Covid-19 CDC Data Analysis

1- Abstract

This report presents an exploratory data analysis performed on data of some COVID-19 patients in the United States. The data was obtained from the Center for Disease Control and Prevention CDC. The goal of our analysis is to demonstrate the relationships between the patients' mortality rate and the following factors: age, race & ethnicity, sex, having medical conditions, patient hospitalization, and patient admission to the intensive care unit ICU. In order to exclude the effects of the newly innovated vaccines and the new strains of SARS-CoV-2, this report uses only the confirmed cases of the data that have been captured during the period [01/01/2020, 02/12/2020]. The report concludes that all the mentioned factors have significant effects on the patients' mortality rate.

2- Introduction

The Center for Disease Control and Prevention CDC shares important parts of a database called "COVID-19 case surveillance" with public. The part that will be used in this report is called **COVID-19 Case Surveillance Public Use Data**. The mentioned part contains 11 data elements (i.e. columns or features). The database is updated monthly, the 11 data elements of our data are:

Table 1:

Column Name	Description	Type
cdc_report_dt	Date case was first reported to the CDC. Calculated date. This date was populated using the date at which a case record was first submitted to the database. If missing, then the report date entered on the case report form was used. If missing, then the date at which the case first appeared in the database was used. If none available, then left blank.	Date & Time
pos_spec_dt	Date of first positive specimen collection (Case Report Form)	Date & Time
onset_dt	Symptom onset date, if symptomatic (Case Report Form)	Date & Time
current_status	Case Status (Case Report Form: What is the current status of this person?) -- Values: Laboratory-confirmed case; Probable case;	Plain Text
Sex	Sex (Case Report Form): Male; Female; Unknown; Other; Missing; NA	Plain Text

Column Name	Description	Type
age_group	Age Group: 0 - 9 Years; 10 - 19 Years; 20 - 39 Years; 40 - 49 Years; 50 - 59 Years; 60 - 69 Years; 70 - 79 Years; 80 + Years;Unknown, Missing; NA; The age group categorizations were populated using the age value that was reported on the case report form. Date of birth was used to fill in missing/unknown age values using the difference in time between date of birth and onset date.	Plain Text
race_ethnicity_combined	Race and ethnicity (combined): American Indian/Alaska Native, Non-Hispanic; Asian, Non-Hispanic; Black, Non-Hispanic; Multiple/Other, Non-Hispanic; Native Hawaiian/Other Pacific Islander, Non-Hispanic; White, Non-Hispanic; Hispanic/Latino; Unknown; Missing; NA. If more than race was reported, race was categorized into multiple/other races.	Plain Text
hosp_yn	Hospitalization status (Case Report Form: Was the patient hospitalized?) -- Values: Yes; No; Unknown; Missing;	Plain Text
icu_yn	ICU admission status (Case Report Form: Was the patient admitted to an intensive care unit (ICU)?) -- Values: Yes; No; Unknown; Missing;	Plain Text
death_yn	Death status (Case Report Form: Did the patient die as a result of this illness?) -- Values: Yes; No; Unknown; Missing;	Plain Text
medcond_yn	Presence of underlying comorbidity or disease (Case Report Form: Pre-existing medical conditions?) -- Values: Yes; No; Unknown; Missing;	Plain Text

For more details about the data kindly refer to COVID-19 Case Surveillance Public Use Data.

The goal of this report is, **for the confirmed cases**, to demonstrate the relationships between the "non-datetime" data elements of the data set and the **death_yn** data element. As mentioned before, in order to avoid the effects of the invented vaccines and the new strains of SARS-CoV-2 on the relationships that are being studied, this report does not take into account the whole available data in **COVID-19 Case Surveillance Public Use Data** data set, but considers only the data of the period **[01/01/2020, 02/12/2020]**.

The remaining sections of this report are organized as follows: The third section explains importing the data, the fourth section explains the data cleaning process, the fifth section contains the univariate analysis, the sixth section demonstrates the multivariate analysis, the seventh section tests some proposed hypotheses, the 8th section discusses the limitations of the data and the analysis, and the last section summarizes the results and talks about the future works.

3- Importing The Data

Firstly, the data have been downloaded from CDC in csv format, and after that the data have been imported into our python environment.

P.S. As CDC updates the data monthly, we have to metion that the last update of the downloaded data was on **December 2, 2020**. Also we have to mention that the downloaded data contains about ~8.4M rows and only 11 columns. In the Table Below you can find detailed info about the imported data:

Table 2:

#	Column	Non-Null Count	Dtype
---	------	--------------	-----
0	cdc_report_dt	8405079 non-null	object
1	pos_spec_dt	2870789 non-null	object
2	onset_dt	4395957 non-null	object
3	current_status	8405079 non-null	object
4	sex	8405061 non-null	object
5	age_group	8404990 non-null	object
6	Race and ethnicity (combined)	8405072 non-null	object
7	hosp_yn	8405079 non-null	object
8	icu_yn	8405079 non-null	object
9	death_yn	8405079 non-null	object
10	medcond_yn	8405079 non-null	object

And in the table below you can find the number of unique values in each column of the imported data:

Table 3:

cdc_report_dt	321
pos_spec_dt	313
onset_dt	338
current_status	2
sex	5
age_group	10
Race and ethnicity (combined)	9
hosp_yn	4
icu_yn	4
death_yn	4
medcond_yn	4

Also in the table below you can find the first five rows of the imported data:

Table 4:

	cdc_report_dt	pos_spec_dt	onset_dt	current_status	sex	age_group	Race and ethnicity (combined)	hosp_yn	icu_yn	death_yn	medcond_yn
0	2020/11/10	2020/11/10	NaN	Laboratory-confirmed case	Male	10 - 19 Years	Black, Non-Hispanic	No	Unknown	No	No
1	2020/11/14	2020/11/10	2020/11/10	Laboratory-confirmed case	Male	10 - 19 Years	Black, Non-Hispanic	No	No	No	No
2	2020/11/19	2020/11/10	2020/11/09	Laboratory-confirmed case	Male	10 - 19 Years	Black, Non-Hispanic	No	No	No	No
3	2020/11/14	2020/11/10	NaN	Laboratory-confirmed case	Male	10 - 19 Years	Black, Non-Hispanic	Missing	Missing	No	Missing
4	2020/11/13	2020/11/10	2020/11/10	Laboratory-confirmed case	Male	10 - 19 Years	Black, Non-Hispanic	No	No	No	Yes

4- Data Cleaning

Firstly, the confirmed cases have been kept and all the other cases have been removed from the data set. After that, all the confirmed cases without corresponding first positive specimen have been removed, because this data could be inaccurate.

The next step was removing all the columns that we were not interested in anymore, i.e. the date columns and the 'current_status' column. The resulting data set info, after applying the previous step, was as clarified in the table below:

Table 5:

#	Column	Non-Null Count	Dtype
---	------	--------------	-----
0	sex	2704683 non-null	object
1	age_group	2704639 non-null	object
2	Race and ethnicity (combined)	2704687 non-null	object
3	hosp_yn	2704693 non-null	object
4	icu_yn	2704693 non-null	object
5	death_yn	2704693 non-null	object
6	medcond_yn	2704693 non-null	object

From the table above, we can notice that almost all the remained data cells didn't contain null values, so we removed all the rows that contain null values. The resulting data set info, after applying the previous step, was as clarified in the table below:

Table 6:

#	Column	Non-Null Count	Dtype
---	------	--------------	-----
0	sex	2704639 non-null	object
1	age_group	2704639 non-null	object
2	Race and ethnicity (combined)	2704639 non-null	object
3	hosp_yn	2704639 non-null	object
4	icu_yn	2704639 non-null	object
5	death_yn	2704639 non-null	object
6	medcond_yn	2704639 non-null	object

The table above shows that the data seems to be consistent, but we were still facing some issues as all the remaining rows contains values like "Unknown" or "Missing". Those values are useless for our purpose, so we deleted them.

P.S. Deleting these values may bias our sample, but this report assumes that the cases that are fully recorded, without null, missing or unknown values are more accurate. Moreover, deleting these values simplifies the analysis. Especially when we have such a volume of data that requires huge computational capabilities.

The resulting data set info after applying the previous step was as clarified in the table below:

Table 7:

#	Column	Non-Null Count	Dtype
---	------	--------------	-----
0	sex	324387 non-null	object
1	age_group	324387 non-null	object
2	Race and ethnicity (combined)	324387 non-null	object
3	hosp_yn	324387 non-null	object
4	icu_yn	324387 non-null	object
5	death_yn	324387 non-null	object
6	medcond_yn	324387 non-null	object

The previous step was the last step of our data cleaning process. And although, the data size had been shrinked from ~8.4Mx11 to 324387x7, the data were still big enough to serve our purpose.

