CSDS 313: Assignment 1

Shaochen (Henry) ZHONG, sxz517
Ningjia HUANG, nsh239

Due and submitted on 09/17/2020
Issued by Dr. Koyutürk

(1)

```
print(df.shape[0])
```

There are **11** columns and **38283** rows in the spreadsheet.

(2)

```
print(df['countriesAndTerritories'].nunique())
```

There are **210** countries in total. Note this number does include "countries and Territories" like Cases_on_an_international_conveyance_Japan. We decide to not to exclude it as there might be (and in fact, are) other listings under the column that are by definition not a country nor a territory. For the sake of consitency, we include everything listed in the data set.

```
print(df.dateRep.min())
```

The earliest date recorded is **12/31/2019**.

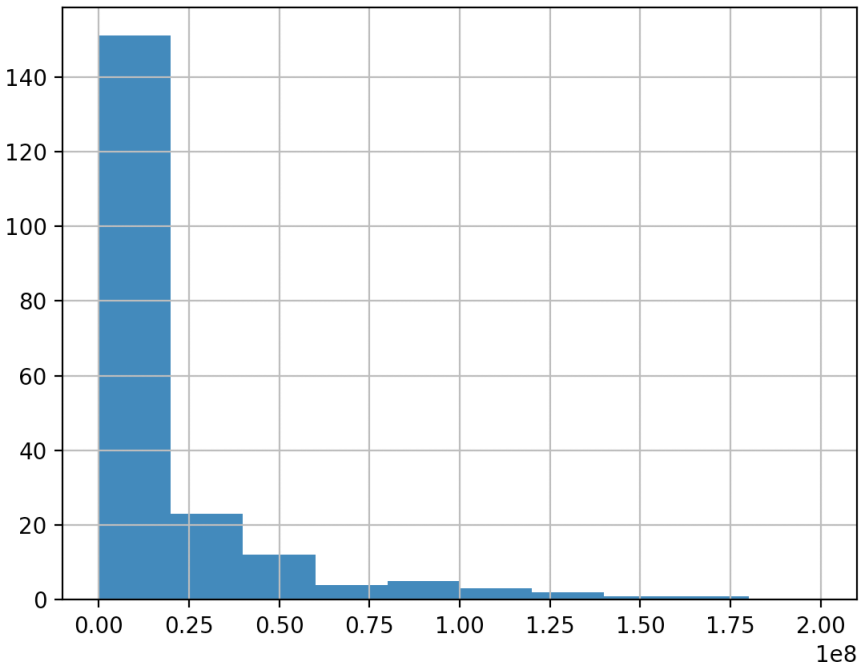
```
print(df.dateRep.max())
```

The latest date recorded is **8/24/2020**.

(3)

```
i = 0
countries = []
populations = []
for country in df['countriesAndTerritories']:
```

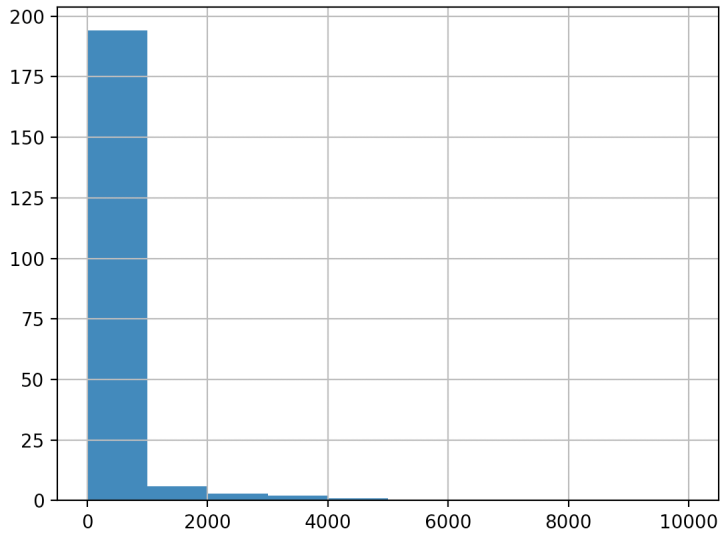
```
    if country not in countries:
        countries.append(country)
        pop = df.at[df['countriesAndTerritories'].eq(country).idxmax(),
        'popData2019'].astype('float')
        populations.append(pop)
        i += 1
df2 = pd.DataFrame({'country': countries, 'population': populations})
mean = df2['population'].mean()
print(mean)
std = df2['population'].std()
print(std)
df2['population'].hist(range = [0,200000000])
plt.show()
```



