

Actividad Evaluable 3: Mapas de calor y boxplots

Equipo 1

Objetivo: Determinar si existen correlaciones significativas entre las variables de la base de datos "covid19_tweets.csv".

Importar librerías, leer csv y obtener un vistazo

```
In [ ]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

data = pd.read_csv("covid19_tweets.csv")
data.head(5)
```

	user_name	user_location	user_description	user_created	user_followers	user_friends	user_favourites	user_verified	date	text	hashtags	source	is_re
0	٧١٥٩٤٦	astroworld	wednesday addams as a disney princess keepin i...	2017-05-26 05:46:42	624	950	18775	False	2020-07-25 12:27:21	If I smelled the scent of hand sanitizers toda...		NaN	Twitter for iPhone
1	Tom Basile 🇺🇸	New York, NY	Husband, Father, Columnist & Commentator. Auth...	2009-04-16 20:06:23	2253	1677	24	True	2020-07-25 12:27:17	Hey @Yankees @YankeesPR and @MLB - wouldn't it...		NaN	Twitter for Android
2	Time4fisticuffs	Pewee Valley, KY	#Christian #Catholic #Conservative #Reagan #Re...	2009-02-28 18:57:41	9275	9525	7254	False	2020-07-25 12:27:14	@diane3443 @wdunlap @realDonaldTrump Trump nev...	['COVID19']		Twitter for Android
3	ethel mertz	Stuck in the Middle	#Browns #Indians #ClevelandProud #[] #Cavs ...	2019-03-07 01:45:06	197	987	1488	False	2020-07-25 12:27:10	@brookbanktv The one gift #COVID19 has give me...	['COVID19']		Twitter for iPhone
4	DIPR-J&K	Jammu and Kashmir	Official Twitter handle of Department of Inf...	2017-02-12 06:45:15	101009	168	101	False	2020-07-25 12:27:08	25 July : Media Bulletin on Novel #CoronaVirus...	['CoronaVirusUpdates', 'COVID19']		Twitter for Android

```
In [ ]: data.describe().transpose()
```

	count	mean	std	min	25%	50%	75%	max
user_followers	179108.0	109055.528184	841467.000703	0.0	172.0	992.0	5284.00	49442559.0
user_friends	179108.0	2121.701566	9162.553072	0.0	148.0	542.0	1725.25	497363.0
user_favourites	179108.0	14444.105663	44522.698958	0.0	206.0	1791.0	9388.00	2047197.0

Realiza el análisis de las variables usando diagramas de cajas y bigotes, histogramas y mapas de calor.