5- Univariate Analysis

This section analyzes each data element (Column or Feature) independently, using the descriptive statistics and the graphical tools.

5.1- sex Feature:

The distribution of the sex values is described in the figure below:

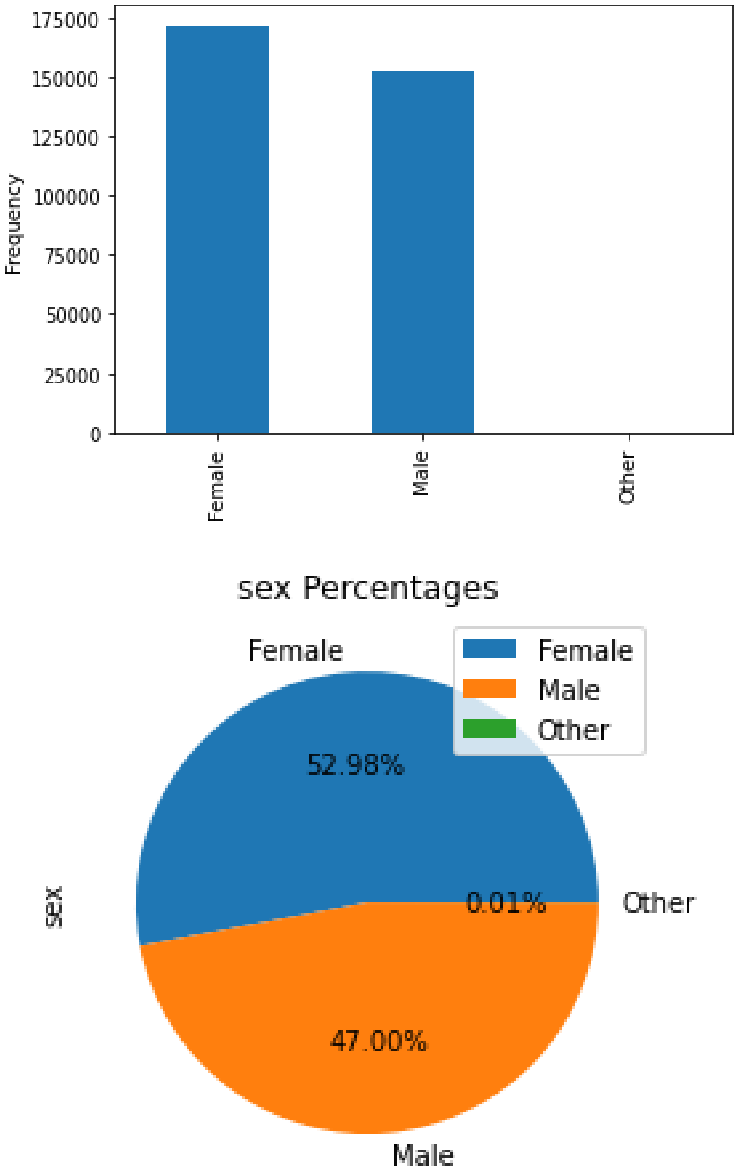


Figure 1

We can see that the sex feature is imbalanced (i.e. there is a considerable difference in the counts of the Females and the Males in our sample). But this difference can be caused by a sampling error or bias, in addition to that we have to know that male/female ratio in USA is 0.96 Wikipedia. Also maybe the females are more likely to be infected with covid-19.

5.2- age_group Feature

The distribution of the age_group values is described in the figure below:

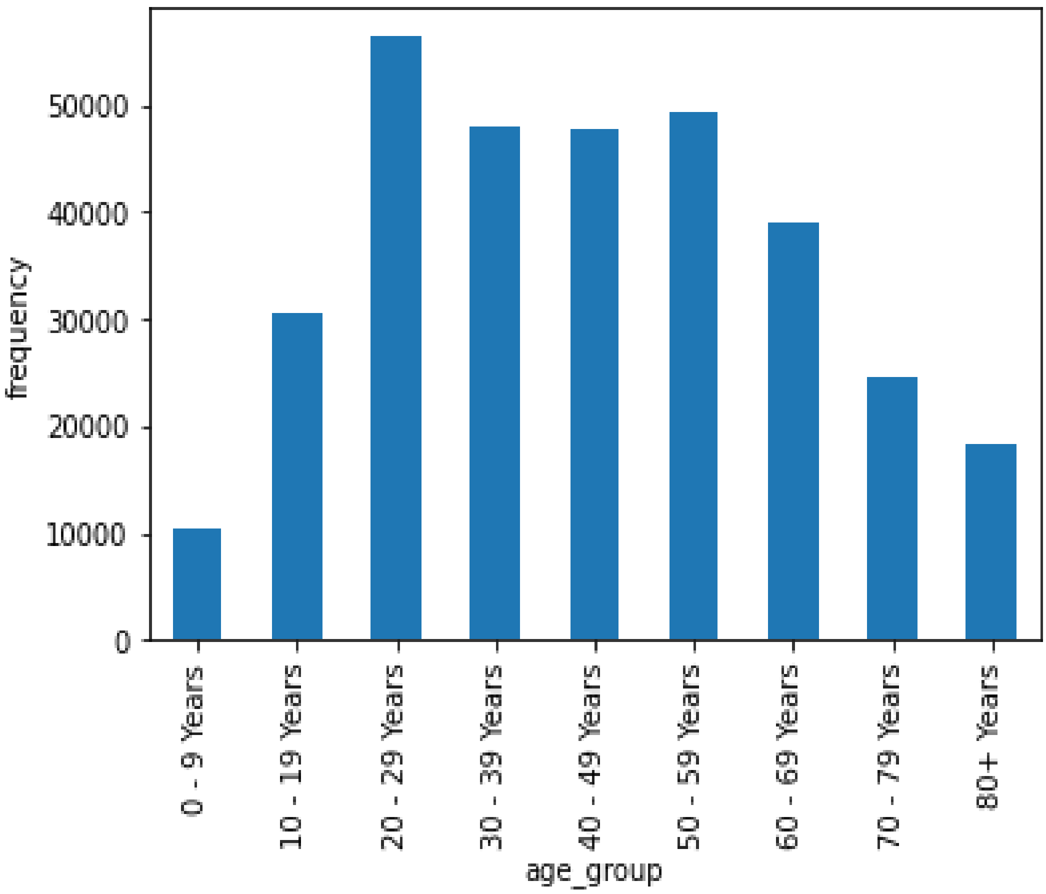
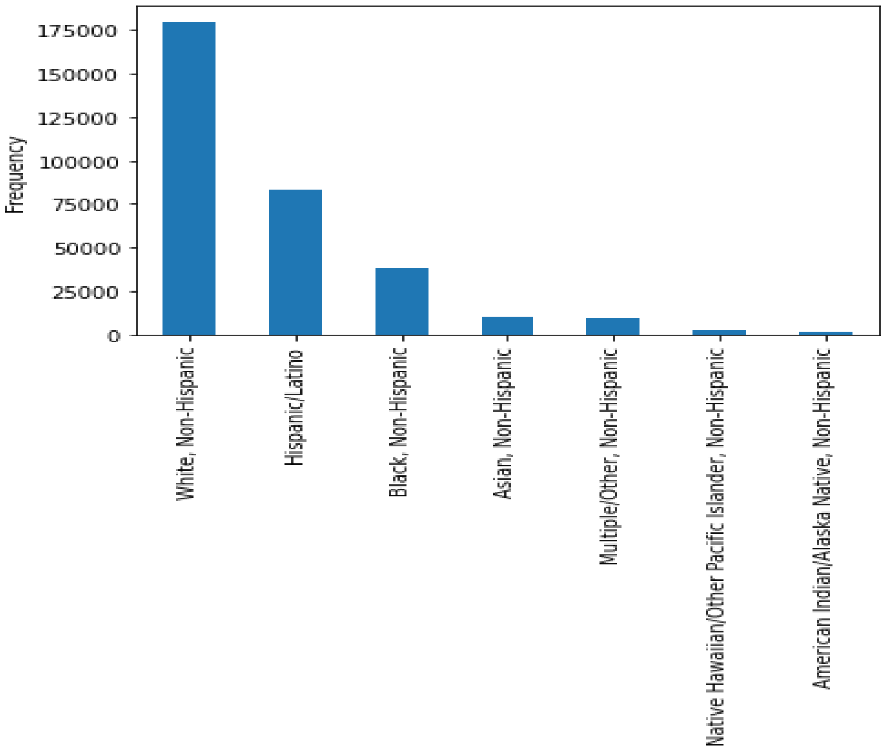


Figure 2

The distribution above looks to be a normal distribution and it is not similar to the USA population distribution by age which is clarified here. The difference seems to be considerable especially for both the children and the old people, because the frequency of the children in our sample seems to be lower while the frequency of the old people seems to be higher. Many reasons might have caused the mentioned effect, maybe because the old people are more likely to be infected with covid-19, or because they are more likely to have serious symptoms, so they have to go to hospital which means that their cases are more likely to be confirmed, i.e. they are more likely to appear in our sample. Also maybe there are other reasons like sampling error or sampling bias.

