Nombre : Josue Alejandro Sauca Pucha

Fecha : 16-06-2023

1. Elija y descargue un data set desde un repositorio libre y realice un análisis de correlación de sus variables.

In [67]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

data = pd.read_csv('covid-05-17-2020.csv')
```

In [68]:
```
#Se presenta los datos de la base de datos
data
```

Out[68]:

	FIPS	Province_State	Country_Region	Last_Update	Lat	Long_	Confirmed
0	45001.0	South Carolina	US	5/18/2020 2:32	34.223334	-82.461707	35
1	22001.0	Louisiana	US	5/18/2020 2:32	30.295065	-92.414197	198
2	51001.0	Virginia	US	5/18/2020 2:32	37.767072	-75.632346	688
3	16001.0	Idaho	US	5/18/2020 2:32	43.452658	-116.241552	773
4	19001.0	Iowa	US	5/18/2020 2:32	41.330756	-94.471059	5
...	...	...	...	...	...	...	...
3302	NaN	NaN	West Bank and Gaza	5/18/2020 2:32	31.952200	35.233200	381
3303	NaN	NaN	Western Sahara	5/18/2020 2:32	24.215500	-12.885800	6
3304	NaN	NaN	Yemen	5/18/2020 2:32	15.552727	48.516388	128
3305	NaN	NaN	Zambia	5/18/2020 2:32	-13.133897	27.849332	753
3306	NaN	NaN	Zimbabwe	5/18/2020 2:32	-19.015438	29.154857	44

3307 rows × 10 columns

In [69]:
```
#Se verifica si existen datos nulos dentro de la misma
data.isnull().sum()
```

Out[69]:
```
FIPS              329
Province_State    181
Country_Region      0
Last_Update         0
Lat                67
Long_              67
Confirmed           0
Deaths              0
Active              0
Combined_Key        0
dtype: int64
```

In [70]:
```
#Se procede a remplaza los valores nulos con NaN
datos = data.replace('NaN', np.nan)
```

In [71]:
```
#Se presenta la conversion de la tabla
datos
```

Out[71]:

	FIPS	Province_State	Country_Region	Last_Update	Lat	Long_	Confirmed
0	45001.0	South Carolina	US	5/18/2020 2:32	34.223334	-82.461707	35
1	22001.0	Louisiana	US	5/18/2020 2:32	30.295065	-92.414197	198
2	51001.0	Virginia	US	5/18/2020 2:32	37.767072	-75.632346	688
3	16001.0	Idaho	US	5/18/2020 2:32	43.452658	-116.241552	773
4	19001.0	Iowa	US	5/18/2020 2:32	41.330756	-94.471059	5
...	...	...	...	...	...	...	...
3302	NaN	NaN	West Bank and Gaza	5/18/2020 2:32	31.952200	35.233200	381
3303	NaN	NaN	Western Sahara	5/18/2020 2:32	24.215500	-12.885800	6
3304	NaN	NaN	Yemen	5/18/2020 2:32	15.552727	48.516388	128
3305	NaN	NaN	Zambia	5/18/2020 2:32	-13.133897	27.849332	753
3306	NaN	NaN	Zimbabwe	5/18/2020 2:32	-19.015438	29.154857	44

3307 rows × 10 columns

In [72]:
```
#Se llena los datos nulos mediante el metodo ffill, es un metodo que
#toma la ultima observacion y la propaga
datos = datos.fillna(method='ffill')
```

In [73]:
```
#Como se ve el dataset ya no esta nulo
datos.isnull().sum()
```

Out[73]:
```
FIPS              0
Province_State    0
Country_Region    0
Last_Update       0
Lat               0
Long_             0
Confirmed         0
Deaths            0
Active            0
Combined_Key      0
dtype: int64
```

In [74]:
```
#Se presenta el nuevo dataset con los datos rellenados
datos
```

Out[74]:

	FIPS	Province_State	Country_Region	Last_Update	Lat	Long_	Confirmed
0	45001.0	South Carolina	US	5/18/2020 2:32	34.223334	-82.461707	35
1	22001.0	Louisiana	US	5/18/2020 2:32	30.295065	-92.414197	198
2	51001.0	Virginia	US	5/18/2020 2:32	37.767072	-75.632346	688
3	16001.0	Idaho	US	5/18/2020 2:32	43.452658	-116.241552	773
4	19001.0	Iowa	US	5/18/2020 2:32	41.330756	-94.471059	5
...	...	...	...	...	...	...	...
