DSC530: Project Milestone 1

Conditions contributing to deaths involving Covid19

For this project, I have decided to work with Covid19 data. Given the high impact this pandemic has had over all of us, I am confident that I can find plenty of data and interesting questions that I would like to answer.

I was able to find a data set in the CDC website that I would like to work with. This data set contains data about conditions contributing to deaths involving coronavirus cases. The most up-to-date information on this data has 12.4k rows and 11 columns. The columns are described as follows:

| Column Name | Description | Type |
| --- | --- | --- |
| **Data as of** | Date of analysis | Date & Time |  |
| **Start Week** | First week-ending date of data period | Date & Time |  |
| **End Week** | Last week-ending date of data period | Date & Time |  |
| **State** | Jurisdiction of occurrence | Plain Text |  |
| **Condition Group** | Condition Group | Plain Text |  |
| **Condition** | Condition contributing to deaths involving COVID-19 | Plain Text |  |
| **ICD10\_codes** | ICD-10 code for condition | Plain Text |  |
| **Age Group** | Age group | Plain Text |  |
| **COVID-19 Deaths** | COVID 19 Deaths | Number |  |
| **Number of Mentions** | Number of mentions | Number |  |
| **Flag** | Counts less than 10 suppressed. | Plain Text |  |

There are a few things to take into account when analyzing this data set. It is important to have assumptions clear up front. Luckily, the CDC website has some clear statements regarding Covid19 data that I think are worth mentioning:

* Number of conditions reported in this table are tabulated from deaths received and coded as of the date of analysis and do not represent all deaths that occurred in that period. Data during this period are incomplete because of the lag in time between when the death occurred and when the death certificate is completed, submitted to NCHS and processed for reporting purposes. This delay can range from 1 week to 8 weeks or more. Conditions contributing to the death were identified using the International Classification of Diseases, Tenth Revision (ICD-10). Deaths involving more than one condition (e.g., deaths involving both diabetes and respiratory arrest) were counted in both totals. To avoid counting the same death multiple times, the numbers for different conditions should not be summated. Deaths with confirmed or presumed COVID-19, coded to ICD–10 code U07.1
* CDC reports aggregate counts of COVID-19 cases and death numbers daily online. Data on the COVID-19 website and CDC’s COVID Data Tracker are based on these most recent numbers reported by states, territories, and other jurisdictions. This data set of “United States COVID-19 Cases and Deaths by State over Time” combines this information. However, data are dependent on jurisdictions’ timely and accurate reporting.
* CDC tracks COVID-19 illnesses, hospitalizations, and deaths to track trends, detect outbreaks, and monitor whether public health measures are working. However, counting exact numbers of COVID-19 cases is not possible. COVID-19 can cause mild illness, symptoms might not appear immediately, there are delays in reporting and testing, not everyone who is infected gets tested or seeks medical care, and there are differences in how completely states and territories report their cases.
* COVID-19 is one of about 120 diseases or conditions health departments voluntarily report to CDC. State, local, and territorial public health departments verify and report cases to CDC. When there are differences between numbers of cases reported by CDC versus by health departments, data reported by health departments should be considered the most up to date. Health departments may update case data over time when they receive more complete and accurate information. The number of new cases reported each day fluctuates. There is generally less reporting on the weekends and holidays.

My plan for this data set is to start by saving it in a csv file and reading it in Python. I will then run some basic cleaning procedures for blanks, duplicates, outliers, and incomplete data. Given that this is a CDC data set, I presume it will be relatively “clean”, but I will still go through the regular cleaning process and make sure it is as good as needed for my analysis. After that, I would like start running some basic statistical functions in Python to get a summary of my data set and the different counts. I will create scatter plots and histograms as needed and attempt to identify any relevant information. I will attempt to identify if the data follows a particular distribution and see if I can model some of the variables.

I am curious to answer the following questions:

1. Is there an underlying health condition that appears more than other conditions in covid19 diseased patients?
2. Is diabetes, for example, a primary health condition present in patients who have died while infected (or presumably infected) by covid19?
3. In the winter months, has there been an increase in respiratory conditions associated with infected (or presumably infected) covid19 patients?
4. Are there health conditions more commonly present in certain age groups than others?
5. Are there health conditions more commonly present in certain states than others?

I am sure that as I go through some of the steps, I might come up with other questions I would like to answer and at the same time I might not be able to answer all the questions I have stated above. However, I am really curious to see how much information and insight I can get from this data set. I am also looking forward to creating some nice visualizations to present my findings at the end of my analysis.

References:

https://data.cdc.gov/NCHS/Conditions-contributing-to-deaths-involving-corona/hk9y-quqm/data