Reto 1

Visualización de información televisiva para la elección de contenidos

Grupo 9
Santiago Gamboa

Índice

Propuesta

Mediante el uso del análisis de datos se quiere desarrollar una herramienta de visualización de las parrillas ofrecidas por distintas empresas televisivas al público y algunas características de estos contenidos, con el fin de que el usuario pueda seleccionar los contenidos audiovisuales que desea consumir con mayor facilidad.

Desarrollo de la solución

Análisis de los datos

En este paso se lleva a cabo un análisis del conjunto de datos utilizado para realizar correcciones y ajustes

- Se hace tratamiento de valores N/A, principalmente en el campo de lenguas nativas.
- Se realizan correcciones en los nombres de las empresas del conjunto de datos, las cuales contienen errores de escritura.
- Para el campo "TIPO" se corrigen errores de escritura y se realiza una generalización debido a la gran cantidad de valores únicos de este campo.

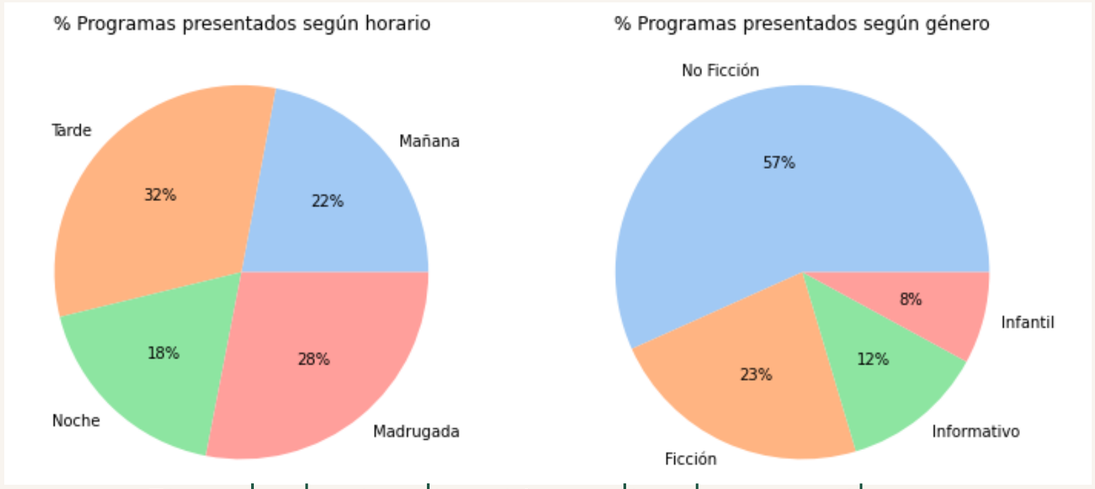
Transformación de datos

Teniendo en cuenta la hora de inicio de los programas televisivos se definen cuatro horarios en los cuales se categorizan los datos:

1. Madrugada (00:00 a. m. - 07:30 a. m.)
2. Mañana (07:30 a. m. - 12:30 p. m.)
3. Tarde (12:30 p. m. - 07:30 p. m.)
4. Noche (07:30 p. m. - 11:59 p. m.)

Exportación de datos y creación de gráficos

Una vez se ha realizado todo el proceso de limpieza general de datos, se inicia el proceso de exportación de datos para ser usados en la herramienta de Power BI. Para esto se realiza un último paso donde se eliminan caracteres como ';' y '\n' que causan errores en la exportación a formatos como csv.



Ejemplo de visualización realizado con python.

Buenas prácticas

Se hace tratamiento de valores nulos y de errores de tipografía en los datos

Se prioriza la calidad del entregable hacia el usuario final por lo que se hacen correcciones que permitan dar un mayor entendimiento de la información por parte de los stakeholders.

```
c = a.where(a<2800).dropna()

df.loc[df["TIPO"].str.contains('|'.join(c.index.to_list()), regex=True), "TIPO"] = 'OTRO'
```

Se contempla el uso de NLP como alternativa para la corrección de errores de escritura

```
for row in df.TIPO.value_counts().index.to_list():
    temp = [(jaccard_distance(set(ngrams(row, 3)), set(ngrams(w, 3))),w) for w in correct_spellings if w[0]==row[0]]
    s.append(sorted(temp, key = lambda val:val[0])[0][1])
```

Resultados

Clickea el link para ver los resultados!

Muchas gracias!