Diagrama de cajas y bigotes

El diagrama de cajas y bigotes permite identificar las zonas donde más datos se concentran, así como en las que hay menos. Para poder hacer diagramas de cajas y bigotes, es necesario identificar variables con valores numéricos, las cuales son user_followers, user_friends, user_favourites. A continuación se muestran las gráficas para las variables, respectivamente.

```
In [ ]: # Encontrar columnas del dataframe
df = data[['user_name', 'user_location', 'user_description', 'user_created', 'user_followers', 'user_friends', 'user_favourites', 'user_verified', 'date', 'text', 'has
print(df)
```

```

0      user_name      user_location \
1      Tom Basile 🇺🇸      New York, NY
2      Time4fisticuffs      Pewee Valley, KY
3      ethel mertz      Stuck in the Middle
4      DIPR-J&K      Jammu and Kashmir
...
179103 AJIMATI AbdulRahman O.      Ilorin, Nigeria
179104      Jason      Ontario
179105      BEEHEMOTH 🇨🇦      Canada
179106      Gary DeLPonte      New York City
179107      TUKY II      Aliwal North, South Africa

0      user_description \
1      wednesday addams as a disney princess keepin i...
2      Husband, Father, Columnist & Commentator. Auth...
3      #Christian #Catholic #Conservative #Reagan #Re...
4      #Browns #Indians #ClevelandProud #[]_[] #Cavs ...
...
179103      Animal Scientist|| Muslim|| Real Madrid/Chelsea
179104      When your cat has more baking soda than Ninja ...
179105      🍷 The Architects of Free Trade 🍷 Really Did ...
179106      Global UX UI Visual Designer. StoryTeller, Mus...
179107      TOKELO SEKHOPA | TUKY II | LAST BORN | EISH TU...

0      user_created      user_followers      user_friends      user_favourites \
1      2017-05-26 05:46:42      624      950      18775
2      2009-04-16 20:06:23      2253      1677      24
3      2009-02-28 18:57:41      9275      9525      7254
4      2019-03-07 01:45:06      197      987      1488
...
179103      2013-12-30 18:59:19      412      1609      1062
179104      2011-12-21 04:41:30      150      182      7295
179105      2016-07-13 17:21:59      1623      2160      98000
179106      2009-10-27 17:43:13      1338      1111      0
179107      2018-04-14 17:30:07      97      1697      566

0      user_verified      date \
1      False      2020-07-25 12:27:21
2      True      2020-07-25 12:27:17
3      False      2020-07-25 12:27:14
4      False      2020-07-25 12:27:10
...
179103      False      2020-08-29 19:44:21
179104      False      2020-08-29 19:44:16
179105      False      2020-08-29 19:44:15
179106      False      2020-08-29 19:44:14
179107      False      2020-08-29 19:44:08

0      text \
1      If I smelled the scent of hand sanitizers toda...
2      Hey @Yankees @YankeesPR and @MLB - wouldn't it...
3      @diane3443 @wdunlap @realDonaldTrump Trump nev...
4      @brookbanktv The one gift #COVID19 has give me...
...
179103      Thanks @IamOhmai for nominating me for the @WH...
179104      2020! The year of insanity! Lol! #COVID19 http...
179105      @CTVNews A powerful painting by Juan Lucena. I...
179106      More than 1,200 students test positive for #CO...
179107      I stop when I see a Stop\n\n@SABCNews\n@Izinda...

0      hashtags      source      is_retweet
1      NaN      Twitter for iPhone      False
2      NaN      Twitter for Android      False
3      ['COVID19']      Twitter for Android      False
4      ['COVID19']      Twitter for iPhone      False
...
179103      ['CoronaVirusUpdates', 'COVID19']      Twitter for Android      False
179104      ['WearAMask']      Twitter for Android      False
179105      ['COVID19']      Twitter for Android      False
179106      NaN      Twitter Web App      False
179107      ['COVID19']      Twitter for iPhone      False
179108      NaN      Twitter for Android      False

[179108 rows x 13 columns]

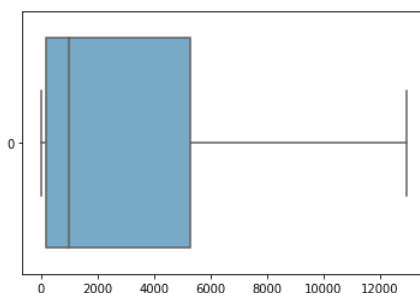
```

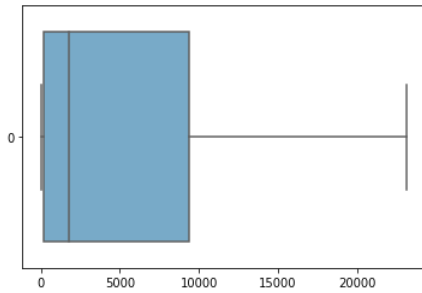
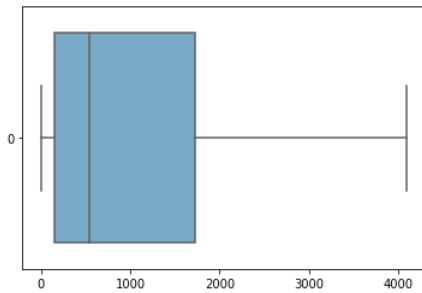
```

In [ ]: # Depurar columnas para solo obtener datos numéricos
numerics = ['int16', 'int32', 'int64', 'float16', 'float32', 'float64']
newdf = data.select_dtypes(include=numerics)

# Graficar
for i in range(len(newdf.columns)):
    sns.boxplot(data=newdf[newdf.columns[i]], orient="h", palette='Blues', showfliers=False)
    plt.show()

