PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA
Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data

Bárbara Lagoeiro Moreira

Séries: seus atributos, avaliações, nomeações e premiações

Belo Horizonte

2019

Bárbara Lagoeiro Moreira

Séries: seus atributos, avaliações, nomeações e premiações

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Ciência de Dados e Big Data como requisito parcial à obtenção do título de especialista.

Belo Horizonte
2019

SUMÁRIO

1. Introdução

1.1. Contextualização

Aproveitando a possibilidade da escolha de tema para o trabalho de conclusão da pós-graduação, optei por um tema leve de meu interesse que considerei ter muitos dados disponíveis *online*: **séries**.

As séries referenciadas nesse trabalho seguem o conceito atualmente definido na *Wikipedia* para **série de televisão**: “Série de televisão, série televisiva, série de TV ou telessérie é um tipo de programa televisivo ou programa online com um número pré-definido de capítulos por temporada, chamadas episódios.[i]”

Com o objetivo de descobrir mais sobre o universo das séries e prever as chances de uma série ter recebido menção em premiações, nesse trabalho foi realizada a coleta ou obtenção de dados de séries em 4 bases diferentes:

- Bases TMDB e IMDB: Duas bases com informações gerais sobre as séries e suas avaliações:
- Bases EMMY e *Golden Globe*: Bases com informações sobre nomeações e premiações de séries nesses eventos. Os Prêmios Globo de Ouro são premiações entregues anualmente aos melhores profissionais do cinema e da televisão dentro e fora dos Estados Unidos. O Emmy é um prestigioso prêmio atribuído a programas e profissionais de televisão.

Esses dados foram processados, tratados e as bases foram analisadas em separado e em conjunto. Também foram aplicados algoritmos de classificação na base final tratada com a junção de todos os dados visando a previsão de nomeação de uma série em um dos eventos de premiação. Os resultados dessas análises são apresentados, assim como todo o procedimento realizado para chegar até eles.

Todos os arquivos referentes a bases utilizadas, scripts de coleta e processamento e notebooks com tratamento e análise de dados são disponibilizados em repositório informado ao fim deste documento.

1.2. O problema proposto

Nesse trabalho é proposta uma análise exploratória dos dados referentes a séries e a nomeações e premiações dessas no EMMY e Globo de Ouro, ao longo de anos. A seguir são respondidas as perguntas propostas pelo 5-W:

- **Por que realizar análises e extrair informações sobre dados de séries?**

Dentro do contexto do trabalho final de pós-graduação, será possível realizar um trabalho mais rico, uma vez que para o tema escolhido – séries – existe ampla disponibilidade de dados públicos que podem ser obtidos via coleta ou *download*.

O tema também é interessante e atual: as séries se tornaram uma nova versão de entretenimento presente no cotidiano das pessoas. O investimento em séries vem crescendo, principalmente com o aperfeiçoamento e popularização dos serviços de *streaming*. Produções permitem que o público acompanhe construções complexas de personagens e roteiros por um período prolongado – abrangendo meses, até anos.

As análises permitirão entender melhor essa forma de entretenimento tão atual e visualizar sua evolução ao longo do tempo.

- **Quais dados serão usados e quais as suas fontes?**

Foram pesquisados, coletados e tratados alguns dados **disponíveis publicamente *online*** sobre o tema escolhido. As seguintes bases foram construídas / obtidas e trabalhadas:

1. *Dataset* de séries do TMDB: como não foram encontrados *datasets* já prontos exclusivamente com dados de séries, nesse trabalho foram codificados um coletor e um *parser* para construir um *dataset* utilizando a API[1] disponibilizada pelo site TMDB.
2. *Datasets* disponibilizados pelo IMDB: subconjuntos de dados do IMDB[2] que são disponibilizados pelo site para uso pessoal e não comercial. Esses *datasets* apresentam dados de filmes, de séries e de outras categorias de produção misturados.
3. *Datasets* de nomeações / vencedores dos prêmios EMMY[3] e Globo de Ouro[4]: obtidos das plataformas *Kaggle* e *AggData.*

- **Quais objetivos das análises realizadas?**

As análises procurarão extrair dos dados informações de padrões, métricas e

[1]https://www.themoviedb.org/documentation/api
[2]https://www.imdb.com/interfaces/
[3]https://www.kaggle.com/pmagda/primetime-emmy-awards
[4]https://www.aggdata.com/awards/golden_globes

tendências que aumentarão nosso conhecimento sobre o tema e caracterizarão as bases trabalhadas. Além disso, quando combinados os dados de séries com os dados dos eventos de premiação, poderemos realizar análises indicativas do que faz uma série de sucesso.

- **Quais os aspectos geográficos que dizem respeito às análises**

Pode-se dizer que as bases utilizadas apresentam um "viés geográfico", uma vez que são centrados largamente em produções ocidentais, especialmente americanas.

- **A qual período se referem os dados analisados?**

As bases do EMMY e Globo de Ouro apresentam dados premiações desde a década de 1940 até 2017 e 2011, respectivamente. As bases de dados de séries produzidas vão aproximadamente da década de 1910 até os dias atuais, mais produções previstas.

A seguir é listado o período compreendido por cada um dos *datasets* utilizados:

- *Dataset* de séries do TMDB: dados desde 1914 com as séries de Charlie Chaplin até séries previstas para 2021.
- *Datasets* disponibilizados pelo IMDB: os dados de "série" mais antiga – de 1897 – se referem ao jornal diário "Palm Beach Daily News". Aqui vem o problema da base de conteúdo aberto que por vezes têm conteúdo cadastrado sem controle ou revisão. Há ainda outros registros incompletos ou referentes a jornais ou desenhos animados. Em 1906 temos o registro da série "Grand Prix Motor Racing". Há registros de séries previstas até 2022.
- *Dataset* do prêmio EMMY: nomeações de 1949 até 2017. Nem todas as categorias de premiação se referem a séries.
- *Dataset* do prêmio Globo de Ouro: nomeações de 1944 até 2011. Nem todas as categorias de premiação se referem a séries.

2. Coleta de Dados

Como citado anteriormente, nesse trabalho foram utilizadas 4 fontes de dados, sendo uma delas foi coletada e as outras obtidas. A seguir são detalhados a origem, o formato e a estrutura dos *datasets.* Os códigos utilizados para obtenção de informações podem ser visualizados no *Jupyter Notebook* 1 ***Exploração inicial dos datasets*** localizado no repositório do trabalho.

2.1 *Dataset* do TMDB

O *dataset* final obtido do TMDB apresenta 29 variáveis (colunas) e 84.141 registros.

Esse *dataset* foi construído por meio da execução dos *scripts* ***coletor.py*** e ***parser.py*** que podem ser vistos no repositório do trabalho. A API[5] disponibilizada pelo TMDB foi utilizada, mas é preciso deixar explícito de acordo com as diretivas de uso da API que este trabalho não é endossado ou certificado pelo TMDB.

A estratégia utilizada foi executar a recuperação de dados de séries via API, informando os identificadores dessas séries. A lista de identificadores de séries também é disponibilizada diariamente pelo site TMDB. A lista utilizada foi do dia *22/08/2019*.

Com a API mencionada, arquivos no formato json com descritores das séries foram recuperados (*script* ***coletor.py***). Após o fim da coleta, tais arquivos foram lidos e convertidos em formato csv (*script* ***parser.py)***.

A seguir são apresentados um exemplo do arquivo json da série “Friends” e a listagem das variáveis, sua descrição, seus tipos identificados automaticamente e seu nível de preenchimento. Destaca-se que muitas das variáveis são também definidas no formato json, sendo seus conteúdos uma lista de novas variáveis com seus atributos. No formato CSV obtido após a execução do *parser*, o conteúdo das colunas gerados por essas variáveis apresenta o texto em formato json.

Exemplo do arquivo json descritivo de uma série: “Friends”


```
{
    "backdrop_path": "/efiX8iir6GEBWCD0uCFIi5NAyYA.jpg",
    "created_by": [
        {
            "credit_id": "525710bf19c295731c03280b",
```


[5]https://www.themoviedb.org/documentation/api

```
            "gender": 1,
            "id": 163461,
            "name": "Marta Kauffman",
            "profile_path": "/muXoOXOxyFlCsCXdda25voeoc3d.jpg"
        },
        {
            "credit_id": "525710bf19c295731c032811",
            "gender": 2,
            "id": 1216352,
            "name": "David Crane",
            "profile_path": "/wXb8gWv24sX24LFuE6rSpJCjWwE.jpg"
        }
    ],
    "episode_run_time": [
        22
    ],
    "first_air_date": "1994-09-22",
    "genres": [
        {
            "id": 35,
            "name": "Comedy"
        }
    ],
    "homepage": "",
    "id": 1668,
    "in_production": false,
    "languages": [
        "en"
    ],
    "last_air_date": "2004-05-06",
    "last_episode_to_air": {
        "air_date": "2004-05-06",
        "episode_number": 18,
        "id": 87634,
        "name": "The Last One (2)",
        "overview": "Ross and Phoebe chase Rachel to the airport, but end up at the wrong
one.  They narrowly make it to the right airport, where Ross declares his love for Rachel,
but she gets on the plane anyway.  A rejected Ross returns home, where he finds a message on
```

```
his answering machine, revealing that Rachel got off the plane, just as the real thing shows
up behind him.  They decide to be together.  Meanwhile, Chandler and Monica finish packing
for their move to the suburbs, and Joey loses Chick Jr. and Duck Jr. in the foosball table.",

        "production_code": "176267",

        "season_number": 10,

        "show_id": 1668,

        "still_path": "/pbMQgT7XHSAKE2oBk548wRt7jbL.jpg",

        "vote_average": 7.611,

        "vote_count": 9

    },

    "name": "Friends",

    "networks": [

        {

            "id": 6,

            "logo_path": "/o3OedEP0f9mfZr33jz2BfXOUK5.png",

            "name": "NBC",

            "origin_country": "US"

        }

    ],

    "next_episode_to_air": null,

    "number_of_episodes": 236,

    "number_of_seasons": 10,

    "origin_country": [

        "US"

    ],

    "original_language": "en",

    "original_name": "Friends",

    "overview": "The misadventures of a group of friends as they navigate the pitfalls of
work, life and love in Manhattan.",

    "popularity": 78.567,

    "poster_path": "/7buCWBTpiPrCF5Lt023dSC60rgS.jpg",

    "production_companies": [

        {

            "id": 1957,

            "logo_path": "/nmcNfPq03WLtOyufJzQbiPu2Enc.png",

            "name": "Warner Bros. Television",

            "origin_country": "US"

        },
```

```
        {
            "id": 31810,
            "logo_path": null,
            "name": "Bright/Kauffman/Crane Productions",
            "origin_country": "US"
        }
    ],
    "seasons": [
        {
            "air_date": "2001-02-15",
            "episode_count": 7,
            "id": 4583,
            "name": "Specials",
            "overview": "",
            "poster_path": "/xaEj0Vw0LOmp7kBeX2vmYPb5sTg.jpg",
            "season_number": 0
        },
        {
            "air_date": "1994-09-22",
            "episode_count": 24,
            "id": 4573,
            "name": "Season 1",
            "overview": "",
            "poster_path": "/nvyx1LNbdQbq9xzU9LIuBcNAr2v.jpg",
            "season_number": 1
        },
        {
            "air_date": "1995-09-21",
            "episode_count": 24,
            "id": 4574,
            "name": "Season 2",
            "overview": "",
            "poster_path": "/4pKU6MfTjfNZS4RIby4Q3ukguLo.jpg",
            "season_number": 2
        },
        {
```

```
            "air_date": "1996-09-16",
            "episode_count": 25,
            "id": 4575,
            "name": "Season 3",
            "overview": "",
            "poster_path": "/mHmIOieHNRDWlvtuLqU2jSkWViZ.jpg",
            "season_number": 3
        },
        {
            "air_date": "1997-09-25",
            "episode_count": 24,
            "id": 4576,
            "name": "Season 4",
            "overview": "",
            "poster_path": "/kXcjUuAjEHGbV1ZJ0fu267dxJ5I.jpg",
            "season_number": 4
        },
        {
            "air_date": "1998-09-24",
            "episode_count": 24,
            "id": 4577,
            "name": "Season 5",
            "overview": "",
            "poster_path": "/kiz7JXtVmt6a1hnzXAST74jOMB4.jpg",
            "season_number": 5
        },
        {
            "air_date": "1999-09-23",
            "episode_count": 25,
            "id": 4578,
            "name": "Season 6",
            "overview": "",
            "poster_path": "/4phhQajxpTEStsoKlKW9wrwWFTv.jpg",
            "season_number": 6
        },
        {
```

```
            "air_date": "2000-10-12",
            "episode_count": 24,
            "id": 4579,
            "name": "Season 7",
            "overview": "",
            "poster_path": "/o7JaYswXab8RNidPk2OXdZzIoc7.jpg",
            "season_number": 7
        },
        {
            "air_date": "2001-09-27",
            "episode_count": 24,
            "id": 4580,
            "name": "Season 8",
            "overview": "",
            "poster_path": "/v6uNAPavJva8gPqIMd4FUmfRa5G.jpg",
            "season_number": 8
        },
        {
            "air_date": "2002-09-23",
            "episode_count": 24,
            "id": 4581,
            "name": "Season 9",
            "overview": "",
            "poster_path": "/6OPCnXw5bSQH7B8quLpbYfqNPT3.jpg",
            "season_number": 9
        },
        {
            "air_date": "2003-09-25",
            "episode_count": 18,
            "id": 4582,
            "name": "Season 10",
            "overview": "",
            "poster_path": "/cgUOlNYyXGjzrSIVoTMuJgFiYSq.jpg",
            "season_number": 10
        }
    ],
```

```
"status": "Ended",
"type": "Scripted",
"vote_average": 7.9,
"vote_count": 1672
```


Separadamente e utilizando a mesma API foram coletados em outro arquivo os identificadores externos dos títulos, como por exemplo, o identificador da série no IMDB. Os arquivos de dados de séries e o arquivo dos identificadores externos foram unidos para posterior processamento (vide Jupyter notebook "**1 Exploração inicial dos datasets utilizados**") visando a incorporação do identificador da série na base do IMDB (outro repositório do qual também foram obtidos dados de séries). Abaixo são mostradas as descrições das variáveis utilizadas do *dataset* final obtido do TMDB:

Listagem descritiva das colunas:

	Nome da coluna/variável	Descrição	Tipo (identificação automática)
1	**backdrop_path**	Armazena nome do arquivo de imagem do *backdrop* da série, que pode ser recuperado quando concatenado com url padrão e parametrização de tamanho. Mais de 66% dos registros apresenta esse campo nulo.	object
2	**created_by**	Lista de criadores da série. Formato json. Além de identificadores, apresenta os seguintes atributos para cada criador: **gênero, nome e caminho para foto de perfil.** Não apresenta valores nulos.	object
3	**episode_run_time**	Tempo em minutos de duração padrão do episódio da série. Aceita uma lista de valores, pois a série pode apresentar mais de um tempo de duração padrão para seus episódios. Não apresenta valores nulos.	object
4	**first_air_date**	Data em que o primeiro episódio entrou no ar, no formato YYYY-MM-DD Mais de 37% dos registros apresentam esse campo nulo.	object
5	**genres**	Lista de gêneros da série. Cada gênero tem	object

		um **id** e um **nome**. Não apresenta valores nulos.	
6	**homepage**	Endereço online da série. Mais de 73% dos registros apresentam esse campo nulo.	object
7	**id**	Identificador da série na base do TMDB. Não apresenta valores nulos.	int64
8	**in_production**	Booleano que informa se a série ainda está sendo produzida. Não apresenta valores nulos.	bool
9	**languages**	Lista de línguas da série. Não apresenta valores nulos.	object
10	**last_air_date**	Data em que a série saiu do ar, no formato YYYY-MM-DD Mais de 37% dos registros apresentam esse campo nulo.	object
11	**last_episode_to_air**	Descrição (em formato json) do último episódio que foi ao ar. Os atributos do episódio trazem os seguintes dados: data que foi ao ar, número do episódio, id, nome, sinopse, código de produção, nº da temporada, id do show, imagem do episódio, média de votos, nº de votos. Mais de 37% dos registros apresentam esse campo nulo.	object
12	**name**	Nome da série. 6 registros apresentam esse campo nulo.	object
13	**next_episode_to_air**	Próximo episódio que será transmitido e seus atributos. Mais de 99,25% dos registros apresentam esse campo nulo.	object
14	**networks**	Lista de redes de TV da série. Os atributos de uma rede de TV trazem os seguintes dados: id, caminho para imagem de logomarca, nome e país de origem. Não apresenta valores nulos.	object
15	**number_of_episodes**	Número total de episódios da série. 1,7% dos registros apresentam esse campo nulo.	float64
16	**number_of_seasons**	Número total de temporadas da série. Não apresenta valores nulos.	int64
17	**origin_country**	Lista de países de origem da série. Não apresenta valores nulos.	object
18	**original_language**	Língua original da série. Não apresenta valores nulos.	object
19	**original_name**	Nome original da série. 6 registros apresentam esse valor nulo.	object
20	**overview**	Texto que dá visão geral do que a série se trata. Mais de 40% dos registros apresentam esse campo nulo.	object
21	**popularity**	Índice de popularidade calculado pelo TMDB[6] Não apresenta valores nulos.	float64

[6]https://developers.themoviedb.org/3/getting-started/popularity

22	**poster_path**	Armazena nome do arquivo de imagem do *pôster* da série, que pode ser recuperado quando concatenado com url padrão e parametrização de tamanho. Mais de 57,22% dos registros apresentam esse campo nulo.	object
23	**production_companies**	Lista das produtoras da série. Os atributos de uma produtora trazem os seguintes dados: id, caminho para imagem da logomarca, nome e país de origem. Não apresenta valores nulos.	object
24	**seasons**	Lista das temporadas da série. Os atributos de uma temporada trazem os seguintes dados: data que entrou no ar, número de episódios, id, nome, sinopse, caminho para imagem do pôster, número da temporada. Não apresenta valores nulos.	object
25	**status**	Estado da série. Pode apresentar um dos seguintes valores: 'Ended', 'Canceled', 'Returning Series', 'In Production', 'Pilot','Planned'. Não apresenta valores nulos.	object
26	**type**	Tipo da série. Pode apresentar um dos seguintes valores: 'Scripted', 'Reality', 'Talk Show', 'News', 'Miniseries','Documentary', 'Video' Não apresenta valores nulos.	object
27	**vote_average**	Média da nota da série no TMDB. Não apresenta valores nulos.	float64
28	**vote_count**	Número de votos recebidos pela série no TMDB. Não apresenta valores nulos.	float64
29	**imdb_id**	Identificador da série na base do repositório do IMDB. Tal identificador poderá ser utilizado para junção com os *datasets* do IMDB. Mais de 56,28% dos registros apresentam esse campo nulo.	object

Observamos dessa análise inicial que os registros de séries a seguir devem ser filtrados do dataset inicial, para melhorar a qualidade da análise:

- Registros sem a informação do nome da série
- Registros de séries que estão em planejamento: campo status igual a 'Pilot' ou 'Planned'.

2.2 *Dataset* do IMDB

Foram utilizados subconjuntos de dados do IMDB[7] que são disponibilizados

[7]https://www.imdb.com/interfaces/

pelo site para uso pessoal e não comercial. Esses *datasets* apresentam dados de filmes, de séries e de outras categorias de produção misturados. Os arquivos foram obtidos na data **16/07/2019**.

Os arquivos disponibilizados que são utilizados nesse trabalho são os seguintes: **title.basics.tsv.gz** e **title.ratings.tsv.gz.** Esses arquivos serão unidos em um só *dataframe* representativo do IMDB para análises. O *dataset* de junção dos dois arquivos apresenta 11 variáveis (colunas) e 6.008.017 registros.
A descrição dos arquivos e suas variáveis está a seguir:

- **title.basics.tsv.gz** – Contém algumas informações básicas sobre os títulos:
    - **tconst (string – identificado automaticamente por object)** – identificador único alfanumérico para o título. Não apresenta valores nulos.
    - **titleType (string – identificado automaticamente por object)** – o tipo ou formato do título. Pode apresentar os seguintes valores: ['short', 'movie', 'tvMovie', 'tvSeries', 'tvEpisode', 'tvShort', 'tvMiniSeries', 'tvSpecial', 'video', 'videoGame']**.** Não apresenta valores nulos.
    - **primaryTitle (string – identificado automaticamente por object)** – o nome mais popular, usado em materiais promocionais do título. Apresenta 7 valores nulos.
    - **originalTitle (string – identificado automaticamente por object)** – o nome original do título, na língua de origem. Apresenta 182 valores nulos.
    - **isAdult (boolean – identificado automaticamente por int64)** - 0: título não adulto; 1: título adulto**.** Não apresenta valores nulos.
    - **startYear (YYYY – identificado automaticamente por float64)** – o ano de lançamento do título. No caso de séries equivale ao ano em que a série foi pela primeira vez ao ar. Apresenta mais de 5,6% de valores nulos.
    - **endYear (YYYY – identificado automaticamente por float64)** – Ano em que série terminou. '\N' para os outros tipos de título. Apresenta mais de 99,15% de valores nulos.
    - **runtimeMinutes(identificado automaticamente por objetc)**– principal tempo padrão de duração do título, em minutos. Apresenta mais de 70% de valores nulos.
    - **genres (string array – identificado automaticamente por object)** – até 3 gêneros associados ao título. Apresenta mais de 7,95% de valores nulos.

- **title.ratings.tsv.gz** – Contém as avaliações e votos do título no IMDB:

- **tconst (string - identificado automaticamente por object) -** identificador único alfanumérico para o título. Não apresenta valores nulos.
- **averageRating** (**identificado automaticamente por float64**)– nota média das avaliações recebidas para o título no IMDB. Apresenta mais de 84,14% de valores nulos.
- **numVotes (identificado automaticamente por float64)** – número total de avaliações recebidas pelo título no IMDB. Apresenta mais de 84,14% de valores nulos.

Observamos da análise inicial dos dados do IMDB que os registros de séries devem ser filtrados do dataset inicial:

- O *dataset* inicial apresenta registros de filmes e outras categorias de produção misturados às séries. Para filtrar os registros de séries podemos utilizar o campo **titleType**.
- Registros sem a informação do **primaryTitle** da série devem ser removidos também, uma vez que não apresentam o nome da série para identificação.

2.3 *Datasets* de nomeações e premiações

- EMMY: obtido da plataforma *Kaggle*[8] em **17/07/2019**. Contém nomeações e premiações do EMMY de 1949 a 2017. O dataset inicialmente apresenta 5 variáveis (colunas) e 19.239 registros. Pelos valores das 1043 categorias diferentes é possível ver que nem todas as premiações se referem a séries(vide Jupyter notebook “1 Exploração inicial dos datasets utilizados”). Mas não vamos filtrá-las para não correr o risco de perder informações. Vamos tentar filtrar os *datasets* do TMDB e IMDB de forma mais objetiva para no momento da junção de bases ficarmos apenas com as nomeações e premiações de séries. A seguir a descrição das colunas do arquivo:

Nome da coluna/variável	Descrição	Tipo (Detectado automaticamente)
year	Ano do EMMY. Exemplo: 1949. Não apresenta valores nulos.	int64
category	Categoria da premiação. De 1949 até 2017 existem **1043** categorias de premiação distintas! Exemplos: Technical Award , Best Film Made For Television Apresenta 14 valores nulos.	object

[8]https://www.kaggle.com/pmagda/primetime-emmy-awards

winner	1 indica vencedor, 0 não vencedor. Não apresenta valores nulos.	int64
nominee	Nome do título ou pessoa que foi nomeado ao prêmio. Apresenta 9 valores nulos.	object
detail	Quando uma pessoa é nomeada, muitas vezes apresenta o nome do título. Quando um título é nomeado, muitas vezes apresenta o nome da rede (CBS, Netflix, HBO) que disponibiliza o título. Apresenta 266 valores nulos.	object

- Globo de Ouro: obtido da plataforma AggData[9] em **26/08/2019**. Contém nomeações e premiações de 1944 até 2011. O dataset apresenta 4 variáveis (colunas) e 6.890 registros. Nem todas as premiações se referem a séries, pois o evento premia cinema e televisão. Adotaremos a mesma política definida em relação às categorias do EMMY: não vamos filtrá-las para não correr o risco de perder informações. Vamos tentar filtrar os *datasets* do TMDB e IMDB de forma mais objetiva para no momento da junção de bases ficarmos apenas com as nomeações e premiações de séries. A seguir a descrição das colunas do arquivo:

Nome da coluna/variável	Descrição	Tipo (Detectado automaticamente)
Year	Ano do Globo de Ouro. Exemplo: 2011. Não apresenta campos nulos.	int64
Category	Categoria da premiação. De 1944 até 2011 existem **75** categorias de premiação distintas! Exemplos: Best Motion Picture - Drama, Best Television Series – Drama Não apresenta campos nulos.	object
Nominee	Nome do título ou pessoa que foi nomeado ao prêmio. Apresentam 6 valores nulos.	object
Won?	“yes” indica vencedor, “no” não vencedor. Não apresenta campos nulos.	object

3. Processamento/Tratamento de Dados

Para processamento, tratamento e análise de dados continuaremos a utilizar a linguagem Python, via *Jupyter Notebooks* para facilitar o acompanhamento e reprodução do que foi realizado.

[9]https://www.aggdata.com/awards/golden_globes

A seguir, primeiramente serão apresentados os tratamentos individuais dos 4 *datasets* previamente construídos ou obtidos. Posteriormente, será realizada junção desses datasets, e novo processamento de dados será realizado.

3.1 *Dataset* do TMDB

Para acompanhar os processamentos aqui descritos, vide *Jupyter Notebook* **2 TMDB – Processamento e Tratamento de Dados.**

Originalmente o dataset apresenta 29 variáveis e 84.141 registros de séries.

Tratamento de dados ausentes

Destaca-se que vários campos do *dataset* TMDB contêm dados em formato de lista, alguns sendo listas de dados no formato JSON (para permitir a captura de múltiplos valores em uma única coluna CSV). Para mais correta análise de valores ausentes, ao ler a base de dados para o *dataframe*, adicionamos o valor '[]' para ser reconhecido como “null value”, uma vez que define listas vazias em vários campos. Sem esse valor adicionado entre os valores considerados nulos todos os campos de lista (**created_by, genres, networks, production_companies, episode_run_time, languages e origin_contry**) tinham preenchimento 100%, o que é um fato distante da realidade como vemos na tabela abaixo.

A seguir o status inicial do *dataset* mostrando o fator de preenchimento em porcentagem para cada uma de suas colunas, considerando o valor '[]' como nulo, e uma descrição para cada coluna sobre tratativas a serem executadas ou observações:

Nome da coluna/variável	% de preenchimento	Observação
next_episode_to_air	0,74	**Será excluído das análises.** Considerado sem relevância a informação de qual será o próximo episódio, além do percentual de nulidade ser elevadíssimo.
created_by (lista)	20,04	Trás informações interessantes como o nome e o sexo do criador da série. Porém apenas 20,04% dos registros apresenta essa informação preenchida.
production_companies (lista)	20,11	Trás informações interessantes como o nome da produtora da série. Apenas 20,11% dos registros apresenta essa

		informação preenchida.
homepage	25,75	Muitas séries não têm um site oficial informado nos dados.
backdrop_path	33,55	**Será excluído das análises.** Considerado sem relevância. Se quisermos ilustrar com imagens, usamos a do pôster que é mais presente.
poster_path	42,77	Será mantido para possíveis ilustrações de *dashboard*.
imdb_id	43,71	É grande o percentual de séries SEM o id correspondente no IMDB. Importante considerar esse dado ao trabalhar com os dois *datasets* em conjunto.
genres (lista)	48,51	Trás informações dos gêneros da série.
networks (lista)	54,48	Trás informações das redes da série.
episode_run_time (lista)	57,30	Trás a informação dos tempos médios padrão de duração de um episódio da série.
languages (lista)	57,74	Línguas faladas na série.
overview	59,88	Sinopse da série. Será mantido para possíveis construções relativas a análise de conteúdo.
last_air_date	62,28	Esse campo será nulo para as séries que ainda não acabaram…então o percentual de nulidade por falta de informação existente pode ser baixo.
last_episode_to_air	62,28	**Será excluído das análises.** Considerado sem relevância a informação de qual foi o último episódio.
first_air_date	62,75	Podemos verificar o campo *startYear* no dataset do IMDB para ver se complementamos as informações faltantes de alguma forma.
seasons (lista)	66,91	**Será excluído das análises.** Considerado sem relevância a informação específica dos episódios por temporada.
origin_country (lista)	67,15	
number_of_episodes	98,30	
original_name	99,99	**Será excluído das análises.** Optamos por usar o campo *name*, que se refere ao nome mais popular da série. No notebook é possível ver como que o **original_name** da série é o nome na língua original da série. Para análise com as demais bases coletadas, nos interessa o nome traduzido.
name	99,99	Esse percentual de nulos será excluído das análises, pois consistem em 6 registros, sem informação de nome ou nome original.

in_production	100	
number_of_seasons	100	
popularity	100	
vote_count	100	
id	100	
status	100	
type	100	
vote_average	100	
original_language	100	

Alguns dos campos apresentam alto índice de nulidade e em análise inicial verificamos que devem ser excluídos da base:

- Colunas: **next_episode_to_air, backdrop_path, last_episode_to_air, seasons, original_name**
- Registros sem a informação do nome da série: Os 6 registros com nome nulo foram analisados. Não apresentam dados de nome original (como demonstrado no notebook), somente um deles tem id do IMDB e esse id se refere a um filme, não a uma série. Foi feita a opção pela exclusão desses registros.

