Hiba Talat

Data Science

Final Project Report

About the Data

The data is collected from healthdata.gov that provides provisional counts of deaths by the month the deaths occurred, by age group, sex, and race/ethnicity, for select underlying causes of death for 2020-2021.

View Data

Filter

	Date.Of.Death.Year	Date.Of.Death.Month	Sex	Race.Ethnicity	AgeGroup	AllCause	NaturalCause	Septicemia..A40.A41.	Malignant.neoplasms..C00.C97
1	2019	1	Female	Hispanic	0–4 years	182	162	NA	
2	2019	1	Female	Hispanic	5–14 years	44	28	NA	
3	2019	1	Female	Hispanic	15–24 years	122	45	0	
4	2019	1	Female	Hispanic	25–34 years	198	100	NA	
5	2019	1	Female	Hispanic	35–44 years	334	260	NA	
6	2019	1	Female	Hispanic	45–54 years	585	500	NA	
7	2019	1	Female	Hispanic	55–64 years	990	942	20	
8	2019	1	Female	Hispanic	65–74 years	1355	1311	22	
9	2019	1	Female	Hispanic	75–84 years	1951	1908	33	
10	2019	1	Female	Hispanic	85 years and over	2720	2663	28	
11	2019	1	Female	Non-Hispanic American Indian or Alaska Native	0–4 years	17	15	0	
12	2019	1	Female	Non-Hispanic American Indian or Alaska Native	5–14 years	NA	NA	0	
13	2019	1	Female	Non-Hispanic American Indian or Alaska Native	15–24 years	12	NA	0	
14	2019	1	Female	Non-Hispanic American Indian or Alaska Native	25–34 years	43	21	0	
15	2019	1	Female	Non-Hispanic American Indian or Alaska Native	35–44 years	55	38	0	
16	2019	1	Female	Non-Hispanic American Indian or Alaska Native	45–54 years	68	53	NA	
17	2019	1	Female	Non-Hispanic American Indian or Alaska Native	55–64 years	129	119	NA	
18	2019	1	Female	Non-Hispanic American Indian or Alaska Native	65–74 years	149	143	NA	
19	2019	1	Female	Non-Hispanic American Indian or Alaska Native	75–84 years	148	140	NA	
20	2019	1	Female	Non-Hispanic American Indian or Alaska Native	85 years and over	150	143	NA	
21	2019	1	Female	Non-Hispanic Asian	0–4 years	NA	NA	0	
22	2019	1	Female	Non-Hispanic Asian	5–14 years	NA	NA	0	
23	2019	1	Female	Non-Hispanic Asian	15–24 years	NA	NA	0	
24	2019	1	Female	Non-Hispanic Asian	25–34 years	13	12	0	
25	2019	1	Female	Non-Hispanic Asian	35–44 years	12	NA	0	
26	2019	1	Female	Non-Hispanic Asian	45–54 years	18	17	0	
27	2019	1	Female	Non-Hispanic Asian	55–64 years	47	43	0	

Showing 1 to 27 of 3,000 entries, 40 total columns

Data Structure

```
> str(healthdata)
'data.frame':	3000 obs. of  40 variables:
 $ Date.Of.Death.Year                                                               : int  2019 2019 2019 2019 2019 2019 2019 2019 2019 2019 ...
 $ Date.Of.Death.Month                                                              : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Sex                                                                              : chr  "Female" "Female" "Female" "Female" ...
 $ Race.Ethnicity                                                                   : chr  "Hispanic" "Hispanic" "Hispanic" "Hispanic" ...
 $ AgeGroup                                                                         : chr  "0-4 years" "5-14 years" "15-24 years" "25-34 years" ...
 $ AllCause                                                                         : int  182 44 122 198 334 585 990 1355 1951 2720 ...
 $ NaturalCause                                                                     : int  162 28 45 100 260 500 942 1311 1908 2663 ...
 $ Septicemia..A40.A41.                                                             : int  NA NA 0 NA NA NA 20 22 33 28 ...
 $ Malignant.neoplasms..C00.C97.                                                    : int  NA NA NA 29 96 209 368 382 363 275 ...
 $ Diabetes.mellitus..E10.E14.                                                      : int  0 NA NA NA NA 40 62 87 95 83 ...
 $ Alzheimer.disease..G30.                                                          : int  0 0 0 0 0 NA NA 32 126 374 ...
 $ Influenza.and.pneumonia..J09.J18.                                                : int  NA NA 0 NA 11 15 32 40 55 93 ...
 $ Chronic.lower.respiratory.diseases..J40.J47.                                     : int  0 NA NA NA NA NA 24 43 77 114 ...
 $ Other.diseases.of.respiratory.system..J00.J06.J30.J39.J67.J70.J98.               : int  NA 0 NA NA NA NA 26 38 58 38 ...
 $ Nephritis..nephrotic.syndrome.and.nephrosis..N00.N07.N17.N19.N25.N27.            : int  0 0 NA 0 NA 10 21 54 53 44 ...
 $ Symptoms..signs.and.abnormal.clinical.and.laboratory.findings..not.elsewhere.classified..R00.R99.: int  22 0 NA NA NA NA NA NA NA 19 ...
 $ Diseases.of.heart..I00.I09.I11.I13.I20.I51.                                      : int  NA 0 NA NA 25 63 146 249 417 745 ...
 $ Cerebrovascular.diseases..I60.I69.                                               : int  0 0 NA NA 10 28 35 76 146 240 ...
 $ COVID.19..U071..Multiple.Cause.of.Death.                                         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ COVID.19..U071..Underlying.Cause.of.Death.                                       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ AnalysisDate                                                                     : chr  "2/9/2021" "2/9/2021" "2/9/2021" "2/9/2021" ...
 $ Note                                                                             : chr  "" "" "" "" ...
 $ flag_allcause                                                                    : chr  "" "" "" "" ...
 $ flag_natcause                                                                    : chr  "" "" "" "" ...
 $ flag_sept                                                                        : chr  "One or more data cells have counts between 1-9 and have been supp
ressed in accordance with NCHS confidentiality standards." "One or more data cells have counts between 1-9 and have been suppressed in accordance with NCHS confidentiality s
tandards." "" "One or more data cells have counts between 1-9 and have been suppressed in accordance with NCHS confidentiality standards." ...
 $ flag_neopl                                                                       : chr  "One or more data cells have counts between 1-9 and have been supp
ressed in accordance with NCHS confidentiality standards." "One or more data cells have counts between 1-9 and have been suppressed in accordance with NCHS confidentiality s
tandards." "One or more data cells have counts between 1-9 and have been suppressed in accordance with NCHS confidentiality standards." "" ...
 $ flag_diab                                                                        : chr  "" "One or more data cells have counts between 1-9 and have been s
uppressed in accordance with NCHS confidentiality standards." "One or more data cells have counts between 1-9 and have been suppressed in accordance with NCHS confidentialit
y standards." "One or more data cells have counts between 1-9 and have been suppressed in accordance with NCHS confidentiality standards." ...
 $ flag_alz                                                                         : chr  "" "" "" "" ...
 $ flag_inflpn                                                                      : chr  "One or more data cells have counts between 1-9 and have been supp
ressed in accordance with NCHS confidentiality standards." "One or more data cells have counts between 1-9 and have been suppressed in accordance with NCHS confidentiality s
tandards." "" "One or more data cells have counts between 1-9 and have been suppressed in accordance with NCHS confidentiality standards." ...
```

