# Alexis Chavez Juarez A01657486

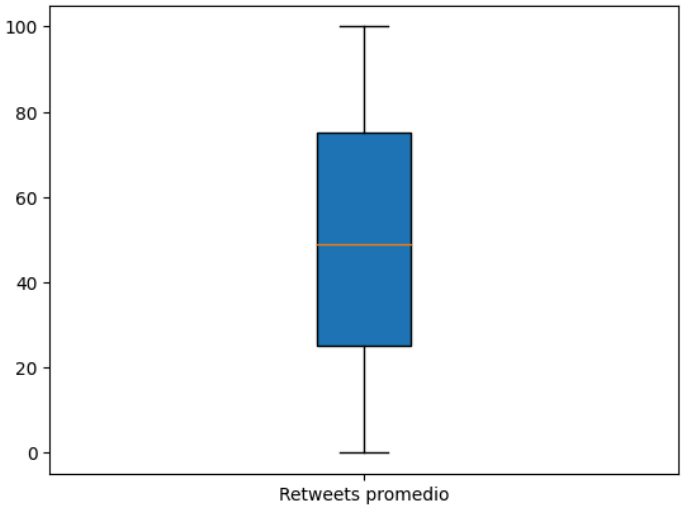
## Mapas de calor y boxplots

```
1 import numpy as np
2 import pandas as pd
3 import csv
4 import matplotlib.pyplot as plt
5 import seaborn as sns
6
7 df = pd.read_csv('twitter_dataset.csv')
8 df
```

| | Tweet_ID | Username | Text | Retweets | Likes | Timestamp |
|---|---|---|---|---|---|---|
| **0** | 1 | julie81 | Party least receive say or single. Prevent pre... | 2 | 25 | 2023-01-30 11:00:51 |
| **1** | 2 | richardhester | Hotel still Congress may member staff. Media d... | 35 | 29 | 2023-01-02 22:45:58 |
| **2** | 3 | williamsjoseph | Nice be her debate industry that year. Film wh... | 51 | 25 | 2023-01-18 11:25:19 |
| **3** | 4 | danielsmary | Laugh explain situation career occur serious. ... | 37 | 18 | 2023-04-10 22:06:29 |
| **4** | 5 | carlwarren | Involve sense former often approach government... | 27 | 80 | 2023-01-24 07:12:21 |
| **...** | ... | ... | ... | ... | ... | ... |
| **9995** | 9996 | ntate | Agree reflect military box ability ever hold. ... | 81 | 86 | 2023-01-15 11:46:20 |
| **9996** | 9997 | garrisonjoshua | Born which push still. Degree sometimes contro... | 73 | 100 | 2023-05-06 00:46:54 |

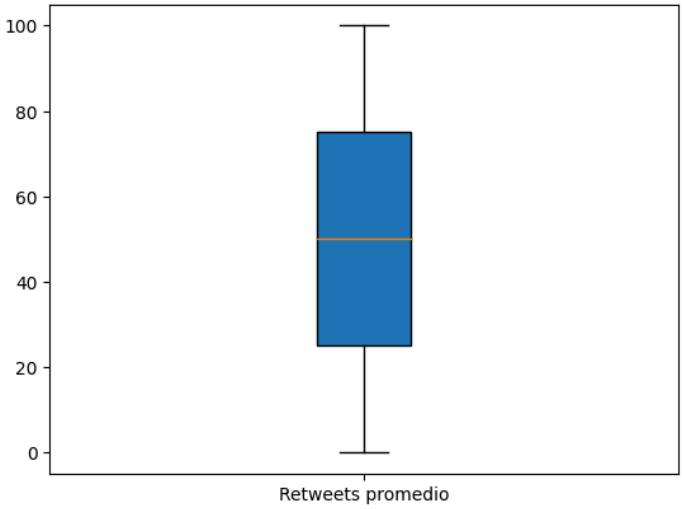
```
1 plt.boxplot([df["Retweets"]], patch_artist=True, labels=["Retweets promedio"])
```

```
{'whiskers': [<matplotlib.lines.Line2D at 0x7a245b4442b0>,
  <matplotlib.lines.Line2D at 0x7a245b444550>],
 'caps': [<matplotlib.lines.Line2D at 0x7a245b4447f0>,
  <matplotlib.lines.Line2D at 0x7a245b444a90>],
 'boxes': [<matplotlib.patches.PathPatch at 0x7a245d4d3ee0>],
 'medians': [<matplotlib.lines.Line2D at 0x7a245b444d30>],
 'fliers': [<matplotlib.lines.Line2D at 0x7a245b444fd0>],
 'means': []}
```