5.3- race and ethnicity (combined) Feature

The distribution of the age_group values is described in the figure below:


Frequency
175000
150000
125000
100000
75000
50000
25000
0
White, Non-Hispanic
Hispanic/Latino
Black, Non-Hispanic
Asian, Non-Hispanic
Multiple/Other, Non-Hispanic
Native Hawaiian/Other Pacific Islander, Non-Hispanic
American Indian/Alaska Native, Non-Hispanic

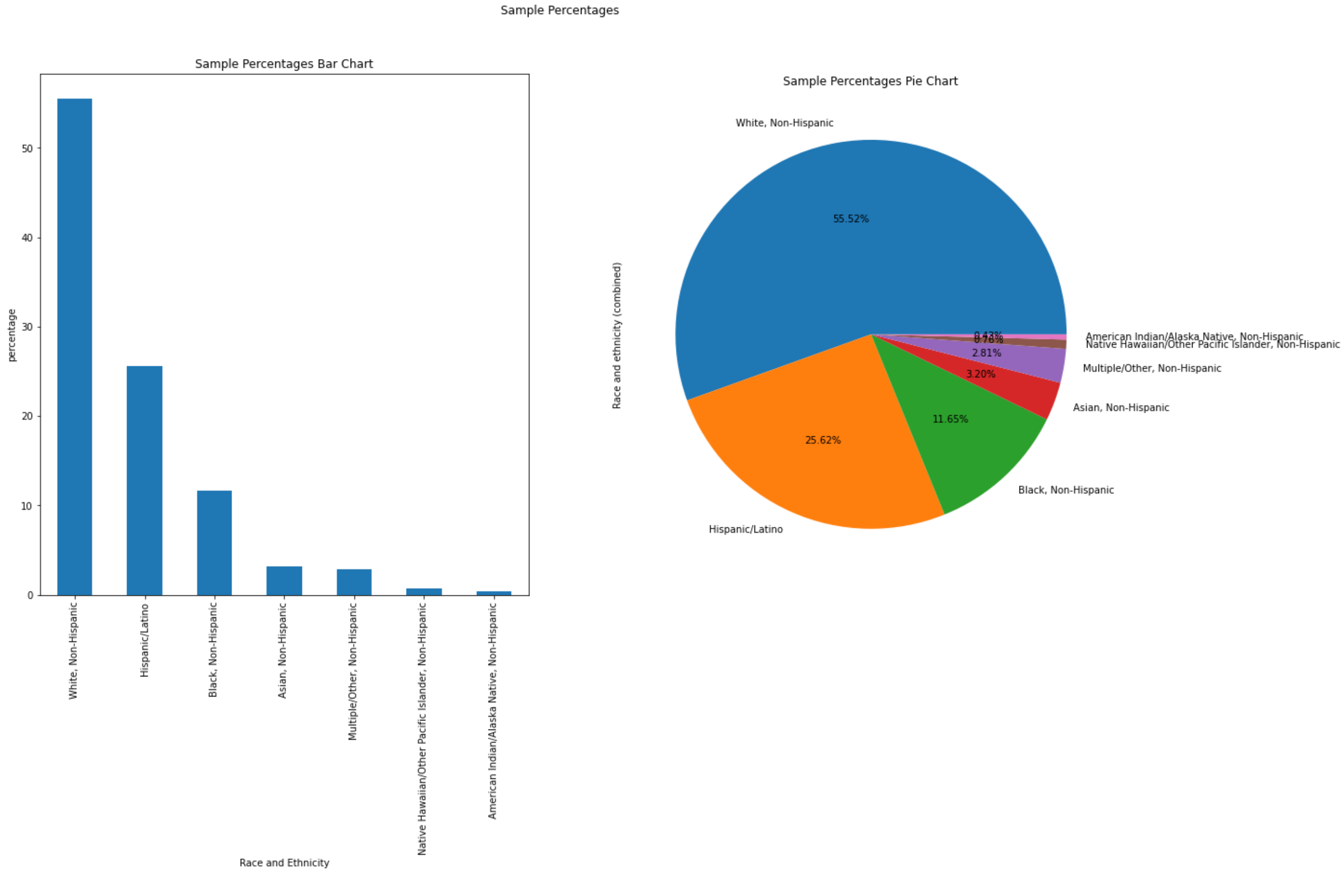

Sample Percentages
Sample Percentages Bar Chart
percentage
50
40
30
20
10
0
White, Non-Hispanic
Hispanic/Latino
Black, Non-Hispanic
Asian, Non-Hispanic
Multiple/Other, Non-Hispanic
Native Hawaiian/Other Pacific Islander, Non-Hispanic
American Indian/Alaska Native, Non-Hispanic
Race and Ethnicity
Sample Percentages Pie Chart
Race and ethnicity (combined)
White, Non-Hispanic
55.52%
American Indian/Alaska Native, Non-Hispanic
Native Hawaiian/Other Pacific Islander, Non-Hispanic
0.43%
0.78%
2.81%
Multiple/Other, Non-Hispanic
3.20%
Asian, Non-Hispanic
11.65%
Black, Non-Hispanic
25.62%
Hispanic/Latino

Figure 3

Actually we don't have exact information about the distribution of the races and ethnicity in USA in 2020. but this link can tell us some information. From the link we can notice that the biggest three parts of the United States population are: the white people (62%), the Hispanic people (~17%) and the black people (12.6%).

From the figures above, it can be easily noticed that the sample contains more Hispanic/Latino people (almost 26%) than the Hispanic people in the American population, and it contains less white people (almost 56%) than the population, but it contains almost the same percentage of the black people (almost 12%).Many reasons might be responsible for the mentioned effect, for example: sampling bias, sampling error, measurement error (i.e. bad classification of the people, or for example, maybe the Hispanic/Latino category is different from the Hispanic category that mentioned here), and maybe the Hispanic/Latino people are more likely and the White Non-Hispanic people are less likely to be infected with covide-12 or get serious symptoms

5.4- hosp_yn Feature

The distribution of the values is described in the figure below:

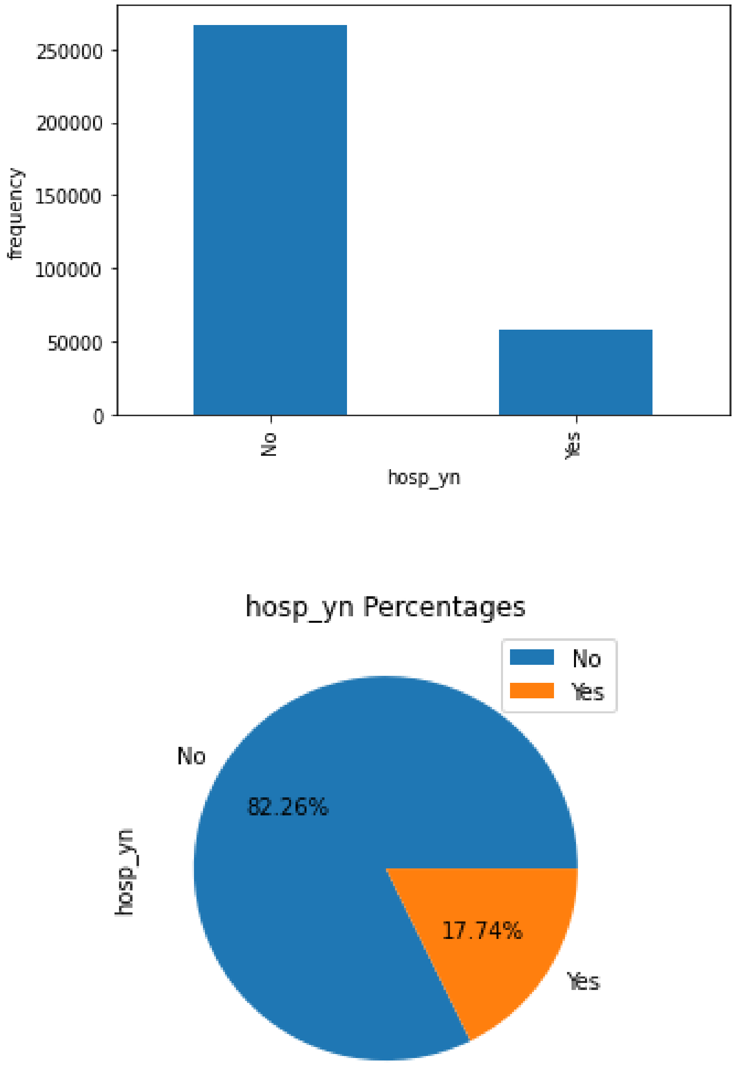


Figure 4

From the results above we can see that almost only about 17.7% of the patients of the sample were hospitalized.

