**A picture containing jewelled headdress, chandelier

Description automatically generated**

STUDENT ID: GAU20508418

DATA SCIENCE

COURSEWORK- 1

DAtaset: COVID-19 SEVERITY CHECKER

name: MEHRAJ HARISH GAUD

**TABLE OF CONTENTS**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SR NO. | | CONTENT | PAGE NO. | |
| 1. | **ABOUT DATASET** | | | **2** |
| 2. | | **HYPOTHESIS** | **2** | |
| 3. | **DATA PRE-PROCESSING** | | | **3** |
| 4. | | **TECHNIQUE USED** | **5** | |

**ABOUT DATASET:**

COVID-19 also known as coronavirus was first identified in December 2019 in Wuhan, China. It is a highly contagious respiratory illness caused by the novel coronavirus SARS CoV 2. It can spread in numerous ways such as an infected person talking, coughing, and sneezing. Also, by touching infected areas and then using the same hands to touch one’s nose, mouth, eyes, and ears. COVID-19 has been one of the most fatal outspreads in history accounting for millions of deaths around the globe.

In this report, we will discuss the dataset about COVID-19 which covers all the possible signs and symptoms of the disease, the severity of the disease, and many other related factors. According to WHO, there are five main symptoms of COVID-19 fever, dry cough, difficulty in breathing, sore throat, and tiredness. This data set is from the following URL:

<https://www.kaggle.com/datasets/iamhungundji/covid19-symptoms-checker>

and was published by **Bilal Hungund**.