$\mu = 36694813.1722488$
$\sigma = 141822871.65179774$
About 150 out of 209 countries are in the first population bin. About 21 countries are in the second population bin. As we go through the x-axis, the number of countries within the ranges remains low. The number of countries drops drastically from the first bin to the second bin, which makes the distribution seems to be power-law distribution since very few countries contribute to a large percent of populations and most countries have relatively small population size.

(4)

```
mask = (df['dateRep'] == '2020-05-04')
df2 = df.loc[mask]
print(df2['cases'].median())
q1 = df2['cases'].quantile(0.25)
q3 = df2['cases'].quantile(0.75)
print(q3 - q1)
df2['cases'].hist(range = [0,10000])
plt.show()
```



$Median = 5.0$
$IQR = Q_3 - Q_1 = 90.75 - 0.00 = 90.75$ About 190 out of 209 countries are in the first cases reported bin. As we go through the x-axis, the number of reported cases decreases drastically. Since the first bin is significantly larger than the others, this distribution seems to be power-law distribution.

(5)

```
mask = (df['dateRep'] >= '2020-06-01') & (df['dateRep'] <= '2020-07-01')
df2 = df.loc[mask]
df2 = df2[['countriesAndTerritories','cases']]
df2 = df2.groupby(['countriesAndTerritories']).sum()
df2 = df2.sort_values('cases',ascending=False)
print(df2.head())
```

Brazil had the greatest increase in the number of cases from June 1st to July 1st. The increase is $903,601$ cases. Note we interpret this question as which country have the greatest number of new cases tested from June 1st to July 1st.

(6)

```
mask = (df['dateRep'] >= '2020-06-01') & (df['dateRep'] <= '2020-07-01')
df2 = df.loc[mask]
df2 = df2[['countriesAndTerritories','cases']]
df2 = df2.groupby(['countriesAndTerritories']).sum()
df2['cases'] = df2['cases']/31
df2 = df2.sort_values('cases',ascending=False)
print(df2)
```

Brazil had the greatest average increase in the number of cases per day from June 1st to July 1st. The average increase is $29,148.419355$.

(7)

```
mask = (df['dateRep'] >= '2020-06-01') & (df['dateRep'] <= '2020-07-01')
df2 = df.loc[mask]
df2 = df2[['countriesAndTerritories','cases','popData2019']]
df2 = df2.groupby(['countriesAndTerritories','popData2019'],as_index=False).sum()
df2["result"] = ""
df2['result'] = df2['cases']/df2['popData2019']*10000
df2 = df2.sort_values('result',ascending=False)
print(df2)
```

Qatar had the greatest increase in average cases per 10,000 people per day from June 1st to July 1st. The average is 144.15599.

(8)

```
df2 = df[['dateRep','countriesAndTerritories','cases','deaths']]
df2 = df2.groupby(['dateRep']).sum().sort_values('cases',ascending=False)
print(df2)
```

On **July 30th, 2020**, the world had the greatest number of reported cases($298,094$ cases).