5.5- icu_yn Feature

The distribution of the values is described in the figure below:

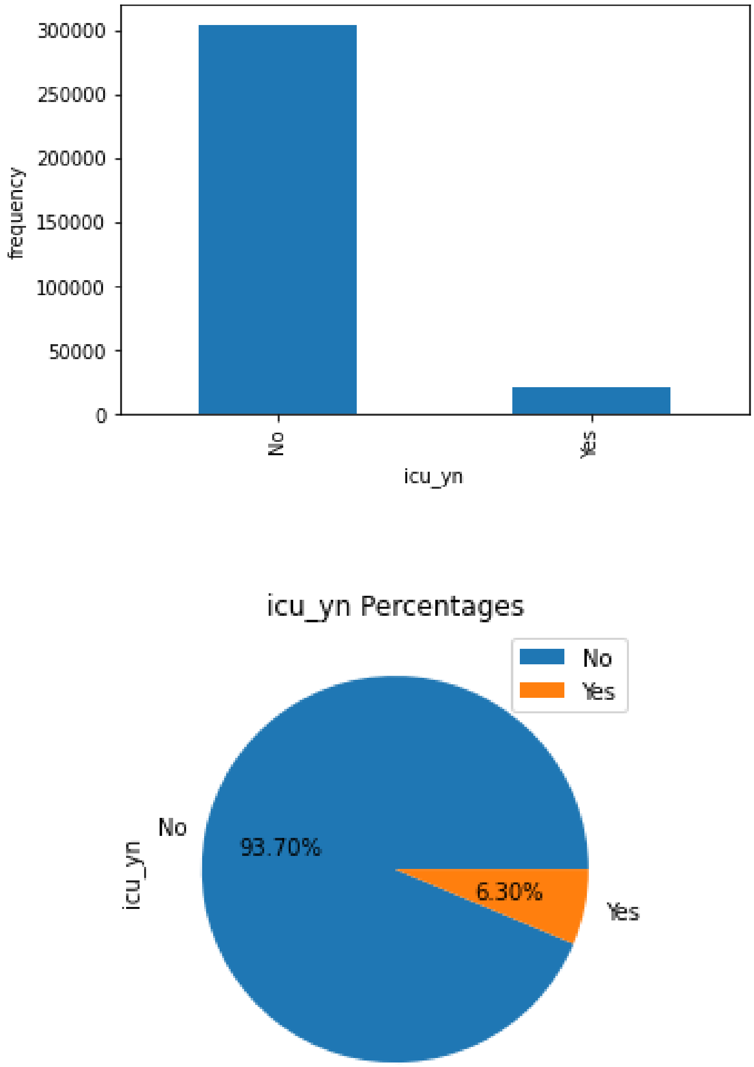


Figure 5

The results above show that almost about 6.3% of the sample patients were ICU admitted.

5.6- death_yn Feature

The distribution of the values is described in the figure below:

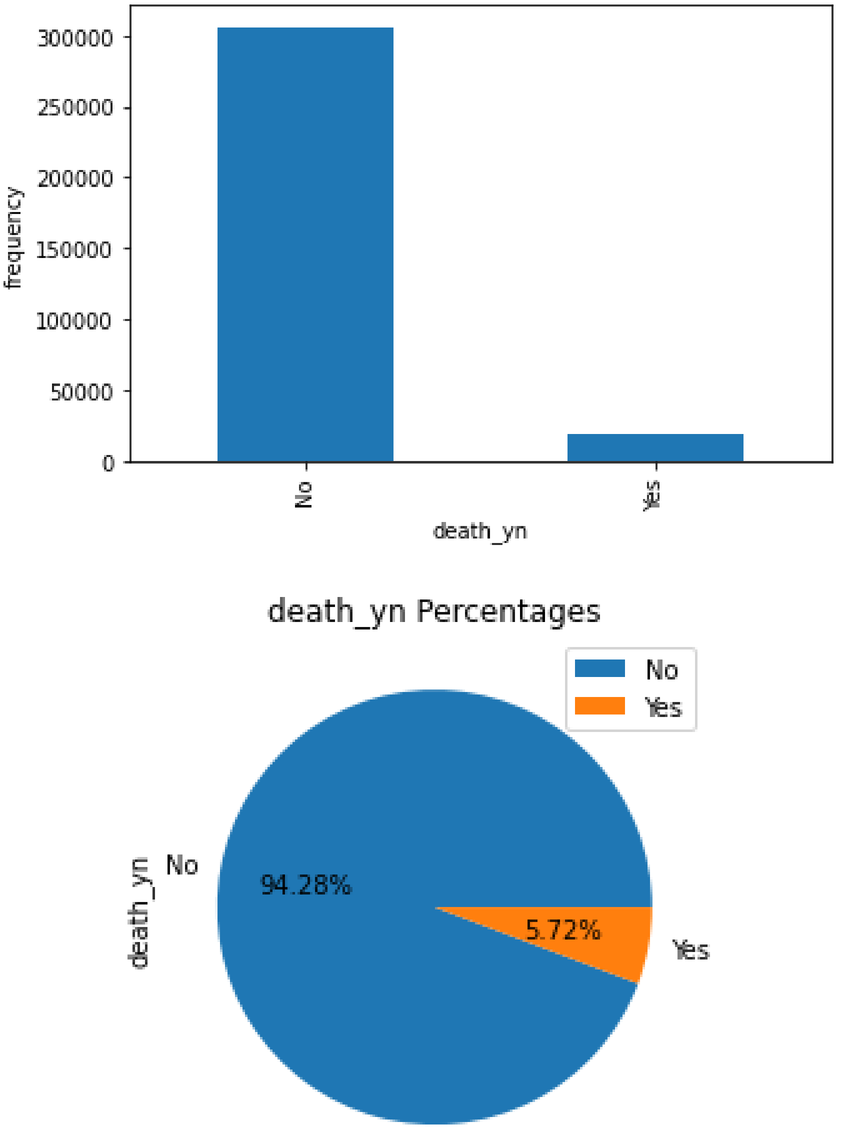


Figure 6

The results above show that the mortality rate of the patients in our sample is 5.72%.