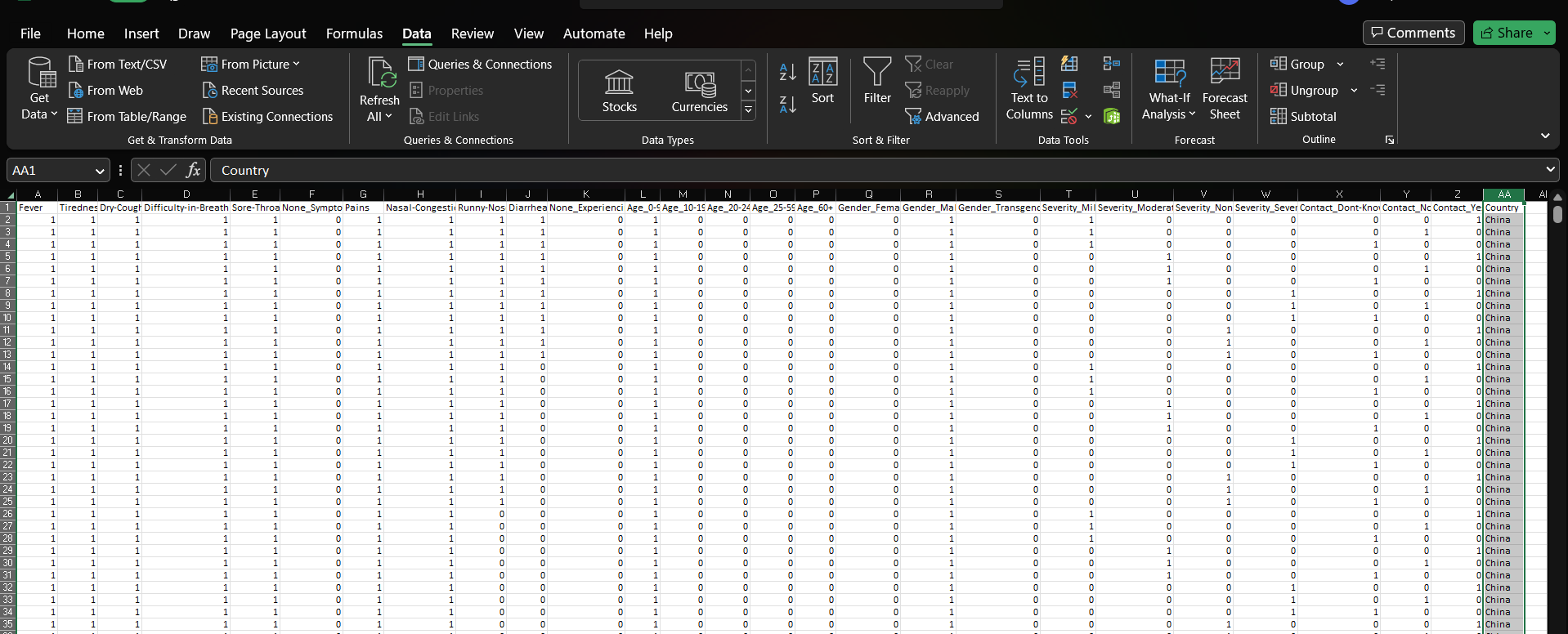
**HYPOTHESIS:**

|  |  |
| --- | --- |
| GOAL | Predict the severity of COVID-19. |
| HYPOTHESIS 1 | Medical experts are trying to find which age and gender are most prone to the severity of this disease and can alert those aged and gender to prevent the spread of the disease and can provide them with regular self-check-up kits. |
| HYPOTHESIS 2 | Medical experts are trying to find the major symptoms which can eventually lead to severe COVID, and therefore can’t be ignored. |
| HYPOTHESIS 3 | Medical experts are also trying to determine whether the virus is spreading through contact so that contact with things and people can be reduced or controlled through different measures like wearing masks, gloves, glasses, sanitizing, etc. |

