COVID-19: Exploratory Data Analysis of Coronavirus Outbreak

Opeoluwa Oluwatunmise Follow
Mar 14 · 10 min read

Opeoluwa Tunmise, Mutiu Samiyu, Taiwo Bashir, Amarachi Mbakwe

The outbreak of the Coronavirus Disease 2019 (**COVID-19**) began in Wuhan City, Hubei Province, China. The upsurge of COVID-19 is fast becoming a major global crisis, which made the World Health Organisation (WHO) declare the **COVID-19** as a pandemic.

We performed an exploratory data analysis on the **COVID-19** dataset and made inferences based on the dataset. The dataset used for our study is on John Hopkin GitHub repository; it consists of the cumulative number of confirmed cases, recovered cases, and death cases. It also includes the Province/State, Observation date (the date of observation), and Country of the infected patients as of March 11, 2020.

Let's import the COVID-19 dataset and the needed python libraries:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import plotly.graph_objects as go
import plotly.express as px
import seaborn as sns

sns.set()
sns.set_style("darkgrid")
sns.despine()

Corona_Data =
pd.read_csv(r"C:\Users\ooluw\OneDrive\Desktop\Corona\covid_19_data.cs
v", parse_dates = ['ObservationDate', 'Last Update'])
```

Let's print the first five rows of the dataset:

```
Corona_Data.head()
```

	SNo	ObservationDate	Province/State	Country/Region	Last Update	Confirmed	Deaths	Recovered
0	1	2020-01-22	Anhui	Mainland China	2020-01-22 17:00:00	1	0	0
1	2	2020-01-22	Beijing	Mainland China	2020-01-22 17:00:00	14	0	0
2	3	2020-01-22	Chongqing	Mainland China	2020-01-22 17:00:00	6	0	0
3	4	2020-01-22	Fujian	Mainland China	2020-01-22 17:00:00	1	0	0
4	5	2020-01-22	Gansu	Mainland China	2020-01-22 17:00:00	0	0	0

Then, we performed a data cleaning process on the COVID-19 dataset, which involves dropping irrelevant columns and removing missing values or replacing them computationally. After that, we manipulated the data to know the cumulative number of cases of patients that are still infected, while some are dead or have recovered.

Hence, we have an updated dataset:

Since we now have an updated dataset, we can begin the data visualization process, which would give us better insights into the COVID-19 dataset.

Let's group the cases based on Observation date:

```
Corona_DateBasis = Corona_Cleaned_Data.groupby('ObservationDate')
['Confirmed','Deaths', 'Recovered', 'Still_Infected'].sum()
Corona_DateBasis = Corona_DateBasis.reset_index()
Corona_DateBasis = Corona_DateBasis.sort_values('ObservationDate',
```

```
ascending = False)
Corona_DateBasis.head().style.background_gradient(cmap = 'PRGn')
```

From the above table, we can infer that in nearly three months, the spread of the virus rose from 555 confirmed cases on January 22, 2020, to 125,865 cases as of March 11, 2020. If the range continues in this proportion, then by May 11, 2020, we would have at least 225, 865 confirmed cases. That's preposterous.

We can also use the line and scatter plot to get a more in-depth insight into the COVID-19 dataset.

We can see from the graph plotted above that the number of confirmed cases keeps increasing, thereby increasing the number of death cases, the number of recovered instances, and the number of patients still infected. Fortunately, there is a higher likelihood of recovering than dying as a result of COVID-19.

Besides, we can also see that between February 16, 2020, and March 8, 2020, there was a decline in the recovery rate of infected patients. We may recommend that the government of the affected nations need to put measures in place to make sure that there is a significant increase in the number of recovered patients over a very long time frame.

We added another column to our COVID-19 dataset from another dataset.

```
country_continent =
pd.read_csv(r"C:\Users\ooluw\OneDrive\Desktop\countryContinent.csv",
encoding="iso-8859-1")
country_continent = country_continent[["country", "continent"]]

country_continent.rename(columns={"country": "Country", "continent":
"Continent"}, inplace=True)
Added_continents_info = pd.DataFrame([["Macau", "Asia"], ["Ivory
Coast", "Africa"], ["North Ireland", "Europe"], ["North Macedonia",
"Europe"], ["UK", "Europe"], ["Iran", "Asia"], [" Azerbaijan",
"Asia"], ["Others", "Other"], ["Russia", "Europe"], ["Taiwan",
"Asia"], ["US", "Americas"], ["South Korea", "Asia"], ["Vietnam",
"Asia"]], columns=["Country", "Continent"])

country_continent = country_continent.append(Added_continents_info)

Corona_Cleaned_Data = pd.merge(Corona_Cleaned_Data,
country_continent, on="Country", how="left")

Corona_Cleaned_Data.info()
```

Let’s group the Confirmed and Deaths cases based on Continents:

```
latest = Corona_Cleaned_Data.groupby(["Province/State",
"Country"]).last().reset_index().copy()
first = Corona_Cleaned_Data.loc[Corona_Cleaned_Data["Confirmed"] > 0,
["ObservationDate", "Country",
"Confirmed"]].groupby(["Country"]).first().reset_index().copy()
global_daily =
Corona_Cleaned_Data.groupby("ObservationDate").sum().reset_index().co
py()

cmap = plt.get_cmap("tab20")
colors_dict = {}
colors_dict["Confirmed"] = cmap(2)
colors_dict["Still_Infected"] = cmap(0)
colors_dict["Deaths"] = cmap(6)
colors_dict["DeathRatio"] = cmap(6)
colors_dict["Recovered"] = cmap(4)

#print(colors_dict)

continents = latest.groupby(["Continent"]).sum().reset_index().copy()

fig, axes = plt.subplots(1, 2)
fig.set_size_inches(11,5)

colors = cmap(np.array([2, 4, 6, 0, 19, 15]))

axes[0].pie(continents["Confirmed"], startangle=90, radius=1,
colors=colors, wedgeprops=dict(width=0.3, edgecolor='w'))
axes[0].set_title('Confirmed cases per continent', fontsize=17)

axes[1].pie(continents["Deaths"], startangle=90, radius=1,
colors=colors, wedgeprops=dict(width=0.3, edgecolor='w'))
axes[1].set_title('Deaths per continent', fontsize=17)

handles = axes[1].get_legend_handles_labels()

fig.legend(labels = continents["Continent"].unique(), loc='center',
frameon=False)

plt.show()
```

It's no surprise that the continent of Asia has more confirmed and death cases than any other continent, which we think is as a result of the outbreak that started in Wuhan City, China, a country in the continent of Asia, and how fast COVID-19 spreads. We also noticed that the spread of COVID-19 is relatively low in the continent of Africa. The question that comes to mind is, why?

We can also deduce from the graph above that COVID-19 is spreading exponentially in the continent of Europe, which puts the citizenry in danger. Again, the question that comes to mind is, why?

Let's now consider the state of patients in different regions?

```
conditions = ["Recovered", "Still_Infected", "Deaths"]

hubei_cases = latest.loc[ latest["Province/State"] == "Hubei",
conditions ].sum()
china_cases = latest.loc[ latest["Country"] == "China", conditions
].sum()
USA_cases = latest.loc[ latest["Country"] == "US", conditions ].sum()
asia_cases = latest.loc[ (latest["Continent"] == "Asia") &
(latest["Country"] != "China"), conditions ].sum()
europe_cases = latest.loc[ latest["Continent"] == "Europe",
conditions ].sum()
nonchina_cases = latest.loc[ latest["Country"] != "China", conditions
].sum()
world_cases = latest[ conditions ].sum()

cases = [hubei_cases, europe_cases, nonchina_cases, china_cases,
USA_cases, asia_cases, world_cases]
titles = ["Hubei", "Europe", "Non China", "China","US", "Asia without
China", "World"]

fig, axes = plt.subplots(2, 3)
fig.set_size_inches(16,8)

colors = cmap(np.array([4, 0, 6]))
```

```
for i in np.arange(2):
    for j in np.arange(3):
        labels = []
        freq = round(cases[3*i+j] * 100 / cases[3*i+j].sum(),1)
        for c in conditions:
            labels.append(c + " (" + str(freq[c]) + "%)")

        axes[i, j].pie(cases[3*i+j], labels=labels, startangle=90,
radius=1, colors=colors, wedgeprops=dict(width=0.3, edgecolor='w'))
        axes[i, j].set_title(titles[3*i+j], fontsize=17)

fig.suptitle("State of patients in different regions", fontsize=20)

plt.show()
```


. . .

From the plots above, we can observe that recovery rates in China (76.1%) and Hubei (72.5%) are relatively higher than those in other regions. We can generally infer that recovery rates are higher than death rates except in the United States. The United States recorded a 3.0% death rate, which exceeds their recovery rate of 1.1%.

Why do we have a higher death rate (3%) compared to the recovery rate (1.1%) in the United States?

Daily Confirmed Cases; Globally