```





```
In [ ]: # Encontrando correlaciones entre variables de la base de datos
```

```
Datos=data.iloc[:,[4,7]]
Datos.corr()
```

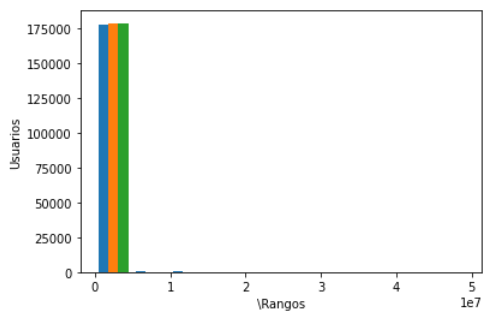
```
Out[ ]:
```

	user_followers	user_verified
user_followers	1.000000	0.320957
user_verified	0.320957	1.000000


Histograma

```
In [ ]: ax = plt.subplot(111)

(counts, bins, patches) = plt.hist(newdf)
plt.xlabel("Rangos")
plt.ylabel("Usuarios")
plt.show()
```



```
In [ ]: dfq= df[df["user_followers"]<199&df["user_favourites"]]
df2 = len(dfq)
dfq
```

	user_name	user_location	user_description	user_created	user_followers	user_friends	user_favourites	user_verified	date	text	hashtags	source
18	Dorian Aur	NaN	NaN	2011-01-30 18:40:16	46	108	453	False	2020-07-25 12:26:43	It is during our darkest moments that we must ...	['light']	Twitter Web App
19	Coronavirus Law	Florida, USA	COVID-19 Practice of Lechner Law	2019-12-03 19:00:11	14	24	74	False	2020-07-25 12:26:39	COVID Update: The infection rate in Florida is...	NaN	Twitter for iPad
27	beatnikgeek the soothsayer	Manhattan, NY	These days, I expose colonizers & exploits @ t...	2008-02-23 19:02:29	86	259	9412	False	2020-07-25 12:26:26	I can imagine the same people profiting off th...	['COVID19']	Twitter for Android
29	 TAX Reform	I ❤️ I ❤️ I ❤️ I	https://t.co /VDCHungubm @1984Vivika https:...	2019-12-06 14:42:15	28	1449	1196	False	2020-07-25 12:26:21	@ratasjuri TAX Reform\n\ntax-free minimum:\nFo...	NaN	Twitter for iPad
32	Beautify Data	Miami, FL	We beautify data to learn and gain insight fro...	2019-02-18 17:11:24	82	92	1152	False	2020-07-25 12:26:17	An update on the total #covid19 cases, recover...	['covid19', 'Africa']	Twitter Web App
...
179068	Eugenio Zuccarelli	New York, USA	Data Scientist Fulbright Scholar @MIT, @Ha...	2020-05-01 22:03:49	26	143	143	False	2020-08-29 19:45:41	Our latest paper on estimating the risk of #CO...	['COVID19', 'MachineLearning']	Twitter for Android
179089	Fzeroone	Lockdown will last for months	"Always look on the bright side of death\njust...	2020-03-20 19:00:59	17	0	737	False	2020-08-29 19:44:56	@jridgway23 The world would be a better place ...	['Covid19']	Twitter Web App
179092	claudia fernandez	NaN	NaN	2015-07-02 23:07:03	68	192	2455	False	2020-08-29 19:44:49	@politvidchannel The Trump Administration's In...	['COVID']	Twitter Web App
179098	John Geer	NaN	#StayAtHome #StayAtHomeSaveLives #MaskUp \nFor...	2020-04-18 01:55:14	61	168	10817	False	2020-08-29 19:44:34	Report #COVID19 outbreaks in K-12 schools here...	['COVID19', 'CloseTheSchools', 'KeepTheSchools...']	Twitter Web App
179099	amyracecar	la playa, el mar .. mi corazón	culinary wizard, auto mechanic, and botanist	2014-02-06 00:55:53	128	542	3506	False	2020-08-29 19:44:34	I have NOTHING BUT 🍀 for the @NBA these days.....	['Covid19']	Twitter Web App