5.7- medcond_yn Feature

The distribution of the values is described in the figure below:

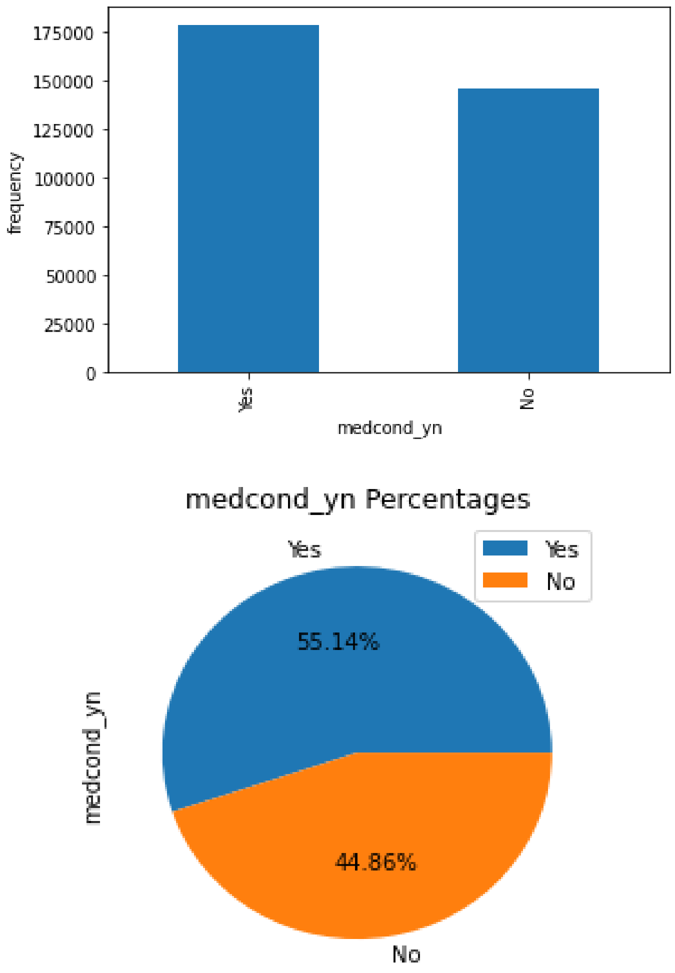


Figure 7

From the results above, we can see that a huge number of the patients in our sample have medical conditions (almost about 55%). Many reasons might have caused this effect: maybe the people who have medical conditions are more likely to be infected, or more likely to have serious symptoms (i.e. They have to go to hospital, which means that their cases are more likely to be confirmed). Or maybe because the American population contains a lot of people who have medical conditions. Or maybe the health care providers in USA make a lot of tests for the people who have medical conditions (like in nursing homes) to check if they have covid-19. Or maybe the data of the patients who have medical conditions are well recorded (i.e. without null, missing or unknown values) so their percentage increased in our sample., or...

6- Multivariate Analysis

As our goal is studying the relationships between our features and the mortality rate, in this section we will study only the relationships between death_yn and all the other features.

In order to simplify the study, we will not study all the other features pairwise relationships.

6.1- sex and death_yn

To understand the relationship between the sex and the death_yn which are categorical variables (Boolean), we checked the mortality rate depending on each possible value of the sex.

P.S. mortality rate is the death rate. In other words, in a sample of patients equals to the number of deaths in the sample divide by the sample size. And this rate can be visualized using probability mass function.

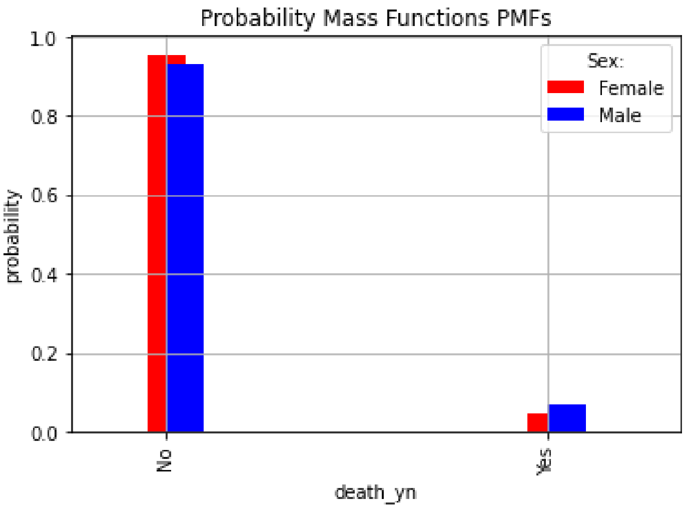


Figure 8

From the results above, we can see that even though that the females have appeared more in our sample (as explained in section 5.1) they might have a lower mortality rate.

6.2- age_group and death_yn

To understand the relationship between the age_group and the death_yn which are category variables, we checked the mortality rate depending on each possible value of the age_group.

The figure below shows the relation between the mortality rate and the age_group:

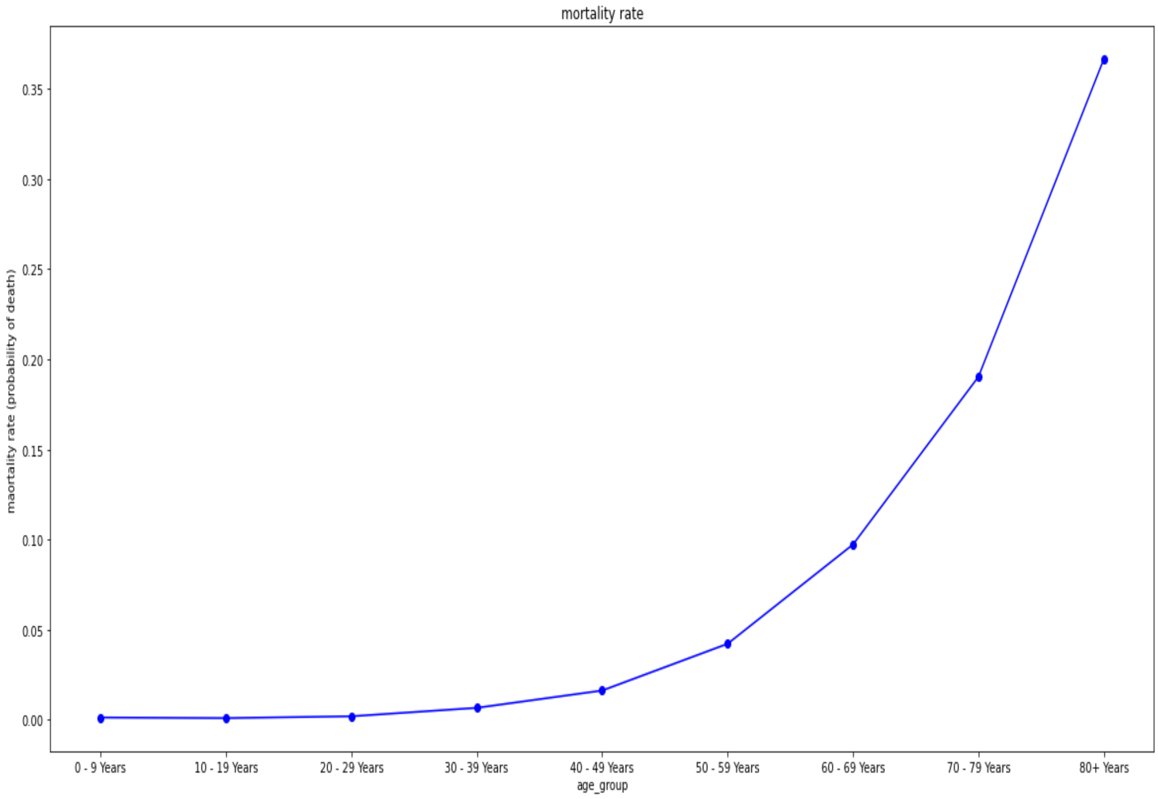


Figure 9

From the figure above we can see that the mortality rate curve for our sample is almost flattened for the young ages, but for the older patients (starting from the '40-49 years' slice) the mortality rate increase exponentially.

6.3- Race and ethnicity (combined) vs death_yn

In this section we plotted the mortality rate by race and ethnicity group, and the result was the figure below:

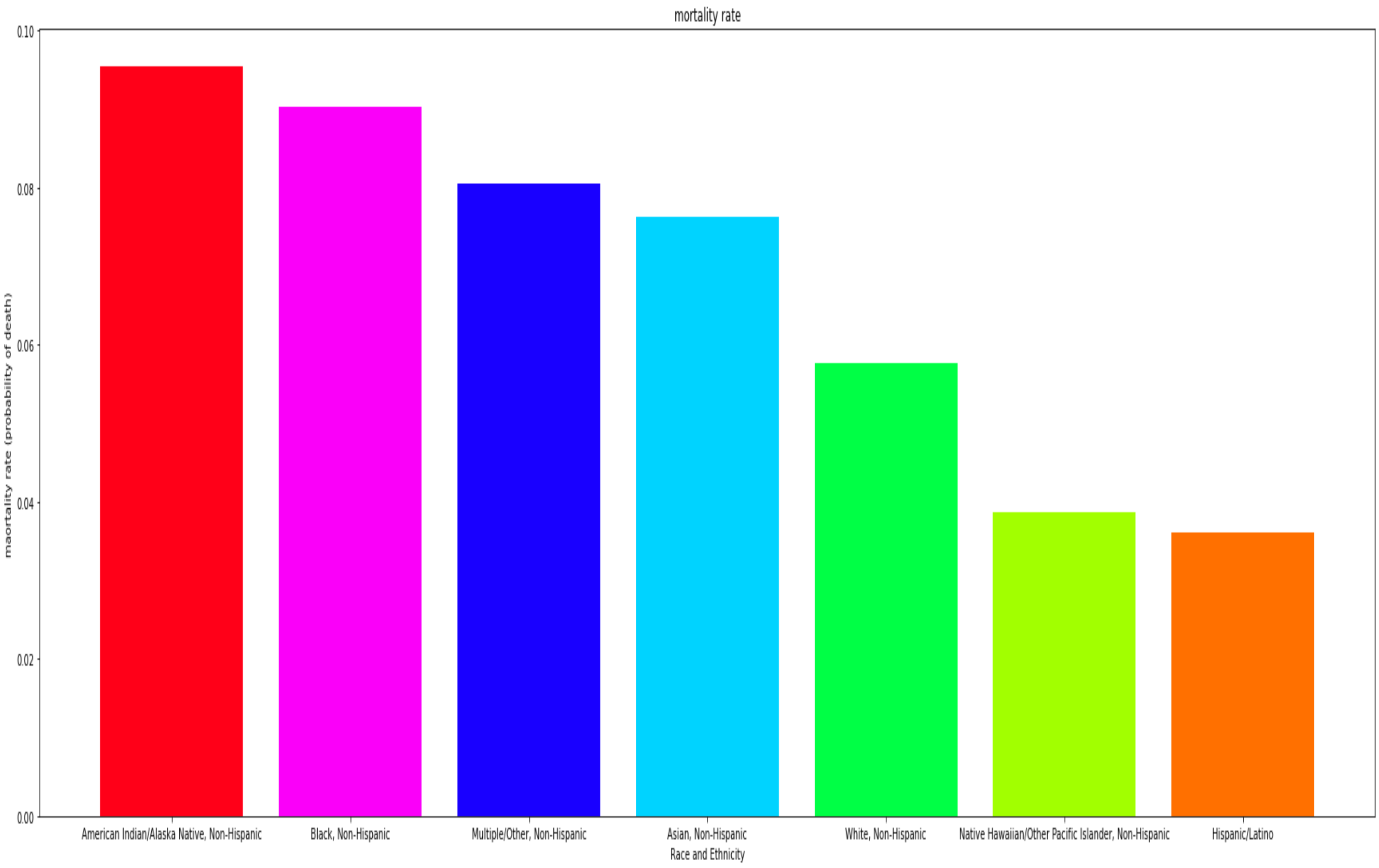


Figure 10

Apparently, the figure above tells us that the "American Indian/Alaska Native, Non-Hispanic" and the "Black Non-Hispanic" have a higher mortality rate than the others. Also it tells us that the "Hispanic/ Latino" group has the lowest mortality rate in our sample.

6.4- hosp_yn and death_yn

To understand the relationship between the hosp_yn and the death_yn which are Boolean variables, we checked the mortality rate depending on the values of the hosp_yn.

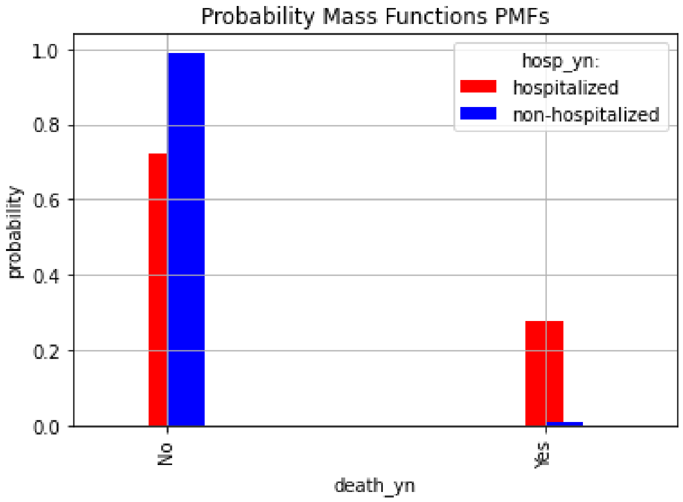


Figure 11

The result above tells us that the patients who don't need to be hospitalized might be less likely to die. That could be because that most the patients who die might have serious symptoms requires the hospitalization.

6.5- icu_yn and death_yn

To understand the relationship between the icu_yn and the death_yn which are Boolean variables, we checked the mortality rate depending on the values of the icu_yn.

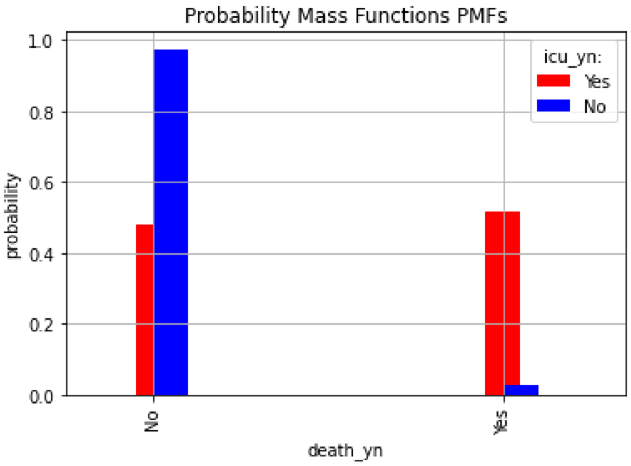


Figure 12

From the results above, in our sample, the mortality rate of the patients that have been admitted to ICU is very high ~52%, while it is less than 3% for the non-admitted ones.

6.6- medcond_yn and death_yn

To understand the relationship between the medcond_yn and the death_yn which are Boolean variables, we checked the mortality rate depending on the values of the medcond_yn.

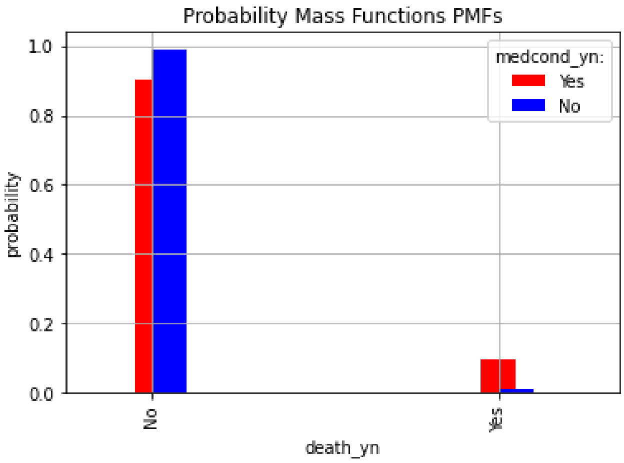


Figure 13

The result above tells us that the patients without medical conditions might be less likely to die.

7- Hypothesis Testing

In the previous section, we noticed that the mortality rate of the patients, in our sample has been affected by all the studied factors.

In this section we will propose and test some hypotheses about the effects that have mentioned above.

In other words, we will try to answer the following question: "**Given a sample and an apparent effect, what is the probability of seeing such an effect by chance?**"

7.1- Hypothesis: patients who are older than 60 years have a higher mortality rate

- **Our Hypothesis**: covied-19 patients who are older than 60 years have a higher mortality rate than the others.
- **Test Statistic**: For a **sub-sample** which has the size **C(sub-sample)** and the number of dead patients **D(sub-sample)**, the test statistic of this sub-sample is: mortality rate = **TS(sub-sample) = D(sub-sample)/C(sub-sample)**. where **C(sub-sample)** in our case is always equal to the count of the patients who are older than 60 years in our original cleaned sample.
- **Null Hypothesis**: patients who are older than 60 years have the same mortality rate of the other patients.
- **Computing P-Value**: our test is one-tailed. To compute the P-Value we generated the distribution of our test statistic **TS** by repeating taking a random 'sub-sample" of size **C** from our original cleaned sample and calculate the corresponding **TS** of this random "sub-sample". After generating the mentioned distribution, we

calculated **P-Value = (100-percentile rank(TSH))/100**, where **TSH** is the test statistic that corresponds to the patients who are older than 60 years in our original cleaned sample, i.e. "**TSH=TS (patients who are older than 60 years in our original cleaned sample)**".

- **The Results of the Test**: the figure below shows the generated cumulative density function CDF of the test statistic and the calculated test statistic of the patients who are older than 60 years in our sample (TSH):

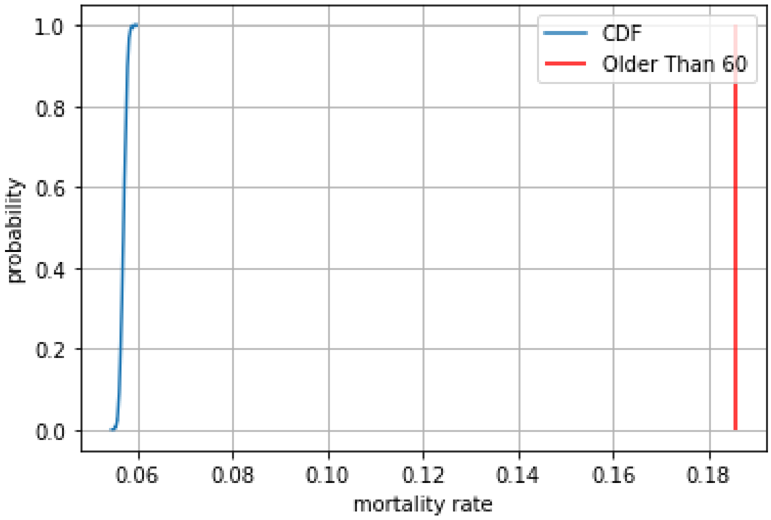


Figure 14

In the figure above we can see that the test statistic that corresponds to the patients who are older than 60 years in our sample (TSH) is far away from the distribution of the test statistics (TS) of the samples that have the same size. And by calculating P-Value we find that P-Value=0, which means the effect is significant and we have to accept our hypothesis.

7.2- Hypothesis: Males have a higher mortality rate

- **Our Hypothesis**: Male covied-19 patients have a higher mortality rate than the others.
- **Test Statistic**: For a **sub-sample** which has the size **C(sub-sample)** and the number of dead patients **D(sub-sample)**, the test statistic of this sub-sample is: mortality rate=**TS(sub-sample) = D(sub-sample)/C(sub-sample)**. where **C(sub-sample)** in our case is always equal to the count of the male patients in our original cleaned sample.
- **Null Hypothesis**: male patients have the same mortality rate of the other patients.
- **Computing P-Value**: our test is one-tailed. To compute the P-Value we will generate the distribution of our test statistic **TS** by repeating taking a random 'sub-sample" of size **C** from our original cleaned sample and calculate the corresponding **TS** of this random "sub-sample". After generating the mentioned distribution we can calculate **P-Value = (100-percentile rank(TSH))/100**, where **TSH** is the test statistic that corresponds to the male patients in our original cleaned sample, i.e. "**TSH = TS(male patients in our original cleaned sample)**".
- **The Results of the Test**: the figure below shows the generated cumulative density function CDF of the test statistic and the calculated test statistic of the Male patients (TSH):

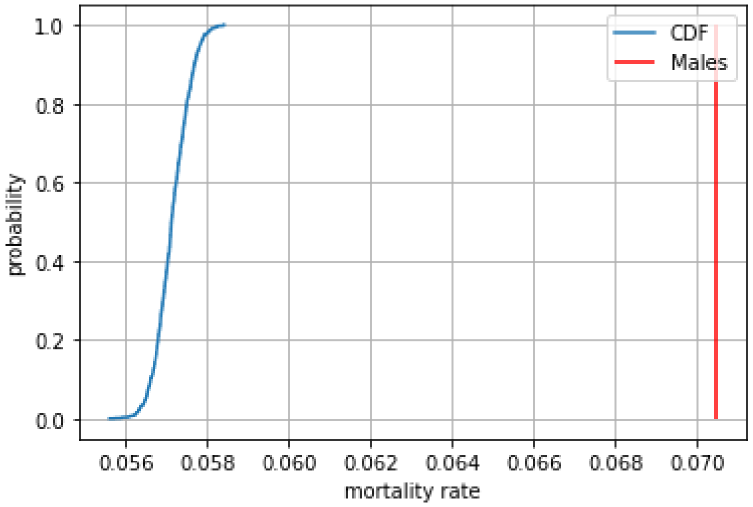


Figure 15

In the figure above we can see that the test statistic that corresponds to the male patients in our sample (TSH) is far away from the distribution of the test statistics (TS) of the samples that have the same size. And by calculating P-Value we find that P-Value=0, which means the effect is significant and we have to accept our hypothesis.

7.3- Hypothesis: "Hispanic/Latino" patients have a lower mortality rate

- **Our Hypothesis**: "Hispanic/Latino" covied-19 patients have a higher mortality rate than the others.
- **Test Statistic**: For a **sub-sample** which has the size **C(sub-sample)** and the number of dead patients **D(sub-sample)**, the test statistic of this sub-sample is: mortality rate=**TS(sub-sample) = D(sub-sample)/C(sub-sample)**. where **C(sub-sample)** in our case is always equal to the count of the "Hispanic/Latino" patients in our original cleaned sample.
- **Null Hypothesis**: "Hispanic/Latino" patients have the same mortality rate of the other patients.
- **Computing P-Value**: our test is one-tailed. To compute the P-Value we will generate the distribution of our test statistic **TS** by repeating taking a random 'sub-sample" of size **C** from our original cleaned sample and calculate the corresponding **TS** of this random "sub-sample". After generating the mentioned distribution we can calculate **P-Value = (percentile rank(TSH))/100**, where **TSH** is the test statistic that corresponds to the "Hispanic/Latino" patients in our original cleaned sample, i.e. "**TSH = TS("Hispanic/Latino" patients in our original cleaned sample)**".
- **The Results of the Test**: the figure below shows the generated cumulative density function CDF of the test statistic and the calculated test statistic of the "Hispanic/Latino" patients (TSH):

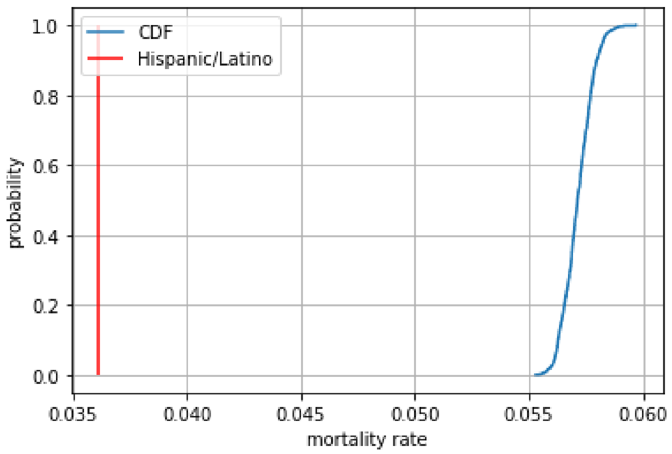


Figure 16

In the figure above we can see that the test statistic that corresponds to the "Hispanic/Latino" patients in our sample (TSH) is far away from the distribution of the test statistics (TS) of the samples that have the same size. And by calculating P-Value we find that P-Value=0, which means the effect is significant and we have to accept our hypothesis.

7.4- Hypothesis: Patients who have medical conditions have a higher mortality rate

- **Our Hypothesis**: covide-19 patients who have medical conditions have a higher mortality rate than the others.
- **Test Statistic**: For a **sub-sample** which has the size **C(sub-sample)** and the number of dead patients **D(sub-sample)**, the test statistic of this sub-sample is: mortality rate=**TS(sub-sample) = D(sub-sample)/C(sub-sample)**. where **C(sub-sample)** in our case is always equal to the count of the covide-19 patients who have medical conditions in our original cleaned sample.
- **Null Hypothesis**: covide-19 patients who have medical conditions have the same mortality rate of the other patients.
- **Computing P-Value**: our test is one-tailed. To compute the P-Value we will generate the distribution of our test statistic **TS** by repeating taking a random 'sub-sample" of size **C** from our original cleaned sample and calculate the corresponding **TS** of this random "sub-sample". After generating the mentioned distribution we can calculate **P-Value = (100 - percentile rank(TSH))/100**, where **TSH** is the test statistic that corresponds to the covide-19 patients who have medical conditions in our original cleaned sample, i.e. "**TSH = TS(the patients who have medical conditions in our original cleaned sample)**".
- **The Results of the Test**: the figure below shows the generated cumulative density function CDF of the test statistic and the calculated test statistic of the patients who have medical conditions (TSH):

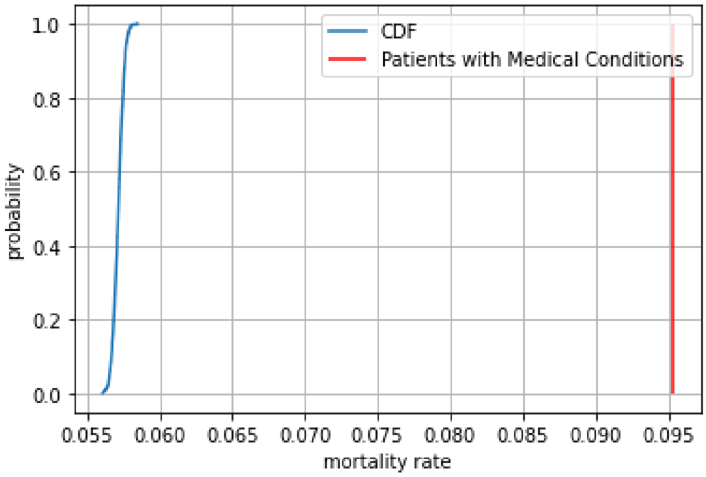


Figure 17

In the figure above we can see that the test statistic that corresponds to the patients who have medical conditions in our sample (TSH) is far away from the distribution of the test statistics (TS) of the samples that have the same size. And by calculating P-Value we find that P-Value=0, which means the effect is significant and we have to accept our hypothesis.