```
world_new =
Corona_Cleaned_Data.groupby("ObservationDate").sum().reset_index().se
t_index("ObservationDate").copy()  world_new =
world_new["Confirmed"].diff().dropna()  plt.figure(figsize=(15,5)) ax
= sns.lineplot(x=world_new.index, y=world_new.values, marker="o",
color=colors_dict["Confirmed"])  sns.despine()
plt.xticks(rotation='vertical') ax.set_ylim(0, world_new.max()*1.1)
ax.set_xlabel("") ax.set_ylabel("") ax.set_title("New confirmed cases
by day", fontsize=20) ax.grid(False, axis="x")
```


From the time plot above, it shows that the highest number of new cases occurred between February 11, 2020, and February 15, 2020, with almost 16000 instances. Starting from February 23, 2020, the number of new cases has been growing exponentially.

Daily Confirmed Cases; China

```
china_new = Corona_Cleaned_Data[ Corona_Cleaned_Data["Country"] ==
"China"
].groupby("ObservationDate").sum().reset_index().set_index("Observati
onDate").copy()

china_new = china_new["Confirmed"].diff().dropna()

plt.figure(figsize=(15,5))
ax = sns.lineplot(x=china_new.index, y=china_new.values, marker="o",
color=colors_dict["Confirmed"])
```

```
sns.despine()
plt.xticks(rotation='vertical')
ax.set_ylim(0, china_new.max()*1.1)
ax.set_xlabel("")
ax.set_ylabel("")
ax.set_title("New confirmed cases in China by day", fontsize=20)
ax.grid(False, axis="x")
```


This plot also shows a spike of new cases in China between February 11, 2020, and February 15, 2020, which correlates with the previous graph of global new confirmed cases. However, after the upsurge, there is a decline in the number of new cases of the virus in China.
What is responsible for the decline?

Daily Confirmed Cases; Outside China

```
nonchina_new = Corona_Cleaned_Data[ Corona_Cleaned_Data["Country"] !=
"China"
].groupby("ObservationDate").sum().reset_index().set_index("Observati
onDate").copy()

nonchina_new = nonchina_new["Confirmed"].diff().dropna()

plt.figure(figsize=(15,5))
ax = sns.lineplot(x=nonchina_new.index, y=nonchina_new.values,
marker="o", color=colors_dict["Confirmed"])

sns.despine()
plt.xticks(rotation='vertical')
```

```
ax.set_ylim(0, china_new.max()*1.1)
ax.set_xlabel("")
ax.set_ylabel("")
ax.set_title("New confirmed cases outside China by day", fontsize=20)
ax.grid(False, axis="x")
```


There was no significant growth of new cases of the virus outside China until February 23, 2020. Hereafter, there was a steady growth till early March, followed by a rapid increase after March 8, 2020.
Which Country outside China is much responsible for the sharp growth?

Countries With Highest Cases; Outside China

```
top_10 = latest[ latest["Country"] != "China"
].groupby("Country").sum().sort_values("Confirmed",
ascending=False).reset_index().head(10).copy()  top_10.head()
top_10_wide = pd.melt(top_10[["Country", "Recovered", "Deaths",
"Still_Infected"]], id_vars=["Country"], value_vars=
["Still_Infected","Recovered","Deaths"])  top_10_wide.head()
```

Frequency table of Outside China Cases

```
plt.figure(figsize=(15,8)) ax = sns.barplot(y="Country", x="value",
hue="variable", data=top_10_wide, palette=colors_dict)
ax.set_xlabel("") ax.set_ylabel("") ax.set_title("Countries with
highest number of cases Outside China", fontsize=20)
ax.legend(frameon=False).set_title('')
```

Italy has the most significant number of people still infected with COVID-19, and the highest number of deaths, with approximately equal mortality and recovery rates. Iran showed the most recovery cases, but it has a higher number of still infected cases. The number of infected cases in the US and Germany (France recorded approximately doubled instances) are almost equal. Meanwhile, the US developed more deaths records as compared to Germany.

New Cases as of March 11