20119 rows × 13 columns

```
In [ ]: # Encontrando correlaciones entre variables de la base de datos

Datos1=data.iloc[:,[4,5]]
print(Datos1.corr())
print("\n")

Datos2=data.iloc[:,[4,6]]
print(Datos2.corr())
print("\n")

Datos3=data.iloc[:,[4,7]]
print(Datos3.corr())
print("\n")

Datos4=data.iloc[:,[5,6]]
print(Datos4.corr())
print("\n")

Datos5=data.iloc[:,[5,7]]
print(Datos5.corr())
print("\n")

Datos6=data.iloc[:,[6,7]]
print(Datos6.corr())
print("\n")

# Realizamos un Heatmap de las correlaciones más altas

corrMatrix = Datos3.corr()
sns.heatmap(corrMatrix, annot=True)
plt.show()

corrMatrix = Datos4.corr()
sns.heatmap(corrMatrix, annot=True)
plt.show()
```

	user_followers	user_friends
user_followers	1.000000	-0.002722
user_friends	-0.002722	1.000000

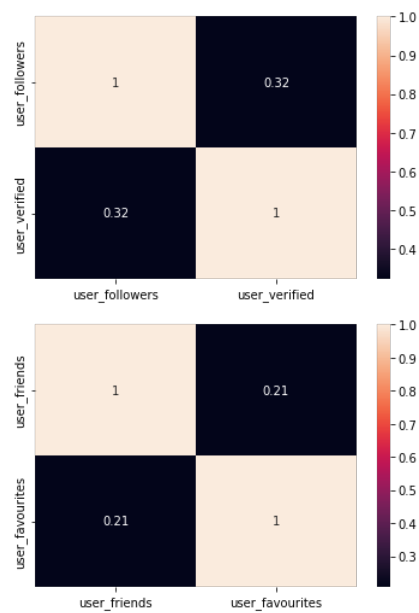
	user_followers	user_favourites
user_followers	1.000000	-0.028724
user_favourites	-0.028724	1.000000

	user_followers	user_verified
user_followers	1.000000	0.322896
user_verified	0.322896	1.000000

	user_friends	user_favourites
user_friends	1.000000	0.207825
user_favourites	0.207825	1.000000

	user_friends	user_verified
user_friends	1.000000	0.013099
user_verified	0.013099	1.000000

	user_favourites	user_verified
user_favourites	1.000000	-0.060316
user_verified	-0.060316	1.000000



Respuestas a preguntas

¿Hay alguna variable que no aporta información?

Andrés Tavera Mihailide: Creemos que todas aportan información indispensable para poder conocer a los usuarios y poder crear un perfil de cada usuario.

Karen Paula Mayorga Guerrero: Basado en el análisis individual de cada variable, se encontró que todas estas pueden ser útiles según diferentes propósitos de investigación.

Alejandro Mariacca Santin: Desde mi punto de vista, primero habría que definir a qué queremos aportar para saber si una variable lo hace o no. Sin embargo, desde un punto de vista muy general, podríamos decir que la variable del nombre de usuario ‘user_name’ no aporta ningún valor. El nombre de usuario rara vez está conectado al resto de los datos ya que no hay ningún tipo de relación entre las variables.

José Antonio Pacheco: Al analizar las variables de la base de datos, llegamos a la conclusión de que todas estas aportan información importante. Obviamente la relevancia de las variables depende de lo que se busque analizar, pero no encontramos variables que fácilmente puedan obtenerse o deducirse de otras. Cada variable aporta información adicional que no se encuentra en las otras.

Duran Sanchez Pablo Ricardo: Analizando solamente la tabla me puedo dar cuenta que no todas las columnas son necesarias, por ejemplo el nombre, no necesitas saber el nombre para el tipo de análisis que realizamos pero puede que para un análisis con diferente objetivos

Si tuvieras que eliminar variables, ¿Cuáles quitarías y por qué?

