Statistical Analysis of Covid-19(SARS-Cov-2) Patients Data of Karnataka, India

Ravi Sharma (ravis.cs.19@nitj.ac.in)

Dr. B. R. Ambedkar National Institute of Technology, Jalandhar, Punjab, India https://orcid.org/0000-0002-9047-3279

Nonita Sharma

Dr. B. R. Ambedkar National Institute of Technology, Jalandhar, Punjab, India

Research Article

Keywords: Covid-19, Chi-Square test, SARS-Cov-2, Statistical analysis.

Posted Date: September 18th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-72912/v1

Abstract

Cases of coronavirus disease 2019 (Covid-19) in India is increasing day by day. Severe acute respiratory syndrome coronavirus 2 (SARS-Cov-2) is a new virus of coronavirus family therefore not much information available, thus making it very difficult task to make medicine or vaccine for this virus as early as possible. So, it is very important to analyse the data and find meaningful insight in data so graph of cases that is increasing day by day can be flatten out. For current study, Karnataka state data has taken and Chi square test is performed to find relationship between gender (male and female), age group (less than 18, 19 to 40, 41 to 65 and greater than 65) and current status (recovered, hospitalized and deceased). Our results show that gender is independent of current status and age group is dependent upon current status and age group and gender relationship is also dependent.

Introduction

Currently, solving any problem data plays a very important role. In industry, to reduce cost and maximize profit, data analysis is very useful. Covid-19 cases are increasing daily. This virus is a new type of virus, so little information is available related to this virus, thus making it very difficult to make medicine or vaccine for this virus as early as possible. At the time of writing this research paper, cases of Covid-19 are reaching nearly 15.01 million worldwide [1]. In India, the total number of cases is approximately 1.2 million, out of which 440k cases are still active [2]. The number of positive cases in a day is also increasing rapidly. Therefore, it is very important to analyse these data and find meaningful insight into the data so that graphs of cases that are increasing daily can be flattened out.

Literature Review

In [3], researchers provide data analysis and prediction for Covid-19 cases all over the world. They collected data from DingXiangYuan, John Hopkin University and WHO. These data are uploaded on the CoronaTracker[4] website. For prediction, the Susceptible-Exposed-Infected-Removed (SEIR) model was used. They also provide sentiment analysis of news on Covid-19, and they found 561 positive articles and 2548 negative articles.

In [5] this analysis, researchers provided an effect of comorbidity on Covid-19 patients. They analysed 1590 confirmed cases in China hospitalized in different hospitals. A total of 686 female patients, 399 patients had comorbidities. In this research, they found that comorbidity plays a crucial role in clinical treatment, and patients with comorbidities have poor clinical outcomes. Another study [6] showed that the most common symptoms of covid-19 were fever, cough, expectoration, headache and myalgia or fatigue.

[7] performed a clinical prediction of mortality of covid-19 based on 150 patients in Wuhan city, China. Of these 150 cases, 68 and 82 were deaths and discharges, respectively. In this study, they found that there

is a significant difference between age in death cases and discharge cases. Forty-three out of 68 deaths had comorbidities, and in discharge cases, 34 out of 82 had comorbidities. Sixty-three patients died due to respiratory failure or myocardial damage. Only 5 patients died without any known cause.

Data Analysis

In this analysis, we have taken Covid-19 data from Karnataka, India. The dataset for this study was downloaded from Kaggle [8]. In this dataset, a list of covid-19 cases of each state is provided, but most of the attribute values were missing. For this study, we are mainly focused on three attributes: gender, age and current status (recovered, hospitalized and deceased). This dataset has many missing values, and directly applying analysis on the dataset is not possible because it will not provide accurate results and a high chance of biasedness.

Therefore, we first perform data preprocessing. In this step, we check for missing values based on state and check that there is very missing value for some particular time interval. Based on these two conditions, we select Karnataka. After selecting the target data, we thoroughly analysed the data and finalized our research question.

Research Question

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?

For this analysis, we used IBM SPSS[9] software.

Dataset

- The dataset has been taken from Kaggle.[8]
- There are a total of 17 attributes in the dataset.
- Except for age, all attribute data types are strings.
- In SPSS, we cannot perform any type of analysis on the string datatype.
- Therefore, we replace the value of gender, transmission type and current status with nominal data.

