Cause of Deaths 2020

Research Question

- Is the Coronavirus a real global pandemic?

How was the data collected?

- Weekly counts of deaths by jurisdiction and cause of death
  - Provided by the CDC. Downloaded and accessed as a CSV file. Can be accessed here.
  - 329988 rows × 15 columns. Variables include jurisdiction, cause group, and number of deaths.
  - Used to calculate the number of deaths per U.S. state per year, from 2015 – 2020.

Data Cleaning

- The dataset used contained unneeded variables that were not needed to answer the research question, so these variables were dropped.
- Missing values for number of deaths variable were handled by replacing the missing values with the mean(average) of that jurisdiction for all years.

Data Analysis

Let's look at the first few rows of the dataset we are working with. We only need the columns "Jurisdiction", "Year", "Cause.Group", and "Number.of.Deaths".

```
##   Jurisdiction Year                    Cause.Group Number.of.Deaths
## 1      Alabama 2015 Alzheimer disease and dementia              120
## 2      Alabama 2015 Alzheimer disease and dementia              120
## 3      Alabama 2016 Alzheimer disease and dementia               76
## 4      Alabama 2016 Alzheimer disease and dementia               76
## 5      Alabama 2017 Alzheimer disease and dementia               96
## 6      Alabama 2017 Alzheimer disease and dementia               96
```

Let’s check for null values.

```
sum(is.na(data$Number.of.Deaths))

## [1] 34
```

There are 34 null values. Lets look at some of these null values.

```
head(filter(data, is.na((data$Number.of.Deaths))))

##     Jurisdiction Year                    Cause.Group Number.of.Deaths
## 1        Indiana 2020 Alzheimer disease and dementia               NA
## 2 North Carolina 2020 Alzheimer disease and dementia               NA
## 3 North Carolina 2020 Alzheimer disease and dementia               NA
## 4 North Carolina 2020 Alzheimer disease and dementia               NA
## 5        Indiana 2020           Circulatory diseases               NA
## 6 North Carolina 2020           Circulatory diseases               NA
```

Let’s fill these values with the mean per state per cause of all years and check for null values again.

```
sum(is.na(data$Number.of.Deaths))

## [1] 0
```

Since there are no more null values in our data, we can continue with our analysis. There are rows in our data where Jurisdiction is “United States”. Lets look at some of these values.

```
head(filter(data, data$Jurisdiction == "United States"))

##    Jurisdiction Year                    Cause.Group Number.of.Deaths
## 1 United States 2015 Alzheimer disease and dementia             6187
## 2 United States 2015 Alzheimer disease and dementia             6187
## 3 United States 2016 Alzheimer disease and dementia             5155
## 4 United States 2016 Alzheimer disease and dementia             5155
## 5 United States 2017 Alzheimer disease and dementia             5844
## 6 United States 2017 Alzheimer disease and dementia             5844

nrow(filter(data, data$Jurisdiction == "United States"))

## [1] 8138
```

We need specific Jurisdiction locations so we will drop these rows.

```
## Number of rows before drop:  329988
## Number of rows after drop:  321850
```

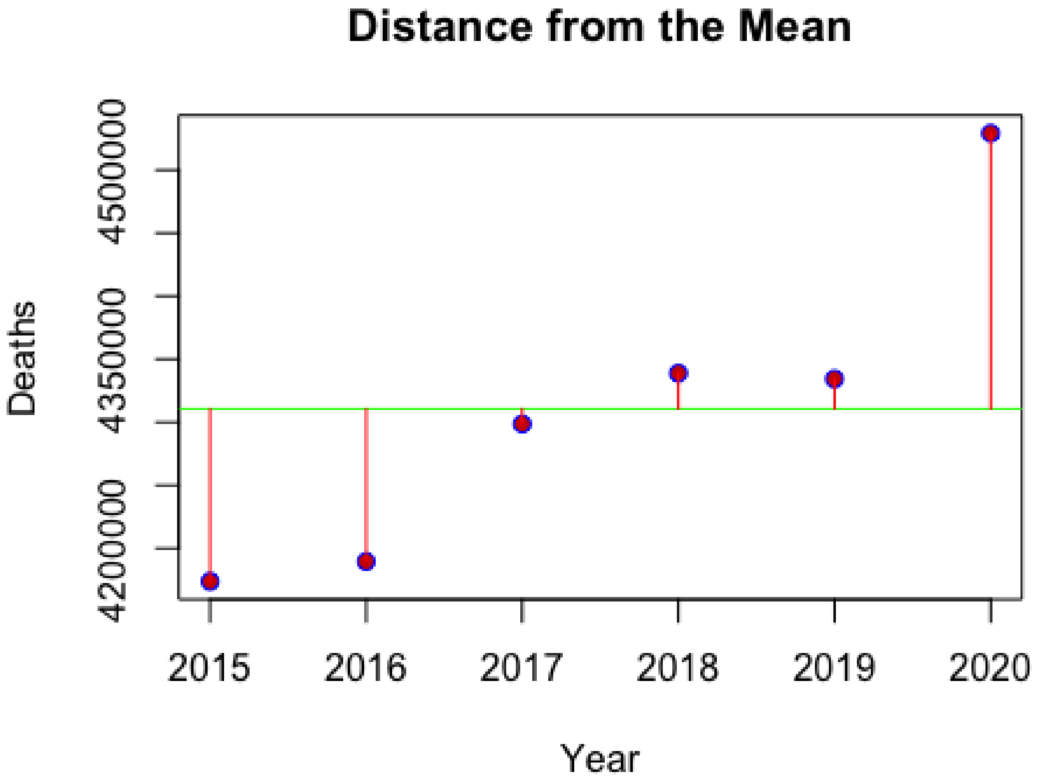
Let’s sum the total deaths and total deaths per year.

```
## Total Deaths:  25864451
```

Let’s look at the new dataframe

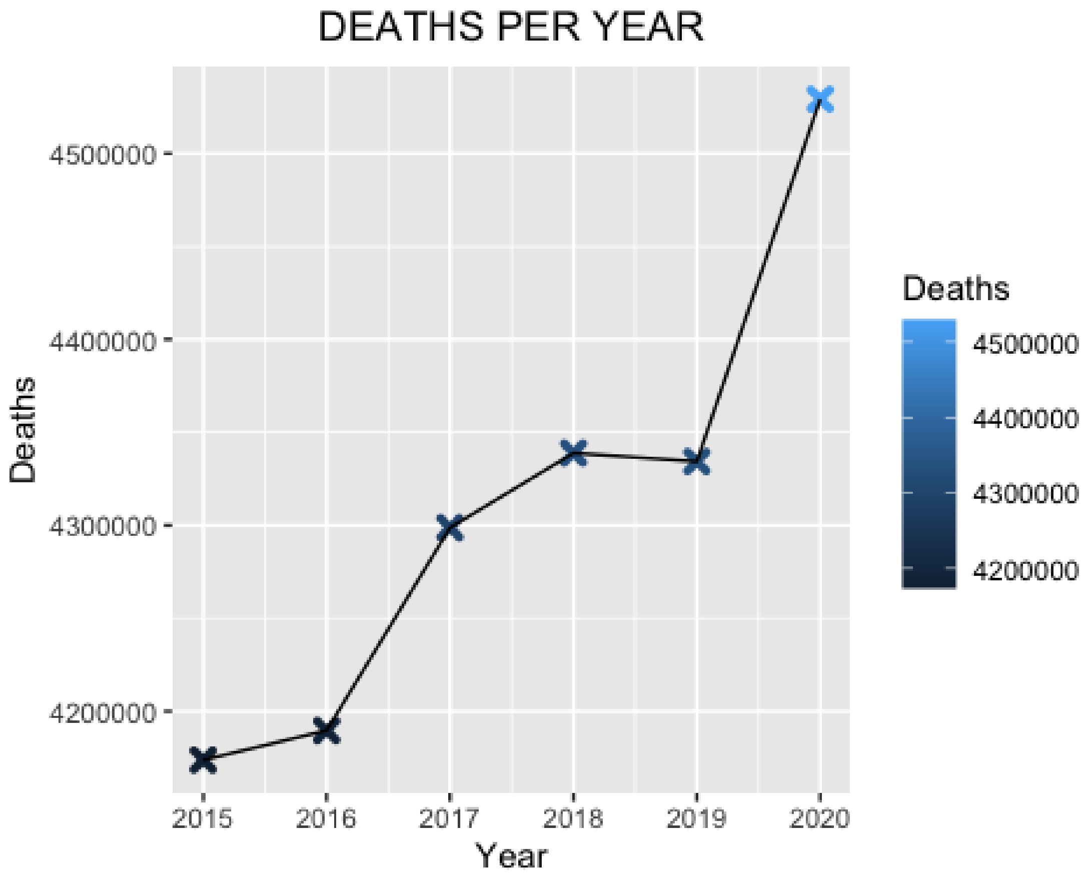
```
##   Year  Deaths Death Rate
## 1 2015 4173810  0.1613725
## 2 2016 4189572  0.1619819
## 3 2017 4298704  0.1662012
## 4 2018 4338856  0.1677536
## 5 2019 4334354  0.1675796
## 6 2020 4529155  0.1751112
```

Let’s plot the distance from the mean for number of deaths per year and the total deaths per year.



```
##   Year    Mean Deaths Above or Below the Mean
## 1 2015 4310742                     -136931.83
## 2 2016 4310742                     -121169.83
## 3 2017 4310742                      -12037.83
## 4 2018 4310742                       28114.17
## 5 2019 4310742                       23612.17
## 6 2020 4310742                      218413.17
```

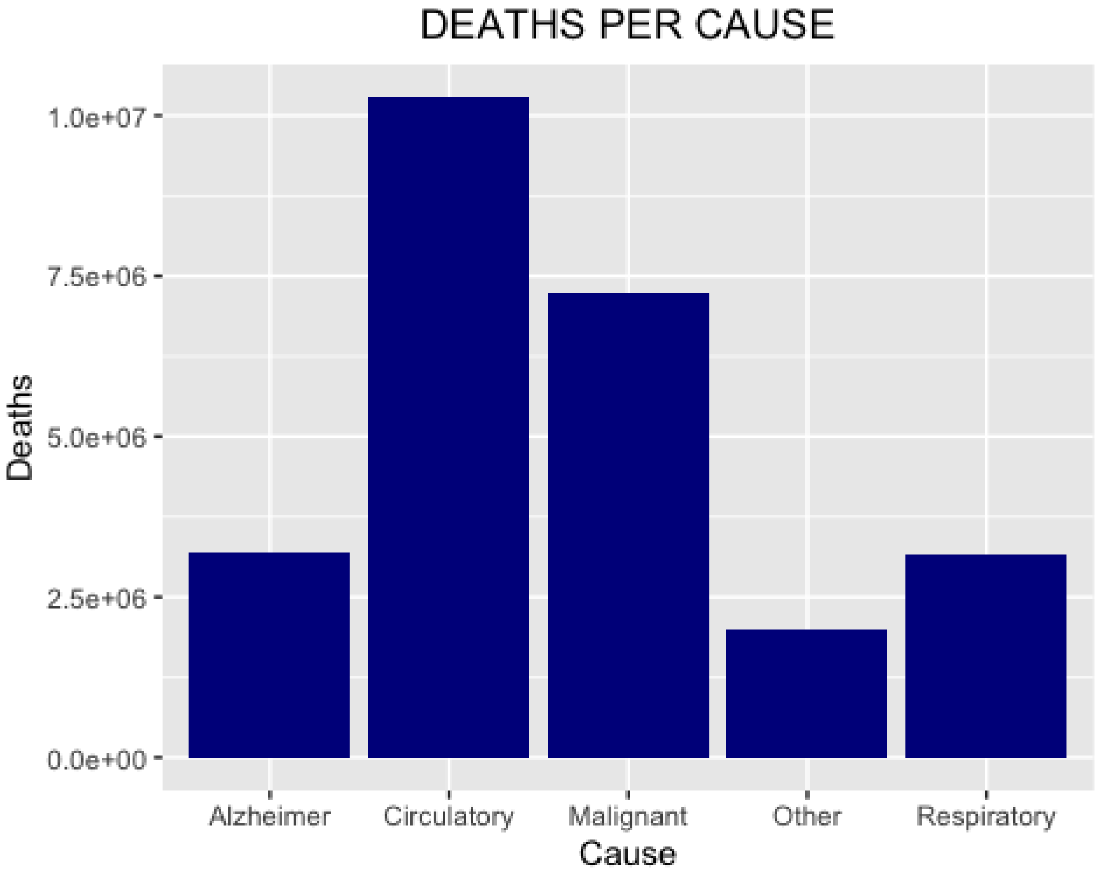
The plot below shows that the number of deaths increase every year. I am just guessing but I believe that this is because the population also increase every year. We will prove that in a different analysis.



Let’s look at the total deaths per cause, with the death rate, and the rank.

```
##               Deaths Death.Rate Rank
## Circulatory 10286544  0.3977097    1
## Malignant    7232838  0.2796440    2
## Alzheimer    3204864  0.1239100    3
## Respiratory  3162371  0.1222671    4
## Other        1977834  0.0764692    5
```

Let’s plot the total deaths by cause.

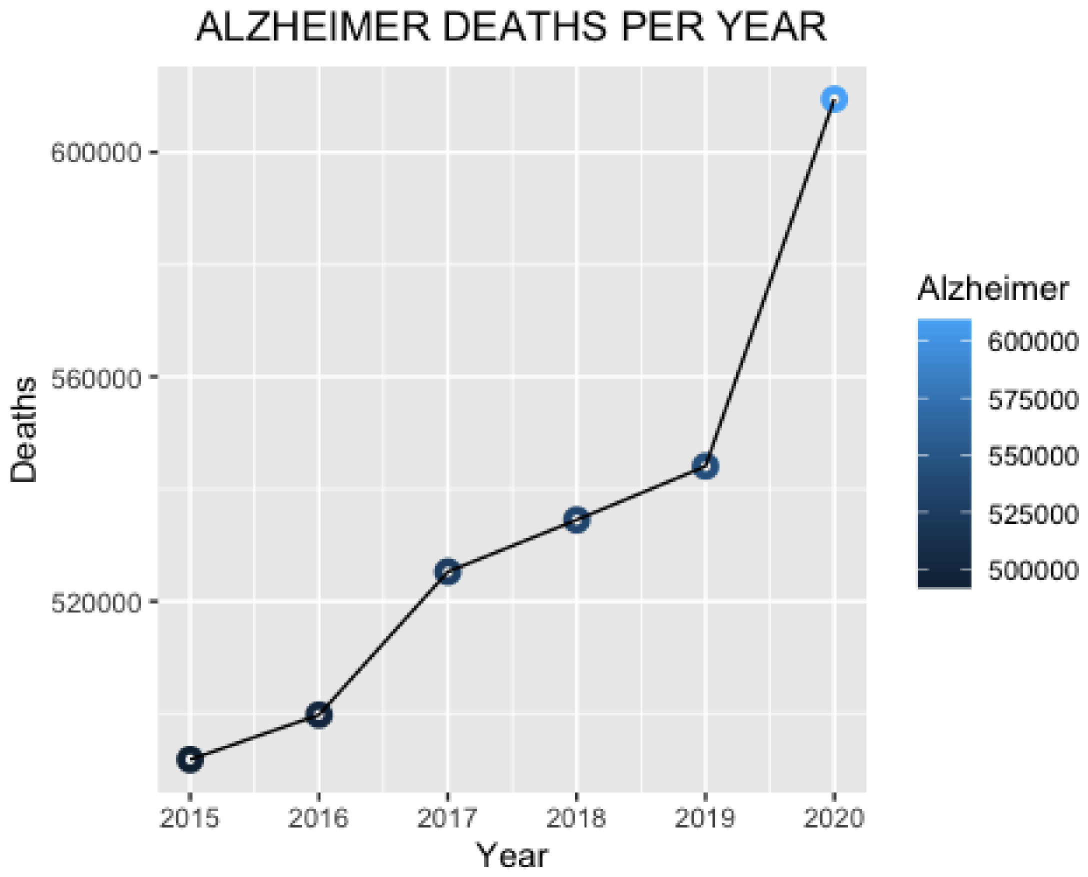


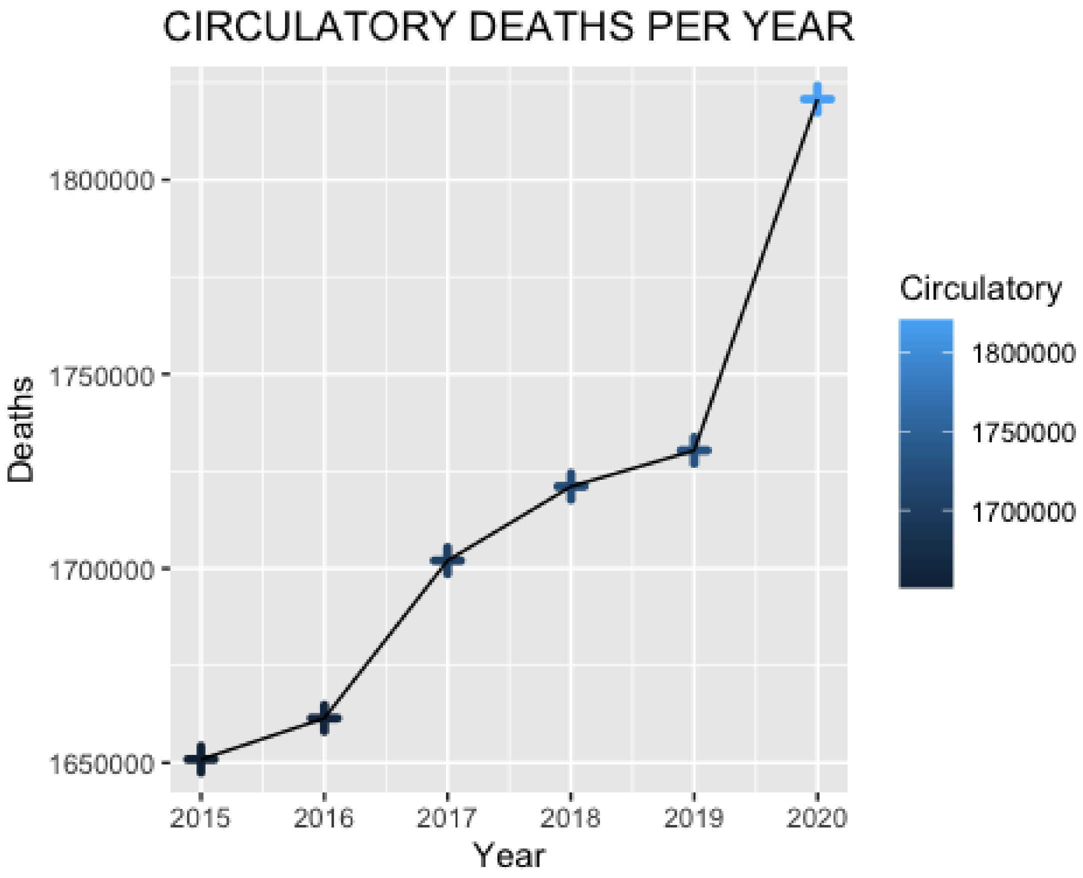
The bar plot above shows that Circulatory is the top cause of deaths, followed by Malignant. Alzheimer and Respiratory are really close but not even close to Circulatory and Malignant.

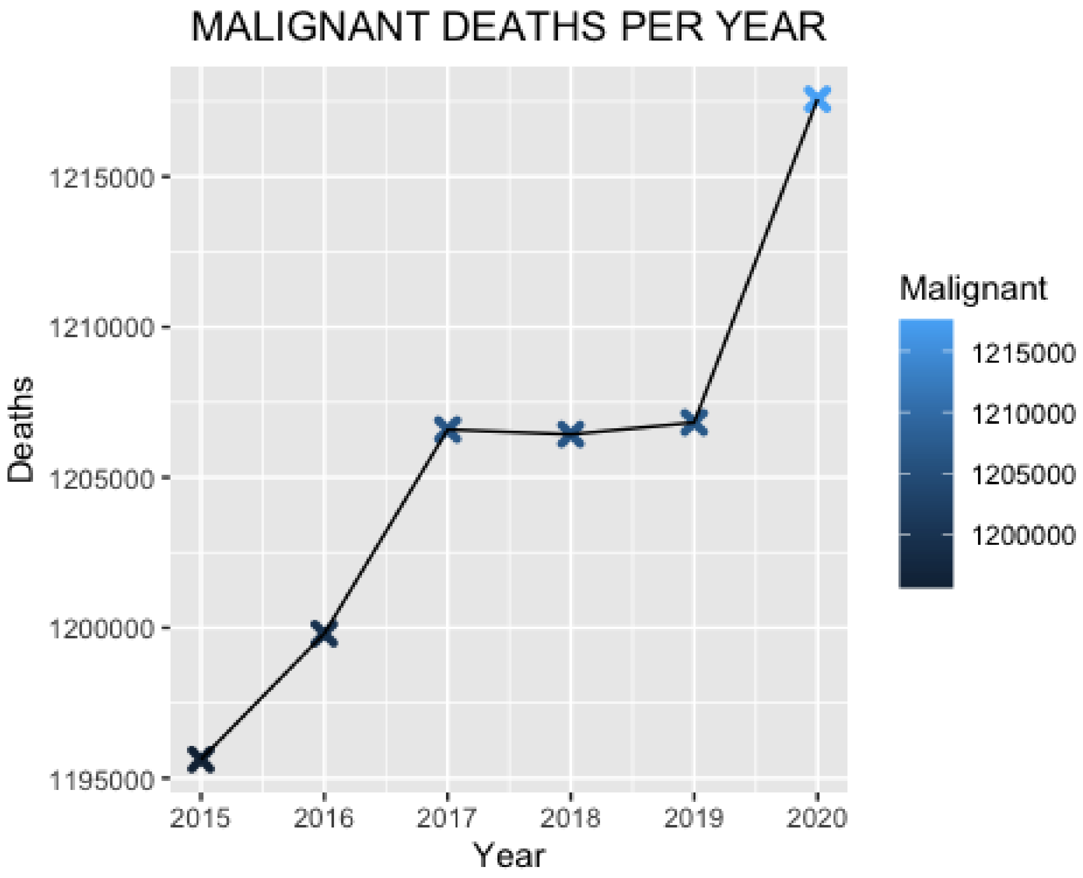
Let’s look at the deaths by cause and year.

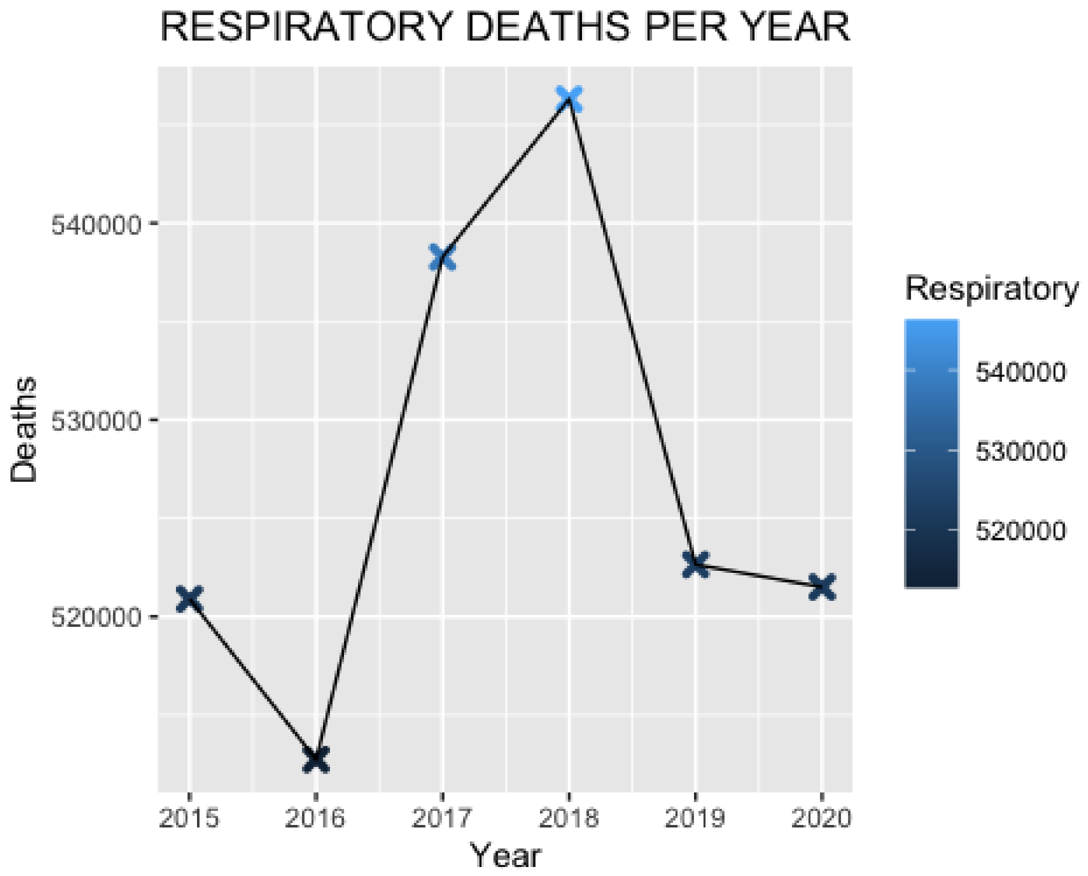
```
##   Year Alzheimer Circulatory Malignant Respiratory  Other
## 1 2015    491824     1650874   1195622      520900 314590
## 2 2016    499756     1661462   1199804      512750 315800
## 3 2017    525260     1702034   1206588      538272 326550
## 4 2018    534502     1721088   1206428      546294 330544
## 5 2019    544070     1730348   1206824      522650 330462
## 6 2020    609452     1820738   1217572      521505 359888
```

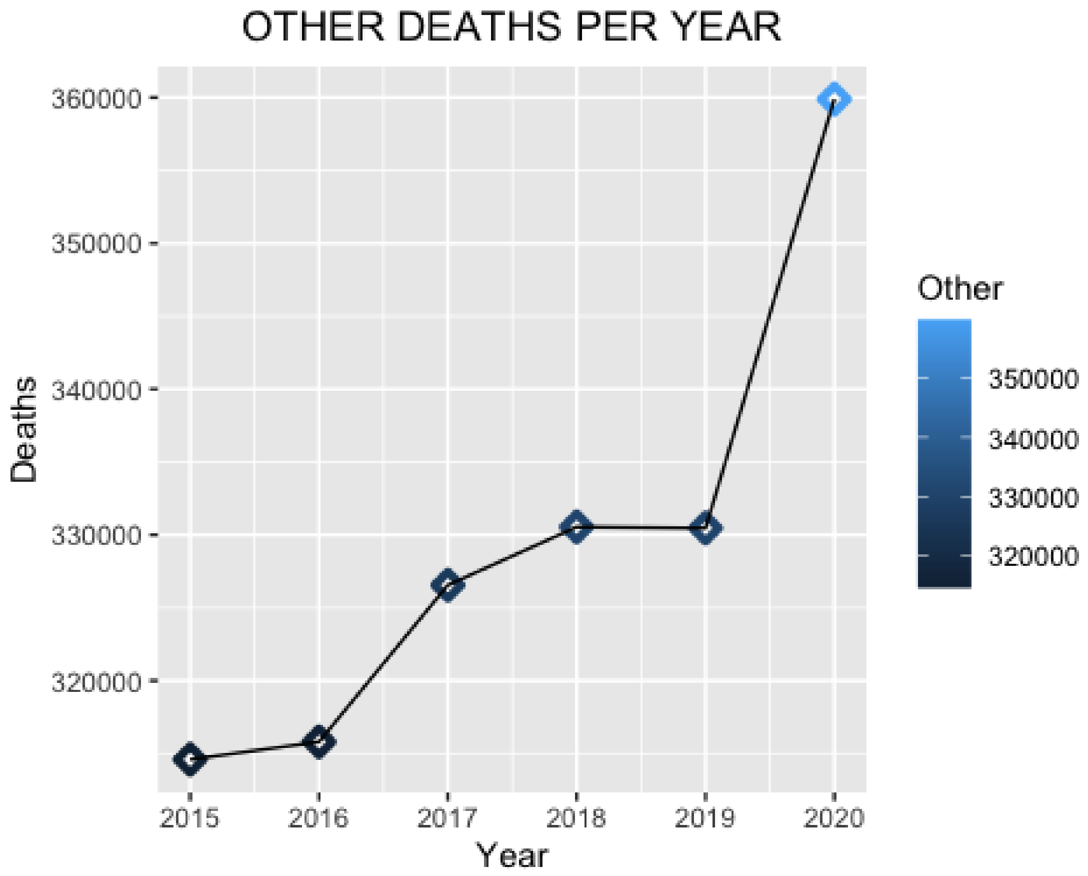
Let’s plot every cause of death by year.




CIRCULATORY DEATHS PER YEAR
Deaths
1800000
1750000
1700000
1650000
2015
2016
2017
2018
2019
2020
Year
Circulatory
1800000
1750000
1700000


MALIGNANT DEATHS PER YEAR
Deaths
1215000
1210000
1205000
1200000
1195000
2015
2016
2017
2018
2019
2020
Year
Malignant
1215000
1210000
1205000
1200000


RESPIRATORY DEATHS PER YEAR
Deaths
540000
530000
520000
2015
2016
2017
2018
2019
2020
Year
Respiratory
540000
530000
520000



Four of the five plots above all have something in common. Every year, the number of deaths for each cause increases except for one plot. The respiratory deaths plot is the only plot that decreases from the year 2018 to 2019 and then decreases a little bit more from 2019 to 2020. How is this possible? Wasn't there a global pandemic for a respiratory virus? Most of the country was on lockdown restrictions because of what was called "The Coronavirus Pandemic". Covid-19 was said to be a deadly virus and that a lot of people were dying from the virus. If this claim is true, then why does the data from this analysis show otherwise? The things that make you go hmmm…..