Andrés Tavera Mihailide: Ninguna. La única que se deriva de otra son los hashtags; sin embargo, es información muy importante para poder recomendar tweets. Si se pueden obtener, pero el tiempo necesario para hacerlo sería muy poco benéfico.

Karen Paula Mayorga Guerrero: En la columna texto se encuentran también los hashtags, por lo que podría ser conveniente eliminar esta columna hashtags. Sin embargo, si en un análisis se requieren utilizar solo los hashtags sin texto, entonces sería eficiente dejar esta columna.

Duran Sanchez Pablo Ricardo: El nombre porque para nuestro estudio no fue requerido, además en búsquedas más específicas se usa el ID no tanto el nombre.

Alejandro Mariacca Santin: Yo personalmente quitaría la variable de nombre de usuario, por privacidad y respeto al individuo que posee la cuenta.

José Antonio Pacheco: Personalmente, la única variable que eliminaría es la de los “hashtags” de la columna 10, puesto que los mismos hashtags pueden encontrarse en la sección de “text”. Sin embargo, si nos interesa conocer únicamente los hashtags de un tweet (mas no el tweet en cuestión), nos conviene tener la columna de “hashtags” para obtenerlos directamente en lugar de buscarlos en “text”.

¿Existen variables que tengan datos extraños?

Andrés Tavera Mihailide: Se pueden encontrar caracteres extraños en los nombres, estos caracteres no siempre son de [a-z] y pueden contener emojis o diversos ASCII. También hay registros que no contienen información. Karen Paula Mayorga Guerrero: Dependiendo de la interfaz que utilice el usuario pueden existir diferentes representaciones de emojis, por lo tanto esto podría ser causa de la presencia de caracteres extraños. Por otro lado, también el concepto de extraño se puede utilizar para algo que para uno no es común pero para otras personas en el planeta si es, entonces puede ser que en otro país, dadas las diferencias de lenguaje ocupen otros caracteres.

Alejandro Mariacca Santin: A primera vista no parece que hayan datos extraños, pues cada uno de ellos tienen un tipo de dato definido y están limitados por el mismo. Sin embargo, analizando más de cerca la información, es posible observar datos atípicos en la variable de ‘user_location’, que indica la localización del usuario al momento de subir su tweet. Dichos datos son extraños porque no pertenecen a lugares reales.

José Antonio Pacheco: Encontramos datos extraños en las columnas de “user name”, “user description” y “user location”. Existen caracteres de texto atípicos en los datos de estas variables. Esto se puede deber a que la base de datos no reconoce caracteres especiales que los usuarios hayan podido escribir en estas secciones, como por ejemplo, los emojis.

Duran Sanchez Pablo Ricardo: Existen caracteres extraños que pueden afectar la búsqueda de datos en las variables de user_location

Si comparas las variables, ¿todas están en rangos similares?

Andrés Tavera Mihailide: No hay rangos similares, los followers pueden variar entre 0 y 49442559; los “user_friends” entre 0 y 497363 y los “user_favourites” entre 2047197. Cabe destacar la gran diferencia entre el máximo de personas que un usuario sigue y el máximo de personas que siguen a una persona varían notablemente.

Karen Paula Mayorga Guerrero: No están en rango similares, por ejemplo, la variable user_followers tiene un rango de [0,49 442 559.0], la variable user_friends tiene un rango de [0, 497 363.0], mientras que la variable tiene un rango de [0,2 047 197.0].

Alejandro Mariacca Santin: Para las variables numéricas, es posible ver que todas tienen un mínimo de 0, que podría indicar cuentas nuevas o nunca usadas, mientras que los máximos varían mucho y es ahí donde se diferencian las variables entre sí. Por ejemplo, el máximo de la variable user_friends es apenas el 1% del máximo de la variable user_followers.

José Antonio Pacheco: A pesar de que el rango mínimo de las variables numéricas (“user_followers”, “user_friends”, y “user_favorites”) es igual a 0 en todas, el rango máximo es más variable para cada una. Este valor máximo para “user_followers” es 49442559, para “user_friends” es 497363, y para “user_favorites” es de 2047197.