```
confirmed_wide = Corona_Cleaned_Data[["ObservationDate", "Confirmed",
"Province/State", "Country"]].groupby(["Country",
"ObservationDate"]).sum().reset_index().copy()  confirmed_wide =
pd.pivot(confirmed_wide, columns="Country", index="ObservationDate",
```

```
values="Confirmed").fillna(0.0)  confirmed_wide =
confirmed_wide.diff().transpose().iloc[:,-1].sort_values(ascending=Fa
lse).head(10)  plt.figure(figsize=(15,8)) ax =
sns.barplot(x=confirmed_wide.values, y=confirmed_wide.index,
palette=sns.color_palette("GnBu_d", 10))  ax.set_xlabel("")
ax.set_ylabel("") ax.set_title("New Cases as of 11th March, 2020",
fontsize=20)  ax.legend(frameon=False).set_title('')
```

The plot above showed the trend of the new cases of COVID-19 in some majorly affected countries globally. Based on those greatly affected countries we considered in this research, Italy has the highest number of new cases with over 2000, while Denmark has a lower number of recorded new instances. We can carefully observe that China has no significant number of new cases here; this means that recently, there were very small or no confirmed new cases in China.

Spread of COVID-19 in China

```
china = Corona_Cleaned_Data[ Corona_Cleaned_Data["Country"] ==
"China" ].groupby("ObservationDate").sum().reset_index().copy()
china_long = pd.melt(china[["ObservationDate", "Confirmed", "Deaths",
"Still_Infected", "Recovered"]], id_vars=["ObservationDate"],
value_vars=["Confirmed","Deaths", "Still_Infected", "Recovered"])
plt.figure(figsize=(15,5)) ax = sns.lineplot(x="ObservationDate",
y="value", hue="variable", data=china_long, marker="o",
palette=colors_dict) sns.despine() plt.xticks(rotation='vertical')
```

```
ax.set_ylim(0, china["Confirmed"].max()*1.1) ax.set_xlabel("")
ax.set_ylabel("") ax.set_title("Spread of COVID-19 in China",
fontsize=20) ax.legend(labels=["Confirmed","Deaths",
"Still_Infected", "Recovered"], frameon=False) ax.grid(False,
axis="x")
```


Admittedly, there are more confirmed cases in China until March 1, 2020, when the growth remains steady. The Recovery cases are strictly increasing while the infected cases are declining. The death cases also remain stable in China over this period.

Spread of COVID-19 in Italy

```
italy = Corona_Cleaned_Data[ Corona_Cleaned_Data["Country"] ==
"Italy" ].reset_index().copy()

italy_long = pd.melt(italy[["ObservationDate", "Confirmed", "Deaths",
"Still_Infected", "Recovered"]], id_vars=["ObservationDate"],
value_vars=["Confirmed","Deaths", "Still_Infected", "Recovered"])

plt.figure(figsize=(15,5))
ax = sns.lineplot(x="ObservationDate", y="value", hue="variable",
data=italy_long, marker="o", palette=colors_dict)
sns.despine()
plt.xticks(rotation='vertical')
ax.set_ylim(0, italy["Confirmed"].max()*1.1)
ax.set_xlabel("")
ax.set_ylabel("")
ax.set_title("Spread of COVID-19 in Italy", fontsize=20)
ax.legend(labels=["Confirmed","Deaths", "Still_Infected",
```

```
"Recovered"], frameon=False)
ax.grid(False, axis="x")
```


There is an exponential growth of confirmed cases in Italy from February 22, 2020. While the recovery rates and the mortality rates are on the same growth level, there are more infected instances recorded.

Spread of COVID-19 in Iran