7.5- Hypothesis: Hospitalized patients have a higher mortality rate

- **Our Hypothesis**: covied-19 patients who have been hospitalized have a higher mortality rate than the others.
- **Test Statistic**: For a **sub-sample** which has the size **C(sub-sample)** and the number of dead patients **D(sub-sample)** the test statistic of this sub-sample is: mortality rate=**TS(sub-sample) = D(sub-sample)/C(sub-sample)**. where **C(sub-sample)** in our case is always equal to the count of the patients who have been hospitalized in our original cleaned sample.
- **Null Hypothesis**: patients who have been hospitalized have the same mortality rate of the other patients.
- **Computing P-Value**: our test is one-tailed. To compute the P-Value we will generate the distribution of our test statistic **TS** by repeating taking a random 'sub-sample" of size **C** from our original cleaned sample and calculate the corresponding **TS** of this random "sub-sample". After generating the mentioned distribution we can calculate **P-Value = (100 - percentile rank(TSH))/100**, where **TSH** is the test statistic that corresponds to the patients who have been hospitalized in our original cleaned sample, i.e. "**TSH = TS(patients who have been hospitalized in our original cleaned sample)**".
- **The Results of the Test**: the figure below shows the generated cumulative density function CDF of the test statistic and the calculated test statistic of the patients who have medical conditions (TSH):

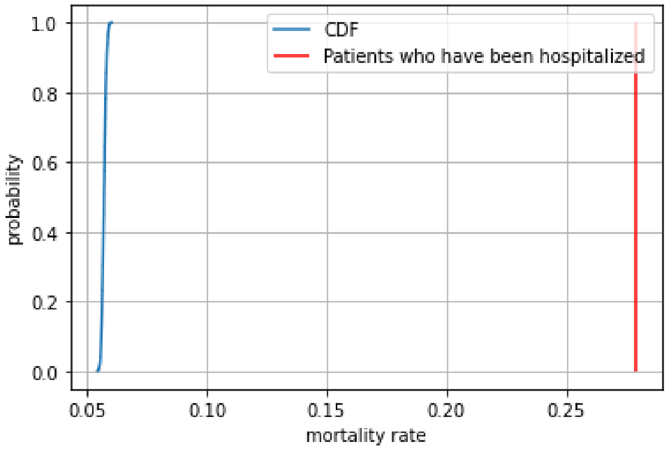


Figure 18

In the figure above we can see that the test statistic that corresponds to the patients who have been hospitalized in our sample (TSH) is far away from the distribution of the test statistics (TS) of the samples that have the same size. And by calculating P-Value we find that P-Value=0, which means the effect is significant and we have to accept our hypothesis.

7.6- Hypothesis: Patients who have been admitted to ICU have a higher mortality rate.

- **Our Hypothesis**: covied-19 patients who have been admitted to ICU have a higher mortality rate than the others.
- **Test Statistic**: For a **sub-sample** which has the size **C(sub-sample)** and the number of dead patients **D(sub-sample)** the test statistic of this sub-sample is: mortality rate=**TS(sub-sample) = D(sub-sample)/C(sub-sample)**. where **C(sub-sample)** in our case is always equal to the count of the patients who have been admitted to ICU in our original cleaned sample.
- **Null Hypothesis**: patients who have been admitted to ICU have the same mortality rate of the other patients.
- **Computing P-Value**: our test is one-tailed. To compute the P-Value we will generate the distribution of our test statistic **TS** by repeating taking a random 'sub-sample" of size **C** from our original cleaned sample and calculate the corresponding **TS** of this random "sub-sample". After generating the mentioned distribution we can calculate **P-Value = (100-percentile rank(TSH))/100**, where **TSH** is the test statistic that corresponds to the patients who have been admitted to ICU in our original cleaned sample, i.e. "**TSH = TS(patients who have been** admitted **to ICU in our original cleaned sample)**".
- **The Results of the Test**: the figure below shows the generated cumulative density function CDF of the test statistic and the calculated test statistic of the patients who have been admitted to ICU (TSH):

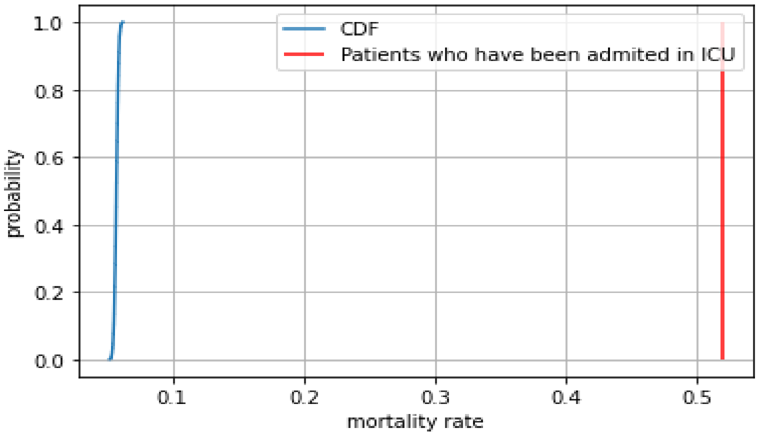


Figure 19

In the figure above we can see that the test statistic that corresponds to the patients who have been admitted to ICU in our sample (TSH) is far away from the distribution of the test statistics (TS) of the samples that have the same size. And by calculating P-Value we find that P-Value=0, which means the effect is significant and we have to accept our hypothesis.

Finally, we have to mention that, in all the six hypothesis tests above, in order to calculate the P-Value, we have generated the distribution of the test statistic TS.
But it was possible to get rid of this generation process of the distribution, by depending on the central limit theorem. i.e. our test statistic TS in all the six hypothesis represents the mortality rate in the test sample which is actually the mean of the binomial variable death_yn (if we considered this variable a binomial one with 0 and 1 values).
Accordingly, TS is normally distributed with a mean equals to the mean of our final cleaned sample and a standard deviation equals to the standard deviation of our final cleaned sample divided by the squared root of the test sample size.

8- Limitations

The limitations of our study are:

1. The first limitation: there are a lot of covid-19 patients who don't have symptoms, which means that they are less likely to appear in our original CDC data and their cases are less likely to be confirmed. And this is a big issue, because it can cause a considerable sampling bias. For example, it might cause increasing the mortality rate in the sample. In addition to this issue there are other limitation of our original CDC data as mentioned on CDC website.
2. The second limitation: in our study we ignored the unknown and missing value which also might bias our final cleaned sample.
3. The third limitation: despite of the fact that we have some highly correlated features (like icu_yn and hosp_yn), we didn't study the pairwise relationships between the features because this issue was not one of the goals of this study.

9- Conclusion and Future Work

Despite the limitations of this study, this report concludes that the mortality rate is significantly related to the age, sex, race and ethnicity, patient hospitalization and patient admission to ICU.

Our future work is studying the pairwise relationships between the features. And after that we might apply some feature engineering, for example we have to merge the hierarchically related features into one feature, i.e. hosp_yn and icu_yn must be merged in one feature. After that we think we can build a decision tree based approach to predict the death_yn of a patient depending on the resulting features. But here we have to worry about the limitations that mentioned in the previous section, so collecting another data from the population in a more random and a more accurate manner will provide us with a more accurate and a more unbiased prediction than what we expect to get by using the current CDC data.