3302	78.0	Zhejiang	West Bank and Gaza	5/18/2020 2:32	31.952200	35.233200	381
3303	78.0	Zhejiang	Western Sahara	5/18/2020 2:32	24.215500	-12.885800	6
3304	78.0	Zhejiang	Yemen	5/18/2020 2:32	15.552727	48.516388	128
3305	78.0	Zhejiang	Zambia	5/18/2020 2:32	-13.133897	27.849332	753
3306	78.0	Zhejiang	Zimbabwe	5/18/2020 2:32	-19.015438	29.154857	44

3307 rows × 10 columns

In [75]:
```
datos.describe()
```

Out[75]:

	FIPS	Lat	Long_	Confirmed	Deaths	Active
count	3307.000000	3307.000000	3307.000000	3307.000000	3307.000000	3307.000000
mean	29210.367705	37.027505	-80.857069	1425.346235	95.308437	804.931358
std	18991.635514	9.855225	40.855055	10807.399817	1017.738128	8730.230598
min	66.000000	-51.796300	-164.035380	0.000000	0.000000	-274829.000000
25%	16058.000000	34.068572	-96.769289	9.000000	0.000000	8.000000
50%	28085.000000	38.147175	-88.265644	41.000000	1.000000	35.000000
75%	45080.000000	41.864593	-81.143642	217.000000	8.000000	178.500000
max	99999.000000	71.706900	178.065000	281752.000000	34636.000000	211748.000000

In [76]:
```
#Para separar los datos se va a tomar la columna Pronvicia_State, los valores
#Virginia y California para realiza la correlacion
datosVirginia = datos[(datos.Province_State == 'Virginia')]
datosCalifornia = datos[(datos.Province_State == 'California')]

#Se ve los datos de vignina de la columna Province_State
datosVirginia.head()
```

Out[76]:

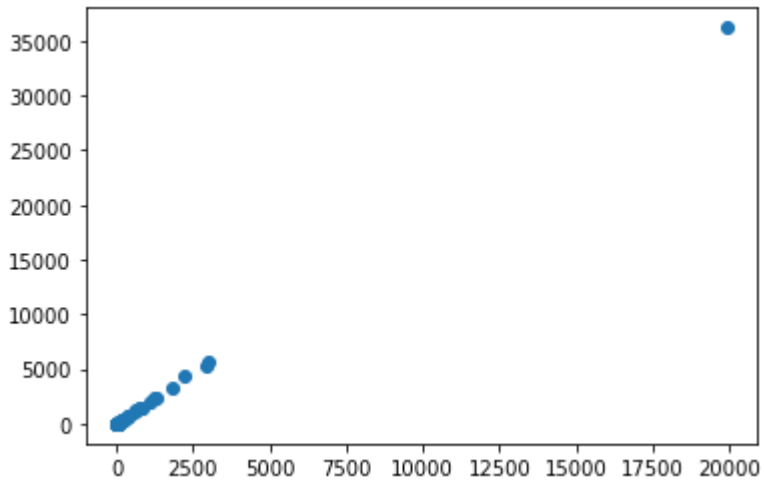
	FIPS	Province_State	Country_Region	Last_Update	Lat	Long_	Confirmed	D
2	51001.0	Virginia	US	5/18/2020 2:32	37.767072	-75.632346	688	
28	51003.0	Virginia	US	5/18/2020 2:32	38.020807	-78.554811	138	
