**Analyzing Corona virus deaths based on pre-existing illness:**

**Vishwabharan**, CB.EN.U4ECE17065.

**Viswavaageesh Chandramouli**, CB.EN.U4ECE17066

***SARS-CoV-2***, a novel corona virus which causes ***COVID-19***, the respiratory illness responsible for the COVID-19 pandemic. The World Health Organization declared the outbreak a Public Health Emergency of International Concern on 30 January 2020, and a pandemic on 11 March 2020. Until today, the virus has infected over 25 million, and claimed the life of nearly 800,000 people.

Thus, in an attempt to analyse the effects of pre-existing morbidities such as pneumonia,

Diabetes, hypertension, Obesity, Tobacco on deaths related to ***COVID-19*** we have designed this project. This is done by fitting a Linear regression model and getting the susceptibility percentage with confidence value (i.e. a range of plausible values for an unknown parameter) of 95% because a **95**% **confidence interval** is a range of values that you can be **95**% certain contains the true **mean** of the population

Data is being collected by the Mexican government, it contains a large number of anonymized patient-related information such as pre-existing conditions, gender, age, symptom date and entry date. The data contains patient information around the world

The project is divided into three phases:

***Phase I:***

**Introduction of the data set:**

Before doing any analysis, data schema is looked at carefully , format and missing values.

The data set contained 566602 records but was reduced because of size constraints , the data dictionary is provided as below:

1. **id**: The identification number of the patient.

2. **sex**: Identify gender of the patient, 1 as female and 2 as male.

3. **patient\_type**: Type of patient, 1 for not hospitalized and 2 for hosptalized.

4. **entry\_date**: The date that the patient went to the hospital.

5. **date\_symptoms**: The date that the patient started to show symptoms.

6. **date\_died**: The date that the patient died, "9999-99-99" stands for recovered.

7. **intubed**: Intubation is a procedure that's used when you can't breathe on your own. Your doctor puts a tube down your throat and into your windpipe to make it easier to get air into and out of your lungs. A machine called a ventilator pumps in air with extra oxygen. Then it helps you breathe out air that’s full of carbon dioxide (CO2). "1" denotes that the patient used ventilator and "2" denotes that the patient did not, "97" "98" "99" means not specified.

8. **pneumonia**: Indicates whether the patient already have air sacs inflammation or not "1" for yes, "2" for no, "97" "98" "99" means not specified.

9. **age**: Specifies the age of the patient.

10. **pregnancy**: Indicates whether the patient is pregnant or not, "1" for yes, "2" for no, "97" "98" "99" means not specified.

11. **diabetes**: Indicates whether the patient has diabetes or not, "1" for yes, "2" for no, "97" "98" "99" means not specified.

12. **copd**: Indicates whether the patient has Chronic obstructive pulmonary disease (COPD) or not, "1" for yes, "2" for no, "97" "98" "99" means not specified.

13. **asthma**: Indicates whether the patient has asthma or not, "1" for yes, "2" for no, "97" "98" "99" means not specified.

14. **inmsupr**: Indicates whether the patient is immunosuppressed or not, "1" for yes, "2" for no, "97" "98" "99" means not specified.

15. **hypertension**: Indicates whether the patient has hypertension or not, "1" for yes, "2" for no, "97" "98" "99" means not specified.

16. **other\_disease**: Indicates whether the patient has other disease or not, "1" for yes, "2" for no, "97" "98" "99" means not specified.

17. **cardiovascular**: Indicates whether if the patient has heart or blood vessels realted disease, "1" for yes, "2" for no, "97" "98" "99" means not specified.

18. **obesity**: Indicates whether the patient is obese or not,  "1" for yes, "2" for no, "97" "98" "99" means not specified.

19. **renal\_chronic**: Indicates whether the patient has chronic renal disease or not,  "1" for yes, "2" for no, "97" "98" "99" means not specified.

20. **tobacco**: Indicates whether if the patient is a tobacco user, "1" for yes, "2" for no, "97" "98" "99" means not specified.

21. **contact\_other\_covid**: Indicates whether if the patient has contacted another covid19 patient.

22. **icu**: Indicates whether the if the patient had been admitted to an Intensive Care Unit (ICU), "1" for yes, "2" for no, "97" "98" "99" means not specified.

Data is looked at carefully and a new columns are created namely:

* Duration: We define duration as the time (day) between the day the symptoms started showing and the day the patient went to the hospital for check up