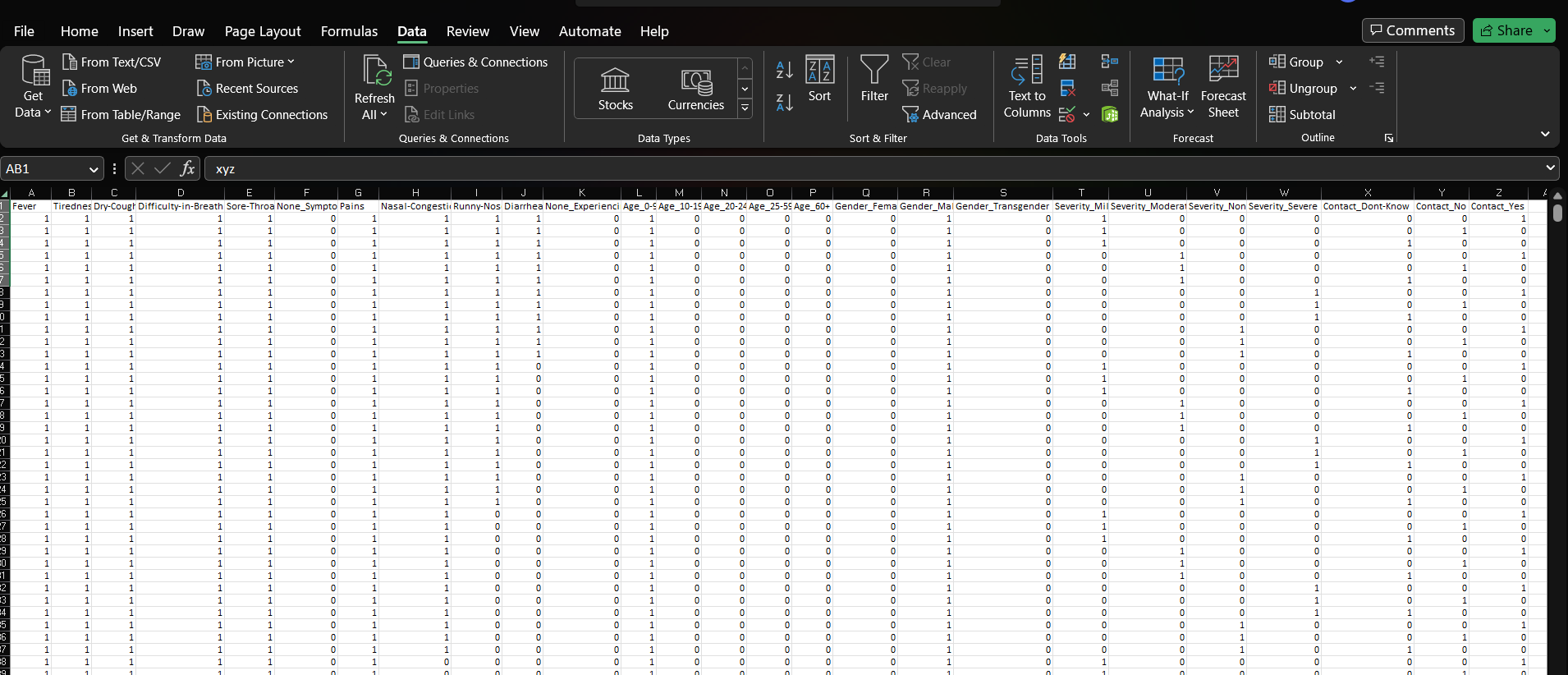
**DATA PRE-PROCESSING:**

In this dataset, we use several pre-processing steps to remove unnecessary data which might not be useful for our analysis or to conclude what we want. So, the following are the steps taken:

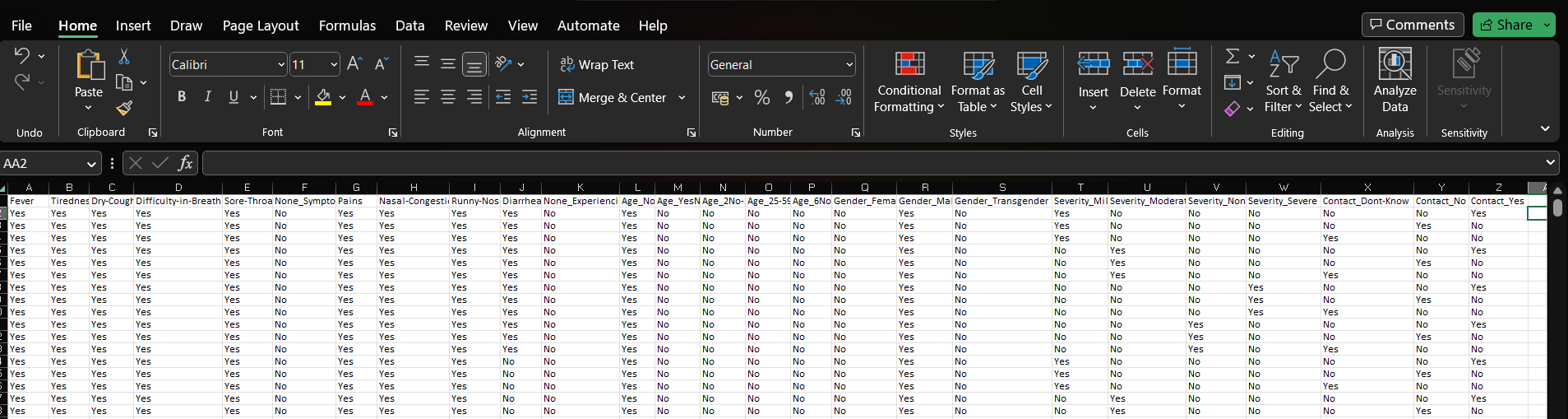
1. I have removed the country column from the dataset as it didn’t help me predict the necessary result.



