P03

June 6, 2019

1 Reporte de práctica 3: Estadística descriptiva básica

1.1 Objetivos

En esta práctica se terminan de limpiar los datos de la práctica anterior, también se realiza la categorización de los datos algunos conteos, promedios y correlaciones.

1.2 Limpieza

Revisando los CSV de la práctica anterior, se pudo observar que las diferencias ortográficas en los nombres de las columnas me generaban columnas duplicadas y además muchos espacios vacíos, se busca investigar cómo renombrar las columnas de un DataFrame.

Primero necesitábamos obtener los nombres de las columnas cargadas al DataFrame, y para eso ocupamos el código:

```
In [ ]: list(df2015)
        list(df2016)
        list(df2017)
        list(df2018)
```

Cada línea nos arroja una lista de los nombres de las columnas del DataFrame. Para no tener que unir las columnas, realicé los pasos de la práctica anterior hasta antes de la combinación de los datos, para asegurarse de que todas las columnas se llamen igual y evitar errores. En todos los DataFrame tenemos seis columnas de información que filtramos en la práctica dos, en todos los DataFrame están en el mismo orden, gracias a esto, podemos usar la función set_axis.

```
In [ ]: df2015.set_axis(['Año','Categoría','País','Género','£Cómo se enteró?','Referencia Dispo
        df2016[0].set_axis(['Año','Categoría','País','Género','£Cómo se enteró?','Referencia Di
        df2016[1].set_axis(['Año','Categoría','País','Género','£Cómo se enteró?','Referencia Di
        df2016[2].set_axis(['Año','Categoría','País','Género','£Cómo se enteró?','Referencia Di
        df2016[3].set_axis(['Año','Categoría','País','Género','£Cómo se enteró?','Referencia Di
        df2016[4].set_axis(['Año','Categoría','País','Género','£Cómo se enteró?','Referencia Di
        df2017[0].set_axis(['Año','Categoría','País','Género','£Cómo se enteró?','Referencia Di
        df2017[1].set_axis(['Año','Categoría','País','Género','£Cómo se enteró?','Referencia Di
        df2017[2].set_axis(['Año','Categoría','País','Género','£Cómo se enteró?','Referencia Di
        df2017[3].set_axis(['Año','Categoría','País','Género','£Cómo se enteró?','Referencia Di
        df2018.set_axis(['Año','Categoría','País','Género','£Cómo se enteró?','Referencia Dispo
        df2018mx.set_axis(['Año','Categoría','País','Género','£Cómo se enteró?','Referencia Dis
```

Ahora podemos proceder a combinar los conjuntos de datos sin tener columnas duplicadas y asegurándonos de que las columnas se llaman igual. Después de usar las funciones para combinarlos y exportar a csv, guardé toda la información en el archivo "datosLimpiosCine.csv", y ahora sabemos que tenemos 2735 registros en total.

1.2.1 Corrección de Nombres de Países

Al aplicar la función unique() en los nombres de los países, nos damos cuenta que escribieron los nombres en Inglés y Español, además de escribir con puntos y espacios donde no van. Como los datos son muy variados aplique el orden alfabético al array resultado y obtuve un error ya que en el año 2017 ponían en la columna País "Si" si era colombiano y "No" si no lo era pero como no especificaron el país, puse "Internacional", también a la colaboración con Colombia.

```
In [ ]: array(['Colombia', 'Irán', 'Georgia/Colombia', 'España', 'Venezuela',
                'México', 'Ecuador', 'Francia', 'Argentina', 'Perú',
                'Estados Unidos', nan, 'Canadá', 'Honduras', 'Brasil', 'Cuba',
                'Uruguay', 'Alemania', 'COLOMBIA', 'colombia', 'bogota',
                'colombia y mexico', 'Bogota', 'Colombia.', 'Marruecos',
                'COLO MBIA', 'República de Colombia', 'colombia ,',
                'Colombia, Brasil y Panamá', 'Republica Dominicana', 'Colomba',
                'ÂtColombia', 'Venezuela/Colombia', 'francia', 'colomBIA',
                'Brazil', 'Belgium', 'Canada', 'Colombia - China',
                'Colombia - Estados Unidos', 'Antioquia,', 'Si', 'No', '1',
                'Afganistán', 'Comoras', 'Afganistan', 'Spain', 'Mexico',
                'United States of America', 'Comoros'], dtype=object)
```