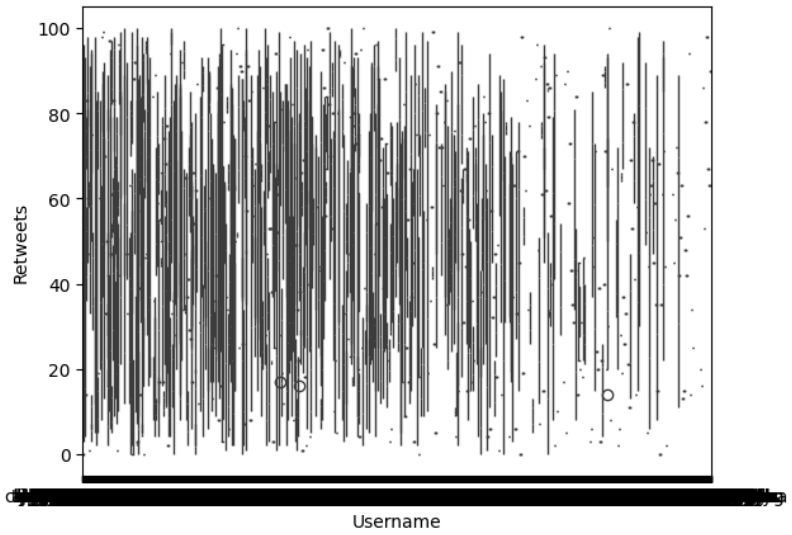

```
1 plt.boxplot([df["Likes"]], patch_artist=True, labels=["Retweets promedio"])
```

```
{'whiskers': [<matplotlib.lines.Line2D at 0x7a242e91df90>,
  <matplotlib.lines.Line2D at 0x7a242e91ea10>],
 'caps': [<matplotlib.lines.Line2D at 0x7a242e91f010>,
  <matplotlib.lines.Line2D at 0x7a242e91fc10>],
 'boxes': [<matplotlib.patches.PathPatch at 0x7a242e91d2d0>],
 'medians': [<matplotlib.lines.Line2D at 0x7a24291bd960>],
 'fliers': [<matplotlib.lines.Line2D at 0x7a24291bdc00>],
 'means': []}
```



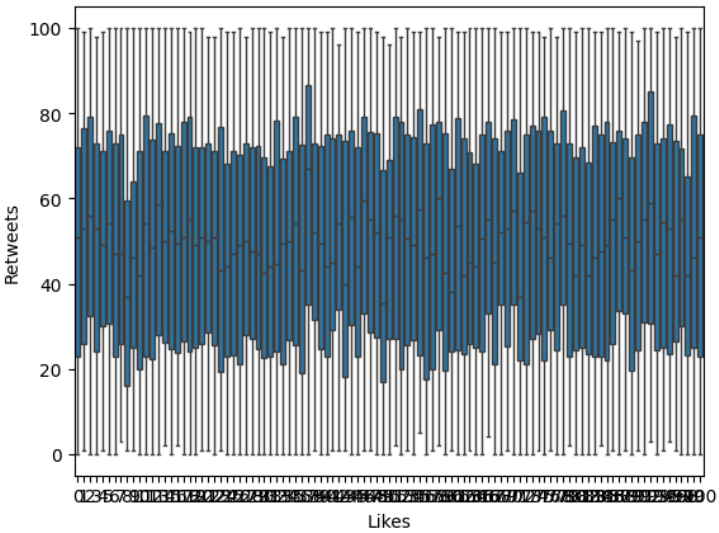

```
1 sns.boxplot(data=df, y="Retweets", x="Username")
```

```
<Axes: xlabel='Username', ylabel='Retweets'>
```