Filtragem

- Registros de séries que estão em planejamento: campo status igual a 'Pilot' (86 registros) ou 'Planned' (124 registros) serão excluídos das análises também, pois o objetivo final é a análise de nomeações e premiações, e esse tipo de série não tem condição de ter sido nomeada.
- Registros com data de início posterior à atual também serão excluídos, pois o objetivo final é a análise de nomeações e premiações, e esse tipo de série não tem condição de ter sido nomeada.

Quanto aos outros campos com valores nulos, para aqueles que forem considerados relevantes para as análises, quando da junção do *dataset* TMDB com o *dataset* IMDB, veremos o que pode ser possível complementar.

Tratamento de dados duplicados

Na base coletada não existe duplicidade exata de registro, ou registros com o mesmo identificador id. No entanto, verificando a duplicidade de nome de série, é encontrado um valor alto. Analisando as entradas com nome duplicado vemos que por vezes a recorrência do nome é o caso de uma série que teve versões diferentes em vários países, como por exemplo, 1 vs. 100. Ou uma série que teve muitas versões ao longo do tempo, como *Superman*. Dado esse contexto, foi restrita a análise de duplicidade para considerar além do nome, a data de estreia, a língua e o país de origem da série.

Considerando esses 4 atributos para definir a unicidade do registro, foram obtidos ainda 1241 registros com duplicidade entre si. Investigando essas 1241 linhas com mesmo nome, data de início e língua / país de origem temos que muitos registros apresentam dados da data de início nulos e os demais dados vazios! Optou-se então por manter desses registros duplicados apenas o registro que tiver maior número de dados preenchidos. Ao final dessas exclusões, não restaram mais duplicatas considerando os atributos (nome, data de início e língua / país de origem) da série e as novas dimensões do *dataset* resultante foram (83.227, 24).

Correções e complementações de dados

1. Ao analisar as séries mais antigas e as mais recentes do *dataset* para validar a formatação das datas, observamos que para os anos 2000 o *dataset* do TMDB para alguns registros adotou o padrão 0000 em vez de 2000. Isso foi tratado para não interferir nas análises temporais. Sete registros tiveram suas datas de início corrigidas.
2. Complementar dados:
    1. A coluna que contém referência para o pôster foi alterada para conter o caminho completo de uma URL válida no site do TMDB. Assim poderá ser utilizada para ilustrar as análises a serem apresentadas.
    2. A coluna de data de estreia da série foi quebrada para serem armazenados e analisados em separado o ano, o mês e o dia de lançamento da série.

Após todos os procedimentos acima descritos, o *dataset* do TMDB passou a apresentar 25 colunas e 83.228 registros (Atenção! Esses valores podem variar em novas execuções do notebook, pois há um método que exclui registros baseando-se na data atual).

3. 2 *Dataset* do IMDB

Para acompanhar os processamentos aqui descritos, vide *Jupyter Notebook* **3 IMDB – Processamento e Tratamento de Dados.**
Originalmente o dataset apresenta 11 variáveis e 6.008.017 registros de séries, filmes e outros tipos de produção misturados.

Filtragem

O *dataset* inicial apresenta registros de filmes e outras categorias de produção misturados às séries. Para filtrar os registros de séries utilizamos o campo **titleType**, que não apresenta valores nulos. Esse campo pode apresentar os seguintes valores: 'short', 'movie', 'tvMovie', 'tvSeries', 'tvEpisode', 'tvShort', 'tvMiniSeries', 'tvSpecial', 'video', 'videoGame'. Vamos considerar apenas os registros com "titleType" igual a "tvSeries" ou "tvMiniSeries". Após essa filtragem, que pode ser executada pelo notebook, temos um *dataset* resultante de 195.040 registros.

Além disso serão filtrados também registros com data de início posterior à atual, pois o objetivo final é a análise de nomeações e premiações, e esse tipo de série não tem condição de ter sido nomeada. Após essa filtragem, que pode ser executada pelo notebook, temos um *dataset* resultante de 194.812 registros.

Tratamento de dados ausentes

Após a filtragem, temos o seguinte status de dados ausentes para as colunas do dataframe do IMDB:

Nome da coluna/variável	% de preenchimento	Observação
endYear	26,14	Pode ser usado em complementação com **last_air_date** da base TMDB.

averageRating	37,28	
numVotes	37,28	
runtimeMinutes	42,62	Pode ser usado em complementação com **episode_run_time** da base TMDB.
genres	90,29	Pode ser usado em complementação com **genres** da base TMDB.
startYear	95,03	Pode ser usado em complementação com **start_air_date** da base TMDB.
originalTitle	99,99	Será excluído para pois está em redundância com o campo **primaryTitle**, que será utilizado.
tconst	100	Identificador que será usado na junção dessa base com a base do TMDB.
titleType	100	Informa a categoria da série. Pode ser 'tvSeries' ou 'tvMiniSeries'
primaryTitle	100	
isAdult	100	

Quatro colunas poderão ser usadas em complementação de informações de registros quando da junção com a base de dados do TMDB: endYear, runtimeMinutes, genres e startYear.

Observa-se que para registros de séries temos preenchimento completo para a variável "primaryTitle". Já "originalTitle" é redundante com o “primaryTitle” em quase 97% da base e tem 18 nulidades (verificações podem ser vistas no *notebook*). Por isso será excluída.

Tratamento de dados duplicados

Na base coletada não existe duplicidade exata de registro, ou registros com o mesmo identificador id. No entanto, verificando a duplicidade de nome de série, da mesma forma que aconteceu na base do TMDB, é encontrado um valor alto. Foi restrita a análise de duplicidade para considerar além do nome, a data de estreia e a data de fim da série. Dessa foram encontradas 1.658 entradas duplicadas. No entanto, ao analisá-las (vide notebook), vimos que algumas das duplicidades podem se referir a produções de países diferentes (exemplo: X Factor). Como não

buscamos os dados de região do IMDB vamos deixar para fazer essa análise após junção com a base do TMDB.

Dimensões do dataset após execução dos processamentos iniciais: 194.812 e 10.

3.3 *Dataset de séries – junção TMDB e IMDB*

Após processamento em separados das bases do TMDB e do IMDB chegou o momento de unificar as bases. Para acompanhar os processamentos aqui descritos, vide Jupyter Notebook **4 Junções TMDB IMDB – Processamento e Tratamento de Dados.**

Originalmente os *datasets* que serão unificados apresentam:

- IMDB: 194.812 registros e 10 colunas.
- TMDB: 83.228 registros e 25 colunas.

Nessa unificação vamos:

1. Unir as bases no modo OUTER JOIN, ou seja, vamos considerar toda e qualquer série, independentemente do registro se referir a apenas uma das bases.
2. Tratar valores ausentes após a junção.
3. Tratar valores duplicados após a junção.

Junção das bases

Na junção consideraremos todos os registros, independente de existirem somente em uma das bases. Com isso inicialmente obtivemos um dataframe com 37 colunas e 244.790 registros de séries.

No *notebook* é possível acompanhar o processamento inicial desse *dataframe*:

- A exclusão de colunas duplicadas (referentes a identificadores gerados)
- A mudança de nome das colunas que têm o mesmo significado mas apresentam origens diferentes, seguindo um padrão. Exemplo:

numVotesTMDB e numVotesIMDB, averageRatingTMDB e averageRatingIMDB.

Tratamento de dados ausentes

Uma vez que o dataframe de junção foi obtido unindo as bases independentemente da coexistência de uma dada série nas duas bases, obtivemos um *dataset* mais esparso, com muitos valores nulos. A figura a seguir mostra o número de valores ausentes e a porcentagem de preenchimento de cada coluna do *dataset*:

	column_name	missing_count	filling_factor
0	created_by	228057	6.835655
16	production_companies	227993	6.861800
3	homepage	222374	9.157237
15	poster_path	208941	14.644798
21	imdb_id	208195	14.949549
2	genres_TMDB	204169	16.594224
8	networks	199221	18.615548
1	episodeRuntimeTMDB	196778	19.613546
6	languages	196496	19.728747
13	overview	194755	20.439969
29	endYearIMDB	193857	20.806814
24	endYearTMDB	192495	21.363209
23	startMonthTMDB	192219	21.475959
22	startYearTMDB	192219	21.475959
11	origin_country	188690	22.917603
32	averageRatingIMDB	172151	29.674006
33	numVotesIMDB	172151	29.674006
9	number_of_episodes	162927	33.442134
30	episodeRuntimeIMDB	161758	33.919686
12	original_language	161562	33.999755
4	tmdb_id	161562	33.999755
14	popularity	161562	33.999755
5	in_production	161562	33.999755
17	status	161562	33.999755
10	number_of_seasons	161562	33.999755
20	numVotesTMDB	161562	33.999755
19	averageRatingTMDB	161562	33.999755
18	type	161562	33.999755
7	name	161562	33.999755
31	genres_IMDB	68889	71.857919
28	startYearIMDB	59646	75.633809
26	primaryTitle	49978	79.583316
27	isAdult	49978	79.583316
25	titleType	49978	79.583316

Combinação de colunas TMDB – IMDB

O primeiro tratamento que será feito sobre valores ausentes será a combinação das colunas de diferentes origens (TMDB e IMDB) de mesmo significado, que possam se complementar. Combinar alguns atributos de mesmo significado nas diferentes bases para uma mesma série. Isso trará benefícios no preenchimento de dados faltantes pois aumentará o preenchimento com dados reais! Segue a lista dos atributos que serão combinados em um:

- *name e primaryTitle*: o nome da série nas duas bases será unificado em um só campo 'name'.
  - Inicialmente o percentual de preenchimento dos campos de primaryTitle e name eram de 79.583316 e 33.999755, respectivamente. Após a combinação das informações no campo 'name', temos o percentual de preenchimento aumentado para 100.00.
  - A combinação ocorreu da seguinte maneira: se o campo do IMDB for nulo, o novo campo assume o valor do campo do TMDB, senão, assume o campo do IMDB.
- *endYearIMDB e endYearTMDB*:
  - Inicialmente o percentual de preenchimento dos campos de endYearIMDB e endYearTMDB eram de 20.806814 e 21.363209, respectivamente. Após a combinação das informações no campo 'endSeriesYear', temos o percentual de preenchimento aumentado para 37.836921.
  - A combinação ocorreu da seguinte maneira: se o campo do IMDB for nulo, o novo campo assume o valor do campo do TMDB, senão, assume o campo do IMDB.
- startYearIMDB e startYearTMDB:
  - Inicialmente o percentual de preenchimento dos campos de startYearIMDB e startYearTMDB eram de 75.633809 e 21.475959, respectivamente. Após a combinação das informações no campo 'startSeriesYear', temos o percentual de preenchimento aumentado para 85.622370.

- A combinação ocorreu da seguinte maneira: se o campo do IMDB for nulo, o novo campo assume o valor do campo do TMDB, senão, assume o campo do IMDB.

- episodeRuntimeIMDB e episodeRuntimeTMDB:
  - Inicialmente o percentual de preenchimento dos campos de episodeRuntimeIMDB e episodeRuntimeTMDB eram de 33.919686 e 19.613546, respectivamente. Após a combinação das informações no campo 'episodeRuntimeSeries', temos o percentual de preenchimento aumentado para 46.244536.
  - A combinação ocorreu da seguinte maneira: Esse campo na base do TMDB é uma lista e no IMDB um valor único. Para o campo que representará o tempo médio padrão final dos episódios seguiremos a regra: se existir o campo no IMDB será o valor do IMDB, senão será uma média dos valores da lista, se houver o valor do TMDB.
- genres_IMDB e genres_TMDB:
  - Inicialmente o percentual de preenchimento dos campos de genres_IMDB e genres_TMDB eram de 71.857919 e 16.594224, respectivamente. Após a combinação das informações no campo 'genres', temos o percentual de preenchimento aumentado para 78.499122.
  - A combinação ocorreu da seguinte maneira: A lista de gêneros da série no TMDB é uma lista de JSONs e no IMDB uma lista de Strings separadas por vírgula. Como os valores podem ser diversos, literalmente e em significado, vamos optar por utilizar a união das duas informações. O formato final seguirá o padrão do IMDB, de lista de nomes de gêneros separados por vírgula. Utilizaremos a função set para eliminar possíveis duplicatas.
  - Uma observação importante é que a denominação de um gênero às vezes é diferente dependendo da base de origem, por exemplo: 'Action' ou 'Action & Adventure'.

Preenchimento de nulos relativos a métricas de avaliação

Os campos relativos às avaliações das séries não serão combinados, pois deseja-se verificar diferenças de impacto das bases IMDB e TMDB. Os valores nulos para os campos numVotesIMDB, averageRatingIMDB, numVotesTMDB, averageRatingTMDB e popularity serão preenchidos com o valor zero, pois esse valor reflete mais fielmente a ausência de avaliação.

Preenchimento de nulos relativos a métricas de tamanho

Os campos relativos ao tamanho das séries terão os valores nulos preenchidos com o valor médio definido para as outras séries. A ausência do valor claramente não indica o tamanho zero, então optou-se por utilizar o valor médio existente para o campo. Os campos que se enquadram nessa categoria são: number_of_episodes, number_of_seasons e episodeRuntimeSeries.

Preenchimento de nulos relativos a valores de listas em formato json

Como estratégia de preenchimento dos valores nulos de colunas que têm listas (em formato json ou não) como valor vamos:

1. Analisar as ocorrências de valores distintos dessas colunas
2. Buscar os 3 valores mais recorrentes
3. Criar colunas booleanas novos para informar qual desses valores está presente para cada registro. Por exemplo, no caso da lista de gêneros, comédia está entre os gêneros mais recorrentes, então será criada e preenchida a coluna ehComedia.
4. A coluna com valores de lista será removida do dataframe.

