Submitted by

Mujtaba & Zartashia

Cause of death in Brazil, 2019-2020-Report

Abstract: Data was obtained from kaggle. We used python libraries and visual studio code for exploratory data analysis. We concluded that "SP" state was most stuck by pandemic during year 2019-2020. Similarly, male were more prone to deaths and among other colors, White people were most affected while Indigenous people were least.The highest ratio of "others" in "cause" column signifies the need of more clarification in dataset.

1- Objective

The objective of this project is to create a data analysis that will help us to understand the number and causes of deaths while comparing people of different states, age, color and gender.

2- Data Acquisition and hypothesis

Data about the cause of death in Brazil within the time frame of 2019 to 2020 was obtained from Kaggle. Brazilian registry offices officially collected data after Corona Pandemic in order to visualize causes of death and influence with respect to age and skin color. In this exploratory data analysis we worked on the following hypothetical questions.

1. To check the influence of coronavirus on the deaths of a specific race and color.
2. Find out about the state having the most mortality rate in case of coronavirus attack and find out the most common cause of mortality.
3. Data visualization for a better understanding of trends in each category with respect to age, race, state, gender, and color and shed a light on the possible causes.

3- Potential Scope: These insights can actually help out to increase the Healthcare facilities in particular to any category of state, age, gender, or color and even direct us to expand the budget for medication and treatment of a particular mortality cause.