33	51510.0	Virginia	US	5/18/2020 2:32	38.814003	-77.081831	1476	
40	51005.0	Virginia	US	5/18/2020 2:32	37.786361	-80.002225	7	
50	51007.0	Virginia	US	5/18/2020 2:32	37.340810	-77.985846	18	

In [77]:
```
#Se ve los datos de Californi de la columna Province_State
datosCalifornia.head()
```

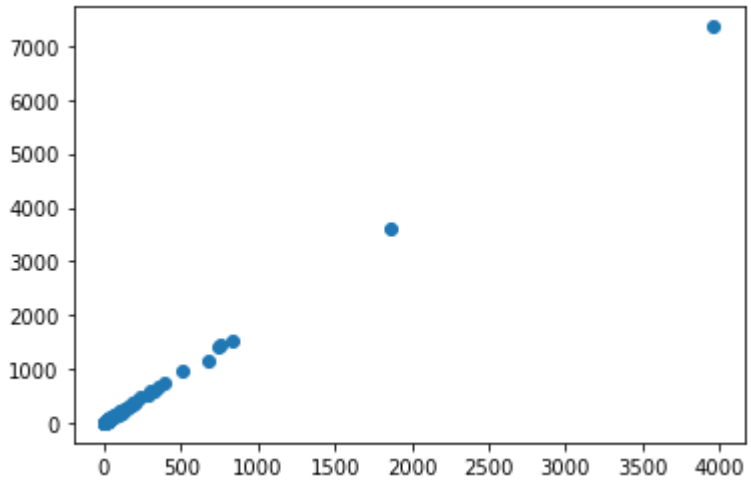
Out[77]:

	FIPS	Province_State	Country_Region	Last_Update	Lat	Long_	Confirmed	D
24	6001.0	California	US	5/18/2020 2:32	37.646294	-121.892927	2392	
48	6003.0	California	US	5/18/2020 2:32	38.596786	-119.822359	2	
49	6005.0	California	US	5/18/2020 2:32	38.445831	-120.656960	9	
306	6007.0	California	US	5/18/2020 2:32	39.667278	-121.600525	22	
313	6009.0	California	US	5/18/2020 2:32	38.205371	-120.552913	13	

In [96]:
```
"""
Se presenta una grafica respecto a los datos de covid en el estado de Californ
primer parametro los datos confirmados de casos, las muertes, luego se va a to
casos activos y se va a graficar
"""
datosObtenidosCalifornia = datosCalifornia[["Confirmed", "Deaths"]].mean(axis=
plt.scatter(datosObtenidosCalifornia, datosCalifornia["Active"])
plt.show()
```



In [97]:
```
"""
Se presenta una grafica respecto a los datos de covid en el estado de Virginia
primer parametro los datos confirmados de casos, las muertes, luego se va a to
casos activos y se va a graficar
"""
datosObtenidosVirginia = datosVirginia[["Confirmed", "Deaths"]].mean(axis=1)
plt.scatter(datosObtenidosVirginia, datosVirginia["Active"])
plt.show()
```



In [80]:
```
datos.corr(method='pearson')
```

Out[80]:

	FIPS	Lat	Long_	Confirmed	Deaths	Active
