Statistical analysis of Covid-19 (SARS-Cov-2)

Mainak Mukherjee

4th Semester

West Bengal State University

Department of Statistics

20/07/2022

Objective :

- This dataset has many missing values, and directly applying analysis on the dataset is not possible because it will not provide accurate results and there will be a high chance of biasedness.
- Therefore, we first perform data preprocessing. In this step, we will check for missing values based on state and check that there is any missing value for some particular time interval.
- Next we will also try to find the relationship between gender (male and female), age group (less than 18-19 to 40-41 to 65 and greater than 65) and current status i.e, (recovered, hospitalized and deceased).
- Further, we will try to find out the following dependencies of the said attributes :

1. Is there any relationship between gender and patient status?
2. Is there any relationship between patient age and patient status?
3. Is there any relationship between patient age and patient gender?

About the Dataset :

1. The dataset has been taken from Kaggle.
2. There are total of 15 attributes in the dataset.
3. Except for age, all attribute data types are strings.
4. In SPSS, we cannot perform any type of analysis on the string datatype.
5. Therefore, we replace the value of gender, transmission type and current status with nominal data.

The following table will show the change of string value into nominal:

Table 1

	LABEL	VALUE
	MALE	1
GENDER	FEMALE	2
	RECOVERED	1
CURRENT STATUS	HOSPITALIZED	2
	DECEASED	3

- Age data are available in integer format,but the value of age ranges between 0 and 100, so it is very difficult to visualize such data. We also divided this attributes into categories and made a new age attribute.
- Next, Table 2 will show the change of age value into age group and nominal.

Table 2

Age Range	Age Group	Value
0 to 18	<18	1
19 to 40	19-40	2
41 to 65	41-65	3
Greater than 65	>65	4

- After filtering the data state wise, we checked the data for missing values. There were a total of 975 cases out of which 183 cases had missing age and gender values. Here, in this analysis we removed these values. To remove the missing values, first we check in which date range we have less missing values. After visualizing the data, we found that from 25/03/2020 to 27/04/2020, and there is very less missing values.

Table 3

- Table 3 will show us the total cases in Maharashtra.

VALID		FREQUENCY	PERCENTAGE
	NALE	464	53.0
	FEMALE	237	27.1
	TOTAL	701	80.1
MISSING		183	19.9
TOTAL		975	100.0

Table 4

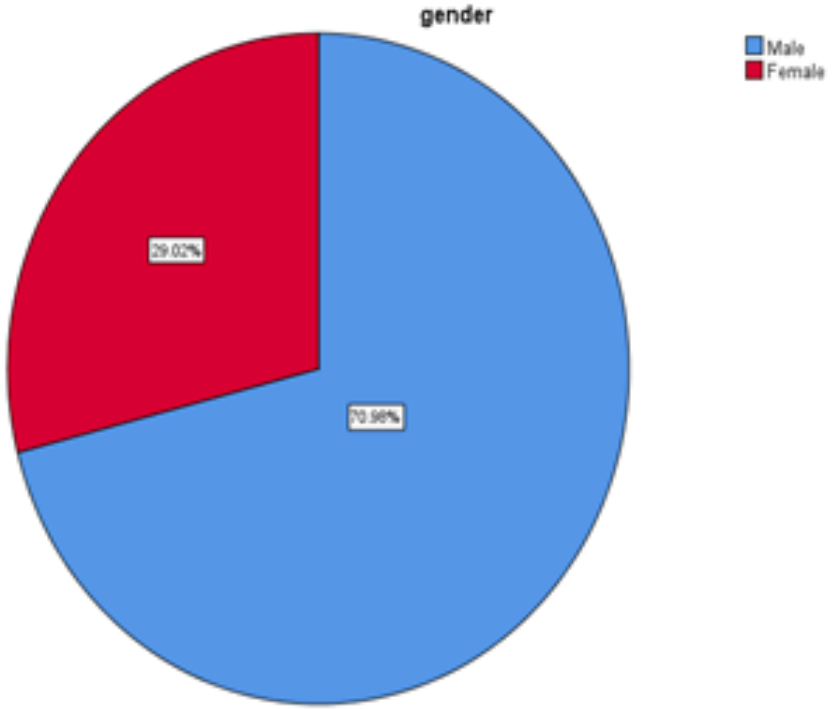
Table 4 showing the total cases remaining after filtering data with date.