Abaixo e no notebook (com mais detalhes) é possível acompanhar a análise e tratamento dos valores nulos para cada uma das colunas a seguir, que apresenta como valores lista em formato json:

- *created_by:* Número de criadores distintos: 14.544. Top 3: 'ABS-CBN', com 109 registros, 'John de Mol', com 65 registros e 'Simon Fuller' com 64 registros. Os top 3 se referem a uma rede de televisão filipina ('ABS-CBN'), ao criador do Big Brother e outros *Reality Shows* ('John de Mol') e ao produtor de TV britânico de *Idols* e *So You Think You Can Dance* ('Simon Fuller')
- *production_companies*: Número de produtoras distintas: 8.897. Top 3: 'BBC', com 451 registros; 'Warner Bros. Television', com 254 registros e 'Universal Television' com 239 registros.
- *networks*: Número de networks distintos: 1.862. Top 3: 'ABC', com 1.881 registros, 'BBC One', com 1.712 registros e 'NBC', com 1.442 registros.

Foram criados os atributos:

- para substituir o atributo production_companies: ehBBC, ehWarner, ehUniversal
- para substituir o atributo created_by: ehABS_CBN, ehJohnDeMol, ehSimonFuller
- para substituir o atributo networks: ehABC, ehBBC, ehNBC

A substituição dos valores nulos para as colunas também de listas 'genres' e 'languages' será feita posteriormente, após as análises do próximo notebook e seção do presente trabalho.

Preenchimento de nulos relativos a valores que podem ser convertidos em Booleanos

A substituição dos valores nulos para as colunas 'homepage' e 'overview' será feita posteriormente, após as análises do próximo notebook e seção do presente trabalho. As colunas passarão a apresentar valores booleanos, informando se há 'homepage' ou se há 'overview' cadastradas para a série.

Exclusão de campos pouco informativos para análise ou redundantes

Por enquanto serão excluídos os campos identificadores das bases IMDB e TMDB e o campo in_production. Após mais análises, poderemos verificar a necessidade de mais exclusões.

Tratamento de dados duplicados

Como já foram realizadas análises nas bases em separado vamos direto para a análise de duplicidade pelos dados de nome, data de estreia e país de origem da série: 5.303 registros com duplicidade encontrada em um dataframe de atuais 244.790 registros e 31 colunas. Como das outras vezes, vamos optar por excluir os registros duplicados que apresentem mais vazios ou nulos (aplicação da estratégia pode ser vista no *notebook*).

Dimensões do dataset após tratamento: **242.090 registros e 31 colunas**.

3.4 *Datasets* de nomeações e premiações

Para acompanhar os processamentos aqui descritos, vide *Jupyter Notebook* **5 EMMY e GG – Processamento e Tratamento de Dados.**

Originalmente os *datasets* apresentam:

- *Golden Globe*: 4 variáveis ('Nominne', 'Year', 'Category' e 'Won?') e 19.239 registros de nomeações.
- *Emmy*: 5 variáveis ('nominne', 'year', 'category', 'winner' e 'detail') e 6.890 registros de nomeações.

Nesse processamento, vamos padronizar as colunas dos *datasets* de premiação e realizar uma junção de seus dados em um só *dataset*.

Filtragem

Os *datasets* de premiações são conjuntos pequenos de dados, por isso escolhemos NÃO filtrar as categorias de premiação, apesar de ser possível definir uma heurística para isso (por exemplo, categorias que contenham a palavra "series"). Essa “filtragem” ocorrerá automaticamente quando unirmos os *datasets* de premiações ao *dataset* de séries. Essa estratégia maximizará os relacionamentos possíveis.

Tratamento de dados ausentes

Golden Globe

O único tratamento a ser realizado será a exclusão dos 6 registros cujo campo 'Nominee' é nulo, pois não há como saber a que série se referem (se for o caso de se referirem a uma série). Na tabela abaixo observa-se que o preenchimento dos campos fica em 100% após essa exclusão.

Nome da coluna/variável	% de preenchimento	Observação
Nominee	99,91	Os 6 registros nulos serão excluídos, uma vez que a coluna com o nome da série nomeada não poderia ser obtido.
Year	100	
Category	100	
Won?	100	

Emmy

Também serão excluídos 9 registros cujo campo 'nominee' é nulo e 14 registros cujo campo 'category' é nulo, pois não há como saber a que série se referem (se for o caso de se referirem a uma série). Na tabela abaixo observa-se que o preenchimento dos campos fica em 100% exceto pela coluna 'detail' que será excluída após a criação da nova coluna 'seriesName' que será feita um pouco mais a frente.

nome da coluna/variável	% de preenchimento	Observação
nominee	99,95	Os 9 registros nulos serão excluídos, uma vez que a coluna com o nome da série nomeada não poderia ser obtido.
year	100	
category	99,93	Existem 14 registros com a categoria nula. Esses registros também serão excluídos, uma vez que a coluna com o nome da série nomeada não poderia ser obtido no momento do cálculo da coluna 'seriesName'
winner	100	
detail	98,62	Coluna será excluída após obtenção da coluna 'seriesName'

Tratamento de dados duplicados

Os registros duplicados encontrados se referem a nomeações realmente diferentes para as premiações. São o caso de mais de uma nomeação na mesma categoria, para mesma série, mesmo ano, mas a pessoa (ator ou atriz) nomeada é diferente. Exemplo: no Golden Globe, em 2001 e 2000 *West Wing* teve duas nomeações para categoria "Best Performance by an Actor In A Television Series - Drama". Um nomeado foi o ator Rob Lowe (ganhador em 2001) e o outro Martin Sheen.

Dessa forma NÃO haverá exclusão das duplicidades, pois cada registro se refere a uma nomeação diferente. No notebook são mostrados mais exemplos disso, para os dois *datasets*.

Correções e complementações de dados

Foram realizados processamentos nos *datasets* objetivando padronizar as colunas, para junção das bases e principalmente definir a coluna com **nome da série** (coluna 'seriesName') que tenha recebido nomeação ou premiação.

Cálculo da coluna com nome da série:

- No caso do *dataset* Golden Globe:
    - Se a categoria premiar um ator ou atriz, o nome da série se encontra após a palavra 'in' na coluna 'nominee'
    - Senão: o nome da série se encontra na coluna 'nominee'.
- No caso do *dataset* Emmy:
    - Se a categoria premiar um ator ou atriz ou performance, o nome da série se encontra na coluna 'detail'
    - Senão: o nome da série se encontra na coluna 'nominee'.

Cálculo do número de nomeações e prêmios por série e criação de uma base unificada

Realizamos o agrupamento dos registros pelo nome da série, para obtermos um arquivo que relacione um nome de série e o número de nomeações e de premiações dessa série no EMMY e no GG. Esse arquivo é que será futuramente analisado junto às informações do arquivo de séries. A '*Criação de base de premiação agrupada e unificada*' pode ser vista no notebook.

Ao final desses processamentos, os dois *datasets* passaram por agrupamentos de seus registros e foram unidos com a estrutura apresentada a seguir, totalmente preenchida e sem nulos. Como todas as colunas se referem a quantidades de nomeações ou premiações, os valores nulos foram preenchidos com o valor zero, para facilitar também os somatórios nas análises.

Dimensões do dataset obtido: 7 atributos, com 10.563 registros. Os atributos são os que seguem:

- seriesName: nome da série que foi obtido como explicado anteriormente e utilizado para a junção das informações das bases do EMMY e do Golden Globe.
- nominees_EMMY: quantidade de nomeações que a série recebeu no EMMY, ao longo do tempo.
- nominees_GG: quantidade de nomeações que a série recebeu no Gonden Globe, ao longo do tempo.
- prizes_EMMY: quantidade de premiações que a série recebeu no EMMY, ao longo do tempo.
- prizes_GG: quantidade de premiações que a série recebeu no Gonden Globe, ao longo do tempo.
- total_nominees: quantidade de nomeações que a série recebeu no EMMY ou no Gonden Globe, ao longo do tempo.
- total_prizes: quantidade de premiações que a série recebeu no EMMY ou no Gonden Globe, ao longo do tempo.

4. Análise e Exploração dos Dados

Para acompanhar os processamentos aqui descritos, vide *Jupyter Notebook* **6 Análise de Dados.** Nessa seção vamos explorar os dados das bases já trabalhadas anteriormente, analisando as ocorrências, padrões, *rankings,* buscando informações que possamos obter dos *datasets*.

4.1 Dataset de Séries (obtido com a junção das bases IMDB e TMDB)

Análise de métricas – votos, nota média e popularidade

Nessa análise vamos procurar entender quais são os tipos de séries mais bem avaliados pelas métricas presentes no *dataset*.

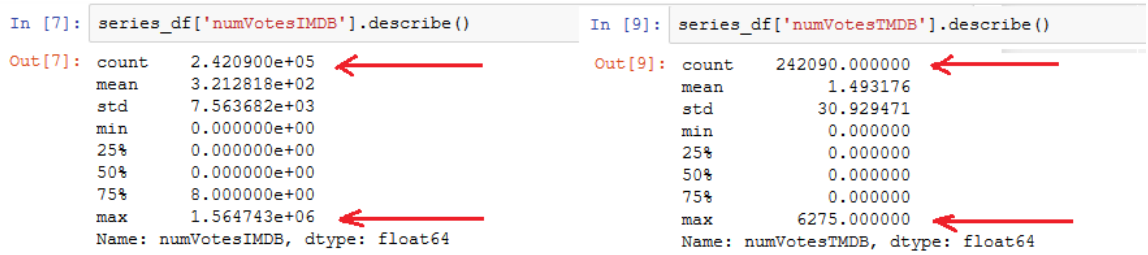
- ***Popularity***: É uma métrica do TMDB baseada em vários quesitos: nº de votos no dia, nº de visualizações no dia, nº de usuários que 'favoritaram' a série

naquele dia, nº de usuários que adicionaram a série à "*watchlist*" no dia, data de lançamento do próximo ou último episódio, número total de votos, pontuação em dias anteriores. É uma métrica baseada na popularidade do momento atual, pois leva em consideração muitos quesitos referentes ao dia, mais do que quesitos acumulados ao longo do tempo.

A pontuação de popularidade parece ser uma quantidade extremamente distorcida, com uma média de apenas 0,57, mas valores máximos atingindo 318.6. No entanto, como pode ser visto no gráfico de distribuição a seguir, quase todos os filmes têm uma pontuação de popularidade menor que 1 (o 75º percentil está em 0,6). Somente registros provenientes da base do TMDB apresentam valores não zerados de popularidade. O tipo de série mais bem avaliado pela popularidade é a série “do momento”. Dentre as top 10 em 'popularity' temos 4 com a temática super herói: *The Flash*, em primeiro, *Arrow*, em terceiro lugar, *The Boys*, em sétimo lugar e *Marvel's Agents of S.H.I.E.L.D*. em décimo. Temos duas séries de desenho: *Family Guy* e *The Simpsons* (pode ser visto pelo notebook).

```
In [4]: series_df['popularity'].describe()

Out[4]: count    242090.000000
        mean          0.571287
        std           3.069344
        min           0.000000
        25%           0.000000
        50%           0.000000
        75%           0.600000
        max         318.617000
        Name: popularity, dtype: float64
```

- ***Número de votos – TMDB e IMDB***: Observa-se que a quantidade de votos no IMDB é muitas ordens de grandeza maior que a quantidade de votos no TMDB. A base do IMDB é bem mais popular e suas votações bem mais expressivas. Podemos ver também que no TMDB (até o 3º quartil), mas também no IMDB (até o 2º), grande número de séries NÃO receberam votos. Lembrando que esse dataset é a junção (via OUTER join) das duas bases e que essas métricas são específicas de cada base.

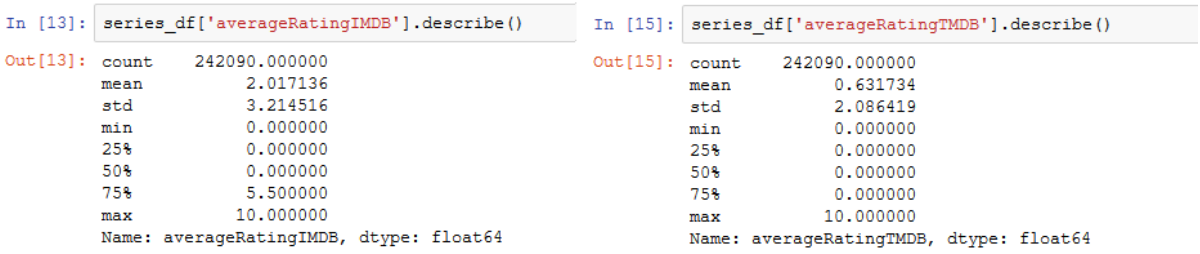


```
In [7]: series_df['numVotesIMDB'].describe()

Out[7]: count    2.420900e+05
        mean     3.212818e+02
        std      7.563682e+03
        min      0.000000e+00
        25%      0.000000e+00
        50%      0.000000e+00
        75%      8.000000e+00
        max      1.564743e+06
        Name: numVotesIMDB, dtype: float64
```

```
In [9]: series_df['numVotesTMDB'].describe()

Out[9]: count    242090.000000
        mean          1.493176
        std          30.929471
        min           0.000000
        25%           0.000000
        50%           0.000000
        75%           0.000000
        max        6275.000000
        Name: numVotesTMDB, dtype: float64
```

Verificando as 10 séries mais votadas pelos dois parâmetros, vemos que 5 das 10 séries mais votadas são coincidentes e que o primeiro lugar é o mesmo “Game Of Thrones”(vide notebook para mais detalhes).

- ***Nota média – TMDB e IMDB***: A nota média dos sites também é baixa, sendo que a do IMDB é mais expressiva (0,63 para TMDB e 2.02 no IMDB) pelo motivo já citado anteriormente: a junção das bases ter sido feita no modo *OUTER join* para trabalharmos com maior número de dados fez com que essas métricas específicas de cada base apresentassem muitos zeros .



```
In [13]: series_df['averageRatingIMDB'].describe()

Out[13]: count    242090.000000
         mean          2.017136
         std           3.214516
         min           0.000000
         25%           0.000000
         50%           0.000000
         75%           5.500000
         max          10.000000
         Name: averageRatingIMDB, dtype: float64
```

```
In [15]: series_df['averageRatingTMDB'].describe()

Out[15]: count    242090.000000
         mean          0.631734
         std           2.086419
         min           0.000000
         25%           0.000000
         50%           0.000000
         75%           0.000000
         max          10.000000
         Name: averageRatingTMDB, dtype: float64
```

Ao verificar quais são as 10 séries mais bem avaliadas pela crítica restringiremos a avaliação às séries que tiveram um número mínimo de votos: 1000 no caso do TMDB e 10000 no caso do IMDB, porque ter uma média alta com poucos avaliadores é fácil.