FIPS	1.000000	0.245692	-0.326275	-0.116693	-0.079414	-0.061950
Lat	0.245692	1.000000	-0.411557	-0.018917	0.020007	0.053804
Long_	-0.326275	-0.411557	1.000000	0.212000	0.134324	0.069030
Confirmed	-0.116693	-0.018917	0.212000	1.000000	0.809726	0.779910
Deaths	-0.079414	0.020007	0.134324	0.809726	1.000000	0.645875
Active	-0.061950	0.053804	0.069030	0.779910	0.645875	1.000000

In [81]:
```
#Se utiliza la correlacion para los datos de virginia
datosVirginia.corr(method='pearson')
```

Out[81]:

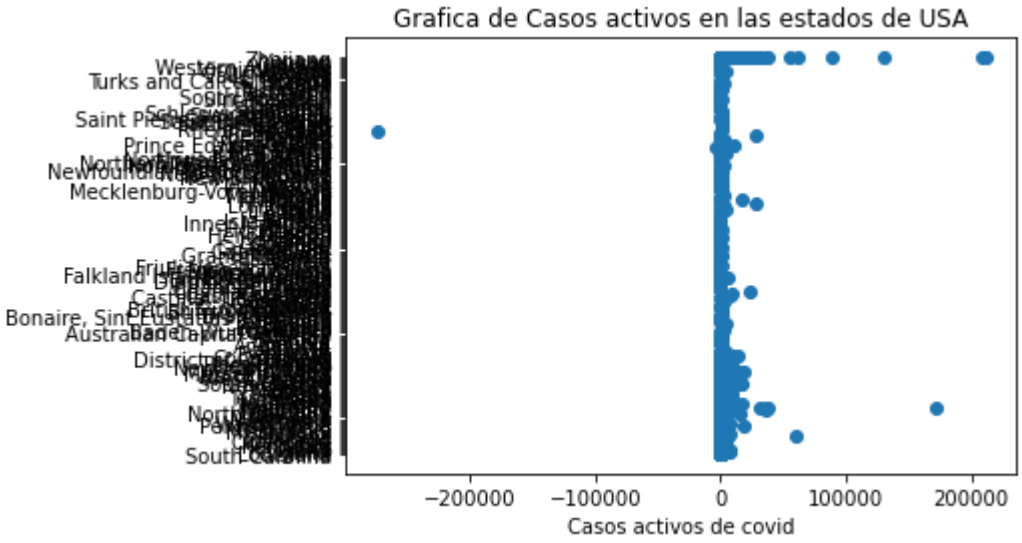
	FIPS	Lat	Long_	Confirmed	Deaths	Active
FIPS	1.000000	0.695533	-0.889928	-0.029846	-0.028965	-0.029815
Lat	0.695533	1.000000	-0.536918	0.199934	0.158149	0.201076
Long_	-0.889928	-0.536918	1.000000	0.104056	0.097264	0.104088
Confirmed	-0.029846	0.199934	0.104056	1.000000	0.944897	0.999924
Deaths	-0.028965	0.158149	0.097264	0.944897	1.000000	0.940794
Active	-0.029815	0.201076	0.104088	0.999924	0.940794	1.000000

In [82]:
```
#Se utiliza la correlacion para los datos de california
datosCalifornia.corr(method='pearson')
```

Out[82]:

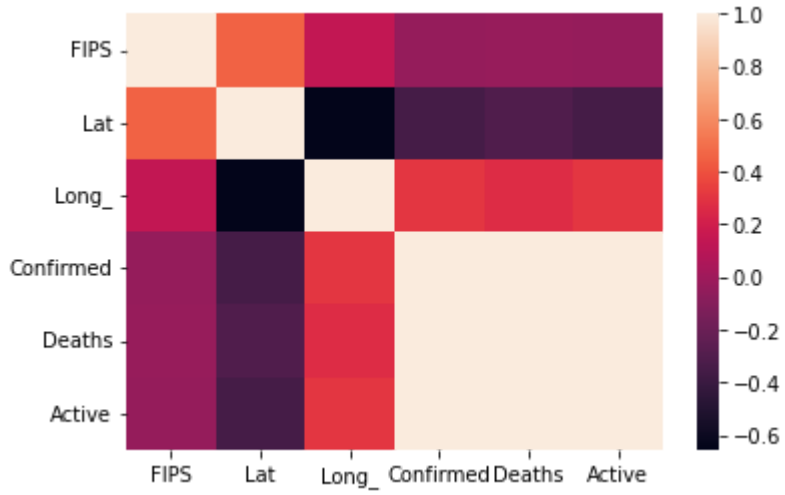
	FIPS	Lat	Long_	Confirmed	Deaths	Active
FIPS	1.000000	0.453286	0.135112	-0.037988	-0.032268	-0.038265
Lat	0.453286	1.000000	-0.656226	-0.355686	-0.308355	-0.357969
Long_	0.135112	-0.656226	1.000000	0.301422	0.268275	0.303009
Confirmed	-0.037988	-0.355686	0.301422	1.000000	0.995605	0.999989
Deaths	-0.032268	-0.308355	0.268275	0.995605	1.000000	0.995156
Active	-0.038265	-0.357969	0.303009	0.999989	0.995156	1.000000

In [83]:
```
plt.scatter(datos['Active'],datos['Province_State'])