Resultados después de la limpieza:

```
In [ ]: array(['Colombia', 'Irán', 'España', 'Venezuela', 'México', 'Ecuador',
                'Francia', 'Argentina', 'Perú', 'Estados Unidos', nan, 'Canadá',
                'Honduras', 'Brasil', 'Cuba', 'Uruguay', 'Alemania',
                'Internacional', 'Marruecos', 'Republica Dominicana', 'Bélgica',
                'Afganistán', 'Comoras'], dtype=object)
```

Todas estas modificaciones las realicé en una columna nueva basada en "País", llamada "Países", para no arruinar la información y tener que volver a empezar si me equivocaba en algo. También limpiamos las categorías y pasamos de esto:

```
In [ ]: array(['Aficionado', 'Profesional', 'Infantil', nan, 'Juvenil', 'SmarTIC',
                'AFICIONADO', 'CRONICAS', 'JUVENIL', 'FAMILIAR',
                'SMARTIC INCLUYENTE', 'PROFESIONAL', 'HORROR', 'HUMOR'],
               dtype=object)
```

Debido a que en México confundieron el concepto de categoría con género del corto, reemplacé los valores por una categoría llamada "nan" que es mejor que inventar una categoría. Y pasamos a esto:

```
In [ ]: array(['Aficionado', 'Profesional', 'Infantil', nan, 'Juvenil', 'SmarTIC'],
               dtype=object)
```

Aprovechando que son los campos con menos errores, también limpiamos el campo de £Cómo se enteró? y pasamos de esto:

```
In [ ]: array(['Convocatoria', 'tv', 'internet', 'amigo', 'Amigo', 'Internet',
                'redes', nan, 'prensa', 'Prensa', 'TV', 'Redes', 'radio', 'Radio',
                'blanco', 'MSM', 'Redes sociales', 'Un(a) amigo(a) me contó',
                'Televisión', 'TelevisiÃşn', 'Un(a) amigo(a) me contÃş',
                'Mensaje de texto', 'Otra'], dtype=object)
```

A esto:

```
In [ ]: array(['Convocatoria', 'Televisión', 'Internet', 'Amigo',
                'Redes Sociales', nan, 'Prensa', 'Radio', 'Otra',
                'Mensaje de texto'], dtype=object)
```

Toda la información corregida la realice usando el comando replace con el siguiente formato:

```
In [ ]: dataframe['columnatemporal'] = dataframe.NombreColumna.replace('Ruido','DatoCoherente'
```

Como se podrá observar en los arreglos iniciales y como terminaron eliminando ruido, la cantidad de veces que ejecuté la función replace fueron muchas y no las alcancé a documentar, pero pueden darse una idea de como y cuantas veces lo utilice. Al final, después de asegurarse que los datos estaban correctos, reemplacé la columna original con los valores de la columna temporal.

También hice un segundo backup de los datos.

```
In [ ]: cine.to_csv('datosLimpiosCine2.csv', sep='\\', index=False)
```

1.3 Conteo y Promedio

Ahora la información que tenemos limpia es: * Año * País * Categoría * £Cómo se enteró?

