Presultados prueba técnica

Data Enginner: Juan David

Contenido

- Revisión response de Api
- Revisión paso de json → DataFrame, teniendo integridad en data
- Revisión profit de DataFrame
- Generación de modelo de Datos
- Posibles preguntas a responder en Base de datos construida

Response base

```
[
    {
        "id": 2730586,
        "url": "https://www.tvmaze.com/episodes/2730586/neznost-2x01-seria-1",
        "name": "\u0421\u0435\u0440\u0438\u044f 1",
        "season": 2,
        "number": 1,
        "type": "regular",
        "airdate": "2024-01-01",
        "airtime": "",
        "airstamp": "2024-01-01T00:00:00+00:00",
        "runtime": 23,
        "rating": {
            "average": null
        },
        "image": null,
        "summary": null,
        "_links": {
            "self": {
                "href": "https://api.tvmaze.com/episodes/2730586"
```

```
    },
    "_embedded": {
        "show": {
            "id": 51908,
            "url": "https://www.tvmaze.com/shows/51908/neznost",
            "name": "\u041d\u0435\u0436\u043d\u043e\u0441\u0442\u044c",
            "type": "Scripted",
            "language": "Russian",
            "genres": [
                "Drama",
                "Comedy",
                "Romance"
            ],
            "status": "Ended",
            "runtime": null,
            "averageRuntime": 19,
            "premiered": "2020-11-12",
            "ended": "2024-01-01",
```

Lista de diccionarios anidados, json, donde cada id representa un tv show y sus características asociadas, particularmente el _embedded, contiene la información del show.

Revisión paso de json → DataFrame, teniendo integridad en data

- Normalizando o aplanando el json, podemos reconstruir un DataFrame con todos los show de tv y sus propiedades asociadas y las siguientes particularidades:

Overview | Alerts 55 | Reproduction

Dataset statistics

Number of variables	66
Number of observations	4904
Missing cells	129164
Missing cells (%)	39.9%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	17.2 MiB
Average record size in memory	3.6 KiB

Variable types

Numeric	15
URL	13
Text	12
Categorical	12
DateTime	6
Unsupported	8

Particularidades

- Variables categóricas con poca o nula variabilidad: type, season, _embedded.show.image.original, _embedded.show._links.self.href,
- _embedded.show.webChannel,…..
- Variables almamente correlacionadas que indican la misma información con dos nombres diferentres.

Modelo de datos en base al tipo de preguntas de “Negocio”

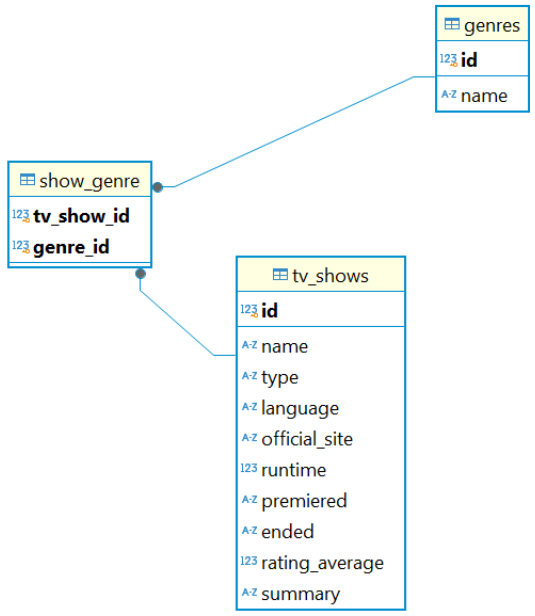
- Con el profit podemos depurar algunas columnas redundades o innecesarias y a la vez poder construir un modelo de datos que permita fácilmente responder interrogantes de dos objetos claros que se encuentra en el dataframe: show y géneros.

- Los show de tv pueden tener varios géneros asociados o no tener ninguno, esto implica una relación de uno a muchos lo cual nos permite proponer el siguiente modelo de datos que garantice la integridad

Modelo de datos

- Donde show_genre es una tabla puente que permite

Mantener las relaciones entre un show y diferentes géneros



Preguntas a responder con Python o Sql desde el modelo o el dataframe original

- Runtime promedio (averageRuntime).
- Conteo de shows de tv por género.
- Listar los dominios únicos (web) del sitio oficial de los show
- Ranking promedio o show con promedio más alto
- Show con mayor cantidad de géneros o show para un publico general.
- Ver si hay correlación entre la cantidad de géneros que tiene un show y su puntuación promedio.

- …..