plt.xlabel('Estados de USA')
plt.xlabel('Casos activos de covid')
plt.title('Grafica de Casos activos en las estados de USA')
plt.show()
```

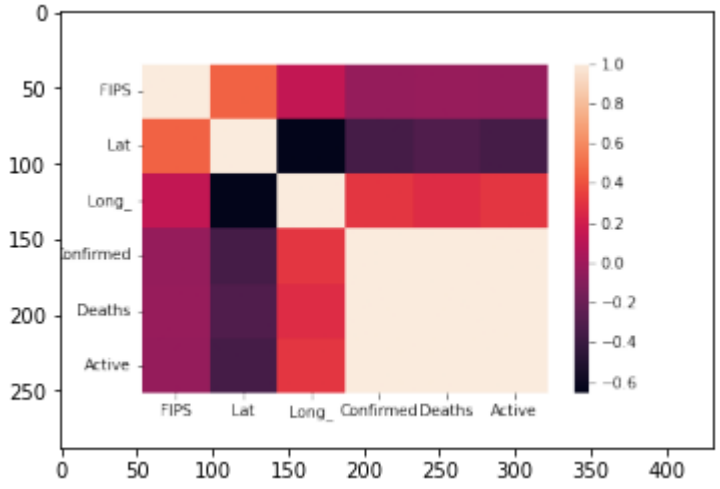


In [84]:
```
import seaborn as sns

californiaDF = datosCalifornia.drop('Province_State', axis=1)
virginiaDF = datosVirginia.drop('Province_State', axis=1)

californiaDF = californiaDF[californiaDF.columns]
virginiaDF = virginiaDF[virginiaDF.columns]
```

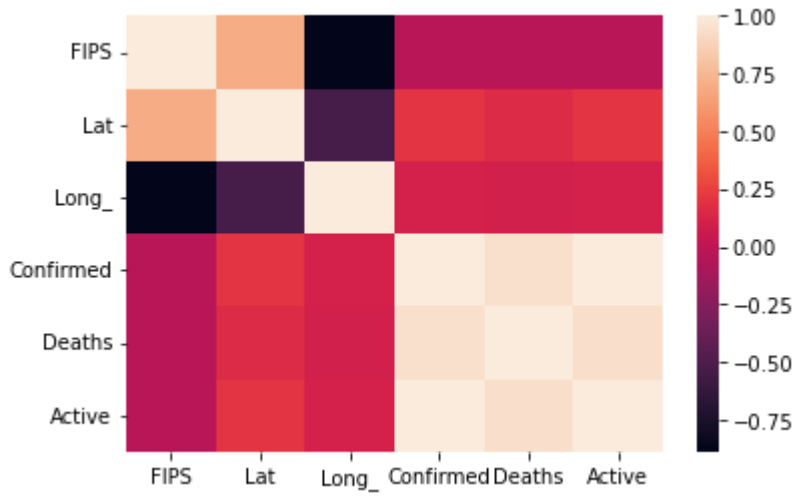
In [86]:
```
wcorr = californiaDF.corr()
sns.heatmap(wcorr)
plt.savefig('attribute_correlation_california.png')
```



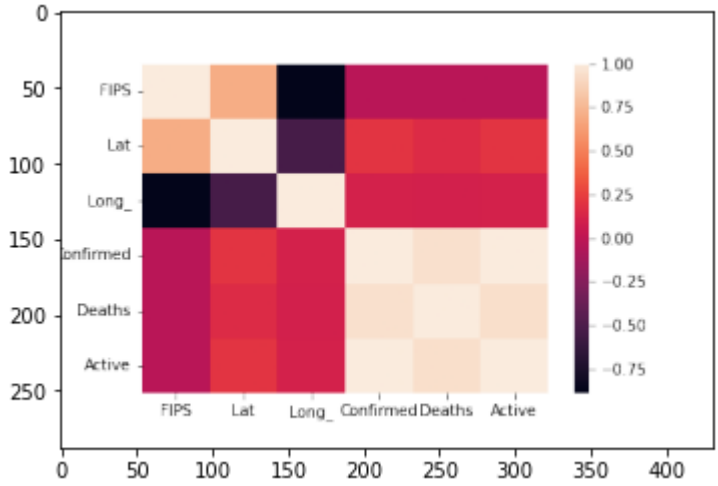
In [88]:
```
im = plt.imread('./attribute_correlation_california.png')
plt.imshow(im)
plt.show()
```



In [87]:
```
wcorr = virginiaDF.corr()
sns.heatmap(wcorr)
plt.savefig('attribute_correlation_virginia.png')
```



In [89]:
```
im = plt.imread('./attribute_correlation_virginia.png')
plt.imshow(im)
plt.show()
```



In [92]:
```
datos.to_csv('nuevo_database.csv')
```

In [93]:
```
!head nuevo_database.csv
```

```
,FIPS,Province_State,Country_Region,Last_Update,Lat,Long_,Confirmed,Deaths,Ac
tive,Combined_Key
0,45001.0,South Carolina,US,5/18/2020 2:32,34.22333378,-82.46170658,35,0,3
5,"Abbeville, South Carolina, US"
1,22001.0,Louisiana,US,5/18/2020 2:32,30.2950649,-92.41419698,198,12,186,"Aca
dia, Louisiana, US"
2,51001.0,Virginia,US,5/18/2020 2:32,37.76707161,-75.63234615,688,9,679,"Acco
mack, Virginia, US"
3,16001.0,Idaho,US,5/18/2020 2:32,43.4526575,-116.2415516,773,22,751,"Ada, Id
aho, US"
4,19001.0,Iowa,US,5/18/2020 2:32,41.33075609,-94.47105874,5,0,5,"Adair, Iowa,