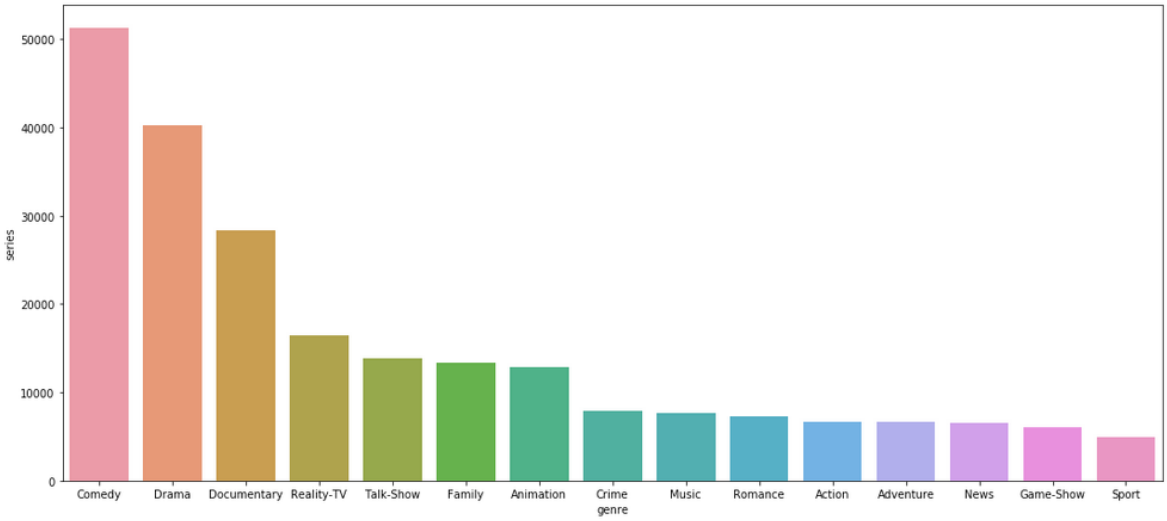
Nas duas tabelas resultantes que podem ser consultadas no notebook, as séries foram ranqueadas primeiro pela nota média no IMDB e depois pela nota média do TMDB. Em ambas as tabelas mostramos informações de métricas dos dois sites para comparação.

Uma descoberta interessante: observamos que para 4 séries do top 10 nota média do IMDB (com no mínimo 10.000 votos no site) temos uma votação insignificante no site do TMDB. Essas 4 séries (The Ants ou Koombiyo, The Filthy Frank Show, College Romance e Yeh Meri Family) são asiáticas, duas indianas, o primeiro lugar é do Sri Lanka e o segundo lugar é uma série de um canal de um

rapaz japonês no Youtube, algo que não havia sido previsto no início desse trabalho. Diante dessa observação podemos ponderar que o IMDB é uma base divulgada mundialmente, apresentando conteúdo e avaliações que ultrapassam o comum americano e englobam produções extremamente independentes e variadas, tais quais as presentes nos canais do Youtube. Já o TMDB é uma base mais focada em conteúdo americano e produtoras tradicionais.

Análise de incidências – gêneros, origem, tempo por episódio, nº de episódios e temporadas

Gêneros: TMDB e IMDB definem 38 denominações diferentes de gêneros. A maior incidência com considerável distância dos demais gêneros é de dramas e comédias. Não sem propósito, nas premiações temos várias categorias exclusivas para premiar séries desses dois gêneros. Abaixo gráfico mostrando os 15 gêneros com maior incidência na base trabalhada (vide notebook para mais detalhes):



Uma vez que a análise dos gêneros foi realizada, vamos aproveitar o momento para tratar a ausência da coluna. Serão criados os atributos ehComedia, ehDrama e ehDocumentario para substituir o atributo 'genres' (vide notebook).

País de origem: no dataset de junção encontramos 130 países de origem diferentes. Maiores representantes são EUA, Grão Bretanha e Japão. O Brasil está em 17º lugar, com 494 séries. Levantamos aqui uma hipótese: o número expressivo de

séries japonesas tem como 'network' o YouTube? Para verificar, atualizamos o notebook 4 para termos um atributo ehYouTube no dataframe e analisar. O resultado descartou a hipótese, uma vez que dos 4.997 registros de origem japonesa apenas 22 são do YouTube (mais detalhes no notebook 6).

Nº de episódios e nº de temporadas: Esses dois atributos parecem relacionados: quanto mais temporadas, mais episódios. Analisando o coeficiente de correlação no dataset vemos que não é alto: é de 0,43.

O fato de se ter mais temporadas apresenta um diferencial, pois uma temporada representa um novo ciclo contratual da série, o que é um indicativo de que a série está indo bem, é algo bem significativo. Vejamos o resumo de ocorrências desses dois atributos e vejamos um pouco sobre quais registros apresentam maiores números de episódios e temporadas no dataset (mais detalhes no notebook):

series_df['number_of_episodes'].value_counts().head(10)	series_df['number_of_seasons'].value_counts().head(10)
20.847526 160363 0.000000 28812 1.000000 6493 6.000000 3734 13.000000 2702 10.000000 2460 12.000000 2424 2.000000 2424 8.000000 2315 3.000000 2249	1.31833 159004 1.00000 39601 0.00000 27544 2.00000 6923 3.00000 3139 4.00000 1714 5.00000 1074 6.00000 708 7.00000 472 8.00000 406

Pelo resumo de ocorrências da coluna 'number_of_episodes' vemos que a muitas séries foi atribuído o valor zero de episódios. Zero é o segundo valor mais frequente, sendo o primeiro um valor não inteiro 20.847526 que é o resultado da estratégia de preenchimento de campos nulos que foi aplicada no notebook 4. No notebook recuperamos as séries com maior número de episódios: muitos são programas diários. Temos também novelas e shows de perguntas e respostas.

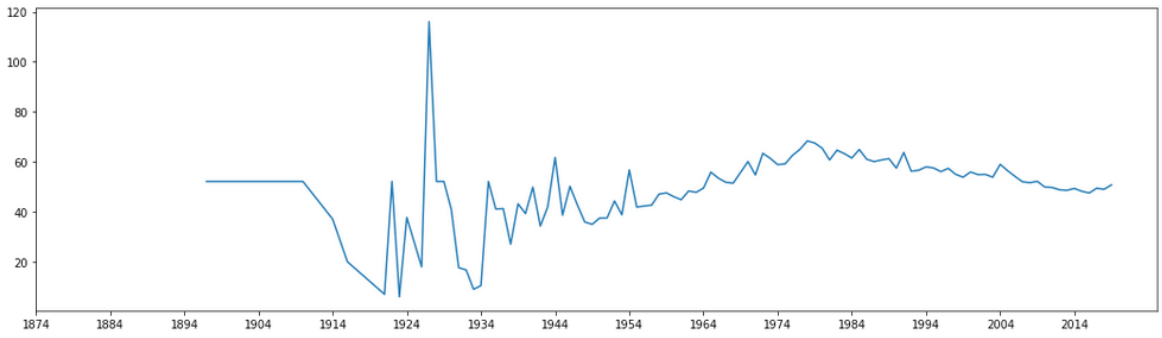
Pelo resumo de ocorrências da coluna 'number_of_seasons' vemos que a muitas séries foi atribuído o valor zero ou um de temporadas. Zero é o terceiro valor mais frequente, sendo o segundo o valor um e o primeiro um valor não inteiro

1.31833 que é o resultado da estratégia de preenchimento de campos nulos que foi aplicada no notebook 4. Exemplos das séries com mais temporadas:

- Fitil é uma popular série de curtas-metragens satíricos / comédias de televisão soviética / russa. Alguns dos episódios foram dirigidos a crianças. Cada edição continha os poucos segmentos curtos: documentário, fictício e animado.
- A Play School é um programa de televisão educacional premiado pela Australian Gold Logie para crianças produzido pela Australian Broadcasting Corporation. É o show infantil de maior duração na Austrália e o segundo show infantil de maior duração do mundo.

Tempo médio por episódio: o tempo médio mais recorrente é um tempo fracionado, obtido pela estratégia de preenchimento de campos nulos que foi aplicada no notebook 4. Em seguida temos séries de 30 e 60 minutos e depois de 45 e 25 minutos (vide notebook para mais detalhes).

Observemos se a duração média dos episódios das séries seguiu algum padrão ou apresenta alguma tendência ao longo do tempo:



É possível perceber que demorou um pouco para ocorrer a formação de um padrão na duração média dos episódios das séries, mas que por volta dos anos 60 esse padrão se definiu, entre os 40 e 60 minutos. Vimos que existem muitas séries com episódios curtos de 30 minutos, mas as séries com mais de 50 minutos são mais numerosas. O início do gráfico, principalmente entre 1924 e 1934, apresenta uma variação bem grande de tempo médio de episódio. Explorando mais o dataset (vide notebook) vimos que existem muitas séries com duração de 1 minuto!

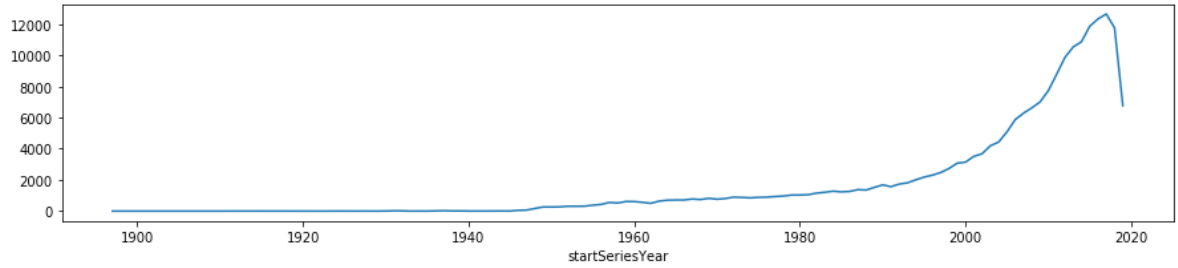
Exemplo de série de duração de 1 minuto: Minutos do Bicentenário foi uma série de curtos segmentos educacionais da televisão americana comemorando o bicentenário da Revolução Americana. *Snipets* era uma série de curtas-metragens de 30 segundos a um minuto para crianças.

Ao analisar as séries mais longas foi descoberto um problema de interpretação do campo. Por exemplo, no caso da série com tempo de episódio mais longo, o tempo de execução está 8.400 min, mas esse é referente aos 280 episódios da série e não a cada um. Consideramos a possibilidade de erro de interpretação do campo no momento da junção das bases do TMDB e IMDB. No entanto, analisando os casos obtidos pela análise do notebook, verificamos que o erro é proveniente da base de origem mesmo (IMDB, no caso). Por vezes o tempo cadastrado se refere a todos os episódios e por vezes se refere a um tempo por episódio. Estimamos o tamanho desse problema analisando quantos registros apresentam tempo médio superior a 4 horas e foi verificado que menos de 1% dos registros apresenta essa condição.

Línguas: Analisando o número de línguas que cada registro de série apresenta, vemos que o atributo em questão está vazio para grande maioria dos registros. São raras as séries com mais 3 línguas. A série com mais línguas registradas na base, 17 línguas, é "Total Drama: Revenge of the Island" que é a quarta temporada da franquia "Total Drama". É um desenho que faz uma paródia de reality shows. A segunda na lista é "Cold War": uma série de documentários televisivos de 24 episódios sobre a Guerra Fria que foi ao ar em 1998. Apresenta várias entrevistas e cenas dos eventos que moldaram as relações tensas entre a União Soviética e os Estados Unidos. O caráter global do evento tratado e o fato de ser um documentário repleto de entrevistas justifica a diversidade linguística da série.
Ao final dessa análise tratamos a substituição do atributo 'languages' pelo número de línguas que apresenta, evitando assim os valores vazios.

Anos de estreia das séries: O gráfico de distribuição de números de séries produzidas ao longo do tempo mostra a forte tendência de aumento na produção de séries, o que corrobora para a motivação desse trabalho (analisar uma forma de entretenimento atual, e crescente).

```
<matplotlib.axes._subplots.AxesSubplot at 0x1c653ee7cc0>
```



Analisando a lista de séries mais antigas (de 1897 a 1916) vemos programas diários, shows do Charlie Chaplin e um seriado francês (Judex). Vide *notebook* para detalhes.

Verificação de colunas categóricas: Analisamos as ocorrências de valores para os atributos status, type, titleType e isAdult:

```
In [38]: series_df['status'].value_counts()

Out[38]: Ended               61072
         Returning Series    20345
         Canceled             1104
         In Production         480
         Pilot                  85
         Name: status, dtype: int64
```

```
In [39]: series_df['type'].value_counts()

Out[39]: Scripted       72901
         Miniseries      3544
         Documentary     2977
         Reality         2577
         Talk Show        787
         News             151
         Video            149
         Name: type, dtype: int64
```

```
In [69]: series_df['titleType'].value_counts()

Out[69]: tvSeries        165839
         tvMiniSeries     26404
         Name: titleType, dtype: int64
```

```
In [70]: series_df['isAdult'].value_counts()

Out[70]: 0.0    191260
         1.0       983
         Name: isAdult, dtype: int64
```

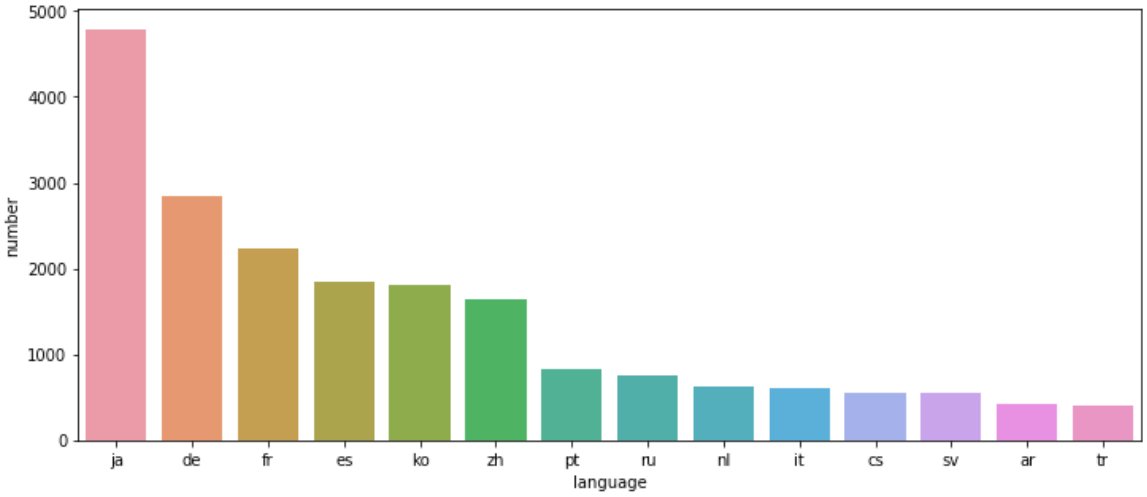
Observando a distribuição das séries em relação a esses atributos, é possível notar que:

- Como era de se esperar, temos mais séries finalizadas do que em andamento.
- Em relação ao tipo temos uma forte presença das séries roteirizadas, que contam uma história ao longo de temporadas que vão sendo desenvolvidas ao longo do tempo. Em seguida, séries no formato mini-série, com início e fim pré-determinados. Depois, séries de "reality" - shows de tv que têm suas temporadas, mas que não contam histórias. Temos também documentários, talk shows, news e video.

