```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

```
from google.colab import drive
drive.mount('/content/drive')
```

```
    Mounted at /content/drive
```

DATASET

```
PP = pd.read_csv('/content/drive/MyDrive/Datas/adhdwomen-comment.csv')
PP.head()
display(PP)
```

	body	id	score	created_utc	created_datetime
0	I'd like to see this sub be more active, too. ...	cqowxhs	1	1430023102	2015-04-26 04:38:22
1	I've found people are more receptive when you ...	cvzg3v2	1	1444835103	2015-10-14 15:05:03
2	Thank you so much. I have been trying to use m...	cw65vo8	1	1445326215	2015-10-20 07:30:15
3	[deleted]	d2tscyn	1	1462457040	2016-05-05 14:04:00
4	Sooooo, not sure why you were told it was 24 h...	d38enqz	1	1463457224	2016-05-17 03:53:44
...	...	...	...	...	...
202653	I love this, thank you!\n\nMy life is worlds b...	gwut5l8	1	1620091758	2021-05-04 01:29:18
202654	Man you sound exactly like me lol, except for ...	gwutlrw	4	1620091974	2021-05-04 01:32:54
202655	I get noise rage too. I'll cast another vote f...	gwutuxo	3	1620092094	2021-05-04 01:34:54
202656	I too have found non-Covid uses for my Covid m...	gwutvrm	2	1620092105	2021-05-04 01:35:05
202657	I mix this In with my chore schedule. On Satur...	gwuu5jy	2	1620092235	2021-05-04 01:37:15

202658 rows × 5 columns

ELIMINAR DUPLICADOS

```
PP_SD = PP.drop_duplicates(['id'])
display(PP_SD)
```

	body	id	score	created_utc	created_datetime
0	I'd like to see this sub be more active, too. ...	cqowxhs	1	1430023102	2015-04-26 04:38:22
1	I've found people are more receptive when you ...	cvzg3v2	1	1444835103	2015-10-14 15:05:03
2	Thank you so much. I have been trying to use m...	cw65vo8	1	1445326215	2015-10-20 07:30:15
3	[deleted]	d2tscyn	1	1462457040	2016-05-05 14:04:00
4	Sooooo, not sure why you were told it was 24 h...	d38enqz	1	1463457224	2016-05-17 03:53:44
...	...	...	...	...	...
202653	I love this, thank you!\n\nMy life is worlds b...	gwut5l8	1	1620091758	2021-05-04 01:29:18
202654	Man you sound exactly like me lol, except for ...	gwutlrw	4	1620091974	2021-05-04 01:32:54
202655	I get noise rage too. I'll cast another vote f...	gwutuxo	3	1620092094	2021-05-04 01:34:54
202656	I too have found non-Covid uses for my Covid m...	gwutvrm	2	1620092105	2021-05-04 01:35:05
202657	I mix this In with my chore schedule. On Satur...	gwuu5jy	2	1620092235	2021-05-04 01:37:15

202658 rows × 5 columns

ELIMINAR LOS TEXTOS [deleted]

```
PP_SDE = PP_SD[PP_SD['body'] == '[deleted]'].index.to_list()
dfADHD = PP_SD.drop(index=(PP_SDE))
```

```
display(dfADHD)
```

	body	id	score	created_utc	created_datetime
0	I'd like to see this sub be more active, too. ...	cqowxhs	1	1430023102	2015-04-26 04:38:22
1	I've found people are more receptive when you ...	cvzg3v2	1	1444835103	2015-10-14 15:05:03
2	Thank you so much. I have been trying to use m...	cw65vo8	1	1445326215	2015-10-20 07:30:15
4	Sooooo, not sure why you were told it was 24 h...	d38enqz	1	1463457224	2016-05-17 03:53:44
5	My doctor is reluctant to give me a fast actin...	d38f9y3	1	1463458428	2016-05-17 04:13:48
...	...	...	...	...	...
202653	I love this, thank you!\n\nMy life is worlds b...	gwut5l8	1	1620091758	2021-05-04 01:29:18
202654	Man you sound exactly like me lol, except for ...	gwutlrw	4	1620091974	2021-05-04 01:32:54
202655	I get noise rage too. I'll cast another vote f...	gwutuxo	3	1620092094	2021-05-04 01:34:54
202656	I too have found non-Covid uses for my Covid m...	gwutvrm	2	1620092105	2021-05-04 01:35:05
202657	I mix this In with my chore schedule. On Satur...	gwuu5jy	2	1620092235	2021-05-04 01:37:15

201478 rows × 5 columns

ELIMINAR COLUMNAS INNECESARIAS