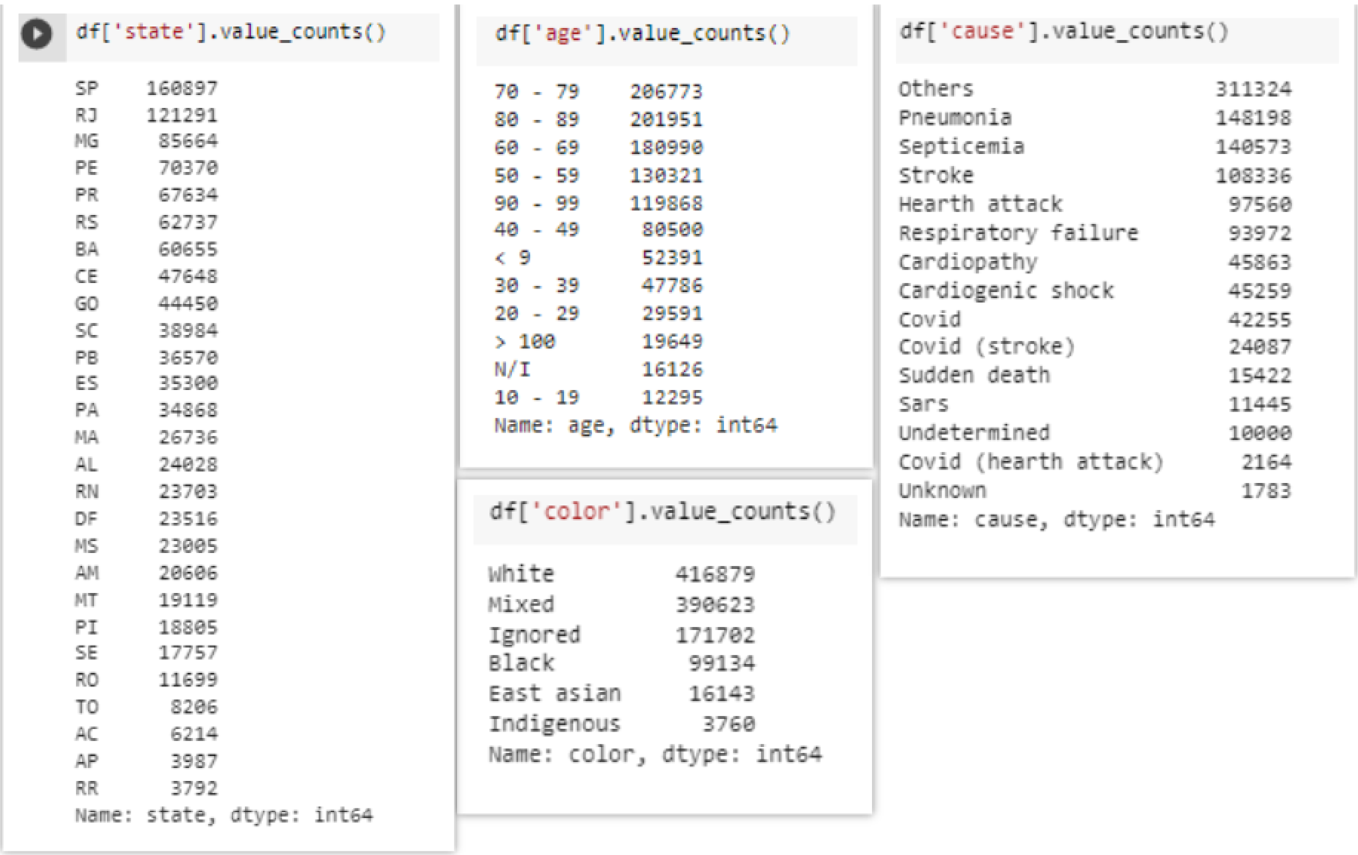
4-Bird Eye View of data

1. To have a glimpse of data we used pandas function and got shape and info. Dataset have 1098241 number of rows and 7 number of Columns namely "date","state" (27 Unique values), "gender" (2 unique values), "age" (12 unique values), "color" (6 unique values), "cause" (15 unique values) and "total" (43 unique values).
2. The colour column was categorized based on the fact that in Brazil, pardo is a race/skin colour category used by the Brazilian Institute of Geography and Statistics (IBGE). General categories include: branco ("**White**"), preto ("**Black**"), amarelo ("yellow", meaning **East Asians**) and indígena ("indigene" or "**indigenous** person", meaning Amerindians
3. Only the "total" column is numeric rest of the column are of object type. So we found the need to type caste and split variables to get maximum out of the data. We used `isnull.sum( )` and found that

data is clean and have no null values in it.

5- Unique values

We figured the data values trend in each column by getting the value counts of each unique value of eah column in descending order. The summary stats of datatset only got us descriptive statistics of "total" column since it was the only numerical column so we procedeed with type casting.



```
df['state'].value_counts()

SP    160897
RJ    121291
MG     85664
PE     70370
PR     67634
RS     62737
BA     60655
CE     47648
GO     44450
SC     38984
PB     36570
ES     35300
PA     34868
MA     26736
AL     24028
RN     23703
DF     23516
MS     23005
AM     20606
MT     19119
PI     18805
SE     17757
RO     11699
TO      8206
AC      6214
AP      3987
RR      3792
Name: state, dtype: int64
```

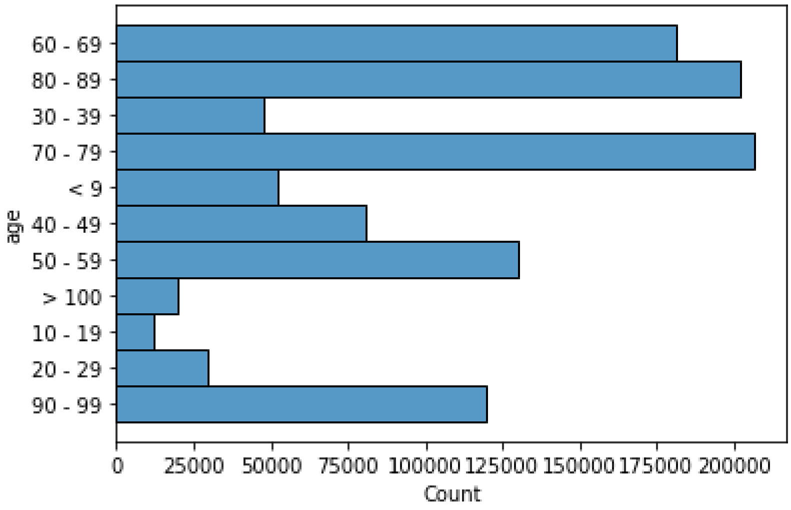
```
df['age'].value_counts()

70 - 79     206773
80 - 89     201951
60 - 69     180990
50 - 59     130321
90 - 99     119868
40 - 49      80500
< 9          52391
30 - 39      47786
20 - 29      29591
> 100        19649
N/I          16126
10 - 19      12295
Name: age, dtype: int64
```

```
df['color'].value_counts()

White          416879
Mixed          390623
Ignored        171702
Black           99134
East asian      16143
Indigenous       3760
Name: color, dtype: int64
```

```
df['cause'].value_counts()

Others                     311324
Pneumonia                  148198
Septicemia                 140573
Stroke                     108336
Hearth attack               97560
Respiratory failure         93972
Cardiopathy                 45863
Cardiogenic shock           45259
Covid                       42255
Covid (stroke)              24087
Sudden death                15422
Sars                        11445
Undetermined                10000
Covid (hearth attack)        2164
Unknown                      1783
Name: cause, dtype: int64
```