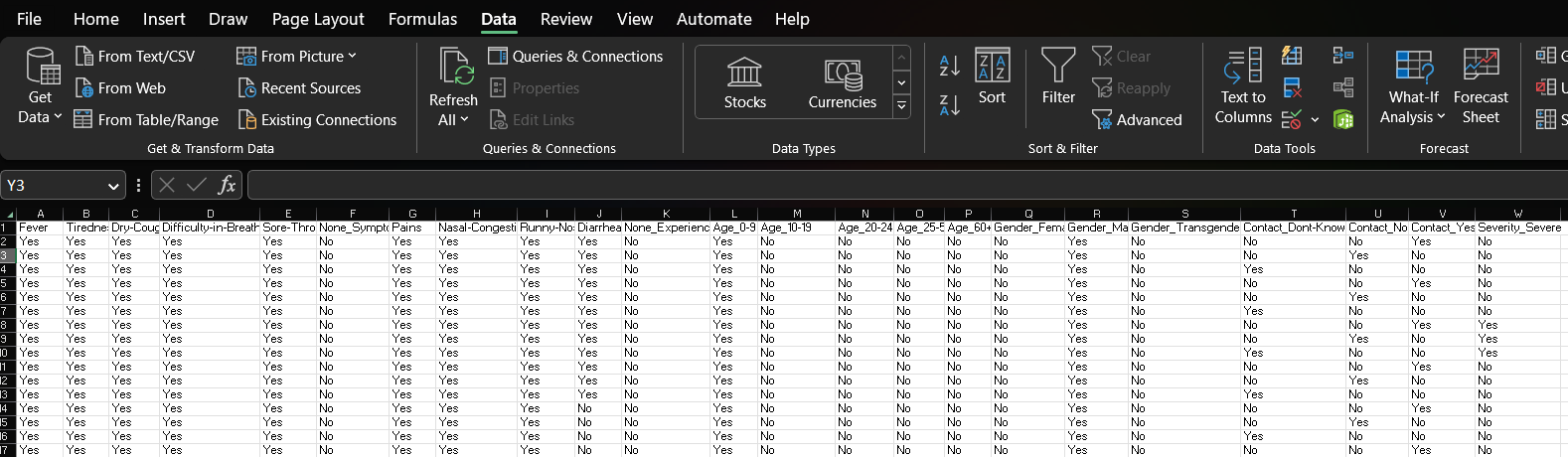
1. I normalized the data by filling up some of the empty remaining columns to make it normal.



1. I have changed all the 1’s to Yes and all the 0’s to No for a better understanding of the user.



1. I have removed the mild, moderate, and none severity columns leaving the severe column only used for prediction as there can only be one labelled column that can be used for prediction.



**TECHNIQUE -USED:**

To achieve the solution for the proposed hypothesis, we can use a Supervised Machine Learning Technique called Naïve Bayes algorithm. This technique uses probabilistic measurement to predict the result based on the previous conditions and events that took place. I choose this technique for this dataset because it is categorical data that can be best solved or predicted using probabilistic measures. The predictors used in this dataset have binary values which are yes or no values with which we can infer whether the person having such signs and symptoms has severe COVID or not.

Diagram

Description automatically generated

**Hypothesis 1:**

Talking about the first hypothesis, we can use this technique to find out which gender or age group is more prone to the severity of this disease using Naïve Bayes classifier/ conditional probability based on the age group and gender data given where we can see ‘Yes’ and ‘No’ under their column. And when we are given such gender or age group, we can predict their severity using their previous probabilistic measure.

**Hypothesis 2:**

Considering the second hypothesis where we are trying to find the symptoms that lead to severe COVID which can be fatal for any age group and gender. We can calculate the probabilistic occurrence of different signs and symptoms from the given data and can find out which sign(s) or symptom(s) has a major impact on the severity of COVID so that when we are given those signs and symptoms, we can use this probability to predict the severity.

**Hypothesis 3:**

Finally, we want to predict whether or not coming in contact with different things and people can lead to severe COVID or can contact be one of the major factors to spread COVID, and if the probability for the spread is higher due to contact, then the government can give out rules and regulations to wear equipment that can avoid contact and stop the spread of the virus and use of appropriate sanitizing liquids to kill the virus.