As colunas isAdult e titleType, provenientes do IMDB, apresentam 79.40% de preenchimento. Já type e status, provenientes do TMDB, apresentam 34.32% de preenchimento. Para futuras análises em que aplicaremos classificadores, vamos:

- preencher os valores nulos das colunas type, titleType e status com o valor 'unknown'
- excluir o atributo isAdult, pois este se mostra irrelevante para a análise em questão.

Língua original: Existem 88 línguas de origem diferentes no dataset. Como esperado, o inglês é fortemente dominante. O japonês assume a segunda posição (mais detalhes no notebook). O gráfico abaixo exclui o inglês para podermos perceber melhor a diferença de incidência entre as outras linguagens.



O número de séries japonesas realmente é preponderante (desconsiderando o inglês). Vemos dentre alguns exemplos dessas séries japonesas, séries de desenho e super-heróis, como por exemplo 'National Kid' e 'Ultraman' (mais exemplos no notebook).

4.2 Análise das bases de premiações

Número de categorias: vamos analisar o número de categorias de cada premiação ao longo do tempo para entender um pouco mais dos eventos Emmy e Golden Globe. Vide gráficos abaixo:

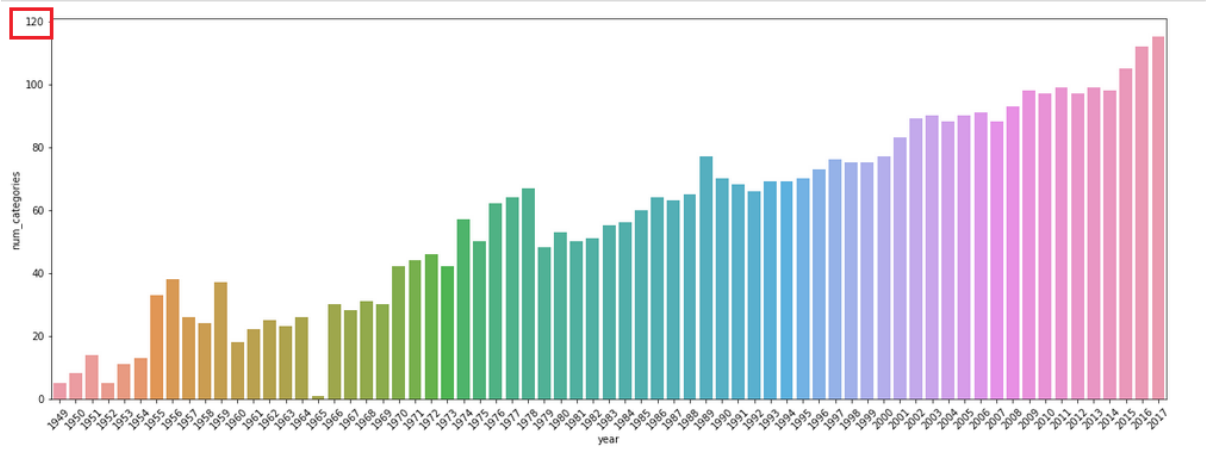


Figura 1: Categorias EMMY

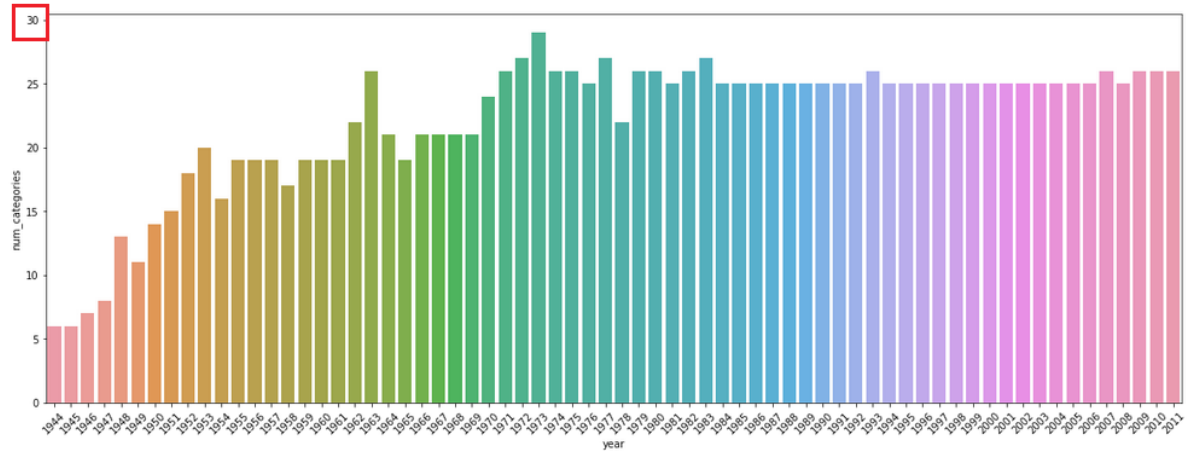


Figura 2: Categorias Golden Globe

Vemos que o Emmy apresenta um número máximo de categorias bem maior, quase 4 vezes maior, que o Golden Globe. Além disso:

- para o dataset do EMMY: as categorias aparecem desde o ano de 1949 até o ano de 2017 e, apesar de variarem consideravelmente em alguns anos, apresentam tendência de crescimento. 1965 é um ano que se destaca com o menor número de categorias. Pesquisando, encontramos a informação que nesse ano a estrutura da cerimônia bem diferente dos anos anteriores. As categorias foram simplificadas para que houvesse apenas quatro categorias principais (o ano anterior tinha 20 categorias principais). Como resultado, apenas cinco shows ganharam um prêmio. O novo formato foi descartado, e no ano seguinte o Emmy voltou ao formato tradicional.
- para o dataset do Golden Globe: as categorias aparecem no intervalo desde o ano de 1944 até o ano de 2011. Regra geral a tendência também é de crescimento do número de categorias mas com intervalos de estabilização.

Número de nomeações e premiações por série: por meio de agrupamento pelo nome da série (similar ao que foi feito também no notebook 5) obtivemos num mesmo dataframe os números de nomeações e premiações de uma série em cada um dos eventos (Emmy e Golden Globe).

- A série mais premiada do Emmy é '*Frasier*', com 103 nomeações e 37 premiações. '*Frasier*' é uma comédia americana transmitida na NBC por 11 temporadas, estreiando em 16 de setembro de 1993 e terminando em 13 de maio de 2004.
- A série mais nomeada do Emmy é '*Saturday Night Life'*, com 131 nomeações e 27 premiações. '*Saturday Night Live'* (abreviado como *SNL*) é um programa de televisão semanal de comédia que está no ar há mais de 3 décadas.
- A série mais premiada do Golden Globe é '*M*a*S*H*', com 26 nomeações e 9 premiações ou o programa 'The Carol Burnett Show', com 29 nomeações e 9 premiações.
    - *M*A*S*H* é uma premiada série de televisão americana. Foi exibida pela CBS entre setembro de 1970 e fevereiro de 1983.
    - 'The Carol Burnett Show' é um programa de comédia de variedade estrelado por Carol Burnett. Os episódios originais ocorreram de 1967 a

1978.

- A série mais nomeada do Golden Globe é 'Cheers', com 32 nomeações e 6 premiações. Cheers é a sitcom americana. A série foi transmitida pela primeira vez a 30 de Setembro, 1982, até 20 de Maio, 1993, sendo uma das séries de televisão mais longas, com 11 temporadas e 273 episódios.

Realizando a junção dos datasets agrupados por série, para as duas premiações, descobrimos que, considerando nomeações e premiações dos dois eventos, as séries mais nomeadas e mais premiadas acompanham os resultados obtidos quando analisando somente o Emmy (consequência do grande número de categorias desse evento):

- A série mais premiada no geral é '*Frasier*', com 127 nomeações e 39 premiações.
- A série mais nomeada no geral é '*Saturday Night Live'*, com 131 nomeações e 27 premiações (todas do Emmy).

Junção de dataset agrupado de premiações por série com dataset de séries: vamos unir os datasets de séries e o dataset de nomeações e premiações agrupados por séries (para o dois eventos – Emmy e Golden Globe). Com essa junção estamos restringindo os registros de nomeação e de premiações a séries que existam na base construída a partir do TMDB e IMDB. Foram feitos 3.731 relacionamentos de nomeações / premiações com as séries do dataset de séries (vide notebook para mais detalhes).
Como a junção das bases foi feita com base no valor exato do nome da série, podemos ter perdido algumas combinações de registros. Nesse trabalho, vamos optar por trabalhar com esses registros em que ocorre o casamento total do padrão. Vamos salvar a base para posterior trabalho de classificação.

Nome das séries – nuvem de palavras: para comparação, geramos a nuvem de palavras que ilustra a recorrência de termos nos títulos das séries em geral (base junção TMDB e IMDB) e dos títulos que tiveram nomeações no EMMY e no GOLDEN GLOBE.

Figura 3: Nuvem - título de séries em geral

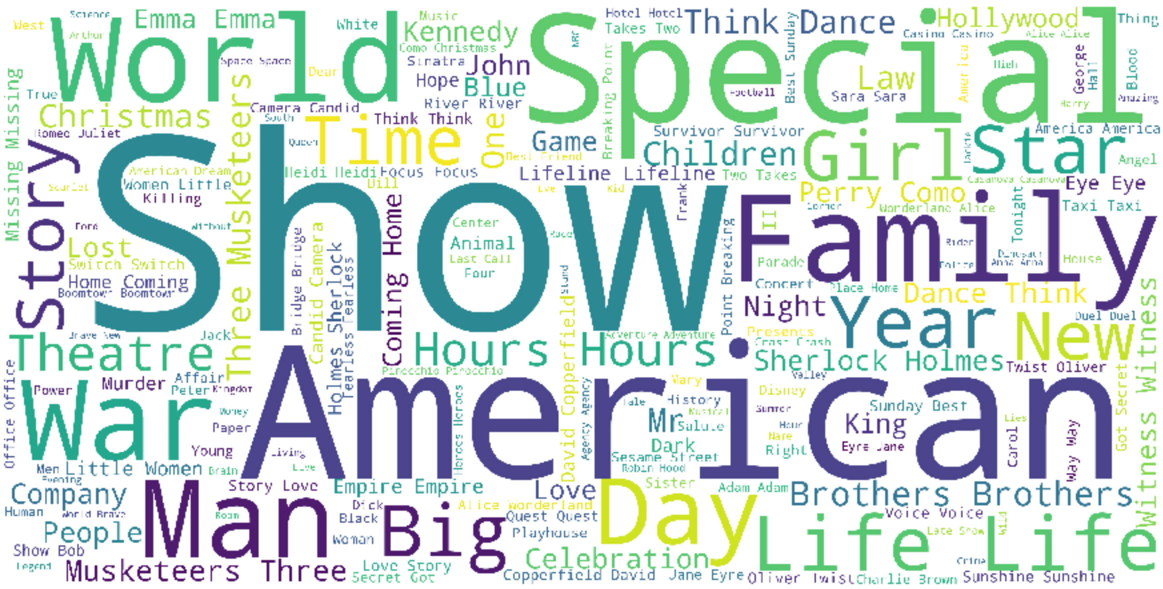


Figura 4: Nuvem - título de séries nomeadas no Emmy ou Golden Globe

A nuvem de palavras de séries em geral mostra termos bem recorrentes nos nomes de programas de tv (TV, Show, Live, Life, Love, New, World, etc). Já nuvem de palavras de séries que foram pelo menos nomeadas ao Emmy ou ao Golden Globe

também apresenta algumas dessas palavras em destaque (Show, World) mas apresenta forte recorrência nas palavras (American, Family, Special, War). Tais palavras estão presentes em títulos de séries nomeados nos eventos Emmy e Golden Globe. Nessa nuvem também é possível ler títulos inteiros de séries, como “Sherlock Holmes”.

5. Criação de Modelos de *Machine Learning*

Para acompanhar os processamentos aqui descritos, vide *Jupyter Notebook* **7 Classificação.**

Para explorar mais um pouco o *dataset* trabalhado até aqui, vamos aplicar alguns algoritmos de classificação sobre os dados, visando identificar quais fatores indicariam a nomeação de uma série no EMMY e ou Golden Globe.

Vamos usar o *dataset* previamente gerado que combina os dados de nomeações e premiações com os dados de séries do IMDB e do TMDB, mas removeremos antes as duplicidades por nome, pois ao efetuar a junção desses *datasets* toda vez que tivermos o mesmo nome de série com nomeação teremos a contabilização de nomeações (e talvez premiações) duplicadas. Por exemplo: entradas diferentes para o nome de '*Saturday Night Live*' (dos EUA e do Brasil) receberiam duplicadas as nomeações e premiações concedidas ao programa americano.

A classe que trataremos é binária, refletindo a nomeação ou não da série. Definindo as duas classes no *dataset* já sem as duplicidades, temos a seguinte representatividade em cada uma delas:

- Não nomeados: 211.765
- Nomeados: 2.237

Temos 2.237 séries nomeadas nos eventos de premiação. O restante das séries não foi mencionada nas premiações Golden Globe ou EMMY (211.765). Há um desbalanceamento entre a quantidade de séries que tiveram menção nas premiações e as que não tiveram.