- In SPSS, there are many useful commands that can be used to handle missing values. To remove missing values in gender we have used the following command : (gender=1 or gender=2)
- This command selects only those rows where we have gender value either 1 or 2 and all the other rows remain unselected.

VALID		FREQUENCY	PERCENTAGE
	MALE	362	70.7
	FEMALE	148	28.9
	TOTAL	510	99.6
MISSING		2	4
TOTAL		512	100.0

Table 5 : Final dataset used for analysis.

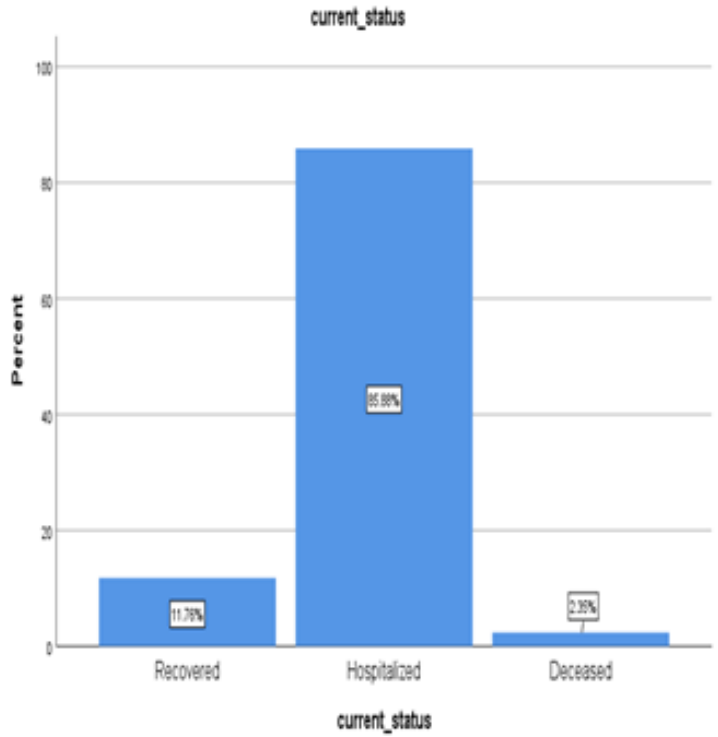
- Table 5 shows the statistics after removing all the missing values.
- Fig. 1 shows the pie chart of male and female cases.



		FREQ.	PERCENTAGE
VALID	MALE	362	71.0
	FEMALE	148	29.0
	TOTAL	510	100.0

Table 6 : Current status attributes

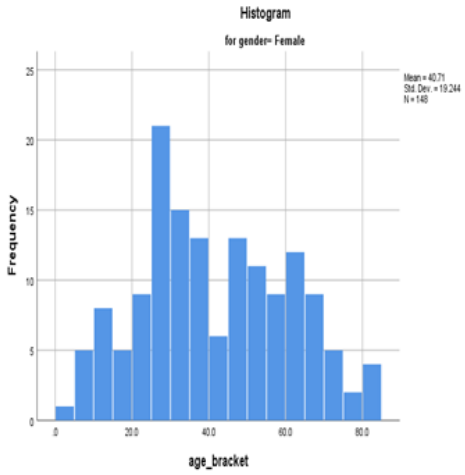
- Table 6 provides us the information related to current status attributes. There are no missing value attributes.
- Fig. 2. Bar chart for current status and it can be clearly seen from the bar chart that the majority of cases are hospitalized.

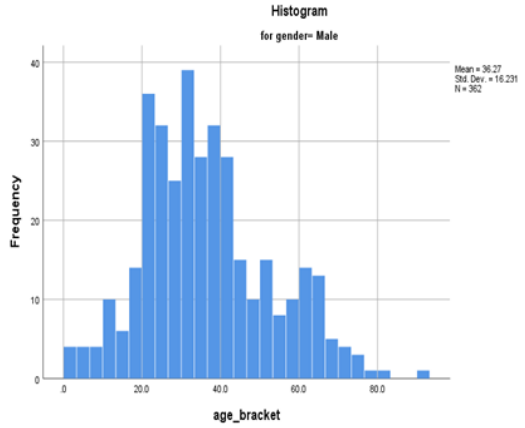


CURRENT STATUS	
VALID	510
MISSING	0

Table 7 :
Valid and missing values in age bracket

- Table 7 provides the details of the age value in the dataset.
- Fig. 3. shows the histogram for male and females respectively.