Primero decidí instalar la librería "tabulate" de Python para obtener resultados ordenados en tablas agradables a la vista. Después corrí el siguiente script para obtener la información de la participación de países divididos por año.

```
In [ ]: listaPaises = []
        for country in paises:
            listaPaises.append(country,cine[(cine['País']==country) & (cine['Año']==2015)].cou
        print(tabulate(listaPaises, tablefmt="github"))
```

Utilicé el count()['Año'] debido a que es el único campo que no tiene NaN y me da un conteo exacto de los registros, pareciera que el contador no funcionó con los NaN por que me arrojó ceros y la cantidad de registros bajó a 2732. Obtuve estos resultados:

```
In [ ]: | País                |2015 |2016 |2017 |2018 |Total |
        |---------------------|-----|-----|-----|-----|------|
        | Colombia            | 477 | 846 | 641 | 438 | 2402 |
        | Irán                |   1 |   0 |   0 |   0 |    1 |
        | España              |  26 |  17 |   1 |   6 |   50 |
        | Venezuela           |   5 |   1 |   1 |   0 |    7 |
        | México              |   8 |   3 |   2 | 171 |  184 |
```

Ecuador	2	3	0	1	6
Francia	2	1	0	0	3
Argentina	1	4	3	1	9
Perú	1	0	0	0	1
Estados Unidos	3	4	3	3	13
nan	0	2	0	1	3
Canadá	1	1	0	0	2
Honduras	1	0	0	0	1
Brasil	5	1	0	0	6
Cuba	1	0	0	0	1
Uruguay	2	1	0	0	3
Alemania	1	1	0	0	2
Internacional	0	5	11	0	16
Marruecos	0	1	0	0	1
Republica Dominicana	0	1	0	0	1
Bélgica	0	1	0	1	2
Afganistán	0	0	19	0	19
Comoras	0	0	1	1	2
Suma	537	893	682	623	2735

Usé también un script para obtener los registros por Categoría y participación por Año. El código fue el siguiente:

```
In [ ]: listaCategoria = []
        >>> for contry in categorias:
        ...     listaCategoria.append([contry,cine[(cine['Categoría']==contry) & (cine['Año']==
        ...
        >>> print(tabulate(listaCategoria, tablefmt="github"))
```

Me di cuenta que efectivamente el NaN no es contado por mi código, me doy cuenta de ésto, debido a que en 2017 tengo varios elementos que yo mismo convertí a NaN porque al momento de capturar la Categoría en México cometieron el error de escribir el Género del video, comprobé que en 2016 me hacen falta 2 video en NaN la tabla anterior y en ésta me hacen falta 165 registros en NaN. Obtuve los siguientes resultados:

```
In [ ]: | Categoría   |2015 |2016 |2017 |2018 |Total |
        |-------------|-----|-----|-----|-----|------|
        | Aficionado  | 313 | 522 | 377 | 241 | 1453 |
        | Profesional | 147 | 151 |  85 |  53 |  436 |
        | Infantil    |  77 |  42 |   0 |   0 |  119 |
        | nan         |   0 |   0 |   0 | 165 |    0 |
        | Juvenil     |   0 | 123 | 132 |  83 |  338 |
        | SmarTIC     |   0 |  55 |  88 |  81 |  224 |
        |-------------|-----|-----|-----|-----|------|
        | Suma        | 537 | 893 | 682 | 623 | 2735 |
```

1.4 Conclusiones

En esta práctica me topé con que se pudo mejorar la limpieza de la práctica pasada. Aprendí a identificar escrituras únicas con la función unique() y a reemplazar string dentro de los DataFrame con la función replace(), reemplacé los nombres de las columnas de todos los DataFrames antes de unirlos y además me di cuenta que era innecesario el DataFrames df2017[4] debido a que solo fue un evento de "Reto al Guion" no había referencia de videos ni de categorías, nada que lo relaciona a las otras tablas de registros.

A los datos les realicé 2 backups, uno después de la limpieza y el otro después de reemplazar el ruido en los campos.

Busqué la forma de sacar promedios y realizar correlaciones, pero en la práctica la Dra. Elisa dijo: "Todas las respuestas abiertas de texto quedan fuera del alcance de esta práctica." Y en mis datos todo, salvo el año, son categorías de texto. Así que doy por concluida mi práctica.

--04 de junio 2019-- Luis Angel Gutierrez Rodríguez 1484412