Data Quality Report

Overview

Within this report I would outline my Initial findings within the dataset covid19-cdc-23200672, which is covid cases reported during the years from 2020-23.Our target prepare, clean and address the data integrity issues. The final column death_yn defines if the person died due to the virus or not and if so, this dataset help to summarise the relations of each columns to the death. Please do refer the appendix for column details which given all the way below.

Initial impression of the dataset is it has scope for very less continuous columns and majorly dominated by categorical columns. There are outliers and empty spaces in most of the columns and seems to be taken into account after cleaning. Moreover we have seen a series of duplicates, and its hard to determine if the duplicates are actually data entered twice as there exists no ID. And also the dataset fails a series of integrity checks as well.

Summary

There existed a significant number of duplicate rows(4661), but no columns. There exist a question of whether if they are actually duplicates or not.

Continuous: For case_positive_specimen_interval it is 136 and for case_onset_interval its 700. And also, for the same there are significant outliers as well. For case_positive_specimen_interval, and case_onset_interval it is the difference of weeks between the first positive specimen/symptoms to the earlier days.It does not make sense to have a negative number as earlier days representing the exposure comes before the testing part/symptoms. But for arguments sake we can the case was reported after the positive specimen was collected/ symptoms were shown, but as we are not clear which is the case, we are considering the negative numbers as outliers for time being and they are being treated based on the row count. The case_positive_specimen_interval had only 136 rows with negative numbers so they are dropped as they are significantly less but for case_onset_interval there were 700 rows so they are not dropped instead we set the earlier day as the day of symptom and replace the number by 0.

Categorical: Further there are columns such as process, exposure_yn, icu_yn, underlying_conditions_yn whose 90%(approx) of the data is missing. Moreover for the continuous columns there exist a series of negative numbers. There are rows such as sex, race, process, exposure_yn, symptomstatus_yn, hosp_yn, icu_yn who has some of the rows having both Missing and Unknown, which is later redifed all as Missing and taken into account while checking the null %.

There exist a series of outliers in the feature set and a detailed analysis is provided below

Review Logical Integrity

A number of additional checks will be performed to see if the data makes sense. Depending on the amount of data affected we may decide to drop those rows or replace the values upon consultation with a domain expert.

- **Test 1**: Check if any dates are of wrong format or is are lesser than the year 2020 as Covid only started by 2020

0 cases found

- **Test 2**: Check if case_positive_specimen_interval has numbers less than 0, as it seems to be an abnormality

136 cases found

- **Test 3**: Check if case_onset_interval has numbers less than 0, as it seems to be an abnormality

700 cases found

- **Test 4**: case_positive_specimen_interval defines Weeks between earliest date and date of first positive specimen collection. Normally if a person effected by covid, the virus will take its effects within weeks and will have his specimen reported as positive within atmost 7-8 weeks which is almost 2 months.Lets, take into consideraton +- a week more. Crossing to almost more than that is far less likely is what I understood from refering to the data provided by WHO.

92 cases found

- **Test 5**: case_onset_interval defines Weeks between earliest date and date of symptom onset.

Normally if a person effected by covid, the virus will take its effects within weeks and will have his symptoms reflected within atmost 7-8 weeks which is almost 2 months.Lets, take into consideraton +- a week more. Crossing to almost more than that is far less likely is what I understood from refering to the data provided by WHO.

38 cases found

- **Test 6**: If icu_yn is yes then hosp_yn should not be no, as ICU is in the hospital

6 cases found

- **Test 7**: Process who is under clinical_evaluation or lab reported cant have a current status of probable_case

573 cases found

- **Test 8**: Does exposure_yn has any no

0 cases found

- **Test 9**. If exposure_yn is yes , current_status has to be either Laboratory-confirmed case or Probable case

0 cases found

- **Test 10**. If symptom_status is Asymptomatic then icu_yn cannot be yes, he has to atleast show some symptoms of covid to be in intensive care unit

8 cases found

- **Test 11**. If symptom_status is Asymptomatic then hosp_yn cannot be yes, he has to atleast show some symptoms of covid to be hospitalized

55 cases found

Review Continuous Features

1. **Negative values in case_positive_specimen_interval and case_onset_interval**
   - There exist negative values within both these columns and even though it creates a confusion on weather this is an irregular data or if it was because the recording was done after the cases.For timebeing we are not doing anything for the negative numbers
2. **Empty values for case_positive_specimen_interval and case_onset_interval**
   - There exist empty section for these two columns and this creates significant effects on the data, thus I am adding the mean to cover the vacent sections for better data manipulation as the type being continuous.

3. Outliers in case_positive_specimen_interval and case_onset_interval

- There clearly exist outliers within both case_positive_specimen_interval and case_onset_interval columns. They initailly look somewhat plausible but will need to be investigated further. If they don't make sense they will be removed.

Histogram - Continuous Data

- **case_positive_specimen_interval**: There are clearly outliers or abnormalities in the plot data in case_positive_specimen_interval, as the min is -109.0 and the max is 142.0.
- **case_onset_interval**: Outliers are found in the plot data as there exist a deviation from -73.0 as min to 65.0 as max

Box plot - Continuous Data

- **case_positive_specimen_interval**: Here the data summarizes the first quartile and the third quartile as zero making it having no body for the box plot.
- **case_onset_interval**: Here the data summarizes the first quartile and the third quartile as zero making it having no body for the box plot.

Review Categorical Features

- **Columns such as state_fips_code and county_fips_code plays very less role**
    - There are a clear mismatch between county_fips_code and res_county columns as both have diffent amount of unique values.And there is a question of the need of fibs_code columns, Needs further investigation.
- **Both the terms Unknown and Missing in some of the columns**
    - There are columns whose values are both Unknown or Missing, thinking on this both means the same. The columns are sex,race,process,exposure_yn,symptomstatus_yn,hosp_yn,icu_yn.Better option would be to replace all the 'Unknown' as 'Missing'
- **If icu_yn is yes then hosp_yn should not be no**
    - The person would be hospitalized to be in the icu, so those rows seems to have an abiguity, would be better to drop them.
- **Process who is under clinical_evaluation or lab reported cant have a current status of probable_case**
    - If the process is clinical_evaluation then its a definite case and not a probable case, so those rows seems to have an ambiguity, would be better to drop them.
- **If symptom_status is Asymptomatic then icu_yn need not be yes, person has to atleast show some symptoms of covid to be in intensive care unit**
    - You would not be moved to ICU until you have atleast a symptom, so those rows seems to have an ambiguity, would be better to drop them.
- **If symptom_status is Asymptomatic then icu_yn cannot be yes, person has to atleast show some symptoms of covid to be hospitalized**

  - You would not be admitted to the hospital without any symptoms usually, there is a minimum criteria to require hospital care, so those rows seems to have an ambiguity ,would be better to drop them.

Bar plot - Categorical Data

- case_month : Total count of 45339 among with 40 unique data the more than 5000 happens on 2022-01, no outliers
- res_state : Total count of 45338 among with 49 unique data where most commonly found cases in Ny, no outliers
- state_fips_code : Total count of 45338.0 among with 49.0 unique data where fibs code of Ny which is 36 the highest, no outliers,
- res_county : Total count of 45338.0 among with 963 unique data where MIAMI-DADE the highest, no outliers
- county_fips_code : Total count of 45338.0 among with 1385.0 unique data where 12086.0 the highest, no outliers
- age_group : Highest found between age group 18 to 49 years, 17939 cases found
- sex : Cases reported among Female the highest, 23107 cases.

- race : Cases reported to mostly white with the total cases of 27609.
- ethnicity : Highest reported among Non-Hispanic/Latino about 27135 cases.
- process : Minimal data on the process done to report the virus and there are a total of 40960 Missing data
- exposure_yn : Minimal data regrading exposure as well as there is a total of 38566 Missing data
- current_status : There are 37894 cases which are Laboratory-confirmed case after testing
- symptom_status : Amongst the effected crowd 21562 were Symptomatic
- hosp_yn : Majority of the effected people was not hospitalized , about 23619 were non hospitalized
- icu_yn : Total of 35193 Missing data regarding ICU status
- death_yn : About 36657 survived and recovered
- underlying_conditions_yn : Cases of underlining conditions are higher with about 4019 people having them

Action to take

1. county_fibs_code
   - Drop the entire column due to outliers.
2. res_county
   - 5% so Do Nothing due to Null Values or empty cells.
3. sex
   - 2% so Do Nothing due to Null Values or empty cells.
   - Replaced all Unknown to Missing due to Has both Missing and Unknown.
4. race
   - Do Nothing due to Null Values or empty cells.
   - Replaced all Unknown to Missing due to Has both Missing and Unknown.
5. ethnicity
   - Do Nothing due to Null Values or empty cells.
6. process
   - Do Nothing due to Null Values or empty cells.
   - Replace with null due to Outliers.
   - Replaced all Unknown to Missing due to Has both Missing and Unknown.
7. case_positive_specimen_interval
   - Replace with interpolated mean due to Null Values or empty cells.
   - Replace with null due to Negative values.
8. case_onset_interval
   - Replace with interpolated mean due to Null Values or empty cells.

- Replace with 0 due to Negative values.
- Replace with interpolated mean due to Outliers.

9. exposure_yn
    - Do Nothing due to Null Values or empty cells.
    - Remove Outliers due to Outliers.
    - Replaced all Unknown to Missing due to Has both Missing and Unknown.
10. hosp_yn
    - Do Nothing due to Null Values or empty cells.
    - Replace with null due to Outliers.
    - Replaced all Unknown to Missing due to Has both Missing and Unknown.
11. icu_yn
    - Do Nothing due to Outliers.
    - Do Nothing due to Null Values or empty cells.
    - Replaced all Unknown to Missing due to Has both Missing and Unknown.
12. current_status
    - Replace with null due to Outliers.
13. symptom_status
    - Do Nothing due to Outliers.
    - Do Nothing due to Null Values or empty cells.
    - Replaced all Unknown to Missing due to Has both Missing and Unknown.
14. underlying_conditions_yn
    - Do Nothing due to Null Values or empty cells.

Appendix: Description of Variables

case_month

- The earlier of month the Clinical Date (date related to the illness or specimen collection) or the Date Received by CDC
- Data Type: Plain Text

res_state

- State of residence
- Data Type: Plain Text

state_fips_code

- State FIPS code
- Data Type: Plain Text

res_county

- County of residence
- Data Type: Plain Text

county_fips_code

- County FIPS code
- Data Type: Plain Text

age_group

- Age group [0 - 17 years; 18 - 49 years; 50 - 64 years; 65 + years; Unknown; Missing; NA, if value suppressed for privacy protection.]
- Data Type: Plain Text

sex

- Sex [Female; Male; Other; Unknown; Missing; NA, if value suppressed for privacy protection.]
- Data Type: Plain Text

race

- Race [American Indian/Alaska Native; Asian; Black; Multiple/Other; Native Hawaiian/Other Pacific Islander; White; Unknown; Missing; NA, if value suppressed for privacy protection.]
- Data Type: Plain Text

ethnicity

- Ethnicity [Hispanic; Non-Hispanic; Unknown; Missing; NA, if value suppressed for privacy protection.]
- Data Type: Plain Text

case_positive_specimen_interval

- Weeks between earliest date and date of first positive specimen collection
- Data Type: Number

case_onset_interval

- Weeks between earliest date and date of symptom onset.
- Data Type: Number

process

- Under what process was the case first identified? [Clinical evaluation; Routine surveillance; Contact tracing of case patient; Multiple; Other; Unknown; Missing]
- Data Type: Plain Text

exposure_yn

- In the 14 days prior to illness onset, did the patient have any of the following known exposures: domestic travel, international travel, cruise ship or vessel travel as a passenger or crew member, workplace, airport/airplane, adult congregate living facility (nursing, assisted

living, or long-term care facility), school/university/childcare center, correctional facility, community event/mass gathering, animal with confirmed or suspected COVID-19, other exposure, contact with a known COVID-19 case? [Yes, Unknown, Missing]
- Data Type: Plain Text

current_status

- What is the current status of this person? [Laboratory-confirmed case, Probable case]
- Data Type: Plain Text

symptom_status

- What is the symptom status of this person? [Asymptomatic, Symptomatic, Unknown, Missing]
- Data Type: Plain Text

hosp_yn

- Was the patient hospitalized? [Yes, No, Unknown, Missing]
- Data Type: Plain Text

icu_yn

- Was the patient admitted to an intensive care unit (ICU)? [Yes, No, Unknown, Missing]
- Data Type: Plain Text

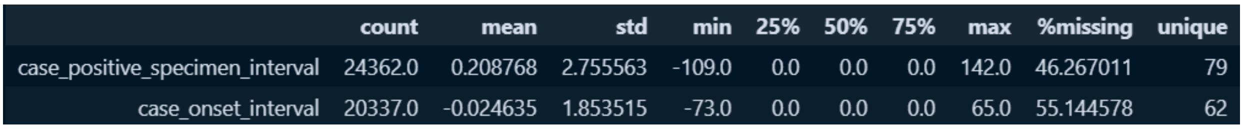
death_yn

- Did the patient die as a result of this illness? [Yes; No; Unknown; Missing; NA, if value suppressed for privacy protection.]
- Data Type: Plain Text