```
dataset = dfADHD.drop(['id','created_utc','score'], axis =1)
dataset.head()
```

	body	created_datetime
0	I'd like to see this sub be more active, too. ...	2015-04-26 04:38:22
1	I've found people are more receptive when you ...	2015-10-14 15:05:03
2	Thank you so much. I have been trying to use m...	2015-10-20 07:30:15
4	Sooooo, not sure why you were told it was 24 h...	2016-05-17 03:53:44
5	My doctor is reluctant to give me a fast actin...	2016-05-17 04:13:48

CREAR COLUMNAS DE AÑO Y MES

```
from datetime import datetime
ss=pd.to_datetime(dataset.created_datetime)
dataset['Year'] = pd.to_datetime(dataset.created_datetime)
dataset['Year'] = dataset['Year'].dt.year
dataset['Month'] = pd.to_datetime(dataset.created_datetime)
dataset['Month'] = dataset['Month'].dt.month
dataset.head()
```

	body	created_datetime	Year	Month
0	I'd like to see this sub be more active, too. ...	2015-04-26 04:38:22	2015	4
1	I've found people are more receptive when you ...	2015-10-14 15:05:03	2015	10
2	Thank you so much. I have been trying to use m...	2015-10-20 07:30:15	2015	10
4	Sooooo, not sure why you were told it was 24 h...	2016-05-17 03:53:44	2016	5
5	My doctor is reluctant to give me a fast actin...	2016-05-17 04:13:48	2016	5

CREAR MUESTRA ALEATORIO

```
dataset = dataset.sample(n=60000)  # Obtener una muestra aleatoria de 50000 filas
len(dataset)
```

```
60000
```

LIMPIEZA DE CARACTERES ESPECIALES

```
import re

# Define a function to clean the text
def clean(body):
    # Removes all special characters and numericals leaving the alphabets
    body = re.sub('[^A-Za-z]+', ' ', body)
    return body

# Cleaning the text in the review column
dataset['Cleaned Body'] = dataset['body'].apply(clean)
dataset.head()
```

	body	created_datetime	Year	Month	Cleaned Body
166382	When I'm struggling to schedule a full workout...	2021-02-25 19:37:54	2021	2	When I m struggling to schedule a full workout...
74485	It was fun to do in an afternoon of hyper focu...	2020-09-09 21:26:53	2020	9	It was fun to do in an afternoon of hyper focu...
114975	Thank you! I've just increased to 30 today so ...	2020-11-28 15:52:54	2020	11	Thank you I ve just increased to today so hope...
[illegible]	What kind of provider did you see?	[illegible]	[illegible]	[illegible]	What kind of provider did you see

```
import numpy as np
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
```

```
# Creamos un DataFrame
df = pd.DataFrame(dataset, columns=["Cleaned Body"])

# Vectorizamos los comentarios usando TF-IDF
vectorizer = TfidfVectorizer()
X = vectorizer.fit_transform(df["Cleaned Body"])

# Aplicamos el algoritmo de K-Means para agrupar los comentarios en 2 clusters
n_clusters = 6
kmeans = KMeans(n_clusters=n_clusters, random_state=42)
kmeans.fit(X)

# Agregamos las etiquetas de los clusters al DataFrame
df["Cluster"] = kmeans.labels_

# Visualizamos los resultados
for cluster_id in range(n_clusters):
    cluster_comments = df[df["Cluster"] == cluster_id]["Cleaned Body"].tolist()
    print(f"Cluster {cluster_id + 1}:")
    for comment in cluster_comments:
        print(f"  - {comment}")

# Para visualizar los centroides de los clusters
#centroids = kmeans.cluster_centers_
#plt.scatter(centroids[:, 0], centroids[:, 1], marker='x', s=150, linewidths=3, color='r')
#plt.title("Centroides de los Clusters")
#plt.show()
df
```

```
display(df)
```

```
                                                 Cleaned Body  Cluster
166382     When I m struggling to schedule a full workout...        1
74485       It was fun to do in an afternoon of hyper focu...        4
114975     Thank you I ve just increased to today so hope...        5
170813     What kind of provider did you see How did the ...        5
94275                               Ufo plant Growing like hell        3
...                                                       ...      ...
14827     Ahahah amazing I regularly switch shoes a coup...        4
150760     I ve been taking stimulants for about months a...        1
49104            I completely relate to all of that I honestly ...        1
119016      When the pandemic hit I was in a thus far fail...        4
180007     who here has rewatched their favorite show a t...        3

60000 rows × 2 columns
```