No notebook, preparamos os atributos que serão usados na classificação, descartando os que não serão utilizados, preenchendo os últimos dados omissos e transformando alguns. Aplicamos diferentes classificadores para tentar resolver o problema (mais detalhes no *notebook*):

- *DummyClassifier*: aplicado para mostrar como uma classificação randômica se comporta em relação aos outros classificadores. Apresenta acurácia total alta, mas valores extremamente baixos para as métricas da classificação da classe 1:

```
#### DummyClassifier####
Acurácia (base de treinamento): 0.9789311978317884
Acurácia de previsão: 0.9791593654353871
[[41906   448]
 [  444     3]]
             precision    recall  f1-score   support

          0       0.99      0.99      0.99     42354
          1       0.01      0.01      0.01       447

avg / total       0.98      0.98      0.98     42801
```

- *DecisionTreeClassifier*: Apresenta valores bem melhores que o classificador *Dummy* no caso da classe 1, mas ainda muito baixos:

```
#### DecisionTreeClassifier####
Acurácia (base de treinamento): 0.9968224484670066
Acurácia de previsão: 0.9852106259199551
[[42082   272]
 [  361    86]]
             precision    recall  f1-score   support

          0       0.99      0.99      0.99     42354
          1       0.24      0.19      0.21       447

avg / total       0.98      0.99      0.98     42801
```

- *KneighborsClassifier*: Valor de precisão da classe 1 um pouco melhor do que o algoritmo anterior. Revocação um pouco pior.

```
#### KNeighborsClassifier####
Acurácia (base de treinamento): 0.9900759925467725
Acurácia de previsão: 0.9888787645148478
[[42293    61]
 [  415    32]]
             precision    recall  f1-score   support

          0       0.99      1.00      0.99     42354
          1       0.34      0.07      0.12       447

avg / total       0.98      0.99      0.99     42801
```

- LinearDiscriminantAnalysis: uma das melhores revocações encontradas para a classe 1 dentre os experimentos feitos.

```
#### LinearDiscriminantAnalysis####
Acurácia (base de treinamento): 0.9779440540651048
Acurácia de previsão: 0.9790659096750076
[[41740   614]
 [  282   165]]
             precision    recall  f1-score   support

          0       0.99      0.99      0.99     42354
          1       0.21      0.37      0.27       447
```

```
avg / total       0.99      0.98      0.98     42801
```

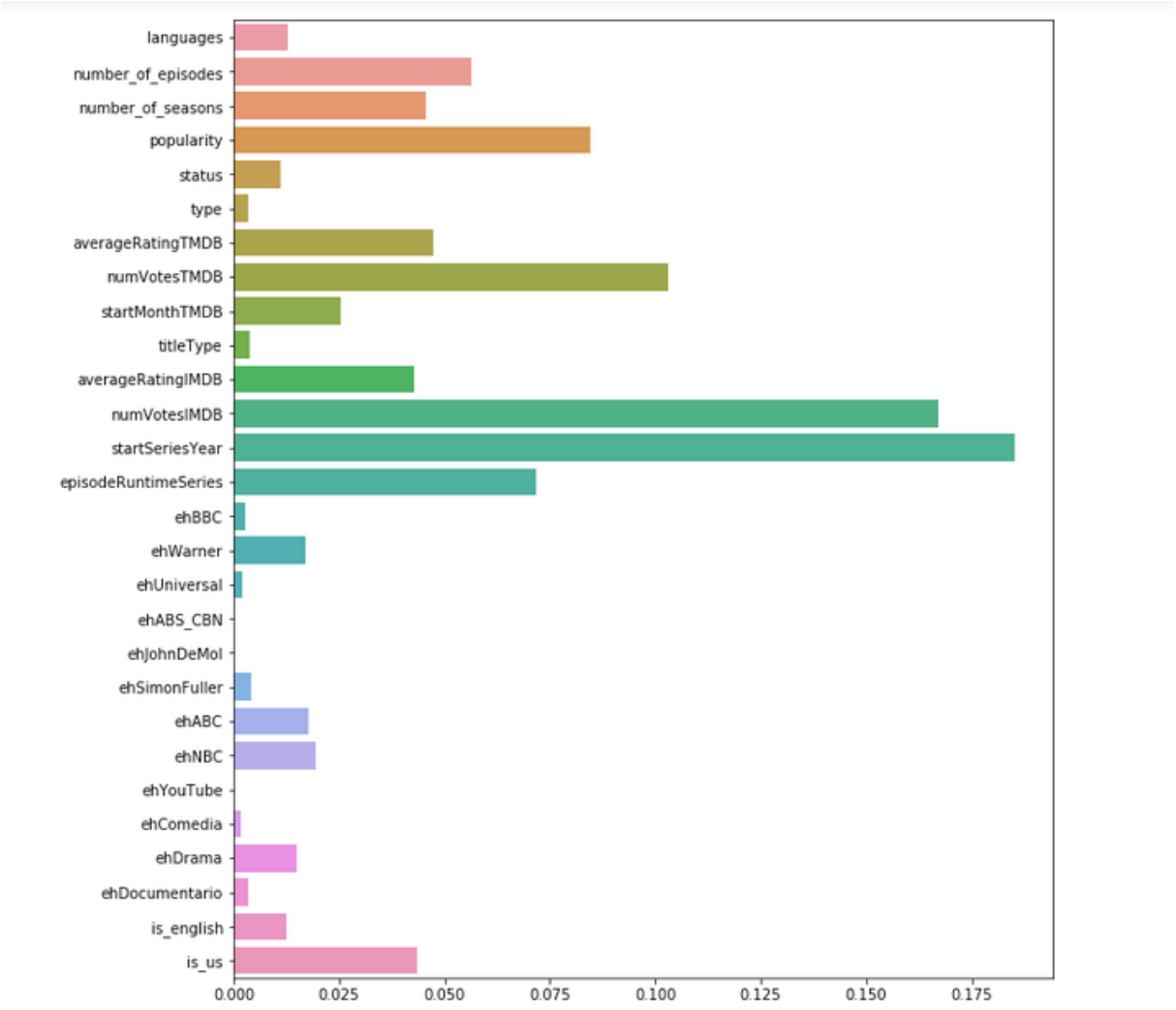
- GaussianNB: melhor valor de revocação obtido para classe 1, mas ainda baixo.

```
#### GaussianNB####
Acurácia (base de treinamento): 0.9732828663383976
Acurácia de previsão: 0.9737622952734749
[[41506   848]
 [  275   172]]
             precision    recall  f1-score   support

          0       0.99      0.98      0.99     42354
          1       0.17      0.38      0.23       447

avg / total       0.98      0.97      0.98     42801
```

- GradientBoostingClassifier: melhores valores de precisão obtidos para a classe 1, acima de 0.5. Revocação da classe 1 baixa. Realizamos algumas investidas para tentar melhorar esse resultado:

```
#### GradientBoostingClassifier####
Acurácia (base de treinamento): 0.9907827641193685
Acurácia de previsão: 0.9896731384780729
[[42298    56]
 [  386    61]]
             precision    recall  f1-score   support

          0       0.99      1.00      0.99     42354
          1       0.52      0.14      0.22       447

avg / total       0.99      0.99      0.99     42801
```

Foi feito um gráfico (detalhes no notebook) mostrando a importância dos atributos utilizados para classificação do classificador que teve maior precisão na definição da classe alvo:



Reduzimos os atributos, excluindo os atributos de baixa importância e também aqueles que poderiam apresentar alguma redundância como:

- is_english – série de língua inglesa - redundante com is_us – série estadunidense

- number_of_episodes, redundante com number_of_seasons

Ao final, o conjunto de atributos utilizados para a classificação foi:

number_of_seasons: nulos foram preenchidos com o valor médio (mostrado em seções anteriores)	popularity: nulos foram preenchidos com o valor zero (mostrado em seções anteriores)
status: nulos foram preenchidos com o valor 'unknown' (mostrado em seções anteriores)	averageRatingTMDB: nulos foram preenchidos com o valor zero (mostrado em seções anteriores)

StartSeriesYear: nulos foram preenchidos com o valor médio.	numVotesTMDB: nulos foram preenchidos com o valor zero (mostrado em seções anteriores)
EpisodeRuntimeSeries: nulos foram preenchidos com o valor médio (mostrado em seções anteriores)	averageRatingIMDB: nulos foram preenchidos com o valor zero (mostrado em seções anteriores)
EhWarner: valor booleano derivado em seções anteriores	numVotesIMDB: nulos foram preenchidos com o valor zero (mostrado em seções anteriores)
ehABC: valor booleano derivado em seções anteriores	ehNBC: valor booleano derivado em seções anteriores
ehDrama: valor booleano derivado em seções anteriores	is_us: valor booleano derivado do campo 'origin_country', indicando se a série é estadunidense ou não.

Além disso, foi feita uma pesquisa sobre os melhores parâmetros para inicialmente serem aplicados junto a esse classificador[10]. Os parâmetros indicados abaixo foram utilizados:

- **min_samples_split:** A indicação do artigo é que deve ser entre 0,5-1% do total de registros. Como é um problema de classes desbalanceadas, sugere usar um valor pequeno no intervalo. Optamos por 1200 (0,5%).
- **min_samples_leaf:** Parâmetro para previne *overfitting.* Valores pequenos são recomendados para classes desbalanceadas.
- **max_depth:** Deve ser escolhido entre 5 e 8, dependendo do número de observações.
- **subsample = 0.8 :** Valor comumente usado.

Foram realizadas algumas variações desses parâmetros, e o melhor valor de precisão encontrado para a classe 1 foi de 0,58 e de revocação 0.13 (vide notebook para mais detalhes).

```
#### GradientBoostingClassifier####
Acurácia (base de treinamento): 0.9908470160805135
Acurácia de previsão: 0.9899301418191164
[[42314    40]
```

[10]https://www.analyticsvidhya.com/blog/2016/02/complete-guide-parameter-tuning-gradient-boosting-gbm-python/

```
[ 391   56]]
             precision    recall  f1-score   support

          0       0.99      1.00      0.99     42354
          1       0.58      0.13      0.21       447

avg / total       0.99      0.99      0.99     42801
```

Conclui-se que a acurácia (número de previsões corretas / número de previsões) é alta para todos os classificadores, inclusive o *DummyClassifier*. Isso se dá porque o dataset é desbalanceado, é fácil prever uma das classes pois ela é muito recorrente. Temos apenas 1,05% de classe 1. O restante é classe 0. Quando avaliamos as métricas POR CLASSE vemos claramente que o desafio está em classificar corretamente a classe 1 (nomeados).

Em relação à classificação da classe 1:

- A precisão (número de *true positives* / (número de *true positives* + número de *false positives*)) é baixa. Ou seja, é baixa a porcentagem de acertos da avaliação positiva da nomeação dentre todas as avaliações como ocorrência positiva. Dentre os avaliados, o classificador com melhor precisão é o *GradientBoostingClassifier*. Melhoramos um pouco os resultados iniciais desse classificador. Vimos pelo gráfico de importância de '*features*' que os atributos relativos a avaliação das séries, principalmente os relativos ao tamanho da votação, apresentaram a maior significância quando da definição das classes pelo algoritmo.
- A revocação (número de *true positives* / (número de *true positives* + número de *false negatives*)) é baixa. Ou seja, é baixa a porcentagem de acertos da avaliação positiva da nomeação dentre todas as ocorrências REAIS de positivos. Dentre os avaliados, o classificador com melhor revocação foi o GaussianNB (0,25).

6. Apresentação dos Resultados

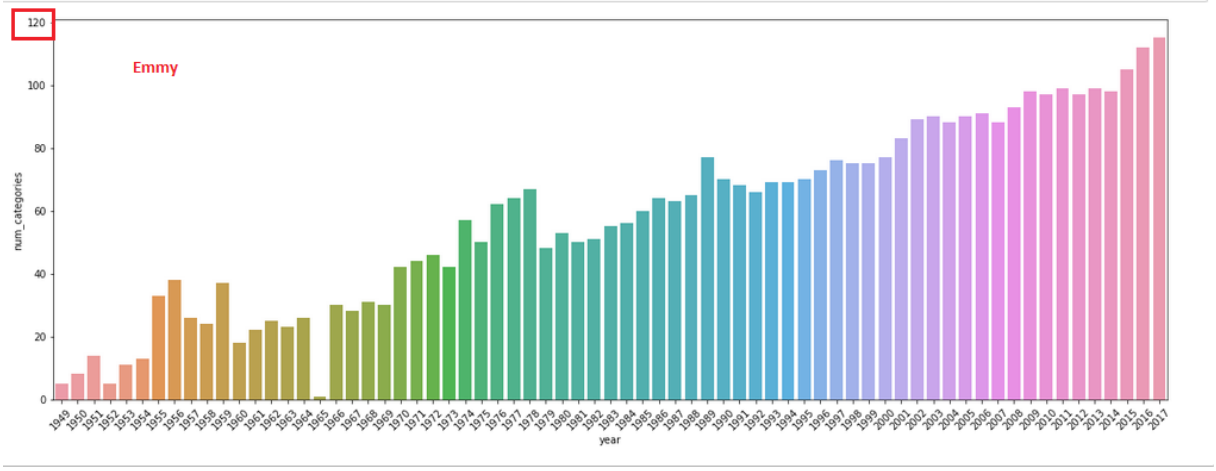
Conforme orientação da pós-graduação, primeiramente será apresentado o preenchimento do *workflow* motivador desse trabalho seguindo modelo de *canvas* proposto por Vasandani (pode ser visto aqui).