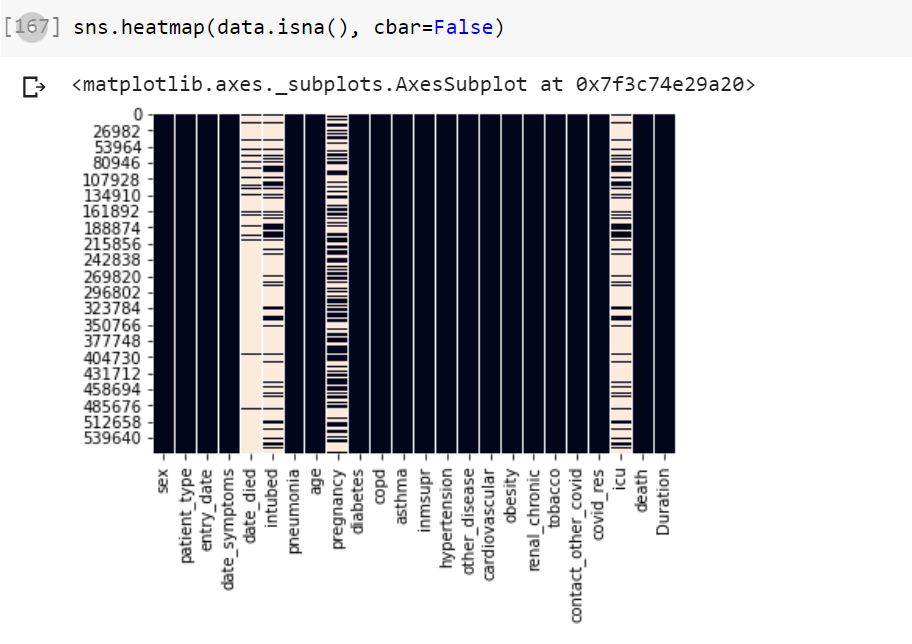
*Death:* this variable takes value of **1** if the patient died and **0** otherwise

***For Phase II:***

**Data cleaning and visualization:**

Data is then further cleaned as several variables have values that are not useful for processing which will be discorded and a data visualization of missing data is done to have an idea of which variable to use for analysis based on missing and erroneous data.

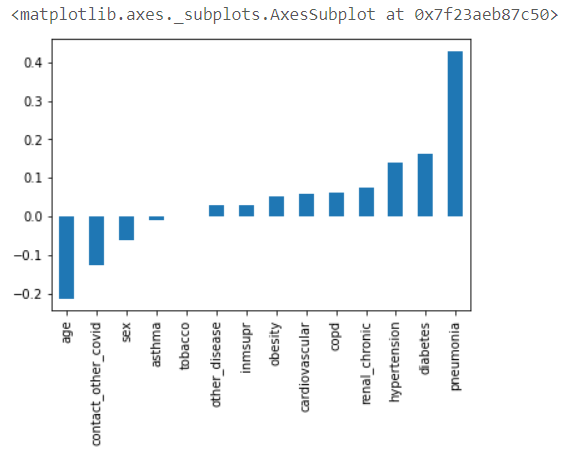
Data Visualization is done using the SNS Heat map and a histogram of missing data.



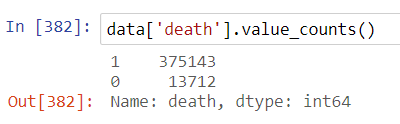
Its clearly visible *that intubed, pregnancy, icu, Date\_symptoms and date\_died* have the most missing values and thus these columns are removed.

**Correlation between different parameters to death:**

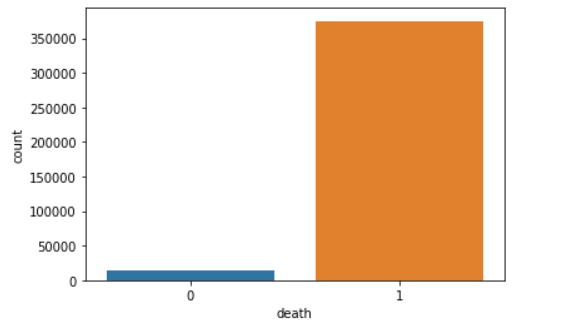
After removing all the NA values the correlation between the different parameters to death is found. This is accomplished by using corr() function available which finds the pairwise correlation of all columns in the data frame.



**Visualization of number of people who died:**

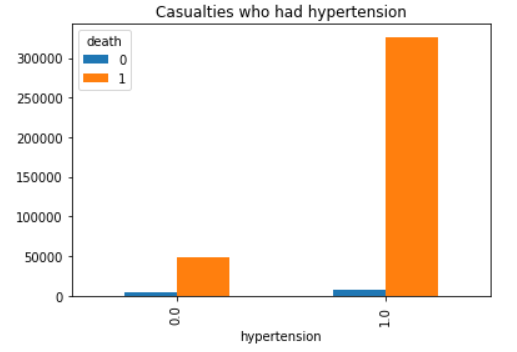






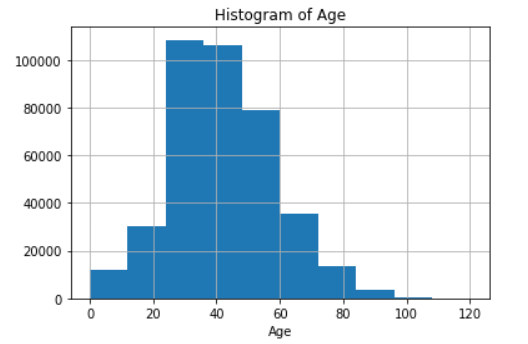
**Particular factors role in death:**

This is done by cross tabulating two factors. For example let us take casualties who had hypertension.



**Histogram of death varied with age:**

We plot a histogram of deaths varied with age.



From this histogram we can infer that most of the deaths is between the age group 24 to 45.

***For Phase III:***

**Fitting the logistic regression model and interpretation**

In this project we used logistic regression to model the relationship between independent variables and response variable.

95% Confidence interval for odds difference between levels (in percentage) is calculated for variables that we found to have proper data and further interpretation of parameters are done.

**Precision Report:**

Accuracy of logistic regression classifier on test set is 0.96



The dotted line represents the ROC curve of a purely random classifier. A good classifier stays far away from that line as possible that is to say, toward the top-left corner.