Table 1. Change of String value into Nominal

	Label	Value
Gender	Male	1
	Female	2
Current Status	Recovered	1
	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 attribute into categories and made a new age attribute.

Table 2. Change of Age value into age group and Nominal

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 875 cases, out of which 174 cases had missing age and gender values. In this analysis, we removed these values. To remove missing values first, we check in which date range we have less missing value. After visualizing the data, we found that from 09 March 2020 to 27 April 2020, there is very less missing value. Table 4 shows that after filtering data with the date range, we have only 2 missing values.

Table 3. Total cases in Karnataka

		Frequency	Percent
Valid	**Male**	464	53.0
	Female	237	27.1
	Total	701	80.1
Missing		174	19.9
Total		875	100.0

Table 4. Total cases remaining after filtering data with date

		Frequency	Percent
Valid	**Male**	362	70.7
	Female	148	28.9
	Total	510	99.6
Missing		2	.4
Total		512	100.0

In SPSS, there are many useful commands that can be used to handle missing value. To remove missing value in gender, we use 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.

Table 5. Final dataset used for analysis

		Frequency	Percent
Valid	**Male**	362	71.0
	Female	148	29.0
	Total	510	100.0

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

Table 6 provides 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.

Valid and missing values in the current status

Current Status		
N	Valid	510
	Missing	0

Table 7 provides details of the age value in the dataset. Fig. 3 shows the histogram, mean age and standard deviation for age. Fig. 4 and Fig. 5 show histograms for males and females, respectively.

Table 7. Valid and missing values in Age Bracket

Age		
N	Valid	510
	Missing	0

Cases according to age group, current status and gender are represented in graphical form in Fig. 6.

To solve the research question, we perform a chi square test. This test is used when we are dealing with nominal or ordinal data and want to find the relationship between two variables.

Results

Chi Square Test

In the chi square test, we assume two hypotheses, the null hypothesis (h_0) and the alternate hypothesis (h_a).

- Null hypothesis (h_0): there is no relationship between variables.
- Alternate hypothesis (h_a): There is a significant relationship between variables.

If the p value (asymptotic significance) is less than .05, then we reject our null hypothesis, and if the value is greater than .05, then we cannot reject our null hypothesis.

RQ. 1. Is there any relationship between gender and patient status?

A chi-square test was performed to determine the relationship between gender and current status. In this test, we want to check whether gender (male or female) has any dependencies on current status (recovered, hospitalized and deceased) and vice versa. 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 .05, 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. 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

RQ2. Is there any relationship between Age group and Patient status?

Similar to the RQ1 Chi square test, the relationship between age group and current status was also determined. In this test, we want to check whether the age group has any dependencies on current status and 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 chi square value is calculated, and it is .000, which is less than .05, 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 Group	Current Status			Total
	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

RQ. 3. Is there any relationship between age group and gender?

A chi-square test was 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. In

Table 11. The chi square value is calculated, and it is .007, which is less than .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 Karnataka

	p Value
Gender and Current Status	.494
Age Group and Current Status	.000
Age Group and Gender	.007

Conclusion

Covid-19 cases is increasing daily, and it is very important to analyse these data. In this study, Karnataka state Covid-19 patients' data were analysed to determine the relationship between different variables. In Table 11. Karnataka results show that there are dependences in age group and current status and in age group and gender only in gender, and current status variables are independent.

References

1. WHO Coronavirus Disease (COVID-19) Dashboard | WHO Coronavirus Disease (COVID-19) Dashboard, https://covid19.who.int/, last accessed 2020/07/24.
2. MoHFW | Home, https://www.mohfw.gov.in/, last accessed 2020/07/24.
3. Amira, F., Hamzah, B., Lau, C.H., Nazri, H., Ligot, D.V., Lee, G., Liang Tan, C., Khursani Bin, M., Shaib, M., Hasanah, U., Zaidon, B., Abdullah, A.B., Chung, M.H., Ong, C.H., Chew, P.Y., Salunga, R.E., Hamzah,

A.B.: CoronaTracker: World-wide COVID-19 Outbreak Data Analysis and Prediction CoronaTracker Community Research Group Correspondence to Fairoza. https://doi.org/10.2471/BLT.20.251561.

4. COVID-19 Corona Tracker, https://www.coronatracker.com/, last accessed 2020/07/13.
5. Guan, W.J., Liang, W.H., Zhao, Y., Liang, H.R., Chen, Z.S., Li, Y.M., Liu, X.Q., Chen, R.C., Tang, C.L., Wang, T., Ou, C.Q., Li, L., Chen, P.Y., Sang, L., Wang, W., Li, J.F., Li, C.C., Ou, L.M., Cheng, B., Xiong, S., Ni, Z.Y., Xiang, J., Hu, Y., Liu, L., Shan, H., Lei, C.L., Peng, Y.X., Wei, L., Liu, Y., Hu, Y.H., Peng, P., Wang, J.M., Liu, J.Y., Chen, Z., Li, G., Zheng, Z.J., Qiu, S.Q., Luo, J., Ye, C.J., Zhu, S.Y., Cheng, L.L., Ye, F., Li, S.Y., Zheng, J.P., Zhang, N.F., Zhong, N.S., He, J.X.: Comorbidity and its impact on 1,590 patients with Covid-19 in China: A nationwide analysis. Eur. Respir. J. 55, (2020). https://doi.org/10.1183/13993003.00547-2020.
6. Xu, X.W., Wu, X.X., Jiang, X.G., Xu, K.J., Ying, L.J., Ma, C.L., Li, S.B., Wang, H.Y., Zhang, S., Gao, H.N., Sheng, J.F., Cai, H.L., Qiu, Y.Q., Li, L.J.: Clinical findings in a group of patients infected with the 2019 novel coronavirus (SARS-Cov-2) outside of Wuhan, China: Retrospective case series. BMJ. 368, (2020). https://doi.org/10.1136/bmj.m606.
7. Ruan, Q., Yang, K., Wang, W., Jiang, L., Song, J.: Clinical predictors of mortality due to COVID-19 based on an analysis of data of 150 patients from Wuhan, China, https://doi.org/10.1007/s00134-020-05991-x, (2020). https://doi.org/10.1007/s00134-020-05991-x.
8. COVID-19 Corona Virus India Dataset | Kaggle, https://www.kaggle.com/imdevskp/covid19-corona-virus-india-dataset?select=patients_data.csv, last accessed 2020/07/07.
9. SPSS Software - India | IBM, https://www.ibm.com/in-en/analytics/spss-statistics-software, last accessed 2020/07/07.

Figures

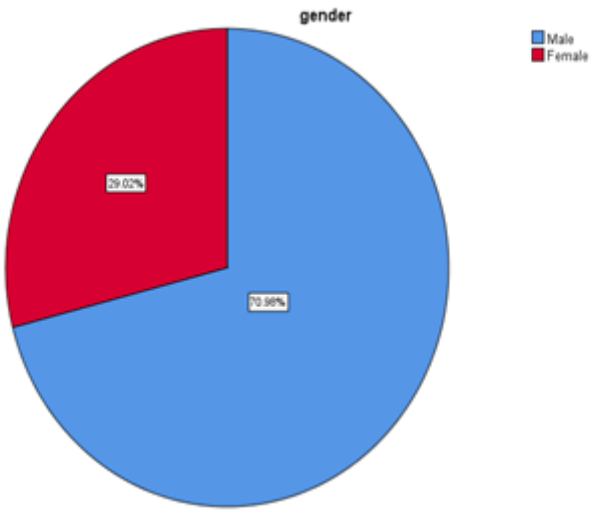

Figure 1

Pie chart for Gender

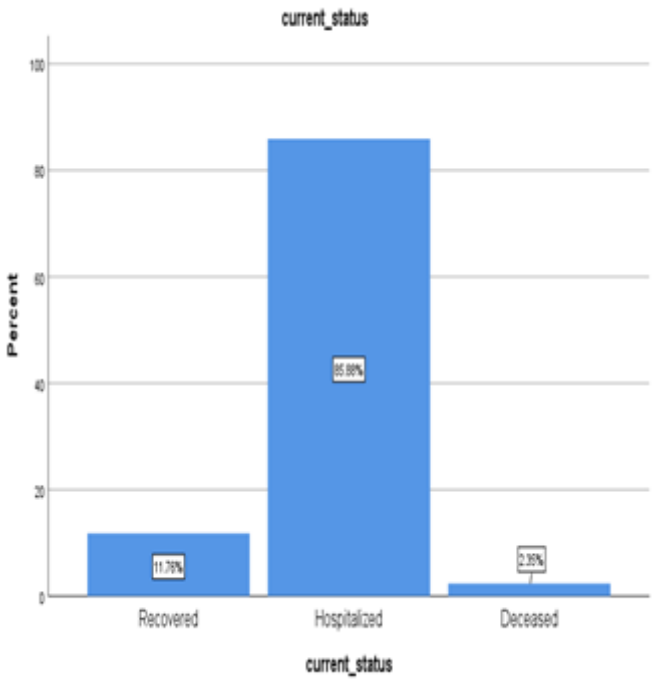

Figure 2

Bar Chart for Current Status

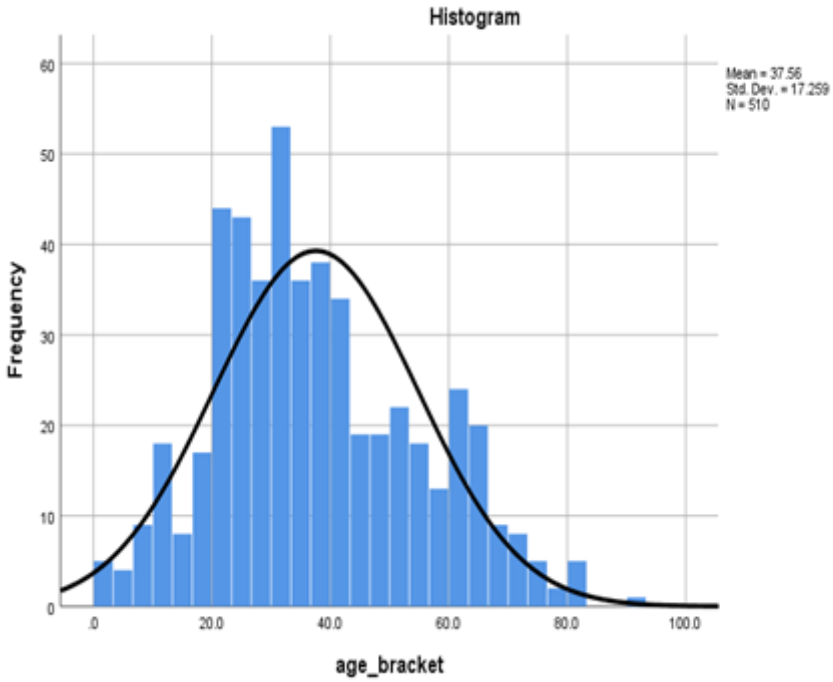

Figure 3

Histogram for age

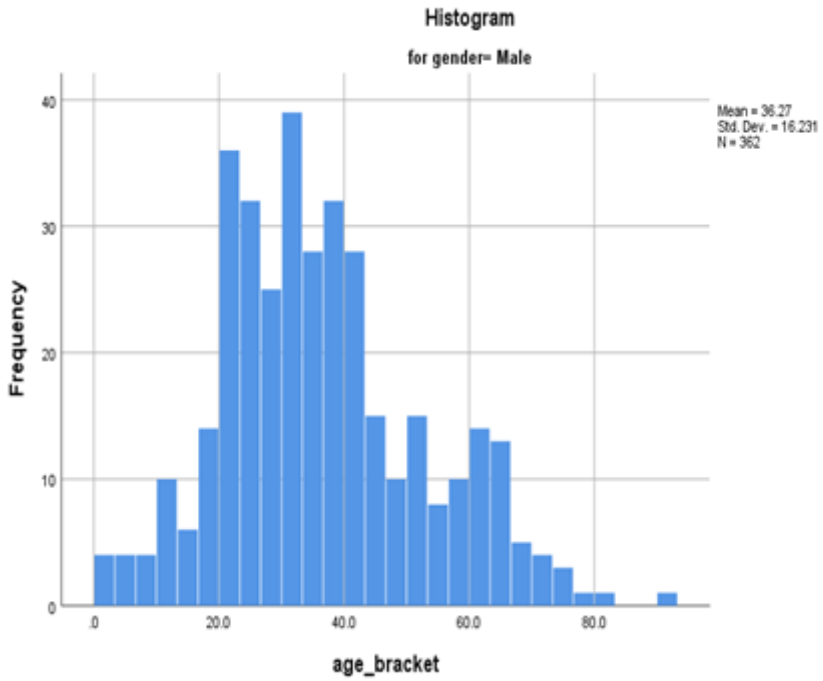

Figure 4

Histogram for age w.r.t. male

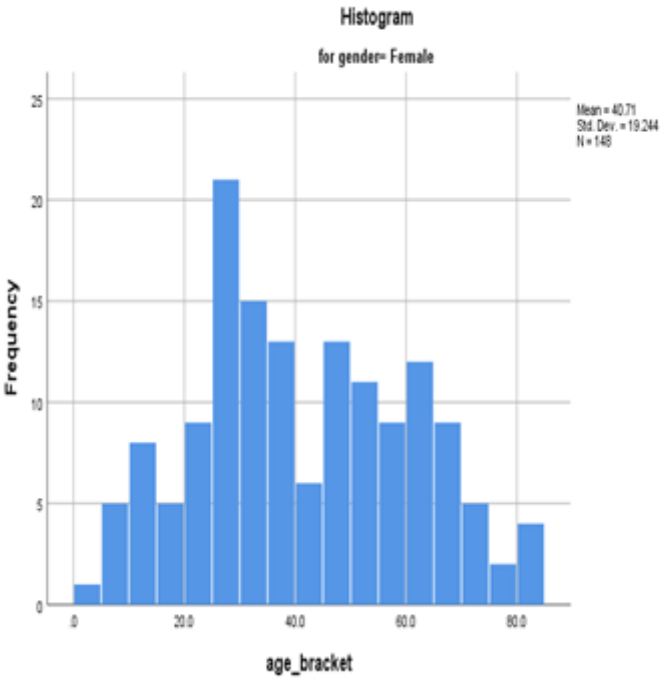

Figure 5

Histogram for age w.r.t. female

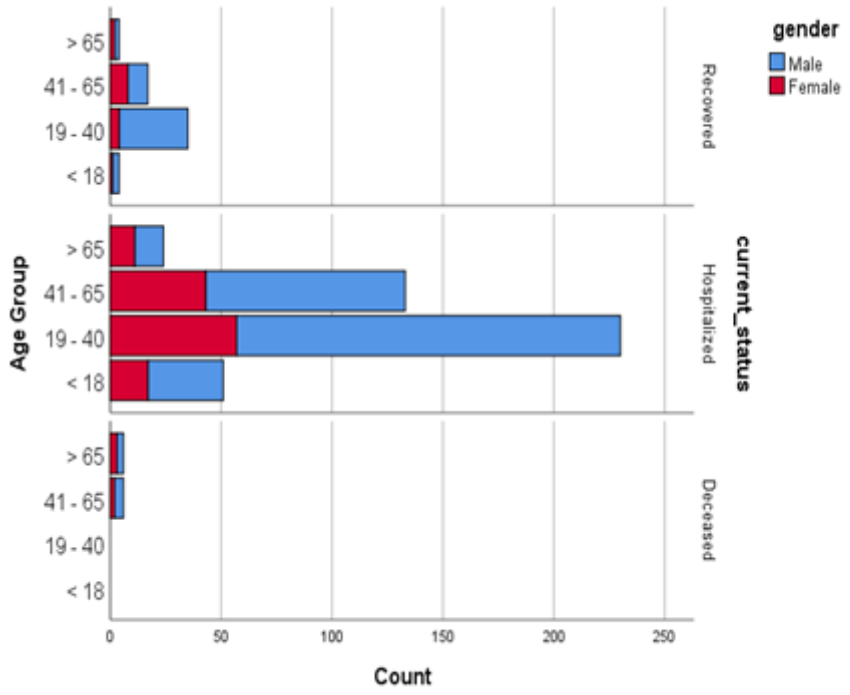

Figure 6

Bar Graph representing Male and female in different age group with current status