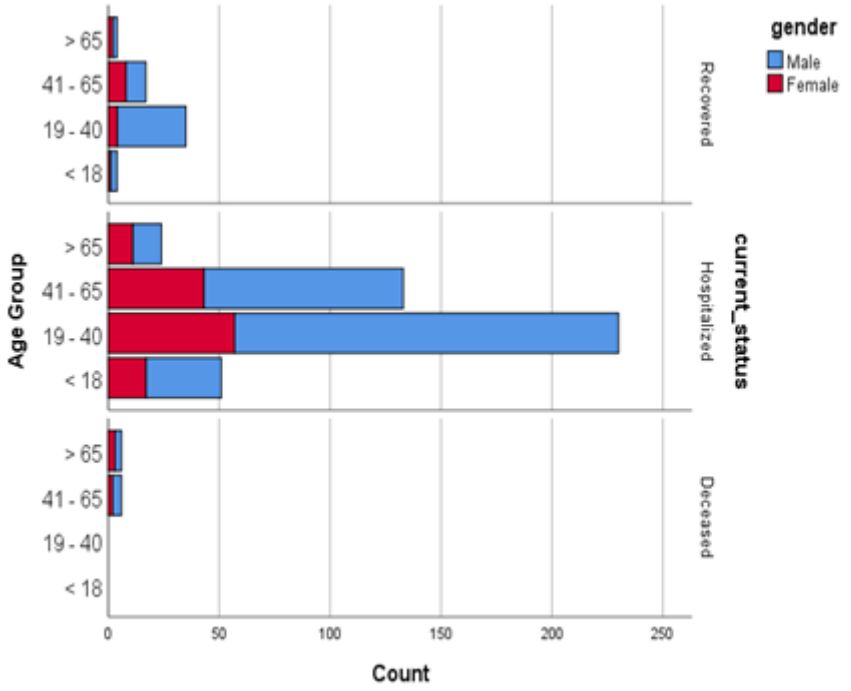




AGE	
VALID	510
MISSING	0

Fig. 6 :

- The cases according to the Age group, current status and gender are represented in graphical form in Fig.6.



APPROACH

- To solve the research questions, we performed a Chi-square test.
- This test is used when we are dealing with nominal or ordinal data and want to find the relationship between the variables.

RESULTS :
Chi Square Test

- In the Chi square test, we assume two hypothesis, the null hypothesis (h_0) and the alternative hypothesis (h_a).
- Null hypothesis (h_0) : there is no relationship between the variables.
- Alternative hypothesis (h_a) : there is a significant relationship between the variables.
- If the p value (asymptotic significance) is less than 0.05 then we reject our null hypothesis and if the value is greater than 0.05 then we cannot reject our null hypothesis

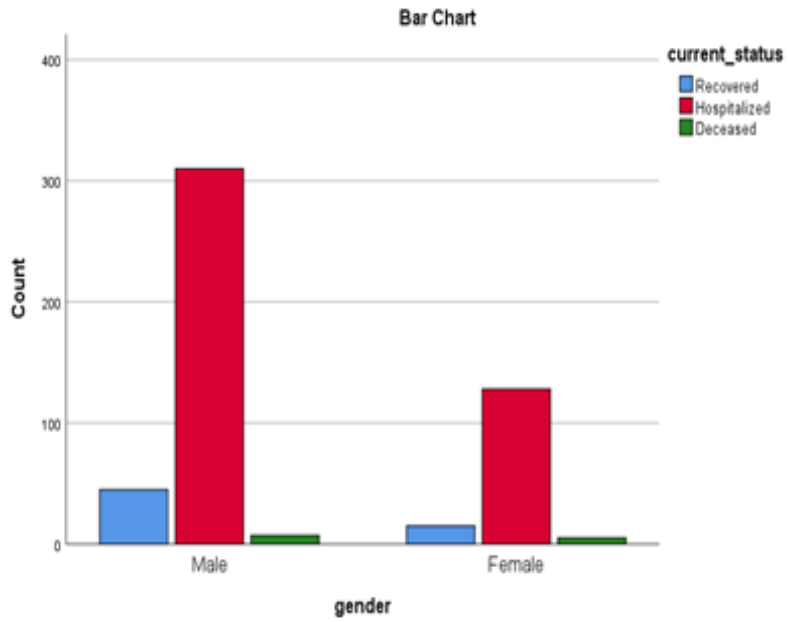
Research Questions :

Q1. Is there any relationship between gender and patient status?

A chi square test was performed to determine the relationship between the gender and the patient status. In this test, we want to check whether gender (male and female) has any dependence on current status and vice versa. Now table 8 gives a cross tabulation of gender and current status and Fig. 7 represents the graphical representation. In table 11, the chi square value is calculated and it is 494 which is much higher than 0.05 and so we cannot reject our null hypothesis. **We can say that there is no effect of gender on the current status of the patient and vice versa. And in other words current status does not depend upon whether a patient is male or female.**

Table 8 :
Cross table of gender and current status.

GENDER	CURRENT STATUS			TOTAL
	RECOVERED	HOSPITALIZED	DECEASED	
MALE	45	310	7	362
FEMALE	15	128	5	148
TOTAL	60	438	12	510



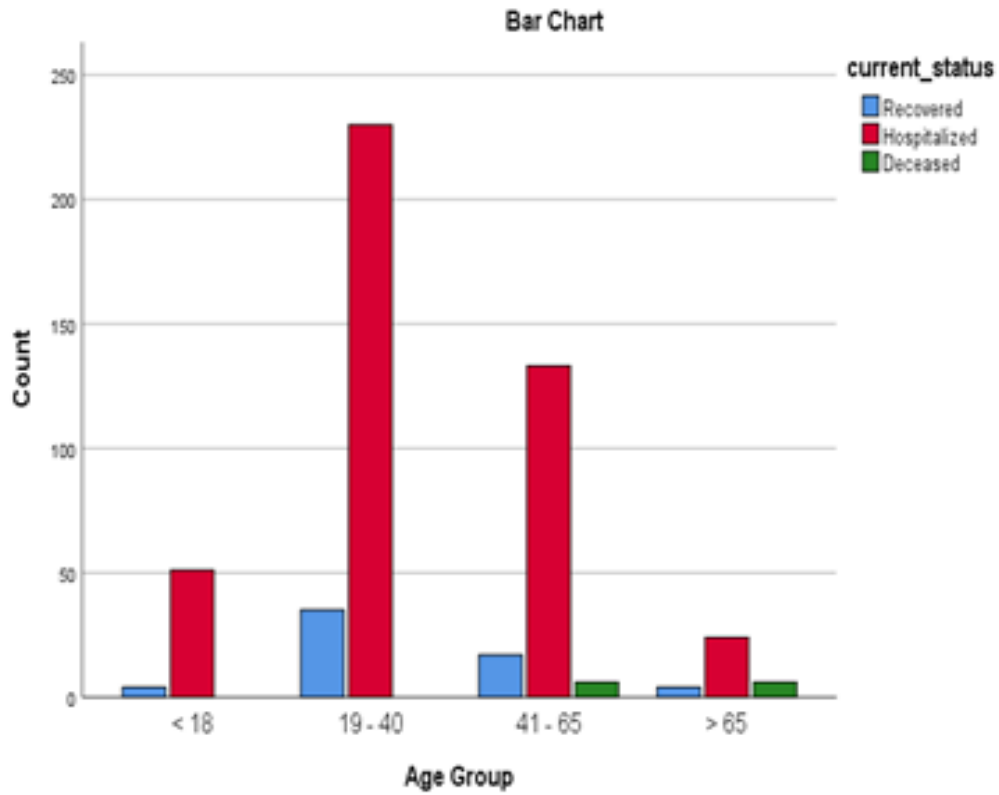
Research Questions

Q2. Is there any relationship between Age group and patient status?