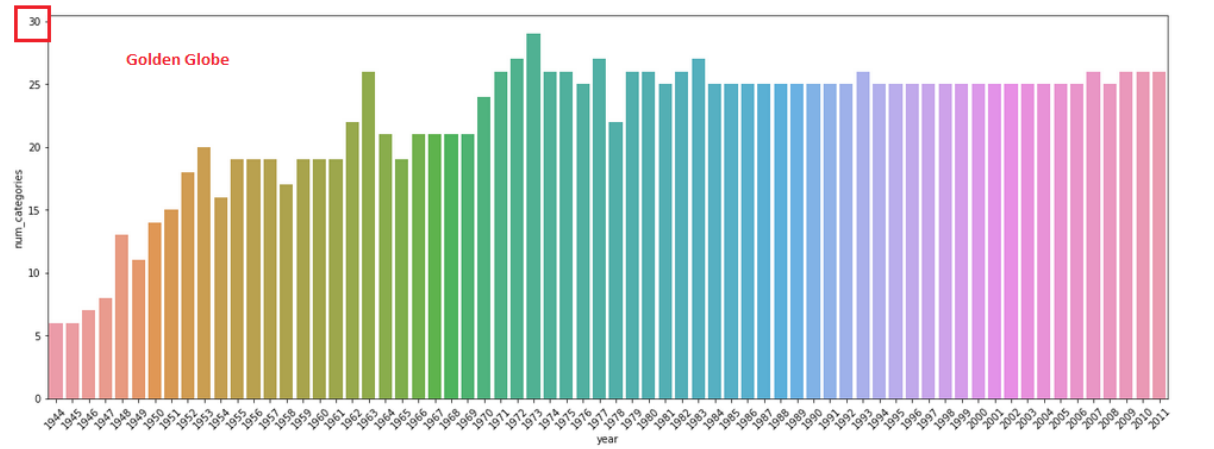
Título: **Séries: seus atributos, avaliações, nomeações e premiações**		
- Definição do problema:Investigar *datasets* referentes a séries de TV e suas premiações para descobrir informações interessantes sobre o tema.	**Resultados e previsões:** Objetiva-se a verificação de especificidades de cada uma das bases e a previsão de nomeação de uma série no EMMY e no GOLDEN GLOBE a partir de atributos das bases do TMDB e IMDB. Hipótese: Boas avaliações e popularidade nessas bases estão relacionadas à nomeação nesses eventos.	**Aquisição de dados:** Os dados foram obtidos de 4 origens: coleta de dados no TMDB e *datasets* disponibilizados pelo IMDB, AggData (dados Golden Globe) e Kaggle (dados Emmy). Questão existente: datasets das premiações são referentes a períodos distintos (Golden Globe vai até 2011 e Emmy até 2017)
Modelagem: São realizadas várias análises exploratórias dos dados em Python até a obtenção de um dataset a ser usado para aplicação de diferentes classificadores da biblioteca sklearn.	**Avaliação do modelo:** Uma vez que temos classes desbalanceadas, os classificadores são avaliados por meio da matriz de confusão e do relatório de classificação com foco nas métricas por classe.	**Preparação dos dados:** Dados ausentes e duplicados foram tratados antes e depois de junções entre os datasets. Alguns dados também precisaram ser agrupados ou corrigidos.

A exploração dos dados realizada possibilitou caracterizar melhor o tipo de informação disponível nas bases do IMDB e TMDB e entender melhor as premiações Emmy e Golden Globe:

- A base do IMDB é bem mais popular e suas votações bem mais expressivas. O IMDB é uma base divulgada mundialmente, apresentando conteúdo e avaliações que ultrapassam o comum americano e englobam produções extremamente independentes e variadas, tais quais as presentes nos canais do Youtube. Já o TMDB é uma base mais focada em conteúdo americano e produtoras tradicionais.

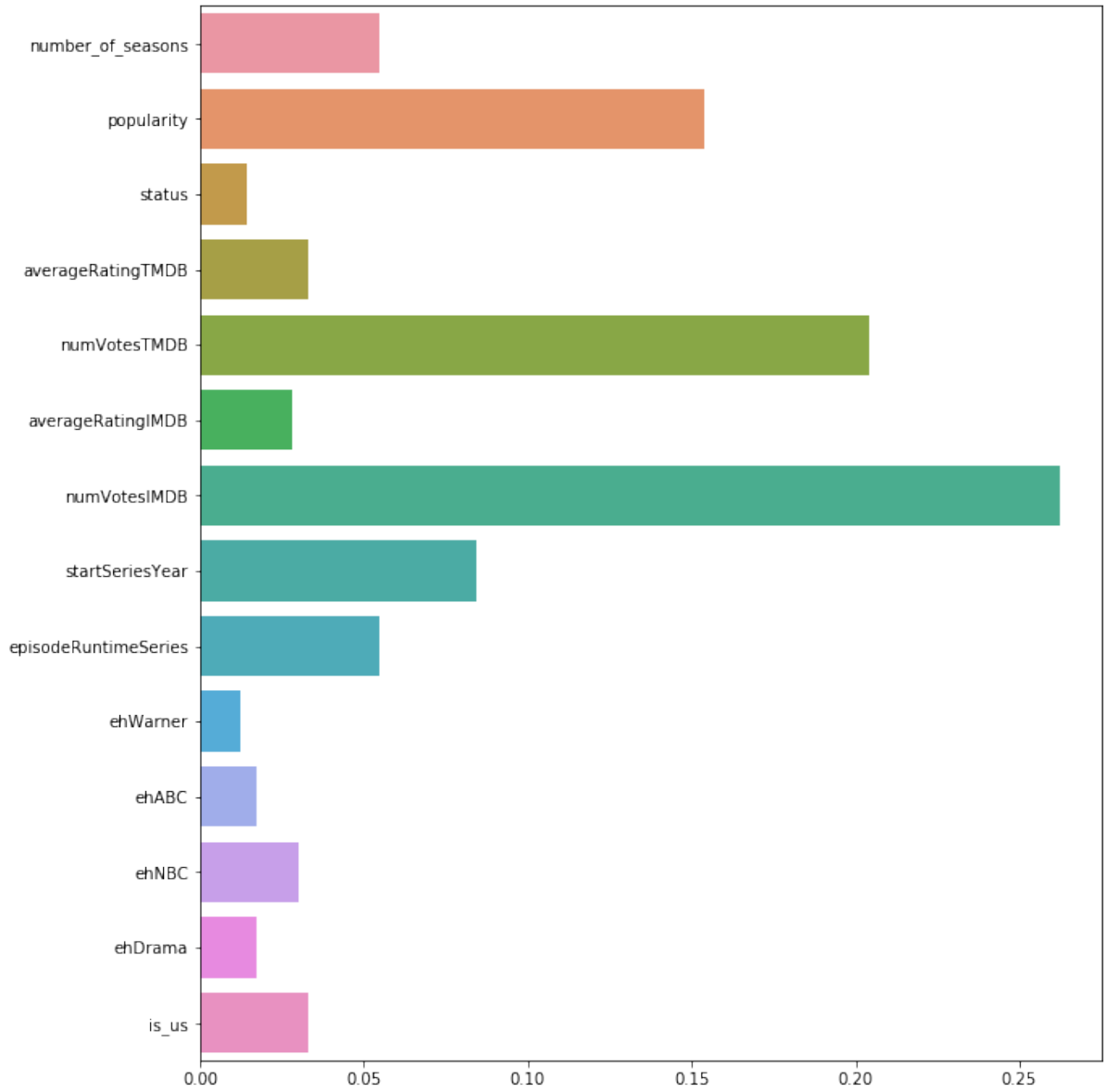
- A base de dados do Emmy apresenta maior peso sobre as análises, pois a cada ano apresenta um número bem maior de categorias de premiação que o Golden Globe, conforme gráficos abaixo já apresentados anteriormente:





A combinação das bases propiciou análises interessantes sobre as séries mais nomeadas e premiadas, pois foi possível listar e identificar os tipos de dados dessas séries. O *dataset* final construído com dados das séries e suas nomeações e premiações, sem dados ausentes ou duplicados, pode ser submetido a um conjunto de algoritmos classificadores.

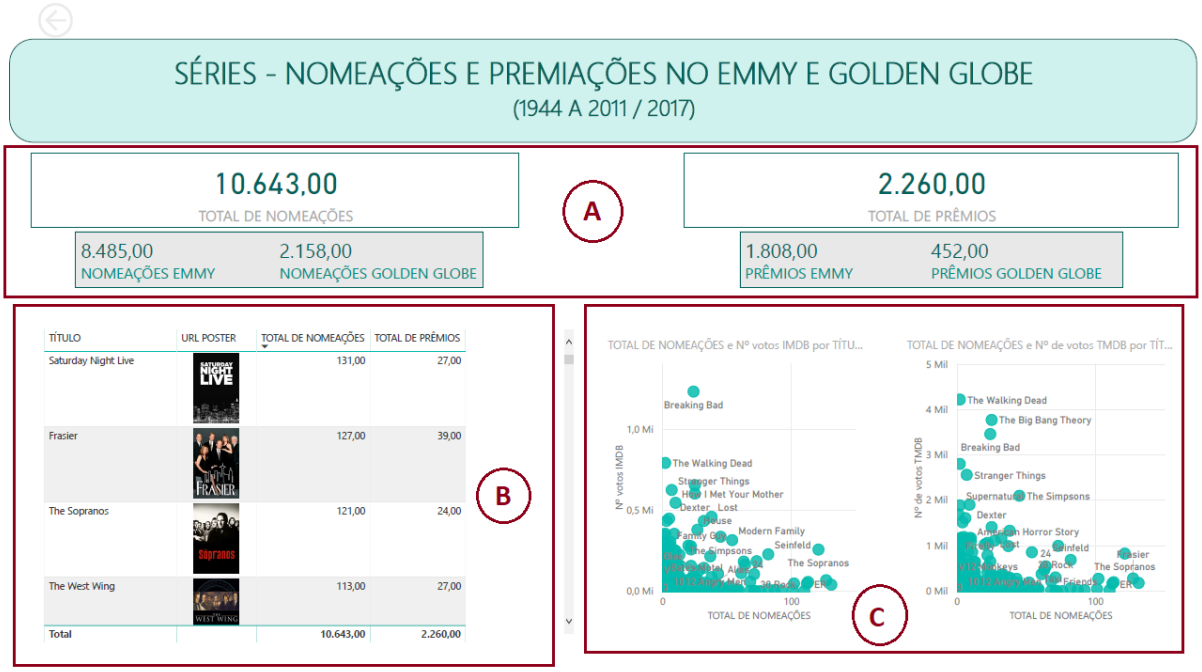
A classificação mais precisa das séries nomeadas, realizada pelo GradientBoostingClassifier, apresentou como atributos mais importantes aqueles referentes ao número de votações recebidos pela série nas bases IMDB e TMDB (vide gráfico a seguir). Em seguida, outra métrica, a de popularidade no TMDB e só depois, com grande diferença, o ano de início da série. Uma hipótese de raciocínio é que essa última *feature* importante poderia ser derivada do aumento do número de categorias de premiação ao longo do tempo: mais categorias, mais nomeações. Logo, o ano da série influenciaria na sua possibilidade de premiação.



As métricas por classe de precisão e revocação não indicaram valores altos. É difícil classificar quando uma classe é nomeada, pois o *dataset* é desbalanceado, apresentando apenas 1,05% de séries nomeadas.

Foram produzidos um dashboard e um infográfico para apresentação dos resultados:

- Dashboard: A seguir são apresentadas imagem e explicação do dashboard que pode ser baixado no repositório do trabalho.



A) Totalizadores de quantidade de nomeações e premiações das séries: de uma forma geral, nos retângulos superiores e por evento (Emmy ou Golden Globe) nos retângulos mais abaixo.

B) Listagem de título, pôster, total de nomeações e total de prêmios das séries que receberam maior número de nomeações. Para visualizar a discriminação por evento de premiação basta selecionar a série e observar os totalizadores do quadrante A.

C) Foram feitos gráficos de dispersão que relacionam o número de votos (do IMDB no 1º gráfico e do TMDB no 2º gráfico) e o total de nomeações para cada título de série. Curiosamente, a relação entre as duas variáveis tem tendência mais inversa: muitos títulos com muita votação não foram nomeados ou tiveram poucas nomeações e títulos com muitas nomeações foram pouco votados. Para visualizar as nomeações e premiações de um ponto no gráfico (que apresenta identificação do título) basta clicar sobre ele e observar os totalizadores do quadrante A.

Por fim é apresentado o infográfico que resume o realizado nesse trabalho:

RESUMINDO O TCC SOBRE SÉRIES

COLETA E OBTENÇÃO DE DADOS

Dados foram coletados ou obtidos de 4 fontes diferentes:
- TMDB
- IMDB
- AggData (Golden Globe)
- Kaggle (Emmy)

TRATAMENTO E PROCESSAMENTO DE DADOS

As bases obtidas foram tratadas e processadas de forma individual e também em conjunto. Ausência de valores e duplicidades em registros foram detectados, analisados e mitigados via estratégias diferenciadas, avaliadas para cada caso.

ANÁLISE DE DADOS

Os datasets de séries e premiações foram analisados, melhor compreendidos e foram encontradas informações interessantes que caracterizam as bases.

APLICAÇÃO DE CLASSIFICADORES

Um dataset foi preparado para a aplicação de uma série de classificadores em busca da definição das séries que poderiam ser nomeadas às premiações. O resultado das classificações foi avaliados via métrica por classe, uma vez que as classes em avaliação são desbalanceadas.

RESULTADOS

A exploração dos dados possibilitou caracterizar melhor o tipo de informação disponível nas bases do IMDB e do TMDB. Também foi possível verificar as diferenças significativas entre as premiações Emmy e Golden Globe. A combinação das bases propiciou análises interessantes sobre as séries mais nomeadas e premiadas.

CONCLUSÃO

No trabalho foram aplicados vários dos conhecimentos adquiridos na pós-graduação para obter, tratar, processar e explorar analiticamente os datasets, além de metodologias de estruturação do trabalho e apresentação de resultados.

FONTE: TCC DA BÁRBARA

7. Links

O link para o repositório desse trabalho é o seguinte:
Repositório TCC

Abaixo a descrição dos arquivos e diretórios da raiz do repositório indicado:

- **video_apresentacao**: vídeo com apresentação de 5 minutos
- **bases_originais**: bases coletadas (TMDB) ou obtidas (IMDB, EMMY e GOLDEN GLOBE) para a realização desse trabalho.
- **jupiterNotebooks**: notebooks com os diversos processamentos que são referenciados no texto do relatório para acompanhamento mais completo dos tratamentos e análises.
- **bases_tratadas**: bases intermediárias ou finais que foram sendo geradas pelos processamentos realizados nos notebooks.
- **scripts**: scripts desenvolvidos para realizar a coleta de dados na base do TMDB. Para arquivos auxiliares (lista de identificadores) e estrutura de diretórios esperada para execução desses scripts: repositorio\bases_originais\TMDB
- **dashboard***:* arquivos do *dashboard* gerado para esse trabalho
- **TCC INFO**: arquivo do infográfico desenvolvido para apresentar o trabalho
- **RelatorioFinal.pdf**: o presente relatório de trabalho.

i

SÉRIE DE TELEVISÃO. In: WIKIPÉDIA, a enciclopédia livre. Flórida: Wikimedia Foundation, 2019. Disponível em: <https://pt.wikipedia.org/w/index.php?title=S%C3%A9rie_de_televis%C3%A3o&oldid=55764866>. Acesso em: 18 jul. 2019.