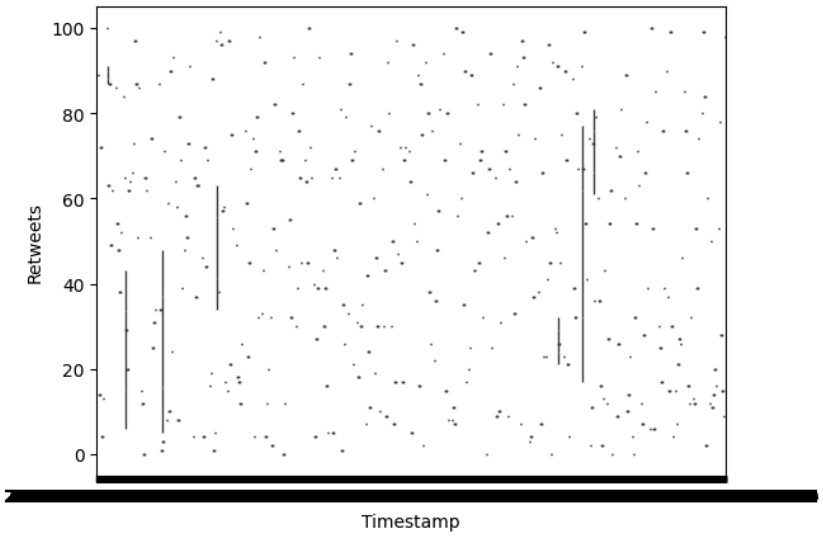

```
1 sns.boxplot(data=df, y="Retweets", x="Likes")
```

```
<Axes: xlabel='Likes', ylabel='Retweets'>
```



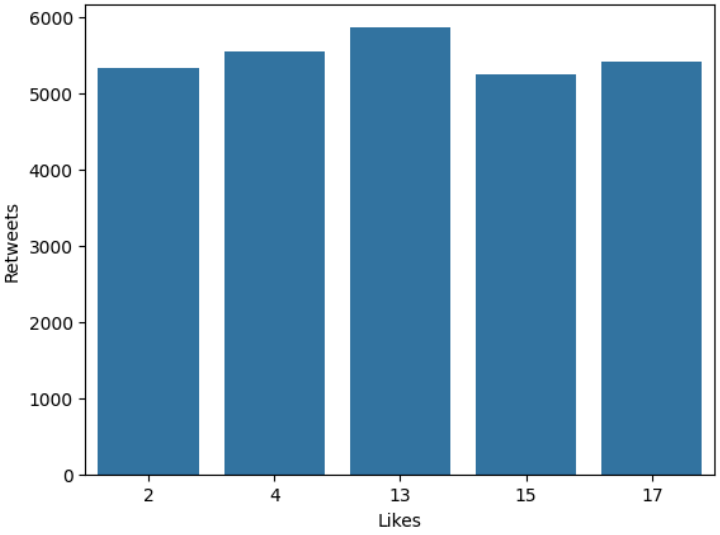

```
1 sns.boxplot(data=df, y="Retweets", x="Timestamp")
```

```
<Axes: xlabel='Timestamp', ylabel='Retweets'>
```