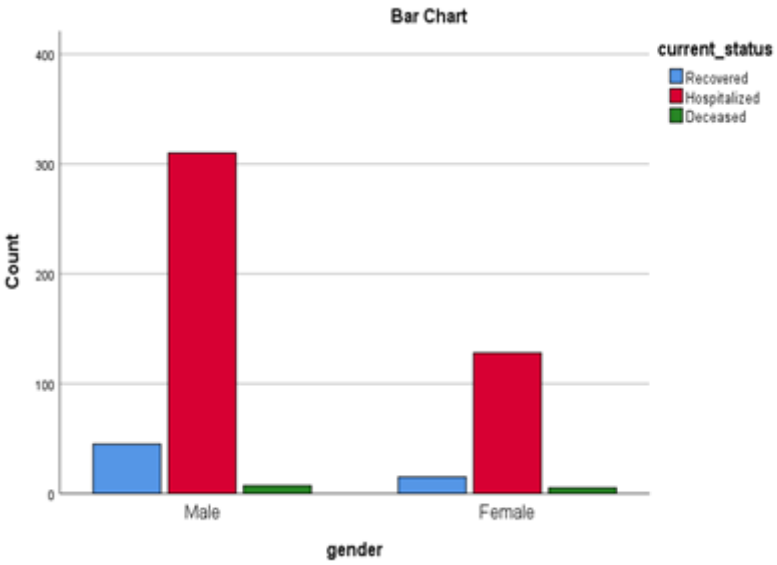

Figure 7

Bar Chart for gender w.r.t. current status

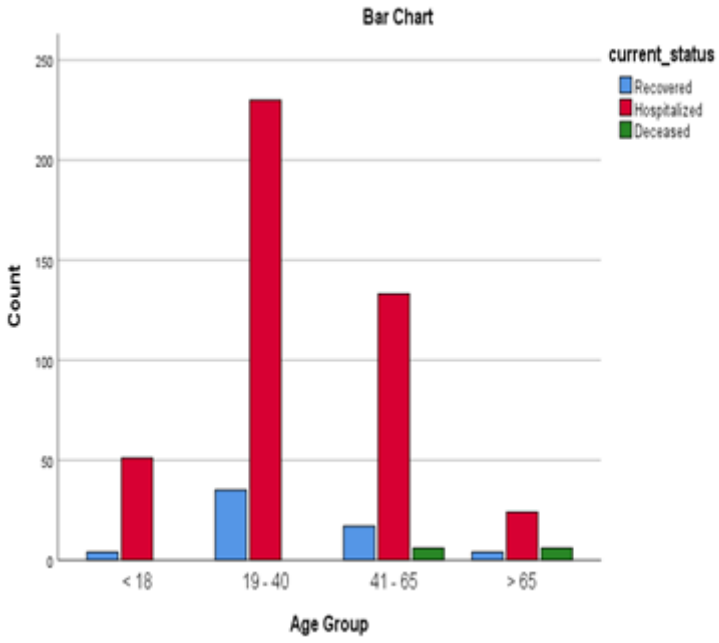

Figure 8

Bar Chart for age group w.r.t. current status

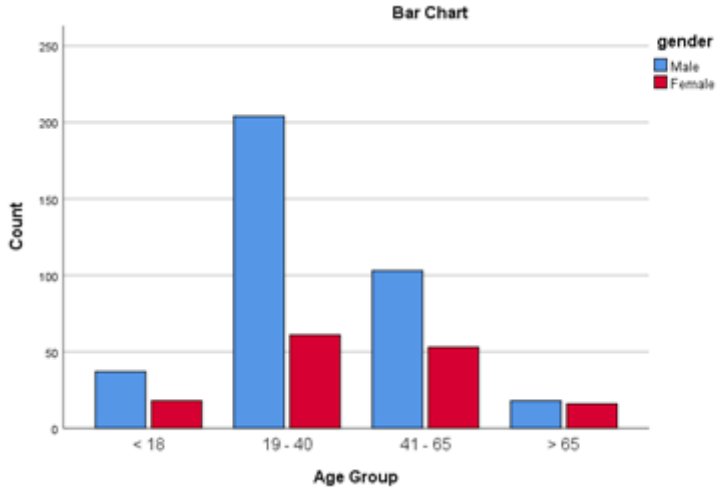

Figure 9

Bar Chart for age group w.r.t. gender