Duran Sanchez Pablo Ricardo: No, en general no todas las variables son numéricas, hay booleanas y cadena de caracteres, y las que sí son numéricas tienen rangos muy distintos.

¿Crees que esto afecte el análisis de los datos?

Andrés Tavera Mihailide: Lo único que notamos es la poca correlación entre las personas que siguen a una persona, comparado con cuantas personas lo siguen. Vimos que está correlación es de un 2%. Más que afectar el análisis de datos, te guía para saber dónde puedes buscar correlaciones.

Karen Paula Mayorga Guerrero: El rango de datos no tiene inferencia en el análisis, pues solo se busca una asociación entre los datos.

Alejandro Mariacca Santin: No usamos mínimos ni máximos para nuestros análisis.

José Antonio Pacheco: Dado que los rangos de las variables numéricas son muy distintos en tamaño, realmente no existirá mucha correlación de unas con otras. Encontramos que las correlaciones más altas que obtuvimos fueron de 0.21 con las variables de “user_friends” y “user_favorites”, y de 0.32 con las variables de “user_followers” y “user_verified”. Dichos valores no representan una correlación muy grande entre estas variables, por lo que podemos concluir que no dependen una de otra.

Duran Sanchez Pablo Ricardo: No porque pues para nuestro análisis no requerimos de alguna limitante como el rango de valores.

¿Puedes encontrar grupos que se parezcan? ¿Qué grupos son estos?

Andrés Tavera Mihailide: Hay 20119 que tienen menos de 200 followers tienen user favourites, esto nos indica que hay ciertos representantes que tienen pocos seguidores pero aún así son activos en la red social.

Karen Paula Mayorga Guerrero: Después de analizar diferentes casos de correlación no hubo un índice alto. Por lo tanto, este podría ser un motivo para que no se parezcan.

Alejandro Mariacca Santin: Dentro de nuestro análisis, no encontramos grandes relaciones. Por ejemplo, calculamos la correlación entre user_followers y user_verified y el resultado

Alejandro Manacca Santini: Dentro de nuestro análisis, no encontramos grandes relaciones. Por ejemplo, calculamos la correlación entre user_followers y user_verified y el resultado fue de 0.32, por lo que concluimos que no se parecen.

Duran Sanchez Pablo Ricardo: En nuestro tipo de análisis no encontramos grupos que se parezcan o relacionen. En si los datos que analizamos ya son un grupo de gente que tiene relación al COVID 19.

José Antonio Pacheco: De acuerdo con nuestro análisis, no encontramos similitudes relevantes en nuestros grupos. Esto puede deberse a que no hay variables que presenten una correlación muy elevada, y que los rangos de valores de dichas variables son muy dispares.

1. ¿Hay alguna variable que no aporta información?

Encontramos algunas variables que no nos aportaban información como lo fueron el nombre de los usuarios y la localización

1. Si tuvieras que eliminar variables, ¿Cuáles quitarías y por qué?

Quitaría el nombre de usuario y localización ya que estos datos no pueden ser analizados debido a los diferentes tipos de caracteres que pueden estar en un mismo texto.

1. ¿Existen variables que tengan datos extraños?

Las variables de nombre y localización pueden contener datos extraños como lo son al momento de escribir el nombre y de donde están ubicados pueden poner datos extraños que nos resultarían difíciles de analizar.

1. Si comparas las variables, ¿todas están en rangos similares?

Se puede observar que no hay rangos similares ya que los datos que utilizamos entre los favoritos de los usuarios y aigos no tiene de similitud

1. ¿Crees que esto afecte el análisis de los datos?

No creo que las variables que quitamos afecten al análisis de nuestros datos ya que hay algunos que contienen datos que simplemente no aportarían nada de importancia al análisis

2. ¿Puedes encontrar grupos que se parezcan? ¿Qué grupos son estos?

Después de analizar los grupos ninguno se tienen nada de parentesco ya que no hay correlación entre los grupos y es por eso que podemos decir que no hay grupos

In []: