Entregable2

March 17, 2021

1 Entregable 2. Obtención de Estadísticas Descriptivas

Equipo 6:

A01706095 - Naomi Estefanía Nieto Vega

A01706189 - Alejandro Angel Calderon Berges

A01706596 - Carlos Soria de la Cabada

Paso 1. Cargar los datos de nuestra base de datos con ayuda de pandas

En esta sección únicamente se leen los datos, como podemos ver es acerca de unos artículos publicados en la web.

[41]:
```
import pandas as pd #Aqui importamos la libreria
import csv

df=pd.read_csv("articulos_ml.csv") #Aqui leemos el archivo
print(df)
```

```
                                                 Title  \
0    What is Machine Learning and how do we use it …
1     10 Companies Using Machine Learning in Cool Ways
2    How Artificial Intelligence Is Revolutionizing…
3    Dbrain and the Blockchain of Artificial Intell…
4    Nasa finds entire solar system filled with eig…
..                                                 …
156  [Log] 83: How Google Uses Machine Learning And…
157  [Log] 84: Zuck Knows If You've Been Bad Or Goo…
158  [Log] 85: Microsoft Improves Windows Phone Voi…
159  [Log] 86: How Google's Acquisition Of DNNresea…
160   [Log] 87: Google's Cloud Is Eating Apple's Lunch

                                                   url  Word count  \
0    https://blog.signals.network/what-is-machine-l…        1888
1                                                  NaN        1742
2                                                  NaN         962
3                                                  NaN        1221
4                                                  NaN        2039
..                                                 …           …
```

```
156  [Log] 83: http://feedproxy.google.com/~r/Techc…        3239
157  [Log] 84: http://feedproxy.google.com/~r/Techc…        2566
158  [Log] 85: http://feedproxy.google.com/~r/Techc…        2089
159  [Log] 86: http://feedproxy.google.com/~r/Techc…        1530
160  [Log] 87: http://feedproxy.google.com/~r/Techc…         953
```

	# of Links	# of comments	# Images video	Elapsed days	# Shares
0	1	2.0	2	34	200000
1	9	NaN	9	5	25000
2	6	0.0	1	10	42000
3	3	NaN	2	68	200000
4	1	104.0	4	131	200000
..	…	…	…	…	…
156	3	11.0	1	84	3239
157	3	8.0	4	85	25019
158	4	4.0	1	86	49614
159	4	12.0	3	87	33660
160	6	13.0	2	88	5956

```
[161 rows x 8 columns]
```

[2]:
```
import pandas as pd #Aqui importamos la libreria
import csv

df=pd.read_csv("articulos_ml.csv") #Aqui leemos el archivo
df.describe() #Aqui le decimos que nos muestre los analiticos en la tabla de␣
 ↪abajo
```

[2]:

	Word count	# of Links	# of comments	# Images video	Elapsed days	\
count	161.000000	161.000000	129.000000	161.000000	161.000000	
mean	1808.260870	9.739130	8.782946	3.670807	98.124224	
std	1141.919385	47.271625	13.142822	3.418290	114.337535	
min	250.000000	0.000000	0.000000	1.000000	1.000000	
25%	990.000000	3.000000	2.000000	1.000000	31.000000	
50%	1674.000000	5.000000	6.000000	3.000000	62.000000	
75%	2369.000000	7.000000	12.000000	5.000000	124.000000	
max	8401.000000	600.000000	104.000000	22.000000	1002.000000	

	# Shares
count	161.000000
mean	27948.347826
std	43408.006839
min	0.000000
25%	2800.000000
50%	16458.000000
75%	35691.000000
max	350000.000000

Paso 2. Verificar la cantidad de datos, las variables que contiene cada vector de datos e identifica el tipo de variables.

En general tenemos 162 datos, de los cuales existen las variables o columas con datos de cantidad de palabras, número de links, número de comentarios, número de imagénes de video, días transcurridos y cantidad de veces que ha sido compartido.

El tipo de las variables es float en su mayoría.

Paso 3. Analizar las variables para saber que representa cada una y en que rangos se encuentran.

En este caso las variables como ya mencioné antes representan una cantidad de las veces que ha ocurrido un evento, que bien puede ser compartir, comentar, etc. Los rangos para cada variable son los siguientes:

[38]:
```
import pandas as pd #Importamos de nuevo la librería para este fragmento de␣
 ↪código.

# initialize list of lists
data = [['word count', 250, 8401],['num of links', 0, 600],['num of comments',␣
 ↪0, 104],['num of images video', 1, 22],['elapsed days', 1, 1002],['num of␣
 ↪shares', 0, 350000]]

# Create the pandas DataFrame
df = pd.DataFrame(data, columns = ['Variables', 'Min Range','Max Range'])

# print dataframe.
print(df)
```

```
             Variables  Min Range  Max Range
0           word count        250       8401
1        num of links          0        600
2      num of comments          0        104
3  num of images video          1         22
4         elapsed days          1       1002
5        num of shares          0     350000
```

Paso 4. Conclusiones

[40]:
```
import pandas as pd #Aqui importamos la libreria

data = [['word count',161.000000,1808.260870,1141.919385],
        ['num of links',161.000000,9.739130,47.271625],
        ['num of comments',129.000000,8.782946,13.142822],
        ['num of images video',161.000000,3.670807,3.418290],
        ['elapsed days',161.000000,98.124224,114.337535],
        ['num of shares',161.000000,27948.347826,43408.006839]]

df = pd.DataFrame(data, columns = ['Variables', 'count','mean','std'])
print(df)
```

	Variables	count	mean	std
0	word count	161.0	1808.260870	1141.919385
1	num of links	161.0	9.739130	47.271625
2	num of comments	129.0	8.782946	13.142822
3	num of images video	161.0	3.670807	3.418290
4	elapsed days	161.0	98.124224	114.337535
5	num of shares	161.0	27948.347826	43408.006839

Como conclusión de acuerdo a las variables podemos asumir que hay un promedio de cantidad de veces que se comparte un artículo bastante alto, siendo este de 27,948; esto podría decirnos que a la mayoría de los usuarios que interactuan les gusta ese artículo. Además es importante observar que tenemos un número alto en la desviación estándar esto quiere decir que existe una gran dispersión en la población de los datos con respecto a la media. La cuenta en general para las variables es de 161, a excepción de la cantidad de comentarios que es de 129. Asimismo tenemos una media aritmética de 1808, esto es la cantidad promedio de palabras que están escritas en los artículos.

[ ]: