Week9_FinalProject2_KoppulaVeera

Veera Koppula

August 04 2021

##Data importing and cleaning steps are explained in the text and follow a logical process. Outline your data preparation and cleansing steps. My original data set collected from CDC website has 27.9M rows. Here are the steps I have followed. I have downloaded a CSV format from the resource location https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data/vbim-akqf Following are the data columns that are present.

```
## New names:
## * `` -> ...2
## * `` -> ...3
## * `` -> ...4
## * `` -> ...5
## * `` -> ...6
```

```
## # A tibble: 13 x 1
##    ...2
##    <chr>
##  1 Variable
##  2 cdc_report_dt
##  3 cdc_case_earliest_dt
##  4 pos_spec_dt
##  5 onset_dt
##  6 current_status
##  7 sex
##  8 age_group
##  9 race_ethnicity_combined
## 10 hosp_yn
## 11 icu_yn
## 12 death_yn
## 13 medcond_yn
```

I have removed from the original dataset where data is either empty, marked as "Missing"/"Uknown" from the columns cdc_case_earliest_dt,pos_spec_dt,onset_dt,current_status,sex,age_group,race_ethinicity_combined,hosp_yn,icu and medcond_yn. Essentially I have removed all null,missing or unknown (uncollected values). This brought up the total rows down to 481,929 rows. This data has been saved into a csv format as filename "COVID-19_Case_Surveillance_Public_Use_Data.csv"

After the filtering out of the missing(uncollected valies), current_status column has same value "Labroratiry-confirmed case" in all rows. per the data dictionary provided by cdc data source, cdc recommends to use "cdc_case_earliest_dt" instead of "cdc_report_dt" also, I fee; "pos_spec_dt" which shows when first speciman was collected, has a real no bearing on the analysis of impact of confounding factors on the outcome of covid infection. as to further clean my data, I have decided to drop these 3 columns from my data set

```
##  [1] "cdc_case_earliest_dt"    "cdc_report_dt"
##  [3] "pos_spec_dt"             "onset_dt"
##  [5] "current_status"          "sex"
##  [7] "age_group"               "race_ethnicity_combined"
##  [9] "hosp_yn"                 "icu_yn"
## [11] "death_yn"                "medcond_yn"
```

```
## [1] "cdc_case_earliest_dt"    "onset_dt"
## [3] "sex"                     "age_group"
## [5] "race_ethnicity_combined" "hosp_yn"
## [7] "icu_yn"                  "death_yn"
## [9] "medcond_yn"
```

##With a clean dataset, show what the final data set looks like. However, do not print off a data frame with 200+ rows; show me the data in the most condensed form possible.

```
## [1] "cdc_case_earliest_dt"    "onset_dt"
## [3] "sex"                     "age_group"
## [5] "race_ethnicity_combined" "hosp_yn"
## [7] "icu_yn"                  "death_yn"
## [9] "medcond_yn"
```

```
## 'data.frame':    481929 obs. of  9 variables:
##  $ cdc_case_earliest_dt   : chr  "2020/12/28" "2021/01/11" "2021/01/06" "2021/01/07" ...
##  $ onset_dt               : chr  "2020/12/28" "2021/01/11" "2021/01/06" "2021/01/07" ...
##  $ sex                    : chr  "Female" "Male" "Female" "Female" ...
##  $ age_group              : chr  "10 - 19 Years" "20 - 29 Years" "50 - 59 Years" "20 - 29 Years" ...
##  $ race_ethnicity_combined: chr  "White, Non-Hispanic" "Hispanic/Latino" "Black, Non-Hispanic" "White
##  $ hosp_yn                : chr  "No" "No" "No" "No" ...
##  $ icu_yn                 : chr  "No" "No" "No" "No" ...
##  $ death_yn               : chr  "No" "No" "No" "No" ...
##  $ medcond_yn             : chr  "No" "No" "Yes" "No" ...
```

```
##   cdc_case_earliest_dt   onset_dt    sex     age_group race_ethnicity_combined
## 1           2020/12/28 2020/12/28 Female 10 - 19 Years     White, Non-Hispanic
## 2           2021/01/11 2021/01/11   Male 20 - 29 Years         Hispanic/Latino
## 3           2021/01/06 2021/01/06 Female 50 - 59 Years     Black, Non-Hispanic
## 4           2021/01/07 2021/01/07 Female 20 - 29 Years     White, Non-Hispanic
## 5           2021/01/13 2021/01/13   Male 50 - 59 Years         Hispanic/Latino
## 6           2021/01/12 2021/01/12   Male 50 - 59 Years     Black, Non-Hispanic
##   hosp_yn icu_yn death_yn medcond_yn
## 1      No     No       No         No
## 2      No     No       No         No
## 3      No     No       No        Yes
## 4      No     No       No         No
## 5      No     No       No        Yes
## 6      No     No       No        Yes
```

##What do you not know how to do right now that you need to learn to import and cleanup your dataset? I have managed to cleanup most of the data and removed the null values. I am trying to look at summarizing values to extract metadata that represents the combination of elements properly to calculate outcomes of hospitalization(hosp_yn), ICU admittance (icu_yn) and death (deat_yn)

##Discuss how you plan to uncover new information in the data that is not self-evident.

There are some complicating elements like Mask mandate, Social distancing, lockdowns and Vaccination could have overlap impact on outcome of cases. I am looking on to filterout the information that could confound the information in the collected data source, that might have got impacted outcome of the cases with same sax/age/race and comorbidities.

##What are different ways you could look at this data to answer the questions you want to answer?

I could look at this data to collect the factors that could influence serious outcomes or at the same time probably factors that would have prevented serious outcomes. However given the complex factors that could contribute tp serious outcomes (such as the social/economic status of the patient), geographical location, unknown progression of novel virus etc could probably make the summary of data as inconclusive or provide partial model that would give predictability of outcomes.

##Do you plan to slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information? Explain.

I am planning to slice the information by age, race and commodities factors to map possible outcome. I am looking other data sources to from a data frame, that could be used to overlay on the sliced summary over infection detection/onset date, to see if any possible implications factors that could have changed outcome of infections for certain patients.

##How could you summarize your data to answer key questions?

I could slice and summarize the data frame by the predictors such as sex/age and comorbities along with the date of infection identification (with possible overlay factors) to answer the key question of predictability of Hospitalization/ICU or death.

##What types of plots and tables will help you to illustrate the findings to your questions? Ensure that all graph plots have axis titles, legend if necessary, scales are appropriate, appropriate geoms used, etc.).

I am looking to do a scatterplot of the information thats sliced, overlaying with any information from the confounding datapoint identified from factors that are not captured in current data set

##What do you not know how to do right now that you need to learn to answer your questions?

finding another data set and gleaning the information to connect from the new data source that contains the factors that could influence virus spread/intesity of the infection to multiple rows in the cdc captured data set from above.

##Do you plan on incorporating any machine learning techniques to answer your research questions? Explain. I am not looking to incorporate machine learning techniques to answer my research question. However we could generate dataset that could drive a machine learning model to help predict the outcomes better in a much larger data set of 27.9M rows that we cleaned to fill out the unknowns/missing captured information.