underlying_conditions_yn

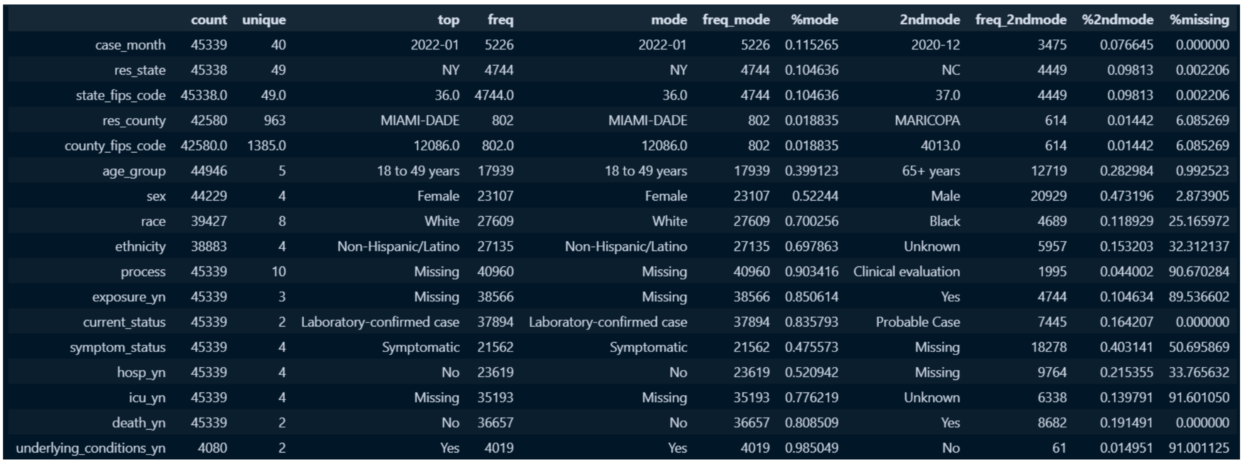
- Did the patient have one or more of the underlying medical conditions and risk behaviors: diabetes mellitus, hypertension, severe obesity (BMI>40), cardiovascular disease, chronic renal disease, chronic liver disease, chronic lung disease, other chronic diseases, immunosuppressive condition, autoimmune condition, current smoker, former smoker, substance abuse or misuse, disability, psychological/psychiatric, pregnancy, other. [Yes, No, blank]
- Data Type: Plain Text

Continuous Statistics



	count	mean	std	min	25%	50%	75%	max	%missing	unique
case_positive_specimen_interval	24362.0	0.208768	2.755563	-109.0	0.0	0.0	0.0	142.0	46.267011	79
case_onset_interval	20337.0	-0.024635	1.853515	-73.0	0.0	0.0	0.0	65.0	55.144578	62

Categorical Statistics



	count	unique	top	freq	mode	freq_mode	%mode	2ndmode	freq_2ndmode	%2ndmode	%missing
case_month	45339	40	2022-01	5226	2022-01	5226	0.115265	2020-12	3475	0.076645	0.000000
res_state	45338	49	NY	4744	NY	4744	0.104636	NC	4449	0.09813	0.002206
state_fips_code	45338.0	49.0	36.0	4744.0	36.0	4744	0.104636	37.0	4449	0.09813	0.002206
res_county	42580	963	MIAMI-DADE	802	MIAMI-DADE	802	0.018835	MARICOPA	614	0.01442	6.085269
county_fips_code	42580.0	1385.0	12086.0	802.0	12086.0	802	0.018835	4013.0	614	0.01442	6.085269
age_group	44946	5	18 to 49 years	17939	18 to 49 years	17939	0.399123	65+ years	12719	0.282984	0.992523
sex	44229	4	Female	23107	Female	23107	0.52244	Male	20929	0.473196	2.873905
race	39427	8	White	27609	White	27609	0.700256	Black	4689	0.118929	25.165972
ethnicity	38883	4	Non-Hispanic/Latino	27135	Non-Hispanic/Latino	27135	0.697863	Unknown	5957	0.153203	32.312137
process	45339	10	Missing	40960	Missing	40960	0.903416	Clinical evaluation	1995	0.044002	90.670284
exposure_yn	45339	3	Missing	38566	Missing	38566	0.850614	Yes	4744	0.104634	89.536602
current_status	45339	2	Laboratory-confirmed case	37894	Laboratory-confirmed case	37894	0.835793	Probable Case	7445	0.164207	0.000000
symptom_status	45339	4	Symptomatic	21562	Symptomatic	21562	0.475573	Missing	18278	0.403141	50.695869
hosp_yn	45339	4	No	23619	No	23619	0.520942	Missing	9764	0.215355	33.765632
icu_yn	45339	4	Missing	35193	Missing	35193	0.776219	Unknown	6338	0.139791	91.601050
death_yn	45339	2	No	36657	No	36657	0.808509	Yes	8682	0.191491	0.000000
underlying_conditions_yn	4080	2	Yes	4019	Yes	4019	0.985049	No	61	0.014951	91.001125

For Graphs and Plots are attached as reference along with this PDF attached