Install the packages

I am using three packages :

- Dplyr: The package contains a set of functions that perform common data manipulation operations
- Ggplot2: ggplot2 is a data visualization package for the statistical programming language R.
- Naniar: It provides data structures and functions that facilitate the plotting of missing values and examination of imputations.

```
> install.packages(c("dplyr","ggplot2","naniar"))
trying URL 'https://cran.rstudio.com/bin/macosx/contrib/4.0/dplyr_1.0.5.tgz'
Content type 'application/x-gzip' length 1251016 bytes (1.2 MB)
==================================================
downloaded 1.2 MB

trying URL 'https://cran.rstudio.com/bin/macosx/contrib/4.0/ggplot2_3.3.3.tgz'
Content type 'application/x-gzip' length 4068756 bytes (3.9 MB)
==================================================
downloaded 3.9 MB

trying URL 'https://cran.rstudio.com/bin/macosx/contrib/4.0/naniar_0.6.0.tgz'
Content type 'application/x-gzip' length 2688536 bytes (2.6 MB)
==================================================
downloaded 2.6 MB


The downloaded binary packages are in
        /var/folders/7j/97t94x2x6nv9pk275mmh03800000gn/T//RtmpOE5zj8/downloaded_packages
```

```
> library(dplyr)

Attaching package: ‘dplyr’

The following objects are masked from ‘package:stats’:

    filter, lag

The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union

> library(naniar)
> library(visdat)
> library(ggplot2)
Learn more about the underlying theory at https://ggplot2-book.org/
```

Create a new data frame to store the useful information

```
> df <- healthdata[,c(1:15,17,18,19,20,38,39)]
```

Structure of df

```
> str(df)
'data.frame':   3000 obs. of  21 variables:
 $ Date.Of.Death.Year                                          : int  2019 2019 2019 2019 2019 2019 2019 2019 2019 2019 ...
 $ Date.Of.Death.Month                                         : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Sex                                                         : chr  "Female" "Female" "Female" "Female" ...
 $ Race.Ethnicity                                              : chr  "Hispanic" "Hispanic" "Hispanic" "Hispanic" ...
 $ AgeGroup                                                    : chr  "0-4 years" "5-14 years" "15-24 years" "25-34 years" ...
 $ AllCause                                                    : int  182 44 122 198 334 585 990 1355 1951 2720 ...
 $ NaturalCause                                                : int  162 28 45 100 260 500 942 1311 1908 2663 ...
 $ Septicemia..A40.A41.                                        : int  NA NA 0 NA NA NA 20 22 33 28 ...
 $ Malignant.neoplasms..C00.C97.                               : int  NA NA NA 29 96 209 368 382 363 275 ...
 $ Diabetes.mellitus..E10.E14.                                 : int  0 NA NA NA NA 40 62 87 95 83 ...
 $ Alzheimer.disease..G30.                                     : int  0 0 0 0 0 0 NA NA 32 126 374 ...
 $ Influenza.and.pneumonia..J09.J18.                           : int  NA NA 0 NA 11 15 32 40 55 93 ...
 $ Chronic.lower.respiratory.diseases..J40.J47.                : int  0 NA NA NA NA NA 24 43 77 114 ...
 $ Other.diseases.of.respiratory.system..J00.J06.J30.J39.J67.J70.J98.   : int  NA 0 NA NA NA NA 26 38 58 38 ...
 $ Nephritis..nephrotic.syndrome.and.nephrosis..N00.N07.N17.N19.N25.N27.: int  0 0 NA 0 NA 10 21 54 53 44 ...
 $ Diseases.of.heart..I00.I09.I11.I13.I20.I51.                 : int  NA 0 NA NA 25 63 146 249 417 745 ...
 $ Cerebrovascular.diseases..I60.I69.                          : int  0 0 NA NA 10 28 35 76 146 240 ...
 $ COVID.19..U071..Multiple.Cause.of.Death.                    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ COVID.19..U071..Underlying.Cause.of.Death.                  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Start.Date                                                  : chr  "01/01/2019" "01/01/2019" "01/01/2019" "01/01/2019" ...
 $ End.Date                                                    : chr  "01/31/2019" "01/31/2019" "01/31/2019" "01/31/2019" ...
```

Head & Tail rows of df

```
> head(df)
  Date.Of.Death.Year Date.Of.Death.Month    Sex Race.Ethnicity    AgeGroup AllCause NaturalCause Septicemia..A40.A41. Malignant.neoplasms..C00.C97.
1               2019                   1 Female       Hispanic   0-4 years      182          162                  NA                           NA
2               2019                   1 Female       Hispanic  5-14 years       44           28                  NA                           NA
3               2019                   1 Female       Hispanic 15-24 years      122           45                   0                           NA
4               2019                   1 Female       Hispanic 25-34 years      198          100                  NA                           29
5               2019                   1 Female       Hispanic 35-44 years      334          260                  NA                           96
6               2019                   1 Female       Hispanic 45-54 years      585          500                  NA                          209
  Diabetes.mellitus..E10.E14. Alzheimer.disease..G30. Influenza.and.pneumonia..J09.J18. Chronic.lower.respiratory.diseases..J40.J47.
1                           0                       0                              NA                                          0
2                          NA                       0                              NA                                         NA
3                          NA                       0                               0                                         NA
4                          NA                       0                              NA                                         NA
5                          NA                       0                              11                                         NA
6                          40                      NA                              15                                         NA
  Other.diseases.of.respiratory.system..J00.J06.J30.J39.J67.J70.J98. Nephritis..nephrotic.syndrome.and.nephrosis..N00.N07.N17.N19.N25.N27.
1                                                                NA                                                                    0
2                                                                 0                                                                    0
3                                                                NA                                                                   NA
4                                                                NA                                                                    0
5                                                                NA                                                                   NA
6                                                                NA                                                                   10
  Diseases.of.heart..I00.I09.I11.I13.I20.I51. Cerebrovascular.diseases..I60.I69. COVID.19..U071..Multiple.Cause.of.Death.
1                                          NA                                  0                                        0
2                                           0                                  0                                        0
3                                          NA                                 NA                                        0
4                                          NA                                 NA                                        0
5                                          25                                 10                                        0
6                                          63                                 28                                        0
  COVID.19..U071..Underlying.Cause.of.Death. Start.Date   End.Date
1                                          0 01/01/2019 01/31/2019
2                                          0 01/01/2019 01/31/2019
```

```
> tail(df)
     Date.Of.Death.Year Date.Of.Death.Month  Sex Race.Ethnicity          AgeGroup AllCause NaturalCause Septicemia..A40.A41. Malignant.neoplasms..C00.C97.
2995               2021                  1 Male          Other       35-44 years       70          57                  0                            NA
2996               2021                  1 Male          Other       45-54 years      112         107                  0                            NA
2997               2021                  1 Male          Other       55-64 years      238         235                 NA                            43
2998               2021                  1 Male          Other       65-74 years      271         262                 NA                            44
2999               2021                  1 Male          Other       75-84 years      221         220                 NA                            31
3000               2021                  1 Male          Other 85 years and over      120         120                 NA                            NA
     Diabetes.mellitus..E10.E14. Alzheimer.disease..G30. Influenza.and.pneumonia..J09.J18. Chronic.lower.respiratory.diseases..J40.J47.
2995                          NA                       0                               0                                        NA
2996                          NA                       0                              NA                                        NA
2997                          10                       0                              NA                                        NA
2998                          NA                      NA                              NA                                        NA
2999                          NA                      NA                              NA                                        NA
3000                          NA                      NA                              NA                                        NA
     Other.diseases.of.respiratory.system..J00.J06.J30.J39.J67.J70.J98. Nephritis..nephrotic.syndrome.and.nephrosis..N00.N07.N17.N19.N25.N27.
2995                                                                  0                                                                  0
2996                                                                 NA                                                                 NA
2997                                                                 NA                                                                  0
2998                                                                  0                                                                 NA
2999                                                                 NA                                                                 13
3000                                                                 NA                                                                 NA
     Diseases.of.heart..I00.I09.I11.I13.I20.I51. Cerebrovascular.diseases..I60.I69. COVID.19..U071..Multiple.Cause.of.Death.
2995                                          NA                                NA                                       12
2996                                          15                                 0                                       20
2997                                          40                                NA                                       54
2998                                          48                                NA                                       85
2999                                          39                                NA                                       70
3000                                          28                                 0                                       47
     COVID.19..U071..Underlying.Cause.of.Death. Start.Date   End.Date
2995                                         11 01/01/2021 01/31/2021
2996                                         19 01/01/2021 01/31/2021
2997                                         51 01/01/2021 01/31/2021
2998                                         83 01/01/2021 01/31/2021
2999                                         66 01/01/2021 01/31/2021
3000                                         45 01/01/2021 01/31/2021
```

Columns in df

```
> colnames(df)
 [1] "DateOfDeathYear"                                  "DateOfDeathMonth"
 [3] "Sex"                                              "RaceEthnicity"
 [5] "AgeGroup"                                         "AllCause"
 [7] "NaturalCause"                                     "SepticemiaA40A41"
 [9] "MalignantneoplasmsC00C97"                         "DiabetesmellitusE10E14"
[11] "AlzheimerdiseaseG30"                              "InfluenzaandpneumoniaJ09J18"
[13] "ChroniclowerrespiratorydiseasesJ40J47"            "OtherdiseasesofrespiratorysystemJ00J06J30J39J67J70J98"
[15] "NephritisnephroticsyndromeandnephrosisN00N07N17N19N25N27" "DiseasesofheartI00I09I11I13I20I51"
[17] "CerebrovasculardiseasesI60I69"                    "COVID19U071MultipleCauseofDeath"
[19] "COVID19U071UnderlyingCauseofDeath"                "StartDate"
[21] "EndDate"
```

We can see that the column names are not very informative.

Clean the column names

```
> colnames(df) <- gsub("\\.","",colnames(df))
> #Changing the names of the cloumns
> names(df)[names(df) == "AllCause"] <- "AllCauses"
> names(df)[names(df) == "SepticemiaA40A41"] <- "Septicemia"
> names(df)[names(df) == "MalignantneoplasmsC00C97"] <- "MalignantNeoplasms"
> names(df)[names(df) == "DiabetesmellitusE10E14"] <- "DiabetesMellitus"
> names(df)[names(df) == "AlzheimerdiseaseG30"   ] <- "AlzheimerDisease"
> names(df)[names(df) == "InfluenzaandpneumoniaJ09J18"] <- "Influenza&Pneumonia"
> names(df)[names(df) == "ChroniclowerrespiratorydiseasesJ40J47" ] <- "ChronicLowerRespiratoryDiseases"
> names(df)[names(df) == "OtherdiseasesofrespiratorysystemJ00J06J30J39J67J70J98" ] <- "OtherDiseasesofRespiratorySystem"
> names(df)[names(df) =="NephritisnephroticsyndromeandnephrosisN00N07N17N19N25N27"] <- "NephritisNephroticSyndromeAndNephrosis"
> names(df)[names(df) =="DiseasesofheartI00I09I11I13I20I51"] <- "HeartDiseases"
> names(df)[names(df) =="CerebrovasculardiseasesI60I69" ] <- "CerebrovascularDiseases"
> names(df)[names(df) =="COVID19U071UnderlyingCauseofDeath"] <- "COVID19UnderlyingCauseofDeath"
> names(df)[names(df) =="COVID19U071MultipleCauseofDeath"] <- "COVID19MultipleCausesofDeath"
> colnames(df)
 [1] "DateOfDeathYear"                "DateOfDeathMonth"                "Sex"
 [4] "RaceEthnicity"                  "AgeGroup"                        "AllCauses"
 [7] "NaturalCause"                   "Septicemia"                      "MalignantNeoplasms"
[10] "DiabetesMellitus"               "AlzheimerDisease"                "Influenza&Pneumonia"
[13] "ChronicLowerRespiratoryDiseases" "OtherDiseasesofRespiratorySystem" "NephritisNephroticSyndromeAndNephrosis"
[16] "HeartDiseases"                  "CerebrovascularDiseases"         "COVID19MultipleCausesofDeath"
[19] "COVID19UnderlyingCauseofDeath"  "StartDate"                       "EndDate"
```

View the df

Filter

	DateOfDeathYear	DateOfDeathMonth	Sex	RaceEthnicity	AgeGroup	AllCauses	NaturalCause	Septicemia	MalignantNeoplasms	DiabetesMellitus
1	2019	1	Female	Hispanic	0–4 years	182	162	NA	NA	0
2	2019	1	Female	Hispanic	5–14 years	44	28	NA	NA	NA
3	2019	1	Female	Hispanic	15–24 years	122	45	0	NA	NA
4	2019	1	Female	Hispanic	25–34 years	198	100	NA	29	NA
5	2019	1	Female	Hispanic	35–44 years	334	260	NA	96	NA
6	2019	1	Female	Hispanic	45–54 years	585	500	NA	209	40
7	2019	1	Female	Hispanic	55–64 years	990	942	20	368	62
8	2019	1	Female	Hispanic	65–74 years	1355	1311	22	382	87
9	2019	1	Female	Hispanic	75–84 years	1951	1908	33	363	95
10	2019	1	Female	Hispanic	85 years and over	2720	2663	28	275	83
11	2019	1	Female	Non-Hispanic American Indian or Alaska Native	0–4 years	17	15	0	0	0
12	2019	1	Female	Non-Hispanic American Indian or Alaska Native	5–14 years	NA	NA	0	0	0
13	2019	1	Female	Non-Hispanic American Indian or Alaska Native	15–24 years	12	NA	0	0	0
14	2019	1	Female	Non-Hispanic American Indian or Alaska Native	25–34 years	43	21	0	0	NA
15	2019	1	Female	Non-Hispanic American Indian or Alaska Native	35–44 years	55	38	0	NA	NA
16	2019	1	Female	Non-Hispanic American Indian or Alaska Native	45–54 years	68	53	NA	NA	NA
17	2019	1	Female	Non-Hispanic American Indian or Alaska Native	55–64 years	129	119	NA	29	10
18	2019	1	Female	Non-Hispanic American Indian or Alaska Native	65–74 years	149	143	NA	37	13
19	2019	1	Female	Non-Hispanic American Indian or Alaska Native	75–84 years	148	140	NA	28	NA
20	2019	1	Female	Non-Hispanic American Indian or Alaska Native	85 years and over	150	143	NA	NA	NA
21	2019	1	Female	Non-Hispanic Asian	0–4 years	NA	NA	0	0	0
22	2019	1	Female	Non-Hispanic Asian	5–14 years	NA	NA	0	NA	0
23	2019	1	Female	Non-Hispanic Asian	15–24 years	NA	NA	0	0	0
24	2019	1	Female	Non-Hispanic Asian	25–34 years	13	12	0	NA	0
25	2019	1	Female	Non-Hispanic Asian	35–44 years	12	NA	0	NA	0
26	2019	1	Female	Non-Hispanic Asian	45–54 years	18	17	0	NA	0
27	2019	1	Female	Non-Hispanic Asian	55–64 years	47	43	0	16	NA
28	2019	1	Female	Non-Hispanic Asian	65–74 years	67	65	NA	22	NA
29	2019	1	Female	Non-Hispanic Asian	75–84 years	87	84	NA	14	NA
30	2019	1	Female	Non-Hispanic Asian	85 years and over	98	96	NA	12	NA
31	2019	1	Female	Non-Hispanic Black	0–4 years	261	232	NA	NA	0

Showing 1 to 30 of 3,000 entries, 21 total columns

DiabetesMellitus	AlzheimerDisease	Influenza&Pneumonia	ChronicLowerRespiratoryDiseases	OtherDiseasesofRespiratorySystem	NephritisNephroticSyndromeAndNephrosis	HeartDiseases	C
0	0	NA	0	NA	0	NA	
NA	0	NA	NA	0	0	0	
NA	0	0	NA	NA	NA	NA	
NA	0	NA	NA	NA	0	NA	
NA	0	11	NA	NA	NA	25	
40	NA	15	NA	NA	10	63	
62	NA	32	24	26	21	146	
87	32	40	43	38	54	249	
95	126	55	77	58	53	417	
83	374	93	114	38	44	745	
0	0	0	0	0	0	0	
0	0	0	0	0	0	0	
0	0	NA	0	0	0	0	
NA	0	NA	0	0	NA	NA	
NA	0	NA	0	NA	NA	NA	
NA	0	NA	NA	0	0	NA	
10	0	NA	NA	NA	NA	23	
13	NA	NA	NA	NA	NA	30	
NA	NA	NA	NA	NA	NA	34	
NA	19	NA	NA	NA	NA	45	
0	0	NA	0	0	0	0	
0	0	0	0	0	0	0	
0	0	NA	0	0	0	0	
0	0	0	0	NA	0	0	
0	0	0	0	0	0	NA	
0	0	0	0	NA	0	NA	
NA	NA	NA	0	0	NA	NA	
NA	0	NA	NA	NA	NA	12	
NA	NA	NA	NA	NA	NA	26	
NA	NA	NA	NA	NA	NA	26	
0	0	NA	NA	NA	0	NA	

howing 1 to 30 of 3,000 entries, 21 total columns

NephritisNephroticSyndromeAndNephrosis	HeartDiseases	CerebrovascularDiseases	COVID19MultipleCausesofDeath	COVID19UnderlyingCauseofDeath	StartDate	EndDate
0	NA	0	0	0	01/01/2019	01/31/2019
0	0	0	0	0	01/01/2019	01/31/2019
NA	NA	NA	0	0	01/01/2019	01/31/2019
0	NA	NA	0	0	01/01/2019	01/31/2019
NA	25	10	0	0	01/01/2019	01/31/2019
10	63	28	0	0	01/01/2019	01/31/2019
21	146	35	0	0	01/01/2019	01/31/2019
54	249	76	0	0	01/01/2019	01/31/2019
53	417	146	0	0	01/01/2019	01/31/2019
44	745	240	0	0	01/01/2019	01/31/2019
0	0	0	0	0	01/01/2019	01/31/2019
0	0	0	0	0	01/01/2019	01/31/2019
0	0	0	0	0	01/01/2019	01/31/2019
NA	NA	0	0	0	01/01/2019	01/31/2019
NA	NA	NA	0	0	01/01/2019	01/31/2019
0	NA	NA	0	0	01/01/2019	01/31/2019
NA	23	NA	0	0	01/01/2019	01/31/2019
NA	30	NA	0	0	01/01/2019	01/31/2019
NA	34	12	0	0	01/01/2019	01/31/2019
NA	45	NA	0	0	01/01/2019	01/31/2019
0	0	0	0	0	01/01/2019	01/31/2019
0	0	0	0	0	01/01/2019	01/31/2019
0	0	0	0	0	01/01/2019	01/31/2019
0	0	0	0	0	01/01/2019	01/31/2019
0	NA	NA	0	0	01/01/2019	01/31/2019
0	NA	0	0	0	01/01/2019	01/31/2019
NA	NA	NA	0	0	01/01/2019	01/31/2019
NA	12	NA	0	0	01/01/2019	01/31/2019
NA	26	NA	0	0	01/01/2019	01/31/2019
NA	26	11	0	0	01/01/2019	01/31/2019
0	NA	NA	0	0	01/01/2019	01/31/2019

3,000 entries, 21 total columns

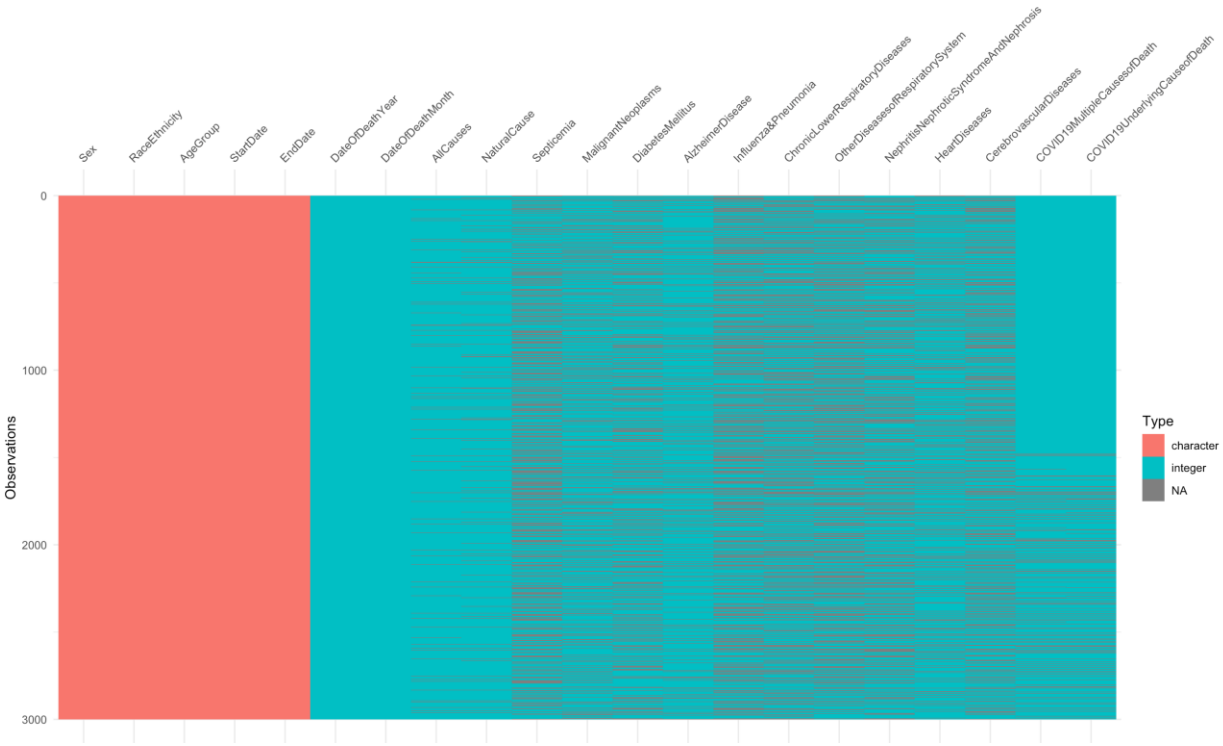
Na Values

NA values are the missing values that we can see in the above figures. Let's visually see them again below:

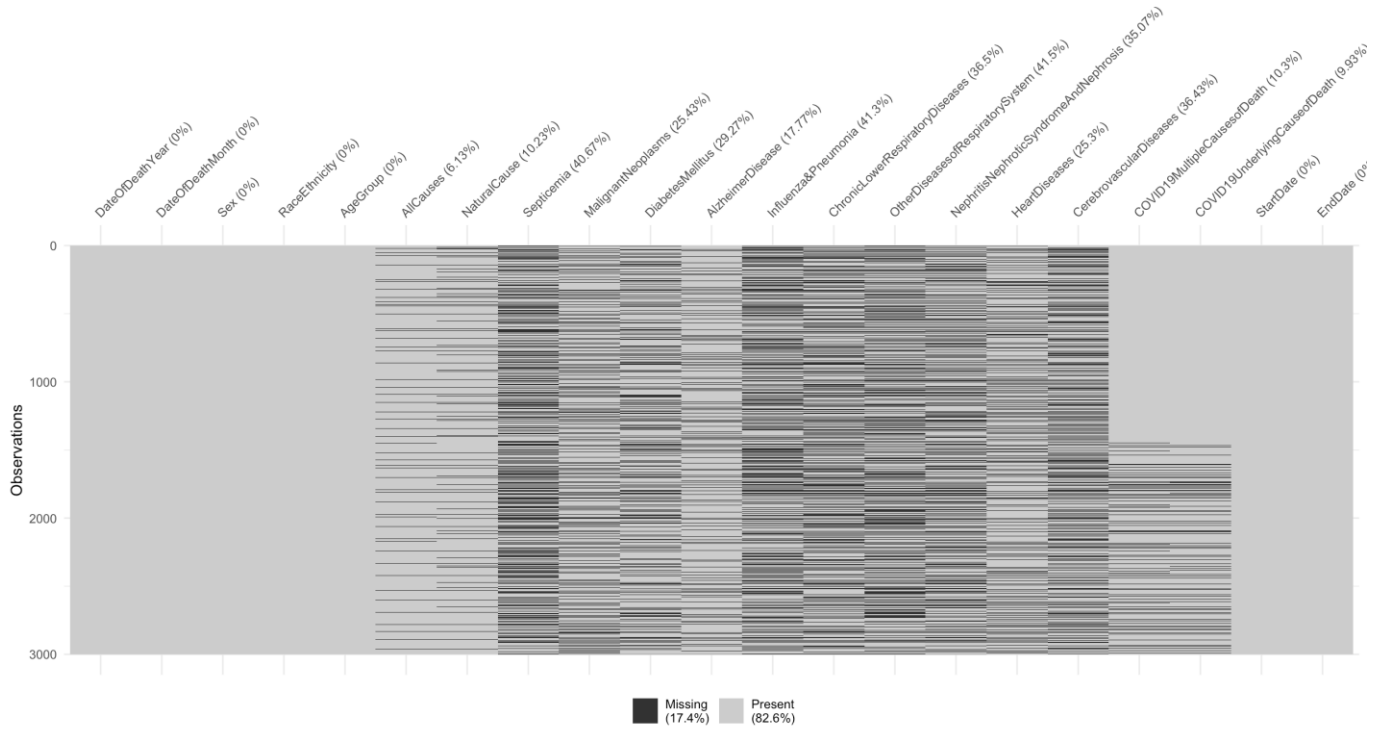
```
> sum(is.na(df))
```

10975

> vis_dat(df)

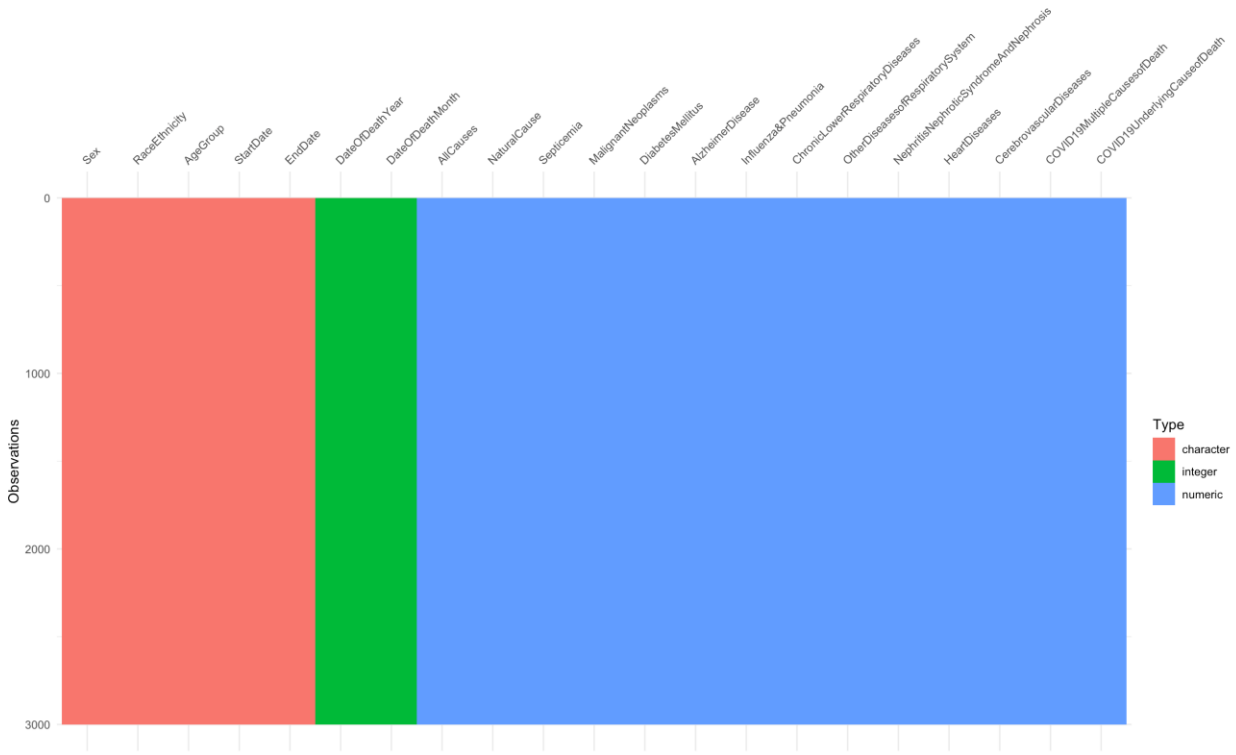


> vis_miss(df)

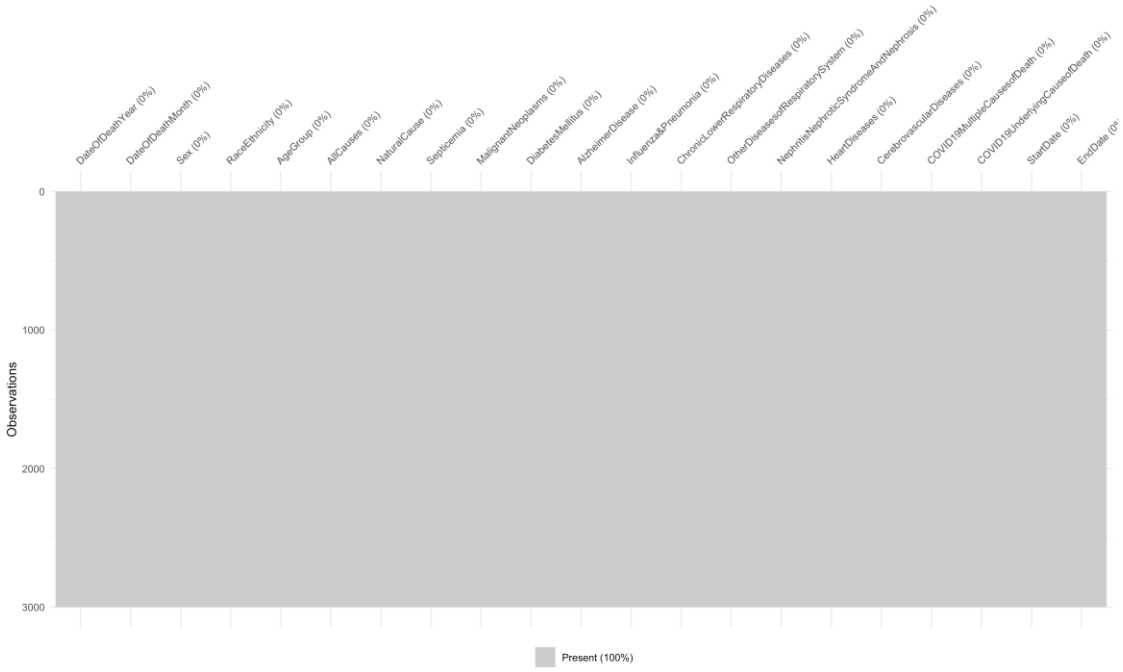


Replace NA with 0

```
> df[is.na(df)] = 0

> sum(is.na(df))

0

>vis_dat(df)
```



vis_miss(df)


DateOfDeathYear (0%)
DateOfDeathMonth (0%)
Sex (0%)
RaceEthnicity (0%)
AgeGroup (0%)
AllCauses (0%)
NaturalCause (0%)
Septicemia (0%)
MalignantNeoplasms (0%)
DiabetesMellitus (0%)
AlzheimerDisease (0%)
Influenza&Pneumonia (0%)
ChronicLowerRespiratoryDiseases (0%)
OtherDiseasesofRespiratorySystem (0%)
NephritisNephroticSyndromeAndNephrosis (0%)
HeartDiseases (0%)
CerebrovascularDiseases (0%)
COVID19MultipleCausesofDeath (0%)
COVID19UnderlyingCausesofDeath (0%)
StartDate (0%)
0
1000
2000
3000
Observations
Present (100%)

Count the number variables in Sex, Race Ethnicity and Age Group to see how the values are distributed

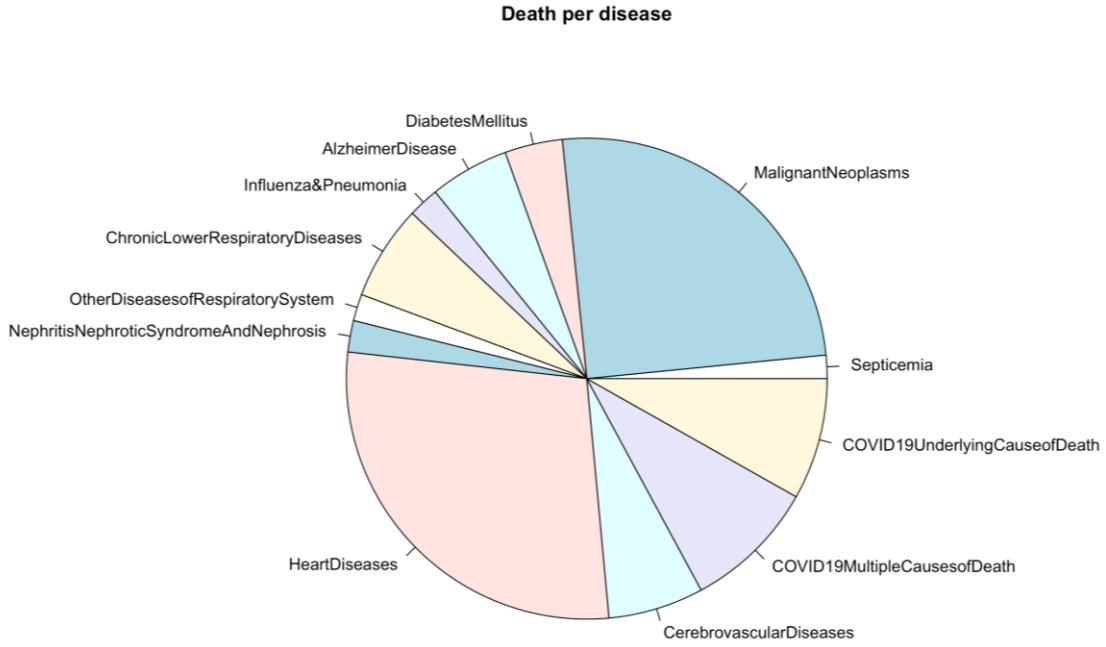
```
> count(df, df$Sex)
  df$Sex    n
1 Female 1500
2   Male 1500
> count(df,df$AgeGroup)
         df$AgeGroup   n
1          0-4 years 300
2        15-24 years 300
3        25-34 years 300
4        35-44 years 300
5        45-54 years 300
6         5-14 years 300
7        55-64 years 300
8        65-74 years 300
9        75-84 years 300
10 85 years and over 300
> count(df,df$RaceEthnicity)
                                df$RaceEthnicity   n
1                                       Hispanic 500
2 Non-Hispanic American Indian or Alaska Native 500
3                             Non-Hispanic Asian 500
4                             Non-Hispanic Black 500
5                             Non-Hispanic White 500
6                                         Other 500
>
```

Observation

The values are equally distributed among variables in all three of columns.

Pie Chart to see which disease caused more deaths

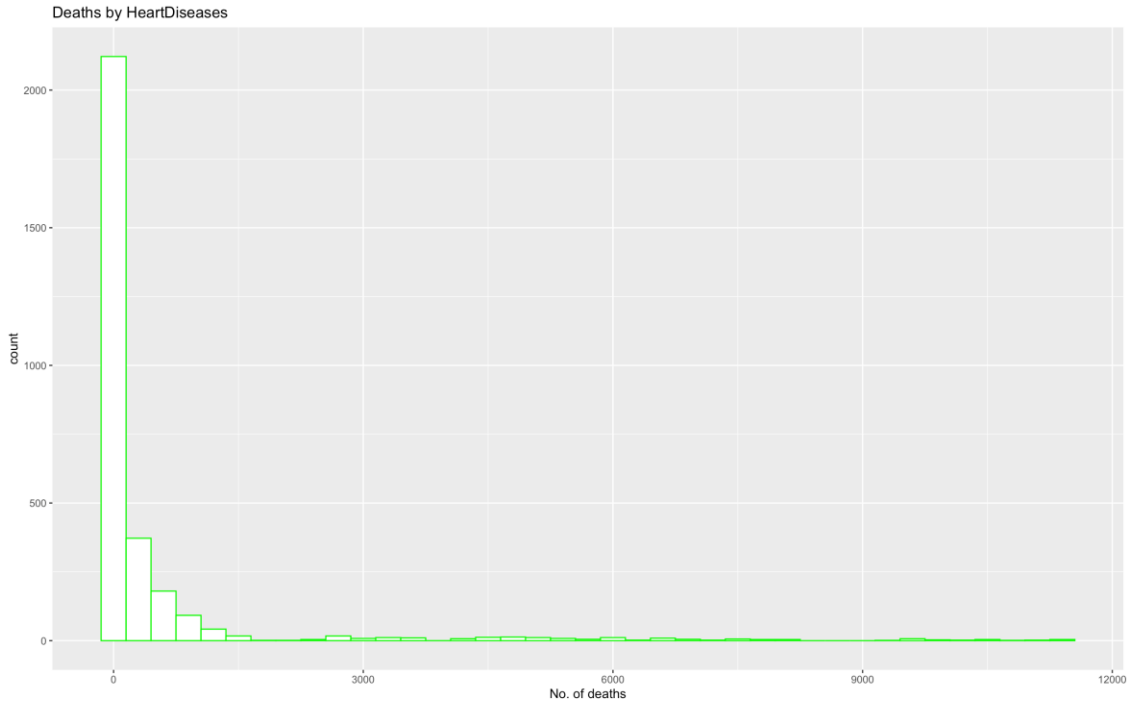
```
> #Visuals
> #install.packages("ggplot2")
> #piechart
> #Find some of the diseases causes death
> sept<- sum(df$Septicemia)
> mal <-sum(df$MalignantNeoplasms)
> dia<-sum(df$DiabetesMellitus)
> alz <- sum(df$AlzheimerDisease)
> inf <- sum(df$`Influenza&Pneumonia`)
> chr <- sum(df$ChronicLowerRespiratoryDiseases)
> oth <- sum(df$OtherDiseasesofRespiratorySystem)
> nep <- sum(df$NephritisNephroticSyndromeAndNephrosis)
> heart <- sum(df$HeartDiseases)
> cere <- sum(df$CerebrovascularDiseases)
> comul <- sum(df$COVID19MultipleCausesofDeath)
> cov <- sum(df$COVID19UnderlyingCauseofDeath)
> pc <- c(sept,mal,dia,alz,inf,chr,oth,nep,heart,cere,comul,cov)
> lab <- c("Septicemia","MalignantNeoplasms","DiabetesMellitus","AlzheimerDisease","Influenza&Pneumoni
a","ChronicLowerRespiratoryDiseases","OtherDiseasesofRespiratorySystem","NephritisNephroticSyndromeAnd
Nephrosis","HeartDiseases"
+         ,"CerebrovascularDiseases","COVID19MultipleCausesofDeath","COVID19UnderlyingCauseofDeath")
> pc
 [1]   76705 1224757  189914  261075  101713  311887   87690  103475 1380463  313590  436611
[12]  399432
> pie(pc,labels = lab, main ="Death per disease")
>
```

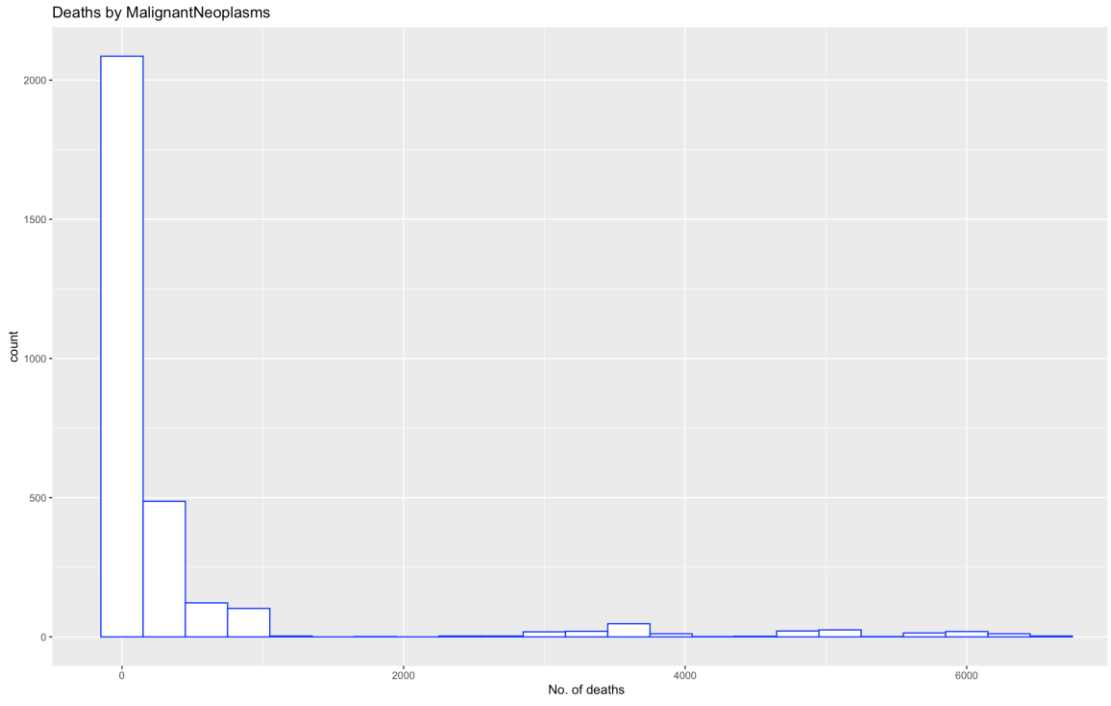


Observation

We can see Malignant Neoplasms and Heart Diseases caused more deaths.

```
> hd_hist<-ggplot(data= df, aes(x=HeartDiseases)) + geom_histogram(binwidth = 300,color="green", fill
="white")+labs(title="Deaths by HeartDiseases", x="No. of deaths")
> hd_hist
> mal_hist<-ggplot(df, aes(x=MalignantNeoplasms)) + geom_histogram(binwidth=300,color="blue", fill="wh
ite")+labs(title="Deaths by MalingnantNeoplasms", x="No. of deaths")
> mal_hist
>
```


Deaths by HeartDiseases
count
2000
1500
1000
500
0
0
3000
6000
9000
12000
No. of deaths


Deaths by MalignantNeoplasms
count
2000
1500
1000
500
0
0
2000
4000
6000
No. of deaths

Observation

No. Of deaths by heart disease is more than 11000

No. Of deaths by Malignant Neoplasms is mor than 6500

Finding highest death in Sex, Race/Ethnicity and Age Group due to Heart Disease

```
> #Finding Sex,Age group and Ethnicity at highest HeartDisease
> hd_sex<-tapply(df$HeartDiseases,df$Sex, max)
> hd_race<-tapply(df$HeartDiseases, df$RaceEthnicity, max)
> hd_age<-tapply(df$HeartDiseases,df$AgeGroup, max)
> View(sort(hd_sex,decreasing = TRUE))
> View(sort(hd_race,decreasing = TRUE))
> View(sort(hd_age, decreasing = TRUE))
```

sort(hd_sex, decreasin…	double [2]	11502 8236
Female	double [1]	11502
Male	double [1]	8236

Name	Type	Value
sort(hd_race, decreasi…	double [6]	11502 1477 978 462 407 52
Non-Hispanic White	double [1]	11502
Non-Hispanic Black	double [1]	1477
Hispanic	double [1]	978
Non-Hispanic Asian	double [1]	462
Other	double [1]	407
Non-Hispanic Ameri…	double [1]	52

Name	Type	Value
sort(hd_age, decreasin…	double [10]	11502 6703 5538 3640 1277 420 …
85 years and over	double [1]	11502
75-84 years	double [1]	6703
65-74 years	double [1]	5538
55-64 years	double [1]	3640
45-54 years	double [1]	1277
35-44 years	double [1]	420
25-34 years	double [1]	128
15-24 years	double [1]	34
0-4 years	double [1]	14
5-14 years	double [1]	0

Observation:

Maximum deaths due to heart disease is ‘11502’

Female has more deaths due to heart disease

Non-Hispanic white has more deaths due to heart disease

85 years and over age group has more deaths due to heart disease

Finding highest death in Sex, Race/Ethnicity and Age Group due to Malignant Neoplasms

```
> ##Finding Sex,Age group and Ethnicity at highest MalignantNeoplasms
> mn_sex<-tapply(df$MalignantNeoplasms,df$Sex, max)
> mn_race<-tapply(df$MalignantNeoplasms, df$RaceEthnicity, max)
> mn_age<-tapply(df$MalignantNeoplasms,df$AgeGroup, max)
> View(sort(mn_sex,decreasing = TRUE))
> View(sort(mn_race,decreasing = TRUE))
> View(sort(mn_age, decreasing = TRUE))
```

Name	Type	Value
sort(mn_sex, decreasi...	double [2]	6498 5217
Male	double [1]	6498
Female	double [1]	5217

Name	Type	Value
sort(mn_race, decreasi...	double [6]	6498 1074 593 252 232 56
Non-Hispanic White	double [1]	6498
Non-Hispanic Black	double [1]	1074
Hispanic	double [1]	593
Other	double [1]	252
Non-Hispanic Asian	double [1]	232
Non-Hispanic Ameri...	double [1]	56

(No selection)

Name	Type	Value
sort(mn_age, decreasi...	double [10]	6498 6208 3896 3873 1057 320 ...
65-74 years	double [1]	6498
75-84 years	double [1]	6208
85 years and over	double [1]	3896
55-64 years	double [1]	3873
45-54 years	double [1]	1057
35-44 years	double [1]	320
25-34 years	double [1]	104
15-24 years	double [1]	45
5-14 years	double [1]	25
0-4 years	double [1]	16

Observation:

Maximum deaths due to heart disease is ‘6498’

Male has more deaths due to Malignant Neoplasms

Non-Hispanic white has more deaths due to Malignant Neoplasms

65-74 age group has more deaths due to Malignant Neoplasms

Challenges faced during the Final Project

The project was fun to do once the things were cleared. The first challenge was to find the right data. It took me almost 1 and a half week to finally land on the data of my interest. Next challenge was to come with the right questions. In my opinion, this was more tough than finding the data. As the data is all scattered and one can come with so many possible questions. Therefore, I decided to pull out the diseases and narrow down the scope by finding the diseases that caused highest deaths.

Finally, visualization every graph is not for all the data. Pick a right visualization for the data was a task took me 2 weeks and finalize the graphics that will be going into the Final report.

Appendix A: Important steps

Follow the following steps to run the code

- Change the directory to the location where the data file is saved.

Session > Set Working Directory > Choose Directory...

- After installing the packages make sure run the following commands

```
library(visdat)
    library(naniar)
    library(dplyr)
    library(ggplot2)
```

Appendix B: Function Definitions

Function	Definition
Head()	**head**() returns the first n rows
Tail()	**Tail**() returns the first n rows
Colnames()	**colnames**() function is used to set the names to columns of a matrix.
View()	**View()** function is used to view the data set
Sum()	**sum** returns the **sum** of all the values present in its arguments.
Count()	**count**() lets you quickly **count** the unique values of one or more variables
Vis_dat()	**vis_dat()** helps explore the data class structure and missingness
Vis_miss()	**vis_miss()**function provides a custom plot for missing data.
tapply()	**tapply**() is used to apply a function over subsets of a vector.

References

Using visdat. (n.d.). Cran. https://cran.r-project.org/web/packages/visdat/vignettes/using_visdat.html

data camp. (n.d.). Data Camp. https://www.datacamp.com/community/tutorials/make-histogram-ggplot2

dplyr tutorial. (n.d.). Dplyr Tutorial. https://genomicsclass.github.io/book/pages/dplyr_tutorial.html

naniar: Data Structures, Summaries, and Visualisations for Missing Data. (n.d.). Naniar. https://cran.r-project.org/web/packages/naniar/index.html

Health data. (n.d.). Centers for Disease Control and Prevention. https://healthdata.gov/dataset/monthly-provisional-counts-deaths-age-group-sex-and-raceethnicity-select-causes-death