US"
5,21001.0,Kentucky,US,5/18/2020 2:32,37.10459774,-85.28129668,92,14,78,"Adai
r, Kentucky, US"
6,29001.0,Missouri,US,5/18/2020 2:32,40.19058551,-92.60078167,26,0,26,"Adair,
Missouri, US"
7,40001.0,Oklahoma,US,5/18/2020 2:32,35.88494195,-94.65859267,75,3,72,"Adair,
Oklahoma, US"
8,8001.0,Colorado,US,5/18/2020 2:32,39.87432092,-104.3362578,2613,102,2511,"A
dams, Colorado, US"
```

2. Responda las siguientes preguntas:

1. Qué variables de su dataset tienen una correlación alta y porque cree que se da este fenómeno en su dataset?

En este caso las variables que mas van a tener relacion serán las de país con los casos activos ya que por ejemplo en un país como USA con una población de 331 millones de habitantes va a tener mas casos frente a un país como Ecuador que cuenta con 17 millones de habitantes.

2. Qué variables de su dataset tienen una correlación baj y porque cree que se da este fenómeno en su dataset?

En este caso las variables que van a tener una baja correlación serán las variables de latitud y longitud ya que estos datos son referente a la posición del país dentro del planeta tierra y no afectan nada a los casos confimados, muertes dentro del covid.

3. De las variables que tiene correlación se puede decir que una es causa de a otra?.

En el dataset presentado la variable Combined_Key(Llave Combinada), es una mezcla de la ciudad de un país y el paus en general, esto quiere decir que se combinan con el fin de mostrar la información mas commpleta de dicho campo.

****4. Suba el arvhivo ipynb y pdf por separado no comprimido.****