| **COVID-19 Survey - Data Definition and Cleaning Steps** | | | |
| --- | --- | --- | --- |
| **General challenges faced:**   1. Finding what could be the unique id – the given dataset doesn’t have a unique ID since it’s a survey distributed among the public and not taken from a hospital. 2. Brainstormed on whether to remove duplicates considering the potential of losing information as the data points may not be from the same responder. 3. Deciding on how to group columns with more than one value. 4. Finding the week number’s equivalence to month in month column. 5. Validating probable and vulnerable columns based on sanitization script. 6. How to classify records based on Schema. We found this by grouping the age & month column. Data in Schema 1- 263639,Schema 2-14,932 & Schema 3 - 15,534. 7. How to optimize the number of Applied Steps to minimal steps. (For e.g., Using Advance editor, we reduced multiple replacements in all the columns; Created custom functions to clean month, symptoms, conditions columns). | | | |
| **Original/New Column Name** | **Data Definition** | **Observations/Key Findings** | **Applied/Cleaning Steps** |
| **Common changes done:**   * Imported the CSV file and started cleaning the dataset using Transform Data in **Power BI.** * Removed Blank Rows if any using Home → Reduce Rows → Remove Rows → Remove Blank Rows. * Renamed the column names to a common nomenclature format. | | | |
| **Schema(New column)** | This column gives the details about which schema the row belongs to. | Observed that age for schema 1 was recorded as y and n. For schema1 and 2, the month was recorded as week number and for schema 3 the month was recorded as the name of the month. | Created a custom column based on the condition: If age is y or n then schema 1, schema 2 is based on week number(15,16) and schema 3 is based on month name. |
| week\_month(Month) | This column specifies the month/week number when the survey was filled. Dataset is based on survey responses collected from 294,105 Canadians from March 23rd until July 30th 2020. | The column has a week number and month name as data. Need to change to a single month format. | Created a new function(fnDateFromWeekNumber) that takes week-number as input parameter and returns a month. |
| fsa(FSA) | Forward Sortation Area (FSA) is a way to designate a geographical unit based on the first three characters in a Canadian postal code. The first character in FSA identifies the province or territory, second character identifies whether the area is urban or rural, third character is a letter that, in combination with the first two characters, identifies geographic districts. | 14932 blanks were found which are from Schema 2. |  |
| probable(Probable) | Probable cases are those cases that are not positive by a confirmatory lab test method but have met one of the following: a) have a positive test result using an antigen test method,b) have symptoms of COVID-19 AND known exposure to COVID-19 and no molecular or antigen test was performed or c) have COVID-19 listed on the death certificate. | Values are y/n. We cross verified the data using the formula: **q3 OR (q1 AND (q2 OR q6)) OR (q6 AND (q2 OR q3)) OR q7** from sanitisation script. | Validated fever/chills/shakes, cough, shortness of breath columns and made sure the data is correct. Then used the columns in the formula. There was a mismatch of 324 records and we replaced the values to the correct value. Changed values to 1/0 and data type changed to whole number. |
| vulnerable(Vulnerable) | A survey respondent was considered to be “vulnerable” to COVID-19 if they were aged 65 or more or if they had at least one of the following: diabetes, cancer, high BP, heart disease, asthma or other breathing-related illness, immunocompromising condition, kidney disease, history of stroke. | Values are y/n. We cross verified the data using the algorithm to determine vulnerable: **q4 OR q5** from the sanitisation script. | Validated any\_medical\_conditions column based on the condition: **(response\_dict['conditions'] == [] or response\_dict['conditions'] is not ['other']).** Updated to the correct value. Used any\_medical\_conditions and age columns to verify vulnerable columns. |
| fever\_chills\_shakes(Fever\_Chills\_Shakes) | Fever, Chills and Shakes are common symptoms of COVID 19. | The column specifies whether the person has/had one of these or not. | Cross verified the data against the symptoms column & changed values to 1/0 and data type to whole number. |
| cough(Cough) | A cough is the most common symptom which is often a dry cough, although phlegm can also be present. | The column specifies whether the person has/had a cough or not. | Cross verified the data against the symptoms column & changed values to 1/0 and data type to whole number. |
| shortness\_of\_breath(Shortness\_Of\_Breath) | Shortness of breath is a common COVID symptom It can appear in both mild/severe COVID infections. | The column specifies whether the person has/had Shortness of breath. | Cross verified the data against the symptoms column & changed values to 1/0 and data type to whole number. |
| any\_medical\_conditions(Any\_Medical\_Conditions) | The data in this column specifies if the survey participant has any underlying medical condition or not. | Verify against the condition **(response\_dict['conditions'] == [] or response\_dict['conditions'] is not ['other']).** | Replaced the mismatch values.Changed values to 1/0 and data type to whole number |
| travel\_outside\_canada(Travel\_Outside\_Canada) | This column specifies if the survey participant has traveled outside of Canada in the past 14 days. |  | Changed values to 1/0 and data type to whole number |
| contact\_with\_illness(Contact\_With\_Illness) | The data in this column specifies if they suspect they have been in contact with someone infected with COVID-19 in the past 14 days. | Verify contact\_with\_illness with contact\_in\_household column.When y in household, make it y in illness and remove  contact\_in\_household column after merging the data. | Updated the mismatch records. Changed values to 1/0 and data type to whole number |
| contact\_in\_household(Removed) | This column specifies if the survey respondent was in contact with the people in the house | Removed this column. |
| tested(Tested) | This column refers if the survey respondent had been tested using laboratory test method |  | Changed values to 1/0 and data type to whole number |
| covid\_results\_date(Covid\_Results\_Date) | This column refers to when the survey respondent got to know about the result. |  | Changed column name with respect to convention. |
| covid\_positive(Covid\_Status) | This column refers to survey participant test results. | **SANITISATION\_MAPPINGS** = {"yes": "y", "no": "n", "positive": "y", "negative": "n", "awaiting\_results": "n","male": "m", "female": "f", "preferNotToRespond": "na"} | Used the observation de and changed positively to positive, negatively to negative, n to Awaiting\_Results. |
| symptoms(Symptoms) | This column refers to the symptoms the survey respondent has/had while they filled the survey. | Found redundant data as fever,chills,shakes, cough,shortness of breath has its own columns. | Removed fever, chills, shakes,cough, shortness of breath and none of the above. Capitalized each word. |
| conditions(Conditions) | This column refers to the health problem they had/has while they filled the survey. | Found a few none of the above values which we don’t need. | Removed none of the above. Capitalized each word. |
| ethnicity(Ethnicity) | Ethnicity is used as a matter of cultural identity of a group, often based on shared ancestry, language, and cultural traditions. | This column has notable columns with more than one value in a cell which can be grouped to a single value "mixed\_group". | Grouped and renamed all mixed ethnicity values under one single category **mixed\_ethnic**. Capitalized each word. |
| sex(Sex) | Denotes the Gender |  | Changed f to Female, m to Male and na to NA. |
| needs(Needs) | Pandemic-induced most pressing needs (i.e., food, medical, financial, emotional, other). |  | Capitalized each word. |
| age\_1(Age) | Respondent’s age at the time of survey. | Based on Physionet, we infer that the Schema 1 survey did not have minute details. So the values in age are based on the Q4. Are you 60 or older?. We cross verified ages with values y and n with the month column. | Changed y to >65 and n to <65 to align with the already existing condition “>65” in the dataset. |
| mental\_health\_impact(Mental\_Health\_Impact) | Covid and quarantine affected both physical and mental health. So it's very important to consider how many people have an impact on their mental health. | This column refers to the person’s mental health impact during the survey during covid. | Changed positively to positive, negatively to negative, noImpact to NoImpact. |
| travel\_work\_school(Travel\_Status) | This column specifies whether the person had traveled before covid or not and if they need to travel. | Based on the observation, we see more travel categories. | Grouped them into Y/N. |
| self\_isolating(SelfIsolation\_Status) | This column specifies whether the person is in isolation or not. |  | Changed values to Y/N. |
| media\_channels(Active\_On\_SocialMedia) | This column specifies how the person is socially active and the medium through which they are active. | Considering all categories as media,planning to make it into one category as Y. | All available values were changed to Y and blanks are left as it is. |
| financial\_obligations\_impact(Financial\_Obligations\_Impact) | This column specifies the financial obligations impact on the responders. | Considering all categories,planning to make it into one category as Y. | Changed the values to Y in case of financial needs. |
| tobacco\_usage(Removed) | The data in this column specifies if the survey participant is using tobacco or not. | This column has very few values which cannot be used for analysis | Removed the column. |
| is\_most\_recent(Most\_Recent\_Survey) | A survey response is considered most recent if it is the most recent submission from the respondent. |  | Changed values to Y/N. |