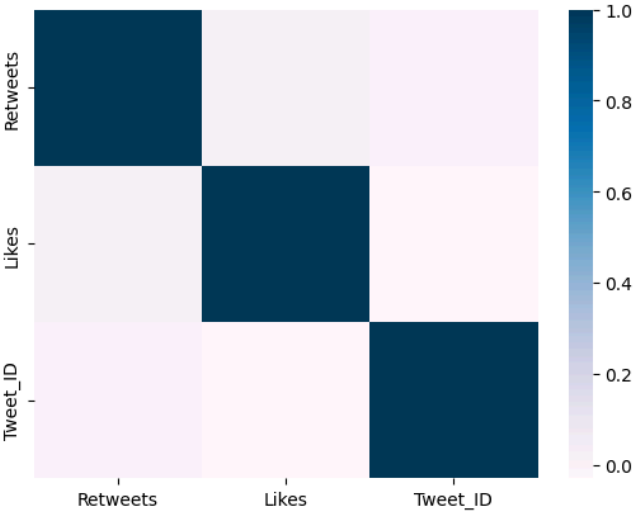

```
1 ydf=df[df["Likes"]<=20]
2 ydf=ydf.groupby(["Likes"]).sum()
3 ydf=ydf.sort_values("Retweets",ascending=False)
4 sns.barplot(data=ydf.head(5),y="Retweets", x="Likes")
```

```
<Axes: xlabel='Likes', ylabel='Retweets'>
```



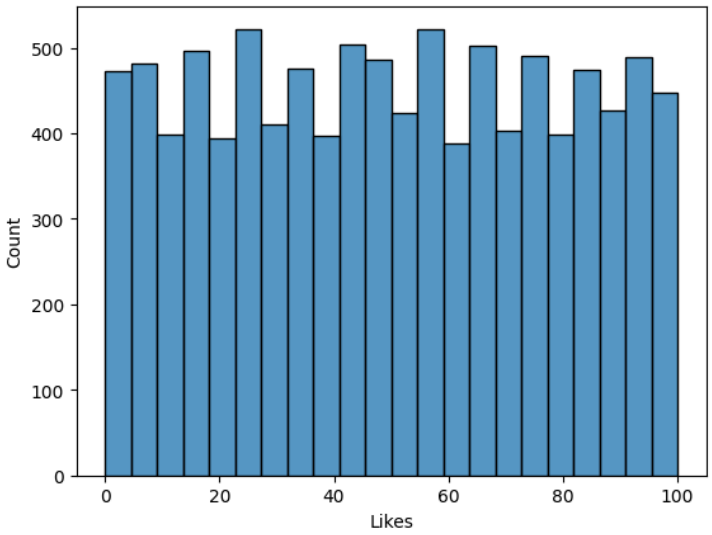

```
1 dfcor=df[['Retweets','Likes','Tweet_ID']].corr()
2 sns.heatmap(dfcor,cmap="PuBu")
```

```
<Axes: >
```




```
1 sns.histplot(data=df,x="Likes")
```

```
<Axes: xlabel='Likes', ylabel='Count'>
```




## Resolución de preguntas:

### ¿Hay alguna variable que no aporta información?

Sí, hay 2 variables que no aportan información porque no se pueden comparar numericamente, las cuales son el Tweet_ID y el texto del tweet. El tweet ID se puede sacar sumándole uno al índice de cada tweet, entonces no es de ningún valor, solo es informativo si se quiere saber que tweet es. Y por otro lado, el texto del tweet no lo estamos comparando, entonces no es necesario en este caso.

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

Cómo se menciono en la pregunta anterior, quitaría el texto del tweet, el tweet_ID y dependiendo que se quiera comparar, también el timestamp. Por ejemplo, en el mapa de calor no se puede comparar, pero en las cajas y bigotes sí.

### ¿Existen variables que tengan datos extraños?

Considero que en este caso, no hay ninguna variable con valor extraño, solo que son difíciles de comparar.

### Si comparas las variables, ¿todas están en rangos similares? ¿Crees que esto afecte?

Las únicas que se pueden comparar son tweet likes y retweets. Entonces si están en rangos similares. Excepto las demás variables que no tienen ninguna relación ni valor similar.

### ¿Puedes encontrar grupos qué se parezcan? ¿Qué grupos son estos?

Los grupos numéricos y los grupos no numéricos. Que por un lado son: tweet_ID, likes y retweets. Y en no numéricos son username, text y timestamp.

Link de github: link