkciu,
- čistí zátvorkové poznámky typu „(uncredited)“, „(voice)“ a podobne,
- **odstraňuje bežné hodnostné a funkčné prefixy** (napr. „Captain“, „Lt.“, „Cmdr.“, „Ensign“ vrátane variantu „jg – junior grade“, alebo tituly „Mr.“, „Dr.“ a „Doctor“ (s výnimkou prípadu postavy „The Doctor“)),

Pre vybrané prípady, kde sa vyskytujú stabilné aliasy, sa uplatňuje explicitné mapovanie na jeden cieľový tvar — zoberme si prípad hlavnej postavy, kapitána Jamesa T. Kirka, ktorý bol v pôvodnej databáze uvedený ako:

Captain James T. Kirk	...	83-krát
Capt. Kirk	...	22-krát
James T. Kirk	...	1-krát
Lieutenant James T. Kirk	...	1-krát
Lt. Cmdr. James T. Kirk	...	1-krát
Lieutenant Jim Kirk	...	1-krát

Vo vyčistenej tabuľke sa vyskytuje meno **James T. Kirk** 109-krát. Táto normalizácia nám pomáha najmä v tom, že teraz budeme presne vedieť, o akú postavu sa jedná — napr. nebude v 40 záznamoch postava s menom „Spock“ a v 102 riadkoch postava „Mr. Spock“, keďže sa očividne jedná o jednu a tú istú postavu. V našich dátach bolo zaznamenaných **313 typov zmien**, pričom najčastejšie išlo o odstránenie hodnotných prefixov (napr. „Captain Jean-Luc Picard → Jean-Luc Picard“ alebo „Lieutenant Worf → Worf“).

Následne skript vykoná tzv. deduplikáciu obsadenia v rámci epizód: dáta agreguje na úrovni (`proj_ep_id`, `nconst`, `character_name_norm`), ponechá najnižší `billing_order` a `is_main_cast` nastaví ako maximum. Výsledkom je súbor `cast_clean.csv` s 9683 riadkami, ktorý predstavuje očistenú tabuľku výskytov postáv vhodnú na ďalšie analýzy.

Kód `normalization_control.py` nadväzuje na predchádzajúce výstupy a vykonáva dve základné kontroly. Po prvé, načíta `episode.csv`, prepočíta (alebo doplní) chýbajúce hodnoty `proj_ep_id` podľa definície 7-miestneho kódu (trojmiestny index seriálu, dvojmiestne číslo sezóny, dvojmiestne číslo epizódy v sezóne) a uloží výsledok do `episode_clean.csv`. Zároveň kontroluje zdvojenia `proj_ep_id`, keďže ide o jednoznačný identifikátor epizódy v rámci projektu.

Po druhé, skript sa pozerá na chýbajúce hodnôt v epizódach a vypíše riadky, kde chýba `air_date` alebo hodnotenie (`avg_rating`, `num_votes`). V našom prípade bolo identifikovaných 16 epizód s chýbajúcimi údajmi, pričom všetky patria k seriálu *Strange New Worlds* v sezónach 4–5. Takéto epizódy ponechávame v datase, aby sme predišli strate záznamov, avšak pri analýzach, ktoré vyžadujú dátum vysielania alebo rating, budú tieto riadky filtrované.

4.2 Výber “relevantných” postáv

Keďže tabuľka `cast_clean.csv` obsahuje veľké množstvo epizodických a jednorazových postáv, zaviedli sme jednoduché kritérium „relevantnosti“ založené na **počte epizód, v ktorých sa postava vyskytuje**. Výber realizuje program `relevant_cast.py`.

Skript načíta `cast_clean.csv` a pripojí ku každému výskytu postavy skratku seriálu (`series`) cez tabuľku epizód `episode_clean.csv` (join cez `proj_ep_id`). Následne pre každú postavu agreguje:

- `n_episodes` – počet unikátnych epizód (`proj_ep_id`), v ktorých sa postava vyskytuje,
- `n_rows` – počet riadkov (výskytov) v `cast_clean.csv`

Ako prah relevancie sme zvolili `MIN_EPISODES = 3`. Výsledok prahovania je rozdelený do dvoch výstupných súborov:

- `relevant_characters.csv` – postavy s `n_episodes` ≥ 3 ,
- `filtered_out_characters.csv` – postavy s `n_episodes` < 3 .

V našich dátach to viedlo k **237 relevantným postavám** a 1578 vyradeným postavám.

Prahovanie slúži na redukciu šumu spôsobeného množstvom jednorazových postáv, ktoré by pri modelovaní zbytočne zvyšovali dimenzionalitu a zhoršovali interpretovateľnosť. Zároveň sme zámerne zvolili nízky prah, pretože aj postavy s výskytom v niekoľkých epizódach môžu byť pre fanúšikovský kontext a interpretáciu výsledkov dôležité (napr. majú samostatné profily na fanúšikovských wiki stránkach), a preto ich nechceme automaticky zahodiť. Typickými príkladmi sú vedľajšie, no opakujúce sa postavy ako *George Samuel Kirk*, *Zhaban* alebo rôzne pomenované funkčné role na lodi (napr. *Conn Officer*).

Vyfiltrované postavy nezahadzujeme avšak úplne, keďže v neskoršej analýze môže byť dôvodom, prečo bola epizóda napr. obľúbenejšia než zvyčajne, práve nejaká vedľajšia alebo epizodická postava, ktorá sa fanúšikom mimoriadne páčila, prípadne výskyt známeho herca v jednorazovej úlohe (tzv. „guest star“). Typickým príkladom je napríklad epizóda „Tsunkatse“ (z *Voyager*), v ktorej hostuje Dwayne „The Rock“ Johnson; ďalej sú to herci ako Tom Hardy, Kirsten Dunst, Christopher Lloyd či dokonca aj Stephen Hawking.

5 Metriky popularity epizód

V tejto časti projektu definujeme, ako budeme merať, že epizóda bola publikom „obľúbená“ alebo „neobľúbená“. Keďže ide o nepriamo pozorovateľný pojem, počítame ho pomocou dát dostupných v IMDb – predovšetkým priemerného hodnotenia epizódy a počtu hlasov. Cieľom je zvoliť takú metriku popularity, ktorá je prakticky použiteľná ako cieľová premenná v neskoršej analýze a zároveň je čo najviac robustná voči rozdielom medzi seriálmi a časovým posunom (napr. odlišná veľkosť publika a iné hodnotiace správanie naprieč dekadami).

Najprv preto zdefinujeme základné cieľové premenné (`avg_rating` a `num_votes`) a ich obmedzenia. Následne porovnáme viacero kandidátnych metrík popularity, ktoré tieto obmedzenia riešia rôznymi spôsobmi, a zdôvodníme ich vhodnosť pre podmienky nášho projektu. Informácia o tom, ktoré postavy sa v epizódach vyskytujú a ktoré považujeme za relevantné, je riešená samostatne v časti o výbere a reprezentácii postáv.

5.1 Popularita epizódy (cieľové premenné)

V tomto projekte je pojem „popularita“ interpretovaný ako reakcia publika meraná cez používateľské hodnotenia v databáze / systéme IMDb. Hlavnou cieľovou premennou

je **priemerné hodnotenie** zadané divákmi na stupnici od 1 (najhoršie) po 10 (najlepšie). Ide o najpriamejší aj najbežnejší spôsob hodnotenia epizód a je dostupné pre každú znamenajúcu epizódu, ktorú analyzujeme (v tabuľke `episode` v stĺpci `avg_rating`). Doplnkovou cieľovou premennou bude **počet hlasov**, ktorý zachytáva mieru pozornosti publika – t. j. koľko ľudí epizódu hodnotilo, čo môže nepriamo indikovať aj jej sledovanosť. Túto hodnotu máme uvedenú pre každú epizódu v tabuľke `episode` v stĺpci `num_votes`.

Vyskytuje sa však problém porovnateľnosti v čase a medzi seriálmi – rating aj hlasovanie sa môžu naprieč dekadami výrazne meniť. Star Trek je navyše dlhodobá franchise, takže jeho publikum sa od prvého vysielania (1966) postupne menilo a zároveň rástla aj veľkosť databázy a spôsob, akým používatelia IMDb hodnotia. Novšie epizódy tiež nemali toľko času nazbierať hlasy ako staršie epizódy, čo môže skresľovať porovnania podľa `num_votes`. Pri počte hlasov býva rozdelenie výrazne šikmé (s dlhým pravým chvostom), preto sa často používa logaritmická transformácia – pričom rátame s tým, že žiadna epizóda nemá menej ako 1 hlas:

$$\log(\text{num_votes})$$

Tento problém budeme riešiť tým, že analýzu budeme robiť zvlášť pre každý seriál (jeden model pre každý seriál), keďže jednotlivé seriály mohli byť publikom prijaté odlišne. Pre porovnávanie epizód v rámci konkrétneho seriálu budeme pracovať aj s *normalizovaným* hodnotením (napr. odchýlka od priemeru/mediánu seriálu alebo sezóny), aby sme zohľadnili rozdiely v ratingových baselinoch. Zároveň musíme brať do úvahy, že niektoré epizódy majú menej hlasov, čo zvyšuje neistotu ich hodnotenia; preto bude vhodné uvažovať o miernej korekcii alebo váhovaní podľa počtu hlasov v neskorších modeloch.

5.2 Návrhy metrík popularity

Sústredíme sa teraz na návrh konkrétnej metriky, ktorou budeme merať popularitu epizódy na základe hodnotenia a počtu hlasov. Ponúka sa viacero možností – pri každej je potrebné zvážiť jej vhodnosť v rámci nášho projektu, praktickú použiteľnosť v analýze a jej potenciálne nevýhody:

1. **Bayesovsky „zmenšený“ (shrinkage) rating** – v podstate ide o to, že priemerné hodnotenie je len odhad „skutočnej“ obľúbenosti epizódy a spoľahlivosť tohto odhadu závisí od počtu hlasov. Empirical Bayes (shrinkage)⁷ pristupuje k problému tak, že zabráňuje tomu, aby niekoľko extrémnych hlasov neprimerane ovplyvnilo celkové priemerné hodnotenie epizódy. Dosiahne to tým, že hodnotenie epizódy „priťahuje“ smerom k typickej hodnote (napr. k priemeru v rámci celého seriálu). Sila tohto priťahovania je však malá, ak má epizóda veľa hlasov (teda je menej pravdepodobné, že výsledok je spôsobený iba 1–2 extrémnymi hodnoteniami; skôr ide o konzistentný názor naprieč používateľmi).

Formálne sa to dá zapísať tak, že výsledný upravený rating je vážený priemer *priemerného hodnotenia epizódy* a *priemerného hodnotenia seriálu*, pričom váhy závisia

⁷Intuitívne vysvetlenie metódy Empirical Bayes shrinkage - <https://kiwidamien.github.io/shrinkage-and-empirical-bayes-to-improve-inference.html>.

od počtu hlasov:

$$\frac{\text{počet hlasov epizódy}}{\text{počet hlasov epizódy} + m} \cdot \text{priemerné hodnotenie epizódy} + \\ + \frac{m}{\text{počet hlasov epizódy} + m} \cdot \text{priemerné hodnotenie seriálu}$$

Teda - čím viac hlasov epizóda má, tým viac sa upravené hodnotenie približuje k jej vlastnému priemernému hodnoteniu; čím menej hlasov má, tým viac sa výsledok približuje k priemeru seriálu.

Nevýhodou tejto metriky je, že obsahuje voľbu hyperparametra (často označovaného ako m) – v podstate ide o pridanie m „virtuálnych“ hlasov s priemerným hodnotením seriálu. Tento hyperparameter určuje, aká prísna je korekcia: ak je m malé, shrinkage je slabý (výsledok sa veľmi nezmení); naopak ak je m veľké, vedie to k silnému priťahovaniu k priemeru. Je teda potrebné vhodne zvoliť m , prípadne použiť viacero hodnôt a experimentálne (robustne) overiť citlivosť výsledkov.

Pre účely nášho projektu je Bayesovsky zmenšený rating vhodný najmä vtedy, ak časť epizód má nízky počet hlasov alebo ak chceme znížiť vplyv šumu a extrémnych hodnotení na cieľovú premennú.

2. **Normalizovaný rating v rámci seriálu** – keďže seriály môžu mať rôznu „baseline“, teda jeden seriál je všeobecne hodnotený vyššie a iný nižšie, potrebujeme zaviesť relatívnu metriku, ktorá sa prispôsobí každému seriálu osobitne. Normalizáciu môžeme vykonať viacerými spôsobmi:

- **z-score** = hodnotenie prevedieme na jednotky štandardných odchýlok od priemeru (v rámci seriálu)⁸
- **odchýlka od mediánu** = hodnotenie – medián sezóny
- **percentil** = poradie epizódy v rámci seriálu

Hlavnou výhodou je porovnateľnosť a kontextualizácia (t. j. postavy porovnávame prirodzene na úrovni seriálu, nie globálne). Normalizácia zároveň znižuje vplyv rozdielov medzi seriálmi v rámci celej franchise a umožňuje interpretovať popularitu epizód relatívne voči tomu, čo je pre daný seriál typické.

Nevýhodou je interpretovateľnosť v absolútnych jednotkách: výsledkom nebude, že epizóda má hodnotenie 8.6, ale napríklad že epizóda je +0.7 štandardnej odchýlky nad priemerom seriálu alebo že sa nachádza v 85. percentile. Táto vlastnosť sa však dá ľahko vysvetliť pri záverečnej interpretácii výsledkov. Celkovo je táto metóda veľmi jednoduchá na implementáciu a nevyžaduje žiadne externé parametre (na

⁸Zhrnutie poznatkov o z-score - <https://www.datacamp.com/tutorial/z-score>

rozdiel od napr. shrinkage).

3. **Reziduál popularity (odchýlka od baseline)** – kým normalizácia iba v podstate preškáluje hodnoty v rámci vybranej skupiny, reziduál explicitne odpočíta očakávanú popularitu epizódy podľa známych trendov a kontextu. Po určení baseline definujeme reziduál ako rozdiel medzi skutočnou popularitou a baseline očakávaním: *“O koľko je epizóda horšia/lepšia, než by sme očakávali (podľa trendu)?”*⁹

Intuitívne si baseline môžeme predstaviť ako predikciu popularity, ktorú by epizóda dostala aj bez toho, aby sme vôbec brali do úvahy konkrétne postavy. Baseline môže zohľadňovať napríklad sezónu, poradie epizódy, rok vysielania, prípadne to, či ide o premiéru alebo finále sezóny. Následne reziduál zachytáva to, čo „ostane“ po odčítaní týchto systémových efektov.

Formálne môžeme reziduál zapísať ako:

skutočná popularita epizódy – očakávaná popularita podľa baseline

Ak ako popularitu používame napríklad priemerné hodnotenie epizódy, dostaneme:

$$\varepsilon_e = \text{rating}_e - \widehat{\text{rating}}_e,$$

kde $\widehat{\text{rating}}_e$ je baseline odhad (očakávané hodnotenie) pre epizódu e .

Baseline môže mať rôznu úroveň zložitosti. Najjednoduchší príklad je baseline daný **priemerom sezóny**:

$$\widehat{\text{rating}}_e = \overline{\text{rating}}_{\text{sezóna}(e)}, \quad \varepsilon_e = \text{rating}_e - \overline{\text{rating}}_{\text{sezóna}(e)}.$$

V tomto prípade sa pýtame: „Je táto epizóda nadpriemerná alebo podpriemerná vzhľadom na svoju sezónu?“

O niečo flexibilnejší baseline je trend podľa **poradia epizódy**, ktorý zohľadní, že začiatky seriálov alebo niektoré obdobia môžu mať systematicky iné hodnotenia. Zmysel je však stále rovnaký: baseline sa snaží popísať „bežný“ vývoj popularity a reziduál je „odchýlka“ od tohto bežného vývoja.

Reziduál popularity je užitočný najmä vtedy, keď vieme, že hodnotenia majú výrazné systematické vzory, ktoré nesúvisia priamo s postavami (napr. sezónnosť, rozdielne prijatie seriálu v rôznych obdobiach, alebo epizódy s prirodzene vyššou pozornosťou ako finále). Ak tieto vzory neodfiltrujeme, model by mohol mylne pripísať „vplyv“ postave, ktorá sa jednoducho častejšie vyskytuje v určitom type epizód (napr. v neskorších sezónach alebo vo finále). Reziduál preto zvyšuje férovosť porovnaní a zlepšuje interpretáciu výsledkov: postavy potom vysvetľujú skôr „nezvyčajnosť“ epizódy, nie jej časové alebo štrukturálne zaradenie.

⁹Prehľad definícií reziduálov - <https://www.displayr.com/learn-what-are-residuals/>

Reziduál závisí od toho, ako baseline zvolíme. Príliš jednoduchý baseline nemusí odstrániť všetky relevantné trendy. Naopak príliš zložitý baseline môže odobrať aj časť variability, ktorú by sme chceli pripísať postavám (t. j. môže „prečistiť“ dáta až príliš). Reziduál popularity je dobrý najmä preto, že franšíza Star Trek pokrýva viac dekád a viacero seriálov, kde očakávame rozdiely v publiku aj v kontexte vzniku epizód. Odpočítanie baseline (napr. aspoň na úrovni sezón) nám umožní sústrediť sa na otázku, či sa v epizódach s konkrétnymi postavami objavuje systematicky pozitívna alebo negatívna odchýlka od očakávaného hodnotenia.

5.3 Výber najvhodnejšej metriky

Keďže jednotlivé seriály (a často aj jednotlivé sezóny) majú odlišnú „baseline“ úroveň hodnotení, priame porovnávanie pôvodného IMDb ratingu medzi epizódami môže byť zavádzajúce. Pre účely nášho projektu preto zavádzame **relatívnu metriku popularity**, ktorá vyjadruje, o koľko je epizóda lepšie alebo horšie hodnotená oproti typickej úrovni „svojho kontextu“. Ako primárnu metriku zvolíme **z-score normalizáciu ratingu** v rámci skupiny (napr. *seriál*).

Z-score je definované ako počet štandardných odchýlok, o ktoré sa hodnotenie epizódy odlišuje od priemeru svojej skupiny:

$$z_e = \frac{\text{rating}_e - \mu_{g(e)}}{\sigma_{g(e)}},$$

kde $\mu_{g(e)}$ a $\sigma_{g(e)}$ sú priemer a štandardná odchýlka ratingov v skupine $g(e)$ (napr. konkrétneho seriálu). Intuitívne, $z_e = 0$ znamená „typickú“ epizódu danej skupiny, $z_e = +1$ znamená epizódu približne o jednu štandardnú odchýlku nad priemerom skupiny, a $z_e = -1$ analogicky pod priemerom.

Z-score umožňuje férovu porovnávať epizódy v rámci jedného seriálu/sezóny bez toho, aby vo výsledkoch dominovali rozdiely v globálnej obľúbenosti rôznych seriálov; poskytuje priamo interpretovateľnú škálu „nadpriemernosti“ v jednotkách štandardných odchýlok, čo sa dobre používa v regresných modeloch a pri vizualizácii efektov postáv; nevyžaduje žiadne externé hyperparametre (na rozdiel od shrinkage).

Je citlivejšie na extrémne hodnoty (outliery) než napr. odchýlka od mediánu. Ak má niektorá skupina veľmi málo epizód alebo takmer nulový rozptyl ratingov, odhad σ môže byť nestabilný (v extrémne $\sigma = 0$). V implementácii preto ošetríme prípady s nulovým alebo chýbajúcim rozptylom a z-score pre takú skupinu nastavíme ako chýbajúce (NA), prípadne použijeme sekundárnu normalizáciu na úrovni celého seriálu.

V kóde `popularity.py` definujeme a vypočítame cieľovú premennú `popularity`, ktorá reprezentuje relatívnu obľúbenosť epizódy v rámci jej seriálu. Vstupom je tabuľka `episode_clean.csv`, ktorá obsahuje (okrem iného) stĺpce `series` a `avg_rating`. Pre každý seriál vypočítame **priemer a štandardnú odchýlku hodnotení naprieč jeho epizódami a použijeme z-score normalizáciu**. V implementácii navyše ošetrujeme prípady, keď je štandardná odchýlka v rámci seriálu nulová alebo chýbajúca (nulová alebo NaN), pretože v takom

prípade by z-score nebolo definované. Pre takéto skupiny nastavíme `popularity` na `NaN`. Výstupom skriptu je súbor `episode_popularity.csv`, ktorý je identický so vstupnou tabuľkou, ale obsahuje nový stĺpec `popularity`. Na záver skript vypíše základné deskriptívne štatistiky tejto premennej:

<code>count</code>	9.420000e+02
<code>mean</code>	2.847451e-16
<code>std</code>	9.941380e-01
<code>min</code>	-4.250781e+00
<code>25%</code>	-7.024317e-01
<code>50%</code>	-1.683954e-02
<code>75%</code>	6.629700e-01
<code>max</code>	3.154904e+00

Teda popularitu sa podarilo vypočítať pre 942 epizód, v rámci každej skupiny (seriálu) má normalizovaná premenná priemer približne nula a `std` = 0.994 \approx 1 odpovedá tomu, že z-score je škálované štandardnou odchýlkou; hodnota blízka 1 potvrdzuje, že normalizácia bola úspešná. Kvartily ukazujú typický rozsah odchýlok: medián je približne 0 (−0.0168), 50% epizód sa nachádza približne medzi −0.70 a +0.66. To znamená, že väčšina epizód je *hodnotená relatívne blízko priemeru svojho seriálu*. Minimum (−4.25) a maximum (+3.15) indikujú existenciu niekoľkých výrazných extrémov: epizódy, ktoré sú v rámci svojho seriálu výrazne podpriemerné (viac než 4 štandardné odchýlky pod priemerom) alebo výrazne nadpriemerné (viac než 3 štandardné odchýlky nad priemerom). Takéto epizódy sú kandidátmi na neskoršiu analýzu, pretože môžu súvisieť so špecifickými prvkami (napr. výrazné dejové udalosti, hosťujúci herci / „guest stars“ alebo konkrétne postavy).

6 Analytické metódy a príprava na modelovanie

V tejto časti opisujeme analytické postupy, pomocou ktorých skúmame vzťah medzi výskytom konkrétnych postáv a popularitou epizód. Cieľom nie je iba nájsť korelácie, ale navrhnúť metodiku, ktorá zohľadňuje špecifiká seriálovej produkcie (časový trend, rozdiely medzi seriálmi, zmeny obsadenia) a umožňuje interpretovať výsledky aj v kontexte dátovej vedy.

Analýzu vedíme od jednoduchých deskriptívnych pohľadov k modelom, ktoré kvantifikujú asociáciu postáv s popularitou epizód a zároveň kontrolujú potenciálne konfundéry. Základnou jednotkou analýzy je epizóda; pre každú epizódu pracujeme s cieľovou premennou popularity (normalizované hodnotenie v rámci seriálu) a s množinou vysvetľujúcich premenných reprezentujúcich postavy (napr. binárna prítomnosť postavy v epizóde alebo agregované charakteristiky obsadenia či relevantnosť).

Metódy sú rozdelené do logických krokov: najprv exploratívne overujeme základné vzory v dátach (rozptyl popularity, distribúcie výskytu postáv, časové trendy). Následne definujeme baseline prístupy, ktoré poskytujú referenčnú úroveň výkonu a slúžia ako kontrola proti preceňovaniu náhodných efektov. V ďalšom kroku používame modely s explicitnými premennými pre postavy (od jednoduchých lineárnych modelov až po regularizované

modely vhodné pre vysokú dimenziu), aby sme získali odhady „vplyvu“ postáv. Napokon rozoberieme konfundéry a robustnosť: overujeme, do akej miery je pozorovaný efekt stabilný po započítaní seriálu, sezóny, poradia epizódy, roku vysielania a ďalších faktorov, ktoré môžu ovplyvňovať hodnotenia nezávisle od postáv.

Výsledkom tejto časti je súbor interpretovateľných metrík a modelových odhadov, ktoré umožňujú porovnať postavy medzi sebou a identifikovať postavy asociované s nadpriemerne hodnotenými (alebo podpriemerne hodnotenými) epizódami.

6.1 Exploratívna analýza dát (EDA) s grafmi

Exploratívna analýza dát (Explorative Data Analysis - skrátene EDA) predstavuje podstatnú fázu práce s dátami, ktorej cieľom je získať **prvý kvalitný obraz** o dátovej množine ešte pred aplikáciou formálnych modelov. V praxi ide o kombináciu deskriptívnych štatistík a vizualizácií, pomocou ktorých sumarizujeme hlavné charakteristiky dát (napr. rozdelenia premenných), identifikujeme vzory a vzťahy, odhaľujeme anomálie (outliery) a zároveň diagnostikujeme problémy kvality dát (chýbajúce hodnoty, nekonzistentné formáty, duplicitné záznamy a pod.).¹⁰

V kontexte tohto projektu EDA slúži ako ukazovateľ pre návrh modelov vplyvu postáv tým, že ukáže:

- typické rozsahy a rozdelenia hodnotení v rámci jednotlivých seriálov,
- rozdiely v počtoch hlasov medzi epizódami a seriálmi,
- časové trendy (zmeny hodnotení v priebehu sezón a rokov),
- výskyt extrémne populárnych/nepopulárnych epizód, ktoré môžu byť spojené so špecifickými postavami alebo udalosťami v príbehu.

V ďalšom kroku preto prejdeme od všeobecných tvrdení ku **konkrétnym vizualizáciám** nad pripravenými tabuľkami `episode_popularity`, `director` a `relevant_characters`. Použijeme jednoduché a interpretovateľné grafy (histogramy, boxploty, scatterploty, stĺpcové grafy).

Rozdelenie popularity epizód - ako prvý krok analyzujeme základnú cieľovú premennú — `popularity`. Ide o z-score štandardizáciu v rámci seriálu, preto očakávame, že väčšina epizód bude sústredená okolo hodnoty 0 a rozdelenie bude mať približne jednotkový rozptyl. Na Obr. 1 je histogram popularity pre všetky epizódy spolu.

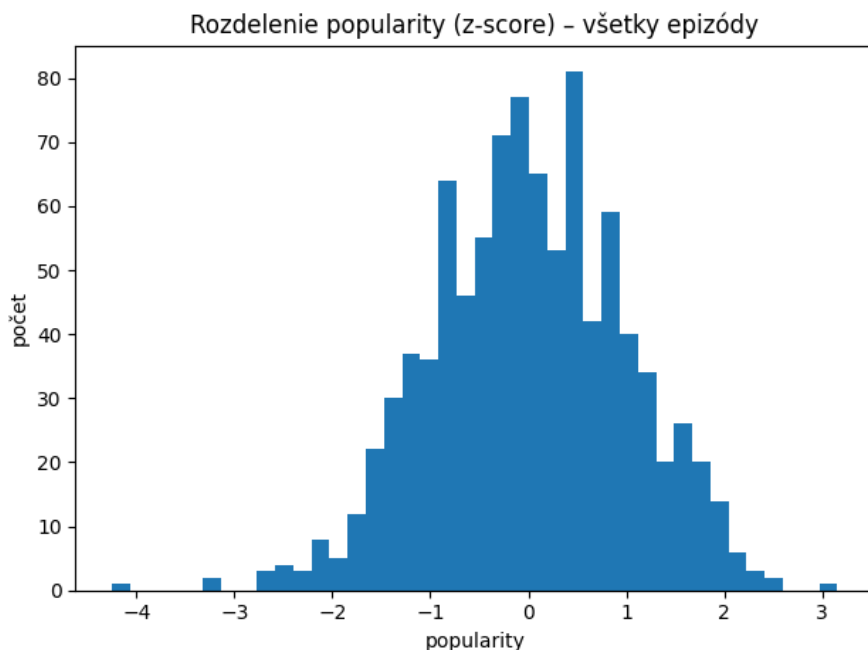
Z histogramu vidíme, že:

- rozdelenie je **unimodálne** a sústredené okolo 0, čo je konzistentné s použitou normalizáciou v rámci seriálov,

¹⁰Články opisujúce implementácie EDA:

<https://medium.com/data-science/a-data-scientists-essential-guide-to-exploratory-data-analysis-25637>

<https://www.analyticsvidhya.com/blog/2021/04/mastering-exploratory-data-analysiseda-for-data-science>



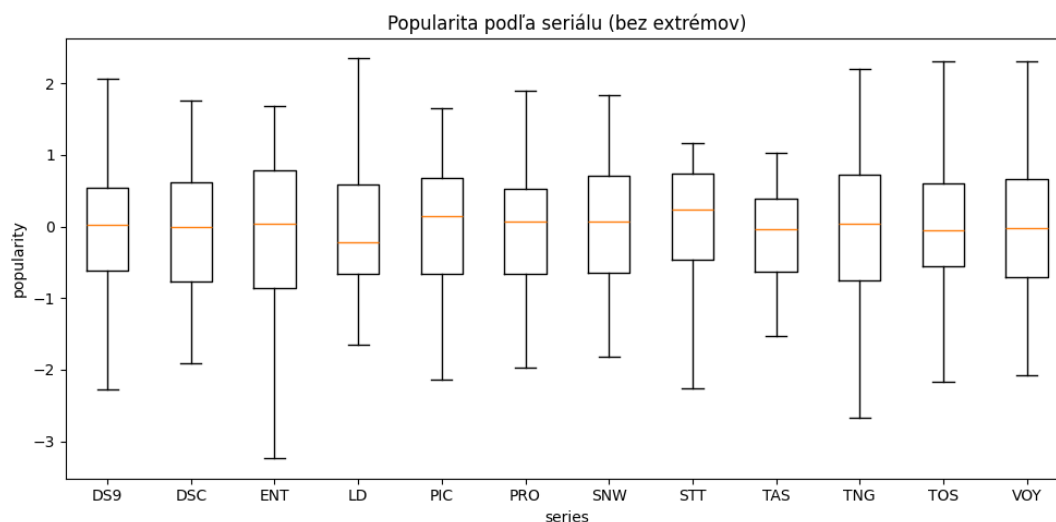
Obr. 1: Rozdelenie popularity (z-score) pre všetky epizódy

- väčšina epizód leží v intervale približne $[-2, 2]$, teda nie viac než dve smerodajné odchýlky od priemeru svojho seriálu,
- v dátach sa vyskytujú aj **výrazné odľahlé hodnoty** (napr. okolo -4 a okolo $+3$), ktoré reprezentujú epizódy hodnotené extrémne nízko/vysoko vzhľadom na svoj seriál; práve tieto prípady sú dôležité pre neskoršiu interpretáciu a kontrolu robustnosti modelov.

Popularita podľa seriálu – ako ďalší krok porovnávame rozdelenie normalizovanej popularity medzi jednotlivými seriálmi pomocou boxplotu. Tento graf slúži ako kontrola toho, či z-score štandardizácia v rámci seriálu vedie k porovnateľným rozdeleniam a zároveň umožňuje rýchlo identifikovať seriály s väčšou variabilitou alebo s výraznejšími odľahlými epizódami. Na Obr. 2 je zobrazený boxplot popularity podľa seriálu, pričom extrémne hodnoty (outliery) sú v grafe skryté (`showfliers=False`), aby bolo lepšie vidieť typické rozdelenie.

Z boxplotu vidíme, že:

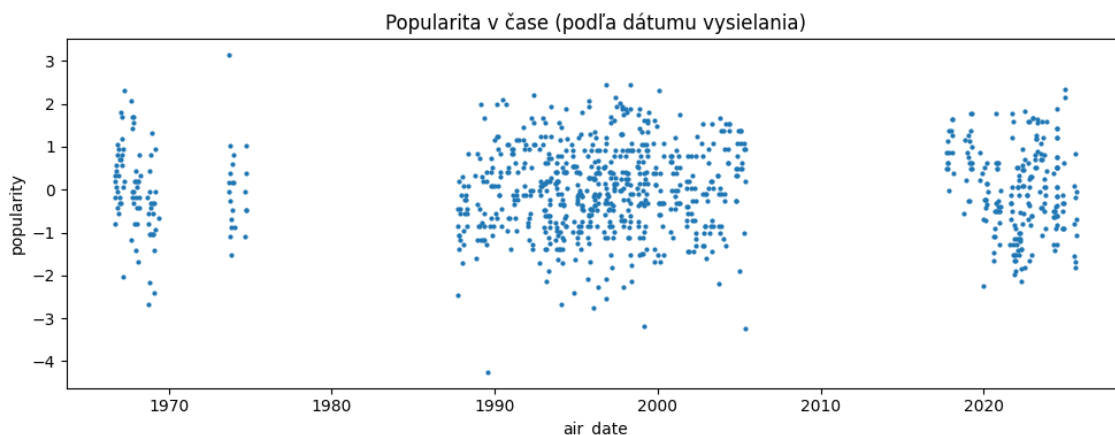
- mediány popularity sú pri väčšine seriálov **blízko nuly**, čo zodpovedá očakávaniu pri štandardizácii v rámci seriálu,
- šírka boxu (interkvartilové rozpätie) je síce vo väčšine seriálov porovnateľná, no viditeľne širší box má napr. ENT a TNG (väčšia variabilita „typických“ epizód), zatiaľ čo užší box pozorujeme napr. pri TAS (kompaktnejšie rozdelenie),
- dĺžky „chvostov/fúzov“ sa líšia najmä pri seriáloch TNG, TOS a VOY, kde fúzy siahajú ďalej do oboch smerov (viac epizód výraznejšie nad/pod priemerom seriálu), kým pri TAS a STT sú fúzy kratšie, čo naznačuje menší rozsah odchýlok od priemeru,



Obr. 2: Popularita podľa seriálu (boxplot bez extrémov)

- **asymetria** je výraznejšia napr. pri ENT a STT, kde je dolný fúz dlhší než horný (silnejší „chvost“ smerom k nízko hodnoteným epizódam), zatiaľ čo pri LD, VOY a TOS je naopak mierne dominantný horný fúz (relatívne viac epizód nad priemerom než hlboko pod priemerom).

Popularita v čase (podľa dátumu vysielaania) – v ďalšom kroku sledujeme, ako sa hodnota popularity správa v čase podľa dátumu prvého vysielaania epizódy (**air_date**). Každý bod v grafe predstavuje jednu epizódu a jej normalizovanú popularitu. Cieľom tejto vizualizácie je zachytiť časové štruktúry v dátach.



Obr. 3: Popularita v čase podľa dátumu vysielania (`air_date`).

Z grafu vidíme, že epizódy sú prirodzene zoskupené do niekoľkých časových **blokov**, ktoré zodpovedajú hlavným obdobiam vysielania seriálov *Star Trek*:

- koniec 60. rokov (TOS) – menší počet bodov, ale stále s výraznou variabilitou popularity,
- 90. roky a začiatok 2000 / nultých rokov (TNG, DS9, VOY, ENT) – najhustejšia časť grafu s veľkým počtom epizód a širokým rozpätím popularity,
- obdobie po roku 2017 (moderné seriály ako DSC, PIC, SNW, PRO a ďalšie) – opäť samostatný blok s vlastným rozptylom.

Dôležité je, že **čas nie je v dátach rovnomerne pokrytý**: medzi týmito blokmi sú veľké medzery (napr. 70.–80. roky alebo približne 2006–2016). Zároveň v niektorých obdobiach pozorujeme aj extrémne nízke alebo extrémne vysoké hodnoty popularity (napr. jednotlivé body okolo -3 alebo $+2$), ktoré budú vhodné neskôr preskúmať detailnejšie (napr. konkrétne epizódy, nízky počet hlasov, alebo špecifický kontext seriálu).

Popularita podľa poradia epizódy – okrem dátumu vysielania sledujeme aj jednoduchšiu časovú štruktúru: poradie epizódy v sezóne (stĺpec `order_ep`). Tento graf je užitočný najmä na odhalenie tzv. „position effects“ – napr. či majú prvé alebo posledné epizódy v sezónach tendenciu byť hodnotené odlišne (premiéry, finále), prípadne či sa v určitých častiach sezón častejšie vyskytujú extrémny.

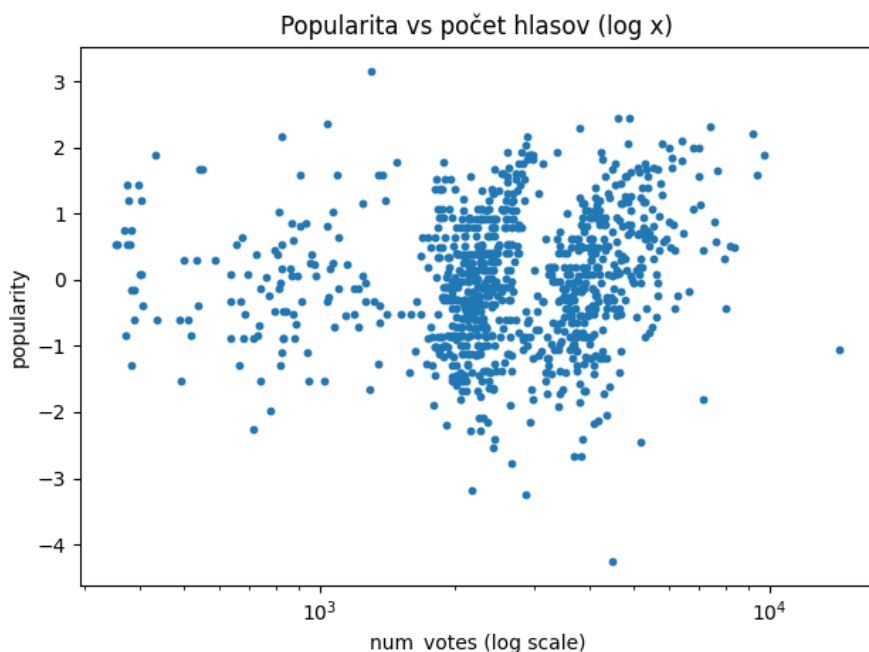
Z grafu na Obr.4 vidíme, že vertikálne „stĺpce“ bodov sú prirodzené: `order_ep` nadobúda iba diskkrétne hodnoty (1, 2, 3, ...), pričom každá hodnota agreguje epizódy zo všetkých seriálov a sezón, ktoré majú epizódu na danej pozícii. Tiež vieme vyčítať, že:

- hodnoty popularity sú pri väčšine poradí epizód sústredené okolo nuly a nevykazujú jednoznačný monotónny trend (t. j. nevyzerá to tak, že by epizódy systematicky „stúpali“ alebo „klesali“ s poradím),
- pri niektorých poradových pozíciách sa objavujú výrazné extrémny - najmä v strednej až neskoršej časti sezóny: najnižšie hodnoty (okolo -4) vidíme približne pri `order_ep` ≈ 22 , a ďalšie výrazne negatívne body (pod -3) sa objavujú aj okolo `order_ep` ≈ 18 až 23. Na opačnej strane sa ojedinelé veľmi vysoké hodnoty (nad $+2$) vyskytujú skôr v skorých pozíciách (`order_ep` ≈ 2).



Obr. 4: Popularita podľa poradia epizódy v sezóne (`order_ep`).

Popularita a. počet hlasov (`num_votes`) – ďalšou dôležitou témou je vzťah medzi normalizovanou popularitou epizódy a počtom hlasov na IMDb. Počet hlasov je významný najmä preto, že epizódy s veľmi nízkym `num_votes` môžu mať nestabilnejšie hodnotenie (vyšší šum) a extrémny popularity môžu byť čiastočne spôsobené malou vzorkou hodnotiacich. Na Obr. 5 je zobrazený scatterplot `popularity` oproti `num_votes`, pričom os x je v logaritmickej mierke.



Obr. 5: Vzťah popularity epizódy a počtu hlasov (`num_votes`)

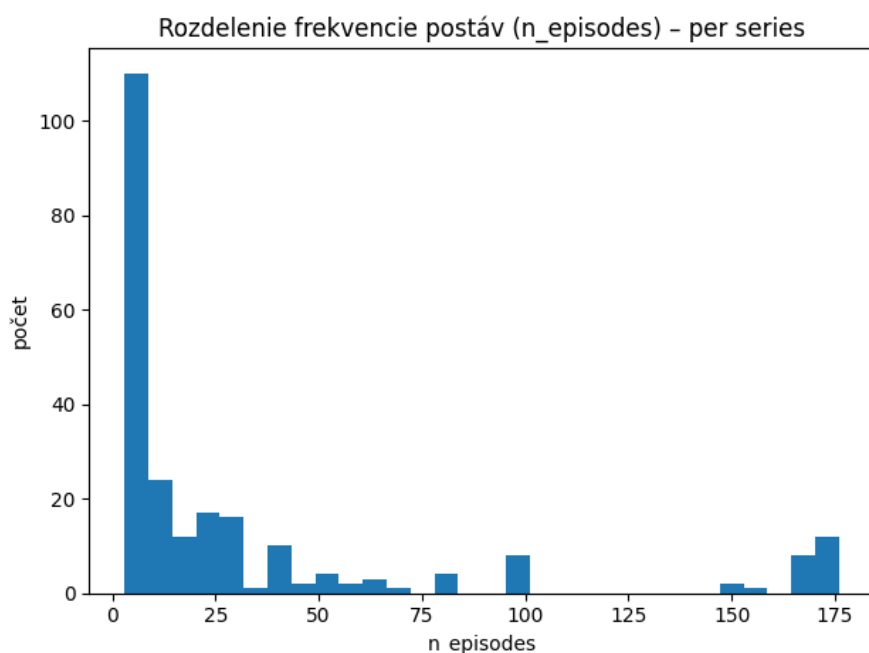
Na grafe pozorujeme:

- väčšina epizód je koncentrovaná v pásme približne 10^3 až 10^4 hlasov, pričom v tomto rozsahu sú hodnoty popularity rozptýlené okolo nuly bez výrazného monotónneho trendu,
- pri nižších hodnotách `num_votes` (vľavo) je rozptyl popularity vizuálne o trochu

väčší a častejšie sa objavujú extrémny - ale smerom na **kladnú** stranu popularity, čo je konzistentné s hypotézou vyššej neistoty hodnotenia pri menšom počte hodnotiacich,

- extrémne negatívne epizódy (napr. okolo -4) sa vyskytujú aj pri stredných až vyšších počtoch hlasov, teda nejde iba o trend malého `num_votes`; takéto epizódy môžu reprezentovať skutočne „kontroverzné“ alebo široko neoblíbené diely.

Rozdelenie frekvencie postáv – po analýze cieľovej premennej popularity sa zameriavame na štruktúru vysvetľujúcich premenných, konkrétne na to, ako často sa relevantné postavy objavujú v epizódach. Premenná `n_episodes` v tabuľke `relevant_characters` vyjadruje počet epizód, v ktorých sa daná postava vyskytuje **v rámci konkrétneho seriálu**. Obr. 6 zobrazuje histogram týchto frekvencií.



Obr. 6: Rozdelenie frekvencie relevantných postáv podľa počtu epizód (`n_episodes`) v rámci seriálu

Histogram vykazuje tzv. **long-tail** správanie: väčšina postáv sa vyskytuje iba v malom počte epizód a len malá časť postáv tvorí „hlavné obsadenie“ s výskytom v desiatkach až stovkách epizód. Tento obraz potvrdzujú aj deskriptívne štatistiky:

- počet relevantných záznamov (postava–seriál) je 237,
- medián je 10 epizód, pričom dolný kvartil je 5 epizód,
- priemer (≈ 33.6) je výrazne vyšší než medián, čo je typické práve pri dlhom „chvoste“ smerom k vysokým hodnotám,
- maximum je 176 epizód, čo zodpovedá postavám prítomným takmer v celom seriáli.

Popri tom sledujeme aj konkrétne epizódy, ktoré majú súčasne **nízky počet hlasov** a **extrémnu popularitu**, keďže práve pri nich môže byť hodnotenie najviac ovplyvnené šumom. V našich dátach je 5% kvantil `num_votes` rovný 762 a v tejto kategórii sa objavila epizóda:

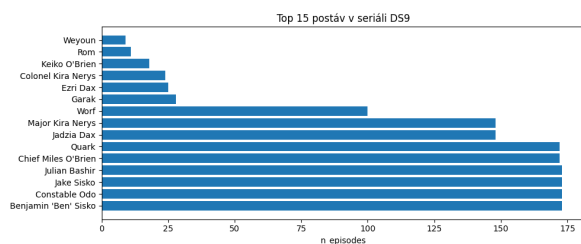
- *The Girl Who Made the Stars* (STT, S2E5) s `num_votes`=712 a `popularity` ≈ -2.25 .

Tento príklad ilustruje, že extrémne hodnoty popularity sa môžu vyskytovať aj pri epizódach s relatívne nízkym počtom hodnotení; v ďalších krokoch preto budeme brať `num_votes` do úvahy pri robustnostných kontrolách a interpretácii výsledkov.

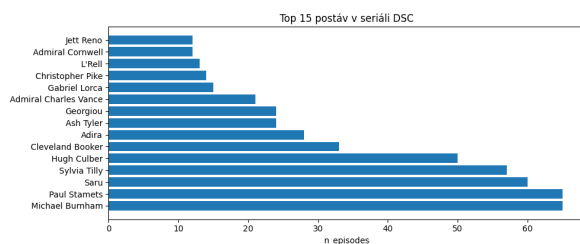
Top 15 postáv v jednotlivých seriáloch – po zobrazení celkového rozdelenia frekvencií postáv (Obr. 6) sa zameriavame na to, ktoré konkrétne postavy tvoria jadro obsadenia v rámci jednotlivých seriálov. Obr. 7 zobrazuje pre každý seriál 15 postáv s najvyšším počtom výskytov v epizódach (`n_episodes`), pričom ide o počty **v rámci daného seriálu** (nie naprieč celou franchise). Tento prehľad je dôležitý pri modelovaní vplyvu postáv: postavy s vysokým `n_episodes` poskytujú veľa pozorovaní a umožňujú stabilnejší odhad efektu, zatiaľ čo zriedkavé postavy budú mať odhady prirodzene neisté.

Z grafov pozorujeme niekoľko výrazných rozdielov medzi seriálmi:

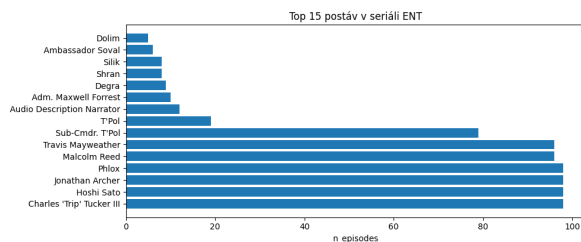
- *dlhé seriály s pevným hlavným obsadením* (DS9, TNG, VOY) majú viacero postáv s veľmi vysokými hodnotami `n_episodes` (typicky ≈ 170), čo zodpovedá tomu, že hlavné postavy sa objavujú takmer v každej epizóde. Zároveň je viditeľný **ostrý zlom** medzi hlavným obsadením a vedľajšími postavami (napr. v DS9 po dominantných postavách nasleduje výraznejší pokles pri postavách ako Garak alebo Worf; vo VOY je podobný zlom viditeľný pri prechode od hlavnej skupiny k postave Kes a ďalej);
- *kratšie moderné seriály* (PIC, SNW, DSC) majú maximálne prirodzene nižšie (desiatky epizód), ale aj tu je viditeľné jadro postáv: v PIC dominuje Jean-Luc Picard spolu s Raffi Musiker a Seven of Nine; v SNW sú najčastejšie postavy členovia posádky (Spock, Una Chin-Riley, Christopher Pike a ďalší);
- *animované seriály* majú odlišnú štruktúru: v LD sú najvyššie štyri hlavné postavy okolo ≈ 50 výskytov a potom nasleduje rýchly pokles; v TAS je jadro tvorené klasickými postavami (Kirk, Spock, McCoy, Sulu, Uhura), pričom počty sú limitované kratšou dĺžkou seriálu (okolo ≈ 20);
- *špecifický prípad STT* – v našich dátach sa pre **STT** objavila iba jedna relevantná postava (Christopher Pike) s `n_episodes` = 3. To naznačuje veľmi nízke pokrytie tohto seriálu v tabuľke `relevant_characters` (napr. chýbajúce záznamy o epizódach, málo postav celkovo), a pri ďalšom modelovaní je vhodné s týmto obmedzením počítat.



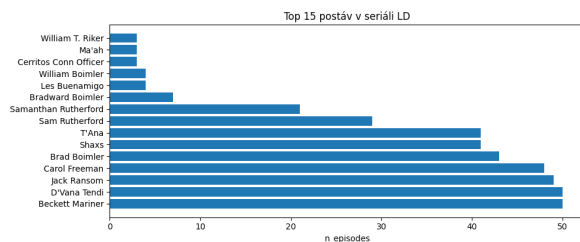
(a) DS9



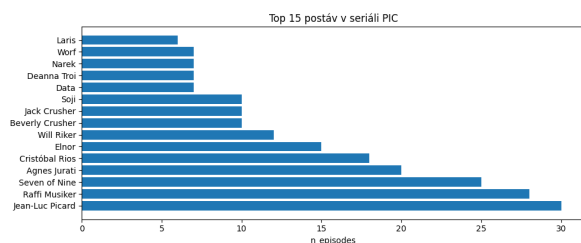
(b) DSC



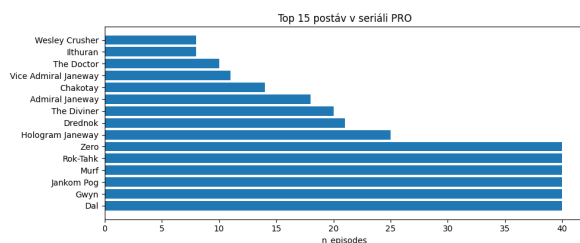
(c) ENT



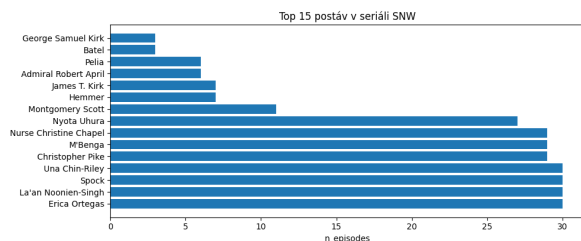
(d) LD



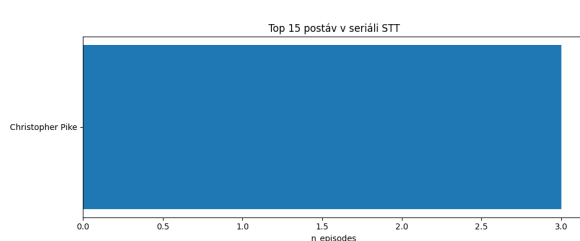
(e) PIC



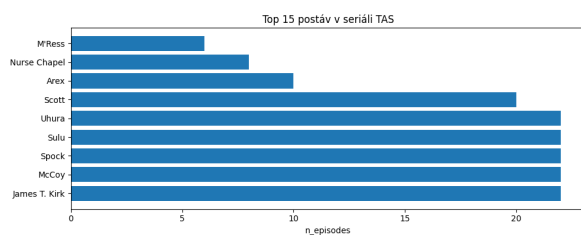
(f) PRO



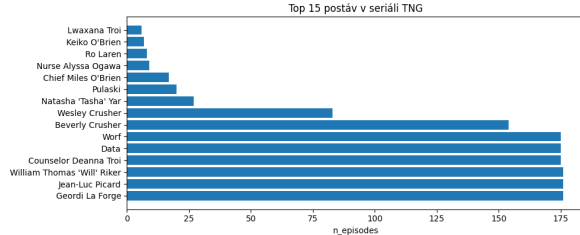
(g) SNW



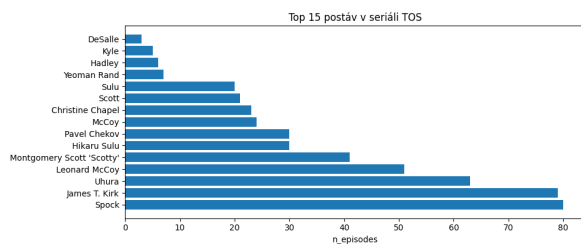
(h) STT



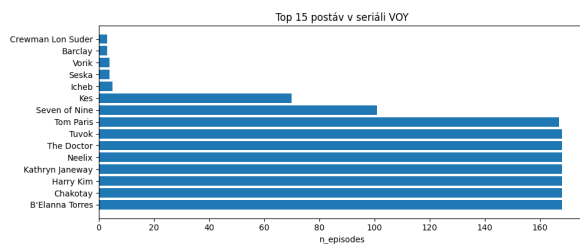
(i) TAS



(j) TNG

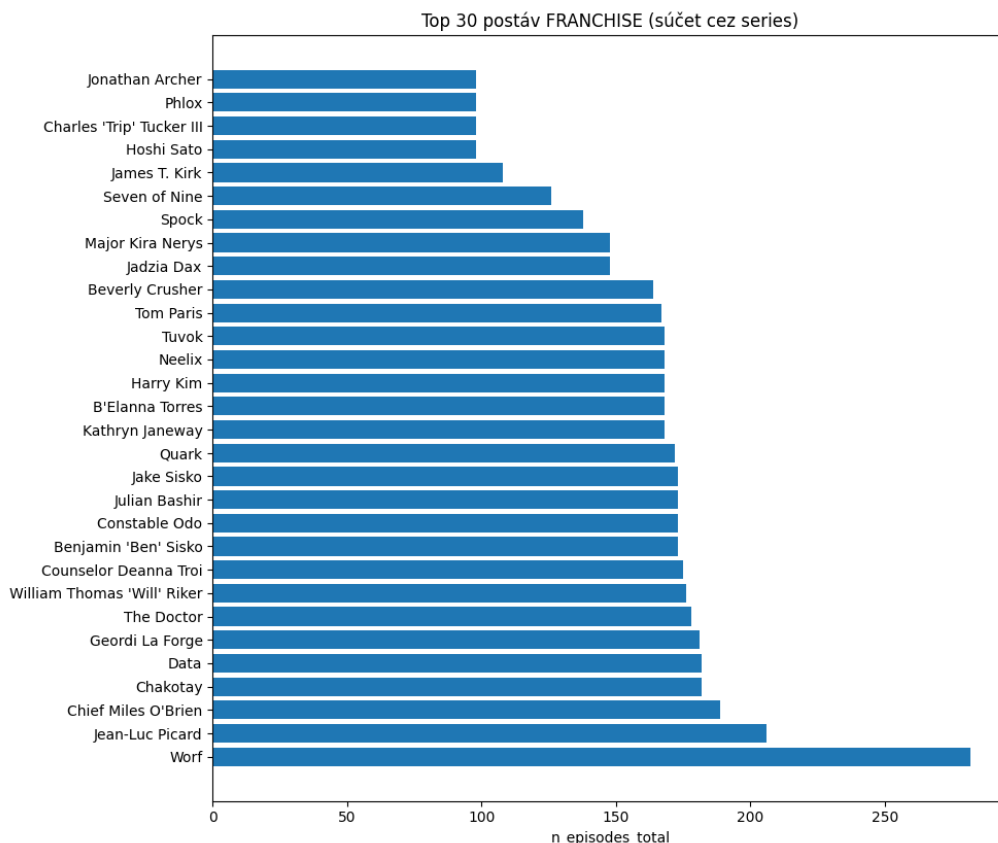


(k) TOS



(l) VOY

Top 30 postáv naprieč celou franchise – okrem pohľadu „v rámci jedného seriálu“ je užitočné pozrieť sa aj na agregovaný pohľad naprieč celým *Star Trek* univerzom. V tomto prípade pre každú postavu sčítame počet výskytov `n_episodes` cez všetky seriály, v ktorých sa objavila, a získame premennú `n_episodes_total`. Obr. 8 zobrazuje 30 postáv s najvyšším celkovým počtom výskytov.



Obr. 8: Top 30 postáv naprieč franchise podľa celkového počtu epizód

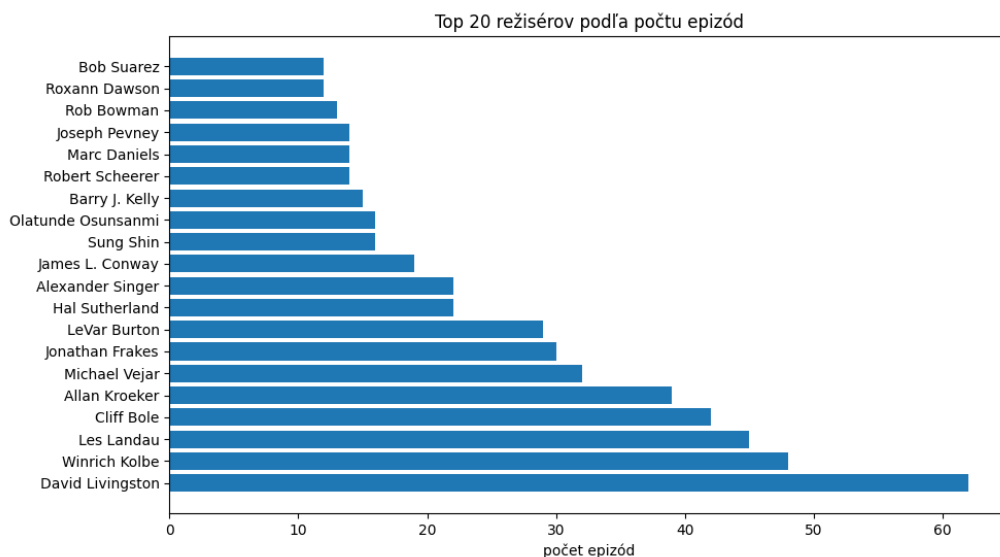
Z grafu vidíme, že:

- absolútne dominuje **Worf** (výrazne najvyšší `n_episodes_total` ≈ 300), čo je konzistentné s tým, že ide o postavu prítomnú vo viacerých seriáloch (najmä TNG a DS9, prípadne aj ďalšie výskyty),
- medzi najvyššie postavy patria aj Jean-Luc Picard (TNG + PIC), Chief Miles O'Brien (TNG + DS9) a ďalšie postavy z dlhých seriálov 90. rokov (napr. Data, Geordi La Forge, Riker, Deanna Troi),
- výrazne sú zastúpené aj postavy z VOY (napr. The Doctor, Janeway, Chakotay, Tuvok, Tom Paris, Harry Kim, B'Elanna Torres), čo odráža vysoký počet epizód tohto seriálu a stabilné obsadenie,
- klasické postavy ako **Spock** a **James T. Kirk** sa nachádzajú v top 30, ale nie na úplnom vrchole, keďže ich výskyty sú rozdelené medzi kratšie seriály (TOS, TAS a prípadné ďalšie moderné výskyty).

Tento „franchise“ pohľad je dôležitý, lebo ukazuje, že celková frekvencia postavy nie je iba funkciou „popularity“ postavy, ale aj kombináciou dĺžky seriálov a toho, či sa postava

objavuje naprieč viacerými seriálmi.

Top 20 režisérov podľa počtu epizód – keďže kvalita a štýl epizód môže byť ovplyvnený aj tvorivým tímom, v EDA analyzujeme aj režisérov. Tabuľka `director` priradzuje epizódam meno režiséra, čo umožňuje zistiť, ktorí režiséri sa v dátach vyskytujú najčastejšie.



Obr. 9: Top 20 režisérov podľa počtu epizód

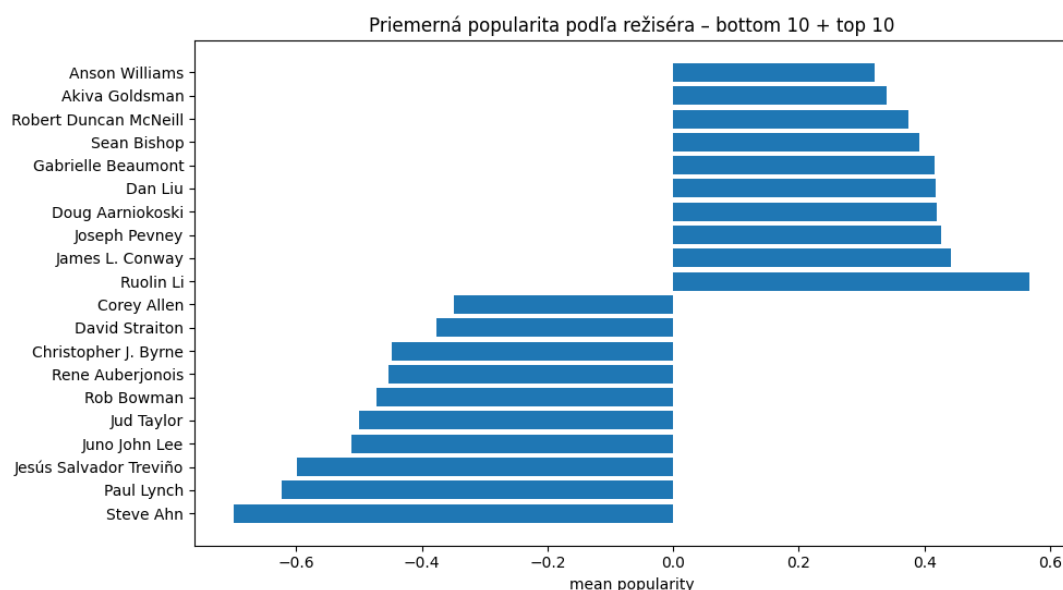
Z grafu vidíme, že:

- najväčší počet epizód v dátach má David Livingston (výrazne viac oproti ostatným), čo naznačuje, že ide o jedného z kľúčových režisérov v rámci dlhších seriálov,
- vysoké počty epizód majú aj Winrich Kolbe, Les Landau a Cliff Bole, ktorí taktiež patria medzi často sa opakujúcich režisérov,
- v rebríčku sa objavujú aj mená hercov, ktorí sa podieľali na réžii epizód (Jonathan Frakes, LeVar Burton),
- rozdiely v počtoch epizód sú výrazné: po niekoľkých najčastejších režiséroch nasleduje skupina s približne porovnateľnými hodnotami, čo znamená, že réžia je rozdelená medzi viacero opakujúcich sa mien, ale s rôznou intenzitou.

Priemerná popularita podľa režiséra – samotná frekvencia režisérov (Obr. 9) ešte nehovorí nič o tom, či epizódy konkrétneho režiséra bývajú hodnotené lepšie alebo horšie. Preto analyzujeme aj priemernú hodnotu `popularity` pre jednotlivých režisérov, pričom do porovnania zahrnieme iba režisérov s dostatočným počtom epizód (v programe `eda.py` filtrované podmienkou $n \geq 5$).

Z grafu vidíme, že:

- rozpätie priemernej popularity medzi režisérmí je približne od -0.7 po $+0.6$, čo znamená, že niektorí režiséri majú epizódy systematicky hodnotené vyššie alebo nižšie vzhľadom na priemer svojich seriálov,



Obr. 10: Priemerná popularita podľa režiséra (s $n \geq 5$ epizód): bottom 10 a top 10

- na negatívnej strane (bottom 10) sú napr. Steve Ahn, Paul Lynch alebo Jesús Salvador Treviño, ktorých epizódy majú v priemere zápornú **popularity** (t. j. skôr pod priemerom seriálu),
- na pozitívnej strane (top 10) sú napr. Ruolin Li, James L. Conway alebo Joseph Pevney, ktorých epizódy majú v priemere kladnú **popularity** (t. j. skôr nad priemerom seriálu).

6.2 Konfundéry (confounders)

V našom projekte chceme odhadnúť, do akej miery je popularita epizódy asociovaná s prítomnosťou konkrétnych postáv. Pri takomto zámere však hrozí, že zdanlivý „vplyv postavy“ bude v skutočnosti spôsobený inou premennou, ktorá súvisí aj s prítomnosťou postavy aj s hodnotením epizódy. Takúto premennú nazývame **konfundér** (confounder). Kontrola konfundérov je kľúčová pre interpretovateľnosť výsledkov: bez nej by sme mohli pripísať postavám efekt, ktorý patrí produkčným faktorom alebo štruktúre seriálu.¹¹

Na základe EDA (grafy popularity v čase, podľa poradia epizódy, a diagnostika `num_votes` a režisérov) identifikujeme najmä tieto skupiny konfundérov:

- **Seriál a časové obdobie (éra)**: jednotlivé seriály vznikali v odlišných obdobiach a s odlišným publikom. Aj keď **popularity** je normalizované z-score v rámci seriálu, časové bloky a rozdiely medzi érami môžu ovplyvňovať rozptyl hodnotení, dostupnosť hlasov a typ epizód.
- **Sezóna a poradie epizódy (season, order_ep)**: premiéry a finále môžu byť systematicky hodnotené inak než „bežné“ epizódy; navyše viaceré postavy sa objavujú typicky až od určitej sezóny alebo iba v časti seriálu.

¹¹Vysvetlenie konfundérov v kontexte dátovej vedy: <https://causalai.causalens.com/resources/blog/confounders-machine-learnings-blindspot/>

- **Počet hlasov (num_votes):** epizódy s nízkym počtom hlasov majú potenciálne vyššiu variabilitu a extrémny; zároveň moderné seriály a špecifické epizódy môžu mať odlišné distribúcie hlasov. Preto `num_votes` (často v log mierke) považujeme za dôležitú kontrolnú premennú.
- **Režisér (director):** EDA ukázala, že režiséri sa líšia nielen frekvenciou, ale aj priemernou popularitou svojich epizód. Ak sa konkrétne postavy vyskytujú častejšie v epizódach určitých režisérov, nekontrolovaný režisérsky efekt by sa mohol zameniť za efekt postáv.

Konfundéry budeme zohľadňovať priamo v modeloch ako kontrolné premenné a zároveň cez baseline porovnania. Baseline bez postáv bude obsahovať iba confundéry (napr. `log(num_votes)`, `season`, `order_ep`, prípadne časové premenné a režiséra). Modely s postavami budeme hodnotiť cez **prírastok** oproti tomuto baseline. Premenné ako **series** môžu byť zahrnuté ako kategórie, aby zachytili systematické rozdiely medzi seriálmi a sezónami. Režisérov s veľmi malým počtom epizód agregujeme do kategórie „other“, alebo režiséra použijeme iba pre skupinu s $n \geq 5$ epizód, aby sme znížili šum z malých vzoriek. Pri citlivých výsledkoch skontrolujeme stabilitu odhadov po filtrovaní epizód s nízkym `num_votes` (napr. pod 5% kvantil) alebo po použití váženia podľa `num_votes`.

Aj pri kontrole confundérov ide primárne o **asociácie**, nie o kauzálne tvrdenia. Výsledky budeme interpretovať ako „epizódy s postavou X majú po kontrole premenných Y typicky vyššiu/nížšiu popularitu“, pričom cieľom je porovnať relatívny vplyv postáv voči produkčným a časovým faktorom.

6.3 Baseline (modely)

Predtým než budeme modelovať „vplyv postáv“ na popularitu epizód, je potrebné definovať baseline prístupy. Baseline predstavuje *minimálnu úroveň výkonu alebo vysvetlenia, ktorú musí každý „model s postavami“ prekonať, aby sme mohli tvrdiť, že informácia o postavách pridáva skutočnú prediktívnu alebo vysvetľovaciu hodnotu*¹². Bez baseline by sme nevedeli rozlíšiť, či pozorované zlepšenie pochádza zo znalosti postáv, alebo iba z jednoduchých štrukturálnych faktorov (napr. rozdiely medzi seriálmi, časové trendy, počet hlasov).

V našom projekte a danom kontexte budeme porovnávať modely podľa toho, ako dobre vysvetlia alebo predikujú cieľovú premennú **popularity** (z-score v rámci seriálu). Baseline zároveň slúži aj ako kontrola confundérov: ak už jednoduchý model bez postáv vysvetľuje významnú časť variability popularity, potom efekt postáv treba interpretovať opatrnejšie.

V práci použijeme viacero úrovní baseline, od najjednoduchších po informatívnejšie:

- **B0: Konštantný model (naivný baseline)** - predikcia je vždy 0 (resp. priemer popularity v tréningovej množine). Keďže **popularity** je štandardizovaná v rámci seriálu, očakávame priemer blízko nule; tento baseline určuje úplné minimum.

¹²Články vysvetľujúce baseline v machine learningu:

<https://towardsdatascience.com/baseline-models-your-guide-for-model-building-1ec3aa244b8d/>
<https://www.sciencedirect.com/topics/computer-science/baseline-model>

- **B1: Model bez postáv so základnými metadátami epizódy** - predikcia popularity využíva iba premenné, ktoré nie sú o postavách, napr. `num_votes` (v log mierke), `season`, `order_ep` a prípadne čas (`air_date`). Tento baseline zachytáva efekt „pozície epizódy“ a kvalitu odhadu hodnotenia pri rôznych počtoch hlasov.
- **B2: Baseline s fixnými efektami seriálu/sezóny** - aj keď popularity už je normalizovaná v rámci seriálu, v dátach stále môžu existovať rozdiely v štruktúre (rôzne éry produkcie, rozdielna distribúcia hlasov, odlišné správanie divákov). Preto používame aj *jednoduchý regresný model* s kategóriami `series`, ktorý poskytne stabilnú referenciu pri porovnaní s modelmi postáv.
- **B3: Režisér ako kontrolná premenná** - režisér môže byť dôležitým konfundérom, preto určíme baseline, ktorý pridá `director` (napr. len pre režisérov s $n \geq 5$ epizód, inak kategória „other“). Tento baseline testuje, akú časť variability popularity vysvetľuje tvorivý tím bez explicitnej informácie o postavách.

Baselinové modely budeme hodnotiť rovnakým spôsobom ako neskoršie modely s postavami. Pre kontinuálnu popularity sú prirodzené metriky napr. MAE/MSE alebo R^2 ; pre robustnejšie porovnanie použijeme aj validáciu, ktorá rešpektuje štruktúru dát (napr. delenie podľa seriálu alebo podľa časových blokov, aby sme minimalizovali „leakage“). Kľúčovým kritériom bude, či modely s postavami prinášajú **dodatočné zlepšenie** oproti baseline pri porovnateľnej komplexite.

Baseline teda tvorí **referenčnú os** celej analýzy: ak model s postavami neprekoná baseline s metadátami (a konfundérmi), potom je tvrdenie o „vplyve postáv“ slabé alebo neodôvodnené.

6.4 Modely s postavami (od jednoduchých po silné)

Po definovaní baseline a identifikovaní konfundérov prechádzame k hlavnému cieľu projektu: navrhnuť modely, ktoré explicitne využívajú informáciu o postavách a umožnia kvantifikovať, či prítomnosť konkrétnych postáv súvisí s vyššou alebo nižšou popularitou epizód. Postavy budeme reprezentovať ako vysvetľujúce premenné (features) a budeme sledovať, o koľko sa zlepší vysvetlenie/predikcia popularity oproti baseline (t. j. oproti modelom bez postáv).

Základnou reprezentáciou bude binárna indikácia prítomnosti postavy v epizóde (napr. 1 = postava csa v epizóde vyskytuje). V praxi tak pre každú epizódu vytvoríme vektor príznakov pre vybranú množinu „relevantných“ postáv. Okrem samotnej prítomnosti však vieme využiť aj metadátovú informáciu o type postavy, konkrétne atribút `is_main_cast`. Ten umožňuje rozlíšiť, či ide o „hlavnú“ postavu seriálu alebo skôr hosťujúcu/vedľajšiu postavu.

Preto budeme reprezentáciu postáv rozširovať o jednoduché agregované príznaky odvodené z `is_main_cast`, napríklad:

- počet prítomných hlavných postáv v epizóde (`#main`),
- počet prítomných nehlavných postáv (`#non-main`),

- podiel hlavných postáv medzi všetkými prítomnými postavami,

Aj pri tejto rozšírenej reprezentácii zostáva kľúčovým problémom tzv. **long-tail** správanie: veľká časť vedľajších postáv sa objaví iba v malom počte epizód. Preto budeme pracovať buď s prahom minimálnej frekvencie (napr. zahrnúť len postavy s $n_episodes \geq k$), alebo s regularizáciou (L1/L2/Elastic Net), ktorá stabilizuje odhady pre zriedkavé postavy a zároveň znižuje riziko preučenia.

Použijeme viacero modelov s rastúcou komplexitou, aby bolo jasné, čo presne pridáva informácia o postavách:

- **M1: Jednopostavový efekt (univariate)** - pre každú postavu samostatne porovnáme epizódy „s postavou“ vs. „bez postavy“ a otestujeme rozdiel v priemernej **popularity** (napr. rozdiel priemerov alebo jednoduchá lineárna regresia s jednou binárnou premennou + konfundéry). Tento krok slúži ako rýchly screening a poskytne intuitívne interpretovateľné výsledky.
- **M2: Viacpostavový lineárny model + konfundéry** - zahrnieme viac postáv naraz do jedného regresného modelu:

$$\text{popularity} \sim (\text{konfundéry}) + \sum_{c \in \mathcal{C}} \beta_c \cdot 1[c]$$

Tento model umožňuje odhadnúť „parciálny efekt“ postavy pri kontrole ostatných postáv a konfundérov. Keďže počet postáv môže byť vysoký, použijeme *regularizáciu*.

- **M3: Regularizované modely (LASSO / Ridge / Elastic Net)** - regularizácia rieši dva problémy: veľký počet postáv (*feature dimension*) a *kolinearitu* (postavy sa často vyskytujú spolu). LASSO (L1) navyše vykonáva výber premenných a poskytuje riedke riešenie (len malý počet postáv s nenulovým koeficientom), Ridge (L2) je stabilnejší pri silnej korelácii príznakov a Elastic Net kombinuje oba efekty.
- **M4: Modely s interakciami alebo zoskupeniami** - môžeme testovať aj interakcie (napr. dvojice postáv, ktoré sa často vyskytujú spolu) alebo zoskupenie postáv do tematických blokov (posádka, antagonisti, hostia). Tento krok by zachytil skupinové efekty, ktoré jednoduchý lineárny súčet koeficientov nevidí.
- **M5: Nelineárne modely** - ako doplnok môžeme porovnať výsledky s jednoduchými stromovými metódami (napr. Random Forest / Gradient Boosting), ktoré dokážu zachytiť nelinearity a interakcie automaticky. V tomto projekte ich však budeme používať skôr ako *robustnostnú kontrolu interpretácie*, keďže interpretácia efektov postáv je v stromových modeloch menej priamočiara.

Modely budeme porovnávať voči baseline a medzi sebou:

- **prírastok vysvetlenej variability** (napr. R^2) alebo zlepšenie chyby (MAE/MSE) oproti baseline,
- **stabilita efektov postáv** pri rôznych špecifikáciách (napr. s/bez režiséra, filtrovanie nízkych hlasov),

- **interpretácia koeficientov** (najmä pri lineárnych/regularizovaných modeloch): znamienko a veľkosť β_c interpretujeme ako asociáciu prítomnosti postavy s odchýlkou popularity od priemeru seriálu, po kontrole konfundérov.

Takto postupujeme od jednoduchých, ľahko interpretovateľných testov až po silnejšie modely, ktoré dokážu naraz pracovať s veľkým počtom postáv a kontrolovať konfundéry. Cieľom nie je len dosiahnuť čo najlepší prediktívny výkon, ale najmä získať *interpretovateľný odhad*, či a ktoré postavy sú spojené s vyššou alebo nižšou popularitou epizód.

7 Implementácia modelov

V tejto časti popisujeme implementáciu a vyhodnotenie modelov, ktoré majú kvantifikovať vzťah medzi prítomnosťou konkrétnych postáv v epizóde a jej popularitou (definovanou v predchádzajúcej časti). Cieľom implementácie je zostrojiť dizajnové matice pre rôzne množiny vysvetľujúcich premenných, aplikovať rovnaký validačný postup a umožniť férové porovnanie viacerých modelových skupín.

Modely budujeme postupne od jednoduchých baseline prístupov až po regularizované regresné modely s vysokodimenziálnymi vstupmi (napr. binárne indikátory postáv). Dôraz kladieme na to, aby všetky varianty používali identické cieľové premenné, rovnaké rozdelenie dát v rámci validácie a rovnaké metriky hodnotenia (napr. RMSE, MAE, R^2). Takto vieme interpretovať rozdiely vo výkonnosti primárne ako dôsledok pridania/odoberania konkrétnych typov informácie (metadáta epizódy, postavy, prípadne ďalšie kontextové premenné).

7.1 Konštrukcia programov

V adresári `modely` používame podadresár `setup`, ktorý združuje spoločnú infraštruktúru zdieľanú všetkými modelmi: jednotné nastavenia ciest k dátam a stĺpcov (`config.py`) a načítanie a zjednotenie tabuliek (CSV s fallbackom na SQLite), tvorbu dizajnovej matice X , cieľa y a skupín pre validáciu (`common.py`).

Hlavným spúšťacím skriptom experimentov je `run_models.py`, ktorý postupne vykoná tréning a vyhodnotenie všetkých definovaných modelových variantov a agreguje metriky do spoločného výstupu.

Baseline modely:

- B0: konštantný baseline (`DummyRegressor`)
- B1: lineárna regresia na metadátach epizódy
- B2: ridge regresia na metadátach ($\alpha = 1$)
- B3: ridge na metadátach s výberom α cez jednoduchý grid-search

Modely s postavami:

- M1: ridge na metadátach + indikátoroch relevantných postáv ($\alpha = 1$)
- M2: ridge na metadátach + postavách s výberom α
- M3: ridge na metadátach + postavách + režisérovi ($\alpha = 1$)
- M4: ridge na metadátach + postavách + režisérovi s výberom α
- M5: lasso (L1) so štandardizáciou vstupov a výberom α

Program `run_models.py` pre každú sériu načíta dáta, zostrojí príslušné dizajnové matice podľa konfigurácie modelu a vykoná jednotné vyhodnotenie modelov. Kľúčovou časťou je funkcia `run_logo_cv`, ktorá realizuje skupinovú krížovú validáciu *Leave-One-Group-Out*¹³ (s `group_col = season`), t.j. v každom kroku necháva jednu sezónu ako testovaciu a trénuje na zvyšných sezónach; ak séria obsahuje iba jednu sezónu, použije sa fallback na `KFold`. Pre každý model sa ukladajú OOF predikcie (out-of-fold) a počítajú sa metriky RMSE, MAE a R^2 . Prejdenie zo „skupina = seriál“ na „skupina = séria“ bolo realizované kvôli lepšej presnosti výsledkov a odstráneniu šumu.

Kód ďalej implementuje jednoduchý výber hyperparametra α pre ridge a lasso (`pick_best_alpha_ri` a `pick_best_alpha_lasso`) na pevne zvolenej mriežke hodnôt. Okrem agregovaných metrických výsledkov generuje aj diagnostické výstupy: grafy „predikcia vs. skutočnosť“ a histogram rezíduí pre každý model, porovnávací graf RMSE v rámci série a (pre modely s postavami) vizualizácie najvplyvnejších postáv na základe absolútnej hodnoty koeficientov po natrénovaní modelu na celých dátach danej série; tieto výsledky sa ukladajú do `output_models/<SERIES>/` spolu s CSV exportom top koeficientov.

7.2 Vyhodnotenie modelov

Pomocou skriptu `analyze_models.py` pre každú sériu vyberieme najlepšie model podľa troch metrík: RMSE (minimum), MAE (minimum) a R^2 (maximum). Výsledky ukazujú, že prínos informácie o postavách (modely typu M*) je výrazne závislý od konkrétnej série.

- série, kde vyhrali modely s postavami (M*):
 - *PRO* (najlepší M4_ridge_meta+chars+dir_best_a10),
 - *PIC* (najlepší M5_lasso_meta+chars+dir_best_a0.001),
 - *DS9* (najlepší M5_lasso_meta+chars+dir_best_a0.1),
 - *ENT* (najlepší podľa RMSE a R^2 je M2_ridge_meta+chars_best_a10; podľa MAE vyhráva B3_ridge_meta_best_a10)
- série, kde vyhrali baseline modely (B*):
 - *DSC* (najlepší B1_linear_meta),
 - *LD* (najlepší B3_ridge_meta_best_a100)

¹³Overview: <https://machinelearningmastery.com/loocv-for-evaluating-machine-learning-algorithms/>

- série, kde bol najlepší triviálny baseline `B0_dummy0`:

- *STT*,
- *TAS*,
- *SNW*,
- *TOS*,
- *TNG*,
- *VOY*

V týchto prípadoch modely s pridanými vysvetľujúcimi premennými nedokázali prekonať predikciu konštantou (priemerom), čo sa prejavuje aj $R^2 \approx 0$ (resp. numericky blízko nule).

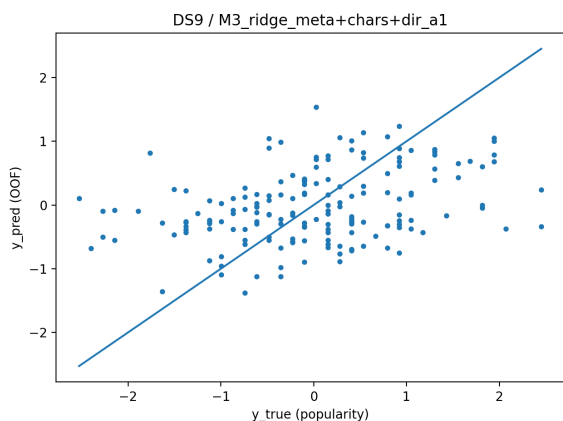
Z hľadiska interpretácie možno výsledky čítať tak, že v niektorých sériách majú epizódové metadáta a prítomnosť relevantných postáv (a prípadne režiséra) merateľnú predikčnú hodnotu, zatiaľ čo v iných sériách je variabilita hodnotení buď nízka, alebo je vysvetlená faktormi, ktoré v použitých premenných nie sú zachytené.

7.3 Grafy modelov

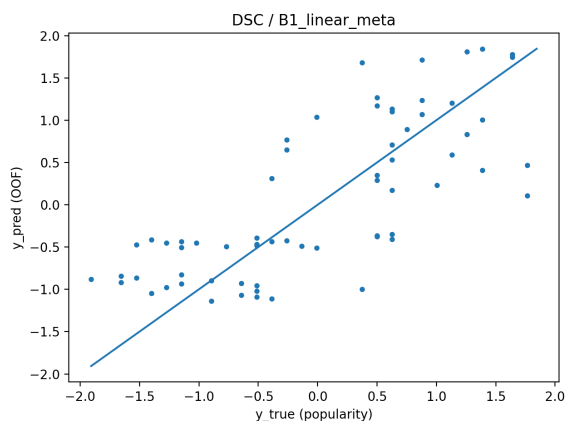
V rámci vyhodnotenia sme pre každú sériu a každý model generovali niekoľko typov grafov, ktoré slúžia na porovnanie kvality predikcie, diagnostiku chýb a interpretáciu vplyvu postáv. Keďže počet výstupov rýchlo rastie (viac modelov \times viac sérií), v reporte uvádzame iba vybrané, najinformatívnejšie príklady; kompletná sada grafov je uložená v adresári `output_models/<SERIES>/`.

Konkrétne sme používali tieto typy grafov:

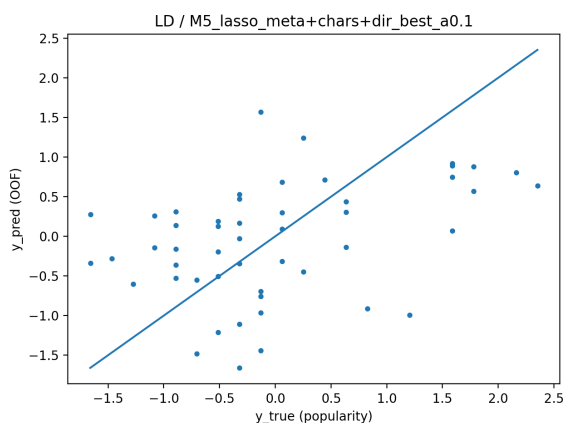
1. **Predikcia vs. skutočnosť (OOF)** - Obr. 11: bodový graf \hat{y} proti y s referenčnou diagonálou $y = \hat{y}$. Používame OOF predikcie (out-of-fold), teda predikcie na epizódach, ktoré neboli v danom folde použité na tréning. Graf umožňuje vizuálne posúdiť, či model zachytáva trend (body približne pri diagonále) alebo sa predikcie „zlievajú“ do úzkeho pásma okolo nuly.
2. **Histogram rezíduí** - Obr. 12: rozdelenie $e = y - \hat{y}$, ktoré ukazuje typickú veľkosť chyby, prípadnú asymetriu a outliery. Pri rozumnom modeli očakávame rezíduá približne centrované okolo nuly bez výrazného posunu.
3. **Najvplyvnejšie postavy** - Obr. 13: stĺpcový graf koeficientov pre vybrané postavy (napr. top- k podľa $|\beta|$) po natrénovaní modelu na celej sérii. Znamienko koeficientu interpretuje smer asociácie (pozitívny/negatívny vplyv na popularitu v rámci štandardizovanej cieľovej premennej), veľkosť koeficientu jeho silu v rámci daného modelu. Pre každú sériu sme vybrali *jeden reprezentatívny* model typu M^* (ten s najnižším RMSE v danej sérii) a zobrazili sme top- k postáv podľa $|\beta|$.



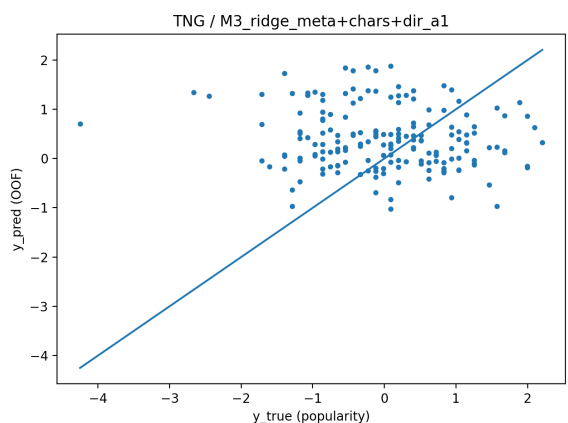
(a) **DS9 / M3 (ridge + meta + postavy + režisér)**. Vidno iba mierny trend; predikcie sú čiastočne stlačené smerom k nule, čo naznačuje obmedzenú vysvetliteľnosť variability popularity.



(b) **DSC / B1 (lineárny model na metadátach)**. Body sa viac približujú diagonále, model lepšie rozlišuje epizódy s nižšou a vyššou popularitou.

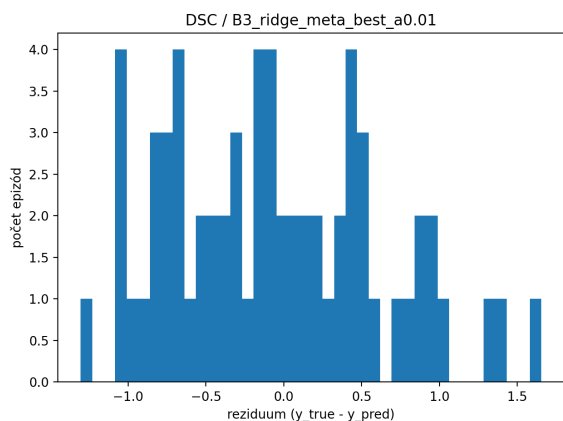


(c) **LD / M5 (lasso + meta + postavy + režisér)**. Predikcie majú tendenciu byť konzervatívne (stlačené okolo stredu), čo je typické pri silnejšej regulárizácii a môže viesť k podpredikovaniu extrémov.

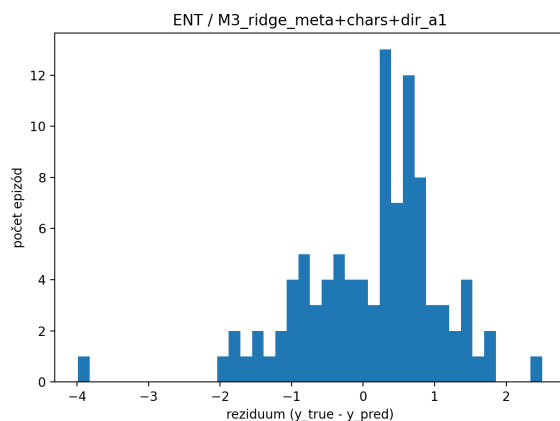


(d) **TNG / M3 (ridge + meta + postavy + režisér)**. Silnejší rozptyl okolo diagonály a koncentrácia predikcií naznačujú slabý signál; je to konzistentné s tým, že pre TNG vychádza ako najlepší triviálny baseline.

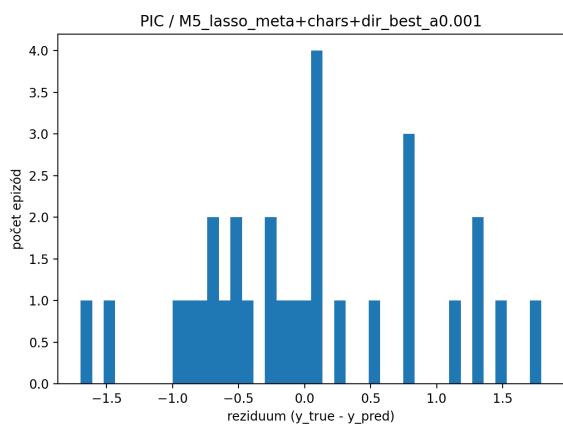
Obr. 11: Reprezentatívne príklady grafov „predikcia vs. skutočnosť“ (OOF). Diagonála $y = \hat{y}$ slúži ako referencia ideálnej predikcie; odchýlky a „stláčanie“ predikcií okolo nuly typicky indikujú podfitovanie alebo nízku vysvetliteľnosť cieľa použitými premennými.



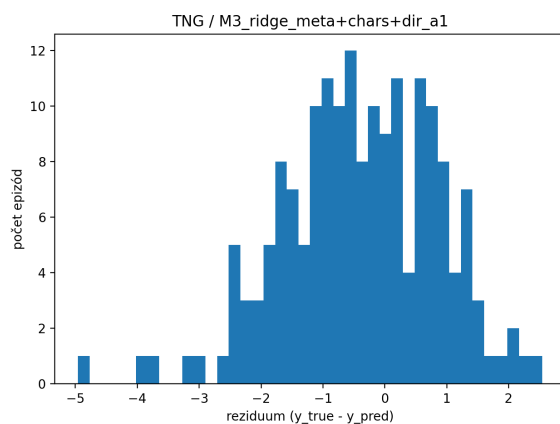
(a) **DSC / B3 (ridge na metadátach).** Rezíduá sú relatívne koncentrované a bez extrémne dlhých chvostov; model robí väčšinu chýb v primeranom rozsahu.



(b) **ENT / M3 (ridge + meta + postavy + režisér).** Viditeľné sú aj väčšie odchýlky a chvosty, čo naznačuje existenciu epizód, ktoré model nedokáže dobre vysvetliť.

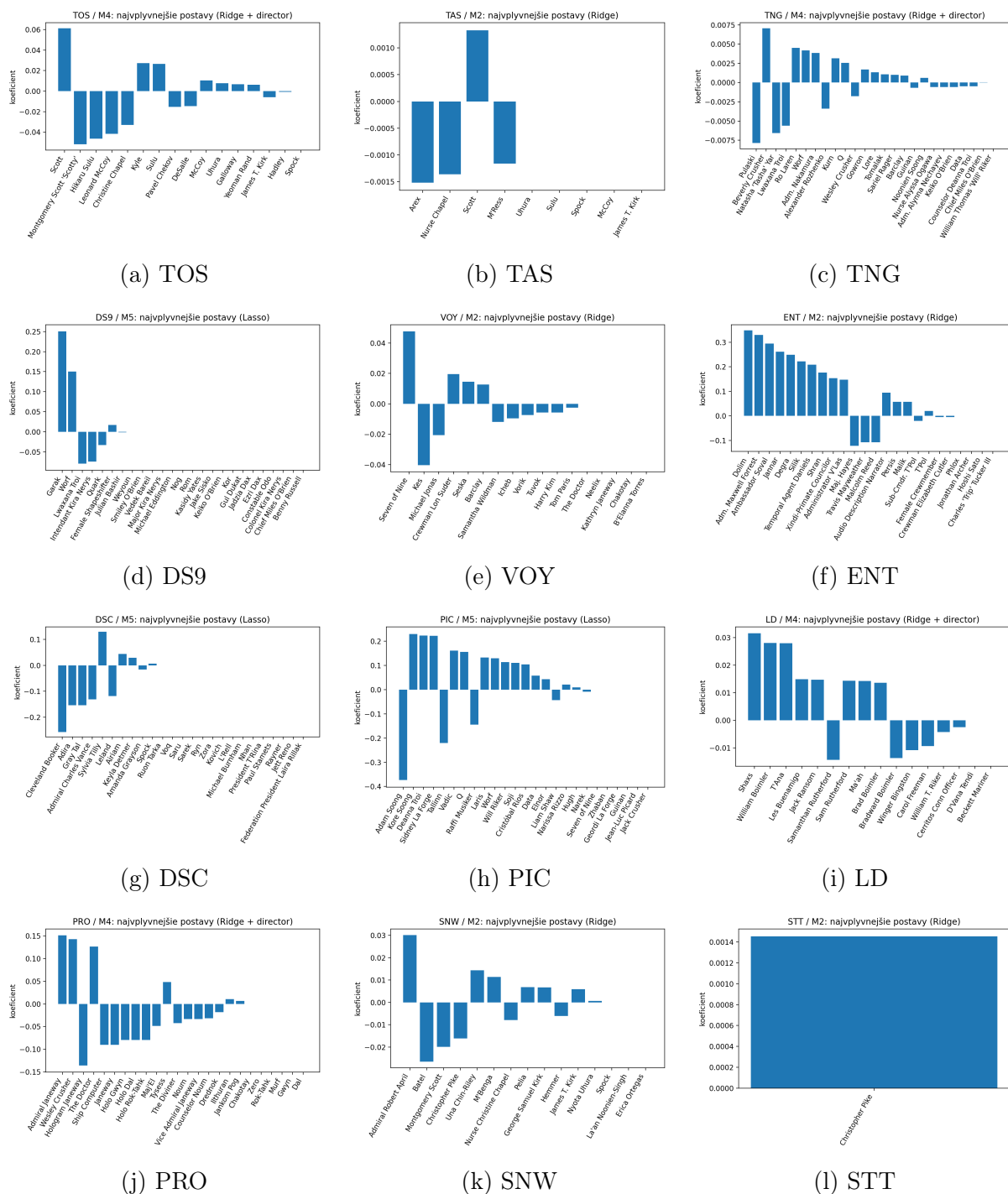


(c) **PIC / M5 (lasso + meta + postavy + režisér).** Pri lasso môže byť rozdelenie rezíduí ovplyvnené silnejšou regulárizáciou; typicky sa prejaví konzervatívnejšia predikcia a horšie zachytávanie extrémov.



(d) **TNG / M3 (ridge + meta + postavy + režisér).** Široké rezíduá a dlhé chvosty sú konzistentné so slabším prediktívnym signálom; v tejto sérii často vychádza ako najlepší triviálny baseline.

Obr. 12: Reprezentatívne histogramy rezíduí $e = y - \hat{y}$ (OOF). Šírka a tvar rozdelenia pomáhajú posúdiť stabilitu modelu, prítomnosť biasu (posun od nuly) a výskyt epizód s výraznou chybou (chvosty/outliery).



Obr. 13: Najvplyvnejšie postavy pre každú sériu: vždy výsledok z modelu s najnižším RMSE.

8 Interpretácia výsledkov

Táto časť sumarizuje, čo vieme vyčítať z tabuľky najlepších modelov pre každú sériu (RMSE/MAE/R²), diagnostických grafov „predikcia vs. skutočnosť“ a histogramov rezíduí, a grafov „najvplyvnejších postáv“ (Obr. 13).

Čo znamená, že niekde vyhral jednoduchý baseline (**B0/B1**)? Ak v niektorej sérii vyjde ako najlepší veľmi jednoduchý model (napr. B0 alebo lineárny model len s meta-

dátami), typicky to znamená, že v dátach je slabý alebo nestabilný signál pre jemnejšie vysvetlenie popularity. Prakticky ide o kombináciu týchto efektov:

- málo epizód / málo variability cieľa - pri malých sériách alebo pri cieľovej premennej s nízkym rozptylom model prirodzene končí pri „predikcii priemeru“ (v našom prípade okolo 0, keďže popularita je z-score v rámci série).
- viac šumu než signálu - epizódové hodnotenia môžu mať vyšší šum (napr. vplyv doby, nostalgia, sampling bias), takže detailné premenné (postavy, režisér) nepridajú konzistentnú informáciu naprieč foldmi.
- nedostatok vysvetľujúcich premenných - ak je prítomnosť postáv alebo režisérske dáta neúplná, bohatší model má viac stupňov voľnosti, no nemá čo stabilne „naučiť“ a teda výkon sa nezlepší, niekedy sa zhorší.

V sériách, kde vyhrávajú modely s postavami (**M2/M3**) alebo s režisérom (**M4/M5**), je signál spravidla stabilnejší: viac epizód, väčšia diverzita obsadenia alebo opakujúci sa štýl tvorcov. Režisér vie fungovať ako proxy pre „štýl“ (dynamika epizódy, tón, typ príbehu), postavy zas zachytia, že niektoré kombinácie/úseky seriálu bývajú systematicky hodnotené vyššie či nižšie.

Pri interpretácii používame najmä R^2 :

- $R^2 > 0$ znamená, že model je lepší než triviálne predikovať priemer (v našom prípade približne 0). To naznačuje, že v dátach existuje *opakovateľný* vzor.
- $R^2 \approx 0$ znamená veľmi slabú predikovateľnosť (predikcie sa zlievajú k priemeru).
- $R^2 < 0$ znamená, že model je horší než predikcia priemeru; často ide o šum alebo o preučenie.

V našich výsledkoch (tabuľka s CV metrikami `series_cv_results.csv`) sa série delia na dve skupiny:

1. Série s **reálnou predikovateľnosťou** ($R^2 > 0$) – modely dokážu aspoň čiastočne zachytiť opakovateľný vzor:
 - DSC: $R^2 \approx 0.54$ (najlepšie), RMSE ≈ 0.68
 - LD: $R^2 \approx 0.27$, RMSE ≈ 0.84
 - PRO: $R^2 \approx 0.25$, RMSE ≈ 0.85
 - PIC: $R^2 \approx 0.24$, RMSE ≈ 0.86
 - DS9: $R^2 \approx 0.16$, RMSE ≈ 0.91
 - ENT: $R^2 \approx 0.07$, RMSE ≈ 0.96
2. Série **bez jasného signálu** ($R^2 \approx 0$ alebo mierne < 0) – najlepší je v podstate baseline (B0), t. j. predikcie sa zlievajú k nule:
 - TNG, VOY, TOS, TAS, STT: $R^2 = 0$ (baseline vyhral)
 - SNW: $R^2 \approx -2.2 \cdot 10^{-16}$

Bodový graf \hat{y} proti y s diagonálou $y = \hat{y}$ vizuálne ukazuje, či model zachytáva trend. Typické scenáre:

- **body približne pozdĺž diagonály:** model zachytáva monotónny vzťah (aspoň hrubý trend)
- **zlievanie do pásma okolo nuly:** model regresuje k priemeru, nevie rozlíšiť epizódy s vyššou/nížšou popularitou - toto je typické pri slabom signáli alebo pri veľmi silnej regularizácii
- **stlačené extrémny:** aj keď je trend prítomný, predikcie majú menší rozsah než skutočné y (podhodnotený maximá a nadhodnotený minimá) - časté pri ridge/lasso a šume v cieľovej premennej

Rezíduum definujem ako $r = y_{\text{true}} - y_{\text{pred}}$. Histogramy pomáhajú odhaliť systematické chyby:

- **symetria okolo 0:** bez výrazného biasu (model v priemere nepreceňuje ani nepodceňuje)
- **posun doprava (kladné rezíduá):** model skôr podhodnocuje (true je vyššie než pred)
- **posun doľava (záporné rezíduá):** model skôr nadhodnocuje
- **ťažké chvosty / odľahlé hodnoty:** epizódy, ktoré model typicky „nevie“ (špeciály, piloty, finále alebo epizódy s atypickým obsadením)

Obr. 13 zobrazuje pre každú sériu stĺpcový graf koeficientov vybraných postáv (top- k podľa $|\beta|$) z **najlepšieho modelu** (najnižšie RMSE). Zmysel týchto koeficientov je:

- **Asociácia v rámci série:** keďže cieľ y je štandardizovaný (z-score popularity v rámci série), koeficienty vyjadrujú, s akou (lineárnou) zmenou očakávanej popularity je spojená vyššia hodnota príslušnej premennej.
- **Znamienko koeficientu:** pozitívne β znamená, že vyššia prítomnosť/aktivita postavy (pri ostatných premenných fixných) je spojená s vyššou očakávanou popularitou; negatívne β analogicky s nižšou.
- **Veľkosť koeficientu:** väčšie $|\beta|$ znamená silnejšiu asociáciu *v rámci daného modelu a danej série*.

Zároveň je dôležité jasne uviesť, čo z toho **nevyplýva**: nie je to kauzalita - koeficient nehovorí, že postava „spôsobuje“ vyššie hodnotenia; len že sa s nimi štatisticky spája v daných dátach. Koeficienty sú modelovo závislé - teda ridge rozdeľuje vplyv medzi korelované premenné, lasso môže vybrať jednu z nich a ostatné potlačiť na nulu. Ak postavy často vystupujú spolu, model nevie jednoznačne priradiť „zásluhu“ jednej postave; vplyv sa rozleje alebo sa náhodne priradí jednej (najmä pri lasso). Škálu koeficientov nemožno porovnávať medzi sériami (rôzne série majú inú variabilitu, iné rozdelenia vstupov a často aj iné nastavenia regularizácie; porovnávame najmä tvar, nie absolútne hodnoty).

Z grafov na Obr. 13 je vidieť typický kontrast: **Ridge (M2/M4)** dáva „husté“ riešenie: veľa nenulových koeficientov a často postupný „chvost“ smerom k nule (napr. TNG

alebo ENT majú viacero menších koeficientov, ktoré spolu skladajú predikciu). **Lasso (M5)** dáva „riedke“ riešenie: často dominuje niekoľko stĺpcov a veľa postáv má koeficient presne 0 (napr. DS9/DSC/PIC majú viditeľne „selektívnejší“ profil).

8.1 Kontext seriálu

Teraz poďme porovnať naše výsledky s kontextom seriálov, teda či najvplyvnejšie postavy mali ozačstný vplyv na hodnotenia. Pre každý seriál zhrnieme kladné a záporné postavy (t.j. postavy s kladným / záporným vplyvom na hodnotenia) a potom z fanúšikovských zdrojov potvrdíme, resp. vyvrátíme tieto tvrdenia. Zo zdrojov uvedených v kapitole 2.4 Doplnujúce zdroje pre získanie dodatočných informácií budeme filtrovať meno postavy pre nájdenie podkladu (ne)obľúbenosti postavy a uvedieme relevantné citáty:

- DS9

- kladné postavy:

- * Garak – vysoko obľúbená postava s viacerými epizódami, kde účinkuje v hlavnej roli, napr. v epizóde centrovanej na Garaka „The Wire“ (2. sezóna, 22. epizóda):

- “This is the only episode of the season meriting 4 stars. Performance, mood, character and story are in perfect balance here, and Garak proves his worth to the cast beyond any doubt—his character is more interesting, nuanced, sophisticated and elegantly portrayed than any of the main cast except maybe Odo. Bravo, Mr Robinson.”*

- Alebo opis postavy Garaka ako:

- “Quite possibly one of the best Star Trek characters ever conceived, Garak is played to perfection by Andrew Robinson as an anti-hero who can carry the main storyline or fall back to supporting role with complete ease.”*

- * Worf – obľúbená postava, často párovaná s už spomenutým Garakom, čiže môžeme predpokladať, že tu hrá rolu skôr skupinovú obľúbenosť – ako spomína tento používateľ vo svojom hodnotení (epizóda „In Purgatory’s Shadow“, 5. sezóna, 14. epizóda):

- “And honestly, is there any character on this series Garak does *not* interact well with? Worf and Garak in the runabout = classic.”*

- záporné postavy:

- * Quark – z dostupných zdrojov vyplýva, že ide skôr o kladnú postavu, no vyskytuje sa aj v slabo exekúovaných epizódach, ktoré fanúšikov až tak neupútali – napr. pre epizódu „Prophet Motive“ (3. sezóna, 16. epizóda):

- “Here, Quark is cardboardedly transparent.”*

- DSC

- kladné postavy:

- * Sylvia Tilly – fanúšikmi opísaná ako:

“The best cadet-now-ensign since Nog...”

– záporné postavy:

- * Adira – objavuje sa v epizóde „Forget Me Not“ (3. sezóna, 4. epizóda). Fanúšikovská kontroverzia sa točí okolo hereckého obsadenia, keďže do role Adiry obsadili nebinárnu osobu, no niekoľkokrát v seriáli referujú na postavu ako „ona“.
- * Leland – jeden opis inej postavy spomína Lelanda ako:
“Just as annoying, but probably more capable than Leland T. Lynch...”

• ENT

– kladné postavy:

- * Maxwell Forrest – vedľajšia postava pri kapitánovi, obľúbená kvôli jeho dobročinnosti, ako je opísané v epizóde „In a Mirror, Darkly, Part I“ (4. sezóna, 18. epizóda):
“Captain Forrest, at the very least, demonstrated a spark of altruism and genuinely honorable behaviour during the evacuation...”
- * Degra – vedľajšia postava vyskytujúca sa v epizóde „The Forgotten“ (3. sezóna, 20. epizóda):
“Degra is very well rounded character and played really subtle.”

– záporné postavy:

- * Travis Mayweather – nedomyslená postava, ktorú by sledovatelia radi videli viac; bohužiaľ mu autori neprípísali mnoho akcie či dialógu, ako spomína jedna recenzia týkajúca sa epizódy „Horizon“ (2. sezóna, 20. epizóda):
“Travis doesn’t annoy me the same way because he’s scarcely given the chance to grow or to not grow; the writers have no idea who this guy is because they refuse to give him anything to do or any semblance of a personality. He’s an empty shell of a character usually used as a tool of the plot.”

• LD

– kladné postavy:

- * T’Ana – menšia postava, no komická a zábavná pre veľa sledovateľov – recenzia o epizóde „We’ll Always Have Tom Paris“ (2. sezóna, 3. epizóda):
“Another delightfully silly and thoroughly enjoyable episode, with some actual laugh-out-loud moments (most notably T’Ana when she finally got her wooden box).”
- * Brad Boimler – postava, ktorá hrala celkom podstatné roly v niekoľkých epizódach (napr. epizóda „Much Ado About Boimler“), kde fanúšikovia majú skôr problém s interakciami iných postáv než s touto postavou. Preto sú reakcie zmiešané a môže sa vyskytnúť v grafoch ako kladná alebo záporná.

– záporné postavy:

- * Samantha Rutherford – taktiež komická postava, no bohužiaľ mala niektoré slabé epizódy („Strange Energies“, 2. sezóna, 1. epizóda):

“...but the Tendi/Rutherford subplot got tiring fast, which is a shame as it started out very promisingly.”

- PIC

- kladné postavy:

- * Adam a Kora Soong – otec a dcéra rodiny Soong, ktorá je dosť podstatná v tomto seriáli; fanúšikovia sú rozdelení názorovo, ale celkovo zaujati ich dynamikou.

- * Deanna Troi – jedna z opakujúcich sa postáv, ktorá sa objavila v tomto seriáli napr. v epizóde „Nepenthe“ (1. sezóna, 7. epizóda):

- “I especially liked how well they wrote Troi’s in this episode—so much better than how they used her in the series.”*

- záporné postavy:

- * Tallinn – menšia rola, nedotiahnutá dokonca:

- “Tallinn still isn’t doing it for me. They dropped another few hints that she’s actually a Romulan, but...she doesn’t seem like a person to me still, just a plot conveyance.”*

- PRO

- kladné postavy:

- * Admiral Janeway – jedna z hlavných postáv seriálu Voyager, ktorej prítomnosť v tomto seriáli je fanúšikmi uvítaná a herecké stvárnenie ocenené, napr. v epizóde „Mindwalk“ (1. sezóna, 18. epizóda):

- “I found the highlight of the episode to be Vice Admiral Janeway. Her portrayal in the episode seemed like a natural extension of what was previously offered in Voyager, and it was heartening to see the crew finally gain an ally within the Federation.”*

- * The Doctor – medicínsky hologram, ktorý pomáhal posádke a rozvinul sa do charakteru obľúbeného mnohými.

- záporné postavy:

- * Hologram Janeway – hologram reálnej postavy admirálky Janeway je zápornou postavou v seriáli, pričom mnoho fanúšikov je polarizovaných ohľadom jej charakteru (niektorí ju ako negatívnu postavu majú radi, iní diskutujú o téme hologramov v rámci Star Treku).

- SNW

- kladné postavy:

- * Admiral Robert April – stabilná postava admirála naprieč seriálom; naprieč epizódami sa vyskytuje stabilný pozitívny sentiment, dobre zhrnutý komentárom:

- “Nice to see April again.”*

- záporné postavy:

- * Batel – postava, ktorá trpí skôr z hľadiska režisérov kvôli negatívne hodnoteným smerom, ktorým sa jej postava vydala – ako hodnotí jeden fanúšik v epizóde „New Life and New Civilizations“ (3. sezóna, 10. epizóda):

“Well, they solved the issue of Marie Batel by turning her into a statue. Great job, writers. (sarcasm) I am struggling to see how she got chosen to be a Beholder.”

- * Christopher Pike – veľká časť tohto seriálu je zameraná na rozvoj postavy Christophera Pikea, čo prináša so sebou aj veľa kritizmu a kontroverzie. Preto je veľa diskusií v mnohých epizódach, kde používatelia majú rozmanité názory, napr.:

“Well I guess this is going to throw fuel on the ‘Pike is incompetent’ fire some people here are building. I had this same thought, as the episode does little to cast Pike’s decision-making in a positive light.”

- STT

- kladné postavy:

- * Christopher Pike – kvôli nedostatku epizód tohto novšieho (ešte nedokončeného) seriálu je v grafe najvplyvnejších postáv uvedená iba jedna postava, ktorá pozitívne vplýva na hodnotenia. Keďže je však jedinou relevantnou, nemôžeme na základe toho nič usudzovať ohľadom jeho charakteru.

- TAS

- záporné postavy:

- * Arex – vedľajšia, menšia rola pomocníka, ktorý pomáhal posádke navigovať; v recenziách nie je spomenutý inak ako pomocník, preto je záporný vplyv skôr náhodou.
- * Nurse Chapel – fanúšikovia mali problém s touto postavou skôr kvôli jej stvárneniu, ktorým jej charakter istým spôsobom „skazili“; dokonca niektorí o jej stvárnení hovoria ako o „asasinácii“ postavy Chapel:

“... with Christine Chapel being unfairly portrayed as something of a dunce.”

“Aside from the character assassination that makes Chapel look hopelessly unprofessional...”

- TNG

- kladné postavy:

- * Worf – stabilná postava obľúbená naprieč celou franšízou.
- * Admiral Nakamura – postava admirála, ktorý je obľúbený, no vyskytne sa v seriáli iba v pár epizódach.

- záporné postavy:

- * Pulaski – komunita fanúšikov sa rozdeľuje na veľmi polarizujúce skupiny s odlišnými názormi, pričom niektorí citujú doktorku Pulaski ako hviezdu epizód, v ktorých účinkuje, iní jej postavu neoceňujú vôbec – epizóda „Unnatural Selection“ (2. sezóna, 7. epizóda):

“Very weak episode. There was way too much telling and not enough showing when it came to Pulaski’s character.”

“...this episode is another proof that Pulaski could have been a great character if she stayed in the show also in further seasons.”

- * Natasha „Tasha“ Yar – prvá zmienka o postave, ktorá nebola fanúšikmi prijatá úplne pozitívne – epizóda „Legacy“ (4. sezóna, 6. epizóda):

“I feel like this was a poor first attempt to start trying to leverage the Tasha Yar character’s story.”

- TOS

- kladné postavy:

- * Scotty – jedna z hlavných postáv, často v hlavnej skupine postáv vo všetkých epizódach s posádkou, ako napr. epizóda „Mirror, Mirror“ (2. sezóna, 4. epizóda):

- “Scotty is brilliant...”*

- * McCoy – podobne ako pri postave Scottyho, jedna z centrálnych postáv, opísaná ako:

- “He’s not a miracle worker, a bricklayer, or any of a dozen other things. He is, however, the poster physician for the irascible healer with the heart of gold.”*

- záporné postavy:

- * Hikaru Sulu – typická záporná postava v kontexte seriálu so zmiešanými, no nie zriedkavými pozitívnymi komentármi.

- VOY

- kladné postavy:

- * Seven of Nine – ikonická postava tohto seriálu, na ktorú sa aj väčšina epizód zameriava – príklad epizódy „The Gift“ (4. sezóna, 2. epizóda):

- “Honestly I really like Seven of Nine, Jeri Ryan knocked it out of the park so hard it’s damned near impossible not to...”*

- * Barclay – obľúbená postava; hlavne svojím charizmatickým prístupom si vyslúžil obdiv a „lásku“ fanúšikov, ktorí ho opisujú ako:

- “There’s a certain affection you can’t help but have for a guy who struggles the way Barclay does. He’s sort of a bumbling goof when it comes to talking to other people, kind of like Rom on DS9 ... except likable, believable, and with genuine depth.”*

- záporné postavy:

- * Kes – dobrá postava, zlá exekúcia – alebo ako by to zhrnul tento opis postavy:

- “What if we had someone who was just incredibly nice? Oh, and she has untapped psychic powers? And let’s not have her do much of anything, okay?”*

8.2 Záver a odpovede na položené otázky

V projekte sme pracovali s metadátami epizód (seriál, sezóna, poradie epizódy, dátum vysielania) a s informáciami o obsadení (prítomnosť postáv) a režiséroch. Popularitu epizódy sme definovali ako z-score hodnotenia v rámci seriálu. Vplyv postáv sme vyhodnocovali cez porovnanie baseline modelov (bez postáv) a modelov s postavami / režisérom pomocou skupinovej krížovej validácie po sezónach a cez interpretáciu koeficientov regularizovaných lineárnych modelov (ridge/lasso).

Zhrňme si teda odpovede na položené otázky:

1. *Vzťahy medzi postavami a popularitou epizód* - priame porovnanie priemerov (epizódy s postavou X vs. bez X) sme nerobili ako samostatný test, ale asociácie postáv s popularitou sme odhadovali cez koeficienty v regresných modeloch. Tým pádom vieme identifikovať postavy, ktoré sú v rámci daného seriálu v *najlepšom modeli* najviac asociované s vyššou/nížšou popularitou (top- k podľa $|\beta|$; Obr. 13).
2. *Robustnosť a vplyv vonkajších faktorov* - robustnosť naprieč sezónami sme riešili nepriamo tým, že vyhodnotenie prebiehalo po sezónach (generalizácia „na iné sezóny“ v rámci toho istého seriálu). Z toho vyplýva, že v niektorých seriáloch existuje opakovateľný signál aj pri testovaní na iných sezónach, kým v iných nie.
3. *Porovnanie naprieč seriálmi a sezónami* - modely sme trénovali a vyhodnocovali separátne pre každý seriál, takže sme porovnávali najmä to, *v ktorých seriáloch* je popularita predikovateľná zo zvolených premenných. Naše výsledky ukazujú dve skupiny: seriály s **reálnou predikovateľnosťou** ($R^2 > 0$), kde model zachytí opakovateľný vzor (najmä DSC, LD, PRO, PIC, DS9, ENT; najlepšie približne DSC $R^2 \approx 0.54$), seriály **bez jasného signálu** ($R^2 \approx 0$), kde vyhráva baseline a predikcie sa zlievajú k nule (TNG, VOY, TOS, TAS, STT, SNW).
4. *Predikcia a praktická využiteľnosť modelov* - porovnanie s baseline sme spravili priamo: v časti seriálov pridanie postáv a/alebo režiséra zlepšilo predikciu oproti baseline (pozitívne R^2), zatiaľ čo v iných seriáloch neprinieslo konzistentný zisk (baseline vyhral, $R^2 \approx 0$). Dôležité prediktory sme identifikovali cez koeficienty interpretovateľných modelov (top- k postáv podľa $|\beta|$ v najlepšom modeli).
5. *Extrémy, „fan service“ a anomálie* - Z diagnostických grafov (predikcia vs. skutočnosť, histogramy rezíduí) vieme identifikovať epizódy s veľkými rezíduami ako kandidátov na „outliers“ (epizódy, ktoré model typicky nevie).

Výsledky koeficientov interpretujeme ako **asociácie v rámci seriálu** (cieľ je štandardizovaný z-score), nie ako kauzálne efekty. Koeficienty sú **modelovo závislé** (ridge rozdeľuje vplyv medzi korelované postavy, lasso môže vybrať iba podmnožinu) a **nie sú priamo porovnateľné medzi seriálmi** (odlišná variabilita a rozdelenia vstupov). Najsilnejší praktický záver je preto na úrovni: „v ktorých seriáloch má obsadenie/režisér stabilný prediktívny signál“ a „ktoré postavy vychádzajú ako najviac asociované s popularitou v rámci najlepšieho modelu daného seriálu“. Kvalitatívna časť dopĺňa tieto štatistické zistenia o kontext jednotlivých seriálov.