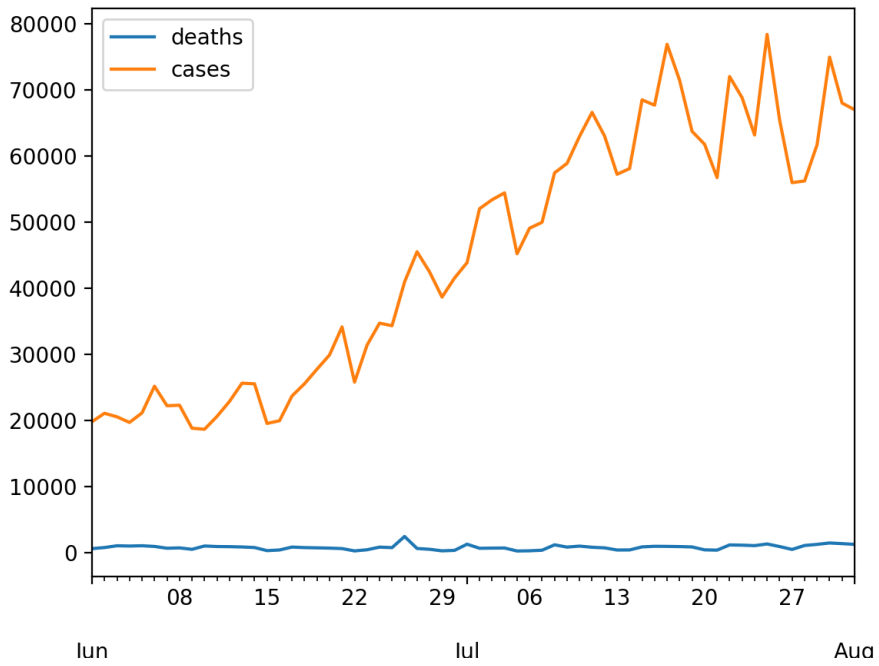
```
df2 = df2.sort_values('deaths',ascending=False)
print(df2)
```

On **April 16th, 2020**, the world had the greatest number of reported deaths($10,542$ cases).

(9)

```
df_holder = (df['dateRep'] >= '2020-06-01') & (df['dateRep'] <= '2020-08-01')
df_holder = df.loc[df_holder]
df_holder = df_holder[['dateRep','countriesAndTerritories','cases','deaths']]
df_us_case = df_holder.loc[df['countriesAndTerritories'] ==
    'United_States_of_America']

df_us_case.plot(x = 'dateRep', y = ['deaths', 'cases'])
plt.show()
```



It seems depite the significant increase on daily new cases, the amount of deaths – altough still sad to say – is actually rather steady on the graph. Which our first thought is the large Y-axis scale on `cases` makes growth of `deaths` hardly noticeable, but with close examination it is actually not the case. Thus, our guess it is that during June to Augest, the pressure on US medical system is not as hard, so that health care works can actually "control" the death of COVID-19 to a certain extent.

(10)

```
df_holder = (df['dateRep'] >= '2020-04-01') & (df['dateRep'] <= '2020-4-30')
df_holder = df.loc[df_holder]
df_holder = df_holder[['dateRep','countriesAndTerritories','cases','popData2019',
    'continentExp']]
```

```
df_case = df_holder.groupby(['countriesAndTerritories','popData2019'], as_index =
    False )['cases'].sum()

df_case['case_per_pop'] = df_case['cases'] / df_case['popData2019'] * 100
df_case = df_case.sort_values(by = ['case_per_pop'], ascending = False)
print(df_case.head())
```

We have `San Marino` to be the most infected among all country/territory during April interm of cases per population, as around 0.97% of its population was infected by new cases tested in April. Although this is mostly because `San Marino` has a relatively low population of `34453`.

(11)

```
df_holder = df[['dateRep', 'countriesAndTerritories','cases','popData2019',
    'continentExp']]
df_continent_case = df_holder.groupby(['continentExp'], as_index = False
    )['cases'].sum()

df_country_pop = df_holder.groupby(['countriesAndTerritories', 'continentExp'],
    as_index = False )['popData2019'].first()
df_continent_pop = df_country_pop.groupby(['continentExp'], as_index =
    False)['popData2019'].sum()
df_continent = pd.merge(df_continent_case, df_continent_pop, on = 'continentExp')
df_continent['case_per_10k'] = df_continent['cases'] / df_continent['popData2019'] *
    10000

df_continent = df_continent.sort_values(by = ['case_per_10k'], ascending = False)
df_continent_case = df_continent_case.sort_values(by = ['cases'], ascending = False)
print(df_continent_case.head())
print(df_continent.head())
```

We have `America` to be both the most infected continent (`12553567` cases) and the most infected continent per million (around `123.85` cases per 10,000 people).