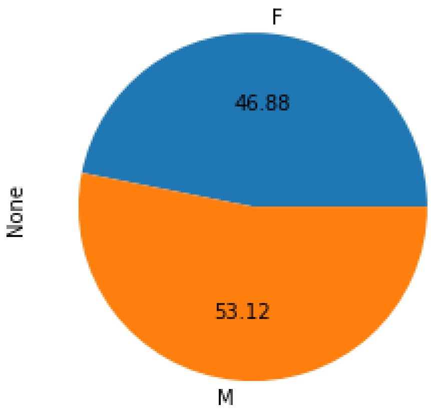
6- Treating "age" column

The countplot of "age" column showed the least impact of a category named "N/I" i.e., "Not/Identified" so we dropped those values from column to cut the clutter and get more sense of our data. Histplot also showed a largely varying distrubution of data which we can expect from a time-series dataset.



7- Understanding "gender "column

Piechart of gender showed that almost 46.88 percent females got died while 53.12 percent males died as a result of corona virus in 2019-2020.



8- Splitting date column

We splitted the date column into "year_of_death", "month_of_death", "day" and then dropped "month_of_death"and "day".

`dtypes` check again indicated the presence of all object type column except "total".So we type casted two columns year_of_death and month_of_death from object to int.

9- Encoding data

We encoded gender column and created a new column with their respective encoded values i.e., 'M': 0, 'F': 1.

After splitting and type casting we got 4 numerical values i.e., total (int64), year_of_death (int32), month_of_death (int32) and gender_in_number (int64)

For a better representation of data we grouped numeric values together and repeated the same for object type.

We used replace function to better understand the trend in age column. Originally there are categories of age range to make it more meaningful we changed "< 9", '10 - 19','20 - 29', "30 - 39","40 - 49", '50 - 59', '60 - 69', '70 - 79', '80 - 89', '90 - 99', "> 100" into "kids",'teens','adults', "adults", "old", "old", 'old', 'old', 'old', 'old', 'old'.

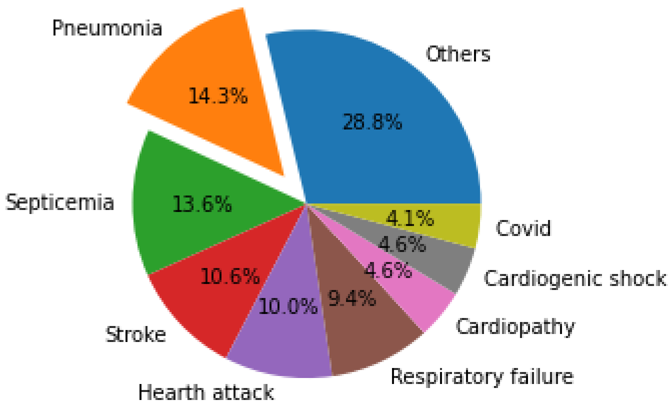
Histplot of 'month_of_death' showed a display of mortality rate in each month. Boxplot of "total" column actually depicted the presence of unique values with respect to each category. It is fair that If we see the the rows of datset, "total" is infact the count of each row telling the occurrence in that particular category of 'date', 'age', 'state', 'colour' ,and 'cause'.

Conclusions

Data filtering helped us to get meaningful answers for our hypothetical questions.

