DATA VISUALIZATION AND ANALYSIS OF THE EFFECTS OF CORONAVIRUS IN CHINA

BY SHYAM SANKAR

FEB 20, 2020

1.INTRODUCTION

1.1 BACKGROUND

Coronaviruses (CoV) are a large family of viruses that cause illness ranging from the common cold to more severe diseases such as Middle East Respiratory Syndrome and Severe Acute Respiratory Syndrome. A novel coronavirus is a new strain that has not been previously identified in humans.

Coronavirus outbreak is an ongoing epidemic of coronavirus disease 2019 (COVID-19) caused by SARS-CoV-2, which started in December 2019. It was first identified in Wuhan, capital of Hubei province China, after 41 people presented with pneumonia of no clear cause. It can spread between people, with the time from exposure to onset of symptoms generally between 2 and 14 days. Symptoms may include fever, cough, and shortness of breath. Complications may include pneumonia and acute respiratory distress syndrome. There is no vaccine or specific antiviral treatment, with efforts typically to manage symptoms and support functioning. Hand washing is recommended to prevent the spread of the disease. Anyone who is suspected of carrying the virus is advised to monitor their health for two weeks, wear a mask, and seek medical advice by calling a doctor rather than directly visiting a clinic.

1.2 PROBLEM

Coronavirus is considered to have a spreading rate that is similar to that of common-flu. The virus can spread easily through fluids. This fast spreading rate builds a curiosity in me to question a relation between the coronavirus confirmed rates in different provinces of China and the population density of the respective provinces.

1.3 INTEREST

Coronavirus is a fast spreading epidemic that has struck us hard as we enter into a new decade. We as data scientists are capable of contributing a lot in making people aware of its effects and providing analytic information to whoever that needs it, whether it be research or just to gain knowledge about it.

2. DATA ACQUISITION AND CLEANING

2.1 DATA SOURCES

The major data set that has been used for the analysis was acquired from kaggle datasets(here). However data regarding the population and population density of provinces in Mainland China had to be collected from other sources. The data on population of each province was collected from here. The population density of each province/state was collected from here.

2.2.DATA CLEANING

The initial csv file acquired from kaggle contained data on province,country, update date, confirmed cases of coronavirus, suspected, recovered, deaths. Information on population density and province names in Chinese were added into the csv file before loading into the notebook and converted into pandas dataframe.

The data frame was filtered to only keep rows containing information as acquired from the latest date ie , the 26th of January 2020.

The "Nan" values in the confirmed, recovered, suspected and deaths columns were updated with zero as an approximation for unhindered analysis.

Since analysis was only being carried out for provinces that are part of mainland china the other columns that had country value set to any other was removed.

The above mentioned second dataset was used to add the longitudes, latitudes and the population of each province in the country.

3. VISUALIZED DATA