Similar to our First question, we will also use the chi square test to determine the **relationship between age group and current status.** In this test, we want to check whether the age group has any dependencies on current status or vice versa. Table 9 gives the cross tabulation of age group and current status and Fig. 8 represents the graphical representation. In table 11, the value of chi square is calculated and it is 0.000, which is less than 0.05 and so we reject our null hypothesis. **We can say that there is an effect of age group on the current status of the patient and vice versa.**

Table 9 :
Cross table of age group and current status.

AGE	CURRENT STATUS			TOTAL
GROUP	RECOVERED	HOSPITALIZED	DECEASED	
<18	4	51	0	55
19-40	35	230	0	265
41-65	17	133	6	156
>65	4	24	6	34
TOTAL	60	438	12	510



Research Question

Q3. Is there any relationship between age group and gender?

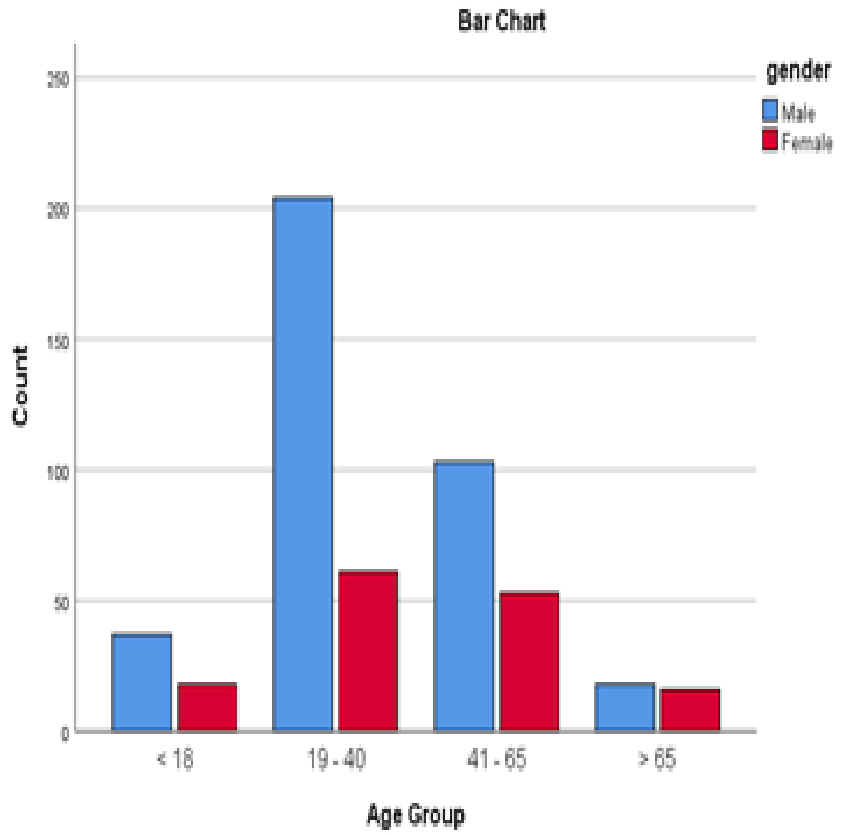
A chi square test is also performed to determine the relationship between age group and gender. In this test, we want to check whether the age group has any dependencies on gender and vice versa. Table 10 gives the cross tabulation of age group and gender and Fig. 9 represents the graphical representation. And in table 11 the chi square value is calculated and it is 0.007 which is less than 0.05, so we reject our null hypothesis. **We can say that there is an effect of age group on the gender of the patient and vice versa.**

Table 10 :
Cross table of age group and gender.

AGE GROUP	GENDER		TOTAL
	MALE	FEMALE	
<18	37	18	55
19-40	204	61	265
41-65	103	53	156
>65	18	16	34
TOTAL	362	148	510

Table 11 :
Chi square value for Maharashtra

	p VALUE
GENDER AND CURRENT STATUS	0.494
AGE GROUP AND CURRENT STATUS	0.000
AGE GROUP AND GENDER	0.007



Conclusion :

- Covid-19 is increasing daily, and it is very important to analyse these data. In this study, Maharashtra state covid-19 patients data were analysed to determine the relationship between different variables. In table 11 Maharashtra results shows that there are dependencies in age group and current status and in age group and gender only in gender, and current status variables are independent.