Top 8 causes of death

```
label = ["Others","Pneumonia","Septicemia","Stroke","Hearth attack", "Respiratory
failure", "Cardiopathy", "Cardiogenic shock", "Covid "]

# pie chart

plt.pie(big_causes, labels=label, autopct='%1.1f%%', explode=
(0,0.2,0,0,0,0,0,0,0))
```



From pie chart we can see that top 8 causes of death were

> Others>Pneumonia (1982) > Septicemia(1893) > Stroke(1476) > Hearth attack(1394) > Respiratory failure (1303) > Covid (638)> Cardiopathy(632), Cardiogenic shock

State wise mortality rate

Top 5 states most stuck with pandemic were:

```
plt.pie(big_causes, labels=label, autopct='%1.1f%%', explode=
(0,0.2,0,0,0,0,0,0,0))df.groupby('state').total.sum().sort_values(ascending=False)
.head(5)
```

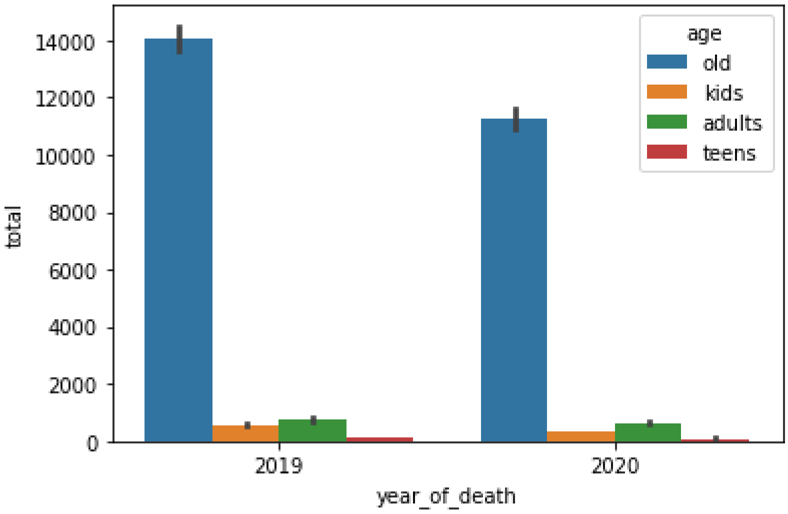
Insight:

> SP (7365)> RJ (3425)> MG (2683) >RS (1953) > PR (1580)

Year wise data insights

2019 has the highest number of deaths (grouping by age).

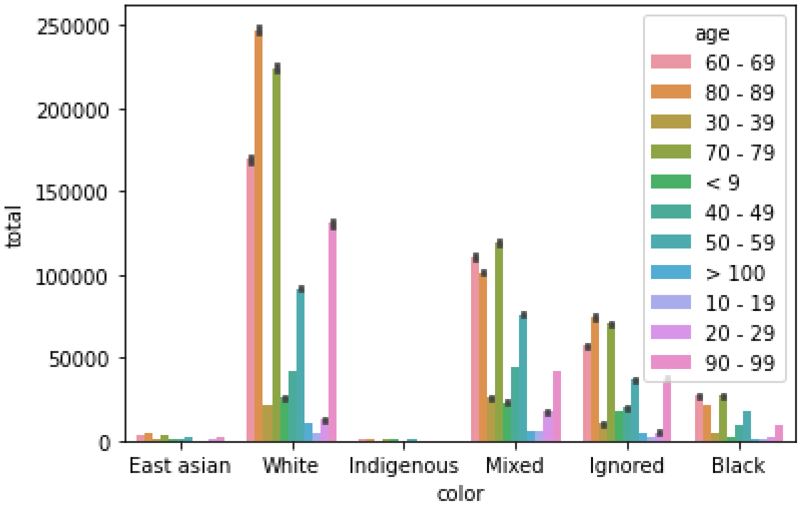
```
sns.barplot(x="year_of_death", y="total", data=df, hue="age" ,estimator=np.sum)
```



Color wise data insights

```
df.groupby('color').total.sum().sort_values(ascending=False).head(5)
```

> White (980971) > Mixed (571663) > Ignored (336754) > Black (123384) > East asian (17609) > Indigenous (3760)



10- Recommendations

Data would have been more revealing if a separate column telling about the previous medical history of that individual would have been added in the dataset so that we could have a better idea whether the heart attack, cardiopathy ,cardiogenic shock, stroke and other reasons were the outcome of previous underlying medical health conditions or coronavirus is the actual culprit. Further the most frequent category of "others" in the "cause" column is really broad with the count values of 3996 have no specific attribution of any disease or list of ailments.

Zartashia Afzal

Email: chemistzartashiaafzal@gmail.com

Mujtaba Choudhary

Email: mujtabachoudhry4@gmail.com