```
iran = Corona_Cleaned_Data[ Corona_Cleaned_Data["Country"] == "Iran"
].groupby("ObservationDate").sum().reset_index().copy()  iran_long =
pd.melt(iran[["ObservationDate", "Confirmed", "Deaths",
"Still_Infected", "Recovered"]], id_vars=["ObservationDate"],
value_vars=["Confirmed","Deaths", "Still_Infected", "Recovered"])
plt.figure(figsize=(15,5)) ax = sns.lineplot(x="ObservationDate",
y="value", hue="variable", data=iran_long, marker="o",
palette=colors_dict) sns.despine() plt.xticks(rotation='vertical')
ax.set_ylim(0, iran["Confirmed"].max()*1.1) ax.set_xlabel("")
ax.set_ylabel("") ax.set_title("Spread of COVID-19 in Iran",
fontsize=20) ax.legend(labels=["Confirmed","Deaths",
"Still_Infected", "Recovered"], frameon=False) ax.grid(False,
axis="x")
```

There is a rapid growth in the number of confirmed cases in Iran. Both the recovered and still infected cases showed a similar increase, but there are more still infected cases than recovered cases. The death cases recorded have been growing since early March.

Spread of COVID-19 in South Korea

```
korea = Corona_Cleaned_Data[ Corona_Cleaned_Data["Country"] == "South
Korea" ].groupby("ObservationDate").sum().reset_index().copy()
korea_long = pd.melt(korea[["ObservationDate", "Confirmed", "Deaths",
"Still_Infected", "Recovered"]], id_vars=["ObservationDate"],
value_vars=["Confirmed","Deaths", "Still_Infected", "Recovered"])
plt.figure(figsize=(15,5)) ax = sns.lineplot(x="ObservationDate",
y="value", hue="variable", data=korea_long, marker="o",
palette=colors_dict) sns.despine() plt.xticks(rotation='vertical')
ax.set_ylim(0, korea["Confirmed"].max()*1.1) ax.set_xlabel("")
ax.set_ylabel("") ax.set_title("Spread of COVID-19 in South Korea",
fontsize=20) ax.legend(labels=["Confirmed","Deaths",
"Still_Infected", "Recovered"], frameon=False) ax.grid(False,
axis="x")
```

In South Korea, there is a sharp growth in the number of confirmed cases (almost 8000) with insignificant recovery cases (below 1000) and death cases. Meanwhile, we observed that there are more infected cases.

Spread of COVID-19 in the United States

```
US = Corona_Cleaned_Data[ Corona_Cleaned_Data["Country"] == "US"
].groupby("ObservationDate").sum().reset_index().copy()

US_long = pd.melt(US[["ObservationDate", "Confirmed", "Deaths",
"Still_Infected", "Recovered"]], id_vars=["ObservationDate"],
value_vars=["Confirmed","Deaths", "Still_Infected", "Recovered"])

plt.figure(figsize=(15,5))
ax = sns.lineplot(x="ObservationDate", y="value", hue="variable",
data=US_long, marker="o", palette=colors_dict)
sns.despine()
plt.xticks(rotation='vertical')
ax.set_ylim(0, US["Confirmed"].max()*1.1)
ax.set_xlabel("")
ax.set_ylabel("")
ax.set_title("Spread of COVID-19 in United States", fontsize=20)
ax.legend(labels=["Confirmed","Deaths", "Still_Infected",
"Recovered"], frameon=False)
ax.grid(False, axis="x"
```


Over time, there have been a minimal number of confirmed cases in the United States until late February, 2020, where we observed a sudden growth. The recorded confirmed

cases increased significantly from that period till the period of this analysis. The death cases documented, as shown in the plot above, are increasing with fewer recovery cases.

Global Mortality Rates For Confirmed Cases

```
mortality = global_daily[["ObservationDate", "Confirmed",
"Deaths"]].copy()
mortality["DeathRatio"] = (mortality["Deaths"] /
mortality["Confirmed"]) * 100

plt.figure(figsize=(15,5))
ax = sns.lineplot(x="ObservationDate", y="DeathRatio",
data=mortality, marker="o", palette=colors_dict)
sns.despine()
plt.xticks(rotation='vertical')
ax.set_ylim(0, mortality["DeathRatio"].max()*1.1)
ax.set_xlabel("")
ax.set_ylabel("")
ax.set_title("Global Mortality Rates with respect to Confirmed Cases
[%]", fontsize=20)
ax.grid(False, axis="x")
```


The plot above shows the global mortality trend based on the confirmed cases of COVID-19 as at the period of this analysis. We can observe that COVID-19 has caused more death between mid-February and early March.

Further Research:

- Analysis based on gender and ages of COVID-19 patients
- Survival Analysis of COVID-19

Coronavirus Covid 19 Data Visualization Data Analysis Pandemic