CSDS 313: Assignment 1

Shaochen (Henry) ZHONG, sxz517
Ningjia Huang, nsh239

Due and submitted on 08/26/2020
CSDS 455, Dr. Connamacher

(1)

```
df = pd.read_excel('covid_data.xlsx')
```

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

(2)

```
L = []
i = 0
for country in df['countriesAndTerritories']:
if country not in L:
    L.append(country)
    print(country)
i += 1
print(i)
```

There are **209** countries in total.

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

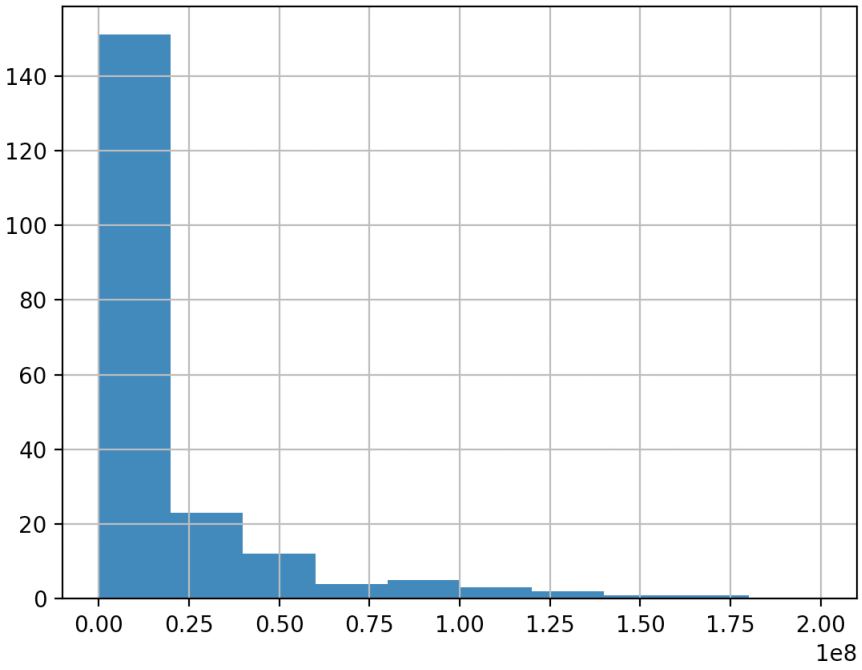
The earliest date recorded is **12/31/2019**.

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

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

(3)

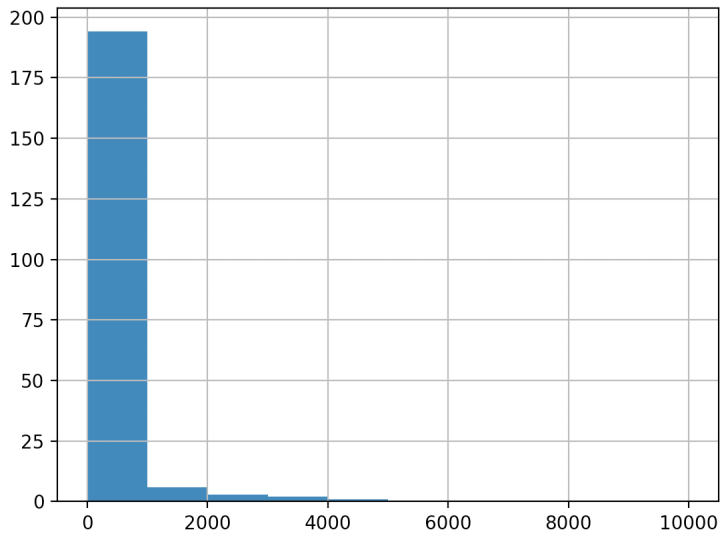
```
i = 0
countries = []
populations = []
for country in df['countriesAndTerritories']:
    if country not in countries:
        countries.append(country)
        pop = df.at[df['countriesAndTerritories'].eq(country).idxmax(),
        'popData2019'].astype('float')
        populations.append(pop)
        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.sort_values('cases',ascending=False)
print(df2)
```

Brazil had the greatest increase in the number of cases from June 1st to July 1st. The increase is $54,771$.

(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 29148.419355.

(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).